Cloud Computing
Project Deliverable:
Data Processing: Dataflow - Apache

March 29, 2022

Group 7
Owais Quadri 100697281
Tegveer Singh 100730432
Danial Asghar 100671850
Shayan Sepasdar 100722542

GitHub Link
https://github.com/danialasghar/Cloud-Group-7

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 targeting various different consumer ETL needs include DataProc, DataFlow and AWS Kinesis. 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].
- AWS Kinesis: “makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.” .

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.

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.

AWS Kinesis is Amazon’s p-a-a-s offering for collecting, ingesting and analyzing video and data event streams. Kinesis is responsible for rapid and continuous data intake and aggregation. Data managed by Kinesis ensures the system has durability and elasticity. Kinesis can be used in any of the following examples: real-time metrics or monitoring, data analytics, stream processing, accelerated data feed intake and more [4].

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 or AWS. These services’ characteristics are summarized in the table below:

Criteria	**DataProc**	**DataFlow**	**AWS Kinesis**
Main Difference	Create a cluster of servers to perform ETL tasks, very similar to how Hadoop would work	The ETL tasks are serverless and use Java/Python/Go Apache Beam SDK to perform Batch and Stream processing tasks	Real-time processing for large amounts of data and runs on AWS EC2 instances
Compatibility	Works by using Spark and Hadoop clusters and can be integrated with other GCP services like BigQuery, BigTable	Works with open source programming framework Apache Beam	Can be integrated with S3, DynamoDB and RedShift
Limits	7500 requests per minute	Max 1000 workers with 15000 messages/30 seconds per worker	Higher transaction limits than GCP services. One call retrieves 10000 records (10MB)
Advantages	- Very low overhead and can handle a large amount of data - Clusters that are not needed can be removed - Integrated with Spark for Machine Learning and Data Science algorithms	- Provides both Stream and Batch Processing - Does not run on top of Hadoop reducing operational overhead - Can Integrate with other services AWS, Hadoop, etc - Can process datasets with unknown sizes due to Autoscaling	- Support large number of consumers - Quick data processing with real-time graphs and analytics - Very high read and write throughput - Kinesis Client Library available for languages like Java
Disadvantages	- Reduction in performance as data increases - High Dependency on Hadoop leading to low portability - No Stream processing - Limited to datasets	- Slightly more expensive in comparison to DataProc - More bound to google technologies	- Higher cost than DataFlow and DataProc as charged per shard per hour - Capability limited by the EC2 instance Kinesis is running on

	of known size		

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.

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

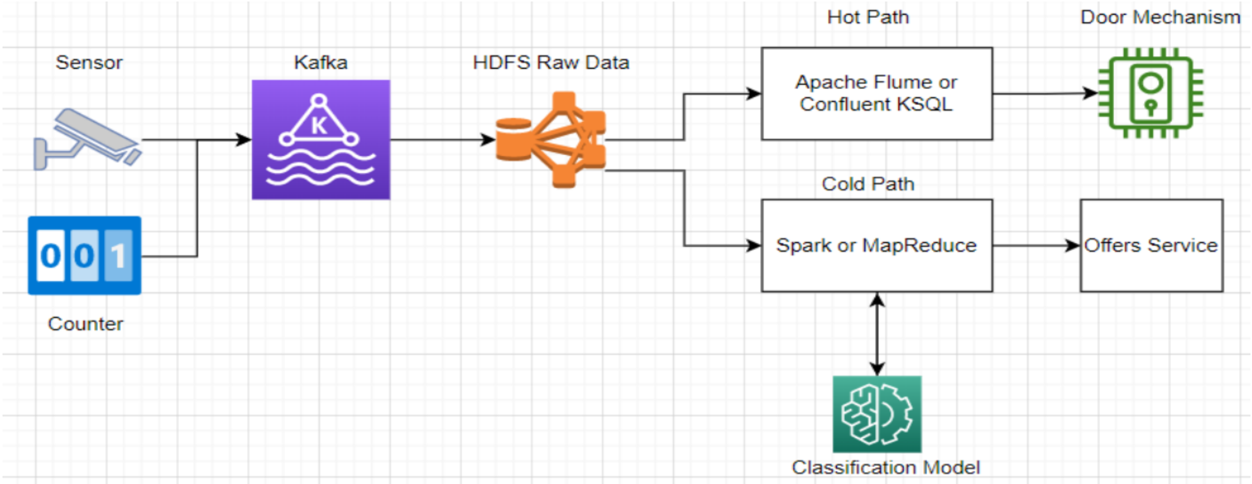
Application

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 realtime 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] "Amazon Kinesis", Amazon Cloud, 2022. [Online]. Available: https://aws.amazon.com/kinesis/. [Accessed: 29- Mar- 2022].

[4] “What is Amazon Kinesis", Amazon Cloud, 2022. [Online]. Available: https://docs.aws.amazon.com/streams/latest/dev/introduction.html. [Accessed: 29- Mar- 2022].