ClickHouse как альтернатива Google BigQuery

Рассказываем, почему Spotify и Uber выбрали российскую разработку для своих проектов

Что общего у Spotify, eBay, Uber, VK, Яндекс и Deutsche Bank?

Все эти компании используют в своей работе ClickHouse — колоночную систему управления базами данных от Яндекса с открытым исходным кодом. В этой статье мы подробно рассмотрим плюсы и минусы ClickHouse, сравним его с Google BigQuery, а также расскажем о миграции на российскую СУБД.

Что такое ClickHouse и чем он хорош?

ClickHouse — это высокопроизводительная колоночная аналитическая СУБД с открытым исходным кодом, разработанная Яндексом.

Предположим, что у вас несколько больших интернет-магазинов. Перед вами стоит задача управлять огромным объемом данных — параметрами визитов всех людей, которые заходили на ваш сайт и делали заказы. Как следствие, Яндекс Метрика и Google Analytics пробрасывают примерно 140 000 000 событий в вашу базу данных в месяц. И за весь период работы у вас накопились терабайты данных.

Вообще, вы можете строить отчеты на любой системе (как платной, так и бесплатной) — вот только когда объем достигнет миллиарда строк, картина в любом случае перестанет быть радужной. Аналитические запросы в БД будут исполняться десятками минут.

Для того чтобы ускорить такие процессы, были созданы особые СУБД — колоночные. Разумеется, в мире существуют и другие разновидности систем управления базами данных: реляционные, объектные, документные, сетевые, функциональные, их различные сочетания. Но именно колоночные обладают самой высокой скоростью работы с большими объемами информации (если сравнивать со строковыми).

ClickHouse — колоночная СУБД, которая также является реляционной. Она была создана Яндексом для того, чтобы анализировать и читать данные по множеству значений максимально быстро: во время разработки код проекта постоянно оптимизировался с приоритетом на производительность.
После релиза многие действительно были впечатлены возможностями масштабирования и скоростью работы ClickHouse. При использовании других популярных СУБД, запрос легко мог выполняться по несколько минут и пользователи уходили пить кофе. В ClickHouse же все происходило за секунды — решение от Яндекса “проглатывало” тысячи строк в секунду и петабайты данных с дисков.

Вот что может ClickHouse в цифрах:
https://habrastorage.org/r/w1560/files/304/ba4/7b2/304ba47b201b4827a3d0a6e76445dab4.png

Давайте разберемся, почему все это стало возможно. И сравним ClickHouse с одним из его главных конкурентов — Google BigQuery.

ClickHouse vs Google BigQuery

1) Google BigQuery

Это облачное PaaS решение, созданное Google еще в 2011 году. Оно позволяет обрабатывать и хранить большие массивы данных без настройки выделенного сервера. В Google BigQuery реализованы большинство функций реляционной СУБД. Например, пользователи могут загружать большой объем табличных данных и обращаться к ним SQL-запросами. Разумеется, можно сохранять и выгружать результаты таких запросов.

А еще это составная часть Google Cloud Platform, в рамках которой вам становятся доступны несколько десятков различных дополнительных инструментов по анализу, работе с данными, визуализации и т.д.

Плюсы Google BigQuery

- Высокая скорость. Google BigQuery умеет обрабатывать 100 миллиардов строк за считанные секнуды: https://cloud.google.com/blog/products/bigquery/anatomy-of-a-bigquery-query
- Работа с SQL. Помимо стандартного, поддерживается диалект Legacy SQL.
- Продвинутый функционал по изменению и обработке данных.
- Экономическая доступность при небольших объемах данных.
- Гибкие настройки доступа. Можно за несколько кликов предоставить ограниченный доступ к базе данных.
- Отличные возможности для различных интеграций (главное, чтобы языки поддерживали REST API). Важно отдельно отметить интеграцию с Google Analytics.
- Простота для пользователя.

Минусы Google BigQuery

Идея этой статьи появилась у нас еще до возникновения угрозы ухода некоторых сервисов Google с российского рынка. Мы не знаем, произойдет это или нет. Но риск законодательных ограничений на передачу данных в зарубежное облако вполне очевиден.

