DataBench

Big Data
Technical Benchmarking

Arne J. Berre, SINTEF,
Todor Ivanov, Univ. Frankfurt,
Tomas Pariente Lobo, Atos

BDVe – Databench Webinar, October 9, 2018

IDC
ANALYZE THE FUTURE

Atos

Institut "Jožef Stefan", Ljubljana, Slovenija

SINTEF

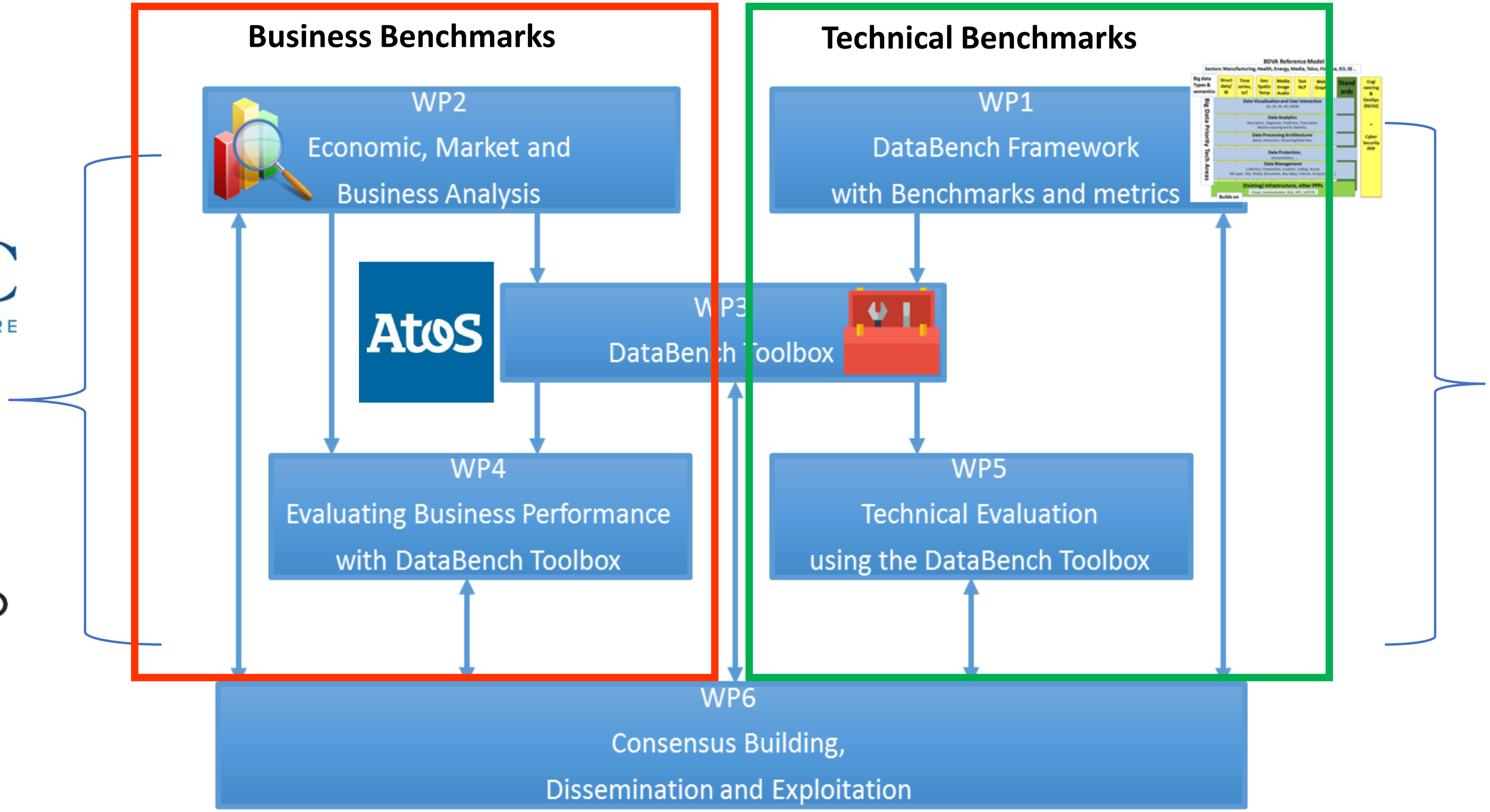
POLITECNICO
MILANO 1863

Frankfurt Big Data Lab
GOETHE UNIVERSITÄT

Technical Benchmarks in Databench Workflow

IDC
ANALYZE THE FUTURE

POLITECNICO
MILANO 1863



Business Benchmarks

Technical Benchmarks

WP2
Economic, Market and
Business Analysis

WP1
DataBench Framework
with Benchmarks and metrics

Atos

WP3
DataBench Toolbox

WP4
Evaluating Business Performance
with DataBench Toolbox

WP5
Technical Evaluation
using the DataBench Toolbox

WP6
Consensus Building,
Dissemination and Exploitation

SINTEF

Frankfurt Big Data Lab
Databases and Information Systems (DBIS)

IJS

IDC
ANALYZE THE FUTURE

Atos

SINTEF

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.

