SOFE 4630U: Cloud Computing
February 15, 2022
Group 11 - Group Report
Project Milestone - Data Storage Implementation: KV + relational

Fajer Zayed (100672347), Ireni Ruthirakuhan (100657302), Raveenth Maheswaran (100704540), Yale Wang (100673933)

GitHub Link: https://github.com/fzayed/Project-Milestone-Group-11.git

Sink and Source Connectors

- Sink Connector
    - A type of connector that connects to a Kafka topic, where the data from that topic can be exported to a relational database
    - Ability to export data from topic(s) to a relational database
    - Don't need to know where the data comes from
        - Loose coupling
    - Gets generic representation of the data and then sink connector plugin writes it to the target system
- Source Connector
    - A type of connector that connects to a relational database, and imports data from that database into a Kafka topic
    - Ability to import data/information from a relational database to a topic
    - Source doesn't need to know where the data is being synced to
        - Loose coupling
    - Interfaces with the source API and extracts the schema and then creates an external object (connect record = an object within the API) and passes it on

The applications/advantages of using Kafka Connectors with data storage.

- Applications use Kafka Connectors in the case where you need to receive data from external systems
- Advantages of using Kafka Connectors are:
    - Flexibility
        - Decouples source and target
        - Changes in source/sink can be done without impacting the other
    - Scalability
        - Buffer for the data
        - When there is too much happening it provides a basic queuing functionality
    - Fault tolerance
        - Connection to sink/source may go down, with knowing that you are still producing data
            - Data is stored into kafka
    - Building pipelines
        - With the data in some exterior location and you want to through kafka to another location → transactional db to an object store

How do Kafka connectors maintain availability?

- Kafka connectors maintain availability through taking in large amounts of data from databases into a topic. This is where data then would become available towards stream processing
- Since data flow through Kafka Connect, the connectors wouldn't be congested with high amounts of data

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

- Converters receive connect records and turns them into bytes
    - Writes it as key and value into a kafka cluster
- Serializer in a regular kafka producer

- Popular converters
    - Json Schema
        - The data can be serialized or deserialized through JSON, which is commonly used for formatting and grouping data that is highly used in common platforms
        - Advantages include that it generates clear and readable documentation
    - Avro
        - Advantages:
            - It's binary format
            - Fast as it doesn't require code generation
            - Flexibility as it has wide variety of programming languages
    - ProtoBuf
        - Advantages include
            - that it is faster, simpler and smaller
            - Has RPC support
            - Structure validation allows a predefined + larger structure

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

- Stores messages as key and value pairs
    - Ability to store, retrieve and update data
- Non relational database
- The key that is associated with the value can then be used for various reasons
    - CRUD operations
    - Provide horizontal scaling
    - Can be highly partitional

What are KV databases' advantages and disadvantages?

- Advantages of Key-Value databases include:
    - Scalable
        - Increase on database load/data has no negative impact on performance
        - Infinitely scalable horizontally
    - Speed/responsiveness
    - Reliability
    - Flexibility
        - DB can be easily relocated w/o change in structure
    - Easy-to-use
        - Easy to implement KV database, compared to other types of databases
- Disadvantages of Key-Value databases include:
    - Single key value

- - No Query language
    - May not be able to import data into a different KV database
    - Not optimized for lookup
  - Values cannot be filtered
  - Performance with big data
    - The more complex the queries and data is, the more it affects the performance

List some popular KV databases.

- Amazon DynamoDB
- Aerospike
- Redis
- BerkeleyDB
- NoSQL Database (Oracle)

Video Link

https://drive.google.com/file/d/1AHV-axfHvXO0PNC1w0eL6AVXv5Hetz_Z/view?usp=sharing

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

- Ensure a robot is working
- Measure accuracy of implemented robotics algorithms
- Adjust robot movements on the fly