Если вы видите что-то необычное, просто сообщите мне.

Подключение Kafka к PostgreSQL

Инструкция поможет вам взять на себя ответственность без проблем и без потери эффектисновсти. Цель статьи в создании процесса экспорта данных настолько гладко, насколько это возможно.

В конце статьи вы сможете успешно подключать Kafka к PostgreSQL, плавно передавать данные потребителю по выбору, для полноценного анализа в реальном времени. В дальнейшем это повзолит пострить гибкий ETL(дословно «извлечение, преобразование, загрузка») конвеер для вашей организации. Из стати вы узнаете более глубокое понимание инструментов и техник и таким образом оно поможет вам отточить ваши умения дальше.

Требования

Для лучшего пониманиния статьи, требуется понимание следующего списка тем:

- Знания PostgreSQL.
- Знания Kafka
- Kafka и PostgreSQL установленны на хосте.

Введение в Kafka

Apache Kafka это продукт с открытым исходным кодом, который помогает публиковать и подписываться на большие по объему сообщения в распределенной системе. Kafka использует идею лидер-последователь, позволяя пользователя копировать сообщения в

независимые от падения, и в дальнейшем позволядет делить и хранить сообщения в Kafka топиках в зависимости от темы сообщения. Kafka позволяет настраивать в реальном времеи потоки данных и приложения для изменения данных и потоков от источника к цели.

Ключевые особенности Kafka:

- Масштабируемость: Kafka имеет исключительную масштабируемость и может быть отмасштабированно без времени простоя.
- Изменение данных: Kafka предлагает KStream и KSQL(в случае Confluent Kafka) для изменению данных на лету.
- Отказоустойчивость: Kafka использует посредников для копирования данных и постоянства данных, для создания отказоустойчивых систем.
- Безопасность: Kafka может быть объеденина с различными метриками безопасности такими как Kerberos, для передачи информации конфиденциально.
- Производительность: Kafka распределенна, разделена и имеет очень высокую пропускную способность для публикации и подписки на сообщения.

Для более подробного описания, можно обратиться на официальный сайт разработчиков Kafka

Введение в PostgreSQL.

PostgreSQL это мощное, производственного класса, с открытым исходным кодом СУБД которая использует стандартные SQL запросы связанных данных и JSON для запросов несвязанных данных хранящихся в базе данных. PostgreSQL имеет отличную поддержку для всех операционных систем. Он поддерживает расширенные типы данных и оптимизацию операций, которые можно найти в комерческих проектах каа Oracle, SQL Server и т.д.

Ключевые особенности PostgreSQL:

- Имеет расширенную поддержку для сложных запросов.
- Предоставляет отличную поддержку для географических объектов и следовательно он может быть использован для географической информационной системы и сервисе на основе положения.
- Предоставляет поддержку для клиент-серверной сетевой технологии

- Упреждающая журнализация(write-ahead-logging (WAL)) позвляет быть базе данных отказоустойчивой.

Для большей информации по PostgreSQL, можно посмотреть официальный вебсайт.

Процесс ручной настройки Kafka и PostgreSLQ интеграции

Kafka поддерживает подключение с PostgreSQL и различными другими базами данных с помощью различных встроенных подключений. Эти коннекторы помогают передавать данные от источника в Kafka и затем передать потоком в целевой сервис с помощью выбора топиков Kafka. Так же, есть множество подключений для PostgreSQL, которые помогают установить подключение к Kafka.

1. Установка Kafka

Чтобы подключить Kafka к PostgreSQL, для начала нужно скачать и установить Kafka.

2. Старт Kafka, PostgreSQL и Debezium сервер

Confluent предоставляется пользователям с различным набором встроенных подключений которые действуют как источники и сток данных, и помогает пользователям передавать их данные через Kafka. Один из таких подключений/образов которые позволяют подключать Kafka к PostgreSQL - Debezium PostgreSQL Docker образ.

Чтобы установить Debezium Docker который поддерживает. подключение к PostgreSQL с Kafka, обратимся к официальному проекту Debezium Docker и склониурем проект на нашу локальную систему.

