ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.651

ОПТИМИЗАЦИЯ ХРАНЕНИЯ И ОБРАБОТКИ ФИНАНСОВЫХ ОТЧЁТОВ МАРКЕТПЛЕЙСА В DUCKDB

Серпинский Роман Эдуардович,
Студент группы ИУК5-11М
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
serpinskiyrea@student.bmstu.ru

Аскеров Салех Теймур оглы,
Студент группы ИУК5-11М
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
askerovst1@student.bmstu.ru

Буракова Мария Сергеевна,
ассистент кафедры ИУК5 КФ
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
m.burakova@bmstu.ru

Аннотация

В данной работе проводится экспериментальное исследование методов оптимизации хранения и обработки финансовых отчётов маркетплейса в аналитическом пайплайне на базе DuckDB. Рассматриваются две модели организации данных: монолитное хранение в виде одного оптимизированного Parquet-файла и партиционированное хранение в виде набора файлов, разделённых по временным интервалам. Анализируются особенности каждой модели, влияние файловой структуры на производительность аналитических запросов и стоимость этапов подготовки данных. Экспериментальная часть включает тестирование типовых финансовых запросов (агрегации, ранжирование, оконные функции, временные срезы) в режимах холодного и прогретого доступа к данным. Особое внимание уделяется оценке медианного и перцентильного времени выполнения запросов, а также стабильности работы при увеличении объёма данных. На основе полученных результатов сформулированы рекомендации по выбору модели хранения финансовых данных для задач аналитики и построения отчётности.

Ключевые слова: финансовые отчёты, DuckDB, Parquet, аналитическая обработка данных, оптимизация хранения, файловая структура базы данных, производительность запросов, cold и warm режимы, партиционирование, агрегатные запросы.

RE-OPTIMIZATION OF STORAGE AND PROCESSING OF MARKETPLACE FINANCIAL REPORTS IN DUCKDB

Serpinskiy Roman Eduardovich,
Student of the group IUK5-11M
Kaluga branch of the Moscow State Technical University named after N.E. Bauman
serpinskiyrea@student.bmstu.ru

Askerov Saleh Teymur oglu,
Student of the group IUK5-11M
Kaluga branch of the Moscow State Technical University named after N.E. Bauman
askerovst1@student.bmstu.ru

Burakova Maria Sergeevna,
Assistant of the Department of IUK5 KF
Kaluga branch of the Moscow State Technical University named after N.E. Bauman
m.burakova@bmstu.ru

ABSTRACT

This paper presents an experimental study of methods for optimizing the storage and processing of marketplace financial reports in an analytical pipeline based on DuckDB. Two data organization models are considered: monolithic storage in the form of a single optimized Parquet file and partitioned storage in the form of a set of files divided by time intervals. The features of each model are analyzed, as well as the impact of the file structure on the performance of analytical queries and the cost of data preparation stages. The experimental part includes testing of typical financial queries (aggregations, ranking, window functions, and time slices) in cold and warm data access modes. Special attention is paid to evaluating the median and percentile query execution times, as well as the stability of performance when the volume of data increases. Based on the results obtained, recommendations are formulated for choosing a financial data storage model for analytics tasks and reporting.

Keywords: financial reports, DuckDB, Parquet, analytical data processing, storage optimization, database file structure, query performance, cold and warm modes, partitioning, aggregate queries.

Введение

В условиях цифровой экономики объёмы финансовых данных, формируемых маркетплейсами и электронными торговыми площадками, постоянно увеличиваются. Финансовые отчёты, содержащие сведения о продажах, комиссиях, логистических и рекламных расходах, используются для аналитической обработки, построения отчётности и принятия управленческих решений. Эффективность работы с такими данными во многом определяется выбранной моделью их хранения и обработки. На практике финансовые данные часто предоставляются в виде Excel-отчётов, что затрудняет их использование в автоматизированных аналитических системах и приводит к

дополнительным затратам времени на подготовку данных. Переход к колоночным форматам хранения, таким как Parquet, и использование аналитических СУБД, в частности DuckDB, позволяют повысить производительность аналитических запросов, однако результат существенно зависит от физической организации данных. Существуют различные подходы к организации хранения финансовых данных, включая монолитное хранение в виде одного оптимизированного файла и партиционированное хранение в виде набора файлов. Выбор между этими подходами особенно важен для задач аналитики, так как они по-разному проявляют себя в режимах холодного и прогретого доступа к данным. Актуальность данной работы заключается в необходимости экспериментального сравнения указанных моделей хранения применительно к задачам финансовой аналитики. Целью исследования является анализ и оптимизация работы с финансовыми данными маркетплейса на основе оценки производительности типовых аналитических запросов в аналитическом пайплайне на базе DuckDB.

