Horizon 2020 Program (2014-2020)
Big data PPP

Research addressing main technology challenges of the data economy.
ICT-16-2017

Industrial-Driven Big Data as a Self-Service Solution [†]

Deliverable D1.3: Positioning of *I-BiDaaS*

Abstract: This deliverable provides a detailed account of the activities of Workpackage 1, towards the positioning of *I-BiDaaS*. It is driven by the business objectives and user requirements of the specific use cases of the project industrial partners, as well as by the state-of-the-art in Big Data as-a-Self-Service research and tools. Following systematic work, this has led to the specification of the requirements, their analysis and mapping to the architecture of the *I-BiDaaS* platform. Furthermore, this deliverable describes an initial design of the experimental protocol, towards the evaluation, in quantitative and qualitative terms, of the research and technological innovation proposed by *I-BiDaaS*.

Contractual Date of Delivery	*31/08/2018*
Actual Date of Delivery	*21/09/2018*
Deliverable Dissemination Level	Public
Editors	*Iliada Eleftheriou, Evangelia Kavakli, Thomas Lambert, Rizos Sakellariou (UNIMAN)*
Contributors	All *I-BiDaaS* partners
Quality Assurance	*Dusan Jakovetic (UNSPMF), Vassilis Prevelakis (AEGIS)*

[†] The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement n° 780787.

The *I-BiDaaS* Consortium

Foundation for Research and Technology – Hellas (FORTH)	Coordinator	Greece
Barcelona Supercomputing Center (BSC)	Principal Contractor	Spain
IBM Israel – Science and Technology LTD (IBM)	Principal Contractor	Israel
Centro Ricerche FIAT (FCA/CRF)	Principal Contractor	Italy
Software AG (SAG)	Principal Contractor	Germany
Caixabank S.A. (CAIXA)	Principal Contractor	Spain
University of Manchester (UNIMAN)	Principal Contractor	United Kingdom
Ecole Nationale des Ponts et Chaussees (ENPC)	Principal Contractor	France
ATOS Spain S.A. (ATOS)	Principal Contractor	Spain
Aegis IT Research LTD (AEGIS)	Principal Contractor	United Kingdom
Information Technology for Market Leadership (ITML)	Principal Contractor	Greece
University of Novi Sad Faculty of Sciences (UNSPMF)	Principal Contractor	Serbia
Telefonica Investigation y Desarrollo S.A. (TID)	Principal Contractor	Spain

Document Revisions & Quality Assurance

Internal Reviewers

1. *Dusan Jakovetic (UNSPMF)*
2. *Vassilis Prevelakis (AEGIS)*

Revisions

Ver.	Date	By	Overview
1.0	21/09/2018	*Quality assurance, Internal reviewers*	Final review and approval.
0.5	20/09/2018	*E. Kavakli*	Incorporation of the final comments.
0.4	20/09/2018	*Quality assurance, Internal reviewers*	Comments on the final draft.
0.3	13/09/2018	*E. Kavakli, R. Sakellariou*	Final draft.
0.2	07/09/2018	*Quality assurance, Internal reviewers*	Comments on the first draft.
0.1	30/08/2018	*Editors*	First draft.
0.0.3	20/06/2018	*R. Sakellariou*	Revised Table of Contents.
0.0.2	25/05/2018	*E. Kavakli*	Revised Table of Contents.
0.0.1	21/04/2018	*E. Kavakli*	Table of Contents.

Contents

List of Abbreviations

AI	Artificial Intelligence
BDV	Big Data Value
BDVA SRIA	European Big Data Value Strategic Research and Innovation Agenda
CEP	Complex Event Processing
CPS	Cyber-Physical Systems
DFP	Data Fabrication Platform
DoA	Description of Action
FR	Functional Requirement
GPU	Graphics Processing Unit
HPC	High Performance Computing
IoT	Internet of Things
KPI	Key Performance Indicator
ML	Machine Learning
MPP	Massively Parallel Processing
MVP	Minimum Viable Product
NFR	Non-Functional Requirements
NIST	National Institute of Standards and Technology
RE	Requirements Engineering
SPEC	Standard Performance Evaluation Corporation
TCP	Transaction Processing Performance Council
QoS	Quality of Service
UM	Universal Messaging
UML	Unified Modelling Language

List of Figures

List of Tables

Executive Summary

This deliverable documents the project's activities towards the Positioning of *I-BiDaaS*. It completes the *Baseline Phase* of *I-BiDaaS*, that is, the work in Work Package 1 (WP1). This is the driving force behind the project providing direct input to the subsequent phases of the project life cycle. It also serves as a check list for the whole project work plan.

In particular, based on the results of the *Project Set Up* activity, Positioning of *I-BiDaaS* aims to specify the test cases for the experiments (in terms of the business, data and technical requirements of the different *I-BiDaaS* use cases). It includes a verification and validation approach (based on the user-specific indicators and industrial-validated benchmarks) and develops a mapping of the *I-BiDaaS* conceptual architectures components with the consolidated requirements of all cases.

The methodological approach followed extends relevant work in the area of Big Data Requirements Engineering and is intended to systematically progress from high level goals and business key performance indicators and the characteristics of the specific data sets and associated analytic capabilities to the requirements of a combined big data solution. This represents an end-to-end Big Data as-a-self-service platform which complies with and extends existing Big Data Reference Architectures.

Overall, deliverable *D1.3* contributes to the *I-BiDaaS* success (in terms of covering the needs of the industrial partners), innovation (in terms of highlighting the sate-of-the-art gaps in the field of Big Data-as-a-self-service solutions), and emphasizes the areas where it advances current research in the context of requirements elicitation, architecture specification and experimental validation of industrial Big Data solutions.

1 Introduction

1.1 Overview

This deliverable presents the results of *WP1* namely, "Setting the scene: Baseline framework". It reports on both the detailed functionality of the *I-BiDaaS* platform, according to the users' needs, and the state-of-the-art in data analytics solutions over extremely large numbers of high volume streams that are produced in projects industrial domains.

In particular, *D1.3* describes the methodology for eliciting and reporting industrial Big Data analytics requirements. This methodology takes into consideration the general challenges and requirements relevant to the *I-BiDaaS* industrial domains (reported in detail in [48]), as well as the literature on Requirements Engineering (RE) approaches specifically for Big Data applications. The application of the methodology for defining the requirements of the specific user cases, is also presented. These requirements are taken into consideration for the definition of the *I-BiDaaS* technical solution, an overview of which is also provided in this document.

In order to validate the new architecture and the different use cases in a business context (*I-BiDaaS* experiments) an experimental protocol has been defined. The protocol aims at grasping the high degree of heterogeneity and complexity of the use cases, taking also into consideration existing industrial and academic Big Data benchmarks. The proposed experimentation process consist of two phases. The first phase is designed to provide quantitative evaluation of the *I-BiDaaS* architecture against workloads that reflect the complexity of each use case, using appropriate benchmarks. The second phase aims at the qualitative evaluation of the innovation of the *I-BiDaaS* solution in the business context and against the identified business requirements.

1.2 Relation to other Tasks and Work Packages

There is a close interrelation between this deliverable and all the tasks in the current Workpackage *WP1*, in that *D1.3* synthesizes the results of all *WP1* tasks and integrates the alternative perspectives developed in each task into a coherent framework that forms the baseline of subsequent *I-BiDaaS* developments. In particular, based on a thorough understanding of the challenges, technologies and the state of the art in terms of data economy and Big Data processing (performed in T1.1), *D1.3* details a number of industrial test cases and associated user requirements (performed in T1.2), which guide the conceptual specification of the *I-BiDaaS* architecture (performed in T1.3) and specifies the experimental protocol against which the *I-BiDaaS* industrial experiments will be validated (performed in T1.4).

Furthermore, the data requirements detailed in D1.3 are used to define the data that are not available for the experiments, in order to be fabricated by IBM in *WP2* (Task 2.1. Definition of data assets nature and format).

Finally, the industrial benchmarks and the experimental protocol defined in *D1.3* will guide the definition of the appropriate test cases to drive the technical evaluation of the solution (Task 5.3. Integration towards Big-Data-as-a-Self-Service) and will be further elaborated and aligned in WP6 (Task 6.1. Experimental protocol alignment), respectively.

This dependency between the current deliverable and WP1, WP2, WP5 and WP6 tasks is further detailed in Figure 1.1, showing how *D1.3* feeds into the tasks T2.1, T5.3 and T6.1 into consideration.

1.3 Contribution to the Scientific and Business Objectives

With respect to the overall project objectives, *D1.3* contributes towards **the development of data processing tools and techniques applicable in real-world settings**, that demonstrate significant increase of speed of data throughput and access (Objective 4). In particular, the detailed analysis of the user needs, culminating in the requirements elicitation framework presented in this document, serves as the baseline for identifying and establishing those concepts that would be applicable to all use cases and for investigating the case for a generic conceptualisation of industrial Big Data analytics as a service solution.

In addition the experimental protocol, specifies the experimentation variables (e.g., performance, accuracy, usability, maintainability) and experimental subjects (including their role) to validate the *I-BiDaaS* tools and services and thus contributes towards the **construction of a safe environment for methodological Big Data experimentation, for the development of new products, services, and tools** (Objective 3).

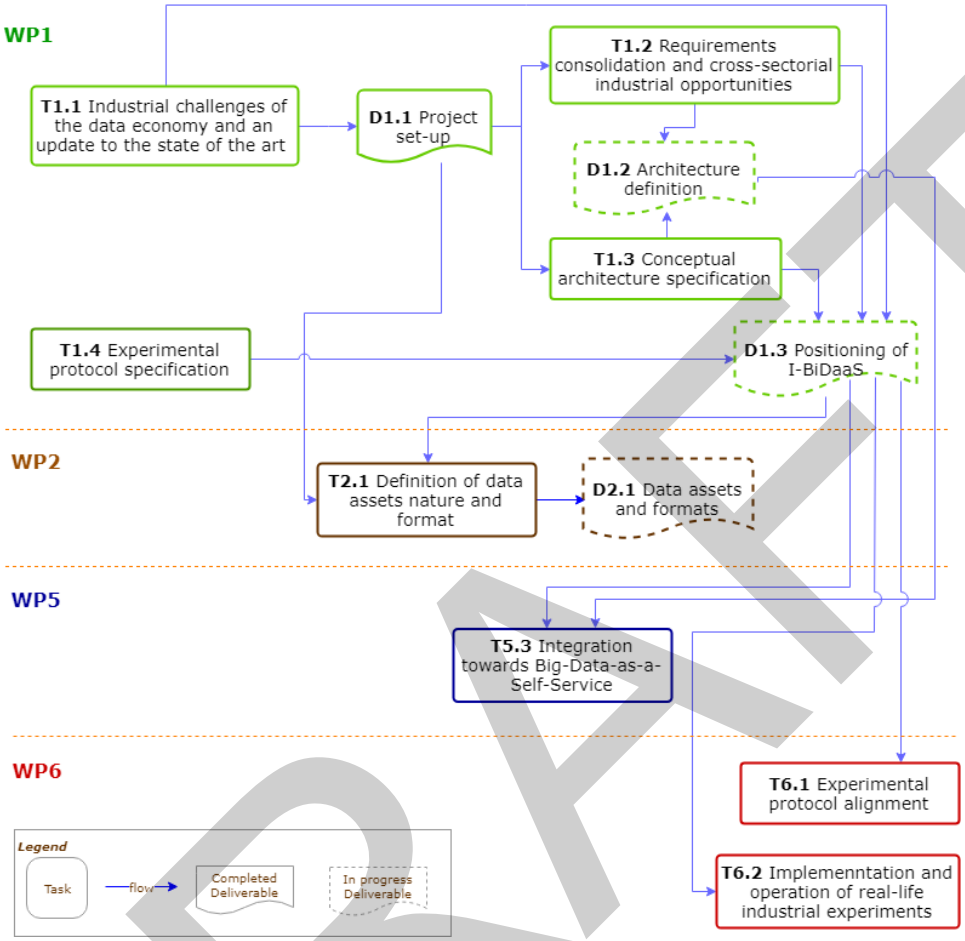

Figure 1.1: Overview of WP1 Tasks and their dependencies to other WPs

With respect to the *WP1* objectives, *D1.3* reports on the process of capturing, representing and consolidating and the requirements for all *I-BiDaaS* use cases. Furthermore it describes how these have been taken into account towards the specification of the *I-BiDaaS* conceptual architecture, and for the definition of the *I-BiDaaS* experimental process. Thus, *D1.3* contributes to the *WP1* objective **to specify the test cases for the experiments including the verification and validation approach**.

1.4 Structure of the Document

Section 2 positions *I-BiDaaS* describing its primary objectives, identifying the relevant target audience and their needs as well as current technological solutions and challenges based on the state-of-the-art. Section 3 provides an overview of the *I-BiDaaS* industrial test cases. Furthermore, it describes the

common methodology used for eliciting and documenting the requirements of all cases, and its application which resulted in a coherent list of *I-BiDaaS* system requirements. These formed the key input for defining the functionality and conceptual architectures of the *I-BiDaaS* platform (described in Section 4). Section 5 describes the methodology that will be used for validating *I-BiDaaS* test cases against user requirements and industry-validated benchmarks. Finally, Section 6 concludes this report with a summary of achievements of *WP1*, a discussion on how the work to date has met the objectives of this part of the project and an introduction to the work planned as detailed in the project's description of action (DoA).

2 I-BiDaaS Positioning

The aim of this section is to describe the position of the *I-BiDaaS* solution. To this end, it provides an overview of the *I-BiDaaS* objectives and defines its target groups and related needs. Furthermore, it sets the context of related solutions and associated state-of-the-art gaps that *I-BiDaaS* aims to address.

2.1 Overview of *I-BiDaaS* Solution

Today, business organizations leverage data pools to drive value, while it is variety, not volume or velocity, which drives big-data investments. The convergence of Internet of Things (IoT), Cloud, and Big Data technologies creates new opportunities towards a novel paradigm for Big Data analytics.

Big-Data-as-a-Self-Service represents the approach of an extensible platform that can provide, as-a-service, analytical capabilities over a variety of industries and use cases. From a functional perspective, such an approach covers the end-to-end capabilities of a fully prescribed analytical solution, from data acquisition to end-user visualization, reporting and interaction.

A self-service solution will have a transformative effect for organizations as it will empower their employees with the right knowledge, and give decision-makers the insights they need to make well-informed, appropriate decisions. It will shift the power balance within an organization, increase efficiency, reduce costs, enhance employee empowerment, and increase profitability.

In this context, the *I-BiDaaS* project aims to empower non-expert Big Data users to easily utilize and interact with Big Data technologies, by designing, building, and demonstrating, a unified solution that significantly increases the speed of data analysis while coping with the rate of data asset growth, and facilitates cross-domain data-flow towards a thriving data-driven EU economy.

To this end, *I-BiDaaS* aims to achieve the following:

- **Objective 1:** Develop, validate, demonstrate, and support, a complete and solid Big Data solution that can be easily configured and adapted by practitioners.
- **Objective 2:** Break inter- and intra-sectorial data-silos, create a data market and offer new business opportunities, and support data sharing, exchange and interoperability.
- **Objective 3:** Construct a safe environment for methodological Big Data experimentation, for the development of new products, services, and tools.
- **Objective 4:** Develop data processing tools and techniques applicable in real-world settings and demonstrate a significant increase of speed of data throughput and access.
- **Objective 5:** Develop technologies that will increase the efficiency and competitiveness of all EU companies and organisations that need to manage vast and complex amount of data.
- **Objective 6:** Demonstrate how data fabrication can help companies in their Big Data experimentation and testing.

2.2 Target Groups

Amid the proliferation of Big Data, not only in terms of datasets but also technologies able to capture, store, manage and analyse large and variable collections of data, it is possible to identify in the data economy vertical submarkets, hereby referred to as target groups, for which the *I-BiDaaS* solution will carry beneficial value.

The main vertical submarkets that will constitute target groups for the *I-BiDaaS* solution are hereinafter listed [81]:

- Automotive, Aerospace & Transportation
- Banking & Securities
- Insurance
- Defence & Intelligence
- Public Safety & Homeland Security
- Education
- Healthcare & Pharmaceutical
- Smart Cities & Intelligent Buildings

- Public Services
- Retail, Wholesale & Hospitality
- Manufacturing & Natural Resources
- Utilities & Energy
- Web, Media & Entertainment
- Telecommunications

The intent of this section is to further analyse why these submarkets form potential target groups for *I-BiDaaS*.

In the **Automotive, Aerospace & Transportation sector**, for instance, possible key applications of Big Data will be autonomous and semi autonomous driving, streamlining vehicle recalls and warranty management, fleet management, intelligent transportation, UBI (Usage Based Insurance), predictive aircraft maintenance, fuel optimization, and air traffic control [65]. Starting to make this envisaged future a reality has been the case with engine manufacturers such as: Boeing, which benefits from the use of Big Data, as flying becomes more efficient; BMW, which is eliminating defects in new vehicle models; Ford Motor Company for efficient transportation decisions; Groupe Renault, using Big Data to boost driver safety; and Honda Motor Company, to improve F1 performance and fuel efficiency [61].

In the **Banking & Securities** sector, Big Data is used for customer retention and personalized products, risk management, fraud detection, and credit scoring. It is proving useful to avoid regulatory penalties, improve business processes, and reduce loan defaults, as per HSBC, JPMorgan Chase & Co., and OTP Bank case studies. Other banks, such as CBA (Commonwealth Bank of Australia), use Big Data even to provide personalized services. Similar applications, including but not limited to claims fraud mitigation, customer retention and profiling, and risk management have been followed in the **Insurance sector** as well. The Zurich Insurance Group proved enhancement of risk management with Big Data, RSA Group improved customer relations, and Primerica upgraded sales force productivity [63].

Furthermore, in **Defense & Intelligence**, Big Data is applied in intelligence gathering, battlefield analytics, energy saving opportunities and even preventing injuries in the battlefield. The U.S. Air Force, Royal Navy, NSA (National Security Agency), the Ministry of State Security of China, French DGSE (General Directorate for External Security) are all examples of organizations that already capitalized on Big Data to provide actionable intelligence to warfighters, empower submarine warfare, detect threats, improve predictive policing, and enhance intelligence. Similarly, the **Public Safety &**

Homeland Security sector invested in Big Data mostly for cyber crime mitigation, crime prediction analytics, video analytics and situational awareness. DHS (Department of Homeland Security), for instance, was able to identify threats through Big Data and Dubai Police located wanted vehicles more efficiently [34].

Tremendous opportunities for Big Data are also present in the educational field, although academia are not known for being early adopters. The **Education sector** has been making use of Big Data potential to enhance information integration, identify learning patterns, and enable student directed learning. For example, Purdue University, located in West Lafayette, Indiana, developed Course Signals, a system that helps predict academic and behavioral issues and notifies teachers as well as students when action is required [29]. The system ensures students achieve their maximum potential as well as it decreases dropouts and failing rates. Course Signals is generally viewed as the best practice of how analytics can be applied in higher education to help improve student results assisting them to graduate in a timely manner. Similar Learner Analytics (LA) have been also used in Nottingham Trent University [75].

In **Healthcare & Pharmaceuticals**, Big Data can be exploited for drug discovery, design and development, clinical development and trials, population health management, personalized healthcare and targeted treatments, proactive and remote patient monitoring, and preventive care and health interventions. Applications have been already pursued at AstraZeneca for analytics-driven drug development; in Bangkok Hospital Group and United-Health Group to transform the patient experience and enhance their care and value; and in Novartis to digitize healthcare. Other cases of successful application of Big Data in the Healthcare sector are Pfizer, which proved its ability to develop effective and targeted therapies with Big Data and Sanofi, which invested on Big Data for proactive diabetes care [53].

Moving further, on energy optimization and fault detection, intelligent building analytics, urban transportation management, water and urban waste management are all key applications of Big Data in **Smart Cities & Intelligent Buildings**. Glasgow City Council promoted smart city efforts with Big Data. The Open Glasgow project aimed to integrate a range of city systems and data to deliver improved services and facilitate wider engagement with citizens. Open Glasgow hosted a Data Repository, a 'Big Data' store that integrated a range of datasets. The aim was to provide a platform to handle real-time information, analytics and models, city dashboards to present live city data, and enhance the existing MyGlasgow app, and a City Observatory [41]. On the other hand, OVG Real Estate developer of the Amsterdam headquarters building of consulting firm Deloitte, exploits Big Data technologies in order to optimize, measure and inform both the user experience and the buildings environmental performance. The building named

"The Edge", is an example how the use of sensors, big data and connectivity changes the way companies manage office buildings [80].

Among other possible applications for **Public Services**, Big Data stands out for public sentiment and economic analysis, tax collection and fraud detection, and prediction and mitigation of disasters. For instance, the ONS (Office for National Statistics) began exploring the UK economy of Big Data, whereas the New York State Department of Taxation and Finance was able to increase tax revenue. Use of Big Data for market analysis has been followed also in **Retail, Wholesale & Hospitality**. Here, achievements have been reached with regard to customer sentiment analysis, customer and branch segmentation, price optimization, personalized marketing, optimization and monitoring of the supply chain and in-field sales analytics. Tesco reduced supermarket energy bills with Big Data, while the Walt Disney Company utilized them for theme park management.

In the sector of **Manufacturing & Natural Resources**, while companies like Intel Corporation demonstrated that it is possible to cut manufacturing costs with Big Data, others like Dow Chemical Company also optimized chemical manufacturing. In addition to asset maintenance and downtime reduction, Big Data proved to be helpful also to save natural resources and to optimize supply chain. Even in **Utilities & Energy**, Big Data brought a significant number of advantages in terms of customer retention, forecasting energy, billing analytics, predictive maintenance, maximizing the potential of drilling, and production optimization. Royal Dutch Shell developed data-driven oil fields, while British Gas improved customer service and Oncor Electric Delivery turned to Big Data for intelligent power grid management [62].

Finally, in **Web, Media & Entertainment**, it is possible to identify as applications of Big Data audience and advertising optimization, channel optimization, recommendation engines, optimized search, live sports event analytics, and outsourcing Big Data analytics to other verticals [59]. Twitter cracked down on abusive content, while Netflix improved viewership. Likewise, in the **Telecommunications sector**, investment potential of Big Data is found in network performance and coverage optimization, customer churn prevention, personalized marketing, tailored location based services, and fraud detection. BT Group hunted down nuisance callers and AT&T used Big Data for smart network management [56].

2.2.1 *I-BiDaaS* End-Users

Within the above target group the *I-BiDaaS* end-users include *non-technical business end-users* (strategic or operational managers), who only consume the analytics results, as well as more "technical" roles that configure analytic services and data flows (subject matter experts, *data scientists*).

Furthermore, *I-BiDaaS* users include *data integrators* responsible for interfacing the existing internal enterprise systems with the *I-BiDaaS* system, *administrators* of such systems and *application developers* responsible for the provision or implementation of some analytic application. Such users can be employees of the company, but they could also work for a special service provider.

2.3 Target Group Needs

The *I-BiDaaS* solution aims to address specific needs of the above target groups. These needs fall in the following categories [40, 18, 73], as reported in existing studies:

1. **Data collection and aggregation:** Industrial companies are faced with a rapidly increasing mountain of data and devices that is growing in both quantity and complexity. The platform must allow the collection and aggregation of data and information from the widest possible range of industrial devices and software systems, as well as those from enterprise information systems. It must integrate and normalize different data types (streaming sensor data versus transactional enterprise data), different response times (once per ten milliseconds versus once per day), and different business requirements, reconciling their use at different levels of analysis.

2. **Advanced analytics at the point of need:** Analytic requirements may range from embedded systems that analyze data and respond in milliseconds to complex, predictive modelling analytics, deployed on massive cloud computing infrastructures that can classify terabytes of widely disparate data. The platform also needs to provide a library of standard analytics that allow operators, if they desire, to design, test, and rapidly deploy new analytics.

3. **New deployment models:** Underlying these technology changes are new deployment models, based on cloud and high performance computing, that are further changing the economics of computing and business. Industrial companies need a highly flexible deployment architecture that allows them to mix and match technology deployment methods – and avoid vendor lock-in - as their needs and technological options change.

4. **Extensibility and customizability:** Taking advantage of new revenue opportunities requires industrial operators to adapt quickly to changing customer requirements and dynamic competitive and regulatory environments. This requires the industrial Big Data platform to be highly extensible and based on standardized APIs and data models

that allow it to adapt to new capabilities, new devices, new data types, and new resources as they become available, while still preserving the capabilities that continue to impart value. The platform also has to support an ecosystem of in-house and third party developers that can enhance existing industrial solutions and innovate new ones. The ability of the platform to leverage commonalities of the entire industrial sector must be tempered with the ability to customize the platform to meet specific company, geographical, and/or vertical requirements.

5. **Orchestration:** The industrial Big Data platform must support the orchestration of information, analytics algorithms, and people in order to ensure that the different components of the industrial Big Data world interoperate effectively.

6. **Modern user experience:** This means that the industrial Big Data platform has to deliver the above components within the context of a modern user experience that is no longer bound to a desktop. This includes supporting a wide range of mobile devices and user interaction models, as well as ensuring that the user experience is tailored to the individuals role and requirements at a particular time and location.

2.4 *I-BiDaaS* Position

It becomes evident that a large number of vertical submarkets are already proving to be affected by the Big Data technological advancements not only for necessities of adaptation to the fast-track market but also proactively investing in the present sector to exploit and be ready to future opportunities.

Although several Big Data solutions can be used to address specific needs in the above markets, one can identify a number of common cross-sectorial cases such as Customer Analysis, Security Intelligence, Fraud Detection, Operational Efficiency, Predictive Maintenance, Data-Driven Products and Services. To reap the full benefit of data analytics, organisations require new solutions that will allow more of these enterprise organizations to simplify deployment of data analytics and business intelligence tools, and scale their architecture as needed.

To this end, *I-BiDaaS* will offer a unified Big Data *as-a-service* solution that will be easily configured and adopted empowering non-expert users to easily utilize and interact with Big Data technologies, coping with the rate of data asset growth while at the same time increasing the speed of data analysis.

