Evidence Based Big Data Benchmarking to Improve Business Performance

DataBench Toolbox - supporting Big Data and AI Benchmarking

Dr. Arne J. Berre, SINTEF
Todor Ivanov, LeadConsult
Tomas Pariente, Atos

Bench´20, November 16th, 2020

www.databench.eu
DataBench Toolbox – supporting Big Data and AI Benchmarking

Arne Berre
Chief Scientist, SINTEF

DataBench Framework for Benchmarks and project pipeline examples

Todor Ivanov
Senior consultant, Lead Consult

DataBench Benchmarks and Communities

Tomás Pariente
Associate Head of AI, Data & Robotics Unit, Atos Research and Innovation

The DataBench Toolbox – Demonstration
+ further relationships also to the ISO SC42 AI and Big Data Reference Models
Top level Generic Big Data and AI Pipeline pattern

(For all Benchmarks and Project pipeline to be related to)
Big Data Analytics
=
Big Opportunities for EU Companies

Catalogue of technical Benchmarks
DataBench has compiled a list of technical benchmarks that can help you with measuring your Big Data system, compare it with others and help in fine-tuning it.

About DataBench Toolbox
Based on existing efforts in big data benchmarking and enabling inclusion of new benchmarks that could arise in the future, the DataBench Toolbox provides a unique environment to search, select and deploy big data benchmarking tools, giving the possibility to generate unified technical metrics and, most importantly, going the extra mile and derive business KPIs for your organization.

User journeys: What type of user are you?
Whether you are more interested in the technical aspects of benchmarking, or your focus lays more on the business aspects we have prepared a set of user-journeys ready to help you while working with this platform.

Just select from the titles below the one that you are more interested in to see a page with advices

- Technical
- Business
- Benchmark provider
### Benchmark catalogue

**BigBench V2**
The BigBench V2 benchmark addresses some of the limitation of the BigBench (TPC-DS) 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 differ...

**HiBench**
A comprehensive benchmark suite consisting of multiple workloads including both synthetic micro-benchmarks and real-world applications. HiBench features several ready-to-use benchmarks from 4 categories: micro benchmarks, Web search, Machine Learning, and HDFS benchmarks. It is used for both stream...

**Yahoo Streaming Benchmark (YSB)**
It is an end-to-end pipeline that simulates a real-world advertisement analytics pipeline. Currently implemented in Kafka, Storm, Spark, Flink and Redis. Yahoo reported the following as background of why they developed YSB: “At Yahoo we have adopted Apache Storm as our stream processing p..." 

**Yahoo Cloud Serving Benchmark (YCSB)**
A benchmark designed to compare emerging cloud serving systems like Cassandra, HBase, MongoDB, Riak and many more, which do not support ACID. It provides a core package of 6 pre-defined workloads A-F, which simulate a cloud OLTP application. Web references https://github.com/brianfrancisco/...

**ABench**
ABench is a big data architecture stack benchmark. It aims to evaluate big data system across multiple layers of big data architecture, including cloud services, data storage, batch processing, interactive processing, streaming and machine learning. The benchmark supports re-using of existing bi...

**AdBench**
It combines Ad Serving, Streaming Analytics on Ad serving logs, streaming ingestion and updates of various data entities, batch oriented analytics (e.g. 1st Billing), Ad Hoc analytical queries, and Machine learning for Ad targeting. While this benchmark is specific to modern Web or Mobile advertising...

**AI Bench**
AI Bench is an industry standard Internet service AI benchmark suite, designed specifically for modern Internet services with microservice-based architecture. The benchmark spans sixteen AI problem domains from the most widely used Internet service domains: search engine, social network, and e-com...
<table>
<thead>
<tr>
<th>PI-DV</th>
<th>PI-DA</th>
<th>PI-DS</th>
<th>PI-DI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Benchmarks

<table>
<thead>
<tr>
<th>Year</th>
<th>TPC-H</th>
<th>TPC-DS v1</th>
<th>TPC-DS v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Mapping between the Pipeline Steps and the Benchmark Ecosystem

Matrix available in the DataBench ToolBox.
DataBench has devised a generic architectural blueprint mapped to a top level pipeline along the Data Value Chain covering the steps of Data Acquisition/Collection (including data ingestion, processing, streaming, extraction and ingestion storage), Data Preparation/Storage (including storage retrieval/queries, data protection, curation, integration and publication), Data Analytics (including data processing for analysis, AI and Machine Learning) and Data Visualization/Interaction (including data presentation, environment/boundary/user action and interaction).

By clicking on the image below on one of the four steps of the pipeline, or in one of the specific elements from the generic blueprint, this search interface will help you discovery benchmarks and associated knowledge (nuggets).
Perspectives on Architectural pipelines

- Data Acquisition/Collection
  - Batch
  - Real time
  - Interactive

- Data Storage/Preparation
  - Storage Retrieval/Access/Queries
  - Data Protection, Curation, Integration, Publication

- Data Analytics/ML
  - Data processing for analysis, AI and Machine Learning

- Data Visualisation, Action/Interaction
  - Data presentation environment/boundary/user action and interaction

- Processing types
  - Batch, Real time, Interactive

- Data types
  - Batch, Real time, Interactive

- Pipeline steps
  - Benchmarks
  - Tools/Components
  - Standards

- Trans-continuum (Edge to Cloud)
Examples of Horizon 2000 – Big Data PPP projects – with Pipelines and relevant benchmarks

• DataBio
• I-BiDaaS
• TBFY
• Track&Know
• DeepHealth
IoT – Real time Data Pipeline
Example: Banking Experiments Workflow

**Project setup**
- Batch & Streaming Data

**Data Selection**
- Identify data collected by the different sources (ATMs, online banking services, employees’ workstations, external providers’ activity, network devices, etc.), stored in CAIXA datapool

**Data preparation**
- Data Fabrication
- Data Integration
- Data Anonymization
- Data Encryption

**Experiment execution**
- Clustering
- Graph Based Analysis

**Results visualization**
- Static charts for batch analysis results
- Real-time interactive graphics for analysis results of constantly incoming streaming data

---

**Data Acquisition/Collection**
- (including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)

**Data Storage/Preparation**
- (including Storage Retrieval/Access/Queries
- Data Protection, Curation, Integration, Publication)

**Data Analytics/ML**
- (including data processing for analysis, AI and Machine Learning)

**Data Visualisation, Action/Interaction**
- (including data presentation environment/boundary/user action and interaction)

DataBench Pipeline
Benchmark case #1 + #2: Data Ingestion

- Benchmarking
  - Daily log files for each step
  - Identify bottlenecks

Benchmark case #2: Query/visualization

- SPARQL Query Performance
- Which triple store (RDF) database to choose?
- DataBench ToolBox
  - LDBC Semantic Publishing

© by the TBFY-consortium, 2020
Track & Know

- Big Data Analytics
- ML based components
- Scalable
- Online Processing
- Offline Processing
- Versatile
- Easy to integrate (Kafka Topics)

Data Collection / Collection
Live data streaming from devices (GPS/Sensors)

Data Storage / Preparation
Cleansing, enrichment, storage in NoSQL stores

Data Analytics / ML
Pattern recognition
Location forecasting
Clustering
Mobility networks

Data Visualization Action / Interaction
Visual analytics
End user GUI
The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825111.