Assignment on Version Management in Data Lakes

Due: April 15th, 2025

Drawing inspiration from the topics covered in class, this assignment challenges students to design, implement, and evaluate an original solution for one of the two following tasks. You may work alone or with a partner.

1 External Version Explanation

Extend the approach of [Shraga and Miller, 2023] to consider external changes. Specifically, for attribute additions for which no valid explanation can be found, use table search to find a set of candidate tables that can explain the addition using a join. Propose a way to rank the tables and select a best candidate to use along with an appropriate transformation (join, left join, outer-join, etc.) You may use any join search method. Two that have open code include [Zhu et al., 2016, 2019].

It is not required that you consider row additions, but you certainly may, especially if you are interested in publishing. Notice that it is not sufficient to just use union search for this as you want tables containing specific values.

There is no benchmark for this task, so you will need to create your own, perhaps by extending the benchmarks from [Shraga and Miller, 2023] or [Fox et al., 2024].

2 Version Search

We discussed possible ways to train a model to perform version search. One would be to use contrastive learning directly, mimicking the approach used in Starmie Fan et al. [2023], but replacing the augmentation operator with version transformations. A second might be to fine-tune the Starmie model (https://github.com/megagonlabs/starmie) on versions. Pick an approach (even a different one) and evaluate your search over a data late that perhaps includes all the version benchmarks from [Fox et al., 2024].

Deliverables

1. Technical Report (PDF or shared overleaf)
    - Problem Analysis: Challenges in external explaination or version search.
    - Methodology: Detailed description of your approach.
    - Results: Quantitative evaluation (tables/graphs of metrics) and qualitative examples. Include an analysis of good outcomes and poor outcomes.
    - Include your github link.
2. GitHub Repository
    - Code: Fully documented, executable implementation.
    - Data: Instructions to reproduce your experiments

References

G. Fan, J. Wang, Y. Li, D. Zhang, and R. J. Miller. Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning. *Proc. VLDB Endow.*, 16(7):1726–1739, 2023. doi: 10.14778/3587136.3587146. URL https://www.vldb.org/pvldb/vol16/p1726-fan.pdf.

D. C. Fox, A. Khatiwada, and R. Shraga. A generative benchmark creation framework for detecting common data table versions. In E. Serra and F. Spezzano, editors, *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 5365–5369. ACM, 2024. doi: 10.1145/3627673.3679157. URL https://doi.org/10.1145/3627673.3679157.

R. Shraga and R. J. Miller. Explaining dataset changes for semantic data versioning with explain-da-v. *Proc. VLDB Endow.*, 16(6):1587–1600, 2023. doi: 10.14778/3583140.3583169. URL https://www.vldb.org/pvldb/vol16/p1587-shraga.pdf.

E. Zhu, F. Nargesian, K. Q. Pu, and R. J. Miller. LSH ensemble: Internet-scale domain search. *Proc. VLDB Endow.*, 9(12):1185–1196, 2016. doi: 10.14778/2994509.2994534. URL http://www.vldb.org/pvldb/vol9/p1185-zhu.pdf.

E. Zhu, D. Deng, F. Nargesian, and R. J. Miller. JOSIE: overlap set similarity search for finding joinable tables in data lakes. In P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, editors, *SIGMOD*, pages 847–864. ACM, 2019. doi: 10.1145/3299869.3300065. URL https://doi.org/10.1145/3299869.3300065.