DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

DataBench Toolbox Demo

BDV Meet-Up

27 June 2019, Riga

Tomas Pariente Lobo, Iván Martínez ATOS

IDC
ANALYZE THE FUTURE

Atos

Institut "Jožef Stefan", Ljubljana, Slovenija

SINTEF

POLITECNICO
MILANO 1863

Frankfurt Big Data Lab
GOETHE UNIVERSITÄT

Toobox 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, if possible, integrate **with existing benchmarking initiatives** and resources where possible.

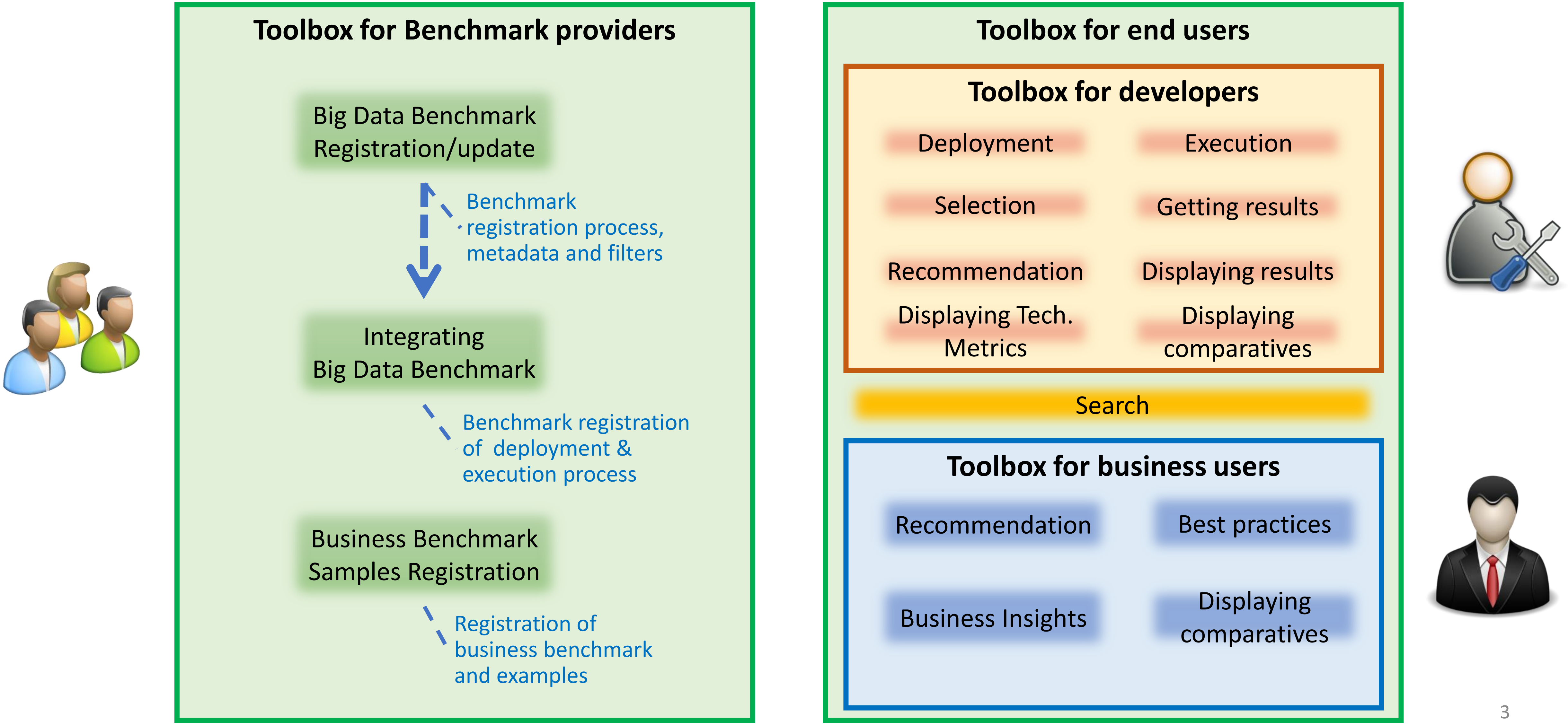
Homogenising metrics

- The Toolbox will implement ways, to emerge **Big Data benchmarking technical metrics and business insights**

Web user interface

- It will include a web-based visualization layer to **assist to the final users to specify their benchmarking requirements** to help them to search, select, deploy, run and getting benchmarks technical results and business insights.

Toolbox usage: 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

Toolbox for developers

Deployment	Execution
Selection	Getting results
Recommendation	Displaying results
Displaying Tech. Metrics	Displaying comparatives

