Evidence Based Big Data Benchmarking to Improve Business Performance

D5.5 Final report on methodology for evaluation of industrial analytic projects scenarios

Abstract

The objective of this document is to evaluate the usage of the DataBench Toolbox. This, is done with both an Analytic and a Project based Evaluation of the Toolbox, followed by a Use case-based Evaluation of the DataBench Blueprints and a Project data analysis with the DataBench Observatory. With a foundation in these evaluations a strategy for further sustainable evolution of the DataBench Toolbox is suggested. The report first provides an Analytic evaluation of the DataBench Toolbox followed by a project-based evaluation. The project-based evaluation has a focus on how projects can use pipelines and blueprints for the analysis of potential Big Data and AI technologies and corresponding benchmarks and also how project experiences with their pipelines and blueprints can be reported and shared as knowledge nuggets in the DataBench Toolbox. A Use case-based Evaluation of the DataBench blueprints is based on the three domains of Agriculture, Heavy Equipment optimization, Smart manufacturing and Healthcare, diagnostic systems. A Project data analysis with the DataBench Observatory is done with an initial Cordis project data description and further Big Data PPP project data descriptions in progress. The report concludes with a description of the strategy to support further current and future projects through a continued sustainability strategy for the DataBench toolbox usage and evolution.

Deliverable D5.5	Final report on methodology for evaluation of industrial analytic projects scenarios
Work package	WP5
Task	5.3
Due date	31/12/2020
Submission date	23/12/2020
Deliverable lead	Lead Consult
Version	2.0
Authors	Lead Consult (Todor Ivanov, Diana Stancheva) SINTEF (Arne J. Berre, Aphrodite Tsalgatidou, Brian Elvesæter) POLIMI (Chiara Francalanci, Giulio Costa, Gianmarco Ruggiero) JSI (Inna Novalija, Marko Grobelnik, Besher M.Massri) ATOS (Tomás Pariente, Iván Martínez and Ricardo Ruiz)
Reviewers	Ricardo Ruiz, Inna Novalija

Keywords

Benchmarking, Evaluation, Big Data, AI, Pipeline, Framework, Blueprint, Methodology, Toolbox, Observatory

Disclaimer

This document reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information this document contains.

Copyright Notice

Table of Contents

List of Figures

List of Tables

Executive Summary

The objectives of the work in work package 5 "Technical Evaluation using the DataBench Toolbox", is to provide a framework and methodology for the usage of the Toolbox with an associated validation and assessment approach as a basis for further sustainable evolution of the DataBench Toolbox.

The objective of this document, D5.5 "Final report on methodology for evaluation of industrial analytic projects scenarios" is to evaluate the usage of the DataBench Toolbox, which is done with both an Analytic and a Project based Evaluation of the Toolbox, followed by a Use case-based Evaluation of the DataBench Blueprints and a Project data analysis with the DataBench Observatory. With a foundation in these evaluations a strategy for further sustainable evolution of the DataBench Toolbox is suggested.

The report first provides an Analytic evaluation of the DataBench Toolbox. This includes an analytic methodology with evaluation criteria with a rank and a score, and a corresponding evaluation and discussion of the outcome with suggestion for features and further improvement possibilities.

Further a Project based evaluation of the DataBench Toolbox is presented. This has a focus on how projects can use Pipelines and Blueprints for the analysis of potential Big Data and AI technologies and corresponding benchmarks and how project experiences with their pipelines and blueprints can be reported and shared as knowledge nuggets in the DataBench Toolbox. The project-based evaluation has been supported by two campaigns through the ReachOut beta testing system: "Generation of Architectural Pipelines and Blueprints" and "Finding the right benchmarks for technical and business users". The results from 6 example Big Data PPP projects based on this are reported. The DataBench methodology and Toolbox have been applied to/used by the following projects: I-BiDaas, TheyBuyForYou (TBFY), DeepHealth, DataBio, Track&Know and CLASS. For each of these projects, their description, the project objectives and architecture, are described and illustrated with figures for how the steps of the pipelines and blueprints area used in each project and how they are related to the steps of the DataBench pipeline and Blueprints. For each project, relevant benchmarks and nuggets are retrieved with the use of the DataBench Toolbox.

A Use case-based Evaluation of the DataBench Blueprints follows. This includes use cases from the three domains of Agriculture, Heavy Equipment optimization, Smart manufacturing and Healthcare, diagnostic systems

Finally, a Project data analysis with the DataBench Observatory is done, with an initial Cordis Project Data Description and further Big Data PPP Project Data Descriptions in progress.

The report concludes with a description of the strategy to support further current and future projects through a continued sustainability strategy for the DataBench Toolbox usage and evolution.

1. Introduction

1.1. Objective

The objectives of WP5 " Technical Evaluation using the DataBench Toolbox " is to provide a framework and methodology for the usage of the DataBench Toolbox with an associated validation and assessment approach as a basis for further sustainable evolution of the DataBench Toolbox.

This report, D5.5 "Final report on methodology for evaluation of industrial analytic projects scenarios" reports on the approach to evaluate the usage of the DataBench Toolbox with both an Analytic and a Project based Evaluation of the Toolbox, followed by a Use case-based Evaluation of the DataBench Blueprints and a Project data analysis with the DataBench Observatory. With a foundation in these evaluations a strategy for further sustainable evolution of the DataBench Toolbox is suggested.

The WP5 work package is using the results from WP1, WP2 and WP3 and the result of this work package will be the technical validation of the DataBench framework from WP1 and WP2 and the Toolbox from WP3 – both with possible extensions based on the usage and feedback from actual projects. This includes how to validate and assess the correspondence of the technical metrics and the resulting benchmarks collected and refined in WP1 and WP2 and integrated in the Toolbox developed in WP3, to make sure that they effectively correspond to the intentions of the original tools and needs of the project communities.

The initial WP5 deliverables D5.1 "Initial Evaluation of DataBench Metrics", D5.2 "Final evaluation of DataBench metrics" and D5.3 "Assessment of technical usability, relevance, scale and complexity" provide the input for the deliverable D5.4 “Analytic modelling relationships between metrics, data and project methodologies". D5.4 provides the methodology and setup which is being supported by the DataBench Toolbox.

The DataBench framework is accompanied by a Handbook (D4.4) and the DataBench Toolbox, which aim to support industrial users and European technology developers who need to make informed decisions on Big Data Technologies investments by optimizing technical and business performance. The Handbook presents and explains the main reference models used for technical benchmarking analysis. The Toolbox is a software tool that provides access to benchmarking services; it helps stakeholders (i) to identify the use cases where they can achieve the highest possible business benefit and return on investment, so they can prioritize their investments; (ii) to select the best technical benchmark to measure the performance of the technical solution of their choice; and, (iii) to assess their business performance by comparing their business impacts with those of their peers, so they can revise their choices or their organization if they find they are achieving less results than median benchmarks for their industry and company size. Therefore, the services provided by the Toolbox and the Handbook support users in all phases of their users’ journey (before, during and in the ex-post evaluation of their Big Data and AI technology investment) and from both the technical and business viewpoint.

The DataBench Toolbox is further evaluated in this deliverable D5.5 "Final report on methodology for evaluation of industrial analytic projects scenarios".

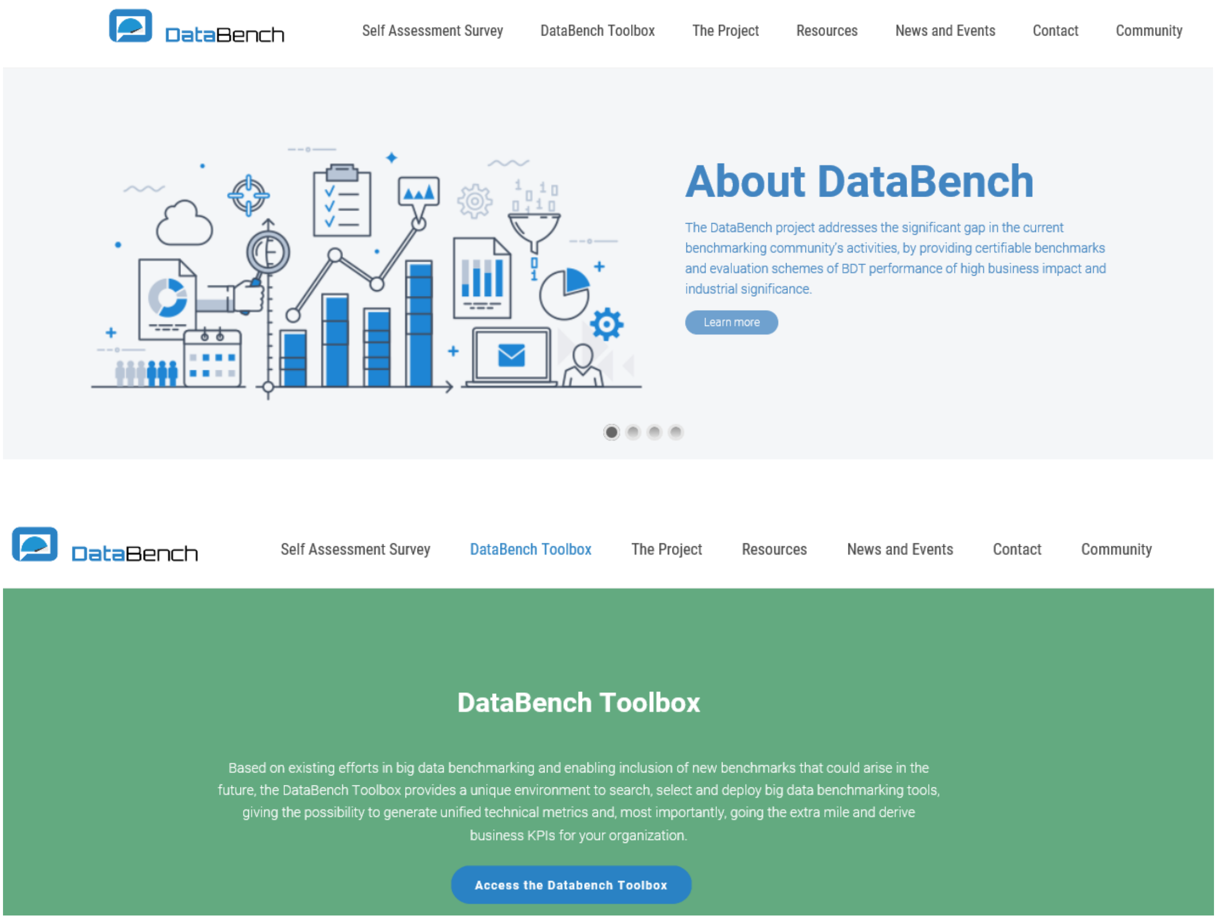
1.2. Structure of the Report

The report is structured as follows:

- Chapter 1 Introduction presents the main objective of the deliverable and the structure of the report.
- Chapter 2 presents the DataBench Toolbox evaluation approach consisting of both an analytics and project based evaluation of the DataBench Toolbox, accompanied by a use-case based evaluation of the DataBench Blueprints and a Project data analysis with the DataBench Observatory.
- Chapter 3 describes an Analytic evaluation of the DataBench Toolbox. This includes an analytic methodology with evaluation criteria with scoring and range, a corresponding evaluation and discussion of the outcome with suggestion for features and further improvement possibilities.
- Chapter 4 presents a Project based evaluation of the DataBench Toolbox, with a focus on how projects can use Pipelines and Blueprints for the analysis of potential Big Data and AI technologies and corresponding benchmark and also how project experiences with their pipelines and blueprints can be reported and shared as knowledge nuggets in the DataBench Toolbox. This has been supported by two campaigns through the ReachOut beta testing system: 1) Generation of Architectural Pipelines – Blueprints and 2) Finding the right benchmarks for technical and business users. The results from 6 example Big Data PPP projects based on this is reported. This is reported with the common structure of Project description, Project Objectives and Architecture, Mapping to DataBench Pipeline and Blueprints and Relevant Benchmarks and Nuggets from DataBench Toolbox. The reported projects are I-BiDaaS, TheyBuyForYou, DeepHealth, DataBio, Track&Know, CLASS. Further projects are in progress as part of the sustainability strategy of DataBench.
- Chapter 5 presents a Use case-based Evaluation of the DataBench Blueprints. This includes in particular use cases from three domains of Agriculture, Heavy Equipment optimization, Smart manufacturing and Healthcare, diagnostic systems
- Chapter 6 describes Project data analysis with the DataBench Observatory, with an initial Cordis Project Data Description and further Big Data PPP Project Data Descriptions in progress.
- Chapter 7 concludes with a description of the strategy to support further current and future projects through a continued sustainability strategy for the DataBench toolbox usage and evolution.

2. Methodology for evaluation of industrial analytic projects scenarios

This document is reporting on the final evaluation of the use of the DataBench Toolbox with the support of the methodology approach from the document D5.4 “Analytic modelling relationships between metrics, data and project methodologies".



About DataBench

The DataBench project addresses the significant gap in the current benchmarking community's activities, by providing certifiable benchmarks and evaluation schemes of BDT performance of high business impact and industrial significance.

DataBench Toolbox

Based on existing efforts in big data benchmarking and enabling inclusion of new benchmarks that could arise in the future, the DataBench Toolbox provides a unique environment to search, select and deploy big data benchmarking tools, giving the possibility to generate unified technical metrics and, most importantly, going the extra mile and derive business KPIs for your organization.

Figure 1: Accessing the DataBench Toolbox from the DataBench website

The functionality of the DataBench Toolbox is described further in the D4.4 DataBench Handbook, but it is also embedded into the support services in the Toolbox itself.

The following Figure 2 depicts the DataBench Big Data and AI Toolbox project methodology from D5.4 "Analytic modelling relationships between metrics, data and project methodologies".

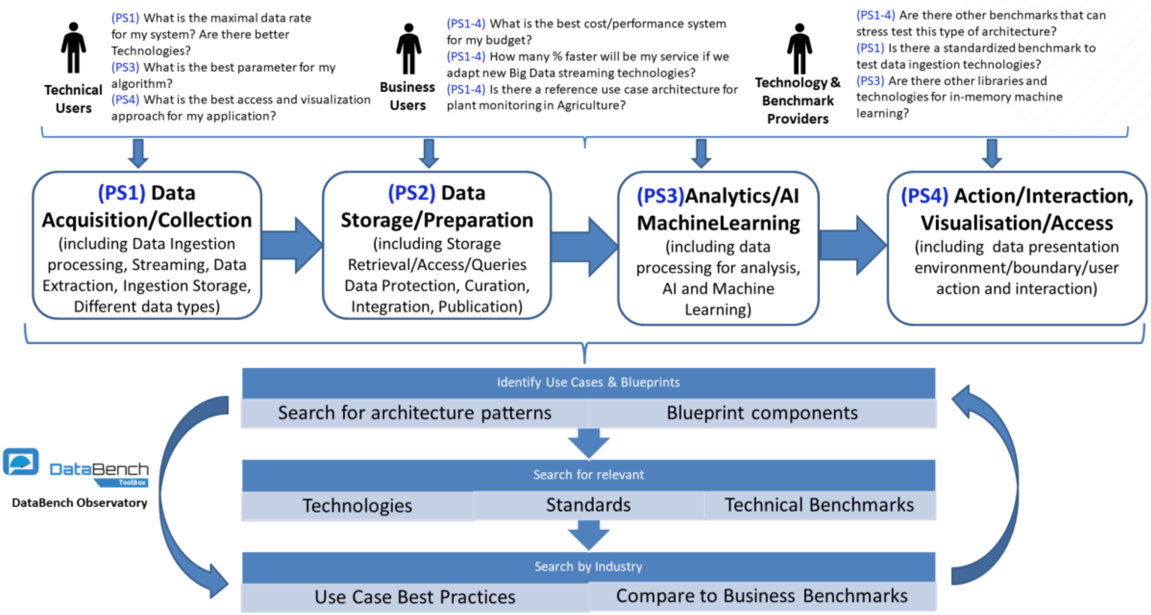

Technical Users
(PS1) What is the maximal data rate for my system? Are there better Technologies?
(PS3) What is the best parameter for my algorithm?
(PS4) What is the best access and visualization approach for my application?
Business Users
(PS1-4) What is the best cost/performance system for my budget?
(PS1-4) How many % faster will be my service if we adapt new Big Data streaming technologies?
(PS1-4) Is there a reference use case architecture for plant monitoring in Agriculture?
Technology & Benchmark Providers
(PS1-4) Are there other benchmarks that can stress test this type of architecture?
(PS1) Is there a standardized benchmark to test data ingestion technologies?
(PS3) Are there other libraries and technologies for in-memory machine learning?
(PS1) Data Acquisition/Collection (including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
(PS2) Data Storage/Preparation (including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
(PS3)Analytics/AI MachineLearning (including data processing for analysis, AI and Machine Learning)
(PS4) Action/Interaction, Visualisation/Access (including data presentation environment/boundary/user action and interaction)
DataBench ToolBox
DataBench Observatory
Identify Use Cases & Blueprints
Search for architecture patterns
Blueprint components
Search for relevant
Technologies
Standards
Technical Benchmarks
Search by Industry
Use Case Best Practices
Compare to Business Benchmarks

Figure 2: Big Data and AI steps related to identification of blueprints, benchmarks, technologies and standards

The following checkpoint list has been suggested for projects that want to take advantage of the DataBench Toolbox for selection and reporting of use of Big Data and AI technologies:

A. Identify Application area and domain (D4.4, D4.3), Business Objectives/KPIs – Consider to do the DataBench Self-assessment Survey from the perspective of any applications (D4.4, D4.3).

B. Identify key technical objectives (BDV Reference model area scope), Identify and map to the 4 pipeline steps, Identify relevant Big data types and processing types + architectural patterns.

C. Identify any relevant Blueprints - use case independent and domain / use-case specific.

D. Identify relevant standards for big data and AI technologies (ISO SC42), Identify relevant technologies - consider appropriateness - open source.

E. Identify relevant benchmarks - consider analysis of technologies. Consider use of relevant benchmarks for the analysis of relevant/provided tools and components.

F. Analyse and report results - for the project exploitation - and possible for DataBench knowledge nuggets (marketing).

These are suggested steps embedded into the search functionalities of the DataBench Toolbox. The search can be done in different ways - as described in the D4.4 DataBench Handbook as well as supported through the Toolbox User Journey support. It is possible to use different user journeys as starting point to help navigating the tool. This is supported also through use of the main two catalogues: the benchmarks catalogue (tools for big data and AI benchmarking), and the knowledge nuggets catalogue (providing information about technical and business aspects related to benchmarking and big data technologies).

3. Analytic Evaluation of the DataBench Toolbox

Existing methodologies for evaluation of software projects and tools mainly consider the process of selecting software products referring to cost, deployment, support, maintenance, and buying a software tool [1]. The few approaches which focus on measurement of the successful implementation of end products do not provide appropriate assessment methods that are applicable to all project types [4]. The reason behind this outcome is that different evaluation methods are better suited for different scenarios and projects [5]. The success of the project is quite often calculated only considering metrics on operational level that assess if the project was delivered on time and on budget.

Currently there is no such generic methodology to fit all various software projects and to be used for the purposes of the evaluation of the DataBench Toolbox. Deliverable D5.3 [8] presented a set of interactive evaluation reports depicting multiple technical metrics and statistics for the different user-roles in the DataBench Toolbox. In contrast to this document, which evaluates the Toolbox platform with respect to usage and utilization, the proposed methodology in this section assesses the Toolbox as a standalone, ready to use software product.

Furthermore, the measurement of the successful implementation of software projects and tools is a complex task, involving analyzing different aspects of the software implementation and considering various user groups [2, 3]. This chapter focuses on the evaluation of the DataBench Toolbox and provides systematic and objective methodology for assessment of the project. As a result, a final judgment statement defining the level of achievement of product requirements, usability, efficiency, and effectiveness is provided.

The following sections present in greater detail the evaluation methodology, which was specifically designed and implemented for the purposes of the DataBench Project. The assessment process considers the project itself and its implementation, as well as end-user satisfaction and overall functionality of the tool. Success criteria, properly set out in measurable terms, are predefined and aligned with the needs and constraints of the project.

3.1. Methodology and Definitions for evaluation of the DataBench Toolbox

This implemented evaluation methodology provides well structured, balanced, and fair measurement of the DataBench software tool. A criteria-based assessment is developed to fulfill a quantitative measurement of the project in terms of sustainability, functionality, maintainability, and usability. The evaluation investigates whether the project conforms to numerous characteristics that are prerequisite for having a sustainable software. It provides a qualitative list of analyzed, structured, and categorized metrics. The approach presents traceability between project requirements, user groups and supporting evidence for the proper evaluation of the DataBench Toolbox. As a result, the applied methodology delivers a single value judgment as an outcome of the assessment process, as well as captures valuable information for further development and improvement of the benchmarking software solution.

Software evaluation is a process whereby the actual outcomes of predefined functional and technical requirements are assessed. The measurement of the DataBench Toolbox is performed against evaluative criteria and accomplishes five consequence steps. *Figure 3* represents these steps, which are discussed in further details in the following subsections.

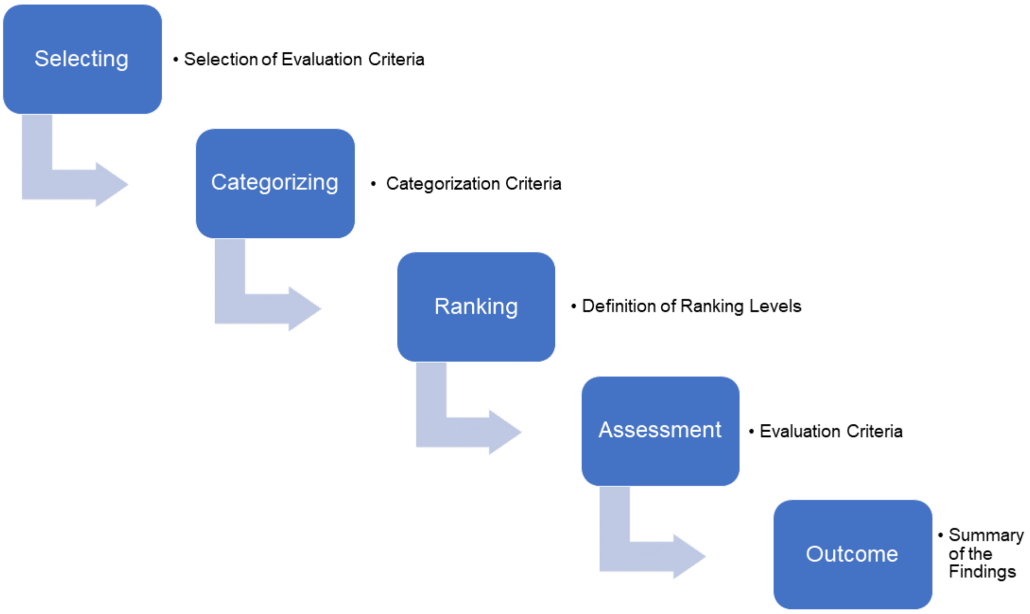

Selecting
• Selection of Evaluation Criteria
Categorizing
• Categorization Criteria
Ranking
• Definition of Ranking Levels
Assessment
• Evaluation Criteria
Outcome
• Summary of the Findings

Figure 3: The process of evaluating the DataBench Toolbox

Selection: The initial step of the evaluation methodology is to specify a broad spectrum of criteria which are important and useful for the assessment of the project. A comprehensive analysis was provided to accurately select the best fitted and most important measures. The research focused and considered questions such as:

- Which criteria should be selected among hundreds of metrics?
- Which criteria are applicable for the purposes of this evaluation?
- Are all user perspectives considered and evaluated?
- What is the importance of each criterion on the overall assessment?
- Are there criteria without which the project cannot function properly?
- What is an appropriate final number of metrics that should be considered in order to accurately evaluate the DataBench Tool?

Categorizing: The goal of this step is to structure and group criteria in a meaningful way suitable for the purposes of this evaluation methodology. Each criterion is carefully analyzed and assigned to the best fitted category. Metrics are refined and examined in greater details during this step and inapplicable metrics are rejected. The following categories are defined:

- *Operational level*: metrics, which measure the overall functionality of the project, are accumulated in this category. Examples of such metrics are:
  - the software is functional;
  - the software does not crash or throw errors when running;
  - the project was completed on time.
- *Identity:* this category examines integrity related metrics such as:

- the project offers functionalities that are not found elsewhere;
- the DataBench Tool is distinguishable from other similar web tools;
- the DataBench software tool has its own domain name.

- *Supportability*: includes criteria, which evaluate the level of support that DataBench Toolbox offers to its consumers. Typical example criteria for this category include:
  - web site has page describing how to get support;
  - project has an e-mail address;
  - web site has site map or index.
- *Copyright:* measures how easy and straightforward it is to see who owns the project and who is responsible for its development.
- *Usability*: among others this category includes the following criteria:
  - high-level description of what the software does is available;
  - high-level description of what/who the software is for is available;
  - website is easy to navigate
- *Learnability*: evaluates in what details the DataBench Toolbox assists its users to learn how to use the software by providing well-structured getting started guides, use cases and/or Frequently Asked Questions section
- *Documentation*: handle metrics such as:
  - the information is presented objectively;
  - the documentation is reliable and free of errors;
  - consists of clear, step-by-step instructions;
  - partitioned into sections for different user groups (business, technical, benchmarking providers).

Ranking: After carefully and accurately selecting and categorizing criteria a ranking procedure on these metrics is applied. Points are given for each ranking based on the importance of each level – the greater the importance, the higher score is given. For the purposes of this methodology the following ranking levels are introduced:

- *Mandatory:* Evaluation criteria that the DataBench Toolbox should accomplish. These are necessary measures without which fulfillment of the DataBench Toolbox should be evaluated as not functional and not ready for public use. Example of such criteria, ranked with 5 points, include:
  - the software is functional;
  - the DataBench Toolbox has its own domain name;
  - it is clearly stated who owns the software's copyright;
  - the software does not crash or throw errors when running.

- *Desired*: Criteria that are nice to have and are very useful. Such criteria increase the *overall* value of the software tool and have a direct impact on the final output of the evaluation judgment. Despite the mandatory criteria, without which the software is evaluated as non-functional, the absence of these criteria leads to lower evaluation of the DataBench Toolbox. The project website is updated on a regular basis; high-level description of what/who the software is for is available; instructions are provided for many basic use cases are some of the criteria ranked as desired and weighted with 3 points.
- *Perspective*: these are features and ideas, or issues that could be considered suitable for further improvement of the DataBench Toolbox. Since these metrics are not mandatory or desired and do not prohibit or affect in any way the correct functionality of the project, they are ranked with 1 point. Examples of such criteria include:
  - instructions are provided supporting all use cases;
  - provides examples of what the user can see in each step e.g. screenshots or command-line excerpts.

Assessment: The DataBench Toolbox is analyzed and evaluated on each criterion. The following well defined weighted scoring approach is applied in order to achieve accurate and fair assessment of the tool:

- *Ranking*: each criterion is weighted based on its importance and impact on the overall performance of the project. These weights are predefined and immutable. As explained in greater details in the Criteria ranking section three levels are defined: Mandatory weighted with 5 points, desired weighted with 3 points and perspective weighted with 1 point
- *Scoring*: The DataBench Toolbox is analyzed against each criterion separately and a corresponding score is assigned. Criteria which are fulfilled and completely implemented are given the highest score of 3 points. Criteria with open issues or not entirely implemented receive a score of 2 points and criteria which are only partially implemented or barely satisfy the criterion are evaluated with 1 point.
- *Analysis of Criteria*: Criteria are analyzed one at a time. During this phase, the scoring process is applied. The weighted measurement for each metric is calculated by multiplying its rank and score. Finally, all results are added together, and a final score is calculated. A maximum score of 351 points could be reached when all criteria are fully satisfied and scored with the maximum number of points. Therefore, the following ranges for measuring the successful implementation of the DataBench Tool are defined:
  - <u>300 points or more</u>: Excellent. The DataBench Toolbox is evaluated as successfully implemented, completely fulfilling both functional and technical requirements
  - <u>200 to 300 points</u>: Average. The tool is functional and achieves its goals. Unfortunately, it possesses weaknesses and has a lot of open issues, bugs and problems that should be resolved in order to deliver a more successful and sustainable software tool.
  - <u>200 points and below</u>: Fail. The project does not reach the ISO/IEC 9126-1 Software engineering — Product quality [6, 7] and therefore is not considered as functional.

Outcome: At this step, all individual results and measurements are gathered, analyzed, and weighted in order to provide a single conclusion of the evaluation process. As a result, a judgment summarizing the outcome of the applied evaluation methodology is provided.

3.2. Evaluation

The underlying evaluation methodology measures the quality of the DataBench Toolbox in various aspects. These areas are derived from ISO/IEC 9126-1 Software engineering — Product quality and include usability, sustainability, and maintainability. The criteria-based assessment evaluates the software tool and checks if it conforms to numerous metrics that every successful project should fulfill.

The proposed mechanism consists of three phases: pre-evaluation, evaluation, and post-evaluation phase. The chosen high level of abstraction enables the introduction of standardization without loss of local flexibility. The three phases are shown on Figure 4 and described as follows:

Pre-evaluation phase
Evaluation phase
Post-evaluation phase

Figure 4: DataBench Toolbox Evaluation phases

- ***Pre-evaluation phase*:** The initial phase of the assessment is to generate a list of criteria against which the DataBench Toolbox will be evaluated.
- ***Evaluation phase*:** The evaluation phase considers the predefined metrics defined in the previous phase and thoroughly evaluates them against the tool.
- ***Post-evaluation phase*:** this phase deals with the definition of a final value judgment and provides a final report of the evaluation outcome.

Table 1 represents in a tabular form the evaluation criteria explicitly designed for the purposes of the evaluation of the DataBench Toolbox. It consists of *Category*, *Criteria* and *Rank* columns which are defined during the pre-evaluation phase and a *Score* column being the outcome of the evaluation phase.

Category	Criteria	Rank	Score
Operational level	The project was completed on time.	3	3
	The project was completed within budget.	3	3

	The project was completed within the agreed scope.	5	3
	The software is functional.	5	1
	The software does not crash or throw errors when running.	5	1
Identity	The DataBench software tool has its own domain name.	5	3
	Project/software has a logo.	3	3
	The project website is updated on a regular basis.	3	3
	All links are current and working properly.	5	1
	The project offers functionalities that are not found elsewhere?	3	3
	The DataBench Tool is distinguishable from other similar sites.	3	3
Supportability	Web site has a page describing how to get support.	3	3
	Project has an e-mail address.	3	3
	Project e-mail address has project domain name.	5	3
	Project has a ticketing system.	1	3
	Website has a sitemap or index.	1	3

	Website has a search facility.	3	3
Copyright	It is clearly stated who owns the software's copyright.	5	3
	Website states who developed the DataBench Tool.	3	3
Usability	High-level description of what/who the software is for is available.	3	3
	High-level description of what the software does is available.	3	3
	High-level description of how the software works is available.	3	2
	The project's website is easy to navigate.	3	2
	Architectural overview, with diagrams, is available.	2	3
Learnability	A getting started guide is provided outlining a basic example of using the DataBench software tool.	3	3
	Instructions are provided for many basic use cases.	3	3
	Instructions are provided supporting all use cases.	1	2
	Provides a Frequently asked questions page.	2	3

Documentation	Provides a high-level overview of the software.	3	3
	Partitioned into sections for different user groups (business, technical, benchmarking providers)	3	3
	Does not require user background and expertise (for each class of user).	3	3
	Consists of clear, step-by-step instructions.	3	2
	Provides examples of what the user can see at each step e.g. screenshots or command-line excerpts.	1	2
	Is on the project web site.	3	3
	The documentation is reliable and free of error.	5	3
	The information is presented objectively.	5	3
	Lists resources for further information.	1	3

Table 1: Evaluation criteria organized by categories

3.3. Outcome

The purpose of this evaluation process is to generate an accurate assessment of the DataBench Toolbox which is the main result of the DataBench project. By applying the proposed methodology, it was determined that all mandatory criteria are met and almost all of them are given the highest score possible. The fulfillment of all mandatory requirements evaluates the project as functional and ready to be released for public use.

