ML4DM ‘24: The fourth workshop on the emerging applications of machine learning in modern data management

Jarek Szlichta
Electrical Engineering & Computer Science
York University
Toronto, ON, Canada
szlichta@yorku.ca

Calisto Zuzarte
IBM
Toronto, ON, Canada
calisto@ca.ibm.com

Verena Kantere
Electrical Engineering & Computer Science
University of Ottawa
Ottawa, ON, Canada
vkantere@uottawa.ca

Amin Kamali
Digital Transformation & Innovation
University of Ottawa
Ottawa, ON, Canada
skama043@uottawa.ca

Andrew Chai
Electrical Engineering & Computer Science
York University
Toronto, ON, Canada
abfchai@yorku.ca

Xiaohui Yu
Electrical Engineering & Computer Science
York University
Toronto, ON, Canada
xhyu@yorku.ca

Chang Liu
Electrical Engineering & Computer Science
University of Ottawa
Ottawa, ON, Canada
cliu162@uottawa.ca

Baoming Chang
Electrical Engineering & Computer Science
University of Ottawa
Ottawa, ON, Canada
bchan081@uottawa.ca

ABSTRACT

Machine Learning (ML) has gained prominence across various fields, including data management. Rule-based components are being replaced by ML-driven counterparts that extract rules from experience. Statistical methods are giving way to approaches that learn functional dependencies, correlations, and data skewness. Learning-based techniques offer advantages, such as reducing the cost of developing and maintaining complex classical modules while tailoring behavior to individual system needs. This workshop brings together leaders from research projects and audiences from academia and industry to explore examples of utilizing ML to modernize data management. The discussed instances span five categories: Robust Plan Selection, Database Knobs Tuning, Data Acquisition, Join Order Selection, and Query Plan Representation.

1 RATIONALE

Effective maintenance and optimization of Database Management Systems (DBMS) heavily rely on the expertise of highly experienced professionals. Achieving peak performance is a significant challenge due to the intricate complexities, diversities, and interdependencies inherent in these systems. Additionally, the constantly shifting landscape of DBMS, characterized by dynamic data updates, evolving workloads, and changing computing environments, further compounds this challenge.

As a result, maintaining optimal system performance becomes a constant endeavor, necessitating frequent recalibrations. To confront these issues, the database community has recently embarked on a transformative journey, harnessing the power of ML (ML4DM). This paradigm shift seeks to automate the laborious tasks associated with the development and maintenance of DBMSs.

Numerous data management tasks require evaluating options within a search space, where problem complexity hinders exhaustive searches within reasonable time frames, and conventional methods prove inadequate and susceptible to local optima entrapment. Reinforcement learning and other ML-driven search techniques have emerged as solutions for such problems. Examples include join order selection [10], knobs tuning [3,4,12], data partitioning [7], plan optimization [5], index selection [1], view selection [2], and query rewrite [13].

Other tasks require precise estimations of target values, where conventional methods struggle with inaccuracies due to simplifying assumptions. Examples include estimating cardinalities [8], estimating cost and latency [6], predicting workload characteristics and resource consumption [11], and learning indexes [9].

2 WORKSHOP OVERVIEW

This is the fourth occurrance of the ML4DM workshop [1]. In its fourth occurrence, this workshop takes a fresh look at five research projects in the field of ML4DM. The teams demonstrate their advancements in tackling difficult problems in the field. The workshop comprise two sessions in half day. The first session included talks, each discussing ML-based methods in one of the following areas: Robust Plan Evaluation, Knobs Tuning, Data Acquisition, Join Order Selection, and Explainable Plan

[1] Links to long abstracts: ML4DM ‘21, ML4DM ‘22, and ML4DM ‘23.

Representation. The second session featured a panel discussion, facilitating the exchange of questions, ideas, and viewpoints between participants and speakers.

