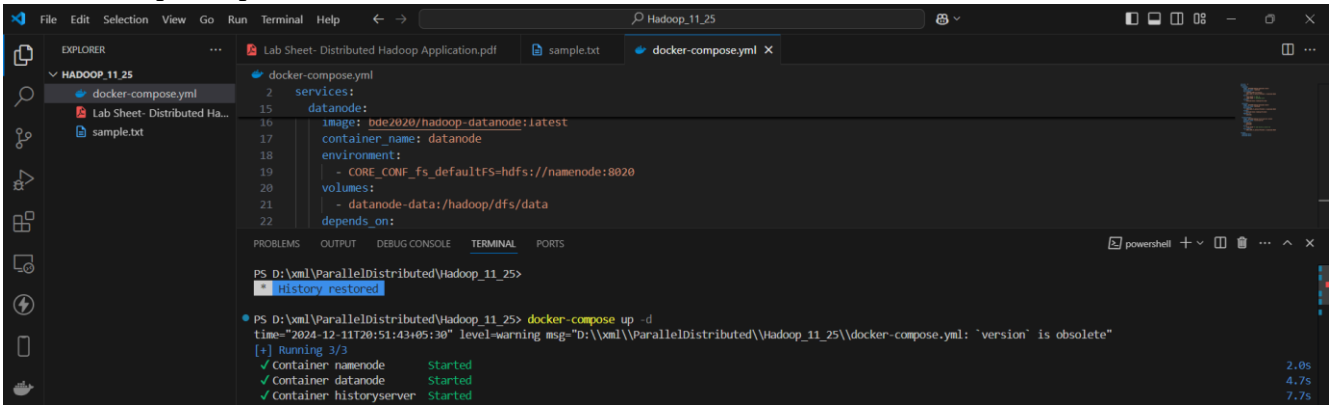
SE6103 Parallel and Distributed Systems

Lab Sheet: Distributed Hadoop Application

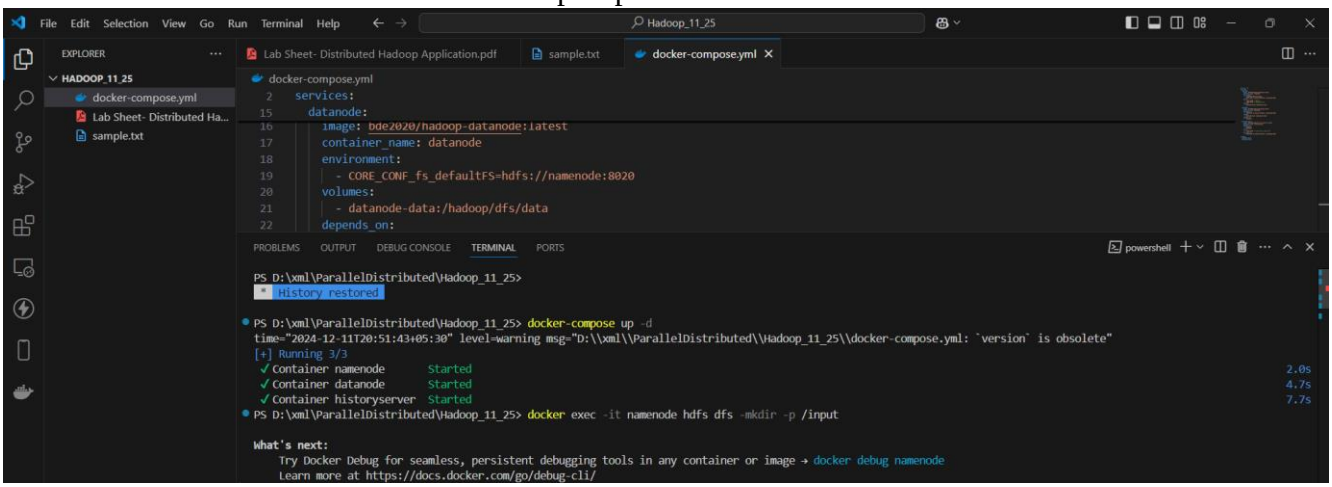
19APSE4325 – Tharindu Lakmal