2.5 *I-BiDaaS* Context

In this section we review and present different Big Data as a self-service solutions that are in relation with *I-BiDaaS* goals and concerns. Furthermore, we discuss the contribution that *I-BiDaaS* platform can bring to the current state-of-the-art. The focus is mainly on open source tools for Big Data that address one or several concerns along the Big Data Value Chain, according to the Reference Model of the European Public-Private Partnership on Big Data Value (BDV) [18]. These concerns are:

- *Data Visualization and User Interaction:* Advanced visualization approaches for improved user experience.
- *Data Analytics:* Data analytics to improve data understanding, deep learning and the meaningfulness of data.
- *Data Processing Architectures:* Optimized and scalable architectures for analytics of both data-at-rest and data-in-motion, with low latency delivering real-time analytics.
- *Data Protection:* Privacy and anonymisation mechanisms to facilitate data protection. This is shown related to data management and processing as there is a strong link here, but it can also be associated with the area of cybersecurity.
- *Data Management:* Principles and techniques for data management.
- *The Cloud and High Performance Computing (HPC):* Effective Big Data processing and data management might imply the effective usage of Cloud and High Performance Computing infrastructures.
- *Internet of Things (IoT), Cyber-Physical Systems (CPS), Edge and Fog Computing:* A main source of Big Data is sensor data from an IoT context and actuator interaction in Cyber-Physical Systems. In order to meet real-time needs it will often be necessary to handle Big Data aspects at the edge of the system.

The discussion is organized in two subsections. The first focuses on research efforts and technical solutions developed in the context of European research and innovation initiatives. The second describes software tools and platforms stemming mainly from the developers community. Although, there exist commercial big data as a self-service tools, e.g. Datameer [30] or Alteryx [4], the discussion in this section is primarily focused on non-commercial approaches aiming to position the *I-BiDaaS* platform as an open source tool for big data as a self-service. It should also be noted that in this section we provide an overview of these efforts; a more detailed discussion can be found in [48].

2.5.1 Related Projects

In this section we have included a number of projects aiming to increase the efficiency and facilitate management of Big Data analytics pipelines also addressing, data protection and user interaction concerns. A distinction is made between: (a) projects having a strong IT focus in that they aim to provide software tools to IT experts (IT infrastructure administrators and application developers), listed in Table 2.1; and (b) projects that have a stronger end-user focus, that is they aim to deliver software solutions that feature user-friendly characteristics enabling non IT experts (data analysts, business users) to access and analyze data contained and managed within the big data ecosystem, listed in Table 2.2.

As *I-BiDaaS* aims to provide an end-to-end big data solution, it shares the concerns and will take into consideration the architectural solutions developed in the above projects especially related to cloud management and orchestration (e.g., projects 2-9 in Table 2.1) as well as to privacy preservation using commodity software (e.g., project 10 in Table 2.1) and will investigate their potential application to the *I-BiDaaS* architecture. Furthermore, *I-BiDaaS* can take advantage of the data analytics programming frameworks also used in these projects. An importance difference with the projects shown in Table 2.1 is that *I-BiDaaS* aims to facilitate analytics for non-expert and as such its objectives are more aligned with the EU projects shown in Table 2.2. Although the application sectors of some of these projects are more specific and different to the ones of the *I-BiDaaS* industrial partners, the analytics and visualization functionalities reported in these projects will also be considered in *I-BiDaaS*.

However, the major difference of *I-BiDaaS* lies in that one of its objectives is to construct a safe experimentation environment for the methodical development of new Big Data products, services, and tools. To this end, it will incorporate the function of fabricating realistic synthetic data in the platform, which can increase the degree of adoption of Big Data technologies in data privacy sensitive sectors like banking. Another difference is that *I-BiDaaS* can further facilitate and speed up times for developing new, less standard machine learning and analytics applications through a sequential programming paradigm, reducing at the developers' side the burden of worrying about parallelization of their tasks.

2.5.2 Big Data Solutions

In this section we focus on existing big data software platforms, presenting a representative subset of solutions that have mutually different focus. In particular, Table 2.3 briefly describes each solution indicating its (main) concerns (according to the BDV Reference Model [18]). A more elaborate description can be found in [48]. As can be seen in this list, with the excep-

tion of Zoe Analytics [101] that also targets data scientists, these platforms aim to address the needs of IT experts, administrators and application developers. The software solutions listed in Table 2.3 will be considered for the design of the *I-BiDaaS* platform as possible approaches for some of its functionalities, more precisely: (a) simpler usage of ML functions; (b) virtualization of analytics; (c) simplifying the execution of programming tasks; and (d) near-real time analysis of data by unifying, in particular for storage, batch analytics and streaming. It should be noted that the above table does not provide an inclusive list as a large number of individual software solutions exists, that could potentially be integrated in a unified open big data platform like the one proposed in *I-BiDaaS*, ranging from data visualization and programming frameworks for batch big data analytics and machine learning toolkits to real-time analytics and resource management solutions. An extensive list of such tools and their relation to the *I-BiDaaS* solution is presented in [48].

Table 2.1: Summary of Big Data EU initiatives targeting IT experts

No	Name	Main Concern	Relevant Objective
1	**IOStack Software Defined Storage for Big Data [52]**	Data Management	To provide a software solution that enables efficient execution of virtualized analytics applications over virtualized storage resources thanks to flexible, automated, and low cost data management models based on software-defined storage (SDS)
2	**FIWARE Future Internet Ware [36]**	The Cloud and High Performance Computing (HPC); IoT, CPS, Edge and Fog Computing	To develop a market-ready open source software, combining components that enable the connection to IoT with Context Information Management and Big Data services in the Cloud.
3	**MClouDaaS (Multi-Cloud Data Analytics as a Service) [66]**	The Cloud and High Performance Computing (HPC)	To enable to use Big Data Analytics in Multi-Cloud scenarios.
4	**mF2C Towards an Open, Secure, Decentralised and Coordinated Fog-to-Cloud Management Ecosystem [67]**	IoT, CPS, Edge and Fog Computing	To address many detailed technical problems at the intersection of fog computing and cloud computing, with a unifying objective of coordinated, open, secure, decentralized and extensible management.
5	**SEA Clouds (Seamless adaptive multi-cloud management of service-based applications) [82]**	The Cloud and High Performance Computing (HPC)	To provide the tools for Modelling, Planning and Controlling cloud applications regardless the underlying provider.
6	**PANACEA [77]**	The Cloud and High Performance Computing (HPC)	To enable proactive autonomic management of cloud resources
7	**Deep Hybrid Data Cloud [32]**	The Cloud and High Performance Computing (HPC)	To integrate the intensive computing services under a Hybrid Cloud approach.
8	**CLASS Edge and Cloud Computation: a Highly Distributed Software for Big Data Analytics [25]**	The Cloud and High Performance Computing (HPC); IoT, CPS, Edge and Fog Computing	To develop a novel software architecture framework to help big data developers to efficiently distributing data analytics workloads along the compute continuum (from edge to cloud).
9	**datACRON [28]**	Data Visualisation and User Interaction; Data Analytics; Data Processing Architectures; Data Management	To provide a set of solutions aiming to increase our abilities to acquire, integrate, process, analyze and visualize data-in-motion and data-at-rest.
10	**SecureCloud [83]**	The Cloud and High Performance Computing (HPC); Data Protection	To provide an ecosystem of cloud facilities characterized by superior security guarantees, providing protection from attacks by privileged users (e.g. the cloud provider or the system administrator) and software (e.g. the hypervisor).

Table 2.2: Summary of Big Data EU initiatives targeting Non-experts

No	Name	Main Concern	Relevant Objective
1	**Big Data Europe Empowering Communities with Big Data Technologies [21]**	Data Visualisation and User Interaction; Data Processing Architectures; Data Management	To deliver an open source Integrator Platform to let organizations in the Climate, Energy, Food, Health, Transport, Security, and Social Sciences to experiment with different big data technologies with minimal effort.
2	**DataBio [27]**	Data Analytics; Data Processing Architectures; Data Management; The Cloud and High Performance Computing (HPC); IoT, CPS, Edge and Fog Computing	To build and pilot a big data technology platform, based on existing technologies and data sets, in collaboration with the end users and proceed to verify the concept through several pilotings in the agriculture, forestry and fishery/aquaculture sectors.
3	**TT Transforming [95]**	Data Analytics	To provide concrete, measurable and verifiable evidence of data value that can be achieved in mobility and logistics.
4	**SPECIAL (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance) [88]**	Data Visualisation and User Interaction; Data Protection	To develop a Big Data Policy Engine, which will extend big data architectures to handle Linked Data, harness them with sticky policies as well as scalable queryable encryption, and develop advanced user interaction and control features.
5	**AEGIS [2]**	Data Visualisation and User Interaction; Data Analytics; Data Management; Data Protection	To deliver an open, secure, privacy-respectful, configurable, scalable cloud based Big Data infrastructure as a Service to the Public Safety and Personal Security related industries.
6	**Toreador [93]**	Data Visualisation and User Interaction; Data Analytics; Data Management; Data Protection	To develop a Big Data Analytics as-a-service approach based on model-driven set-up and management of Big Data analytics processes.

Table 2.3: Summary of software systems and solutions for Big Data

No	Name	Main Concern	Description
1	**Zoe Analytics [101]**	Data Visualization and User Interaction	An open-source user-facing application that sits on top of a container engine and makes it simple for data scientists and system administrators to efficiently use available resources and run the latest data intensive frameworks.
2	**Apache Samoa [12]**	Data Analytics	A distributed streaming machine learning (ML) framework that contains a programming abstraction for distributed streaming ML algorithms.
3	**Apache OpenWhisk [11]**	Data Processing Architectures	Open source, scalable, reliable, and robust serverless platform.
4	**Pravega [79]**	Data Management	Open source storage primitive.

3 I-BiDaaS Requirements

This section focuses on the process of capturing and consolidating the requirements of all *I-BiDaaS* use cases that will guide the technical solution of the project. To this end, it first provides an overview of the *I-BiDaaS* use cases. Following it presents the requirements engineering methodology that was used in *I-BiDaaS*. Finally, it presents the consolidated list of system requirements that resulted from the application of this methodology in the *I-BiDaaS* use cases.

3.1 *I-BiDaaS* Use Cases

The *I-BiDaaS* design is driven by a number of real use case scenarios (summarised in Table 3.1) within the three industrial sectors where the project's industrial partners are involved, namely: telecommunications sector, financial sector and manufacturing sector. Although company-specific, these scenarios deploy Big Data analytics in order to address problems that cross sectorial borders such as customer (social) analysis (use cases 1, 6, 7), data-driven services (use case 3), fraud detection (use case 4), operational efficiency (use cases 2 and 9) , predictive maintenance (use case 8). Similarly, the scope of the use cases cover generic objectives which relate to: introducing Big Data analytics to improve current operations (use cases 3,8 and 9), improve efficiency of Big Data analytics solutions (use cases 1 and 2), improving decision making with respect to implementing Big Data analytics solutions (use cases 4-7).

The following sections 3.1.1 - 3.1.3, provide a presentation of the 9 industrial test cases whose consolidated list of user requirements will be transformed to pilot implementation of the *I-BiDaaS* platform. A detailed description of the dataset for each use case is provided in [50].

Table 3.1: Overview of *I-BiDaaS* use cases

No	Use Case Title	Data Set	Big Data Deployment	Vertical Area	Scope
1	Accurate location prediction with high traffic and visibility	Mobility data from customer phones;	Customer Analysis	Telecom	Improve efficiency of Big Data analytics
2	Optimization of placement of telecommunication equipment	Mobility data from customer phones;	Operational efficiency	Telecom	Improve efficiency of Big Data analytics
3	Employment of bots in call centre	Call centre data containing hours of recorded speech	Data driven services	Telecom	Introduce Big Data analytics to improve current operation
4	Enhance control over third party agencies	Security and Event Management data	Detection of illicit usage of enterprise services	Financial	Improve decision making with respect to Big Data analytics solutions
5	Advanced analysis of bank transfer payment in financial terminal	Bank transfer data	Fraud detection	Financial	Improve decision making with respect to Big Data analytics solutions
6	Analysis of relationships through IP address	IP address data about any operation of a customer can do in a bank session	Customer social analysis	Financial	Improve decision making with respect to Big Data analytics solutions
7	Building of a social graph	Social graph data, related to the relationships between customers	Customer social analysis	Financial	Improve decision making with respect to Big Data analytics solutions
8	Maintenance and monitoring of production assets	SCADA data and Manufacturing Execution System data	Predictive maintenance	Manufacturing	Introduce Big Data analytics to improve current operation
9	Production process of Aluminium Casting	Data is collected from various sensors and Operators data	Operational efficiency	Manufacturing	Introduce Big Data analytics to improve current operation

3.1.1 Telecommunication Industry Cases

With the rapid expansion of smart phones and other connected mobile devices, telecommunication service providers need to rapidly process, store, and derive insights from the diverse volume of data travelling across their networks. Big Data analytics can help providers improve profitability by optimizing network services/usage, enhancing customer experience, and improving security.

Data-driven improvement of services or product is a key: telecoms need to share data between cell towers, users and processing centres, and due to the sheer volume of this data, it is important to process it near the source and then efficiently transfer it to various data centres for further use. New distributed messaging systems are required to effectively transport huge amounts of data and to make this data available with reliable geo-distributed replication across multiple data-centres. Additional challenges include real-time deep packet inspection to optimise traffic routing and steer network quality of service, real-time call data record analysis to identify fraud, event-based marketing campaigns that use geo-location and social media, allowing differentiated responses, and many more.

In the *I-BiDaaS* project we will be working with the Telefonica I&D (TID) company. TID is the main innovation catalyst inside the Telefonica Group, one of the world's largest telecom companies. TID conducts applied investigation activities, conceptualises new uses for communication and develops innovative products and services for the Telefonica Group with the aim of strengthening the Group's competitiveness through technological innovation. The TID test cases to be used in the *I-BiDaaS* project are described below:

1. **Accurate location prediction with high traffic and visibility:**

   Telefonica wants to predict places with high traffic and congestion events in order to optimise their resource distribution. Telefonica has a large amount of data on so-called *cell network events* generated by transactions of mobile phones. A transaction can be, for instance, placing or receiving a call, sending or receiving an SMS, asking for a specific URL in your mobile phone browser, or sending a text message or a data transaction from/to any mobile phone app. These events are picked up by the antennas that are closer to the mobile phone, thus providing an approximate location of the device.

   In order to predict high traffic locations, Telefonica wants to group aforementioned data by user and sort the events chronologically. Dealing with such large-scale, variable-length, sequences in a timely fashion represents a significant challenge. In addition, sparsity represents an added problem for this type of data: not all timestamps contain data; there can be several "temporal holes" in the event sequences

that any stream processing system needs to handle. Two important tasks that stem from these challenges are to: (1) interpolate missing events to recover plausible event trajectories, and (2) forecast immediately next events to anticipate movements at scale. Ideally, Telefonica would like to construct a model that can perform the two functions while compacting the information into its (hopefully fewer) internal parameters. By the previous process we envisage minimum processing time with growing data sizes, while maintaining real-time requirements. The expected results would be valuable insights on the audience, understanding the behaviour of local and non-local customers over various periods of time (e.g holidays), and extract insight on the behavioural patterns of groups of people, enabling Telefonica to optimise their value propositions.

2. **Optimization of placement of telecommunication equipment:** One of Telefonica's aims is the optimisation of the network operations by providing caches and identifying optimal antenna locations. While some effort towards optimisation is already in place, the requirement would be to further analyze the data and provide new insight into how the network could be further optimised given the provided data from customer usage. The objective is to further analyse the dataset described in the previous test case in order to improve the routing and placement of the telecommunication equipment that is already in place, or to arrange accordingly the new equipment obtained.

3. **Employment of bots in call center:** One of the main aims of Telefonica is to improve the customer experience by using advanced machine learning techniques. As part of improving the customer experience, Telefonica employed voice activated bots in their call centers to help customers accomplish tasks related to network configuration and operation. Telefonica would like to improve upon the current implementation of bots at the Call Centers by utilizing and analysing the customer Call Center conversations and thus creating bots that are more helpful, more anthropomorphic, and realistic. There is a database created aiming at developing speech technologies that transform audio calls into relevant information for the Call Center that can be used to assess its performance and/or to automatically screen phone calls. By facilitating the results of the project, Telefonica plans to improve the number of audio calls that can be processed per time unit.

3.1.2 Financial Industry Cases

The financial industry is sitting on a vast reservoir of data and information that can be leveraged for product development, personalized marketing,

and advisory benefits. Key areas in this industry are fraud and risk management. In this context, advanced analytic algorithms on Big Data can be used to predict and avoid security breaches. With cyber security more important than ever, falling behind in the use of data for security purposes is not an option. Also, real-time view and analysis is critical towards competitive advantage in the financial/banking sector.

The company we are working with to analyse and study this industry is CAIXA bank. CAIXA and other companies within the financial industry are facing risk of loss of control when deploying strategic services outside their premises. A phenomenon often documented as the emergence of new technologies, such as cloud computing, mobile banking, Big Data, social networks, etc., require maximum level of specialization on several areas (like IT, legal, risk analysis and management, innovation planning, etc.) that can drain the existing workflows of companies. The internal departments of CAIXA cannot cope with such a great demand of skills and the competitive costs of the aforementioned technologies. As a result, part of the services provided by companies are outsourced to third party agencies.

To retain control over third party agencies, we need to analyse technical logs of applications and technical controls, such as firewalls, routers, authentication servers, application servers, which requires analysis of very large size of information. All this information is coming from very different sources. This increases the variety of formats and huge amount of information is produced, so velocity is an issue as well.

Another area of interest is social modelling of users in order to detect fraud in money transfers. In the daily operation of a bank, a significant problem to avoid is the potential fraud related to money transfers among users. That is, while the vast majority of money transfers are actually consented by the two users involved, a small fraction of them may be a result of illicit activity. Certainly, CAIXA invests resources to prevent and detect potentially illicit money transfers in order to avoid money loss, compensations, and reputational damage.

To this end, a graph model can be constructed that represents relationships among users. In such a graph, relationships or links among users are built based on different information sources. For instance, they can link users depending on whether they have performed transfers among them in the past, if they belong to the same family, or even if they access to CAIXA online services with the same IP address. Therefore, instead of looking for suspicious transfers, CAIXA can find relationships or "paths between users, which indicate that a given money transfer may be consented and does not represent a problem. This approach filters out the vast majority of daily money transfers. Afterwards, a small number of money transfers that are not supported by any of the modelled trust relationships among users can be examined to infer whether they may be a source of fraud or not.

In this context, the CAIXA test cases representing the financial industry are the following:

1. **Enhance control over third party agencies:** As mentioned above, Caixa bank has a lot of providers offering different services, hence it is crucial to control any third party activities. Currently, Caixa bank has well-defined policies on these activities, nevertheless the monitoring of them is critical in order to test and check security requirements. More specifically, this use case is focused on detecting if two different providers are connected, or trying to connect to Caixa bank using the same credentials. This use case will be performed over synthetic data.

2. **Advanced analysis of bank transfer payments in financial terminal:** The financial terminal is an internal application used by Caixa bank employees to manage customers accounts and provides them several services, such as transfer between banks. The objective of this use case is to detect fraud transactions by identifying differences between reliable transfers and possible fraudulent cases, for example when the transfer is done outside normal physical office hours. This use case will be performed over synthetic data.

3. **Analysis of relationships through IP address:** Caixa bank stores information of their customers and the operations they do (i.e. bank transfers, checking of accounts, etc.) using several channels, such as mobile apps and online banking. Caixa intends to use this information in order to identify new kinds of relationships between customers using their IP addresses that might indicate whether there is suspicion of potential fraud. This use case will be performed over synthetic data.

4. **Building a social graph:** Caixa bank wants to build a social graph to test the performance of databases oriented to graph analysis. Due to internal policies, it is hard to evaluate which technological solutions are the best to be used, so the objective is to develop, in a systematic and consistent way, a benchmark of graph databases. This use case will be performed over synthetic data.

3.1.3 Manufacturing Industry Cases

Automated processes and mechanisation of the manufacturing industry resulted in the generation of a large amount of data. Despite the potential benefits many companies cannot fully exploit Big Data to transform their operations as they don't have the knowledge and experience to utilise the data they are producing.

In manufacturing, companies can use advanced analytic tools to take a deep dive into historical process data, identify patterns and relationships

among discrete process steps and inputs, and then optimise the factors that prove to have the greatest effect. Many global manufacturers in a range of industries and locations now have an abundance of real-time shop-floor data and the capability to conduct sophisticated statistical assessments. They just have to take previously isolated data sets, aggregate them, and analyse them to reveal important insights. Insights that can transform their operations, optimise their processes and increase their effectiveness and efficiency.

The manufacturing company that will be studied in the *I-BiDaaS* project is the CENTRO RICERCHE FIAT (CRF). CRF is one of the leading applied research centres in the automotive industry. The CRF test cases to be used in the *I-BiDaaS* project are described below:

1. **Maintenance and monitoring of production assets:** The main objective of CRF is to use Big Data to efficiently produce/reconfigure plans (in case of orders, pre-orders, and other incidents). Regarding production, a re-plan is generated (up to several times per day) about the sequence of vehicles going through 'body in white', 'painting' and 'assembly' phases. A possible interruption may occur due to material delay, material quality issues, or material unavailability. There are also possible interruptions due to process issues such as machine breakdown, and unscheduled maintenance. The information about the quantity of the materials, back-up systems and their availability or usage is always kept in our systems. CRF envisages that the proposed solution will be able to combine all this information and proceed to real-time re-planning of the procedures needed so that the manufacturing process continues without interruption and avoid excessing costs and financial damage. Data analytics would have to work in near real-time speed. The final system should be able to produce real-time automated decisions (e.g. re-planning).

2. **Production process of Aluminium Casting:** (engine casting) This use case concerns the quality issues on the cylinder engine. There are many sensors installed on each engine. This is aluminium casting, which is much more difficult than traditional casting. The main objective is to link all the data during the production process to engine defects. Each produced engine is completely evaluated and visually inspected. The goal is to correlate defects with the production process characteristics.

3.2 Methodology Used for User Requirements in *I-BiDaaS*

This section describes the *I-BiDaaS* way-of-working towards elicitation, analysis and reporting of the requirements for all experiments and types of stake-

holders. It also provides a classification (typology) of all key concepts, for describing the *I-BiDaaS* requirements.

Requirements engineering (RE) is "the process of discovering that purpose, by identifying stakeholders and their needs, and documenting these in a form that is amenable to analysis, communication, and subsequent implementation" [76]. The phase of requirements is of critical importance for the *I-BiDaaS* project aiming to ensure that the project will deliver a technical solution whose components are aligned with the business goals of the industrial partners.

To this end, we first reviewed existing literature on RE methods for Big Data applications. Based on this analysis we then devised a framework to elicit different requirement categories, namely business requirements, user requirements and system requirements described bellow.

3.2.1 Big Data Requirements Engineering

Despite a large coverage of the requirements gathering domain in the literature, only a limited number of research papers focuses on gathering requirements specifically for Big Data applications. In traditional software development, the development processes are usually well-established. On the other hand, the processes for the development of applications involving Big Data are not yet well established in the scientific literature. This is because new challenges emanate when developing applications with Big Data in mind. Considering the computing complexity involved in Big Data applications and data characteristics such as volume, variety, veracity, and velocity make requirements engineering an emerging research area.

In this section we discuss the reasons why requirements engineering in the context of Big Data applications is yet under represented in the literature. Then, we review related research on the topics that can help us formulate a framework for capturing and analyzing the requirements of *I-BiDaaS* end users.

Requirements Engineering (RE) for Big Data Applications is a challenging task due to:

- the characteristics of Big Data (e.g., volume, velocity, variety, veracity, value) [60],
- the rapid technological changes,
- the difficulty of selecting Big Data technologies [24],
- the complex integration of new and old systems,
- the difficulty in matching available frameworks and technologies with system requirements [6].

Addressing Big Data challenges is prompting organisations to rethink their basic assumptions about the relationship between business and IT and their respective roles [31]. Data needs to be processed in real-time otherwise incoming data could become lost and/or obsolete (velocity); mountain-ranges of historical data may exist (volume) and the end-user application system needs to be scalable to be able to cope with increasing size; data can be structured or unstructured (variety) and needs to be associated or aggregated in innovative ways to create new value for business; historical or streams of data may need to be cleaned up prior to analysing it (veracity).

Arruda *et al.* in [17] investigate the types of requirements in Big Data applications. They found requirements falling under the categories of functional, quality and data type of requirements.

Functional requirements describe what a system should do, how the system should react to potential inputs, and how the system should behave in particular situations [57].

Quality requirements cover the quality attributes that a Big Data application must address, such as privacy and security, performance, availability, scalability, consistency, elasticity and low latency.

Having the right specification of data requirements is important for defining some of the systems functional requirements (e.g., systems needs to support diversified output file formats for visualization, rendering, and reporting; systems needs to support legacy, large, and advanced distributed data storage, etc.).

Orthogonal to the above is the dimension of the requirements elicitation process. Recent literature, suggest the use of a goal-oriented process [70, 35], whereby the elicitation of Big Data requirements is seen as the systematic transformation of high-level goals to specific requirements that operationalise these goals. Such approaches aim to a better alignment of analytics with business strategies, a lack of which can result in unclear expectations of how analytics contribute to business strategies, lack of executive sponsorship, and analytics project failures.

3.2.2 The Requirements Analysis Framework

Given the aforementioned challenges in Big Data RE, a framework was devised aiming to capture and document the requirements of the end-users of the *I-BiDaaS* platform.

Figure 3.1 encapsulates our framework on requirements analysis in three interrelated categories (or views): business requirements, user requirements and system requirements.

