Part 1:

Creating an input directory

hdfs dfs -mkdir /home/13student13/input_Project

Copying the Project Data to input directory

hdfs dfs -copyFromLocal Project_Data /home/13student13/input_Project

Running the MRjob on cluster

python MedianWind.py -r hadoop hdfs:///home/13student13/input_Project/Project_Data --output-dir=hdfs:///home/13student13/output_ProjectPart13

```
   map 86% reduce 26%
   map 88% reduce 29%
   map 90% reduce 29%
   map 94% reduce 29%
   map 98% reduce 31%
   map 100% reduce 31%
   map 100% reduce 100%
  Job job_1715113870348_2242 completed successfully
  Output directory: hdfs:///home/13student13/output_ProjectPart13
Counters: 49
        File Input Format Counters
                Bytes Read=415313
        File Output Format Counters
                Bytes Written=1430
        File System Counters
                FILE: Number of bytes read=423888
                FILE: Number of bytes written=11364459
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=422563
                HDFS: Number of bytes written=1430
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=153
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=50
                Launched map tasks=50
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=362373120
                Total megabyte-milliseconds taken by all reduce tasks=50324480
                Total time spent by all map tasks (ms)=353880
                Total time spent by all maps in occupied slots (ms)=353880
                Total time spent by all reduce tasks (ms)=49145
                Total time spent by all reduces in occupied slots (ms)=49145
                Total vcore-milliseconds taken by all map tasks=353880
                Total vcore-milliseconds taken by all reduce tasks=49145
        Map-Reduce Framework
                CPU time spent (ms)=34720
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=11570
                Input split bytes=7250
                Map input records=36404
                Map output bytes=366682
                Map output materialized bytes=424182
                Map output records=28600
                Merged Map outputs=50
                Physical memory (bytes) snapshot=17737654272
                Reduce input groups=110
                Reduce input records=28600
                Reduce output records=110
                Reduce shuffle bytes=424182
                Shuffled Maps =50
                Spilled Records=57200
                Total committed heap usage (bytes)=15689318400
                Virtual memory (bytes) snapshot=154156142592
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///home/13student13/output_ProjectPart13
Removing HDFS temp directory hdfs:///user/student13/tmp/mrjob/MedianWind.student13.20241206.051550.622077...
Removing temp directory /tmp/MedianWind.student13.20241206.051550.622077...
```

Listing the contents of output files

hdfs dfs -ls /home/13student13/output_ProjectPart13

```
[student13@msba-hadoop-name ~]$ hdfs dfs -ls /home/13student13/output_ProjectPart13
Found 2 items
-rw-r--r--   5 student13 supergroup          0 2024-12-05 21:17 /home/13student13/output_ProjectPart13/_SUCCESS
-rw-r--r--   5 student13 supergroup       1430 2024-12-05 21:17 /home/13student13/output_ProjectPart13/part-00000
```

Output:

hdfs dfs -cat /home/13student13/output_ProjectPart13/part-00000

```
[student13@msba-hadoop-name ~]$ hdfs dfs -cat /home/13student13/output_ProjectPart13/part-00000
"192101"        230
"192102"        230
"192103"        230
"192104"        230
"192105"        230
"192106"        230
"192107"        250
"192108"        180
"192109"        270
"192110"        250
"192111"        230
"192112"        230
"192201"        140
"192202"        180
"192203"        270
"192204"        180
"192205"        230
"192206"        200
"192207"        200
"192208"        200
"192209"        230
"192210"        270
"192211"        230
"192212"        230
"192301"        200
"192302"        180
"192303"        230
"192304"        230
"192305"        200
"192306"        230
"192307"        230
"192308"        200
"192309"        230
"192310"        200
"192311"        180
"192312"        180
"192401"        180
"192402"        200
"192403"        200
"192404"        200
"192405"        230
"192406"        230
"192407"        230
"192408"        200
"192409"        230
"192410"        200
"192411"        270
"192412"        230
"192501"        250
"192502"        200
"192503"        200
"192504"        200
"192505"        200
```

