**ReadMe:**

-mapper.py and reducer.py are mapper and reducer files

-Using AWS instance

Login command for EC2 instance:

ssh -i hadoopcluster.pem [ubuntu@ec2-52-27-143-160.us-west-2.compute.amazonaws.com](mailto:ubuntu@ec2-52-26-15-146.us-west-2.compute.amazonaws.com)

To put file on hdf from EC2 instance:

Hadoop fs –put stations stations

**Note**: I have sent different sample output files in attachment (e.g., m1r1 🡪 for 1 mapper and 1 reducer ; m2r2\_3a, m2r2\_3b 🡪 2 mappers and 2 reducers for instance 3)

**Various cases:**

For all commands below m1r1 means mapper 1 and reducer 1, same as m10r2 should be mapper 10 & reducer 2

1) Mapper: 1, Reducer: 1

Command:

hadoop jar hadoop-streaming-2.6.0.jar -D mapreduce.job.maps=1 -D mapreduce.job.reduces=1 -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input stations -output m1r1

**Time taken:**

Instance 1: 50 sec

Instance 2: 45 sec

Instance 3: 36 sec

2) Mapper: 2, Reducer: 1

Command:

hadoop jar hadoop-streaming-2.6.0.jar -D mapreduce.job.maps=2 -D mapreduce.job.reduces=1 -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input stations -output m2r1

**Time taken:**

Instance 1: 50 sec

Instance 2: 46 sec

Instance 3: 33 sec

3) Mapper: 2, Reducer: 2

Command:

hadoop jar hadoop-streaming-2.6.0.jar -D mapreduce.job.maps=2 -D mapreduce.job.reduces=2 -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input stations -output m2r2

**Time taken:**

Instance 1: 43 sec

Instance 2: 45 sec

Instance 3: 31 sec

4) Mapper: 10, Reducer: 1

Command:

hadoop jar hadoop-streaming-2.6.0.jar -D mapreduce.job.maps=10 -D mapreduce.job.reduces=1 -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input stations -output m10r1

**Time taken:**

Instance 1: 1 min, 1 sec

Instance 2: 1 min, 19 sec

Instance 3: 1 min, 13 sec

5) Mapper: 10, Reducer: 2

Command:

hadoop jar hadoop-streaming-2.6.0.jar -D mapreduce.job.maps=10 -D mapreduce.job.reduces=2 -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input stations -output m10r2

