OntarioTech UNIVERSITY

SOFE4630 Cloud Computing (Winter 2022 - Dr. M. El-darieby)

Lab 3: Project Milestone-- Data Storage Implementation: KV + relational

Date: 03/18/2022

Student Names:

Alexander Pitcher (100693749)

Mitul Patel (100700131)

Clarissa Branje (100716458)

Amin Khakpour(100669547)

Lab 3: Kafka connects for Redis and MySQL

Objective:

- **Deploy Tabular and key-Value data storage to GKE.**
- **Get familiar with Key-Value data storage**
- **Get familiar with Kafka Connectors and their configuration.**
- **Configure and use Kafka source connector to Redis.**
- **Configure and user MySQL sink and source Kafka connectors.**

- **Sink and Source connectors?**

A source connector is responsible for collecting data from a system. This could be a database, stream tables, and even message brokers. Furthermore, source connector is able to collect metrics from application servers into kafka topics, which allows the data available for stream processing with low latency.

A sink connector provides data from kafka topics to other systems which could be Elasticsearch, batch systems like Hadoop or any database.

- Source Connectors provide data to kafka, whereas sink connectors take

  data out of kafka.

- **The applications/advantages of using Kafka Connectors with data storage?**

There are several benefits of using kafka connectors with data storage. Several advantages include using the following:

- Data Centric Pipeline
- Flexibility and Scalability
- Reusability and Extensibility

The pipeline allows for connecting meaningful data abstractions to pull and push data to kafka. Connection runs stream and batch systems which can scale to organization-wide service. Connections extend them to tailor the user's needs along with using existing connectors.

The application of kafka connectors is that they can be used to

import/export data from outside systems. The main advantage of using

Kafka connectors are flexible and modular.

- **How do Kafka connectors maintain availability?**

Apache Kafka connector maintains availability from its built-in support for parallelism and scalable data copying. Furthermore kafka connectors also have task rebalancing and fault tolerance for kafka connect. If a worker fails unexpectedly, every other worker will detect it automatically and redistribute connectors and tasks across available workers.

- **List the popular Kafka converters for values and the properties/advantages of each.**

There are several types of kafka converters. A few examples are the following:

- Avro: Sterilizes record keys and values, it is very compact and efficient.
- Protobuf: Ensures signals dont get lost between other apps, it processes
- information very quickly
- String: able to define conversion between strings and objects and control
- their behavior.
- JSON: defines which JSON converter is used to convert an object.
- JSON Schema

- Byte Array

Kafka converters take in a specific record set from different types of data scheme and use a converter to convert the data to format it into different types for data ingestion. There are several guidelines for choosing a serialization formats such as:

- Schema (ie. provides contract between services) Message formats such as Avro and Protobuf have strong schema support. JSON and delimited strings do not have much schema support.
- Ecosystem compatibility (Avro, JSON, and Protobuf typically first-class citizens on Confluent Platform.
- Message size (JSON relies on compression where Avro and Protobuf are binary formats and have smaller message size
- Language support (Avro is strongly Java based and with Protobuf, Go is typically more used.

What's a Key-Value (KV) database?

A key value database is typically a nonrelational database which uses keys to store data. It works almost like a hashmap where each key has a unique identifier. Keys and values can be anything from simple objects to complex compound objects.

What are KV databases' advantages and disadvantages?

KV database has the following advantages:

- Simple data format which means fast read and write operations
- Values can be practically anything including JSON along with flexible schemas.
- It uses a key-value method to store data, it can store, retrieve, and

  manage arrays of data, essentially its a hash table

KV database has the following disadvantages:

- Optimized only for data given only a single key and value.
- Not ideal for lookup when key is not present. Each lookup where the key is not known would require scanning the entire collection.

List some popular KV databases.

- Amazon DynamoDB
- Amazon ElasticCache
- Redis
- Couchbase
- ScyllaDB

What are KV databases' advantages and disadvantages?

- Advantages
- Scalability: scalable because its able to take a lot of requests
- Speed: able to process constant requests for read and writes
- Flexibility

○ **Disadvantages**

- Only optimized for data with a single key and value
- Not very well made for reading/looking up data

List some popular KV databases.

○ Redis

○ ScyllaDB

○ Amazon DynamoDB

○ Azure table storage

9. List some possible applications that can be implemented by using the uploaded dataset.

a. This dataset could be usable in various artificial intelligence applications, such as detecting weather patterns, detecting moving obstacles, etc. uploading the dataset to the cloud helps by allowing anyone from anywhere in the world to access this dataset.

All the steps listed from step 4 to 6 are captured in the videos bellow

Video Link 1

https://drive.google.com/file/d/1xKWV5tuUGVSzRFFYwpcoMC4Ej1bHWmVN/view?usp=sharing

Video Link 2

https://drive.google.com/file/d/1CX2R8Fy-pRHCvH52WXeZz-biDvJlppu5/view?usp=sharing

Video Link 3

https://drive.google.com/file/d/1LlymU5CsAxJHDhBAnl4Kbvx2crXGYjcp/view?usp=sharing

Video Link 4

https://drive.google.com/file/d/1Z9zVGNKGTgdBEkj99v_35YwMak9fwUla/view?usp=sharing