Part 2-

Running the python file on Spark local

spark-submit --master local Range.py

```
24/12/04 13:15:07 INFO executor.Executor: Running task 46.0 in stage 2.0 (TID 147)
24/12/04 13:15:07 INFO scheduler.TaskSetManager: Finished task 42.0 in stage 2.0 (TID 146) in 110 ms on localhost (executor driver) (47/50)
24/12/04 13:15:07 INFO io.HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoop.mapred.FileOutputCommitter
24/12/04 13:15:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/12/04 13:15:07 INFO storage.ShuffleBlockFetcherIterator: Getting 6 non-empty blocks including 6 local blocks and 0 remote blocks
24/12/04 13:15:07 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/12/04 13:15:07 INFO python.PythonRunner: Times: total = 40, boot = -20, init = 60, finish = 0
24/12/04 13:15:07 INFO storage.ShuffleBlockFetcherIterator: Getting 6 non-empty blocks including 6 local blocks and 0 remote blocks
24/12/04 13:15:07 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/12/04 13:15:07 INFO python.PythonRunner: Times: total = 40, boot = 1, init = 39, finish = 0
24/12/04 13:15:07 INFO python.PythonRunner: Times: total = 87, boot = -20, init = 106, finish = 1
24/12/04 13:15:07 INFO output.FileOutputCommitter: Saved output of task 'attempt_20241204131457_0015_m_000046_0' to hdfs://msba-hadoop-name:9000/ho
me/13student13/outputProjectPart24/_temporary/0/task_20241204131457_0015_m_000046
24/12/04 13:15:07 INFO mapred.SparkHadoopMapRedUtil: attempt_20241204131457_0015_m_000046_0: Committed
24/12/04 13:15:07 INFO executor.Executor: Finished task 46.0 in stage 2.0 (TID 147). 2024 bytes result sent to driver
24/12/04 13:15:07 INFO scheduler.TaskSetManager: Starting task 47.0 in stage 2.0 (TID 148, localhost, executor driver, partition 47, ANY, 7852 byte
s)
24/12/04 13:15:07 INFO executor.Executor: Running task 47.0 in stage 2.0 (TID 148)
24/12/04 13:15:07 INFO scheduler.TaskSetManager: Finished task 46.0 in stage 2.0 (TID 147) in 112 ms on localhost (executor driver) (48/50)
24/12/04 13:15:07 INFO io.HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoop.mapred.FileOutputCommitter
24/12/04 13:15:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/12/04 13:15:07 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks including 1 local blocks and 0 remote blocks
24/12/04 13:15:07 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/12/04 13:15:08 INFO python.PythonRunner: Times: total = 40, boot = -26, init = 66, finish = 0
24/12/04 13:15:08 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks including 1 local blocks and 0 remote blocks
24/12/04 13:15:08 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/12/04 13:15:08 INFO python.PythonRunner: Times: total = 41, boot = 1, init = 39, finish = 1
24/12/04 13:15:08 INFO python.PythonRunner: Times: total = 87, boot = -25, init = 112, finish = 0
24/12/04 13:15:08 INFO output.FileOutputCommitter: Saved output of task 'attempt_20241204131457_0015_m_000047_0' to hdfs://msba-hadoop-name:9000/ho
me/13student13/outputProjectPart24/_temporary/0/task_20241204131457_0015_m_000047
24/12/04 13:15:08 INFO mapred.SparkHadoopMapRedUtil: attempt_20241204131457_0015_m_000047_0: Committed
24/12/04 13:15:08 INFO executor.Executor: Finished task 47.0 in stage 2.0 (TID 148). 2024 bytes result sent to driver