Business requirements describe specific needs of a company that must be addressed in order to achieve an objective. They relate to business' objectives, vision and goals. They provide the scope of a business need or

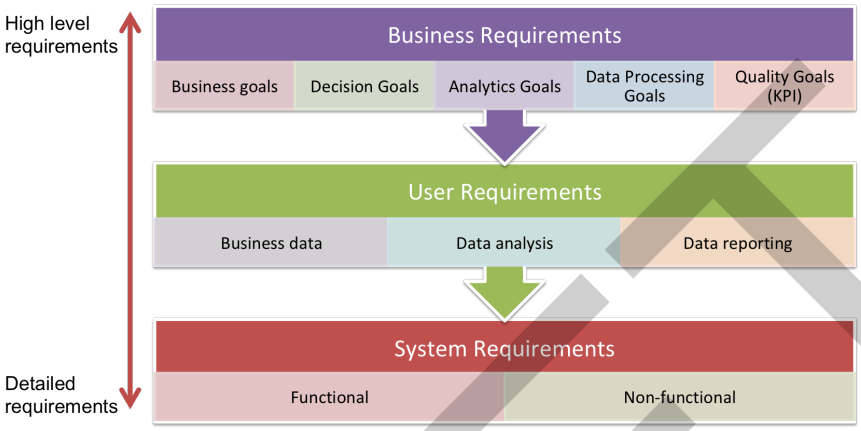

Figure 3.1: Requirements analysis framework

problem that needs to be addressed through a specific Big Data analytics activity or project [57].

The business requirements view aims to facilitate the elicitation and clarification of requirements in the business context and ensure the alignment of business and Big Data analytics strategies. The main elements of this view starting from high-level business goals and moving down to the analytics goals that satisfy them are:

- **Business Goals:** Express company vision. They may refer to strategic objectives (e.g., *to improve customer satisfaction*) or operational goals (e.g., *to improve routing and placement of telecommunication equipment*)
- **Decision Goals:** Describe how Big Data will be used in order to achieve business goals, e.g., *Use Big Data to employ bots in call centres*
- **Analytics Goals:** Capture the type of analysis (algorithm) required in order to realise decision goals, e.g., *Analyse the behaviour of local and non local customers over periods of time*
- **Data Processing Goals:** Similar to analytics goals. Focus on the Big Data lifecycle activities (i.e., acquisition, preparation, analysis and Visualisation). Examples include *Automate audio call transformation and metadata generation* and *Enable the automatic extraction of behavioural patterns of customers*
- **Quality Goals:** Quality properties (such as security, performance, scalability, etc.) that may relate to any other type of goal, e.g., *Improve performance of audio call processing*.

- **Key Performance Indicators (KPIs):** Indicators (e.g., computational cost, infrastructure cost, time, etc.) for evaluating the achievement of all objectives, often led by quality goals. For example, the *Data processed per time unit* indicator relates to the quality goal mentioned above.

The user requirements view describe the needs of a particular stakeholder or group of stakeholders. In the context of Big Data analytics, stakeholders include [72]:

- **Data Source Providers:** Introduce new data or information feeds into the Big Data analytics system.
- **Big Data Capability Providers:** They include Big Data application providers, the system orchestrator and the Big Data framework providers. Application providers supply the applications that execute the Big Data life cycle activities. The system orchestrator, integrates the different data application activities into an operational vertical system. Finally, the Big Data framework provider, provides the necessary infrastructure (computing, storage, network, etc) to establish the computing environment in which to execute Big Data applications.
- **Data Consumers:** Business users or other systems that use the Big Data analytics services and results.

Based on the above classification, the requirements in this view are divided into:

- **Data source requirements:** requirements relevant to the origin and characteristics of data (including the 3Vs), e.g., *Audio meta-data will be stored in a unified message format*.
- **Analytics capability provider requirements:** Describe requirements with respect to the data analytics services, e.g., *Interpolate missing events to recover plausible events trajectories*.
- **Data consumer requirements:** Describes user interface requirements, e.g., *Analytics results will be viewed using desktop/laptop devices*.

The final view relates to the system requirements, i.e., descriptions of the behaviour that the system (or a system component) should expose, or the capabilities it should own in order to realize the intentions of its users. This behaviour is expressed in terms of the system **functional requirements (FRs)**, stating what the system should do (e.g., *The system should enable the integration of attribute and transaction level data from a variety of internal sources*), as well as **non-functional requirements (NFRs)**, denoting

the conditions under which the behaviour indicated by the functional requirements will be executed, e.g., *The system should assure the protection of classified and personally identifiable information*.

FRs and NFRs result from the structured analysis of the previous requirements categories, as described in Section 3.4. These are both used to drive the system architecture, an overview of which is provided in Section 4 and detailed in [49].

3.2.2.1 Typology and Key Concepts

An important aspect of the RE framework is the requirements typology shown in the concept map of Figure 3.2. It provides a uniform representation of all key RE concepts, thus facilitating the consolidation of user requirements and the identification of cross-sectorial industrial challenges.

The requirements typology is a classification of key RE concepts according to their characteristics. It was constructed using mainly a top-down strategy, based on the concept categories and associated characteristics that were identified through the review of relevant literature [74, 71, 73, 98]. The concept categories were further revised in a bottom-up manner through the generalization of the data gathered by the users.

The following is a detailed list of the concepts categories and their definition.

- **A - High level (business) requirements:** Describe what the organization wishes to achieve.
  - *Business Goals:* Express company vision. They can be Strategic or Operational.
  - *Decision Goals:* Describe how Big Data will be used in order to achieve business goals. They can be of the following type: One-time analysis for a specific business decision; Open "blue sky" exploration of "interesting" data; Ongoing reporting and business intelligence.
  - *Analytics Goals:* Relate to the type of analysis (algorithm) required in order to realize decision goals. They can fall in one of the following categories: Historical analysis and reporting (descriptive / diagnostic); or Predictive analysis and reporting (predictive / prescriptive).
  - *Data processing Goals:* Similar to analytics goals. They focus on the data processing lifecycle activities, i.e., Acquisition, Preparation, Analysis and Visualisation.
  - *Quality Goals:* Quality properties that may relate to any other type of goal. They may express concerns such as security, performance, scalability, reliability, availability, etc.

  - *KPIs:* Indicators for evaluating the achievement of all objectives, often led by quality goals. Examples are: Computational Cost, Infrastructure Cost, Time, etc.

- **B - User requirements:** Describe the requirements of the different stakeholders (data owners, capability providers, data consumers).

  - **B1 - Data source requirements:** Describe requirements relevant to the origin and characteristics of data (including the 3Vs). They might differ for different data sources.

    * *Data source:* Describes the origin of data. This might include: Internal IT applications, Log files, Sensors, Mobile devices, Social Media, Data feeds, Government and special interest groups, Commercial Data Providers.
    * *Data owner:* Differentiates between internal and external data sources.
    * *Data content:* Describes the kind of data contained in the data set. E.g., attribute level data (data that can be used to identify an entity), or transaction level data (data generated by an entity).
    * *Data type:* Refers to the nature of data, e.g., structured, semi-structured, unstructured.
    * *Data format:* Refers to the style of data, such as: relational tables, spreadsheets, XML, JSON, NoSQL, plain text, e-mail, social media, website, mobile data, communications, media files, business apps files.
    * *Data storage:* Refers to the way data is currently stored (e.g., SQL database, NoSQL DB, based on Files, other), specifying the exact name of the database or other system used (if applicable).
    * *Data consumption*: Refers to the way data can be acquired from source(s), e.g., NFS, SQL query, HDFS.
    * *Data distribution:* Indicates whether data source(s) is real-time, distributed, centralized, local, or remote.
    * *Big Data characteristics:* Describe the properties of the (raw) data.
      · *Volume:* Refers to the amount and size of data that should be analyzed and processed. Possible values: NULL (not applicable), 0, Terabytes, Petabytes, Exabytes.
      · *Velocity:* Refers to the speed that data is produced. Possible values: NULL, 0, Periodically Batch, (Near) Real-time, Streaming.

- · *Variety:* Refers to the diversity of the data (formats). For example if data is stored in XML and CSV, then the number of formats is 2. Possible values: NULL, 0, 5-10 Formats, 10-25 Formats, 25+ Formats.
- · *Variability:* Refers to the amplitude of the data flow rate over time (data rate). It describes the inconsistent speed at which Big Data is generated. An example of highly variable data includes social media, where sentiments and trending topics change quickly and often. Possible values: NULL, 0, Low, Medium, High.
- · *Volatility:* Similar to variability. Refers to the structural changes of data over time (nature of data). An example of highly volatile data includes natural language data. Possible values: NULL, 0, Low, Medium, High.
- · *Consistency:* Refers to consistency between multiple copies of data on multiple machines in distributed systems, or between same data on multiple sources. Possible values: NULL, 0, Strict consistency, Eventual consistency, General weak consistency.

* *Data sensitivity:* May refer to classified information, personally identifiable information, etc.
* *Degree of privacy required:* Possible values include: Data can be revealed only to a customer, data stays inside your business organization, data can be shared with partner companies, data can be made publicly available.

– **B2 - Analytics capability provider requirements:** Describe requirements with respect to the data analytics services.

* *Type of analysis:* Refers broadly to tools and algorithms currently in use / considered for use for processing the data at any stage of analysis.
  - · *Data to information stage:* Refers to converting raw data into information. It includes traditional data preparation algorithms such as: Reduction, Cleaning, Transformation, Integration.
  - · *Information to knowledge stage:* Refers to extracting knowledge from information. It includes machine learning algorithms such as prediction, classification and validation.
  - · *Knowledge to wisdom stage:* Refers to applying knowledge to make decisions. It may include: simulation, optimization, decision support.
* *Data Jurisdiction:* Possible values include: Company, Country, European, extra -territorial.

    - *Data interoperability:* Refers to integration requirements with other sources.
    - *System interoperability:* Refers to interoperability requirements with other systems.
  - **B3 - Data consumer requirements:** Describe user interface requirements.
    - *Access control:* Specifies access control requirements, such as: basic role based access control, access control at application level, access control at resource level, additional privacy aware access control; whereby different roles may include: public user, VIP user, internal/external developer, internal/external support staff, internal operator, etc.
    - *Visualization:* Refers to the way the data is viewed by an analyst making decisions based on the data. Possible values: Table, Chart, Graph, Map, Network layout.
    - *Type of interaction:* Refers to the way decision makers interact with the data, e.g., Overview, Zoom, Filter, Details on Demand, Streaming results.
    - *Device:* Refers to accessing devices including mobile devices.
    - *Speed of delivery:* E.g., daily, hourly, near-real time.
  - **B4 - Additional requirements:** Allows for the inclusion of requirements that do not fall under any of the above categories.

3.3 Requirements Analysis Way-of-Working

A common way-of-working towards the analysis of user requirements was adopted in all use cases, involving three activities, namely: scoping, elicitation and validation. The first activity focuses on describing the business context to discover requirements in the large by documenting the high-level business goals and associated KPIs as terms of reference. The aim of this activity is to focus users' attention on where the boundaries of the solution under investigation should lie, and helps to identify at least an initial scope for the system.

Elicitation, follows a mostly top-down approach whereby business goals are further refined/decomposed in order to identify specific user requirements, whose analysis results in the definition of system requirements.

Finally, during validation the documented requirements were critiqued during face-to-face meetings and teleconferencing sessions with the participation of requirements analysts, users and technology providers in order to getting all parties involved to understand the implications of the requirements specification and then agree, i.e. validate, that it accurately reflects their wishes.

The above activities were carried out in an iterative manner resulting in a stepwise refinement of the results being produced.

The whole process was facilitated by the use of appropriate Questionnaire (see Appendix A), that allowed us to collect information from several stakeholders in relatively short time, especially as stakeholders, whose input was needed to establish the requirements were spread out geographically.

Additionally, examples of the requested fields to fill, as well as guiding questions, were provided in order to ensure overall comprehension of what is needed to be captured. Finally, in the cases that information on the requirements was available (collected in the context of *D1.1 Project setup* [48]) this was used to partly pre-fill the questionnaires and minimise end-users' effort. Evidently, they were asked to check pre-filled fields and ensure that documented information is valid and accurate.

3.4 Consolidated Requirements of the *I-BiDaaS* Platform

The process of requirements' consolidation involved two steps. The first step has been to classify the information gathered using the RE questionnaire according to the RE concepts typology, as shown in Figure 3.3.

Figure 3.3 shows an excerpt of the Data source requirements elicited using the RE questionnaires. Shown in italics, each use case (UC) specific requirement is classified using the typology. For example, in UC1 "*Data come from customer phones*" is classified as *mobile devices*, whilst UC8 "MES" is classified as *Internal IT applications* and "*SCADA sensor data*" as *Sensors*.

The second step is based on the classification to aggregate each use cases specific requirements into generalized system requirements that are sector neutral and technology agnostic. For example, the previous user requirements can be generalized into the following system requirement: "*The system should enable aggregation of data coming from a variety of data sources*".

The above process resulted in the generation of 10 generic system requirements (6 FRs and 4 NFRs) shown in Table 3.2, summarizing over 300 use case specific requirements from 9 use cases.

A detailed list of all requirements is provided in Appendix B. It should be noted that for devising the consolidated system requirements several revisions/additions were made to the original list of requirements and their classifications, in multiple iterations between users and analysts, through a creative and cooperative process.

3.5 Summary

Requirements elicitation has resulted in a set of requirements in three categories that define the required system capabilities, for achieving the user requirements.

Following this stage the design of the *I-BiDaaS* system involves the identification of the the functional components seeking to achieve these requirements and their interactions and this is the subject matter of the following section 4.

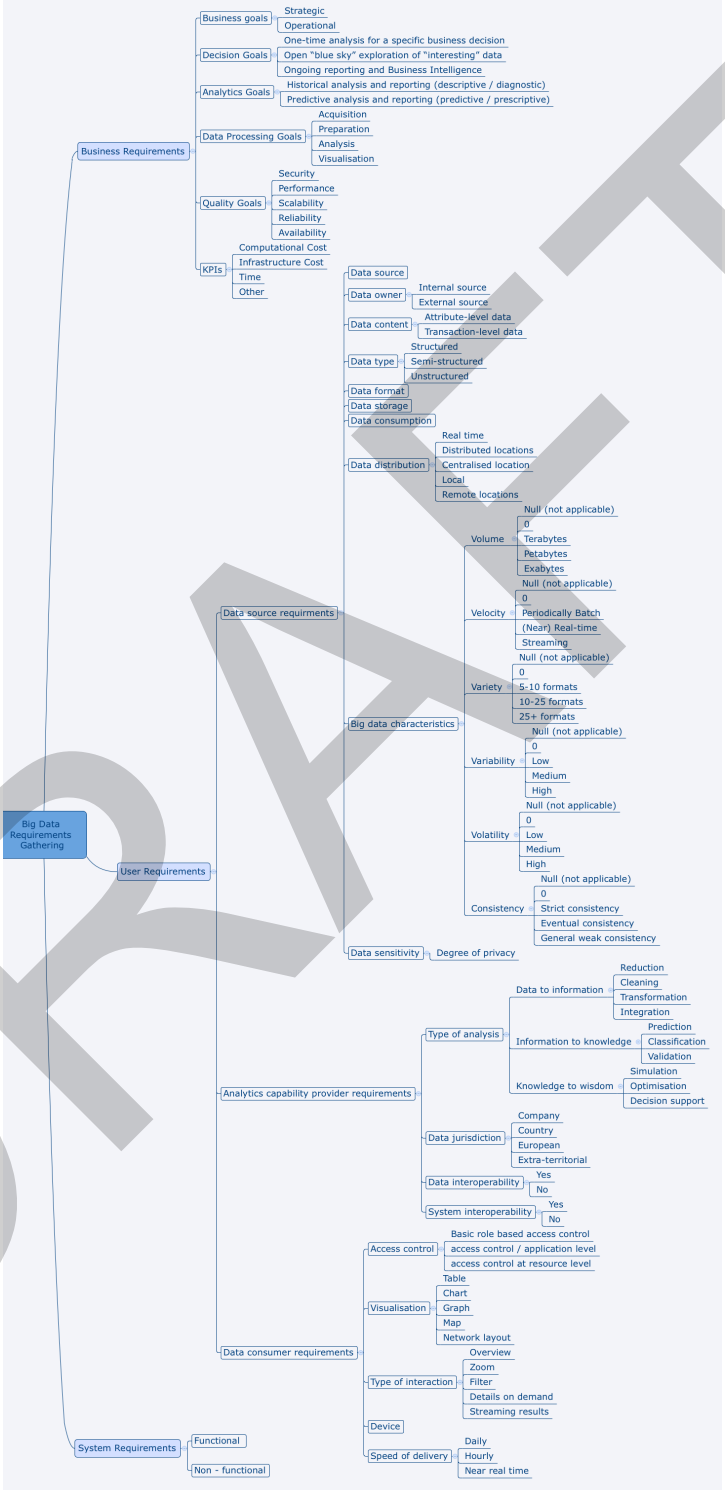

Figure 3.2: Concept map encapsulating the RE typology

Use Case Title	Data Source requirements							
	Data origin		**Data Owner**		**Data Content**		**Data Type**	
UC 1: Accurate location prediction with high traffic and visibility	Data come from customer phones.	*Mobile devices*	TID	*Internal*	Event Data	*attribute level data, and transaction level data*	event data type	*structured*
UC 5: Advanced Analysis of bank transfer payment in financial terminal	The data come from a dataset that is maintained at CaixaBank Warehouse.	*Internal IT applications*	Data are owned by CaixaBank.	*Internal*	Synthetic data, ensuring the privacy of customers, will be generated with CaixaBank Warehouse containing content such as: Customer information, Value of the transference, Office, Etc.	*attribute level data and transaction level data*	Structured data.	*structured*
UC 8: Maintenance and monitoring of production assets	MES (Manufacturing Execution System)	*Internal IT applications*	FCA/CRF	*Internal*	Data about the production line containing the id of the vehicle, the time that the vehicle arrived at each phase of the production line, the characteristics of the vehicle.	*attribute level data and transaction level data*	Structured data	*structured*
	SCADA sensors data	*Sensors*	FCA/CRF	*Internal*	Sensor data	*transaction level data*	Structured data	*structured*

Figure 3.3: Requirements Consolidation using the RE typology

Table 3.2: Consolidated system requirements

No	System Requirement	Type
Data source requirements		
FR1	The system should enable the generation of anonymized and synthetic data to enable safe experimentation and testing	FR
FR2	The system should enable aggregation of both attribute level and transaction level data coming from a variety of internal data sources and in multiple formats	FR
FR3	The system should be able to accommodate data sets that feature high volume, high velocity, high variety, high variability, high volatility and high data sparsity	FR
NFR1	The system should ensure security of sensitive data	NFR
Analytics capability provider requirements		
FR4	The system should support diversified, analytic processing, machine learning and decision-support techniques to support multiple stages of analysis.	FR
FR5	The system should support interactive data analysis.	FR
NFR2	The system should support near real time analytics performance	NFR
Data consumer requirements		
FR6	The system should support diversified visualization and interaction of results on both desktop and mobile devices.	FR
NFR3	The system should support near real-time updating of results	NFR
NFR4	The system should support multilevel access control at resource and application level	NFR

4 I-BiDaaS Architecture

4.1 Overview

The design of a software system architecture involves partitioning a system into functional modules seeking to satisfy the system functional and non-functional requirements, taking into account additional quality attributes and constraints specified in best practices in software and distributed systems development (such as portability, maintainability, loosely coupled, component based, service-oriented, among others). Once the main modules are identified, their interactions (interoperability) are defined and each module can be developed separately.

The architecture vision of the *I-BiDaaS* platform is guided by the desire to support Big Data as a self-service and is driven by the requirements of 9 industrial-validated, real life cases.

In light of the above, the *I-BiDaaS* project has identified potential state-of-the-art gaps in tools and technologies for Big Data as-a-self-service and carried out a detailed analysis of end user requirements, as described in sections 2 and 3, respectively. These are the key input of the following task which is the software architecture design (described bellow), which then guides the implementation of the software system. It should be noted that the architecture design presented in the following sections may be refined and updated at later stages of the project, as needed.

4.2 Conceptual Architecture of *I-BiDaaS* Platform

The starting point of architectural design is the conceptual architecture. In the conceptual architecture the required system functions are organised according to the selected architectural style(s).

The *I-BiDaaS* conceptual architecture, shown in Figure 4.1 follows a layered architectural style consisting of three principal layers: the *infrastructure*

layer, the *distributed large-scale layer*, and the *application layer*. This division into "vertical" layers is conceptual and serves to better organize the involved complex processes and entities. The list of modules that the platform involves, their respective layer and role in the platform, as well as the current development status, is presented in Table 4.1 below. Optional, additional modules will be considered, such as the Terracotta database by SAG [92] (as a possible alternative for Hecuba tools), and the Integration Server by SAG [51].

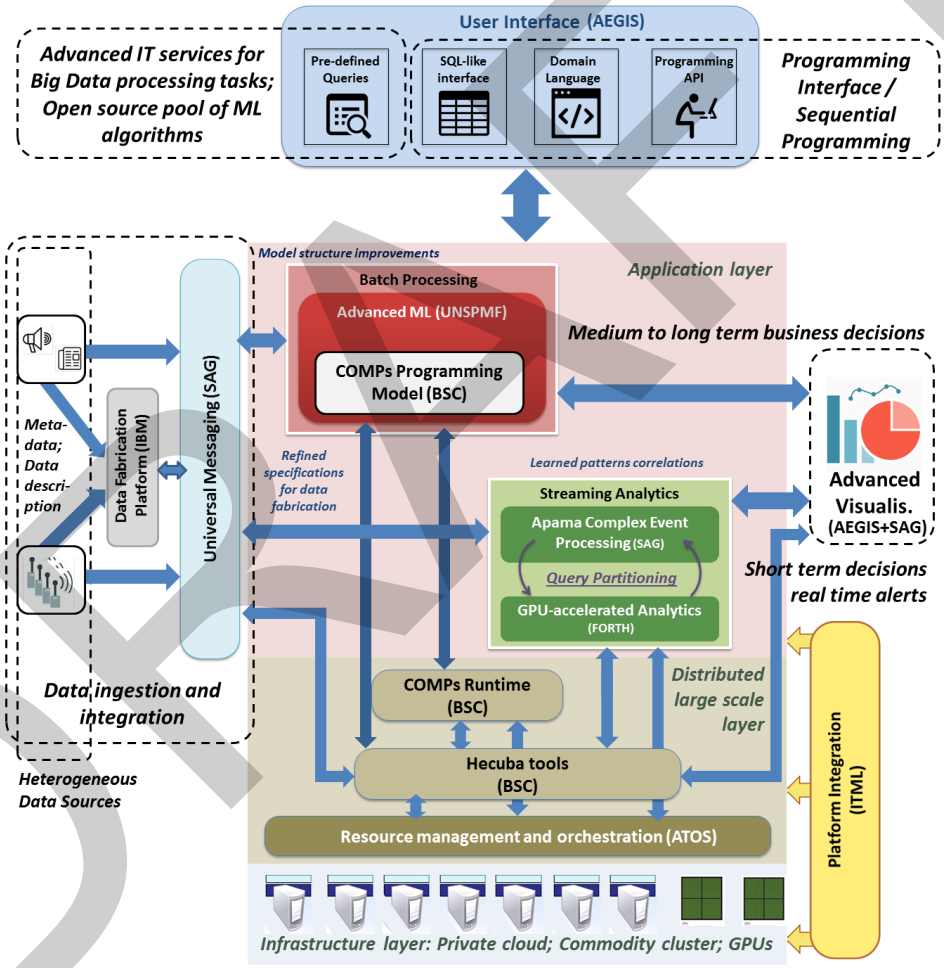

Figure 4.1: The *I-BiDaaS* platform

We proceed by briefly describing each of the three layers. Beyond the conceptual level discussed here, details on the inbound/outbound inter-

Table 4.1: The *I-BiDaaS* platform modules

Platform module and partner owner	**Platform layer**	**Brief description of functionalities/role in the platform**	**Current status**
ATOS cloud	Infrastructure layer	Computational and storage resources	Available
FORTH commodity cluster	Infrastructure layer	Computational and storage resources with a GPU cluster; enables functionalities of privacy preservation through commodity hardware (Intel SGX)	Available
Resource management and orchestration module (ATOS)	Distributed large scale layer	Performs management and orchestration of resources	Under development
COMPSs runtime module (BSC)	Distributed large scale layer	Data-driven task scheduling and data movement optimization	Available
Hecuba module (BSC)	Distributed large scale layer	Consists of Hecuba interface sub-module and Hecuba database sub-module. Hecuba interface sub-module provides interface between python (or Java, or similar) applications and the Hecuba database. Hecuba database refers to a key-value datastore (Cassandra, ScyllaDB, or similar) and Qbeast an indexation system that supports interactive analysis and arbitrary approximate queries.	Available; extensions planned during the project
Data fabrication platform DFP (IBM)	Application layer	Platform for generating user modelled realistic synthetic data for testing and development	Available; extensions planned during the project
Universal messaging UM (SAG)	Application layer	Passive publish/subscribe message oriented-middleware; primarily used to ingest, prepare, and integrate data from various sources and make it available for batch processing and streaming analytics	Available
Batch analytics module (UNSPMF+BSC)	Application layer	Contains two sub-modules, Advanced machine learning sub-module (UNSPMF), and COMPSs programming model sub-module (BSC). The latter is a sequential programming model able to exploit the inherent parallelism of the applications developed with it; used for implementation of the advanced machine learning sub-module. The former is a python implementation (using COMPSs programming model) of a pool of machine learning algorithms for batch analytics	COMPSs programming model sub-module available; Advanced machine learning sub-module under development.
Streaming analytics module (SAG+FORTH)	Application layer	Consists of two sub-modules, Apama complex event processing sub-module (SAG), and GPU-accelerated streaming analytics and pattern matching sub-module (FORTH). The former is a complex event processing infrastructure for streaming analytics. The latter is a system for optimizing the pattern matching procedures (i.e., string searching and regular expression matching) taking advantage of the benefits that GPU accelerators (FORTH cluster) offer.	Apama sub-module available; GPU-accelerated streaming analytics and pattern matching sub-module under development.
Advanced visualization module (SAG+AEGIS)	Application layer	Consists of the MashZone sub-module (SAG) and Advanced Visualizations Toolkit AVT (AEGIS).	MashZone available; AVT under development.

faces, integration of different modules, and physical aspects of the platform are presented in [49].

