OntarioTech UNIVERSITY

Project Milestone

Data Processing: Dataflow- apache beam

Name: Mihir Patel
Student Number: 100702168
Date: January 31, 2022

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.**

- The three most common data processing services that Google Cloud platform provides are DataFlow, DataPrep and DataProc.
  - DataFlow is a cloud-based data processing solution that can handle batch as well as real-time data streaming. It has capabilities such as resource auto-scaling and dynamic work rebalancing, as well as flexible scheduling and ready-to-use real-time AI patterns.
  - Google DataPrep is a data service that allows you to explore, clean, and prepare structured and unstructured data in the cloud. The main feachers are Active profiling, optimized processing throughput and Data quality rules.
  - Google Cloud DataProc is a managed service for Spark and ApacheHadoop jobs that enables batch processing, querying, streaming, and machine learning using open source data technologies.

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**

Dataset

- The dataset for autonomous driving cars will be taken from Google Open Images. The dataset includes both labeled and unlabeled objects such as cars, pedestrians, traffic lights, fire hydrants, buses, trucks, and signs, among other things. Over 400,000 photos are included in the dataset, which depict a wide range of environmental changes. Sensor data is also included in the dataset, which is essential for calculating the distance between the objects and the device. This dataset can be utilized in the creation and design of object detection systems.

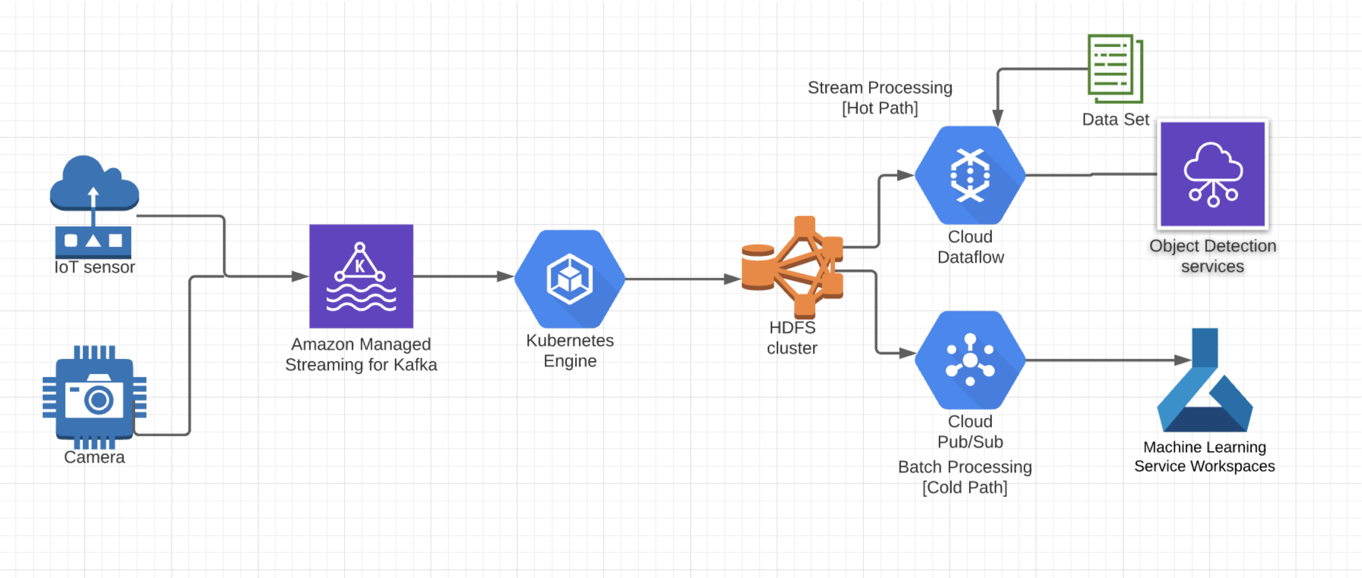
The application

- Using stream and batch processing, the above dataset will be used to create a portable object detection system that may be used in cars to assist drivers in driving safely by detecting objects in real time.

Its impact

- Object detection is becoming more important as the frequency of car accidents in Ontario continues to rise. According to statistics, there is a 26% increase in car accidents every year. As a result, object detection methods are required to resolve the dilemma. The device would employ Big Data to construct an object detection model that would be able to successfully detect objects and assist drivers in driving safely on the road.The steam processing [hot path] will be utilized to detect objects and alert users when they approach them. To accomplish this the system will use the linked sensors to process data such as distance and speed. As of batch processing [cold path], the primary goal is to gather data that can be used to improve object detection. It will also be utilized to provide system upgrades, which will improve the user experience.

A graph showing the proposed pipeline(s)


Stream Processing [Hot Path]
Data Set
IoT sensor
Amazon Managed Streaming for Kafka
Kubernetes Engine
HDFS cluster
Cloud Dataflow
Object Detection services
Camera
Cloud Pub/Sub
Batch Processing [Cold Path]
Machine Learning Service Workspaces

List of other tools

- Machine learning library for example tensorflow and tensorflow lite
- Big Data, Apache Beam, Dataflow
- HDFS for handling large data sets running on commodity hardware