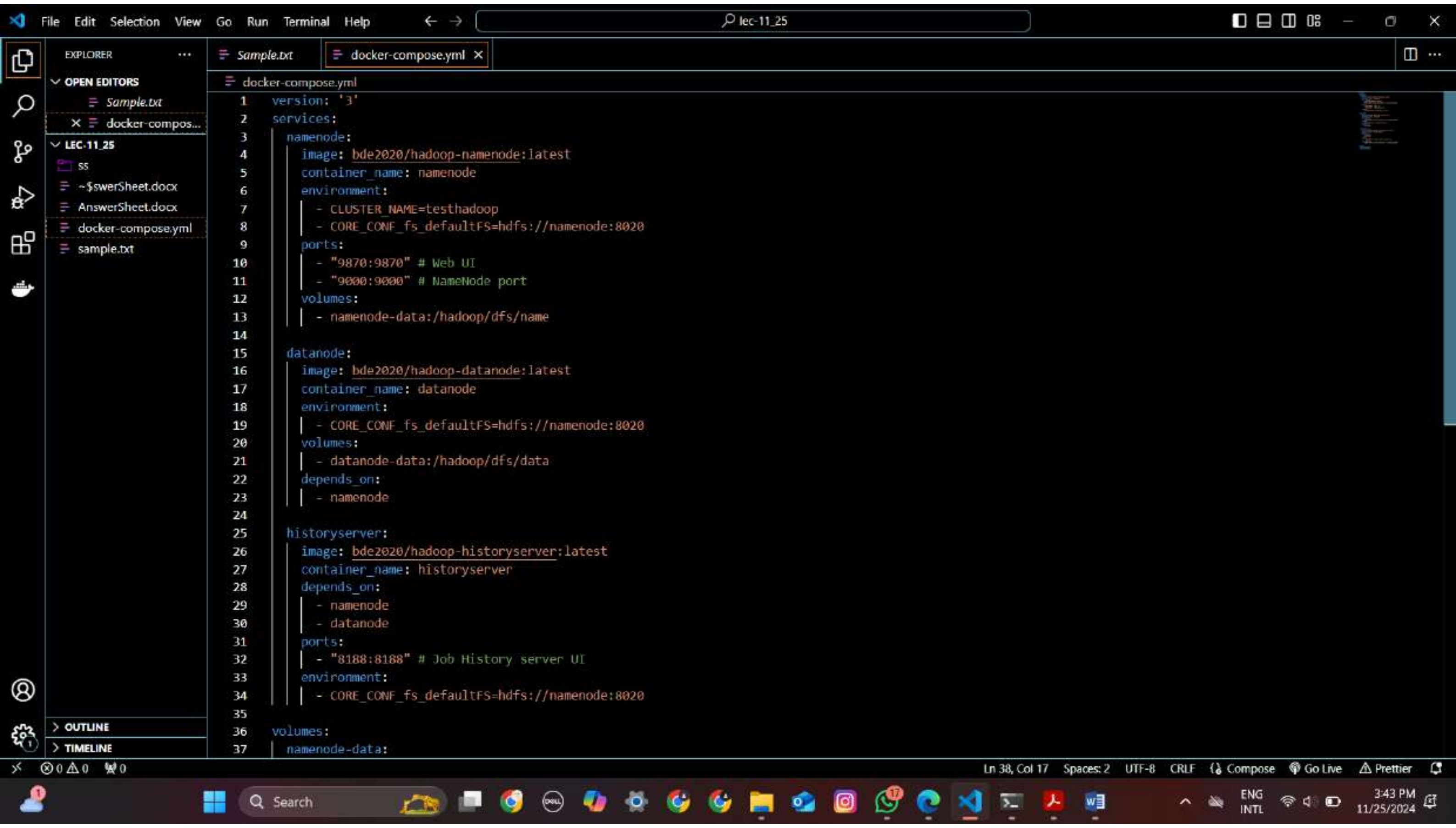
Task 1: Setting Up the Distributed Hadoop Cluster

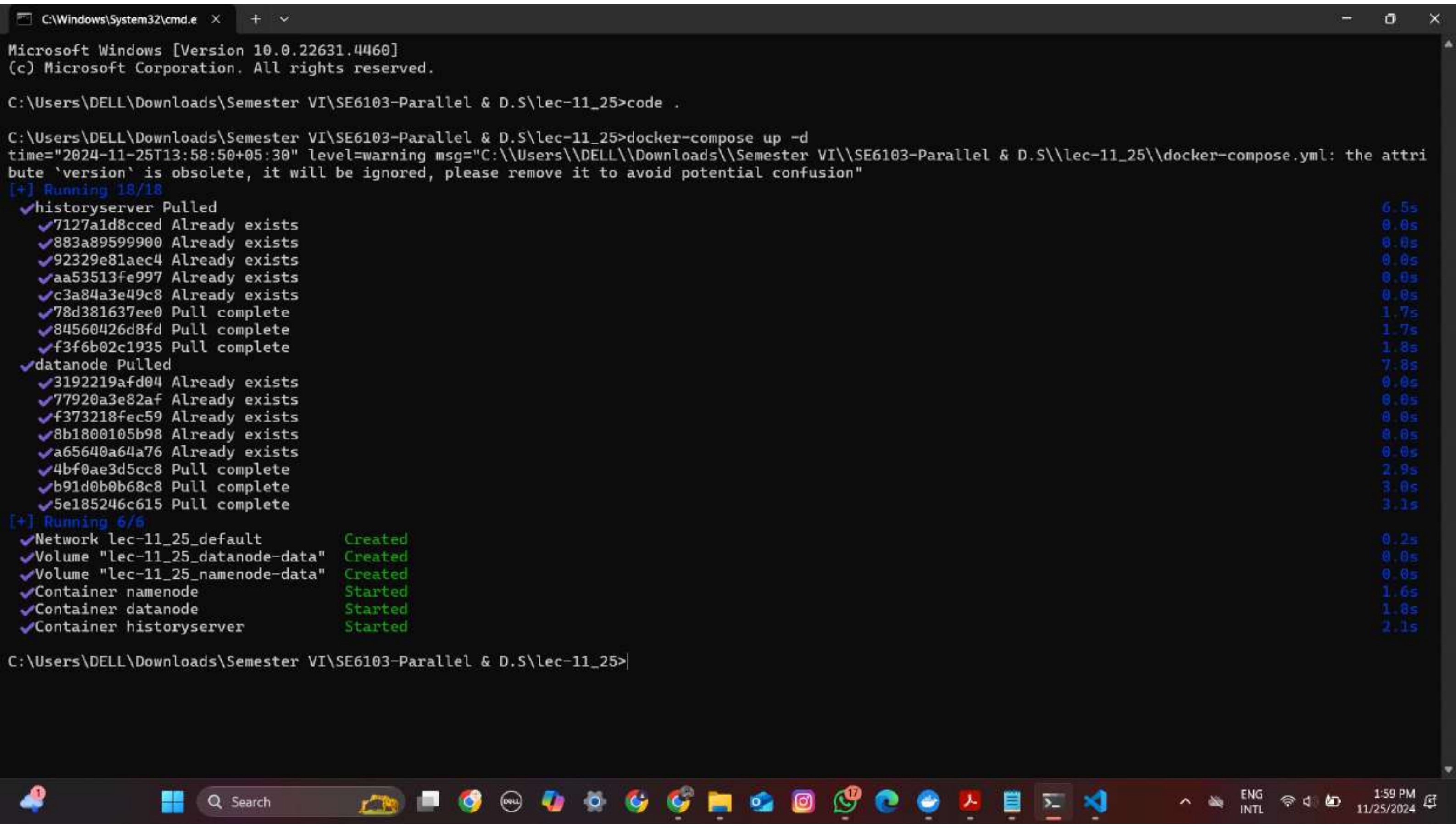
1. Step 1: Prepare the Docker Compose File



