DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

D4.4 DataBench Benchmarking Handbook

Abstract

This is the DataBench Handbook and final report of DataBench WP4. Its main goal is to support the final exploitation and sustainability of the project's results by providing information and guidelines to the future users of the DataBench Toolbox. The DataBench Toolbox is a software tool which will provide access to benchmarking services, KPIs and various types of knowledge: the DataBench Handbook plays a complementary role to the Toolbox by providing a comprehensive view of how the project developed and validated the benchmarks offered in the Toolbox, of how technical and business benchmarking are linked in the project's research and providing guidelines to the data and information contained in the Toolbox. In addition, the Handbook provides references to the main project's deliverables for users interested to dwelve more in depth into the project's results. The DataBench Handbool and Toolbx are aimed at industrial users and European technology developers who need to make informed decisions on Big Data Technologies investments by optimizing technical and business performance. This Handbook demonstrates how DataBench achieved its goal to design a benchmarking process helping European organizations developing Big Data Technologies (BDT) to reach for excellence and constantly improve their performance, by measuring their technology development activity against parameters of high business relevance. In this way DataBench has filled a gap in BDT knowledge and understanding.

Deliverable D4.4	DataBench Benchmarking Handbook
Work package	WP4
Task	4.4
Due date	31/10/2020
Submission date	
Deliverable lead	IDC
Version	1.0
Authors	IDC (Gabriella Cattaneo, Richard Stevens, Cristina Pepato, Erica Spinoni) Polimi (Chiara Francalanci) ATOS (Tomas Pariente Lobo, Ricardo Ruiz) Sintef (Arne Berre)
Reviewers	Polimi (Sergio Gusmeroli), LEAD Consult (Todor Ivanov)

Keywords

Disclaimer

Copyright Notice

Table of Contents

Table of Figures

Table of Tables

Executive Summary

This is the DataBench Handbook and final report of DataBench WP4. Its main goal is to support the final exploitation and sustainability of the project's results by providing information and guidelines to the future users of the DataBench Toolbox. The DataBench Toolbox is a software tool which will provide access to benchmarking services, KPIs and various types of knowledge: the DataBench Handbook plays a complementary role to the Toolbox by providing a comprehensive view of how the project developed and validated the benchmarks offered in the Toolbox, of how technical and business benchmarking are linked in the project's research and providing guidelines to the data and information contained in the Toolbox. In addition, the Handbook provides references to the main project's deliverables for users interested to dwelve more in depth into the project's results.

DataBench Value Proposition

The DataBench Handbool and Toolbx are aimed at industrial users and European technology developers who need to make informed decisions on Big Data Technologies investments by optimizing technical and business performance. The DataBench toolbox helps stakeholders to identify the use cases where they can achieve the highest possible business benefit and return on investment, so they can prioritize their investments; to select the best technical benchmark to measure the performance of the technical solution of their choice; to assess their business performance by comparing their business impacts with those of their peers, so they can revise their choices or their organization if they find they are achieving less results than median benchmarks for their industry and company size. Therefore, the services provided by the Toolbox and the Handbook will support users in all phases of their users' journey (before, during and in the ex-post evaluation of their BDT investment) and from both the technical and business viewpoint.

This Handbook demonstrates how DataBench achieved its goal to design a benchmarking process helping European organizations developing Big Data Technologies (BDT) to reach for excellence and constantly improve their performance, by measuring their technology development activity against parameters of high business relevance. In this way DataBench has filled a gap in BDT knowledge and understanding.

Conceptual Framework

The conceptual framework of DataBench research links business and technical evaluation of BDT benchmarking and relies on a systematic ecosystem of indicators. The research process of the project (Figure 1) shows how the conceptual framework was developed on the basis of an analysis of the state of the art of benchmarking and the development of indicators (WP1) followed by in-depth data collection and research of industrial users' needs and measurement of KPIs on business needs, as well as case studies (WP2 and WP4). The DataBench Toolbox was developed in several iterations feeding from and interacting with these activities and validated through the technical work of WP5.

The research process of the project (Figure 1) shows how the conceptual framework was developed on the basis of an analysis of the state of the art of benchmarking and the development of indicators (WP1) followed by in-depth data collection and research of industrial users' needs and measurement of KPIs on business needs, as well as case studies (WP2 and WP4). The DataBench Toolbox was developed in several iterations feeding from and interacting with these activities and validated through the technical work of WP5.

Stakeholder analysis

DataBench carried out an in-depth analysis of industrial users' needs for the adoption of BDT in order to tailor our benchmarking services to demand requirements. Data collection was focused on actual and potential BDT users. Business organisations realize the importance of benchmarking the business impacts of BDT, as shown by Figure 10 below, only 10% of them dismiss it as not at all or slightly important: moreover, 45% consider benchmarking very or extremely important.

The perception of the importance of benchmarking is positively correlated with the level of adoption (actual users evaluate it very highly) and the company size (with large companies more appreciative of its importance than small ones).

From the case study analysis, we have seen that it is important to make technical choices that can support long-term change in order to enable greater business benefits. From the evidence that has been collected so far from case studies, an important lesson learnt is that most companies believe that technical benchmarking requires highly specialized skills – skills that are not currently present in the company – and considerable investments. There is a general agreement on the fact that BDTs are diverse and complex and that technical choices are not simple and are potentially impactful. Even if companies do not perform benchmarking activities, they have been found to rely on trusted external entities to compare technologies, such as IT consultants and system integrators.

Technical Benchmarks

The DataBench Framework for Big Data and AI Benchmarks is based on the BDVA (Big Data Value Association) Reference Architecture. In order to have an overall perspective on Big Data and AI systems the usage of a top level generic pipeline has been introduced. The Handbook presents and explains the main reference models used for technical benchmarking analysis.

Business Benchmarks

The Handbook describes the BDT business benchmarks by industry, and by company size, which are also accessible for users in the Toolbox in the Knowledge Nugget section. The Handbook provides the main benchmarks values and explanatory comments.

The business benchmarks are calculated on the basis of 8 business KPIs. Thanks to our methodological approach, the business KPIs selected by the project are valid metrics and can be used as benchmarks for comparative purposes by researchers or business users for each of the industry and company-size segments measured. These indicators are:

- Benchmarks, because they represent the average improvement achieved by business users and can be used for comparative purposes, as a target or as a best performance metric.
- Of industrial significance, because they apply to the actual and emerging needs of specific industries and specific company-size segments.
- Of European economic significance, because the benchmarks are measured for all the relevant European industries and company-size segments in which Big Data can have the highest impacts.

- Useful for linking technical and business benchmarking, because they are also measured for the main use cases, consisting of the application of Big Data technology to particular business processes and/or application domains, thus enabling the user to match the expected business improvements with the type of technology performance needed to achieve the business goals.

Presentation of the Toolbox

The DataBench Toolbox is the main technical result of the DataBench project. The Toolbox provides access to a knowledge base of big data benchmarking related artefacts, ranging from metadata about existing benchmarking tools and initiatives in the community to heterogeneous information and studies performed by the project about benchmarking encapsulated in what we call "knowledge nuggets".

The Handbook present an overview of the DataBench Toolbox and guidelines to understand its structure and access its main services.

Next steps

The Handbook and the DataBench toolbox are essential components of DataBench exploitation plan. The project is negotiating an agreement to deliver the Toolbox and the Handbook to the Big Data Innovation Hubs network through the IA EUHubs4Data.

The Handbook will have a second release at the end of the project to include the final results of the Toolbox validation which is due to be concluded in month 35 while this deliverable is due in month 34.

1. Introduction

1.1 Objective

This report is the DataBench Handbook and the final report of DataBench WP4. Its main goal is to support the final exploitation and sustainability of the project's results by providing information and guidelines to the future users of the DataBench Toolbox. This report will have a second release at the end of the project to include the final results of the Toolbox validation which is due to be concluded in month 35 while this deliverable is due in month 34.

The DataBench Toolbox is a software tool which will provide access to benchmarking services, KPIs and various types of knowledge: the DataBench Handbook plays a complementary role to the Toolbox by providing a comprehensive view of how the project developed and validated the benchmarks offered in the Toolbox, of how technical and business benchmarking are linked in the project's research and providing guidelines to the data and information contained in the Toolbox. In addition, the Handbook provides references to the main project's deliverables for users interested to dwelve more in depth into the project's results. In this deliverable we pull together the two main tracks of the project's research, the technical and business tracks, and provide a summary of the final results of this research.

The DataBench Handbool and Toolbx are aimed at industrial users and European technology developers who need to make informed decisions on Big Data Technologies investments by optimizing technical and business performance. This Handbook demonstrates how DataBench achieved its goal to design a benchmarking process helping European organizations developing Big Data Technologies (BDT) to reach for excellence and constantly improve their performance, by measuring their technology development activity against parameters of high business relevance. In this way DataBench has filled a gap in BDT knowledge and understanding.

The Handbook illustrates the conceptual framework developed by the project to investigate existing Big Data benchmarking tools and projects, identify main gaps and provide a robust set of metrics to compare technical results coming from those tools. The Handbook explains and provides the scientific and methodological background of the benchmarks provided by the Toolbox and links it to the scientific state of the art by main benchmarking communities, thereby providing a sound basis to inspire trust and confidence in the users of the DataBench Toolbox results and services. The Handbook and the Toolbox together will serve as a decision-support tool for companies approaching BDT application and as a practical source of quantitative performance targets for companies assessing their actual or future investments in BDT applications. This is the main legacy of the DataBench project.

In the 3 years of the project, the emergence of AI has driven attention to the use of Big Data for AI technologies and tools. Therefore, in the final phase of the project the technical benchmarking analysis has been extended to include considerations about data and AI benchmarking.

1.2 Structure of the Report

The report is structured as follows:

- Chapter 1 (Introduction) outlines the main objectives and the structure of the report.
- Chapter 2 outlines the conceptual framework, methodology and data sources behind the project's benchmarking services, demonstrating the relevance of technical and business benchmarking of BDT, which represent the sound basis of the tools and services offered by the Toolbox.
- Chapter 3 demonstrates how the DataBench benchmarks respond to European users' industrial needs and provides data showing their correspondence to stakeholder categories by industry and company size, which is at the basis of the Toolbox users' profiling and users' journeys. The chapter describes the main use cases of BDT, which have been leveraged to connect technical and business benchmarking, and the case studies used to validate the value of benchmarks.
- Chapter 4 presents an overview of the framework of technical benchmarks analysed by the project, correlated with the BDVA reference architecture, and the pipelines considered by the project.
- Chapter 5 illustrates the business benchmarks developed by the project, based on 7 business impacts KPIs validated through data collection and interaction with BDT users, which are included in the Toolbox Knowledge Nuggets component.
- Chapter 6 presents the structure and main components of the Toolbox and the support offered by user category.
- Chapter 7 explains how the Handbook and Toolbox represent a key component of DataBench exploitation plans and how they will be made available to users after the project's end.

2. Conceptual Framework

2.1. Overview

The conceptual framework of DataBench research links business and technical evaluation of BDT benchmarking and relies on a systematic ecosystem of indicators. The process and the framework are illustrated in the following figures, with more in-depth documentation provided by D1.1, *Industry Requirements with Benchmark Metrics and KPIs*. Concerning the evaluation of business impacts, the methodology is illustrated in detail in D2.1, *Economic and Market Analysis Methodology*, and *D.4.1 Data Collection*.

The research process of the project (Figure 1) shows how the conceptual framework was developed on the basis of an analysis of the state of the art of benchmarking and the development of indicators (WP1) followed by in-depth data collection and research of industrial users' needs and measurement of KPIs on business needs, as well as case studies (WP2 and WP4). The DataBench Toolbox was developed in several iterations feeding from and interacting with these activities and validated through the technical work of WP5.

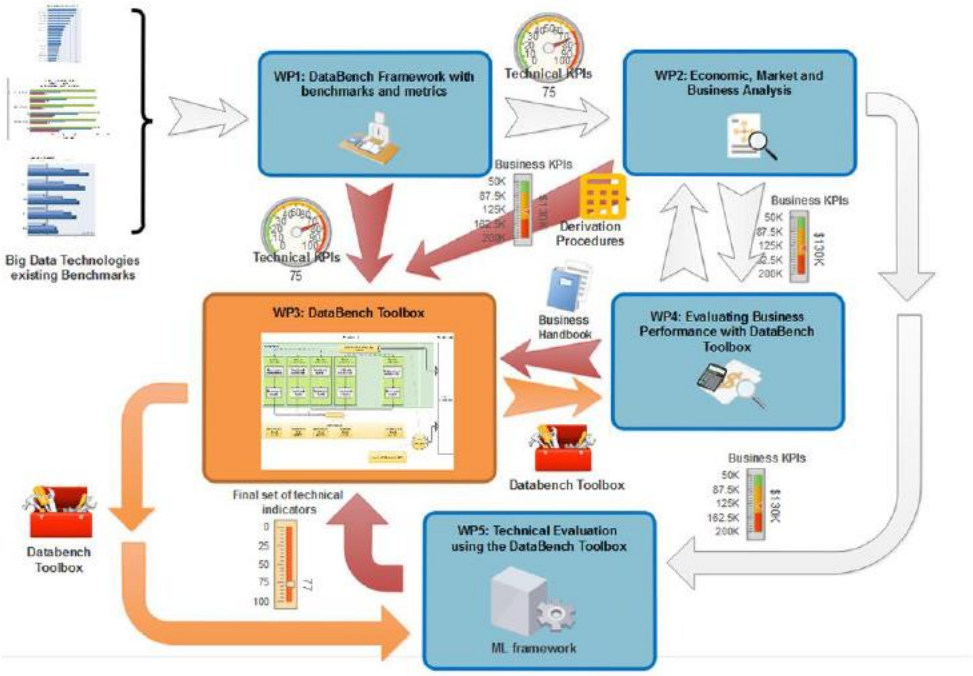

Figure 1 - DataBench Research Process and Outcomes
Source: D.1.1 Industry Requirements with Benchmark Metrics and KPIs.

More specifically, Figure 2 provides a snapshot of the conceptual framework linking technical and business benchmarking in this project. As shown below, the comprehensive analysis of the main BDT (Big Data Technologies) requirements by industry and technology (top layer of the figure) covered business features, BDA application features, platform and architecture features, and of course technical benchmarking features. All these aspects were classified and measured through the DataBench ecosystem of indicators (Figure 3) and fed to the Toolbox. The benchmarking tool in turn is structured around the main data pipelines and performance metrics of the different technical benchmarks included in the tool, helping users to navigate in the benchmark library of the project and select the optimal BDT benchmarking approaches by type of implementation. By type of implementation we mean the implementation of BDT solutions in specific business processes/use cases, which have been also identified and classified. The business impacts of these BDT use cases has been measured through the 7 business KPIs selected by the project, based on the industrial users' survey and case studies, which measure aspects such as revenues and profit growth, customer satisfaction, product and/or service innovation.

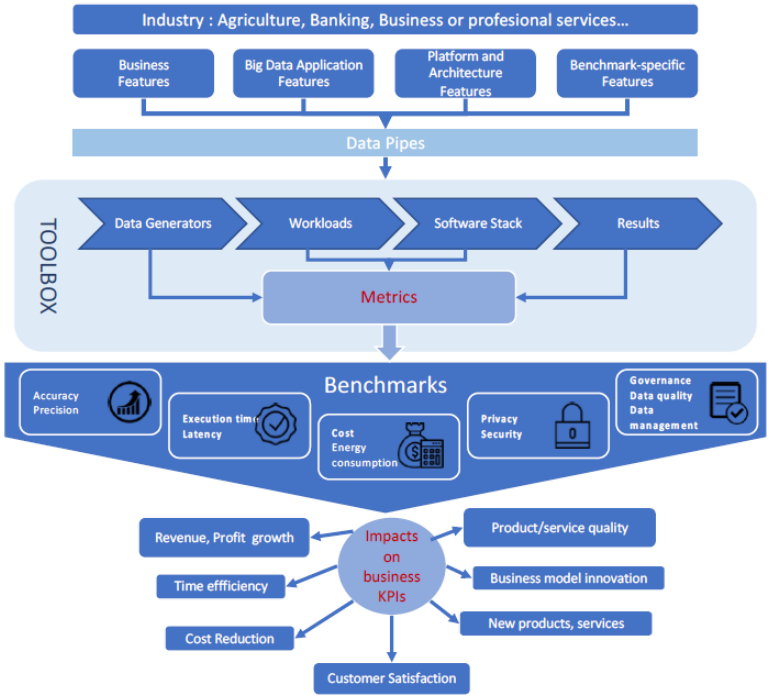

Figure 2 – Technical and Business Benchmarking Framework
Source: DataBench D1.1 "Industry Requirements with benchmark metrics and KPIs"

2.2. Value proposition for stakeholders

It is important remarking that from the start DataBench targeted three main categories of stakeholders, which are addressed differently by the DataBench Toolbox as illustrated in chapter 6:

- Benchmarking Providers: Organizations that own a particular benchmark. They can be the actual developers of the benchmark or the organizations that maintain them. DataBench interacted with them to identify and collect benchmarks. In the Toolbox, these users can register and update their benchmarks.
- Technical Users: Users that would like to search and potentially execute a technical benchmark. DataBench interacted with them to investigate their needs and requirements for benchmarking, and through the Toolbox designed services meeting their needs.
- Business Users: Users that would like to search and understand the business value of specific big data solutions when making choices about BDT investments. The Toolbox provides them with data about potential business benefits for the main use cases and access to business benchmarks by industry and company size. This allows them to compare themselves with their peers and their business achievements.

The DataBench toolbox helps stakeholders to identify the use cases where they can achieve the highest possible business benefit and return on investment, so they can prioritize their investments; to select the best technical benchmark to measure the performance of the technical solution of their choice; to assess their business performance by comparing their business impacts with those of their peers, so they can revise their choices or their organization if they find they are achieving less results than median benchmarks for their industry and company size. Therefore, the services provided by the Toolbox and the

Handbook will support users in all phases of their users' journey (before, during and in the ex-post evaluation of their BDT investment) and from both the technical and business viewpoint.

2.3. The Ecosystem of indicators

The ecosystem of indicators developed by DataBench is shown in the figure 3 below and is aligned to the state of the art of research and best practice from the point of view of economic and market research, as documented in D.2.1, *Economic and Market Analysis Methodology*. As anticipated, the indicators are grouped in 4 main categories (business features, BDA Application Features, Platform and Architecture features and Benchmark specific features).

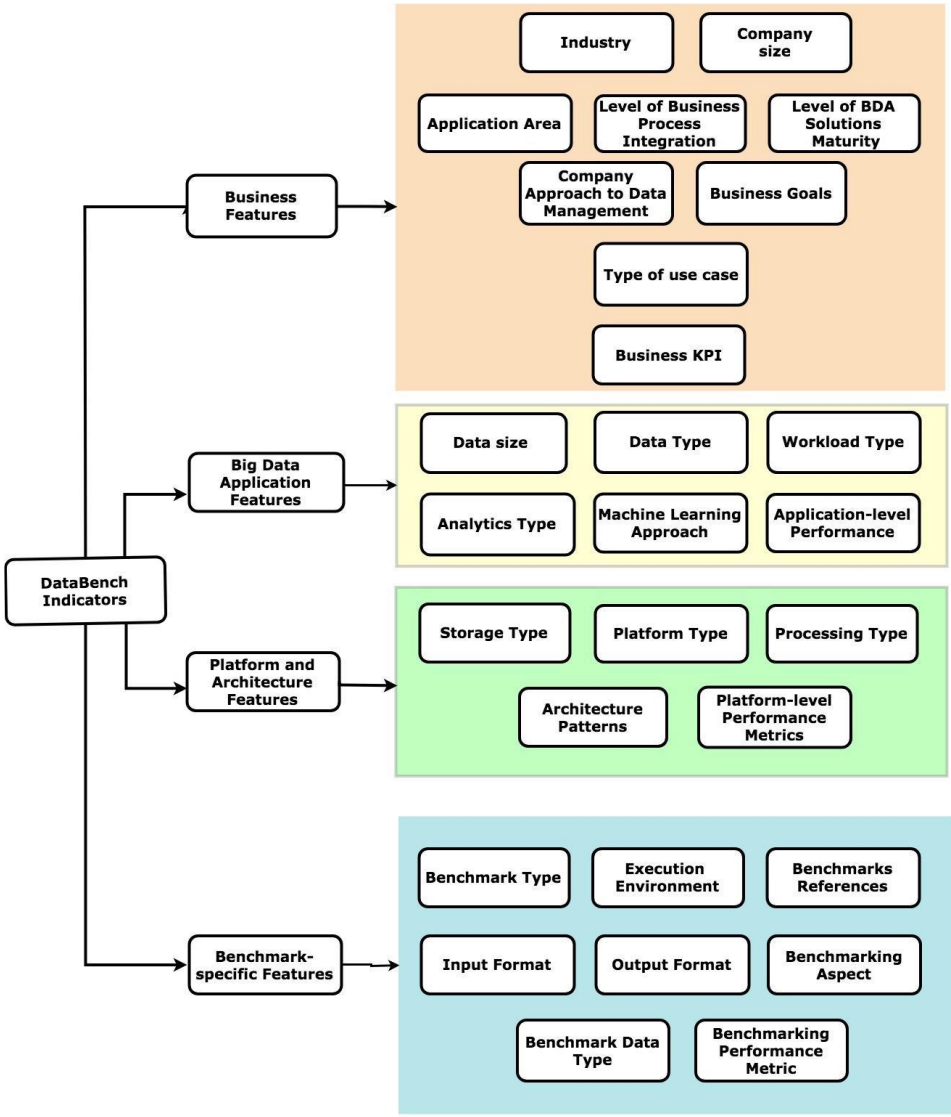

Figure 3 - DataBench Indicators
Source: DataBench D1.1 "Industry Requirements with benchmark metrics and KPIs"

2.3.1. Business Indicators

The business feature indicators can be divided into the following main subgroups:

1. The classification of business users (industry and company size)
2. The type of BDA implementation (application area, level of business process integration, level of BDA solutions maturity, company approach to data management, and main business goals)
3. The type of use case (cross-industry and industry-specific)
4. Business impact KPIs, which correspond to industrial benchmarks

Groups 1, 2, and 3 are semantic indicators measured through simple nominal questions in the survey (business users select the category in which they belong) to classify users. The survey results are measured as frequencies of respondents by category. Descriptive parameters can be used to measure the correlation between the type of user and the type of application and, in turn, the type of business impact. They will be used in the benchmarking tool as a user interface to guide users to identify themselves and their type of BDA application and, in turn, to look for the type of technical benchmark most relevant for them.

The business KPIs (group 4) are different from the others because they are impact indicators. They represent 8 categories of business factor, selected on the basis of business literature and IDC research of technology vendors and users as the most relevant for measuring the impacts of innovative technology investments on business performance. For example, these factors are most often used to evaluate the results of pilots of new technology investments.

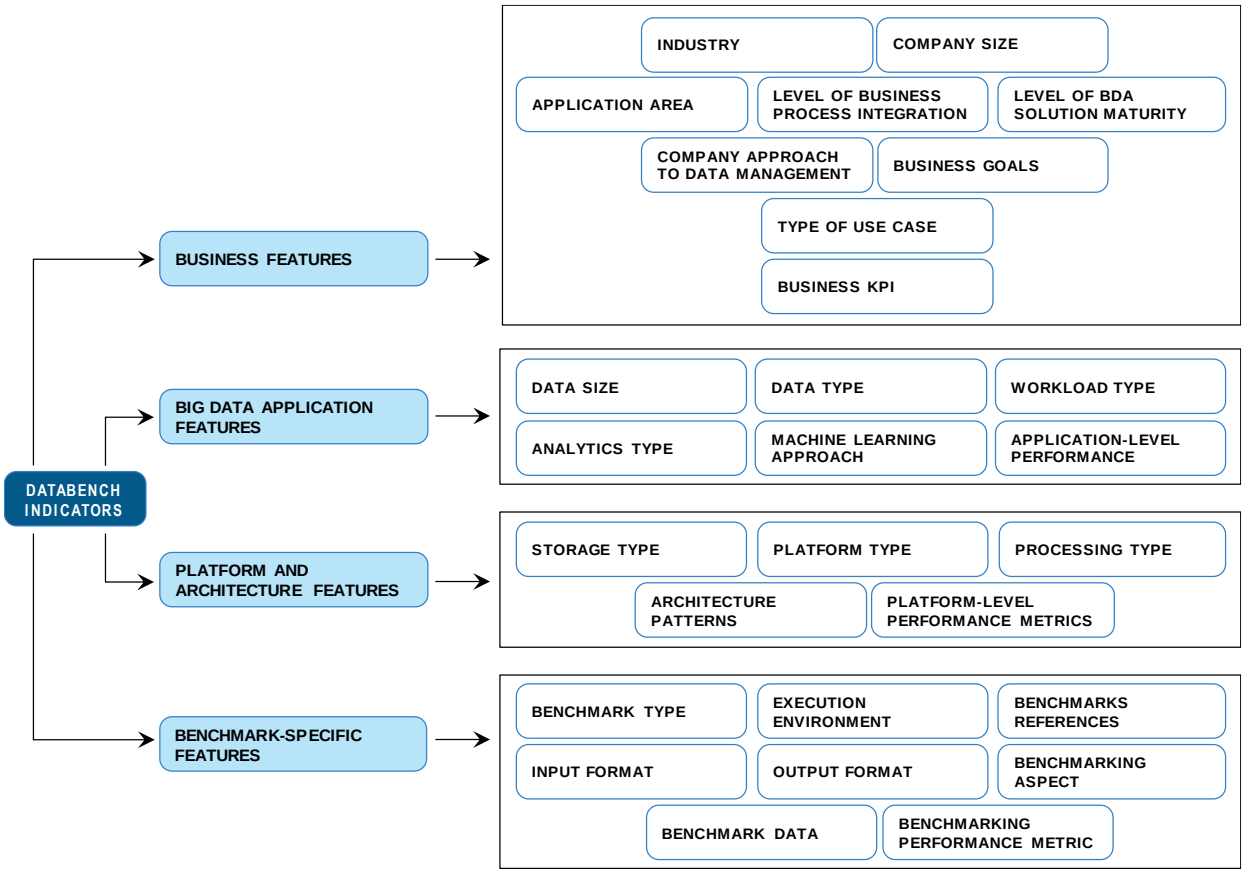

Figure 4 - DataBench Business Indicators.
Source: DataBench D1.1 "Industry Requirements with benchmark metrics and KPIs"

2.3.2. Classification of Use Cases

To bridge the gap between technical and business benchmarking we focus on the identification of use cases, which in this project we define as:

a discretely funded effort designed to accomplish a particular business goal or objective through the application of big data technology to particular business processes and/or application domains, employing line-of-business and IT resources.

