Assignment 02

Sabaragamuwa University of Sri Lanka

Faculty of Computing

Department of Software Engineering

Parallel & Distributed Systems

SE6103

Name Janith Charutha

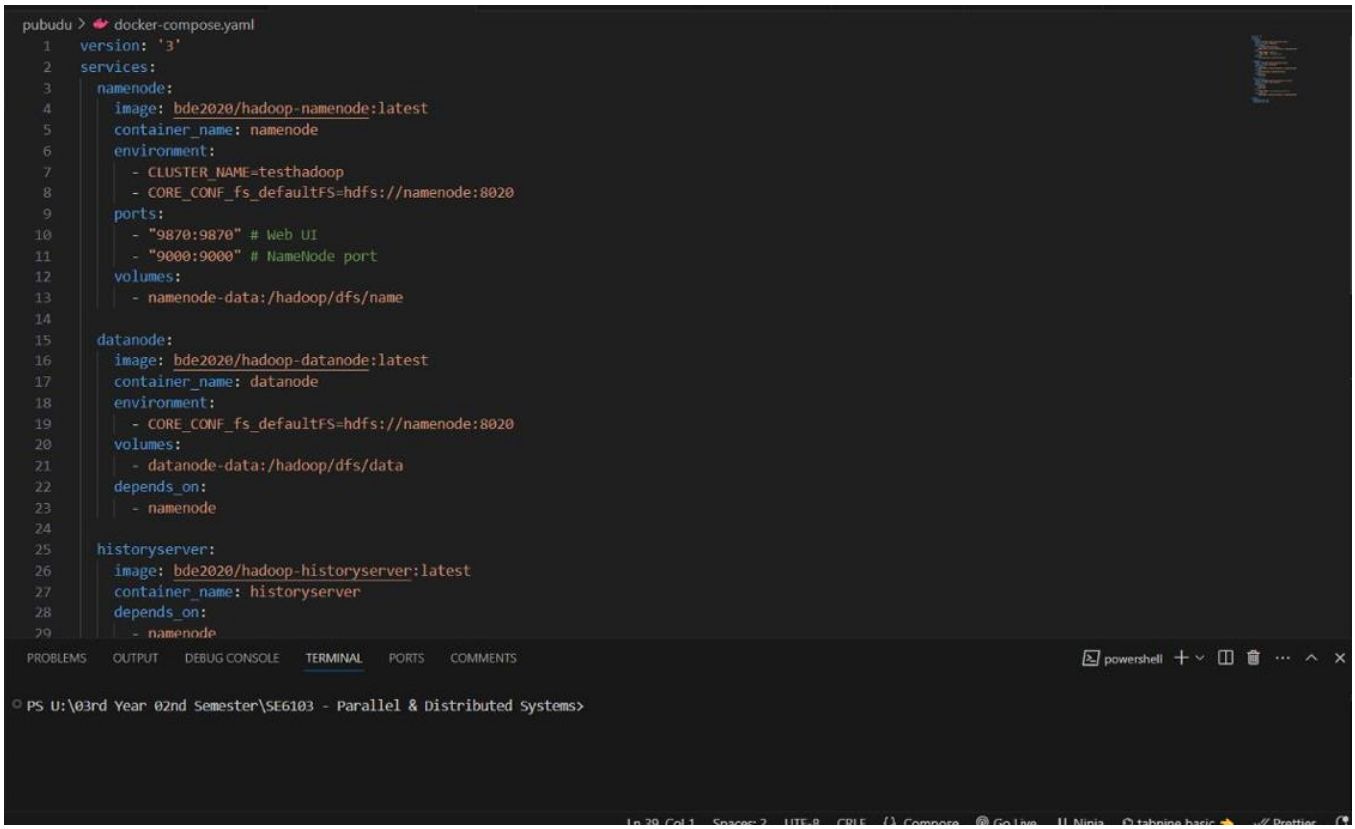
Reg.no 19APSE4273

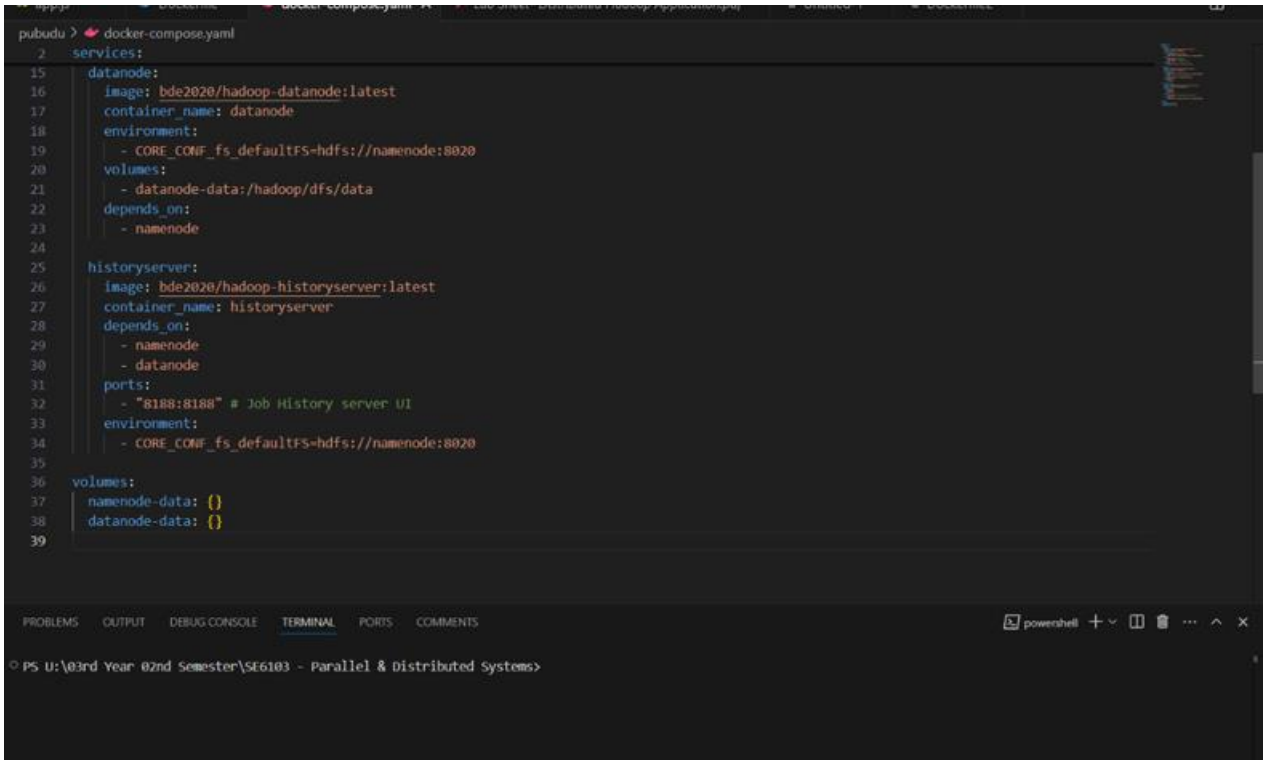
