eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Departamento de Lenguajes y Sistemas Informáticos
Lengoaia eta Sistema Informatikoak Saila
Department of Computer Languages and Systems

A proposal for the management of data-driven services in Smart Manufacturing scenarios

Mikel Niño Bartolomé

Supervised by

José Miguel Blanco Arbe &
Arantza Illarramendi Echave

A dissertation submitted to the Department of Computer Languages
and Systems of the University of the Basque Country UPV/EHU
for the degree of Doctor of Philosophy (Ph.D.) in Informatics Engineering

Donostia-San Sebastián, May 2017

A mi familia

Acknowledgments

I would like to begin this dissertation expressing my gratitude to several people who have played a very important role in the completion of this PhD thesis.

First of all, I wanted to thank my supervisors, José Miguel Blanco and Arantza Illarramendi, for their continuous support and for giving me plenty of research independence. It has been a pleasure and privilege to have their advice and supervision throughout the whole PhD process.

Many thanks to all my colleagues in the BDI research group, especially to Kevin Villalobos for his invaluable help in the field testing of data reduction techniques conducted in this research work. Thanks also to my colleagues here in the Department of Computer Languages and Systems and also in the University of La Rioja, and very special thanks to my colleagues in the Frankfurt Big Data Lab for accepting me as a guest researcher, for their hospitality and their help during my stay and afterwards.

This work would not have been possible without the collaboration of the industrial partners involved in this project. Very special thanks to Fernando, Ángel, Esti and Dominique for their availability and tremendous generosity.

Finally, my biggest gratitude and love goes to my friends, my family, and especially to Paola for their unconditional care and affection and for accompanying me throughout this journey.

Funding

The author of this dissertation wants to thank the Spanish Ministry of Economy and Competitiveness for their support funding this research.

Contents

Chapter 1

Introduction

The so-called *Big Data* and, by extension, data processing and exploitation technologies constitute one of the most relevant global trends in Information Technologies (IT) since the early 2010s. While the antecedents of data analytics techniques date back several decades ago and the first Big Data technologies were developed during the 2000s decade, it has been along the 2010s decade when the popularization of Big Data [MCB$^+$11] has led to the promotion and interest in using these technologies across many application fields. The cross-sector applicability of data processing and exploitation technologies, favored by the intensive promotion of Big Data tools and other synergistic technologies such as *Cloud Computing* and the *Internet of Things* (IoT), has led to coin the concept of "data-driven economy" [Eur14] as one of the cornerstones of economic development at a global scale. According to the report published by the European Commission in 2017 [IO17], the value of the EU data market, i.e. the exchange of data-related products and services, was estimated in almost EUR 60 Billion in 2016, and is expected to grow to more than EUR 106 Billion by 2020. Similarly, the total number of companies in the EU whose main activity is the production and delivery of data-related products or services is expected to grow from 255,000 units in 2016 to 360,000 units in 2020, and the aggregated impacts of the data market itself on the EU economy as a whole will grow from almost 2% of the EU GDP in 2016 to a 4% in 2020.

One of the strategic focuses where this data-driven economy is being deployed worldwide is the manufacturing industry, as a means to revitalize the global competitiveness of this sector, given its impact in many countries' economies, and to reverse a trend towards deindustrialization. For instance, according to the European Commission, industrial production accounts for 17% of Europe's GDP and 75% of the EU's exports are manufactured products. Moreover, it remains a key driver for growth and job creation, as each job in manufacturing generates at least an additional job in services [Eur17a]. The instantiation of this data-driven economy in the manufacturing industry has led to the development of *Smart Manufacturing*, as a global-scale overarching term for different initiatives and strategies addressing the use of data exploitation for optimizing and transforming manufacturing businesses. Indeed, the main initiatives worldwide promoting the adoption of Smart Manufacturing approaches coincide in time with the popularization of Big Data along the 2010s decade. Smart Manufacturing is defined [DEP$^+$12] upon two main concepts: the compilation of *manufacturing records* of products, with data about their history, state, quality and characteristics, and the application of *manufacturing intelligence* to those records, so that the exploitation of those data allows manufacturers to predict, plan and manage specific circumstances in order to optimize their production. This enables important business opportunities for the manufacturers, either to internally apply this approach or to *servitize* their business in order to help other manufacturers to shift towards a Smart Manufacturing-oriented operation of their production.

By its very own definition, the deployment of Smart Manufacturing approaches demands the introduction of data-related IT and digital platforms that support the achievement of the goals established for Smart Manufacturing. The appropriate design and implementation of such platforms faces diverse research and innovation challenges regarding the required technological enablers, including, among others, the following [Eur16]: improved methods of gathering valuable machine data and data integration across different sources; *cyberization* of legacy machines and integration of new IoT compliant machines with legacy production lines; data architectures matching industrial needs, provision of the right information, to the right person at the right time; tools for forecasting, monitoring and visualizing; implementation of data analytics methods in order to correlate product, process and business related information, and to forecast the product qualities performance indicators; etc.

Given the wide spectrum of these technological challenges and their complexity, the adoption of the required data-related IT by manufacturers aiming at shifting their businesses towards Smart Manufacturing demands the support of technology suppliers [Eur17b] specialized in these *Industrial Big Data Services* (IBDS). Thus, the risk of this technology adoption process is reduced for manufacturers and, at the same time, it enables a new market for technology suppliers deploying the required innovative solutions to support the adoption of Smart Manufacturing. This specialization of technology suppliers, i.e. the *IBDS Providers*, and their challenges designing their business around the supply of these technologies constitute the focus of this research work.

1.1 IBDS Providers: A Fundamental Agent for the Effective Development of Smart Manufacturing

IBDS Providers derive from a specialization of Information Technology Services (ITS) Providers, i.e. those companies whose business is focused on supplying IT services and enterprise software to companies demanding those services. The confluence of the technological and manufacturing business factors described above has led to the emergence of IBDS Providers, as the specialized ITS Providers supplying manufacturers with the required technology and services to *smartize* manufacturing businesses. Thus, IBDS Providers constitute a fundamental agent in industrial scenarios where there is an interest in adopting Smart Manufacturing approaches. Barring big manufacturing companies, the business core of most manufacturers does not include the technological skills and specialized team to develop and deploy the Industrial Big Data solutions required for adopting Smart Manufacturing approaches or for transforming their business model via data-driven servitization. Therefore, the effective development of Smart Manufacturing promotion policies and their extensive adoption by manufacturers -including SMEs- as a means to strengthen the competitiveness of the manufacturing sector cannot be achieved without these specialized technological agents. This context facilitates establishing strategic partnerships between IBDS Providers and manufacturing companies, aiming at designing the required *smart services* that allow these manufacturers to leverage the potential of Smart Manufacturing to transform their businesses or the operation of the production processes.

IBDS Providers represent the focus of this research work and where our contributions are targeted at. In this context of global-scale promotion of Smart Manufacturing and related initiatives in different countries and regions worldwide, we adopt the perspective of IBDS Providers and their strategic aim at developing and consolidating a sustainable and scalable business providing their services in Smart Manufacturing scenarios. This focus also allows us to frame the goal of our contributions within the existing proposals in the fields of Smart Manufacturing and data-driven projects. Thus, the overarching goal of this research work is to provide contributions that (a) help the business sector of IBDS Providers to develop effective and efficient data-driven services for the development of Smart Manufacturing and its strategic economic goals, and (b) adapt and extend existing conceptual, methodological, and technological proposals in order to include those practical elements that facilitate their use in business contexts.

Indeed, the observation of the Smart Manufacturing scenarios where IBDS Providers aim at supplying their services facilitates the identification of opportunities for relevant and purposeful contributions extending existing approaches. For instance, many conceptual proposals regarding the development of technological platforms for Smart Manufacturing offer a holistic approach and are aimed at an agent that has the capability to design from scratch or completely redesign the required infrastructure. However, in the real-world scenarios where IBDS Providers supply their services, there are running manufacturing businesses with an Operational Technology (OT) infrastructures already deployed and functioning. Therefore, for an IBDS Provider's business value proposition to be easily accepted, they must aim at deploying additional technology so that it integrates into the existing one and does not interfere with the current operation of the manufacturing business.

On a related matter, most of the main methodological and conceptual approaches supporting a data exploitation lifecycle assume a starting stage where there are indeed some *new data* available to be processed. Nevertheless, this is not the case when an IBDS Provider aims at supplying their services to manufacturing companies, as most data-generating devices currently deployed in their facilities have been designed for automation and internal supervision, and not to convey their data to an external platform where they can be processed, exploited and analyzed. Therefore, the technology deployed by an IBDS Provider must save that *gap* in order to extract the data and feed them as new data into a repository where they can be accumulated for their exploitation. Moreover, the design of that technological solution must be aligned with the sustainable development of the IBDS Provider's business, and not as *ad hoc* projects for each manufacturing facility to be monitored.

1.2 Scope and Method for this Research Work

Among the different opportunities that arise in the previously described context for relevant contributions aimed at facilitating the sustainability and scalability of an IBDS Provider's business, we highlight three specific challenges related to the early stages of the data lifecycle. These are the stages that ensure the availability of new data coming from monitored manufacturing facilities, whose owners are interested in exploiting those data in order to smartize their businesses. Thus, the three challenges on which we focus our research are the following:

1. Devising a more efficient data storage strategy that reduces the costs of the cloud infrastructure required by an IBDS Provider to centralize and accumulate the massive-scale amounts of data from the supervised manufacturing facilities.
2. Designing the required architecture for the data capturing and integration infrastructure that sustains an IBDS Provider's platform. This architecture must ensure a non-intrusive integration with the OT infrastructure currently functioning in monitored facilities and the progressive extension of the platform's functionalities to supply services to increasingly more scenarios.
3. The collaborative design process with partnering manufacturers of the required smart services for a specific manufacturing sector. This collaboration sustains the strategic partnerships with manufacturers in the targeted scenarios and reinforces the business value proposition of an IBDS Provider to supply their services in this market.

The research scope outlined by the aforementioned challenges points at an important characteristic of this work: instead of being driven by a specific research and knowledge area, it is driven by a wider analysis focus around the requirements related to Information Systems (IS) for IBDS Providers to design a sustainable and scalable business in these scenarios. This implies a research work that analyzes (a) the Smart Manufacturing scenarios where an IBDS Provider supplies their services, in order to characterize all relevant agents involved and their business strategies and IS-related requirements, and (b) the identification of research and knowledge areas where related work can be analyzed, so that synergies can be drawn with relevant references and limitations can be discovered as an opportunity for pertinent contributions.

In order to accomplish those goals, the method followed in this work is based on two main methodological approaches: *Design Science Research* and *Case Study Research*. On one hand, Design Science Research provides a methodology for research in IS. It aims at building purposeful *design artifacts* that are based on (a) the needs and requirements of the identified business problem in the analyzed application domain, and (b) the identification of synergies and opportunities with respect to existing knowledge in the related research areas. These foundations ensure the *rigor* and *relevance* of the design artifacts, so that they are valid research contributions for the academic audience and useful contributions for the practitioner audience and their environment. On the other hand, Case Study Research allows IS researchers to learn by studying the innovations put in place by practitioners and capturing knowledge from it, so that they can later formalize this knowledge. This is particularly appropriate for practice-based problems where both the experiences of actors and the context of action are critical. Conducting a case study is especially adequate for our research work, as its focus requires the direct observation of a real-world business setting where the relevant agents to all levels interact with each other to build the required smart services, according to their respective business strategies.

A case study sustains two crucial elements for this research work. First, it allows us to capture a more detailed characterization of the targeted Smart Manufacturing scenarios, through the analysis of a relevant instance of those scenarios and the multiple agents involved in them. This refines the scope of our research work and the specific scenarios where our contributions are aimed at, based on the practical requirements and business needs of the agents interacting in these scenarios around IBDS Providers. Leveraging these identified requirements and needs as input for the design science research process is what ensures the relevance of the proposed design artifacts. Second, it provides the ground for a field validation of the design artifacts' core components in a real-world setting. Indeed, the contributions of design science research are assessed as they are applied to the targeted business need in an appropriate environment. A successful contrast in this environment is what enables their addition as new relevant content in the knowledge base of the related research areas, for further research and practice.

Thus, in order to conduct our case study, we integrated ourselves in the real-world business setting of an IBDS Provider supplying their services to diverse Smart Manufacturing scenarios. This gave us the opportunity to observe the market of IBDS Providers in general as well as the different types of manufacturing companies and sectors where the services of IBDS Providers are deployed. Moreover, it also granted us the access to ongoing *smartization projects* where the services supplied by the IBDS Providers where being deployed in specific manufacturing sectors. In particular, we thoroughly examined the strategic partnership that this IBDS Provider had established with a Capital Equipment Manufacturer (CEM) deploying a data-driven servitization strategy in a chemical manufacturing sector distributed worldwide, and accompanied them throughout the deployment of a smartization project for one of this CEM's international customers. This allowed us to interact with relevant stakeholders in the involved companies and to access to the raw data coming from the monitored facilities and the technology deployed to capture and process those data. These real-world elements reinforced the characterization of the targeted scenarios and enabled the field validation of the core components of our research contributions, aimed at specific roles in an IBDS Provider's team involved in the analyzed smartization projects.

Furthermore, the observation and analysis of the main characteristics of the targeted scenarios and the requirements for an effective solution of the posed challenges allowed us to identify the relevant research and knowledge areas to examine and integrate in this multidisciplinary work: techniques and strategies for time-series data reduction, architectures and frameworks for data platforms in Smart Manufacturing scenarios, operational infrastructure in manufacturing facilities, conceptual models for Big Data systems, process models for Knowledge Discovery and Data Mining, project management, stakeholders management, requirements elicitation and business model design. Thus, the presented contributions are sustained by the examination and identification of synergies with relevant proposals in these areas, as well as the discovery of opportunities to overcome their limitations to address practical aspects of the real-world scenarios where IBDS Providers supply their services.

1.3 Main Contributions of this Research Work

The development of the previously outlined research method, accompanying and interacting with the different management and technical roles in all the organizations involved in the real-world business setting of our case study, allowed us to extract valuable insights to characterize the global market of IBDS Providers, the general requirements of the main agents in these Smart Manufacturing scenarios and the particular needs of the roles in the project team that an IBDS Provider establishes for their smartization projects in diverse manufacturing sectors. These requirements and needs, extracted from the strategic, tactical and operational reality of these companies, together with the examination of the adaptations and extensions necessary for the proposals in the related research and knowledge areas to effectively answer to that reality, sustain the relevance and rigor of the three main contributions of this research work.

The *first* main contribution is a *procedural and architectural design for the planning and execution of time-series data reduction analysis.* This contribution is focused on the duty of a given IBDS Provider's data engineer in charge of analyzing the reduction of the highly heterogeneous types of time series that constitute the data to be captured from monitored facilities where smartization projects are conducted. The relevance of this contribution is linked to (a) the costs of the cloud storage services that IBDS Providers require in order to deploy and run their platform and how these cloud resources impact the scope of the data exploitation services offered to manufacturers, and (b) the internal costs in allocated time and resources to explore the data reduction possibilities of the captured raw time-series data. Thus, this contribution represents the process (including the architecture of the IT artifacts to automate most of its steps) that efficiently guides the analysis of the data engineer and prioritizes allocating analysis resources to those time-series data with higher expected impact in storage space savings. The application of this design for an efficient reduction analysis allows obtaining the specification of reduction solution to be deployed in the IBDS Provider's platform, i.e. which reduction techniques must be applied to which time-series data, so that data storage is optimized without compromising their later exploitation. Moreover, as the data engineer uses an instantiation of the proposed design to analyze further application scenarios, the characterization of time series into families and their association with recommended reduction techniques is refined. This refinement supports an efficient knowledge management process of the insights and lessons learned extracted from different deployments and enables the savings in resources and time allocated for successive reduction analyses.

The *second* main contribution is the *design of a distributed hybrid architecture for the data capturing and integration platform of an IBDS Provider.* This architectural model complements existing popular paradigms for Big Data systems by describing the architectural components that save the gap between an initial state where no data is yet extracted from manufacturing facilities and the eventual availability of a centralized data repository on top of which different exploitation functionalities can be designed. The components of this architecture effectively combine Industrial IoT (IIoT) and Cloud Computing elements to provide an efficient answer to the volume, velocity and variety of data found in real-world manufacturing business settings. The main differential point of the proposed design is that the architecture is not conceived as a solution to migrate the whole industrial infrastructure of those settings demanding a shift towards Smart Manufacturing. Instead, it is conceived as a solution that support the business of an IBDS Provider, based on facilitating that shift to manufacturers with a non-intrusive, integrative approach with respect to already running OT infrastructures. Furthermore, it facilitates the successive upgrade of the supported functionalities to cover more application scenarios and to progressively support more data transformation steps towards the provision of smart services.

The *third* main contribution is the design of a *process model for a business stakeholders-driven characterization of data exploitation requirements in smart services.* This contribution is sustained by the integration of relevant knowledge from research areas such as stakeholders management, business model design and interview analysis to overcome the shortcomings identified in *Knowledge Discovery and Data Mining (KDDM) process models* and *requirements elicitation* in order to design smart services for the targeted Smart Manufacturing scenarios. Thus, this contribution extends KDDM process models with an incremental approach, designed as a *spiral process model* for the integration of *business understanding* into the data lifecycle to be covered, and facilitates the interaction with business stakeholders in order to elicit and characterize data exploitation requirements. This characterization is captured in a template, coined as the *BRIDGE canvas*, that connects business requirements and their impact into relevant KDDM process steps, so that those requirements can be leveraged as input for the relevant data lifecycle steps. These contributions are aimed at the project manager role supplied by the IBDS Provider for the smartization projects conducted in the targeted scenarios.

1.4 Dissertation Outline

This dissertation is divided into eight chapters, being this introduction the first of them. After that, the main content of this dissertation can be divided into two parts. A first part (chapters 2 to 4) covers the definition of the relevant background, method and scope that settle the ground for our contributions, and a second part (chapters 5 to 7) presents in detail the three main contributions of this research work.

Chapter 2 provides a detailed background on the context and antecedents of the focus of our research and contributions, i.e. IBDS Providers and their supply of services for Smart Manufacturing scenarios. The chapter details both the technological and manufacturing business backgrounds that lead to the emergence of IBDS Providers, and frames the research challenges around this agent where we focus on.

Chapter 3 describes the research method followed in this work, based on Design Science Research and Case Study Research. As well as presenting both methodologies, the chapter also describes how they have been combined to support this research work. The presentation of our three main contributions along chapters 5, 6 and 7 follows the schema of method steps outlined in this chapter.

Chapter 4 provides the characterization of the Smart Manufacturing scenarios where we target our contributions for IBDS Providers, as well as the real-world setting where we conducted our case study, as a relevant instance of those scenarios. The characterization of the targeted scenarios, the involved agents and their main strategies, requirements and needs contributes to delimit the scenarios included in the scope of our research.

Chapter 5 presents the first of our contributions, i.e. the procedural and architectural design for the planning and execution of time-series data reduction analysis. As a motivation for this contribution, the chapter begins describing the problem of data storage in an IBDS Provider's business and the need for efficient data storage strategies, as well as detailing related work on time-series data reduction in order to verify the opportunity for a methodological approach to assist the reduction analysis by a data engineer. The chapter continues describing the field validation of the core hypotheses of our contribution in the real-world business setting where we conducted our case study. Finally, it presents the contributed design, composed of procedural and architectural models for planning and executing time-series data reduction analysis. Besides, Appendices A and B present, respectively, the detailed results of the conducted field validation and the low-level design of the IT artifacts supporting the proposed design.

Chapter 6 presents the second of our contributions, i.e. the design of a Distributed Hybrid Architecture (DHA) for the data capturing and integration platform of an IBDS Provider. This contribution is sustained by the main non-functional requirements derived from the targeted Smart Manufacturing scenarios where IBDS Providers supply their services. By analyzing these requirements and the data capturing and integration technology deployed in the more than sixty manufacturing facilities in the analyzed business setting, we identified the core components that contributed to fulfill those requirements. Thus, after the analysis of diverse references on architectural proposals, identifying limitations to overcome and drawing synergies with the identified core components, we present the designed DHA and the two level of data management it encompasses: one IIoT-based level for the local management of raw data at each connected manufacturing facility, and another cloud-based level for the centralized management of a Big Data Lake with data from all connected and monitored facilities.

Chapter 7 presents the third of our contributions, i.e. the design of a process model for a business stakeholders-driven characterization of data exploitation requirements in smart services. It analyzes related work in the different knowledge areas integrated in the design of this contribution: requirements elicitation in data-related projects, KDDM process models, interview analysis, stakeholders management and business model design. The chapter continues describing the field validation of the core hypotheses of our contribution. This involved the design of a requirements capture process and its supporting tools, as well as their contrast in the real-world business setting where we conducted our case study. After the conclusions of this field validation, we present the refined design of the elicitation interviewing process as a spiral process model, and the supporting canvas tool to capture business requirements and translate them into technical, KDDM-oriented requirements.

Finally, chapter 8 presents the global conclusions after conducting this research work and the opportunities for further extensions of the research lines where we designed our contributions.

Chapter 2

Context and Antecedents

This research work and its contributions are focused on the business context of an *Industrial Big Data Services Provider* (IBDS Provider). This role describes an *Information Technology Services Provider* (ITS Provider) whose business strategy is targeted at providing the required IT support and data-related services for manufacturing companies, including *Capital Equipment Manufacturers* (CEMs) who pursue a data-driven servitization approach. For that purpose, the IBDS Provider offers its own *Platform-as-a-Service* (supported in *Big Data* technologies, *Cloud Computing* and *Industrial Internet of Things*) as a transversal solution for the data gathering and integration needs in diverse manufacturing markets. Besides, they establish partnerships with manufacturing companies to deploy their solution in specific manufacturing markets and collaboratively design the vertical, sector-specific solutions (to be provided in a *Software-as-a-Service* model) in order to provide *smart services* as a means for CEMs to deploy their data-driven servitization strategy and for manufacturers to transform the operation of their production processes.

The role of IBDS Provider emerges in a business context that is the consequence of the evolution and trends in different areas. On one hand, the evolution of the provision of IT services and the technological breakthroughs since the start of the new millennium have progressively transformed the technological ground that sustains the business models of ITS Providers. In this sense, the adoption of *Cloud Computing* approaches and, above all, the rise of *Big Data* technologies and the resurgence of *data analytics* have shifted the focus of many ITS Providers and that of the companies demanding their services. On the other hand, there has been an intensive development of the concept of *Smart Manufacturing* at all levels of the manufacturing industry, including various public and private initiatives worldwide promoting its adoption. This has enabled the opportunity for manufacturers to shift towards a more Smart Manufacturing-oriented approach and for CEMs to increase their competitiveness by shifting towards a data-driven servitization strategy and aiming at providing their customers with the required services to adopt a Smart Manufacturing approach. The accomplishment of such a shift towards Smart Manufacturing demands that CEMs and their customers establish partnerships with ITS Providers specialized in the required technologies, which grounds the motivation for the emergence of IBDS Providers.

This chapter presents the technological background that has led to the evolution of ITS Providers, with a special focus on the rise of Big Data technologies and the resurgence of data analytics approaches, given their relevance for a IBDS Provider. Besides, the relevant milestones that have led to the emergence of Smart Manufacturing are detailed, and it is presented how smart services are a predominant focus in servitization strategies by CEMs. This context sustains the business opportunity that leads to the emergence of IBDS Providers, whose challenges that are relevant for our research are also introduced.

2.1 Technological Background

The provision of IT Services to companies, based on the supply of enterprise information systems, has evolved along with major technological milestones and breakthroughs. This has shaped the focus of ITS Providers (from which IBDS Providers are a specialized category) and how they approached the provision of IT services to companies demanding such services. In this sense, the progressive adoption of *Cloud Computing* has enabled different models, changing both the means to provide IT services and the nature of the provided IT services themselves, depending on whether they are focused on providing infrastructure, platform or software (i.e. fully fledged solutions).

Besides, in order to explain the emergence of IBDS Providers, the rise of *Big Data* technologies and the resurgence of *data analytics* play a crucial role. The concept of Big Data has been one of the main technological trends worldwide during the 2010s decade. However, the idea of exploiting datasets with a business vision and many of the technologies used for it date back a few decades ago. Since then, different factors have contributed to the evolution of those first approaches for business data analytics until the modern development of Big Data technologies, and to the massive popularization of Big Data as a technological trend.

This section details the evolution of the technological ground for ITS Providers. Given the major relevance of Big Data for the role of IBDS Provider, we devote a specific subsection to analyze the antecedents, origins and rise of Big Data technologies.

2.1.1 Evolution of Technological Ground for ITS Providers

The role of *Information Technology Services Provider (ITS Provider)* describes an IT company whose business model is focused on the supply of *enterprise information systems* (also referred as enterprise software) and related services to companies that demand such IT services. The enterprise information systems supplied by ITS Providers encompass transversal business operations, such as accounting or business intelligence, usually integrated as enterprise resource planning (ERP) suites, as well as vertical, industry-specific solutions [McL13].

The market for ITS Providers emerged as those companies not having IT development as part of their business core decided in favor of outsourcing IT services to a third party. This decision was motivated by relevant benefits for these companies [Dha12]: on one hand, it allowed them to keep their business focus and increase efficiency to develop and market their products or services earlier to the market; on the other hand, they benefited from the specialized IT skills, higher standards, better integration practices and the economy of scale possessed by the ITS Provider thanks to their knowledge and experience in diverse industries. Indeed, companies tended to establish strategic partnerships with their ITS Providers when this helped them gaining competitive advantage to develop new strategic applications in their markets.

The evolution of ITS Providers and the technological focus that sustained their provision of IT services has been strongly shaped by various influential milestones in the history of information technologies. In particular, two major milestones during the beginning of the new millennium defined the evolution of the landscape for ITS Providers during the subsequent years [Zue11]: the need for ITS Providers (mainly ERP Providers) to refocus their businesses when their volume of activity plummeted after the “Y2K effect”, and the massive adoption of the World Wide Web that, despite the burst of the “dot-com bubble”, had transformed the perception of how information technologies would be used by people and companies towards a more Internet-based approach, where the World Wide Web would provide the ground upon which to generate and transact businesses [OK10][Zue11].

One of the first new ideas to emerge in this context was the concept of *Application Service Providers* (ASPs) [Bia00]. While the traditional model of software provision required companies to purchase licenses as well as host and maintain the software in-house [Zue11], the ASP model allowed traditional software vendors to offer hosted versions of their software running on off-premise data centers, so that end users accessed it via a client software that provided a *seamless* perception of the use of the IT solution [Bas17][Zue11].

The evolution of this model was boosted by two new technological concepts that were developed in parallel during the 2000s decade: the idea of *Web 2.0* to label the progressive transformation of how providers and end users started producing and consuming services via web-based interfaces towards a more people-centric collaboration and interaction approach [MPF10], and the emergence of *virtualization* as the approach to enable users to simulate and run multiple virtual computers on one physical computer, thus sharing computational resources and reducing the costs of system administration [MPF10][Zue11]. The combination of virtualization and the transition from single-tenant to multi-tenant solutions (providing more flexible scalability and resource balancing), together with the massive adoption of web-based provision and interaction with IT services, led to the evolution of the initial concept of ASPs and to the emergence of the approach that was eventually coined as *Cloud Computing*. The term alludes to how virtualization abstracts the technology layer and hides it from the end user behind some “cloud” [Bas17][MPF10][Zue11].

The progressive increment of interest by companies in virtualization and, eventually, cloud computing solutions during the second half of the 2000s decade is clearly shown in studies that evaluate the technology priorities by CIOs of companies worldwide [Gar05][Gar10][Gar12][Gar17]. In these studies it is shown that, while in 2005 [Gar05] *virtualization* started emerging as the tenth most prioritized technology, in 2010 [Gar10] it was ranked in the first place together with the emerging concept of *Cloud Computin*g in second place and *Web 2.0* as third. From that year on [Gar12][McL13], cloud computing displaced virtualization, consolidating itself as the trending term, and it has been since then until 2017 [Gar17] among the three top ranked technology priorities, along with Big Data/data analytics and mobile technologies.

The increasing adoption of the Cloud Computing model has introduced noticeable changes in the way ITS Providers design their supply of IT services and business applications. Indeed, not only the means of providing the IT service has changed, but also the nature of the IT services themselves to be provided has suffered a significant transformation [Dha12]. Thus, three main models for the provision of IT services have been enabled with the adoption of Cloud Computing [MG11]:

- *Infrastructure-as-a-Service* (IaaS). *IaaS Providers* focus on the abstraction of IT infrastructure resources, such as storage space and computing power, and on providing them as a service for those who want to deploy and run different applications on that infrastructure.
- *Platform-as-a-Service* (PaaS). At this level the abstraction does not only cover the essential technical resources, but also some essential application services that enable the development of purpose-specific solutions.
- *Software-as-a-Service* (SaaS). This level encompasses the provision of fully fledged, purpose-specific solutions for end users.

In any of those three models, regardless of whether the nature of the service is providing infrastructure, platform of software, the approach is based on the virtualization of IT infrastructure, multi-tenancy for flexibility and scalability, and web-based provision of services on a pay-per-use basis [Dha12]. In fact, the rising adoption of these three models reinforces each other, given that the ability to implement platform and infrastructure services in the cloud in a time- and cost-efficient way boosts the creation of scalable SaaS applications [McL13].

From the supplied user’s or company’s perspective, one of the main benefits of this approach is that the solution is hosted, maintained and upgraded by the service provider [Zue11]. Moreover, this approach changes fundamentally the cost structure of the consumption of IT solutions, as it shifts from capital expenditure to operating expenditure [Dha12]. The acquisition of traditional on-site software licenses (and the hardware to run it) required a relevant upfront investment of time and capital, while these new models facilitate the process to the point of assimilating the purchase of software to a subscription-based service [OK10]. However, Cloud Computing-based models also present relevant drawbacks, mainly related to the concerns about security and privacy, as well as the different applicable requirements and regulations on data sovereignty depending on which country the data center is located in [Dha12][McL13]. Nevertheless, the advantages in terms of cost savings, scalability, accessibility, easy upgrades and resilience are increasingly attractive for many organizations [McL13]. As a consequence, this type of solutions has been adopted in all areas of enterprise information systems. Traditional software vendors are also migrating their provision approach towards this model [McL13][Zue11], and an increasing number of user companies expect customer-specific innovative IT solutions supplied by ITS Providers adopting one of these models [Dha12].

2.1.2 The Rise of Big Data Technologies

The origin of Big Data technologies is grounded on the different approaches for data analytics and their application in business contexts, which dates back to many decades ago [NnI15]. The origin of Big Data technologies is motivated by the application of these data-related techniques and tools for the use case of a specific profile of organizations: those major technological companies founded in late 1990s and focused on developing their business around the World Wide Web and the provision of search engines. Once the initial Big Data technologies were developed, their progressive use and evolution led to an ecosystem of numerous tools that have been increasingly adopted in diverse industries. Moreover, the synergies with other technological breakthroughs, the availability of conceptual constructs to develop Big Data systems, and the mainstreaming of the *Big Data* concept as a technological trend have massively contributed to the resurgence of data analytics and its renewed popularization among researchers, practitioners and users.

2.1.2.1 Antecedents: Data Analytics prior to Big Data Technologies

One of the first key terms that we can find in the field of business data analytics is *Business Intelligence*, whose first reference dates back to 1958 by Hans Peter Luhn [Luh58], researcher at IBM. However, this first definition of the term was quite different from the evolution it suffered subsequently, with the progressive computerization of business processes. After the development in this field in the following years, it was in the 80s decade when the concept of Business Intelligence (mainly with the definition proposed by Howard Dresner [Mar06]) became established, in order to define a set of software systems designed to support business decision making, based on the gathering and analysis of *facts* (i.e. data). These systems were focused on a descriptive analysis, consulting historical data in an aggregated way and cross-matching indicators to obtain a better vision of what has happened and is happening in the organization.

Hence, the Business Intelligence approach left aside a predictive type of analysis, which aimed at extracting knowledge from data in the form of patterns, trends or models that provided a degree of certainty about the outcome of potential future actions. In order to refer to this type of analysis, at the end of the 80s the expression *Data Mining* was coined. The origin of this term comes from an analogy with mining techniques, where a valuable material (in this case, *knowledge*) is extracted from mining deposits (*data banks*). Along with Data Mining, which is arguably the most known and extended term to refer to this type of analysis among a group of similar expressions [HK06], around the same time the expression *Knowledge Discovery in Databases* (KDD) started being used. On many occasions both terms were used interchangeably, although the term *Data Mining* was also used to refer specifically to the analytics step in a KDD process. In fact, the first academic seminar on KDD held in 1989 [PS91] led to the First International Conference on *Knowledge Discovery and Data Mining* (KDDM) in 1995 [FU95].

The development of KDDM projects to search for and exploit patterns in data banks, using *Machine Learning* techniques [Mit97] to build predictive models, began to spread among business contexts during the 90s decade. This led to an increasing interest in data mining applications during those years (see Figure 2.1). Banking firms and insurance companies stood out in this application, aiming at leveraging the outcome of this type of analysis to facilitate decision-making processes linked to their products (for instance, fraud detection by insurance companies, or the authorization or denial of credit applications).

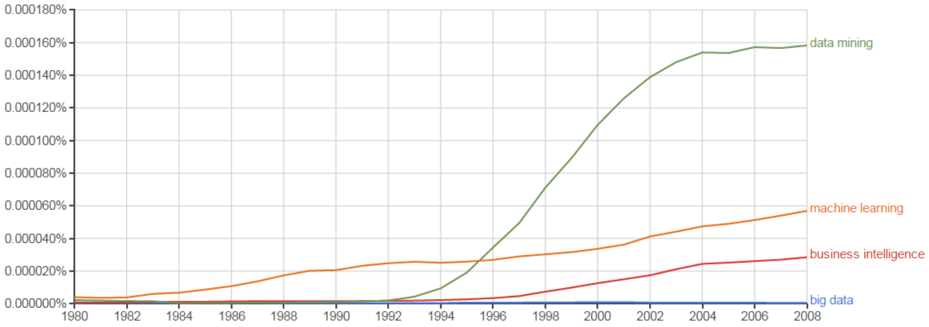


Figure 2.1: Historical frequency of occurrence of relevant terms on data analytics between 1980 and 2008[1]

It is in this context where proposals for a reference model to conduct KDDM projects began to appear. The foundational schema of KDDM phases (see Figure 2.2) was proposed in 1996 [FPSS96] by the academics organizing the KDDM-related seminars and conferences mentioned above. After that, various additional KDDM process models were proposed [KM06]. Among those proposals, the *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) reference model [She00] stands out given its acceptance among KDDM practitioners. Indeed, although the version 1.0 of CRISP-DM was published in 2000 and there is no current effort to publish a new version, it is still cited as the most often used

[1] Extracted from Google Books Ngram Viewer (https://books.google.com/ngrams) on March 2017, as the result obtained of the query comparing the historical frequency of occurrence of "big data", "business intelligence", "data mining" and "machine learning" in this service's English corpus from 1980 to 2008.

methodology to manage data mining projects [PS14].

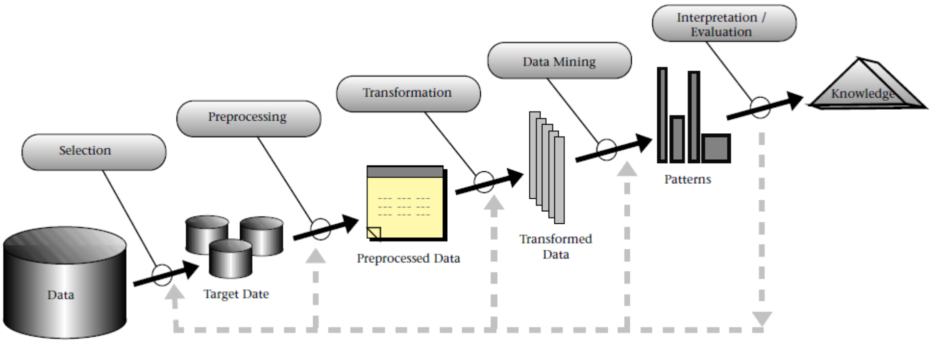


Figure 2.2: Foundational schema of KDDM phases (extracted from [FPSS96])

2.1.2.2 The Development of Big Data Technologies

The need for the technologies that have been defined afterwards as *Big Data* had its origin in the application of data analytics by the big technological companies arising (mainly in Silicon Valley) along with the emergence and popularization of the World Wide Web between the end of the 90s decade and the beginning of millennium. The problem faced by these companies did not differ from the one described before in the case of banking firms and insurance companies: boost their business exploiting their data banks. The key differential element emerges when we compare the dimension of data in both scenarios: while in the previous examples the amount of data could be processed using the tools and capabilities provided by conventional computers, in the case of big technological companies the large volumes of data to be analyzed made unfeasible in practice their processing via traditional techniques. In fact, it was in those years when volume, velocity and variety in data [Lan01] (a model later known as "3 V") began to be analyzed as key aspects in any strategy for an optimal management of data in business contexts.

The pioneer in this new scenario was Google, which initially faced this problem [LRU14] to efficiently process their PageRank algorithm [PBMW98] when applied to large volumes of data coming from the analysis of a multitude of Web sites. As an alternative to existing solutions and strategies for parallel processing of large volumes of data, which were based on using high-performance machines (High-Performance Computing, HPC) with a large amount of processing cores, Google opted for developing their own solution with a different strategy. They focused on an efficient automation of most of the work involved in dividing the processing of large volumes of data among a set of distributed machines, each with an individual performance far more modest than that of a machine used in HPC. This solution was built upon two essential elements: a *distributed file system* to manage the storage of large-scale data in a partitioned and replicated way among the set of distributed machines (nodes in a cluster) [GGL03] and a software providing efficient implementations of the most complex tasks to be executed by those distributed applications dealing with the processing of data stored in such a system. This software and, by extension, the programming model enabled by it received the name of *MapReduce* [DG04], marking the main milestone in the origin of Big Data technologies.

The dissemination by Google of the details about their distributed file system and the MapReduce programming model served as inspiration for other projects that aimed at solving similar problems. In particular, the academic papers published by Google served Doug Cutting [Cut09] to improve the project to develop a web search engine in which he was involved at that time. This project provided the grounds for the work that Cutting developed later when he joined Yahoo, building a system implementing MapReduce with the capability to process in a distributed way the enormous volumes of data required by a major global search engine. Thus, the open-source system called *Apache Hadoop* was born, whose two main modules where the *Distributed File System* (HDFS, the open-source implementation of the distributed file system described by Google some years before) and *Hadoop MapReduce* (implemented upon the aforementioned HDFS).

The availability of an open-source solution like Apache Hadoop facilitated the adoption of Big Data technologies, favoring at the same time the creation over the next years of additional tools around this platform that boosted its usefulness and the later emergence of alternative platforms such as Apache Spark [ZCF$^+$10]. Figure 2.3 summarizes some of the main milestones until 2015 regarding the development of these technologies, including the antecedents on data analytics detailed above.

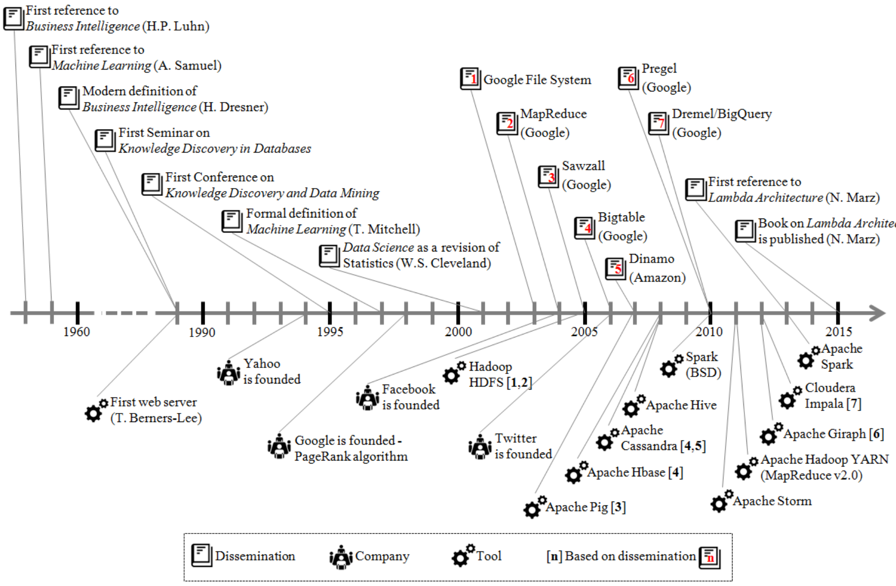

First reference to *Business Intelligence* (H.P. Luhn)
First reference to *Machine Learning* (A. Samuel)
Modern definition of *Business Intelligence* (H. Dresner)
First Seminar on *Knowledge Discovery in Databases*
First Conference on *Knowledge Discovery and Data Mining*
Formal definition of *Machine Learning* (T. Mitchell)
Data Science as a revision of Statistics (W.S. Cleveland)
1 Google File System
2 MapReduce (Google)
3 Sawzall (Google)
4 Bigtable (Google)
5 Dinamo (Amazon)
6 Pregel (Google)
7 Dremel/BigQuery (Google)
First reference to *Lambda Architecture* (N. Marz)
Book on *Lambda Architecture* is published (N. Marz)
1960
1990
1995
2000
2005
2010
2015
First web server (T. Berners-Lee)
Yahoo is founded
Google is founded - PageRank algorithm
Facebook is founded
Hadoop HDFS [1,2]
Twitter is founded
Apache Pig [3]
Apache Hbase [4]
Apache Cassandra [4,5]
Apache Hive
Spark (BSD)
Apache Storm
Apache Hadoop YARN (MapReduce v2.0)
Apache Giraph [6]
Cloudera Impala [7]
Apache Spark
Dissemination
Company
Tool
[n] Based on dissemination n

Figure 2.3: Chronology of milestones related to antecedents and development of Big Data technologies (extracted from [Nn15])

2.1.2.3 Main Factors Boosting the Adoption of Big Data Systems

Several technological factors support the increasing adoption of Big Data technologies during the last decade, related to the strong synergies between Big Data and two other *Key Enabling Technologies* (KET) in this area: *Cloud Computing* [ZZCW10] and the *Internet of Things* (IoT) [XHL14].

The progressive development of *Cloud Computing* systems allowed accessing big clusters of machines in a dynamic-renting mode, i.e. only requiring the payment for the computing and storage resources consumed at any particular time. This enabled a substantial reduction in the economic barrier to access these technologies (together with the availability of many open-source implementations of the required tools) and, therefore, more affordable approaches for Big Data system developers to be equipped with the infrastructure to store and process large-scale volumes of data, using various technologies deployed for that purpose.

As well as this, the proliferation of all kinds of devices capturing and sharing data over the Internet (IoT) opened the possibility for many different off-line application fields (where data were generated in a physical environment and therefore not originally available in an Internet platform) to centralize their data in a Cloud Computing system and apply various analytics approaches to those data.

While diverse Big Data technologies for different functionalities and abstraction levels were easily available after their progressive development during the 2010s decade, the proposal of various conceptual constructs guiding the design of Big Data systems and the integration of the available technologies also facilitated their adoption. Among these constructs we can highlight the concept of *(Big) Data Lake* [O'L14] and the architectural design pattern named as *Lambda Architecture* [MW15].

The notion of *Data Lake* was first coined by James Dixon in 2010 [Dix10] describing an approach for the centralized storage of the structured, semi-structured or unstructured data coming from diverse sources that were required to build the intended Big Data system. Given that such a system should support diverse analytical use cases from different user profiles, and considering that those use cases are typically not characterized in detail beforehand, those data should be stored in their raw format, i.e. not having applied any filtering or processing to make those data fit any particular schema. The subsequent detailed characterization of the data analytics needs by different users would allow identifying the transformations to be later performed on the accumulated data, in order to generate the required filtered and processed data views. Other authors have proposed a renaming of the concept as *Data Reservoir* [Blo14].

The *Lambda Architecture* (outlined in Figure 2.4) consists of a design pattern for Big Data systems aiming at reducing their complexity and providing better fault tolerance. One of the basic principles of this proposal is the *immutable data* approach, which has direct synergies with the concept of *Data Lake* described above: the accumulated data are not modified as new data are generated and stored; instead, all data are accumulated in their raw format as they are generated. On top of the accumulated data, different layers are defined to organize the system components that provide the data services: a *batch layer* pre-calculates the required operations on the *master dataset* of accumulated raw data to generate elaborated, transformed data views for different needs; a *service layer* provides efficient access to those batch views, so that queries on those views are resolved with low response times; a *speed layer* resolves the incremental processing of real-time data (and their integration in the queries requiring them) as long as they have not been accumulated yet in the master dataset from which batch views are prepared.

In parallel to these developments, the significant media interest in many of the technological breakthroughs generated in environments like Silicon Valley led to the popularization of Big Data not only in specialized contexts but also to the mainstreaming of the "Big Data" concept among the general public. Indeed, the interest in Big Data began to peak during the second half of 2011 (see Figure 2.5), well after the creation of the main technological contributions that led to Apache Hadoop and related technologies. This was mainly caused by the publication in June 2011 of a McKinsey Global Institute's report on global technological trends pointing at Big Data as a technological breakthrough for "innovation, competition and productivity" [MCB+11].

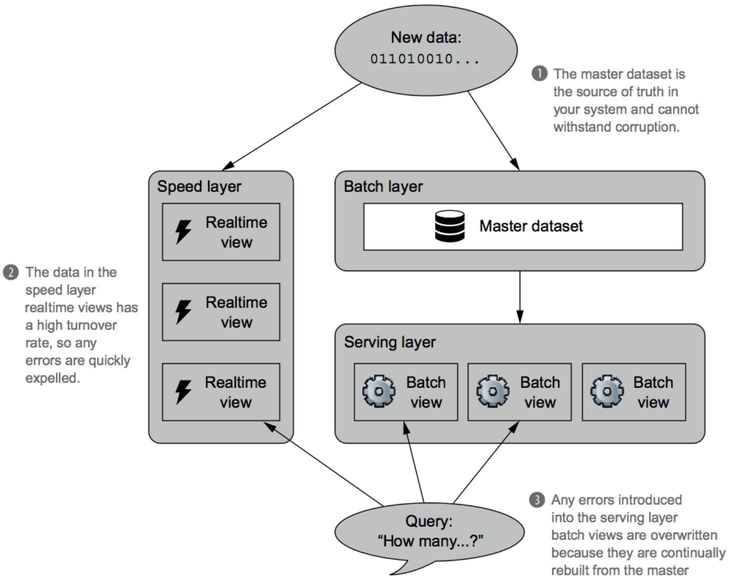

New data:
011010010...
1 The master dataset is the source of truth in your system and cannot withstand corruption.
Speed layer
Realtime view
Realtime view
Realtime view
Batch layer
Master dataset
Serving layer
Batch view
Batch view
Batch view
2 The data in the speed layer realtime views has a high turnover rate, so any errors are quickly expelled.
Query:
"How many...?"
3 Any errors introduced into the serving layer batch views are overwritten because they are continually rebuilt from the master dataset.

Figure 2.4: Lambda Architecture diagram (extracted from [MW15])

Although McKinsey's report did mention those Big Data technologies that were developed during precedent years, the specific data analysis techniques and the use cases described were mainly related to predictive analytics (i.e. Data Mining) applications. This caused a popularization of diverse predictive analytics use scenarios (and the consequent resurgence of Machine Learning techniques that, as presented in 2.1.2.1, were already applied in business contexts in the 90s) along with the "Big Data" tag and, eventually, led to a widespread misconception of Big Data. That is, instead of understanding Big Data as an additional technological layer for an efficient processing and analysis of large-scale volumes of data, the term began to be massively used to substitute "Data Mining" in order to refer to those popularized applications. In spite of this misleading use of the term "Big Data", it should remain clear that not every Data Mining application is about Big Data and vice versa. Furthermore, it should be differentiated whether the problem to be solved requires or not specific technologies to process and analyze large volumes of data [vdL15].

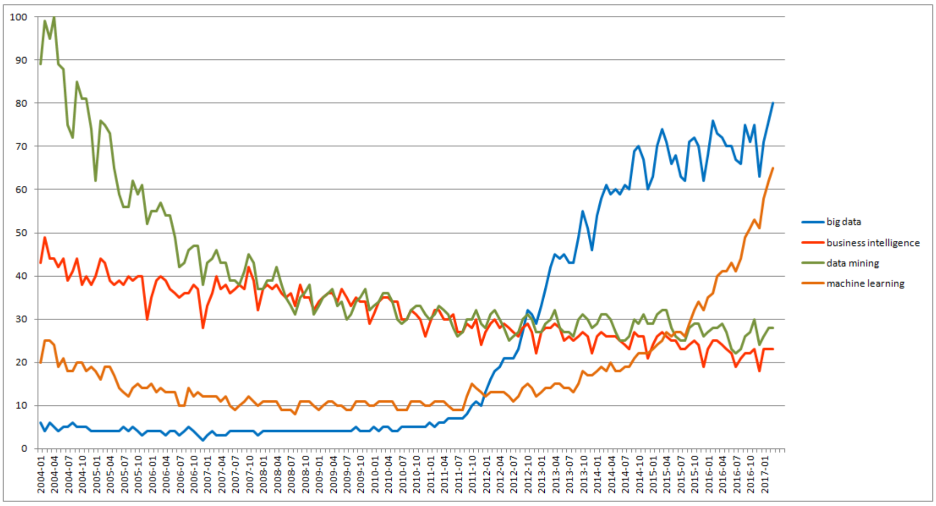


Figure 2.5: Interest over time (based on worldwide volume of Google searches) in relevant terms on data analytics between 2004 and 2017[2]

2.2 Manufacturing Business Background

As noted by Harding et al. in [HSSK06], the adoption of Data Mining and data analytics applications in the manufacturing industry began in the 1990s and gradually extended its adoption to different areas in manufacturing engineering. Nevertheless, the resurgence of data analytics (with the popularization of Big Data and related technologies as one of its main enablers) motivated a new wave of interest in these applications as a driver for manufacturing business transformation and a cornerstone of new strategies for the competitiveness of the manufacturing industry in many countries. In particular, it led to the emergence of *Smart Manufacturing* and the creation of data-driven smart services as the core of *manufacturing servitization* strategies.

This section details the most relevant milestones in the origin of the concept Smart Manufacturing and the strategic initiatives that have boosted their promotion and adoption among manufacturing companies. Besides, the specific case of the data-driven servitization of CEMs is analyzed, as a particularly relevant

[2] Data extracted from Google Trends (https://trends.google.com) on March 2017, as the result obtained of the query comparing the worldwide volume of Google searches for "big data", "business intelligence", "data mining" and "machine learning" from 2004 to 2017.

business context where the provision of smart services drives the servitization strategy and where IBDS Providers play a crucial role.

2.2.1 The Emergence of Smart Manufacturing

Apart from naming an academic conference organized in 2008 [IK08], the concept of Smart Manufacturing was first analyzed in detail in a technical report named "Smart Process Manufacturing: an Operations and Technology Roadmap" [DED+09]. While this report was published in 2009, the information forming the basis of that report was generated during a workshop held in April 2008 [Int08], involving different industry and academic experts from the USA. The aforementioned report put the focus on the technological and economical trends affecting the process manufacturing industry in a global economy, and presented *Smart Process Manufacturing* (SPM) as an approach to address the challenges and opportunities derived from those trends. The report defined the vision of SPM as "an integrated, knowledge-enabled, model-rich enterprise in which all operating actions are determined and executed proactively applying the best possible information and a wide range of performance metrics". As well as providing that definition, the report also acknowledged the importance of *smart technologies* and *cyber-infrastructures* to support the SPM vision, comprising technological areas such as data interoperability, networked sensors and multi-level security, among others.

The work initiated with the referred report was extended with a new workshop of similar characteristics held in September 2010 [Dav10]. This workshop led to the constitution of the *Smart Manufacturing Leadership Coalition* (SMLC) and to the publication in 2011 of the technical report "Implementing 21st Century Smart Manufacturing" [Sma11]. This new report detailed the vision and goals for the Smart Manufacturing enterprise, based on integrating data capture and exploitation throughout the entire product life cycle, so that the manufacturing process gains in flexibility and can rapidly react to specific circumstances with reduced costs, thus optimizing performance and efficiency. Besides, an action plan is detailed with priority actions classified into different categories, one of them being related to ensuring affordable industrial data collection and management systems. Indeed, the processing of the large-scale volumes of data generated by machine controllers, sensors, etc. (i.e. "Industrial Big Data") and their processing into useful information is the key of Smart Manufacturing approaches [LKY14].

The main advances of the work done by SMLC were compiled in a journal paper in 2012 [DEP+12], where Smart Manufacturing was defined as "the dramatically intensified application of *manufacturing intelligence* throughout the manufacturing and supply chain enterprise". This *manufacturing intelligence* comprises the "real-time understanding, reasoning, planning and management of all aspects of the enterprise manufacturing process and is facilitated by the pervasive use of advanced sensor-based data analytics, modeling, and simulation". Indeed, Smart Manufacturing systems agilely adapt to new situations by using real-time data for intelligent decision-making, as well as predicting and preventing failures proactively [JML$^+$15]. This leads to a fundamental business transformation towards e.g. *performance-based enterprises* and *demand-driven supply chain services.*

Thus, the concept of Smart Manufacturing as proposed by SMLC aimed at enabling a “game-changing”, advanced manufacturing model for the 21st century, differentiating itself from the previous technological advances deployed in manufacturing industries. Two core ideas that support this differentiation [DEP$^+$12] are:

- The compilation of a *manufacturing record* for each product, with data from sensors, procedures, specifications, tasks records and other observations. This creates a record for each product with data about its history, state, quality and characteristics.

- The application of *manufacturing intelligence*, thanks to the availability of product records and the ability to apply particular requirements more flexibly. Thus, manufacturing companies can adjust their production more flexibly and produce models of their processes that can be used to predict, plan and manage specific circumstances in order to optimize their production.

The progressive development of these core ideas led to different Smart Manufacturing applications with different scopes and approaches for using that *manufacturing intelligence*[BKM$^+$14][LNR14]: Manufacturing System Control, Manufacturing Quality Control, Fault Diagnosis of Manufacturing Equipment, Predictive Maintenance of Manufacturing Equipment, Decision-Support Systems, Decision-Guidance Systems, etc. Besides, the integration of technologies such as Cloud Computing and the Internet of Things in Smart Manufacturing solutions led to the proposal of new paradigms to provide guidelines for these applications, such as *Cloud Manufacturing* [ZLT$^+$14], *Internet of Manufacturing Things* [ZZW$^+$15] or *Internet of Things for Modern Manufacturing* [BXW14].

2.2.2 Strategic Initiatives on Industrial Competitiveness related to Smart Manufacturing

The rising interest in Smart Manufacturing applications has been boosted during the 2010s decade by the appearance of various public and private initiatives promoting their adoption. Some of these initiatives have been launched with direct involvement of national governments as a result of their strategies to accelerate the development of Smart Manufacturing in order to revitalize their manufacturing industry [DEP$^+$12]. Next, we detail the three most representative instances of these strategic initiatives: *Advanced Manufacturing* [Pre11], *Industrial Internet* [EA12] and *Industrie 4.0* [KLW11].

2.2.2.1 Advanced Manufacturing (USA)

In June 2011 it was presented the report “Ensuring American Leadership in Advanced Manufacturing” [Pre11] by the USA President’s Council of Advisors on Science and Technology (PCAST). This report recommended the launching of an innovation policy built on the concept of Advanced Manufacturing (related to the use of emergent new technologies to transform the creation of existing products and enable new products), in order to ensure the strategic development of the manufacturing industry in the USA. As a consequence of this recommendation, it was established the *Interagency working group on Advanced Manufacturing* (IAM), who developed the report “A National Strategic Plan for Advanced Manufacturing” [Nat12], published in February 2012. The concept of Advanced Manufacturing was detailed in this report and included the “use and coordination of information, automation, computation, software, sensing, and networking”. The strategic goal of the actions described in the report aimed at closing the existing gap between research and development activities and the deployment of technological innovations in production environments. The efforts derived from those actions led to a preliminary design in January 2013 of the *National Network for Manufacturing Innovation* (NNMI) Program and to the *Revitalize American Manufacturing and Innovation* (RAMI) Act in December 2014. The strategic plan for the NNMI Program was presented in a report [Nat16] published in February 2016. All these milestones are summarized in Figure 2.6.

