SOFE 4630 Winter 2022 - Cloud Computing

Project Milestone#4 : Data Processing: Dataflow- apache beam

Course Group No: 12

Group Members:

Ashwin Shanmugam - 100700236
Faazil Shaikh - 100707829
Evans Mosomi - 100719552
Samuel Rantetoding - 100694161

Q3.

- Video Demo: https://drive.google.com/file/d/1RUnxRnh4xs3QYBYnyluIzAvNI9cZw_Pz/view?usp=sharing

Q4.

- Video Demo: https://drive.google.com/file/d/1WSGmnLMNNuYuQs0-34VSjbbJHQ_lb1J3/view?usp=sharing

Q6.

- The other processing service is DataPrep.
- Differences:
  - Automation of a cluster is done by DataPrep and DataFlow, while DataProc relies on manual setup for a cluster.
  - Suitable for BigQuery, while the other processing service is more suitable for Apache and Hadoop.
- Advantages:
  - Improve data visualization as the user interface is displayed in a tabular format.
  - High security as the dataset is protected due to the access management applied.
  - Utilized machine learning to remove unnecessary data to reduce the size of the datasets.
- Disadvantage:
  - Not easy to use as DataFlow.
  - Heavily used for systems that are focused on user interface.
  - Not working well with Apache Hadoop for the processing service.
- Limitations:
  - There is maximum of 1000 datasets can be stored in the database
  - There is a lack of memory size that can be used for storing the datasets.
  - Does not support UTF-32.

Q7.

- Application:
  - The LIDAR dataset is useful for vulnerability assessment, such as coastal vulnerability analysis.
  - It can help scientists to trace any possibilities of disasters, like a tsunami could occur at any moment.
- Impact:
  - The impact of this is that it could help prevent any disasters that could happen near the coastline, which could lead to the harmful effects to the population near the coastline.

        - The system could be applied by first applying a coastal modeling to study a simple rendering of a real coastline.
        - This could help the scientists to find any possibility that an improvement can be made to the coastline area or any possibility of a disaster that could happen at the coastline.
- Datasets:
    - The dataset used is a long-term vision and LIDAR dataset called NCLT dataset.
    - It is a dataset that utilizes long term vision and can be helpful for object detection.
- Tools:
    - Climate Change Vulnerability Assessment Tool for Coastal Habitats, which is useful to gather additional information for vulnerability assessment.
    - HDFS, which is used to keep the data that wants to be used by the application.

Upload the used files, reports, and videos (links) into your repository.

- Repository Link: https://github.com/ashwin1609/CloudComputing-Milestone4