



Evidence Based Big Data Benchmarking to Improve Business Performance

Benchmarking Big Data

20/06/2019, Open Expo Europe, Madrid

Tomas Pariente Lobo, ATOS



Atos



POLITECNICO
MILANO 1863

Frankfurt Big Data Lab
GOETHE  UNIVERSITÄT

Agenda

1. Big Data Benchmarking: Introduction and Motivation
2. Big Data Technical and Business Benchmarking
3. DataBench Toolbox
4. Next Steps



Big Data Benchmarking: Introduction and Motivation

Benchmarking

- The Term Benchmark:
 - *A benchmark is a measured „best-in-class“ achievement recognised as the standard of excellence for that business process. (APQC 1993)*
- Two main types of benchmarks:
 - **Business Performance Benchmarking** – comparison of performance measures for the purpose of determining how good one’s own company is compared to others.
 - **A software benchmark** is a program used for comparison of software products/tools executing on a pre-configured hardware environment.

Example Use Case

- Which system will have **best price/performance** for my application?
➔ **Need to use a benchmark**
- What is more important the **minimal execution time or lower total cost**?
➔ **Depends on the business KPIs**
- The company decide to introduce Machine Learning model to improve product recommendations to customers. Different ML models have different Accuracy. How important is the accuracy? Should the company invest in improving the model accuracy?
➔ **Business decision**

System A	System B
4 Nodes (servers)	30 Nodes (servers)
4 hours (execution time)	4 minutes (execution time)
5500 Euro (total cost)	50 000 Euro (total cost)

DataBench

One stop shop for Big Data Benchmarking



2

Big Data Technical and Business
Benchmarking

Building a bridge between Big Data Technical and Business Benchmarking

Main activities

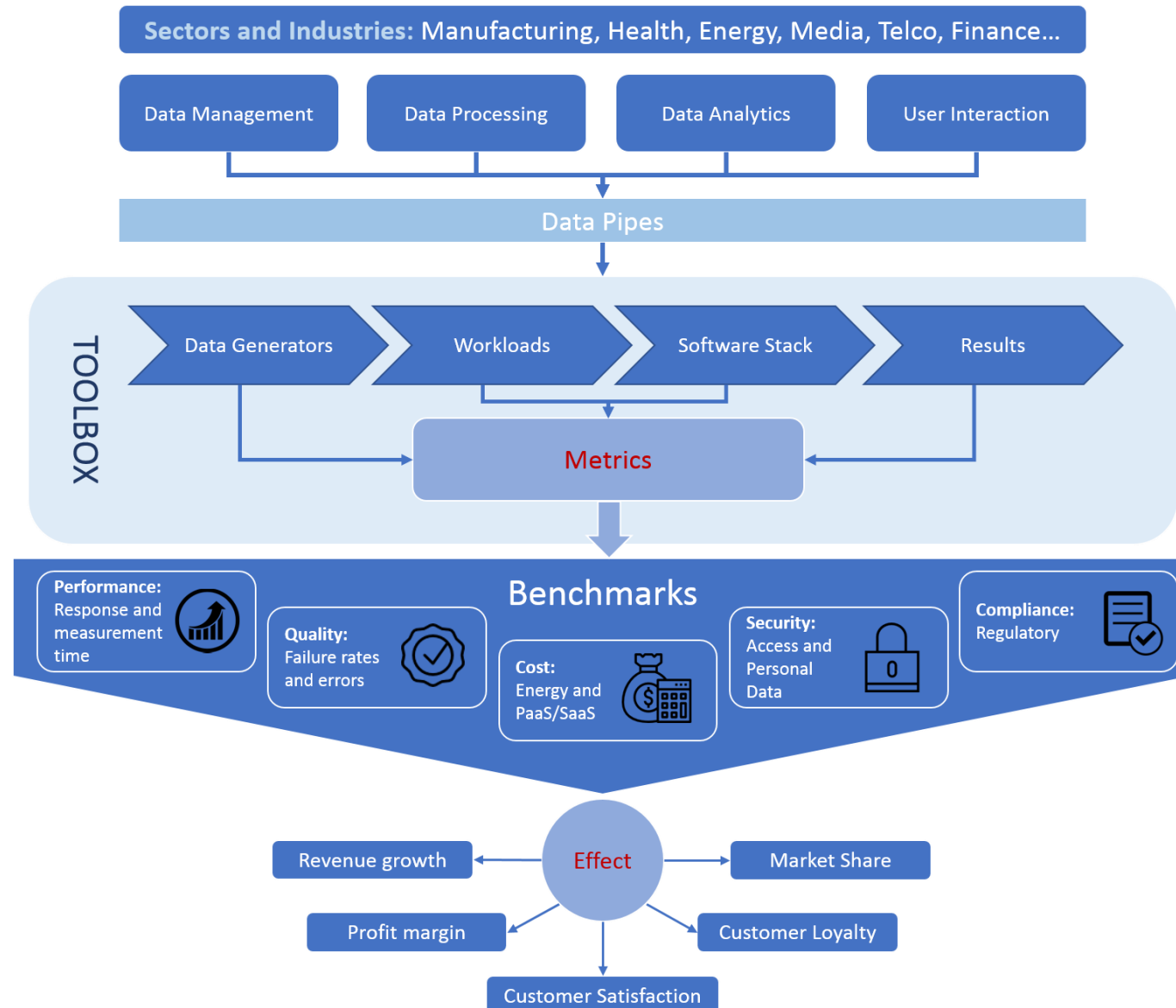
- Classify the main use cases of BDT by industry
- Compile and assess technical benchmarks
- Perform economic and market analysis to assess industrial needs
- Evaluate business performance in selected use cases