24/12/04 13:15:08 INFO scheduler.TaskSetManager: Starting task 49.0 in stage 2.0 (TID 149, localhost, executor driver, partition 49, ANY, 7852 byte
s)
24/12/04 13:15:08 INFO scheduler.TaskSetManager: Finished task 47.0 in stage 2.0 (TID 148) in 112 ms on localhost (executor driver) (49/50)
24/12/04 13:15:08 INFO executor.Executor: Running task 49.0 in stage 2.0 (TID 149)
24/12/04 13:15:08 INFO io.HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoop.mapred.FileOutputCommitter
24/12/04 13:15:08 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/12/04 13:15:08 INFO storage.ShuffleBlockFetcherIterator: Getting 6 non-empty blocks including 6 local blocks and 0 remote blocks
24/12/04 13:15:08 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/12/04 13:15:08 INFO python.PythonRunner: Times: total = 40, boot = -22, init = 62, finish = 0
24/12/04 13:15:08 INFO storage.ShuffleBlockFetcherIterator: Getting 6 non-empty blocks including 6 local blocks and 0 remote blocks
24/12/04 13:15:08 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/12/04 13:15:08 INFO python.PythonRunner: Times: total = 40, boot = 2, init = 38, finish = 0
24/12/04 13:15:08 INFO python.PythonRunner: Times: total = 87, boot = -20, init = 106, finish = 1
24/12/04 13:15:08 INFO output.FileOutputCommitter: Saved output of task 'attempt_20241204131457_0015_m_000049_0' to hdfs://msba-hadoop-name:9000/ho
me/13student13/outputProjectPart24/_temporary/0/task_20241204131457_0015_m_000049
24/12/04 13:15:08 INFO mapred.SparkHadoopMapRedUtil: attempt_20241204131457_0015_m_000049_0: Committed
24/12/04 13:15:08 INFO executor.Executor: Finished task 49.0 in stage 2.0 (TID 149). 2024 bytes result sent to driver
24/12/04 13:15:08 INFO scheduler.TaskSetManager: Finished task 49.0 in stage 2.0 (TID 149) in 109 ms on localhost (executor driver) (50/50)
24/12/04 13:15:08 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
24/12/04 13:15:08 INFO scheduler.DAGScheduler: ResultStage 2 (runJob at SparkHadoopWriter.scala:78) finished in 5.406 s
24/12/04 13:15:08 INFO scheduler.DAGScheduler: Job 0 finished: runJob at SparkHadoopWriter.scala:78, took 10.621600 s
24/12/04 13:15:08 INFO io.SparkHadoopWriter: Job job_20241204131457_0015 committed.
24/12/04 13:15:08 INFO server.AbstractConnector: Stopped Spark@21a2a510{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
24/12/04 13:15:08 INFO ui.SparkUI: Stopped Spark web UI at http://msba-hadoop-name.csueastbay.edu:4040
24/12/04 13:15:08 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/04 13:15:08 INFO memory.MemoryStore: MemoryStore cleared
24/12/04 13:15:08 INFO storage.BlockManager: BlockManager stopped
24/12/04 13:15:08 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/12/04 13:15:08 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/04 13:15:08 INFO spark.SparkContext: Successfully stopped SparkContext
24/12/04 13:15:09 INFO util.ShutdownHookManager: Shutdown hook called
24/12/04 13:15:09 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-409758a9-504b-4e30-b42c-7dfeddf23847
24/12/04 13:15:09 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-01114864-187f-444a-aa1a-5da5cd83c1f9
24/12/04 13:15:09 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-01114864-187f-444a-aa1a-5da5cd83c1f9/pyspark-678c2b40-6e9a-4e0d-a3d9
-8060e09848b6
```