2.1. Robust Plan Evaluation Selection

Query optimizers in RDBMSs often select suboptimal execution plans due to inaccurate parameter estimates and assumptions that may not hold at runtime, leading to inadequate support for robust query optimization. With the growing interest in using ML to enhance data systems, promising advances have been made in query optimization. Inspired by these, we propose Roq (Robust Query Optimizer), a holistic framework using a risk-aware learning approach. Roq introduces a novel formalization of robustness in query optimization, a method for quantifying and measuring it through approximate probabilistic ML, and innovative strategies for query plan evaluation and selection. Our experiments show that Roq significantly improves robust query optimization compared to state-of-the-art methods.

2.2. Db2une: Tuning Under Pressure via DL

Modern database systems, such as IBM Db2, have numerous parameters or "knobs" that need precise configuration for optimal workload performance. Manually tuning these knobs is challenging, even for experts. We introduce Db2une, an automatic query-aware tuning system that uses deep learning to enhance performance while reducing resource usage. Db2une features a specialized transformer-based query-embedding pipeline called QBERT, which generates context-aware representations of query workloads. These representations are input into a stability-oriented, on-policy deep reinforcement learning model. Our system employs a multi-phased, database metadata-driven training approach, incorporating cost estimates, interpolation of these costs, and database statistics. This method allows Db2une to discover optimal tuning configurations efficiently without executing queries, making it scalable to very large workloads where query execution would be prohibitively expensive.

2.3. Data Acquisition for Improving Model Confidence

In the area of machine learning, the quality of training data is critical for model performance, yet the impact on model confidence has often been overlooked. Our work addresses this by enhancing the data acquisition process to improve model confidence, focusing on scenarios where only a limited number of samples can be selected from a large data pool. We propose two key methodologies, Bulk Acquisition (BA) and Sequential Acquisition (SA), along with efficient approximations (kNN-BA and kNN-SA) to target the most promising subsets. In addition, we introduce a Distribution-based Acquisition approach to generalize our methods across different settings. Extensive experiments demonstrate the effectiveness of our approaches, outperforming alternative baselines across various tasks and datasets.

2.4. Join Order Selection via Dueling Deep Q-Networks

Join order selection is a complex problem due to the exponentially growing search space as the number of tables increases. Traditional methods use heuristic algorithms to prune the search space but face limitations due to static heuristics, resulting in inefficient query optimizers that fail to adapt to feedback from the DBMS. Recent research suggests using Deep Reinforcement Learning (DRL) models to overcome these limitations by utilizing feedback. However, most studies focus on join order representations without extensively comparing DRL methods. This research proposes GTDD, a novel framework that integrates Graph Neural Networks (GNN), Tree-structured LSTM, and Dueling-DQN. GTDD demonstrates more stable training and superior join order discovery compared to traditional DBMS execution plans.

2.5. BiGG: A Novel Technique for Query Plan Representation Based on Graph Neural Nets

Learning representations for query plans play a pivotal role in machine learning-based query optimizers of database management systems. To this end, specific architectures known as tree models have been proposed in the literature to transform tree-structured query plans into representations that are learnable by downstream machine learning models. The quality of these representations significantly impacts the performance of subsequent tasks. In this context, to enhance the capabilities of tree models in representing query plans, we introduce graph neural networks (GNNs) to tree models for the first time and propose a novel tree model, BiGG, which utilizes Bidirectional GNN aggregated by Gated recurrent units (GRUs). Our experimental results demonstrate that BiGG markedly improves performance in cost estimation tasks and provides outstanding plan selection performance compared to state-of-the-art tree models.

3 ORGANIZERS AND PARTICIPANTS

This workshop is organized by Calisto Zuzarte, Verena Kantere, Jarek Szlichta, and Amin Kamali. The organizers additionally invited Xiaohui Yu, Andrew Chai, Chang Liu, and Baoming Chang to join the organizers to deliver talks, discuss, and exchange ideas with the audience on different topics related to the theme of the workshop.

Calisto Zuzarte is a Senior Technical Staff Member (STSM) in the Db2 development organization in IBM. His expertise is in database query optimization and has 60+ patents and 60+ research publications related to this area. His current interest is in the application of machine learning in query optimization and optimization in the Lakehouse environment.

