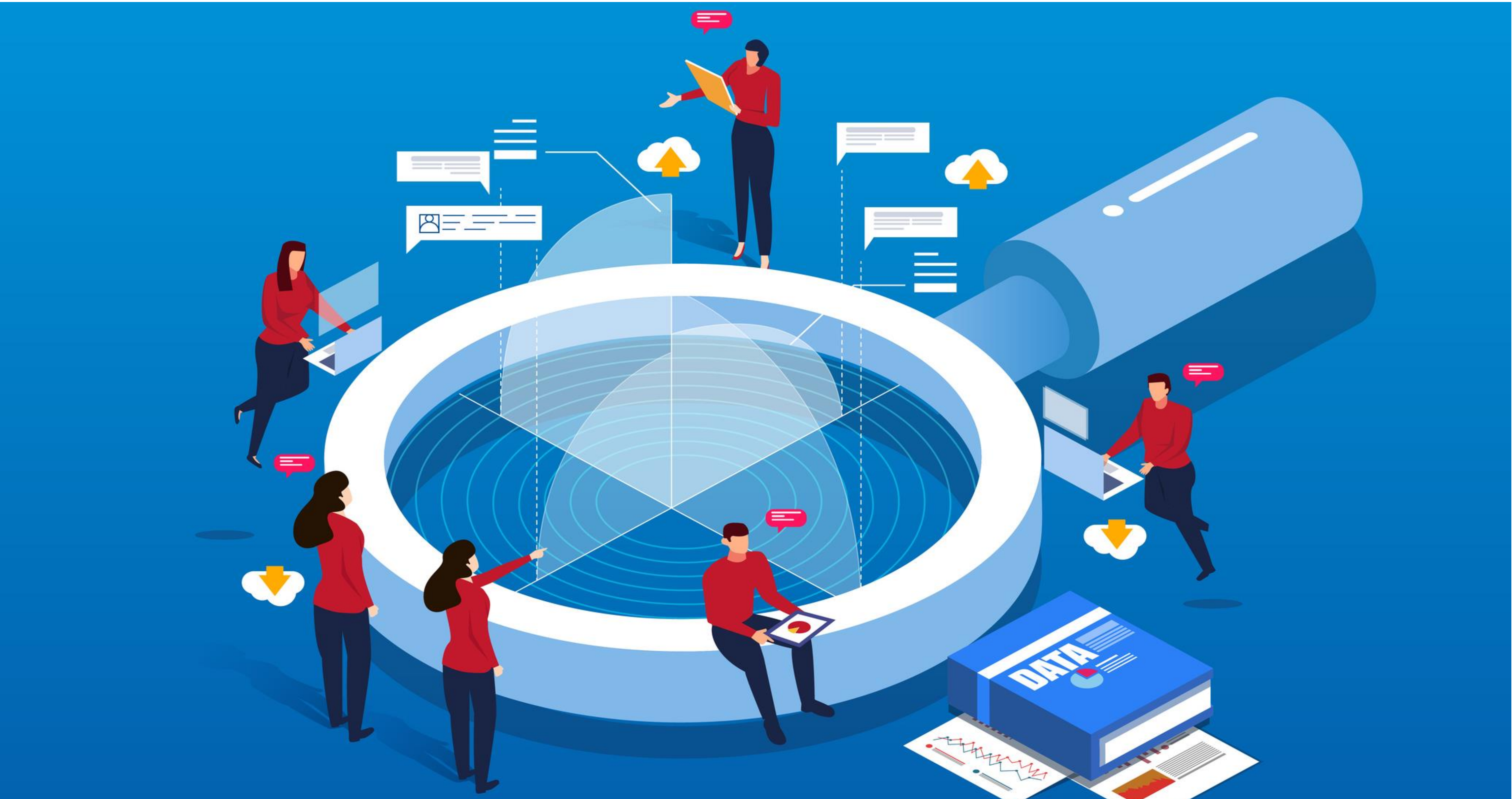
DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

DataBench Toolbox Architecture



About Databench

Organisations rely on evidence from the Benchmarking domain to provide answers on how their processes are performing. There is extensive information on how and why to perform technical benchmarks for the specific management and analytics processes, but there is a lack of objective, evidence-based methods to measure the correlation between Big Data Technology (BDT) benchmarks and an organisation's business benchmarks and demonstrate return on investment (ROI).

The DataBench project addresses this 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.

Based on existing efforts in big data benchmarking and enabling inclusion of new benchmarks that could arise in the future, the DataBench Toolbox will provide a unique environment to search, select and deploy big data benchmarking tools, giving the possibility to generate unified technical metrics and derive business KPIs.

Abstract

This white paper reports on the current view of the DataBench Toolbox architecture and main functional elements as described in the DataBench deliverable D3.1. The goal of the DataBench Toolbox is to provide a way of reusing existing big data benchmarking efforts under a common framework, providing therefore a way to select, download and homogenize technical and business indicators.

