LDBC

Collaborative Project

FP7 – 317548

D1.1.7 Final Benchmarks Integration And Release

Coordinator: Venelin Kotsev (ONTO)
With contributions from : Vladimir Alexiev (ONTO), Arnau Prat (UPC), Alex Averbuch (NEO)

Quality reviewer: Josep Lluís Larriba Pey (UPC)

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	M30
Actual delivery date:	M30
Version:	1.0
Total number of pages:	18
Keywords:	LDBC, SPB, SNB, Semantic Publishing Benchmark, Social Network Benchmark, Benchmarking Ontology, benchmark, RDF, graph, database

Abstract

This document reports on the final integration and release of the benchmark software developed by the LDBC Council. The two benchmarks: The Semantic Publishing Benchmark and The Social Network Benchmark have been adopted by the LDBC Council and have been made available to the public.

Descriptions of current status and latest improvements of both benchmarks as well as their location at public software repositories have been given. Also description of benchmarking ontology developed for the purposes of storing and presenting benchmark results has been provided.

Executive summary

This document reports on the final integration and release of the LDBC benchmarks. Two benchmarks have been developed and adopted by the LDBC Council - The Semantic Publishing Benchmark and The Social Network Benchmark. They have been published and made open to the public.

Access to Benchmark Software section provides general information on how to access and download the software.

Next two sections provide details about current state and final integration of the benchmark software - their latest improvements and status.

Final section of the document provides information about the LDBC's benchmarking ontology developed for the purposes of representing and storing the benchmarks results.

Document Information

IST Project Number	FP7 - 317548	**Acronym**	LDBC
Full Title	LDBC		
Project URL	http://www.ldbc.eu/		
Document URL	http://www.ldbc.eu:8090/display/PROJECT/Deliverables/		
EU Project Officer	Carola Carstens		

Deliverable	**Number**	D1.1.7	**Title**	Final Benchmarks Integration And Release
Work Package	**Number**	WP1	**Title**	Common Benchmark Methodology

Date of Delivery	**Contractual**	M30	**Actual**	M30
Status	version 1.0		final □	
Nature	prototype □ report ☑ dissemination □			
Dissemination level	public ☑ consortium □			

Authors (Partner)				
Responsible Author	**Name**	Venelin Kotsev	**E-mail**	venelin.kotsev@ontotext.com
	Partner	ONTO	**Phone**	+359 889 955 288

Abstract (for dissemination)	This document reports on the final integration and release of the LDBC benchmarks. Two benchmarks have been developed and adopted by the LDBC Council - The Semantic Publishing Benchmark and The Social Network Benchmark. They have been published and made open to the public. Access to Benchmark Software section provides general information on how to access and download the software. Next two sections provide details about current state and final integration of the benchmark software - their latest improvements and status. Final section of the document provides information about the LDBC's benchmarking ontology developed for the purposes of representing and storing the benchmarks results.
Keywords	LDBC, Semantic Publishing Benchmark, Social Network Benchmark, SPB, SNB, Benchmarking Ontology, benchmark, RDF, graph, database

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
23.03.2015	0.1	Venelin Kotsev	Initial version of the document
26.04.2015	0.2	Venelin Kotsev	Second version of the document
01.04.2015	0.3	Venelin Kotsev	Third version of the document
06.04.2015	0.4	Venelin Kotsev	Fourth version of the document
07.04.2015	1.0	Venelin Kotsev	Final version of the document

Table of Contents

List Of Figures

List Of Tables

Abbreviations

API - Application Programming Interface
LDBC - Linked Data Benchmark Council
PDF - Portable Document Format
RDF - Resource Description Format
SNB - Social Network Benchmark
SPB - Semantic Publishing Benchmark
SUT - System Under Test
URL - Uniform Resource Locator
VCS - Version Control System

1 Introduction

This deliverable focuses on the final results coming from the technical work packages of the collaborative project: LDBC (FP7 - 317548).

