Выбор и описание СУБД для проекта прогнозирования спроса FMCG-компании

Санджар Джаводов

17 декабря 2024 г.

Введение

В рамках проекта по разработке системы прогнозирования спроса для FMCG-компании была выбрана система управления базами данных (СУБД) ClickHouse. Этот документ подробно описывает характеристики ClickHouse и обосновывает ее выбор для нашего проекта.

Общее описание ClickHouse

ClickHouse - это колоночная система управления базами данных с открытым исходным кодом, разработанная компанией Yandex. Она специально оптимизирована для выполнения аналитических запросов над большими объемами данных в режиме реального времени.

Ключевые характеристики ClickHouse

Колоночное хранение данных

ClickHouse использует колоночное хранение данных, что обеспечивает:

- Эффективное сжатие данных (до 10 раз по сравнению с обычными СУБД)
- Быстрое выполнение агрегатных запросов
- Оптимальную работу с временными рядами

Высокая производительность

- Скорость обработки запросов достигает миллиардов строк в секунду на одном сервере
- Линейная масштабируемость производительности при добавлении новых серверов

- Параллельная обработка запросов на многоядерных процессорах

Масштабируемость

- Поддержка распределенных запросов
- Возможность горизонтального масштабирования путем добавления новых узлов
- Автоматическое шардирование и репликация данных

SQL-совместимость и расширения

- Поддержка стандартного SQL с дополнительными расширениями
- Широкий набор встроенных функций для анализа данных
- Возможность создания пользовательских агрегатных функций на C++

Типы данных и индексы

- Поддержка широкого спектра типов данных, включая массивы и вложенные структуры
- Специальные типы данных для работы с IP-адресами и географическими координатами
- Возможность создания вторичных индексов для ускорения запросов

Интеграция с инструментами ML

- Встроенные функции для работы с моделями машинного обучения
- Возможность интеграции с популярными библиотеками ML через внешние функции
- Поддержка экспорта данных в форматы, удобные для обработки ML-алгоритмами

Преимущества ClickHouse для нашего проекта

Эффективная работа с временными рядами

Колоночное хранение и оптимизированные алгоритмы сжатия идеально подходят для хранения и анализа исторических данных о продажах.

Быстрое выполнение аналитических запросов

Высокая скорость обработки запросов позволит оперативно анализировать большие объемы данных для прогнозирования спроса.

Масштабируемость

По мере роста объема данных и усложнения аналитических задач, ClickHouse позволит легко масштабировать систему.

Поддержка ML-операций

Встроенные функции и возможность интеграции с ML-библиотеками упростят процесс разработки и внедрения моделей прогнозирования.

Схема базы данных

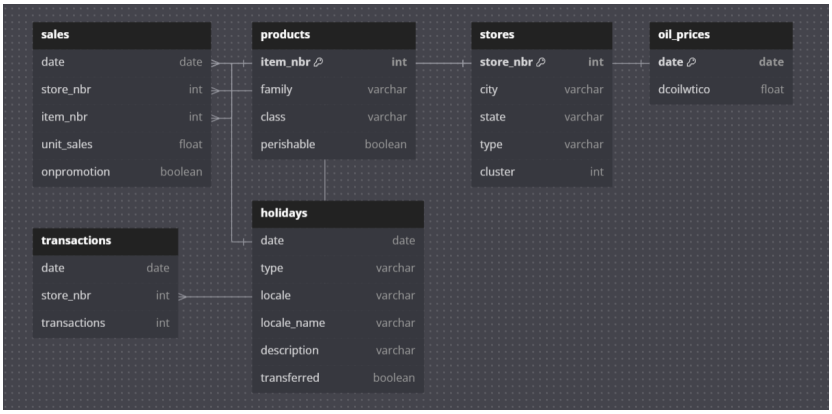

Рис. 1: Схема

Для эффективного хранения и обработки данных в ClickHouse была разработана следующая схема базы данных:

Таблицы и их структура

- **sales**: Центральная таблица с данными о продажах
  - date: Date
  - store_nbr: UInt32
  - item_nbr: UInt32
  - unit_sales: Float32
  - onpromotion: UInt8
- **products**: Информация о товарах
  - item_nbr: UInt32 (Первичный ключ)
  - family: String

  - class: String
  - perishable: UInt8

- **stores**: Данные о магазинах
  - store_nbr: UInt32 (Первичный ключ)
  - city: String
  - state: String
  - type: String
  - cluster: UInt32

- **oil_prices**: Ежедневные цены на нефть
  - date: Date (Первичный ключ)
  - dcoilwtico: Float32

- **transactions**: Количество транзакций в магазинах
  - date: Date
  - store_nbr: UInt32
  - transactions: UInt32

- **holidays**: Информация о праздниках и событиях
  - date: Date
  - type: String
  - locale: String
  - locale_name: String
  - description: String
  - transferred: UInt8

Особенности схемы

- **Денормализация**: Схема частично денормализована для оптимизации производительности аналитических запросов в ClickHouse.
- **Типы данных**: Использованы эффективные типы данных ClickHouse, такие как UInt32 для целочисленных идентификаторов и Float32 для числовых значений с плавающей точкой.
- **Сортировка**: Таблицы отсортированы по ключевым полям для ускорения запросов, например, таблица sales отсортирована по (date, store_nbr, item_nbr).
- **Движок таблиц**: Используется движок MergeTree для всех таблиц, что обеспечивает высокую производительность вставки и запросов.

Преимущества разработанной схемы

- **Оптимизация для аналитики**: Структура таблиц оптимизирована для выполнения аналитических запросов, характерных для задач прогнозирования спроса.
- **Гибкость**: Схема позволяет легко добавлять новые атрибуты и таблицы по мере развития проекта.
- **Эффективное хранение**: Использование специализированных типов данных ClickHouse обеспечивает эффективное хранение и быструю обработку данных.
- **Поддержка временных рядов**: Структура таблиц, особенно sales и oil_prices, оптимизирована для работы с временными рядами.

Заключение

Выбор ClickHouse в качестве СУБД для нашего проекта прогнозирования спроса в FMCG-компании обеспечивает необходимую производительность, масштабируемость и гибкость. Эта СУБД позволит эффективно хранить и анализировать большие объемы данных о продажах, что критически важно для точного прогнозирования спроса и оптимизации бизнес-процессов компании.