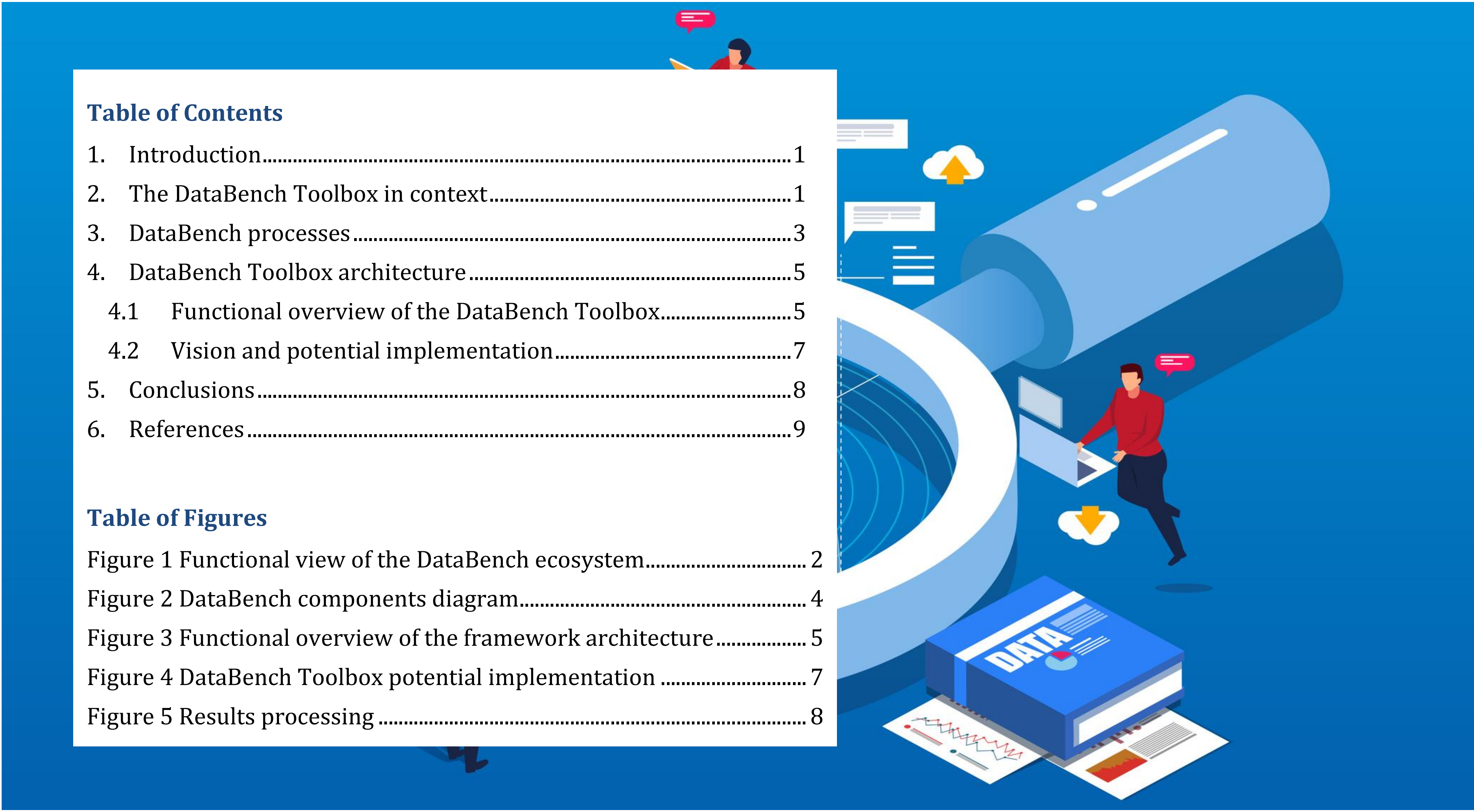


Table of Contents

Table of Figures

1. Introduction

In recent years, there has been an increasing interest in benchmarking solutions to the relatively new field of big data. Existing big data benchmarks cover different aspects of the big data value chain. Some of these benchmarks are targeting specific big data engines such as HiBench [Huang et al.,2010], [HiBench Suite, 2018] or SparkBench [Agrawal et al., 2015], [SparkBench, 2018]. Others are targeting specific database workloads, such as YCSB (Yahoo! Cloud Serving Benchmark) [Cooper et al., 2010], TPCx-IoT [TPCx-IoT, 2018], or Yahoo Streaming Benchmark (YSB) [Chintapalli et al., 2016]. There are attempts to cover some analytic processes [YSB, 2018], or even a more abstract, end-to-end, technology agnostic, application-level, analytics benchmark, such in the cases of BigBench [Ghazal et al., 2013], [BigBench, 2018] and BigBench V2 [Ghazal et al, 2017]. Frameworks such as ABench [Ivanov and Singhal, 2018] goes an extra mile providing Big Data Architecture Stack benchmark based on covering multiple Big Data application scenarios. Finally, a benchmarking framework such as the one proposed by the HOBBIT project provides a way to abstract the generation of specific benchmarks based on a dockized and extensible architecture.

