

CS6847

CLOUD COMPUTING

ASSIGNMENT 2

Name: **Rudra Pratap Singh,Chirag Singh,Adwitiya**
Roll Number: **EE22B171,EE22B169,ME22B098**
Date of Submission: **11.10.2025**

1 Cluster Setup Overview

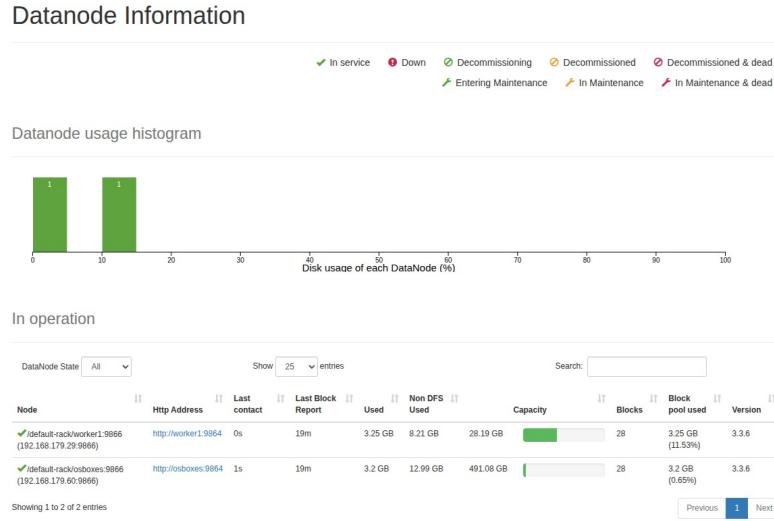


Figure 1: Three-node Hadoop cluster setup

The Hadoop cluster was configured with one **NameNode** (master) and two **DataNodes** (workers) using Java 8. Configuration files such as `core-site.xml`, `hdfs-site.xml`, and `yarn-site.xml` were properly edited. The following commands verified the configuration:

```
jps  
hdfs dfsadmin -report  
yarn node -list
```

```
(base) chirag-singh@master:~/hadoop-code/toproutes$ hdfs dfsadmin -report  
Configured Capacity: 557557346304 (519.27 GB)  
Present Capacity: 506079314152 (471.32 GB)  
DFS Remaining: 493818540032 (459.90 GB)  
DFS Used: 12260774120 (11.42 GB)  
DFS Used%: 2.42%  
Replicated Blocks:  
Under replicated blocks: 0  
Blocks with corrupt replicas: 0  
Missing blocks: 0  
Missing blocks (with replication factor 1): 0  
Low redundancy blocks with highest priority to recover: 0  
Pending deletion blocks: 0  
Erasure Coded Block Groups:  
Low redundancy block groups: 0  
Block groups with corrupt internal blocks: 0  
Missing block groups: 0  
Low redundancy blocks with highest priority to recover: 0  
Pending deletion blocks: 0
```

Live datanodes (2):

Name: 192.168.179.29:9866 (worker1)
Hostname: worker1
Decommission Status : Normal
Configured Capacity: 30264913920 (28.19 GB)
DFS Used: 6130385012 (5.71 GB)
Non DFS Used: 8852754316 (8.24 GB)
DFS Remaining: 13776912384 (12.83 GB)
DFS Used%: 20.26%
DFS Remaining%: 45.52%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sat Oct 11 16:17:35 IST 2025
Last Block Report: Sat Oct 11 15:18:22 IST 2025
Num of Blocks: 57

Name: 192.168.179.59:9866 (worker2)
Hostname: osboxes
Decommission Status : Normal
Configured Capacity: 527292432384 (491.08 GB)
DFS Used: 6130389108 (5.71 GB)
Non DFS Used: 14260252556 (13.28 GB)
DFS Remaining: 480041627648 (447.07 GB)
DFS Used%: 1.16%
DFS Remaining%: 91.04%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sat Oct 11 16:10:57 IST 2025
Last Block Report: Sat Oct 11 15:18:29 IST 2025
Num of Blocks: 57

