DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

D1.1 Industry Requirements with benchmark metrics and KPIs

Abstract

The DataBench project aims to bridge the gap between technical and business benchmarking of Big Data and Analytics applications. The requirements discussed in this report are the result of the first analysis performed in the project on existing Big Data Benchmarking tools, from the interaction with BDVA (Big Data Value Association) and participation in the development and analysis of results of a first questionnaire developed within BDVA, and from analysis of Big Data technology and benchmarking developed in other Work Packages of the project.

As a result of this analysis, an integrated set of benchmark metrics and KPIs is proposed, as an ecosystem of indicators covering Business features, Big data application features, Platform and architecture features, and Benchmark-specific features.

The deliverable discusses the use of these features in an integrated way, as a basis for a methodological integration, for the development of the DataBench Toolbox, and for relating indicators and building a KPI knowledge graph.

European Commission

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

Deliverable D1.1	Industry Requirements with benchmark metrics and KPIs
Work package	WP1
Task	1.1
Due date	31/12/2018
Submission date	21/12/2018
Deliverable lead	POLIMI
Version	1.0
Authors	POLIMI (Barbara Pernici, Chiara Francalanci, Angela Geronazzo, Lucia Polidori) IDC (Gabriella Cattaneo and Helena Schwenk) JSI (Marko Grobelnik) ATOS (Tomás Pariente, Iván Martínez) GUF (Todor Ivanov) SINTEF (Arne Berre)
Reviewer	IDC (Mike Glennon)

Keywords

Benchmarking, KPIs, indicators, modelling framework

Table of Contents

Table of Figures

Table of Tables

Executive Summary

D1.1 Industry Requirements with benchmark metrics and KPIs documents the collected industrial requirements of European significance with mappings to related vertical and horizontal benchmark metrics and KPIs.

In Task 1.1 we initiated the contacts with representatives of various industry sectors and started establishing industrial requirements based on interviews and interactions for priorities and metrics related to analysis of different use cases from industrial sectors and from the existing ICT14 and ICT15 projects. The objective is to establish an industrial user community that can provide the foundation for holistic end-to-end benchmarks that will go across all the different layers of the Big Data technology stack, according to the BDVA reference model. Existing Big Data benchmarks have primarily focused on the commercial/retail domain related to transaction processing (TPC benchmarks and BigBench) or to applications suitable for graph processing (Hobbit and LDBC – Linked Data Benchmark Council). The analysis of different sectors in the BDVA has concluded that they all use different mixes of the different Big Data Types (Structured data, Time series/IoT, Spatial, Media, Text and Graph). Industrial sector specific benchmarks will thus relate to a selection of important data types, and their corresponding vertical benchmarks, adapted for this sector. The existing holistic industry/application benchmarks have primarily been focusing on structured data and Graph data types and DataBench will in addition be focusing on the industry requirements for time series/IoT, spatial and media and text, from the requirements of different industrial sectors such as manufacturing, transport, bio economies, earth observation, health, energy and many others.

The requirements discussed in this report are the result of the first analysis performed in the project on existing Big Data Benchmarking tools, from the interaction with BDVA (Big Data Value Association) and participation in the development and analysis of results of a first questionnaire developed within BDVA, and from analysis of Big Data technology and benchmarking developed in other Work Packages of the project.

As a result of this analysis, an integrated set of benchmark metrics and KPIs is proposed, as an ecosystem of indicators covering Business features, Big data application features, Platform and architecture features, and Benchmark-specific features.

The deliverable discusses the use of these features in an integrated way, as a basis for a methodological integration, for the development of the DataBench Toolbox, and for relating indicators and building a KPI knowledge graph.

1. Introduction and Objectives

The research work conducted in WP1 has the goal to provide a reference framework for understanding the relationship between business KPIs and technical benchmarks, following the objectives for this Work Package defined in the DoA:

Objective 1. Provide the BDT Stakeholder communities with a comprehensive framework to integrate Business and Technical benchmarking approaches for Big Data Technologies.

Objective 4. Liaise closely with the BDVA, ICT 14, 15 to build consensus and to reach out to key industrial communities, to ensure that benchmarking responds to real needs and problems.

The work presented in this deliverable has been developed during the first year of the project, taking as input also the work in other WPs, and in particular, WP2, WP3, and WP4. WP2 and WP4 are both responsible for identifying and assessing business impacts of benchmarks, both from technical and from organizational points of view. As a basis for this report, the work in WP2 has contributed a framework to investigate the main Big Data use cases implemented by industry; WP4 (D4.1), developing an in-depth case study research, has paved the way to show how the relationships between business KPIs and technical benchmarks is materialized in real cases. WP3 in D3.1 has provided a general description of the DataBench Toolbox and also discussed the role of the business and technical metrics. As shown in Figure 1, WP3 has also the role of the connecting the work developed in all Work Packages, based on the DataBench Framework developed in WP1.

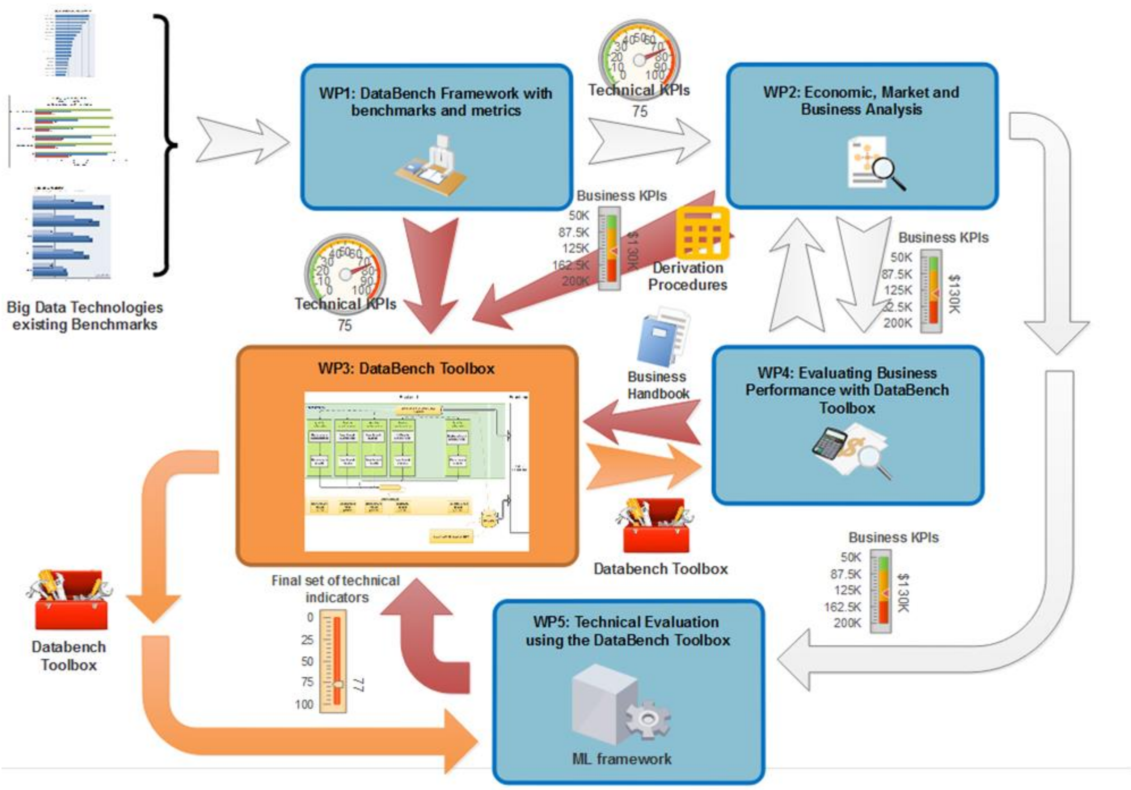

Figure 1 - DataBench WPs

The present report has the main goal of presenting the results of T1.1 to create the basis for a holistic end-to-end benchmarking system for different industry sectors.

This document presents the **state of the art analysis** performed in T1.1:

- We considered the existing benchmarks, identifying the main analysis dimensions.
- We collaborated first with BDVA with a preliminary analysis of relevant indicators, developing an initial questionnaire for ICT14 and ICT15 Big Data Projects (see Section 2.4), and then with WP2 towards the creation of an ecosystem of indicators.

This deliverable contributes to the state of the art presenting the results of the analysis and harmonization of the different indicators in an **ecosystem of indicators** able to capture the different characteristics, from a business perspective to a technical view. The indicators will be the basis for further analyses and for the Toolbox development.

While the structure of the DataBench ecosystem of indicators is being defined in this deliverable, we still expect possible modifications and refinements in the following phases of the project, as the detailed analyses phases continue in WP2 and WP3, and benchmarks analysis and metrics evaluation is performed in WP4 and WP5.

The document is structured as follows:

- Section 1 provides the introduction to the objectives of the deliverable.
- Section 2 contains an overview of the examined benchmarks and their principal characteristics.
- Section 3 dives into a detailed description of the indicators ecosystem for the different perspectives.
- Section 4 presents an integrated framework for using the indicators in the DataBench Toolbox.
- Finally, Section 5 provides the conclusions of the document and outlines the future work on the DataBench metrics.

2. Overview on Big Data Benchmarking

In this chapter we present the state of the art on Big Data benchmarking from three different perspectives: an analysis of benchmarking tools (Sections 2.1), the reference models being developed within BDVA and their use for situating benchmarks (Section 2.2), and a first analysis of BDT (Big Data Technology) and benchmarking developed by the project within BDVA (Section 2.3).

2.1 Benchmarks under Evaluation for the DataBench ToolBox

As already described in D3.1, in WP1 in the first year the DataBench project performed a first survey of big data benchmarking tools. As a result, a set of benchmarks was selected for further in depth analysis, which is ongoing and will be reported within WP3 deliverables, and a number of dimensions for analysing each benchmark was identified and discussed, considering also the recent paper by Han et al., 2018, which discusses benchmarking for Big Data.

In particular, as illustrated in Figure 2, benchmarks are classified according to benchmark categories (Micro- and Application benchmarks), their Year of publication, name, Type and domain, and Data type. Figure 2 provides a summary of the main characteristics of each selected benchmark. In the following sections, each one is described more in detail, according to the following dimensions: Description, Benchmark type and Domain, Workload, Data type and generation, Metrics, Implementation and technology stack, Reported results and usage, Reference papers.

While the work of describing more in detail all the selected benchmarks is ongoing, it is useful to present a summary illustration of each selected benchmark in this deliverable, as the analysis work was the basis for identifying the features and indicators that are proposed in Section 3 and the integrated framework discussed in Section 4 towards providing a description of benchmarking tools in both a business- and technology-related framework.

In the following, the selected benchmarks are described in detail: in Section 2.1 Micro-benchmarks are presented, while Section 2.2. presents Application benchmarks.

In this section in the descriptions the original terms and definitions from the benchmarks are reported.

Category	Year	Name	Type	Domain	Data Type	Metrics
Micro-benchmarks	2010	HiBench	Micro-benchmark Suite	Micro-benchmarks, Machine Learning, SQL, Websearch, Graph, Streaming Benchmarks	Structured, Text, Web Graph	Execution Time, Throughput
	2015	SparkBench	Micro-benchmark Suite	Machine Learning, Graph Computation, SQL, Streaming Application	Structured, Text, Web Graph	Execution Time, Throughput
	2010	YCSB	Micro-benchmark	cloud OLTP operations	Structured	Execution Time, Throughput
	2017	TPCx-IoT	Micro-benchmark	workloads on typical IoT Gateway systems	Structured, IoT	Performance Metric
Application Benchmarks	2015	Yahoo Streaming Benchmark	Application Streaming Benchmark	advertisement analytics pipeline	Structured, Time Series	Execution Time, Throughput
	2013	BigBench/ TPCx-BB	Application End-to-end Benchmark	a fictional product retailer platform	Structured, Text, JSON logs	Performance Metric
	2017	BigBench V2	Application End-to-end Benchmark	a fictional product retailer platform	Structured, Text, JSON logs	Performance Metric
	2018	ABench (Work-in-Progress)	Big Data Architecture Stack Benchmark	set of different workloads	Structured, Text, JSON logs	

Figure 2 - Benchmarks under evaluation

2.1.1 Micro-Benchmarks

HiBench

1. Description

HiBench [Huang, S] is a comprehensive big data benchmark suite for evaluating different big data frameworks. It consists of 19 workloads including both synthetic micro-benchmarks and real-world applications from 6 categories which are: micro, ml (machine learning), sql, graph, websearch and streaming.

2. Benchmark type and domain

Micro-benchmark suite including 6 categories which are micro, ml (machine learning), sql, graph, websearch and streaming.

3. Workload

- Micro Benchmarks: Sort (sort), WordCount (wordcount), TeraSort (terasort), Sleep (sleep), enhanced DFSIO (dfsioe)
- Machine Learning**:** Bayesian Classification (Bayes), K-means clustering (Kmeans), Logistic Regression (LR), Alternating Least Squares (ALS), Gradient Boosting Trees (GBT), Linear Regression (Linear), Latent Dirichlet Allocation (LDA), Principal Components Analysis (PCA), Random Forest (RF), Support Vector Machine (SVM), Singular Value Decomposition (SVD)
- SQL: Scan (scan), Join(join), Aggregate(aggregation)
- Websearch Benchmarks: PageRank (pagerank), Nutch indexing (nutchindexing)
- Graph Benchmark: NWeight (nweight)
- Streaming Benchmarks: Identity (identity), Repartition (repartition), Stateful Wordcount (wordcount), Fixwindow (fixwindow)

4. Data type and generation

Most workloads use synthetic data generated from real data samples. The workloads use structured and semi-structured data.

5. Metrics

The measured metrics are execution time (latency), throughput and system resource utilizations (CPU, Memory, etc.).

6. Implementation and technology stack

HiBench can be executed in Docker containers. It is implemented using the following technologies:

- Hadoop: Apache Hadoop 2.x, CDH5, HDP
- Spark: Spark 1.6.x, Spark 2.0.x, Spark 2.1.x, Spark 2.2.x
- Flink: 1.0.3
- Storm: 1.0.1
- Gearpump: 0.8.1
- Kafka: 0.8.2.2

7. **Reported results and usage:**

- Yi, L., & Dai, J. (2013, July). Experience from hadoop benchmarking with HiBench: from micro-benchmarks toward end-to-end pipelines. In Workshop on Big Data Benchmarks(pp. 43-48). Springer, Cham.
- Ivanov, T., Niemann, R., Izberovic, S., Rosselli, M., Tolle, K., & Zicari, R. V.. (2014). Benchmarking DataStax Enterprise/Cassandra with HiBench. Frankfurt Big Data Laboratory Technical Paper. (http://arxiv.org/ftp/arxiv/papers/1411/1411.4044.pdf
- Ivanov, T., Zicari, R. V., Izberovic, S., & Tolle, K. (2014). Performance Evaluation of Virtualized Hadoop Clusters. Frankfurt Big Data Laboratory Technical Paper. (http://arxiv.org/ftp/arxiv/papers/1411/1411.3811.pdf)
- Alzuru, I., Long, K., Gowda, B., Zimmerman, D., & Li, T. (2015, August). Hadoop Characterization. In Trustcom/BigDataSE/ISPA, 2015 IEEE (Vol. 2, pp. 96-103). IEEE.
- Samadi, Y., Zbakh, M., & Tadonki, C. (2016, May). Comparative study between Hadoop and Spark based on Hibench benchmarks. In Cloud Computing Technologies and Applications (CloudTech), 2016 2nd International Conference on (pp. 267-275). IEEE.
- Ahmed, H., Ismail, M. A., Hyder, M. F., Sheraz, S. M., & Fouq, N. (2016). Performance Comparison of Spark Clusters Configured Conventionally and a Cloud Service. Procedia Computer Science, 82, 99-106.

