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 that will provide access to benchmarking services, KPIs, and various types of knowledge. The DataBench Handbook plays a complementary role to the DataBench Toolbox by providing a comprehensive view of how the project developed and validated the benchmarks offered in the Toolbox and how technical and business benchmarking are linked in the project's research, and by 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 in delving more deeply into the project's results.

The DataBench Handbook and DataBench Toolbox are aimed at industrial users and European technology developers that need to make informed decisions on Big Data technology (BDT) investments to optimise their technical and business performances.

This Handbook demonstrates how DataBench has achieved its goal of designing a benchmarking process to help European organisations that are developing BDT to constantly improve their performances and strive for excellence. DataBench achieves this by measuring technology development activity against highly relevant business parameters. DataBench thus fills a gap in BDT knowledge and understanding.

Deliverable D4.4	DataBench Benchmarking Handbook
Work package	WP4
Task	4.4
Due date	31/12/2020
Submission date	31/05/2021
Deliverable lead	IDC
Version	3.0
Authors	IDC (Gabriella Cattaneo, Richard Stevens, Cristina Pepato, Erica Spinoni) Polimi (Chiara Francalanci) ATOS (Tomas Pariente Lobo, Ricardo Ruiz) Sintef (Arne Berre)
Reviewers	Polimi (Federica Acerbi), LEAD Consult (Todor Ivanov)

Keywords

Disclaimer

Copyright Notice

Table of Contents

Table of Figures

Table of Tables

Acronym and Abbreviations

Acronym	Title
AI	Artificial Intelligence
BDBC	Big Data Benchmarking Community
BDT	Big Data Technologies
BDV	Big Data Value
BDVA	Big Data Value Association
BSS	Business Support System
CMS	Content Management System
CRM	Customer Relationship Management
DIAS	Data Information Access Services
DIH	Digital Innovation Hub
EC	European Commission
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
HIPEAC	High Performance and Embedded Architecture and Compilation
HPC	High Performance Computing
HTML	HyperText Markup Language
ICT	Information and Communication Technologies
IoT	Internet of Things
KN	Knowledge Nuggets
KPI	Key Performance Indicator
LDBC	Linked Data Benchmarking Council

NFV	Network Function Virtualization
OSS	Operational Support System
PPP	Public Private Partnership
RoI	Return on Investments
SRIA	Strategic Research and Innovation Agenda
SME	Small and Medium Enterprise
TF	Task Force
TFP	Total Factor of Productivity
TRL	Technology Readiness Level
WP	Work Package

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 that will provide access to benchmarking services, KPIs, and various types of knowledge. The DataBench Handbook plays a complementary role to the DataBench Toolbox by providing a comprehensive view of how the project developed and validated the benchmarks offered in the Toolbox and how technical and business benchmarking are linked in the project's research, and by 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 in delving more deeply into the project's results.

DataBench Value Proposition

The DataBench Handbook and DataBench Toolbox are aimed at industrial users and European technology developers that need to make informed decisions on Big Data technology (BDT) investments to optimise their technical and business performances. The DataBench Toolbox helps stakeholders: identify use cases whereby they can prioritise their investments and achieve the highest possible business benefits and returns on investment; select the best technical benchmark to measure the performance of the technical solution of their choice; and assess their business performance and compare it with that of their peers so they can revise their choices and organisation if they find they are achieving lower results than the median benchmarks for their industry and company size. The services provided by the Toolbox and the Handbook will therefore support users in all phases of their journeys (before, during, and after, in the post evaluation of their BDT investments), from both a technical and a business viewpoint. The information provided by the Handbook is designed to support the users' in optimizing their use of the Toolbox.

This Handbook demonstrates how DataBench has achieved its goal of designing a benchmarking process to help European organisations that are developing BDT to constantly improve their performances and strive for excellence. DataBench achieves this by measuring technology development activity against highly relevant business parameters. DataBench thus fills a gap in BDT knowledge and understanding.

Conceptual Framework

The conceptual framework of DataBench research links the 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 the analysis of the state of benchmarking and the development of indicators (WP1), followed by data collection, in-depth research into industrial users' needs, and the measurement of business-need-based KPIs – all supported by 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 realise the

importance of benchmarking the business impacts of BDT, as shown in Figure 10, below. Only 10% of them dismiss it as not at all important or slightly important; 45% consider benchmarking very important or extremely important.

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

From case-study analysis, we see 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 specialised skills – skills that are not currently present in the company – and considerable investment. There is general agreement that BDTs are diverse and complex and that technical choices are not simple and are potentially impactful. Even if companies do not perform benchmarking, they have been found to rely on trusted external entities to compare technologies, such as IT consultants and systems integrators.

Technical Benchmarks

The DataBench Framework for Big Data and AI Benchmarks is based on Big Data Value Association (BDVA) 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 company size, and these benchmarks are also accessible for users in the Toolbox, in the Knowledge Nugget section. The Handbook provides the main benchmark 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 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 items, 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' (KNs).

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

Sustainability and Usability of results

This DataBench Handbook is designed to provide a guide to all the tools and services provided by DataBench also after the end of the project in December 2020, thereby ensuring their sustainability and exploitation in time. The main outcome of the project is of course the Databench Toolbox, but the background and in-depth analysis developed by the project will also remain fully available in the projects' deliverables as valuable shared knowledge. The project website and all related materials will be available online and all services, documentation and links will be maintained for at least 2 years from the end of the project on December 31st, 2020. As is the practice of IDC and many of the partners, given the low cost of maintaining the physical hardware, domains and network connectivity, this period will likely be extended. The project's coordinator will manage enquiries by interested users after the end of the project and make sure they will be answered.

Organizations interested in extracting the maximum value from DataBench results should consider the option to integrate the Toolbox in their own IT infrastructures as a benchmarking process and continue feeding updated content to it. DataBench partners have made available the necessary knowledge and data to enable other ICT projects, research organizations, stakeholder organizations or companies to take over and adopt our tools and results in a proactive way. More specifically, the consortium will collaborate with the H2020 Innovation Action EUHubs4Data, just started with 3 years planned duration, so that DataBench tools and results can be leveraged by the Big Data Innovation Hubs network for training and exploitation by European enterprises and possibly updated.

1 Introduction

1.1 Objective

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 that will provide access to benchmarking services, KPIs, and various types of knowledge. The DataBench Handbook plays a complementary role to the DataBench Toolbox by providing a comprehensive view of how the project developed and validated the benchmarks offered in the Toolbox and how technical and business benchmarking are linked in the project's research, and by 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 delving more deeply 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 Handbook and DataBench Toolbox are aimed at industrial users and European technology developers that need to make informed decisions on BDT investments by optimising technical and business performance.

This Handbook demonstrates how DataBench has achieved its goal of designing a benchmarking process to help European organisations that are developing BDT to constantly improve their performances and strive for excellence. DataBench achieves this by measuring technology development activity against highly relevant business parameters. DataBench thus fills 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 the technical results from those tools. The Handbook explains and provides the scientific and methodological background of the benchmarks provided in the Toolbox and links it to the latest techniques used in the main benchmarking communities, thereby providing a sound basis to inspire the users' trust and confidence in the DataBench Toolbox results and services. The Handbook and the Toolbox together will serve as a decision-support tool for companies approaching BDT applications 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 three years of the project, the emergence of artificial intelligence (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 the technical and business benchmarking of BDT, which represents the sound basis of the tools and services offered by the Toolbox.
- Chapter 3 demonstrates how DataBench benchmarks respond to European users' industrial needs and provides data to show their correspondence with stakeholder categories by industry and company size, which is at the basis of the Toolbox users' profiling and BDT 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 the 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 impact 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 key components of the DataBench Exploitation Plan 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 the business and technical evaluation of Big Data technology (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 in DataBench deliverable D1.1, *Industry Requirements with Benchmark Metrics and KPIs* [1]. Concerning the evaluation of business impacts, the methodology is illustrated in detail in D2.1, *Economic and Market Analysis Methodology* [2], and in D4.1, *Data Collection* [3].

The research process of the project (Figure 1) shows how the conceptual framework was developed on the basis of the analysis of the state of benchmarking and the development of indicators (WP1), followed by data collection, in-depth research into industrial users' needs, and the measurement of business-need-based KPIs – all supported by 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.

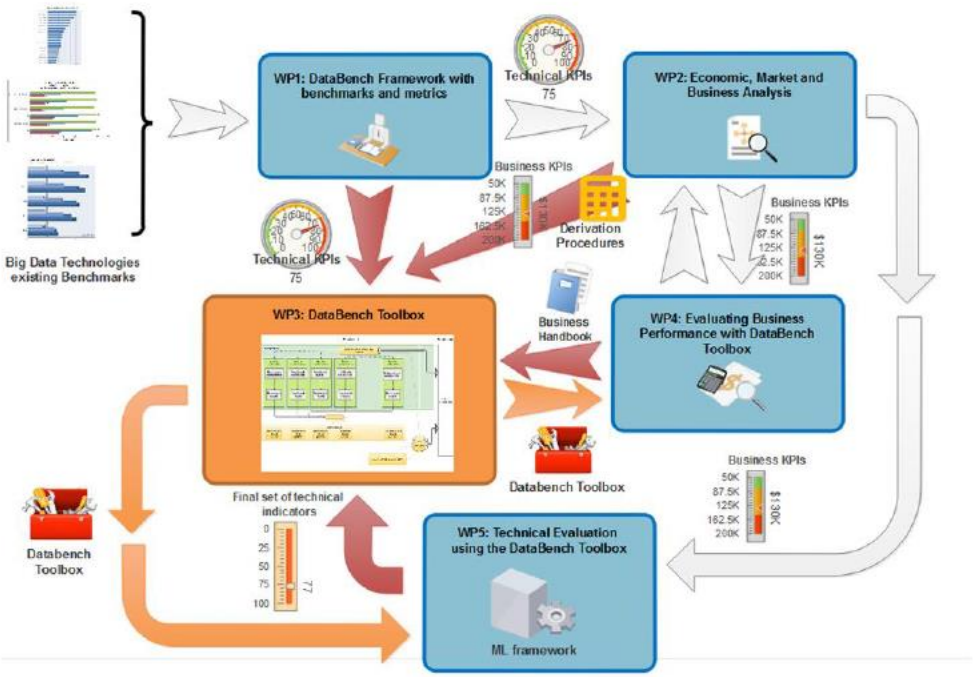

Figure 1, DataBench Research Process and Outcomes
Source: D.1.1, *Industry Requirements with Benchmark Metrics and KPIs* [1]

More specifically, Figure 2 provides a snapshot of the conceptual framework that links the technical and business benchmarking used in this project. As shown below, in Figure 2, the comprehensive analysis of the main Big Data technology 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 is in turn 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 which we mean the implementation of BDT solutions in specific business processes/use cases, which have also been identified and classified. The business impacts of these BDT use cases have been measured through the seven business KPIs selected by the project, based on the industrial users' survey and case studies, which measure aspects such as revenue and profit growth, customer satisfaction, and product and/or service innovation.

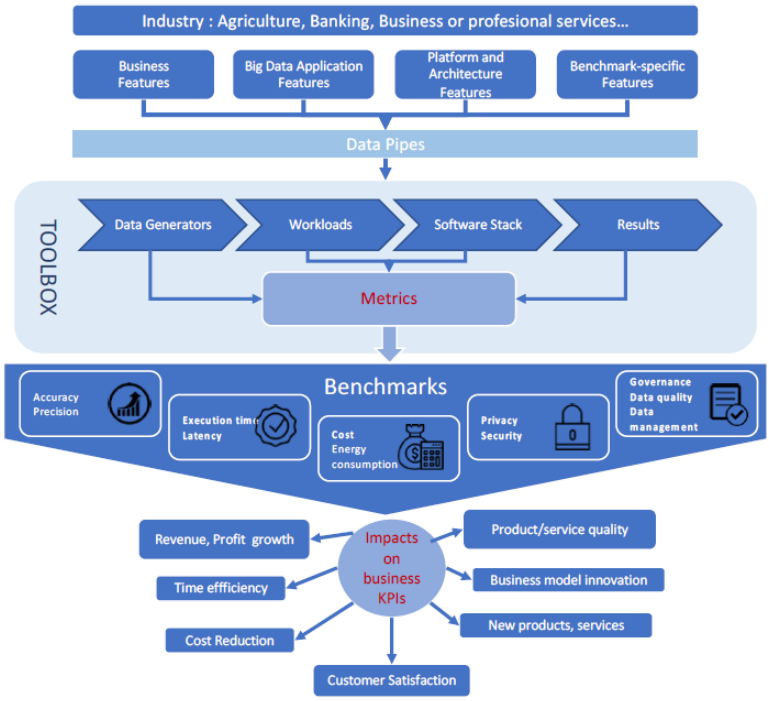

Figure 2, Technical and Business Benchmarking Framework
Source: DataBench D1.1, *Industry Requirements with Benchmark Metrics and KPIs* [1]

2.2 Value Proposition for Stakeholders

It is important to say that, from the start, DataBench targeted three main stakeholder categories, which are addressed differently by the DataBench Toolbox (as illustrated in Chapter 6):

- **Benchmarking Providers:** These are organisations that own a particular benchmark. They can be the actual developers of the benchmarks or the organisations 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:** These are users who would like to search for, and potentially execute, a technical benchmark. DataBench interacted with them to investigate their needs and requirements for benchmarking and designed services through the Toolbox to meet their needs.

- **Business Users:** These are users who would like to search for and understand the business value of specific Big Data solutions when making choices about BDT investments. The Toolbox provides them with data about the potential business benefits of the main use cases and access to business benchmarks by industry and company size. This enables the business users to compare themselves with their peers and their peers' business achievements.

The DataBench Toolbox helps stakeholders: identify use cases whereby they can prioritise their investments and achieve the highest possible business benefits and returns on investment; select the best technical benchmark to measure the performance of the technical solution of their choice; and assess their business performance and compare it with that of their peers so they can revise their choices and organisation if they find they are achieving lower results than the median benchmarks for their industry and company size. The services provided by the Toolbox and the Handbook will therefore support users in all phases of their journeys (before, during, and after, in the post evaluation of their BDT investments), from both a technical and a business viewpoint.

2.3 The Ecosystem of Indicators

The ecosystem of indicators developed by DataBench is shown in the Figure 3, below, and aligns with the latest research methodologies and best practices from the economic and market research viewpoints, 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).

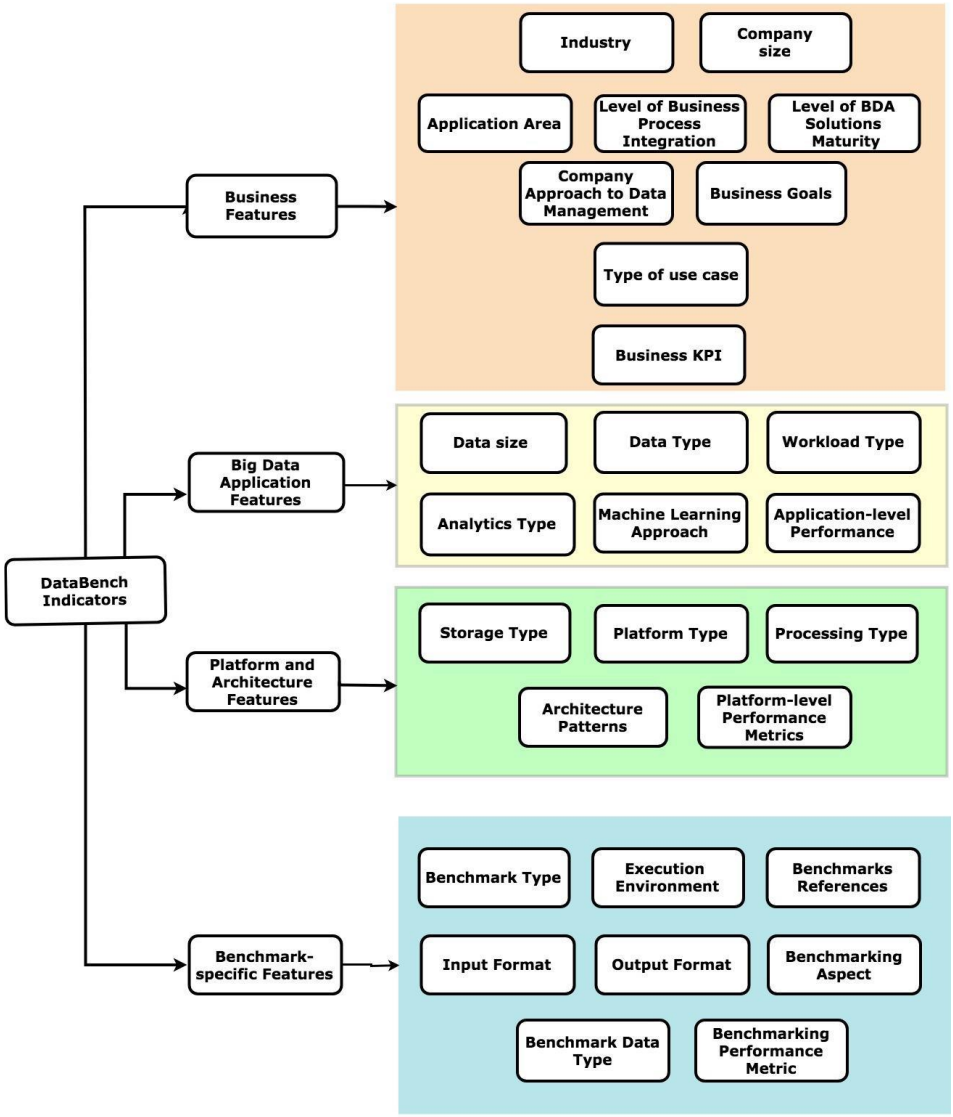

Figure 3, DataBench Indicators
Source: DataBench D1.1, *Industry Requirements with Benchmark Metrics and KPIs*

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 for prospective new technology investments.

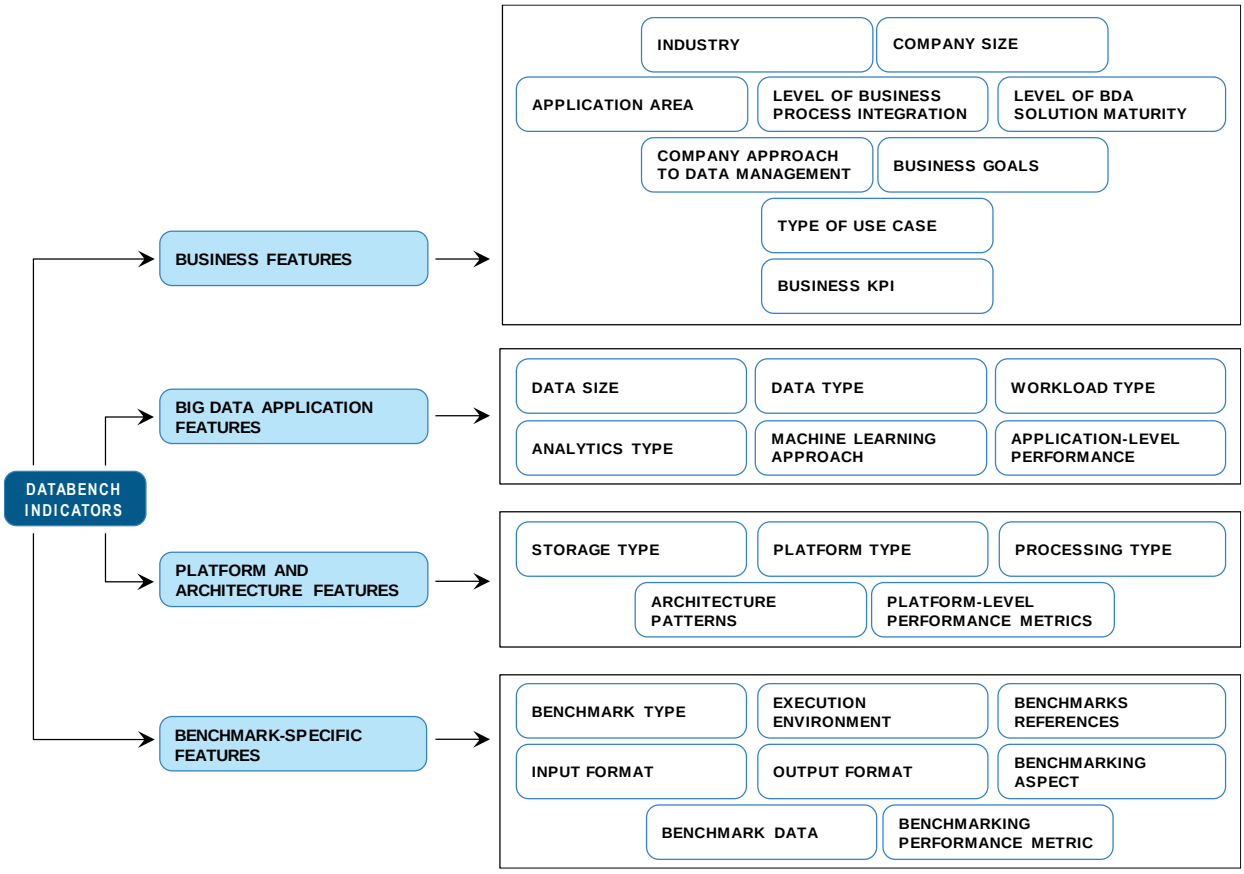

Figure 4, DataBench Business Indicators.
Source: DataBench D1.1, *Industry Requirements with Benchmark Metrics and KPIs*

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 we define in this project 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, and 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 also easily correlates with business impacts, it provides a way of evaluating how technology requirements may influence business outcomes. The classification of use cases measured in this project is part of the ecosystem of indicators, as 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 conducted 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), which are 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 BDT decision makers and influencers, 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 benchmark values. The following figures outline the composition of the interviews used to calculate the business benchmarks.

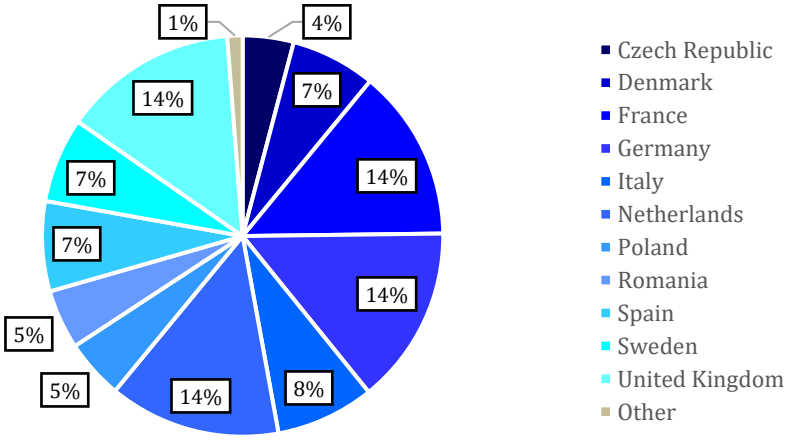

Figure 5, Respondents by Country. Source: DataBench Survey, June 2020

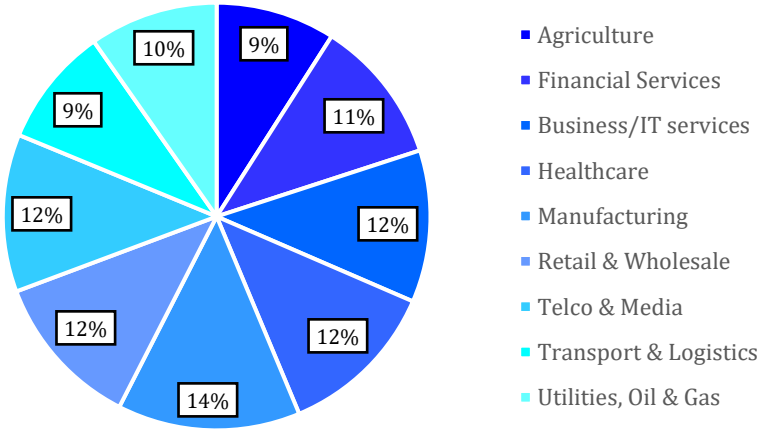

Figure 6, Respondents by Industry. Source: DataBench Survey, June 2020

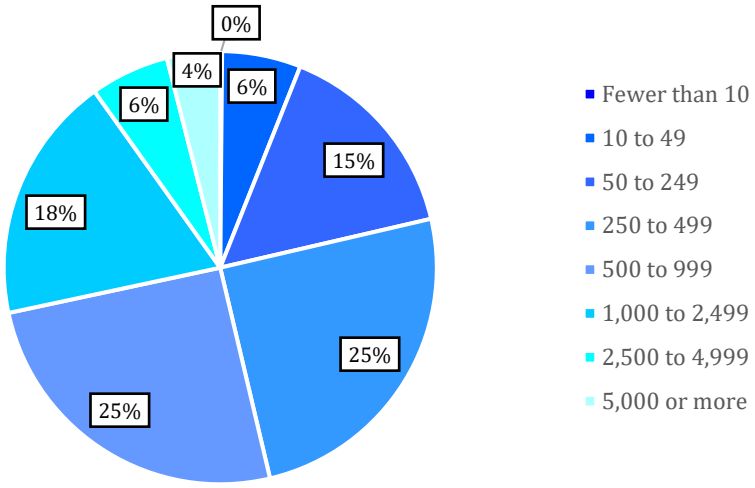

Figure 7, Respondents by Company Size. Source: DataBench Survey, June 2020

3 Stakeholder Analysis

3.1 Overview

DataBench conducted in-depth analysis into industrial users' needs for the adoption of BDT in order to tailor benchmarking services to demand requirements. Data collection 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 in Figure 8 (which displays data by cross-company function/activity), below, the priority business goals driving BDT investment concern the improvement of business processes and market understanding. Several other business goals were also mentioned by a high share of respondents, confirming that BDT is relevant for multiple business objectives. It is therefore logical that measuring BDT's business impacts is also relevant – namely, 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 to improve the customer and/or employee experience. Enterprises need to fight the fall of demand and provide safe (often contactless) customer experiences. This trend requires even higher investments in Big Data supported by AI powering user-friendly interfaces (e.g. conversational AI). In this context, the considerations presented here about industrial users' needs remain more than valid.