Listing the contents of the output

hdfs dfs -ls /home/13student13/outputProjectPart24

```
[student13@msba-hadoop-name ~]$ hdfs dfs -ls /home/13student13/outputProjectPart24
Found 51 items
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/_SUCCESS
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00000
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00001
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00002
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00003
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00004
-rw-r--r--   5 student13 supergroup         15 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00005
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00006
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00007
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00008
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00009
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00010
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00011
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00012
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00013
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00014
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00015
-rw-r--r--   5 student13 supergroup         34 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00016
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00017
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00018
-rw-r--r--   5 student13 supergroup         15 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00019
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00020
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00021
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00022
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00023
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00024
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00025
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00026
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00027
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00028
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00029
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00030
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00031
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00032
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00033
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00034
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00035
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00036
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00037
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00038
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00039
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00040
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00041
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00042
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00043
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00044
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00045
-rw-r--r--   5 student13 supergroup         15 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00046
-rw-r--r--   5 student13 supergroup         19 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00047
-rw-r--r--   5 student13 supergroup          0 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00048
-rw-r--r--   5 student13 supergroup         15 2024-12-04 13:15 /home/13student13/outputProjectPart24/part-00049
```

Output

hdfs dfs -cat /home/13student13/outputProjectPart24/part-*

```
[student13@msba-hadoop-name ~]$ hdfs dfs -cat /home/13student13/outputProjectPart24/part-*
(u'028360', 0)
(u'014270', 49500)
(u'029700', 0)
(u'033020', 50000)
(u'029350', 0)
(u'012620', 50000)
(u'034970', 50000)
(u'014030', 50000)
(u'011060', 49800)
(u'030910', 49950)
(u'038040', 49950)
(u'023610', 50000)
(u'028970', 0)
(u'032620', 20000)
(u'029110', 0)
```

Part 3:

Running the MRjob on cluster

python MRStationVisibility.py -r hadoop hdfs:///home/13student13/input_Project/Project_Data --output-dir=hdfs:///home/13student13/output_ProjectPart3

```
                HDFS: Number of bytes written=413501
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=153
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=50
                Launched map tasks=50
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=357467136
                Total megabyte-milliseconds taken by all reduce tasks=49252352
                Total time spent by all map tasks (ms)=349089
                Total time spent by all maps in occupied slots (ms)=349089
                Total time spent by all reduce tasks (ms)=48098
                Total time spent by all reduces in occupied slots (ms)=48098
                Total vcore-milliseconds taken by all map tasks=349089
                Total vcore-milliseconds taken by all reduce tasks=48098
        Map-Reduce Framework
                CPU time spent (ms)=35720
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=11426
                Input split bytes=7250
                Map input records=36404
                Map output bytes=413501
                Map output materialized bytes=486411
                Map output records=36305
                Merged Map outputs=50
                Physical memory (bytes) snapshot=17671667712
                Reduce input groups=15
                Reduce input records=36305
                Reduce output records=36305
                Reduce shuffle bytes=486411
                Shuffled Maps =50
                Spilled Records=72610
                Total committed heap usage (bytes)=16377184256
                Virtual memory (bytes) snapshot=154173276160
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///home/13student13/output_ProjectPart3
Removing HDFS temp directory hdfs:///user/student13/tmp/mrjob/MRStationVisibility.student13.20241206.020657.133262...
Removing temp directory /tmp/MRStationVisibility.student13.20241206.020657.133262...
```

Listing the contents of the output file

hdfs dfs -ls /home/13student13/output_ProjectPart3

```
[student13@msba-hadoop-name ~]$ hdfs dfs -ls /home/13student13/output_ProjectPart3
Found 2 items
-rw-r--r--   5 student13 supergroup          0 2024-12-05 18:08 /home/13student13/output_ProjectPart3/_SUCCESS
-rw-r--r--   5 student13 supergroup     413501 2024-12-05 18:08 /home/13student13/output_ProjectPart3/part-00000
```