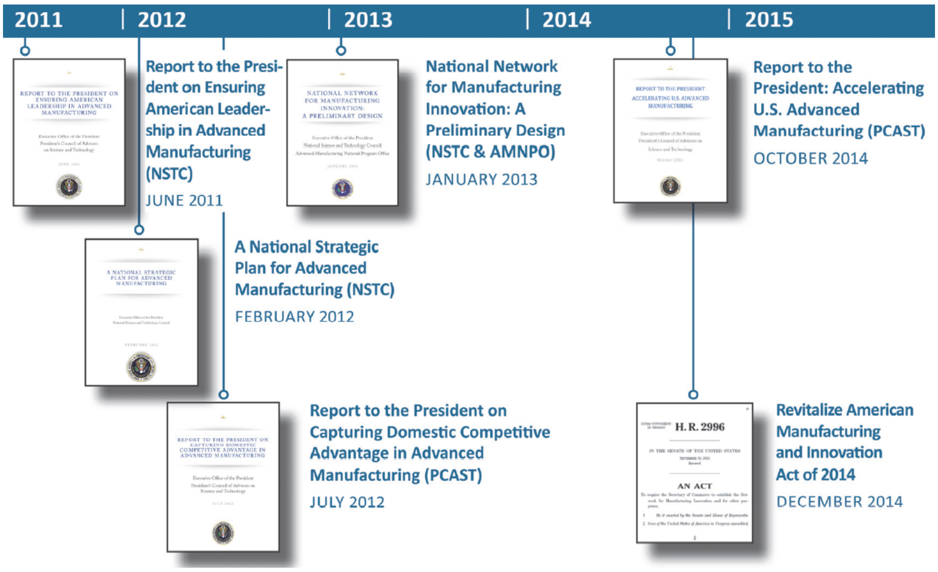

2011
2012
2013
2014
2015
Report to the President on Ensuring American Leadership in Advanced Manufacturing (NSTC)
JUNE 2011
National Network for Manufacturing Innovation: A Preliminary Design (NSTC & AMNPO)
JANUARY 2013
Report to the President: Accelerating U.S. Advanced Manufacturing (PCAST)
OCTOBER 2014
A National Strategic Plan for Advanced Manufacturing (NSTC)
FEBRUARY 2012
Report to the President on Capturing Domestic Competitive Advantage in Advanced Manufacturing (PCAST)
JULY 2012
Revitalize American Manufacturing and Innovation Act of 2014
DECEMBER 2014

Figure 2.6: Timeline for the Creation of the NNMI Program (extracted from [Nat16])

2.2.2.2 Industrial Internet (USA)

In parallel to the US Government's strategy around Advanced Manufacturing, the concept of *Industrial Internet* started being promoted and developed by major US corporations. The company leading this initiative was General Electric, publishing the report "Industrial Internet: Pushing the Boundaries of Minds and Machines" [EA12] in November 2012. This report developed the concept of Industrial Internet as the strategic use of technological breakthroughs related to connectivity and data analysis, applying them to the equipment in diverse industrial sectors. The strategy was supported by three key elements: interconnected *smart machines* distributed worldwide and equipped with measuring and controlling ability thanks to the use of digital technologies; *advanced analytics* using predictive algorithms on data generated by those machines; and the use of those elements to *facilitate decision-making processes*. The definition of this strategy eventually led to the foundation of the Industrial Internet Consortium (IIC) in March 2014, with AT&T, Cisco Systems, General Electric, IBM and Intel as founding members [Ind15]. The IIC has since then published different technical papers [Ind17b] describing, among other elements, a reference architecture [Ind17c] and a security framework for the Industrial Internet. They have also developed *testbeds* [Ind17a] to demonstrate the real-world implementation of Industrial Internet solutions.

2.2.2.3 Industrie 4.0 (Germany)

Among the various initiatives launched across Europe (see Figure 2.7) promoting the adoption of Smart Manufacturing approaches, Germany's *Industrie 4.0* (Industry 4.0) has become the most popular worldwide. The concept of *Industrie 4.0* was coined in 2011 [KLW11] as an initiative promoted by German public and private agents, joining academic and industrial experts. The initiative was supported by the German government as a long-term strategy to reinforce the competitiveness of German manufacturing industry by means of a progressive adoption of technologies, with the concept of the Internet of Things/Services as the main exponent of the technological breakthroughs to adopt. Based on those premises, the *Industrie 4.0 Working Group* was created to develop the main lines of that strategy, which were compiled in their final report [KWH13] published in April 2013. As it is described in that report, the integration of the Internet of Things/Services into the elements of a manufacturing plant would lead to *Cyber-Physical Systems* (CPS). This concept would encompass different production elements provided by *intelligence* and the ability to store and exchange data. This would enable CPS to register their historic logs, self-diagnose their states, and autonomously demand and activate actions involving other interconnected elements. The term *Smart Factory* would name a factory built on that foundation.

After the report was published, various German industrial associations constituted the *Plattform Industrie 4.0* (Industry 4.0 Platform) for the further development of the strategy. Among their produced outputs, we can highlight the *Reference Architectural Model Industrie 4.0* (RAMI 4.0) [Pla16] and their collab-

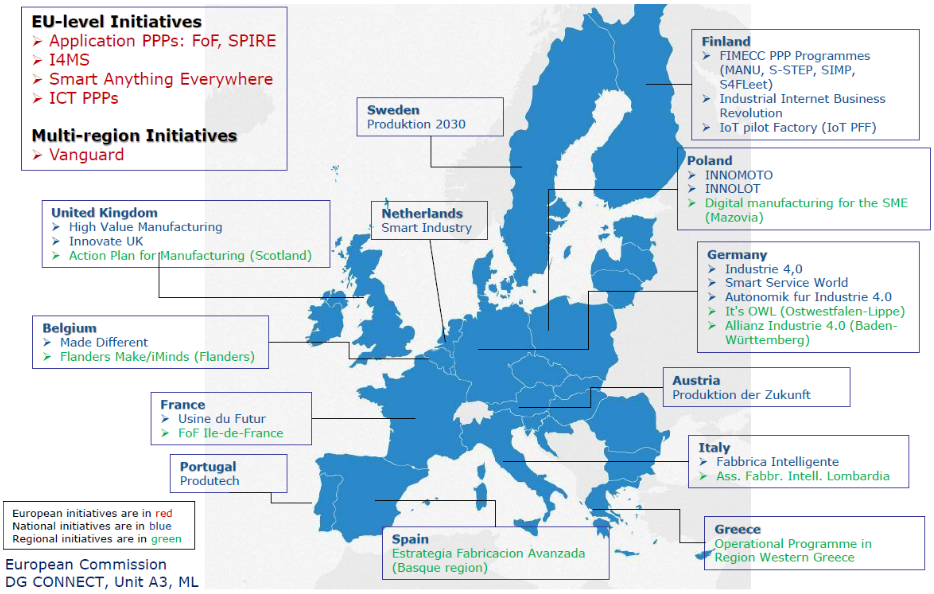

EU-level Initiatives
➢ Application PPPs: FoF, SPIRE
➢ I4MS
➢ Smart Anything Everywhere
➢ ICT PPPs
Multi-region Initiatives
➢ Vanguard
Sweden
Produktion 2030
Finland
➢ FIMECC PPP Programmes (MANU, S-STEP, SIMP, S4FLeet)
➢ Industrial Internet Business Revolution
➢ IoT pilot Factory (IoT PFF)
Poland
➢ INNOMOTO
➢ INNOLOT
➢ Digital manufacturing for the SME (Mazovia)
United Kingdom
➢ High Value Manufacturing
➢ Innovate UK
➢ Action Plan for Manufacturing (Scotland)
Netherlands
Smart Industry
Germany
➢ Industrie 4,0
➢ Smart Service World
➢ Autonomik fur Industrie 4.0
➢ It's OWL (Ostwestfalen-Lippe)
➢ Allianz Industrie 4.0 (Baden-Württemberg)
Belgium
➢ Made Different
➢ Flanders Make/iMinds (Flanders)
Austria
Produktion der Zukunft
France
➢ Usine du Futur
➢ FoF Ile-de-France
Italy
➢ Fabbrica Intelligente
➢ Ass. Fabbr. Intell. Lombardia
Portugal
Produtech
European initiatives are in red
National initiatives are in blue
Regional initiatives are in green
European Commission
DG CONNECT, Unit A3, ML
Spain
Estrategia Fabricacion Avanzada (Basque region)
Greece
Operational Programme in Region Western Greece

Figure 2.7: Overview of Digital Manufacturing Initiatives across Europe in January 2015 (extracted from [Eur15])

oration with other related initiatives such as the Industrial Data Space [OJS+16]. Their featured report "Industrie 4.0 in a Global Context" [KAG+16], published by Germany's National Academy of Science and Engineering (Acatech) and produced by some of the members of the original *Industrie 4.0 Working Group*, highlights the differences between the approach followed by *Industrie 4.0* and other similar initiatives worldwide. For instance, they describe Germany's strategic focus on "integrating information, communication and manufacturing technologies in smart, self-organizing factories", while USA's focus (and increasingly also China's) is on smart products, large Internet-based platform ecosystems and the new data-driven business models that are based on them.

2.2.3 Capital Equipment Manufacturers' Servitization in Smart Manufacturing Contexts

The increasing attention received by Smart Manufacturing applications and the possibilities they enable to transform manufacturing companies generate an important opportunity for Capital Equipment Manufacturers (CEMs, also referred as Capital Goods Manufacturers/Companies) to launch innovative business models thanks to a data-driven servitization approach. CEMs produce machine tools or infrastructure integrated in a larger production process run by a third manufacturing party, i.e. their customers. As these customers become interested in transforming the operation of their businesses towards a more Smart Manufacturing-oriented approach, CEMs can design value-added services [AAB13] that support their customers in that transformation. Next, we will examine the concept of *servitization* and its application in manufacturing, the key concept of *Smart Services* as the foundation for a data-driven servitization approach by CEMs and the technological challenges that this approach involves.

2.2.3.1 Servitization in Manufacturing

The concept of servitization was originally coined in 1988 [VR88] to refer to a trending interest by corporations in offering "bundles" [VR88] of customer-focused combinations of goods, services, support, self-service and knowledge, with services playing the lead role in those integrated systems. Since its coining, servitization has gained a lot of attention from the manufacturing industry, with a notably growing interest among these companies to adapt their business models to include an element of service provision, attaching value-added services to their products and thus increasing the value provided to customers [DUM$^+$15][LKY14]. This is mainly due to the need for a strategic change in manufacturing business models to be able to compete in a global market. The need for differentiation has increased dramatically in this globalization context, and servitization strategies help strengthening relationships with customers, hence locking out competitors [AAAS15][KA14].

Different research communities have studied this topic and contributed with knowledge related to the servitization of manufacturing [BLBK09]: services marketing, service management, operations management, product-service systems (PPS) and service science. This has also led to the proposal of diverse terms related to the transition from products to service-based solutions [AAAS15][Ni15] [PGL12]: "integrated product and services offering", "service infusion in manufacturing", "service-oriented value chain", etc. All these terms reflect the same idea of increasing competitiveness by transforming the business model into a services-based one.

2.2.3.2 Smart Services as the Focus for Capital Equipment Manufacturers' Servitization

The current context of intensive promotion of Smart Manufacturing, as previously described in this section, generates a strong opportunity and motivation for CEMs to shift their business models towards a servitization approach. Besides, these scenarios are focused on the creation of a new service by the CEM to an existing market, i.e. the customers operating in the manufacturing business sector where the CEM has traditionally marketed their products. Focusing on an existing market has the advantage that key components of the market (such as customers or competitors) are already familiar, which adds to the motivation of CEMs to launch servitization strategies. The servitization strategy of a CEM can be based on different types of product-service systems. Among the alternatives presented in [AAAS15], the "product and processes focused" type refers to those servitization approaches where the CEM offers services aiming at optimizing customer processes, which leads to increase efficiency and effectiveness of customer's operations. This category includes the data-driven services where the CEM helps their customers to evolve towards Smart Manufacturing.

The term *Smart Services* [KRH$^+$14] has been coined to designate these highly IT-based services where the growing volume of generated data is being captured and exploited to (among other uses) make product and process performance more visible and to design services dealing with their optimization [MSA15].

2.2.3.3 Technological Challenges for CEM's Data-driven Servitization

The provision of smart services involves important challenges for the CEMs, especially when it comes to the use of the key information technologies (IT) acting as drivers for Smart Manufacturing. These challenges are related not only to the understanding of how customer requirements impact the definition of those smart services, but also to the integration of required IT. This increases the difficulty in the development and launch of these IT-based services to the market [MSA15], as CEMs need to integrate new capabilities in order to effectively design those smart services. Dinges et al. [DUM$^+$15] present a survey on which technologies play a more important role when CEMs design their servitization approach. The answers given by the panel representing CEMs from different manufacturing sectors showed a high level of consensus over their top ranking: predictive analytics [LNR14]; analysis of existing datasets; remote communications to adjust, fix or update equipment or products; dashboard technologies; and case-based reasoning for pattern recognition and analysis. Advances in these technologies provide an important support for business model innovations among CEMs. The application of data-related technologies in order to monitor equipment and processes and to provide information about performance, equipment condition or usage enables the provision of data-driven services that extend the value provided by servitized CEMs [HEVY15].

However, apart from the struggles with service innovation frequently faced by product-centric companies [AAAS15], it is challenging for CEMs to keep pace with emerging opportunities arising from advanced technological development [DUM$^+$15]. Moreover, the survey conducted in [AAAS15] presents the main obstacles perceived by CEMs for their effective servitization, highlighting the "difficulty to monitor the product usage conditions and related data". These shortcomings result into the need for technological partners who are specialized in the involved key IT, so that their expertise can be combined with the CEM's knowledge of the targeted manufacturing sector to design the described smart services. This is the context of business opportunity where the role of *Industrial Big Data Services (IBDS) Provider* arises.

2.3 The Profile of Industrial Big Data Services Provider in Smart Manufacturing Scenarios

This section details further the business opportunity that motivates the emergence of the role of IBDS Provider and how their business model is integrated in the value chain that builds smart services for manufacturers who want to shift their business towards a Smart Manufacturing approach. The role that plays an IBDS Provider in this context and how their technological support must sustain the data lifecycle for those smart services allows us to identify relevant challenges for the ITS Providers aiming at building a business as IBDS Providers for the manufacturing market. These challenges constitute the motivation for our research focus in this work.

2.3.1 The Business Opportunity that Motivates the Emergence of IBDS Providers

The role of *IBDS Provider* describes a specialization of ITS Providers whose technological expertise is focused on the key enabling technologies laying the foundation for the data lifecycle in Smart Manufacturing applications. This data lifecycle is connected to the two core differential ideas that defined Smart Manufacturing's overarching goals: compiling the *manufacturing record* for each product and applying *manufacturing intelligence* to those compiled data. Nevertheless, the involved key technologies are not only linked to Big Data (as manufacturing indeed is an industry sector generating large-scale volumes of data), but also to the required solutions to capture and export data from manufacturing facilities (related to the concept of *Industrial Internet of Things*) and to centralize those data in massive cloud-based computing infrastructures for their subsequent processing (related to *Cloud Manufacturing*).

The specialized profile of an IBDS Provider enables a potentially strong synergy with those CEMs aiming at providing smart services as the means for their data-driven servitization. On one hand, most CEMs (especially manufacturing SMEs) don't possess the specialized know-how on the involved technologies and therefore require technological partners for the effective design and deployment of their data-driven servitization. On the other hand, the IBDS Provider can design *horizontal* solutions based on the required data-related technologies, i.e. solutions with cross-sector applicability (as opposed to sector-specific, i.e. *vertical* solutions). Thus, these solutions can be deployed in different manufacturing sectors by reaching agreements with CEMs that aim at providing smart services or directly with manufacturers demanding those services. Besides, as each CEM provides access to each particular sector and the specialized know-how on that manufacturing market, the IBDS Provider gains access to multiple manufacturing sectors and to a high replicability potential for the deployment of their IT solution (given that each CEM would aim at providing smart services for their various customers, and each customer would potentially own multiple facilities to be engaged in the use of the provided service).

In order to deploy the outlined business strategy, the IBDS Provider usually adopts the *Platform-as-a-Service* (PaaS) model to design their own horizontal solution (integrating Big Data technologies, Cloud Computing and Industrial Internet of Things) to solve the data gathering and integration needs in diverse manufacturing markets. In order to build and maintain this PaaS solution, the IBDS Provider can be supplied by a *Cloud Services Provider*, another specialization of ITS Providers that is focused on the provision of cloud-based infrastructure (mainly storage and computing resources) using an *Infrastructure-as-a-Service* (IaaS) model. The partnerships with manufacturers to deploy their platform in specific manufacturing markets allow the IBDS Provider to co-design with those manufacturers *sector-specific solutions* for their markets. Thus, the result of each partnership is a *vertical solution*, generally provided in a *software-as-a-service* (SaaS) model, with which smart services are provided for different manufacturing sectors. The integration of the described strategies is outlined in Figure 2.8.

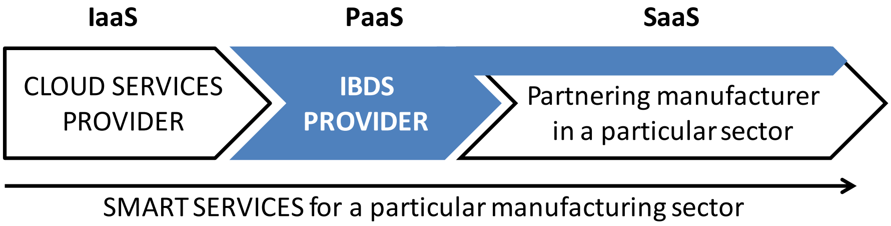

IaaS
PaaS
SaaS
CLOUD SERVICES PROVIDER
IBDS PROVIDER
Partnering manufacturer in a particular sector
SMART SERVICES for a particular manufacturing sector

Figure 2.8: Role played by an IBDS Provider in the provision of Smart Services

2.3.2 Challenges of Building a Business on Providing IBDS for the Manufacturing Industry

In order to design and build their Industrial Big Data Services for the manufacturing industry (i.e. the horizontal solutions described above), an IBDS Provider must face challenges that arise from two main sources: (a) the correct solving, from a technological point of view, of the IT solution that supports the data lifecycle in these environments and that will eventually enable the provision of smart services built on data, and (b) the creation of a business and the development of its strategy, integrating the requirements derived from that business context into the design of the required artifacts to compose the IT solution.

Reference models for a KDDM process provide a first approach to the data lifecycle that must be covered in order to design the required smart services built on data. Nevertheless, the foundational schema for a KDDM process [FPSS96] parts from a stage where indeed there exist some data to be processed. This is definitely not the case when the goal is to provide a service to owners of multiple manufacturing plants where most data-generating devices have only been designed for internal setting and supervision purposes, and where the deployed operational technologies lack the capabilities to export those data to other processing environments. Therefore, there is an important gap to be saved in order to ensure that there exists a repository of data to be exploited, i.e. even before the required data are available to initiate a KDDM process. Moreover, the main conceptual constructs proposed to guide the design of Big Data systems also suffer from a similar problem. They focus on providing efficient solutions for data processing in order to give an answer to diverse analytical use cases, but they assume a starting point where *new data* is already arriving to the repository (see

Figures 2.2 and 2.4) and therefore do not cover that previous stage in the data lifecycle.

Furthermore, this gap must not be solved in a way that only considers technological or analytical requirements. The application of these solutions in such a business context, where all involved agents (including the IBDS Provider) must solve their different business strategies, demands proposals that take the requirements derived from those strategies into account. For instance, CRISP-DM reference model [She00] provides a useful resource with the tight relationship between the *business understanding* and the *data understanding* steps, as outlined in Figure 2.9, but it does not cover the business reality of the IBDS Provider and the requirements derived from it that impact the way the solution is designed.

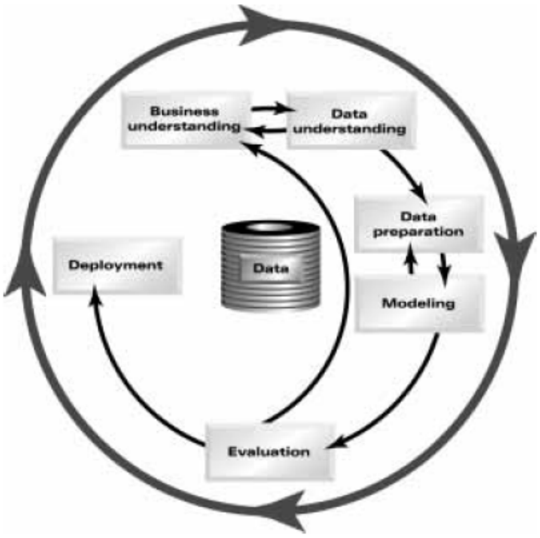


Figure 2.9: Phases of the CRISP-DM Reference Model (extracted from [She00])

Among the different challenges for the IBDS Provider that arise from the context described above, we highlight three specific challenges related to the early stages of the data lifecycle (i.e. before the availability of a data repository to be exploited by different processes with different approaches). These three challenges provide the focus for this research work and for the way our research method is designed in order to contribute to those challenges.

The *first challenge* is related to the source of costs that supposes for an IBDS Provider the need for being supplied by the *cloud services provider* of the required IT resources. This involves a substantial expense for the IBDS Provider, given the requirement of managing massive-scale amounts of data coming from all the manufacturing facilities where their solution is deployed. Therefore, the IBDS Provider needs to design an efficient data storage strategy that does not hamper the sustainability of their business and, at the same time, guarantees the resolution of the required smart services.

The *second challenge* is related to the required architecture for an IBDS Provider to design their platform. This platform must be composed of the necessary data capturing and integration infrastructure that (a) must be deployed in the analyzed manufacturing facilities where all relevant data must be extracted and (b) must allow the centralization of all captured data into a cloud-based repository for their later exploitation.

Finally, the *third challenge* is related to how an IBDS Provider contributes to the design of the required smart services for a specific manufacturing sector. This allows identifying how relevant stakeholders pose different data exploitation requirements to be solved by smart services and how the combination of those requirements impacts the different stages of the data lifecycle (and, in particular, those stages directly linked to the two previous challenges).

Chapter 3

Research Method

In order to produce relevant research contributions that address the selected challenges for an IBDS Provider, the research method followed to conduct this research work draws from two main sources with a solid conceptualization as research methodologies: *Design Science Research* [HMPR04][Hev07] and *Case Study Research* [Bas17][Eis89]. Moreover, both methodologies facilitate their straightforward integration with each other in order to sustain our research method.

On one hand, Design Science Research provides a methodology for research in Information Systems that aims at building purposeful design artifacts for the analyzed application domain that are both grounded in existing knowledge and codified as additions (i.e. contributions) to the knowledge base. The design process of those artifacts is based on the needs and requirements of the identified business problem (*relevance*) and the identification of synergies and opportunities regarding related work in the existing knowledge base (*rigor*).

On the other hand, Case Study Research facilitates guidelines to interact with a real-world business setting and their agents, so that the characterization of relevant stakeholders, strategies, needs and requirements is extracted and leveraged to provide a more detailed vision of the application domain and its business problems and opportunities. These elements allow giving purpose to the design artifacts to be produced as contributions. Moreover, the interaction with a real-world business setting facilitates the *field testing* (in terms of Design Science Research) of the artifacts and their core elements, in order to validate their applicability in real-world scenarios.

Thanks to the foundation provided by these two methodologies, we designed our research method combining key elements from both approaches. The designed method guided the construction of purposeful design artifacts for the research goals stated in the Introduction chapter, ensuring the relevance, rigor and applicability of our contributions. This chapter presents the aforementioned research methodologies and how they have been integrated to design a method to support the conducted research.

3.1 Methodological Grounding

The integration of Design Science Research (DSR) and Case Study Research (CSR) provides us with the required grounding to conduct our research. On one hand, DSR methodology is sustained by elements that match our research focus, where IT solutions play a crucial role as enablers of business strategies, and where there is a need for characterizing (a) the problems and opportunities of an application domain and the requirements that the proposed solutions must fulfill to be relevant for that domain, and (b) the required grounding on existing proposals to leverage what has already been proposed and to contribute where there is a clear opportunity to extend existing knowledge. On the other hand, CSR provides the guide to observe a real business setting, to extract relevant knowledge to characterize the application domain and to provide scenarios to conduct field testings of the proposed solutions. This section presents the fundamentals of both methodologies and how they can be integrated in our research method.

3.1.1 Design Science Research as a Methodological Ground for the Research Method

Design Science Research (DSR) is a research methodology based on the application of the design-science paradigm to the research on Information Systems (IS). IS research is focused on the interaction of business strategy, IT strategy, organizational infrastructure, and IS infrastructure. IS research is especially relevant for scenarios where IT solutions are enablers of business strategies [HMPR04].

Design-science is a problem-solving paradigm. In this context, *design* should be understood as the "act of creating an explicitly applicable solution to a problem". Therefore, design science addresses research through the *design of artifacts*[1] in order to meet the business or organizational needs identified as the starting point. In the design-science paradigm, knowledge and understanding of a problem domain are achieved in the design process of artifacts [HMPR04].

The main elements of DSR are thoroughly described in [HMPR04][Hev07], and summarized in Figure 3.1. Next, the main concepts of the three *cycles of activities* that constitute this methodology are summarized, based on their detailed descriptions presented in [HMPR04][Hev07].

The central piece of the DSR methodology is the *design cycle*, which addresses the *building* and *evaluation* of *design artifacts*. Artifacts constructed in DSR are rarely full-grown information systems. Instead, four types of design artifacts are identified as potential results to be produced by design science research in information systems:

- *Constructs.* They provide the language in which problems and solutions are defined and communicated.

[1] The result of design science research are *design artifacts*, which must not be confused with *IT artifacts*, i.e. deployed implementations of IT solutions.

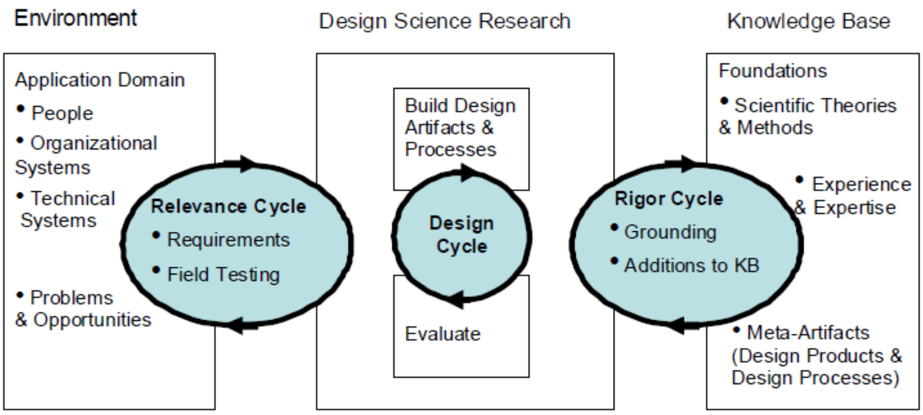


Figure 3.1: Design Science Research methodology (extracted from [Hev07])

- *Models.* They use constructs to represent a real-world situation and the connection between problem and solution components.
- *Methods.* They provide guidance on how to solve problems. They can range from formal (e.g. mathematical algorithms) to informal (e.g. textual descriptions) approaches, with combined possibilities between both ends.
- *Instantiations.* They show that constructs, models and methods can be implemented in a working system. They enable a more concrete assessment of an artifact's suitability to its intended purpose.

Two main processes are identified with respect to the *design artifacts* to be produced: *build* and *evaluate*. Thus, a *design process* (*design cycle*) is composed of the activities in order to build a design artifact and to evaluate it. Artifacts are considered purposeful as long as they help address the stated problem. The evaluation provides a better understanding of the problem and an assessment of the adequacy of the artifact (its applicability to the problem and the utility it provides to solve it) and the conducted design process.

The contributions of design-science research are assessed as (a) they are applied to the business need in an appropriate environment and (b) they add to the content of the knowledge base for further research and practice. These two ideas are linked to the *relevance* and *rigor* cycles in the design science research methodology.

The *environment* where the design artifacts are aimed at is characterized as an *application domain* where different people, organizational systems and technical systems interact with each other. The identification of problems and opportunities (i.e. business needs) is crucial for design science research, as it is motivated by the desire to improve the environment by the introduction of innovative design artifacts. This is why framing research activities to address business needs ensures *relevance*. Thus, the *relevance cycle* connects the contextual environment of the research project with the design science activities and provides *requirements* (i.e. those business needs identified as problems and opportunities) as input for the research, as well as the acceptance criteria with which to assess the applicability of the proposed design artifacts.

Meanwhile, the *rigor cycle* provides existing knowledge (grounding theories and methods, experience and expertise, meta-artifacts) to the research project in order to ensure its innovation and to guarantee that the proposed design artifacts are research contributions. Indeed, the rigor cycle allows establishing the difference between design science research and routine design. The key differentiator is the clear identification of a contribution to the archival *Knowledge Base* (KB) of foundations and methodologies. While the state of the art in the application domain (i.e. extant literature and related work) and the existing artifacts and processes provide *grounding* for the proposals, the proposed design artifacts (DSR results) must constitute a relevant addition to the KB.

Therefore, the joint assurance of *rigor* and *relevance* is what ensures that design science research results are, on one hand, valid research contributions for the academic audience and, on the other hand, useful contributions for the practitioner audience and their environment (application domain).

3.1.2 Case Study Research as the Source of Relevance and Applicability of Contributions

From the perspective of research in Information Systems (IS), Case Study Research (CSR) has traditionally been considered an approach that allows IS researchers to learn by studying the innovations put in place by practitioners and capturing knowledge from it, so that they can later formalize this knowledge. Indeed, CSR is especially suited to IS research because this research field typically addresses recent technological breakthroughs and their interest from a organizational, rather than technical, point of view [Bas17].

CSR is particularly appropriate for practice-based problems where both the experiences of actors and the context of action are critical, and is considered a viable IS research strategy when the researcher can study IS in their natural setting (i.e. without controlling or manipulating subjects or events), in order to understand the nature and complexity of the processes taking place. In this sense, it differentiates from Action Research in that CSR refers to research efforts where *research questions* (i.e. the focus of the research effort) are specified prior to the study by researchers who take the role of observers rather than practitioners [Bas17].

Next, the five main elements that define how to conduct CSR are summarized, based on their detailed descriptions presented in [Bas17][Eis89].

Unit of Analysis. It has to be determined the most appropriate unit of analysis for the research project (individuals, groups, an entire organization) and what generalizations (i.e. to other organizations, individuals, etc.) are expected to obtain.

Single/Multiple Case. While most research efforts require multiple cases, various scenarios justify the usefulness and appropriateness of a single case, e.g. if the situation has previously been inaccessible to scientific investigation, or if it represents a critical or unique case.

Site (Setting) Selection. In the event of research on organization-level phenomena, the setting selection should be based on the characteristic of firms, i.e. industry, company size, vertical or horizontal integration, etc. A well-defined research focus and the initial definition of the research questions allow the researcher to specify the kind of organization to be approached. An important criterion is that the problem of interest should be observable as transparently as possible in the selected setting. The researcher must contact the individual with enough authority (according to the topic of study) to approve the project. The cooperation must be sustained by two key points: ensuring confidentiality and providing benefits to the organization.

Data Collection Methods. Multiple data collection methods are typically used in CSR, with the goal of obtaining a rich set of data surrounding the research problem and capturing its contextual complexity. Specific data to be collected will depend on the research questions and the unit of analysis.

Data Analysis and Exposition. The analysis of case data depends heavily on the integrative powers of the researcher. As much as possible, the contextual richness of the case study should be presented. The research should move from objectives and questions to assumptions and design choices and, finally, to results and conclusions. The emergent concepts in these results should be developed along with their contrast with existing literature, in order to identify similarities, contradictions, synergies and opportunities.

3.1.3 Integrating Design Science and Case Study in our Research Method

The application of DSR as a foundation for our research method has clear synergies with our research focus. This can be drawn from the clear role of IT solutions as enablers of the business strategies of all agents (as outlined along chapter 2) and our aim to contribute with design artifacts for the identified challenges, in order to help the targeted organizations (i.e. IBDS Providers) to meet their business needs.

Thus, the DSR methodology outlined in Figure 3.1 provides us with the following methodological foundation:

- The *environment* is characterized by the Smart Manufacturing scenarios around IBDS Providers. The extraction of real-world *requirements* from a business setting that represents a significant instance of those scenarios is what will confer the practical foundation on the *design artifacts* to be built.
- The revision of the *knowledge base* (i.e. extant literature and work related

to the addressed areas) provides *grounding* for the *design artifacts*, in the shape of synergies and differences with related work.

- The *design artifacts* (i.e. constructs, model, methods, instantiations) will be *built* upon the foundation provided by (a) the *requirements* characterizing the targeted scenarios and (b) the synergies with related work and the identified gaps in the knowledge base that create the opportunity for new *additions to the KB* together with the contrast of their applicability in the targeted domain.

On the other hand, the application of CSR is particularly appropriate for our research focus, given that the problem to observe is not meant to be analyzed from an isolated, laboratory perspective. Instead, it is required a first-hand observation of a real-world business setting where the relevant agents to all levels (IBDS Providers, CEMs with servitization strategies and manufacturers pursuing a Smart Manufacturing approach) interact with each other to build the required services, according to their respective business strategies. This approach allows us to identify and understand all the practical requirements derived from these settings. Identifying these requirements is crucial in order to fulfill two goals:

- Injecting the necessary real-world features into the characterization of the targeted scenarios serving as *environment* (according to the DSR methodology).
- Integrating those practical requirements into our proposed design artifacts in order to ensure the *relevance* of our contributions.

Moreover, conducting a case study in a real-world business setting provides a scenario to assess the applicability of the proposed design artifacts.

3.2 Requirements for the Real-World Business Setting of our Case Study

Given the research challenges where this work is focused in, IBDS Providers constitute the focus of our contributions. They are ITS Providers whose expertise is specialized on the technological foundation for the data lifecycle in Smart Manufacturing applications. Our goal therefore is to contribute with purposeful design artifacts aimed at the profile of IBDS Providers and the Smart Manufacturing scenarios where they can develop their business strategy and supply their services. The scope of our research is focused on those profiles of IBDS Provider with their own business strategy, based on establishing strategic partnerships with either CEMs pursuing a data-driven servitization strategy or directly with manufacturing companies, in order to deploy smart services for specific manufacturing sectors.

The selection of the business setting for our case study had two main goals. The first goal was *characterizing the Smart Manufacturing scenarios* where IBDS

Providers supply their services and the main manufacturing agents involved in those scenarios. This would provide the characterization of the *environment*, in terms of Design Science Research, in which information systems-related problems are analyzed, and from where requirements are extracted in order to sustain the relevance of contributions proposed to solve those problems. The second goal was accessing a real-world business setting where to conduct *field validations* of the core components designed for our contributions, in order to contrast their applicability and practical utility in such scenarios.

This required that the setting should allow us to conduct a *two-level case study.* On one hand, the analysis should focus on the business context around an IBDS Provider, observing diverse types of Smart Manufacturing scenarios where an IBDS Provider supplies their services and collaborates with different profiles of manufacturing companies leveraging these *Industrial Big Data Services.* This would allow us to characterize the Smart Manufacturing scenarios that would constitute the *application domain*, i.e. the environment depicted in the DSR methodology where to orient our contributions to. Moreover, it should provide requirements derived from the business strategy of an IBDS Provider and their partners, in order to supply them as input to the *design process*, so that we could guarantee the *relevance* of our contributions. In this sense, targeting a company with the profile of SME would facilitate the goal of accessing the top-level management and enabling a direct access to their business strategy, so that we could capture it better for our characterization of the environment.

On the other hand, it should grant direct access to the collaboration projects involving IBDS Providers, servitized CEMs and manufacturers interested in adopting a Smart Manufacturing approach for the operation of their production process. This would allow us to observe directly the initial steps of these projects and the deployment of the required IT solutions for the capture, visualization and analysis of the data generated along the production process in monitored facilities. Furthermore, it would give us the opportunity to conduct field validations in order to integrate and contrast our proposals for those IBDS Provider's roles that could leverage them when conducting these projects. Those field validations would involve working with real organizations in global-scale scenarios, business requirements and real data coming from operating factories, something that solves one of the main challenges that has been historically faced by the research on Intelligent Manufacturing Systems [MVF$^+$07]. In order to ensure these goals, several key decision factors were established for the selection of the appropriate setting: the openness of the top-level management representatives of the involved organizations, the accessibility to their companies and facilities, and the possibility to characterize diverse Smart Manufacturing scenarios from different sectors and involving multiple monitored facilities worldwide, which ensures a high degree of representativeness. Thus, we could conduct the aforementioned characterization and field validations as thoroughly as possible.

However, accessing such a real-world business setting in these conditions also comes with a compromise: gaining full access to such a complex and interconnected business context hampers the possibility of conducting a case study with more than one IBDS Provider (given that different IBDS Providers might be potential competitors again each other). This is closely linked to the requirement of confidentiality when conducting a case study in order to gain cooperation with the observed organizations, to access their business strategies and to ensure a more transparent observation of the problem of interest.

Nevertheless, we decided in favor of conducting our case study with one IBDS Provider with whom we could access to the aimed business context with full detail. This decision was also supported by the fact that our analysis was not only focused on the IBDS Provider as an organization, but on the business context around the IBDS Provider (and where they must offer their solutions). Therefore, even collaborating with one IBDS Provider, we would gain access to a rich business setting with multiple instances of manufacturing sectors where they establish partnerships with CEMs developing their data-driven servitization strategy and with manufacturers demanding smart services. Moreover, we would gain access to the competitive market of IBDS Providers in general, allowing us to characterize different types of organizations fitting the profile of IBDS Provider.

3.3 Method to Build Design Artifacts as Contributions for the Research Challenges

Parting from the definition of the two-level case study, we organized our research according to a method sustained by the main elements described in the DSR methodology, so that we could *design* purposeful contributions to the selected research challenges. This section outlines our method steps, whose results are presented in detail in the following chapters.

The first step in our research method will be to extract the relevant features of the analyzed real-world business setting (via the two-level case study) that can be integrated into a better characterization of the targeted Smart Manufacturing scenarios. For that purpose, the conducted observation will focus on:

- Identifying the relevant roles and stakeholders interacting in these scenarios.
- Characterizing the business strategies of the main agents in such scenarios, as well as the needs and requirements that are derived from those strategies and how each relevant agent's requirements are affected or related to those of other agents.

The result of this observation will allow us to consolidate a more practical vision of the targeted Smart Manufacturing scenarios. These scenarios will constitute the *application domain* where we characterize *problems* and *opportunities* to build *design artifacts* as our contributions. Thus, the contributions proposed in this work will be oriented to provide solutions to IBDS Providers taking into account the main practical requirements of the scenarios where they supply their services.

Then, in order to create our contributions (*design artifacts*) for each of the identified challenges, we will conduct the following set of steps for each goal (outlined in Figure 3.2) supported by the key elements in the DSR methodology:

1. Once we compile the characterization of the targeted Smart Manufacturing scenarios, we will extract those needs and requirements particularly relevant to the addressed research goal. This will not only provide a more detailed vision of the problem and the basic elements of the required solution, but also a guarantee of its *relevance* for the characterized environment.
2. Once the problem and the type of solution are framed, we will review the *archival knowledge base* to examine existing work related to the vision, technologies, etc. identified as required elements for the solution. This revision will allow us to identify synergies, as well as gaps and differences, with existing proposals. This will provide the rigorous contrast to verify the opportunity for relevant, well-grounded contributions to the knowledge base.
3. Given that input of *relevance* and *rigor*, we will conduct a *design cycle* with three steps:
    a. A first *build* step where we will extract the core concepts and elements that will sustain our proposed design artifact.
    b. An *evaluate* step in order to validate the applicability of those core concepts and elements, through a *field testing* in the real-world business setting where we conduct our case study research.
    c. A second build step, once the applicability has been validated, to formalize a *design artifact* as the proposed contribution (*addition to KB*) sustained by those contrasted concepts and elements.

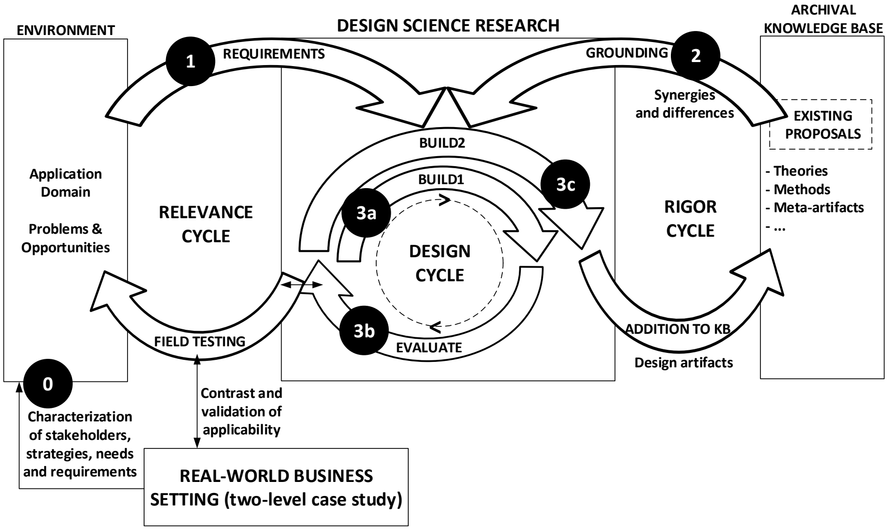


Figure 3.2: Steps to build our contributions based on DSR

This method will be applied to contribute with design artifacts for the *three challenges* for an IBDS Provider posed in the previous chapter, related to:

1. A more efficient data storage strategy that ensures a sustainable platform for their business.
2. The required abstract architecture for the data capturing and integration infrastructure to sustain their platform.
3. The collaborative design process with their partners of the required smart services for a specific manufacturing sector.

Chapter 4

Characterization of Targeted Smart Manufacturing Scenarios

One of the key milestones in this research work was the selection of a real-world business setting were to conduct our case study. The opportunity to access such a setting was facilitated by an IBDS Provider with which our research group had maintained contact during the previous years. Once we started knowing in detail the business context around this IBDS Provider and the manufacturing sectors where they had established partnerships to supply their services and deploy their solutions, we concluded that this setting could provide a very useful scenario to conduct our case study. This decision was supported by the openness of the top manager of this IBDS Provider to gain full access to their business context (i.e. customers, competitors, providers, etc.) and their strategy towards the targeted market. Moreover, it also granted access to one of their collaboration projects with a particular CEM, facilitated by the availability of the top managers of this CEM to access their servitization strategy as well. Therefore, the setting was also identified as extremely valuable to observe and analyze the provision of Industrial Big Data Services to a servitized CEM, and on how this CEM defines the provision of smart services towards their customers.

The valuable insights extracted from the observation of this business setting enabled the abstraction and characterization of the main agents involved in these scenarios, and the delimitation of our research context in terms of the Smart Manufacturing scenarios where our contributions will be targeted at. Thus, we can differentiate:

- The characterization of the targeted Smart Manufacturing scenarios around IBDS Providers, i.e. those generic application scenarios towards our contributions will be targeted at. The characterization of these scenarios is based on the interaction between three main agents, i.e. *IBDS Providers*, *servi-*

tized CEMs and *smartized manufacturers*, and the collaboration projects in which they engage in order to fulfill their business strategies.

- The real-world business setting where we conduct our case study, as a relevant instance of those scenarios. In this sense, this setting provided us with the opportunity to access the early stages of collaboration between the relevant agents in these scenarios. Furthermore, this also gave us direct access to the deployment of the required IT solutions in one of the targeted facilities, where we conducted *field testings* of the core components of our contributions.

This chapter focuses on presenting the characterization of the Smart Manufacturing scenarios where we target our contributions for IBDS Providers. First, the chapter presents the real-world business setting where we conducted our case study, as a relevant instance of the targeted scenarios. The second section of this chapter presents a characterization of the main agents involved in these scenarios. This characterization details their business strategies, the main features of their collaboration projects and the main roles involved in them, as well as practical requirements for the effective deployment of these projects. The chapter is closed with some conclusions linking this characterization with the provision of requirements and field validation for the contributions of this research work.

4.1 Characterization of Analyzed Agents in our Case Study

Our case study allowed us to integrate ourselves in the real-world business setting around an IBDS Provider[1] supplying their services to diverse Smart Manufacturing scenarios. Furthermore, it also granted a direct access to a particular case of collaboration between this IBDS Provider and a servitized CEM leveraging these solutions to deploy their data-driven servitization strategy, and to the initial steps of the deployment of smart services for a manufacturing facility owned by one of this CEM's customers, i.e. a *smartized manufacturer*. Thus, this setting provided important advantages:

- From the business point of view, it gave us the opportunity to observe the genesis of a servitization strategy, the deployment of the IT solution to support it and the practical problems faced along the process. It also gave us access to interact with the main business stakeholders in these manufacturing sectors. Furthermore, as the involved IBDS Provider and the servitized CEM were both SMEs[2], we could interact more easily with their top-management representatives and access the details of their business strategies.

[1] For business confidentiality purposes, the names of the companies involved in this research work will not be disclosed in this dissertation.

[2] SME: Small and Medium Enterprise

- From the technical point of view, this also enabled the opportunity to access the raw data to be captured and integrated in one of these real-world industrial environments (i.e. real data coming from operating factories), as well as to familiarize ourselves with the challenges that this task poses to an IBDS Provider.

We accompanied these companies during 30 months with detailed access to the IBDS Provider's business context in general and to the manufacturing sector of the aforementioned servitized CEM in particular. During this period of time different data collection methods were used to observe and interact with all involved agents (periodical interviews with managers and technicians from all involved organizations, direct observation of various business and technical meetings, visit and field work in one of the monitorized manufacturing plants). This section presents with further detail the characterization of the aforementioned IBDS Provider, servitized CEM and smartized manufacturer, as well as their interaction (outlined in Figure 4.1) in the analyzed case study.

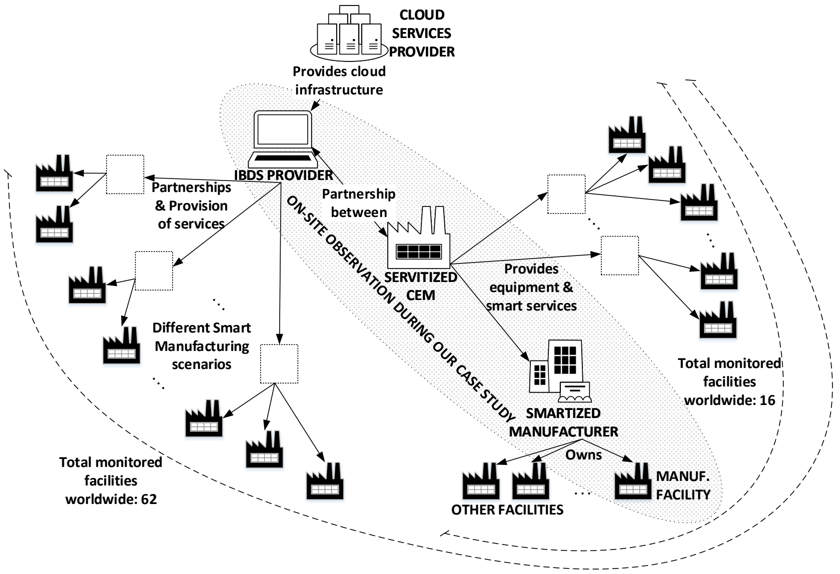

CLOUD SERVICES PROVIDER
Provides cloud infrastructure
IBDS PROVIDER
Partnerships & Provision of services
Partnership between
ON-SITE OBSERVATION DURING OUR CASE STUDY
SERVITIZED CEM
Provides equipment & smart services
Different Smart Manufacturing scenarios
SMARTIZED MANUFACTURER
Owns
OTHER FACILITIES
MANUF. FACILITY
Total monitored facilities worldwide: 62
Total monitored facilities worldwide: 16

Figure 4.1: Schema of the organizations interacting in the business setting analyzed in our case study

4.1.1 The Analyzed IBDS Provider

The IBDS Provider around which the case study is conducted is an IT-based SME whose business model is focused on the deployment of IT solutions based on three of the main Key Enabling Technologies (KET) supporting smart manufacturing: Big Data, Internet of Things and Cloud Computing. They deploy

Industrial Internet of Things (IIoT) devices that connect to the low-level IT infrastructure operating in manufacturing plants, in order to capture raw data generated by industrial sensors regarding some magnitudes or indicators of interest. These captured raw data (time series generated by the continuous operation of the manufacturing process or equipment to be analyzed) are automatically transmitted to a cloud computing environment, where the IBDS Provider supplies different exploitation functionalities on those data. The cloud computing infrastructure is provided by a *cloud services provider*, supplying the required worldwide accessibility, computing power and different types of storage for the centralized data. Although this cloud services provider owns data centers in different countries of Europe and America, the specific data center supporting the services supplied to the IBDS Provider is located less than 200 km away from the IBDS Provider's premises. This choice is motivated by this provider's higher security standards and an easier accessibility in order to convey a trust guarantee to partners, even offering them the possibility to visit the premises where their data are securely stored.

The market strategy of this IBDS Provider is mainly aimed at CEMs from different sectors that sell their equipment to manufacturing customers worldwide and, therefore, deploy their equipment in manufacturing plants all over the world. The IT solutions deployed by this IBDS Provider allow these CEMs to adopt data-enabled servitization strategies, aimed at providing their customers not only with equipment but also with value-added services based on the exploitation of data generated by that equipment and by other components integrated along the manufacturing process. Given the global scale of these customers (potentially owning multiple plants worldwide), the scenarios where CEMs aim at offering these data-enabled services are characterized by the need for gathering and processing massive, distributed data to analyze a manufacturing process (or a particular step of that process) under different settings. Depending on the specific manufacturing business sector where these CEMs operate and on the specificities of their servitization strategy and the data-enabled services to be provided, the massive data to be gathered might be related to areas such as the control of product quality or process efficiency, fault diagnosis, predictive maintenance of equipment, etc.

Thanks to the IT solutions supplied by this IBDS Provider, CEMs from different sectors are being provided with the tools to servitize their business models. As of March 2017, more than 60 manufacturing facilities worldwide are currently being provided with different exploitation capabilities for the large-scale data they generate (approximately 400 new GB of data every week). Table 4.1 summarizes the application domains where this IBDS Provider has currently deployed their IT solutions.

In order to deploy their IT solutions for the servitized CEM's customers demanding smart services, there are three main roles played by the IBDS Provider's personnel: *project manager*, *deployment technician* and *data engineer*. The project manager is in charge of the different collaboration projects with servitized CEMs, in order to manage and supervise the provision of services for smartized manufacturers. Whenever some field work is required in one of the monitored facilities owned by the smartized manufacturer, the deployment technician trav-

Manufacturing sector / Application domain	**Monitored processes and indicators**	**Smart Manufacturing goal**	**Number of monitored facilities worldwide**
Aerospace and Railways / Machining and grinding	Global process monitoring	Assessment of equipment condition, remote analysis, prediction of remaining useful life	3
Broaching and cold forming	Global process monitoring	Assessment of equipment condition	1
Electrical machining	Global process and equipment monitoring	Optimization of equipment uptime	2
High-precision machining	Global process monitoring, interoperability with vibration analysis systems	Assessment of equipment condition and impact of vibration on the overall process	3
High-precision milling and broaching	Global process and vibration monitoring	Assessment of equipment condition and process optimization	20
Industrial cleaning	Global process monitoring	Process optimization	5
Industrial professional training	Global process and equipment monitoring	Training application for new maintenance strategies and process optimization	1
Laser cutting and high-precision grinding	Global process and equipment monitoring	Assessment of equipment condition, process optimization, failure prediction	6
Paper processing	Vibration monitoring	Assessment of equipment condition	1
Polyurethane foam production	Global process monitoring, interoperability with facility management systems	Process optimization	16
Processing of metallic coils	Global process control	Assessment of equipment condition	3
Stamping waste management	Vibration monitoring	Assessment of impact of vibration on the overall process	1

Table 4.1: Application domains where the IT solutions supplied by the analyzed IBDS Provider are deployed

els to that location. This is mainly required when the necessary adjustments to connect to all relevant low-level sources cannot be completed in a remote way. Last, the data engineer is in charge of supervising the technological platform capturing and storing data from monitored facilities, as well as the quality and correct visualization of these data.

4.1.2 The Analyzed Servitized CEM

The collaboration with the analyzed IBDS Provider, apart from an overall perspective of their partnerships with companies in diverse manufacturing sectors and the deployment of their IT solution in manufacturing facilities worlwide, also facilitated the direct access to the particular case of one of the CEMs establishing a partnership with the IBDS Provider in order to transform their business model via servitization.

The analyzed CEM is a manufacturing SME, so far focused on selling equipment and storage infrastructure for larger manufacturing companies in the chemical manufacturing sector of polyurethane foam production. The manufacturing process for which this CEM provides their equipment is focused on the transformation of raw materials (petroleum derivatives) into foam blocks of different physical features and dimensions, which will be later machined into specific shapes and sizes. This CEM's customers, i.e. the manufacturers executing that process, are medium-size companies, producing 5-15 million kg of foam blocks per year with an estimated annual profit of 1 million euros on average. This chemical manufacturing sector is spread worldwide and manufacturing plants are built close to the locations where the product is going to be bought and used. The same company may own several manufacturing plants, each in a different country. Therefore, the equipment provided by this CEM is used in tens of manufacturing plants all around the world.

The center piece of this chemical manufacturing sector is a continuous production process, with similar high-level phases among the plants executing it. This process involves different chemical and mechanical subprocesses to transform raw materials into the final product. Nevertheless, depending on the specific plant, these subprocesses might be implemented with equipment from different providers and with different setting features. The degree of automation varies along these phases: some are highly automated (the core of the chemical transformations involved in the process) whereas the mechanical phases combine automated and manual operation.

The core idea in this CEM's servitization strategy is to offer services to increase the value of their customers' production systems thanks to an optimized performance. In terms of evolution scenarios for automated production systems [VHFST15], the motivation is a new target for production performance via an extension of their capabilities. In this chemical manufacturing sector it is estimated that 80-85% of total costs along the process are due to the acquisition of raw materials. Therefore, even a small optimization in these costs might generate massive savings at the end of the year. Nevertheless, the management of key parameters controlling the process is done without a solid scientific foundation or analytic formula. This results in considerable non-quality extra cost in the final product, due to waste or impurities.

This scenario enabled a business opportunity for this CEM to start offering smart services based on the capture, analysis and exploitation of relevant data along this manufacturing process. These services are targeted at helping their customers making their production more efficient and increasing the quality of produced goods. For that purpose, this CEM establishes partnerships with, on one hand, the IDBS Provider for the supply of the required IT solutions and, on the other hand, those of their customers interested in a more Smart Manufacturing-oriented and optimized operation of their manufacturing process. The CEM provides a *plant engineer* to coordinate the provision of the IT-based services of the IBDS Provider with the deployment of this CEM's equipment. Furthermore, in order to develop the required smart services, this CEM also hired a *data scientist* to develop specific analytical models upon the captured and integrated data by the IBDS Provider's platform.

Thanks to these partnerships, 16 foam block production facilities worldwide owned by different companies are currently being monitored. More than 400 indicators generated by sensors along the foam block production process are being continuously monitored in each facility, as outlined in Figure 4.2. This results in 2 million raw measurements per hour and plant. The goal is to gather and analyze all these indicators along the process in different facilities, in order to build a global view of the whole process and to relate potentially influencing process parameters with the quality of produced goods, which can be later exploited to improve this process' efficiency.

4.1.3 The Analyzed Smartized Manufacturer

The business setting analyzed in our case study also granted us direct access to one of the collaborations between the IBDS Provider, the servitized CEM and one of the latter's customers interested in adopting a Smart Manufacturing approach for the operation of the foam block production process in their facilities.

This manufacturer produces different products, all of them based on the use of polyurethane foam, for different markets. They own facilities focused on different manufacturing processes: foam block production, foam block machining for different market uses, and fabrication of the required components to be integrated with foam shapes in order to produce furniture for large chain stores worldwide. A total of 2,500 employees work along their facilities.

One of these facilities, founded in 2015, owned by this manufacturer is focused on the foam block production process previously outlined. The equipment provided by the analyzed servitized CEM is deployed in this facility. The collaboration between the three analyzed organizations enabled an agreement with the manufacturing company owner in order to design and deploy smart services for this manufacturing company, focusing first on this facility.

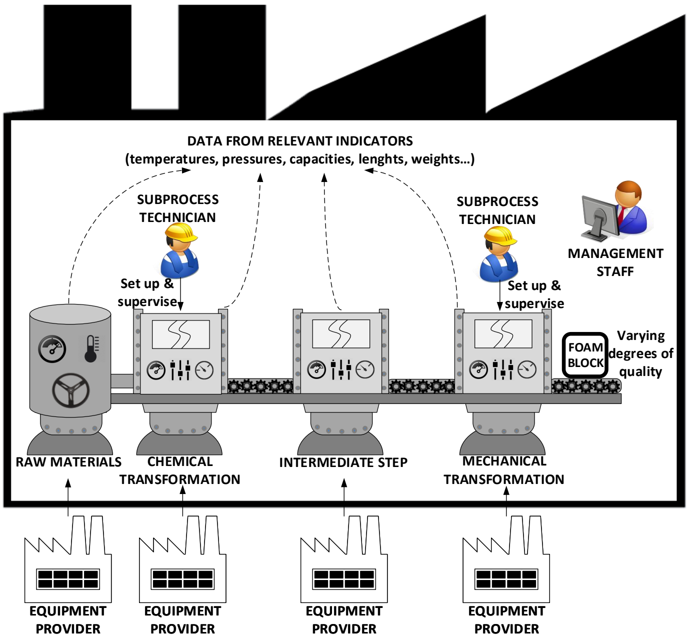


Figure 4.2: Schema of the manufacturing process in a foam block production facility

4.2 Smart Manufacturing Scenarios Targeted in this Research Work

The real-world business setting presented in the previous section represents an instance of the Smart Manufacturing scenarios where the contributions of this research work are targeted at. We base our conceptualization of Smart Manufacturing scenarios on the definition provided by [DEP$^+$12] for the concept of Smart Manufacturing, as presented in chapter 2. Thus, the two core ideas that sustain a Smart Manufacturing approach are (a) the compilation of data from sensors and other observations to create a *manufacturing record* for each product, and (b) the application of *manufacturing intelligence* on those data in order to generate analytics models that can be leveraged to optimize production. Parting from there, our targeted Smart Manufacturing scenarios are defined by *three main types of agents* that interact in the provision and consume of smart services. These agents are *IBDS Providers*, *servitized CEMs* and *smartized manufacturers*, whose strategies, needs and requirements are described along this section.