Two benchmarks have been released by the LDBC Council:

- LDBC Semantic Publishing Benchmark (LDBC-SPB) - testing the performance of RDF database systems
- LDBC Social Network Benchmark (LDBC-SNB) - testing the performance of graph-like database systems

Both benchmarks have been adopted by the LDBC Council and have been made available to the public on the Internet. They have been designed to test different types database technologies (RDF and Graph). Nevertheless intention of the LDBC Council has been to keep certain development standards by following common practices and methodologies. Those common practices have been implemented not only in software but also in benchmarks' documentation. Intention to facilitate the users as well as the contributors of the benchmarks has been a primary goal of the consortium and it reflects on all aspects of the design and documentation.

Following section will give a general overview on how to access the benchmark software. Next sections on this document provide descriptions of latest improvements and current status of both benchmarks. Also details about LDBC's benchmark ontology developed for storing and representing benchmark results have been given in the last section of this document.

2 Access to Benchmark Software

The LDBC Council has provided public access to all benchmark software and related tools and documentation. We have chosen a public source code repository GitHub [1] as one of the largest and most popular locations for sharing and collaboration on open-source software projects.

All of LDBC's software modules have been located in a common namespace called "LDBC" on GitHub - it is also a starting point to benchmarks' related materials and software. Accessing and downloading LDBC's benchmark modules requires no additional software but a simple web-browser. Of course using specialized software tools (e.g. a client for the Git [2] Version Control System) would give the advanced user additional capabilities e.g. tracking of versions, fixes, new features, etc which can be useful in certain cases.

Each of the modules in LDBC's namespace have integrated quick-reference documentation for introducing the benchmark user with the software (e.g. short description, build, install, quick configuration, etc.) without the cost of spending much time in reading the detailed documentation. That quick-reference documentation is a part of each module and can be accessed from each module's entry point.

Figure 1 shows the user interface of GitHub and in particular the entry point for the Semantic Publishing Benchmark module accessed from a web-browser.

