Benchmarking for Big Data Applications with the DataBench Framework

Dr. Arne J. Berre, SINTEF

Supported by
Gabriella Cattaneo, IDC,
Barbara Pernici,Politecnico di Milano
Tomas Pariente Lobo, ATOS
Todor Ivanov, Univ. Frankfurt,
Roberto Zicari, Univ. Frankfurt

IEEE Big Data Conference
The Second IEEE International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications (BPOD)
December 10, 2018

DataBench

IDC ANALYZE THE FUTURE

Atos

Institut "Jožef Stefan", Ljubljana, Slovenija

SINTEF

POLITECNICO MILANO 1863

Frankfurt Big Data Lab
GOETHE UNIVERSITÄT

250 pm

January 1st, 2018-December 31st, 2020

www.databench.eu

Outline

- DataBench
- Business Benchmarking
- Technical Benchmarking
- BDVA Reference Model – Digital Platforms
- Overview of Benchmarks

What is Benchmarking?

https://www.shopify.com/encyclopedia/benchmarking

- Benchmarking is a process of measuring the performance of a company's products, services, or processes against those of another business considered to be the best in the industry, aka "best in class."
- The point of benchmarking is to identify internal opportunities for improvement. By studying companies with ***superior performance***, ***breaking down what makes such superior performance possible***, and then ***comparing those processes*** to how your business operates, you can implement changes that will yield significant improvements.

→ **Business Benchmarking**

Technical Benchmarks

- In short, we define that a software benchmark is a ***program used for comparison of software products/tools executing on a pre-configured hardware environment***.

- ***Jim Gray*** (Gray 1992) describes the benchmarking as follows:

”This quantitative comparison starts with the definition of a benchmark or workload. The benchmark is run on several different systems, and the performance and price of each system is measured and recorded. Performance is typically a throughput metric (work/second) and price is typically a five-year cost-of-ownership metric. Together, they give a price/performance ratio.”

Building a bridge between technical and business benchmarking

DataBench

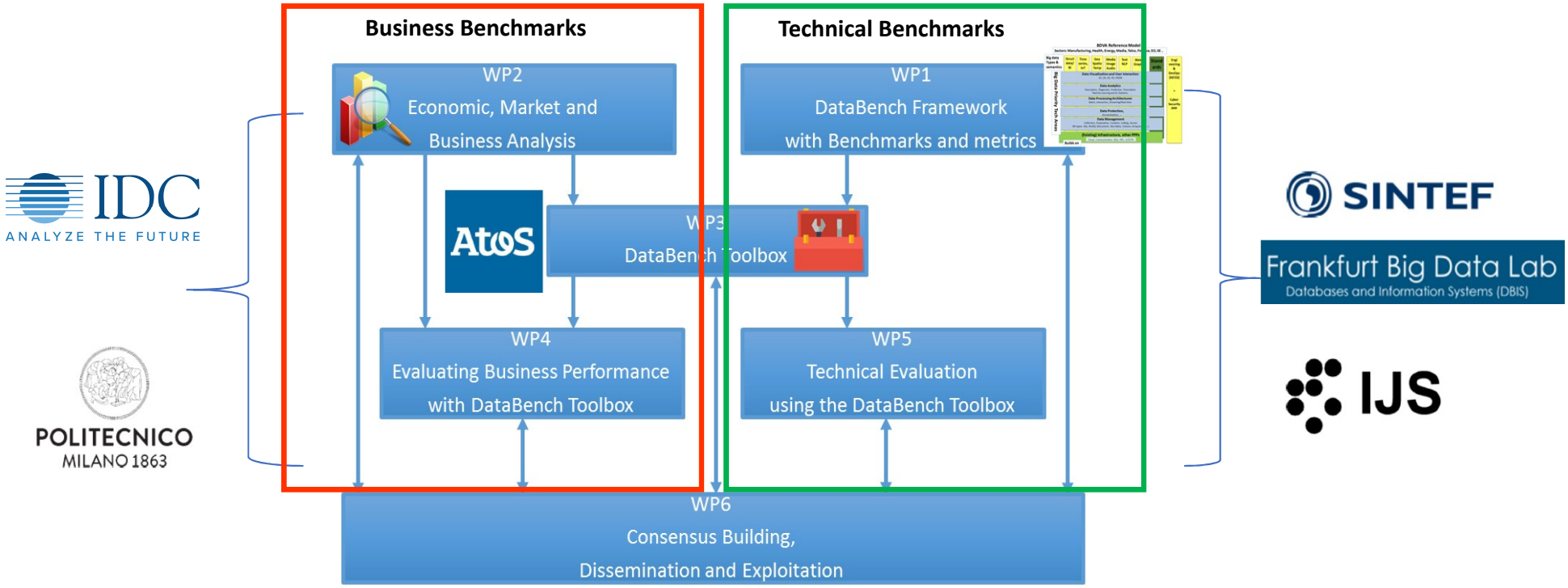
Main Activities

- Classify the main use cases of BDT by industry
- Compile and assess technical benchmarks
- Perform economic and market analysis to assess industrial needs
- Evaluate business performance in selected use cases

Expected Results

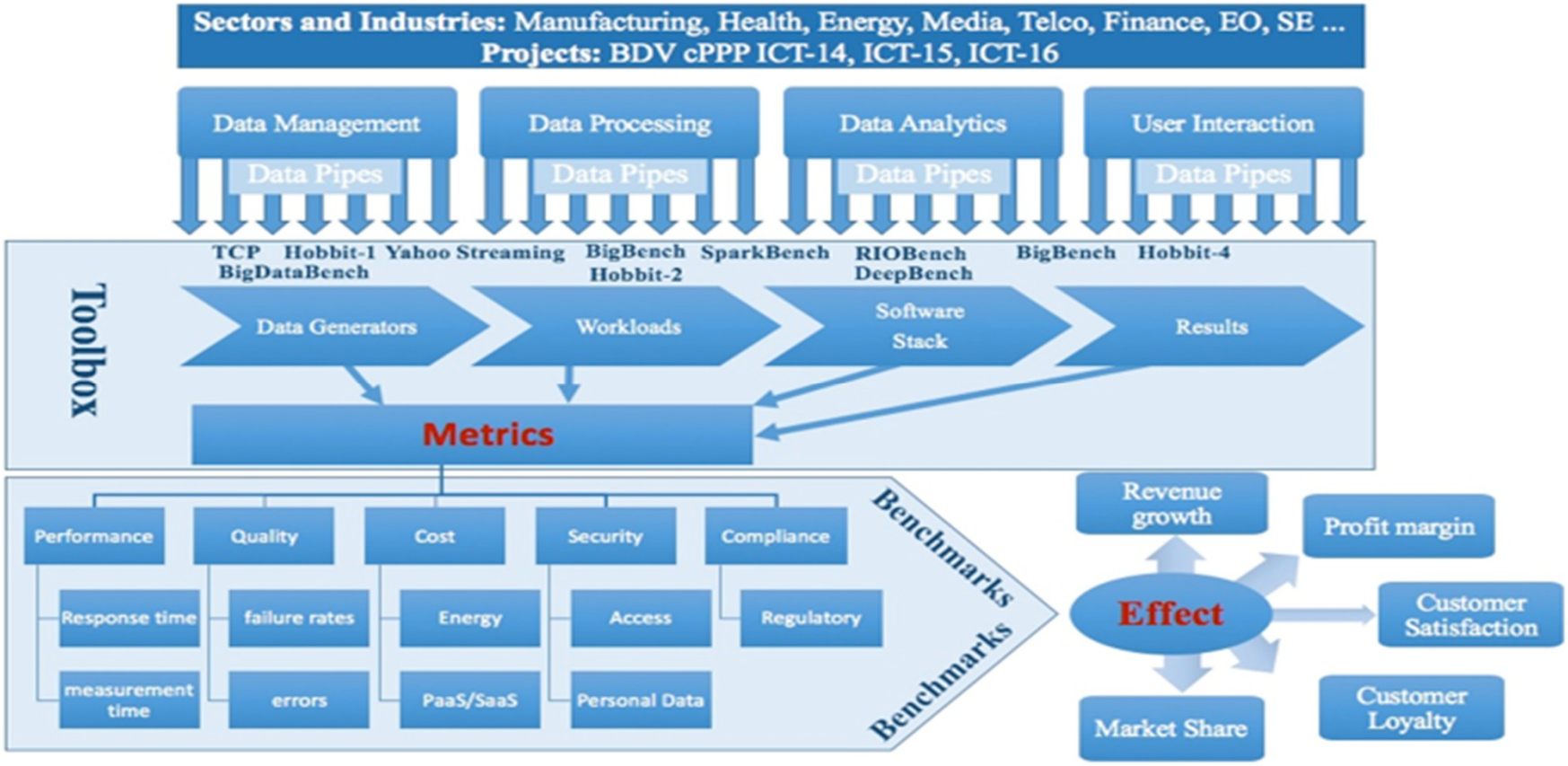
- A conceptual framework linking technical and business benchmarks
- European industrial and performance benchmarks
- A toolbox measuring optimal benchmarking approaches
- A handbook to guide the use of benchmarks

