Collaborative Project

Holistic Benchmarking of Big Linked Data

Project Number: 688227
Start Date of Project: 2015/12/01, **Duration:** 36 months

Deliverable 8.4
Standardization Report

Dissemination Level	Public
Due Date of Deliverable	Month 36, 30/11/2018
Actual Submission Date	Month 36, 30/11/2018
Work Package	WP8 - Dissemination
Task	T8.1
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	7
Filename	D8.4_Standardization_Report.pdf

Abstract: This deliverable discusses the dissemination actions taken and the results achieved for the duration of the project.

The information in this document reflects only the author's views and the European Commission is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688227.

History

Version	Date	Reason	Revised by
0.1	15/11/2018	First Draft	Axel Ngonga (INFAI)
0.2	22/11/2018	Second Draft	Axel Ngonga (INFAI)
0.3	30/11/2018	Feedback	Michael Röder (INFAI)
1.0	30/11/2018	Revision and final draft	Axel Ngonga (INFAI)

Author List

Organization	Name	Contact Information
INFAI	Axel Ngonga	ngonga@infai.org
INFAI	Michael Röder	roeder@informatik.uni-leipzig.de

Executive Summary

We monitored the development of standards for benchmarking Big Linked Data throughout the course of the HOBBIT project. While benchmarking has been a core concern of a number of organizations over the last years, no *standards* (i.e., protocols or platforms) have been established pertaining to benchmarking Big Linked Data. Still, a number of Linked Data and Big Data benchmarking platforms—which often reuse a small set of technologies—have seen the light of day over the last four years. We give a brief overview of some of the most widely known benchmarking platforms and point to their technical components. The special issue of the Semantic Web Journal on benchmarking—which was managed by members of the HOBBIT consortium—conveys an overview of relevant technologies in the domain of the Semantic Web.

1 Introduction

The core technical tasks of the HOBBIT project (as reflected in WPs 2-6) were the development of a benchmarking platform for Big Linked Data and of the corresponding benchmarks. The monitoring of the state of the art in benchmarking (especially w.r.t. standardization) was hence of utmost importance to the project. With this monitoring, we aimed to ensure that the platform offered state-of-the-art capabilities w.r.t. benchmarking. *No benchmarking standards specific to Big Linked Data were developed over the three years of the project.* Still, some relevant Big Data benchmarking platforms were developed between the grant proposal writing phase and the project completion. We hence give a brief overview of some of the most relevant of these platforms. Details on the platform can be found at the links and references provided. Some of the descriptions provided below are summaries of the platform overviews provided by the platform developers.

2 Linked Data Benchmarking

Several benchmarks have been developed in the area of RDF datasets. For a large proportion of existing benchmarks and benchmark generators (e.g., LUBM [9], SP^{2}Bench [14], BSBM [1], SRBench [21], DBSBM [10] and FEASIBLE [13]), the focus has currently on creating frameworks able to generate data and query loads [9, 1, 10, 13] able to stress triple stores. IGUANA [4] is the first benchmarking framework for the unified execution of these data and query loads. However, IGUANA focuses exclusively on benchmarking storage solutions. Moreover, it is not designed to scale up to distributed processing.

Knowledge Extraction—especially Named Entity Recognition and Linking—has also seen the rise of a large number of benchmarks [12]. Several conferences and workshops aiming at the comparison of information extraction systems (including the Message Understanding Conference [15] and the Conference on Computational Natural Language Learning [16]) have created benchmarks for this task. In 2014, Carmel et al. [2] introduced one of the first Web-based evaluation systems for Named Entity Recognition and Linking. The BAT benchmarking framework [5] was also designed to facilitate benchmarking based on these datasets by combining seven Wikipedia-based systems and five datasets. The GERBIL framework [12] extended this idea by being knowledge-base-agnostic and addressing the NIL error problem in the formal model behind the BAT framework.While these systems all allow for benchmarking knowledge extraction solutions, they do not scale up to the requirements of distributed systems.

In the area of Question Answering using Linked Data, challenges such as BioASQ [18], and the Question Answering over Linked Data (QALD) [19] have aimed to provide benchmarks for retrieving answers to human-generated questions. The GERBIL-QA platform [20] is the first open benchmarking platform for question answering which abides by the FAIR principles. However, like its knowledge extraction companion, it is not designed to scale up to large data and task loads.