Expected Results

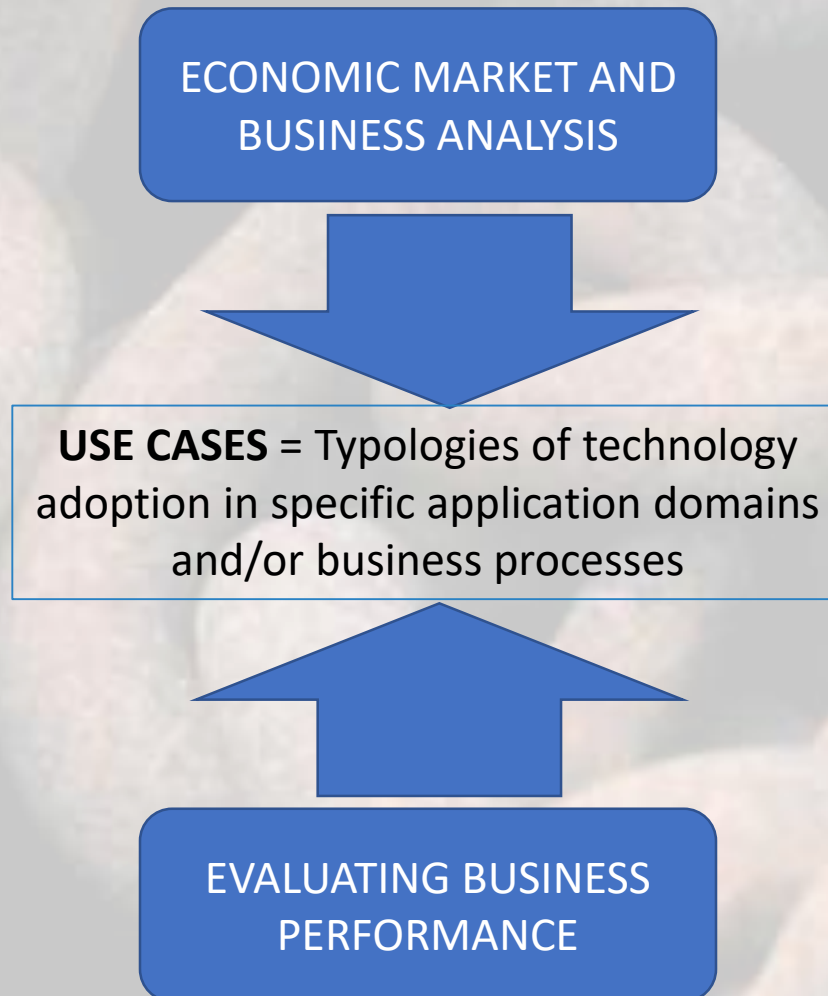
- A conceptual framework linking technical and business benchmarks
- European industrial and performance benchmarks
- A Toolbox measuring optimal benchmarking approaches
- A handbook to guide the use of benchmarks

Technical and Business Benchmarking Framework



How to link Technical and Business Benchmarking

- **Focus on** economic and industry analysis and the **EU Big Data market**
- Classify leading **Big Data technologies use cases by industry**
- **Analyse industrial users benchmarking needs** and assess their relative **importance for EU economy** and the main industries
- Demonstrate the scalability, **European significance** (high potential **economic impact**) and **industrial relevance** (responding to primary needs of users) of the benchmarks



- Focus on **data collection and identification of use cases** to be monitored and measured
- Evaluation of **business performance of specific Big Data initiatives**
- Leverage **DataBench Toolbox**
- Provide specific industrial benchmarks
- Produce the **DataBench Handbook**, a manual supporting the application of the DataBench Toolbox

Big Data Business Benchmarking Tools

- Questionnaires to stakeholders
- Studies per sector
- Self-assessment tool
- Handbook
- Use cases insights

 DataBench
Evidence Based Big Data Benchmarking to Improve Business Performance

Measuring the impact of Big Data and Analytics on business results

SELF-ASSESSMENT REPORT

Name

Organization

Project

In this customized report your answers to the DataBench Big Data survey are compared with respondents from your same industry and company size class to help you compare your Big Data business KPIs with those of your peers. This will help you gain inspiration and insights about how best to implement data-driven innovation.

