




Assignment 3

# PROBLEM1-TASK1

Manraj Singh

B00877934



## Problem-1

### Task-1

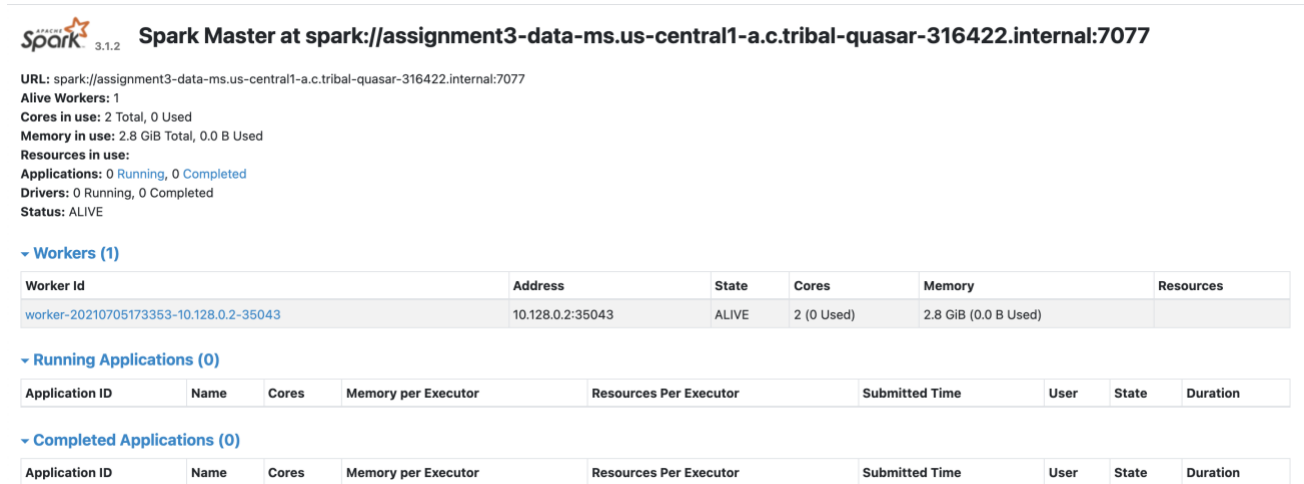
- I created a new compute engine instance on GCP(VM) [2]
- For configuration I used Ubuntu 20.04 LTS and BootDisk.
- For connection I allowed all traffic under firewall (HTTP, HTTPs)
- Updated VPC network for HTTP and HTTPs and allowed all protocols and ports.
- Further configuration was done on browser cloud shell.
- Installed Scala, java and GIT for Apache Spark. [4]
  - sudo apt-get update
  - sudo apt install default-jdk -y
  - sudo apt install scala -y
  - Sudo apt install git -y
- Installed Apache Spark using below steps: [4]
  - Created directory for Apache-Spark in home.
    - i. mkdir Apache\_Spark
    - ii. cd Apache\_Spark
  - Downloaded and installed latest version of Apache-Spark
    - i. wget <https://downloads.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz> [4]
  - extracted the zip and moved to spark folder
    - i. tar xvf spark-3.1.2-bin-hadoop3.2.tgz
    - ii. sudo mv spark-3.1.2-bin-hadoop3.2 /opt/spark
- Configured path for Apache Spark inside ~/.profile file:
  - sudo vi ~/.profile
  - added below paths as shown in [figure 1](#):

```
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
export PYSPARK_PYTHON=/usr/bin/python3
```

*[Figure1 : Paths added to the ~/.profile file \[2\]](#)*

- Saved the file.
- Updated the source file:
  - Source ~/.profile
- Started Master Node:
  - Start-master.sh
  - Logs generated in /opt/spark/logs/spark-smanraj54-org.apache.spark.deploy.master.Master-1-assignment3-data-ms.out

- Master URL : spark://assignment3-data-ms.us-central1-a.c.tribal-quasar-316422.internal:7077
- Started Worker Node:
  - Start-worker.sh <Master URL(mentioned above in start master)>
  - Logs generated in /opt/spark/logs/spark-smanraj54-org.apache.spark.deploy.worker.Worker-1-assignment3-data-ms.out
- VM Dashboard : [1]



*Figure 2: VM Dashboard:[1]*

## References:

- [1] 35.226.207.33. 2021. *Spark Master at spark://assignment3-data-ms.us-central1-a.c.tribal-quasar-316422.internal:7077*. [online] Available at: <<http://35.226.207.33:8080/>> [Accessed 5 July 2021].
- [2] Ca.bbcollab.com. 2021. *Bb Collaborate-LAB 6*. [online] Available at: <[https://ca.bbcollab.com/collab/ui/session/playback/load/d0d7c403df7f4f4f810901c3ddac8fff?authToken=eyJhbGciOiJIUzI1NiJ9.eyJzdWIiOiJiYkNvbGxhYkFwaSIsInJlY29yZGlhZlVpZCI6ImQwZDdjNDZGY3ZjRmNGY4MTA5MDFjM2RkYWY4ZmZmIiwiaXNzIjoibmJDb2xsYWJBcGkiLCJ0eXBIIjoibG9jaXNzIiwiaWF0IjE2MDc3ODksImh0dCI6MTYyNTUwNzQ4OSwiY29uc3VtZXIiOiJGQkZEREZDMEZENEQxQTQxQjQ2QTQxOTICRjU2QjQ2NCJ9.NMd7-iFPdZXSzGBgkzU-0K\\_bnnXEKxyiaH-aYK6PThc](https://ca.bbcollab.com/collab/ui/session/playback/load/d0d7c403df7f4f4f810901c3ddac8fff?authToken=eyJhbGciOiJIUzI1NiJ9.eyJzdWIiOiJiYkNvbGxhYkFwaSIsInJlY29yZGlhZlVpZCI6ImQwZDdjNDZGY3ZjRmNGY4MTA5MDFjM2RkYWY4ZmZmIiwiaXNzIjoibmJDb2xsYWJBcGkiLCJ0eXBIIjoibG9jaXNzIiwiaWF0IjE2MDc3ODksImh0dCI6MTYyNTUwNzQ4OSwiY29uc3VtZXIiOiJGQkZEREZDMEZENEQxQTQxQjQ2QTQxOTICRjU2QjQ2NCJ9.NMd7-iFPdZXSzGBgkzU-0K_bnnXEKxyiaH-aYK6PThc)> [Accessed 5 July 2021].
- [3] Console.cloud.google.com. 2021. *Google Cloud Platform*. [online] Available at: <[https://console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/assignment3-data-ms?project=tribal-quasar-316422&rif\\_reserved](https://console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/assignment3-data-ms?project=tribal-quasar-316422&rif_reserved)> [Accessed 5 July 2021].
- [4] Downloads.apache.org. 2021. [online] Available at: <<https://downloads.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz>> [Accessed 5 July 2021].