

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
Starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark
sat3812-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
Starting org.apache.spark.deploy.worker.Worker, logging to /opt/spark/logs/spark
sat3812-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`

Spark Master at spark://hadoop1:8080

Spark Master 8080

Spark Master at spark://hadoop1:7077

URL: spark://hadoop1:7077
Alive Workers: 1
Cores in use: 1 Total, 0 Used
Memory in use: 6.7 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▼ Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241106004039-192.168.13.122-41583	192.168.13.122:41583	ALIVE	1 (0 Used)	6.7 GiB (0.0 B Used)	

▼ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

▼ Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241106004116-0000	StudentPerformance	1	1024.0 MiB		2024/11/06 00:41:16	root	FINISHED	17 min

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
Starting 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.

```
root@hadoop2 sat3812]# jps
32 Worker
06 Jps
root@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.

Spark Master at spark://hadoop1:8080

hadoop1:8080

Spark Master 8080

Spark 3.5.0

Spark Master at spark://hadoop1:7077

URL: spark://hadoop1:7077

Alive Workers: 2

Cores in use: 2 Total, 0 Used

Memory in use: 13.5 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 1 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241106041542-192.168.13.123-33647	192.168.13.123:33647	ALIVE	1 (0 Used)	6.7 GiB (0.0 B Used)	
worker-20241106041546-192.168.13.122-36369	192.168.13.122:36369	ALIVE	1 (0 Used)	6.7 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241106041754-0000	StudentPerformance	2	1024.0 MiB		2024/11/06 04:17:54	root	FINISHED	15 min