LearnKartS

A Training Services Company

PROBLEM STATEMENT

BIG DATA ANALYTICS FOR ANALYZING LARGE DATA SETS

Scenario:

An organization needs a tool for big data analytics for analyzing its large data set. Before the organization decides to buy, you have been asked to do a POC and demo the COE through your screenshots. They have planned to go with Cloud Managed servers in AWS and a Platform as a service from the cloud provider. So the company does not have to work on fixing the servers or applications when they are down and licensing as needed.

Requirements:

1. You are to set it up from scratch.
2. Create a Key new keypair for logging into the servers.
3. Create a Security Group that provides access to SSH into the server.
4. Create an S3 bucket and a folder in it and integrate it with the EMR it can be used as the Data Store for EMR.
5. Use the Default VPC
6. Create logging for the EMR jobs
7. Create a Table and load data in it and have it stored in the S3 data store.
8. Query the Data in EMR
9. Check the EMR cluster performance and resource usage through the URL
10. Check each application usage through the console
11. Protect the EMR from being viewed by other users.