Besides the inherent issue of having multiple systems for big data benchmarking that do not cover the entire needs of many organizations, a major problem with this kind of approaches is that they do not offer a homogenized set of metrics, nor business insights.

DataBench aims to tackle these issues by offering a methodological approach and a set of tools to enable the reuse of existing big data benchmarking tools, as well as means to derive business metrics from the execution of the benchmarks learning from past experiences.

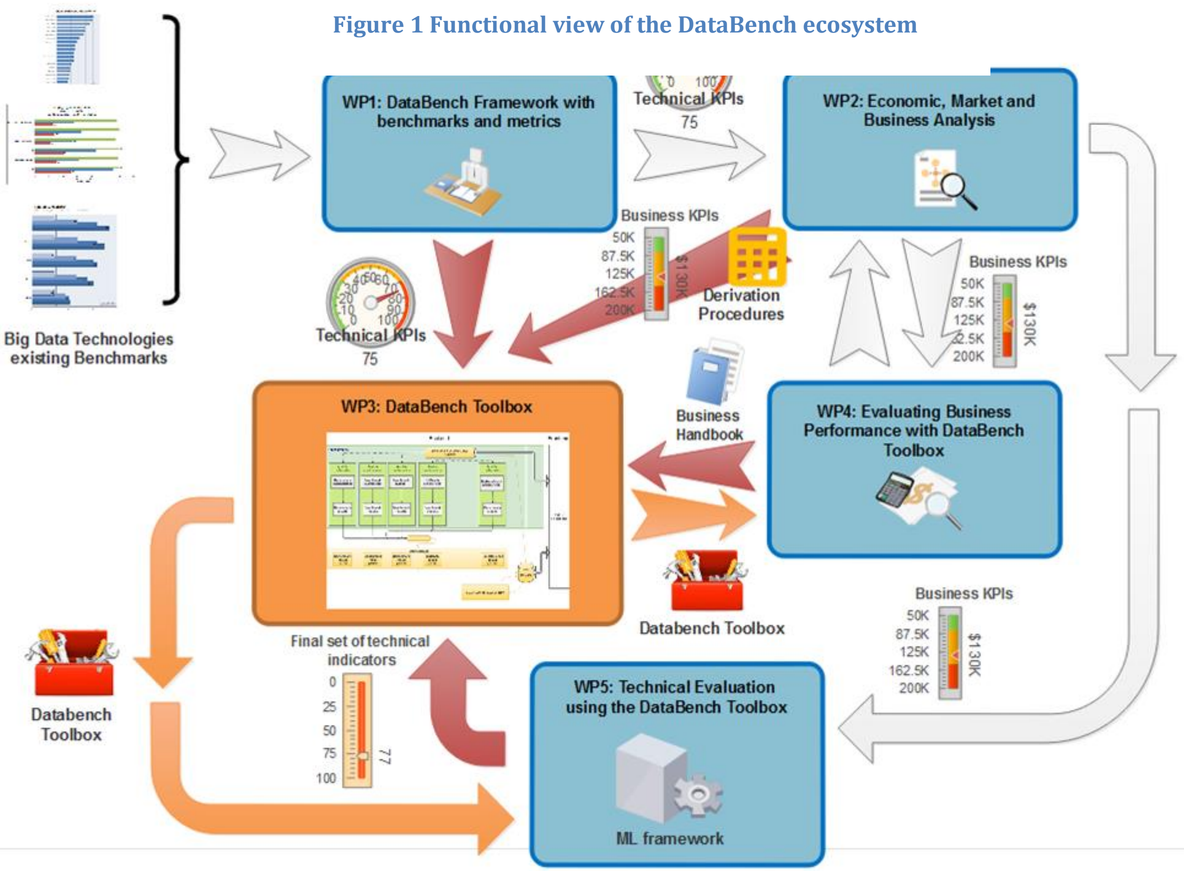
This white paper explains briefly the current thoughts about the architecture of the DataBench Toolbox, which is the main IT component of the DataBench framework. The Toolbox will allow users to select from the plethora of benchmarks the ones that suit their needs, deploy and use them, as well as giving the possibility to upload the results of the benchmark execution to get a homogenized set of metrics, including business insights.

This document is a summary and therefore takes content from the DataBench deliverable D3.1 DataBench Architecture [D3.1].

