INFSCI 2750: Cloud Computing

Big Data Hadoop System: Mini Project 1

Team-
Tazeen Khan - tfk12@pitt.edu
Sweta Rawal - swr22@pitt.edu
Neha Shah - nes95@pitt.edu

Part 1: Setting up Hadoop

We set up the yarn-sit.xml, core-site.xml, hdfs-site.xml, and mapred-site.xml to build up the cluster in hadoop. Starting up the yarn services and hdfs services by using the following command: `sbin/start-dfs.sh, sbin/start-yarn.sh`

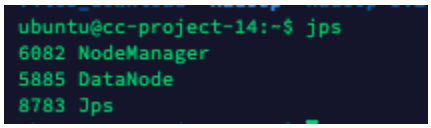
cc-project-13 - Name node

cc-project-14 - Data node

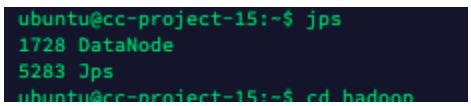
cc-project-15 - Data node

In namenode, we get the jps in below:

In datanode 1: cc-project-14



In datanode 2: cc-project-15



Report details.

```
Live datanodes (2):

Name: 10.11.10.35:9866 (cc-project-14.cc-s21-project.is2750-pg0.utah.cloudlab.us)
Hostname: cc-project-14.cc-s21-project.is2750-pg0.utah.cloudlab.us
Decommission Status : Normal
Configured Capacity: 41442029568 (38.60 GB)
DFS Used: 106496 (104 KB)
Non DFS Used: 4567945216 (4.25 GB)
DFS Remaining: 36857200640 (34.33 GB)
DFS Used%: 0.00%
DFS Remaining%: 88.94%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Fri Feb 26 22:13:58 UTC 2021
Last Block Report: Fri Feb 26 22:08:34 UTC 2021
Num of Blocks: 8


Name: 10.11.13.132:9866 (cc-project-15.cc-s21-project.is2750-pg0.utah.cloudlab.us)
Hostname: cc-project-15.cc-s21-project.is2750-pg0.utah.cloudlab.us
Decommission Status : Normal
Configured Capacity: 41442029568 (38.60 GB)
DFS Used: 106496 (104 KB)
Non DFS Used: 4080025600 (3.80 GB)
DFS Remaining: 37345120256 (34.78 GB)
DFS Used%: 0.00%
DFS Remaining%: 90.11%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Fri Feb 26 22:13:58 UTC 2021
Last Block Report: Fri Feb 26 22:08:34 UTC 2021
Num of Blocks: 8
```

Part 2: Docker image

Hadoop-Docker.zip contains the following:

1. Dockerfile: to create image
2. Config files to be copied to hadoop (core-site.xml, mapred-site.xml, hdfs-site.xml, yarn-site.xml)
3. Bootstrap.sh script for onstartup execution.

Step to create Docker image and run word count example

1. Unzip Hadoop-Docker
2. In terminal, change directory to Hadoop-Docker
3. To build image run,

    docker build -f Dockerfile -t <yourImageName> .

