DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

D4.1 Data Collection Plan

Abstract

This deliverable presents the plan of the data collection activities executed as part of WP4. The general goal of WP4 is to evaluate the impact of BDT (Big Data Technology) on business performance in key use cases adopting advanced big data and analytics technologies. The work in WP4 is based upon a case study approach. From a methodological standpoint, the case study analysis is inherently in-depth and bottom up, and, as such, it represents a natural complement to the extensive and top down research instruments adopted in WP2. In the framework of the DataBench project, case studies are considered necessary since the relationship between technical choices in organisations and the business KPIs these organisations have established for themselves is complex and difficult to model, as well as to measure. There is a lack of empirical evidence and benchmarks of the business benefits that can be achieved using BDTs, despite the general agreement that they provide a high level of innovation and potential to impact business. Our goal in WP4 is to help fill this gap by providing evidence of the business benefits of BDTs within the general framework of the DataBench project. This deliverable explains the methodology that will be used for the case study analysis, including the classification and selection of the case studies, the recruitment, piloting and analysis phases.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966

Deliverable D4.1	Data Collection Plan
Work package	WP4
Task	4.1
Due date	31/08/2018
Submission date	31/08/2018
Deliverable lead	POLIMI
Version	1.0
Authors	IDC (Gabriella Cattaneo, Mike Glennon and Helena Schwenk) POLIMI (Chiara Francalanci, Marta Pinzone, Barbara Pernici, Angela Geronazzo, Sergio Gusmeroli)
Reviewers	SINTEF (Arne Berre) JSI (Marko Grobelnik) Richard Stevens (IDC)

Keywords

Benchmarking, big data, big data technologies, business performance, economic indicator, European significance, industrial relevance, performance metrics, tool, use cases

Disclaimer

Copyright Notice

Table of Contents

Table of Figures

Table of Tables

Executive Summary

This deliverable explains the methodology that will be used for the case study analysis. The analysis will start from collecting evidence from a variety of sources (this activity is referred to as *desk analysis*), including the business literature, other relevant projects known by the consortium, as well as data collection from relevant businesses that can constitute a reference for the industry. In addition to this, it will analyze the work done by ICT–14 and ICT–15 projects, with a focus on benchmarks that may be available from the activities that these projects have conducted or are conducting to assess the impact of their research and experimentation efforts.

The following task of WP4 is to select, recruit and investigate a sample of case studies based on objective, evidence-based criteria, in order to collect measurements of business KPIs which can be used to extrapolate the benchmarks of industrial significance, which is the final objective. The classification of the case studies is used to identify the main features differentiating them and enabling the definition of a qualitative sample representative of the European industry.

The deliverable discusses how case study recruitment will be performed, by engaging companies in the DataBench cooperative effort. It also discusses how case studies are analysed, explain how interviews, documentation and follow-ups will be performed.

The interaction between the case study analysis and the activities in other WPs is discussed, focusing within WP2 and WP3. The schedule of research activities for the case study analysis is then presented. Conclusions are drawn in Section 4.

1 Introduction and Objectives

This deliverable presents the plan of the data collection activities executed as part of WP4. The general goal of WP4 is to evaluate BDT (Big Data Technology) impact on business performance in key use cases adopting advanced big data and analytics technologies. From a methodological standpoint, the case study analysis is inherently *in-depth* and *bottom up*, and, as such, it represents a natural complement to the extensive and top down research instruments adopted in WP2 (see D2.1 and [1]). In the framework of the DataBench project, case studies are considered necessary as the relationship between technical choices and business KPIs is complex and difficult both to model and to measure. There is a recognized lack of evidence of the business benefits of BDTs, despite the general agreement on their potential business innovation impact [2, 3]. This lack of tangible measures of business KPIs represents an issue for managers who have to make investment decisions, a concern for the policy makers, as well as a call for joint academic and industry research in this direction. Our goal in WP4 is to help fill this gap by providing evidence of the business benefits of BDTs within the general framework of the DataBench project.

As a consequence of their complementarity in research method and common objectives, the relationship between WP2 and WP4 is tight and the selection of industries, companies, and use cases in WP4 is partly driven by the results of the research activities conducted in WP2 and, particularly, by the results of the industrial needs survey. The time frame of data collection activities in WP2 extends over a two-year period and will represent a continuous source of insights influencing the course of research in WP4, with consequent continuous adjustments. However, to minimize these changes and make WP4 activities as efficient as possible, preliminary results from the survey are planned to be delivered at the end of Month 10 (internal deadline that is significantly earlier than the actual deadline in month 12). This deliverable explains how these preliminary results will be exploited to finalize the data collection plan together with the insights from a first pilot case study which will be conducted between Month 9 and Month 10 (see Section 2.5.3). Figure 1 shows the main data collection activities, their timing and their interrelations.

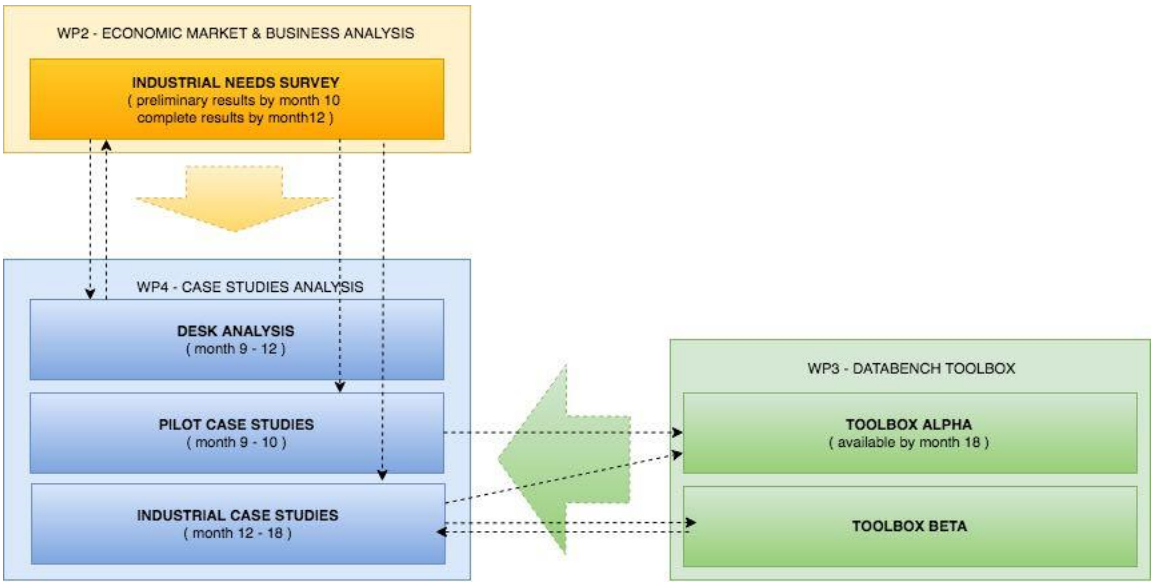

Figure 1 Main Data Collection Activities, Timing and Interrelations

The relationship between WP4 and WP3 is highly synchronized. The DataBench Toolbox designed and implemented in WP3 will represent a key component of the value proposition for companies to take part in the WP4 case study analysis. A first prototype of the Toolbox will be available at the end of month 18 and selected companies (possibly partners of ICT–14 and/or ICT–15 projects) involved in WP4 case study analysis can play the role of beta testers and, at the same time, be allowed to use the DataBench Toolbox to support the selection and execution of technical benchmarking initiatives, according to their needs. Before month 18, case studies can provide important inputs to WP3 in terms of actual business needs, key variables driving the selection of benchmarks, key Q&A (questions and answers) to be incorporated in the Toolbox to support the design of technical benchmarking initiatives.

T4.1 will also collect evidence from a variety of sources (this activity is referred to as *desk analysis*), including the business literature, other relevant projects known by the consortium, as well as data collection from relevant businesses that can constitute a reference for the industry. In addition to this, T4.1 will analyse the work done by ICT–14 and ICT–15 projects, with a focus on benchmarks that may be available from the activities that these projects have conducted or are conducting to assess the impact of their research and experimentation efforts.

2 Data Collection Methodology

Section 2.1 provides an overview of the data collection methodology for our case study analysis. A challenge of our case study analysis is the involvement and commitment of companies and ICT–14/15 projects (this also represents a typical risk of any case study analysis, see also D2.1). In the next section, the actions that will be taken to address this challenge are discussed. The inputs from WP2 and WP3 are explained in detail in Sections 2.2 and 2.3. The approach to the desk analysis is discussed in Section 2.4, providing a preliminary framework that will be applied in our case study analysis and, particularly, in our pilot case to be executed in the month 9 – month 10 time frame. The case study methodology is then explained in Section 2.5.

2.1 Overview of the Methodology

The case study analysis conducted in WP4 will be aimed to the following objectives:

- To gather in-depth knowledge on case studies by understanding not only "what" has been done, but also "how" it has been done, from both a technical and business perspective.
- To unveil the link between technical choices and business KPIs in different case studies (different industries, different technologies, different applications of technology).
- To understand how technical benchmarks can help make better technical choices and thus have a greater impact on business KPIs.
- For companies that have executed a self-assessment with technical benchmarks: to collect data on the effectiveness of their technical choices and the actual impact on measurable business KPIs.
- For companies that have not executed a self-assessment with technical benchmarks: to provide suggestions on how to conduct a self-assessment and, if possible, on how they can use the DataBench Toolbox to support their self-assessment. It should be noted that case studies represent an opportunity to gather information *with* the Toolbox and feed it into the DataBench back-end for various purposes including exploratory experimentation with machine learning, see D3.1.

To reach our goals, we will provide the following benefits for participating companies (our *value proposition*):

- We will provide an introductory session showing our classification of big data and analytics use cases by industry and by technology from WP4 desk analysis and WP2 survey.
- We will help the company to monitor their business KPIs and matching them with benchmarks of progress determined during project activities.
- We will gauge the company against the data (anonymized or average per industry) provided by WP2 survey.
- We will provide access to the alpha and beta versions of the DataBench Toolbox to support self-assessment through technical benchmarking.
- We will provide a copy of the DataBench Handbook at the end of the project.

