The LDBC Social Network Benchmark

Gábor Szárnyas
CWI

DSDSD, 13 May 2022
Graph data management systems

- Graph data model
- Graph query language
MATCH
  sp = shortestPath((p1:Person {name: 'Bob'}) - [:knows*] - (p2:Person)),
  (p2)-[:interestedIn]->(:Tag)-[:subclassOf*]->(:Tag {name: 'Art'})
RETURN sp

Query formulated in Cypher.
MATCH
sp=shortestPath((p1:Person {name: 'Bob'})-[[:knows*]-(p2:Person)),
(p2)-[:interestedIn]->(:Tag)-[:subclassOf*]->(:Tag {name: 'Art'}))
RETURN  sp

Query formulated in Cypher.

property graph
Graph data management systems

GDMSs provide a graph-aware UI and support graph processing features.

- Graph data model
- Graph query language
- Graph visualization
Graph data management systems

GDMSs provide a graph-aware UI and support graph processing features.
GDMSs have performance problems

Performance- and maturity issues are common.

→ EDBT 2022 keynote

The graph data management space needs to evolve faster.
TPC
- Stimulate competition
- Clear goals
- Clear rules
- Governance

TPC price/perf trend 1990-2005: improved 58% per year
prices have declined 37%/y

TPC-A
price/perf
37% per year

TPC-C
price/perf
37% per year
LDBC = Linked Data Benchmark Council

**GOAL**
Accelerate progress in graph data management

**METHOD**
Create *standard benchmarks* and supervise their use

**APPROACH**
*Neutral* – members include DBMS, hardware & cloud vendors

GDMS benchmark: Social Network Benchmark suite
Social network data set
Example graph

Main entities:
- Person-knows-Person network
- Forums
- Message threads

Correlations:
- Structure-level
- Attribute-level

Dynamic graph
Lifespan management

The generator sees the entire temporal graph at once.

Plus: flashmobs events
**Dynamic graph**

Initial snapshot and insert/delete batches

---

*Supporting dynamic graphs in SNB Datagen*, GRADES-NDA 2020
Deleting a Forum
SNB Business Intelligence

🔗 Specification

- Analytical workload
- Subgraph queries
- Path finding queries
- Periodic refreshes
Business Intelligence workload

Workload: Ad-hoc graph OLAP queries with daily updates

Batches: 33 days of W/R operations

- W: apply one day’s worth of updates
- R: 20 complex read queries with different parameters

≈35 months

2012-Nov-29

2012-Nov-30

2012-Dec-31

load

W

R

W+R

power test

throughput test

throughput test
Parameter curation

Parameter selection is especially important for skewed and correlated data sets:

- starting a query from a person with a low degree vs. a high degree
- cost of reachability queries if there is a path vs. no path
Umbra SF10: naïve vs. curated parameters
Q11: Triangle query – WCOJs are beneficial

Parameters: Only big countries, similar intervals
Q14: Correlations – Different runtimes/query plans

For each pair of countries, calculate the cost as a sum of cases #1-4. Cases that have a match add to the final score with the specified value. Each case only counts once, multiple matches do not increase to the score.

Parameters: (A) close countries, (B) far-away countries
Q18: Diamond query – WCOJs are beneficial

Parameters: pick tags with similar amount of persons
Q20: Single-source weighted shortest path

Parameters:
1. compute same-university-knows edges during the measurement window
2. pick similarly-sized companies
3. determine company-person2 pairs where there is a path
## BI implementations

Complete reference implementations

<table>
<thead>
<tr>
<th>system</th>
<th>data model</th>
<th>language</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neo4j</td>
<td>graph</td>
<td>Cypher</td>
<td>500</td>
</tr>
<tr>
<td>TigerGraph</td>
<td>graph</td>
<td>GSQL</td>
<td>1300</td>
</tr>
<tr>
<td>Umbra</td>
<td>relational</td>
<td>SQL</td>
<td>850</td>
</tr>
</tbody>
</table>
Design decisions to prevent misuse
[PREVENT] Immature systems

Signs of immaturity: append-only storage, limited data model, basic optimizer

- large-scale data sets
- date time, arrays, Unicode strings
- complex queries, ~30 choke points
- deletes → garbage collection
[PREVENT] Benchmark mistakes and cheating

Auditing process

- auditors inspect the code
- full disclosure report for reproducibility
- audited results on ldbcouncil.org

Trademarking and fair use terms

- LDBC benchmark results is trademarked
- non-audited benchmarks require a disclaimer
- we still encourage to use LDBC on immature systems
Summary
LDBC timeline

- **EU FP7 project**
- **TUC meeting**
- **Papers**
- **Audit**

- **data generator**
  - TPCTC

- **SNB Interactive**
  - SIGMOD

- **SNB BI beta**
  - GRADES-NDA

- Two-day TUC

- ACM SIGMOD PODS 2022
  - Philadelphia, PA, USA
Conclusion

LDBC SNB is a comprehensive benchmark suite for graph processing.

Ongoing developments:

- **Finalizing SNB BI v1.0**, supporting the adoption of the SNB.
- **Updated benchmark: SNB Interactive v2.0**
  - transactional workload
  - deletes and larger scale factors are backported
- **New benchmark: Financial Benchmark**
  - multi-graph of financial transactions
  - low-latency queries (< 20ms)