Hadoop Testing Guide

NOTE: If you are creating a new VM please allocate sufficient resources. Recommended: Memory-4GB Disk-50GB

Step 1 Start Hadoop services

Command : start-all.sh

```
bigdata:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as bigdata in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bigdata-virtual-machine]
Starting resourcemanager
Starting nodemanagers
bigdata:~$
```

Make sure 6 nodes are running:

Command : jps

```
bigdata:~$ jps
7040 NameNode
12659 Jps
7187 DataNode
7405 SecondaryNameNode
7773 NodeManager
7647 ResourceManager
bigdata:~$
```

Step 2: Make Mapper and Reducer Executable

Command : sudo chmod +x <directory of mapper.py>

Command : sudo chmod +x <directory of reducer.py>

```
/hadoop-handson$ chmod +x mapper.py
/hadoop-handson$ chmod +x reducer.py
```

Step 3: Upload Data to HDFS

Create a directory in HDFS and upload the ‘sample_data.json’ file.

Command : hdfs dfs -mkdir -p /input/

Command : hdfs dfs -put <directory of wordcount.txt> /input/

```
/hadoop-handson$ hdfs dfs -mkdir -p /input/
```

```
/hadoop-handson$ hdfs dfs -put '/home/abhinav/Desktop/hadoop-handson/wordcount.txt' /input/
```

NOTE: Use wordcount.txt (1.3 MB) as the sample dataset. And for larger dataset use wikisent2.txt (934.57 MB)

Step 4: dos2unix

Use the ‘dos2unix’ command to convert a file's line endings when transferring it from Windows to Unix systems to ensure compatibility.

Command: sudo apt install dos2unix

Command: dos2unix mapper.py

Command: dos2unix reducer.py

```
/hive-installation$ sudo apt install dos2unix
```

```
abhinav@abhinav:~/Desktop/hadoop-handson$ dos2unix mapper.py
dos2unix: converting file mapper.py to Unix format...
abhinav@abhinav:~/Desktop/hadoop-handson$ dos2unix reducer.py
dos2unix: converting file reducer.py to Unix format...
abhinav@abhinav:~/Desktop/hadoop-handson$
```

Step 5: Run the Hadoop Streaming Job

Command :

hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \

-mapper <path to mapper.py> \

-reducer <path to reducer.py> \

-input <path to directory of wordcount.txt inside HDFS> \

-output <path to output directory>

```
abhinav@abhinav:~/hadoop-3.4.0/bin$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
> -mapper /home/abhinav/Desktop/hadoop-handson/mapper.py \
> -reducer /home/abhinav/Desktop/hadoop-handson/reducer.py \
> -input /input \
> -output /output/op1
2024-08-12 15:00:01,770 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
packageJobJar: [/tmp/hadoop-unjar5682426764359697417/] [] /tmp/streamjob8470366794307531512.jar tmpDir=null
2024-08-12 15:00:02,362 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2024-08-12 15:00:02,471 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2024-08-12 15:00:03,082 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/abhinav/.staging/job_172345451
8852_0001
2024-08-12 15:00:04,156 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-12 15:00:04,172 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-08-12 15:00:05,191 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-12 15:00:06,091 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1723454518852_0001
2024-08-12 15:00:06,091 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-12 15:00:06,637 INFO conf.Configuration: resource-types.xml not found
2024-08-12 15:00:06,638 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-12 15:00:07,590 INFO impl.YarnClientImpl: Submitted application application_1723454518852_0001
2024-08-12 15:00:07,677 INFO mapreduce.Job: The url to track the job: http://abhinav:8088/proxy/application_1723454518852_0001/
2024-08-12 15:00:07,684 INFO mapreduce.Job: Running job: job_1723454518852_0001
2024-08-12 15:00:20,097 INFO mapreduce.Job: Job job_1723454518852_0001 running in uber mode : false
2024-08-12 15:00:20,098 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-12 15:00:38,393 INFO mapreduce.Job:  map 28% reduce 0%
2024-08-12 15:00:39,438 INFO mapreduce.Job:  map 57% reduce 0%
2024-08-12 15:00:43,570 INFO mapreduce.Job:  map 66% reduce 0%
2024-08-12 15:00:44,581 INFO mapreduce.Job:  map 76% reduce 0%
2024-08-12 15:00:46,653 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-12 15:01:02,998 INFO mapreduce.Job:  map 100% reduce 97%
2024-08-12 15:01:05,053 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-12 15:01:08,168 INFO mapreduce.Job: Job job_1723454518852_0001 completed successfully
```

