Manideep Elasagaram

Net ID: is6496

The Course Project

Plagiarism and cheating will not be tolerated for this individual and independent project assignment. All submissions will undergo a thorough plagiarism check. Sharing your work with other students, copying others' work, or using AI tools such as ChatGPT is strictly prohibited. Any instances of plagiarism or cheating will result in a grade of 0 for the entire project submission.

The course project consists of four parts. Parts 1 and 2 require the use of the project data zip file, while parts 3 and 4 require the use of a text file named “skyceiling.txt”.

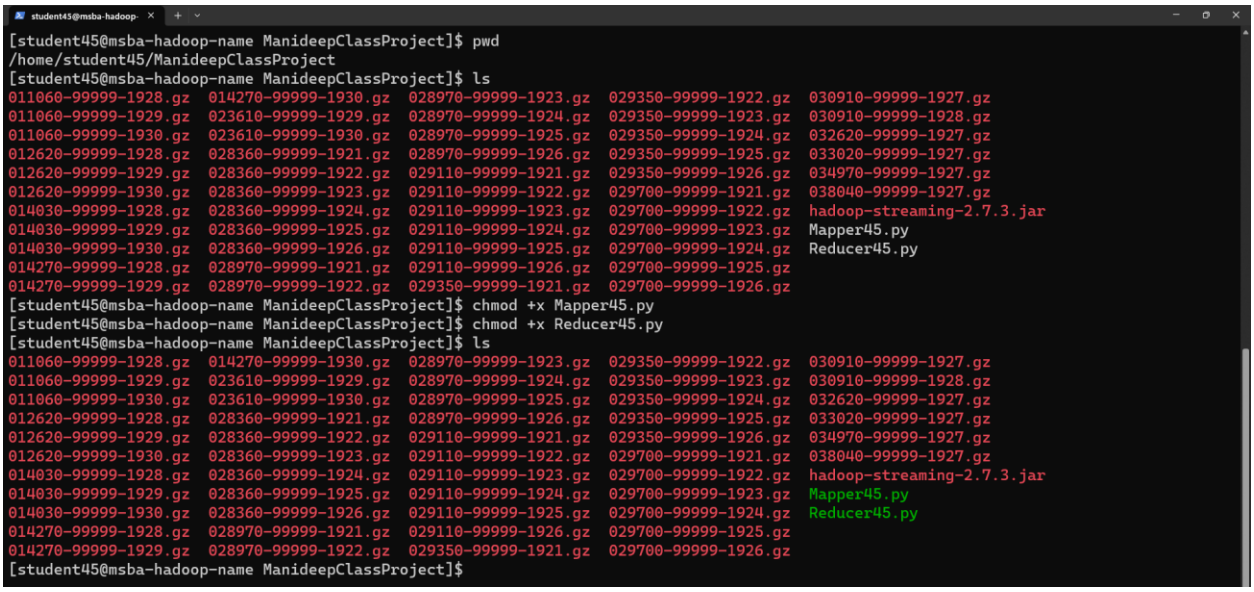
Part 1:

The first part is to develop a Mapper and Reducer application to calculate the *range* (the difference between max and min values) of *sky ceiling height* (meters) for *each observation month* from NCDC records (note: 99999 indicates missing value, and [01459] indicate good quality value).

Step:1

chmod +x Mapper45.py

chmod +x Reducer45.py



Step:2

Unzipping the files:

gunzip *-99999-*.gz

```
[student45@msba-hadoop-name ManideepClassProject]$ gunzip *-99999-*.gz
[student45@msba-hadoop-name ManideepClassProject]$ ls
011060-99999-1928  014270-99999-1928  028360-99999-1925  029110-99999-1922  029350-99999-1925  030910-99999-1928
011060-99999-1929  014270-99999-1929  028360-99999-1926  029110-99999-1923  029350-99999-1926  032620-99999-1927
011060-99999-1930  014270-99999-1930  028970-99999-1921  029110-99999-1924  029700-99999-1921  033020-99999-1927
012620-99999-1928  023610-99999-1929  028970-99999-1922  029110-99999-1925  029700-99999-1922  034970-99999-1927
012620-99999-1929  023610-99999-1930  028970-99999-1923  029110-99999-1926  029700-99999-1923  038040-99999-1927
012620-99999-1930  028360-99999-1921  028970-99999-1924  029350-99999-1921  029700-99999-1924  hadoop-streaming-2.7.3.jar
014030-99999-1928  028360-99999-1922  028970-99999-1925  029350-99999-1922  029700-99999-1925  Mapper45.py
014030-99999-1929  028360-99999-1923  028970-99999-1926  029350-99999-1923  029700-99999-1926  Reducer45.py
014030-99999-1930  028360-99999-1924  029110-99999-1921  029350-99999-1924  030910-99999-1927
[student45@msba-hadoop-name ManideepClassProject]$
```

