

David Beck
CSC555
Project Phase 1
5/19/19

Part 1:

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-172-31-7-237.us-east-2.compute.internal (172.31.7.237:50010)	0	In Service	29.99 GB	4 KB	2.32 GB	27.67 GB	0	4 KB (0%)	0	2.6.4
ip-172-31-3-232.us-east-2.compute.internal (172.31.3.232:50010)	0	In Service	7.99 GB	4 KB	2 GB	5.99 GB	0	4 KB (0%)	0	2.6.4
ip-172-31-7-147.us-east-2.compute.internal (172.31.7.147:50010)	0	In Service	7.99 GB	4 KB	2 GB	5.99 GB	0	4 KB (0%)	0	2.6.4

Decomissioning

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

Hadoop, 2014.

Legacy
UI

```
ec2-user@ip-172-31-7-237:~  
File System Counters  
  FILE: Number of bytes read=59605201  
  FILE: Number of bytes written=86827979  
  FILE: Number of read operations=0  
  FILE: Number of large read operations=0  
  FILE: Number of write operations=0  
  HDFS: Number of bytes read=231153307  
  HDFS: Number of bytes written=20056175  
  HDFS: Number of read operations=9  
  HDFS: Number of large read operations=0  
  HDFS: Number of write operations=2  
Job Counters  
  Launched map tasks=2  
  Launched reduce tasks=1  
  Data-local map tasks=2  
  Total time spent by all maps in occupied slots (ms)=46402  
  Total time spent by all reduces in occupied slots (ms)=6150  
  Total time spent by all map tasks (ms)=46402  
  Total time spent by all reduce tasks (ms)=6150  
  Total vcore-milliseconds taken by all map tasks=46402  
  Total vcore-milliseconds taken by all reduce tasks=6150  
  Total megabyte-milliseconds taken by all map tasks=47515648  
  Total megabyte-milliseconds taken by all reduce tasks=6297600  
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)=553  
  CPU time spent (ms)=40280  
  Physical memory (bytes) snapshot=716718080  
  Virtual memory (bytes) snapshot=6453735424  
  Total committed heap usage (bytes)=503316480  
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    0m40.353s  
user    0m3.823s  
sys     0m0.337s  
[ec2-user@ip-172-31-7-237 ~]$
```

Discussion: The 3-node cluster was measurably faster than the single node at mapping the bioproject.xml data (40 seconds compared to 1 minute 10 seconds). That's a time savings of 30 seconds. The reason the 3-node cluster was able to process this data faster was because the master node could better delegate and divide up the work to the worker nodes and thereby speed up the mapreduce job overall.

Part 2:

Query 1.2 time: 25.096 seconds

```
ec2-user@ip-172-31-7-237:~/apache-hive-2.0.1-bin
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1558293028766_0003, Tracking URL = http://ip-172-31-7-237.us-east-2.compute.internal:8088/proxy/application_1558293028766_0003/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1558293028766_0003
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2019-05-19 19:54:27,214 Stage-2 map = 0%, reduce = 0%
2019-05-19 19:54:33,406 Stage-2 map = 33%, reduce = 0%, Cumulative CPU 2.66 sec
2019-05-19 19:54:38,568 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 13.79 sec
2019-05-19 19:54:39,597 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 15.42 sec
MapReduce Total cumulative CPU time: 15 seconds 420 msec
Ended Job = job_1558293028766_0003
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3 Reduce: 1 Cumulative CPU: 15.42 sec HDFS Read: 594368
452 HDFS Write: 12 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 420 msec
OK
14215822897
Time taken: 25.095 seconds, Fetched: 1 row(s)
hive>
```

Query 1.3 time: 25.796 seconds

ec2-user@ip-172-31-7-237:~/apache-hive-2.0.1-bin

```
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1558293028766_0004, Tracking URL = http://ip-172-31-7-237.us-east-2.compute.internal:8088/proxy/application_1558293028766_0004/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1558293028766_0004
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2019-05-19 19:57:07,594 Stage-2 map = 0%, reduce = 0%
2019-05-19 19:57:15,833 Stage-2 map = 67%, reduce = 0%, Cumulative CPU 8.99 sec
2019-05-19 19:57:17,896 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 14.09 sec
2019-05-19 19:57:20,973 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 15.72 sec
MapReduce Total cumulative CPU time: 15 seconds 720 msec
Ended Job = job_1558293028766_0004
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3 Reduce: 1 Cumulative CPU: 15.72 sec HDFS Read: 594368
557 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 720 msec
OK
4435791464
Time taken: 25.769 seconds, Fetched: 1 row(s)
hive>
```