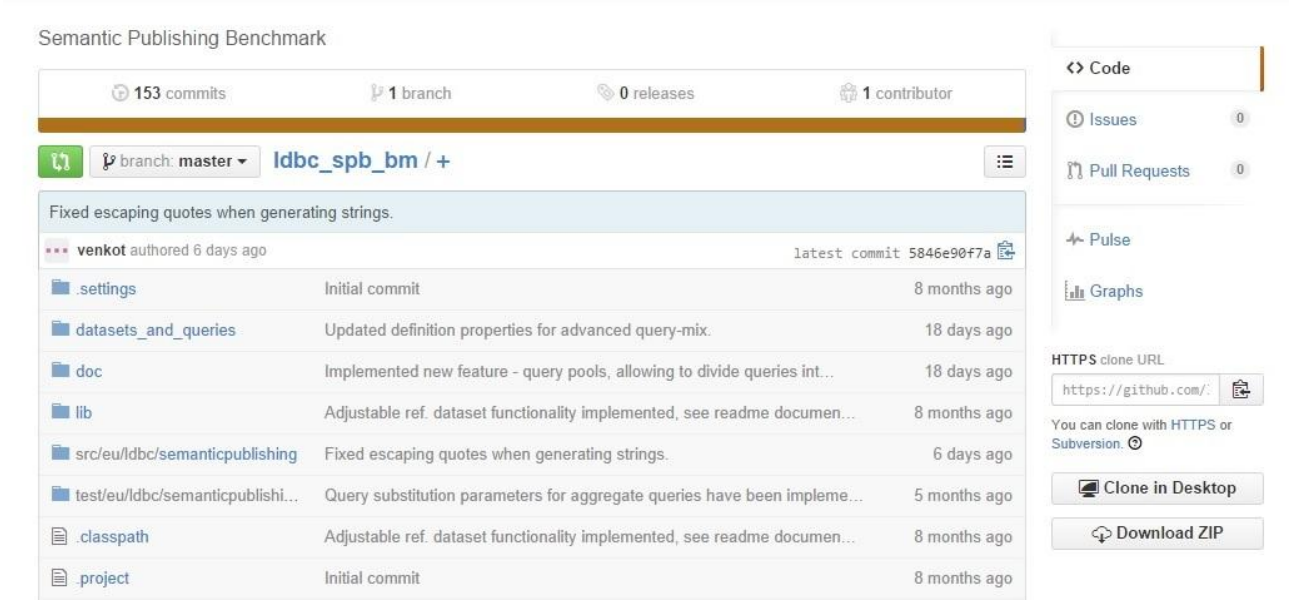

Figure 1: LDBC-SPB's repository entry point on GitHub

More information about GitHub has been given in deliverable D1.1.6 - "Initial Benchmarks Integration and Release" [3] where overview of GitHub's features and tools have been provided.

Following two sections provide details about current status of the benchmarks developed by the LDBC Council - LDBC Semantic Publishing Benchmark and LDBC Social Network Benchmark. New features implemented and their final integration.

3 LDBC Semantic Publishing Benchmark v2.0

LDBC Semantic Publishing Benchmark (SPB) is a benchmark for RDF databases inspired by the Media/Publishing industry. SPB simulates the interactions with RDF data such as constant updates and querying which would typically occur in every-day lifecycle of a publishing or media organization. Such lifecycle would typically consist of a constant stream of running updates - annotation or tagging entities with meta-data while consumption of that data occurs simultaneously. The term data in this case consists of a 1) reference data which is a set of curated pieces of information (entities) in certain area of our life - e.g. persons in politics, sports, organizations etc, and 2) meta-data - a layer of additional information which defines relations between entities and adds provenance information about them.

A new version 2.0 of the LDBC SPB has been released recently. That release has been inspired by continuous and on-going evolution of the semantic technologies in the publishing domain and can be described with a following list of improvements:

- **Enriched reference datasets:** data from DBpedia [4] adding entities for companies, events and persons which are interconnected and relate to each other in various aspects. Previous version of SPB has utilized around 127 000 statements for reference data, while current one uses about 23 million statements
- **Enriched Geonames locations:** locations from entire Europe have been added as a standard dataset, along with mappings for corresponding DBpedia locations (using owl:sameAs). A proper implementation of an RDF engine would be able to recognize both as an equivalent entities
- **Enriched Ontologies:** added Geonames ontology which is used for exploring the transitiveness of locations, e.g. certain location is a part of an administrative region
- **Optimized generation of data:** generated data now consists of real popular entities which are now existing in reference data resembling real datasets and inter-connectedness between them. RDFRank has been used to calculate popularity of entities vs. random selection of entities in previous version.
- **Enhanced use of inference:** currently SPB2.0 uses primitives from both RDFS and OWL. Added TransitiveProperty, SymmetricProperty and sameAs, allowing a more relevant use of potential of RDF systems
- **Changes in generated data:** generated meta-data (so called CreativeWorks) keeps the initial structure. Enhanced number of relations have been added, effectively tripling the number of references compared to ones in previous version of SPB, pushing further the abilities of RDF engines to handle effectively queries that utilize such relations
- **New queries:** two additional queries have been added which explore the relations of reference data as well as verify that proper inference by the RDF database is supported
- **Enhanced validation:** on top of existing validation mechanisms, additional layer of validation has been implemented which verifies that RDF engine can handle queries with more complex inference patterns correctly
- **Benchmarking ontology:** the benchmark driver can generate results that are conformant to LDBC's benchmarking ontology, thus allowing exploration various aspects of the benchmark results

The new version of SPB 2.0 builds on top of the existing one. It is an evolution in the direction of a larger and more inter-connected datasets; use of more complex inference patterns; deeper exploration of scalability.