Step:3

Creating new directory to save input files, copying the input file to the directory:

hdfs dfs -mkdir /home/45student45/input_project

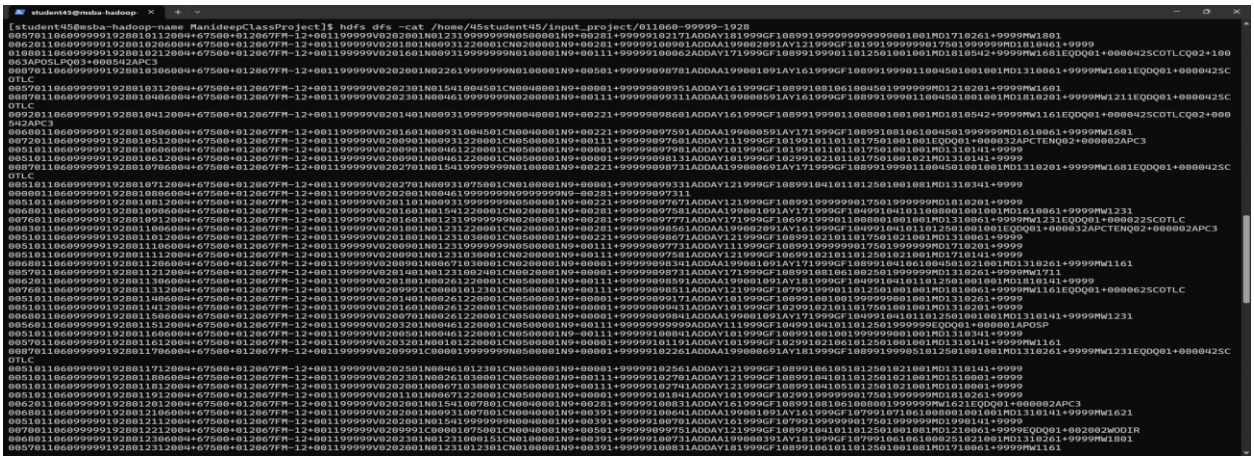
hdfs dfs -copyFromLocal /home/student45/ManideepClassProject/*-99999-* /home/45student45/input_project/

```
[student45@msba-hadoop-name ManideepClassProject]$ hdfs dfs -mkdir /home/45student45/input_project/
[student45@msba-hadoop-name ManideepClassProject]$ hdfs dfs -copyFromLocal /home/student45/ManideepClassProject/*-99999-* /home/45student45/input_project/
[student45@msba-hadoop-name ManideepClassProject]$ hdfs dfs -ls /home/45student45/input_project/
Found 50 items
-rw-r--r--   5 student45 supergroup      18946 2023-05-05 22:34 /home/45student45/input_project/011060-99999-1928
-rw-r--r--   5 student45 supergroup      19934 2023-05-05 22:34 /home/45student45/input_project/011060-99999-1929
-rw-r--r--   5 student45 supergroup      19971 2023-05-05 22:34 /home/45student45/input_project/011060-99999-1930
-rw-r--r--   5 student45 supergroup       8726 2023-05-05 22:34 /home/45student45/input_project/012620-99999-1928
-rw-r--r--   5 student45 supergroup       9657 2023-05-05 22:34 /home/45student45/input_project/012620-99999-1929
-rw-r--r--   5 student45 supergroup      10054 2023-05-05 22:34 /home/45student45/input_project/012620-99999-1930
-rw-r--r--   5 student45 supergroup      13033 2023-05-05 22:34 /home/45student45/input_project/014030-99999-1928
-rw-r--r--   5 student45 supergroup      59900 2023-05-05 22:34 /home/45student45/input_project/014030-99999-1929
-rw-r--r--   5 student45 supergroup     119256 2023-05-05 22:34 /home/45student45/input_project/014030-99999-1930
-rw-r--r--   5 student45 supergroup       9948 2023-05-05 22:34 /home/45student45/input_project/014270-99999-1928
-rw-r--r--   5 student45 supergroup       9442 2023-05-05 22:34 /home/45student45/input_project/014270-99999-1929
-rw-r--r--   5 student45 supergroup       9998 2023-05-05 22:34 /home/45student45/input_project/014270-99999-1930
-rw-r--r--   5 student45 supergroup      37353 2023-05-05 22:34 /home/45student45/input_project/023610-99999-1929
-rw-r--r--   5 student45 supergroup      87958 2023-05-05 22:34 /home/45student45/input_project/023610-99999-1930
-rw-r--r--   5 student45 supergroup     151540 2023-05-05 22:34 /home/45student45/input_project/028360-99999-1921
-rw-r--r--   5 student45 supergroup     152226 2023-05-05 22:34 /home/45student45/input_project/028360-99999-1922
-rw-r--r--   5 student45 supergroup     151749 2023-05-05 22:34 /home/45student45/input_project/028360-99999-1923
-rw-r--r--   5 student45 supergroup     153031 2023-05-05 22:34 /home/45student45/input_project/028360-99999-1924
-rw-r--r--   5 student45 supergroup     150831 2023-05-05 22:34 /home/45student45/input_project/028360-99999-1925
-rw-r--r--   5 student45 supergroup     151838 2023-05-05 22:34 /home/45student45/input_project/028360-99999-1926
-rw-r--r--   5 student45 supergroup     147469 2023-05-05 22:34 /home/45student45/input_project/028970-99999-1921
-rw-r--r--   5 student45 supergroup     148648 2023-05-05 22:34 /home/45student45/input_project/028970-99999-1922
-rw-r--r--   5 student45 supergroup     151594 2023-05-05 22:34 /home/45student45/input_project/028970-99999-1923
-rw-r--r--   5 student45 supergroup     152011 2023-05-05 22:34 /home/45student45/input_project/028970-99999-1924
-rw-r--r--   5 student45 supergroup     151430 2023-05-05 22:34 /home/45student45/input_project/028970-99999-1925
-rw-r--r--   5 student45 supergroup     119414 2023-05-05 22:34 /home/45student45/input_project/028970-99999-1926
-rw-r--r--   5 student45 supergroup     150855 2023-05-05 22:34 /home/45student45/input_project/029110-99999-1921
-rw-r--r--   5 student45 supergroup     150762 2023-05-05 22:34 /home/45student45/input_project/029110-99999-1922
-rw-r--r--   5 student45 supergroup     151938 2023-05-05 22:34 /home/45student45/input_project/029110-99999-1923
-rw-r--r--   5 student45 supergroup     136425 2023-05-05 22:34 /home/45student45/input_project/029110-99999-1924
-rw-r--r--   5 student45 supergroup     148719 2023-05-05 22:34 /home/45student45/input_project/029110-99999-1925
-rw-r--r--   5 student45 supergroup     148629 2023-05-05 22:34 /home/45student45/input_project/029110-99999-1926
-rw-r--r--   5 student45 supergroup     151947 2023-05-05 22:34 /home/45student45/input_project/029350-99999-1921
-rw-r--r--   5 student45 supergroup     149122 2023-05-05 22:34 /home/45student45/input_project/029350-99999-1922
-rw-r--r--   5 student45 supergroup     148971 2023-05-05 22:34 /home/45student45/input_project/029350-99999-1923
-rw-r--r--   5 student45 supergroup     149467 2023-05-05 22:34 /home/45student45/input_project/029350-99999-1924
-rw-r--r--   5 student45 supergroup     152100 2023-05-05 22:34 /home/45student45/input_project/029350-99999-1925
-rw-r--r--   5 student45 supergroup     152487 2023-05-05 22:34 /home/45student45/input_project/029350-99999-1926
-rw-r--r--   5 student45 supergroup     152001 2023-05-05 22:34 /home/45student45/input_project/029700-99999-1921
-rw-r--r--   5 student45 supergroup     151532 2023-05-05 22:34 /home/45student45/input_project/029700-99999-1922
-rw-r--r--   5 student45 supergroup     152747 2023-05-05 22:34 /home/45student45/input_project/029700-99999-1923
-rw-r--r--   5 student45 supergroup     152754 2023-05-05 22:34 /home/45student45/input_project/029700-99999-1924
-rw-r--r--   5 student45 supergroup     152978 2023-05-05 22:34 /home/45student45/input_project/029700-99999-1925
-rw-r--r--   5 student45 supergroup     152522 2023-05-05 22:34 /home/45student45/input_project/029700-99999-1926
-rw-r--r--   5 student45 supergroup      30057 2023-05-05 22:34 /home/45student45/input_project/030910-99999-1927
```

