Benchmarking Linked Open Data Management Systems

by Renzo Angles, Minh-Duc Pham and Peter Boncz

With inherent support for storing and analysing highly interconnected data, graph and RDF databases appear as natural solutions for developing Linked Open Data applications. However, current benchmarks for these database technologies do not fully attain the desirable characteristics in industrial-strength benchmarks [1] (e.g. relevance, verifiability, etc.) and typically do not model scenarios characterized by complex queries over skewed and highly correlated data [2]. The Linked Data Benchmark Council (LDBC) is an EU FP7 ICT project that brings together a community of academic researchers and industry, whose main objective is the development of industrial-strength benchmarks for graph and RDF databases.

Objective, well-designed and good quality benchmarks are important to fairly compare the performance of software products and uncover useful insights related to their strengths as well as their limitations. They encourage the advancement of technology by providing both academy and industry with clear targets for performance and functionality.

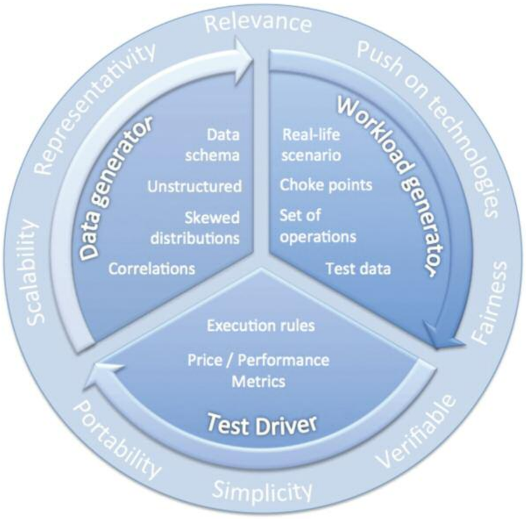

Figure 1: Elements and features of the LDBC Social Network Benchmark.

The Linked Data Benchmark Council (LDBC) aims to create benchmarks following principles including relevance, simplicity, fairness and sustainability. In particular, a goal of LDBC is to develop benchmarks that test critical usability features more thoroughly than the benchmarks so far produced by academia, including provision of a mechanism that ensures benchmarking results are reviewed for conformance by independent auditing. To this end, LDBC will provide open source benchmarks, developed by task forces integrated by expert architects who know the critical functionalities inside data management engines ("choke points"), and supported by a Technical User Community (TUC) that provides use-cases and feedback.

The Social Network Benchmark (SNB) is an LDBC benchmark intended to test various functionalities of systems used for graph data management. The scenario of this benchmark, a social network, is chosen with the following goals in mind: it should be understandable to a large audience, and this audience should also understand the relevance of managing such data; the scenario in the benchmark should cover a complete range of interesting challenges, according to the benchmark scope; and the query challenges in it should be realistic in the sense that, though synthetic, similar data and workloads are encountered in practice.

The SNB is composed of three main elements: a data generator, which allows creation of data according to a given data schema; a workload generator, which defines the set of operations that the System Under Test (SUT) has to perform; and a test driver, which is used to execute the workload over the SUT, and measure its performance according to well-defined performance metrics and execution rules. The features of these elements are summarized in Figure 1.

The SNB data generator is being designed to create synthetic data with the following characteristics: the schema must be representative of a real social network; the generation method must consider properties of real-life data, including data correlations and distributions; and the software generator must be easy-to-use, configurable and scalable. By leveraging parallelism through Hadoop, the current version of the data generator (based on the S3G2 generator [3]) ensures fast and scalable generation of huge datasets, allowing a social network structure with millions of user profiles, enriched with interests/tags, posts, and comments (see an example in Figure 2). Additionally, the generated data exhibits interesting realistic value correlations (e.g. German people having predominantly German names), structural correlations (e.g. friends being mostly people living close to one another), and statistical distributions (e.g. the friendship relationship between people follows a power-law distribution).

Aiming at covering all the main aspects of social network data management, the SNB provides three different workloads: an interactive workload, oriented to test the throughput of the systems with relatively simple queries and concurrent updates; a business intelligence workload, consisting of complex structured queries for analysing online behaviour of users for marketing purposes; and a graph analytics workload, thought to test the functionality and scalability of the systems for graph analytics that typically cannot be expressed in a query language. Each workload will be designed based on well-identified key technical challenges called "choke points". The objective is to ensure that the workload stresses important technical functionalities of actual systems.