A recent detailed comparison of instance matching benchmarks can be found in [6]. The authors show that there are several benchmarks using either real or synthetically generated datasets. SEALS [1] is the best-known platform for benchmarking link discovery frameworks. It offers the flexible addition of datasets and measures for benchmarking link discovery. However, the platform was not designed to scale and can thus not deal with datasets which demand distributed processing.

The HOBBIT platform is the first benchmarking framework which supports all steps of the LD life cycle which can be benchmarked automatically (see Table 1).[2] In addition, it is the first benchmarking

[1] http://www.seals-project.eu/

[2] We are not aware of the existence of an automatic benchmark for the quality analysis step. However, the platform

platform for Linked Data which scales up to the requirements of Big Data platforms through horizontal scaling. The comparability of HOBBIT's benchmarking results are ensured by the cluster underlying the open instantiation of the platform.[3]

Table 1: Comparison of benchmarking frameworks

	Year	Extraction	Storage	Manual Revision	Linking	Enrichment	Quality Analysis	Evolution	Exploration	Scalable Data	Scalable Tasks	Fair Benchmarking
BAT	2013	✓		—								
GERBIL	2014	✓		—								✓
IGUANA	2017		✓	—							✓	
HOBBIT	2017	✓	✓	—	✓	✓	(✓)	✓	✓	✓	✓	✓

3 Other Benchmarking Platforms

A number of benchmarking platforms for Big Data have been developed over the last years, of which we present a few below. Surveys on Big Data benchmarking can be found in [7, 17].

3.1 Peel

The Peel framework[4] supports the automation of experiments on Big Data infrastructure. These experiments comprise a system that is executed using input data that is either provided in a directory or from a piece of code that is executed. However, the framework only supports systems that can be executed on one of the supported Big Data solutions like Flink or Spark.[5] The main drawback of the framework is that the results it generates are not transparent as the execution of systems and benchmarks is hidden from the users, making it unclear how the resource allocation for benchmark and systems was. Similar observations hold for Plug and Play Bench [3], which was designed to evaluate different hardware settings.

3.2 BigBench

Also relevant according to the literature are novel Big Data benchmarks for benchmarking relational databases, e.g., BigBench [8]. BigBench aims to be technology-agnostic by relying on specifications combined with an open-source reference implementation kit. The kit is designed to lower the barrier of entry to benchmarking for users from industry and academia. The open-source character of the

itself would support such a benchmark.

[3] See http://master.project-hobbit.eu.

[4] http://peel-framework.org

[5] The complete list can be found at https://github.com/peelframework/peel#supported-systems.

platform targets consistency and comparability. These targets are shared with the Hobbit platform but implemented in a manner not compatible with the FAIR principles,[6] which are at the core of the design of Hobbit.

3.3 HiBench

Intel's HiBench[7] is designed to evaluate the runtime performance, throughput and resource use of benchmarking platforms. At the time of writing, the suite contained 19 workloads pertaining to sorting, counting, search and indexing as well as machine learning. The platform can be deployed on Hadoop, Flink, Storm, Gearpump and Kafka. While the platform is diverse, it is not designed to support Linked Data. For example, it provides no workload for SPARQL. Moreover, it does not cover an important portion of the lifecycle of data (e.g., extraction, visualization, browsing, etc.). Still, the YAML-based deployment and the compatibility with a plethora of means for Big Data deployment make it a viable contribution for Big Data benchmarking.

3.4 BigDataBench

The design of this platform[8] is based on the concept of *data motifs*. These are fundamentally atomic building blocks of processing pipelines for data analytics. BigDataBench contains 13 datasets which reflect workloads from domains such as search engines and social networks. While the 47 benchmarks (incl. micro- and macro-benchmarks) included in the suite are all related Bid Data analytics, they allow for benchmarking various aspects of this important step in Big Data processing. Like HiBench, the platform can be deployed using Hadoop, Spark and Flink. In addition, it supports MPI. Still, the suite is designed for exactly one step of the Big Data lifecycle and does not cover Linked Data.

3.5 CloudSuite

Now available in version 3.0, CloudSuite[9] targets tasks related to Cloud Services (e.g., servers). It focuses on KPIs such as runtime, resources needed and degree of parallelism due to the intrinsic nature of Cloud-based solutions. The workloads contained in the benchmarking suite range from graph analytics to media stream and is being integrated into Google's PerfKitBenchmarker.[10] The technology stack supported by CloudSuite includes Hadoop and Spark.

