Databases for Big Data 2025/2026
Reading Assignment I
Dmytro Tupkalenko (89252101)

Title: An Empirical Evaluation of Columnar Storage Formats
Authors: Xinyu Zeng, Yulong Hui, Jiahong Shen, Andrew Pavlo, Wes McKinney, Huanchen Zhang

Overview of the main idea presented

The paper compares two column-based formats *Parquet* and *ORC*. This is done in two parts, the first one is the descriptional, where the specific feature implementations are described. The second one is the benchmarking, which is done by comparing the performance of the two formats for specific features (e.g. block compression) and for performance in specific workloads (e.g. ML).

Key findings/takeaways

The findings provide evidence-based insights to performance of the two formats across different scenarios - which can be used when making a choice among the two formats if the usage is known in advance. Additionally, the comparison identifies the strong sides of each of the formats - which can be used in designing future generation of the columnar formats.

Description of the system discussed and how it wasmodified/extended

The formats were kept without any changes.

Workloads/benchmarks used in evaluation

The authors designed benchmarking scheme, that tests performance on synthetic data that is generated according to some specific configuration (e.g. how sorted is the data, what is the ratio of null values, ...).