Clickhouse دیتابیسی برای فصول تحلیلی!

ارائه توسط: محمد عرب انواری
با حمایت : DataHobbies

زمستان 1401

راجع به چه چیزهایی قراره صحبت بشه؟

- یک مقدمه کوتاه راجع به کلیک هاوس
- تعریف یک چالش واقعی
- راه‌های حل چالش مورد نظر
- مقایسه عملی کلیک هاوس و MariaDB
- کمی عمیق‌تر راجع به کلیک هاوس

یک مقدمه کوتاه راجع به کلیک هاوس

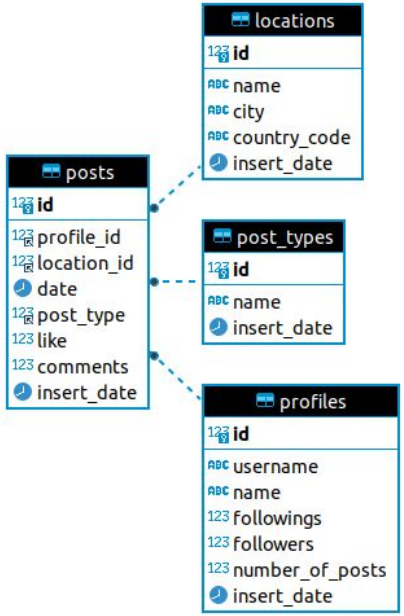
کلیک هاوس چیه واقعا؟

- یک دیتابیس column-oriented برای مصارف تحلیلی
- پشتیبانی کامل از SQL
- سرعت بالا برای استفاده‌های OLAP

تعریف یک چالش واقعی

تعریف مسئله

- نیاز به تحلیل داده‌های گردآوری شده از اینستاگرام
- دیتاست اینستاگرام (لینک Kaggle)
- ایجاد جداول از روی فایل‌های CSV



تعریف چالش

- نیاز به استخراج information از دل دیتای خام
- درخواست ها
  - جمع تعداد لایک و کامنت با توجه به نوع پست
  - مکانی برای هر کاربر که بیشتر در آنجا پست گذاشته
  - Engagement Rate هر کاربر (میانگین همه پست ها)
- ویژگی کوئری های OLAP

راه‌های حل چالش مورد نظر

pandas

استفاده از راه حل‌های سطح کد (Pandas)

نقاط روشن

- همه دیتا روی رم لود میشه -> سرعت بالا
- توابع به کمک Cython پیاده سازی شدن و سرعت بالایی دارن
- پشتیبانی از محاسبات vectorize

نقاط تاریک

- همه دیتا روی رم لود میشه -> نیاز به رم بالا برای دیتاهای حجیم

استفاده از ابزارهای بیگ دیتا (مثلا Spark و HDFS)

نقاط روشن

- چون دقیقا برای همین منظور طراحی شده این سیستم‌ها و به خوبی از پس این کوئری ها بر میاد.
  - شبیه به pandas ولی بهینه برای دیتاهای حجیم

نقاط تاریک

- عدم پشتیبانی native از SQL
- یک اکوسیستم جدا داره که نیاز به یادگیری داره

دیتابیس‌های رایج مثل MariaDB یا Postgre

نقاط روشن

- سهولت استفاده
- جامعه کاربری بزرگ

نقاط تاریک

- سرعت پایین
  - نگهداری سطری داده
- طراحی شده برای نیازهای متفاوت

ClickHouse

استفاده از دیتابیس‌های ستون محور : Clickhouse

نقاط روشن

- طراحی شده برای محاسبه کوئری‌های تحلیلی در کمترین زمان برای داده حجیم
  - Column-Oriented
  - فشرده سازی
- پشتیبانی کامل از SQL
- سهولت استفاده