Infrastructure layer: The infrastructure, "lower-most vertical" layer includes the actual underlying storage and processing infrastructure of the *I-BiDaaS* solution, nominally provided and managed by ATOS and FORTH. This includes: 1) ATOS cloud, 2) FORTH GPU (Graphics Processing Unit) cluster (see Figure 4.1). We note that the *I-BiDaaS* solution will be deployable to other infrastructure premises as well; for instance, *I-BiDaaS* use cases 4-6 will be deployed within the CAIXAs proprietary private cloud.

The **Distributed large-scale layer** is responsible for the orchestration and management of the underlying physical computational and storage infrastructure. It allows the effective and efficient use of the cloud infrastructure and enables the application layer to provide effective Big Data analytics. The distributed large-scale layer manages several tasks, including the following: 1) task and data dependency capturing; 2) data transfer optimization: 3) task and data scheduling; 4) resource provisioning and management; and 5) management of data storage and databases of the platform. It consists of the Resource management and orchestration module (ATOS), COMPS runtime (BSC), and Hecuba tools (BSC).

The **Application layer** sits on top of the distributed large-scale layer and involves multiple modules as shown in Table 4.1. It refers to the architecture aspects and modules that are involved in the actual workflow of extracting actionable knowledge from the Big Data, starting from data ingestion, preparation, and fabrication, to batch and streaming analytics, to visualization and delivering analytics results for supporting decision making. Figure 4.1 depicts the knowledge extraction workflow that includes several highly innovative aspects. The workflow is as follows; The data from heterogeneous sources are ingested in the solution. For early development scenarios when not sufficient real data is available, or when regulative restrictions prohibit usage of real data, we use the IBMs data fabrication platform (DFP) to generate realistic synthetic data. Two different modes of operation are envisioned, depending on the privacy constraints of the Big Data pipeline in question: 1) the DFP is installed at end users local premises, realistic synthetic data is generated, and subsequently uploaded to the *I-BiDaaS* nominal infrastructure (ATOS cloud and FORTH cluster) for experimentation and testing; or 2) the necessary components of the *I-BiDaaS* platform are installed locally in end users premises. The data gets ingested into the batch processing module and the streaming analytics module via the Universal Messaging module UM (SAG) [96].If data transformation and pre-processing is needed, the integration server by SAG can be used. UM is a message oriented middleware that acts as intermediary to facilitate communication between multiple platform modules.

The streaming analytics module, consisting of Apama complex event processing sub-module (SAG) [15] and GPU-accelerated streaming analyt-

ics and pattern matching sub-module (FORTH), perform analytics on the ingested streaming data, also referencing historic information where necessary, to identify business patterns that have happened or are about to happen. The batch and real time analytics results are fed to the advanced visualization module (AEGIS AVT and SAG MashZone sub-modules). The collected data can be stored in the Hecuba module (Hecuba DB sub-module). The data from the Hecuba DB are picked up by the Advanced Machine Learning module (UNSPMF) for batch processing. This module will contain a pool of *I-BiDaaS* machine learning algorithms, implemented in COMPSs (COMP Superscalar) programming model [19] with necessary external additions, which will be continuously enriched as the platform lives. The batch analytics results are forwarded to the Advanced Visualization Module to support decision making.

4.2.1 Innovations within the *I-BiDaaS* Platform

The *I-BiDaaS* platform incorporates several innovations. First, the correlations produced by the batch analysis are fed back to DFP, to be used for training and to help building rules that will be used for future data generation purposes. Furthermore, regarding stream processing, we allow to offload parts of the streaming analytics that can be parallelized to the GPU-accelerated streaming analytics and pattern matching sub-module (FORTH); this gives us the opportunity to partition the analytics queries, between the high-level stream processing engine (Apama) and the low-level, hardware optimized implementation (FORTH GPUs). By carefully performing part of the queries at the lowest level (especially for filtering), only the required data will be forwarded to the stream-processing engine for a more sophisticated analysis, while the remainder will be ignored at the earliest possible. The partition of the queries can be done either statically (i.e., during the implementation of a specific user query) or dynamically at run-time (i.e., by monitoring the execution of a user-defined query, and deciding if the offloading to a many core processor would lead to better performance). Finally, the system is envisioned to introduce feedback from the analytics results to problem modelling; in other words, models (e.g., for structured (non)convex optimization) will be upgraded for each application class (e.g., through introducing non-zero weights to additional regularization functions) based on historical analytics performance scores.

4.3 *I-BiDaaS* Functionality

The following scenario corresponds to the *I-BiDaaS* typical workflow. First, users define the analytics they want to perform on the requested data sources. To do so, a graphical interface will be provided to define a Big Data pipeline.

The requested analytics can use the batch processing mode, the streaming processing mode, or both; the desired mode is specified by the user. Additionally, the user can generate data for training, testing, and experimentation in early development stages, through the DFP. It can also perform data sampling through the Hecuba DB sub-module (Qbeast).

To this end, the platform will offer the following functionalities (F):

- **(F1) Data ingestion and integration.** The non-IT user specifies the location of the proprietary data (e.g., large volumes of diverse corporate data) and optionally specifies any external open data sources (e.g., data related to the industrial and non-industrial environment) that can be used to enrich the proprietary data with contextual information; this is performed through a simple, web-based, interface. Subsequently, the platform integrates and ingests the data in the platform. In other words, the platform allows for a seamless integration of data as well as incorporation of the real time, event-triggered, streaming data (including for example sensor data) into the platform, ready to be analysed. This functionality will primarily be achieved through the UM module (SAG).

- **(F2) Fabricating realistic synthetic data for testing.** The user can test its novel algorithms (e.g., machine learning methods) or novel software products by fabricating data for testing and experimentation, when real data may not be available. This is important, e.g., in early development stages when real data may not be available, or when regulatory and other constraints disable uploading real proprietary data to external resources. Experimentation and testing over realistic synthetic data within cloud resources (here nominally, ATOS cloud) can significantly shorten the time for testing and experimentation of novel Big Data tools, algorithms, products, etc. This functionality will be mainly achieved by DFP (IBM).

- **(F3) Advanced IT services for Big Data processing tasks.** The non-IT users can select the task they would like to perform over the data through a simple, web-based interface. There will be a pre-defined pool of machine learning algorithms/built-in tasks, as well as a list of modes of operation such as: a) batchturn on the analytics and wait for the result, the result is not considered urgent, b) interactivethe user periodically gets further questions from the platform, c) continuous real-time trackingturn on the application and receive alerts upon detecting critical events, etc.).

- **(F4) Sequential programming.** The IT user (developer) implements a novel task (e.g., a novel machine learning algorithm), possibly to be added to the platforms pool of tasks, in a standard programming

language (e.g., Python) through sequential programming; at the runtime, the platform automatically 'parallelizes' the task over the actual distributed infrastructure. Removing the burden of worrying about resources from the IT user is mainly achieved through the COMPSs runtime module (BSC) and the Resource management and orchestration module (ATOS). This can significantly shorten the time for the implementation of new customized algorithms.

- **(F5) Open source repository for Big Data processing tasks.** When developing a novel application task, the IT-user can reuse the code of the existing task within the platforms open source repository (this will correspond to the Github page of the *I-BiDaaS* knowledge repository [1]).

- **(F6) Data sampling and interactive querying.** Users will be able to sample or interactively query data to facilitate experimentation, data exploration, and arbitrarily increase or decrease the level of details to test both the algorithms' correctness and scalability. This will be achieved through the Qbeast indexing system (BSC), a Hecuba DB sub-module that provides data-thinning and multidimensional indexing capabilities.

- **(F7) Advanced visualizations and monitoring.** The users observe results through advanced visualization, dashboards, aggregates, etc., to support the effective decision making (through AVT by AEGIS and MashZone by SAG).

In conclusion, we can see that system functional requirements (FR 1-5) detailed in 3 and state of the art gaps (SoAG 1-3) detailed in 2 are well addressed by the *I-BiDaaS* platform. Specifically, F1 is the main enabler for FR1; FR2 and FR3 are mainly addressed by F2; FR4 by F3-5; FR5 by F6; and FR6 by F7. Clearly, SoAG1 is addressed by all functionalities F1-F7, i.e., by the platform as a whole; SoAG2 by F4; SoAG3 by F2.

This mapping between system requirements and *I-BiDaaS* functions and associated components, is illustrated in Table 4.2. Note that the ATOS cloud module and the FORTH commodity cluster module constitute the infrastructure layer and hence implicitly contribute to all functionalities F1-F7.

4.4 *I-BiDaaS* Architecture Specification

This section provides an initial technical specification and model of the overall *I-BiDaaS* platform.

Table 4.2: Mapping of system requirements to *I-BiDaaS* components and associated functions

System Requirement	***I-BiDaaS* Component**	***I-BiDaaS* Functionality**
FR1	DFP (IBM)	F1: Fabricating realistic synthetic data for testing
FR2, FR3	UM Module (SAG)	F2: Data ingestion and integration
FR4	Batch analytics module (UNSPMF+BSC); Streaming analytics module (SAG+FORTH)	F3: Advanced IT services for Big Data processing tasks
FR4	COMPS runtime module (BSC), COMPS programming model sub-module (BSC), Resource management and orchestration module (ATOS)	F4: Sequential programming
FR4	*I-BiDaaS* knowledge repository; Advanced ML sub-module (UNSPMF)	F5: Open source repository for Big Data processing tasks
FR5	Hecuba DB sub-module, Qbeast indexing system (BSC)	F6: Data sampling and interactive querying
FR6	Advanced visualization module (SAG+AEGIS)	F7: Advanced visualisation and monitoring

In order to deal with the system complexity we use a *viewpoint-oriented* approach, whereby the architectural description is partitioned into a number of separate, though complementary, representations, each reflecting specific concerns held by one or more of its stakeholders, using certain conventions such as notations, modelling methods, analysis techniques, etc. To this end, several architecture frameworks have been proposed, which are essentially viewpoint classification schemes. The classification by Kruchten [58], proposed in 1995, is probably the best known classification and is still widely used. The viewpoint-oriented paradigm has also been adopted by ISO/IEC 42010:2007, later revised by the ISO/IEC 42010:2011 Standard "Systems and software engineering–architecture description" [54].

The proposed *I-BiDaaS* architecture specification comprises of the following viewpoints, each depicted using appropriate UML diagrams [5] (see Figures 4.2, 4.3, 4.4 and 4.5):

- **Logical viewpoint.** Describes the high level structure of the architecture in terms of the functional components (modules) that collaboratively deliver the required functionality. The emphasis is on the relationships between the components realized through external interfaces and not on the internal structure of components. This viewpoint concerns all stakeholders that are interested in the platform. Figures 4.2 and 4.3 present the *I-BiDaaS* platform through the informational

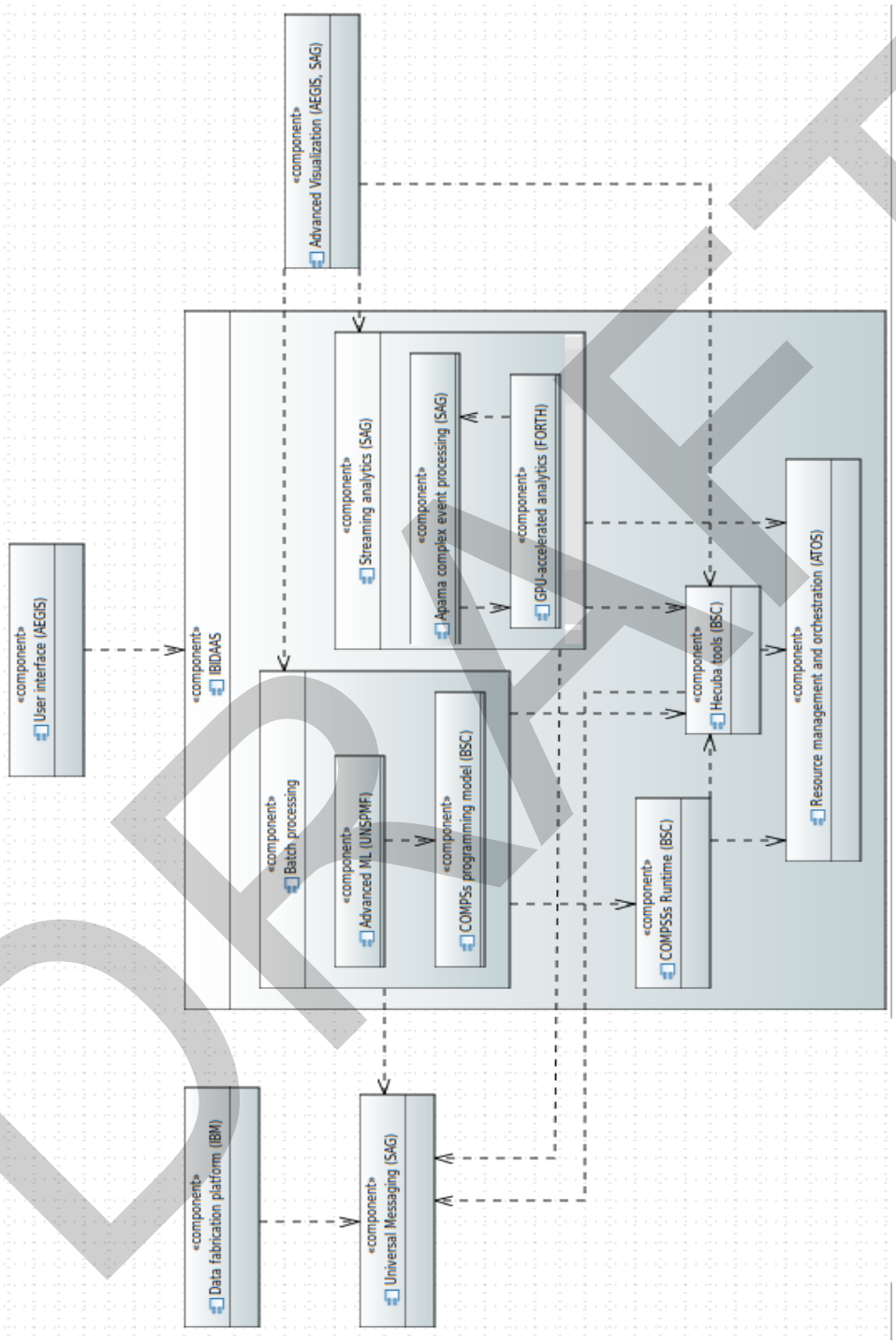

Figure 4.2: UML component diagram for the *I-BiDaaS* integral platform

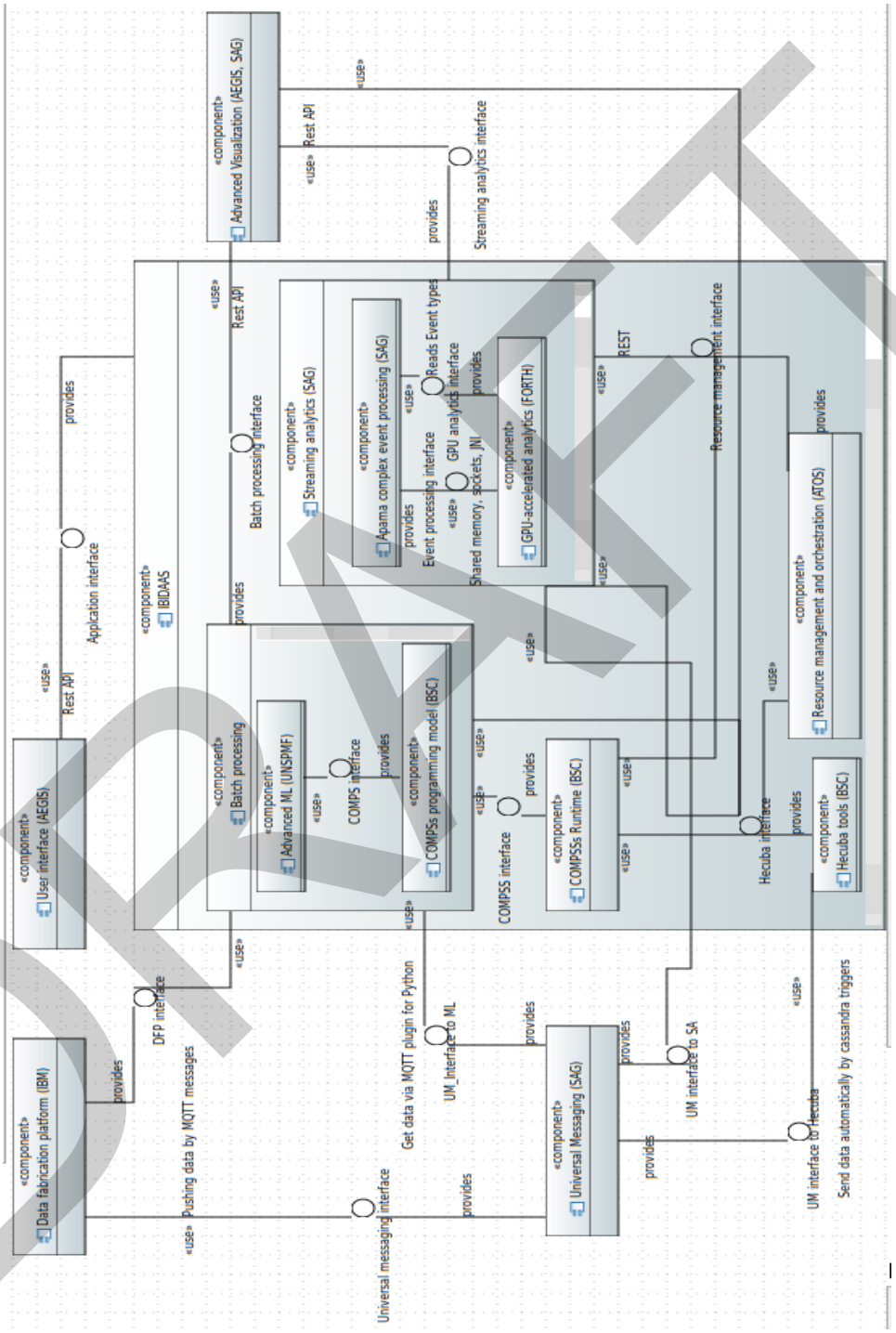

Figure 4.3: UML component diagram with interfaces for the *I-BiDaaS* integral platform

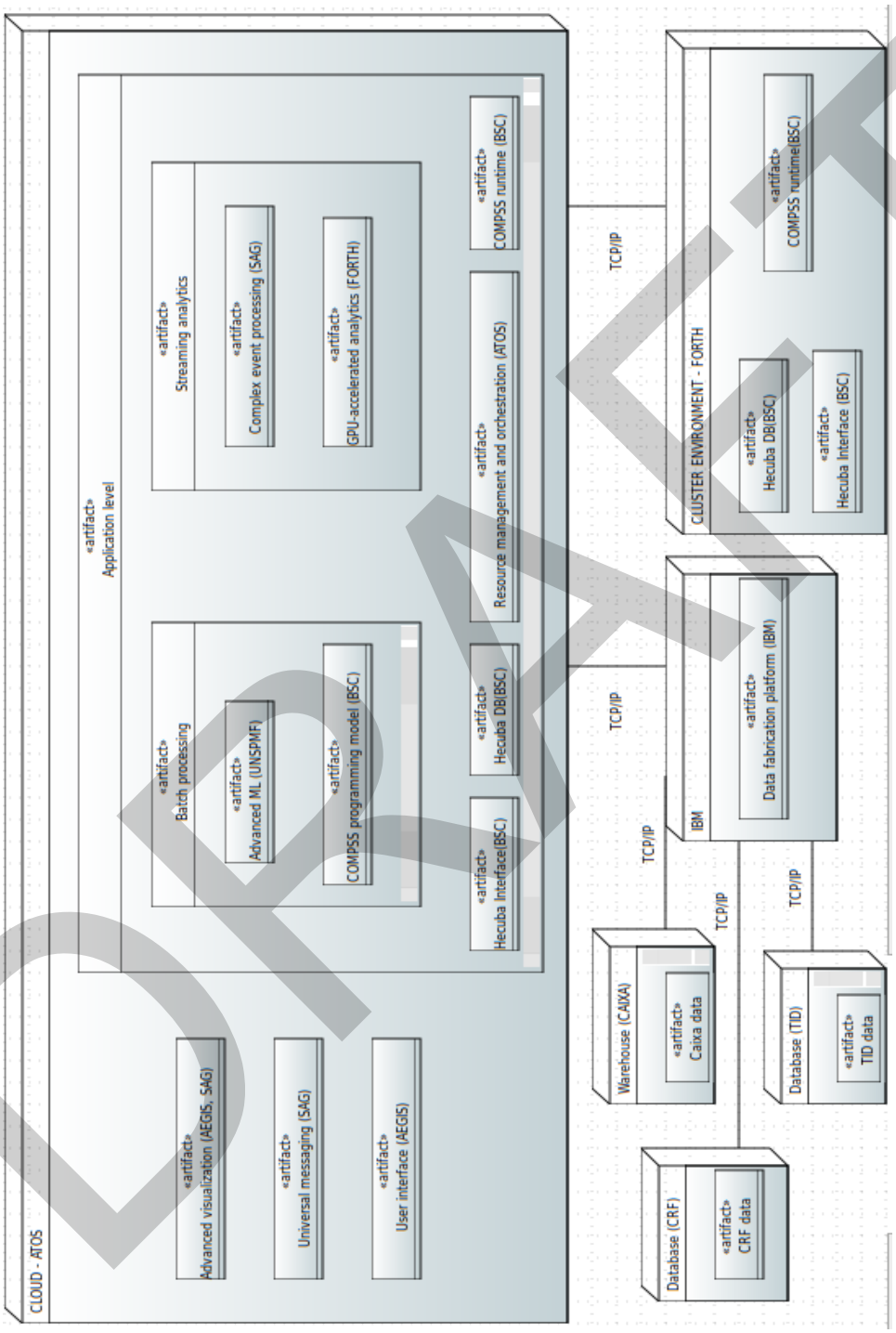

Figure 4.4: UML deployment diagram for the *I-BiDaaS* integral platform

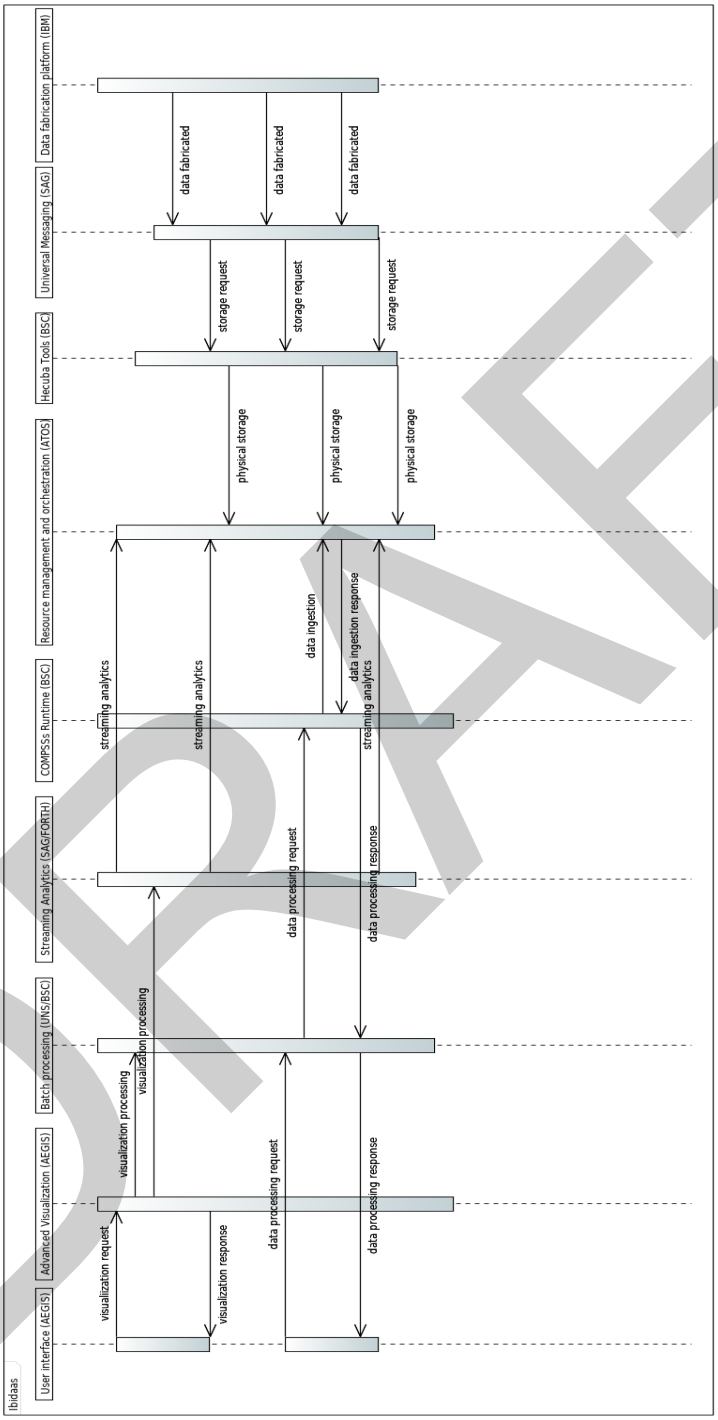

Figure 4.5: UML sequence diagram for the *I-BiDaaS* integral platform

viewpoint via a UML component diagram and a UML component diagram with interfaces, respectively.

- **Informational viewpoint.** Describes runtime information flow between system components and places emphasis on the circulation of information at a high level of abstraction. The stakeholders who are interested in this viewpoint are primarily users (both IT and non-IT), acquirers, developers, and maintainers, but most of them have some level of interest. Figure 4.5 depicts the *I-BiDaaS* platform from the informational point of view via a UML sequence diagram.