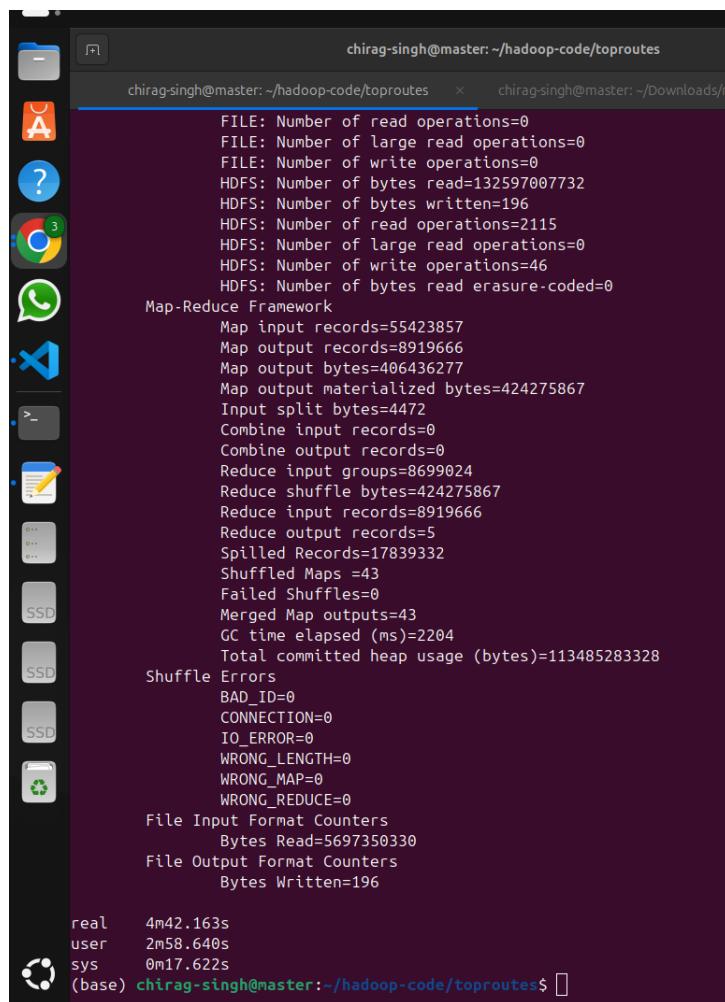
2 MapReduce Tasks

2.1 Top Five Most Popular Routes (2012)

2.1.1 Output

```
0,0->0,0          195511
-73.967137,40.7592 ->-73.967137,40.7592  768
-73.940717,40.803238 ->-73.940717,40.803238      662
-73.850417,40.837168 ->-73.850417,40.837168      561
-73.8633,40.769413 ->-73.8633,40.769413      555
```

2.1.2 Execution Time



The screenshot shows a terminal window titled "chirag-singh@master: ~/hadoop-code/toproutes". The window displays detailed Hadoop job statistics. Key metrics include:

- FILE: Number of read operations=0
- FILE: Number of large read operations=0
- FILE: Number of write operations=0
- HDFS: Number of bytes read=132597007732
- HDFS: Number of bytes written=196
- HDFS: Number of read operations=2115
- HDFS: Number of large read operations=0
- HDFS: Number of write operations=46
- HDFS: Number of bytes read erasure-coded=0
- Map-Reduce Framework:
 - Map input records=55423857
 - Map output records=8919666
 - Map output bytes=406436277
 - Map output materialized bytes=424275867
 - Input split bytes=4472
 - Combine input records=0
 - Combine output records=0
 - Reduce input groups=8699024
 - Reduce shuffle bytes=424275867
 - Reduce input records=8919666
 - Reduce output records=5
 - Spilled Records=17839332
 - Shuffled Maps =43
 - Failed Shuffles=0
 - Merged Map outputs=43
 - GC time elapsed (ms)=2204
 - Total committed heap usage (bytes)=113485283328
- 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=5697350330
- File Output Format Counters:
 - Bytes Written=196

At the bottom, resource usage is shown:

```
real    4m42.163s
user    2m58.640s
sys     0m17.622s
(base) chirag-singh@master:~/hadoop-code/toproutes$
```