Two types of Smart Manufacturing scenarios are defined around IBDS Providers (see Figure 4.3), depending on whether the provision of smart services to smartized manufacturers is done via a direct partnership between those two agents or via a partnership with a servitized CEM, i.e a *servitization* scenario. Indeed, the goal of this research work is to provide IBDS Providers with contributions that can be leveraged to facilitate the development of their business strategy and the deployment of their services in any of those types of Smart Manufacturing scenarios.

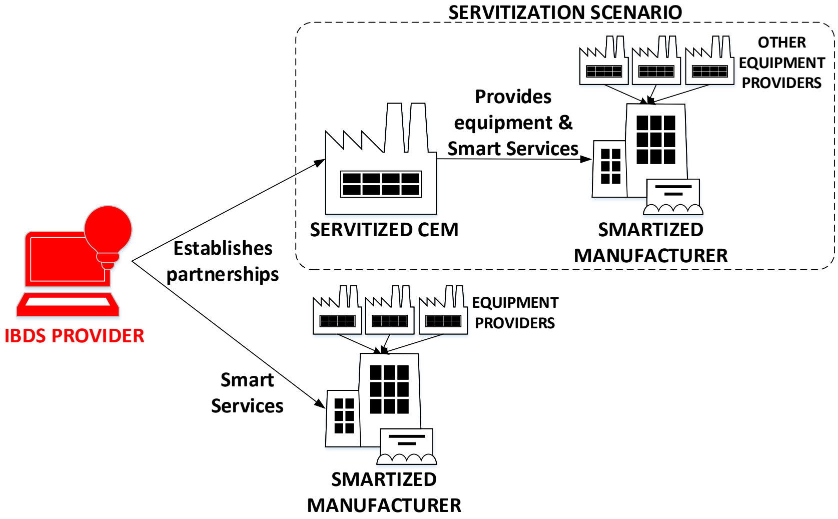


Figure 4.3: Differentiation between servitization and non-servitization scenarios

Servitized CEMs are those companies who have been focused so far on selling equipment to manufacturers in a particular sector and now aim at transforming their business model adopting a servitization approach. Their new business model is based on the provision of smart services for their customers, so that they can shift the operation of their manufacturing process towards a Smart Manufacturing-oriented approach. The definition previously specified for Smart Manufacturing scenarios, focused on the compilation and exploitation of manufacturing records of products, allows us to characterize the servitization scenarios where our research is aimed at. Some servitization scenarios might be focused on the exploitation of data only related to the equipment provided by the CEM, e.g. to attach a predictive maintenance service to the equipment sold to their customers. However, based on the aforementioned focus, the analyzed scenarios in our research include those where the smart services to be provided aim at supporting the exploitation and analysis of data from *the whole manufacturing process* operated by customers, so that they can compile, analyze and exploit manufacturing records for each product unit.

Smartized manufacturers are the manufacturing companies who want to shift the operation of their manufacturing process towards a more Smart Manufacturing-oriented approach. Based on the definition provided by [DEP$^+$12], they aim at extracting value from the data generated along the operated process. These data compose the manufacturing record (sometimes also referred as *digital twin* or *cyber-twin* [LBK15]) for each product. The analysis and exploitation of manufacturing records provide the input for decision-support and decision-guidance systems for production optimization [BKM$^+$14].

4.2.1 IBDS Providers

Attending to the typology and nature of organizations playing the role of IBDS Provider in real-world business settings, we can identify two main types of scenarios:

(A) Where the IBDS Provider is an *independent organization.* In this type of scenarios, an IT-focused organization (an IT-based company or a research center/institute) develops their own business strategy to provide the required platform and services for independent CEMs so that they can *servitize* their business model, or for smartized manufacturers to directly leverage those services in their facilities.

(B) Where *a CEM integrates the required IT capabilities.* In these scenarios we find a medium- or big-size equipment manufacturer that also possesses the required resources and skills (in their own organization, as a shareholder of a specialized provider or as a member of the same holding group) to develop *Industrial Big Data Services* to supply smart services to their customers. In this case, we focus on the scenarios where there exists a subsidiary or spin-off organization developing their own business strategy as an IBDS Provider. Thus, this organization can provide these services to other companies in different manufacturing sectors, much in a similar way to the scenarios grouped above as (A).

Thus, the market strategy of an IBDS Provider is aimed at establishing partnerships:

- With CEMs from different manufacturing sectors that want to transform their business model with a data-driven servitization approach, in order to offer smart services to those of their customers wanting to adopt a Smart Manufacturing approach, or
- Directly with those manufacturing companies who want to shift the operation of their facilities towards a more Smart Manufacturing-oriented approach, in order to optimize their production process along their facilities worldwide.

The business value proposition of an IBDS Provider towards their partners is sustained by the provision of a *horizontal Industrial Big Data platform.* This platform integrates the required technologies for the capture, integration and visualization of relevant data from manufacturing facilities. The integration of these technologies supports the stages in the data lifecycle since (a) data are generated in a manufacturing production environment and available only to those components managing the production process (i.e. PLCs or SCADA systems) until (b) they are made available to ubiquitous data exploitation processes as "new data" [MW15] to be leveraged by different data-enabled services for different stakeholders using different analytical approaches. The *horizontality* of the platform is linked to the fact that it must facilitate data exploitation in diverse manufacturing sectors, depending on which market their partners operate their business in. Moreover, the global-scale activity of their partners, be it servitized

CEMs deploying their equipment and services to facilities worldwide or manufacturers owning those multiple, distributed facilities, leads to multiple deployments for an IBDS Provider and a high replication potential for their services.

The multiplicity of deployments for various servitized CEMs or manufacturers in their respective manufacturing sectors implies a global-scale set of targeted scenarios, as outlined in Figure 4.4. Therefore, an efficient management of costs and investments required for building and deploying the aforementioned platform is crucial in order to ensure a sustainable business model. In this sense, leveraging a cloud computing infrastructure in this architectural approach minimizes the use of dedicated resources and provides the flexibility to scale the storage and computing power necessary to process all the integrated data, while transferring the associated costs to customers via the adequate service fee. For that purpose, a *Cloud Services Provider* supplies the IBDS Provider with the required cloud computing resources. This represents, however, an important internal cost for an IBDS Provider in terms of data storage, transmission and processing resources to handle all data from all the connected facilities. Nevertheless, the need for cloud services must not involve being dependant of a specific provider. This implies that the IBDS Provider designs their technological solution in a *Platform-as-a-Service* model, so that all layers of the solution are built (and, therefore, owned) by the IBDS Provider on top of a generic cloud computing infrastructure. This provides the required flexibility to use the services of different cloud services providers, or many of them at the same time, depending on the deployment requirements imposed by specific projects, which facilitates the migration and avoiding the need of adapting the solution to specific platforms from different providers. Indeed, depending on the requirements of each deployment project and the country of origin of the data owner, data might be required to be physically stored in data centers located in specific geographical areas in order to comply with specific regulatory requirements about data sovereignty.

In such a context, the IBDS Provider can play a crucial role that is not limited to provide the platform that will sustain the data lifecycle along all its stages until the provision of smart services. Indeed, in increasingly more business scenarios where IT services are provided by a third party, customers expect innovations or the identification of customer-specific innovative solutions from their outsourcing service providers [Dha12]. Thus, an IBDS Provider can collaborate with their partners in the design process of the required smart services, involving themselves directly in the *smartization projects* conducted with engaged customers. This *strengthens their business value proposition*, as they would not only be a technological supplier but also a business partner collaborating in the design of the intended smart services. Indeed, IBDS Provider can provide important value for their partners with the capture of business requirements related to the design of those smart services and how the characterization of those requirements helps defining the right data capture and processing steps.

This approach has two major advantages. On one hand, the IBDS Provider combines their accumulated knowledge on both the business and the technical sides, which facilitates the connection between the business use scenarios and the technical, data exploitation-related requirements. On the other hand, they can leverage a collaboration management process that is replicable in each targeted

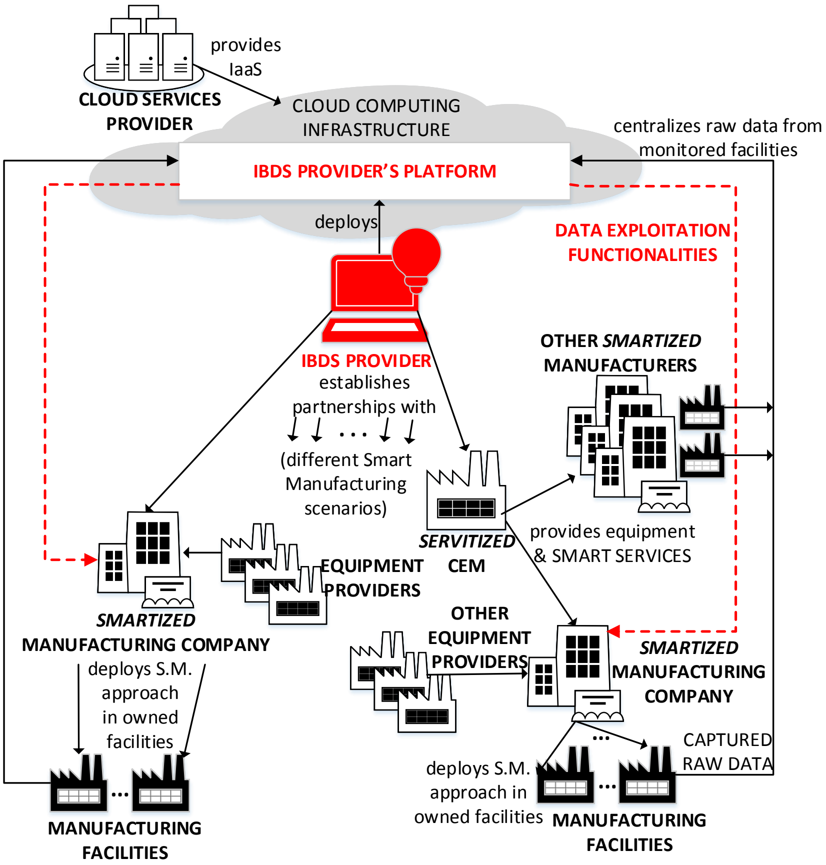

CLOUD SERVICES PROVIDER
provides IaaS
CLOUD COMPUTING INFRASTRUCTURE
IBDS PROVIDER'S PLATFORM
centralizes raw data from monitored facilities
deploys
DATA EXPLOITATION FUNCTIONALITIES
IBDS PROVIDER
establishes partnerships with
...
(different Smart Manufacturing scenarios)
OTHER SMARTIZED MANUFACTURERS
SERVITIZED CEM
provides equipment & SMART SERVICES
EQUIPMENT PROVIDERS
SMARTIZED MANUFACTURING COMPANY
deploys S.M. approach in owned facilities
...
MANUFACTURING FACILITIES
OTHER EQUIPMENT PROVIDERS
SMARTIZED MANUFACTURING COMPANY
CAPTURED RAW DATA
deploys S.M. approach in owned facilities
...
MANUFACTURING FACILITIES

Figure 4.4: Schema of the targeted scenarios for IBDS Providers

manufacturing sector where they aim at establishing a partnership to supply their services.

In order to undertake those smartization projects, the IBDS Provider can contribute with a project team combining diverse roles. The scope of this research work is focused on providing contributions for the following two roles:

- The *project manager* with the required combination of skills to collaborate with the top-level management staff of their manufacturing partners in the design of smart services based on successive smartization projects. They will also drive the capture of business requirements for the appropriate design of smart services.
- The *data engineer* in charge of managing the appropriate integration of technologies to support the data capture and integration platform sustaining the IBDS Provider's services. They must supervise that the technological platform complies with the IBDS Provider's goal of a scalable and sustainable global business, and that it fulfills the requirements to smoothly integrate with the technology already operating in manufacturing facilities in order to extract relevant data to be monitored and exploited.

According to the roles identified in the scenarios analyzed in our case study and the main roles proposed for data science teams [CPL16], we can differentiate the roles of *data engineer* and *data scientist*. Depending on the IBDS Provider's strategy and the specific targeted scenarios, the IBDS Provider can include in their project team a data scientist to develop analytic solutions for sector-specific use cases. The decision will depend on the strategic part this role can play for the IBDS Provider's business and a balance between a more controlled and a more complex management of the project team. Nevertheless, the role of data scientist and their duties developing analytical models is left out of the scope of this research work.

On a related matter, the design of the IBDS Provider's platform must integrate those built-in services that facilitate the collaborative development of these smartization projects. In this sense, it is of valuable help the availability of a *multi-purpose dashboard* that provides a real-time visualization of all raw data captured in a manufacturing facility newly engaged in a smartization project. Thus, the visualization and analysis of raw data in the early stages of these projects provides progressively increasing knowledge on the nature of each indicator involved, and enables the successive deployment of incremental improvements via preprocessing components that increase data quality.

The feedback cycle of this incremental approach is beneficial both for the IBDS Provider and their partner, but with different scopes. That is, each deployment derives a feedback cycle with a more general scope for the IBDS Provider (identifying those elements that can improve other deployments in any manufacturing sector using their solution) and, in parallel, there is another feedback cycle whose scope is focused on the specific sector of their partner and the smartization projects conducted in that sector.

4.2.2 Servitized CEMs

IBDS Providers can establish partnerships with servitized CEMs in order to supply their services to support a CEM's data-driven servitization strategy, leading to a Smart Manufacturing servitization scenario. The customer market for a servitized CEM in such an scenario is defined by those larger manufacturing companies operating the manufacturing process where the equipment supplied by that CEM is integrated to support or automate a particular step in that process. This CEM faces a scenario where (a) they compete with other CEMs providing the same type of equipment for that specific step of the manufacturing process, and (b) they share the same market in a non-competitive way with those CEMs providing the required equipment for the rest of the steps in the manufacturing process. This means that those customers where this CEM deploys their equipment may hire the provision of equipment for the rest of their process from different providers depending on the specific customer. Moreover, manufacturers operating that process may own several facilities distributed worldwide. This leads to a global-scale servitization scenario for the servitized CEM as outlined in Figure 4.5.

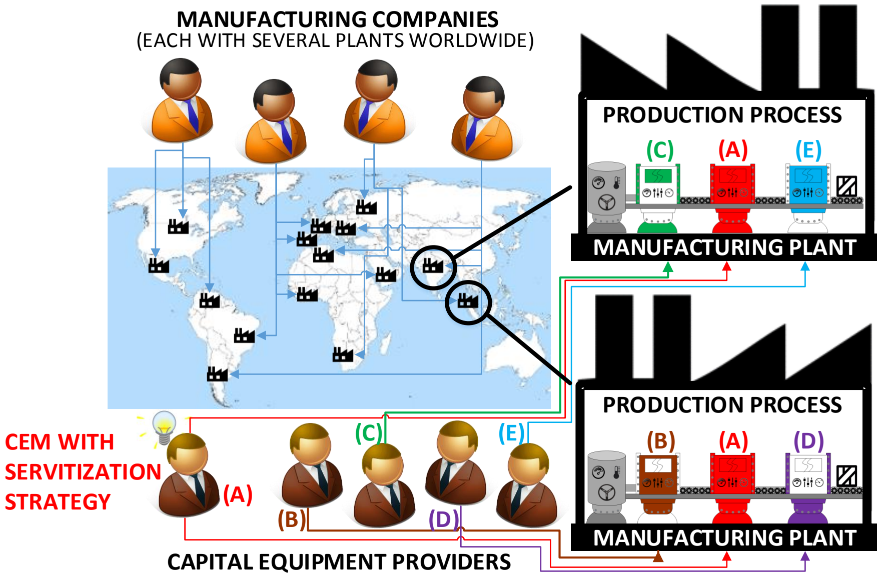


Figure 4.5: Schema of a servitization scenario for a particular servitized CEM

The servitized CEM wants to transform their business strategy towards a data-driven servitization approach, based on offering smart services to those of their customers aiming at evolving the operation of their businesses towards Smart Manufacturing. Moreover, given the global scale of these customers, potentially owning multiple plants worldwide, the scenarios where the servitized CEM aims at offering these data-enabled services are characterized by the need for gathering and processing massive, distributed data to analyze a manufacturing process under different settings. Thus, in this "product and processes focused" servitization approach [AAAS15] the CEM offers value-added services aiming at optimizing customer processes, which leads to increase efficiency and effectiveness of customer's operations. This increases the value of their customers' production systems thanks to an optimized performance.

The servitization scenario described above is focused on the creation of a new service by the CEM to an existing market, i.e. the customers operating in the manufacturing business sector where the CEM has traditionally marketed their products. Focusing on the existing market has the advantage that key components of the market (such as customers or competitors) are already familiar, which facilitates the access to relevant stakeholders and the communication with them.

The transformation of the CEM's business model via a data-enabled servitization can be sustained by establishing a strategic partnership with an IBDS Provider. This enables the combination of (a) the knowledge of the targeted manufacturing sector and the access to customers and relevant stakeholders provided by the CEM, and (b) the technological solutions and expertise in data capture and exploitation supplied by the IBDS Provider.

From the perspective of the servitized CEM, the use of the IBDS Provider's solutions and the supply of smart services based on this partnership must fulfill a main non-functional, business requirement: adopting the IBDS Provider's solution must allow the servitized CEM to incur an *incremental investment.* They must be able to progressively transfer the costs of that investment to those of their customers engaging in the use of the data-enabled value-added services (and, therefore, also obtaining progressive returns of their respective investment). In other words, the servitized CEM must not incur a considerable investment in a technological solution to support their servitization before obtaining some first returns from the market they target their services at. Indeed, this is one of the main challenges derived from transitioning to a servitization approach, as the expanded focus on service provision might increase costs without generating an immediate increase in returns [DUM$^+$15].

The way the servitized CEM integrates this need for an incremental approach in the provision of smart services is by designing *smartization projects* as the means to progressively integrate new customers. Thus, together with the provision of their equipment, the servitized CEM can deploy and refine smart services by engaging customers in the launching of pilot projects with a reduced scope and a limited amount of involved facilities and generated data. This facilitates the necessary incremental investment and enables the refinement and scaling of provided services (leveraging the lessons learned from previous deployments) as more facilities and new customers are engaged in its use.

A crucial requirement for the appropriate design of these smartization projects is that the smart services to be provided might not be limited to the process step supported by the equipment provided by the servitized CEM. Instead, they should be flexible enough to consider the exploitation and analysis of data from the whole manufacturing process operated by customers. This implies integrating different manufacturing process steps executed by equipment from different providers and supervised by specialized technicians for each step or subprocess, leading to a highly complex map of business stakeholders that implies a multi-view elicitation process. Therefore, the right design of the smart services should be supported by a detailed characterization and classification of the main stakeholders in engaged customers. This characterization maps onto the general schema of main business stakeholders for a manufacturing business context outlined in Figure 4.6. Thus, beginning with their direct interlocutor in the customer company, i.e. the owner who hires the value-added service, the proposed smart services must be capable of solving multi-view analytics needs (depending on the particular data-based insights required by each stakeholder in the customer company) based on different data exploitation approaches not fully characterized beforehand.

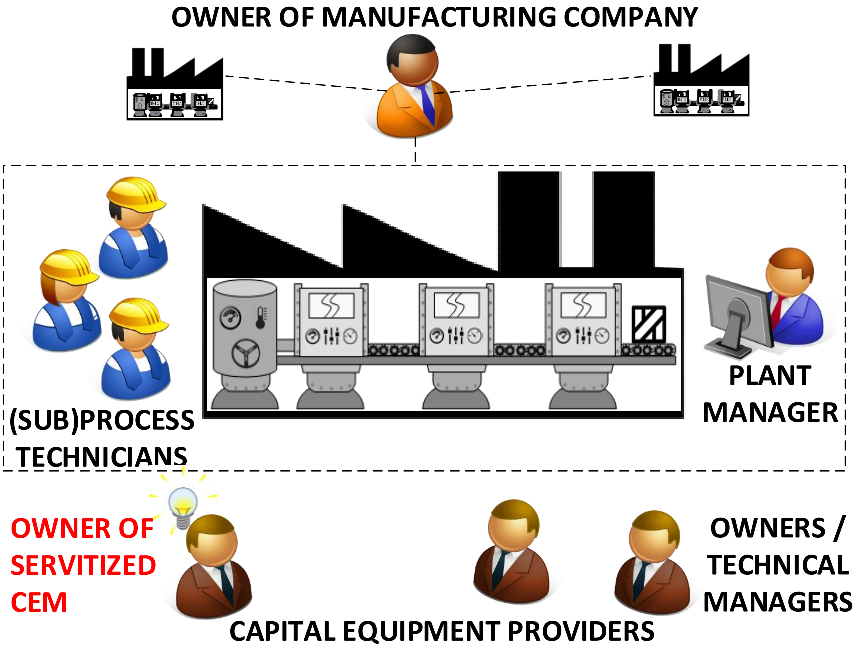

OWNER OF MANUFACTURING COMPANY
(SUB)PROCESS TECHNICIANS
PLANT MANAGER
OWNER OF SERVITIZED CEM
OWNERS / TECHNICAL MANAGERS
CAPITAL EQUIPMENT PROVIDERS

Figure 4.6: Schema of main stakeholders for the design of smart services in a servitization scenario

When smartization projects are launched in these servitization scenarios, i.e. as a result of the services provided by a partnership between an IBDS Provider and a servitized CEM, the IBDS Provider's project manager can leverage the stakeholder identification provided in Figure 4.6 for the initial project step of stakeholder analysis [Pro13]. Furthermore, in order to manage communications and organize the necessary interactions during the initial steps of smartization projects, the stakeholders presented in Figure 4.6 can be organized in five different levels of project influence, as listed in Table 4.2.

Stakeholder	Description
1. Owners (top-level representatives) of servitized CEM	The involvement of the top-level management staff from the servitized CEM is a crucial requirement for smartization projects in these servitization scenarios, given their direct access to customer companies, their knowledge of customer needs and their facilitation to access relevant stakeholders.
2. Owners of manufacturing companies	These are the customers targeted by the servitized CEM, to whom they have direct access that provides insights on the business strategies of these companies and their interest in Smart Manufacturing approaches.
3. Plant managers of manufacturing facilities	The customer companies may own different plants around the world, each of which is managed by a different person hired by the owner to be in charge of that plant. Each plant could have a different organizational schema and might implement their production process using different equipment.
4. Subprocess technicians of manufacturing facilities	Each plant manager hires different expert technicians to supervise specific phases or subprocesses along the manufacturing process.
5. Other capital equipment providers (owners / technical managers)	The rest of process steps in each plant are implemented using specialized equipment from other different providers. This equipment will provide relevant data to be captured in order to compile the *manufacturing record* of each product. For this purpose, the relevant interactions will be usually conducted with the *technical managers* in charge of their deployments for that particular manufacturer.

Table 4.2: Key stakeholders shown in Figure 4.6 ordered by their level of project influence

4.2.3 Smartized Manufacturers

The main objective of the manufacturing company interested in shifting towards a Smart Manufacturing approach is to extract value from the vast amount of data generated along the operated manufacturing process. Thus, these data could not only be used internally in the components and equipment along the process for its automated control, but also for optimizing process efficiency and product quality.

The data to be captured and exploited are generated along the *fabrication* process in a manufacturing facility (see Figure 4.7). Thus, the manufacturing record around which to deploy manufacturing intelligence is focused on these fabrication data. Other Smart Manufacturing scenarios also contemplate the integration of data captured during the *use* of the produced goods by the market consuming them. Nevertheless, this approach is considered out of the scope of this research work.

The manufacturing company may own different facilities, potentially dis-

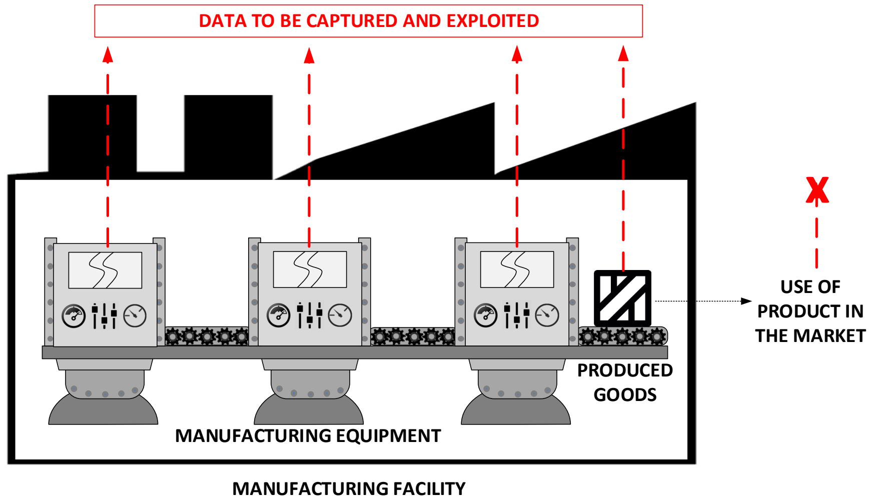

DATA TO BE CAPTURED AND EXPLOITED
X
USE OF PRODUCT IN THE MARKET
PRODUCED GOODS
MANUFACTURING EQUIPMENT
MANUFACTURING FACILITY

Figure 4.7: Schema of data to capture and exploit in a manufacturing facility

tributed worldwide, where they operate their fabrication process. For that purpose, they deploy different types of equipment that automate or assist in the execution of the different steps of the process. This equipment may be supplied by different providers, i.e. CEMs, specialized in some particular step of the fabrication process.

Apart from the servitization scenarios that involve a servitized CEM, smartization projects can be conducted as a result of a direct collaboration between an IBDS Provider and a smartized manufacturer. In these scenarios, as for relevant stakeholders and their level of project influence, the characterization is similar to the one presented in Figure 4.6. The only slight modification is that there would not be a servitized CEM as the stakeholder driving the project and, therefore, the IBDS Provider's project manager would focus the stakeholder analysis in those stakeholders from level 2 to 5 in Table 4.2.

Besides the functional requirements directly related to the goal of data analysis and exploitation activities, there are a number of non-functional requirements to be taken into account in order to design the appropriate smart services for these manufacturers. The most relevant ones are the following:

Assurance of a short-term value as an immediate return of their investment. Manufacturing company managers may tend to perceive a low return-to-effort ratio during the first phases of these data-driven projects [OLBO15]. Therefore, it is required to yield a progressive return of these manufacturers' investment when they engage in the use of smart services. The expected long-term savings depend on the potential success of the predictive models to be built. Therefore, it is necessary that the deployed services offer a basic and sustainable service in the short-term, while waiting for the potential added value obtained in the medium-long term from the predictive analytics. Thus, the manufacturer will perceive an adequate return-to-effort ratio, as the solution deployment will not require an excessive effort before starting obtaining a minimum value from the exploitation of their data. This facilitates the commitment by managers in order to develop later stages (i.e. further smart services) of such projects.

Avoid interference with the current manufacturing process operation. In order to facilitate customers' acceptance of smart services, the operating infrastructure should be kept intact as much as possible, leveraging current data export capabilities and not requiring additional IT projects. The deployment of the solution must demand a very limited effort from the customer side, at least not until some value is offered thanks to the capture, analysis and exploitation of their data.

Adequate contractual coverage of the use of data. As the owner of the data that are going to be captured and analyzed, the contractual agreement to use smart services must incorporate specific clauses dealing with a delimitation of the use that the provider is allowed to do with the data (whose property is retained by the smartized manufacturer).

Adequate contractual coverage of the location of data. The smartized manufacturer must be given assurance that their data, once transmitted outside the manufacturing facilities where they were captured, won't be stored in any other location that is not covered by the contract. An abstract concept like *the cloud*, although familiar and easily understandable by the IT community, does not convey the required clarity and precision to answer a recurring worry on "where are my data" by customers in any manufacturing context. Answering this question with a clear indication on where the data center is located (even the possibility to visit it) and the security measures deployed in it contributes dramatically to increase the customer's trust in the offered smart services.

Appropriate security mechanisms. The necessary security considerations must be taken into account when deploying new IT infrastructure in each manufacturing facility that can exchange data through a gateway to the Internet. In particular, the contract also must give guarantees on the security mechanisms controlling that no other infrastructure (apart from the one deployed to offer the data-enabled service) will have access to the data and the facility's infrastructure.

4.3 Conclusions

Our case study has allowed extracting important features of those targeted scenarios as the environment for our research. Those are the scenarios where, following the design science paradigm, we identify information systems-related problems and the requirements for purposeful solutions. The targeted scenarios are characterized by the interaction and collaboration of different main agents. Regarding our research, IBDS Providers represent the main agent at whom our contributions are targeted, so that they can leverage them when deploying their services in Smart Manufacturing scenarios and interacting with the other two relevant agents: servitized CEMs and smartized manufacturers. These agents pose relevant requirements to be fulfilled by the appropriate design of the technology and services to be supplied by IBDS Providers.

Apart from an Industrial Big Data Platform to deploy the required IT solutions in Smart Manufacturing scenarios, the market strategy of an IBDS Provider is based on the development of smartization projects for the different manufacturing customers demanding these solutions. The launching of smartization projects to deploy smart services in these manufacturing companies can be initiated through two different paths: either as a result of the services provided by a partnership between an IBDS Provider and a servitized CEM, i.e. one of the equipment providers in the targeted manufacturing sector, or as the result of a direct collaboration between the smartized manufacturer and an IBDS Provider. The case study has also allowed us to identify the main roles that IBDS Providers must consider when forming the team for one of these smartization projects. Indeed, the features of these projects, sustained by the collaboration among agents with different business strategies, demand a project team with different roles that must address a complex process. Our contributions aim at helping two specific roles, the data engineer and the project manager, to develop their respective duties in such projects (linked to the three research challenges specified in 2.3.2): on one hand, the collaborative design of smart services with partnering manufacturers and, on the other hand, the design, update and optimization of the data capturing and integration infrastructure supporting a cost-sustainable platform for an IBDS Provider's business.

The case study also granted access to a real-world business setting that represents a relevant and valuable instance of the targeted Smart Manufacturing scenarios, as it enabled the possibility to observe the solutions deployed in more than 60 locations worldwide, corresponding to Smart Manufacturing scenarios in different sectors. Moreover, it gave us the opportunity to conduct field validations in order to integrate and contrast our proposals for those IBDS Provider's roles that could leverage them when conducting smartization projects.

Chapter 5

A Procedural and Architectural Model for the Planning and Execution of Time-Series Data Reduction Analysis

From the characterization of the Smart Manufacturing scenarios for an IBDS Provider presented along the previous chapter, it is derived one of the main problems around which we pose a research goal. This problem is related to the considerable internal costs associated to the storage resources for the massive amount of data to be stored in such a context (i.e. 24x7 time-series data coming from the sensors in all the monitored facilities worldwide owned by each manufacturer leveraging the Industrial Big Data Platform supplied by the IBDS Provider), which hampers business sustainability and scalability for IBDS Providers.

In this sense, data reduction techniques represent a resource with potential to overcome that handicap. The field of data reduction has a solid conceptualization as part the data preprocessing step [GLH15] in Knowledge Discovery and Data Mining processes [KM06]. In the Smart Manufacturing scenarios discussed here, however, the application of data reduction techniques would not only be focused on preprocessing the input for data mining algorithms, but also on fulfilling non-functional requirements such as internal cost optimization in order to ensure business sustainability while enabling a wide range of approaches for later exploitation. In the specific case of time-series data, several categories for reduction and approximation techniques have been defined [Fu11] and compared using data from different contexts [WMD$^+$13]. Still, it has not been proposed a systematic approach that enables (a) the analysis of the combined potential of these techniques in industrial application scenarios liked the one mentioned above, (b) linking their reduction potential to the technical performance requirements derived from the business setting where they are intended to be deployed, and (c) optimizing the constrained time and resources that can be devoted to this analysis in these business-oriented scenarios.

This chapter presents one of the three main contributions of this dissertation: a *procedural and architectural model of the reduction analysis* to be carried out by a data engineer in charge of analyzing the reduction of the hundreds of time series that can be found in each of these application scenarios. The reduction analysis aims at obtaining the specification of the time-series data reduction solution that provides an optimized representation to efficiently store those raw time-series data in a Big Data Lake for their later exploitation in different processes. The main benefits of the procedural and architectural model designed for the planning and execution of this analysis are twofold: on one hand, it represents the process (including the architecture of the IT artifacts to automate most of its steps) that efficiently guides the analysis of the data engineer and prioritizes allocating resources to the analysis of those time-series data with higher expected impact in storage space savings; on the other hand, it combines the analysis of different families of data reduction techniques to provide a better fit for the heterogeneity found among the time series in the analyzed manufacturing scenarios.

The procedural and architectural models are the result of a *design process* preceded by a *field testing*, as outlined in section 3.3. The field testing was conducted in the business setting of our case study. The aim of this field testing was to contrast the applicability of the ground ideas supporting the proposed model in a real-world business setting, prior to its design as an artifact, and to validate that the desired reduction results could be obtained following that approach. Given the positive results obtained by this field testing, we conceptualized the validated approach as a design artifact (i.e. the proposed processes and architecture), so that this approach could be added as a new contribution to the existing knowledge and leveraged by practitioners in order to implement their reduction analysis solutions.

5.1 Motivation and Analysis of Related Work

The characterization of the targeted Smart Manufacturing scenarios and the role played by an IBDS Provider in them provides an understanding of how the capability of storing the more data the better from monitored facilities is related to the internal costs of the an IBDS Provider's platform. Moreover, it also has an impact on the service that can be supplied by an IBDS Provider and the exploitation capabilities on captured data. This section describes this interrelation as a motivation for a contribution based on the systematic use of data reduction techniques. Besides, related work on time-series data reduction is analyzed in order to identify relevant techniques and the analysis of their performance, as well as to reinforce the motivation for a systematic approach to apply these techniques in the targeted scenarios.

5.1.1 The Problem of Data Storage and the Need for Efficient Data Storage Strategies

One of the most important requirements from the perspective of an IBDS Provider is their need for a progressive, incremental investment in computing and storage resources. This is necessary in order to avoid a high volume of fixed costs due to *a priori* dedicated resources to store the massive amount of data coming from all the connected manufacturing plants. Hiring cloud-based computing and storage resources from a cloud services provider guarantees the fulfillment of this goal. Thus, as an IBDS Provider engages in new deployments of their solution, the costs corresponding to the storage resources required for the volume of data to be stored can be transferred via the adequate service fees. This introduces, however, the practical requirement of establishing limits with respect to the *time window* of data (i.e. how long historic data are kept before freeing storage space for new incoming data) as one of the service terms that an IBDS Provider agrees with a customer. This is an important parameter that greatly influences the competitiveness of an IBDS Provider, as the perceived value of their solution will be directly linked to the exploitation potential of the more data, the better.

Nevertheless, a more thorough understanding of the type of data to be captured and exploited in these Smart Manufacturing scenarios leads to identify untapped opportunities that an IBDS Provider can leverage to devise a more efficient data storage approach. Indeed, dealing with raw time-series data from industrial sensors operating in real-world factories introduces several inefficiencies for their later centralized storage, given that their original deployment was mainly for internal management purposes and not to support data export and exploitation processes like the ones described here. On one hand, these raw data come with noise (wrong measurements) to be filtered out and with missing values (errors in the measuring or transmission processes) to be filled in. On the other hand, in many cases industrial machine controllers are programmed in an inefficient way in terms of capturing data for analytical purposes. Sometimes they may be sending a constant value for several hours to indicate that the machine is turned off, but those data are captured and stored anyway, occupying space that increases data storage costs. The first problem (improving raw data quality via noise cleaning and missing value treatment) is left out of the scope of this research work. The second problem is, precisely, the key that motivates this contribution and where data reduction techniques can play a crucial role.

The use of data reduction techniques allows optimizing the storage space of the accumulated data. This widens the time window of data that can be accumulated in the Big Data Lake maintained by an IBDS Provider with the same storage resources and, therefore, costs. This would enable the exploitation of more (older) data instances. Besides, the adequate combination of lossless and approximate reduction techniques can provide more flexibility when defining the terms of service for customers. A first level of optimization can be achieved by using lossless techniques only. Thus, maintaining the same time window of data would have lower internal costs and this could be transferred to more reduced fees or higher margin, which would in any case lead to more competitiveness for the IBDS Provider. Besides, the use of approximate reduction techniques (i.e. incurring in some reconstruction error) could allow achieving an even higher cost reduction, which could be offered as an alternative to the customer (i.e. a *standard* fee for lossless storage and a *reduced* fee for approximate storage up to some error threshold).

Having said that, although numerous data reduction techniques are documented in existing literature [Fu11][GGB12][PVK$^+$04][WMD$^+$13], it is important to note that their efficient application in scenarios like the ones targeted in this research work is not straightforward. The intrinsic heterogeneity of the monitored indicators in each manufacturing process leads to time-series data of very different nature, susceptible to be reduced by various families of techniques, and with diverse reduction potential. The data engineer in charge of exploring the reduction potential of these indicators (time-series data) in diverse scenarios needs a more efficient approach than a case-by-case effort. It is necessary a systematic approach that provides the data engineer with guidelines about how to conduct this analysis, the type of time series that they can find, their estimated reduction potential and the most appropriate techniques to achieve that reduction. Such an approach would guarantee to optimize the constrained time and resources that can be devoted to this analysis in these business-oriented scenarios and to obtain the maximum benefit possible in terms of savings in storage resources. Moreover, it should be generic enough so that the data engineer could leverage it in different scenarios, given that the platform must facilitate the adoption of a Smart Manufacturing approach in diverse manufacturing sectors, with different types of time-series data and with different analytical use cases in mind for their later exploitation. The fulfillment of these goals motivates the procedural and architectural model for time-series data reduction analysis that is presented as a contribution of this research work.

5.1.2 Related Work on Time-Series Data Reduction

Different previous works address the application of reduction and approximation techniques to time-series data. In fact, the inefficiency of storing large volumes of raw time-series data has been explicitly stated as a strong motivation for this type of analyses [EEC$^+$09][PVK$^+$04]. In this subsection we review this background to draw potential synergies and identify gaps that reinforce the motivation to propose a solution aligned with the goals presented in 5.1.1. We focus this revision on the groups of reduction techniques commonly used in comparisons and evaluations, the different types of time series analyzed, and the details on frameworks or methods to conduct these analyses in industrial application scenarios and to deploy their results.

In [Fu11] it is provided a very thorough classification of different techniques for the reduced representation of time-series data, grouping them in families and identifying the most representative techniques in each family. Reference [WMD$^+$13] also provides a hierarchy of time series representation methods, which includes the main techniques already compiled in [Fu11] with the exception of the technique known as *Perceptually Important Points* (PIP) [CFLN02]. Indeed, the selection of reduction and approximation techniques that are analyzed and compared is similar across various references discussing time-series data mining [GGB12][PVK$^+$04][WMD$^+$13]. This provides a solid foundation to identify the main reduction techniques to consider in our analysis.

Nevertheless, despite the recurrent use of reduction techniques from different families (according to the reviewed classifications [Fu11][WMD$^+$13]) in all these references, there is a lack of a more holistic view of the various types of time series that are present in the same application scenario. Such is the case in the manufacturing setting analyzed in our work, given the heterogeneity in the syntactic features of the hundreds of captured time series. Indeed, one important foundation for our contributions in this work is that they have been drawn from the heterogeneity in the actual time-series data (and, therefore, in the required reduction techniques) that are being generated in manufacturing plants.

This heterogeneity implies a need for considering techniques beyond those usually analyzed families, such as *lossless data compression algorithms*, that may be appropriate for specific types of time series (e.g. those generated by *binary* indicators, frequently found in these application scenarios) and for some of the requirements to guarantee for their later exploitation. The only found reference that also integrates these data compression algorithms in the analysis they present is [BFL13], where *Run-Length Encoding* (RLE) [RC67] is assessed at the same time as PIP and piecewise representations [Keo97].

Regarding methodological approaches, reference [uRCBW16] proposes a "big data reduction framework" for early data reduction at the customer and enterprise ends, i.e. data preprocessing before centralizing data in cloud computing infrastructures. However, that early data reduction is actually focused on analyzing raw data and solving analytical use cases by creating "knowledge patterns" to be exploited locally. Therefore, while this early data reduction indeed contributes in decreasing the cost of cloud-based resources for the subsequent centralized storage, this reduction approach does not guarantee the required genericity in the reduced data to be later exploited by different processes with different analytical approaches. Furthermore, it does not cover specific types of raw data such as time series (which is the predominant raw data in manufacturing application scenarios) or techniques to identify the best reduction approach for the data to be processed.

No reference has been found that provides details towards a method that can assist the task of a data engineer when analyzing which reduction techniques are the most suitable ones for which of the data to process in the application scenario. Indeed, such a method would facilitate an efficient use of the time and resources that can be devoted to that task, given the practical constraints found in business scenarios. This strongly reinforces the motivation to contribute with design artifacts that facilitate the solution of the described data reduction problem.

5.2 Hypothesis Validation prior to Creating Design Artifacts

Once we confirmed the relevance of the problem and the possibility to contribute with innovative approaches that can be added to the existing knowledge on the area, we focused our research work on contributing with a design artifact built on *two key ideas*: (a) allowing the application of the most suitable reduction techniques to the different types of time-series data found in one of these manufacturing scenarios, and (b) prioritizing the reduction analysis on those time series with higher reduction potential, in order to optimize the time and resources to allocate for such an analysis.

Nevertheless, as a prerequisite to design such a contribution, we established the need to validate beforehand the applicability and effectiveness of an approach based on those key ideas, once applied to a real-world manufacturing business setting. Obtaining that validation would provide a more solid grounding and motivation for our contribution. Therefore, we posed the following two hypotheses that needed to be validated in order to confirm the suitability of the conceived approach for the design artifact:

- *Hypothesis 1: Obtaining of substantial reductions.* The systematic application of different reduction techniques to the time-series data captured from the sensors in a manufacturing plant can lead to substantial savings in storage costs to the IBDS Provider, while preserving the possibility of reconstructing them when needed for later exploitation.
- *Hypothesis 2: Heterogeneity in obtained reduction per technique and time series.* If we identify which technique gives the best reduction performance for each of the time series captured in a given scenario and the obtained reduction in each case, we find notable differences in which are the best techniques depending on the time series and in which reduction is obtained for each time series in the same scenario.

In order to verify these hypotheses, we established a *field testing* in the manufacturing business setting where we conducted our case study. The work method conceived for this field testing consisted of three steps, each based on the structure of a design cycle (build-evaluate) and each focused on one of the three main areas covered in this analysis: *time-series data*, *reduction techniques* and *reduction performance criteria*. For each of these areas, it was conducted (a) the conceptualization and extraction of elements to leverage from relevant references, (b) the *building* of testing-oriented IT artifacts that implement the approach to validate, and (c) the use of this artifacts in the analyzed manufacturing business setting in order to *evaluate* their applicability and effectiveness. Thus, the three-step method to conduct the field testing was organized as follows:

1. *Time-series data.* It consisted of (a) the conceptualization of the time-series data capture in a manufacturing process, (b) the implementation of testing-oriented IT artifacts supporting the classification of time-series

data according to an initial set of families, and (c) the application of those artifacts in order to extract a sample of the time-series data generated for all indicators in the analyzed manufacturing setting and define a preliminary classification of the extracted time-series according to the defined families.

2. *Reduction techniques.* It consisted of (a) the conceptualization of the application of reduction techniques to time-series data and the extraction of relevant techniques based on the related work and on the heterogeneity in the extracted time-series families, (b) the implementation of testing-oriented IT artifacts supporting the application of the selected reduction techniques, the storage of the reduced and reconstructed versions of each time series and the assignment of recommended techniques to time-series families, and (c) the application of those artifacts to assign the recommended techniques to the identified time-series families in the analyzed manufacturing setting.
3. *Reduction performance criteria.* It consisted of (a) the conceptualization of the reduction performance criteria to assess, (b) the implementation of testing-oriented IT artifacts supporting the assessment of the identified performance criteria during the application of reduction techniques and the visualization of the obtained performance, and (c) the application of those artifacts in order to assess the reduction of the time-series families using the recommended techniques.

The rest of this section details the execution and results of these three steps, the combination of their results to design the field testing in the business setting of our case study and the final conclusions, based on the outcome of the field testing, in order to validate/refute the formulated hypotheses.

5.2.1 Time-Series Data Captured from a Manufacturing Process

The objects of our analysis are the *time-series data* representing the relevant *indicators* to measure along the *manufacturing process* (structured as a sequence of *steps*) of a particular manufactured *product.* The main concepts in this analysis are specified in Table 5.1 and their relationship is outlined in Figure 5.1.

The field testing was conducted in one of the manufacturing plants of the real-world business setting for our case study. In the analyzed plant (property of one of the customers of the analyzed CEM) there were a total of 442 indicators connected to the data capture system for their monitoring. These indicators registered time-series data with a continuous measurement (one measurement per second) of a variety of equipment setting parameters and physical magnitudes (temperatures, lengths, weights, capacities, etc.) related to the produced goods and environmental conditions. From those 442 indicators, it was taken a sample consisting of the time-series data generated during a complete week of operation of the analyzed plant. It was observed that 128 indicators were returning the same measurement during the whole time interval. Therefore, as their lossless reduction was straightforward, they were not included in the field testing, resulting in a set of 314 indicators to analyze.

Concept	**Description**
Product unit (P_u)	Each of the instances of manufactured product, after completing all the steps in the manufacturing process.
Sensor	The various machines and equipment executing the steps completing the manufacturing process are fitted with sensors that continuously register the values for a set of variables (physical magnitudes related to the product and its environment, setting parameters of the equipment, etc.) that characterize the state of the step where they are located along the manufacturing process.
Indicator (I_i *for i=1..IND*)	Each of the variables that is measured by sensors along the analyzed manufacturing process. IND denotes the total number of indicators.
Measurement (v)	A value for an indicator registered by a sensor at a specific time.
Timestamp (t)	The time corresponding to a measurement.
Time series (TS_i)	As each sensor is continuously registering values (i.e. measurements) for a set of variables (i.e. indicators), this log of measurements can be viewed as time-series data.
Time series collection ($\{TS_i\}$)	This concept describes a set of time series registering values for the same indicator, or for indicators measuring the same phenomenon or sharing the same syntactic characteristics.
Time series segment (subseries)	A subset of continuous measurements extracted from the time series registered for an indicator.

Table 5.1: Main concepts related to the capture of time-series data in a manufacturing process

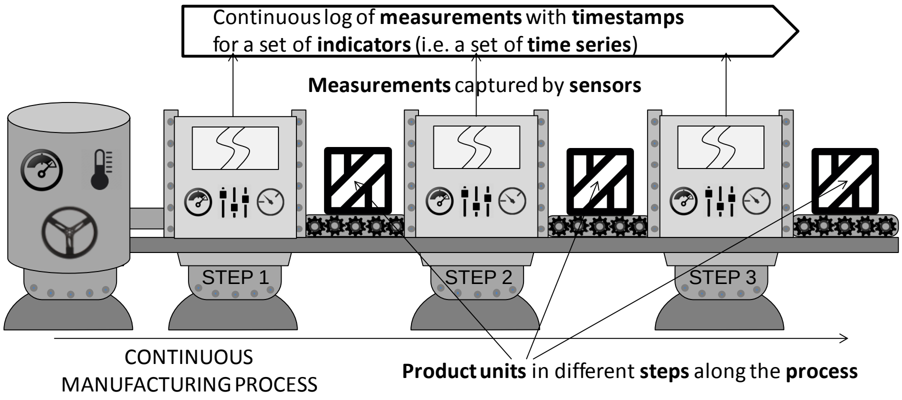

Continuous log of **measurements** with **timestamps** for a set of **indicators** (i.e. a set of **time series**)
Measurements captured by **sensors**
STEP 1
STEP 2
STEP 3
CONTINUOUS MANUFACTURING PROCESS
Product units in different **steps** along the **process**

Figure 5.1: Conceptual schema of time-series data capture in a manufacturing process

After examining the indicators and the type of time-series data that were captured, we could identify two main syntactic families among time series, depending on whether they were representing *continuous* or *discrete* data. Furthermore, among those indicators represented by continuous time-series data, two main subgroups were identified (see Table 5.2), depending on their temporal relationship with the progress of product units along the manufacturing process:

- *Product-driven indicators.* These continuous indicators register a magnitude measured directly on each product unit. As different units progress through the step of the manufacturing process where that indicator is measured, the generated time series repeats a similar pattern of measurements for each product unit. Therefore, this type of time series will be segmented according to the repeated pattern and these segments (subseries) will be grouped and analyzed as a time series collection.
- *Product-undriven indicators.* The temporal progress of the measurements in these indicators is not directly related to the repeated advance of successive product units through the process step where the indicator is being registered. Therefore, they do not reflect any repeated pattern and are analyzed as a whole.

5.2.2 Selection of Reduction Techniques

With respect to the application of reduction techniques to time-series data, the conceptualization led to the identification of the three main concepts specified next.

*Generic reduction technique (g*RED*)*
Each reduction algorithm considered for the analysis. Each generic reduction

Group	Num. of indicators	% of disk space
Discrete binary	146	44.35%
Discrete n-ary, n>2	85	26.69%
Continuous, product-undriven	31	11.46%
Continuous, product-driven	52	17.50%

Table 5.2: Initial classification of indicators

technique (gRED) has a formal, numeric parameter p that adjusts a bigger or smaller dimensionality for the reduced representation to be obtained by the application of that technique. Thus, for the technique to be applicable to a time series TS_i in order to obtain a specific reduced representation of TS_i, a specific value (an actual parameter z) must be assigned to the formal parameter p.

*Reduced time series (TS*RED*)*
The reduced representation for a time series obtained by the application of a reduction technique. The specific format of *TS*RED will depend on the technique.

*Reconstructed time series (TS*REC*)*
Each reduction technique has an associated *reconstruction* function. While the reduction technique transforms a time series *TS* into its reduced representation *TS*RED, the reconstruction function transforms *TS*RED into a *reconstructed representation TS*REC with the same format as the original time series.

Based on this conceptualization, we built a set of testing-oriented IT Artifacts that covered the implementation of reduction techniques and their application to time-series data in order to generate their reduced and their subsequent reconstructed version. In that regard, the first step was to identify the different alternatives of gRED to be parameterized and analyzed in the field testing. The heterogeneity in the time series found in the analyzed application scenarios implied the need for a broader set of reduction techniques than those that are usually analyzed at the same time in the experimental settings described in existing references. In particular, it implied the need for combining techniques for both *continuous* and *discrete* time-series data.

With respect to *continuous* time-series data, the relevant technique families described in [Fu11] and the specific techniques used in [WMD$^+$13] can be leveraged for the initial selection of reduction techniques. Based on these references, we selected the techniques listed in Table 5.3, which also includes the meaning of the formal parameter that adjusts a bigger or smaller dimensionality for the reduced representation obtained by the application of each technique.

Regarding *discrete* time-series data, there is an important feature that influences the type of reduction algorithms to apply. These data allow identifying

Generic reduction technique (*gRED*)	**Formal parameter (*p*) of each technique**
Sampling (SAM) [Aas69] Piecewise Aggregate Approximation (PAA) [KCPM01] Adaptive Piecewise Constant Approximation (APCA) [CKMP02] Perceptually Important Points (PIP) [CFLN02]	n = Num. of selected points for the reduced representation
Piecewise Linear Representation (PLR) [Keo97]	s= Num. of segments to be approximated by linear regression
Polynomial Regression (PRE) [Sti74]	d = Degree of the polynomial
Chebyshev Polynomials (CHEB) [CKMP02]	c = Num. of Chebyshev coefficients considered
Discrete Wavelet Transformation (DWT) [CF99] using the Haar filter [SS99]	l = Resolution level of the Haar transform

Table 5.3: Selected reduction techniques for continuous data

Generic reduction technique (*gRED*)	**Formal parameter (*p*) of each technique**
Run-Length Encoding (RLE) [RC67] LempelZivWelch (LZW) [Wel84]	No parameter required

Table 5.4: Selected reduction techniques for discrete data

different operation modes of the production equipment, which are necessary to delimit the steps in the process and to guide the identification of which data from continuous indicators correspond to which step. Therefore, in order not to hamper the right assignment of data segments to process steps (which would result into incorrect data views), the application of *lossless* reduction algorithms is required in these cases. Given this requirement, the algorithms to be used to analyze their reduction were selected accordingly. They are listed in Table 5.4.

The *reduction techniques* in Table 5.3 were marked as the recommended ones to be analyzed with continuous data and those in Table 5.4 were marked as the recommended ones for discrete data. This led to extending the initial classification of indicators in Table 5.2 with the selected techniques to be analyzed with each group, as reflected in Table 5.5.

5.2.3 Reduction Performance Requirements

Technical performance requirements constitute another core element of the analysis of data reduction in these scenarios. The two main concepts at this level

Group	**Num. of indicators**	**% of disk space**	**Selected reduction techniques**
Discrete binary	146	44.35%	LZW, RLE (see Table 5.4)
Discrete n-ary, n>2	85	26.69%	
Continuous, product-undriven	31	11.46%	APCA, CHEB, DWT, PAA, PIP, PLR, PRE, SAM (see Table 5.3)
Continuous, product-driven	52	17.50%	

Table 5.5: Initial assignment of selected techniques

are specified next.

Performance dimension for a reduction technique (PER_f)
The application of a reduction technique (and its associated reconstruction function) to a time series TS_i, while fulfilling the goal of producing a reduced representation of TS_i, has an associated performance that is assessed according to different dimensions. This allows comparing the performance of different techniques in order to select the technique that best fulfills the *performance requirements* established in the application scenario.

Performance requirement (R_q)
It compares the value for a performance dimension PER_f with a threshold T using the comparison operator *OP*. For instance, a performance requirement like *(Compression Ratio in Disk < 25%)* implies that the application of the reduction technique must achieve a reduced representation for the time series that occupies less than 25% of the disk space occupied by its original representation.

Based on this conceptualization, we built a set of testing-oriented IT Artifacts focused on the evaluation of performance requirements when applying reduction techniques to the analyzed time-series data. In the conducted field testing two main performance dimensions guided the assessment of reduction performance:

- *Ratio on Error (RTERR).* It allows defining a performance requirement determining a threshold for a *maximum assumable RTERR* when reconstructing the reduced time series with respect to the original one.
- *Compression Ratio in Disk (COMPD).* It is expressed as the ratio between the disk space occupied by a reduced representation and the space occupied by the original representation. It allows defining a performance requirement determining a threshold for a *compression ratio not to be exceeded* by the reduced time series with respect to the original one.

In order to identify the best parameterization for the analyzed techniques in the case of continuous time series, a threshold for a maximum assumable *RTERR* in the reconstructed time series was set to a root mean squared error equal to 1% of the average measurement for each indicator. For each analyzed technique by the *reduction analyzer*, among those parameterizations fulfilling the requirement on *RTERR*, the one providing the best *COMPD* was selected.

Apart from these two main performance dimensions, another two dimensions were also registered during the field testing: *Reduction Computing Time (REDCT)* and *Reconstruction Computing Time (RECCT)*. Although they were not directly used in order to identify the best parameterization in the field testing, these performance dimensions might have an impact on the selection of recommended reduction techniques and on the decision on how these techniques should be deployed in application scenarios.

5.2.4 Results of the Field Testing

Accounting for all the time series segments analyzed and the parameters applied to generic reduction techniques, an approximate total of 470,000 applications of different versions of fRED were executed during the field testing and assessed with respect to *RTERR* and *COMPD*. The reduction analysis provided substantial results in terms of savings of storage resources. Furthermore, it also produced important insights that validated the hypotheses posed prior to the field testing. The main results of the conducted tests (detailed further in Appendix A) are summarized next.

5.2.4.1 Discrete Binary (DB) Data

This group of 146 indicators is mainly comprised of parameters with two possible states for the operation mode in different equipment along the process, i.e. whether some operation mode is active or not, whether a crane is moving upwards or backwards, whether a conveyor is moving forwards or backwards, etc.

For this group of indicators the two lossless algorithms listed in Table 5.4 (RLE and LZW) were analyzed. The best *COMPD* was always obtained by RLE by a consistent margin: in average, the reduced representation obtained by RLE for a given time series occupied a 12.05% (±1.55% for α=0.05) of the disk space occupied by the reduced representation obtained by LZW. The average COMPD obtained by the best reduced representation (i.e. RLE) was **0.0485%** (±0.01% for α=0.05).