Search

Toolbox for business users

Recommendation	Best practices
Business Insights	Displaying comparatives

Alpha version of the Toolbox already available for Alpha-testers

Searchable

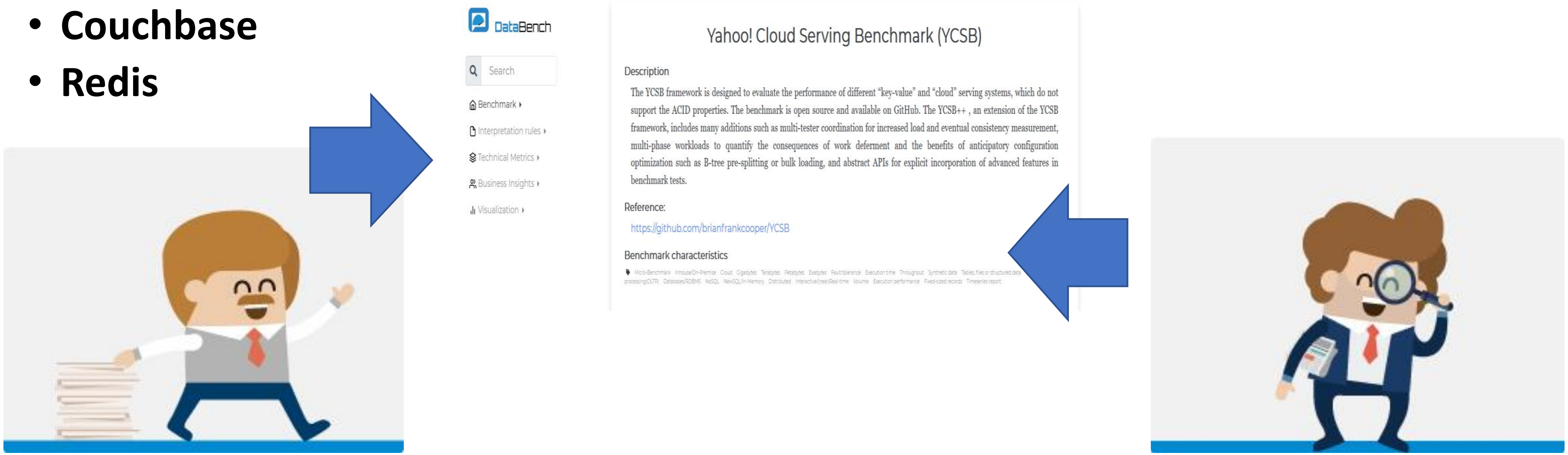
- HiBench
- SparkBench
- YCSB
- TPCx-IoT
- Yahoo Streaming Benchmark
- BigBench V2
- TPC-H
- TPC-DS
- Hadoop Workload Examples
- PigMix
- Social Network Benchmark
- WatDiv
- Sanzu
- BigDataBench
- CLASS Benchmark

Runnable

- HiBench
- YCSB
- Yahoo! streaming
- CLASS (in progress)

CherryData Use Case

- The company needs to benchmark the following low latency databases:
  - **Arango**
  - **Orient**
  - **Couchbase**
  - **Redis**



Yahoo! Cloud Serving Benchmark (YCSB)

Description

The YCSB framework is designed to evaluate the performance of different "key-value" and "cloud" serving systems, which do not support the ACID properties. The benchmark is open source and available on GitHub. The YCSB++ , an extension of the YCSB framework, includes many additions such as multi-tester coordination for increased load and eventual consistency measurement, multi-phase workloads to quantify the consequences of work deferment and the benefits of anticipatory configuration optimization such as B-tree pre-splitting or bulk loading, and abstract APIs for explicit incorporation of advanced features in benchmark tests.

Reference:

https://github.com/brianfrankcooper/YCSB

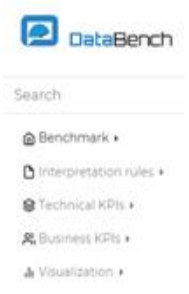
Benchmark characteristics

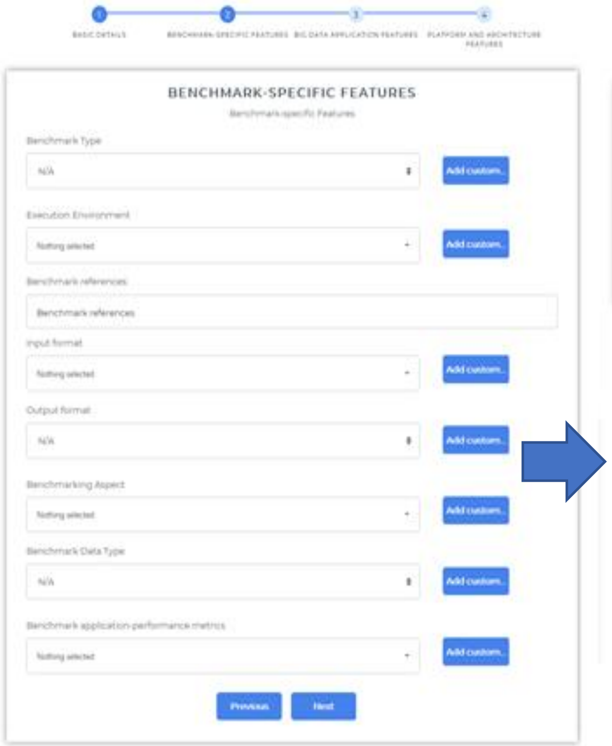
Micro-Benchmark Inhouse/On-Premise Cloud Gigabytes Terabytes Petabytes Exabytes Fault tolerance Execution time Throughput Synthetic data Tables, files or structured data processing(OLTP) Databases/RDBMS NoSQL NewSQL/In-Memory Distributed Interactive/(near)Real-time Volume Execution performance Fixed-sized records Timeseries report