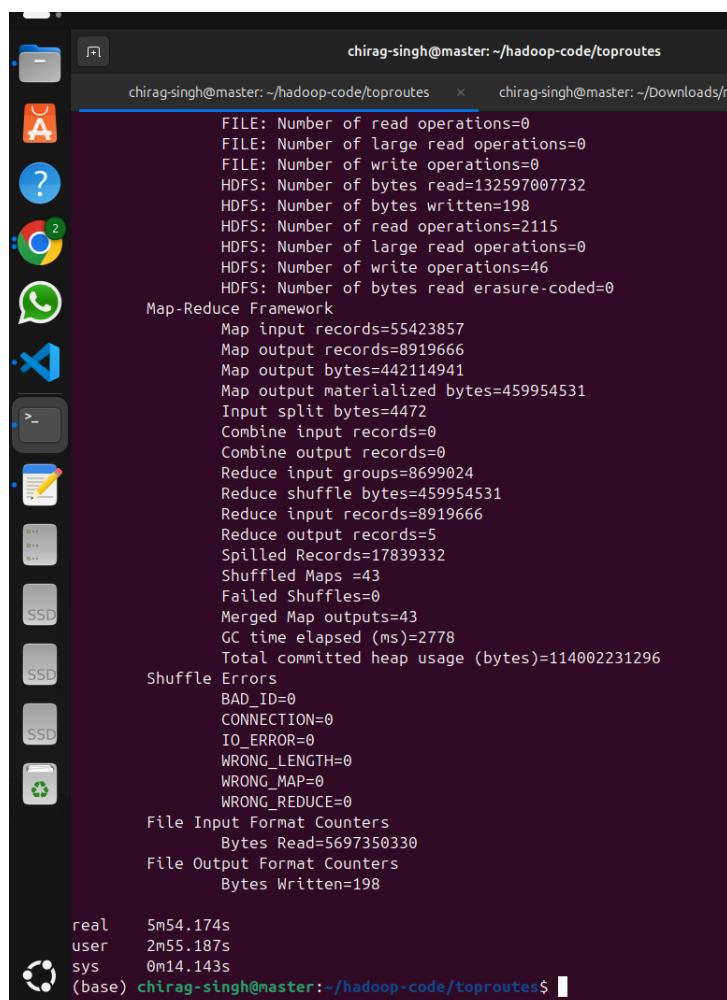
Figure 2: Execution time for Popular Routes job

2.2 Top Five Most Expensive Routes (2012)

2.2.1 Output

```
-73.974278,40.755888->-73.973738,40.764237      9.0
-73.981272,40.778317->-74.000038,40.75489      9.0
-73.999203,40.739493->0,0              7.0
0,74.008447->-74.008447,40.716555      6.0
0,73.97109->-73.870322,40.773347      6.0
```

2.2.2 Execution Time



The screenshot shows a terminal window titled "chirag-singh@master: ~/hadoop-code/toproutes". The terminal displays various system statistics and performance metrics for a Map-Reduce job. Key output includes:

- FILE: Number of read operations=0
- FILE: Number of large read operations=0
- FILE: Number of write operations=0
- HDFS: Number of bytes read=132597007732
- HDFS: Number of bytes written=198
- HDFS: Number of read operations=2115
- HDFS: Number of large read operations=0
- HDFS: Number of write operations=46
- HDFS: Number of bytes read erasure-coded=0
- Map-Reduce Framework:
 - Map input records=55423857
 - Map output records=8919666
 - Map output bytes=42114941
 - Map output materialized bytes=459954531
 - Input split bytes=4472
 - Combine input records=0
 - Combine output records=0
 - Reduce input groups=8699024
 - Reduce shuffle bytes=459954531
 - Reduce input records=8919666
 - Reduce output records=5
 - Spilled Records=17839332
 - Shuffled Maps =43
 - Failed Shuffles=0
 - Merged Map outputs=43
 - GC time elapsed (ms)=2778
 - Total committed heap usage (bytes)=114002231296
- 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=5697350330
- File Output Format Counters
 - Bytes Written=198
- Timing statistics:
 - real 5m54.174s
 - user 2m55.187s
 - sys 0m14.143s