5.2.4.2 Discrete N-ary (DN) Data

Among the indicators grouped in this category, given the different results obtained in terms of *COMPD* and the performance of the two assessed techniques, two subgroups were identified.

Subgroup DN-1 (25 indicators). The indicators in this subgroup are similar to the binary category, but with more than two possible states: multiple operation modes, positions for a component, number of spaces occupied in a warehouse, etc. In this group RLE outperformed LZW in terms of COMPD in all cases, even with a higher margin: in average, the reduced representation obtained by RLE for a given time series occupied a 3.32% (±1.59% for α=0.05) of the disk space occupied by the reduced representation obtained by LZW. The average COMPD obtained by RLE was better than with binary data: **0.0127%** (±0.006% for α=0.05).

Subgroup DN-2 (60 indicators). This subgroup comprises indicators registering the operational speed of the different conveyors along the process. These indicators have a prefixed set of possible values for this speed, measured in revolutions per minute (rpm). In this group the outperformance of RLE is not as clear as in the previous cases. RLE provides the best *COMPD* in 52 of the 60 indicators. In those cases, in average, the reduced representation obtained by RLE for a given time series occupied a 50.68% (±6.61% for α=0.05) of the representation obtained by LZW. There are 8 cases, however, where LZW outperforms RLE. In those 8 cases, in average, the reduced representation obtained by LZW for a given time series occupied an 82.86% (±9.02% for α=0.05) of the disk space occupied by that obtained by RLE. Taking into account the best *COMPD* for each indicator, regardless of which one of the two techniques provided it, the average obtained was **0.2488%** (±0.04% for α=0.05).

5.2.4.3 Continuous, Product-Undriven (CPU) Data

This group of 31 indicators registers the content level in the raw materials tanks, i.e. those tanks containing the different materials that are supplied to the chemical transformation step in the production of polyurethane foam. The values of these time series are highly stable for long stretches, with punctual changes over time. Therefore, the time series registered for these indicators are composed of two types of subseries: (a) long subseries with *near-zero difference between successive measurements*, i.e. where the series composed of the measurements' differences has near-zero average and deviation, and (b) short subseries with *sharp monotonic variation*, some decreasing (when tanks are emptied) and some others increasing (when tanks are filled in). Considering the best *COMPD* obtained by each technique with a parameterization that guarantees an error rate not exceeding the required threshold, PIP obtains the best *COMPD* in 85% of the cases. Taking into account the best *COMPD* for each indicator, regardless of the technique providing it, the average obtained was **0.002987%** (±0.0013% for α=0.05).

5.2.4.4 Continuous, Product-Driven (CPD) Data

In this category four subgroups were identified, mainly based on the specific magnitude they register for each product unit as it evolves in time (or as each product unit is conveyed through some measuring frame). Each time series is segmented into subseries, corresponding to the measurements for each foam block.

Subgroup CPD-1 (32 indicators). All these indicators register the temperature evolution of the exothermic chemical reaction taking place among the mixed raw materials while foam blocks are being formed. In all the analyzed instances it is observed that, parting from a maximum value obtained in the first segment of measurements, they show a *monotonically decreasing series* of approximately 250,000 measurements. Furthermore, although the value is changing constantly, the decrement between consecutive measurements is not sharp in any point, i.e. the series composed of the difference between successive measurements have a low average and low deviation. Therefore, these series show a *clear and stable trend over time.* Considering the best *COMPD* obtained by each technique with a parameterization that guarantees an error rate not exceeding the required threshold, the average best *COMPD* is **0.014%** (±0.002% for α=0.05).

Subgroup CPD-2 (9 indicators). This subgroup comprises the different indicators along the process that register the height of each block. This height is registered when a conveyor belt transports each produced block through a measuring frame. These time series reflect the irregularities in this magnitude, given that all blocks present an irregular surface that has to be trimmed out, thus providing an important inefficiency indicator. Therefore, they show *no monotonic trend*, combining increasing and decreasing patterns. Moreover, those *increments and decrements show a high variability*, both along a particular time series instance and among instances, and it can be observed a combination of *high and low differences between maximum and minimum values.* Combining PAA and PIP it is obtained the best *COMPD* in 79% of the cases. The average best *COMPD* is **36.5327%** (±1.7125% for α=0.05).

Subgroup CPD-3 (3 indicators). In this subgroup the time series register the width of each produced block in different steps along the process. Although these time series show similar syntactic characteristics to those in the previous subgroup (CPD-2), we analyzed these two subgroups separately, in order to verify whether both subgroups offered similar compression results regardless of the different measured magnitude. Considering the best *COMPD* obtained by each technique with a parameterization that guarantees an error rate not exceeding the required threshold, PIP obtains the best *COMPD* in 94% of the cases. Taking into account the best *COMPD* for each indicator, regardless of the technique providing it, the average obtained was **41.0993%** (±2.477% for α=0.05).

Subgroup CPD-4 (8 indicators). In this case the weight of different blocks is registered. While the time series show no monotonic trend, the magnitude of the increments and decrements is considerably reduced with respect to the cases in CPD-2 and CPD-3, and the difference between maximum and minimum values is shorter. For the analyzed time series in this subgroup, PIP obtains the best *COMPD* in all cases. The average *COMPD* is **1.0247%** (±0.545% for α=0.05).

Finally, we refined the initial classification of indicators and assignment of selected techniques in Table 5.5, according to the obtained subgroups and the performance of the analyzed techniques. Table 5.6 summarizes the average *COMPD* (from best to worse) obtained in all subgroups of indicators, as a refinement of the original groups. In this final summary, the number of indicators and original disk space are now divided according to established subgroups. Besides, the selection of reduction techniques is refined for each subgroup, excluding those that

Refined group-subgroup	**Num. of indicators**	**Original disk space (% of total)**	**Average COMPD (reduced/ original disk space)**	**Refined selection of reduction techniques**
(CPU) Continuous, product-undriven	31	11.46%	0.002987%	PIP, SAM, CHEB, PRE, PLR, PAA
(DN-1) Discrete n-ary - subgroup 1 -	25	7.85%	0.0127%	RLE
(CPD-1) Continuous, product-driven - subgroup 1 -	32	10.77%	0.014%	PIP, PRE, CHEB
(DB) Discrete binary	146	44.35%	0.0485%	RLE
(DN-2) Discrete n-ary - subgroup 2 -	60	18.84%	0.2488%	RLE, LZW
(CPD-4) Continuous, product-driven - subgroup 4 -	8	2.69%	1.02%	PIP, PRE, CHEB, SAM, PLR
(CPD-2) Continuous, product-driven - subgroup 2 -	9	3.03%	36.53%	PAA, PIP, APCA, CHEB, PRE
(CPD-3) Continuous, product-driven - subgroup 3 -	3	1.01%	41.10%	PIP, APCA, CHEB, SAM, PAA, PRE

Table 5.6: Final summary of the reduction analysis

obtained in average a *COMPD* that at least doubles the best one.

5.2.5 Conclusions Derived from the Field Testing

The obtained results allowed us to validate the two hypotheses posed prior to the field testing. Regarding *hypothesis 1*, based on the results summarized in Table 5.6, we calculated the weighted average of the obtained COMPD for each subgroup of indicators, according to their occupied disk space with respect to the total. The result was an overall compression ratio of 1.62%. This implied that *98.38% of the storage used* (and the corresponding costs) *could be saved* ensuring a lossless compression of 71% of all data, while ensuring that for the remaining 29% of data the reconstruction error did not exceed a 1% of the average measurement for each indicator.

With respect to *hypothesis 2*, its validation is sustained by the heterogeneity in the results. On one hand, various subgroups could be identified among the captured time-series data, based on the patterns in the performance obtained by the analyzed reduction techniques. The identified subgroups constituted new specialized time series families (subgroups CPD-2 and CPD-3 could be merged into a single family, given their similar syntactic characterization and reduction analysis results), and each family was assigned different recommended reduction techniques. This recommendation could be leveraged for future analyses of similar time-series data. Thus, the characterization of each family's syntactic features (i.e. the presence of monotonic trends, the variability in increments and decrements, the difference in the range of measurements along instances of the same family, etc.) could be used to match new indicators found in application scenarios and leverage the information associated with that family and the refined selection of appropriate reduction techniques could be used as the recommendation for the techniques to analyze. The initial, more general time series families with an unrefined selection of techniques would remain as a guide for the analysis of new indicators not matching the newly added subgroups.

On the other hand, there were noteworthy differences in the reduction obtained by the analyzed techniques among different families of time series. This would serve as a basis for the concept of reduction potential ranking as a recommendation for the data engineer. The average *COMPD* obtained for each family could be used as their potential reduction ranking, in order to prioritize their analysis. Thus, the data engineer could establish prioritizations in order to invest the allocated analysis time and resources in the time series families with more potential.

5.3 Modeling the Planning and Execution of a Reduction Analysis

By validating the hypotheses that grounded our approach, we confirmed that time-series data reduction techniques are indeed a valuable resource for these application scenarios, given their effectiveness in providing significant savings in data storage for the type of data that are captured. However, the heterogeneity in terms of recommended techniques and the potentially obtainable reduction for different time-series families must be properly addressed in real-world business scenarios with practical constraints. Indeed, our field testing scenario, where we can meticulously test many different reduction techniques for all the time-series data available, does not have to deal with the same practical constraints as the data engineer of an IBDS Provider demanding solutions for their data reduction analysis. In such a business scenario, the optimization of the syntactic representation of data used for their storage is an analysis work that competes for time and resources with other important duties of the data engineer in an IBDS Provider. Therefore, a data engineer would clearly benefit from a planning for this analysis work that allows them to prioritize those data with higher reduction potential and to delimit the range of techniques and parameterizations to explore with them. Thus, the data engineer could execute the analysis following that prioritized sequence of *analysis jobs* (each focused on a specific collection of time series with similar syntactic features and reduction potential), in order to guarantee the maximum reduction potential is obtained for the time and resources allocated for the reduction analysis.

The previous reasoning provided us with the focus for the design artifacts to be produced after the field testing. Thus, we modeled the time-series data reduction analysis as two processes to be conducted one after the other: the *planning* of the reduction analysis and its subsequent *execution*. For each of these processes, the work of the data engineer is supported by an IT artifact: the *reduction analysis planner* (for the first process) and the *reduction analysis executor* (for the second one). Our contribution is to provide the *modeling for these two processes and for the architecture of these two artifacts*. The proposed process models utilize the constructs and graphical representation provided by Object Management Group's standard *Business Process Model and Notation* (BPMN 2.0) [Obj11]. This procedural and architectural model formally encapsulates the approach validated in the field testing and the specification and purpose of the testing-oriented IT artifacts built for that evaluation. Thus, this model could be instantiated in application scenarios with similar requirements demanding data reduction analysis solutions.

5.3.1 A Global View of Reduction Analysis in the Context of an Application Scenario

The *reduction analysis* (composed of two main processes, *planning* and *execution*) is carried out by a data engineer in charge of analyzing the optimal storage representation for the time-series data generated in a particular Smart Manufacturing scenario, i.e the *application scenario*. The strategy to conduct this analysis must guarantee that (a) the obtained reduced representation does not hamper the use of those data for the adequate resolution of the use cases in the application scenario, (b) the syntactic specificities of each time series are taken into account in the analysis process, and (c) the data engineer prioritizes the analysis of those indicators with bigger reduction potential (i.e. bigger impact in storage cost savings) given the practical constraints on the time and resources to allocate for this analysis.

The context in which the reduction analysis is conducted is outlined in Figure 5.2. The input to the reduction analysis consists of two elements:

- The *raw time-series data* from the relevant indicators captured in the application scenario. In order to obtain this input, two important milestones must have been completed beforehand: (a) the relevant indicators whose data must be captured must have been identified through the interaction with the business stakeholders in the application scenario and the characterization of the data-enabled processes that solve the business use cases; (b) the required infrastructure for the capture of the raw time-series data for those indicators must have been deployed (where appropriate, noise elimination and missing values treatment techniques will have been applied beforehand in order to ensure data quality).
- The different *technical performance requirements* that the assessed reduction techniques must meet in order to be deployed in the application scenario. The characterization of these requirements has been previously de-

rived from the specification of the use cases to be solved in this application scenario.

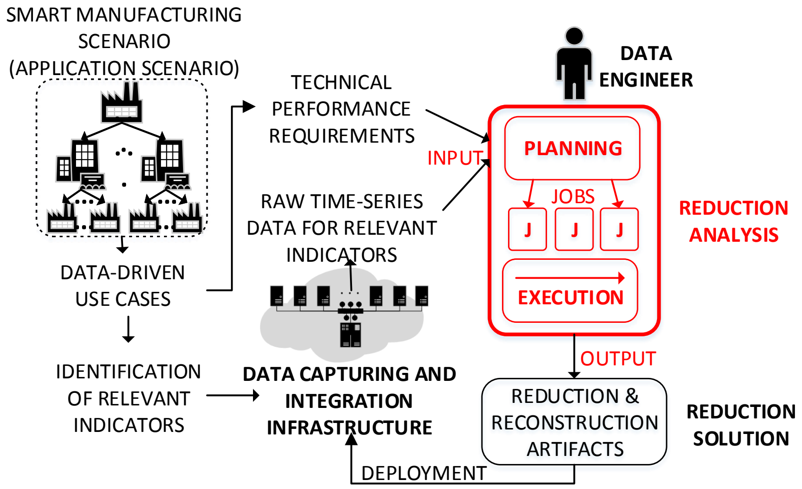

SMART MANUFACTURING SCENARIO (APPLICATION SCENARIO)
DATA-DRIVEN USE CASES
IDENTIFICATION OF RELEVANT INDICATORS
TECHNICAL PERFORMANCE REQUIREMENTS
RAW TIME-SERIES DATA FOR RELEVANT INDICATORS
DATA CAPTURING AND INTEGRATION INFRASTRUCTURE
INPUT
DATA ENGINEER
PLANNING
JOBS
J
J
J
EXECUTION
REDUCTION ANALYSIS
OUTPUT
REDUCTION & RECONSTRUCTION ARTIFACTS
REDUCTION SOLUTION
DEPLOYMENT

Figure 5.2: A view of reduction analysis contextualized in an application scenario

The *output* of the reduction analysis is the specification of the *reduction solution*, i.e. which reduction techniques to apply to which indicators. This will transform the original representation of time-series data for each indicator into an optimal reduced representation for their storage. The deployment of reduction techniques and their application according to this specification will lead to the syntactic optimization for all the time-series data generated in the application scenario.

Although the execution of the reduction analysis is based on two main performance dimensions, *RTERR* and *COMPD* (as defined in 5.2.3), other dimensions related to computing time of reduction and reconstruction techniques may be considered in order to enhance the specification of the reduction solution. Indeed, depending on the requirements for real-time provision of services in the targeted application scenario, the technical performance related to computing time would influence in which component of the data capturing and integration infrastructure the reduction solution should be deployed.

5.3.2 A Procedural and Architectural Modeling of Reduction Analysis Planning

The first goal of the data engineer is to obtain the *reduction analysis plan*, which is a sequence of *reduction analysis jobs*. Each job contains three elements: (a) a *collection of time series* that will be analyzed together given their similar syntactic features (i.e. they belong to the same *time series family*); (b) a selection of *recommended reduction techniques* to analyze with that collection of time series, based on the expected performance in terms of compression ratio; and (c) the *expected compression ratio* to be obtained in that collection. This expected compression ratio, together with the volume of data (i.e. time series) pertaining to that family, are used as the criteria to prioritize the jobs and form a sequence (i.e. the plan) with them.

The data engineer obtains the reduction analysis plan with the support of an IT artifact, the *Reduction Analysis Planner* (RAP). The high-level architecture of the RAP is outlined in Figure 5.3. A description of the main modules composing the RAP is presented next, whereas their internal design is presented with further detail in Appendix B.

Data Loading Module. This module facilitates the entry point for the technical performance requirements and the data input, i.e. the collection of all the time-series data from different indicators for which the reduction potential is analyzed.

Syntactic Characterization Module. Given a time series, this module provides the functionality of extracting those syntactic features that are used to characterize the time series families in the knowledge base managed by the RAP. For each syntactic feature that is relevant to characterize a given time series (such as the examples outlined in 5.2.4), a function is provided to compute that feature over all instances in the analyzed time series and extract the average and deviation. These values, computed for all features, compose the syntactic characterization assigned to the analyzed time series.

Recommendation Module. This module manages the *syntactic characterization knowledge base*, which contains a typification of time series according to their syntactic features. For each type of time series, i.e. time series family, it contains the recommendation of the most suitable reduction techniques in terms of expected compression ratio to be obtained for a time series belonging to that family. Thus, a matching component (*matcher*) queries the knowledge base to search for the particular characterization that the *Syntactic Characterization Module* has assigned to a time series, and retrieves the data related to the time series family corresponding to that characterization, i.e. the recommended reduction techniques and the expected compression ratio. These recommendations are also filtered according to the given technical performance requirements for the application scenario.

Plan Scheduling Module. It groups all time series belonging to the same family in a time series collection, together with their related data obtained from the knowledge base, and generates the sequence of reduction analysis jobs following an order based on their expected compression ratio and the volume of time series in that collection.

Data Storage Module. This module implements the data persistence along the workflow involving the previously outlined modules. It stores the analyzed time series, their characterization and their assigned families and recommended techniques.

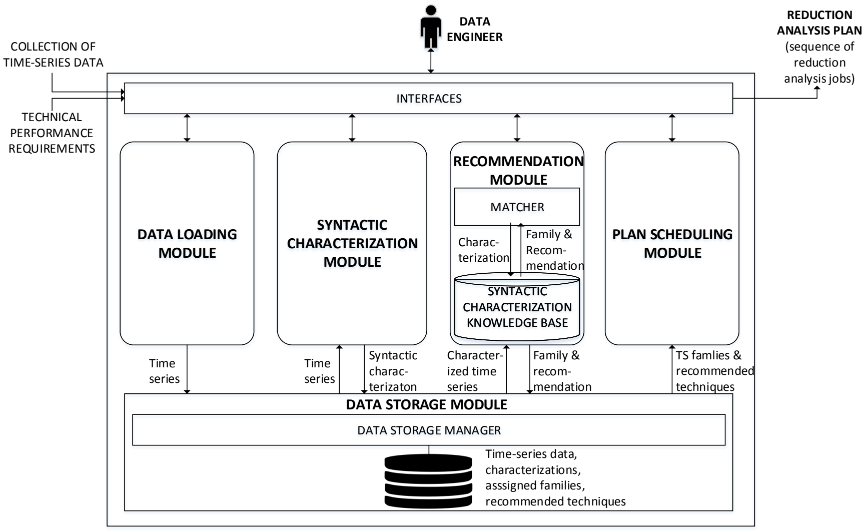

DATA ENGINEER
REDUCTION ANALYSIS PLAN
(sequence of reduction analysis jobs)
COLLECTION OF TIME-SERIES DATA
TECHNICAL PERFORMANCE REQUIREMENTS
INTERFACES
DATA LOADING MODULE
SYNTACTIC CHARACTERIZATION MODULE
RECOMMENDATION MODULE
MATCHER
Charac-terization
Family & Recom-mendation
SYNTACTIC CHARACTERIZATION KNOWLEDGE BASE
PLAN SCHEDULING MODULE
Time series
Time series
Syntactic charac-terizaton
Character-ized time series
Family & recom-mendation
TS famlies & recommended techniques
DATA STORAGE MODULE
DATA STORAGE MANAGER
Time-series data, characterizations, asssigned families, recommended techniques

Figure 5.3: High-level architecture model of the Reduction Analysis Planner

The high-level algorithm that is implemented by the combination of the modules presented above is described next.

```
Load $\{TS_i\}$ the collection of time series $TS_i$ to analyze
Load $\{R_q\}$ the set of technical performance requirements for the application scenario
Load $\{SC_f\}$ the set of functions to compute the syntactic characterization features registered in the knowledge base
for all $TS_i$ to analyze in the application scenario do
    Obtain $\{TS_{ij}\}$ the set of instances for the given time series $TS_i$ to analyze
    for each $SC_f$ in $\{SC_f\}$ do
        for each $TS_{ij}$ in $\{TS_{ij}\}$ do
            Compute $C_{fij} = SC_f(TS_{ij})$ {The application of $SC_f$ to $TS_{ij}$}
        end for
        Compute $C_{fi}Avg$ as the average of $SC_f(TS_{ij})$ for all $TS_{ij}$
        Compute $C_{fi}Std$ as the standard deviation of $SC_f(TS_{ij})$ for all $TS_{ij}$
    end for
    Compute the matching model in the Knowledge Base KB for an entry composed of $C_{fi}Avg$ and $C_{fi}Std$ for all $SC_f$
    Obtain $TS$FAM the assigned time series family from KB
    Obtain $\{f$RED$\}$ the recommended reduction techniques for $TS$FAM from KB
    Filter $\{f$RED$\}$ according to $\{R_q\}$
end for
Group $TS_i$ with the same assigned family $TS$FAM
Return a reduction analysis job for each group
```

We also modeled the use of the RAP by the data engineer using the constructs for process modeling provided by BPMN 2.0. Figure 5.4 presents this process model, represented as a collaboration diagram between two processes, one for the data engineer and one for the RAP. For the sake of simplification, the internal data persistence provided by the Data Storage Module is omitted in this diagram. In terms of the constructs provided by BPMN 2.0, the *pool* for the RAP contains different *lanes* for its other four main modules, thus representing which module is responsible for each *activity* in the process. In this process model all activities are represented as *tasks*. For those modeled as *loop tasks*, an *annotation* is included to indicate the iteration condition.

5.3.3 A Procedural and Architectural Modeling of Reduction Analysis Execution

Once the *reduction analysis plan* (i.e. the sequence of reduction analysis jobs) is obtained, as well as the *reduction performance requirements* for the particular application scenario, the data engineer follows the prioritization order specified in the plan in order to execute reduction analysis jobs within the constraints of time and resources allocated for this analysis.

Figure 5.5 represents the model for this execution process. When the data engineer is presented with a reduction analysis job, they can decide on its *assisted execution* (detailed below) or on its *automatic resolution.* Given that the job contains a selection of recommended techniques and expected compression ratio, based on the accumulated knowledge from previous analyses, the data engineer can decide on directly accept the recommended technique with the best expected performance. This option is highly useful, for instance, when the IBDS Provider is facing successive deployments for the same manufacturer or in the same manufacturing sector. In this case, the facilities to be monitored will be executing a highly similar manufacturing process, and most of the indicators to capture will be the same across facilities. Thus, the accumulated knowledge that led the RAP to propose its recommendation for a given time series family enables a direct application of this recommendation in the case of similar indicators.

On the other hand, in order to carry out an *assisted execution* of a specific reduction analysis job, the data engineer utilizes the Reduction Analysis Executor (RAE), an IT artifact that allows reducing time series using different techniques, evaluating the outcome of such reduction, comparing that outcome with the given performance requirements and presenting the results of this analysis. The assisted execution of a specific reduction analysis job is modeled as a subprocess (*execute job*) that is further detailed along this subsection.

Parting from the reduction analysis job to execute, the data engineer provides a *reduction analysis context* as input for the RAE. A reduction analysis context is composed of: the time series collection $\{TS_j\}$ provided in the job, one of the reduction techniques $g\textsc{red}$ recommended in the job, a set of actual parameters $\{z_p\}$ in order to analyze the performance of $g\textsc{red}$ once parameterized with each different z_p, and the set of performance requirements $\{R_q\}$. There are two main

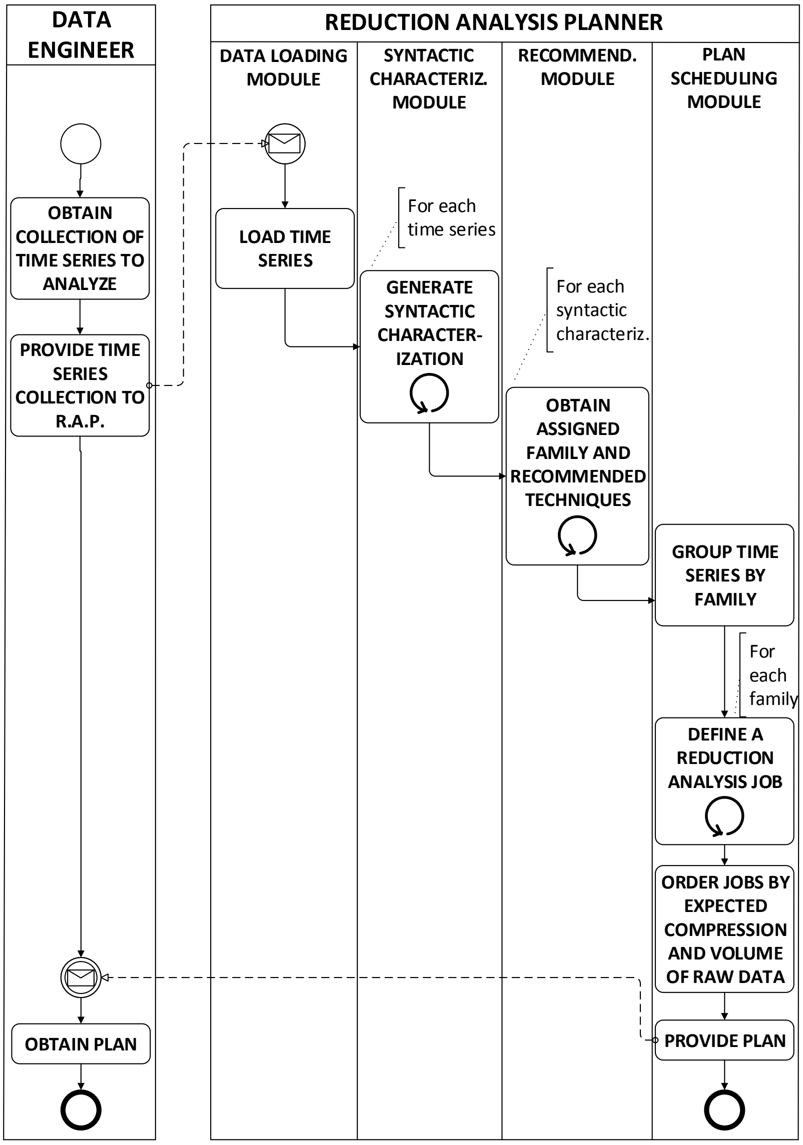

DATA ENGINEER
REDUCTION ANALYSIS PLANNER
DATA LOADING MODULE
SYNTACTIC CHARACTERIZ. MODULE
RECOMMEND. MODULE
PLAN SCHEDULING MODULE
OBTAIN COLLECTION OF TIME SERIES TO ANALYZE
PROVIDE TIME SERIES COLLECTION TO R.A.P.
LOAD TIME SERIES
For each time series
GENERATE SYNTACTIC CHARACTER-IZATION
For each syntactic characteriz.
OBTAIN ASSIGNED FAMILY AND RECOMMENDED TECHNIQUES
GROUP TIME SERIES BY FAMILY
For each family
DEFINE A REDUCTION ANALYSIS JOB
ORDER JOBS BY EXPECTED COMPRESSION AND VOLUME OF RAW DATA
PROVIDE PLAN
OBTAIN PLAN

Figure 5.4: Process model for the planning of reduction analysis

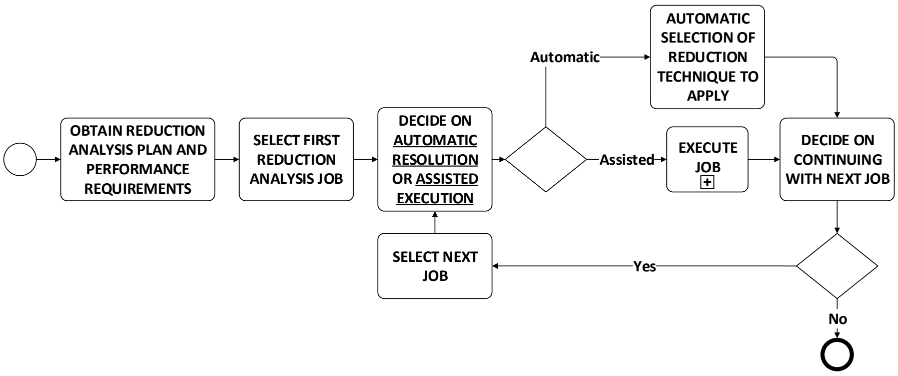

OBTAIN REDUCTION ANALYSIS PLAN AND PERFORMANCE REQUIREMENTS
SELECT FIRST REDUCTION ANALYSIS JOB
DECIDE ON AUTOMATIC RESOLUTION OR ASSISTED EXECUTION
Automatic
AUTOMATIC SELECTION OF REDUCTION TECHNIQUE TO APPLY
Assisted
EXECUTE JOB
DECIDE ON CONTINUING WITH NEXT JOB
SELECT NEXT JOB
Yes
No

Figure 5.5: Process model for the execution of the reduction analysis plan

performance dimensions considered in this set, as the main focus that delimits the analysis of reduction techniques: *RTERR* and *COMPD* (as defined in 5.2.3). Figure 5.6 summarizes the relationship and differences between the concepts of *job* (as extracted from the reduction analysis plan) and *context* (input for the RAE).

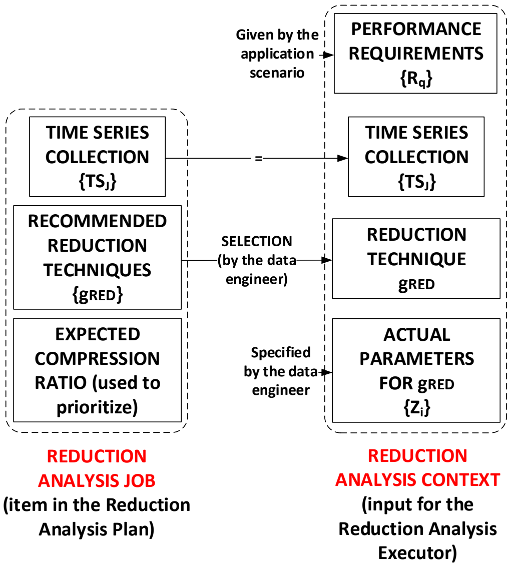

Given by the application scenario
PERFORMANCE REQUIREMENTS {R_q}
TIME SERIES COLLECTION {TS_J}
=
TIME SERIES COLLECTION {TS_J}
RECOMMENDED REDUCTION TECHNIQUES {gRED}
SELECTION (by the data engineer)
REDUCTION TECHNIQUE gRED
EXPECTED COMPRESSION RATIO (used to prioritize)
Specified by the data engineer
ACTUAL PARAMETERS FOR gRED {Z_i}
REDUCTION ANALYSIS JOB (item in the Reduction Analysis Plan)
REDUCTION ANALYSIS CONTEXT (input for the Reduction Analysis Executor)

Figure 5.6: Relationship between a *reduction analysis job* and a *reduction analysis context*

As the output of the analysis, the RAE updates an *exploration chart* of the state of the analysis. The different XY points to be shown in that chart correspond to the obtained performance (in terms of *RTERR* and *COMPD*) by the different parameterized reduction techniques derived from a generic technique.

Besides, the RAE generates a *summary table* for each generic technique gRED to be analyzed, where each cell contains the value obtained for each performance dimension (column) while applying gRED with the actual parameter z (row) to the analyzed time-series collection.

The internal architecture of the RAE (outlined in Figure 5.7). A description of the main modules composing the RAE is presented next, whereas their internal design is presented with further detail in Appendix C.

Data Loading Module. This module facilitates the entry point for the data input provided by the data engineer, i.e. a *reduction analysis context* specifying the time series collection, the reduction technique and the actual parameters to be analyzed, as well as the performance requirements to assess the results.

Reduction and Reconstruction Engine. This module manages the application of the appropriate reduction and reconstruction functions on the time-series data, according to the specified technique, thus obtaining the reduced and reconstructed versions of the provided data. When a job demands the analysis of a particular reduction technique, the Reduction and Reconstruction Engine will retrieve its implementation from the *Technique Library*, i.e. a library containing the implementation for all the reduction techniques that might be included in the reduction analysis jobs provided as input for the RAE.

Evaluation Module. This module contrasts the results of the reduction and reconstruction process with the performance requirements specified for the reduction analysis job, so that the values for the different performance dimensions are obtained and compared with the required thresholds.

Output Renderer. It refreshes the information presented to the data engineer via the exploration chart and the summary table, with the results of successive reduction analyses and their performance evaluation.

Data Storage Module. This module implements the data persistence along the workflow involving the previously outlined modules. It stores the original, reduced and reconstructed versions of the analyzed time series, the parameterization applied to analyzed techniques and their performance evaluation.

The *execute job* subprocess, where the data engineer uses the RAE to execute a reduction analysis job, is modeled as presented in Figure 5.8. In the same way as with the planning process, this process model represents a collaboration diagram between two processes, one for the data engineer and one for the RAE, and, for the sake of simplification, the internal data persistence provided by the Data Storage Module is omitted in this diagram. Thus, the *pool* for the RAE contains different *lanes* for its other four main modules, representing which module is responsible for each *activity* in the process. A textual description of the *execute job* subprocess, from the perspective of the data engineer (DE), is presented next.

As the first configuration step to start the reduction analysis of the time series collection $\{TS_j\}$ provided in the selected job, the DE selects one of the generic reduction techniques gRED in $\{g\text{RED}\}$. Apart from selecting gRED, the

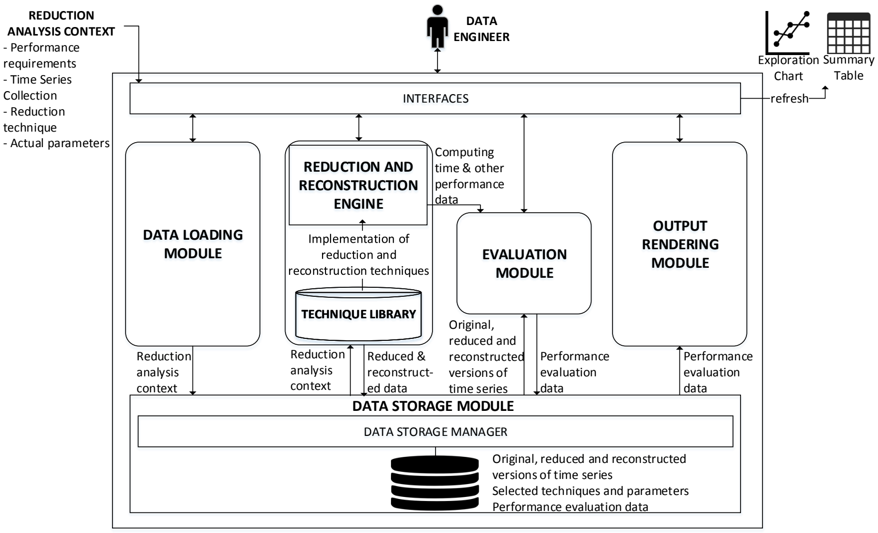

REDUCTION ANALYSIS CONTEXT
- Performance requirements
- Time Series Collection
- Reduction technique
- Actual parameters
DATA ENGINEER
Exploration Chart
Summary Table
INTERFACES
refresh
DATA LOADING MODULE
REDUCTION AND RECONSTRUCTION ENGINE
Implementation of reduction and reconstruction techniques
TECHNIQUE LIBRARY
Computing time & other performance data
EVALUATION MODULE
OUTPUT RENDERING MODULE
Reduction analysis context
Reduction analysis context
Reduced & reconstruct-ed data
Original, reduced and reconstructed versions of time series
Performance evaluation data
Performance evaluation data
DATA STORAGE MODULE
DATA STORAGE MANAGER
Original, reduced and reconstructed versions of time series
Selected techniques and parameters
Performance evaluation data

Figure 5.7: High-level architecture model of the Reduction Analysis Executor

DE specifies two actual parameters $\{z_1, z_2\}$ so that the RA program analyzes the application of gRED with both parameters to $\{TS_j\}$. The reason for selecting two actual parameters to begin the analysis is twofold: (a) it is adequate to begin exploring as less parameterizations as possible to obtain a prompt measurement of the running time of the application of a parameterized reduction technique derived from gRED to $\{TS_j\}$, and (b) starting with two actual parameterizations allows obtaining a measurement of the improvement, in terms of *RTERR* and *COMPD*, that is provided by the difference between the results with z_1 and z_2. This selection, together with the performance requirements, specifies the reduction analysis context to provide as input for the RAE.

The output generated by the RAE for this first analysis provides the DE with information to decide whether to continue analyzing the selected gRED and, if that be the case, define a new range of actual parameters for gRED to be analyzed. The values for that new range of actual parameters are estimated according to the results of the previous analyses (i.e. where they are located in the *parameterization exploration area* of the exploration chart, depending on the obtained values for *RTERR* and *COMPD*), in order to obtain results which are closer to the limits defined by the thresholds. How many actual parameters (i.e. parameterized reduction techniques derived from gRED) will be given as input to be analyzed is estimated based on the running time observed for the previous analyses and the availability of time and resources for the analysis task.

While the RAE is executing the analysis for each selected actual parameterization, the results are refreshed on the exploration chart and the summary table, so that they can be observed by the DE while the RAE concludes each analysis. This also allows the DE to cancel some or all of the already programmed analyses but not executed yet (i.e. the remaining actual parameterizations to be

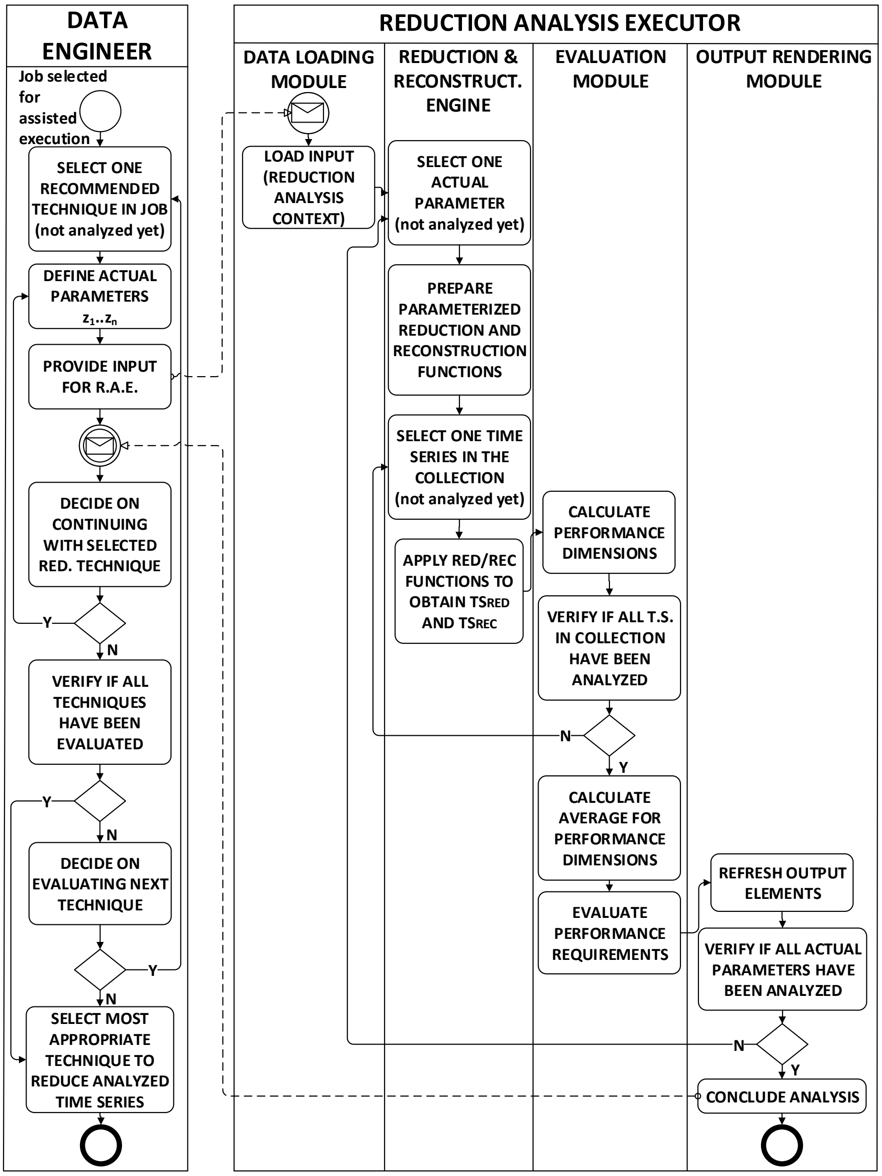

DATA ENGINEER
Job selected for assisted execution
SELECT ONE RECOMMENDED TECHNIQUE IN JOB (not analyzed yet)
DEFINE ACTUAL PARAMETERS z1..zn
PROVIDE INPUT FOR R.A.E.
DECIDE ON CONTINUING WITH SELECTED RED. TECHNIQUE
Y
N
VERIFY IF ALL TECHNIQUES HAVE BEEN EVALUATED
Y
N
DECIDE ON EVALUATING NEXT TECHNIQUE
Y
N
SELECT MOST APPROPRIATE TECHNIQUE TO REDUCE ANALYZED TIME SERIES
REDUCTION ANALYSIS EXECUTOR
DATA LOADING MODULE
LOAD INPUT (REDUCTION ANALYSIS CONTEXT)
REDUCTION & RECONSTRUCT. ENGINE
SELECT ONE ACTUAL PARAMETER (not analyzed yet)
PREPARE PARAMETERIZED REDUCTION AND RECONSTRUCTION FUNCTIONS
SELECT ONE TIME SERIES IN THE COLLECTION (not analyzed yet)
APPLY RED/REC FUNCTIONS TO OBTAIN TSRED AND TSREC
EVALUATION MODULE
CALCULATE PERFORMANCE DIMENSIONS
VERIFY IF ALL T.S. IN COLLECTION HAVE BEEN ANALYZED
N
Y
CALCULATE AVERAGE FOR PERFORMANCE DIMENSIONS
EVALUATE PERFORMANCE REQUIREMENTS
OUTPUT RENDERING MODULE
REFRESH OUTPUT ELEMENTS
VERIFY IF ALL ACTUAL PARAMETERS HAVE BEEN ANALYZED
N
Y
CONCLUDE ANALYSIS

Figure 5.8: Process model for the *execute job* subprocess

analyzed), if the DE observes the results of the ongoing analyses and concludes that the obtained *RTERR* and *COMPD* are too far from the thresholds. Thus, the DE can adjust a new set of actual parameters to analyze.

By repeating the previous tasks, the information provided by the RAE as the result of each new analysis helps the DE to take decisions on the new parameterizations to analyze, looking for obtaining results that are closer to the thresholds delimiting the *parameterization exploration area* (see Figure 5.9).

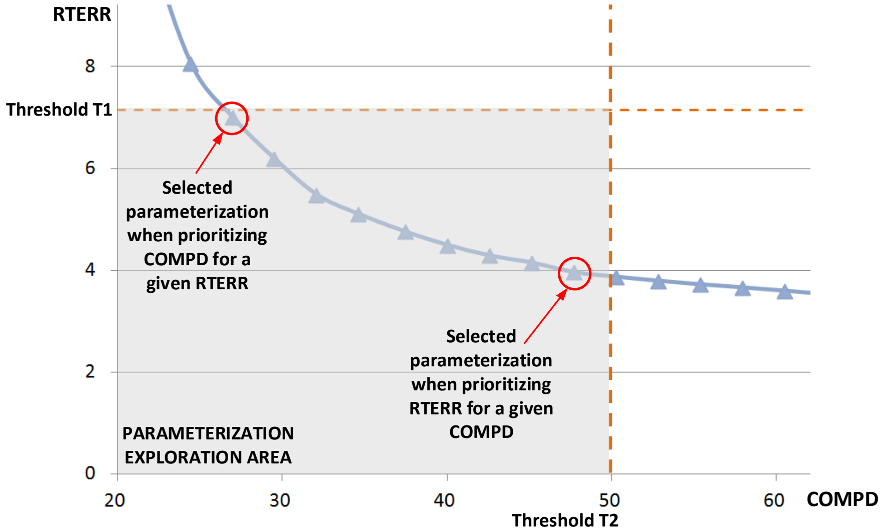


Figure 5.9: Detail of an exploration chart

After some successive repetitions, the decision on continuing the analysis of the selected *g*RED is based on whether one of the following two conditions is met:

A) The allocated time to explore the reduction potential of *g*RED for the time series collection $\{TS_j\}$ has been consumed without finding any result inside the *parameterization exploration area*. In this case the technique can be discarded or, if the consumption of extra resources can be assumed, the allocated time can be extended so that the exploration process can continue.

B) There are some obtained results that provide a characterization of how the parameterized reduction techniques derived from *g*RED behave (i.e. which results they provide in terms of *RTERR* and *COMPD*) in the closest area to the limits defined by the thresholds for those two performance dimensions. If this condition is met, as a final step to conclude the exploration of this reduction technique for $\{TS_j\}$, the DE selects the optimal parameterization for this reduction technique when applied to the analyzed time series collection.

Next, the DE decides whether to continue the analysis with another of the recommended reduction techniques for $\{TS_j\}$. This is repeated until all recommended techniques have been analyzed or until the DE decides not to continue with the selected reduction analysis job, because the allocated time and resources have been consumed. As the final task for the *execute job* subprocess, given the obtained reduction performance assessment for the analyzed techniques, the DE selects the most appropriate technique and parameterization to reduce the time series collection provided in the reduction analysis job.

The successive repetition of the *execute job* subprocess for the reduction analysis jobs in the plan (within the given time and resource constraints for the analysis) allows the DE to specify the *reduction solution* to be deployed in the application scenario, i.e. the final output of the reduction analysis. This reduction solution is specified as a series of associations between (a) a collection of time series captured for some indicators in the application scenario and (b) the most appropriate reduction technique and parameterization to apply to those data in order to obtain their reduced representation. The deployment of a reduction solution following that specification into the data capturing and integration infrastructure will apply the appropriate reduction technique for each captured indicator. Moreover, if the constraints have not allowed completing the analysis for all reduction analysis jobs, the approach ensures that the solution built so far prioritizes the reduction of those time-series data where the highest benefit (in terms of savings in storage costs) is obtained.

5.4 Conclusions

The proposed design is sustained by the main principles that design science research poses for a purposeful design artifact: the identification of the business problem and needs provides *relevance* to the proposed solution; the grounding on the identified synergies and gaps with respect to the existing knowledge base of related work provides *rigor* and motivates the opportunity for a proposal; the *design cycle* parts from a conceptualization of the key areas to address and leads to *building* testing-oriented IT artifacts to conduct a *field testing* in a real-world business setting in order to *evaluate* the hypotheses that ground our approach. This setting provided access to real production data generated in these environments and illustrated the suitability of the approach for its use in industrial application scenarios and how its use contributes to considerable savings in storage costs for IT companies developing Big Data services for manufacturing business settings.

The instantiation and application of the proposed design not only fulfills the goal of obtaining the best overall reduction possible thanks to the combination of different families of reduction techniques; it also allows refining the association between time series families with different syntactic features and the recommended reduction techniques, in order to increase the efficiency in subsequent analyses for similar scenarios. Indeed, the successive application of the approach in more scenarios where new indicators with different syntactic features are captured, as well as the inclusion of new reduction techniques, will give continuity to this refinement. As a data engineer uses an instantiation of the proposed design to analyze increasingly more scenarios, the refinement of the syntactic characterization knowledge base supports an efficient knowledge management process of the insights and lessons learned extracted from different deployments. This accumulated knowledge from previous deployments and analyses enables the savings in resources and time allocated for successive reduction analyses. This is sustained in the presented proposal by the automated resolution of analysis jobs or by their assisted execution with the proposed IT artifacts.

Chapter 6

A Decentralized Hybrid Architecture for the Data Capturing and Integration Platform of IBDS Providers

Another of the challenges where we focused our research was the required architecture for IBDS Providers to design the platform sustaining their business in a global context like the one detailed in chapter 4. This platform should integrate the required key enabling technologies in order to obtain the data to be exploited through smart services, according to the diverse use cases for each targeted sector. Indeed, it is in such a platform where the reduction solution obtained as the result of the reduction analysis presented in chapter 5 would be deployed, in order to ensure a better cost-sustainability for an IBDS Provider's platform and, therefore, for their business.

Many of the existing conceptual proposals to design Big Data systems part from the availability of a *raw data repository* and, therefore, focus on the design of the required architecture for effective and efficient data exploitation processes. Nevertheless, there is a non-trivial gap to be covered between the demand for such a data repository in a manufacturing application scenario (where the IBDS Provider aims at supplying their services) and its eventual availability. This gap presents a series of practical requirements that the platform must fulfill [NnSBI16] in order to sustain an IBDS Provider's business and, at the same time, to be aligned with the business strategies of (a) servitized CEMs with whom the IBDS Provider establishes partnerships to supply smart services for specific manufacturing sectors, and (b) manufacturers in those sectors aiming at leveraging the supplied solution to shift their manufacturing processes towards a Smart Manufacturing approach. It is in that gap (up to the assurance of the availability of the data repository to exploit) where we contribute with a design artifact modeling the architecture with which to deploy the platform sustaining an IBDS Provider's business, in order to give an appropriate answer to all those requirements.

In order to contribute to solve this particular challenge, we leveraged our access to a real-world business setting. This allowed us to examine how more than 60 manufacturing facilities worldwide from different sectors were using a combination of technologies to capture, integrate and monitor relevant production data. The examination of these real-world manufacturing scenarios allowed us to identify several core components whose purpose and specification contributed to fulfill the main non-functional requirements identified for the targeted scenarios.

Thus, we conceived a *Distributed Hybrid Architecture* (DHA) as a design artifact modeling the data capturing and integration platform of IBDS Providers. The DHA comprises two levels of data management: one IIoT-based level is oriented to the local management of raw data at each connected manufacturing facility, and another cloud-based level is oriented to the management of a Big Data Lake with data from all connected and analyzed facilities. While there exist other proposals (presented along section 6.2) that are also based on the combination of IIoT and Cloud Computing, the innovative features of the proposed architectural design allow fulfilling the non-functional requirements for an Industrial Big Data platform to sustain the business strategy of an IBDS Provider.

This chapter details the design of this DHA, beginning with the identification of the main requirements derived from the targeted Smart Manufacturing scenarios that influence its design. It also presents an analysis of the related references on architectural proposals for Smart Manufacturing scenarios, in order to identify limitations to overcome and synergies to leverage. These synergies, as well as the observation and analysis in the conducted case study, ground the design of the DHA and how its internal modules are combined to solve the main requirements derived from the targeted scenarios.

6.1 Requirements Derived from the Targeted Smart Manufacturing Scenarios

The analysis of non-functional requirements from the perspectives of the different agents, as presented in the characterization of the targeted Smart Manufacturing scenarios in chapter 4, derives the main requirements for an IBDS Provider to design their platform. These requirements are presented along this section and summarized in Table 6.1.

In order to provide an architectural design for an IBDS Provider's platform, one critical aspect is that it has to be conceived to support a global-scale service to be marketed to manufacturers in different sectors, not as a customized project for a few plants. Answering market's demand should abide by a properly dimensioned investment. Deployments in each facilitity should involve a *restricted volume of fieldwork* and *ad hoc* configuration to ensure a sustainable business model, given the global scale of the service.

There is also the need for *flexibility to integrate the solution into different* *industrial Operational Technology (OT) infrastructures* already running in the manufacturing facilities to be monitored. Most manufacturing companies have been deploying some OT infrastructure over the years towards a progressive automated management of their manufacturing processes. However, this infrastructure was not necessarily designed to facilitate an efficient data export outside the plant. The integration of the solution must be designed according to the *technological reality* and the predominant standards among manufacturing companies. For that purpose, the solution must be able to capture industrial data from different production environments from diverse manufacturing sectors. This involves dealing with equipment and industrial components with very different capabilities for data export, as well as with different qualities of connectivity. Therefore, it is necessary to integrate the solution with the data export capabilities that are available in each case and to include the necessary components in the solution to address this heterogeneity and to overcome performance shortcomings in those capabilities.

Following an *incremental approach* is of paramount importance in order to facilitate scalability and to support a progressive investment and partial returns. On one hand, the architecture must facilitate the launching of initial projects with a small amount of plants and reduced-scale volumes of data. On the other hand, it must progressively scale to work with large-scale data as more facilities are connected and insights from prior deployments are leveraged. Indeed, a high multiplicity of deployments allows an IBDS Provider to benefit from an *economy of scale* at two different levels. First, the know-how derived from a deployment in a particular manufacturing sector, enabled by the partnership with manufacturers operating in that sector, can be leveraged in the subsequent deployments for that same sector (*in-sector*). Second, some *cross-sector* elements can also be leveraged in deployments in other sectors. These cross-sector elements are related to the common components in OT infrastructures in manufacturing facilities (for instance, field buses to connect to) and common strategies for data optimization regarding their quality and their cost in terms of storage resources.

From the customers' perspective, it is required to yield a *progressive return of investment* when manufacturers engage in the use of the solution. The expected long-term savings for manufacturers depend on the potential success of future smart services based on predictive models. Therefore, it is necessary that the deployed architecture offers a basic and sustainable *short-term value as an immediate return of investment*, while waiting for the added value to be potentially obtained in the medium-long term.

Acceptance of the solution by manufacturers would be facilitated by a *non-intrusive approach* that *avoids interference with their manufacturing process operation.* The OT infrastructure should be kept intact as much as possible, leveraging current data export capabilities and not requiring additional IT projects. The deployment of the solution must demand a very limited effort from the customer side, at least not until some value is offered thanks to the storage, processing and analysis of their data.

Last, the *appropriate security mechanisms* must be taken into account when deploying new IT infrastructure that can exchange data through a gateway to the

Restricted volume of fieldwork and *ad hoc* configuration
Flexibility for the integration with different OT infrastructures
Incremental approach to facilitate scalability
In-sector and cross-sector economy of scale
Progressive return of investments for customers
Assurance of short-term value as an immediate return of investment
Non-intrusive approach to facilitate acceptance by manufacturers
Avoid interference with current manufacturing process operation
Appropriate security mechanisms for data exchange and for keeping the infrastructure safe from external threats

Table 6.1: Summary of requirements for the architecture of the solution derived from the characterization of the targeted scenarios

Internet. Those security mechanisms must control that no other infrastructure, apart from the one deployed to offer the data-enabled service, will have access to the data and the OT infrastructure of the monitored facility.

6.2 Analysis of Related Work

This section analyzes diverse proposals of architecuture models and conceptual frameworks for data integration platforms in manufacturing application scenarios. These approaches have emerged with the rise of Smart Manufacturing and the different initiatives promoting its adoption among manufacturers. The analyzed approaches vary in their degree of abstractness and generic nature with respect to the composing elements and targeted manufacturing scenarios. The analysis of these proposals has two main goals. The first goal is to verify to which extent the practical requirements identified as motivation are covered in existing proposals and, if substantial limitations are identified, to reinforce our motivation to propose a contribution that extends and complements existing work in order to fulfill such requirements. The second goal is to identify synergies with the analyzed proposals in order to integrate relevant components in our contribution and to establish a tighter connection with existing work.

6.2.1 Relevant References on Architectures and Generic Frameworks

Among the relevant Internet of Things (IoT) reference architecture models [WE16], a relevant milestone in this area is the Reference Architecture Model for Industry 4.0 (RAMI 4.0) [Pla16] that guides the development of Industry 4.0 applications in a standardized way. Reference [FKF16] proposes an architecture model based on RAMI 4.0 for a Socio-Cyber-Physical System. In [LRPn16] the key concepts of RAMI 4.0 are detailed and it is presented a model of a PLC as an Industry 4.0 component, based on the structure for such a component proposed by RAMI 4.0. In [LBK15] a five-level conceptual framework is proposed as a guideline to implement Cyber-Physical Systems (CPS) in Industry 4.0-based manufacturing systems, and in [WTS+16] a prototype platform and a software-defined architecture are defined for IoT in the context of Industry 4.0.

The paradigm of Cloud Manufacturing [ZLT+14] provides the foundation for a five-level architecture [TZXZ14] to enable intelligent perception and access of manufacturing resources via this paradigm and IoT technologies. In [SBS16] it is presented a research agenda in order to develop practical methodologies and instrumentation to deploy Cloud Manufacturing systems.

We can also highlight some other relevant proposals of generic frameworks. In [HVH15] it is presented a model-based framework to integrate data elements of distributed data systems and sources, merging XML-based integration technologies with the concept of Enterprise Application Integration. In [SYM+15] it is introduced a framework based on the formal language SystemJ to design and implement distributed manufacturing automation systems. In [MB14] a brief introduction is presented on a generic architecture for IoT applications and services in manufacturing industry, connecting manufacturing systems with cloud computing environments.

