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

Lab 4: Project Milestone-- Data Processing: Dataflow- apache beam

Mitul Patel

Lab 4: Project Milestone-- Data Processing: Dataflow- apache beam

1. Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Other types of processing services are the following:

- Oracle Database
    - Good support and service
    - Easy to deploy
    - Good contracting and evaluation
- Apache Spark
    - Scalable across multiple machine
    - Flexible with programming languages
- Databricks Lakehouse Platform
    - Detailed documentation
    - Data visualization
- Neo4j

Main difference between Dataflow and DataProc is that DataProc is primarily more automated, scales on-demand and UI-driven compared to Dataflow which follows batch and stream processing on a given data. A Dataflow is responsible for creating new pipelines for resources and to process data on-demand. Furthermore, Apache Spark is

very quick and has a large unified engine. It is fairly easy to use when compared to other data processing services. There are also some disadvantages to using Apache spark where if you have small files or do not require real-time processing. Lastly, Apache spark is also limited by its iterative processing along with its use of fewer algorithms.

2.

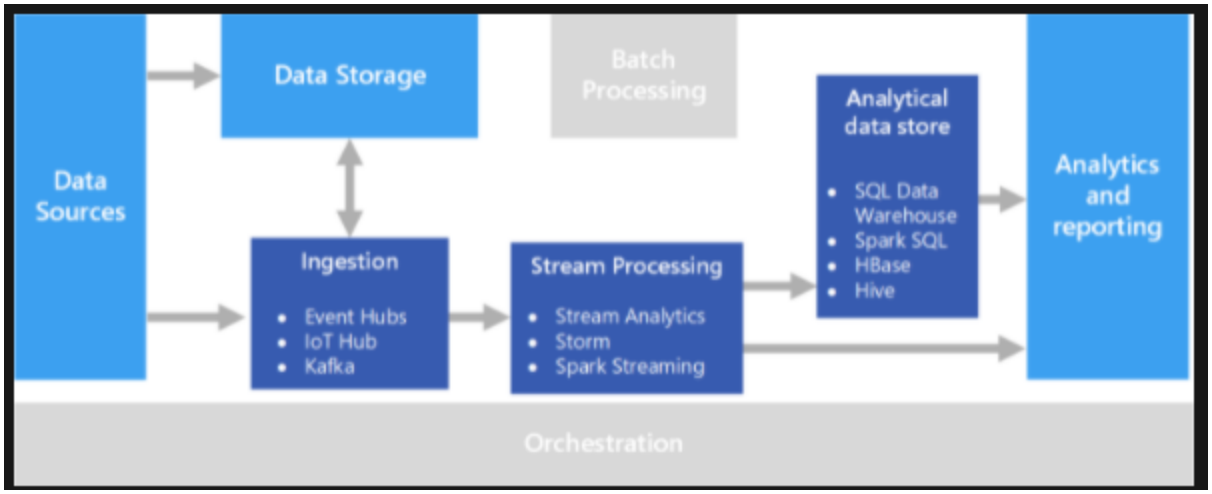
Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to

- The application.
- Its impact.
- The used dataset (size, schema/structure).
- A graph showing the proposed pipeline(s).
- List of other tools (AI, clustering,…) needed to implement that application.

A practical application use case for both batch and stream processing is a smart parking lot. There are multiple sensors which relay the information to a gateway. Next the gateway will forward the information to a kafka cluster where the data will be processed using both stream and batch processing. The stream processing would be required for real time analysis where it would update users on availability within a parking lot. The batch processing would be responsible for gathering long term data for treads and other types of microservices.

The impact for an application of this scale would allow multiple users to interact with one another and allow many different types of data to be processed and stored. The impact of using stream processing would provide real time updates to a social media application. The real time processing can include message delivery and status. A batch processing on a dataset would include something where data is collected on users time spent on each post or the type of posts. Furthermore, using the type of data the user browses on the application, ads could be generated and displayed to the user.

The previous dataset would be good for a social media application where several different types of data are collected and generated through batch and stream processing with tons and tons of data volume, velocity and variety. An example for a social media application such as facebook would be pdf, images, all types of files, messages, json data, videos and many more. The size of the dataset would be massive as millions of users would be uploading millions of data all round the world.



This image above could be used as a proposed pipeline. Where any data generated from the source would be saved and ingested and sent to a stream processing for stream analytics. This is where the real time data would be generated. After which data would be fed into analytical data stores such as warehouses, and many more to provide analytics and reporting.

Lastly, machine learning could be leveraged to perform all types of pattern recognition. An example for a social media application would be that data generated from the users would be processed using unsupervised machine learning to gather information on what types of ads to provide to each user. A profile for each user would be generated.