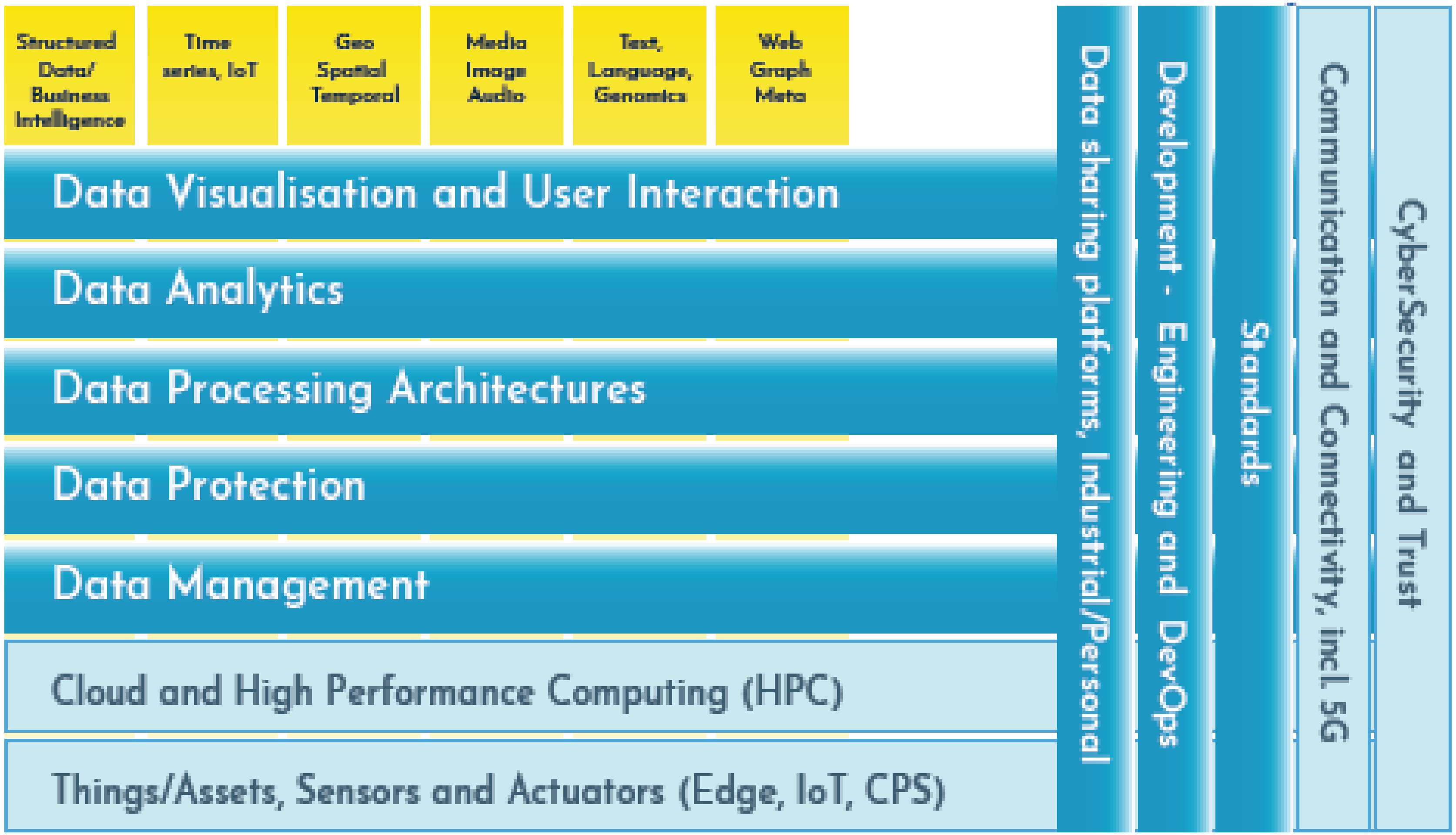
BDV Reference Model

KEY

Core BDVA

In Collaboration

Data Types



Structured Data/ Business Intelligence

Time series, IoT

Geo Spatial Temporal

Media Image Audio

Text, Language, Genomics

Web Graph Meta

Data Visualisation and User Interaction

Data Analytics

Data Processing Architectures

Data Protection

Data Management

Cloud and High Performance Computing (HPC)

Things/Assets, Sensors and Actuators (Edge, IoT, CPS)