Цель исследования — определить наиболее эффективную модель хранения финансовых данных для аналитических задач в DuckDB.

Используемая система и критерии сравнения

В качестве аналитической системы в работе используется DuckDB [1] — встраиваемая аналитическая СУБД, предназначенная для выполнения OLAP-запросов и работы с колоночными форматами данных, в частности Parquet. DuckDB позволяет выполнять SQL-запросы напрямую над файловыми данными без развёртывания серверной инфраструктуры, что делает её удобной для построения локальных аналитических пайплайнов и проведения экспериментальных исследований производительности.

Вопрос выбора формата хранения данных для аналитических нагрузок подробно рассматривается в современных исследованиях аналитических СУБД. В работе Liu C. и соавторов показано, что колоночные форматы хранения, такие как Apache Parquet, обеспечивают значительный выигрыш производительности при выполнении агрегирующих и сканирующих запросов за счёт снижения объёма считываемых данных и эффективного использования векторизованной обработки [7].

Результаты, полученные в рамках данной работы, подтверждают выводы исследования [8]: при выполнении типовых OLAP-запросов над финансовыми данными маркетплейса DuckDB демонстрирует высокую стабильность времени выполнения и низкие накладные расходы, особенно в сценариях, не требующих развёртывания распределённой инфраструктуры. Это делает выбранную СУБД целесообразной основой для аналитических витрин и отчётных систем, ориентированных на оперативный анализ данных.

Для сравнения моделей хранения финансовых данных используются следующие метрики.

В качестве ETL-метрик измеряется время выполнения подготовительных операций:

create_views — время создания представлений над Parquet-файлами, характеризующее накладные расходы на инициализацию аналитического пайплайна;

analyze_and_warm — время материализации данных и построения статистик оптимизатора, отражающее стоимость прогрева данных перед серией аналитических запросов.

Основным критерием сравнения является время выполнения аналитических запросов. Для оценки производительности используются типовые финансовые запросы (агрегации, ранжирование, оконные функции, временные срезы), выполняемые в двух режимах:

cold — без предварительного прогрева данных;

warm — после материализации и построения статистик.

Для каждого запроса рассчитываются следующие статистические показатели:

avg — среднее время выполнения запроса;

p50 — медианное время выполнения, характеризующее типичную производительность;

p95 — 95-й перцентиль времени выполнения, отражающий стабильность работы системы и наличие редких задержек.

Использование указанных метрик позволяет комплексно оценить влияние модели хранения данных на производительность аналитического пайплайна в различных режимах доступа к финансовым данным.

Практическая часть

В рамках практической части была реализована и проведена серия экспериментальных исследований, направленных на сравнение двух моделей хранения финансовых данных: монолитной (SINGLE) [2] и партиционированной (MULTI). На основе исходных финансовых Excel-отчётов маркетплейса [3] была выполнена подготовка данных, включающая нормализацию структуры, приведение наименований полей к единому формату и заполнение отсутствующих значений.

Подготовленные данные были сохранены в формате Parquet [4] в двух вариантах: SINGLE — один оптимизированный файл с сортировкой по ключевым полям; MULTI — набор Parquet-файлов, партиционированных по году и номеру недели.

Для оценки влияния объёма данных на производительность был дополнительно сформирован увеличенный датасет путём синтетического масштабирования исходных данных с временным сдвигом. Это позволило сохранить структуру и распределение данных при контролируемом росте объёма.

ETL-метрики [5] используются для оценки накладных расходов, связанных с подготовкой данных и инициализацией аналитического пайплайна.

Метрика create_views

Позволяет оценить стоимость начальной инициализации аналитического пайплайна при подключении к данным.

Формула расчёта:

$$T_{create\_views} = t_{after} - t_{before}$$

где:

t_{before} — момент начала создания представлений,

t_{after} — момент завершения операции.