There are also some noteworthy proposals of architectural solutions referring to the use of Big Data technologies in specific manufacturing applications or environments. In [PGL12] it is outlined an architecture for a data ingestion system integrating different Big Data open-source technologies to gather and store high-throughput machine logs from a set of milling machines. Reference [OLBO15] presents an embedded study in a large-scale manufacturing facility to identify the requirements for a system model to integrate, process and analyze industrial equipment data for maintenance applications in such an environment. Reference [BM12] presents a framework based on Hadoop to analyze machine maintenance data collected from sensors embedded in industrial machines, in a cloud computing environment. In [YPC+14] a system architecture based on Hadoop is presented for manufacturing process analysis. Reference [KWL15] presents a case study where an architecture of layers and functional building blocks is proposed as a blueprint for prescriptive enterprise systems in the process manufacturing industry. In [RTKM16] the EU-funded research project Proteus is outlined. It aims at using Apache Flink for scalable analytics and visualization in Industry 4.0 and will analyze a use case in the steel manufacturing sector.

Regarding the use of Big Data technologies, besides these proposals focused specifically on manufacturing, it is also worth mentioning the transversal proposal of the Lambda Architecture [MW15]. This generic, abstract architecture guides the design of a Big Data system as a set of layers that implement different batch or stream processing steps to create the required views on top of massive-scale data.

There are other worth-mentioning EU-funded research projects in this same area. Reference [KCK+15] describes an IoT platform designed with a four-level architecture and the related prototypes for the car manufacturing industry developed within the Ebbits project. Reference [SGWR14] presents the cloud-based system developed within the iMAIN project for stress and condition monitoring, planned to be demonstrated on forming presses. In [JZFV16] it is presented the generic data processing architecture to be used in the ongoing Mantis project to predict the wear of machinery components.

6.2.2 Limitations and Potential Synergies of Analyzed Work

From the perspective of an IBDS Provider who wants to leverage existing proposals to design the platform that sustains their global-scale business, there are *two major limitations* in the analyzed proposals that are related to the identified requirements for the solution.

The first major limitation is related to the fact that reviewed proposals mostly remain at a conceptual level, proposing integral approaches as a vision of future scenarios. Furthermore, they envision solutions *as a whole*, i.e. to be deployed in scenarios where the deploying party has total control of the infrastructure and therefore can redesign it completely following the proposed approach. This is definitely not the case in the targeted Smart Manufacturing scenarios where an IBDS Provider deploys their solution, as they aim at supplying smart services for manufacturers *who have a running manufacturing business supported in an already deployed OT infrastructure*. IBDS Providers must integrate their solution into the operating infrastructure of the manufacturing facilities where data must be captured and exploited. Therefore, for the solution to the accepted by manufacturers, it is of great importance to adopt a non-intrusive approach that integrates smoothly and requires to alter as minimum as possible the existing infrastructure and the operation of the manufacturing process. The success of an IBDS Provider's business is highly dependent on offering a *manufacturer-friendly* transition to Smart Manufacturing. Moreover, for a given particular customer, it is necessary to use an incremental approach, starting with a reduced scope that can be extended once some first visible outcome is ensured in order to provide value.

The second major limitation is related to those proposals detailing approaches for specific types of application scenarios. Given their focus on specific scenarios or use cases, these proposals mostly part from a predefined set of industrial data source types, i.e. the industrial components that are present in the application scenario and generate the data to be captured. Therefore, the proposals design their data ingestion components accordingly. However, for those scenarios where an IBDS Provider aims at supplying their services, it should not be presupposed a closed set of industrial data source types. The platform should have the flexibility to *evolve and adapt their data capturing functionalities* in order to extract data from new industrial components, as new manufacturers from different sectors manifest their interest in smart services and thus provide new application scenarios with different technical requirements for data extraction. These requirements will depend on the characteristics of the OT infrastructure already deployed in the facilities to be monitored. Furthermore, the adaptation to include new functionalities and data extraction protocols should not require a major reorganization of the platform or costly field work, so that it does not hamper the sustainability of a global-scale business to provide smart services worldwide.

On a related matter, in those cases where some implementation is presented, it is in a very preliminary or prototype state, tested in simplified or simulated scenarios with synthetically-generated data. They lack case studies in real-world manufacturing business scenarios that impose specific requirements related to a sustainable business model or the need to provide progressive valuable returns for customers in the short term while waiting for a medium-long term value. Some of the analyzed EU-funded projects outline promising case studies [RTKM16][JZFV16], but they still are in a very early stage of analysis. Furthermore, although several cases are presented as designed for distributed scenarios, this is mainly due to data being gathered from independent machines implementing steps of a particular manufacturing process. Therefore, it does not imply a global-scale business context with different companies and facilities from various manufacturing sectors distributed worldwide.

Nevertheless, some of the analyzed references present interesting concepts that are closely related to necessary elements in the Smart Manufacturing scenarios where an IBDS Provider can supply their services. For instance, the extended view of a PLC presented in [LRPn16] introduces the idea of a component that provides the process data of the PLC controller to the IP network in a reliable and secure way. Such an approach could also address one of the challenges highlighted in [SGWR14], which is the development and integration of embedded devices with data preprocessing capabilities in order to capture relevant information. Furthermore, the integration of cloud computing environments is a valuable resource in our case, not as the integration of single, distributed steps of an instance of manufacturing process [SBS16] but as a way to centralize the massive-scale data from all the analyzed manufacturing facilities where smart services are to be provided.

Moreover, this idea of massive-scale centralized data allows us to draw important synergies and complementarities between our approach and two main constructs regarding Big Data systems: the *Lambda Architecture* paradigm [MW15] and the concept of *Big Data Lake* [O'L14]. Our contribution is focused on the architecture for the platform that ensures the availability of the massive-scale raw data repository. This encompasses two different types of data management: the distributed capture of raw data from all analyzed manufacturing facilities, and the accumulation of those data in a centralized repository. This centralized raw data repository resembles the concept of Big Data Lake, in the sense that raw data are accumulated from their original sources with no schema-based transformation prior to their exploitation. Besides, this repository also can play the role of *master dataset* upon which different data exploitation layers are deployed following the principles of the Lambda Architecture, i.e. processing the accumulated data in the lake in order to enable the required views for further exploitation by end users. Figure 6.1 frames the focus of our contribution in the context of its potential synergies with Lambda Architecture and Big Data Lakes, and outlines how it supports the lifecycle of raw data since they are generated in a component of a manufacturing facility distributed worldwide until they are centralized and accumulated for their later exploitation.

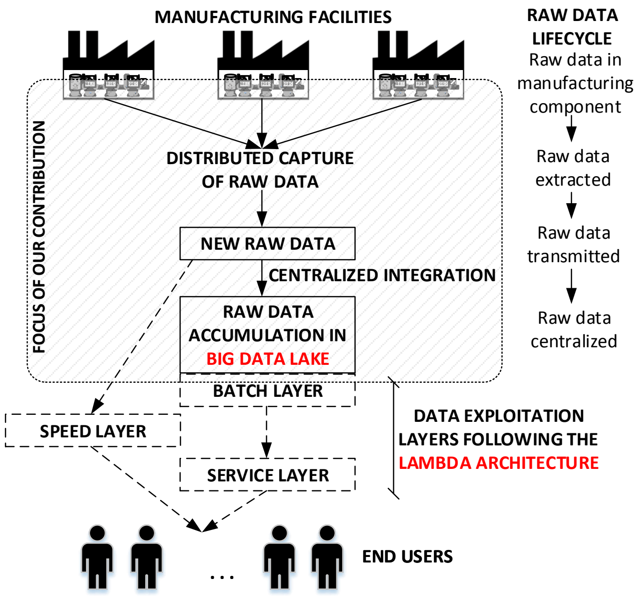

MANUFACTURING FACILITIES
RAW DATA LIFECYCLE
Raw data in manufacturing component
Raw data extracted
Raw data transmitted
Raw data centralized
FOCUS OF OUR CONTRIBUTION
DISTRIBUTED CAPTURE OF RAW DATA
NEW RAW DATA
CENTRALIZED INTEGRATION
RAW DATA ACCUMULATION IN BIG DATA LAKE
BATCH LAYER
SPEED LAYER
SERVICE LAYER
DATA EXPLOITATION LAYERS FOLLOWING THE LAMBDA ARCHITECTURE
END USERS
...

Figure 6.1: Relationship between our contribution, Big Data Lake and Lambda Architecture

6.3 Grounding for the Proposed Design Artifact

Our goal is to propose a design artifact modeling the architecture of a data capturing and integration platform that fulfills the main non-functional requirements identified in section 6.1 for the targeted Smart Manufacturing scenarios. Thanks to the access to a real-world business setting in our case study, we examined the cases of more than 60 manufacturing facilities and the combination of technologies used in those cases to capture and integrate relevant production data. Thus, we identified core technological components that contributed to fulfill the posed requirements and we integrated their purpose and specification in our design artifact.

As well as this, we also leveraged the synergies with various conceptual and methodological proposals: the Lambda Architecture and Big Data Lakes (as outlined in section 6.2), the idea of *bringing computation to data* and its relationship with the concept of *Fog Computing*, and the possibilities that these concepts enable to efficiently cover the different steps of the data lifecycle along the phases of a KDDM process. The combination of all these elements grounded the proposal of a *Decentralized Hybrid Architecture* (DHA), whose general design is outlined at the end of this section.

6.3.1 Observation and Contrast with the Real-World Business Setting of our Case Study

In order to conceive a design artifact as a contribution for this challenge, we had the advantage of counting on the real-world business setting where we conducted our two-level case study. This setting gave us access to more than 60 cases of manufacturing facilities worldwide from different sectors (processing of metallic coils, high-precision machining, high-precision milling and broaching, etc.) and allowed us to examine how diverse technologies for capturing and integrating real production data had been combined and deployed in those real-world business scenarios. Among the deployed technologies we identified and extracted core components that contributed to give answer to the posed requirements for our contribution. As long as we put them in relationship with those requirements, we would validate and reinforce the applicability of those components as a solution to the requirements that are not fulfilled in existing proposals. Therefore, having established a link between those components and the fulfilled requirements, we integrated them into our conceptualization of an architecture model, i.e. the proposed design artifact. Thus, this design artifact would constitute a contribution extending and complementing existing proposals in order to give an appropriate answer to the identified requirements.

We identified the following four core components among the deployed technologies as crucial for the fulfillment of the posed requirements: the combination of IIoT and Cloud Computing for local and global data management, the connection to existing OT infrastructures in manufacturing facilities and the use of local data persistence, a secured communication between local and global levels of the platform, and the cloud-based components for data exploitation. Their relationship with the requirements is summarized in Table 6.2. Thus, we integrated their purpose and specification in the design of the proposed architecture, given their applicability to solve the posed requirements.

6.3.2 Synergies with other Data-related Conceptual and Methodological Proposals

Apart from the synergies with conceptual proposals such as the Lambda Architecture and Big Data Lake, we also drew from other relevant conceptual and methodological proposals in order to conceive the contributed design artifact.

The philosophy behind the design of the architecture devised in this work is strongly aligned with the key ideas that motivated the origin of Big Data technologies [NnI15] and their close relationship with Cloud Computing solutions. When Google faced the problem of computing efficiently their PageRank algorithm with large-scale data, they devised a solution where those data were divided into *chunks* and stored across several nodes in a cluster. These nodes were commodity servers sharing replicas of those chunks of data to ensure fault tolerance. The computation model devised to process these data (coined as MapReduce [DG04]) was derived from the idea of *bringing computation to data*, i.e. a specific processing task dealing with some subset of the data was assigned to the cluster

Core component	**Requirements derived from the targeted scenarios fulfilled by the component**
Combination of IIoT and Cloud Computing for local and global data management	Incremental approach to facilitate scalability. Restricted volume of fieldwork and *ad hoc* configuration. Non-intrusive approach to facilitate acceptance by manufacturers. Avoid interference with current manufacturing process operation.
Connection to existing OT infrastructure and local data persistence	Flexibility for the integration with different OT infrastructures. In-sector and cross-sector economy of scale.
Secured communication between local computing devices and the cloud environment	Appropriate security mechanisms for data exchange and for keeping the infrastructure safe from external threats.
Cloud-based components for data exploitation	Progressive return of investments for customers. Assurance of short-term value as an immediate return of investment.

Table 6.2: Correspondence between analyzed core components and fulfilled requirements

node where that subset was stored.

This key idea is closely linked to the concept of *Fog Computing* [BMZA12], which was firstly proposed in the context of connected vehicles [Bon11]. Fog Computing proposes leveraging the computing power of distributed computing nodes. These computing nodes are not deployed in a cloud infrastructure but closer to the field elements where data-related computation is required. In the targeted scenarios, this approach can be applied to the deployment of computing nodes into the manufacturing facilities to be monitored, so that all necessary data-related computation is solved by an efficient combination of distributed and centralized, cloud-based nodes.

This enables a powerful synergy with the data lifecycle and the stages in a Knowledge Discovery and Data Mining (KDDM) process, particularly with data preprocessing. In terms of a KDDM process, data preprocessing is usually presented as a phase focused on preparing and/or reducing data to create a data view that fulfills the requirements of a particular data mining problem or as input for a particular data mining algorithm [GLH15][KM06]. Instead, the potential synergy in this scenario with the locally distributed nodes is to enable a local data preprocessing step that aims at enabling an efficient data transmission and subsequent centralized storage, being non-dependant on any particular data analytics need that wants to be solved in the application scenario. In other words, the scope of data preprocessing in locally distributed nodes is to help solving more efficiently the sustainability requirements for an IBDS Provider's business, while providing an optimized version of raw data that still can be leveraged to solve the elicited data exploitation needs for the application scenario.

6.3.3 Design of the Decentralized Hybrid Architecture

As mentioned earlier, the design of the DHA adapts the key concept of bringing computation to data to this context's requirements. In order to fit the characteristics of the context of an IBDS Provider, the DHA combines *two different levels of data management*. On one hand, there is a *local level*, where computing nodes are deployed and integrated into the OT infrastructure of the manufacturing facilities worldwide whose data are intended to be captured. This brings the first steps of computation and data processing closer to where data are originated. The design of these nodes leverages the purpose and specification of the use of IIoT technologies in the cases analyzed in the real-world business setting. Thus, the architecture integrates IIoT technologies in those local computing nodes, which are able to capture raw data from each relevant indicator in the monitored manufacturing facilities and send them over the Internet. Besides, the progressive upgrade of their functionalities enables the preprocessing of those raw data for their efficient transmission and subsequent centralized storage.

On the other hand, there is a *global level*, based on a Cloud Computing environment (i.e. a cluster of computing nodes supplied by a cloud services provider) for the centralization of all captured data. This cloud-based level enables the subsequent development and deployment of exploitation solutions on those data. The cloud computing environment contains the tools for monitoring and managing the correct functioning of all the architecture. It also centralizes the preprocessed data from manufacturing plants in a Big Data Lake. Several functionalities are enabled in the cloud computing environment to exploit the *lake*, including a built-in service for the real-time and historic visualization of each monitored indicator.

Therefore, the architecture is considered *hybrid* in the sense that it combines local and global approaches [BM12], i.e. it is a two-level cluster (see Figure 6.2) composed of a *decentralized pool of local computing nodes* and a cluster of nodes in a *cloud computing environment*. These two levels constitute the design artifact that integrates the key functionalities of the core components analyzed in the real-world business setting of our case study. Thus, the proposed design artifact can be leveraged by practitioners who need to fulfill the identified requirements in these scenarios.

Other existing architecture proposals are also based on the combination of IIoT and Cloud Computing levels. However, the *main differential point* of our proposal is the inclusion of the necessary architecture components and operation processes (deployment, monitoring, upgrading) that allow fulfilling the non-functional requirements for an Industrial Big Data platform to sustain the business strategy of an IBDS Provider. Thus, a platform designed according to the proposed architecture allows an incremental and non-intrusive deployment of the platform on OT infrastructures already running in manufacturing facilities, as well as the successive upgrade of the supported functionalities to cover more ap-

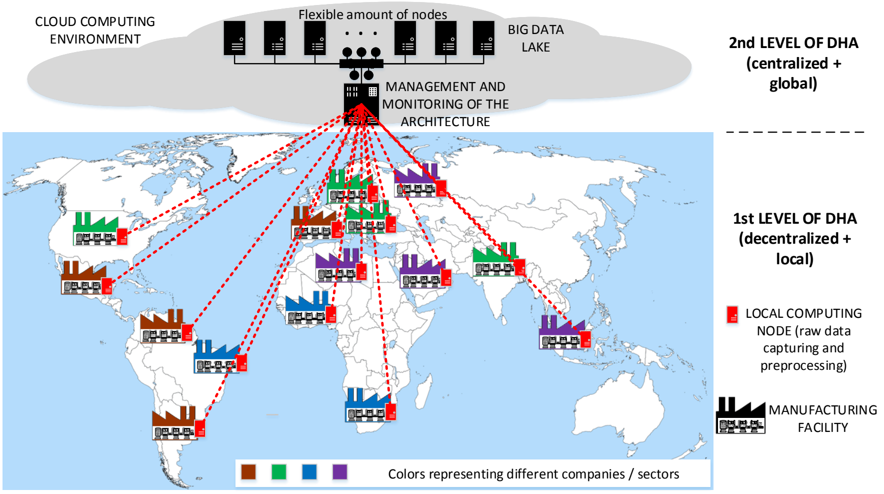

CLOUD COMPUTING ENVIRONMENT
Flexible amount of nodes
BIG DATA LAKE
MANAGEMENT AND MONITORING OF THE ARCHITECTURE
2nd LEVEL OF DHA (centralized + global)
1st LEVEL OF DHA (decentralized + local)
LOCAL COMPUTING NODE (raw data capturing and preprocessing)
MANUFACTURING FACILITY
Colors representing different companies / sectors

Figure 6.2: High-level schema of the Decentralized Hybrid Architecture

plication scenarios and to progressively support more data transformation steps towards the provision of smart services. These differential aspects constitute the main innovative features of the proposed architectural design with respect to other analyzed proposals and give an effective answer to the main requirements of the scenarios where an IBDS Provider can supply their services.

6.4 Design of Local Computing Nodes

Local computing nodes encapsulate all functionalities regarding to the extraction of raw data from manufacturing components and the transmission of captured data over the Internet to the centralized repository. Besides, they can be delivered to the manufacturing facility where they must be deployed and, once connected, they can be remotely set up to start functioning, removing the need for on-site deployment work. These features contribute to fulfill two important requirements: a sustainable deployment that *does not require costly field work*, and a *non-intrusive deployment approach* that facilitates the acceptance of the solution by manufacturers. Moreover, the progressive deployment of local nodes as new agreements are reached with new manufacturing customers complements the scalability of the cloud computing and facilitates the required *incremental approach* for the *sustainability* of IBDS Provider's platform and, therefore, of their business. Figure 6.3 outlines the integration of a local computing node into the existing infrastructure in a monitored manufacturing facility.

The design of local computing nodes abides by the following principles:

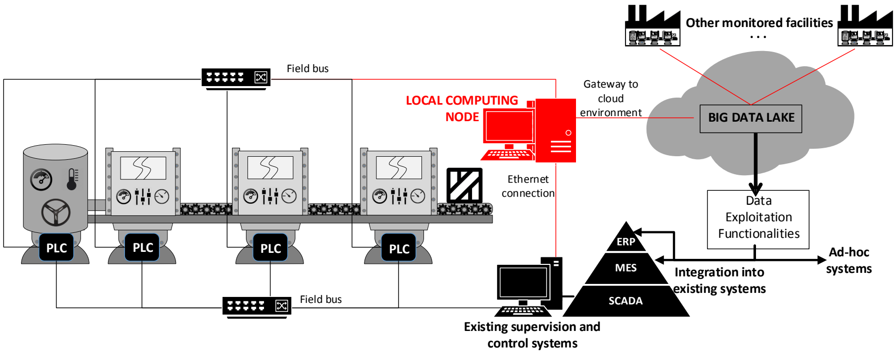

Other monitored facilities
...
Field bus
Gateway to cloud environment
LOCAL COMPUTING NODE
BIG DATA LAKE
Ethernet connection
Data Exploitation Functionalities
ERP
MES
SCADA
Ad-hoc systems
Integration into existing systems
PLC
PLC
PLC
PLC
Field bus
Existing supervision and control systems

Figure 6.3: Schema of the integration of a local computing node into the infrastructure of a manufacturing facility

- Integration with the already operating OT infrastructure in a manufacturing facility, to provide that facility with the required IIoT functionalities so that relevant raw data can be extracted for their later exploitation.
- Flexibility to extract raw data from different industrial components via diverse low-level connections and protocols.
- Assurance of eventual transmission of all captured raw data in scenarios with varying conditions on the quality of connectivity systems.
- Assurance of security in all data transmissions to and from outside the OT infrastructure.
- Assurance of the data supply for a first level of data exploitation service based on real-time visualization of monitored indicators, available via SaaS for any manufacturing facility right after a local computing node is deployed in that facility.
- Capability to upgrade their functionalities without interfering with the normal operation of the monitored manufacturing facility. The periodic upgrade of functionalities not only allows covering the data extraction from more industrial components. As well as that, it allows evolving the data lifecyle stages that are covered by deploying data preprocessing components.

These local computing nodes can be deployed either as a stand-alone device or as a virtual machine installed in an already deployed computer with all required connections. Their internal high-level architecture (outlined in Figure 6.4) is composed of the modules supporting the fulfillment of the aforementioned principles. Those main modules are detailed along this section. Besides, the local computing node also captures internal data from the hardware components, (i.e. CPU, memory, hard disk, internal temperature, etc.) in order to convey them to the cloud computing environment for the monitoring of local nodes correct functioning.

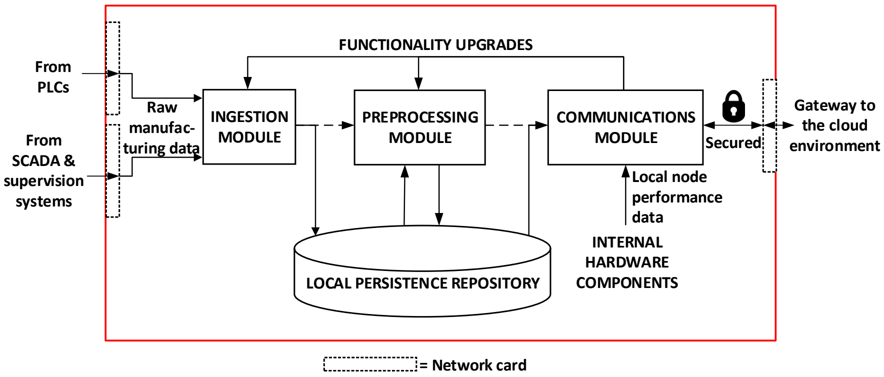

FUNCTIONALITY UPGRADES
From PLCs
From SCADA & supervision systems
Raw manufac-turing data
INGESTION MODULE
PREPROCESSING MODULE
COMMUNICATIONS MODULE
Secured
Gateway to the cloud environment
Local node performance data
INTERNAL HARDWARE COMPONENTS
LOCAL PERSISTENCE REPOSITORY
= Network card

Figure 6.4: High-level internal architecture of a local computing node

6.4.1 Ingestion Module

One of the main goals of the architecture is to ensure the *low-latency capture of raw data*, with which the service of real-time visualization will be immediately available right after a local node is deployed. The *crawler* executed in the Ingestion Module, together with the *Local Persistence Repository* described in the next subsection, contribute to ensure this goal.

The *crawler*, whose internal architecture is outlined in Figure 6.5, continuously executes a crawling algorithm to read raw data from the interconnected industrial components along the monitored manufacturing facility. For that purpose, it makes use of different low-level connection libraries, which act as wrappers to connect to the raw data sources via different low-level protocols and types of network cards. Thanks to this internal structure, each local computing node is prepared to capture raw data either via field bus directly from Programmable Logic Controllers (PLC) across the different phases of the process using standard industrial protocols, or via local network from control and supervision systems (SCADA[1] and others) already deployed in the manufacturing facility. The implementation of the crawling algorithm can leverage existing proposals of open-source tools [LIX14][QLT$^+$15], implementation patterns [SS13] and models [JSS$^+$16] for real-time data ingestion.

Most analyzed manufacturing scenarios generate data at a sampling rate of one sample per second. Still, some analyzed scenarios present more demanding needs and required capturing and transmitting up to 60,000 data samples per second. The low-level connection to the OT infrastructure guarantees a low latency in the data capture. Moreover, the modular capabilities for data capture, together with the periodic upgrade of low-level connection libraries via a remote management process (see 6.5.3), allow covering increasingly more scenarios with different types of internal field buses (Modbus TCP, Ethernet IP, Profibus/Profinet, FINS[2], etc.), OPC protocols and other data exchange func-

[1] SCADA: Supervisory Control and Data Acquisition
[2] FINS: Factory Interface Network Service

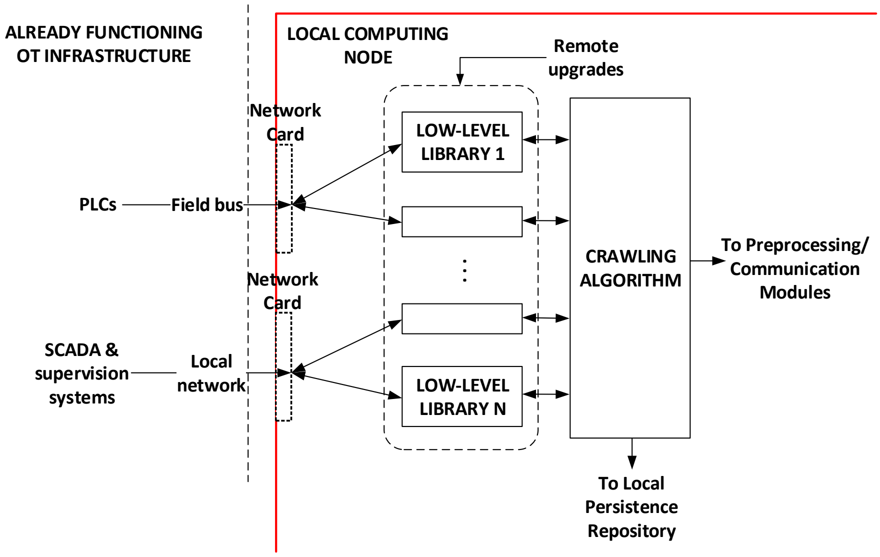


Figure 6.5: High-level internal architecture of the *crawler* in the Ingestion Module

tionalities based on Web Services and IP-based protocols. This approach is motivated by the diversity in types of industrial data sources that can be found in deployment scenarios. For instance, in the field testing conducted in one of the analyzed manufacturing plants regarding data reduction analysis, the more than 300 analyzed indicators corresponded to sensors from very different industrial components: conveyor belts, measuring frames with infrared sensors, tanks, cranes, weighing scales, temperature sensors, purpose-specific equipment exporting raw data via their internal PLCs, etc. This constitutes only a small sample of the various cases to be covered by the Ingestion Module, as it just represents one particular plant from a specific manufacturing sector. The coverage of all monitored manufacturing facilities, together with the flexibility to cover more in many other sectors in the future, imply an even higher degree of heterogeneity in the access to raw data. The *flexibility* provided by the automatic, periodic update of all local devices with a firmware image containing all low-level libraries for raw data access guarantees an adequate coverage of this syntactic *heterogeneity* in the data access features. Moreover, the automatic transference of any new low-level data access library to all deployed devices provides an important *economy of scale* in the development of these data access components.

6.4.2 Local Persistence Repository

The use of a *local persistence repository* has the main goal of providing a preventive storage of captured raw data samples while these are being sent through the *communication module* (see 6.4.4). Taking into account that the already deployed OT infrastructure in most manufacturing environments has not been designed with the goal of efficiently exporting data outside the facility, the preventive storage of captured raw data overcomes potential shortcomings in the deployed Internet connection. This ensures that all captured data are eventually transmitted to the cloud-based layer of the platform. Thus, data transmission will not be affected by networks with *difficult connectivity conditions* such as high latency or *jitter*, i.e. variance in latency over time.

In terms of *hardware* requirements for this preventive storage, the use of solid-state disk drives is highly recommended when setting up local computing nodes, in order to ensure higher operational speed. Regarding *software* requirements, the pattern of data transactions in these scenarios is characterized by the need for managing a very high volume of small data transactions, rather than managing a few batches of big *data chunks* (as is more common in map-reduce-based operations). This requirement, together with the absence of purely relational (e.g. join) operations in this specific task, points at *NoSQL (non-relational) databases* as the data management system solution to implement this local persistence repository. In order to select which specific NoSQL system is best for this purpose, again, the best operational speed should be the criterion to prioritize.

6.4.3 Preprocessing Module

The components to be considered for the preprocessing module are those solving two main types of preprocessing needs: (a) *cleaning* noise and treating missing values, and (b) obtaining a *reduced representation* of captured data via the *reduction solution* obtained as a result of the reduction analysis (see chapter 5). Nevertheless, depending on the technical requirements of the particular application scenario, the architecture provides the flexibility to decide whether to activate this preprocessing module in local computing nodes or not, i.e. postponing preprocessing for a later state once raw data are centralized in the cloud environment. This will depend on the technical requirements in the particular application scenario and on the assessment of preprocessing components' performance. This is related to the assessment of performance dimensions such as reduction and reconstruction compute time, as outlined in 5.3.1. Thus, the decision can be different for each application scenario, depending on the contrast between the operational speed of preprocessing components and the requirements for real-time data exploitation in the specific scenario.

In the case of enabling the local use of this module, when a local computing node is initially deployed, it will begin its operation by transmitting the raw data in the format that is directly gathered by the ingestion module. The contrast of the first visualizations of the captured raw data with the technical and business representatives of the monitored manufacturing facility, as well as the reduction analysis of the captured raw data, will provide the required insights that lead to a progressive *fine tuning* of the preprocessing components. Thus, they will fit the specificities of the data indicators captured in that particular scenario. The preprocessing components will be incrementally activated and upgraded, so that raw data will be *cleaned* and *reduced* by them before being transmitted. The upgrading of the modules in the local computing node and the supported stages in the data lifecycle is outlined in Figure 6.6. The preprocessing will be a continuous process, executed as the ingestion module continuously captures new raw data. This will enable the centralization of a more efficient representation of raw data, which will also be compliant with the functional and non-functional requirements of the application scenario.

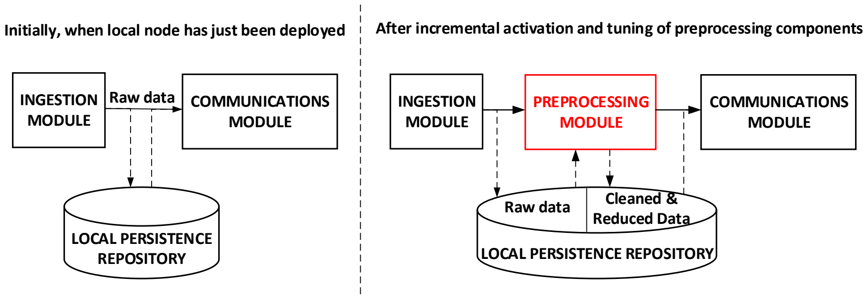

Initially, when local node has just been deployed
INGESTION MODULE
Raw data
COMMUNICATIONS MODULE
LOCAL PERSISTENCE REPOSITORY
After incremental activation and tuning of preprocessing components
INGESTION MODULE
PREPROCESSING MODULE
COMMUNICATIONS MODULE
Raw data
Cleaned & Reduced Data
LOCAL PERSISTENCE REPOSITORY

Figure 6.6: Upgrading of a local computing node with a preprocessing module

6.4.4 Communication Module

This module manages the *data exchange* and communication with the cloud environment and the *security mechanisms* to validate the identification of both ends using encrypted credentials. This ensures that monitored OT infrastructure is not directly connected to the Internet and that the local node acts as an intermediate barrier that protects the points where OT and IT infrastructures converge, thus keeping the OT infrastructure safe from potential security threats over the Internet.

The data exchange with the cloud environment includes:

- Transmission of manufacturing data to be centralized for their later exploitation.
- Transmission of data concerning local node performance (i.e. CPU, RAM, hard drive usage, etc.) to be remotely monitored.
- Reception of updates/upgrades to be deployed in different components of the local node. This includes new *low-level connection libraries* for the *Ingestion Module* and the activation and tuning of the *preprocessing components*.

Each local computing node only stores and processes data corresponding to the manufacturing facility where it is deployed. A local node does not share data with other deployed local nodes, and consequently it does not use its computing power to process data from a different facility.

In order to implement an internal communication protocol using certified credentials for both ends, the specification of the Transport Layer Security (TLS) protocol [DR08] provides a valuable resource. Thus, an *internal handshake protocol* based on TLS must be implemented in order to establish the communication channel as secure. Instead of using a credential validation via certified authorities, each local computing node would store the current *fingerprint* of the server side, i.e. the cloud environment. This fingerprint would be periodically renewed and broadcasted to the local devices in a secure way. Besides, as part of the security management, the required functionalities to renew or revoke *local device credentials* in a remote way must be included.

6.5 Design of the Cloud Computing Environment

The cloud computing environment is composed of *front-end* and *back-end* layers, outlined in Figure 6.7. The *front-end layer* encompasses the functionalities addressing the secured transmission and reception of data to and from the local computing nodes deployed in the monitored manufacturing facilities. The *back-end layer* provides the required functionalities so that smart services can be implemented on the centralized data, as well as the global management of the platform and its functioning.

The back-end layer also includes an important functionality: a *built-in service for real-time and historic visualization* of the time-series data captured in monitored facilities. Every facility where a local computing node is deployed can have immediate access to this visualization service. The inclusion of this horizontal service for monitored facilities in all supplied manufacturing sectors constitutes the SaaS element of the business value proposition of an IBDS Provider and fulfills *two crucial goals* for the success of an IBDS Provider's business:

1. The first goal is the provision of short-term value for manufacturers deploying this solution, which was identified as an important requirement for the immediate return of their investment (see 4.2.3). Indeed, deploying this visualization service in the cloud environment enables an easy remote monitoring of the current functioning of each connected facility. Figure 6.8 provides an example of this remote real-time monitoring.

2. The second goal is the provision of valuable information in order to refine the design of smart services for a particular manufacturing sector. The visualization of raw data can provide insights to design the required solutions to support the successive transformation steps of captured data (e.g. the identification of noise to be filtered out or the presence of missing values to be filled in) and the tuning of the smart services to be provided.

This section details the functionality of the modules for *Data Flow Reception* and *Data Load Balancing*, in the front-end layer, as well as two main components of the back-end layer: the *Big Data Lake* to centralize the data captured from all monitored facilities where the IBDS Provider supplies their services and the *Monitoring and Management tools* to supervise the correct functioning of the platform and update its functionalities.

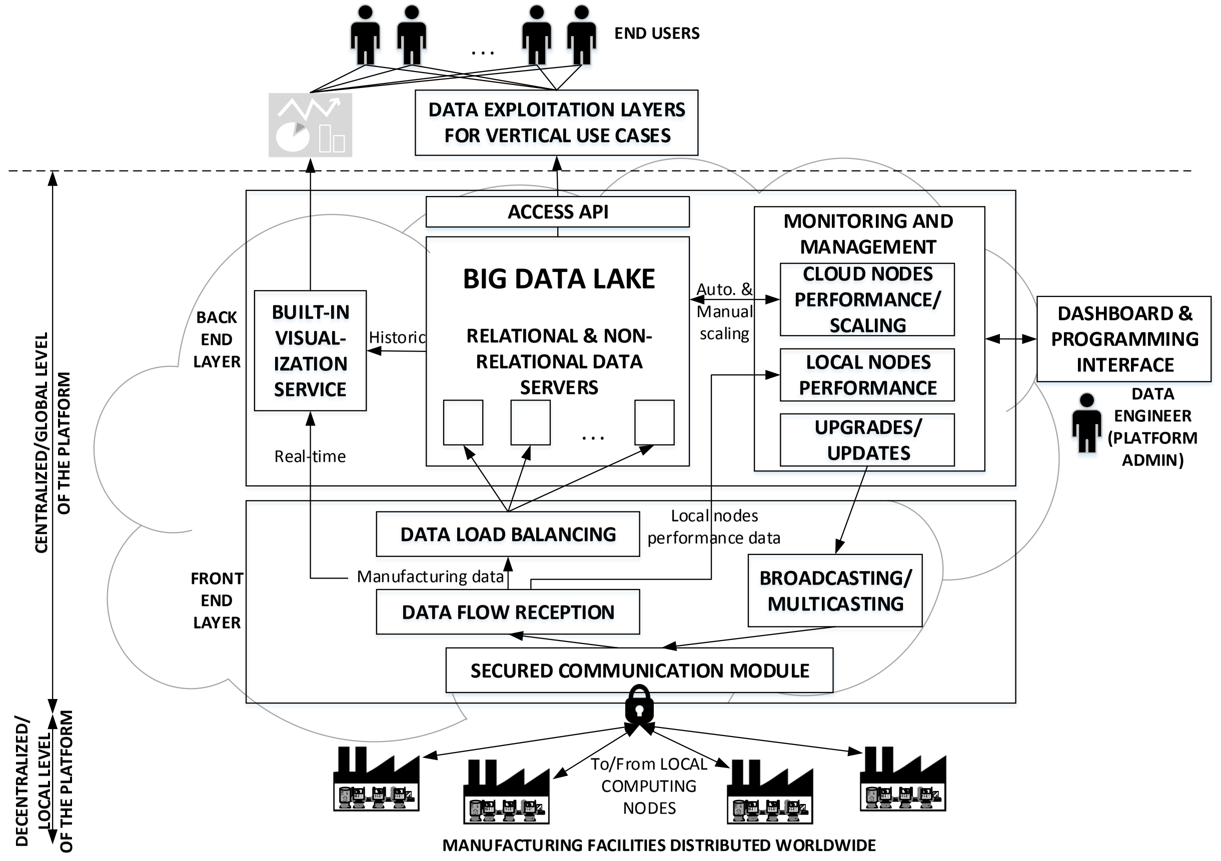

END USERS
...
DATA EXPLOITATION LAYERS FOR VERTICAL USE CASES
ACCESS API
MONITORING AND MANAGEMENT
CLOUD NODES PERFORMANCE/ SCALING
LOCAL NODES PERFORMANCE
UPGRADES/ UPDATES
BIG DATA LAKE
RELATIONAL & NON-RELATIONAL DATA SERVERS
...
Auto. & Manual scaling
DASHBOARD & PROGRAMMING INTERFACE
DATA ENGINEER (PLATFORM ADMIN)
BACK END LAYER
BUILT-IN VISUAL-IZATION SERVICE
Historic
Real-time
FRONT END LAYER
DATA LOAD BALANCING
Local nodes performance data
Manufacturing data
DATA FLOW RECEPTION
BROADCASTING/ MULTICASTING
SECURED COMMUNICATION MODULE
CENTRALIZED/GLOBAL LEVEL OF THE PLATFORM
DECENTRALIZED/ LOCAL LEVEL OF THE PLATFORM
To/From LOCAL COMPUTING NODES
MANUFACTURING FACILITIES DISTRIBUTED WORLDWIDE

Figure 6.7: High-level internal architecture of the cloud computing environment

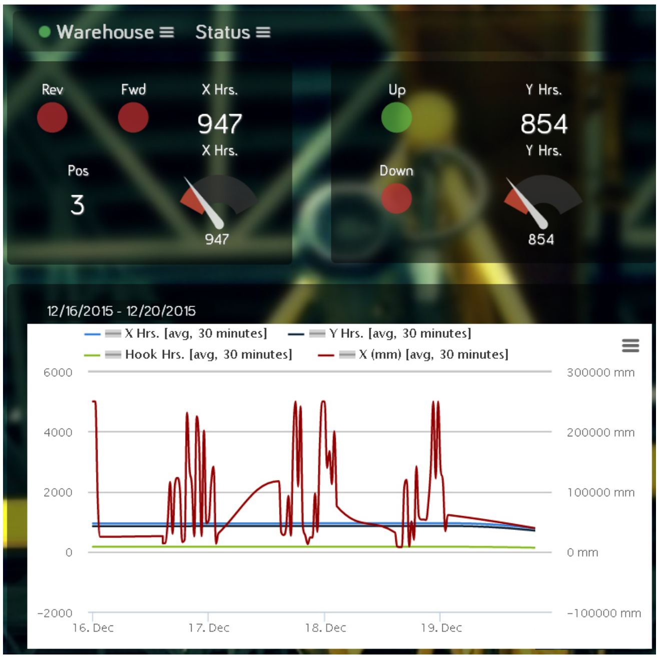


Figure 6.8: Example of the visualization panel for warehouse indicators in one of the analyzed manufacturing facilities

6.5.1 Data Flow Reception and Data Load Balancing

The data reception and transmission from the cloud computing environment share the same security and validation mechanisms included in the local computing nodes. Thus, all data communications are channeled through a *secured communications module* and both ends are identified using encrypted credentials.

There are two types of data received from local computing nodes: the *manufacturing data* captured from the different industrial components in the monitored facilities, and the *local node performance data* related to the monitoring of the local computing nodes themselves. The *Data Flow Reception Module* is in charge of channeling the incoming data flow to the appropriate module, depending on the type of received data. Data related to the monitoring of local computing nodes and their performance is redirected to the platform management modules in the back-end layer. Manufacturing data, which constitute the sustain for the different services to be provided, follow two paths in parallel. On one hand, manufacturing data are supplied to a *Data Load Balancing Module*, which manages the storage of the received data across the cluster of storage server nodes in the back-end constituting the *Big Data Lake*. On the other hand, manufacturing data are also supplied to the back-end module that manages the built-in service that is horizontally available for all supplied customers and manufacturing sectors: the visualization of the raw time-series data for each indicator captured in the monitored facilities. Indeed, the direct streaming supply of captured manufacturing data in parallel to their storage in the Big Data Lake enables their real-time visualization with the minimum possible delay, mimicking the direct supply of incoming data to the streaming layer in the Lambda Architecture [MW15].

6.5.2 Big Data Lake

The centralized accumulation of manufacturing data captured from all monitored facilities in their original raw format or, at most, cleaned and reduced for their later reconstruction prior to their exploitation, constitutes the *Big Data Lake* of the platform. In terms of the synergies of our proposal with the Lambda Architecture paradigm, as outlined in Figure 6.1, it constitutes the master dataset on top of which to design the data exploitation layers according to the vertical use cases (i.e. the intended smart services) for each targeted manufacturing sector. These data exploitation layers are subject to be designed following the abstract layers provided by the Lambda Architecture paradigm.

Given that this Big Data Lake accumulates all data managed by an IBDS Provider and therefore it covers different application scenarios involving different customers, the supply of data from the lake to the exploitation layers must be controlled by an *Access API* that manages the access rights from users and applications to the appropriate data. Such an API must offer a nested view of the centralized data, organized according to different perspectives: the customer company (owner of data), the manufacturing facility where data were produced, and the specific machine or equipment generating those data. This allows fulfilling two goals: on one hand, it enables a fine-grained control of data access depending on their ownership; on the other hand, it offers more flexibility than the SCADA systems deployed in each facility in order to integrate a multi-facility view for the same customer and to build personalized applications exploiting those integrated data. For instance, the consumption of data by the built-in visualization service follows the same principles of the access API, controlling the appropriate access depending on specific users and their rights on data.

Moreover, a combination of relational and non-relational technologies can be considered in the implementation of the lake. This combination allows answering different demands in the transformation and mining of captured data, depending on the different smart services to be implemented in the supplied scenarios.

6.5.3 Monitoring and Management

The cloud-based level must also provide the management tools to be used by the data engineer administering the platform. These tools are directly related to three processes that are crucial for the use of the platform as the cornerstone of an IBDS Provider's business and their provision of services: the *deployment* of the platform for a particular scenario/facility, the *monitoring* of its correct performance, and the *update/upgrade* of the functionalities in local computing nodes deployed worldwide.

Management tools include the functionalities to convey configuration instructions during the deployment of local nodes, the supervision of their correct deployment and the subsequent monitoring of an adequate performance (see Figure 6.9 for an example of performance monitoring). Besides, they include the functionalities to deliver different *updates and upgrades* to the firmware image used in some or all local nodes, guaranteeing that every node is running the latest and most appropriate version of its functionalities for the scenario where it is deployed. This progressive upgrading of local node functionalities enables a more effective capture of raw data from industrial components via the required protocols, as well as a more efficient data preprocessing using the appropriate techniques.

Moreover, the performance of the cloud nodes with the pool of data servers constituting the Big Data Lake can also be monitored, and the management tools include the functionalities to supervise the automatic *scaling* of the required computing resources and also to manually scale them, depending on the required performance conditions.

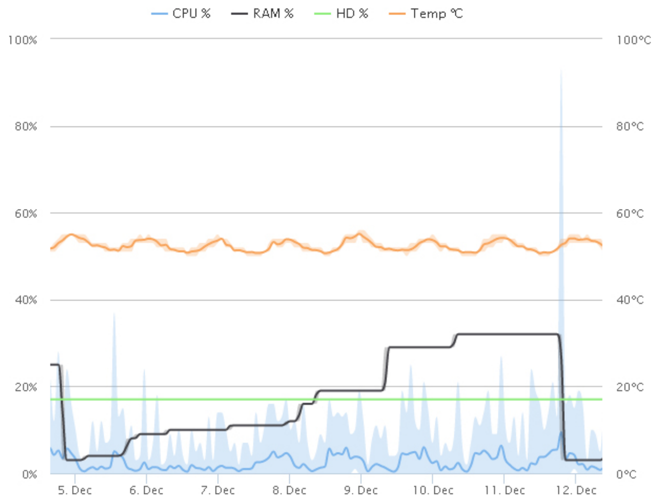


Figure 6.9: Example of performance monitoring for a local computing node

6.6 Conclusions

The presented design artifact models a proposal for the architecture of Industrial Big Data platforms that can sustain the business of IBDS Providers, effectively combining IIoT and Cloud Computing components and providing an efficient answer to the volume, velocity and variety of data found in real-world manufacturing business settings. The main differential contribution of the proposed design is that the architecture is not conceived as a solution to migrate the whole industrial infrastructure of those settings demanding a shift towards Smart Manufacturing. Instead, it is conceived as a solution that support the business of an IBDS Provider (be it an independent IT-based company or a specialized unit of a large manufacturing organization) who wants to supply services that facilitate that shift to others with a non-intrusive, integrative approach with respect to already running OT infrastructures. Moreover, the proposed design facilitates the sustainability and scalability of the *business value proposition* of IBDS Providers.

The presented DHA goes one step further than most of the conceptual proposals related to Smart Manufacturing architectures, including the required components that fulfill the main non-functional requirements derived from the scenarios where an IBDS Provider develops their business. It also complements existing popular paradigms for Big Data systems such as the Lambda Architecture, by describing the architectural components that save the gap between an initial state where no data are extracted yet from manufacturing facilities and the eventual availability of a centralized data repository on top of which different exploitation functionalities can be designed according to the layers in the Lambda Architecture. Furthermore, the presented DHA provides a more flexible approach than those proposals focused on specific use cases and sectors, enabling the progressive upgrade of its modules to increasingly cover more application scenarios and more data transformation stages.

This proposal of a DHA constitutes a valuable complement for the conceptual frameworks proposed to deploy Big Data systems in Smart Manufacturing scenarios. In this regard, the key elements of the DHA provide additional guidelines when implementing a solution based on one of those conceptual frameworks for Smart Manufacturing scenarios. At the same time, the DHA puts the spotlight on business-oriented, practical aspects derived from a hands-on experience with real-world manufacturing business settings. Thus, those aspects could be taken into account when devising future versions or extensions of these frameworks.

Chapter 7

Business Stakeholders-driven Characterization of Data Exploitation Requirements for Smart Services

In the Smart Manufacturing scenarios analyzed in this research work, the supply of *smart services* for a particular manufacturing sector is based on the partnership between IBDS Providers and manufacturing agents specialized in the targeted sector. The value proposition of an IBDS Provider is based, on one hand, on their *horizontal* Industrial Big Data platform to capture, integrate and visualize relevant data from the monitored facilities and, on the other hand, on the collaborative design of smart services together with their manufacturing partners. This provides an interesting business context for an IBDS Provider, given the possibility of multiple deployments for various engaged customers in the same sector. Indeed, in Smart Manufacturing servitization scenarios, the partnership with a servitized CEM opens the possibility to access a market of multiple interested manufacturers, i.e. that CEM's customers. This multiplicity may also be present in scenarios where an IBDS Provider collaborates directly with a smartized manufacturer and this manufacturer also aims at expanding their business by offering services to other agents in their sector. In any of those types of scenarios, the design of smart services is sustained by the development of *smartization projects* for those manufacturers that want to shift the operation of their businesses towards a more Smart Manufacturing-oriented approach.

Such a design process based on the development of smartization projects with different customers has a notable parallelism with the design of a new business model or new services to be offered to a market. In that sense, it is crucial to characterize that market and their data-related needs. Therefore, relevant knowledge has to be elicited from business stakeholders, so that it is built a deep understanding of the business problem, the data exploitation needs, the relevant processes that can potentially leverage the outcome of data exploitation and the interfaces among those processes, information upon which to make business decisions, influential variables, etc.

Moreover, the elicited knowledge should be directly linked to the suitable data capture and processing step where that knowledge could be used as input to better plan and manage the technological support for that step. This is where the IBDS Provider's data-related technological know-how can be more valuable, in order to *translate* the business requirements to data-related tasks. The definition of the required data-related tasks has a clear parallelism with the methodological support that KDDM process models provide, so that the appropriate KDDM process is executed to capture and preprocess relevant variables, create the required analytic models that extract value from data and integrate these models into the operation of manufacturing systems.

Nevertheless, it has to be noted that the design of such a KDDM process entails a complex problem from the point of view of stakeholders analysis. This complexity is due to the business context in which smart services are built. For instance, a servitized CEM aims at leveraging data exploitation not as a means for internal optimization, but as the core of a brand new value-added service for their customers. Therefore, the IBDS Provider should not follow a KDDM process for an *ad hoc*, one-time project for a particular organization, given that the CEM is not the only organization to characterize. Instead, it must fit the design of new data-driven services for a market of potentially many different companies, i.e. the customers to whom the CEM has been supplying their equipment so far. Therefore, in order to build the right smart services, the smartization projects must capture and characterize the data exploitation needs from these customers and the different levels of stakeholders in their respective organizations. The design of the elicitation process and the interaction with stakeholders must take into account this multi-view scenario. Indeed, all business stakeholders in this complex map will have very different informational needs and will provide very different requirements. These requirements should all be taken into account when planning the KDDM process, and integrated when managing it.

The presented context demands a more flexible approach in order to capture knowledge from the complex map of relevant business stakeholders to whom the smart services are aimed at. Indeed, the characterization of data exploitation needs and the design and deployment of the appropriate services should follow an *incremental approach* where the scope is progressively refined and widened. This approach may begin with developing reduced-scale pilot projects with a limited number of customers and monitored facilities, so that the initial design and deployment of smart services is improved with the insights of these pilot projects and with the increased learning from successive projects with new business stakeholders.

This chapter presents our contribution to (a) extend KDDM process models with an incremental approach for the integration of *business understanding*

[She00] into the data lifecycle to be covered, and (b) conduct the interaction with business stakeholders in order to elicit and characterize data exploitation requirements, so that these requirements can be leveraged as input for the relevant data lifecycle steps. These contributions are aimed at the project manager role supplied by the IBDS Provider in these smartization projects.

The contributions are sustained by the identification on shortcomings in *KDDM process models* and *requirements elicitation in data-related projects*, and the integration of knowledge from relevant research areas such as *interview analysis*, *stakeholders management* and *business model design* to overcome those shortcomings. In order to validate the necessity of these new contributions to support smartization projects in the targeted scenarios, we built a validation-oriented version of the process to conduct elicitation interviews and the template to capture business requirements and their impact into the technological support for data lifecycle steps. These components were contrasted in a field validation in the real-world manufacturing business setting where we conducted our case study. After the validation of their utility in this relevant instance of the targeted scenarios, we refined our proposal in order to contribute with design artifacts modeling a *spiral process model* to conduct the business stakeholders-driven characterization of smart services and the *template* to capture and characterize the connection between business requirements and their impact into relevant KDDM process steps.

7.1 Analysis of Related Work

In order to achieve a good understanding of relevant work related to the characterization of requirements in data-related projects, two main knowledge areas where initially analyzed: *requirements elicitation* (as part of requirements engineering) and *KDDM process models.*

Requirements Engineering (RE) is a crucial stage in software design and development, concerned with the identification of goals and constraints for a *system* and the assignment of responsibilities for the resulting requirements [AW05]. The system context provides the basic conditions for RE, in the form of different facets belonging to the *business perspective* or the *technical perspective* of the information system to be developed [SWW11]. That is, RE has to *translate* solution-independent target requirements “written in the language of the stakeholders” to solution-oriented technical design requirements “composed in the language of the developers” [BLK11]. In our case, we are mainly interested in *Requirements Elicitation* -a core activity in a RE process [ZC05][SWW11]-, interacting with key stakeholders from the business side of the problem. This interaction is crucial in order to elicit all requirements emanating from customer needs and their value creation processes [BLK11]. There exist different approaches to guide a RE process in data-driven projects, mainly with a goal-oriented focus, including a comparative study [CLSM$^+$14] identifying the techniques used for elicitation, specification and validation of requirements in several approaches [CLMT13][GRG08][PG08]. However, the scenarios they discuss are mainly centered on representing captured requirements as an interrelated hierarchy of informational goals to be solved by data warehouses.

Focusing on Requirements Elicitation and the set of techniques used in it, conducting interviews is arguably the most common one [CLSM$^+$14]. It has to be taken into account, nevertheless, that it is a very resource-demanding method [HA05]. Besides, the participation of top-level managers from stakeholder organizations -characterized by their lack of availability- is crucial to elicit valuable requirements, which adds complexity to the scheduling of interviews. Therefore, in order to make these interviews more efficient by reducing their cost and effort, it would help to leverage a guide on how to arrange and manage them based on practical experience conducting elicitation processes with stakeholders in a real business scenario. However, software engineering researchers reporting studies in which interviews have been used to collect requirements often fail to describe how they were conducted [HA05]. As the lack of effective communication between the research/developing side and the business side is often cited as an obstacle for proper RE [CLSM$^+$14], providing details on how to conduct this process could be a valuable resource for researchers and developers.

Regarding the major KDDM model proposals, they do include a first step covering application domain understanding and business requirements identification, which are later converted into data mining goals [KM06]. Nevertheless, these proposals mostly approach the problem focusing on a single data mining problem type or a single application of an analytical model, leading to a limited identification of relevant stakeholders and the potential interaction among their respective informational needs. There exists a model proposal with four generic user roles for a data mining scenario [XJW$^+$14], but this model treats a person and their organization as the same role, so it does not provide an appropriate characterization for complex scenarios with heterogeneous data exploitation needs. There also exists a proposal for a multi-view KDDM process [ZBO$^+$14] but, apart from not presenting any case study, it does not detail a requirements elicitation process or a well-defined model for those business stakeholders from whom to elicit requirements. Indeed, the different data exploitation needs by all identified stakeholders in our analyzed scenarios leads to a multi-view requirements elicitation [SWW11].

We can make a further distinction among the proposals for KDDM process models analyzed in [KM06]. While the foundational schema for a KDDM process [FPSS96] is tightly linked to the lifecycle that data go through in order to create an analytical model, a proposal like CRISP-DM [CCK$^+$00][She00] incorporates further details on other relevant aspects that have to be managed from the perspective of a fully fledged project to build an analytical model in an organizational context. Indeed, CRISP-DM includes the concept of the *business perspective* of a data-related project and the vision of the organization that wants to leverage the analytical model according to some business goals. Nevertheless, the approach proposed by CRISP-DM is focused on the provision of analytical models as a result of an internal project for an organization, and not as the development of a new service aimed at a market of various potential customers with different levels of stakeholders. Besides, it lacks a more detailed reference on how the knowledge obtained in the *business understanding* phase can be used as input for the subsequent design of the different data lifecycle stages.

Thus, there are two major shortcomings in the analyzed knowledge areas in order to be leveraged in our targeted business context. On one hand, there is a need for an *incremental approach* to capture and characterize requirements from such a multi-view scenario, where different business stakeholders are progressively engaged in smartization projects and their data exploitation needs allows refining the design of the required smart services. On the other hand, the map of business stakeholders involved in these scenarios leads to a complex interaction that requires guidelines on how to conduct elicitation interviews, capture business requirements and characterize their *translation* as technical, KDDM-oriented requirements. In order to overcome these shortcomings, proposals from other knowledge areas such as *stakeholders management* and *business model design* were identified to be integrated in the design of contributions.

While we can consider CRISP-DM a reference model for KDDM-oriented project management, general reference models for project management [Pro13] also provide valuable resources linked to the analyzed problem, as they include guidelines for stakeholder management. Indeed, stakeholder analysis is considered a crucial front-end step for knowledge elicitation [Pou97]. An important first task [Pou97][Pro13] to consider in this regard is to identify relevant stakeholders. It is proposed to address stakeholder identification as an iterative process, in which the knowledge extracted from initial stakeholders guide the subsequent steps with new stakeholders, thus leading to a continuous development and refinement of the expression of user needs and the knowledge representation of the analyzed domain [Pou97]. This fits our targeted scenarios, as not only business stakeholders in a given organization point to other relevant stakeholders to consider, but also the successive addressing of new organizations (i.e. new potential customers for the smart services) provides new stakeholders to be considered in the analysis. Moreover, the incremental nature of this process also has clear synergies with the proposal of spiral models [Boe88] for software development and enhancement.