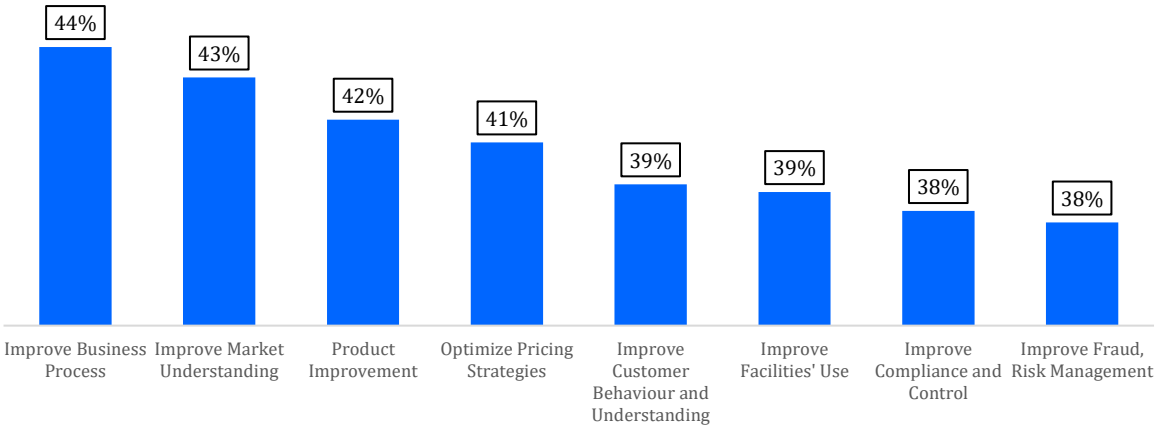

Figure 8, Business Goals Driving BDT Adoption (% of respondents).
Source: DataBench Survey 2018, n = 700, Deliverable D.2.2 [4]

DataBench results show a high level of BDT satisfaction or expectation, 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 piloting to scaling up BDT. Respondents still considering or evaluating BDT are more conservative in their expectations: The majority expects low or medium benefits from BDT.

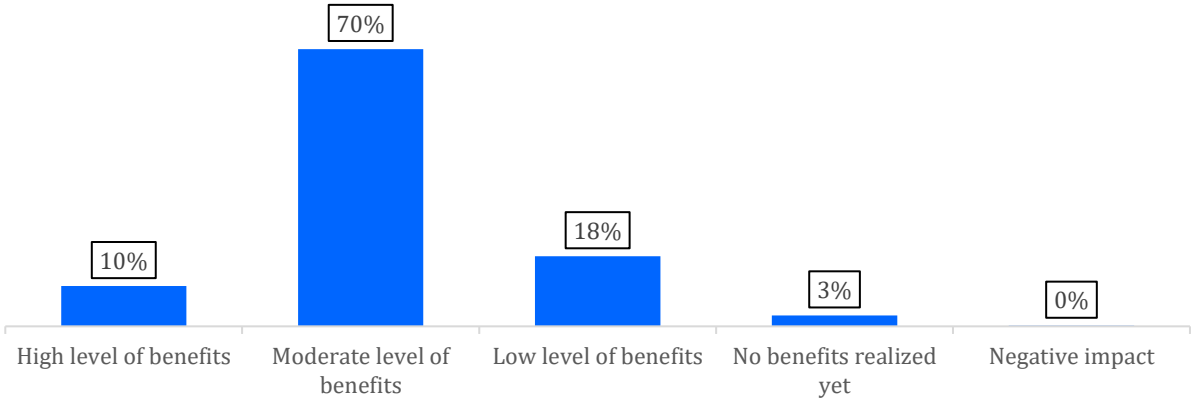

Figure 9, Expected Benefits of BDT Adoption (% of Respondents).
Source: DataBench Survey 2018, n = 700, Deliverable D.2.2 [4]

Business organisations realise the importance of benchmarking the business impacts of BDT. As shown in Figure 10, below, only 10% of them dismiss it as not at all important or slightly important; 45% consider benchmarking very important or extremely important.

The perception of the importance of benchmarking positively correlates with the level of adoption (actual users evaluate it very highly) and company size (with large companies more appreciative of its importance than small ones). Organisations in industries that are more advanced and sophisticated in the use of BDT (finance, retail, and telecom & media) again evaluate the importance of benchmarking more highly than others, while entities in laggard industries (healthcare and agriculture) have a higher share of respondents not particularly interested in benchmarking. The obvious deduction is that benchmarking becomes relevant when organisations 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.

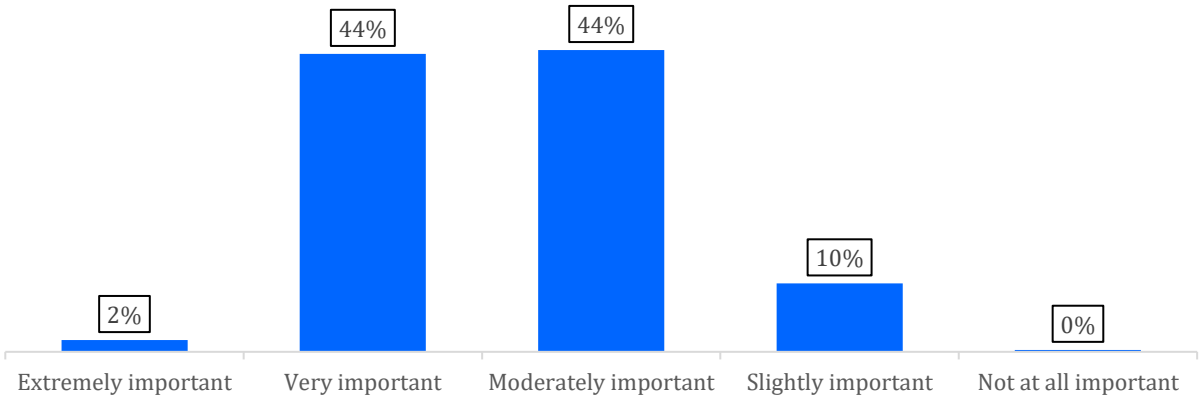

Figure 10, Importance of Benchmarking BDT Impacts (% of Respondents).
Source: DataBench Survey 2018, n = 700, Deliverable D.2.2 [4]

3.2 Adoption of BDT

Maturity around the adoption of BDT is highly variable by industry and company size, with the increasing intensity of adoption positively correlating with company size (figures 11 and 12). Adoption by industry also varies considerably, even though it is increasing fast in all sectors. Finance, business/IT services, and telecom & media lead in adoption rates, but retail, utilities, and manufacturing also have a relevant share of advanced users. Organisations in sectors where IT investments were historically lower, such as agriculture and healthcare, require BDT investment choices to be backed by strong business cases. They weigh their choices carefully and therefore have good reason to be interested in benchmarking business impacts.

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

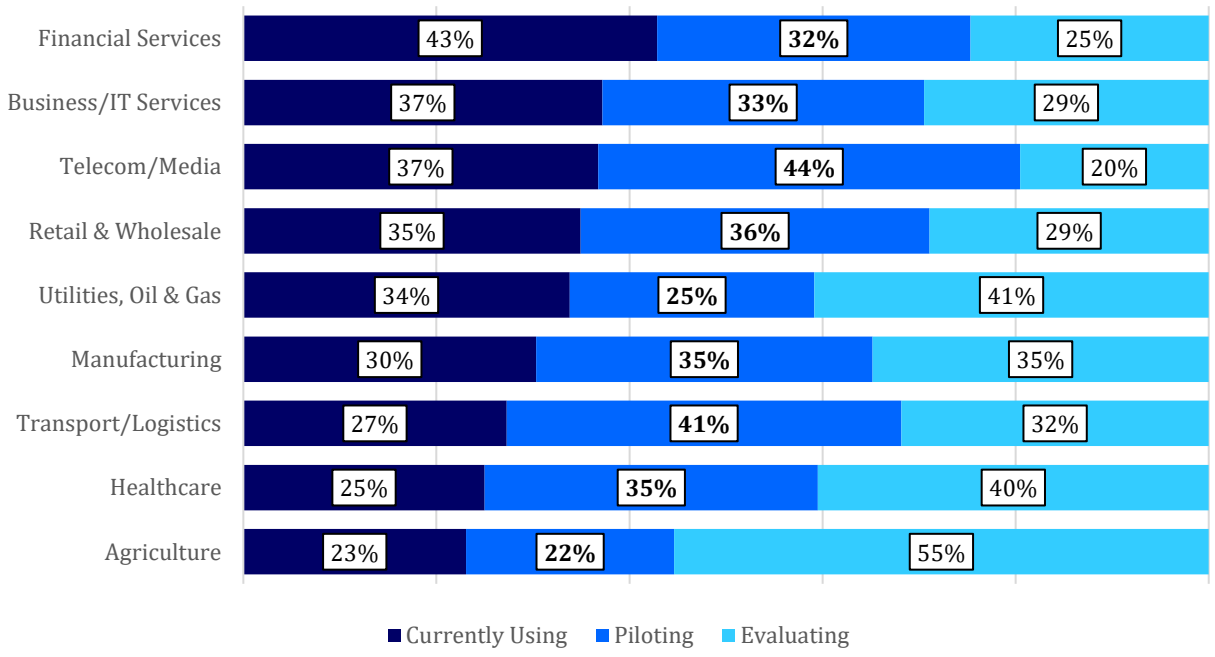

Figure 11, Current and Planned BDA Adoption by Industry.
Source: DataBench D2.2, *Preliminary Benchmarks of European and Industrial Significance* [4]

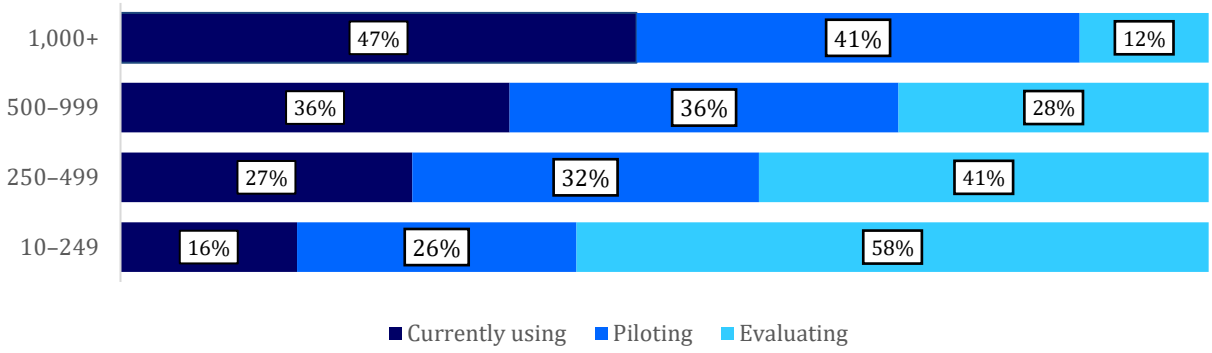

Figure 12, BDA Current and Planned Adoption by Company Size
Source: DataBench D2.2, *Preliminary Benchmarks of European and Industrial Significance*

3.3 Main BDT Use Cases

Within this project, use cases represent the link between technical solutions and business goals and help the collection of data for BDA exploitation typology – the main use-case types. The information about business benchmarking is stored in the Toolbox in what we have defined as Knowledge Nuggets (KN). In the Knowledge Nugget section of the Toolbox, users can search for business benchmarks by industry and use-case type in order to compare themselves with their peers or collect information about business impacts.

The DataBench survey investigated the adoption and maturity of 35 use cases identified through desk and previous field research as relevant, to have a pragmatic and realistic view of the footprint of BDT adoption in Europe. The use cases hereby defined pertain to storing, transforming, and analysing data and harnessing Big Data technologies as a way of enabling organisations to extract value from data to achieve the main business goals. The 35 use cases investigated in the survey include cross-industry use cases (potentially adopted across all industries or groups of industries) and industry-specific use cases (for example, prediction of patients' admission for healthcare) which were asked only to respondents of the relevant industries. The potential number of users (number of respondents in the survey) of industry-specific use cases is obviously lower than the potential number of users of cross-industry use cases. It is important to measure both the absolute number of users and the level of take-up out of the total potential users, which varies by use case. The first assesses the value of a use case for the whole European industry, the second assesses the value for specific industries.

This information is shown in Figure 13 below. The figure shows the ranking of use cases in terms of absolute take-up (the number of respondents currently evaluating or using each one, the shadow bars) and the number of potential users (the dark blue bars). For example, the number one case is Risk assessment with 381 users, representing a 54% share of the total potential users who were 700 (the whole sample, because this is a cross-industry use case). The use case ranked number 5, Automated customer service, has 210 users, but with a 55% level of take-up because the number of potential users (respondents) was only 380. Customer profiling is number 7 for absolute users' number, but first by level of take-up (70% of potential users).

Focusing the analysis on the top five use cases, we notice a mix of ways to use BDT oriented within and outside the organisation and common patterns across industries.

- **Risk Exposure Assessment:** This first use case is extremely relevant across all industries, especially for organisations in the process of evaluating current processes, services, and products, but also for those evaluating the introduction of new products/services. When business processes are under evaluation, risk exposure assessment can come into play, too, providing information on opportunities and risks related to new processes.
- **New Product Development:** New product (and service) development progresses with the adoption of Big Data technology because it helps organisations to (re)shape products (and services) according to customer needs and interests. This also links with product and offer personalisation/customisation.
- **Price Optimisation:** Product and service price optimisation is a complex mechanism that can be undertaken only once a BDA platform is in place and functions well, as price/offer optimisation involves profiling and targeting specific customer segments and tailoring offers and prices accordingly.
- **Regulatory Intelligence:** Big Data solutions and technologies are helpful in setting and managing regulatory compliance strategies and in building a regulatory-savvy yet data-centric company. A Big Data platform (or solution) helps an organisation capitalise on potential and real value by ensuring the use of data and adherence to next-generation regulatory compliance. Big Data technologies help businesses update and modernise their compliance processes, making them more precise and effective, follow regulatory changes easily (both domestically and internationally), and improve decision making processes.
- **Automated Customer Service:** In automating customer services, organisations can optimise responses to customers in terms of both timing – from call reception to handling and forecasting service completion – and cost. This process not only improves customer satisfaction; it also lowers call handling costs and human errors and their related 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 across all industries as a pivotal solution in digital transformation and the achievement of digital business objectives. The volume, variety, and velocity of data from multiple sources (both internal and external) is increasing, and the opportunity to exploit this data to gain a better understanding of current performances and areas of possible improvement is clear and valid.

The exploitation of Big Data is helpful for a large number of use cases, embracing both internal and external processes, such as optimising conversion rates, detecting and avoiding risks, streamlining operations, and monitoring customer behaviour, among others.

Enhancing decision-making processes is an ongoing effort for many organisations, and Big Data analytics can contribute to achieving business goals and profitable results. Big Data analytics solutions vary by vertical market; organisations in each sector implement Big Data solutions for purposes and business models specific to that sector.

3.4 Case Studies

Case study analysis is an important goal of WP4. Deliverable D4.2, *Data Collection Results* [5], provides a detailed description of the methodology used in case study analysis and a description of the status of research in WP4. It also draws preliminary conclusions. The final research results are presented in DataBench deliverable D4.3, *Evaluation of Business Performance* [6]. In the scope of the DataBench project, we have collected more than 700 articles, gathered from three main source types:

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

Each of these articles was tagged with different metadata, such as data size/magnitude, velocity, and source types. This metadata was thoroughly discussed in D4.2 and is 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 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 openness of the different companies regarding discussion and cooperation. Case studies involve a considerable effort. As a consequence, the goal is not to reach statistical significance and generality per se, but to provide insightful qualitative explanations of findings from extensive surveys (such as the DataBench survey and desk analysis), as well as to provide indications for subsequent research.

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

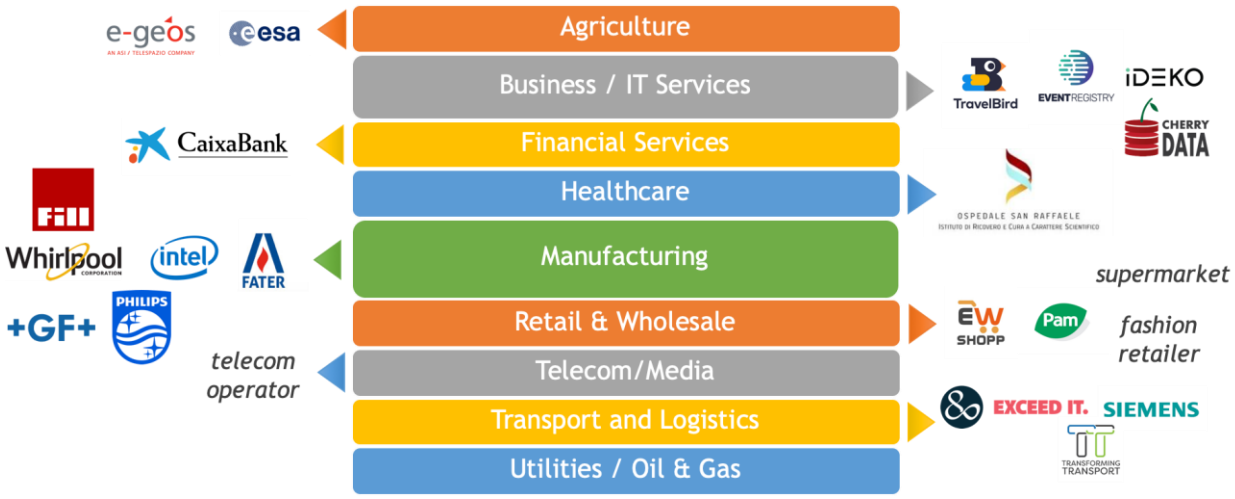

Figure 14, Case Studies (Total 22)
Source: DataBench Deliverable D4.3, *Evaluation of Business Performance* [6]

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 recognised as having better performance and obtaining 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 – the automation of manual work – and a threat – scepticism about the superiority of machines in decision making.
- **High awareness of opportunities**: Interest is significant in learning about the most frequent use cases in different industries and in taking part to projects to gain market competitive insights and obtain tools to benchmark performances for the industry.
- **Data the starting point**: From a technical standpoint, data- and processing-intensive analytics requires 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 must make an initial investment, choose a technology stack, and create a data lake to store their data. Thereafter, achieving a high level of data quality is complex and difficult. So, the business benefits from BDA are not low-hanging fruit, but rather the outcome of process changes that start from the data itself.
- **Data governance a concern**: This is a common concern and represents a general concept with organisational, technical, and legal implications. As governance means data security, this tight link can act as an obstacle but also as opportunity for technical innovation.
- **Deployment stage project limited:** Only a few projects reach the deployment stage due to a combination of technical, human, and organisational factors. Undoubtedly, the benefits associated with BDT are observable, even at early adoption stages. A lack of business benefits is not a hindrance to BDT uptake for various deep-rooted reasons.

Furthermore, a subset of our case studies provided us with business-KPI data. Overall, evidence from case studies aligns with the results from the DataBench survey and positions business impact in the 4–8% range. The companies that have measured a positive business impact have all developed their own software, focusing on select use cases and 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 pilots; others have already reached full-scale deployment (more detailed information in DataBench deliverable D4.3, *Data collection results* [6]).

As determined from the case-study analysis, 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 specialised skills – skills that are not currently present in the company – and considerable investment. It is generally agreed that Big Data technologies are diverse and complex and that technical choices are not simple and are potentially impactful. Even if companies do not perform benchmarking, they have been found to rely on trusted external entities to compare technologies, such as IT consultants and systems integrators.

On the other hand, companies acknowledge the variety and complexity of technical solutions for Big Data and envision the following risks:

- Realising they have chosen a technology that proves non-scalable over time
- Relying on cloud technologies that might create a lock-in
- Discovering that cloud services are expensive, especially as a consequence of scalability

From a technical benchmarking perspective, it is important that benchmarking is supported with tools that reduce complexity by guiding users along predefined 'user journeys' towards the identification and execution of benchmarks. The Toolbox addresses these needs with a large collection of knowledge nuggets that can help companies in framing their technical choices. The link between use cases, technical needs, and technical benchmarks is realised through the concept of a technology pipeline, as discussed in the next section. Different stages of the pipelines have been associated with technologies based on a reference technical blueprint, shown in Figure 15. The blueprint is general, as it indicates the building blocks of a BDA architecture, with no reference to specific technical solutions. For example, it reports NoSQL technologies, but specifies no particular NoSQL database, such as MongoDB, Cassandra, or HBase. Companies can use benchmarks for NoSQL databases to support software selection based on the technical performance parameters obtained through benchmarking. If different technologies perform differently, this software selection can be critical to ensure the overall performance of the architecture and to obtain the best cost-to-performance ratio. A detailed description of the methodology that can be used for sizing and costing is provided in D4.3 [6].

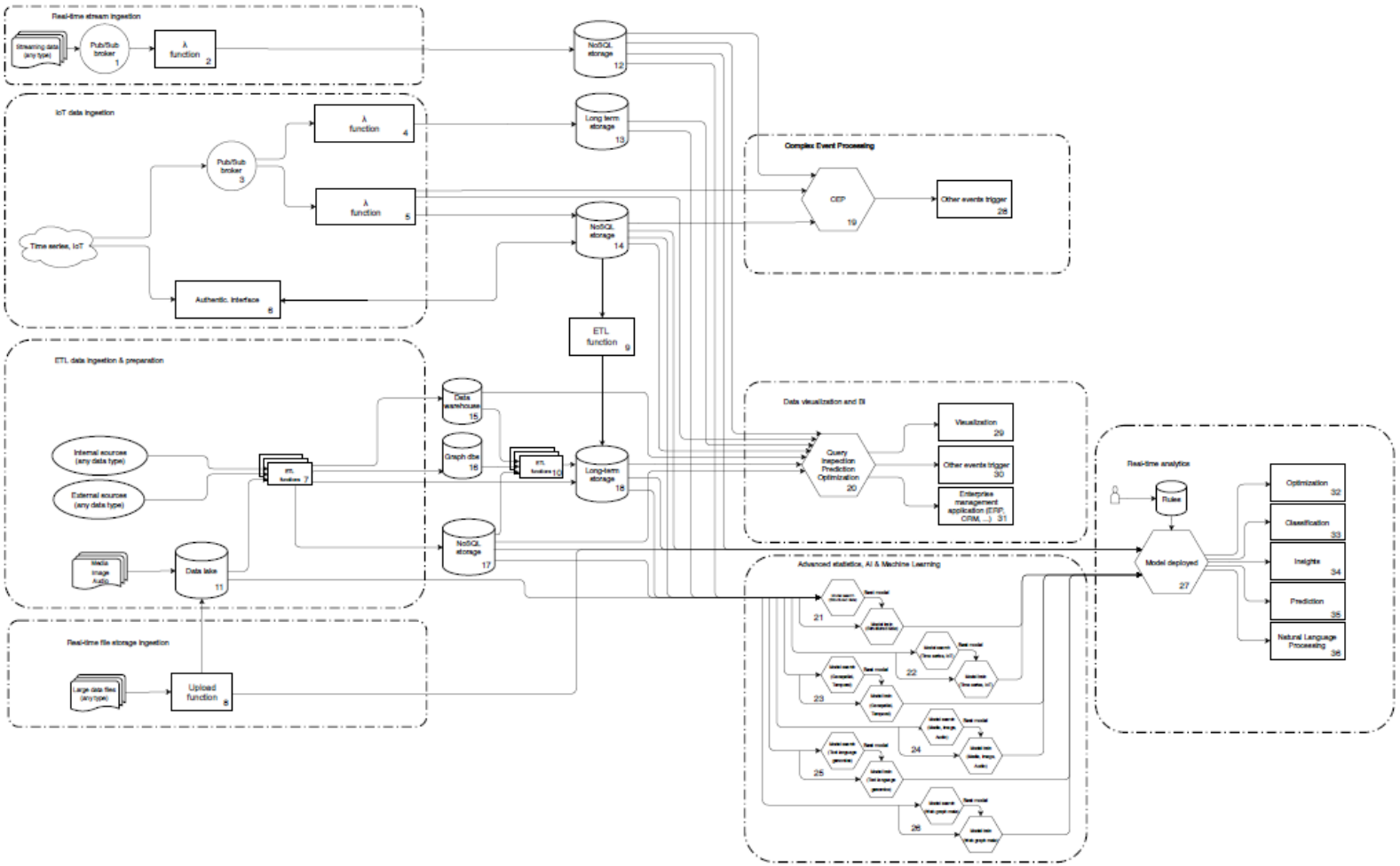

Figure 15, Reference Technical Blueprint

4 Technical Benchmarks

4.1 Overview

The DataBench Framework for Big Data and AI Benchmarks is based on Big Data Value Association (BDVA) reference architecture. A top-level generic pipeline has been introduced to provide an overall usage perspective on Big Data and AI systems. This generic pipeline helps the user understand the connections between and flow across the various parts of Big Data and AI systems and applications.

The following figure 16 depicts a top-level Big Data and AI value chain pipeline in the context of technical benchmarks and business benchmarks. It is abstract enough to be customisable for specific pipelines, depending on data type and processing used (e.g. IoT data and real-time processing). The top-level pipeline depicted in Figure 16 contains four major steps, which are analysed below.

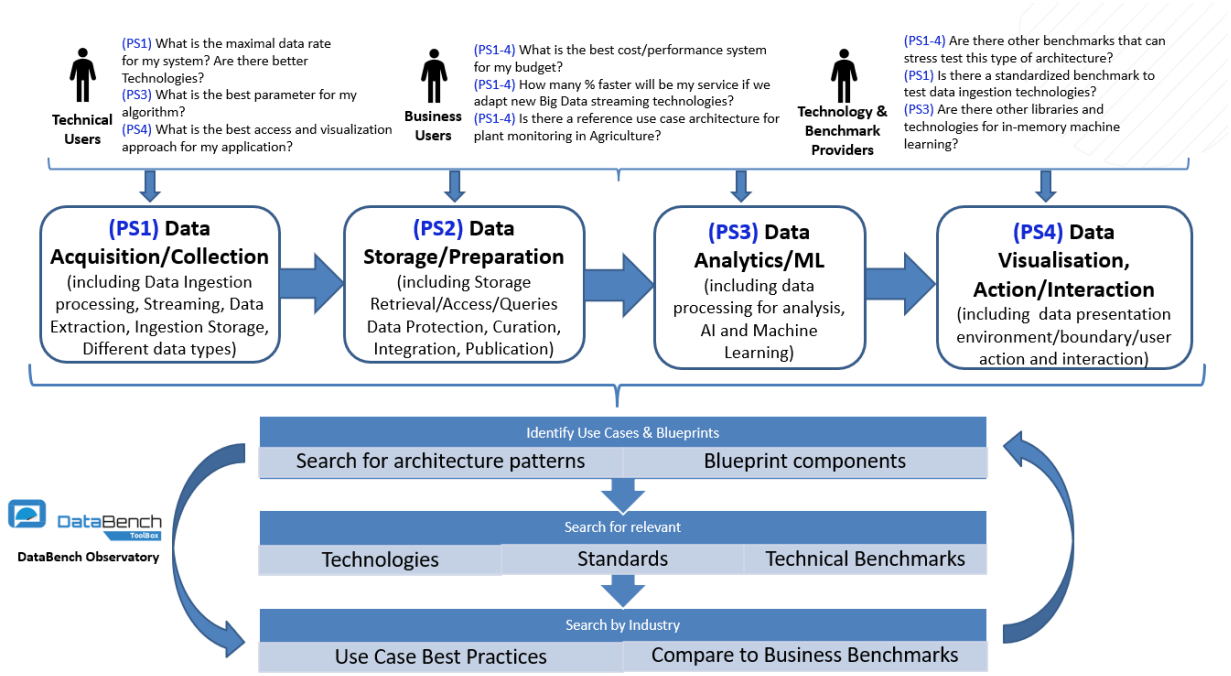

Figure 16, Top-Level Generic Pipeline, Including Methodology Steps (from D5.4 [13])

These steps are in compliance with the activities described in the reference architecture for Big Data applications.