Example:

```
Tazeens-MacBook-Air:Hadoop-Docker tazeen$ docker build -f Dockerfile -t test1 .
[+] Building 5.5s (23/23) FINISHED
 => [internal] load build definition from Dockerfile                                                                                0.1s
 => => transferring dockerfile: 168B                                                                                                0.0s
 => [internal] load .dockerignore                                                                                                   0.1s
 => => transferring context: 2B                                                                                                     0.0s
 => [internal] load metadata for docker.io/library/ubuntu:20.04                                                                     1.7s
 => [ 1/18] FROM docker.io/library/ubuntu:20.04@sha256:703218c0465075f4425e58fac086e09e1de5c340b12976ab9eb8ad26615c3715             0.0s
 => [internal] load build context                                                                                                   0.1s
 => => transferring context: 1.40kB                                                                                                 0.0s
 => CACHED [ 2/18] RUN apt-get update                                                                                               0.0s
 => CACHED [ 3/18] RUN apt-get install -y curl tar sudo openssh-server openssh-client rsync                                         0.0s
 => CACHED [ 4/18] RUN apt-get update && apt-get install -y openjdk-8-jdk wget                                                      0.0s
 => CACHED [ 5/18] RUN rm -f /etc/ssh/ssh_host_dsa_key /etc/ssh/ssh_host_rsa_key /root/.ssh/id_rsa                                  0.0s
 => CACHED [ 6/18] RUN ssh-keygen -q -N "" -t dsa -f /etc/ssh/ssh_host_dsa_key                                                      0.0s
 => CACHED [ 7/18] RUN ssh-keygen -q -N "" -t rsa -f /etc/ssh/ssh_host_rsa_key                                                      0.0s
 => CACHED [ 8/18] RUN ssh-keygen -q -N "" -t rsa -f /root/.ssh/id_rsa                                                              0.0s
 => CACHED [ 9/18] RUN cp /root/.ssh/id_rsa.pub /root/.ssh/authorized_keys                                                          0.0s
 => CACHED [10/18] RUN     wget https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz &&     tar -xzf hadoop-3  0.0s
 => CACHED [11/18] ADD core-site.xml /usr/local/hadoop/etc/hadoop/core-site.xml                                                     0.0s
 => CACHED [12/18] ADD hdfs-site.xml /usr/local/hadoop/etc/hadoop/hdfs-site.xml                                                     0.0s
 => CACHED [13/18] ADD mapred-site.xml /usr/local/hadoop/etc/hadoop/mapred-site.xml                                                 0.0s
 => CACHED [14/18] ADD yarn-site.xml /usr/local/hadoop/etc/hadoop/yarn-site.xml                                                     0.0s
 => CACHED [15/18] RUN /usr/local/hadoop/bin/hdfs namenode -format                                                                  0.0s
 => [16/18] ADD bootstrap.sh /etc/bootstrap.sh                                                                                      0.1s
 => [17/18] RUN chown root:root /etc/bootstrap.sh                                                                                   2.1s
 => [18/18] RUN chmod 700 /etc/bootstrap.sh                                                                                         0.8s
 => exporting to image                                                                                                              0.3s
 => => exporting layers                                                                                                             0.2s
 => => writing image sha256:f6e3a48ec4ec02efd0d3a04b0bb37269f0baeeb8d1750c411250148ef311fdc5                                        0.0s
 => => naming to docker.io/library/test1                                                                                            0.0s
```

4. To create container from image run,

   docker run -it -d -p 52025:25 <yourImageName

Example:

```
Tazeens-MacBook-Air:Hadoop-Docker tazeen$ docker run -it -d -p 52025:25 test1
43dccdff22ab669696e9f3667291bf72f4fe816ccfd71033266a1f68687e5692
```

To run word count example in container

5. Open container terminal
6. Run: ***cd to usr/local/hadoop***
7. Run: ***hdfs dfs -mkdir -p /user/root/input***
8. Run: ***hdfs dfs -put LICENSE.txt /user/root/input/***
9. Run: ***hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar wordcount /user/root/input output***

```
# hdfs dfs -mkdir -p /user/root/input
# hdfs dfs -put LICENSE.txt /user/root/input/
2021-02-24 20:38:04,367 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
# hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar wordcoun
t /user/root/input output
```

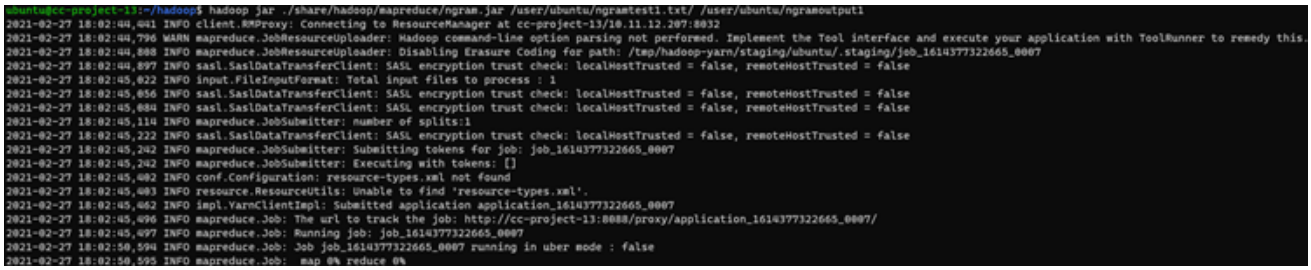
10. Run: ***hdfs dfs -ls /user/root/output***
11. Run: ***hdfs dfs -cat /user/root/output/part-r-00000***

```
                Bytes Written=55524
# hdfs dfs -cat /user/root/output/part-r-00000



2021-02-26 23:40:15,862 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
""AS    2
"AS     25
"AS-IS" 1
"Adaptation"    1
"COPYRIGHTS     1
"Collection"    1
"Collective     1
"Contribution"  2
"Contributor"   2
"Creative       1
"Derivative     2
"Distribute"    1
"French 2
```

