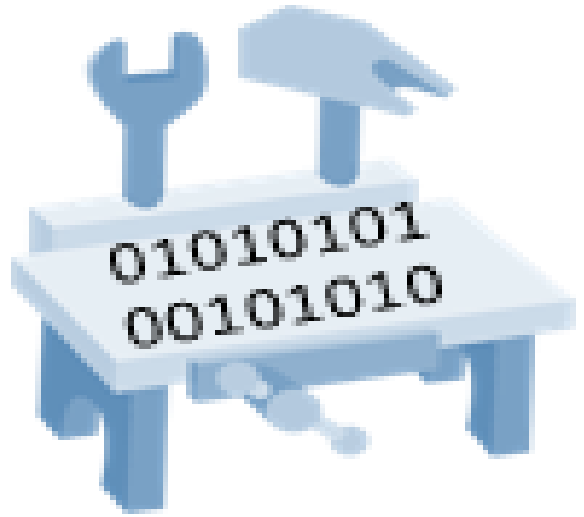
DATABENCH - Evidence Based Big Data Benchmarking to Improve Business Performance

ICT-17-2016-2017 - Big data PPP: Support, industrial skills, benchmarking and evaluation

RIA - January 1st, 2018-December 31st, 2020



Atos IDC ANALYZE THE FUTURE IJS SINTEF POLITECNICO MILANO 1863 Frankfurt Big Data Lab Databases and Information Systems (DBIS)