Databench Workflow



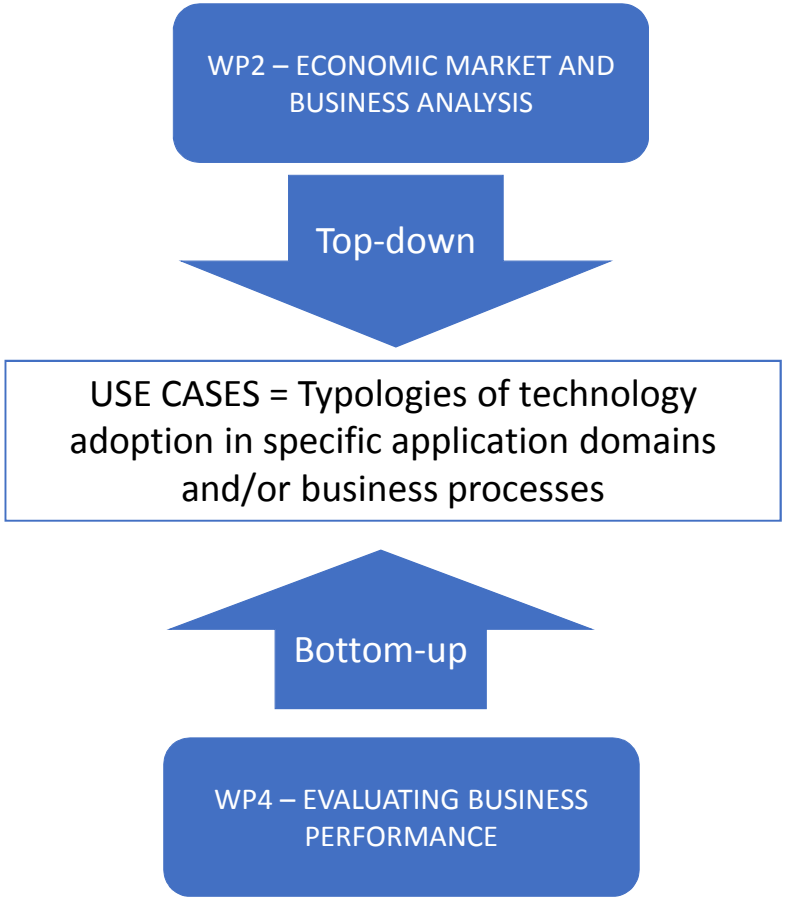
Approach

A way to reuse existing benchmarks & derive technical and business KPIs



How to link technical and business benchmarking

- Focus on economic and industry analysis and the EU Big Data market
- Classify leading Big Data technologies use cases by industry
- Analyse industrial users benchmarking needs and assess their relative importance for EU economy and the main industries
- Demonstrate the scalability, European significance (high potential economic impact) and industrial relevance (responding to primary needs of users) of the benchmarks



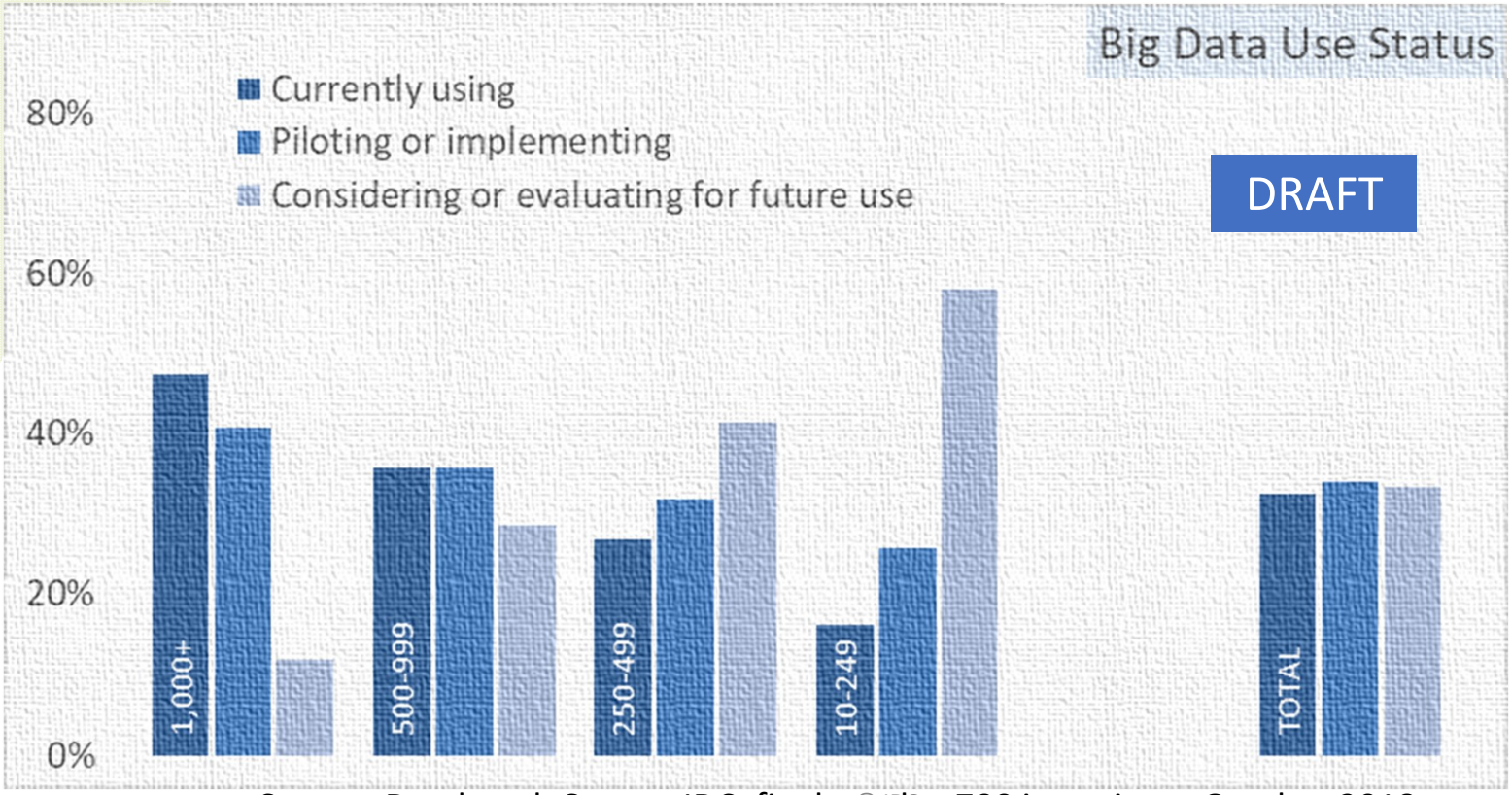
- Focus on data collection and identification of use cases to be monitored and measured
- Evaluation of business performance of specific Big Data initiatives
- Leverage Databench toolbox
- Provide the specific industrial benchmarks to WP"
- Produce the Databench Handbook, a manual supporting the application of the Databench toolbox