Examples of use cases are predictive maintenance in manufacturing, risk assessment in multiple industries, or industry-specific applications such as Yield monitoring and prediction in agriculture. Since a use case is based on a specific technology solution with specific technology performances, but at the same time it is easily correlated with business impacts, it provides a way to evaluate how technology requirements may influence business outcomes. The classification of use cases measured in this project is part of the ecosystem of indicators and presented below.

Industry	Specific Use Cases
Agriculture	Precision agriculture Yield monitoring and prediction Field mapping & crop scouting Heavy equipment utilization
Banking	Cyberthreat & detection
Insurance	Usage based insurance
Other Financial Services	Cyberthreat & detection
Business or Professional services	Social media analytics
Healthcare	Illness/disease diagnosis and progression Personalized treatment via comprehensive evaluation of health records Patient admission and re-admission predictions Quality of care optimization
Manufacturing Process	Smart warehousing Asset management Quality management investigation
Manufacturing Discrete	Smart warehousing Asset management Quality management investigation Connected vehicles optimization

Industry	Specific Use Cases
Retail Trade	Intelligent Fulfillment
Wholesale Trade	Intelligent Fulfillment Increase productivity and efficiency of DCs/warehouses
Telecommunications	Network analytics and optimization
Media	Ad Targeting Scheduling optimisation
Transport & Logistics	Connected vehicles optimization Logistics and package delivery management
Utilities	Field service optimization Energy consumption analysis and prediction
Oil & Gas	Field service optimization Energy consumption analysis and prediction

Table 1 List of Cross-industry BDT use cases

Use Case	Industries
Price optimization	All
New product development	All
Risk exposure assessment	All
Regulatory intelligence	All ((excluding Agriculture)
Customer profiling, targeting, and optimization of offers	Banking, Insurance, Other Finance, Business or Professional services, IT services, Retail Trade, Telecommunications, Media, Utilities
Customer scoring and/or churn mitigation	Banking, Insurance, Other Finance, Telecommunications, Utilities
Fraud prevention and detection	Banking, Insurance, Other Finance, Business or Professional services, IT services, Healthcare, Telecommunications
Product & Service Recommendation systems	Banking, Insurance, Other Finance, Business or Professional services, IT services, Retail Trade, Telecommunications, Media
Automated Customer Service	Banking, Insurance, Other Finance, Business or Professional services, IT services, Healthcare, Retail Trade, Telecommunications, Media
Supply chain optimization	Agriculture, Manufacturing Process and Discrete, Retail Trade, Wholesale Trade, Transport & Logistics, Utilities, Oil & Gas
Predictive Maintenance	Agriculture, Manufacturing Process and Discrete, Wholesale Trade, Transport & Logistics, Utilities, Oil & Gas
Inventory and service parts optimization	Agriculture, Manufacturing Process and Discrete, Wholesale Trade, Transport & Logistics, Oil & Gas

Table 2 List of industry-specific BDT use cases

2.4. Data Sources

The project carried out ad-hoc data collection through a survey of European business organisations in 11 member states and a second wave survey with the business partners of Horizon 2020 ICT14 and ICT15 projects carrying out BDT pilots resulting in a dataset of 730 valid interviews. The industry classification is based on Eurostat's NACE REV. 2 code. The survey excluded micro-enterprises with fewer than 10 employees (unlikely to be advanced adopters of BDT). The answers were used to calculate the value of the business KPIs.

The survey was conducted in the local language by experienced interviewers, targeted senior decision makers and influencers for BDTs, and screened respondents on the basis of

their actual and planned use of BDA. Business organisations not using and not interested in using BDTs were excluded.

In addition, the project carried out desk research on over 700 use cases documented in literature and 22 case studies based on direct interviews, which were used as validation of the business KPIs and the business benchmarks values. The following figures outline the composition of the interviews used to calculate the business benchmarks.

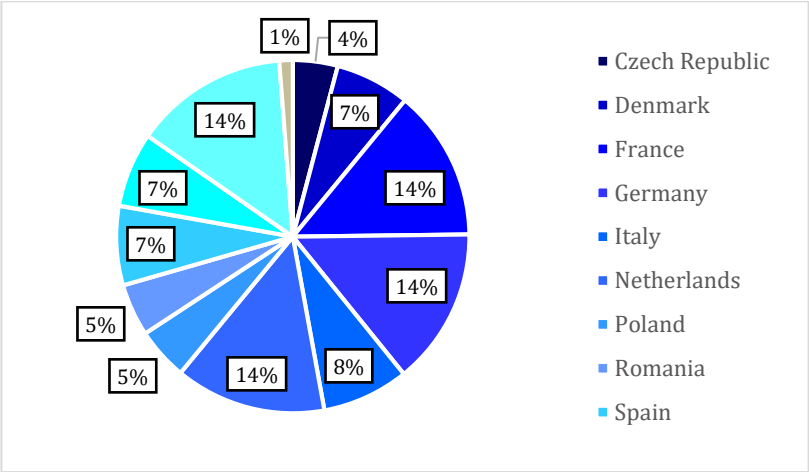

Figure 5 - Respondents by country. Source: DataBench Survey, June 2020

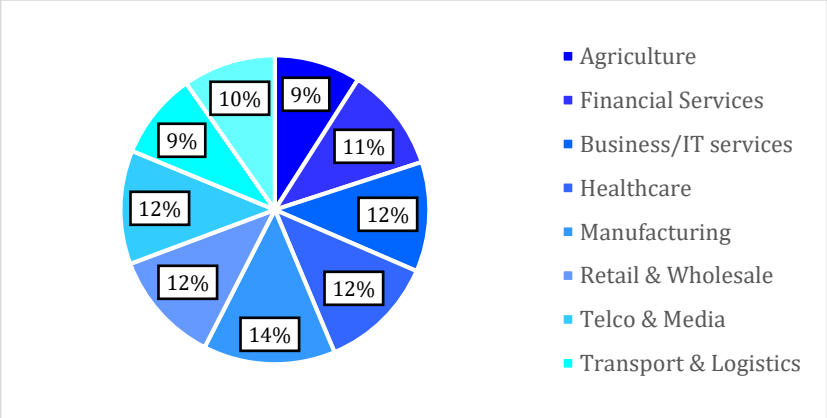

Figure 6 - Respondents by industry. Source: DataBench Survey, June 2020

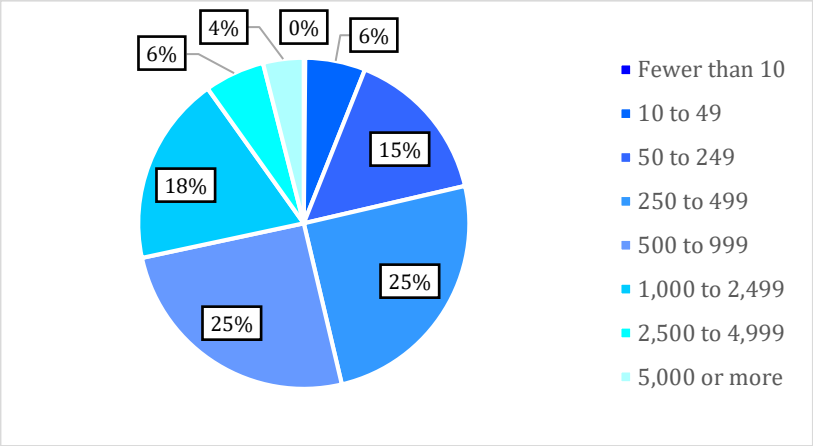

Figure 7 -Respondents by company size. Source: DataBench Survey, June 2020

3. Stakeholder Analysis

3.1 Overview

DataBench carried out an in-depth analysis of industrial users' needs for the adoption of BDT in order to tailor our benchmarking services to demand requirements. Data collection was focused on actual and potential BDT users so in this paragraph we do not focus on take-up but on the BDT implementation process. As shown by Figure 8 below, the priority business goals driving BDT investment concern the improvement of business processes and market understanding, showing the value of data for cross-company functions and activities. Several other business goals are also mentioned by a high share of respondents, confirming that BDT are relevant for multiple business objectives. It is therefore logical that measuring BDT's business impacts is also relevant to assess the effectiveness of these investments.

Recent IDC research on the impacts of the Covid-19 pandemic on the ICT market shows a rise in relevance of technology investments addressed to improve the customer or employee experience. Enterprises need to fight the fall of demand and provide safe (often contactless) customer experiences. This trend requires even higher investments on Big Data correlated with AI powering user friendly interfaces (for example so-called conversational AI). In this context the considerations presented here about industrial users' needs remain more than valid.

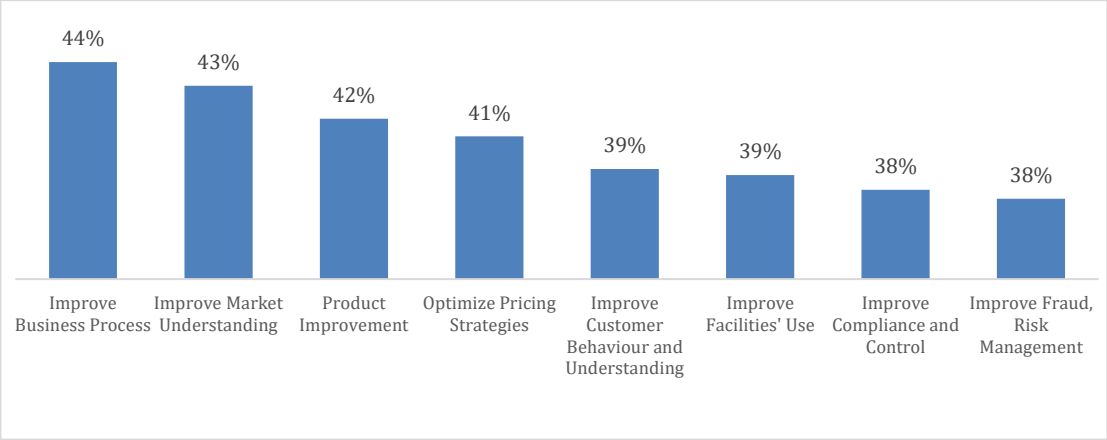

Figure 8 - Business Goals driving BDT adoption (% of respondents).
Source: DataBench survey 2018, n=700, deliverable D.2.2

DataBench results show a high level of satisfaction or expectation with BDT, since 80% of first wave survey respondents declare to have achieved or expect moderate or high benefits (Figure 9) and none have seen negative impacts. Moreover, positive impacts are stronger for actual users, of which 15% have achieved a high level of benefits and 80% a moderate level. This points to a positive dynamic of growing benefits as users progress from the piloting to the scaling up of BDT. Respondents still considering or evaluating BDT are more conservative in their expectations: the majority expects low or medium benefits from BDT.

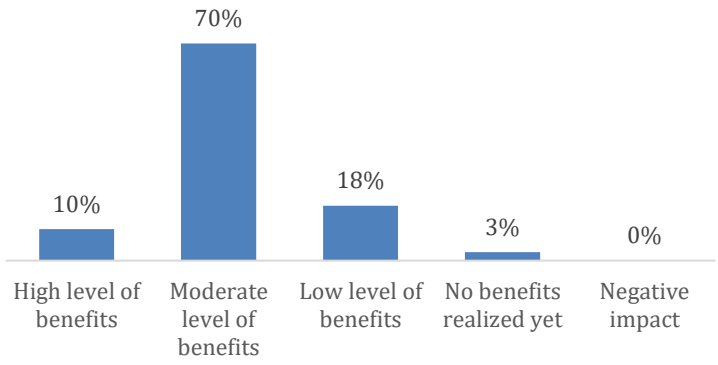

Figure 9 – Expected Benefits from BDT adoption (% of respondents).
Source: DataBench survey 2018, n=700, deliverable D.2.2

Business organisations realize the importance of benchmarking the business impacts of BDT, as shown by Figure 10 below, only 10% of them dismiss it as not at all or slightly important: moreover, 45% consider benchmarking very or extremely important.

The perception of the importance of benchmarking is positively correlated with the level of adoption (actual users evaluate it very highly) and the company size (with large companies more appreciative of its importance than small ones). The industries more advanced and sophisticated in the use of BDT (Finance, Retail, Telecom-Media) again show a higher evaluation of the importance of benchmarking than the others, while laggard industries (Healthcare, Agriculture) show a higher share of respondents not particularly interested in benchmarking. The obvious deduction is that benchmarking becomes relevant when organizations are engaged in practice with BDT. But this also confirms that awareness of BDT business benchmarking is low among SMEs and industries with lower adoption, and the availability of evidence-based benchmarks would be likely to increase awareness and help to make better business decisions.

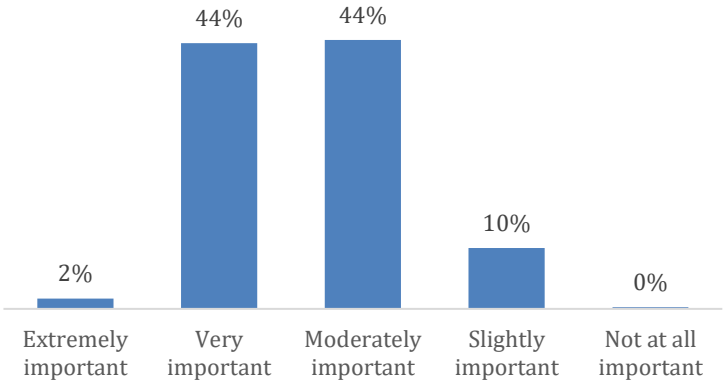

Figure 10 – Importance of benchmarking BDT impacts (% of respondents).
Source: DataBench survey 2018, n=700, deliverable D.2.2

3.3 Adoption of BDT

The maturity around the adoption of BDT is highly variable by industry and company size, with increasing intensity of adoption positively correlated with company size (Figure 11, 12). Adoption by industry also varies considerably even though it is increasing fast in all sectors. Finance, Business/IT services and Telecom media lead by intensity of adoption but retail, utilities and manufacturing also have a relevant share of advanced users.

Organizations in sectors where IT investments were historically lower, for example agriculture and healthcare, require BDT investment choices to be backed by strong business cases and weigh carefully their choices, therefore have good reason to be interested in benchmarking business impacts.

The consequences of the Covid-19 pandemic are likely to drive higher BDT investments exactly in the sectors previously lagging behind, particularly the public sector and healthcare. Governments and healthcare stakeholders will need to leverage BDT and AI to manage more efficiently the treatment of Covid patients, the track and tracing process of contagions and of course to support pharmaceuticals companies in developing and quickly producing and marketing tests and vaccines. Overall this underlines the continuing relevance of DataBench results.

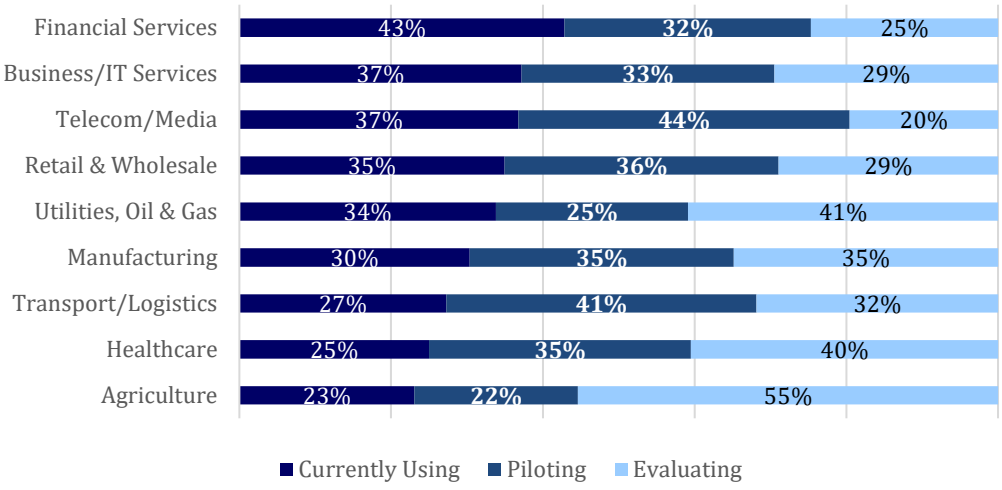

Figure 11 - BDA current and planned adoption by industry.
Source: DataBench deliverable D2.2 "Preliminary Benchmarks of European and Industrial Significance"

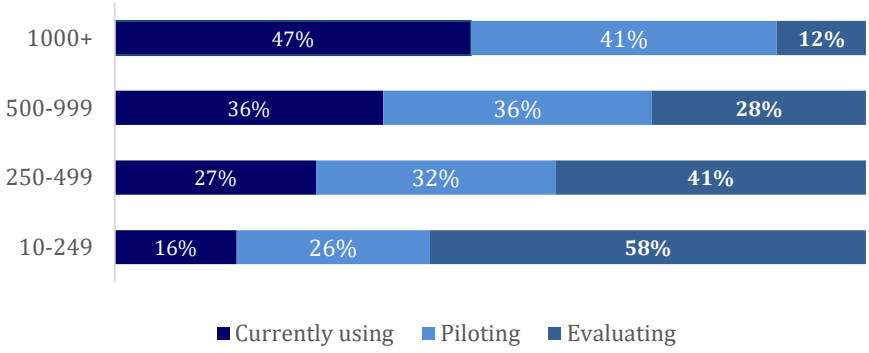

Figure 12 - BDA current and planned adoption by company size.
Source: DataBench D2.2 "Preliminary Benchmarks of European and Industrial Significance"

3.4 Main BDT Use Cases

Within this project, the use cases represent the link between technical solutions and business goals and help the collection of data on the main concrete typologies of BDA's exploitation. In the Knowledge Nuggets section of the Toolbox users can find business benchmarks by industry and type of use case.

We have identified 35 use cases within the DataBench survey to assess the adoption, the maturity of the use cases across different industries and to have a pragmatic and realistic view of the footprint of BDA adoption. Some of the use cases are adopted cross-industries, while other are shared and common to only some of the industries, while some others are really industry specific. The use cases hereby defined are pertinent to storing, transforming, analysing, and harnessing Big Data technologies as a method of enabling organizations to extract value from data to achieve the main business goals.

The figure below shows the ranking of use cases in terms of absolute numbers of respondents. The shadow bars in the figure show the size of respondents' sample for the specific use case, currently using or evaluating them. The top three use cases are the ones that are in common to all industries, risk exposure assessment, new product development and price optimization. However, looking at responses by industry (table below), we can observe that some industries do not consider all three of them as top three use cases as they prioritize other activities. But considering the total sample, they became highly relevant. Broadening the analysis to the top 5 use cases, a mix of internally and externally oriented use cases is presented. In the top five we observe automated customer service as one of the most deployed use case and when considering only the use cases that are currently in use (light blue bar), it is evident that customer understanding, targeting and optimization of offers is more relevant and falls in the top five.

Taking a closer look to the top 5 use cases we find common patterns across industries.

- Risk exposure assessment. This first use case is extremely relevant to all the industries, especially when they are in the process to evaluate current processes, services, and products, but also to evaluate the introduction of new products/services in the portfolio. When instead the business process is under evaluation, the risk exposure assessment can come into play too, providing information on opportunities and risks related to the new processes.
- New product development. New product (and service) development is progressed by the adoption of Big Data because it helps organizations to (re-)shape products (and services) according to customers needs and interests. This is also something linked with the personalization/customization of products and offers.
- Price optimization. Optimization of products and services' prices is a complex mechanism that can be undertaken only after having in place a well-functioning BDA platform, able to profile and target specific customers segments with tailored offers and prices.
- Regulatory intelligence. Big Data and BDTs are helpful in setting and managing the regulatory compliance strategy and in building a regulatory-savvy but data-centric company. The big data platform (or solution) helps organizations to capitalize on the potential and real value in ensuring the usage of data but following what can be defined as next-gen regulatory compliance. Technologies help in updating and modernizing the compliance process and simply follow the regulatory changes (within the origin country but also foreign regulations), improve decision making process and make the compliance more stringent and effective.

- Automated customer service. In automating the customer services, organizations can optimize the response to customer both in terms of timing – from call reception to handling and forecasting service completion – and costs. This process not only improve the customer satisfaction but also lower call handling costs, human errors, and human costs.

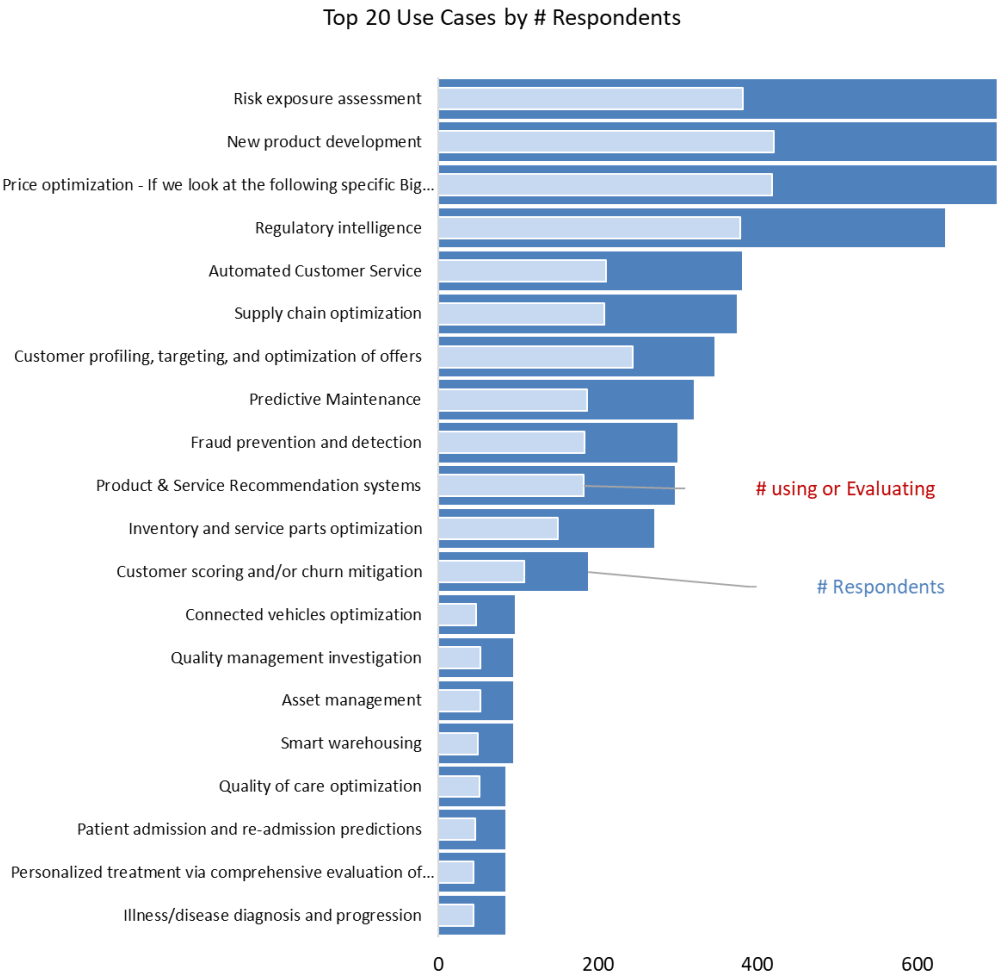

Figure 13 - Top 20 use cases by number of respondents.
Source: DataBench D2.2 "Preliminary Benchmarks of European and Industrial Significance"

Big Data is considered by all industries as a pivotal solution toward Digital Transformation and the achievement of digital business objectives. With the increasing volume, variety and velocity of data, coming from multiple source and carrying information from both internal business processes and external data sources, there is a valid opportunity to exploit this data to gain better understanding on current performances and improvement areas.

The exploitation of Big Data is helpful for a large number of use cases, embracing both internal and external processes, such as optimizing conversion rates, detecting and avoiding risks, streamlining operation, monitor customer behaviour, etc.

Enhancing the decision-making process is an ongoing effort for many organizations and big data and analytics can contribute toward achieving business goals and profitable results. Big Data adoption varies according to specific vertical markets' goals, meaning that each sector implements big data according to its own specific purpose and business model.

3.5 Case Studies

Case study analysis is an important goal of WP4. Deliverable D4.2, *Data Collection Results*, provides a detailed description of the methodology that has been used in case study analysis and of the status of research in WP4, drawing preliminary conclusions, while the final research results are presented in D4.3, *Evaluation of Business Performance*. In the scope of the DataBench project, we have collected more than 700 articles, gathered from three main types of sources:

- the scientific literature,
- European research projects (including ICT 14-15 projects),
- customer success stories of the most important BDT providers.

Each of these articles was tagged with different metadata, e.g., the magnitude of data size, the velocity, the type of sources. These metadata have been thoroughly discussed in D4.2 and are reported here in Figures 2 and 3, for the sake of clarity.

In D4.1, *Data Collection Plan*, and D4.2, *Data Collection Results*, we have defined a methodology for the analysis of case studies. The depth of the analysis depends on the case study, on the outcome of the first interview and on the openness to discussion and cooperation of the different companies. Case studies involve a considerable effort and, as a consequence, the goal is not to reach statistical significance and generality per se, but to provide qualitative, insightful explanations to findings from extensive surveys (such as the DataBench survey and the desk analysis) as well as indications for subsequent research.

We have performed a total of 22 case studies distributed across 8 industries and 7 countries. Figure 14 shows the companies that have participated in the DataBench case-study analysis. All companies have gone through the first interview, 15 have provided documentation, 9 have accepted to perform a second interview and 6 have provided data and involved the DataBench team to be supported in their decision processes. Not all companies have consented to disclosing the information that they have shared, 3 have requested to remain anonymous (see Figure below).

