



Clickhouse دیتابسی برای فصول تحلیلی!

ارائه توسط: محمد عرب انواری
با حمایت : [DataHobbies](https://DataHobbies.com)

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؟

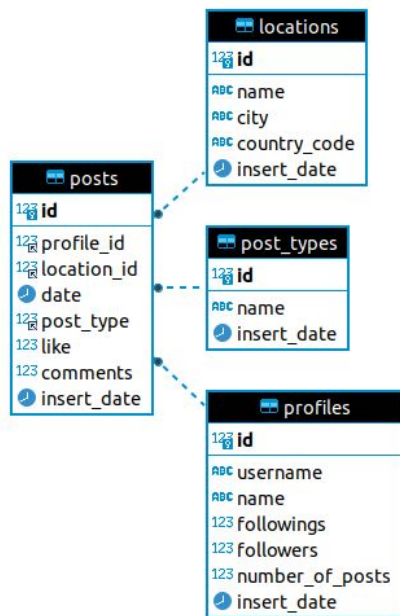
- یک دیتابیس column-oriented برای مصارف تحلیلی
- پشتیبانی کامل از SQL
- سرعت بالا برای استفاده‌های OLAP

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

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 کوئری‌ها 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

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

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

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



نتایج نهایی مقایسه / فشرده سازی

MariaDB

mariadb_present - 127.0.0.1:3309

Databases

instagram

Tables

> locations	130M
> post_types	16K
> posts	6.8G
> profiles	648M

Clickhouse

clickhouse_present - localhost:8123

default

instagram

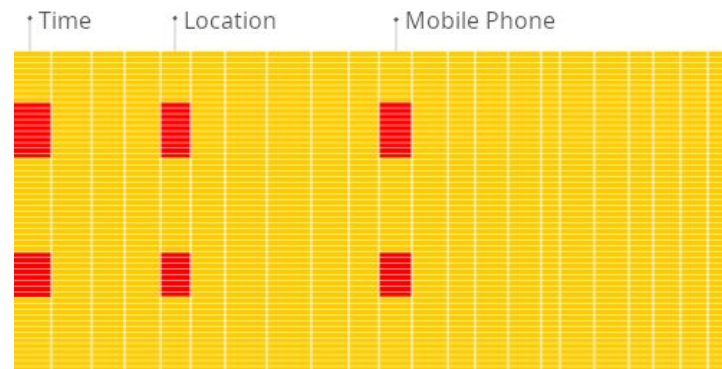
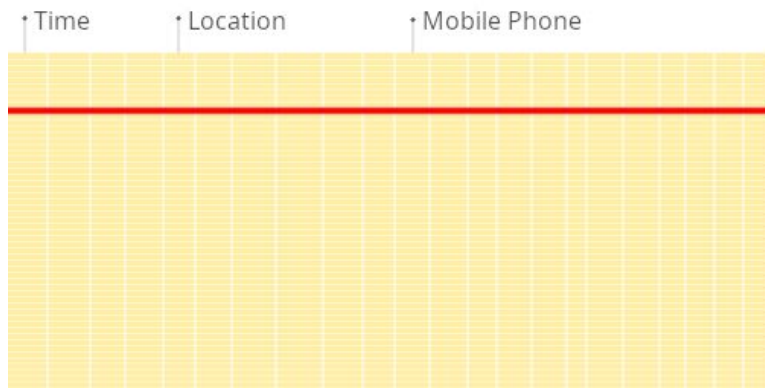
Tables

> locations	38M
> post_types	230
> posts	498M
> profiles	169M

کمی عمیق تر راجع به کلیک هاوس

ستون محور بودن کلیک هاوس

- نحوه ذخیره سازی
- فشرده سازی



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



OpenSource بودن و Community فعال

The screenshot displays the GitHub repository for ClickHouse. The main content area shows a list of files and directories with their commit history. The right sidebar provides an overview of the repository's activity and community metrics.