1.2 Метрика analyze_and_warm

Характеризует стоимость прогрева данных перед выполнением серии аналитических запросов и позволяет сравнить накладные расходы различных моделей хранения.

Формула расчёта:

$$T_{warm} = t_{after\_warm} - t_{before\_warm}$$

Таблица 1 — Результаты ETL-метрик

model	step	ms
SINGLE	create_views	2.399

SINGLE	analyze_and_warm	2.96
MULTI	create_views	2.409
MULTI	analyze_and_warm	5.249

2 Метрики выполнения аналитических запросов

Для обеих моделей хранения был сформирован единый набор типовых аналитических SQL-запросов [6], отражающих реальные сценарии финансовой аналитики: агрегирование показателей, ранжирование по прибыли, расчёт суммарных затрат, использование оконных функций и временных срезов. Каждый запрос выполняется 5 раз в двух режимах: cold и warm.

2.1 Среднее время выполнения (avg)

Используется для общей оценки производительности, но чувствительно к выбросам.

Формула расчета:

$$avg = \frac{1}{N}\sum_{i=1}^{N} t_i$$

2.2. Медианное время выполнения (p50)

Характеризует типичную производительность аналитического запроса и является основной метрикой сравнения.

Формула расчета:

$$p50 = \text{median}(t_1, t_2, ..., t_N)$$

2.3. 95-й перцентиль (p95)

Позволяет оценить стабильность выполнения запросов и наличие редких, но существенных задержек.

Формула расчета:

$$p95 = t_{\lceil 0.95 \cdot (N-1) \rceil}$$

Эксперименты проводились в двух режимах:

cold — выполнение запросов без предварительного прогрева данных;

warm — выполнение запросов после материализации данных и построения статистик оптимизатора.

Каждый запрос выполнялся многократно, после чего фиксировались временные характеристики выполнения.

Таблица 2 — Результаты метрик запросов

model	query_id	mode	n	avg_ms	p50_ms	p95_ms
MULTI	Q1	cold	5	4.739	4.712	4.997
MULTI	Q1	warm	5	1.144	1.152	1.237
MULTI	Q2	cold	5	4.515	4.515	4.637
MULTI	Q2	warm	5	0.862	0.872	0.875

MULTI	Q3	cold	5	4.976	4.692	5.481
MULTI	Q3	warm	5	1.014	0.968	1.123
MULTI	Q4	cold	5	6.893	6.966	7.087
MULTI	Q4	warm	5	3.528	3.394	4.164
MULTI	Q5	cold	5	8.338	8.369	8.744
MULTI	Q5	warm	5	1.565	1.551	1.68
SINGLE	Q1	cold	5	2.153	2.101	2.301
SINGLE	Q1	warm	5	1.103	1.066	1.207
SINGLE	Q2	cold	5	1.747	1.771	1.815
SINGLE	Q2	warm	5	0.843	0.827	1.023
SINGLE	Q3	cold	5	1.965	1.909	2.176
SINGLE	Q3	warm	5	1.053	1.031	1.174
SINGLE	Q4	cold	5	4.254	4.243	4.547
SINGLE	Q4	warm	5	3.286	3.233	3.419
SINGLE	Q5	cold	5	2.974	2.959	3.025
SINGLE	Q5	warm	5	1.387	1.38	1.442

Заключение

В ходе выполнения научно-исследовательской работы было проведено экспериментальное сравнение двух моделей хранения финансовых данных маркетплейса в аналитическом пайплайне на базе DuckDB: монолитной модели хранения в виде одного оптимизированного Parquet-файла (SINGLE) и партиционированной модели хранения в виде набора Parquet-файлов (MULTI).

На основании полученных экспериментальных данных установлено, что модель SINGLE характеризуется меньшими накладными расходами на этапах подготовки и прогрева данных. Значения ETL-метрик показывают, что время материализации и построения статистик оптимизатора для модели MULTI выше, что обусловлено необходимостью обработки большего числа файлов и дополнительных операций агрегации.

Анализ временных характеристик выполнения аналитических запросов показал, что в режиме холодного доступа модель SINGLE стабильно превосходит модель MULTI по медианному и перцентильному времени выполнения. В зависимости от типа запроса ускорение составляет от полутора до более чем двух раз. Данный эффект объясняется снижением накладных расходов на файловую систему и более эффективным последовательным чтением данных из одного файла.