Benchmark Provider

I am going to register into DATABENCH the **YCSB**

Registering a benchmark in the Toolbox




1 BASIC DETAILS
2 BENCHMARK-SPECIFIC FEATURES
3 BIG DATA APPLICATION FEATURES
4 PLATFORM AND ARCHITECTURE FEATURES

BENCHMARK-SPECIFIC FEATURES

Benchmark-specific Features

Benchmark Type: N/A

Execution Environment: Nothing selected

Benchmark references: Benchmark references

Input format: Nothing selected

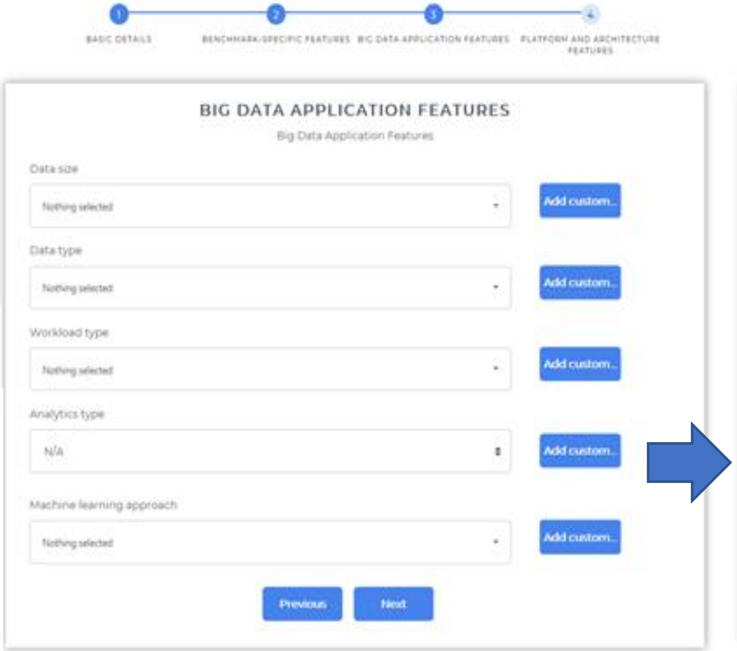
Output format: N/A

Benchmarking Aspect: Nothing selected

Benchmark Data Type: N/A

Benchmark application-performance metrics: Nothing selected

Add custom...
Previous
Next


1 BASIC DETAILS
2 BENCHMARK-SPECIFIC FEATURES
3 BIG DATA APPLICATION FEATURES
4 PLATFORM AND ARCHITECTURE FEATURES

BIG DATA APPLICATION FEATURES

Big Data Application Features

Data size: Nothing selected

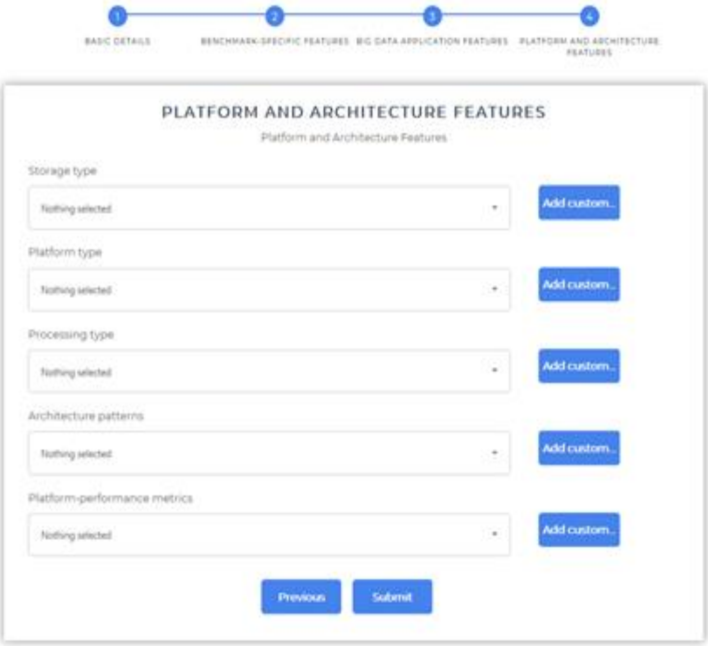
Data type: Nothing selected

Workload type: Nothing selected

Analytics type: N/A

Machine learning approach: Nothing selected

Add custom...
Previous
Next


1 BASIC DETAILS
2 BENCHMARK-SPECIFIC FEATURES
3 BIG DATA APPLICATION FEATURES
4 PLATFORM AND ARCHITECTURE FEATURES

PLATFORM AND ARCHITECTURE FEATURES

Platform and Architecture Features

Storage type: Nothing selected

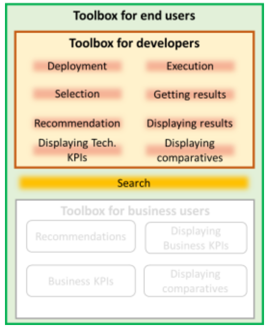
Platform type: Nothing selected

Processing type: Nothing selected

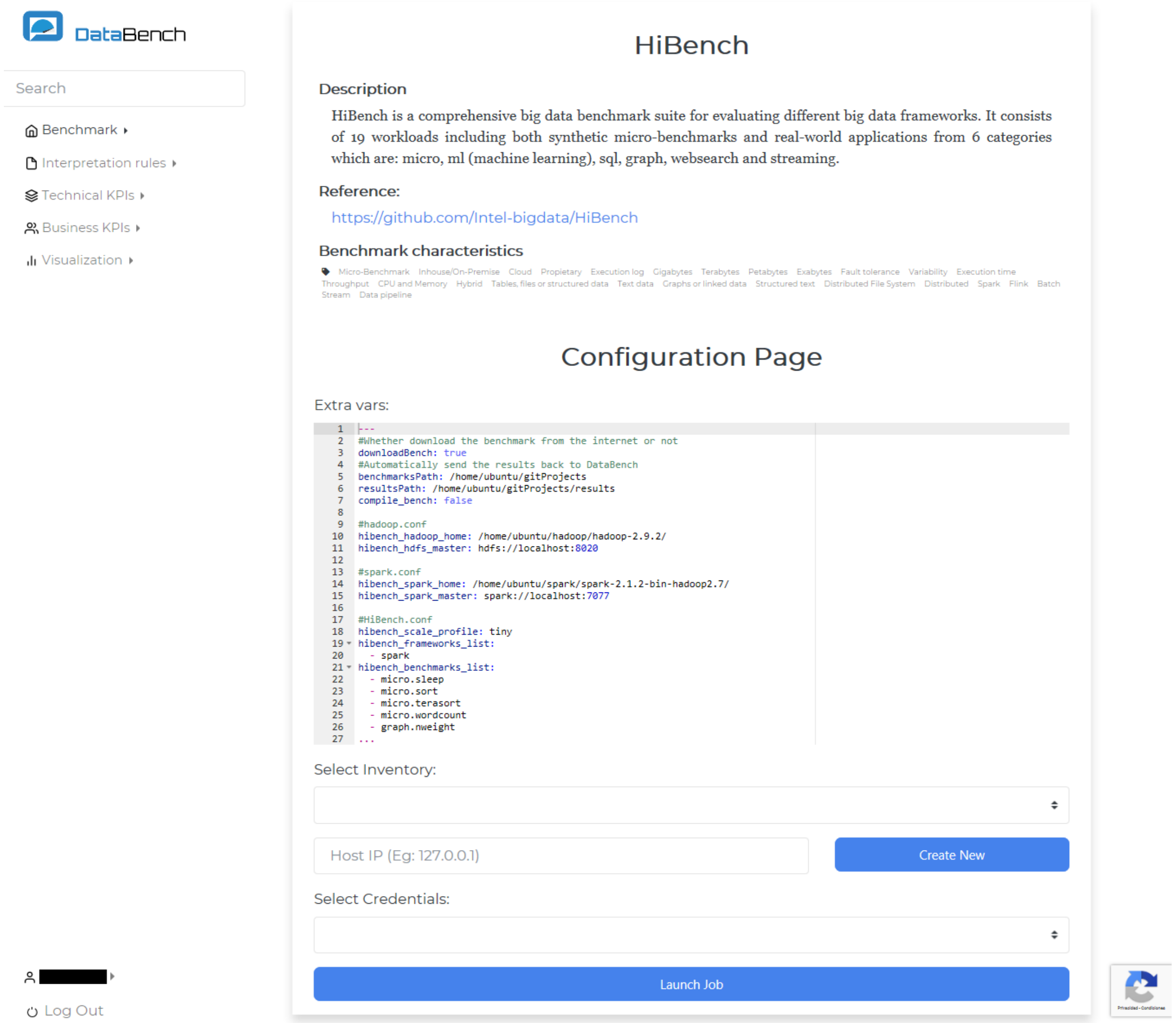
Architecture patterns: Nothing selected

Platform-performance metrics: Nothing selected

Add custom...
Previous
Submit



Adding configuration for benchmark deployment and run



HiBench

Description

HiBench is a comprehensive big data benchmark suite for evaluating different big data frameworks. It consists of 19 workloads including both synthetic micro-benchmarks and real-world applications from 6 categories which are: micro, ml (machine learning), sql, graph, websearch and streaming.