A brief description of the steps in the Big Data and AI pipeline and the related BDV Reference model areas is presented in the following paragraphs. (More details can be found in the DataBench deliverables D5.5 *Final report on methodology for evaluation of industrial analytic projects scenarios* [7] and *D1.2 DataBench Framework* [8].

4.1.1 Data Acquisition and Collection

This step includes acquisition and collection from various sources, including both streamed data and data extraction from the relevant external data sources and data spaces. It includes support for handling all relevant data types, as well as the relevant data protection handling for this step. This step often relates to the use of both real-time and batch data collection and the associated streaming and messaging systems. It uses enabling technologies to collect streamed data in motion from things/assets, sensors, and actuators and connects to existing data sources for data at rest. This step often also includes the use of relevant communication and messaging technologies.

4.1.2 Data Storage and Preparation

This step includes the use of appropriate storage systems and data preparation and curation for further data use and processing. Data storage includes storage and retrieval in various databases systems, both SQL and NoSQL – more specifically, key-value column-based storage, document and graph storage, and storage structures, such as file systems. Historically, many benchmarks exist in this area for testing and comparing data storage alternatives. Tasks performed in this step also include data annotation, publication, and presentation for discovery, reuse, and preservation, and interaction with various data platforms and data spaces for broader data management and governance. Handling data protection and all aspects thereof is another part of this step.

4.1.3 Analytics, AI, and Machine Learning

This step consists of data analytics and related processes, including descriptive, predictive, and prescriptive analytics and use of AI/machine learning and algorithms to support decision making and knowledge transfer. In terms of machine learning, this step includes the subtasks for necessary model training and model verification/validation and testing – before actual operation with input data. In this context, the previous step of data storage and preparation provides data input both for training and validation and for test data, as well as operational input data.

4.1.4 Action and Interaction, Visualisation, and Access

This step (which includes the data presentation environment, boundaries, and user action and interaction) identifies the boundaries of the environment for action/interaction – typically, through a visual interface with various data visualisation techniques for human users and through an API or an interaction interface for system boundaries. Interaction boundaries are between machines and objects, machines and machines, people and machines, and environments and machines. Action/Interaction across system boundaries typically impacts the data acquisition/collection step, in which system boundary inputs are collected.

4.1.5 Following actions

These 4 methodology steps can be customised based on the different data types used in the various applications or projects' pilots and can be set up differently for each processing architecture, whether batch, real-time/streamed, or interactive. They are also suitable for Machine Learning which starts with training data and then uses operational data. The steps of the Big Data and AI Pipeline Framework align with ISO SC42 AI Committee Standards –

in particular, with the collection, preparation, analytics, and visualisation/access steps in the Big Data application layer, as covered by recent international standard ISO 20547-3 Big Data Reference Architecture (more specifically, the functional components of the Big Data Reference Architecture, [18].

4.2 DataBench Framework

The DataBench Framework is based on the structure of the BDVA architecture model and includes vertical and horizontal benchmarks, in line with this model, along with business-oriented benchmarks.

Industry-based use cases are analysed to derive examples and metrics for each of the Big Data types. The idea is to reuse and adapt the established benchmarks for structural data (BigBench, BigDataBench, TPC, and others) and graph data/linked data (Hobbit I-IV and LDBC 1–3) and incorporate benchmark proposals related to time series/IoT (Yahoo Stream Benchmark, RIoTBench, StreamBench, and others) and inputs from DataBench partners' research benchmarks on streaming sensor data (ABench [UFRA] and SenseMark [SINTEF]).

Similarly, image, audio, video, text/NLP, and other media data types are used in analytics and process benchmarking and are relevant for machine learning (DeepBench, DeepMark, and others). A final relevant area for vertical benchmarks is the effect of technology support on data privacy and security. A set of projects related to supporting data privacy has been started under 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 that align with the following Big Data types:

- Structured data benchmarks
- IoT/Time-series benchmarks
- Spatiotemporal benchmarks
- Media/Image benchmarks
- Text/NLP Benchmarks
- Graph/Metadata benchmarks

BDV Reference Model [19], as shown in Figure 17, has been developed by BDVA and takes into account inputs 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 ensure logical connections with other areas of the digital platform, such as cloud, high-performance computing (HPC), IoT, networks/5G, and cybersecurity, among others.

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

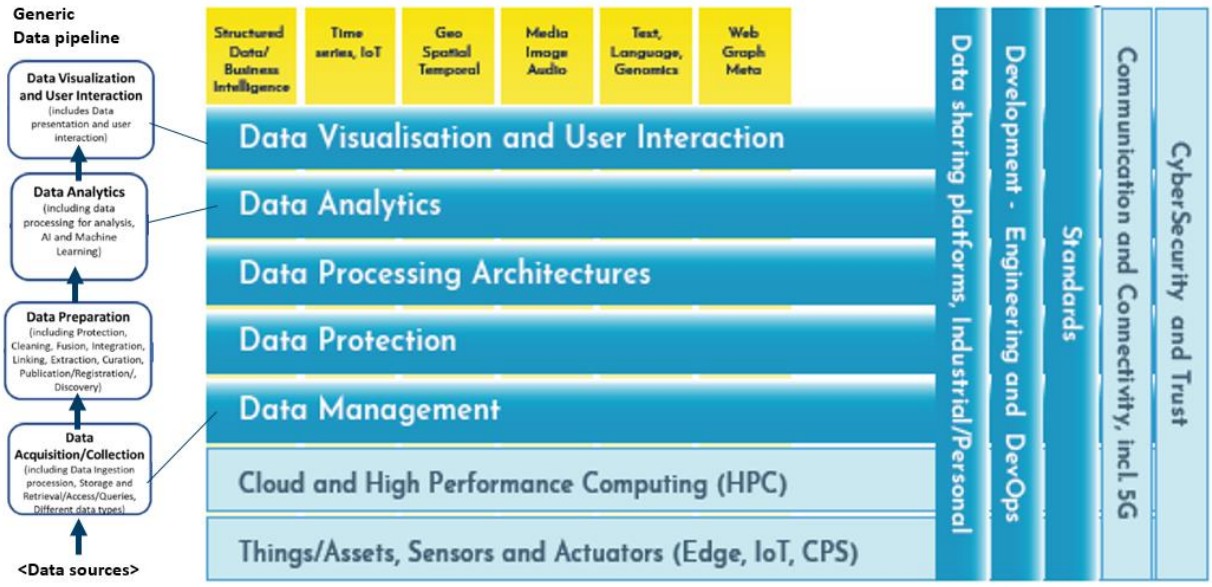

Figure 17, BDV Reference Model as a Foundation for the DataBench Framework (Related to the Generic Pipeline)

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 visualisation. It should be noted that the horizontal concerns do not imply a layered architecture. As an example, data visualisation may be applied directly to collected data (data management aspect) without the need for data processing or analytics. Further data analytics might take place in the IoT area (e.g. edge analytics). This shows logical areas, but they may execute in different physical layers.
- **Vertical concerns** address cross-cutting issues that may both affect horizontal concerns and involve non-technical aspects (e.g., standardisation as both a technical and a non-technical concern).

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, which is now being further developed in ISO JTC1 SC42 Artificial Intelligence.

The following technical priorities, as expressed in the BDV Reference Model, are covered 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 often relates to the usage of various combinations of the identified Big Data types described in the vertical concerns.

- **Data Visualisation and User Interaction**: Advanced visualisation approaches for improved user experience.
- **Data Analytics**: Data analytics to improve data understanding, deep learning, and the meaningfulness of data.
- **Data Processing Architectures**: Optimised 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 of encryption. This area is also associated with the areas of cybersecurity, risk, and trust.
- **Data Management**: Principles and techniques for data management, including both data lifecycle 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 covered separately, in conjunction 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: 1) structured data; 2) times-series data; 3) geospatial data, 4) image, video, audio, and other media data; 5) text data, including NLP data and genomics representations; 6) graph data, network/web data, and metadata. These types often demand the use of different techniques and mechanisms in the horizontal concerns – for example, in terms of data analytics and data storage. In addition, it is important to support both the syntactic and semantic aspects of the data for all Big Data types.
- **Standards**: The standardisation of Big Data technology areas to facilitate data integration, sharing, and interoperability.
- **Communication and Connectivity**: Effective communication and connectivity mechanisms are necessary to provide support for Big Data. This area is covered separately, in conjunction with various communication communities, such as the 5G community.
- **Cybersecurity**: Big Data often needs 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 covered separately, in conjunction with the cybersecurity PPP community.
- **Engineering and DevOps**: For building Big Data Value systems. This area is covered with the Networked European Software and Service Initiative (NESSI) community.
- **Data Platforms**: Marketplaces, IDP/PDP, and 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, such as research data platforms (RDPs) and urban/city data platforms (UDPs). These platforms include the efficient use of a number of the horizontal and vertical Big Data areas – most notably, the areas of data management, data processing, data protection, and cybersecurity.

- **AI Platforms**: In the context of the relationship between AI and Big Data, the BDV Reference Model is being refined to reflect data platforms' support of AI solutions and their compatibility with Machine Learning, analytics, visualisation, and processing solutions, among others (e.g. in the upper technology areas), as well as their compatibility with all Big Data types.

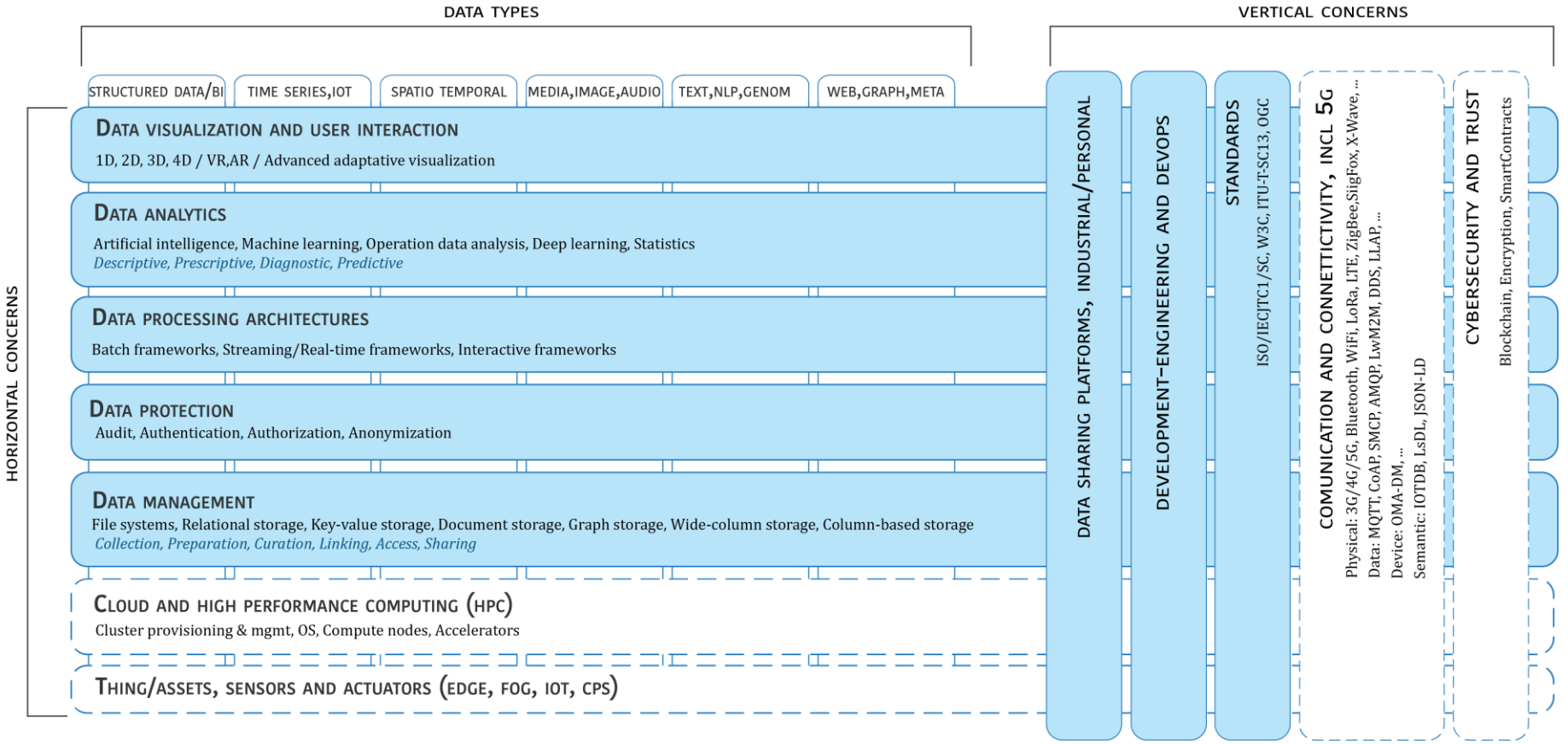

Figure 18, Refinement of the BDVA Reference Model – Horizonal Concerns

The BDV Reference Model may serve as common reference framework to locate Big Data technologies in the overall IT stack. The 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 is based on defining sub-categories within each of the reference model areas, in alignment with the refinement of the respective areas in the ISO SC42 suite of standards and technical reports currently in progress. The subcategories describe typical technology types within each of the areas, which is relevant in benchmarking context.

The modelling approach in the figure is at the top level so as to describe the logical technical areas within a wider Big Data and AI platform and the relevant subcategories within each area. In addition to the technical subcategories, it identifies the typical process steps in a Big Data pipeline that are relevant for the various areas. The project has consolidated and unified the models, metamodels, and ontologies from D1.1 [1], D3.1[9], D5.1 [10], and D1.2 [8]. and the companion D1.3[11] and D1.4[12] public deliverables. The results are provided in D.5.4 *Analytic modelling relationships between metrics, data and project methodologies* [13] and D.5.5 *Final report on methodology for evaluation of industrial analytic projects scenarios* [7].

4.2.3 Data Visualisation and the User Interaction Layer

This layer incorporates research areas related to the science of analytical reasoning, assisted by advanced visualisation 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 searches 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 visualisations (including zooms, dynamic filtering, and annotation) for the quick and correct interpretation of data. A focus on the relevance and relatedness of information for efficient search and exploration.
- **Cross-Platform Mechanisms for Data Exploration, Discovery, and Querying:** Uniform data visualisation on a range of devices.
- **Innovating Reporting:** Innovative multi-device reports and dashboards (dynamic, 3D, augmented-reality dimensions, etc.).
- **Domain-Specific Data Visualisation Techniques:** Innovative techniques and approaches to visualise data from specific domain (graphic, geospatial, sensor, and mobile data, among others).

4.2.4 DataBench Pipeline, Framework, and Available Benchmarks

The DataBench Framework is based on the structure of the BDVA architecture model and includes vertical and horizontal benchmarks, in line with this model, along with business-oriented benchmarks. This is illustrated in Figure 19, below.

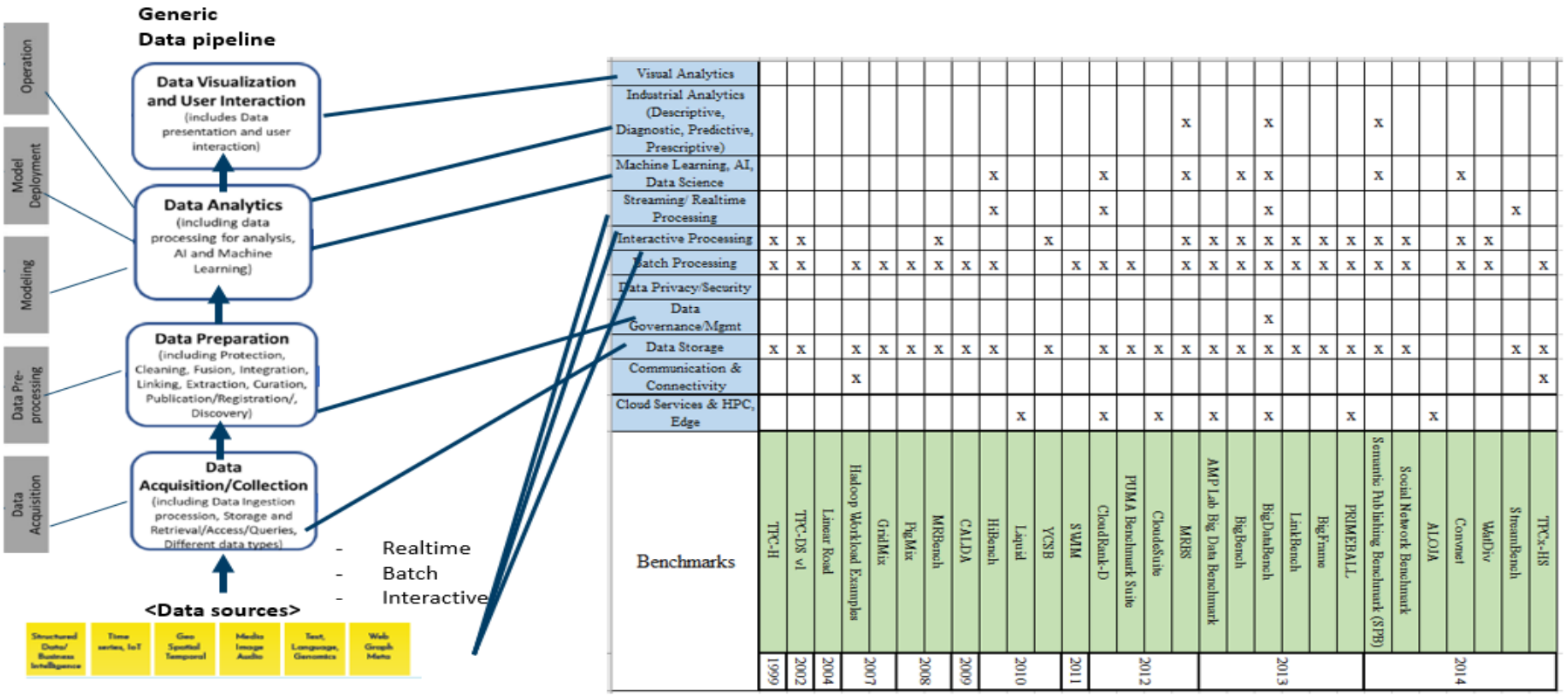

Figure 19, Refinement of the BDVA Reference Model – Generic Data Pipeline

4.2.5 Examples of Related Generic Pipelines

This paragraphs provides examples of the Big Data and AI Pipelines leveraged in DataBench and illustrated more in detail in D.5.4 [13].

4.2.5.1 Pipeline for IoT Data Real-Time Processing and Decision Making

The "Pipeline for IoT data real-time processing and decision making" has been applied to three pilots in the DataBio [17] project from the agriculture and fishery domain. Since it is quite generic, it can also be applied to other domains. The main characteristic of this pipeline is the collection of real-time data from IoT devices to generate insights for operational decision making by applying real-time data analytics to the collected data.

Figure 20 shows the steps of the pipeline for real-time IoT data processing and decision making that we have just described and their mapping to the steps of top-level Generic Big Data and AI Pipeline pattern.

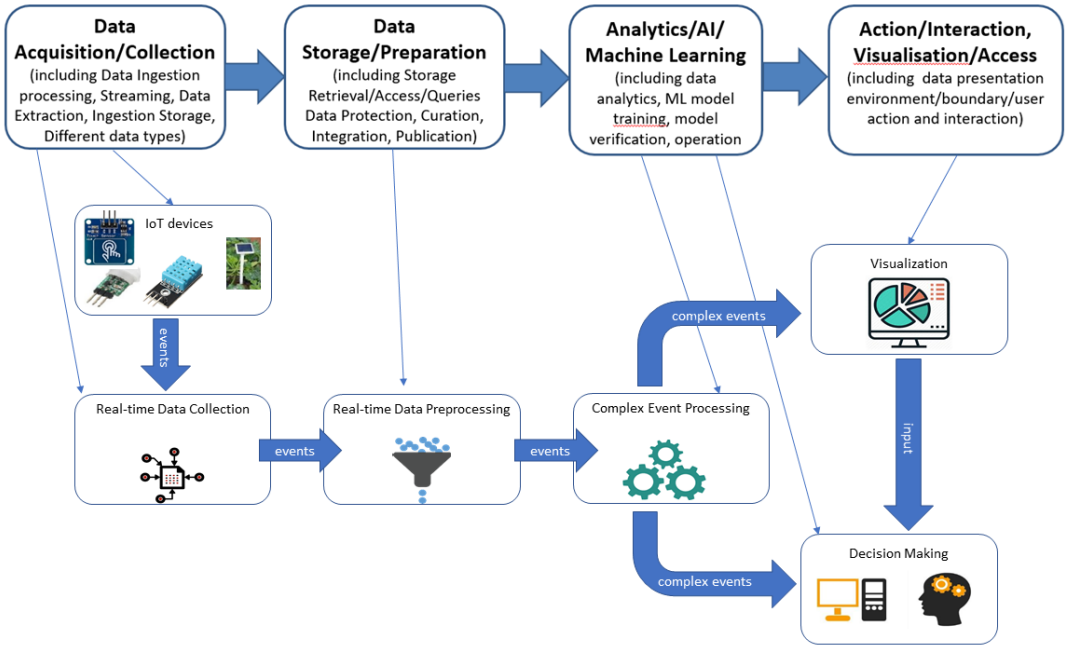

Figure 20, Example of an IoT Pipeline Pattern – Source D.5.4 page 19 [13]

4.2.5.2 Pipeline for Linked Data Integration and Publication

In the DataBio project [17] and some other agri-food projects, linked data has been extensively used as a federated layer to support the large-scale harmonisation and integration of a large variety of data collected from various heterogeneous sources and to provide an integrated view of this data. A triplestore populated with linked data during the course of the DataBio project (and a few other related projects) resulted in a repository of over 1 billion triples – one of the largest semantic repositories related to agriculture, as recognised by the EC innovation radar, which named it the "Arable Farming Data Integrator for Smart Farming". Projects like DataBio have helped deploy endpoints, providing access to dynamic data sources in their native linked-data format through an additional virtual semantic layer.

This action has been realised in the DataBio project through instantiations in 'the pipeline for the publication and integration of linked data', which has been applied in different uses cases related to bio-economy sectors.

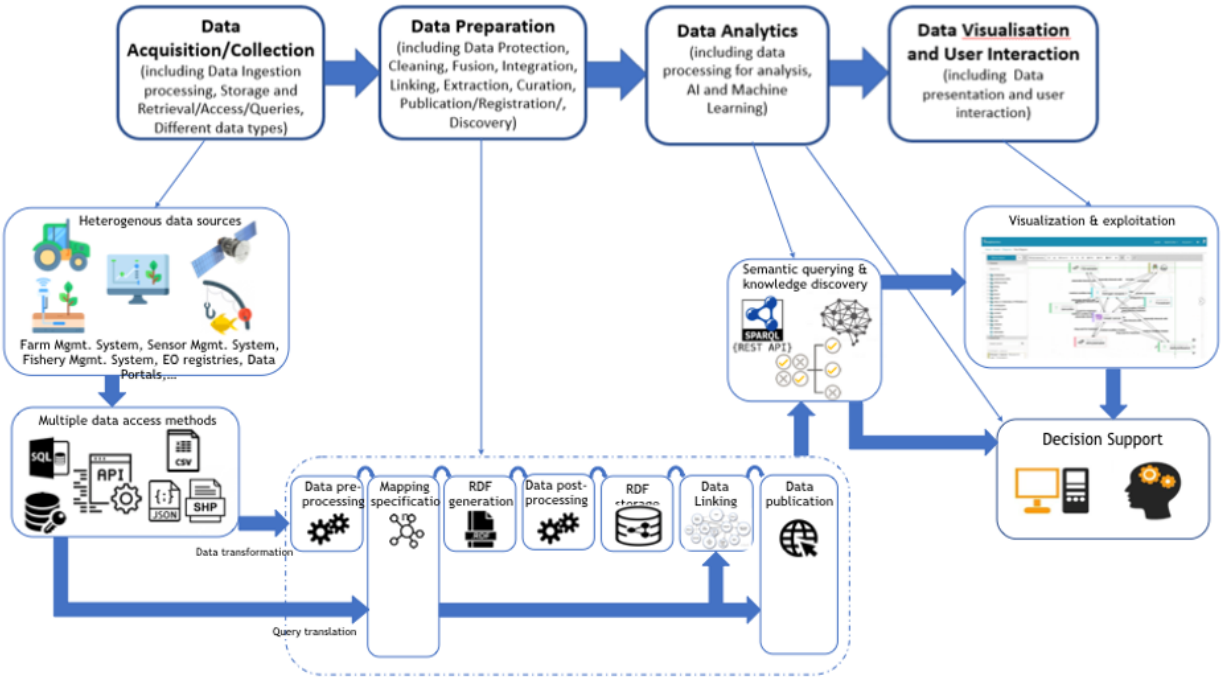

Figure 21, Example of Graph/Linked Data Pipeline Pattern Source:_D.5.4 page 20 [13]

The top-level Generic Big Data and AI Pipeline pattern discussed in the previous chapters has been used as a reference to build architectural blueprints that specify the technical systems/components needed at different stages in a pipeline. For example, in the data acquisition phase, a software broker to synchronise the data source and the destination is needed. A data acquisition broker then sends the data to a lambda function that transforms the data in a format that can be stored in a database. In DataBench, we have identified and classified all these technical components. This classification work has been performed with an empirical bottom-up approach, starting from Big Data analytics (BDA) use cases and then recognising the commonalities among the technical requirements of different use cases and designing a general architectural blueprint, as depicted in Figure 22.