3

DataBench Toolbox

Toolbox 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.

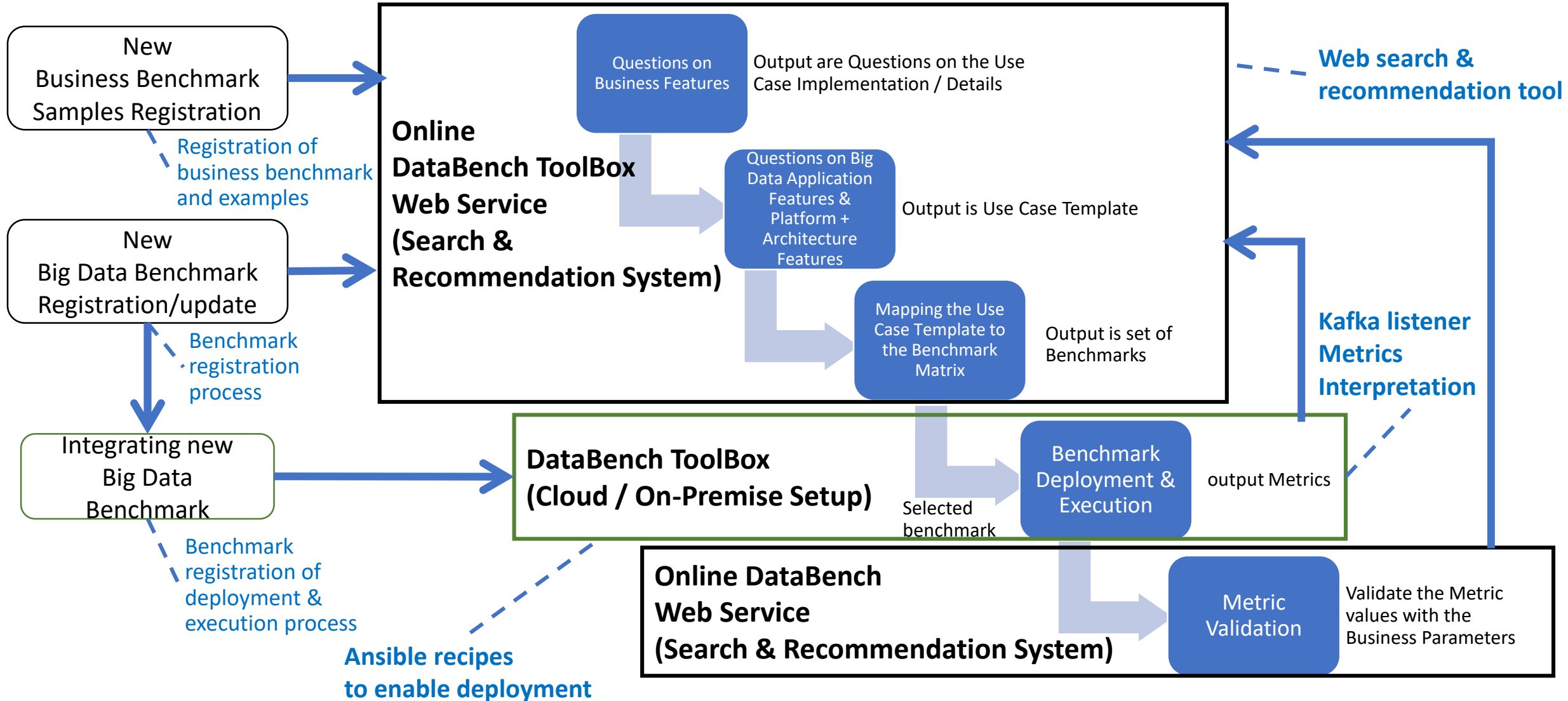
Homogenising metrics

- 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.