Reference:

https://github.com/Intel-bigdata/HiBench

Benchmark characteristics

Micro-Benchmark Inhouse/On-Premise Cloud Propietary Execution log Gigabytes Terabytes Petabytes Exabytes Fault tolerance Variability Execution time Throughput CPU and Memory Hybrid Tables, files or structured data Text data Graphs or linked data Structured text Distributed File System Distributed Spark Flink Batch Stream Data pipeline

Configuration Page

Extra vars:

```yaml
---
#Whether download the benchmark from the internet or not
downloadBench: true
#Automatically send the results back to DataBench
benchmarksPath: /home/ubuntu/gitProjects
resultsPath: /home/ubuntu/gitProjects/results
compile_bench: false

#hadoop.conf
hibench_hadoop_home: /home/ubuntu/hadoop/hadoop-2.9.2/
hibench_hdfs_master: hdfs://localhost:8020

#spark.conf
hibench_spark_home: /home/ubuntu/spark/spark-2.1.2-bin-hadoop2.7/
hibench_spark_master: spark://localhost:7077

#HiBench.conf
hibench_scale_profile: tiny
hibench_frameworks_list:
  - spark
hibench_benchmarks_list:
  - micro.sleep
  - micro.sort
  - micro.terasort
  - micro.wordcount
  - graph.nweight
...
```

Select Inventory:

Host IP (Eg: 127.0.0.1)

Select Credentials:

Preparing an Ansible Playbook

Steps:

1) Ansible template to be filled by benchmark providers
2) Upload the playbook to Toolbox Git
3) Create a job template in Ansible AWX for that playbook
4) Link the benchmark with the template so it can be run from the platform

Technical End User

I am Paolo Ravanelli, CTO of **Cherrydata**. My company needs to benchmark the following low latency databases:

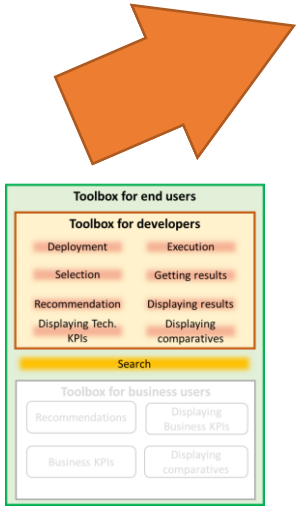
- **Arango**
- **Orient**
- **Couchbase**
- **Redis**

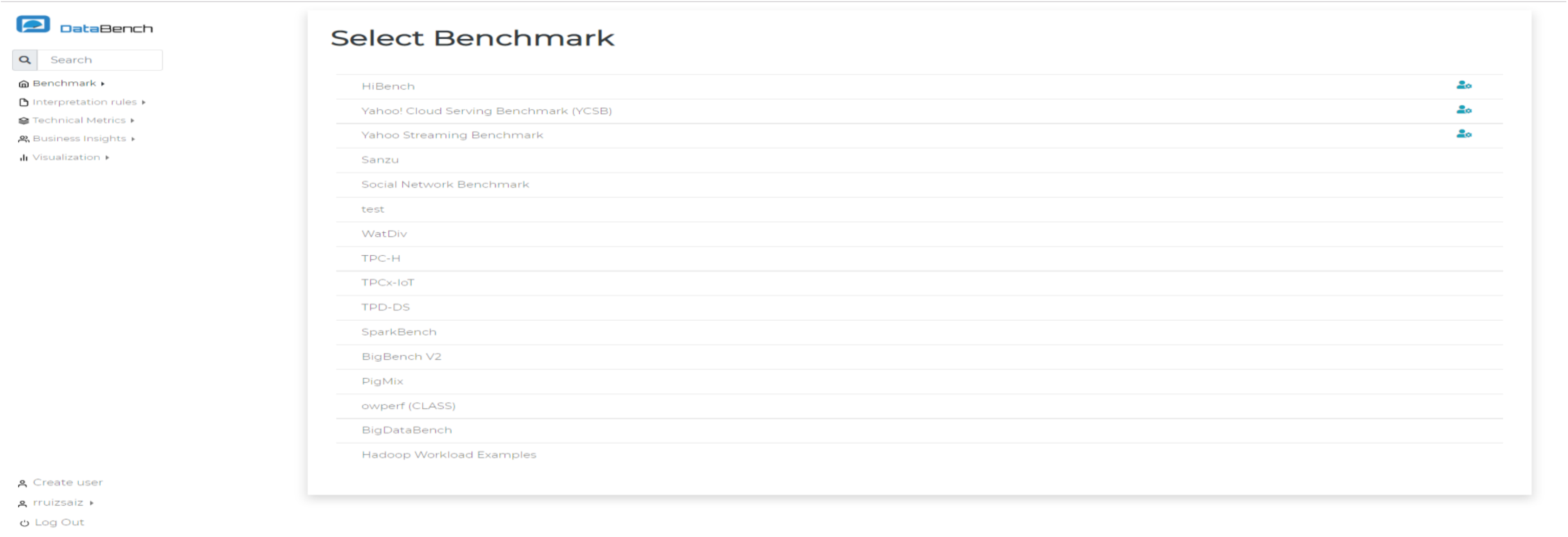
The company is a strong believer in benchmarking and your suggestion to use **YCSB** has already been very useful.

