HOBBIT

Collaborative Project

Holistic Benchmarking of Big Linked Data

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

Deliverable 1.3.2
Final Association Mission Statement and Business Scenarios

Dissemination Level	Public
Due Date of Deliverable	Month 34, 30/09/2018
Actual Submission Date	Month 34, 30/09/2018
Work Package	WP1 - Requirements Elicitation and Community Building
Task	T1.3
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	13

Abstract: This deliverable describes the final mission statement of the HOBBIT association. As per the review meeting, the previous plans for the association were amended. The HOBBIT association was integrated into the special group 7 of task force 6 of the Big Data Value Association. The first set of business cases pinpointed in the previous version of the deliverable remained unchanged. However, the association setting was altered. The details of this setting are described herein.

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
0.0	14/09/2018	First draft	Axel-Cyrille Ngonga Ngomo (InfAI)
0.2	21/09/2018	Revision of first draft	Henrik Oppermann (USU)
0.3	24/09/2018	Updates and corrections	Axel-Cyrille Ngonga Ngomo (InfAI)
1.0	29/09/2017	Quality checks	Eliza Shaw (InfAI)
1.0	30/09/2017	Final version submitted	Nadine Jochimsen (InfAI)

Author List

Organization	Name	Contact Information
InfAI	Axel-Cyrille Ngonga Ngomo	ngonga@infai.org

Executive Summary

This deliverable presents the final version of the mission statement of the HOBBIT association. The association was designed to be as complementary as possible to existing organizations to ensure that synergies with other associations can be achieved. To maximize its impact, it was established within the Big Data Value Association in collaboration with DataBench, another Big Data benchmarking project of the European Commission. The five preliminary business cases remained unchanged as they reflected discussions with potential end users at events such as EDF, Apache Big Data, ESWC and further HOBBIT and DataBench events as well as with technology vendors in and outside of the consortium.

Contents

List of Tables

1 Introduction

During the second half of the HOBBIT project, the consortium has continued pushing towards establishing itself as the provider of a benchmarking platform for industry and academia with a focus on Big Linked Data technologies. In accordance with the suggestions of the commission, the consortium sought an arrangement with the Big Data Value Association to continue the mission of the project after its completion. These efforts culminated in the formal integration of HOBBIT into a novel Special Group (SG7) of the task force 6 of BDVA. SG7 has the mission of leading data benchmarking in Europe from the technical perspective, which is at the core of TF6's activities.

In the following, we detail the structure and mandates of SG7.

2 Motivation and Rationale

Big Data is one of the key assets of the future. However, the cost and effort required for introducing Big Data technology in a value chain is significant. Mastering the creation of value from Big Data will enhance European competitiveness, result in economic growth and jobs and deliver societal benefit. It is thus of utmost importance to reduce the costs and hurdles required to introduce Big Data processing into the European industry. A key step towards abolishing the barriers to the adoption and deployment of Big Data is to provide European companies with open benchmarking reports that allow them to assess the fitness of existing solutions for their purposes. However, achieving this goal demands:

1. The deployment of benchmarks on data that reflects reality within realistic settings.
2. The provision of corresponding industry-relevant key performance indicators (KPIs).
3. The computation of comparable results on standardized hardware.
4. The institution of an independent and thus bias-free organization to conduct regular benchmarks and provide the European industry with up-to-date performance results.

It is also a motivation that the technical benchmarks will provide a foundation for the better analysis of business level benchmarks and KPIs related to the adoption and usage of big data technologies. For this there will be an interaction with Business focused TFs/SGs in BDVA.