Academic year 3nd Year 2st Semester

Due Date 25-11-2024

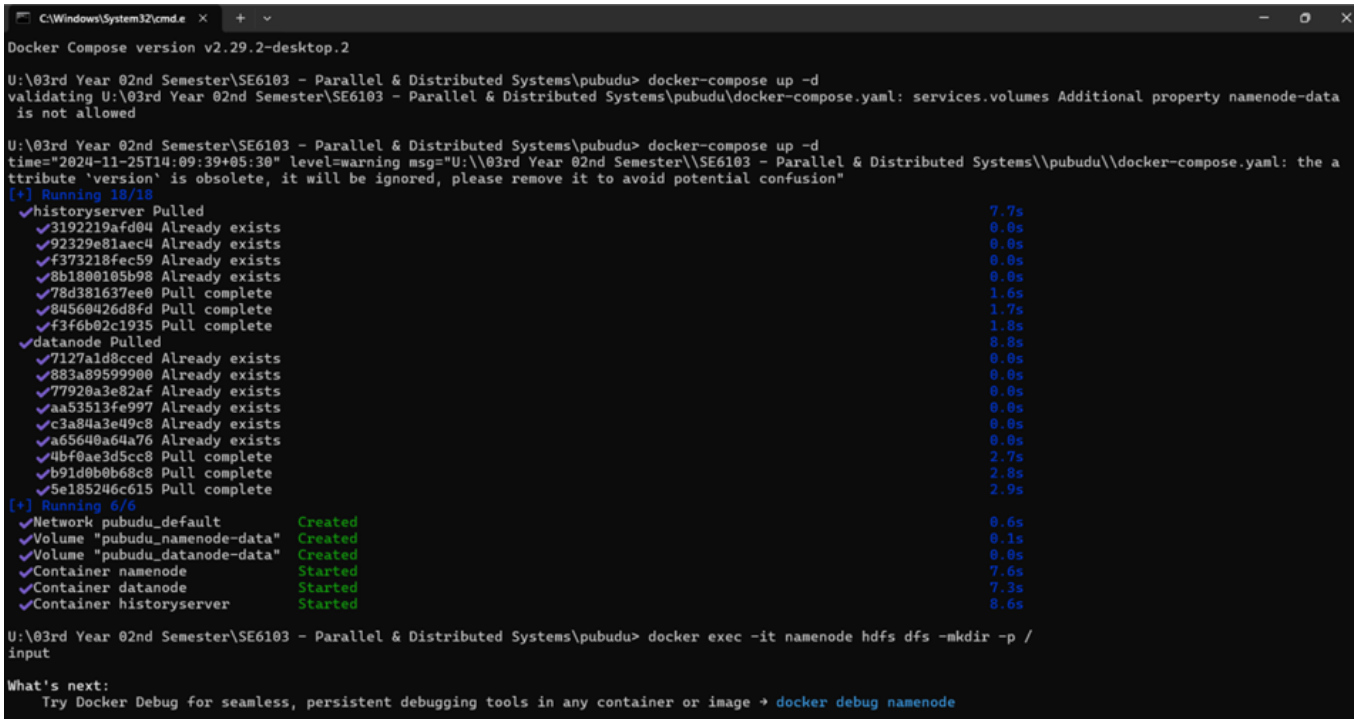
Task 01. Setting Up the Distributed Hadoop Cluster