1. docker-compose.yml

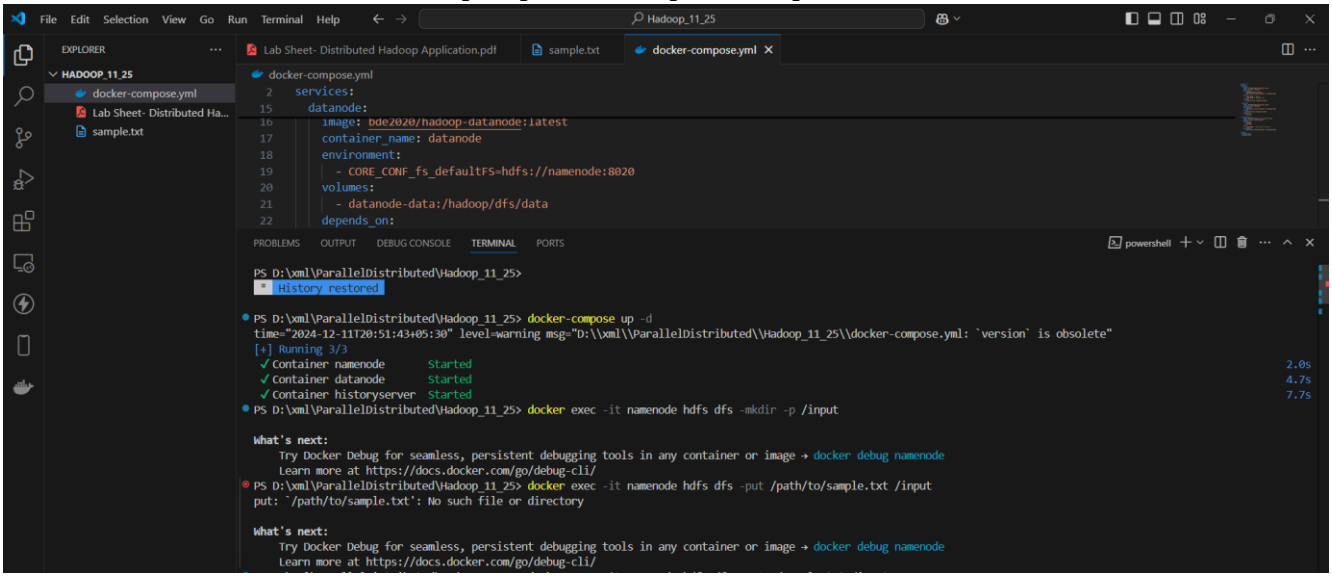
2. docker-compose up -d



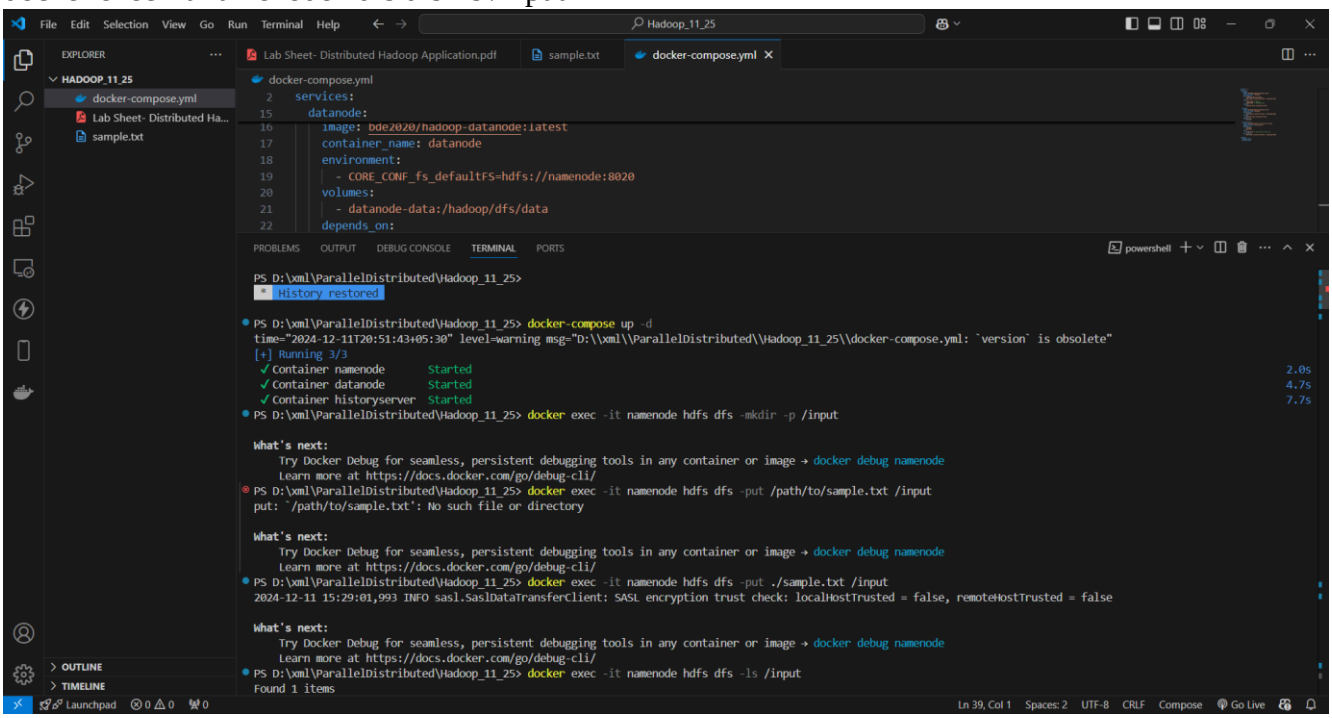
3. docker exec -it namenode hdfs dfs -mkdir -p /input