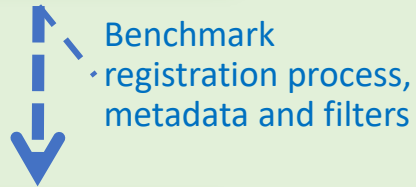
Methodology Workflow & Implementation



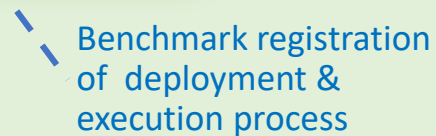
Toolbox usage: General Overview

Toolbox for Benchmark providers

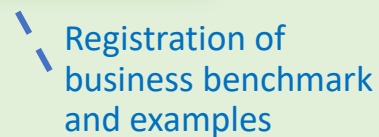
Big Data Benchmark
Registration/update



Integrating
Big Data Benchmark



Business Benchmark
Samples Registration



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

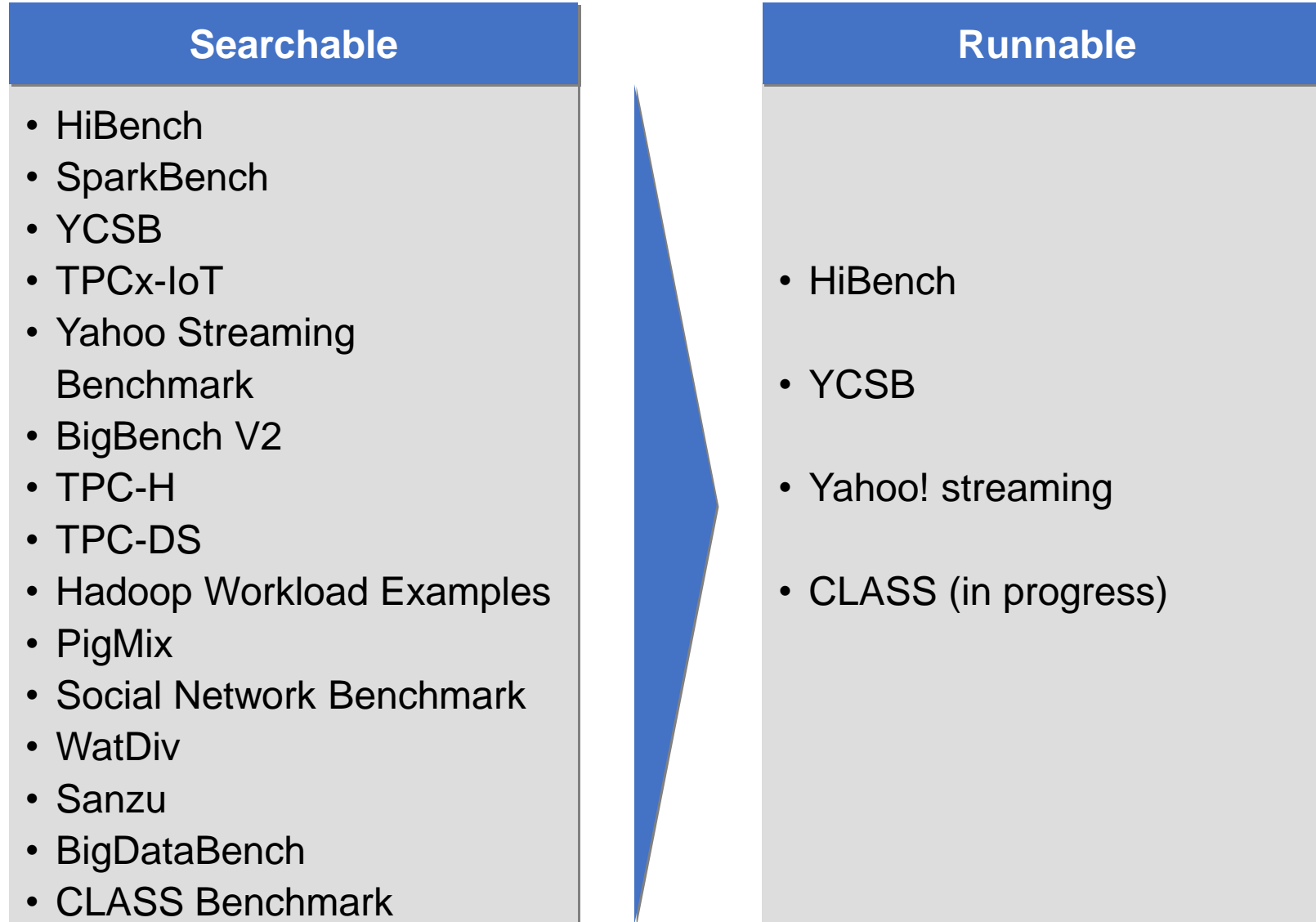
Best practices

Business Insights

Displaying
comparatives



Alpha version of the Toolbox already available for Alpha-testers



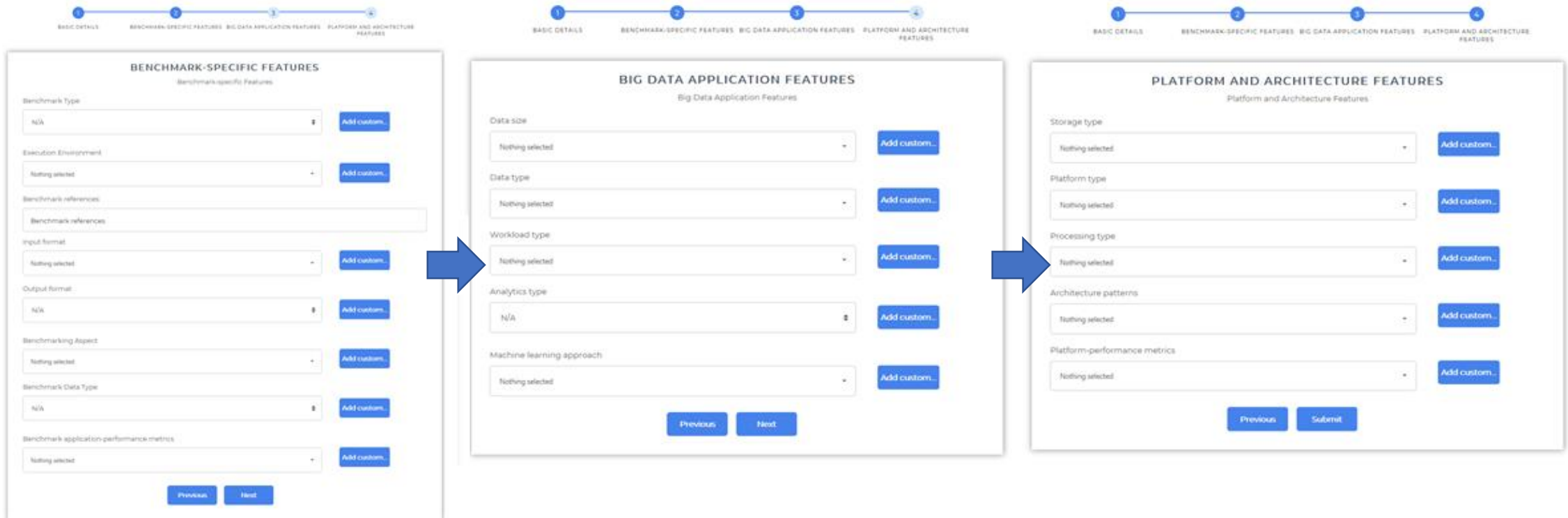
Registering a Benchmark in the Toolbox

Search

- Benchmark
- Interpretation rules
- Technical KPIs
- Business KPIs
- Visualization

h






+



The registration process is shown in three sequential steps, each with a progress indicator at the top:

- Step 1: BENCHMARK-SPECIFIC FEATURES**
 - Benchmark Type: N/A (Add custom...)
 - Execution Environment: Nothing selected (Add custom...)
 - Benchmark references: Benchmark references (Add custom...)
 - Input format: Nothing selected (Add custom...)
 - Output format: N/A (Add custom...)
 - Benchmarking Aspect: Nothing selected (Add custom...)
 - Benchmark Data Type: N/A (Add custom...)
 - Benchmark application performance metrics: Nothing selected (Add custom...)
 - Buttons: Previous, Next
- Step 2: BIG DATA APPLICATION FEATURES**
 - Data size: Nothing selected (Add custom...)
 - Data type: Nothing selected (Add custom...)
 - Workload type: Nothing selected (Add custom...)
 - Analytics type: N/A (Add custom...)
 - Machine learning approach: Nothing selected (Add custom...)
 - Buttons: Previous, Next
- Step 3: PLATFORM AND ARCHITECTURE FEATURES**
 - Storage type: Nothing selected (Add custom...)
 - Platform type: Nothing selected (Add custom...)
 - Processing type: Nothing selected (Add custom...)
 - Architecture patterns: Nothing selected (Add custom...)
 - Platform-performance metrics: Nothing selected (Add custom...)
 - Buttons: Previous, Submit

Adding configuration for Benchmark deployment and run

-  Benchmark ▶
-  Interpretation rules ▶
-  Technical KPIs ▶
-  Business KPIs ▶
-  Visualization ▶

