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

Project Deliverable 4: Data Processing - Data Flow

Group 5

Date: 03/29/2022
Student Names:
Alexander Pitcher (100693749)
Mitul Patel (100700131)
Clarissa Branje (100716458)
Amin Khakpour(100669547)

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.

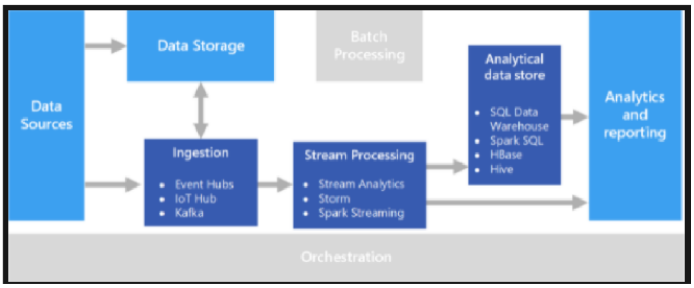
- Apache Spark
    - Some advantages
        - Quick
        - Part of one large unified engine
        - Very easy to use compared to other processing services
    - Some disadvantages
        - Cannot do real time data processing
        - Mainly optimized for large files
        - Expensive
    - Limitations
        - Not good at processing small files
        - Cannot process in real time
        - Iterative processing
        - Uses fewer algorithms compared to other services
- Apache Hadoop
    - Some advantages
        - Scalability
        - Performance
        - Ease of use

        - Cost effective
    - Some disadvantages
        - Security
        - Processing overhead
        - Iterative Processing
    - Limitations
        - Struggles with small files
        - Slow processing speed
        - No real time data processing
- Oracle Database
    - Some advantages
        - Good support and service
        - Easy to deploy
    - Some disadvantages
        - Extremely High cost of licensing
    - Limitations
        - Various limitations with size of files used
- Azure HDInsight
    - Some advantages
        - Able to keep data size and volume separate from the size of cluster
        - Low cost
    - Some disadvantages
        - Complexity
        - Not very good support
        - Requires expertise with the platform
    - Limitations
        - Does Not support some Apache services

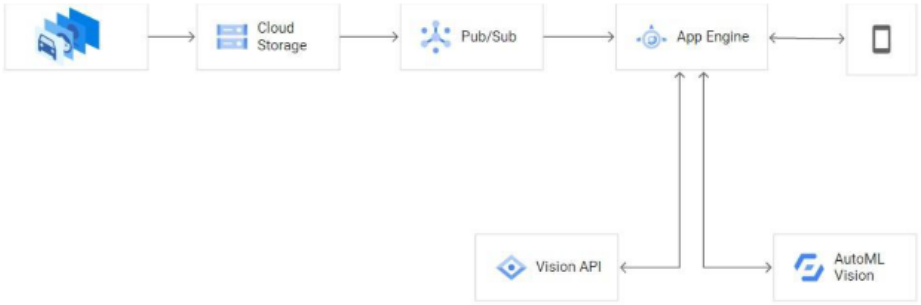
The main difference between Dataflow and DataProc is that DataProc is primarily more automated, scales on-demand and UI-driven compared to Dataflow which follows batch and stream processing on a given data. A dataflow is responsible for creating new pipelines for resources and to process data on-demand.

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

- Social Media Platform
    - The impact
        - A social media platform that is consistently taking in large amounts of data and needs to have it processed quickly. Using a mix of stream and batch, the processing speeds will be high

    - The dataset used
        - Social Influence on Shopping(163.59 KB, using .csv),
    - List of tools needed to implement the platform
        - MySql, Oracle, Microsoft Data Program, etc.
- Smart Parking Lot
    - The Impact
        - This would allow multiple users to interact with one another and allow many different types of data to be processed and stored.
    - The dataset used
        - Previous Dataset from last lab
    - Graph of pipelines
    - 

    - List of tools needed to implement the platform
        - Machine learning could be used to perform all types of pattern recognition
- Streaming Service (i.e. Netflix)
    - The Impact
        - Could be used to update (Adding/removing movies) the current movies, as well as process real-time transactions such as creating new accounts, or processing media requests
    - The dataset used
        - The Movie Dataset(239Mb)
    - Features
        - Posters, backdrops, budget, revenue, release dates, languages, productions countries, and companies
- Predicting Appearance Change
    - The Impact
        - The purpose of this can be predicting the change of the surface of the objects or the environment based on the time of the day or the season of the year
    - The dataset used
        - The dataset from the previous lab
    - Graph showing the proposed pipelines



-

- List of Tools
    - Machine learning that helps to correct the learning process
    - AI
    - Dataflow because it needs clustering