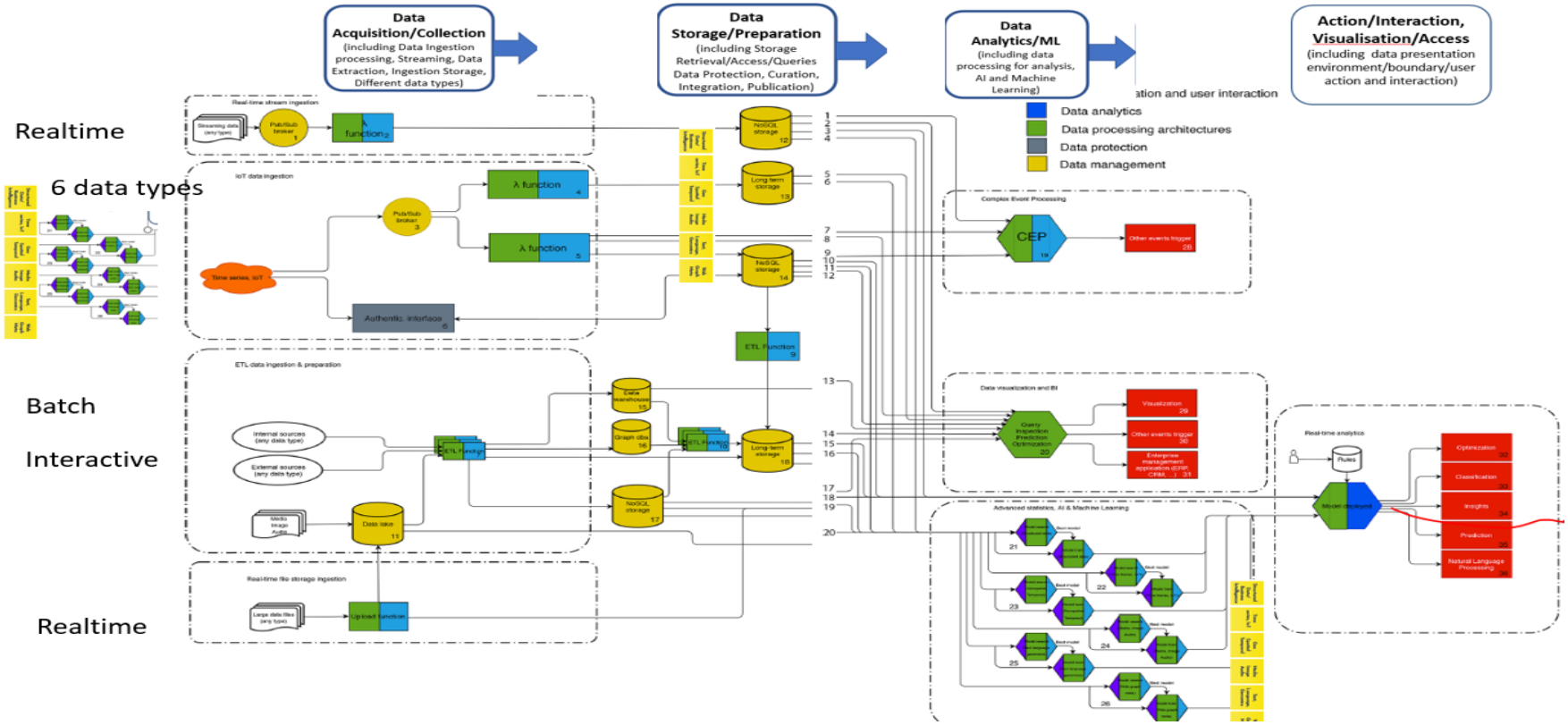

Figure 22, DataBench General Architectural Blueprint – source: D.5.4 page 23 [13]

From a technical benchmarking perspective, it is important that benchmarking is supported with tools that reduce complexity by guiding users along predefined user journeys towards the identification and execution of benchmarks. The Toolbox addresses these needs with a large collection of knowledge nuggets that can help companies frame their technical choices. The link between use cases, technical needs, and technical benchmarks is realised through the concept of a technology pipeline, as discussed in the next section. Different stages in the pipelines have been associated with technologies based on a reference technical blueprint, as shown in Figure 16. The blueprint is general, as it indicates the building blocks of a BDA architecture, with no reference to specific technical solutions. For example, it reports NoSQL technologies, but does not specify any particular NoSQL database, such as MongoDB, Cassandra, or HBase. Companies can use benchmarks for NoSQL databases to support software selection based on technical performance parameters obtained through benchmarking. If different technologies have different performances, this software selection can be critical to ensure the overall performance of the architecture and to obtain the best cost-to-performance ratio. A detailed description of the methodology for sizing and cost assessment is provided in D4.3 [6].

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 operationalised 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 ensure 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 for 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

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; the second batch comprises soft qualitative KPIs that cannot be easily or 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 23, Benchmarks Overview

Source: DataBench Deliverable D.2.4 [14]

5.2 Business Benchmarks by Industry

This chapter describes the BDT business benchmarks by industry, which are presented in the framework of 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 Deliverable D.2.4, *Benchmarks of European and industrial significance* [14].

5.2.1 Agriculture

As charted below, organisations in the agriculture sector evaluate Big Data solutions in terms of profit increases, even though margins are currently small. Increasing the precision of agriculture and yield predictions could boost profits and broaden margins. In having a precise view of production, farmers can optimise the use of land and seeds, pushing to the full the exploitation of their capabilities. For now, investments are mostly undertaken by larger agricultural organisations. But, with technology's dropping prices, even smaller agricultural concerns will be able to access these technologies.

The ability to plant seeds in an optimal/efficient way through precision agriculture (distancing seeds based on potential plant growth) – controlling the productivity of seeds and soil and forecasting it (yield monitoring and prediction) – and the application of predictive machinery maintenance will help organisations to organise activities better within the year (exact date to plant and harvest, the optimisation of crop rotation, maintenance schedules, etc.), improve revenue streams, and increase margins.

In this regard, the qualitative KPIs of product/service quality and time efficiency play relevant roles. Providing higher-quality and more nutritious produce on time will increase bargaining power when setting yield prices. Hence, farmers are not so interested in finding new sources of revenue or inventing completely new services, but they are expecting better value from data to enable them to change traditional/old-fashioned gut feeling-based business decisions.

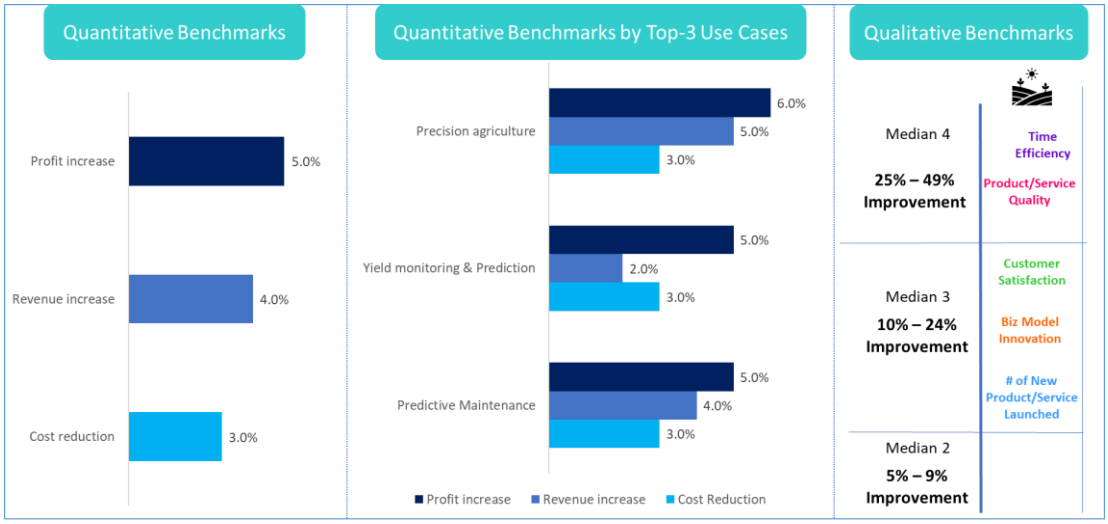

Figure 24, BDT Benchmarks: Agriculture
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.2 Financial Services

The finance 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 for the variety of service providers operating in this sector, from high street banks and insurance companies to FinTechs and investment management providers.

The top two use cases — customer scoring and/or churn mitigation (to apply credit ratings and/or to predict customer churn) and customer profiling & targeting and offer optimisation — are essentially about increasing revenue and profit. The third top-rated use case, fraud prevention and detection (predicting whether new customers are potential fraudsters and assessing whether specific transactions are legal), relates more to cost reduction than to profit and margin increases. This is because being able to assess potential scams quickly and avoid false positives helps organisations reduce financial losses.

In analysing qualitative benchmarks, the most relevant is customer satisfaction, which aligns with the top two use cases for the finance industry. Following the same reasoning, in increasing customer satisfaction, organisations increase customer stickiness (reduce churn). In addition, highly satisfied customers are more willing to provide and share additional data (partly due to loyalty and length of custom), which can be used to develop target strategies for upsales and cross-sales.

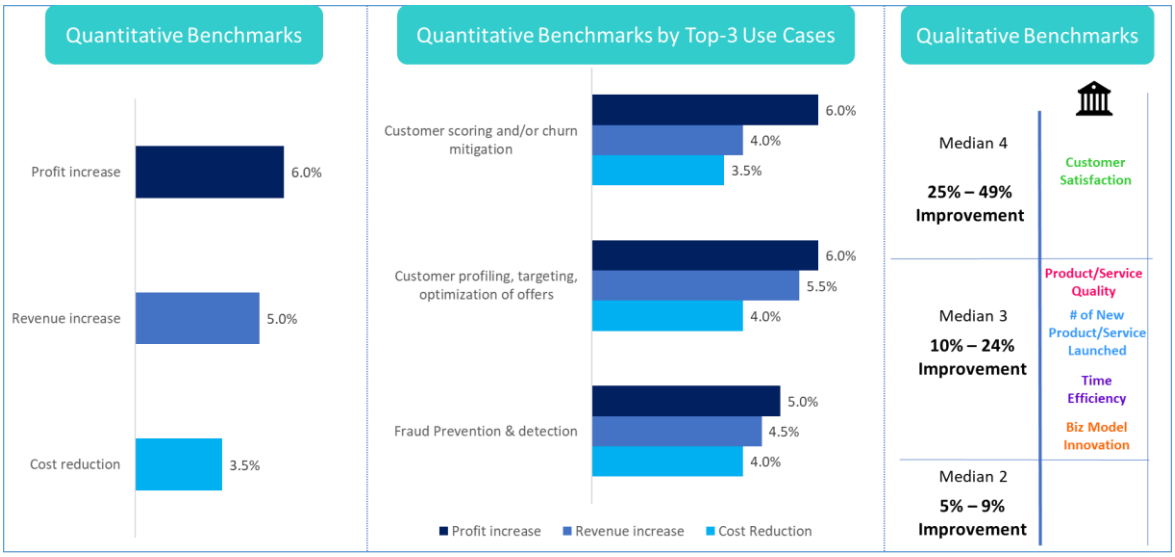

Figure 25, BDT Benchmarks: Financial Services
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.3 Business & IT Services

The business & IT services sector is the traditional leader in Big Data adoption and exploitation and in using Big Data solutions to achieve large profit increases. This is evident when we analyse the top-rated use cases. For instance, risk exposure assessment (predicting the potential future loss of business) and customer profiling & targeting and offer optimisation have the potential to increase profits (by up to 7% and 6% per year, respectively). While the customer profiling and new product development use cases relate

strictly to profit and margin increases, the link between profit increase and risk exposure assessment is clear. Risk exposure assessment is an activity that, per se, relates 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, an organisation can decide whether to follow up on it or simply to drop it. In so doing, the organisation is able to assess and choose only (or mostly) more profitable activities, disregarding those with high levels of risk.

Analysing the qualitative benchmarks, it is easy to understand why customer satisfaction and product/service quality are the KPIs that benefit the most from BDA adoption. Customer satisfaction relates strictly both to profit increase and to customer profiling & targeting and offer optimisation, while product/service quality delivers greater customer satisfaction and links with new product development.

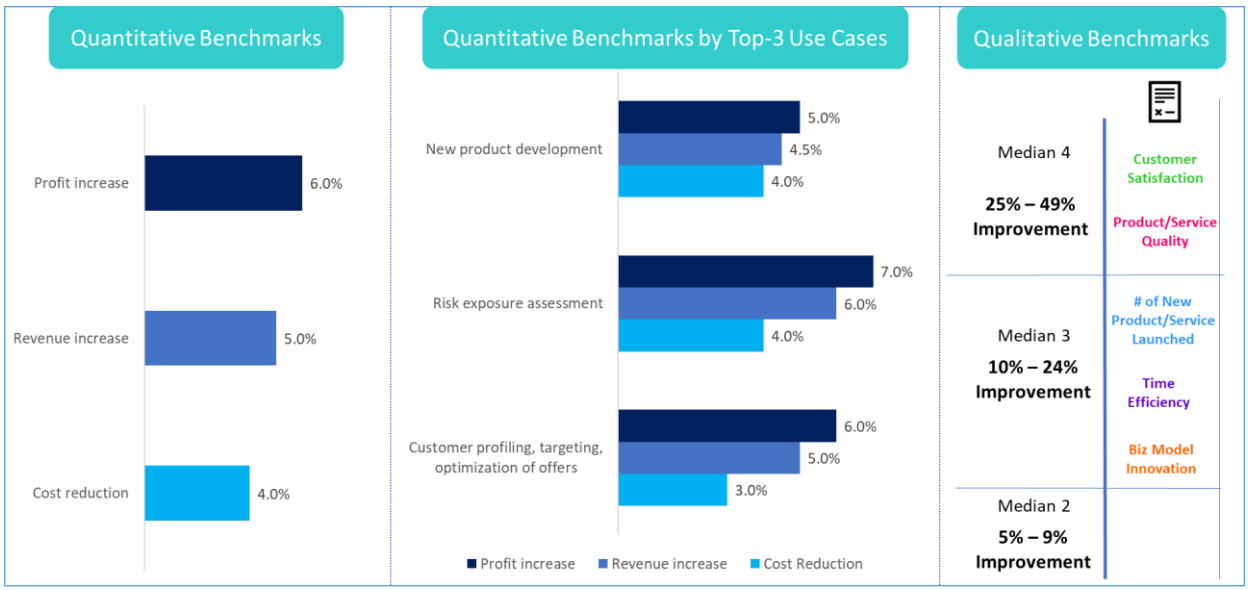

Figure 26, BDT Benchmarks: Business & IT Services
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.4 Healthcare

The healthcare sector is characterised by a long-standing data abundance. Up until recently, however, this invaluable asset has not been used due to the data's high sensitivity – aside from the privacy issues related to analysing it due to potential reidentification of patients. Despite limited data usage, the sector sees a wide range of beneficial applications for BDA adoption – related strictly to patients (quality of care optimisation, illness/disease diagnosis and progression, etc.) and more generally to regulatory intelligence (acting upon the understanding of regulations to legally use data) and to fraud prevention and detection. Within quality of care optimisation, there are several different sub use cases, from availability of hospital beds and the management of treatment slots to resource allocation (people and equipment).

In this context, the increased availability of data and computational resources and the current development of privacy preserving technologies can potentially vastly improve resource usage and optimisation. As healthcare services and structures are often government owned, cost reduction was the only relevant KPI in the past, with little interest

in increasing profits and margins. But, with shrinking budgets, the creation of revenue streams is also essential for the healthcare sector. Among the most relevant use cases, and bridging use cases and KPIs, we find quality of care optimisation. This use case is largely evaluated by profit increase because quality of care relates to patients' satisfaction (i.e. customer satisfaction), the optimisation of services, and avoiding the waste of important resources. In addition, an ability to predict equipment faults/the need for maintenance helps improve general infrastructure management. Analysing qualitative KPIs reveals that all KPIs are relevant in assessing BDT usage.

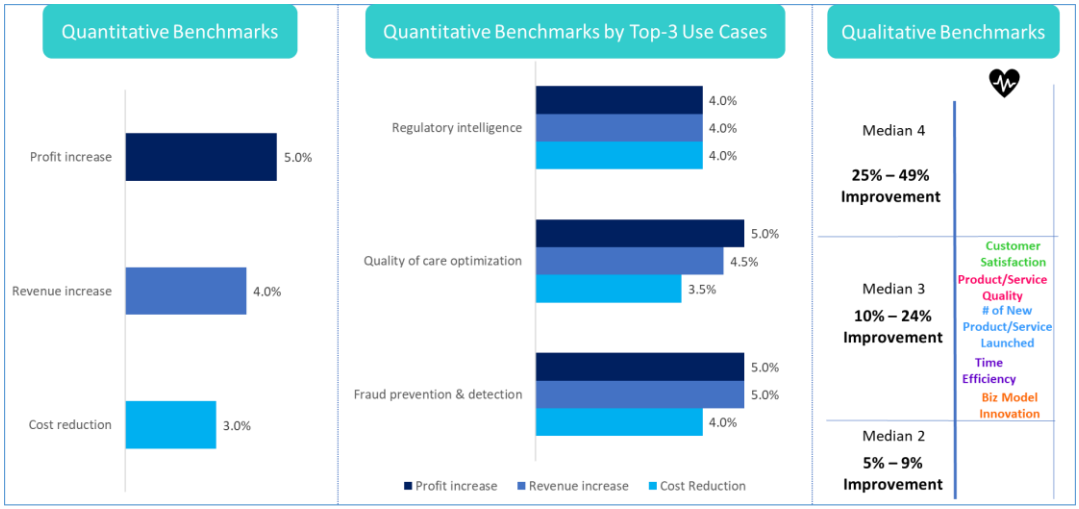

Figure 27, BDT Benchmarks: Healthcare
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.5 Manufacturing

Manufacturing is traditionally data-abundant sector, with batch data from traditional IT systems (inventory, production, sales, etc.) and new data streams, such as those produced by the Industrial Internet of Things (IIoT) – an area of increasing importance.

The manufacturing sector, despite its diversity, has the same common needs – supply chain optimisation (a topic never more relevant than today due to the pandemic-induced collapse of supply chains worldwide), predictive maintenance, and product development. The first two use cases relate strongly to cost reduction, but they also deliver increased profits. In optimising 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 full-speed production.

In avoiding supply chain disruptions, organisations can maintain production levels, profitability, and high margins. Predictive maintenance (determining exactly when machinery needs to be stopped for necessary maintenance work) – and, to extend this concept, predictive failure – helps organisations to accurately allocate maintenance hours to avoid impacting production schedules and lines. The ability to maintain full-speed production helps organisations deliver higher margins and profits. The third most important use case is new product development, which relates to profit increase. An organisation's use of Big Data to understand customers' needs and desires and to translate them into new products or product upgrades increases its ability to achieve greater profits.

In summary, no specific qualitative benchmark is considered more relevant than any other. Nevertheless, as an outlier, business model innovation is a less important benchmark in BDA activities. This is because Big Data is already exploited for other, more relevant, activities, and some business model innovation activities (e.g. data monetisation) are still currently under evaluation by manufacturers.

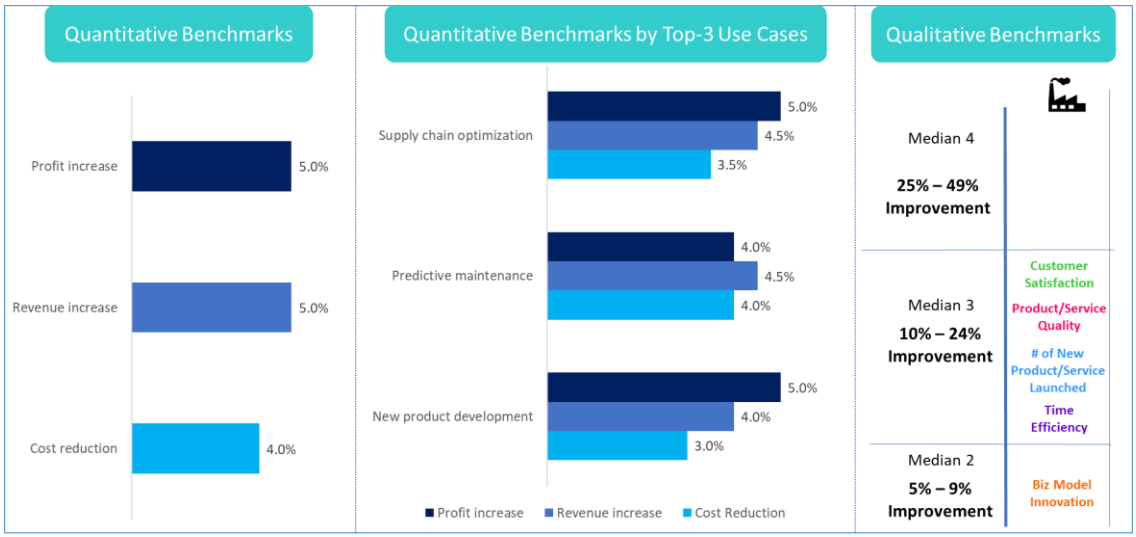

Figure 28, BDT Benchmarks: Manufacturing
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.6 Retail & Wholesale

Retail & wholesale is another sector with abundant data. It was one of the first sectors to see the adoption of BDA solutions. On a general level, profit increase is the preferred benchmark, both in general terms and when analysing specific use cases. Regarding the top three use cases, they are all preferably evaluated with profit increase as the main benchmark. The optimisation of the supply chain (and logistics as a part of it) can lead to profit and margin increases. In creating a more transparent and interconnected view of the supply chain, retailers and wholesalers can potentially avoid disruptions. In applying BDTs to supply chain data, it is easier to forecast and predict potential issues and act upon them in a timely manner. Price optimisation is mainly regarded in terms of profit increase. An ability to properly segment customers and target them with different prices helps organisations increase margins and thus increase profits.

The other top use case, new product development, also has a connection with the preferred qualitative benchmark, product/service quality. Retailers and wholesalers are interested in creating new products to improve product quality and related services and thus grow profits. Raising the quality of existing products and services (delivery, customer care, insurance, etc.) increases customer retention. With increased retention, an organisation can collect further information and data on customers and offer new products (cross-selling and up-selling) to boost revenues. The other qualitative KPIs are all regarded as moderately important in some way, offering a good range of improvements, but they are still not as highly regarded as product/service quality.

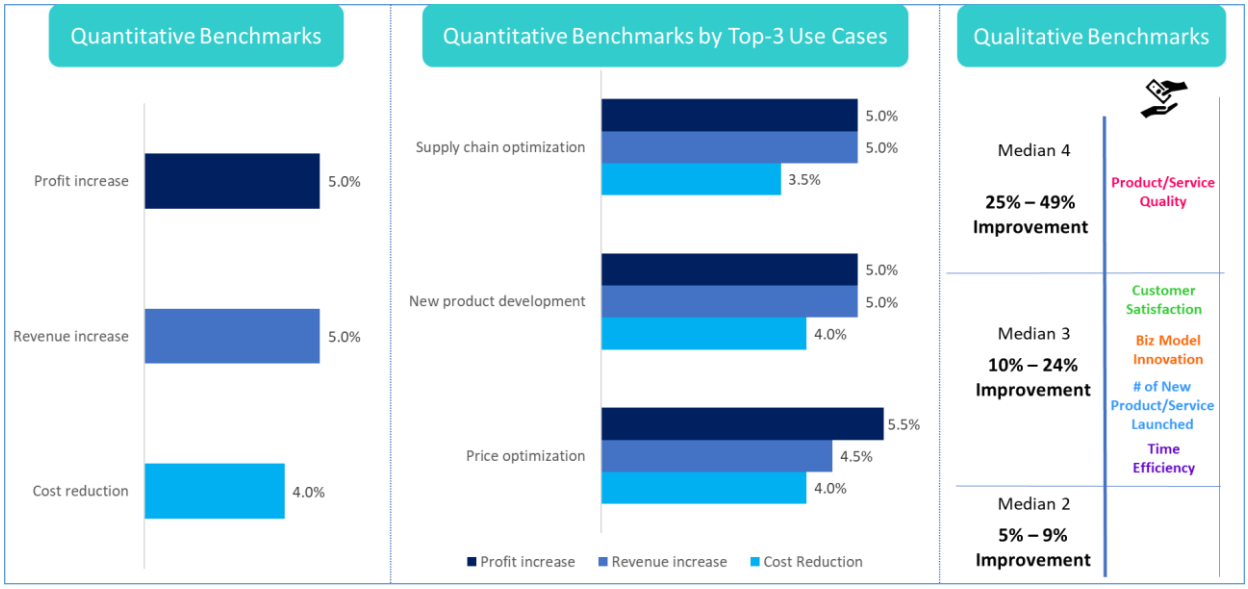

Figure 29, BDT Benchmarks: Retail & Wholesale
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.7 Telecom & Media

The telecommunications & media sector is facing dramatic challenges in consumer choices. Now more than ever before, customers have little loyalty to services providers and are able and willing to switch provider on the fly. The quality of service and customer satisfaction in this challenge play pivotal roles, which 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 in terms of providing 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 telecommunications providers and media companies, we find automated customer service and customer profiling & targeting and offer optimisation. Both use cases are benchmarked using the profit increase (6% for both) and revenue increase quantitative benchmarks. The top use case, a product and service recommendation system, is measured in terms of profit increase and is linked with customer satisfaction. With this use case – semi-automated systems recommending specific products/services to customers and users based on their profiles and needs – it is possible to leverage customer satisfaction and customer stickiness to boost profits. Concluding on telecommunications and media, the most impactful KPI is profit increase.

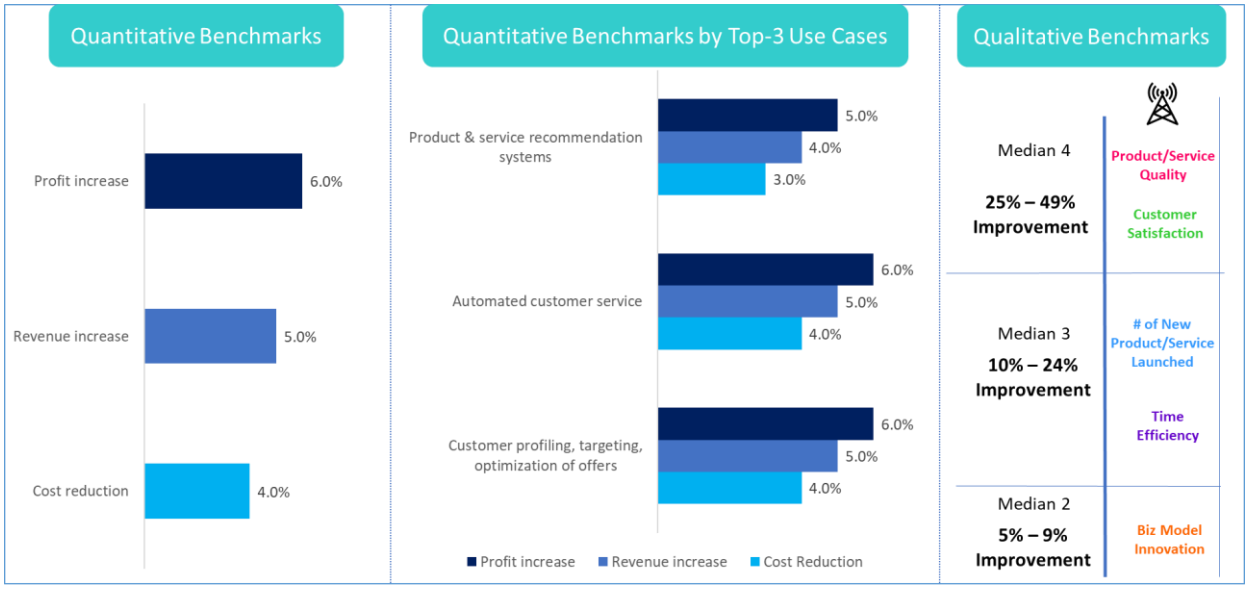

Figure 30, BDT Benchmarks: Telecom & Media
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.8 Transport & Logistics

The transport & logistic sector is already highly reliant on the use of streamed data to track parcels and trucks and provide high service quality. And the use of BDA to provide added value is increasingly playing a pivotal role beyond just an additional option, albeit a significant one. The real-time tracking of deliveries from suppliers to clients (B2B) or consumers (B2C) has been either an operational management resource and a customer benefit for a long time, and now it is a core service. BDT enables this information flow to be analysed much more effectively and in real time for the optimisation of delivery and communication with 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 optimisation and inventory and service parts optimisation. All three use cases are highly rated, as they are able to deliver both large profit increases (especially price optimisation, at 6%) and relevant margin increases. Cost reduction is less relevant on all fronts.

