LDBC

Collaborative Project

FP7 – 317548

D1.1.6 Initial Benchmarks Integration And Release

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

1st Quality reviewer: Josep Lluís Larriba Pey (UPC)
2nd Quality reviewer: Alex Averbuch (NEO)

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	M24
Actual delivery date:	M24
Version:	1.0
Total number of pages:	24
Keywords:	LDBC, SPB, SNB, Semantic Publishing Benchmark, Social Network Benchmark, benchmark, RDF, graph, database

Abstract

This document reports on the initial 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 reached a state of ‘draft publication’ and have been made available to the public. Descriptions of the benchmark components, location at public software repositories and documentation have been given. Also definitions of the execution, reporting and auditing rules, defined by the DLBC have been provided.

Executive summary

This deliverable reports on the initial integration and release of the LDBC benchmarks software. Two benchmark software modules have been developed by the LDBC Council and have reached the state of ‘draft publication’ - The Semantic Publishing Benchmark and The Social Network Benchmark. They have been made available to the public on a public source code repository on the Internet.

Descriptions of the benchmark software components are given - access to the software, documentation, data generators, datasets, rules for auditing, reporting and execution.

Packaging and Access to Benchmark Software section describes the public source code repository on the Internet chosen by the LDBC Council to store the benchmark software projects. This section also provides information on the methods for accessing and downloading the software.

Next two sections give details about the general structure of the benchmark software (Documentation, Datasets, Data Generators) for The Semantic Publishing and The Social Network Benchmarks respectively.

Final section of the document provides information about the execution, reporting and auditing rules that the LDBC Council has defined in order to have consistent and reproducible results with both benchmarks.

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**	D1.1.6	**Title**	Initial Benchmarks Integration And Release
Work Package	**Number**	WP1	**Title**	Common Benchmark Methodology

Date of Delivery	**Contractual**	M24	**Actual**	M24
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 initial 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 reached a state of ‘draft publication’ and have been made available to the public. Descriptions of the benchmark components, location at public software repositories and documentation have been given. Also definitions of the execution, reporting and auditing rules, defined by the DLBC have been provided.
Keywords	LDBC, Semantic Publishing Benchmark, Social Network Benchmark, SPB, SNB, benchmark, RDF, graph, database

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
19.09.2014	0.1	Venelin Kotsev	Initial version of the document
22.09.2014	0.2	Venelin Kotsev	Updated version after first remarks from UPC
23.09.2014	0.3	Venelin Kotsev	Updated version after initial remarks from NEO
24.09.2014	0.4	Venelin Kotsev	Updated version after second remarks from NEO
26.09.2014	0.5	Venelin Kotsev	Final updates
01.10.2014	1.0	Venelin Kotsev	Final version

Table of Contents

List Of Figures

List Of Tables

Abbreviations

ACID	- Atomicity, Consistency, Isolation, Durability
API	- Application Programming Interface
FDR	- Full Disclosure Report
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 results coming from the technical work packages of the collaborative project: LDBC (FP7 - 317548).

Two benchmark software modules have been developed and 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 reached the status of ‘draft publication’ and have been made available to the public via the Internet.

The document will describe items related to each of the benchmark software modules : packaging (location in source code repository), access to the software, internal structure of the benchmark modules (datasets, data generators, documentation), reporting, execution and auditing rules defined by the LDBC Council.

While both benchmarks have been designed to test different types database technologies (RDF and Graph) - intention of the LDBC is to keep and develop both, following common practices and methodologies. Next sections in this document reflect that intention and provide a unified by structure information about each of those benchmarks.

2 Packaging and Access to Benchmark Software

The LDBC Council has decided to keep the benchmark software into a public source code repository GitHub [1]. Being one of the largest hosts for source code, GitHub provides a set of features vital for any software project like :