Verena Kantere has held academic positions as: Professor at the School of EECS of the University of Ottawa, Assistant Professor at the School of ECE of the National Technical University of Athens (NTUA), Maître d'Enseignement et de

Recherche at the Centre Universitaire d' Informatique of the University of Geneva and Junior Assistant Professor at the Department of Electrical Engineering and Information Technology at the Cyprus University of Technology. She has conducted research for many years in the domain of data management, showing results in Peer-to-Peer systems, scientific data management, cloud data management, Big Data management, and analysis. She has received an Engineering Diploma, a Ph.D. from the NTUA, and a M.Sc. from the University of Toronto.

Jarek Szlichta is an Associate Professor and a Research Enhanced Faculty in responsible data science and AI under the Connected Minds at York University. He is also a Research Faculty Fellow at IBM Centre for Advanced Studies (CAS) and an Adjunct Professor at University of Waterloo. He serves as the co-director of the Data & AI Lab. His research concerns various topics in data science with special interests in self-driven data systems, graph data, large-scale machine learning, and responsible AI to obtain trustworthy insights from data. He is a recipient of the IBM CAS Faculty of the Year Award for tuning of the IBM Db2 system with machine learning.

Amin Kamali is a Data Scientist and a Ph.D. candidate in Digital Transformation and Innovation at the University of Ottawa. Over the past decade, he has occupied diverse roles at IBM, all centered around Data and AI. These roles have spanned a spectrum from Business Intelligence to Digital Transformation, Data Science, and Machine Learning. His primary research interest revolves around delving into the latest AI breakthroughs and their potential to transform data systems. Amin has been privileged to organize multiple workshops and tutorials for various audiences and to present at several academic conferences.

Xiaohui Yu is a Professor and the Graduate Program Director in the School of Information Technology, York University, Canada. He obtained his PhD degree from the University of Toronto. His research interests lie in the broad area of data science, with a particular focus on the intersection of data management and machine learning (ML). The results of his research have been published in top data science journals and conferences, such as SIGMOD, VLDB, ICDE, and TKDE. He regularly serves on the program committees of leading conferences and is an Associate/Area Editor for the IEEE Transactions on Knowledge and Data Engineering (TKDE), the ACM Transactions on Knowledge Discovery in Data (TKDD), and Information Systems. He is a General Co-Chair for the KDD 2025 conference. He has collaborated regularly with industry partners, and some research results have been incorporated into large-scale production systems.

Andrew Chai is an MSc student in Computer Science at York University, starting in 2023, supervised by Dr. Jarek Szlichta. He is also a Research Fellow at the IBM Centre for Advanced Studies. His research interests include self-driven data systems and explainable AI. His research results have been published in top data science conferences, including VLDB and SIGMOD.

Chang Liu is an MSc thesis student in Computer Science at the University of Ottawa. His research focuses on Join Order Selection using Deep Reinforcement Learning (DRL), employing various neural networks to capture information at different levels of a query. He is particularly interested in exploring techniques to address the unstable training processes often encountered in DRL.

Baoming Chang is an MSc thesis student in Computer Science at the University of Ottawa, focusing his research on machine learning for query optimization. His specific areas of interest include query plan representation utilizing graph neural networks, as well as enhancing the explainability and robustness of cost models. He holds his Bachelor of Engineering in Computer Science from the East China University of Science and Technology.

4 OUTCOMES

The workshop speakers present and share insights on applying ML to tackle data management issues. Discussions cover challenges, motivations, methods, and performance evaluations compared to traditional approaches. A panel discussion follows, allowing for an exchange of ideas, questions, and comments between the speakers and the audience. The event highlights that ML-enhanced data management goes beyond academic interest, showcasing its promising potential in real-world applications.