- **Physical viewpoint.** Details the allocation of the system components to different physical containers, and describes the hardware required, network requirements, and physical constraints. System administrators, developers, and testers are the stakeholders primarily interested in this viewpoint. Figure 4.4 depicts the *I-BiDaaS* platform from the physical point of view via a UML deployment diagram.

With reference to Figures 4.2, 4.3, 4.4 and 4.5, we now make a more precise description of the architecture with respect to the conceptual one in Section 4.2, including the information on data flow and how the different modules communicate.

Specifically, the data flow is as follows:[1] First, DFP generates realistic synthetic data based on the rules it receives from a data provider. Then, it ”pushes” the fabricated data to UM via a Message Queuing Telemetry Transport (MQTT) messaging protocol [69]. Next, Hecuba DB gets data from UM. This can be achieved via a MQTT plugin for Python (see [33]). At the same time, Apama gets data from UM in a streaming fashion via the existing standard interface between Apama and UM. Then, the streaming analytics results from Apama output are sent to another UM channel and are then written into Hecuba DB. Apama can forward data to the GPU streaming analytics module by FORTH through an inter-process communication mechanism like shared memory, sockets or the Java Native Interface (JNI). From Apama, data goes to the advanced visualization module. This is achieved via the standard connection between Apama and MashZone [64], the communication is done only to notify changes in the dataset results (data will be read through Hecuba Interface). If network traffic is an issue, multiple instances of UM will be used; for example, in case of the *I-BiDaaS* use cases (Table **??**), one UM will be used per one data provider (CAIXA, TID, CRF). The feedback from the batch processing module to DFP can be achieved through UM, similarly to the feedback from the Streaming analytics module to the batch processing module.

[1] We focus only on a use case where a realistic synthetic dataset is considered, while the operation is similar for other data sources as well.

From Figure 4.4, it can be seen that most of the modules will be physically placed where the infrastructure is (ATOS cloud and FORTH cluster). The main database system will be Hecuba DB, and it will be located at both ATOS and FORTH. The data providers datasets are stored at their local premises; they send to the DFP the data fabrication rules via a secure channel, which subsequently fabricates the realistic synthetic data and sends it to the ATOS and FORTH infrastructures. Alternatively, fabrication via DFP can be executed directly at either ATOS cloud or FORTH cluster, where DFP can be installed, while the fabrication rules are sent to DFP via a secure channel.

4.5 Discussion and Summary

This section identified the main *I-BiDaaS* system modules and outlined modules functionality. These *I-BiDaaS* components closely match the functional components of the Big Data Interoperability Framework of the National Institute of Standards and Technology (NIST) [72], shown in Table 4.3. In fact, the *I-BiDaaS* platform includes an additional activity, namely fabrication of realistic synthetic data, which facilitates experimentation and testing for privacy issues sensitive sectors like, e.g., banking.

Furthermore, the *I-BiDaaS* architecture addresses most of the horizontal concerns identified in the European Big Data Value Strategic Research and Innovation Agenda (BDVA SRIA), in the context of the BDV reference model [18], as detailed in Table 4.4. Thus, it provides and end-to-end Big Data solution.

Focusing on the BDVA SRIA expected high priority advances in data processing architectures (item 3 in Table 4.4), *I-BiDaaS* addresses several of them, namely: Techniques and tools for processing real-time heterogeneous data sources; and Real-time architectures for data in motion. The two aspects are addressed in I-BiDaaS by the integration and intelligent query and task partitioning between Apama Complex Event Processing (CEP) and FORTH GPU-accelerated streaming analytics for stream processing. In addition, distributed architectures advances highlighted in BDVA SRIA relate to the *I-BiDaaS* federation of cloud and GPU cluster resources.

Finally, it should be noted that, vertical concerns expressed in The BDV Reference Model (i.e., cross-cutting issues, which may affect all the horizontal concerns, such as Big Data Types and Semantics, Standards, Cybersecurity, etc.) are also addressed by other *I-BiDaaS* project aspects (e.g., standardization efforts).

Subsequent work during the course of the project will provide guidelines, recommendations, and descriptions on how the *I-BiDaaS* proprietary modules can be replaced with open source alternatives.

Table 4.3: Mapping between NIST reference architecture activities and *I-BiDaaS* modules and effort (project tasks) that help achieving these activities within the *I-BiDaaS* platform

No	NIST reference architecture activity	***I-BiDaaS* module or effort**
System Orchestrator		
1	System architecture (defining the requirements to be met by the system and establishing technical guidelines to fulfill the requirements)	T1.2; T1.3; T1.4 (Requirements elicitation; architecture definition)
2	Data science (define requirements for individual algorithms)	T1.4 (Experimental protocol)
3	Security/privacy (control of access, ensure privacy and security of personal or corporate information)	FORTH commodity cluster module privacy preservation through commodity hardware; DFP (IBM); authentication and authorization within I-BiDaaS user interface (AEGIS)
Big Data application provider		
4	Data collection; data preparation	UM (SAG)
5	Analytics (extract knowledge from data based on the requirements of the vertical application)	Batch analytics module (UNSPMF+BSC); Streaming analytics module (SAG+FORTH)
6	Visualization	Advanced visualization module (AEGIS+SAG)
7	Access (communication/interaction with data consumer)	User interface (AEGIS)
Big Data framework provider		
8	Infrastructure activities	UM (SAG) MQTT messaging; FORTH commodity cluster GPU implementation for high parallelization and low latency tasks
9	Platform activities (organization and distribution of data; data movement, etc.)	COMPSs runtime (BSC); ATOS resource management and orchestration module
10	Processing activities	Batch analytics module (UNSPMF+BSC); Streaming analytics module (FORTH+SAG)
Management fabric activities		
11	System management (configuration, resource management and provisioning, etc.)	ATOS resource management and orchestration module
12	Big Data lifecycle management (data recovery, preservation, accessibility, etc.)	ATOS resource management and orchestration module
Security and privacy fabric activities		
13	Authentication, authorization, auditing	All *I-BiDaaS* layers; User interface (AEGIS)

Table 4.4: Mapping between BDV reference models horizontal concerns and *I-BiDaaS* modules and effort (project tasks)

No	**BDV reference model horizontal concern**	***I-BiDaaS* module or platform as a whole**
1	Data visualization and user interaction	Advanced visualization module (AEGIS+SAG); User interface (AEGIS)
2	Data analytics	Batch processing module (UNSPMF+BSC); Streaming analytics module (SAG+FORTH)
3	Data processing architectures expected advances according to BDVA SRIA	*I-BiDaaS* platform
4	Data protection	FORTH commodity cluster privacy preservation through commodity hardware (Intel SGX); DFP (IBM) for generation of realistic synthetic data when real data cannot be uploaded to cloud or similar systems
5	Data management	COMPSs runtime (BSC); ATOS resource management and orchestration module
6	The Cloud and HPC	(efficient usage of Cloud) ATOS resource management and orchestration module

5 Experimental Protocol Specification

This section describes the approach that will be used towards the experimental evaluation and validation of the *I-BiDaaS* platform. It analyzes the technological and business aspects, which must be taken into account for the development of a comprehensive experimental protocol. Furthermore, it investigates existing Big Data benchmarks which could be used to evaluate the new concepts and algorithms proposed in *I-BiDaaS* and for assessing the performance of the *I-BiDaaS* platform.

In particular, the use case requirements and architecture specification described in sections 3 and 4 respectively, guide the definition of the experimentation and evaluation process, consisting of two phases. The first focuses on the evaluation of the *I-BiDaaS* architecture and its individual components, while the second phase focuses on the evaluation of the application of the *I-BiDaaS* platform in the different use cases in the business context and against the identified high-level business requirements.

5.1 Overview

Traditionally, evaluation protocols for big data applications rely on Big Data benchmarking [55]. In *I-BiDaaS* we take a wider perspective whereby the experimental protocol takes also into consideration additional aspects, namely:

1. the requirements of the *I-BiDaaS* use cases, described previously in section 3,

2. the properties of the innovative *I-BiDaaS* architecture, described in 4, that builds upon related work, existing reference frameworks as well as past experiences of the project partners with Big Data processing solutions and platforms,

3. the existence of advanced Big Data Benchmarking protocols, used in industrial and academic communities that are aligned with the *I-BiDaaS* objectives and make it possible to quantitatively and qualitatively evaluate this development with existing systems and academic works, and

4. best-practice software testing guidelines, which take into account specific Non-Functional, Quality of Service and Quality of Experience requirements (e.g. for example, scalability, availability, reliability, response time, data throughput, resources utilisation, etc.).

We claim that considering all the above issues is necessary in order to achieve end users' and stakeholders' acceptance of the *I-BiDaaS* architecture, and allow for its long term sustainability.

Figure 5.1 illustrates the proposed approach as a two phase experimentation process whereby, we move from the evaluation of the characteristics of *I-BiDaaS* the overall architecture based on quantitative and qualitative variables and benchmarks towards the evaluation of its particular deployments in specific industrial settings. To this end, the second step of the experimental protocol aims to highlight the *I-BiDaaS* innovation in actual industrial settings.

To identify the evaluation variables used in both phases that capture the concerns of both industrial partners and technology providers we have used a mixed-mode survey which combined structured interviews and questionnaires (see Appendix C).

Figure 5.1: Two phases of the *I-BiDaaS* experimental protocol

In the following sections, we first discuss related work on Big Data Benchmarking protocols and then we describe the proposed experimental process.

5.2 Background

As discussed previously, the *I-BiDaaS* use cases present a combination of big data characteristics, in terms of volume, velocity, variety, variability, volatility and sparsity, that require diversified, analytic processing capabilities. At the same time, the complexity of the *I-BiDaaS* platform integrating a number of different modules that realize different system capabilities, makes the selection of an objective Big Data benchmark, that covers all relevant characteristics, a complex task. Therefore, *I-BiDaaS* necessitates the definition of multiple benchmarks which should be used jointly in order to address all possible use cases.

Today, several initiatives and international consortia exist that concentrate on the definition of benchmarks for Big Data and Cloud processing, notably the Transaction Processing Performance Council (TPC) [94] and the Standard Performance Evaluation Corporation (SPEC) [86]. Various industry leaders have their own benchmarks, usually complemented with White Papers explaining their vision on Big Data analytics platforms. Examples of such benchmarks include:

- VMMark [97], which is designed to measure the performance and scalability of virtualization frameworks.
- Terasort [91] as a popular benchmark that measures the amount of time to sort one terabyte of randomly distributed data on a given computer system. The Terasort benchmark is commonly used to measure MapReduce performance of an Apache Hadoop [7] cluster.
- Graph500 [42] is a set of benchmarks for graph related problems.

Additional benchmarks have been designed by researchers aiming at analysing various properties of new and innovative, Linked Data, IoT, and sensor-data based smart applications. The following list describes some advanced benchmarks emerging from academic communities, which may prove useful in the evaluation of the *I-BiDaaS* architecture.

- *Berkeley Big Data Benchmark* [20] has been used to compare quantitatively and qualitatively five existing Big Data systems: (1) Redshift - a hosted MPP database offered by Amazon.com based on the ParAccel data warehouse. The authors of the benchmark have tested Redshift on the Dense Storage (HDD) instance family; (2) Hive - a Hadoop-based data warehousing system; (3) Shark - a Hive-compatible SQL engine, which runs on top of the Spark computing framework; (4) Impala - a Hive-compatible SQL engine with its own Massively Parallel Processing (MPP)-like execution engine; and (5) Stinger/Tez, which is a next generation Hadoop execution engine.

- *HiBench* [45] is a brand new benchmark suite that is used for Hadoop. It consists of a set of Hadoop programs, including both synthetic micro-benchmarks and real-world Hadoop applications. The HiBench has been recently used to characterize the Hadoop framework in terms of speed (e.g., job running time), throughput (e.g., the number of tasks completed per minute), HDFS bandwidth, system resource (e.g., CPU, memory and I/O) utilizations, and data access patterns.

- *The BigDataBench* [22] is a a Big Data and Artificial Intelligence (AI) Benchmark Suite developed by the Chinese Academy of Sciences. This advanced benchmark considers Big Data and AI workload as a pipeline of one or more classes of units of computation performed on different initial or intermediate data inputs. It is a rapidly evolving benchmark which has been successfully applied to evaluate multiple systems and architectures. In its current version BigDataBench 4.0 provides 13 representative real-world data sets and 47 benchmarks.

- *The Hobbit platform* [46] is a distributed FAIR [99] benchmarking platform for the Linked Data lifecycle developed in the context of the European research and innovation project Hobbit. It is open source and it can be downloaded and executed locally. It can also be accessed through its online instance [1] which can be used for a) running public challenges and b) making sure that even people without the required infrastructure are able to run the benchmarks they are interested in. So far Hobbit has delivered the initial versions of the following benchmarks:

  - *Generation Acquisition* benchmarks measure the performance of SPARQL query processing systems when faced with streams of data from industrial machinery in terms of efficiency and completeness.
  - *Analysis Processing* benchmarks focus on testing the performance of link discovery systems and machine learning methods (supervised and unsupervised) for data analytics.
  - *Storage Curation* benchmarks aim at testing the performance of data storage and versioning systems for Linked Data.
  - *Visualization Services* benchmarks aim at testing the performance of query answering and faceted browsing systems for Linked Data without involving users.

From the above discussion it can be seen that Big Data benchmarks, can be either technology-specific solutions and technology-agnostic solutions.

[1] accessible at master.project-hobbit.eu

The former address a set of representative applications in the targeted domains and generally mandate the use of a specific technology to implement the solution. The goal is to test the efficiency of a selected technology in the context of a realistic operational scenario. The latter aim at creating a level playing field for any number of technologies to compete in providing the most efficient implementation of a realistic application scenario within the targeted application domain. No assumption is made about which technology choice will best satisfy the real world demands at a solution level. This allows the analysis of the results collected under experimentation and compare them with other existing (baseline) systems. Therefore, apart from the applicability to a particular solution under test, in terms of its functionality, the choice of a particular benchmark depends on the benchmarking goals. An additional concern regarding the technical execution of the experimental protocol, is the cost of preparing and running particular benchmarks. Various studies have shown that setting up benchmarks requires large installations, and significant amount of time for preparation.

Currently, there is extensive information on how and why to perform technical benchmarks for the specific Big Data management and analytics processes [55], however there is a lack of objective, evidence-based methods to measure the correlation between between Big Data benchmarks and and organisations business benchmarks [89]. In order to assist European organizations developing Big Data technologies to determine which benchmarking solution best responds to their needs, the European project *DataBench* [26] currently underway, aims to investigate existing Big Data benchmarking tools and projects and provide a robust set of metrics to compare those tools. Rather than creating another benchmark, DataBench aims at supporting the efficient usage, evolution and synergy of available Big Data benchmarks. To this end, the project will build the *DataBench Toolbox*, a tool which will connect and evaluate external benchmarking initiatives.

5.3 The Experimentation Process

Figure 5.2 illustrates the proposed process that aims at evaluating and validating (1) the *I-BiDaaS* architecture, and (2) its implementation in the specific project use cases. The basic steps of the experimentation process and the activities involved in each step are described bellow.

Scoping is the first step, where we scope the experiment in terms of business KPIs for evaluating the achievement of the use case objectives. Such KPIs have been defined in the user requirements elicitation phase (see requirements questionnaires).

An example of KPIs in CAIXA's Enhance control over third party agencies use case include: Data charging time and Time to get analytics results. Furthermore, known industrial sector KPIs have also been considered. In

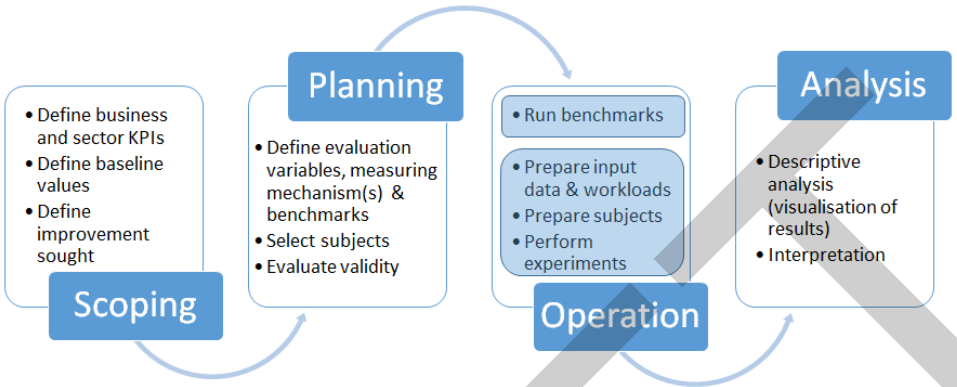

Figure 5.2: Experimentation process

addition, for each KPI the baseline (current) value and improvement sought is also defined at this stage. The scoping step is designed, so that it is able to calculate the KPIs based on the measurements performed in each experiment.

Planning is the next phase of the experimentation process, where the design of the experiment is determined, in terms of the instrumentation (variables to be measured and associated measuring mechanism) and evaluation subjects that will perform the measurement. It should be noted that evaluation variables may be relevant to one, more or all of the system functionalities.

Furthermore, in this step, any factors affecting the outcome of the experiment are to be acknowledged at this phase and appropriate measures should be identified in order to ensure validity of measurements.

Operation of the experiment follows from the design. In the operational step, measurements are recorded which then are analysed and interpreted (with respect to the improvements sought in the scoping step) in the final *Analysis* step.

In particular, operation considers two phases. The first phase aims at quantitative validation of the *I-BiDaaS* architecture in parts and as a whole by using specific benchmarks. It proceeds as follows. For each of the identified use cases sets of specific workload types are selected. The workload types, such as graph analytics, artificial intelligence, data warehouse, NoSQL, streaming etc., are well documented and freely available as part of the existing benchmarks (see section 5.6). Then, the experimentation continues by a gradual increase of the complexity through: (a) definition and experimentation with Micro Benchmarks; (b) combinations of workloads which are more representative of the specific use cases; and (c) complex end-to-end benchmarking, which involves high-level and business aspects of the specific use cases. The second phase mainly aims at qualitative eval-

uation of the innovative *I-BiDaaS* architecture in the business context and against the identified business requirements.

5.4 Collecting Relevant Information from Users and Technology Providers

In agreement with the industrial and business partners of the project, we decided to prepare a detailed questionnaire, aiming to document all use case experimentation needs and plans. In order to prepare appropriate questions at our ongoing project meetings we collected feedback, using structured interviews, on the following topics.

1. *Evaluation in relation to the expected project level innovation and achievements*

   We intended to define important project level evaluation variables, including the definitions of usability, innovation, robustness and performance. Moreover, we intended to map various qualitative aspects, such as those just mentioned to specific components of the *I-BiDaaS* architecture. Additional discussions included the steps that must be taken in order to verify that a certain variable (utility, innovation, robustness, performance) has been achieved. Benchmarking needs have also been discussed as well as KPIs for evaluating the achievement of the project objectives.

2. *Evaluation from the use cases viewpoint*

   Here we concentrated on questions relevant to how the partners intend to evaluate each use case. The aim of these discussions was to specify a step by step scenario that can be performed for the verification of each *I-BiDaaS* use case. The discussions also concentrated on the expected benefits from applying the *I-BiDaaS* solution in each use case. Generally, since *I-BiDaaS* is an integrated architecture, it makes sense that the individual use cases rely as much as possible on the architecture as a whole and not on subsets of architectural components.

3. *Evaluation from the viewpoint of subjects participating in the experiments*

   Here we concentrated on methods to collect feedback from the subjects that will use the *I-BiDaaS* solution (i.e. the MVP) as a whole (or its components). We discussed the potential of involving subjects such as quality assurance and control managers, data analysts, financial administrators, human resource officers, infrastructure engineers, IT security personnel, decision executives in the production, resource

planning managers, call centre employees, network infrastructure engineers and market and business analysts. The partners were also asked to fill in the number of persons that can directly engaged in the experimentation process and help find additional persons from their working environment that can engage in the experimentation process.

Additional questions were related to the (essential) user interfaces that will be provided by the *I-BiDaaS* platform to its users. They aimed at investigating the way of using such interfaces, the number of users necessary to test functional and non-functional properties of the *I-BiDaaS* platform and similar.

In addition to this, we identified methods of engagement of the participants (i.e. the persons to be engaged in experimentation and testing). These include:

(a) workshops (physical or virtual) organized by the data providers (T1.2),

(b) INFO days organized by the consortium industrial partners (CAIXA, CRF, TID)

(c) starting from M12 when the MVP will be released, an open session that will be organized for the experimental subjects co-located with the project's General Assembly meeting, and

(d) by means already put in place by the consortium, namely TID (WAYRA), TID (AURA) and CRF (Campus-Melfi).

4. *Industrial benchmarking needs, validating the solution as a whole*

This type of discussions were necessary in order to gain better insight into the benchmarking needs, the functional properties that must be verified, the QoS, any QoS metrics that must be monitored by a monitoring system, and methods to assure comparability of the obtained results.

The complete list of interview questions is provided in Appendix C. The outcome of these interviews guided the construction of the questionnaire (also included in Appendix C that was used in order to prepare for both experimentation phases.

5.5 Experimentation Setup

In this section we summarise the above mentioned discussions and results of the survey, which represent our experimentation setup. This summary focuses mainly on the scoping and planning of the experiments, as the next steps are to be further elaborated in subsequent phase of the project.

5.5.1 Scoping - use case providers

This section was generated based on input from the organisations that provide use cases to the project. Sector specific KPIs that will be used to measure the achievement of the objectives of the use case have been specified. Comprehensive use cases are provided including KPIs. See Tables 5.1 and 5.2.

5.5.2 Planning - all partners

This section aims at the identification of the evaluation variables and the measuring mechanism. Due to the necessity for collaborative work, it is filled in by all project partners.

Using detailed tables, the partners were required to indicate evaluation parameters and associated variables to be measured in the context of the experiments, removing irrelevant ones / adding new ones where applicable. They were asked to be as much as possible specific, with respect to the measuring mechanism that has to be used, for example, to name specific standards, benchmarks and testing tools to be used in the process.

In this context, usability is the degree to which a software system can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use. In addition to that aspects such as performance, scalability, availability and replication will be tested.

Our approach towards testing non-functional aspects of the system is presented schematically in Table 5.3.

The partners also indicated a number of high-level non-functional aspects that should be tested in the experimentation process. These are presented in Table 5.4. Innovation refers to the implementation of a new or significantly changed product or process, and this is priority to all of the partners. Robustness is the ability of a computer system to cope with errors during execution and cope with erroneous input. Accountability is an assurance that an individual or an organization will be evaluated on their performance or behaviour related to something for which they are responsible.

The functional parts of the *I-BiDaaS* system will also be tested, as an integrated solution (MVP) as well as individual components. This is presented in the following Tables 5.5, 5.6 and 5.7.

5.5.3 Validity Evaluation - technology providers

In this section we document potential threats to the validity in terms of the measurements indicated in the tables in the previous steps of the process. We indicate actions that will be taken to assure validity.

Table 5.1: KPIs for the use cases - subject to refinement in the next project phase - part 1

No	Use Case Name	KPIs	Baseline value	Improvement sought
1	Building of a social graph (CAIXA)	Data charging time. Time to get analytics results. Count number of nodes by category. Count number of edges by category. Price of the solution technologies. Time to generate business rules. Time of granting permits for data access at the start of a project with an external provider.	To be determined. 3-4 weeks.	To be determined. 1-2 weeks. (50% of time reduction).
2	Enhance control over third party agencies (CAIXA)	Data charging time. Time to get analytics results. Time to validate previous results of real data analytics. Confusion matrix. Price of technologies. Time to generate business rules. Time of granting permits for data access at the start of a project with an external provider.	To be determined. 3-4 weeks.	To be determined. 1-2 weeks. (50% of time reduction).
3	Advanced Analysis of bank transfer payment in financial terminal (CAIXA)	Data charging time. Time to get analytics results. Confusion matrix. Price of technologies. Time to generate business rules. Time of granting permits for data access at the start of a project with an external provider.	To be determined. 3-4 weeks.	To be determined. 1-2 weeks. (50% of time reduction).
4	Analysis of relationships through IP address (CAIXA)	Comparison of real results vs results from synthetic data. Accuracy. Recall. True Positive Rate. True Negative Rate. Confusion matrix. Time of granting permits for data access at the start of a project with an external provider.	To be determined. 3-4 weeks.	To be determined. 1-2 weeks. (50% of time reduction).

Table 5.2: KPIs for the use cases, subject to refinement in the next project phase - part 2

No	Use Case Name	KPIs	Baseline value	Improvement sought
5	Employment of bots in call center (TID)	% of low customer satisfaction index (CSI) audio calls analyzed / per time unit	~10%/month	~20%/month (i.e., 100% increase in low CSI audio calls analyzed
		% of audio calls automatically transcribed	~0%	~100% (i.e., enable full transcription of all audio calls for further analysis)
6	Accurate location prediction with high traffic and visibility (TID)	Data processed / time unit	~3TB / 15 min	~3.5-4TB / 15 min (2-4% increase)
7	Optimization of placement of telecommunication equipment (TID)	Computation cost, storage costs	60 min to process the daily aggregated data (15TB of web transactions and TCP statistics + 3 TB of signal events from the Mobility Management Entity)	45 min to process the daily aggregated data (15TB of web transactions and TCP statistics + 3 TB of signal events from the Mobility Management Entity)
8	Predictive maintenance (CRF)	Overall equipment Effectiveness (OEE)	To be determined.	To be determined.
		Job per hour (jph)		
		Maintenance costs		
9	Production process of Aluminium	Scrap percentage	To be determined.	To be determined.
		Cost per unit		
		Quality control and repair costs		

In particular, the following validity threats have been identified:

- Scalability and Operational Performance variables can be affected by noise in the system (i.e. running other processes at the same time). A proper system isolation must be ensured to correctly measure the corresponding metrics.

- Advanced visualisations might introduce a level of complexity in the resulted user interface. This might have an effect on the users' perception of the usefulness and added value of the visualisations. Actions to be taken against this threat involve the presence of helping functions such as in-page guides, explanatory texts, usage hints on the various visualization elements and so on.

