

УДК 004.651.4

Влияние метаданных на скорость чтения и записи LSM-дерева в Apache Paimon

Сурпин Вадим Павлович^(*)

Surpin.V.Pa@sberbank.ru

ПАО «Сбербанк», Москва, Россия

Аннотация. Рассмотрен процесс регулярного обновления таблицы данных в формате LSM-дерева в реализации Apache Paimon при помощи ETL-инструмента Apache Spark, моделирующий обновление данных в корпоративном хранилище (Lakehouse, DWH). Обнаружен рост времени обновления таблиц при выполнении множественных итераций записи и исследованы причины роста: исключено влияние сборщика мусора (Garbage Collector) Java-машины, влияние количества версий (snapshot), а также сжатия данных (compaction). Установлен линейный характер зависимости скорости записи Paimon от количества метаданных, а также слабая зависимость скорости чтения от него. Найден способ решения проблемы роста времени записи штатными средствами Paimon — кэшированием метаданных в памяти. Проведены начальные этапы исследования влияния количества бакетов на скорость работы с Paimon.

Ключевые слова: хранилища данных, DWH, lakehouse, базы данных, структуры данных, LSM-дерево, Apache Paimon, Apache Spark, Java, Compaction, метаданные

Введение. Хранилища данных прошли путь развития от классических DWH на основе реляционных и MPP баз данных, до data lake и lake house архитектур [1], которые обеспечивают хранение и обработку как структурированных, так и неструктурированных данных, а также обладают лучшими показателями масштабируемости на больших объёмах данных по сравнению с классическим подходом. В основе современных lake house решений лежат открытые табличные форматы (OTF). Перспективным OTF форматом является Apache Paimon, который появился в рамках проекта Apache Flink и рассчитан на потоковую обработку и использует хранение данных в виде LSM-дерева, которое также считается перспективным для HTAP-нагрузок. В данной работе исследуется его применение также в пакетном режиме для обеспечения совместимости с существующими витринами. Обнаружено влияние метаданных на скорость работы и рекомендованы меры по стабилизации скорости.

Методы и материалы; результаты. В работе исследовалась работа Open Table Forma-a Apache Paimon [2] в пакетном режиме при помощи фреймворка Apache Spark [3]. Задача состояла в многократном обновлении таблицы с первичным ключом, которая хранится в виде LSM-дерева [4] поверх формата parquet, и измерении времени обновления и последующего чтения таблицы в зависимости от количества проведённых итераций обновления. Исследование должно было подтвердить гипотезу о том, что время работы витрины стабильно и не зависит от количества итераций записи,

за исключением небольших колебаний при проведении compaction. Однако в процессе исследования были обнаружены существенные (порядка 50 %) колебания времени записи. Было проведено исследований причин их появления: влияния compaction, влияния количества файлов и количества версий данных, сборку мусора Java [5], ограничений количества открытых файловых дескрипторов. В результате была установлена зависимость от количества метаданных: оказалось, что именно они даже более существенно влияют на скорость работы Apache Paimon, чем количество данных и compaction. В результате обнаружено, что стабилизация работы витрины достигается при кэшировании метаданных в памяти пишущего процесса и уменьшении количества метаданных за счёт сокращения числа секций и бакетов. Последнее утверждение интересно тем, что обычные рекомендации по оптимизации скорости работы с данными в Paimon указывают как раз на увеличение количества секций и бакетов, но не исследуют верхние границы.

Заключение. В работе установлена существенная зависимость скорости записи витрин данных на основе Apache Paimon от количества метаданных. Также обнаружено, что влияние на чтение существенно меньше, чем на запись. Предложены рекомендации по ускорению и стабилизации времени работы за счет хранения метаданных в оперативной памяти. Предложены ограничения по количеству секций и бакетов в данных, которые уменьшат объем метаданных и ускорят работу витрин.

Список источников

- [1] Data Warehouse, Data Lake, Data Lakehouse, Data Fabric, Data Mesh — что это такое, и в чем разница между концепциями. URL: <https://habr.com/ru/articles/846296/> (дата обращения 26.09.2025).
- [2] Apache Paimon. A lake format that enables building a Realtime Lakehouse Architecture with Flink and Spark for both streaming and batch operations. URL: <https://paimon.apache.org/> (дата обращения 26.09.2025).
- [3] Apache Spark. Unified engine for large-scale data analytics. URL: <https://spark.apache.org/> (дата обращения 26.09.2025).
- [4] Understanding Apache Paimon's Consistency Model Part 1. URL: <https://jack-vanlightly.com/analyses/2024/7/3/understanding-apache-paimon-consistency-model-part-1> (дата обращения 26.09.2025).
- [5] Java Garbage Collection Basics. URL: <https://www.oracle.com/webfolder/technetwork/tutorials/obe/java/gc01/index.html> (дата обращения 26.09.2025).