REFERENCES

[1] Bailu Ding, Sudipto Das, Ryan Marcus, Wentao Wu, Surajit Chaudhuri, and Vivek R. Narasayya. 2019. AI Meets AI: Leveraging Query Executions to Improve Index Recommendations. In *Proceedings of the 2019 International Conference on Management of Data* (SIGMOD '19), Association for Computing Machinery, Amsterdam, Netherlands, 1241–1258. DOI:https://doi.org/10.1145/3299869.3324957

[2] Yue Han, Guoliang Li, Haitao Yuan, and Ji Sun. 2022. AutoView: An Autonomous Materialized View Management System with Encoder-Reducer. *IEEE Trans. Knowl. Data Eng.* (2022), 1–1. DOI:https://doi.org/10.1109/TKDE.2022.3163195

[3] Connor Henderson, Spencer Bryson, Vincent Corvinelli, Parke Godfrey, Piotr Mierzejewski, Jaroslaw Szlichta, and Calisto Zuzarte. 2022. BLUTune: Query-informed Multi-stage IBM Db2 Tuning via ML. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (ACM CIKM '22), Association for Computing Machinery, New York, NY, USA, 3162–3171. DOI:https://doi.org/10.1145/3511808.3557117

[4] Connor Henderson, Vincent Corvinelli, Parke Godfrey, Piotr Mierzejewski, Jaroslaw Szlichta, and Calisto Zuzarte. 2023. BLUTune: Tuning Up IBM Db2 with ML. In *2023 IEEE 39th International Conference on Data Engineering (ICDE),* 3615-3618. DOI:https://doi.org/10.1109/ICDE55515.2023.00281

[5] Axel Hertzschuch, Claudio Hartmann, Dirk Habich, and Wolfgang Lehner. 2022. Turbo-Charging SPJ Query Plans with Learned Physical Join Operator Selections. *PVLDB* 15, 11 (2022), 2706–2718.

[6] Benjamin Hilprecht and Carsten Binnig. 2022. One Model to Rule them All: Towards Zero-Shot Learning for Databases. *ArXiv210500642 Cs* (January 2022). Retrieved March 27, 2022 from http://arxiv.org/abs/2105.00642

[7] Benjamin Hilprecht, Carsten Binnig, and Uwe Röhm. 2020. Learning a Partitioning Advisor for Cloud Databases. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (SIGMOD '20), Association for Computing Machinery, New York, NY, USA, 143–157. DOI:https://doi.org/10.1145/3318464.3389704

[8] Henry Liu, Mingbin Xu, Ziting Yu, Vincent Corvinelli, and Calisto Zuzarte. 2015. Cardinality Estimation Using Neural Networks. In *Proceedings of the 25th Annual International Conference on Computer Science and Software Engineering* (CASCON '15), IBM Corp., Riverton, NJ, USA, 53–59. Retrieved November 12, 2019 from

http://dl.acm.org/citation.cfm?id=2886444.2886453

[9] Chaohong Ma, Xiaohui Yu, Yifan Li, Xiaofeng Meng, and Aishan Maoliniyazi. 2022. FILM: A Fully Learned Index for Larger-Than-Memory Databases. *Proc. VLDB Endow.* 16, 3 (November 2022), 561–573. DOI:https://doi.org/10.14778/3570690.3570704

[10] Xiang Yu, Guoliang Li, Chengliang Chai, and N. Tang. 2020. Reinforcement Learning with Tree-LSTM for Join Order Selection. *2020 IEEE 36th Int. Conf. Data Eng. ICDE* (2020). DOI:https://doi.org/10.1109/ICDE48307.2020.00116

[11] Lixi Zhang, Chengliang Chai, Xuanhe Zhou, and Guoliang Li. 2022. LearnedSQLGen: Constraint-aware SQL Generation using Reinforcement Learning. In *Proceedings of the 2022 International Conference on Management of Data* (SIGMOD '22), Association for Computing Machinery, New York, NY, USA, 945–958. DOI:https://doi.org/10.1145/3514221.3526155

[12] Alexander Bianchi, Andrew Chai, Vincent Corvinelli, Parke Godfrey, Jarek Szlichta, Calisto Zuzarte. Db2une: Tuning Under Pressure via Deep Learning. *Proc. VLDB Endow.* accepted (to appear), 14 pages, 2024.

[13] Xuanhe Zhou, Lianyuan Jin, Ji Sun, Xinyang Zhao, Xiang Yu, Jianhua Feng, Shifu Li, Tianqing Wang, Kun Li, and Luyang Liu. 2021. DBMind: a self-driving platform in openGauss. *Proc. VLDB Endow.* 14, 12 (July 2021), 2743–2746. DOI:https://doi.org/10.14778/3476311.3476334