In exchange for the benefits described above, we expect from companies participating in our case study analysis:

- Participation in IDC survey.
- Interview to achieve a thorough understanding of the big data and analytics pilots/projects that they have done or have plans to do:
  - from a technical standpoint and
  - from a business standpoint.
- Outputs of technical benchmarking outputs, if available, both qualitative and quantitative.
- Measures of business KPIs, if available, both before and after the different big data and analytics pilots/projects that are surveyed during the interview.
- Call for solutions, formalized as business and technical requirements as an input to innovators (BDVA).

2.2 Inputs from WP2

2.2.1 Economic, Market and Business Parameters

WP2 involves performing an economic and market analysis to assess the "European economic significance" of benchmarking tools and performance parameters and validating the business impacts of BDT benchmarks of performance parameters of industrial significance.

The aim of this methodology is to provide an estimate of the potential "footprint" in the European economy of the BDT benchmarks and identify the industry sectors where the perspective BDT business benchmarks will generate the highest potential **economic impact**.

The work will involve the analysis of a number of different economic indicators with BDT market indicators, including indicators of the diffusion of BDT spending in Europe by country and industry sourced from IDC research in order to help identify the economic impact. For example, the high potential economic impact may concern a very large sector (e.g. Manufacturing) where a relatively small BDT business impact if scaled up to the whole sector may result in very large gains for Europe; or it may concern a smaller sector (e.g. Utilities) where a very high BDT impact on business processes could scale up to substantial gains for the overall European economy even if applied to a smaller number of enterprises than in the case of Manufacturing. In particular, economic impact analysis will leverage IDC's research on BDT spending by industry and other indicators on the European Data Market and Data Economy, as detailed in Table 1.

Relevant Data	Source
Gross Value Add by Industry	Eurostat
No. Employees by Industry, Country	Eurostat
No. Companies by Country, Industry	Eurostat
Big Data Spending Guide	IDC
Black Book ICT Spending by Country	IDC

European Data Market Monitoring tool	IDC

Table 1: Economic Analysis Main Sources of Data

A corollary to this work involves assessing the **industrial significance** of the performance parameters to be benchmarked by identifying the BDT business benchmarks that map the actual and emerging needs of industrial users, with the highest potential impact on business processes. This task will involve investigating the main Big Data use cases by industry with the economic relevance of industries as detailed in 2.2.2.

In summary, the economic and market analysis methodology will involve the following steps:

Phase 1

1. Desk research of main public sources (mainly Eurostat and OECD) to select the most relevant economic indicators;
2. Extraction of relevant data from IDC databases and ongoing research on BDT and the European data market;
3. Elaboration of data to identify the most economic significant industries and those with the highest potential impact of Big Data. The potential impact of big data by industry is evaluated from the share and growth of big data by industry for each of the economically significant industries selected.
4. Preliminary classification of main use cases by industry and business process leveraging IDC research (see next chapter).
5. Definition of preliminary benchmarks of economic and industrial significance.

Phase 2

1. Collection of inputs from D.2.3 (analysis of actual and emerging users' needs) and the evaluation of business cases (WP4);
2. Assessment of scalability and feasibility of the main business impact KPIs measured in WP4 to the European dimension;
3. Assessment of potential economic and industrial impacts;
4. Definition of final benchmarks of economic and industrial significance.

2.2.2 The Industrial Needs Survey

As part of the work to measure **industrial significance**, WP2 will investigate the main Big Data (BD) cases implemented by industry. To achieve this goal, the project will first perform a preliminary analysis of uses cases and industries by leveraging pertinent data about end-user investment priorities and the most frequent BDT use cases implemented by industry, measured by IDC's annual survey of IT users by vertical market. Identifying BD use cases by industry this way will ensure that the benchmarks identified respond to actual business needs and acceptance and recognition by the industrial community.

Next, the project will carry out an industrial needs survey of a representative sample of European organizations. Building on a preliminary identification of BDT use cases by industry, the user survey will collect evidence and data about the actual and emerging needs of industrial users. A standard survey will be developed focused on the identification of the BD use cases prioritized in each industry, actual and planned, the KPIs used, why they are used, the potential impacts on business processes and their relevance for business strategies and objectives.

Identifying BD use cases by industry allows the project to correlate Big Data technical performance with potential business process impact. For example, the implementation of predictive analytics in customer churn analysis is a leading use case in the Telecoms and Utilities with clearly relevant business impact on customer loyalty and retention processes. The survey will collect data on the type and relevance of BDT use cases planned/implemented by industry, showing investment priorities and business needs. Based on this data, a preliminary assessment of the BDT performance parameters of industrial significance linked with potential business will be performed. By investigating the actual and emerging industrial requirements, it is also possible to identify the areas of activity without a benchmarking/evaluation scheme which are also industrial priorities and will target them in its activities.

In addition to the main user survey, the project will collect additional evidence about emerging industrial needs by surveying the industrial partners of ICT-14 and 15 H2020 projects, a sample of BDVA company associates and a sample of Big Data users representing the most relevant industries. It is worth noticing that especially projects in ICT-15 have to run processes to understand the impact of the deployment of BDT in their use cases. Furthermore, they had to predict the potential impact and as a consequence they have already defined KPIs that will be measured along the project duration.

The resulting in-depth analysis of user survey data will identify the main use cases classified by industry and their contribution to potential business impacts. This analysis will provide the main parameters of industrial significance for the evaluation of benchmarks by WP4 and 5.

In summary, the methodology for the industrial needs survey will incorporate the following main steps:

1. Definition of the scope of the analysis: type of users, type of needs to be investigated, type of use cases;
2. Development of the survey sample and questionnaire;
3. Implementation of the survey;
4. Elaboration of results;
5. In-depth analysis of results;
6. Production of deliverable.

2.3 Inputs from WP3

The work in WP3 is related to the conception and implementation of the DataBench Toolbox. As explained in D3.1 [6], the Toolbox aims to provide support to industries and big data practitioners for reusing big data benchmarks in a unified manner. Users will be able to perform selection, download and execution of the desired benchmarking frameworks, as well as give feedback to the Toolbox about their benchmarking execution results to be able to get access to more standardized and comparable technical metrics and hints about their potential business value. It is precisely in this last aspect, the business perspective, where WP3 will rely on the work done in the scope of WP4.

D3.1 summarized the status of our big data benchmarking tools survey performed mainly in the scope of WP1. The survey includes several dimensions:

- On the one hand, we collected the different benchmarking tools ordered from the oldest to the newest. This allows us to see in a single table a timeline of the evolution of the benchmarks.
- On the other hand, the table provides a set of characteristics aligned heavily to the BDVA Reference Model, but also to the current features offered by the benchmarks analysed:
    - Vertical sectors (Business, Transport, Manufacturing, Energy, Bioinformatics, Health, Telecom, Finance) along with other categories related to verticals but not belonging to the typical industry classification, such as Social Media, General micro-benchmarks or standard benchmarks.
    - Data types (metadata, graph data, text -NLP, web-, image and audio, spatio-temporal, structured -BI-).
    - Data value chain (data storage, data management, data processing -batch, streaming and interactive-, industrial analytics -descriptive, diagnostic, predictive, prescriptive-, visualization, security and privacy, communications and connectivity).

The table provides an overview of the support of the existing benchmarks for all these aspects, therefore providing a powerful insight on what these benchmarks can offer from the technical point of view. This technical overview is a very important input for WP4. WP1 will continue working on this survey and provide a complete view by M12.

On the other hand, the Toolbox, in collaboration with WP1, will take the technical metrics provided by the different benchmarks, map and unify them in order to provide comparable results. This is in principle independent of the vertical domains and case studies, but nevertheless understanding the set of technical metrics the project is dealing with is an important input for WP4.

2.4 Desk Analysis

In this section, the data collection methodology will be explained. The idea is to collect information on big data and analytics use cases from the academic and industry literature as a basis for a first understanding of the link between technical benchmarks and business KPIs. From a preliminary inspection of the scientific literature and the BDV Reference Model (www.bdva.eu and [5]) two perspectives emerge as commonly used to categorize use cases:

- *business process* and
- *data features*.

The type of *technical benchmark* that is used to support BDT technical choices is seldom mentioned without a clear explanation of the decision process that has led to the use of that specific benchmark. Therefore, a fundamental objective of the desk analysis is to find the dimensions of business processes and data features that are useful to select the most appropriate technical benchmark that addresses key technical choices in that use case that can affect business KPIs.

Our preliminary desk analysis has considered 20 use cases selected to cover all BDV data features according to the BDV Reference Model from the sources for desk analysis reported in Annex (1). The full list of the 20 use cases is reported in Appendix (2). We have used these use cases to infer dimensions for business processes, data features and technical benchmarks (the dimensions for technical benchmarks is also largely based on a

preliminary survey of the technical benchmarks selected in D3.1). The full desk analysis will map all use cases on these dimensions to validate and possibly extend them. The next sections report the results of this preliminary desk analysis.

2.4.1 The Three Macro-Areas in Detail

Business process

In this section, the star diagram summarizing the dimensions of business processes is shown. These dimensions are listed in the following.