The outcome of the criteria-based assessment is calculated to be equal to 310 points out of 351. Based on the scoring criteria introduced earlier in this chapter, it is determined that the Toolbox is evaluated as ***Excellent***. A precise evaluation for each category was also provided - six out of all seven categories achieved more than 85% of all possible points for its category.

One of the critical evaluation categories, namely the Operation level category, was evaluated with 43 out of 63 points, or 68% successful rate. Although the overall assessment of the project is determined to be Excellent a deeper analysis of the outcome for this category was provided and discussed in further details in the following section 4. In contrast, two of the categories, namely Usability and Copyright received the perfect score, meaning that the project completely satisfies the derived requirements for these categories from the ISO/IEC 9126-1 Software engineering standards. The project received 96% and 95% of all possible points in the Learnability and Documentation categories respectively, pointing out that new users can easily and intuitively start using the DataBench Toolbox nevertheless they have business, technical or benchmarking background. In the Usability category the project was evaluated with 36 out of 42 possible points, thus implying that there exist some improvements possibilities for this category. The same can be seen for the Identity category.

The final score of 85% gathered from all categories indicates a well-structured, fully functional, and successfully implemented software product. The project implements both technical and functional requirements and thus satisfies the ISO/IEC 9126-1 Software engineering — Product quality norms in terms of usability, sustainability, and maintainability.

3.4. Features and further improvement possibilities

During the evaluation of the DataBench Toolbox not only a narrow examination of the web tool was implemented, but also improvement possibilities were detected and proposed. It was analyzed that two of the metrics from the Operational Level Category, namely “The software is functional”, and “The software does not crash or throw errors when running”, received the minimum score. It is investigated that in a few places the software throws Null Pointer exception or does not launch at all. Since these issues do not fully stop the DataBench Toolbox to function, but only prohibit the execution of some functionalities it was decided that the project will not be evaluated as non-functional.

Additional improvements possibilities were proposed for the navigation of the web tool, where more straightforward and accurate menu navigation could be achieved. Furthermore, a detailed step by step use case of how to use the DataBench Toolbox and further execution examples for all three user groups (business, technical and benchmarking provider) are recommended and taken into consideration.

3.5. Concluding remarks

The DataBench Toolbox is a powerful tool that enables users to investigate, select and evaluate Big Data benchmarking tools. It offers a wide catalogue with both business and technical benchmarking, considers users with various backgrounds (business, technical and benchmarking providers) and gives insights for both business relevance (such as organizations KPIs) and technical aspects (such as performance and latency metrics). Being developed for several years now, the project provides possibilities to determine optimal Big Data benchmarking approaches.

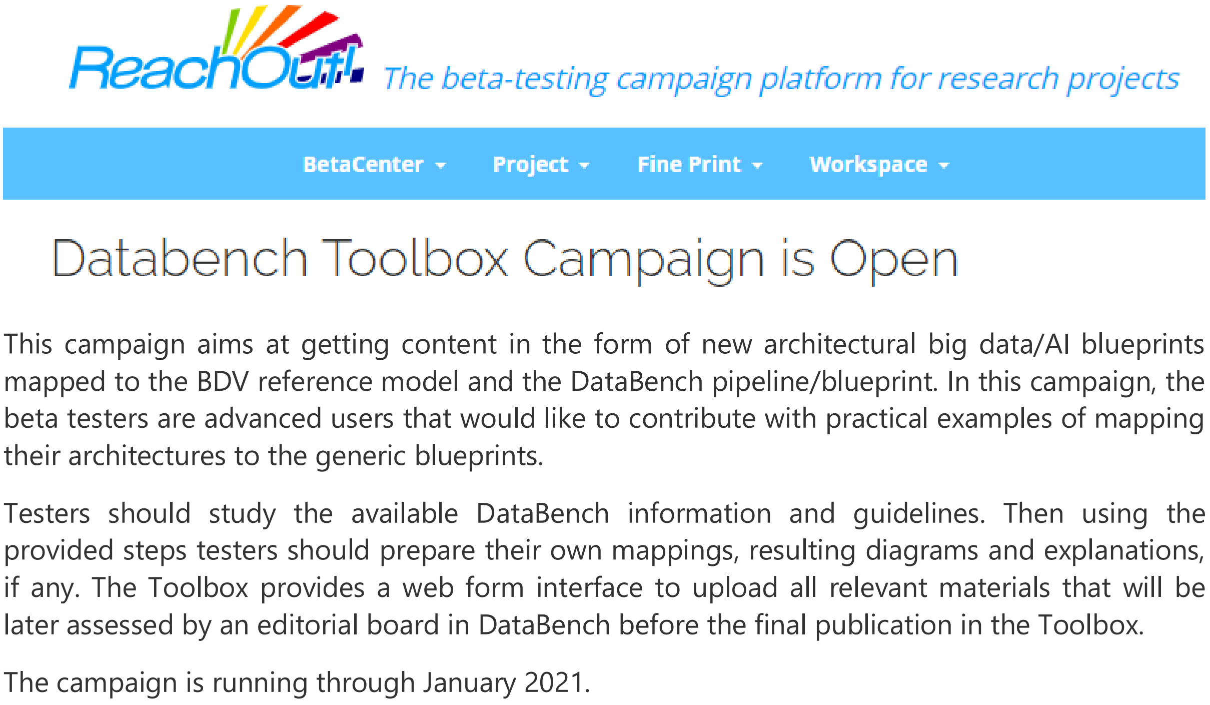
The outcome of the methodology applied to evaluate the DataBench Toolbox reflects a good overall score. A comprehensive and in-depth analysis, developed especially for the purposes of the assessment of the toolbox, was implemented and applied. First, a wide range of evaluation metrics, which investigate both functional and technical aspects were thoroughly gathered and organized into categories. Second, a ranking mechanism was applied, whereas criteria are given one of the three levels of importance – mandatory, desired, and perspective.

Accurate and fair assessment of the tool was guaranteed by implementing a scoring system, which examines and evaluates each criterion separately. Finally, a conclusion of the applied methodology was provided. By achieving a relatively high score for each category, the outcome of the assessment process evaluated the DataBench Toolbox in total in the highest category of Excellent. It is concluded that the toolbox was developed as a sustainable software, which conforms the ISO/IEC 9126-1 Software engineering — Product quality software requirements for usability, sustainability, and maintainability.

4. Project usage – DataBench Toolbox Campaign

4.1. ReachOut Campaign – DataBench Toolbox

During the fall of 2020 the DataBench project has initiated a beta test of the DataBench Toolbox as a ReachOut campaign[1] (open until end of January 2021).



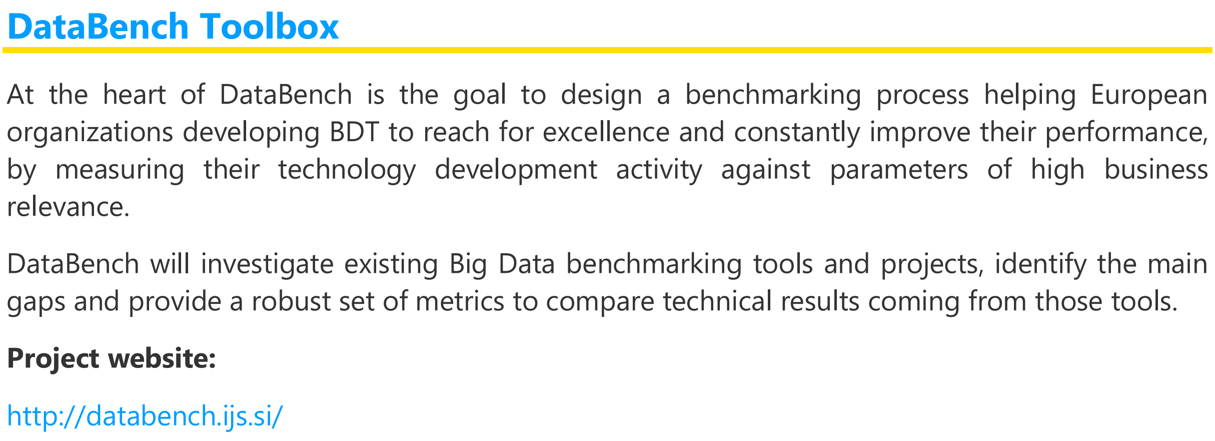
Databench Toolbox Campaign is Open

This campaign aims at getting content in the form of new architectural big data/AI blueprints mapped to the BDV reference model and the DataBench pipeline/blueprint. In this campaign, the beta testers are advanced users that would like to contribute with practical examples of mapping their architectures to the generic blueprints.

Testers should study the available DataBench information and guidelines. Then using the provided steps testers should prepare their own mappings, resulting diagrams and explanations, if any. The Toolbox provides a web form interface to upload all relevant materials that will be later assessed by an editorial board in DataBench before the final publication in the Toolbox.

The campaign is running through January 2021.

The campaign motivates for the use of the DataBench Toolbox



DataBench Toolbox

At the heart of DataBench is the goal to design a benchmarking process helping European organizations developing BDT to reach for excellence and constantly improve their performance, by measuring their technology development activity against parameters of high business relevance.

DataBench will investigate existing Big Data benchmarking tools and projects, identify the main gaps and provide a robust set of metrics to compare technical results coming from those tools.

Project website:

http://databench.ijs.si/

[1] https://www.reachout-project.eu/view/Events/Databench_Toolbox

The campaign contains two sub-campaigns:

ReachOut Campaign 1 – **Generation of architectural Pipelines-Blueprints**

This campaign aims at getting content in the form of new architectural big data/AI blueprints mapped to the BDV reference model and the DataBench pipeline/blueprint. In this campaign we focus mainly on advanced users that would like to contribute with practical examples of mapping their architectures to the generic blueprints. The results will be published in the DataBench Toolbox acknowledging the ownership and can be used by the owners for their own purposes in their projects/organizations to claim their efforts in mapping with existing standardization efforts in the community. *This campaign is described in more detail in Annex A*

ReachOut Campaign 2 – **Benchmarks for technical and business users**

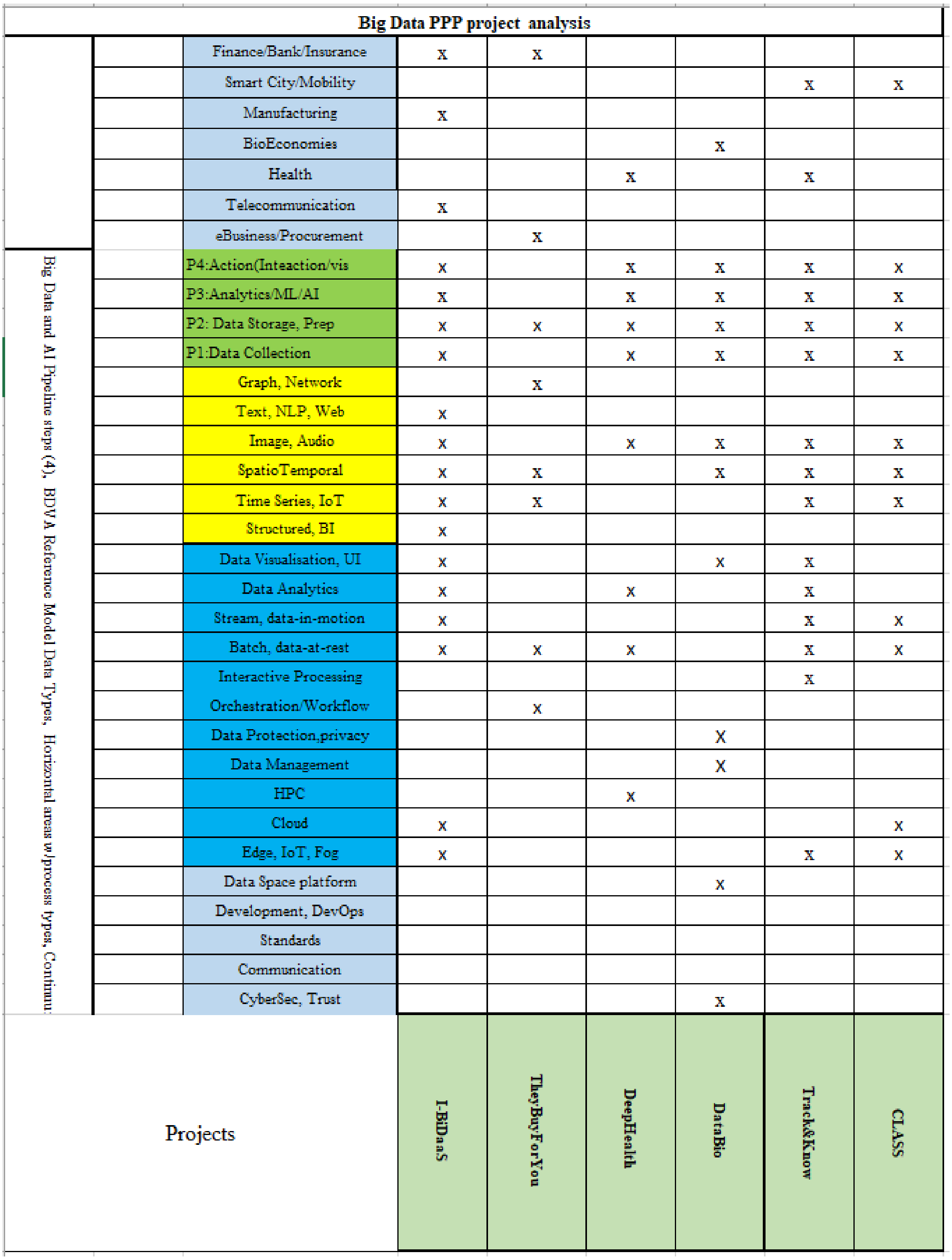
This campaign aims at getting feedback of the usage of the Tool and the user interface of the web front-end of the Toolbox. The Toolbox provides a set of user journeys, or suggestions, for three kind of users: 1) Technical user (people interested in technical benchmarking), 2) Business users (interested in finding facts, tools, examples and solutions to make business choices), and 3) Benchmark providers (users from benchmarking communities or that generated their own benchmarks). In this campaign we focus mainly on technical and business users. We provide some minimal instructions for these two types of users to understand if finding information in the Toolbox is not a cumbersome process and getting your feedback. The idea is to use the user journeys drafted in the Toolbox to drive this search process and understand if users find this information enough to kick-start the process of finding the right benchmark and knowledge they were looking for. *This campaign is described in more detail in Annex B*

Annex C contains the template for the campaign feedback questionnaire.

The following sections present project based evaluation of the DataBench Toolbox, based on the two campaign areas of 1) Pipelines and Blueprints and 2) Technical Benchmarks.
This is with a focus on how projects can use Pipelines and Blueprints for the analysis of potential Big Data and AI technologies and corresponding benchmark and also how project experiences with their pipelines and blueprints can be reported and shared as knowledge nuggets in the DataBench Toolbox.

This has been supported by two campaigns through the ReachOut beta testing system: 1) Generation of Architectural Pipelines – Blueprints and 2) Finding the right benchmarks for technical and business users.

The results from 6 example Big Data PPP projects based on this is reported. This is reported with the common structure of Project description, Project Objectives and Architecture, Mapping to DataBench Pipeline and Blueprints and Relevant Benchmarks and Nuggets from DataBench Toolbox. The reported projects are I-BiDaaS, TheyBuyForYou, DeepHealth, DataBio, Track&Know and CLASS. Further projects are in progress as part of the sustainability strategy of DataBench.



Big Data PPP project analysis								
		Finance/Bank/Insurance	x	x				
		Smart City/Mobility					x	x
		Manufacturing	x					
		BioEconomies				x		
		Health			x		x	
		Telecommunication	x					
		eBusiness/Procurement		x				
Big Data and AI Pipeline steps (4), BDVA Reference Model Data Types, Horizontal areas w/process types, Continuu		P4:Action(Inteaction/vis	x		x	x	x	x
		P3:Analytics/ML/AI	x		x	x	x	x
		P2: Data Storage, Prep	x	x	x	x	x	x
		P1:Data Collection	x		x	x	x	x
		Graph, Network		x				
		Text, NLP, Web	x					
		Image, Audio	x		x	x	x	x
		SpatioTemporal	x	x		x	x	x
		Time Series, IoT	x	x			x	x
		Structured, BI	x					
		Data Visualisation, UI	x			x	x	
		Data Analytics	x		x		x	
		Stream, data-in-motion	x				x	x
		Batch, data-at-rest	x	x	x		x	x
		Interactive Processing					x	
		Orchestration/Workflow		x				
		Data Protection,privacy				X		
		Data Management				X		
		HPC			x			
		Cloud	x					x
		Edge, IoT, Fog	x				x	x
		Data Space platform				x		
		Development, DevOps						
		Standards						
		Communication						
		CyberSec, Trust				x		
		Projects	I-BiDaaS	TheyBuyForYou	DeepHealth	DataBio	Track&Know	CLASS

Table 2 Big Data PPP projects mapped into the different areas of the DataBench Framework

Table 2 Big Data PPP project shows the six projects described in the following sections, mapped into the different areas in the DataBench Framework.

4.2. I-BiDaaS

4.2.1. Project description

I-BiDaaS[2] **(2018-2020) - Industrial-Driven Big Data as a Self-Service Solution** is an EU-funded Big Data PPP project that aims to empower IT and non-IT big data experts to easily utilize and interact with big data technologies. I-BiDaaS is proposing a unified solution that significantly increases the speed of data analysis and facilitates cross-domain dataflow towards a thriving data-driven EU economy.

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

I-BiDaaS will achieve its goals following a methodical approach. As a first step, it has guaranteed access to real-world industry big data.

I-BiDaaS will proceed with breaking inter and intra-sectorial data-silos, and support data sharing, exchange, and interoperability. Having done so, it will support methodical big data experimentation by putting in place a safe data processing environment.

Vision

I-BiDaaS is a self-service solution, aiming to empower users to easily utilize and interact with big data technologies by designing, building and demonstrating a unified framework that significantly increases the speed of data analysis while coping with the rate of data asset growth and facilitates cross-domain data-flow towards a thriving data-driven EU economy.

I-BiDaaS will shift the power balance within an organisation, increase efficiency, reduce costs, improve employee empowerment and increase profitability. Moreover, I-BiDaaS will deliver a full array of big data business analytics solutions for structured, unstructured, noisy and potentially synthetic data for companies in multiple industries that are more accessible, cost effective and employee-empowering than existing solutions, which gives companies the confidence to deploy Big Data Self-Service solutions across the organisation, from consumer-facing employees with little IT experience or expertise to top management and helps companies to optimize decision-making at the tactical, operational and strategic levels.

The domains that can exploit such self-service solutions are numerous; I-BiDaaS will explore three critical ones with significant challenges and requirements: banking, manufacturing, and telecommunications.

4.2.2. Project objectives and Architecture

The I-BiDaaS Objectives are:

- To develop, validate, demonstrate and support, a complete and solid big data solution that can be easily configured and adopted by practitioners.
- To break inter- and intra-sectorial data-silos, create a data market, offer new business opportunities and support data sharing, exchange as well as interoperability.

[2] https://www.ibidaas.eu/

- To construct a safe environment for methodological big data experimentation for the development of new products, services, tools.
- To develop data processing tool and techniques applicable in real-world settings and demonstrate significant increase of speed of data throughput and access.
- To develop technologies that will increase the efficiency and competitiveness of all EU companies and organisations that need to manage vast and complex amounts of data.

The detailed description of the I-BiDaaS architecture (the six workflow steps together with layered system and User interface) is presented below:

Step 1: The data (real or synthetic produced via data fabrication tool) are ingested into the batch processing and the streaming analytics modules via the Universal Message broker.

Step 2: The analytic modules perform analytics on the ingested streaming data, also referencing historic information where necessary, to identify business patterns that have happened or are about to happen.

Step 3 (I-BiDaaS innovation): The batch and real time analytic results are fed to the advanced visualization tools. An innovation is that part of the analytics can be offloaded to the parallel GPU-accelerated engine to further speed-up the execution of streaming analytics.

Step 4: The collected data can be stored in Hecuba, that uses the Apache Cassandra as a back-end, which can then be processed by the COMPSs pool of distributed machine learning algorithms.

Step 5 (I-BiDaaS innovation): The correlations produced by the analysis can fed back to the data fabrication platform, to be used for training and help building rules that will be used for future data generation purposes.

Step 6 (I-BiDaaS innovation): The real-time processing module feeds the batch-processing module with inputs that enable periodic refinements of models used in machine learning methods. The proposed solution goes beyond the traditional lamda architecture in terms of interleaving batch and stream processing.

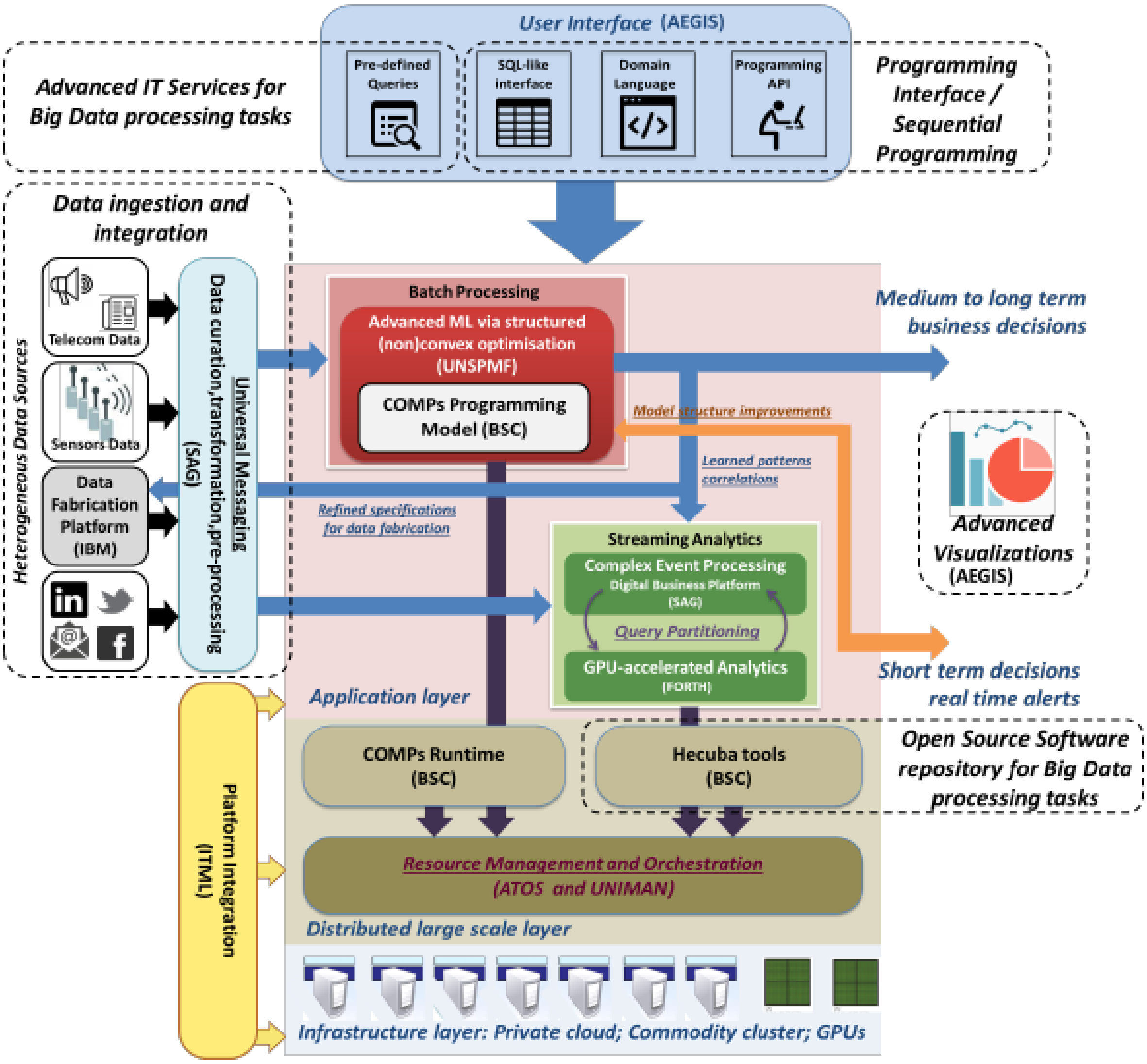

User Interface (AEGIS)
Advanced IT Services for Big Data processing tasks
Pre-defined Queries
SQL-like interface
Domain Language
Programming API
Programming Interface / Sequential Programming
Data ingestion and integration
Heterogeneous Data Sources
Telecom Data
Sensors Data
Data Fabrication Platform (IBM)
Data curation,transformation,pre-processing
Universal Messaging (SAG)
Batch Processing
Advanced ML via structured (non)convex optimisation (UNSPMF)
COMPs Programming Model (BSC)
Medium to long term business decisions
Model structure improvements
Learned patterns correlations
Refined specifications for data fabrication
Advanced Visualizations (AEGIS)
Streaming Analytics
Complex Event Processing
Digital Business Platform (SAG)
Query Partitioning
GPU-accelerated Analytics (FORTH)
Short term decisions real time alerts
Application layer
COMPs Runtime (BSC)
Hecuba tools (BSC)
Open Source Software repository for Big Data processing tasks
Platform Integration (ITML)
Resource Management and Orchestration (ATOS and UNIMAN)
Distributed large scale layer
Infrastructure layer: Private cloud; Commodity cluster; GPUs

Figure 5: I-BiDaaS project architecture

I-BiDaaS User Interface: The proposed solution offers also a multi-purpose interface which can be used by different categories of users. It provides different levels of abstractions such as **Programming API** (providing access to every level of our software stack); **Domain language** (providing access to the application layer); **Pre-defined analytics** (providing simplest form to non-IT users to easily combine and multiplex with the desired data sources, to form a data processing pipeline).

I-BiDaaS layered system: The I-BiDaaS architecture composes of three principal layers: **the infrastructure layer** (providing the actual underlying storage and processing infrastructure of the I-BiDaaS solution); **the distributed large-scale layer** (controlling the orchestration and management of the underlying physical computational and storage infrastructure) and **application layer** (referring to the architecture aspects and components that are involved in the actual workflow of extracting actionable knowledge from the big data, starting from data preparation, to analytics, to delivering analytics results for supporting decision making).

4.2.3. Mapping to DataBench Pipeline and Blueprints

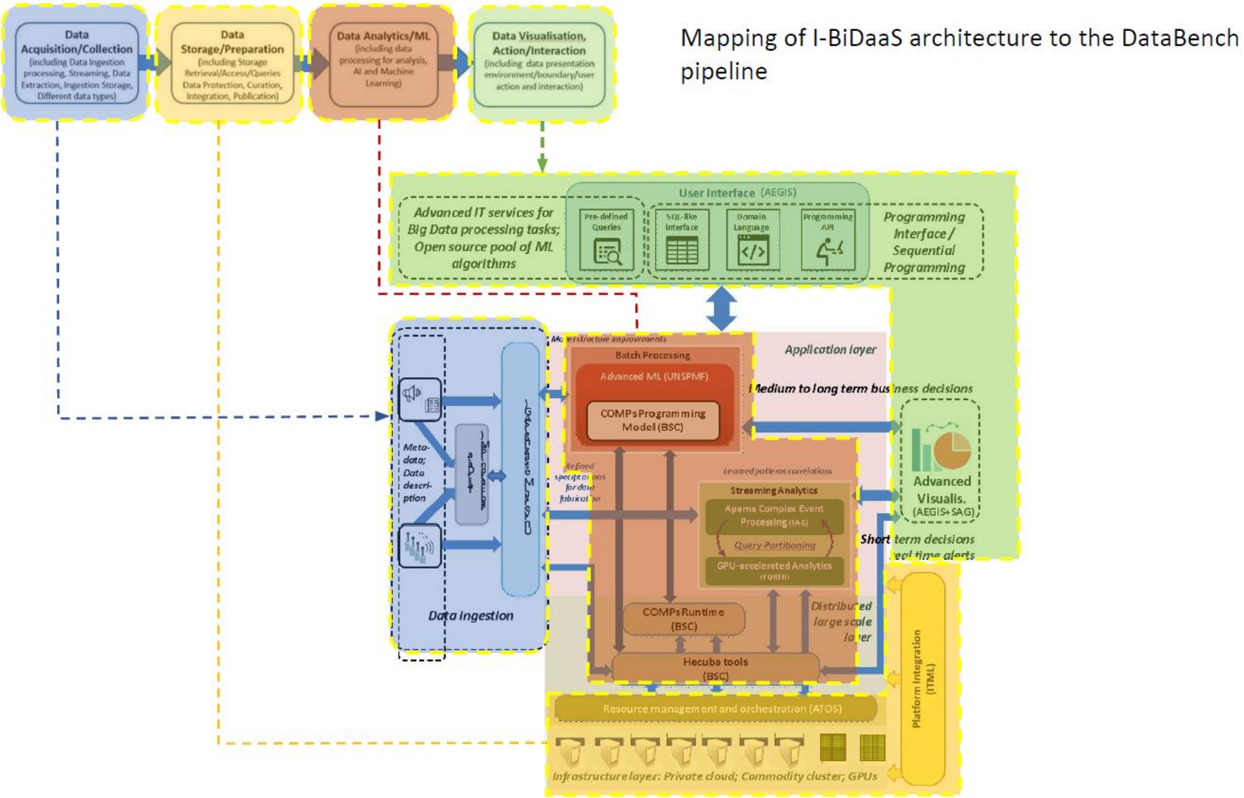
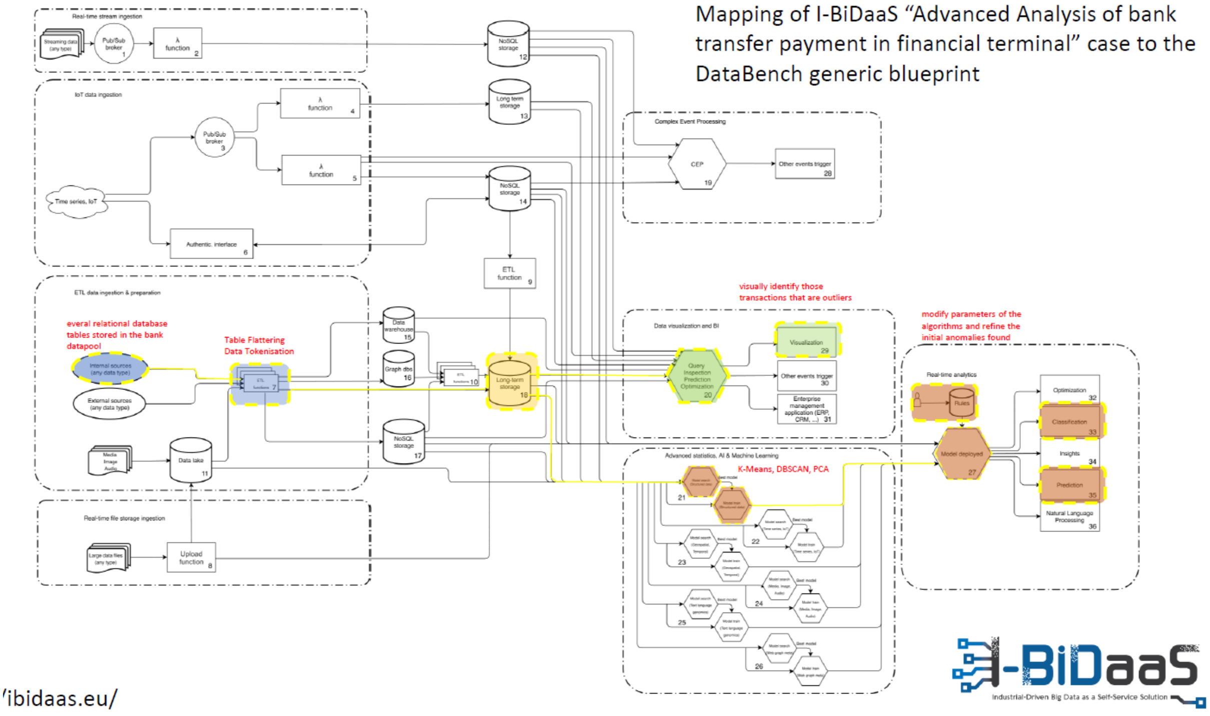

Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)
Data Visualisation, Action/Interaction
(including data presentation environment/boundary/user action and interaction)
Mapping of I-BiDaaS architecture to the DataBench pipeline
User Interface (AEGIS)
Advanced IT services for Big Data processing tasks; Open source pool of ML algorithms
Pre-defined Queries
SQL-like interface
Domain Language
Programming API
Programming Interface / Sequential Programming
Application layer
Batch Processing
Advanced ML (UNSPMF)
COMPs Programming Model (BSC)
Medium to long term business decisions
Advanced Visualis. (AEGIS+SAG)
Meta-data; Data description
Refined specifications for data fabrication
Learned patterns correlations
Streaming Analytics
Apama Complex Event Processing (SAG)
Query Partitioning
GPU-accelerated Analytics (FORTH)
Short term decisions real time alerts
COMPs Runtime (BSC)
Distributed large scale layer
Hecuba tools (BSC)
Data ingestion
Resource management and orchestration (ATOS)
Infrastructure layer: Private cloud; Commodity cluster; GPUs
Platform Integration (ITML)

Figure 6: Mapping of I-BiDaaS architecture to the DataBench pipeline

I-BiDaaS supports Big Data as a self-service and provides a safe experimentation environment for the methodical development of new Big Data products, services, and tools. The I-BiDaaS system has been applied in a number of industrial use cases in the telecommunication, manufacturing and banking sectors. This blueprint (a) presents the I-BiDaaS system architecture, (b) links this architecture to the 4 steps of the DataBench generic pipeline and (c) particularises the DataBench Generic Big Data Analytics Blueprint, in the context of one of the I-BiDaaS use cases in the financial sector, namely, 'Advanced Analysis of bank transfer payment in financial terminal'.