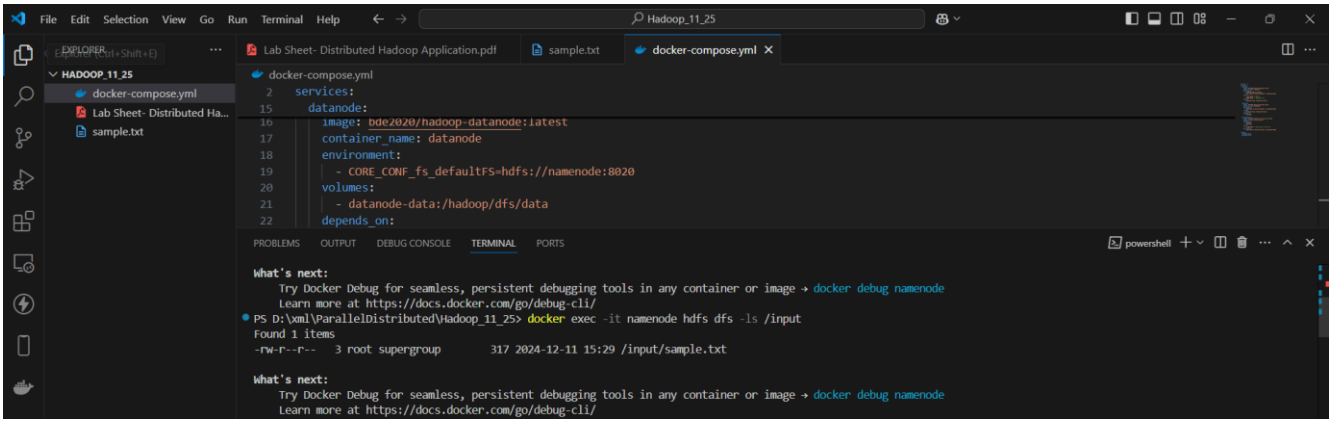


4. docker exec -it namenode hdfs dfs -put /path/to/sample.txt /input

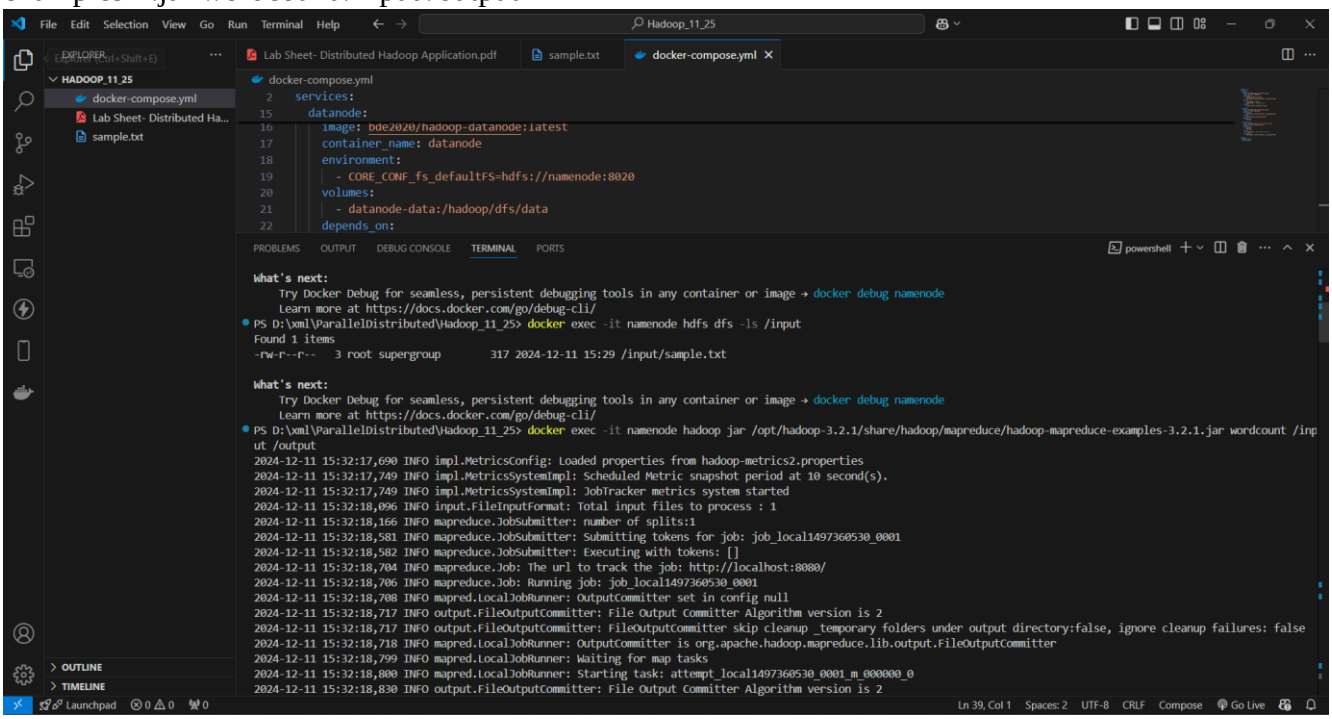


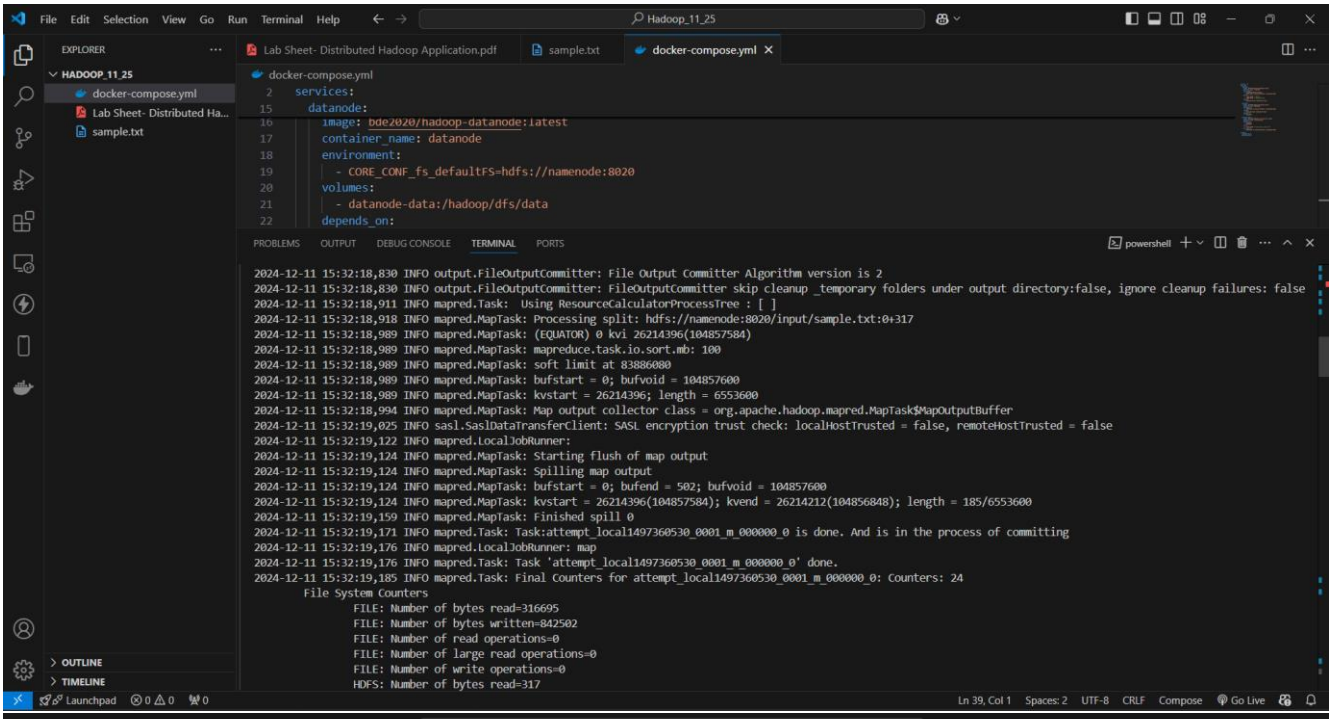
5. docker exec -it namenode hdfs dfs -ls /input