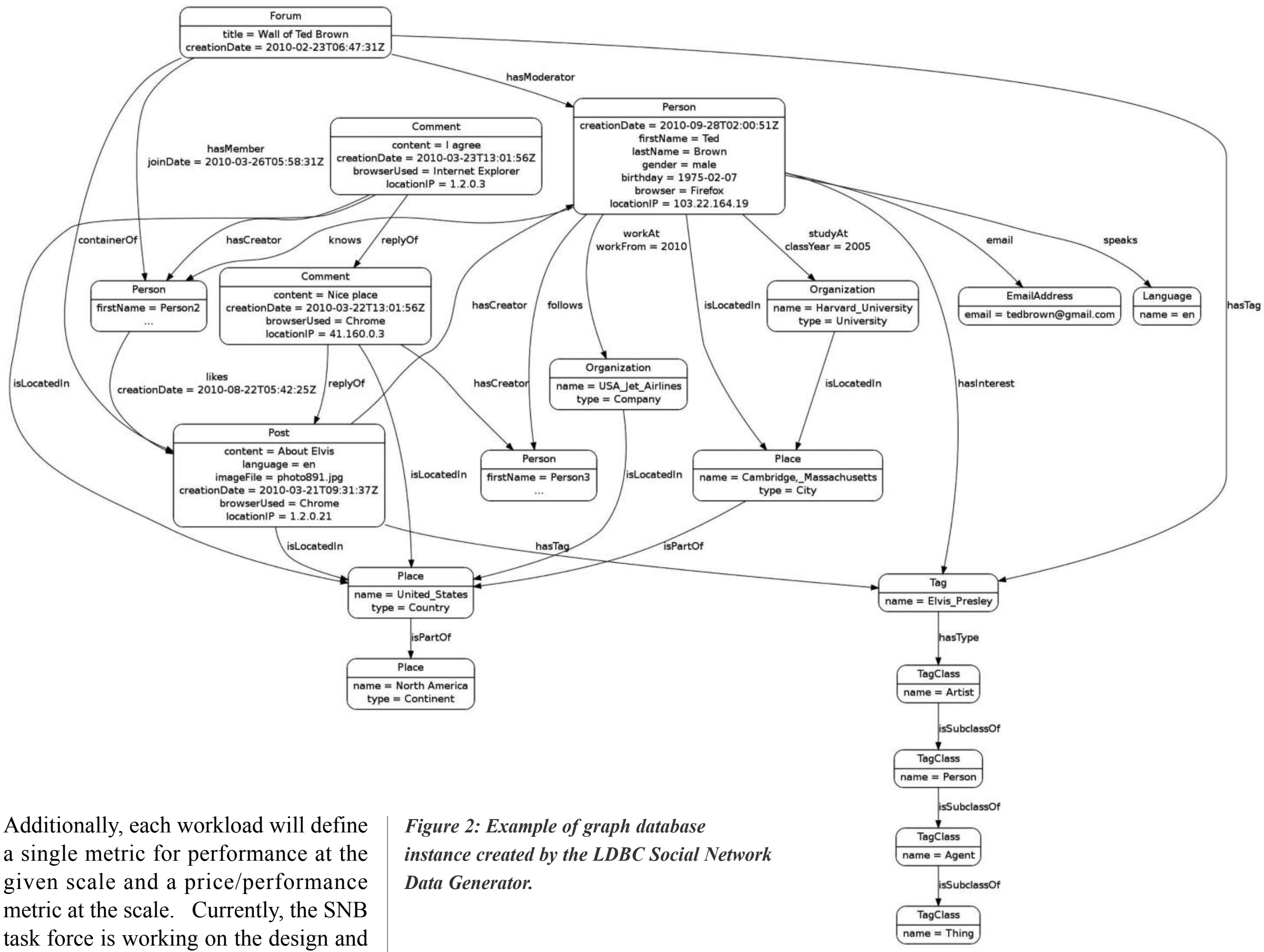

Figure 2: Example of graph database instance created by the LDBC Social Network Data Generator.

Additionally, each workload will define a single metric for performance at the given scale and a price/performance metric at the scale. Currently, the SNB task force is working on the design and implementation of the interactive workload generator. The first version of the interactive workload specification consists of twelve queries, inspired by a collection of choke points, which include "classical" complex operations (e.g. aggregated queries) and "non-traditional" complex operations (e.g. graph traversals). Besides the operations, it is also relevant to define smart methods for test data selection (ie data specifically selected to be used as substitution parameters for the operations). Test data must be carefully selected to obtain comparable results, and hence ensure the repeatability and fairness of the benchmark.

The SNB can be downloaded from GitHub and more information about its development is available in the TUC Wiki (see Links). We would like to invite readers to join the LDBC community initiative by sharing their user experience, testing their systems and participating in the LDBC-related events.

Links:
http://www.ldbc.eu
ldbc.eu:8090/display/TUC
github.com/ldbc

References:

[1] K. Huppler: "The Art of Building a Good Benchmark", in: TPCTC, 2009.
[2] S. Duan et al.: "Apples and Oranges: A Comparison of RDF Benchmarks and Real RDF Datasets", in: ACM SIGMOD, 2011.
[3] M.-D. Pham et al.: "A Scalable Structure-Correlated Social Graph Generator", in TPCTC, 2012.

Please contact:

Renzo Angles
VU University Amsterdam (Netherlands) / Universidad de Talca (Chile)
E-mail: r.anglesrojas@vu.nl