Early Results from the Databench Business users Survey

Business Dimensions

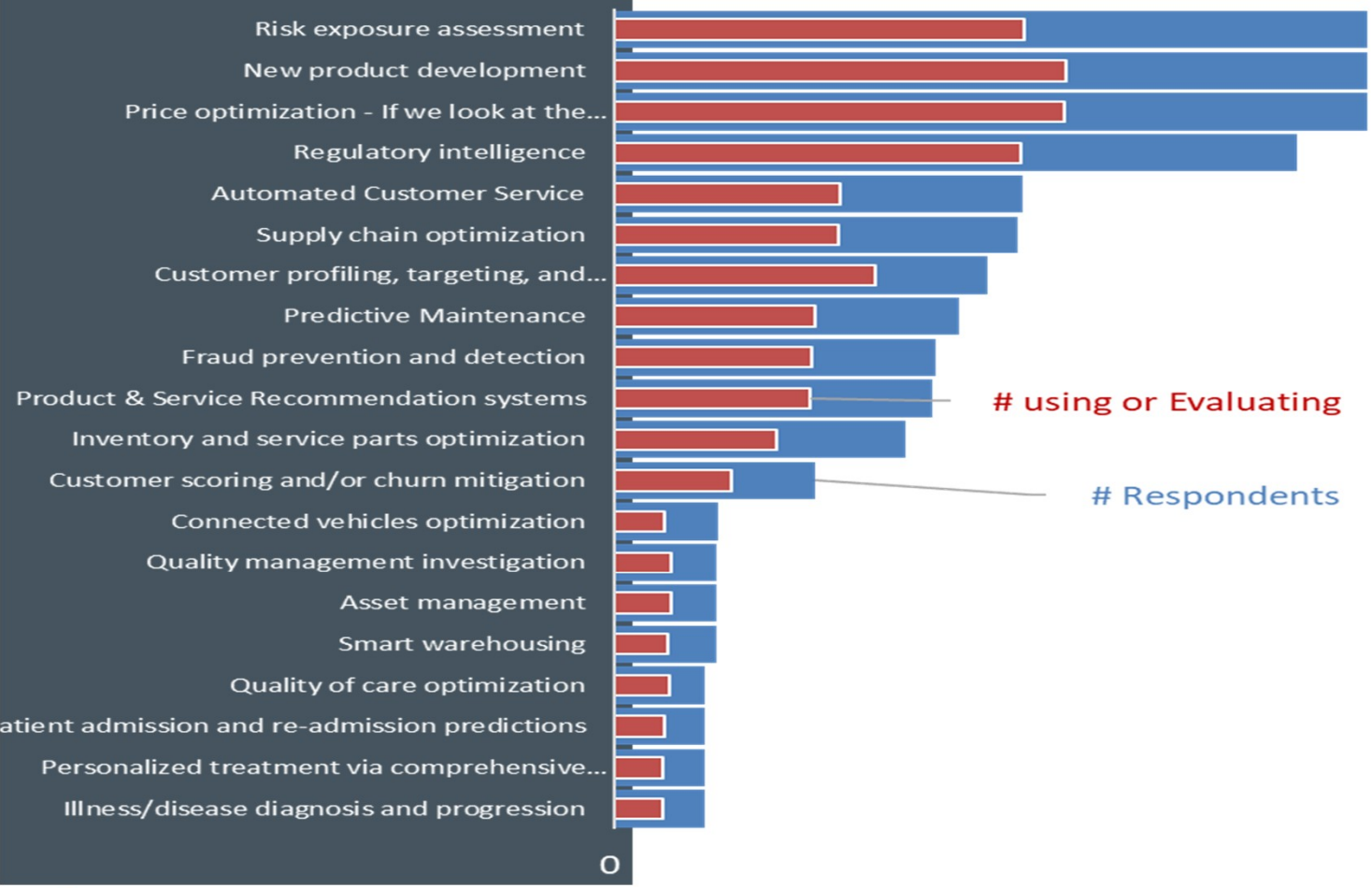
European Companies use of Big Data

By **Size Band**



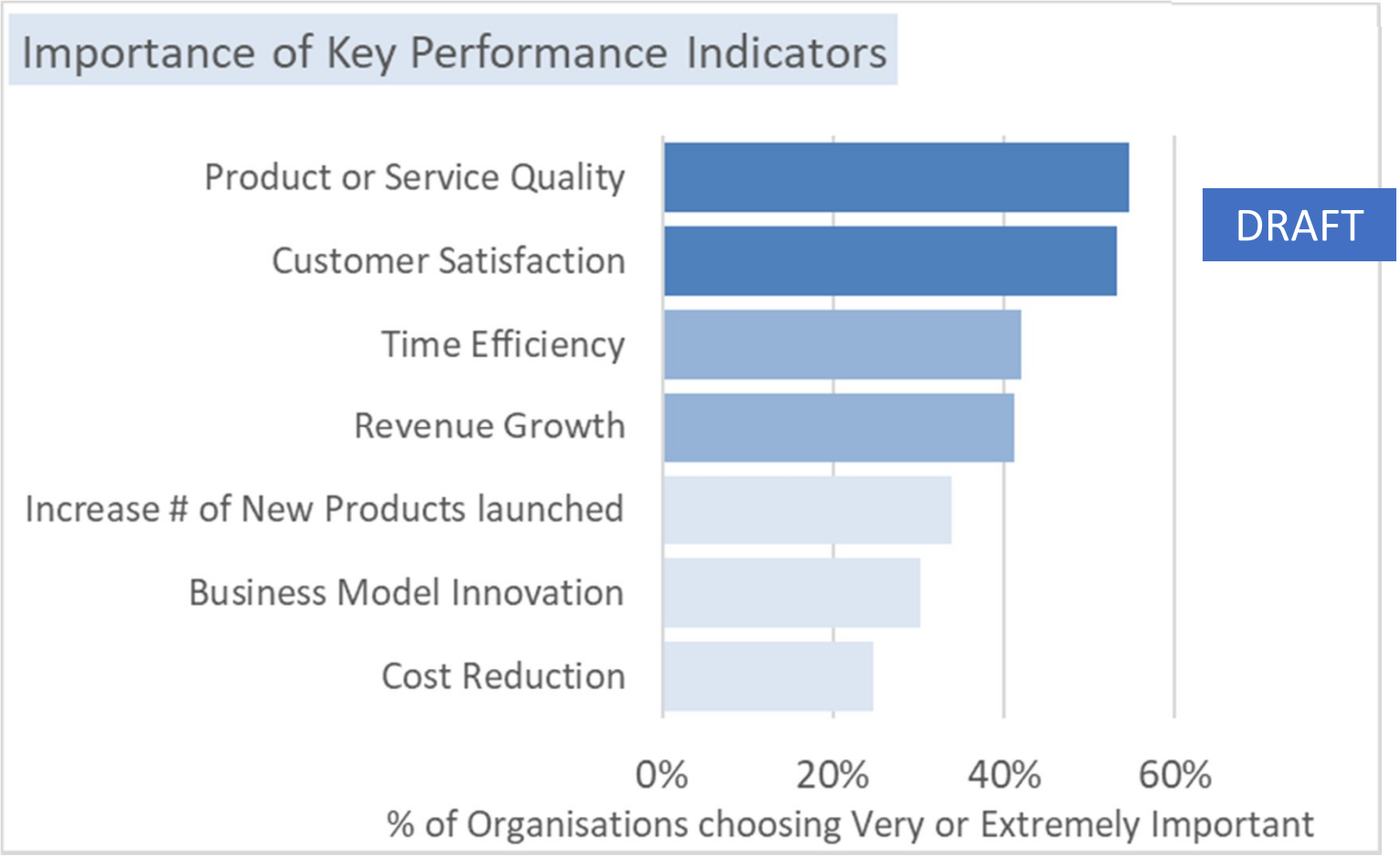
Source: Databench Survey, IDC, final results, 700 interviews, October 2018

Big Data Key Use Cases



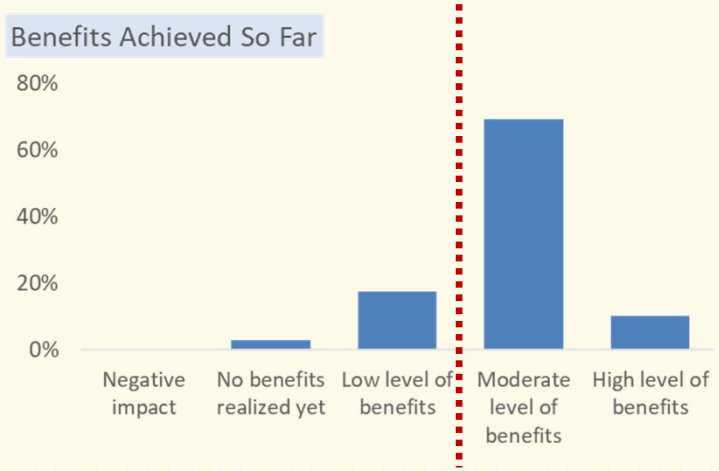
Source: IDC DataBench Survey, October 2018 (n=700 European Companies)

Key Performance Indicators in Users'view



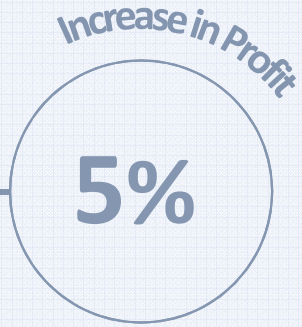
Source: Databench Survey, IDC, final results, 700 interviews, October 2018

Big Data is Worth the Investment

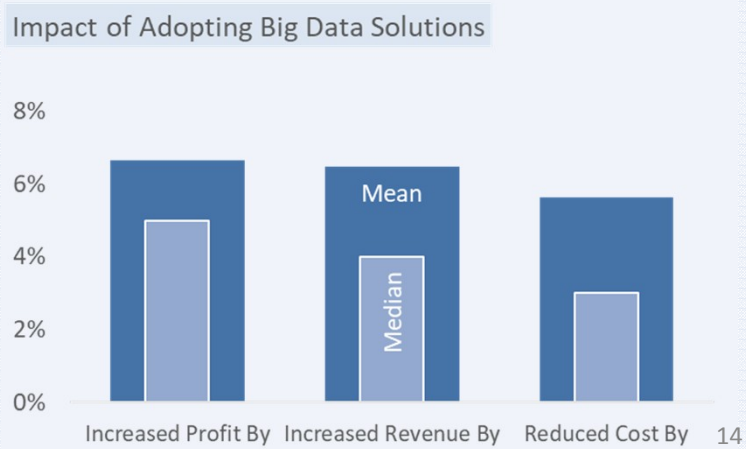


Nearly **90%** of businesses saw **moderate** or **high** levels of benefit in their Big Data implementation

DRAFT



Adopting Big Data Solutions increased **profit** by 5%, **revenue** by 4%, and reduced **cost** by 3%



Source: IDC DataBench Survey, October 2018 (n=700 European Companies)

What can DataBench do for you?