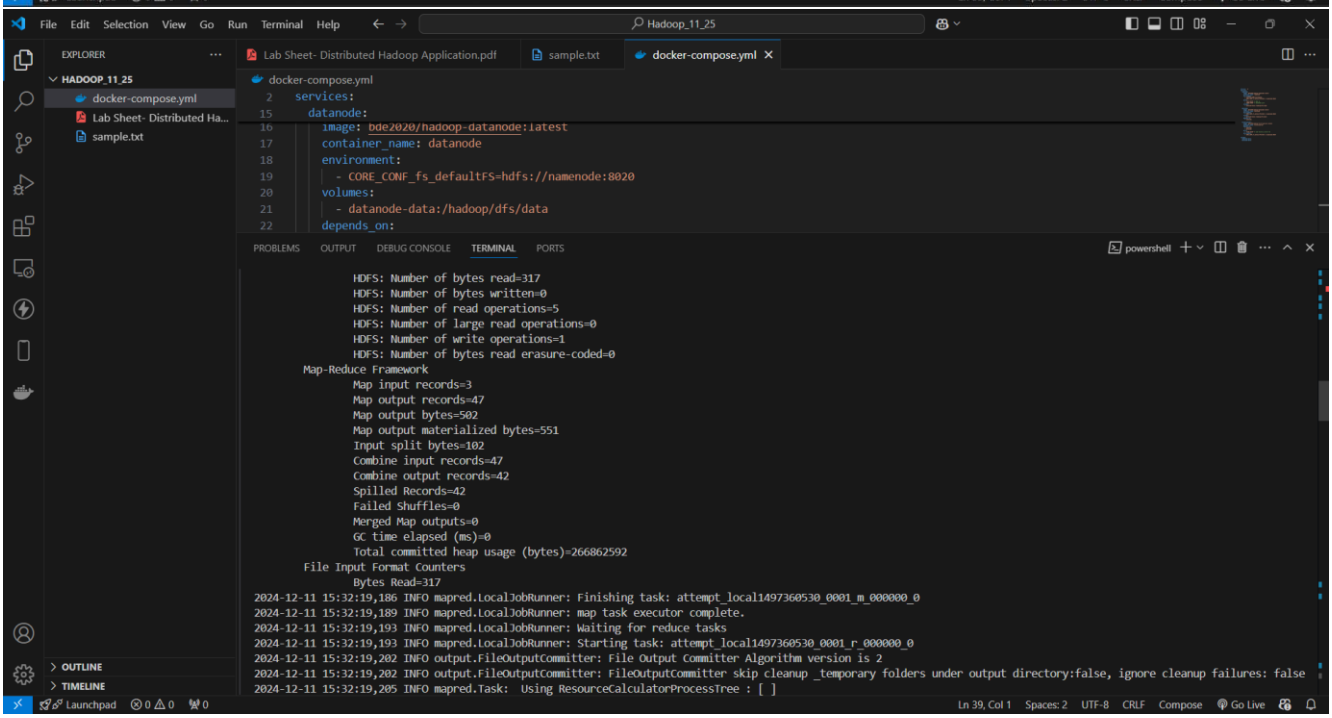


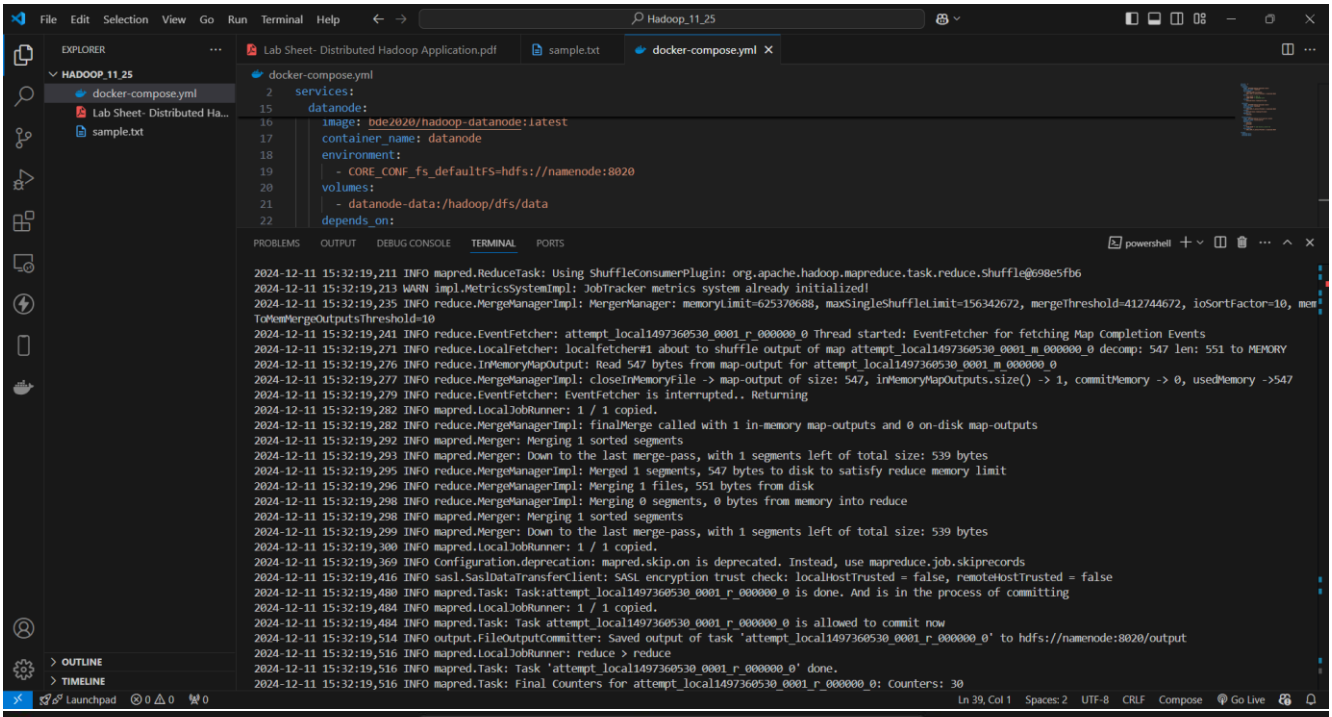


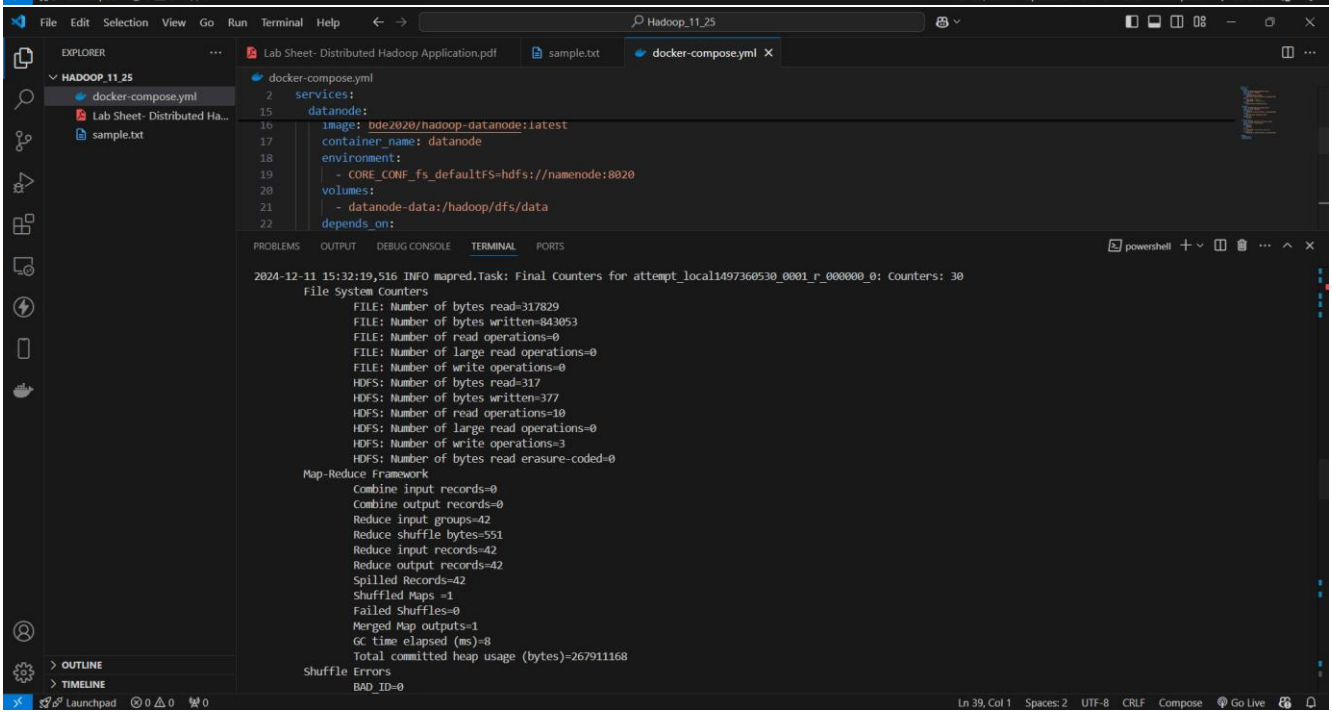
6. docker exec -it namenode hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar wordcount /input /output

