Evidence Based Big Data Benchmarking to Improve Business Performance

Benchmarking Big Data

20/06/2019, Open Expo Europe, Madrid

Tomas Pariente Lobo, ATOS
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**

<table>
<thead>
<tr>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Nodes (servers)</td>
<td>30 Nodes (servers)</td>
</tr>
<tr>
<td>4 hours (execution time)</td>
<td>4 minutes (execution time)</td>
</tr>
<tr>
<td>5500 Euro (total cost)</td>
<td>50 000 Euro (total cost)</td>
</tr>
</tbody>
</table>
Many Big Data Benchmarks around

**SUVA Reference Model**

<table>
<thead>
<tr>
<th>Domain/Sector</th>
<th>Business</th>
<th>Industry</th>
<th>Research</th>
<th>Government</th>
<th>Work-in-progress (WP1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retail</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufacturing</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

**Benchmarks**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TPC-H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-COL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-CB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBase-OF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-DS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-CD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBase-OF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-CBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-CBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBase-OF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-CBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC-CBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

21/06/2019
DataBench
One stop shop for Big Data Benchmarking
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
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

**ECONOMIC MARKET AND BUSINESS ANALYSIS**

**USE CASES** = Typologies of technology adoption in specific application domains and/or business processes

**EVALUATING BUSINESS PERFORMANCE**

- 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 Toolbox
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 get benchmarks technical results and business insights.
Methodology Workflow & Implementation

Online DataBench ToolBox Web Service (Search & Recommendation System)

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

DataBench ToolBox (Cloud / On-Premise Setup)

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

Selected benchmark

Benchmark Deployment & Execution
- output Metrics

Online DataBench Web Service (Search & Recommendation System)

Ansible recipes to enable deployment

Web search & recommendation tool

Kafka listener Metrics Interpretation

Metric Validation
- Validate the Metric values with the Business Parameters
Toolbox usage: General Overview

Toolbox for Benchmark providers

- Big Data Benchmark Registration/update
- 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
  - Selection
  - Recommendation
  - Displaying Tech. Metrics
  - Execution
  - Getting results
  - Displaying results

- Toolbox for business users
  - Recommendation
  - Best practices
  - Business Insights
  - Displaying comparatives
Alpha version of the Toolbox already available for Alpha-testers

<table>
<thead>
<tr>
<th>Searchable</th>
<th>Runnable</th>
</tr>
</thead>
<tbody>
<tr>
<td>• HiBench</td>
<td>• HiBench</td>
</tr>
<tr>
<td>• SparkBench</td>
<td>• YCSB</td>
</tr>
<tr>
<td>• YCSB</td>
<td>• Yahoo! streaming</td>
</tr>
<tr>
<td>• TPCx-IoT</td>
<td>• CLASS (in progress)</td>
</tr>
<tr>
<td>• Yahoo Streaming Benchmark</td>
<td></td>
</tr>
<tr>
<td>• BigBench V2</td>
<td></td>
</tr>
<tr>
<td>• TPC-H</td>
<td></td>
</tr>
<tr>
<td>• TPC-DS</td>
<td></td>
</tr>
<tr>
<td>• Hadoop Workload Examples</td>
<td></td>
</tr>
<tr>
<td>• PigMix</td>
<td></td>
</tr>
<tr>
<td>• Social Network Benchmark</td>
<td></td>
</tr>
<tr>
<td>• WatDiv</td>
<td></td>
</tr>
<tr>
<td>• Sanzu</td>
<td></td>
</tr>
<tr>
<td>• BigDataBench</td>
<td></td>
</tr>
<tr>
<td>• CLASS Benchmark</td>
<td></td>
</tr>
</tbody>
</table>
Registering a Benchmark in the Toolbox
Adding configuration for Benchmark deployment and run

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 form Git the DataBench Toolbox
  2. Fill in the variables files with the data of the target system
  3. Run de playbook.

Variables:

```
#Hbench Specific variables
compile_bench: false

#hadoop.conf
hbench_hadoop_home: /home/ubuntu/hadoop/hadoop-2.9.2/
hbench_hdfs_master: hdfs://XXXXX:8820

#spark.conf
hbench_spark_home: /home/ubuntu/spark/spark-2.1.2-bin-hadoop2.7/
hbench_spark_master: spark://localhost:7077

#Hbench.conf
hbench_scale_profile: tiny
hbench_frameworks_list:
  - spark
hbench_benchmarks_list:
  - micro.sleep
  - micro.sort
  - micro.terasort
  - micro.wordcount
  - graph.weight
```

Playbook running:
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...
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/](https://www.databench.eu/)
Evidence Based Big Data Benchmarking to Improve Business Performance

tomas.parientelobo@atos.net
@tpariente