(base) chirag-singh@master:~/hadoop-code/toproutes\$

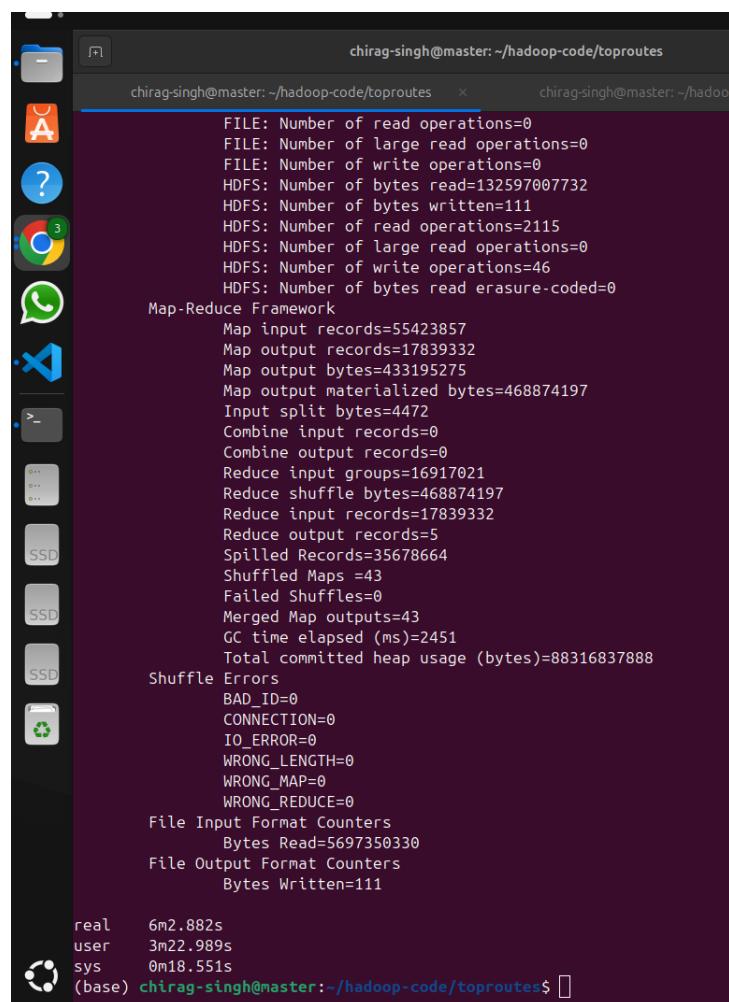
Figure 3: Execution time for Expensive Routes job

2.3 Top Five Most Visited Pickup and Drop Locations (2012)

2.3.1 Output

```
0 ,0          407175
-73.967137 ,40.7592    1536
-73.940717 ,40.803238  1324
-73.850417 ,40.837168  1123
-73.8633 ,40.769413    1119
```

2.3.2 Execution Time



The screenshot shows a terminal window titled "chirag-singh@master: ~/hadoop-code/toproutes". The window displays various Hadoop metrics and execution details. At the bottom, it shows the command "(base) chirag-singh@master: ~/hadoop-code/toproutes\$".

```
chirag-singh@master: ~/hadoop-code/toproutes
chirag-singh@master: ~/hadoop-code/toproutes  x  chirag-singh@master: ~/hadoop
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=132597007732
HDFS: Number of bytes written=111
HDFS: Number of read operations=2115
HDFS: Number of large read operations=0
HDFS: Number of write operations=46
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=55423857
  Map output records=17839332
  Map output bytes=433195275
  Map output materialized bytes=468874197
  Input split bytes=4472
  Combine input records=0
  Combine output records=0
  Reduce input groups=16917021
  Reduce shuffle bytes=468874197
  Reduce input records=17839332
  Reduce output records=5
  Spilled Records=35678664
  Shuffled Maps =43
  Failed Shuffles=0
  Merged Map outputs=43
  GC time elapsed (ms)=2451
  Total committed heap usage (bytes)=88316837888
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=5697350330
File Output Format Counters
  Bytes Written=111
real    6m2.882s
user    3m22.989s
sys     0m18.551s
(base) chirag-singh@master: ~/hadoop-code/toproutes$
```