docker-compose.yml

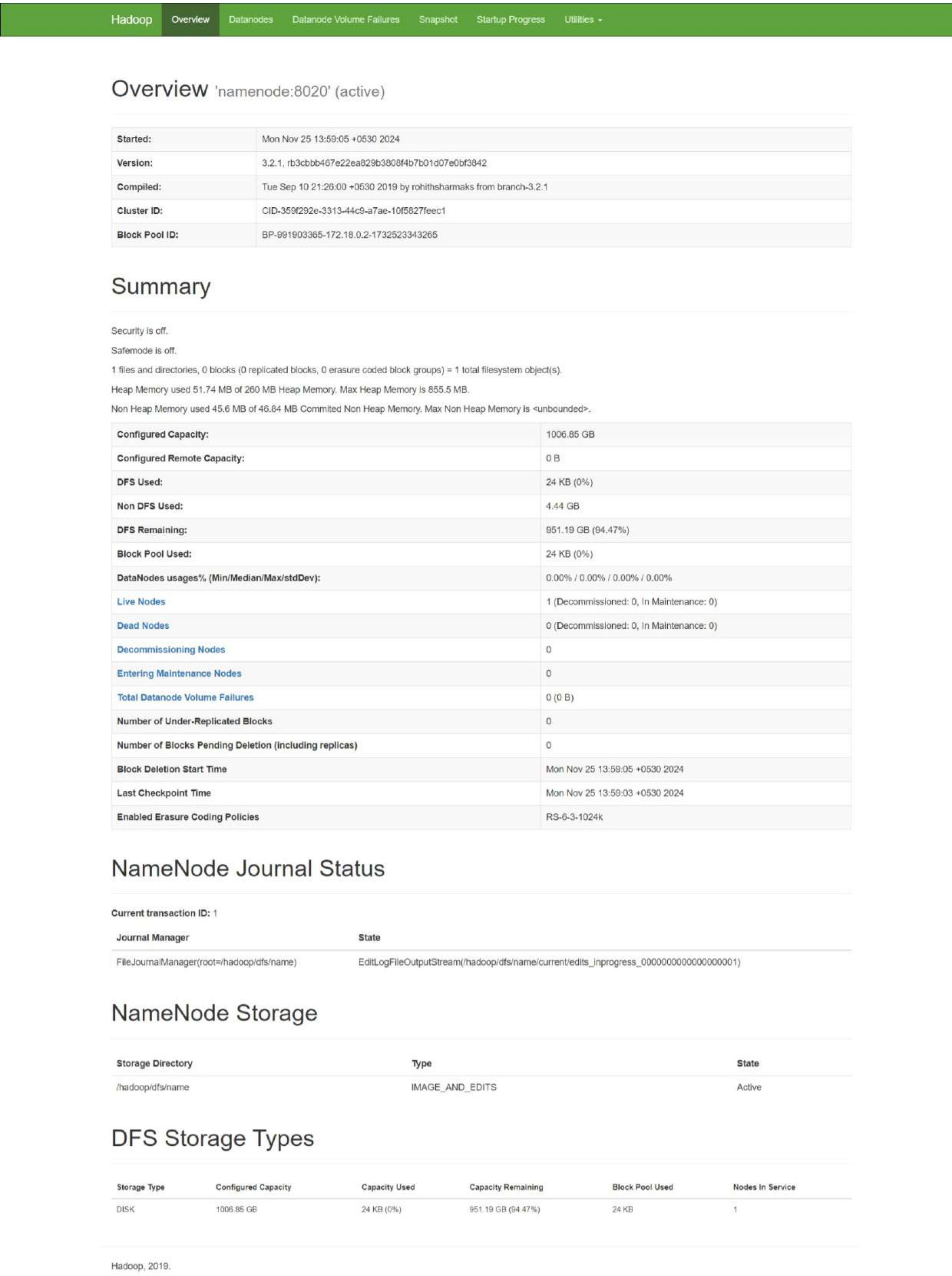
```yaml
version: '3'
services:
  namenode:
    image: bde2020/hadoop-namenode:latest
    container_name: namenode
    environment:
      - CLUSTER_NAME=testhadoop
      - CORE_CONF_fs_defaultFS=hdfs://namenode:8020
    ports:
      - "9870:9870" # Web UI
      - "9000:9000" # NameNode port
    volumes:
      - namenode-data:/hadoop/dfs/name

  datanode:
    image: bde2020/hadoop-datanode:latest
    container_name: datanode
    environment:
      - CORE_CONF_fs_defaultFS=hdfs://namenode:8020
    volumes:
      - datanode-data:/hadoop/dfs/data
    depends_on:
      - namenode

  historyserver:
    image: bde2020/hadoop-historyserver:latest
    container_name: historyserver
    depends_on:
      - namenode
      - datanode
    ports:
      - "8188:8188" # Job History server UI
    environment:
      - CORE_CONF_fs_defaultFS=hdfs://namenode:8020

volumes:
  namenode-data:
```

Step 2: Deploy the Cluster



```
Microsoft Windows [Version 10.0.22631.4460]
(c) Microsoft Corporation. All rights reserved.

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>code .

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker-compose up -d
time="2024-11-25T13:58:50+05:30" level=warning msg="C:\\Users\\DELL\\Downloads\\Semester VI\\SE6103-Parallel & D.S\\lec-11_25\\docker-compose.yml: the attri
bute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion"
[+] Running 18/18
  ✔historyserver Pulled                                     6.5s
    ✔7127a1d8cced Already exists                            0.0s
    ✔883a89599900 Already exists                            0.0s
    ✔92329e81aec4 Already exists                            0.0s
    ✔aa53513fe997 Already exists                            0.0s
    ✔c3a84a3e49c8 Already exists                            0.0s
    ✔78d381637ee0 Pull complete                             1.7s
    ✔84560426d8fd Pull complete                             1.7s
    ✔f3f6b02c1935 Pull complete                             1.8s
  ✔datanode Pulled                                          7.8s
    ✔3192219afd04 Already exists                            0.0s
    ✔77920a3e82af Already exists                            0.0s
    ✔f373218fec59 Already exists                            0.0s
    ✔8b1800105b98 Already exists                            0.0s
    ✔a65640a64a76 Already exists                            0.0s
    ✔4bf0ae3d5cc8 Pull complete                             2.9s
    ✔b91d0b0b68c8 Pull complete                             3.0s
    ✔5e185246c615 Pull complete                             3.1s
[+] Running 6/6
  ✔Network lec-11_25_default         Created                0.2s
  ✔Volume "lec-11_25_datanode-data"  Created                0.0s
  ✔Volume "lec-11_25_namenode-data"  Created                0.0s
  ✔Container namenode                Started                1.6s
  ✔Container datanode                Started                1.8s
  ✔Container historyserver           Started                2.1s

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>
```

Step 3: Verify Cluster Status



Overview 'namenode:8020' (active)

Started:	Mon Nov 25 13:59:05 +0530 2024
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 21:26:00 +0530 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-359f292e-3313-44c9-a7ae-10f5827feec1
Block Pool ID:	BP-991903365-172.18.0.2-1732523343265

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 51.74 MB of 260 MB Heap Memory. Max Heap Memory is 855.5 MB.

Non Heap Memory used 45.6 MB of 46.84 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1006.85 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	4.44 GB
DFS Remaining:	951.19 GB (94.47%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Mon Nov 25 13:59:05 +0530 2024
Last Checkpoint Time	Mon Nov 25 13:59:03 +0530 2024
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 1

Journal Manager	State
FileJournalManager(root=/hadoop/dfs/name)	EditLogFileOutputStream(/hadoop/dfs/name/current/edits_inprogress_0000000000000000001)

NameNode Storage

Storage Directory	Type	State
/hadoop/dfs/name	IMAGE_AND_EDITS	Active

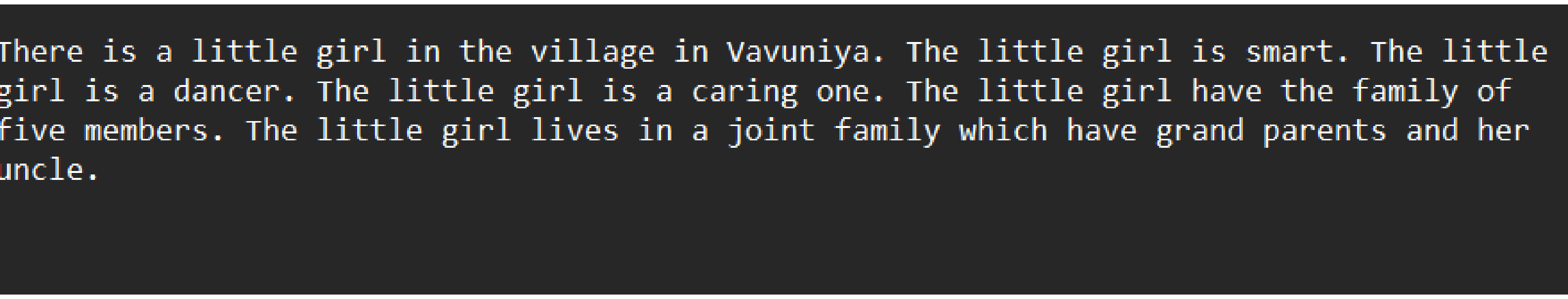
DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	1006.85 GB	24 KB (0%)	951.19 GB (94.47%)	24 KB	1

Hadoop, 2019.