When considering qualitative KPIs, we would expect customer satisfaction, product/service quality, and time efficiency to be the most relevant in providing marked improvements, but survey data tells a different story. All the qualitative KPIs are regarded as providing medium-size improvements (in a range of 10–24%). This clearly indicates that organisations in this industry are highly sophisticated in their ICT usage. The use of BDA is no longer seen as relevant to ensure the quality of the services offered, but BDA can and will provide considerable added value to all activities in the industry.

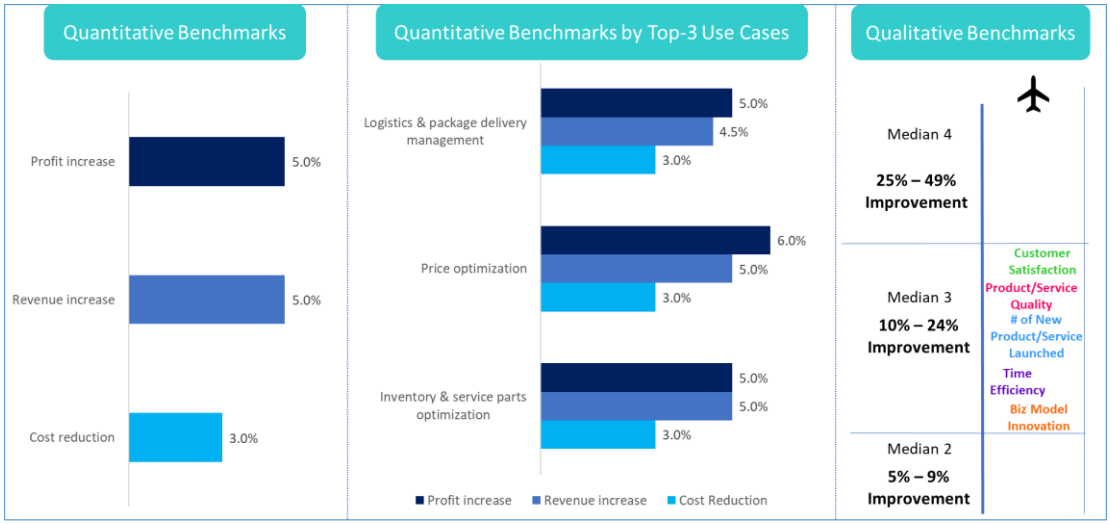

Figure 31, BDT Benchmarks: Transport/Logistics
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.2.9 Utilities and Oil & Gas

The petrochemical sector is a pioneer in using advanced analytics for resource exploration. With advances in BDT and computing 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 involving digital and core technologies.

The main priorities in terms of use cases are risk exposure assessment (of new activities and services), the predictive maintenance of machineries and oil/gas pipelines, 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 with predictive maintenance. The quality of service in the utility sector is evaluated by the non-discontinuation of service and avoiding service outages, while the predictive maintenance of machinery and distribution networks is essential continued service (avoiding outages). The number of new products and services launched relates directly to optimal and well-performing risk exposure assessment analysis, which helps organisations to understand whether a new product or service will be a viable solution or something to be eliminated from the innovation channel. Another extremely important qualitative KPI is customer satisfaction, as the use of BDT to manage customer relationships is a primary activity in such organisations.

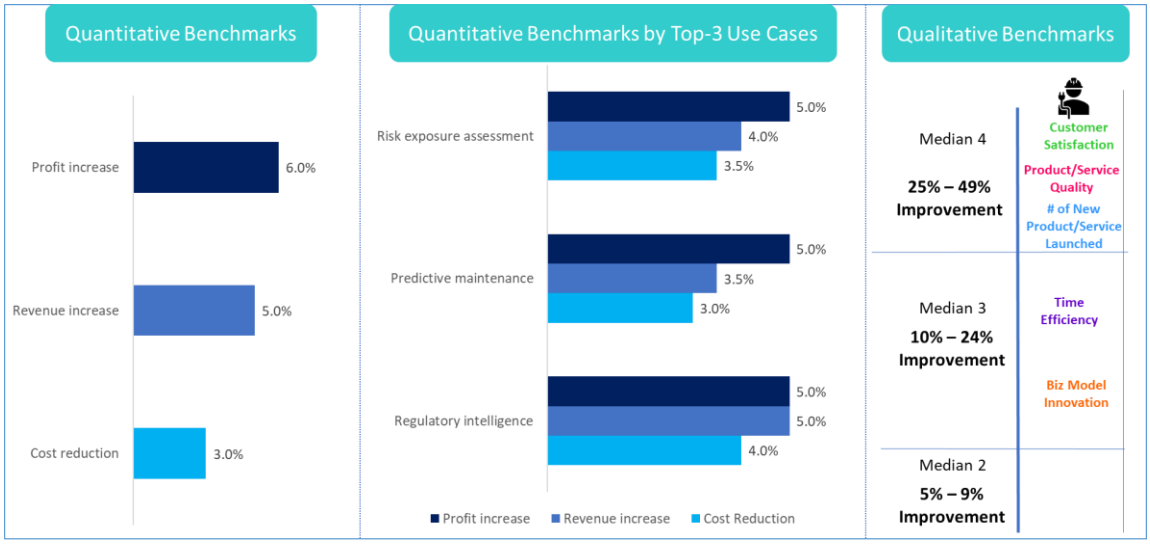

Figure 32, BDT Benchmarks: Utilities and Oil & Gas
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.3 Business Benchmarks by Company Size

This chapter describes the BDT business benchmarks by industry, which are presented in the Toolbox's Knowledge Nugget section through the user interface. The comments below provide some background on 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*. [14]

5.3.1 Small and Medium-Size Enterprises (SMEs)

SMEs (50–249 employees) evaluate profit increase with the highest impact, followed by revenue increase, and cost reduction. Despite demonstrating the greater impacts in terms of profit increase, when analysing the potential benchmarks against the top three use cases, we observe that profit increase is never the best benchmark for SMEs. Contrary to what we saw previously, in the industry analysis, and what we will see in the upcoming size segments, risk exposure assessment and price optimisation are equally benchmarked using cost reduction and revenue increase, while regulatory intelligence is mostly benchmarked using revenue increase.

Responses from SMEs are sketchy, with low percentages, offering no clear perspective. Moreover, the qualitative benchmarks that offer the biggest improvements (still only medium-large) are not clearly linked with use cases. In addition, a discordance is evident between business goals and achievements resulting from BDA solutions.

It is worth noting that, although we are comparing companies within the same segment size here, these companies are from very diverse sectors, which undermines the results. This does not mean, however, that SMEs cannot benefit as much from adopting the right use cases and deploying BDA solutions as large enterprises.

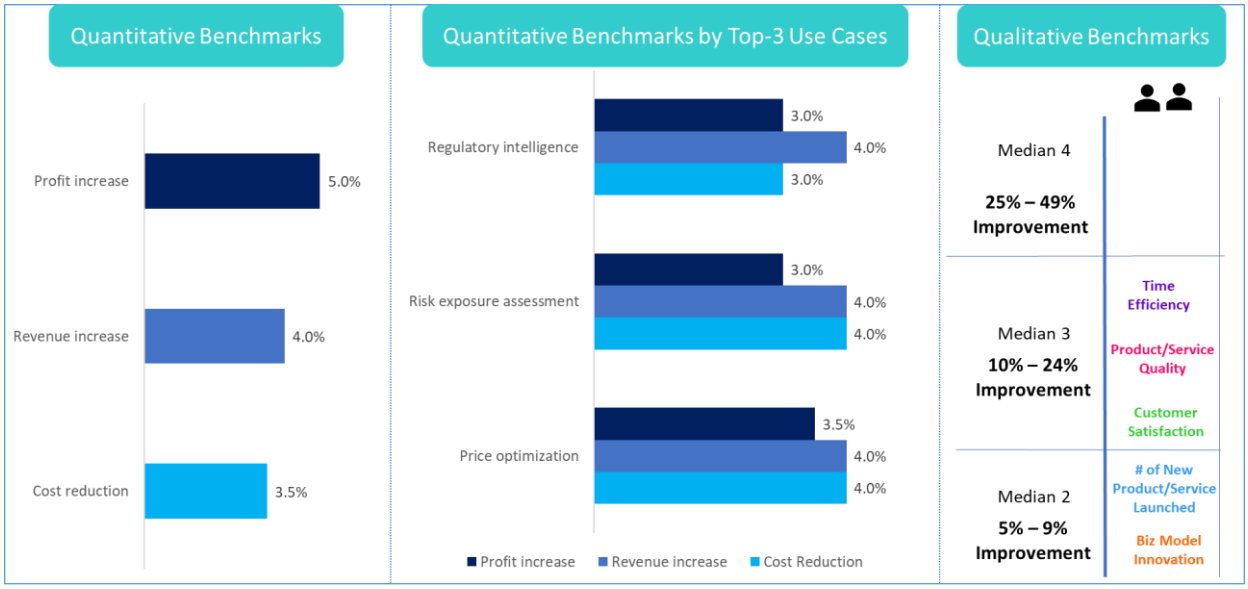

Figure 33, BDT Benchmarks: SMEs
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.3.2 Medium-Large Enterprises

Medium-large enterprises (250–499 employees) demonstrate greater BDT impacts on profit and margin increases than on cost reduction. This is because organisations within this segment-size already had in place the right architectures 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, with revenue growth, while regulatory intelligence is mostly benchmarked with revenue growth. From a qualitative KPI perspective, analysis by company size regarding improvements offers no specific benchmarks when, with all the presented benchmarks rated in a medium range (10–24%).

Overall, medium-size companies adopt BDT to grow. These organisations can use BDA to analyse their existing customer bases and pricing to identify upselling opportunities. Critically, they can also analyse the market to identify new customers. Diversification offers another growth opportunity for these organisations, whereby BDA helps them understand market needs and invest effectively to develop new products and services and/or to modify existing ones.

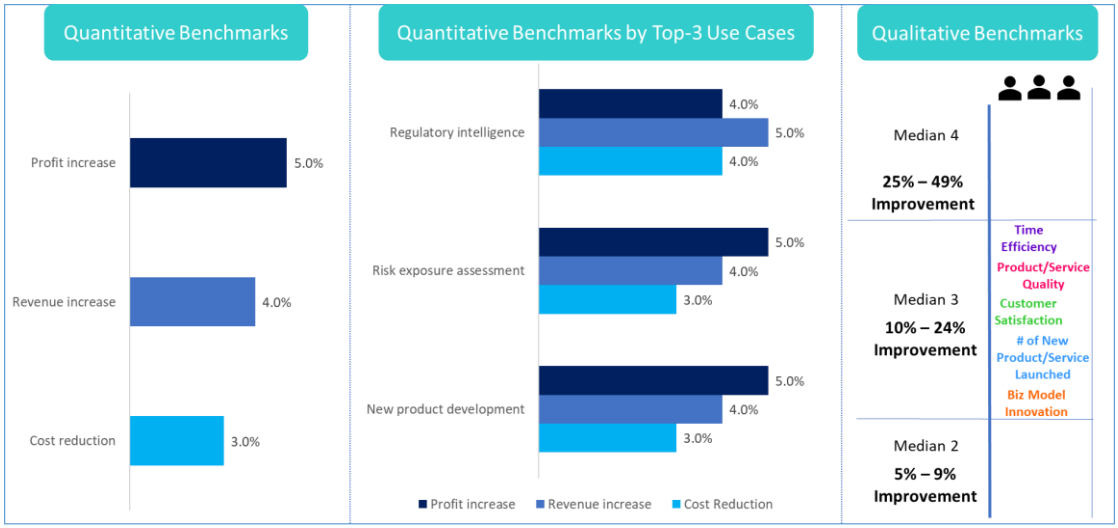

Figure 34, BDT Benchmarks: Medium-Large Enterprises
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.3.3 Large Enterprises

Large enterprises (500–999 employees) are often better positioned than smaller ones to consider investing in new products and services, whether via internal R&D, mergers & acquisition, or partnerships with OEMs and resellers. The ability the BDA offers to analyse existing internal product sales and the financial and market prospects of potential M&A partners makes BDA a valuable decision support tool. Detailed analysis of market prospects can be expected from external partners or investors. Quantitative KPI benchmarks overlap precisely with use case evaluations. Risk exposure assessment, price optimisation, and new products development are equally benchmarked by profit and revenue increases, with cost reduction playing a smaller role. Considering the opening statement, we can clearly see how these top three use cases perfectly fit the segment description.

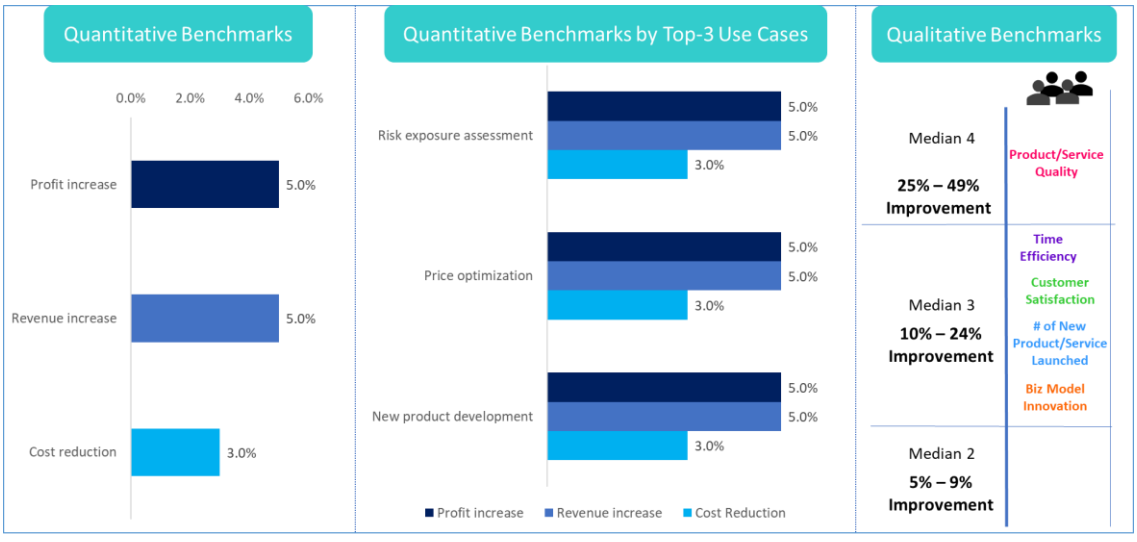

Figure 35, BDT Benchmarks: Large Enterprises
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.3.4 Very Large Enterprises

Very large enterprises (1,000+ employees) have the resources to invest in BDA, and BDT is now able to integrate information that was previously in silos to enable consolidated views of both the company's own processes, which can be optimised, and the current or potential market for products and services. Large companies have access to far more internal customer data, but they also gather external data, which they integrate, leading to significant opportunities to improve customer relations and to upsell and cross-sell using carefully targeted offerings.

Very large organisations rely heavily on profit as the main benchmark, rather than revenue increases and cost reduction. A similar situation is evident when evaluating the top-three use cases. Customer profiling & targeting and offer optimisation and regulatory intelligence have the same values in terms of quantitative benchmarks. New product development differs slightly from the other two, with equal benchmarking values for profit and revenue increases. Qualitative benchmarks are in line with the top-three use cases presented. In support of the values and results presented for the first and third use cases, product/service quality and customer satisfaction are extremely relevant, recording big improvements (25–49%).

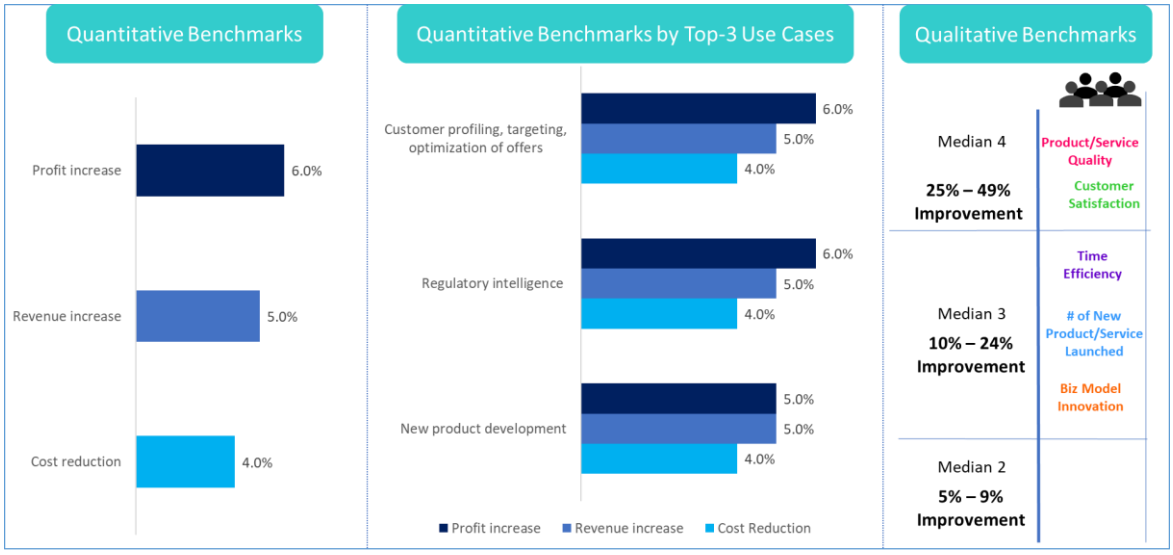

Figure 36, BDT Benchmarks: Very-Large Enterprises
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

5.4 Star Performers

Star performers are organisations with the best achievements in terms of business impacts from the use of BDT. Out of the sample of 730 enterprises interviewed, we found 35 organisations falling in this category (roughly 5% of the total sample). Analysing star performers helps other organisation to identify the upper boundaries of potential achievements – what they can aim for if they maximise the effectiveness and efficiency of their BDT deployments.

The majority of the star performers group, unsurprisingly, are very large and large enterprises. Big data technologies are sophisticated, and their adoption requires both

significant investment and good internal information-system infrastructure to ensure datasets are 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 the leading industries regarding the use of data – namely, business & IT services (22%), retail (19%), and financial services (14%). Generally speaking, these enterprises were early adopters of BDT, hold more ambitious plans and expectations than the rest of the sample, and tend to be more eager in taking risks on innovation investments.

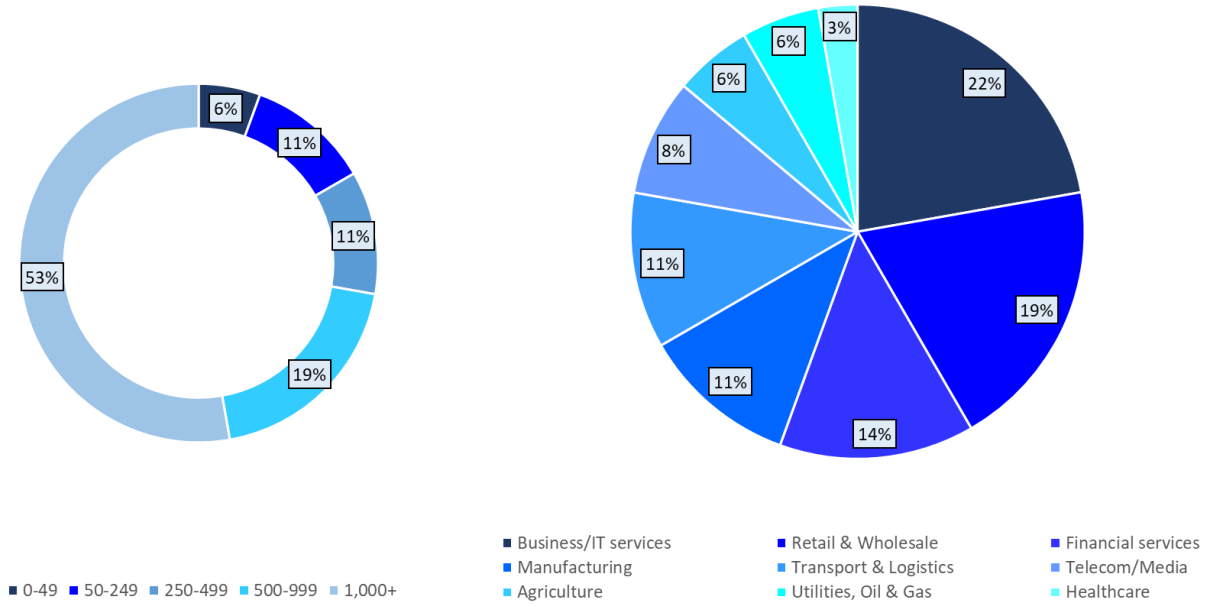

Figure 37, Star Performers Group Composition
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

BDT business impacts for star performers are markedly higher than those for the rest of the business users' sample, as shown in Figure 38, below. In terms of profit and revenue increases, star performers achieve a mean value of an 8% increase, compared with 5% for profit and 4% for revenue increases for the rest of the sample. Star performers regard cost reduction as less important than the other two quantitative benchmarks; however, the best performers achieve slightly greater cost reduction than the market average. Concerning qualitative KPIs, the share of star performers with the greatest improvements from BDT (over 25%) is three times higher than is the case for the rest of the sample regarding time efficiency, product/service quality, and customer satisfaction. The results for the launch of new products and services are better, but close to those for the rest of the sample. Strangely, the majority of star performers declare low or medium improvements in the case of the adoption of innovative business models. Perhaps this situation results from company size: Large corporations are notoriously reluctant to introduce disruptive new business models.

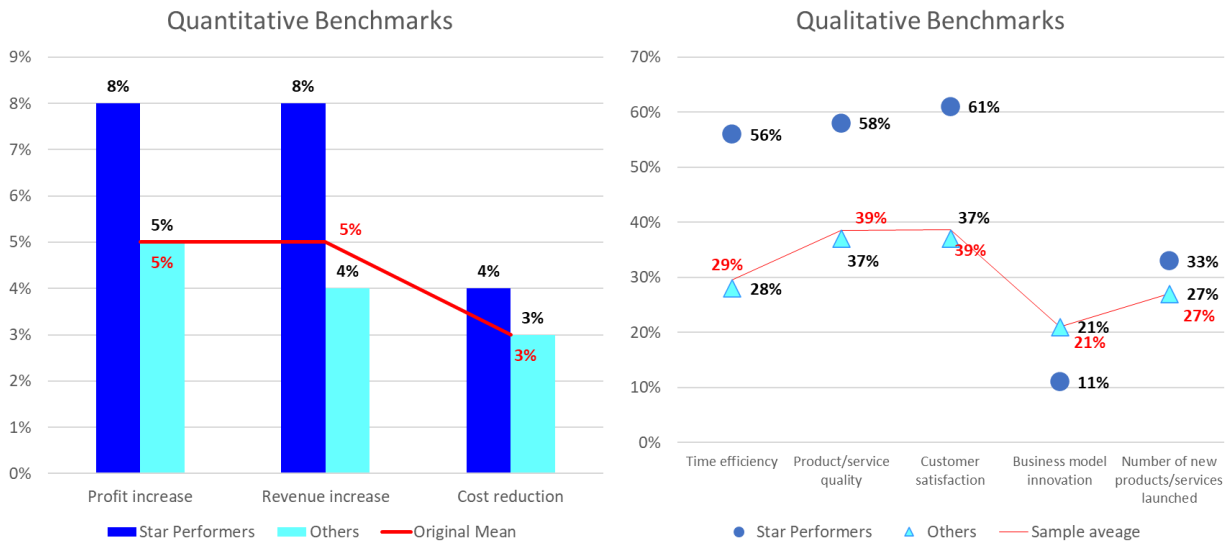

Figure 38, BDT Benchmarks: Star Performers
Qualitative Benchmarks: Share of Respondents with High Improvements (>25%)
Source: D2.4, *Benchmarks of European and Industrial Significance* [14]

Finally, star performers are also ahead of the market in terms of the adoption of advanced and sophisticated BDT use cases, as shown in Figure 39, below.

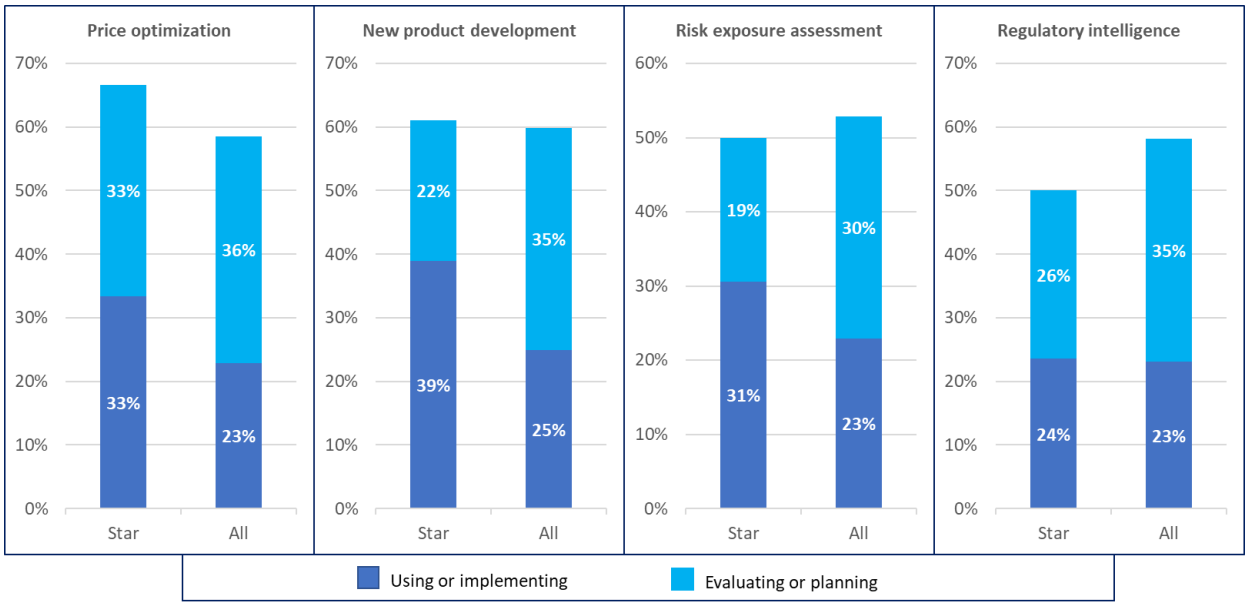

Figure 39, Star Performers' Top Use Cases
Source: DataBench Survey, June 2020 – D.2.4 [14]

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 items, 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 an entry point 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: This enables 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: The catalogue provides a list and metadata about the most relevant technical Big Data benchmarking tools. The initial list is based on the deliverables provided in the scope of DataBench WP1 – in particular, D1.2[8], D1.3[11], and D1.4[12]. The catalogue is extensible in that new benchmarks can be added, following the editorial procedure explained later in this section.
- Knowledge Nugget (KN) Catalogue: The aim of this knowledge base is to offer to the community a better understanding of the business value of Big Data benchmarking. As already mentioned, KNs 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 to add new KNs – also following an editorial procedure.
- Searching features: These enable searches throughout the catalogues of our knowledge base and the display of the information found.
- Tag-based navigation between resources: The front-end of the Toolbox enables navigation through the various resources (benchmarks and KNs), using annotations and tags. By clicking on a specific tag, the set of resources annotated with the same tag will be prompted and accessed, giving the possibility to browse resources at will.
- User journeys: As will be explained later in this section, the Toolbox provides tips and advice for different user types on how to use and navigate the entire Toolbox. These are encompassed in so-called 'user journeys' (as envisaged) and are intended to help users navigate, which can be achieved in multiple ways.
- Usage statistics: Statistics on the usage of the tool (visibility only to administrators, although select statistics are open to everyone).

