Building the DataBench Workflow and Architecture

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

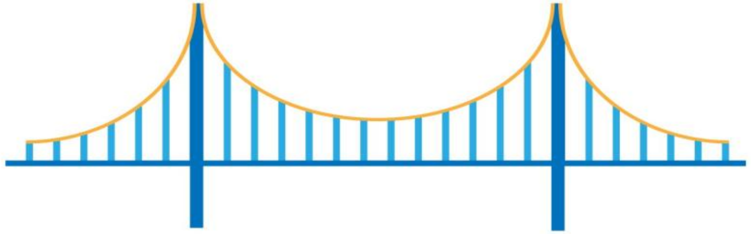
2019 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench'19)
Denver, Colorado, USA
Nov 14-16, 2019

Agenda

1. Project Overview
2. DataBench Workflow
3. DataBench Architecture
4. Next Steps

Building a bridge between technical and business benchmarking

Mapping and assessing technical benchmarks



Evaluating business performance and benchmarks

Develop a Benchmarking Toolbox and Handbook

IDC ANALYZE THE FUTURE POLITECNICO MILANO 1863 SINTEF AtoS IJS Frankfurt Big Data Lab GOETHE UNIVERSITÄT

DataBench Project

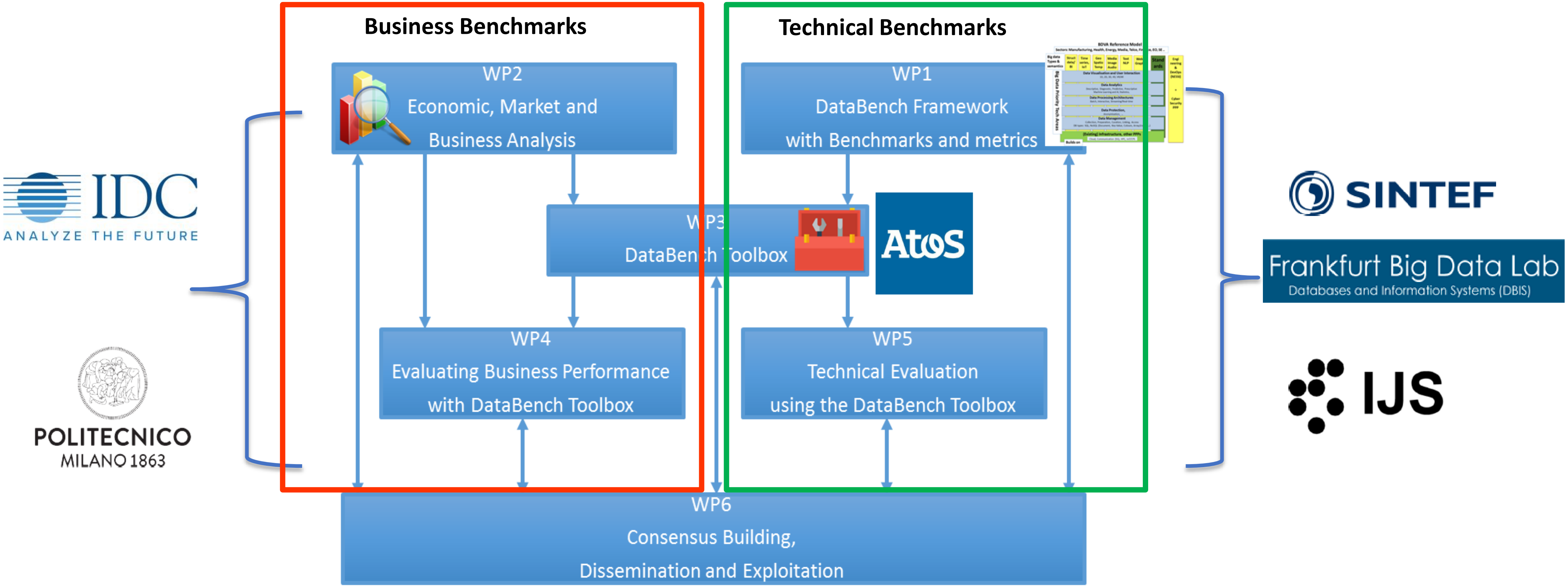
DataBench (Project ID: 780966) is a three year EU H2020 project (started in January 2018) that ***investigates existing Big Data benchmarking tools and projects***, identifies the main gaps and provides a ***robust set of metrics to compare technical results coming from those tools.***

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.

DataBench Work Packages

DataBench



Business Benchmarks

WP2
Economic, Market and
Business Analysis

WP4
Evaluating Business Performance
with DataBench Toolbox

Technical Benchmarks

WP1
DataBench Framework
with Benchmarks and metrics

WP3
DataBench Toolbox

WP5
Technical Evaluation
using the DataBench Toolbox

WP6
Consensus Building,
Dissemination and Exploitation

IDC ANALYZE THE FUTURE

POLITECNICO MILANO 1863

Atos

SINTEF

Frankfurt Big Data Lab
Databases and Information Systems (DBIS)

IJS

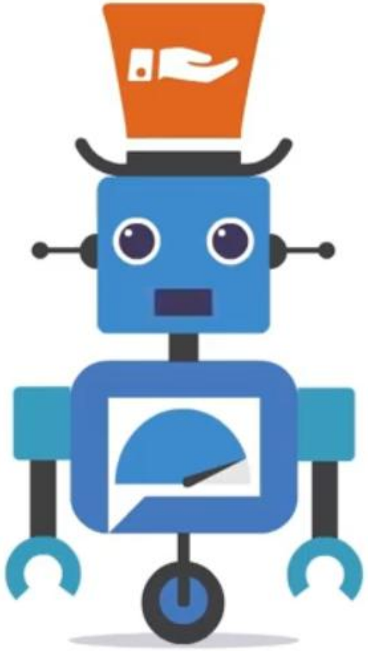
IDC ANALYZE THE FUTURE

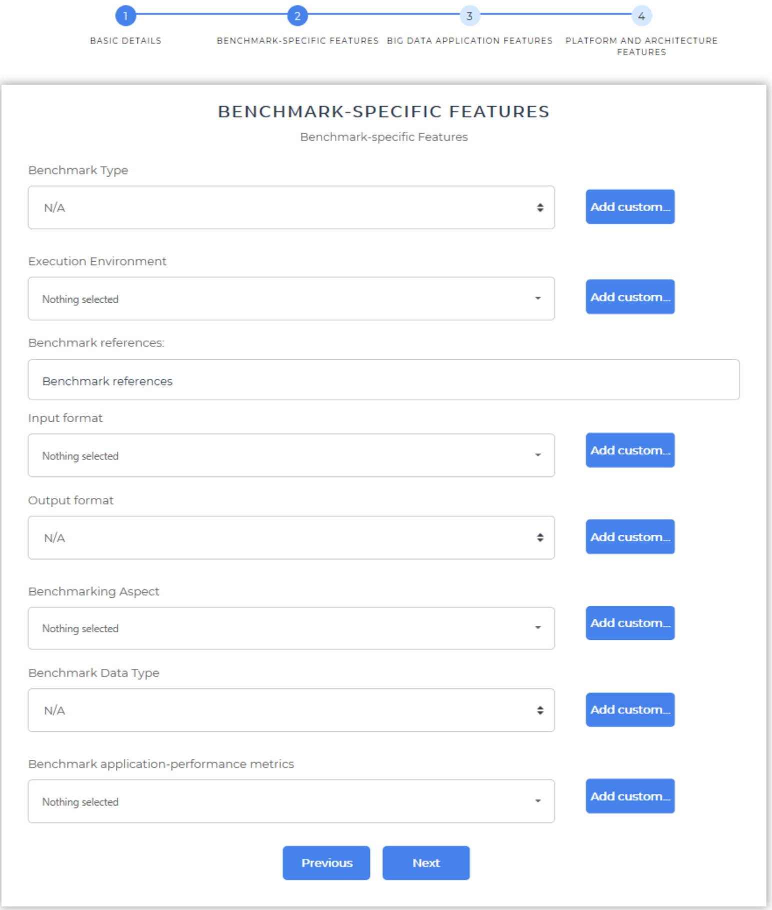
Atos

SINTEF

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





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

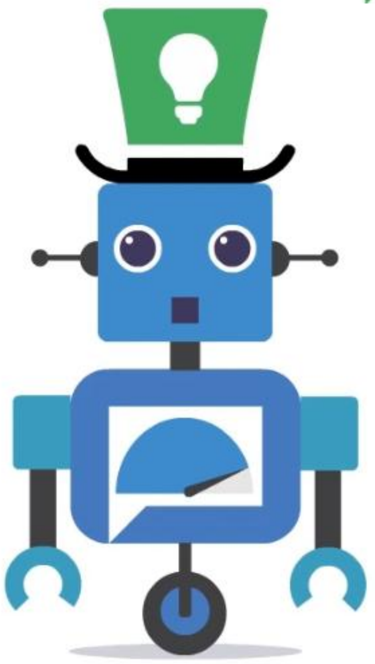
Benefits

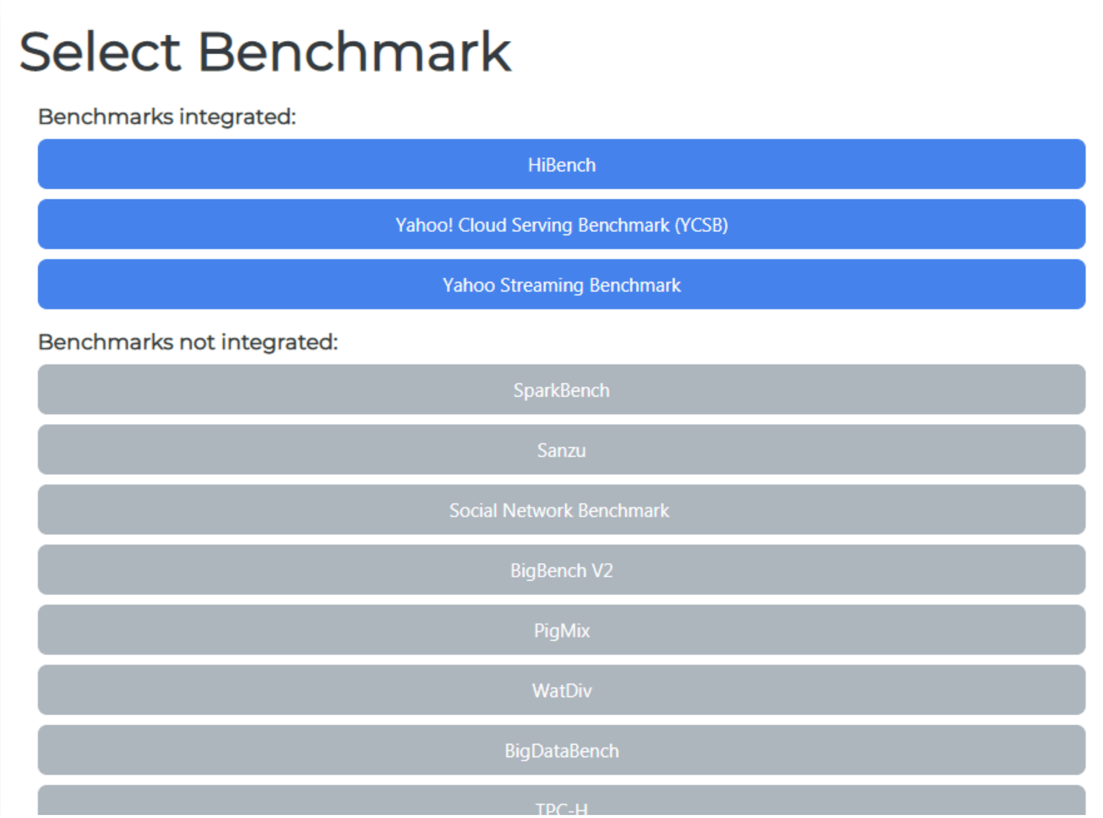
- ✓ Your benchmark will be accessible in the Toolbox catalog
- ✓ People can discover, access, consult and execute your Benchmark easily
- ✓ With the automation procedures for integration, Toolbox users' can run the benchmark directly and obtain useful information about performance and business insights