2. The DataBench Toolbox in context

The DataBench project is composed of several work packages, each of which is in charge of delivering a set of results of different nature (software, reports, handbooks, etc.). From the technical perspective, the DataBench Toolbox, to be delivered in WP3, is the main software element to be provided to the users, but it is not in isolation as depicted in .

Figure 1 Functional view of the DataBench ecosystem


Big Data Technologies existing Benchmarks
WP1: DataBench Framework with benchmarks and metrics
Technical KPIs
75
WP2: Economic, Market and Business Analysis
Business KPIs
50K
87.5K
125K
162.5K
200K
$130K
Derivation Procedures
Technical KPIs
75
Business KPIs
50K
87.5K
125K
162.5K
200K
$130K
WP3: DataBench Toolbox
Business Handbook
WP4: Evaluating Business Performance with DataBench Toolbox
Databench Toolbox
Business KPIs
50K
87.5K
125K
162.5K
200K
$130K
Final set of technical indicators
0
25
50
75
100
77
Databench Toolbox
WP5: Technical Evaluation using the DataBench Toolbox
ML framework

The different elements of the DataBench ecosystem listed in can be divided into a set of major elements of the framework:

- **DataBench Toolbox:** The DataBench Toolbox is the core technical component of the DataBench Framework. It will be the entry point for users that would like to perform big data benchmarking and will ultimately deliver recommendations and business insights. It will include features to reuse existing big data benchmarks, and will help users to search, select, download, execute and get a set of homogenized results. The Toolbox takes as input most of the work done in the rest of the project's work packages.
- **Benchmarks to be integrated into the Toolbox:** Located on the left hand side of , external big data benchmarks are the input to the whole DataBench Framework, and in particular to the DataBench Toolbox. These benchmarks will be made available to the users from the Toolbox user interface, as well as the possibility to get the results of the benchmarking execution runs and homogenize them into a common model in order to get homogenized set of metrics. The degree of integration of the Toolbox with the existing benchmarks will vary depending on the integration issues that may arise.
- **Process to derive business KPIs:** Deriving business KPIs from the results of running big data benchmarks is not a straightforward process. It will depend on the business context and therefore it is not completely clear yet to what extend it could be automated. However, different approaches to derive business insights are currently under study and will be reported in future deliverables. This process is a collaborative work among the different work packages. WP1 is investigating the different existing benchmarks in order to devise a coherent framework and providing input on available technical KPIs and mappings to

existing big data reference models and to the data value chain in order to cluster them. This work will be applied in WP4 to the desk analysis and to the in-depth case studies, which will result into an analysis of the relations between technical and business KPIs, use cases and a Handbook to guide users to select and interpret the results. WP2 on the other hand, will provide guidance in terms of market value.

- The AI framework: WP5 will provide a ML framework that will serve to enable recommendations for the benchmarking community based on past experiences. It will be integrated with the Toolbox.

3. DataBench processes

The diagrams in UML are classified according to their structure or behaviour, the latter refer to the way in which instructions are executed or how activities are given within the system. The structure diagrams define how the software is structured. An example of this is the component diagram that defines the different parts that make up the system to work.

This section aims to provide the DataBench component diagram showing the different subsystems. The global DataBench component diagram is shown in Figure 2.

The subsystems are the following:

- The Accessing subsystem will be in charge of the authentication, access authorization and auditing.
- The User Intentions subsystem will enable users to specify their requirements related to the usage of a big data benchmark.
- The Setup and Runtime subsystem will support the overall workflow associated to the deployment and execution of one the benchmarks registered in the Toolbox. The workflow will include three main steps, namely i) setup and configuration; ii) deployment and execution; and finally iii) injection of the results back into the Toolbox. The Toolbox will provide mechanisms either to download the Toolbox itself to be able to install it into the user premises (isolated or In-house deployment mode), or to deploy directly the benchmark in public clouds (Sandbox mode).
- The Analytics and KPIs management subsystem will manage the homogenization of technical and business metrics and the associated derivation procedures.
- The Benchmark management subsystem will provide the means to add new benchmarks to the Toolbox and manage the necessary metadata to enable their ulterior usage.
- Finally, the Visualization and Reporting subsystem will provide searching over the technical and derived business metrics related to a benchmark execution and the graphical user interface to interact with the system.

