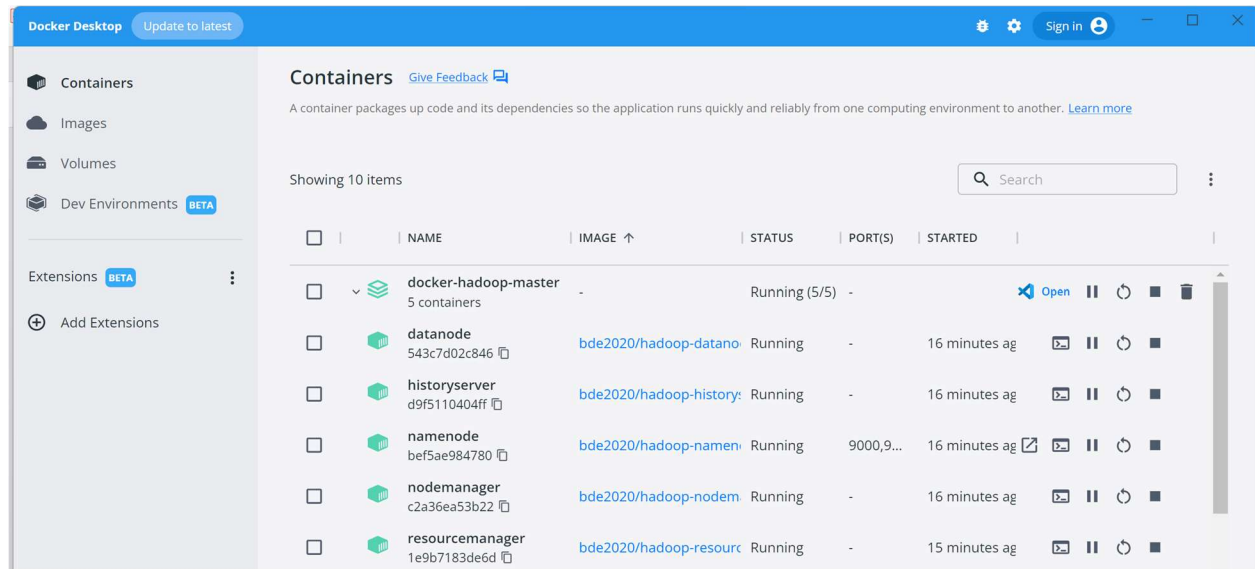
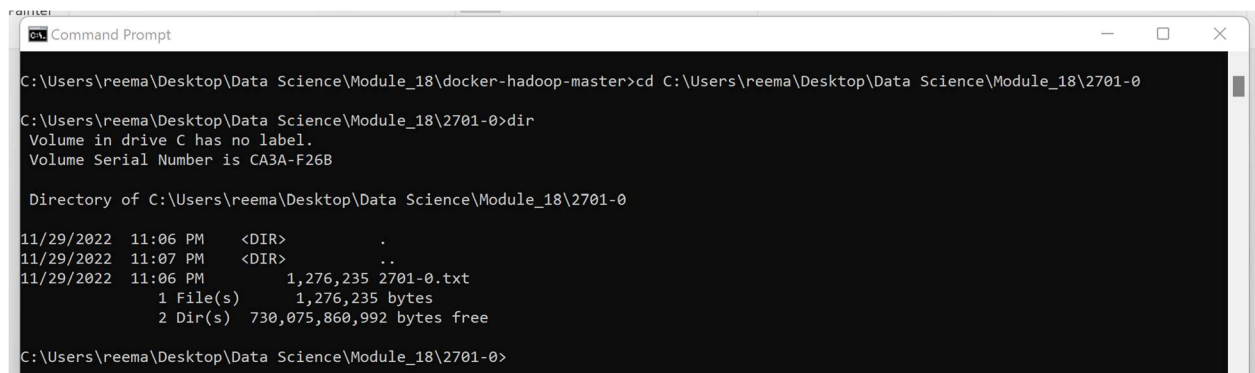


## 1. Deployed Docker desktop to set up Hadoop



## 2. Downloaded the Moby Dick .zip file and listed all files under folder '2701-0' on local computer



3. Created the input folder in the namenode *container*.

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
/
# mkdir input
# ls -l
total 416
-rw-r--r-- 1 root root 325083 Feb  4 2020 KEYS
drwxr-xr-x 1 root root 4096 Feb  4 2020 bin
drwxr-xr-x 2 root root 4096 Sep  8 2019 boot
drwxr-xr-x 5 root root 340 Nov 30 06:40 dev
-rwxr-xr-x 1 root root 4155 Feb  4 2020 entrypoint.sh
drwxr-xr-x 1 root root 4096 Nov 30 06:40 etc
drwxr-xr-x 3 root root 4096 Feb  4 2020 hadoop
drwxr-xr-x 2 root root 4096 Feb  4 2020 hadoop-data
drwxr-xr-x 2 root root 4096 Sep  8 2019 home
drwxr-xr-x 2 root root 4096 Nov 30 07:11 input
drwxr-xr-x 1 root root 4096 Jan 30 2020 lib
drwxr-xr-x 2 root root 4096 Jan 30 2020 lib64
drwxr-xr-x 2 root root 4096 Jan 30 2020 media
drwxr-xr-x 2 root root 4096 Jan 30 2020 mnt
drwxr-xr-x 1 root root 4096 Feb  4 2020 opt
dr-xr-xr-x 261 root root  0 Nov 30 06:40 proc
drwx----- 1 root root 4096 Feb  4 2020 root
drwxr-xr-x 3 root root 4096 Jan 30 2020 run
-rwxr-xr-x 1 root root 494 Feb  4 2020 run.sh
drwxr-xr-x 1 root root 4096 Feb  4 2020/sbin
drwxr-xr-x 2 root root 4096 Jan 30 2020/srv
dr-xr-xr-x 11 root root  0 Nov 30 06:40/sys
drwxrwxrwt 1 root root 4096 Nov 30 06:40/tmp
drwxr-xr-x 1 root root 4096 Jan 30 2020/usr
drwxr-xr-x 1 root root 4096 Jan 30 2020/var
#
```

4. Copied the .txt file to the namenode *container*.

