Predicting Memory Demand for a Batch of Queries

Ghadeer Abuoda

Electrical Engineering and Computer Science
Lassonde School of Engineering
York University
gha@yorku.ca

1 Introduction

Estimating the resource demand of database queries is essential for various database operations and decision-making processes, including admission control, workload management, and capacity planning. "working memory" refers to a portion of system memory where the database management system (DBMS) carries out in-memory operations, such as sorting and aggregation, during query execution. With limited system memory, the DBMS must determine the optimal timing for scheduling and executing a finite number of in-memory tasks (i.e., queries). If the DBMS inaccurately estimates a query's working memory requirements, it may either under-allocate or over-allocate memory. This can prevent the DBMS from achieving optimal query performance, such as faster execution and higher throughput, and may lead to query failures. To ensure high performance, the DBMS needs precise working memory estimations for queries before admitting them for execution.

2 Our Research

In our recent work, we propose a novel approach for estimating the memory demand of a batch of database queries, known as a workload [3]. Unlike traditional methods that estimate the resource usage of individual queries—particularly for tasks like cardinality estimation [1, 2], which is separate from working memory estimation but often leads to inaccuracies—our approach focuses on the collective resource demands of query batches. By modeling the memory requirements of an entire workload, we aim to achieve more accurate resource estimations. First, queries with similar plan characteristics and estimated cardinalities tend to have similar memory demands. Using this intuition, we analyze historical query data to identify query templates, which group queries with similar memory usage patterns. Second, we divide training queries into fixed-size workloads and represent each workload as a histogram—a distribution of query templates.

Through a comprehensive experimental evaluation, we show that our approach reduces the memory estimation error of the state-of-the-practice methods. Compared to an alternative single-query model, our model and its variants were faster during training and inferencing. Overall, the results demonstrate the advantages of the LearnedWMP approach and its potential for a broader impact on query performance optimization.

3 Presentation Structure

In this presentation, we will outline the problem of predicting the working memory needed for a batch of queries before their execution in the DBMS, the state of the practice limitations, and the importance of tackling this problem. We will present our recent work on an approach for efficiently predicting the working memory demand for a batch of queries and possible integration with the DBMS.

References

[1] Yuxing Han, Ziniu Wu, Peizhi Wu, Rong Zhu, Jingyi Yang, Liang Wei Tan, Kai Zeng, Gao Cong, Yanzhao Qin, Andreas Pfadler, Zhengping Qian, Jingren Zhou, Jiangneng Li, and Bin Cui. Cardinality estimation in dbms: a comprehensive benchmark evaluation. *Proc. VLDB Endow.*, 15(4):752–765, 2021.

[2] Kyoungmin Kim, Jisung Jung, In Seo, Wook-Shin Han, Kangwoo Choi, and Jaehyok Chong. Learned cardinality estimation: An in-depth study. In *Proc. of the 2022 International Conference on Management of Data*, pages 1214–1227, 2022.

[3] Shaikh Quader, Andres Jaramillo, Sumona Mukhopadhyay, Ghadeer Abuoda, Calisto Zuzarte, David Kalmuk, Marin Litoiu, and Manos Papagelis. LearnedWMP: Workload Memory Prediction Using Distribution of Query Templates. *arXiv preprint arXiv:2401.12103*, 2024.