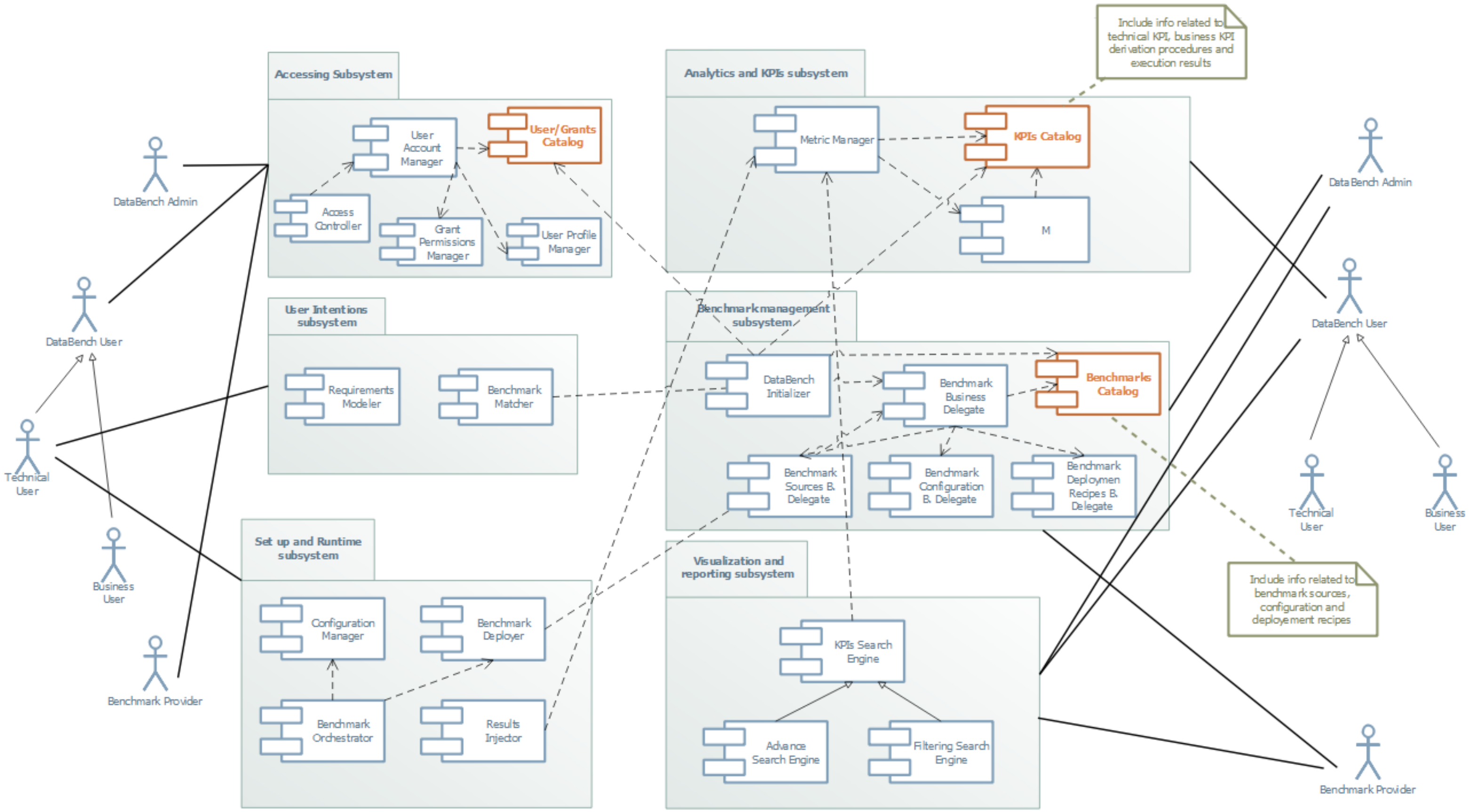


Figure 2 DataBench components diagram

4. DataBench Toolbox architecture

In this section we provide a high level overview of the proposed architecture of the DataBench Toolbox. First, we will present a functional description of the main modules that will form the Toolbox in a technology agnostic manner, as well as their interconnection pointing out the main functionality of each one. Secondly, we relate the BDVA reference model with the described Toolbox architecture and identify the existing gaps. Finally, we will present a potential technical implementation of the functional architecture, describing in more detail each of the modules with their possible implementation and the technologies behind them.

