Building the DataBench Workflow and Architecture

Todor Ivanov (todor@dbis.cs.uni-frankfurt.de), Timo Eichhorn, Arne Jørgen Berre, Tomas Pariente Lobo, Ivan Martinez Rodríguez, Ricardo Ruiz Saiz, Barbara Pernici, Chiara Francalanci

2019 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench'19)
Denver, Colorado, USA
Nov 14-16, 2019
Agenda

1. Project Overview
2. DataBench Workflow
3. DataBench Architecture
4. Next Steps
Evidence Based Big Data Benchmarking to Improve Business Performance

Building a bridge between technical and business benchmarking

Mapping and assessing technical benchmarks

Evaluating business performance and benchmarks

Develop a Benchmarking Toolbox and Handbook
DataBench Project

DataBench (Project ID: 780966) is a three year EU H2020 project (started in January 2018) 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.

Project Outcomes:

• **DataBench Framework** - Including a complete set of metric for BDT assessment.

• **Multiple Analysis** - Assessing the European and industrial significance of the BDT examined by the project.

• **DataBench Toolbox** - A tool to connect and evaluate external benchmarks.

• **DataBench Handbook** - Providing guidelines to the use of the project’s results, Framework & Toolbox, describing metrics implementation and benchmarks.
Benchmark Providers

I’m a benchmark provider and my team has developed a benchmark for Big Data, that I would like to list in the DataBench Toolbox

Benchmark-Specific Features

- Benchmark Type: N/A
- Execution Environment: Nothing selected
- Benchmark references: Nothing selected
- Input Format: Nothing selected
- Output Format: N/A
- Benchmarking aspect: Nothing selected
- Benchmark Data Type: N/A
- Benchmark application-performance metrics: Nothing selected

Benefits

- Your benchmark will be accessible in the Toolbox catalog
- People can discover, access, consult and execute your Benchmark easily
- With the automation procedures for integration, Toolbox users’ can run the benchmark directly and obtain useful information about performance and business insights

More details in video: https://www.youtube.com/watch?v=VLw7GGE673Y

11/14/2019
Today I’m wearing a new hat and I would like to search for a benchmark to test specific big data tools, apps, or machine learning methods. I will try the DataBench Toolbox to see which is the best one suiting my needs.

Select Benchmark

Benchmarks integrated:
- Hilibench
- Yahoo! Cloud Serving Benchmark (YCSB)
- Yahoo Streaming Benchmark

Benchmarks not integrated:
- SparkBench
- Sanou
- Social Network Benchmark
- BigBench V2
- PigMix
- WapDiv
- BigDataLbench
- TPC-H

Benefits

- One-stop-shop for benchmarks
- Automated deployment and execution of many of the existing benchmarks
- Obtain performance metrics
- Compare performance metrics and results

More details in video: https://www.youtube.com/watch?v=cKxA_Oyl180
Business Users

I just changed my hat for the last time, to be a more business-oriented person interested in getting Business Insights out of the Big Data Benchmarks integrated into the DataBench Toolbox.

Benefits

- Navigate and get a plethora of knowledge around technical and business benchmarks in a specific sector.
- Obtain performance metrics and compare with others.
- Get business insights and recommendations about Big Data apps, tools, Artificial Intelligence Methods, among many others.

More details in video: https://www.youtube.com/watch?v=1hZnQ40YWZI
DataBench Toolbox - General Overview

**Toolbox for Benchmark providers**
- Big Data Benchmark Registration/update
  - Benchmark registration process, metadata and filters
- Integrating Big Data Benchmark
  - Benchmark registration of deployment & execution process
- Business Benchmark Samples Registration
  - Registration of business benchmark and examples

**Toolbox for end users**
- **Search**
  - Deployment
  - Execution
  - Selection
  - Getting results
  - Recommendation
  - Displaying results
  - Displaying Tech. Metrics
  - Displaying comparatives

**Toolbox for developers**
- Recommendation
- Displaying comparatives

**Toolbox for business users**
- Recommendation
- Best practices
- Business Insights
- Displaying comparatives
The New Business Benchmark Samples Registration captures domain and industry specific best practices and blueprints associated to concrete business key performance indicators (KPIs).
The New Big Data Benchmark Registration captures the **necessary meta-data** and **features about technical benchmarks** to enable the search and recommendation processes, and to enable the **automation of the deployment** and the **interpretation of the results of the execution** of the benchmarks (Integrating new Big Data Benchmark component).
Benchmark Meta-data
Registering a benchmark in the Toolbox
The Search and Recommendation System shows the steps to define the search criteria (technical, business, application or platform features) as well as associated material (blueprints, best practices in sectors, etc.), a user could pose to the system with the aim to select a benchmark that suits their needs.
The Toolbox Alpha version is already available

- Classified around **65 benchmarks** developed between 1999 and 2018!
- **More than 30** are already searchable in the Toolbox!

