

Project Milestone 4: Data Processing

Danial Asghar

100671850

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.

The three data processing services that Google Cloud Platform offers are the following: DataProc, DataFlow and DataPrep, targeting various different consumer ETL needs. These services are advertised by Google as the following:

- DataProc: “a fully managed and highly scalable service for running Apache Spark, Apache Flink, Presto and other opensource tools and frameworks... Used for data lake modernization, ETL, and secure data science” [1].
- DataFlow: “a fully managed streaming analytics service that minimizes latency, processing time, and cost through autoscaling and batch processing” [2].
- DataPrep: “an intelligent cloud data service to visually explore, clean, and prepare data for analysis and machine learning” [3].

Firstly, DataProc is GCP’s offering for customers to provide a managed Spark/Hadoop service. Customers can migrate their full Spark/Hadoop clusters to DataFlow which allows for creation and management of the clusters. DataProc allows consumers to use MapReduce and process large amounts of data quickly. It is a flexible service, offering serverless deployments, logging, and monitoring tools which lets users focus on their business logic, instead of having to manage infrastructure. It allows users to containerize their Spark Jobs, with full management using GKE. DataProc offers Spark ML libraries and DataScience allowing users to create classification algorithms.

Secondly, DataFlow is GCP’s offering for customers to execute different data processing patterns. It uses Apache Beam SDK allowing users to develop stream and batch processing jobs. The service’s main purpose is to simplify management and operations of Big Data. It offers streaming data analytics with lower data latency. Autoscaling ensures customers only pay per use and jobs can be scaled to match spiky loads.

Lastly, DataPrep is a tool created to solve the problems of lacking data visualization, slow processing of data and managing redundant data. It allows users to view data by transforming it into tabular form, i.e. JSON, CSV, etc. DataPrep is a middleware service for processing data to be used in other services like GCP’s BigQuery. It is an elastic service because it is serverless, works at scale and doesn’t require managing any infrastructure. DataPrep has the ability to automatically “detect schemas, data types, possible joins, and anomalies such as missing values, outliers, and duplicates” [3].

All the previously explained services are quite similar in nature, as they all target the operation, management, and analysis of Big Data. They are all platform-as-a-service managed offerings by Google in GCP. Although, there are some large differences in the services. DataProc is Dev-ops focused, DataFlow is serverless approach focused and DataPrep is UI-driven. Differences are summarized in the following chart:

Points of Difference	DataProc	DataFlow	DataPrep
Provisioning	Manually	Serverless, automated	Fully automated
System Integration	Apache Spark and Hadoop	Apache Beam	BigTable and BigQuery
Ease of Use	Simple	Relatively difficult	Easy
Approach	Hands on, Dev-ops approach	Fully managed	Fully managed
Unique ability	Data science / ML	Batch and stream processing	UI driven data processing

Table 1: Listing the service differences [4]

Suggest a practical application using both stream and batch processing that can be applied to a given dataset.

Dataset

The dataset to be used will be the North Campus Long-Term (NCLT) dataset which contains many longer-term dynamic elements like pedestrians, construction, vehicle traffic, weather conditions, etc. This dataset provides a spatially and temporally large-scale data set, with large quantities of lidar data of indoor and outdoor coverage, observing a variety of environment changes. It includes omnidirectional imagery, 3D lidar imagery, planar lidar, GPS, and proprioceptive sensors for odometry. The sensor data is provided as separate CSV files, per sensor but does not label the columns. The image data set provides a large set of unlabelled images. This dataset is a great resource for any of the following tasks:

- Object recognition and computer vision
- Obstacle detection and tracking
- Weather detection

