

Multi node Hadoop cluster

Overview

'ip-172-31-75-50.ec2.internal:8020' (active)

Started:	Wed Oct 21 06:19:05 UTC 2020
Version:	2.6.4, r5082c73637530b0b7e115f9625ed7fac69f937e6
Compiled:	2016-02-12T09:45Z by jenkins from (detached from 5082c73)
Cluster ID:	CID-5c76bd73-6a80-4d1d-ae92-8472fe1eb839
Block Pool ID:	BP-2098190691-172.31.75.50-1603260610575

Summary

Security is off.

Safemode is off.

7 files and directories, 0 blocks = 7 total filesystem object(s).

Heap Memory used 105.26 MB of 197.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 34.59 MB of 36.19 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	89.96 GB
DFS Used:	12 KB
Non DFS Used:	6.67 GB
DFS Remaining:	83.3 GB
DFS Used%:	0%
DFS Remaining%:	92.59%
Block Pool Used:	12 KB
Block Pool Used%:	0%
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	3 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	10/21/2020, 1:19:05 AM

NameNode Journal Status

Current transaction ID: 10

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-ec2-user/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-ec2-user/dfs/name/current/edits_inprogress_000000000000000010)

NameNode Storage

Storage Directory	Type	State
/tmp/hadoop-ec2-user/dfs/name	IMAGE_AND_EDITS	Active

Hadoop, 2014.

Legacy UI

[Hadoop](#)
[Overview](#)
[Datanodes](#)
[Snapshot](#)
[Startup Progress](#)
[Utilities](#)

Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
ip-172-31-65-188.ec2.internal (172.31.65.188:50010)	2	In Service	29.99 GB	4 KB	2.16 GB	27.83 GB	0	4 KB (0%)	0	2.8.4
ip-172-31-75-50.ec2.internal (172.31.75.50:50010)	2	In Service	29.99 GB	4 KB	2.35 GB	27.64 GB	0	4 KB (0%)	0	2.8.4
ip-172-31-72-225.ec2.internal (172.31.72.225:50010)	2	In Service	29.99 GB	4 KB	2.16 GB	27.83 GB	0	4 KB (0%)	0	2.8.4

Decommissioning

Node	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction
------	--------------	-------------------------	------------------------------	--

[Hadoop](#)
[Overview](#)
[Datanodes](#)
[Snapshot](#)
[Startup Progress](#)
[Utilities](#)

Startup Progress

Elapsed Time: 0 sec, Percent Complete: 100%

Phase	Completion	Elapsed Time
Loading fsimage /tmp/hadoop-ec2-user/dfs/name/current/fsimage_00000000000000000000 355 B	100%	0 sec
inodes (0/0)	100%	
delegation tokens (0/0)	100%	
cache pools (0/0)	100%	
Loading edits	100%	0 sec
/tmp/hadoop-ec2-user/dfs/name/current/edits_00000000000000000001-00000000000000000008 1 MB (8/8)	100%	
/tmp/hadoop-ec2-user/dfs/name/current/edits_00000000000000000009-00000000000000000009 1 MB (1/1)	100%	
Saving checkpoint	100%	0 sec
Safe mode	100%	0 sec
awaiting reported blocks (0/0)	100%	

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
<I>holarctica</I> 28
<I>holarctica</I><B>. 8
<I>holarctica</I>, 1
<I>palearctica</I> 4
<i>holarctica</i> 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</i></b>.&#x0D; 2
antarctica, 4
antarctica</Name> 10
antarctica</OrganismName> 11
antarctica</Title> 1
antarcticum 32
antarcticum</Name> 3

```

```

antarcticum</OrganismName>      3
antarcticus      31
antarcticus</i>      4
antarcticus</i></b>.      1
antarcticus).      1
antarcticus,      1
antarcticus</Name>      5
antarcticus</OrganismName>      5
arctic      21
arctica      27
arctica</I>)      2
arctica</i>      3
arctica</i>,      1
arctica.</Description>      2
arctica</Name>      5
arctica</OrganismName>      5
arcticus      31
arcticus</i>      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.