ABench: Big Data Architecture Stack Benchmark [Vision Paper]

Todor Ivanov todor@dbis.cs.uni-frankfurt.de

Frankfurt Big Data Lab
GOETHE UNIVERSITÄT

Goethe University Frankfurt am Main, Germany
http://www.bigdata.uni-frankfurt.de/

Rekha Singhal rekha.singhal@tcs.com

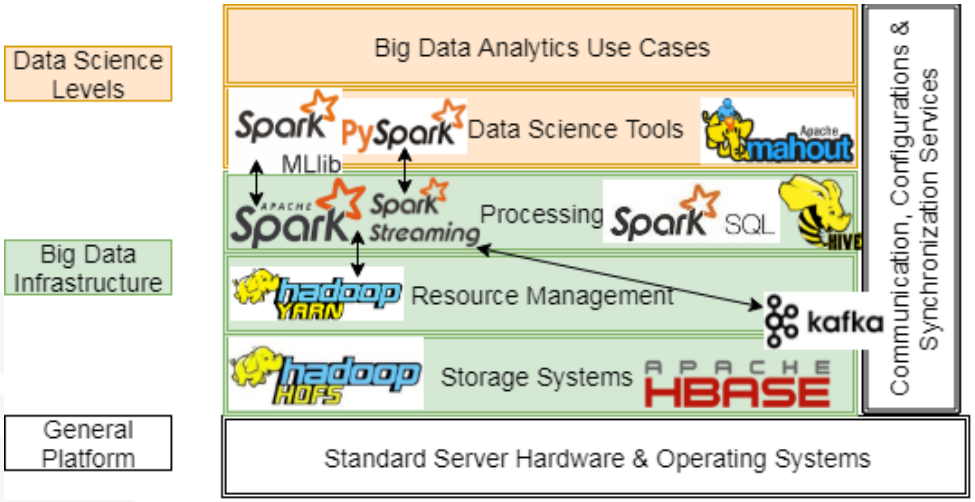
TATA CONSULTANCY SERVICES

TCS Research – Mumbai, India
http://www.tcs.com

Motivation

- Growing number of new ***Big Data technologies*** and ***connectors*** in the Big Data Stacks

  → Challenges for Solution Architects, Data Engineers, Data Scientist, Developers, etc.



- Missing benchmarks for ***each technology***, ***connector*** or a ***combination of them***

- Consequence → ***Increasing complexity in the Big Data Architecture Stacks***

- Our approach → ***ABench: Big Data Architecture Stack Benchmark***

ABench Features

- 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

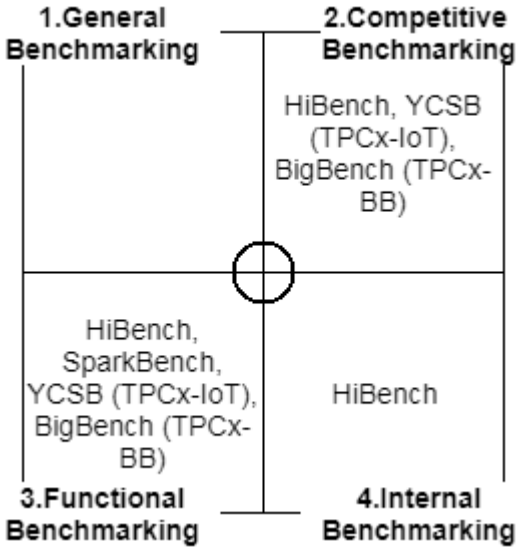
Benchmarks Types (adapted from Andersen and Pettersen [1])

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

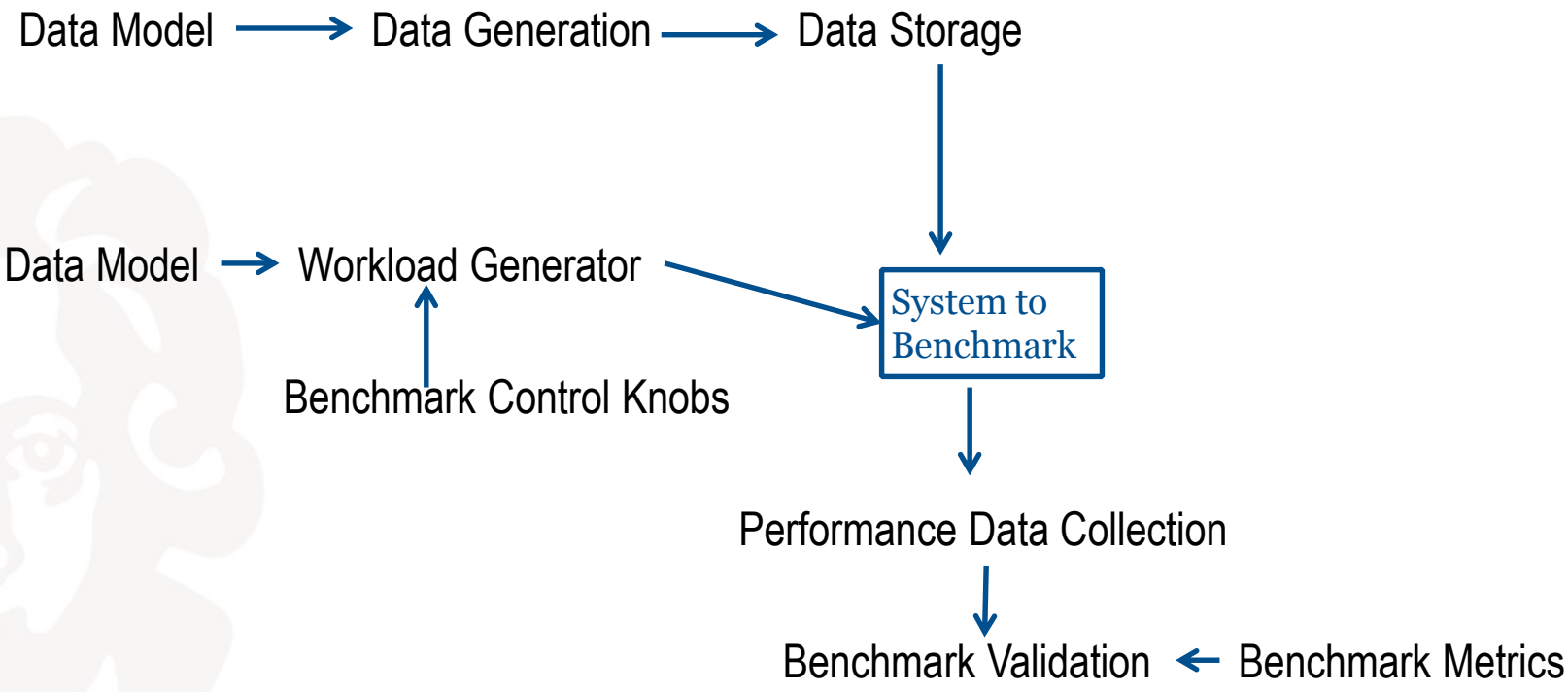
2. **Competitive Benchmarking:** is a *performance comparison* between the best tools on the platform layer that offer similar functionality (*e.g., throughput of MapReduce vs. Spark vs. Flink*).

