Tutorial on Benchmarking Big Data Analytics Systems

Todor Ivanov
todor@dbis.cs.uni-frankfurt.de
Frankfurt Big Data Lab, Goethe University
Frankfurt am Main, Germany

Rekha Singhal
rekha.singhal@tcs.com
TCS Research
Mumbai, India

ABSTRACT
The proliferation of big data technology and faster computing systems led to pervasions of AI based solutions in our life. There is need to understand how to benchmark systems used to build AI based solutions that have a complex pipeline of pre-processing, statistical analysis, machine learning and deep learning on data to build prediction models. Solution architects, engineers and researchers may use open-source technology or proprietary systems based on desired performance requirements. The performance metrics may be data pre-processing time, model training time and model inference time. We do not see a single benchmark answering all questions of solution architects and researchers. This tutorial covers both practical and research questions on relevant Big Data and Analytics benchmarks.

CCS CONCEPTS
- Computer systems organization → Distributed architectures.

KEYWORDS
Big Data, Analytics, ML, AI, Benchmarking

ACM Reference Format:

1 INTRODUCTION
In the age of Big Data, often characterized by the so called 3Vs (Volume, Velocity and Variety) [31], it is essential to use the right tools and best practices when implementing AI applications. Traditionally, benchmarking tools and methodologies have been used to compare different technologies both in terms of performance and functionality [16]. With the growing number of open source and enterprise tools in the Big Data Ecosystem [17], the need of standardized Big Data Benchmarks that provide accurate comparison between these new technologies has become very important [5]. The use of big data technologies for building machine-learning [39] and deep-learning [50] pipelines, and models has introduced more complexity in choosing the right model building framework, libraries and hardware architecture.

AI [3] has motivated the advances in hardware development, such as new hardware accelerators and configurable components [36] (e.g. NVMs (Non-Volatile Memory) [4], GPUs (Graphics Processing Unit) [42], FPGAs (Field Programmable Gate Array) [48, 49], TPUs (Tensor Processing Units) [9] and more), which suggest a complete rewriting of the existing software stack [29]. Such major changes in the backend systems impact both the processing and storage layers. In order to optimize and validate the benefits of the new software stack, suitable and standardized Big Data benchmarks comprising of machine-learning and deep-learning workloads are necessary.

Historically, technical benchmarking can be seen as a process of applying transparent and common methodologies to compare systems or software technologies. Jim Gray back in 1992 [16] described 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.”

In short, we can summarize that a software benchmark is a program used for comparison of software products/tools executing on a pre-configured hardware environment. There are different types of benchmarks that focus on specific functionalities:

Micro-benchmarks are either a program or routine to measure and test the performance of a single component or task [38]. They are used to evaluate either individual system components or specific system behaviors (or functions of codes) [19]. Micro-benchmarks report simple and well-defined quantities such as elapsed time, rate of operations, bandwidth, or latency [38]. Typically, they are developed for a specific technology, which reduces their complexity and development overhead. Popular micro-benchmark examples also part of the Hadoop binaries are WordCount, TestDFSIO, Pi, K-means, HiveBench and many others.

Application-level benchmarks also known as End-to-end benchmarks are designed to evaluate the entire system using typical application scenarios, each scenario corresponds to a collection of related workloads [19]. Typically, these type of benchmarks are more complex and are implemented using multiple technologies, which makes them significantly harder to develop. For example application-level Big Data benchmarks are the one standardized by the Transaction Processing Performance Council (TPC) [47] such as TPC-H, TPC-DS, BigBench(TPCx-BB) and many others.

Benchmark suites are combinations of different micro and/or end-to-end (application-level) benchmarks and these suites aim to provide comprehensive benchmarking solutions [19]. Examples for Big Data benchmark suites are HiBench [24], SparkBench [33], CloudSuite [8], BigDataBench [22], PUMA [1] and many others.
Another important distinction between benchmarks is if they are standardized by an official organization (like SPEC [45] or TPC [47]) or not standardized (typically developed by a vendor or research organization).

2 TUTORIAL STRUCTURE
The tutorial is organized for 3 hours with half an hour break in between. The tutorial covers the research questions and available benchmarks in the big data analytics domain. It aims to answer relevant questions such as:

- What benchmark to pick for a particular application?
- How to distinguish different available benchmarks for a given application?
- How do we address the different application requirements in terms of schema, heterogeneity of data types and technologies?
- How to pick the right benchmark type (micro-benchmark, application and suites) for particular use case?

The tutorial will focus on the following aspects of the big-data-analytic-systems-benchmarks:

- Summarize the Big Data challenges, requirements and features that emerging new technologies and benchmarks should address.
- Present an extensive overview of the current benchmark initiatives and organizations developed as part of the DataBench project classified according to relevant areas, technologies and architecture stacks.
- Present in detail popular and representative Machine Learning and Big Data benchmarks.
- Outline popular tools and methodologies for evaluating technologies and platforms by utilizing existing benchmarks.