نقاط تاریک

- یک سری از امکانات دیتابیس‌های OLTP رو از ما میگیره

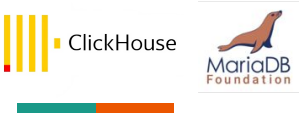
مقایسه عملی کلیک هاوس و MariaDB

راه اندازی Clickhouse و MariaDB

- استفاده از docker برای راه اندازی دیتابیس‌ها
  - ساختار docker compose
  - نحوه مانیتور کردن سرویس‌ها
- اتصال به دیتابیس از طریق dbeaver

کد توسعه داده شده

- توسعه کد به صورت ماژولار
  - ماژول های utils
  - ماژول های متفرقه
  - فایل main و نحوه اجرای برنامه
- خروجی نهایی کد توسعه داده شده



نتایج نهایی مقایسه / زمان پاسخ

Clickhouse

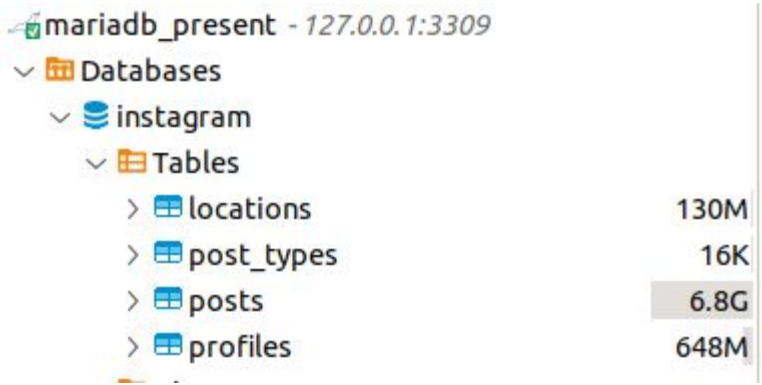
- زمان ورود داده‌ها: 29 دقیقه
- زمان پاسخ به کوئری تحلیلی اول: 0.98 ثانیه
- زمان پاسخ به کوئری تحلیلی دوم: 8.46 ثانیه
- زمان پاسخ به کوئری تحلیلی سوم: 8.44 ثانیه

MariaDB

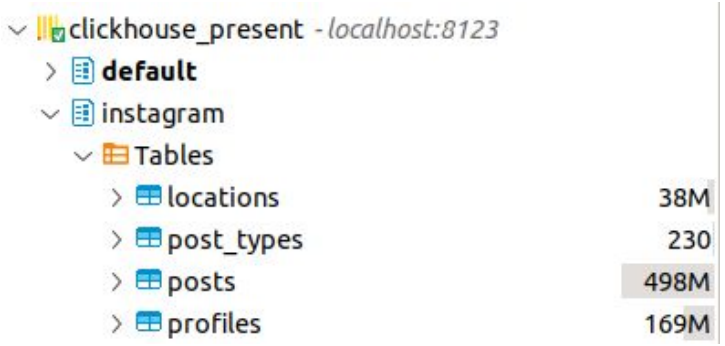
- زمان ورود داده‌ها: 9 ساعت
- زمان پاسخ به کوئری تحلیلی اول: 15 دقیقه
- زمان پاسخ به کوئری تحلیلی دوم: بیش از 8 ساعت!
- زمان پاسخ به کوئری تحلیلی سوم: 24 دقیقه