Step 01: Prepare the Docker Compose File

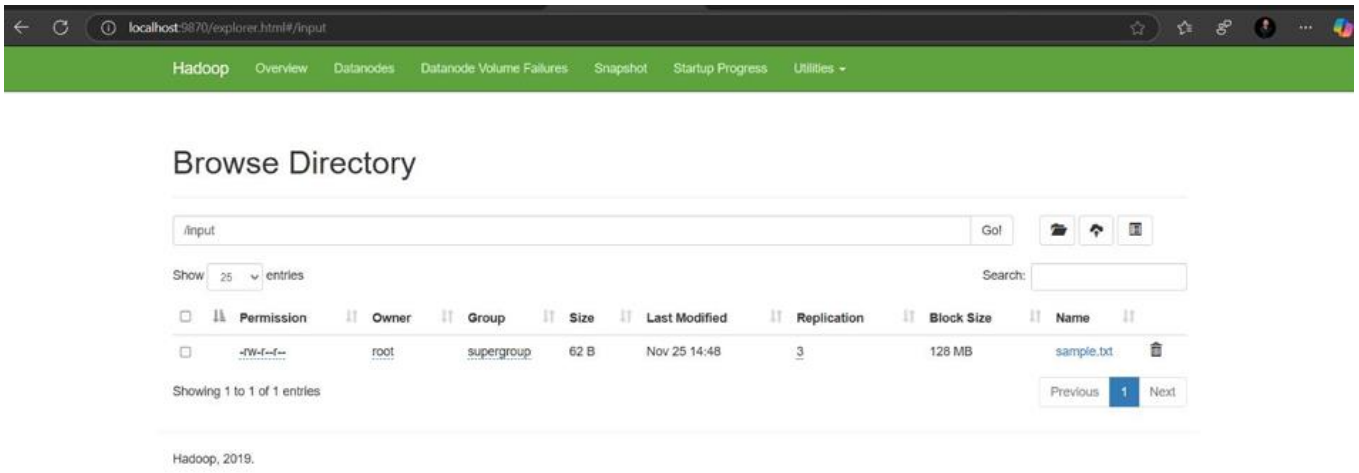




Step 02: Deploy the Cluster

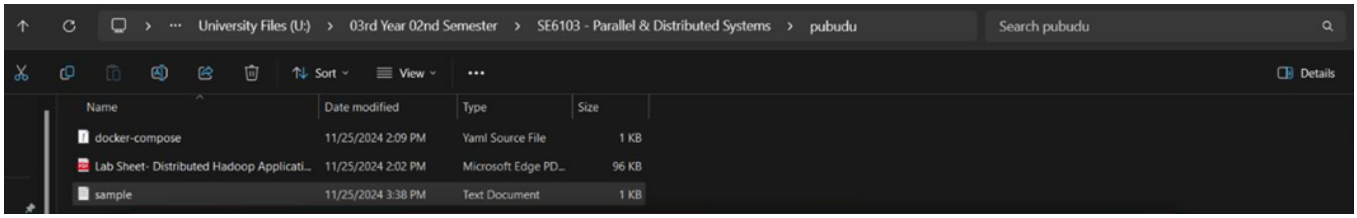


Step 03: Verify Cluster Status



Task 02: Uploading Data to HDFS

Step 01: Download Sample Data



Step 02: Upload Data to HDFS

```
U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu>docker cp ./sample.txt namenode:/sample.txt
Successfully copied 2.05kB to namenode:/sample.txt

U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu>docker exec -it namenode hdfs dfs -put sample.
txt /input
2024-11-25 09:18:16,817 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

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/

U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu>docker exec -it namenode hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples
-*.jar wordcount/input/output
JAR does not exist or is not a normal file: /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar

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/
                                                        docker exec -it namenode hadoop jar /opt/hadoo
educe-examples-*.jar wordcount/input/outputl & Distributed Systems\pubudu>docker exec -it namenode hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapr
JAR does not exist or is not a normal file: /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar

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/

U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu>docker exec -it namenode hadoop jar /opt/hadoop-3.2.1/share/hadoop/mapreduce/hadoop-mapreduce-ex
amples-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.
```