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

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

A screenshot of the homepage of the Toolbox is shown in Figure 40.

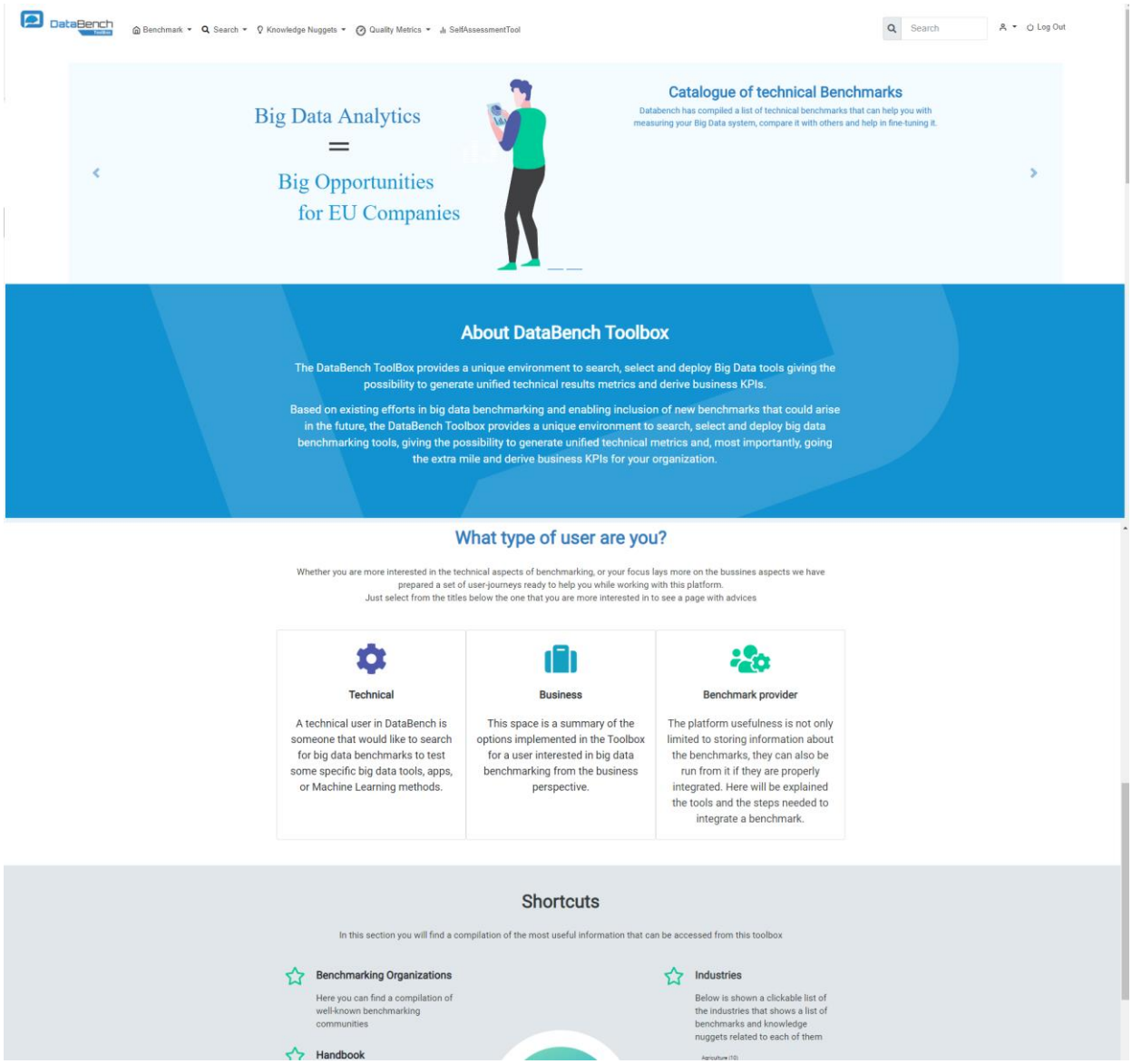

Figure 40, Homepage of the Toolbox
Source: DataBench Toolbox

6.2 Architecture Building Blocks of the DataBench Toolbox

The main building blocks of the DataBench Toolbox from a high-level technical perspective are depicted in Figure 41. This functional architecture comprises the main building blocks of the Toolbox – namely, the DataBench Toolbox Web user interface, the Toolbox Catalogues, and the Toolbox Benchmarking Automation Framework. This last block is a bridge to the Execution of Benchmarks building block, located outside of the Toolbox. The intended users of the Toolbox are explained in more detail in section 6.3.

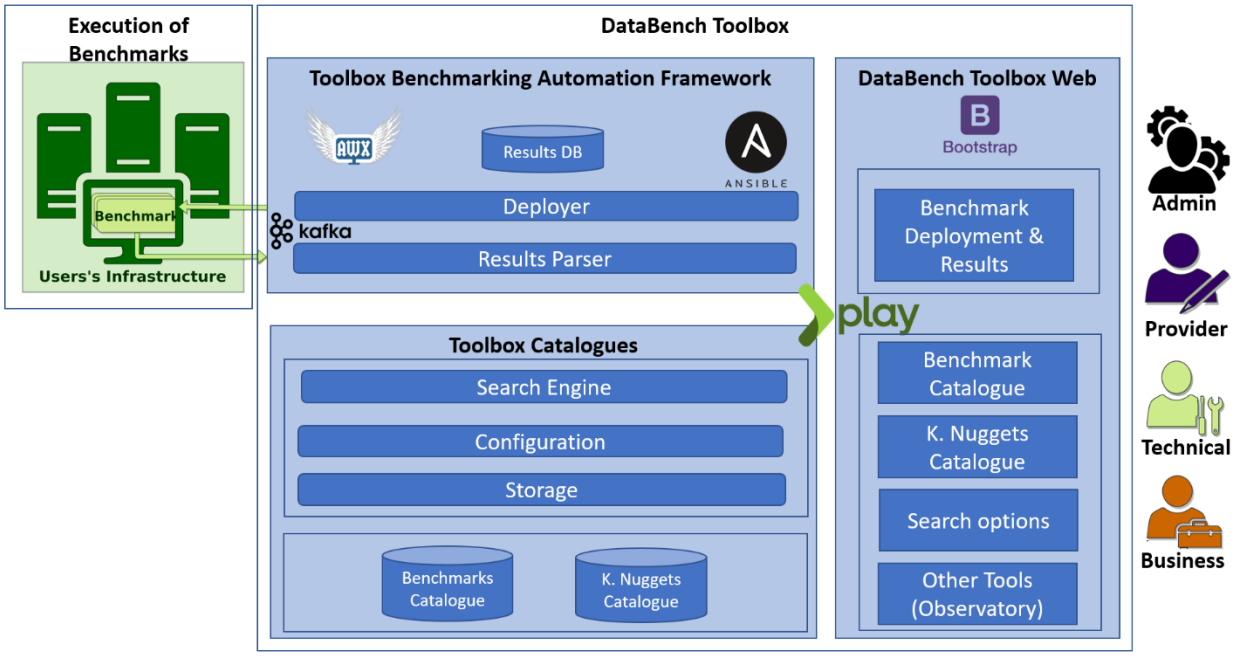

Figure 41, DataBench Toolbox Functional architecture

The main building blocks depicted in Figure 41 are as follows:

- **The DataBench Toolbox Web building block** is the main entry point for users. Figure 40 shows the homepage and is explained in more detail in section 686.5.
- **The Toolbox Catalogues building block** provides the backend functionality associated with the management, search, and browsing of knowledge nuggets and benchmarking tools, as well as the repositories in which this information is stored.
- **The Toolbox Benchmarking Automation Framework building block** is the automation bridge to enable the configuration and deployment of technical benchmarks fully integrated (not only listed in the Catalogue) with the Toolbox. This building block enables the execution of some selected benchmarks and offers the possibility to the users to provide the results back to the system by means of a Kafka service, or uploading the results file. It enables deployment to the Execution of Benchmarks building block.
- **The Execution of Benchmarks building block**. The Toolbox does not provide any infrastructure or playground in which to deploy and execute benchmarks. This building block is therefore located in the infrastructure provided by the user, outside of the Toolbox, and can be either in-house or cloud infrastructure. The user must define in advance the HOSTS and credentials of where the deployment will take place within a selected benchmark description. Currently, this feature is enabled only for the first benchmarks listed in the Benchmark Catalogue, as depicted in Figure 42 (using the icon on the right) – namely, BigBench V2, HiBench, OWPERF, YSB, and YCSB – but can be extended in the future to include other compatible benchmarks. More information about configuration and usage can be found in section 6.5.3. How to add new integrated benchmarks is explained in more detail in section 6.6.2.

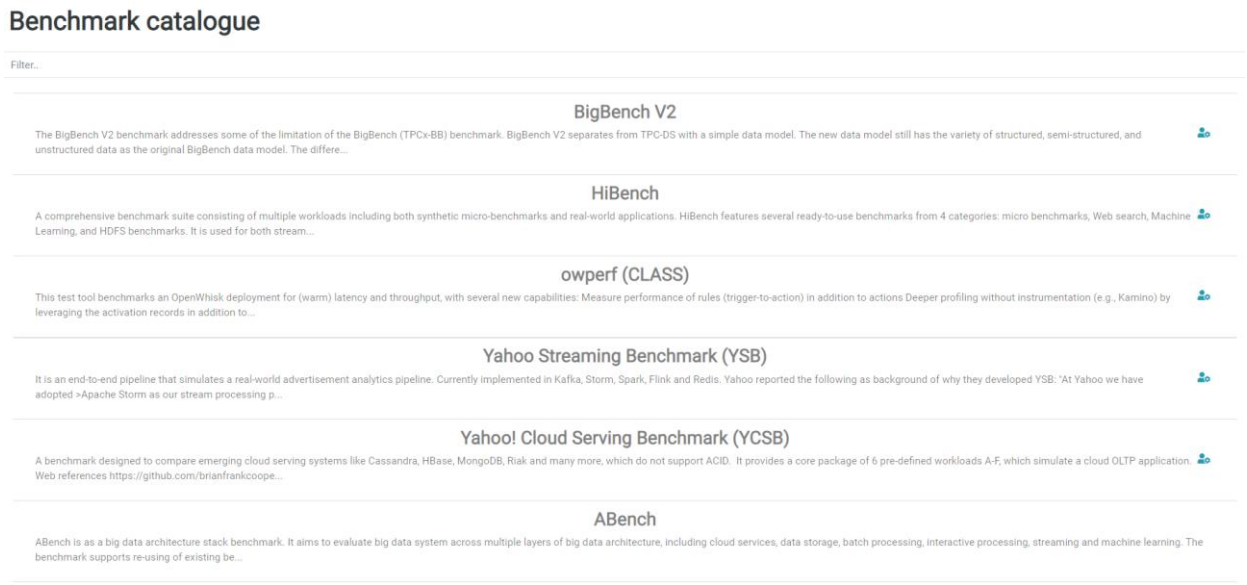

Figure 42, List of Integrated Benchmarks
Source: DataBench Toolbox

6.3 Intended Users of the Toolbox

As reported in deliverable D3.4 [1], 'the DataBench Toolbox sits at the core of DataBench results, providing access for different user types to the resources made available by the project and by external initiatives.' A summary of the users of the Toolbox follows:

- **DataBench administrators**: As with any tool supported by a knowledge base, the DataBench Toolbox needs some housekeeping. Toolbox admins oversee the Toolbox, ensuring that it operates well. They are responsible for user management and can add to and maintain the contents of the knowledge base. They oversee the approval or rejection of proposed content (benchmarks and knowledge nuggets) from the community or from DataBench experts. Administrators can also visualise the Toolbox usage statistics.
- **Technical users**: Technical users encompass all potential stakeholders interested in benchmarking Big Data solutions and the different skills and needs involved. The Toolbox suggests different paths, depending on the user's needs, via a standard set of options. Technical users include the following:
    - Casual users, who are interested in searching and browsing existing benchmarks and info about them
    - Benchmarking experts, who may belong to specific benchmarking organisations and are interested in finding new benchmarks and eventually deploying and executing them to evaluate the technical indicators of Big Data solutions
    - IT experts from industrial organisations, who might not necessarily be experts on benchmarking, and who may need some help to understand how to benchmark a Big Data solution, evaluate alternatives, and/or visualise what others have done in the past
    - Special DataBench users of interest, who are IT people of EU R&D projects on Big Data – more specifically, of projects funded under Big Data Value Public-Private Partnerships (BDV PPPs) – who would like either to check what the alternatives are or benchmark their own results to support their choices.

A technical user might belong to more than one of the above subcategories, but all have IT expertise (developers, testers, system administrators, etc.) and an interest in Big Data architecture blueprints. Their main interest is therefore in finding the right benchmarks, information, and knowledge to ease their benchmarking tasks.

- **Business users**: As in the case of technical users, business users might have different profiles and interests, but they all share common interests – namely, aspects such as performing business benchmarking, as opposed to pure technical benchmarking, and understanding the business implications of selecting a Big Data system and how this choice influences their business indicators. Business users might also fall into some of the technical user subcategories, but leaning more towards the business side of benchmarking. These include:
    - Causal users, who may be interested in searching, navigating, and browsing knowledge about benchmarking and finding similar Big Data application use cases in their industry or architecture blueprint references for their domain or use case
    - People in charge of the exploitation and sustainability of Big Data solutions in EU Big Data R&D projects, who are interested in knowledge to back their decisions about the use of Big Data in a particular sector, industry, or market

  In general, all business users are more interested in navigating the Knowledge Nuggets Catalogue than the Technical Benchmark Catalogue.

- **Benchmark providers:** Benchmark providers are those who have developed a specific Big Data benchmark and would like it to be a part of the Technical Benchmark Catalogue of the Toolbox. These users might belong to specific benchmarking organisations (e.g. the TPC[1]) or be developers of a specific Big Data benchmark (e.g. the people behind YCSB[2]). The Toolbox allows these users to add their benchmarks to the Toolbox. New additions are subject to an editorial procedure before being listed in the Toolbox. Some of these benchmarks may go a step further and enable automated deployment and testing directly from the Toolbox itself. This last step requires engaging with the Toolbox administrators.

The main benefits for the users in the other three categories (i.e. DataBench administrators aside) is summarised in Figure 43.

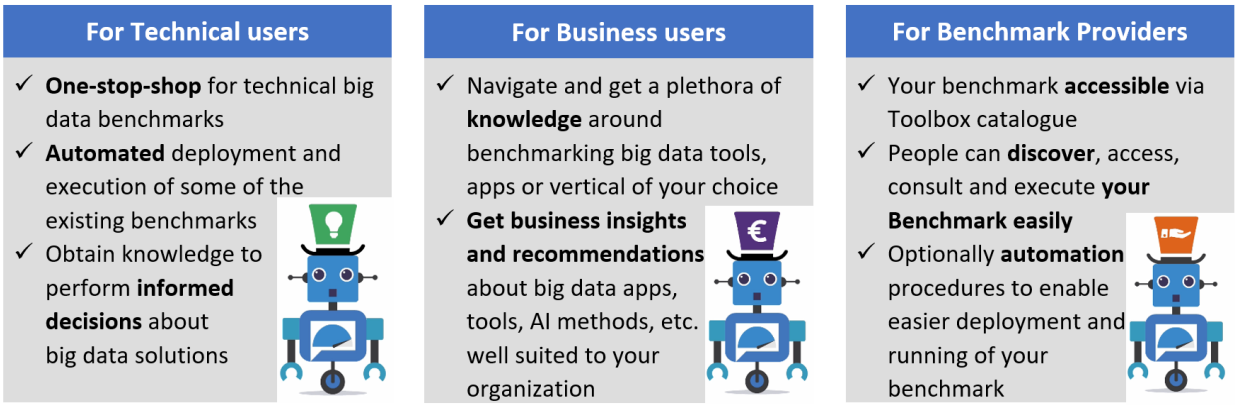

Figure 43, Summary of Main Benefits for the Users of the DataBench Toolbox

1 http://www.tpc.org/

2 https://github.com/brianfrankcooper/YCSB/wiki

6.4 Toolbox User Interface

Toolbox users can access the different tools and elements explained in the overview of the Toolbox from the main page, as shown in Figure 40. In this section, we explain the different options of the Toolbox available in the front end. The options seen on the main page are as follows:

Menu

The menu provides options to access to the Technical Benchmark Catalogue and Knowledge Nuggets Catalogue, search options, and links to other tools, such as self-assessment and quality metrics. Some of these options are only available to registered users; for example, in the submenus of the Benchmarks option, only registered users with the role of Benchmark Provider can suggest new benchmarks for inclusion in the Toolbox.

Search

As explained in D3.4[1], 'The Toolbox provides three search types: 1) a search box, in which users can type a title or resource part, tag, or word; 2) an advanced search, which enables the selection of tags to navigate to the available resources; and 3) a search interface, which uses the Big Data Value Reference Model, as explained in BDV SRIA[16], to navigate to specific technical benchmarks covering the horizontal and vertical layers of the model.' This means that, in addition to the text search enabled via the search box located in the header, the menu of the Toolbox offers two search type options – a guided search and a search via the BDV Reference Model.

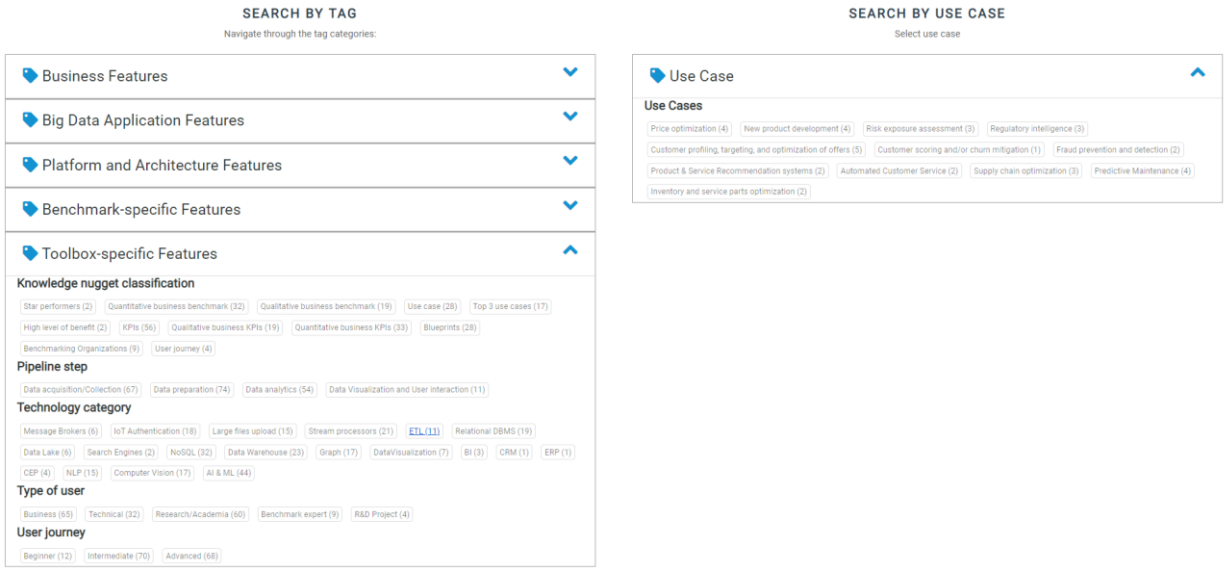

Figure 44, Guided Search
Source:DataBench Toolbox

Figure 44 shows the guided search page. This page allows the user to search by the different tags used to annotate resources (technical benchmarks and knowledge nuggets). Users can expand the different tag categories and select a tag. (For example, in Figure 44, above, the Toolbox-Specific Features category is expanded.) Once a tag has been selected, the resources annotated with that tag are listed.

Some tags relate to technical aspects, such as those under the categories of Big Data Application Features, Platform and Architecture Features, and Benchmark Specific Features; while tags under the Toolbox-Specific Features category relate more to knowledge nuggets. Under the Business Features category, the tags relate to both technical aspects and knowledge nuggets.

The next option provides access by way of a graphic depiction of the BDV Reference Model, as shown in Figure 45.

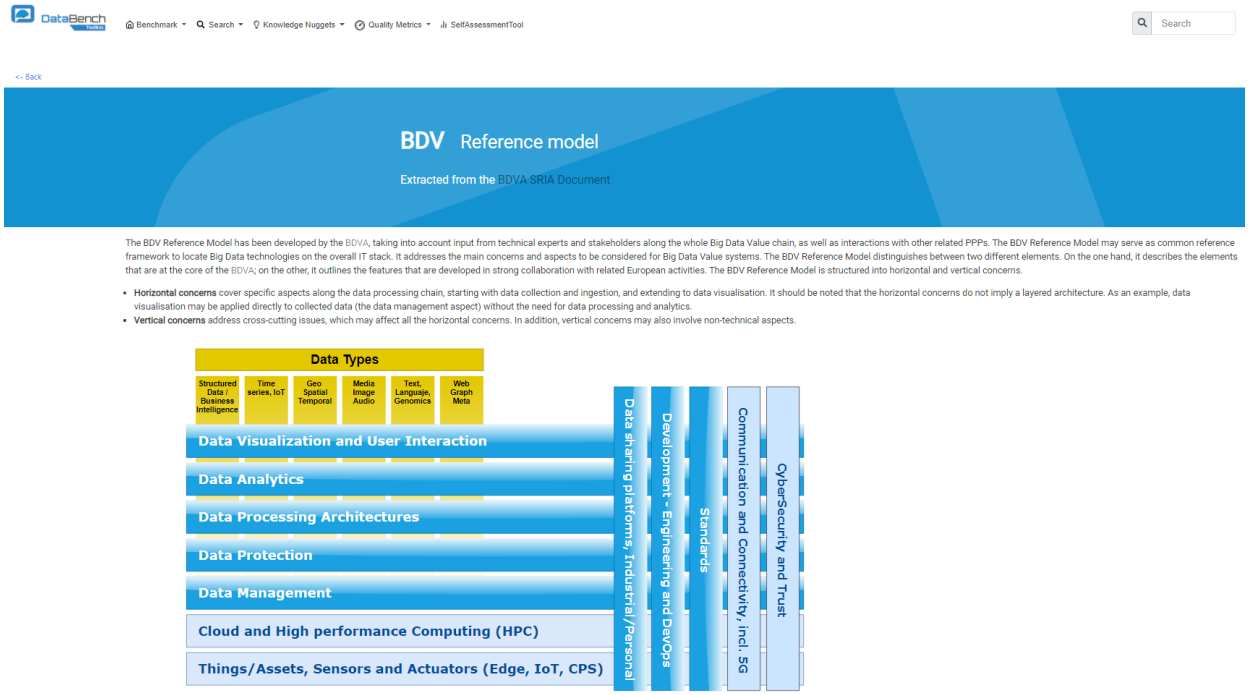

Figure 45, Search by BDV Reference Model
Source: DataBench Toolbox

When a user clicks on a layer of the model, technical benchmarks annotated to that layer appear. Note that not all benchmarks are annotated with the layers of the model, as the match between benchmarks and layers is not perfect. However, it is a good starting point to locate benchmarks that may be of help in a specific area.

A third advanced search is called Search by Blueprint/Pipeline, as shown in Figure 46. This figure depicts a pipeline with four steps and its relation to the generic blueprint defined in DataBench.

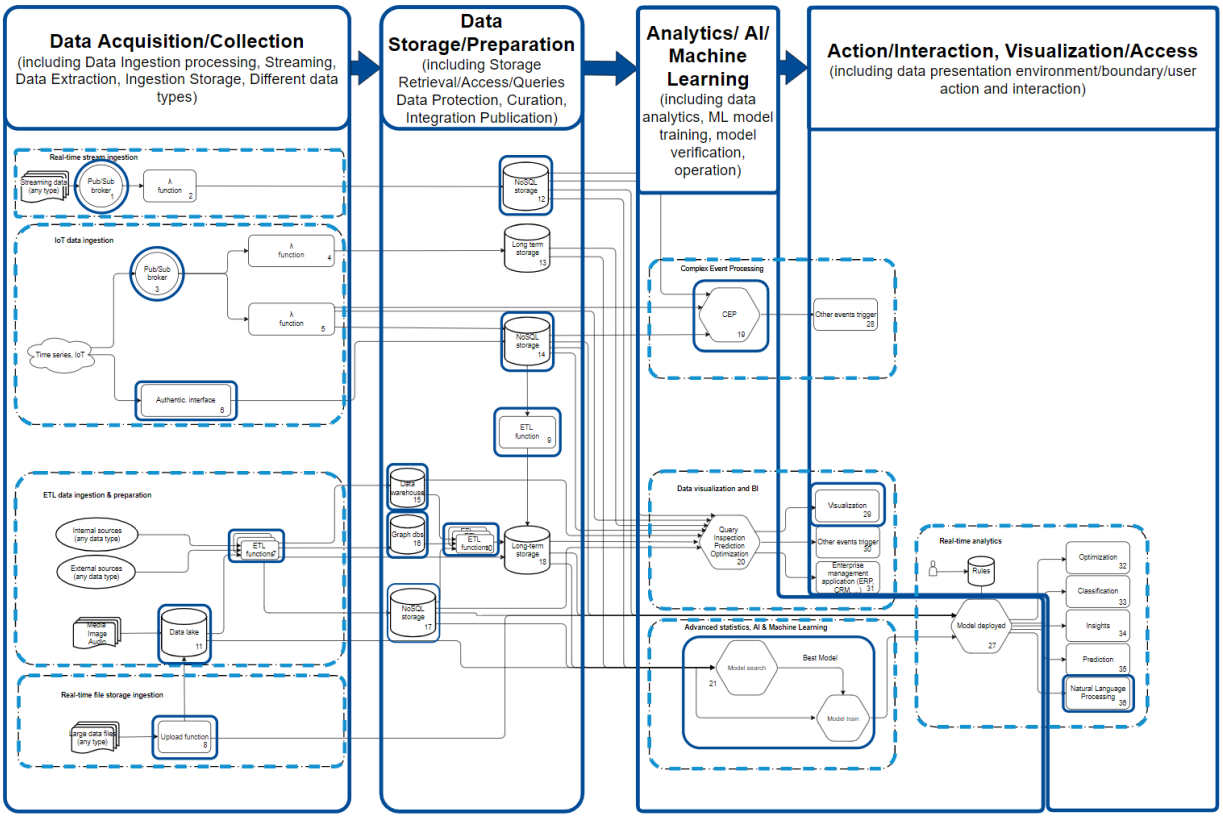

Figure 46, Search by Blueprint/Pipeline
Source: DataBench Toolbox

Users can click inside the blue boxes to search for the benchmarks annotated to a specific element of the blueprint or on one of the four steps. Note that not all benchmarks are annotated with the specific elements of the blueprint, but almost all are tagged with one or more of the four steps of the pipeline.