Data sharing platforms, Industrial/Personal

Development - Engineering and DevOps

Standards

Communication and Connectivity, incl. 5G

CyberSecurity and Trust

BDV BIG DATA VALUE ASSOCIATION

BDV SRIA

European Big Data Value

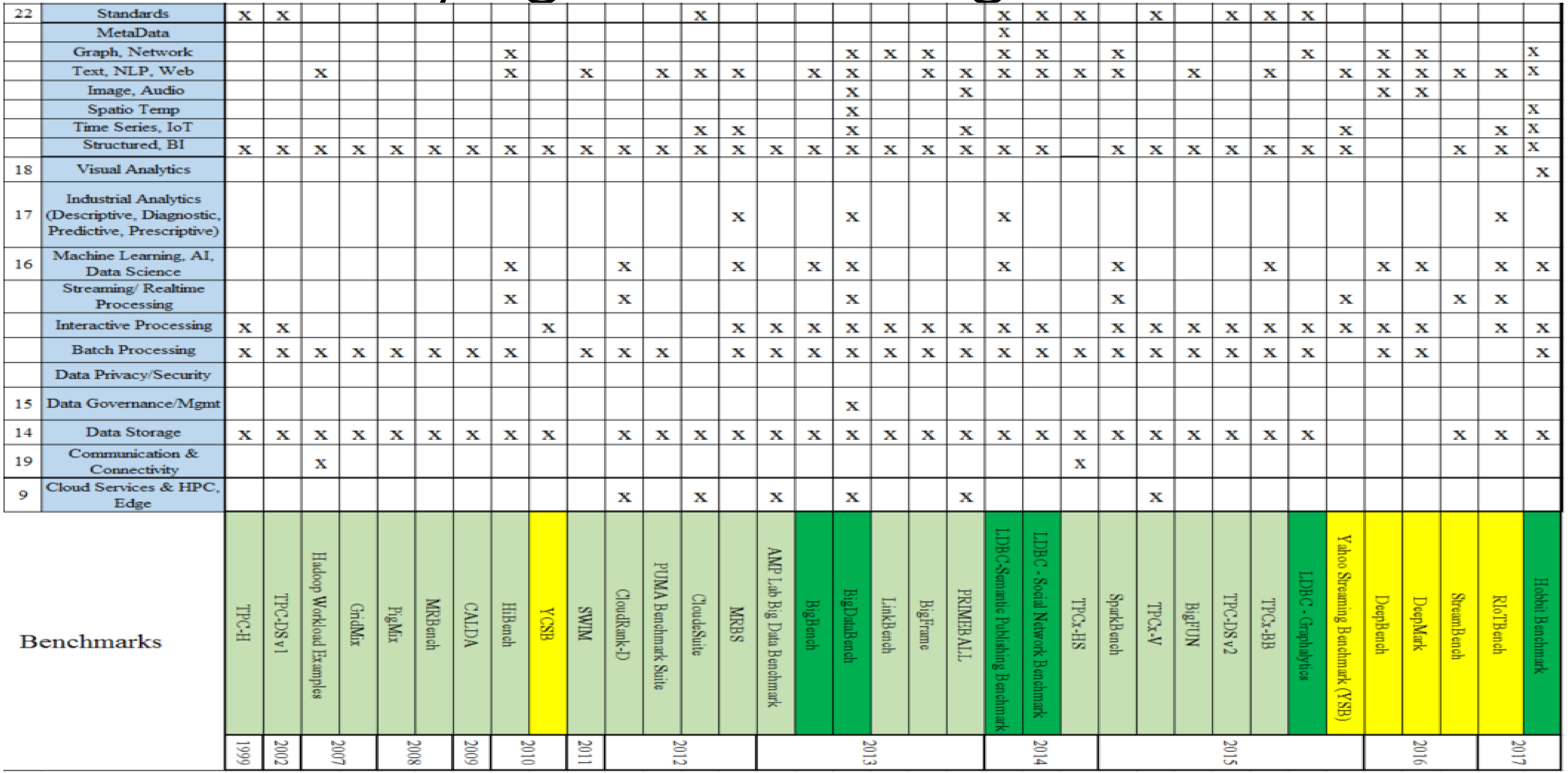
Strategic Research and Innovation Agenda

Version 4.0 October 2017

Accelerating Data-Driven Innovation in Europe

www.bdva.eu

Identifying and Selecting Benchmarks



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										
	Benchmarks	TPC-H	TPC-DSv1	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-DSv2	TPCx-BB	LDBC - Graphalytics	Yahoo Streaming Benchmark (YSB)	DeepBench	DeepMark	StreamBench	RIoTBench	Hobbit Benchmark
		1999	2002	2007		2008		2009	2010		2011	2012				2013						2014			2015							2016			2017	