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
  Proprietary
  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:

```

1  |--
2  #Whether download the benchmark from the internet or not
3  downloadBench: true
4  #Automatically send the results back to DataBench
5  benchmarksPath: /home/ubuntu/gitProjects
6  resultsPath: /home/ubuntu/gitProjects/results
7  compile_bench: false
8
9  #hadoop.conf
10 hibench_hadoop_home: /home/ubuntu/hadoop/hadoop-2.9.2/
11 hibench_hdfs_master: hdfs://localhost:8020
12
13 #spark.conf
14 hibench_spark_home: /home/ubuntu/spark/spark-2.1.2-bin-hadoop2.7/
15 hibench_spark_master: spark://localhost:7077
16
17 #HiBench.conf
18 hibench_scale_profile: tiny
19 hibench_frameworks_list:
20   - spark
21 hibench_benchmarks_list:
22   - micro.sleep
23   - micro.sort
24   - micro.teasort
25   - micro.wordcount
26   - graph.nweight
27   ...
    
```

Select Inventory:

Host IP (Eg: 127.0.0.1)

Create New

Select Credentials:

Launch Job



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 users: Executing Benchmarks

- Deployment in-house
 1. Download from Git the DataBench Toolbox
 2. Fill in the variables files with the data of the target system
 3. Run the playbook.

Variables:

```

---
#Hibench Specific variables
compile_bench: false

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

```

Playbook running:

```

rrulz@rrulz-VirtualBox:~/gitProjects/DataBench/ansiblePlaybooks$ ansible-playbook -i variables.yml Benchmark_HiBench.yml
PLAY [local] *****
TASK [Gathering Facts] *****
ok: [local]

TASK [Ensures ../results/HiBench/ dir exists] *****
ok: [local]

PLAY [edge] *****
TASK [Gathering Facts] *****
ok: [edge]

TASK [git] *****
changed: [edge]

TASK [BC is needed by HiBench to create the report file, checking it is installed] ***
ok: [edge]

TASK [Compile HiBench Benchmark suite] *****
skipping: [edge]

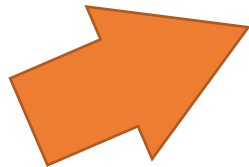
TASK [debug] *****
ok: [edge] => {
  "hibench_compile_stdout_lines": "VARIABLE IS NOT DEFINED!"
}

TASK [Create the hadoop config file with the selected options] *****
ok: [edge]

TASK [Create the spark config file with the selected options] *****
ok: [edge]

TASK [Create the hibench config file with the selected options] *****

```



Toolbox for end users

Toolbox for developers

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

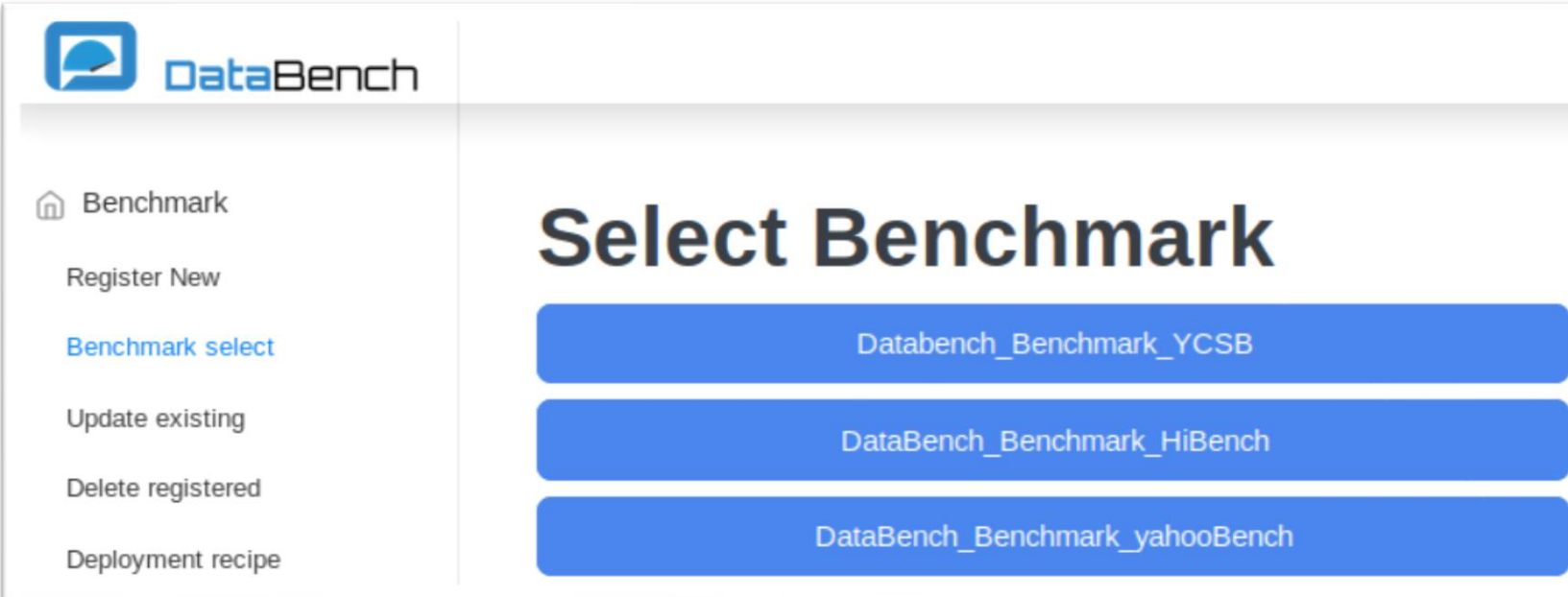
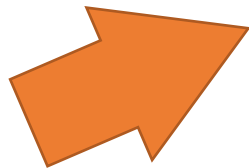
Search