Finally, the search box located near the menu (Figure 47) provides a full-text search of the benchmark and knowledge-nugget descriptions.

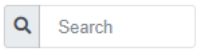

Figure 47, Full-Text Search Box
Source: DataBench Toolbox

Figure 48, below, shows an example of the results of a search (i.e. a list of resources) irrespective of the search type used.

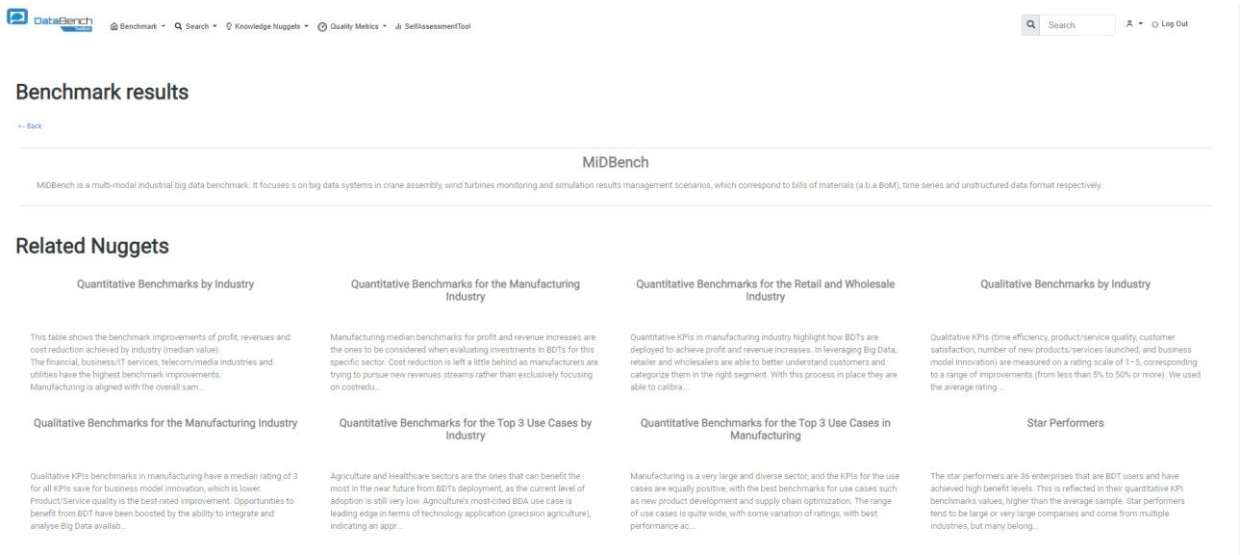

Figure 48, Search Results
Source: DataBench Toolbox

6.5 User Journeys

To help users navigate the entire the DataBench Toolbox, minimise the learning curve and entry barriers, and maximise the chances of finding and using the right benchmark solutions or knowledge nugget, the Toolbox offers tips and advice, as encompassed in 'user journeys'.

By their very nature, 'user journeys' will evolve. More tips and dedicated pages in the Toolbox will be shared as we receive feedback and learn from users' usage. The information presented in this section should therefore be considered as the starting point of the information presented to users in the Toolbox.

6.5.1 Support for Casual Users

As explained before in this section, casual users may be either technical or business users with interest in searching, navigating, and browsing knowledge or benchmarks. Their initial goal might not be to execute a specific benchmark, but rather to browse information about Big Data benchmarking to find what others are doing, such as similar use cases, industry examples, and architecture blueprints.

This type of users is what the Toolbox calls beginners. 'User journeys' for beginners can be found for business and technical users by clicking on the respective section on the homepage, shown in detail in Figure 49.

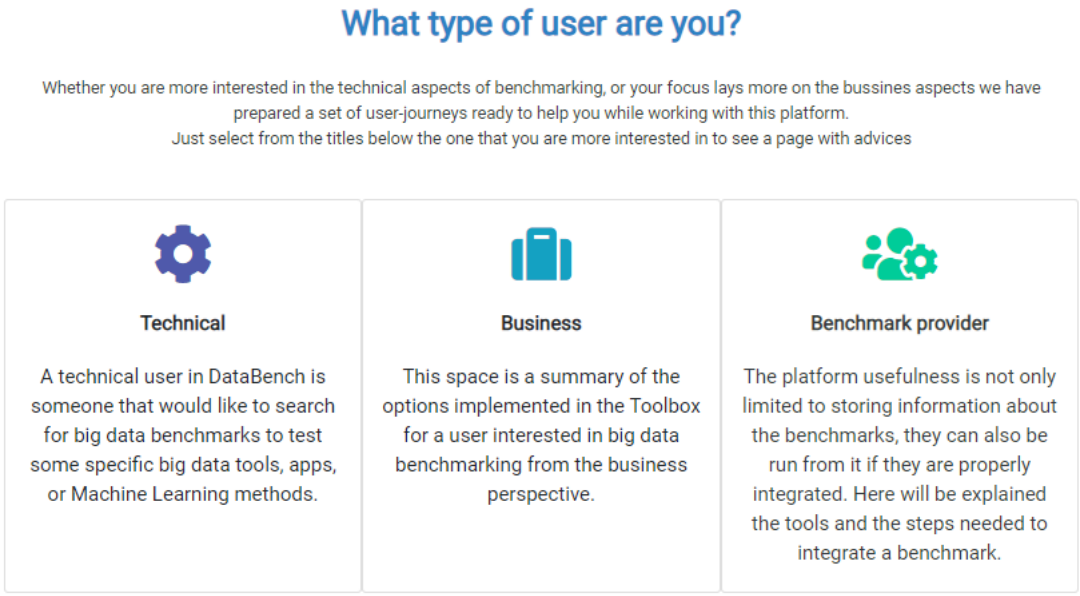

Figure 49, 'User Journey' Section from the Toolbox Homepage
Source: DataBench Toolbox

From the technical perspective, the tool provides recommendations for casual users' first steps (becoming acquainted with the Toolbox) towards finding the right information, as shown in Figure 50.

Figure 50, Technical 'User Journeys'
Source: DataBench Toolbox

First of all, it is good to understand some of the main Big Data benchmarking concepts, such as what benchmarking is and how it differs from technical validation and testing. These terms are often used interchangeably, even though they are different things in reality. The Toolbox therefore provides some knowledge nuggets, explaining what Big Data benchmarking is and what can be expected from benchmarking. The first recommendation for the casual user or beginner is to visit the Frequently Asked Question (FAQ) section of the Toolbox, which includes definitions and links to other pieces of knowledge, as shown in Figure 51.

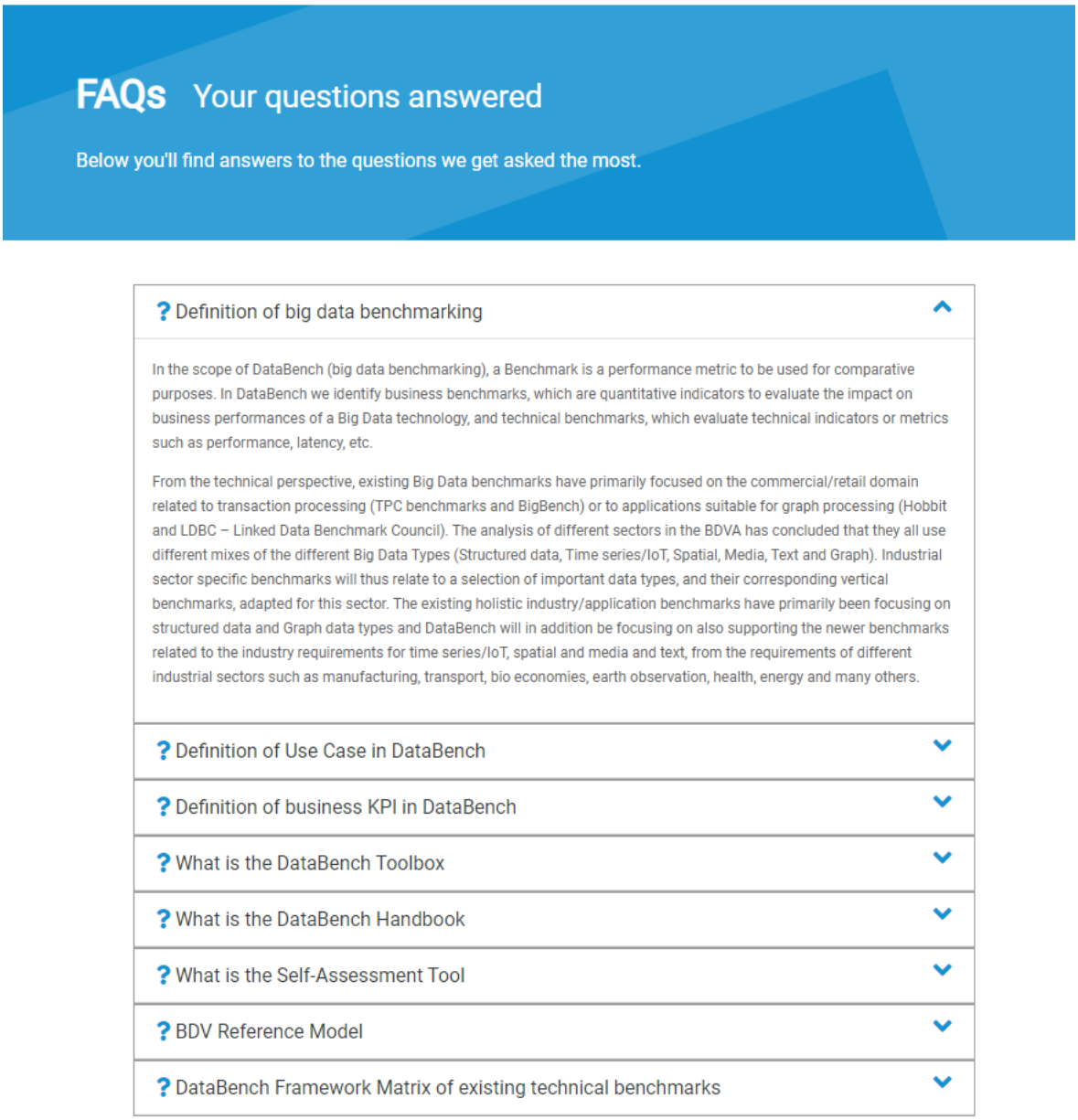

Figure 51, FAQ Section of the Toolbox
Source: DataBench Toolbox

As mentioned above, 'user journeys' will be updated after receiving feedback from users. At the time of writing this document, 'user journeys' for beginners and casual technical users include the following information:

- The use of search functionalities: As explained in section 6.4. – full-text search, search by tag (guided search), and search by BDV Reference Model and search by Blueprint/Pipeline.
- How to browse the entire Knowledge Nugget Catalogue and Technical Benchmark Catalogue by selecting the appropriate options in the menu located in the header: Once located, the user can select one technical benchmark and navigate to that page to look for more content. Benchmarks are marked with metadata to indicate their main features.

- Browsing Big Data architecture blueprints: Various Big Data usage patterns for different domains have been gathered in the form of architecture blueprints. These can help the user understand what others are doing. The suggestion is to type 'blueprint' into the search box to find them, but blueprints can be found in other ways, such as by navigating from other nuggets or typing the specific industry in the search box.

In a similar way, the Toolbox offers 'user journeys' for advanced business users, as shown in Figure 52.

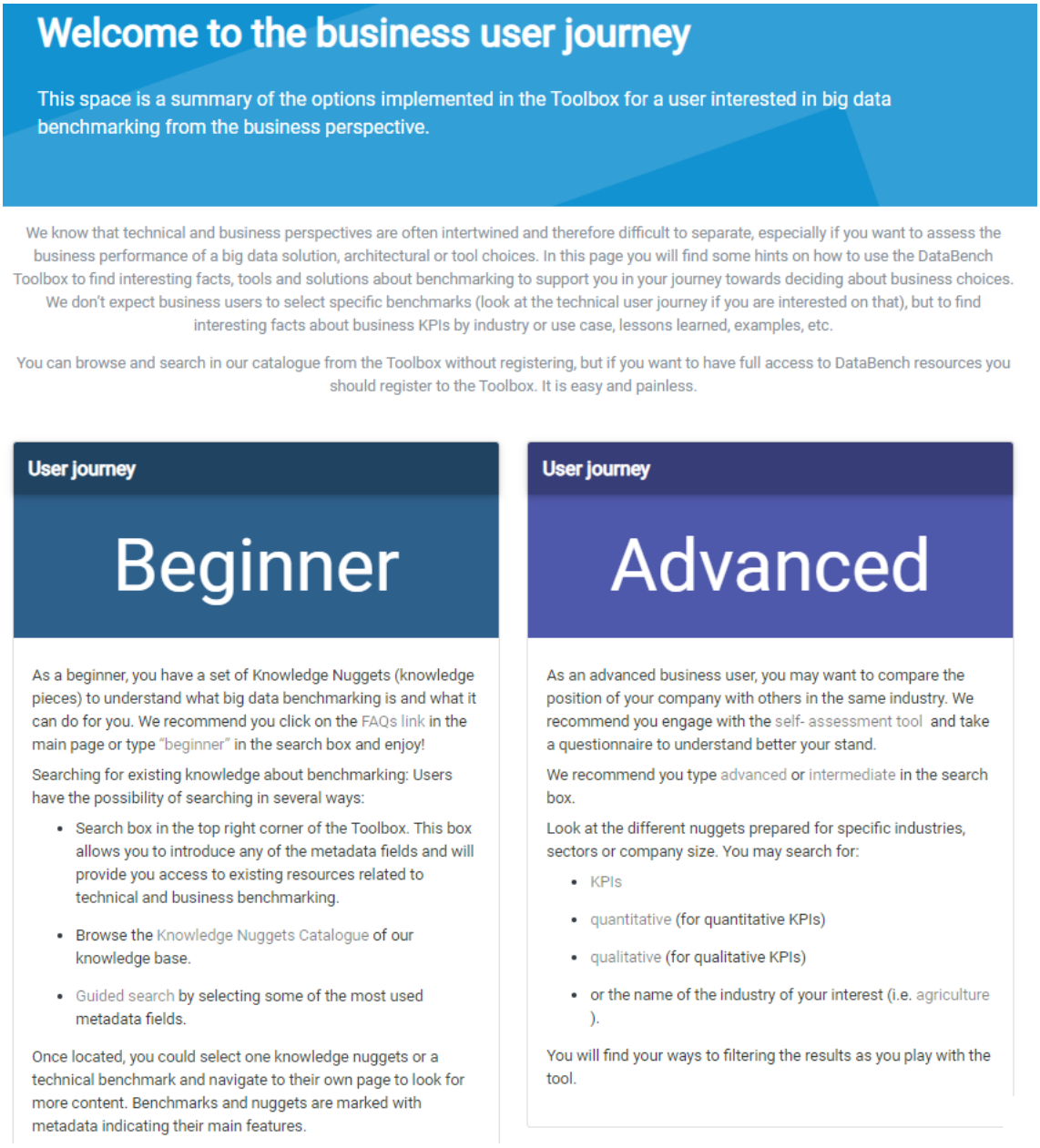

Figure 52, Business User Journeys
Source: DataBench Toolbox

Figure 50 and Figure 52, respectively, show the current 'user journeys' for more advanced technical users and business users. As mentioned above, these 'user journeys' will be updated and upgraded based on feedback received from users.

6.5.2 Support for Benchmarking Providers

Sections 6.6.1 and 6.6.2. provide an overview of how the tool supports benchmark providers when they want to add new benchmarks to the platform. As in previous cases, the 'User Journey' section of the Toolbox homepage shown in Figure 49 gives access to information related to adding new benchmarks. It is divided in two routes:

- Adding a new benchmark to the catalogue
- Integrating a benchmark for automated deployment and execution

Figure 53 displays 'user journey' information for these two different paths. This information is very similar to that discussed in sections 6.6.1 and 6.6.2, complemented with more information gathered from deliverable D3.4[1].

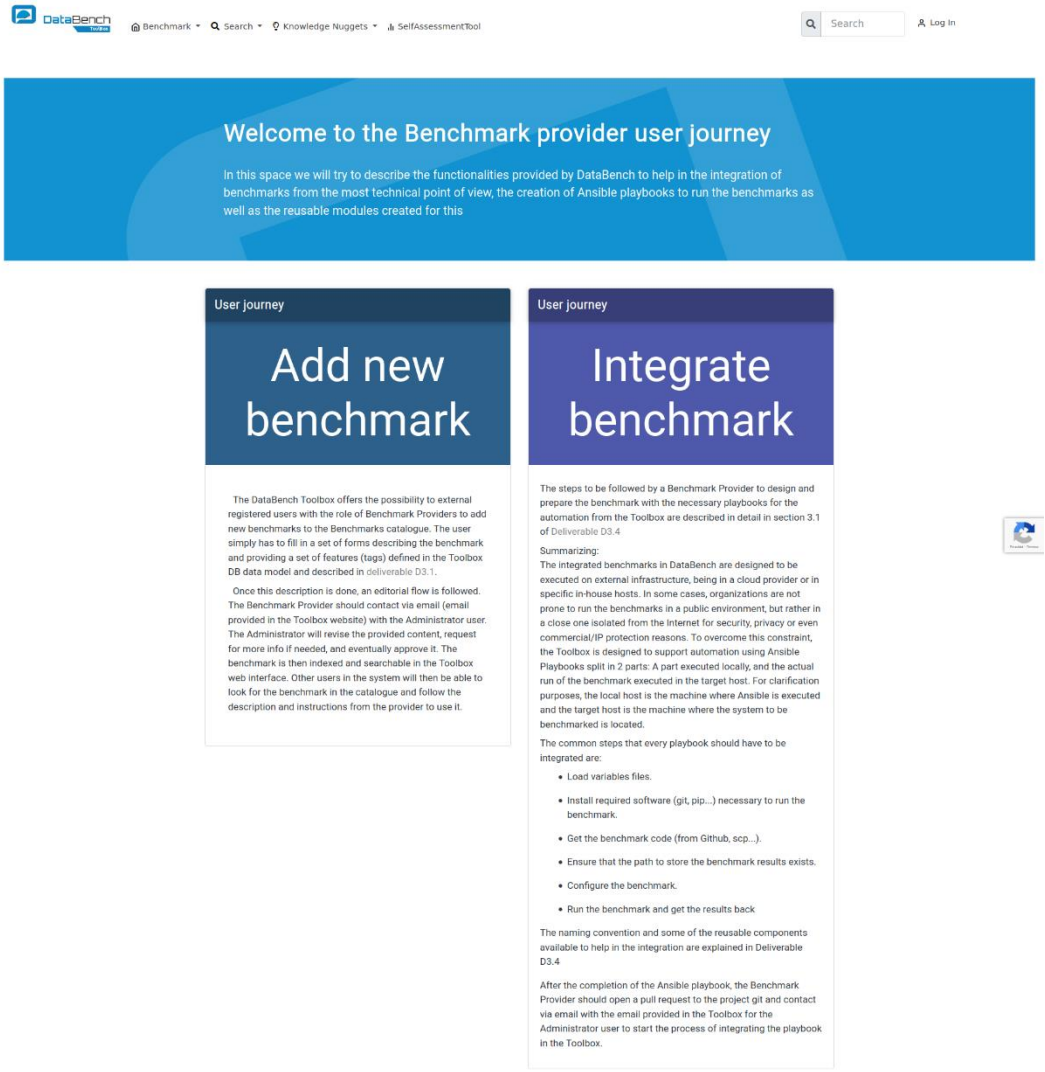

Figure 53, 'User Journeys' for Benchmark Providers
Source: DataBench Toolbox

6.5.3 Support for Benchmarking Experts

In addition to the 'user journeys' offered for technical users, as shown in Figure 50, 'user journeys' are offered specifically for advanced technical users who are experts on Big Data benchmarking. These users belong to, or are aware of, the work done in existing benchmarking initiatives (e.g. TPC, Bench Council, and STAC Benchmark Council, among others). More information about these organisations can be found on the DataBench website[3].

However, with new benchmarking initiatives and different benchmarks popping up almost every other day, maintaining a comprehensive and up-to-date list of benchmarks is quite difficult – even for benchmarking experts. The Toolbox provides such a list, with a link to each benchmark in the Benchmark Catalogue, which is accessible from the menu of the Toolbox (using the Benchmark/Benchmark Catalogue option). A screenshot of the catalogue is shown in Figure 42 (par.6.2).

The catalogue provides access to the list of benchmarks introduced in the Toolbox. Each benchmark is annotated with tags, which enable searching and navigation using the search functions described in this document. The benchmarks are selectable by clicking on them, and each one shows the tags annotated to it, as shown in the example in Figure 54 of the YCSB benchmark.

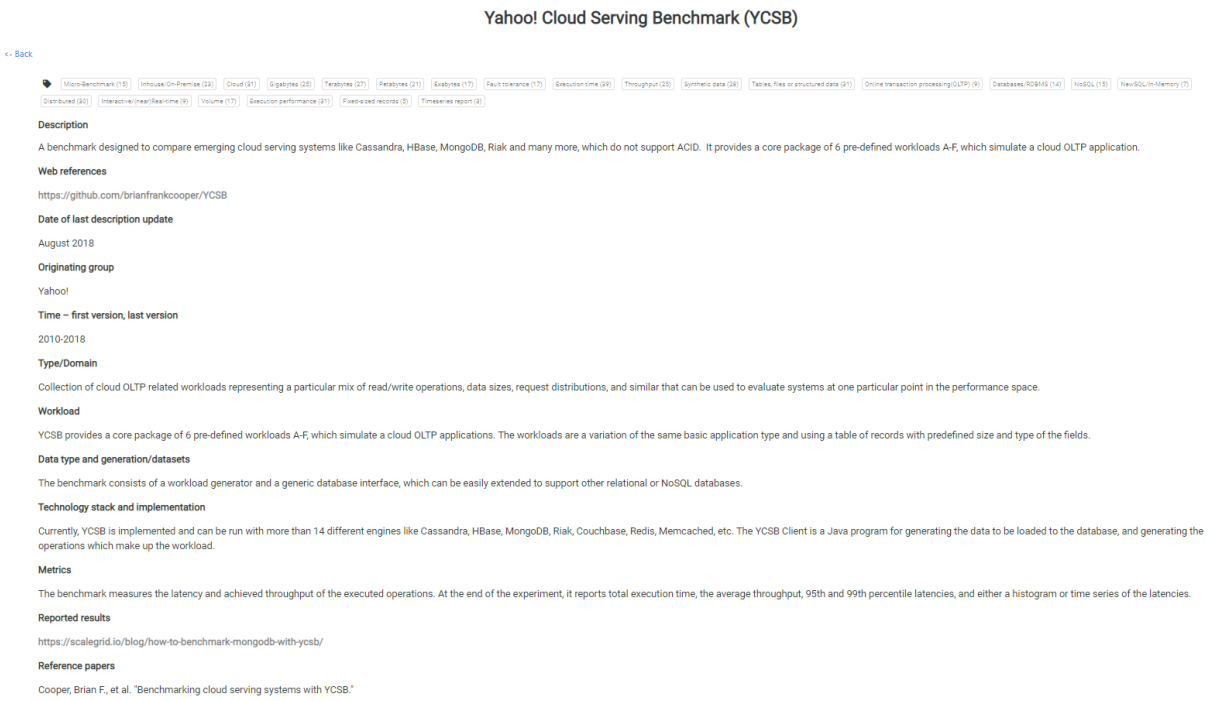

Figure 54, Browsing a Specific Benchmark
Source: DataBench Toolbox

3 https://www.databench.eu/community/

Note that the page provides both well-defined metadata (description, web reference, dates, domain, workloads, metrics, etc.) and tags above the description to navigate to other benchmarks or to knowledge nuggets that relate to the particular benchmark. This shows the navigational approach followed by the Toolbox.

Some benchmarks (e.g. YCSB) have more options for advanced users. Registered users with the technical user role see more interactive information than that shown in Figure 51. In such cases, benchmark providers and DataBench administrators have enabled the integration, automation, deployment, and execution, as explained in section 6.5.2., above. With these specific benchmarks, the registered user sees a deployment and execution panel, as shown in Figure 55.

Note that this advanced feature of the Toolbox is only available for a few select benchmarking tools, although benchmark providers have the option to automate their own benchmarks in the same way. Users can always follow the weblinks and the instructions on the webpages of the benchmark providers and install these benchmarks manually, as is the case with the benchmarks that are not fully integrated.

An explanation of how to configure, deploy, and execute integrated benchmarks is provided in deliverable D3.4[1]. Here is a summary:

- For automated deployment and execution, the Toolbox relies on the use of Ansible[4]. Ansible is an automation engine that allows the user to automate the configuration, deployment, orchestration, and even execution of IT tasks.
- The option to execute an integrated benchmark from the Toolbox is only available to registered users with the roles of technical user and administrator. Registered users of the Toolbox with these roles do not have to use the integrated benchmarks or Ansible at all; they can customise the host machines to deploy the benchmark and customise the configuration file proposed by the Toolbox for the benchmark to be able to perform the necessary actions.
- The host machines on which the benchmark should run can be any hosts available to the user, identified by IP addresses and credentials. The user may create the inventory on the fly via the configuration file and store it for further usage (i.e. to select in successive runs). The credentials to access this infrastructure can be created from the user profile to avoid security issues (i.e. via the 'Credentials' option under the user profile icon, on the right-hand side of the page). The machines are provided by the user, either in house or via a 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 particular benchmark, the outputs and metrics measured, the elements to compare, etc. In the example shown in Figure 55, the YCSB benchmark asks for variables such as the host name, databases to compare, users to access the selected databases, path to retrieve the results, etc. The user tailors the variables in the file.

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

- Once these steps have been completed, the technical user clicks on the 'Launch Job' button at the foot of the page, and the system automatically deploys and runs the benchmark.