Step:4

Displaying contents of the sample file:

hdfs dfs -cat /home/45student45/input_project/011060-99999-1928



Step:5

Executing python file using the jar file

hadoop jar hadoop-streaming-2.7.3.jar \

-file /home/student45/ManideepClassProject/Mapper45.py \

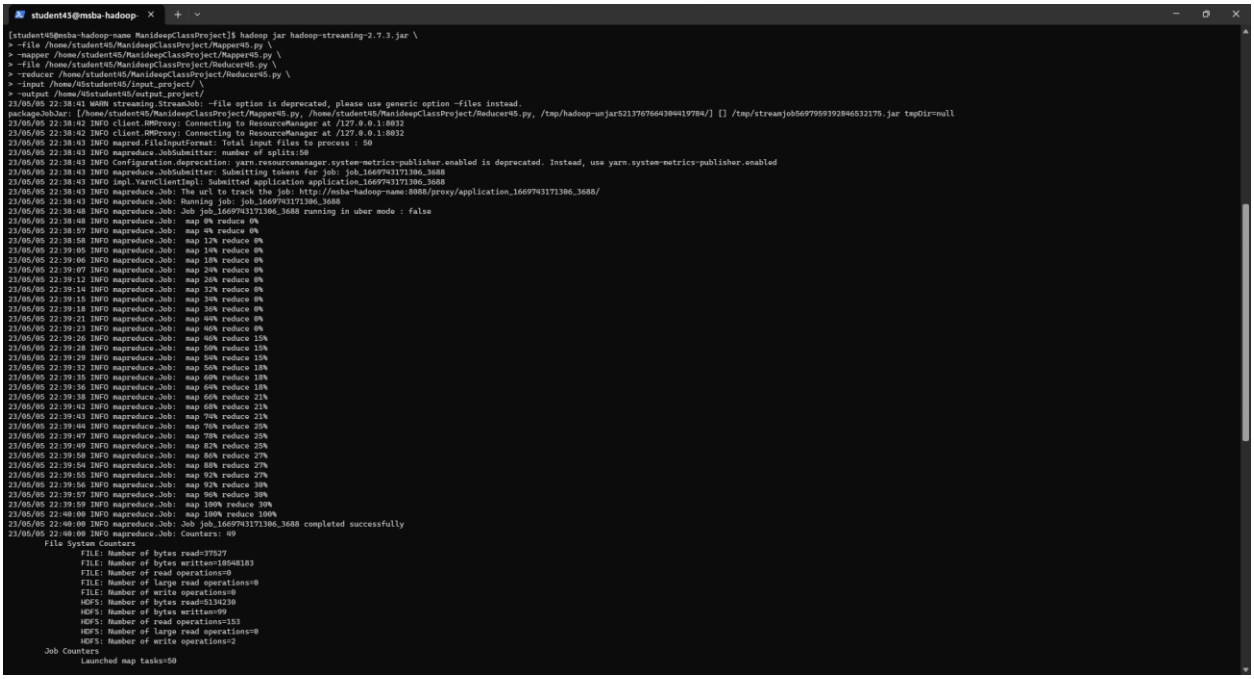
-mapper /home/student45/ManideepClassProject/Mapper45.py \