Table 5.3: Non-functional aspects to be tested during experimentation - subject to refinement in the next project phase

Evaluation parameter	**Evaluation variable**	**Measuring mechanism**	**Relevant Functionality**					**Stakeholders**	
			Data ingestion & preparation	**Data Fabrication**	**Batch Analytics**	**Streaming Analytics**	**Data Visualisation**	**End Users**	**IT specialist**
Usability	Effectiveness (accuracy and completeness)	Participant completion rate	X	X	X	X	X	Data analysts	Network Infrastructure engineers
		Number of errors	X	X	X	X	X		
		Software Improvement	X	X	X	X	X		
	Efficiency	Task time efficiency	X	X	X	X	X	Call center employees	Network Infrastructure engineers
		Perception of time required to accomplish a task	X	X	X	X	X	Decision executives in the production	
		Perception of task completion quality	X	X	X	X	X	Quality assurance and control managers.	
	Satisfaction	Degree to which user needs are satisfied – look and feel	X	X	X	X	X	Call center employees	Network Infrastructure engineers
			X	X	X	X	X	Decision executives in the production.	
			X	X	X	X	X	Quality assurance and control managers.	
Performance	Throughput	Standard definitions	X	X	X	X	X		Network Infrastructure engineers
	Latency		X	X	X	X	X		
Scalability	Scale-up		X	X	X	X	X		
	Scale-in		X	X	X	X	X		
	Scale-out		X	X	X	X	X		
	Elasticity		X	X	X	X	X		
Availability	Failure takeover		X	X	X	X	X		
	Error recovery		X	X	X	X	X		
Replication	Cost/Performance tradeoffs		X	X	X	X	X		
	Freshness / Consistency		X	X	X	X	X		

Table 5.4: High-level non-functional aspects to be evaluated during the experimentation process

Evaluation parameter	Evaluation variable	Measuring mechanism	To be recorded by	
			End user	IT specialist
Innovation	Additional functionality beyond the system currently in use	Measured against the industrial requirements	Data analysts	Infrastructure engineers
				IT security personnel
				Network engineers
	Beyond the standard technologies	Measured against the latest	Data analysts	Infrastructure engineers
		Strategic Research and Innovation Agenda (SRIA) released by BDVA		Network engineers
Robustness	Efficient data integration	Number of different heterogeneous datasets	Data analysts	Network infrastructure engineers
		Capability of integration and processing– error free		
		Measured against industrial-validated benchmarks	Quality assurance and control managers	Infrastructure engineers
Accountability	Performance against commercial or current solutions	Accountability metrics: A-PPL12	Financial administrators	Network engineers
			Market and business analysts	

5.5.4 Human Resources in the Evaluation Process

In the following Table 5.8 we present the initial number of human resources that will take part in the experimentation and evaluation process.

With respect to expected ways to access and engage the required number of experimental subjects, TID will make use of internal email lists and employee forums to recruit participants for these experimental tests. There are thousands of employees in such lists and it will be made possible to access and engage them in a meaningful way. Furthermore, the use-cases proposed by TID are in coordination with internal business units. Therefore, TID has already tentative access to employees that are related to each of the use-cases, and each of the required participant roles. Thus, it will be straightforward to access and engage them in the experiments.

CAIXA on the other hand, will engage several data analysts from third party agencies, in order to evaluate the different specified use-cases. IT security personnel, infrastructure engineers, quality assurance and financial administrators from CAIXA will also be involved in the evaluation of innovation, processes optimization and flexibility on accessing CAIXA's datasets.

Table 5.5: Testing quality at MVP system level

Evaluation parameter	Evaluation variable	Measuring mechanism	To be recorded by		
			End User	IT User	Technology Provider
IBiDaaS Integrated Solution - MVP					
IBiDaaS Integrated solution (MVP)	Scalability		X	X	X
	Operational Performance	Response time	X	X	X
		Data throughput	X	X	X
		Resources utilisation	X	X	X
	Availability		X	X	X
	Reliability	Data failure	X	X	X
	Data Security		X	X	X
	Privacy		X	X	X
	Compliance	Measured against relevant standards	X	X	X
	Cost		X	X	X

5.5.5 Operation

This section describes the steps to be taken in order to prepare subjects and data sets in order to perform the experiments and collect the relevant quantitative and qualitative data. As mentioned earlier, this stage involves two phases. The former uses specific benchmarks in order to collect quantitative data regarding the performance of the *I-BiDaaS* platform and its components (further analyzed in section 5.6). The second phase concerns the collection of qualitative data from industrial users regarding the application of the *I-BiDaaS* solution in real settings (described bellow). It should be noted that the described work flow is only indicative as further alignment will be necessary depending on the specific experiments that will be performed in the context of each use case.

In particular, the subjects will be first asked to experiment with the pipeline already in place (without *I-BiDaaS* components). Then they will be asked to fill in an on-line questionnaire that addresses dimensions of importance, such as impact, usability, robustness, efficiency, satisfaction and effectiveness.

This questionnaire will allow us to assess the performance experienced by the participants during their interaction with the system as is. Then, they will be given a short training on the new pipeline with the new components in place, and will be asked to repeat some tasks, as executed earlier, but with the new functionalities.

Table 5.6: Testing individual components quality - subject to refinement in the next project phase - part 1

Evaluation parameter	Evaluation variable	Measuring mechanism	To be recorded by		
			End User	IT User	Technology Provider
Sub-component specific performance parameters - DFP, Universal Messaging, Advanced ML, CEP Engine					
DFP (IBM)	Scalability	Linear (in number of DFP instances) speedup in generated data records			X
	Validity	Generated data must fit the data model			X
	Performance	Number of generated records per time unit			X
	Compliance	Measured against relevant standards			X
	Availability	No crashes.			X
Universal Messaging (SAG)	Scalability	Response time			X
		Data throughput			
		Resources utilization			
	Operational Performance				X
	Availability				X
	Reliability	Data failure			X
	Data Security				X
	Compliance	Measured against relevant standards, e.g. JMS, MQTT, AMQP			X
Advanced ML (UNSPMF)	Metrics for quality of results (e.g., including a subset of the following: Classification Accuracy; Value of empirical loss function; Confusion Matrix; Area under Curve; F1 Score; Mean Absolute Error; Mean Squared Error; R∧2 score);				X
CEP Engine (SAG)	Scalability	Response time			X
		Data throughput			
		Resources utilization			
	Operational Performance				X
	Availability				X
	Reliability				X

The questionnaire will be repeated and the new performance experienced will be recorded. In this way, we will be able to assess the performance of the platform given the newly deployed methods and tools.

The datasets will be prepared for the two experiments, one without *I-BiDaaS*, and one with *I-BiDaaS*. It is important to use the same datasets for the 'before' and 'after' evaluations, in order to make the experiments comparable.

Table 5.7: Testing individual components quality - subject to refinement in the next project phase - part 2

Evaluation parameter	Evaluation variable	Measuring mechanism	To be recorded by		
			End User	IT User	Technology Provider
Sub-component specific performance parameters - GPU-accelerated analytics, Hecuba DB System, Qbeast, Visualisation Tool					
GPU-accelerated analytics	Performance	Throughput			X
		Latency			X
Hecuba Database System (BSC)	Scalability	Speedup			X
	Operational Performance	Response time			
		IOPS			
		Disk usage			
	Availability	Percentage of timeouts			
	Reliability	Fault tolerance to down nodes (amount of missed data and slowdown during the recovery process)			
Qbeast (BSC)	Scalability	Speedup			X
	Operational Performance	Response time			
		IOPS			
		Disk usage			
	Availability	Percentage of timeouts			
	Reliability	Fault tolerance to down nodes (amount of missed data and slowdown during the recovery process)			
Visualization Tool (AEGIS)	Operational Performance	Response time			X
		Data throuput			
		Resources utilization			
	Scalability	Data volume visualised			X
	Availability	Uptime			X
	Reliability	Data failure			X
	Usability (Efficiency)	Task time efficiency	X		
		Perception of time required to accomplish a task			
		Perception of task completion quality			
	Usability (Satisfaction)	Degree to which user needs are satisfied – look and feel	X		

5.6 Benchmarking

As mentioned in Section 5.2, a number of existing benchmarking solutions might be used for testing the *I-BiDaaS* platform. In this section, we describe one such indicative solution BigDataBench 4.0.[22], as a comprehensive benchmark which has been successfully applied to evaluate multiple sys-

Table 5.8: Human resources to take part in the experimentation process

Subjects	CAIXA	CRF	TID	Other
Quality assurance and control managers	1	-	-	1
Data analysts	3	2	3	2
Financial administrators	1	-	-	1
Human resource officers		1	-	1
Infrastructure engineers	1	1	1	2
IT security personnel	2	-	-	1
Decision executives in the production	-	2	1	1
Resource planning managers	-	2	-	1
Call centre employees	-	-	2	1
Network infrastructure engineers	-	1	2	1
Market and business analysts	-	1		2

tems and architectures [2] along the different horizontal concerns of the BDV Reference Model [16]. It should be noted that the selection of a specific benchmark will be further elaborated in the course of the project (in the context of Workpackage 6).

In the remaining section we provide a more elaborate description of the BigDataBench 4.0 in order to highlight its potential application in the context of the *I-BiDaaS*.

The BigDataBench 4.0 has been developed having in mind modern Big Data and AI workloads, which are not only complex, but also fast challenging and expanding. Complex workloads may aggravate the cost of porting benchmarks to individual sites, and across different architectures and systems. This also raises the difficulty in reproducibility of the data collected. The concept of *dwarfs*, which are abstractions of frequently-appearing units of computation may be regarded as an important step towards achieving scalability of the tests, while keeping them relatively simple to implement and reproduce. In the context of *I-BiDaaS*, each dwarf captures the common requirements of each class of unit of computation, which has been identified by the project.

BigDataBench 4.0 is a benchmarking approach that intends to cover a wide spectrum of potential applications, which include the use of diverse Big Data and Artificial Intelligence (AI) workloads. Actually, there exist eight so-called dwarfs, which represent the diversity of combinations. In the current version, 13 representative real-world data sets are provided, on top of which 47 Big Data and AI benchmarks can be used. This makes ample possibilities for comparisons with related Big Data architectures such as the Lambda architecture. The types of workloads supported include online services, offline analytics, graph analytics, AI, data warehouse, NoSQL, and streaming. As such, it covers all initially use cases jointly discussed by the project partners for the testing of the MVP architecture.

[2] A detailed analysis of the applicability and usage process of the BigDataBench 4.0 is provided in [39].

5.6.1 Benchmarking Methodology

The employed methodology follows the BigDataBench 4.0 methodology and its three principles. First, the specification of the benchmarks is separated from the implementation. Under this methodology each use case will be modelled as an independent use case from the underlying Big Data Systems. Second, the benchmark should apply to state-of-the-art AI and other (e.g. graph analysis) algorithms that are used by the individual use cases. This must be agreed with the use cases providers. And, third, the existing data sets must be carefully selected to fit in style and scope as close as possible to the data sets that are actually used by the project partners.

5.6.2 Big Data and AI Dwarfs

The types of computations that can be involved in the Big Data and AI dwarfs are currently eight: Matrix (involving vector-vector, vector-matrix and matrix-matrix computations), Sampling (involving selection of a subset of original data), Transform (indicates the conversion from an original domain to another domain, example, Fast Fourier Transformation), Graph (uses nodes representing entities and edges representing dependencies), Logic (bit manipulation), Set (operations on one or more collections of data), Sort and Basic (statistic) computations. The dwarf set caters, but does not include the basic operations of get, put, post, and delete, which are used in online services. These have been defined as Basic operations following previous works [44].

5.6.3 Benchmarking Process

The best way to view this process is from the top of the pyramid, where individual dwarfs are being tested for against the individual use cases (see Figure 5.3). The already designed micro-benchmarks concentrate on basic individual features of the use cases, including their CPU-intensive, memory-intensive and I/O intensive aspects.

Following this, combinations of such dwarfs are formed and may be regarded as component-level benchmarks. These can be understood as representative workloads in different application domains. A DAG-like structure can be used to describe them. For example, Googlenet is a combination of five different dwarfs, including Matrix, Transform, Sampling, Logic, and Basic statistics.

This process will eventually culminate with the analysis of our more complex end-to-end applications. These will cater for additional non technological aspects, such as organisational aspects, which are related to the introduction of the *I-BiDaaS* methodology. Such end-to-end scenarios will

be drafted during the following project period and will aim at evaluation of the technology at the premises of the consortium partners.

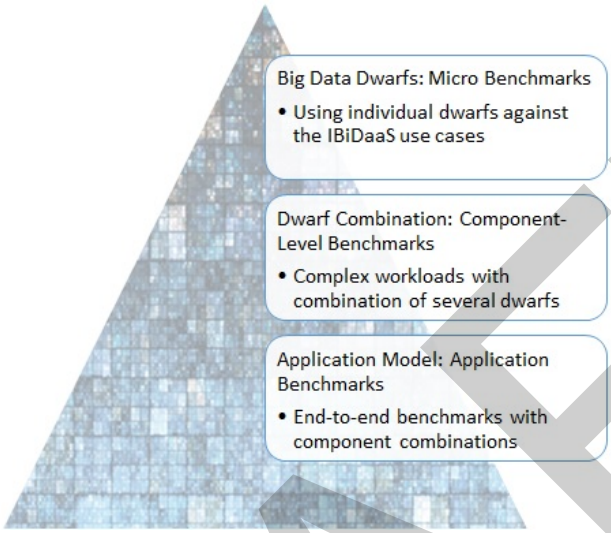

Figure 5.3: *I-BiDaaS* process view

The initial set of experiments conducted with the BigDataBench 4.0 implies that the designed pipeline is very efficient compared to traditional benchmarks, such as SPEC CPU 2006 [87], PARSEC [78], and HPCC [47].

The authors of the BigDataBench 4.0 have compared the new benchmarking approach with existing methods and provide detailed characterisation of the obtained results. Their comparisons (with other existing benchmarks) have concentrated on aspects, such as *retiring*, which is the slots fraction used by useful work and has been assessed at an average of 22.9%. Furthermore, they show that the the different workload types reflect diverse pipeline behaviours.

Bad speculation is a characterisation of the benchmark from the view point of the slots fraction wasted due to incorrect speculations, including branch misprediction and machine clears, which is related to the computing resource (processor). The authors have assessed that the benchmark has a relatively small fraction of bad speculation.

Back-end bound and *front-end bound* are two additional characterisation approaches for the benchmarks. Back-end bound occurs when the back-end has not enough required resources to accept new operations in the micro-operation cache, and can be divided into back-end core bound and back-end memory bound. Back-end core bound refers to hardware being lack of resources (e.g. divider unit) or underutilization of an execution port. The back-end memory bound represents the stalls due to load or store instructions. The front-end bound occurs when front-end undersupplies the back-end in a cycle. It is composed of two categories, front-end latency

bound (i.e. delivers no uops) and front-end bandwidth bound (i.e. delivers non-optimal amount of uops). From the provided data, we can observe, for example, that eight out of twelve Spark-based benchmarks have low front-end bound percentages, only occupying less than 8% on average. NoSQL and data warehouse have 35% and 25% front-end bound traffic.

5.6.4 Benchmarking in relation to AI Applications

AI benchmarks always need some training, which requires hundreds of iterations. From architecture viewpoint AI benchmarks are deemed to be too time consuming, even if running on GPUs. The authors of the benchmark have evaluated the impact of the number of iterations on fifty micro-architectural characteristics of the AI methods. Experiments have been conducted by using small, medium and large numbers of iterations, e.g. for training five different neural networks. The conclusions of their work is that small number of iterations are enough for micro-architectural evaluation of AI benchmarks. This reduces the time necessary for benchmarking AI (neural networks) related approaches.

5.6.5 Matching Workload Types to the Use Cases

The BigDataBench 4.0 features seven workload types, particularly, online service, offline analytics, graph analytics, artificial intelligence (AI), data warehouse, NoSQL and streaming. We have identified several workload types that could potentially be used with our use cases (see Table 5.9).

Table 5.9: *I-BiDaaS* use cases against the BigDataBench 4.0 workload types (all potentially applicable)

No	Use Case Name	Applicable BigDataBench 4.0 workload types
1	Building of a social graph (CAIXA)	graph analytics, artificial intelligence, data warehouse, NoSQL, streaming
2	Enhance control over third party agencies (CAIXA)	online service, offline analytics, data warehouse
3	Advanced analysis of bank transfer payment in financial terminal (CAIXA)	online service, offline analytics, graph analytics, data warehouse
4	Analysis of relationships through IP address (CAIXA)	graph analytics, offline analytics, streaming
5	Employment of bots in call center (TID)	artificial intelligence, data warehouse, NoSQL, streaming
6	Accurate location prediction with high traffic and visibility (TID)	artificial intelligence, graph analytics, streaming
7	Optimization of placement of telecommunication equipment (TID)	offline analytics, artificial intelligence, data warehouse
8	Predictive maintenance (CRF)	artificial intelligence, data warehouse
9	Production process of Aluminium Casting (CRF)	online services, data warehouse, NoSQL, streaming

5.6.6 Computing Resources - Infrastructure Characterisation

Illustration of the overall benchmarking methodology and process has been done by using a reference, typical state-of-practice processor, an Intel Xeon E5-2620 V3 processor.

The above mentioned seven workload types have been already characterised by using the Top-Down method [100]. Namely, in order to obtain comparative results, it is necessary to take into account the specifics of the underlying processor and the possibilities for its optimisation for specific workloads. The Top-Down method is a low-cost analysis method and is already featured in production systems. It requires just eight simple new performance events to be added to a traditional Performance Monitoring Unit (PMUs). This method is also comprehensive, it has no restriction to predefined set of performance issues. The method is now well established, and also accounts for granular bottlenecks in super-scalar cores.

5.6.7 Reference Data Sets

The benchmark includes 13 representative data sets, including:

1. a Wikipedia unstructured data set consisting of 4,300,000 English articles,
2. an Amazon semi-structured dataset consisting of 7,911,684 reviews on 889,176 movies by 253,059 users,
3. an unstructured Google Web Graph containing 875,713 nodes representing Web pages and 5,105,039 edges representing the links these between Web pages,
4. a Facebook social graph consisting of 4,039 nodes, which represent users, and 88,234 edges, which represent friendship between the users,
5. an e-commerce structured transaction data set, which is from a Web site (kept as anonymous). This data set is from an e-commerce web site, which we keep anonymous by request,
6. a data set, which is a semi-structured vertical search engine for scientists, consisting of 278,956 resums automatically extracted from 20,000,000 web pages of about 200 universities and research institutions,
7. a tiny image data set, which has 60,000 colour images with the dimension of 3232. They are classified into 10 classes and each class has 6,000 examples,
8. an image database organized according to the WordNet hierarchy,

9. a data set containing about one million labelled images, classified into 10 scene categories and 20 object categories,
10. a data set from translated TED talks, provided by the IWSLT evaluation campaign,
11. an unstructured data set including corpus and search query data, including an index and segment data with the total size of 4.98 GB,
12. a data set of handwritten digits, with a training set of 60,000 examples, and a test set of 10,000 examples,
13. a data set representing scores for movies, with 9,518,231 training examples and 386,835 test examples (semi-structured text).

All these data sets can effectively be used in the exploration of the *I-BiDaaS* use cases. Their applicability is currently being analysed by the project partners.

5.6.8 Reference Software Implementations

In order to facilitate comparisons with existing Big Data analytics platforms, the BigDataBench provides reference implementation of various state-of-the-art solutions. These include the following:

- Hadoop [7], Spark [13], Flink[37] and Message Passing Interface (MPI) implementations for offline analytics,
- Hadoop[7], Spark GraphX [84], Flink Gelly [38], and GraphLab [43] for for graph analytics,
- TensorFlow [90] and Caffe [23] for AI methods,
- Hive [9], Spark-SQL [85] and Impala [10] for data warehousing,
- MongoDB [68] and HBase [8] for NoSQL,
- Spark Streaming [14] and JStorm [3] for streaming.

5.7 Summary

Following the elicitation of the use case requirements, which guided the specification of the *I-BiDaaS* solution, this section focused on the definition of the appropriate process, variables and mechanisms for testing the performance of this solution and for evaluating its applicability to the different *I-BiDaaS* use cases. This was based on the information that was collected by user partners and technology providers as well as on the review of existing

big data benchmarks. In fact, bearing in mind the comprehensive process, which starts from small individual micro-benchmarks and gradually extends to full scale end-to-end evaluation scenarios, the number of data sets, which are suitable for almost all use case scenarios envisaged by the project partners, the set of reference software implementations, that can be directly used, the BigDataBench 4.0 was described as a potential candidate to be used in the *I-BiDaaS* experimentation process.

So far, we have performed the scoping and planning phases of the experimental process. With respect to future steps (i.e., operation and analysis), these will take place following the delivery of the integrated *I-BiDaaS* solution (MVP) in *WP5* (corresponding to the 1st phase of the exeprimental protocol). Next, the implementation and operation of real-life industrial experiments in *WP6* (corresponding to the 2nd phase of the experimental protocol) will take place. It should be noted that, this might result in additional elaboration of the defined experimental protocol variables and mechanisms as described under the task *T6.1* (Experimental protocol alignment).

6 Concluding Remarks

This deliverable *D1.3*, has provided an overview of the work carried out within all tasks of *WP1. Setting the scene: Baseline framework*. As such it forms the baseline providing direct input to the next project phases. At the same time it forms an evolving artifact that is expected to form a foundation stone that will be refined and revised in subsequent system implementation and experimentation phases as per the *I-BiDaaS* DoA.

It is important to note that three of WP1's tasks, T1.2 (i.e., requirements elicitation), T1.3 (i.e., architecture specification) and T1.4 (i.e., experimental protocol definition) were running essentially in parallel, forming a challenging feedback loop, especially between the first two tasks, that has been visited several times to obtain a coherent result.

We consider our past activities and this deliverable as an important step towards the elicitation and validation of requirements in the context of Big Data applications. Special emphasis has been paid to the description and elicitation of functional and non-functional requirements at both technical and business (high-level), which may serve as a reference process. The requirements have been distilled and mapped to a comprehensive experimentation protocol, which starts with small experiments and gradually broadens and extends its complexity towards evaluation in the business settings of the industrial partners.

D1.3 integrates the results from a significant amount of preliminary coordination in order to: a) define the nature and format of the experimentation data assets; b) elicit and document the *I-BiDaaS* platform business and technical requirements of common interest; c) investigate the *I-BiDaaS* KPIs and the industrial-validated benchmarks that will continuously monitor projects achievements and d) define the *I-BiDaaS* conceptual architecture and technologies updates.

As such it contributes to the achievement of most project objectives and especially the development and validation of a complete and solid Big Data as a self-service solution.

In particular, the *I-BiDaaS* use cases reported in *D1.3* have facilitated the identification of the data that is not available for the experiments in order to be fabricated (in WP2). In the following project phases, the requirements and architectural specifications will guide the development of the integrated solution of the project (in WP5), and combined with the experimental protocol will guide the execution of the cross-sectorial experiments (in WP6).

A Requirements Questionnaire

This appendix provides the template of the questionnaire we used to gather requirements from the three companies representing the financial, manufacturing and telecommunications industries. We asked the representatives to fill in a questionnaire for each test case they requested. The test cases of each industry are described in Section 3.

Along with the template questionnaire we provided to the end users a description of the requested fields of the questionnaire and possible values (according to the requirements typology).