Query 2.1 time: 104.589 seconds

```
ec2-user@ip-172-31-7-237:~/apache-hive-2.0.1-bin
419415707      1998      MFGR#1226
358466340      1998      MFGR#1227
251549955      1998      MFGR#1228
383138860      1998      MFGR#1229
296330561      1998      MFGR#123
437181243      1998      MFGR#1230
398944492      1998      MFGR#1231
424062455      1998      MFGR#1232
406967188      1998      MFGR#1233
428867240      1998      MFGR#1234
352277781      1998      MFGR#1235
361827086      1998      MFGR#1236
341618569      1998      MFGR#1237
244739231      1998      MFGR#1238
414151803      1998      MFGR#1239
330082371      1998      MFGR#124
415312453      1998      MFGR#1240
360289624      1998      MFGR#125
341657580      1998      MFGR#126
377507061      1998      MFGR#127
361416497      1998      MFGR#128
318769573      1998      MFGR#129
Time taken: 104.589 seconds, Fetched: 280 row(s)
hive>
```

Select transform results:

```
ec2-user@ip-172-31-7-237:~/apache-hive-2.0.1-bin
29979  NtELMSuf      IRAN      #5
29980  307eqWcw      CANADA   #8
29981  063ubxlt      ETHIOPIA #8
29982  H0wQlCf1      JAPAN    #4
29983  RfRDlhMV      IRAQ     #9
29984  lt3KPiAe      ALGERIA  #4
29985  nfApZl D      SAUDI AR #0
29986  9Z 73TOV      ARGENTIN #9
29987  X7 HDnPz      IRAN     #5
29988  rpO5NeDq      JORDAN   #8
29989  B7rtGkAD      UNITED S #8
29990  PTBjVjnj      RUSSIA   #9
29991  jLtJzbqU      MOZAMBIQ #3
29992  ORRP4d7x      IRAQ     #8
29993  tzfH0l2G      INDONESI #4
29994  XyRnbUhh      EGYPT    #0
29995  Bbaooa RUSSIA  #8
29996  qDl5k2Iq      GERMANY  #5
29997  G7Pn24Na      INDONESI #1
29998  2uuIxo x      UNITED K #4
29999  pxbqW7BK      JAPAN    #4
30000  3I5hj95      RUSSIA   #7
Time taken: 0.569 seconds, Fetched: 30000 row(s)
hive>
```

Total Query 0.1-0.3 time: 3 minutes, 30 seconds and 225 milliseconds

[illegible]

Part 4:

```
ec2-user@ip-172-31-7-237:~/hadoop-2.6.4
FILE: Number of bytes read=88660834
FILE: Number of bytes written=183162881
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=594518205
HDFS: Number of bytes written=727
HDFS: Number of read operations=159
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
Job Counters
  Failed reduce tasks=1
  Killed map tasks=1
  Killed reduce tasks=1
  Launched map tasks=50
  Launched reduce tasks=5
  Data-local map tasks=47
  Rack-local map tasks=3
  Total time spent by all maps in occupied slots (ms)=821838
  Total time spent by all reduces in occupied slots (ms)=110474
  Total time spent by all map tasks (ms)=821838
  Total time spent by all reduce tasks (ms)=110474
  Total vcore-milliseconds taken by all map tasks=821838
  Total vcore-milliseconds taken by all reduce tasks=110474
  Total megabyte-milliseconds taken by all map tasks=841562112
  Total megabyte-milliseconds taken by all reduce tasks=113125376
Map-Reduce Framework
  Map input records=6001215
  Map output records=6001215
  Map output bytes=76658386
  Map output materialized bytes=88661716
  Input split bytes=4500
  Combine input records=0
  Combine output records=0
  Reduce input groups=50
  Reduce shuffle bytes=88661716
  Reduce input records=6001215
  Reduce output records=47
  Spilled Records=12002430
  Shuffled Maps =150
  Failed Shuffles=0
  Merged Map outputs=150
  GC time elapsed (ms)=10244
  CPU time spent (ms)=88560
  Physical memory (bytes) snapshot=12378845184
  Virtual memory (bytes) snapshot=113760616448
  Total committed heap usage (bytes)=9403629568
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=594513705
File Output Format Counters
  Bytes Written=727
19/05/20 19:15:21 INFO streaming.StreamJob: Output directory: /data/reduced_line
order
real    1m13.493s
user    0m4.180s
sys     0m0.336s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```