-file /home/student45/ManideepClassProject/Reducer45.py \

-reducer /home/student45/ManideepClassProject/Reducer45.py \

-input /home/45student45/input_project/ \

-output /home/45student45/output_project/


student45@msba-hadoop-
[student45@msba-hadoop-name ManideepClassProject]$ hadoop jar hadoop-streaming-2.7.3.jar \
> -file /home/student45/ManideepClassProject/Mapper45.py \
> -mapper /home/student45/ManideepClassProject/Mapper45.py \
> -file /home/student45/ManideepClassProject/Reducer45.py \
> -reducer /home/student45/ManideepClassProject/Reducer45.py \
> -input /home/45student45/input_project/ \
> -output /home/45student45/output_project/
23/05/05 22:38:41 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/student45/ManideepClassProject/Mapper45.py, /home/student45/ManideepClassProject/Reducer45.py, /tmp/hadoop-unjar5213767664304419784/] [] /tmp/streamjob5697959392846532175.jar tmpDir=null
23/05/05 22:38:42 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
23/05/05 22:38:42 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
23/05/05 22:38:43 INFO mapred.FileInputFormat: Total input files to process : 50
23/05/05 22:38:43 INFO mapreduce.JobSubmitter: number of splits:50
23/05/05 22:38:43 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/05/05 22:38:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669743171306_3688
23/05/05 22:38:43 INFO impl.YarnClientImpl: Submitted application application_1669743171306_3688
23/05/05 22:38:43 INFO mapreduce.Job: The url to track the job: http://msba-hadoop-name:8088/proxy/application_1669743171306_3688/
23/05/05 22:38:43 INFO mapreduce.Job: Running job: job_1669743171306_3688
23/05/05 22:38:48 INFO mapreduce.Job: Job job_1669743171306_3688 running in uber mode : false
23/05/05 22:38:48 INFO mapreduce.Job: map 0% reduce 0%
23/05/05 22:38:57 INFO mapreduce.Job: map 4% reduce 0%
23/05/05 22:38:58 INFO mapreduce.Job: map 12% reduce 0%
23/05/05 22:39:05 INFO mapreduce.Job: map 14% reduce 0%
23/05/05 22:39:06 INFO mapreduce.Job: map 18% reduce 0%
23/05/05 22:39:07 INFO mapreduce.Job: map 24% reduce 0%
23/05/05 22:39:12 INFO mapreduce.Job: map 26% reduce 0%
23/05/05 22:39:14 INFO mapreduce.Job: map 32% reduce 0%
23/05/05 22:39:15 INFO mapreduce.Job: map 34% reduce 0%
23/05/05 22:39:18 INFO mapreduce.Job: map 36% reduce 0%
23/05/05 22:39:21 INFO mapreduce.Job: map 44% reduce 0%
23/05/05 22:39:23 INFO mapreduce.Job: map 46% reduce 0%
23/05/05 22:39:26 INFO mapreduce.Job: map 46% reduce 15%
23/05/05 22:39:28 INFO mapreduce.Job: map 50% reduce 15%
23/05/05 22:39:29 INFO mapreduce.Job: map 54% reduce 15%
23/05/05 22:39:32 INFO mapreduce.Job: map 56% reduce 18%
23/05/05 22:39:35 INFO mapreduce.Job: map 60% reduce 18%
23/05/05 22:39:36 INFO mapreduce.Job: map 64% reduce 18%
23/05/05 22:39:38 INFO mapreduce.Job: map 66% reduce 21%
23/05/05 22:39:42 INFO mapreduce.Job: map 68% reduce 21%
23/05/05 22:39:43 INFO mapreduce.Job: map 74% reduce 21%
23/05/05 22:39:44 INFO mapreduce.Job: map 76% reduce 25%
23/05/05 22:39:47 INFO mapreduce.Job: map 78% reduce 25%
23/05/05 22:39:49 INFO mapreduce.Job: map 82% reduce 25%
23/05/05 22:39:50 INFO mapreduce.Job: map 86% reduce 27%
23/05/05 22:39:54 INFO mapreduce.Job: map 88% reduce 27%
23/05/05 22:39:55 INFO mapreduce.Job: map 92% reduce 27%
23/05/05 22:39:56 INFO mapreduce.Job: map 92% reduce 30%
23/05/05 22:39:57 INFO mapreduce.Job: map 96% reduce 30%
23/05/05 22:39:59 INFO mapreduce.Job: map 100% reduce 30%
23/05/05 22:40:00 INFO mapreduce.Job: map 100% reduce 100%
23/05/05 22:40:00 INFO mapreduce.Job: Job job_1669743171306_3688 completed successfully
23/05/05 22:40:00 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=37527
FILE: Number of bytes written=10548183
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5134230
HDFS: Number of bytes written=99
HDFS: Number of read operations=153
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=50