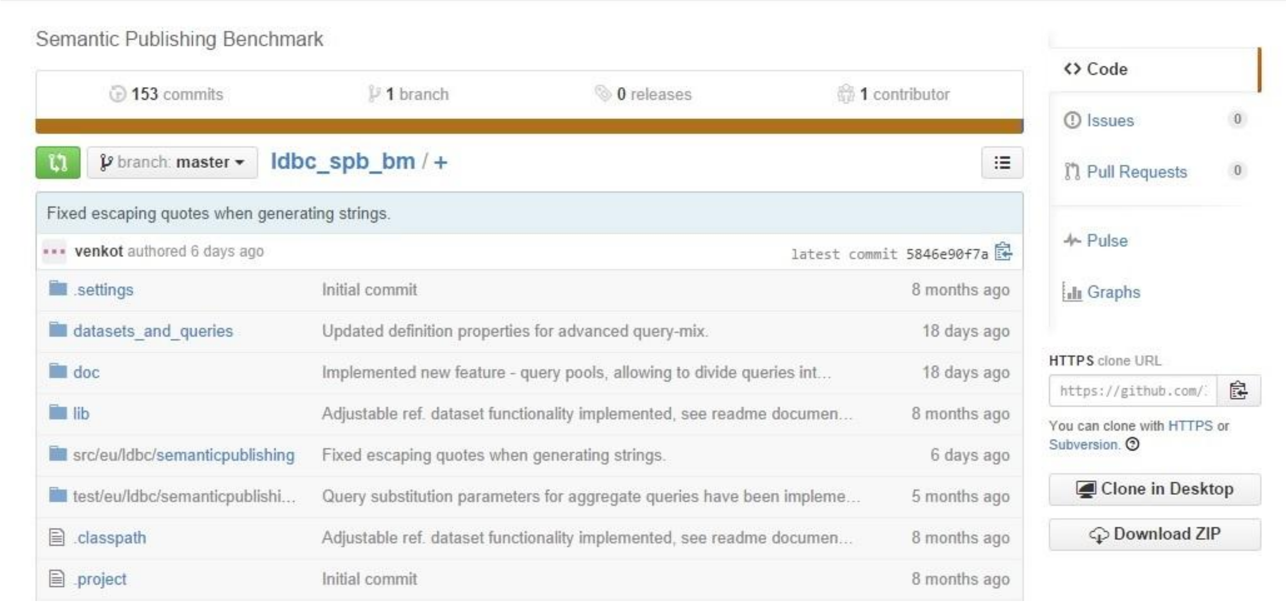
- **tools** - giving the ability to easy access and collaborate on a project
- **integrated issue tracking** - providing easier management of any obstacles that a software project can come upon during its life-cycle
- **collaborative code review -** provides means for collaboration between members of the project as well as the public
- **management of teams** - provides easy management for members and organisations
- **integration -** provides integration with various software and operating systems
- **public access** - access to repositories is public and free

The LDBC benchmark software modules have been located in GitHub at address : https://github.com/ldbc/.

A common namespace called '*ldbc*' has been set-up containing a number of repositories that host all developed software components where each component playing a specific role in the benchmark's test process, e.g. Data Generator produces synthetic large data, Datasets contain reference data used by the Data Generator component, Driver component executes a workload on a System Under Test (SUT) and measures its performance.

All of the software components can be downloaded from their corresponding repositories by using a web browser or by using an appropriate software tool.

The use of a web browser requires no additional software and the benchmark software can be downloaded by simply opening the corresponding repository's web page and selecting the 'download' option (Figure 1). As a result of that operation, a software package is downloaded as a single compressed file whose contents are ready to work with after extraction.



Semantic Publishing Benchmark

153 commits 1 branch 0 releases 1 contributor

branch: master ldbc_spb_bm / +

Fixed escaping quotes when generating strings.

venkot authored 6 days ago latest commit 5846e90f7a

.settings	Initial commit	8 months ago
datasets_and_queries	Updated definition properties for advanced query-mix.	18 days ago
doc	Implemented new feature - query pools, allowing to divide queries int...	18 days ago
lib	Adjustable ref. dataset functionality implemented, see readme documen...	8 months ago
src/eu/ldbc/semanticpublishing	Fixed escaping quotes when generating strings.	6 days ago
test/eu/ldbc/semanticpublishi...	Query substitution parameters for aggregate queries have been impleme...	5 months ago
.classpath	Adjustable ref. dataset functionality implemented, see readme documen...	8 months ago
.project	Initial commit	8 months ago

HTTPS clone URL

https://github.com/

You can clone with HTTPS or Subversion.