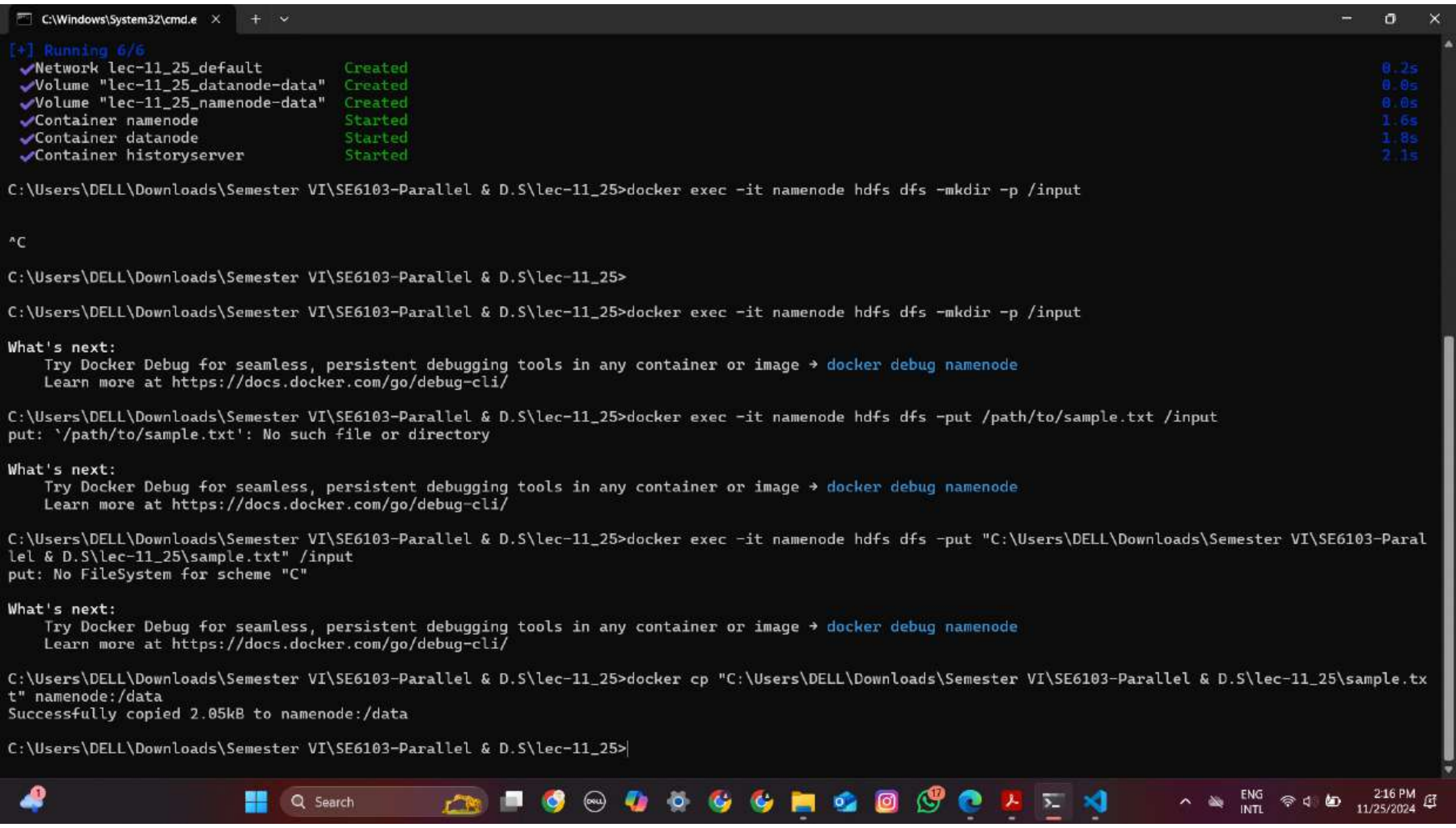
Task 2: Uploading Data to HDFS

1. Data Set-sample.txt

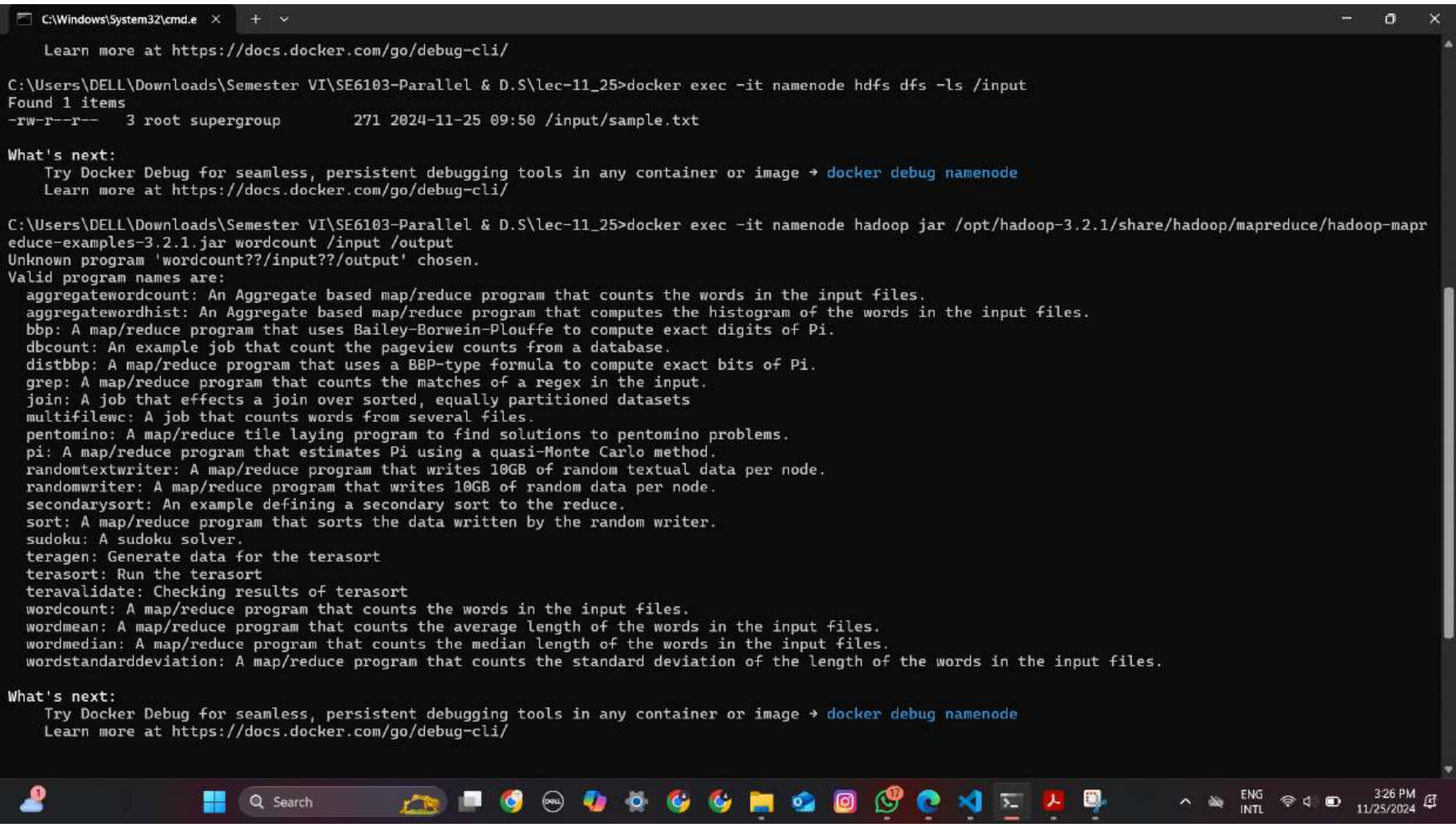


```
There is a little girl in the village in Vavuniya. The little girl is smart. The little
girl is a dancer. The little girl is a caring one. The little girl have the family of
five members. The little girl lives in a joint family which have grand parents and her
uncle.
```