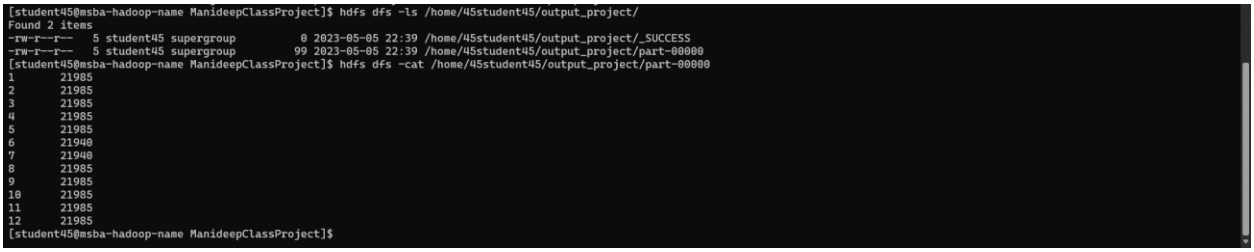

student45@msba-hadoop-
23/05/05 22:39:50 INFO mapreduce.Job: map 86% reduce 27%
23/05/05 22:39:54 INFO mapreduce.Job: map 88% reduce 27%
23/05/05 22:39:55 INFO mapreduce.Job: map 92% reduce 27%
23/05/05 22:39:56 INFO mapreduce.Job: map 92% reduce 30%
23/05/05 22:39:57 INFO mapreduce.Job: map 96% reduce 30%
23/05/05 22:39:59 INFO mapreduce.Job: map 100% reduce 30%
23/05/05 22:40:00 INFO mapreduce.Job: map 100% reduce 100%
23/05/05 22:40:00 INFO mapreduce.Job: Job job_1669743171306_3688 completed successfully
23/05/05 22:40:00 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=37527
FILE: Number of bytes written=10548183
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5134230
HDFS: Number of bytes written=99
HDFS: Number of read operations=153
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=50
Launched reduce tasks=1
Data-local map tasks=50
Total time spent by all maps in occupied slots (ms)=292464
Total time spent by all reduces in occupied slots (ms)=52034
Total time spent by all map tasks (ms)=292464
Total time spent by all reduce tasks (ms)=52034
Total vcore-milliseconds taken by all map tasks=292464
Total vcore-milliseconds taken by all reduce tasks=52034
Total megabyte-milliseconds taken by all map tasks=299483136
Total megabyte-milliseconds taken by all reduce tasks=53282816
Map-Reduce Framework
Map input records=364404
Map output records=3411
Map output bytes=30699
Map output materialized bytes=37821
Input split bytes=6450
Combine input records=0
Combine output records=0
Reduce input groups=12
Reduce shuffle bytes=37821
Reduce input records=3411
Reduce output records=12
Spilled Records=6822
Shuffled Maps =50
Failed Shuffles=0
Merged Map outputs=50
GC time elapsed (ms)=10963
CPU time spent (ms)=35080
Physical memory (bytes) snapshot=17571983360
Virtual memory (bytes) snapshot=154110947328
Total committed heap usage (bytes)=15611199488
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=5127780
File Output Format Counters
Bytes Written=99
23/05/05 22:40:00 INFO streaming.StreamJob: Output directory: /home/45student45/output_project/

Step:6

Displaying the final results:

hdfs dfs -ls /home/45student45/output_project/

hdfs dfs -cat /home/45student45/output_project/part-00000



Part 2:

The second part is to develop a python application that can be implemented in PySpark to calculate the *average visibility distance* (meters) for *each USAF weather station ID* from NCDC records (note: 999999 indicates missing value, and [01459] indicate good quality value).

Step:1

Unzipping sample files and creating a new directory:

gunzip *-99999-*.gz

hdfs dfs -mkdir /user/hadoop/input_2

```
Last login: Fri May  5 23:00:27 on ttys003
(base) manideepelasagaram@Manideeps-MacBook-Pro ~ % cd Desktop
(base) manideepelasagaram@Manideeps-MacBook-Pro Desktop % ssh -i BAN632.pem hadoop@ec2-34-205-90-229.compute-1.amazonaws.com

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
6 package(s) needed for security, out of 18 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M::::::M         M::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M:::::::M       M:::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M::::::::M     M::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-4-169 ~]$ gunzip *-99999-*.gz
[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -mkdir /user/hadoop/input_2
```

Step:2

Copying sample files from local to Hadoop:

hdfs dfs -copyFromLocal *-99999-* /user/hadoop/input_2

```
[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -copyFromLocal *-99999-* /user/hadoop/input_2
```