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*)

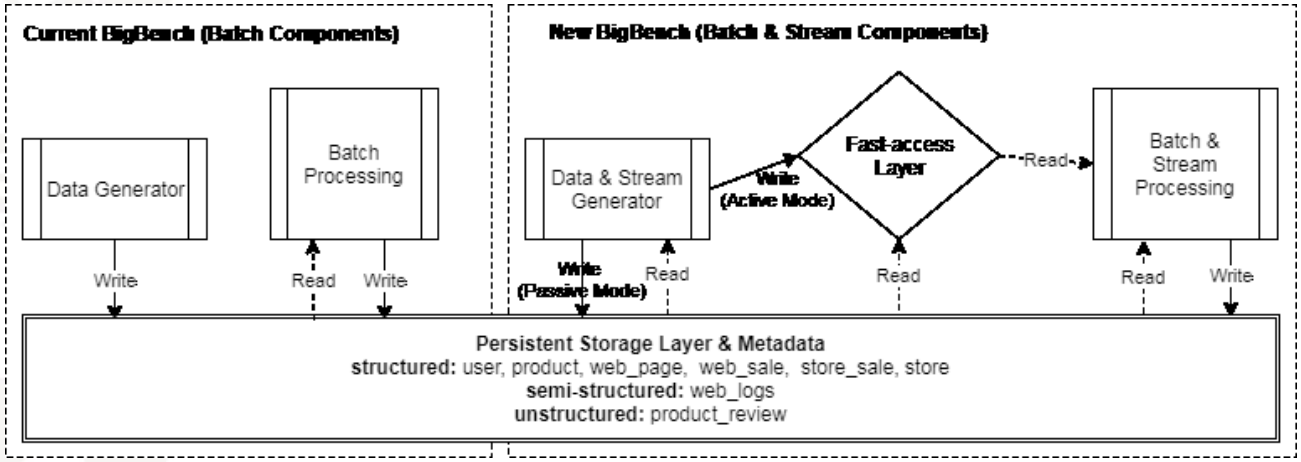


ABench Framework



Stream Processing Benchmark – Use Case

- Adding stream processing to BigBench [2,3]
- Reuse of the web click logs in JSON format from BigBench V2 [3]
- 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 [2]
  - five (Q5, Q20, Q25, Q26 and Q28) out of the 30 queries cover common ML algorithms

- Proposal by Sweta Singh (IBM)[4] for new workload with Collaborative Filtering using Matrix Factorization implementation in Spark MLlib via the Alternating Least Squares (ALS)

- Other types of advanced analytics inspired by Gartner [5]:
  - *descriptive analytics*
  - *diagnostic analytics*
  - *predictive analytics*
  - *prescriptive analytics*

- Introduce new ML metrics for scalability and accuracy

Next Steps

- Building express version of the benchmark framework
- Provide open source implementation of the Use Case benchmarks to stress test the existing Big Data Architecture Stacks
- Enable the comparison of the most popular technologies (e.g., Kafka, Spark, etc.)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Thank you for your attention!

This research has been supported by the Research Group of the Standard Performance Evaluation Corporation (SPEC).

REFERENCES

[1] Bjørn Andersen and P-G Pettersen. 1995. Benchmarking handbook. Champman & Hall.

[2] Ahmad Ghazal, Todor Ivanov, Pekka Kostamaa, Alain Crolotte, Ryan Voong, Mohammed Al-Kateb, Waleed Ghazal, and Roberto V. Zicari. 2017. BigBench V2: The New and Improved BigBench. In ICDE 2017, San Diego, CA, USA, April 19-22.

[3] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. BigBench: Towards An Industry Standard Benchmark for Big Data Analytics. In SIGMOD 2013. 1197–1208.

[4] Sweta Singh. 2016. Benchmarking Spark Machine Learning Using BigBench. In 8th TPC Technology Conference, TPCTC 2016, New Delhi, India, September 5-9, 2016.

[5] Gartner 2017, https://www.gartner.com/doc/3471553/-planning-guide-data-analytics