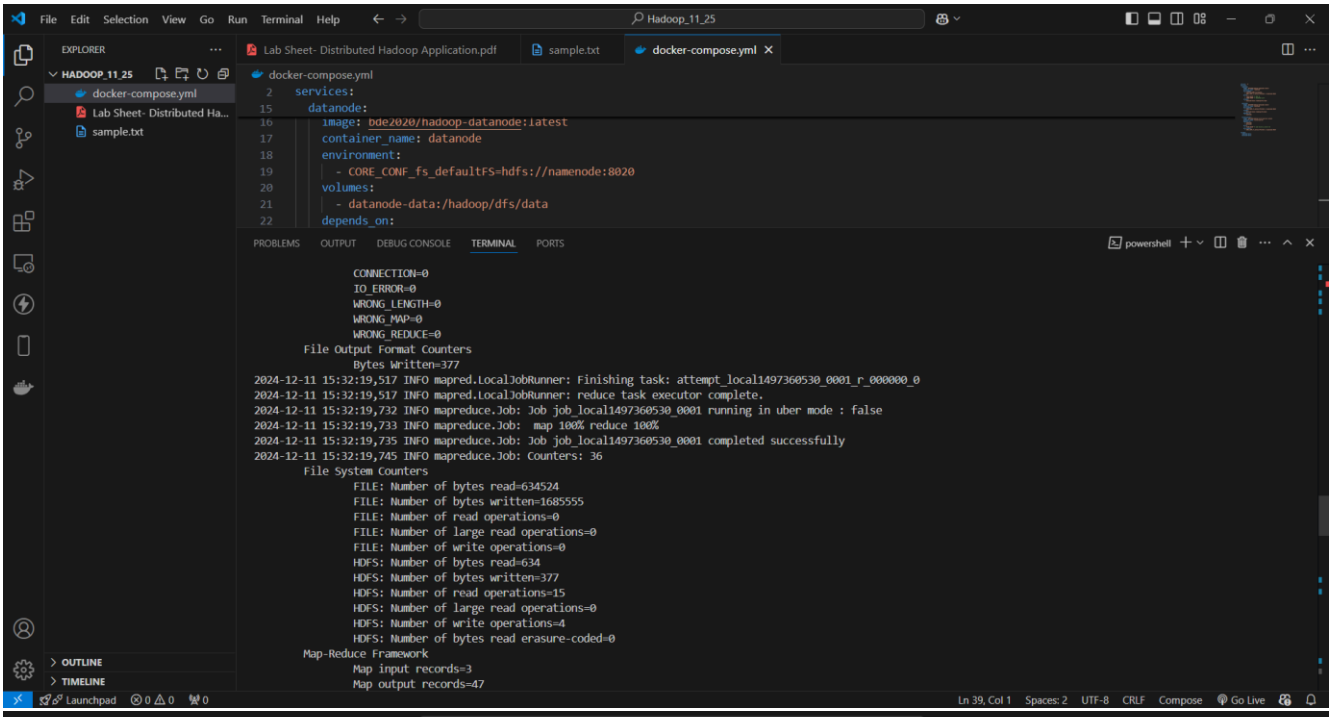


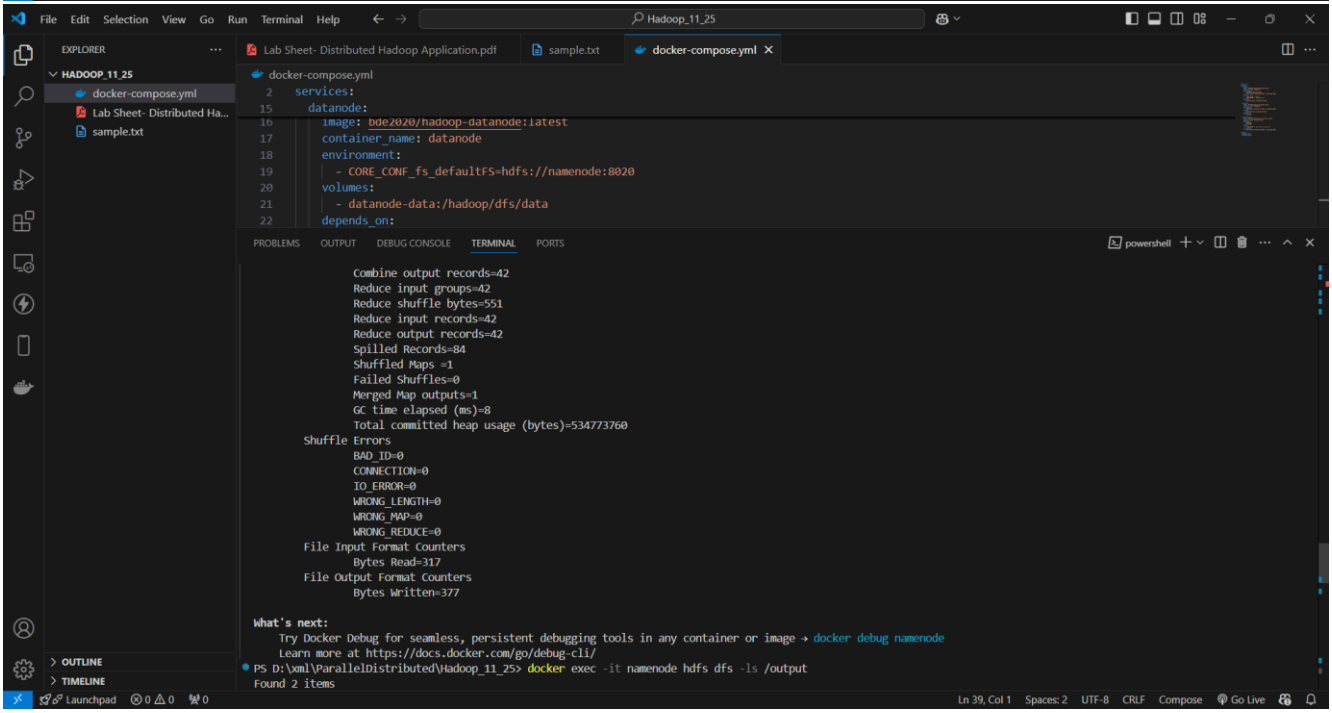

File Edit Selection View Go Run Terminal Help
Hadoop_11_25
EXPLORER
HADOOP_11_25
docker-compose.yml
Lab Sheet- Distributed Ha...
sample.txt
Lab Sheet- Distributed Hadoop Application.pdf
sample.txt
docker-compose.yml
docker-compose.yml
services:
datanode:
image: bde2020/hadoop-datanode:latest
container_name: datanode
environment:
- CORE_CONF_fs_defaultFS=hdfs://namenode:8020
volumes:
- datanode-data:/hadoop/dfs/data
depends_on:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
powershell
2024-12-11 15:32:18,830 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-12-11 15:32:18,830 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-12-11 15:32:18,911 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2024-12-11 15:32:18,918 INFO mapred.MapTask: Processing split: hdfs://namenode:8020/input/sample.txt:0+317
2024-12-11 15:32:18,989 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-12-11 15:32:18,989 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-12-11 15:32:18,989 INFO mapred.MapTask: soft limit at 83886080
2024-12-11 15:32:18,989 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-12-11 15:32:18,989 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-12-11 15:32:18,994 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-12-11 15:32:19,025 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-11 15:32:19,122 INFO mapred.LocalJobRunner:
2024-12-11 15:32:19,124 INFO mapred.MapTask: Starting flush of map output
2024-12-11 15:32:19,124 INFO mapred.MapTask: Spilling map output
2024-12-11 15:32:19,124 INFO mapred.MapTask: bufstart = 0; bufend = 502; bufvoid = 104857600
2024-12-11 15:32:19,124 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214212(104856848); length = 185/6553600
2024-12-11 15:32:19,159 INFO mapred.MapTask: Finished spill 0
2024-12-11 15:32:19,171 INFO mapred.Task: Task:attempt_local1497360530_0001_m_000000_0 is done. And is in the process of committing
2024-12-11 15:32:19,176 INFO mapred.LocalJobRunner: map
2024-12-11 15:32:19,176 INFO mapred.Task: Task 'attempt_local1497360530_0001_m_000000_0' done.
2024-12-11 15:32:19,185 INFO mapred.Task: Final Counters for attempt_local1497360530_0001_m_000000_0: Counters: 24
File System Counters
FILE: Number of bytes read=316695
FILE: Number of bytes written=842502
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=317
OUTLINE
TIMELINE
Launchpad
Ln 39, Col 1 Spaces: 2 UTF-8 CRLF Compose Go Live


File Edit Selection View Go Run Terminal Help
Hadoop_11_25
EXPLORER
HADOOP_11_25
docker-compose.yml
Lab Sheet- Distributed Ha...
sample.txt
Lab Sheet- Distributed Hadoop Application.pdf
sample.txt
docker-compose.yml
docker-compose.yml
services:
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:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
powershell
HDFS: Number of bytes read=317
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=3
Map output records=47
Map output bytes=502
Map output materialized bytes=551
Input split bytes=102
Combine input records=47
Combine output records=42
Spilled Records=42
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=266862592
File Input Format Counters
Bytes Read=317
2024-12-11 15:32:19,186 INFO mapred.LocalJobRunner: Finishing task: attempt_local1497360530_0001_m_000000_0
2024-12-11 15:32:19,189 INFO mapred.LocalJobRunner: map task executor complete.
2024-12-11 15:32:19,193 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2024-12-11 15:32:19,193 INFO mapred.LocalJobRunner: Starting task: attempt_local1497360530_0001_r_000000_0