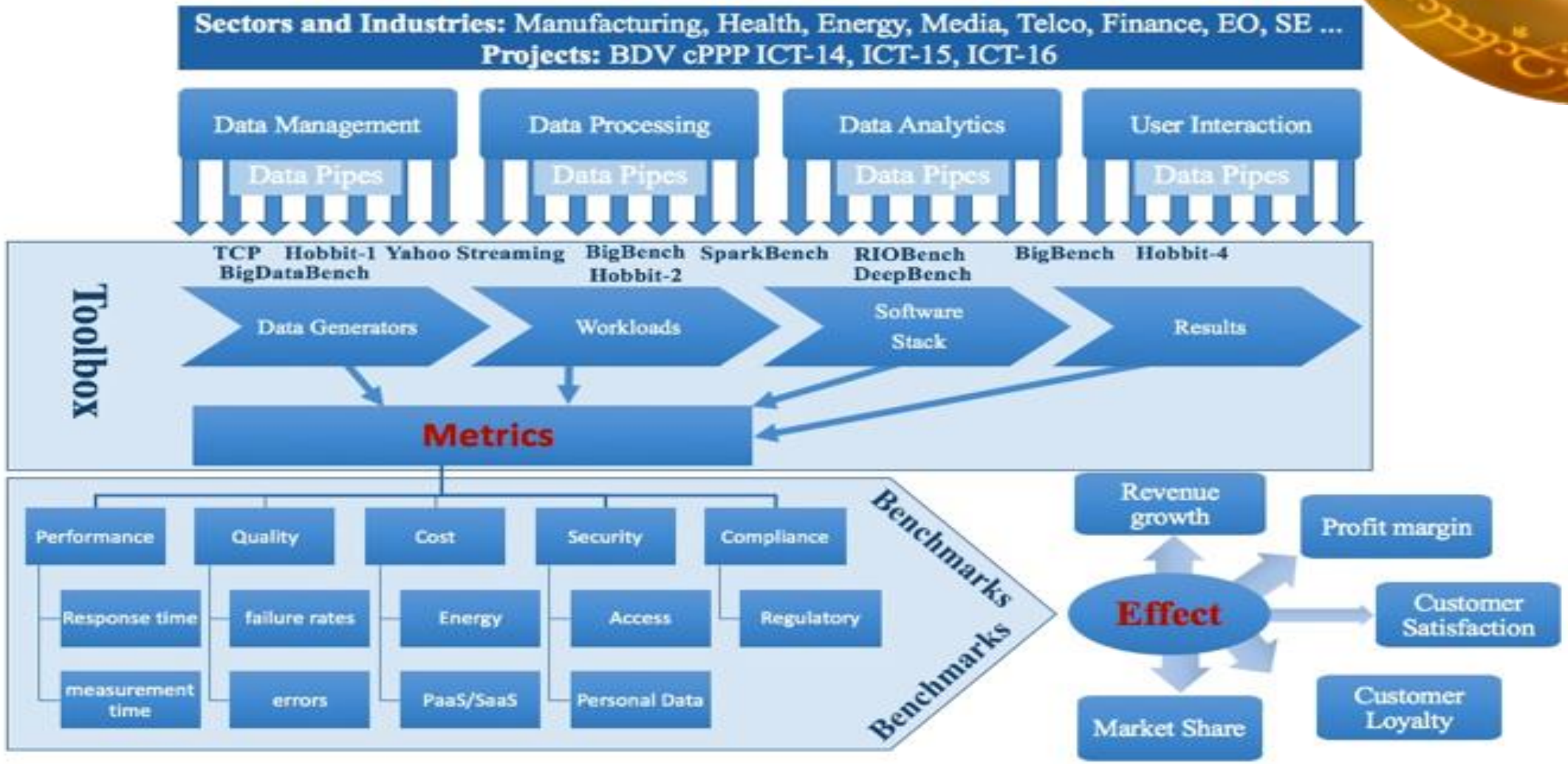
Main Objectives

1. Provide the Big Data Technologies (BDT) stakeholder communities with a **comprehensive framework to integrate business and technical benchmarking** approaches for BDT
2. Perform **economic and market analysis to assess the "European economic significance"** of benchmarking tools and performance parameters
3. Evaluate the **business impacts of BDT** benchmarks of performance parameters of industrial significance
4. **Develop a tool** applying methodologies to determine optimal BDT benchmarking approaches
5. **Evaluation of the DataBench Framework and Toolbox** in representative industries, data experimentation/integration initiatives (ICT-14) and Large Scale Pilot (ICT-15)
6. **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

Approach

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

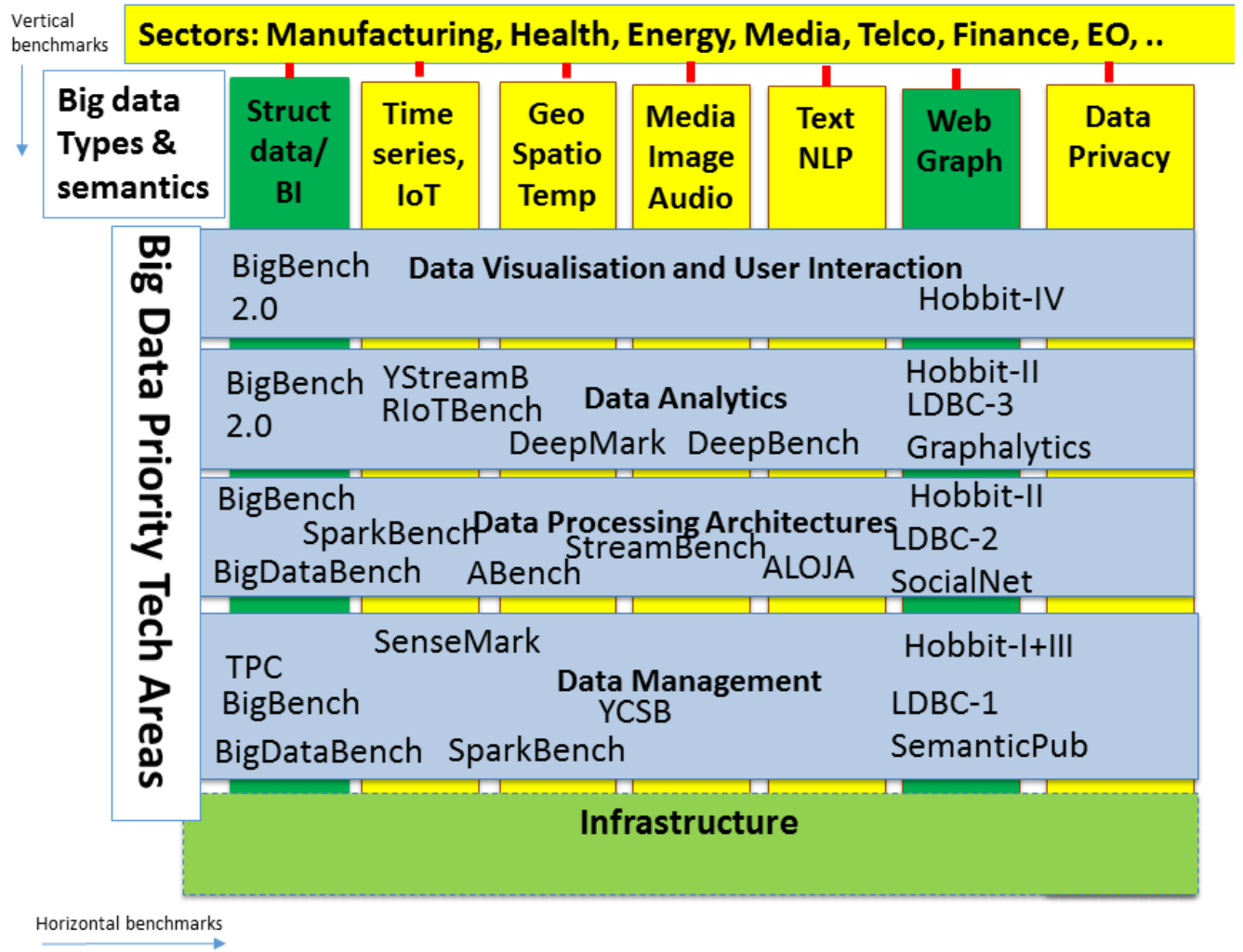
One **ToolBox to rule them all**,
One ToolBox to find them,
One ToolBox to bring them all
and to DataBench bind them