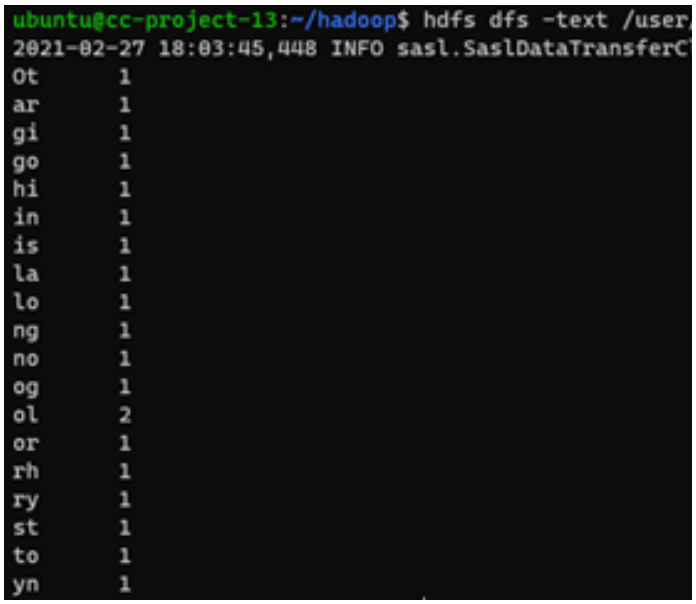
Part 3: Developing a Hadoop program

N-Grams Program

1. Move to hdfs:***hdfs dfs -put '/home/ubuntu/ngramtest1.txt' /user/ubuntu***
2. Run to check the file in hdfs: ***hdfs dfs -ls***
3. Run: ***hadoop jar ./share/hadoop/mapreduce/ngram.jar /user/ubuntu/ngramtest1.txt/ /user/ubuntu/ngramoutput1***
4. Check the output: ***hdfs dfs -text /user/ubuntu/ngramoutput1/part-r-00000***



Output:



Part 4: Developing a Hadoop program to analyze real logs

1. How many hits were made to the website item “/assets/img/home-logo.png”?
   Ans: **98776**

2. How many hits were made from the IP: 10.153.239.5
   Ans. **547**

3. Which path in the website has been hit most? How many hits were made to the path?
   Ans: ‘**/assets/css/combined.css’, 117348**
4. Which IP accesses the website most? How many accesses were made by it?