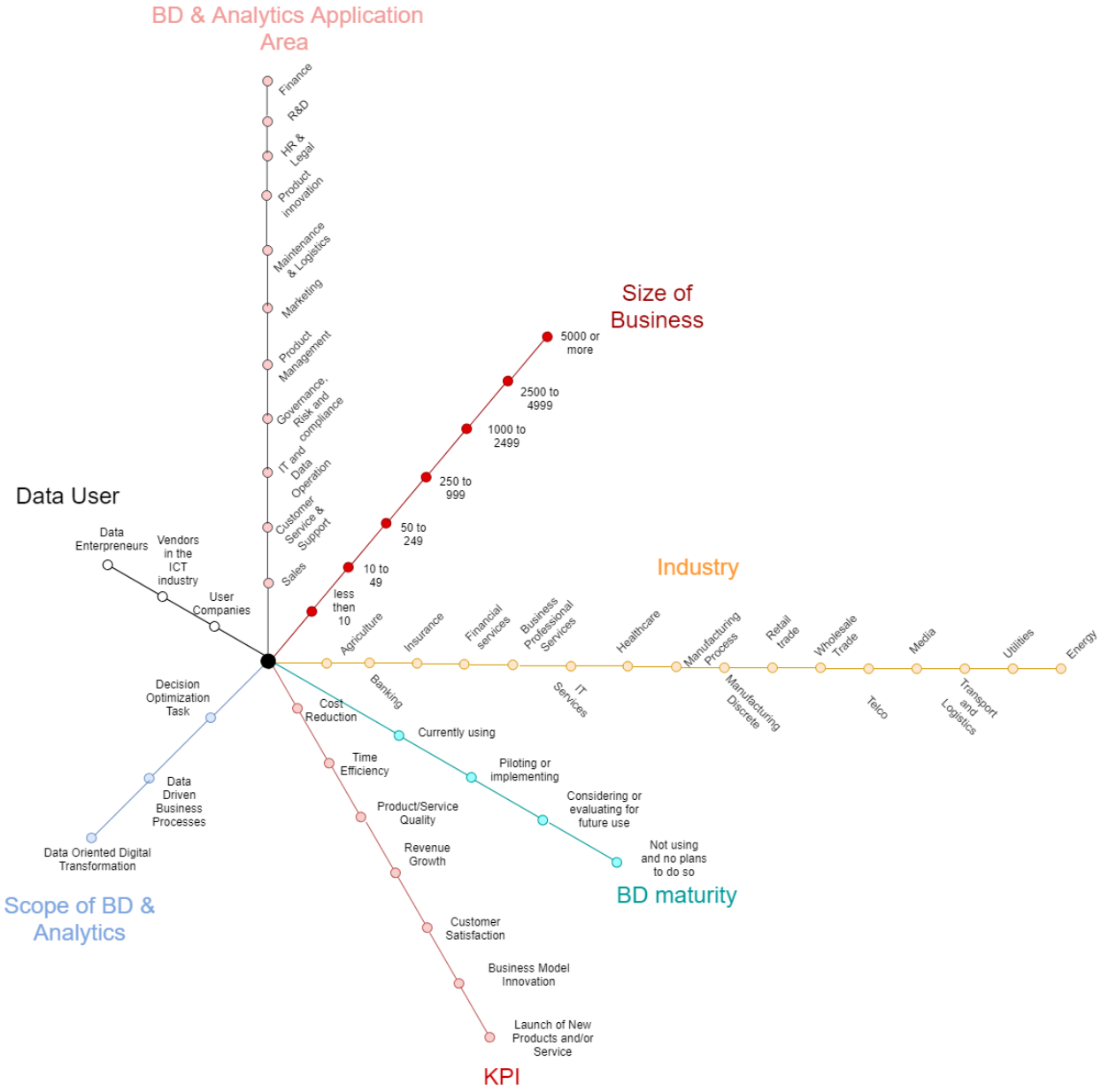

Figure 2 Business Process Dimensions – Star Diagram

- Industry: *Agriculture, Banking, Insurance, Financial Services, Business Professional Services, IT Services, Healthcare, Manufacturing Process, Manufacturing Discrete,*

Retail Trade, Wholesale Trade, Telco (Telecommunications), Media, Transport and Logistics, Utilities, Energy.

The selection of the industries has been done through an accurate work of integration between the industries considered in BDVA (BioTech – AgriFood, Transport – Mobility, Health – Ageing, Manufacturing, Energy, Smart Cities, Earth – Obs and GeO, Telecom, Retail, Finance, others), the Hobbit's (https://project-hobbit.eu/) set of industries (Construction, Public Administration, Education, Energy, Professional Services, Utilities, Digital marketing, Retail and Wholesale, Financial Services, Telecom-IT- Media, Healthcare, Food Agriculture, Manufacturing, Transport Services) and IDC's standard industries classification (Finance, Manufacturing and Resources, Distribution and Services, Infrastructure, Public Sector, Consumer) IDC's industry classification is also aligned with Eurostat NACE II codes and this is important to use comparable data from IDC databases and surveys as well as other public sources.

As for every project, DataBench needs to balance resources and scope and make hard decisions to maximise its efficiency. In the case of the selection of sectors to be analysed, even if in principle all of them are relevant, we made the decision to exclude the following sectors:

- The public sector (Government and Education) was excluded for the following reasons:
  - DataBench primary focus is on industry: Government and Education are not industry. The dynamics of the no-profit public sector are very different from the private sector. "Business" KPIs are not comparable with the other sectors and would require additional, different analytical work.
  - It is difficult to provide comparable data for Government because the number of government agencies is counted differently from industries and varies wildly by country. To some extent this is also true for Education.
  - Healthcare was instead retained because of its relevance for data-driven innovation and its organization closer to private industry dynamics. Besides, healthcare organizations are included in industry statistics.

The Construction industry was not included because of its fragmentation and low level of maturity in the adoption of data-driven innovation. According to the European Data Market Monitor study[1], Construction is the industry with the smallest share of data market value (0.5% in 2017 versus for example 11% of Retail) and also the lowest share of data market investment on total ICT spending (7.3% versus a total EU average of 10.4%).

- BD Maturity: *Currently using, piloting or implementing, considering or evaluating for future use, not using and no plans to do so.*

  The BD Maturity is the level of maturity of the utilization of BD & Analytics of an organization, and particularly in the context of the use case. Companies that are already using big data analysis for this dimension will assume the value "currently using". Companies, however, that do not perform big data analytics yet but who are

1 Updating the European Data Market Monitoring Tool, 1st Report on Facts and Figures, February 2018, http://datalandscape.eu/

carrying out tests in order to be able to use them soon will assume the "piloting or implementing" value for this dimension, and so on.

- KPI: *Cost Reduction, Time Efficiency, Product/Service Quality, Revenue Growth, Churn, Customer Satisfaction, Business Model Innovation, Launch of New Products and/or Service.*
  These KPIs have been selected as preliminary general-purpose set of high-level dimensions of benefits. Context-specific KPIs may emerge during the case study analysis, which would be more closely related to the characteristics of different initiatives. These additional KPIs will be considered and our initial list will be extended, if needed. It should be noted that the KPIs in our preliminary list are high level and are likely to represent a good generalization of a number of context-specific measures.

  The business model innovation KPI has been separated from the product innovation KPI since the first indicates a more structural organizational change that has a broader impact.

- Scope of BD & Analytics: *Decision Optimization Task, Data Driven Business Processes, Data Oriented Digital Transformation.*

  This dimension measures the main purpose of the big data and analytics that will be performed in the context of the use case. The “decision optimization task” value of this dimension indicates that the objective of the analysis is to optimize one or more decisions concerning the type of business of that specific use case. “Data-driven business value” indicates that the purpose of the use case is to create a process that is data driven, therefore business decisions will be made basing on the results of the studies performed on these large amounts of data. The “data-driven digital transformation” is a longer-term digital transformation of the company.

- Data User: *Data Entrepreneurs, Vendors in the ICT Industry, User Companies.*

  The Data User is the main actor of the use case. It can be a “data entrepreneur”, when an entrepreneur owns a data-driven company.

- BD & Analytics Application Area: *Sales, Customer Service & Support, IT and Data Operation, Governance Risk and Compliance, Product Management, Marketing, Maintenance & Logistics, Product Innovation, HR & Legal, R&D, Finance.*

- Size of Business: *5000 or more, 2500 to 4999, 1000 to 2499, 250 to 999, 50 to 249, 10 to 49, less than 10.*

Data features

In this section, the star diagram concerning the data features of the use case is discussed. The objective of this star diagram is to collect information about data storage, machine learning approach, type of data involved, dataset size, datasource, type of analytics, processing paradigm and performance metrics measured.

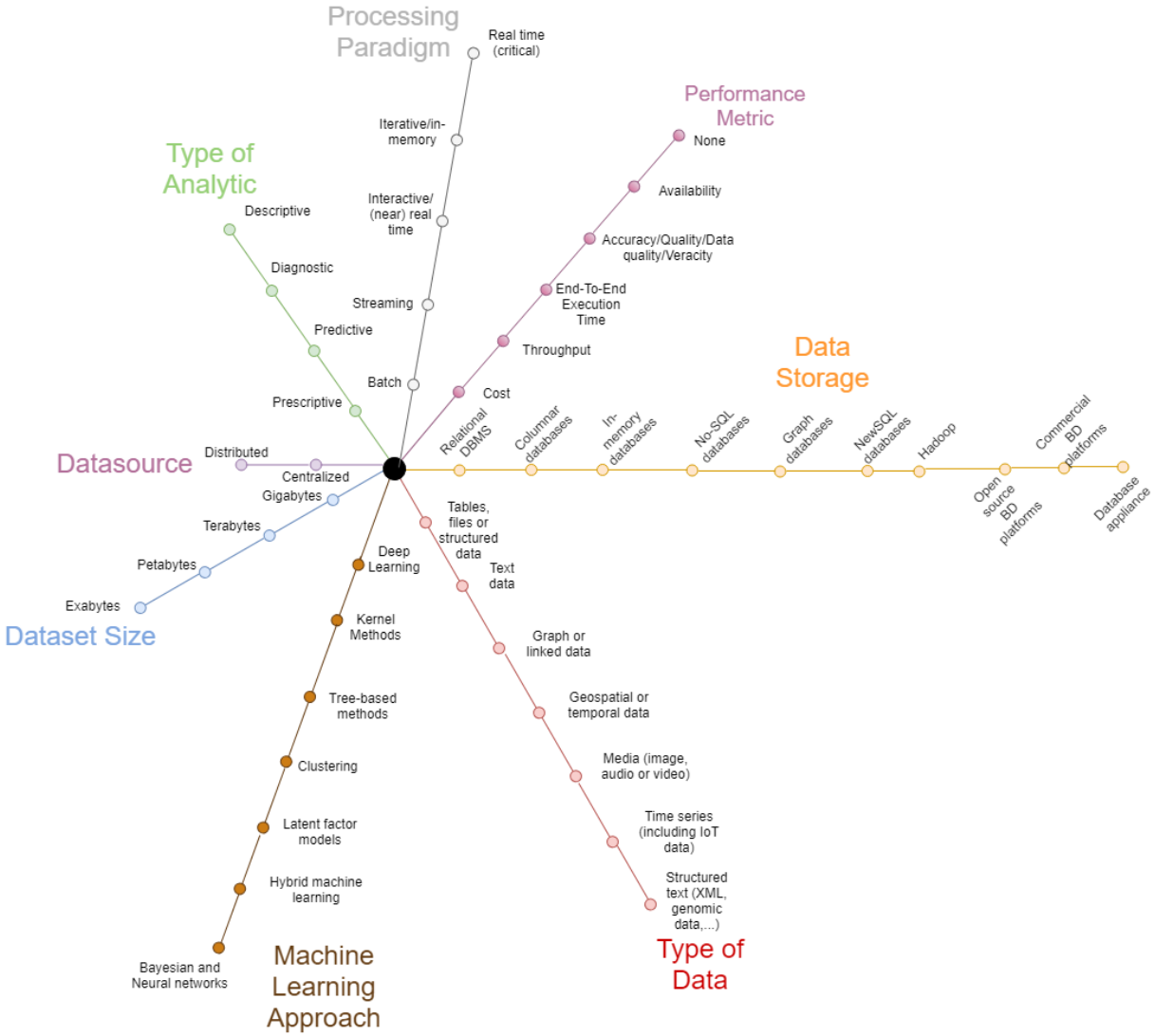

Figure 3 Data Features Dimensions – Star Diagram

- Data Storage**:** *Relational DBMS, Columnar databases, In-memory databases, No-SQL databases, Graph databases, NewSQL databases, Hadoop, Open source BD platforms, Commercial BD platforms, database appliance, Industrial Data Platforms for data sharing and/or data exchange (i.e. International Data Space(s), https://www.internationaldataspaces.org/en/*

  With the Data Storage dimension, we aim to identify the way in which data are stored in the context of the use case examined. These dimensions heavily draw from the BDV Reference model [5], with a few extensions from our preliminary desk analysis focused on the data size and performance metric dimensions.

