Short Research Project Description

Title: Access Path Selection in Modern Columnar DBMSs
Responsible mentor: Teona Bagashvili and Junchang Wang

Background: Efficient data retrieval is critical for analytical DBMSs. Notable mechanisms include sequential scans (using Zone Maps), sparse indexing (e.g., Columns Sketches [1]), and the recently emerging bitmap indexing [2,3]. Each method possesses distinct strengths and is optimal for specific queries and data characteristics, necessitating the selection of the appropriate approach in DBMSs.

Existing research on retrieval approaches has mainly focused on pure columnar DBMSs [4]. This overlooks the evolution of modern analytical DBMSs, which have widely adopted PAX-based formats [5] for both storage and in-memory column data. Therefore, selecting the optimal retrieval strategy for modern PAX-based vectorized columnar DBMSs remains an open question with practical implications.

Objective: This project aims to give students a good understanding of data retrieval mechanisms, guiding them to a solution on access path selection in modern analytical DBMSs. Using a customized DuckDB [6] with state-of-the-art retrieval mechanisms implemented by TAs, students will investigate the dimensions influencing various mechanism, evaluate their trade-offs, and propose a comprehensive solution to the open question of optimal retrieval strategy. The workflow is as follows:

(a) Download the custom DuckDB (using branch *Experiments*), study the data retrieval procedures in a modern columnar DBMS with PAX-like storage, and analyze the codebase to understand how a scan-heavy query (e.g., TPC-H Q6) is executed.
(b) Review the implementations of different retrieval mechanisms (including Scan with zone maps, Column Sketches with AVX-512 acceleration, and CUBIT/RABIT).
(c) Read the papers on these approaches to identify potential parameters affecting query performance, such as query type, selectivity, and column data characteristics (e.g., cardinality, distribution, and data types).
(d) Generate workloads and verify/refine your theoretical models, ultimately providing comprehensive guidance on access path selection in modern columnar DBMSs.

References

[1] B. Hentschel, M. S. Kester, and S. Idreos. 2018. Column Sketches: A Scan Accelerator for Rapid and Robust Predicate Evaluation. In Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18).
[2] J. Wang and M. Athanassoulis. 2024. CUBIT: Concurrent Updatable Bitmap Indexing. Proc. VLDB Endow. 18(2), pp. 399–412.
[3] J. Wang, F. Xiao, and M. Athanassoulis. 2025. RABIT: Efficient Range Queries with Bitmap Indexing. Proc. ACM Manag. Data. 3(6), pp. 354.

[4] Daniel J Abadi, Daniel S Myers, David J DeWitt, and Samuel R Madden. 2006. Materialization strategies in a column-oriented DBMS. In 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 466–475.
[5] A. Ailamaki, D. J. DeWitt, M. D. Hill, and M. Skounakis. 2001. Weaving Relations for Cache Performance. In Proceedings of the 27th International Conference on Very Large Data Bases (VLDB '01), pp. 169–180.
[6] https://github.com/junchangwang/CS561-Access-Path-Selection.git