SPB software also comes with a comprehensive set of documentation describing various aspects of the software and operation e.g.:

- **Requirements to run the Benchmark:** defines the requirements that a system under test must meet in order to run the benchmark
- **Datasets and Data Generator:** describes the reference datasets and ontologies used by the benchmark's Data Generator. Provides information on the data generation process
- **Instructions and Configuration:** detailed descriptions of configuration and definitions properties used to control the behaviour of SPB
- **Execution Rules:** defines the rules that would guarantee a valid benchmark result

A brief "how-to" documentation has also been added to help the new users quickly on how to configure, set-up and run the benchmark.

The complete reference documentation for the SPB software can be found at the *ldbc_spb_bm_2.0* repository (LDBC_SPB_v2.0.docx [5]).

Following Table 1 gives a short description of the Semantic Publishing Benchmark v2.0 software in terms of packaging and distribution:

GitHub Repository	**Description**
ldbc_spb_bm_2.0	Location: https://github.com/ldbc/ldbc_spb_bm_2.0 Contains: • Data Generator: produces synthetic output data at different scales and models real-world correlations in data • The benchmark test driver: generates and executes a workloads, measures performance, validates correctness and reports results • Documentation describing in full detail all aspects of the benchmark software • Reference datasets and ontologies required by the Data Generator for producing the synthetic output data

Table 1 : Description of components of The Semantic Publishing Benchmark v2.0

4 LDBC Social Network Benchmark

The LDBC Social Network Benchmark (LDBC-SNB) aims at being comprehensive benchmark setting the rules for the evaluation of graph-like data management technologies. LDBC-SNB is designed to have a plausible look-alike of all the aspects of operating a social network site, as one of the most representative and relevant use case of modern graph-like applications.

LDBC-SNB is designed to be flexible and to have an affordable entry point. From small single node and in memory systems to large distributed multi-node clusters have its own place in LDBC-SNB.

Recent improvements have been added to Data Generator, it has been completely refactored, by rewriting a major part of the code to make it more maintainable, readable, robust and fast [6]. Beside these refactorings, new features have been added to make it more appealing to the users, not only the benchmark implementers. Here is the list of new features:

- **Serialization:** added interfaces to allow users to implement their own serializers. This allows them to output data in the best suited format for their systems, even allowing to directly feed the data into the database instead of outputting it as a file
- **Data Generation:** Added interfaces to override the generation of degree distributions. The user has now been allowed to decide which degree distribution to use when generating the data. Built-in distribution generators have been provided e.g. Zeta, Geometric, etc., but anyone can implement its own. This is to promote the usage of DATAGEN beyond the scope of the benchmark. Of course, the benchmark rules state that to have a valid run one must use the Facebook Distribution
- **Parameter consistency:** Normalized parameter names to be more consistent
- **Partitioning of generated output:** Added the possibility to create partitions of the output files
- **Updates related to newest library components:** updated the code to use the latest versions of Hadoop: 2.6.0. This implies a more robust and reliable implementation
- **Usability:** simplified usage of data generator and simplified/improved the documentation to avoid having duplicated stuff

Improvements to SNB's queries:

- **New queries:** added seven short reads to be able to balance the ratio between reads and writes. Writes are very frequent, especially the larger the dataset. This implies that in order to keep a good ratio between reads and writes (like 80-20), complex reads have to be issued very frequently. The problem is that complex reads are complex, and keeping up with a good throughput with just them is difficult, making the benchmark less appealing. Furthermore, in a real system, users do not only perform complex queries but also short queries that retrieve small amounts of data.

- **Business Intelligence Queries:** 24 queries have been defined for the business intelligence workload, with their reference SQL implementations. They can be found in the following URL [7]

Improvements to SNB's benchmark driver:

- **Workload improvement:** with the new workload containing short reads now, once a complex read is issued, a sequence of short reads is performed in order to simulate the behavior of a user in the social network. There are two types of short reads sequences: person centric and message centric. Depending on the complex read, person or message centric sequences of short reads are issued. The number of sequences is determined by a probability, that decreases as long as the short read sequences are performed. This way, a better benchmark has been provided, that combines both complex and short queries, that rewards both throughput and latency, and that contains updates as in a real system.