- Type of Data: *Tables, files or structured data, Text data, Graph or linked data, Geospatial or temporal data, Media (image, audio or video), Time series (including IoT data), Structured text (XML, genomic data, etc.).*

Data can be structured, such as tables or files, or non-structured, such as text and media.

- Machine Learning Approach**:** *Deep Learning/AI, Kernel Methods, Tree-based methods, Clustering, Latent factor models, Hybrid machine learning, Bayesian and Neural Networks.*

  These techniques have been selected bottom up from case studies and represent the techniques that are explicitly mentioned among a larger set of possible machine learning techniques.

- Dataset Size: *Gigabytes, Terabytes, Petabytes, Exabytes.*

- Datasource**:** *Distributed, Centralized.*

  The data sources are centralized when they are located, stored, and maintained in a single location.

- Type of Analytics: *Descriptive, Diagnostic, Predictive, Prescriptive.*

  This represents a common categorization of analytic tasks.

- Processing Paradigm**:** *Batch, Streaming, Interactive/(near) real time, Iterative/in-memory, Real time (critical).*

  This dimension refers to the timing of data processing.

- Performance Metric**:** *Cost, Throughput, End-To-End Execution Time, Accuracy/Quality/Data Quality/ Veracity, Availability, None.*

  The performance metric represents a key measurable data characteristic in the use case.

Technical benchmarks

The star diagram that will be used to classify the technical benchmarks is shown below. As noted before, this diagram heavily draws from the classification of technical benchmarks provided in D3.1 and will be used as a basis to associate appropriate technical benchmarks with use cases, consistent with the business and data characteristics of different use cases.

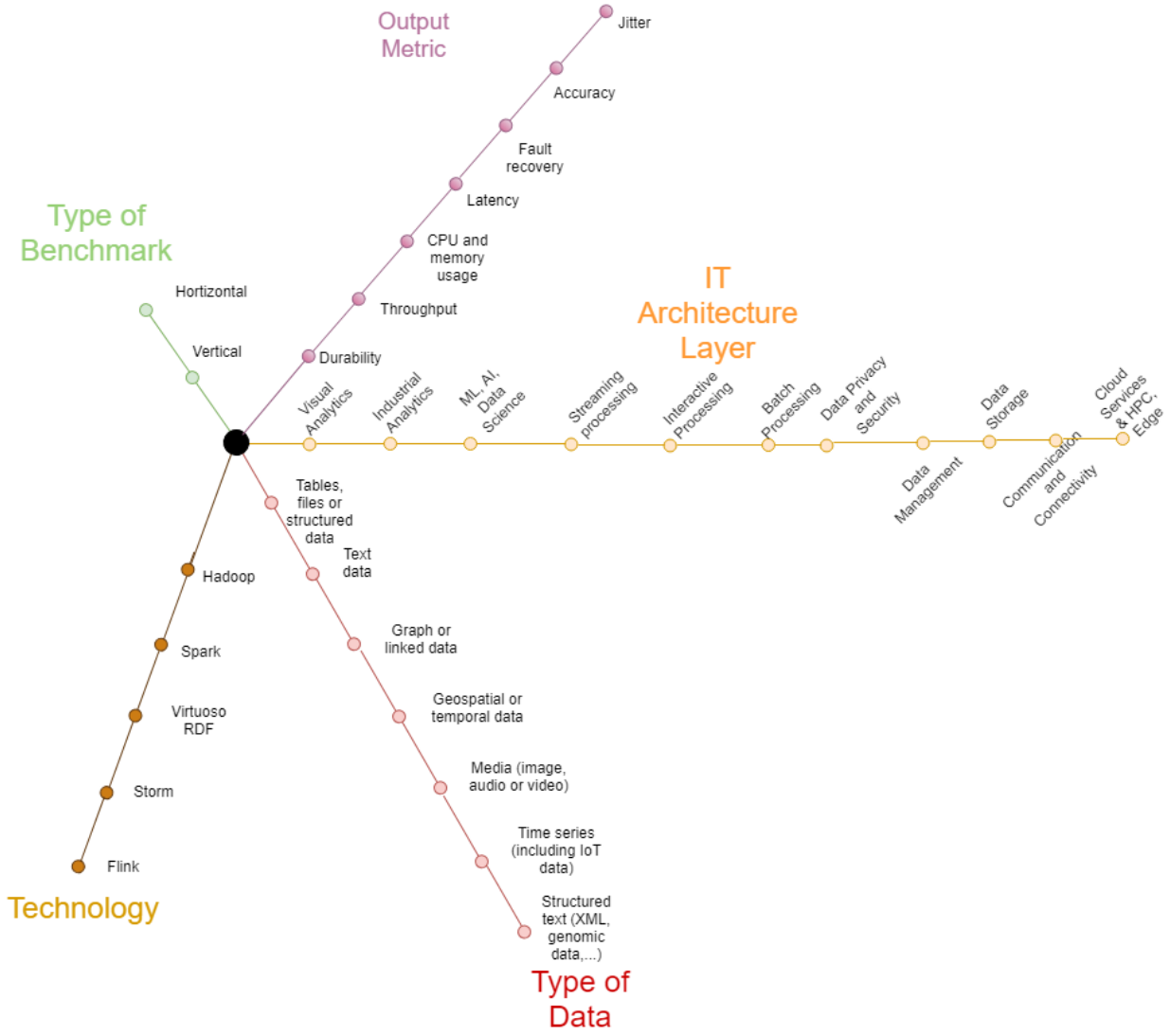

Figure 4 Technical Benchmarks Dimensions – Star Diagram

- Type of Benchmark: *Horizontal, Vertical.*

  The "horizontal" benchmarks test end-to-end big data and analytics architectural layers, while "vertical" benchmarks test only one or two architectural layers.

- Output Metric: *Durability, Throughput, CPU and Memory Utilisation, Latency, Fault Recovery, Accuracy, Jitter.*

  See D3.1 and reference technical manuals of benchmarks for a definition of these features.

- IT Architecture Layer: *Visual Analytics, Industrial Analytics, ML, AI, Data Science, Streaming Processing, Interactive Processing, Batch Processing, Data Privacy and Security, Data Management, Data Storage, Communication and Connectivity, Cloud Services & HPC, Edge.*

- Type of Data: *Tables, files or structured data, Text data, Graph or linked data, Geospatial or temporal data, Media (image, audio or video), Time series (including IoT data), Structured text (XML, genomic data, etc.).*

  See the data features dimension.

- Technology: *Virtuoso RDF, Hadoop, Spark, Storm, Flink…*

  This represent a preliminary list of big data technologies that will be completed bottom up based on the case study desk analysis.

2.5 Case Study Methodology

A key task of WP4 is to select, recruit and investigate a sample of case studies based on objective, evidence-based criteria, in order to collect measurements of business KPIs which can be used to extrapolate the benchmarks of industrial significance, which is the final objective.
This task faces several challenges and risks including:

- Difficulty to identify and select case studies of actual implementation of Big data technologies according to the parameters described in the previous paragraphs;
- Need to provide a value proposition for the interviewees to make them willing to collaborate with the interview process and share their data;
- Need to define a sample of case studies reflecting the main features of the European industry, the most relevant and innovative industries, the business processes where big data can indicatively provide the highest value added, to provide significant results of broader significance for the European economy.

To manage these challenges the case study methodology is articulated in 5 main steps which will be described in the following paragraphs:

- Classification and selection of a long list of potential case study candidates;
- Recruitment of the candidates to convince them to agree to be a case study;
- Implementation of the case studies (starting with a pilot case to test the methodology);
- Analysis of the case study results and write-up of the individual case, including feedback from the industry;
- Longitudinal analysis of results and extrapolation of results to the European industry, eliminating idiosyncratic factors correlated with individual experiences while identifying common factors leading to the definition of benchmarks of European significance.

2.5.1 Classification and Selection of Case Studies

The classification of the case studies is used to identify the main features differentiating them and enabling the definition of a qualitative sample representative of the European industry. From this point of view, and building on the analysis carried out in WP2, we have selected two main parameters as the key criteria for the identification of the sample:

- Type of industry (indicatively 5 different industries);
- Type of use case (indicatively 4-5 main typologies of use cases).

The selection of industries will be based on their share of value added and their intensity of use of BDA so that the case studies represent a relevant share of the EU industry.

The selection of use cases will be based on the results of the user survey carried out by WP2 and the desk research and will focus on the most frequent use cases affecting relevant business processes with potentially the highest impacts. By highest impacts we consider use cases applicable to more than one industry and with potential relevant consequences on the bottom line and competitiveness of a high number of companies. Niche applications specific to individual industries or limited to small groups of industries can be very interesting but are not a priority for the objectives of this project.

The target sample of case studies is presented in the matrix below (Table 2). It should be noticed that it will also be possible for a single case study to cover more than one use case and/or more than one industry (as there are value chains which may associate more than one industry as defined by statistical criteria: for example, the agri-food value chain links together Agriculture, Manufacturing and Wholesale/Retail). Our case studies are focused on innovation and therefore may show how the traditional value chains are evolving, but we will always need to maintain a link with the standard statistics in order to be able to use Eurostat data to extrapolate results.