**Time taken:**

Instance 1: 1 min, 6 sec

Instance 2: 1 min, 27 sec

Instance 3: 1 min, 11 sec

**Proof of some of the outputs of timing shown above:**

**Instance 1: 1 mapper; 1 reducer**

****

**Instance 2: 1 mapper; 1 reducer**

****

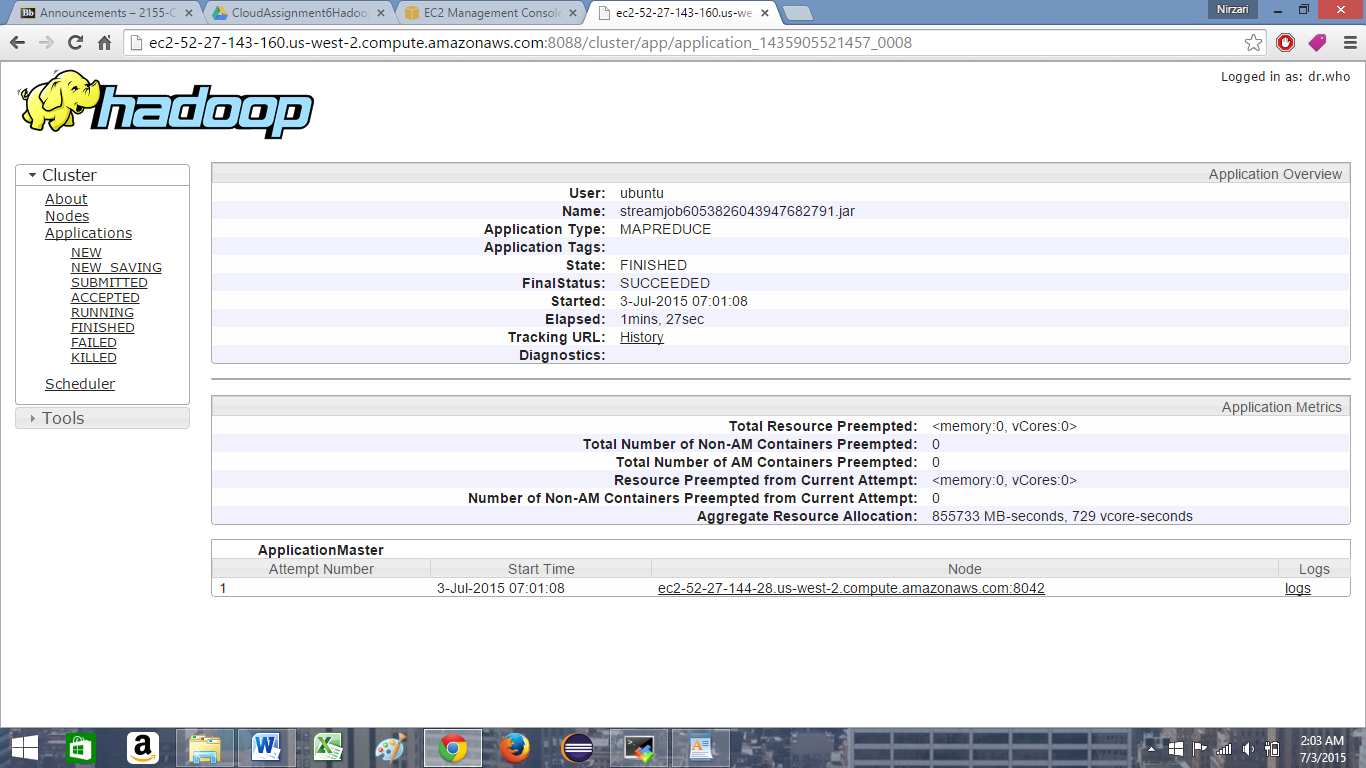
**Instance 3: 1 mapper; 1 reducer**

****

**Instance 1: 10 mapper; 