Mapping of I-BiDaaS “Advanced Analysis of bank transfer payment in financial terminal” case to the DataBench generic blueprint

Real-time stream ingestion
Streaming data (any type)
Pub/Sub broker 1
λ function 2
NoSQL storage 12
IoT data ingestion
λ function 4
Long term storage 13
Pub/Sub broker 3
λ function 5
NoSQL storage 14
Time series, IoT
Authentic. interface 6
ETL function 9
ETL data ingestion & preparation
everal relational database tables stored in the bank datapool
Table Flattering Data Tokenisation
Internal sources (any data type)
External sources (any data type)
ETL functions 7
Data warehouse 15
Graph dbs 16
ETL functions 10
Long-term storage 18
Media Image Audio
Data lake 11
NoSQL storage 17
Real-time file storage ingestion
Large data files (any type)
Upload function 8
Complex Event Processing
CEP 19
Other events trigger 28
visually identify those transactions that are outliers
Data visualization and BI
Query Inspection Prediction Optimization 20
Visualization 29
Other events trigger 30
Enterprise management application (ERP, CRM, ...) 31
modify parameters of the algorithms and refine the initial anomalies found
Real-time analytics
Rules
Model deployed 27
Optimization 32
Classification 33
Insights 34
Prediction 35
Natural Language Processing 36
Advanced statistics, AI & Machine Learning
K-Means, DBSCAN, PCA
Best model
21
22
23
24
25
26
'ibidaas.eu/
I-BiDaaS
Industrial-Driven Big Data as a Self-Service Solution

Figure 7: Mapping of I-BiDaaS “Advanced Analysis of bank transfer” case to the DataBench blueprint

The above mapping to the DataBench generic blueprint has been provided as a Knowledge Nugget in the DataBench Toolbox.

4.2.4. Relevant Benchmarks and Nuggets from DataBench Toolbox

This section provides suggestions for relevant benchmarks and knowledge nuggets that include Big Data and AI technologies, architectural patterns and blueprints as well as business benchmarks for comparison. The following tables represent multiple searches that were performed with the current version of the DataBench Toolbox and illustrate its current results. The filtering criteria applied in the searches are based on the project descriptions from the previous sections.

Table 3 shows the top 10 technical benchmarks and top 5 knowledge nuggets resulting from a BDV Reference Model Search for all of the six different data types as all of them are relevant for the I-BiDaaS platform and services.

Go to Search --> BDV Reference Model and select one of the Data Types:					
Structured Data/ Business Intelligence	Time series, IoT	Geo Spatial Temporal	Media Image Audio	Text, Language, Genomics	Web Graph Meta

The search returns the following list of benchmarks (only the top 10):

BigBench V2	owperf (CLASS)	IDEBench	AIBench	BigBench V2	HiBench
HiBench	Yahoo Streaming Benchmark (YSB)	MLPerf	AIMatrix	HiBench	Berlin SPARQL Benchmark (BSBM)
Yahoo Streaming Benchmark (YSB)	AIoTBench	Training Benchmark for DNNs (TBD)	BigDataBench	AIBench	BigFrame
Yahoo! Cloud Serving Benchmark (YCSB)	BenchIoT		CloudSuite	AIMatrix	CloudRank-D
AMP Lab Big Data Benchmark	CityBench		DAWNBench	BigDataBench	gMark
BigDataBench	CloudRank-D		Deep Learning Benchmarking Suite (DLBS)	CloudRank-D	Graphalytics
BigFrame	CloudSuite		DeepMark (Convnet)	CloudSuite	Hobbit Benchmark
BigFUN	DeepMark (Convnet)		Edge AI Bench	DAWNBench	LinkBench
CALDA	Edge AI Bench		Fathom	DeepMark (Convnet)	LIQUID
CloudRank-D	HERMIT		HPC AI500	Edge AI Bench	MiDBench

The search returns the following list of knowledge nuggets (only the top 5):

NoSQL - Key-Value DB	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Linked Data Integration and Publication Pipeline pattern
Benchmarks features matrix	Data features Star Diagram	Data features Star Diagram	Computer Vision	Data features Star Diagram	Benchmarks features matrix
Data features Star Diagram	IoT Ingestion and Authentication	Earth Observation and Geospatial Pipeline pattern	Data features Star Diagram	DataBench Generic Data Pipeline (4 steps data value chain)	Data features Star Diagram
Data Lake	IoT Pipeline pattern	Technical benchmarks Star Diagram	Technical benchmarks Star Diagram	Natural Language Processing (NLP)	Linked Data Benchmark Council (LDBC)
Data Warehouse	Technical benchmarks Star Diagram	Transaction Processing Performance Council		Technical benchmarks Star Diagram	NoSQL - Graph DB

Table 3: Results from BDV Reference Model Search

The first four columns of Table 3 present the top 10 technical benchmarks and top 5 knowledge nuggets obtained from a BDV Reference Model search based on the Horizontal Layers relevant for the I-BiDaaS architecture and functionalities. The last two columns are results from a Guided Benchmark search filtering the two most relevant data processing types: stream (data-in-motion) and batch (data-at-rest).

Go to Search --> BDV Reference Model and select one of the Horizontal Layers:				**Go to Search --> Guided Benchmark Search and select the Platform and Architecture Features drop-down menu. Then click on the following tags:**	
Data Visualization and User Interaction	Data Analytics	Cloud and High Performance computing (HPC)	Things/Assets, Sensors and Actuators (Edge, IoT, CPS)	Stream (data-in-motion)	Batch (data-at-rest)
The search returns the following list of benchmarks (only the top 10):				**The search returns the following list of benchmarks (only the top 10):**	
ABench	BigBench V2	CloudSuite	owperf (CLASS)	BigBench V2	BigBench V2

AIBench	HiBench	GARDENIA	AIoTBench	HiBench	HiBench
Hobbit Benchmark	ABench	HPC AI500	BenchIoT	Yahoo Streaming Benchmark (YSB)	ABench
IDEBench	AdBench	MLBench Services	CloudSuite	ABench	AMP Lab Big Data Benchmark
NNBench-X	AIBench	PUMA Benchmark Suite	Edge AI Bench	AIM Benchmark	Berlin SPARQL Benchmark (BSBM)
Penn Machine Learning Benchmark (PMLB)	AIM Benchmark	TPCx-V	HERMIT	BigDataBench	BigBench
VisualRoad	AIMatrix		IoTABench	CityBench	BigDataBench
	AIoTBench		IoTBench	CloudRank-D	CALDA
	ALOJA		Linear Road	CloudSuite	CloudRank-D
	Benchip		MiDBench	DAWNBench	CloudSuite
The search returns the following list of knowledge nuggets (only the top 5):				**The search returns the following list of knowledge nuggets (only the top 5):**	
Benchmarks features matrix	Agriculture Architectural Blueprints	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix
Data Visualization tools	AI&ML Developmemt Platforms (IDEs)	Technical benchmarks Star Diagram	IoT Ingestion and Authentication	Classification of data processing architectures	Classification of data processing architectures
Search engines	AI&ML Frameworks	The DataBench Framework and the BDV Reference Model	IoT Pipeline pattern	Data features Star Diagram	Data features Star Diagram
Use case independent blueprints - Data Processing and Exploitation Systems	AI&ML Libraries			Message brokers and Pub-Sub	Project ALOJA
Use case independent blueprints - Data processing and exploitation systems - Data visualization and business intelligence architecture	AI&ML Platforms			Stream Processing engines	

Table 4: Results from BDV Reference Model Search by Horizontal Layers and Guided Benchmark Search filtering by Platform and Architecture Features

Finally, Table 4 illustrates the top 5 most relevant knowledge nuggets resulting from a Guided Benchmark search selecting Business Features for three key industry domains in I-BiDaaS: banking, manufacturing and telecommunications.

Go to Search --> Guided Benchmark Search and select the Business Features drop-down menu. Then click on the following tags:		
Finance/ Bank/ Insurance	Manufacturing process/ discrete	Telecommunication
The search returns the following list of knowledge nuggets (only the top 5 most relevant):		
Benchmarks features matrix	Evidence of business process performance - Production quality (Manufacturing)	Practical example of creating a blueprint and derived cost-effectiveness analysis: Targeting the Telecommunications Industry
Financial services - Fraud detection Architectural Blueprints	Evidence of business process performance - Service innovation (Manufacturing)	Qualitative Benchmarks for the Telecom and Media Industry
Qualitative Benchmarks for the Financial Services Industry	Qualitative Benchmarks for the Manufacturing Industry	Quantitative Benchmarks for the Telecom and Media Industry
Qualitative Benchmarks by Industry	Quantitative Benchmarks for the Manufacturing Industry	Telecommunication - Churn prediction and custom promotions Blueprint
Securities Technology Analysis Center (STAC)	Quantitative Benchmarks for the Top 3 Use Cases in Manufacturing	Telecommunications - Network resource and capacity optimization Architectural Blueprints

Table 5: Results from Guided Benchmark Search filtering by Business Features

Based on the DataBench Toolbox search results listed in the above tables and performing a cross-selections using the main filtering criteria, we summarized a list of the following technical benchmarks classified in the three business industries:

Finance

- Securities Technology Analysis Center (STAC) [9]

Telecommunication

- AIM Benchmark [10]
- Edge AIBench [11]
- AIBench [12]
- Semantic Publishing Benchmark (SPB) [13]

Manufacturing

- Senska [14]
- TPCx-IoT [15]
- MiDBench [16]

4.3. TheyBuyForYou (TBFY)

4.3.1. Project description

The TheyBuyForYou (TBFY) project[3] (2018-2020) has developed a Knowledge Graph (KG) based platform and end-user tools for integrating and reconciling procurement and company data from distributed data sources, including analytics tools such as anomaly detection and cross-lingual search. The figure below illustrates the process of creating the KG.

- Procurement and company data underlying the KG is provided by two main data providers: OpenOpps[4] for procurement data (e.g., tenders and contracts) and OpenCorporates[5] for supplier data (i.e., companies). OpenOpps has gathered over three million tender documents from more than 685 publishers through Web scraping and by using open APIs, while OpenCorporates currently has 140 million entities collected from national registers. For the period January 2019 until December 2020 the data sources counted more than 1.9 million JSON files totaling 17 GB in size.

- We integrated the two high-quality data sets according to an ontology network to form a Knowledge Graph. The ontology network includes an ontology for representing procurement data based on Open Contracting Data Standard (OCDS), namely the OCDS ontology [17], and another ontology for representing company data, namely the euBusinessGraph ontology [18]. Mapping the JSON source data to RDF format according to the ontology is done through daily batch processing that on average processes around 3000 files per day. The resulting RDF data is published to a triple store. As of December 2020, the KG currently contains more than 147 million triples.

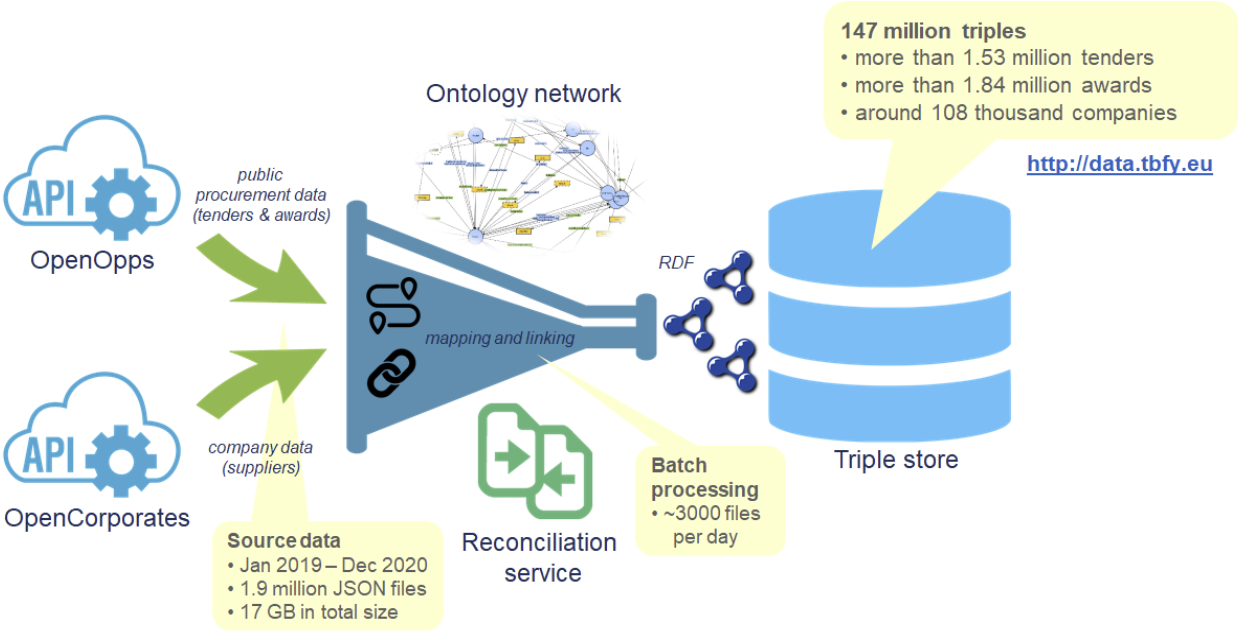


Figure 8: Creating the TBFY Knowledge Graph

[3] https://theybuyforyou.eu

[4] https://openopps.com

[5] https://opencorporates.com

4.3.2. Project objectives and Architecture

Application area and domain

Business / IT services for public procurement.

Business objectives

TheyBuyForYou will explore how procurement and public spending data, paired with data management, analytics, and interaction design, could be used to innovate four key areas:

1. **Economic development:** Delivering better economic outcomes from public spending, in particular for SMEs. SMEs should be able to get better access to public tenders, competing with more established players.
2. **Demand management:** Spotting trends in spending and supplier management to achieve long term goals such as cost savings and efficiency gains.
3. **Competitive markets:** Identifying areas for cost cuts through healthier competition.
4. **Procurement intelligence:** Producing advanced analytics to inform decision support, risk monitoring and supply market analysis for procurement and purchasing managers.

Key technical objective - The key technical objective of TBFY is to build a technology platform, consisting of a set of modular, Web-based services and APIs to publish, curate, integrate, analyse, and visualize a comprehensive, cross-border and cross-lingual procurement knowledge graph, including public spending and corporate data from multiple sources across the EU.

4.3.3. Mapping to DataBench Pipeline and Blueprints

Three architectural aspects of the TBFY platform were mapped to the four steps of the DataBench Generic Data Pipeline [19]:

1. The **TBFY platform architecture**, which can be used as a blueprint for how to build a Knowledge Graph (KG) based platform.
2. The **TBFY data ingestion pipeline**, which can be used as a blueprint for how to ingest JSON data files through an REST API, reconcile/extend the JSON data, map the JSON data to RDF according to an ontology, and finally publish the data as a Knowledge Graph (KG).
3. **Value-added services built on top of the TBFY platform**, which can be seen as example tools and applications that can be built using the KG data.

TBFY platform architecture

In TBFY we have been working on a layered architecture [20] of data services, ontologies, core APIs and tools that allows different levels of access and use of our procurement knowledge graph. A layer-based architecture allows separating the different services so that most of the interaction occurs only between adjacent layers and any change in a technology does not affect the rest of the services. The TBFY platform architecture shown in the figure below covers the Data Acquisition/Collection and Data Storage/Preparation. On the left-hand side, we include the ETL processes that are being used to incorporate the data sources into the KG. On the right-hand side we provide an overview of the main data storage mechanisms, including a triple store for the generated RDF-based data and a document store for the documents associated to public procurement (tender notices, award notices, etc.), whose URLs are accessible via specific properties of the KG (using rdfs:seeAlso). For those specific cases where a URI is also available in the original data sources (from OpenOpps and OpenCorporates), such URI is provided in the KG using a statement with owl:sameAs. The KG is accessible via a core REST API. The API catalogue is mostly focused on providing access mechanisms to those who want to make use of the knowledge graph, particularly software developers. Therefore, they are mostly focused on providing access to the KG. The API catalogue is organised around the main entities that are relevant for public procurement, such as contracting processes, awards, and contracts. Since the KG is stored as RDF in a triple store, there is also a SPARQL endpoint for executing ad-hoc queries. Finally, the platform provides business cases with a set of value-added and analytics services that uses and analyses the data from the knowledge graph.

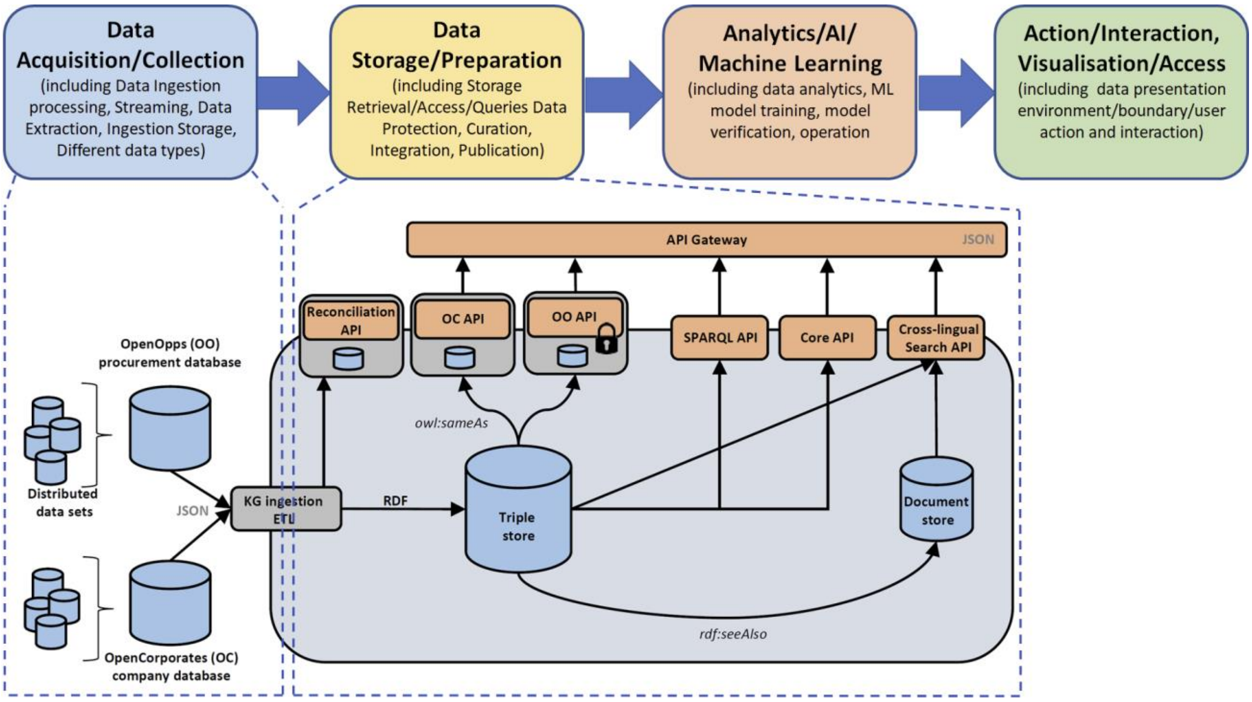

Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Analytics/AI/ Machine Learning
(including data analytics, ML model training, model verification, operation
Action/Interaction, Visualisation/Access
(including data presentation environment/boundary/user action and interaction)
API Gateway
JSON
Reconciliation API
OC API
OO API
SPARQL API
Core API
Cross-lingual Search API
OpenOpps (OO) procurement database
Distributed data sets
JSON
KG ingestion ETL
RDF
owl:sameAs
Triple store
Document store
rdf:seeAlso
OpenCorporates (OC) company database

Figure 9: TBFY platform architecture mapped to the DataBench Generic Data Pipeline

TBFY data ingestion pipeline

The TBFY data ingestion pipeline consists of 6 steps:

1. **Download procurement data:** Downloads procurement data from the OpenOpps OCDS API as JSON data files.
2. **Reconcile supplier data:** Matches supplier records in awards using the OpenCorporates Reconciliation API. The matching company data is downloaded using the OpenCorporates Company API as JSON data files.
3. **Enrich JSON data:** Enriches the JSON data files downloaded in steps 1 and 2, e.g., adding new properties to support the mapping to RDF (e.g., fixing missing identifiers).
4. **Convert JSON to XML:** Converts the JSON data files from step 3 into corresponding XML data files. Due to limitations in JSONPath, i.e., lack of operations for accessing parent or sibling nodes from a given node, we prefer to use XPath as the query language in RML.
5. **Map XML data to RDF:** Runs RML Mapper on the enriched XML data files from step 4 and produces N-Triples files.
6. **Publish RDF to database:** Stores the RDF (NTriples) files from step 5 to Apache Jena Fuseki and Apache Jena TBD.

The first two steps correspond to the **Data Acquisition/Collection**, while the four last steps correspond to the **Data Storage/Preparation**.

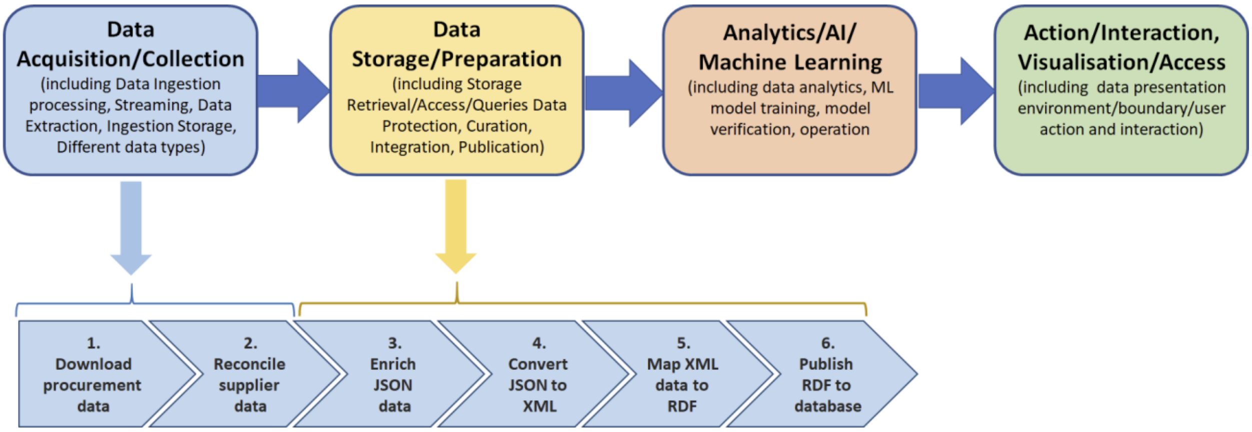

Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Analytics/AI/ Machine Learning
(including data analytics, ML model training, model verification, operation
Action/Interaction, Visualisation/Access
(including data presentation environment/boundary/user action and interaction)
1. Download procurement data
2. Reconcile supplier data
3. Enrich JSON data
4. Convert JSON to XML
5. Map XML data to RDF
6. Publish RDF to database

Figure 10: TBFY data ingestion pipeline mapped to the DataBench Generic Data Pipeline

Value-added services built on top of the TBFY platform

In the value-added services layer of the TBFY platform we find services and tools that provide features beyond the core platform functionality. Examples of such services are 1) cross-lingual search API [21], 2) knowledge graph queries, statistics and visualisation [22], and 3) anomaly detection that are based on machine learning models [23]. As such the value-added services also extends the platform to cover the **Analytics/AI/Machine Learning** and **Action/Interaction, Visualisation/Access.**

As shown in the figure below, the benchmarking conducted was related to data access/query as in the **Data Storage/Preparation** phase and data presentation/visualization in the **Action/Interaction, Visualization/Access** phase of the DataBench pipeline.

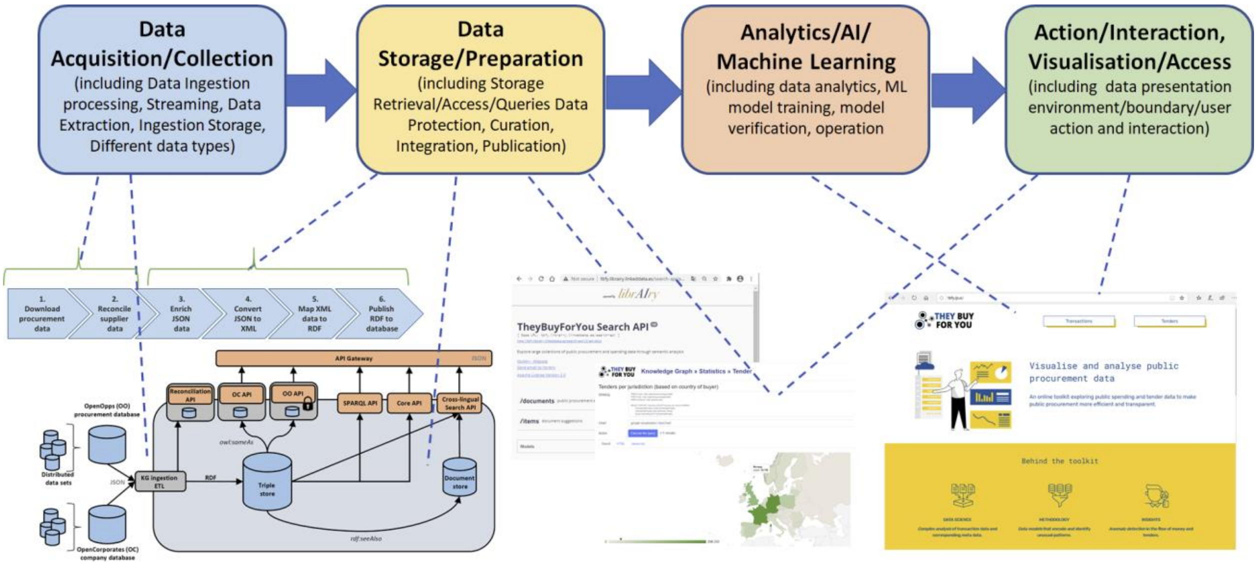

Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Analytics/AI/ Machine Learning
(including data analytics, ML model training, model verification, operation
Action/Interaction, Visualisation/Access
(including data presentation environment/boundary/user action and interaction)
1. Download procurement data
2. Reconcile supplier data
3. Enrich JSON data
4. Convert JSON to XML
5. Map XML data to RDF
6. Publish RDF to database
API Gateway
JSON
Reconciliation API
OC API
OO API
SPARQL API
Core API
Cross-lingual Search API
OpenOpps (OO) procurement database
Distributed data sets
JSON
KG ingestion ETL
RDF
owl:sameAs
Triple store
Document store
rdf:seeAlso
OpenCorporates (OC) company database
librAIry
TheyBuyForYou Search API
/documents
/items
Models
THEY BUY FOR YOU
Knowledge Graph » Statistics » Tender
Tenders per jurisdiction (based on country of buyer)
Transactions
Tenders
Visualise and analyse public procurement data
An online toolkit exploring public spending and tender data to make public procurement more efficient and transparent.
Behind the toolkit

Figure 11: TBFY platform with value-added services mapped to the DataBench Generic Data Pipeline

Relevant blueprints - use case independent and domain / use-case specific

The DataBench toolbox contains some relevant blueprints related to the horizontal concern Data Management [24] of the BDV Reference Model [25], including:

- The DataBench Framework and the BDV Reference Model
- DataBench Generic Data Pipeline
- Generic Big Data Analytics Blueprint
- Use case independent blueprints - Data Management Systems - Extract-Transform-Load storage architecture
- ETL (Extract-Transform-Load)

However, we did not find any specific blueprints related to constructing knowledge graphs involving mapping source data to RDF.

Mapping to the Generic Big Data Analytics Blueprint

In the Generic Big Data Analytics Blueprint [26] figure below, we have highlighted the components covered by the extended TBFY platform, which includes the value-added services built on top. The colour highlighting matches the colours for the 4 steps that we used in the mapping to the DataBench Generic Data Pipeline described above.