Как только вы склонировали проект вам нужно запустить Zookeper сервис который хранит настройки Kafka, настройки топиков, и упревление нодами Kafka. Это всё запускается следующей командой:

docker run -it --rm --name zookeeper -p 2181:2181 -p 2888:2888 -p 3888:3888 debezium/zookeeper:0.10

Теперь с работающим Zookeper, вам нужно запустить Kafka сервер. Чтобы сделать это откройте консоль и выполните следующую команду:

docker run -it --rm --name kafka -p 9092:9092 --link zookeeper:zookeeper debezium/kafka:0.10

Как только вы запустили Kafka и Zookeeper, теперь запускаем PostgreSQL сервер, его мы будем подключать к Kafka. Это можно выполнить следующей командой:

docker run — name postgres -p 5000:5432 debezium/postgres

Теперь стартуем Debezium. Для этого выполним следующую команду:

docker run -it — name connect -p 8083:8083 -e GROUP_ID=1 -e CONFIG_STORAGE_TOPIC=my-connect-configs -e OFFSET_STORAGE_TOPIC=my-connect-offsets -e ADVERTISED_HOST_NAME=$(echo $DOCKER_HOST | cut -f3 -d’/’ | cut -f1 -d’:’) — link zookeeper:zookeeper — link postgres:postgres — link kafka:kafka debezium/connect

Как только вы запустили все эти сервера, логинимся в командную оболочку PostgreSQL используя следующие команды

psql -h localhost -p 5000 -U postgres

3. Создаем базу данных в PostgreSQL

Как только вы вошли в PostgreSQL, вам необходимо создать базуданных. Для примера если вы хотите создать базуданных с именем `emp`, вы можете использовать следующую команду:

CREATE DATABASE emp;

В готовой базе, создадим таблицу, которая будет хранть информацию. Для этого выполним:

CREATE TABLE employee(emp_id int, emp_name VARCHAR);

Теперь нужно добавить данные или несколько записей в таблицу. Для этого выполните выполните команды как указано ниже:

INSERT INTO employee(emp_id, emp_name) VALUES(1, 'Richard') INSERT INTO employee(emp_id, emp_name) VALUES(2, 'Alex') INSERT INTO employee(emp_id, emp_name) VALUES(3, 'Sam')

Таким образом вы можете создать postgreSQL базу данных и вставить в неё значение, для того чтобы настроить подключение между Kafka и PostgreSQL.

4. Поднятие подключения Kafka-PostgreSQL

Как только вы настроили PostgreSQL базу данных, вам нужно поднять Kafka-Postgres подключение, которое позволить вам тянуть данные из PostgreSQL в Kafka топик. Для этого вы можете создать Kafka подключение используя следующий скрипт:

curl -X POST -H “Accept:application/json” -H “Content-Type:application/json” localhost:8083/connectors/ -d ‘ { “name”: “emp-connector”, “config”: { “connector.class”: “io.debezium.connector.postgresql.PostgresConnector”, “tasks.max”: “1”, “database.hostname”: “postgres”, “database.port”: “5432”, “database.user”: “postgres”, “database.password”: “postgres”, “database.dbname” : “emp”, “database.server.name”: “dbserver1”, “database.whitelist”: “emp”, “database.history.kafka.bootstrap.servers”: “kafka:9092”, “database.history.kafka.topic”: “schema-changes.emp” } }’

Чтобы проверить что подключение прошло успешно воспользуйтесь командой:

curl -X GET -H “Accept:application/json” localhost:8083/connectors/emp-connector

Для того, чтобы проверить что Kafka получил данные из PostgreSQL или нет, нужно ключить Kafka Console Consumer, используя следующую команду:

docker run -it -name watcher -rm - link zookeeper:zookeeper debezium/kafka watch-topic -a -k dbserver1.emp.employee

Команда выше теперь отобразит вашу базу данных PostgreSQL в консоли. После того как убедимся что данные полученны в Kafka верно, можно воспользоваться KSQL/KStream или Spark поток для произвдения действий ETL над данными.

Revision #1

Created 2022-10-13 14:12:49 UTC by gasick

Updated 2022-10-13 14:13:03 UTC by gasick