Курсовая работа
Программный проект на тему

"Полиморфные куски данных в таблицах типа MergeTree в ClickHouse"

Выполнил студент группы 176, 3 курса, Попов Антон Дмитриевич
Руководитель КР: Руководитель группы разработки ClickHouse Миловидов Алексей Николаевич

ClickHouse

- ClickHouse - распределенная колоночная СУБД для аналитических запросов
- Исходный код написан на C++ и находится в открытом доступе

MergeTree

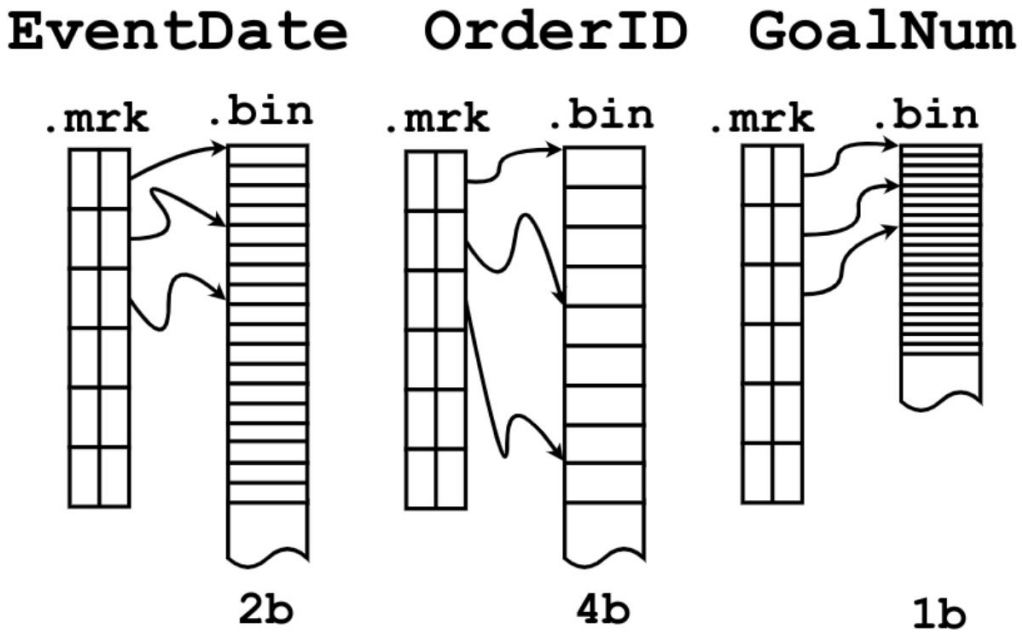
- Поддерживает индексы, партиционирование, репликацию
- Данные хранятся в кусках, отсортированных по первичному ключу
- При каждой вставке создается новый кусок
- Далее происходят фоновые слияния для более оптимального хранения данных

MergeTree

- Кусок разделяется на гранулы. Затем первая строка каждой гранулы помечается соответствующим значением первичного ключа и засечкой
- Засечка хранит смещения, по которым можно прочитать гранулу, начиная с первой строки, а также количество строк в грануле
- В кусках каждый столбец хранится в отдельном файле
- Засечки пишутся для каждого файла с данными.

MergeTree

Визуализация хранения данных в кусках MergeTree



Особенности поколоночного хранения

- Быстрые манипуляции со столбцами
- Более оптимальное сжатие данных
- Более оптимальное чтение одного столбца

Недостаток: вставки маленького количества строк в таблицы с большим количеством столбцов является неэффективным, поскольку требуют создания большого количества файлов.

Цель и актуальность

Цель работы - улучшение производительности мелких вставок в таблицы семейства MergeTree при помощи создания новых форматов хранения кусков данных.

Работа является актуальной, потому что решает одну из важных проблем ClickHouse, с которой сталкивается множество пользователей.

Существующие решения

- Буферизация данных с помощью сторонних приложений
- Встроенные в ClickHouse решения: Buffer таблицы, интеграция с Apache Kafka

Новые форматы хранения данных

- Компактный. Все столбцы хранятся в одном файле. Засечки также записываются в один файл.
- Формат с буферизацией данных в оперативной памяти со сбросом на диск.

Из одного формата в другой куски переходят во время слияний при достижении порогов по размеру, выраженному в количестве строк или байт.

Компактный формат

- Каждая гранула записывается в виде набора подряд сериализованных столбцов
- Гранулы записываются подряд в один файл
- Одна засечка представляет собой количество строк в грануле и массив смещений для каждого столбца
- Константное количество файлов вместо пропорционального количеству столбцов

Формат в оперативной памяти

- Кусок хранит сортированный блок, находящийся в оперативной памяти
- Этот блок является единственной гранулой
- Реализована опциональная поддержка Write Ahead Log (WAL)
- В WAL записываются все приходящие при INSERT-е, а также при полученные при репликации, блоки в Native формате ClickHouse
- WAL позволяет восстанавливать данные в случае аварийного завершения работы сервера

Этапы реализации

- Рефакторинг, выделение интерфейсов для чтения и записи куска, а также хранения его метаинформации
- Реализация данных интерфейсов для кусков в новых форматах.
- Реализация логики для ALTER запросов.
- Тестирование. Особенно важно проверить различные краевые случаи в реплицированных таблицах

Результаты

- Реализованы два новых формата хранения данных в кусках таблиц семейства MergeTree. Были добавлены настройки таблиц, регулирующие формат хранения данных.
- У пользователей во многих случаях появится возможность отказаться от дополнительнительных приложений для буферизации данных

Сравнение производительности

Данные вставлялись в таблицу hits тестового датасета Яндекс Метрики, которая имеет 133 столбца. Измерялся показатель QPS (Query Per Second) в зависимости от количества вставляемых строк и количества потоков.

	1 строка, 1 поток	100, 1	1000, 1	1, 4	100, 4	1000, 4
Широкий формат	118 QPS	92	80	13	11	16
Компактный формат	432	391	258	1233	1151	773
Формат в памяти	423	464	242	1126	1352	705
Buffer таблица	688	531	334	2391	1718	563