Output

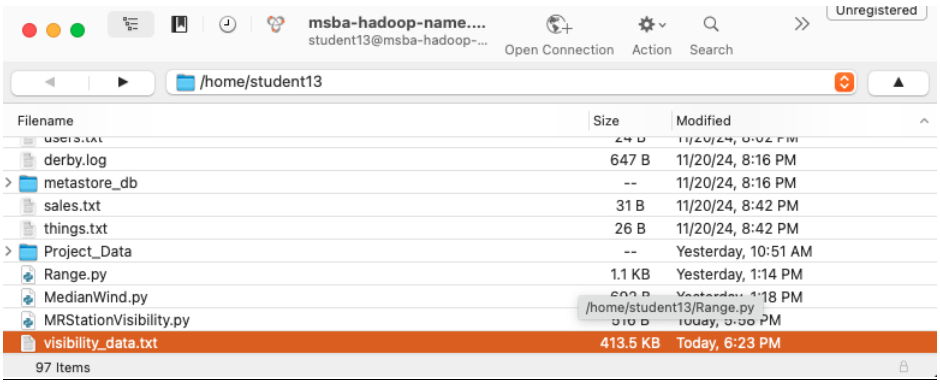
hdfs dfs -cat /home/13student13/output_ProjectPart3/part-00000

```
[student13@msba-hadoop-name ~]$ hdfs dfs -cat /home/13student13/output_ProjectPart3/part-00000
"011060"        10000
"011060"        4000
"011060"        2000
"011060"        4000
"011060"        50000
"011060"        10000
"011060"        20000
"011060"        4000
"011060"        4000
"011060"        4000
"011060"        10000
"011060"        4000
"011060"        10000
"011060"        10000
"011060"        2000
"011060"        50000
"011060"        4000
"011060"        50000
"011060"        50000
"011060"        50000
"011060"        10000
"011060"        4000
"011060"        1000
"011060"        4000
"011060"        2000
"011060"        20000
"011060"        50000
"011060"        50000
"011060"        50000
"011060"        50000
"011060"        50000
"011060"        50000
"011060"        50000
"011060"        4000
"011060"        50000
"011060"        50000
"011060"        1000
"011060"        20000
"011060"        50000
"011060"        50000
"011060"        10000
"011060"        4000
"011060"        4000
"011060"        20000
"011060"        10000
"011060"        20000
"011060"        4000
"011060"        20000
"011060"        10000
"011060"        10000
"011060"        10000
"011060"        50000
"011060"        20000
```

Writing the output to a .txt file

hdfs dfs -getmerge /home/13student13/output_ProjectPart3 visibility_data.txt

cat visibility_data.txt



Part 4 a:

Loading the data to pig

```
records = LOAD 'visibility_data.txt'
 AS (usaf_id:chararray, visibility:int);
```

Checking the loaded data

```
DUMP records;
```

```
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local936328822_0001


2024-12-05 19:53:07,792 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
annot initialize JVM Metrics with processName=JobTracker, sessionId= - already in
itialized
2024-12-05 19:53:07,793 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
annot initialize JVM Metrics with processName=JobTracker, sessionId= - already in
itialized
2024-12-05 19:53:07,794 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
annot initialize JVM Metrics with processName=JobTracker, sessionId= - already in
itialized
2024-12-05 19:53:07,799 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
2024-12-05 19:53:07,802 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-12-05 19:53:07,802 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sch
emaTupleBackend has already been initialized
2024-12-05 19:53:07,813 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileIn
putFormat - Total input files to process : 1
2024-12-05 19:53:07,813 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.util.MapRedUtil - Total input paths to process : 1
("011060",10000)
("011060",4000)
("011060",2000)
("011060",4000)
("011060",50000)
("011060",10000)
("011060",20000)
("011060",4000)
("011060",4000)
("011060",4000)
("011060",10000)
("011060",4000)
("011060",10000)
("011060",10000)
("011060",2000)
("011060",50000)
("011060",4000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",10000)
("011060",4000)
```

DESCRIBE records;

```
grunt> DESCRIBE records;
records: {usaf_id: chararray,visibility: int}
```

filtered_records = FILTER records BY visibility != 999999;

