Project Milestone-- Data Processing: Dataflow- apache beam

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.

Azure HDInsight is a Hadoop component distribution on the cloud. Azure HDInsight enables processing vast volumes of data in a configurable environment simple, quick, and cost-effective. You can utilise Hadoop, Spark, Hive, LLAP, Kafka, Storm, R, and other popular open-source frameworks.

DataProc features:

- You can work on a massive quantity of data each day with your existing MapReduce without any overhead concerns.
- You may communicate cluster data to your apps using the built-in monitoring system. You can obtain rapid reports from the system, and you can also save data in Google's BigQuery.
- To customise and execute categorization algorithms, use Spark Machine Learning Libraries and Data Science.

DataFlow features:

- At the same time, ETL (extract, convert, and load) data into several data warehouses.
- Data science approaches are used to process massive volumes of data for study and forecasts. For example, genetics, weather, and financial information.
- To handle a large number of parallelization processes, Dataflow is considered a MapReduce substitute.
  Real-time, user, managerial, financial, and retail sales data may all be scanned.

Differences:

- Dataproc allows manual cluster provisioning, but Dataflow enables automated cluster provisioning.
- If dev-ops approach is the issue, dataproc is a better choice, but dataflow is a serverless approach.
- Dataproc is a better option for real-time data collection, whereas The data lake, data collection, cleaning, cloud, and workload processing are highly rated for the Dataflow.

Similarities:

- Both are recognized as Big Data processing.
- Both are products of google

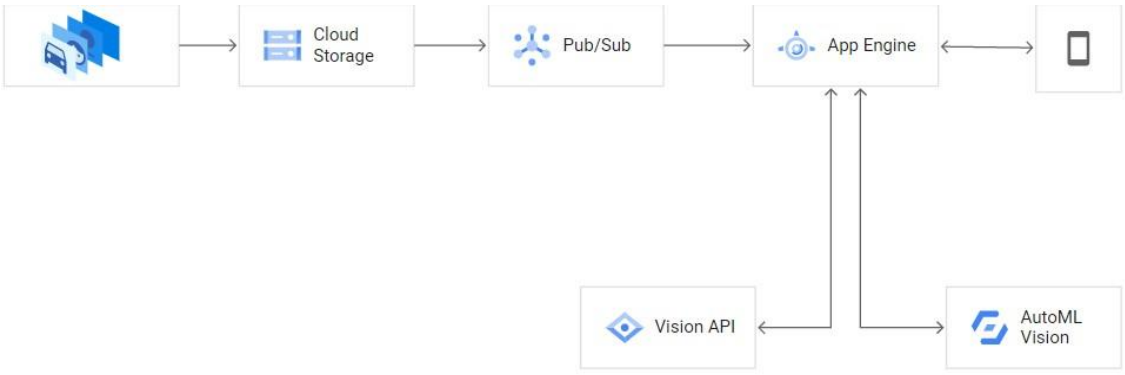
Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decided to use another dataset, It should maintain both variety and huge volume. Your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

The dataset that is being used is The University of Michigan North Campus Long-Term Vision and LIDAR Dataset (NCLT). The NCLT dataset contains 34.9 hours of logs covering 147.4 km of robot trajectory, and was collected in 27 discrete mapping sessions, Fig. 1(b). Each session covers roughly the entire mapped area and contains both indoor and outdoor environments.

The application.
Predicting appearance change
Its impact.
The purpose of this can be predicting the change of the surface of the objects or the environment bsed on the time of the day or the season of the year.
The used dataset (size, schema/structure).
To achieve the goal of the application we are going to use Images and Sensors from this dataset

2012-01-08	lb3.tar.gz (106 GB)	sen.tar.gz (114 MB)

A graph showing the proposed pipeline(s).



In this approach I used Snaowflake because of Ease of Implementation and Cloud-First Approach

List of other tools (AI, clustering,…) needed to implement that application.

Other tool that it needs is Machine learning Module that helps to correct the learning process
The other part is implementing AI
Because of the dataflow it needs clustering as well.