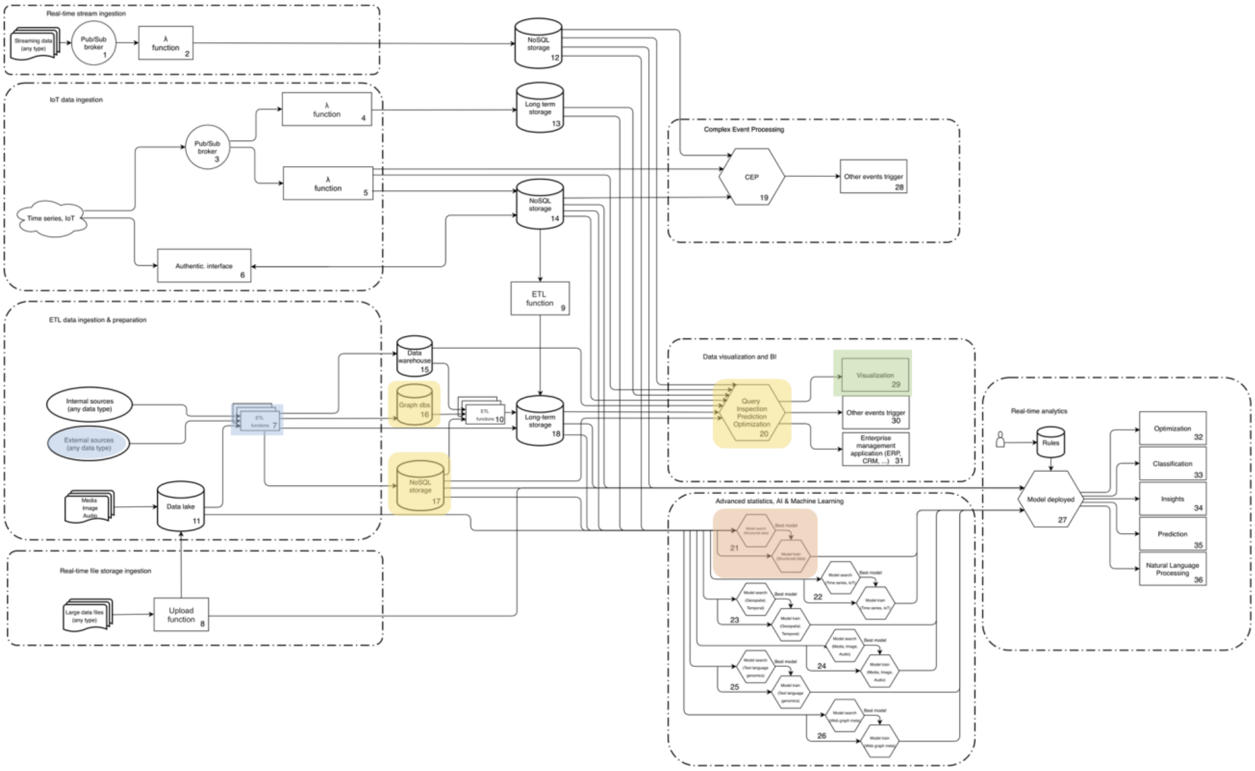

Real-time stream ingestion
Streaming data (any type)
Pub/Sub broker 1
λ function 2
NoSQL storage 12
IoT data ingestion
λ function 4
Long term storage 13
Pub/Sub broker 3
λ function 5
Time series, IoT
NoSQL storage 14
Authentic. interface 6
ETL function 9
ETL data ingestion & preparation
Data warehouse 15
Internal sources (any data type)
External sources (any data type)
ETL functions 7
Graph dbs 16
ETL functions 10
Long-term storage 18
NoSQL storage 17
Media Image Audio
Data lake 11
Real-time file storage ingestion
Large data files (any type)
Upload function 8
Complex Event Processing
CEP 19
Other events trigger 28
Data visualization and BI
Visualization 29
Query Inspection Prediction Optimization 20
Other events trigger 30
Enterprise management application (ERP, CRM, ...) 31
Real-time analytics
Rules
Model deployed 27
Optimization 32
Classification 33
Insights 34
Prediction 35
Natural Language Processing 36
Advanced statistics, AI & Machine Learning
Model search (Structured data)
Best model
Model train (Structured data)
21
Model search (Time series, IoT)
Best model
22
Model train (Time series, IoT)
Model search (Geospatial, Temporal)
Best model
23
Model train (Geospatial, Temporal)
Model search (Media, Image, Audio)
Best model
24
Model train (Media, Image, Audio)
Model search (Text language genomics)
Best model
25
Model train (Text language genomics)
Model search (Web graph meta)
Best model
26
Model train (Web graph meta)

Figure 12: TBFY platform with value-added services mapped to the Generic Big Data Analytics Blueprint

Relevant standards for big data and AI technologies (ISO SC42)

The ongoing artificial intelligence standardisation in ISO, i.e. the SC42 [27], is not readily available for free. At the time of writing, 6 standards are published [28] and 21 standards are under development [29].

For the Knowledge Graph based approach of the TBFY project, standardisation related to data models for procurement data was of key importance. Based on the Open Contracting Data Standard (OCDS) [30] we developed an OCDS ontology [31] to represent procurement data in the Knowledge Graph based TBFY platform.

4.3.4. Relevant Benchmarks and Nuggets from DataBench Toolbox

Two key components of the TBFY platform were subject for benchmarking:

1. The **TBFY data ingestion pipeline** that downloads procurement and company data, links the data, enriches the data, transforms the data to RDF format, and finally publishes RDF data.
2. The **TBFY Knowledge Graph (KG) triple store** where the published RDF data is made available through APIs and SPARQL endpoints.

<u>TBFY data ingestion pipeline</u>

Benchmark objective

- Measure performance.

- Identify bottlenecks.

Relevant benchmarks

The DataBench Toolbox contains several benchmarks related to batch processing architectures that provides a good starting point for identifying relevant benchmarks for the TBFY data ingestion pipeline. Table 5 depicts the top 10 technical benchmarks from the BDV Reference Model search for three relevant selected data types and batch processing architectures. The last column represents the 5 most relevant knowledge nuggets resulting from a Guided Benchmark search filtered by the retail trade tag as part of the Business Features.

Go to Search --> BDV Reference Model and				**Go to Search --> Guided Benchmark Search and select the Business Features drop-down menu. Then click on the following tags:**
select one of the Data Types:			**or select one of the Horizontal Layers:**	
Time series, IoT	Geo Spatial Temporal	Web Graph Meta	Data Processing Architectures: Batch (data-at-rest)	Retail trade
The search returns the following list of benchmarks (only the top 10):				
owperf (CLASS)	IDEBench	HiBench	BigBench V2	
Yahoo Streaming Benchmark (YSB)	MLPerf	Berlin SPARQL Benchmark (BSBM)	HiBench	
AIoTBench	Training Benchmark for DNNs (TBD)	BigFrame	ABench	
BenchIoT		CloudRank-D	AMP Lab Big Data Benchmark	
CityBench		gMark	Berlin SPARQL Benchmark (BSBM)	
CloudRank-D		Graphalytics	BigBench	
CloudSuite		Hobbit Benchmark	BigDataBench	
DeepMark (Convnet)		LinkBench	CALDA	
Edge AI Bench		LIQUID	CloudRank-D	
HERMIT		MiDBench	CloudSuite	
The search returns the following list of knowledge nuggets (only the top 5):				
Benchmarks features matrix	Benchmarks features matrix	Linked Data Integration and Publication Pipeline pattern	Benchmarks features matrix	Evidence of business process performance - Recommendation system (Retail, eCommerce)
Data features Star Diagram	Data features Star Diagram	Benchmarks features matrix	Classification of data processing architectures	Qualitative Benchmarks for the Retail and Wholesale Industry

IoT Ingestion and Authentication	Earth Observation and Geospatial Pipeline pattern	Data features Star Diagram	Data features Star Diagram	Quantitative Benchmarks for the Retail and Wholesale Industry
IoT Pipeline pattern	Technical benchmarks Star Diagram	Linked Data Benchmark Council (LDBC)	Project ALOJA	Quantitative Benchmarks for the Top 3 Use Cases in Retail and Wholesale
Technical benchmarks Star Diagram	Transaction Processing Performance Council	NoSQL - Graph DB		Retail - price optimization and customized promotions Blueprint

Table 6: Results from BDV Reference Model Search by Data Type and Horizontal Layers and Guided Benchmark Search filtering by Business Features (TBFY project)

TBFY Knowledge Graph (KG) triple store

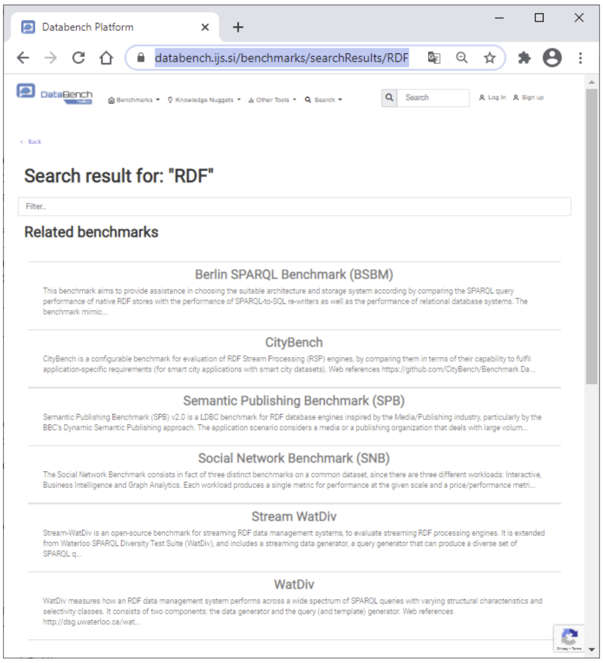
Benchmark objective

- Measure query performance.
- Decide which triple store to choose.
- Decide cloud hosting plan (cost vs performance).
- Tune database parameters.

Relevant benchmarks

The DataBench toolbox provides 4 types of searches, i.e., 1) free-text search, 2) guided search [32], 3) blueprint/pipeline search [33], and 4) BDV reference model search [34]. The table below shows screenshots of the search results for each of these 4 types selecting “RDF”, “Platform and Architecture Features → Storage type → Graph Databases”, “Data Storage/Preparation → Graph dbs” and “Data Types → Web Graph Meta”, respectively.

Free-text search for “RDF” [35]	Guided search selecting “Platform and Architecture Features → Storage type →Graph Databases” [36]



Search result for: "RDF"

Related benchmarks

Berlin SPARQL Benchmark (BSBM)

This benchmark aims to provide assistance in choosing the suitable architecture and storage system according by comparing the SPARQL query performance of native RDF stores with the performance of SPARQL-to-SQL re-writers as well as the performance of relational database systems. The benchmark mimic...

CityBench

CityBench is a configurable benchmark for evaluation of RDF Stream Processing (RSP) engines, by comparing them in terms of their capability to fulfil application-specific requirements (for smart city applications with smart city datasets). Web references https://github.com/CityBench/Benchmark Da...

Semantic Publishing Benchmark (SPB)

Semantic Publishing Benchmark (SPB) v2.0 is a LDBC benchmark for RDF database engines inspired by the Media/Publishing industry, particularly by the BBC's Dynamic Semantic Publishing approach. The application scenario considers a media or a publishing organization that deals with large volum...

Social Network Benchmark (SNB)

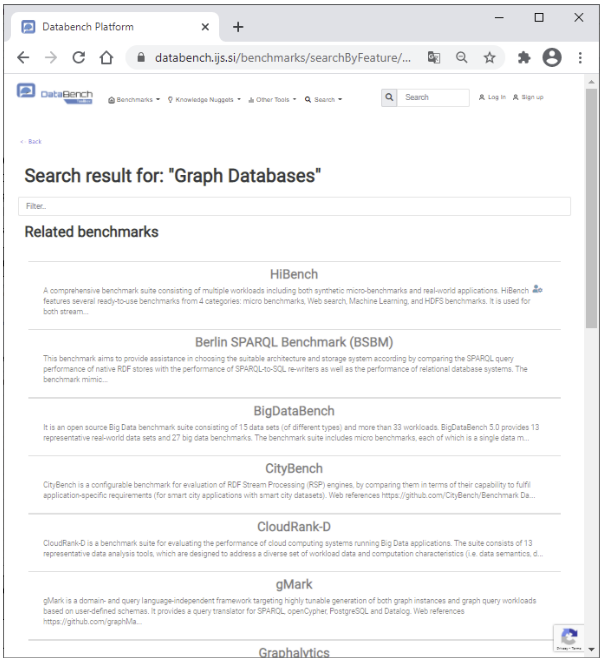
The Social Network Benchmark consists in fact of three distinct benchmarks on a common dataset, since there are three different workloads: Interactive, Business Intelligence and Graph Analytics. Each workload produces a single metric for performance at the given scale and a price/performance metri...

Stream WatDiv

Stream-WatDiv is an open-source benchmark for streaming RDF data management systems, to evaluate streaming RDF processing engines. It is extended from Waterloo SPARQL Diversity Test Suite (WatDiv), and includes a streaming data generator, a query generator that can produce a diverse set of SPARQL q...

WatDiv

WatDiv measures how an RDF data management system performs across a wide spectrum of SPARQL queries with varying structural characteristics and selectivity classes. It consists of two components: the data generator and the query (and template) generator. Web references http://dsg.uwaterloo.ca/wat...



Search result for: "Graph Databases"

Related benchmarks

HiBench

A comprehensive benchmark suite consisting of multiple workloads including both synthetic micro-benchmarks and real-world applications. HiBench features several ready-to-use benchmarks from 4 categories: micro benchmarks, Web search, Machine Learning, and HDFS benchmarks. It is used for both stream...

Berlin SPARQL Benchmark (BSBM)

This benchmark aims to provide assistance in choosing the suitable architecture and storage system according by comparing the SPARQL query performance of native RDF stores with the performance of SPARQL-to-SQL re-writers as well as the performance of relational database systems. The benchmark mimic...

BigDataBench

It is an open source Big Data benchmark suite consisting of 15 data sets (of different types) and more than 33 workloads. BigDataBench 5.0 provides 13 representative real-world data sets and 27 big data benchmarks. The benchmark suite includes micro benchmarks, each of which is a single data m...

CityBench

CityBench is a configurable benchmark for evaluation of RDF Stream Processing (RSP) engines, by comparing them in terms of their capability to fulfil application-specific requirements (for smart city applications with smart city datasets). Web references https://github.com/CityBench/Benchmark Da...

CloudRank-D

CloudRank-D is a benchmark suite for evaluating the performance of cloud computing systems running Big Data applications. The suite consists of 13 representative data analysis tools, which are designed to address a diverse set of workload data and computation characteristics (i.e. data semantics, d...

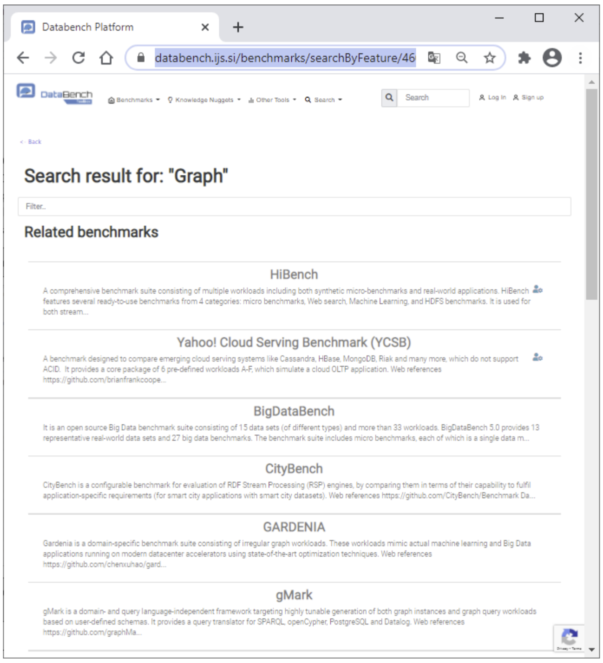
gMark

gMark is a domain- and query language-independent framework targeting highly tunable generation of both graph instances and graph query workloads based on user-defined schemas. It provides a query translator for SPARQL, openCypher, PostgreSQL and Datalog. Web references https://github.com/graphMa...

Graphalytics

Blueprint/pipeline search selecting “Data Storage/Preparation → Graph dbs” [37]

BDV reference model search selecting “Data Types → Web Graph Meta” [38]



Search result for: "Graph"

Related benchmarks

HiBench

A comprehensive benchmark suite consisting of multiple workloads including both synthetic micro-benchmarks and real-world applications. HiBench features several ready-to-use benchmarks from 4 categories: micro benchmarks, Web search, Machine Learning, and HDFS benchmarks. It is used for both stream...

Yahoo! Cloud Serving Benchmark (YCSB)

A benchmark designed to compare emerging cloud serving systems like Cassandra, HBase, MongoDB, Riak and many more, which do not support ACID. It provides a core package of 6 pre-defined workloads A-F, which simulate a cloud OLTP application. Web references https://github.com/brianfrankcoope...

BigDataBench

It is an open source Big Data benchmark suite consisting of 15 data sets (of different types) and more than 33 workloads. BigDataBench 5.0 provides 13 representative real-world data sets and 27 big data benchmarks. The benchmark suite includes micro benchmarks, each of which is a single data m...

CityBench

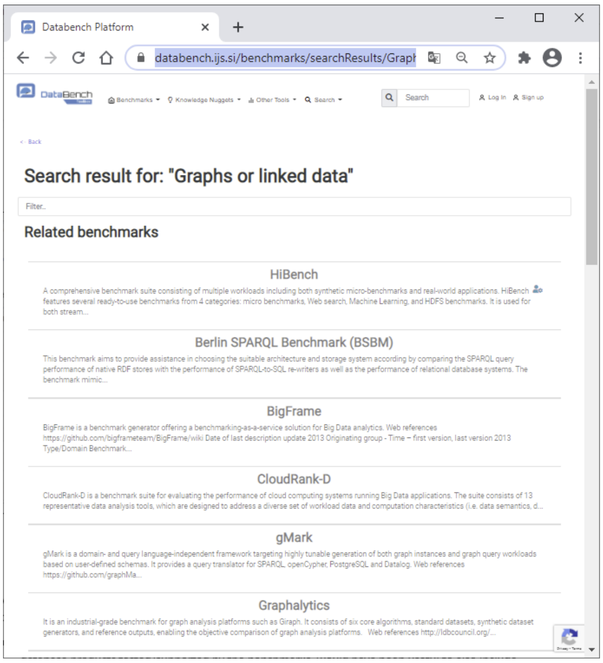
CityBench is a configurable benchmark for evaluation of RDF Stream Processing (RSP) engines, by comparing them in terms of their capability to fulfil application-specific requirements (for smart city applications with smart city datasets). Web references https://github.com/CityBench/Benchmark Da...

GARDENIA

Gardenia is a domain-specific benchmark suite consisting of irregular graph workloads. These workloads mimic actual machine learning and Big Data applications running on modern datacenter accelerators using state-of-the-art optimization techniques. Web references https://github.com/chenxuhao/gard...

gMark

gMark is a domain- and query language-independent framework targeting highly tunable generation of both graph instances and graph query workloads based on user-defined schemas. It provides a query translator for SPARQL, openCypher, PostgreSQL and Datalog. Web references https://github.com/graphMa...



Search result for: "Graphs or linked data"

Related benchmarks

HiBench

A comprehensive benchmark suite consisting of multiple workloads including both synthetic micro-benchmarks and real-world applications. HiBench features several ready-to-use benchmarks from 4 categories: micro benchmarks, Web search, Machine Learning, and HDFS benchmarks. It is used for both stream...

Berlin SPARQL Benchmark (BSBM)

This benchmark aims to provide assistance in choosing the suitable architecture and storage system according by comparing the SPARQL query performance of native RDF stores with the performance of SPARQL-to-SQL re-writers as well as the performance of relational database systems. The benchmark mimic...

BigFrame

BigFrame is a benchmark generator offering a benchmarking-as-a-service solution for Big Data analytics. Web references https://github.com/bigframeteam/BigFrame/wiki Date of last description update 2013 Originating group - Time – first version, last version 2013 Type/Domain Benchmark...

CloudRank-D

CloudRank-D is a benchmark suite for evaluating the performance of cloud computing systems running Big Data applications. The suite consists of 13 representative data analysis tools, which are designed to address a diverse set of workload data and computation characteristics (i.e. data semantics, d...

gMark

gMark is a domain- and query language-independent framework targeting highly tunable generation of both graph instances and graph query workloads based on user-defined schemas. It provides a query translator for SPARQL, openCypher, PostgreSQL and Datalog. Web references https://github.com/graphMa...

Graphalytics

It is an industrial-grade benchmark for graph analysis platforms such as Giraph. It consists of six core algorithms, standard datasets, synthetic dataset generators, and reference outputs, enabling the objective comparison of graph analysis platforms. Web references http://ldbcouncil.org/...

Figure 13: Search results for relevant benchmarks for TBFY

As it can be seen, several relevant benchmarks were suggested. This provided a really good starting point for selecting a benchmark that matched the needs in the TBFY project. However, further review of the suggested benchmarks still needs to be done by the end-user in order to narrow down the results. As part of this review, additional information such as ranking of benchmarks, direct links to installation and user guides, list of triple stores products tested/supported by the benchmarks, would have been useful information to further help narrowing down the search.

4.3.5. Analysis and benchmark results

The TBFY platform architecture and the TBFY data ingestion pipeline as described in this section were submitted as a suggestion for a new pipeline/blueprint to be included as a knowledge nugget in the DataBench toolbox as part of the DataBench campaign “Generation of architectural Pipelines-Blueprints” on ReachOut [39].

<u>TBFY data ingestion pipeline</u>

Benchmark results

- Bottlenecks identified.
- Revised implementation with improved performance.

The data ingestion pipeline was implemented in Python [40]. While the DataBench toolbox provided suggestions for relevant benchmarks, after a closer review of the suggestions we did not find any specific benchmark that perfectly matched our needs. Thus we decided to develop our own benchmarking approach by adding log files for each step. Running benchmark scripts on the logs we were able to identify bottlenecks in step 2. By implementing caching of reconciliation and company lookups we were able to significantly improve the performance of the pipeline.

<u>TBFY Knowledge Graph (KG) triple store</u>

Benchmark results

- Benchmarking of two different triple stores.
- Upgraded the cloud hosting (i.e., migrated to a more powerful Amazon EC2 instance).
- Database parameters adjusted (e.g., the JVM heap size).

After reviewing the benchmarks related to graph databases in the DataBench toolbox, we decided on using the LDBC Semantic Publishing Benchmark (SPB) [41] as this supported SPARQL 1.1 compliant databases that we were using in TBFY. We were in particular interested in the SPARQL query performance, as we had developed a RESTful API on top of the TBFY Knowledge Graph (KG). The RESTful API uses SPARQL queries to access the data from the KG and returns the results formatted as JSON. Additionally, we also set up a website [42] that runs live SPARQL queries to get statistics from the KG that are visualized as different charts, e.g., bar charts and geo charts.

We ran the SPB benchmarks on two different triple stores that we had been using in the TBFY project for the Knowledge Graph. The benchmarking helped us decide which triple store to choose, as well as finetune some deployment parameters such as the JVM heap size. Finally, we upgraded the cloud hosting, i.e., migrated to a more powerful Amazon EC2 instance to improve the overall performance.

4.4. DeepHealth

4.4.1. Project description

The DeepHealth[6] (2019-2021) project is one of the Big Data PPP projects. Its aim is to offer a unified framework completely adapted to exploit underlying heterogeneous HPC and Big Data architectures; and assembled with state-of-the-art techniques in Deep Learning and Computer Vision.

Deep-Learning and HPC to Boost Biomedical Applications for Health (DeepHealth) project is funded by the EC under the topic ICT-11-2018-2019 "HPC and Big Data enabled Large-scale Test-beds and Applications". DeepHealth is a 3-year project, kicked-off in mid January 2019 and is expected to conclude its work in December 2021. The aim of DeepHealth is to offer a unified framework completely adapted to exploit underlying heterogeneous HPC and Big Data architectures; and assembled with state-of-the-art techniques in Deep Learning and Computer Vision.

4.4.2. Project objectives and Architecture

The DeepHealth project will combine High-Performance Computing (HPC) infrastructures with Deep Learning (DL) and Artificial Intelligence (AI) techniques to support biomedical applications that require the analysis of large and complex biomedical datasets and thus, new and more efficient ways of diagnosis, monitoring and treatment of diseases. Moreover, two new libraries, the European Distributed Deep Learning Library (EDDLL) and the European Computer Vision Library (ECVL), will be developed and incorporated in the DeepHealth framework for manipulating and processing the images in a more efficient way and thus, for increasing the productivity of professionals working on biomedical images. The resulting enhanced diagnosis will significantly improve the health service provided to the society, making public health systems more efficient and profitable for everyone.

[6] https://deephealth-project.eu/

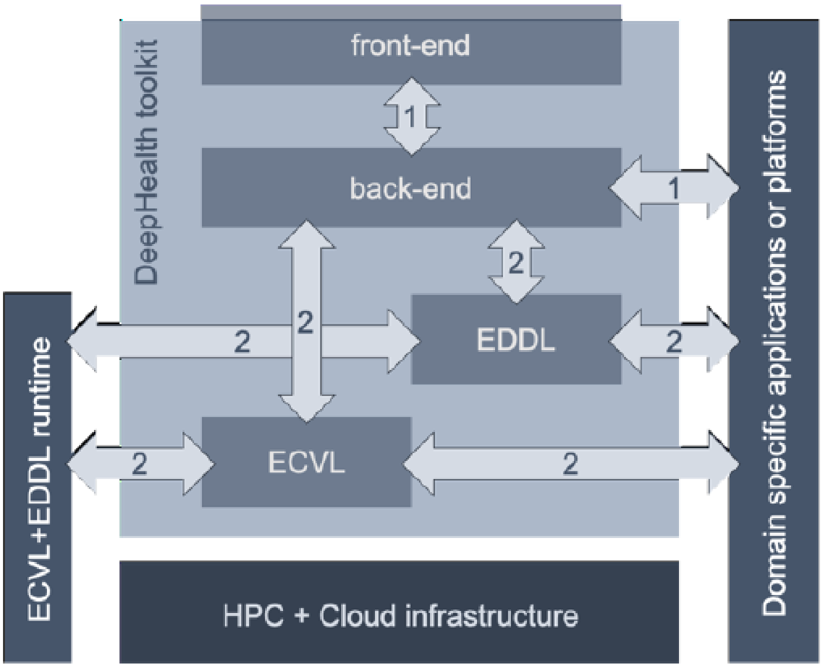


Figure 14: DeepHealth Architecture

4.4.3. Mapping to DataBench Pipeline

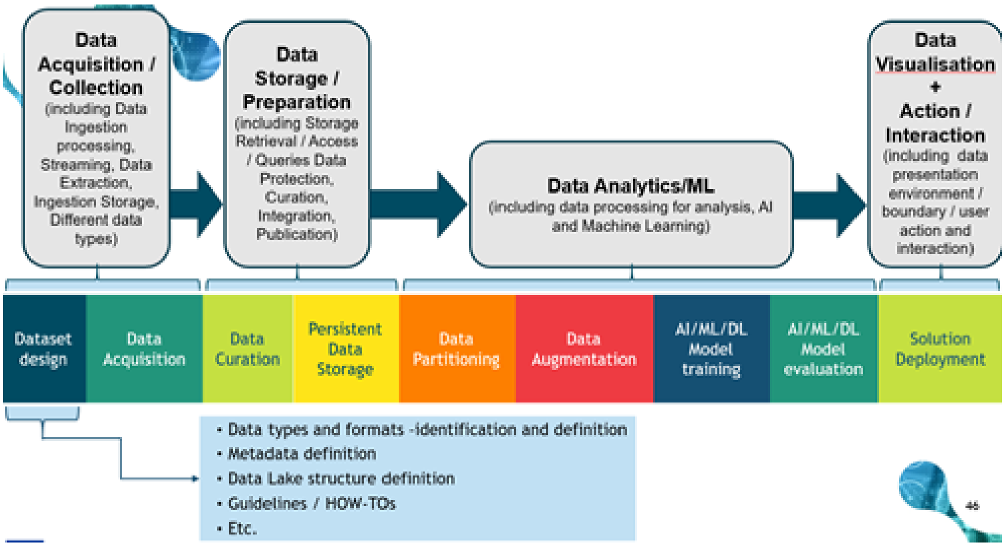

Data Acquisition / Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage / Preparation
(including Storage Retrieval / Access / Queries Data Protection, Curation, Integration, Publication)
Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)
Data Visualisation + Action / Interaction
(including data presentation environment / boundary / user action and interaction)
Dataset design
Data Acquisition
Data Curation
Persistent Data Storage
Data Partitioning
Data Augmentation
AI/ML/DL Model training
AI/ML/DL Model evaluation
Solution Deployment
• Data types and formats -identification and definition
• Metadata definition
• Data Lake structure definition
• Guidelines / HOW-TOs
• Etc.
46

Figure 15: Mapping of DeepHealth pipeline to DataBench pipeline

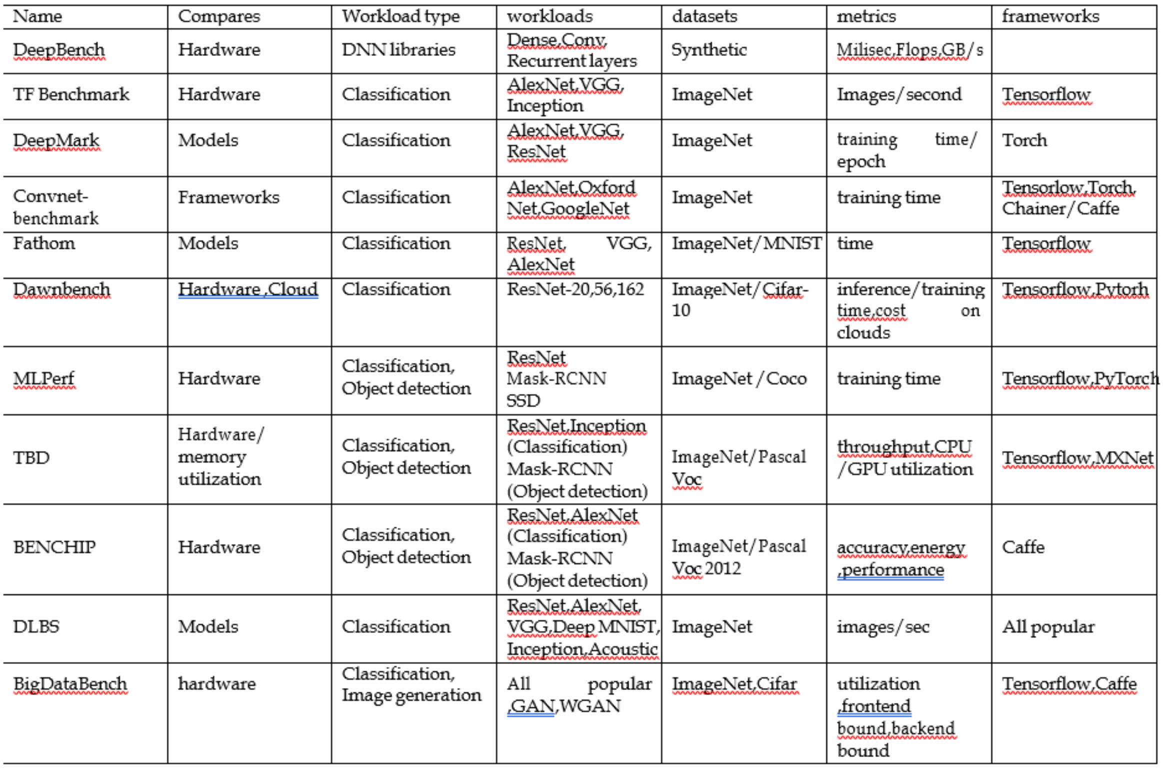
4.4.4. Relevant Benchmarks and Nuggets from DataBench Toolbox

This section provides suggestions for relevant benchmarks and knowledge nuggets that include Big Data and AI technologies, architectural patterns and blueprints as well as business benchmarks for comparison. Figure 6 below represents the top 10 technical benchmarks and knowledge nuggets resulting from BDV Reference Model searches by data type and three different Horizontal layers. The last column showing the top 5 most relevant knowledge nugget results from the Guided Benchmark Search in the Business Features selecting the Health industry tag.