DUMP filtered_records;

```
2024-12-05 19:58:26,161 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - C
annot initialize JVM Metrics with processName=JobTracker, sessionId= - already in
itialized
2024-12-05 19:58:26,164 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
2024-12-05 19:58:26,165 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-12-05 19:58:26,165 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sch
emaTupleBackend has already been initialized
2024-12-05 19:58:26,296 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileIn
putFormat - Total input files to process : 1
2024-12-05 19:58:26,296 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.util.MapRedUtil - Total input paths to process : 1
("011060",10000)
("011060",4000)
("011060",2000)
("011060",4000)
("011060",50000)
("011060",10000)
("011060",20000)
("011060",4000)
("011060",4000)
("011060",4000)
("011060",10000)
("011060",4000)
("011060",10000)
("011060",10000)
("011060",2000)
("011060",50000)
("011060",4000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",10000)
("011060",4000)
("011060",1000)
("011060",4000)
("011060",2000)
("011060",20000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",50000)
("011060",4000)
("011060",50000)
("011060",50000)
```

grouped_records = GROUP filtered_records BY usaf_id;

DUMP grouped_records;

```
e.mapReduceLayer.MapReduceLauncher - Success!
2024-12-05 19:58:54,315 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-12-05 19:58:54,315 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sch
emaTupleBackend has already been initialized
2024-12-05 19:58:54,325 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileIn
putFormat - Total input files to process : 1
2024-12-05 19:58:54,325 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.util.MapRedUtil - Total input paths to process : 1
("011060",{("011060",50000),("011060",20000),("011060",4000),("011060",20000),("0
11060",4000),("011060",4000),("011060",50000),("011060",20000),("011060",10000),(
"011060",10000),("011060",4000),("011060",20000),("011060",4000),("011060",4000),
("011060",50000),("011060",50000),("011060",50000),("011060",10000),("011060",100
00),("011060",50000),("011060",20000),("011060",20000),("011060",20000),("011060"
,50000),("011060",50000),("011060",20000),("011060",20000),("011060",2000),("0110
60",50000),("011060",50000),("011060",50000),("011060",50000),("011060",50000),("
011060",50000),("011060",50000),("011060",50000),("011060",50000),("011060",50000
),("011060",50000),("011060",50000),("011060",50000),("011060",4000),("011060",40
00),("011060",4000),("011060",4000),("011060",10000),("011060",10000),("011060",2
0000),("011060",20000),("011060",50000),("011060",10000),("011060",50000),("01106
0",20000),("011060",50000),("011060",50000),("011060",50000),("011060",50000),("0
11060",50000),("011060",50000),("011060",50000),("011060",50000),("011060",50000)
,("011060",50000),("011060",20000),("011060",50000),("011060",10000),("011060",10
000),("011060",10000),("011060",20000),("011060",4000),("011060",20000),("011060"
,10000),("011060",20000),("011060",4000),("011060",4000),("011060",10000),("01106
0",50000),("011060",50000),("011060",20000),("011060",1000),("011060",50000),("01
1060",50000),("011060",4000),("011060",50000),("011060",50000),("011060",50000),(
"011060",50000),("011060",50000),("011060",50000),("011060",50000),("011060",2000
0),("011060",2000),("011060",4000),("011060",1000),("011060",4000),("011060",1000
0),("011060",50000),("011060",50000),("011060",50000),("011060",4000),("011060",5
0000),("011060",2000),("011060",10000),("011060",10000),("011060",4000),("011060"
,10000),("011060",4000),("011060",4000),("011060",4000),("011060",20000),("011060
",10000),("011060",50000),("011060",4000),("011060",2000),("011060",4000),("01106
0",10000),("011060",50000),("011060",4000),("011060",4000),("011060",10000),("011
060",10000),("011060",10000),("011060",20000),("011060",20000),("011060",10000),(
"011060",20000),("011060",20000),("011060",20000),("011060",50000),("011060",5000
0),("011060",20000),("011060",20000),("011060",10000),("011060",20000),("011060",
50000),("011060",20000),("011060",20000),("011060",10000),("011060",2000),("01106
0",10000),("011060",500),("011060",20000),("011060",10000),("011060",1000),("0110
60",20000),("011060",50000),("011060",200),("011060",1000),("011060",50000),("011
060",50000),("011060",2000),("011060",50000),("011060",20000),("011060",50000),("
011060",50000),("011060",1000),("011060",10000),("011060",4000),("011060",50000),
("011060",50000),("011060",4000),("011060",20000),("011060",50000),("011060",4000
),("011060",20000),("011060",20000),("011060",50000),("011060",20000),("011060",5
0000),("011060",50000),("011060",50000),("011060",20000),("011060",10000),("01106
0",20000),("011060",10000),("011060",4000),("011060",4000),("011060",50000),("011
060",50000),("011060",4000),("011060",4000),("011060",4000),("011060",4000),("011
060",4000),("011060",4000),("011060",4000),("011060",4000),("011060",20000),("011
060",50000),("011060",20000),("011060",10000),("011060",10000),("011060",10000),(
```