Based on the BDVA Reference Model

Aligned to BDVA:

Analysis of benchmarks for different data types, steps of the Data Value Chain and verticals of industrial relevance



DataBench success - Takeaways

- Not reinventing the wheel, but using wheels to build a new car
  - DataBench toolbox as integration of multiple existing benchmarks
  - Filling gaps
- Maximize Impact
  - Translation from performance & technical to business KPIs
- Sustainability
  - Sustainable and globally supported and recognized Big Data benchmarks
  - Aligned with BDVA and other existing initiatives (i.e. BDVe and the Hobbit project)
- Right partners

THANKS

- Blanca Jordán
- blanca.jordan@atos.net

BACK-UP SLIDES

Identifying and Selecting 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	RIoTBench	Hobbit Benchmark
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
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										
	Benchmarks	1999	2002	2007		2008		2009	2010		2011	2012				2013						2014			2015						2016			2017	

Expected Impacts

Availability of solid, relevant, consistent and comparable metrics for measuring progress in Big Data processing and analytics performance

Availability of metrics for measuring the quality, diversity and value of data assets

Sustainable and globally supported and recognized Big Data benchmarks of

Definition of methodology and metrics

WP1	DataBench Framework with benchmarks and metrics	Lead
T1.1	Holistic end-to-end Benchmarks – for Industry sectors	POLIMI
T1.2	DataBench Framework – with Vertical Big Data Type benchmarks	SINTEF
T1.3	Horizontal Benchmarks – Analytics and Processing	JSI
T1.4	Horizontal Benchmarks – Data Management	ATOS
D.1.1	Industry Requirements and sector specific specifications and KPIs	POLIMI
D.1.2	DataBench Framework – with Vertical Big Data type benchmarks	SINTEF
D.1.3	Horizontal Benchmarks – Analytics and Processing	JSI
D.1.4	Horizontal Benchmarks – Data Management	ATOS

Economic, Market and Business Analysis

WP2	Economic, Market and Business Analysis	Lead	16	2	10	3	3	4
T2.1	Development of the economic, market and business analysis methodology	IDC	3	1	1	1	1	1
T2.2	Preliminary assessment of European and industrial significance	IDC	3	1	2	1	0	1
T2.3	Analysis of actual and emerging needs of industrial users	POLIMI	2	0	1	0	1	1
T2.4	Identification, evaluation and mapping of BDA use cases by industry	POLIMI	3	0	4	0	1	1
T2.5	Benchmarks to Assess European Industrial Significance	IDC	5	0	2	1	0	0

Objective I

Provide the BDT stakeholder communities with a comprehensive framework to integrate business and technical benchmarking approaches for Big Data Technologies

The first objective of the proposal is to develop a BDT framework bringing together diverse BDT benchmarking solutions to provide a comprehensive benchmarking system able to respond to the real needs of European businesses, technology providers and the research community. DataBench will identify and unify the numerous existing BDT benchmarking initiatives and their business and technical metrics into a common structure based on the BDVA reference model. DataBench will investigate and deliver a single model to import and assess the technical requirements and data coming existing benchmarking tools and platforms based on the BDVA reference model and provide recommended benchmarks for dimensions from Big Data analytics through processing to data management, covering various Big Data types from structured data through time series/real-time streaming. The objective is to provide a model which correlates technical benchmarks to performance and business needs of different sectors and domains.