Use Case Title		***Please specify***
Vertical (area)		*Please specify*
Partner name		*Please specify*
Corresponding person (email)		*Please specify*
A - Understanding business objectives		High level (business) requirements
Please describe what your organisation wishes to achieve in the context of I-BiDaaS. Try to focus on specific measurable objectives that will shape the data analytics services required in your company.		
What are the objectives that the use case serves?	Business Goals	*Please specify*
	Decision Goals	*Please specify*
	Analytics Goals	*Please specify*
	Data Processing Goals	*Please specify*
	Quality Goals	*Please specify*
	KPIs	*Please specify*
B- Understanding user requirements		
B1 - Understanding business data		Data source requirements
Please describe (as detailed as possible) the different data sets/sources that are planned to be used, both the data sets/sources internal to the company and external data. (For example, call logs from this period to that period, etc.) *Please complete this section for all datasets involved, repeating the rows as appropriate.*		
Dataset name:	*Please indicate the name of the dataset (if applicable)*	
Where will the source data come from?	*Please indicate the origin of the data, e.g., internal IT applications, Log files, Sensors, etc*	
Data owner	*Please describe whether the data is owned by your organisation (internal) or not (external)*	
Data content	*Please describe the kind of data contained in the dataset*	
Data type	*Are the data structured / semi-structured / unstructured?*	
Data format	*List / describe data formats*	
How is the data stored?	*Are databases used? If yes, what specific databases are used? Is there a data model that can be shared with the consortium?*	
How can the data be consumed?	*Please describe whether data can be consumed directly from the source or downloaded over a network or internet API or if it is available only as a real-time stream.*	
Is the data stored at geographically different locations?	*If yes, please specify.*	
What are the characteristics of the data?	Volume	*Specify quantitative volume of data handled in the use case*
	Velocity	*Specify the rate of flow at which the data is created*
	Variety	*Does the data come from multiple sources and multiple formats?*
	Variability	*How does the data rate varies in time? Please describe.*
	Volatility	*How does the nature of data varies in time? Please describe.*
	Consistency	*Is data consistent along the different data sources? Please describe.*

Does the data contain sensitive data?	Classified Information	*If yes please specify*		
	Personally Identifiable Information	*If yes please specify*		
	Other "sensitive" data	*If yes please specify*		
	Degree of privacy required	*Who can the data be revealed to and under which conditions?*		
B2 - Understanding data analysis		Analytics capability provider requirements		
What type of analysis will be considered?	<Stage of analysis>	<Current Tool/Algorithm used>	<Required Tool / Algorithm considered>	
	Data to Information	*Please indicate any data preparation used*	*Please indicate any data preparation required*	
	Information to knowledge analysis	*Please indicate any machine learning applications used*	*Please indicate any machine learning applications required*	
	Knowledge to wisdom analysis	*Please indicate any decision support applications used*	*Please indicate any decision support applications required*	
What are performance parameters that you wish to improve?	<Parameter name>	<Current value>	<Improvement sought>	<Priority>
	Please specify	*Please specify*	*Please specify*	*Please specify*
Where will the results be stored? **Do they have to stay within specific jurisdiction?**	*Please indicate if the results should remain within the company, country, EU, other*			
Do you want to combine the results with data from other sources?	*If yes please specify which ones*			
Do you want to integrate with an existing Business Intelligence system?	*Do you already use some kind of a business intelligence system; if yes, which one is that?* *Would you like to integrate it with the I-BiDaaS platform (bearing in mind licensing, contractual, and other restrictions)?*			
B3 - Understanding data reporting		Data consumer requirements		
Who will use analytics reports and how?	*Please indicate the type of access control required*			
How will the results be visualised?	*Please indicate the type of visualisation foreseen*			
What type of interaction is envisaged?	*Please indicate the way decision makers will interact with the data*			
What device will be used for viewing the results?	*Please indicate all possible devices (desktop, mobile, other)*			
What is the requirement for speed/frequency in delivering the results?	*What is the requirement for speed/frequency in delivering the results?*			
Other issues (please add):		Additional requirements		
Please specify any issues not covered by the above questions				

Use Case Title: Title provided by the use case author
Vertical Area: Intended to categorize the use cases. E.g. Financial, Manufacturing, Telecommunications
Partner name: Name of partner organisation submitting the use case
Corresponding person: email of the person(s) submitting the use case

- **A - High level (business) requirements.** Describe what the organisation wishes to achieve
    - **Business Goals**: Express company vision. They can be: Strategic; Operational
    - **Decision Goals**: Describe how big data will be used in order to achieve business goals. They can be of the following type: One-time analysis for a specific business decision; Open "blue sky" exploration of "interesting" data; On-going reporting and BI
    - **Analytics Goals**: Relate to the type of analysis (algorithm) required in order to realise decision goals. They can fall in one of the following categories: Historical analysis and reporting (descriptive / diagnostic); Predictive analysis and reporting (predictive / prescriptive)
    - **Data processing Goals**: Similar to analytics goals. Focus on the data processing lifecycle activities: Acquisition, Preparation, Analysis, Visualisation
    - **Quality Goals**: Quality properties that may relate to any other type of goal. They may relate to: Security, Performance, Scalability, Reliability, Availability
    - **KPIs**: Indicators for evaluating the achievement of all objectives, often led by quality goals. Examples are: Computational Cost, Infrastructure Cost, Time, etc.
- **B- User requirements**. Describe the requirements of the different stakeholders (data owners, capability providers, data consumers)
- **B1- Data source requirements**. Describes requirements relevant to the origin and characteristics of data (including the 3Vs). They might differ for different data sources.
    - **Dataset name**: Dataset identifier name, if applicable.
    - **Data source**: Describes the origin of data. This might include: Internal IT applications, Log files, Sensors, Mobile devices, Social Media, Data feeds, Government and special interest groups, Commercial Data Providers
    - **Data owner**: Differentiates between internal and external data sources
    - **Data content**: Describes the kind of data contained in the data set. For example: attribute level data (data that can be used to identify an entity), or transaction level data (data generated by an entity)
    - **Data type**: Refers to the nature of data, such as: structured, semi-structured, unstructured
    - **Data format**: Refers to the style of data, such as: relational tables, spreadsheets, XML, JSON, NoSQL, plain text, e-mail, social media, website, mobile data, communications, media files, business apps files
    - **How is data stored**: Refers to the way data is *currently* stored (e.g. SQL database, NoSQL DB, based on Files, other), specifying the exact name of the database or other system used (if applicable)
    - **How can data be consumed**: Refers to the way data can be acquired from source(s)
    - **Is data stored in different locations**: Indicates whether data source(s) is distributed, centralized, local, or remote.
    - **Data characteristics:** Describe the properties of the (raw) data
        - **Volume**: Refers to the amount and size of data that should be analyzed and processed. Possible values: NULL (not applicable), 0, Terabytes, Petabytes, Exabytes

        - **Velocity**: Refers to the speed that data is produced. Possible values: NULL, 0, Periodically Batch, (Near) Real-time, Streaming
        - **Variety**: Refers to the diversity of the data (formats). For example if data is stored in XML and CSV, then the number of formats is 2. Possible values: NULL, 0, 5-10 Formats, 10-25 Formats, 25+ Formats. Please note that all different formats are listed in the **Data format** field earlier in the questionnaire.
        - **Variability**: Refers to the amplitude of the data flow rate over time (data rate). It describes the inconsistent speed at which big data is generated. An example of highly variable data includes social media, where sentiments and trending topics change quickly and often. Possible values: NULL, 0, Low, Medium, High
        - **Volatility**: Similar to variability. Refers to the structural changes of data over time (nature of data). An example of highly volatile data includes natural language data. Possible values: NULL, 0, Low, Medium, High
        - **Consistency**: Refers to consistency between multiple copies of data on multiple machines in distributed systems or between same data on multiple sources. Possible values: NULL, 0, Strict consistency, Eventual consistency, General weak consistency
    - **Sensitive data**: may refer to classified information, personally identifiable information, other
        - **Degree of privacy required**: data can be revealed only to a customer, data stays inside your business organization, data can be shared with partner companies, data can be made publicly available
- **B2 - Analytics capability provider requirements**. Describes requirements with respect to the data analytics services
    - **Type of analysis**: Refers broadly to tools and algorithms currently in use / considered for use for processing the data at any stage of analysis.
        - **Data** to **information stage**: Refers to converting raw data into information: It includes traditional data preparation algorithms such as: Reduction, Cleaning, Transformation, Integration
        - **Information to knowledge stage**: Refers to extracting knowledge from information, it includes machine learning algorithms such as clustering, classification, pattern discovery
        - **Knowledge to wisdom stage**: Refers to applying knowledge to make decisions, it may include: simulation, optimization, decision support
    - **Analytics performance parameters**: Describe non-functional requirements relevant to the data analytics services
        - **Prioritisation**: helps to identify the most valuable performance requirements from the given set by distinguishing the critical few from the trivial many. Possible values: High (very valuable to the company), Medium (valuable but not urgent), Low (low value)
    - **Data Jurisdiction**: Company, Country, European, extra-territorial
    - **Combination with other data sources**: Refers to integration requirements with other sources
    - **Combination with existing BI system**: Refers to interoperability requirements with other systems
- **B3 – Data consumer requirements**. Describes user interface requirements
    - **Who will use the reports and how.** Specifies Access control requirements, such as: Basic role based access control, access control / application level, access control at resource level additional privacy

aware access control; whereby different roles may include: public user, VIP user, internal/external developer, internal/external support staff, internal operator, etc.
    - **Visualization**: Refers to the way the data is viewed by an analyst making decisions based on the data. Possible values: Table, Chart, Graph, Map, Network layout
    - **Type of interaction**: Refers to the way decision makers interact with the data, e.g.: Overview, Zoom, Filter, Details on Demand, Streaming results
    - **Device**: Refers to accessing devices including mobile devices
    - **Speed of delivery**: E.g. near-real time
- **B4 – Additional requirements.** Allows for the addition of requirements not covered by existing questions

B

Elicited Requirements

This appendix incorporates the list of all use case requirements elicited by *I-BiDaaS* industrial partners, using the requirements questionnaire. All gathered information is categorised according to the categories in the requirements typology (in red).

Use case	Business Goals		Decision Goals		Analytics Goals		Data Processing Goals		Quality Goals		KPIs	
TID-1: Employment of bots in call centre	Improve customer experience	Strategic	Employ advance Machine Learning techniques		Develop speech technologies that transform audio calls (customer Call Centre conversations) into relevant information for the Call Centre that can be used to assess its performance and/or to screen phone calls automatically	Historical analysis and reporting (descriptive / diagnostic)	Automate procedure for the audio call transformation and metadata generation and analysis	Acquisition	Improve performance of audio calls processing	Performance	number of audio calls processed / per time unit	Other
			Use of big data to Employ bots in call centres	Ongoing reporting and BI							% of audio calls monitored	Other
TID-2: Accurate location prediction with high traffic and visibility	Improve the customer experience	Strategic	Employ advance Machine Learning techniques		Analyse the behaviour of local and non-local customers over various periods of time (e.g. holidays), and extract insights on the behavioural patterns of groups of people, enabling TID to optimize their value propositions	Historical analysis and reporting (descriptive / diagnostic)	Enable the automatic extraction of behavioural patterns of customers	Analysis	Minimise processing time with growing data sizes while maintaining real-time requirements	Performance	Data processed / time unit	Other
			Use of big data for accurate location prediction with high traffic and visibility	Ongoing reporting and BI								
TID-3: Optimization of placement of telecommunication equipment	Improve routing and placement of the telecommunication equipment	Operational	Use of big data to optimize the placement of telecommunication equipment	Ongoing reporting and BI	Based on the data retrieved by the customers, their usage and the location, execute algorithms and further improve the routing and placement of the telecommunication equipment that is already in place or arrange accordingly the new	Historical analysis and reporting (descriptive / diagnostic)	Deal with data sparsity	Analysis	Reduce high computation and storage costs	Efficiency	Computation cost	Computational Cost
											Storage cost	Infrastructure Cost
CAIXA 1: Building of a social graph	Speeding up the implementation of Big Data solutions, particularly graph databases, inside bank infrastructure.	Operational	Take advantage about which technology is the most suitable, taking into account the volume of customers of the bank.	One-time analysis for a specific business decision	Academic operations on a graph, checking metrics related to computational cost.	Historical analysis and reporting (descriptive / diagnostic)	Automate procedure for partial dataset generation depending on the specific case.	Acquisition	Data Analytics performance.	Performance	Data charging time.	
	Review the impact of cost infrastructure with different volume of data.	Operational	Evaluate the future usage of proposed project tools (e.g. IBM Data Fabrication Tool Solution)	One-time analysis for a specific business decision	Test some well-known algorithms oriented to graphs.	Historical analysis and reporting (descriptive / diagnostic)	Improve processing of charged data.	Analysis	Test the scalability and response time of the graph solution (the system has to take into account that the graph may increase exponentially).	Scalability	Time to get analytics results.	Time
	Enable safe experimentation and testing of biga data analytics tools	Strategic							Reduce time when complex relationships are request.	Efficiency	Count number of nodes by category.	Other
							Improve information extraction from charged data.	Analysis	The response of the system must be brief (seconds).	Efficiency	Count number of edges by category.	Other
											Price of technologies.	Infrastructure Cost
											Time to generate business rules.	Time
CAIXA 2: Enhance control over third party agencies	Avoid not authorized connections.	Operational	Inspect whether employees of third-party providers follow CaixaBank's policies.	Ongoing reporting and BI	Validation of the analytical process, the accuracy and reliability.	Historical analysis and reporting (descriptive / diagnostic)	Automate procedure for partial dataset generation depending on the specific case.	Acquisition	Improve security of outsourced services.	Security	Data charging time.	Time
	Minimise loss of control when outsourcing tasks to third party agencies.	Strategic	Reinforce insights in fraud prevention department.	Ongoing reporting and BI			Improve processing of charged data.	Analysis	Improve efficiency and security practices.	Efficiency	Time to get analytics results.	Time
	Improve detection and avoidance of misbehaviours and illicit usage of CaixaBank services.	Operational	Evaluate the future usage of proposed project tools (e.g. IBM Data Fabrication Tool Solution).	One-time analysis for a specific business decision			Improve information extraction from charged data.	Analysis	Reduce costs from illicit activities from clients and/or third-party companies.	Efficiency	Time to validate previous results of real data analytics.	Time
	Enable safe experimentation and testing of biga data analytics tools	Strategic										
	Get a tool for third parties to implement work in CaixaBank environment without accessing real data	Operational							Maintain CaixaBank security standards.	Security	Price of technologies.	Infrastructure Cost
	Work without real data in order to speed up the implementation of new technologies.	Operational							Maintain CaixaBank data confidentiality standards.	Security	Time to generate business rules.	Time

Use case	Business Goals		Decision Goals		Analytics Goals		Data Processing Goals		Quality Goals		KPIs	
CAIXA3: 3 Advanced Analysis of bank transfer payment in financial terminal	Avoid potential fraud related to money transfers among customers.	Strategic	Perform analytics of large volumes of data to automatically improve the current fraud systems of the bank.	One-time analysis for a specific business decision	Validation of the analytical process, the accuracy and reliability.	Historical analysis and reporting (descriptive / diagnostic)	Automate procedure for partial dataset generation depending on the specific case.	Acquisition	Ensure that generated data provides the same insights than real data.	Efficiency	Data charging time	Time
	Avoid fraud committed by employees.	Strategic	Reinforce insights in fraud prevention department.	Ongoing reporting and BI			Processing the bank transfer from outside office time.	Analysis	The process must guarantee the result before making the analysis.	Efficiency	Time to get analytics results	Time
	Enable safe experimentation and testing of biga data analytics tools	Strategic										
	Get a tool for third parties to implement work in CaixaBank environment without accessing real data.	Operational	Evaluate the future usage of proposed project tools (e.g. IBM Data Fabrication Tool Solution).	One-time analysis for a specific business decision					Reduce costs from illicit activities.	Efficiency	Confusion matrix.	Other
	Work without real data in order to speed up the implementation of new technologies.	Operational									Price of technologies.	Infrastructure Cost
											Time to generate business rules.	Time
CAIXA 4: Analysis of relationships through IP address	Develop Proof of Concepts.	Strategic	Speeding up the implementation in production environments of new tools and algorithms.	One-time analysis for a specific business decision	Discover the same (or new) relationships from synthetic data.	Historical analysis and reporting (descriptive / diagnostic)	Automate procedure for partial dataset generation depending on the specific case.	Acquisition	Reliability of the insights generated (the relationships must be valid).	Reliability	Comparison of real results vs results from synthetic data	Other
	Test new technologies.	Strategic	Evaluate the future usage of proposed project tools (e.g. IBM Data Fabrication Tool Solution).	One-time analysis for a specific business decision	Detect hidden patterns related to IP address of customers.	Historical analysis and reporting (descriptive / diagnostic)			Validate the types of relationships discovered.	Other	Accuracy	Other
	Work without real data in order to speed up the implementation of new technologies.	Operational			Validation of the analytical process, the accuracy and reliability.	Historical analysis and reporting (descriptive / diagnostic)			Ensure that generated data provides the same insights than real data.	Efficiency	Recall	Other
					Find relationships between users from IP address.	Historical analysis and reporting (descriptive / diagnostic)			The process must guarantee the result before making the analysis.	Efficiency	TP rate	Other
									Reduce costs from illicit activities.	Efficiency	TN rate	Other
											Confusion matrix	Other
CRF1: Maintenance and monitoring of production assets	Be competitive in a very competitive sector	Strategic	Use Big Data for improving production line	Ongoing reporting and BI	Analyse sensor-generated data	Predictive analysis and reporting (predictive / prescriptive)	Multichannel Acquisition (data gathering)	Acquisition	Reduce storage costs - and to avoid the unsold	Efficiency	Overall equipment Effectiveness(OEE)	Other
	Improve mnitoring and maintenance of production assets	Operational			Integrate sensor-generated data with Manufacturing Execution System (MES).	Predictive analysis and reporting (predictive / prescriptive)	Innovative Analysis and reactive visualization (wisdom)	Analysis	Decision making should be automated or possible to do in Real Time.	Performance	Job per hour (jph)	Other
					Automate decision making	Predictive analysis and reporting (predictive / prescriptive)		Visualisation	Improve maintenance efficiency	Efficiency	Maintenance costs	Other
					Analyse sensor-generated data	Predictive analysis and reporting (predictive / prescriptive)	Smart Preparation (data availability)	Preparation				

Use case	Business Goals		Decision Goals		Analytics Goals		Data Processing Goals		Quality Goals		KPIs	
CRF2: Production process of Aluminium Casting	Improve anticipation of defects	Operational	Use Big Data for improving quality to the production process	Ongoing reporting and BI	Study production error rates	Historical analysis and reporting (descriptive / diagnostic)	Multichannel Acquisition	Acquisition	Decision making should be automated or possible to do in Real Time.	Performance	Scrap percentage	Other
	Be competitive in a very competitive sector	Strategic			We want to understand the quality of the process.	Predictive analysis and reporting (predictive / prescriptive)	Innovative Analysis and reactive visualization	Preparation			Cost per unit	Other
					We want to understand the relationship between the quality of the product and correlate it with the data sensors	Predictive analysis and reporting (predictive / prescriptive)	Innovative Analysis and reactive visualization	Visualisation			Quality control and repair costs	Other
							Smart Preparation	Preparation				

Data Source Requirements

Use Case	Source		Owner		Content		Type		Format		Storage		Consumption		Location	
TID-1: Employment of bots in call centre	TID Call Centers	Internal IT applications	TID	Internal	hours of recorded speech	transaction level data	A mixture of structured and unstructured data.	unstructured	Audio data is recorded in Mono, based on ITU G.711 codec at 8kHz.	media files	The data are stored in separate audio and data files	based on Files	The data can be consumed using network file system (NFS).	NFS	Source data are captured in TID datacentres.	distributed
				Internal	metadata	attribute level data	metadata	structured	Meta-data will be structured in a unified message format	JSON						
TID-2: Accurate location prediction with high traffic and visibility	Data come from customer phones.	Mobile devices	TID	Internal	Event Data	transaction level data	event data type	structured	All events will be structured in a unified message (JSON-like) formats	JSON	Data files stored in HDFS	based on Files	Using HDFS connectors.	HDFS	The data are being collected real-time, from different geographic locations.	real-time
						attribute level data										
TID 3: Optimization of placement of telecommunication equipment	Data come from customer phones.	Mobile devices	TID	Internal	Cell network events	transaction level data	tabular form	structured	All events will be structured in a unified message (JSON-like) formats	JSON	Data files stored in HDFS	based on Files	Using HDFS connectors	HDFS	The data are being collected real-time, from different geographic locations.	real-time
CAIXA 1: Building of a social graph	CaixaBank Warehouse	Internal IT applications	CaixaBank	Internal	Synthetic data	attribute level data	structured data	structured	Standard relational database	relational tables	Relational database	SQL database	Standard SQL language	SQL query	Locally	local
CAIXA 2: Enhance control over third party agencies	Existing Security and Event Management System (IBM QRadar)	Log files	CaixaBank	Internal	Synthetic data		log files are unstructured text files	unstructured							Locally within CaixaBank facilities	local
CAIXA3: 3 Advanced Analysis of bank transfer payment in financial terminal	The data come from a dataset that is maintained at CaixaBank Warehouse.	Internal IT applications	Data are owned by CaixaBank.	Internal	Synthetic data, ensuring the privacy of customers, will be generated with CaixaBank Warehouse containing content such as: Customer information, Value of the transference, Office, Etc.	attribute level data	Structured data.	structured	More than 300 different sources of structured and unstructured data. A model is converting unstructured data to structured (Tables).	relational tables	Relational Database.	SQL database	Through standard SQL language.	SQL query	The CAIXA Data Warehouse dataset is stored locally within CaixaBank.	local

Data Source Requirements

Use Case	Source		Owner		Content		Type		Format		Storage		Consumption		Location	
CAIXA 4: Analysis of relationships through IP address	The data come from a dataset that is maintained at CaixaBank Warehouse.	Internal IT applications	Data are owned by CaixaBank	Internal	Synthetic data about any operation of a customer can do in a bank session.	transaction level data	Structured data.	structured	The format of the data is the format of a standard relational databases.	relational tables	Relational Database.	SQL database	Through standard SQL language.	SQL query	CAIXA Data Warehouse dataset is stored locally within CaixaBank	local
CRF1: Maintenance and monitoring of production assets	MES (Manufacturing Execution System)	Internal IT applications	FCA/CRF	Internal	Data about the production line and contains the id of the vehicle, the time that the vehicle arrived at each phase of the production line, the characteristics of the vehicle.	attribute level data	Structured data	structured	Excel files	spreadsheets	local repository	based on Files	directly from the source or will be placed in external cloud	other	locally in private clouds.	distributed
	SCADA sensors data	Sensors	FCA/CRF	Internal	Sensor data	transaction level data	Structured data	structured	Excel files	spreadsheets	local repository	based on Files	directly from the source or will be placed in external cloud	other	locally in private clouds.	distributed
CRF2: Production process of Aluminium Casting	Sensor data and Operators' data	Sensors Data feeds	FCA/CRF	Internal	Data of temperature Data of flow Data of pressure Data on quality of the product and the process	transaction level data	Structured data	structured	Excel files	spreadsheets		based on Files		other	Most of the data are stored locally in private clouds	distributed

Data Source Requirements

Use Case	Volume	Velocity		Variety		Variability		Volatility		Consistency		Sensitive Data		Degree of privacy required	
TID-1: Employment of bots in call centre	200 hours of speech per day for a month, plus some metadata per call, Terabytes. Terabytes	Some logs are collected real-time within seconds, others once per day.	Periodically Batch	A mixture of structured and unstructured data (see above in data type), 5-10 formats.	5-10 Formats	Medium	Medium	High	High	Strict	Strict consistency	Name, customer details that are included in the billing account.	personally identifiable information	Data stays inside the TID organization	data stays inside your business organization
		Some logs are collected real-time within seconds, others once per day.	(Near) Real-time												
TID-2: Accurate location prediction with high traffic and visibility	4TB per day (that can optionally be compressed to 15GB per day). Terabytes	Some logs are collected within seconds, other once per day. The majority of data is achieved in less than 5 minutes.	(Near) Real-time	3 formats	0	High	High	Low	Low	Strict	Strict consistency	Customer information (such as contract information from the user)	personally identifiable information	Data stays inside the TID organization	data stays inside your business organization
TID 3: Optimization of placement of telecommunication equipment	4TB per day (that can optionally be compressed to 15GB per day). Terabytes	Some logs are collected within seconds, other once per day. The majority of data is achieved in less than 5 minutes.	(Near) Real-time	3 formats	5-10 Formats	High	High	Low	Low	Strict	Strict consistency	Customer information (such as contract information from the user).	personally identifiable information	Data stays inside the TID organization	data stays inside your business organization
CAIXA 1: Building of a social graph															
CAIXA 2: Enhance control over third party agencies															
CAIXA3: 3 Advanced Analysis of bank transfer payment in financial terminal	2PB presently, growing 300.000.000 rows per day. Petabytes	Combination of real-time and periodic batch	Periodically Batch	Structured data in the following data models:	5-10 Formats	LOW	Low	LOW	Low	HIGH	Strict consistency	The dataset will be synthetically generated from corporate and customer information.	personally identifiable information	Data are owned by CaixaBank and are not publicly available.	data stays inside your business organization

Data Source Requirements

Use Case	Volume	Velocity	Variety	Variability	Volatility	Consistency	Sensitive Data	Degree of privacy required
		Streaming					classified information	
CAIXA 4: Analysis of relationships through IP address	7.000kk of rows per month	Depends on the request but in general fast. (Near) Real-time	Medium 5-10 Formats	LOW Low	LOW Low	HIGH Strict consistency	Corporate and customer information. classified information Customers' personal information. personally identifiable information	Data are owned by CaixaBank and are not publicly available. data stays inside your business organization
CRF1: Maintenance and monitoring of production assets	TB Terabytes	Real time streaming Streaming	All the data have the same format in an excel file	0	0	0 NULL	YES classified information	In general all proprietary data should remain within the company data stays inside your business organization
	TB Terabytes	Real time streaming Streaming	All the data have the same format in an excel file	0	0	0 NULL	YES classified information	In general all proprietary data should remain within the company data stays inside your business organization
CRF2: Production process of Aluminium Casting	GB 0	Real time streaming (Near) Real-time		0	0	0 NULL	YES classified information	All proprietary data should remain within CRF data stays inside your business organization

Use Case	Type of Analysis	Data to information		Information to knowledge		Knowledge to wisdom		Performance Parameters	Jurisdiction
TID-1: Employment of bots in call centre	Multiple	YES		YES		NO		% of low customer satisfaction index (CSI) customer audio calls analyzed / per time unit Customer satisfaction. Such a performance parameter can encapsulate the efficiency and accuracy of a Call Centre.	The results will remain in Telefonica data centers.
TID-2: Accurate location prediction with high traffic and visibility	Multiple	Data aggregation Group given data by user and sort events chronologically	Integration	Interpolate missing events to recover plausible event trajectories	prediction	Forecast immediately next events to anticipate movements at scale.	optimization	Data processing volume Data processing time	Telefonica Data centers
TID-3: Optimization of placement of telecommunication equipment	Multiple	Data aggregation Group given data by user and sort events chronologically	Integration	Interpolate missing events to recover plausible event trajectories	prediction	Forecast immediately next events to anticipate movements at scale.	optimization	Data processing volume Data processing time	Telefonica Data centers
CAIXA 1: Building of a social graph	Information to knowledge	NO		Any tool or algorithm valid to check performance will be of high value.	prediction	NO		N/A	All the results must be stored in CaixaBank's infrastructure
CAIXA 2: Enhance control over third party agencies	Information to knowledge	NO		Techniques related to log analysis such as process mining algorithms or similar.	classification prediction	NO		N/A	All the results must be stored in CaixaBank's infrastructure
CAIXA3: 3 Advanced Analysis of bank transfer payment in financial terminal	Information to knowledge	NO		Generation of synthetic data rules to validate the current analytics.	validation	NO		N/A	All the results must be stored in CaixaBank's infrastructure
CAIXA 4: Analysis of relationships through IP address	Information to knowledge	NO		Generation of synthetic data rules to validate the current analytics.	validation	NO		N/A	All the results must be stored in CaixaBank's infrastructure
CRF1: Maintenance and monitoring of production assets	Multiple	Integrate the two datasets (MES and SCADA	Integration	Combine information about the quantity of the materials, back-up systems and their availability or usage	prediction	The final system should be able to produce real-time automated decisions	decision support	Analytics performance	The results will be sent back to the local repository
CRF2: Production process of Aluminium Casting	Multiple	NO		Combine information about the quantity of the materials, back-up systems and their availability or usage So that we can understand the parameters that affect quality of the process.	prediction	The final system should be able to produce real-time automated decisions	decision support	Analytics Performance	The results will be sent back to the local repository