2024-12-11 15:32:19,202 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-12-11 15:32:19,202 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-12-11 15:32:19,205 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
OUTLINE
TIMELINE
Launchpad
Ln 39, Col 1 Spaces: 2 UTF-8 CRLF Compose Go Live

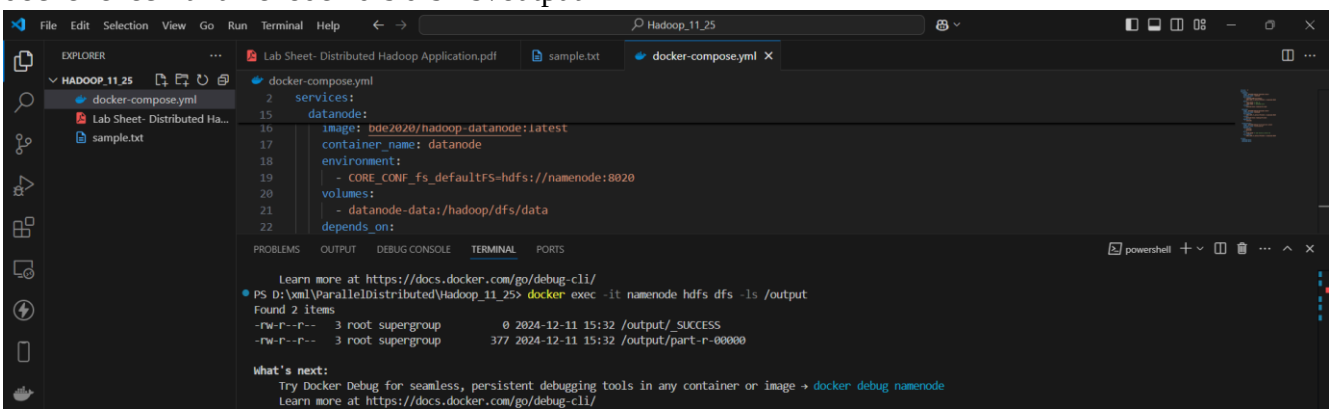

File Edit Selection View Go Run Terminal Help
Hadoop_11_25
EXPLORER
HADOOP_11_25
docker-compose.yml
Lab Sheet- Distributed Ha...
sample.txt
Lab Sheet- Distributed Hadoop Application.pdf
sample.txt
docker-compose.yml
services:
datanode:
image: bde2020/hadoop-datanode:latest
container_name: datanode
environment:
- CORE_CONF_fs_defaultFS=hdfs://namenode:8020
volumes:
- datanode-data:/hadoop/dfs/data
depends_on:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
powershell
2024-12-11 15:32:19,211 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@698e5fb6
2024-12-11 15:32:19,213 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-12-11 15:32:19,235 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=625370688, maxSingleShuffleLimit=156342672, mergeThreshold=412744672, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2024-12-11 15:32:19,241 INFO reduce.EventFetcher: attempt_local1497360530_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2024-12-11 15:32:19,271 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1497360530_0001_m_000000_0 decomp: 547 len: 551 to MEMORY
2024-12-11 15:32:19,276 INFO reduce.InMemoryMapOutput: Read 547 bytes from map-output for attempt_local1497360530_0001_m_000000_0
2024-12-11 15:32:19,277 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 547, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->547
2024-12-11 15:32:19,279 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2024-12-11 15:32:19,282 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-12-11 15:32:19,282 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2024-12-11 15:32:19,292 INFO mapred.Merger: Merging 1 sorted segments
2024-12-11 15:32:19,293 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 539 bytes
2024-12-11 15:32:19,295 INFO reduce.MergeManagerImpl: Merged 1 segments, 547 bytes to disk to satisfy reduce memory limit
2024-12-11 15:32:19,296 INFO reduce.MergeManagerImpl: Merging 1 files, 551 bytes from disk
2024-12-11 15:32:19,298 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2024-12-11 15:32:19,298 INFO mapred.Merger: Merging 1 sorted segments
2024-12-11 15:32:19,299 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 539 bytes
2024-12-11 15:32:19,300 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-12-11 15:32:19,369 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2024-12-11 15:32:19,416 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-11 15:32:19,480 INFO mapred.Task: Task:attempt_local1497360530_0001_r_000000_0 is done. And is in the process of committing
2024-12-11 15:32:19,484 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-12-11 15:32:19,484 INFO mapred.Task: Task attempt_local1497360530_0001_r_000000_0 is allowed to commit now
2024-12-11 15:32:19,514 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1497360530_0001_r_000000_0' to hdfs://namenode:8020/output
2024-12-11 15:32:19,516 INFO mapred.LocalJobRunner: reduce > reduce
2024-12-11 15:32:19,516 INFO mapred.Task: Task 'attempt_local1497360530_0001_r_000000_0' done.
2024-12-11 15:32:19,516 INFO mapred.Task: Final Counters for attempt_local1497360530_0001_r_000000_0: Counters: 30
OUTLINE
TIMELINE
Launchpad
Ln 39, Col 1 Spaces: 2 UTF-8 CRLF Compose Go Live