8. **Reference papers:**

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

SparkBench

1. Description

Spark-Bench is a flexible system for benchmarking and simulating Spark jobs. It consists of multiple workloads organized in 4 categories.

2. Benchmark type and domain

Spark-Bench is a Spark specific benchmarking suite to help developers and researchers to evaluate and analyze the performance of their systems in order to optimize the configurations. It consists of 10 workloads organized in 4 different categories.

3. Workload

The atomic unit of organization in Spark-Bench is the workload. Workloads are standalone Spark jobs that read their input data, if any, from disk, and write their output, if the user wants it, out to disk. Workload suites are collections of one or more workloads. The workloads in a suite can be run serially or in parallel. The 4 categories of workloads are:

- Machine Learning: logistic regression (LogRes), support vector machine (SVM) and matrix factorization (MF).
- Graph Computation: PageRank, collaborative filtering model (SVD++) and a fundamental graph analytics algorithm (TriangleCount (TC)).
- SQL Query: select, aggregate and join in HiveQL and RDDRelation.
- Streaming Application: Twitter popular tag and PageView

4. Data type and generation

The data type and generation is depending on the different workload. The LogRes and SVM use the Wikipedia data set. The MF, SVD++ and TriangleCount use the Amazon Movie Review data set. The PageRank uses Google Web Graph data and respectively Twitter uses Twitter data. The SQL Queries workloads use E-commerce data. Finally, the PageView uses PageView DataGen to generate synthetic data.

5. Metrics

SparkBench defines a number of metrics facilitating users to compare between various Spark optimizations, configurations and cluster provisioning options:

- Job Execution Time(s) of each workload
- Data Process Rate (MB/seconds)
- Shuffle Data Size

6. Implementation and technology stack

Spark-Bench is currently compiled against the Spark 2.1.1 jars and should work with Spark 2.x. It is written using Scala 2.11.8.

7. Reported results and usage

- Hema, N., Srinivasa, K. G., Chidambaram, S., Saraswat, S., Saraswati, S., Ramachandra, R., & Huttanagoudar, J. B. (2016, August). Performance Analysis of Java Virtual Machine for Machine Learning Workloads using Apache Spark. In Proceedings of the International Conference on Informatics and Analytics (p. 125). ACM.
- Liang, Y., Chang, S., & Su, C. (2017, December). A Workload-Specific Memory Capacity Configuration Approach for In-Memory Data Analytic Platforms. In Ubiquitous Computing and Communications (ISPA/IUCC), 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on (pp. 486-490). IEEE.

8. Reference papers:

- Min Li, Jian Tan, Yandong Wang, Li Zhang, Valentina Salapura: SparkBench: a spark benchmarking suite characterizing large-scale in-memory data analytics. Cluster Computing 20(3): 2575-2589 (2017)
- 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
- SparkBench, https://github.com/CODAIT/spark-bench

Yahoo! Cloud Serving Benchmark (YCSB)

1. Description

The YCSB framework is designed to evaluate the performance of different "key-value" and "cloud" serving systems, which do not support the ACID properties. The benchmark is open source and available on GitHub. The YCSB++ , an extension of the YCSB framework, includes many additions such as multi-tester coordination for increased load and eventual consistency measurement, multi-phase workloads to quantify the consequences of work deferment and the benefits of anticipatory configuration optimization such as B-tree pre-splitting or bulk loading, and abstract APIs for explicit incorporation of advanced features in benchmark tests.

2. Benchmark type and domain

The framework is a collection of cloud OLTP related workloads representing a particular mix of read/write operations, data sizes, request distributions, and similar that can be used to evaluate systems at one particular point in the performance space.

3. Workload

YCSB provides a core package of 6 pre-defined workloads A-F, which simulate cloud OLTP applications. The workloads are a variation of the same basic application type and using a table of records with predefined size and type of the fields. Each operation against the data store is randomly chosen to be one of:

- **Insert**: insert a new record.
- **Update**: update a record by replacing the value of one field.
- **Read**: read a record, either one randomly chosen field or all fields.
- **Scan**: scan records in order, starting at a randomly chosen record key. The number of records to scan is randomly chosen.

The YCSB workload consists of random operations defined by one of the several built-in distributions:

- **Uniform**: choose an item uniformly at random.
- **Zipfian**: choose an item according to the Zipfian distribution.
- **Latest**: like the Zipfian distribution, except that the most recently inserted records are in the head of the distribution.
- **Multinomial**: probabilities for each item can be specified.

4. Data type and generation

The benchmark consists of a workload generator and a generic database interface, which can be easily extended to support other relational or NoSQL databases.

5. Metrics

The benchmark measures the latency and achieved throughput of the executed operations. At the end of the experiment, it reports total execution time, the average throughput, 95th and 99th percentile latencies, and either a histogram or time series of the latencies.

6. **Implementation and technology stack**

Currently, YCSB is implemented and can be run with more than 14 different engines like Cassandra, HBase, MongoDB, Riak, Couchbase, Redis, Memcached, etc. The YCSB Client is a Java program for generating the data to be loaded to the database, and generating the operations which make up the workload.

7. **Reported results and usage:**

- Abubakar, Y., Adeyi, T. S., & Auta, I. G. (2014). Performance evaluation of NoSQL systems using YCSB in a resource austere environment. Performance Evaluation, 7(8), 23-27.
- Kumar, S. P., Lefebvre, S., Chiky, R., & Soudan, E. G. (2014, November). Evaluating consistency on the fly using YCSB. In Computational Intelligence for Multimedia Understanding (IWCIM), 2014 International Workshop on (pp. 1-6). IEEE.
- Rosselli, M., Niemann, R., Ivanov, T., Tolle, K., & Zicari, R. V.. (2015). Benchmarking the Availability and Fault Tolerance of Cassandra. Paper presented at the Big Data Benchmarking – 6th International Workshop, WBDB 2015, Toronto, ON, Canada, June 16-17, 2015.
- Fan, H., Ramaraju, A., McKenzie, M., Golab, W., & Wong, B. (2015). Understanding the causes of consistency anomalies in Apache Cassandra. Proceedings of the VLDB Endowment, 8(7), 810-813.

8. **Reference papers:**

- 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.
- Swapnil Patil, Milo Polte, Kai Ren, Wittawat Tantisiriroj, Lin Xiao, Julio López, Garth Gibson, Adam Fuchs, Billie Rinaldi: YCSB++: benchmarking and performance debugging advanced features in scalable table stores. ACM Symposium on Cloud Computing in conjunction with SOSP 2011, SOCC '11, Cascais, Portugal, October 26-28, 2011.
- YCSB, https://github.com/brianfrankcooper/YCSB

TPCx-IoT

1. **Description**

The TPC Benchmark IoT (TPCx-IoT) benchmark workload is designed based on Yahoo Cloud Serving Benchmark (YCSB). It is not comparable to YCSB due to significant changes. The TPCx-IoT workloads consists of data ingestion and concurrent queries simulating workloads on typical IoT Gateway systems. The dataset represents data from sensors from electric power station(s).

2. **Benchmark type and domain**

TPCx-IoT was developed to provide the industry with an objective measure of the hardware, operating system, data storage and data management systems for IoT Gateway systems. The

TPCx-IoT benchmark models a continuous system availability of 24 hours a day, 7 days a week.

3. Workload

The System Under Test (SUT) must run a data management platform that is commercially available and data must be persisted in a non-volatile durable media with a minimum of two-way replication. The workload represents data inject into the SUT with analytics queries in the background. The analytic queries retrieve the readings of a randomly selected sensor for two 30 second time intervals, TI_1 and TI_2. The first time interval TI_1 is defined between the timestamp the query was started T_s and the timestamp 5 seconds prior to T_S, i.e. $TI_1 = [T_S - 5, T_S]$. The second time interval is a randomly selected 5 seconds time interval TI_2 within the 1800 seconds time interval prior to the start of the first query, $T_S - 5$. If T_S <=1810, prior to the start of the first query, $T_S - 5$.

4. Data type and generation

Each record generated consists of driver system id, sensor name, time stamp, sensor reading and padding to a 1 Kbyte size. The driver system id represents a power station. The dataset represents data from 200 different types of sensors.

5. Metrics

TPCx-IoT was specifically designed to provide verifiable performance, price-performance and availability metrics for commercially available systems that typically ingest massive amounts of data from large numbers of devices. TPCx-IoT defines the following primary metrics:

- IoTps as the performance metric
- $/IoTps as the price-performance metric
- system availability date

6. Implementation and technology stack

The benchmark currently supports the HBase 1.2.1 and Couchbase-Server 5.0 NoSQL databases. A guide providing instructions on how to add new databases is also available.

7. Reference papers:

- TPCx-IoT, http://www.tpc.org/tpc_documents_current_versions/pdf/tpcx-iot_v1.0.3.pdf
- Nambiar, R.: Introducing the First Benchmark Standard for IoT - https://blogs.cisco.com/datacenter/tpc-iot
- Raghunath Nambiar, Meikel Poess: Reinventing the TPC: From Traditional to Big Data to Internet of Things. TPCTC 2015: 1-7.

2.1.2 Application Level Benchmarks

Yahoo Streaming Benchmark (YSB)

1. Description

The YSB benchmark is a simple advertisement application. There are a number of advertising campaigns, and a number of advertisements for each campaign. The benchmark reads the events in JSON format, processes and stores them into a key-value store. These steps attempt to probe some common operations performed on data streams.

2. Benchmark type and domain

The Yahoo Streaming Benchmark is a streaming application benchmark simulating an advertisement analytics pipeline.

3. Workload

The analytics pipeline processes a number of advertising campaigns, and a number of advertisements for each campaign. The job of the benchmark is to read various JSON events from Kafka, identify the relevant events, and store a windowed count of relevant events per campaign into Redis. The benchmark simulates common operations performed on data streams:

1. Read an event from Kafka.
2. Deserialize the JSON string.
3. Filter out irrelevant events (based on event_type field)
4. Take a projection of the relevant fields (ad_id and event_time)
5. Join each event by ad_id with its associated campaign_id. This information is stored in Redis.
6. Take a windowed count of events per campaign and store each window in Redis along with a timestamp of the time the window was last updated in Redis. This step must be able to handle late events.

4. Data type and generation

The data schema consists of seven attributes and is stored in JSON format:

- user_id: UUID
- page_id: UUID
- ad_id: UUID
- ad_type: String in {banner, modal, sponsored-search, mail, mobile}
- event_type: String in {view, click, purchase}
- event_time: Timestamp
- ip_address: String

5. Metrics

The reported metrics by the benchmark are:

- Latency as window.final_event_latency = (window.last_updated_at – window.timestamp) – window.duration
- Aggregate System Throughput

6. Implementation and technology stack

The YSB benchmark is implemented using Apache Storm, Spark, Flink, Apex, Kafka and Redis.

7. Reported results and usage (reference papers)

- Perera, S., Perera, A., & Hakimzadeh, K. (2016). Reproducible experiments for comparing apache flink and apache spark on public clouds. arXiv preprint arXiv:1610.04493.
- Venkataraman, S., Panda, A., Ousterhout, K., Armbrust, M., Ghodsi, A., Franklin, M. J., & Stoica, I. (2017, October). Drizzle: Fast and adaptable stream processing at scale. In Proceedings of the 26th Symposium on Operating Systems Principles (pp. 374-389). ACM.

8. Reference papers:

- 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 Workshops2016: 1789-1792.
- YSB, https://github.com/yahoo/streaming-benchmarks
- YSB Blog description, https://yahooeng.tumblr.com/post/135321837876/benchmarking-streaming-computation-engines-at

BigBench/TPCx-BB

1. Description

BigBench is an end-to-end big data benchmark that represents a data model simulating the volume, velocity and variety characteristics of a big data system, together with a synthetic data generator for structured, semi-structured and unstructured data. The structured part of the retail data model is adopted from the TPC-DS benchmark and further extended with semi-structured (registered and guest user clicks) and unstructured data (product reviews). In 2016, BigBench was standardized as TPCx-BB by the Transaction Processing Performance Council (TPC).

2. Benchmark type and domain

BigBench is an end-to-end, technology agnostic, application-level benchmark that tests the analytical capabilities of a Big Data platform. It is based on a fictional product retailer business model.

3. Workload

The business model and a large portion of the data model's structured part is derived from the TPC-DS benchmark. The structured part was extended with a table for the prices of the retailer's competitors, the semi-structured part was added represented by a table with website logs and the unstructured part was added by a table showing product reviews. The simulated workload is based on a set of 30 queries covering the different aspects of big data analytics proposed by McKinsey.

4. Data type and generation

The data generator can scale the amount of data based on a scale factor. Due to parallel processing of the data generator, it runs efficiently for large scale factors. The benchmark consists of four key steps: (i) System setup; (ii) Data generation; (iii) Data load; and (iv) Execute application workload.

5. Metrics

TPCx-BB defines the following primary metrics:

- BBQpm@SF, the performance metric, reflecting the TPCx-BB Queries per minute throughput; where SF is the Scale Factor.
- $/BBQpm@SF, the price/performance metric
- System Availability Date as defined by the TPC Pricing Specification

6. Implementation and technology stack

Since the BigBench specification is general and technology agnostic, it should be implemented specifically for each Big Data system. The initial implementation of BigBench was made for the Teradata Aster platform. It was done in the Aster's SQL-MR syntax served - additionally to a description in the English language - as an initial specification of BigBench's workloads. Meanwhile, BigBench is implemented for Hadoop, using the MapReduce engine and other components like Hive, Mahout, Spark SQL, Spakr MLlib and OpenNLP from the Hadoop Ecosystem.

7. Reported results and usage (reference papers)

- Todor Ivanov, Max-Georg Beer: Evaluating Hive and Spark SQL with BigBench.Frankfurt Big Data Laboratory Technical Paper. (http://arxiv.org/ftp/arxiv/papers/1512/1512.08417.pdf)
- Alzuru, I., Long, K., Gowda, B., Zimmerman, D., & Li, T. (2015, August). Hadoop Characterization. In Trustcom/BigDataSE/ISPA, 2015 IEEE (Vol. 2, pp. 96-103). IEEE.
- Singh, S. (2016, September). Benchmarking Spark Machine Learning Using BigBench. In Technology Conference on Performance Evaluation and Benchmarking (pp. 45-60). Springer, Cham.
- Nicolás Poggi, Alejandro Montero, David Carrera: Characterizing BigBench Queries, Hive, and Spark in Multi-cloud Environments. TPCTC 2017: 55-74.
- Nguyen, V. Q., & Kim, K. (2017). Performance Evaluation between Hive on MapKeduce and Spark SQL with BigBench and PAT. In Proceedings of KISM Spring Conference April (pp. 28-29).
- Richins, D., Ahmed, T., Clapp, R., & Reddi, V. J. (2018, February). Amdahl's Law in Big Data Analytics: Alive and Kicking in TPCx-BB (BigBench). In 2018 IEEE International

Symposium on High Performance Computer Architecture (HPCA) (pp. 630-642). IEEE.