Use Case		Combination with other data sources		Combination with existing BI system	
TID-1: Employment of bots in call centre	Company	NO	NO	NO	NO
TID-2: Accurate location prediction with high traffic and visibility	Company	NO	NO	NO	NO
TID-3: Optimization of placement of telecommunication equipment	Company	NO	NO	NO	NO
CAIXA 1: Building of a social graph	European	Any information about performance of graph databases may be considered.	YES	NO	NO
CAIXA 2: Enhance control over third party agencies	European	YES	YES	NO	NO
CAIXA3: 3 Advanced Analysis of bank transfer payment in financial terminal	European	NO	NO	Not in an initial phase, however depending on results, it may be interesting to integrate them into our BI systems.	YES
CAIXA 4: Analysis of relationships through IP address	European	NO	NO	NO	NO
CRF1: Maintenance and monitoring of production assets	Company	In the future	YES	NO	NO
CRF2: Production process of Aluminium Casting	Company	In the future	YES	NO	NO

Data Consumer Requirements

Use Case	Access Control		Visualisation		Interaction		Device		Speed of delivery	
TID-1: Employment of bots in call centre	Access and Reuse of Personal or corporate Datasets is limited to TID as its owner. Reports will be access to TID's staff.	access control at resource level	Table, Chart	Table Chart	Overview, Filter, Details on Demand	Overview Filter Details on Demand	Desktop/laptop devices	PC	Hourly or Daily updated results	Hourly Daily
TID-2: Accurate location prediction with high traffic and visibility	Access and Reuse of Personal or corporate Datasets is limited to TID as its owner. Results and reports will be accessible to TID's staff.	access control at resource level	Tables, Charts, Maps	Table Chart Map	Overview, Filter, Details on demand	Overview Filter Details on Demand	Desktop/laptop device	PC	Hourly or daily updates	Hourly Daily
TID-3: Optimization of placement of telecommunication equipment	Access and Reuse of Personal or corporate Datasets is limited to TID as its owner. Results and reports will be accessible to TID's staff.	access control at resource level	Tables, Charts, Maps	Table Chart Map	Overview, Filter, Details on demand	Overview Filter Details on Demand	Desktop/laptop device	PC	Hourly or daily updates	Hourly Daily
CAIXA 1: Building of a social graph	CaixaBank's departments and partners working with Data Analytics and CaixaBank's Social Graph.	Basic role based access control	N/A		N/A		N/A		N/A	
CAIXA 2: Enhance control over third party agencies	Caixabank IT department and collaborating Caixabank third parties.	Basic role based access control	Currently not defined		N/A		N/A		N/A	
CAIXA3: 3 Advanced Analysis of bank transfer payment in financial terminal	The reports will be mainly consumed by iSOC (integrated Security Operation Center), however, there will be more departments involved.	Basic role based access control	Currently not defined		N/A		N/A		N/A	
CAIXA 4: Analysis of relationships through IP address	CaixaBank IT deparment.	Basic role based access control	N/A		N/A		N/A		N/A	
CRF1: Maintenance and monitoring of production assets	Operator on the line Data scientist Process manager	Basic role based access control	Interactive interface	Multiple	Operator on the line: only visualization Data scientist: can customize and then analyse data Process manager: collaborate with the data scientist and then takes decision on the action to actuate as a consequence of the analysis results	Multiple	Tablet / computer	PC Mobile	Near real time	Near real time
CRF2: Production process of Aluminium Casting	Operator on the line Data scientist Process manager	Basic role based access control	Interactive interface	Multiple	Operator on the line: only visualization Data scientist: can customize and then analyse data Process manager: collaborate with the data scientist and then takes decision on the action to actuate as a consequence of the analysis results	Multiple	Tablet / computer	PC Mobile	Near real time	Near real time

C Experimental Protocol Instruments

This appendix describes the instruments that have been used in order to collect data, from industrial users and technology providers, regarding the definition of qualitative and quantitative parameters that will be used in the experimental protocol. In particular a mixed-mode method was used involving a combination of semi-structured interviews and questionnaires. The former provides in-depth insights on participant attitudes, thoughts, and actions, whilst the latter produces generalisable results.

In the following paragraphs we first provide a complete list of questions that were asked in written form and at project meetings to inform the design of the *I-BiDaaS* experimental protocol. Next we provide the template of the questionnaire that was in addition to these questions. Once again known values were pre-filled to be validated and revised by participants. It should be noted that these questions are still valid and will be used also at a later stage of the project to refine details of the experimental protocol as part of *WP6* activities.

Interview Questions

1. In relation to expected project level achievements
   (a) How do you define Usability in relation to your contribution to the project?
   (b) How do you define Innovation (including in relation to i-v of the Technical Annex) in relation to your contribution to the project?
   (c) How do you define Robustness in relation to your contribution to the project?
   (d) How do you define Performance in relation to your contribution to the project?

(e) Your contributed architectural component applies to which of the above variables and how? Please elaborate on one or more of the following components: Pre-defined Queries, SQL-like interface, Domain language Programming API, Advanced ML via structured (non) convex optimisation, COMPs Programming Model, Complex Event Processing, GPU-accelerated analytics, Hecuba tools, Cassandra, Resource Management and Orchestration, Integrated platform, Universal Messaging (Data curation, transformation, pre-processing), Telecom data, Sensor Data, Data Fabrication Platform, Social Networks, Real time alerts, Advanced visualizations.

(f) Is any essential architectural component missing from the above list? Explain what it is and how it contributes to the project.

(g) Which steps will be taken to verify that certain variable (utility, innovation, robustness, performance) has been achieved and how? Explain the verification process.

(h) Which goals (variables) are essential to be achieved?

(i) Provide a list of use cases that can be supported with your technology, and explain how will you apply your technology? (Consider making a table.)

2. Use cases view point

   (a) How do you evaluate your use case as a whole? Please specify all processes, step by step, to be followed by your use case.

   (b) Specify step by step scenario that can be performed for verification of your use case, that utilizes components of the architecture.

   (c) What benefits are expected from the specified process?

   (d) Which architectural component do you think would be beneficial for your use case? (Consider making a table.)

3. View point of the subjects participating in the experiments

   (a) Which subjects will use the integrated solution (i.e. the MVP) as a whole (or in parts), and how?

   (b) Consider the following subjects: quality assurance and control managers, data analysts, financial administrators, human resource officers, infrastructure engineers, IT security personnel, decision executives in the production, resource planning managers, call centre employees, network infrastructure engineers, market and business analysts.

   (c) What type of interfaces will be used? See from the list: predefined queries, SQL-like interface, domain language, Programming API.

(d) What (essential) user interfaces must be provided by I-BiDaaS for its users?

(e) How would the users use the interfaces?

(f) How many users are required for testing functional/non-functional properties?

(g) Specify how many experimental subjects do you plan to involve through activities of your organisation including a) workshops (physical or virtual) organized by the data providers (T1.2), b) three INFO days organized by the consortium (CAIXA, CRF, TID), c) starting from M12 when the MVP will be released, through an open session that will be organized for the experimental subjects co-located with the projects General Assembly meeting, d) means already in place by the consortium, namely TID (WAYRA), TID (AURA) and CRF (Campus-Melfi).

4. Validation of the solution as a whole, industry benchmarks

   (a) What industry validation benchmarks would you use?

   (b) What functional properties will be verified and how?

   (c) What QoS properties will be verified and how?

   (d) Which monitoring system will be used to obtain QoS metrics?

   (e) How do you assure comparability of the obtained results?

Questionnaire

The following questions intend to facilitate (mainly) the scoping and planning of the experiments. Some questions (indicated in brackets) are addressed to either end users or technology providers, whilst others concern both. Please fill / revise the following tables focusing on the use case or *I-BiDaaS* platform component that you are responsible for.

General

Participant's Name	
Participant's Organisation	
Relevant use case(s)	
Relevant project component(s)	

Scoping: (End Users)

Please define relevant business and sector KPIs that can be used to measure the achievement of the objectives of each use case (as recorded through the user requirements questionnaires). Please indicate additional KPIs that you feel are relevant. ***Please note that it should be possible to link these KPIs to the experimental measurements***.

No	Use Case Name	KPI	Baseline value[1]	Improvement sought	Benchmark[2]
1	Building of a social graph (CAIXA)	- Data charging time. - Time to get analytics results. - Count number of nodes by category. - Count number of edges by category. - Price of the solution technologies. - Time to generate business rules			
2	Enhance control over third party agencies (CAIXA)	- Data charging time - Time to get analytics results. - Time to validate previous results of real data analytics. - Confusion matrix. - Price of technologies. - Time to generate business rules.			
3	Advanced Analysis of bank transfer	- Data charging time, - Time to get analytics results			

[1] Current performance
[2] Performance best practices to compare with

No	Use Case Name	KPI	Baseline value[1]	Improvement sought	Benchmark[2]
	payment in financial terminal (CAIXA)	- Confusion matrix. - Price of technologies. - Time to generate business rules.			
4	Analysis of relationships through IP address (CAIXA)	- Comparison of real results vs results from synthetic data. - Accuracy. - Recall. - TP rate. - TN rate. - Confusion matrix.			
5	Enhance control over third party agencies (CAIXA)	- Data charging time, - Time to get analytics results - Types of data handled - Cost			
6	Employment of bots in call centre (TID)	- number of audio calls processed / per time unit - % of audio calls monitored			
7	Accurate location prediction with high traffic and visibility (TID)	- Data processed / time unit			
8	Optimization of placement of telecommunication equipment (TID)	- Computation cost - Storage cost			
9	Predictive maintenance (CRF)	- Overall equipment Effectiveness(OEE) - Job per hour (jph) - Maintenance costs			
10	Production process of Aluminium Casting (CRF)	- Scrap percentage - Cost per unit - Quality control and repair costs			

Planning: (All)

Identify evaluation variables and measuring mechanism:

Using the following tables, please indicate the evaluation parameters and associated variables to be measured in the context of the experiments, removing irrelevant ones / adding new ones where applicable. *Please try to be as specific as possible with respect to the measuring mechanism to be used (e.g. name specific standards, benchmarks, testing tools to be used)*.

Table 1. Testing System[3] Functionality (All)

Evaluation parameter	Evaluation variable	Measuring mechanism	Relevant Functionality[4]						To be recorded by[5]		
			Data ingestion & preparation	Data Fabrication	Batch Analytics	Streaming Analytics	Data Visualisation	Programming Framework	End User[6]	IT User	Technology Provider
Usability[7]	Effectiveness (accuracy and completeness)	Participant completion rate Number of errors Software Improvement									
	Efficiency	Task time efficiency Perception of time required to accomplish a task Perception of task completion quality									
	Satisfaction	Degree to which user needs are satisfied – look and feel									

[3] Integrated solution
[4] Select all that apply
[5] Select all that apply
[6] See subjects table bellow
[7] Usability is the degree to which a software system can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use.

Table 2. Testing System Quality (All)

Evaluation parameter	Evaluation variable	Measuring mechanism	To be recorded by[8]		
			End User[9]	IT User	Technology Provider
Innovation[10]	Additional functionality beyond the system currently in use	Measured against the industrial requirements			
	Beyond the standard technologies	Measured against the latest Strategic Research and Innovation Agenda (SRIA) released by BDVA			
Robustness[11]	Efficient data integration	Number of different heterogeneous datasets Capability of integration and processing– error free			
Accountability[12]	Performance against commercial or current solutions	Measured against industrial-validated benchmarks Accountability metrics: A-PPL12			

[8] Select all that apply
[9] See subjects table bellow
[10] Innovation refers to the implementation of a new or significantly changed product or process.
[11] Robustness is the ability of a computer system to cope with errors during execution and cope with erroneous input
[12] Accountability is an assurance that an individual or an organization will be evaluated on their performance or behavior related to something for which they are responsible.

Table 3. Testing System Performance as whole and per component (Technology Providers)

Evaluation parameter	Evaluation variable	Measuring mechanism	To be recorded by[13]		
			End User[14]	IT User	Technology Provider
Integrated solution[15] (ITML)	Scalability				
	Operational Performance	Response time Data throughput Resources utilisation			
	Availability				
	Reliability	Data failure			
	Data Security				
	Privacy				
	Compliance	Measured against relevant standards			
	Cost	Measured against ?			
Sub-component specific performance parameters[16]					
DFP (IBM)					
Universal messaging (SAG)					
Advanced ML (UNSPMF)					
COMPSs programming framework (BSC)					

[13] Select all that apply
[14] See subjects table bellow
[15] Minimum Viable Product
[16] Add additional component specific evaluation parameters variables for testing each sub-components in isolation, if necessary

CEP engine (SAG)					
GPU-accelerated analytics (FORTH)					
Hecuba database system (BSC)					
Qbeast (BSC)					
Visualization tool (AEGIS)					
Resource management and orchestration module (ATOS)					
COMPSs runtime (BSC)					

Validity evaluation: (Technology providers)

Please identify any threats to validity in terms of the measurements indicated in the tables in previous step and indicate actions that need be taken to assure validity.

Evaluation Subjects: (End users)

Which subjects will take part in the experiments? Please ammend the following table as applicable. *In case there is a differentiation between use cases, please indicate adding additional collumns.*

Table 4. Experimental Subjects

Subjects	CAIXA	CRF	TID	OTHER
Quality assurance and control managers	**1**	**-**	**-**	**1**
Data analysts	**3**	**2**	**2**	**2**
Financial administrators	**1**	**-**	**-**	**1**
Human resource officers	**1**	**1**	**1**	**1**
Infrastructure engineers	**1**	**1**	**1**	**2**
IT security personnel	**2**	**-**	**-**	**1**
Decision executives in the production	**-**	**2**	**-**	**1**
Resource planning managers	**-**	**2**	**-**	**1**
Call centre employees	**-**	**-**	**3**	**1**
Network infrastructure engineers	**-**	**1**	**2**	**1**
Market and business analysts	**-**	**1**		**2**

Describe the ways to access and engage the required number of experimental subjects.

Operation (End Users)

Experiment preparation:

Describe the steps taken to prepare subjects, data sets and anything else involved in performing the experiment.

Bibliography

[1] *I-BiDaaS* knowledge repository. https://github.com/ibidaas/knowledge_repository.

[2] AEGIS. https://www.aegis-bigdata.eu/.

[3] Alibaba JStorm. http://jstorm.io/.

[4] Alteryx. https://www.alteryx.com/.

[5] S. W. Ambler. *The Object Primer*. Cambridge University Press, 2004.

[6] K. M. Anderson. Embrace the challenges: software engineering in a big data world. In *Big Data Software Engineering (BIGDSE), 2015 IEEE/ACM 1st International Workshop on*, pages 19–25. IEEE, 2015.

[7] Apache Hadoop. http://hadoop.apache.org/.

[8] Apache HBase. https://hbase.apache.org/.

[9] Apache Hive. https://hive.apache.org/.

[10] Apache Impala. https://impala.apache.org/.

[11] Apache OpenWhisk. https://openwhisk.apache.org/.

[12] Apache Samoa. https://samoa.incubator.apache.org/.

[13] Apache Spark. https://spark.apache.org/.

[14] Apache Spark Streaming. http://spark.apache.org/streaming/.

[15] Apama. ttps://www.softwareag.com/corporate/products/apama_webmethods/analytics/apama_predictive_analytics/default.

[16] Arne J. Berre. Big Data Value Reference model from BDVA (SRIA 4.0) with Machine Learning/AI, 2018. https://www.european-big-data-value-forum.eu/wp-content/uploads/2018/01/Arne-BerreSintef-EBDVF17.pdf.

[17] D. Arruda and N. H. Madhavji. State of requirements engineering research in the context of big data applications. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 307–323. Springer, 2018.

[18] B. D. V. Association. European Big Data Value Strategic Research and Innovation Agenda (SRIA), version 4.0. Technical report, 2017.

[19] R. M. Badia, J. Conejero, C. Diaz, J. Ejarque, D. Lezzi, F. Lordan, C. Ramon-Cortes, and R. Sirvent. COMP Superscalar, an Interoperable Programming Framework. *SoftwareX*, 3:32–36, 2015.

[20] Berkeley Big Data Benchmark. https://amplab.cs.berkeley.edu/benchmark/.

[21] Big Data Europe. https://www.big-data-europe.eu/.

[22] BigDataBench. http://prof.ict.ac.cn/BigDataBench.

[23] Caffe. http://caffe.berkeleyvision.org/.

[24] H. Chen, R. Kazman, S. Haziyev, and O. Hrytsay. Big data system development: An embedded case study with a global outsourcing firm. In *2015 IEEE/ACM 1st International Workshop on Big Data Software Engineering*, pages 44–50, May 2015.

[25] CLASS. https://class-project.eu.

[26] DataBench. https://www.databench.eu.

[27] DATABIO. https://www.databio.eu/en/.

[28] datACRON. http://datacron-project.eu.

[29] Datafloq. Purdue University Achieves Remarkable Results With Big Data. https://datafloq.com/read/purdue-university-achieves-remarkable-results-data/489.

[30] Datameer. https://www.datameer.com/.

[31] T. H. Davenport, P. Barth, and R. Bean. *How'big data'is different*. MIT Sloan Management Review, 2012.

[32] DeepHybridDataCloud. https://deep-hybrid-datacloud.eu/.

[33] Eclipse Paho. https://pypi.org/project/paho-mqtt/.

[34] EMC. Empowering the U.S. Air Force. Toward Joint IT Transformation., 2013. https://www.emc.com/collateral/solution-overview/h13595-so-empowering-the-us-air-force.pdf.

[35] H. Eridaputra, B. Hendradjaya, and W. D. Sunindyo. Modeling the requirements for big data application using goal oriented approach. In *2014 International Conference on Data and Software Engineering (ICODSE)*, pages 1–6, Nov 2014.

[36] FIWARE. http://www.fiware.org.

[37] Flink. https://flink.apache.org/.

[38] Flink Gelly. https://flink.apache.org/news/2015/08/24/introducing-flink-gelly.html.

[39] W. Gao, J. Zhan, and L. Wang. Bigdatabench: A dwarf-based big data and ai benchmark suite. Technical report, 2018.

[40] GE Software. The Case for an Industrial Big Data Platform. https://www.ge.com/digital/sites/default/files/Industrial_Big_Data_Platform.pdf.

[41] Glasgow City Council. Future City Glasgow Evaluation, 2016. http://futurecity.glasgow.gov.uk/reports/12826M_FutureCityGlasgow_Evaluation_Final_v10.0.pdf.

[42] Graph500. https://graph500.org/.

[43] GraphLab-Create 2.1. https://turi.com/.

[44] D. Guinard, V. Trifa, and E. Wilde. A resource oriented architecture for the web of things. In *2010 Internet of Things (IOT)*, pages 1–8, Nov 2010.

[45] HiBench. https://github.com/intel-hadoop/HiBench.

[46] The Hobbit platform. https://project-hobbit.eu/outcomes/hobbit-platform/.

[47] HPCC. http://icl.cs.utk.edu/hpcc.

[48] I-BiDaaS Consortium. D1.1: Project Set-up, 2018. Project report.

[49] I-BiDaaS Consortium. D1.2: Architecture definition, 2018. Project report.

[50] I-BiDaaS Consortium. D8.1: Data Management Plan, 2018. Project report.

[51] Integration Server. https://www.softwareag.com/corporate/products/webmethods_integration/integration/integration_server/default.

[52] IOStack. http://iostack.eu/.

[53] IQPC. Data Analytics for Pharma Development. https://www.bigdatainpharma.com/.

[54] ISO/IEC/IEEE 42010:2011. Systems and software engineering — Architecture description.

[55] T. Ivanov, T. Rabl, M. Poess, A. Queralt, J. Poelman, N. Poggi, and J. Buell. Big data benchmark compendium. In R. Nambiar and M. Poess, editors, *Performance Evaluation and Benchmarking: Traditional to Big Data to Internet of Things*, pages 135–155, Cham, 2016. Springer International Publishing.

[56] Jacques Bughin. Telcos: The untapped promise of big data, 2016. Telcos: Theuntappedpromiseofbigdata.

[57] G. Kotonya and I. Sommerville. *Requirements engineering: processes and techniques*. Wiley Publishing, 1998.

[58] P. B. Kruchten. The 4+1 view model of architecture. *IEEE Software*, 12(6):42–50, Nov 1995.

[59] H. Lippell. *Big Data in the Media and Entertainment Sectors*, pages 245–259. Springer International Publishing, Cham, 2016.

[60] N. H. Madhavji, A. Miranskyy, and K. Kontogiannis. Big picture of big data software engineering: with example research challenges. In *Proceedings of the First International Workshop on BIG Data Software Engineering*, pages 11–14. IEEE Press, 2015.

[61] Maire, S., and Spafford, C. The Data Science Revolution Thats Transforming Aviation, 2017. https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/#325cf1a47f6c.

[62] Marr, B. Big Data In Big Oil: How Shell Uses Analytics To Drive Business Success., 2015. https://www.forbes.com/sites/bernardmarr/2015/05/26/big-data-in-big-oil-how-shell-uses-analytics-to-drive-business-success/#5f9b1192229e.

[63] Marr, B. How Big Data Is Changing Insurance Forever, 2015. https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/#19cbc3a9289b.

[64] MashZone NextGen. https://www.softwareag.com/pl/products/apama_webmethods/mashzone_nextgen/default.

[65] McKinsey Quarterly. Interview Ted Colbert. Data as jet fuel: An interview with Boeings CIO, 2018. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/data-as-jet-fuel-an-interview-with-boeings-cio.

[66] MCloudDaaS. http://booklet.atosresearch.eu/node/1603.

[67] mF2C. http://www.mf2c-project.eu/.

[68] MongoDB. https://www.mongodb.com/.

[69] MQTT protocol. http://mqtt.org/.

[70] S. Nalchigar and E. Yu. Conceptual modeling for business analytics: a framework and potential benefits. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, pages 369–378. IEEE, 2017.

[71] S. Nalchigar and E. Yu. Conceptual modeling for business analytics: A framework and potential benefits. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 01, pages 369–378, July 2017.

[72] NIST Big Data Public Working Group. NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. Technical report, National Institute of Standards and Technology, Special Publication 1500-6r1, 2018.

[73] NIST Big Data Public Working Group: Use Cases Requirements Subgroup. National Institute of Standards and Technology (NIST) Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements. Technical report, National Institute of Standards and Technology, Special Publication 1500-3, 2015.

[74] I. Noorwali, D. Arruda, and N. H. Madhavji. Understanding quality requirements in the context of big data systems. In *Proceedings of the 2Nd International Workshop on BIG Data Software Engineering*, BIGDSE '16, pages 76–79, New York, NY, USA, 2016. ACM.

[75] NTU. NTU Student Dashboard. https://www4.ntu.ac.uk/adq/running_a_course/student-success/ntu-student-dashboard/index.html.

[76] B. Nuseibeh and S. Easterbrook. Requirements engineering: A roadmap. In *Proceedings of the Conference on The Future of Software Engineering*, ICSE '00, pages 35–46, New York, NY, USA, 2000. ACM.

[77] PANACEA. http://projects.laas.fr/panacea-cloud.

[78] PARSEC. http://parsec.cs.princeton.edu.

[79] Pravega. http://pravega.io/.

[80] Randall T. The Smartest Building in the World., 2015. https://www.bloomberg.com/features/2015-the-edge-the-worlds-greenest-building/.

[81] Research and Markets. The Big Data Market: 2018 - 2030 - Opportunities, Challenges, Strategies, Industry Verticals Forecasts, 2018. https://www.researchandmarkets.com/reports/4564313/the-big-data-market-2018-2030-opportunities.

[82] SEA Clouds. http://www.seaclouds-project.eu/.

[83] SecureCloud. https://www.securecloudproject.eu.

[84] Spark GraphX. https://spark.apache.org/graphx/.

[85] Spark SQL. https://spark.apache.org/sql/.

[86] Standard Performance Evaluation Corporation. https://www.spec.org/benchmarks.html.

[87] SPEC CPU 2006. https://www.spec.org/cpu2006.

[88] SPECIAL. https://www.specialprivacy.eu/.

[89] R. Stevens, G. Cattaneo, H. Schwenk, C. Pepato, C. Ostberg, H. Nuria De Lama Sanchez, T. Pariente Lobo, C. Francalanci, B. Pernici, A. Geronazzo, P. Giacomazzi, L. Polidori, A. Jrgen Berre, V. Hoffman, M. Grobelnik, J. Hodson, T. Ivanov, and R. V. Zicari. Databench: Evidence based big data benchmarking to improve business performance. In *The 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD18)*. ACM, 2018.

[90] Tensorflow. https://www.tensorflow.org/.

[91] Terasort. https://mapr.com/resources/terasort-benchmark-comparison-yarn/.

[92] Terracotta. http://www.terracotta.org/.

[93] Toreador project. http://www.toreador-project.eu/.

[94] Transaction Processing Performance Council. http://www.tpc.org/information/benchmarks.asp.

[95] Transforming Transport. https://transformingtransport.eu.

[96] Universal Messaging. https://www.softwareag.com/corporate/products/az/universal_messaging/default.

[97] VMMark. https://www.vmware.com/au/products/vmmark.html.

[98] M. Volk, S. Hart, S. Bosse, and K. Turowski. How much is big data? a classification framework for it projects and technologies. In *22nd Americas Conference on Information Systems, AMCIS 2016*. Association for Information Systems, August, 11-14 2016.

[99] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, and N. Blomberg. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, page "160018".

[100] A. Yasin. A top-down method for performance analysis and counters architecture. In *2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 35–44, March 2014.

[101] Zoe Analytics. http://zoe-analytics.eu/.