In addition to the two main parameters, the classification of case studies will also take into account several other variables already discussed in the paragraph 2.4.1:

- Type of data user: balanced distribution between data entrepreneurs, vendors, users;
- Scope of BDA: balanced distribution between decision optimization tasks, data driven business processes, data oriented digital transformation;
- BD maturity: our case studies will focus only on organization currently using BDA, or at best piloting/implementing (if they can provide inputs on business KPIs measurements);
- Company size: indicatively 40% large companies (250 to 999 employees); 40% very large companies (over 1000 employees); 20% small companies (up to 249 employees). It will be much more difficult to find small companies users of BDAs and fitting our other criteria, as currently the diffusion of data innovation is prevalent in larger companies.
- BDA application area: balanced distribution of the main application areas depending on the typology of use cases selected (privileging the most frequent areas with potentially the largest impacts).
- Geography: case studies should come from all of Europe, with a balance between North-South, East-West regions. The penetration of BDA is lower in Eastern Europe so we don't expect to achieve the same number of case studies from that region as from Western Europe but it should not be overlooked.

The study team will strive to collect a sample of case study corresponding to all these ideal criteria, even though it will not be possible to achieve all of them. The closest we can get to this ideal sample, the more it will be representative of the European industry. However, the first and most important objective is to collect evidence about business KPIs of BDA so this is the main defining selection criterion, while the others must take second place.

	Use Case 1	Use Case 2	Use Case 3	Use Case 4	Use Case 5	Total
Industry 1	1	1	1	1	1	5
Industry 2	1	1	1	1	1	5
Industry 3	1	1	1	1	1	5
Industry 4	1	1	1	1	1	5
Industry 5	1	1	1	1	1	5
Total	5	5	5	5	5	25

Table 2: Target Sample of Case Studies

In practice, the study team will collect information about organizations implementing BDA at such a level of maturity to be able to measure impacts, classify them based on the parameters identified above, and build a long list of potential case studies, prioritized on the basis of their correspondence to our ideal criteria.

The main sources of the long list of case studies will be:

- The pilot cases run by industrial partners of ICT-14 and 15 projects and others funded by H2020, for example the IoT Large-Scale Pilots;
- The BDVA community through the common events and communication to members;
- The respondents to the user survey run by WP2 who will accept to be contacted (there will be a specific question on this, otherwise the respondents are anonymous);
- The industry contacts of the partners, for example end users known by IDC's Big data and Digital Transformation research practices;
- Desk research on the web and other sources of leading BDA implementations.

Our target will be a long list of approximately 50-75 potential case studies. The list will be a living document regularly updated during the project (since the period of implementation of the case studies will be several months it will be possible to add cases).

We will also publish a call for volunteer case studies on the DataBench website and promote this option in the various events attended by partners where there may be interested organizations.

2.5.2 Recruitment Phase

The recruitment phase is critical since we will need to ensure the active collaboration of the case studies organizations and their trust to share their data with us.

Once developed the long list, the study team will contact the candidates through email and send them an invitation which will include:

- A clear explanation of the request
- What will be expected from them
- What they will gain from their participation to the case study

- Links to further information

This will probably need to be done in two steps, a shorter initial communication focused on the "value proposition" for the organization and, if they respond and are interested, a longer and more specific communication. This will likely be followed by a call where a member of the study team will explain more precisely the request. This will include investigating if other organizations partnering with the first contacted should be involved to complete the case study.

The study team will develop a dossier with the key information to be sent to all potential candidates. This will also be available for download from the website.

In order to convince them to participate, the DataBench value proposition will include:

- Support to measure and evaluate the business impacts of their BDA innovation, using our KPIs methodology;
- Information on best practice BDA technical benchmarking;
- Access to the DataBench tool;
- Provision of advice and consulting on BDA best practice implementations (even though we should not promise ad hoc strategic consulting or technical implementation support for free, which would be misleading);
- Participation in the DataBench community;
- Full respect of confidentiality as requested by the organization, using data only in aggregated or anonymised form if so requested;
- Visibility as good practice or pioneer case study, if they are interested;
- Access to a short report with the main results of case studies to enable business benchmarking (respecting confidentiality as requested by the participants).

Once the organization accepts to participate, we will enter into the case study implementation and analysis phase which is described in the following paragraph.

2.5.3 Piloting

The Piloting phase aims at testing the methodology in a case study with a twofold goal:

1. evaluating BDT impact on business performance, consistent with the general objectives of WP4, and
2. understanding how the selected company ties their technical choices with business KPIs to gather practical insights and requirements for the design of the Toolbox. The final aim is to derive correlations between technical/business KPIs to be included in the Handbook (D4.4), to drive enterprises towards the choice of the best BDT benchmark through the DataBench Toolbox and, thus, support technical choices that can enable or maximize business benefits.

Technical KPIs are an input from WP3. They are represented as entity and indicate which Technical Benchmarks can measure with each KPI, in which domain(s) and based on which data types (see D3.1). Three main common steps have been identified in a generic Business Process and for each of them a set of typical technical KPIs have been defined. Examples are given below:

- **Pre-processing phase**. According to the BDVA Reference Model, this includes the capturing, extraction, cleaning and all other processing steps necessary to feed the following phase. Usually this includes Time-related (preparation, processing times),

Cost-related (CPU, memory, storage but also human effort), Quality-related (data gaps, data reliability, data duplications) technical KPIs.

- **Processing phase**. This phase is typically performed by intense processing (data in motion and/or data at rest) either performed at the edge, in the cloud or even in a HPC specialized facility. This is the core of many Technical Benchmarks, including Time/Cost related KPIs (processing time, costs) as well as Quality-related KPIs (e.g. accuracy of an estimation / forecast; goodness of a plan or schedule). Architecture and configuration (e.g. number of nodes, of CPUs, processing-storage power) are also variables which could influence the choice of technical benchmarks.
- **Post-processing phase**. This phase is more heterogeneous than the previous two and includes more quantitative (e.g. querying time/cost; updating time/cost) and more qualitative KPIs (user experience/satisfaction, likeability, friendliness, decisional support).

From the Industrial Cases (and their reference projects), **business KPIs** will be also defined. For instance, in the Manufacturing Industry domain, leveraging on the BOOST 4.0 Lighthouse project, the five application domains (Smart Digital Engineering; Smart Production Planning and Management; Smart Operations and Digital Workplaces; Smart Connected Production; Smart Maintenance and Services) will define each 3-5 major Business (generic) KPIs which could be considered as archetypes. For instance, Time-to-Market (digital engineering); Total Cost of Ownership (production planning); Mean Absolute Percentage Error (operations); Value Chain Carbon Footprint (connected production); Mean Time Between Failures or Mean Time To Repair (maintenance).

The piloting phase will be implemented in one company which is a target for DataBench in terms of industry, size and geographical location (see D2.1) and can be considered a leader in the exploitation of BDTs. In the pilot case, the DataBench researcher responsible for the analysis, called DataBench Ambassador will proceed to the application of the Case Study methodology (see Section 2.5.4), with a specific focus on:

- Modelling of the use case according to its three major phases (pre-processing, processing and post-processing) and relevant technical KPIs;
- Defining Business KPIs at the Strategic, Tactical and Operational levels (see, for example, the ECOGRAI method[2]);
- Specifying in the process model the activities responsible for measuring both Technical and Business KPIs;
- Discussing with the company (whenever possible) alternative technical architectures and key technical choices, to help understanding the criteria for the selection of the most relevant technical benchmark;
- Model possible correlations between technical and business KPIs for the specific case study.

Typical qualitative correlations will establish links, weak or strong, between technical and business KPIs for a certain industry. For instance, in a typical new product Development process, the correlation between Processing Time and Time-to-Market is WEAK; while in a typical predictive Maintenance process, the correlation between Failure Prediction Accuracy and Mean Time Between Failure is VERY STRONG.

2 https://link.springer.com/chapter/10.1007/978-0-387-34847-6_39

If needed/possible, the job roles involved in the use case process, their activities and decisions, and key competencies will be also identified and analysed.

Furthermore, if needed/possible, this bottom-up process will be then completed with a maturity assessment (§2.5.5) and with the top-down survey performed in WP2, in order to provide use cases with added-value feedback regarding their positioning with respect to their full exploitation of BDT business benefits for their business.

The pilot company will be granted access to the DataBench ToolBox (from Month 18), to obtain suggestions of the most suitable Technical Benchmarking (through the Toolbox), to maximize the business benefits derived from the introduction of BDT into their business.

An important outcome of the pilot phase will be **a set of recommendations** on how to conduct the analysis of the other case study, with respect to the initial blueprint provided in the next section.

2.5.4 Case Study Analysis

The case study analysis will start with an on-site visit or a call with the DataBeanch team. The call/visit aims at collecting information and details on the case study by investigating the current solution adopted, the technical challenges, the adoption of technical benchmarking solutions and the expected short-term and long-term benefits. The call/visit will involve company's managers with business as well as technical background to provide an as broad as possible picture of the case study. During the onsite visit the DataBench team will collect information about:

- the company, its business and its IT infrastructure;
- the case study goals;
- the case study issues and challenges regarding the data processing and IT infrastructure;
- expected benefits enabled by the case study;
- relevant business KPIs.

In order to get an overview of the company IT infrastructure, the DataBench team will collect blueprints of the IT infrastructure, schemas of the data sources and characteristics of the data streams. In particular, the analysis will focus on the parts of the IT infrastructure relevant to the case study. If the case study is already implemented by the company, details about the current solution will be investigated with a specific focus on its issues and strengths.

The team will inquire the structure of the data streams useful to the case study with a particular emphasis on data quality metrics and data fusion issues derived by the management of heterogeneous data sources. Furthermore, the team will examine data processing and analytics characteristics required by the case study, with a specific focus on the data characteristics, such as volatility and veracity, and on the requirements and solutions evaluated by the company for data storage, processing, visualization and analytics.

The team will investigate whether the company has adopted technical benchmarks in the specific case study and/or other case studies and will get insights about technical performance metrics evaluated by the company to ease the selection of an appropriate benchmark.

Moreover, the team will collect information about the business benefits enabled by the case study, in a short- and in a long-term scenario, and with a specific focus on measurable business KPIs.

After the first call/visit, the DataBench team will continue to work with the company, possibly with additional calls or visits, to collect all the information needed to complete the case study analysis.