The specific area of smart services for manufacturing companies has been analyzed in order to propose guidelines for the development of such smart services. In particular, the reference framework presented in [MSA15] describes a process-activity model for the development of smart services, highlighting a relevant requirements analysis phase prior to the service design. It also lists some relevant tools to be used in that requirements analysis phase, such as interviews, workshops, requirements list, etc. However, it does not provide a detailed link between the results of the requirements analysis phase and the input for the subsequent phases of service design, test, implementation and launch, or a description of how the listed methods and tools should be used while extracting relevant knowledge from business stakeholders during requirements analysis. Moreover, the proposed phases follow a linear schema, without capturing the incremental nature inherent to the engagement of relevant stakeholders in new service development processes. Indeed, it is important to design a requirements analysis proposal that considers interacting with customers along the entire process of smart service development, so that requirements are not only elicited but also validated and verified in cooperation with customers [BLK11].

On a related knowledge area, the design of new services and the exploration of new business models has been boosted by several interconnected proposals

[BD12][OP10][Rie11] strongly based on the following key principles: the contrast of value proposals with direct feedback from the market since the early stages of service design and conceptualization, the use of interviews with relevant business stakeholders for a first-hand discovery of needs and requirements and a constant contrast of the proposals, an incremental approach where the scope of the proposals is iteratively refined thanks to the feedback and learning from prior contrasts, the use of pilot projects as an actionable tool for the extraction of validated learning, and the use of predefined templates to capture and successively refine the knowledge and requirements captured in these interactions and contrast with relevant stakeholders. Given the strong synergies with some of the key characteristics of the scenarios where the intended smart services have to be designed, all these elements constitute useful resources to be leveraged in the design of our contribution.

7.2 Hypothesis Formulation and Design of Validation-Oriented Artifacts

Given the identified shortcomings in existing proposals related to KDDM process models and requirements elicitation, and the potentially valuable contributions to be leveraged from other related knowledge areas, we formulated the following *three hypotheses*:

1. In order to ensure an effective requirement elicitation in the smartization projects conducted in these scenarios, it is necessary to *integrate new components in order to extend current KDDM proposals* and to provide additional tools for the management of an elicitation process with business stakeholders.

2. A predefined *template* that facilitates the progressive capture of business requirements and the characterization of their impact in the different data lifecycle steps constitutes a valuable resource to be leveraged in smartization projects in order to overcome the shortcomings in current proposals.

3. A *process to manage elicitation interviews* with relevant business stakeholders, as the core element of an incremental approach to progressively refine the design of smart services, constitutes a valuable resource to be leveraged in smartization projects in order to overcome the shortcomings in current proposals.

In order to validate these hypotheses, a method with *two main steps* was followed. First, we built a validation-oriented version of the two components whose suitability and applicability we wanted to validate, i.e. the *template* supporting the characterization of relevant business and technical requirements and the *process* to manage elicitation interviews with business stakeholders. Then, we conducted a field validation in the business setting of our case study, where we contrasted the applicability of those two components in order to validate out hypotheses. In this section we describe the first step of the followed method, i.e. the creation of validation-oriented versions of the previously mentioned two components.

7.2.1 Capture of Requirements during the Elicitation Process

Requirements must not only be extracted in terms of a business perspective, but must also be expressed as solution-oriented technical design requirements [BLK11]. In the case of smart services, the technical aspects are linked to the data lifecycle and the stages depicted in KDDM process models. Therefore, the tools proposed to support the elicitation process must not only cover the gathering of business requirements, but also their *translation* in terms of technical input for the different KDDM process steps. Thus, in order to identify which relevant information to capture during an elicitation interview, two levels of information items (detailed in Table 7.1) were defined to characterize a data analytics need: *Business perspective* and *KDDM perspective* [SWW11].

1) Business perspective [CCK$^+$00][She00]: These elements characterize a *use scenario* for data analytics, using concepts (stakeholder-process-indicators) that could guide more easily the interaction with business stakeholders.

2) KDDM perspective: They are the translation of the business perspective elicited about a data analytics need, in terms of the information that can be extracted from those items and used as input in the different KDDM process steps. The foundational schema for a KDDM process [FPSS96] was used to establish KDDM phases, with two slight adaptations. First, the selection step was extended to a capture step (extraction of raw data from the manufacturing process). Second, the evaluation step was expanded to a deployment and use step [CCK$^+$00][She00], as the analytical model has to be integrated into existing systems and exploited as part of the analyzed process.

When designing the information items composing the KDDM perspective, we took advantage of ideas and concepts coming from different proposals. For instance, the schema of KDDM phases to consider was based on the foundational one [FPSS96], with slight variations inspired by practical considerations in CRISP-DM [CCK$^+$00][She00] concerning the characterization of both business and KDDM perspectives and the deployment and use of analytical models. Another inspirational reference was the Lambda Architecture [MW15], where the ideas of capturing a massive raw data repository and creating different data views from it for different purposes inspired the information items for the capture and transformation phases.

The separation in two perspectives allows focusing the interactions with business stakeholders on the business perspective, facilitating a more effective communication during the elicitation process. This requires that the *interviewer* has skills for effective business communication, as well as a detailed understanding of a KDDM process in order to establish a clear relationship between business requirements into data mining goals and to *translate* elicited knowledge into items

Business perspective (use scenario)	**KDDM perspective (link to KDDM phases)**
⇒ Specific *stakeholder* in the organizational architecture demanding data analytics ⇒ Which *processes* they are accountable for ⇒ Which information or *Key Performance Indicators* (KPI) they want to monitor when supervising those processes ⇒ In which *use context* (e.g. temporal or operational restrictions) they need that information	⇒ *Capture*: Check if all indicators relevant for the use scenario are already being captured with the existing infrastructure and, if not, plan necessary actions to do so. Characterize all relevant components composing the implementation of the manufacturing process in each particular plant, and ensure that the architecture is prepared to extract raw data from those specific components. ⇒ *Preprocessing*: Analyze visually how raw data are being captured for all considered indicators and identify necessary techniques for data cleaning (noise, missing values) and reduction. Evaluate which particular preprocessing techniques are more efficient with each raw indicator or data source. ⇒ *Transformation (create data views)*: Define the required data transformation and integration (with potentially additional external sources) to create the data view needed in that use scenario; identify constraints on schema. Create a unified/federated schema that can integrate data from different plants, taking into account their different implementations and the possible differences between their data schemas. ⇒ *Mining (create analytical models)*: Identify the required approach (descriptive, predictive, prescriptive), the temporal constraints (whether it has to be built on real time or it can be built on batch) and additional constraints on required tools and algorithms (e.g. depending on the type of outcome variable). ⇒ *Deployment and use (integration with existing systems and processes)*: Define aspects such as whether the analytical model must provide real-time support or it will be sporadically used, whether its results must be deployed automatically or used as a support for decision-making processes, or whether the model must evolve and update itself autonomously or improved versions will be released periodically.

Table 7.1: Information items composing the *Business Perspective* and the *KDDM Perspective*, to characterize a data analytics need

for the KDDM perspective.

7.2.2 Process to Organize and Conduct Elicitation Interviews

So as to capture the potential use of data analytics in a given manufacturing sector, it is necessary to interact with the main business stakeholders in one of these scenarios (see Figure 4.6 in chapter 4) and to elicit the requirements for the smart services to be marketed in that manufacturing sector. Our goal in this case was to provide guidelines on how to conduct these interviews, something often cited as missing in requirements elicitation literature [CLSM$^+$14][HA05]. Thus, a common approach would be replicated when interacting and extracting knowledge and requirements from different stakeholders. In order to design such interviewing approach, studies from social sciences on qualitative methods of interview analysis [BM09][Tri09] and modern proposals that use interviewing as the core technique for new business model exploration [BD12] were leveraged. These proposals provided valuable guidelines for the kind of interviews (market-oriented interactions to elicit knowledge from managers) to be conducted in this field testing.

In order to organize and conduct elicitation interviews with these business stakeholders, we designed the process outlined in Figure 7.1, which is described next:

1. The starting point is the template with the two groups of information items described in Table 7.1 to characterize data analytics needs.

2. When a business stakeholder is selected to be interviewed during the elicitation process, the interviewer sends beforehand a document describing the goals and mechanics for the interview. They are described in terms of the business perspective, which will be the central point of the elicitation during the interview. In this previous communication it is also explained that the interview will not be conducted as a survey or checklist to fill in, but an exploration around the key elements in the document (i.e. semi-structured interview). This serves as context for the business stakeholder, so that they can prepare better the interview.

3. The interview is conducted as an exploration, where the business stakeholder has freedom to explain their view on the requested information. The information items in the business perspective guide the conversation. The interviewer has the necessary questions in mind to put to the exploration and to further characterize these items, so that they can be later linked to the KDDM phases.

4. Depending on how many combinations of use scenarios (stakeholder-process-indicators) are discussed during the interview, different data analytics needs are characterized. With this information the template for both perspectives (business and KDDM) is filled in.

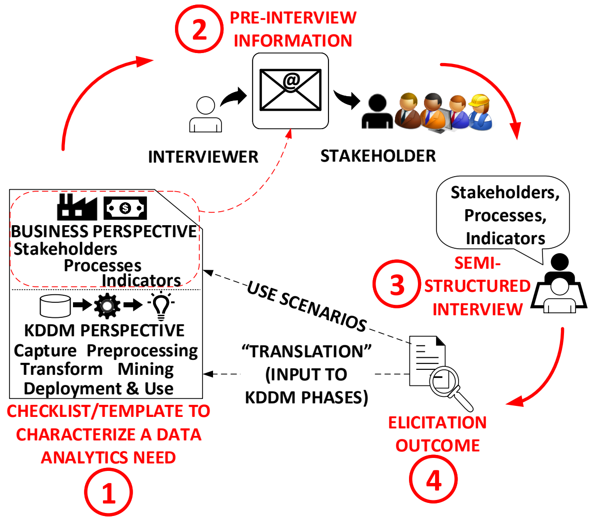


Figure 7.1: Process to organize and conduct an elicitation interview

The successive execution of elicitation interviews following this procedure would constitute the core of an incremental approach, so that the initial design and deployment of smart services is progressively refined thanks to the new insights obtained as new potential customers and their stakeholders are engaged in this elicitation process.

7.3 Field Validation in the Real-World Business Setting of our Case Study

Following the research method established to validate the posed hypotheses, we conducted a field validation in the real-world business setting of our case study, focusing on the case of the servitized CEM designing smart services for the manufacturing sector of polyurethane foam production. We leveraged the opportunity to observe the first steps of a smartization project for the smartized manufacturer presented in section 4.1, i.e. a polyurethane foam production company, and the integration process of a brand new facility to be monitored, owned by this company.

During this research process, we collaborated with the representatives of the IBDS Provider conducting several interviews with business stakeholders from the servitized CEM and the servitized manufacturer, with whom the CEM had an agreement to deploy a pilot project to develop smart services. Besides, the first-hand observation of this manufacturing facility (aligned with the "get out of the building" motto from the Customer Development model [BD12]) provided additional insights to better understand the physical environment and the production process around which smart services should be built. Moreover, it provided access

Company profile	Role profile	Num. of interviews
Servitized CEM	General Manager	6
Servitized CEM	President	2
Polyurethane foam production company (customer 1)	Plant Manager	2
Polyurethane foam production company (customer 1)	Subprocess Technician (mechanical transformation)	1
Capital equipment provider for a different process phase 1	Technical Manager	1
Capital equipment provider for a different process phase 2	Technical Manager	1

Table 7.2: Interviews with business stakeholders during the validation process

to relevant stakeholders in their own working environment, as well as to technical managers from equipment providers for other steps of the manufacturing process.

In order to conduct these interviews we used the validation-oriented components presented in the previous section. Thus, the validation of the applicability of these components and the learning from the field testing would ground our contribution with design artifacts integrating these components.

7.3.1 Outcome of the Interviewing Process

The analyzed interviewing process span through 14 months, during which we collaborated with the IBDS Provider in several interviews with selected representatives, one at a time, from the relevant business stakeholders. The number of analyzed interviews along this period of time is summarized in Table 7.2, detailing the specific stakeholders who were interviewed.

Top-level management staff from the servitized CEM was established as the main source to characterize the manufacturing sector and its requirements. Their General Manager, apart from the business vision, provided a solid technical and engineering understanding of the manufacturing processes involved and the relevant variables to be taken into account. Besides, their President has long experience in this chemical manufacturing sector, as well as a solid economic background on the financial management of this type of manufacturing companies. Therefore, we conducted a first series of interactions with the top management of the servitized CEM to map relevant business stakeholders in client organizations into the general business stakeholders of these servitization scenarios (see Table 4.2 in chapter 4). Thus, we obtained a first list or relevant stakeholders in targeted companies, i.e. polyurethane foam production companies:

- At manager level: *foam production company owner*, *foam production plant* *manager.*

- At technical level: *chemical transformation process technician*, *mechanical transformation process technician.*
- Other equipment providers: *chemical transformation equipment provider*, *mechanical transformation equipment provider.*

The interviews with the servitized CEM clearly benefited from their vested interest in the successful design of the smart services. Although the limited availability of top-level managers was still an issue, it was easier to arrange meetings with them than with the rest of business stakeholders. Besides, there was a need for conducting more interviews with them, as there are different kinds of key knowledge to extract from those interactions:

- First and foremost, they provided a complete insight about the global business scenario they are operating in, their servitization strategy and how smart services could help achieve it.
- They also provided the vision of the owners of the manufacturing companies they work with. While it is highly difficult to access these owners, the direct access that this CEM has had to them along the years provided the required insights on these companies' business goals and market demands, and how they are related to data analytics needs.
- The interviews with the CEM's representatives also provided key knowledge to help preparing the future interviews with customers (the manager of the plant analyzed as a pilot case), in order to have a set of elements to contrast and validate with them. These elements were essential to guide those interviews in a more efficient way.

While the interviews with CEM's representatives were more abundant and initially more exploratory, the meetings with the rest of business stakeholders were more *straight to the point*, as they didn't have the same predisposition and availability (at least not until the pilot project's outcome would start providing them with real value). For this reason, these meetings benefited from preparing detailed information to send in advance, to set the right context for the interview. This information prepared and sent in advance also homogenized the focus among interviews with potential customers and thus helped capturing insights that could be more easily grouped, compared and synthesized.

The outcome of the conducted interviews was organized in two main deliverables:

1. A general context for the smart services to be developed in order to address the identified informational needs in this chemical manufacturing sector. This general context provided global guidelines of the CEM's strategy towards their market and a prioritization of data analytics needs from all user profiles.

2. The characterization of data analytics needs for various business stakeholders. This was documented using the characterization items described in Table 7.1.

Regarding the general context for the smart services, the size of targeted customer companies facilitates the CEM a direct access to the *foam production company owner*. The CEM prioritizes providing value to this stakeholder. Therefore, owners' informational needs have to be solved first and foremost. All performance indicators and informational needs demanded by other stakeholders have to be dependent on the owner's ones. For instance, the personnel supervising specific phases of the production system might look for local efficiency in the subprocess they are accountable for, but the actions to achieve those partial goals could be detrimental to the global efficiency goals for the plant or the company.

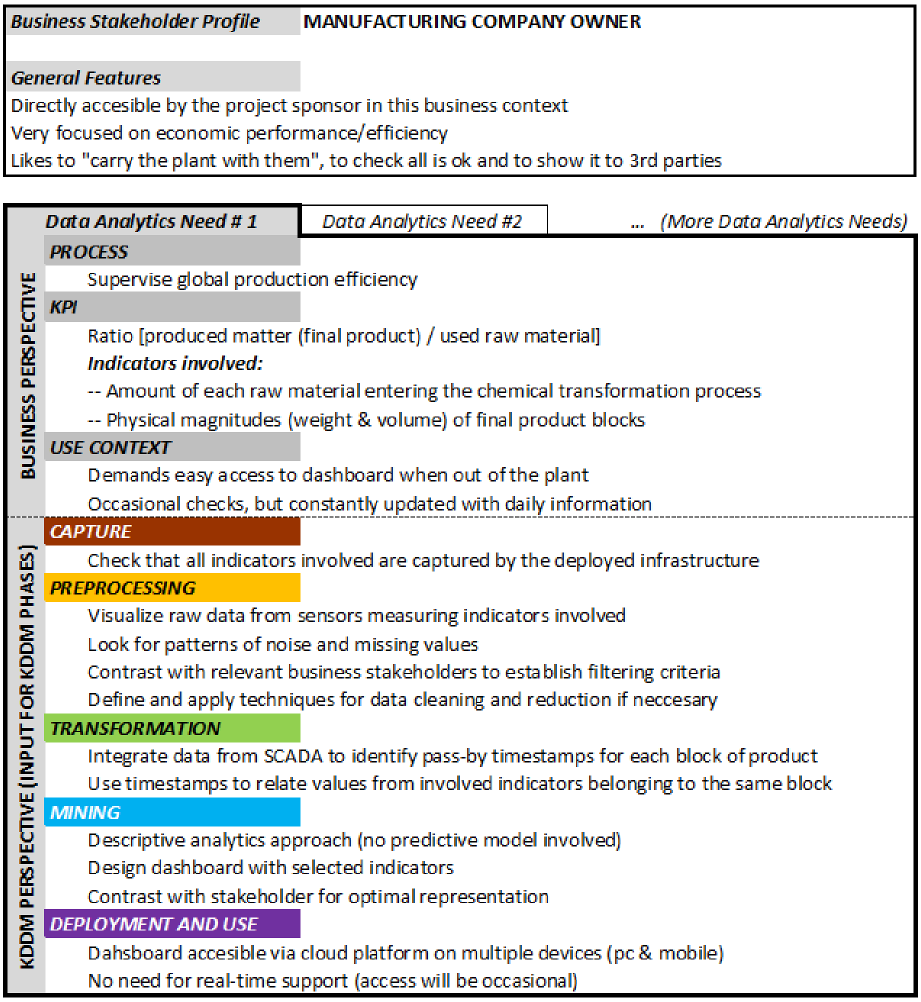
Thanks to their direct contact with *foam production company owners*, the CEM's representatives supplied insights on the most important areas where to provide value in this scenario:

- *Global production efficiency*: Ratio between produced matter (final product) and used raw material.
- *Global financial efficiency*: Ratio between earnings from sold products and costs to produce and sell them.
- *Quantitative vision of provided value*: The smart services must also provide the owner with information on the estimated savings (reduced waste, optimized efficiency) due to the application of the different data analytics outcomes on all subprocesses in the different plants using the system.

The role of *chemical transformation process technicians* illustrates a case where data analytics needs have highly different features. They manage the *set* values for diverse parameters (amount of raw materials, temperatures, some mechanical elements) of the equipment executing the chemical transformation. Besides, the equipment also controls via internal sensors the *actual* values of these magnitudes, which may differ from the set values. The task of a process technician is to tune the set values so that the actual values are the desired ones for the type of product to produce. In this regard, smart services can provide real-time recommendations for the process technician on the best possible tuning for these settings (i.e. prescriptive analytics), according to an evaluation of the whole process (not only this phase) and the expected global efficiency given the current actual values.

All data analytics needs were documented using a template with the items described in Table 7.1. The characterization for the *foam production company owner* is presented in Figure 7.2.

The interviews with technicians from other equipment providers (those companies providing equipment for the other steps in the manufacturing process) followed a slightly different structure, as they were focused on understanding the technical details of the data export capabilities of their equipment and ensuring



Business Stakeholder Profile **MANUFACTURING COMPANY OWNER**

General Features

Directly accesible by the project sponsor in this business context

Very focused on economic performance/efficiency

Likes to "carry the plant with them", to check all is ok and to show it to 3rd parties

Data Analytics Need # 1 | *Data Analytics Need #2* | ... *(More Data Analytics Needs)*

BUSINESS PERSPECTIVE

PROCESS

Supervise global production efficiency

KPI

Ratio [produced matter (final product) / used raw material]

Indicators involved:

-- Amount of each raw material entering the chemical transformation process

-- Physical magnitudes (weight & volume) of final product blocks

USE CONTEXT

Demands easy access to dashboard when out of the plant

Occasional checks, but constantly updated with daily information

KDDM PERSPECTIVE (INPUT FOR KDDM PHASES)

CAPTURE

Check that all indicators involved are captured by the deployed infrastructure

PREPROCESSING

Visualize raw data from sensors measuring indicators involved

Look for patterns of noise and missing values

Contrast with relevant business stakeholders to establish filtering criteria

Define and apply techniques for data cleaning and reduction if neccesary

TRANSFORMATION

Integrate data from SCADA to identify pass-by timestamps for each block of product

Use timestamps to relate values from involved indicators belonging to the same block

MINING

Descriptive analytics approach (no predictive model involved)

Design dashboard with selected indicators

Contrast with stakeholder for optimal representation

DEPLOYMENT AND USE

Dahsboard accesible via cloud platform on multiple devices (pc & mobile)

No need for real-time support (access will be occasional)

Figure 7.2: Example of characterization of a data analytics need for the *foam production company owner*

the connection to the raw data sources that have been identified as relevant for the use scenarios characterized during previous interviews. Thus, their outcome was not a characterization of business requirements that needed to be translated into technical requirements, but a refinement and enhancement of the *technical requirements* of the initial steps in the data lifecycle, mainly focused on the capture step in order to ensure that the required connection to the raw data from the involved equipment was available in the local level of the IBDS Provider's platform. In this sense, the short-term goal was to enable the visualization of these raw data via the built-in visualization service included in the cloud level of the platform.

Furthermore, this visualization of the captured raw data was also leveraged as input for further interviews with the representatives of the servitized CEM, both to further detail technical requirements and also to refine the characterization of the smart services to be provided from a business point of view. On one hand, the visualization led to identify different types of noise and missing values in captured raw data and to characterize the necessary filtering and preprocessing components to be deployed in the data capturing and integration platform. On the other hand, it also led to the identification of new possibilities for data exploitation in this manufacturing sector.

7.3.2 Conclusions of Field Validation

The field validation provided valuable conclusions and validated learning related to the two contrasted core components, in order to integrate them into our contribution with design artifacts for the requirement elicitation and analysis in the smartization projects conducted in these scenarios.

7.3.2.1 Design of an Elicitation Interviewing Process

The use of semi-structured interviews with a market-oriented approach [BD12] proved to be a suitable technique for this kind of business contexts. The interviewing meetings tended to be longer (1.5-2 hours) than other documented uses of semi-structured interviews for software engineering [HA05]. The goal of this type of exploration contributed to this, because smart services must be designed not as an *ad hoc*, one-time project, but as a product to be marketed. Therefore, more detailed business vision and market orientation had to be captured during the interviews. This was especially the case in the initial interviews with the CEM's representatives.

The initial design of the set of interviews was based on the features of the complex map of business stakeholders and our *a priori* understanding of exploratory interviews with potential customers for a new service [BD12]. We also identified useful synergies with studies from social sciences on qualitative methods of interview analysis [BM09][Tri09].

The interviews with the CEM's representatives, for instance, evolved from an initial “exploratory” [BM09] approach (aiming at establishing an initial orientation) to “systematizing” [BM09] interviews, trying to obtain more detailed information about specific topics identified in advance. With plant managers, however, the insights and prior knowledge captured in the meetings with the CEM’s representatives were leveraged to prepare more focused, systematizing interviews since the beginning.

The information sent beforehand to business stakeholders about the goals for the interview (centered in the business perspective previously described, as the focus for elicitation) highly contributed to conduct more efficient and goal-oriented interviews. This information plays the role of an “elaborate topic guide” [BM09] to gain access to the interviewee’s knowledge in a systematizing interview. Besides, as this information was not a closed checklist, the interviewed stakeholders felt more free to explain their own views and explore different possibilities, which resulted in a very rich characterization of data analytics needs.

Sending this pre-interview information also helps prepare the interview taking into account the type of interactions and communications a manager is used to in their corporate environment. It is recommended [Tri09] to open the interview with a more guided schedule, as managers are more used to this type of interaction in their corporate environment.

The interaction with stakeholders from the providers of the equipment supporting other steps in the observed manufacturing process led to identify the need for different approaches for interviews. Thus, conducted interviews were not only focused on the capture and translation of business perspective but also on the refinement of the KDDM perspective.

7.3.2.2 Translation from Business into Technical, KDDM-oriented Requirements

The differentiation of information items in two levels, i.e. business and KDDM perspectives, and the use of a template to maintain a record of the characterization at both levels provided multiple benefits. Apart from keeping the traceability of which technical requirements are needed to satisfy which business requirements, it also helped support different focuses of interviews. Indeed, while most interviews were business-oriented and demanded a capture of requirements expressed in terms of the business perspective and their subsequent translation into the KDDM perspective, some other interviews were prepared and conducted with a focus on the refinement of the information captured in the KDDM perspective. In the case of business-oriented interviews, the differentiation of these two information levels also helped keep the business-oriented focus during the interviews (something the business stakeholders felt more comfortable with) and leave the KDDM-oriented reflection for post-interview work.

Also related to business-oriented interviews, it was concluded that the right profile of interviewer is of foremost importance, as the interviewing process and the knowledge of KDDM processes are closely intertwined. Therefore, for the IBDS Provider to leverage this approach and to conduct effective interactions in smartization projects, it is essential to allocate a project manager with the required combination of skills for effective business communication and detailed understanding of a KDDM process. Besides, the interviewee's perception of the interviewer's competences and interests greatly influences interviewee's answers and the interaction model during the interview [BM09]. This is why it was important that the interviewer was familiar with general business and corporate aspects, so that a more balanced interaction could be achieved in these non-technical aspects. This background had to be combined with more domain-specific information captured from the interviews with CEM's representatives. This was crucial to effectively conduct a discursive, argumentative interview with a manager [Tri09].

On a related matter, the direct observation of the manufacturing plant where the servitized CEM was conducting a pilot project with one of their customers was a very valuable resource to understand the physical production environment and the specificities of the equipment generating the data to be captured and analyzed. While other data analytics projects may have a more abstract approach, in these smartization projects for Smart Manufacturing scenarios it is crucial to combine interviews with the direct observation of the source of data, i.e. the production environments. Besides, this observation and the *in situ* interaction with other equipment providers generated key insights to understand the potential heterogeneity to be managed in the project due to differences in equipment among plants. Thanks to this, the appropriate items were added to the KDDM perspective in Table 7.1.

Also, regarding the translation of captured knowledge to KDDM-oriented requirements, it was verified that data-driven services for these business contexts should integrate diverse data models with different analytical approaches (descriptive, predictive or prescriptive; batch or real-time; applied automatically or supporting decision-making processes; etc.) in order to answer the needs from all relevant stakeholders and to support different subsystems along the production process.

In this sense, the availability of the raw time-series data visualization contributes to the success of the project in different dimensions. First, the identification of all relevant indicators is a starting point that helps channeling the interactions with business stakeholders during the first elicitation steps. Second, it supports a first level of value-added service for potential customers in the short term, focused on descriptive analytics of relevant indicators. As previously mentioned in 4.2.3, this facilitates the commitment by manufacturing company owners in order to engage in these projects. Last, it provides a valuable resource for more detailed interactions with stakeholders. Subsequent rounds of interviews can leverage an early visualization of these raw data, so that it is used as an item for discussion. This would provide insights that can be later translated into valuable input for the KDDM process. For instance, regarding the preprocessing phase, it would help establish appropriate criteria for filtering out noise and filling in missing values. It would also help co-design the final dashboards for descriptive analytics needs.

7.4 Proposal of Design Artifacts for the Characterization of Data Exploitation Requirements

The contrasted elements in the field validation can be leveraged by IBDS Providers in their collaboration with manufacturing partners in order to design smart services for a particular manufacturing sector. This approach implies that different levels of stakeholders have to be engaged in a requirements elicitation process. The complexity of the targeted scenario also determines how many levels of stakeholders should be analyzed in the project. For instance, some scenarios demand the capture and processing of distributed data to analyze a continuous production process under different settings. In these cases the complexity of the map of stakeholders will be similar to the scenario analyzed in the field validation. Other scenarios present less complexity, e.g. when the object of data analysis is the particular equipment manufactured by a servitized CEM and not the whole process where it is integrated. This is the case e.g. of a predictive maintenance service for a particular equipment item sold to different companies.

In any case, all these different types of scenarios share the need for identifying key business stakeholders and for designing set of elicitation interviews. Therefore, the contrasted elements can be leveraged to extract relevant knowledge from the appropriate stakeholders in the different Smart Manufacturing scenarios where the IBDS Provider supplies their services. In order to facilitate their use, we contribute with the design of (a) a process model representing an incremental approach for the business stakeholders-driven characterization of requirements for smart services, and (b) a template for practitioners to fill in the characterization of the information items in the business perspective and the KDDM perspective. These design artifacts constitute a valuable contribution that extends existing approaches dealing with requirements elicitation and KDDM process models.

7.4.1 A Spiral Process Model for Business Stakeholders-driven Characterization of Smart Services for a Manufacturing Sector

Parting from the identification of relevant business stakeholders, the process to organize and conduct elicitation interviews, and the requirements capture drawing the connection between the business perspective and the KDDM perspective, we designed a process model integrating these contributions.

Figure 7.3 outlines the integration of the aforementioned contributions. Starting from the *business scenario of the targeted manufacturing sector* (upper-left corner in Figure 7.3) where smart services are to be supplied, two different paths are represented. On the right side it is shown the data lifecycle along several stages, where the two first stages are supported by the data capturing and integration platform designed according to the *Distributed Hybrid Architecture* described along chapter 6. This data lifecycle begins with the automated *capture* of raw data from the monitored manufacturing facilities supported by the aforementioned platform, where the appropriate *preprocessing* (data cleaning and reduction) components are also deployed in order to manage a more efficient *Big Data Lake* centralizing data from all connected facilities. Then, the subsequent phases of a KDDM process (*transformation*, *mining*, and *deployment*) produce different intermediate versions of the dataset and the corresponding data views. This life cycle ends with the analytical models that compose the smart systems to be integrated into the existing manufacturing systems and processes.

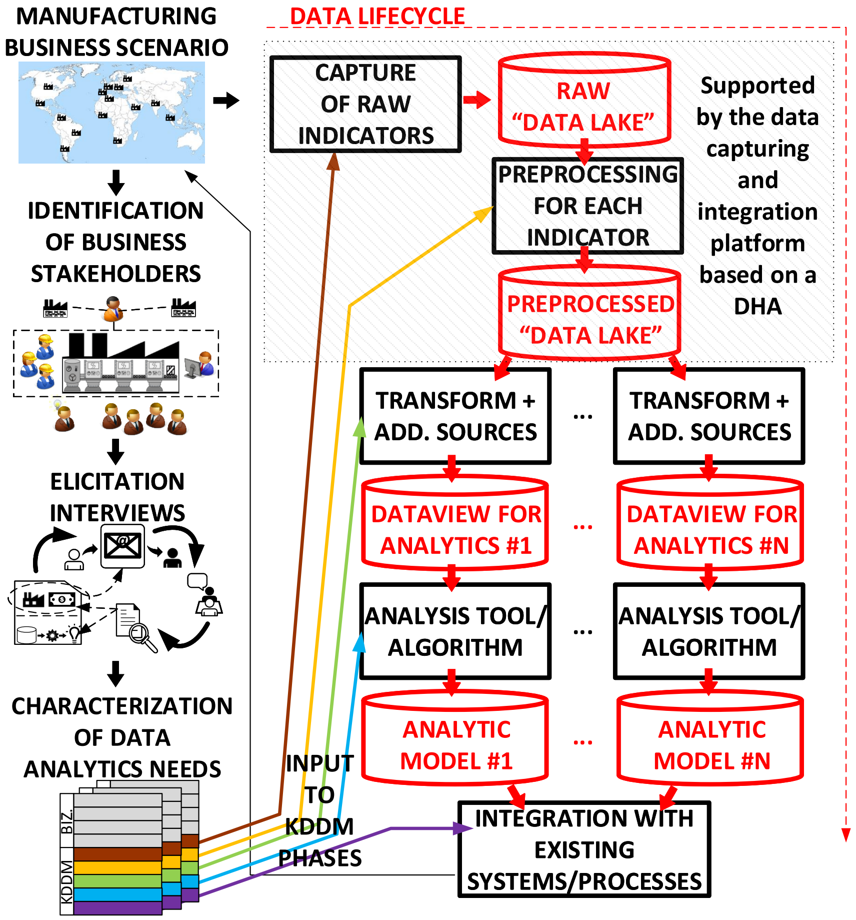

MANUFACTURING BUSINESS SCENARIO
IDENTIFICATION OF BUSINESS STAKEHOLDERS
ELICITATION INTERVIEWS
CHARACTERIZATION OF DATA ANALYTICS NEEDS
BIZ.
KDDM
INPUT TO KDDM PHASES
DATA LIFECYCLE
CAPTURE OF RAW INDICATORS
RAW "DATA LAKE"
Supported by the data capturing and integration platform based on a DHA
PREPROCESSING FOR EACH INDICATOR
PREPROCESSED "DATA LAKE"
TRANSFORM + ADD. SOURCES
...
TRANSFORM + ADD. SOURCES
DATAVIEW FOR ANALYTICS #1
DATAVIEW FOR ANALYTICS #N
ANALYSIS TOOL/ ALGORITHM
ANALYSIS TOOL/ ALGORITHM
ANALYTIC MODEL #1
ANALYTIC MODEL #N
INTEGRATION WITH EXISTING SYSTEMS/PROCESSES

Figure 7.3: Integration of business stakeholders-driven characterization of requirements into the data lifecycle

On the left side of the figure, three elements derived from the field validation are represented: the *identification of relevant business stakeholders* in the analyzed manufacturing scenario, the execution of an *elicitation process* via semi-structured interviews with business stakeholders, and the characterization of data analytics needs not only as use scenarios in business terms, but also as KDDM-oriented requirements. This provides a direct connection to KDDM process phases, so that captured requirements are linked to the phase where they can be used as valuable input to design the appropriate data-related solution.

Nevertheless, the integration of these elements as it is outlined in Figure 7.3 requires an enhancement in order to capture the incremental approach that these scenarios demand. Indeed, the identification of relevant business stakeholders, their engagement in elicitation interviews and the translation of captured requirements as relevant input for the steps in the data lifecycle do not follow a linear, one-run process. Instead, they must be progressively refined and enhanced, given that:

- New relevant business stakeholders are progressively identified in the targeted manufacturing sector, as new customers are interested in the supplied services and the characterization of the business scenario is refined.

- The conduction of additional elicitation interviews with already and newly identified stakeholders leads to a progressive refinement of the characterization of data analytics requirements driving the design of smart services for the targeted sector. Besides, the outcome of pilot projects and first deployments of data exploitation solutions provide insights that can be leveraged and analyzed in additional interviews.

- As new customers, stakeholders and requirements are integrated into the process, the KDDM steps covered along the data lifecycle are enhanced, both refining the solution deployed to solve already covered steps and advancing to further steps.

Figure 7.4 presents a graphical synthesis of an incremental approach integrating the progressive refinement and enhancement of the key components in the proposed stakeholders-driven characterization of data analytics requirements. It is represented as a *spiral process model*, inspired by the proposals of spiral lifecycles for software development [Boe88] and the incremental proposals for the exploration of new business models [BD12][Rie11].

At the center of the diagram in Figure 7.4 it is represented the business scenario of the targeted manufacturing sector, as the starting point much in the same way as in Figure 7.3. The analysis of this business scenario and the successive smartization projects conducted with engaged customers facilitate a progressive identification of relevant business stakeholders. These stakeholders are engaged in an elicitation interviewing process that leads to the characterization of data-related requirements and their translation into relevant input for the KDDM process steps. Most interviews would be conducted by the *smartization project manager* allocated by the IBDS Provider, who combines the skills to conduct a business-oriented interaction and the effective translation of business requirements into technical requirements. Other interviews would directly focus on refining KDDM-oriented requirements with technical stakeholders and therefore would be conducted by data engineers and scientists. Thus, the resolution of the data lifecycle steps that can be solved up to that point can be accomplished with the required correspondence with business requirements.

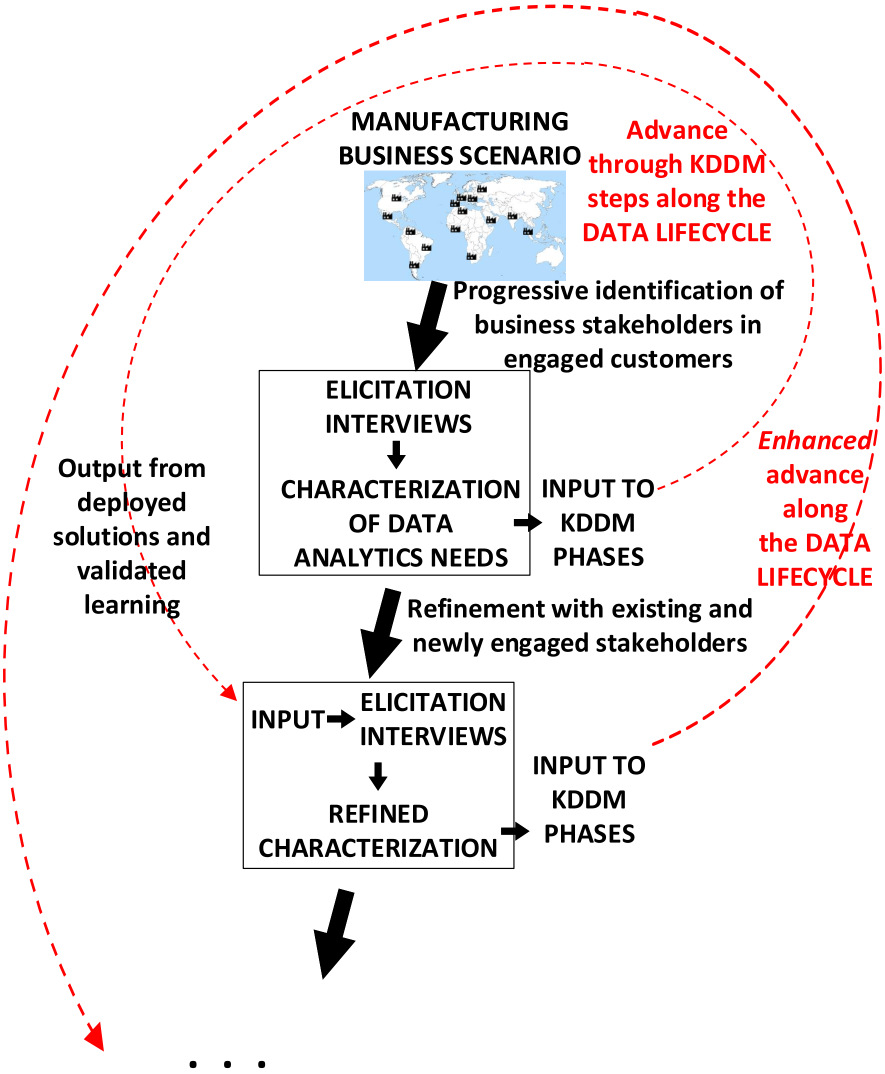


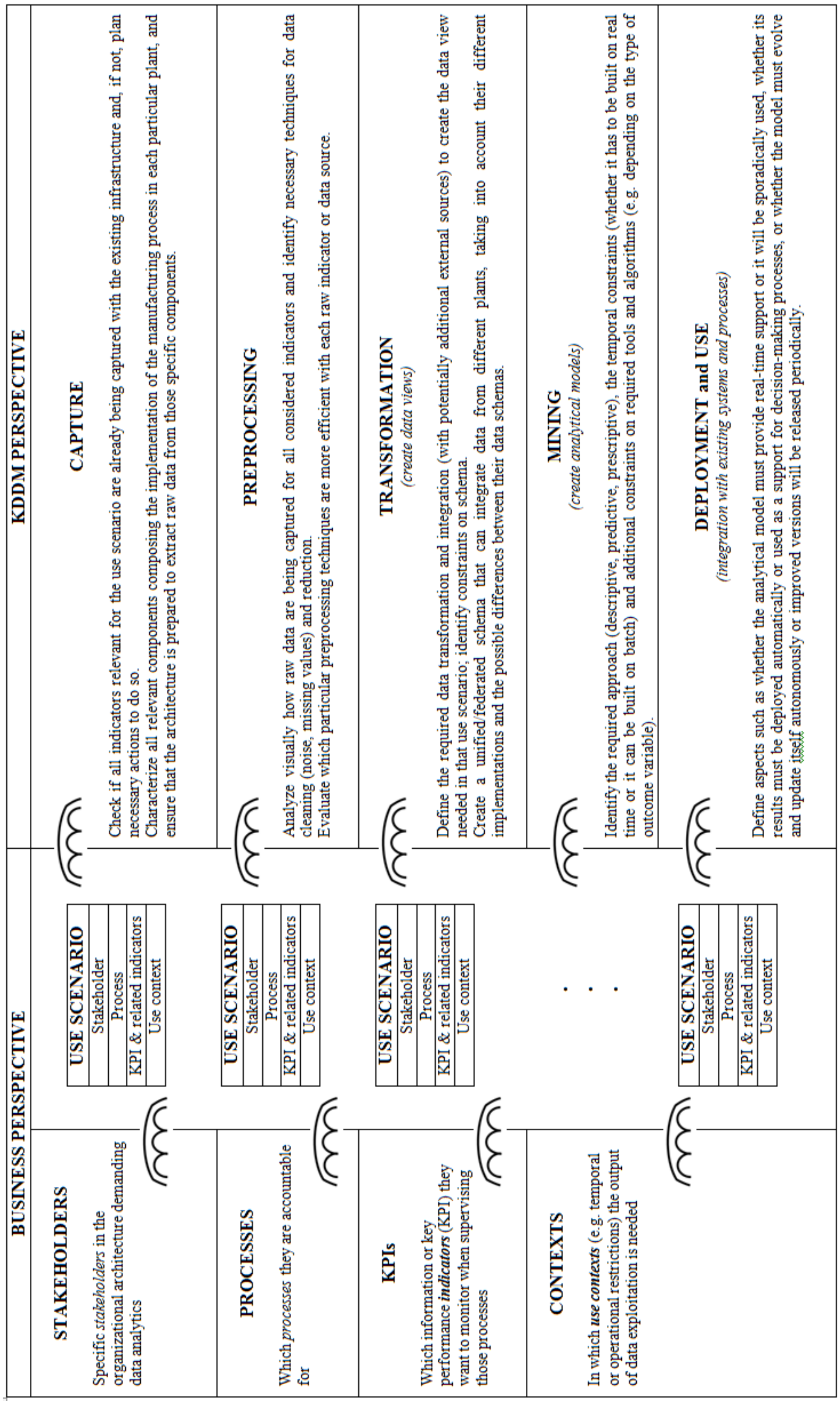
Figure 7.4: Spiral process model for business stakeholders-driven characterization of smart services

The output provided by the steps that have been covered so far in the data lifecycle (e.g. visualization of raw data, in the first iterations) is leveraged as input for new rounds of interviews with business stakeholders. As a refined characterization of requirements is obtained by these interviews and new stakeholders from newly engaged customers are also interviewed, the advance through the data lifecycles is enhanced, both by refining an already covered step (e.g. including more relevant indicators in the capture step) and by advancing further in the lifecycle (e.g. defining and deploying the appropriate preprocessing mechanisms). The successive iterations of this process allow progressively refining the characterization of requirements and the implementation of the technological solutions supporting the data lifecycle and, therefore, the required smart services.

7.4.2 The BRIDGE Canvas: a Template to Capture the Business Requirements' Impact on Data Gathering and Exploitation

Based on the business perspective and the KDDM perspective characterized in Table 7.1 and on their contrasted application during the field validation, we have designed a template to facilitate the capture and successive refinement of the information items contained in both perspectives. The format of this template is based on the type of templates proposed by Osterwalder [Ost04] and later popularized [OP10] and massively adopted among entrepreneurial contexts as a tool to characterize the main features of a business model. In the context of building smart services for a particular manufacturing sector, this canvas template is used (a) to capture the knowledge about the business perspective (relevant stakeholders, process, KPIs, etc.) captured during the elicitation process and (b) to draw a *bridge* (a connection) from the use scenarios identified in the business perspective to the KDDM perspective, based on the impact and implications of these use scenarios on KDDM process steps. Thus, we coined the template as the ***B****usiness* ***R****equirements' **I**mpact on **D**ata **G**athering and **E**xploitation (BRIDGE) canvas.* Figure 7.5 presents the structure and contents of the *BRIDGE canvas*, and Figure 7.6 shows an example of a filled-in BRIDGE canvas with information on use scenarios characterized in the field validation.

The left half of the canvas contains the *business perspective*, i.e. the characterization from the business point of view of the elements that compose the different *use scenarios* for data analytics among the relevant stakeholders in the targeted manufacturing sector. The business perspective is divided in two parts, with the left-most part also subdivided in four areas, one for each of the basic components of use scenarios: *stakeholders* demanding data analytics, *processes* to be enhanced with data analytics, *KPIs* to supervise and optimize, and *use context* for data analytics. Relevant information items for each component can be gathered and listed in the corresponding area. Thus, the right-most part of the business perspective is used to register specific use scenarios to be solved in this market, formed of combinations of elements from the basic components. On the other hand, the right half of the canvas contains KDDM perspective with five areas, one for each of the data lifecycle steps that were represented in Table 7.1. Thus, the relevant input that derives from the characterized used scenarios can



BUSINESS PERSPECTIVE

KDDM PERSPECTIVE

STAKEHOLDERS

Specific *stakeholders* in the organizational architecture demanding data analytics

PROCESSES

Which *processes* they are accountable for

KPIs

Which information or key performance ***indicators*** (KPI) they want to monitor when supervising those processes

CONTEXTS

In which ***use contexts*** (e.g. temporal or operational restrictions) the output of data exploitation is needed

CAPTURE

Check if all indicators relevant for the use scenario are already being captured with the existing infrastructure and, if not, plan necessary actions to do so.

Characterize all relevant components composing the implementation of the manufacturing process in each particular plant, and ensure that the architecture is prepared to extract raw data from those specific components.

PREPROCESSING

Analyze visually how raw data are being captured for all considered indicators and identify necessary techniques for data cleaning (noise, missing values) and reduction.

Evaluate which particular preprocessing techniques are more efficient with each raw indicator or data source.

TRANSFORMATION

(create data views)

Define the required data transformation and integration (with potentially additional external sources) to create the data view needed in that use scenario; identify constraints on schema.

Create a unified/federated schema that can integrate data from different plants, taking into account their different implementations and the possible differences between their data schemas.

MINING

(create analytical models)

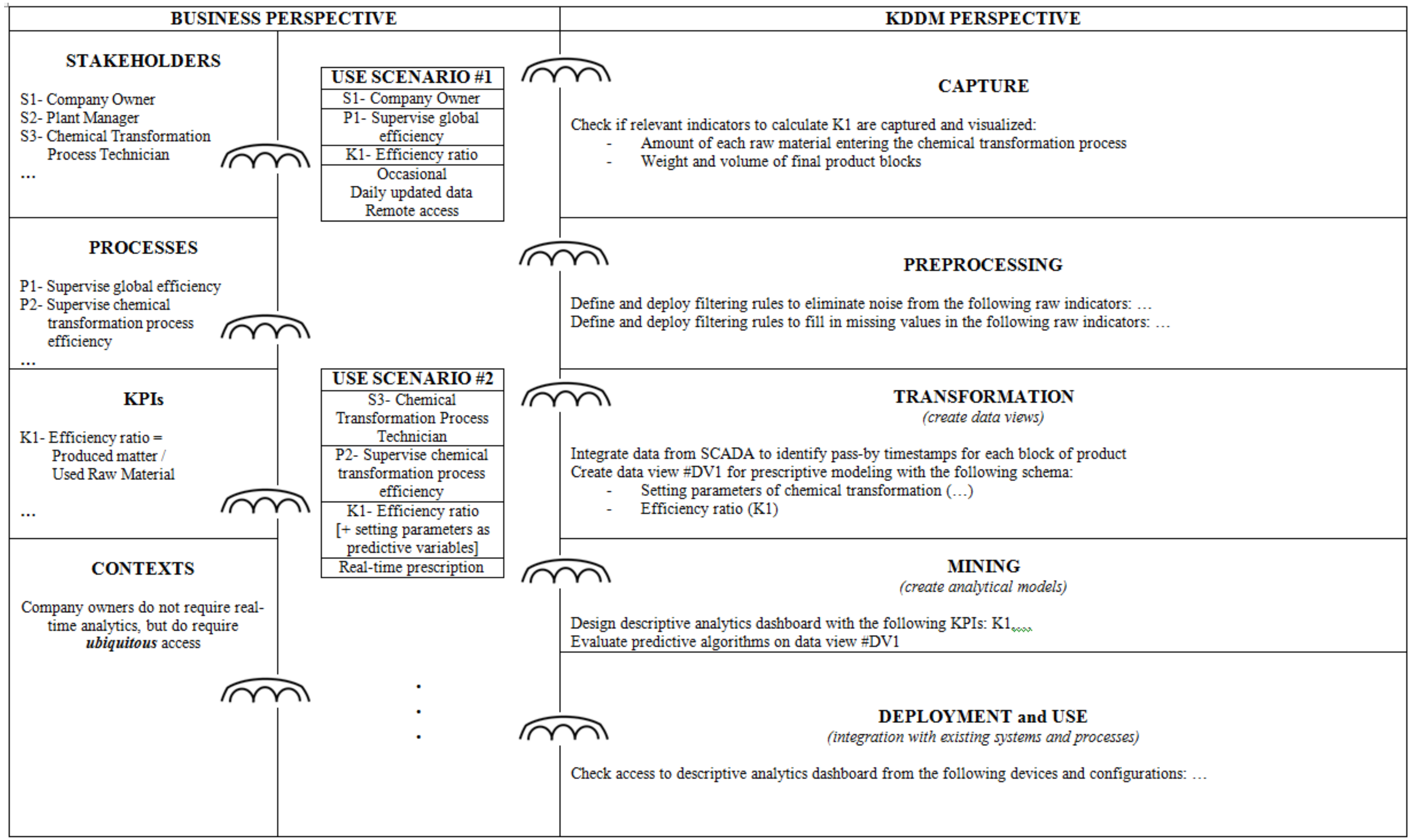
Identify the required approach (descriptive, predictive, prescriptive), the temporal constraints (whether it has to be built on real time or it can be built on batch) and additional constraints on required tools and algorithms (e.g. depending on the type of outcome variable).

DEPLOYMENT and USE

(integration with existing systems and processes)

Define aspects such as whether the analytical model must provide real-time support or it will be sporadically used, whether its results must be deployed automatically or used as a support for decision-making processes, or whether the model must evolve and update itself autonomously or improved versions will be released periodically.

Figure 7.5: Template for the *Business Requirements' Impact on Data Gathering and Exploitation (BRIDGE)* canvas



BUSINESS PERSPECTIVE

STAKEHOLDERS

S1- Company Owner
S2- Plant Manager
S3- Chemical Transformation Process Technician
…

PROCESSES

P1- Supervise global efficiency
P2- Supervise chemical transformation process efficiency
…

KPIs

K1- Efficiency ratio = Produced matter / Used Raw Material
…

CONTEXTS

Company owners do not require real-time analytics, but do require ***ubiquitous*** access

USE SCENARIO #1

S1- Company Owner
P1- Supervise global efficiency
K1- Efficiency ratio
Occasional
Daily updated data
Remote access

USE SCENARIO #2

S3- Chemical Transformation Process Technician
P2- Supervise chemical transformation process efficiency
K1- Efficiency ratio [+ setting parameters as predictive variables]
Real-time prescription

. . .

KDDM PERSPECTIVE

CAPTURE

Check if relevant indicators to calculate K1 are captured and visualized:
- Amount of each raw material entering the chemical transformation process
- Weight and volume of final product blocks

PREPROCESSING

Define and deploy filtering rules to eliminate noise from the following raw indicators: …
Define and deploy filtering rules to fill in missing values in the following raw indicators: …

TRANSFORMATION
(create data views)

Integrate data from SCADA to identify pass-by timestamps for each block of product
Create data view #DV1 for prescriptive modeling with the following schema:
- Setting parameters of chemical transformation (…)
- Efficiency ratio (K1)

MINING
(create analytical models)

Design descriptive analytics dashboard with the following KPIs: K1
Evaluate predictive algorithms on data view #DV1

DEPLOYMENT and USE
(integration with existing systems and processes)

Check access to descriptive analytics dashboard from the following devices and configurations: …

Figure 7.6: Example of BRIDGE canvas with use scenarios characterized in the field validation

be assigned to the appropriate lifecycle step, so that it is leveraged for the design of the technological support for each step. Moreover, those interviews that are directly focused on KDDM aspects of smartization projects (e.g. the interviews with technical managers from other equipment providers in the field validation) would contribute to refine the information captured in the right half of the canvas. As new stakeholders are engaged in the elicitation process of smartization projects, the information in the BRIDGE canvas is progressively refined to reflect in the most accurate possible way the data-related needs to be solved by smart services in that manufacturing sector and the requirements for the data-driven technological solution to support them.

The information items in the business perspective also guide the preparation of elicitation interviews, as it was shown in the field validation. Indeed, the information on stakeholders, processes and indicators in the analyzed manufacturing business scenario can be leveraged to prepare the information to send beforehand to interviewed stakeholders.

7.5 Conclusions

The development of smart services to evolve manufacturing production systems in these Smart Manufacturing scenarios demands extensions and adaptations of existing KDDM process models. This is due to the fact that the project goal is not to build an internal tool, but to build a knowledge-based product to be later commercialized as part of a value-added service for manufacturing companies who want to shift towards a Smart Manufacturing approach. This adds to the complexity in the characterization of business needs and goals, as well as their impact on the KDDM aspects of the deployed technological solution, as these data-related needs correspond to a multi-view scenario integrating data exploitation requirements from multiple stakeholder profiles.

The proposed *spiral process model* and the supporting *BRIDGE canvas* are the outcome of a design science research process that ensures the contribution of purposeful design artifacts for business scenarios with the aforementioned characteristics. Apart from the fulfilling of requirements that ensure the relevance of contributions, the grounding of the proposed artifacts is based on the combined synergies of a diversity of knowledge areas: requirements engineering and elicitation, interview analysis, KDDM process models, stakeholders management, and the design and development of smart services in particular and new services and business models in general. The field testing conducted in the real-world business setting of our case study allowed us to contrast the validity and applicability of the proposed approach and its practical elements: an incremental approach to organize and conduct an elicitation interviewing process with relevant business stakeholders, and the use of a supporting tool to capture the requirements during the elicitation process and to establish the link between business requirements and their impact in KDDM process stages. Furthermore, the proposed contributions are aligned with proposed enhancement potentials for existing approaches dealing with requirements analysis for analytical information systems [SWW11].

Contributing with the proposed design artifacts opens the possibility of a contrast with further works analyzing other kinds of smart manufacturing scenarios. Such works are indeed arising given the trending interest in this research field. This contrast would contribute to consolidate a methodology, enhancing existing approaches to deal with the new challenges in this type of projects. This methodological support will provide a more complete vision of project milestones, stakeholders to involve, a timeline of expected outcomes and the required steps to achieve them.

Chapter 8

Conclusions

The progressive transformation of manufacturing industry with the adoption of Smart Manufacturing-related business strategies represents one of the most important focuses of economic development worldwide during the 2010s decade. The interest by manufacturing companies in Smart Manufacturing, boosted by diverse initiatives and policies worldwide promoting its adoption, is based on the possibilities to transform their production processes and their business models. On one hand, significant gains in the efficiency of automated production systems, the quality of produced goods and profit in general are expected via the adoption of these data-driven approaches and the value extracted from data insights. On the other hand, it enables a shift towards data-driven servitization strategies for those equipment manufacturers that want to transform their business models via the supply of value-added services to their manufacturing customers. The expected benefits of these different approaches have led to diverse goals for Smart Manufacturing applications: production system control, product quality control, decision-support systems, fault diagnosis and predictive maintenance of equipment, etc.

This context has led to the emergence of a specialization among providers of IT services, focused on the supply of *Industrial Big Data Services* (IBDS). These technological services are related to the data capturing and exploitation solutions that are required for the effective development of Smart Manufacturing approaches. In order to supply these data-driven services, *IBDS Providers* establish partnerships with manufacturers in different sectors and markets and develop *smartization projects* for the deployment of the required solutions in the facilities owned by engaged manufacturers. These projects are developed in parallel in various sectors and aim at progressively deploy and refine the smart services required by each scenario. The management of these smartization projects entails important challenges for IBDS Providers regarding (a) organizational aspects linked to the required roles in the team carrying out those projects and (b) technological aspects related to the design of the required data capturing and integration platform sustaining the worldwide deployment of multiple projects in parallel. Furthermore, all these aspects must be aligned with IBDS Providers' business strategy and also with the requirements and needs of the various manufacturers with whom they establish partnerships across multiple sectors.