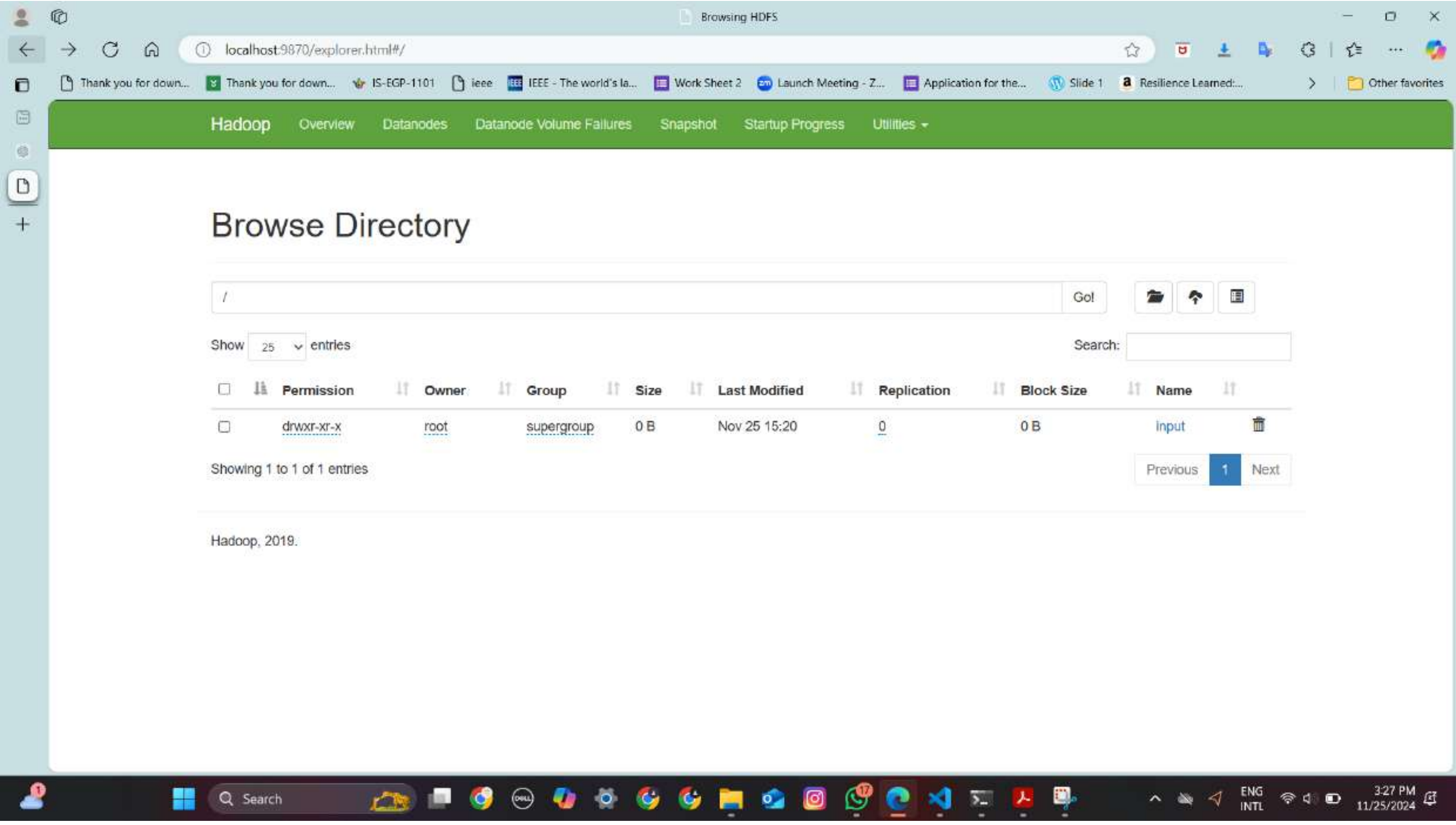
2. Upload Data to HDFS



```
[+] Running 6/6
 ✔Network lec-11_25_default         Created                                                                          0.2s
 ✔Volume "lec-11_25_datanode-data"  Created                                                                          0.0s
 ✔Volume "lec-11_25_namenode-data"  Created                                                                          0.0s
 ✔Container namenode                Started                                                                          1.6s
 ✔Container datanode                Started                                                                          1.8s
 ✔Container historyserver           Started                                                                          2.1s

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -mkdir -p /input


^C

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -mkdir -p /input

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -put /path/to/sample.txt /input
put: `/path/to/sample.txt': No such file or directory

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -put "C:\Users\DELL\Downloads\Semester VI\SE6103-Paral
lel & D.S\lec-11_25\sample.txt" /input
put: No FileSystem for scheme "C"

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker cp "C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25\sample.tx
t" namenode:/data
Successfully copied 2.05kB to namenode:/data

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>
```



```
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -ls /input
Found 1 items
-rw-r--r--   3 root supergroup        271 2024-11-25 09:50 /input/sample.txt

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hadoop jar /opt/hadoop-3.2.1/share/hadoop/mapreduce/hadoop-mapr
educe-examples-3.2.1.jar wordcount /input /output
Unknown program 'wordcount??/input??/output' chosen.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
  terasort: Run the terasort
  teravalidate: Checking results of terasort
  wordcount: A map/reduce program that counts the words in the input files.
  wordmean: A map/reduce program that counts the average length of the words in the input files.
  wordmedian: A map/reduce program that counts the median length of the words in the input files.
  wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the input files.

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/
```