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

Technical Users

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:

- HiBench
- Yahoo! Cloud Serving Benchmark (YCSB)
- Yahoo Streaming Benchmark

Benchmarks not integrated:

- SparkBench
- Sanzu
- Social Network Benchmark
- BigBench V2
- PigMix
- WatDiv
- BigDataBench
- TPC-H

Benefits

✓ One-stop-shop for benchmarks

✓ Automated deployment and execution of many of the existing benchmarks

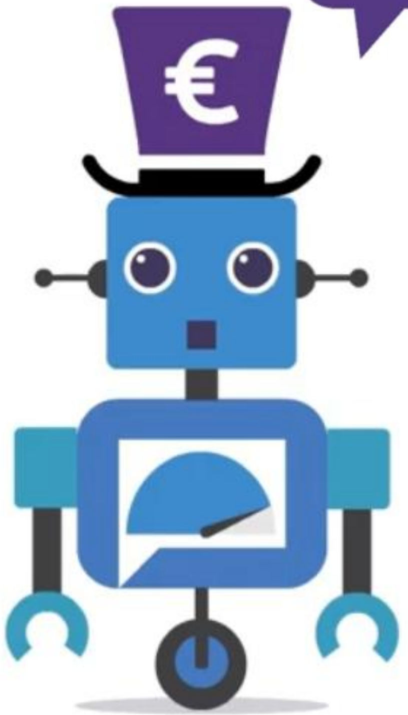
✓ Obtain performance metrics

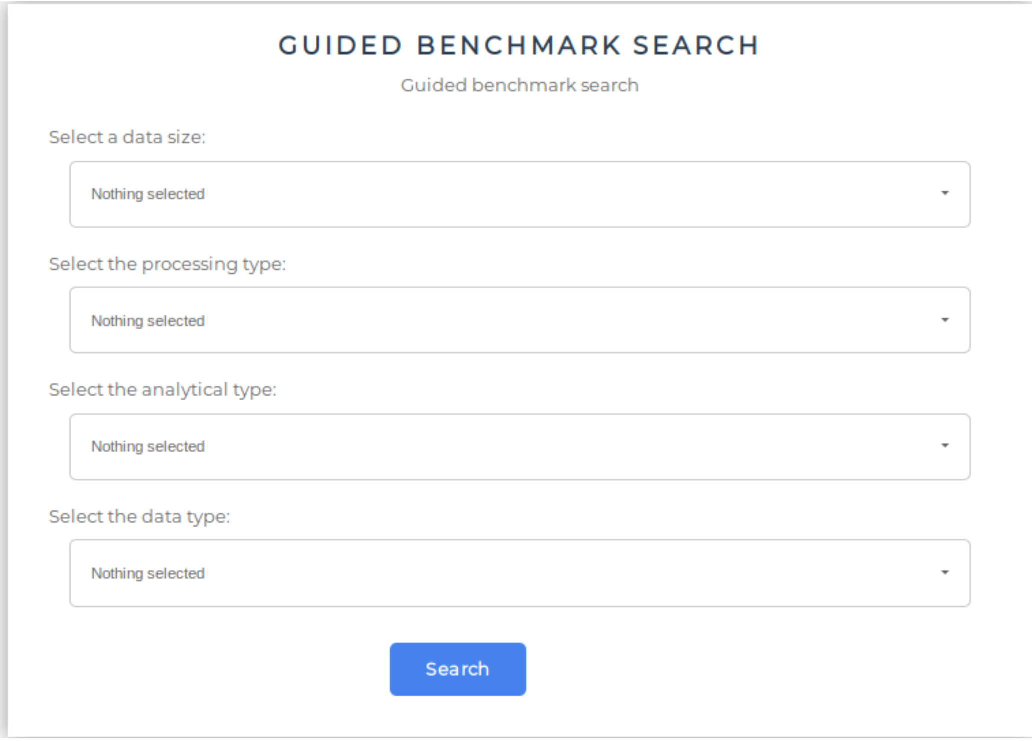
✓ Compare performance metrics and results

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





GUIDED BENCHMARK SEARCH

Guided benchmark search

Select a data size:

Nothing selected

Select the processing type:

Nothing selected

Select the analytical type:

Nothing selected

Select the data type:

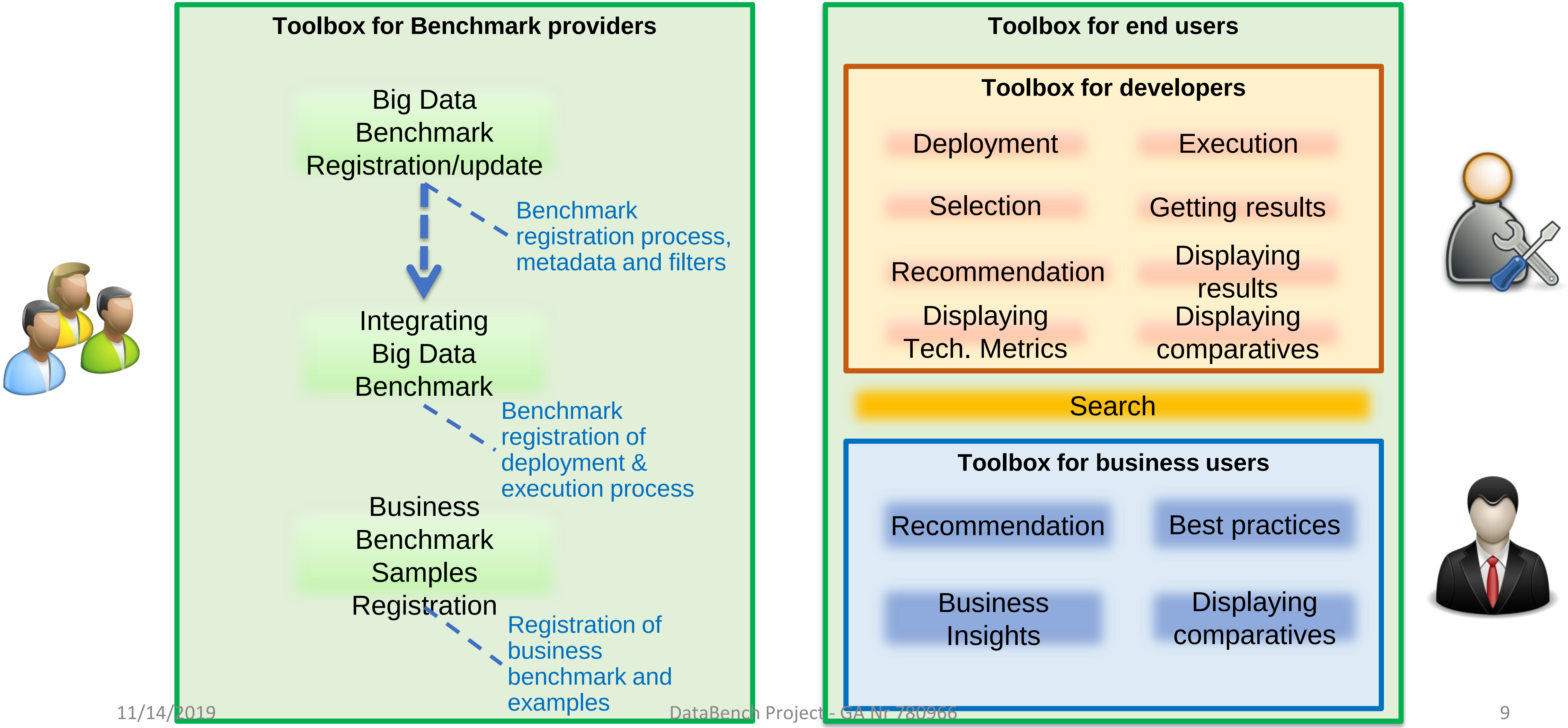
Nothing selected

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

Toolbox for developers

Deployment

Execution

Selection

Getting results

Recommendation

Displaying results

Displaying Tech. Metrics

Displaying comparatives

Search

Toolbox for business users

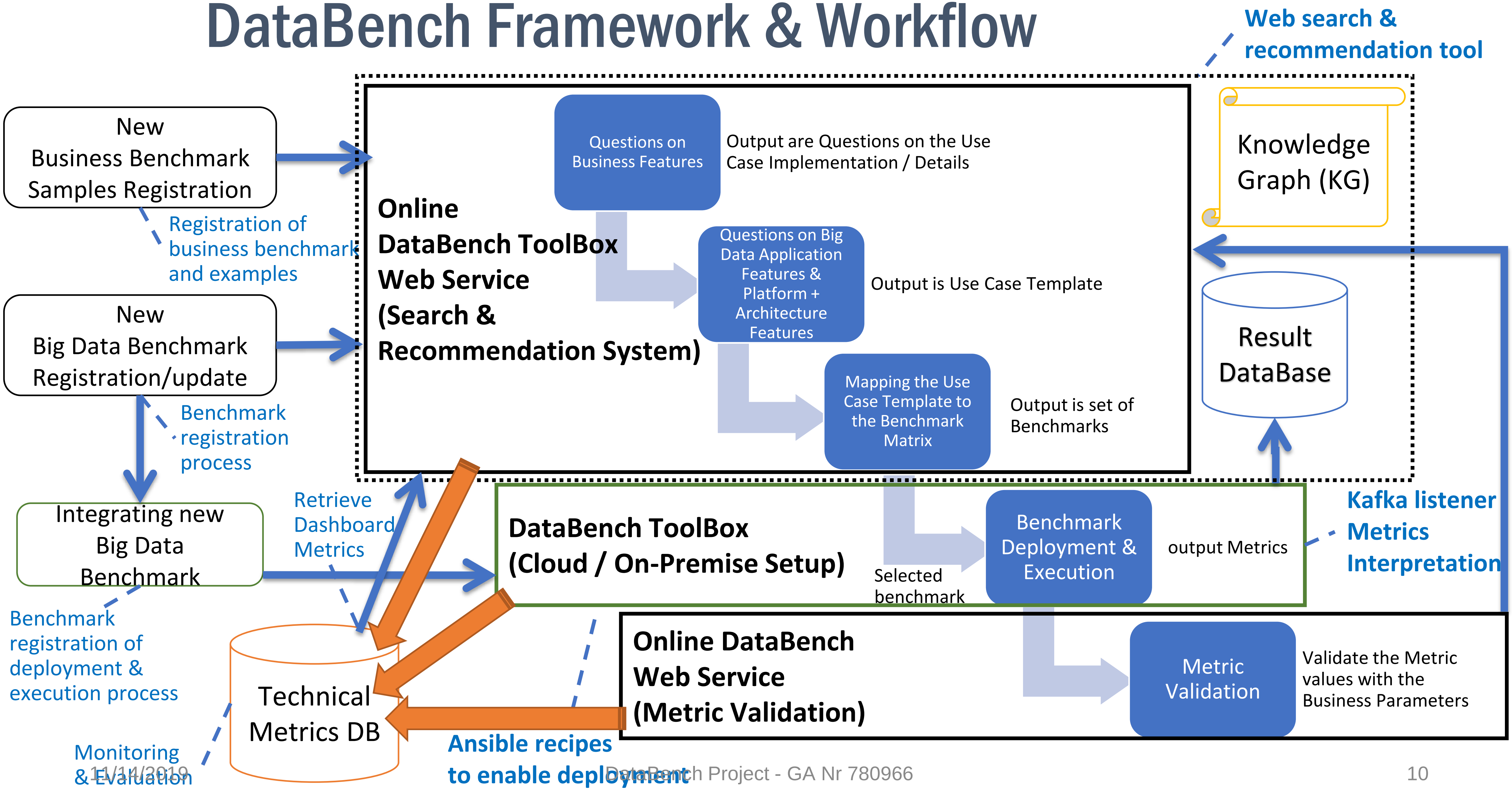
Recommendation

Best practices

Business Insights

Displaying comparatives

DataBench Framework & Workflow



Web search & recommendation tool

New Business Benchmark Samples Registration

Registration of business benchmark and examples

New Big Data Benchmark Registration/update

Benchmark registration process

Integrating new Big Data Benchmark

Benchmark registration of deployment & execution process

Technical Metrics DB

Monitoring & Evaluation

Retrieve Dashboard Metrics

Online DataBench ToolBox Web Service (Search & Recommendation System)

Questions on Business Features

Output are Questions on the Use Case Implementation / Details

Questions on Big Data Application Features & Platform + Architecture Features

Output is Use Case Template

Mapping the Use Case Template to the Benchmark Matrix

Output is set of Benchmarks

Knowledge Graph (KG)

Result DataBase

DataBench ToolBox (Cloud / On-Premise Setup)

Selected benchmark

Benchmark Deployment & Execution

output Metrics

Kafka listener Metrics Interpretation

Online DataBench Web Service (Metric Validation)

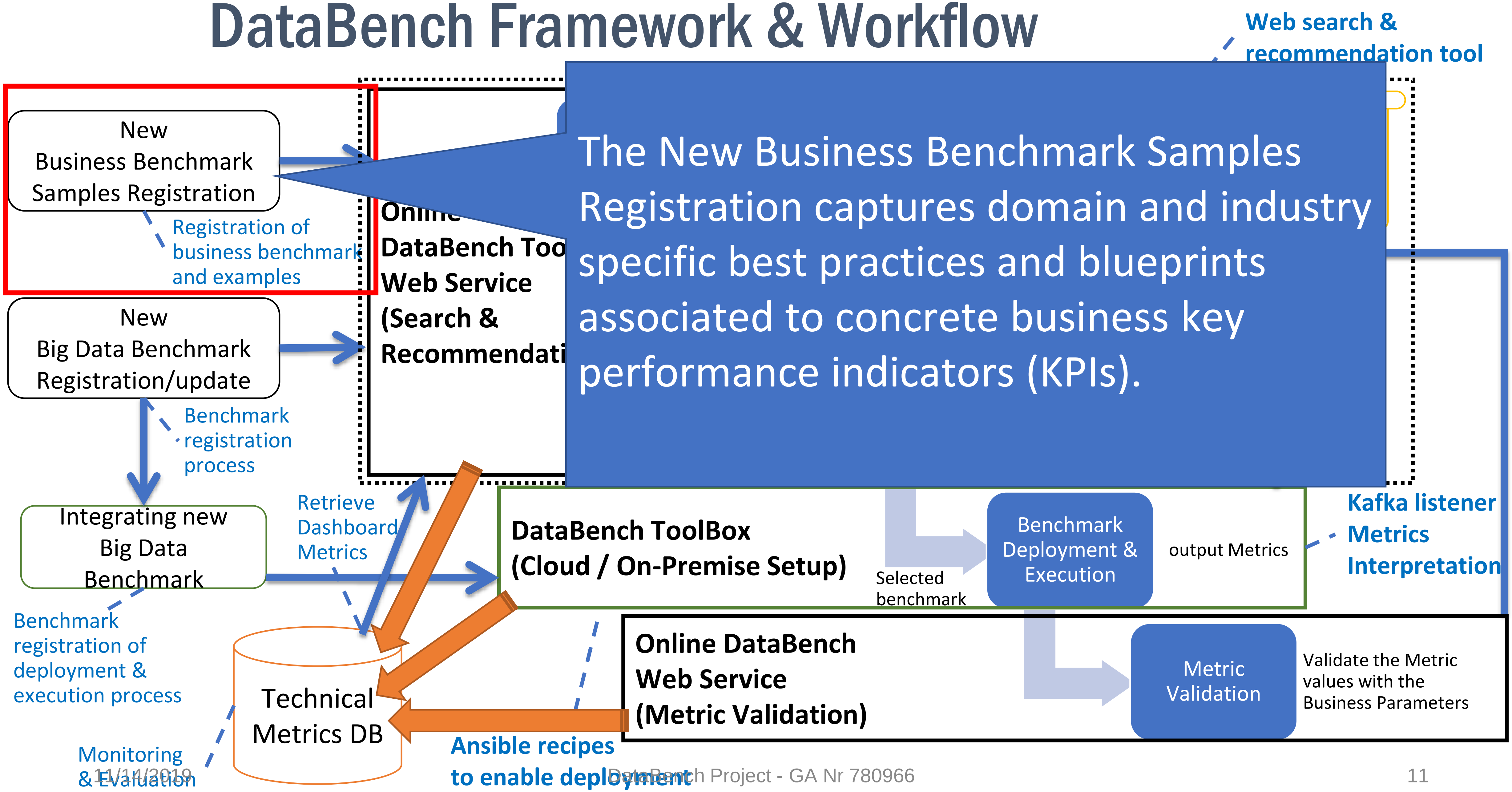
Metric Validation

Validate the Metric values with the Business Parameters

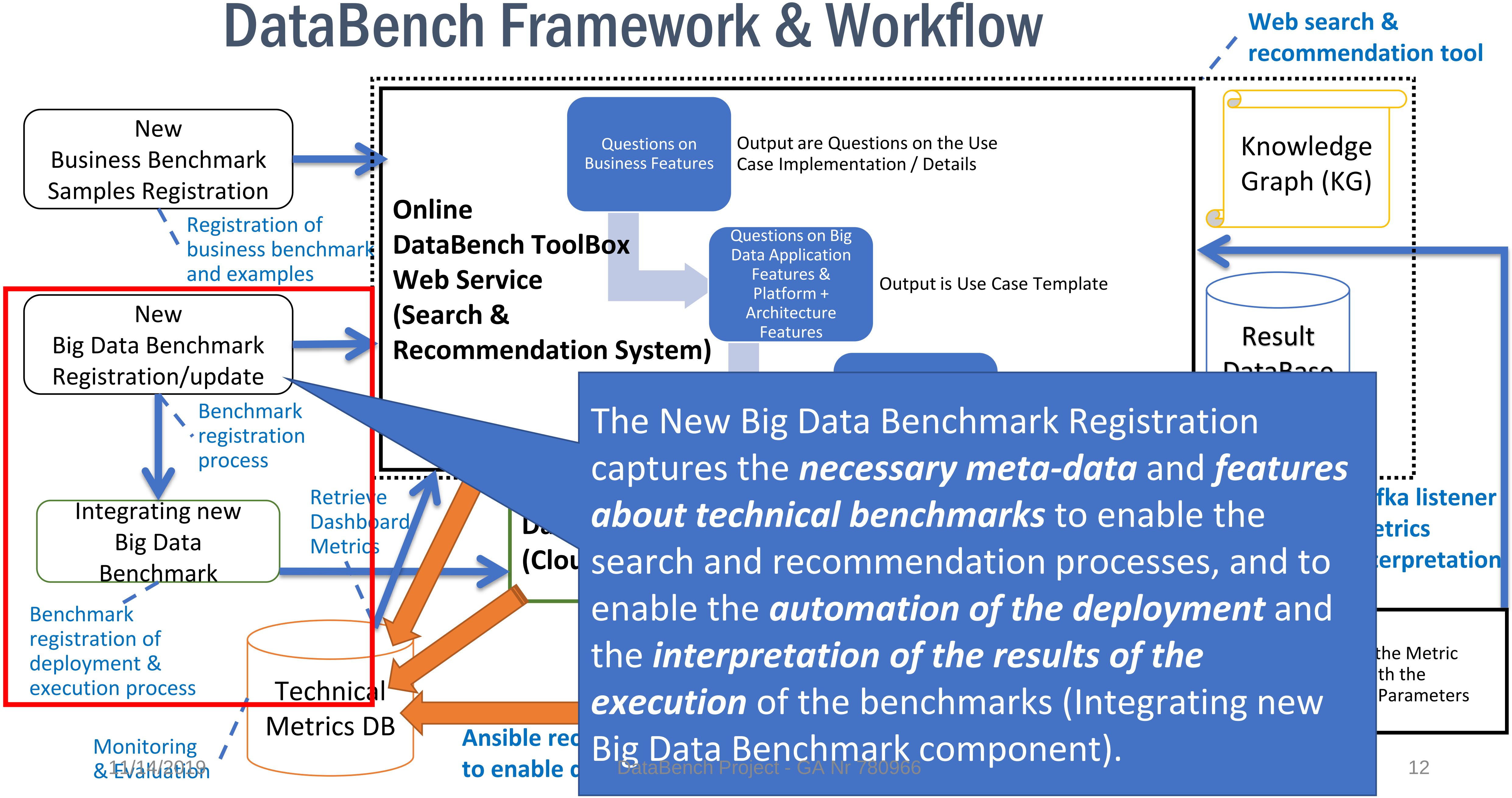
Ansible recipes to enable deployment

DataBench Framework & Workflow

The New Business Benchmark Samples Registration captures domain and industry specific best practices and blueprints associated to concrete business key performance indicators (KPIs).