Toolbox for business users

- Recommendations
- Displaying Business KPIs
- Business KPIs
- Displaying comparatives

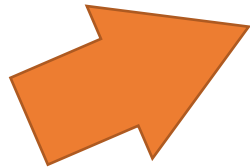
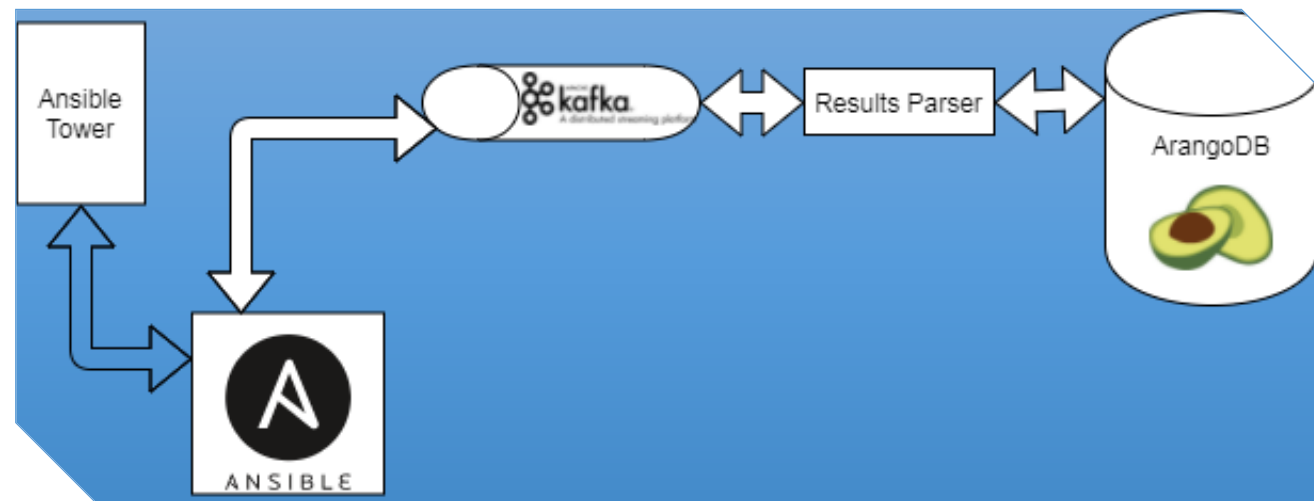
Technical users: Selecting and executing Bench.(II)

- 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)



