Performance of the program

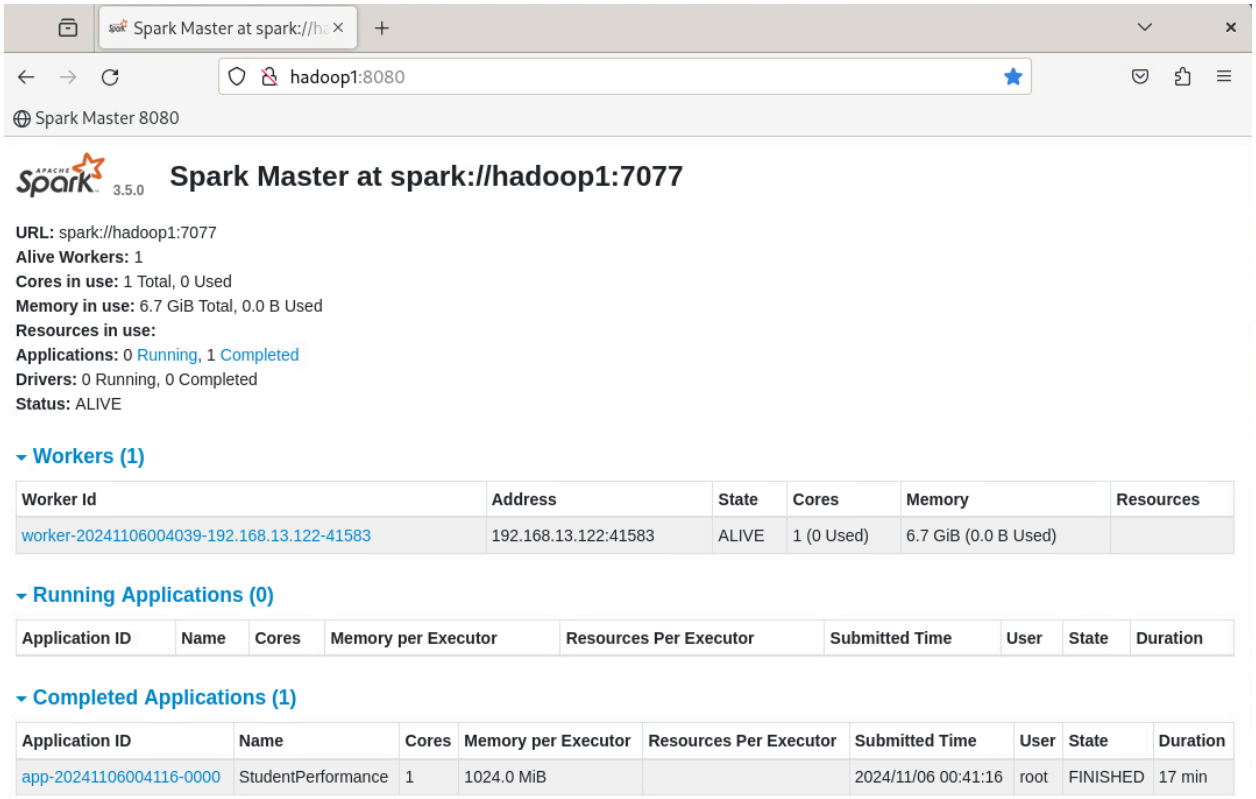
1. Submit a spark job for only one vm (hadoop1)

I started the master VM using command- /opt/spark/sbin/start-master.sh
Then I started the worker VM (only hadoop 1) using this command
/opt/spark/sbin/start-worker.sh spark://hadoop1:7077

```
ot@hadoop1 sat3812]# /opt/spark/sbin/start-master.sh
rting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark
t3812-org.apache.spark.deploy.master.Master-1-hadoop1.out
ot@hadoop1 sat3812]# jps
5 Master
4 Jps
ot@hadoop1 sat3812]# /opt/spark/sbin/start-worker.sh spark://hadoop1:7077
rting org.apache.spark.deploy.worker.Worker, logging to /opt/spark/logs/spark
t3812-org.apache.spark.deploy.worker.Worker-1-hadoop1.out
ot@hadoop1 sat3812]# jps
5 Master
4 Jps
0 Worker
```

Now I am running my python code using only on this vm(hadoop1) using command

opt/spark/bin/spark-submit --master spark://hadoop1:7077 /opt/ml.py



I observed that the duration for completing the action using only 1 Vm was approximately 17 minutes.

Now Let's check how this duration will be changed if we use 2 workers (VMs).

2. Submit a spark job for two Vms (hadoop1 and hadoop2)

I started all (master and worker) Vms using command - /opt/spark/sbin/start-all.sh

```
ot@hadoop1 ~]# /opt/spark/sbin/start-all.sh
rting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark
t3812-org.apache.spark.deploy.master.Master-1-hadoop1.out
.168.13.123: starting org.apache.spark.deploy.worker.Worker, logging to /opt/
rk/logs/spark-root-org.apache.spark.deploy.worker.Worker-1-hadoop2.out
.168.13.122: starting org.apache.spark.deploy.worker.Worker, logging to /opt/
rk/logs/spark-root-org.apache.spark.deploy.worker.Worker-1-hadoop1.out
ot@hadoop1 ~]# jps
6 Master
5 Jps
4 Worker
ot@hadoop1 ~]#
```

I also checked for my second vm (hadoop2) using jps, and it shows that it started as a worker.

```
ot@hadoop2 sat3812]# jps
2 Worker
06 Jps
ot@hadoop2 sat3812]#
```

Here, I am considering 2 Vms (2 workers - hadoop1 and hadoop2)

The duration for completing a specific action using two VMs was about 15 minutes. This is 2 minutes earlier than the time taken when using only one VM, which was around 17. Running the Python code using both VMs (Hadoop 1 and Hadoop 2) demonstrates a significant improvement in processing speed.