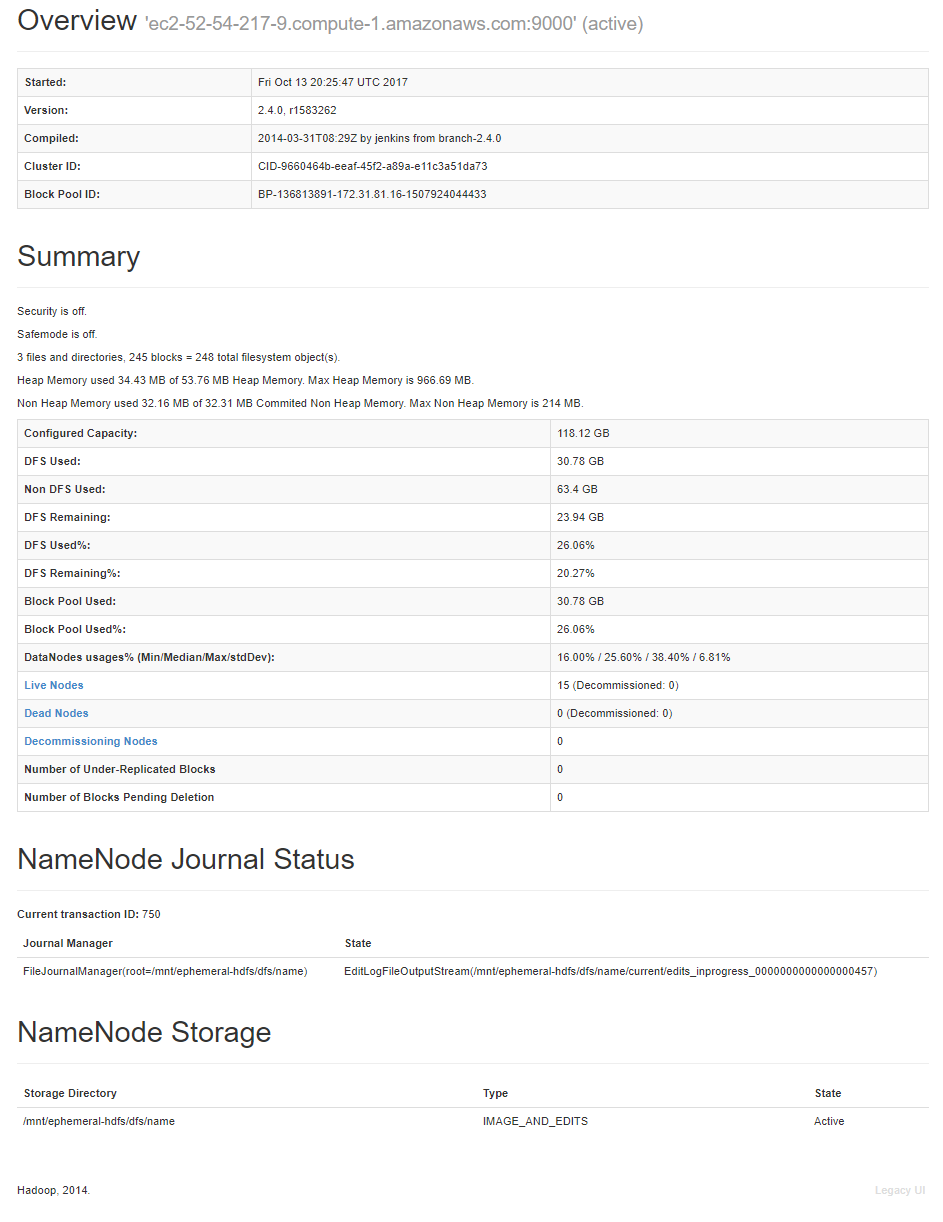
Assignment2 Report

1.launch spark cluster using spark-ec2

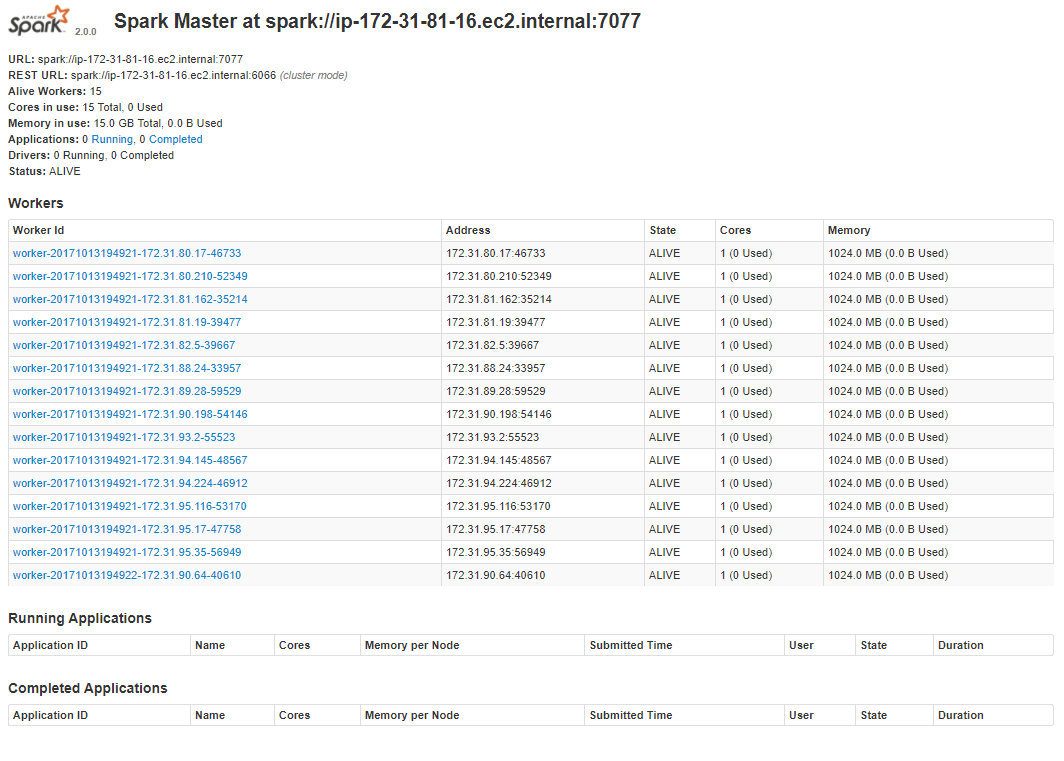
EC2 Instance I chose is t2.micro(1 core, 1G memory). Cluster contains 15 slaves

./spark-ec2 -i pageRank -k pageRank.pem -s 15 launch PageRank

Hadoop page



Spark Page



2. Put data onto HDFS

sudo ./hadoop fs -put ~/data/rawd/articles.tsv

3. Run java code

sudo /root/spark/bin/spark-submit --class pageRank.PageRank ~/data/jar/source.jar /data/articles.tsv

4. Run

sudo /root/spark/bin/spark-submit --class pageRank.PageRank ~/data/jar/GraphX.jar /data/articles.tsv

4. Result

Top 100 pages (5 iteration) : see Top100(5).txt

Top 100 pages (10 iteration) : see Top100(10).txt

Top 100 university : see Top100-University.txt

Top 100 alumni: see Top100-alumni.txt

5.

I choose 2 iteration times, which is 5 and 10. The result of two is slightly different, which means after 5 iterations, the weights of each pages are about to converge. But more iterations can still make result more accurate.

Using GraphX can greatly fast the process. The time to run 10 iteration pageRank takes about 5 min. However, it takes about 24 min for origin map-Reduce to finish the job.