Ссылка на статью:
// Ученые записки УлГУ. Серия Математика и информационные технологии. 2024, № 1, с. 17-21.

Поступила: 15.12.2023
Окончательный вариант: 03.06.2024

УДК 004.65

Реализация сервиса сбора статистических данных в ClickHouse

Дергунов А.В.

adergunov@webpower.ru
Ульяновский государственный университет, Россия

В работе описывается сервис, разработанный с использованием языка программирования GO и столбцовой OLAP СУБД ClickHouse. Сервис поддерживает интегрирование с любыми веб приложениями, а также с программными комплексами, поддерживающими передачу данных по HTTP протоколу.

Ключевые слова: *ClickHouse, программный комплекс, сервис статистики, базы данных*

Введение

В современном мире можно сказать, что данные – это новая нефть. Данные помогают систематизировать информацию, а аналитика на ее основе — выстроить оптимальную стратегию продвижения бизнеса, а также большие данные нужны для обучения моделей искусственного интеллекта, который в будущем будет обладать невообразимыми возможностями.

Первичные данные сами по себе являются достаточно бесполезными. Данные приобретают ценность в результате переработки — анализа, сравнения, сортировки, фильтрации и применения. Для таких операций очень важно иметь быстрый процессор, много памяти и хорошую систему управления базами данных, способную моментально выполнять запросы на структурированных больших данных. Тут на помощь приходит ClickHouse [1]. Clickhouse — это колоночно-ориентированная система управления базами данных (СУБД) с открытым исходным кодом, используемая для онлайн-аналитической обработки (OLAP), созданная Яндексом. В настоящее время он поддерживает вторую по величине платформу веб-аналитики — Яндекс Метрику [2].

В сегодняшних технологически продвинутых программных комплексах возникает вопрос надежности работы, когда все функции выполняются в пределах одного

приложения. Такой подход называется монолитной архитектурой. При отказе части функционала, есть риск отключения всей системы. Для решения данной проблемы используется архитектура разработки продукта, основанная на микросервисах [5].

В данной работе рассматривается проблема оптимизации сбора статистических данных в приложении для аренды жилья, в котором используется MySQL. Цель работы - создать инструмент, обеспечивающий отказоустойчивость системы и повышение скорости работы. В качестве примера в статье описывается реализация сервиса сбора статистики, обрабатывающий и сохраняющий данные в базу, независимо от состояния других частей программного комплекса.

1. Архитектура и инструменты

На рис. 1 схематически изображена архитектура приложения.

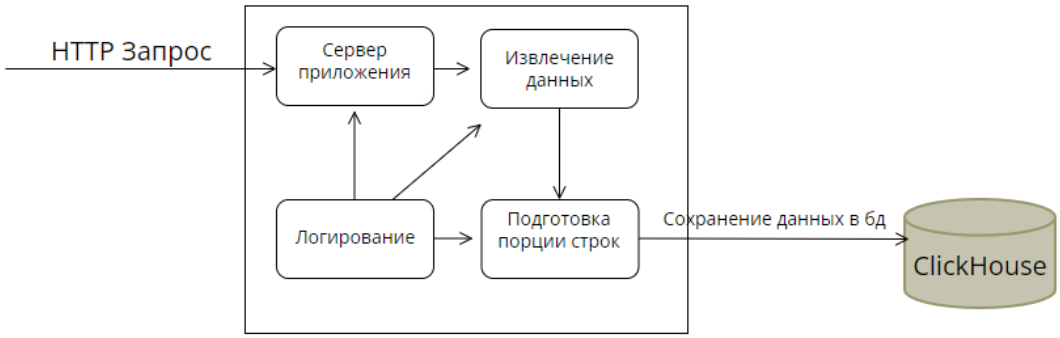

Рис. 1. Архитектура приложения

Как видно на рис. 1, система работает в 3 шага: получение запроса с данными, подготовка к записи в приложении и запись. Такой подход позволяет отделить часть функционала системы в отдельное приложение, которое будет работать параллельно на отдельном сервере. Клиент быстрее получит результат выполнения программы, так как не нужно ждать завершения обработки и сохранения статистических данных. В свою очередь работа приложения также происходит в несколько этапов: обработка сервером запроса, извлечение данных с параметров запроса и накапливание в памяти приложения строк для, так называемой пакетной записи в базу данных. Пакетная вставка (метод добавления большого количества записей одновременно) используется для оптимизации производительности массовой загрузки данных. Преимущества пакетной вставки: увеличение производительности, экономия ресурсов, уменьшение нагрузки на базу данных, что очень важно при сборе статистики, так как в больших программных системах в пике нагрузка может достигать до миллиона запросов в секунду и больше. Информация выполнения каждого этапа будет записываться в лог файл, чтобы можно было отследить ошибки в случае неисправности сервиса [3]-[4].