After successfully running the code the output should be as below:

```
                Combine input records=0
                Combine output records=0
                Reduce input groups=1229395
                Reduce shuffle bytes=198975623
                Reduce input records=15136661
                Reduce output records=1229395
                Spilled Records=45409983
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1060
                CPU time spent (ms)=89250
                Physical memory (bytes) snapshot=1483849728
                Virtual memory (bytes) snapshot=7661563904
                Total committed heap usage (bytes)=1411383296
                Peak Map Physical memory (bytes)=491249664
                Peak Map Virtual memory (bytes)=2569289728
                Peak Reduce Physical memory (bytes)=615206912
                Peak Reduce Virtual memory (bytes)=2665054208
        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=233621022
        File Output Format Counters
                Bytes Written=15235270
2024-08-08 13:15:17,167 INFO streaming.StreamJob: Output directory: /output/op1
abhinav@abhinav:~/hadoop-3.4.0/bin$
```

Here it can be observed how many number of mappers were used to parse the file (larger the file size more number of mapper splits will be observed)

```
abhinav@abhinav:~/hadoop-3.4.0/bin$
abhinav@abhinav:~/hadoop-3.4.0/bin$ hadoop jar '/home/abhinav/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0
.jar' -mapper '/home/abhinav/Desktop/hadoop-handson/mapper.py' -reducer '/home/abhinav/Desktop/hadoop-handson/reducer.py
' -input /input -output /output/op1
2024-08-05 16:48:21,546 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
tin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar8565240259454139362/] [] /tmp/streamjob2367572725436720725.jar tmpDir=null
2024-08-05 16:48:22,907 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2024-08-05 16:48:23,258 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2024-08-05 16:48:23,765 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
abhinav/.staging/job_1722851617514_0004
2024-08-05 16:48:24,400 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-05 16:48:25,304 INFO mapreduce.JobSubmitter: number of splits:7
2024-08-05 16:48:26,045 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1722851617514_0004
2024-08-05 16:48:26,046 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-05 16:48:26,609 INFO conf.Configuration: resource-types.xml not found
2024-08-05 16:48:26,610 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-05 16:48:26,844 INFO impl.YarnClientImpl: Submitted application application_1722851617514_0004
2024-08-05 16:48:27,004 INFO mapreduce.Job: The url to track the job: http://abhinav:8088/proxy/application_172285161751
4_0004/
2024-08-05 16:48:27,006 INFO mapreduce.Job: Running job: job_1722851617514_0004
2024-08-05 16:48:37,721 INFO mapreduce.Job: Job job_1722851617514_0004 running in uber mode : false
2024-08-05 16:48:37,723 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-05 16:49:00,281 INFO mapreduce.Job:  map 4% reduce 0%
2024-08-05 16:49:02,099 INFO mapreduce.Job:  map 21% reduce 0%
2024-08-05 16:49:08,643 INFO mapreduce.Job:  map 24% reduce 0%
2024-08-05 16:49:14,310 INFO mapreduce.Job:  map 31% reduce 0%
2024-08-05 16:49:21,106 INFO mapreduce.Job:  map 42% reduce 0%
2024-08-05 16:49:27,497 INFO mapreduce.Job:  map 45% reduce 0%
2024-08-05 16:49:34,865 INFO mapreduce.Job:  map 53% reduce 0%
2024-08-05 16:49:40,634 INFO mapreduce.Job:  map 57% reduce 0%
```

Here it is taking 7 mapper splits to perform wordcount on wikisent2.txt (900MB) text file