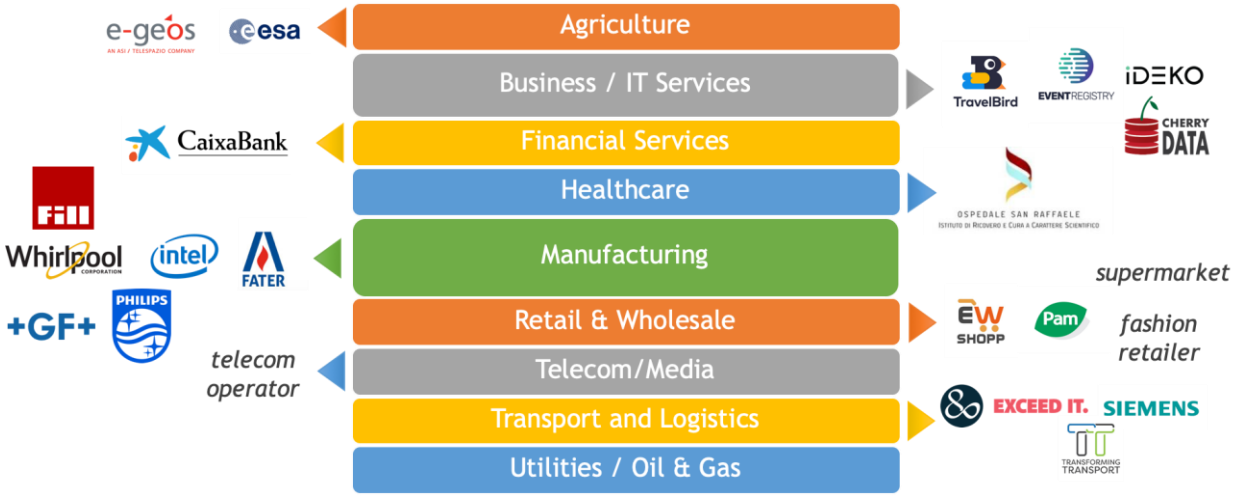

Figure 14 - Case Studies (total 22).
Source: DataBench deliverable D4.3 "Evaluation of Business Performance"

From the analysis of the case studies, we can identify some commonalities across the entire BDT space.

- **A new approach to business intelligence and decision-making**: with prescriptive analytics, decision-making can be delegated to machines whenever machines are recognized to have better performance and obtain greater business benefits. Decisions supported by an accurate prediction are usually better than decisions supported by descriptive statistics only. Hence, BDA's ultimate goal is the automation of decision making. This is seen as both an opportunity – automation of manual work – and a threat – skeptical about superiority of machines in decision-making.
- **High awareness for opportunities**: There is a significant interest in learning about the most frequent use cases in different industries and in taking part to projects to gain market competitive insights and to obtain tools to benchmark performances for the industry.
- **Data is the starting point**: from a technical standpoint, data- and processing-intensive analytics require a dedicated database, with its own hardware, data management technologies and data design. Setting up this infrastructure requires time and a significant investment but represents and enables subsequent application-level implementations. Companies have accepted the idea that they have to make an initial investment, choose a technology stack and create a so called 'data lake' to store their data. What can be subsequently done is reaching a certain level of data quality, complex both to implement and reach. So, business benefits from BDA do not seem to be low-hanging fruits, but rather the outcome of a change process that stats from the data itself.
- **Data governance is a concern**: this is a common concern and represents a general concept, with organizational, technical, and legal implications. As governance means data security, this tight link can act as an obstacle but also as opportunity for technical innovation.
- **Only a few projects reach deployment stage** because of a mixed combination of technical, human, and organizational factors. Undoubtedly, the benefits associated with BDTs are observable even at early adoption stages and lack of business benefits is not a hindering motivation for the scarce technology uptake, and reasons have deeper roots.

Furthermore, from the case studies analysis we have evidence of business KPIs for a subset of our case studies. Overall, evidence from case studies is aligned with results from the DataBench survey and positions business impact is in the 4-8% range. Companies that have measured a positive business impact have all developed their own software, focusing on selected use cases with a practical approach. They had a clear view of the business issues to be tackled and of the potential benefits of advanced analytics techniques. Some of them are currently working on the large-scale deployment of their pilots, others have already reached full-scale deployment (more detailed information in DataBench deliverable D4.3 "*Data collection results*").

From the case study analysis, we have seen that it is important to make technical choices that can support long-term change in order to enable greater business benefits. From the evidence that has been collected so far from case studies, an important lesson learnt is that most companies believe that technical benchmarking requires highly specialized skills – skills that are not currently present in the company – and considerable investments. There

is a general agreement on the fact that BDTs are diverse and complex and that technical choices are not simple and are potentially impactful. Even if companies do not perform benchmarking activities, they have been found to rely on trusted external entities to compare technologies, such as IT consultants and system integrators.

4. Technical Benchmarks

4.1 Overview

The DataBench Framework for Big Data and AI Benchmarks is based on the BDVA (Big Data Value Association) Reference Architecture. In order to have an overall perspective on Big Data and AI systems the usage of a top level generic pipeline has been introduced.

The following figure depicts a top-level pipeline, following the Big Data and AI Value chain, that is abstract enough, so that it can be specialised in order to describe more specific pipelines, depending on the type of data and the type of processing (e.g. IoT data and real-time processing). This top-level pipeline contains the four major steps which are depicted in Figure 15 and analysed next.

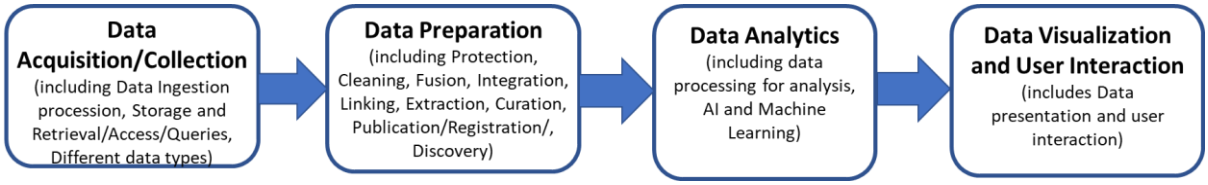

Figure 15 - Top Level Generic Pipeline

These steps are in compliance with the activities described in the Reference Architecture for Big Data Application.

A brief description of these steps follows next.

4.1.1. Data Acquisition/Collection

In general, this step handles the interface with the data providers and includes the transportation of data from various sources to a storage medium where it can be accessed, used, and analysed by an organization. Tasks in this step, depending on application implementations, include accepting or performing specific collections of data, pulling data or receiving pushes of data from data providers and storing or buffering data. The cycle Extract, Transform, Load (ETL)/Extract, Load, Transform (ELT) can also be included in this step. At the initial collection stage, sets of data are collected and combined. Initial metadata can also be created to facilitate subsequent aggregation or look-up methods. Security and privacy considerations may also be included in this step, since authentication and authorization activities as well as recording and maintaining data provenance activities are usually performed during data collection. Last, we would like to note that tasks in this step may vary, depending on the type of the collected data.

4.1.2. Data Preparation

Tasks performed in this step include data validation, like for example checking formats, data cleansing, such as removing outliers or bad fields, extraction of useful information, organization and integration of data collected from various sources, leveraging metadata keys to create an expanded and enhanced dataset, annotation, publication and presentation of the data in order to be available for discovery, reuse and preservation, standardization and reformatting, or encapsulating. Also, in this step, source data are frequently persisted to archive storage and provenance data are verified or associated. The transformation part of the ETL/ELT cycle could also be performed in this step, although advanced transformation is usually included in the next step which is related with data analytics. Optimization of data through manipulations, such as data deduplication and indexing, could also be included here in order to optimize the analytics process.

4.1.3. Data Analytics

In this step, new patterns and relationships, which might be invisible, are discovered so as to provide new insights. The extraction of knowledge from the data is based on the requirements of the vertical application which specify the data processing algorithms. This step can be considered as the most important step as it explores meaningful values, and thus, it is the basis for giving suggestions and making decisions. Hashing, indexing and parallel computing are some of the methods used for Big Data analysis. Machine learning techniques and Artificial Intelligence are also used here, depending on the application requirements.

4.1.4. Data Visualisation and User Interaction

Data can have no value without being interpreted. Visualization assists in the interpretation of data by creating graphical representations of the information conveyed, and thus adding more value to data. This is due to the fact that the human brain processes information much better when it is presented in charts or graphs rather than on spreadsheets or reports. Thus, visualization is an essential step as it assists users to comprehend large amounts of complex data, interact with them, and make decisions according to the results. It is worth to note that effective data visualization needs to keep a balance between the visuals it provides and the way it provides them so that it attracts users' attention and conveys the right messages.

In the following sections, the above steps of the top level pipeline are specialised based on the different data types used in the various project pilots, and are set up differently based on different processing architectures, such as batch, real-time/streaming or interactive. Also, with Machine learning there will be a cycle starting from training data and later using operational data.

4.2. DataBench Framework

The DataBench Framework is based on the structure of the BDVA Architecture Model – and is focusing on both vertical and horizontal benchmarks according to this model – further related to business-oriented benchmarks.

The industry-based use cases are analysed in order to derive examples and metrics that can be related to each of the Big Data types. The focus is on reusing and adapting the established benchmarks for structural data (BigBench, BigDataBench, TPC and others) and graph data/linked data (Hobbit I-IV and LDBC 1-3) and in particular on incorporating benchmark proposals related to Time series/IoT (Yahoo Stream Benchmark, RIoTBench, StreamBench and others) and also input from DataBench partners research benchmarks on streaming sensor data, ABench (UFRA) and SenseMark (SINTEF).

Similarly, there will be a focus on the data types of Image/Audio/Media and Text/NLP where also analytic and processing benchmarks for machine learning (DeepBench, DeepMark and others) are relevant. A final relevant area for vertical benchmarks is on the effect of technology support for data privacy and security. A set of projects related to how to support data privacy has been started under the Big Data PPP ICT18 and a benchmark approach for analysing and understanding the use of these techniques has been requested from the user community.

The vertical dimension is based on benchmarks according to the following Big Data types:

- Structured Data Benchmarks
- IoT/Time Series Benchmarks
- SpatioTemporal Benchmarks
- Media/Image Benchmarks
- Text/NLP Benchmarks
- Graph/Metadata Benchmarks

The BDV Reference Model1 shown in Figure 16 has been developed by the BDVA, taking into account input from technical experts and stakeholders along the whole Big Data Value chain as well as interactions with other related PPPs. An explicit aim of the BDV Reference Model in the SRIA 4.0 document is to also include logical relationships to other areas of a digital platform such as Cloud, High Performance Computing (HPC), IoT, Networks/5G, CyberSecurity etc.

The BDV Reference Model may serve as common reference framework to locate Big Data technologies on the overall IT stack. It addresses the main concerns and aspects to be considered for Big Data Value systems.

[1] http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf (page 37)

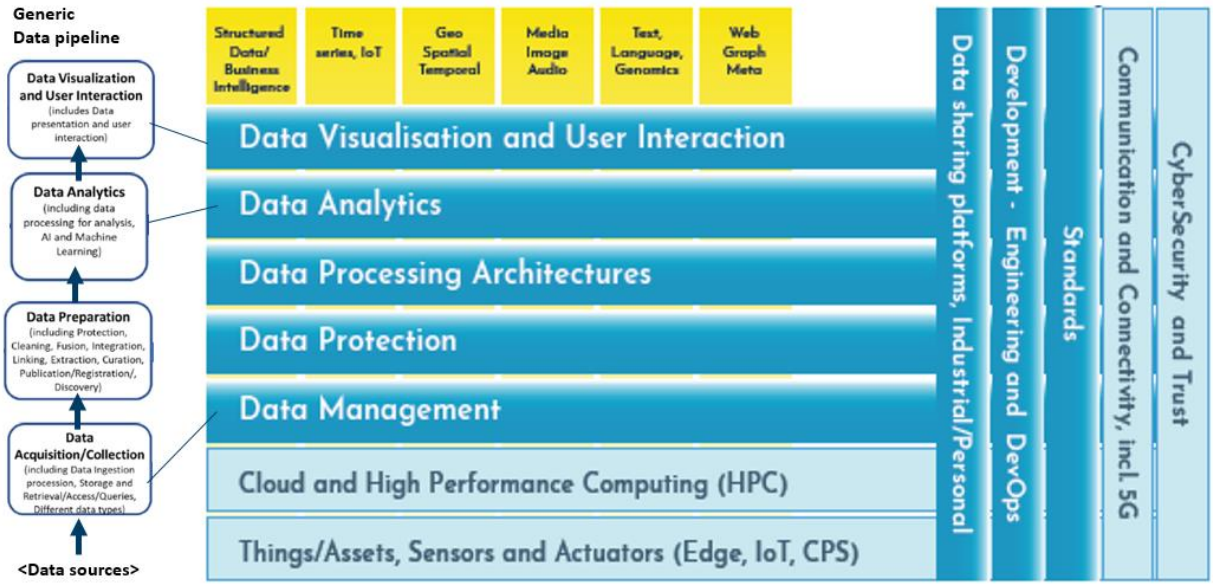

Figure 16 - BDV Reference Model as a foundation for the DataBench Framework – related also to the Generic pipeline

The BDV Reference Model is structured into horizontal and vertical concerns.

- **Horizontal concerns** cover specific aspects along the data processing chain, starting with data collection and ingestion, reaching up to data visualization. It should be noted, that the horizontal concerns do not imply a layered architecture. As an example, data visualization may be applied directly to collected data (data management aspect) without the need for data processing and analytics. Further data analytics might take place in the IoT area – i.e. Edge Analytics. This shows logical areas – but they might execute in different physical layers.
- **Vertical concerns** address cross-cutting issues, which may affect all the horizontal concerns. In addition, verticals may also involve non-technical aspects (e.g., standardization as technical concerns, but also non-technical ones).

Given the purpose of the BDV Reference Model to act as a reference framework to locate Big Data technologies, it is purposefully chosen to be as simple and easy to understand as possible. It thus does not have the ambition to serve as a full technical reference architecture. However, the BDV Reference Model is compatible with such reference architectures, most notably the emerging ISO JTC1 WG9 Big Data Reference Architecture – now being further developed in ISO JTC1 SC42 Artificial Intelligence.

The following technical priorities as expressed in the BDV Reference Model are elaborated in the remainder of this section:

4.2.1. Horizontal concerns

- **Big Data Applications**: Solutions supporting Big Data within various domains will often consider the creation of domain specific usages and possible extensions to the various horizontal and vertical areas. This is often related to the usage of various combinations of the identified Big Data types described in the vertical concerns.
- **Data Visualisation** and User Interaction: Advanced visualization approaches for improved user experience.

- **Data Analytics**: Data analytics to improve data understanding, deep learning, and 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. It also has links to trust mechanisms like Blockchain technologies, smart contracts and various forms for encryption. This area is also associated with the area of CyberSecurity, Risk and Trust.
- **Data Management**: Principles and techniques for data management including both data life cycle management and usage of data lakes and data spaces, as well as underlying data storage services.
- **Cloud and High Performance Computing (HPC):** Effective Big Data processing and data management might imply effective usage of Cloud and High Performance Computing infrastructures. This area is separately elaborated further in collaboration with the Cloud and High Performance Computing (ETP4HPC) communities.
- **IoT, 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.

4.2.2. Vertical concerns

- **Big Data Types and semantics**: The following six Big Data types have been identified – based on the fact that they often lead to the use different techniques and mechanisms in the horizontal concerns, which should be considered, for instance for data analytics and data storage: *1) Structured data; 2) Times series data; 3) GeoSpatial data, 4) Media, Image, Video and Audio data; 5) Text data, including Natural Language Processing data and Genomics representations; 6) Graph data, Network/Web data and Meta data*. In addition, it is important to support both the syntactical and semantic aspects of data for all Big Data types.
- **Standards**: Standardisation of Big Data technology areas to facilitate data integration, sharing and interoperability.
- **Communication and Connectivity**: Effective communication and connectivity mechanisms are necessary for providing support for Big Data. This area is separately elaborated further with various communication communities, such as the 5G community.
- **Cybersecurity**: Big Data often need support to maintain security and trust beyond privacy and anonymisation. The aspect of trust frequently has links to trust mechanisms such as blockchain technologies, smart contracts and various forms of encryption. The CyberSecurity area is separately elaborated further with the CyberSecurity PPP community.
- **Engineering and DevOps**: for building Big Data Value systems. This area is also elaborated further with the NESSI (Networked European Software and Service Initiative) Software and Service community.

- **Data Platforms**: Marketplaces, IDP/PDP, Ecosystems for Data Sharing and Innovation support. Data Platforms for Data Sharing include in particular Industrial Data Platforms (IDPs) and Personal Data Platforms (PDPs), but also include other data sharing platforms like Research Data Platforms (RDPs) and Urban/City Data Platforms (UDPs). These platforms include efficient usage of a number of the horizontal and vertical Big Data areas, most notably the areas for data management, data processing, data protection and CyberSecurity.
- **AI platforms**: In the context of the relationship between AI and Big Data there is an evolving refinement of the BDV Reference Model – showing how AI platforms typically include support for Machine Learning, Analytics, visualization, processing etc. in the upper technology areas supported by data platforms – for all of the various Big Data types.

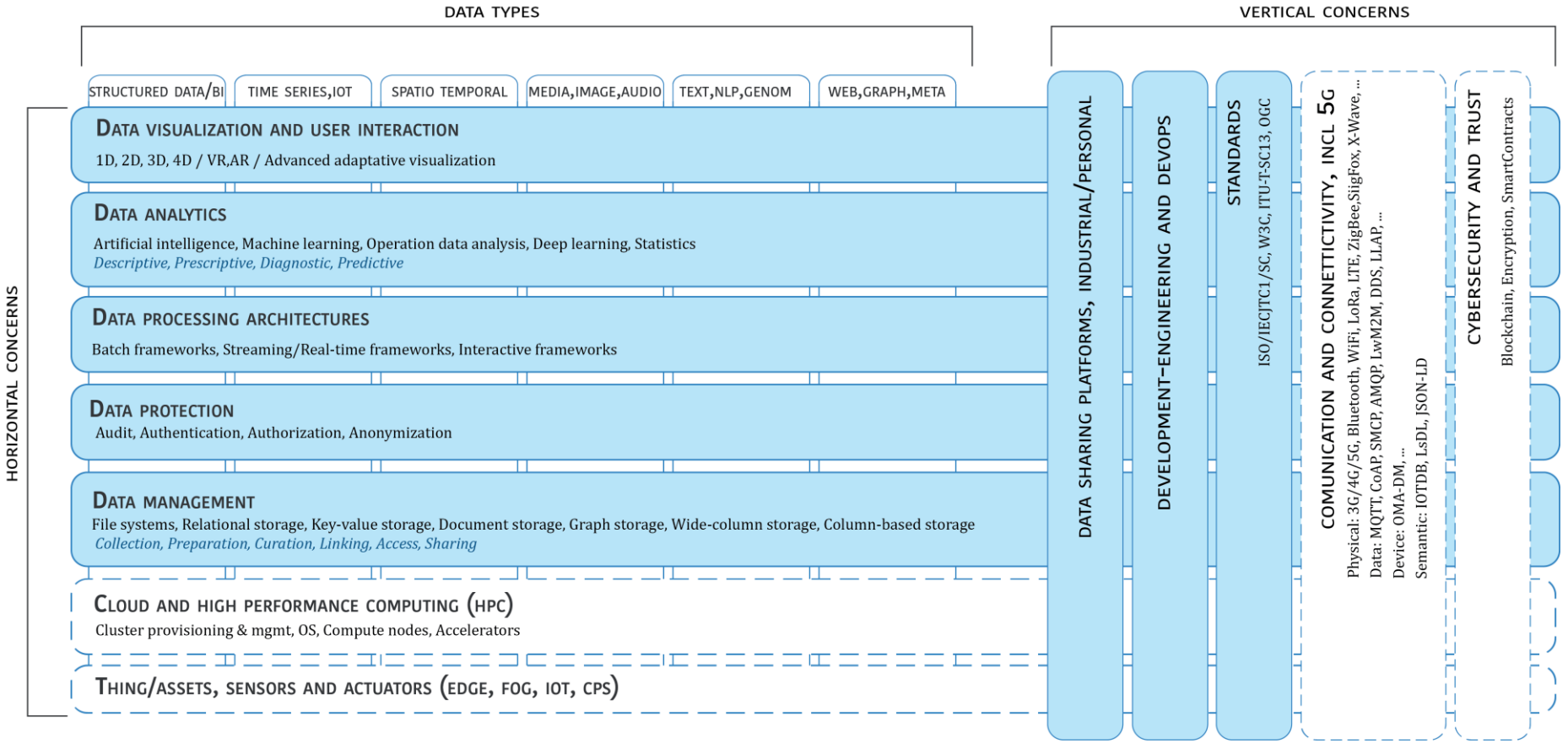

Figure 17 - Refinement of the BDVA Reference Model

The BDV Reference Model may serve as common reference framework to locate Big Data technologies on the overall IT stack. BDV Reference Model is compatible with such reference architectures, most notably the ISO JTC1 WG9 Big Data Reference Architecture which now has become part of the ISO SC42 AI (and Big Data) standard ISO 12345 XX.

The refinement of the BDVA Reference Model has been based on defining sub-categories within each of the reference model areas based on the refinement of the respective areas in the ISO SC42 suite of standards and technical reports currently in progress. The sub-categories describe typical technology types within each of the areas, relevant in benchmarking context.

The modeling approach in the figure is on the top level to describe logical technical areas within a wider Big Data and AI platform, and within each of the areas, relevant subcategories within this area. In addition to technical subcategories it has also been identified typical process steps in a Big Data pipeline relevant for the various areas. Work has started to consolidate and unify the models, metamodels and ontologies from D1.1, D3.1, D5.1 and D1.2 and the companion D1.3 and D1.4 public deliverables.

Data Visualization and User Interaction Layer:

This layer incorporates the research areas related to science of analytical reasoning assisted by advanced visualization and user interaction approaches. Major concern areas include:

Visual data discovery_ Proactive extraction of relevant information through visual data discovery techniques.

Interactive visual analytics of multiple scale data_ Facilitating empirical search for acceptable scales of analysis and the verifications of results.

Collaborative, intuitive and interactive visual interfaces_ Exploiting advanced discovery aspects of Big Data Analytics to enable collaborative decision-making processes. Carefully designed presentations and digital visualizations (including zooms, dynamic filtering, annotation*)* for quick and correct interpretation of data, Focus on relevance and relatedness of information for efficient search and exploration.

Cross-platform mechanisms for data exploration, discovery and querying_ Uniform data visualization on a range of devices.

Innovating reporting_ Innovative multi-device reports and dashboards (including dynamic, 3D, augmented-reality dimensions, etc.).

Domain-specific data visualization techniques_ Innovative techniques and approaches to visualize data coming from specific domain (e.g. graphs, geospatial, sensor, mobile data, etc.).

4.2.3. DataBench Pipeline, Framework and available Benchmarks

The DataBench Framework is based on the structure of the BDVA Architecture Model – and is focusing on both vertical and horizontal benchmarks according to this model – further related to business-oriented benchmarks. This is illustrated in Figure 18 below.

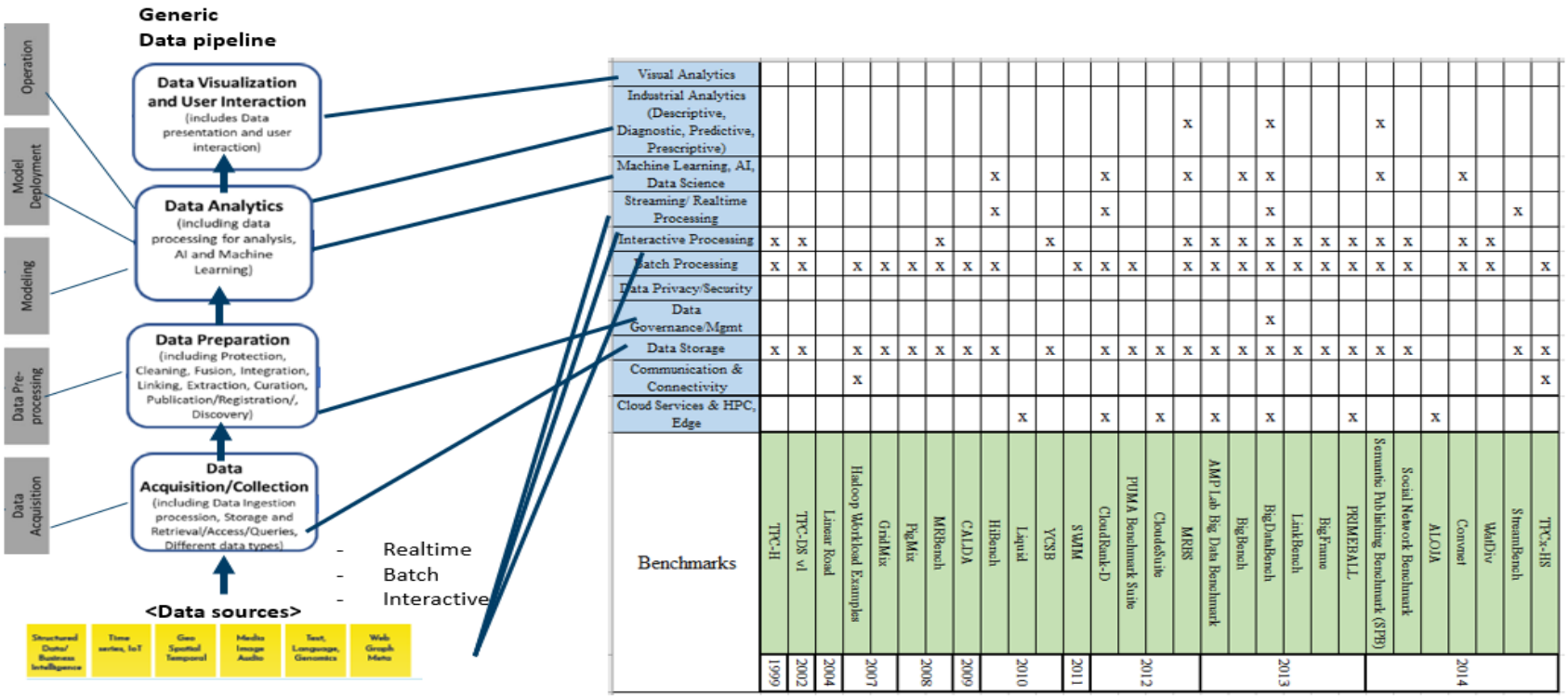

Figure 18- Refinement of the BDVA Reference Model

4.2.4. Examples of relation Generic Pipelines

The following Figures provide example of the generic pipelines used.