4.1 Functional overview of the DataBench Toolbox

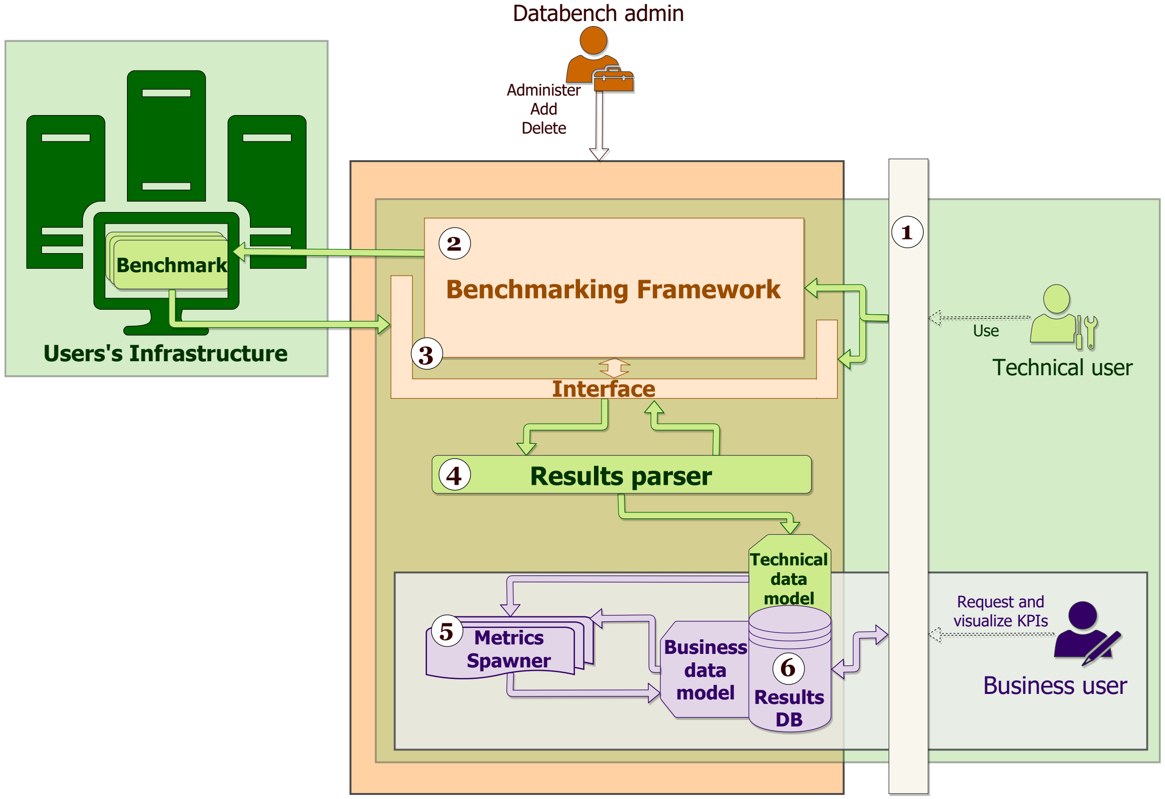


Figure 3 Functional overview of the framework architecture

The architecture of the DataBench Toolbox will follow a modular approach based on templates complemented with a web interface from where the user will be able to interact with the system, as can be seen in .

The architecture is composed of the following interconnected modules:

- **Web interface:** This web interface module will provide the functionality to the different users of the tool to choose which benchmarks they want to run. It will also be in charge of providing a layer of configuration that the users can fill in to preconfigure the templates and the benchmarks to be run later on. Moreover, it will show in the results of the benchmark executions and the derived metrics and business insights.
- **Benchmarking framework:** This component provides the main backend functionality of the Toolbox. It will act as a repository of recipes to be used to generate a template that will be provided to the technical user for the configuration in their in-house installation, or to actually deploy the benchmark in a public infrastructure, depending on the mode selected (in-house or Sandbox).
- **Results interface:** Once the execution of the benchmarks is finalized, the results should be transferred back to the Toolbox through the results interface, to be parsed into the defined technical data model in order to be able to get a set of homogenized technical metrics. The Results interface will be the interconnection point for any benchmark result added to the tool. In this way we will have a single point of connection for the result parsers to read the data from simplifying the whole architecture of the framework.
- **Results parser:** Closely related to the Results interface, this component is in charge of converting these heterogeneous results into a homogenized technical metrics data model. Moreover, the project will provide guidelines for any benchmark provider to generate its own parser and therefore enabling the extensibility of the Toolbox to integrate new benchmarks.
- **Metrics spawner:** This module will be mainly connected to the Results DB module so it can work with the technical metrics data model and work on the derivation of potential business KPIs. Note that the process of deriving business metrics is still under discussion at the time of writing this document. The derivation process could be cumbersome and in many cases impossible without more background knowledge. Therefore, this module will most probably combine different techniques and approaches, not always automated.
- **Results DB:** This module will be the repository of homogenized runs of benchmark executions, including both technical metrics and derived business insights.

4.2 Vision and potential implementation

The Toolbox will be implemented in several iterations, adding more functionality as it grows. The first iteration will cover the integration of some of the most widely used benchmarks and the initial version of most of the components listed above. Most of the functionality related to the web interface and the potential generalization of the derivation of business metrics will be gradually developed and integrated in future iterations. **Error! Reference source not found.** shows a potential implementation of the DataBench Toolbox.