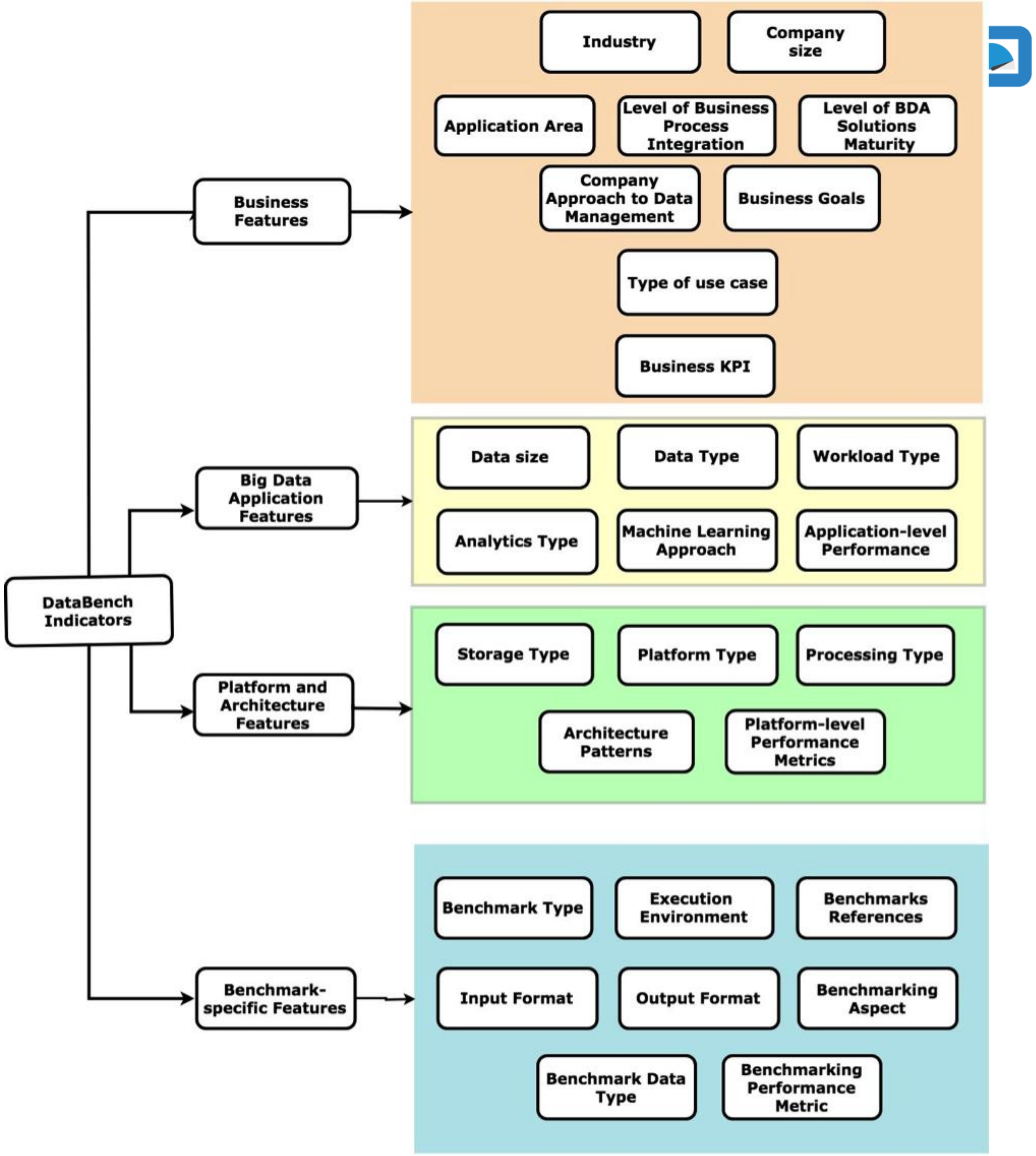

Web search & recommendation tool
New Business Benchmark Samples Registration
Registration of business benchmark and examples
New Big Data Benchmark Registration/update
Benchmark registration process
Onli
DataBench Too
Web Service
(Search &
Recommendati
Integrating new Big Data Benchmark
Benchmark registration of deployment & execution process
Retrieve Dashboard Metrics
Technical Metrics DB
Monitoring & Evaluation
DataBench ToolBox (Cloud / On-Premise Setup)
Selected benchmark
Benchmark Deployment & Execution
output Metrics
Kafka listener Metrics Interpretation
Online DataBench Web Service (Metric Validation)
Metric Validation
Validate the Metric values with the Business Parameters
Ansible recipes to enable deployment

DataBench Framework & Workflow

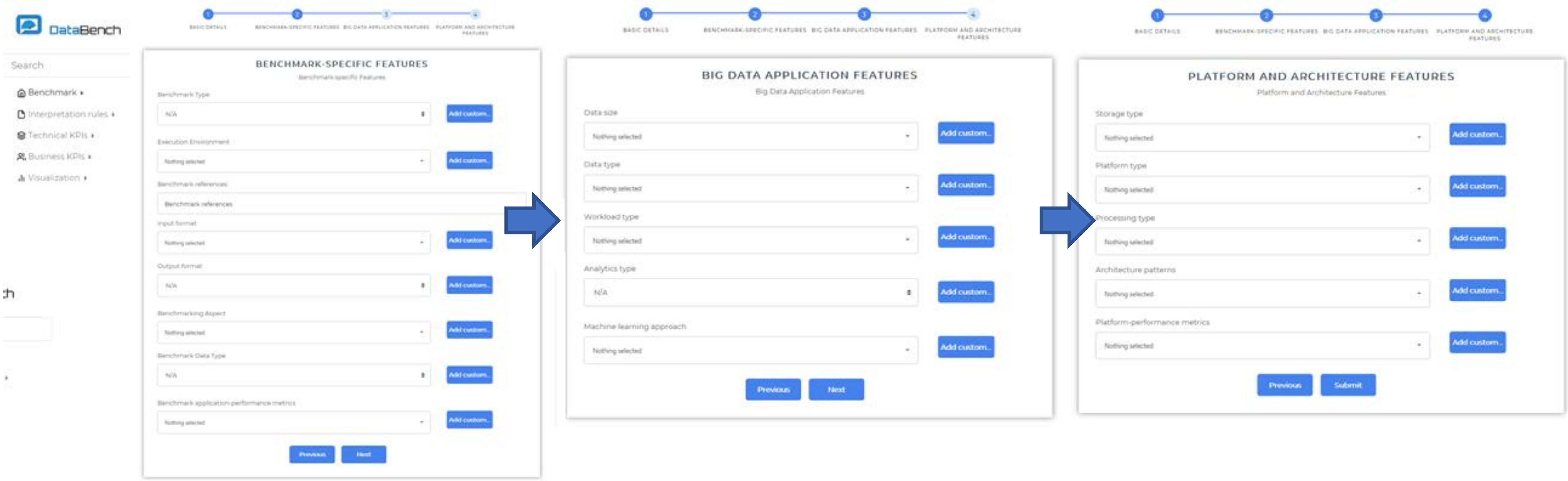

Web search & recommendation tool
New Business Benchmark Samples Registration
Registration of business benchmark and examples
New Big Data Benchmark Registration/update
Benchmark registration process
Integrating new Big Data Benchmark
Benchmark registration of deployment & execution process
Monitoring & Evaluation
Technical Metrics DB
Retrieve Dashboard Metrics
Online DataBench ToolBox Web Service (Search & Recommendation System)
Questions on Business Features
Output are Questions on the Use Case Implementation / Details
Questions on Big Data Application Features & Platform + Architecture Features
Output is Use Case Template
Knowledge Graph (KG)
Result DataBase
(Clou
Ansible rec
to enable c
fka listener
etrics
erpretation
the Metric
th the
Parameters