- Provide methodologies and tools to help assess and maximise the business benefits of BDT adoption
- Provide criteria for the selection of the most appropriate BDTs solutions
- Provide benchmarks of European and industrial significance
- Provide a questionnaire tool comparing your choices and your KPIs with your peers

What we want from you?

- Expression of interest to become a case study and monitoring your Big Data KPIs
- Answer a survey on your Big Data experiences

DataBench

Big Data
Technical Benchmarking

IDC
ANALYZE THE FUTURE

Atos

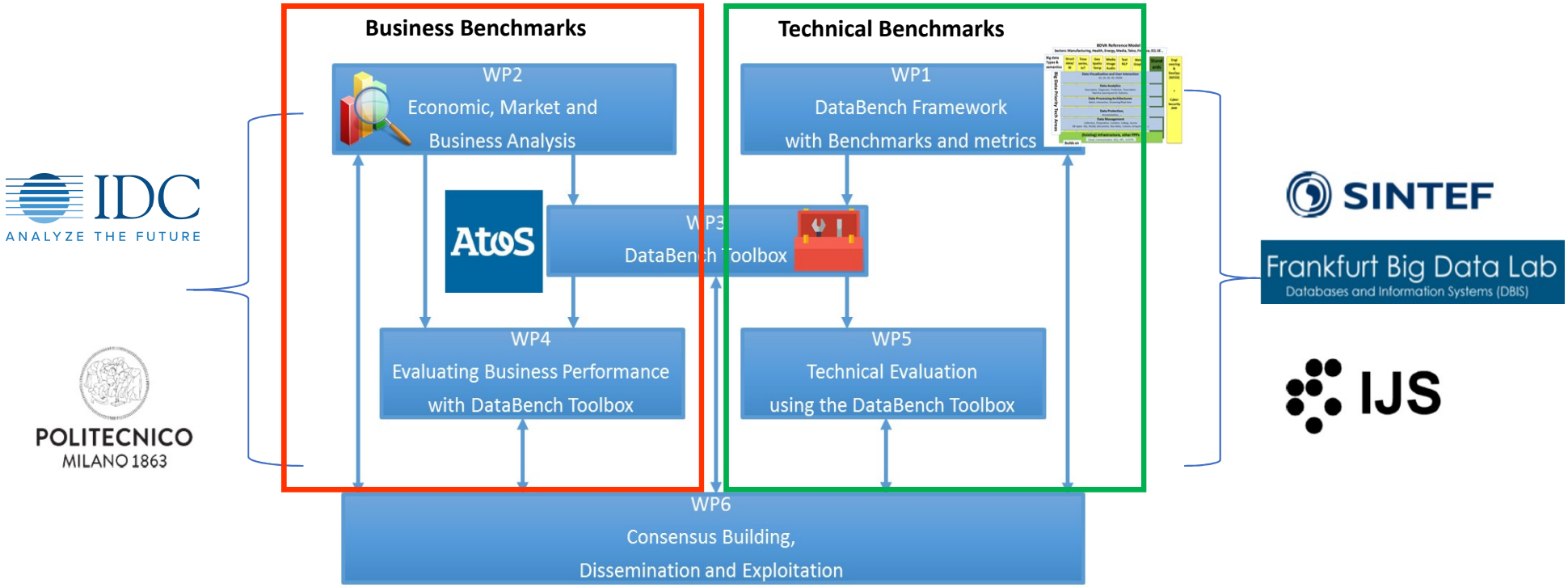
Institut "Jožef Stefan", Ljubljana, Slovenija

SINTEF

POLITECNICO
MILANO 1863

Frankfurt Big Data Lab
GOETHE UNIVERSITÄT

Technical Benchmarks in Databench Workflow



Goals & Objectives

Holistic benchmarking approach for big data

- The DataBench Toolbox will be a **component-based system** of both **vertical** (holistic/business/data type driven) **and horizontal** (technical area based) **big data benchmarks**. **following** the layered architecture provide by **the BDVA reference model**.

Not reinventing the wheel, but use wheels to build a new car

- It should be able to **work** or integrate **with existing benchmarking initiatives** and resources where possible.

Filling gaps

- The Toolbox will investigate **gaps of industrial significance** in the big data benchmarking field and contribute to overcome them.

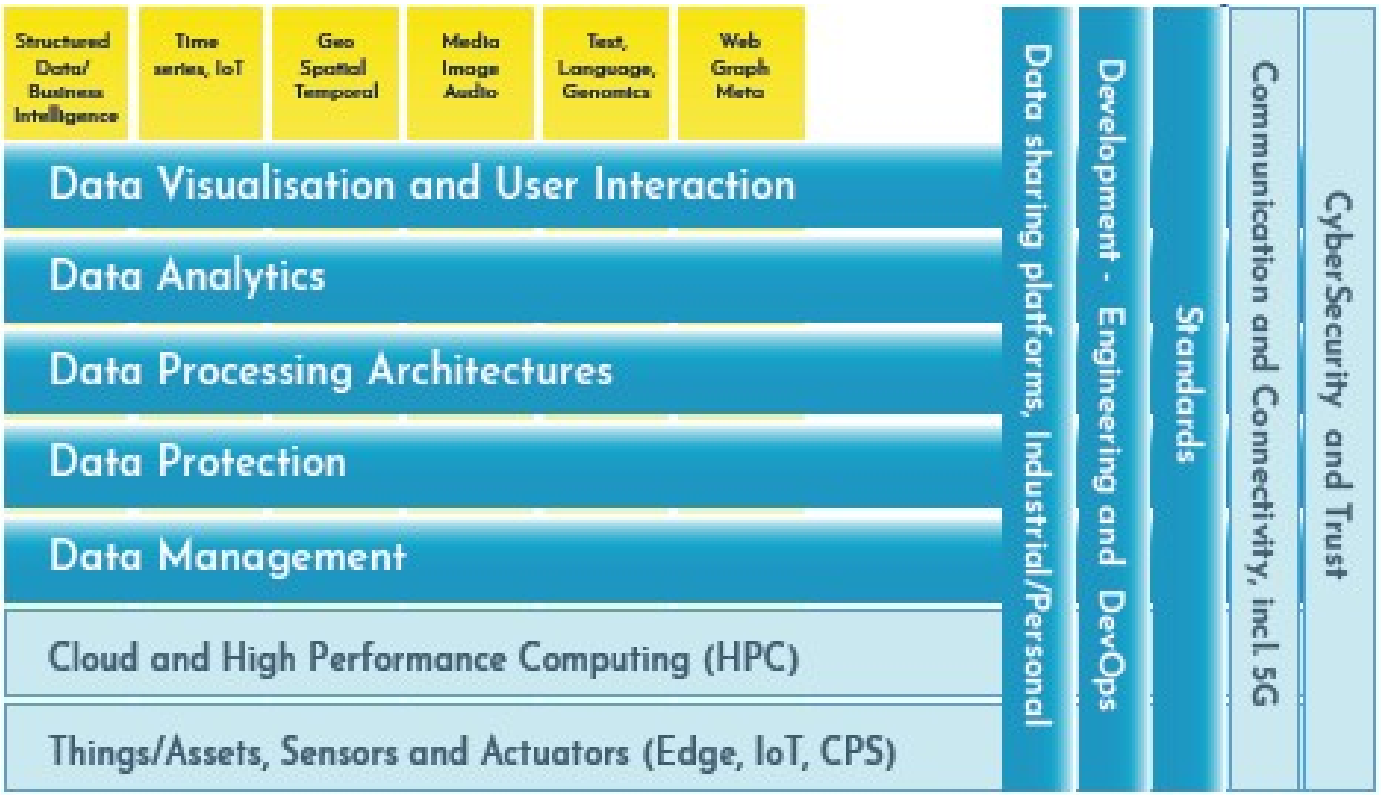
Homogenising metrics

- The Toolbox will implement ways to derive as much as possible **the DataBench technical metrics and business KPIs** from the metrics extracted from the integrated benchmarking.

Web user interface

- It will include a web-based visualization layer to **assist to the final users to specify their benchmarking requirements**, such as selected benchmark, data generators, workloads, metrics and the preferred data, volume and velocity, **as well as searching and monitoring** capabilities.