3 Proposers and Sponsors

We collaborated with DataBench[1] to set up the association. The proposers of the association were as follows:

[1] https://www.databench.eu/

Organisation	Name	Email	Status
InfAI	Axel-Cyrille Ngonga Ngomo	ngonga@infai.org	Lead
IMEC	Gayane Sedrakyan	gayane.sedrakyan@ugent.be	
FORTH	Irini Fundulaki	fundul@ics.forth.gr	
SINTEF	Arne J. Berre	Arne.J.Berre@sintef.no	Lead
Univ. Frankfurt	Todor Ivanov	todor@dbis.uni-frankfurt.de	
ATOS	Tomas Pariente Lobo	tomas.parientelobo@atos.net	
Fondazione Politecnico di Milano	Barbara Pernici	barbara.pernici@polimi.it	

4 Initial Members

The initial members of the association are as follows:

1. SINTEF
2. Frankfurt University
3. ATOS
4. IDC
5. Fondazione Politecnico di Milano
6. IMEC
7. InfAI
8. FORTH
9. NCSR Demokritos
10. FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV
11. USU Software AG
12. OPENLINK Group Limited
13. AGT GROUP (R&D) GMBH
14. TOMTOM POLSKA SP ZOO

We plan to grow the group iteratively through the BDVA events.

5 Goals

5.1 Activities

- Provide benchmarks, key performance indicators, benchmarking tools and services for the independent and repeatable benchmarking of big data technologies
- Facilitate the systematic evaluation, improvement and objective comparison of scalable big data solutions
- Perform generalization of knowledge from open-source benchmarking technologies
- Detect potential use cases and categories of users
- Detect potential synergies with benchmarking organizations, other big data benchmarking activities
- Perform requirement specifications from the association
- Produce open benchmarking reports

5.2 Targets

- Synergies, use case and datasets for big data benchmarks to enhance benchmarking framework and domains
- Ensure synergy of results from Big Data PPP Benchmarking projects like HOBBIT and DataBench related to the requirements and needs of the BDVA members and the Big Data community in general
- Promote the use of the HOBBIT framework for linked data, and also consider this as input for benchmarking of other big data types
- Generalized best practices, guidelines and standards to be offered as tutorials and support for the community

6 Business Cases

The initial business use cases proposed by the HOBBIT consortium were regarded as fitting for the association. Currently, five main business cases are foreseen (see D1.3.1), of which an overview is given in Table 1.

6.1 Smoke Tests

Smoke tests (also known as sanity tests) [1] are commonly defined as simple tests designed to reveal failures of such severity that they can impede a piece of software from running accurately. A common example for the need of such tests is result completeness. A large number of tools guarantee that they return complete results w.r.t. to the input of the user (e.g., a SPARQL query). However, the completeness claims of many tools have been shown to be wrong (see, e.g., [2]). The idea behind this

use case is to define a series of smoke tests that can be used along the Linked Data lifecycle and lead to a software certification. With a "Checked by HOBBIT" logo, software applications could guarantee that they pass all the tests in a relevant suite. Establishing such certifications would allow the promotion of the HOBBIT brand at low costs and attract a large number of interested parties to see the other offers of HOBBIT. We are considering whether such a certification could be made available against a nominal fee used to maintain the tests and update them regularly. The regular updates are deemed important, as they would prevent engines from displaying behavior tailored towards the smoke tests (see Volkswagen incident). HOBBIT's large number of datasets will be critical in this sense to ensure that the smoke tests present the tools to be evaluated with a variety of tasks.

The **strengths** of this business case lie at hand. While the W3C and a few code bases provide simple tests for some of the Semantic Web functionality (e.g., checking the outputs of RDF libraries for whether they conforms to specifications), an exhaustive library of smoke tests for the Linked Data lifecycle does not exist. However, such a library would clearly be of benefit for application developers as it would allow tools they make available on the market to be certified. The library could be easily built to be one-size-fits-all and to rely exclusively on standards (SPARQL, HTTP, XML, etc.). Therewith, it would be a "cookie-cutter"-type product. Finally, the smoke tests could be derived directly from the HOBBIT benchmarks, making the product easy to build.

The **weaknesses** of this business case lie in its commercial viability. Given that one would basically sell a brand, it would be important for the HOBBIT branding to be known and accepted. Hence, the acceptance of the branding would depend on the efforts made during and after the project to popularize the HOBBIT brand. This will most probably be combined with financial investments that will have to be carried before revenue can be made. Still, the **opportunites** linked to this case are clear: Implementing this business case would establish HOBBIT across the whole Linked Data community as the reference organization for checking the viability of implementations. Hence, it would allow to secure a large number of members, who would want to influence the smoke tests through their membership. This large number of members would ensure that even more relevant smoke tests are carried out and that corner cases are not forgotten, hence leading to a self-funded ecosystem of use cases and HOBBIT members.