8. Reference papers:

- 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
- Chaitanya K. Baru, Milind A. Bhandarkar, Carlo Curino, Manuel Danisch, Michael Frank, Bhaskar Gowda, Hans-Arno Jacobsen, Huang Jie, Dileep Kumar, Raghunath Othayoth Nambiar, Meikel Poess, Francois Raab, Tilmann Rabl, Nishkam Ravi, Kai Sachs, Saptak Sen, Lan Yi, Choonhan Youn: Discussion of BigBench: A Proposed Industry Standard Performance Benchmark for Big Data. TPCTC 2014: 44-63
- BigBench, https://github.com/intel-hadoop/Big-Data-Benchmark-for-Big-Bench
- TPCx-BB, http://www.tpc.org/tpc_documents_current_versions/pdf/tpcx-bb_v1.2.0.pdf

BigBench V2

1. Description

The BigBench V2 benchmark addresses some of the limitation of the BigBench (TPCx-BB) benchmark. BigBench V2 separates from TPC-DS with a simple data model. The new data model still has the variety of structured, semi-structured, and unstructured data as the original BigBench data model. The difference is that the structured part has only six tables that capture necessary information about users (customers), products, web pages, stores, online sales and store sales. BigBench V2 mandates late binding by requiring query processing to be done directly on key-value web-logs rather than a pre-parsed form of it.

2. Benchmark type and domain

Similar to BigBench, BigBench V2 is an end-to-end, technology agnostic, application-level benchmark that tests the analytical capabilities of a Big Data platform.

3. Workload

All 11 TPC-DS queries on the complex structured part are removed and replaced by simpler queries mostly against the key-value web-logs. The new BigBench V2 queries have only 5 queries on the structured part versus 18 in BigBench. This change has no impact on the coverage of the different business categories done in BigBench. In addition to the removal of TPC-DS queries, BigBench V2 mandates late binding, but it does not impose a specific implementation of it. This requirement means that a system using BigBench V2 can extract the keys and their corresponding values per query at run-time.

4. Data type and generation

A new scale factor-based data generator for the new data model was developed. The web-logs are produced as key-value pairs with two sets of keys. The first set is a small set of keys that represent fields from the structured tables like IDs of users, products, and web pages. The other set of keys is larger and is produced randomly. This set is used to simulate the real life cases of large keys in web-logs that may not be used in actual queries. Product reviews are produced and linked to users and products as in BigBench but the review text is produced synthetically contrary to the Markov chain model used in BigBench. Product reviews are generated in this way because the Markov chain model requires real data sets which limits our options for products and makes the generator hard to scale.

5. **Metrics**

BigBench V2 uses the same metric definition and computation as BigBench:

- BBQpm@SF, the performance metric, reflecting the TPCx-BB Queries per minute throughput; where SF is the Scale Factor.
- $/BBQpm@SF, the price/performance metric
- System Availability Date as defined by the TPC Pricing Specification

6. **Implementation and technology stack**

Similar to BigBench, BigBench V2 is technology agnostic and can be implemented for any system. Query implementations on Hive, Mahout, Spark SQL, Spark MLlib and OpenNLP from the Hadoop Ecosystem were reported in the paper.

7. **Reference papers:**

- 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.

2.2 BDVA framework and benchmarks

The Big Data Value Reference Model developed by BDVA (under the leadership of SINTEF and BDVA TF6) is being used as a foundation for the identification of different relevant areas in the context of benchmarking. The BDVA Reference Model from BDVA SRIA 4.0 is shown first, then we present and describe the extended version including domains and the placement of AI and data platforms that has been worked on during 2018.

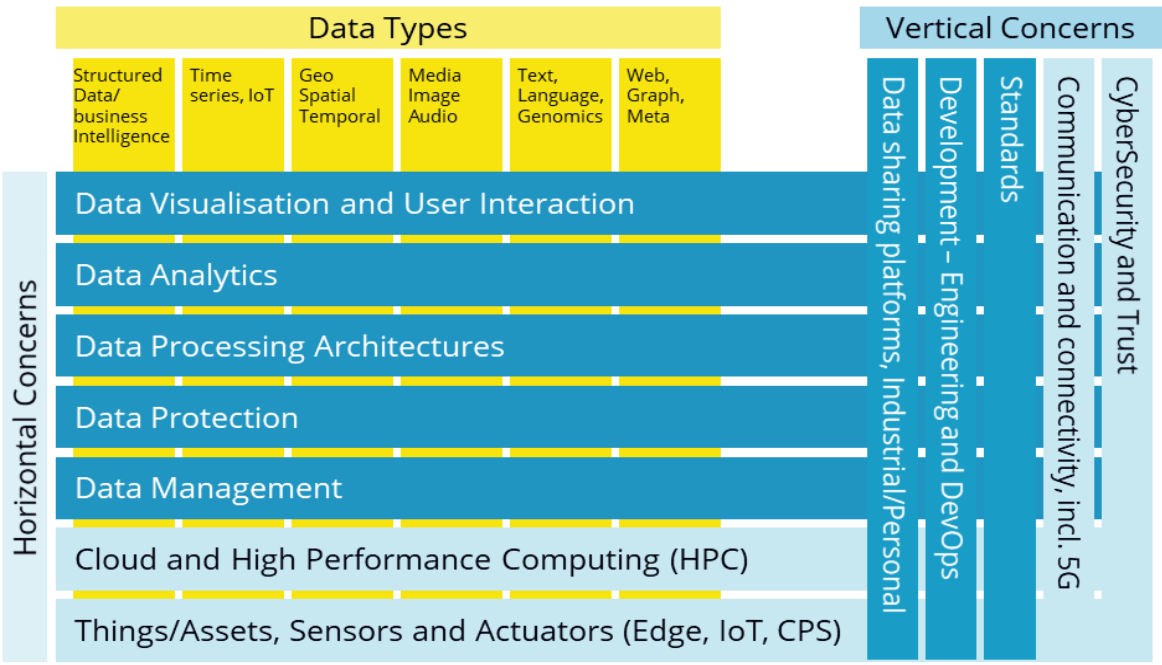

Figure 3 - BDV Reference Model from SRIA 4.0 (January 2018)

The BDV Reference Model illustrated different technical areas that are relevant for technical solutions, standards and potentially benchmarks (Figure 3).

The extended version of the Reference Model worked on during 2018 (Figure 4) extends this with showing domains and also the placement of AI and data platforms in this context.

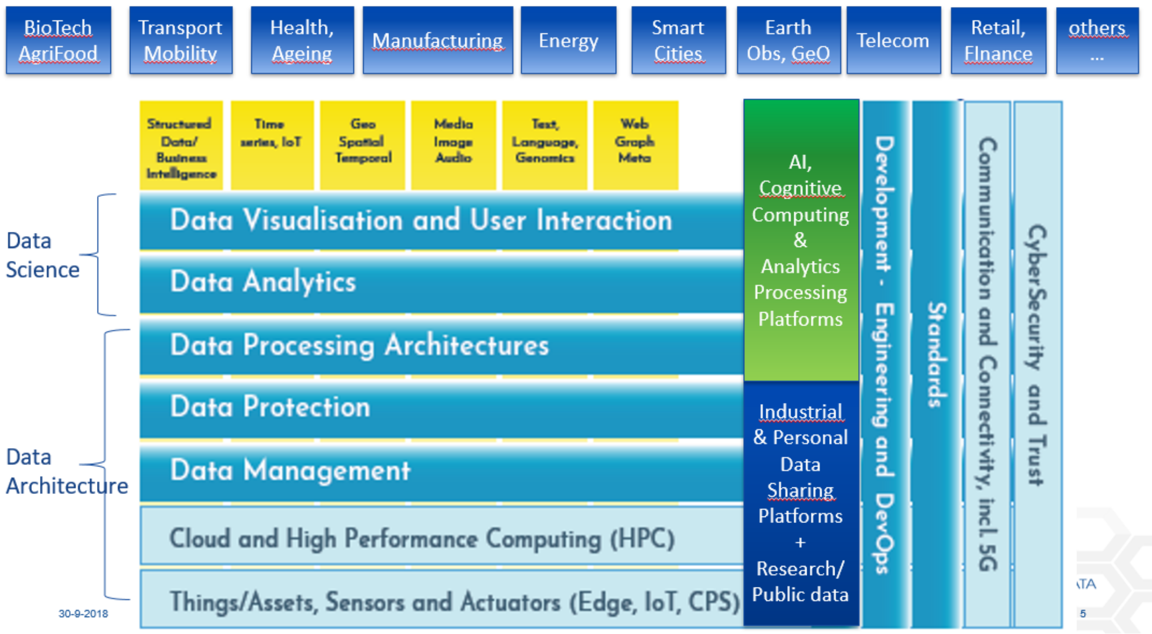

Figure 4 - Big Data Value Reference Model (extended version)

The BDV Reference Model has been developed by the BDVA, taking into account input from technical experts and stakeholders along the whole Big Data Value chain, as well as interactions with other related PPPs. The BDV Reference Model may serve as common reference framework to locate Big Data technologies on the overall IT stack. It addresses the main concerns and aspects to be considered for Big Data Value systems.

The BDV Reference Model distinguishes between two different elements. On the one hand, it describes the elements that are at the core of the BDVA; on the other, it outlines the features that are developed in strong collaboration with related European activities.

The BDV Reference Model shows on the top a number of relevant application domains. It also shows a logical placement of the areas of AI platforms and Data platforms.

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

- *Horizontal concerns* cover specific aspects along the data processing chain, starting with data collection and ingestion, and extending to data visualisation. It should be noted that the horizontal concerns do not imply a layered architecture. As an example, data visualisation may be applied directly to collected data (the data management aspect) without the need for data processing and analytics.
- *Vertical concerns* address cross-cutting issues, which may affect all the horizontal concerns, and also relates to how different big data types cuts across the horizontal areas. In addition, vertical concerns may also involve non-technical aspects.

It should be noted that the BDV Reference Model has no ambition to serve as a technical reference structure. However, the BDV Reference Model is compatible with such reference architectures, most notably the emerging ISO JTC1 SC42 AI and Big Data Reference Architecture.

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

Horizontal concerns

- Data Visualisation and User Interaction: Advanced visualisation approaches for improved user experience.
- Data Analytics: Data analytics to improve data understanding, deep learning and the meaningfulness of data.
- Data Processing Architectures: Optimised and scalable architectures for analytics of both data-at-rest and data-in-motion, with low latency delivering real-time analytics.
- Data Protection: Privacy and anonymisation mechanisms to facilitate data protection. This is shown related to data management and processing as there is a strong link here, but it can also be associated with the area of cybersecurity.
- Data Management: Principles and techniques for data management.
- The Cloud and High Performance Computing (HPC): Effective Big Data processing and data management might imply the effective usage of Cloud and High Performance Computing infrastructures.
- IoT, CPS, Edge and Fog Computing: A main source of Big Data is sensor data from an IoT context and actuator interaction in Cyber Physical Systems. In order to meet real-time needs it will often be necessary to handle Big Data aspects at the edge of the system. This area is separately elaborated further in collaboration with the IoT (Alliance for Internet of Things Innovation (AIOTI)) and CPS communities.

Vertical concerns

- Big Data Types and Semantics: The following 6 Big Data types have been identified, based on the fact that they often lead to the use of different techniques and mechanisms in the horizontal concerns, which should be considered, for instance, for data analytics and data storage: (1) Structured data; (2) Time series data; (3) Geospatial data; (4) Media, Image, Video and Audio data; (5) Text data, including Natural Language Processing data and Genomics representations; and (6) Graph data, Network/Web data and Metadata. In addition, it is important to support both the syntactical and semantic aspects of data for all Big Data types.
- Standards: Standardisation of Big Data technology areas to facilitate data integration, sharing and interoperability.
- Communication and Connectivity: Effective communication and connectivity mechanisms are necessary in providing support for Big Data. This area is separately further elaborated, along with various communication communities, such as the 5G community.
- Cybersecurity: Big Data often need support to maintain security and trust beyond privacy and anonymisation. The aspect of trust frequently has links to trust mechanisms such as blockchain technologies, smart contracts and various forms of encryption. Data protection has been identified as a focused area by BDVA and has

thus received its own horizontal area – with an associated set of ongoing research topics and projects. It could have been grouped also under Cybersecurity, but this has been kept as a separate area also because of the independent European research areas of trust and security and the separate ECSO – European Cyber Security Organisation.

- Engineering and DevOps for building Big Data Value systems: This topic will be elaborated in greater detail along with the NESSI Software and Service community.
- Marketplaces, Industrial Data Platforms and Personal Data Platforms (IDPs/PDPs), Ecosystems for Data Sharing and Innovation Support: Data platforms for data sharing include, in particular, IDPs and PDPs, but also other data sharing platforms like Research Data Platforms (RDPs) and Urban/City Data Platforms (UDPs). These platforms facilitate the efficient usage of a number of the horizontal and vertical Big Data areas, most notably data management, data processing, data protection and cybersecurity.

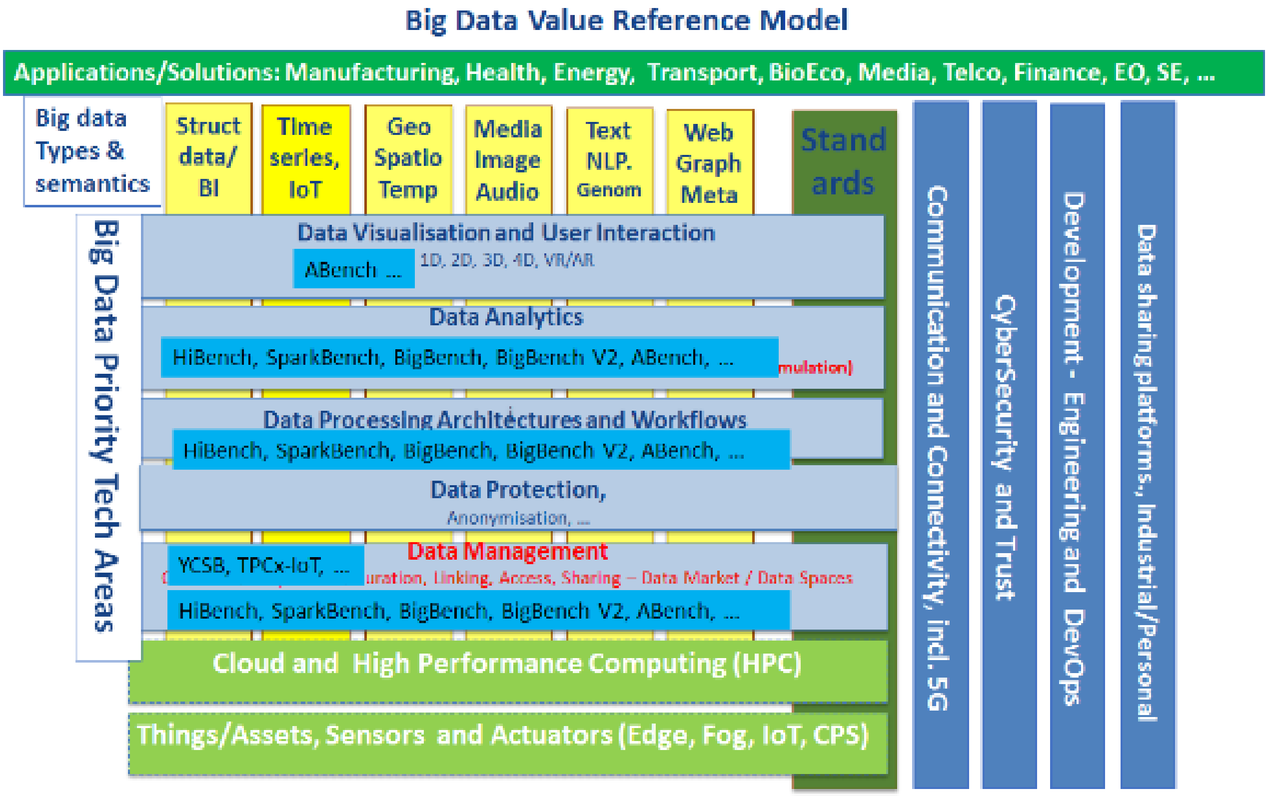

Figure 5 - Big Data Benchmarks mapped into some of the areas of the BDV Reference Model (D3.1)

Figure 5 (from D3.1) illustrates initial work, to be further developed in the project, on how the selected Big Data benchmarks we are investigating in the project can be mapped into some of the areas of the BDV Reference Model. This approach will be followed further in the DataBench Framework worked on in DataBench WP1 and WP3.