BDV Reference Model



Identifying and Selecting Benchmarks

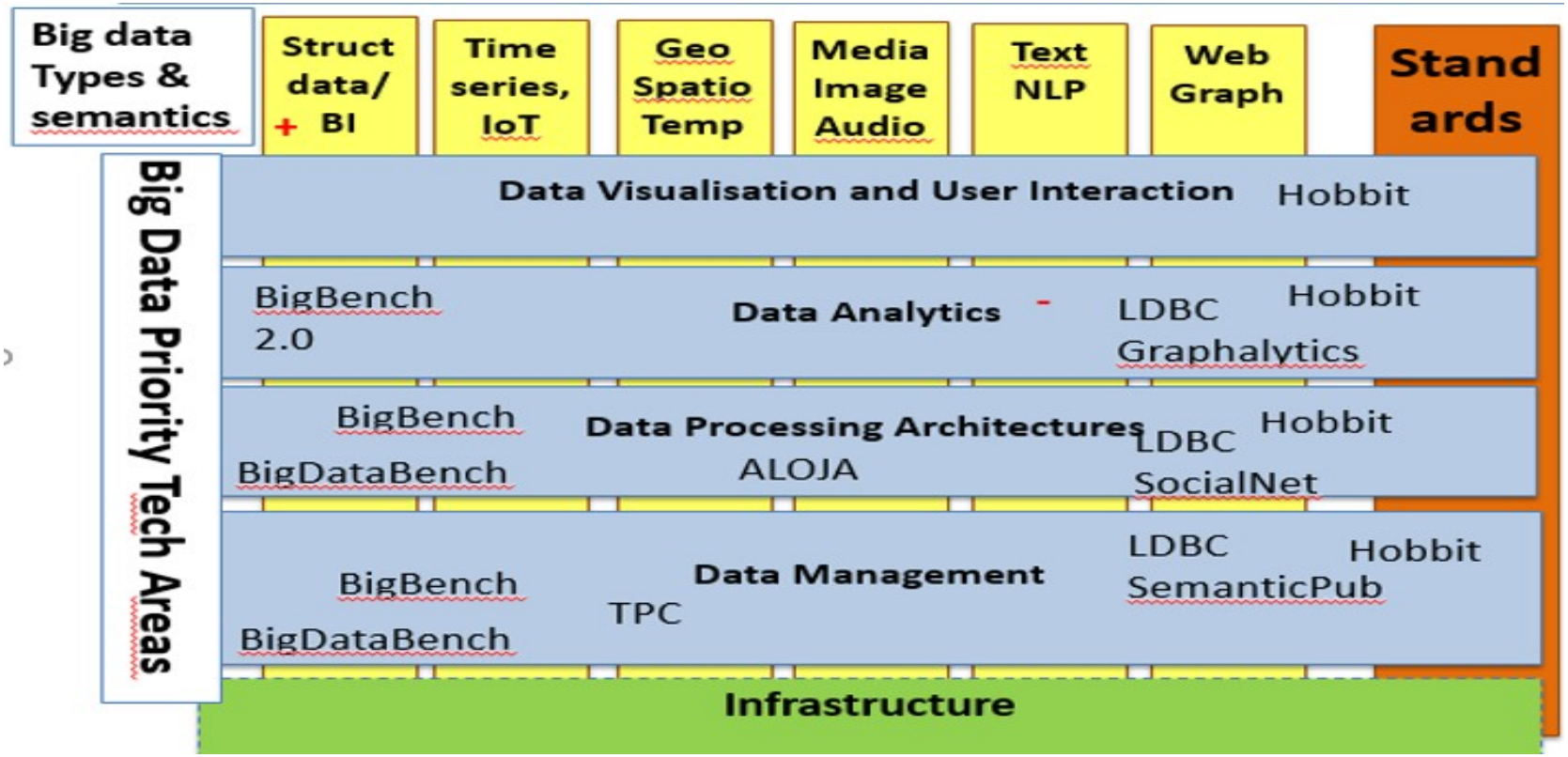
	Benchmarks	TPC-H	TPC-DS v1	Hadoop Workload Examples	GridMix	PigMix	MRBench	CALDA	HiBench	YCSB	SWIM	CloudRank-D	PUMA Benchmark Suite	CloudeSuite	MRBS	AMP Lab Big Data Benchmark	BigBench	BigDataBench	LinkBench	BigFrame	PRIMEBALL	LDBC-Semantic Publishing Benchmark	LDBC - Social Network Benchmark	TPCx-HS	SparkBench	TPCx-V	BigFUN	TPC-DS v2	TPCx-BB	LDBC - Graphalytics	Yahoo Streaming Benchmark (YSB)	DeepBench	DeepMark	StreamBench	RiloTBench	Hobbit Benchmark
		1999	2002	2007		2008		2009	2010		2011	2012				2013						2014			2015							2016			2017	
22	Standards	X	X											X								X	X	X		X		X	X	X						
	MetaData																					X														
	Graph, Network								X									X	X	X		X	X		X					X		X	X			X
	Text, NLP, Web			X					X		X		X	X	X		X	X		X	X	X	X	X	X		X		X		X	X	X	X	X	X
	Image, Audio																	X			X											X	X			
	Spatio Temp																	X																		X
	Time Series, IoT													X	X			X			X										X				X	X
	Structured, BI	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X			X	X	X
18	Visual Analytics																																			X
17	Industrial Analytics (Descriptive, Diagnostic, Predictive, Prescriptive)														X			X				X													X	
16	Machine Learning, AI, Data Science								X			X			X		X	X				X			X				X			X	X		X	X
	Streaming/ Realtime Processing								X			X						X							X						X			X	X	
	Interactive Processing	X	X							X					X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X
	Batch Processing	X	X	X	X	X	X	X	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X			X
	Data Privacy/Security																																			
15	Data Governance/Mgmt																	X																		
14	Data Storage	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X				X	X	X
19	Communication & Connectivity			X																				X												
9	Cloud Services & HPC, Edge											X		X		X		X			X					X										

Updating with new Benchmarks

BDVA Reference Model

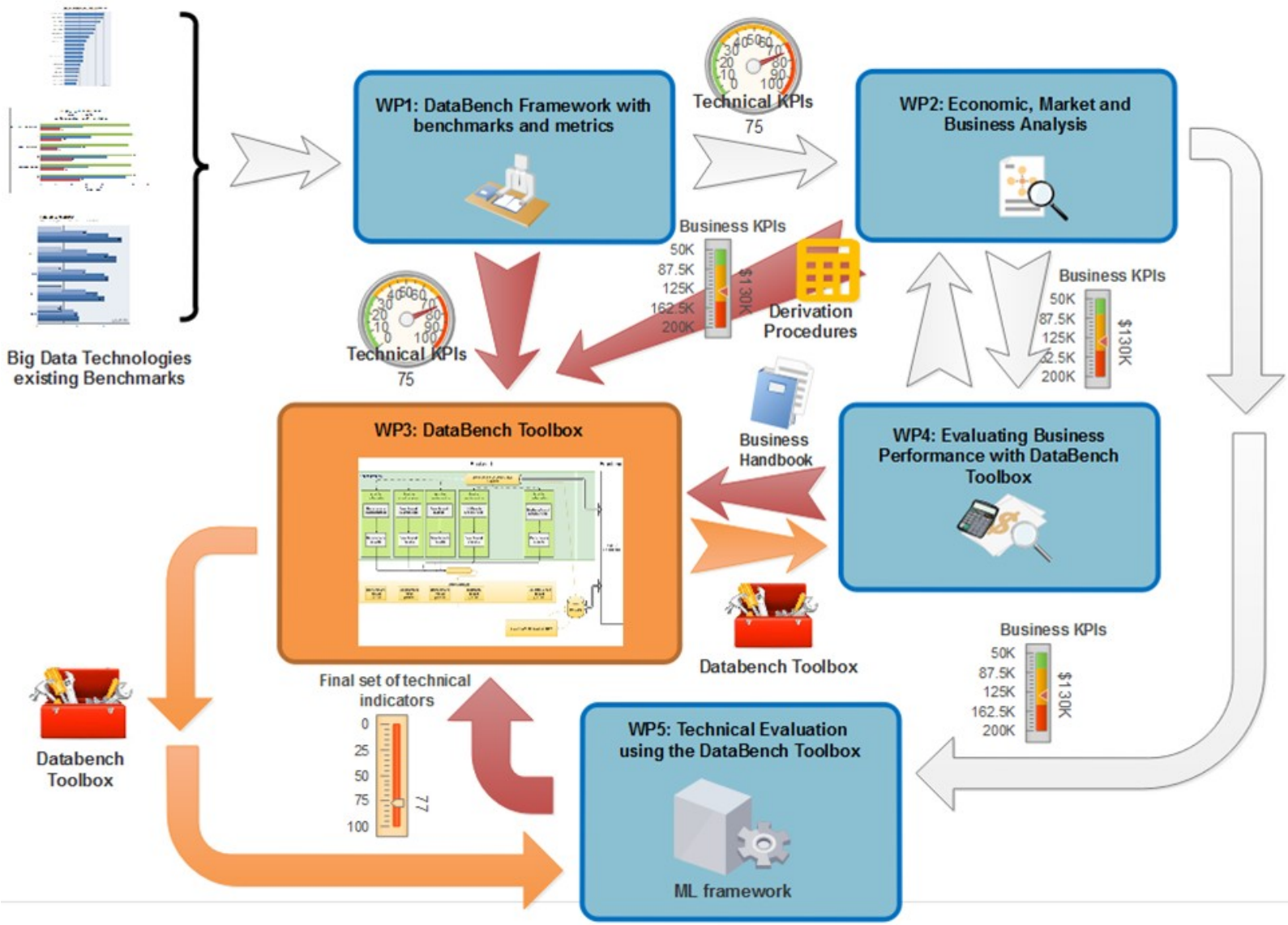
Verticals, incl. Data types	23	Domain/Sector/Business solutions KPIs (Manufact, Transport, Energy,..																																																	
		Business																																																	
		Transport																																																	
		Manufacturing																																																	
		Energy																																																	
		.. Domain X …																																																	
	22	Standards	x	x												x								x	x		x		x		x	x	x																		
		MetaData																						x																											
		Graph, Network									x									x	x	x		x	x			x					x			x	x					x									
		Text, NLP, Web				x					x		x		x	x	x		x	x		x	x	x	x	x	x	x		x			x			x	x				x	x									
		Image, Audio																		x			x													x	x														
		Spatio Temp																		x																					x									
		Time Series, IoT														x	x			x			x											x							x	x									
		Structured, BI	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x							x	x									
Analytics, Processing, Data Management, Infra	18	Visual Analytics																																								x									
	17	Industrial Analytics (Descriptive, Diagnostic, Predictive, Prescriptive)															x			x				x																	x									
	16	Machine Learning, AI, Data Science									x			x			x		x	x				x				x				x			x	x				x	x									
		Streaming/ Realtime Processing									x			x						x						x		x						x						x										
		Interactive Processing	x	x								x					x	x	x	x	x	x	x	x	x			x	x	x	x	x	x	x	x	x				x	x									
		Batch Processing	x	x		x	x	x	x	x	x		x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x		x	x					x									
		Data Privacy/Security																																																	
	15	Data Governance/Mgmt																		x																															
	14	Data Storage	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x							x	x									x
	19	Communication & Connectivity				x																					x																								
	9	Cloud Services & HPC, Edge												x		x		x		x			x						x																					
		Benchmarks	TPC-H	TPC-DS v1	Linear Road	Hadoop Workload Examples	GridMix	PigMix	MRBench	CALDA	HiBench	YCSB	SWIM	CloudRank-D	PUMA Benchmark Suite	CloudeSuite	MRBS	AMP Lab Big Data Benchmark	BigBench	BigDataBench	LinkBench	BigFrame	PRIMEBALL	Semantic Publishing Benchmark (SPB)	Social Network Benchmark	StreamBench	TPCx-HS	SparkBench	TPCx-V	BigFUN	TPC-DS v2	TPCx-BB	Graphalytics	Yahoo Streaming Benchmark (YSB)	DeepBench	DeepMark	TensorFlow Benchmarks	Fathom	AdBench	RIoTBench	Hobbit Benchmark	TPCx-HS v2	BigBench V2	Sanzu	Penn machine learning benchmark (PMLB)	OpenML benchmark suites	Senska	DAWNBench /MLPerf	IDEBench	ABench
			1999	2002	2004	2007		2008		2009	2010		2011	2012				2013						2014				2015						2016								2017							2018