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



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

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

PLATFORM AND ARCHITECTURE FEATURES

Platform and Architecture Features

Storage type: Nothing selected

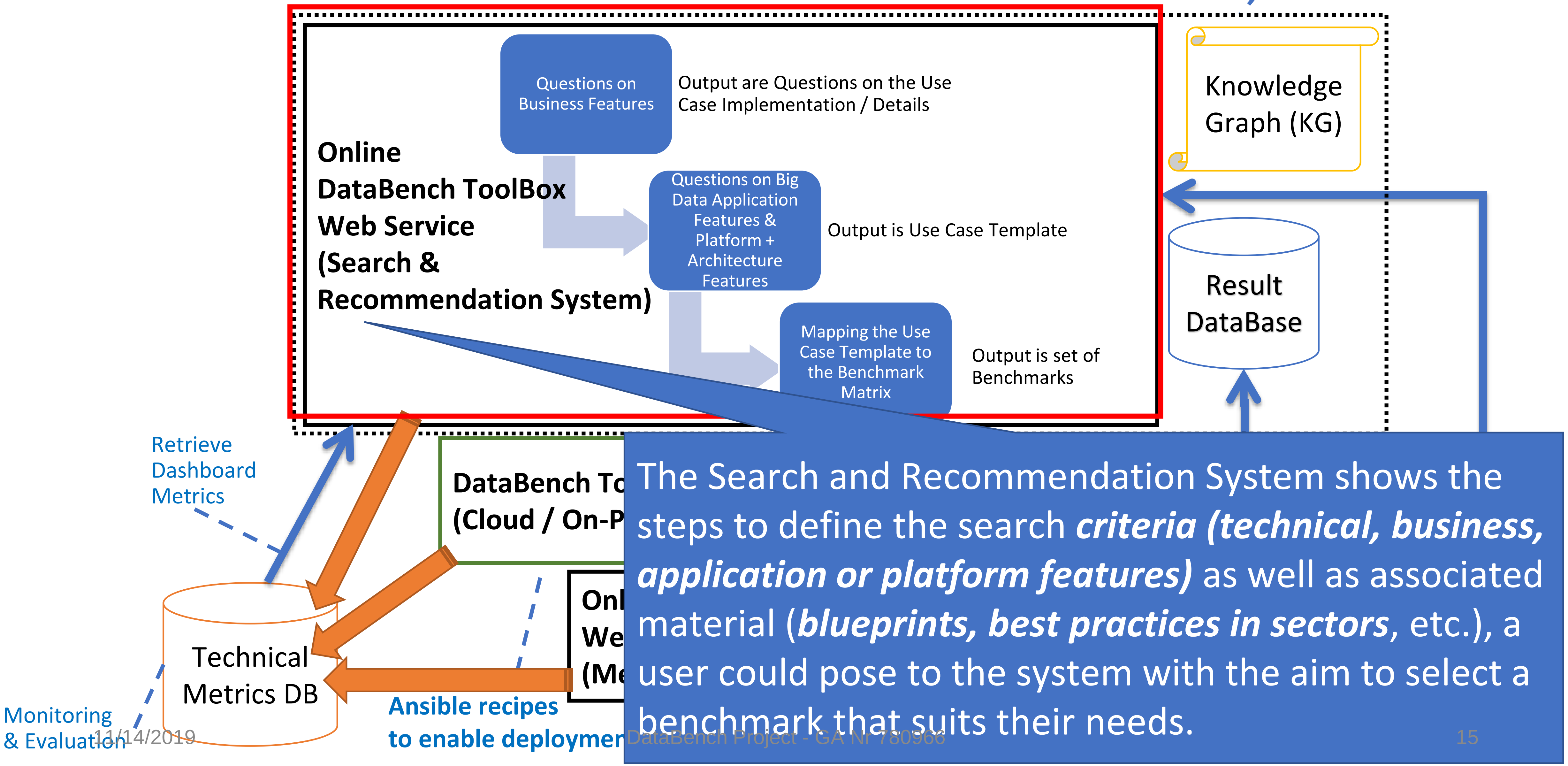
Platform type: Nothing selected

Processing type: Nothing selected

Architecture patterns: Nothing selected

Platform-performance metrics: Nothing selected

DataBench Framework & Workflow

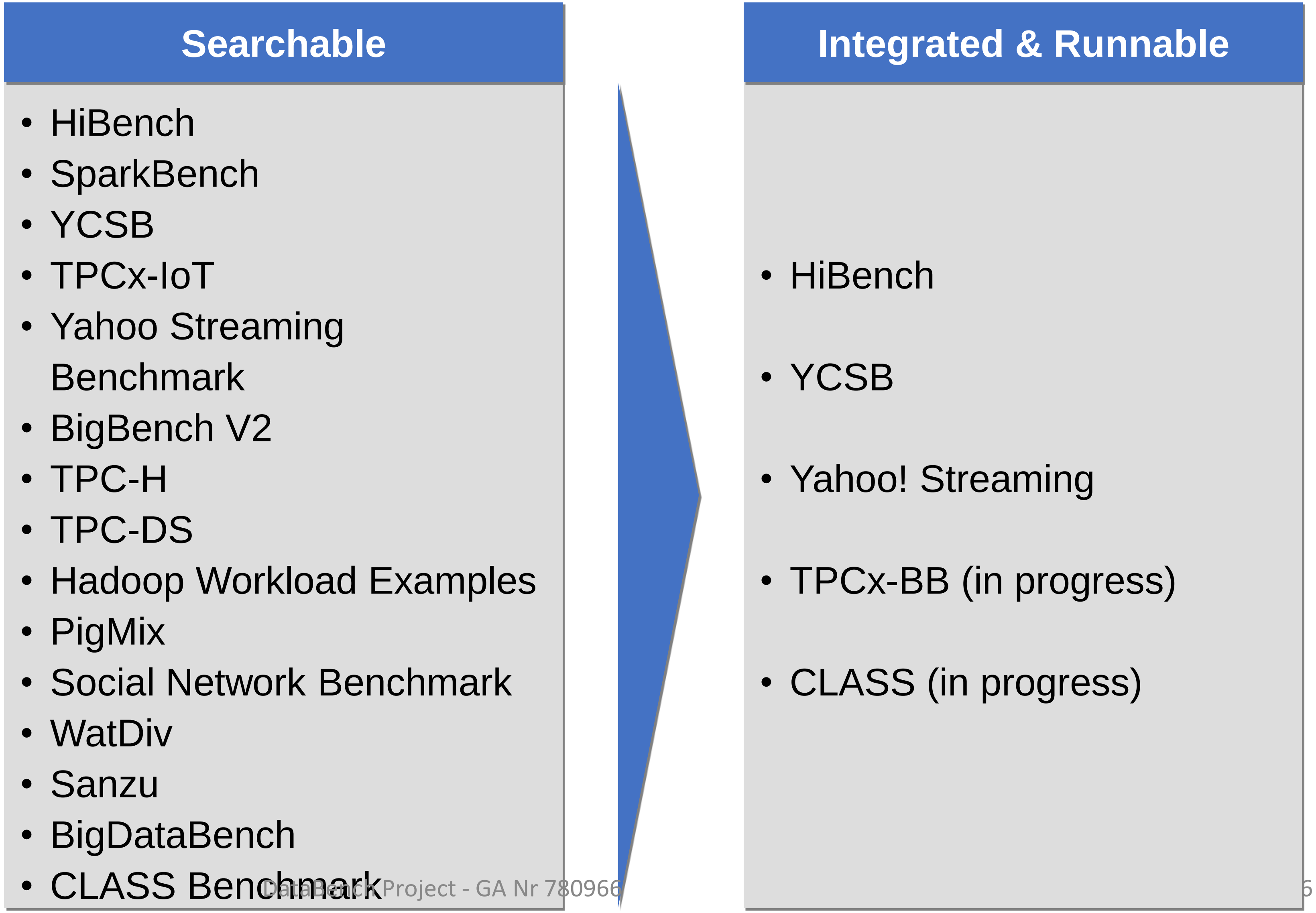


The Search and Recommendation System shows the steps to define the search ***criteria (technical, business, application or platform features)*** as well as associated material (***blueprints, best practices in sectors***, etc.), a user could pose to the system with the aim to select a benchmark that 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

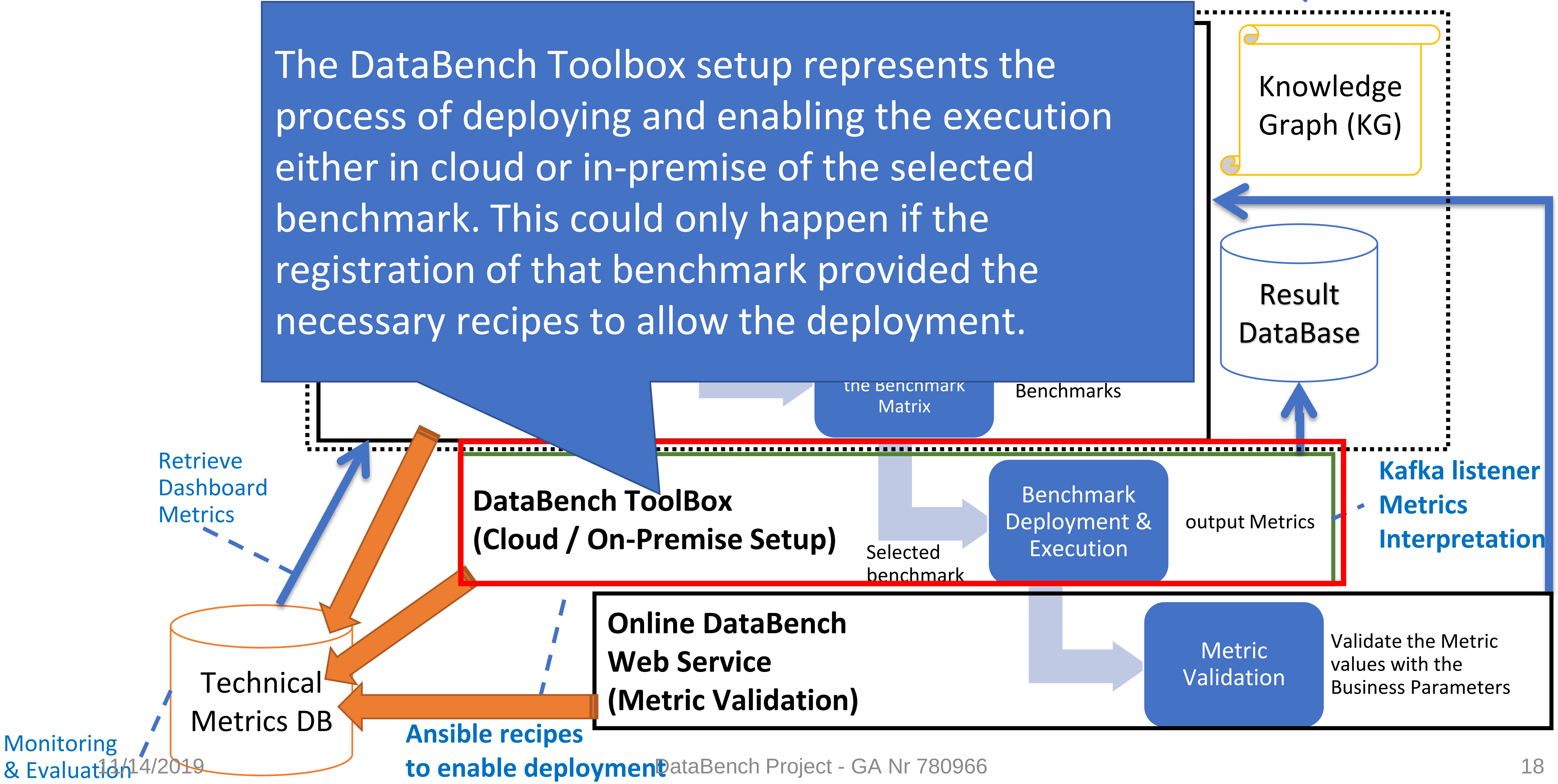
Integrated & Runnable

- HiBench
- YCSB
- Yahoo! Streaming
- TPCx-BB (in progress)
- CLASS (in progress)

Initial benchmarks to be integrated in the Toolbox

Name	Domain	Data Type	Funcitonality	Status
HiBench	Microbenchmark. ML, SQL, Websearch, Graph, Streaming Benchmarks	Structured, Text, Web Graph	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 **micro, machine learning, sql, graph, websearch and streaming.**	**done**
SparkBench	Microbenchmark. ML, Graph Computation, SQL, Streaming	Structured, Text, Web Graph	System for benchmarking and simulating **Spark jobs**. Multiple workloads organized in 4 categories.	**In progress**
YCSB	Microbenchmark. Cloud OLTP operations	Structured	Evaluates performance of different **"key-value" and "cloud" serving systems**, 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.	**done: Arango, Mongo, Orient, Redis**
TPCx-IoT	Microbenchmark. Workloads on typical IoT Gateway systems	Structured, IoT	Based on YCSB. Workloads of data ingestion and concurrent queries simulating workloads on typical **IoT Gateway systems**. Dataset with data from sensors from electric power station(s)	**In progress**
Yahoo Streaming Benchmark	Appl. benchmars. Ad analytics pipeline	Structured, Time Series	The Yahoo Streaming Benchmark is a **streaming application benchmark** simulating an **advertisement analytics** pipeline.	**Integrated, parametrization**
BigBench V1 & V2 / TPCx-BB	Appl. benchmark. Fictional product retailer platform	Structured, Text, JSON logs	End-to-end, technology agnostic, **application-level** benchmark that tests the **analytical capabilities** of a Big Data platform. It is based on a fictional product retailer business model.	**In progress**

DataBench Framework & Workflow



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.