В режиме прогретого доступа различия в производительности между моделями хранения существенно уменьшаются. Для большинства запросов медианное время выполнения становится сопоставимым, что указывает на доминирование вычислительных затрат над затратами ввода-вывода. Однако для запросов с оконными функциями и глобальными агрегациями модель SINGLE сохраняет преимущество по стабильности выполнения, что подтверждается более низкими значениями 95-го перцентиля.

Дополнительные эксперименты с увеличенным объёмом данных показали, что выявленные закономерности сохраняются при масштабировании: накладные расходы партиционированной модели растут быстрее, чем у монолитной, а преимущество SINGLE в холодном режиме остаётся выраженным.

Таким образом, по результатам проведённого исследования можно сделать вывод, что для задач финансовой аналитики, ориентированных на интерактивную работу, построение отчётности и BI-дашбордов, более предпочтительной является модель монолитного хранения данных. Партиционированная модель целесообразна в сценариях работы с большими историческими архивами и выраженной фильтрацией по временным диапазонам, однако требует дополнительных затрат на подготовку и прогрев данных. Полученные результаты позволяют сформулировать практические рекомендации по выбору модели хранения финансовых данных и могут быть использованы при проектировании аналитических витрин и отчётных систем на базе DuckDB.

Список литературы:

1. PostgreSQL Global Development Group. PostgreSQL Documentation. – URL: https://www.postgresql.org/docs/ (дата обращения: 10.12.2025). – Текст: электронный.
2. DuckDB Developers. DuckDB Documentation. – URL: https://duckdb.org/docs/stable/ (дата обращения: 10.12.2025). – Текст: электронный.
3. DuckDB Developers. DuckDB – an in-process SQL OLAP database management system. – URL: https://duckdb.org/ (дата обращения: 11.12.2025). – Текст: электронный.
4. Apache Parquet. Apache Parquet Documentation. – URL: https://parquet.apache.org/docs/ (дата обращения: 11.12.2025). – Текст: электронный.
5. Wildberries. Официальные отчёты и документация для партнёров маркетплейса Wildberries. – URL: https://seller.wildberries.ru/about-portal/ru/ru (дата обращения: 11.12.2025). – Текст: электронный.
6. Talend. ETL Performance Metrics and Monitoring. – URL: https://help.talend.com/r/en-US/8.0/data-integration/etl-performance-metrics (дата обращения: 12.12.2025). – Текст: электронный.
7. Liu C., Pavlenko A., Interlandi M., Haynes B. Data formats in analytical DBMSs: performance trade-offs and future directions // The VLDB Journal. – 2025. – Vol. 34, Article 30.
8. Kohn A. DuckDB-Wasm: Fast analytical processing for the Web // Proceedings of the VLDB Endowment. – 2025. – Vol. 15, pp. 3574–3584.

References:

1. The global PostgreSQL Development Team. PostgreSQL documentation. – URL: https://www.postgresql.org/docs/ (date of request: 10.12.2025). – Text: electronic.
2. DuckDB developers. DuckDB documentation. – URL: https://duckdb.org/docs/stable/ (date of request: 10.12.2025). – Text: electronic.
3. DuckDB developers. DuckDB is an in-process SQL OLAP database management system. – URL: https://duckdb.org/ (date of request: 11.12.2025). – Text: electronic.

4. Apache Parquet. Apache Parquet documentation. – URL: https://parquet.apache.org/docs/ (date of request: 11.12.2025). – Text: electronic.
5. Wildberries. Official reports and documentation for Wildberries marketplace partners. – URL: https://seller.wildberries.ru/about-portal/ru/ru (date of request: 11.12.2025). – Text: electronic.
6. Talend. ETL Performance Metrics and Monitoring. – URL: https://help.talend.com/r/en-US/8.0/data-integration/etl-performance-metrics (date of request: 12.12.2025). – Text: electronic.
7. Liu C., Pavlenko A., Interlandi M., Haynes B. Data formats in analytical DBMSs: performance trade-offs and future directions // The VLDB Journal. – 2025. – Vol. 34, Article 30.
8. Kohn A. DuckDB-Wasm: Fast analytical processing for the Web // Proceedings of the VLDB Endowment. – 2025. – Vol. 15, pp. 3574–3584.