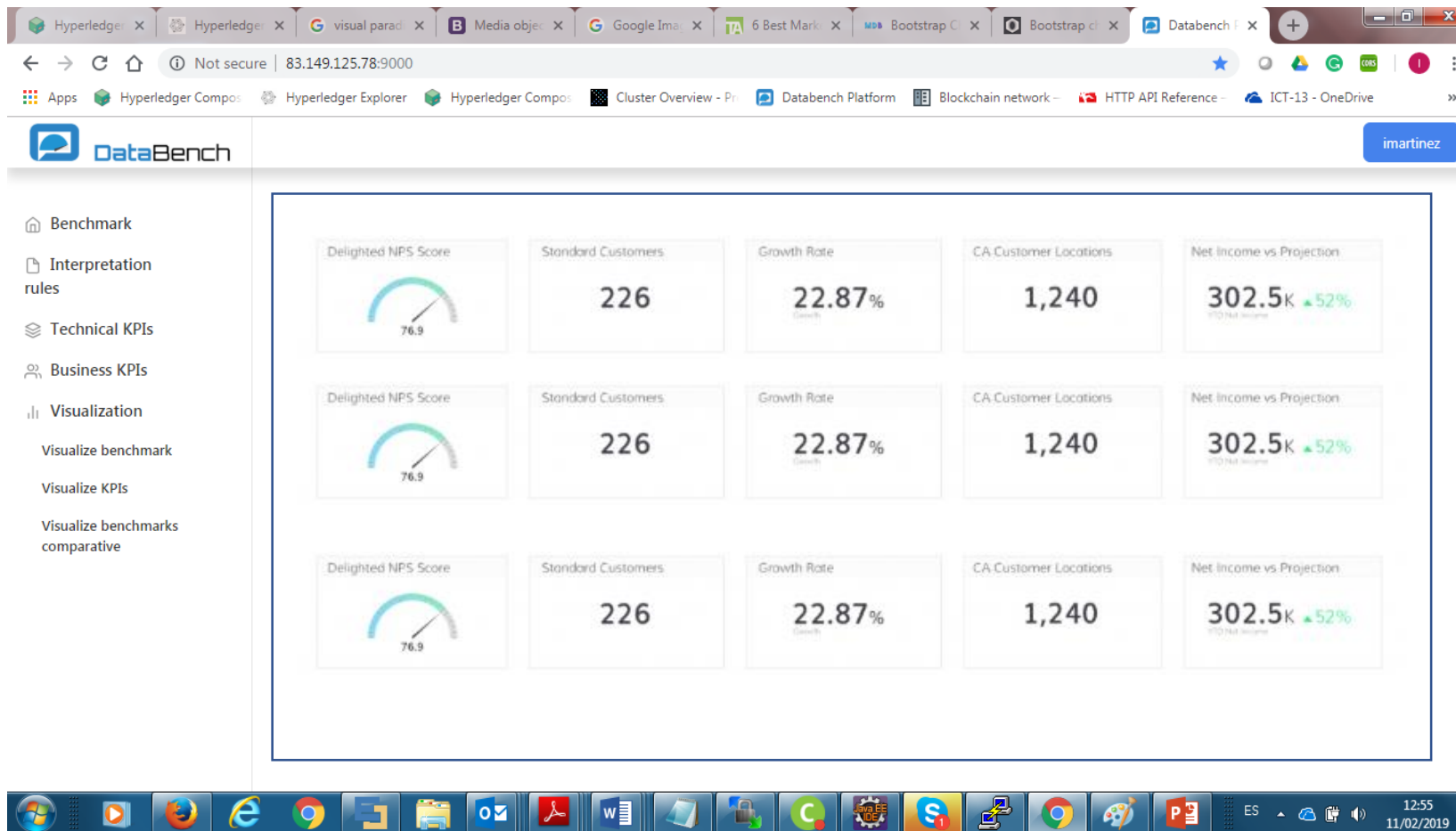
Technical users: Gathering Results

- Execution In-House
 - Connection back to the Toolbox to get results from benchmarks runs to our Kafka instance
 - Methodology and Ansible-playbook template script to easily adapt to new benchmarks
- Execution in the DataBench platform
 - When run in the DataBench based platform, the results are sent back automatically
 - Same script as In-House



Technical users: 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...



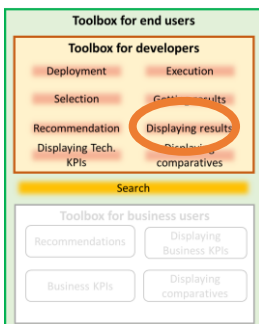
The screenshot shows a web browser window with the DataBench application. The browser tabs include 'Hyperledger', 'visual paradig', 'Media obje', 'Google Ima', '6 Best Mark', 'Bootstrap C', and 'Databench F'. The address bar shows 'Not secure | 83.149.125.78:9000'. The application header features the DataBench logo and a user profile 'imartinez'.

The main content area displays a grid of benchmark results for five KPIs: Delighted NPS Score, Standard Customers, Growth Rate, CA Customer Locations, and Net Income vs Projection. Each KPI is represented by a card with a gauge or numerical value and a trend indicator.

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

The left sidebar contains navigation options: Benchmark, Interpretation rules, Technical KPIs, Business KPIs, Visualization, Visualize benchmark, Visualize KPIs, and Visualize benchmarks comparative. An orange arrow points from the 'Visualization' section to the main content area.

The bottom taskbar shows various application icons including Windows, Edge, Chrome, File Explorer, Outlook, Adobe Reader, Word, and PowerPoint. The system tray shows the time as 12:55 on 11/02/2019.



The diagram shows a 'Toolbox for end users' with three main sections:

- Toolbox for developers:** Includes Deployment, Execution, Selection, Getting results, Recommendation, and Displaying results (circled in red). Below this is a 'Search' bar.
- Toolbox for business users:** Includes Recommendations and Displaying Business KPIs.
- Bottom section:** Includes Business KPIs and Displaying comparatives.



Next Steps

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
- Enlarge the community and sustainability of the tools

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.pariantelobo@atos.net

@tpariente