MLSkip: Data Skipping for ML Filters via Lightweight Metadata

Mihail Stoian*[1], Mark Gerarts*[2]†, Andreas Zimmerer[1], Pascal Ginter[3], Jan Van den Bussche[2], Andreas Kipf[1]

[1] University of Technology Nuremberg
{mihail.stoian,andreas.zimmerer,andreas.kipf}@utn.de
[2] Hasselt University
{mark.gerarts,jan.vandenbussche}@uhasselt.be
[3] Technical University of Munich
pascal.ginter@tum.de

*Equal contribution †Presenter

Database vendors increasingly support filtering tuples on the output of machine learning (ML) models. Traditional data skipping techniques cannot handle these ML-based predicates, raising new data management challenges. We present MLSkip, a study of data skipping techniques for ML filters operating on Parquet data.

Parquet organizes data into *row groups*. Each group contains a metadata section with the minimum and maximum values of its columns. Traditional techniques use this metadata to determine whether a row group can be skipped when evaluating basic predicates. We show that the existing min-max metadata is sufficient for data skipping on ML predicates by framing the problem as a formal verification question and applying techniques from the field of Neural Network Verification (NNV).

Although min-max metadata already yields promising results, we also introduce lightweight metadata alternatives to increase pruning efficiency. Most notably, we present the *bounded convex hull* as a tradeoff between metadata size and pruning effectiveness.

Our preliminary results on ReLU architectures show that, for filters with selectivity below 0.1%, min-max metadata achieves an average pruning effectiveness of 27.4% on TPC-H and TPC-DS tables. The bounded convex hull improves this to 38.31%. We also measure an end-to-end speedup of 1.07× over PyTorch in DuckDB.

We published an extended abstract of this work in the AIDB 2026 workshop [1], but MLSkip remains an active research topic. For this DBDBD talk, the goal is to present the main idea behind MLSkip and discuss our current research of applying MLSkip to semantic filters and joins.

References

[1] Mihail Stoian, Mark Gerarts, Pascal Ginter, Andreas Zimmerer, Jan Van den Bussche, and Andreas Kipf. MLSkip: Data skipping for ML filters via lightweight metadata. In *Proceedings of Workshops at the 52nd International Conference on Very Large Data Bases (VLDB 2026)*, AIDB: The 7th International Workshop on Applied AI for Database Systems and Applications, Boston, MA, USA, 2026.