BDVA Reference Model

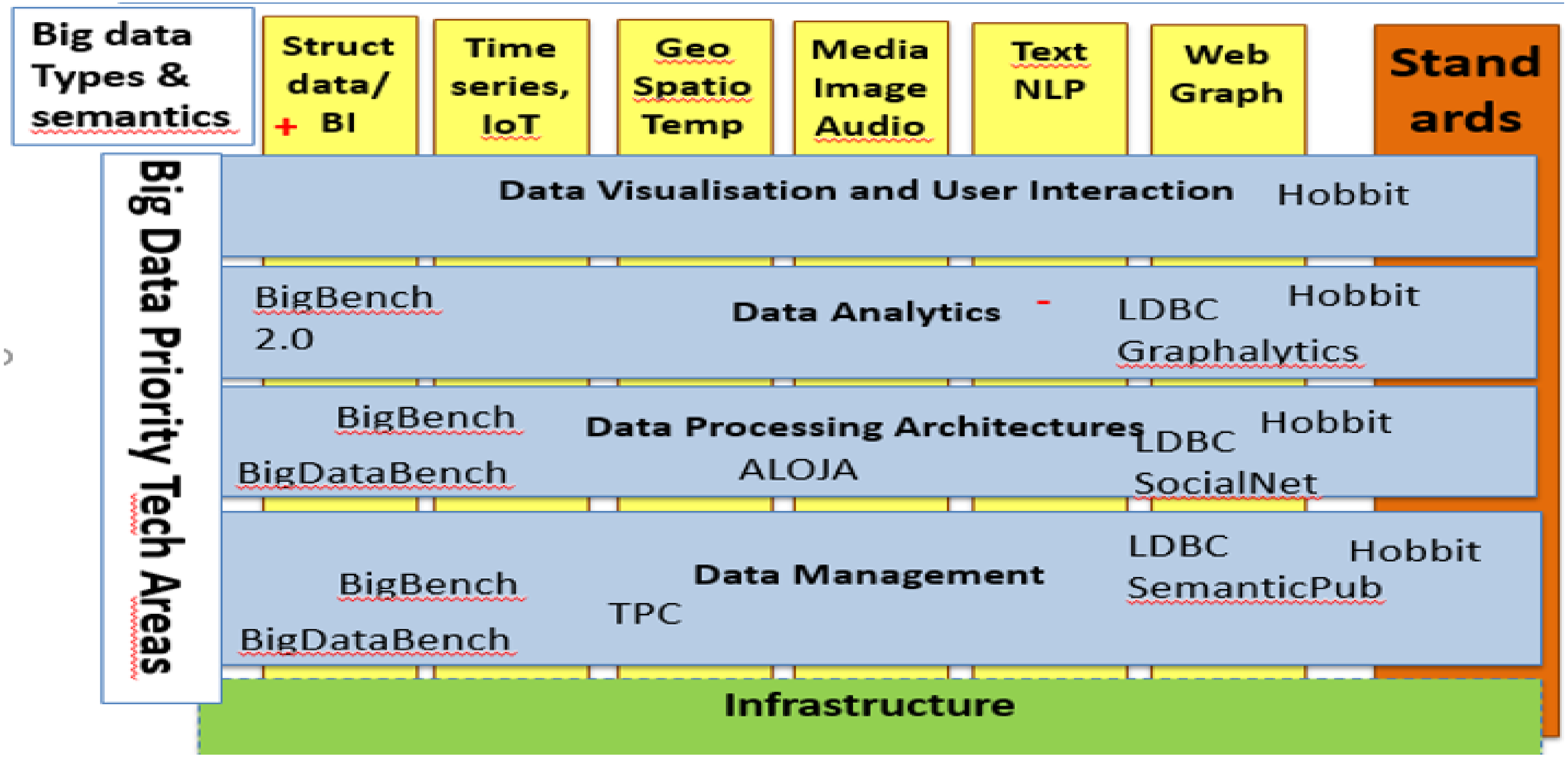
		Benchmarks	TPC-H	TPC-DS v1	Linear Road	Hadoop Workload Examples	GridMix	PigMix	MRBench	CALDA	HiBench	YCSB	SWIM	CloudRank-D	PUMA Benchmark Suite	CloudeSuite	MRBS	AMP Lab Big Data Benchmark
Verticals, incl. Data types	23	Domain/Sector/Business solutions KPIs (Manufact, Transport, Energy,..																
		Business																
		Transport																
		Manufacturing																
		Energy																
		.. Domain X ...																
	22	Standards	x	x												x		
		MetaData																
		Graph, Network									x							
		Text, NLP, Web				x					x		x		x	x	x	
		Image, Audio																
		Spatio Temp																
		Time Series, IoT														x	x	
		Structured, BI	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x
Analytics, Processing, Data Management, Infra	18	Visual Analytics																
	17	Industrial Analytics (Descriptive, Diagnostic, Predictive, Prescriptive)															x	
	16	Machine Learning, AI, Data Science									x			x			x	
		Streaming/ Realtime Processing									x			x				
		Interactive Processing	x	x								x					x	x
		Batch Processing	x	x		x	x	x	x	x	x		x	x	x		x	x
		Data Privacy/Security																
	15	Data Governance/Mgmt																
	14	Data Storage	x	x		x	x	x	x	x	x	x		x	x	x	x	x
	19	Communication & Connectivity				x												
	9	Cloud Services & HPC, Edge												x		x		x
			1999	2002	2004	2007		2008		2009	2010		2011	2012				2013

		Benchmarks	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)
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	
		MetaData						x										
		Graph, Network		x	x	x		x	x			x					x	
		Text, NLP, Web	x	x		x	x	x	x	x	x	x		x		x		x
		Image, Audio		x			x											
		Spatio Temp		x														
		Time Series, IoT		x			x											x
		Structured, BI	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x
Analytics, Processing, Data Management, Infra	18	Visual Analytics																
	17	Industrial Analytics (Descriptive, Diagnostic, Predictive, Prescriptive)		x				x										
	16	Machine Learning, AI, Data Science	x	x				x				x				x		
		Streaming/ Realtime Processing		x						x		x						x
		Interactive Processing	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	
		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	
	19	Communication & Connectivity									x							
	9	Cloud Services & HPC, Edge		x			x						x					
								2014				2015						

