

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



MATCH

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



Graph data management systems

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



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*.





LDBC = Linked Data Benchmark Council

GOAL Accelerate progress in graph data management

Create *standard benchmarks* and supervise their use

APPROACH

METHOD

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

- 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



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



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
 (2) determine the second s

(3) determine company-person2 pairs where there is a path

BI implementations

Complete reference implementations

system	data model	language	LOC
Neo4j	graph	Cypher	500
TigerGraph	graph	GSQL	1 300
Umbra	relational	SQL	850







Design decisions to prevent misuse



[PREVENT] Immature systems

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



[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

data



EU FP7 project | TUC meeting | Papers | Audit

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)



The graph & RDF benchmark reference