Figure 1 : GitHub user interface

Using specialized software tools (e.g. a client for the Git [2] Version Control System) for downloading software require some knowledge of the Git VCS but still the download can be achieved by executing a single line command e.g. :

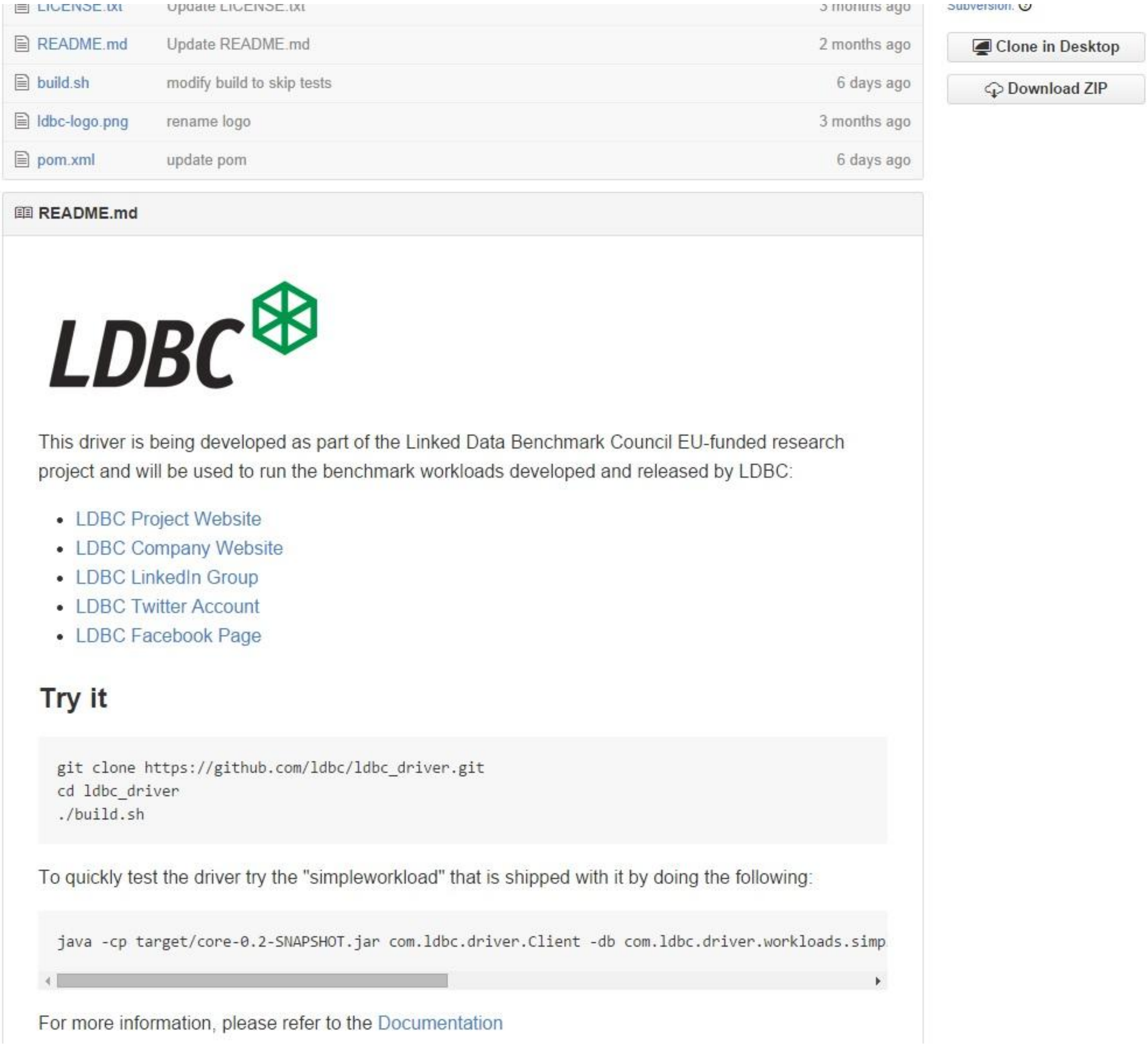
git clone https://github.com/ldbc/ldbc_spb_bm.git

The choice for a method to access the benchmark software depends on the type of user - for advanced users it is more appropriate to access repositories by using specialized software tools while regular users are not required to do so.

Once downloaded, each of the components contains a quick reference documentation for building, installation and use.