range_of_visibility = FOREACH grouped_records GENERATE group AS usaf_id,

>> MAX(filtered_records.visibility) - MIN(filtered_records.visibility) AS visibility_range;

DUMP range_of_visibility;

```
2024-12-05 20:00:04,231 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.util.MapRedUtil - Total input paths to process : 1
("011060",49800)
("012620",50000)
("014030",50000)
("014270",49500)
("023610",50000)
("028360",0)
("028970",0)
("029110",0)
("029350",0)
("029700",0)
("030910",49950)
("032620",20000)
("033020",50000)
("034970",50000)
("038040",49950)
```

ILLUSTRATE range_of_visibility;

```
2024-12-05 20:00:31,993 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMa
pReduce$Reduce - Aliases being processed per job phase (AliasName[line,offset]): M: records[1,10],recor
ds[-1,-1],filtered_records[3,19],grouped_records[4,18] C:  R: range_of_visibility[5,22]
-------------------------------------------------------------
| records     | usaf_id:chararray     | visibility:int     |
-------------------------------------------------------------
|             | "012620"              | 10000              |
|             | "012620"              | 50000              |
|             | "012620"              | 999999             |
-------------------------------------------------------------
-----------------------------------------------------------------------
| filtered_records     | usaf_id:chararray     | visibility:int     |
-----------------------------------------------------------------------
|                      | "012620"              | 10000              |
|                      | "012620"              | 50000              |
-----------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-----------------------
| grouped_records     | group:chararray     | filtered_records:bag{:tuple(usaf_id:chararray,visibility:
int)}                 |
-------------------------------------------------------------------------------------------------------------
-----------------------
|                     | "012620"            | {}
                      |
|                     | "012620"            | {}
                      |
-------------------------------------------------------------------------------------------------------------
-----------------------
-----------------------------------------------------------------------
| range_of_visibility     | usaf_id:chararray     | visibility_range:int     |
-----------------------------------------------------------------------
|                         | "012620"              | 40000                    |
-----------------------------------------------------------------------
```

Part 4 b:

DROP TABLE IF EXISTS project13;

```
hive> DROP TABLE IF EXISTS project13;
OK
```

CREATE TABLE project13 (usaf_id STRING, visibility INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

```
hive> CREATE TABLE project13 (usaf_id STRING, visibility INT)
    > ROW FORMAT DELIMITED
    >   FIELDS TERMINATED BY '\t';
OK
```

LOAD DATA LOCAL INPATH 'visibility_data.txt'

OVERWRITE INTO TABLE project13;

```
hive> LOAD DATA LOCAL INPATH 'visibility_data.txt'
    > OVERWRITE INTO TABLE project13;
Loading data to table default.project13
OK
```

SELECT usaf_id, AVG(visibility)

FROM project13

WHERE visibility != 999999

GROUP BY usaf_id;

```
Total MapReduce CPU Time Spent: 6 seconds 40 msec
OK
"011060"        24848.672566371682
"012620"        26542.331288343557
"014030"        33686.024844720494
"014270"        17137.426900584796
"023610"        37068.553459119496
"028360"        0.0
"028970"        0.0
"029110"        0.0
"029350"        0.0
"029700"        0.0
"030910"        11362.198391420912
"032620"        8316.497461928933
"033020"        12318.483412322275
"034970"        5803.20197044335
"038040"        14158.064516129032
Time taken: 27.488 seconds, Fetched: 15 row(s)
hive>
```