File/Directory	Description	Last Commit
.github	Bump to newer version of debug-action	yesterday
base	fix TSA support	last week
benchmark	Remove old file	6 months ago
cmake	Merge pull request #44828 from ClickHouse/remove-two-lines-of-C...	2 weeks ago
contrib	Merge branch 'master' into text_log_add_pattern	last week
docker	archiving -> achieving (typo)	2 days ago
docs	Merge pull request #44820 from stigsb/system_tables_volume_co...	34 minutes ago
packages	Remove adduser dependency	2 weeks ago
programs	Add <storage_policy> config parameter for system logs	yesterday

Repository Statistics:

- Watch: 693
- Fork: 5.4k
- Starred: 26.8k
- Contributors: 1,162
- Releases: 423
- Issues: 2.5k
- Pull requests: 286
- Discussions: 0
- Actions: 0
- Projects: 0
- Wiki: 0
- Security: 0
- Insights: 0

About: ClickHouse® is a free analytics DBMS for big data. clickhouse.com

Languages:

- C++ 83.8%
- Python 5.6%
- CMake 0.9%
- Other 0.8%
- Assembly 6.0%
- Shell 2.6%
- Go 0.3%



OpenSource بودن و Community فعال

A screenshot of the ClickHouse GitHub repository page. The page shows the repository name "ClickHouse / ClickHouse", a search bar, and navigation tabs for Code, Issues, Pull requests, Discussions, Actions, Projects, Wiki, Security, and Insights. The "Code" tab is selected, showing a list of files and folders with their commit history. The right sidebar displays the "About" section, including the repository description, website link, and a list of languages used in the project. The bottom right sidebar shows the "Contributors" section with a list of contributor avatars and the "Languages" section with a bar chart showing the distribution of languages used in the project.

Search or jump to... Pull requests Issues Codespaces Marketplace Explore

ClickHouse / ClickHouse Public Watch 693 Fork 5.4k Starred 26.8k

<> Code Issues 2.5k Pull requests 286 Discussions Actions Projects Wiki Security Insights

master 6,358 branches 1,599 tags Go to file Add file <> Code

antonio2368 Merge pull request #45320 from stigsb/system_tab... @ad37ad 34 minutes ago 105,779 commits

.github	Bump to newer version of debug-action	yesterday
base	fix TSA support	last week
benchmark	Remove old file	6 months ago
cmake	Merge pull request #44828 from ClickHouse/remove-two-lines-of-C...	2 weeks ago
contrib	Merge branch 'master' into text_log_add_pattern	last week
docker	archiving -> achieving (typo)	2 days ago
docs	Merge pull request #44820 from stigsb/system_tables_volume_co...	34 minutes ago
packages	Remove adduser dependency	2 weeks ago
programs	Add <storage_policy> config parameter for system logs	yesterday

About

ClickHouse® is a free analytics DBMS for big data

clickhouse.com

sql big-data analytics clickhouse dbms olap distributed-database mpp hacktoberfest

Readme Apache-2.0 license Code of conduct Security policy 26.8k stars 693 watching 5.4k forks

Release v22.3.17.13-lts Latest last week + 423 releases

Contributors 1,162 + 1,151 contributors

Languages

C++ 83.8%	Assembly 6.0%
Python 5.6%	Shell 2.6%
CMake 0.9%	Go 0.3%
Other 0.8%	

هزینه بسیار بالای عملیات های delete و update

- فرکانس بالای همچنین عملیات هایی باعث down شدن Clickhouse می شود
- نحوه داشتن جدولی که داده ها در آن اپدیت شوند، شدن نیست اما متفاوت است.
 - کوئری های همچنین جدولی نیز متفاوت است
 - برای مثال پست های اینستاگرام



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](#)

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 می باشد.





ممنون از همراهیتون!

با تشکر از [DataHobbies](https://github.com/anvaari/Clickhouse_VS_MariaDB) 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