Furthermore, bundled with each of the benchmark software modules is a reference documentation which can serve as an in-depth source of information.

Quick reference instructions have been also provided at each of the repositories' web pages (Figure 2) following the common practice for documenting software projects hosted on public source code repositories.



LICENSE.txt	Update LICENSE.txt	3 months ago
README.md	Update README.md	2 months ago
build.sh	modify build to skip tests	6 days ago
ldbc-logo.png	rename logo	3 months ago
pom.xml	update pom	6 days ago

This driver is being developed as part of the Linked Data Benchmark Council EU-funded research project and will be used to run the benchmark workloads developed and released by LDBC:

- LDBC Project Website
- LDBC Company Website
- LDBC LinkedIn Group
- LDBC Twitter Account
- LDBC Facebook Page

Try it

```
git clone https://github.com/ldbc/ldbc_driver.git
cd ldbc_driver
./build.sh
```

To quickly test the driver try the "simpleworkload" that is shipped with it by doing the following:

```
java -cp target/core-0.2-SNAPSHOT.jar com.ldbc.driver.Client -db com.ldbc.driver.workloads.simp
```

For more information, please refer to the Documentation

Figure 2 : Quick reference documentation posted on GitHub repositories

Both benchmark software modules are using open software tools for automating their build process. Depending on the internal structure of the modules each automation tool is a more suitable choice for the purpose it serves than the other. Table 1 contains descriptions of those automation build tools.

Benchmark Software	**Build Automation Software Tool**
LDBC - Semantic Publishing Benchmark	Apache Ant [3] - a Java based build tool as a part of the Apache open-source project.
LDBC - Social Network Benchmark	Apache Maven [4] - a build automation tool for building and managing software projects and dependencies on third-party software libraries.

Table 1 : Description of used build automation tools

Following two sections provide descriptions of the benchmark software modules developed by the LDBC Council - LDBC Semantic Publishing Benchmark and LDBC Social Network Benchmark.

3 LDBC Semantic Publishing Benchmark

LDBC Semantic Publishing Benchmark is a benchmark for RDF databases inspired by the Media/Publishing industry. From a technology standpoint, the benchmark assumes that an RDF database is used to store the benchmark data and supports interactions with that data such as constant updates and querying.

The Semantic Publishing Benchmark Software has been developed in two components each hosted on separate GitHub repositories described in Table 2 :

Repository	**Description**
ldbc_spb_bm	The LDBC-SPB benchmark repository contains : • Data generator used to produce synthetic output data at different scales and models real-world correlations in that generated data • The benchmark test driver that generates and executes a workload against an RDF database, measures performance, validates correctness and reports results • Documentation describing in full detail all aspects of the benchmark software operations • Initial reference datasets and ontologies required by the data generator to produce the synthetic output data Location : https://github.com/ldbc/ldbc_spb_bm
ldbc_spb_optional_datasets	Additional reference datasets, used to further enrich the initial reference data stored in the LDBC-SPB benchmark repository. They are useful when generating a synthetic output data at larger scales, allowing to produce larger diversity in distributions of entities in data. Additional datasets : • Person data from DBPedia • Geo-locations from Geonames Location : https://github.com/ldbc/ldbc_spb_optional_datasets

Table 2 : Description of repositories for the Semantic Publishing Benchmark

3.1 Documentation

Reference documentation about the internals of each of the benchmarks has been written and stored in their public repositories. Contents of both reference documentations for the LDBC-SPB and LDBC-SNB have been made coherent by structure, thus allowing easier access and navigation. Slight deviations in both reference documents are possible due to the technological differences between the benchmarks, but the general structure of documents' content has been kept unified.

