WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

BIG DATA ANALYSIS IN DIGITAL MARKETING RESEARCH

Industry: Research, Education | **Location:** Viena | **Solution:** Clickhouse, a column-store database

KEY HIGHLIGHTS

Challenges

- Conventional open source databases are not optimized for analytical workloads out of the box adoptions
- Find a workflow that would perform well after data size exceeds memory size

Requirements

- Need for affordable and scalable data analysis tools
- Need for privacy and data ownership

Solutions

- Clickhouse, a column-store database

Impact / Benefits

Clickhouse's performance and hardware efficiency is simple amazing, getting started is easy and the community is growing at a fast pace.

- supports plenty of relevant features for marketing research

CUSTOMER STORY

Vienna University of Business and Economics

Christian Hotz-Behofsits, Teaching & Research Associate at Vienna University of Business and Economics, is one of the creators of RClickhouse package for R. Christian Hotz-Behofsits shares his data analysis challenges his group is facing and how ClickHouse helps in their research.

Current Trends

In the last decades, the increasing availability of broadband internet and the accompanying digitalization has had a lasting effect on many industries. Nowadays, it is even common to purchase accounting systems, databases or the whole IT-infrastructure as a service. Those trends are challenges for both managers and marketing departments and thus current topics for digital marketing researchers. At the same time, data storage got cheaper and cheaper and firms started to gather every piece of data they could catch in hopes of gilding them one day.

> “Although data was already available it took a decade till managers had recognized the real value of it and now there is a real demand for data-driven decision support."
>
> *Christian Hotz-Behofsits*

Thus, big data is not an empty phrase anymore, data is available and its analysis is both feasible and reasonable. Nevertheless, mills are slowly working in research and it takes time for innovations in IT (i.e. big data analysis) to arrive at marketing research.

Our Systems

For some time now, I work for the Institute for Interactive Marketing and Social Media. As a digital marketing institute with a strong quantitative focus, we see it as our daily job to change this situation. Thus, we analyze immense datasets (up to 120TB per table) and find business and research-relevant insights by exploiting big data.

Nowadays, our systems are capable of assisting us in answering current marketing research questions. Especially, by aggregating huge data sets, processing geospatial information, merging diverse datasets and providing a powerful backend for visual real-time big data exploration. But this was not always the case. When I started, the typical software-stack was anything except big data ready. A single MySQL-Database was used as temporary web scraping data store. Afterwards, data was exported as CSV-file and some statistical programming language (e.g., R or Stata) was used to load it into memory for further analysis.

The Issues

This workflow works great for small data but hit its limit when data size exceeds memory size. Of course, some smart guys always find solutions (e.g., regressions that work only on subsets at once), but those treat the symptom, not the cause. Happily, times are changing and this is not only the case for our institute, but also marketing research changed. A few months back, high impact journals started to publish special big data issues (e.g., Special Issue of Marketing Science on Big Data), prediction models are not shabbily treated anymore and everyone wants to do machine learning.

Different Services

Of course, all those changes require a rethinking of the complete workflow. So, we started to look for scalable data analysis tools. On one side, there was the open source track led by the Hadoop ecosystem and related projects (e.g., Impala, Spark) and on the other side there are plenty of monthly rentable services (e.g., Google BigQuery, Amazon Athena). As a state university, money is a rare resource and has to be spent wisely. Furthermore, privacy and data ownership are important issues. These and many other points are the reason why we prefer self-hosted open-source software and try to avoid external service providers wherever it is feasible. But even open-source software is not perfect and it was not easy to find a suitable solution. Although Hadoop scales well, it does not optimally exploit the full power of an underlying hardware. It lives form the number of nodes and thus horizontal scaling. On a single node (as we have) a heap of different services and a non-trivial configuration is required to tune it. Because conventional open source databases (e.g., MySQL and Postgres) are not optimized for analytical workloads out of the box adoptions are provided in form of extensions (e.g., cstore_fdw) as well as new engines (e.g., MariaDB ColumnStore).

ClickHouse Solution

Over time, we tested diverse solutions like MariaDB ColumnStore (alpha), Postgres cstore_fdw, Apache Drill and MemSQL. All databases had their upsides, but not a single one was satisfying. By chance we finally came across Clickhouse, a column-store database by the Russian search engine provider Yandex that is not perfect, but on the best way there. Beside some minor flaws like the SQL-standard incompatible join-syntax, there are some weaknesses related to its novelty. As a young project, it is missing an established community and stable full-featured extensions for most programming languages. Nevertheless, we recognized and valued it’s potential. Clickhouse’s performance and hardware efficiency is simple amazing, getting started is easy and the community is growing at a fast pace. It also supports plenty of relevant features for marketing research like functions for geo-data analysis (e.g., PointInPolygon), data-sampling (i.e. MergeTree engine sampling), statistical functions (e.g., varSamp) and JSON support.

> “The open-source ecosystem lives from participation and that’s why we decided to provide and maintain a R-package (i.e., RClickhouse) for this database. Luckily, we are not the only supporters and because of specialized companies like Altinity, the future of this database looks bright and we want to be part of it."
>
> *Christian Hotz-Behofsits*

Written by Christian Hotz-Behofsits, Teaching & Research Associate at Vienna University of Business and Economics

SOLUTION COMPONENTS

ClickHouse streamlines all your data processing. It's easy to use: ingest all your structured data into the system, and it is instantly available for reports. New columns for new properties or dimensions can be easily added to the system at any time without slowing it down.