File Edit Selection View Go Run Terminal Help
Hadoop_11_25
EXPLORER
HADOOP_11_25
docker-compose.yml
Lab Sheet- Distributed Ha...
sample.txt
Lab Sheet- Distributed Hadoop Application.pdf
sample.txt
docker-compose.yml
services:
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:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
powershell
2024-12-11 15:32:19,516 INFO mapred.Task: Final Counters for attempt_local1497360530_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=317829
FILE: Number of bytes written=843053
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=317
HDFS: Number of bytes written=377
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=42
Reduce shuffle bytes=551
Reduce input records=42
Reduce output records=42
Spilled Records=42
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=8
Total committed heap usage (bytes)=267911168
Shuffle Errors
BAD_ID=0
OUTLINE
TIMELINE
Launchpad
Ln 39, Col 1 Spaces: 2 UTF-8 CRLF Compose Go Live

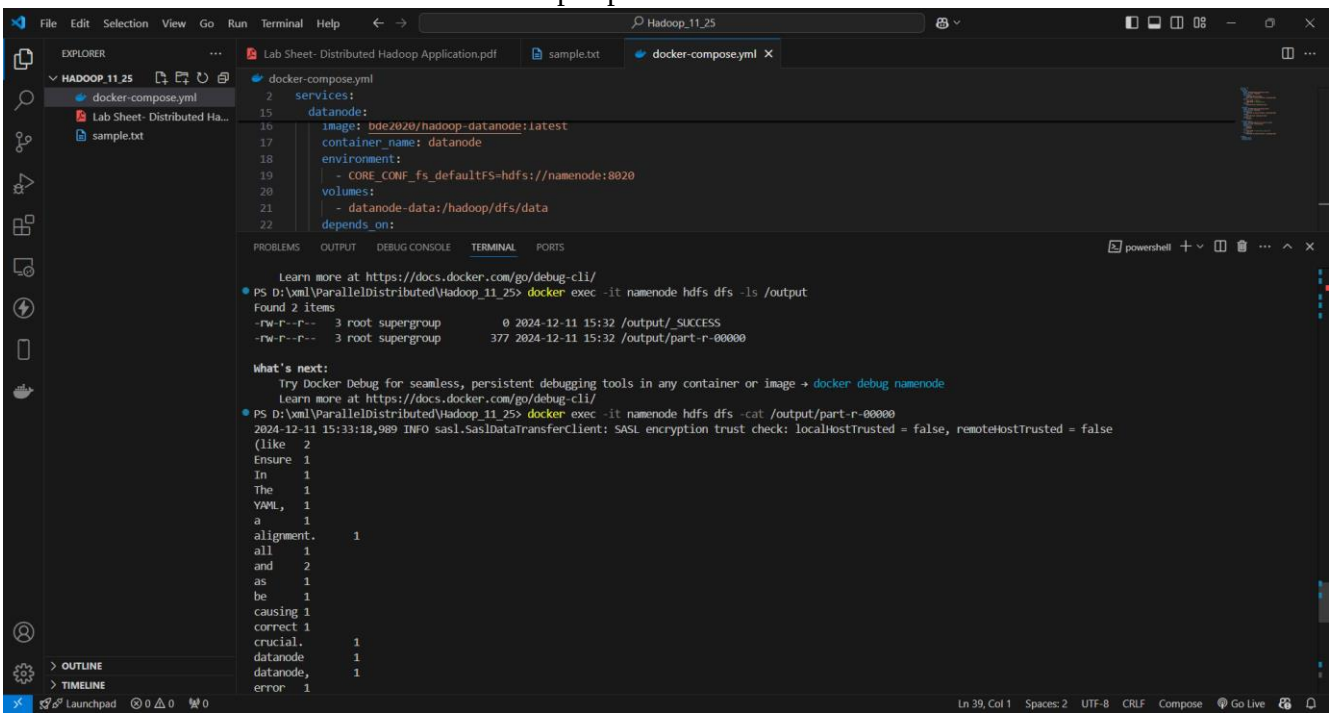

File Edit Selection View Go Run Terminal Help
Hadoop_11_25
EXPLORER
HADOOP_11_25
docker-compose.yml
Lab Sheet- Distributed Ha...
sample.txt
Lab Sheet- Distributed Hadoop Application.pdf
sample.txt
docker-compose.yml
services:
datanode:
image: bde2020/hadoop-datanode:latest
container_name: datanode
environment:
- CORE_CONF_fs_defaultFS=hdfs://namenode:8020
volumes:
- datanode-data:/hadoop/dfs/data
depends_on:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
powershell
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=377
2024-12-11 15:32:19,517 INFO mapred.LocalJobRunner: Finishing task: attempt_local1497360530_0001_r_000000_0
2024-12-11 15:32:19,517 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-12-11 15:32:19,732 INFO mapreduce.Job: Job job_local1497360530_0001 running in uber mode : false
2024-12-11 15:32:19,733 INFO mapreduce.Job: map 100% reduce 100%
2024-12-11 15:32:19,735 INFO mapreduce.Job: Job job_local1497360530_0001 completed successfully
2024-12-11 15:32:19,745 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=634524
FILE: Number of bytes written=1685555
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=634
HDFS: Number of bytes written=377
HDFS: Number of read operations=15
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=3
Map output records=47
OUTLINE
TIMELINE
Launchpad
Ln 39, Col 1 Spaces: 2 UTF-8 CRLF Compose Go Live