Identifying and Selecting Benchmarks



Dimensions of Technical Benchmarks

Metrics	Data Types	Benchmark Data Usage	Storage Type	Processing Type	Analytics Type	Architecture Patterns	Platform Features
Execution time/ Latency	Business Intelligence (Tables, Schema...)	Synthetic data	Distributed File System	Batch	Descriptive	Data Preparation	Fault-tolerance
Throughput	Graphs, Linked Data	Real data	Databases/ RDBMS	Stream	Diagnostic	Data Pipeline	Privacy
Cost	Time Series, IoT	Hybrid (mix of real and synthetic) data	NoSQL	Interactive/(near) Real-time	Predictive	Data Lake	Security
Energy consumption	Geospatial, Temporal		NewSQL/ In-Memory	Iterative/In-memory	Prescriptive	Data Warehouse	Governance
Accuracy	Text (incl. Natural Language text)		Time Series			Lambda Architecture	Data Quality
Precision	Media (Images, Audio and Video)					Kappa Architecture	Veracity
Availability						Unified Batch and Stream architecture	Variability
Durability							Data Management
CPU and Memory Utilization							Data Visualization


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

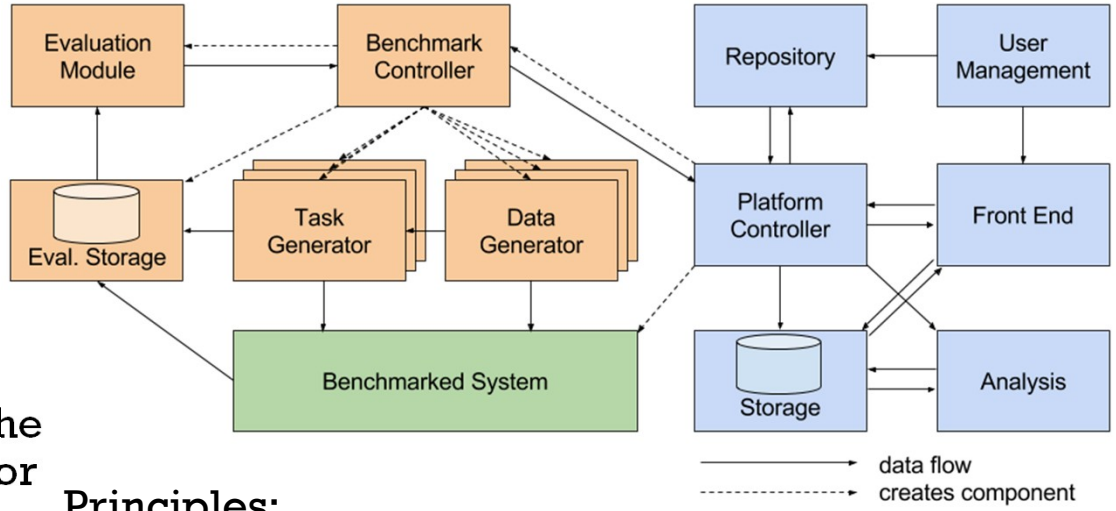
Benchmark Organizations

THE HOBBIT PLATFORM

- Benchmark any step of the Linked Data lifecycle
- Ensure that benchmarking results can be found, accessed, integrated and reused easily (FAIR principles)
- Benchmark Big Data platforms by being the first distributed benchmarking platform for Linked data.

- The Hobbit platform comprises several components:
  - Single components are implemented as independent containers.
  - Communication between these components is done via a message bus.

- Everything is dockerized, from the benchmarked system to all the components



Principles:

- Users can test systems with the HOBBIT benchmarks without having to worry about finding standardized hardware
- New benchmarks can be easily created and added to the platform by third parties.
- The evaluation can be scaled out to large datasets and on distributed architectures.
- The publishing and analysis of the results of different systems can be carried out in a uniform manner across the different benchmarks.

TPC (Transaction Processing Performance Council)

- The TPC (Transaction Processing Performance Council) is a non-profit corporation operating as an industry consortium of vendors that define transaction processing, database and big data system benchmarks.
- TPC was formed on August 10, 1988 by eight companies convinced by Omri Serlin. In November 1989 was published the first standard benchmark TPC-A with 42-pages specification (Gray (1992)). By late 1990, there were 35 member companies.

Active TPC Benchmarks as of 2018:

Benchmark Domain	Specification Name
Transaction Processing (OLTP)	TPC-C, TPC-E
Decision Support (OLAP)	TPC-H, TPC-DS, TPC-DI
Virtualization	TPC-VMS, TPCx-V, TPCx-HCI
Big Data	TPCx-HS V1, TPCx-HS V2, TPCx-BB, TPC-DS V2
IoT	TPCx-IoT
Common Specifications	TPC-Pricing, TPC-Energy

SPEC (Standard Performance Evaluation Corporation)

- The SPEC (Standard Performance Evaluation Corporation) is a non-profit corporation formed to establish, maintain and endorse standardized benchmarks and tools to evaluate performance and energy efficiency for the newest generation of computing systems.
- It was founded in 1988 by a small number of workstation vendors. The SPEC organization is umbrella organization that covers four groups (each with their own benchmark suites, rules and dues structure): the Open Systems Group (OSG), the High-Performance Group (HPG), the Graphics and Workstation Performance Group (GWPG) and the SPEC Research Group (RG).

Active SPEC Benchmarks as of 2018:

Benchmark Domain	Specification Name
Cloud	SPEC Cloud IaaS 2016
CPU	SPEC CPU2006, SPEC CPU2017
Graphics and Workstation Performance	SPECapc for SolidWorks 2015, SPECapc for Siemens NX 9.0 and 10.0, SPECapc for PTC Creo 3.0, SPECapc for 3ds Max 2015, SPECwpc V2.1, SPECviewperf 12.1
High Performance Computing, OpenMP, MPI, OpenACC, OpenCL	SPEC OMP2012, SPEC MPI2007, SPEC ACCEL
Java Client/Server	SPECjvm2008, SPECjms2007, SPECjEnterprise2010, SPECjbb2015
Storage	SPEC SFS2014
Power	SPECpower ssj2008
Virtualization	SPEC VIRT SC 2013

STAC Benchmark Council

- The STAC Benchmark Council consists of over 300 financial institutions and more than 50 vendor organizations whose purpose is to explore technical challenges and solutions in financial services and to develop technology benchmark standards that are useful to financial organizations.
- Since 2007, the council is working on benchmarks targeting Fast Data, Big Data and Big Compute workloads in the finance industry.