Web search & recommendation tool

Knowledge Graph (KG)

Result DataBase

the Benchmark Matrix

Benchmarks

Retrieve Dashboard Metrics

DataBench ToolBox (Cloud / On-Premise Setup)

Selected benchmark

Benchmark Deployment & Execution

output Metrics

Kafka listener Metrics Interpretation

Online DataBench Web Service (Metric Validation)

Metric Validation

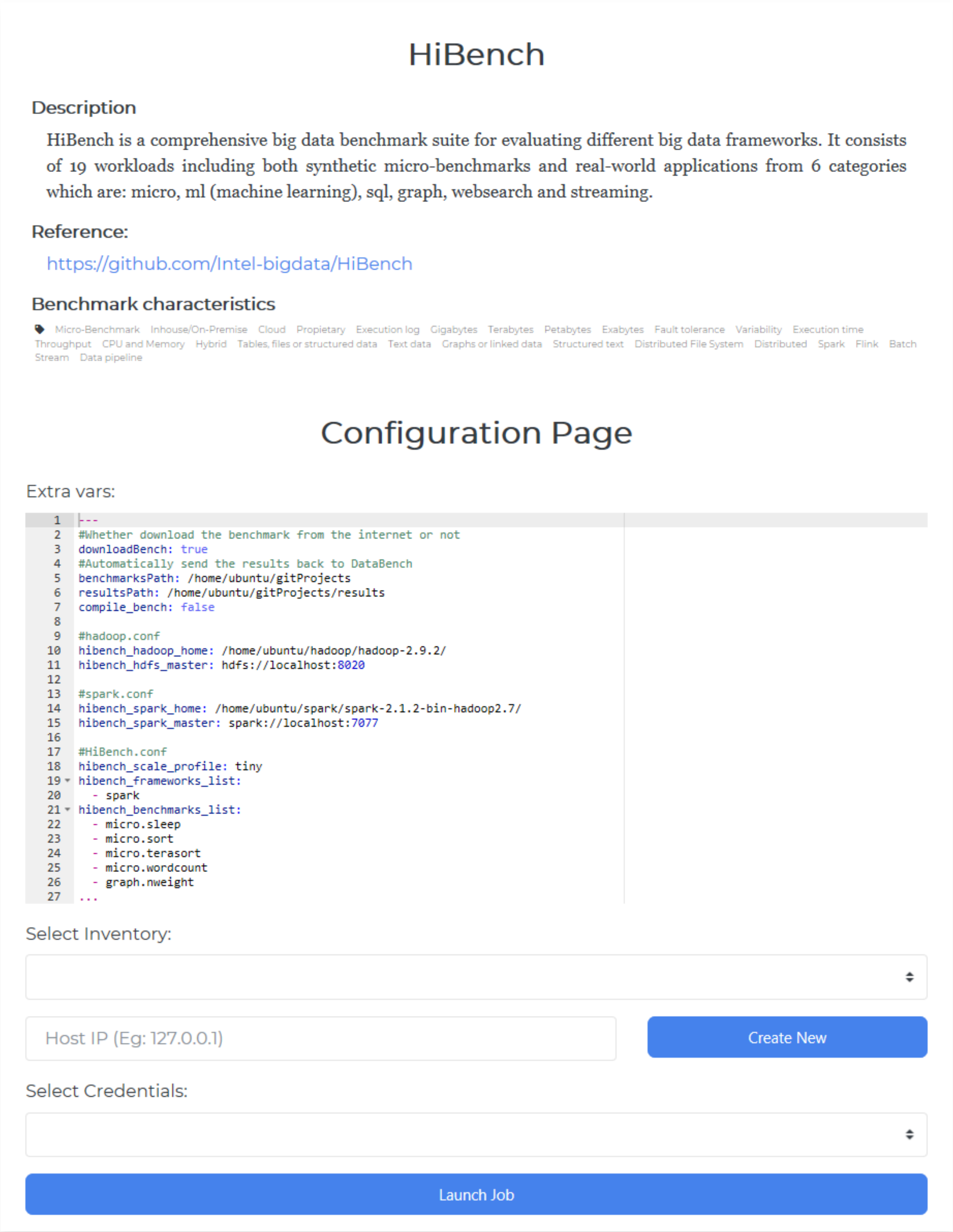
Validate the Metric values with the Business Parameters

Technical Metrics DB

Monitoring & Evaluation

Ansible recipes to enable deployment

Configure benchmark parameters and execute the benchmark!



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:

