LDBC: benchmarks for graph and RDF data management

Peter Boncz
CWI, Amsterdam, The Netherlands;
boncz@cwi.nl,

ABSTRACT

The Linked Data Benchmark Council (LDBC) is an EU project that aims to develop industry-strength benchmarks for graph and RDF data management systems. LDBC introduces a so-called "choke-point" based benchmark development, through which experts identify key technical challenges, and introduce them in the benchmark workload, which we describe in some detail. We also present the status of two LDBC benchmarks currently in development, one targeting graph data management systems using a social network data case, and the other targeting RDF systems using a data publishing case.

1. INTRODUCTION

The Linked Data Benchmark Council (LDBC)[1] is an EU project that brings together a community of academic researchers and industry, whose main objective is the development of open source, yet industrial grade, benchmarks for graph and RDF databases. The founding industry members of LDBC are the graph database companies Neo Technologies and Sparsity Technologies, and the RDF database companies Ontotext and OpenLink Systems. A result of the project will be the LDBC non-profit organization, open for worldwide industry participation, which during an after the end of the EU project will supervise the creation and maintenance of the benchmarks as well as the activities for obtaining, auditing and publishing the benchmarking results.

In this paper we describe LDBC and a process for developing benchmarks based on technical challenges called "choke points", developed by LDBC. This methodology depends on a combination of workload input by end users, and access to true technical experts in the architecture of the systems being benchmarked. The overall goal of the choke-point based approach is to ensure that a benchmark workload covers a spectrum of technical challenges, forcing systems onto a path of technological innovation.

[1]Linked Data Benchmark Council is EU project FP7-317548 – see `http://ldbc.eu`).

2. CHOKE-POINTS

On the surface, a benchmark models a particular scenario, and this should be believable, in the sense that users of the benchmark must be able to understand the scenario and believe that this use-case matches a larger class of use cases appearing in practice. On a deeper – technical – level, however, a benchmark exposes technology to a workload. Here, a benchmark is valuable if its workload stresses important technical functionality of actual systems. This stress on elements of particular technical functionality we call "choke points". To understand benchmarks on this technical level, intimate knowledge of actual system architectures is needed. The LDBC consortium was set-up to gain access to those architects of the initial LDBC industry members, as well as to the architects of database systems RDF-3X, HyPer, MonetDB and Vectorwise. In a recent paper [1], LDBC authors analyzed the relational TPC-H benchmark in terms of 28 different choke points; providing both a good illustration of the choke point concept, and an interesting to-do list for those optimizing a system for TPC-H. Specific examples among those 28 are choke points like exploiting functional dependencies in group-by, foreign-key joins with a low match ratio (to be exploited by e.g. bloom filters), and discovering correlation among key attributes in a clustered index (e.g. using zone maps).

Choke points can be an important design element during benchmark definition. The technical experts in a task force identify choke points relevant for a scenario, and document these explicitly. As the benchmark workload evolves during the process of its definition, a close watch is kept on which queries in the workload test which choke point, aiming for complete coverage using a limited amount of queries. Choke points thus can ensure that existent techniques are present in a system, but can reward future systems that improve performance on still open technical challenges.

3. ONGOING DEVELOPMENT

We shortly summarize the current activities of LDBC benchmark development *task forces*. [2]

The Semantic Publishing Benchmark (SPB) simulates the management and consumption of RDF metadata that describes media assets, or creative works. The scenario is a media organization that maintains RDF descriptions of its catalogue of creative works – for this benchmark very useful input is being provided by actual media organizations which make heavy use of RDF, among which the BBC. The benchmark is designed to reflect a scenario where a large number of aggregation agents provide the heavy query workload, while at the same time a steady stream of creative work description management operations are in progress. This benchmark plainly targets RDF database systems, which support at least basic forms of semantic inference.

A driver workload is generated by a number of concurrently running editorial lookup and update and aggregation queries simulating the workload of a publishing organization. Choke points may arise in cases where the RDF database engine is not able to: decide which type of join to use; run in parallel UNIONs and DISTINCTs; choose the right query plan based on the selectivity of the DISTINCT; or identify common parts of correlated subqueries.

The Social Network Benchmark (SNB) is designed for evaluating a broad range of technologies for tackling graph data management workloads. The systems targeted are quite broad: from graph, RDF, and relational database systems to Pregel-like graph programming frameworks.

SNB includes a data generator that enables the creation of synthetic social network data with the following characteristics: the data schema is representative of a real social network; the data generated includes properties occurring in real data, e.g. irregular structure, structure/value correlations and power-law distributions; and the software generator is easy-to-use, configurable and scalable.

The requirement to generate at scale a complex social graph with special data distributions that at the same time exhibits certain interesting value correlations (e.g. German people having predominantly German names) and structural correlations (e.g. friends being mostly people living near, colleagues or classmates), poses an interesting challenge. The SNB data generator builds on the work on correlated social network generation in S3G2 [2], whose source code has been adapted to the SNB data schema. S3G2 comes with the ability to leverage parallelism through Hadoop, ensuring fast and scalable generation of huge datasets.

SNB splits into three separate workloads:

– **Interactive workload.** This workload tests system throughput with relatively simple queries and concurrent updates. The workloads test ACID features and scalability in an online operational setting. Given the high write intensity, this workload may also be used to let the dataset grow, which will be implemented by pre-generating data in the generator but only importing the data corresponding to one time point in the bulk load, and playing out the rest of the modifications in the update workload. The targeted systems are expected to be those that offer transactional functionality.

– **Business intelligence workload.** This workload consists of complex structured queries for analyzing online behavior of users for marketing purposes. The workload stresses query execution and optimization. The targeted systems are expected to be those that offer an abstract query language. Queries typically touch a large fraction of the data and do not require repeatable read.

– **Graph Analytics Workload.** This workload tests the functionality and scalability of the systems for graph analytics that typically cannot be expressed in a query language. The analytics is done on most of the data in the graph as a single operation and produces large intermediate results. The analysis is not expected to be transactional or need isolation. This workload targets graph programming frameworks, though systems with a query-language might compete using iterative implementations that repeatedly fire queries and keep intermediate results in temporary data structures.

The SNB choke-points include handling transactional conflicts, graph traversals and shortest paths, result size estimation, and many more.

4. CONCLUSIONS

We presented the Linked Data Benchmarking Council (LDBC), a new initiative towards for benchmarking RDF and Graph data management systems. A main technical advance is its "choke point" driven benchmark design, which ensures that interesting and well-chosen technical challenges will emerge from implementing the benchmarks.

5. REFERENCES

[1] P. Boncz, T. Neumann, and O. Erling. TPC-H Analyzed: Hidden Messages and Lessons Learned from an Influential Benchmark. In *TPCTC*, 2013.

[2] M.-D. Pham, P. A. Boncz, and O. Erling. S3G2: A Scalable Structure-Correlated Social Graph Generator. In *TPCTC*. Springer-Verlag, 2012.

[2]For details, see the LDBC Technical User Community Portal: `http://www.ldbc.eu:8090/display/TUC`.