Following list describes key items in the Semantic Publishing Benchmark reference documentation :

- **Formal Definitions of the SPB Benchmark**
  - **Requirements to Run the Benchmark -** defines the requirements that a system under test must meet in order to measure its performance
  - **Data and Data Generator -** describes the reference datasets and ontologies used by the benchmark's Data Generator. Provides information on the data generation process e.g. steps in the data generation process, correlations that are modelled by the data generator, distributions of entities in generated data, serialization formats of generated data, statistics for various scale factors of generated data etc.
  - **Workloads** - provides information on the workload performed by the SPB consisting of simultaneous execution of 'editorial' (insert, update and delete operations) and 'aggregation' agents (aggregation, search, geo-spatial, faceted search operations etc.). Also provides information on the so called 'choke points' presented by the workload queries which are testing capabilities of the SUT for dealing with various technical challenges - a list of 'choke points' along with their distribution in the queries has been provided; information about query substitution parameters has been given

- **Instructions**
  - **Configuration** - detailed descriptions of configuration and definitions properties has been provided - used to control the benchmark software
  - **Running the Benchmark** - explains how to run various phases of the benchmark - e.g. generate data, validate query results, generate substitution parameters, start benchmark phases e.g. warm-up and the benchmark run itself
  - **Gathering Results -** describes the results produced by the LDBC-SPB, values that are reported and types of files that contain those results
  - **Validation** - provides information about validation of query results and conformance tests that the SPB performs

- **Appendices**
  - **Ontologies description** - provides descriptions of the ontologies (defining the relations of entities in data) used in the benchmark
  - **Query listings -** provides listings of all queries that generate the workload
  - **Samples of Benchmark Results** - a result sample produced by the benchmark run with information describing it

The reference documentation for the SPB software can be found in a PDF document located at subfolder ‘doc/’ of the *ldbc_spb_bm* repository (LDBC_SPB_v0.1.pdf [5]).

3.2 Datasets

Datasets are the initial reference data used by a data generator component to produce the synthetic large output data used in the benchmark run. It consists of ontologies and reference data related to certain domains from every-day life such as : politics, sports, news etc. Ontologies describe the relations between entities in reference data and the reference data contains information about those entities. Both ontologies and reference data have been provided by the BBC and contain entities for domains such as : UK politics, Sports - competitions and teams. Additional reference datasets have been added for further enriching existing data - geo-locations from a public data source : Geonames [6] and data about persons from DBPedia [7].

Description of the datasets used in the SPB are given in Table 2 :

Dataset	**Description**
UK Politics Data	Data about members of the UK Parliament. Data has been provided by the BBC.
Sports and Sports Teams Data	Data about sports and sports teams in football (International, UK and Scottish), Formula 1. That portion of reference data has been provided by the BBC.
Geo-locations Data	Provides data about geo-locations for all European countries. Data has been taken from the public data source - Geonames.
Person Data	Provides data about persons - birth dates, birth places, current occupation, etc.. Data has been taken from the public data source - DBPedia.
CreativeWork Ontology	Defines classes, sub-classes and properties of various types of creative work. Creative work (also called journalistic asset) is the meta-data created about entities found in reference datasets.
Company Ontology	Provides internal terminology about companies, product groups, etc.
Core Concepts Ontology	Gives core concepts about the entities and their relations e.g. persons, places, events, organisations etc.
Person Ontology	Describes people and their roles.
Provenance Ontology	Versioning and change log information for datasets.
CMS Ontology	Used for interpreting locators into various specialised content management systems.
Tagging Ontology	The ontology for connecting CreativeWork instances with concepts from domain ontologies.
Sports Ontology	The ontology for describing sports, competitions, events.
News Ontology	Describes the basic concepts that journalists can tag a Creative Work with.

Table 3 : Description of datasets used in Semantic Publishing Benchmark

All reference data and ontologies have been stored at the *ldbc_spb_bm* repository except the person data datasets from DBPedia and the additional geo-locations data from Geonames. They have been located in a separate repository : *ldbc_spb_optional_datasets*. The reason to use a separate repository is primarily for user's convenience because of the large sizes of those additional datasets.

More information on Datasets can be found in the SPB reference documentation (LDBC_SPB_v0.1.pdf)

3.3 Data Generator

SPB’s Data Generator is an integral part of the LDBC-SPB software module and is located at the : *ldbc_spb_bm* repository. It uses a set of input reference data - reference datasets and ontologies described previously and generates a synthetic output data at various scales. Generated output data consists of meta data about the entities found in reference datasets - the so called 'Creative Works' or journalistic assets. Each Creative Work contains various properties e.g. title, description, date of creation and modification, reference to other entities etc. The data generator attempts to reproduce realistic models of correlations in generated synthetic data by following certain patterns discovered in the real-world data instances.

