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

УДК 004.651

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АРХИТЕКТУРЫ ХРАНЕНИЯ АНАЛИТИЧЕСКИХ ДАННЫХ ДЛЯ СЕРВИСА АНАЛИТИКИ

Серпинский Роман Эдуардович,
Студент группы ИУК5-21М
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
serpinskiyrea@student.bmstu.ru

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

Аскеров Салех Теймур оглы,
Студент группы ИУК5-21М
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
askerovst1@student.bmstu.ru

Аннотация

В данной работе проводится экспериментальное исследование архитектур хранения аналитических данных для сервиса аналитики Wildberries на базе ClickHouse. Рассматриваются четыре подхода к организации данных: Wide-table, Star schema, Normalized schema и модель Raw-Clean-Mart с заранее подготовленными аналитическими витринами. Цель исследования заключается в оценке влияния структуры хранения на скорость загрузки, объём занимаемого дискового пространства, время выполнения аналитических запросов и масштабируемость при росте объёма данных. Экспериментальная часть выполнена на синтетически масштабированных Wildberries-подобных данных объёмом до 500 МБ.

Ключевые слова: Wildberries, ClickHouse, аналитическое хранилище, архитектура данных, wide table, star schema, normalized schema, Raw-Clean-Mart, витрины данных, OLAP, масштабируемость, производительность запросов.

COMPARATIVE ANALYSIS OF THE STORAGE ARCHITECTURE OF ANALYTICAL DATA FOR AN ANALYTICS SERVICE

Roman E. Serpinski,
Student of the IUK5-21M group

Kaluga Branch of the Bauman Moscow State Technical University
serpinskiyrea@student.bmstu.ru

Maria S. Burakova,
Assistant of the Department of IUK5 CF
Kaluga Branch of the Bauman Moscow State Technical University
m.burakova@bmstu.ru

Saleh T. Askerov,
Student of group IUK5-21M
Bauman Moscow State Technical University (Kaluga Branch)
askerovst1@student.bmstu.ru

ABSTRACT

In this paper, we conduct an experimental study of analytical data storage architectures for the ClickHouse-based Wildberries analytics service. Four approaches to data organization are considered: Wide-table, Star schema, Normalized schema and the Raw-Clean-Mart model with pre-prepared analytical storefronts. The purpose of the study is to assess the impact of the storage structure on download speed, the amount of disk space occupied, the execution time of analytical queries, and scalability with increasing data volume. The experimental part was performed on synthetically scaled Wildberries-like data with a volume of up to 500 MB.

Keywords: Wildberries, ClickHouse, analytical storage, data architecture, wide table, star schema, normalized schema, Raw-Clean-Mart, data marts, OLAP, scalability, query performance.

Введение

Аналитические сервисы маркетплейсов работают с большим количеством разнородных данных: продажами, заказами, остатками, рекламными расходами, финансовыми показателями, карточками товаров и складской информацией. Для построения отчётов эти данные необходимо не только загрузить, но и организовать таким образом, чтобы пользовательские запросы выполнялись быстро и стабильно [2].

Выбор архитектуры хранения напрямую влияет на производительность аналитической системы. Одна Wide-table уменьшает количество JOIN, но может приводить к дублированию атрибутов. Star schema разделяет факты и измерения, что удобно для BI-моделей, однако требует соединения таблиц при выполнении запросов. Normalized schema снижает избыточность, но увеличивает сложность запросов. Подход Raw-Clean-Mart добавляет слои подготовки и витрин, ускоряя типовые отчёты ценой дополнительного хранения и более сложного ETL.

Целью работы является экспериментальное сравнение указанных архитектур хранения в ClickHouse применительно к задачам аналитики Wildberries.

Используемая система и критерии сравнения

В качестве аналитической системы используется ClickHouse — колоночная OLAP-СУБД, ориентированная на быстрые аналитические запросы по большим объёмам данных [5]. Для эксперимента применялся локальный режим ClickHouse с физическим хранением таблиц через параметр path. Все основные таблицы создавались на базе семейства MergeTree [1], так как этот механизм является базовым для хранения аналитических данных в

ClickHouse и поддерживает сортировку по ключу, эффективное чтение и фоновую организацию частей данных.

Критерии сравнения:

время создания схемы и загрузки данных;

объём занимаемого дискового пространства;

пиковое потребление оперативной памяти при подготовке данных [3];

среднее время выполнения типовых аналитических запросов;

масштабируемость при росте объёма данных.