2.3 BDVA SG on Benchmarks

With the recognition of the importance of benchmarking within the BDVA community and the Big Data PPP, it was decided in March 2018 to establish a new Special Interest Group within the BDVA TF6 Technical Priorities called SG7 Benchmarking.

The leadership of this group was established as a co-chairing between Axel-C. Ngonga Ngomo of the HOBBIT project and Arne J. Berre of the DataBench project. Arne J. Berre is also the leader of the TF6 Technical Priorities.

One of the early results from this group was the creation and analysis of the questionnaire that is provided in Annex I.

The motivation and rationale for the SG7 Benchmarking group was described as follows:

Big Data is one of the key assets of the future. However, the cost and effort required for introducing Big Data technology in a value chain is significant. Mastering the creation of value from Big Data will enhance European competitiveness, will result in economic growth and jobs and will deliver societal benefit.

It is thus of utmost importance to reduce the costs and hurdles required to introduce Big Data processing into the European industry. A key step towards abolishing the barriers to the adoption and deployment of Big Data is to provide European companies with open benchmarking reports that allow them to assess the fitness of existing solutions for their purposes. However, achieving this goal demands:

- The deployment of benchmarks on data that reflects reality within realistic settings.
- The provision of corresponding industry-relevant key performance indicators (KPIs).
- The computation of comparable results on standardized hardware.
- The institution of an independent and thus bias-free organization to conduct regular benchmarks and provide the European industry with up-to-date performance results.

It is also a motivation that the technical benchmarks will provide a foundation for the better analysis of business level benchmarks and KPIs related to the adoption and usage of big data technologies. For this there will be an interaction with Business focused TFs/SGs in BDVA.

The context for this activity was described as follows:

The background for the proposed SG activity is the benchmarking framework derived from the HOBBIT project and synergies with the new Big Data PPP "DataBench" project and the needs for and experiments with big data technology benchmarking in various other projects and with BDVA member organisations.

The HOBBIT project has already established a set of Big Linked Data benchmarks that can be used in practice for a number of current activities and projects that are using linked data technologies. HOBBIT offers a set of benchmarks for each step of the Big Data Value Chain, namely Generation & Acquisition, Analytics & Processing, Storage & Curation and finally Visualization & Services.

Existing Big Data Benchmarking Communities to which DataBench will be related:

- TPC (http://www.tpc.org/) - Transaction Processing Performance Council
- SPEC (https://www.spec.org/) - Standard Performance Evaluation Corporation
- STAC (https://stacresearch.com/) - STAC Benchmark Council
- LDBC (http://www.ldbcouncil.org/) – Graph and semantic data benchmarks

- Hobbit Community (https://project-hobbit.eu)
- BigDataBench (http://prof.ict.ac.cn/)

There are also emerging communities in particular related to benchmarking of analytics/machine learning/AI that can be interacted with in the future.

There is also a logical link to the project coordination activities of Big Data PPP projects in the BDVe project, and the BDVe benchmarking activity.

The HOBBIT project has been running since 2016 and is ending in December 2018. The DataBench project started in 2018 and will run through 2019 and 2020. There is also an established contact with other international big data benchmarking communities.

The activities and expectations of this group were described as follows:

Activities:

- Provide benchmarks, Key Performance Indicators, benchmarking tools and services for the independent and repeatable benchmarking of big data technologies
- Facilitate the systematic evaluation, improvement and objective comparison of scalable big data solutions
- Generalization of knowledge from open-source benchmarking technologies
- Detect potential use cases and categories of users
- Detect potential synergies with benchmarking organizations, other big data benchmarking activities
- Requirement specifications from the association
- Producing open benchmarking reports

Expectations:

- Synergies, use case and datasets for big data benchmarks to enhance benchmarking framework and domains
- Ensure synergy of results from Big Data PPP Benchmarking projects like HOBBIT and DataBench related to the requirements and needs of the BDVA members and the Big Data community in general
- Promote the use of the HOBBIT framework for linked data, and also consider this as input for benchmarking of other big data types
- Generalized best practices, guidlines and standards to be offered as tutorials and support for the community

Initially planned tasks are as follows:

- Monitoring of European performance in Big Data technologies (e.g., through benchmarking campaigns, open challenges, dedicated benchmarking)
- Creation of high-impact white papers for the European industry on the current state of technology in domains of European importance
- Enhancing the community around big data benchmarking and standards
    - Revenue generation (membership strategies promoted through workshops, tutorials, surveys)
    - Identify Industrial Requirements from different industry sectors, including interviews for priorities and metrics

- Establish vertical holistic benchmarks – end-to-end for different Industry sectors
- Establish vertical benchmarks – Big Data Type specific
- Establish vertical benchmarks related to Data Privacy/ Security
- Analyse and adapt horizontal benchmarks for Analytics and Processing
- Analyse and adapt horizontal benchmarks for Data Management

This activity will relate to other BDVA TF/SG activities for the further detailing of business requirements related to economic, market and business metrics and KPIs for business performance – related also to the overall BDVA KPI measurements.

Multi-channel reach out strategy will be used including but not limited to:

- Dissemination and requirements gathering with surveys and interaction with projects and big data technology communities.
- Activities at BDVA and BDVe meetings and events.
- Distribution of flyers at different events.
- Organization of workshops, challenges.
- Cooperation with relevant projects.
- Social media interaction.

The questionnaire on business, technical, and benchmarking aspects developed within the BDVA Benchmarking group was issued in March 2018 and answers were collected in the period March-May 2018. Respondents were mainly participants in European PPP Big Data projects, for a total of 36 responders, representing 37 different projects.

The analysis of this first questionnaire, synthetically reported in Annex I, has been one of the sources for the assessment of suitable business and technical indicators and for the development of the DataBench survey designed and then performed within DataBench WP2 in Fall 2018.

3. DataBench ecosystem of Key Performance Indicators Classifications

3.1 The DataBench ecosystem of indicators

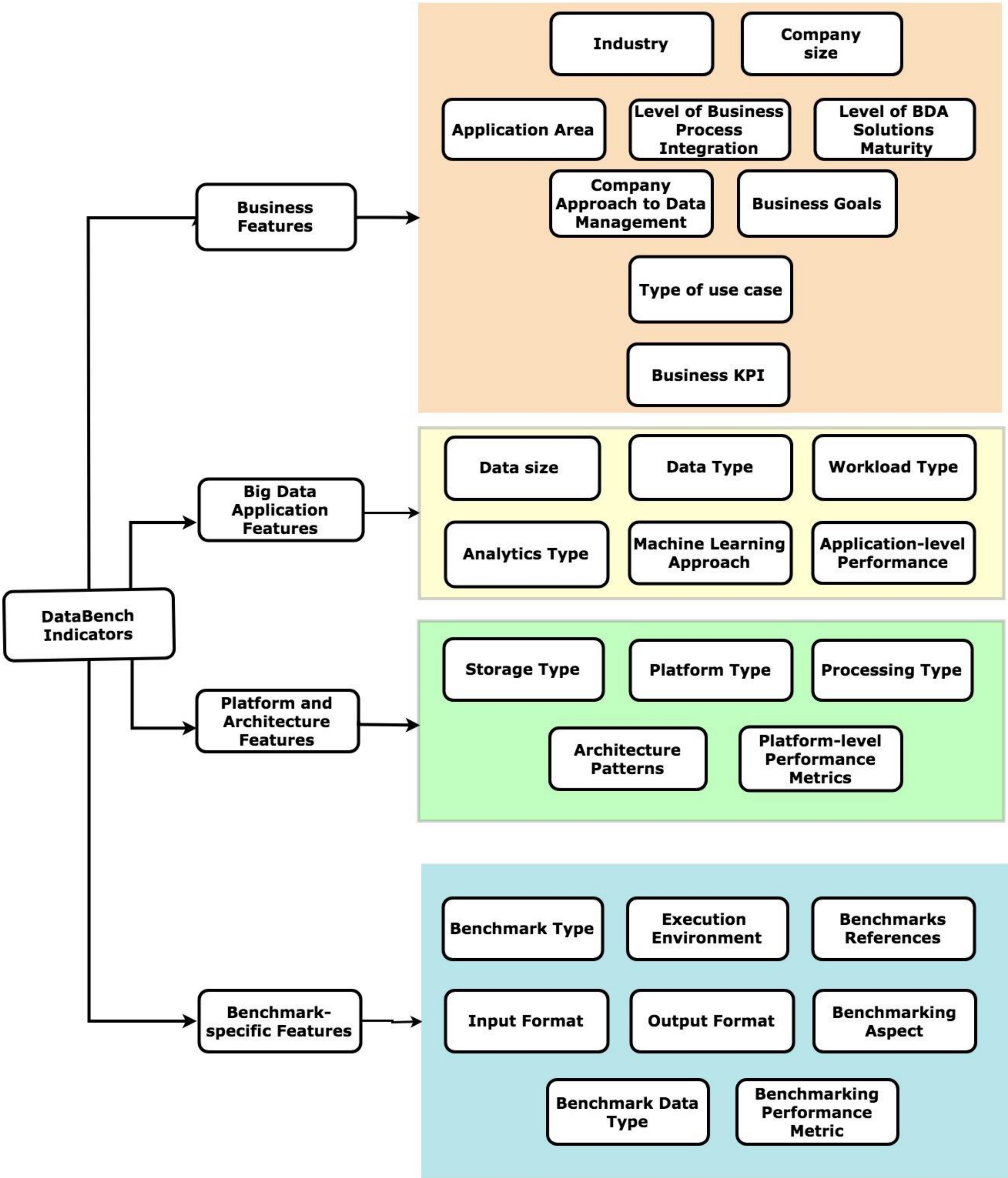

Table 1 - DataBench indicators ecosystem

In this section, we illustrate the ecosystem of indicators that has been derived in DataBench from the state of the art described in Section 2 and from the analysis activities being developed in the other Work Packages of the project.

As several indicators emerged from the analysis, we propose to classify them in four features, grouping relevant indicators from different points of views:

- Business features
- Big Data Application features
- Platform and Architecture features
- Benchmark-specific features.

For each feature, the specific indicators are defined, as illustrated in detail in the following sections. Table 1 - DataBench indicators ecosystem provides an overview of the indicators that have been selected.

For each of the indicators, further refinements can be defined:

- For each indicator, a set of possible values or categories is indicated in the following. This set can be refined and extended in the following of the project.
- More specific subclasses can be defined for each category, for instance Industry categories can be refined in more specific industry subcategories, and cross-industry Use cases can be defined, such as Fraud prevention and detection.
- For values, qualitative or quantitative values can be defined, with values or value ranges; for instance in Business Performance KPIs, for Costs the following qualitative values can be defined: Not at all important / Slightly important / Moderately important / Important / Extremely important.

In the following presentation, the focus is mainly on the features, the indicators for each feature (as illustrated in Table 1), and a description of possible values or categories for each indicator. Possible further refinements are discussed where relevant, and more detailed description are going to emerge in the following phases of the project.

3.2 Business features

3.2.1 Approach

In the DataBench indicators ecosystem, business features correspond to the main parameters used to identify and classify the typologies of Big Data & Analytics implementations in a business organization (use cases) and the performance metrics used to measure their business impacts (business KPIs). This methodology is presented in detail in the previous project deliverable D.2.1 Economic and Market Analysis. This chapter provides a summary description of these parameters in order to explain how they are positioned in the indicators ecosystem and how they will be used to correlate technical and business benchmarking. The description of the indicators is based on the most recent version operationalized in the business needs survey carried out by IDC in October 2018 (to be analysed in forthcoming deliverable D.2.2, due in December 2018).

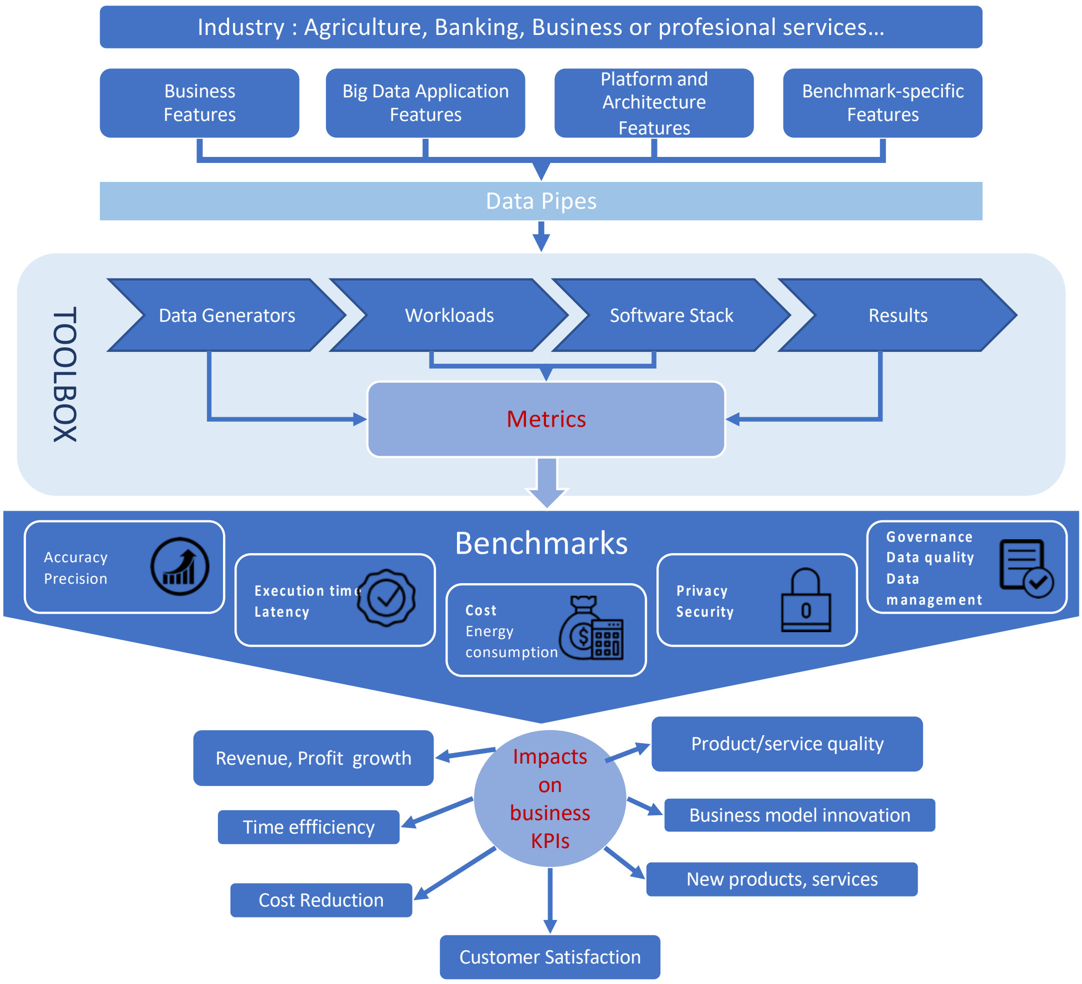

Figure 6 - BDA Technical and Business Benchmarking Framework (Source: DataBench 2018)

As shown in Figure 6 - BDA Technical and Business Benchmarking Framework (Source: DataBench 2018), DataBench will carry out a comprehensive review of the main BDT (Big Data Technologies) benchmarks by industry and technology (top layer of the figure). The analysis will feed into the benchmarking tool designed by the project, which will determine the optimal BDT benchmarking approaches by type of implementation (central layer of the figure). The tool will carry out the technical evaluation of benchmarks defining specific metrics. These metrics will be correlated through the use cases analysis and the case studies with their impact on the main business KPIs, such as revenues and profit growth, customer satisfaction, product and/or service innovation.