نتایج نهایی مقایسه / فشرده‌سازی

MariaDB



Clickhouse

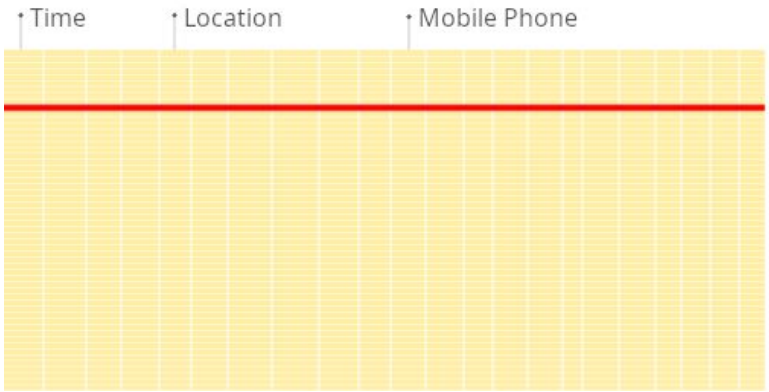


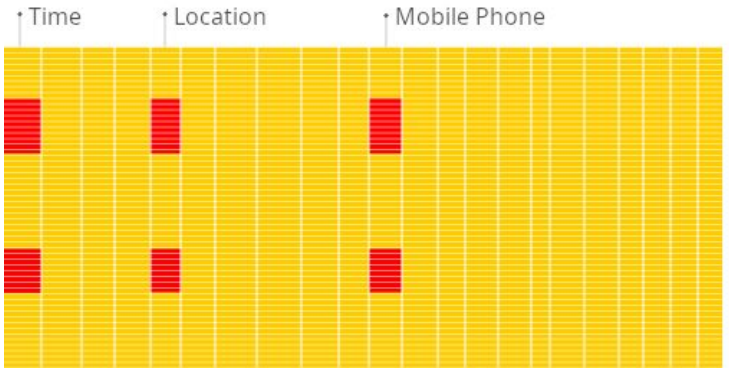
کمی عمیق تر راجع به کلیک هاوس

ClickHouse

ستون محور بودن کلیک هاوس

- نحوه ذخیره‌سازی
- فشرده‌سازی





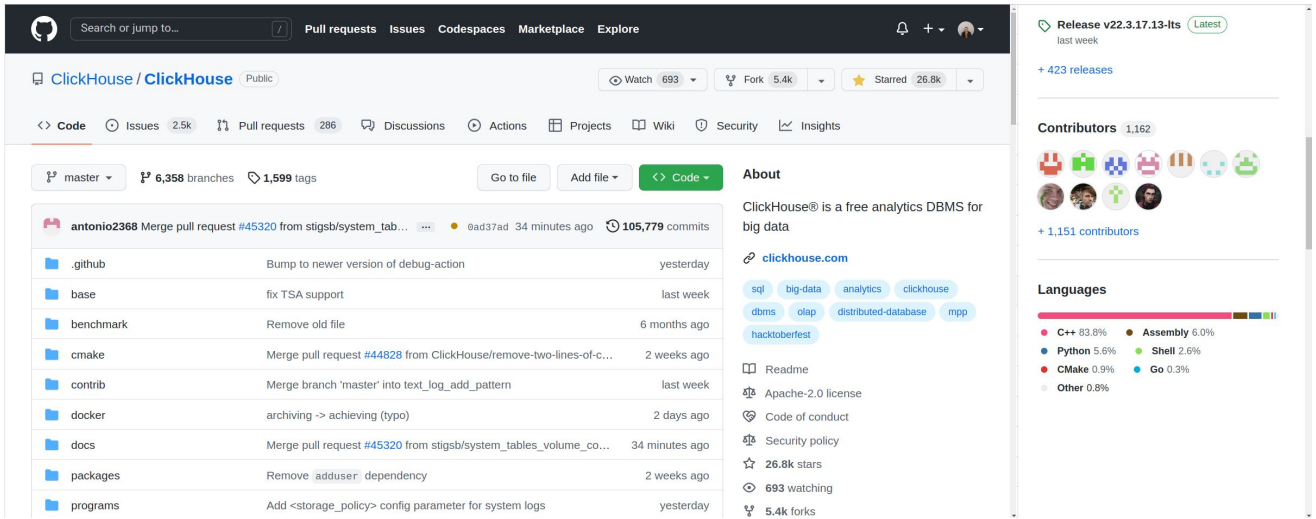
ClickHouse

پشتیبانی کامل از SQL

- ارائه یک سری امکانات بیشتر در دستورهای SQL
  - ASOF Join
- ارائه توابع متنوع و مفید
  - توابع محاسبات اماری
    - STD, Median, Quartile
  - توابع Aggregation
    - ArrayGroup
    - argMax