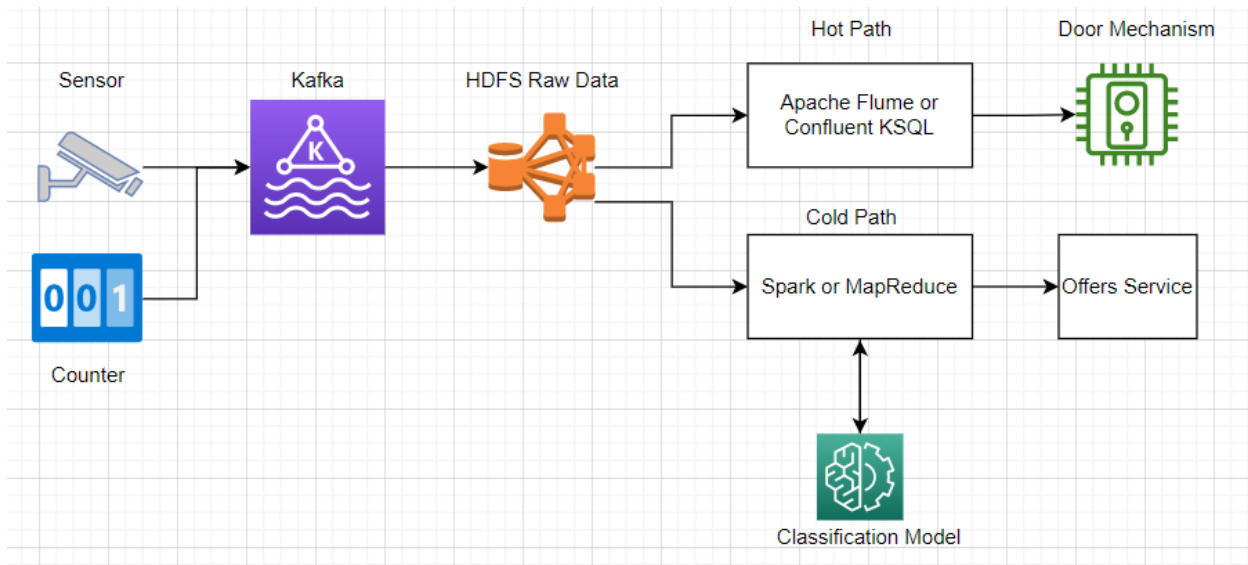
Application

The application to be built will be using object recognition to determine how many people are within a location at a given time.

Impact

The need for people management started in 2020 with COVID. Due to government regulations most stores, malls and restaurants were limited to only a certain percentage of patronage. The application would use Big Data to create a classification model to determine whether a male or female adult or child entered the mall. For countries with ongoing COVID restrictions, the hot path streaming analytics would ensure that stores could have automated gates which allowed only a specific amount of people in the store at one time, determining them using object recognition. The cold path analysis could be used for marketing and offers because the model could be used to determine what kind of items were popular at what time and send those to visitors at a specific time of day.

Pipelines



The application pipeline's rough architecture is displayed in the image above. The pieces are as follows:

- **Sensors:** cameras are required to capture video data when people enter and exit. A proximity counter can be used as trigger as well to determine if someone entered/exited the location.
- **Kafka:** Event streaming platform required to ingest the data into the Cloud backend.
- **HDFS:** Hadoop distributed file system to store raw, immutable data files.
- **Hot Path:** Apache Flume or Confluent KSQL would be used here to do near real-time processing of sensor data. Spark Streaming would be a viable alternative as well but that processes data in small batches and might not be fast enough to keep the counter real-time in busy times with lots of traffic. Depending on how fast the system is, it could be used to directly control gate control mechanism, reducing the need of a security guard to watch people enter and exit.
- **Cold Path:** Spark or MapReduce would be used to conduct batch processing. This path would also rely on a classification model, i.e. Deep neural network, to determine characteristics like gender and age of the visitors. Once the predictions are done then specific marketing offers can be targeted to the customers once the cold path is processed at a specific time.

Tools

Tools and services which would be required for this system include the following:

- **Machine learning component** – for example, deployed Multilayer perceptron classification neural net.
- **HDFS** – data lake required to store raw data.
- **Hardware libraries** – interfaces to interact with all the sensors

References

- [1]"Dataproc | Google Cloud", *Google Cloud*, 2022. [Online]. Available: <https://cloud.google.com/dataproc>. [Accessed: 29- Mar- 2022].
- [2]"Dataflow | Google Cloud", *Google Cloud*, 2022. [Online]. Available: <https://cloud.google.com/dataflow>. [Accessed: 29- Mar- 2022].
- [3]"Dataprep by Trifacta | Google Cloud", *Google Cloud*, 2022. [Online]. Available: <https://cloud.google.com/dataprep>. [Accessed: 29- Mar- 2022].
- [4]D. difference?, "What's the Difference Between Dataproc, Dataflow & Dataprep?", *WisdomPlexus*, 2022. [Online]. Available: <https://wisdomplexus.com/blogs/dataproc-vs-dataflow-vs-dataprep/>. [Accessed: 29- Mar- 2022].