Detailed information on the Data generator and data generation process has been provided in the SPB reference documentation (LDBC_SPB_v0.1.pdf), following list outlines key features of the SPB Data generator :

- **Scalable, Deterministic, Usable, Realistic** - the data generator produces scalable in size synthetic data; deterministic - for each run, the same consistent data output is generated; usable - easy to configure and run; realistic - generated data contains correlation models found in the real-world data
- **Data generation process** - defined steps are followed when producing the synthetic output data including the modelling of correlations between entities
- **Input data** - uses reference data and ontologies with distributions of entities found in the real-world datasets
- **Output data** - various data serialization formats are available allowing integration with the majority of RDF database systems, various configuration properties and allocations of the Data generator have been used to control the generation of synthetic output data
- **Parallel data generation** - data generation processes can be executed simultaneously for a faster data output
- **Scale factors** - synthetic output data can be generated at different scale factors

Quick reference information on how to operate with the Data generator can also be found on the main web page of the *ldbc_spb_bm* repository.

4 LDBC Social Network Benchmark

LDBC's Social Network Benchmark tests various functionalities of systems used for graph-like data management. It consists of three sub-benchmarks, or workloads, that focus on different functionalities : Interactive, Business Intelligence and Graph Analytics.

The Social Network Benchmark Software has been developed into four components each hosted on a separate repository in GitHub. Table 3 provides descriptions of the components and their repositories :

Repository	**Description**
ldbc_driver	A benchmark test driver - load testing tool that generates a workload, executes it and measures performance against a database system and optionally validates correctness of query execution and reports the results of the benchmark upon completion Location : https://github.com/ldbc/ldbc_driver
ldbc_snb_datagen	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 Location : https://github.com/ldbc/ldbc_snb_datagen
ldbc_snb_implementations	Contains implementations of the workload components for the test driver. Currently implementations have been added by two database vendors : Virtuoso [8] and Neo4J [9] Location : https://github.com/ldbc/ldbc_snb_implementations
ldbc_snb_docs	Reference documentation describing in full detail all aspects of the benchmark software Location : https://github.com/ldbc/ldbc_snb_docs

Table 4 : Description of repositories for The Social Network Benchmark

4.1 Documentation

The Social Network Benchmark has been distributed with a reference documentation which provides detailed information on the benchmark internals. As mentioned previously, the reference document follows a unified document content structure with possible slight deviations specific for each of the benchmarks.

Following list describes the key items in the reference documentation of the LDBC-SNB :

- **Formal Definitions of the SNB Benchmark**
  - **Requirements to Run the Benchmark -** provides definitions of the requirements that the database system must meet in order to be capable of running the benchmark
  - **Data and Data Generation** - provides descriptions of the data types used; data schema - defining the structure of the data in terms of entities and their relations as well as descriptions of each type of entity; data generation - information about the data generation process and modelling of correlations in data; datasets - information about the set of dictionary and resource files used in the data generation process; supported serialization formats for the generated synthetic output data; definitions of the scale factors for generated synthetic data
  - **Workloads** - provides information on the workload that the LDBC-SNB benchmark driver performs : gives descriptions on the various 'choke points' that queries are stressing the SUT with; defines a query description format so that more than one implementation of the query set can be used; gives a list of query descriptions in plain language along with query parameters and expected results; describes the substitution parameters that are generated and used during the benchmark run; describes the load definition of the benchmark test - how execution of queries of different complexity have been interleaved by empirically determined values

- **Instructions**
  - **Configuration** - information on how to configure the data generator component and the benchmark test driver - description of various configuration properties; configuration of output data serializers
  - **Running the Benchmark** - provides information on how to run the benchmark : prerequisites before starting the workload execution i.e. details on implementing a vendor specific database connector component as well as information on configuring the driver and workload specific properties; gives detailed descriptions of general driver and workload properties as well as how to set up and run the benchmark
  - **Gathering Results -** provides description of the results **-** file format as well as a sample data
  - **Validation -** provides information about validation of query results with sample data

- **Appendices**
  - **Query Set Implementations** - provides query listings for different implementations of the benchmark query sets i.e. Virtuoso SPARQL and SQL, Neo Cypher
  - **Scale Factor Statistics -** provides various statistics about the distribution and relations of entities in generated data for different scale factors

The reference documentation of the SNB software can be found in the *ldbc_snb_docs* repository which contains all recent versions of the document e.g. : LDBC_SNB_v0.1.3.pdf [10].