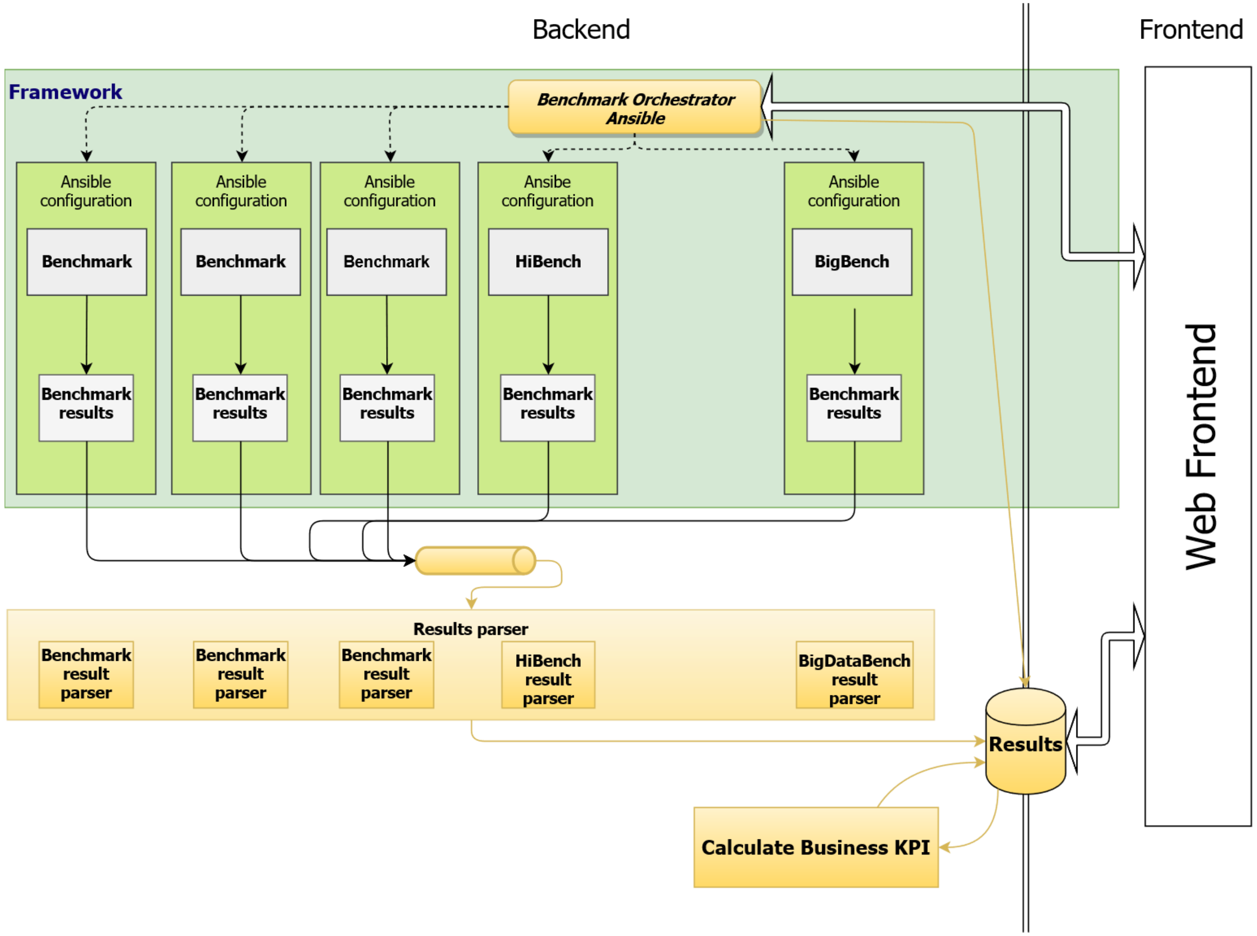


Figure 4 DataBench Toolbox potential implementation

This implementation is based on *Ansible* [Ansible]. *Ansible* is an orchestration, configuration and deployment tool, based on templates called playbooks that simplify the process of deploying and configuring applications in different hosts.

The idea is to generate an *Ansible* playbook to download and configure each of the benchmarks that will be integrated in the Toolbox. These playbooks will be stored in a playbook/configuration GIT repository. With this playbook repository in place, the Toolbox will have a simple web frontend where the user can choose the desired benchmarks to run. This information will be used to provide the requested playbooks to the user so they can configure them according to their requirements and run them in their environment.