Go to Search --> BDV Reference Model and				**Go to Search --> Guided Benchmark Search and select the Business Features drop-down menu. Then click on the following tags:**
select one of the Data Types:	**or select one of the Horizontal Layers:**			
Media Image Audio	Data Analytics	Data Processing Architectures: Batch (data-at-rest)	Cloud and High Performance computing (HPC)	Health
The search returns the following list of benchmarks (only the top 10):				
AIBench	BigBench V2	BigBench V2	CloudSuite	
AIMatrix	HiBench	HiBench	GARDENIA	
BigDataBench	ABench	ABench	HPC AI500	
CloudSuite	AdBench	AMP Lab Big Data Benchmark	MLBench Services	
DAWNBench	AIBench	Berlin SPARQL Benchmark (BSBM)	PUMA Benchmark Suite	
Deep Learning Benchmarking Suite (DLBS)	AIM Benchmark	BigBench	TPCx-V	
DeepMark (Convnet)	AIMatrix	BigDataBench		
Edge AI Bench	AIoTBench	CALDA		
Fathom	ALOJA	CloudRank-D		
HPC AI500	Benchip	CloudSuite		
The search returns the following list of knowledge nuggets (only the top 5):				
Benchmarks features matrix	Agriculture Architectural Blueprints	Benchmarks features matrix	Benchmarks features matrix	Genomic Pipeline pattern
Computer Vision	AI&ML Developmemt Platforms (IDEs)	Classification of data processing architectures	Technical benchmarks Star Diagram	Healthcare - Patient monitoring Blueprint
Data features Star Diagram	AI&ML Frameworks	Data features Star Diagram	The DataBench Framework and the BDV Reference Model	Qualitative Benchmarks for the Healthcare Industry
Technical benchmarks Star Diagram	AI&ML Libraries	Project ALOJA		Quantitative Benchmarks for the Healthcare Industry
	AI&ML Platforms			Quantitative Benchmarks for the Top 3 Use Cases in Healthcare

Table 7: Results from BDV Reference Model Search by Data Type and Horizontal Layers and Guided Benchmark Search filtering by Business Features for the DeepHealth project

Figure 7 below summarizes a list of technical AI benchmarks, framework and technologies that was assembled based on the Toolbox search results and knowledge nuggets, which contain lists of multiple technologies and libraries.



Name	Compares	Workload type	workloads	datasets	metrics	frameworks
DeepBench	Hardware	DNN libraries	Dense,Conv, Recurrent layers	Synthetic	Milisec,Flops,GB/s	
TF Benchmark	Hardware	Classification	AlexNet,VGG, Inception	ImageNet	Images/second	Tensorflow
DeepMark	Models	Classification	AlexNet,VGG, ResNet	ImageNet	training time/ epoch	Torch
Convnet-benchmark	Frameworks	Classification	AlexNet,Oxford Net,GoogleNet	ImageNet	training time	Tensorlow,Torch, Chainer/Caffe
Fathom	Models	Classification	ResNet, VGG, AlexNet	ImageNet/MNIST	time	Tensorflow
Dawnbench	Hardware ,Cloud	Classification	ResNet-20,56,162	ImageNet/Cifar-10	inference/training time,cost on clouds	Tensorflow,Pytorh
MLPerf	Hardware	Classification, Object detection	ResNet Mask-RCNN SSD	ImageNet / Coco	training time	Tensorflow,PyTorch
TBD	Hardware/ memory utilization	Classification, Object detection	ResNet,Inception (Classification) Mask-RCNN (Object detection)	ImageNet/Pascal Voc	throughput,CPU /GPU utilization	Tensorflow,MXNet
BENCHIP	Hardware	Classification, Object detection	ResNet,AlexNet (Classification) Mask-RCNN (Object detection)	ImageNet/Pascal Voc 2012	accuracy,energy ,performance	Caffe
DLBS	Models	Classification	ResNet,AlexNet, VGG,Deep MNIST, Inception,Acoustic	ImageNet	images/sec	All popular
BigDataBench	hardware	Classification, Image generation	All popular ,GAN,WGAN	ImageNet,Cifar	utilization ,frontend bound,backend bound	Tensorflow,Caffe

Table 8: Summary of relevant AI Benchmarks with links to Deep Learning Frameworks

4.5. DataBio

4.5.1. Project description

DataBio[7] (2017-2019) is one of the first Big Data PPP H2020 lighthouse projects - focusing on utilizing Big Data to contribute to the production of the best possible raw materials from agriculture, forestry, and fishery/aquaculture for the bioeconomy industry in order to produce food, energy and biomaterials, also taking into account responsibility and sustainability issues.

DataBio has deployed state-of-the-art Big Data technologies taking advantage of existing partners' infrastructure and solutions. These solutions aggregate Big Data from the three identified sectors (agriculture, forestry, and fishery) and intelligently process, analyse and visualize them. The DataBio software environment allows the three sectors to selectively utilize numerous software components, pipelines and datasets, according to their requirements. The execution has been through continuous cooperation of end-users and technology provider companies, bioeconomy and technology research institutes, and stakeholders from the EU´s Big Data Value PPP programme.

[7] https://www.databio.eu/

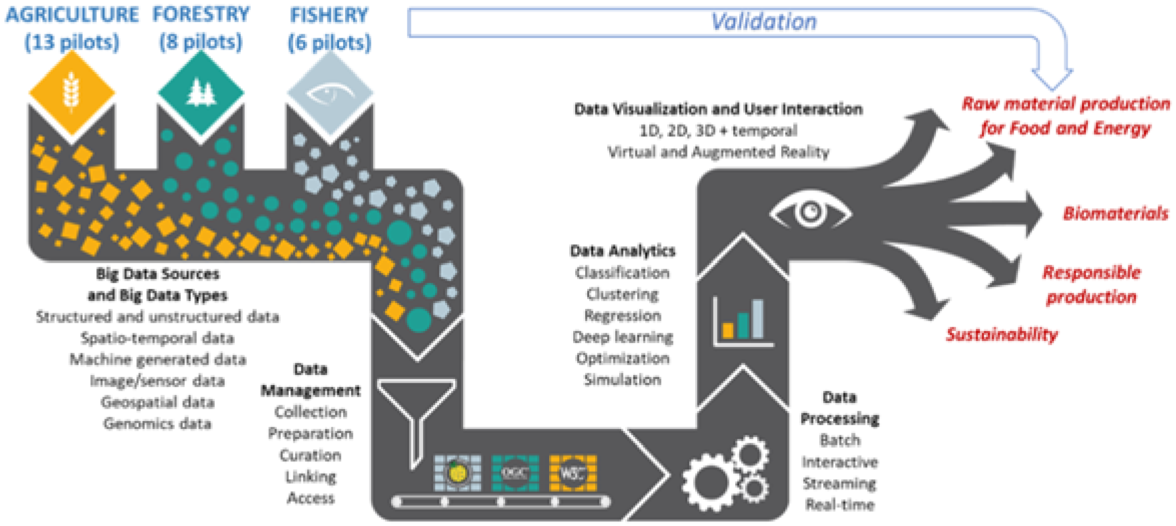


Figure 16: DataBio project overview

DataBio has been driven by the development, use and evaluation of 27 pilots, where also associated partners and additional stakeholders have been involved. The selected pilot concepts have been transformed into pilot implementations utilizing co-innovative methods and tools. Through intensive matchmaking with the technology partners in DataBio, the pilots have selected and utilized market-ready or near market-ready ICT, Big Data and Earth Observation methods, technologies, tools, datasets and services, mainly provided by the partners within DataBio, in order to offer added-value services in their domain.

Based on the developed technologies and the pilot results, new solutions and new business opportunities are emerging. DataBio has organized a series of stakeholder events, hackathons and trainings to support result take-up and to enable developers outside the consortium to design and develop new tools, services and applications based on the DataBio results.

4.5.2. Project objectives and Architecture

The DataBio objectives are:

1. Build a versatile DataBio platform suitable for different industries and user profiles
2. Ensure effective utilization of existing data sets
3. Ensure a wide-spread use of the DataBio platform technologies in the agriculture, forestry and fishery sectors
4. Opening the possibilities for European ICT industry including SMEs to participate actively on European and WorldWide Bioeconomy Big Data market
5. Opening the possibilities for European Earth Observation industry including SMEs offering their new Bioeconomy related services in Europe and World Wide
6. Ensure interoperability and easy setup of new multivendor applications utilizing Big DataBio platform

Data handling in DataBio specifically aimed at the following bio economy sectors:

Agriculture: The main goal was to develop smart agriculture solutions that boost the production of raw materials for the agri-food chain in Europe while making farming sustainable. This includes optimized irrigation, and use of fertilizers and pesticides, prediction of yield and diseases, identification of crops and assessment of damages. Such smart agriculture solutions are based on data from satellites, drones, IoT sensors, weather stations and genomic data.

Forestry: Big Data methods are expected to bring the possibility to both increase the value of the forests as well as to decrease the costs within sustainability limits set by natural growth and ecological aspects. The key technology is to gather more and more accurate information about the trees from a host of sensors including new generations of satellites, UAV images, laser scanning, crowdsourced data collected by mobile devices and data collected by machines operating in the forests.

Fisheries: The ambition is to herald and promote the use of Big Data analytical tools to improve the ecological and economic sustainability, such as improved analysis of operational data for engine fault detection and fuel reduction, tools for planning and operational choices for fuel reduction when searching and choosing fishing grounds, as well as crowdsourcing methods for fish stock estimation.

One of the key technical objectives of Databio is to build a versatile DataBio platform suitable for different industries and user profiles and ensure effective utilization of existing data sets. The DataBio platform consists of a development environment, software components used and developed by DataBio partners and pipelines connecting the components to services. The Big Data *toolset* provided by the project offers functionalities primarily for *services* in the domains of agriculture, forestry and fishery. The functionalities enable new software *components* to be easily and effectively combined with open-source, standards-based Big Data, and proprietary components and infrastructures based on the use of generic and domain-specific components.

4.5.3. Mapping to DataBench Pipeline and Blueprints

The DataBio toolset supports the forming of reusable and deployable *pipelines* of interoperable components (mostly provided by partners), thus extending the impact of DataBio to new bioeconomy projects as well as to other business areas.

In the following we present two of the DataBio pipelines that were constructed during the project and we show the mapping of their steps to the steps of the Top level Generic Big Data and AI Pipeline pattern, which has been described in deliverable “D5.4 Analytic modelling relationships between metrics, data and project methodologies”. The first pipeline that appears below is the Generic pipeline for the Forest data ecosystem data processing and decision-making. The main characteristic of this generic pipeline is the collection of standardized data from multiple data sources to generate a forest data ecosystem where the decision-making can be based on a combination of different data types via the standardized API integrations. This pipeline has been derived from three pilots in forestry.

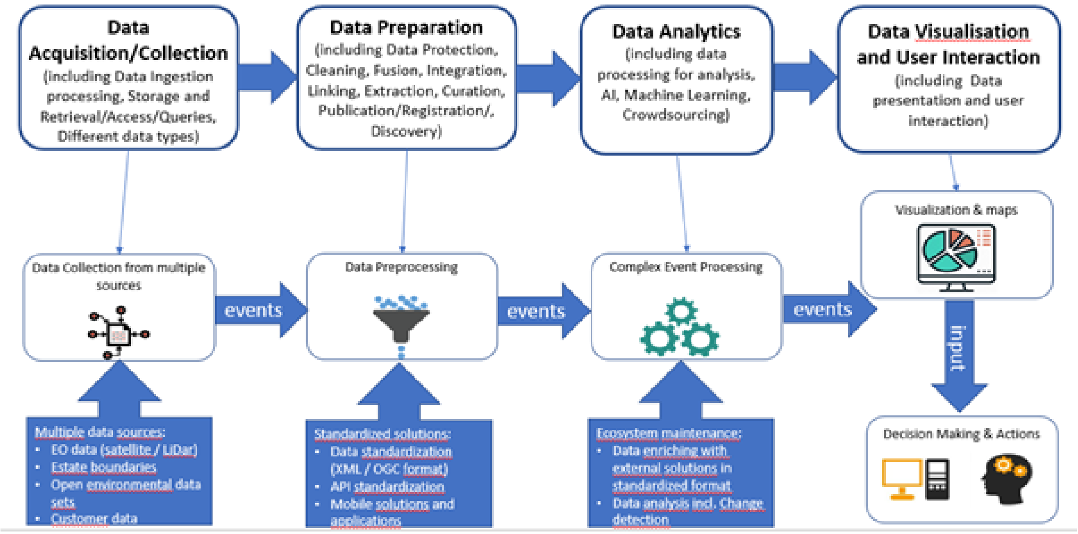

Data Acquisition/Collection
(including Data Ingestion processing, Storage and Retrieval/Access/Queries, Different data types)
Data Preparation
(including Data Protection, Cleaning, Fusion, Integration, Linking, Extraction, Curation, Publication/Registration/, Discovery)
Data Analytics
(including data processing for analysis, AI, Machine Learning, Crowdsourcing)
Data Visualisation and User Interaction
(including Data presentation and user interaction)
Data Collection from multiple sources
events
Data Preprocessing
events
Complex Event Processing
events
Visualization & maps
input
Decision Making & Actions
Multiple data sources:
EO data (satellite / LiDar)
Estate boundaries
Open environmental data sets
Customer data
Standardized solutions:
Data standardization (XML / OGC format)
API standardization
Mobile solutions and applications
Ecosystem maintenance:
Data enriching with external solutions in standardized format
Data analysis incl. Change detection

Figure 17: Mapping of the Generic pipeline for the Forest data ecosystem data processing and decision-making to the top-level pipeline

Conceptually, the same approach could be applied to other domains beyond forestry. Basically, all the use cases from any domain in which data is collected in a standardized format from the different data sources and furthermore enriched with additional operations via the standardized API's can nicely fit into this generic pipeline. For example: the observation data regarding the environmental and physical phenomena collected by the citizens or industry professionals by mobile applications can be further adopted as a part of the forest data ecosystem. This data can be used to analyse the possible magnitude of the forest damages further as well as for preventing the spread of the damage and additional implications such as industry profit losses or ecosystem damages. Another use case is related to the field data such as sample plot data collected by the mobile solution end-users, which can be further combined with other data sources and utilized as a reference data for the satellite data analysis and related data products.

Another domain that was exercised in the DataBio project is the fishery domain. A deciding factor for the efficiency of fishing vessels is the ability to find the best suitable fishing grounds, combined with the best suitable fishing methods and tools. Such decisions are today based on the skills and knowledge of the captain and crew in combining public data (weather, earth observation and public catch data) with private data (catch history, locations and price) to optimize the fisheries. There is a strong need for decision support applications in the planning and execution of open sea fisheries, and the application of Big Data technology in catch and process yield optimization have largely untapped potential in this industry. This challenge goes beyond the need of the agriculture and forestry applications, as it is not normally known where the fish is located when the vessel leaves the port.

The distances travelled involved in finding and harvesting the biomass is of quite another scale than land-based production. The search area is huge, for example, in whitefish trawling a typically three-week fishing trip can circumvent the entire Barents' sea before returning to port in Troms or Vesterålen, a single trip of many thousands of kilometers. With respect to this aspect, fisheries can be said to have more in common with hunting than harvesting. Big

Data technology, and in particular large-scale prediction analytics that can help locate the fishing grounds is therefore of key interest to the fishery sector.

A generic pipeline for fisheries data and analytics visualization purposes was devised based on the common reusable (sub) pipeline of four small pelagic fishery pilots in DataBio. This generalized pipeline represents a valuable and exploitable asset applicable to other use-cases.

Figure 18 shows the end-to-end flow from collecting public and private datasets, with very varying data rates, e.g. from yearly and quarterly statistical datasets to daily, hourly and real-time vessel data like private catch data and sensor data such as vessel location, motion and orientation. These datasets are preprocessed, e.g., filtered, converted and collated, to standardized formats that are mapped, cross-filtered and charted in a graphical user interface and also used to train machine learning models for catch prediction. The output recommendations of the prediction models are displayed in the map GUI together with the collated datasets. It also shows the mapping of these steps to the generic top-level pipeline.

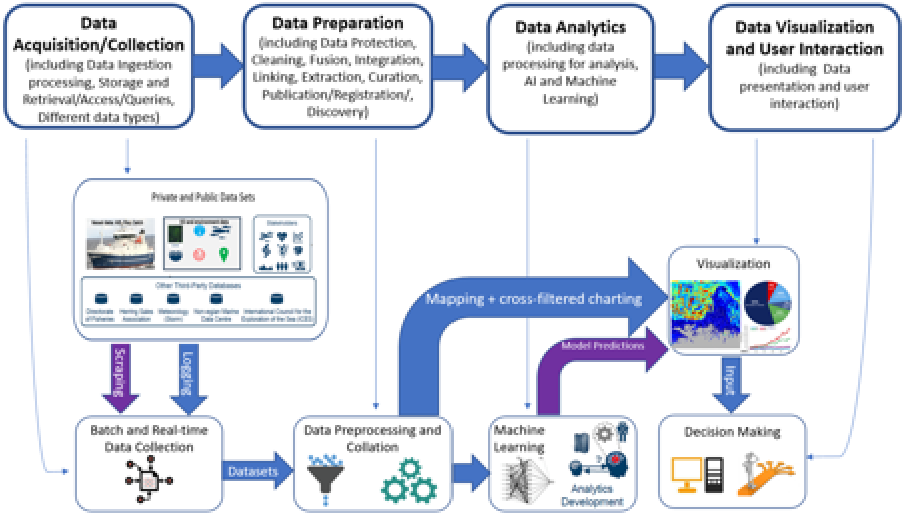

Data Acquisition/Collection (including Data Ingestion processing, Storage and Retrieval/Access/Queries, Different data types)
Data Preparation (including Data Protection, Cleaning, Fusion, Integration, Linking, Extraction, Curation, Publication/Registration/, Discovery)
Data Analytics (including data processing for analysis, AI and Machine Learning)
Data Visualization and User Interaction (including Data presentation and user interaction)
Private and Public Data Sets
Other Third-Party Databases
Scraping
Logging
Batch and Real-time Data Collection
Datasets
Data Preprocessing and Collation
Mapping + cross-filtered charting
Model Predictions
Machine Learning
Analytics Development
Visualization
Input
Decision Making

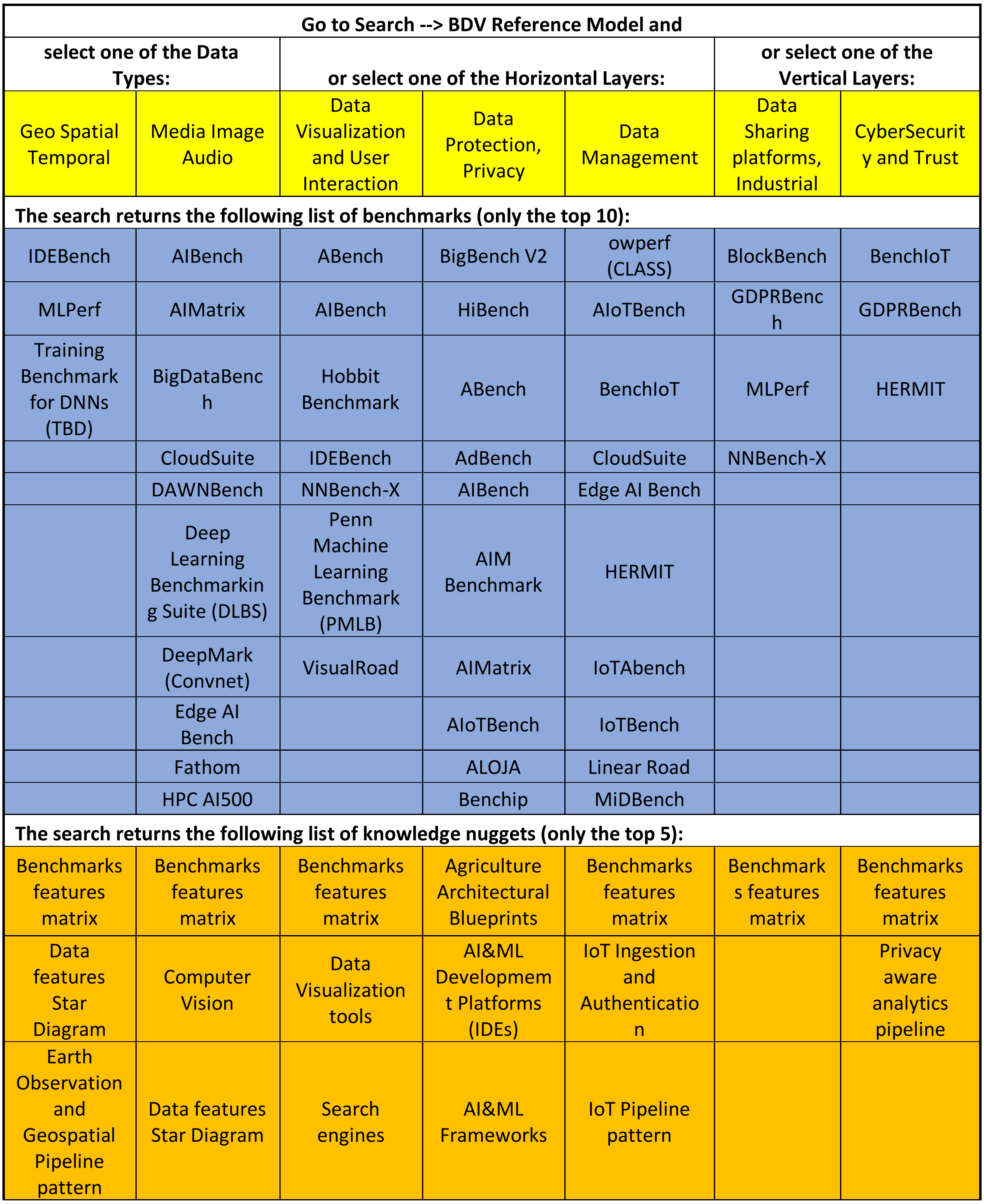
Figure 18: Generic pipeline for processing heterogeneous datasets for fish catch prediction and its mapping to the Generic top-level pipeline

The work in DataBio is in compliance with existing standards in the area of Big Data, such as the NIST Big Data Interoperability Framework: Volume 6, Reference Architecture [43], the ISO JTC1 SC42 AI and Big Data Reference Architecture, the OGC 10-157r4 “Earth Observation Profile of Observations and Measurements (O&M)” standard supported by European and Canadian Space Agencies [44].

4.5.4. Relevant Benchmarks and Nuggets from DataBench Toolbox

The section provides suggestions for relevant benchmarks and knowledge nuggets that include Big Data and AI technologies, architectural patterns and blueprints as well as business benchmarks for comparison. The following tables represent multiple searches that were performed with the current version of the DataBench Toolbox and illustrate its current results.

Figure 8 shows the top 10 technical benchmarks and top 5 knowledge nuggets resulting from a BDV Reference Model Search for three selected data types, three selected Horizontal Layers and two selected Vertical Layers. All the filtering criteria applied in the searches are based on the DataBio project descriptions provided above.



Go to Search --> BDV Reference Model and						
select one of the Data Types:		or select one of the Horizontal Layers:			or select one of the Vertical Layers:	
Geo Spatial Temporal	Media Image Audio	Data Visualization and User Interaction	Data Protection, Privacy	Data Management	Data Sharing platforms, Industrial	CyberSecurity and Trust
The search returns the following list of benchmarks (only the top 10):						
IDEBench	AIBench	ABench	BigBench V2	owperf (CLASS)	BlockBench	BenchIoT
MLPerf	AIMatrix	AIBench	HiBench	AIoTBench	GDPRBench	GDPRBench
Training Benchmark for DNNs (TBD)	BigDataBench	Hobbit Benchmark	ABench	BenchIoT	MLPerf	HERMIT
	CloudSuite	IDEBench	AdBench	CloudSuite	NNBench-X	
	DAWNBench	NNBench-X	AIBench	Edge AI Bench		
	Deep Learning Benchmarking Suite (DLBS)	Penn Machine Learning Benchmark (PMLB)	AIM Benchmark	HERMIT		
	DeepMark (Convnet)	VisualRoad	AIMatrix	IoTAbench		
	Edge AI Bench		AIoTBench	IoTBench		
	Fathom		ALOJA	Linear Road		
	HPC AI500		Benchip	MiDBench		
The search returns the following list of knowledge nuggets (only the top 5):						
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Agriculture Architectural Blueprints	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix
Data features Star Diagram	Computer Vision	Data Visualization tools	AI&ML Development Platforms (IDEs)	IoT Ingestion and Authentication		Privacy aware analytics pipeline
Earth Observation and Geospatial Pipeline pattern	Data features Star Diagram	Search engines	AI&ML Frameworks	IoT Pipeline pattern		

Technical benchmarks Star Diagram	Technical benchmarks Star Diagram	Use case independent blueprints - Data Processing and Exploitation Systems	AI&ML Libraries			
Transaction Processing Performance Council		Use case independent blueprints - Data processing and exploitation systems - Data visualization and business intelligence architecture	AI&ML Platforms			

Table 9: Results from BDV Reference Model Search by Data Type, Horizontal and Vertical Layers for the DataBio Project

Figure 9 below shows the 5 most relevant knowledge nuggets obtained from Guided Benchmark search for Business Features filtered by manufacturing and health industries.

Go to Search --> Guided Benchmark Search and select the Business Features drop-down menu. Then click on the following tags:	
Manufacturing process/ discrete	Health
The search returns the following list of knowledge nuggets (only the top 5 most relevant):	
Evidence of business process performance - Production quality (Manufacturing)	Genomic Pipeline pattern
Evidence of business process performance - Service innovation (Manufacturing)	Healthcare - Patient monitoring Blueprint
Qualitative Benchmarks for the Manufacturing Industry	Qualitative Benchmarks for the Healthcare Industry
Quantitative Benchmarks for the Manufacturing Industry	Quantitative Benchmarks for the Healthcare Industry
Quantitative Benchmarks for the Top 3 Use Cases in Manufacturing	Quantitative Benchmarks for the Top 3 Use Cases in Healthcare

Table 10: Results from Guided Benchmark Search filtering by Business Features for the DataBio Project

4.6. Track&Know

4.6.1. Project description

The Track & Know[8] (2018-2022) Big Data PPP project researches, develops and exploits modern software frameworks that aim to increase the efficiency of Big Data.

A variety of toolboxes (that contain specific methods/functions/algorithms for various types of data aggregation, manipulation and further analysis) are developed within the project and integrated in a software platform.

A Big Data Processing (BDP) toolbox is developed to implement data acquisition technology that captures data from heterogeneous data sources. The BDP toolbox extends the current solutions and delivers a tool for efficient access, indexing, partitioning and load balancing for Big spatiotemporal data.

A Complex Event Recognition (CER) toolbox detects complex event occurrences by analysing patterns in simple events. To do that, it uses contextual information and results from the Big Data Analytics (BDA) toolbox. For example: the toolbox may infer a complex event (such as dangerous driving or non-economical driving) by analysing patterns based on vehicle speed, direction, driver events, fuel consumption and other contextual information such as weather etc.

The BDA toolbox is developed to analyse heterogeneous data and to draw conclusions about the spatiotemporal distribution of mobility patterns. The BDA toolbox delivers scalable data mining techniques (such as clustering, sequence mining, hot-spot analysis) for voluminous offline and streaming trajectory data.

A Visual Analytics (VA) toolbox develops interactive and scalable methodologies to visualise data at all steps of analysis. The VA toolbox can efficiently handle both historical and streaming spatiotemporal data originating from different sources, with varying levels of resolution and quality.

The project integrates the toolboxes in a platform and tests them in pilot cases organized in the three domains (insurance sector, health sector and fleet management) with two common links: service optimisation and driver behaviour.

4.6.2. Project objectives and Architecture

The Track & Know – project focuses on five objectives:

Objective 1

- Create the Track & Know framework of best practices, technology implementation patterns and toolsets. This framework will map the end user industry analytics needs and corresponding datasets in transportation, health and insurance to the BDVA reference model.

Objective 2

- Produce a scalable, fault-tolerant platform for Big Data by collecting, integrating and processing streams of data

[8] https://trackandknowproject.eu/

- Build efficient, interoperable and scalable Track & Know Toolboxes*with Big Data Software stacks and integrate them into the Big Mobility Data Integrator
- Create toolsets for efficient distributed management, process mining querying and the visualisation of Big Data (BDP Toolbox)

Objective 3

- Produce the Track & Know Analytics Toolboxes. These Toolboxes include the real-time detection and forecasting of the Big Mobility Data Analytics Toolbox and the Predictive Complex Event Recognition Toolbox
- Produce the Track & Know Analytics Toolboxes:
  1. The real-time detection and forecasting of Big Mobility Data Analytics (BDA) Toolbox;
  2. The Predictive Complex Event Recognition (CER) Toolbox;
  3. And the Real-time Interactive Visual Analytics (VA) Toolbox.

Objective 4

- Test, validate and evaluate the Track & Know Toolboxes addressing different industrial domains, such as the automotive mobility, health and insurance

Objective 5

- Ensure scale up trough wide dissemination, exploitation actions, liaison, clustering and correlation with other European and large-scale pilots and projects.

To reach the proposed objectives, Big Data Research requires a multi-disciplinary approach. This approach will bring together research actors, customer’s demand and business field experts, provisioning infrastructure operators, and software industry actors in order to transform the voluminous and incomprehensible data into intelligence and knowledge.

A high level architecture of the Track&Know project is depicted below. This architecture fulfils Big data requirements, by considering the data diversity, volume, availability in terms of extremely large and complex collections and the detailed use case project scenarios. The architecture consists of:

- Data sources which represent the structured and unstructured data streams to be made available and be connected to the platform.
- Data store which represent the batch and interactive data sources that will be made available and will be connected to the platform.
- Connectors together with the Communication platform, that connect external Data sources and the Data store and make them available to the platform, Toolboxes and Pilots.
- Underlying Infrastructure providing all the necessary Big data tools.