4.2 Datasets

Datasets used by the LDBC-SNB benchmark are a set of resource files and dictionaries extracted from the public data source DBPedia. Datasets are used by the data generator to produce synthetic large output data in such way that information mimics real-world data found in social networks not only by data structure (by following a defined data schema for entities and their relations), but also by content, e.g. person names for certain country will match the real ones.

Following Table 4 describes resource files containing dictionary properties and rankings for each entry :

Resource	Description
Browsers	A list of web browsers and their probability to be used. It is used to set the browsers used by the users.
Cities by country	A list of cities and the country they belong. It is used to assign cities to users and universities.
Companies by country	A list of companies and their country. It is used to set the countries where companies operate.
Countries	A List of countries and their population.
Emails	A List of email providers. It is used to generate the email accounts of persons.
IP Zones	A list of IP ranges assigned to each country. It is used to assign the IP addresses to users.
Languages by country	A set of languages spoken in each country. It is used to set the languages spoken by each user.
Name by country	A set of names and the probability to appear in each country. It is used to assign names to persons, correlated with their countries.
Popular places by country	A set of popular places in each country. These are used to set where images attached to posts are taken from.
Surnames’ country	A set of surnames and the probability to appear in each country. It is used to assign surnames to persons, correlated with their countries.
Tags by Country	A set of tags and their probability to appear in each country. It is used to assign the interests to persons and forums.
Tag Classes	Contains for each tag, the classes it belongs to.
Tag Hierarchies	Contains for each tagClass their parent tagClass.
Tag Matrix	Contains for each tag, the correlation probability with the other tags. It is used to

	enrich the tags associated to messages.
Tag Text	Contains a text for each tag. It is used to generate the text for messages.
Universities by city	A set of universities per city. It is used to set the cities where universities operate.

Table 5 : Description of datasets used in Social Network Benchmark

All resource and dictionaries are located at *ldbc_snb_datagen* repository as they initially will serve as a source of data for the Data generator.

More information on Datasets can be found in the SNB reference documentation (LDBC_SNB_v0.1.3.pdf)

4.3 Data Generator

The data generator of the LDBC-SNB is located in a dedicated repository : *ldbc_snb_datagen*, and is independent of the SNB test driver. It uses the set of dictionary and resource files described in the previous section to generate synthetic output data following a property dictionary model defined by : a dictionary, a ranking function and a probability function. The idea to have a separate ranking and probability function is motivated by the need of generating correlated values : for example in the case of a dictionary of property *firstName*, the popularity of first names might depend on the gender, country, birth date.

Details on the Data generator have been described in the SNB reference documentation (LDBC_SNB_v0.1.3.pdf), following list outlines some features :

- **Realistic, Scalable, Deterministic, Usable** - realistic - produced synthetic output data mimics the real-data found on the social networks not only by structure but also by content; scalable - data can be produced at various scales; deterministic - each run of the Data generator produces the same consistent data; usable - easy to configure and run
- **Data generation process** - defines the steps that are followed in the process of generating the synthetic output data, modeling of correlations in generated data, possibility to implement custom vendor's connector module
- **Input data** - resource and dictionary files as well as the schema used by the Data generator
- **Output data** - describes available data serialization formats; properties of the data - static part - data to be bulk loaded, update stream part - consisting of update events to the system
- **Parallel data generation** - data generation is executed simultaneously for a faster output of synthetic data
- **Scale factors** - a set of scale factors have been defined, targeting systems of different sizes and budgets

A quick reference on how to configure and start the data generation process can also be found on the main web page of the *ldbc_snb_datagen* repository in GitHub.

The SNB Data generator requires an open-source software component Apache Hadoop [11] - an Apache Software Foundation project for scalable and distributed computing.

5 Execution, Reporting and Auditing Rules

The purpose of benchmark auditing is to improve the credibility and reproducibility of benchmark claims by involving a set of detailed execution and reporting rules and providing third party verification of compliance with these. Rules may exist separately of auditing but auditing is not meaningful unless the rules are adequately precise. The credibility of the entire benchmark process relies on clear understanding of what a benchmark is expected to demonstrate and on the auditor being capable of understanding the process. Also relies on auditor's capability of verifying that the benchmark execution is fair and does not abuse the rules of that benchmark.