		Benchmarks	DeepBench	DeepMark	TensorFlow Benchmarks	Fathom	AdBench	RIoTBench	Hobbit Benchmark	TPCx-HS v2	BigBench V2	Sanzu	Penn machine learning benchmark (PML	OpenML benchmark suites	Senska	DAWNBench /MLPerf	IDEBench	ABench
Verticals, incl. Data types	23	Domain/Sector/Business solutions KPIs (Manufact, Transport, Energy,..																
		Business																
		Transport																
		Manufacturing																
		Energy																
		.. Domain X ...																
	22	Standards																
		MetaData																
		Graph, Network	x	x					x									
		Text, NLP, Web	x	x				x	x									
		Image, Audio	x	x														
		Spatio Temp							x									
		Time Series, IoT						x	x									
		Structured, BI						x	x									
Analytics, Processing, Data Management, Infra	18	Visual Analytics							x									
	17	Industrial Analytics (Descriptive, Diagnostic, Predictive, Prescriptive)						x										
	16	Machine Learning, AI, Data Science	x	x				x	x									
		Streaming/ Realtime Processing						x										
		Interactive Processing	x	x				x	x									
		Batch Processing	x	x					x									
		Data Privacy/Security																
	15	Data Governance/Mgmt																
	14	Data Storage						x	x									x
	19	Communication & Connectivity																
	9	Cloud Services & HPC, Edge																
			2016								2017							2018

Updating with new Benchmarks

Identifying and Selecting Benchmarks



Big data Types & semantics

Struct data/ + BI

Time series, IoT

Geo Spatio Temp

Media Image Audio

Text NLP

Web Graph

Standards

Big Data Priority Tech Areas

Data Visualisation and User Interaction
Hobbit

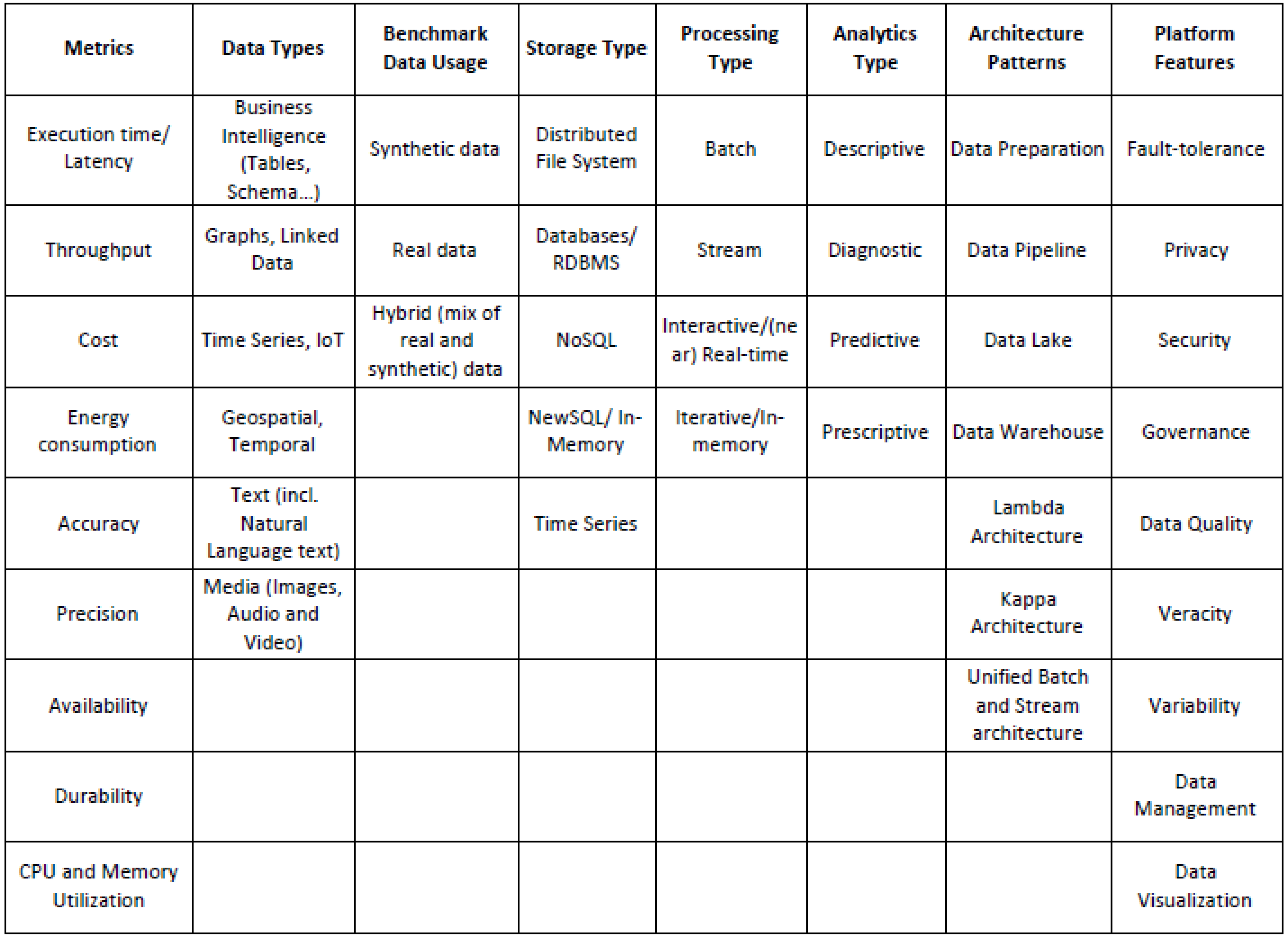
BigBench 2.0
Data Analytics -
LDBC Graphalytics
Hobbit