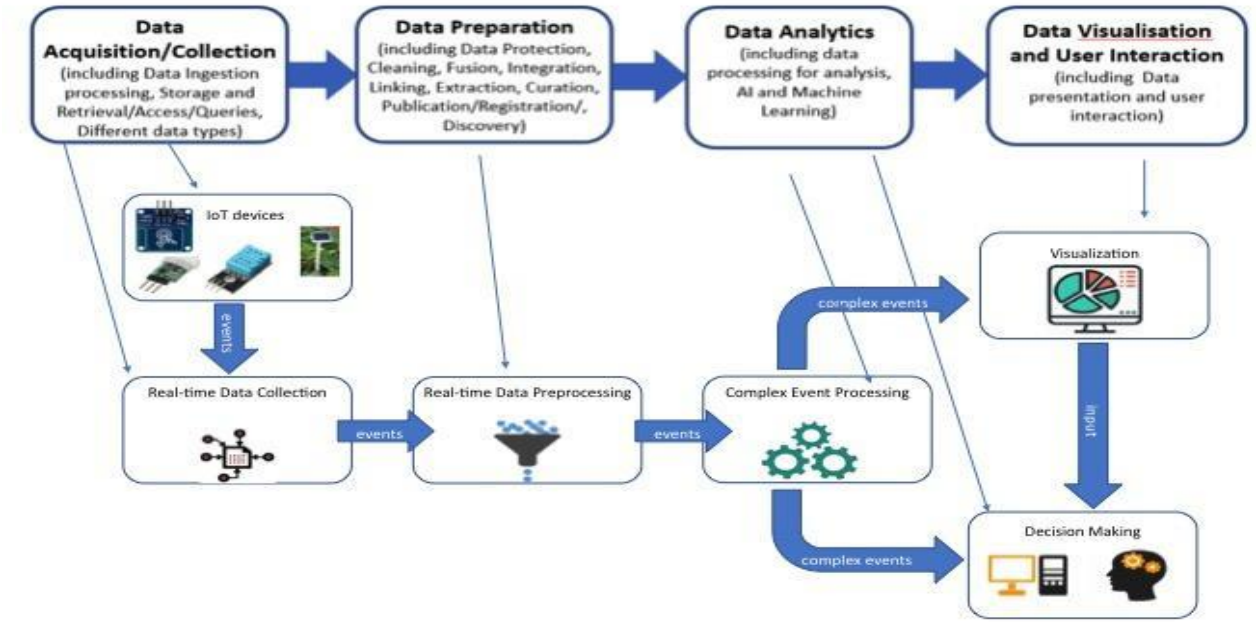

Figure 19 - Example of IoT pipeline pattern

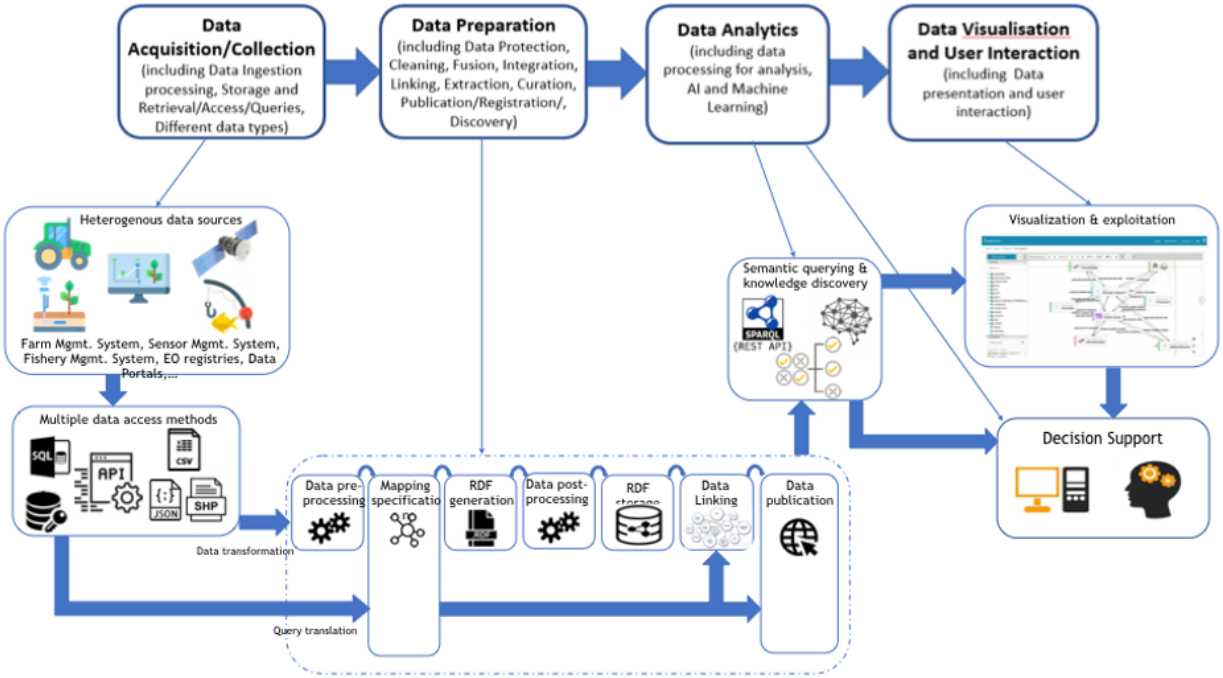

Figure 20 – Example of Graph/Linked Data pipeline pattern

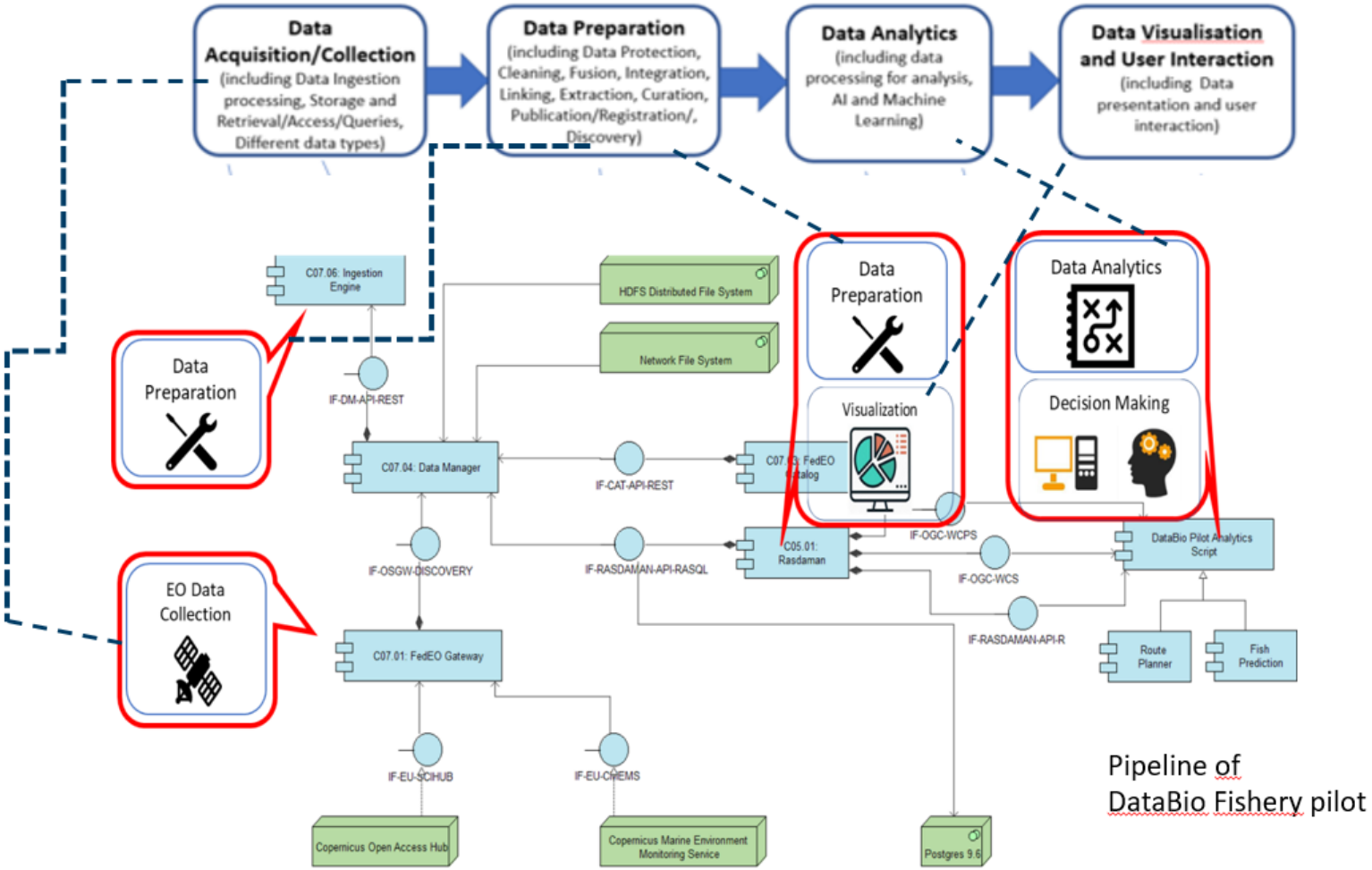

Figure 21 – Example of system pipeline from DataBio project pilot

5. Business Benchmarks

5.1 Overview: from KPIs to Benchmarks

The business KPI definitions are based on business and marketing literature, but these definitions have been simplified and operationalized to allow measurement through business surveys. This approach is one of several options for the measurement of technology business impacts, an approach chosen for its applicability to an objective of the project – namely, the need to estimate business-impact-related industrial benchmarks that are valid for European industry and differentiated by sector and company size. The data collection process is illustrated in the next paragraph.

Since IDC focuses on emerging technologies and market forecasting, we have developed a methodology based on business surveys that enables us to collect data about the overall average impacts of technology investments based on companies' own evaluations. Since companies do not carry out investments without an economic or business rationale, this data has a sound basis, even though it is technically a result of the opinions of respondents. To make sure these opinions are valuable, and fact based, we have employed several methods, including:

- The careful selection of the role and responsibility of the survey respondent (who must have the relevant knowledge)
- The careful quality control of survey data, discarding incoherent and unbelievable answers, as well as the careful management of the survey itself (for example, rotating answer options so that no ranking bias exists)
- Statistical elaboration techniques, discarding outliers and extreme values, by checking the maximum and minimum data points
- Long experience in survey management and a reliance on experienced and well-known interviewers
- Comparative analysis of the resulting data with literature and other sources about the business impacts of technology innovation

All these methods have been employed in this project to define and collect data about the business impacts of BDA and to calculate industrial benchmarks. Table 1 and Table 2, below, provide details of each KPI, its metrics, and the measurement results.

Thanks to our methodological approach, the business KPIs selected by the project are valid metrics and can be used as benchmarks for comparative purposes by researchers or business users for each of the industry and company-size segments measured. These indicators are:

- Benchmarks, because they represent the average improvement achieved by business users and can be used for comparative purposes, as a target or as a best performance metric.
- Of industrial significance, because they apply to the actual and emerging needs of specific industries and specific company-size segments.

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

- Of European economic significance, because the benchmarks are measured for all the relevant European industries and company-size segments in which Big Data can have the highest impacts.
- Useful for linking technical and business benchmarking, because they are also measured for the main use cases, consisting of the application of Big Data technology to particular business processes and/or application domains, thus enabling the user to match the expected business improvements with the type of technology performance needed to achieve the business goals.

We differentiate the KPIs into two different categories, according to how they are evaluated and what they measure. The first batch measures the quantitative business KPIs, which are profit increase, revenue increase and cost reduction; while the second batch is composed by soft qualitative KPIs that cannot be easily and straightforwardly calculated, so a range of expected improvements is provided. Within this second set of KPIs, we include time efficiency, product/service quality, customer satisfaction, new products/services launched, and business model innovation.

KPI	Definition	Metrics
Revenues increase	Increase in company revenues thanks to the adoption of BDA	**Quantitative benchmarks**: % increase measured as median of the sample
Profit increase	Increase in company profit thanks to the adoption of BDA	
Cost reduction	Reduction in process costs thanks to the introduction of BDA	
Time efficiency	Efficient use of time in business processes	**Qualitative benchmarks:** average rating on a scale of 1–5 based on the following ratings: • Less than 5% improvement = 1 • 5–9% = 2 • 10–24% = 3 • 25–49% = 4 • 50% or more = 5
Product/Service quality	Product/Service features corresponding to users' implied or stated needs and impacting their satisfaction	
Customer satisfaction	A measure of customers' positive or negative feeling about a product or service compared with their expectations	
New Products/ Services launched	A measure of the number of new products and/or services enabled by data-driven innovation and launched by the company after engaging in the Big Data investment	
Business model innovation	Novel ways of mediating between companies' product and economic value creation (for example, moving from traditional sales to service subscription models)	

Figure 22 - Benchmarks overview.
Source: DataBench D2.4 deliverable "Benchmarks of European and Industrial Significance"

5.3 Business Benchmarks by Industry

This chapter describes the BDT business benchmarks by industry, which are presented by the Toolbox in the Knowledge Nugget section through the user interface. The comments below provide some background about the significance of the business benchmarks by sector. More detailed benchmarks by use case are provided in D.2.4 *Benchmarks of European and industrial significance.*

5.3.1 Agriculture

As charted below, it is possible to see that organizations in the agriculture sector are evaluating big data solutions in terms of profit increase as currently relegated to small margins. Increasing the precision of the agriculture and yield prediction could boost profit and broad margins. In having a precise view on production, farmers can optimize the use of

lands and seeds, pushing to the full exploitation of their capabilities. To be noted, investments as such can be mostly done by larger agricultural organizations than smaller ones, but with technology's dropping prices, even smaller users will be able to access these technologies. The ability to plant seeds in an optimal/efficient way (precision agriculture) – distancing seeds according the potential plant growth -, controlling the productivity of seeds and soil and forecasting it (yield monitoring and prediction), and predictive maintenance of the machineries and tractors will help organizations to better organize activities within the year (exact date to plant and harvest, optimization of crop rotation, schedule the maintenance, etc.) will improve revenues stream and also bolstering margins. Along this reasoning, we also find that among the qualitative KPIs, product/service quality and time efficiency, play a relevant role. Being able to provide better, on time and more nutrient products will raise the bargaining power in setting prices of yields. Hence, farmers are not so much interested in finding new sources of revenue or inventing completely new services, but they are expecting increased value from data changing traditional, old-fashioned and gut feeling based business processes.

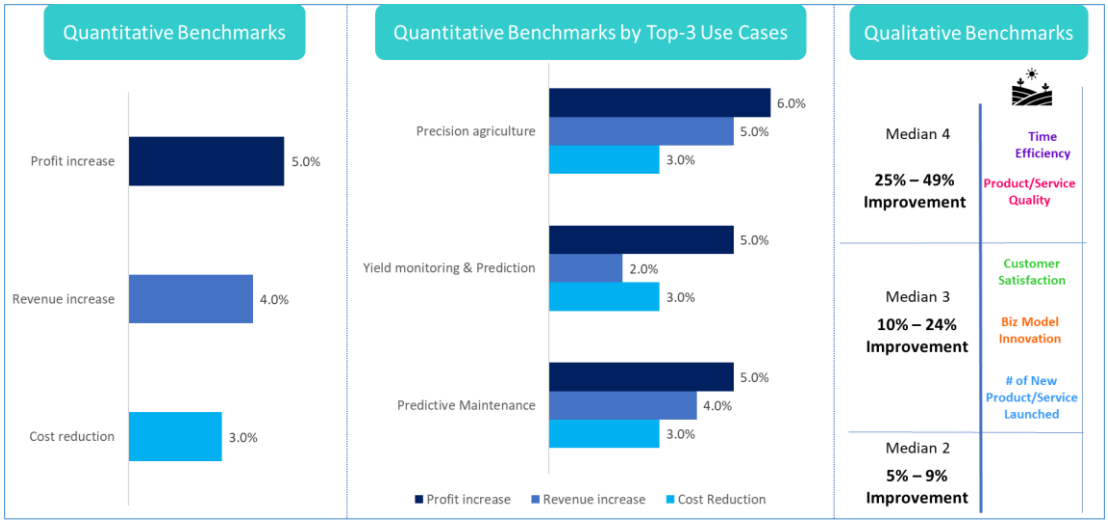

Figure 23 - BDT Benchmarks: Agriculture
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.2 Financial Services

The financial service industry (banks and insurance companies) is a traditionally data abundant sector, which has always tried to play this ace to be a leader in technical innovation. On top of this, customer satisfaction is a real competitive differentiation between the variety of services providers playing in this sector, from high street banks, to insurance companies, FinTechs, or investment management providers.

Looking at the top two use cases, customer scoring, and/or churn mitigation (ability to predict whether a customer is more likely to drop off the service or just to score the worthiness of a customer) and customer profiling, targeting, and optimization of offers, it is clear that the main outcome of these two processes, if carried out properly, is to increase revenues and profits. While the third top rated use case, fraud prevention and detection, hence the ability of a bank or insurer to predict if a new customer is a potential fraudster or a specific transaction is legal, is more likely to be related to cost reduction rather than to profit and margins increase. This is because being able to quickly assess potential scums, avoiding false positive results, helps organizations in reducing money losses. In analyzing the qualitative benchmarks, the one that plays a more relevant role is customer satisfaction,

which is again in line with the top two use cases for the financial industry. Following the same reasoning, in increasing the customer satisfaction, organizations see a higher customer stickiness (hance a lower probability to churn out); on top, a highly satisfied customer is also more willing to provide and share additional data (also thanks to the length of the loyalty and service use), sharing with organizations more data on which they can build targeted strategies creating opportunities for up-selling and cross-selling activities.

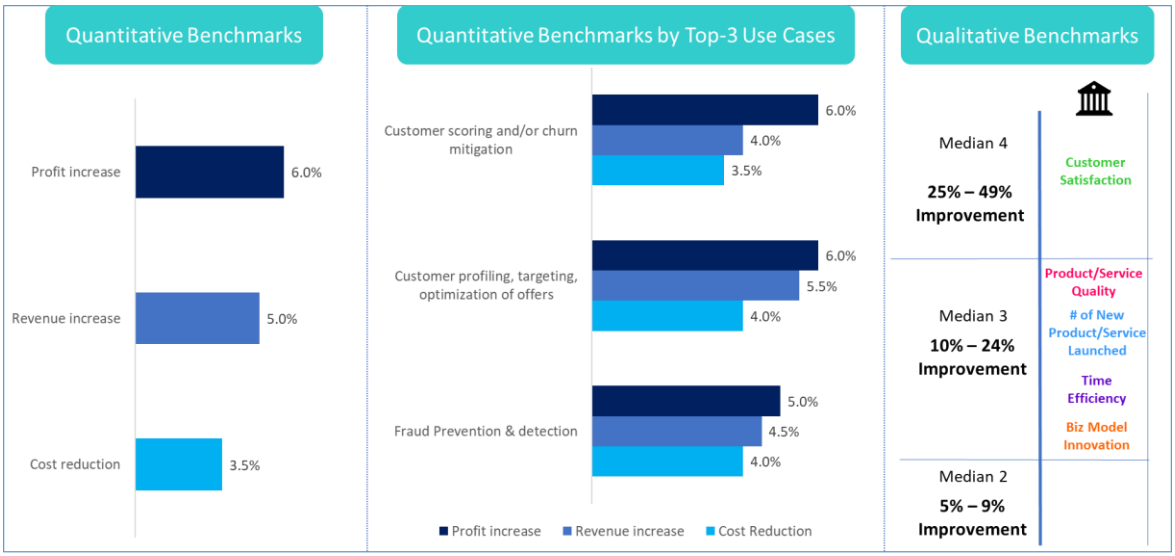

Figure 24 - BDT Benchmarks: Financial Services
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.3 Business/IT Services

Business/IT services sector is the traditional leader in big data adoption and exploitation showing large increase in profits. This is true when analyzing the most rated use cases. For instance, we can observe that risk exposure assessment (evaluate the potential future loss of an activity or event) and customer profiling, targeting and optimization of offers present large profit increase (respectively 7% and 6%). While customer profiling and new product development use cases are activities and use cases strictly related to profit and margins increase, links to profit increase and risk exposure assessment need to be deeply considered. Risk exposure assessment is an activity that per se seems to be related to cost reduction, but as it deals with potential losses (and not real ones), this cannot be considered a cost reduction related activity. In exactly forecasting the risk related to an activity, the organization can decide whether to follow up on it or simply drop it; in performing this, the organization is able to look up and choose only (or mostly) more profitable activities, disregarding high-risk ones. Analyzing the qualitative benchmarks, it is straightforward understanding why customer satisfaction and product/service quality are the KPIs showing the larger improvement from BDA adoption. Customer satisfaction is strictly related both to profit increase and the interest in the customer profiling, targeting and optimization of offers activities; while product/service quality delivers larger customer satisfaction and links with new product development.

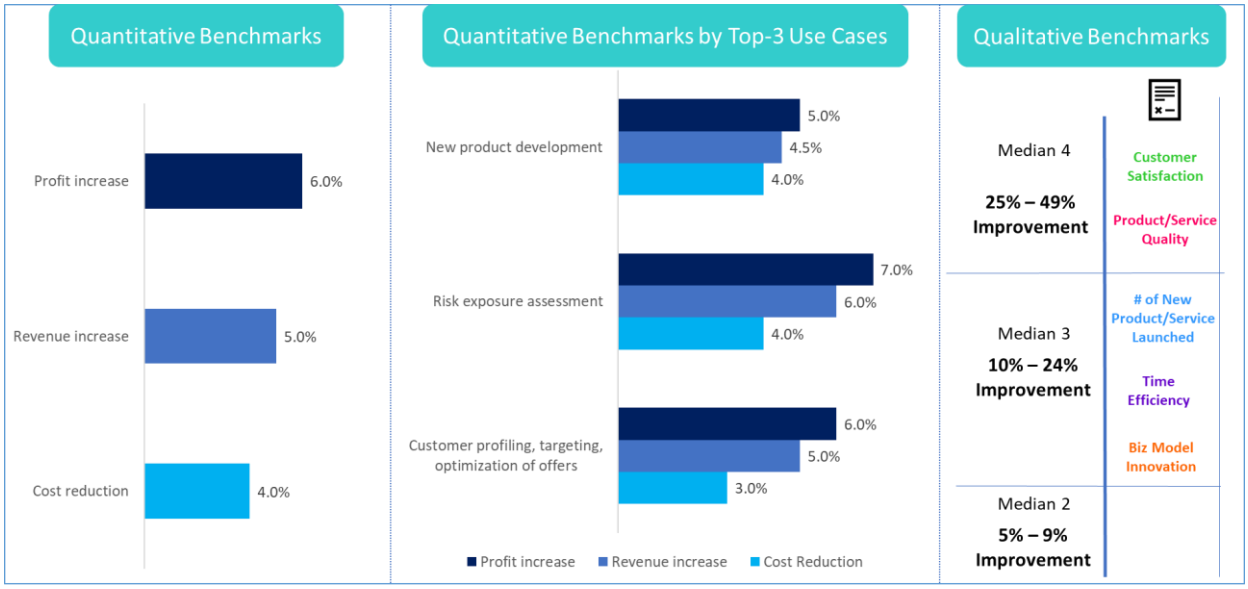

Figure 25 - BDT Benchmarks: Business/IT services
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.4 Healthcare

The healthcare sector is characterized by a long standing data abundance but up to now, this invaluable source of data hasn't been used due to the high sensitivity of this data, aside the privacy issues in analyzing it due to the potential re-identification of patients. Despite the scarce usage, the sector sees a wide range of beneficial applications for BDA adoption both strictly related to patients (quality of care optimization, illness/disease diagnosis and progression, etc.) but also more general ones related to regulatory intelligence (acting upon the understanding of regulation to legally treat data) and fraud prevention and detection. Within quality of care optimization there are several different sub use cases, from availability of hospital beds, to management of treatment's slots, to resource allocation – people and technical sources). Within this context, the increased availability of data, computational resources and the current development of privacy preserving technologies, can potentially and largely improve resource usage and optimization. As often healthcare services and structures are government owned, cost reduction was the only relevant KPI in the past, with scarce interest in increasing profits and margins, but with shrinking budgets also the potential revenue stream creation is essential to the healthcare sector. Among the most relevant use cases, and bridging use cases with KPIs, we find quality of care optimization. This use case is largely evaluated by profit increase because the quality of care is both related to patients' satisfaction (like customer satisfaction) but also to optimization of services and avoiding waste of important resources. In addition, the ability to predict machinery faults/maintenance helps a better management of the general infrastructure. In spending some time on the understanding of the qualitative KPIs, it is clear that there is no KPIs more relevant than the others in assessing BDTs usage, as they are all relatively relevant.

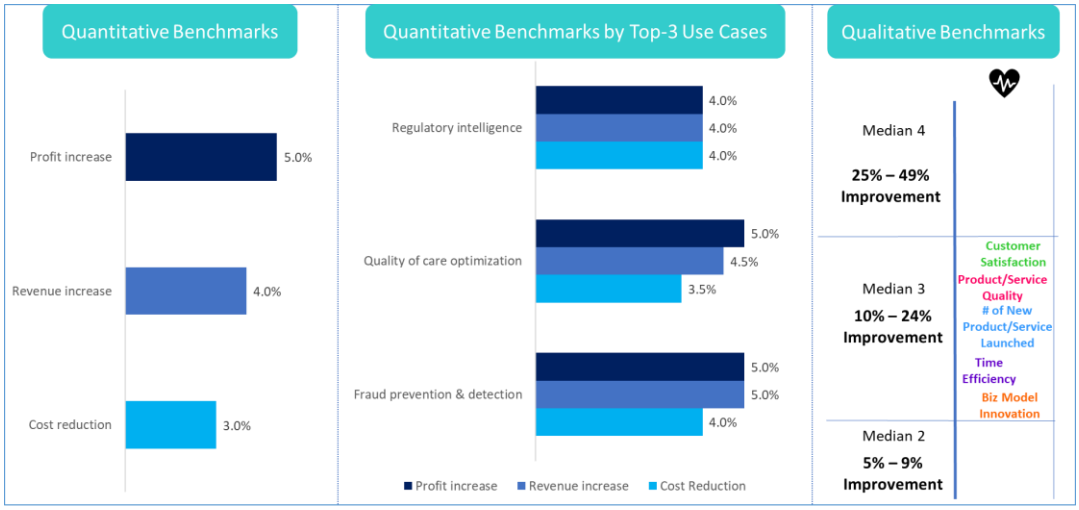

Figure 26 - BDT Benchmarks: Healthcare
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.5 Manufacturing

Manufacturers play in a sector that can be considered traditionally data abundant, with batch data from existing sources such as traditional IT systems – inventory, production, sales, etc. – but also new streams of data – with the increasing importance of IIoT (Industrial Internet of Things).