Для разработки сервиса сбора статистики использовался язык программирования Go. Это компилируемый многопоточный язык с открытым исходным кодом. В основном его применяют в веб-сервисах и клиент-серверных приложениях. В конце 2021 года

Golang вошёл в топ-5 востребованных языков. Авторы языка попытались объединить лёгкость разработки на Python и скорость исполнения программ на C и C++, поэтому сделали Go компилируемым.

В качестве базы данных для хранения статистических данных используется ClickHouse, по причине наличия таких преимуществ как:

- наличие линейной масштабируемости, пета байтных кластеров и возможность работы в разных дата центрах;
- сжатие данных происходит очень эффективно благодаря колоночной архитектуре;
- в ClickHouse используется SQL, но с дополнительными функциями для различных типов данных;
- внешние словари упрощают работу с данными и позволяют использовать разные источники данных;
- скорость работы ClickHouse значительно выше, чем у других баз данных, что меняет подход к работе с данными;
- интерфейсы для работы с ClickHouse включают консольный клиент, HTTP и JDBC драйвер, а также коннекторов для разных языков программирования.

Для тестирования будет использоваться программа Insomnia. Это сервис для создания, тестирования, документирования, публикации и обслуживания API. Он позволяет создавать коллекции запросов к любому API, применять к ним разные окружения, настраивать мок-серверы, писать автоматические тесты, анализировать и визуализировать результаты запросов.

2. Тестирование сервиса

На рис. 2 представлена структура таблицы, где первичным ключом является колонка “ts” – расшифровывается как timestamp, то есть временная метка.

Название	#	Тип Данных	Длина
123 ts	1	Int64	20
ABC ip	2	String	
ABC age	3	String	
ABC browser	4	String	
ABC country	5	String	
ABC city	6	String	
ABC gender	7	String	
ABC phone	8	String	
ABC lang	9	String	
ABC balance	10	String	
ABC p_id	11	String	
ABC p_name	12	String	
ABC p_cost	13	String	
ABC p_rating	14	String	
ABC reviews_cnt	15	String	
123 credit	16	UInt8	3
123 delivery	17	UInt8	3
123 is_auth	18	UInt8	3
ABC url	19	String	

Рис. 2. Список параметров

После запуска сервера ClickHouse и приложения прописываются тестовые данные параметров запроса в Insomnia.

После успешного сохранения данных, сервер возвращает код 200. Также можно подряд отправлять несколько запросов, данные будут сохранятся пачками каждые 10 секунд. На рис. 3 видно, что указаны не все параметры, представленные на рис. 2. Пропущенные значения записываются в базе данных как NULL, при необходимости их можно отфильтровать в выборках или вовсе удалить.

Так как тестирование проводилось на локальном компьютере, поля country и city не определяются по ip и остаются пустыми.