The complexity of these projects carried out by IBDS Providers motivates and provides the focus for this research work. The three main contributions presented in this dissertation aim at providing valuable solutions for specific challenges in these smartization projects, particularly in the design of the required smart services in a collaborative way with partnering manufacturers and in the technological support for the early stages in the data lifecycle that enable the availability of manufacturing data to be exploited. The targeted challenges are specifically related to the duty of two of the involved roles in IBDS Providers: the *project manager* that drives the interaction with relevant stakeholders from engaged manufacturers and the elicitation of requirements for smart services, and the *data engineer* in charge of the design, update and optimization of the data capturing and integration platform.

With respect to the duty of the project manager, this research work contributes with the design of a *spiral process model* representing an incremental approach for the business stakeholders-driven characterization of requirements for smart services, and the *BRIDGE canvas* as the template to capture the business requirements for these smart services and their connection and implications for the data lifecycle steps in a KDDM process model. These design artifacts support the progressive identification of relevant stakeholders in the targeted manufacturing sector and the elicitation of requirements from them, as new customers are engaged in the supplied services and the characterization of the business scenario is refined. The spiral process model and the BRIDGE canvas constitute a valuable contribution that extends existing approaches dealing with requirements elicitation and KDDM process models, based on the combined synergies with knowledge areas such as project and stakeholders management, interview analysis and business model design.

Regarding the duty of the data engineer, two main contributions are proposed. On one hand, it is presented the design of a *Decentralized Hybrid Architecture* (DHA) for the data capturing and integration platform of an IBDS Provider. The design of DHA leverages the analysis of the Industrial IoT and Cloud Computing components deployed for more than 60 manufacturing facilities distributed worldwide where data are captured and centralized for their later exploitation via different services. The DHA fulfills the main non-functional, business requirements derived from the Smart Manufacturing scenarios where IBDS Providers supply their services. It ensures a non-intrusive and scalable deployment in manufacturing facilities with already operating infrastructures, enabling the progressive upgrade of its modules to increasingly cover more application scenarios and more data transformation stages. It also draws synergies and complements existing Big Data-related paradigms, saving the gap between an initial state where no data are extracted yet from manufacturing facilities and the eventual availability of a centralized data repository, conceived as a *Big Data Lake*, on top of which diverse exploitation functionalities may be designed following the *Lambda Architecture*.

On the other hand, it is presented the design of the *planning and execution of the time-series data reduction analysis* to be carried out by the data engineer.

This reduction analysis addresses the optimization of one of the most relevant internal costs for an IBDS Provider and their platform: the cloud storage resources for the accumulated manufacturing data captured as time series from different sensors and production equipment in all the monitored facilities worldwide. The proposed procedural and architectural modeling of reduction analysis planning and execution allows the data engineer to optimize the time and resources they can devote to compose the reduction solution, i.e. which reduction techniques to apply to which time-series data, to be deployed into the preprocessing components of the DHA. This contribution helps the data engineer obtain the best overall reduction possible thanks to the combination of different families of reduction techniques, and manage the accumulated knowledge from previous analyses to sustain the optimization in storage costs savings.

These three main contributions integrate key practical elements derived from the direct observation and hands-on experience developed in our case study. In this regard, the real-world business setting where we conducted our case study has constituted an immensely valuable resource. It has granted us direct access to organizations developing their business strategies in the targeted Smart Manufacturing scenarios, allowing us to observe the complexity of these real-world scenarios and the practical issues and challenges to face [NnBI15] when developing smartization projects that aim at connecting data-related technological solutions to the reality of the manufacturing industry and their operational technology. Moreover, it has provided us with insightful knowledge of how these solutions drive the servitization strategies of equipment providers, giving us direct access to an instance of manufacturing sector distributed worldwide, the facilities where data-driven services are deployed and the stakeholders from diverse organizations involved in this context. The representativeness of the analyzed organizations, stakeholders and technology has facilitated a rich characterization of these scenarios and the identification of the relevant aspects to take into account when building and deploying data-driven services in real-life manufacturing business scenarios. This characterization evidences the relevance of the presented contributions and of others that can address more challenges and requirements derived from these scenarios.

8.1 Future Work

The integration of different disciplines in order to build our contributions for the smartization projects developed by IBDS Providers, together with the characterization of the targeted Smart Manufacturing scenarios, their main agents and their respective requirements and needs, constitute valuable consolidated knowledge that opens the possibility for diverse lines of further research works. We group those lines in two main general directions. First, the integrative approach developed along this dissertation, sustained by a research method supported by Design Science Research and Case Study Research, can be applied to an extension of the scope of this research work, either extending the targeted Smart Manufacturing scenarios (e.g. extending the monitored data to those generated during the use of the manufactured product) or covering more business and technological challenges that IBDS Providers face in these scenarios, related to further stages in the data lifecycle. Indeed, given that this research work focuses on those data lifecycle stages ensuring the availability of new manufacturing data for the data exploitation layers, a similar research approach can be followed to analyze further steps that implement the analytics steps on those data. This will also extend the roles in IBDS Providers' smartization projects that could leverage the proposed contributions. Second, the followed multidisciplinary approach opens the possibility for the examined disciplines and research areas to delve into their potential respective contributions for the analyzed problem and the targeted scenarios. Thus, the contributions presented along this dissertation can be enriched by further specialized contributions in the integrated areas.

Focusing on this last identified direction for future work, there are several potential lines to extend different components of the proposed contributions. Regarding the *stakeholders-driven characterization of data exploitation requirements*, the integration of knowledge from various areas opens the possibility to further work with different focuses. For instance, the presented contribution has close relationship with KDDM process models such as CRISP-DM, which introduces the concept of *specialized process model* for versions derived from the CRISP-DM general model that include particular elements for specific application scenarios. Thus, the proposed design artifacts can be integrated into a specialized process model using the same constructs as CRISP-DM, e.g. the differentiation between generic and specialized tasks, mapping of generic models, etc. On a related matter, researchers from the requirements engineering area can leverage the presented concepts, such as the details from the KDDM perspective, to extend and specialize their requirements engineering proposals for data-driven projects. With respect to elicitation interviews, the interviewing process can be formally designed according to a specific lifecycle, where the practical constraints and the stakeholder characterization in each project have to be mapped into a plan of elicitation interviews. This plan would integrate different interviewing approaches and goals as progressively more stakeholders are engaged.

With respect to the *architectural proposal for the data capturing and integration platform*, the flexibility in the adopted approach for extending its functionalities facilitates the future integration of new communication and secure connectivity standards, as they become part of the technological reality observed in the targeted industrial scenarios. Thus, there is an open opportunity for research works that advance those future scenarios and detail integration schemas with new communication and connectivity proposals. Regarding the extension to cover further steps in the data lifecycle, a relevant issue to be explored is the best approach to integrate the result of the exploitation and analytics steps, e.g. predictive models once tested and validated, back into the infrastructure deployed in the monitored facilities. On this matter, the use of standard representation formats, such as the Predictive Model Markup Language and the Portable Format for Analytics by the Data Mining Group[1], provide valuable mechanisms for an easier deployment and portability of predictive models.

Regarding the proposed design for *time-series data reduction analysis*, apart from the inclusion and contrast of new families of time series, new reduction

[1] http://dmg.org/pfa/docs/motivation/

techniques and new performance dimensions, an important open research line is the integration with of ontology-based formal representations for sensors and their observations, so that the conceptual model supporting our proposal can be represented as an extension of these ontologies. This would facilitate their integration with other data representation systems. On a different matter, in order to provide specific implementations of the proposed approach to facilitate its application, there is ongoing work to integrate implemented algorithms [SL17] for time series classification.

8.2 Overall Conclusion

The main differential value of the contributions presented in this dissertation is that they map adequately to an identification of real-world problems in the analyzed business scenarios. Indeed, the solution for those problems requires the assurance of practical requirements that are drawn from the analysis of such scenarios and for which existing proposals need to be adapted and extended. In this regard, an additional value of this work is its multidisciplinary approach integrating knowledge from many different research areas, drawing synergies and identifying limitations as an opportunity for valuable contributions. The utility and applicability of these contributions have been contrasted and validated in a real-world business setting as a relevant instance of the Smart Manufacturing scenarios where these contributions are targeted at.

Furthermore, the proposed contributions can constitute a valuable resource for *both practitioners and researchers*. On one hand, they provide a global benefit for IBDS Providers and, by extension, for the manufacturing industry aiming at increasing their competitiveness thanks to the adoption of Smart Manufacturing approaches. These contributions also enhance the role of IBDS Providers as necessary agents in the strategic development of the manufacturing industry and the effective deployment of Smart Manufacturing adoption policies. On the other hand, they integrate and extend existing conceptual, methodological and technological proposals in diverse knowledge and research areas. In this regard, we aim at putting the spotlight on practical aspects that are required for leveraging these proposals in real-world scenarios where IBDS Providers supply their services, so that these aspects can be taken into account when devising future versions of these proposals.

Bibliography

[AAAS15] Federico Adrodegari, Andrea Alghisi, Marco Ardolino, and Nicola Saccani. From Ownership to Service-oriented Business Models: A Survey in Capital Goods Companies and a PSS Typology. *Procedia CIRP*, 30:245–250, 2015. DOI:10.1016/j.procir.2015.02.105.

[AAB13] Federico Adrodegari, Andrea Alghisi, and Andrea Bacchetti. Servitization of Capital Good Manufacturers: an empirical research in Italian machinery sector. In *Proceedings of the 18th International Symposium on Logistics*, pages 34–43, Vienna, Austria, July 2013. DOI:10.13140/2.1.4279.1365.

[Aas69] K. J. Aastroem. On the choice of sampling rates in parametric identification of time series. *Information Sciences*, 1(3):273–278, July 1969. DOI:10.1016/S0020-0255(69)80013-7.

[AW05] Aybke Aurum and Claes Wohlin. Requirements Engineering: Setting the Context. In *Engineering and Managing Software Requirements*, pages 1–15. Springer-Verlag, Berlin, Germany, 2005. DOI:10.1007/3-540-28244-0_1.

[Bas17] Brian Bassett. A Brief History of Enterprise Software - Part 2, Cloud City and Open Source Makin' It Rain, January 2017. URL:http://corgibytes.com/blog/2017/01/05/enterprise-software-pt2/.

[BD12] Steve Blank and Bob Dorf. *The Startup Owner's Manual: The Step-By-Step Guide for Building a Great Company.* K&S Ranch, 2012.

[BFL13] Giuseppe Burtini, Scott Fazackerley, and Ramon Lawrence. Time series compression for adaptive chart generation. In *Proceedings of the 2013 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering*, pages 1–6, Regina, Saskatchewan, Canada, May 2013. DOI:10.1109/CCECE.2013.6567840.

[Bia00] Alessandra Bianchi. Upstarts: ASPs, April 2000. URL:http://www.inc.com/magazine/20000401/18093.html.

[BKM+14] Alexander Brodsky, Mohan Krishnamoorthy, Daniel A Menascé, Guodong Shao, and Sudarsan Rachuri. Toward smart manufacturing using decision analytics. In *Proceedings of 2014 IEEE International Conference on Big Data*, pages 967–977, Washington, DC, USA, October 2014. DOI:10.1109/BigData.2014.7004330.

[BLBK09] T. S. Baines, H. W. Lightfoot, O. Benedettini, and J. M. Kay. The servitization of manufacturing: A review of literature and reflection on future challenges. *Journal of Manufacturing Technology Management*, 20(5):547–567, 2009. DOI:10.1108/17410380910960984.

[BLK11] Marina Berkovich, Jan Marco Leimeister, and Helmut Krcmar. Requirements Engineering for Product Service Systems: A State of the Art Analysis. *Business & Information Systems Engineering*, 3(6):369–380, December 2011. DOI:10.1007/s12599-011-0192-2.

[Blo14] Robin Bloor. It's Not a Data Lake, It's a Data Reservoir, July 2014. URL:http://insideanalysis.com/2014/07/its-not-a-data-lake-its-a-data-reservoir/.

[BM09] Alexander Bogner and Wolfgang Menz. The Theory-Generating Expert Interview: Epistemological Interest, Forms of Knowledge, Interaction. In *Interviewing Experts*, pages 43–80. Palgrave Macmillan, Basingstoke, UK, 2009. DOI:10.1057/9780230244276_3.

[BM12] Arshdeep Bahga and Vijay K. Madisetti. Analyzing Massive Machine Maintenance Data in a Computing Cloud. *IEEE Transactions on Parallel and Distributed Systems*, 23(10):1831–1843, October 2012. DOI:10.1109/TPDS.2011.306.

[BMZA12] Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. Fog Computing and Its Role in the Internet of Things. In *Proceedings of the 1st edition of the MCC workshop on Mobile cloud computing*, pages 13–15, Helsinki, Finland, August 2012. DOI:10.1145/2342509.2342513.

[Boe88] Barry W. Boehm. A Spiral Model of Software Development and Enhancement. *Computer*, 21(5):61–72, May 1988. DOI:10.1109/2.59.

[Bon11] Flavio Bonomi. Connected Vehicles, the Internet of Things, and Fog Computing. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, Las Vegas, NV, USA, September 2011. URL:https://www.sigmobile.org/mobicom/2011/vanet2011/program.html.

[BXW14] Zhuming Bi, Li Da Xu, and Chengen Wang. Internet of Things for Enterprise Systems of Modern Manufacturing. *IEEE Transactions on Industrial Informatics*, 10(2):1537–1546, May 2014. DOI:10.1109/TII.2014.2300338.

[CCK+00] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rdiger Wirth. CRISP-DM

1.0: Step-by-step data mining guide. Technical report, SPSS, 2000. URL:ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.

[CF99] Kin-Pong Chan and Ada Wai-Chee Fu. Efficient Time Series Matching by Wavelets. In *Proceedings of the 15th International Conference on Data Engineering*, pages 126–133, Sydney, Australia, March 1999. DOI:10.1109/ICDE.1999.754915.

[CFLN02] Fu-Lai Chung, Tak-Chung Fu, Robert Luk, and Vincent Ng. Evolutionary Time Series Segmentation for Stock Data Mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 83–90, Maebashi City, Japan, December 2002. DOI:10.1109/ICDM.2002.1183889.

[CKMP02] Kaushik Chakrabarti, Eamonn Keogh, Sharad Mehrotra, and Michael Pazzani. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. *ACM Transactions on Database Systems*, 27(2):188–228, June 2002. DOI:10.1145/568518.568520.

[CLMT13] Ania Cravero-Leal, José Norberto Mazón, and Juan Trujillo. A business-oriented approach to data warehouse development. *Ingeniería e Investigación*, 33(1):59–65, April 2013. URL:http://hdl.handle.net/10045/33413.

[CLSM$^+$14] Ania Cravero-Leal, Samuel Seplveda, Alejandro Maté, José Norberto Mazón, and Juan Trujillo. Goal oriented requirements engineering in data warehouses: a comparative study. *Ingeniería e Investigación*, 34(2):66–70, August 2014. DOI:10.15446/ing.investig.v34n2.44708.

[CPL16] Brian Caffo, Roger D. Peng, and Jeffrey Leek. *Executive Data Science: A Guide to Training and Managing the Best Data Scientists*. Leanpub, May 2016. URL:http://leanpub.com/eds.

[Cut09] Douglass Read Cutting. Joining Cloudera, August 2009. URL:https://cutting.wordpress.com/2009/08/10/joining-cloudera/.

[Dav10] Jim Davis. Implementing 21st Century Smart Manufacturing. Technical report, Smart Manufacturing Leadership Coalition, September 2010. URL:https://smartmanufacturingcoalition.org/sites/default/files/meaningful_use_priorities_and_metrics_recommendations_on_public-private_partnership_programs_1.pdf.

[DED$^+$09] Jim Davis, Tom Edgar, Yiannis Dimitratos, Jerry Gipson, Ignacio Grossmann, Peggy Hewitt, Ric Jackson, Kevin Seavey, Jim Porter, Rex Reklaitis, and Bruce Strupp. Smart Process Manufacturing: An Operations and Technology Roadmap. Technical report, Smart Process Manufacturing Engineering Virtual Organization, November 2009.

URL:https://www.smartmanufacturingcoalition.org/sites/default/files/spm_-_an_operations_and_technology_roadmap.pdf.

[DEP+12] Jim Davis, Thomas Edgar, James Porter, John Bernaden, and Michael Sarli. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Computers & Chemical Engineering*, 47:145–156, December 2012. DOI:10.1016/j.compchemeng.2012.06.037.

[DG04] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation (OSDI'04)*, volume 6, pages 10:1–13, San Francisco, CA, USA, December 2004.

[Dha12] Subhankar Dhar. From outsourcing to Cloud computing: evolution of IT services. *Management Research Review*, 35(8):664–675, 2012. DOI:10.1108/01409171211247677.

[Dix10] James Dixon. Pentaho, Hadoop, and Data Lakes, October 2010. URL:https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

[DR08] T. Dierks and E. Rescorla. The Transport Layer Security (TLS) Protocol - Version 1.2. Technical report, Internet Engineering Task Force, August 2008. URL:https://tools.ietf.org/html/rfc5246.

[DUM+15] Veit Dinges, Florian Urmetzer, Verónica Martínez, Mohamed Zaki, and Andy Neely. The future of servitization: Technologies that will make a difference. Technical report, University of Cambridge, June 2015. URL:http://cambridgeservicealliance.eng.cam.ac.uk/resources/Downloads/Monthly%20Papers/150623FutureTechnologiesinServitization.pdf.

[EA12] Peter C. Evans and Marco Annunziata. Industrial Internet: Pushing the Boundaries of Minds and Machines. Technical report, General Electrics, November 2012. URL:http://www.ge.com/sites/default/files/Industrial_Internet.pdf.

[EEC+09] Hazem Elmeleegy, Ahmed K. Elmagarmid, Emmanuel Cecchet, Walid G. Aref, and Willy Zwaenepoel. Online Piece-wise Linear Approximation of Numerical Streams with Precision Guarantees. *Proceedings of the VLDB Endowment*, 2(1):145–156, August 2009. DOI:10.14778/1687627.1687645.

[Eis89] Kathleen M. Eisenhardt. Building Theories from Case Study Research. *The Academy of Management Review*, 14(4):532–550, October 1989.

[Eur14] European Commission. Towards a thriving data-driven economy. Technical report, Towards a thriving data-driven economy, July 2014. URL:https://ec.europa.eu/digital-single-market/news/communication-data-driven-economy.

[Eur15] European Commission. Report from the Workshop on Innovation in Digital Manufacturing. Technical report, European Commission, February 2015. URL:http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=8736.

[Eur16] European Factories of the Future Research Association. Factories 4.0 and Beyond. Technical report, EFFRA, September 2016. URL:http://effra.eu/attachments/article/129/Factories40_Beyond__v30_public.pdf.

[Eur17a] European Commission. The Factories of the Future, May 2017. URL:https://ec.europa.eu/digital-single-market/en/smart-manufacturing-0.

[Eur17b] European Commission. Innovation: ICT for Manufacturing SMEs, May 2017. URL:https://ec.europa.eu/digital-single-market/en/smart-manufacturing-1.

[FKF16] Hans Fleischmann, Johannes Kohl, and Jrg Franke. A Reference Architecture for the Development of Socio-Cyber-Physical Condition Monitoring Systems. In *Proceedings of the 11th System of Systems Engineering Conference*, pages 1–6, Kongsberg, Norway, June 2016. DOI:10.1109/SYSOSE.2016.7542963.

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37–54, 1996.

[FU95] Usama Fayyad and Ramasamy Uthurusamy. Preface. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, Montréal, Québec, Canada, August 1995.

[Fu11] Tak-Chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, February 2011. DOI:j.engappai.2010.09.007.

[Gar05] Gartner. Gartner Survey of 1,300 CIOs Shows IT Budgets to Increase by 2.5 Percent in 2005 (Press Release), January 2005. URL:http://www.gartner.com/newsroom/id/492096.

[Gar10] Gartner. Gartner EXP Worldwide Survey of Nearly 1,600 CIOs Shows IT Budgets in 2010 to be at 2005 Levels, January 2010. URL:http://www.gartner.com/newsroom/id/1283413.

[Gar12] Gartner. Insights From The 2012 Gartner CIO Agenda Report. Technical report, Gartner, 2012. URL:http://imagesrv.gartner.com/cio/pdf/cio_agenda_insights.pdf.

[Gar17] Gartner. Insights From The 2017 CIO Agenda Report. Technical report, Gartner, 2017. URL:http://www.gartner.com/imagesrv/cio/pdf/Gartner_CIO_Agenda_2017.pdf.

[GGB12] Juozas Gordevicius, Johann Gamper, and Michael Bhlen. Parsimonious temporal aggregation. *The VLDB Journal*, 21(3):309–332, June 2012. DOI:10.1007/s00778-011-0243-9.

[GGL03] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google File System. In *Proceedings of the 19th ACM symposium on Operating systems principles*, pages 29–43, Bolton Landing, NY, USA, October 2003.

[GLH15] Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*. Springer International Publishing, Switzerland, 2015. DOI:10.1007/978-3-319-10247-4.

[GRG08] Paolo Giorgini, Stefano Rizzi, and Maddalena Garzetti. GRAnD: A goal-oriented approach to requirement analysis in data warehouses. *Decision Support Systems*, 45(1):4–21, April 2008. DOI:10.1016/j.dss.2006.12.001.

[HA05] Siw Elisabeth Hove and Bente Anda. Experiences from Conducting Semi-Structured Interviews in Empirical Software Engineering Research. In *Proceedings of the 11th IEEE International Software Metrics Symposium*, Como, Italy, September 2005. DOI:10.1109/METRICS.2005.24.

[Hev07] Alan R. Hevner. A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2):87–92, 2007.

[HEVY15] María Holgado, Steve Evans, Doroteya Vladimirova, and Miying Yang. An internal perspective of business model innovation in manufacturing companies. In *Proceedings of the 2015 IEEE 17th Conference on Business Informatics*, pages 9–16, Lisbon, Portugal, July 2015. DOI:10.1109/CBI.2015.42.

[HK06] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, San Francisco, CA, USA, 2nd edition, 2006.

[HMPR04] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design Science in Information Systems Research. *Management Information Systems Quarterly*, 28(1):75–105, March 2004.

[HSSK06] Jennifer Anne Harding, Muhammad Shahbaz, S Srinivas, and Andrew Kusiak. Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering*, 128(4):969–976, November 2006. DOI:10.1115/1.2194554.

[HVH15] Johann Hufnagel and Birgit Vogel-Heuser. Data Integration in Manufacturing Industry: Model-Based Integration of Data Distributed from ERP to PLC. In *Proceedings of the IEEE 13th International Conference on Industrial Informatics*, pages 275–281, Cambridge, UK, July 2015. DOI:10.1109/INDIN.2015.7281747.

[IK08] Institute of Control, Robotics and Systems and Korea Machine Tool Manufacturers' Association. ICSMA 2008. In *Proceedings of the International Conference on Smart Manufacturing Application*, Gyeonggi-do, Korea, April 2008.

[Ind15] Industrial Internet Consortium. Fact Sheet. Technical report, Industrial Internet Consortium, September 2015. URL:https://www.iiconsortium.org/docs/IIC_FACT_SHEET.pdf.

[Ind17a] Industrial Internet Consortium. The Industrial Internet of Things Volume G1: Reference Architecture. Technical report, Industrial Internet Consortium, January 2017. URL:http://www.iiconsortium.org/IIRA.htm.

[Ind17b] Industrial Internet Consortium. Technical Papers, Publications, and White Papers, 2017. URL:http://www.iiconsortium.org/white-papers.htm.

[Ind17c] Industrial Internet Consortium. Testbeds, 2017. URL:http://www.iiconsortium.org/test-beds.htm.

[Int08] Integrated Manufacturing Technology Initiative. Smart Process Manufacturing Workshop Report. Technical report, Integrated Manufacturing Technology Initiative, May 2008. URL:https://www.smartmanufacturingcoalition.org/sites/default/files/spm-workshop-report.pdf.

[IO17] IDC and Open Evidence. The European Data Market Study: Final Report. Technical report, DataLandscape, February 2017. URL:http://www.datalandscape.eu/study-reports.

[JML$^+$15] Kiwook Jung, K. C. Morris, Kevin W. Lyons, Swee Leong, and Hyunbo Cho. Mapping Strategic Goals and Operational Performance Metrics for Smart Manufacturing Systems. *Procedia Computer Science*, 44:184–193, 2015. DOI:10.1016/j.procs.2015.03.051.

[JSS$^+$16] Cun Ji, Qingshi Shao, Jiao Sun, Shijun Liu, Li Pan, Lei Wu, and Chenglei Yang. Device Data Ingestion for Industrial Big Data Platforms with a Case Study. *Sensors*, 16(3):279:1–15, March 2016. DOI:10.3390/s16030279.

[JZFV16] Erkki Jantunen, Urko Zurutuza, Luis Lino Ferreira, and Pal Varga. Optimising Maintenance: What are the expectations for Cyber Physical Systems. In *Proceedings of the 3rd International Workshop on Emerging Ideas and Trends in Engineering of Cyber-Physical Systems*, pages 53–58, Vienna, Austria, April 2016. DOI:10.1109/EITEC.2016.7503697.

[KA14] Bart Kamp and Henar Alcalde. Servitization in the Basque Economy. *Strategic Change*, 23(5-6):359–374, August 2014. DOI:10.1002/jsc.1982.

[KAG$^+$16] Henning Kagermann, Reiner Anderl, Jrgen Gausemeier, Gnther Schuh, and Wolfgang Wahlster. Industrie 4.0 in a Global Context: Strategies for Cooperating with International Partners. Technical report, Acatech, November 2016. URL:http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Publikationen/Projektberichte/acatech_eng_STUDIE_Industrie40_global_Web.pdf.

[KCK$^+$15] Hussein Khaleel, Davide Conzon, Prabhakaran Kasinathan, Paolo Brizzi, Claudio Pastrone, Ferry Pramudianto, Markus Eisenhauer, Pietro A. Cultrona, Fulvio Rusina, Gabriel Lukac, and Marek Paralic. Heterogeneous Applications, Tools, and Methodologies in the Car Manufacturing Industry Through an IoT Approach. *IEEE Systems Journal*, PP(99):1–12, September 2015. DOI:10.1109/JSYST.2015.2469681.

[KCPM01] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, 3(3):263–286, August 2001. DOI:10.1007/PL00011669.

[Keo97] Eamonn Keogh. Fast Similarity Search in the Presence of Longitudinal Scaling in Time Series Databases. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, pages 578–584, Newport Beach, CA, USA, November 1997. DOI:10.1109/TAI.1997.632306.

[KLW11] Henning Kagermann, Wolf-Dieter Lukas, and Wolfgang Wahlster. Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution, April 2011. URL:http://www.vdi-nachrichten.com/Technik-Gesellschaft/Industrie-40-Mit-Internet-Dinge-Weg-4-industriellen-Revolution.

[KM06] Lukasz A. Kurgan and Petr Musilek. A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, 21(1):1–24, March 2006. DOI:10.1017/S0269888906000737.

[KRH$^+$14] Henning Kagermann, Frank Riemensperger, Dirk Hoke, Johannes Helbig, Dirk Stocksmeier, Wolfgang Wahlster, August-Wilhelm Scheer, and Dieter Schweer. Smart Service Welt: Recommendations for the Strategic Initiative Web-based Services for Businesses. Technical report, Acatech, March 2014. URL:http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Projekte/Laufende_Projekte/Smart_Service_Welt/BerichtSmartService_engl.pdf.

[KWH13] Henning Kagermann, Wolfgang Wahlster, and Johannes Helbig. Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Final report of the Industrie 4.0 Working Group. Technical report, Acatech, April 2013. URL:http://www.acatech.de/de/publikationen/stellungnahmen/kooperationen/detail/artikel/recommendations-for-implementing-the-strategic-initiative-industrie-40-final-report-of-the-industr.html.

[KWL15] Julian Krumeich, Dirk Werth, and Peter Loos. Prescriptive Control of Business Processes - New Potentials Through Predictive Analytics of Big Data in the Process Manufacturing Industry. *Business & Information Systems Engineering*, 58(4):261–280, December 2015. DOI:10.1007/s12599-015-0412-2.

[Lan01] Doug Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety, February 2001. URL:http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[LBK15] Jay Lee, Behrad Bagheri, and Hung-An Kao. A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3:18–23, January 2015. DOI:10.1016/j.mfglet.2014.12.001.

[LIX14] Xiufeng Liu, Nadeem Iftikhar, and Xike Xie. Survey of Real-time Processing Systems for Big Data. In *Proceedings of the 18th International Database Engineering and Applications Symposium*, pages 356–361, Porto, Portugal, July 2014. DOI:10.1145/2628194.2628251.

[LKY14] Jay Lee, Hung-An Kao, and Shanhu Yang. Service innovation and smart analytics for Industry 4.0 and big data environment. In *Procedia CIRP 16. Product Services Systems and Value Creation. Proceedings of the 6th CIRP Conference on Industrial Product-Service Systems*, pages 3–8, Windsor, Ontario, Canada, May 2014. Elsevier. DOI:10.1016/j.procir.2014.02.001.

[LNR14] David Lechevalier, Anantha Narayanan, and Sudarsan Rachuri. Towards a domain-specific framework for predictive analytics in manufacturing. In *Proceedings of 2014 IEEE International Conference on Big Data*, pages 987–995, Washington, DC, USA, October 2014. DOI:10.1109/BigData.2014.7004332.

[LRPn16] Reinhard Langmann and Leandro F. Rojas-Peña. A PLC as an Industry 4.0 component. In *Proceedings of the 13th International Conference on Remote Engineering and Virtual Instrumentation*, pages 10–15, Madrid, Spain, February 2016. DOI:10.1109/REV.2016.7444433.

[LRU14] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, Cambridge, UK, 2nd edition, 2014.

[Luh58] Hans Peter Luhn. A Business Intelligence System. *IBM Journal of Research and Development*, 2(4):314–319, October 1958.

[Mar06] China Martens. BI at age 17, October 2006. URL:http://www.computerworld.com/article/2554088/business-intelligence/bi-at-age-17.html.

[MB14] Mircea Murar and Stelian Brad. Monitoring and controlling of smart equipments using Android compatible devices towards IoT applications and services in manufacturing industry. In *Proceedings of the 2014 IEEE International Conference on Automation, Quality and Testing, Robotics*, pages 1–5, Cluj-Napoca, Romania, May 2014. DOI:10.1109/AQTR.2014.6857841.

[MCB$^+$11] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung-Byers. Big Data: The next frontier for innovation, competition and productivity. Technical report, McKinsey Global Institute, May 2011. URL:http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[McL13] Charles McLellan. The Evolution of Enterprise Software: An overview, May 2013. URL:http://www.zdnet.com/article/the-evolution-of-enterprise-software-an-overview/.

[MG11] Peter Mell and Timothy Grance. The NIST Definition of Cloud Computing. Technical report, National Institute of Standards and Technology Special Publication 800-145, September 2011. URL:http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, USA, 1st edition, 1997.

[MPF10] Andy Mulholland, Jon Pyke, and Peter Fingar. *Enterprise Cloud Computing*. Meghan-Kiffer Press, 2010.

[MSA15] Thomas Meiren, Nicola Saccani, and Andrea Alghisi. Development of smart services in manufacturing companies. In *Proceedings of the 25th Annual European Association for Research on Services -RESER- Conference 2015*, pages 1–14, Copenhagen, Denmark, September 2015.

[MVF$^+$07] Gérard Morel, Paul Valckenaers, Jean-Marc Faure, Carlos E. Pereira, and Christian Diedrich. Manufacturing plant control challenges and issues. *Control Engineering Practice*, 15(11):1321–1331, November 2007. DOI:10.1016/j.conengprac.2007.05.005.

[MW15] Nathan Marz and James Warren. *Big Data: Principles and best practices of scalable real-time data systems*. Manning Publications Co., 1st edition, April 2015.

[Nat12] National Science and Technology Council. A National Strategic Plan for Advanced Manufacturing. Technical report, National Science and Technology Council, February 2012. URL:https://energy.gov/sites/prod/files/2013/11/f4/nstc_feb2012.pdf.

[Nat16] National Science and Technology Council. National Network for Manufacturing Innovation Program: Strategic Plan. Technical report, National Science and Technology Council, February

2016. URL:https://www.manufacturing.gov/files/2016/02/2015-NNMI-Strategic-Plan.pdf.

[Ni15] Wei-tao Ni. Evolution Analysis of Value Chain in the Process of Manufacturing Servitization. In *Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation*, Tianjin, China, July 2015. DOI:10.2991/978-94-6239-148-2_91.

[Nn15] Mikel Niño. Chronology of antecedents, origin and development of Big Data, September 2015. URL:http://www.mikelnino.com/2016/03/chronology-big-data.html.

[NnBI15] Mikel Niño, José Miguel Blanco, and Arantza Illarramendi. Business Understanding, Challenges and Issues of Big Data Analytics for the Servitization of a Capital Equipment Manufacturer. In *Proceedings of the 2015 IEEE International Conference on Big Data*, pages 1368–1377, Santa Clara, CA, USA, October 2015. DOI:10.1109/BigData.2015.7363897.

[NnI15] Mikel Niño and Arantza Illarramendi. Understanding Big Data: Antecedents, Origin and Later Development. *Dyna New Technologies*, 2(1:14):1–8, January 2015. DOI:10.6036/NT7835.

[NnSBI16] Mikel Niño, Fernando Sáenz, José Miguel Blanco, and Arantza Illarramendi. Requirements for a Big Data capturing and integration architecture in a distributed manufacturing scenario. In *Proceedings of the 2016 IEEE 14th International Conference on Industrial Informatics*, pages 1326–1329, Poitiers, France, July 2016. DOI:10.1109/INDIN.2016.7819372.

[Obj11] Object Management Group. Business Process Model and Notation (BPMN) Version 2.0. Technical report, Object Management Group, January 2011. URL:http://www.omg.org/spec/BPMN/2.0/.

[OJS$^+$16] Boris Otto, Jan Jrjens, Jochen Schon, Sren Auer, Nadja Menz, Sven Wenzel, and Jan Cirullies. Industrial Data Space: Digital Sovereignity over Data. Technical report, Fraunhofer Institute, 2016. URL:https://www.fraunhofer.de/content/dam/zv/en/fields-of-research/industrial-data-space/whitepaper-industrial-data-space-eng.pdf.

[OK10] David L. Olson and Subodh Kesharwani. Enterprise Information System Trends. In *Proceedings of the 12th International Conference on Enterprise Information Systems*, pages 3–14, Funchal-Madeira, Portugal, June 2010. DOI:10.1007/978-3-642-19802-1_1.

[O'L14] Daniel E. O'Leary. Embedding AI and Crowdsourcing in the Big Data Lake. *IEEE Intelligent Systems*, 29(5):70–73, October 2014. DOI:10.1109/MIS.2014.82.

[OLBO15] Peter O'Donovan, K. Leahy, K. Bruton, and D. T. J. O'Sullivan. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of Big Data*, 2(1):25:1–26, November 2015. DOI:10.1186/s40537-015-0034-z.

[OP10] Alexander Osterwalder and Yves Pigneur. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers.* Wiley, 2010.

[Ost04] Alexander Osterwalder. *The Business Model Ontology: a proposition in a design science approach.* PhD thesis, Université de Lausanne, Lausanne, Switzerland, 2004.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, January 1998. URL:http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

[PG08] Naveen Prakash and Anjana Gosain. An approach to engineering the requirements of data warehouses. *Requirements Engineering*, 13(1):49–72, January 2008. DOI:10.1007/s00766-007-0057-x.

[PGL12] Yongtae Park, Youngjung Geum, and Hakyeon Lee. Toward integration of products and services: Taxonomy and typology. *Journal of Engineering and Technology Management*, 29(4):528–545, December 2012. DOI:10.1016/j.jengtecman.2012.08.002.

[Pla16] Plattform Industrie 4.0. Reference Architectural Model Industrie 4.0 (RAMI4.0) - An Introduction. Technical report, Plattform Industrie 4.0, October 2016. URL:http://www.plattform-i40.de/I40/Redaktion/EN/Downloads/Publikation/rami40-an-introduction.html.

[Pou97] Athanasia Pouloudi. Stakeholder Analysis as a Front-End to Knowledge Elicitation. *AI & Society*, 11(1):122–137, March 1997. DOI:10.1007/BF02812443.

[Pre11] President's Council of Advisors on Science and Technology. Report to the President on Ensuring American Leadership in Advanced Manufacturing. Technical report, President's Council of Advisors on Science and Technology, June 2011. URL:https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-advanced-manufacturing-june2011.pdf.

[Pro13] Project Management Institute. *A Guide to the Project Management Body of Knowledge (PMBOK Guide).* Project Management Institute, 5th edition, 2013.

[PS91] Gregory Piatetsky-Shapiro. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine*, 11(5):68–70, January 1991. DOI:10.1609/aimag.v11i4.873.

[PS14] Gregory Piatetsky-Shapiro. CRISP-DM, still the top methodology for analytics, data mining, or data science projects, October 2014. URL:http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[PVK$^+$04] Themistoklis Palpanas, Michail Vlachos, Eamonn Keogh, Dimitrios Gunopulos, and Wagner Truppel. Online Amnesic Approximation of Streaming Time Series. In *Proceedings of the 20th International Conference on Data Engineering*, pages 339–349, Boston, MA, USA, April 2004. DOI:10.1109/ICDE.2004.1320009.

[QLT$^+$15] Lin Qiao, Yinan Li, Sahil Takiar, Ziyang Liu, Narasimha Veeramreddy, Min Tu, Ying Dai, Issac Buenrostro, Kapil Surlaker, Shirshanka Das, and Chavdar Botev. Gobblin: Unifying Data Ingestion for Hadoop. *Proceedings of the VLDB Endowment*, 8(12):1764–1769, August 2015. DOI:10.14778/2824032.2824073.

[RC67] A. H. Robinson and Colin Cherry. Results of a Prototype Television Bandwidth Compression Scheme. *Proceedings of the IEEE*, 55(3):365–364, March 1967. DOI:10.1109/PROC.1967.5493.

[Rie11] Eric Ries. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* Crown Business, 1st edition, 2011.

[RTKM16] Tilmann Rabl, Jonas Traub, Asterios Katsifodimos, and Volker Markl. Apache Flink in current research. *Information Technology*, 58(4):157–165, August 2016. DOI:10.1515/itit-2016-0005.

[SBS16] Olena Skarlat, Michael Borkowski, and Stefan Schulte. Towards a Methodology and Instrumentation Toolset for Cloud Manufacturing. In *Proceedings of the 1st International Workshop on Cyber-Physical Production Systems*, Vienna, Austria, April 2016. DOI:10.1109/CPPS.2016.7483920.

[SGWR14] Diego Salazar, Gerardo Glorioso, Markus Wabner, and Martin Riedel. Maintenance Support Wireless System for Ram of Forming Presses. In *Proceedings of Maintenance Performance Measurement and Management Conference 2014*, pages 89–93, Coimbra, Portugal, September 2014. DOI:10.14195/978-972-8954-42-0_13.

[She00] Colin Shearer. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22, 2000.

[SL17] Patrick Schaefer and Ulf Leser. Fast and Accurate Time Series Classification with WEASEL, January 2017. URL:https://arxiv.org/abs/1701.07681.

[Sma11] Smart Manufacturing Leadership Coalition. Implementing 21st Century Smart Manufacturing: Workshop Summary Report. Technical report, Smart Manufacturing Leadership Coalition, June 2011. URL:https://smartmanufacturingcoalition.org/sites/default/files/implementing_21st_century_smart_manufacturing_report_2011_0.pdf.

[SS99] Zbigniew R. Struzik and Arno Siebes. The Haar Wavelet Transform in the Time Series Similarity Paradigm. In *Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery*, pages 12–22, Prague, Czech Republic, September 1999. DOI:10.1007/978-3-540-48247-5_2.

[SS13] Nitin Sawant and Himanshu Shah. Big Data Ingestion and Streaming Patterns. In *Big Data Application Architecture Q&A*. Apress, Berkeley, CA, December 2013. DOI:10.1007/978-1-4302-6293-0_3.

[Sti74] Stephen M. Stigler. Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Historia Mathematica*, 1(4):431–439, November 1974. DOI:10.1016/0315-0860(74)90033-0.

[SWW11] Florian Stroh, Robert Winter, and Felix Wortmann. Method Support of Information Requirements Analysis for Analytical Information Systems: State of the Art, Practice Requirements, and Research Agenda. *Business & Information Systems Engineering*, 3(1):33–43, February 2011. DOI:10.1007/s12599-010-0138-0.

[SYM$^+$15] Suraksha S. Setty, Humaa Yaqoob, Avinash Malik, Kevin I-Kai Wang, Zoran Salcic, Heejong Park, and Udayanto Dwi Atmojo. A Unified Framework for the Design of Distributed Cyber-Physical Systems Industrial Automation Example. In *Proceedings of the IEEE 10th Conference on Industrial Electronics and Applications*, pages 996–1002, Auckland, New Zealand, June 2015. DOI:10.1109/ICIEA.2015.7334253.

[Tri09] Rainer Trinczek. How to Interview Managers? Methodical and Methodological Aspects of Expert Interviews as a Qualitative Method in Empirical Social Research. In *Interviewing Experts*, pages 203–216. Palgrave Macmillan, Basingstoke, UK, 2009. DOI:10.1057/9780230244276_10 10.1057/9780230244276_10.

[TZXZ14] Fei Tao, Ying Zuo, Li Da Xu, and Lin Zhang. IoT-Based Intelligent Perception and Access of Manufacturing Resource Toward Cloud Manufacturing. *IEEE Transactions on Industrial Informatics*, 10(2):1547–1557, May 2014. DOI:10.1109/TII.2014.2306397.

[uRCBW16] Muhammad Habib ur Rehman, Victor Chang, Aisha Batool, and Teh Ying Wah. Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management*, 36(6):917–928, December 2016. DOI:10.1016/j.ijinfomgt.2016.05.013.

[vdL15] Rick van der Lans. Big Data Myth 2: Analytics Requires Big data, September 2015. URL:http://searchdatamanagement.techtarget.com/blog/The-Wondrous-World-of-Data/Big-Data-Myth-2-Analytics-Requires-Big-data.

[VHFST15] Birgit Vogel-Heuser, Alexander Fay, Ina Schaefer, and Matthias Tichy. Evolution of software in automated production systems:

Challenges and research directions. *Journal of Systems and Software*, 110:54–84, December 2015. DOI:10.1016/j.jss.2015.08.026.

[VR88] Sandra Vandermerwe and Juan Rada. Servitization of business: Adding value by adding services. *European Management Journal*, 6(4):314–324, December 1988. DOI:10.1016/0263-2373(88)90033-3.

[WE16] Michael Weyrich and Christof Ebert. Reference Architectures for the Internet of Things. *IEEE Software*, 33(1):112–116, February 2016. DOI:10.1109/MS.2016.20.

[Wel84] Terry A. Welch. A Technique for High-Performance Data Compression. *Computer*, 17(6):8–19, June 1984. DOI:10.1109/MC.1984.1659158.

[WMD$^+$13] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, March 2013. DOI:10.1007/s10618-012-0250-5.

[WTS$^+$16] Jiafu Wan, Shenglong Tang, Zhaogang Shu, Di Li, Shiyong Wang, Muhammad Imran, and Athanasios V. Vasilakos. Software-Defined Industrial Internet of Things in the Context of Industry 4.0. *IEEE Sensors Journal*, 16(20):7373–7380, October 2016. DOI:10.1109/JSEN.2016.2565621.

[XHL14] Li Da Xu, Wu He, and Shancang Li. Internet of Things in Industries: A Survey. *IEEE Transactions on Industrial Informatics*, 10(4):2233–2243, November 2014. DOI:10.1109/TII.2014.2300753.

[XJW$^+$14] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren. Information Security in Big Data: Privacy and Data Mining. *IEEE Access*, 2:1149–1176, 2014. DOI:10.1109/ACCESS.2014.2362522.

[YPC$^+$14] Hanna Yang, Minjeong Park, Minsu Cho, Minseok Song, and Seongjoo Kim. A System Architecture for Manufacturing Process Analysis based on Big Data and Process Mining Techniques. In *Proceedings of the 2014 IEEE International Conference on Big Data*, pages 1024–1029, Washington DC, USA, October 2014. DOI:10.1109/BigData.2014.7004336.

[ZBO$^+$14] EL Moukhtar Zemmouri, Hicham Behja, Brahim Ouhbi, Brigitte Trousse, Abdelaziz Marzak, and Youssef Benghabrit. Goal Driven Approach to Model Interaction between Viewpoints of a Multi-view KDD Process. *Journal of Mobile Multimedia*, 9(3-4):214–229, March 2014.

[ZC05] Didar Zowghi and Chad Coulin. Requirements Elicitation: A Survey of Techniques, Approaches, and Tools. In *Engineering and Managing Software Requirements*, pages 19–46. Springer-Verlag, 2005. DOI:10.1007/3-540-28244-0_2.

[ZCF+10] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster Computing withWorking Sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, pages 10:1–7, Boston, MA, USA, June 2010.

[ZLT+14] Lin Zhang, Yongliang Luo, Fei Tao, Bo Hu Li, Lei Ren, Xuesong Zhang, Hua Guo, Ying Cheng, Anrui Hu, and Yongkui Liu. Cloud manufacturing: A new manufacturing paradigm. *Enterprise Information Systems*, 8(2):167–187, 2014. DOI:10.1080/17517575.2012.683812.

[Zue11] Lara Zuehlke. Enterprise Software History, Part 4: Dotcom to Today, September 2011. URL:http://blog.softwareadvice.com/articles/enterprise/software-history-part-4-109142011/.

[ZZCW10] Shufen Zhang, Shuai Zhang, Xuebin Chen, and Shangzhuo Wu. Analysis and Research of Cloud Computing System Instance. In *Proceedings of the 2nd International Conference on Future Networks*, pages 88–92, Sanya, Hainan, China, January 2010. DOI:10.1109/ICFN.2010.60.

[ZZW+15] Yingfeng Zhang, Geng Zhang, Junqiang Wang, Shudong Sun, Shubin Si, and Teng Yang. Real-time information capturing and integration framework of the internet of manufacturing things. *International Journal of Computer Integrated Manufacturing*, 28(8):811–822, 2015. DOI:10.1080/0951192X.2014.900874.

List of Figures

List of Tables

Appendix A

Results of Field Testing of Time-Series Data Reduction

This Appendix details the results of the field testing on time-series data reduction techniques which were presented in a summarized way in 5.2.4. This field testing analyzed the application of diverse reduction techniques on the manufacturing time-series data captured from the real-world setting of our case study. An approximate total of 470,000 applications of different versions of parameterized reduction techniques, i.e. a reduction technique with a specific value for the parameter setting the dimensionality for the reduced version of data, were conducted in the field testing. For that purpose, the 314 analyzed indicators were grouped into eight families, according to their basic syntactic features and the registered magnitude:

1. Discrete binary (DB).
2. Discrete n-ary - Subgroup 1 (DN-1).
3. Discrete n-ary - Subgroup 2 (DN-2).
4. Continuous, product-undriven (CPU).
5. Continuous, product-driven - Subgroup 1 (CPD-1).
6. Continuous, product-driven - Subgroup 1 (CPD-2).
7. Continuous, product-driven - Subgroup 1 (CPD-3).
8. Continuous, product-driven - Subgroup 1 (CPD-4).

The main results of these tests were presented along 5.2.4, indicating the best overall compression ratio in disk (*COMPD*) for each time series family and the reduction technique offering the best results. In this Appendix the results for each one of these eight time-series families is presented with further detail.

For families DB, DN-1 and DN-2 (discrete time series), it is presented:

- Number of analyzed time series.
- Number of time series grouped by the technique that offers the best *COMPD*.
- Average *COMPD* obtained by each analyzed technique across all time series in that family, including a 95% confidence interval (α=0.05).
- Average *COMPD* for that family, selecting the best *COMPD* obtained for each time series, regardless of the technique obtaining it.
- Average ratio between the *COMPD* obtained by the two analyzed techniques, grouping the cases where each technique (RLE or LZW) obtained the best results.

For families CPU, CPD-1, CPD-2, CPD-3 and CPD-4 (continuous time series), it is presented:

- Number of analyzed time series.
- Number of time series grouped by the technique that offers the best *COMPD*.
- Average *COMPD* obtained by each analyzed technique across all time series in that family, including a 95% confidence interval (α=0.05). The average for each technique is calculated across those time series where that technique obtains a $COMPD < 100\%$ without exceeding a root mean squared error equal to 1% of the average measurement for each indicator (ratio on error, *RTERR*).
- Average *COMPD* for that family, selecting the best *COMPD* obtained for each time series, regardless of the technique obtaining it.

A.1 Discrete binary (DB)

Number of analyzed indicators	146
Segmentation applied to indicators	No
Number of analyzed time series	146

Number of time series where best COMPD is obtained by:	
RLE [1]	146 (100%)
LZW [2]	0 (0%)

COMPD obtained by	Average for all time series	± for a 95% confidence interval
RLE	0.0485%	±0.01%
LZW	0.352%	±0.0132%
Best COMPD, regardless of the technique	**0.0485%**	**±0.01%**

Ratio between	Average for all time series	± for a 95% confidence interval
COMPD by RLE and COMPD by LZW ([1])	12.05%	±1.55%

A.2 Discrete n-ary - Subgroup 1 (DN-1)

Number of analyzed indicators	25
Segmentation applied to indicators	No
Number of analyzed time series	25

Number of time series where best COMPD is obtained by:	
RLE [1]	25 (100%)
LZW [2]	0 (0%)

COMPD obtained by	Average for all time series	± for a 95% confidence interval
RLE	0.0127%	±0.006%
LZW	0.3631%	±0.0221%
Best COMPD, regardless of the technique	**0.0127%**	**±0.006%**

Ratio between	Average for all time series	± for a 95% confidence interval
COMPD by RLE and COMPD by LZW ([1])	3.32%	±1.59%

A.3 Discrete n-ary - Subgroup 2 (DN-2)

Number of analyzed indicators	60
Segmentation applied to indicators	No
Number of analyzed time series	60

Number of time series where best COMPD is obtained by:	
RLE [1]	52 (87%)
LZW [2]	8 (13%)

COMPD obtained by	Average for all time series	± for a 95% confidence interval
RLE	0.267%	±0.0528%
LZW	0.4029%	±0.0217%
Best COMPD, regardless of the technique	**0.2488%**	**±0.0411%**

Ratio between	Average for all time series	± for a 95% confidence interval
COMPD by RLE and COMPD by LZW ([1])	50.68%	±6.62%
COMPD by LZW and COMPD by RLE ([2])	82.86%	±9.02%

A.4 Continuous, product-undriven (CPU)

Number of analyzed indicators	31
Segmentation applied to indicators	No
Number of analyzed time series	31

Number of time series where best COMPD is obtained by:	
PIP	27 (87.1%)
SAM	3 (9.68%)
CHEB	1 (3.23%)

COMPD obtained without exceeding the 1% threshold for RTERR, by	Average for all time series	± for a 95% confidence interval
PIP	0.0031%	±0.0013%
SAM	0.0036%	±0.0021%
CHEB	0.004%	±0.0021%
PRE	0.0042%	±0.0023%
PLR	0.0045%	±0.002%
PAA	0.0059%	±0.0025%
APCA	0.1516%	±0.0918%
DWT	1.4587%	±1.5561%
Best COMPD, regardless of the technique	**0.002987%**	**±0.0013%**

A.5 Continuous, product-driven - Subgroup 1 (CPD-1)

Number of analyzed indicators	32
Segmentation applied to indicators	Yes
Number of analyzed time series	120

Number of time series where best COMPD is obtained by:	
PIP	58 (48.33%)
PAA	44 (36.67%)
CHEB	9 (7.5%)
PRE	9 (7.5%)

COMPD obtained without exceeding the 1% threshold for RTERR, by	Average for all time series	± for a 95% confidence interval
PIP	0.0152%	±0.0023%
CHEB	0.0202%	±0.0028%
PRE	0.0215%	±0.0035%
PAA	0.0399%	±0.0077%
PLR	0.0575%	±0.0112%
APCA	0.0724%	±0.0147%
SAM	0.1934%	±0.0427%
DWT	7.851%	±1.571%
Best COMPD, regardless of the technique	**0.014%**	**±0.002%**

A.6 Continuous, product-driven - Subgroup 2 (CPD-2)

Number of analyzed indicators	9
Segmentation applied to indicators	Yes
Number of analyzed time series	1109

Number of time series where best COMPD is obtained by:	
PIP	490 (44.18%)
PAA	391 (35.26%)
SAM	102 (9.2%)
CHEB	64 (5.77%)
APCA	32 (2.89%)
PRE	24 (2.16%)
PLR	6 (0.54%)

COMPD obtained without exceeding the 1% threshold for RTERR, by	Number of time series where COMPD<100% is obtained	% over total time series (1109)	Average COMPD for those time series	± for a 95% confidence interval
PAA	821	74.03%	38.9657%	±1.9152%
PIP	923	83.23%	45.0202%	±2.5137%
APCA	622	56.09%	64.6701%	±3.0545%
CHEB	720	64.92%	73.9902%	±3.141%
PRE	730	65.83%	77.7457%	±2.9167%
PLR	676	60.96%	80.7685%	±2.9607%
SAM	894	80.61%	86.8854%	±3.9139%
DWT	-	-	-	-
Best COMPD, regardless of the technique	1109	100%	**36.5327%**	**±1.7125%**

A.7 Continuous, product-driven - Subgroup 3 (CPD-3)

Number of analyzed indicators	3
Segmentation applied to indicators	Yes
Number of analyzed time series	277

Number of time series where best COMPD is obtained by:	
PIP	261 (94.22%)
PRE	7 (2.53%)
SAM	6 (2.17%)
CHEB	3 (1.08%)

COMPD obtained without exceeding the 1% threshold for RTERR, by	Number of time series where COMPD<100% is obtained	% over total time series (277)	Average COMPD for those time series	± for a 95% confidence interval
PIP	271	97.83%	39.917%	±2.2534%
APCA	115	41.52%	59.8317%	±5.0286%
SAM	182	65.7%	61.5722%	±5.8541%
CHEB	130	46.93%	64.9303%	±7.9349%
PAA	182	65.7%	68.3645%	±5.3182%
PRE	222	80.14%	70.556%	±5.5584%
PLR	190	68.59%	82.9987%	±5.3077%
DWT	-	-	-	-
Best COMPD, regardless of the technique	277	100%	**41.0993%**	**±2.477%**

A.8 Continuous, product-driven - Subgroup 4 (CPD-4)

Number of analyzed indicators	8
Segmentation applied to indicators	Yes
Number of analyzed time series	16

Number of time series where best COMPD is obtained by:	
PIP	16 (100%)

COMPD obtained without exceeding the 1% threshold for RTERR, by	Average for all time series	± for a 95% confidence interval
PIP	1.0247%	±0.545%
PRE	1.535%	±0.6817%
CHEB	1.7183%	±0.7834%
SAM	1.8856%	±0.5648%
PLR	1.9501%	±0.8431%
PAA	2.5778%	±1.441%
APCA	2.6492%	±1.359%
DWT	23.3481%	±11.4888%
Best COMPD, regardless of the technique	**1.0247%**	**±0.002%**

Appendix B

Internal Design of the *Reduction Analysis Planner*

This Appendix presents the internal design of the interaction among the main components of the *Reduction Analysis Planner* (RAP) described in 5.3.2. The interaction is represented using UML sequence diagrams for the following internal processes:

1. Time series loading process (figure B.1).
2. Syntactic characterization process (figure B.2).
3. Reduction recommendation process (figure B.3).
4. Reduction plan scheduling process (figure B.4).

Besides, it is detailed a potential implementation schema of the most relevant functions in those components. These functions, which are highlighted in red in the UML sequence diagrams, are the following:

- *characterizeTSs()* and *characterize()*, in the Syntactic Characterization Module.
- *getTimeSeriesFamily()* and *getReductionInformation()*, in the Matcher included in the Recommendation Module.
- *classify()*, in the Time Series Classifier of the Syntactic Characterization Knowledge Base included in the Recommendation Module.
- *getReductionTechniques()*, in the Recommendation Module.
- *getPlanification()*, in the Plan Scheduling Module.