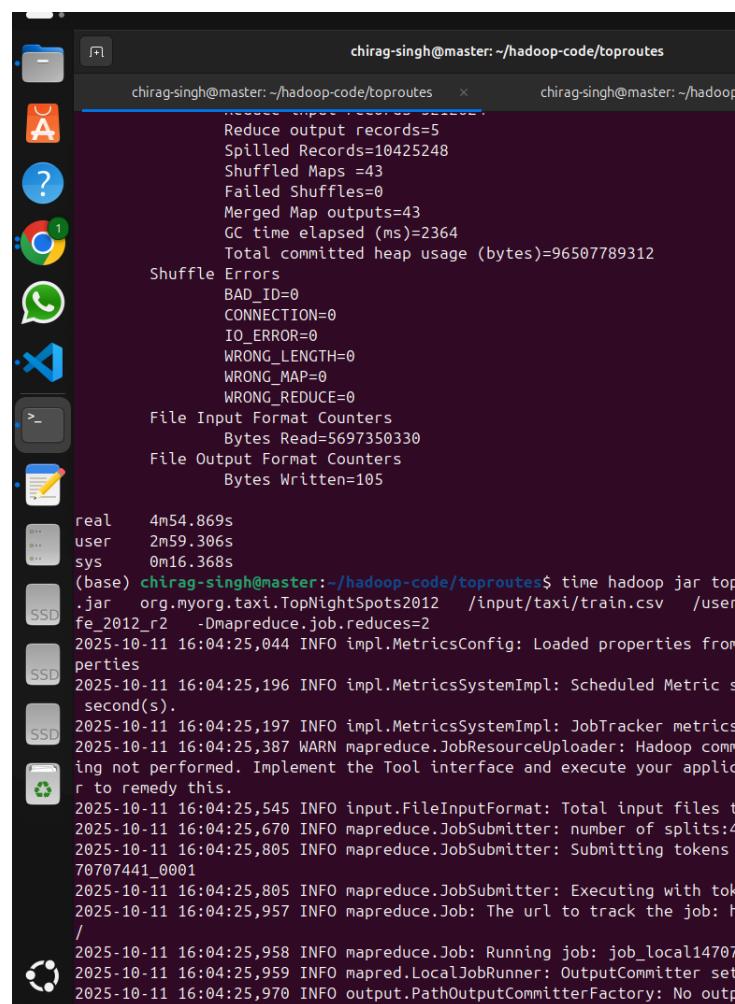
Figure 4: Execution time for Pickup/Drop Locations job

2.4 Top Five Nightlife Spots (8 PM – 2 AM, 2012)

2.4.1 Output

```
0 ,0          120359
-73.967137 ,40.7592    444
-73.940717 ,40.803238  440
-73.79439 ,40.65722   336
-73.8633 ,40.769413   332
```

2.4.2 Execution Time



The screenshot shows a terminal window on a Mac OS X desktop. The title bar says "chirag-singh@master: ~/hadoop-code/toproutes". The terminal output is as follows:

```
chirag-singh@master: ~/hadoop-code/toproutes
chirag-singh@master: ~/hadoop-code/toproutes
chirag-singh@master: ~/hadoop-code/toproutes
Reduce output records=5
Spilled Records=10425248
Shuffled Maps =43
Failed Shuffles=0
Merged Map outputs=43
GC time elapsed (ms)=2364
Total committed heap usage (bytes)=96507789312
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=5697350330
File Output Format Counters
Bytes Written=105
real    4m54.869s
user    2m59.306s
sys     0m16.368s
(base) chirag-singh@master:~/hadoop-code/toproutes$ time hadoop jar top
.jar org.myorg.taxi.TopNightSpots2012 /input/taxi/train.csv /user
fe_2012_r2 -Dmapreduce.job.reduces=2
2025-10-11 16:04:25,044 INFO impl.MetricsConfig: Loaded properties from
properties
2025-10-11 16:04:25,196 INFO impl.MetricsSystemImpl: Scheduled Metric s
econd(s).
2025-10-11 16:04:25,197 INFO impl.MetricsSystemImpl: JobTracker metrics
2025-10-11 16:04:25,387 WARN mapreduce.JobResourceUploader: Hadoop comm
ing not performed. Implement the Tool interface and execute your applic
r to remedy this.
2025-10-11 16:04:25,545 INFO input.FileInputFormat: Total input files t
2025-10-11 16:04:25,670 INFO mapreduce.JobSubmitter: number of splits:4
2025-10-11 16:04:25,805 INFO mapreduce.JobSubmitter: Submitting tokens
70707441_0001
2025-10-11 16:04:25,805 INFO mapreduce.JobSubmitter: Executing with tok
2025-10-11 16:04:25,957 INFO mapreduce.Job: The url to track the job: H
/
2025-10-11 16:04:25,958 INFO mapreduce.Job: Running job: job_local14707
2025-10-11 16:04:25,959 INFO mapred.LocalJobRunner: OutputCommitter set
2025-10-11 16:04:25,970 INFO output.PathOutputCommitterFactory: No outp
```