Selecting and executing benchmarks: YCSB

- For benchmarks ready to run:

1. Search and choose the benchmark you want to run from the list
2. Fill in the variables with the data of the target system (i.e. host IPs)
3. Provide credentials to log into the target system (public key)
4. Let the system run the playbook (deployment and running)

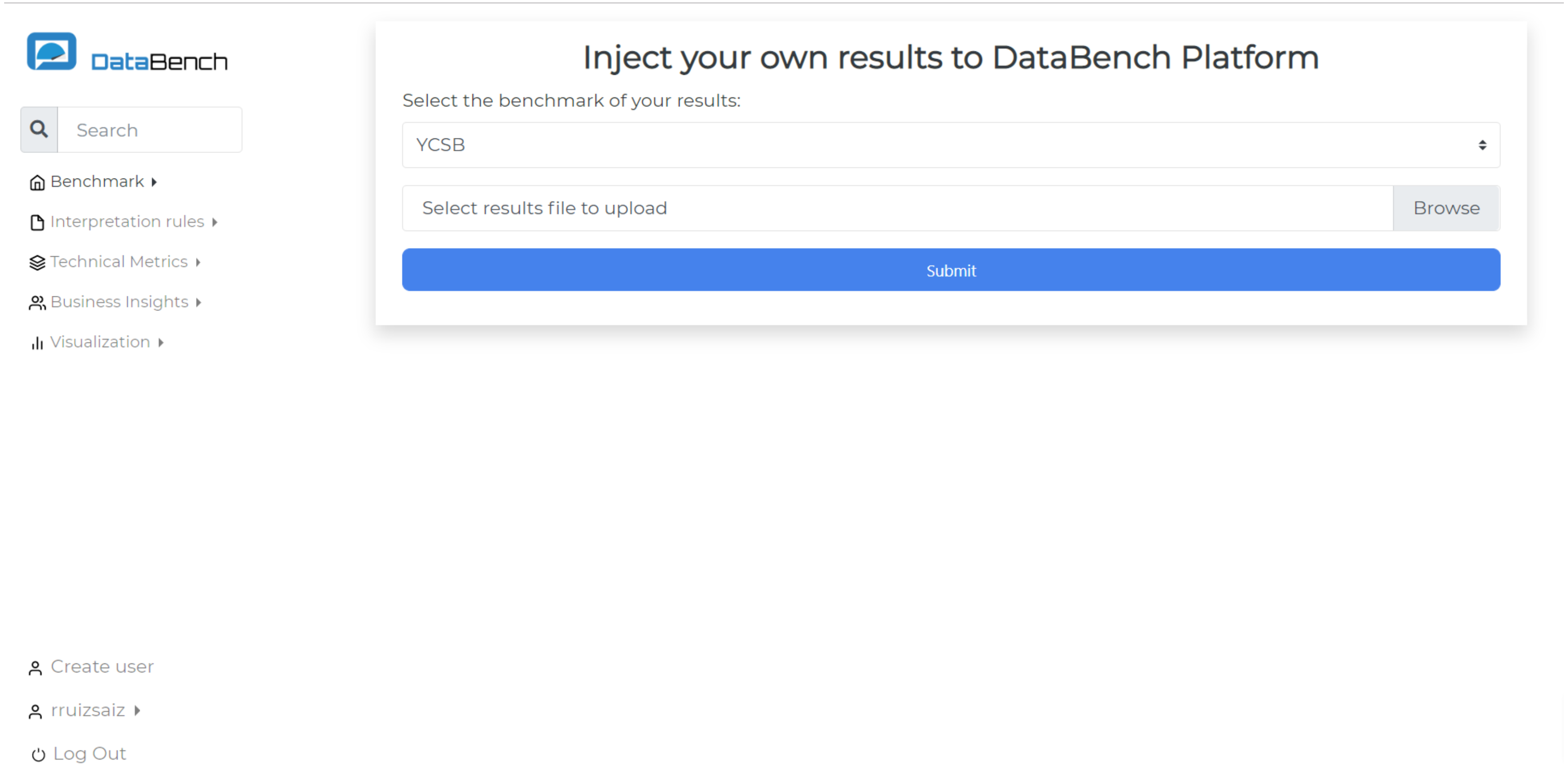




Select Benchmark

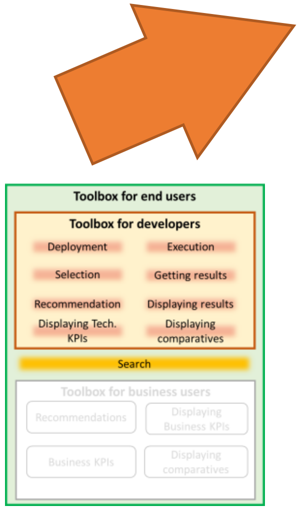
HiBench
Yahoo! Cloud Serving Benchmark (YCSB)
Yahoo Streaming Benchmark
Sanzu
Social Network Benchmark
test
WatDiv
TPC-H
TPCx-IoT
TPD-DS
SparkBench
BigBench V2
PigMix
owperf (CLASS)
BigDataBench
Hadoop Workload Examples