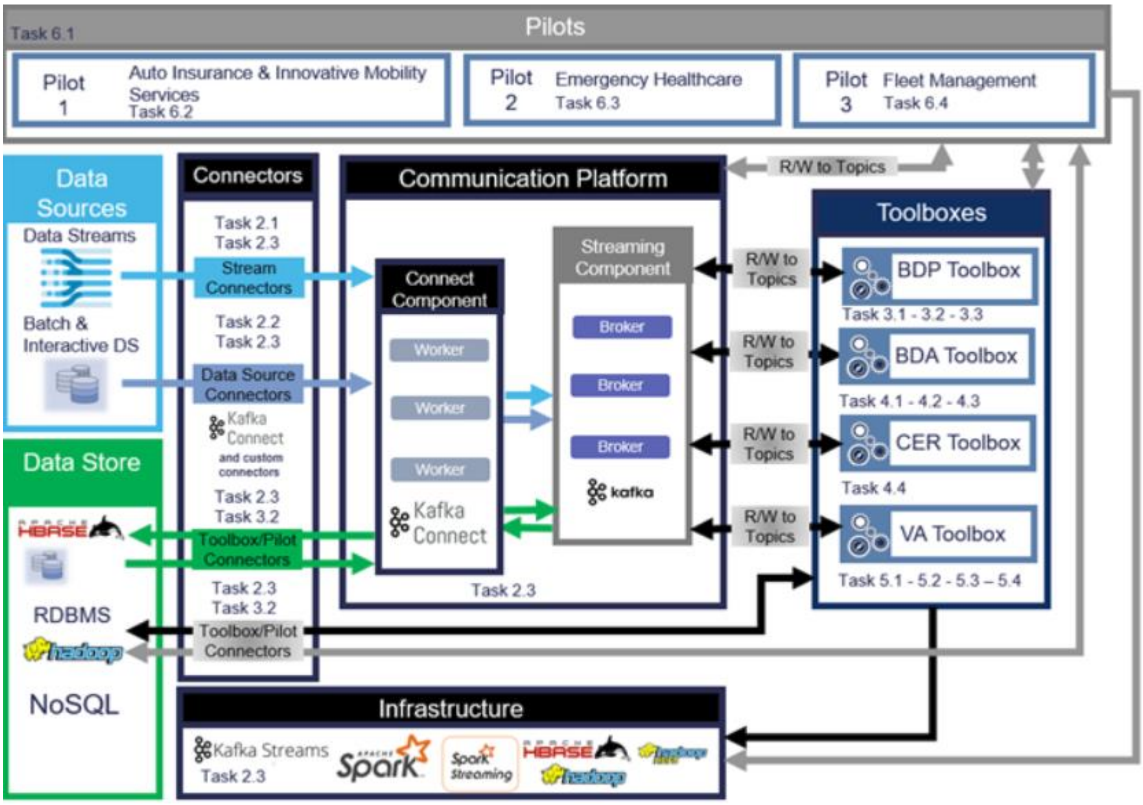
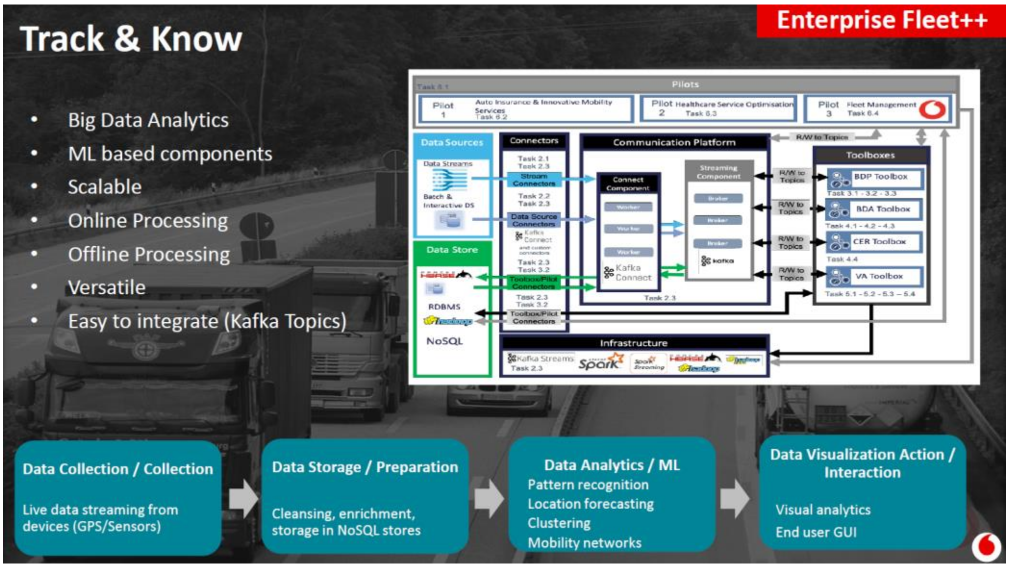


Figure 19: Track&Know High Level Architecture

4.6.3. Mapping to DataBench Pipeline

The following figure depicts the relationship of the various Track&Know steps with the steps of the DataBench pipeline.



Track & Know

Enterprise Fleet++

- Big Data Analytics
- ML based components
- Scalable
- Online Processing
- Offline Processing
- Versatile
- Easy to integrate (Kafka Topics)

Data Collection / Collection

Live data streaming from devices (GPS/Sensors)

Data Storage / Preparation

Cleansing, enrichment, storage in NoSQL stores

Data Analytics / ML

Pattern recognition
Location forecasting
Clustering
Mobility networks

Data Visualization Action / Interaction

Visual analytics
End user GUI

Figure 20: Track&Know steps related to DataBench Pipeline steps

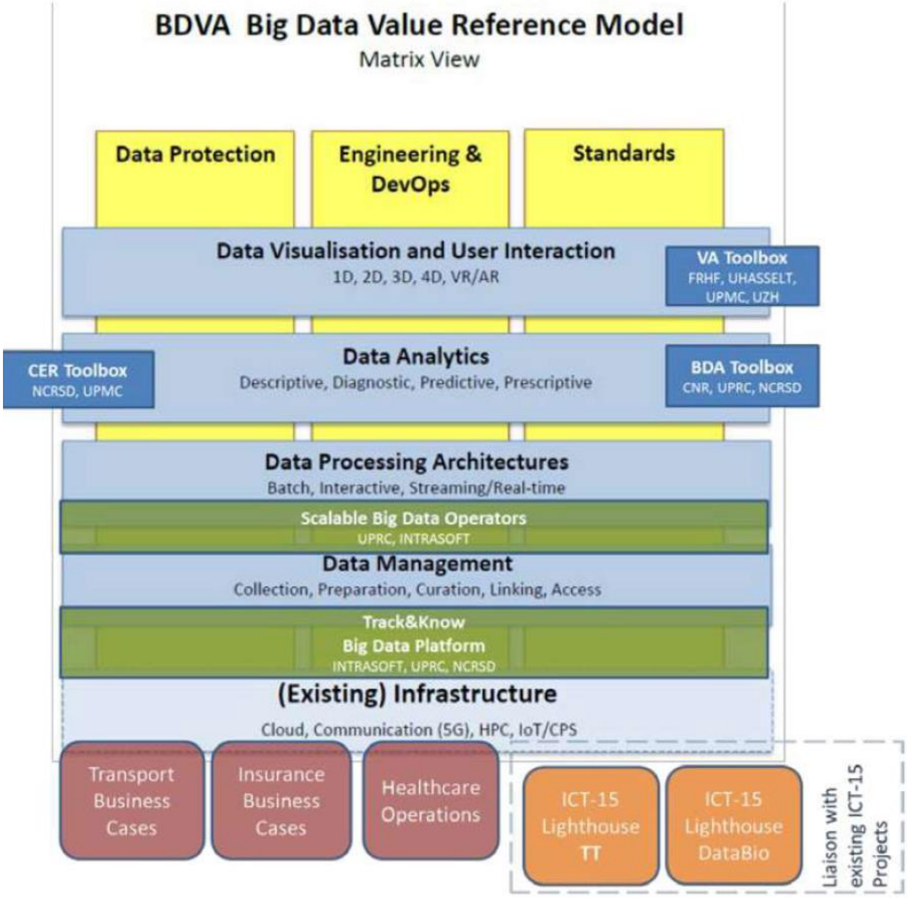

BDVA Big Data Value Reference Model
Matrix View
Data Protection
Engineering & DevOps
Standards
Data Visualisation and User Interaction
1D, 2D, 3D, 4D, VR/AR
VA Toolbox
FRHF, UHASSELT, UPMC, UZH
Data Analytics
Descriptive, Diagnostic, Predictive, Prescriptive
CER Toolbox
NCRSD, UPMC
BDA Toolbox
CNR, UPRC, NCRSD
Data Processing Architectures
Batch, Interactive, Streaming/Real-time
Scalable Big Data Operators
UPRC, INTRASOFT
Data Management
Collection, Preparation, Curation, Linking, Access
Track&Know
Big Data Platform
INTRASOFT, UPRC, NCRSD
(Existing) Infrastructure
Cloud, Communication (5G), HPC, IoT/CPS
Transport Business Cases
Insurance Business Cases
Healthcare Operations
ICT-15 Lighthouse TT
ICT-15 Lighthouse DataBio
Liaison with existing ICT-15 Projects

Figure 21: Components from Track&Know mapped into the BDV Reference Model

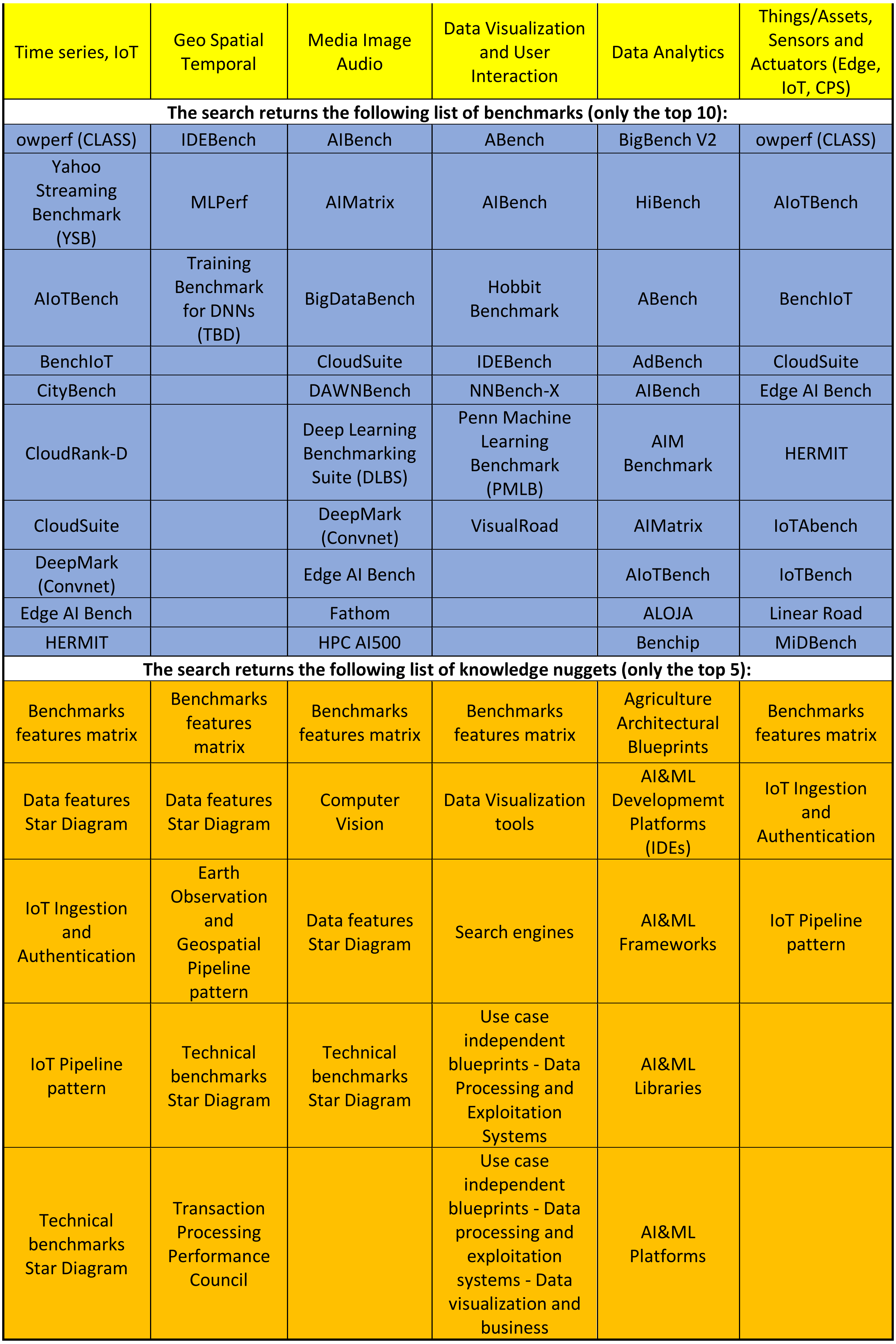
Figure 21: Components from Track&Know mapped into the BDV Reference Modelshows Components from Track&Know mapped into their respective areas in the BDV Reference Model, This kind of mapping has typically been done for all the BDV PPP projects. It is useful and important to highlight targeted focus areas, but it is also useful to have a complementary pipeline view showing how the components are connected, as introduced by the DataBench pipeline perspective introduced in D5.4.

4.6.4. Relevant Benchmarks and Nuggets from DataBench Toolbox

The section provides suggestions for relevant benchmarks and knowledge nuggets that include Big Data and AI technologies, architectural patterns and blueprints as well as business benchmarks for comparison. The following tables present results obtained from the different searches available in the DataBench Toolbox and illustrate its current results.

Table 10 lists the top 10 technical benchmarks (in blue) and the top 5 knowledge nuggets (in orange) resulting from BDV Reference Model search based on three data types and two Horizontal Layers filtered according to the Track&Know description provided above.

Go to Search --> BDV Reference Model and	
select one of the Data Types:	or select one of the Horizontal Layers:



Time series, IoT	Geo Spatial Temporal	Media Image Audio	Data Visualization and User Interaction	Data Analytics	Things/Assets, Sensors and Actuators (Edge, IoT, CPS)
The search returns the following list of benchmarks (only the top 10):					
owperf (CLASS)	IDEBench	AIBench	ABench	BigBench V2	owperf (CLASS)
Yahoo Streaming Benchmark (YSB)	MLPerf	AIMatrix	AIBench	HiBench	AIoTBench
AIoTBench	Training Benchmark for DNNs (TBD)	BigDataBench	Hobbit Benchmark	ABench	BenchIoT
BenchIoT		CloudSuite	IDEBench	AdBench	CloudSuite
CityBench		DAWNBench	NNBench-X	AIBench	Edge AI Bench
CloudRank-D		Deep Learning Benchmarking Suite (DLBS)	Penn Machine Learning Benchmark (PMLB)	AIM Benchmark	HERMIT
CloudSuite		DeepMark (Convnet)	VisualRoad	AIMatrix	IoTAbench
DeepMark (Convnet)		Edge AI Bench		AIoTBench	IoTBench
Edge AI Bench		Fathom		ALOJA	Linear Road
HERMIT		HPC AI500		Benchip	MiDBench
The search returns the following list of knowledge nuggets (only the top 5):					
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Agriculture Architectural Blueprints	Benchmarks features matrix
Data features Star Diagram	Data features Star Diagram	Computer Vision	Data Visualization tools	AI&ML Developmemt Platforms (IDEs)	IoT Ingestion and Authentication
IoT Ingestion and Authentication	Earth Observation and Geospatial Pipeline pattern	Data features Star Diagram	Search engines	AI&ML Frameworks	IoT Pipeline pattern
IoT Pipeline pattern	Technical benchmarks Star Diagram	Technical benchmarks Star Diagram	Use case independent blueprints - Data Processing and Exploitation Systems	AI&ML Libraries	
Technical benchmarks Star Diagram	Transaction Processing Performance Council		Use case independent blueprints - Data processing and exploitation systems - Data visualization and business	AI&ML Platforms	

			intelligence architecture		

Table 11: Results from BDV Reference Model Search by Data Type and Horizontal Layers for the Track&Know Project

Table 11 extends the previous search by using Guided Benchmark search to filter based on the three most popular data processing types: *stream (data-in-motion)*, *batch (data-at-rest)* and *interactive/ real-time*. Similar to the above table, here are depicted the top 10 technical benchmarks and top 5 knowledge nuggets resulting from the Toolbox search.

Go to Search --> Guided Benchmark Search and select the Platform and Architecture Features drop-down menu. Then click on the following tags:		
Stream (data-in-motion)	Batch (data-at-rest)	Interactive/(near) Real-time
The search returns the following list of benchmarks (only the top 10):		
BigBench V2	BigBench V2	Yahoo Streaming Benchmark (YSB)
HiBench	HiBench	Yahoo! Cloud Serving Benchmark (YCSB)
Yahoo Streaming Benchmark (YSB)	ABench	AMP Lab Big Data Benchmark
ABench	AMP Lab Big Data Benchmark	Berlin SPARQL Benchmark (BSBM)
AIM Benchmark	Berlin SPARQL Benchmark (BSBM)	BigFUN
BigDataBench	BigBench	CBench-Dynamo
CityBench	BigDataBench	CityBench
CloudRank-D	CALDA	Graphalytics
CloudSuite	CloudRank-D	IDEBench
DAWNBench	CloudSuite	LinkBench
The search returns the following list of knowledge nuggets (only the top 5):		
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix
Classification of data processing architectures	Classification of data processing architectures	Classification of data processing architectures
Data features Star Diagram	Data features Star Diagram	Data features Star Diagram
Message brokers and Pub-Sub	Project ALOJA	
Stream Processing engines		

Table 12: Results from Guided Benchmark Search filtering by Platform and Architecture Features for the Track&Know Project

Finally, Table 12 provides results related to the Business Features for the three related industries: Finance/ Bank/ Insurance, Smart City / Mobility and Health. For each of three categories are listed the most relevant 5 knowledge nuggets that provide useful comparison information as well as technical and architectural blueprints and technologies.

Go to Search --> Guided Benchmark Search and select the Business Features drop-down menu. Then click on the following tags:		
Finance/ Bank/ Insurance	Smart City/ Mobility	Health
The search returns the following list of knowledge nuggets (only the top 5 most relevant):		

Benchmarks features matrix	Benchmarks features matrix	Genomic Pipeline pattern
Financial services - Fraud detection Architectural Blueprints	Evidence of business process performance - Rail quality of service (Transport)	Healthcare - Patient monitoring Blueprint
Qualitative Benchmarks for the Financial Services Industry	Qualitative Benchmarks for the Transportation and Logistics Industry	Qualitative Benchmarks for the Healthcare Industry
Qualitative Benchmarks by Industry	Quantitative Benchmarks for the Transport and Logistics Industry	Quantitative Benchmarks for the Healthcare Industry
Securities Technology Analysis Center (STAC)	Transports and logistics - Fleet management and path optimization	Quantitative Benchmarks for the Top 3 Use Cases in Healthcare

Table 13: Results from Guided Benchmark Search filtering by Business Features for the Track&Know Project

4.7. CLASS

4.7.1. Project description

CLASS[9] (2018-2020) - Edge and Cloud Computation: A Highly distributed Software for Big Data Analytics is one of the Big Data PPP projects.

Current trends towards the use of big data technologies in the context of smart cities suggest the need for developing novel software development ecosystems upon which advanced mobility functionalities can be developed. CLASS is creating a novel software architecture that allows users to develop and execute advanced big-data applications in an efficient way. The goal of this new software infrastructure is to allow collecting, storing, analysing and processing a vast amount of geographically-distributed data, in order to transform it into valuable knowledge for the public sector, private companies and citizens.

CLASS aims to develop a novel software architecture framework to help big data developers to efficiently distributing data analytics workloads along the compute continuum (from edge to cloud) in a complete and transparent way while providing sound real-time guarantees. This ability opens the door to the use of big data into critical real-time systems, providing them with superior data analytics capabilities to implement more intelligent and autonomous control applications.

Applying big-data technologies to smart cities field applications entails many challenges: from processing data across the compute continuum (from edge to cloud), to predicting real-time responses, and employing a programming model that can mix different application program interfaces (APIs) and models. The CLASS platform is facing these needs by integrating technologies from different computing domains into a single development framework. This allows to efficiently combine data-in-motion and data-at-rest analytics along the compute continuum, while providing real time guarantees.

The capabilities of the CLASS framework is being demonstrated on a real smart-city use case in the City of Modena, featuring a heavy sensor infrastructure to collect real-time data across a wide urban area, and three connected vehicles equipped with heterogeneous sensors/actuators and V2X connectivity to enhance the driving experience.

[9] https://class-project.eu/

4.7.2. Project objectives and Architecture

Features:

- Innovative parallel and distributed programming models and architectures from the high-performance domain
- Timing analysis methods and energy-efficient parallel architectures from the real-time embedded domain.
- Advanced data analytics platforms and programming models from the big-data domain

Areas of Application:

Smart Cities, Connected cars, Urban mobility. In addition to those areas of application, CLASS architecture can be applied to all domains with critical real-time needs.

Market Trends and Opportunities:

CLASS architecture is a solution for Smart Cities and Automotive Smart Areas (ASA), which provides the ASA city-awareness environment needed to support the huge performance requirements of the real-time big data analytics methods needed.

Customer Benefits:

CLASS aims to develop a novel software architecture to help programmers and big data practitioners to combine data-in-motion and data-at-rest analysis, by efficiently distributing data and process mining along the compute continuum, while providing real-time guarantees

Technological novelty:

- A software architecture ecosystem for distributing big-data workloads along the compute continuum while providing real-time guarantees
- Development of data analytics workloads combining data-in-motion and data-at-rest that benefits from distribution capabilities
- A novel distributed sensing/computing infrastructure for the development of advanced urban mobility applications with data analytics and real-time requirements

The CLASS software architecture components can be found in the dedicated CLASS github channel and they include:

- Data Analytics Platform
- Computation Coordination and Distribution Framework
- Cloud Layer
- Edge Layer
- Data Analytics Methods

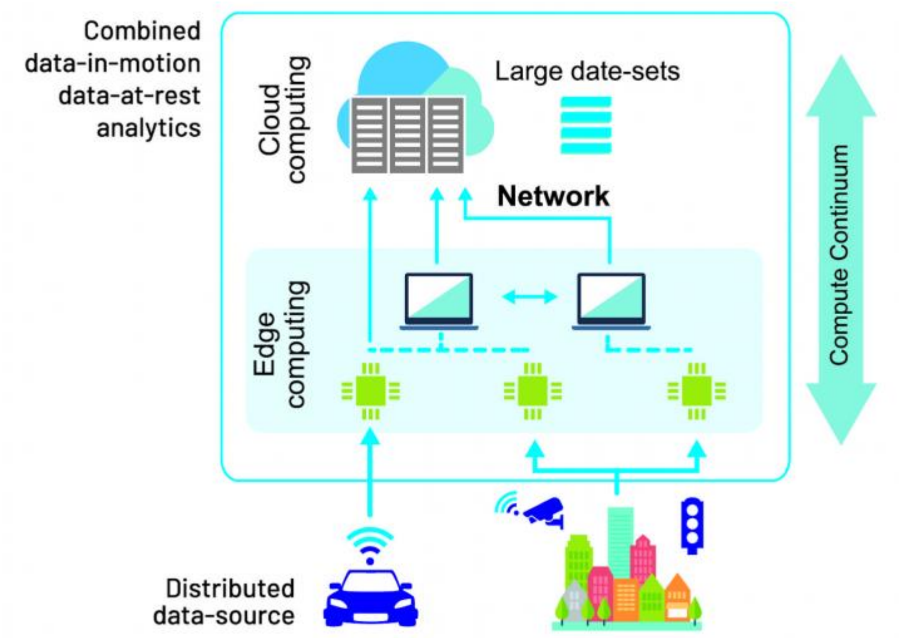


Figure 22: Overview of Edge and Cloud in the CLASS project

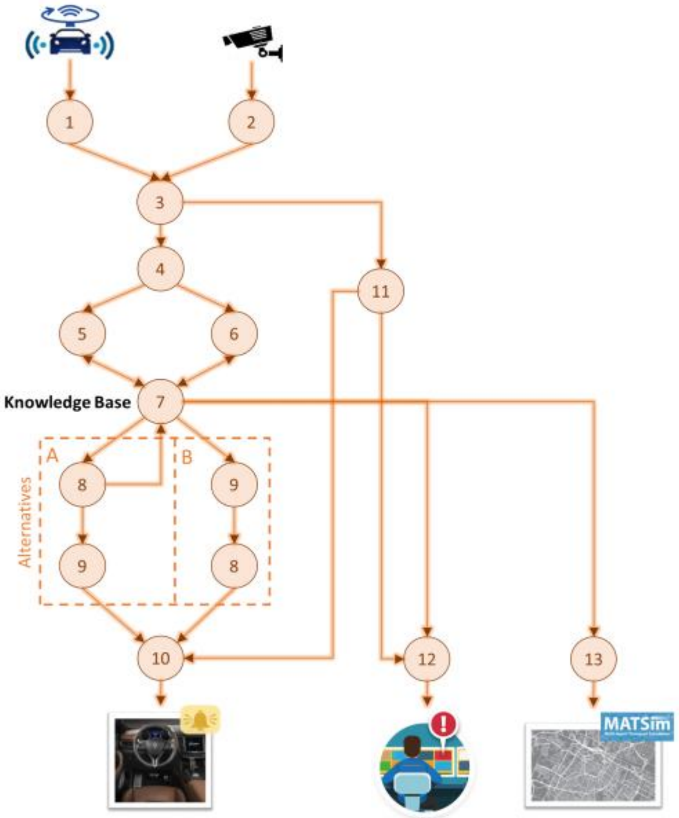


Figure 23: Graphical overview of the main CLASS pipeline

The description of the numbers for each node in Figure 24 can be seen in the Pipeline mapping in Figure 25.

4.7.3. Mapping to DataBench Pipeline and Blueprints

The CLASS pipeline illustrated in Figure 23: Graphical overview of the main CLASS pipeline is mapped to the DataBench Pipeline in Figure 24: Mapping of CLASS activities to the steps of the DataBench pipeline. Behind each of the steps there is one or more components supporting the functionality of the step.

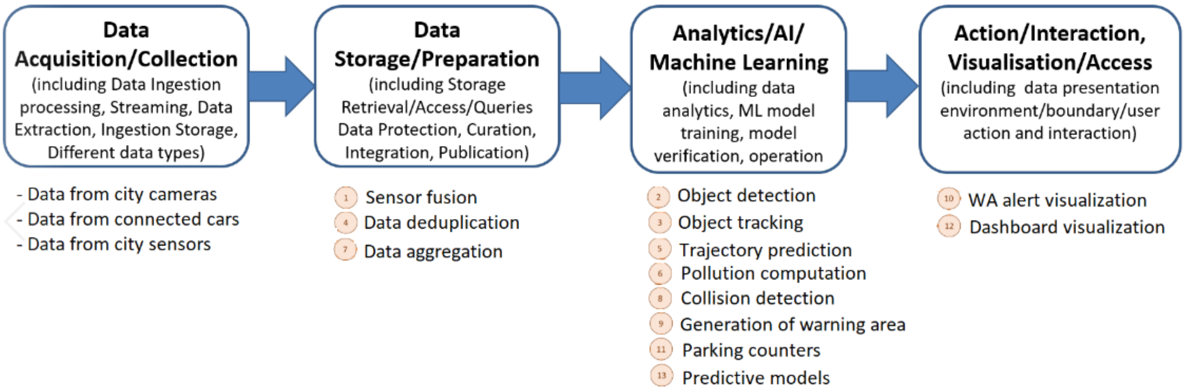

Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Analytics/AI/ Machine Learning
(including data analytics, ML model training, model verification, operation
Action/Interaction, Visualisation/Access
(including data presentation environment/boundary/user action and interaction)
- Data from city cameras
- Data from connected cars
- Data from city sensors
1 Sensor fusion
4 Data deduplication
7 Data aggregation
2 Object detection
3 Object tracking
5 Trajectory prediction
6 Pollution computation
8 Collision detection
9 Generation of warning area
11 Parking counters
13 Predictive models
10 WA alert visualization
12 Dashboard visualization

Figure 24: Mapping of CLASS activities to the steps of the DataBench pipeline

The detailed pipeline connection for the numbers in Figure 25 can be seen in Figure 24.

4.7.4. Relevant Benchmarks and Nuggets from DataBench Toolbox

The section provides suggestions for relevant benchmarks and knowledge nuggets that include Big Data and AI technologies, architectural patterns and blueprints as well as business benchmarks for comparison. The following list of results demonstrates the different searches available in the DataBench Toolbox and illustrates its current results.

Table 13 lists the top 10 technical benchmarks (in blue) and the top 5 knowledge nuggets (in orange) resulting from BDV Reference Model search based on three data types and two Horizontal Layers filtered following the CLASS project features provided above.

Go to Search --> BDV Reference Model and				
select one of the Data Types:			or select one of the Horizontal Layers:	
Time series, IoT	Geo Spatial Temporal	Media Image Audio	Cloud and High Performance computing (HPC)	Things/Assets, Sensors and Actuators (Edge, IoT, CPS)
The search returns the following list of benchmarks (only the top 10):				
owperf (CLASS)	IDEBench	AIBench	CloudSuite	owperf (CLASS)
Yahoo Streaming Benchmark (YSB)	MLPerf	AIMatrix	GARDENIA	AIoTBench
AIoTBench	Training Benchmark for DNNs (TBD)	BigDataBench	HPC AI500	BenchIoT

BenchIoT		CloudSuite	MLBench Services	CloudSuite
CityBench		DAWNBench	PUMA Benchmark Suite	Edge AI Bench
CloudRank-D		Deep Learning Benchmarking Suite (DLBS)	TPCx-V	HERMIT
CloudSuite		DeepMark (Convnet)		IoTAbench
DeepMark (Convnet)		Edge AI Bench		IoTBench
Edge AI Bench		Fathom		Linear Road
HERMIT		HPC AI500		MiDBench
The search returns the following list of knowledge nuggets (only the top 5):				
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix
Data features Star Diagram	Data features Star Diagram	Computer Vision	Technical benchmarks Star Diagram	IoT Ingestion and Authentication
IoT Ingestion and Authentication	Earth Observation and Geospatial Pipeline pattern	Data features Star Diagram	The DataBench Framework and the BDV Reference Model	IoT Pipeline pattern
IoT Pipeline pattern	Technical benchmarks Star Diagram	Technical benchmarks Star Diagram		
Technical benchmarks Star Diagram	Transaction Processing Performance Council			

Table 14: Results from BDV Reference Model Search by Data Type and Horizontal Layers for the CLASS project

In the next table we present the results from a Guided Benchmark search based on two Platform and Architecture Features such as stream and batch. In blue are listed the top 10 technical benchmarks and in orange the top 5 knowledge nuggets. Additionally, the third column provides the 5 most relevant knowledge nuggets for the smart city/ mobility domain that can be used for business benchmark comparison.

Go to Search --> Guided Benchmark Search and select		
the Platform and Architecture Features drop-down menu. Then click on the following tags:		**the Business Features drop-down menu and click on:**
Stream (data-in-motion)	Batch (data-at-rest)	Smart City/ Mobility
The search returns the following list of benchmarks (only the top 10):		
BigBench V2	BigBench V2	
HiBench	HiBench	
Yahoo Streaming Benchmark (YSB)	ABench	
ABench	AMP Lab Big Data Benchmark	
AIM Benchmark	Berlin SPARQL Benchmark (BSBM)	

BigDataBench	BigBench	
CityBench	BigDataBench	
CloudRank-D	CALDA	
CloudSuite	CloudRank-D	
DAWNBench	CloudSuite	
The search returns the following list of knowledge nuggets (only the top 5):		
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix
Classification of data processing architectures	Classification of data processing architectures	Evidence of business process performance - Rail quality of service (Transport)
Data features Star Diagram	Data features Star Diagram	Qualitative Benchmarks for the Transportation and Logistics Industry
Message brokers and Pub-Sub	Project ALOJA	Quantitative Benchmarks for the Transport and Logistics Industry
Stream Processing engines		Transports and logistics - Fleet management and path optimization

Table 15: Results from Guided Benchmark Search filtering by Platform and Architecture and Business Features for the CLASS project

As part of the CLASS project was developed a customized benchmark called owperf (CLASS) [45] which is also available and integrated in the DataBench Toolbox under the owperf (CLASS) benchmark [46].

5. Use case Evaluation of the DataBench Blueprints

As discussed in D4.3, the DataBench case study analysis has found that in some industries/use cases technology costs can be so high to significantly reduce business benefits. This potentially negative economic impact of BDTs can be mitigated with a careful selection of technologies based on quantitative, objective technical performance benchmarks. From this perspective, the blueprints can be used to support the identification of the most critical technical components for different use cases and obtain an estimate of infrastructural costs before committing to a specific technology and/or starting implementation.

In WP5, we have tested the effectiveness of the cost-assessment methodology described in D4.3 on three use cases. Results are reported below.