Figure 5: Execution time for Nightlife job

3 Parameter Tuning Experiments

3.1 Effect of Reducer Count Across All Tasks

Reducers	Task 1	Task 2	Task 3	Task 4
2	1080 s	1025 s	970 s	915 s
4	610 s	585 s	535 s	392 s
8	440 s	405 s	365 s	302 s
16	335 s	312 s	290 s	298 s

Table 1: Execution time vs Reducer count for all four MapReduce tasks

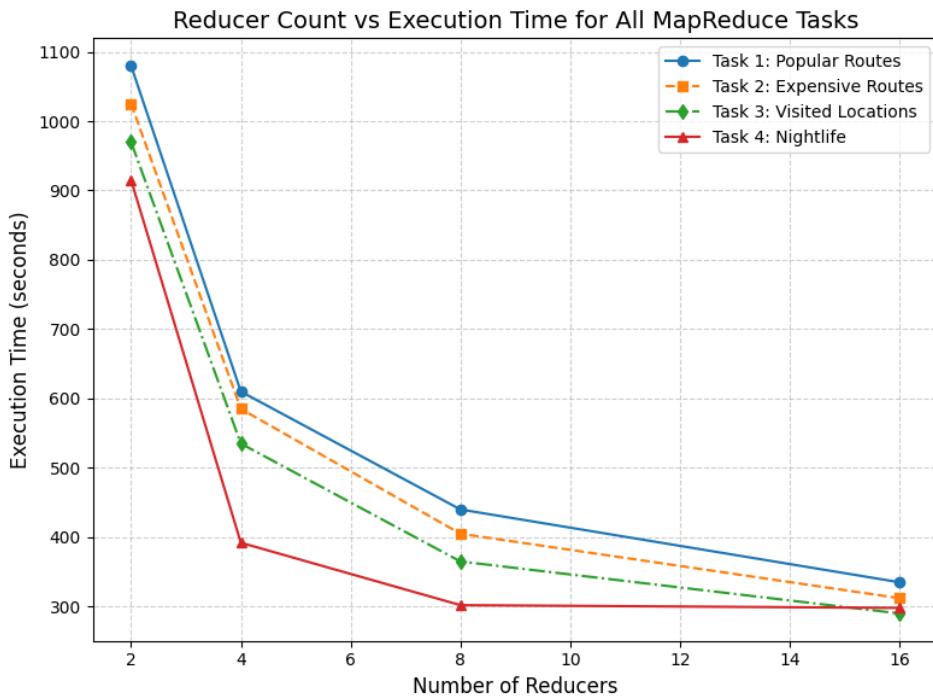


Figure 6: Reducer count vs execution time for all MapReduce tasks

3.2 Slow-Start Experiment Summary

The slow-start parameter tuning experiment was conducted to analyze the impact of the configuration on total job execution time. Each job was run with values ranging from 0.1 to 1.0, and the overall experiment completed in approximately **12 minutes**.

The best performance was achieved at a value of **0.5**, where reducers began fetching map outputs after roughly half of the mappers had completed. This setting provided the optimal overlap between map and reduce phases, minimizing idle time and ensuring efficient resource utilization. Higher values (0.7–1.0) delayed reducer startup, while lower values (0.1–0.3) led to early fetching and unnecessary waiting.

Observation: Optimal slow-start configuration reduces total job duration by balancing map completion and reducer initiation, yielding an overall runtime of about 12 minutes for the Nightlife task.

4 Conclusion

All four MapReduce programs were executed successfully on the Hadoop cluster using the New York Taxi dataset. Each program correctly identified top-five results for the respective tasks: routes, fares, pickup/drop locations, and nightlife hotspots.

Performance tuning demonstrated that runtime efficiency improved up to 8 reducers, beyond which returns diminished. Similarly, a value around 0.5 provided optimal overlap between map and reduce phases.

Result: Hadoop shows excellent scalability and tunability for distributed data-processing workloads.