Sharing results after executing YCSB



Inject your own results to DataBench Platform

Select the benchmark of your results: YCSB



Toolbox for end users

Toolbox for developers

Deployment
Execution
Selection
Getting results
Recommendation
Displaying results
Displaying Tech. KPIs
Displaying comparatives

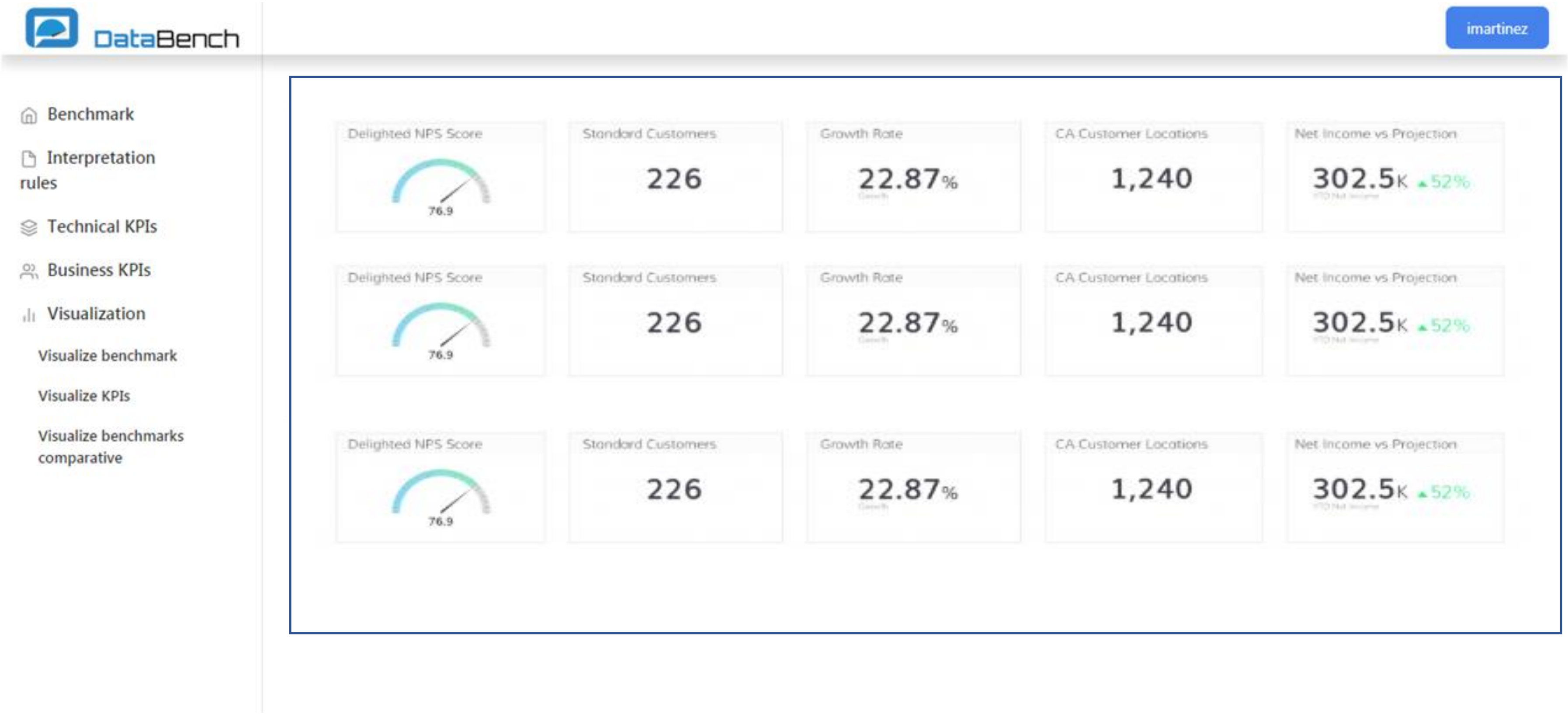
Search

Toolbox for business users

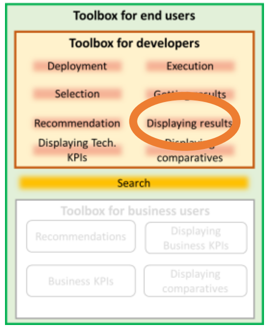
Recommendations
Displaying Business KPIs
Business KPIs
Displaying comparatives

Visualizing Benchmark Results

- Ongoing work (for the Beta version): Investigating visual paradigms to homogenize and show the results of the a given run, comparison with other runs or with other benchmarks…



Delighted NPS Score	Standard Customers	Growth Rate	CA Customer Locations	Net Income vs Projection
76.9	226	22.87%	1,240	302.5K ▲52%
76.9	226	22.87%	1,240	302.5K ▲52%
76.9	226	22.87%	1,240	302.5K ▲52%



