8th Technical User Community meeting - LDBC

Peter Boncz
boncz@cwi.nl
LDBC Organization (non-profit)

+ non-profit members (FORTH, STI2) & personal members
+ Task Forces, volunteers developing benchmarks
+ TUC: Technical User Community (8 workshops, ~40 graph and RDF user case studies, 18 vendor presentations)
What does a benchmark consist of?

• Four main elements:
  – *data & schema*: defines the structure of the data
  – *workloads*: defines the set of operations to perform
  – *performance metrics*: used to measure (quantitatively) the performance of the systems
  – *execution & auditing rules*: defined to assure that the results from different executions of the benchmark are valid and comparable

• Software as Open Source (GitHub)
  – data generator, query drivers, validation tools, ...
Audience

• For developers facing graph processing tasks
  – recognizable scenario to compare merits of different products and technologies

• For vendors of graph database technology
  – checklist of features and performance characteristics

• For researchers, both industrial and academic
  – challenges in multiple choke-point areas such as graph query optimization and (distributed) graph analysis
SPB scope

• The scenario involves a media/publisher organization that maintains semantic metadata about its Journalistic assets (articles, photos, videos, papers, books, etc), also called Creative Works

• The Semantic Publishing Benchmark simulates:
  – Consumption of RDF metadata (Creative Works)
  – Updates of RDF metadata, related to Annotations

• Aims to be an industrially mature RDF database benchmark (SPARQL1.1, some reasoning, text and GIS queries, backup&restore)
LDBC Task Forces

- Semantic Publishing Benchmark Task Force
  - Develops industry-grade RDF benchmark

- Social Network Benchmark Task Force
  - Develops benchmark for graph data management systems
  - Broad coverage: three workloads

- Graph Analytics Task Force
  - Spin-off from the SNB task force

- Graph Query Language Task Force
  - Not strictly about benchmarking
  - Studies features of graph database query languages
Semantic Publishing Benchmark (SPB)

Information on this page will not be updated. Facts were accurate as of August 13, 2012.

Bulgaria

Team GB's Campbell secures medal

Luke Campbell is guaranteed an Olympic medal after beating Bulgaria’s Detelin Dalakliev in his bantamweight semi-final.

6 Aug 12

Bulgaria beat GB volleyball men

MEN'S VOLLEYBALL
29 Jul 12

Great Britain's men produce a battling display on their Olympic debut but are beaten in straight sets by Bulgaria at Earls Court.

Bulgaria Medallists

- Bronze
  Teofil Pulev
  Men's Heavyweight (91kg)

- Silver
  Stanka Zlateva Hristova
  Women's Freestyle 72kg
Social Network Benchmark: schema
Benchmark Workloads

• **Interactive**: tests throughput running short queries while consistently handling concurrent updates
  – *Show all photos posted by my friends that I was tagged in*

• **Business Intelligence**: consists of complex structured queries for analyzing online behavior
  – *Influential people the topic of open source development?*

• **Graph Analytics**: tests the functionality and scalability on most of the data as a single operation
  – *PageRank, Shortest Path(s), Community Detection*
Systems

- **Graph database systems**
  - e.g. Neo4j, InfiniteGraph, DEX, Titan
- **Graph programming frameworks**
  - e.g. Giraph, Signal/Collect, Graphlab, Green Marl, Grappa
- **RDF database systems**
  - e.g. OWLIM, Virtuoso, BigData, Jena TDB, Stardog, Allegrograph
- **Relational database systems**
  - e.g. Postgres, MySQL, Oracle, DB2, SQLServer, Virtuoso, MonetDB, Vectorwise, Vertica
- **noSQL database systems**
  - e.g. HBase, REDIS, MongoDB, CouchDB, or even MapReduce systems like Hadoop and Pig
## Workloads by system

<table>
<thead>
<tr>
<th>System</th>
<th>Interactive</th>
<th>Business Intelligence</th>
<th>Graph Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph databases</td>
<td>Yes</td>
<td>Yes</td>
<td>Maybe</td>
</tr>
<tr>
<td>Graph programming frameworks</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RDF databases</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Relational databases</td>
<td>Yes</td>
<td>Yes</td>
<td>Maybe, by keeping state in temporary tables, and using the functional features of PL-SQL</td>
</tr>
<tr>
<td>NoSQL Key-value</td>
<td>Maybe</td>
<td>Maybe</td>
<td>-</td>
</tr>
<tr>
<td>NoSQL MapReduce</td>
<td>-</td>
<td>Maybe</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Interactive (On-line) Workload

• Test online ACID features and scalability
• The system under test is expected to run in a steady state, providing durable storage
• Updates are typically small
• Updates will conflict a small percentage of the time
• Queries typically touch a small fraction of the database
The LDBC Social Network Benchmark: Interactive Workload