Browse Directory

/

Show 25 entries

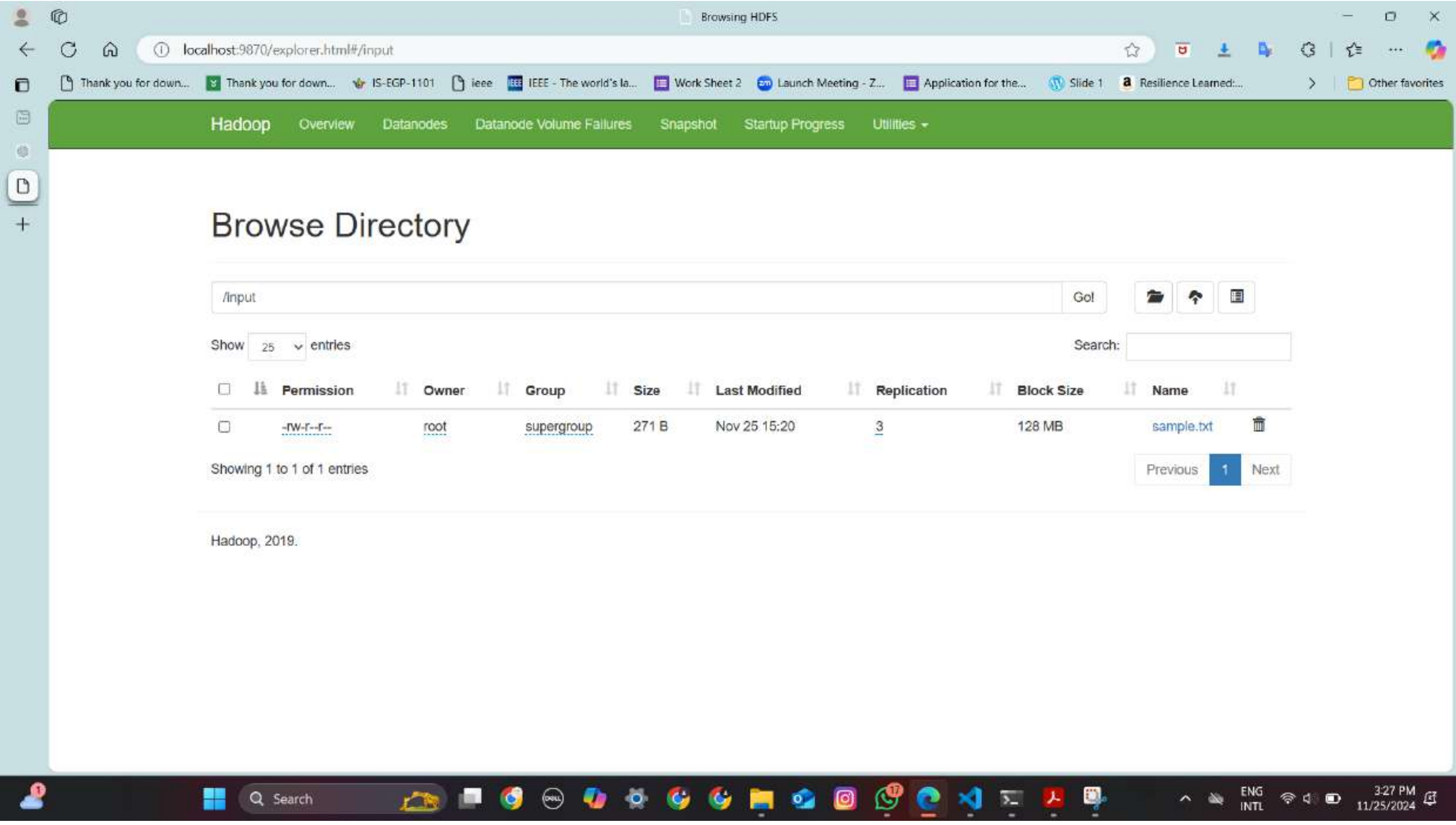
Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	root	supergroup	0 B	Nov 25 15:20	0	0 B	input

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2019.



Browse Directory

/input Go!

Show 25 entries

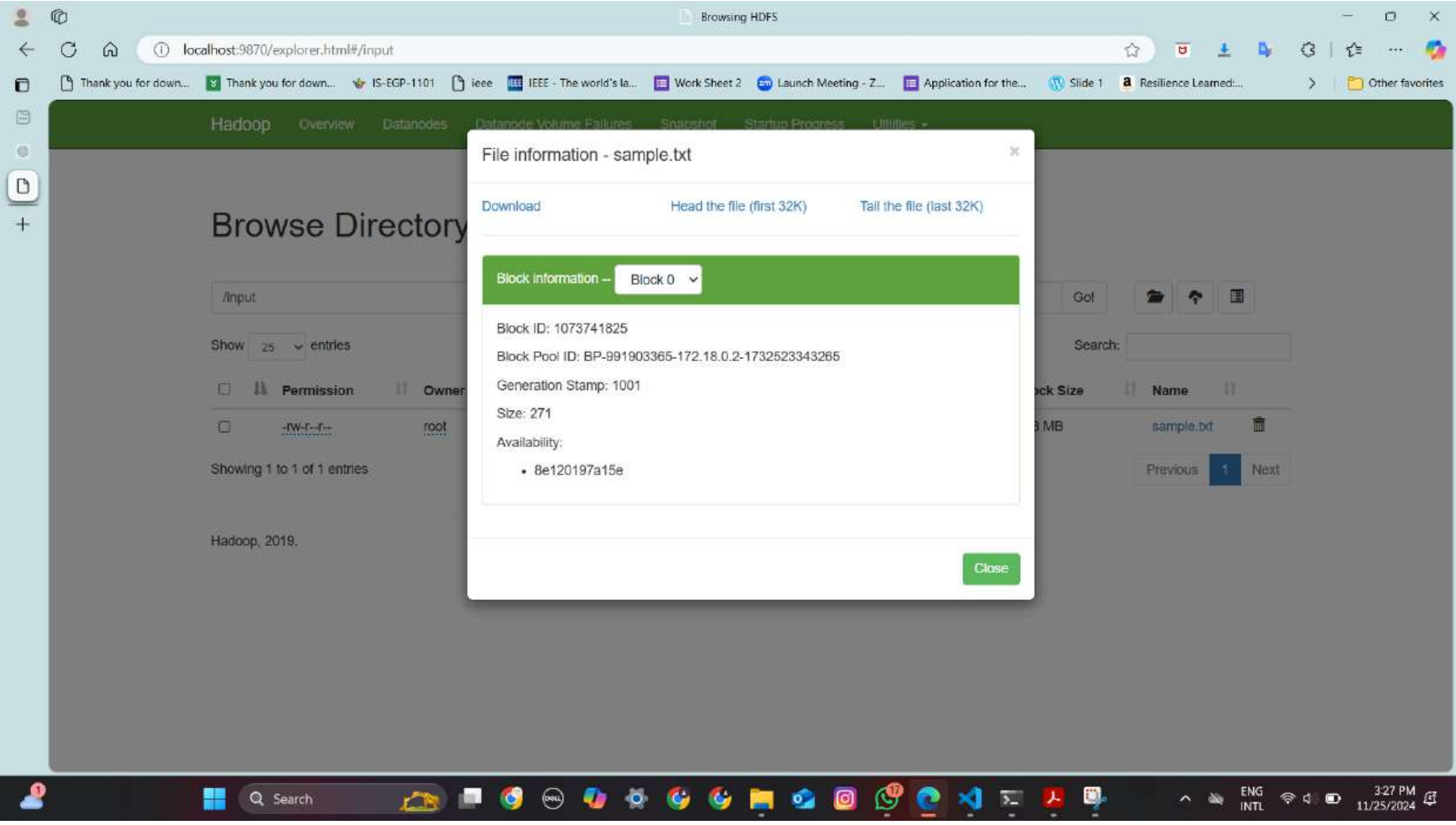
Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	root	supergroup	271 B	Nov 25 15:20	3	128 MB	sample.txt	🗑

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2019.



Browse Directory

/input

Show 25 entries

□	Permission	Owner	...ck Size	Name
□	-rw-r--r--	root	...8 MB	sample.txt

Showing 1 to 1 of 1 entries

Hadoop, 2019.