5.1. Agriculture, Heavy Equipment optimization

A data-intensive component in this use case is the NoSQL storage. Several performance test on Apache Cassandra, Couchbase, HBase and MongoDB NoSQL stores by using YCSB benchmark are described in [47]. These tests have been performed on Amazon Web Services EC2 instances using 2.xlarge class of instances (30.5 GB RAM, 4 CPU cores, and a single volume of 800 GB of SSD local storage) for the database nodes and c3.xlarge class of instances (7.5GB RAM, 4 CPU cores) to drive the test activity. A careful DBMS choice can be performed by using benchmarking results. As discussed in [47], with a balanced read-write workload, which is the likely workload in our use case, there are significant differences among different DBMS technologies (see, for example, Figure 25: NoSQL benchmarking - Balanced Read/Write mix, throughput.).

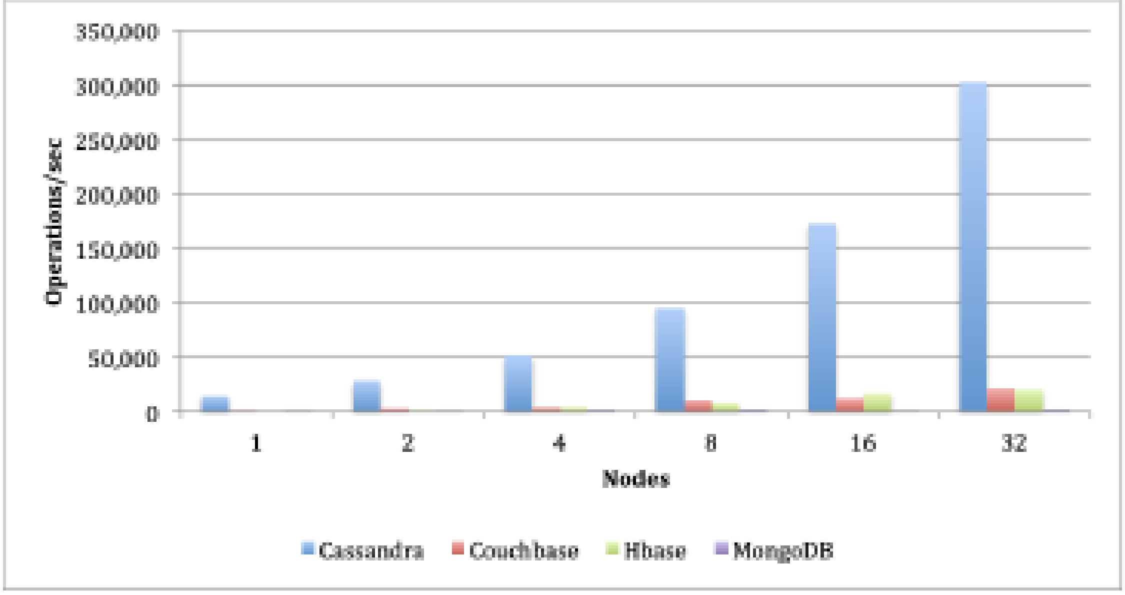


Figure 25: NoSQL benchmarking - Balanced Read/Write mix, throughput.

For our use case, let us consider a scenario where the DBMS needs to process 10K operations per second, balanced between read and write (see also [48] for a more in-depth discussion on sizing the infrastructure in this use case). By hosting the infrastructure on AWS, Cassandra would need just one master node and one worker node for a total price of 0.58$/h. Reaching a sufficient throughput performance would require 16 worker nodes for Couchbase, for a total cost of 3.5655$/h, about six times the cost for the infrastructure required by Cassandra. This represents a cost gap that is certainly significant, especially for a medium-size enterprise and indicates that the use of technical benchmarking can lead to crucially important savings.

5.2. Smart manufacturing

In this use case, we focus on real-time analytics and the choice of a distributed streaming data platform to collect IoT data from large production plants. Apache Storm, Apache Flink and Spark Streaming are candidate technologies for this processing-intensive component. These technologies have been benchmarked in various contexts. We base our analyses on results reported in [49] and summarized in Figure 27.

Let us assume that our smart manufacturing use case requires a throughput of 1 million tuple/s (see, also, [48]). This workload seems reasonable for many smart manufacturing applications, especially when sensors are placed inside manufacturing products. According to these benchmarking tests, to reach a throughput greater than 1 million tuples/s Apache Storm requires eight nodes against one required by Apache Flink. In greater detail, Flink needs one master node and one worker node, which on AWS would cost 0.153$/h. A similar performance could be reached by Storm by relying on 8 worker nodes for a total cost of 0.51$/h, more than three time the cost of Flink. Once again, this cost gap is significant and justifies the need for a careful technology selection based on technical benchmarking.

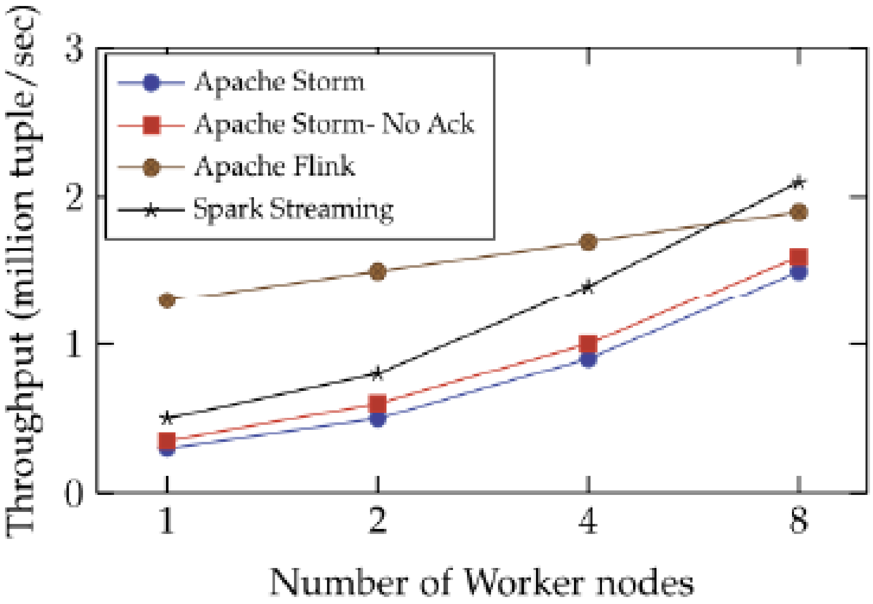


Figure 26: Benchmarking results from [3].

5.3. Healthcare, diagnostic systems

Diagnostic systems can involve the use of deep learning frameworks, such as CNTK, Caffe 2, MXNet and Pytorch in the context of computer vision. These technologies have been benchmarked in [50]. The test was aimed at creating a “Rosetta Stone” of deep-learning frameworks. Deep learning tasks were performed on different configurations of Azure Deep Learning Virtual Machines. Among the results obtained, it is relevant to focus on the evaluation of two different deep learning tasks regarding computer vision:

1) Training time. A convolutional neural network (VGG style, 32 bit) was trained on CIFAR-10 dataset containing 50K images and 10K test images uniformly split among 10 classes.

2) Average time of feature extraction. A pre-trained ResNet 50 model was used to perform the feature extraction of 1000 images and the overall processing time was measured.

Regarding training, no evidence was found of a significant performance gap among the frameworks under benchmarking. In contrast, feature extraction has a considerably different performance across technologies. For example, Caffe 2 is found to perform worse than all the other deep learning frameworks and the choice of Caffe 2 instead of MXNet (the best performing framework) would have brought to processing from 2 to 5 times higher.

There is evidence of usage of feature extraction in medical diagnosis use cases (e.g. in Alzheimer’s disease diagnosis). In order to give a broad estimation of the cost impact of technical choices, let us assume to be in a medical diagnosis use case for which feature extraction of 7500 pictures/minute needs to be performed by using a pre-trained ResNet50 model. With this workload, MXNet framework would require half the processing capacity compared to Caffe2 framework and, hence, would reduce infrastructural costs by 50%.

6. Project data analysis in the DataBench Observatory

As DataBench deliverable D5.2 (Final evaluation of DataBench metrics) describes a methodology for developing DataBench observatory tool, along with DataBench popularity index and a number of dynamically obtained metrics.

In this DataBench deliverable D5.5 (Final report on methodology for evaluation of industrial analytic projects scenarios) we describe the project data analysis (with specific accent on assessment Artificial Intelligence, Big Data and Benchmarking).

The objectives of project data analysis are:

- to establish a dynamic process that allows for automatic analysis of Cordis open project data, is easily maintained and updatable (and can be supported after the end of DataBench project);
- to explore the results of project data analysis, in particular – to define the project component in the DataBench popularity index;
- to observe the ranking and trends for topics, tools and technologies within projects open data;
- to assess project related metrics;
- to provide visualizations and user interaction for EU projects open data within DataBench observatory tool.

Below we proceed with description of project data, provide insights into project data analysis and Cordis projects assessment – and also the ongoing analysis of Big Data PPP projects.

6.1. Big Data PPP Project Data Description

Data is now recently becoming available for the 60 + projects under the Big Data PPP Umbrella. Similar to what is described next on the analysis of project data from Cordis, we are now also initiating a detailed analysis of project data available from the Big Data PPP projects – to become available through the DataBench Observatory.

In particular, the Big Data PPP projects analysis is being performed on the information about projects obtained from the several data sources. The sources of project data from the Big Data PPP projects is as follows:

- BDVe projects [51];
- Individual project websites (referred to in the web reference above);
- EU funded projects on data [52];
- Publicly available project deliverables and publications;
- Solutions from BDVe Marketplace [53]

The goal of the undergoing Big Data PPP project analysis is to tune the methodology related to the development and implementation of the DataBench observatory (described in

DataBench deliverable D5.2) particularly for specified sources and to establish an automatic process of producing analytical outcomes.

6.2. Cordis Project Data Description

Cordis open project data is one of the data sources used for developing DataBench popularity index.

In particular, the dataset accounts for over 30500 H2020 projects and over 25700 FP7 projects obtained from EU open data portal[10],[11]. The data provides publicly available information for each project, such as: project ID (grant agreement number), project acronym, project status, funding programme, topic, project title, project start date and project end date, objective, total cost, EC max contribution, call ID, funding scheme (type of action), coordinator, coordinator country, participants, participant countries etc.

Figure 28 presents a metadata example for the ongoing ADIMITTED project, which intends to adopt a complex hardware architecture to support big data analysis, to involve algorithms for data correlation, time series management and statistical analysis in aviation domain.

[10] https://data.europa.eu/euodp/sl/data/dataset/cordisH2020projects

[11] https://data.europa.eu/euodp/sl/data/dataset/cordisfp7projects

```
{ "projectUrl": "http:\/\/www.admitted-project.eu\/",
  "coordinator": "TXT E-SOLUTIONS SPA",
  "acronym": "ADMITTED",
  "endDate": "2023-11-30",
  "topics": "JTI-CS2-2018-CfP08-FRC-01-18",
  "subjects": [],
  "coordinatorCountry": "IT",
  "title": "Advanced Data Methods for Improved Tiltrotor Test and Design",
  "objective": "Flight testing is an important phase during the development of an aircraft to validate the design.
During flight, data is gathered and design problems are identified and solved. The collected data are fundamental for
the analysis and Aircraft are properly instrumented to generate large amounts of information. Such huge amount of
data needs to be properly evaluated and traditional methods and platforms are no more effective. Flight testing is a
significant cost contributor to the aircraft production life cycle and is still extensively deployed. Flight test
programmes take several years and more prototypes are built to reduce lead times. Strong adherence to rigour
safety and certification requirements and generally unchanged circular advisories inhibit the potential improvement
of flight test designs. Innovative algorithms and statistical estimation are not achieving its full potential in the
industrialized flight testing environment. The methods in this proposal increase the quality and productivity of an
experiment, leading to a required test point reduction or increased predictive capabilities. The purpose of this
project is to define and implement a state-of-the-art platform able to support data analysis. This is achieved by
adopting a complex hardware architecture to support big data analysis and implementing specific algorithms to
support data correlation, time series management and statistical analysis. Furthermore, to support flight test
engineers, novel approaches based on machine learning are provided to support the technicians in detecting specific
flight conditions. The same platform is also adapted to support the development of the Next Generation Civil Tilt
Rotor Technology Demonstrator.",
  "call": "H2020-CS2-CFP08-2018-01",
  "participantCountries": ["NL","CH"],
  "fundingScheme": "CS2-RIA",
  "ecMaxContribution": "1718330",
  "id": "832003",
  "rcn": "220993",
  "programme": "H2020-EU.3.4.5.3.",
  "frameworkProgramme": "H2020",
  "startDate": "2019-02-01",
  "totalCost": "1718330",
  "status": "SIGNED",
  "participants": [
    "STICHTING NATIONAAL LUCHT- EN RUIMTEVAARTLABORATORIUM",
    "SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA"]}
```

Figure 27: Example of Project Metadata

7.1. Cordis Project Data Analysis

For Cordis project data analysis we have performed several tasks, including semantic annotation with Wikipedia concepts and tagging with DataBench ontology (described in the deliverable D5.2), normalization and ranking of topics, tools and technologies, detecting trends in topics, tools and technologies and preparing visualizations that reflect the project data analysis.

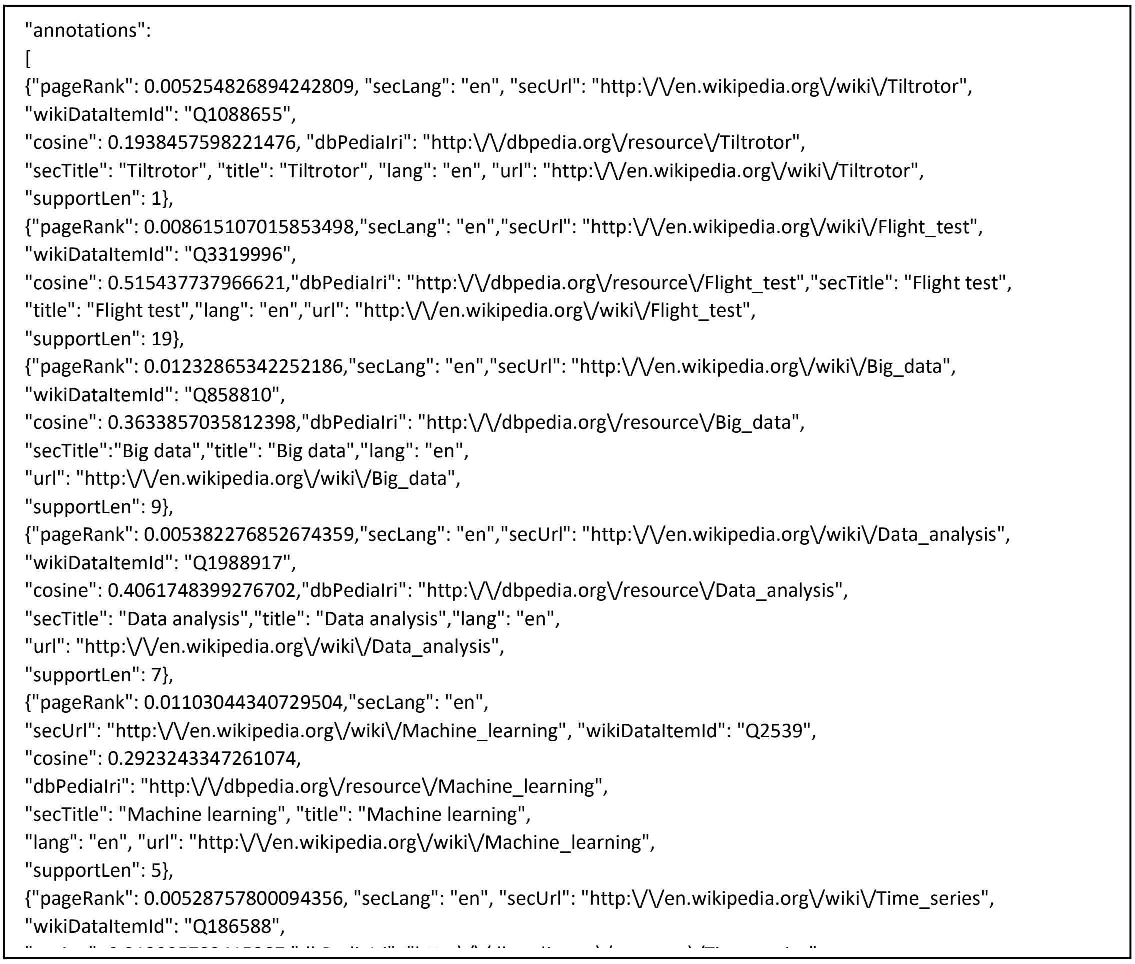
Since DataBench deliverable D5.2 in details describes the methodology for establishing DataBench popularity index and Cordis project data present one of the index component, in this deliverable we are going to provide a brief description of semantic annotation task with examples relevant to project data analysis and assess the analysis results.

As mentioned above, semantic annotation of Cordis project data can be viewed as a task of augmenting the unstructured textual information such as project title, project objective (see

Figure 28: Example of Project Metadata) with Wikipedia concepts, and then tagging with concepts from DataBench ontology.

In the process of wikification the pagerank is computed over the constructed mention-concept graph and a set of relevant concepts is defined. The wikification approach supports operating with multiple languages (currently available in Wikipedia).

In Figure 29: Project Description Semantic Annotation (pageRank>0.005, cosine > 0.2) we present an example of wikification of the project text example above. The wikification is performed with JSI Wikifier tool[12] - a web service that takes a text document as input and annotates it with links to relevant Wikipedia concepts.



```
"annotations":
[
{"pageRank": 0.005254826894242809, "secLang": "en", "secUrl": "http:\/\/en.wikipedia.org\/wiki\/Tiltrotor",
"wikiDataItemId": "Q1088655",
"cosine": 0.1938457598221476, "dbPediaIri": "http:\/\/dbpedia.org\/resource\/Tiltrotor",
"secTitle": "Tiltrotor", "title": "Tiltrotor", "lang": "en", "url": "http:\/\/en.wikipedia.org\/wiki\/Tiltrotor",
"supportLen": 1},
{"pageRank": 0.008615107015853498,"secLang": "en","secUrl": "http:\/\/en.wikipedia.org\/wiki\/Flight_test",
"wikiDataItemId": "Q3319996",
"cosine": 0.515437737966621,"dbPediaIri": "http:\/\/dbpedia.org\/resource\/Flight_test","secTitle": "Flight test",
"title": "Flight test","lang": "en","url": "http:\/\/en.wikipedia.org\/wiki\/Flight_test",
"supportLen": 19},
{"pageRank": 0.01232865342252186,"secLang": "en","secUrl": "http:\/\/en.wikipedia.org\/wiki\/Big_data",
"wikiDataItemId": "Q858810",
"cosine": 0.3633857035812398,"dbPediaIri": "http:\/\/dbpedia.org\/resource\/Big_data",
"secTitle":"Big data","title": "Big data","lang": "en",
"url": "http:\/\/en.wikipedia.org\/wiki\/Big_data",
"supportLen": 9},
{"pageRank": 0.005382276852674359,"secLang": "en","secUrl": "http:\/\/en.wikipedia.org\/wiki\/Data_analysis",
"wikiDataItemId": "Q1988917",
"cosine": 0.4061748399276702,"dbPediaIri": "http:\/\/dbpedia.org\/resource\/Data_analysis",
"secTitle": "Data analysis","title": "Data analysis","lang": "en",
"url": "http:\/\/en.wikipedia.org\/wiki\/Data_analysis",
"supportLen": 7},
{"pageRank": 0.01103044340729504,"secLang": "en",
"secUrl": "http:\/\/en.wikipedia.org\/wiki\/Machine_learning", "wikiDataItemId": "Q2539",
"cosine": 0.2923243347261074,
"dbPediaIri": "http:\/\/dbpedia.org\/resource\/Machine_learning",
"secTitle": "Machine learning", "title": "Machine learning",
"lang": "en", "url": "http:\/\/en.wikipedia.org\/wiki\/Machine_learning",
"supportLen": 5},
{"pageRank": 0.00528757800094356, "secLang": "en", "secUrl": "http:\/\/en.wikipedia.org\/wiki\/Time_series",
"wikiDataItemId": "Q186588",
```

Figure 28: Project Description Semantic Annotation (pageRank>0.005, cosine > 0.2)

It is possible to notice, that wikification has captured the main concepts from the unstructured project textual data, such as “Tiltrotor”, “Flight test”, “Big data”, “Data analysis”, “Time series” etc.

These concepts (when aligned with DataBench ontology), represent the topics, tools and technologies associated with particular project. Since each Cordis project also contains start and end date, the project topics, tools and technologies become part of DataBench popularity index (monthly and overall), as well as represented in trends.

[12] http://www.wikifier.org

7.2. Cordis Projects Assessment

In this section we present visualization snapshots for Cordis project assessment, and we discuss the obtained results. DataBench observatory tool is an application intended for exploring topic, tool and technology popularity, visibility and trends through different data sources (academic, such as Microsoft Academic Graph, research and development, such as Cordis EU projects, industry, such as job advertisements, GitHub, general popularity, such as Google Trends etc.)

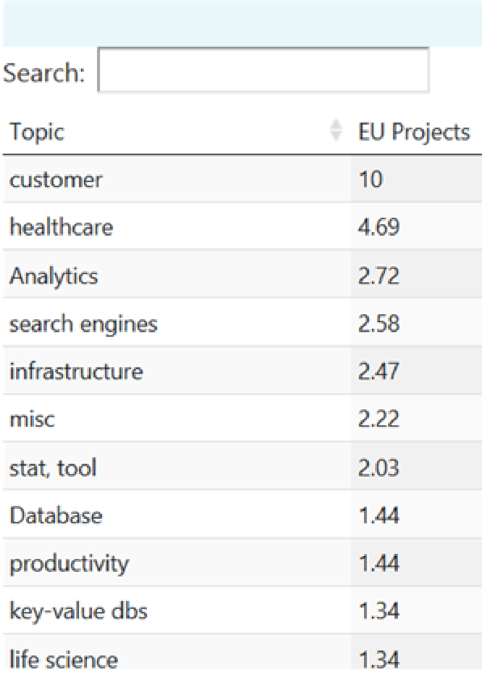
The methodology behind DataBench observatory and the DataBench popularity index measure is described in detail in deliverable D5.2.

DataBench observatory presents a user a possibility to interact with a tool:

- to select index components, presentation parameters,
- to sort topics, tools and technologies as well as
- to search specific topic, tool and technology, explore the ranking components,
- to add the selected topic, tool or technology to the trends and observe its trend.

Figure 30: Popular Topics for Projects (from DataBench Observatory tool, overall) presents a list of most popular topics from Cordis data. The topics notated with small letters represent the extended categories for tools and technologies. The numbers are normalized (from 1 to 10) and it is possible to notice that EU projects are very customer oriented, as well as data oriented. There are a lot of projects from healthcare and life sciences domains and topics related to databases, analytics, productivity, infrastructure, search engines.

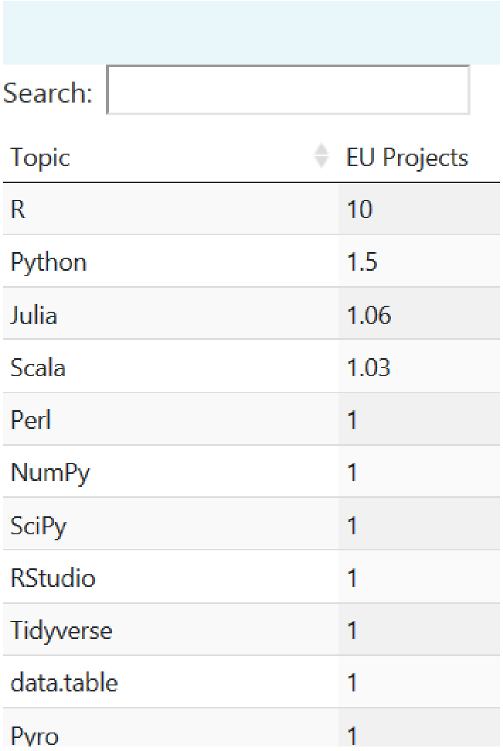
Figure 31: Popular Tools and Technologies for Projects (from DataBench Observatory tool, Category: stat_tool, July 2020) presents popular tools and technologies from July 2020 in statistical tools category. It shows that R dominates in this category, along with several programming languages allowing for statistical calculations, such as Julia, Scala, Python etc.



Search:

Topic	EU Projects
customer	10
healthcare	4.69
Analytics	2.72
search engines	2.58
infrastructure	2.47
misc	2.22
stat, tool	2.03
Database	1.44
productivity	1.44
key-value dbs	1.34
life science	1.34

Figure 29: Popular Topics for Projects (from DataBench Observatory tool, overall)

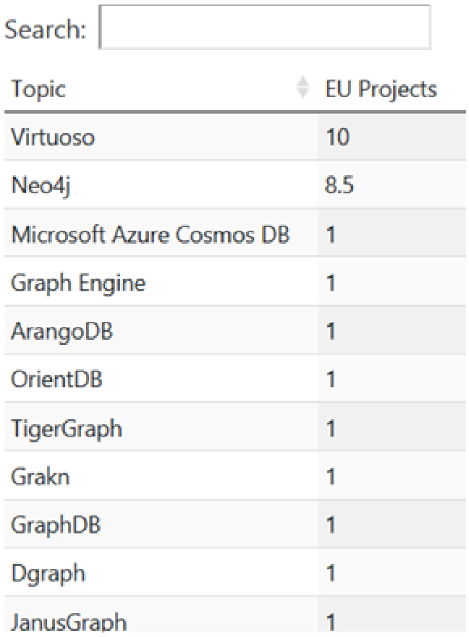


Search:

Topic	EU Projects
R	10
Python	1.5
Julia	1.06
Scala	1.03
Perl	1
NumPy	1
SciPy	1
RStudio	1
Tidyverse	1
data.table	1
Pyro	1

Figure 30: Popular Tools and Technologies for Projects (from DataBench Observatory tool, Category: stat_tool, July 2020)

The most popular tools in the category graph databases within Cordis data are displayed at Figure 32: Popular Tools and Technologies for Projects (from DataBench Observatory tool, Category: graph database, overall). Such databases, as Virtuoso and Neo4j are visibly active in this category.

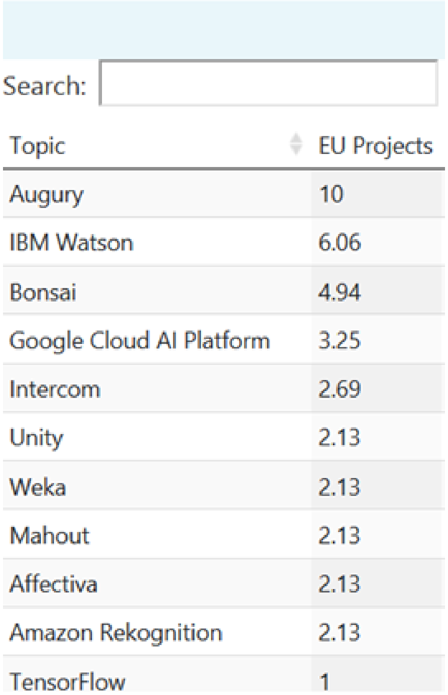


Search:

Topic	EU Projects
Virtuoso	10
Neo4j	8.5
Microsoft Azure Cosmos DB	1
Graph Engine	1
ArangoDB	1
OrientDB	1
TigerGraph	1
Grakn	1
GraphDB	1
Dgraph	1
JanusGraph	1

Figure 31: Popular Tools and Technologies for Projects (from DataBench Observatory tool, Category: graph database, overall)

The machine learning tool and technologies are shown below at Figure 32: Popular Tools and Technologies for Projects (from DataBench Observatory tool, Category: machine learning, overall).



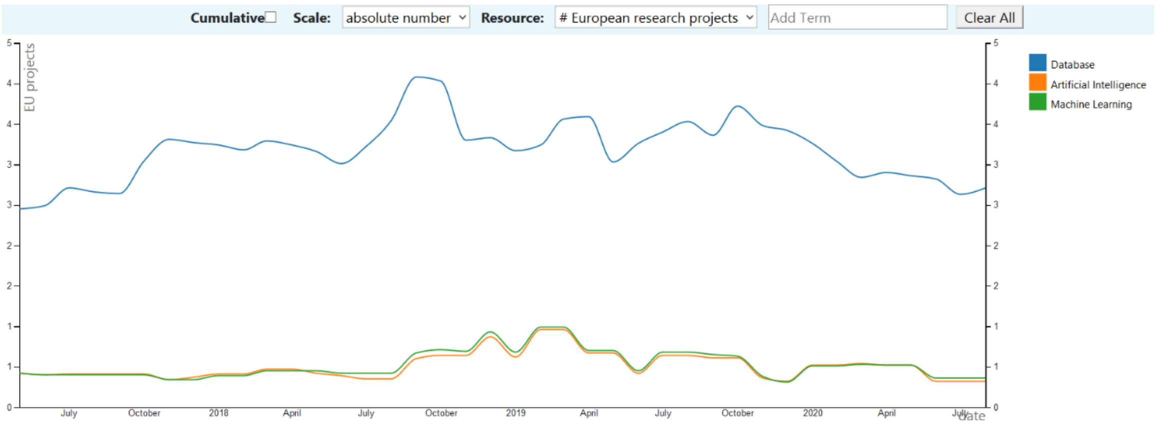
Search:

Topic	EU Projects
Augury	10
IBM Watson	6.06
Bonsai	4.94
Google Cloud AI Platform	3.25
Intercom	2.69
Unity	2.13
Weka	2.13
Mahout	2.13
Affectiva	2.13
Amazon Rekognition	2.13
TensorFlow	1

Figure 32: Popular Tools and Technologies for Projects (from DataBench Observatory tool, Category: machine learning, overall)

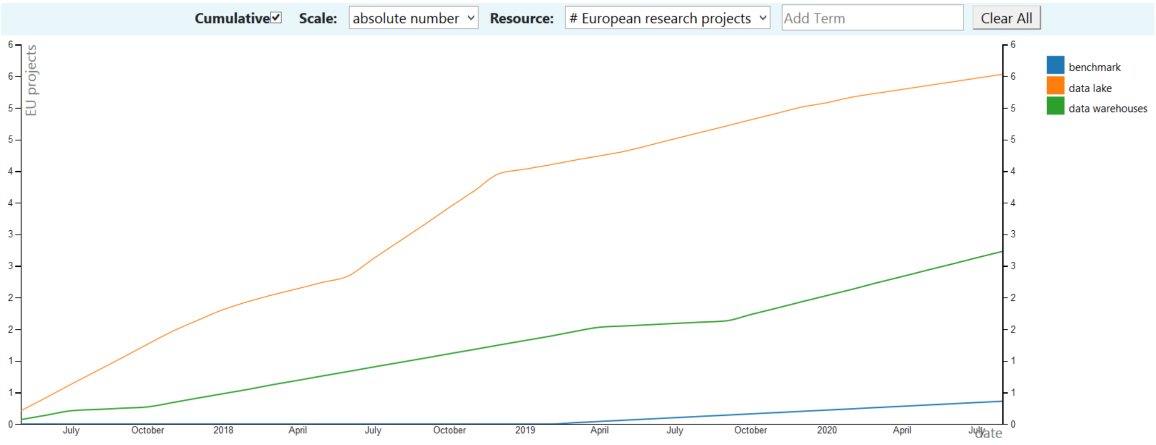
Figure 34: Topic Trends for Projects (Database, AI, Machine Learning) and Figure 8: Topic Trends for Projects (Benchmark, Data Lake, Data Warehouses; Cumulative) provide a view on topic trends for Artificial Intelligence, Machine Learning and Databases, as well as Benchmarking and several data related topics. It is possible to notice that Databases topic is very popular within Cordis data.

In DataBench observatory the user has an option to select the presentation options for each figure – in such way, one of the figures presents a cumulative trend.