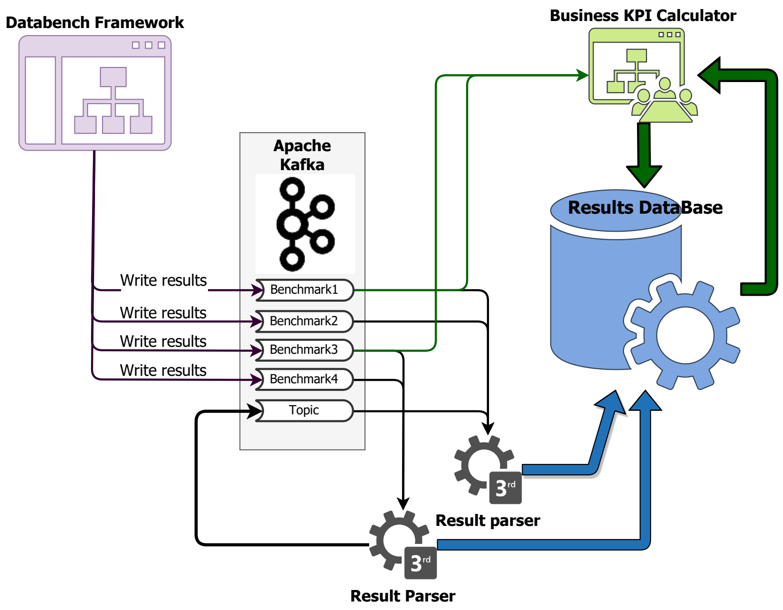


Figure 5 Results processing

Apache Kafka [Apache Kafka] will provide the interface for the interconnection pipeline between the benchmarks and the Result parser as can be seen in Figure 5 Results processing.

Since each benchmark generates its results in a different format, we need some modules in charge of parsing those results into a standardized technical data model that could be then read and interpreted to generate the needed KPIs.

The only requirement for these parsers, as can be interpreted from the diagram in Figure 5 Results processing, is that they should be able to use *Kafka* as a source and/or target and relational databases as target. Since Apache *Kafka* is widely used in the Big Data community, there are lots of frameworks and libraries to use it with almost any programming language or tool so the chosen technology to implement the result parsers is completely open.

As far as the Results DB component is concerned, the expected volume of results from execution of benchmarks is not expected to be large enough to explore highly scalable solutions. The proposed technical data model is relational, linking the benchmark information, the information of the execution and the results obtained.

5. Conclusions

This document is a white paper summarizing the findings presented in D3.1. DataBench Architecture. D3.1 goal was to present the initial architecture of the DataBench Toolbox and associated technical components. This white paper provided a high-level overview of the different elements that form the DataBench ecosystem, putting the DataBench Toolbox, an umbrella framework for big data benchmarking reusing existing efforts well-settled in the community, in context. The document dives in the architecture of the Toolbox explaining the proposed functional elements and subsystems and later providing an example of reference implementation using specific tools and techniques such as Ansible and Apache Kafka.

The alpha version of the tool will be delivered in M18 (July 2019), and enhanced in several iterations in the following 18 months.

6. References

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen: BigBench: towards an industry standard benchmark for big data analytics. SIGMOD Conference 2013: 1197-1208

Ahmad Ghazal, Todor Ivanov, Pekka Kostamaa, Alain Crolotte, Ryan Voong, Mohammed Al-Kateb, Waleed Ghazal, Roberto V. Zicari: BigBench V2: The New and Improved BigBench. ICDE 2017: 1225-1236

Ansible, https://www.ansible.com/

Apache Kafka, https://kafka.apache.org/

BigBench, https://github.com/intel-hadoop/Big-Data-Benchmark-for-Big-Bench, 2018

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

Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, Russell Sears: Benchmarking cloud serving systems with YCSB. Proceedings of the 1st ACM Symposium on Cloud Computing, SoCC 2010, Indianapolis, Indiana, USA, June 10-11, 2010

Dakshi Agrawal, Ali Raza Butt, Kshitij Doshi, Josep-Lluís Larriba-Pey, Min Li, Frederick R. Reiss, Francois Raab, Berni Schiefer, Toyotaro Suzumura, Yinglong Xia: SparkBench - A Spark Performance Testing Suite. TPCTC 2015: 26-44

Deliverable: DataBench D3.1. DataBench Architecture. Pariente, Tomas, 2018

HiBench Suite, https://github.com/intel-hadoop/HiBench, 2018

Huang, S., Huang, J., Dai, J., Xie, T., Huang, B.: The hibench benchmark suite: Characterization of the mapreduce-based data analysis. In: Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA. pp. 41–51 (2010)

Sanket Chintapalli, Derek Dagit, Bobby Evans, Reza Farivar, Thomas Graves, Mark Holderbaugh, Zhuo Liu, Kyle Nusbaum, Kishorkumar Patil, Boyang Peng, Paul Poulosky: Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming. IPDPS Workshops 2016: 1789-1792

SparkBench, https://github.com/CODAIT/spark-bench, 2018

Todor Ivanov, Rekha Singhal: ABench: Big Data Architecture Stack Benchmark. Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE 2018, Berlin, Germany, April 09-13, 2018.

TPCx-IoT, http://www.tpc.org/tpc_documents_current_versions/pdf/tpcx-iot_v1.0.3.pdf , 2018.

YCSB, https://github.com/brianfrankcooper/YCSB, 2018

YSB, https://github.com/yahoo/streaming-benchmarks, 2018

Visit our website and get in touch!

✉ info@databench.eu

@DataBench_eu

@DataBenchEU

DataBench Project

DataBench

DataBench Project

DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

www.databench.eu