К минусам Google BigQuery также можно отнести высокую стоимость при работе с большими объемами информации. Периодически пользователи получают счета на

несколько сотен долларов всего за несколько простых на первый взгляд запросов. При этом, тарифы в любой момент могут измениться и вырасти.

По карману могут ударить стриминговые вставки, обновления данных и ряд других функций.

2) ClickHouse

Благодаря своим особенностям, ClickHouse реактивно отвечает на любые аналитические запросы в реальном времени. По этому параметру он обгоняет Google BigQuery и большинство других конкурентов. Вот слова Виктора Тарнавского, одного из разработчиков ClickHouse.

“В ClickHouse сейчас нет медленно работающих функций — все работают настолько быстро, насколько это возможно, в пределах разумного. Также используется векторная обработка данных — это означает что данные никогда не обрабатываются по строчкам, обрабатываются только колоночками".

ClickHouse способен масштабироваться до десятков триллионов записей и петабайтов данных. Благодаря скорости эту СУБД можно легко использовать в интерактивных приложениях. Вы можете как построить систему с собственными серверами, так и воспользоваться облаком от Яндекса.

Но давайте рассмотрим плюсы и минусы подробнее:

Плюсы ClickHouse

- Максимальная производительность. Как мы уже писали выше, ClickHouse работает быстро по сравнению с другими СУБД. По одному из тестов Яндекса ClickHouse обладает наиболее высокой пропускной способностью на длинных запросах и наиболее низкой задержкой на коротких запросах среди доступных для тестирования систем подобного класса: https://clickhouse.com/docs/ru/introduction/performance/
- Сжатие данных. По своей сути ClickHouse — это поколоночная система. Для нас это важно тем, что сжатие данных в нем работает очень хорошо, а значит, увеличивается скорость вычислений и снижается место для хранения.
- Соответствие законам РФ. Сервис полностью сертифицирован для работы на территории нашей страны. В частности, он соответствует всем критериям Федерального закона «О персональных данных» №152 и с большой долей вероятности будет доступен российским пользователям и дальше (конечно, если вы работаете не в России, это не так актуально).
- Отказоустойчивость. Разработчики специально закладывали в ClickHouse возможность выдерживать падение датацентра. По умолчанию работает асинхронная репликация, которая повышает уровень отказоустойчивости.
- Масштабируемость. Из коробки ClickHouse умеет линейно масштабироваться для постройки баз очень большого размера.

- В отличие от Google BigQuery, ClickHouse может хранить данные не только в облаке.

Минусы ClickHouse

- В ClickHouse нет полноценных транзакций - это существенный недостаток для некоторых бизнесов и сфер применения.
- К сожалению, из коробки вы не сможете изменять или удалять уже записанные данные с низкими задержками и высокой частотой запросов. ClickHouse может предложить лишь массовое удаление и редактирование для очистки ненужного или соответствия GDPR.
- Из-за разреженного индекса ClickHouse плохо умеет выполнять точечные чтения одиночных строк по своим ключам.
- Можно считать недостатком то, что в ClickHouse применяется диалект SQL, а не его стандартный формат (хотя декларативный язык запросов ClickHouse во многом совпадает с SQL стандартом).

Результаты сравнения

Разумеется, Google BigQuery — это проверенная, удобная и, функциональная облачная система. Она поддерживает большинство ключевых возможностей современных СУБД. Ее неоспоримый плюс - легкая интеграция со сторонними платформами. В целом, ее можно назвать более удобной и простой для новичка.

Однако ClickHouse — это хороший выбор для масштабных в плане вычисления запросов. Чем больше запросов SELECT вы делаете на постоянной основе - тем больше смысла в миграции с BigQuery на ClickHouse. Скорее всего такая замена вполне позволит вам сэкономить. Помимо экономических аспектов, ClickHouse есть смысл рассматривать если вам необходима более высокая производительность и большая скорость масштабирования.