The manufacturing sector, despite being wide and embracing a large variety of activities, identifies the same common needs: which are supply chain optimization – a topic that has never been more relevant than today, when during the early days of the pandemic worldwide supply chains collapsed -, predictive maintenance, and product development. The first two use cases can be both assessed by strong cost reduction, but they also deliver relevant increased profits. In optimizing the supply chain, in broadening and consolidating partnerships, and in strengthening the ecosystem, manufacturers are able to promptly respond to unexpected crises, maintaining resiliency and production up to speed. In avoiding disruptions of the supply chain, organizations can maintain unaffected the production level and hence keep profitability and high margins. Predictive maintenance – having a platform that forecasts the exact moment when a machinery needs to be stopped for maintenance, and extending this concept also predictive failure – helps organizations to properly allocate maintenance hours with little impacts on the production schedules and lines. The ability to maintain production up to speed helps organizations in delivering higher margins and profits. The third most important use case is new product development and it is assessed with profit increase. The use of big data to understand customers' needs and desires and translating them into new products or upgrade of existing lines of production strengthen the organizations possibility to achieve larger profits. Summarizing on the qualitative benchmarks, there is no specific benchmark that is valued as more relevant than the others. However, business model innovation is perceived less important as benchmark in BDA activities. This is because right now big data is fairly exploited for other – more relevant activities – and some business model innovation activities (i.e. data monetization) are still currently under evaluation by manufacturers.

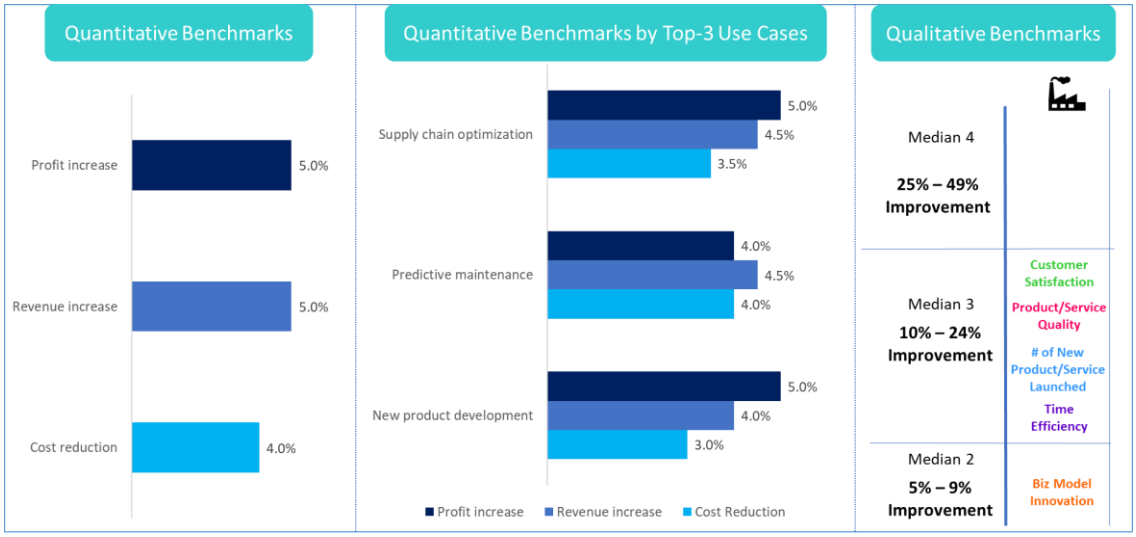

Figure 27 - BDT Benchmarks: Manufacturing
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.6 Retail & Wholesale

The retail and wholesale sector is another sector with abundance of data. It has been one of the first sectors in adopting BDA solutions to leverage the potential of data. On a general level, profit increase is the preferred benchmark, both in general terms but also when analyzing specific use cases. Regarding the top 3 use cases, they are all preferably evaluated with profit increase as main benchmark. The optimization of the supply chain (and logistics behind it) can lead to relevant profit and margins increase. In creating a more transparent and interconnected view of the supply chain, retailers and wholesalers can potentially avoid disruptions, and in applying BDTs to supply chain data it is easier to forecast and predict potential issues and timely act upon them. Price optimization is mainly regarded in terms of profits increase and the ability to properly segment customers and target them with different prices, helps organizations to leverage on margins to increase profits. The other top use case, new product development, has also a connection with the preferred qualitative benchmark, product/service quality. Retailers and wholesalers are interested in creating new products to enlarge profits and improve the quality of the products and related services offered. The quality of already existing products matched with the quality of the service offered (delivery, customer care, insurance, etc.) increase customer fidelity. With increased fidelity the organization can collect further information and data on the customer to service him/her with new products (somehow related also to cross-selling and up-selling) boosting revenues. Other qualitative KPIs are all somehow regarded as moderately important with a good range of improvements, but still not as high as product/service quality.

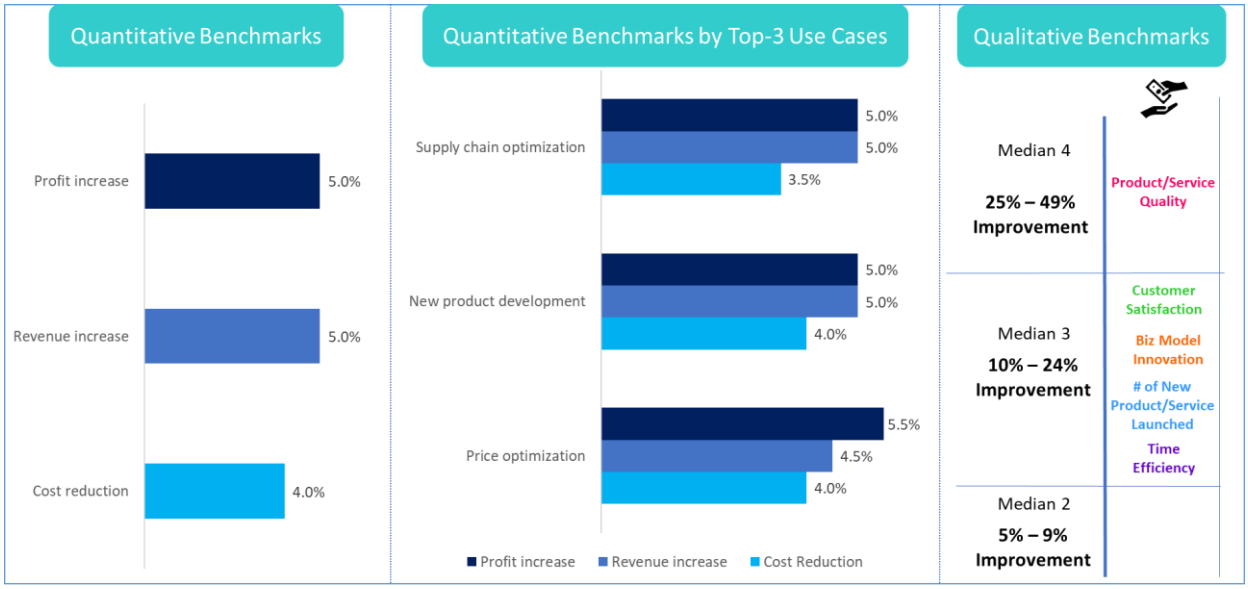

Figure 28 - BDT Benchmarks: Retail/Wholesale
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.7 Telecom/Media

Telecommunication (and media) sector is facing dramatic challenges in consumer choices. Now more than ever before customers have a little loyalty to these services providers and are able and willing to switch providers on the fly. The quality of the service and the customer satisfaction in this challenge play a pivotal role, and this explains why these two are the most important qualitative KPIs used by telecommunication providers and media companies. In this volatile context, real-time and advanced analytics are critical to provide excellent service quality and high-quality customer service (increasing customer satisfaction and loyalty), addressing technical issues in an optimal manner and proactively interacting with customers. This is translated from the most important qualitative KPIs also into the relevance of use cases. Among the top three use cases for telecommunication providers and media companies, we find automated customer service and customer profiling, targeting, and optimization of offers. Both the two use cases are valued with large increase of profits (both 6%) and revenues. The top use case, product and service recommendation system is valued in terms of profit increase and is linked to customer satisfaction. With this use case, semi-automated systems recommending specific products/services to customers and users according to their profile and needs, it is possible to leverage on customer satisfaction and customer stickiness to boost profits. Concluding on telecommunication and media, the most impactful KPI is profit increase.

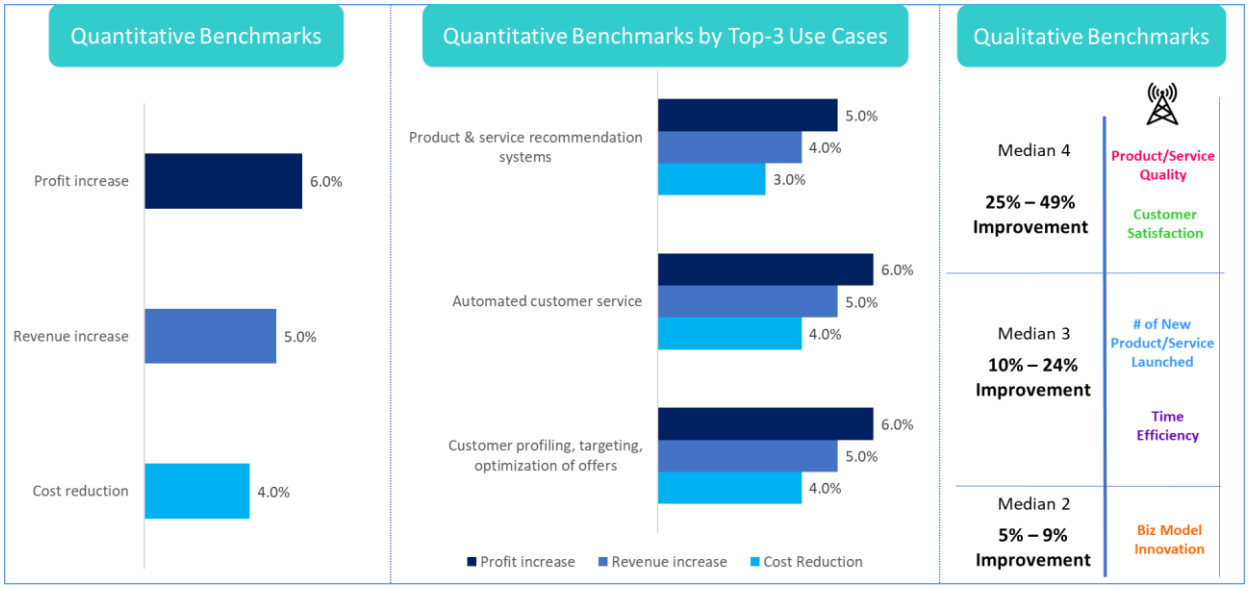

Figure 29 - BDT Benchmarks: Telecom/Media
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.8 Transport/Logistics

Transport/logistic sector is already highly relying on the use of stream data to track parcels and trucks to provide a high service quality. And the use of BDA to provide added value is increasingly play a pivotal role and not being only a significant additional option. Real-time tracking of deliveries from suppliers to client (B2B) or consumer (B2C) has been either an operational management resource and a customer benefit for a long time, and now it is a core service. BDTs enable this information flow to be analyzed much more effectively and in real-time to optimize delivery and communications with the customers. This is reflected in the first of the top three most common use cases, logistics and package delivery management. However, in line with this reasoning the other two use cases identified are price optimization and inventory and service parts optimization. All the three use cases are highly considered as they deliver large profit increase (especially price optimization, 6%) but also relevant margins increase. Less relevant is cost reduction on all fronts.

When considering qualitative KPIs, we would expect customer satisfaction, product/service quality and time efficiency as most relevant ones in providing larger improvements but survey data tells a different story. All the qualitative KPIs, are regarded to provide medium improvements (in a range of 10% to 24%). This is a clear sign that it is an industry with highly sophisticated use of ICT. The use of BDA is not any longer seen relevant to ensure quality of the services offered, but BDA can and will provide a considerable added value to all the industry activities.

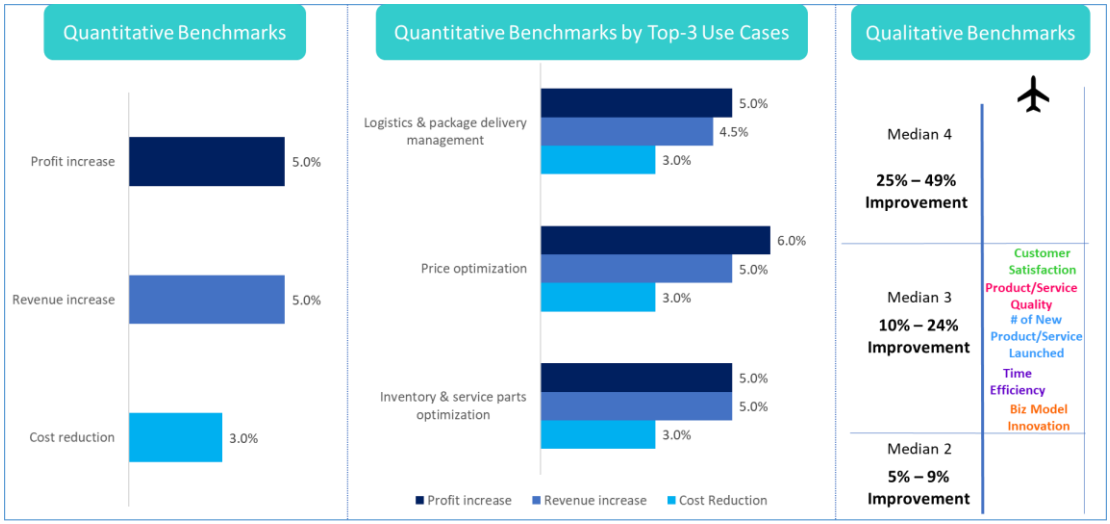

Figure 30 - BDT Benchmarks: Transport/Logistics
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.3.9 Utilities, Oil & Gas

The petrochemical sector is a pioneer in using advanced analytics for resource exploration, and with advancement in the BDTs and computational power, this activity will become highly automated and more efficient. But despite being a pioneer in BDA, this sector is still undergoing a profound transformation process involving digital and core technologies.

The main priorities in terms of use cases are risk exposure assessment (of new activities and services), predictive maintenance of machineries and oil/gas pipes, and regulatory intelligence to better understand and adapt processes to changing regulations. All these use cases are evaluated in terms of profit increase (5% each) while margins appear to be less relevant. The importance of these use cases is also highlighted by the high relevance of some qualitative KPIs. Product/service quality is linked to predictive maintenance: the quality of the service in utilities sector is evaluated by non-discontinuation of the service and avoiding service outages, while predictive maintenance of machineries and pipes is essential to keep proving the service without outbreaks. The number of new products/services launched is directly related with an optimal and well performing risk exposure assessment analysis, which helps organizations to understand whether a new product or service offered will be a viable solution or something to be eliminated from the innovation pipelines. On top of that, another extremely important qualitative KPI is customer satisfaction, as the use of BDTs to manage customer relationship is a primary activity in such organizations.

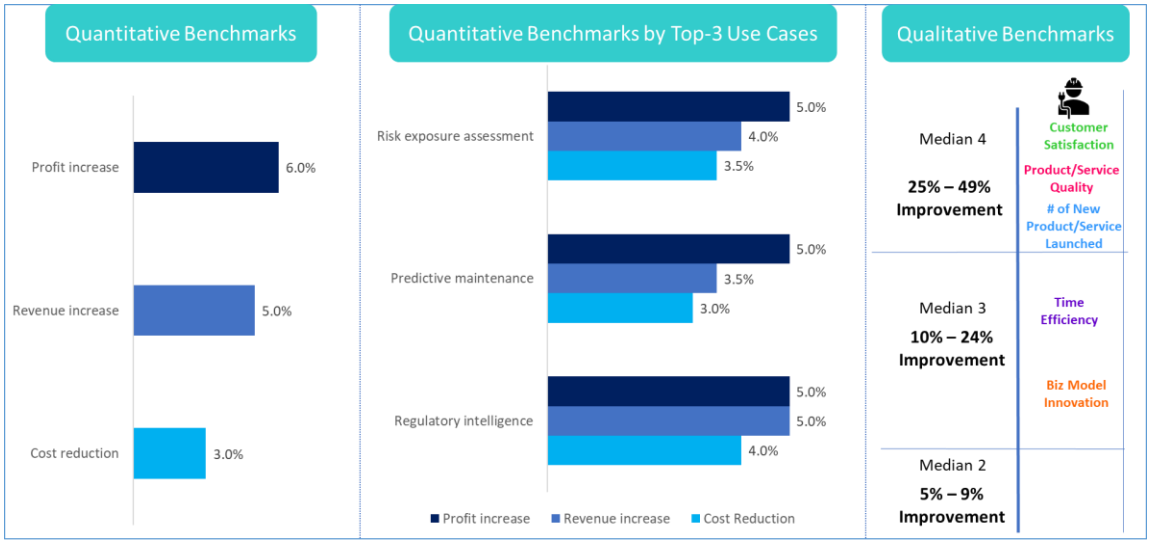

Figure 31 - BDT Benchmarks: Utilities, Oil and Gas
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.4 Business Benchmarks By Company Size

This chapter describes the BDT business benchmarks by industry, which are presented by the Toolbox in the Knowledge Nugget section through the user interface. The comments below provide some background about the significance of the business benchmarks by sector. More detailed benchmarks by use case are provided in D.2.4 *Benchmarks of European and industrial significance.*

5.4.1 Small and Medium Enterprises (SMEs)

SMEs between 50 and 249 employees evaluate profit increase with the highest impact, followed by revenue increase, and cost reduction. Despite showing the larger impacts on profit increase, when analyzing the potential benchmarks for the top three use cases, we can observe that profit increase is never the best benchmark for SMEs. Contrary to what previously seen in the industry analysis – and upcoming size segments -, risk exposure assessment and price optimizations are equally benchmarked with cost reduction and revenue increase, while regulatory intelligence is preferred to be benchmarked with revenue increase. Responses from SMEs are flaky with low percentages and do not clearly provide a unique strategy. Also the qualitative benchmarks providing the highest improvements – still are medium large improvements – are not clearly linked to use cases and on top, it appears a discordance between their business goals and their achievements in adopting BDA solutions.

However, as we are matching company within the same segment size but from highly different sectors, this combination of industry and company size presents discontinuations of the results. But this does not mean that SMEs in adopting the right use case or starting the deployment of BDA solutions cannot perform as well as large enterprises.

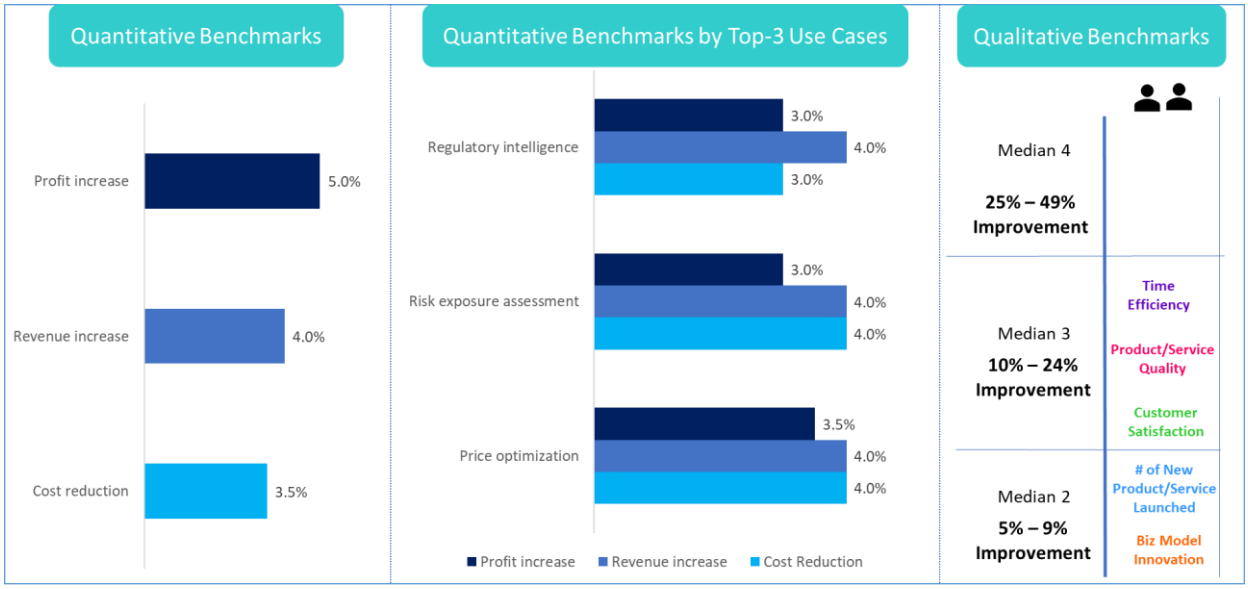

Figure 32 - BDT Benchmarks: SMEs
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.4.2 Mid-large Enterprises

Mid-large enterprises (with 250-499 employees) show larger impacts of BDTs on profit and margins increase than on cost reduction. This is because organizations within this segment-size had already in place the right architecture to achieve cost reductions and are now focusing their attention and efforts on increasing profits and margins. The same is valid when considering specific use cases. Risk exposure assessment and new product development are largely benchmarked with profit increase and secondarily by revenues growth, while regulatory intelligence is mostly benchmarked with revenues growth. From a qualitative KPIs perspective, the company size highlights difficulties in prioritizing a specific benchmark when considering improvements and rate all the presented benchmarks in a medium range with improvements from 10% to 24%.

Overall, mid-sized companies engage with BDTs to boost growth. BDA can enable them to analyze their existing customer base and pricing to identify upselling opportunities and, critically, to provide market analysis to identify new customers. On top, diversification is also a growth opportunity for these organizations as far as they understand the market needs and invest properly to develop new or modify products and services.

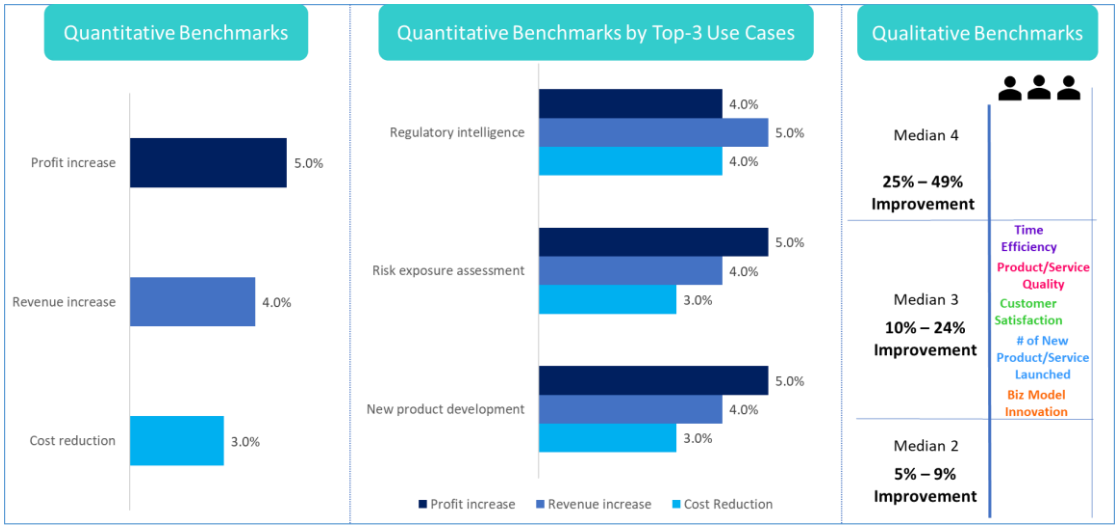

Figure 33- BDT Benchmarks: Mid-Large enterprises
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.4.3 Large Enterprises

Large enterprises (with 500-999 employees) are often better positioned than smaller ones to consider investment in new products and services, whether via internal R&D, mergers and acquisition, or OEM and reseller partnerships. Effective BDA of existing internal products sales and the financial and market prospects of potential M&A targets or partners is a valuable decision support tool. Detailed analysis of market prospects is likely to be expected from external partners or investors. Quantitative KPIs benchmarks exactly overlaps with the evaluation by use case. Risk exposure assessment, price optimization and new products development are equally benchmarked by profit and revenues increase, with cost reduction playing a smaller role. Considering the opening statement, we can clearly see how these top three use cases are perfect fit to the segment description.

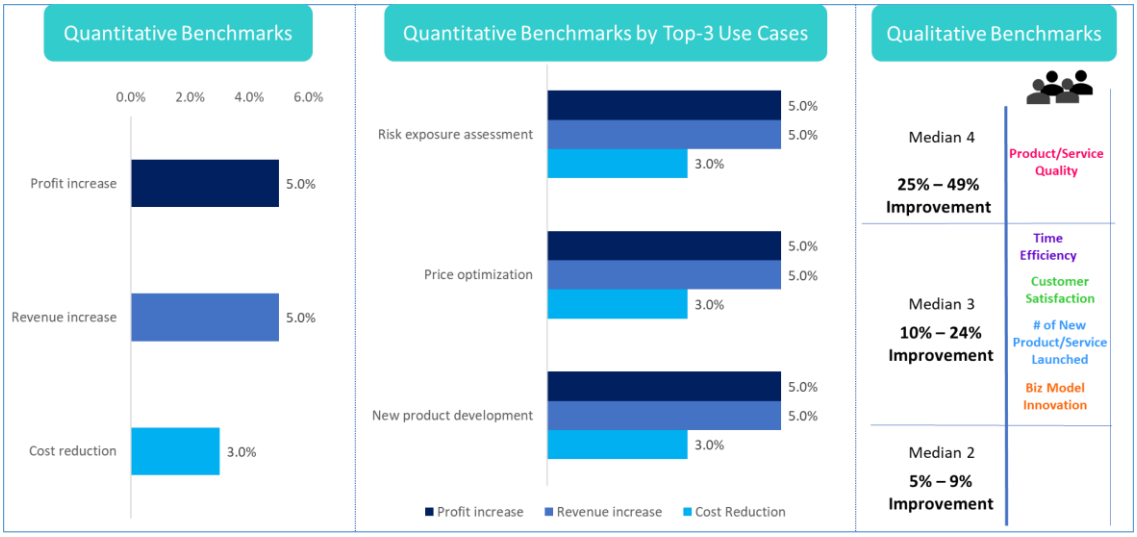

Figure 34 - BDT Benchmarks: Mid-Large enterprises
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.4.4 Very Large Enterprises

Very large enterprises (with over 1000 employees) have the resources to invest in BDA, and BDTs are now able to integrate information that was previously in silos to enable consolidated views of both the company's own processes, which can be optimized, and the current or potential market for products and services. Large companies have access to far more internal customer data, but also gather external data, to integrate leads to significant opportunities in improving customer relations, while at the same time upselling and cross-selling to those customers with carefully targeted offerings.

Very large organizations strongly rely on profit increase as main benchmark rather than revenue increase and cost reduction. Something similar is reflected when evaluating use cases. Customer profiling, targeting, and optimization of offers and regulatory intelligence have the same values presented by the quantitative benchmarks. New product development instead slightly differs from the other two, with equal benchmarking values for profit and revenues increase. Qualitative benchmarks are in line with the top three use cases presented. Extremely relevant, with large 25%-49% improvements, product/service quality and customer satisfaction strengthen the values and results presented with the first and third use cases.