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

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

Examples of use cases are predictive maintenance in manufacturing, risk assessment in multiple industries, or industry-specific applications such as Yield monitoring and prediction in agriculture. Since a use case is based on a specific technology solution with specific technology performances, but at the same time it is easily correlated with business impacts, it provides a way to evaluate how technology requirements may influence business outcomes. Business users think in terms of use cases, not technologies: by using these concepts in its final Benchmarking Handbook, DataBench will be able to satisfy business needs while at the same time maintaining its alignment with scientific and technology best practice.

3.2.2 The survey

To ground the analysis in the European economic and industrial landscape, the study team carried out in September-October 2018 a survey of a casual sample of 700 European business organizations. The size of the sample has been decided in order to allow for an adequate reliability of results (margin of error 3.5% for the whole sample) and the cost (proportional to the overall budget of the project and the relevance of this task compared to the overall workplan). The list of countries surveyed has been selected based on the following criteria:

- Geographical balance (representing all main geographical areas in the EU)
- Country size (mix of large, medium and small Member States)
- IT maturity balance (mix of MS with high, medium and low intensity IT spending)
- Share of Data Market value (the MS selected represent 87% of the European data market value in 2017[1])
- Adequate coverage of the EU economy (the Member States surveyed together represent 76% of the EU GDP in 2017[2])

The geographical distribution of interviews allows extrapolating results to the whole EU28 economy by leveraging clusters of countries with similar socio-economic and Big Data usage characteristics.

1 Source: Update of the European Data Market Study, Facts and Figures report, January 2018, IDC

2 Sources: Eurostat data, EIU, EC EU growth, December 2017

The industry classification is based on Eurostat's NACE REV. 2 code in order to be able to use statistical data on value added and other parameters as well as IDC's Vertical Market databases. The following industries were excluded for the following reasons:

- Government: DataBench is focused on the private sector, government does not use the same business KPIs as the private sector, and the number of government agencies varies substantially from country to country so that Eurostat does not provide comparable statistics by number of entities.
- Education: a mostly public and no profit sector, very different from private industry, with vastly different dynamics of technology adoption by segment (for example, primary school vs research and university). Investigating it would have required a different type of survey and questionnaire.
- Finally, to achieve a reasonable sample size by industry we had to eliminate another industry and our choice fell on the construction industry which according to the EDM Monitoring tool statistics is a low user of BDT, is highly fragmented and would have required high screening efforts to identify data user companies.

The survey sample by company size finally excluded micro-enterprises under 10 employees (unlikely to be advanced adopters of BDT) since the objective was to focus on enterprises having already achieved concrete benefits from the use of Big Data and Analytics.

The results of the survey will be analysed and presented in the forthcoming DataBench D.2.2 "Preliminary benchmarks of European Economic and Industrial significance".

The final survey sample is shown in the Figure below and is adequately balanced.

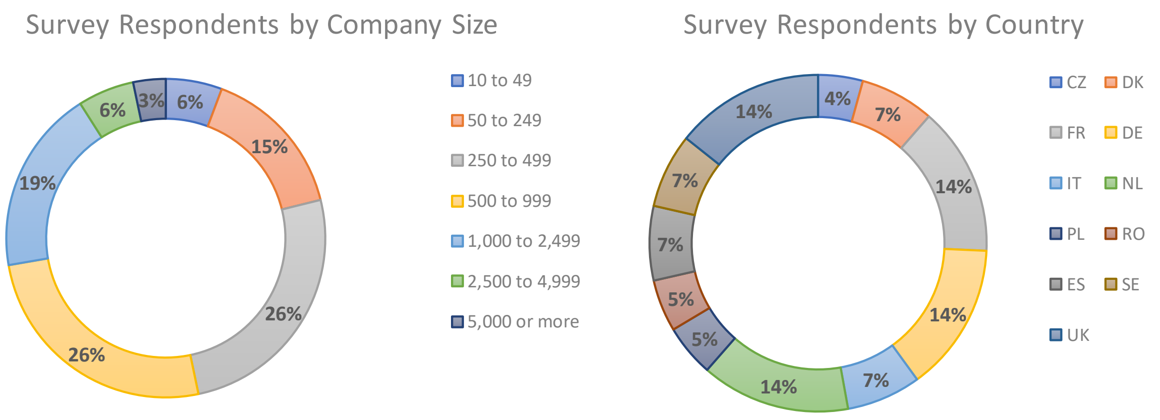

Figure 7 - Composition of the Survey Sample by size and country (Source: IDC, 700 Interviews, October 2018)

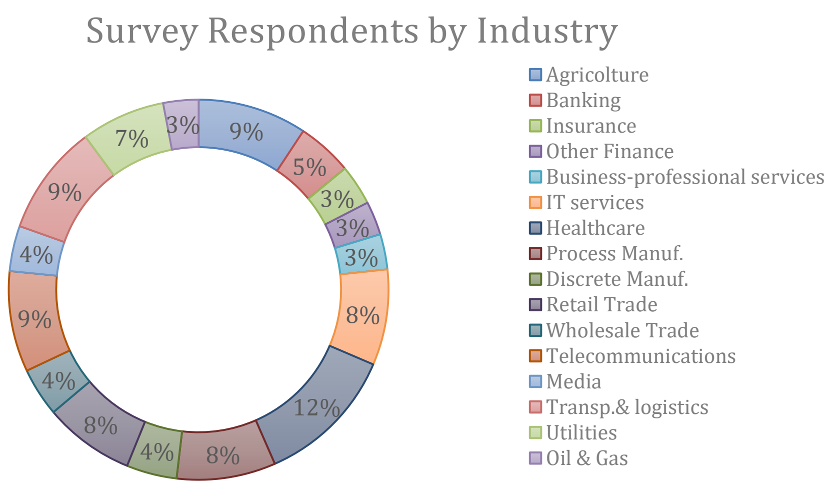

Figure 8 - Composition of the Survey Sample by industry (Source: IDC, 700 Interviews, October 2018)

3.2.3 Business Indicators

The business features indicators can be divided in the following main groups:

1. Classification of business users (industry and company size).
2. Type of BDA implementation (Application area, Level of Business Process integration, Level of BDA Solutions Maturity, Company approach to data management, main business goals).
3. Type of use case (cross-industry and industry-specific).
4. Business Impact KPIs.

The four groups are represented in Table 1, showing the relevant indicators grouped together. The indicators categories are presented in detail in the Figures 6 and 7 below. Groups 1, 2, 3 (Figures 9,10 and Tables 2,3) are semantic indicators measured through simple nominal questions (business users select the category in which they belong) to classify users. The survey results are measured as frequencies of respondents by category. Descriptive parameters can be used to measure the correlation between type of user and type of application and in turn type of business impacts. They will be used in the Benchmarking tool as a user interface to guide users to identify themselves and their type of BDA application, and in turn to look for the type of technical benchmark most relevant for them.

The use cases (group 3, Tables 2 and 3) represent the link between technical solutions and business goals. The potential list is extremely long, with a long tail of specific use cases. For the sake of this project we have selected 12 cross-industry use cases and 23 industry-specific use cases, representing the most frequent and potentially impactful typologies identified so far by IDC research.

The business impact KPIs (group 4) are 7 indicators selected on the basis of business literature and research as the most relevant for measuring innovative technologies impacts.

They are measured as simple numeric values mainly in percentage (percentage of improvement).

Industry

- Agriculture
- Banking, Insurance, Other financial services
- Business or professional services, excluding IT services
- IT services
- Healthcare
- Manufacturing process
- Manufacturing discrete
- Retail trade
- Wholesale trade
- Telecommunications
- Media
- Transport and logistics
- Utilities
- Oil & Gas

Company Size

- 10 to 49 employees
- 50 to 249 employees
- 250 to 499 employees
- 500 to 999 employees
- 1,000 to 2,499 employees
- 2,500 to 4,999 employees
- 5,000 or more employees

Application Area

- Customer service and support
- Engineering
- Research and development (R&D)
- Product innovation (new business initiatives)
- Maintenance and logistics
- Marketing
- Finance
- HR and legal
- Sales
- Product management
- Governance, risk, and compliance
- IT and data operations

Level of Business Process Integration

- High (where there is real time integration with business processes for eg real time fraud detection)
- Medium (where there are mixed levels of integration with business processes eg propensity models available as part of business processes but not scored in real time)
- Low (for eg where big data reports and dashboards are processed in a batch environment and made available the following day)

Figure 9 - Business Parameters: Industry, Application area, Level of business process integration

Level of BDA solutions Maturity

- Currently using
- Piloting or implementing
- Considering or evaluating for future use
- No use and no plans

Business KPI

- Increase in the number of products/services launched
- Customer satisfaction
- Business model innovation
- Revenue and profit growth
- Product/service quality
- Time Efficiency
- Cost reduction

Business Goals

- Better understand customer behavior and expectations
- Optimize our pricing strategies and go-to-market programs
- Product, services, or program improvement and innovation
- Improve our understanding of the market and our competitors
- Improve and optimize our business processes and operations
- Improve our facilities, and equipment design, maintenance, and utilization
- Improve our operational, fraud, and risk management
- Implement better regulatory compliance and financial controls

Approach to data management

- Structured and transactional data is captured and curated within a data warehouse or operational data store
- Structured and unstructured data from different locations, including on-premise and in the cloud, is captured and curate
- Enterprisewide repositories or data lakes are used to capture, organize, and curate data from multiple sources and forma
- Real-time data (for example, log files, social media, and IoT data) is streamed and used alongside other contextual data
- The Big Data platform is available/exposed to external customers/partners and developers to build and extend data-driven

Figure 10 - Business Parameters: Maturity, Business KPI, Business Goals, Approach to Data Management

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

Table 2 – Classification of BDA Cross-industry Use Cases

(Source: IDC User Needs Survey, 2018)

Industry	Specific Use Cases	Industry	Specific Use Cases
Agriculture	Precision agriculture Yield monitoring and prediction Field mapping & crop scouting Heavy equipment utilization	Retail Trade	Intelligent Fulfillment
Banking	Cyberthreat & detection	Wholesale Trade	Intelligent Fulfillment Increase productivity and efficiency of DCs/warehouses
Insurance	Usage based insurance	Telecommunications	Network analytics and optimization
Other Financial Services	Cyberthreat & detection	Media	Ad Targeting Scheduling optimisation
Business or Professional services	Social media analytics	Transport & Logistics	Connected vehicles optimization Logistics and package delivery management
Healthcare	Illness/disease diagnosis and progression Personalized treatment via comprehensive evaluation of health records Patient admission and re-admission predictions Quality of care optimization	Utilities	Field service optimization Energy consumption analysis and prediction
Manufacturing Process	Smart warehousing Asset management Quality management investigation	Oil & Gas	Field service optimization Energy consumption analysis and prediction
Manufacturing Discrete	Smart warehousing Asset management Quality management investigation Connected vehicles optimization		

Table 3 Classification of Industry-Specific BDA Use Cases

(Source: IDC User Needs Survey, 2018)

3.2.4 Scope of BDA: the data-driven company

Finally, based on the combination of technology and business indicators, we aim to provide a synthetic assessment of how the use of Big Data and Analytics impacts the organization business strategy. The assumption to be tested is that a higher level of integration of BDT in business process is correlated with a higher level of benefits, that is higher positive business impacts.

The suggested classification is based on the following stages of development of the implementation of BDT in the organization:

- Ad-hoc BDT implementations optimizing decision-making tasks;
- Implementation of data oriented digital transformation processes: these are the activities that lead an enterprise to be able to adopt a certain BDT and to properly manage data in digital format, which represents a pre-condition to build data-driven business processes. Taking full advantage of a certain BDT implies certain degrees of maturity for the target enterprise and its major resources;
- Implementation of data-driven business processes: organizational processes that include data management activities targeted to data analytics and their integration within other operational business processes.

The validity and usefulness of these BDT implementation stages will be fine-tuned and validated by DataBench particularly through the case studies.[3]

3 More details in D.2.1 Economic and Market Analysis

3.3 Big Data application features

The goal of the Big Data Application features is to describe the exact application environment and its requirements that can be later used in the process of selecting a suitable Big Data benchmark. The features depict properties of the system and implementation properties typical for the top application layer of the architecture.

Data Size	Data Type	Workload Type	Analytics Type	Machine Learning Approach	Application-level Performance
Gigabytes Terabytes Petabytes Exabytes	Tables, files or structured data Text data Graphs or linked data Geospatial or temporal data Media (images, audio, video) Time series (including IoT) Structured text	Online transaction processing (OLTP) Online analytical processing (OLAP) Hybrid transaction/analytical processing (HTAP)	Descriptive Diagnostic Predictive Prescriptive	Deep Learning Kernel Methods Tree-based Methods Clustering Latent Factor Models Hybrid Machine Learning Bayesian and Neural Networks	Cost Throughput End-to-end Execution Time Data quality (Accuracy/quality/data quality/veracity) Availability

Table 4 - Big Data application features

- **Data Size**: measures the data volume of the application data.
- **Data Type**: depicts the type of data that the application is processing and storing.
- **Workload Type**: describes the typical application operations in terms of processing.
- **Analytics Type**: outlines the main analytics category of the application.
- **Machine Learning Approach**: outlines the main approach and algorithms in case of machine learning usage.
- **Application-level Performance**: describes the metrics used to measure and monitor the application performance.

3.4 Platform and Architecture Features

The Platform and Architecture features describe in detail the system backend architecture on which the application is hosted including the processing, storage and management components. Providing details for all features will help to perform a more precise selection process.

