Single node cluster

Word count

Start HDFS, Hadoop and history server

```
[ec2-user@ip-172-31-56-147 ~]$ start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/ec2-user/hadoop-2.6.4/logs/hadoop-ec2-user-namenode-ip-172-31-56-147.ec2.internal.out
localhost: starting datanode, logging to /home/ec2-user/hadoop-2.6.4/logs/hadoop-ec2-user-datanode-ip-172-31-56-147.ec2.internal.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:mVpeeHwOhl0QFARYhYgTl9thepKNoG2OGh1qEGS2BeE.
ECDSA key fingerprint is MD5:34:94:a5:36:e5:4a:15:75:2b:60:75:a0:3a:e3:e1:d0.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/ec2-user/hadoop-2.6.4/logs/hadoop-ec2-user-secondarynamenode-ip-172-31-56-147.ec2.internal.out
[ec2-user@ip-172-31-56-147 ~]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/ec2-user/hadoop-2.6.4/logs/yarn-ec2-user-resourcemanager-ip-172-31-56-147.ec2.internal.out
localhost: starting nodemanager, logging to /home/ec2-user/hadoop-2.6.4/logs/yarn-ec2-user-nodemanager-ip-172-31-56-147.ec2.internal.out
[ec2-user@ip-172-31-56-147 ~]$ mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /home/ec2-user/hadoop-2.6.4/logs/mapred-ec2-user-historyserver-ip-172-31-56-147.ec2.internal.out
[ec2-user@ip-172-31-56-147 ~]$ jps
30387 NameNode
30963 NodeManager
30708 SecondaryNameNode
30533 DataNode
31270 JobHistoryServer
31305 Jps
30842 ResourceManager
[ec2-user@ip-172-31-56-147 ~]$ hadoop fs -mkdir /data
```

Download a large text file

```
[ec2-user@ip-172-31-56-147 ~]$ wget http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/bioproject.xml
--2020-09-20 00:57:13--  http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/bioproject.xml
Resolving rasinsrv07.cstcis.cti.depaul.edu (rasinsrv07.cstcis.cti.depaul.edu)... 140.192.39.95
Connecting to rasinsrv07.cstcis.cti.depaul.edu (rasinsrv07.cstcis.cti.depaul.edu)|140.192.39.95|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 231149003 (220M) [text/xml]
Saving to: 'bioproject.xml'

100%[==========================================================================================================================

2020-09-20 00:57:33 (11.2 MB/s) - 'bioproject.xml' saved [231149003/231149003]
```

Copy the file to HDFS for processing

```
[ec2-user@ip-172-31-56-147 ~]$ hadoop fs -put bioproject.xml /data/
[ec2-user@ip-172-31-56-147 ~]$ hadoop fs -ls /data
Found 1 items
-rw-r--r--   1 ec2-user supergroup  231149003 2020-09-20 00:58 /data/bioproject.xml
```

Job execution time

```
        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)=95257
                Total time spent by all reduces in occupied slots (ms)=12098
                Total time spent by all map tasks (ms)=95257
                Total time spent by all reduce tasks (ms)=12098
                Total vcore-milliseconds taken by all map tasks=95257
                Total vcore-milliseconds taken by all reduce tasks=12098
                Total megabyte-milliseconds taken by all map tasks=97543168
                Total megabyte-milliseconds taken by all reduce tasks=12388352
        Map-Reduce Framework
                Map input records=5284546
                Map output records=18562366
                Map output bytes=279356680
                Map output materialized bytes=26902454
                Input split bytes=202
                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)=982
                CPU time spent (ms)=41510
                Physical memory (bytes) snapshot=562561024
                Virtual memory (bytes) snapshot=6315945984
                Total committed heap usage (bytes)=333602816
        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    1m13.478s
user    0m4.123s
sys     0m0.161s
```

```
[ec2-user@ip-172-31-56-147 ~]$ hadoop fs -du /data/wordcount1/
0          /data/wordcount1/_SUCCESS
20056175   /data/wordcount1/part-r-00000
```

Count of occurrences of “arctic”

```
[ec2-user@ip-172-31-56-147 ~]$ 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
```