6.2 On-Premise Benchmarking

Most companies have no interest in making their data and software tools available for benchmarking. Still, the interest in objective performance measurements grows constantly. This business use case targets this constellation by offering "on-premise" benchmarking services. The potential customers are two-fold:

1. data owners interested in the right tool for their particular data processing task and
2. solution providers interested in knowing how well they perform when compared with the competition.

For the first customer segment, the service offers would include

1. an agent going to the premises of the customers,
2. the installation the benchmarking platform and the corresponding benchmarks and available reference software,

3. the implementation of the necessary system adapters for the platforms of their choice to run with HOBBIT,
4. the execution of experiments and
5. the generation of reports (possible with suggestions for improvement).

The procedure for the second customer segment would be similar, with the slight difference that they would most probably be interested in the baselines (as reference data) and their own piece of software.

In all cases, the **strenght** of this business case is that it requires a significant amount of rare expertise that would only be available to the consortium. This makes the use case at hand difficult to implement for potential competitors at the same costs as for the HOBBIT association. The need for specific functionality (the system adapters) makes the use case more attractive from a financial standpoint. However, it also means that a significant amount of work would have to be performed for each business case of the type. This can still be regarded as an advantage, as a series of such cases would significantly improve the ecosystem around the HOBBIT platform and many adapters could be reused.

The **weaknesses** of the use case at hand are linked to the need for tailored solutions for each of the instantiations of the use case. Still, the **opportunities** clearly make this business case viable, as it would support the growth of the HOBBIT ecosystem across different systems. Moreover, attractive deals linked to HOBBIT memberships would support the extension of the member base for the association.

6.3 Challenge Organization

Organizations such as Kaggle have shown that the organization of challenges can be a viable and attractive business model. In this business use case, we build upon the idea behind Kaggle by supporting customers of the association in the creation of business-relevant challenges: Data-driven companies are constantly faced with challenges that are difficult to solve in-house or where even slight performance improvements could be turned into significant financial gains (e.g., a reduction of the query runtimes of distributed query engine can save space, time and energy). Such organizations often have to buy experts' knowledge without any guarantees of results. HOBBIT could support companies facing this need by offering to support the organization, development, deployment and management of challenges. For the organizations, it would mean the provision of datasets and of benchmark specifications to be used. Baselines could also be provided. The HOBBIT consortium would manage the process from the formulation of the tasks to the selection of winners, including all intermediary steps.

The **strengths** of this business case lie in the business model behind it being known to be viable. Organizations such as Kaggle have shown that supporting the benchmark process can be turned into a profitable endeavor. The platform and the benchmarks build a solid foundation for showcasing the abilities and strengths of the association to potential customers. As in the first use case, this however demands some investments into the HOBBIT brand, which is one of its **weaknesses**. The **opportunities** connected with this business case are neertheless immense, as establishing HOBBIT as the reference organization for the design and deployment of challenges should be able to sustain the organization financially. Moreover, the number of members could be increased by offering reduced fees to association members, a feature that has been foreseen since the beginning of the project.

6.4 Benchmarking Output

Consultants such as Gartner[2] have built an important part of their business on the distribution of documents pertaining to the performance of frameworks and corresponding prognoses. While the HOBBIT association does not plan to reach the business volume of such organizations, the uptake of Linked Data in industry makes the provision of reports on system performances, current bottlenecks and recent advances a potentially viable pillar for the association. This is especially of interest because a significant portion of these reports can be generated automatically but still demands the special expertise of the consortium to be assigned the right interpretation.

