УДК 004.624

СИСТЕМА ПОТОКОВОЙ РЕПЛИКАЦИИ ДАННЫХ ДЛЯ РЕЛЯЦИОННОЙ СУБД

Карапетян Э.А. (ИТМО)
Научный руководитель – Клименков С.В. (ИТМО)

Введение. Существуют различные типы хранилищ данных, которые используются в различных системах. У каждого из видов хранилищ есть свои особенности, однако часто возникает ситуация, когда одни и те же данные нужны в нескольких хранилищах сразу. Если в одних и тех же системах хранения данных, таких как реляционная СУБД, обычно есть возможность репликации между серверами, то между различными хранилищами такой возможности нет. По этим причинам разработчикам зачастую приходится заниматься агрегацией данных в программном коде. Также достаточно большой проблемой является переход с одного хранилища на другое без остановки системы, ведь в современном мире при возникновении подобной задачи необходима достаточная подготовка и зачастую остановка системы, ведь без облачной системы подобные манипуляции данными весьма тяжелы.

Изучение данной проблемы весьма актуально сейчас, так как подобная система может решить сразу несколько проблем – репликация данных для использования их сразу несколькими видами хранилищ, переход на другое хранилище, организации доступности данных в программном коде в реальном времени для обработки или анализа.

Основная часть. Создано прикладное решение, которое представляет из себя систему на базе подхода CDC, с использованием его представителя Debezium, который предоставляет коннекторы к различным хранилищам данных. Данная система предоставляет возможность организации репликации данных реляционной СУБД в иное хранилище данных с возможностью обработки потока данных в программном коде.

Для реализации потока данных используется Apache Kafka, которая с помощью коннекторов Debezium может получать данные и отправлять их в нужное хранилище или Kafka Listener. Чтобы добиться большей производительности системы необходимо было оптимизировать потоки данных и нагрузку на систему. Для этого используется бинарная сериализация AVRO, которая уменьшает объем передаваемых данных, а также передача пакетов данных (batch) в Kafka, которая передает несколько обновлений данных, а не по одной записи.

Для запуска системы используется разработанная программа, которая позволяет сконфигурировать систему для указания источника и стока данных, а также иные конфигурационные данные, такие как название базы данных, схемы, таблицы и прочие параметры.

Выводы. В результате данного исследования получился продукт, который может упростить процессы с манипуляцией данными, такие как репликация, синхронизация хранилищ данных, интеграция систем, а также предоставление доступа к данным для их аналитики в реальном времени. Этот продукт значительно упростит работу с данными и позволит быстрее и эффективнее решать задачи манипуляции данными.

Список использованных источников:

1. Шаталова Юлия Георгиевна, Жиглов Ярослав Владимирович Разработка системы репликации для распределенной базы данных предприятия // Символ науки. 2017. №3. URL: https://cyberleninka.ru/article/n/razrabotka-sistemy-replikatsii-dlya-raspredelennoy-bazy-dannyh-predpriyatiya (дата обращения: 13.02.2024).
2. Воробьев Сергей Петрович, Горобец Виталий Владимирович Исследование модели транзакционной системы с репликацией фрагментов базы данных, построенной по принципам облачной среды // ИВД. 2012. №4-1. URL: https://cyberleninka.ru/article/n/issledovanie-modeli-

tranzaktsionnoy-sistemy-s-replikatsiey-fragmentov-bazy-dannyh-postroennoy-po-printsipam-oblachnoy-sredy (дата обращения: 13.02.2024).

3. Реализация механизма репликации в СУБД Postgre SQL / Р. Ф. Гибадуллин, А. М. Зиннатов, М. Ю. Перухин, Р. Н. Гайнуллин // Вестник Технологического университета. – 2017. – Т. 20, № 24. – С. 100-101. – EDN TAMWXD.

4. Кирносов, В. Ю. Сравнительный анализ механизмов репликаций данных в различных СУБД / В. Ю. Кирносов, Н. М. Куржангулов // Фундаментальные и прикладные исследования в современном мире. – 2017. – № 18-1. – С. 84-91. – EDN YROUGX.