BigBench
BigDataBench
Data Processing Architectures
ALOJA
LDBC SocialNet
Hobbit

BigBench
BigDataBench
TPC
Data Management
LDBC SemanticPub
Hobbit

Infrastructure

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

Summary

DataBench

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)

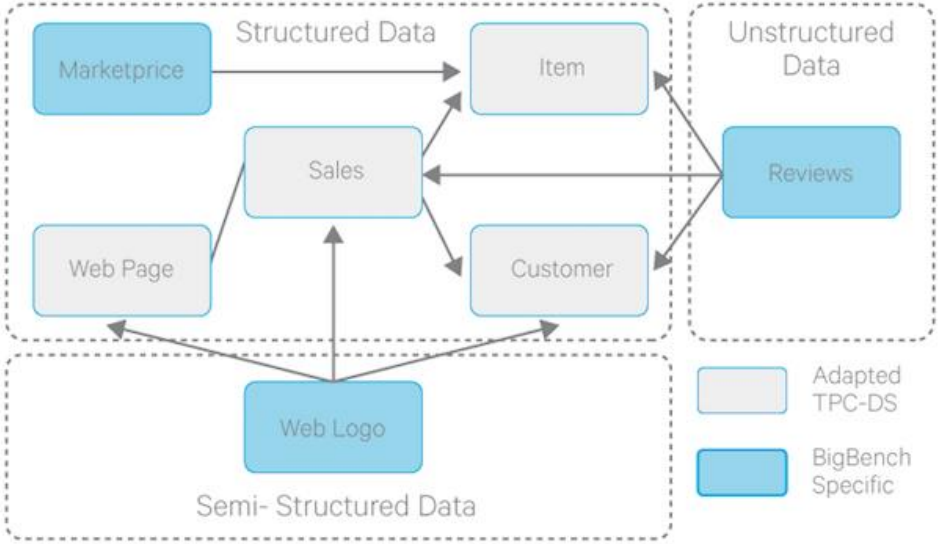
Application-oriented benchmarks:

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**

BIGBENCH

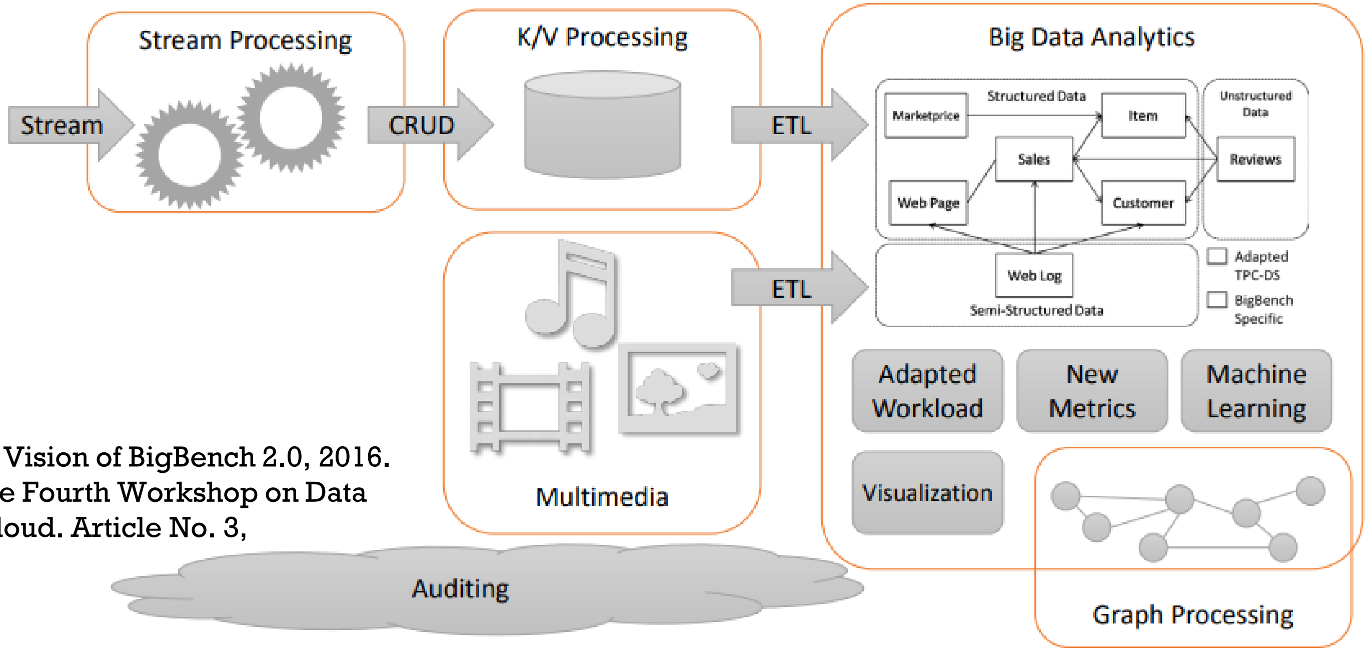
- The BigBench specification comprises two key components:
  - a data model specification
  - a workload/query specification.
- The structured part of the BigBench data model is adopted from the TPC-DS data model
- The data model specification is implemented by a data generator, which is based on an 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