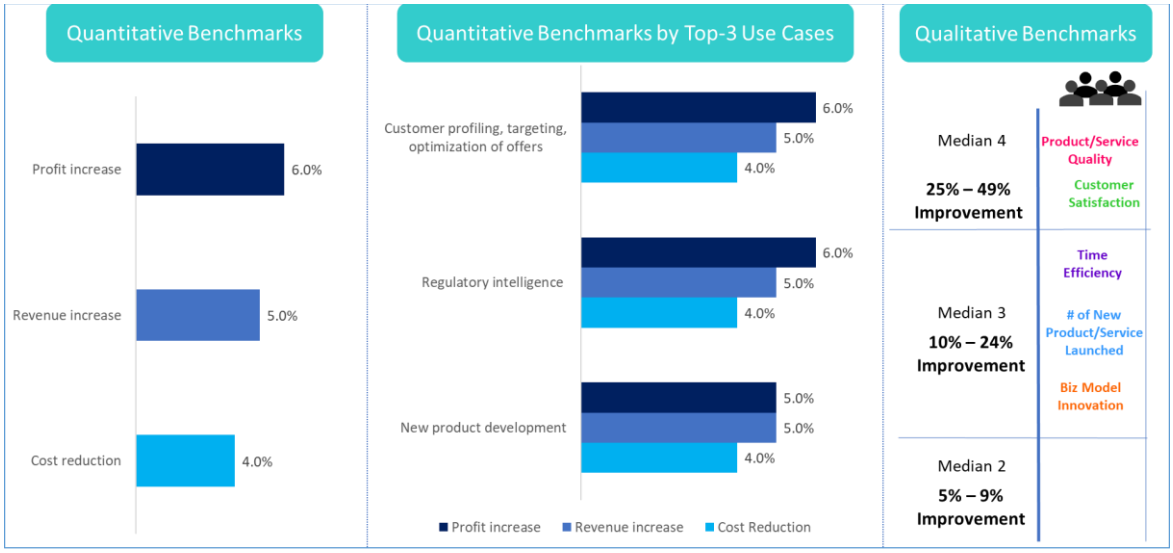

Figure 35 - BDT Benchmarks: Mid-Large enterprises
Source: D2.4 "Benchmarks of European and Industrial Significance"

5.5 Star Performers

Star performers are organizations with the best achievements in terms of business impacts from the use of BDTs. Out of the sample of 730 enterprises interviewed we found 35 organizations falling in this category (roughly 5% of the total sample). Analysing star performers helps other organization to identify the upper boundary of the potential achievements and what they could aim for if they could maximise their effectiveness and efficiency in BDTs.

The star performers group includes (unsurprisingly) a majority of very large and large enterprises. Big data technologies are sophisticated, their adoption requires good investment capability but also a good internal information system infrastructure making datasets available and usable. Large companies started before SMEs in BDT adoption and are today further along in the learning curve of data-driven innovation. For similar reasons, the majority of star performers come from 3 leading industries for the use of data, that is

business/IT services (22%), retail (19%), and financial services (14%). Generally speaking, these enterprises have been early adopters of BDT, show more ambitious plans and expectations than the rest of the sample, and tend to be more eager in taking risks for innovation investments.

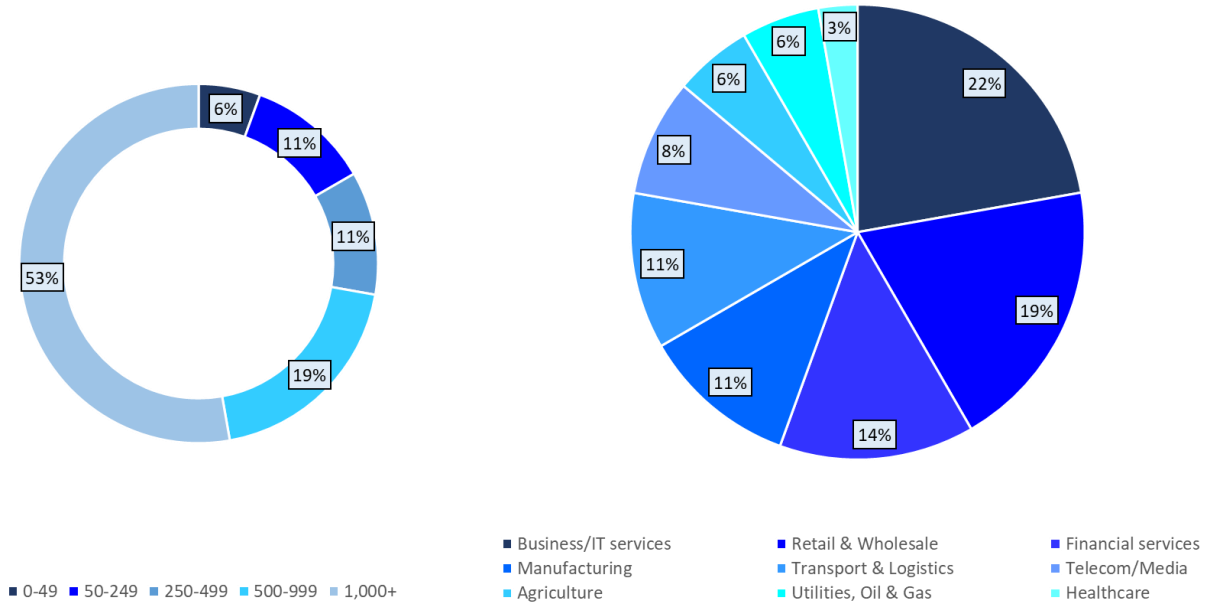

Figure 36 - Star Performers group composition
Source: D2.4 "Benchmarks of European and Industrial Significance"

BDT business impacts for star performers are remarkably higher than the rest of the business users' sample, as shown by Figure 37 below. In terms of profit and revenue increase star performers achieve a mean value of 8% increase against 5% for profits and 4% for revenues for their competitors. Star Performers regard cost reduction as less important than the other two quantitative benchmarks; however, best performers have slightly larger cost reduction than the market average. Concerning qualitative KPIs, the share of star performers with the highest improvement thanks to BDT (over 25%) is three times more numerous than the rest of the sample for time efficiency, product/service quality and customer satisfaction, while the results for the launch of new products and services are better but close to the rest of the sample. Strangely enough, the majority of star performers declare low or medium improvements in the case of adoption of innovative business models. Perhaps this is correlated to company size: large corporations are notoriously reluctant to introduce new disruptive business models.

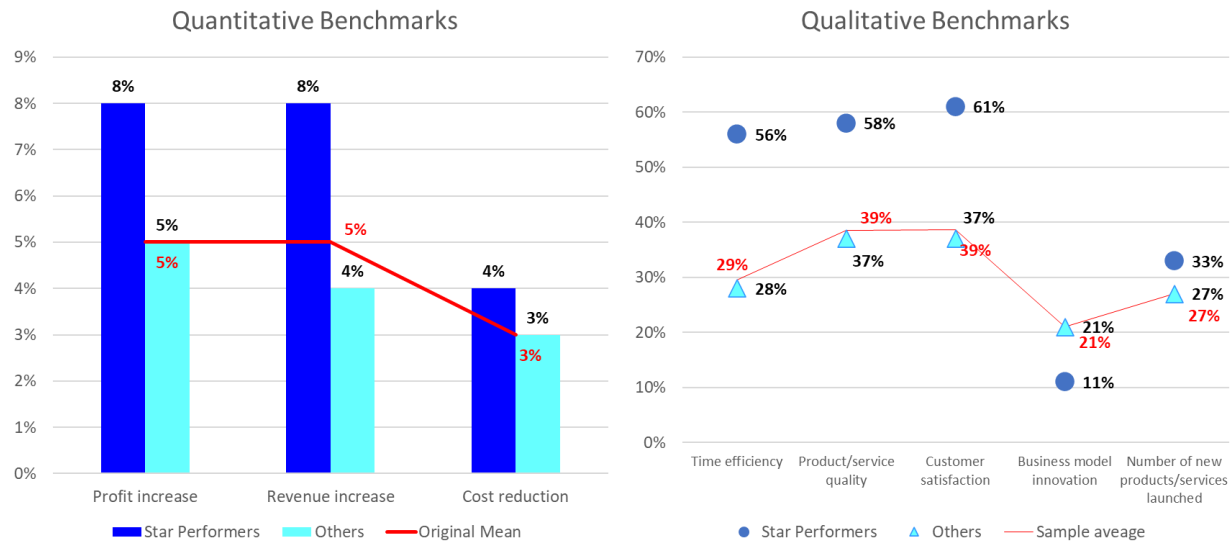

Figure 37 – BDT Benchmarks: Star Performers
Qualitative Benchmarks: share of respondents with high improvements (>25%)
Source: D2.4 "Benchmarks of European and Industrial Significance"

Finally, star performers are ahead of the market also in terms of adoption of advanced and sophisticated BDT use cases, as shown by Figure 38 below.

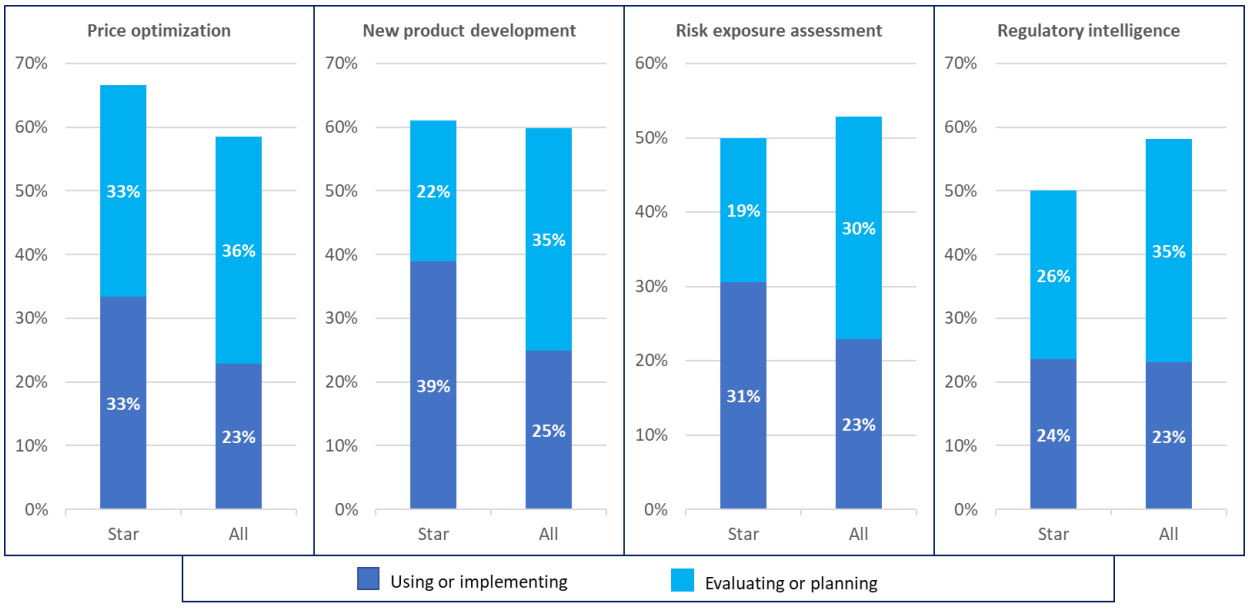

Figure 38 - Star Performers' top use cases

Source: DataBench survey, June 2020

6. Presentation of the Toolbox

6.1 Overview

The DataBench Toolbox is the main technical result of the DataBench project. The Toolbox provides access to a knowledge base of big data benchmarking related artefacts, ranging

from metadata about existing benchmarking tools and initiatives in the community to heterogeneous information and studies performed by the project about benchmarking encapsulated in what we call “knowledge nuggets”.

The DataBench Toolbox is not a single tool, but rather a “box of tools” as its name implies, meaning that instead of being a benchmarking system, it serves as entry point to access to resources about benchmarking: it is intended as a one-stop shop for big data benchmarking. The Toolbox comprises the following elements:

- A web-based front-end enabling access to the main functionalities of the Toolbox. More information about the Toolbox user interface can be found in deliverable D3.4 [1].
- The Big Data Benchmarking Tools Catalogue (BDBT). The BDBT catalogue provides a list and metadata about the most relevant technical big data benchmarking tools. The initial list is based in the deliverables provided in the scope of the DataBench WP1, in particular D1.2[2], D1.3[3] and D1.4[4]. The catalogue is extensible to new benchmarks following the editorial flow explained later in this section.
- A Knowledge Nuggets Catalogue (KN). The aim of this knowledge base is offering to the community a better understanding of the business value of big data benchmarking. As already mentioned, KN are pieces of information initially composed from different elements of the research carried out within the project. However, a KN can be anything that is valuable for the benchmarking community. Therefore, the Toolbox enables the possibility of adding new KN following also an editorial workflow.
- Searching features. To be able to display and search throughout the catalogues of our knowledge base.
- Tag-based navigation between resources. The front-end of the Toolbox allows to navigate through the different resources (benchmarks or KN) by using the annotations or tags. By clicking on a specific tag, the set of resources annotated with the same tag will be prompted and accessed, giving the possibility of browsing resources at will.
- User journeys: As it will be explained afterwards in this section, the Toolbox provides a set of tips and advices to different categories of users on how to use and navigate throughout the Toolbox. The so-called user journeys are envisaged to help users on this, but the navigation can be done in many other multiple ways.
- Statistics of the usage of the tool (visibility only to administrators, although selected statistics are open to anyone).

The Toolbox is accessible via the following URL: http://databench.ijs.si/

The access is open to unregistered users, but some functionalities are only accessible to registered users.

A screenshot of the front page of the Toolbox is shown in Figure 39.

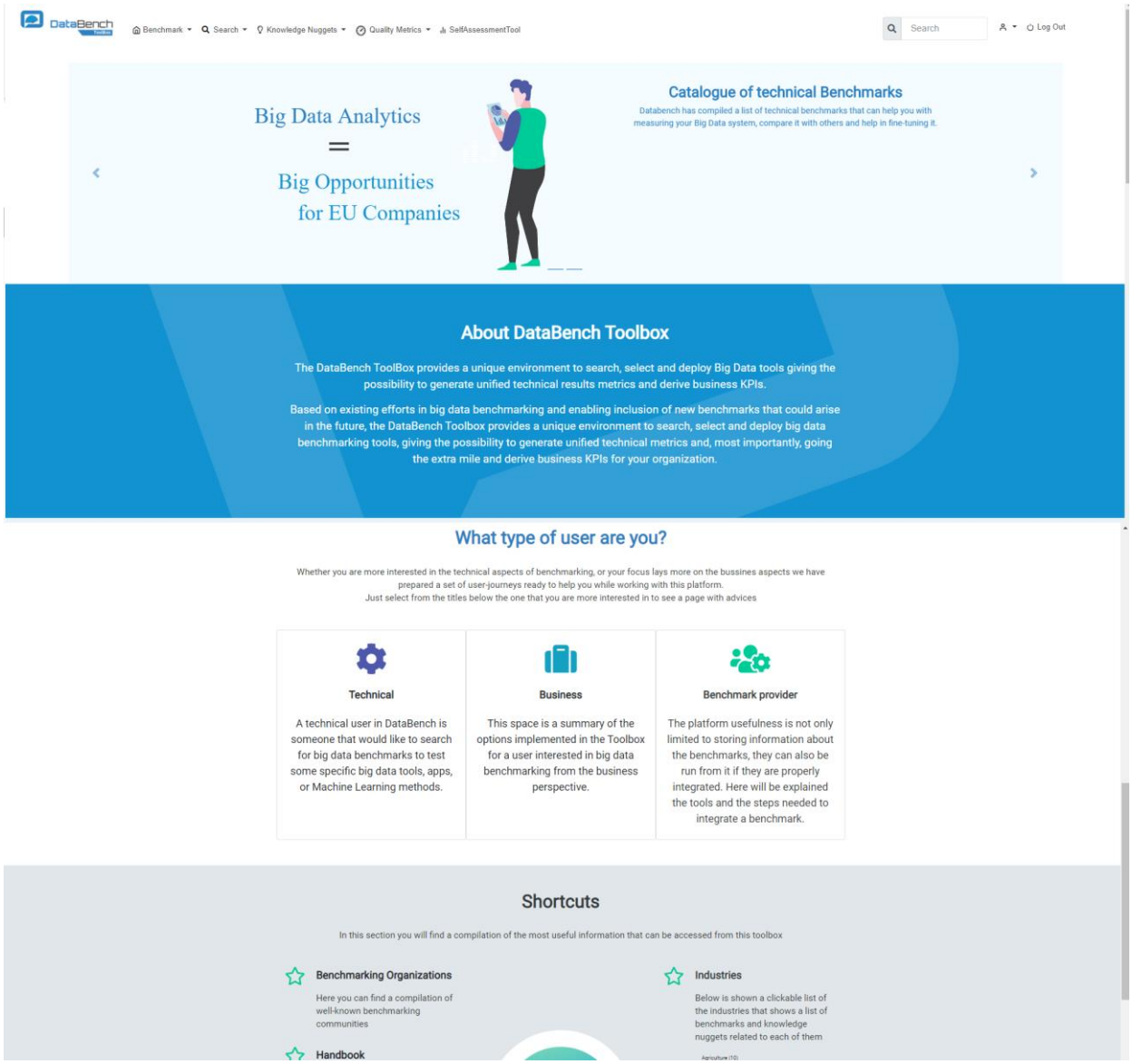

Figure 39 – Front page of the Toolbox

6.2 Intended users of the Toolbox

As reported in deliverable D3.4, "the DataBench Toolbox sits at the core of DataBench results providing access to different types of users to the resources made available by the project and by external initiatives." A summary of the users of the Toolbox is the following:

- **DataBench Administrators**: As any other tool with a knowledge base behind the scenes, the Toolbox needs some housekeeping. The Toolbox admins oversee the Toolbox ensuring that the operations are going well, are responsible of management of users and have functionalities to add and maintain the content of the knowledge base. They are in charge of approving or rejecting new content (benchmarks or knowledge nuggets) provided by the community or by DataBench experts. Administrators also can visualize the statistics of usage of the Toolbox.
- **Technical users**: Technical users is an umbrella englobing different potential stakeholders interested in benchmarking big data solutions with different set of skills and needs. Therefore, the Toolbox offers different suggested paths depending on their

needs, although using the same options to all of them. Examples of technical users could be *casual users*, users who are interested in searching and browsing existing benchmarks and some info about them; *benchmarking experts*, belonging or not to specific benchmarking organizations, who would be interested in finding new benchmarks and eventually deploy and execute them to evaluate technical indicators of big data solutions; *IT experts from industrial organizations*, that might not be necessarily experts on benchmarking, and that may need some help to understand how to benchmark a big data solution, evaluate alternatives, or visualize what others have done in the past; a special user of interest for DataBench are the IT people of *EU R&D projects on big data*, and more particularly projects funded under the BDV PPP, who would like either to check what alternatives are, or benchmark their own results to support their choices. A technical user might belong not only to one, but to some of these categories at the same time, and all should have in common a minimum IT expertise (developers, testers, system administrators) and interest on big data architectural blueprints. Their main interest is therefore in finding the right benchmarks and information and knowledge to ease their benchmarking tasks.

- **Business users**. As in the case of technical users, the business users might have different profiles and interests, although they all share their interest in aspects such as performing business benchmarking rather than on pure technical benchmarking tasks, or understanding the business implications of selecting a big data system and how this choice influence their business indicators. Business users might also fall into some of the subcategories explained for technical users, but leaning more on the business side of benchmarking: *causal users* might be interested on searching, navigating and browsing knowledge about benchmarking, finding similar cases of application of big data in their industry, interest on architectural blueprints of reference for their domain or use case, etc.; people in charge of exploitation and sustainability of big data solutions in *EU R&D projects on big data*, would be interested in finding knowledge to back their decisions about the use of big data in a particular sector, industry or market. In general, all business users would be more interested in navigating through the knowledge base (knowledge nuggets) catalogue rather than the technical benchmark catalogue.
- **Benchmark Provider**. Benchmark providers are those that developed a specific big data benchmark and would like it to be part of the technical benchmark catalogue of the Toolbox. These users might belong to specific benchmarking organizations (i.e. the TPC[2]) or developers a specific big data benchmark (i.e. people behind YCSB[3]). The Toolbox allows these users to add their benchmark to the Toolbox. The new addition will follow an editorial approach before being listed in the Toolbox. Some of these benchmarks might go a step further and enabling the automation for being deployed and run from the Toolbox itself. This last step requires engaging with the Toolbox administrators.

Besides administrators, the main benefits for the three main categories of users is summarized in Figure 40.

[2] http://www.tpc.org/

[3] https://github.com/brianfrankcooper/YCSB/wiki

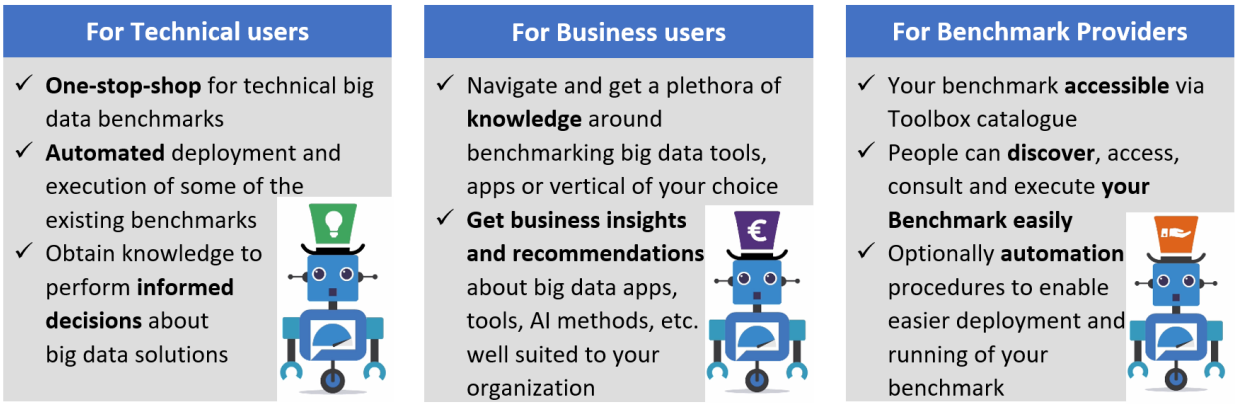

Figure 40 – Summary of main benefits for the users of the DataBench Toolbox

6.3 Toolbox user interface

Users of the Toolbox can access the different tools and elements explained in the overview of the Toolbox from the main page shown in Figure 39. In this section we are going to explain the different options of the Toolbox available in the front-end. The options seen in the main page are the following:

Menu

The menu provides options to access to the Technical Benchmark and Knowledge Nuggets catalogues, search options, as well as links to other tools such as the Self-assessment or quality metrics. Some of these options are only available to registered users (i.e. in the submenus of the "Benchmarks" option, only registered users with role of benchmarking providers could suggest new benchmarks to include in the Toolbox).

Search

As explained in D3.4, "The Toolbox provides three types of search: 1) a search box where users may type parts or any of the existing tags or words of the title or the resources; 2) an advanced search enabling the selection of tags to navigate to the resources; and 3) a search interface using the Big Data Value Reference Model explained in the BDV SRIA[5] to navigate to specific technical benchmarks that provide coverage for the horizontal or vertical layers of the model." This means that besides the text search enabled via the search box located in the header, the menu of the Toolbox allows to select the type of search, with two options: Guided search and search by the BDV Reference Model.

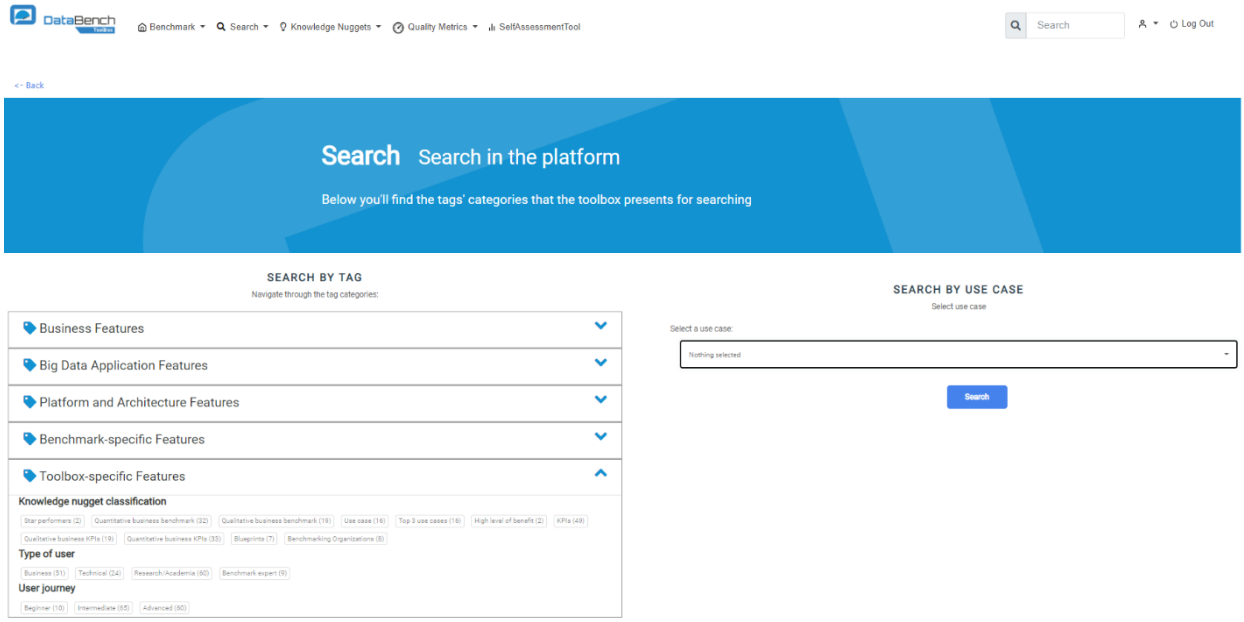

Figure 41 – Guided search

Figure 41 shows the Guided Search page. This page allows to search by the different tags used to annotate resources (technical benchmarks or knowledge nuggets). Users can unfold the different categories of tags and select one (as a matter of example, in the figure the "Toolbox specific features" category is unfolded). Once selected, the resources annotated with that tag will be listed.

Some of the tags are meant for technical aspects (i.e. the tags under the categories of "Big Data Application Features", "Platform and Architecture Features" and "Benchmark Specific Features"), while others are more related to knowledge nuggets (i.e. the "Toolbox-specific Features"), or both (i.e. the "Business Features").