Step:3

Displaying the copied files:

hdfs dfs -ls /user/hadoop/input_2

```
[[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -ls /user/hadoop/input_2
Found 50 items
-rw-r--r--   1 hadoop hdfsadmingroup      18946 2023-05-06 06:07 /user/hadoop/input_2/011060-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup      19934 2023-05-06 06:07 /user/hadoop/input_2/011060-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup      19971 2023-05-06 06:07 /user/hadoop/input_2/011060-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup       8726 2023-05-06 06:07 /user/hadoop/input_2/012620-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup       9657 2023-05-06 06:07 /user/hadoop/input_2/012620-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup      10054 2023-05-06 06:07 /user/hadoop/input_2/012620-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup      13033 2023-05-06 06:07 /user/hadoop/input_2/014030-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup      59900 2023-05-06 06:07 /user/hadoop/input_2/014030-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup     119256 2023-05-06 06:07 /user/hadoop/input_2/014030-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup       9948 2023-05-06 06:07 /user/hadoop/input_2/014270-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup       9442 2023-05-06 06:07 /user/hadoop/input_2/014270-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup       9998 2023-05-06 06:07 /user/hadoop/input_2/014270-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup      37353 2023-05-06 06:07 /user/hadoop/input_2/023610-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup      87958 2023-05-06 06:07 /user/hadoop/input_2/023610-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup     151540 2023-05-06 06:07 /user/hadoop/input_2/028360-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     152226 2023-05-06 06:07 /user/hadoop/input_2/028360-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     151749 2023-05-06 06:07 /user/hadoop/input_2/028360-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     153031 2023-05-06 06:07 /user/hadoop/input_2/028360-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     150831 2023-05-06 06:07 /user/hadoop/input_2/028360-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     151838 2023-05-06 06:07 /user/hadoop/input_2/028360-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup     147469 2023-05-06 06:07 /user/hadoop/input_2/028970-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     148648 2023-05-06 06:07 /user/hadoop/input_2/028970-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     151594 2023-05-06 06:07 /user/hadoop/input_2/028970-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     152011 2023-05-06 06:07 /user/hadoop/input_2/028970-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     151430 2023-05-06 06:07 /user/hadoop/input_2/028970-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     119414 2023-05-06 06:07 /user/hadoop/input_2/028970-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup     150855 2023-05-06 06:07 /user/hadoop/input_2/029110-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     150762 2023-05-06 06:07 /user/hadoop/input_2/029110-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     151938 2023-05-06 06:07 /user/hadoop/input_2/029110-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     136425 2023-05-06 06:07 /user/hadoop/input_2/029110-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     148719 2023-05-06 06:07 /user/hadoop/input_2/029110-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     148629 2023-05-06 06:07 /user/hadoop/input_2/029110-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup     151947 2023-05-06 06:07 /user/hadoop/input_2/029350-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     149122 2023-05-06 06:07 /user/hadoop/input_2/029350-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     148971 2023-05-06 06:07 /user/hadoop/input_2/029350-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     149467 2023-05-06 06:07 /user/hadoop/input_2/029350-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     152100 2023-05-06 06:07 /user/hadoop/input_2/029350-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     152487 2023-05-06 06:07 /user/hadoop/input_2/029350-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup     152001 2023-05-06 06:07 /user/hadoop/input_2/029700-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     151532 2023-05-06 06:07 /user/hadoop/input_2/029700-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     152747 2023-05-06 06:07 /user/hadoop/input_2/029700-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     152754 2023-05-06 06:07 /user/hadoop/input_2/029700-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     152978 2023-05-06 06:07 /user/hadoop/input_2/029700-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     152522 2023-05-06 06:07 /user/hadoop/input_2/029700-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup      30057 2023-05-06 06:07 /user/hadoop/input_2/030910-99999-1927
-rw-r--r--   1 hadoop hdfsadmingroup      34042 2023-05-06 06:07 /user/hadoop/input_2/030910-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup      33359 2023-05-06 06:07 /user/hadoop/input_2/032620-99999-1927
-rw-r--r--   1 hadoop hdfsadmingroup      36616 2023-05-06 06:07 /user/hadoop/input_2/033020-99999-1927
-rw-r--r--   1 hadoop hdfsadmingroup      35171 2023-05-06 06:07 /user/hadoop/input_2/034970-99999-1927
-rw-r--r--   1 hadoop hdfsadmingroup      36622 2023-05-06 06:07 /user/hadoop/input_2/038040-99999-1927
```

Step:4

Executing the python file:

spark-submit --master yarn AVD_Sparkfile.py

Step:5

Displaying the output files:

hdfs dfs -ls /user/hadoop/

```
[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -ls /user/hadoop/
Found 3 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/.sparkStaging
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-06 06:07 /user/hadoop/input_2
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2
[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -ls /user/hadoop/output_2/
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt
[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -ls /user/hadoop/output_2/output_q2.txt/
Found 51 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00002
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00003
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00004
-rw-r--r--   1 hadoop hdfsadmingroup         47 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00005
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00006
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00007
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00008
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00009
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00010
-rw-r--r--   1 hadoop hdfsadmingroup         29 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00011
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00012
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00013
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00014
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00015
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00016
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00017
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00018
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00019
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00020
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00021
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00022
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00023
-rw-r--r--   1 hadoop hdfsadmingroup         30 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00024
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00025
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00026
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00027
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00028
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00029
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00030
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00031
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00032
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00033
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00034
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00035
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00036
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00037
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00038
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00039
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00040
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00041
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00042
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00043
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00044
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00045
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00046
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00047
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00048
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-06 06:20 /user/hadoop/output_2/output_q2.txt/part-00049
```

Step:6

Displaying the contents of the output files combined:

hdfs dfs -cat /user/hadoop/output_2/output_q2.txt/part-000*

```
[[hadoop@ip-172-31-4-169 ~]$ hdfs dfs -cat /user/hadoop/output_2/output_q2.txt/part-000*
('023610', 37068.553459119496)
('029350', 0.0)
('014270', 17137.426900584796)
('034970', 5803.20197044335)
('012620', 26542.331288343557)
('033020', 12318.483412322275)
('029110', 0.0)
('028970', 0.0)
('032620', 8316.497461928933)
('029700', 0.0)
('014030', 33686.024844720494)
('030910', 11362.198391420912)
('028360', 0.0)
('011060', 24848.672566371682)
('038040', 14158.064516129032)
[hadoop@ip-172-31-4-169 ~]$
```

Part 3:

The third part is to load the text file into Pig and get the range of sky ceiling height for each USAF weather station ID.