### Searchable
- HiBench
- SparkBench
- YCSB
- TPCx-IoT
- Yahoo Streaming Benchmark
- BigBench V2
- TPC-H
- TPC-DS
- Hadoop Workload Examples
- PigMix
- Social Network Benchmark
- WatDiv
- Sanzu
- BigDataBench
- CLASS Benchmark

### Integrated & Runnable
- HiBench
- YCSB
- Yahoo! Streaming
- TPCx-BB (in progress)
- CLASS (in progress)
## Initial benchmarks to be integrated in the Toolbox

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Data Type</th>
<th>Functionality</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiBench</td>
<td>Microbenchmark. ML, SQL, Websearch, Graph, Streaming Benchmarks</td>
<td>Structured, Text, Web Graph</td>
<td>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 <em>micro</em>, <em>machine learning</em>, <em>sql</em>, <em>graph</em>, <em>websearch</em> and <em>streaming</em>.</td>
<td>done</td>
</tr>
<tr>
<td>SparkBench</td>
<td>Microbenchmark. ML, Graph Computation, SQL, Streaming</td>
<td>Structured, Text, Web Graph</td>
<td>System for benchmarking and simulating <em>Spark jobs</em>. Multiple workloads organized in 4 categories.</td>
<td>In progress</td>
</tr>
<tr>
<td>YCSB</td>
<td>Microbenchmark. Cloud OLTP operations</td>
<td>Structured</td>
<td>Evaluates performance of different “<em>key-value</em>” and “<em>cloud</em>” <em>serving systems</em>, 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.</td>
<td>done: Arango, Mongo, Orient, Redis</td>
</tr>
<tr>
<td>TPCx-IoT</td>
<td>Microbenchmark. Workloads on typical IoT Gateway systems</td>
<td>Structured, IoT</td>
<td>Based on YCSB. Workloads of data ingestion and concurrent queries simulating workloads on typical <em>IoT Gateway systems</em>. Dataset with data from sensors from electric power station(s)</td>
<td>In progress</td>
</tr>
<tr>
<td>Yahoo Streaming Benchmark</td>
<td>Appl. benchmarks. Ad analytics pipeline</td>
<td>Structured, Time Series</td>
<td>The Yahoo Streaming Benchmark is a <em>streaming application benchmark</em> simulating an <em>advertisement analytics</em> pipeline.</td>
<td>Integrated, parametrization</td>
</tr>
<tr>
<td>BigBench V1 &amp; V2 / TPCx-BB</td>
<td>Appl. benchmark. Fictional product retailer platform</td>
<td>Structured, Text, JSON logs</td>
<td>End-to-end, technology agnostic, <em>application-level</em> benchmark that tests the <em>analytical capabilities</em> of a Big Data platform. It is based on a fictional product retailer business model.</td>
<td>In progress</td>
</tr>
</tbody>
</table>
The DataBench Toolbox setup represents the process of deploying and enabling the execution either in cloud or in-premise of the selected benchmark. This could only happen if the registration of that benchmark provided the necessary recipes to allow the deployment.
Configure benchmark parameters and execute the benchmark!
The validation of the metrics (in development) allows in certain cases the matching of the technical metrics with business insights or key performance indicators (KPIs). The matching process will be realized with the help of Knowledge Graph (KG).
Monitoring and Evaluation (in development) is realized by gathering of multiple metrics and internal component information that are used to monitor the DataBench framework and analyze the different user behavior.
Platform Metrics Dashboard (Static Metrics)
- number of (active) users
- number of implemented benchmarks
- number of benchmark runs
- number of platform environments
- more ...

User (Profile) Metrics Dashboard
- number of benchmark searches
- number of executed benchmarks
- number of submitted benchmark results
- history log of all operations performed by the user in the last 30 days
- more ...

Administrator Metrics Dashboard
- monitor both Platform and User Metrics
- end-to-end platform analysis on the utilization of the platform
  - Single Ease Question, rate of successful tasks, Resource utilization of hosting platform, etc.
- discover patterns and trends in the user searches and most executed operations
- more ...

Platform Metrics Dashboard

- Number of active user (last 30 days)
- Number of implemented benchmarks
- Uptime of DataBench toolbox
- Number of platform environments
- Number of benchmark executions

DataBench Project - GA Nr 780966
1. **Web Interface** connects to the backend of the Toolbox and provides the different users with the functionality to choose which benchmarks they want to run and configure.

2. **Benchmark Framework Interface** module is the main point of interaction for the administrator with the Benchmarking Framework. They are in charge of handling the *integration, addition* and *deletion of the new, updated or modified benchmarks*.

3. **Results Interface** enables the transfer of benchmark results to the framework either automatically by the benchmark run or manually by the user.
4. **Results Parser** converts the benchmark results into standardized data model to enable calculation of the business metrics in the next steps.

5. **Metrics Spawner** connects to the **Results DB** module, so that it can parse the corresponding results from the technical data model and calculate the defined KPIs and at the end, write them back to the **Results DB**.

6. **Results DB** stores persistently the metric data provided by the **Result Parser**.

7. **Metrics DB** is very similar to the Results DB module with the difference that it stores persistently the collected monitoring metrics.

8. **Metrics Dashboards** offer the monitoring and evaluation functionality of the DataBench framework.
1. **Bootstrap**: the GUI of the Alpha version has been developed using the Bootstrap framework.

2. **Play!-Framework** is the backend framework used to implement the web functionality.

3. **AWX project** is the upstream open source project of **Ansible Tower**, which allows controlling the automation deployment of software and tools.

4. **Kafka** is used to act as an interface between Ansible and the Results database.

5. **MySQL** stores the parsed benchmark metrics as well as other meta-data.

6. **Log Files** log all the operations and user actions of the Framework.
DataBench ToolBox Development

Alpha version URL:  [http://83.149.125.78:9000/](http://83.149.125.78:9000/)
Contacts

✉️ info@databench.eu

@DataBench_eu

DataBench

DataBench Project

Visit: www.databench.eu