Reproducibility Report for ACM SIGMOD 2022 Paper: "Conjunctive Queries with Comparisons"

THOMAS NEUMANN, TUM, Germany

The paper introduces a new evaluation strategy for conjunctive queries with predicates, implements it on top of Spark, and demonstrates its effectiveness with different query scenarios. When repeating the experiments, we obtained results similar to those reported in the original paper.

1 INTRODUCTION

The reproduced paper [1] was written by Qichen Wang and Ke Yi from Hong Kong University of Science and Technology. It introduces an efficient evaluation strategy for acyclic queries with inequality predicates, improving the asymptotic complexity for these cases, and implements the strategy atop of Spark to demonstrate practical relevance. We were able to reproduce all experiments reported in the paper except for Figure 10, where we got slightly different results.

2 SUBMISSION

The source code and the scripts for reproducing the paper results are available on GitHub. Submission consists of

- GitHub repository with code and scripts at: https://github.com/hkustDB/SparkCQC
- a detailed readme describing the evaluation process
- source code for most experiments (the TPC-E data generator and the any-k source code is referenced by the README and is downloaded separately)
- scripts to run all experiments and to produce the results

3 HARDWARE AND SOFTWARE ENVIRONMENT

Table 1 a shows the hardware setup used for the experiments.

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	Intel Xeon	Intel X7560
cores	24	60
GHz	2.1	2.2
RAM	416GB	1024GB
Storage	4 disk HDD RAID	6 disk HDD RAID

4 REPRODUCIBILITY EVALUATION

4.1 Process

We ran the experiments using the scripts provided by the authors. The scripts mostly worked out of the box, there were only a few dependencies not mentioned in the readme that we had to add[1]. The overall process is very slow, though, and requires a lot of disk space. When a process is killed due to timeouts the temporary storage is not cleaned up, which caused problems. The default query timeout of 24h is quite excessive, too, in particular given the fact that each query is

[1]we had to add: curl, unzip, make, g++, bc

repeated 10 times. We changed that to a timeout of 1h, which is sufficient for the more interesting approaches. The process still took several days and required manual intervention along the way to avoid overflowing the disk with temporary files.

4.2 Results

Using the provided scripts we could reproduce the results of Figure 7, 8, 9, obtaining very similar results. In Figure 10 we got some slight deviations, in our experiments the 1-D Alternative was $\approx$ 9% slower than SparSQL, while the original paper indicated that it was about 10% faster. Overall both approaches were very similar in this particular experiment, though, the difference might stem from the different hardware used in the experiments.

5 SUMMARY

The evaluation of the reproducibility results shows that it produces similar results compared with the paper results.

REFERENCES

[1] Qichen Wang and Ke Yi. 2022. Conjunctive Queries with Comparisons. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 108–121. https://doi.org/10.1145/3514221.3517830