The **strenghts** of this business case lie in its being one-size-fits-all for each step of the Linked Data lifecycle. Moreover, a large proportion of the generation process can be automated, making it a low-cost product for the association. The necessary expertise to assign the results generated a correct meaning also makes this particular case difficult to implement for other organizations, thus making the potential competition limited. The **weaknesses** of the business case are again related to branding. The association must be well known amongst practitioners for the reports to attain a high monetary value and be bought regularly. The **opportunity** behind this business case is still clear as there are no associations that (1) focus on the whole of the Linked Data lifecycle while (2) aiming to provide insights on the performance of a large number of tools and solutions.

6.5 Tutorials

Tutorials at international conferences and data science meetups will target a broad audience (faculty staff, professional industry educators, practitioners, researchers, students, solution providers and technology users) that want to obtain/improve skills in the domain of big linked data and benchmarking methodologies/solutions. The goal of the tutorials will be to provide practical guidance and hands-on experience on big linked data benchmarking and the use of the HOBBIT platform for research, development, teaching/training purposes. The tutorial sessions will cover the state-of-the-art, theory, review of different technologies, solutions and mimicking algorithms used for big data benchmarking, setup of the HOBBIT benchmarking environment and hands-on experience. The effectiveness of the proposed method and scientific results based on experimental validation will be briefly introduced. The tutorials will also review the challenges and trends for future research and development in this domain. The material used for the tutorials will be offered to the participants as a free download.

The elaborated material will be used to promote new business cases, e.g. online courses such as MOOCS, Coursera trainings, etc. The **strenghts** of this business case lie in the possibility of reaching a broad/diverse audience with the goal to engage them through hands-on experience and relevant starter material for research and development. The **weaknesses** include the fact that revenues from this business case will mostly depend on the number of the subscribed participants and thus the strategy to attract/engage potential participants.

7 Conclusion

In this document, we presented a preliminary version of the mission statement of the HOBBIT organisation. In addition, we presented a brief and preliminary compilation of potential business cases for the HOBBIT association. Throughout the business use cases, the need for HOBBIT to be established as a reference platform became clear. Hence, over the next 18 months, we will intensify our efforts to

[2] http://gartner.com

make the value of the HOBBIT platform and of the HOBBIT results clear to industry and academia. The association will be built up with the aim of gathering members that are not in the consortium but are willing to push the development of high-performance Linked Data driven solutions.

Use Case	Strengths	Weaknesses	Opportunities
Smoke Tests	Easy to implement; portable; cookie-cutter (can be employed on any application with standard I/O, e.g., triple stores); preliminary benchmark results for functionality tests	Need for known brand; preliminary investments	Currently not available (open market); Easy to deploy as service even for remote applications
On-premise benchmarking; Development of domain-specific benchmarks	Tailored tests for dedicated application; benchmarking on premise ensure security of data and is thus commonly accepted by system developers; companies have full control over the setup; high remuneration potential; rare expertise available to the consortium; access to datasets ensures access to novel markets	Potentially small number of customers; need for HOBBIT association to be known	Few services of the sort available at the moment; Tedious
Challenge organization	Existing business model (Kaggle); cookie-cutter through platform, only need to gather relevant data and tests from companies; potentially high remuneration; access to experts in different domains and opening for new business models (e.g., expert search)	Potential competition if successful (Kaggle, etc.)	Access to novel markets; access to supplementary datasets
Benchmarking output	Already available through challenges; derivation of novel insights across domains; consultancy + access to insights derived from raw challenge data	Need for critical mass of users	Interest from other communities
Tutorials	Already available platform, datasets, research papers on the methodologies, mimicking algorithms, scientific findings and overview of challenges for research and development	Need for reachout strategy to engage necessary number of participants	Possibility to reach a broad audience, engaging through hands-on experience

Table 1: Summary of preliminary business cases

References

[1] Steve McConnell. Daily build and smoke test. *IEEE software*, 13(4):144, 1996.

[2] Muhammad Saleem, Yasar Khan, Ali Hasnain, Ivan Ermilov, and Axel-Cyrille Ngonga Ngomo. A fine-grained evaluation of sparql endpoint federation systems. *Semantic Web*, 7(5):493–518, 2016.