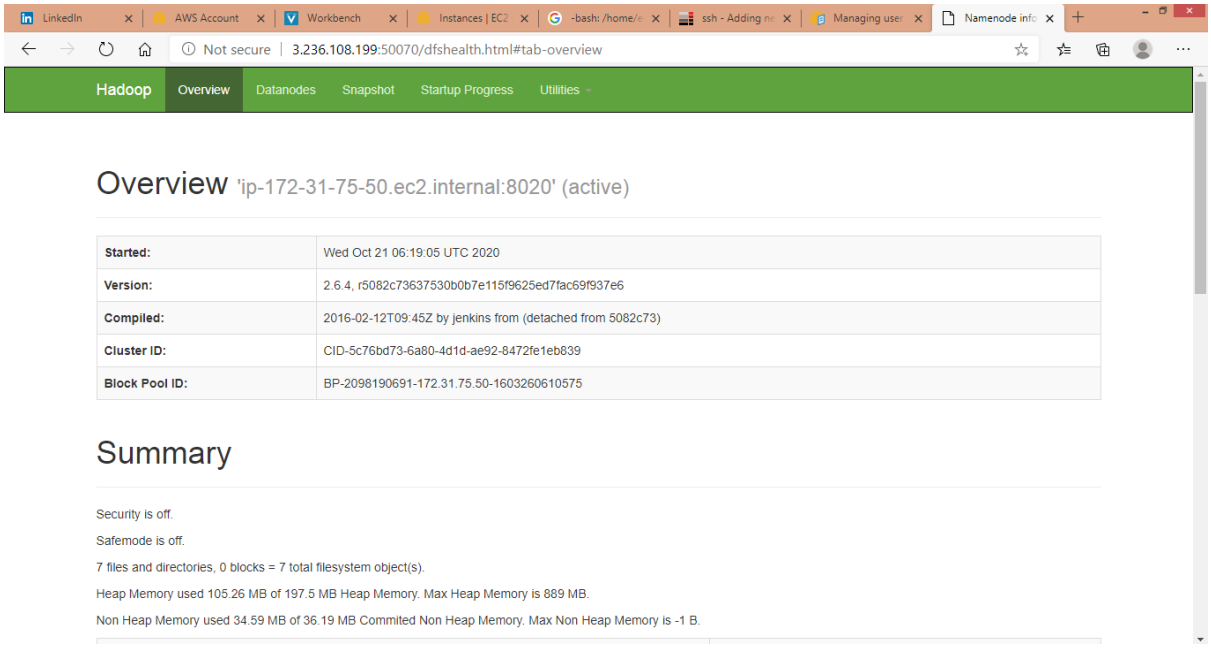
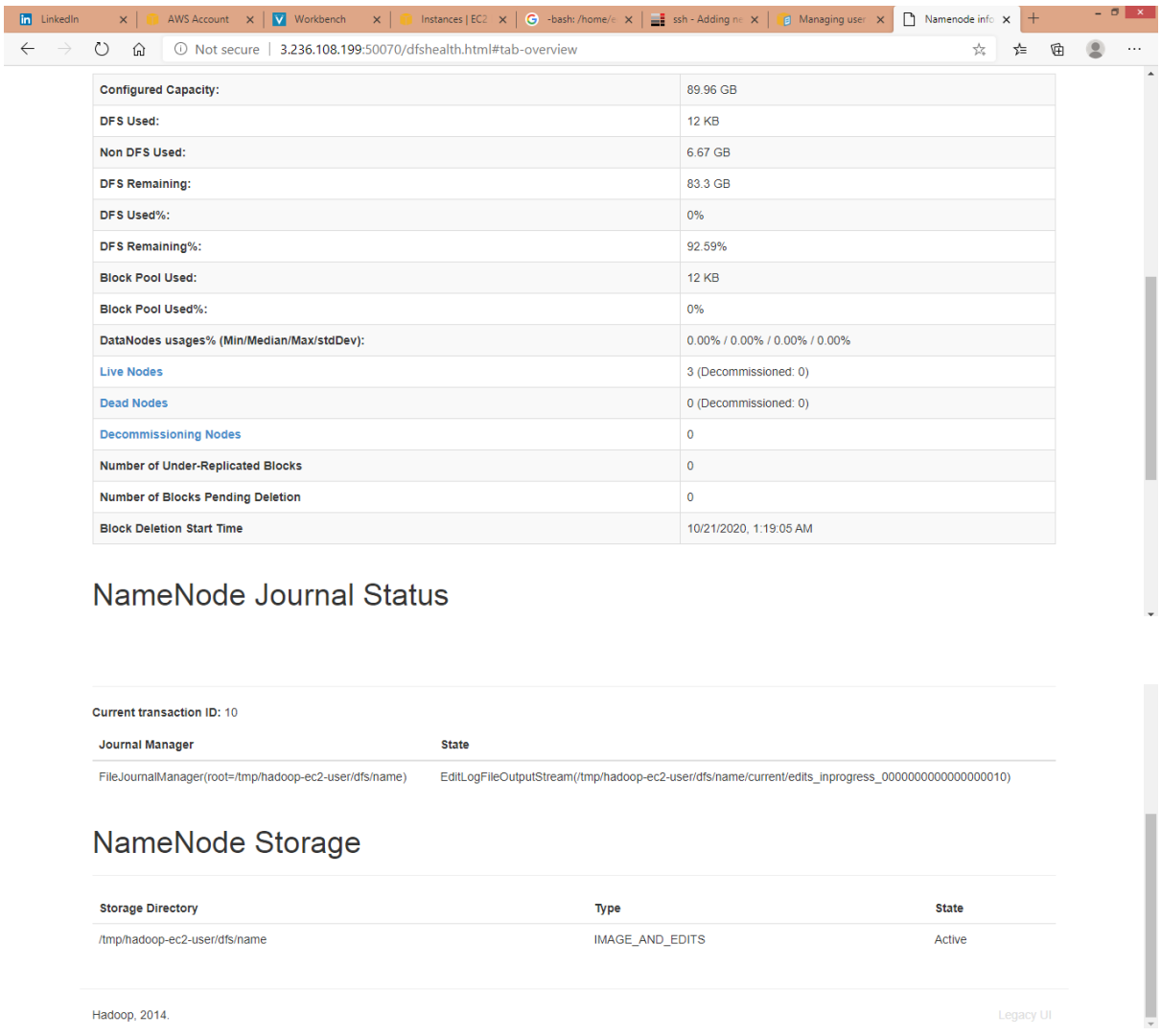
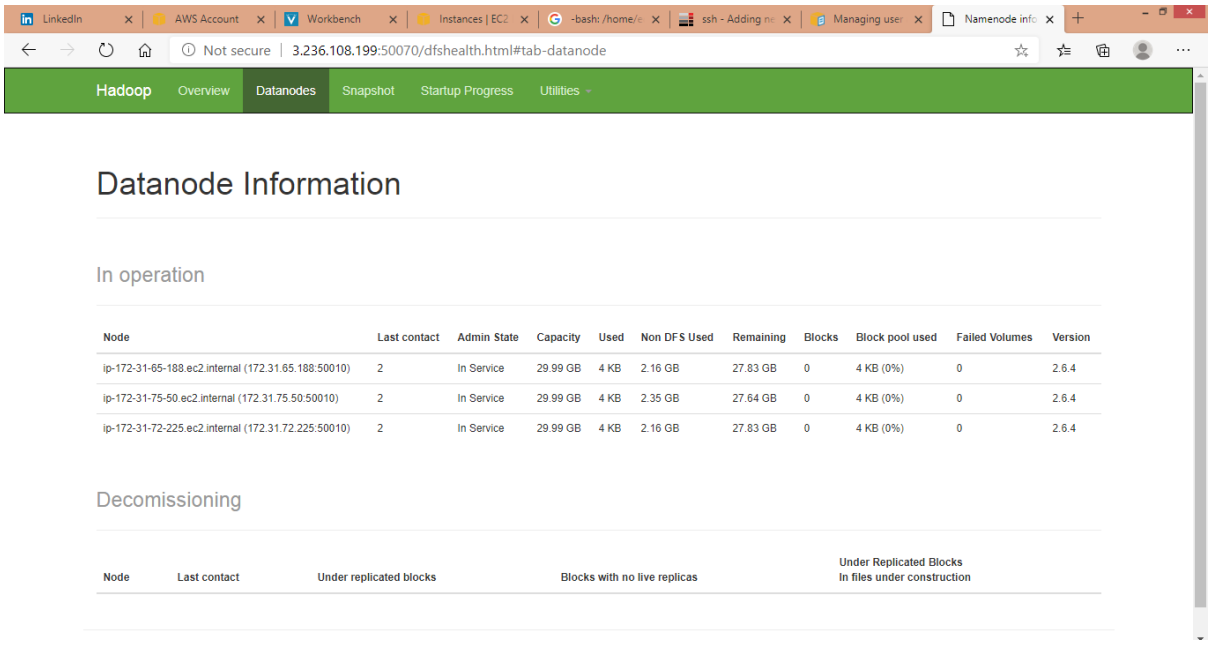
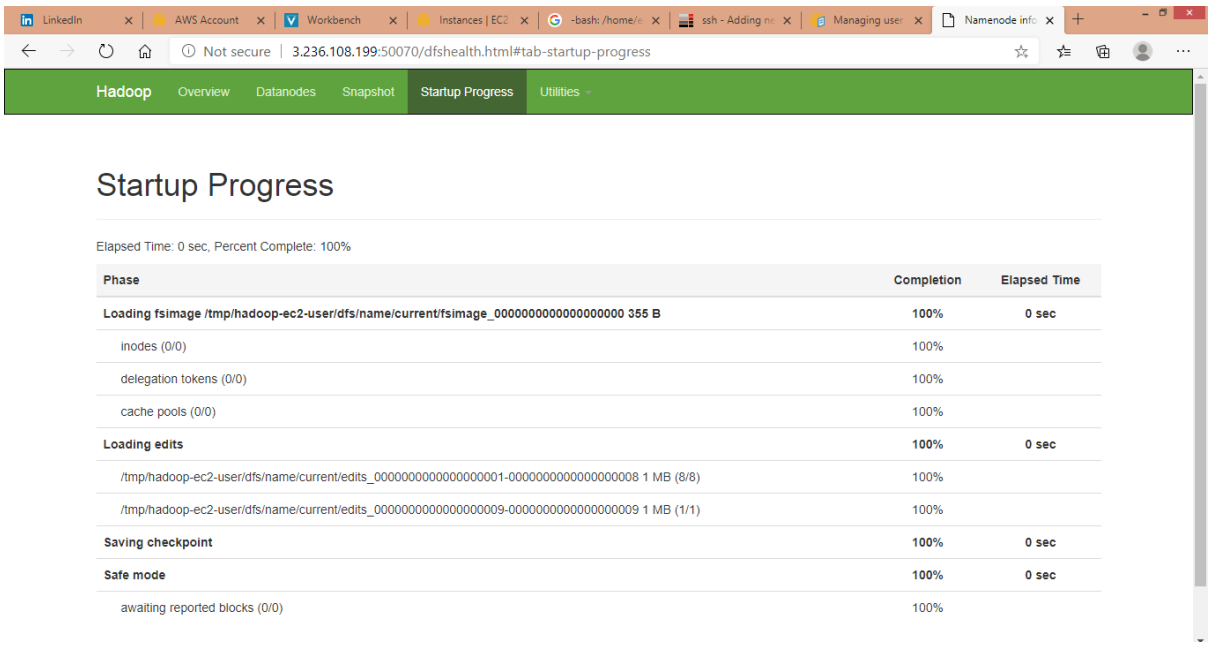
Multi node Hadoop cluster









Single node cluster runtime:

```
real    1m13.478s
user    0m4.123s
sys     0m0.161s
```

Multi node cluster runtime:

```
                Total time spent by all reduces in occupied slots (ms)=7387
                Total time spent by all map tasks (ms)=51442
                Total time spent by all reduce tasks (ms)=7387
                Total vcore-milliseconds taken by all map tasks=51442
                Total vcore-milliseconds taken by all reduce tasks=7387
                Total megabyte-milliseconds taken by all map tasks=52676608
                Total megabyte-milliseconds taken by all reduce tasks=7564288
        Map-Reduce Framework
                Map input records=5284546
                Map output records=18562366
                Map output bytes=279356680
                Map output materialized bytes=26902454
                Input split bytes=208