Task 03: Running a Map Reduce Job

Step 01: Run the Word Count MapReduce Job

```
U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu>docker exec -it namenode hadoop jar /opt/hadoop-3.2.1/share/hadoop/mapreduce/hadoop-mapreduce-ex
amples-3.2.1.jar wordcount /input /output
2024-11-25 10:02:51,733 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties

2024-11-25 10:02:52,163 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-25 10:02:52,163 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-25 10:02:53,924 INFO input.FileInputFormat: Total input files to process : 1
2024-11-25 10:02:54,013 INFO mapreduce.JobSubmitter: number of splits:1
2024-11-25 10:02:54,679 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2112989383_0001
2024-11-25 10:02:54,680 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-25 10:02:54,991 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-11-25 10:02:54,994 INFO mapreduce.Job: Running job: job_local2112989383_0001
2024-11-25 10:02:55,000 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-11-25 10:02:55,024 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-25 10:02:55,025 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-11-25 10:02:55,030 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-11-25 10:02:55,245 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-11-25 10:02:55,247 INFO mapred.LocalJobRunner: Starting task: attempt_local2112989383_0001_m_000000_0
2024-11-25 10:02:55,337 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-25 10:02:55,338 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-11-25 10:02:55,482 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-11-25 10:02:55,496 INFO mapred.MapTask: Processing split: hdfs://namenode:8020/input/sample.txt:0+62
2024-11-25 10:02:55,678 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-11-25 10:02:55,678 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-11-25 10:02:55,678 INFO mapred.MapTask: soft limit at 83886080
2024-11-25 10:02:55,678 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-11-25 10:02:55,678 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-11-25 10:02:55,693 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-11-25 10:02:55,989 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-25 10:02:56,014 INFO mapreduce.Job: Job job_local2112989383_0001 running in uber mode : false
2024-11-25 10:02:56,020 INFO mapreduce.Job:  map 0% reduce 0%
2024-11-25 10:02:57,319 INFO mapred.LocalJobRunner:
2024-11-25 10:02:57,332 INFO mapred.MapTask: Starting flush of map output
2024-11-25 10:02:57,333 INFO mapred.MapTask: Spilling map output
2024-11-25 10:02:57,333 INFO mapred.MapTask: bufstart = 0; bufend = 79; bufvoid = 104857600
2024-11-25 10:02:57,333 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536); length = 13/6553600
2024-11-25 10:02:57,388 INFO mapred.MapTask: Finished spill 0
2024-11-25 10:02:57,422 INFO mapred.Task: Task:attempt_local2112989383_0001_m_000000_0 is done. And is in the process of committing
2024-11-25 10:02:57,436 INFO mapred.LocalJobRunner: map
2024-11-25 10:02:57,437 INFO mapred.Task: Task 'attempt_local2112989383_0001_m_000000_0' done.
2024-11-25 10:02:57,463 INFO mapred.Task: Final Counters for attempt_local2112989383_0001_m_000000_0: Counters: 24
```

```
                FILE: Number of bytes read=633604
                FILE: Number of bytes written=1684177
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=124
                HDFS: Number of bytes written=71
                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=1
                Map output records=4
                Map output bytes=79
                Map output materialized bytes=93
                Input split bytes=102
                Combine input records=4
                Combine output records=4
                Reduce input groups=4
                Reduce shuffle bytes=93
                Reduce input records=4
                Reduce output records=4
                Spilled Records=8
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=79
                Total committed heap usage (bytes)=538968064
        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=62
        File Output Format Counters
                Bytes Written=71

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/
```

Step 03: View Job Output

```
U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu> docker exec -it namenode hdfs dfs -ls /output
Found 2 items
-rw-r--r--   3 root supergroup          0 2024-11-25 10:02 /output/_SUCCESS
-rw-r--r--   3 root supergroup         71 2024-11-25 10:02 /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/
```

```
U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu> docker exec -it namenode hdfs dfs -cat /output/part-r-00000
2024-11-25 10:06:22,227 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
bbbbburrcu      1
eeeeggggjjjjorhhh       1
hhhhhgdojvvvvvvv        1
sddddgggggggggty        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/
```

Task 04: Analyze and Clean Up

Step 01: To clean up the cluster

```
U:\03rd Year 02nd Semester\SE6103 - Parallel & Distributed Systems\pubudu> docker-compose down
time="2024-11-25T15:38:32+05:30" level=warning msg="U:\\03rd Year 02nd Semester\\SE6103 - Parallel & Distributed Systems\\pubudu\\docker-compose.yaml: the attribute `vers
ion` is obsolete, it will be ignored, please remove it to avoid potential confusion"
[+] Running 4/4
 ✔Container historyserver  Removed                                                                                10.8s
 ✔Container datanode       Removed                                                                                10.5s
 ✔Container namenode       Removed                                                                                10.6s
 ✔Network pubudu_default   Removed                                                                                 0.3s
```