GCP Dataflow Project using Python and Apache Beam

Business Overview

Google Cloud is a collection of physical assets, such as computers and hard disk drives, and virtual resources, such as virtual machines (VMs), housed in Google data centers worldwide. This resource distribution has various advantages, including redundancy in a failure and decreased latency by putting resources closer to customers. This release also presents some guidelines for combining resources.
GCP offers a web-based graphical user interface for managing Google Cloud projects and resources. If a user prefers to work at the command line, the G-Cloud command-line tool can handle most Google Cloud activities.

This is the third project in the GCP Roadmap project series, the previous projects utilize services such as PubSub, Compute Engine, Cloud Storage, and BigQuery. In this project, we will explore GCP Dataflow with Apache Beam:

Tech Stack

➔ Language: Python3
➔ Services: Cloud Storage, Dataflow, Apache Beam, BigQuery, G-Cloud SDK, Pub/Sub

Cloud Storage

Cloud Storage is a service that allows users to store their data on the Google Cloud. An object is an immutable piece of data that consists of a file in any format. Objects can be stored in containers known as buckets. All buckets are related to a project, and the user may organize their projects into organizations. After starting a project, users may create Cloud Storage buckets, upload things to the buckets, and get objects. Users can also give rights to make data accessible to certain domains or for specific use cases such as establishing a website.

BigQuery

Google Bigquery is a Cloud Datawarehouse powered by Google, which is Serverless, highly scalable, and cost-effectively designed for making data driven business decisions quickly. It offers both the batch and streaming insertion capabilities and is integrated with Tensorflow as well to perform machine learning using SQL like dialects.

Pub/Sub

Pub/Sub allows services to interact asynchronously and is used in streaming analytics and data integration pipelines to ingest and disseminate data. It, like Kafka, allows users to design systems of event producers and consumers, known as publishers and subscribers. Publishers connect with subscribers asynchronously by disseminating events rather than synchronous remote procedure calls (RPCs). Publishers transmit events to the Pub/Sub service without considering how or when these events will be handled. Pub/Sub then sends events to any services that need to respond to them. This type of asynchronous integration improves the overall flexibility and robustness of the system.

Apache Beam

Apache Beam is a batch and streaming data processing unified programming model. It offers many APIs for interacting with various data sources and processing data using various backends, such as Spark or Dataflow. As a result, the data may be stored elsewhere, and computation can be performed on it in a serverless manner or on a specified backend.

Dataflow

Google Cloud Dataflow is a cloud-based data processing service that can handle batch and real-time data streaming. It allows users to build processing pipelines for integrating, preparing and analyzing massive data sets, which is typical of big data processing.

Approach

- Read JSON encoded messages from the GCS file, transforms the message data, and write the results to BigQuery.
- Read JSON encoded messages from Pub/Sub, transforms the message data, and write the results to BigQuery

Key Takeaways

- Introduction to Apache Beam with Dataflow Runner
- Introduction to Dataflow features
- Creating a demo Beam Pipeline
- Publishing Flights data to Pub/Sub for Dataflow Streaming
- Understanding Dataflow parameters
- Introduction to Clustering in BigQuery
- Streaming Job Execution using Dataflow runner
- Running Batch pipeline using Dataflow
- Using BigQuery to write the transformed data