Retrospective review of publications related to LDBC benchmark standards: 2020 whitepaper (authored by Oracle)

Summary

document	https://www.oracle.com/a/tech/docs/ldbc-graph-benchmark-2020-06-30-neo-only-v3.1.pdf
date	June 2020
workload	SNB Business Intelligence
scale factor(s)/data set size	sf100
code availability	no code available
data generation	SNB Datagen, v0.3.2 (?)
parameter generation	not specified
benchmark driver	custom
systems-under-test	Oracle PGX, Neo4j 3.5.0 Community Edition

1. What does it purport to be? What does the publication state as its goals and results, and does it state any qualifications or limitations to those goals and results?

The *publication compares the performance of Neo4j and Oracle PGX* by using all 25 queries of the LDBC SNB BI workload available as of June 30, 2020. In this retrospective report, we mostly discuss the details of the LDBC SNB BI workload run on Oracle PGX, as results produced by Neo4j are taken from an earlier publication from July, 2019. That execution and the related results are discussed as part of a separate retrospective report.

2. How is it being used, quoted or described in published material available on the web? This should be conducted to a reasonable degree with a focus on statements made by LDBC members.

Only one reference to this report was found during the research:

- Oracle website has a blog entry where the findings from the publication are summarized: https://blogs.oracle.com/oraclespatial/post/oracle-database-property-graph-outperforms-neo4j-in-ldbc-social-network-bi-benchmark

3. How closely does it adhere to the LDBC benchmark standards, in terms of completeness, fidelity, reproducibility etc.? Which changes were made compared to the official benchmark specification (data sets, queries, query parameters, driver/workload)?

The benchmarking rules for the BI workload were not released by the time the report was written (and these rules have not been released as of the writing of this retrospective report), thus there are only a few available existing resources one can rely on when running the BI workload. The GRADES paper *An early look at the LDBC Social Network Benchmark's Business Intelligence workload* from 2018 by LDBC members is one publication that describes a possible benchmark run:

> "*Methodology.* We executed 100 queries for warmup, then executed 250 queries and measured their response time. Queries were selected randomly, following a uniform distribution and were executed one-by-one, i.e. with no interleave between them. For each scale factor/tool/query, we calculated the geometric mean of execution times"

The present publication uses the median of the first 10 executions. Median is a robust metric, although the suggested approach would be to use the geometric mean of measured executions. Furthermore, it is not clear what parameters were used in the query executions. In an LDBC-approved audit, the queries would be provided with different query parameters each time to reduce the impact of trivial data caching effects. Additionally, parameters in audited runs shall be generated using LDBC's "parameter curation" technique to ensure predictable query runtimes.

Furthermore, below is a list of minor issues that would have been problematic during an audited benchmark run with Oracle PGX.

- Exact version numbers are missing of the used LDBC documents and components
    - This is especially problematic in the case of LDBC Datagen – it is possible that the presented benchmark with Oracle PGX used a Datagen (v0.3.2?) that has a different version compared to the one used for obtaining assessment results for Neo4j formerly (v0.2.7)
- The input data (social network graph) is available in several CSV formats, but the report does not say which one of them was used
- License costs are missing from cost estimates

4. How accurate is its reporting and analysis?

The publication has detailed benchmark results in the form of a table with run times. Additionally, runtimes of queries are presented using barcharts with log scale y-axis (Figure 5). Barcharts with log scale are to be avoided because they are confusing; instead some different presentation would be desired (e.g., simply use scatterplots).

Also, it is important to mention that the results from the paper comparing TigerGraph and Neo4j performance evaluation times for Neo4j for SF-100 are unreliable as of the writing of this retrospective report, as clearly there has been a copy-paste error in the original results table. (See in the retrospective report on the TigerGraph vs. Neo4j comparison.)

Finally, it is mentioned in the report that the evaluation was done on equivalent machines, yet they are not exactly the same, which would be the ideal case. Neo4j numbers were produced in Amazon AWS, while Oracle numbers in the Oracle Cloud.

5. How might this publication be categorized if it were produced in the future, after we have adopted policies on how to describe or publish results obtained from audited, non-audited-but-complete, and LDBC derivative-or-inspired tests?

We would rate this publication as derived, as there are some shortcomings, however, these are mostly due to the fact that the BI auditing guidelines are not yet released.

6. Any reactions, prior to distribution to the board, from the publication's authors

Comment from Oskar Van Rest (point of contact at Oracle): I shared this with colleagues at Oracle. We believe that the report correctly points out some deficiencies with the results of our blog post, like missing query parameters, missing versions numbers and missing code for reproducibility.
We believe the “bronze / derived” rating to be fair.
No further comments.