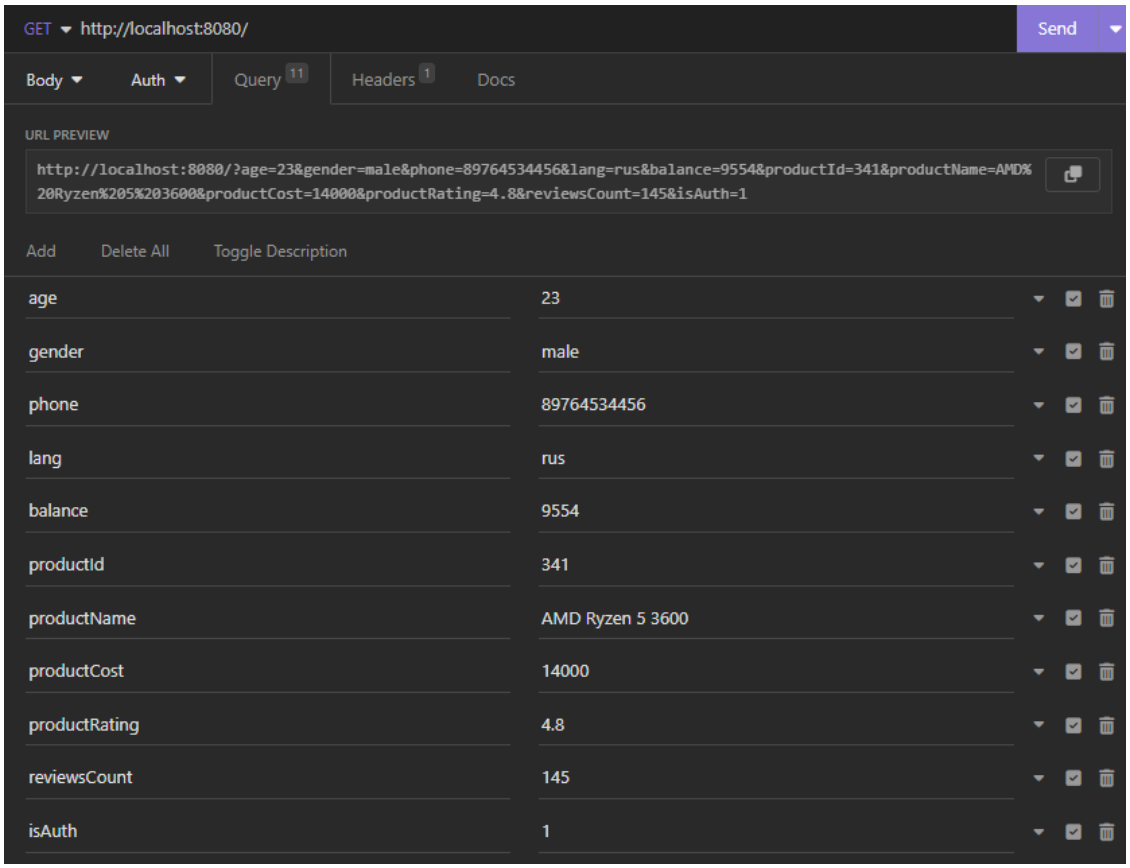

Рис. 3. Запрос к сервису сбора статистики

ts	ip	bro	count	city	gender	phone	lang	balance	p_id
1 698 232 479	127.0.0.1:53896	23	insomnia		male	89764534456	rus	9554	341
1 698 232 842	[::1]:54057	23	Chrome		male	89764534456	rus	9554	341
1 698 232 842	[::1]:54057	23	Chrome		male	89764534456	rus	9554	341
1 698 232 889	[::1]:54057	23	Chrome		male	89764534456	rus	9554	341

p_name	p_cost	p_rating	rev				url
AMD Ryzen 5 3600	14000	4.8	145	0	0	1	/?age=23&gender=male&phone=89764534456&lang=rus&b
AMD Ryzen 5 3600	14000	4.8	145	0	0	1	localhost:8080
AMD Ryzen 5 3600	14000	4.8	145	0	0	1	localhost:8080
AMD Ryzen 5 3600	14000	4.8	145	0	0	1	localhost:8080

Рис. 4. Обзор сохраненных данных в ClickHouse

Для сравнения ClickHouse с MySQL были рассматрены две таблицы с одинаковыми полями, одна – Mysql (650 0000 записей), другая – ClickHouse(~4 500 000 записей). Выполнение ряда запросов с многочисленными условиями и агрегирующими функциями выявило, что ClickHouse работает в среднем в пять раз быстрее, чем MySQL, хотя записей примерно в пять раз больше.

Заключение

В статье описана реализация и тестирование сервиса сбора статистики для улучшения работы веб-приложения по аренде жилья с использованием актуальных технологий, а также сравнение обработки данных двух разных СУБД. Компилируемый язык программирования Go позволяет моментально обрабатывать запросы, а выделение логики работы части системы позволяет не беспокоится, что сохранение данных прекратится при сбое основной серверной программы, так как в сервис можно отправлять данные напрямую с клиента. Также можно отметить, что основная затрата времени выполнения полной процедуры записи (около 5 миллисекунд) уходит на сетевые издержки.

Список литературы

1. Миловидов А. *Clickhouse для DBA.* Яндекс, 2021.
2. Стасышина Т. Л., Стасышин В. М. *Базы данных. Лекции по курсу.* НГТУ, 2021.
3. Willy Richert. *Designing Evolvable Microservices*. O'Reilly, 2017.
4. Mark Powell. Go Programming Language. O'Reilly Media, 2016.
5. Foreman John W. *Data smart*. John Wiley & Sons Limited, 2017.

Implementation of the statistical data collection service in ClickHouse

Dergunov, A.V.

adergunov@webpower.ru

Ulyanovsk State University, Russia

The paper describes a service developed using the GO programming language and the ClickHouse columnar OLAP DBMS. The service supports integration with any web applications, as well as with software packages that support data transmission over the HTTP protocol.

Keywords: *ClickHouse, software package, statistics service, databases*