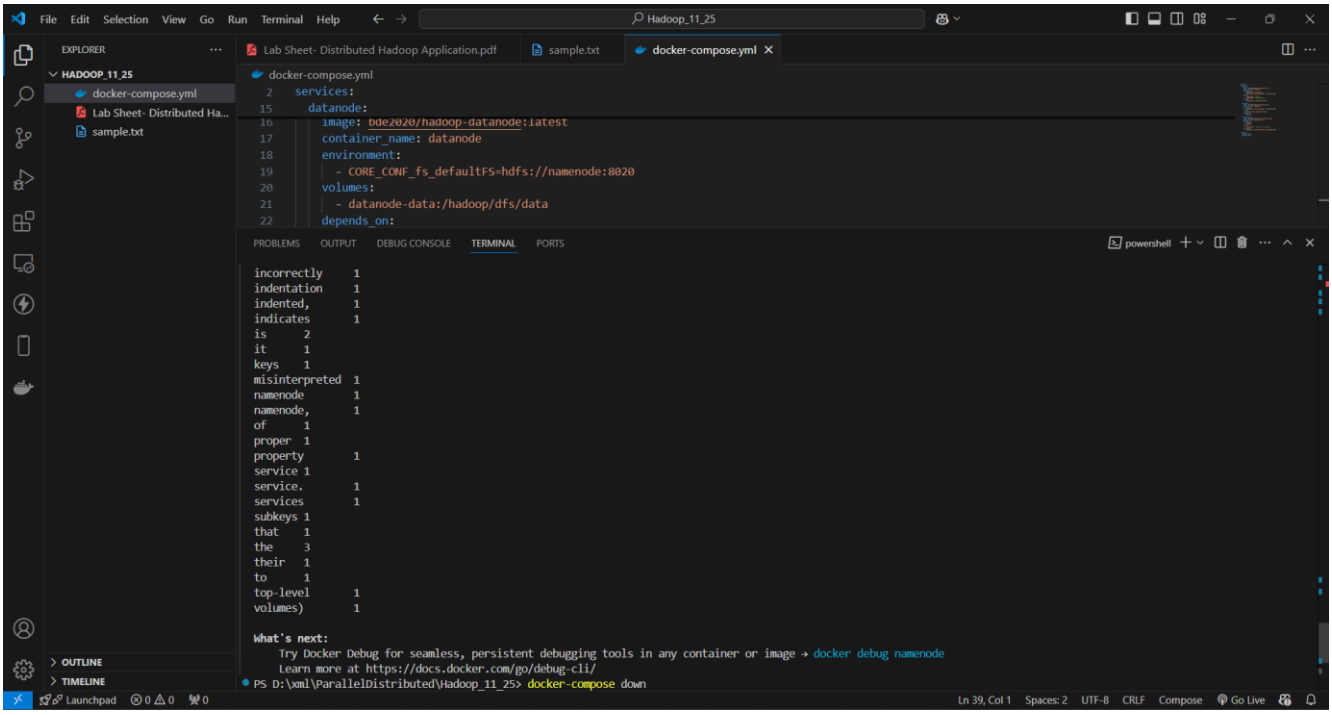
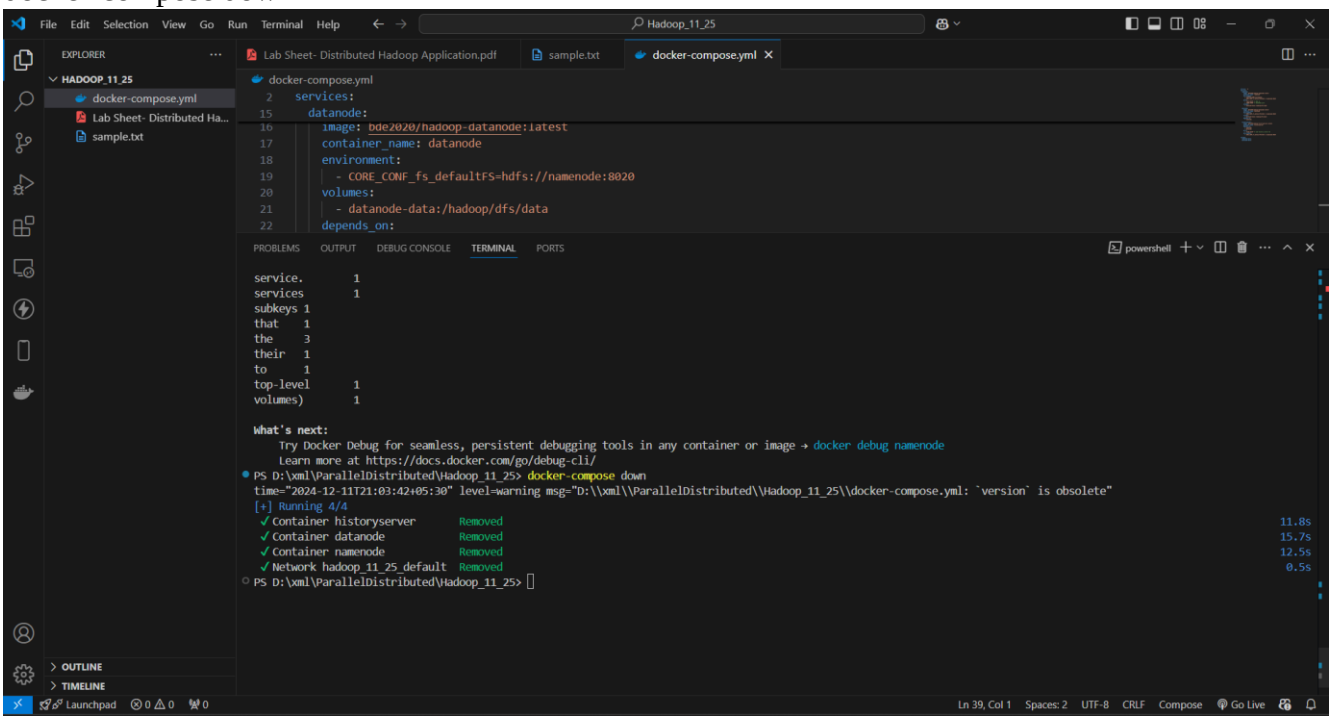

File Edit Selection View Go Run Terminal Help
Hadoop_11_25
EXPLORER
HADOOP_11_25
docker-compose.yml
Lab Sheet- Distributed Ha...
sample.txt
Lab Sheet- Distributed Hadoop Application.pdf
sample.txt
docker-compose.yml
services:
datanode:
image: bde2020/hadoop-datanode:latest
container_name: datanode
environment:
- CORE_CONF_fs_defaultFS=hdfs://namenode:8020
volumes:
- datanode-data:/hadoop/dfs/data
depends_on:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
powershell
Combine output records=42
Reduce input groups=42
Reduce shuffle bytes=551
Reduce input records=42
Reduce output records=42
Spilled Records=84
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=8
Total committed heap usage (bytes)=534773760
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=317
File Output Format Counters
Bytes Written=377
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/
PS D:\xml\ParallelDistributed\Hadoop_11_25> docker exec -it namenode hdfs dfs -ls /output
Found 2 items
OUTLINE
TIMELINE
Launchpad
Ln 39, Col 1 Spaces: 2 UTF-8 CRLF Compose Go Live

7. docker exec -it namenode hdfs dfs -ls /output



8. docker exec -it namenode hdfs dfs -cat /output/part-r-00000





9. docker-compose down