   Ans: **10.216.113.172, 158614**

Running Log Programs

Access Log 1 Program

1. Move to hdfs:***hdfs dfs -put '/home/ubuntu/access_log' /user/ubuntu/log1***
2. Run to check the file in hdfs: ***hdfs dfs -ls***
3. Run: ***hadoop jar ./share/hadoop/mapreduce/Accesslog1.jar /user/ubuntu/access_log/ /user/ubuntu/log1***
4. Check the output: ***hdfs dfs -text /user/ubuntu/log1/part-r-00000***

```
21/02/28 17:43:42 INFO mapred.JobClient: Job complete: job_local1119245266_0001
21/02/28 17:43:42 INFO mapred.JobClient: Counters: 17
21/02/28 17:43:42 INFO mapred.JobClient:   Map-Reduce Framework
21/02/28 17:43:42 INFO mapred.JobClient:     Spilled Records=36
21/02/28 17:43:42 INFO mapred.JobClient:     Map output materialized bytes=570
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce input records=15
21/02/28 17:43:42 INFO mapred.JobClient:     Map input records=4477843
21/02/28 17:43:42 INFO mapred.JobClient:     SPLIT_RAW_BYTES=1830
21/02/28 17:43:42 INFO mapred.JobClient:     Map output bytes=2963280
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce shuffle bytes=0
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce input groups=1
21/02/28 17:43:42 INFO mapred.JobClient:     Combine output records=15
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce output records=1
21/02/28 17:43:42 INFO mapred.JobClient:     Map output records=98776
21/02/28 17:43:42 INFO mapred.JobClient:     Combine input records=98776
21/02/28 17:43:42 INFO mapred.JobClient:     Total committed heap usage (bytes)=3513778176
21/02/28 17:43:42 INFO mapred.JobClient:   File Input Format Counters
21/02/28 17:43:42 INFO mapred.JobClient:     Bytes Read=504998876
21/02/28 17:43:42 INFO mapred.JobClient:   FileSystemCounters
21/02/28 17:43:42 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=794394684
21/02/28 17:43:42 INFO mapred.JobClient:     FILE_BYTES_READ=5343926227
21/02/28 17:43:42 INFO mapred.JobClient:   File Output Format Counters
21/02/28 17:43:42 INFO mapred.JobClient:     Bytes Written=44
```

```
/assets/img/home-logo.png       98776
```

Access Log 2 Program

5. Run: ***hadoop jar ./share/hadoop/mapreduce/Accesslog2.jar /user/ubuntu/access_log/ /user/ubuntu/log2***
6. Check the output: ***hdfs dfs -text /user/ubuntu/log2/part-r-00000***

```
21/02/28 17:43:42 INFO mapred.JobClient: Job complete: job_local1119245266_0001
21/02/28 17:43:42 INFO mapred.JobClient: Counters: 17
21/02/28 17:43:42 INFO mapred.JobClient:   Map-Reduce Framework
21/02/28 17:43:42 INFO mapred.JobClient:     Spilled Records=36
21/02/28 17:43:42 INFO mapred.JobClient:     Map output materialized bytes=570
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce input records=15
21/02/28 17:43:42 INFO mapred.JobClient:     Map input records=4477843
21/02/28 17:43:42 INFO mapred.JobClient:     SPLIT_RAW_BYTES=1830
21/02/28 17:43:42 INFO mapred.JobClient:     Map output bytes=2963280
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce shuffle bytes=0
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce input groups=1
21/02/28 17:43:42 INFO mapred.JobClient:     Combine output records=15
21/02/28 17:43:42 INFO mapred.JobClient:     Reduce output records=1
21/02/28 17:43:42 INFO mapred.JobClient:     Map output records=98776
21/02/28 17:43:42 INFO mapred.JobClient:     Combine input records=98776
21/02/28 17:43:42 INFO mapred.JobClient:     Total committed heap usage (bytes)=3513778176
21/02/28 17:43:42 INFO mapred.JobClient:   File Input Format Counters
21/02/28 17:43:42 INFO mapred.JobClient:     Bytes Read=504998876
21/02/28 17:43:42 INFO mapred.JobClient:   FileSystemCounters
21/02/28 17:43:42 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=794394684
21/02/28 17:43:42 INFO mapred.JobClient:     FILE_BYTES_READ=5343926227
21/02/28 17:43:42 INFO mapred.JobClient:   File Output Format Counters
21/02/28 17:43:42 INFO mapred.JobClient:     Bytes Written=44
```

```
10.153.239.5    547
```