Storage Type	Platform Type	Processing Type	Architecture Patterns	Platform-level Performance
Distributed File System	Distributed	Batch	Data Preparation	Execution time/ Latency
Databases/ RDBMS	Centralized	Stream	Data Pipeline	Throughput
NoSQL	Spark	Interactive/(near) Real-time	Data Lake	Cost
NewSQL/ In-Memory	Flink	Iterative/In-memory)	Data Warehouse	Energy consumption
Time Series Databases			Lambda Architecture	Accuracy
			Kappa Architecture	Precision
			Unified Batch and Stream architecture	Availability
				Durability
				CPU and Memory Utilization

Table 5 – Platform and Architecture features

- **Storage Type**: describes the type of system used to persistently store the application data.
- **Platform type**: indicates the type of platform in terms of category or particular technology stack.
- **Processing Type**: describes what type of processing is supported by the platform.
- **Architecture Patterns**: depicts the type of architecture pattern implemented in the system backend and hosting the application.
- **Platform-level Performance Metrics**: describes the metrics used to measure and monitor the platform and architecture performance.

3.5 Benchmark-specific Features

The Benchmark-specific features extend the Application, Platform and Architecture features defined above to depict a more precise view of the user requirements for a Big Data benchmark. The specific features focus on typical Big Data benchmark characteristics covering the input and output data, execution settings as well as metrics.

Benchmark Type	Execution Environment	Benchmark References	Input Data Format	Output Data Format	Benchmarking aspect	Benchmark Data Type	Benchmarking Perf. Metrics
Micro-benchmark Application benchmark Benchmark suite	Sandbox/ VM Inhouse/ On-premise Cloud	Execution in Toolbox Downloads Links References	JSON XML CSV Proprietary	Execution Log	Fault-tolerance Privacy Security Governance Data Quality Veracity Variability Data Management Data Visualization	Synthetic data Real data Hybrid (mix of real and synthetic) data	Execution time/ Latency Throughput Cost Energy consumptio n Accuracy Precision Availability Durability CPU and Memory Utilization

Table 6 - Benchmark-specific features

- **Benchmark Type**: identifies the category of the benchmark
- **Execution Environment**: describes the environment settings in which the benchmark is typically executed.
- **Configuration**: defines particular configuration properties of the benchmark.
- **Benchmark References**: links and references to existing best practices, how-tos, and experimental papers using the benchmark as well as links to the benchmark home page.
- **Input Data Format**: defines the input data file formats used by the benchmark.
- **Output Data Format**: defines the resulting output data produced and reported by the benchmark.
- **Benchmarking Aspect**: defines the stress test characteristics for which the benchmark can be applied.
- **Benchmark Data Type**: specifies the type of data used by the benchmark.
- **Benchmarking Performance Metrics**: defines the type of metrics that the benchmark measures and reports to the user.

4. Towards an integrated framework

4.1 Methodological integration framework

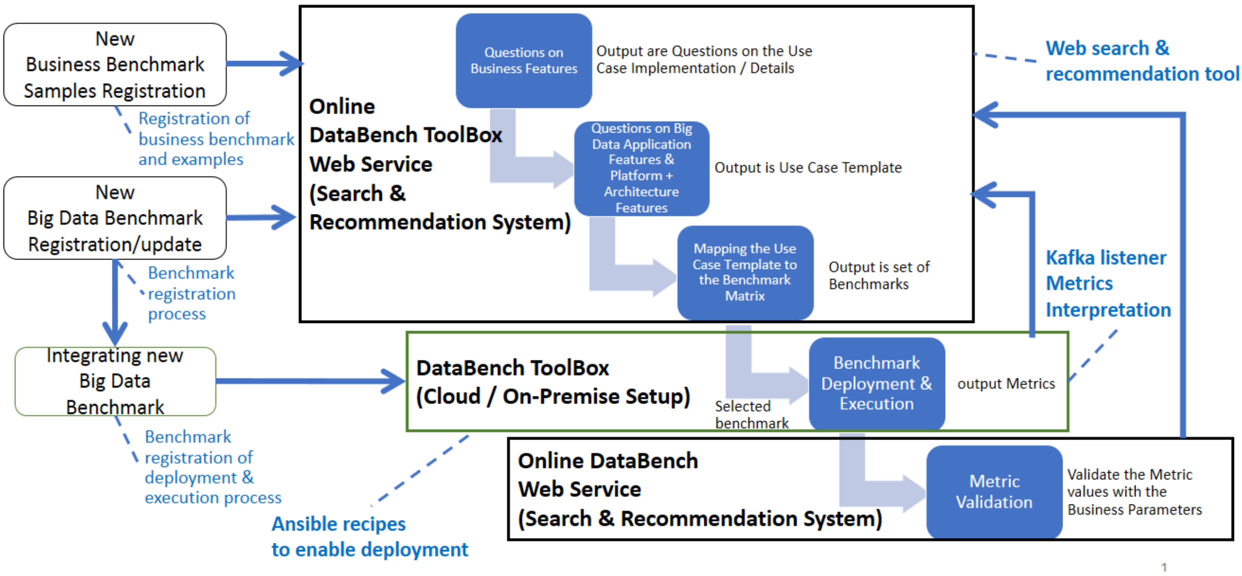

Figure 11 - DataBench methodological framework

Figure 11 shows a schema of processes intended to illustrate different elements of the tooling support to be provided in DataBench to different set of users. A single user may have different roles, as identified in D3.1, initially the following:

- Benchmarking Providers: Organizations that own a particular benchmark. They can be the actual developers of the benchmark or the organizations that maintain them. These users can register and update their benchmarks.
- Technical Users: Users that would like to search and potentially execute a technical benchmark. This includes the possibility of searching, downloading, executing and giving the results of the execution back to the Toolbox.
- Business Users: Users that would like to search and understand the business value of specific big data solutions. These users would not need to run technical benchmarks, but rather search for similar cases, business indicators, etc.
- DataBench Admin: People in charge of the administration of the Toolbox.

There are several processes depicted in Figure 11. On the left-hand side of the figure, the three boxes represent the registration process of two different kinds of benchmarks:

- The registration of data related to business-oriented big data benchmarks. The idea of the component located in the upper left corner of the figure ("New Business Benchmark Samples Registration") is to capture domain and industry specific best practices and blueprints associated to concrete business KPIs.
- The registration of technical benchmarks. The two remaining components on the left represent the way the DataBench Toolbox will capture the necessary metadata and features about technical benchmarks to enable the search and recommendation

processes ("New Big Data Benchmark Registration/Update" component), and to enable the automation of the deployment and the interpretation of the results of the execution of the benchmarks ("Integrating new Big Data Benchmark" component). Note that the registration of the automation provided by the second component is optional, in the sense that it requires the provision of deployment recipes and rules of interpretation of the results of the execution of the benchmarks which could prove a difficult task for some of the benchmarks analysed so far. However, the aim in DataBench is to automate as many as possible technical benchmarks, so the documentation of the process to integrate the automation will be also a key part for future extensibility to other benchmarks.

The components in the center of the Figure 11 show the full process from searching to executing and visualizing the results of benchmarks. The processes related to the DataBench Toolbox have been introduced in deliverable D3.1, while the validation of metrics is going to be introduced in deliverable D5.1. This process is divided into the following steps:

- Search and Recommendation System: The upper central box shows the steps to define the search criteria a user could pose to the system with the aim to select a benchmark that suits their needs. Based on those criteria (technical, business, application or platform features as explained in Section 3), the system will offer a set of potential benchmarks that could fulfil the user needs, as well as associated material (blueprints, best practices in sectors, etc.) that might facilitate the decision of the selection of the right benchmark.
- The DataBench Toolbox setup: The middle central box (in green in Figure 11) represents the process of deploying and enabling the execution either in cloud or in-premise of the selected benchmark. This could only happen if the registration of that benchmark provided the necessary recipes to allow the deployment. After the execution, the results of the benchmark will be sent back to the Toolbox for post-processing.
- The validation of the metrics: This process will allow in certain cases the matching of the technical metrics with business insights or KPIs. The results of the benchmarks will be then visualized and compared to others, giving the user a clear added-value in comparison with the mere technical results that the execution of a technical benchmark may provide.

At the point of writing this document, partners are in the process of agreeing and prototyping the look and feel of the different processes listed in this section. In order to do so, the figures below show mock-ups to describe the registration process, showing examples of how different features listed in Section 3 could be established. These mock-ups are intended as examples of the type of interactions the users registering benchmarks may have, and therefore serve the purpose of illustration of the processes described in this document before starting the actual implementation of the DataBench Toolbox.

For example, Figure 12 shows the beginning of the registration of a new benchmark as the actual realization of the first steps of the component "New Big Data Benchmark Registration/Update" listed in Figure 11. Users performing the registration of the new benchmark, typically the "Benchmarking Provider" or the "DataBench Admin" on their behalf, will go through several web forms to provide the necessary features to describe the benchmark for further search and recommendation purposes. In this particular case, Figure

12 shows some business features such as the industries for which the benchmarks are intended, sectors, degree of maturity, etc. These features may apply or not to a particular benchmark, but overall the idea is to enable the categorization of the new benchmark with the complete set of features to enable search and further recommendation.

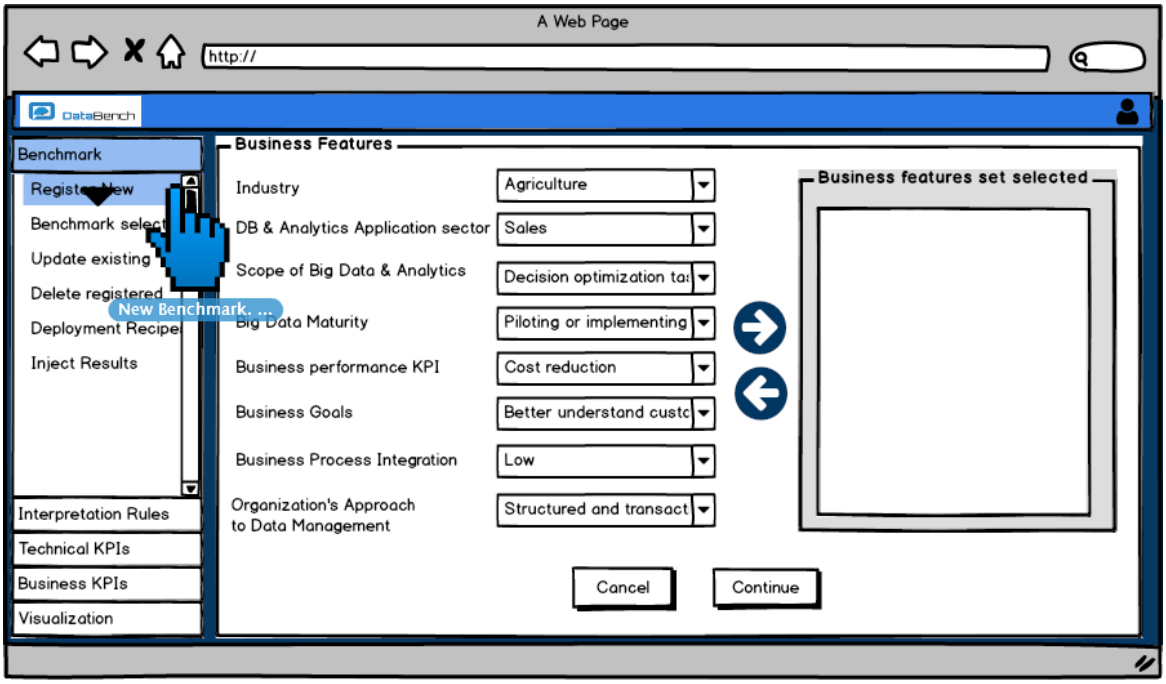

Figure 12 - DataBench mock-up of the start of the registration of a new benchmark

The example process continues until all different types of features listed in Section 3 have been established for the new benchmark. At that point, the initial registration is finished and the benchmark is searchable by end users of the Toolbox.

However, if the benchmark provider wishes to go a step further and automate the process of deploying and enabling the execution of the benchmark from the Toolbox, they should continue providing the rules of interpretation of the results and providing the Ansible recipes for deployment. An example of interpretation rule definition is shown in Figure 13. In this case, the user selects one of the technical output results of the benchmark, in this case "throughput", and associates a certain threshold to qualify the output in a measurable way. In the example shown in Figure 13, a throughput higher that 100 means in this particular benchmark that the throughput is considered high in a scale ranging from 1 (very low) to 5 (very high). The use of normalized scales for specific benchmarks will therefore allow having a way of comparing heterogeneous results from different benchmarks.

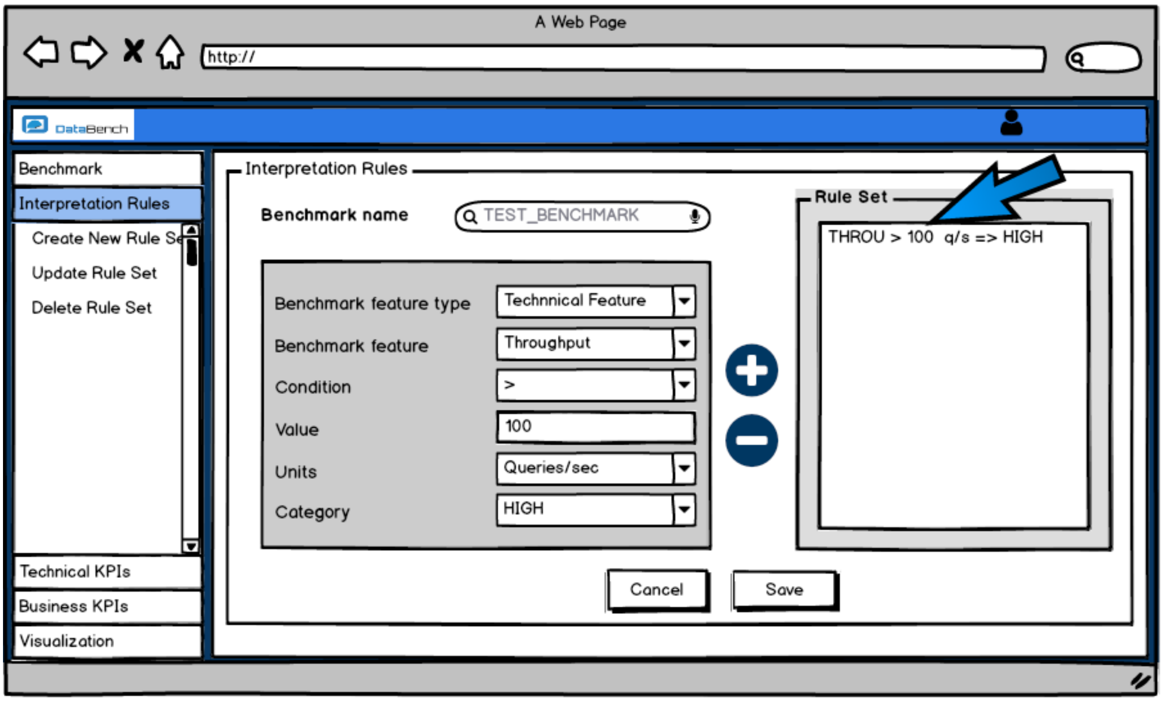

Figure 13 - DataBench mock-up of the adding automation (interpretation rules)

After defining the interpretation rules for all the output results and the recipes for deployment, the benchmark is ready to be automated from the Toolbox. Technical users may therefore use the DataBench search and recommendation engines to find, deploy and execute their benchmarks, and provide the results back to the Toolbox. These results will be validated and processed giving the possibility to be compared with others and derive business insights as added value to both Technical and Business users.

4.2 Relating indicators

In this section, we delineate possible directions to relate indicators, based on the performed analyses. In particular, we focus on the survey performed in Fall 2018 in WP2, on the analysis of different benchmarks, and on the ongoing desk analysis.