A tentative template useful to summarize the collected information and to support the case study analysis is provided in Appendix (1). The template is drawn from the BDVA template adopted for the ICT 14-15 projects and extended to include details about technical benchmarks adopted, business benefits and KPIs. The template will be extended and refined with reference to each specific case study.

The case study analysis will start from the information collected from the company and will possibly focus on:

- discussing issues and challenges of the case study with reference to its industry;
- suggesting technical benchmarks useful to support the development of a new solution or the improvement of the current approach;
- highlighting Big Data specific challenges in a short- and a long-term scenario with reference to the case study and to its industry;
- investigating issues and challenges in generalizing the case study.

The analysis contributions will be refined and possibly extended during the pilot case study. Furthermore, the outcome of the pilot case will provide insights to finalize the interview, the interaction process and the materials to be expected/requested from the company. Moreover, further possible areas of analysis will be evaluated with reference to each specific case study.

2.5.5 Maturity Evaluation

As reported in the paragraph above, when possible, the analysis of the single use case will be enriched by performing a more holistic digital maturity assessment to gather contextual information that may influence the exploitation of BDT and the resulting business benefits.

To this end, DataBench will use the Industry 4.0 Test, which is a digital maturity assessment questionnaire developed by Politecnico di Milano and currently supported by Confindustria Italia. It is suitable both for large enterprises and SMEs.

The Test allows researchers to analyze the current state of business practices and capabilities related to the 8 main process areas that contribute to the creation of value within a company: 1. Design and Engineering; 2. Production Management; 3. Quality Management; 4. Maintenance Management; 5. Logistics Management; 6. Supply Chain Management; 7. Human Resources Management; 8. Marketing, Sales and Customer Care. Furthermore, two orthogonal areas related to the overall digital strategy and the “smart” product (or service) of the company are considered in the evaluation.

The Industry 4.0 Test allows researchers to assess each process area against 4 different dimensions of analysis, providing a detailed assessment of 1. execution and 2. control of the process, 3. use of digital technologies and 4. organizational and people-related aspects.

The company's capabilities are measured along 5 levels of maturity, which are based on the well-known CMMI (Capability Maturity Model Integration) framework. The first level of

maturity is characterized by poorly controlled and reactively managed processes, while the fifth level is characterized by fully digitally-oriented processes.

Within DataBench, whenever feasible, the Industry 4.0 Test will be used to collect data from one (or more) company manager, who will be assisted by the DataBench Ambassador. A tour of the company site will allow an initial impression to be formed of how the processes work. Then, about 1-3 hours will be needed to complete the Test, depending on the number of processes to assess and the number of key informants to interview.

After the interviews, the DataBench Ambassador will analyse the information collected and will provide the company with a synthetic report summarizing its Digital Maturity. Moreover, whenever feasible, a list of strengths and weaknesses will be highlighted. Based on the identified strengths and weaknesses, opportunities will be identified and discussed with the company manager(s), to pinpoint concrete actions to improve the company's digital maturity and move forward toward the data-economy.

The results of the maturity assessments and opportunity identification will be then used to develop a "big-data migration blueprint" highlighting, when possible, common recommendations to help companies make the most out of BDT. In DataBench, two main complementary socio-technical perspectives will be considered in shaping the "big-data migration blueprint":

1. **A Big Data Platform Migration Pathway**, which will provide a flexible methodology to analyse readiness of the enterprise to migrate the 5 BDVA Technical Challenges towards a more mature positioning against Data Management; Data Protection; Data Architecture; Data Analytics and Data Visualisation parameters (Figure 1).
2. **A Big Data People Migration Pathway**. In this case the focus will be on organizational aspects and people competencies, in order to give an overarching view of what constitutes the contemporary professions around Big Data, helping HR and other managers to better organize for big data, to search for better recruitments and develop human capital towards the data-economy.

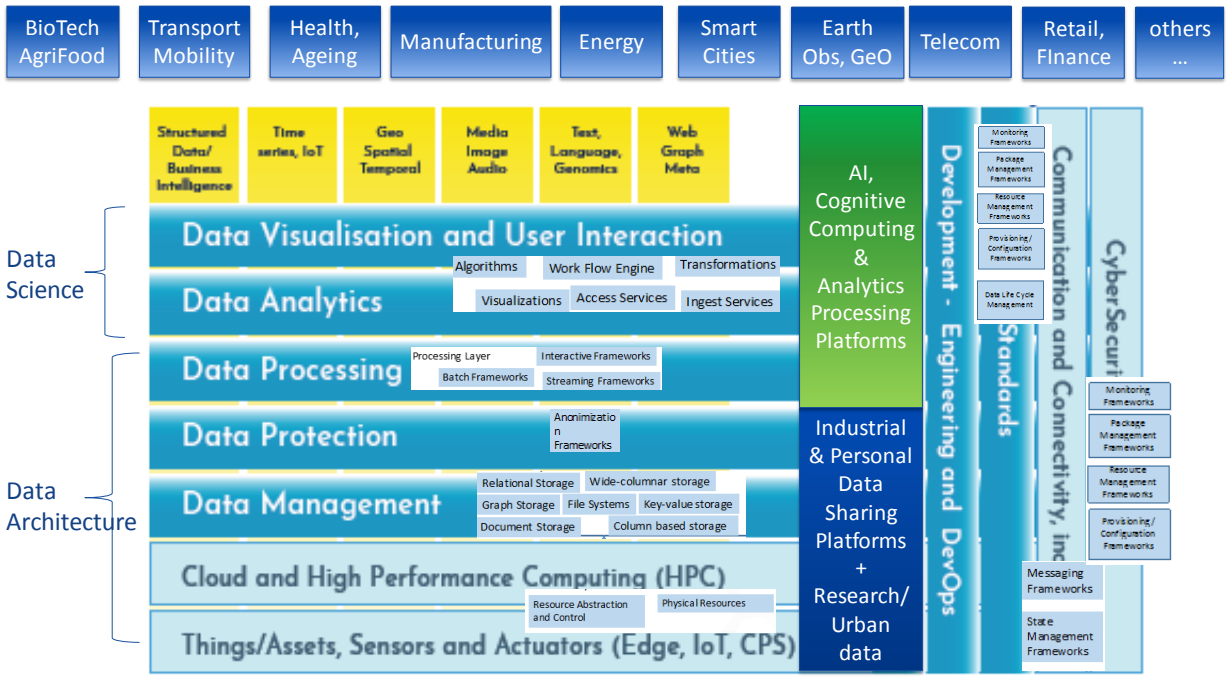

Figure 5 BDV Reference Model – with Industrial Data Platforms and AI Platforms

The BDV Reference Model above from BDVA TF6 work in 2018, shows where the areas of AI/Analytics platforms and Data platforms (Industrial, Personal, Research, Urban/Governmental) are placed in the BDV Reference Model.

In the DataBench analysis we will also identify any usage of AI platforms and Data sharing platforms, as well as links to supporting areas of Cloud and High Performance Computing for data processing and Internet o Things platforms for sensor data collection.

3 Activities and Timing

General activities and timing of WP4 is reported below. The next sections discuss the interrelations and timing of information exchanges between WP4 and other WPs, as well as the timing of the main activities to be performed in WP4.

WP4 - EVALUATING BUSINESS PERFORMACE WITH DATABENCH TOOLBOX	MONTH		2018										2019										
TASK	start	end	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
T 4.1 - Data collection	10	18	D4.1 Data collection plan																							
Case study piloting																										
											D4.2 Data collection results															
Case studies analysis																										
T 4.2 - Evaluation of business performance	18	34									D4.3 Evaluation of business performance															
T 4.3 - Production of DataBenach handbook	24	34									D4.4 DataBench benchmarking handbook															

3.1 Inputs from WP2

WP2 will provide an initial analysis of actual and emerging needs of European industrial users by assessing the value and economic relevance of the EU industry sectors using and developing BDT technologies and the industrial significance of the performance parameters measured by the BDT benchmarks.

As part of this activity, WP2 will run a user survey and complete desk research to focus on on the most frequent use cases affecting relevant business processes with potentially the highest economic impacts. The resulting analysis of the survey will provide WP4 with a classification of industries and uses cases that can be used to hone a long list of potential case study candidates. The selection of industries will be based on their share of value added and their intensity of use of BDA so that the case studies represent a relevant share of the EU industry.

3.2 Inputs from WP3

WP3 will need specific test use cases right after the initial release of the alpha version of the DataBench Toolbox expected by M18. This means that the work done in both WP3 and WP4 should be synchronised to be able on the one hand to select and implement the derivation of business metrics for specific cases, and to test their appropriateness. In order to do that, WP3 would need the following:

- Receive a preliminary specification of the business metrics derivation process for the specific use cases selected in WP4 no later than M13;
- Proceed with the implementation of those specific derivation processes and integrate them into the Toolbox for its alpha version expected in M18;
- Testing of the results from M18 to M20;
- Receive a specification of how to generalize the business metrics generation process as much as possible, or how to handle cases where the process cannot be generalized by M21;
- Implement those processes for the beta version of the Toolbox expected by M24.

After the release of the beta version it is expected to have a final release of the Toolbox including visualization and search components expected by M30. As stated in D3.1, some of the requirements for derivation of business metrics might be achieved using visualization elements as support for the decision of the users, rather than implementing specific algorithms that calculate business metrics. This means that after the 2 intermediate releases these potential decision support visual elements for business users should be discussed between WP3 and WP4 and potentially implemented in the final release.

3.3 Case Study Analysis

The pilot case will be executed in Months 9-10. The pilot case will be selected in the first two weeks of Month 9. The company will be contacted and material from the company will be requested in Month 9. The first interview will be conducted in the first half of Month 10. Possible follow-ups with the company will occur in the second half of Month 10.

The desk analysis will be drafted by Month 10 and then completed by Month 12.

The first draft of the extended list of case studies will be completed by Month 10, with the cooperation of all partners involved in WP4. Companies will be contacted in Months 11-15, scheduling interviews and visits in the Month 11-18 time frame.

The documentation will be produced continuously and will be shared with all partners. Preliminary insights will be shared with partners of WP2 and WP3 by Month 13.

4 Conclusions

4.1 Feedback Information to WP2

WP2 and WP4 are both responsible for identifying and assessing business impacts of benchmarks and, as such, will cooperate closely.

