Exploring Hadoop and Spark Ecosystem

**Objective:**

* To understand Hadoop and Spark components and their usage. Also to get an idea of accessing the required components in the GCP platform.

**Pre-requisites:**

* GCP account created with an active GMAIL/Google account.
* A project, a storage bucket, folders, data files associated with the same.
* Hadoop and Spark cluster associated with the same and turned on (Running stage).

**Note:**

* We have created the above pre-requisites in the previous homework.

**Step 1: Login to GCP console and start the cluster(all the three nodes) if it is not running already.**

* If the cluster is running, it should show like the below screenshot.

**A screenshot of a computer

Description automatically generated**

**Step 2: Explore Hadoop and Spark**

* Select “SSH” and click on “Open in new browser window”.
* Click on “Authorize”. The terminal is now open in a new window.

A screenshot of a computer program

Description automatically generated

* To verify user and present working directory type the below commands in the terminal
  + “whoami”
  + “pwd”

A screenshot of a computer program

Description automatically generated

* Type the below command to see all the processes running on the hadoop cluster
  + “ps -ef | grep -i Hadoop”
* The below screenshot shows the running processes

A black and white screen

Description automatically generated

* There is a setting option in the top right corner of the terminal. Enable scrollbar to see all the processes.

A black and white screen shot

Description automatically generated

**Step 3: Identifying and describing Hadoop Components**

* In the above two screenshots we can see the below listed Hadoop components
  + Hive:
    - process 1003 is running the HiveServer2
    - process 1005 is running the HiveMetaStore
    - **Description:** Hive is used to query data from the HDFS in hadoop ecosystem
  + Mapreduce shown as “mapred”
    - Process 1286 is running the JobHistoryServer
    - **Description:** Mapreduce enables the execution across the distributed system of hadoop
  + Hdfs
    - Process 1292 is running the Name Node
    - Process 1290 is running the Secondary Name Node
    - **Description:**
      * Hdfs stands for Hadoop Distributed File System
      * HDFS has two components Name Node(Master, stores matadata) and Data Node(Slave, stores the actual data)
  + Yarn
    - Process 1310 is running the Application History Server
    - Process 1312 is running the Resource Manager
    - **Description:**
      * Yarn stands for yet another resource manager.
      * It is the resource management unit for Hadoop ecosystem.
      * Yarn has two components, Resource Manager(Master, allocates the resources) and Node Manager(Slave, reports the resource utilization)
  + Root
    - Process 1328 is running the startup script.
    - **Description:** This script starts the program
  + Spark
    - Process 2064 is running the History Server
    - **Description:** Spark is a framework for faster processing of big data.

**Step 4: Stop the cluster**

* Once we observe the Hadoop components, it is time to stop the cluster.
* The below screenshot shows the cluster is now stopped.

A screenshot of a computer

Description automatically generated