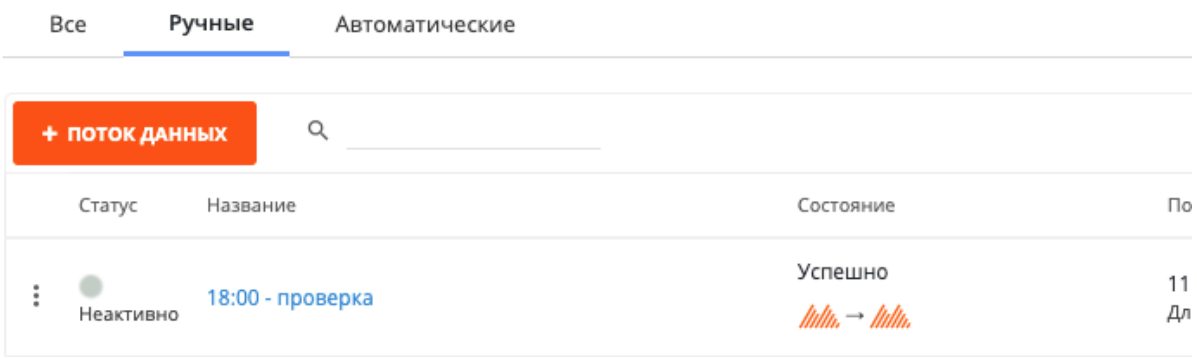
Другой неоспоримый плюс систем на ClickHouse — максимально полное соответствие российскому законодательству. К сожалению, прогнозировать какие последствия могут ждать российских пользователей при использовании Google BigQuery в ближайшее время достаточно сложно. На момент написания статьи уже отключена реклама Google Ads и монетизация YouTube. Мы оптимисты, поэтому искренне верим, что все будет хорошо:) Но, к сожалению, никаких гарантий этого нет.

Я на Google BigQuery, но хочу на ClickHouse. Что делать?

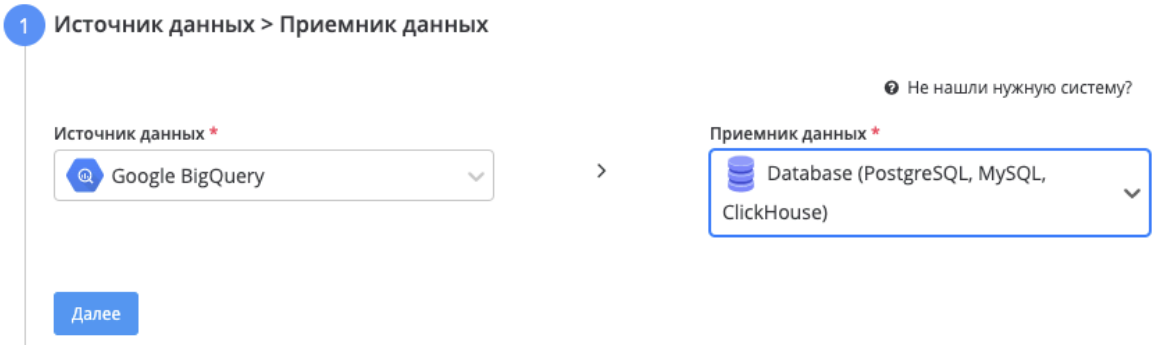
К счастью, в Garpun Feeds есть удобный коннектор, который поможет в несколько простых шагов перейти с одной СУБД на другую. В частности, он позволяет перенести базы с Google BigQuery на ClickHouse (такие кейсы использования нашего сервиса уже есть). Одним из важных его преимуществ является то, что в большинстве случаев перенос можно осуществить без привлечения программистов за несколько кликов.

Итак, если вы хотите перейти на ClickHouse, нужно сделать следующее:

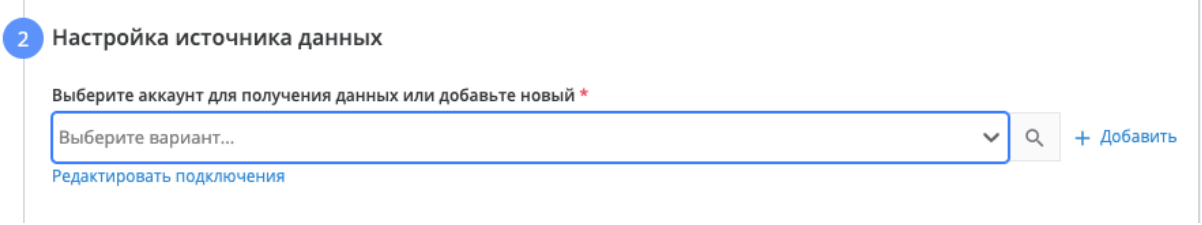
1. Зарегистрироваться в системе Garpun (ссылка)
2. Перейти в сервис Garpun Feeds
3. В разделе “потоки данных” нажать на кнопку “+ поток данных”



4. В качестве источника данных выберите Google BigQuery, в качестве приемника указать Database. Подключение к ClickHouse, MySQL и к PostgreSQL выполняется через один и тот же протокол, так что переживать не стоит.