The main deliverables, objectives and impacts of WP2 are detailed in Table 3:

WP2 Tasks and Deliverables	Relevant Objectives	Relevant Milestones	Timing	Contribution to Impacts
Task 2.1 - D.2.1 Economic, Market and Business Analysis Methodology	Objective II		M3 March 2018	1) Availability of solid, relevant, consistent and comparable metrics for measuring progress in Big Data processing and analytics performance
Task 2.2 - D.2.2 Preliminary Benchmarks of European and Industrial significance	Objective II	**MS06** Delivery of benchmarks to assess European and Industrial Significance	M12 December 2018	2)Sustainable and globally supported and recognized Big Data benchmarks of industrial significance
Task 2.3 – Task 2.4 - D.2.3 Analysis of actual and emerging industrial needs and use case mapping	Objective II, IV		M18 June 2019	3)Improvement of competitiveness for European Industry
Task 2.5-D.2.4 Benchmarks of European and Industrial significance	Objectives I, II and VI	**MS11** Demonstrate the European and Industrial Significance of Benchmarking Technologies	M24 December 2019	2)Sustainable and globally supported and recognized Big Data benchmarks of industrial significance
		MS13 DataBench Handbook available to guide in the use of performance benchmarks	M30 June 2020	1)Availability of solid, relevant, consistent and comparable metrics for measuring progress in Big Data processing and analytics performance

Table 3: Deliverables, Objectives and Impacts of WP2

The central premise that links both WP2 and WP4 together hinges on the identification of use case by industry as shown in Figure 6:

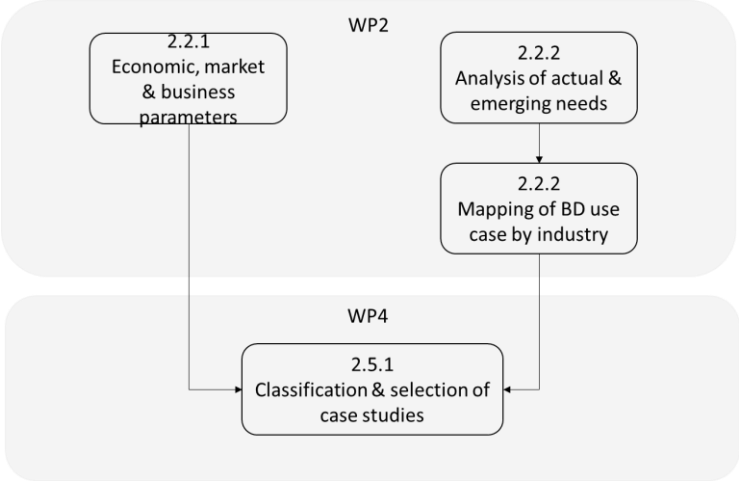

Figure 6 Task Dependencies between WP2 and WP4

WP2 uses a top-down methodological approach, leveraging European economic and industrial analysis to assess the overall relevance and potential impacts of business benchmarks. The analysis of actual and emerging needs of European industrial users pursued by WP2 (outlined in 2.2.) will provide a basis for the assessment of industrial significance and feed into the WP4 activities in this area.

WP4 builds on a bottom-up methodological approach focused on the evaluation of business performance in specific BDT initiatives representing a sample of use cases. WP4 will provide WP2 with the results of business performance measurements in the sample of use cases which will feed back into WP2 for the assessment of the use cases and the development of the final benchmarks.

4.2 Feedback Information to WP3

The main deliverables and impacts of WP3 and the relation with WP4 activities are detailed in Table 4:

WP3 Tasks and Deliverables	Timing	Contribution to Impacts
Task 3.1 - D.3.1 Definition of the DataBench Toolbox architecture	M6 June 2018	1) Definition of the DataBench Toolbox architecture. This is an input to WP4 to understand the type of tooling support is planned from WP3.
Task 3.2 - D.3.2 Alpha version of the DataBench Toolbox	M18 June 2018	2)First version of the DataBench Toolbox. This version will be focused on building the mechanism to automate the reuse and selection of big data benchmarking tool from the Toolbox. WP4 should contribute by M13 with the selection of a specific use case and the way to derive concrete business metrics in order WP3 to provide a proof of concept implementation of the derivation process in the alpha version.

Task 3.3 - D.3.3 DataBench Toolbox - Beta including end-to-end-scenario tool	M24 December 2019	3)Improvement of the DataBench Toolbox by adding the mechanisms to integrate new benchmarking tools. WP4 T4.2 should contribute to the testing of the alpha version of the DataBench Toolbox and the proof of concept for the use case-driven specific business metrics derivation process implemented in the alpha version. This should be done between M18 and M20 to give feedback to WP3 for the second version of the Toolbox.
Task 3.4-D.3.4 Release Version of DataBench Toolbox including visualization and search components	M30 June 2020	4)Final version of the DataBench Toolbox including the search and visualization interfaces as well as the generalization of the derivation procedures to get business insights. Before M30, WP4 D4.3 will test further the DataBench Toolbox to evaluate business performance and give feedback to WP3 in terms of how to derive business metrics. The final evaluation will take place after the final release of the Toolbox expected by M30. WP4 will get input from WP3 to task T4.4 to produce the DataBench Handbook describing how users should make use of the Toolbox.

Table 4: Deliverables, Objectives and Impacts of WP3

As described in Table 4, WP3 and WP4 are tightly linked together. On the one hand, WP3 relies on WP4 for the identification of specific use cases for testing the Toolbox, and the identification of business metrics and their potential derivation process. On the other hand, WP4 will perform business performance testing using the different versions of the DataBench Toolbox giving feedback to WP3 for further improvements. Last but not least, WP3 will provide input to the production of the DataBench Handbook.

4.3 Concluding Remarks

WP4 will investigate and frame the relationship between technical and business KPIs. As shown in Figure 7, the characteristics of business processes represent a fundamental driver of business KPIs, while the features of data and the characteristics of the technical architecture represent a fundamental driver of the technical KPIs and, hence, the choice of the benchmarking tools. The case study analysis will provide several blueprints of this relationship (it should be noted that the desk analysis will be thorough and extensive and will also provide numerous blueprints). Overall, we expect the analysis to be broad and the resulting blueprints to be heterogeneous, as they depend on many variables, including industry, company size, maturity, etc.

As information is collected, our objective is to make a classification effort that simplifies the understanding of this relationship, highlighting the most important variables that drive technical and business KPIs. For example, we may find applications of BDAs that are frequent and consistently successful, possibly cross-industry. In this case, the corresponding blueprint will play an important role in the definition of the summary, high-level framework. We may also find that some use cases can be associated not only with KPIs, but also with numerical estimates of the KPIs. These use cases will also play an important role, while KPIs that are only hypothesized by companies without empirical proof will be

considered less important, especially in association with use cases that are less mature (e.g. pilots) or less frequent (e.g. an original yet unexplored idea of a single company).

This challenging exploratory bottom-up work will be strongly tied to the extensive top-down research performed in WP2. While the large-scale questionnaire performed in WP2 is key to understanding the distribution of use cases across industries and countries, the case study analysis in WP4 will be more likely to obtain estimates of business KPIs and to show how those benefits can be enabled or increased with technical benchmarking and correct architectural choices. Together, extensive research and in-depth case studies can provide invaluable guidelines for companies on how to prioritize and conduct their BDT initiatives.

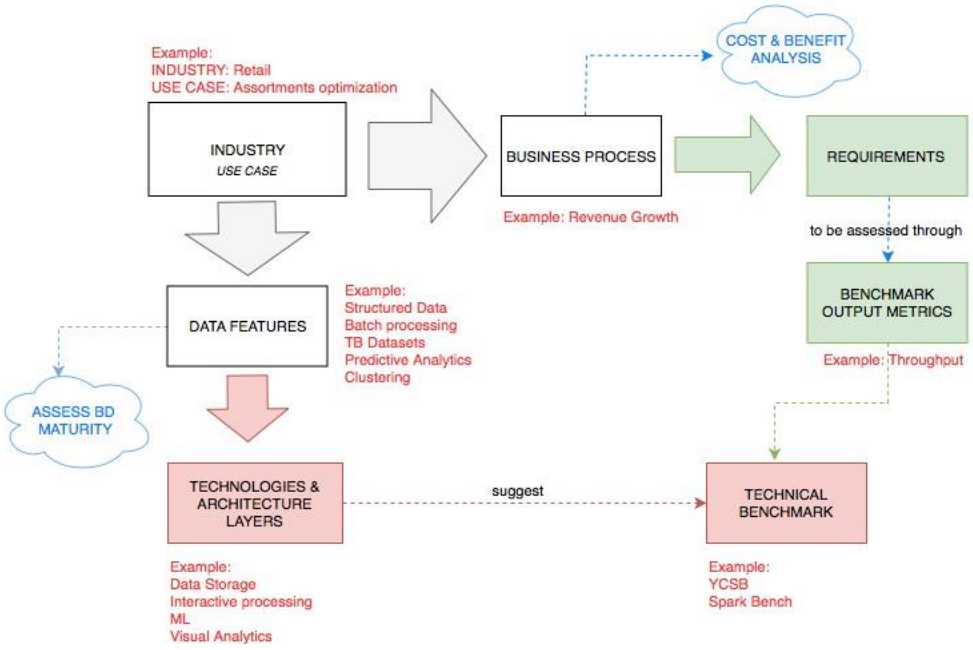

Figure 7 Discovering the Relationship between Technical and Business KPIs

5 Annex I - References

[1] Milfs, A. J., Durepos, G., Wiebe, E. (Eds.), 2010. *Encyclopedia of Case Study Research*, Sage Publications. California. p. xxxi. ISBN 978-1-4129-5670-3.

[2] Ross, J. W., Beath, C. M., Quaadgras, A. 2013. "You may not need bid data after all," *Harvard Business Review*, Dec.

[3] Niebel, T., Rasel, F., Viete, S., 2017. "BIG Data – BIG Gains? Empirical Evidence on the Link Between Big Data Analytics and Innovation," Discussion Paper No. 17-053, *ZEW*, Center for European Economic Research, Mannheim.

[4] Calabrò, A., Lonetti, F., Marchetti, E., 2015. "KPI Evaluation of the Business Process Execution through Event Monitoring Activity," Third International Conference on Enterprise Systems, DOI 10.1109/ES.2015.23.