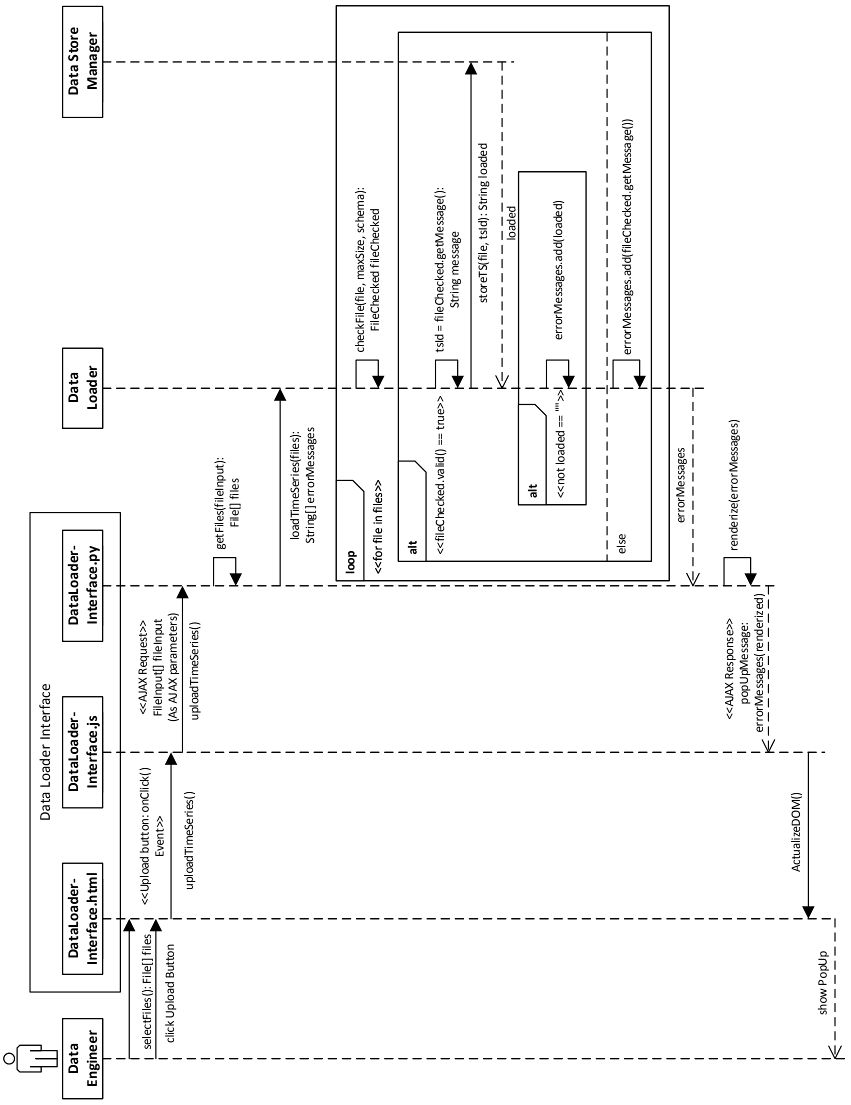

Data Engineer
Data Loader Interface
DataLoader-Interface.html
DataLoader-Interface.js
DataLoader-Interface.py
Data Loader
Data Store Manager
selectFiles(): File[] files
click Upload Button
<<Upload button: onClick() Event>>
uploadTimeSeries()
<<AJAX Request>> FileInput[] fileInput (As AJAX parameters)
uploadTimeSeries()
getFiles(fileInput): File[] files
loadTimeSeries(files): String[] errorMessages
loop
<<for file in files>>
checkFile(file, maxSize, schema): FileChecked fileChecked
alt
<<fileChecked.valid() == true>>
tsId = fileChecked.getMessage(): String message
storeTS(file, tsId): String loaded
loaded
alt
<<not loaded == "">>
errorMessages.add(loaded)
else
errorMessages.add(fileChecked.getMessage())
errorMessages
renderize(errorMessages)
<<AJAX Response>> popUpMessage: errorMessages(renderized)
ActualizeDOM()
show PopUp

Figure B.1: Sequence diagram for the time series loading process

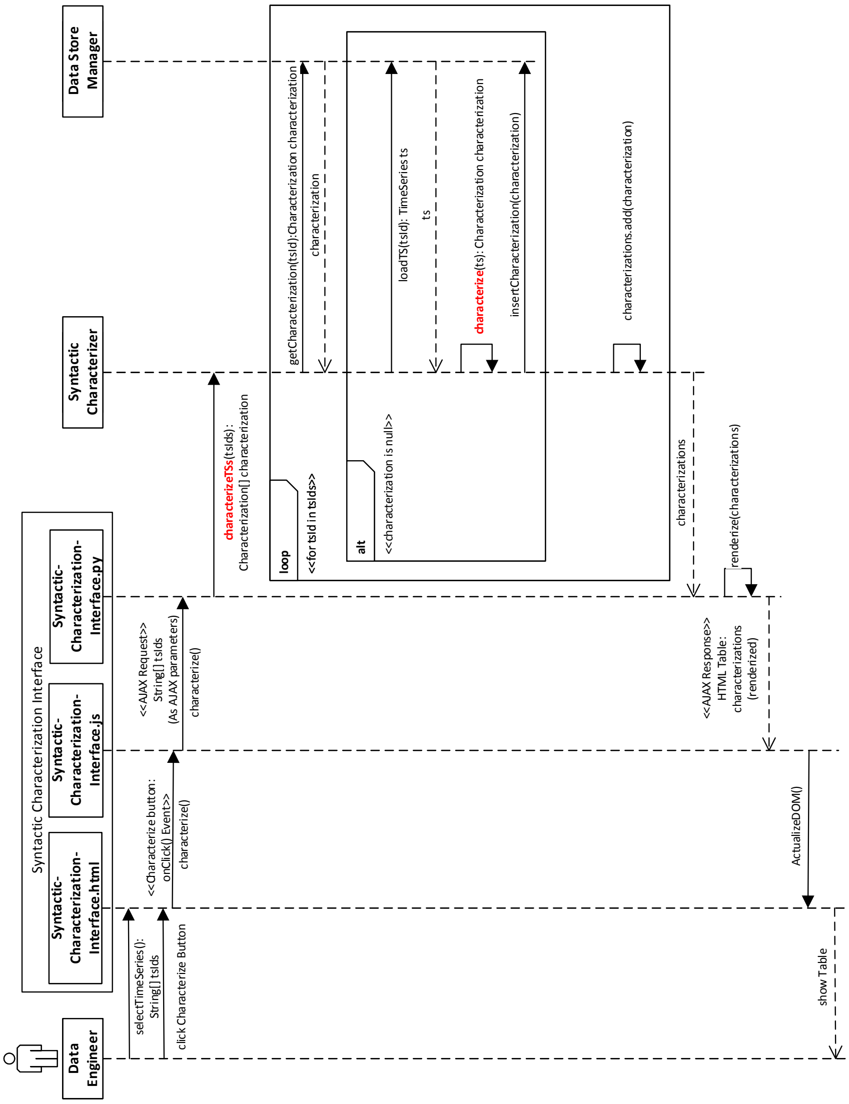

Data Engineer
Syntactic Characterization Interface
Syntactic-Characterization-Interface.html
Syntactic-Characterization-Interface.js
Syntactic-Characterization-Interface.py
Syntactic Characterizer
Data Store Manager
selectTimeSeries(): String[] tsIds
click Characterize Button
<<Characterize button: onClick() Event>> characterize()
<<AJAX Request>> String[] tsIds (As AJAX parameters) characterize()
characterizeTSs(tsIds): Characterization[] characterization
loop
<<for tsId in tsIds>>
getCharacterization(tsId):Characterization characterization
characterization
alt
<<characterization is null>>
loadTS(tsId): TimeSeries ts
ts
characterize(ts): Characterization characterization
insertCharacterization(characterization)
characterizations.add(characterization)
characterizations
renderize(characterizations)
<<AJAX Response>> HTML Table: characterizations (renderized)
ActualizeDOM()
show Table

Figure B.2: Sequence diagram for the syntactic characterization process

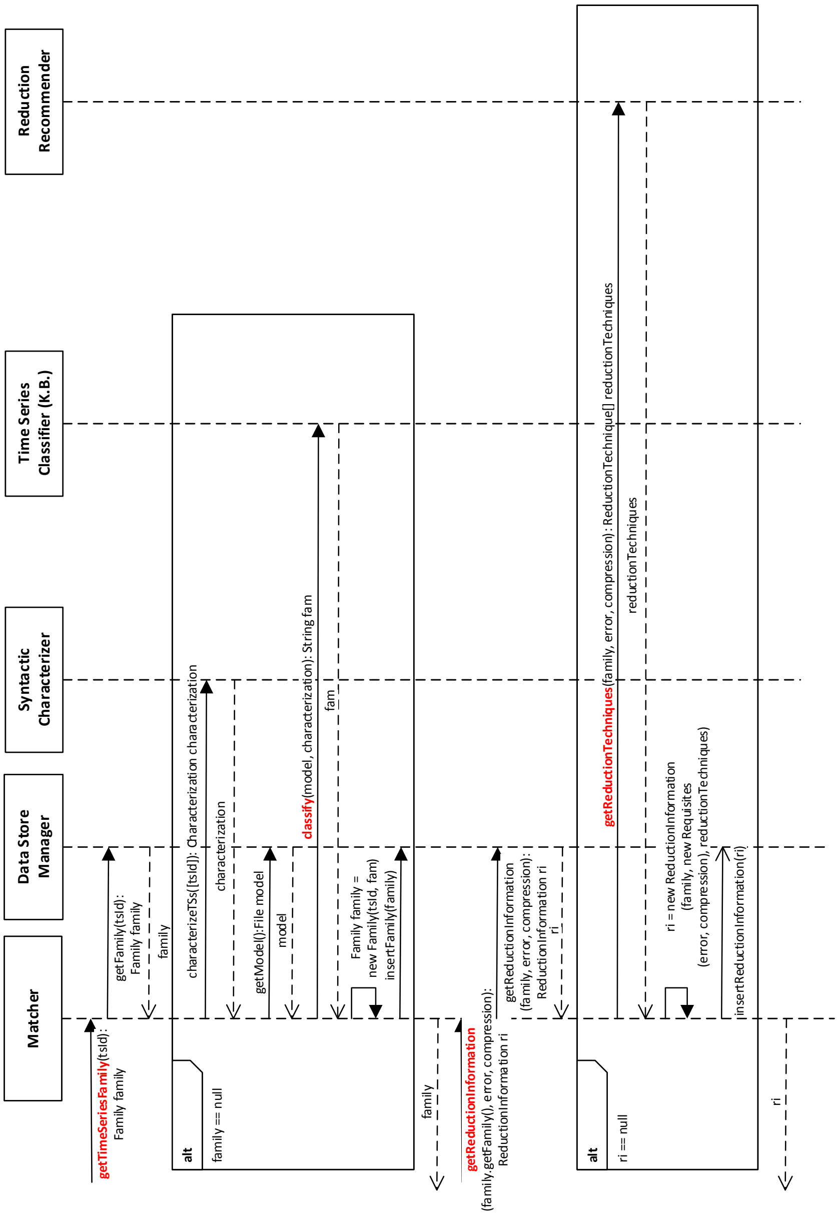

Matcher
Data Store Manager
Syntactic Characterizer
Time Series Classifier (K.B.)
Reduction Recommender
getTimeSeriesFamily(tsId): Family family
getFamily(tsId): Family family
family
alt
family == null
characterizeTSs({tsId}): Characterization characterization
characterization
getModel():File model
model
classify(model, characterization): String fam
fam
Family family = new Family(tsId, fam)
insertFamily(family)
family
getReductionInformation (family.getFamily(), error, compression): ReductionInformation ri
getReductionInformation (family, error, compression): ReductionInformation ri
ri
alt
ri == null
getReductionTechniques(family, error, compression): ReductionTechnique[] reductionTechniques
reductionTechniques
ri = new ReductionInformation (family, new Requisites (error, compression), reductionTechniques)
insertReductionInformation(ri)
ri

Figure B.3: Sequence diagram for the reduction recommendation process

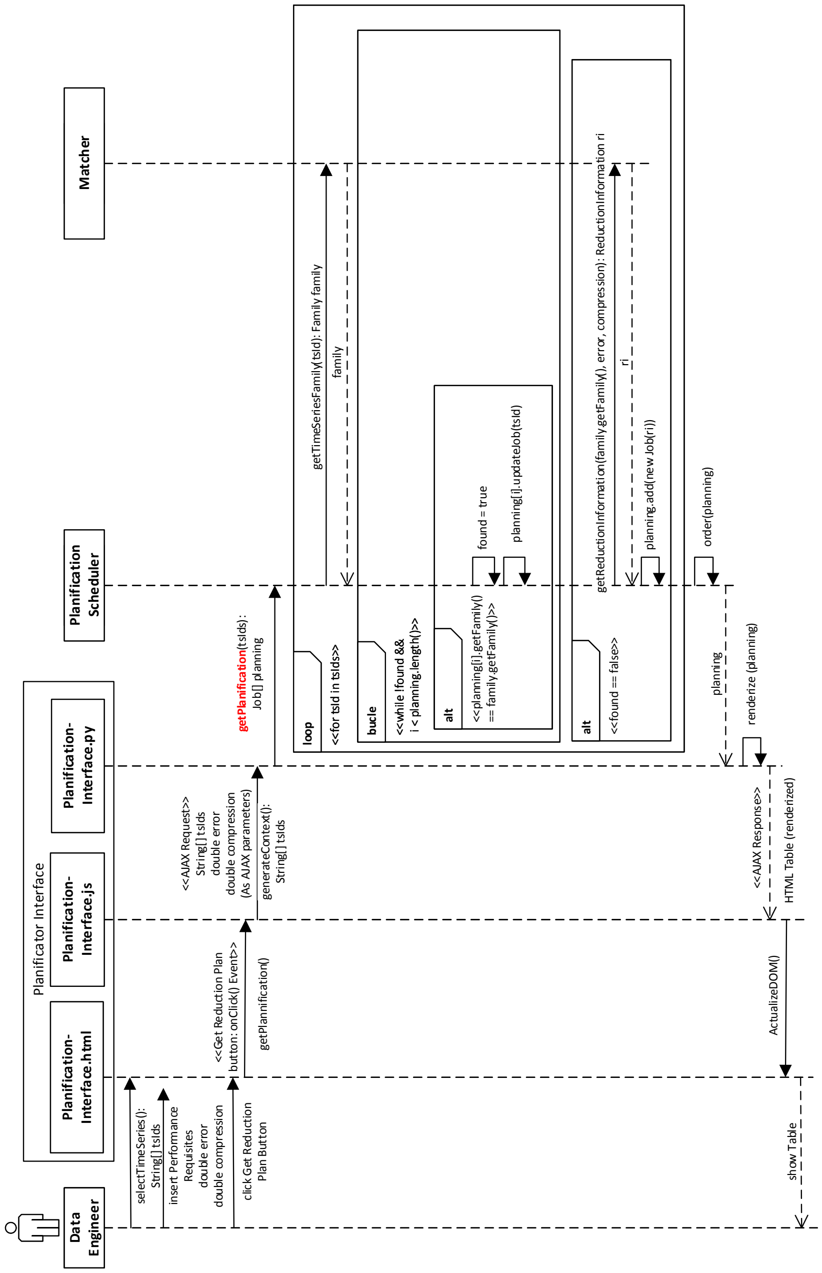

Data Engineer
Planificator Interface
Planification-Interface.html
Planification-Interface.js
Planification-Interface.py
Planification Scheduler
Matcher
selectTimeSeries(): String[] tsIds
insert Perfomance Requisites double error double compression
click Get Reduction Plan Button
<<Get Reduction Plan button: onClick() Event>>
getPlannification()
<<AJAX Request>> String[] tsIds double error double compression (As AJAX parameters)
generateContext(): String[] tsIds
getPlanification(tsIds): Job[] planning
loop
<<for tsId in tsIds>>
getTimeSeriesFamily(tsId): Family family
family
bucle
<<while !found && i < planning.length()>>
alt
<<planning[i].getFamily() == family.getFamily()>>
found = true
planning[i].updateJob(tsId)
alt
<<found == false>>
getReductionInformation(family.getFamily(), error, compression): ReductionInformation ri
ri
planning.add(new Job(ri))
order(planning)
planning
renderize (planning)
<<AJAX Response>> HTML Table (renderized)
ActualizeDOM()
show Table

Figure B.4: Sequence diagram for the reduction plan scheduling process

characterizeTSs()	
Input	String[ ] tsIds
Output	Characterization[ ] characterizations

```
Characterization[] characterizations = []
for tsId in tsIds:
  Characterization characterization = DataStoreManager.
      getCharacterization(tsId)
  if characterization is null:
    TimeSeries ts = DataStoreManager.loadTS(tsId)
    characterization = characterize(ts)
    DataStoreManager.insertCharacterization(characterization)
  characterizations.add(characterization)
return characterizations
```

characterize()	
Input	TimeSeries ts
Output	Characterization characterization

```
String[] characteristics = DataStoreManager.getModelCharacteristics
    ()
Characterization characterization = new Characterization(tsId)
for characteristic in characteristics:
  f = getFunction(characteristic)
  value = execute(f, ts)
  Characteristic c = new Characteristic(characteristic, value)
  characterization.add(c) #Adds the characteristic to the
      characterization
return characterization
```

getTimeSeriesFamily()	
Input	String tsId
Output	String family

```
Family family = DataStoreManager.getFamily(tsId)
if family == null:
  Characterization characterization = SyntacticCharacterizer.
      characterizeTSs([tsId])[1] #Needs a list as parameter
  File model = DataStoreManager.getModel()
  String fam = TimeSeriesClasificator.classify(characterization,
      model)
  Family family = new Family(tsId, fam)
  DataStorage.insertFamily(family)
return family
```

getReductionInformation()	
Input	String family double error double compression
Output	ReductionInformation reductionInformation

```
ReductionInformation ri = DataStoreManager.getRedutionInformation(
    family, error, compression)
if ri == null:
  ReductionTechniques [] reductionTechniques = ReductionRecommender.
      getRedutionTechniques(family, error, compression)
  ri = new ReductionInformation(family, new Requisites(error,
      compression), reductionTechniques)
  DataStoreManager.insertReductionInformation(ri)
return ri
```

classify()	
Input	Characterization characterization Model model
Output	String family

```
family = ""
Instance instance = new Instance()
instance.attributes = []
Characteristic[] characteristics = Characterization.
    getCharacteristics()
for c in characteristics:
  instance.attributes.add(c.getValue()
try:
  model = loadModel(model)
  family = model.classify(instance)
except:
  #Exception handling
finally:
  return family
```

getReductionTechniques()	
Input	String family double error double compression
Output	ReductionTechnique[ ] reductionTechniques

```
import json
ReductionTechnique[] reductionTechniques = []
db = connect2DB()
if db.exists("KnowledgeBase")
  ReductionInformation reductionInformation = db.get("KnowledgeBase"
      , {family= family, threshold <= error, orderBy = threshold})
      [1]
  JSONArray techniques = reductionInformation.get("techniques")
  for i in 1:techniques.lenght():
    JSONObject t = techniques[i]
    reductionInformation.add(new ReductionTechnique(t.get("name"), t
        .get("param"), t.get("error"), t.get("compression"),
        compression)
db.close()
return reductionTechniques
```

getPlanification()	
Input	String[ ] tsIds
Output	Job[ ] planning

```
Job[] planning = Job[]
ReductionInformation ris = []
  for tsId in tsIds:
    Family family = Matcher.getTimeSeriesFamily(tsId)
    #addToPlanning
    boolean found = false
    i = 0
    while !found && i < planning.length()
      if planning[i].getFamily() == family.getFamily():
        found = true
        planning[i].updateJob(tsId)
      i++
    if !found:
      ReductionInformation ri = Matcher.getReductionInformation(
          family.getFamily(), PlannificationInterface.error,
          PlannificationInterface.compression)
      planning.add(new Job(ri)
order(planning)
return planning
```

Appendix C

Internal Design of the *Reduction Analysis Executor*

This Appendix presents the internal design of the interaction among the main components of the *Reduction Analysis Executor* (RAE) described in 5.3.3. This interaction is represented using UML sequence diagrams (figures C.1 and C.2) for the assisted execution process of the reduction analysis plan.

Besides, it is detailed a potential implementation schema of the most relevant functions in those components. These functions, which are highlighted in red in the UML sequence diagrams, are the following:

- *generateContext()*, in the Data Loading Module.
- *executePlan()*, in the Reduction and Reconstruction Engine.
- *evaluate()* and *evaluateAverage()*, in the Evaluation Module.

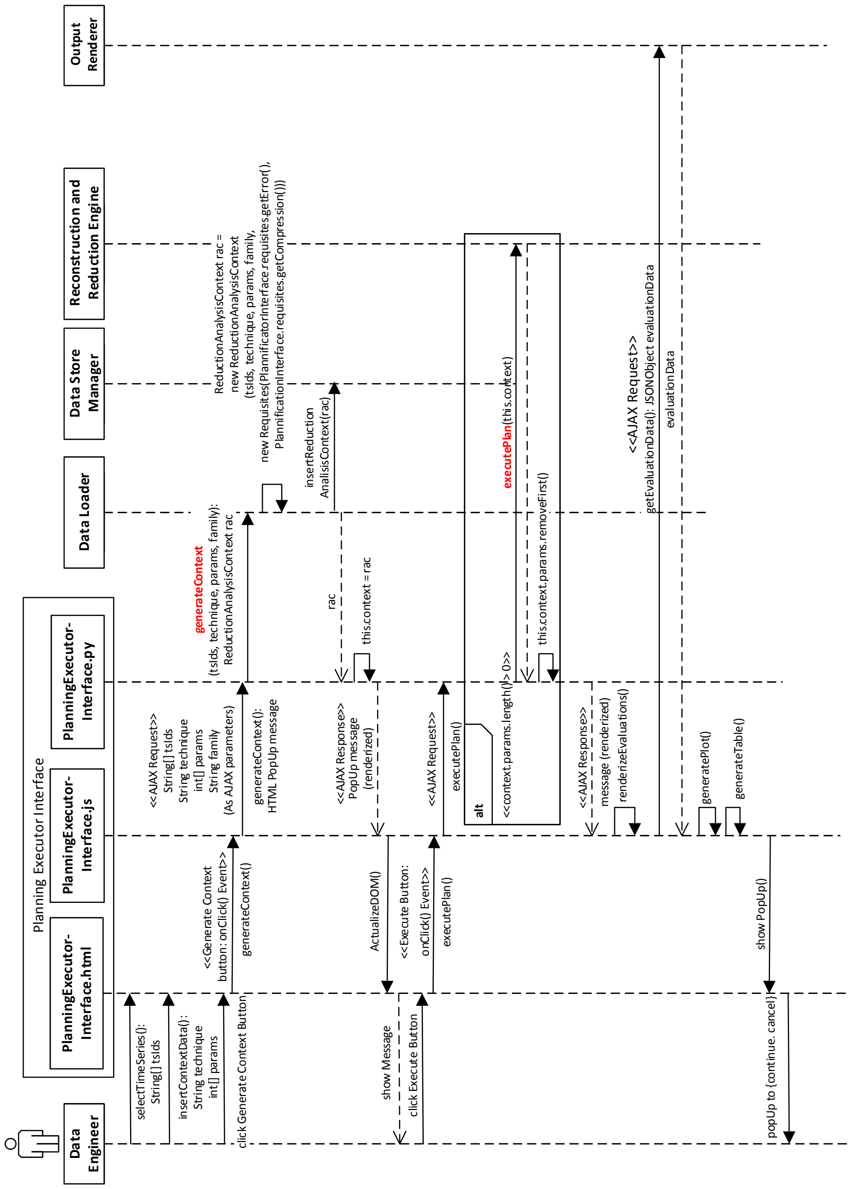

Data Engineer
Planning Executor Interface
PlanningExecutor-Interface.html
PlanningExecutor-Interface.js
PlanningExecutor-Interface.py
Data Loader
Data Store Manager
Reconstruction and Reduction Engine
Output Renderer
selectTimeSeries(): String[] tsIds
insertContextData(): String technique int[] params
click Generate Context Button
<<Generate Context button: onClick() Event>>
generateContext()
<<AJAX Request>> String[] tsIds String technique int[] params String family (As AJAX parameters)
generateContext(): HTML PopUp message
generateContext (tsIds, technique, params, family): ReductionAnalysisContext rac
ReductionAnalysisContext rac = new ReductionAnalysisContext (tsIds, technique, params, family, new Requisites(PlannificatorInterface.requisites.getError(), PlannificationInterface.requisites.getCompression()))
insertReduction AnalisisContext(rac)
rac
this.context = rac
<<AJAX Response>> PopUp message (renderized)
ActualizeDOM()
show Message
click Execute Button
<<Execute Button: onClick() Event>>
executePlan()
<<AJAX Request>>
executePlan()
alt
<<context.params.length() > 0>>
executePlan(this.context)
this.context.params.removeFirst()
<<AJAX Response>>
message (renderized)
renderizeEvaluations()
<<AJAX Request>>
getEvaluationData(): JSONObject evaluationData
evaluationData
generatePlot()
generateTable()
show PopUp()
popUp to {continue, cancel}

Figure C.1: Sequence diagram for the assisted execution of the reduction analysis plan (Part I)

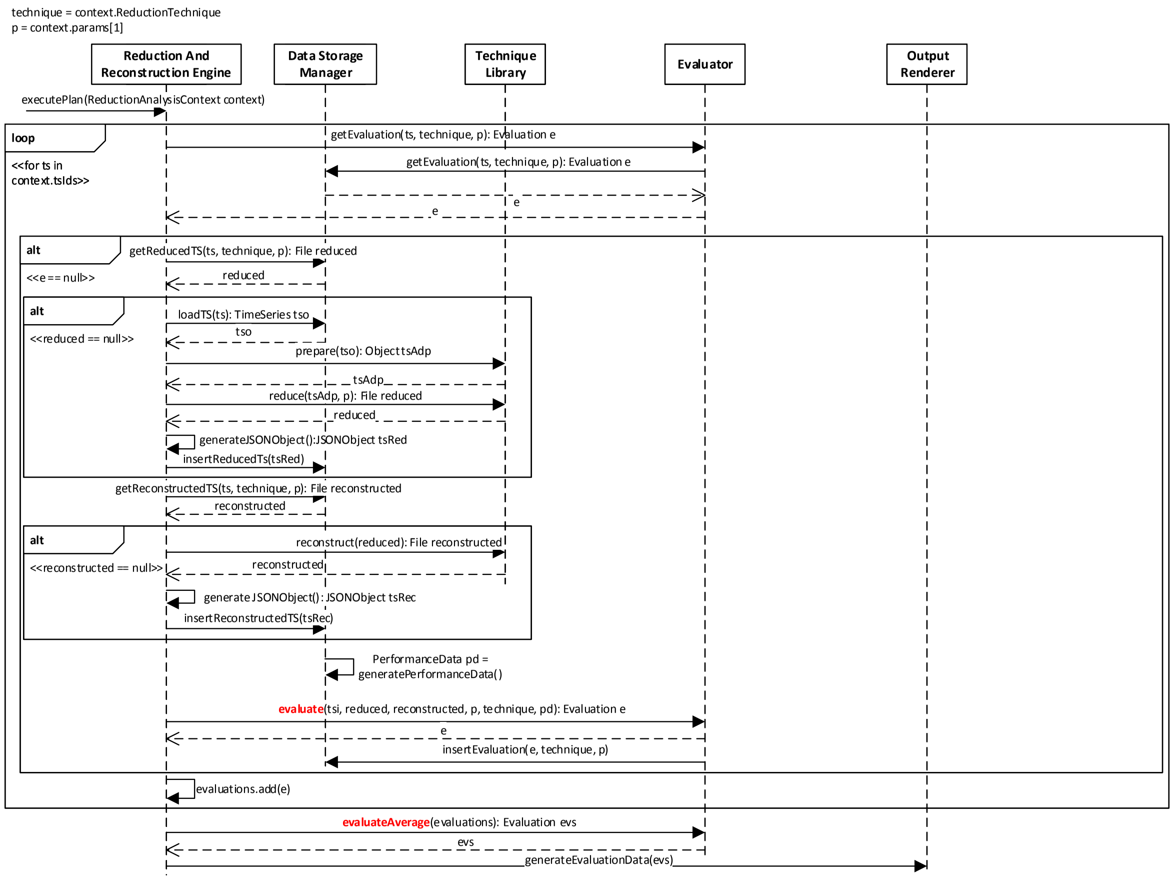

technique = context.ReductionTechnique
p = context.params[1]
Reduction And Reconstruction Engine
Data Storage Manager
Technique Library
Evaluator
Output Renderer
executePlan(ReductionAnalysisContext context)
loop
<<for ts in context.tsIds>>
getEvaluation(ts, technique, p): Evaluation e
getEvaluation(ts, technique, p): Evaluation e
e
e
alt
<<e == null>>
getReducedTS(ts, technique, p): File reduced
reduced
alt
<<reduced == null>>
loadTS(ts): TimeSeries tso
tso
prepare(tso): Object tsAdp
tsAdp
reduce(tsAdp, p): File reduced
reduced
generateJSONObject():JSONObject tsRed
insertReducedTs(tsRed)
getReconstructedTS(ts, technique, p): File reconstructed
reconstructed
alt
<<reconstructed == null>>
reconstruct(reduced): File reconstructed
reconstructed
generate JSONObject() : JSONObject tsRec
insertReconstructedTS(tsRec)
PerformanceData pd = generatePerformanceData()
evaluate(tsi, reduced, reconstructed, p, technique, pd): Evaluation e
e
insertEvaluation(e, technique, p)
evaluations.add(e)
evaluateAverage(evaluations): Evaluation evs
evs
generateEvaluationData(evs)

Figure C.2: Sequence diagram for the assisted execution of the reduction analysis plan (Part II)

generateContext()	
Input	String[ ] tsIds String reductionTechnique int[ ] params String family
Output	ReductionAnalysisContext rac

```
ReductionAnalysisContext rac = null
double error = PlannificationInterface.requisites.getError()
double compression = PlannificationInterface.requisites.
    getCompression()
rac = new ReductionAnalysisContext(tsIds, reductionTechnique, params
    , family, new Requisites(error, compression))
DataStoreManager.insertReductionAnalysisContext(rac)
return rac
```

executePlan()	
Input	ReductionAnalysisContext rac
Output	-

```
String[] tsIds = rac.getTsIds()
Evaluation[] evaluations = []
p = rac.getParams()[1]
Timer t = new Timer()
for ts in tsIds:
  Evaluation e = Evaluator.getEvaluation(ts, reductionTechnique, p)
  if e == null:
    File reduced = DataStoreManager.getReducedTS(ts, reductionTechnique, p)
    if reduced = null:
      prepare = getFunction("prepare", reductionTechnique+".py") # Obtains the ’prepare’ function associated to the selected reduction technique
      TimeSeries tso = DataStoreManager.loadTS(ts)
      Object tsAdp = execute(prepare, tso) #Executes the ’prepare’ function
      reduc = getFunction("reduce", reductionTechnique+".py")
      t.start()
      reduced = execute(reduc, tsAdp, p)
      PerformanceProperty reductionTime = new PerformanceProperty("reductionTime", t.getTime())
      JSONObject tsRed = newJSONObject()
      tsRed.add("reduced", reduced)
      tsRed.add("technique", reductionTechnique)
      tsRed.add("tsId", ts)
      tsRed.add("param", p)
      DataStoreManager.insertReducedTS(tsRed)
    File reconstructed = DataStoreManager.getReconstructedTS(ts, reductionTechnique, p)
    if reconstructed == null:
      reconstruct getFunction("reconstruct", reductionTechnique+".py")
      t.start()
      reconstructed = execute(reconstruct, reduced)
      PerformanceProperty reconstructionTime = new PerformanceProperty("reconstructionTime", t.getTime())
      JSONObject tsRec = newJSONObject()
      tsRec.add("reconstructed", reconstructed)
      tsRec.add("technique", reductionTechnique)
      tsRec.add("tsId", ts)
      tsRec.add("param", p)
      DataStoreManager.insertReducedTS(tsRed)
      PerformanceData pd = new PerformanceData()
      pd.add(reductionTime)
      pd.add(reconstructionTime)
    Evaluation e = Evaluator.evaluate(tso, reduced, reconstructed, p, reductionTechnique, pd)
  evaluations.add(e)
Evaluation evs = Evaluator.evaluateAverage(evaluations)
OutputRenderer.generateEvaluationData(evs)
```

evaluate()	
Input	String tsId File reduced File reconstructed int p String reductionTechnique PerformanceData pd
Output	Evaluation evaluation

```
Evaluation evaluation = null
TimeSeries tso = DataStoreManager.loadTS(tsId)
TimeSeries tsAprox = DataStoreManager.loadReconstructedTS(tsId)
double error = getError(tso, tsAprox)
double compression = reduced.size() / DataStoreManager.getSize (tsId
    )
evaluation = new Evaluation(error, compression, p, ...)
DataStoreManager.insertEvaluation(tsId, evaluation,
    reductionTechnique, p, pd)
return evaluation
```

evaluateAverage()	
Input	Evaluation [ ] evaluations
Output	Evaluation evaluation

```
Evaluation evaluation = null
double error = 0
double compression
for e in evaluations:
  error += e.getError()
  compression += e.getCompression()
error = error/evaluations.length()
compression = compression/evaluations.length()
evaluation = new Evaluation(error, compression)
OutputRenderer.renderizeExplorationChart(evaluation.getError(),
    evaluation.getCompression())
OutputRenderer.renderizeSummaryTable(evaluation.getError(),
    evaluation.getCompression())
return evaluation
```

Appendix D

Resumen en Castellano

El llamado *Big Data* y, por extensión, las tecnologías de procesamiento y explotación de datos constituyen una de las tendencias en Tecnologías de la Información (TI) a nivel global desde comienzo de los años 2010. Aunque los antecedentes de las técnicas de análisis de datos datan de varias décadas atrás y las primeras tecnologías *Big Data* se desarrollaron durante la década de los 2000, a lo largo de la década de los 2010 la popularización del *Big Data* [MCB$^+$11] ha motivado el interés por la aplicación de estas tecnologías en numerosos campos de aplicación. La aplicación en múltiples sectores de las tecnologías para el procesamiento y explotación de datos, favorecida por una promoción intensiva de las herramientas Big Data y de otras tecnologías sinérgicas como la "Computación en la Nube" (*Cloud Computing*) y el "Internet de las Cosas" (*Internet of Things*, IoT), ha derivado el concepto de "economía de los datos" (*data-driven economy*) [Eur14] como uno de los pilares del desarrollo económico a nivel mundial. Según el informe publicado por la Comisión Europea en 2017 [IO17], el valor del mercado de datos en la Unión Europea (UE), esto es, el intercambio de productos y servicios basados en datos, se estima que fue de 60 mil millones de euros en 2016, y se espera que crezca hasta suponer más de 106 mil millones de euros en 2020. De manera similar, se espera que el número total de empresas en la UE cuya actividad principal es el suministro de productos y servicios basados en datos crezca de 255.000 unidades en 2016 a 360.000 unidades en 2020, y que el impacto agregado de este mercado de datos respecto al total de la economía de la UE crezca desde casi un 2 % del PIB de la UE en 2016 a un 4 % en 2020.

Uno de los focos estratégicos donde esta economía de los datos se está desplegando a nivel mundial es la industria de manufactura, como un medio para revitalizar la competitividad global de este sector dada su relevancia para la economía de numerosos países, y para revertir la tendencia hacia la desindustrialización. Por ejemplo, según la Comisión Europea, la producción del sector industrial supone el 17 % del PIB de la UE y el 75 % de sus exportaciones son productos manufacturados. Es más, representa un factor clave en la creación y crecimiento del empleo, dado que cada trabajo en la industria de manufactura genera al menos un trabajo adicional en servicios [Eur17a]. La concreción de esta economía de los datos en la industria de manufactura ha dado lugar al desarrollo de la “Fabricación Avanzada” o “Fabricación Inteligente” (*Smart Manufacturing*), como un término global abarcando diferentes iniciativas y estrategias que abordan la explotación de datos para la optimización y transformación de los negocios de manufactura. No en vano, las principales iniciativas a nivel mundial promoviendo la adopción del *Smart Manufacturing* [EA12][KLW11][Pre11] coinciden en el tiempo con la popularización del *Big Data* a lo largo de los años 2010.

El *Smart Manufacturing* se define [DEP$^+$12] sobre dos conceptos principales: la compilación de registros (*manufacturing records*) sobre los productos fabricados con datos sobre su histórico, estado, calidad y características principales, y la aplicación de “inteligencia” (*manufacturing intelligence*) sobre dichos registros de manera que los fabricantes puedan predecir, planificar y gestionar circunstancias específicas que permitan optimizar la producción. El interés por parte de las empresas de manufactura en el *Smart Manufacturing* se basa en las posibilidades para transformar sus procesos de producción y sus modelos de negocio. Por un lado, la adopción de estas aproximaciones basadas en datos tiene como objetivo que el valor obtenido de la explotación de dichos datos genere incrementos significativos en la eficiencia de los sistemas automatizados de producción, en la calidad de los bienes producidos y en el beneficio de la empresa en general. Por otro lado, habilita la posibilidad de adoptar estrategias de servitización basadas en datos para aquellos fabricantes de bienes de equipo que quieran transformar sus modelos de negocio a través del suministro a sus clientes de servicios de valor añadido basados en la explotación de datos. Los beneficios esperados de estas diferentes aproximaciones han derivado en diferentes objetivos para las aplicaciones del Smart Manufacturing: control de los sistemas de producción, control de la calidad de productos, sistemas de apoyo a la toma de decisiones, diagnóstico de fallos y mantenimiento predictivo del equipamiento, etc.

Por su propia definición, el despliegue del *Smart Manufacturing* demanda la utilización de TI relacionadas con la explotación de datos y de plataformas digitales que faciliten el logro de los objetivos marcados para el *Smart Manufacturing.* El diseño e implementación apropiados de dichas plataformas se enfrenta a diversos retos de investigación e innovación, en relación a las tecnologías habilitadoras necesarias. Esto incluye, entre otros, los siguientes elementos [Eur16]: métodos mejorados para la captura de datos valiosos de las máquinas y la integración de dichos datos capturados de diferentes fuentes; la inclusión de nuevos elementos tecnológicos junto a sistemas heredados y la integración de máquinas IoT en líneas de producción heredadas; arquitecturas de datos que cubran las necesidades industriales y suministren la información correcta a la persona adecuada en el momento preciso; herramientas para la predicción, monitorización y visualización; implementación de métodos de análisis de datos que permitan correlacionar información de producto, proceso y negocio, así como predecir los indicadores de rendimiento y de calidad de producto; etc.

Dado el amplio espectro de estos retos tecnológicos y su complejidad, la adopción de las TI de explotación de datos que las empresas de manufactura necesitan para transformar sus negocios hacia el *Smart Manufacturing* requiere del apoyo de empresas suministradoras de tecnología [Eur17b] que estén especializadas en estos “Servicios de *Big Data* Industrial” (SBDI). De esa manera, las empresas de manufactura reducen el riesgo en la adopción de estas tecnologías y, al mismo tiempo, se habilita un nuevo mercado para los suministradores de tecnología, ligado al despliegue de soluciones tecnológicas innovadoras que faciliten la adopción del *Smart Manufacturing.* Esta especialización en los suministradores de tecnología, es decir, los Proveedores de SBDI, y sus retos al diseñar y gestionar el suministro de estas tecnologías constituyen el foco de este trabajo de investigación.

D.1 Los Proveedores de SBDI: un Agente Fundamental para el Smart Manufacturing

En este contexto de desarrollo a nivel global del *Smart Manufacturing* y de diversas iniciativas promocionando su adopción en diferentes países y regiones, nos situamos en la perspectiva de los Proveedores de SBDI y su objetivo estratégico de desarrollar su negocio suministrando estos servicios en escenarios de *Smart Manufacturing.* Este foco nos permite encuadrar el objetivo de nuestras contribuciones dentro del marco de las propuestas existentes para el *Smart Manufacturing* y los proyectos de explotación de datos. De esa manera, el objetivo general de este trabajo de investigación es proporcionar contribuciones que (a) ayuden al sector de los Proveedores de SBDI a desplegar servicios de datos eficaces para el desarrollo del *Smart Manufacturing* y sus objetivos estratégicos, y (b) adapten y extiendan las propuestas conceptuales, metodológicas y tecnológicas existentes para incorporar los elementos prácticos que faciliten su aprovechamiento en estos contextos de negocio.

Para suministrar sus servicios basados en la explotación de datos, los Proveedores de SBDI establecen alianzas con empresas industriales en diferentes sectores y mercados y desarrollan proyectos que despliegan las soluciones tecnológicas necesarias en las instalaciones de las empresas de manufactura. Estos proyectos se desarrollan en paralelo en varios sectores y tienes como objetivo el despliegue y refinamiento progresivos de los servicios requeridos sobre los datos en cada escenario. La gestión de estos proyectos conlleva retos importantes para los Proveedores de SBDI en relación a (a) aspectos organizacionales ligados a los roles que son necesarios en el equipo que lleva a cabo estos proyectos y (b) aspectos tecnológicos relacionados con el diseño de la plataforma de captura e integración de datos que sustenta el despliegue de múltiples proyectos en paralelo a nivel mundial. Además, todos estos aspectos deben estar alineados con la estrategia de negocio de los Proveedores de SBDI y con los requerimientos y necesidades de las empresas industriales con las que establecen sus alianzas en diferentes sectores de manufactura. La complejidad de estos proyectos motiva y proporciona el foco para este trabajo de investigación. Así, las contribuciones de este trabajo tienen como objetivo proporcionar soluciones valiosas a varios de los retos específicos encontrados en estos proyectos.

La observación y análisis de los escenarios de *Smart Manufacturing* donde los Proveedores de SBDI despliegan sus servicios facilita la identificación de oportunidades para hacer contribuciones relevantes que extiendan las propuestas existentes en las áreas de conocimiento implicadas. Por ejemplo, muchas de las propuestas conceptuales en relación al desarrollo de plataformas tecnológicas para el *Smart Manufacturing* plantean una aproximación holística y se orientan a agentes que tienen la capacidad de diseñar desde cero o rediseñar por completo la infraestructura necesaria. Sin embargo, en los escenarios reales de negocio donde los Proveedores de SBDI suministran sus servicios, estos se encuentran con negocios de manufactura en marcha con una infraestructura de Tecnología de Operación (TO) ya desplegada y funcionando. Por esa razón, para que la propuesta de valor de un Proveedor de SBDI sea aceptada más fácilmente, su objetivo al desplegar tecnología adicional desde ser su integración con la existente sin interferir con la operación en marcha del negocio de manufactura.

Relacionado con lo anterior, gran parte de las principales aproximaciones metodológicas para el ciclo de vida de la explotación de datos asumen un punto de partida donde efectivamente existen nuevos datos disponibles para su procesamiento. No obstante, este no es el caso cuando un Proveedor de SBDI suministra sus servicios a empresas de manufactura, ya que la mayoría de dispositivos generadores de datos que funcionan en sus instalaciones fueron diseñados y desplegados para la automatización y supervisión interna, y no para facilitar la transmisión de dichos datos a una plataforma externa para su procesamiento, explotación y análisis posteriores. Por ese motivo, la tecnología desplegada por un Proveedor de SBDI debe cubrir esa brecha para poder extraer los datos y almacenarlos en el repositorio donde se acumulen para su explotación. Es más, el diseño de esa solución tecnológica debe estar alineado con un desarrollo sostenible del negocio del Proveedor de SBDI, y no como proyectos *ad hoc* para cada instalación industrial a monitorizar.

D.2 Alcance y Método de este Trabajo de Investigación

De entre las diferentes oportunidades que surgen en el contexto descrito anteriormente para contribuciones relevantes que faciliten el desarrollo de los objetivos de negocio de un Proveedor de SBDI, destacamos tres retos específicos relacionados con las etapas iniciales del ciclo de vida de los datos. Estas etapas aseguran la disponibilidad de los nuevos datos a procesar provenientes de las instalaciones industriales monitorizadas, cuyos propietarios buscan explotar dichos datos para acercar sus negocios al *Smart Manufacturing*. Así, los tres retos en los que centramos nuestra investigación son los siguientes:

1. La concepción de una estrategia más eficiente para el almacenamiento de datos que reduzca los costes de la infraestructura “en la nube” que un Proveedor de SBDI necesita para centralizar y acumular la cantidad masiva de datos provenientes de todas las instalaciones industriales a las que presta servicio.

2. El diseño de la arquitectura para la infraestructura de captura e integración de datos que sustenta la plataforma tecnológica de un Proveedor de

SBDI. Esta arquitectura debe asegurar una integración no intrusiva con la infraestructura de TO en funcionamiento en las plantas monitorizadas y una extensión progresiva de las funcionalidades de la plataforma para poder suministrar servicio a cada vez más escenarios.

3. El proceso de diseño colaborativo junto con las empresas de manufactura de los servicios de datos requeridos para un sector industrial concreto. Esta colaboración sustenta las alianzas estratégicas con estas empresas en los escenarios objetivo y refuerza el valor de la propuesta de servicios de los Proveedores de SBDI.

El alcance perfilado por estos retos apunta a una importante característica de esta investigación: en vez de girar en torno a un área específica de investigación y conocimiento, está dirigida por un foco de análisis más amplio en torno a los requerimientos de los sistemas de información con los que los Proveedores de SBDI sustentan su negocio. Esto implica un trabajo de investigación que analice (a) los escenarios de *Smart Manufacturing* donde un Proveedor de SBDI suministra sus servicios, para caracterizar sus agentes relevantes, sus estrategias de negocio y sus requerimientos respecto a los sistemas de información implicados, y (b) la identificación de las áreas de conocimiento donde analizar trabajos de investigación relacionados, para poder trazar sinergias con referencias pertinentes y descubrir limitaciones a modo de oportunidad para contribuciones relevantes.

Para cumplir estos objetivos, el método empleado en este trabajo se basa en dos aproximaciones metodológicas principales: la “Investigación basada en Ciencia del Diseño” (*Design Science Research*, DSR) [HMPR04][Hev07] y la “Investigación basada en Estudio de Casos” (*Case Study Research*, CSR) [Bas17][Eis89]. Por un lado, la DSR proporciona una metodología para la investigación de sistemas de información, con el objetivo de construir artefactos de diseño que estén basados en (a) las necesidades y requerimientos del problema de negocio identificado en el dominio de aplicación analizado, y (b) la identificación de sinergias y oportunidades con respecto al conocimiento existente en las áreas de investigación relacionadas. Esta base asegura el rigor y la relevancia de los artefactos de diseño, de manera que sean contribuciones de investigación validas para la audiencia académica y aportaciones valiosas para la audiencia profesional y su entorno. Por otro lado, la CSR permite a los investigadores de sistemas de información aprender del análisis de las innovaciones puestas en práctica por los profesionales y capturar conocimiento que puedan después formalizar. Este enfoque es particularmente apropiado para problemas basados en la práctica y donde tanto las experiencias de los actores como el contexto de sus acciones sean críticos. La realización de un estudio de casos es especialmente adecuada para nuestro trabajo de investigación, dado que su foco requiere una observación directa de un escenario de negocio real donde los agentes relevantes a todos los niveles interactúen entre sí para la construcción de servicios basados en datos, cumpliendo sus respectivas estrategias de negocio.

Así, la realización de un estudio de casos sustenta dos elementos cruciales de este trabajo. En primer lugar, nos permite capturar una caracterización más detallada de los escenarios de Smart Manufacturing, a través del análisis de una instancia relevante de estos escenarios y de los agentes implicados en ellos. Esto permite refinar la definición del alcance de nuestra investigación y los escenarios específicos a los que se dirigen nuestras contribuciones, basándonos en los requerimientos prácticos y las necesidades de negocio de todos los agentes que interactúan en estos escenarios en torno a los Proveedores de SBDI. El aprovechamiento de estos requerimientos y necesidades como entrada al proceso de DSR es lo que asegura la relevancia de los artefactos de diseño propuestos. En segundo lugar, facilita el terreno para una validación de campo en un contexto real de negocio de los componentes nucleares de los artefactos de diseño. Las contribuciones de un proceso de DSR se evalúan en tanto en cuanto se aplican a las necesidades de negocio de un entorno marcado como objetivo. Un contraste exitoso en dicho entorno es lo que habilita su inclusión como nuevo contenido relevante para la base de conocimiento de las áreas relacionadas, para su posterior puesta en práctica e investigación adicional.

Para poder realizar el estudio de casos, nos integramos en el escenario real de negocio de un Proveedor de SBDI que suministra servicios a diversos escenarios de *Smart Manufacturing*. Esto nos permitió observar el mercado de los Proveedores de SBDI en general, así como los diferentes tipos de empresas industriales y sectores de manufactura donde los servicios de los Proveedores de SBDI se despliegan. Además, nos facilitó el acceso a los proyectos de despliegue de dichos servicios en sectores de manufactura concretos. En particular, estudiamos detalladamente y de primera mano la alianza estratégica establecida entre este Proveedores de SBDI y un fabricante de bienes de equipo desplegando su estrategia de servitización [VR88] basada en datos en un sector de manufactura química distribuido por todo el mundo, y acompañamos a estas empresas a lo largo del despliegue del proyecto lanzado para uno de los clientes internacionales de este fabricante de bienes de equipo. Eso nos permitió interactuar en primera persona con los implicados relevantes de estas empresas y acceder a los datos provenientes de las instalaciones industriales monitorizadas y a la tecnología desplegada para capturar y procesar dichos datos. Todos estos elementos del mundo real reforzaron la caracterización de los escenarios objetivo y permitieron la validación de campo de los componentes nucleares de nuestras contribuciones, dirigidas a roles específicos del equipo que los Proveedores de SBDI organizan para los proyectos de despliegue de sus servicios.

D.3 Contribuciones Principales de este Trabajo de Investigación

El desarrollo del método de trabajo antes descrito, acompañando e interactuando con diferentes roles de gestión y técnicos en las organizaciones implicadas en el escenario de nuestro estudio, nos permitió caracterizar el mercado de los Proveedores de SBDI, los requerimientos generales de los agentes de los escenarios de *Smart Manufacturing* y las necesidades particulares de los roles del equipo que un Proveedor de SBDI establece para sus proyectos de despliegue en diversos sectores de manufactura. Todos estos requerimientos y necesidades, extraídos de la realidad estratégica, táctica y operativa de estas empresas, junto con el análisis de las adaptaciones y extensiones necesarias para que las propuestas en las áreas de conocimiento relacionadas den una respuesta eficaz a dicha realidad, sustentan la relevancia y el rigor de las tres contribuciones principales de esta investigación. Estas contribuciones están específicamente orientadas a dos de los roles del equipo para los proyectos de los Proveedores de SBDI: el director de proyecto que gestiona la interacción con los implicados de las empresas de manufactura y la extracción de requerimientos de negocio para los servicios a lanzar, y el ingeniero de datos a cargo del diseño, actualización y optimización de la plataforma de captura e integración de datos.

La primera contribución principal es un diseño del proceso y arquitectura para la planificación y ejecución del análisis de la reducción de series temporales. Esta contribución se dirige al cometido del ingeniero de datos de un Proveedor de SBDI, a cargo de analizar cómo reducir el espacio de almacenamiento de los tipos altamente heterogéneos de series temporales que constituyen los datos a capturar en las instalaciones industriales donde despliegan sus servicios. La relevancia de esta contribución está ligada a (a) los costes de los servicios de almacenamiento en la nube que un Proveedor de SBDI requiere para desplegar y operar su plataforma, dado el impacto de estos recursos en el alcance de los servicios a ofrecer a las empresas de manufactura, y (b) los costes internos del tiempo y recursos asignados para explorar las posibilidades de reducción de datos en las series temporales capturadas. Así, esta contribución representa el proceso (incluyendo la arquitectura de los artefactos tecnológicos para automatizar la mayor parte de sus pasos) que guía de manera eficiente el trabajo de este ingeniero de datos y prioriza la asignación de recursos de análisis a aquellas series temporales en las que se espera un mayor impacto en ahorro de espacio de almacenamiento. La aplicación de este proceso permite obtener la especificación de la solución de reducción a desplegar en la plataforma del Proveedor de SBDI, es decir, qué técnicas de reducción deben aplicarse sobre qué series temporales, para que el espacio de almacenamiento de los datos se optimice sin comprometer su explotación posterior. Además, a medida que el ingeniero de datos utiliza una implementación del diseño propuesto para analizar escenarios adicionales, se refina progresivamente la caracterización de series temporales, su clasificación en familias y su asociación con técnicas de reducción recomendadas. Este refinado proporciona un proceso eficiente de gestión de conocimiento y lecciones aprendidas en los diversos despliegues y habilita los ahorros en recursos para los sucesivos análisis.

La segunda contribución principal es el diseño de una arquitectura distribuida híbrida para la plataforma de captura e integración de datos de un Proveedor de SBDI. Este modelo de arquitectura complementa los paradigmas existentes para sistemas *Big Data*, describiendo los componentes que cubren la brecha entre un estado inicial en el que aún no se están capturando datos en las instalaciones industriales y el estado donde ya se tiene disponible un repositorio centralizado con esos datos, concebido como un *Big Data Lake* [O'L14] sobre el que se puedan diseñar diferentes capas de funcionalidades para su explotación [MW15]. Los componentes de esta arquitectura combinan eficazmente elementos de IoT industrial y de computación en la nube, analizando su utilización en más de 60 instalaciones industriales distribuidas a nivel mundial, para dar respuesta al volumen, velocidad y variedad de datos encontrados en escenarios reales de negocio de manufactura. El principal punto diferencial del diseño propuesto es que la arquitectura no está concebida como una solución para migrar por completo la infraestructura industrial de aquellos escenarios que quieren migrar hacia el Smart Manufacturing. La arquitectura está diseñada como una solución para el negocio de un Proveedor de SBDI, basado en facilitar esa migración de manera integradora y no intrusiva con respecto a la infraestructura ya en funcionamiento. Por otra parte, facilita al ingeniero de datos la actualización progresiva de las funcionalidades de la plataforma para cubrir más escenarios de aplicación y más pasos de transformación de los datos de cara a la provisión de servicios sobre ellos.

La tercera contribución principal es el diseño de un modelo para el proceso, dirigido por los implicados de negocio, de caracterización de los requerimientos de explotación de datos para los servicios a desplegar, y está dirigida al rol de director de proyecto que el Proveedor de SBDI proporciona en los proyectos de despliegue de servicios realizados en los escenarios objetivo. Esta contribución se basa en la integración de conocimiento relevante de áreas como gestión de implicados, diseño de modelos de negocio y análisis de entrevistas para superar las limitaciones identificadas en los modelos para el proceso de "descubrimiento de conocimiento y minería de datos" (*Knowledge Discovery and Data Mining*, KDDM) [KM06] y para la extracción de requerimientos en proyectos de datos [CLSM$^+$14], de cara a su aplicación en el diseño de los servicios basados en datos para los escenarios objetivo. Así, esta contribución extiende los modelos para el proceso de KDDM con una aproximación incremental, diseñada como un modelo de proceso en espiral para la integración de la comprensión del negocio en el ciclo de vida a cubrir para los datos, y facilita la interacción con implicados de negocio para extraer y caracterizar los requerimientos de explotación de datos. Esta caracterización se captura en una plantilla denominada "lienzo puente" (*BRIDGE canvas*), que conecta los requerimientos de negocio con su impacto en los pasos relevantes del proceso de KDDM, de manera que esos requerimientos puedan ser tenidos en cuenta como entrada para los pasos del ciclo de vida de los datos.

D.4 Conclusión General

Las contribuciones presentadas en este trabajo integran aspectos prácticos clave que se derivan de la observación directa y de la experiencia de primera mano en el estudio de casos realizado. Además, un valor adicional del trabajo es su aproximación multidisciplinar integrando conocimiento de diversas áreas de investigación, estableciendo sinergias con ellas e identificando limitaciones a modo de oportunidades para contribuciones valiosas. A ese respecto, el escenario de negocio real donde se ha realizado el estudio ha representado un recurso de gran valor. Nos ha proporcionado el acceso directo a las organizaciones que desarrollan sus estrategias de negocio dentro de los escenarios de *Smart Manufacturing*, permitiéndonos observar la complejidad de dichos escenarios y los retos prácticos a encarar al desarrollar los proyectos que buscan conectar las soluciones tecnológicas de explotación de datos con la realidad de la industria de manufactura y la tecnología para la operación de sus instalaciones. Además, nos ha facilitado conocimiento sobre cómo estas soluciones permiten a los fabricantes de bienes de equipo el desarrollo de sus estrategias de servitización, analizando el caso de un sector distribuido a nivel mundial, las instalaciones donde se despliegan los servicios basados en datos y los implicados de las diversas organizaciones involucradas. La representatividad de las organizaciones, implicados y tecnologías analizadas ha facilitado la caracterización detallada de estos escenarios y la identificación de los aspectos relevantes a tener en cuenta en el despliegue de estos proyectos. Esta caracterización evidencia la relevancia de las contribuciones presentadas y de otras que puedan abordar más retos y requerimientos que se derivan de estos escenarios.

Estas contribuciones suponen un recurso de valor tanto para profesionales como para el mundo académico. Por una parte, proporcionan un apoyo beneficioso para los Proveedores de SBDI y, por extensión, para la industria de manufactura que busca incrementar su competitividad mediante la adopción de los servicios suministrados por estos Proveedores de SBDI. Así, estas contribuciones refuerzan el rol de los Proveedores de SBDI como agentes necesarios en el desarrollo estratégico de la industria de manufactura y en el despliegue eficaz de las políticas de adopción del *Smart Manufacturing*. Por otra parte, las contribuciones propuestas integran y extienden las propuestas conceptuales, metodológicas y tecnológicas existentes en diversas áreas de conocimiento, poniendo el foco en los aspectos prácticos que es necesario tener en cuenta para que estas propuestas puedan aprovecharse en los escenarios reales de negocio donde los Proveedores de SBDI suministran sus servicios. Así, estos aspectos pueden ser también tenidos en cuenta al idear versiones futuras de estas propuestas.