Сравниваемые архитектуры хранения

В эксперименте были реализованы четыре модели организации данных:

Wide-table — одна широкая таблица sales_wide, содержащая факты и основные атрибуты без необходимости JOIN.

Star schema — fact_sales и таблицы измерений dim_product, dim_campaign, dim_warehouse.

Normalized schema — fact_sales и более детально нормализованные справочники брендов, категорий, кампаний, складов и регионов.

Raw-Clean-Mart — хранение исходного raw-слоя, очищенного clean-слоя и заранее рассчитанных mart-таблиц для типовых отчётов [6].

Практическая часть

Для эксперимента использовался CSV-файл объёмом 500 МБ, содержащий 6671764 строки. Данные имитируют структуру отчётов Wildberries: дата заказа, артикул, бренд, рекламная кампания, склад, показы, клики, рекламные расходы, количество заказов, выручка, себестоимость и остатки. На основе одного и того же источника для каждой архитектуры создавались физические таблицы ClickHouse, после чего выполнялась серия одинаковых аналитических запросов.

Типовые запросы:

Q1 — выручка, рекламные расходы, прибыль и ROAS по дням и брендам;

Q2 — ROAS по рекламным кампаниям;

Q3 — топ товаров по прибыли;

Q4 — складская динамика по дням.

Таблица 1 — Результаты загрузки и хранения

Архитектура	Загрузка, сек	Хранение, МБ	Пиковая RAM, МБ
Wide-table	3,278	217,6	1010,2
Star schema	4,157	217,8	1017,1
Normalized schema	4,642	217,8	1020,2
Raw-Clean-Mart	8,192	529,4	1351,6

Наименьшие накладные расходы на загрузку показала архитектура Wide-table: создание схемы и загрузка данных заняли 3,278 секунды. Это объясняется отсутствием дополнительных преобразований — все данные загружаются в одну денормализованную таблицу без необходимости формирования вспомогательных справочников. Star schema и Normalized schema требуют больше времени из-за формирования справочников и заполнения fact-таблицы [4]. Самой дорогой по подготовке оказалась Raw-Clean-Mart: 8,192 секунды, что объясняется созданием raw-слоя, clean-слоя и нескольких агрегированных витрин. По объёму хранения Wide-table, Star schema и Normalized schema занимают около 218 МБ, тогда как Raw-Clean-Mart занимает 529,4 МБ.

Таблица 2 — Результаты аналитических запросов

Архитектура	Q1, сек	Q2, сек	Q3, сек	Q4, сек
Wide-table	0,299	0,409	0,450	0,402

Star schema	0,315	0,979	0,644	0,480
Normalized schema	0,322	1,032	0,560	0,571
Raw-Clean-Mart	0,309	0,308	0,310	0,310

По результатам запросов наиболее стабильное время показала архитектура Raw-Clean-Mart: все четыре запроса выполнялись примерно за 0,31 секунды. Это связано с тем, что запросы обращаются не к детализированным исходным данным, а к заранее подготовленным аналитическим витринам. Wide-table также показала хорошие результаты и стала наиболее сбалансированным вариантом среди схем без предварительно рассчитанных mart-таблиц. Star schema и Normalized schema заметно проиграли на запросе Q2, где требуется агрегация ROAS по рекламным кампаниям и брендам с использованием JOIN.

Масштабируемость

Для оценки масштабируемости эксперимент дополнительно выполнялся на объёмах 100, 250 и 500 МБ. Результаты показывают важную закономерность: при росте данных в 5 раз среднее время запросов у Raw-Clean-Mart выросло только в 1,13 раза, у Wide-table — в 1,35 раза, а у Star schema и Normalized schema — примерно в 1,85 раза. Следовательно, витринный подход лучше всего масштабируется по скорости пользовательских отчётов, но требует почти пятикратного роста хранения и более дорогого этапа подготовки данных. Wide-table показывает хороший компромисс между масштабируемостью, скоростью и размером хранения. Схемы с JOIN масштабируются хуже именно в аналитических запросах, где нужно соединять fact-таблицу со справочниками.

Заключение

В ходе выполнения научно-исследовательской работы было проведено экспериментальное сравнение четырёх архитектур хранения аналитических данных для сервиса аналитики Wildberries на базе ClickHouse: Wide-table, Star schema, Normalized schema и Raw-Clean-Mart. Эксперимент выполнялся на наборе данных объёмом 500 МБ, а масштабируемость дополнительно оценивалась на наборах 100, 250 и 500 МБ.