2 reducer**

****

**Instance 2: 10 mapper; 2 reducer**

****

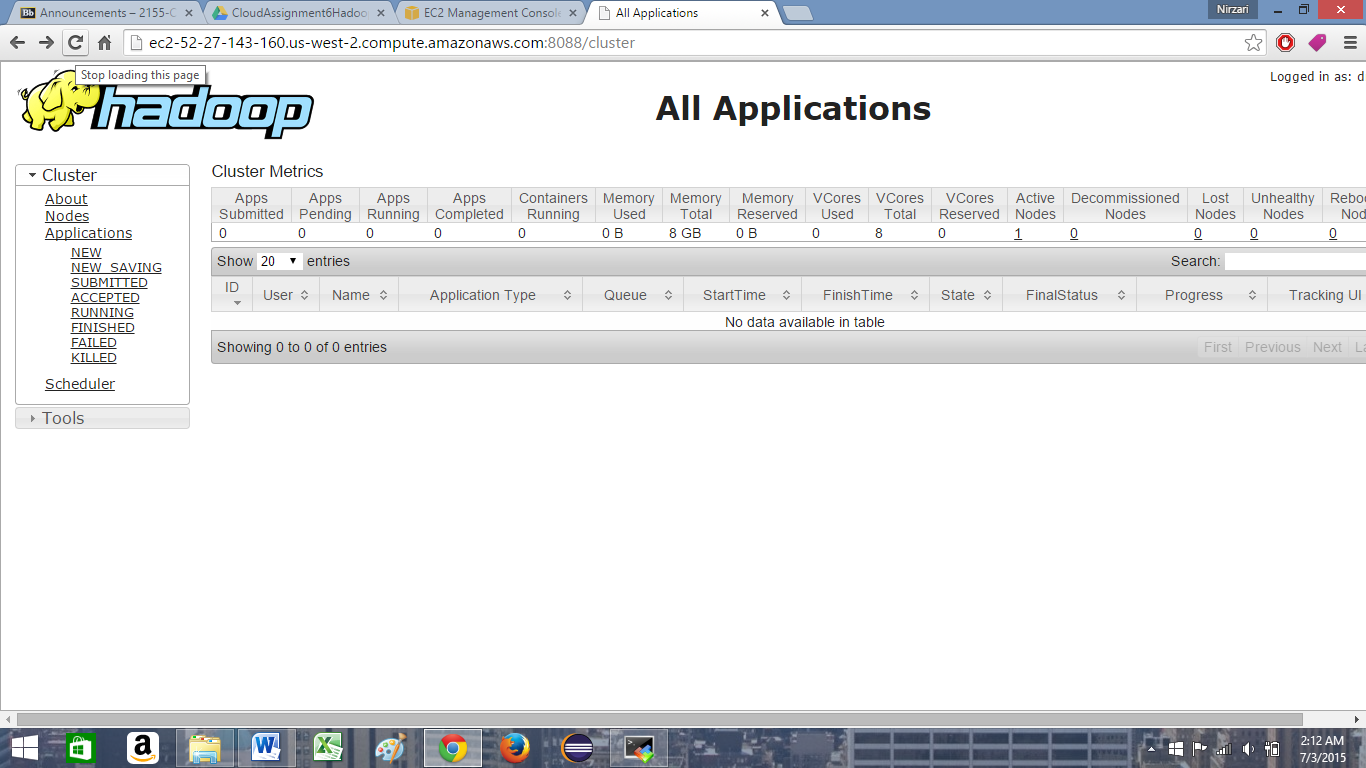
**Instance 3: 10 mapper; 2 reducer**

****

***Proof of using 3 instances:***

Here number of active nodes show number of instances used.

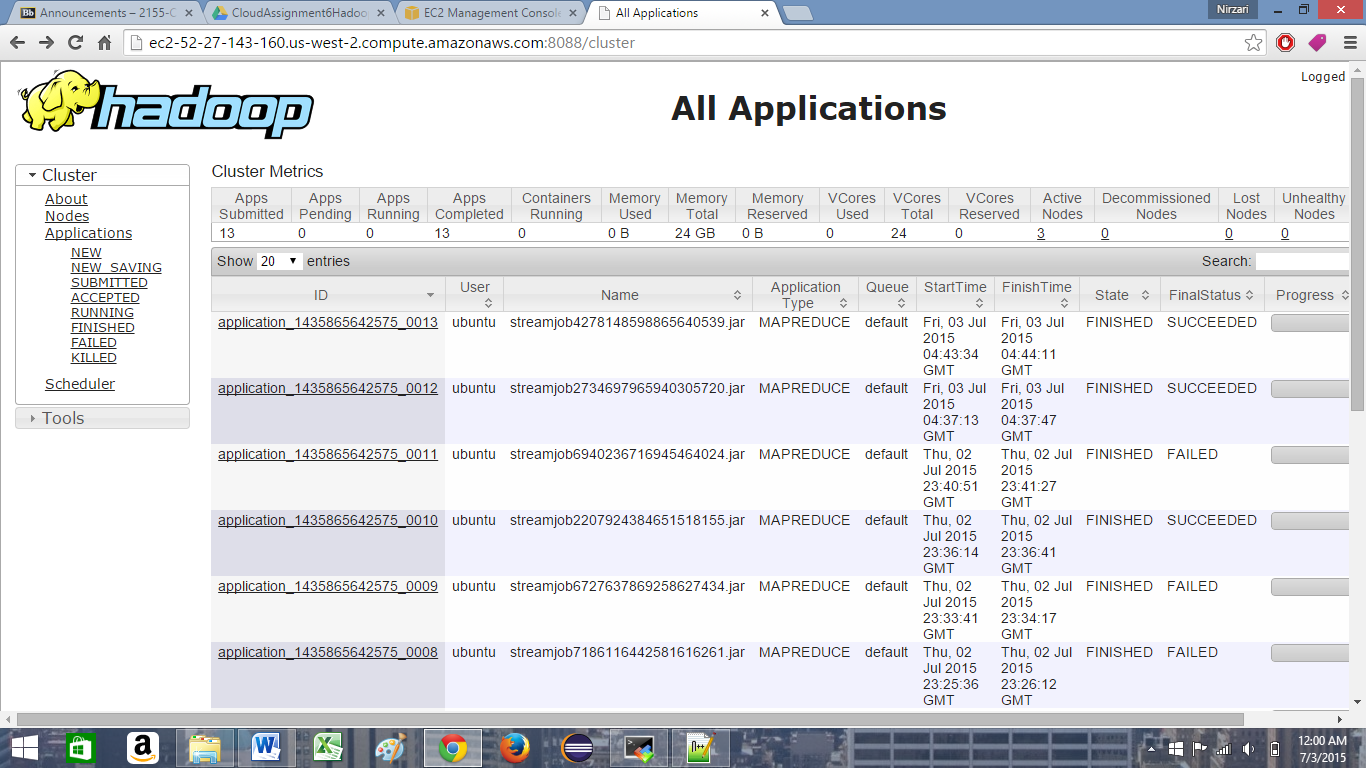
**Instance1:**



**Instance 2:**

****

**Instance 3:**

****