[5] BDV SRIA, Big Data Value association, European Big Data Value - Strategic Research and Innovation Agenda, vers. 4.0, Oct. 2017, http://www.bdva.eu/sria

[6] Deliverable: DataBench D3.1. DataBench Architecture, Pariente, Tomas, 2018.

Web sources of information for desk analysis:

https://blog.bigml.com/

https://www.kaggle.com/

https://archive.ics.uci.edu/ml/index.php

https://resourcewatch.org/

https://it.hortonworks.com/solutions/

https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/#6427ad83289b

https://www.thinkwithgoogle.com/

https://blog.capterra.com/10-ways-commercial-construction-companies-can-use-big-data/

https://www.qubole.com/resources/big-data-media-entertainment/

https://it.teradata.com/Resources?AssetType=Case+Studies

https://blogs.microsoft.com/iot/

Appendix (1) – Reference Model for Case Study Analysis (extended based on BDVA Template)

Case study title	
Industry	
Author/Company/email	
Actors/Stakeholders and their role and responsibilities	
Company description	
Case study description	
Goals	
Current solution	Compute (System)
	Storage
	Networking
	Software
Data characteristics	Data volume
	Data velocity
	Data variety (data types)
	Data variability
Data sharing and exchange platform use	
Data Anonymization and Privacy needs	
Data processing and analytics characteristics	Data volatility
	Data veracity
	Data monetary value
	Data visualization
	Data storage

	Data processing (On premise, Cloud, HPC)
	Analytics and machine learning/AI
Big Data specific challenges	Short term (in 1 year)
	Long term (in 5 years)
Relevant technical performance metrics	
Technical benchmark adoption	Current
	Short term (in 1 year)
	Long term (in 5 years)
Expected benefits	Short term (in 1 year)
	Long term (in 1 year)
Business KPIs	Current (measured)
	Short term (expected)
	Long term (expected)

Appendix (2) – List of Use Cases considered for Preliminary Desk Analysis

NAME	LINK	DESCRIPTION	BUSINESS KPI	INDUSTRY	DIMENSIONS FOUND
Quality control with real-time and historical data on the assembly line	https://it.hortonworks.com/solutions/manufacturing/	Improve process and products to levels that cannot be reached in the absence of sufficient data	CUSTOMER SATISFACTION PRODUCT/SERVICE QUALITY	MANUFACTURING PROCESS	TYPE OF DATA: Time series (including IoT data) BD & ANALYTICS APPLICATION AREA: customer service & support SCOPE OF BD & ANALYTICS: data oriented digital transformation DATA USER: user companies
Listerine	https://www.thinkwithgoogle.com/intl/en-145/success-stories/global-case-studies/case-study-listerines-customized-video-approach/	Customized video	REVENUE GROWTH	RETAIL TRADE	BD & ANALYTICS APPLICATION: marketing BD MATURITY: piloting or implementing DATA USER: vendors in the ICT industry SCOPE OF BD & ANALYTICS: decision optimization task processing paradigm: iterative/in-memory TYPE OF DATA: graph or linked data DATA STORAGE: graph databases
Driving insurance	https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/#6427ad83289b	predictive analysis applied to people behaviour when they drive	REVENUE GROWTH	INSURANCE	TYPE OF ANALYTICS: predictive analytics TYPE OF DATA: media (image, audio or video)
Increased productivity in the pharmaceutical field	https://it.hortonworks.com/solutions/manufacturing/	Maximize the yield	COST REDUCTION	MANUFACTURING PROCESS	SCOPE OF BD & ANALYTICS: data driven business processes DATA USER: data enterpreneurs BD & ANALYTICS APPLICATION AREA: product innovation TYPE OF ANALYTICS: predictive analytics TYPE OF DATA: geospatial or temporal data MACHINE LEARNING APPROACH: deep learning
Analyse the sentiment towards the brand	https://it.hortonworks.com/solutions/retail/	Picture of the opinions on a brand expressed in social media	REVENUE GROWTH CUSTOMER SATISFACTION	RETAIL TRADE	TYPE OF DATA: graph or linked data TYPE OF ANALYTICS: diagnostic analytics BD & ANALYTICS APPLICATION AREA: marketing
Optimize websites	https://it.hortonworks.com/solutions/retail/	Optimize websites	REVENUE GROWTH INNOVATION	RETAIL TRADE	TYPE OF DATA: XML DATA STORAGE: graph databases PROCESSING PARADIGM BD & ANALYTICS APPLICATION AREA: IT and Data Operation BD maturity: currently using
Develop new products	https://it.hortonworks.com/solutions/telecc	Capture in-depth information specific to a product, as well as certain geographic areas and consumer segments	INNOVATION	TELCO	TYPE OF DATA: graph or linked data PROCESSING PARADIGM: iterative/in-memory PERFORMANCE METRIC: availability SCOPE OF BD & ANALYTICS: decision optimization task BD MATURITY: considering or evaluating for future use BD & ANALYTICS APPLICATION AREA: marketing
Permanent storage of data from research in the medical field	https://it.hortonworks.com/solutions/healthcare/	Supporting data sets permanently available	PRODUCT/SERVICE QUALITY	HEALTHCARE	DATASET SIZE: terabytes PROCESSING PARADIGM: real time TYPE OF DATA: structured text BD & ANALYTICS APPLICATION AREA: customer service & support DATA USER: user companies PERFORMANCE METRIC: availability SCOPE OF BD & ANALYTICS: data driven business process BD MATURITY: considering or evaluating for future use TYPE OF ANALYTICS: predictive
Monitoring on equipment, drugs and health workers through RFID data	https://it.hortonworks.com/solutions/healthcare/	Predictive health analysis	PRODUCT/SERVICE QUALITY TIME EFFICIENCY	HEALTHCARE	DATASET SIZE: terabytes PROCESSING PARADIGM: real time TYPE OF DATA: IoT BD & ANALYTICS APPLICATION AREA: customer service & support DATA USER: user companies SCOPE OF BD & ANALYTICS: data driven business process BD MATURITY: considering or evaluating for future use TYPE OF ANALYTICS: predictive
Information on energy exchange, a step ahead of the markets	https://it.hortonworks.com/solutions/energ	Real-time trading technologies enable energy suppliers to respond instantly to market opportunities without exposing their organization to unnecessary legal or financial risks	PRODUCT/SERVICE QUALITY INNOVATION	ENERGY	TYPE OF DATA: IoT, graph or linked data PROCESSING PARADIGM: real time DATASOURCE: distributed TYPE OF ANALYTICS: predictive PERFORMANCE METRIC: end-to-end execution time BD & ANALYTICS APPLICATION AREA: product management DATA USER: user companies BD MATURITY: currently using SCOPE OF BD & ANALYTICS: decision optimization task
AUSOL Load Balancing Pilot (Transforming Transport)	https://transformingtransport.eu/ausol-load-balancing-pilot	Understanding the road traffic Optimize highway operation Guarantee safer roads and make a better use of these roads	TIME EFFICIENCY PRODUCT/SERVICE QUALITY COST REDUCTION INNOVATION	TRANSPORT AND LOGISTICS	BD & ANALYTICS APPLICATION AREA: product innovation DATA USER: vendors in the ICT industry SCOPE OF BD & ANALYTICS: data driven business process TYPE OF DATA: Geospatial or temporal data MACHINE LEARNING APPROACH PROCESSING PARADIGM: real time TYPE OF ANALYTICS: predictive PERFORMANCE METRIC: availability
Precision agriculture in olives, fruits, grapes and vegetables (DataBio)	https://www.databio.eu/en/pilots/ deliverable D1.1	Providing a set of smart farming services to farmer utilizing available precision agriculture techniques	REVENUE GROWTH INNOVATION COST REDUCTION	AGRICULTURE	TYPE OF ANALYTIC: descriptive and prescriptive DATASOURCE: centralized DATASET SIZE: terabytes PROCESSING PARADIGM: real-time MACHINE LEARNING APPROACH: deep learning TYPE OF DATA: time series 8including IoT) DATA STORAGE: cloud BD & ANALYTICS APPLICATION AREA: product management DATA USER: user company SCOPE OF BD & ANALYTICS: data driven business process BD MATURITY: considering
Parking Availability (QROWD)	http://qrowd-project.eu/wp-content/uploads/2018/01/Business-case-requirements-and-design.pdf	Information about the probability to find a parking spot for four- and two-wheeled vehicles	CUSTOMER SATISFACTION	TRANSPORT AND LOGISTICS	TYPE OF ANALYTIC: predictive PROCESSING PARADIGM: real-time MACHINE LEARNING APPROACH: yes TYPE OF DATA: media DATA STORAGE: datasets PERFORMANCE METRIC: availability DATA USER: user company SCOPE OF BD & ANALYTICS: decision optimization task BD MATURITY: considering or evaluating for future use
Advanced time-series analytics in the automotive sector (AEGIS)	https://www.aegis-bigdata.eu/wp-content/uploads/2017/03/AEGIS-D1.2-The-AEGIS-Methodology-and-High-Level-Usage-Scenarios-v1.0.pdf	A research team creates services on top of streaming vehicle data to (a) identify unsafe driving patterns and correlate them with external conditions and (b) to timely detect damages in the road network.	TIME EFFICIENCY INNOVATION	TRANSPORT AND LOGISTICS	PROCESSING PARADIGM: streaming MACHINE LEARNING APPROACH: hybrid machine learning TYPE OF DATA: Time series (including IoT data) PERFORMANCE METRIC: end-to-end execution time BD & ANALYTICS APPLICATION AREA: customer service & support DATA USER: user companies SCOPE OF BD & ANALYTICS: cdata driven business process BD MATURITY: considering or evaluating for future use
Mare Protection (BigDataOcean)	http://www.bigdataocean.eu/site/wp-content/uploads/2016/12/BigDataOcean_D4.1_BigDataOcean_Technology_Requirements_and_User_Stories_v1.00.pdf	This pilot is concerned with enabling the proper handling of pollution accidents in the marine environment through the use of tools as part of a Decision Support System.	INNOVATION	HEALTHCARE	TYPE OF ANALYTIC: predictive DATASOURCE: centralized/distributed DATASET SIZE: terabytes PROCESSING PARADIGM: streaming TYPE OF DATA: relational, noSQL, triplestore DATA STORAGE:relational, noSQL, triplestore PERFORMANCE METRIC: availability BD & ANALYTICS APPLICATION AREA: DATA USER: user company SCOPE OF BD & ANALYTICS: data driven business process BD MATURITY: considering

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966