Active STAC Benchmarks as 2018:

Benchmark Domain	Specification Name
Feed handlers	STAC-M1
Data distribution	STAC-M2
Tick analytics	STAC-M3
Event processing	STAC-A1
Risk computation	STAC-A2
Backtesting	STAC-A3
Trade execution	STAC-E
Tick-to-trade	STAC-T1
Time sync	STAC-TS
Big Data	in-development
Network I/O	STAC-N1, STAC-T0

Relevant Benchmark Platforms & Tools

- Hobbit - https://project-hobbit.eu/
- ALOJA - https://aloja.bsc.es/
- OLTPBench - https://github.com/oltpbenchmark/oltpbench
- PEEL - http://peel-framework.org/
- PAT - https://github.com/intel-hadoop/PAT

Funded by the H2020 Framework Programme of the European Union

Evidence Based Big Data Benchmarking to Improve Business Performance

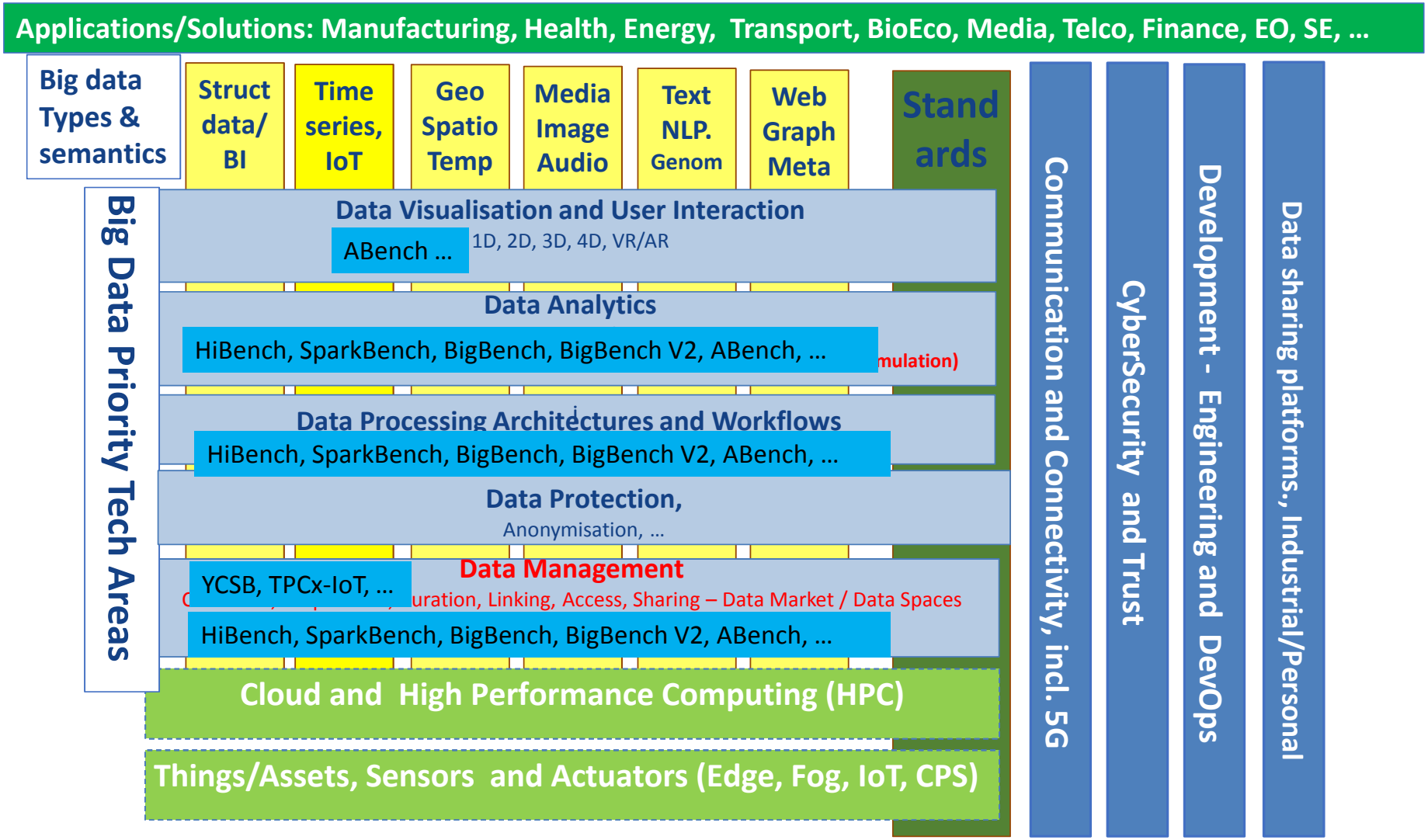
DataBench

The DataBench ToolBox Benchmarks

Stress-testing the Big Data Technology Ecosystem

Todor Ivanov (Frankfurt Big Data Lab)

Big Data Value Reference Model



Summary

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

Some of the benchmarks to integrate (I)

Micro-benchmarks:

Year	Name	Type
2010	HiBench	Big data benchmark suite for evaluating different big data frameworks. 19 workloads including synthetic micro-benchmarks and real-world applications from 6 categories which are **micro, machine learning, sql, graph, websearch and streaming.**
2015	SparkBench	System for benchmarking and simulating **Spark jobs**. Multiple workloads organized in 4 categories.
2010	Yahoo! Cloud System Benchmark (YSCB)	Evaluates performance of different **"key-value" and "cloud" serving systems**, which do not support the ACID properties. The YCSB++ , an extension, includes many additions such as multi-tester coordination for increased load and eventual consistency measurement.
2017	TPCx-IoT	Based on YCSB, but with significant changes. Workloads of data ingestion and concurrent queries simulating workloads on typical **IoT Gateway systems**. Dataset with data from sensors from electric power station(s)

Some of the benchmarks to integrate (II)

Application-oriented benchmarks:

Year	Name	Type
2015	Yahoo Streaming Benchmark (YSB)	The Yahoo Streaming Benchmark is a **streaming application benchmark** simulating an **advertisement analytics** pipeline.
2013	BigBench/TPCx-BB	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.
2017	BigBench V2	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
2018	ABench (Work-in-Progress)	New type of multi-purpose Big Data benchmark covering **many big data scenarios and implementations. Extends other benchmarks such as BigBench**

Contacts

info@databench.eu

@DataBench_eu

DataBench

DataBench Project

DataBench

DataBench Project

DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

gcattaneo@idc.com

rstevens@idc.com

Micro-benchmarks

1 HiBench (version 7.1 - https://github.com/intel-hadoop/HiBench)

DataBench

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

② HiBench (version 7.1 - https://github.com/intel-hadoop/HiBench)

Data Type	Most workloads use synthetic data generated from real data samples. The workloads use structured and semi-structured data.
Metrics	The measured metrics are execution time (latency), throughput and system resource utilizations (CPU, Memory, etc.).
Implementation	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

1 SparkBench (https://github.com/CODAIT/spark-bench)

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

② SparkBench(https://github.com/CODAIT/spark-bench)

Data Type	**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.**
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
Implementation	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.

1 *Yahoo! Cloud System Benchmark (YSCB)*

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 YCSB++ , an extension of the YCSB framework, includes many additions such as multi-tester coordination for increased load and eventual consistency measurement.**
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.
Workload	YCSB provides a core package of 6 pre-defined workloads A-F, which simulate a 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**, **Update**, **Read** and **Scan**. The YCSB workload consists of random operations defined by one of the several built-in distributions: **Uniform, Zipfian, Latest** and **Multinomial.**

2 *Yahoo! Cloud System Benchmark (YSCB)*

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

1 TPCx-IoT

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).
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.
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.

2 TPCx-IoT

Data Type	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.
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
Implementation	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.

Application Benchmarks

1 BigBench/ TPCx-BB

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).**
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.
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.

2 BigBench/ TPCx-BB

Data Type	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.
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
Implementation	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.

1 Yahoo Streaming Benchmark (YSB)

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 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. These steps attempt to probe some common operations performed on data streams.**
Domain	The Yahoo Streaming Benchmark is a streaming application benchmark simulating an advertisement analytics pipeline.
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.

2 Yahoo Streaming Benchmark (YSB)

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

1 BigBench V2

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.**
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.
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.

2 BigBench V2

Data Type	**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. We decided to generate product reviews 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.**
Metrics	BigBench V2 uses the same metric definition and computation as BigBench.
Implementation	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.

ABench: Big Data Architecture Stack Benchmark (*Work-in-Progress*)

- New type of multi-purpose Big Data benchmark covering many big data scenarios and implementations.