- **Validation improvement:** added a more complete validation mode, which now includes both updates and short reads. This mode allows the user to both create a validation dataset and also validate another validation dataset. Passing a validation dataset is necessary to pass the audit process.

- **Validation dataset:** an official validation dataset has been uploaded and ready to be used to validate vendor implementations. This validation dataset includes both data in different formats so it can be used by different vendors [8]

- **Interactive workload query-mix:** the interactive query-mix has been improved based on extensive experiments performed on top of different vendor technologies, to determine the complexity of the different queries. From the conducted experiments, several scaling functions have been developed in order to, given a scale factor, determine the frequencies of the queries

As an on-going effort of LDBC to keep a uniform structure of all presented information - repositories have integrated "how-to" section. It allows users of the benchmarks to easily navigate through repositories and get basic knowledge of their content.

SNB software is also accompanied by a comprehensive set of documentation describing various aspects of the software and operation e.g.: Requirements to run the Benchmark, Datasets, Data Generator**,** Instructions and Configuration, Execution Rules (LDBC_SNB_v0.2.1.pdf [9]).

The Social Network Benchmark Software has been developed into five GitHub repositories, following Table 2 gives a short description of each:

Repository	**Description**
ldbc_driver	Location: https://github.com/ldbc/ldbc_driver A benchmark test driver - load testing tool that generates a workload, executes it, measures performance against and optionally validates correctness of query execution. Reports the results of the benchmark upon completion
ldbc_snb_datagen_0.2	Location: https://github.com/ldbc/ldbc_snb_datagen_0.2 A Data generator tool that produces synthetic output data used by the benchmark driver. Generated data mimics the characteristics of real data and is configurable for different scale sizes
ldbc_snb_implementations	Location: https://github.com/ldbc/ldbc_snb_implementations Contains implementations of the workload components for the test driver. Currently implementations have been added by two database vendors : Virtuoso [10] and Neo4J [11]
ldbc_snb_interactive_validation	Location: https://github.com/ldbc/ldbc_snb_interactive_validation This repository contains the instructions and files necessary to validate LDBC SNB Interactive workload with implemented database connectors
ldbc_snb_docs	Location: https://github.com/ldbc/ldbc_snb_docs Reference documentation describing in full detail all aspects of the benchmark software

Table 2 : Description of repositories for The Social Network Benchmark

5 Benchmarking ontology

A benchmarking ontology (BM) [12] has been developed. It defines a common schema for transforming results coming from all benchmarks into a common schema. Having result data described with that ontology schema allows utilizing the full potential of RDF systems for storing, querying and analyzing data in a more flexible way than common means provided by relational databases.

External ontologies have been studied and some inspiration has been taken from these ontologies or have been reused directly. They are relevant for the scope of areas such as describing the System Under Test (e.g. hardware and price, platform, database etc.), Benchmark definition, Dataset size, Results provenance and detailed log, result metrics etc.

BM ontology uses the RDF Data Cube Vocabulary [13] useful for publishing multi-dimensional data such as statistics in such a way that it can be linked to related data sets and concepts. The Data Cube vocabulary provides means to do this using the W3C RDF (Resource Description Framework) standard. The Data Cube vocabulary is a core foundation which supports extension vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional data sets.

The BM ontology can play several useful roles:

- Result gathering from several vendors (benchmark sponsors), allowing easy data integration and comparability
- Generation of partial Summary Reports and Full Disclosure Reports from captured RDF benchmark data. Most W3C Conformance reports and Implementation reports are now produced from EARL RDF using simple scripts.
- Easy accumulation of historic results for a specific database, enabling performance comparison and charting

The BM ontology and appropriate convertors that can capture benchmark run statistics in RDF can form a useful data integration environment that may help other benchmark authors beyond SNB and SPB.