The LDBC Council has defined a common set of audible properties of systems and benchmark implementations that are being considered when defining execution, reporting and auditing rules.

Following list provides descriptions those properties :

- **ACID Compliance -** outlines transactional behaviors of SUTs which may be verified in the course of auditing a benchmark run. A benchmark specifies transactional semantics that may be required for different parts of the workload.
- **Schema Design -** a system may declare no schema at all, as may be the case with RDF or graph database systems. A benchmark may specify restrictions on schema, in the LDBC context, the matter is more complex since the range of possible SUTs is broader, including diverse combinations of schema-first and schema-less systems and configuration.
- **Qualification and Correctness of Results -** a benchmark should be published with a deterministically re-creatable validation dataset. Validation queries applied to the validation dataset will deterministically produce a set of correct answers - this is used as a pre-benchmark run test for the correctness of a SUT.
- **Data Access Transparency -** it may be specified that a system under test is not allowed the use of explicit access paths. For example explicitly specifying which explicit data structure should be used for any given operation may be prohibited.
- **Query Declarativity -** in the graph database world there is no standard query language and the APIs are in common use. For the API based implementations where join type and join order or other properties are procedurally determined it is not allowed to decide among alternative, semantically equivalent implementations of the same query based on parameters. Generally a declarative system is allowed the choice between parameterized queries and ones where parameters are given as literals.
- **Materialization -** the mix of read and update operations in a workload will determine to which degree precomputation of results is possible or warranted. Precomputation of aggregates or joins is generally forbidden unless explicitly allowed by the benchmark definition.
- **Steady State -** an online workload must be able to indefinitely keep up the reported throughput. The benchmark definition may put specific restrictions on the duration of individual parts of the workload. The SUT should be sized so as not to run out of space for new data for a reasonable duration of time. Each benchmark shall state specific requirements in this respect.
- **Operation Mix -** a benchmark consists of multiple different operations that my vary in frequency and duration. Individual instances of each operation may depend on parameter selection. A benchmark must specify an operation mix and a minimum count of operations that constitutes a compliant benchmark execution.
- **System Configuration and pricing -** a benchmark execution will produce a full disclosure report which specifies the hardware and software of the SUT, the benchmark implementation version and any specifics that are detailed in the benchmark specification.

- **Benchmark Specifics** - each benchmark states specific requirements or elaborates on the audible properties that do not apply to all, due to technological or implementation differences. Definitions of those for the LDBC-SPB and LDBC-SNB benchmarks have been added in deliverable document : D6.6.3 - Auditor Training.

Considering the properties described in the list above, each of the benchmarks (LDBC Semantic Publishing Benchmark and LDBC Social Network Benchmark) have defined a corresponding set of Execution, Reporting and Auditing rules that match their specific requirements. More details on each set of rules can be found in deliverable document : D6.6.3 - Auditor Training.

6 Conclusions

This deliverable reports on the initial benchmark integration and release of the benchmark software developed by the LDBC Council. It describes packaging of benchmark software components - public source code repositories and access, documentation, datasets, data generators. The document also provides information on the definitions of Execution, Reporting and Auditing Rules provided by the LDBC, being an integral part of the benchmark process.

Using a public source code repository for hosting the benchmark software and providing a detailed and accessible information on all aspects of the benchmarks, simplifies community interaction, build-up and integration of the benchmarks developed by the LDBC.

References

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

[2] : A software tool for operating with the Git version control system : http://www.git-scm.com

[3] : A software build tool : http://ant.apache.org/

[4] : A software building tool : http://maven.apache.org/

[5] : Reference documentation for the Semantic Publishing Benchmark :
https://github.com/ldbc/ldbc_spb_bm/blob/master/doc/LDBC_SPB_v0.1.pdf

[6] : Geonames data of geo-locations : http://www.geonames.org/

[7] : Person data from DBPedia : http://wiki.dbpedia.org/Downloads2014?v=10ne#persondata

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

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

[10] : Reference documentation for the Social Network Benchmark :
https://github.com/ldbc/ldbc_snb_docs/blob/master/LDBC_SNB_v0.1.3.pdf

[11] : A distributed computing platform : http://hadoop.apache.org/