- Benchmark Framework
  - Data generators or plugins for custom data generators
  - Include data generator or public data sets to simulate workload that stresses the architecture

- Reuse of existing benchmarks
  - Case study using BigBench (in the next slides, Streaming and Machine Learning)

- Open source implementation and extendable design

- Easy to setup and extend

- Supporting and combining all four types of benchmarks in ABench

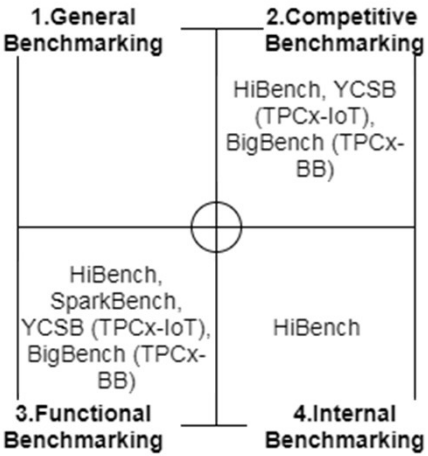
ABench Benchmarks Types (Andersen and Pettersen)

1. **Generic Benchmarking:** checks whether an implementation fulfills given business requirements and specifications (*Is the defined business specification implemented accurately?).*

2. **Competitive Benchmarking:** is a *performance comparison* between the best tools on the platform layer that offer similar functionality (*e.g., throughput of MapReduce vs. Spark vs. Flink*).

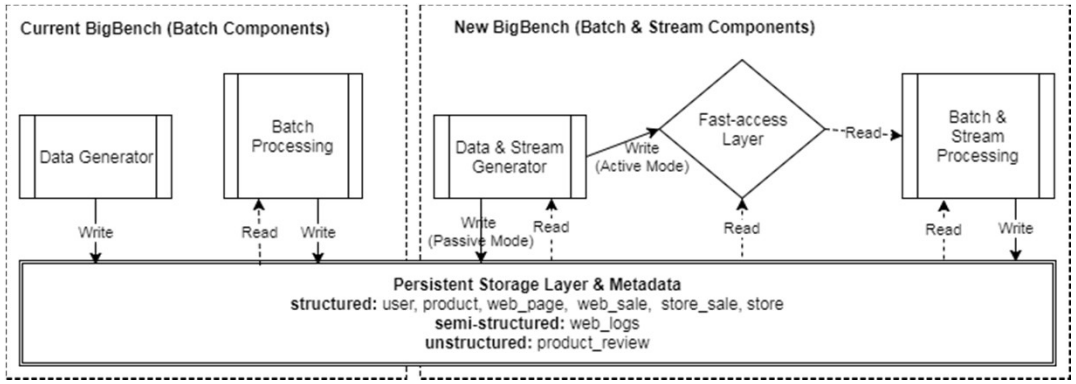
3. **Functional Benchmarking** is a *functional comparison* of the features of the tool against technologies from the same area. (*e.g., Spark Streaming vs. Spark Structured Streaming vs. Flink Streaming*).

4. **Internal Benchmarking:** *comparing different implementations* of a functionality (*e.g., Spark Scala vs. Java vs. R vs. PySpark*)



Stream Processing Benchmark – Use Case

- Adding stream processing to BigBench
- Reuse of the web click logs in JSON format from BigBench V2
- Adding new streaming workloads
  - possibility to execute the queries on a subset of the incoming stream of data
- Provide benchmark implementations based on Spark Streaming and Kafka



- Work In-progress: Exploratory Analysis of Spark Structured Streaming, @PABS 2018, Todor Ivanov and Jason Taaffe

Machine Learning Benchmark – Use Case

- Expanding the type of Machine Learning workloads in BigBench
  - five (Q5, Q20, Q25, Q26 and Q28) out of the 30 queries cover common ML algorithms

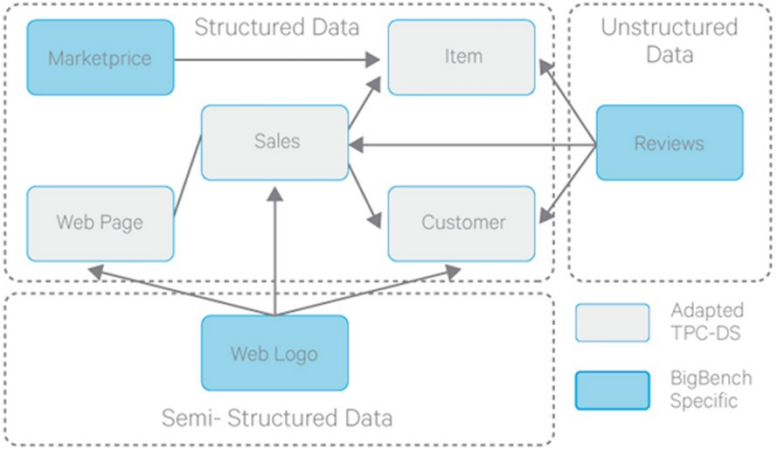
- Other types of advanced analytics inspired by Gartner (https://www.gartner.com/doc/3471553/-planning-guide-data-analytics)
  - descriptive analytics
  - diagnostic analytics
  - predictive analytics
  - prescriptive analytics

- Introduce new ML metrics for scalability and accuracy

BIGBENCH

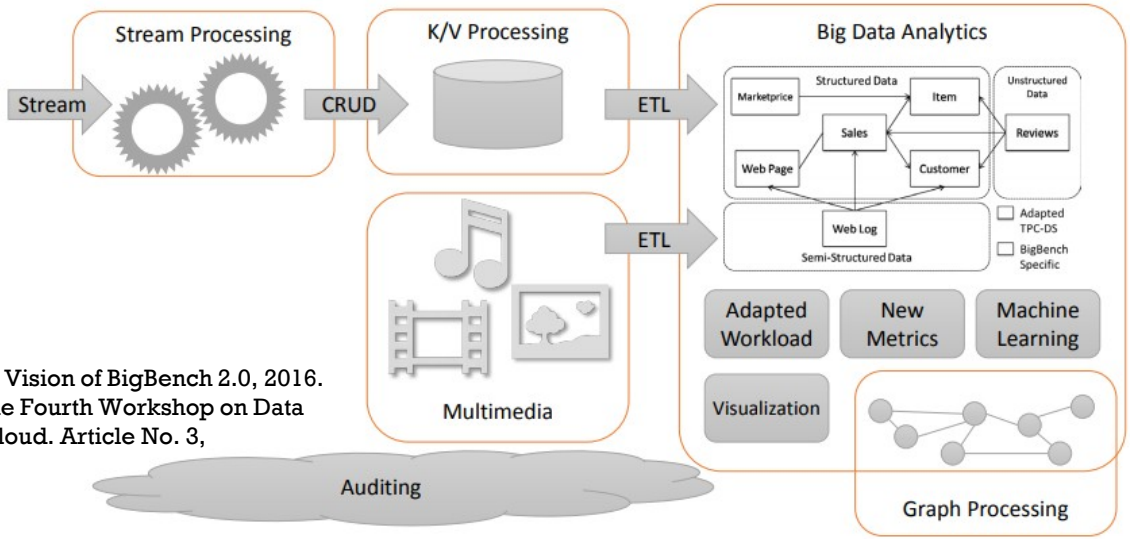
- The BigBench specification comprises two key components:
  - a data model specification
  - a workload/query specification.
- The structured part of the BigBench data model is adopted from the TPC-DS data model
- The data model specification is implemented by a data generator, which is based on an extension of PDGF.
- BigBench 1.0 workload specification consists of 30 queries/workloads (10 structured from TPC-DS, and 20 adapted from a McKinsey report on Big Data use cases and opportunities).
- BigBench 2.0 …

The BigBench data model



http://blog.cloudera.com/blog/2014/11/bigbench-toward-an-industry-standard-benchmark-for-big-data-analytics/

The BigBench 2.0 overview



Rabi T., et al. The Vision of BigBench 2.0, 2016. Proceedings of the Fourth Workshop on Data analytics in the Cloud. Article No. 3,

Summary

- DataBench:
  - A **framework for big data benchmarking** for PPP projects and big data practitioners
  - We will provide **methodology** and **tools**

- Added value:
  - An **umbrella to access to multiple benchmarks**
  - Homogenized **technical metrics**
  - Derived **business KPIs**,
  - A **community around**

- PPP projects, industrial partners (BDVA and beyond) and benchmarking initiatives are welcomed to work with us, either to use our framework or to add new benchmarks

Big Data Benchmark session at EBDVF'2018

Monday November 12th, 1700 – 1830,EBDVF'2018, Vienna

17.00 - 17.05 Introduction - Arne Berre/Axel Ngonga

17.05 - 17.20 Designing Big Data Benchmarks - Irini Fundulaki

17.20 - 17.35 LDBC - Peter Boncz

17.35 - 17.50 DataBench - Gabriella Cattaneo/Tomas P. Lobo

17.50 - 18.05 Holistic Benchmarking - Axel Ngonga/Gayane Sedrakyan

18.05 - 18.15 Using HOBBIT for Industrial applications - Pavel Smirnov (AGT)

18.15 - 18.25 The EU Big Data Inducement price challenge - Kimmo Rossi (EC)

18.25 - 18.30 Summary and discussion

Contacts

info@databench.eu

@DataBench_eu

DataBench

DataBench Project

DataBench

DataBench Project

DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

Arne.J.Berre@sintef.no

todor@dbis.cs.uni-frankfurt.de

tomas.parientelobo@atos.net