Orri Erling
OpenLink Software, UK
oerling@openlinksw.com

Alex Averbuch
Neo Technology, Sweden
alex.averbuch@neotechnology.com

Josep Larriba-Pey
Sparsity Technologies, Spain
larri@sparsity-technologies.com

Hassan Chafi
Oracle Labs, USA
hassan.chafi@oracle.com

Andrey Gubichev
TU Munich, Germany
gubichev@in.tum.de

Arnaud Prat*
Universitat Politècnica de Catalunya, Spain
apratt@ac.upc.edu

Minh-Duc Pham
VU University Amsterdam,
The Netherlands
m.d.pham@vu.nl

Peter Boncz
CWI, Amsterdam, The Netherlands
boncz@wci.nl

ABSTRACT
The Linked Data Benchmark Council (LDBC) is now two years underway and has gathered strong industrial participation for its mission to establish benchmarks, and benchmarking practices for evaluating graph data management systems. The LDBC introduced a new choke-point driven methodology for developing benchmark workloads, which combines user input with input from expert systems architects, which we outline. This paper describes the LDBC Social Network Benchmark (SNB), and presents database benchmarking innovation in terms of graph query function...

tional table, for instance as a table where every row contains an edge, and the start and end vertex of every edge are a foreign key reference (in SQL terms). However, what makes a data management problem a graph problem is that the data analysis is not only about the values of the data items in such a table, but about the connection patterns between the various pieces. SQL-based systems were not originally designed for this – though systems have implemented diverse extensions for navigational and recursive query execution.

In recent years, the database industry has seen a proliferation of new graph-oriented data management technologies.
Business Intelligence Workload

- The workload stresses query execution and optimization
- Queries typically touch a large fraction of the data
- The queries are concurrent with trickle load
- The queries touch more data as the database grows
Graph Analytics Workload (Graphalytics)

- The analytics is done on most of the data in the graph as a single operation
- The analysis itself produces large intermediate results
- The analysis transactional: no need for isolation from possible concurrent updates
Graphalytics Architecture
Graphalytics Algorithms

- **general statistics (STATS)**
  - counts the numbers of vertices and edges in the graph and computes the mean local clustering coefficients
- **breadth-first search (BFS)**
  - traverses the graph starting from a seed vertex, visiting first all the neighbors of a vertex before moving to the neighbors of the neighbors.
- **connected components (CONN) algorithm**
  - determines for each vertex the connected component it belongs to.
- **community detection (CD) algorithm**
  - detects groups of nodes that are connected to each other stronger than they are connected to the rest of the graph
- **graph evolution (EVO)**
  - predicts the evolution of the graph according to the “forest fire” model
LDBC Graphalytics: A Benchmark for Large-Scale Graph Analysis on Parallel and Distributed Platforms


Oracle Labs, Intel Labs, IBM Research, Huawei Research America, Delft University of Technology, UPC Barcelona, Georgia Tech, CWI Amsterdam

ABSTRACT

In this paper we introduce LDBC Graphalytics, a new industrial-grade benchmark for graph analysis platforms. It consists of six deterministic algorithms, standard datasets, synthetic dataset generators, and reference output, that enable the objective comparison of graph analysis platforms. Its test harness produces deep metrics that quantify multiple kinds of system scalability, such as horizontal/vertical and weak/strong, and of robustness, such as failures and performance variability. The benchmark comes with open-source software for generating data and monitoring performance. We describe and analyze six implementations of the benchmark (three from the community, three from the industry), providing insights into the strengths and weaknesses of the platforms. Key to our contribution, vendors perform the tuning and benchmarking of their platforms.
More Information


Blogs
Specifications
Early Result FDRs
Videos of TUC talks
Developer info
Code, Issue Tracking
Graph Query Language Task Force

- Renzo Angles, Universidad de Talca
- **Marcelo Arenas, PUC Chile - task force lead**
- Pablo Barceló, Universidad de Chile
- Peter Boncz, Vrije Universiteit Amsterdam
- George Fletcher, Eindhoven University of Technology
- Irini Fundulaki, Foundation for Research and Technology - Hellas (FORTH)
- Claudio Gutierrez, Universidad de Chile
- Tobias Lindaaker, Neo Technology
- Marcus Paradies, SAP
- Raquel Pau, UPC
- Arnau Prat, UPC
- Tomer Sagi, HP Labs
- Oskar van Rest, Oracle Labs
- Hannes Voigt, TU Dresden
- Yinglong Xia, Huawei Research America
GRADES Invitation (Friday)

- Keynote by Jure Leskovec (Stanford + Pinterest)
  - Known for creating the SNAP dataset collection
- First paper on Graph Frames
  - Spark RDD extension for graph data
  - PGQL: Oracle Graph Query Language
  - And more..