The second search option is the "Search by BDV Reference Model". This option gives access to the image of the model of reference coined in the scope of the Big Data PPP as shown in Figure 42.

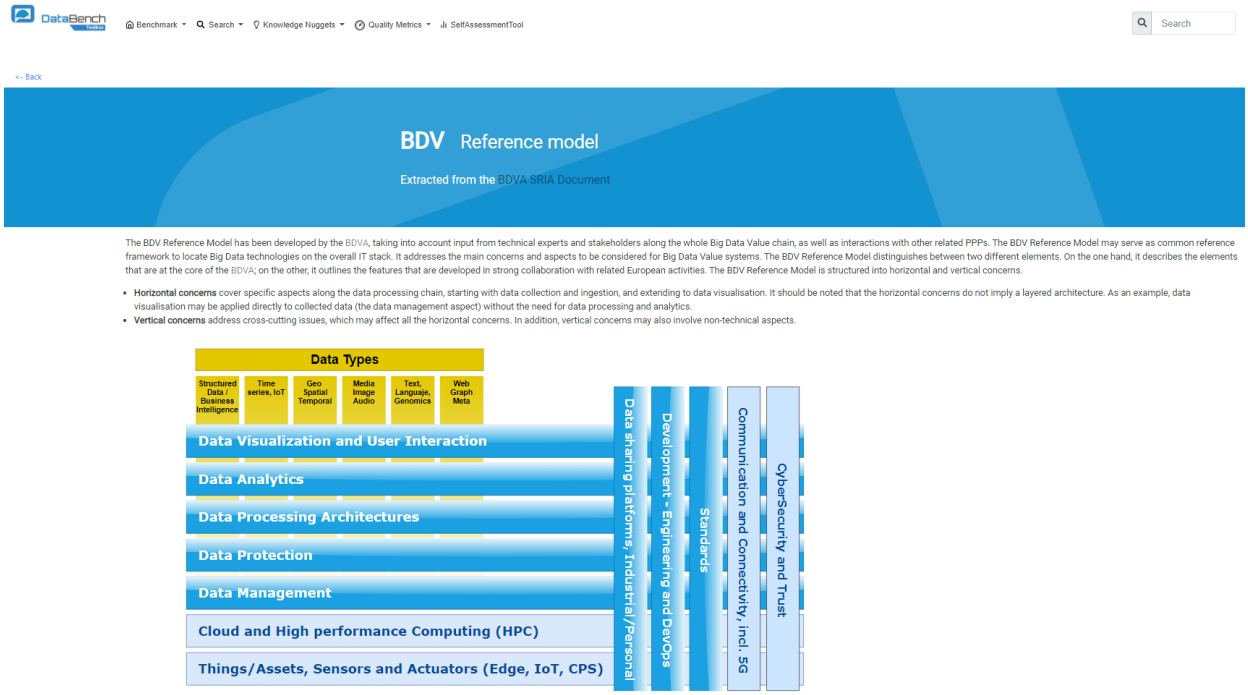

Figure 42 – Search by BDV Reference Model

Users can click on the layers of the model and the list of technical benchmarks annotated to one of the layers will appear. Note that not all benchmarks are annotated with the layers of the model, as there is not perfect match between each benchmark and the layers. However, it is a good starting point to find some of the benchmarks that might help in a specific area.

Finally, the search box located near the menu (Figure 43) provides a full-text search over the description of the benchmarks and knowledge nuggets.

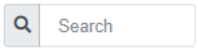

Figure 43 – Full-text search box

An example of the results of any of the search is a list of resources, such the example shown in Figure 44

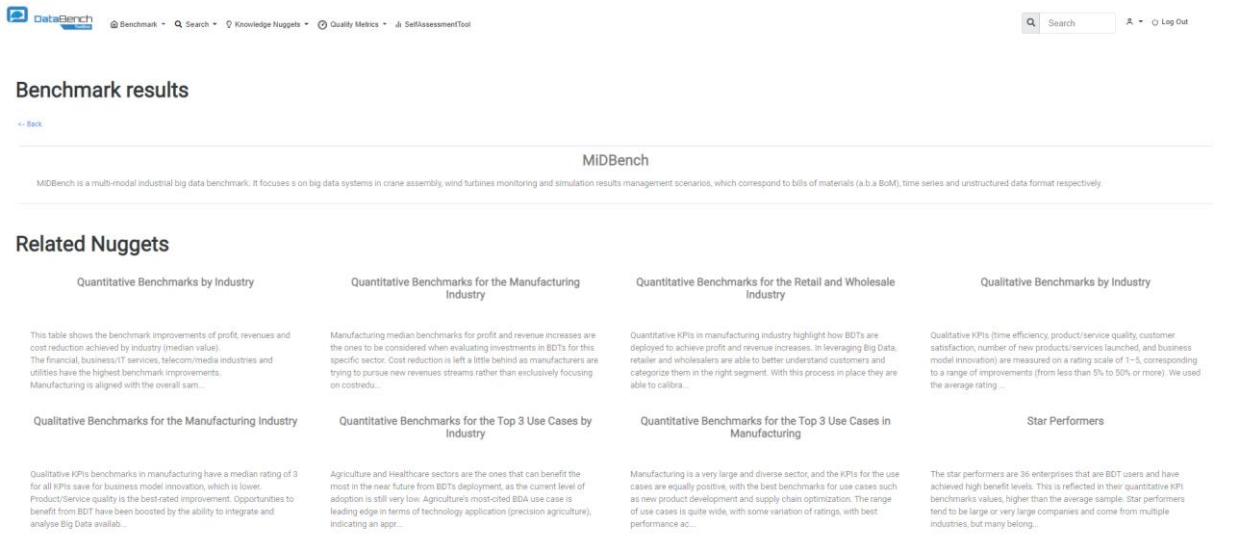

Figure 44 – Search results

6.4 User journeys

In order to help users to navigate throughout the DataBench Toolbox, minimize their learning curve and entry barriers while maximize their chances to find and use the right benchmarking solutions or knowledge, the Toolbox provides a set of tips and advices.

By their very inner nature user journeys are in evolution. More tips and dedicated pages in the Toolbox will be shared as we get feedback and learn from the usage of the users. Therefore, the information presented in this section should be considered as the starting point of the information presented to the users.

6.4.1 Support for casual users

As explained before in this section, casual users might be either technical or business users with interest on searching, navigating and browsing knowledge or benchmarks. Their initial goal might not be executing any specific benchmark, but rather browsing information about big data benchmarking to find what others are doing, such as similar cases of application, industry examples, architectural blueprints, etc.

This type of users is what the Toolbox names “Beginners”. User journeys for beginners can be found either for business or technical users by clicking on the respective section in the front-page, shown in detail in Figure 45.

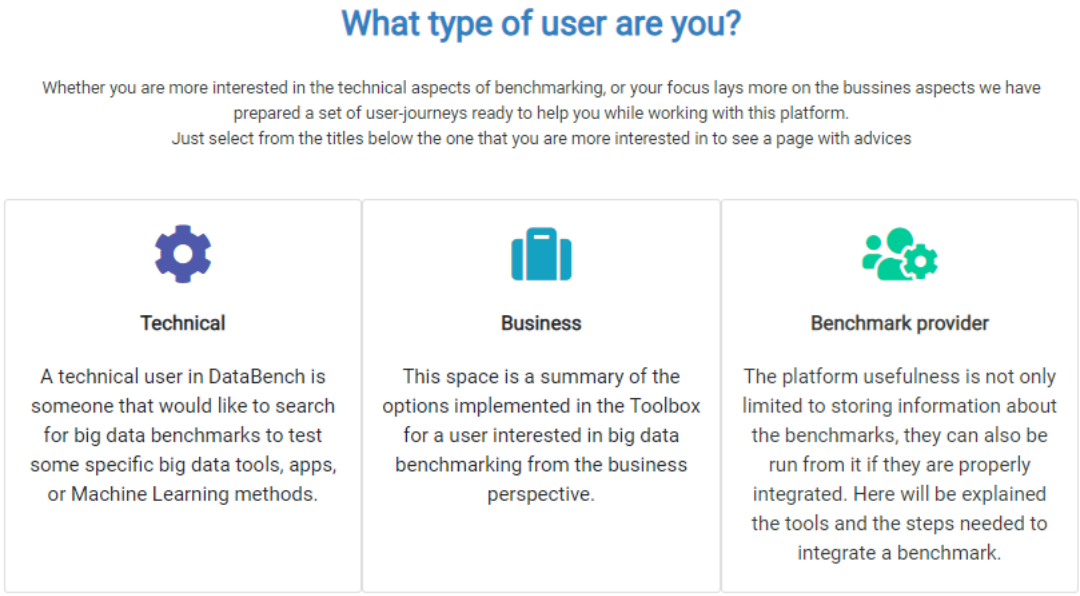

Figure 45 – User journeys section from the Toolbox front-page

From the technical perspective, the tool provides a set of recommendations for casual users on the first steps to get acquainted with the Toolbox and try to find the right information they could be seeking, as shown in Figure 46.

Figure 46 – Technical user journeys

First of all, it is good to understand some of the main concepts about big data benchmarking, such as what it is benchmarking in opposition to technical validation or testing of big data solutions. These are terms often interchanged, while in reality are two different things. Therefore, the Toolbox provides some knowledge nuggets explaining what big data benchmarking is and what can be expected from benchmarking. The first recommendation for casual users or beginners is therefore directing them to the Frequently Asked Question (FAQ) of the Toolbox, where users might find definitions and links to other pieces of knowledge, as shown in Figure 47.

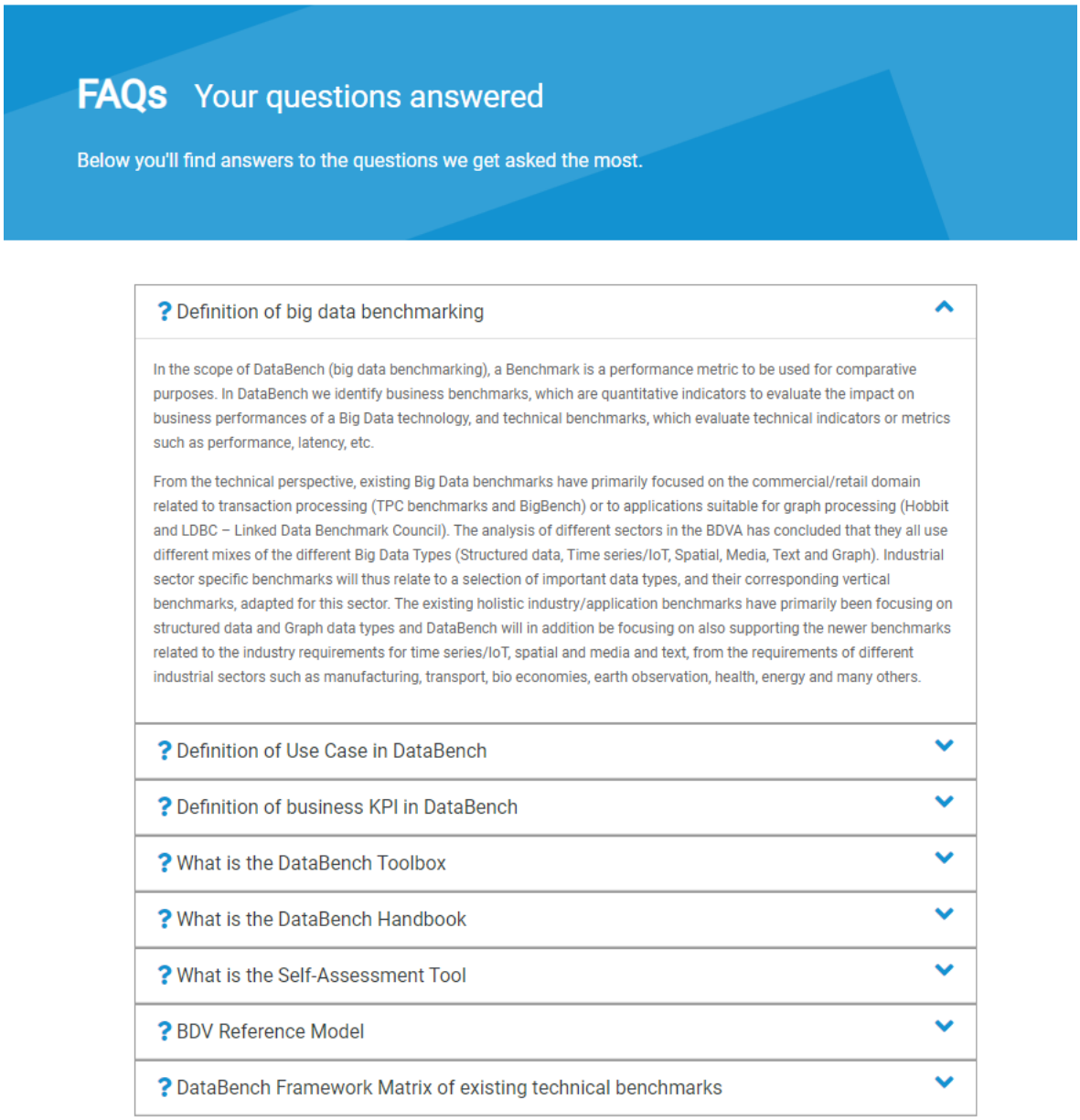

Figure 47 – FAQ section of the Toolbox

As mentioned above, the user journeys will be updated after receiving feedback from users. At the time of writing this document, the user journey for beginners or technical casual users are giving advice about the following subjects:

- Use of the search functionalities: As explained in section 6.3 (full-text search, search by tag – guide search – and search by BDV Reference Model).
- How to browse the entire catalogues of knowledge nuggets or technical benchmarks by selecting the appropriate options of the menu located in the header. Once located, you could select one technical benchmark and navigate to their own page to look for more content. Benchmarks are marked with metadata indicating their main features.
- Browsing big data architectural blueprints. Different archetypical architectural patters of the usage of big data in different domains have been gathered in the form of blueprints. These might be interesting to understand what other are doing. The user journey suggests typing "blueprint" in the search box to find them, but

blueprints might be reached in several other ways (i.e. by navigating from other nuggets or typing the specific industry in the search box).

In a very similar way, the Toolbox provides use journeys for causal business users, as shown in Figure 48.

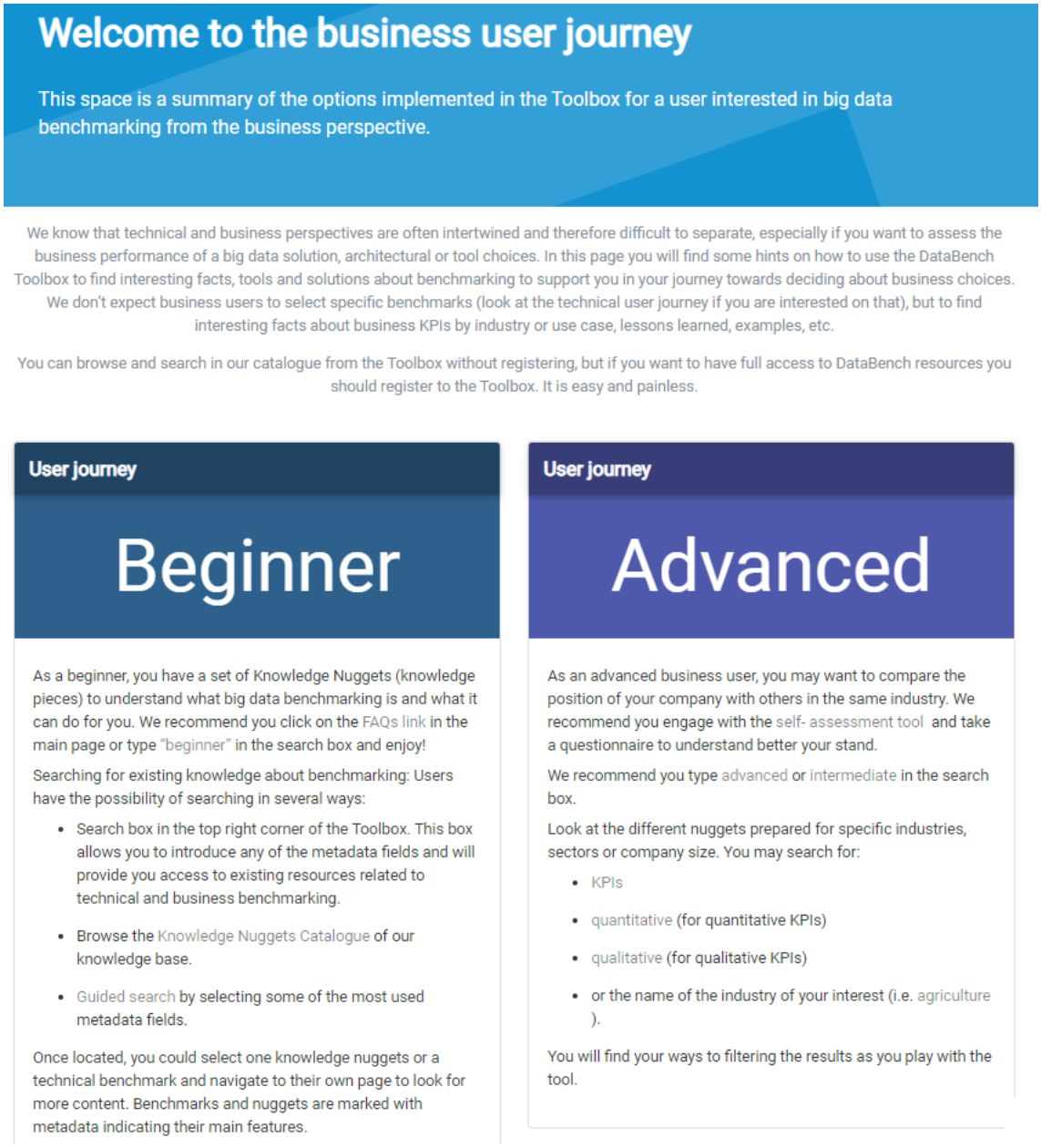

Figure 48 – Business user journeys

Figure 46 and Figure 48 also show the current user journeys for more advanced technical and business users, respectively. As mentioned above, these user journeys will be updated and upgraded based in the feedback received from users.

6.4.2 Support for benchmarking providers

Sections 6.5.1 and 6.5.2 provide an overview on how the tool supports benchmark providers to add new benchmarks into the platform. As in the previous cases, the user journeys section of the Toolbox front page shown in Figure 45 gives access to the information related to add new benchmarks. It is divided in two parts:

- Adding a new benchmark to the catalogue
- Integrating a benchmark for automation of deploying and execution.

Figure 49 shows the information displayed as user journeys for these two different aspects. This information is very similar to what it is discussed in sections 6.5.1 and 6.5.2, complemented with more information gathered from deliverable D3.4.

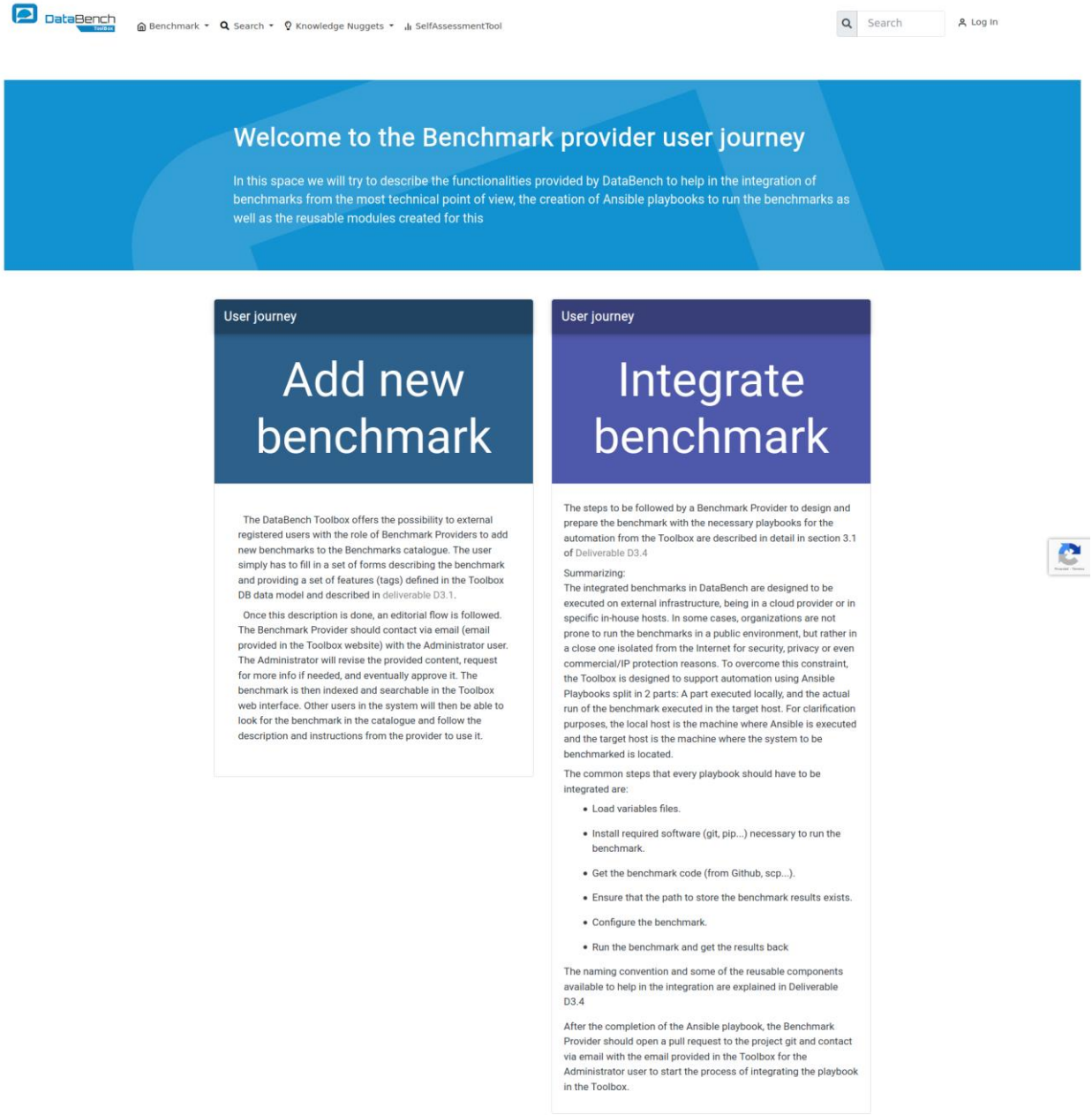

Figure 49 – Benchmarking providers user journeys

6.4.3 Support for benchmarking experts

Besides the user journeys for technical users shown in Figure 46, there are specific user journeys for advanced technical users that are experts on big data benchmarking. These users belong to or are aware of the work done in existing benchmarking initiatives (i.e. TPC, Bench Council, STAC Benchmark Council, or others). More information about these organizations can be found in the DataBench website[4].

However, the amount of benchmarking initiatives and different benchmarks that are popping out every other day makes quite difficult to have a comprehensive and up to date list of benchmarks even for benchmarking experts. The Toolbox intends to provide such a list, with links to each of the benchmarks through the benchmarks catalogue accessible from the menu of the Toolbox (option Benchmark/Benchmark catalogue). A screenshot of the catalogue is shown in Figure 50.

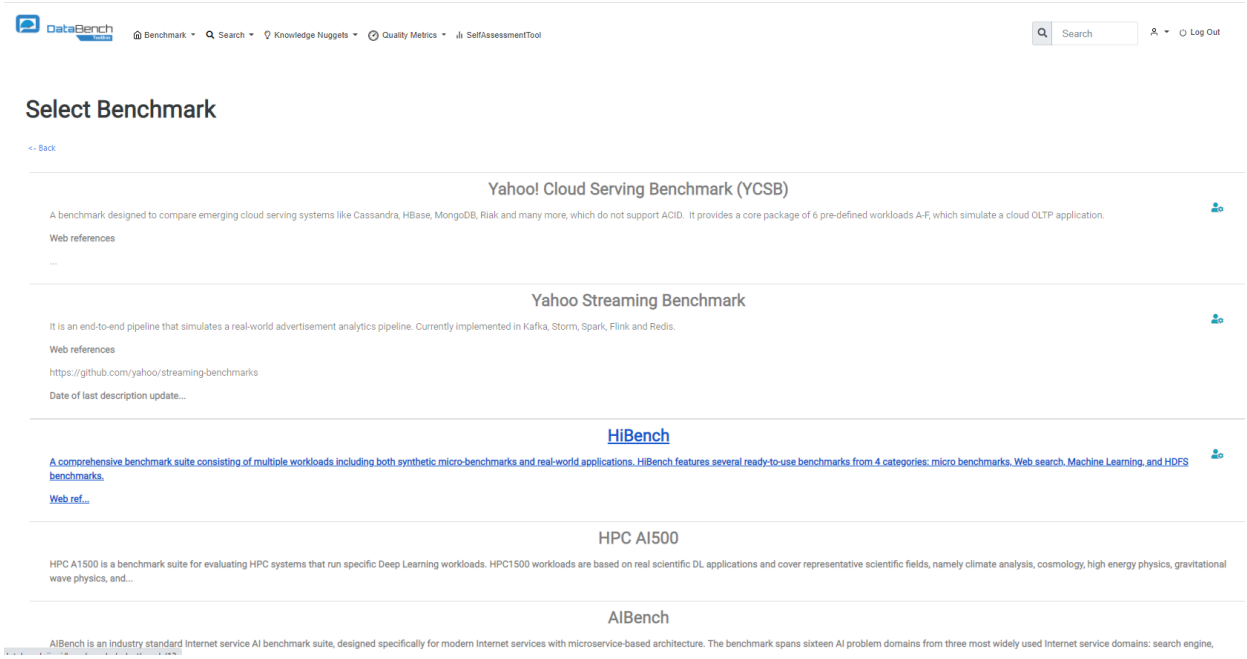

Figure 50 – Benchmark catalogue

The catalogue gives access to the list of benchmarks introduced in the Toolbox. Besides browsing the entire catalogue, each of the benchmarks is annotated with a set of tags that allows searching and navigation using the search functions described in this document. The benchmarks are selectable by clicking on them and shows the tags, as shown in Figure 51 with the example of the YCSB benchmark.