```
Command Prompt
C:\Users\reema\Desktop\Data Science\Module_18\2701-0>docker cp 2701-0.txt namenode:/input/2701-0.txt
C:\Users\reema\Desktop\Data Science\Module_18\2701-0>
```

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
# pwd
/
# cd input
# pwd
/input
# ls -l
total 1248
-rwxr-xr-x 1 root root 1276235 Nov 30 07:06 2701-0.txt
#
```

5. Created an input folder on Hadoop file system

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
# cd ..
# pwd
/
# ls
KEYS  boot  entrypoint.sh  hadoop      home  lib  media  opt  root  run.sh  srv  tmp  var
bin   dev   etc             hadoop-data input  lib64 mnt   proc  run   sbin   sys  usr
# hadoop fs -mkdir -p input
#
```

6. Ran the HDFS command to copy the contents of the local input folder to the HDFS input folder.

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
# hadoop fs -mkdir -p input
# hdfs dfs -put ./input/* input
2022-12-01 02:43:56,753 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
#
```

7. Ran the curl command to download the jar file.

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
# hadoop fs -mkdir -p input
# hdfs dfs -put ./input/* input
2022-12-01 02:43:56,753 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
# curl -L https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-mapreduce-examples/2.7.1/hadoop-mapreduce-examples-2.7.1-sources.jar --output hadoop-mapreduce-examples-2.7.1-sources.jar
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total   Spent    Left   Speed
100  680k  100  680k    0     0  1215k      0  0:00:00  0:00:00  0:00:00  1216k
#
```

## 8. Successfully ran the word count program.

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
# hadoop jar hadoop-mapreduce-examples-2.7.1-sources.jar org.apache.hadoop.examples.WordCount input output
2022-12-01 05:38:07,561 INFO client.RMPProxy: Connecting to ResourceManager at resourcemanager/172.24.0.6:8032
2022-12-01 05:38:07,846 INFO client.AHSPProxy: Connecting to Application History server at historyserver/172.24.0.3:10200
2022-12-01 05:38:08,147 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
root/.staging/job_1669790450621_0001
2022-12-01 05:38:08,396 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2022-12-01 05:38:08,804 INFO input.FileInputFormat: Total input files to process : 1
2022-12-01 05:38:08,900 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2022-12-01 05:38:08,933 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2022-12-01 05:38:08,943 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-01 05:38:09,248 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2022-12-01 05:38:09,309 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669790450621_0001
2022-12-01 05:38:09,309 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-01 05:38:09,565 INFO conf.Configuration: resource-types.xml not found
2022-12-01 05:38:09,566 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-01 05:38:10,221 INFO impl.YarnClientImpl: Submitted application application_1669790450621_0001
2022-12-01 05:38:10,280 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1669
790450621_0001/
2022-12-01 05:38:10,282 INFO mapreduce.Job: Running job: job_1669790450621_0001
2022-12-01 05:38:22,611 INFO mapreduce.Job: Job job_1669790450621_0001 running in uber mode : false
2022-12-01 05:38:22,615 INFO mapreduce.Job: map 0% reduce 0%
2022-12-01 05:38:32,706 INFO mapreduce.Job: map 100% reduce 0%
2022-12-01 05:38:37,736 INFO mapreduce.Job: map 100% reduce 100%
2022-12-01 05:38:37,747 INFO mapreduce.Job: Job job_1669790450621_0001 completed successfully
2022-12-01 05:38:37,846 INFO mapreduce.Job: Counters: 54
File System Counters
```

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
Map-Reduce Framework
  Map input records=22316
  Map output records=215864
  Map output bytes=2113253
  Map output materialized bytes=159938
  Input split bytes=112
  Combine input records=215864
  Combine output records=33568
  Reduce input groups=33568
  Reduce shuffle bytes=159938
  Reduce input records=33568
  Reduce output records=33568
  Spilled Records=67136
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=284
  CPU time spent (ms)=6510
  Physical memory (bytes) snapshot=554610688
  Virtual memory (bytes) snapshot=13569843200
  Total committed heap usage (bytes)=452984832
  Peak Map Physical memory (bytes)=359751680
  Peak Map Virtual memory (bytes)=5111824384
  Peak Reduce Physical memory (bytes)=194859008
  Peak Reduce Virtual memory (bytes)=8458018816
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=1276235
```

9. Inspected results of file by running 'hdfs dfs -cat output/part-r-00000'

```
docker exec -it bef5ae984780ce60c53c1a12571f46a92e971ccfa074b67a23670d29949c27c7 /bin/sh
"Like 1
"Look 1
"Moby 1
"Mr. 1
"My 1
"Nay, 2
"Nay,' 1
"O'h 1
"Say 1
"Shall 1
"Shut 1
"Sink 1
"So 2
"Stern 1
"The 1
"Then 2
"This 1
"Though 1
"Turn 3
"Very 1
"Well 1
"What 3
"Where 1
"Who's 1
"Why 1
"Will 2
"Yes, 1
"You 2
" 'Bout 1
" 'Hind 1
" 'Tis 2
" 'Twill 1
" 'tis 1
#
```