Access Log 3 Program

7. Run: ***hadoop jar ./share/hadoop/mapreduce/Accesslog3.jar /user/ubuntu/access_log/ /user/ubuntu/log3***
8. Check the output: ***hdfs dfs -text /user/ubuntu/log3/part-r-00000***

```
21/02/28 23:30:35 INFO mapred.JobClient:  map 100% reduce 100%
21/02/28 23:30:35 INFO mapred.JobClient: Job complete: job_local1294465364_0001
21/02/28 23:30:35 INFO mapred.JobClient: Counters: 17
21/02/28 23:30:35 INFO mapred.JobClient:   Map-Reduce Framework
21/02/28 23:30:35 INFO mapred.JobClient:     Spilled Records=102
21/02/28 23:30:35 INFO mapred.JobClient:     Map output materialized bytes=1635
21/02/28 23:30:35 INFO mapred.JobClient:     Reduce input records=30
21/02/28 23:30:35 INFO mapred.JobClient:     Map input records=4477843
21/02/28 23:30:35 INFO mapred.JobClient:     SPLIT_RAW_BYTES=1830
21/02/28 23:30:35 INFO mapred.JobClient:     Map output bytes=261044349
21/02/28 23:30:35 INFO mapred.JobClient:     Reduce shuffle bytes=0
21/02/28 23:30:35 INFO mapred.JobClient:     Reduce input groups=10
21/02/28 23:30:35 INFO mapred.JobClient:     Combine output records=30
21/02/28 23:30:35 INFO mapred.JobClient:     Reduce output records=1
21/02/28 23:30:35 INFO mapred.JobClient:     Map output records=4477843
21/02/28 23:30:35 INFO mapred.JobClient:     Combine input records=4477843
21/02/28 23:30:35 INFO mapred.JobClient:     Total committed heap usage (bytes)=3933208576
21/02/28 23:30:35 INFO mapred.JobClient:   File Input Format Counters
21/02/28 23:30:35 INFO mapred.JobClient:     Bytes Read=504998876
21/02/28 23:30:35 INFO mapred.JobClient:   FileSystemCounters
21/02/28 23:30:35 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=1026397139
21/02/28 23:30:35 INFO mapred.JobClient:     FILE_BYTES_READ=5574122435
21/02/28 23:30:35 INFO mapred.JobClient:   File Output Format Counters
21/02/28 23:30:35 INFO mapred.JobClient:     Bytes Written=84
```

```
GET /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg HTTP/1.1       82918
```

Access Log 4 Program

9. Run: ***hadoop jar ./share/hadoop/mapreduce/Accesslog4.jar /user/ubuntu/access_log/ /user/ubuntu/log4***
10. Check the output: ***hdfs dfs -text /user/ubuntu/log4/part-r-00000***

```
21/02/28 23:35:39 INFO mapred.JobClient:  map 100% reduce 100%
21/02/28 23:35:39 INFO mapred.JobClient: Job complete: job_local1815634825_0001
21/02/28 23:35:39 INFO mapred.JobClient: Counters: 17
21/02/28 23:35:39 INFO mapred.JobClient:   Map-Reduce Framework
21/02/28 23:35:39 INFO mapred.JobClient:     Spilled Records=102
21/02/28 23:35:39 INFO mapred.JobClient:     Map output materialized bytes=690
21/02/28 23:35:39 INFO mapred.JobClient:     Reduce input records=30
21/02/28 23:35:39 INFO mapred.JobClient:     Map input records=4477843
21/02/28 23:35:39 INFO mapred.JobClient:     SPLIT_RAW_BYTES=1830
21/02/28 23:35:39 INFO mapred.JobClient:     Map output bytes=79641285
21/02/28 23:35:39 INFO mapred.JobClient:     Reduce shuffle bytes=0
21/02/28 23:35:39 INFO mapred.JobClient:     Reduce input groups=21
21/02/28 23:35:39 INFO mapred.JobClient:     Combine output records=30
21/02/28 23:35:39 INFO mapred.JobClient:     Reduce output records=1
21/02/28 23:35:39 INFO mapred.JobClient:     Map output records=4477843
21/02/28 23:35:39 INFO mapred.JobClient:     Combine input records=4477843
21/02/28 23:35:39 INFO mapred.JobClient:     Total committed heap usage (bytes)=4134535168
21/02/28 23:35:39 INFO mapred.JobClient:   File Input Format Counters
21/02/28 23:35:39 INFO mapred.JobClient:     Bytes Read=504998876
21/02/28 23:35:39 INFO mapred.JobClient:   FileSystemCounters
21/02/28 23:35:39 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=1026381379
21/02/28 23:35:39 INFO mapred.JobClient:     FILE_BYTES_READ=5574113456
21/02/28 23:35:39 INFO mapred.JobClient:   File Output Format Counters
21/02/28 23:35:39 INFO mapred.JobClient:     Bytes Written=34
```

```
10.216.113.172  158614
```