Step:1 records_ch = LOAD 'pig/skyceiling_data.txt'

AS (USAF_stationID:chararray, ceiling_height:int);

```
grunt> records_ch = LOAD 'pig/skyceiling_data.txt'
>>   AS (USAF_stationID:chararray, ceiling_height:int);
2023-05-04 19:54:03,912 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt>
```

Step:2 DUMP records_ch;

Step:3 DESCRIBE records_ch;

Step:4 filtered_records_ch = FILTER records_ch BY ceiling_height != 99999;

DUMP filtered_records_ch;

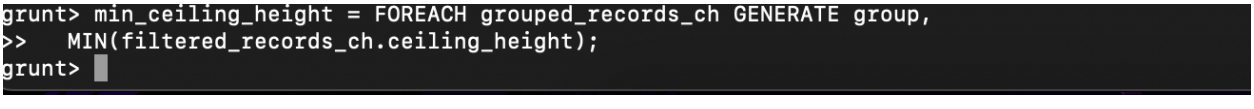
Step:5 grouped_records_ch = GROUP filtered_records_ch BY USAF_stationID;

DUMP grouped_records_ch;

Step:6 DESCRIBE grouped_records_ch;

Step:7 min_ceiling_height = FOREACH grouped_records_ch GENERATE group,

MIN(filtered_records_ch.ceiling_height);



DUMP max_ceiling_height;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.9.0   0.17.0  student45       2023-05-04 20:00:15     2023-05-04 20:00:16     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local11033654270_0004       1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     filtered_records_ch,grouped_records_ch,min_ceiling_height,records_ch     GROUP_BY,COMBINER       file:/tmp/temp-863784154/tmp1102737492,

Input(s):
Successfully read 3391 records from: "file:///home/student45/pig/skyceiling_data.txt"

Output(s):
Successfully stored 10 records in: "file:/tmp/temp-863784154/tmp1102737492"

Counters:
Total records written : 10
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local11033654270_0004


2023-05-04 20:00:16,034 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:00:16,035 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:00:16,036 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:00:16,038 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-05-04 20:00:16,039 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-04 20:00:16,039 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-05-04 20:00:16,054 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-05-04 20:00:16,054 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(011060,15)
(012620,15)
(014030,15)
(014270,15)
(023610,15)
(030910,15)
(032620,240)
(033020,15)
(034970,15)
(038040,60)
grunt>
```

Step:8 max_ceiling_height = FOREACH grouped_records_ch GENERATE group,

MAX(filtered_records_ch.ceiling_height);

```
grunt> max_ceiling_height = FOREACH grouped_records_ch GENERATE group,
>>   MAX(filtered_records_ch.ceiling_height);
```

DUMP max_ceiling_height;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.9.0   0.17.0  student45       2023-05-04 20:01:34     2023-05-04 20:01:34     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1172565997_0005        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     filtered_records_ch,grouped_records_ch,max_ceiling_height,records_ch     GROUP_BY,COMBINER       file:/tmp/temp-863784154/tmp156975719,

Input(s):
Successfully read 3391 records from: "file:///home/student45/pig/skyceiling_data.txt"

Output(s):
Successfully stored 10 records in: "file:/tmp/temp-863784154/tmp156975719"

Counters:
Total records written : 10
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1172565997_0005


2023-05-04 20:01:34,566 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:01:34,567 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:01:34,569 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:01:34,571 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-05-04 20:01:34,572 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-04 20:01:34,572 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-05-04 20:01:34,585 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-05-04 20:01:34,585 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(011060,22000)
(012620,22000)
(014030,22000)
(014270,22000)
(023610,22000)
(030910,22000)
(032620,22000)
(033020,22000)
(034970,22000)
(038040,22000)
grunt>
```

Step:9 f height_range = FOREACH grouped_records_ch GENERATE group AS USAF_stationID, (max_ceiling_height - min_ceiling_height) AS range_ch;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.9.0   0.17.0  student42       2023-05-04 20:03:43     2023-05-04 20:03:44     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1363916344_0002 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Range_height_end,grouped_records,records        GROUP_BY,COMBINER       file:/tmp/temp1415349613/tmp
-109594995,

Input(s):
Successfully read 3391 records from: "file:///home/student42/pig/skyceiling_data.txt"

Output(s):
Successfully stored 10 records in: "file:/tmp/temp1415349613/tmp-109594995"

Counters:
Total records written : 10
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1363916344_0002


2023-05-04 20:03:44,294 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:03:44,295 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:03:44,296 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-04 20:03:44,299 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-05-04 20:03:44,300 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-04 20:03:44,300 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-05-04 20:03:44,311 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-05-04 20:03:44,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(011060,21985)
(012620,21985)
(014030,21985)
(014270,21985)
(023610,21985)
(030910,21985)
(032620,21760)
(033020,21985)
(034970,21985)
(038040,21940)
grunt>
```

Part 4:

The fourth part is to load the text file into Hive and get the average sky ceiling height for each USAF weather station ID.

Step:1 DROP TABLE IF EXISTS records45;