5. Далее, выберите подключенный аккаунт Google BQ или подключите новый с помощью кнопки “Добавить” справа от столбца выбора аккаунта.



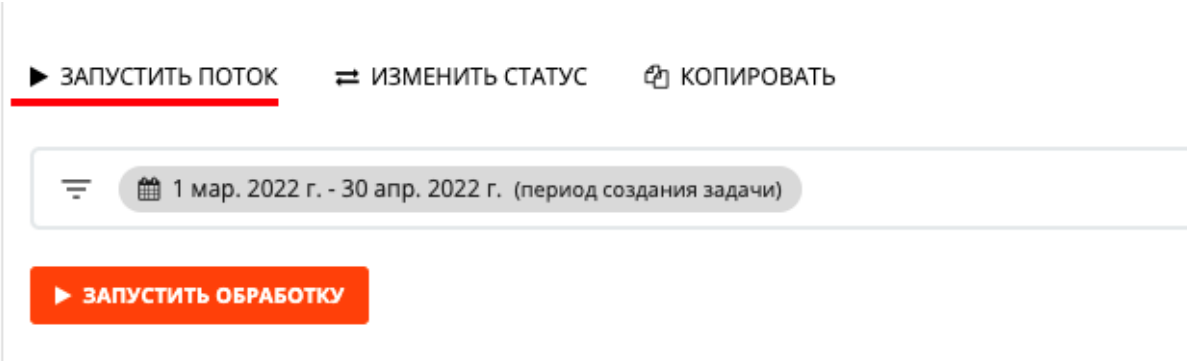
6. После успешного подключения выберите ID проекта в вашем BQ и в поле “Standard SQL Query” напишите какие данные вам нужно загрузить из вашей таблички в новое хранилище. Например, SELECT * FROM “название дата-сета в BQ”, чтобы выгрузить все данные.

7. На следующем шаге, в настройках приемника данных, выберите или подключите вашу базу данных и укажите параметры новой схемы базы данных, выберите способ записи и настройки партицирования, если требуется.

8. Далее, укажите название будущего потока и настройте период автоматического обновления данных. Нажмите “Готово”. После этого поток будет создан.

9. По умолчанию, поток будет запускаться автоматически в указанное время и обновлять данные за нужный период.

Если вам требуется собрать данные за больший период, вы можете сделать это, нажав на кнопку “Запустить поток” и выбрав нужный диапазон дат.



Выводы

В большинстве случаев ClickHouse вполне можно рассматривать как достойную альтернативу Google BigQuery. Он очень быстрый и надежный. Это особенно актуально, если у вас есть большой объем данных. Неслучайно кроме Яндекса ClickHouse используют огромное количество компаний с мировым именем - от AdTech гигантов до крипто-проектов и банков. Например, ClickHouse использует Spotify, eBay, Uber, VK, Яндекс и даже Deutsche Bank.

- Если вы уже используете Google BigQuery, то просто знайте, что вы в любой момент можете поменять эту СУБД на ClickHouse. В случае, если у вас большая база данных, скорее всего это сэкономит вам деньги и время.

- Если вы уже используете Google BigQuery и у вас небольшая база данных, однозначного смысла мигрировать на ClickHouse нет. В случае кризисных обстоятельств, можно будет сменить СУБД.

- Если вы прямо сейчас находитесь в поисках хорошей СУБД - мы настоятельно рекомендуем рассмотреть ClickHouse. А если вы в принципе не уверены, что

именно хотите от системы управления баз данных и какую выбрать - смело пишите нашим специалистам. Мы поможем принять оптимальное решение;)