File information - sample.txt

Block information -- Block 0

Block ID: 1073741825

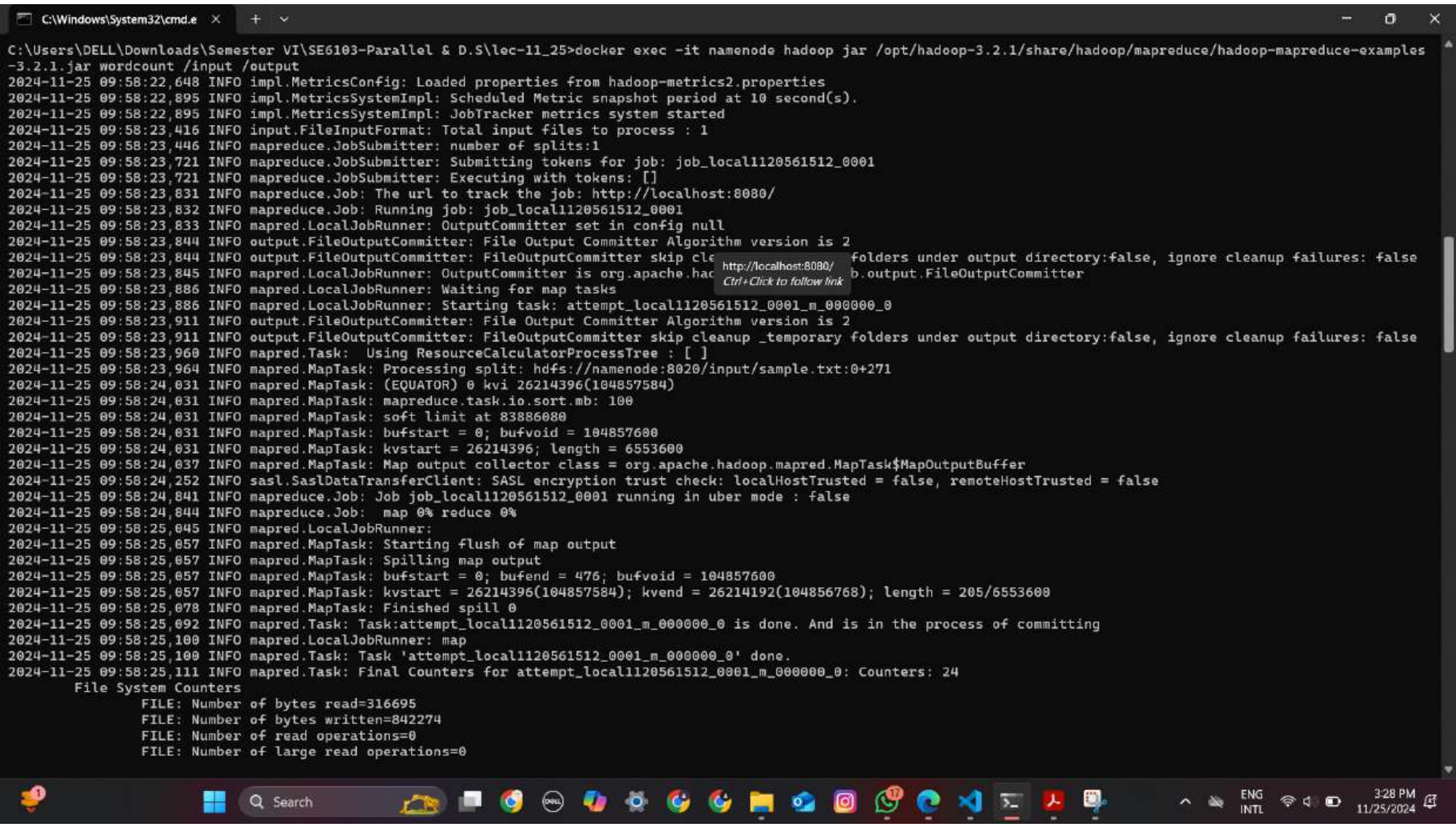
Block Pool ID: BP-991903365-172.18.0.2-1732523343265

Generation Stamp: 1001

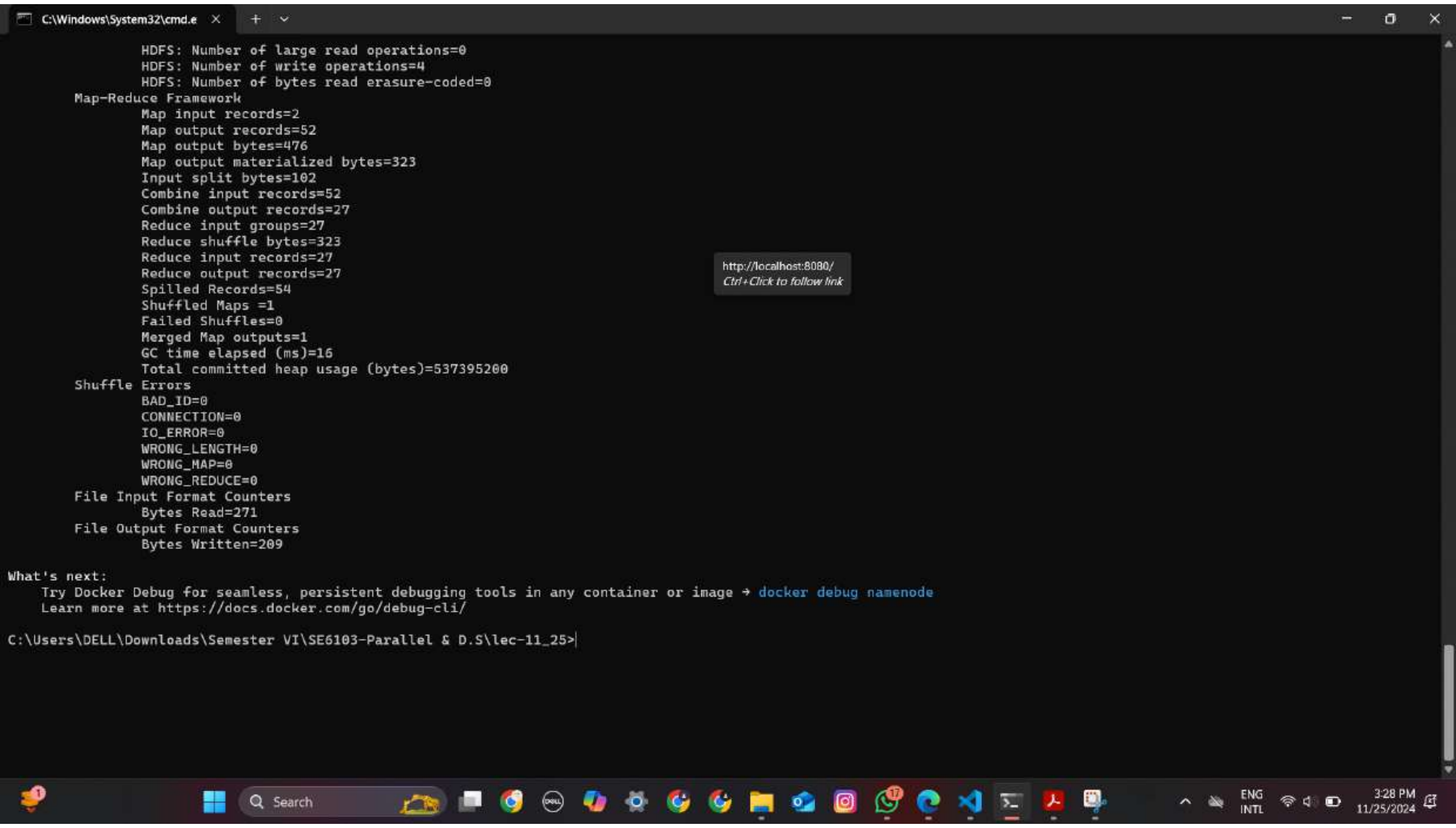
Size: 271

Availability:

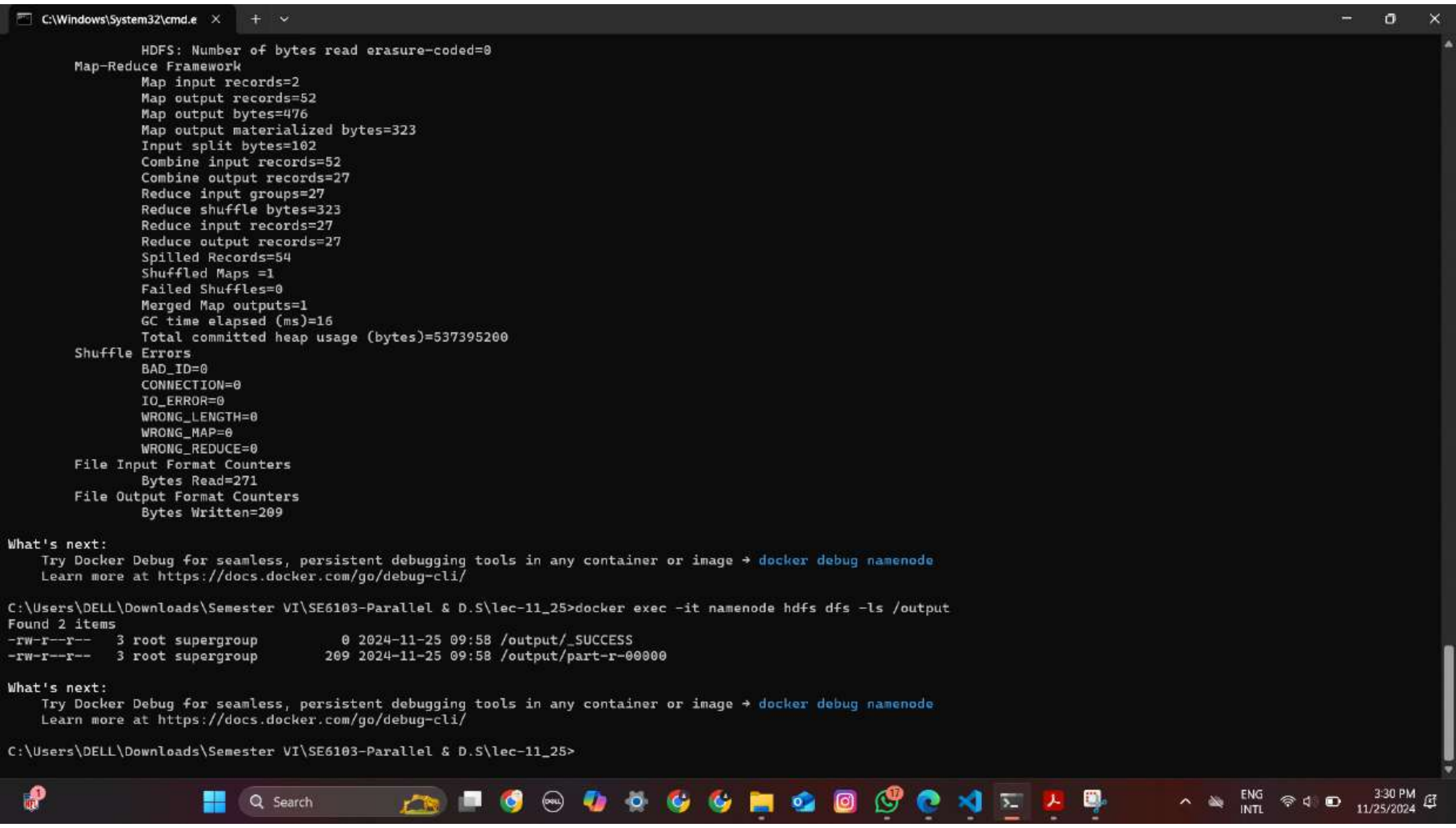
- 8e120197a15e



```
C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hadoop jar /opt/hadoop-3.2.1/share/hadoop/mapreduce/hadoop-mapreduce-examples
-3.2.1.jar wordcount /input /output
2024-11-25 09:58:22,648 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-11-25 09:58:22,895 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-25 09:58:22,895 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-25 09:58:23,416 INFO input.FileInputFormat: Total input files to process : 1
2024-11-25 09:58:23,446 INFO mapreduce.JobSubmitter: number of splits:1
2024-11-25 09:58:23,721 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1120561512_0001
2024-11-25 09:58:23,721 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-25 09:58:23,831 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-11-25 09:58:23,832 INFO mapreduce.Job: Running job: job_local1120561512_0001
2024-11-25 09:58:23,833 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-11-25 09:58:23,844 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-25 09:58:23,844 INFO output.FileOutputCommitter: FileOutputCommitter skip cle                 folders under output directory:false, ignore cleanup failures: false
2024-11-25 09:58:23,845 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hac                 b.output.FileOutputCommitter
2024-11-25 09:58:23,886 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-11-25 09:58:23,886 INFO mapred.LocalJobRunner: Starting task: attempt_local1120561512_0001_m_000000_0
2024-11-25 09:58:23,911 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-25 09:58:23,911 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-11-25 09:58:23,960 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-11-25 09:58:23,964 INFO mapred.MapTask: Processing split: hdfs://namenode:8020/input/sample.txt:0+271
2024-11-25 09:58:24,031 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-11-25 09:58:24,031 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-11-25 09:58:24,031 INFO mapred.MapTask: soft limit at 83886080
2024-11-25 09:58:24,031 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-11-25 09:58:24,031 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-11-25 09:58:24,037 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-11-25 09:58:24,252 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-25 09:58:24,841 INFO mapreduce.Job: Job job_local1120561512_0001 running in uber mode : false
2024-11-25 09:58:24,844 INFO mapreduce.Job:  map 0% reduce 0%
2024-11-25 09:58:25,045 INFO mapred.LocalJobRunner:
2024-11-25 09:58:25,057 INFO mapred.MapTask: Starting flush of map output
2024-11-25 09:58:25,057 INFO mapred.MapTask: Spilling map output
2024-11-25 09:58:25,057 INFO mapred.MapTask: bufstart = 0; bufend = 476; bufvoid = 104857600
2024-11-25 09:58:25,057 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214192(104856768); length = 205/6553600
2024-11-25 09:58:25,078 INFO mapred.MapTask: Finished spill 0
2024-11-25 09:58:25,092 INFO mapred.Task: Task:attempt_local1120561512_0001_m_000000_0 is done. And is in the process of committing
2024-11-25 09:58:25,100 INFO mapred.LocalJobRunner: map
2024-11-25 09:58:25,100 INFO mapred.Task: Task 'attempt_local1120561512_0001_m_000000_0' done.
2024-11-25 09:58:25,111 INFO mapred.Task: Final Counters for attempt_local1120561512_0001_m_000000_0: Counters: 24
        File System Counters
                FILE: Number of bytes read=316695
                FILE: Number of bytes written=842274
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

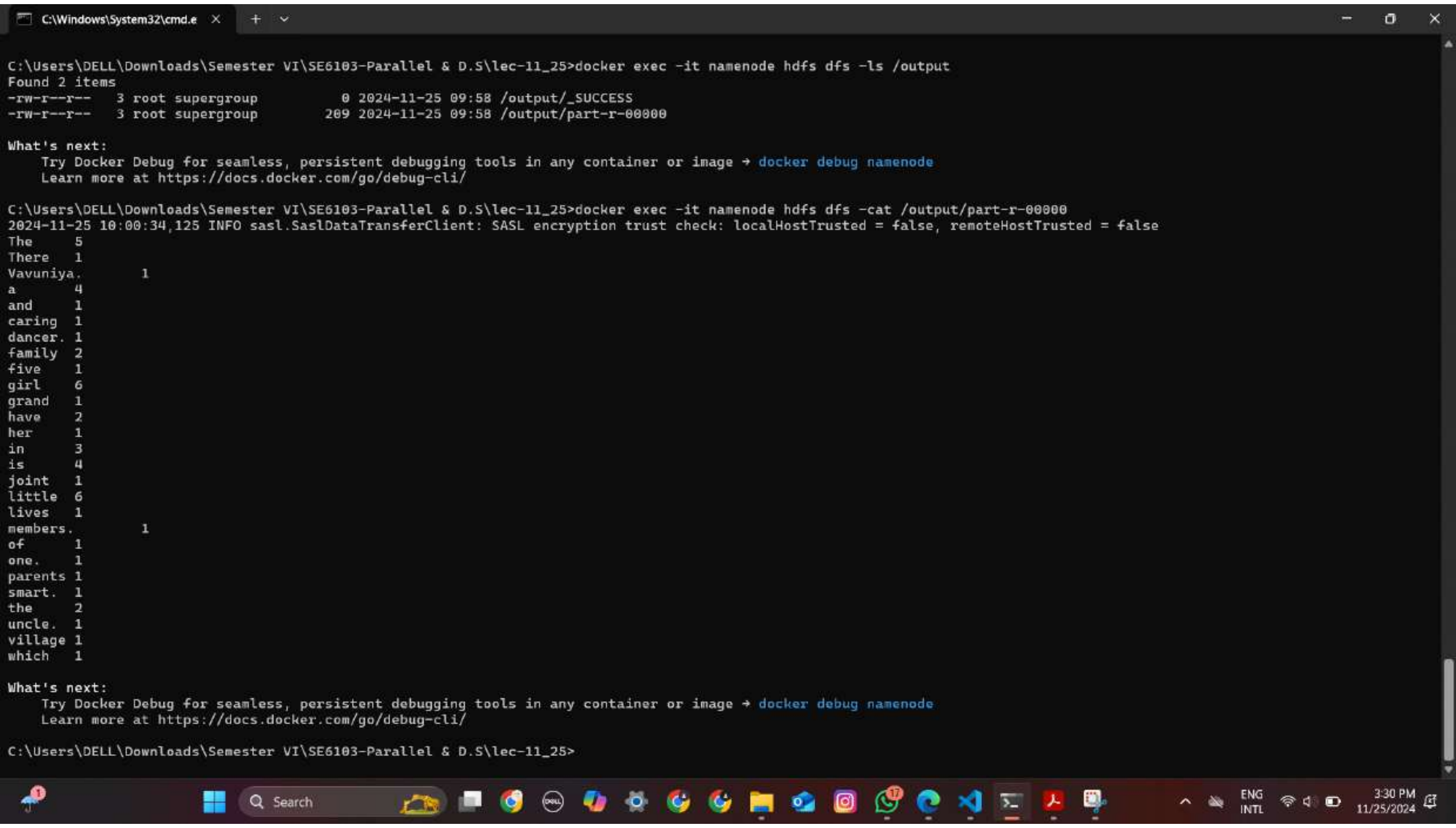