Toolbox for end users

Toolbox for developers

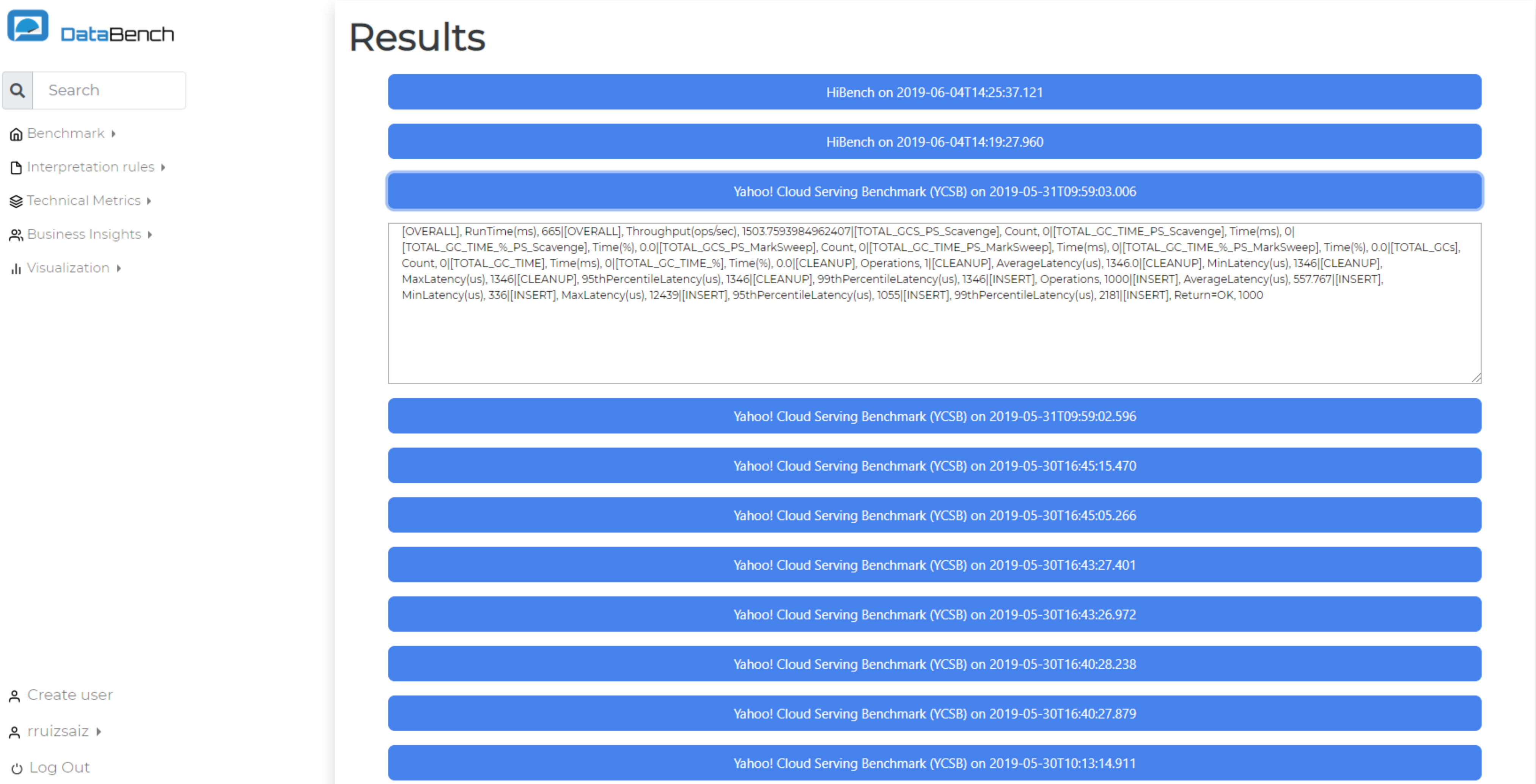
Deployment · Execution · Selection · Getting results · Recommendation · Displaying results · Displaying Tech. KPIs · Displaying comparatives

Search

Toolbox for business users

Recommendations · Displaying Business KPIs · Business KPIs · Displaying comparatives

Visualizing YCSB execution results



Results

HiBench on 2019-06-04T14:25:37.121

HiBench on 2019-06-04T14:19:27.960

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-31T09:59:03.006

```
[OVERALL], RunTime(ms), 665|[OVERALL], Throughput(ops/sec), 1503.7593984962407|[TOTAL_GCS_PS_Scavenge], Count, 0|[TOTAL_GC_TIME_PS_Scavenge], Time(ms), 0|
[TOTAL_GC_TIME_%_PS_Scavenge], Time(%), 0.0|[TOTAL_GCS_PS_MarkSweep], Count, 0|[TOTAL_GC_TIME_PS_MarkSweep], Time(ms), 0|[TOTAL_GC_TIME_%_PS_MarkSweep], Time(%), 0.0|[TOTAL_GCs],
Count, 0|[TOTAL_GC_TIME], Time(ms), 0|[TOTAL_GC_TIME_%], Time(%), 0.0|[CLEANUP], Operations, 1|[CLEANUP], AverageLatency(us), 1346.0|[CLEANUP], MinLatency(us), 1346|[CLEANUP],
MaxLatency(us), 1346|[CLEANUP], 95thPercentileLatency(us), 1346|[CLEANUP], 99thPercentileLatency(us), 1346|[INSERT], Operations, 1000|[INSERT], AverageLatency(us), 557.767|[INSERT],
MinLatency(us), 336|[INSERT], MaxLatency(us), 12439|[INSERT], 95thPercentileLatency(us), 1055|[INSERT], 99thPercentileLatency(us), 2181|[INSERT], Return=OK, 1000
```

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-31T09:59:02.596

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T16:45:15.470

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T16:45:05.266

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T16:43:27.401

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T16:43:26.972

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T16:40:28.238

Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T16:40:27.879

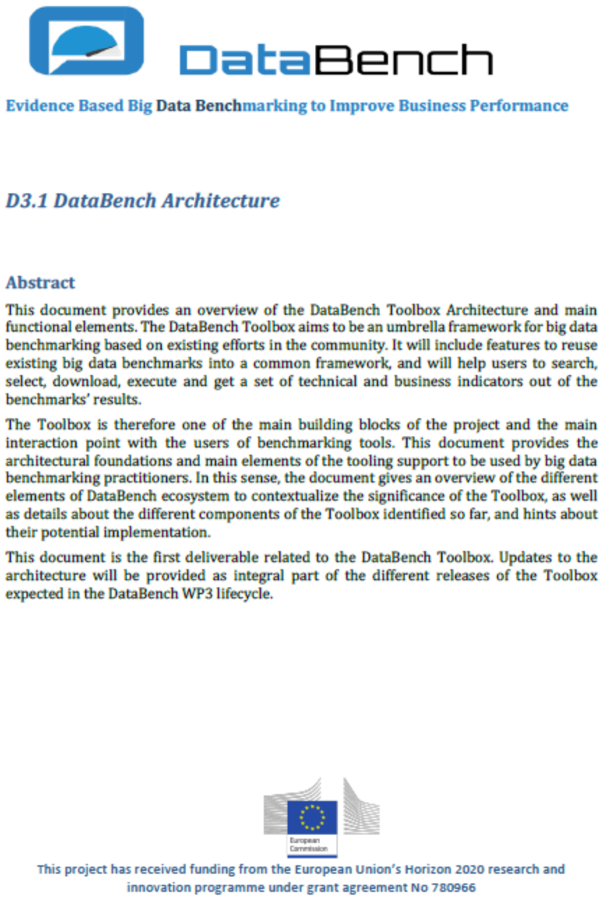
Yahoo! Cloud Serving Benchmark (YCSB) on 2019-05-30T10:13:14.911