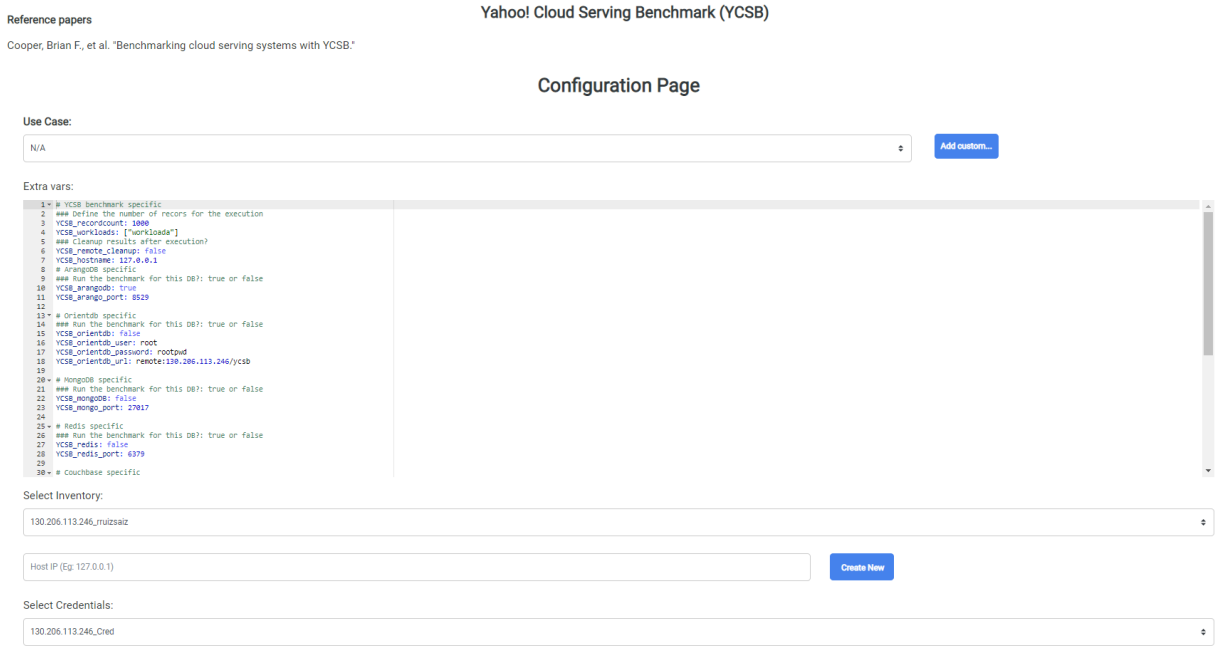

Figure 55, Browsing and Interacting with an Integrated Benchmark
Source: DataBench Toolbox

The results of the run can be accessed either at the host where the installation took place or via the 'Benchmark/Results' option in the menu. This last option enables the user to visualise the results file of all the runs made automatically via the Toolbox. Note that this data is only accessible by the user that initiated the deployment/execution. Visualisation examples are show in Figure 56 (detailed results of a single execution) and Figure 57 (a time series of executions of the same benchmark showing some specific metrics). Note that these visualisations are tailored for each benchmark. When a new benchmark is fully integrated, the benchmark provider should provide a means to visualise the data.

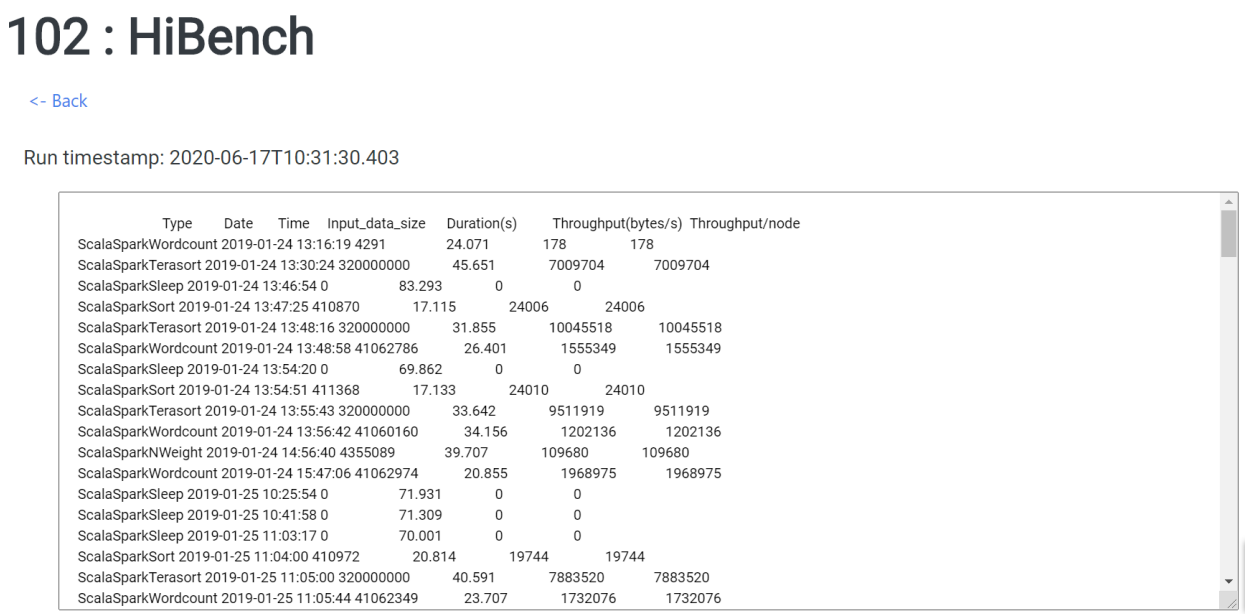

Figure 56, Visualisation Example: Results from an Integrated Benchmark
Source: DataBench Toolbox

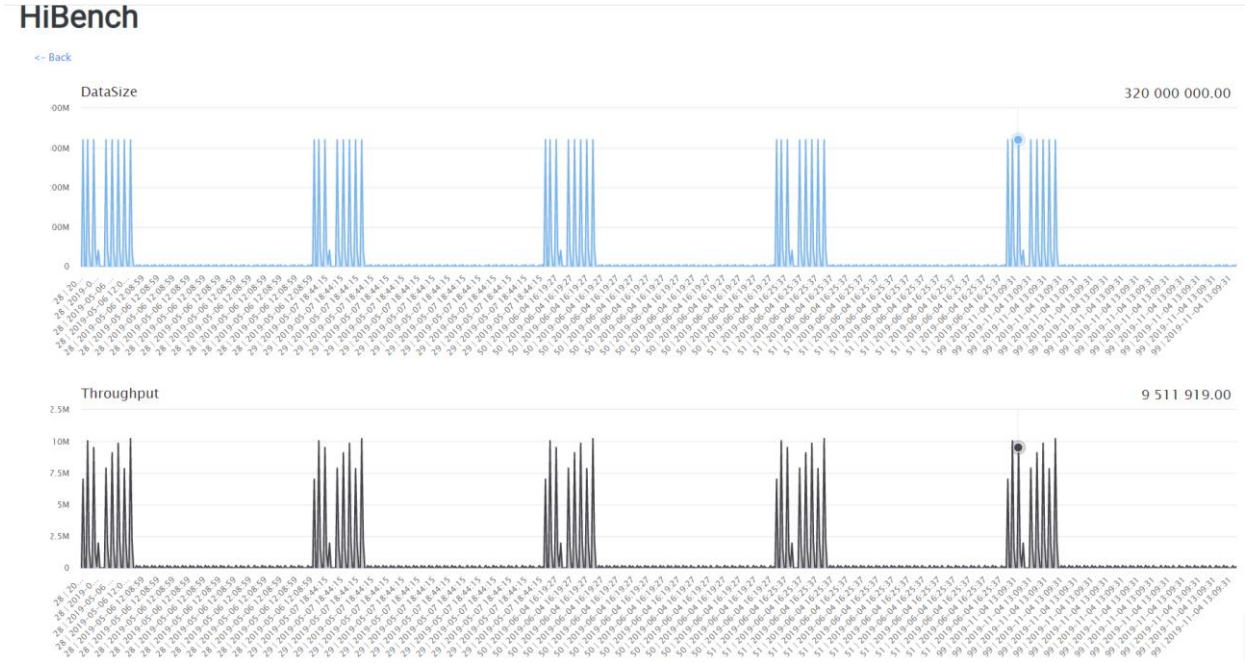

Figure 57, Browsing and Interacting with an Integrated Benchmark
Source: DataBench Toolbox

6.5.4 Support for Big Data R&D Projects

As mentioned above, among the main target users of DataBench are EC-funded Big Data projects – especially those funded under topics related to the BDV PPP. We have thus far

identified two main lines of collaboration with these projects related to their needs at the different stages of their lifecycles:

- The beginning of the project: Projects usually undergo a phase in which they need to assess how advanced the tools and methods are in relation to their goals. In this phase, they need to check different Big Data/AI tools, frameworks, and applications and take informed decisions on the best options to fulfil their needs. At this stage, many of the existing benchmarks may help decision makers compare tools/solutions and make choices. The example of YCSB shown above is clear: It allows the user to compare several performance metrics of different NoSQL databases (ArangoDB, OrientDB, Redis, MongoDB, and Couchbase). Many other benchmarks might enable the user to better understand which tools are available and to execute 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 in obtaining ideas on how the tools and applications implemented in the project perform. A few application benchmarks might help in this regard, but the applications developed in the project often require a more specific benchmark – one closer to the application itself. This entails the development or customisation of new benchmarking tools – normally, by taking and adapting an existing one. It is always best to use an existing benchmark to save time and development resources. In the cases when this is not possible one should think about alternative approaches like benchmark customization. For this to be successful one needs to have a clear idea about the application requirements and then to search for the most related existing benchmark using the Toolbox searches described previously. For these, DataBench offers some guidelines in the form of 'user journeys' and knowledge nuggets. Then when the most relevant benchmark is found and covers the same domain or the same use case scenario of what one plans to implement, one can use it as a best practice implementation and develop customized extensions before starting a new implementation from scratch.

Finally, projects may be interested not only in technical benchmarking, but also in business aspects. The par. 6.5.5 below will likely also be of interest for projects, as will all of the information presented in chapter 5.

6.5.5 Support for Business Users

Figure 52, above, shows the current 'user journeys' offered for more advanced business users. As mentioned above, these 'user journeys' will be updated and upgraded based on feedback received from users.

Knowledge nuggets are provided for all Toolbox user types, but business users may benefit the most from them. These nuggets represent information from both project results and the benchmarking community as a whole. An example of a knowledge nugget can be seen in Figure 58, below.

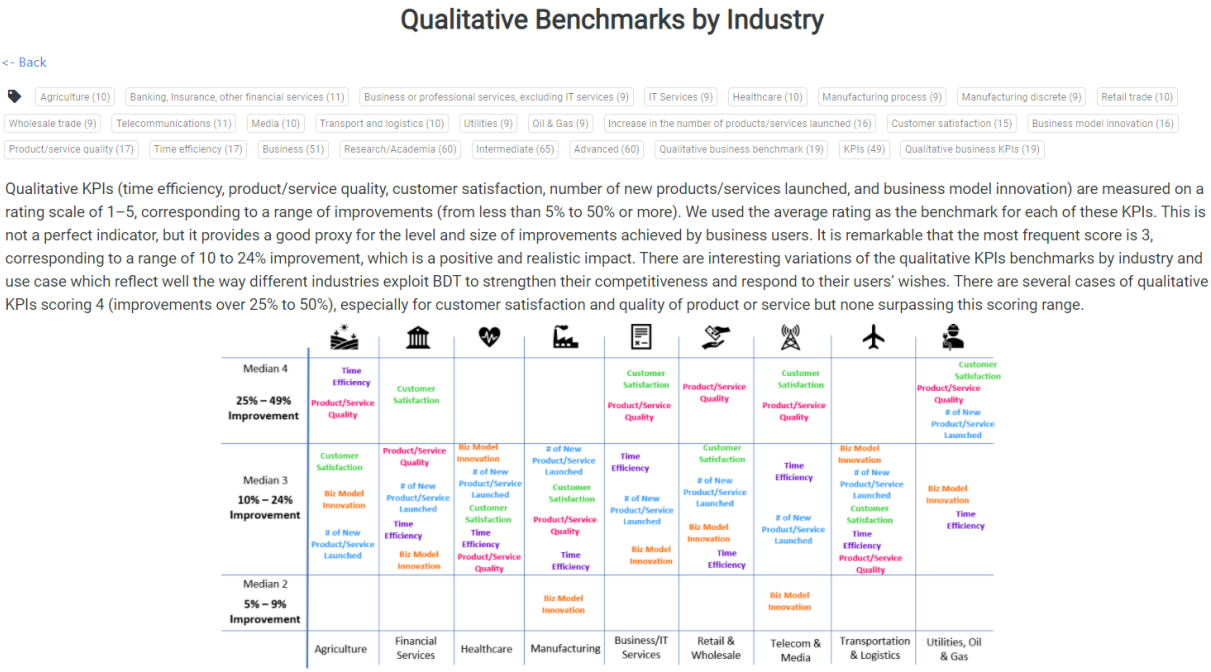

Figure 58, Example of a Knowledge Nugget
Source: DataBench Toolbox

Paragraph 6.6.3. explains how this knowledge base can be extended by any user, thus making it a live component to understand the benchmarking landscape.

We have covered the usage of the Toolbox for casual users above. Business users may want to find information about their sectors, compare certain things with those of similar organisations or competitors, or simply check what others are doing with Big Data and AI. Several knowledge nuggets in the Toolbox are tagged with 'advanced' or 'intermediate', indicating the initial entry point, depending on the user type. By typing these labels in the search box of the Toolbox, a list of knowledge nuggets with some hints will be offered to the user.

Nuggets also relate to specific domains and sectors. Typing the name of the sector (e.g. 'agriculture') provides a list of associated nuggets, as well as the technical benchmarks used in that sector. In fact, searching for nuggets by industry/sector and/or company-size segment is among the main 'user journey' tips offered to business users. These users may find information by typing into the search box terms such as 'KPI', 'quantitative' (for quantitative KPIs), 'qualitative' (for qualitative KPIs), etc.

As mentioned previously, the project receives feedback from users, and it prepares and updates the 'user journeys' offered for business users, including hints and knowledge about business benchmarking and related info.

6.6 Methodology to Add New Knowledge/Benchmarks

The Toolbox is extensible by design, enabling the catalogues to be populated with new and updated benchmarking tools and knowledge nuggets. This section explains how different registered user types can add to the DataBench Toolbox and covers the editorial processes in place to ensure the quality of information provided. Toolbox administrators are the top-

level custodians of editorial content and are responsible for the information accessed in the Toolbox. They are in charge of user management, as well as approving or rejecting the information provided by users.

In DataBench, we have implemented a simple workflow whereby users can get in touch with administrators via email after submitting new additions to the catalogues. Users who have finalised additions or who want to know more about how to do so can find a contact email address at the foot of Toolbox's homepage. Administrators have the possibility to revise the elements pending approval (and the list thereof) and can communicate with the respective users via email.

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

6.6.1 Support for Adding New Benchmarks to the Catalogue

This process was explained in deliverable D3.4[1] (section 3.1). The following text is quoted from that deliverable:

'The DataBench Toolbox offers the possibility to external registered users with the roles of benchmark provider and administrator to add new benchmarks to the Benchmark Catalogue. The user simply fills in a form to describe the benchmark and provides a set of features (tags) defined in the Toolbox data model, as described in deliverable D3.1 [9].

'Once this description is complete, the editorial process starts. The benchmark provider contacts the administrator via email (as provided on the Toolbox website). The Administrator revises the provided content, requests more info if needed, and eventually approves it. The benchmark is then indexed and made searchable in the Toolbox web interface. Other users in the system can then search for the benchmark in the catalogue and follow the description and instructions from the provider on how to use it.' [1]

6.6.2 Support for Integrating New Benchmarks to Be Executed from the Toolbox

This process was explained in detail in deliverable D3.4[1] (section 3.1). In this section, we only summarise the steps to be done without entering into the technical details explained in D3.4[1]:

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 establish the automation process using Ansible (whether totally or partially) to enable the configuration, deployment, and execution of the benchmark and the return of results to the Toolbox database for further visualisation. This automation process requires two main steps:

1. **The creation of an Ansible playbook for the benchmark:** The benchmark provider should provide configuration and automation for their benchmark. D3.4[1] explains the steps to be done in Ansible to enable this preparation, which basically consists of using two Ansible playbooks: one to be executed locally (in the machine on which Ansible is running, the DataBench host machine), and the target host(s), where the deployment and execution will take place. D3.4[1] explains what this entails for the benchmark provider. The Toolbox also offers knowledge nuggets containing such information, as explained in the benchmark provider 'user journey'. The Toolbox offers a set of reusable playbooks to facilitate the process. The

benchmark provider is advised to contact the Toolbox administrator to proceed and/or to obtain information and guidelines on the process.

2. **The integration of the ansible playbook in the toolbox by the administrators:** After the integration of the playbooks, the benchmark provider contacts the administrator to finalise the process. The administrator revises the code and uploads it to a Git repository, creates the template for the new benchmark in AWX, selects the project and the playbook, enables the configuration of the template at runtime from the web, and finalises the integration. All these steps are explained in detail in D3.4[1].

6.6.3 Support for Adding New Knowledge Nuggets

As with 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 is provided with a form, such as the one shown in Figure 59.

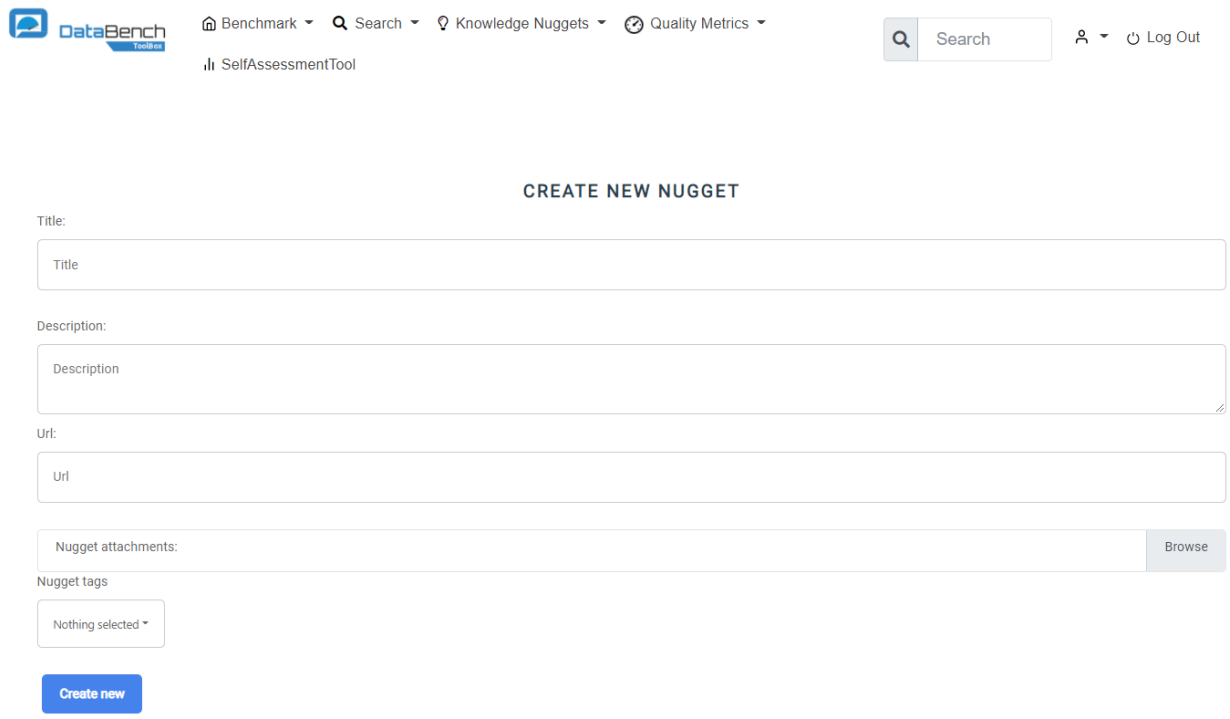

Figure 59, Knowledge Nugget Creation Form
Source: DataBench Toolbox

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) and supported by an external reference link (a URL) and one or more attachments. Image attachments are rendered within the nugget during visualisation, once the nugget has been approved. Users can select tags to annotate the nugget from a set of existing tags and can propose new tags.

Once the content has been submitted, an editorial workflow starts. The user contacts the Toolbox administrators via email. An administrator then revises the content and eventually approves or rejects the nugget.

7 Sustainability and Usability of DataBench' results

This DataBench Handbook is designed to provide a guide to all the tools and services provided by DataBench also after the end of the project in December 2020, thereby ensuring their sustainability and exploitation in time. The main outcome of the project is of course the Databench Toolbox, but the background and in-depth analysis developed by the project will also remain fully available in the projects' deliverables as valuable shared knowledge as explained in the previous chapters.

DataBench' partners have made sure of the continuing availability of this knowledge and results after December 2020. The Databench project was originally conceived in the light of supporting the Big Data and Benchmarking communities, including ICT projects, research organizations, stakeholder organizations or companies that are part of the ecosystem BDA and AI ecosystems. In fact, all the software developed, instructions for using the benchmarks and the toolbox, the market analysis which has been performed, the benchmark validation techniques as well as the surveys of companies actively using these techniques are provided as open data and open source. They are provided free of charge on the project's website and can be freely used or downloaded. The project website and all of the materials will be available online and all services, documentation and links will be maintained for at least 2 years from the end of the project on December 31st, 2020. As is the practice of IDC and many of the partners, given the low cost of maintaining the physical hardware, domains and network connectivity, this period will likely be extended.

However, organizations interested in extracting the maximum value from DataBench results should consider the option to integrate the Toolbox in their own IT infrastructures as a benchmarking process and continue feeding updated content to it. There are three main reasons to do so. First the maintenance of the benchmarks themselves will not be continued after the end of the project. This is important because although the benchmarking methodology is considered to be of high value, the benchmarks (as is the case with all metrics-based analysis) have been measured at a given point in time and their validity will inevitably decrease (more or less slowly) as time goes. The benchmarks are a reflection of the metrics and conditions during the project lifetime and may not be current after that date. For example, if one considers benchmarks for the effective treatment of gram-positive bacteria in wounds during the first world war, one would find they are no longer applicable for benchmarking the treatment of the same bacteria in wounds during the second world war, given the fact that Fleming discovered Penicillin in 1928. The same is true with BigData Analytics, as the types and quantities of data and the analytical procedures are undergoing a moment of significant evolution, while the onset of new AI and computing technologies is transforming the benchmarking tools at a very high rate. The Benchmarks we have provided clearly reflect the situation for European industry today but will invariably loose pertinence as technology and market conditions progress.

Second, benchmarking is a machine intensive analysis, databases are large and computational needs can grow exponentially as requirements are expanded. The computational resources available on the current DataBench demonstration platform may not be sufficient for future operational benchmarking that organisations may require.

Finally, data in organisations and research is commonly proprietary or sensitive. Data used in analytics processing from companies is often considered proprietary pertaining to commercial, technical or supply chain information regarding their operations. Organizations are often using benchmarking to improve their production, maintenance or

delivery processes to obtain a competitive advantage. In many cases they may not be comfortable using DataBench resources to perform their benchmarking and maintain their own industrial or research benchmarks. For these reasons it is understood that organisations will in many cases need to adopt and adapt DataBench tools to their own IT infrastructures.

DataBench partners do in fact understand this and have made available the necessary knowledge and data to enable other ICT projects, research organizations, stakeholder organizations or companies to take over and adopt our tools and results in a proactive way. More specifically, the consortium will collaborate with the H2020 Innovation Action EUHubs4Data, just started with 3 years planned duration, so that DataBench tools and results can be leveraged by the Big Data Innovation Hubs network for training and exploitation by European enterprises and possibly updated.

In conclusion, in order to use this Handbook and the referenced deliverables to learn about and improve Big Data benchmarking, potential users should consider the following steps:

I. Identify the kinds of analytics that they can expect to perform and the types of economic market and business metrics that are important to European industry[5];
II. Identify business objectives and the KPIs for their application areas[6];
III. Perform the DataBench Self-Assessment Survey considering the types of analytics that they intend to use[7];
IV. Identify key technical objectives (BDV Reference model area scope)[8];
V. Identify and map operations expected to be monitored to the specific DataBench pipeline steps[9];
VI. Identify relevant Big data types and processing types and architectural patterns[10];
VII. Identify any relevant use case independent specific use-case Blueprints for their Domain[11];
VIII. Consider appropriateness of relevant standards for big data and AI technologies (ISO SC42);
IX. Identify relevant benchmarks - consider analysis of technologies. Consider use of relevant benchmarks for the analysis of relevant/provided tools and components;[12]
X. Analyse and report results in the DataBench knowledge nuggets database[13];
XI. If deemed appropriate, Organizations should download the application code and follow the instructions to implement the benchmarking platform as described in the system[14].

5 While described in this document, this aspect is fully explored in Deliverable D2.1 Economic, Market and business analysis methodology.

6 These aspects are covered in deliverable D4.3 Evaluation of business performance.

7 Can be found at https://www.databench.eu/self-assessment-survey/

8 Deliverable D1.1 Industry Requirements with benchmark metrics and KPIs provides detailed information

9 These are described in D1.3 Horizontal Benchmarks – Analytics and Processing, D1.4 Horizontal Benchmarks – Data Management

10 Idem.

11 See D4.3 Evaluation of Business Performance

12 Described in D5.4 Analytic modelling relationships between metrics, data and project methodologies

13 Found at https://databench.ijs.si/knowledgeNugget/listKnowledgeNuggets

14 Instructions are included at: https://www.databench.eu/databench-toolbox/

We expect many organizations will want to follow this path and several are currently engaging with a number of organizations, trade representatives and research projects that are in some stage of the process described above. If any organization should want to follow this example, all of the documentation for the various steps is referenced and can be found on the DataBench website. In the case that additional help should be required a number of the project partners are carrying on the research and maintenance of the tools and the coordinator would be more than happy to field inquiries and direct the interested parties to the most appropriate interlocutor or resource. Email inquiries will also be addressed and responded to and all email inquiries should be addressed to while informative videos can be found on the DataBench channel of Youtube. The DataBench website at www.databench.eu contains extensive information and complete documentation and of course access to the ToolBox.

8 References

All DataBench deliverables are retrievable at https://www.databench.eu/public-deliverables/.

[1]. D.1.1 DataBench Deliverable **Industry Requirements with Benchmark Metrics and KPIs.**

[2]. D2.1, DataBench Deliverable **Economic and Market Analysis Methodology**

[3]. D.4.1 DataBench Deliverable **Data Collection**

[4]. D.2.2 DataBench Deliverable **Preliminary Benchmarks of European and Industrial Significance**

[5]. D.4.2 DataBench Deliverable **Data Collection results**

[6]. D.4.3 DataBench Deliverable **Evaluation of Business Performance**

[7]. D.5.5 DataBench Deliverable **Final report on methodology for evaluation of industrial analytic projects scenarios**

[8]. D.1.2 DataBench Deliverable **DataBench Framework, with Vertical Big Data Type Benchmarks**

[9]. D.3.1 DataBench Deliverable DataBench Architecture ver.2, updated November 2020

[10]. D.5.1 DataBench Deliverable Initial Evaluation of DataBench Metrics

[11]. D1.3 DataBench Deliverable D1.3. **Horizontal Benchmarks – Analytics and Processing**

[12]. D1.4 DataBench Deliverable **Horizontal Benchmarks – Data Management**

[13]. D.5.4 DataBench Deliverable **Analytic modelling relationships between metrics, data and project methodologies**

[14]. D.2.4 DataBench Deliverable **Benchmarks of European and Industrial Significance**

[15]. D3.4. DataBench Deliverable **Release Version of DataBench Toolbox, Including Visualisation and Search Components**

[16]. Big Data Value Strategic Research and Innovation Agenda (BDV SRIA). http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf

[17]. DataBio – Data-driven Bioeconomy is a H2020 lighthouse project running pilots on Big data for agriculture, forestry and fishing https://www.databio.eu/en/about-databio/summary/

[18]. ISO/IEC 20547-3:2020, Information technology — Big data reference architecture — Part 3: Reference architecture

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