4 Summary

The Hobbit consortium has monitored the development of benchmarking standards over the last four years (including the time of the proposal writing). While a plethora of benchmarking platforms have seen the light of day, no standards for benchmarking Linked Data exist to date. A number of platforms are often used in industry and academia, including Amplab's benchmark[11] for data analytics, TPCx-HS[12] for SPARK, some of the LDBC datasets which pertain to social network working loads[13] and

[6] https://www.force11.org/group/fairgroup/fairprinciples
[7] https://github.com/intel-hadoop/HiBench
[8] http://prof.ict.ac.cn/BigDataBench/
[9] http://cloudsuite.ch/
[10] https://github.com/GoogleCloudPlatform/PerfKitBenchmarker
[11] https://amplab.cs.berkeley.edu/benchmark/
[12] http://www.tpc.org/tpcx-hs/
[13] http://ldbcouncil.org/

BigDataBench. Most frameworks support Hadoop and SPARK as Big Data technologies. FLINK is regarded by many as important but is not yet supported by a large number of platform. Docker and Docker Swarm are often used to package software and data solutions to ensure an easy deployment and ultimately increase the number of users.

The lack of standards in benchmarking Linked DAta means that this gap is still to be filled. The Hobbit platform is designed to achieve exactly this purpose. In addition to targeting an adoption in more domains through the Hobbit association formed in the Big Data Value Association, the Hobbit consortium has began interacting with the Standardization Group (i.e.g, Special Group 6 of Task Force 6) at BDVA to elucidate questions all around benchmarking standards for Europe.[14]

References

[1] Christian Bizer and Andreas Schultz. The Berlin SPARQL Benchmark. *Int. J. Semantic Web Inf. Syst.*, 5(2), 2009.

[2] David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June (Paul) Hsu, and Kuansan Wang. ERD 2014: Entity recognition and disambiguation challenge. *SIGIR Forum*, 2014.

[3] Sheriffo Ceesay, Adam David Barker, and Blesson Varghese. Plug and Play Bench : simplifying big data benchmarking using containers. In *2017 IEEE International Conference on Big Data*, 2017.

[4] Felix Conrads, Jens Lehmann, Muhammad Saleem, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. IGUANA: A generic framework for benchmarking the read-write performance of triple stores. In *ISWC*. 2017.

[5] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *22nd World Wide Web Conference*, 2013.

[6] Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, and Tzanina Saveta. Instance matching benchmarks in the era of linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2016.

[7] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. Bigbench: towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208. ACM, 2013.

[8] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 1197–1208, New York, NY, USA, 2013. ACM.

[9] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. LUBM: A Benchmark for OWL Knowledge Base Systems. *J. Web Sem.*, 3(2-3), 2005.

[10] Mohamed Morsey, Jens Lehmann, Sören Auer, and Axel-Cyrille Ngonga Ngomo. DBpedia SPARQL benchmark - performance assessment with real queries on real data. In *The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I*, pages 454–469, 2011.

[14] https://www.european-big-data-value-forum.eu/program/benchmarking/

[11] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. A survey of current link discovery frameworks. *Semantic Web*, (Preprint), 2015.

[12] Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. GERBIL – Benchmarking Named Entity Recognition and Linking Consistently. *Semantic Web Journal*, (Preprint):1–19, 2017.

[13] Muhammad Saleem, Qaiser Mehmood, and Axel-Cyrille Ngonga Ngomo. FEASIBLE: A feature-based SPARQL benchmark generation framework. In *14th International Semantic Web Conference*, 2015.

[14] Michael Schmidt, Thomas Hornung, Georg Lausen, and Christoph Pinkel. SP2Bench: A SPARQL performance benchmark. In *International Conference on Data Engineering (ICDE)*. IEEE, 2009.

[15] Beth M. Sundheim. Tipster/MUC-5: Information extraction system evaluation. In *Proceedings of the 5th Conference on Message Understanding*, 1993.

[16] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, 2003.

[17] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V Vasilakos. Big data analytics: a survey. *Journal of Big data*, 2(1):21, 2015.

[18] George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*, 2012.

[19] Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. Question answering over linked data (QALD-5). In *CLEF*, 2015.

[20] Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrad, Jonathan Huthmann, Axel-Cyrille Ngonga-Ngomo, Christian Demmler, and Christina Unger. Benchmarking question answering systems. *Semantic Web Journal*, (Preprint), 2018.

[21] Ying Zhang, Minh-Duc Pham, Oscar Corcho, and Jean-Paul Calbimonte. SRBench: A Streaming RDF/SPARQL Benchmark. In *International Semantic Web Conference (ISWC)*, volume 7649 of *Lecture Notes in Computer Science*. Springer, 2012.