Considering the set of indicators presented in Section 3, some initial considerations may be drawn on the sets of indicators used in the WP2 survey and in the benchmark analysis. As illustrated in Table 7, indicators in the business features category are typical of the business and market analysis of WP2 and Benchmark-specific features are used in the description of the benchmarks. The other features, both for Big Data Applications indicators, and for Platform and architecture features, are common to both analyses. This overlap allows performing further analyses to relate not only business indicators among themselves, as shown for instance in Figure 14, which shows the KPIs that contribute most to business goals, but also the contribution to business KPIs improvement related to technical measures, illustrated in Figure 15 and Figure 16.

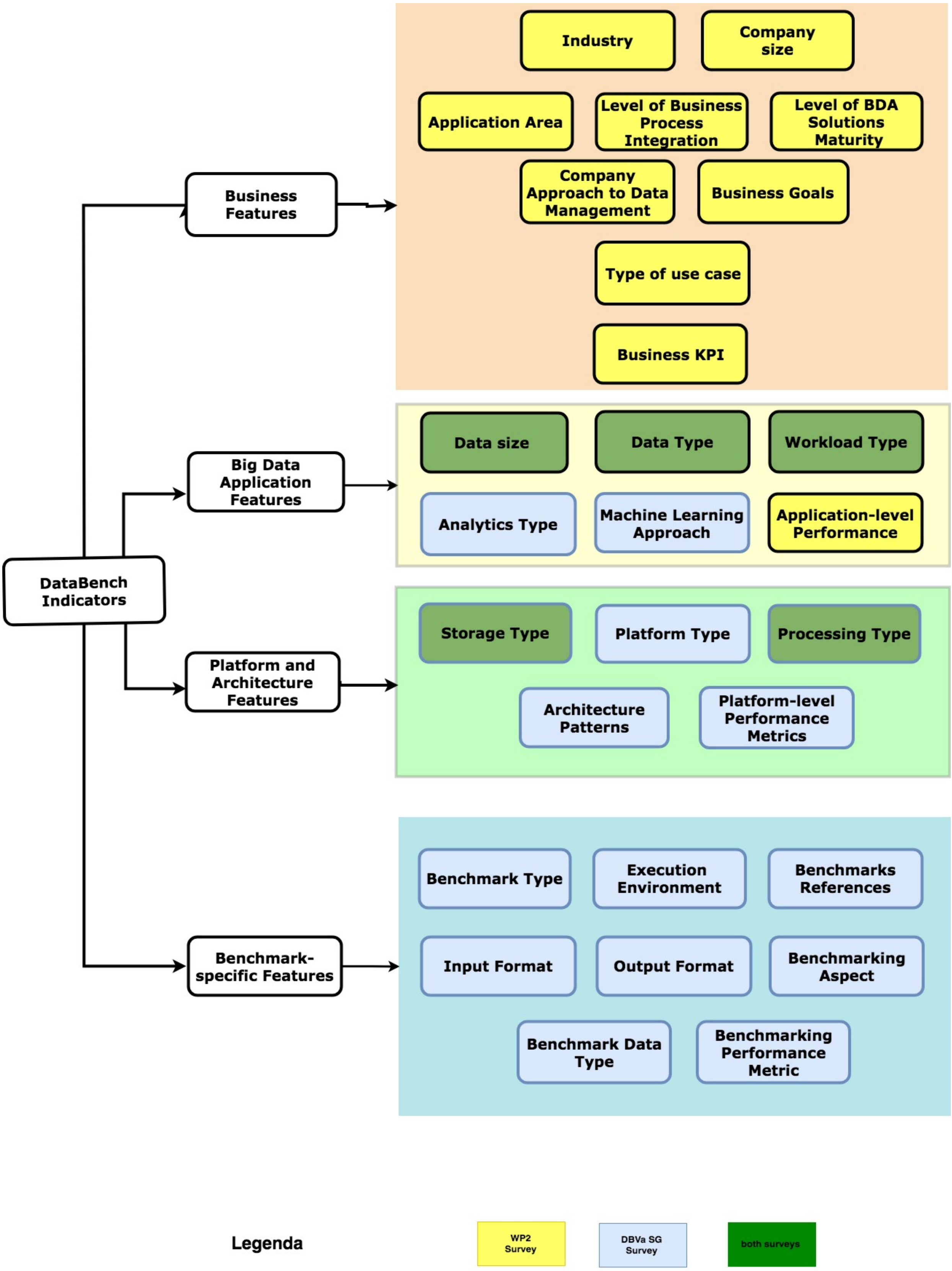

Table 7 – Comparing indicators contained in the WP2 survey and in benchmark descriptions

Rating	Expected Increase
1	None (0%)
2	Less than 5%
3	5%-9%
4	10%-24%
5	25%-49%

Business Goal

Key Performance Indicator	Better understand customer behavior and expectations	Optimize our pricing strategies and go-to-market programs	Product, services, or program improvement and innovation	Improve our understanding of the market and our competitors	Improve and optimize our business processes and operations	Improve our facilities, and equipment design, maintenance, and utilization	Improve our operational, fraud, and risk management	Implement better regulatory compliance and financial controls
Cost reduction	3.31	3.19	3.37	3.23	3.29	3.19	3.22	3.20
Time efficiency	3.76	3.84	3.91	3.87	3.84	3.83	3.83	3.92
Product/service quality	4.12	4.15	3.96	4.28	4.12	4.04	4.14	4.16
Revenue growth	4.03	4.06	3.98	4.09	4.03	3.99	4.11	4.06
Customer satisfaction	4.08	4.20	4.05	4.09	4.16	4.08	4.06	4.11
Business model innovation	3.60	3.64	3.67	3.71	3.65	3.65	3.65	3.78
Number of new products/services launched	3.71	3.78	3.78	4.01	3.88	3.80	3.83	3.94

Lowest — Key Performance Indicator that contributes most to achieving Business Goal — Highest

Figure 14 - KPI that contribute most to business goals (WP2 survey)

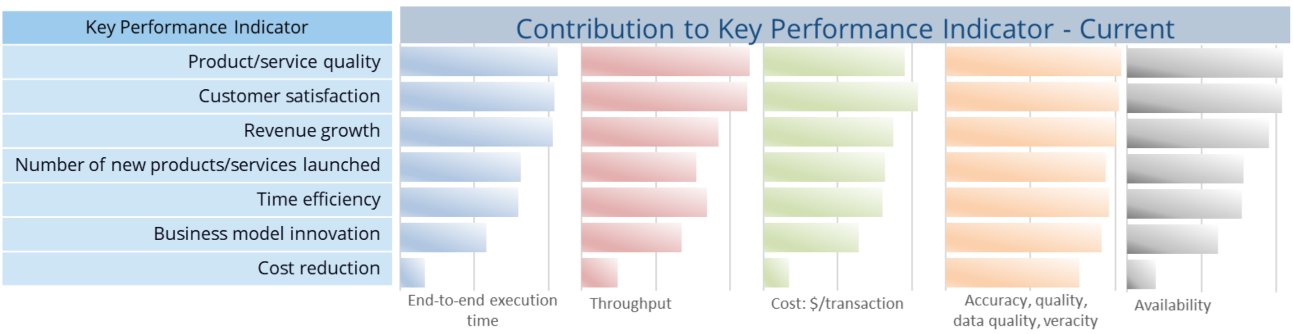

Figure 15 - Contribution to current KPI improvement made by each technical measure

The relative contribution to each KPI from the technical measures is shown in Figure 15, where each KPI is assessed separately. The data is from the survey of 700 respondents, where we asked two specific questions: What is percentage of expected improvement for these specific KPI's, and What are the top technical performance metrics used to measure your BDA environment?

This figure shows the specific improvement in each KPI associated with the technical measure. It is clear from the figure that in most cases Product or Service Quality is the biggest contributor to performance improvement, with the exception of Cost (e.g., $ per transaction), and here, surprisingly, it is customer satisfaction that makes the biggest contribution to improving cost. In most cases – except for Accuracy, Quality, and Veracity – the contribution to the KPI improvement made by cost reduction is notably lower than the other technical measures.

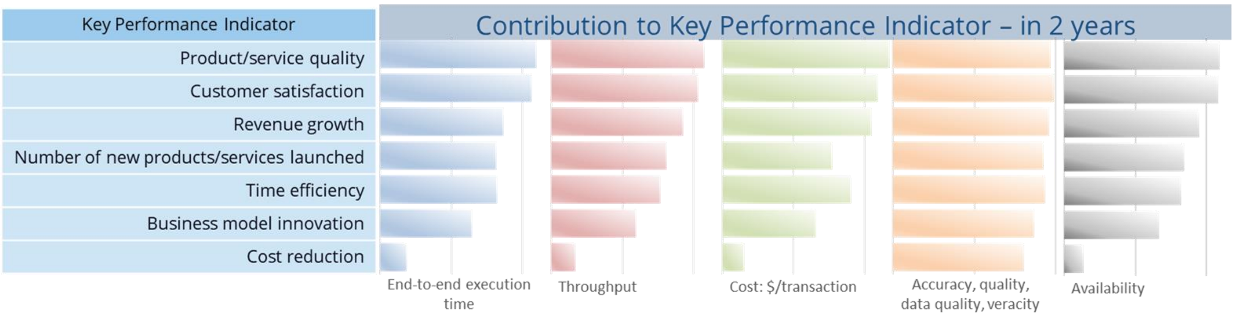

Figure 16 - Contribution to future KPI improvements made by each technical measure

The outlook for future expectations of technical measures' contributions to KPI's is not much different for the leading technical measures, although for the cost KPI customer satisfaction drops slightly its contribution to the KPI, and time efficiency becomes a bigger contributor to KPI success.

These weights give a matrix used to map between the technical measures and the KPI's, and choose appropriate measures and benchmarks for specific use cases.

The ecosystem of KPI classification and, consistently, the outcome of WP2 questionnaire represent also the basis for the activities in WP4. In WP4 we are performing an extensive desk analysis, mapping BDT use cases from the literature based on the DataBench Framework. The complete list of use cases of the extensive desk analysis together with their mapping on the DataBench Framework can be found at the following link; http://78.47.228.66/ecis2019/dimensions_use_cases.htm. The analysis involves industrial use cases and use cases presented by EU ICT 14-15 projects. This extensive data analysis is based on public information with a comprehensive approach to include a broad set of industries and applications of BDTs. The extensive data analysis seems to confirm that the high level of abstraction of the DataBench Framework presented in this deliverable is useful to gather methodological findings from the desk analysis. As an example, as summarized in Figure 17, from a business perspective the desk analysis highlighted customer satisfaction among the top relevant indicators in most industries, with a particular emphasis in industries that provide products/services to consumers, e.g., telco/media, healthcare, banking/insurance/financial services, retail trade/wholesale trade. Conversely, other KPIs appear more strictly related to a specific industry. As an example, *cost reduction* is the most relevant indicator in banking/insurance/financial services and utilities/energy, whereas *revenue growth* is the pivotal KPI in retail trade/wholesale trade, and transport/logistics and healthcare appear to be focused on *product/service quality*. Moreover, some industries are more concerned with *innovation,* e.g., utility/energy and agriculture.

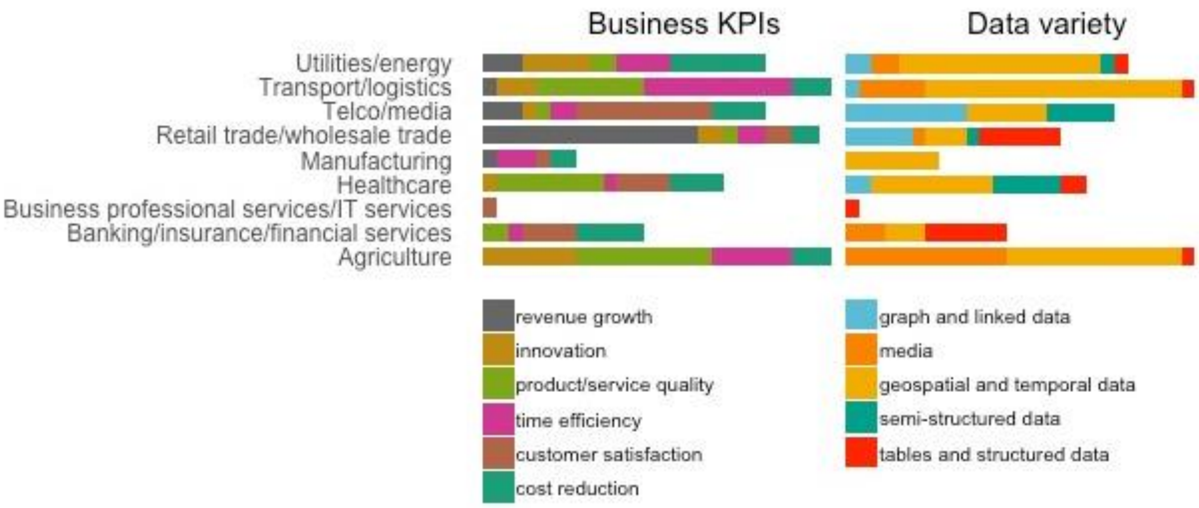

Figure 17 - Quantitative analysis of the desk analysis use cases

From a technical perspective, the desk analysis indicated that *tables and structured data* tend to be present in all industries, although they are predominant in selected industries, such as banking/insurance/financial services. On the contrary, selected industries, including manufacturing, transport/logistics, utilities/energy have specific use cases addressing *geospatial and temporal data* created by IoT devices in monitoring and automation processes. Other types of data, such as *graph and linked data*, are present in all the industries that perform social media analysis.

Overall, from a data analysis perspective it emerged the need to process a growing amount of data by exploiting *predictive/prescriptive* methods with *real-time* constraints, thus making evident the quest for a structured approach able to tackle technical challenges and to support technical choices pivotal to enable business benefits. Moreover, these preliminary findings suggested the relevance of providing blueprints by industry.

4.3 Features selection for profiling by industry sector

Another type of analysis is presented in Figure 18, where we present the profile obtained for the Manufacturing domain, selecting the indicators that have high confidence in the domain, i.e., for which most of the respondents in the sectors indicated an interest. This analysis was performed using the BDVA SG Benchmarking survey results. Respondents were mainly participants in European PPP Big Data projects, for a total of 36 responders, representing 37 different projects.

The questionnaire is synthetically reported in Annex I.