```
hive> DROP TABLE IF EXISTS records45;
OK
Time taken: 6.291 seconds
hive>
```

Step:2 CREATE TABLE records45 (USAF_stationID STRING, ceiling_height INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

```
hive> CREATE TABLE records45 (USAF_stationID STRING, ceiling_height INT)
    > ROW FORMAT DELIMITED
    >   FIELDS TERMINATED BY '\t';
OK
Time taken: 0.56 seconds
hive>
```

Step:3 LOAD DATA LOCAL INPATH 'pig/skyceiling_data.txt'

OVERWRITE INTO TABLE records45;

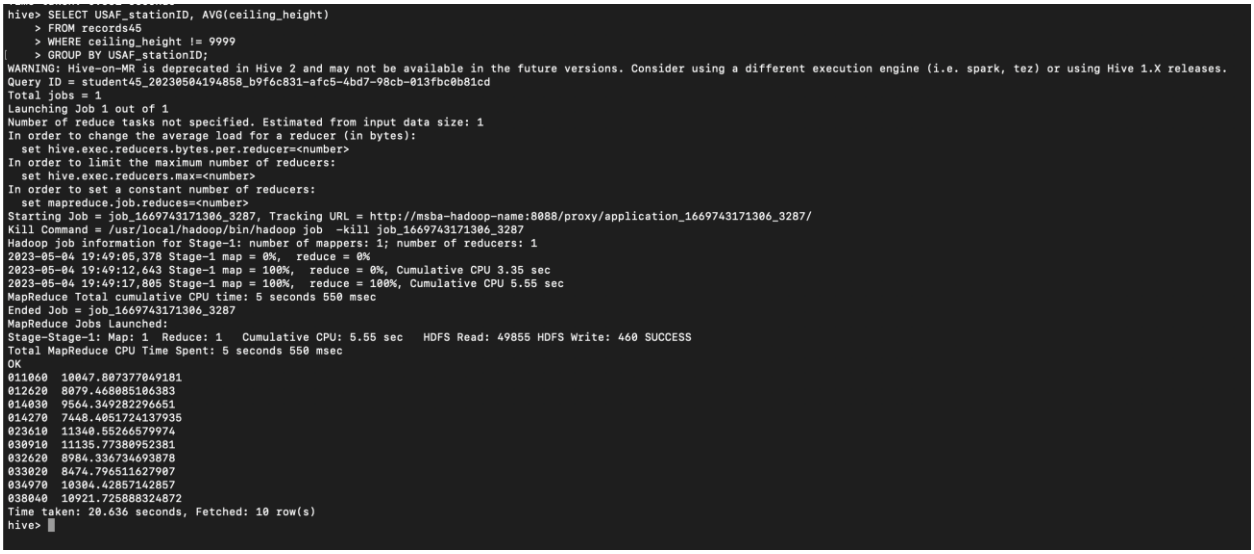
```
hive> LOAD DATA LOCAL INPATH 'pig/skyceiling_data.txt'
    > OVERWRITE INTO TABLE records45;
Loading data to table default.records45
OK
Time taken: 0.531 seconds
hive>
```

Step:4 SELECT USAF_stationID, AVG(ceiling_height)

FROM records45

WHERE ceiling_height != 9999

GROUP BY USAF_stationID;



```
hive> SELECT USAF_stationID, AVG(ceiling_height)
    > FROM records45
    > WHERE ceiling_height != 9999
    > GROUP BY USAF_stationID;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = student45_20230504194858_b9f6c831-afc5-4bd7-98cb-013fbc0b81cd
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1669743171306_3287, Tracking URL = http://msba-hadoop-name:8088/proxy/application_1669743171306_3287/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1669743171306_3287
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-04 19:49:05,378 Stage-1 map = 0%,  reduce = 0%
2023-05-04 19:49:12,643 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.35 sec
2023-05-04 19:49:17,805 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.55 sec
MapReduce Total cumulative CPU time: 5 seconds 550 msec
Ended Job = job_1669743171306_3287
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.55 sec   HDFS Read: 49855 HDFS Write: 460 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 550 msec
OK
011060  10047.807377049181
012620  8079.468085106383
014030  9564.349282296651
014270  7448.4051724137935
023610  11340.55266579974
030910  11135.77380952381
032620  8984.336734693878
033020  8474.796511627907
034970  10304.42857142857
038040  10921.725888324872
Time taken: 20.636 seconds, Fetched: 10 row(s)
hive>
```

You need to turn in:

1) Part 1:

a. *if you are using JAVA to develop the Mapper and Reducer applications:* the three java files (mapper, reducer and main);

b. *if you are using Hadoop streaming jar and developing two python programs (mapper python file and reducer python file)*: the two python files (mapper and reducer);

c. *if you are using mrjob library and developing one python program with two functions:* the python file (with the mapper and reducer functions);

2) Part 2: the python program you developed;

3) the commands from converting java files into a Jar file to running the Jar file in Hadoop, or the commands to execute the python files in Hadoop and in Spark;

4) the step by step commands and screenshots of solutions from all the parts;

The original dataset for this project is available on Blackboard.