                Combine input records=20053191
                Combine output records=2673165
                Reduce input groups=1040390
                Reduce shuffle bytes=26902454
                Reduce input records=1182340
                Reduce output records=1040390
                Spilled Records=3855505
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=752
                CPU time spent (ms)=44400
                Physical memory (bytes) snapshot=768716800
                Virtual memory (bytes) snapshot=6405160960
                Total committed heap usage (bytes)=526385152
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=231153099
        File Output Format Counters
                Bytes Written=20056175

real    0m43.580s
user    0m4.026s
sys     0m0.265s
```

```
[ec2-user@ip-172-31-75-50 ~]$ hadoop fs -du /data/wordcount1/
0         /data/wordcount1/_SUCCESS
20056175  /data/wordcount1/part-r-00000
```

```
[ec2-user@ip-172-31-75-50 ~]$ hadoop fs -cat /data/wordcount1/part-r-00000 | grep arctic
&lt;I&gt;holarctica&lt;/I&gt;   28
&lt;I&gt;holarctica&lt;/I&gt;&lt;/B&gt;.        8
&lt;I&gt;holarctica&lt;/I&gt;,  1
&lt;I&gt;palearctica&lt;/I&gt;  4
&lt;i&gt;holarctica&lt;/i&gt;   1
(Antarctic      3
(Antarctica)    1
(Antarctica),   11
<Label>Antarctic        1
<Name>Antarctic 3
<Name>Antarctica        1
<Strain>Antarctic       1
<Title>Antarctic        5
Antarctic       137
Antarctic,      1
Antarctic.      2
Antarctic.</Description>        1
Antarctic.</Title>      1
Antarctic</Title>       4
Antarctica      16
Antarctica)</Title>     1
Antarctica,     9
Antarctica.     24
Antarctica.&#x0D;       3
Antarctica.</Description>       19
Antarctica</Description>        2
Antarctica</Name>       1
Antarctica</Title>      6
Palearctic      1
Project">Antarctic      1
Subarctic       11
abbr="Antarctic 1
antarctic       5
antarctica      17
antarctica&lt;/i&gt;&lt;/b&gt;.&#x0D;   2
antarctica,     4
antarctica</Name>       10
antarctica</OrganismName>       11
antarctica</Title>      1
antarcticum     32
antarcticum</Name>      3
```

```
antarcticum</OrganismName>      3
antarcticus     31
antarcticus&lt;/i&gt;    4
antarcticus&lt;/i&gt;&lt;/b&gt;.         1
antarcticus).   1
antarcticus,    1
antarcticus</Name>      5
antarcticus</OrganismName>      5
arctic  21
arctica 27
arctica&lt;/I&gt;)       2
arctica&lt;/i&gt;        3
arctica&lt;/i&gt;,       1
arctica.</Description>  2
arctica</Name>  5
arctica</OrganismName>  5
arcticus        31
arcticus&lt;/i&gt;       2
arcticus</Name> 4
arcticus</OrganismName> 4
holarctica      77
humans.Antarctic        1
palearctica     66
palearctica</Name>      1
sub-Antarctic   4
sub-arctic      4
subantarctic    1
subantarcticus  7
subantarcticus</Name>   1
subantarcticus</OrganismName>   1
subarctic       21
```

The single node cluster for word count problem in assignment 1 takes 1 minute 14 seconds to complete while the 3 node cluster completes within ~46 seconds.
Since the MapReduce processes are split across 3 clusters and blocks are read simultaneously, the process tends to run faster. The process completes faster by ~ 28 seconds as compared to 1 node cluster.