GCP-IaC: Building Efficient ETL Pipeline for Financial Data Analytics

Business Overview

This project focuses on creating an efficient ETL (Extract, Transform, Load) pipeline for the Financial Data Set on the Google Cloud Platform (GCP). Critical to this project is using Infrastructure as Code (IaC) principles, specifically GCP Deployment Manager, to automate and manage infrastructure setup.

Automated deployable infrastructure plays a vital role in adhering to best practices in the technology industry. It brings several benefits, including consistency and reproducibility. By defining infrastructure as code, the deployment process becomes consistent and repeatable, eliminating variations caused by manual configuration. This ensures identical environment setups and enables easy testing, staging, or development reproduction.

Infrastructure as Code also enables version control and collaboration. Teams can track changes, review modifications, and roll back to previous versions if needed. This fosters collaboration among team members, ensuring transparency, accountability, and efficient collaboration that reduces time-to-market.

Scalability and agility are other significant advantages of automated deployable infrastructure. With IaC, teams can easily define and deploy additional resources or modify existing infrastructure to adapt to changing business needs. This enables organizations to respond quickly to market dynamics, scale applications or services effectively, and stay ahead of the competition.

Disaster recovery and reliability are crucial considerations. Infrastructure as Code provides a solid foundation for robust disaster recovery strategies. By creating backup and recovery processes, organizations reduce the risk of data loss and ensure business continuity. In the event of an incident or failure, infrastructure can be easily recreated from the code, minimizing downtime and mitigating potential financial and reputational losses.

Cost optimization is another key advantage of automated infrastructure deployment. Teams can define infrastructure templates incorporating cost-saving measures, such as rightsizing instances, using spot instances, or implementing auto-scaling rules. This optimizes resource utilization and reduces unnecessary expenses, leading to significant cost savings.

Organizations can enhance operational efficiency, reduce manual errors, improve collaboration, and achieve greater agility and scalability by adopting best practices in the technology industry, such as automated deployable infrastructure through Infrastructure as Code.

Aim

The aim of this project is to build an ETL pipeline for the Financial Data Set on GCP, enabling the extraction, transformation, and loading of data from a SQL server to BigQuery for analytics purposes.

Dataset Description

This project uses equity financial data from BseIndia, which includes stock price history from various industry segments. A few of the fields included in the dataset are as follows

- Open_price
- High_price
- Low_price
- Close_price
- No_of_shares
- No_of_Trades
- Total_turnover

Approach

1. Use GCP Deployment Manager to create necessary resources like GCS buckets, BigQuery tables, and a virtual machine.
2. Install Apache NiFi on the virtual machine to extract data from the SQL server and dump it into a GCS bucket.
3. Create a cloud function to monitor the bucket for changes and trigger a PySpark job using Cloud Dataproc.
4. Utilize workflow templates to create a Dataproc cluster and execute the PySpark transformation job.
5. Load the transformed data into BigQuery for analytics and optionally store a backup in a cloud storage bucket.

Tech Stack

Language: Python, SQL
Services: SQL Server, AWS RDS, GCP Compute Engine, GCP Cloud Functions, Apache NiFi, GCP Cloud Storage, GCP BigQuery, GCP Dataproc, GCP Deployment Manager

GCP Dataproc

Google Cloud Dataproc is a managed Apache Hadoop and Apache Spark service offered by Google Cloud Platform (GCP). It provides a fully managed and scalable platform for processing big data workloads. With Dataproc, users can easily create and manage clusters, allowing for fast and cost-effective processing of large datasets. It supports various data processing frameworks, including Spark, Hadoop, Hive, and Pig, and offers integration with other GCP services. Dataproc provides automatic scaling, high availability, and easy cluster management, enabling users to focus on their data analysis and insights rather than infrastructure management.

GCP Cloud Functions
Google Cloud Functions is a serverless compute platform provided by Google Cloud Platform (GCP). It allows developers to build and deploy event-driven functions that automatically respond to events from various cloud services. With Cloud Functions, developers can write code in popular programming languages like JavaScript, Python, and Go, without the need to manage servers or infrastructure. Functions can be triggered by events from GCP services, such as Cloud Storage, Pub/Sub, or Firestore, as well as HTTP requests. Cloud Functions offers automatic scaling, pay-per-use pricing, and seamless integration with other GCP services, enabling developers to focus on writing code and delivering business logic.

GCP Deployment Manager
Google Cloud Deployment Manager is a robust infrastructure management service that Google Cloud Platform (GCP) provides. It enables users to automate the creation and management of cloud resources using declarative configurations called deployment templates. With Deployment Manager, users can define their infrastructure as code, specifying the desired state of their resources, such as virtual machines, storage buckets, and networking components. These templates can be version-controlled, allowing for reproducible deployments and easy collaboration. Deployment Manager provides a consistent and reliable way to manage infrastructure, reducing manual errors and streamlining the deployment process. It integrates seamlessly with other GCP services, enabling users to create and manage complex environments efficiently.

GCP BigQuery
Google BigQuery is a fully managed data warehouse and analytics platform that Google Cloud Platform (GCP) offers. It allows users to analyze large datasets quickly and efficiently using a serverless and scalable architecture. BigQuery supports standard SQL queries and provides powerful features like automatic scaling, real-time data ingestion, and built-in machine learning capabilities. With its high-performance columnar storage and distributed query processing, BigQuery can easily handle massive volumes of data. It integrates seamlessly with other GCP services, making it easy to ingest data from various sources and share insights with stakeholders. BigQuery empowers organizations to derive valuable insights and make data-driven decisions at scale.

Key Learning Takeaways:
- Understanding the Project Overview and Architecture
- Understanding the Equity Dataset
- Setting up SQL server on AWS RDS for simulation
- Importing Data into SQL Server
- Introduction to GCP IaC using Deployment Manager
- Using Deployment Manager to spin-up GCP resources
- Installing NiFi on Compute Engine VM
- Setting up ETL pipeline using NiFi Processors
- Loading data into GCP Cloud Storage
- Deploying a trigger on Cloud Functions

- Dataproc automation for transformation using Spark
- Loading transformed data into BigQuery
- Perform stock data analysis on BigQuery to derive insights
- Using Cloud Shell to terminate the resources

Architecture