BlockJoin: Efficient Matrix Partitioning Through Joins

Andreas Kunft, Asterios Katsifodimos, Sebastian Schelter, Tilmann Rabl, Volker Markl

Introduction

Late Materialization

Evaluation

A Context-Aware Operator Pipeline

Join Kernel

Late Materialization

Early Materialization

Goal. Enhance machine learning pipelines on normalized data in distributed settings.

Problem. The naïve approach materializes the join result row-wise, even though the final matrix is represented in a block-wise partitioning.

Solution. Logically fused operator pipeline that avoids materialization of the join result while creating the matrix-index and choses the best suited matrix-block materialization strategy based on a cost model.

A Context-Aware Operator Pipeline

Join Kernel

Late Materialization

Early Materialization

Goal. Enhance machine learning pipelines on normalized data in distributed settings.

Problem. The naïve approach materializes the join result row-wise, even though the final matrix is represented in a block-wise partitioning.

Solution. Logically fused operator pipeline that avoids materialization of the join result while creating the matrix-index and choses the best suited matrix-block materialization strategy based on a cost model.

Evaluation

The cost model validates the observations from the evaluation: More columns in the PK table favors late materialization and vise versa.