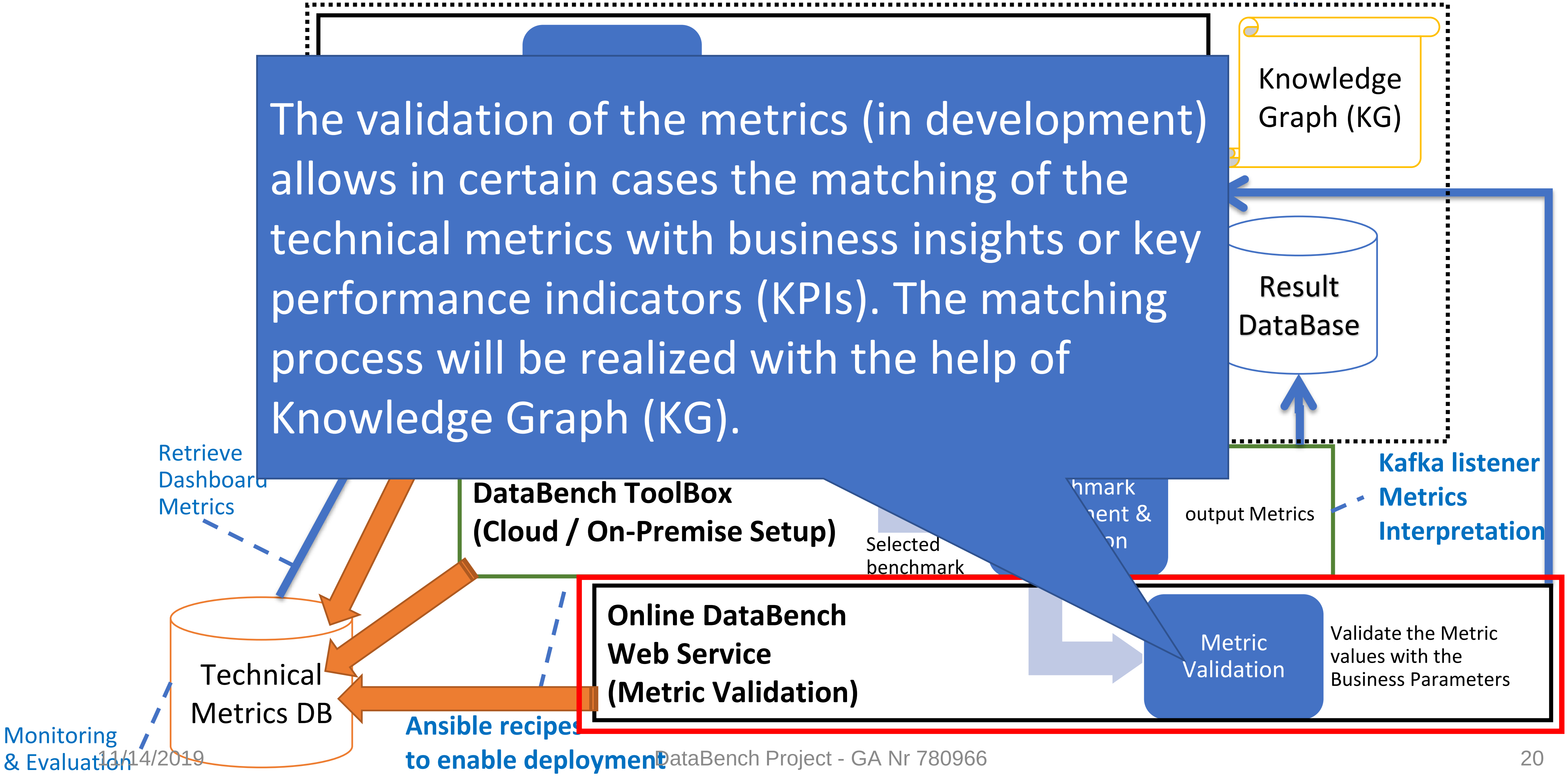
```
---
#Whether download the benchmark from the internet or not
downloadBench: true
#Automatically send the results back to DataBench
benchmarksPath: /home/ubuntu/gitProjects
resultsPath: /home/ubuntu/gitProjects/results
compile_bench: false

#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:

DataBench Framework & Workflow

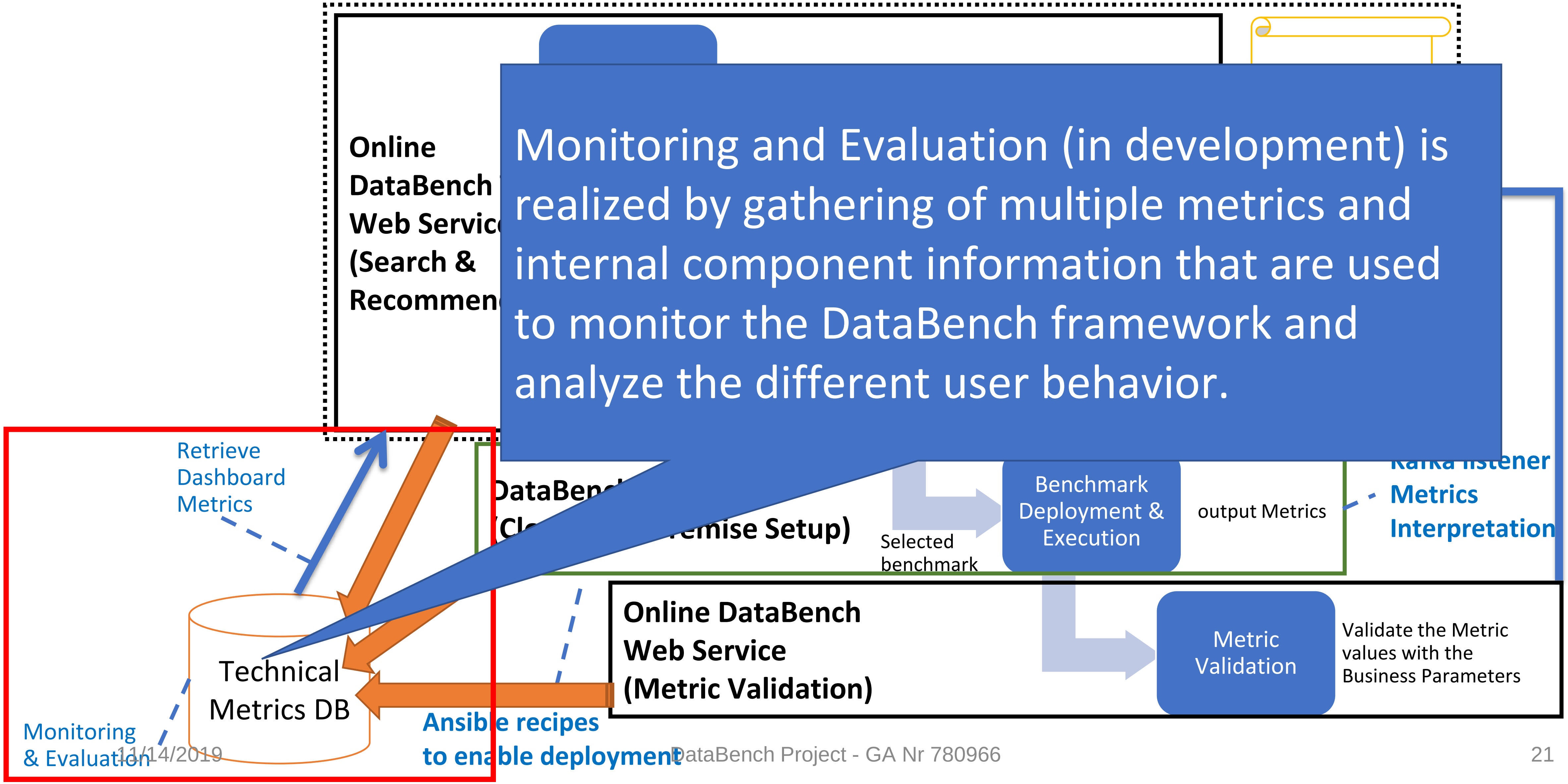


The validation of the metrics (in development) allows in certain cases the matching of the technical metrics with business insights or key performance indicators (KPIs). The matching process will be realized with the help of Knowledge Graph (KG).

Web search & recommendation tool
Knowledge Graph (KG)
Result DataBase
Kafka listener Metrics Interpretation
Retrieve Dashboard Metrics
DataBench ToolBox (Cloud / On-Premise Setup)
hmark
ent &
on
output Metrics
Selected benchmark
Online DataBench Web Service (Metric Validation)
Metric Validation
Validate the Metric values with the Business Parameters
Technical Metrics DB
Monitoring & Evaluation
Ansible recipes to enable deployment

DataBench Framework & Workflow

Web search & recommendation tool



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)

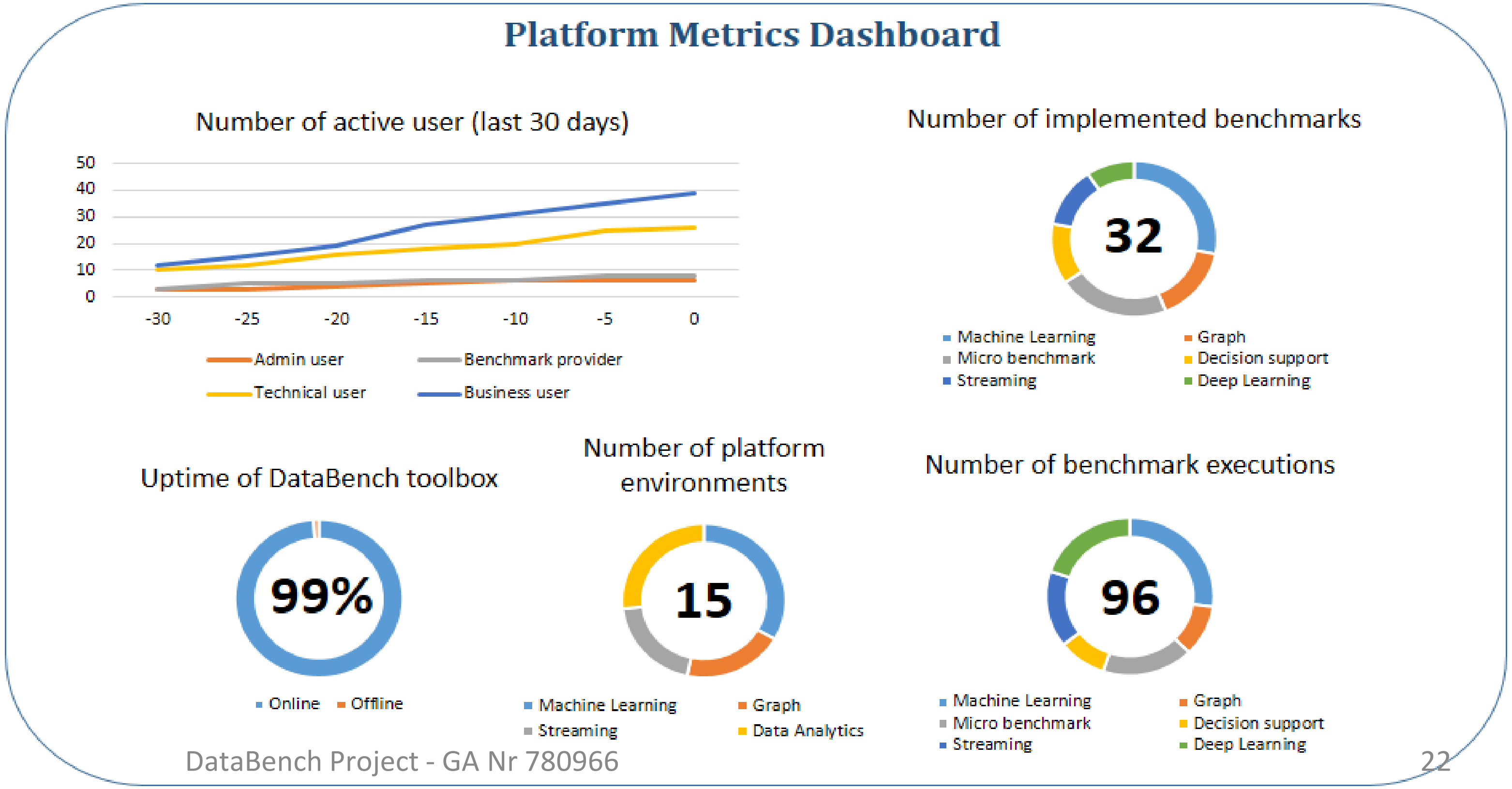
- number of (active) users
- number of implemented benchmarks
- number of benchmark runs
- number of platform environments
- more …

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 …