Cumulative□ Scale: absolute number Resource: # European research projects

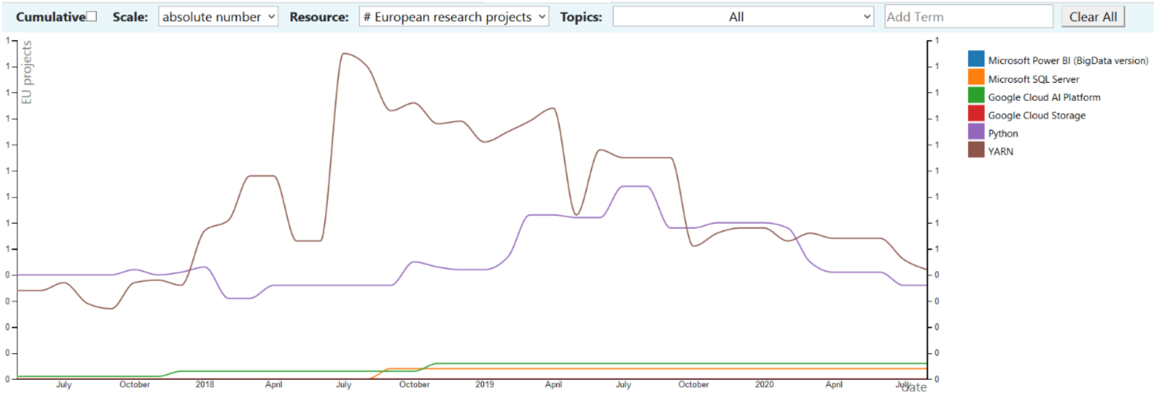
Figure 33: Topic Trends for Projects (Database, AI, Machine Learning)



Cumulative☑ Scale: absolute number Resource: # European research projects

Figure 34: Topic Trends for Projects (Benchmark, Data Lake, Data Warehouses; Cumulative)

Figure 35: Tools and Technologies Trends for Projects (Microsoft Power BI, Microsoft SQL Server, Google Cloud AI, Google Cloud Storage, Python, YARN) shows trends several tools and technologies. In particular, it is possible to observe how Python trend is fluctuating within Cordis project data.



Cumulative□ Scale: absolute number Resource: # European research projects Topics: All Add Term Clear All

Figure 35: Tools and Technologies Trends for Projects (Microsoft Power BI, Microsoft SQL Server, Google Cloud AI, Google Cloud Storage, Python, YARN)

The examples provided above present a practical application for the DataBench observatory tool and the possibility to analyze and visualize topics, trends and technologies within Cordis project open data.

Since the process is automated, the DataBench partners can maintain and regularly update the tool with new data. For each topic, tool or technology from DataBench ontology, within specific time slot, we can identify the number of associated research projects that mention the topic/tool and technology. The DataBench observatory tool allows the user to explore research projects based on semantic analysis of project text description.

7. Conclusions

This report has presented evaluations and examples of the usage of the DataBench Toolbox.

This has been done with both an Analytic and a Project based Evaluation of the Toolbox, followed by a Use case-based Evaluation of the DataBench Blueprints and a Project data analysis with the DataBench Observatory.

The report first provided an Analytic evaluation of the DataBench Toolbox. This included an analytic methodology with evaluation criteria with a rank and a score, and a corresponding evaluation and discussion of the outcome with suggestion for features and further improvement possibilities.

Further a Project based evaluation of the DataBench Toolbox has been presented. This has a focus on how projects can use Pipelines and Blueprints for the analysis of potential Big Data and AI technologies and corresponding benchmarks and how project experiences with their pipelines and blueprints can be reported and shared as knowledge nuggets in the DataBench Toolbox. The project-based evaluation has been supported by two campaigns through the ReachOut beta testing system: "Generation of Architectural Pipelines and Blueprints" and "Finding the right benchmarks for technical and business users". The results from 6 example Big Data PPP projects based on this are reported. The DataBench methodology and Toolbox have been applied to/used by the following projects: I-BiDaas, TheyBuyForYou (TBFY), DeepHealth, DataBio, Track&Know and CLASS.

With a foundation in these evaluations a strategy for further sustainable evolution of the DataBench Toolbox is suggested.

The project analysis has been based on addressing needs and requirements and solutions through the areas of the BDV Reference Model extended with the four pipeline steps of Data Acquisition/Collection and Storage, Data Preparation and Curation, Data Analytics with AI/Machine Learning and Action and Interaction, including Data Visualisation and User Interaction as well as API Access.

This has been found useful for a number of projects and the approach is now being followed up on by a number of current Big Data PPP projects in their description of their Big Data and AI Applications.

A Use case-based Evaluation of the DataBench Blueprints followed. This included use cases from the three domains of Agriculture, Heavy Equipment optimization, Smart manufacturing and Healthcare, diagnostic systems

Finally, a Project data analysis with the DataBench Observatory has been done, with an initial Cordis Project Data Description and further Big Data PPP Project Data Descriptions is in progress.

In this work, we have addressed the sustainability of the DataBench Toolbox. The developed methodology has been refined, adjusted and put in the form to be reusable by external users and suitable for continuous monitoring of the Big Data projects in the future, as part of the DataBench Toolbox.

Sustainability of the DataBench Toolbox is planned through continued support from Big Data and AI i-Spaces and Digital Innovation Hubs, as well as organizations like BDVA and a future AI/Big Data/Robotics PPP organisation, as described further in the DataBench Exploitation and Sustainability plan.

The aim for the DataBench Toolbox is to be helpful for the planning and execution of future Big Data and AI oriented projects, and to serve as a source for the identification and use of relevant technical benchmarks, also including links to a business perspective for applications through identified business KPIs and business benchmarks.

Further work is now related to populating the DataBench Toolbox with additional examples of actual Big Data and AI pipelines realised by different projects, and further updates from existing and emerging technical benchmarks. In particular there is an interaction with the following projects: Boost 4.0, ExtremeEarth, BodyPass, DataPorts, COGNITWIN, EW_Shopp, ELASTIC, MUSKETEER, BigDataStack and LEXIS.

References

1. A Process for COTS Software Product Evaluation, July 2004 - https://apps.dtic.mil/dtic/tr/fulltext/u2/a443491.pdf
2. Software Evaluation: Tutorial-based Assessment - Mike Jackson, Steve Crouch and Rob Baxter - https://www.software.ac.uk/sites/default/files/SSI-SoftwareEvaluationTutorial.pdf
3. Software Product Evaluation - Current status and future needs for customers and industry by Teade Punter , Rini Van Solingen , Jos Trienekens - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.195.7657
4. Guidelines for Project and Programme Evaluations - https://www.mopfi.gov.mm/sites/default/files/upload_pdf/2017/09/Guidelines%20for%20Project%20and%20Programme%20Evaluations_Austrian%20Development%20Cooperation_0.pdf
5. Methods and Practice of Software Evaluation. The Case of the European Academic Software Award, Peter Baumgartner and Sabine Payr - http://www.medidaprix.org/mdd_2001/easa-evaluation.pdf
6. Software Evaluation: Criteria-based Assessment, Mike Jackson, Steve Crouch and Rob Baxter - https://www.software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf
7. ISO/IEC 9126-1:2001 Software engineering — Product quality — Part 1: Quality model - https://www.iso.org/standard/22749.html
8. D5.3 Assessment of Technical Usability, Relevance, Scale and Complexity - https://www.databench.eu/wp-content/uploads/2019/07/d5.3-assessment-of-technical-usability-relevance-scale-and-complexity.pdf
9. https://databench.ijs.si/knowledgeNugget/nugget/71
10. https://databench.ijs.si/benchmarks/selectbench/109
11. https://databench.ijs.si/benchmarks/selectbench/66
12. https://databench.ijs.si/benchmarks/selectbench/65
13. https://databench.ijs.si/benchmarks/selectbench/83
14. https://databench.ijs.si/benchmarks/selectbench/97
15. https://databench.ijs.si/benchmarks/selectbench/23
16. https://databench.ijs.si/benchmarks/selectbench/68
17. OCDS ontology - https://github.com/TBFY/ocds-ontology
18. euBusinessGraph ontology - https://github.com/euBusinessGraph/eubg-data
19. DataBench Generic Data Pipeline - https://databench.ijs.si/knowledgeNugget/nugget/127

20. https://tbfy.github.io/platform/
21. https://tbfy.librairy.linkeddata.es/search-api/api.html
22. http://data.tbfy.eu/statistics/
23. http://tbfy.ijs.si/
24. Data Management -https://databench.ijs.si/benchmarks/searchResults/Data%20management
25. BDV Reference Model - https://databench.ijs.si/bdva
26. Generic Big Data Analytics Blueprint - https://databench.ijs.si/knowledgeNugget/nugget/84
27. https://www.iso.org/committee/6794475.html
28. https://www.iso.org/committee/6794475/x/catalogue/p/1/u/0/w/0/d/0
29. https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0
30. Open Contracting Data Standard (OCDS) - https://standard.open-contracting.org/latest/en/
31. OCDS ontology - https://github.com/TBFY/ocds-ontology
32. https://databench.ijs.si/benchmarks/guidedSearch
33. https://databench.ijs.si/unicum
34. https://databench.ijs.si/bdva
35. https://databench.ijs.si/benchmarks/searchResults/RDF
36. https://databench.ijs.si/benchmarks/searchByFeature/140
37. https://databench.ijs.si/benchmarks/searchByFeature/460
38. https://databench.ijs.si/benchmarks/searchResults/Graphs%20or%20linked%20data
39. https://www.reachout-project.eu/view/Main/#vMmeETzVe7kA
40. https://github.com/TBFY/knowledge-graph
41. https://databench.ijs.si/benchmarks/selectbench/83
42. http://data.tbfy.eu/statistics/
43. https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-6-reference-architecture
44. http://docs.opengeospatial.org/is/10-157r4/10-157r4.html
45. https://class-project.eu/news/benchmarking-real-time-serverless-applications-owperf
46. https://databench.ijs.si/benchmarks/selectbench/38
47. https://www.ennomotive.com/industrial-iot-sensor-prices/
48. G. Costa, “A framework for the evaluation of big data technologies”, Master thesis, Adv. Arne Berre, Chiara Francalanci, Politecnico di Milano, July 2020.
49. Nasiri, H., Nasehi, S., Goudarzi, M., “Evaluation of distributed stream processing frameworks for IoT applications in Smart Cities,” Journal of Big Data, 2019.
50. https://github.com/ilkarman/DeepLearningFrameworks
51. BDVe projects: https://www.big-data-value.eu/our_projects
52. https://ec.europa.eu/digital-single-market/en/programme-and-projects/project-factsheets-data
53. BDVe Marketplace: https://marketplace.big-data-value.eu/solutions

Annex A – ReachOut Campaign – Pipelines and Blueprints

Generation of architectural Pipelines-Blueprints

Starts on:
22/11/2020
Ends on:
31/01/2021
Estimated Test Duration:
30 minutes plus mapping to blueprints that requires desk analysis
Target beta testers profile:
Developers

Beta tester level:
Advanced

Campaign objectives
DataBench has released the DataBench Toolbox, a one-stop shop for big data and AI benchmarking. It offers a catalogue of existing benchmarking tools and information about technical and business benchmarking.

This campaign (extended until the end of January 2021) aims at getting content in the form of new architectural big data/AI blueprints mapped to the BDV reference model and the DataBench pipeline/blueprint. In this campaign we focus mainly on advanced users that would like to contribute with practical examples of mapping their architectures to the generic blueprints. The results will be published in the DataBench Toolbox acknowledging the ownership and can be used by the owners for their own purposes in their projects/organizations to claim their efforts in mapping with existing standardization efforts in the community.

Note that we provide information about the BDV Reference Model, the four steps of the DataBench Generic data pipeline (data acquisition, preparation, analysis and visualization/interaction), and the generic big data blueprint devised in DataBench, as well as some examples and best practices to provide the mappings . Testers should study the available DataBench information and guidelines. Then using the provided steps testers should prepare their own mappings, resulting diagrams and explanations, if any. The Toolbox provides a web form interface to upload all relevant materials that will be later assessed by an editorial board in DataBench before the final publication in the Toolbox.

Requirements for this campaign
- Having a big data/AI architecture in place in your project/organization
- Willingness to provide mappings from your architecture to be part of the DataBench pipeline/blueprints
- Basic Knowledge of web browsing
- Internet connection
- Use preferably Google Chrome

For any inquiry regarding to this campaign, please write an email to databenchtoolbox@gmail.com.

Beta test instructions and scenario

The Toolbox is accessible without the need to log in to the system, but the options are limited to pure search. You can see that without registering the options in the menu are very few. To perform this campaign, we would like all involved users to first sign in into the DataBench Toolbox to get a user profile that you will use throughout the campaign:

- Go to https://databench.ijs.si/ and click on "Sign up" option located at the top of the page on the right side.

- Fill in the form to generate your new user by providing a username and password of your choice, your organization, email, and your user type (at least Technical for this exercise).

Once you have created your user, please sign in with it to the Toolbox. You will be directed to the Toolbox main page again, where you could see that you have more options available.

Besides the options available through the menu, the main page provides:
A) a carrousel with links,
B) User journeys for users of different profiles: Technical, Business and Benchmarks providers,
C) Videos aimed at these 3 types of users explaining briefly the main functionalities offered for each of them,
D) Shortcuts to some of the of the functionalities, such as FAQ, access to the benchmarks or knowledge catalogues, the DataBench Observatory, etc.

A) Get information about DataBench pipelines and blueprints

This campaign aims at providing you the means to search and browse existing data pipelines and the explanations on how to map your own architecture to efforts such as the BDV Reference model, the DataBench Framework and the mappings with existing initiatives.

We encourage you to first go to the Technical user journey accessible from the front-page of the Toolbox, read it and follow the links given to you to get acquainted with the entries related to blueprints and pipelines. In the "Advanced" user journey you will find the following:

- Link to the DataBench Framework and it relation to the BDV Reference Model, where you can find an introduction to the different elements that composes the DataBench approach towards technical benchmarking.

- Link to the DataBench Generic Pipeline , where an explanation of the 4 main steps in data pipelines are explained. These 4 steps are the basic building blocks for the mappings to other blueprints and existing initiatives.

- User Journey - Generic Big Data Analytics Blueprint : This is the main piece of information that you need to understand what we mean by mapping an existing architecture to our pipelines and blueprints. You will find links to the generic pipeline figure.

- Practical example of creating a blueprint and derived cost-effectiveness analysis: Targeting the Telecommunications Industry .

- Ways to report your suggestions for new blueprints, by using the Suggest blueprint/pipeline option under the Knowledge Nuggets menu

Below is a summary of the minimal set of actions we encourage you to do:

1. Go to the User journeys area of the main page and click on "Technical".

2. Go to the link to the User Journey: Generic Big Data Analytics Blueprint at the bottom of the "Advanced" area of the page.

3. Read and understand the different elements of the pipeline (the 4 steps) and the elements of the generic blueprint as described in the previous link.

4. Check examples of already existing blueprints. In order to do that use the search box located at the top right corner and type "blueprint". Browse through the blueprints.

B) Desk analysis

Once you are familiar with the DataBench Toolbox and the main concepts related to the blueprints, you need to do some homework. You should try to map your own architecture to the DataBench pipeline and the generic blueprint. We suggest the following steps:

- Prepare a figure with the architecture you have in mind in your project/organization.

- Create links to the 4 steps of the data pipeline and generate a new figure showing the mapping.

- Create links to the Generic Big Data Analytics Blueprint figure and generate a new figure showing the mappings. In order to do so you might use the generic pipeline figure and particularize to your components as it was done in the example provided for the Telecommunications Industry .

C) Upload your blueprint to the Toolbox

- Upload your files as pdf or images by using the Form of suggestion of blueprints available from the Knowledge Nuggets menu. Try to include a description with a few words about the sector of application of your blueprint, main technical decisions or anything you might find interesting to share.

- The DataBench project will revise the blueprints and publish them into the platform acknowledging your authorship.

Congratulations! You have completed the assignment of this campaign! Go now to fill in the feedback questionnaire. Please note that filling in the questionnaire will be your ticket for incentives.

Feedback questionnaire

When you are done with the testing, please fill in the feedback questionnaire.
Please note that filling in the questionnaire will be your ticket for incentives.

Incentives

As a recognition for your efforts and useful feedback, you will be added as a DataBench contributor within our Website, your blueprint published, and the authorship of your contribution acknowledged in the Toolbox. This offer is limited to the beta testers interacting with the team, by 15 December 2020. You will be contacted individually for contribution opportunities. Please, provide a valid contact email during the survey phase and in the form for suggestions of new blueprints.

Also, Beta Testers will be offered to be added to the ReachOut Hall of fame, will take part in the ReachOut Lottery, and 16 randomly selected beta testers providing a fully returned questionnaire will be awarded a money prize in recognition.

Campaign Mailing List

Please provide your e-mail address below and in the feedback questionnaire, in order to enter the ReachOut incentives programme and to join the mailing list for this campaign, in order to interact with the Campaign Manager. Find out more about Reachout informed consent.

Annex B – ReachOut Campaign – Benchmarks for technical and business users

Finding the right benchmarks for technical and business users

Estimated Test Duration:
30 to 40 minutes
Target beta testers profile:
Business users, Developers

Beta tester level:
Intermediate

Campaign objectives
DataBench has released the DataBench Toolbox, a one-stop shop for big data and AI benchmarking. It offers a catalogue of existing benchmarking tools and information about technical and business benchmarking.

This campaign aims at getting feedback of the usage of the Tool and the user interface of the web front-end of the Toolbox. The Toolbox provides a set of user journeys, or suggestions, for three kind of users: 1) Technical user (people interested in technical benchmarking), 2) Business users (interested in finding facts, tools, examples and solutions to make business choices), and 3) Benchmark providers (users from benchmarking communities or that generated their own benchmarks). In this campaign we focus mainly on technical and business users. We provide some minimal instructions for these two types of users to understand if finding information in the Toolbox is not a cumbersome process and getting your feedback. The idea is to use the user journeys drafted in the Toolbox to drive this search process and understand if users find this information enough to kick-start the process of finding the right benchmark and knowledge they were looking for.

Requirements for this campaign
- Previous knowledge about Big Data or AI
- Basic Knowledge of web browsing
- Internet connection
- Use preferably Google Chrome

For any inquiry regarding a this campaign, please write an email to databenchtoolbox@gmail.com.

Beta test instructions and scenario
The Toolbox is accessible without the need to log in to the system, but the options are limited to pure search. You can see that without registering the options in the menu are very few.

Initial steps to log in as a Toolbox user

To perform this campaign, we would like all involved users to first sign in into the DataBench Toolbox and create a user profile that you will use throughout the campaign:

- Go to http://databench.ijs.si/ and click on "Sign up" option located at the top of the page on the right side.
- Fill in the form to generate your new user by providing an username and password of your choice, your organization, email, and your user type (Technical and/or Business, depending on your preferences and skills).

Once you have created your user, please sign in with it into the Toolbox. You will be directed to the Toolbox main page again, where you could check that you have more options available.

Besides the options available through the menu, the main page provides:
A) a carrousel with links,
B) User journeys for users of different profiles: Technical, Business and Benchmarks providers,
C) Videos aimed at these 3 types of users explaining briefly the main functionalities offered for each of them,
D) Shortcuts to some of the functionalities, such as FAQ, access to the benchmarks or knowledge catalogues, the DataBench Observatory, etc.

A) For Technical Users

This campaign aims at using the user journeys as starting point to help you navigating the tool. We encourage you to click on the Technical user journey, read it and follow the provided links to get acquainted with the tool and what you can do with it. Get used to the main two catalogues: the benchmarks catalogue (tools for big data and AI benchmarking), and the knowledge nuggets catalogue (providing information about technical and business aspects related to benchmarking and big data technologies). Learn about existing big data architectural blueprints and browse some of them.

Additionally, if you already have a goal in your mind (i.e. finding a benchmark for testing a specific ML model, or compare the characteristics of different NoSQL databases), we encourage you to try to find the appropriate benchmark and report your conclusions later in the questionnaire.

Below is a summary of the minimal set of actions we encourage you to do:

1. Go to the User journeys area of the main page and click on "Technical".

2. Read the content of this page, divided into advice for "Beginners" (first-time users) and "Advanced" (providing extra recommendations of what to do next). Focus first on the "Beginners" area and click on the different links to browse to the different options to get used to the tool. We recommend you to come back to the User journey page until you click on all the available options for beginners, but feel free to stray and use the navigation and links from other pages to get used to the tool. After you finish clicking on all the options for beginners, you should have seen the benchmarks and knowledge nuggets catalogues, used some of the search functionalities and browsed some of the existing architectural blueprints. You are now ready to go further!

3. Focus now on the "Advanced" area of the User journey page .

- Here you will find ways to suggest new content via web forms (i.e. new benchmarks you might know that are missing in the catalogue, a version of a big data blueprint you are dealing with in a project, or a new knowledge nugget based on your experience). We are not expecting you to fill-in these forms at this stage, but just acknowledge their potential value (and feel free to contribute any time).

- You will find also links to specific more advanced user journeys or practical examples at the end of the advanced user journeys. Click on the ones that take your attention and start navigating via the links offered by them. From this moment we expect that you know the main options of the Toolbox and how to navigate and browse through it. You should have noted by now that both benchmarks and knowledge nuggets are annotated or categorized with clickable tags, which makes navigation through related items possible.

4. Get used to the search functionalities. The Toolbox offers 4 types of search:
- Search text box located at the top right corner of the pages. This is a full text search. You can enter any text and the results matching that text from both the benchmark and knowledge nuggets catalogues will appear.

- Search by "BDV Reference Model" option from the menu allows you to have a look at the model created by the BDV PPP community (check the BDV SRIA for more details). The model is represented graphically and is clickable. If you click in any of the vertical or horizontal layers of the model you will be directed to the benchmarks and/or knowledge annotated in the Toolbox to these layers. Browse through this search.

- Search by "Guided benchmark search" . In simple terms this is a search by the tags used to annotate benchmarks and knowledge nuggets. These tags range from technical to business aspects. You can click on the categories of tags to find related information. Browse to some of the options of this search.

- Finally, the "Search by Blueprint/Pipeline" option allows a search that presents graphically a generic architectural blueprint developed in DataBench with the most common elements of a big data architecture. The blueprint is aligned with 4 steps of a DataBench Generic data pipeline (data acquisition, preparation, analysis and visualization/interaction). The graphic is clickable both at the level of the four steps of the pipeline and in some of the detailed elements of the blueprint. Click on the parts of the diagram you are interested to find a list of existing benchmarks and nuggets related to it. Browse some of them. There are nuggets that show a summary of existing big data tools for each of the elements of the pipeline. See if you find it easy to browse through the results .

Congratulations! You have completed the assignment of this campaign! Go now to fill in the feedback questionnaire.

NOTE – Some of the available benchmarks can be deployed and run in your premises. Those are listed first in the Benchmark catalogue and when you click on them you will find the configuration file at the bottom of their description. If you want to run any of them, you should have dedicated infrastructure to do so. We are not expecting you to do so in this exercise.

B) For Business users

As for technical users, this campaign aims at using the user journeys as starting point to help you navigating the tool. We encourage you to click on the Business user journey, read it and follow the links given to you to get acquainted with the tool and what you can do with it. Get used to the main two catalogues: the benchmarks catalogue (tools for big data and AI benchmarking), but mainly to the knowledge nuggets catalogue (providing information about technical and business aspects related to benchmarking and big data technologies). Learn about existing big data architectural blueprints and browse to some of them, as they apply to different industries and might be of interest for business purposes.

Additionally, if you already have a goal in your mind (i.e. finding most widely used business KPIs in a specific sector), we encourage you to try to find the appropriate information in the knowledge nugget catalogue and report your conclusions later in the questionnaire.

Below there is a summary of the minimal set of actions we encourage you to do:

1. Go to the User journeys area of the main page and click on "Business".

2. Read the content of this page, divided into advice for "Beginners" (first-time users) and "Advanced" (providing extra recommendations for what to do next). Focus first on the "Beginners" area and click on the different links to browse to the different options to get used to the tool. We recommend you to come back to this User journey page until you click on all the available options for beginners, but feel free to stray and use the navigation and links from other pages to get used to the tool. After finishing clicking on all the options for beginners, you should have seen the benchmarks and knowledge nuggets catalogues, used some of the search functionalities and browsed some of the existing architectural blueprints. You are now ready to go further!

3. Focus now on the "Advanced" area of the User journey page.
- You will find links to different elements, such as nuggets related to business KPIs, by industry, etc. Browse through them and follow the links.

- You will find ways to suggest new content via web forms (i.e. a new knowledge nugget based on your experience). We are not expecting you to fill-in these forms at this stage, but just acknowledge their potential value (but feel free to contribute any time).

- You will find also links to specific more advanced user journeys or practical examples at the end of the advanced user journeys. Click on the ones that take your attention and start navigating via the links offered by them. From this moment we expect that you know the main options of the Toolbox and how to navigate and browse through it. You should have noted by now that both benchmarks and knowledge nuggets are annotated or categorized with clickable tags, which makes navigation through related items possible.

5. Get used to the search functionalities. The Toolbox offers 4 types of search:
- Search text box located at the top right corner of the pages. This is a full text search. You can enter any text and the results, matching that text from both the benchmark and knowledge nuggets catalogues, will appear.

- Search by "BDV Reference Model" option from the menu allows you to have a look at the model created by the BDV PPP community (check the BDV SRIA for more details). The model is represented graphically and is clickable. If you click on any of the vertical or horizontal layers of the model you will be directed to the benchmarks and/or knowledge annotated in the Toolbox to these layers. Browse through this search.

- Search by "Guided benchmark search". In simple terms this is a search by the tags used to annotate benchmarks and knowledge nuggets. These tags range from technical to business aspects. You can click on the categories of tags to find related information. Browse to some of the options of this search.

Finally, the "Search by Blueprint/Pipeline" option allows a search that presents graphically a generic architectural blueprint developed in DataBench with the most common elements of a big data architecture. The blueprint is aligned with 4 steps of a DataBench Generic data pipeline (data acquisition, preparation, analysis and visualization/interaction). The graphic is clickable both at the level of the four steps of the pipeline and in some of the detailed elements of the blueprint. Click on the parts of the diagram you are interested to find a list of existing benchmarks and nuggets related to it. Browse some of them. There are nuggets that show a summary of existing big data tools for each of the elements of the pipeline. See if you find them it easy to browse through the results .
6. This part of the test is not guided, as we expect you to navigate through the options you have seen previously. Once you know how to navigate, try to find information of interest for your industry or area of interest:
• Try to find information about the most widely used KPIs or interesting use cases.
• Try to find information about architectural blueprints for your inspiration.

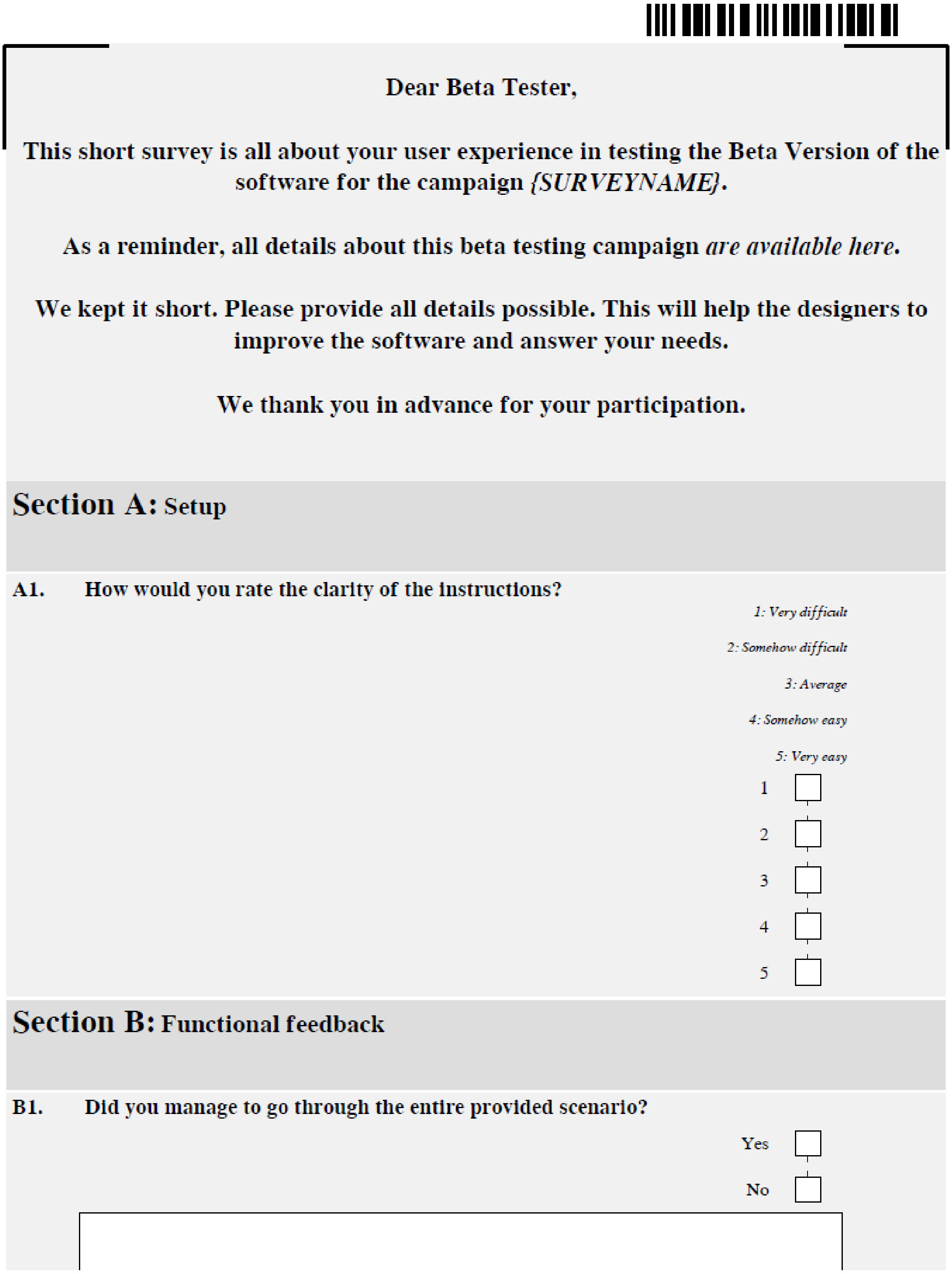
Congratulations! You have completed the assignment of this campaign! Go now to fill in the **feedback questionnaire**.

Incentives
As a recognition for your efforts and useful feedback, you will be added as a DataBench contributor within our Website. This offer is limited to the beta testers interacting with the team, by 6 December 2020. You will be contacted individually for contribution opportunities. Please, provide a valid contact email during the survey phase.

Also, Beta Testers will be offered to be added to the ReachOut Hall of fame, will take part in the ReachOut Lottery, and 16 randomly selected beta testers providing a fully returned questionnaire will be awarded a money prize in recognition.

Annex C – ReachOut Campaign – Feedback questionnaire

(First page of the Campaign Feedback questionnaire)



Dear Beta Tester,

This short survey is all about your user experience in testing the Beta Version of the software for the campaign *{SURVEYNAME}*.

As a reminder, all details about this beta testing campaign *are available here*.

We kept it short. Please provide all details possible. This will help the designers to improve the software and answer your needs.

We thank you in advance for your participation.

Section A: Setup

A1. How would you rate the clarity of the instructions?

1: Very difficult
2: Somehow difficult
3: Average
4: Somehow easy
5: Very easy

1 □
2 □
3 □
4 □
5 □

Section B: Functional feedback

B1. Did you manage to go through the entire provided scenario?

Yes □
No □