The LDBC Benchmarking Ontology intends to capture the key information about a benchmark definition and benchmark run, covering the following areas:

- System Under Test:
  - Hardware And Pricing (to compute dollars per query)
  - Platform And Software, including Operating System
  - Database Make, Version, Configuration Parameters
- Benchmark Definition, Including:
  - Version, Parts (variants)
  - Pointers To Query Definitions
  - Data Generator Parameters And Version
  - Dataset Summary Description (e.g. Scale)
- Test Driver Parameters And Version
- Benchmark Run Provenance: Sponsor, Date Executed, Date Of Report, Date Approved
- Detailed Results Log: timed events that happened during a run
- Result Statistics About The Run (e.g. response time, throughput, at different percentiles, etc)

The total scope of the BM ontology is rather ambitious. The initial version of the BM ontology describes Result statistics, following partner priorities. We have covered:

- CUBE normalization (shortcut expansion) using SPARQL INSERT or Ontotext GraphDB rules
- Construction of a CUBE for capturing benchmark results, discussion of the applicable dimensions

- Examples of mapping SNB results, including a custom extension for capturing SNB-specific measures and attributes
- Sample queries for getting cube data

LDBC Benchmark ontology comes with reference documentation located at GitHub as well as with all necessary supporting files (e.g. Turtle samples, ontology definitions) [14]

Following Table 3 gives a short description of the benchmarking ontology - location and components:

Repository	**Description**
ldbc_bm_ontology	Location: https://github.com/ldbc/ldbc_bm_ontology Ontology implementation accompanied by required external sources and samples.

Table 3 : Description of the repository for LDBC's Benchmarking Ontology

6 Conclusions

This deliverable reports on the final benchmark integration and release of the benchmark software developed by the LDBC Council. It describes the current state of the benchmark software. The document also provides information on the benchmarking ontology implemented for the purposes of representing and storing LDBC's benchmark results. Both the Semantic Publishing and The Social Network Benchmarks have reached a consistent state which allows them to be used for testing performance and publishing official benchmark results both by end-users and database vendors.

Being open-source and encouraging the benchmarking community to actively contribute to both benchmarks gives the opportunity to constantly evolve and keep both benchmarks 'on track' with the future trends and innovations in the database field and in industry.

References

[1] : GitHub - a public source code repository : http://www.github.com

[2] : Git Version Control System - http://git-scm.com/

[3] : D1.1.6 - Initial benchmarks Integration and Release :

http://www.ldbc.eu:8090/download/attachments/1671227/LDBC_D1.1.6_v1.0_final.pdf

[4] : DBpedia: http://www.dbpedia.org

[5] : Reference documentation for the Semantic Publishing Benchmark :

https://github.com/ldbc/ldbc_spb_bm_2.0/blob/master/doc/LDBC_SPB_v2.0.docx

[6] : Repository location for new Data Generator : https://github.com/ldbc/ldbc_snb_datagen_0.2

[7] : SNB Business Intelligence workload :
http://wiki.ldbcouncil.org/display/TUC/Business+Intelligence+Workload

[8] : SNB Validation Dataset : https://github.com/ldbc/ldbc_snb_interactive_validation

[9] : Reference documentation for the Social Network Benchmark :

https://github.com/ldbc/ldbc_snb_docs/blob/master/LDBC_SNB_v0.2.1.pdf

[10] : http://www.openlinksw.com/

[11] : http://www.neo4j.org/

[12] : LDBC Benchmarking ontology : https://github.com/ldbc/ldbc_bm_ontology

[13] : The RDF Data Cube Vocabulary : http://www.w3.org/TR/vocab-data-cube/

[14] : Benchmark ontology documentation :

http://htmlpreview.github.io/?https://github.com/ldbc/ldbc_bm_ontology/blob/master/README.html