3 BIG DATA ANALYTICS TECHNOLOGIES
Due to the growing number of new data platforms like Hybrid Transaction/Analytical Processing (HTAP) ([30, 35]), Distributed Parallel Processing Engines ([41], [18], [43], [13], etc.), Big Data Management ([2]), SQL-on-Hadoop-alike ([20], [44], [23], etc.) and Analytics Systems ([21]) integrating Machine Learning ([34], [32]), Deep Learning ([46]) and more, the emerging benchmarks try to follow the trend to stress these new system features. This makes the currently standardized benchmarks (such as TPC-C, TPC-H, etc.) only partially relevant for the emerging Big Data Management systems as they offer new features that require new analytics benchmarks.

Also, the data-driven nature of machine/deep learning workloads motivated research and development in specialized hardware such as TPU, GPU, etc.. The modern benchmark for big data analytics systems shall encompass heterogeneous middle-ware and hardware architectures.

4 BIG DATA BENCHMARKS
Figure 1 is an attempt to classify and categorize the most popular Big Data and Analytics benchmarks. We divided the benchmarks in six categories according to their workload, data type and use of Big Data technologies. These categories are micro-benchmarks, Big Data and SQL-on-Hadoop benchmarks, streaming benchmarks, machine learning and deep learning benchmarks, graph benchmarks and new emerging benchmarks. Below are summarized popular benchmarks that will be presented in the tutorial:

4.1 BigBench
BigBench [6, 15] is an end-to-end application-level 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). It consists of 30 complex queries. In 2016, TPC standardized BigBench as TPCx-BB.

BigBench V2 [14] 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.

BigBench Streaming Extension [26] extends the BigBench V2 benchmark with a data streaming component that simulates typical statistical and predictive analytics queries in a retail business scenario. The goal is to preserve the existing BigBench design and introduce a new streaming component that supports two data streaming modes: active and passive. In active mode, the data stream generation and processing happen in parallel, whereas in passive mode, the data stream is pre-generated in advance before the actual stream processing. The stream workload consists of five queries inspired by the existing 30 BigBench queries.

ADABench [40] is an end-to-end ML benchmark covering the complete ML lifecycle, from data preparation all the way to inference. It covers 16 real business use cases including different scale ranges. Currently six cases are implemented (in Python and Spark) covering various dimensions of analytics in the retail business vertical.

4.2 BigDataBench
BigDataBench [7] is an open source Big Data benchmark suite [22] consisting of 14 data sets and 33 workloads. Six of the 14 data sets are real-world based, generated using the BDGS [12] data generator. The generated data types include text, graph, and table data, and are fully scalable. The 33 workloads are divided into five common application domains: search engine, social networks, electronic commerce, multimedia analytics, and bioinformatics.

4.3 MLPerf
The MLPerf1 [10, 11] project aims to build a common set of benchmarks that enables the machine learning field to measure system performance for both training and inference from mobile devices.
to cloud services. It aims to collect publicly available data sets and models for the following problems: Image classification, Object detection, Translation, Recommendation, Reinforcement Learning, Speech to text and Sentiment Analysis.

4.4 DataBench

DataBench² [27, 37] is a three year EU-funded project that investigates existing Big Data benchmarking tools and projects, identifies the main gaps and provides a robust set of metrics to compare technical results coming from those tools. The DataBench Toolbox is a one-stop-shop for Big Data Benchmarking, offering multiple benefits for different kind of users and businesses.

4.5 ABench

ABench [28] is Big Data Architecture Stack Benchmark that targets the representation and comparison of different Big Data architecture patterns. The benchmark framework shall stress test the common application business requirements (e.g. retail analytics, retail operational, etc.), big data technologies functionalities and best practice implementation architectures. The benchmark framework should have an open source implementation and extendable design as well as easy to be setup and extend. It should include data generator, public data sets and existing benchmarks to simulate workloads that stress test the best practice Big Data architectures.

5 CONCLUSIONS

AI is pervasive in our life and there is need to understand how to benchmark systems, which are used to build machine-learning and deep-learning based solutions using emerging big data technology stacks. This tutorial covers the research questions and available benchmarks in this domain. We have discussed in detail popular benchmarks such as BigBench, BigDataBench, MLPerf and the DataBench project.

ACKNOWLEDGMENTS

This work has been partially funded by the European Commission H2020 project DataBench - Evidence Based Big Data Benchmarking to Improve Business Performance, under project No. 780966. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained

---

²www.databench.eu
in this work. The authors thank all the participants in the project for discussions and common work.

REFERENCES
[34] MLlib. 2020. spark.apache.org/mllib/