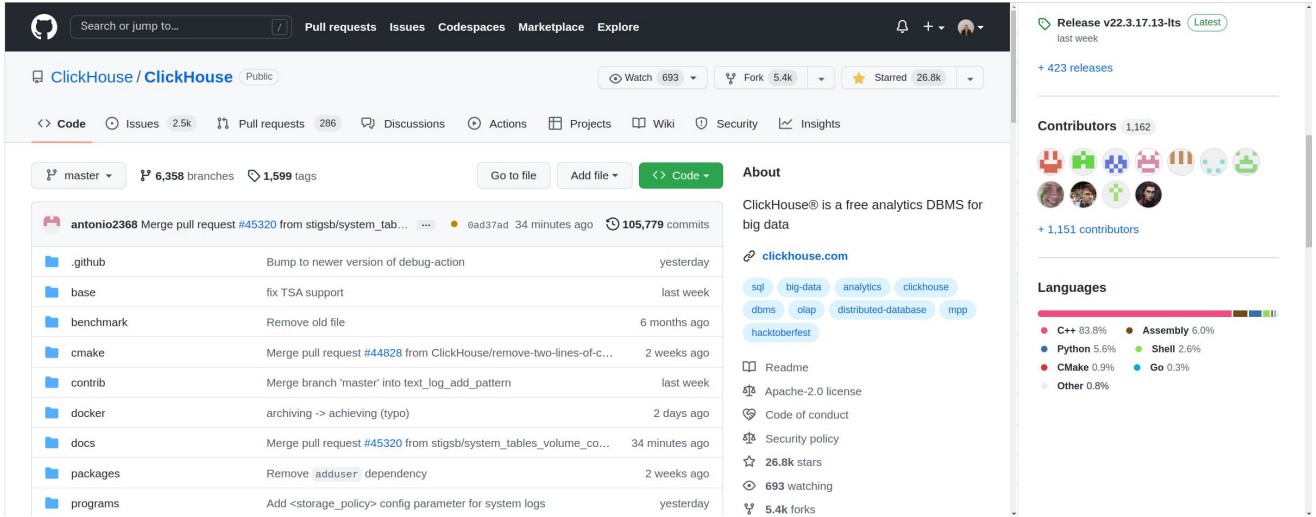
ClickHouse

OpenSource بودن و Community فعال



ClickHouse

OpenSource بودن و Community فعال



ClickHouse

هزینه بسیار بالای عملیات های delete و update

- فرکانس بالای همچنین عملیات هایی باعث down شدن Clickhouse می‌شود
- نحوه داشتن جدولی که داده‌ها در آن اپدیت شوند، شدنیست اما متفاوت است.
  - کوئری‌های همچین جدولی نیز متفاوت است
  - برای مثال پست‌های اینستاگرام

ClickHouse

کلیدها در Clickhouse

- عدم پشتیبانی از Foreign Key
  - نحوه ارتباط بین جداول
- نحوه عملکرد Primary key در کلیک هاوس
  - Granul
  - Sorting Key
  - Primary Key
  - مطالعه بیشتر در مستندات کلیک هاوس و این پست

ClickHouse

نکات تکمیلی

- بهینه نبودن انتخاب ستون‌های زیاد در جداول بزرگ
- فشار زیاد بر روی clickhouse هنگام insert های زیاد در زمان کم
  - راه حل: جداول buffer

و در آخر ...

Clickhouse یک انتخاب مناسب برای مصارف تحلیلی می‌باشد.

ممنون از همراهیتون!

با تشکر از DataHobbies برای برنامه ریزی همراهی در آماده سازی این ارائه

تمامی کدها و فایل ارائه در مخزن گیت‌هاب موجود می‌باشد
https://github.com/anvaari/Clickhouse_VS_MariaDB

خوشحال میشم اگر سوالی داشتین باهام در ارتباط باشین
anvaari@proton.me