[4] https://www.databench.eu/community/

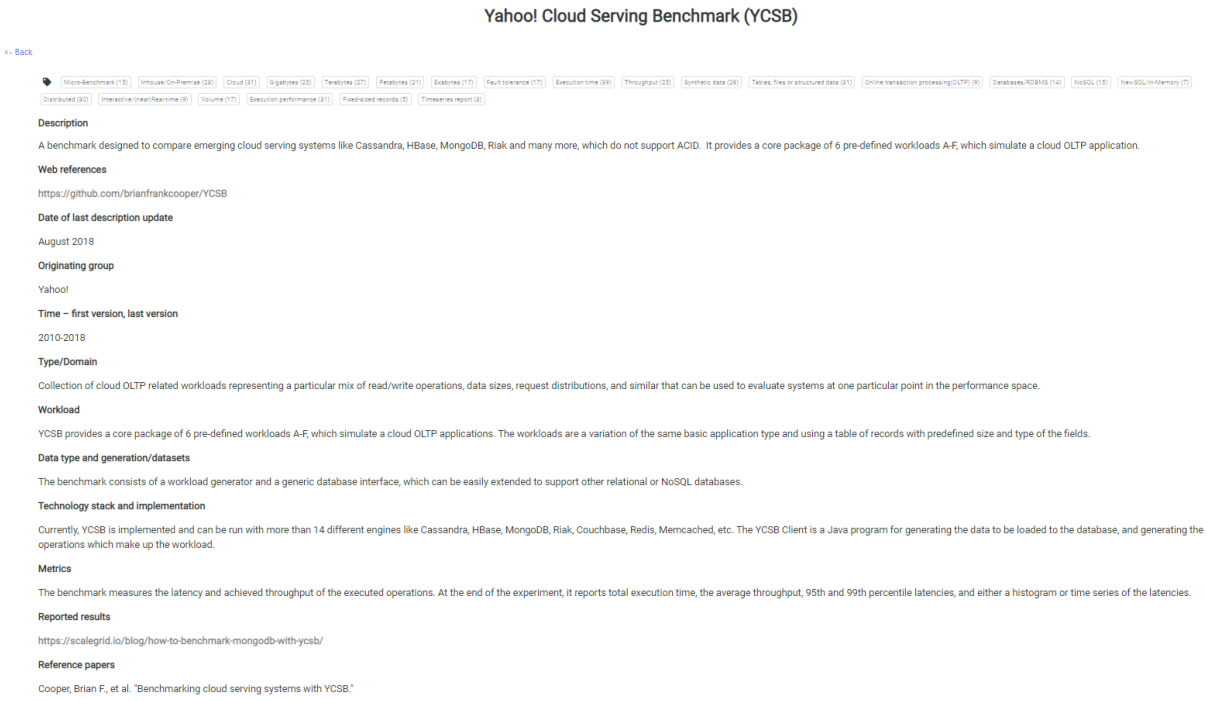

Figure 51 – Browsing a specific benchmark

Note that the page gives the possibility not only of having a set of well-defined metadata such as description, web references, dates, domain, workloads, metrics, etc. but also tags at the top of the description to navigate to other benchmarks or knowledge nuggets with some relation to a particular benchmark. This shows the navigational approach followed by the Toolbox.

Some of the benchmarks (i.e. YCSB) have more options for advanced users. Registered users with the role of technical users will see more interactive information besides the one shown in the previous figure. This requires that benchmark providers or DataBench administrators had done a previous work to enable the integration, automation, deployment and execution of the specific benchmark as explained later in section 6.5.2. If that is the case, the registered user will be prompted with the deployment and execution panel as shown in Figure 52.

Note that this is an advanced feature of the Toolbox only provided for a few selected benchmarking tools, but it can be extended to benchmark providers to automate their own benchmark in the same way. Nevertheless, users might always decide to use the links to the benchmark web page in order to follow the instructions of the benchmark providers to use and install the benchmark manually, as they must do in the benchmarks that are not fully integrated.

If the benchmark is integrated, the way of configuring it for deployment and execution has been explained in deliverable D3.4, and it is summarized here:

- For automation of the deployment and execution the Toolbox relies on the use of Ansible[5]. Ansible is an automation engine that allows to automate several steps of

[5] https://www.ansible.com/

configuration management, deployment, orchestration and even execution of IT tasks.

- The execution of an integrated benchmark from the Toolbox is only allowed for registered users with the role of Technical user or Administrators. Registered users of the Toolbox with those roles wouldn't have to worry to use the integrated benchmarks or Ansible at all, but rather customize the inventory of host(s) machine(s) to deploy the benchmark, and the configuration file proposed by the Toolbox for the benchmark to be able to perform the necessary actions.
- The inventory of host(s) machine(s) where the benchmark should run could be any host available by the user identified by an IP address and credentials. The user may create the inventory on the fly via the configuration file or store it for further usage and select it in successive runs. The credentials to access the inventory can be created from the user profile to avoid security issues via de "Credential" option under the user profile icon on the right-hand side of the page. The machines are machines provided by the user, either in-house, or in any cloud provider using a public IP address (otherwise the Toolbox would not be able to automate the process).
- The variables of the configuration file vary from one benchmark to another, as they depend on the variables accepted by the benchmark in particular, the outputs and metrics measured, elements to compare, etc. In the example shown in Figure 52, the YCSB benchmark ask for variables such as the host name, databases to compare, users to access to the selected databases, path to retrieve the results, etc. The user would need to tailor the variables in the file.
- Once these steps are completed, the technical user should click on the "Launch job" button at the bottom of the page and the system will automatically take care of deploying and running the benchmark.

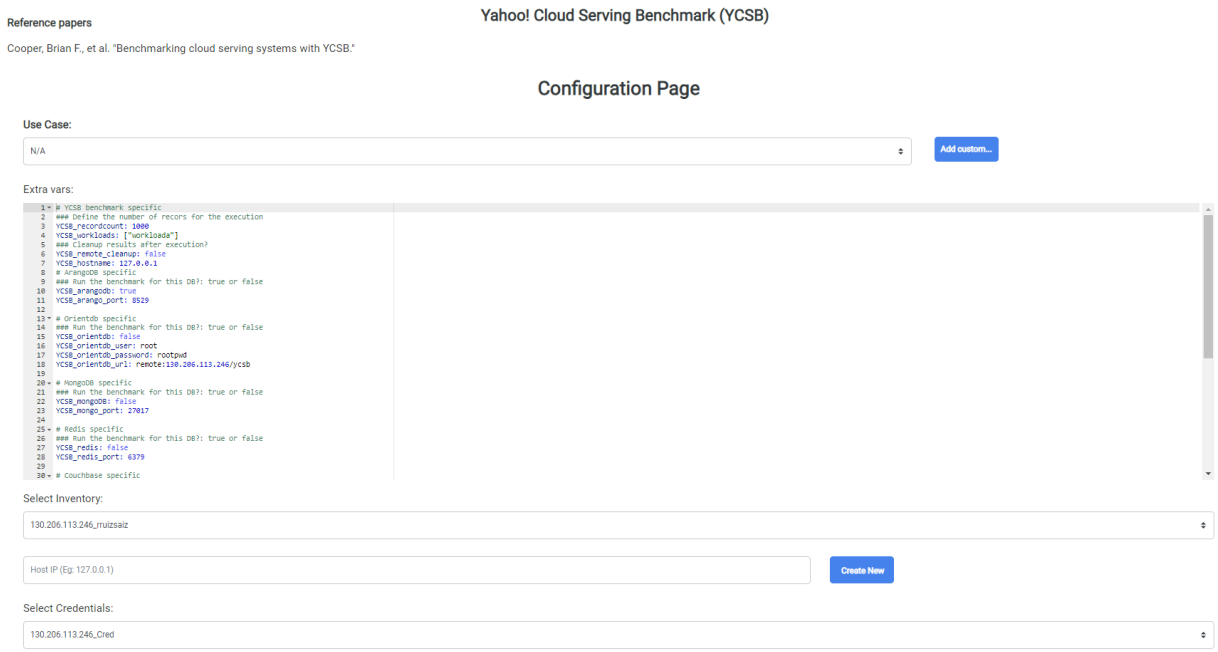

Figure 52 – Browsing and interacting with an integrated benchmark.

The results of the run can be accessed either in the host where the installation took place or via the option from the menu “Benchmark/Results”. This last option gives the possibility to the user to visualize the results file of all the runs made automatically via the Toolbox. Note that these data are only accessible by the user that initiated the deployment/execution. Two examples of visualization can be seen in Figure 53 (detailed results of a single execution) and Figure 54 (time series of executions of the same benchmark showing some specific metrics). Note that these visualizations are tailored for each benchmark, and therefore if new benchmarks are fully integrated the benchmark provider should also provide the way to perform similar visualizations or not.

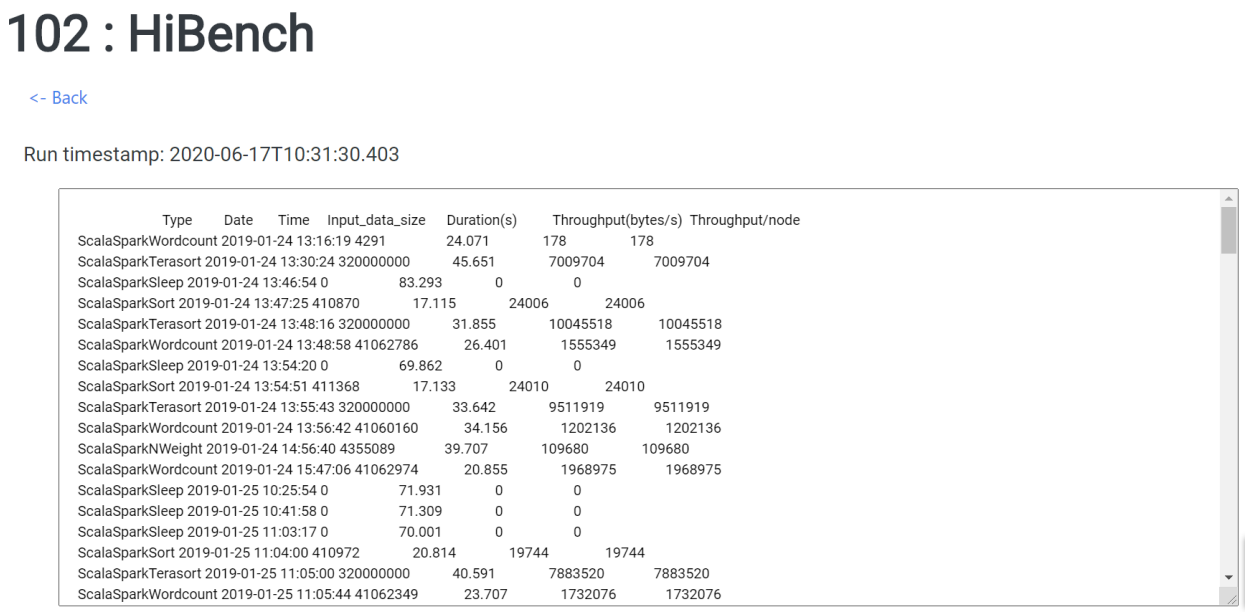

Figure 53 – Example of visualization of results of an integrated benchmark

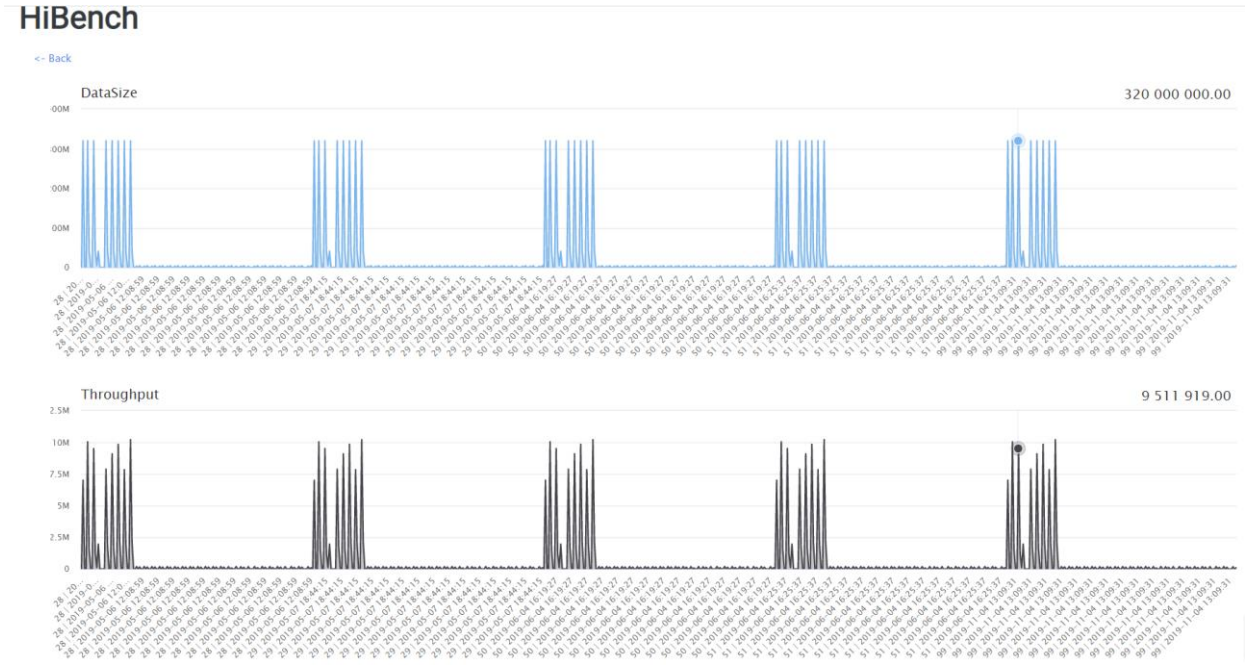

Figure 54 – Browsing and interacting with an integrated benchmark.

6.4.4 Support for big data R&D projects

As mentioned above, one of the main target users of DataBench are EC-funded projects on big data, and especially the ones funded under topics related to the BDV PPP. We have identified so far two main lines of collaboration with these projects related to their needs in the different time of their life cycle:

- Beginning of the project: Projects usually goes through a phase where they need to assess the state-of-the-art of tools and methods related to their goals. In this phase, they need to check different big data or AI tools, frameworks and applications and take an informed decision on what are the best to fulfil their needs. This is the time where many of the existing benchmarks may help to compare and decide among the tools of choice. The example of YCSB shown above is clear: it allows to compare several performance metrics of different NoSQL databases (ArangoDB, OrientDB, Redis, MongoDB and Couchbase). There are many other benchmarks that might allow to understand better which tools are available and executing them in the infrastructure of the project to check in detail how these tools perform.
- Near the end of the project: In this phase, projects are more interested on getting some ideas on how the tools and applications implemented in the project perform. There are a few application benchmarks that might help on this, but often the applications developed in the project require of a more specific benchmark closer to the application itself. This entails the development or customization of new benchmarking tools, normally by taking and adapting an existing one. For these, DataBench offers some guidelines in the form of user journeys and knowledge nuggets.
- Finally, projects might be interested not only on the technical benchmarking side, but rather in the business aspects. The section below would be also of interest for projects, as well as all the information given in sections **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.**.

6.4.5 Support for business users

Figure 48 above already shows the current user journeys for more advanced business users. As mentioned above, these user journeys will be updated and upgraded based in the feedback received from users.

Knowledge nuggets have been thought for all kind of users of the Toolbox, but it is business users who might benefit the most from them. These nuggets represent pieces of information coming not only from project results, but also from the benchmarking community as a whole. An example of knowledge nugget can be seen in Figure 55.

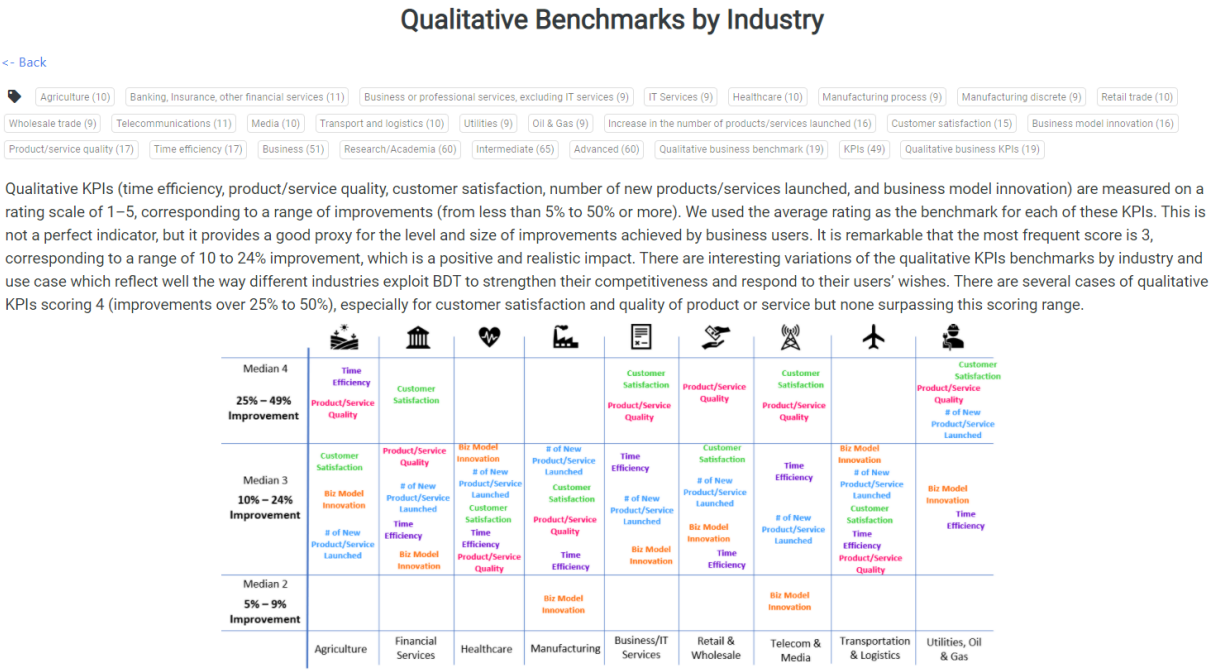

Figure 55 – Example of knowledge nugget.

Section 6.5.3 explains how this knowledge base can be extended by any user, therefore being a live component to understand the benchmarking landscape.

We have covered the usage of the Toolbox for casual users above. In the case of business users, they may want to find information about their sector, compare with similar organizations or competitors or simply check what others are doing in big data and AI. There are several knowledge nuggets in the Toolbox tagged with the labels "advanced" and "intermediate" that might be the initial entry point to this type of users. By typing these labels in the search box of the Toolbox, a list of knowledge nuggets with some hints will be prompted to the user.

There are also nuggets related to specific domains and sectors. By typing the name of the sector (i.e. "agriculture") a list of associated nuggets or even technical benchmarks used in that sector will appear. In fact, looking for the nuggets prepared for different industries, sectors or company size are some of the main suggestions for business users provided in the user journeys. They may find information typing in the search box terms such as "KPI", "quantitative" (for quantitative KPIs), "qualitative" (for qualitative KPIs), etc.

As mentioned in previous cases, the project is getting feedback from users and preparing a set of user journeys for business users giving more hints and knowledge about business benchmarking and related info.

6.5 Methodology to add new knowledge/benchmarks

The Toolbox has been prepared to be extensible in order to populate the catalogues of benchmarking tools and knowledge nuggets. This section explains how to do so by the different type of registered users as well as the editorial workflow in place to ensure the quality of the information. In this sense, the Toolbox administrators are the top-level editorial responsible of the information that is accessed in the Toolbox. They are in charge of the user management as well as approving or rejecting the information provided by the

users. In DataBench we have implemented a simple workflow to get in touch with administrators via email after including any new addition to the catalogues. As for any other communication, users should go to the bottom of the main page of the Toolbox to find the contact email once they have finalized their additions or might want to know more on how to do so. The administrator will have the possibility of revising the list of pending elements to approve, reject, or communicate back via email to the user.

The subsections below explain what users should do to successfully add new elements to the catalogues.

6.5.1 Support for adding new benchmarks to the catalogue

This process has been explained in deliverable D3.4 (section 3.1). The following text is quoted from that deliverable:

"The DataBench Toolbox offers the possibility to external registered users with the role of Benchmark Providers or Administrators to add new benchmarks to the Benchmarks catalogue. The user simply has to fill in a set of forms describing the benchmark and providing a set of features (tags) defined in the Toolbox DB data model and described in deliverable D3.1.

Once this description is done, an editorial flow is followed. The Benchmark Provider should contact via email (email provided in the Toolbox website) with the Administrator user. The Administrator will revise the provided content, request for more info if needed, and eventually approve it. The benchmark is then indexed and searchable in the Toolbox web interface. Other users in the system will then be able to look for the benchmark in the catalogue and follow the description and instructions from the provider to use it." [1]

6.5.2 Support for integrating new benchmarks to be executed from the Toolbox

This process has been explained in detail in deliverable D3.4 (section 3.1). In this section we only summarize the steps to be done without entering in the technical details explained in D3.4:

As mentioned above, the integration of specific benchmarks is optional and in some cases is not possible due to technical constraints. The Benchmark Provider has to enable the automation process using Ansible, totally or partially, to enable the configuration, deployment and execution of the benchmarks, as well as getting the results of the runs back into the Toolbox database for further visualization. This automation process requires two main steps:

- Creation of an Ansible Playbook for the benchmark. The Benchmark Provider should perform the following steps:
    - Preparation of the hosts (external infrastructure, either in-house or in the cloud) where the benchmark is going to run. D3.4 explains the steps to be done in Ansible to enable this preparation, basically consisting on using two Ansible playbooks: One to be executed locally (in the machine where Ansible is running – the DataBench host machine), and the target host(s) where the deployment and execution will take place. D3.4 explains what this entails for Benchmarking Providers. The Toolbox also offers knowledge nuggets explaining this type of information explained in the benchmarking provider user journey. The Toolbox offers a set of reusable playbooks to facilitate this

to the user. The Toolbox administrator should be contacted by the user in order to bootstrap and give advice in this process.

- Integration of the Ansible playbook in the Toolbox by the administrators. After the integration of the playbooks, the benchmark provider should contact the administrator to finalize the process. They will revise the code and upload it to the git repository, create the template for the new benchmark in AWX, select the project and the playbook, allow the configuration of the template at runtime from the web and finalize the integration. All these steps are explained in D3.4 in detail.

6.5.3 Support for adding new knowledge nuggets

As in the case of technical benchmarks, registered users may propose new knowledge nuggets. The procedure is similar, but much simpler. A registered user may add a new nugget by using the appropriate menu option under the "Knowledge Nugget" menu. The user will be prompted with a form such as the one shown in Figure 56.

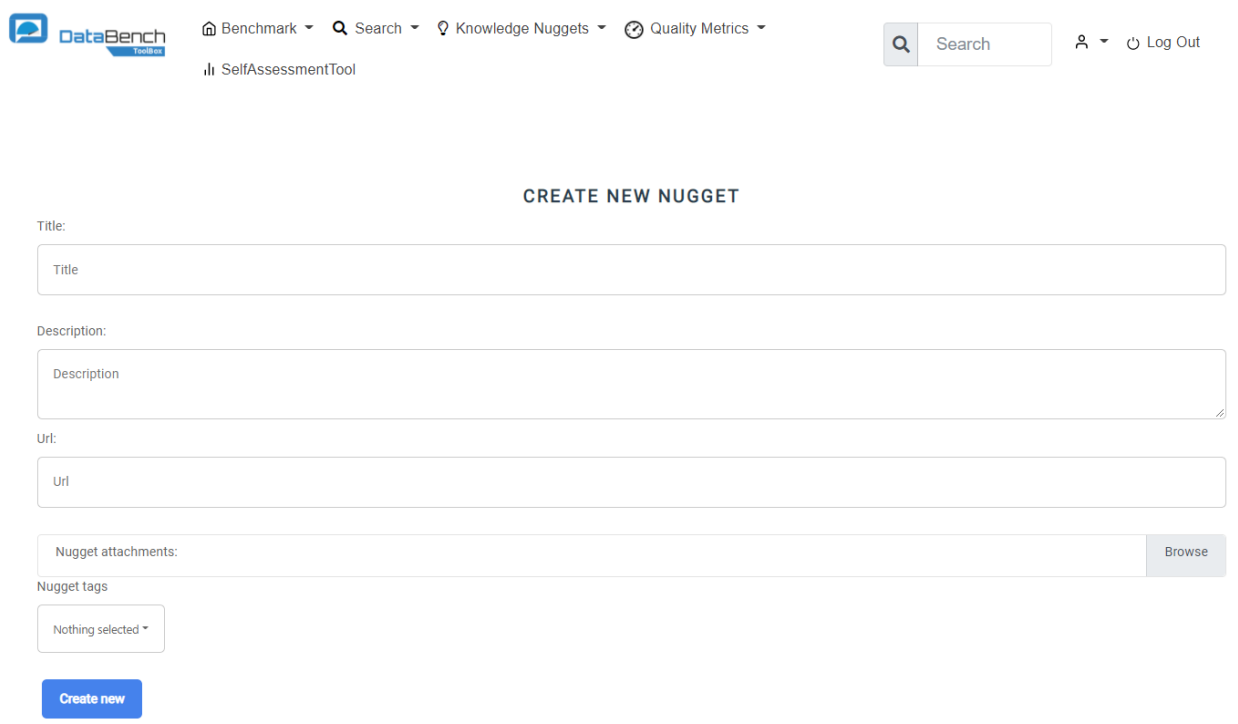

Figure 56 –Knowledge nugget creation form.

Note that the knowledge nugget can be created with a title (to be shown in the list of nuggets) and description (basic HTML tags are accepted), but also pointing to a URL of reference for external links and adding one or several attachments. The images attached will be also rendered within the nugget during in the visualization once the nugget is approved. The user might also propose some tags to annotate the nuggets from the set of existing tags (they might even propose new tags).

Once the content is proposed an editorial workflow starts where the user should contact via email to the Toolbox administrator who would revise the content and eventually approve or reject the nugget.

7. Next Steps

The Handbook and the DataBench toolbox are essential components of DataBench exploitation plan. The project is negotiating an agreement to deliver the Toolbox and the Handbook to the Big Data Innovation Hubs network through the IA EUHubs4Data.

The Handbook will have a second release at the end of the project to include the final results of the Toolbox validation which is due to be concluded in month 35 while this deliverable is due in month 34.

References

[1]. D3.4. Release Version of DataBench Toolbox including visualization and search components
[2]. D1.2. DataBench Deliverable D1.2. DataBench Framework – with Vertical Big Data Type benchmarks
[3]. D1.3 DataBench Deliverable D1.3. Horizontal Benchmarks – Analytics and Processing
[4]. D1.4 DataBench Deliverable D1.4. Horizontal Benchmarks – Data Management
[5]. Big Data Value Strategic Research and Innovation Agenda (BDV SRIA). http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf