Automation of Apache Spark Deployment with Ansible and Terraform on GCP

Overview

This project focuses on automating the deployment of an Apache Spark cluster using Terraform and Ansible on Google Cloud Platform (GCP). Terraform will handle the provisioning of infrastructure such as virtual machines, networking, and security configurations. Ansible will be used to install, configure, and manage Apache Spark across the cluster. The goal is to create a fully automated, repeatable, and secure big data environment.

Description

Terraform will define and deploy several Compute Engine instances within a custom VPC. These instances will include a Spark master node, multiple worker nodes, and an edge node for job submission. Optional components such as a storage node can be added for enhanced functionality. The infrastructure will be defined as code, enabling easy scaling, modification, and teardown.

Ansible playbooks will automate the configuration of Spark and related software. Each instance will be set up with necessary dependencies such as Java and system tools. Spark services will be configured to start automatically, forming a functional distributed cluster. Security best practices, including SSH key-only authentication and IAM-based access control will be implemented to ensure a secure setup.

To validate the deployment, the WordCount application will be executed on the cluster. This test will confirm that Spark is operating correctly and efficiently. Performance will be measured by running WordCount with varying numbers of executors. The results will be documented in the final report.

Deliverables

The final deliverables will include a Git repository containing Terraform and Ansible configurations, documentation, and the WordCount test application. A three-page report will summarize the architecture, methodology, testing results, and conclusions. A live demo will also be conducted to demonstrate the automation pipeline, deployment process, and Spark job execution.