Summary

- Next Toolbox releases
  - Beta Toolbox by December 2019
  - Final release by June 2020
- Generation of a Benchmarking Knowledge Graph supporting technical and business aspects
- Find relations between technical metrics and business insights based in use cases

More info

- Check our website: https://www.databench.eu/



Evidence Based Big Data Benchmarking to Improve Business Performance

D3.1 DataBench Architecture

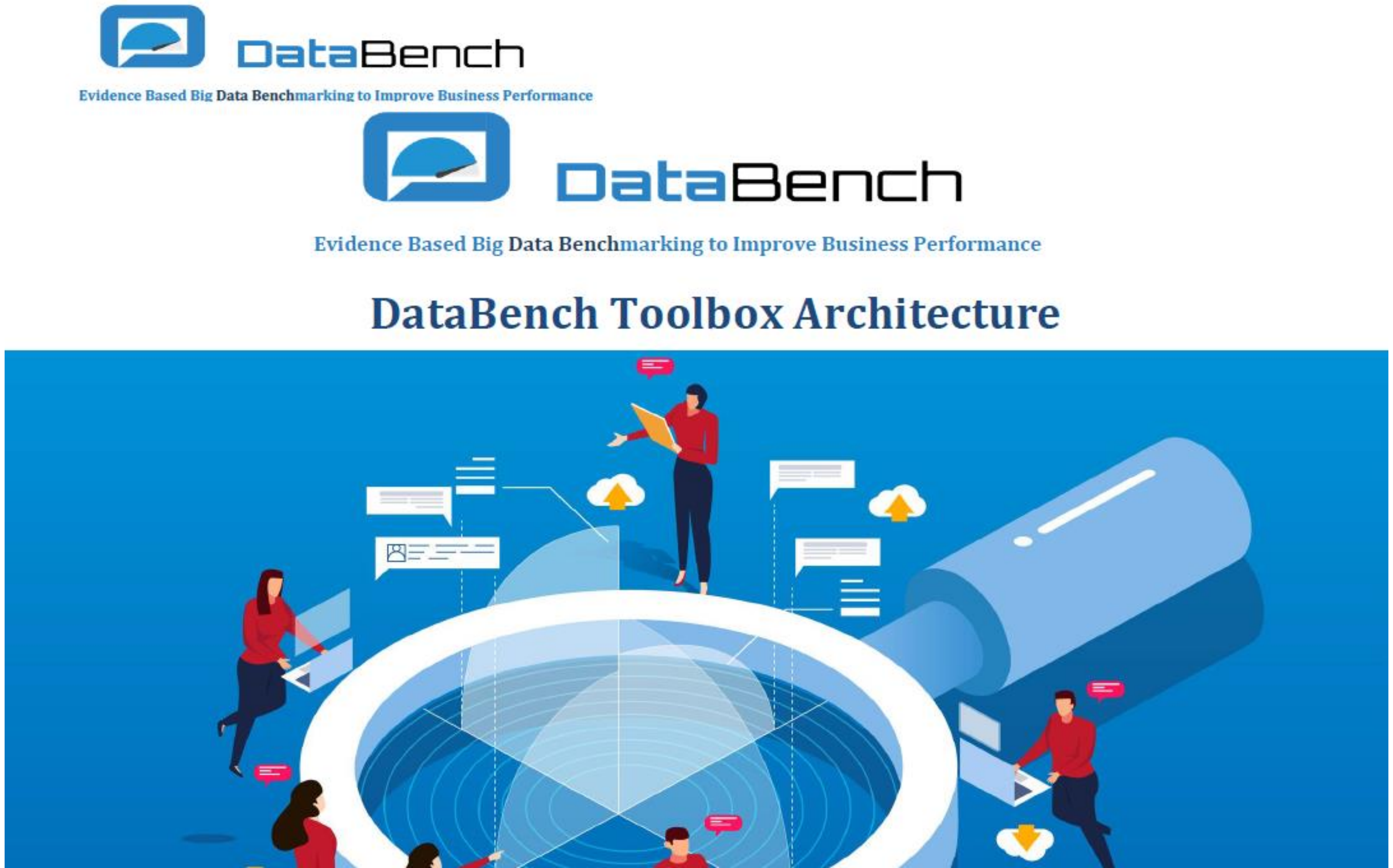
Abstract

This document provides an overview of the DataBench Toolbox Architecture and main functional elements. The DataBench Toolbox aims to be an umbrella framework for big data benchmarking based on existing efforts in the community. It will include features to reuse existing big data benchmarks into a common framework, and will help users to search, select, download, execute and get a set of technical and business indicators out of the benchmarks' results.

The Toolbox is therefore one of the main building blocks of the project and the main interaction point with the users of benchmarking tools. This document provides the architectural foundations and main elements of the tooling support to be used by big data benchmarking practitioners. In this sense, the document gives an overview of the different elements of DataBench ecosystem to contextualize the significance of the Toolbox, as well as details about the different components of the Toolbox identified so far, and hints about their potential implementation.

This document is the first deliverable related to the DataBench Toolbox. Updates to the architecture will be provided as integral part of the different releases of the Toolbox expected in the DataBench WP3 lifecycle.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780966



Evidence Based Big Data Benchmarking to Improve Business Performance

Evidence Based Big Data Benchmarking to Improve Business Performance

DataBench Toolbox Architecture

Contacts

info@databench.eu

@DataBench_eu

DataBench

DataBench Project

DataBench

DataBench Project

DataBench

Evidence Based Big Data Benchmarking to
Improve Business Performance

tomas.parientelobo@atos.net

ivan.martinez@atos.net