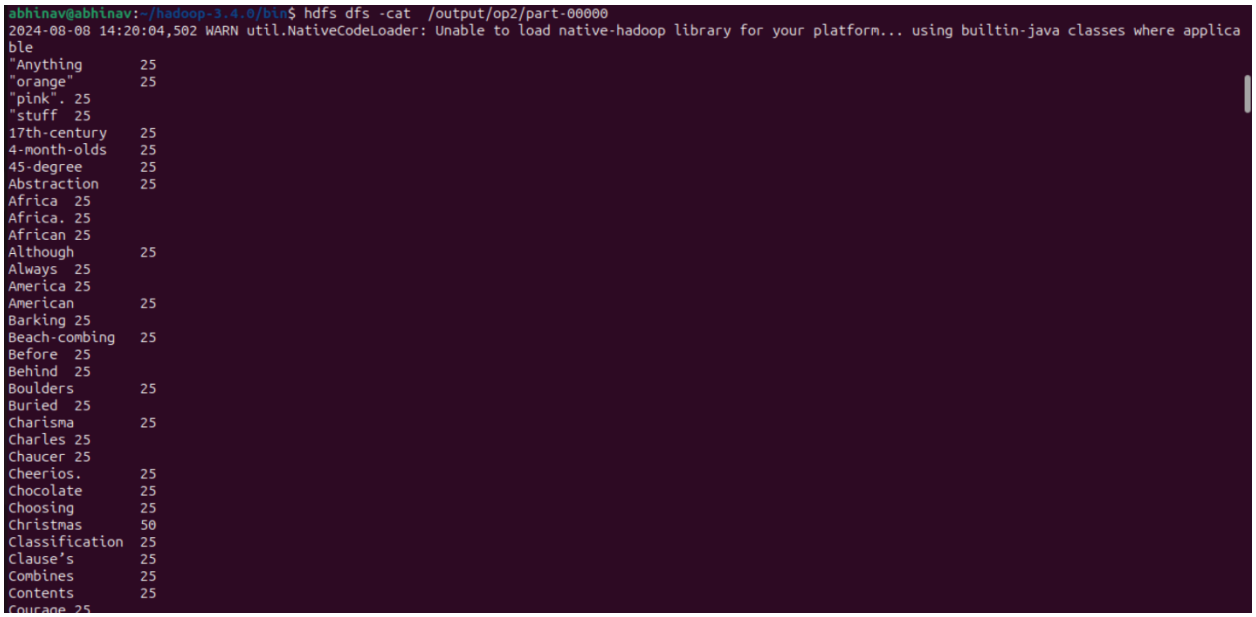
Step 6: Check the Output

Once the job is completed, view the output with

Command : hdfs dfs -ls /output/op1
(this will list all the output files created by Hadoop Map-Reduce operation)

Command: hdfs dfs -cat /output/op1/part-00000
(this will display the output generated by Hadoop Map-Reduce operation)

```
abhinav@abhinav:~/hadoop-3.4.0/bin$
abhinav@abhinav:~/hadoop-3.4.0/bin$ hdfs dfs -ls  /output/op1
2024-08-08 13:17:31,715 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
tin-java classes where applicable
Found 2 items
-rw-r--r--   1 abhinav supergroup          0 2024-08-08 13:15 /output/op1/_SUCCESS
-rw-r--r--   1 abhinav supergroup   15235270 2024-08-08 13:15 /output/op1/part-00000
abhinav@abhinav:~/hadoop-3.4.0/bin$
```



NOTE: The output of the sample dataset is here. And for the wiki dataset is here.

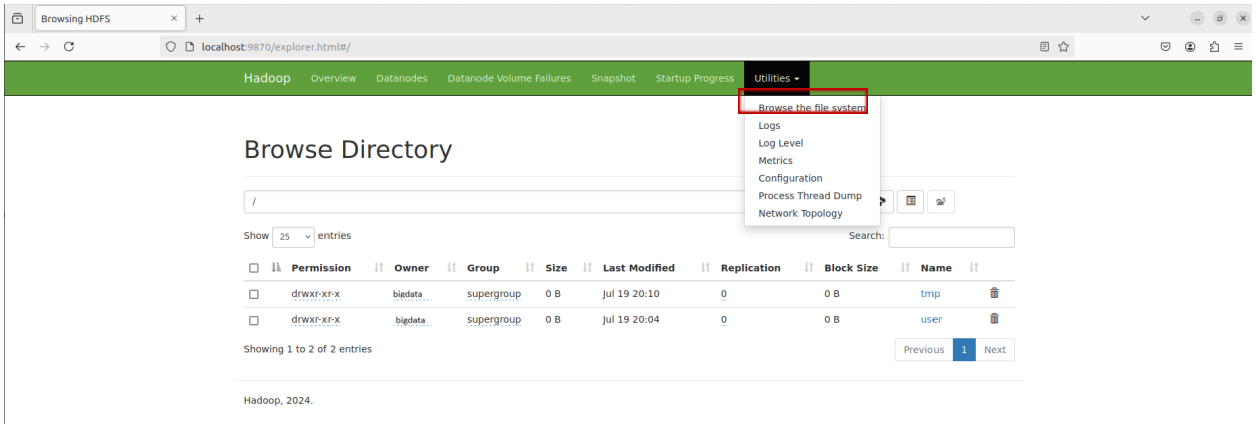
Troubleshooting

1) Check Log Files: If the job fails, Hadoop will generate log files that can help identify the issue.

   You can find these in the Hadoop logs directory, typically located at `$HADOOP_HOME/logs`.

2) Browse the HDFS Directory with Web UI.

Open a browser and type
localhost:9870
you should be able to see all your files



3) If the files are not visible in the UI
   Command :
   stop-all.sh
   sudo rm -rf ~/dfsdata/
   sudo rm -rf ~/tmpdata/
   hdfs namenode -format
   start-all.sh