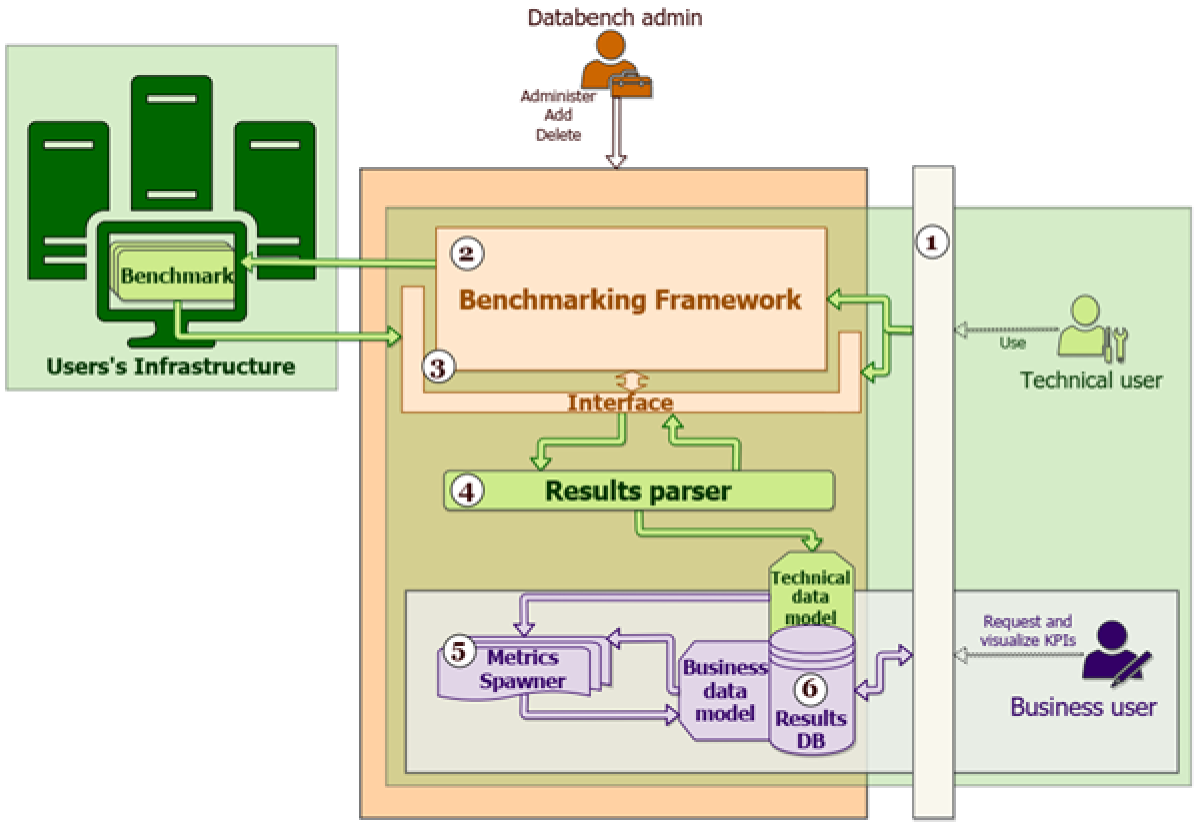
Administrator Metrics Dashboard

- 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 …



DataBench Architecture (1)

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.



DataBench Architecture (2)

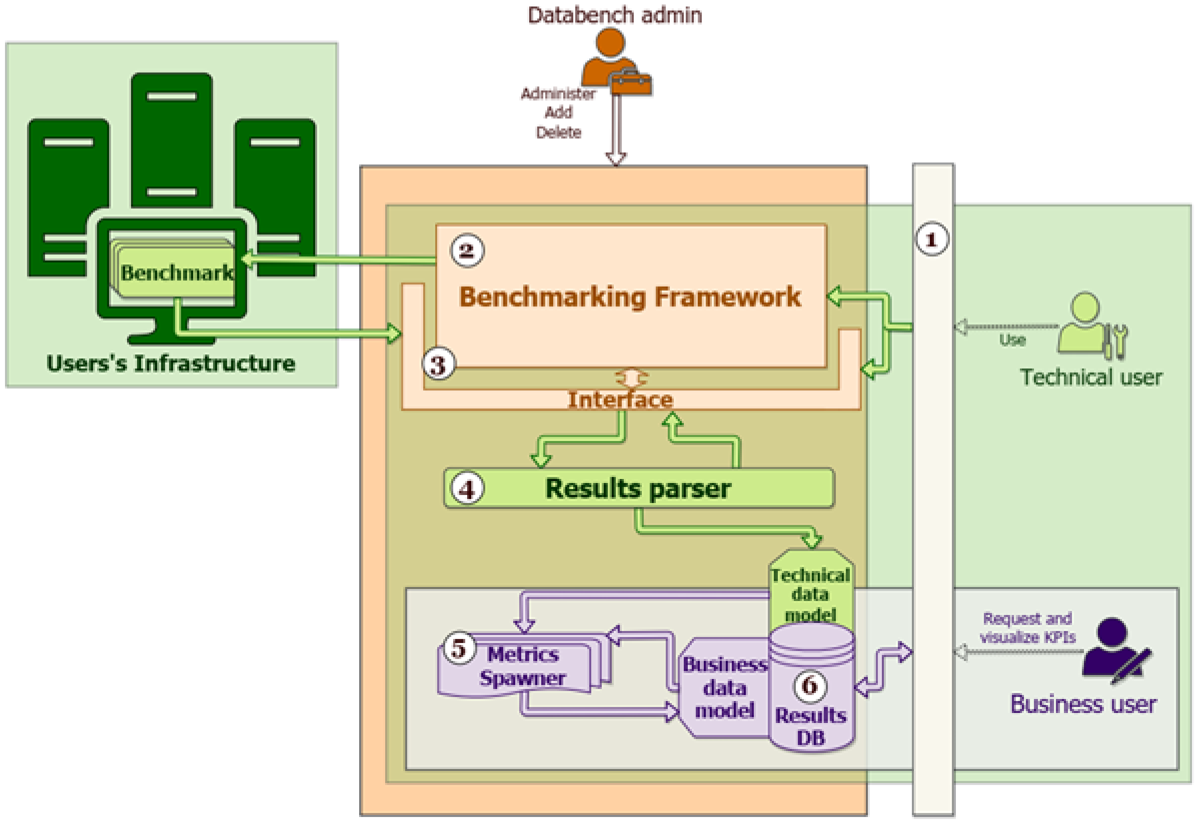
4. ***Results Parser*** converts the benchmark results into standardized data model to enable calculation of the business metrics in the next steps.

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.***

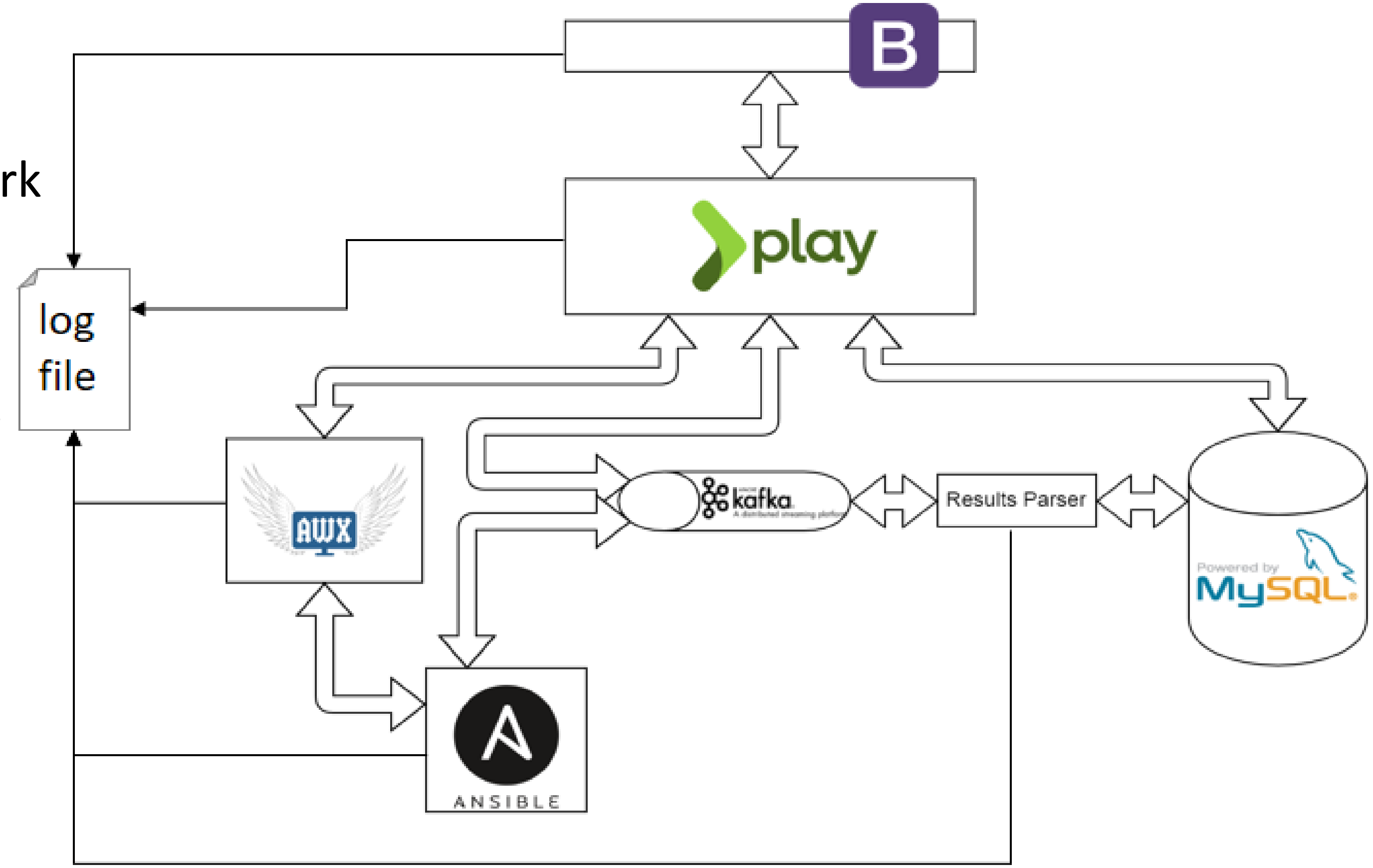
7. ***Metrics DB*** is very similar to the Results DB module with the difference that it stores persistently the collected monitoring metrics.

8. ***Metrics Dashboards*** offer the monitoring and evaluation functionality of the DataBench framework.

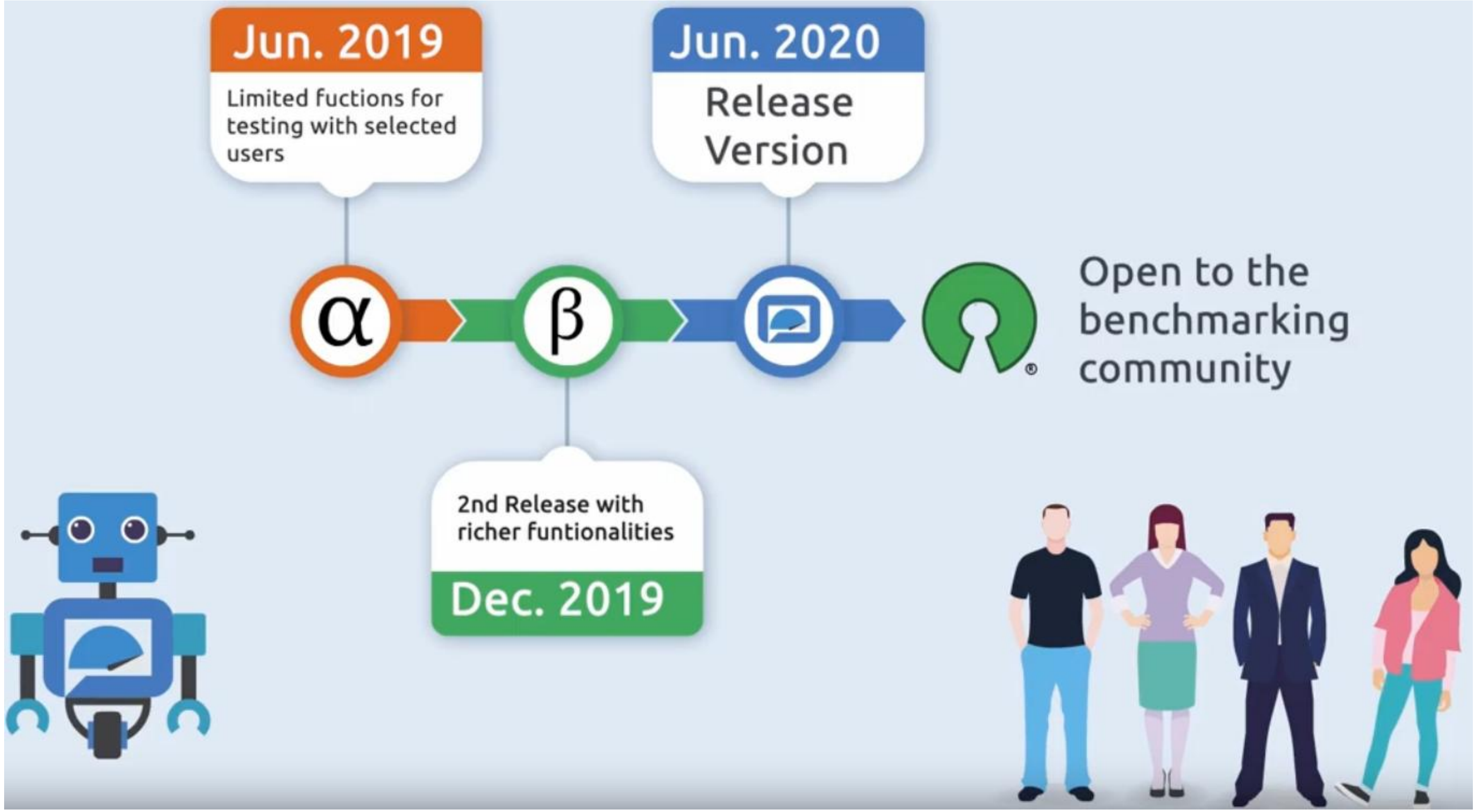


DataBench Implementation (In progress)

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 It 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



Jun. 2019

Limited fuctions for testing with selected users

Jun. 2020

Release Version

Open to the benchmarking community

2nd Release with richer funtionalities

Dec. 2019

Alpha version URL: http://83.149.125.78:9000/
More details in D3.2: https://www.databench.eu/wp-content/uploads/2019/07/d3.2-databench-toolbox-alpha-including-support-for-reusing-of-existing-benchmarks.pdf

Contacts

✉ info@databench.eu

@DataBench_eu

DataBench

DataBench Project

DataBench

Visit: www.databench.eu