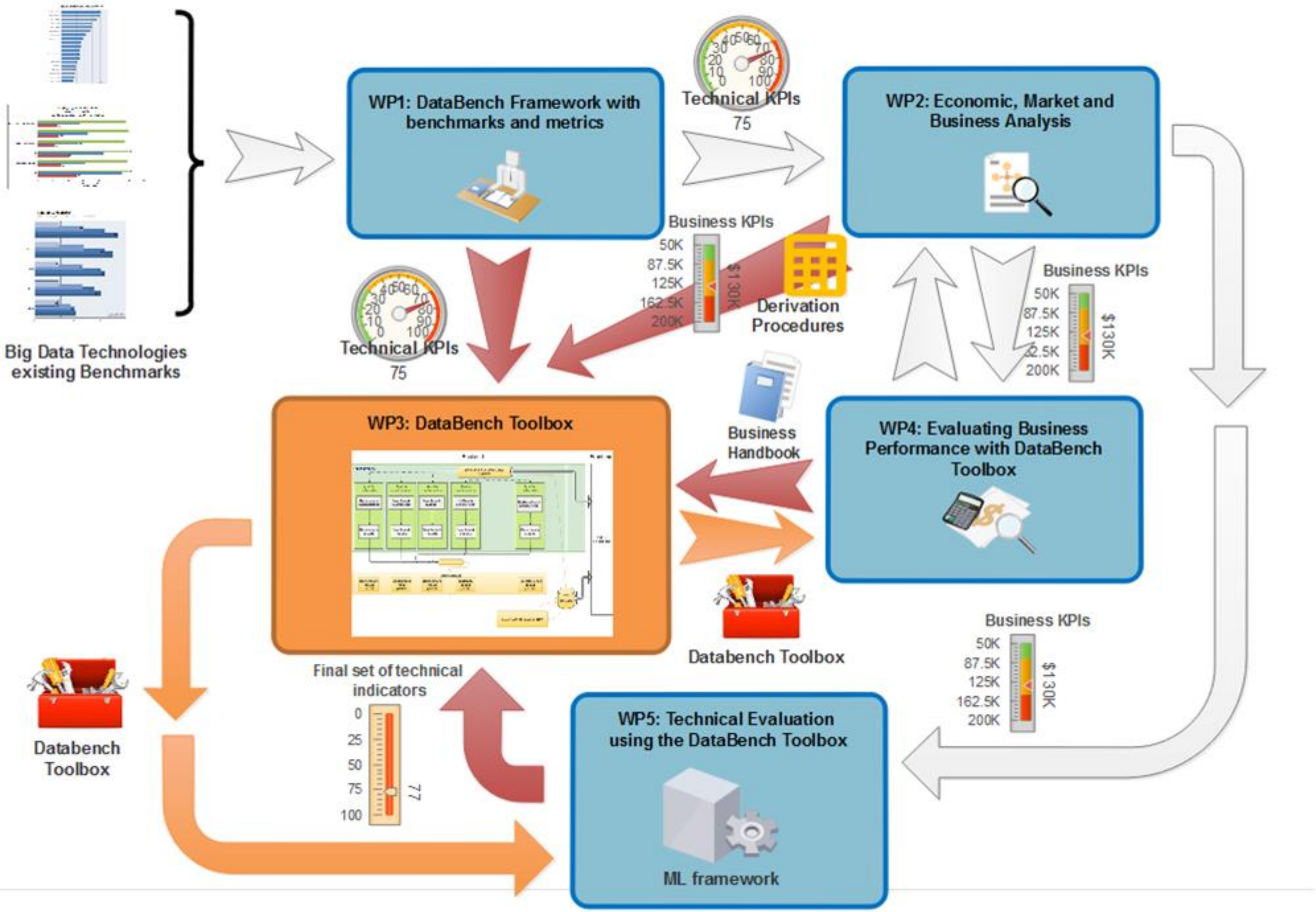


http://blog.cloudera.com/blog/2014/11/bigbench-toward-an-industry-standard-benchmark-for-big-data-analytics/