Atos

Objective II

Perform economic and market analysis to assess the “European economic significance" of benchmarking tools and performance parameters

This objective aims at measuring the relative relevance and impact of the industries developing and/or implementing the BDT benchmarks identified by the project, based on clear, coherent and evidence-based criteria. This will allow to assess the potential “footprint” in the European economy of the BDT benchmarks and provide a way to link the technical progress with the economic impacts. To do so we will leverage IDC’s research on BDT spending by industry and other indicators on the European Data Market and Data Economy developed on behalf of DG CONNECT. This data will be complemented with economic indicators drawn from public sources such as the Eurostat and ISCO on value added, overall turnover and job creation. A model of the European data economy direct, indirect and induced economic impacts will also be leveraged

The industrial significance of the performance parameters measured by the “best of breed” BDT benchmarks identified by DataBench will be measured through the analysis of the main use cases of Big Data Technologies by industry and their main business impacts. This will leverage valuable data about end-user investment priorities and the most frequent BDT use cases implemented by industry, measured by IDC’s annual survey of IT users by vertical market. This will ensure that the benchmarks identified respond to actual business needs and pave the way towards their acceptance and recognition by the industrial community.

Objective III

Evaluate the business impacts of BDT benchmarks of performance parameters of industrial significance

This objective will classify the main BDT use cases and their benchmarks in terms of their potential impacts on the main performance processes for the main manufacturing and services value chains, and assess their likely impact on business parameters such as improved revenues, reduced costs, improved efficiency. By correlating these results to the economic relevance of the sectors measured in Objective II we will be able to assess the scalability of their potential impact on the EU economy and the demonstrate the level of industrial significance of the BDT benchmarks selected. The result will be the development of business and industrial benchmarks of the progress driven by BDT, based on objective and evidence-based criteria.

Objective IV

Develop a tool applying methodologies to determine optimal BDT benchmarking approaches.

The objective is to develop a tool applying methodologies to determine optimal data management and analytics benchmarking approaches. The tool will enable the reutilization of existing accepted benchmarking efforts and leverage on top of them. Specifically, the tool will allow (a) the acquisition of resources by reusing existing benchmark frameworks and artefacts (datasets, workloads and software stacks), (b) the definition of new benchmarks, (c) the import and homogenisation of the results of the benchmarks to an agreed DataBench set of metrics, and (d) an easy to navigate user interface providing searching and visualization capabilities of the metrics and benchmarks suitable for monitoring of the results of Big Data projects or frameworks.

Metrics will be divided into three types: user-perceivable, architecture and business metrics. User perceivable metrics are those that matter for users. Examples of user-perceivable metrics are the duration of a test, request latency, and throughput. While user-perceivable metrics are used to compare performances of workloads of the same category or domain, architecture metrics are designed to compare workloads from different categories. Examples of architecture metrics are million instructions per second (MIPS) and million floating-point operations per second (MFLOPS). In addition, these metrics should not only measure system performance, but also take energy consumption, cost efficiency into consideration. These business metrics will be based on the previous ones to derive the potential impact of the benchmarked technologies from the business perspective. Aspects such as costs, security, quality and compliance will be derived and measured where possible

Objective V

Evaluation of the DataBench Framework and Toolbox in representative industries, data experimentation/integration initiatives (ICT-14) and Large Scale Pilot (ICT-15)

Different sectors have different requirements for Big Data. For example, if a benchmarking approach is effective at correlating Velocity to economic advantage it will be more beneficial to sectors like the financial industries, where the speed in managing and analysing data is important to users to generate competitive advantage. In sectors like pharmaceuticals or healthcare research where even small errors and failures are important, Veracity must be a priority in the benchmarks. Thus, the benchmarks must focus on turning data into Value for the user. This objective is focused on evaluating the performance of the DataBench Framework and Toolbox and ensuring the benchmarks and the system can handle real needs and requirements coming from the research community industry. The approach must first be future proof and be verified against the upcoming innovative Big Data analytics and management technologies coming out of ICT 14 and ICT 15 initiatives with a wide range of representative industries. Our objective is to assess and prove the validity of the approach in a sample of use cases drawn from the 5 leading industries in terms of BDT investment, sourced from IDC's continuously updated statistics of BDT spending, worldwide and in Europe during the activities in WP4 and WP5.

Objective VI

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 principal focus of this objective will be to liaise with data experimentation/integration initiatives and Large Scale Pilots in ICT 14 and 15 to involve them in DataBench activities and maximize synergies, while ensuring that the project and its results are available to other communities like the BDVA and external industry involved in the BDT benchmarking domain. The first part of this objective is to identify which projects and consortium are actively involved in the benchmarking discussion and which consortia are less actively involved or aware of benchmarking progress in Europe. DataBench's objective is to focus on understanding which measurement tools, techniques and processes are being used to assess, measure and compare the results and if the DataBench approach and tools might help them reach their goals. The objective includes reaching out to industry as well, initially through the industries involved in the research and how these projects are involved with industry and their intentions towards market uptake. The final objective to encourage ICT 14 and 15 projects to use the tools that are made available within the project at no cost and with adequate support to understand if the DataBench approach and Toolbox satisfies the real needs in the research and the industrial communities.

Atos