MANUFACTURING [7 respondents]

- [D10] What are your big data benchmarking goals/plans?
  - **Check whether an implementation fulfils given business requirements and specifications.** [86%]
- [D11] Which aspects of Big Data are you benchmarking or planning to benchmark? (ref. BDV Reference Model)
  - **Data Analytics** [100%]
  - **Data Management** [71%]
  - **Data Processing** [71%]
- [D12] What kind of data are you using/planning to use?
  - **Real Data** [86%]
  - **Hybrid** [86%]
- [D15] What type of Data Storage (Storage/Querying/Discovery) are you benchmarking/considering?
  - **Graph Stores** [67%]
  - **NoSQL** [67%]
- [D16] What is the most important type of Data Processing in your platform?
  - **Interactive/(near) Real-time processing** [71%]
- [D17] What types of data problems are you tackling?
  - **Descriptive** [71%]
  - **Predictive** [86%]
- [D18] What types of machine learning approaches do you typically use?
  - **Supervised** [86%]
- [D19] Which modelling techniques do you typically use?
  - **Deep Learning** [67%]
- [D20] What types of data are stored and processed in your system/platform? (Ref. BDV Reference Model types)
  - **Time Series including IoT Data** [86%]
- [D21] What are the technical key performance metrics that you (want to) measure in your system/platform/service?
  - **End-to-end execution time (Runtime)** [100%]
  - **Throughput** [67%]
- [D22] Which of the following qualitative features are important for your application/platform?
  - **Fault-Tolerance** [71%]

Figure 18 - Example of profiling KPIs in the Manufacturing domain (elaboration of the initial questionnaire with BDVA SG on benchmarking, Pernici et al., 2018)

4.4 KPI Knowledge Graph

The information collected at various stages of the project will be organised in a form to be easily accessible, structured and interoperable with other semantic knowledge resources. For that purpose we plan to use a recently popular data structure called 'Knowledge Graph' (KG) (https://en.wikipedia.org/wiki/Knowledge_Graph) allowing flexible data schemas and be scalable for operations like search, aggregation, and in particular interlinked with other relevant global semantic vocabularies and resources like WikiData (https://en.wikipedia.org/wiki/Wikidata) and LinkedOpenData (http://linkeddata.org/).

For the purpose of referring to the specific Knowledge Graph built in the DataBench project, we will call it with the working name as 'DataBenchKG'. In the following paragraphs we are describing constituents of DataBenchKG, its planned implementation and required characteristics.

The envision the information that is going to be considered in the project to be coming from the following sources (but not limited to, in the case of necessity to expand):

- Questionnaires – structured question-answer pairs
- Interviews – structured questions and unstructured answer textual descriptions
- Data science algorithms descriptions – structured descriptions of algorithms used in data science; descriptions will be aligned with an ontology of machine learning and broader data science related algorithms (as a starting point we plan to use W3C Machine Learning Schema https://www.w3.org/community/ml-schema/)
- Data science tools descriptions – structured descriptions of tools used in data science; since such an ontology doesn't exist, we plan to develop 'minimal viable ontology' satisfying the project needs
- Dataset descriptions – structured descriptions of characteristics of datasets which are commonly used in data science in broader in the area of data analytics; there are several approach how to structure the domain of data characteristics and during the course of the project we plan to construct a viable solution for such a schema satisfying the needs of the project; a major objective will be to automate the process of extracting such characteristics from datasets
- Benchmarking tools description – structured descriptions of tools to perform benchmarking with particular focus on the DataBench platform, but being also able to describe benchmarking tools from the similar initiatives (including related H2020 projects)
- Benchmarking experiments outcomes – each benchmarking experiment will measure several KPIs (like time, memory, quality of results, business), which will be recorded and stored in a structured way
- Benchmarking experiments machine learning models – aggregate models built from 'Benchmarking experiments outcomes' data by machine learning algorithms; the purpose of models is to derive analytical understanding on how data science algorithms and tools perform under different datasets and parametrizations. The models will be represented: the most likely candidate to represent machine learning models in an interoperable way is 'Predictive Model Markup Language' / PMML (https://en.wikipedia.org/wiki/Predictive_Model_Markup_Language)

The above listed types of information will be stored in a form of a Knowledge Graph, where corresponding 'knowledge fragments' will be aligned with either external ontologies/schemas or ontologies/schemas will be constructed within the project (due to a non-existence of appropriate pre-existing semantic resources). For general conceptual terms we plan to use stable and clean semantic resources from LinkedOpenData and WikiData. For specific technical concepts, where pre-existing semantic resources exists, we will align with the corresponding semantic ontologies/schemas/vocabularies, like W3C Machine Learning Schema.

The data will be stored conceptually in the Knowledge Graph structure, whereas for the implementation of the actual storage will use one of the proven and scalable graph databases such as Neo4J (https://neo4j.com/), ArangoDB (https://www.arangodb.com/) or similar. The final decision, which graph database to be used for DataBenchKG, will be taken at the start of the implementation phase.

An important property, to be satisfied by DataBenchKG, is aggregation and analytics on the top of the collected data. Most of the data sources (listed above) stored in the DataBenchKG are not of a very large scale and with some limited temporal dynamics, and therefore we don't expect major issues with managing and storing the data. For these data sources we expect for the graph database engine to support operations such as search and basic statistics. The most intensive data source will be coming from the 'Benchmarking experiments outcomes' (generated by the tools from WP5), where we expect tens of thousands (or more) experiments to be performed and stored in the graph data engine, with the specific purpose to aggregate and model the data with machine learning algorithms. For the purpose to be scalable and easily accessible, we might use for this dataset an alternative data storage engine, likely a NoSQL database MongoDB or relational database PostgreSQL. More detailed description of the data intensive part of DataBenchKG is described in D5.1.

5. Concluding remarks

The present report is based on the results of DataBench during the first year of the project, and it collects and harmonizes the indicators that emerged from several points of view in the analysis of the market and case studies and from a classification of benchmarking tools, developed in the following activities:

- WP2 Economic, Market and Business Analysis, and in particular the design of the survey developed in the Work Package and the analysis of the results.
- WP3 DataBench Toolbox, and the Definition of the DataBench Toolbox architecture in Task 3.1.
- WP4 Evaluating Business Performance with DataBench Toolbox and the ongoing data collection in Task 4.1.
- WP5 Technical Evaluation using the DataBench Toolbox, and the initial evaluation of DataBench metrics.

The resulting set of indicators, classified in the following four features: Business features, Big Data Application features, Platform and Architecture features, Benchmark-specific features. Such an ecosystem of indicators is going to be validated in the next months both in the first release of the Toolbox, and in the further data collection, data analysis, and validation activities. The first level indicators described in this report will also be further refined in more specific classes and the relations among them will be studied in detail.

References

DataBench Deliverable D2.1, Economic, Market and Business Analysis Methodology, April 2018

DataBench Deliverable D2.2, Preliminary Benchmarks of European and industrial significance, December 2018 (in preparation)

DataBench Deliverable D3.1, DataBench Architecture, July 2018

DataBench Deliverable D4.1, Data Collection Plan, August 2018

DataBench Deliverable D5.1, Initial Evaluation of DataBench Metrics, December 2018 (in preparation)

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

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

Barbara Pernici, Chiara Francalanci, Angela Geronazzo, Lucia Polidori, Stefano Ray, Leonardo Riva, Arne Jørgen Berre and Todor Ivanov, "Big Data key performance indicators", XV edition of the itAIS conference, Pavia, Italy, Oct. 2018

Rui Han, Lizy Kurian John, Jianfeng Zhan: Benchmarking Big Data Systems: A Review. IEEE Trans. Services Computing 11(3): 580-597, 2018

Annex I – BDVA Questionnaire SG Benchmarking (Spring 2018)

Benchmarking Big Data Benchmarks

By answering this questionnaire, you will help gathering evidence on the use of Big Data technologies and benchmarks. With this survey, we aim to assess how companies could benefit from Big Data benchmarking. The results will be used to build a bridge between technical and business benchmarking. All results will be shared with registered respondents.

What is your current role/position?

- Data Engineer
- Software/Application Developer
- DevOps (development and operations)
- System Administrator
- System Architect
- Data Analyst
- Data Scientist
- Other:

Are you participating in EU research projects? If yes, which ones?

Your answer:

Are you affiliated with an organization? If yes, which one?

Your answer:

Which societal challenges do you target?

- SC1: Health, Demographic Change and Wellbeing
- SC2: Food Security, and the Bioeconomy
- SC3: Secure, Clean and Efficient Energy
- SC4: Smart, Green and Integrated Transport
- SC5: Climate Action, Environment, Resource Efficiency and Raw Materials
- SC6: Inclusive, Innovative and Reflective Societies
- SC7: Secure Societies
- None
- Other:

What are your Big Data application domains?

- Energy
- Financial Services
- Manufacturing
- Construction
- Food Agriculture

- Retail, Wholesale
- Professional Services
- Transport Services
- Public Administration
- Healthcare
- Education
- Telecom, IT, Media
- Utilities
- Other:

Do you use business indicators to measure the performance of your big data & analytics initiatives?

- We do not use them
- We target revenue growth
- We target margin growth
- We target cost reduction
- We target time efficiency
- We target customer satisfaction
- We target product/service quality
- Other:

Are your big data & analytics in real-time and integrated with business processes?

- Yes
- No
- Not yet, but will be in the near future
- I don't know

In which role do you perform benchmarking?

- Technology provider, vendor or system integrator
- Academic researcher
- End user
- None
- Other:

Are you currently evaluating software using benchmarking technologies?

- HOBBIT Benchmarking Platform
- HiBench
- SparkBench
- BigBench / TPCx-BB
- Yahoo! Cloud Serving Benchmark (YCSB) / TPCx-IoT
- Kaggle
- GERBIL
- No

- Other:

What are your big data benchmarking goals/plans?

- Comparing different architectures (e.g., Lambda vs. Data Lakes)
- Comparing different software technologies and stacks (e.g., MapReduce, Spark, Flink)
- Comparing different implementations of a functionality (e.g., Spark Scala, Java, R, PySpark)
- Check whether an implementation fulfills given business requirements and specifications
- Other:

Which aspects of Big Data are you benchmarking or planning to benchmark? (ref. BDV Reference Model)

- Data Storage (Storage/Querying/Discovery – SQL, NoSQL, Column, Key-value, Raster ...)
- Data Management (Extraction, Annotation, Enrichment, Curation, Link/Integration/Federation)
- Data Protection
- Data Processing (Batch, Stream, Interactive/(near) Real-time and Iterative/In-memory processing)
- Data Analytics (Descriptive, Diagnostic, Predictive, Prescriptive) (MachineLearning: Supervised, Un-supervised, Reinforcement learning), Deep Learning
- Data Visualization
- Complete domain application/system/solution ….
- Other:

What kind of data are you using/planning to use?

- Synthetic data
- Real data
- Hybrid (mix of real and synthetic) data
- Other:

Which dataset sizes do you target in your application(s)?

- In Megabytes
- In Gigabytes
- In Terabytes
- In Petabytes
- Other:

Are you willing to be a member of our benchmarking community? Goodies include the results of this survey. If yes, please add your email address below.

What type of Data Storage (Storage/Querying/Discovery) are you benchmarking/considering?

- Relational Database Management Systems
- SQL
- NoSQL
- Column Stores
- Key-Value Stores
- Graph Stores
- In-memory Stores
- Other:

What is the most important type of Data Processing in your platform?

- Batch processing
- Stream processing
- Interactive/(near) Real-time processing
- Iterative/In-memory processing
- Other:

What types of data problems are you tackling?

- Descriptive
- Inferential
- Predictive
- Prescriptive
- Other:

What types of machine learning approaches do you typically use?

- Unsupervised
- Semi-supervised
- Supervised
- Active

Which modelling techniques do you typically use?

- Deep Learning
- Kernel Methods
- Tree-based Methods
- Latent Factor Models
- Clustering
- Other:

What types of data are stored and processed in your system/platform? (Ref. BDV Reference Model types)

- Business intelligence – Tables/Schema
- Structured text – Genomics
- Graphs and Linked Data

- Time series incl. IoT data
- Geospatial or temporal
- Text (incl. natural language)
- Media (images, audio or video)
- Other:

What are the technical key performance metrics that you (want to) measure in your system/platform/service?

- End-to-end execution time (Runtime)
- Throughput
- Specific Performance Metrics (i.e. QphH(TPC-H query-per-Hour)@Size(data size), BBQpm(Big Bench Query-per-minute)@SF (Scale Factor)
- Cost ($/QphH@Size, $/BBQpm@SF)
- Energy Consumption (Watts/QphH(TPC-H query-per-Hour)@Size)
- Accuracy (Precision, Recall, F-measure, Mean Reciprocal Rank)
- Availability (in %)
- Other:

Which of the following qualitative features are important for your application/platform?

- Fault-tolerance
- Privacy
- Security
- Governance - Managing the data lifecycle
- Veracity - Defines data accuracy, how truthful it is, any imprecision or uncertainties.
- Variability - Defines the different interpretations that a certain data can have when put in different contexts.
- Data Quality - Quality of data in terms of coverage, time representation, finely measured, etc.
- Correctness
- Other:

What are the key technologies that you are using in your big data infrastructure? For example, Big Data platforms such as Cloudera, HortonWorks, MapR or others offering Hadoop distributions, Spark, Flink, Storm or similar for batch and stream processing, Hive, Spark SQL, Presto or similar for SQL capabilities on top of Hadoop.

Annex II – Features in WP2 survey (October 2018)

Screening Questions
qs1. In which country is your organization located?
qs2. Approximately how many people are currently employed (full-time or part-time) in your organization in your country, including all branches, divisions, and subsidiaries?
qs3. Which of the following best describes your position within your organization?
qs4. What is your role in decisions regarding your organization's use or potential plans for using Big Data and analytics? [...].
qs5. Which of the following industries best describes your organization's primary business? Please make sure you are referring to your company, not your specific role within the organization.
qs6. What is the status of your organization's use of Big Data and analytics technologies and solutions today?
Core Questions – Business Alignment and KPIs
q1. In which of the following areas has your company implemented or does it plan to implement Big Data and analytics initiatives? [Choose all that apply]
q2. Which of the following business goals are driving adoption or consideration of Big Data and analytics in your organization? [Choose all that apply]
q3. How important is the ability to benchmark the business impact of your organization's Big Data and analytics efforts?
q4. How important are the following business Key Performance Indicators (KPIs) for measuring the impact of your organization`s Big Data and analytics efforts? [...]
Main benefits
q5. What level of benefits has your organisation achieved so far (alt: does your organisation expect to achieve) from the use of a Big Data and analytics environment?
q6a. In percentage terms, what is the actual benefit realised (alt: what benefit do you expect to realise) from the use of Big Data and analytics for the following business KPIs? [...]
q6r. Please try to estimate the benefit (alt: expected benefit) realized from the use of Big Data and analytics for the following business KPIs.
q7. To what extent has your organisation`s deployment of Big Data and analytics impacted (alt: will your organisation's deployment ... be impacted by) the ability to attain the following business KPIs?
q8. For the following business KPIs please estimate what percentage of expected improvement will be linked to the adoption of Big Data and analytics in 2020?
q8a. What was your organization's revenue in <COUNTRY> last year, in <CURRENCY>?
Use Cases
q9. If we look at the following specific Big Data and analytics business use cases, what is your organization's position on each of these?
Technical Questions
q10. How would you describe the level of business process integration currently achieved within your Big Data and analytics environment?
q11. Do you believe that supplying capabilities such as real-time integration with business processes will improve Big Data and analytics' impact on your organization and/or community?
q12. To what extent is your Big Data and analytics environment linked or aligned with other technology investments?
q13. In data storage terms, what measurement is typically used to gauge the size of your Big Data and analytics environment(s)?
q14. What type of data storage do you currently use for your Big Data and analytics environment? [Choose all that apply]
q15. What types of data are stored and processed in your Big Data environment? [Choose all that apply]
q16. Which of the following best describes your organization's current approach to the management of data?
q17. To what extent are the following types of data processing paradigms important in your Big Data environment?
q18. What are the top technical performance metrics currently used to measure your Big Data and analytics environment? How about in two years from now - what will you start using? Choose all that apply.
q19. What is the current state of your organization's use of these different analytic techniques?
q20. Looking at Big Data skills requirements, in which areas — if any — do you have difficulty finding enough resources? [Choose all that apply]