The BigBench 2.0 overview

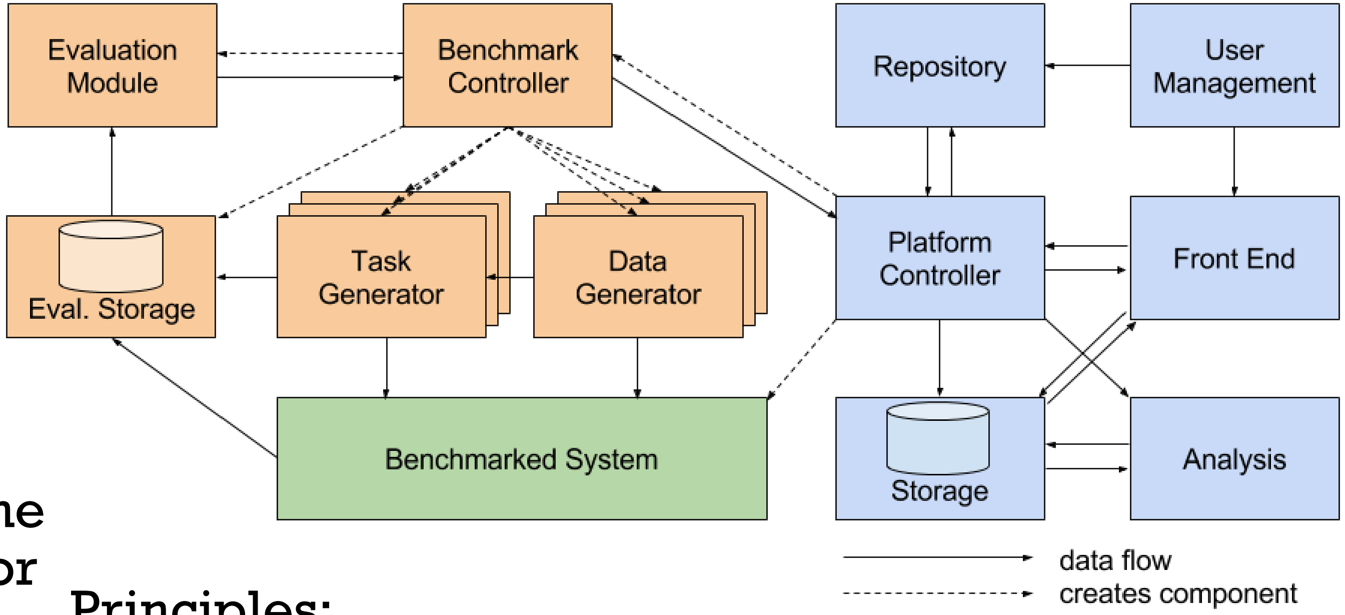


Rabi T., et al. The Vision of BigBench 2.0, 2016. Proceedings of the Fourth Workshop on Data analytics in the Cloud. Article No. 3,



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
Business KPIs
50K
87.5K
125K
162.5K
200K
$130K
Technical KPIs
75
WP3: DataBench Toolbox
Business Handbook
WP4: Evaluating Business Performance with DataBench Toolbox
Databench Toolbox
Business KPIs
50K
87.5K
125K
162.5K
200K
$130K
Final set of technical indicators
0
25
50
75
100
77
Databench Toolbox
WP5: Technical Evaluation using the DataBench Toolbox
ML framework

THE 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.

Summary

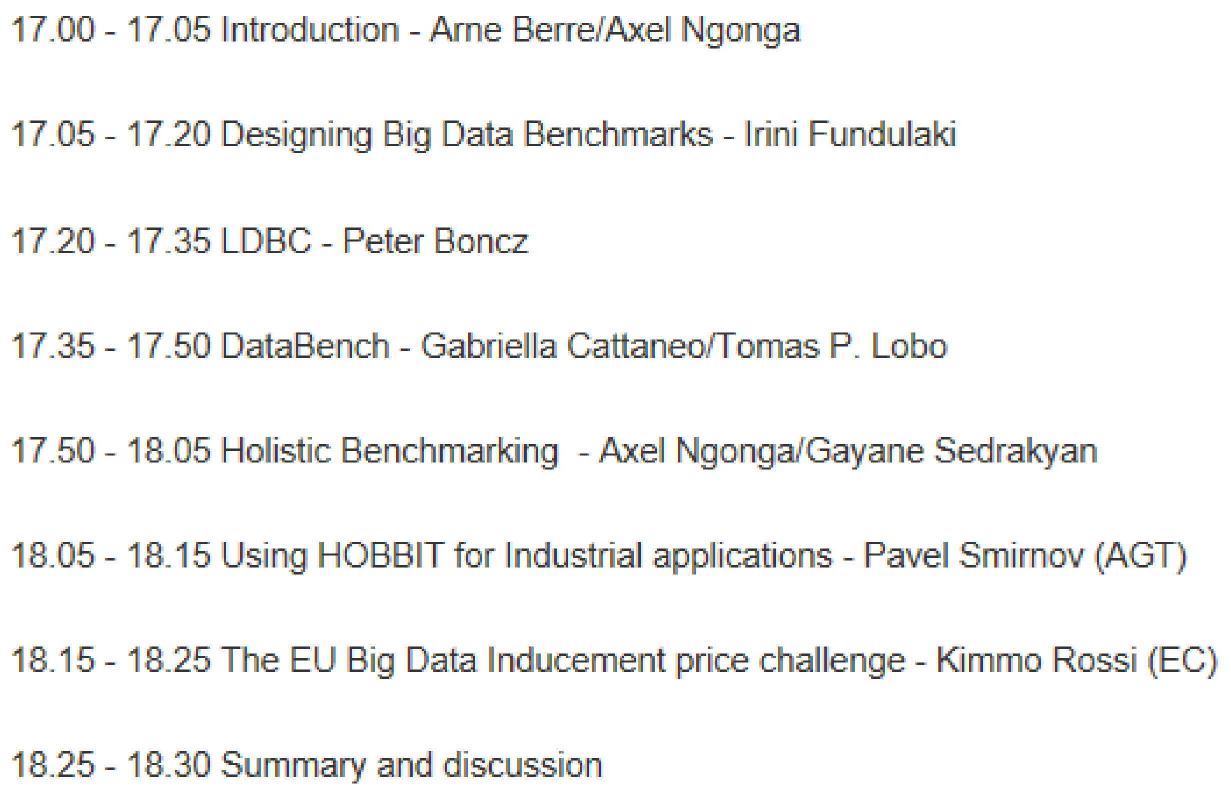
- DataBench:
  - A **framework for big data benchmarking** for PPP projects and big data practitioners
  - We will provide **methodology** and **tools**

- 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