На основании результатов можно сделать вывод, что Raw-Clean-Mart является наиболее эффективной архитектурой для быстрого выполнения типовых пользовательских отчётов и BI-дашбордов. Она обеспечивает стабильное время запросов при росте данных, но требует значительно большего объёма хранения и более дорогой подготовки.

Wide-table является наиболее практичным компромиссом для MVP и средних аналитических систем: она быстро загружается, занимает меньше места и показывает хорошую скорость запросов без сложного ETL. Star schema и Normalized schema целесообразны тогда, когда приоритетом является логическая чистота модели, управляемость справочников и развитие аналитической структуры, однако для интерактивных отчётов они могут уступать из-за затрат на соединение таблиц.

Таким образом, с учётом целей разрабатываемого сервиса аналитики Wildberries, ориентированного на регулярное построение отчётов, BI-дашбордов и быстрый доступ пользователей к подготовленным аналитическим показателям, архитектура Raw-Clean-Mart является наиболее рациональным и обоснованным выбором. Она обеспечивает лучший баланс между производительностью, масштабируемостью, качеством подготовки данных и пригодностью для промышленной эксплуатации, поэтому именно данный подход рекомендуется использовать в качестве основной архитектуры хранения аналитических данных.

Список литературы:

1. Stonebraker M., Abadi D. J., Batkin A. et al. C-Store: A Column-Oriented DBMS // Proceedings of the 31st International Conference on Very Large Data Bases (VLDB). – 2005. – Pp. 553–564. – URL: http://www.vldb.org/conf/2005/papers/p553-stonebraker.pdf.
2. Abadi D. J., Madden S., Ferreira M. C. Integrating Compression and Execution in Column-Oriented Database Systems // Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. – 2006. – Pp. 671–682. – DOI 10.1145/1142473.1142548.
3. Lakshman A., Malik P. Cassandra: A Decentralized Structured Storage System // ACM SIGOPS Operating Systems Review. – 2010. – Vol. 44, No. 2. – Pp. 35–40. – DOI 10.1145/1773912.1773922.
4. Stonebraker M., Madden S. The End of an Architectural Era (It's Time for a Complete Rewrite) // Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB). – 2007. – Pp. 1150–1160. – URL: http://www.vldb.org/conf/2007/papers/industrial/p1150-stonebraker.pdf.
5. Abadi D. J., Boncz P., Harizopoulos S., Idreos S., Madden S. The Design and Implementation of Modern Column-Oriented Database Systems // Foundations and Trends in Databases. – 2012. – Vol. 5, No. 3. – Pp. 197–280. – DOI 10.1561/1900000024.
6. Stonebraker M., Çetintemel U., Zdonik S. The 8 Requirements of Real-Time Stream Processing // ACM SIGMOD Record. – 2005. – Vol. 34, No. 4. – Pp. 42–47. – DOI 10.1145/1107499.1107504.

References:

1. Stonebraker M., Abadi D. J., Batkin A. et al. C-Store: A Column-Oriented DBMS // Proceedings of the 31st International Conference on Very Large Data Bases (VLDB). – 2005. – Pp. 553–564. – URL: http://www.vldb.org/conf/2005/papers/p553-stonebraker.pdf.
2. Abadi D. J., Madden S., Ferreira M. C. Integrating Compression and Execution in Column-Oriented Database Systems // Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. – 2006. – Pp. 671–682. – DOI 10.1145/1142473.1142548.
3. Lakshman A., Malik P. Cassandra: A Decentralized Structured Storage System // ACM SIGOPS Operating Systems Review. – 2010. – Vol. 44, No. 2. – Pp. 35–40. – DOI 10.1145/1773912.1773922.
4. Stonebraker M., Madden S. The End of an Architectural Era (It's Time for a Complete Rewrite) // Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB). – 2007. – Pp. 1150–1160. – URL: http://www.vldb.org/conf/2007/papers/industrial/p1150-stonebraker.pdf.
5. Abadi D. J., Boncz P., Harizopoulos S., Idreos S., Madden S. The Design and Implementation of Modern Column-Oriented Database Systems // Foundations and Trends in Databases. – 2012. – Vol. 5, No. 3. – Pp. 197–280. – DOI 10.1561/1900000024.
6. Stonebraker M., Çetintemel U., Zdonik S. The 8 Requirements of Real-Time Stream Processing // ACM SIGMOD Record. – 2005. – Vol. 34, No. 4. – Pp. 42–47. – DOI 10.1145/1107499.1107504.