```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=2
                Map output records=52
                Map output bytes=476
                Map output materialized bytes=323
                Input split bytes=102
                Combine input records=52
                Combine output records=27
                Reduce input groups=27
                Reduce shuffle bytes=323
                Reduce input records=27
                Reduce output records=27
                Spilled Records=54
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=16
                Total committed heap usage (bytes)=537395200
        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=271
        File Output Format Counters
                Bytes Written=209

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>
```



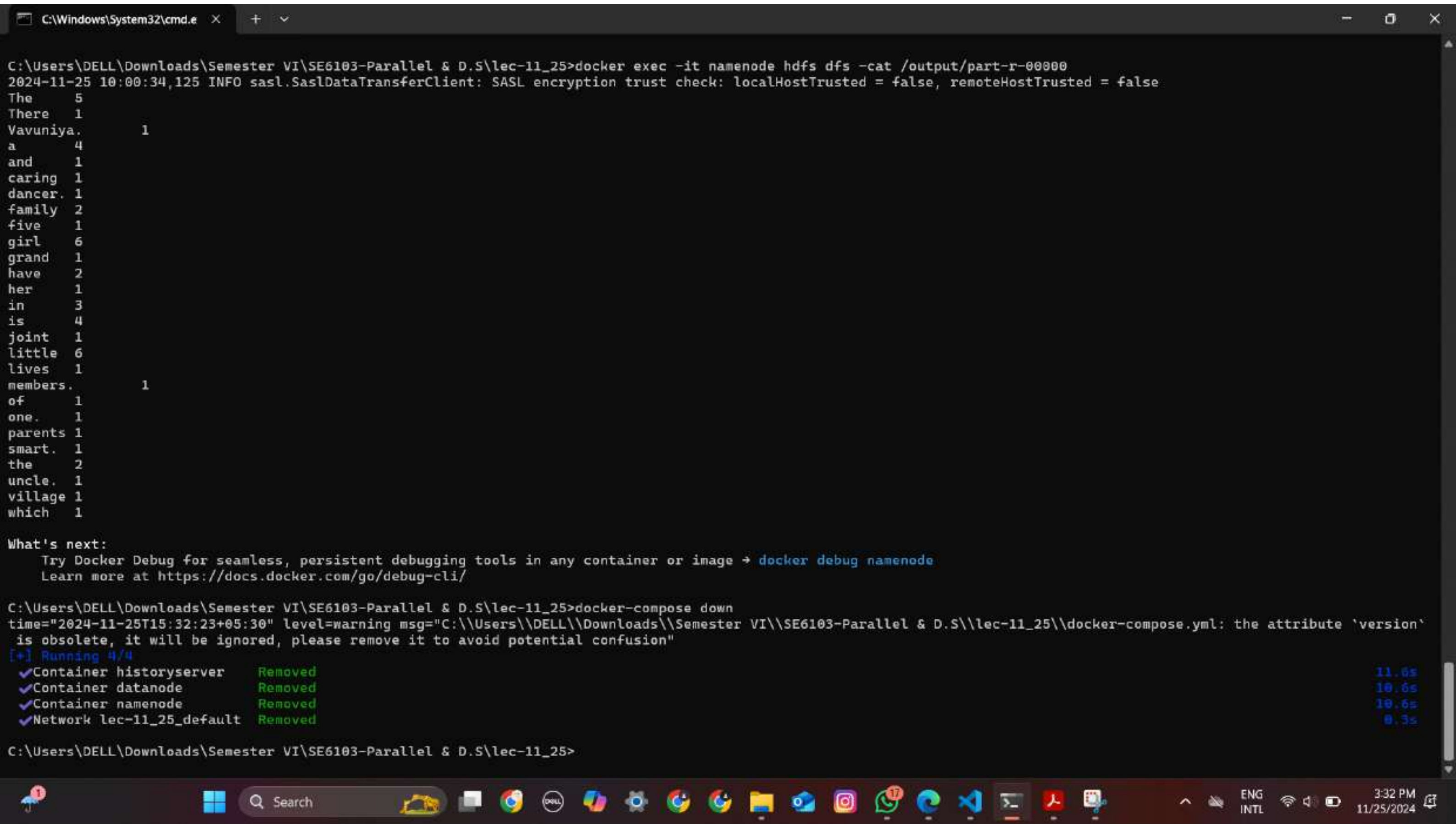
```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=2
                Map output records=52
                Map output bytes=476
                Map output materialized bytes=323
                Input split bytes=102
                Combine input records=52
                Combine output records=27
                Reduce input groups=27
                Reduce shuffle bytes=323
                Reduce input records=27
                Reduce output records=27
                Spilled Records=54
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=16
                Total committed heap usage (bytes)=537395200
        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=271
        File Output Format Counters
                Bytes Written=209

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -ls /output
Found 2 items
-rw-r--r--   3 root supergroup          0 2024-11-25 09:58 /output/_SUCCESS
-rw-r--r--   3 root supergroup        209 2024-11-25 09:58 /output/part-r-00000

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>
```

Task 3: Running a MapReduce Job



```
C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -ls /output
Found 2 items
-rw-r--r--   3 root supergroup          0 2024-11-25 09:58 /output/_SUCCESS
-rw-r--r--   3 root supergroup        209 2024-11-25 09:58 /output/part-r-00000

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -cat /output/part-r-00000
2024-11-25 10:00:34,125 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
The     5
There   1
Vavuniya.       1
a       4
and     1
caring  1
dancer. 1
family  2
five    1
girl    6
grand   1
have    2
her     1
in      3
is      4
joint   1
little  6
lives   1
members.        1
of      1
one.    1
parents 1
smart.  1
the     2
uncle.  1
village 1
which   1

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>
```

Task 4: Analyze and Clean Up



```
C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker exec -it namenode hdfs dfs -cat /output/part-r-00000
2024-11-25 10:00:34,125 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
The     5
There   1
Vavuniya.       1
a       4
and     1
caring  1
dancer. 1
family  2
five    1
girl    6
grand   1
have    2
her     1
in      3
is      4
joint   1
little  6
lives   1
members.        1
of      1
one.    1
parents 1
smart.  1
the     2
uncle.  1
village 1
which   1

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>docker-compose down
time="2024-11-25T15:32:23+05:30" level=warning msg="C:\\Users\\DELL\\Downloads\\Semester VI\\SE6103-Parallel & D.S\\lec-11_25\\docker-compose.yml: the attribute `version`
 is obsolete, it will be ignored, please remove it to avoid potential confusion"
[+] Running 4/4
 ✔Container historyserver    Removed                                                                                              11.6s
 ✔Container datanode         Removed                                                                                              10.6s
 ✔Container namenode         Removed                                                                                              10.6s
 ✔Network lec-11_25_default  Removed                                                                                               0.3s

C:\Users\DELL\Downloads\Semester VI\SE6103-Parallel & D.S\lec-11_25>
```