

# BAN632\_BIG\_DATA

## The Course Project

## 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).

Using **MRJob**, Calculated the range of sky ceiling height (meters) for each observation month from NCDC records.

→ Connect to university Hadoop system:

**ssh student22@msba-hadoop-name.csueastbay.edu**

→ Copy the .gz files to cyber duck.

→ As they are zipped files, unzip the files:

**gunzip \*-99999-\*.gz**

```

Last login: Fri May 5 00:57:51 on ttys001
(base) nekki@Madhus-MacBook-Pro ~ % ssh student22@msba-hadoop-name.csueastbay.edu
student22@msba-hadoop-name.csueastbay.edu's password:
Last failed login: Fri May 5 01:28:41 PDT 2023 from c-24-130-179-77.hsd1.ca.comcast.net on ssh:notty
There were 3 failed login attempts since the last successful login.
Last login: Thu May 4 21:40:59 2023 from c-24-130-179-77.hsd1.ca.comcast.net
[student22@msba-hadoop-name ~]$ cd Project
[student22@msba-hadoop-name Project]$ gunzip *-99999-*.gz
[student22@msba-hadoop-name Project]$

```

[illegible][illegible]

→ Python MRJob file:

```
MRJob_project_1.py
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

MRJob_project_1.py 1 X
Users > nekki > Desktop > Madhu_Project > Q1 > MRJob_project_1.py > ...

1  from mrjob.job import MRJob
2
3  class MRMaxTemp(MRJob):
4      def mapper(self, _, line):
5          input = line.strip()
6          (month, heights, quality) = (input[19:21], input[70:75], input[75:76])
7          if (heights != '99999' and quality in ['0', '1', '4', '5', '9']):
8              yield(month, heights)
9
10     def reducer(self, month, heights):
11         hmax = 0
12         hmin = next(heights)
13         for i in heights:
14             if i > hmax:
15                 hmax = i
16             if i < hmin:
17                 hmin = i
18         hrange = int(hmax) - int(hmin)
19         yield(month, hrange)
20
21 if __name__ == '__main__':
22     MRMaxTemp.run()
23
```

→ Running MRJob with all gun zipped NCDC files:

```
python MRJob_project_1.py -r local *-99999-* --output-dir=./output_q1/
```

```
[student22@msba-hadoop-name Project]$ python MRJob_project_1.py -r local *-99999-* --output-dir=./output_q1/
No configs found; falling back on auto-configuration
No configs specified for local runner
Creating temp directory /tmp/MRJob_project_1.student22.20230505.114739.122690
Running step 1 of 1...
job output is in ./output_q1/
Removing temp directory /tmp/MRJob_project_1.student22.20230505.114739.122690...
[student22@msba-hadoop-name Project]$
```

→ Output file is created:

mika-hadoop@name.cornell.edu - SFTP  
mika-hadoop@name.cornell.edu

homedir\userdir\22\Project

Filename	Size	Modified
output10		Yesterday, 11:31 PM
MRJob_project_Tray		615 B Yesterday, 11:29 PM
MRJob		641 B Yesterday, 11:28 PM
012620-99999-1928	8.7 KB	Today, 4:29 AM
012470-99999-1929	34.4 KB	Today, 4:29 AM
012620-99999-1928	9.7 KB	Today, 4:29 AM
012470-99999-1928	9.9 KB	Today, 4:29 AM
012470-99999-1930	10.0 KB	Today, 4:29 AM
012620-99999-1930	10.1 KB	Today, 4:29 AM
014030-99999-1928	13.0 KB	Today, 4:29 AM
01060-99999-1928	18.9 KB	Today, 4:29 AM
01060-99999-1929	19.9 KB	Today, 4:29 AM
01060-99999-1930	20.0 KB	Today, 4:29 AM
03090-99999-1927	20.1 KB	Today, 4:29 AM
012620-99999-1927	21.4 KB	Today, 4:29 AM
03090-99999-1928	24.0 KB	Today, 4:29 AM
014070-99999-1927	26.2 KB	Today, 4:29 AM
013020-99999-1927	36.6 KB	Today, 4:29 AM
03040-99999-1927	36.6 KB	Today, 4:29 AM
02340-99999-1928	37.4 KB	Today, 4:29 AM
skyming_data.txt	40.9 KB	5/2/23, 10:11 PM
014030-99999-1929	59.9 KB	Today, 4:29 AM
02340-99999-1930	68.0 KB	Today, 4:29 AM
014030-99999-1930	119.3 KB	Today, 4:29 AM
028970-99999-1928	119.4 KB	Today, 4:29 AM
02910-99999-1924	126.4 KB	Today, 4:29 AM
028970-99999-1921	147.5 KB	Today, 4:29 AM
02910-99999-1926	148.6 KB	Today, 4:29 AM
028970-99999-1922	148.6 KB	Today, 4:29 AM
02910-99999-1925	148.7 KB	Today, 4:29 AM
029350-99999-1923	149.0 KB	Today, 4:29 AM
029350-99999-1922	149.1 KB	Today, 4:29 AM
029350-99999-1924	149.5 KB	Today, 4:29 AM
02910-99999-1922	150.8 KB	Today, 4:29 AM
029350-99999-1925	150.8 KB	Today, 4:29 AM
02910-99999-1921	150.9 KB	Today, 4:29 AM

Filename	Size	Modified
part-00000	11 B	Today, 4:47 AM
part-00001	11 B	Today, 4:47 AM
part-00005	11 B	Today, 4:47 AM
part-00004	22 B	Today, 4:47 AM
part-00003	33 B	Today, 4:47 AM
part-00002	44 B	Today, 4:47 AM

→ Displaying contents in terminal using -cat command

**hdfs dfs -copyFromLocal output\_q1 /home/22student22/Project/  
hdfs dfs -ls /home/22student22/project/output\_q1/  
hdfs dfs -cat /home/22student22/Project/output\_q1/part-0000\***

```
[student22@msba-hadoop-name Project]$ hdfs dfs -copyFromLocal output_q1 /home/22student22/Project/
[student22@msba-hadoop-name Project]$ hdfs dfs -ls /home/22student22/Project/
Found 3 items
drwxr-xr-x - student22 supergroup          0 2023-05-04 21:51 /home/22student22/Project/input
drwxr-xr-x - student22 supergroup          0 2023-05-04 23:33 /home/22student22/Project/output
drwxr-xr-x - student22 supergroup          0 2023-05-05 04:50 /home/22student22/Project/output_q1
[student22@msba-hadoop-name Project]$ hdfs dfs -ls /home/22student22/Project/output_q1/
Found 6 items
-rw-r--r--  5 student22 supergroup          11 2023-05-05 04:50 /home/22student22/Project/output_q1/part-00000
-rw-r--r--  5 student22 supergroup          11 2023-05-05 04:50 /home/22student22/Project/output_q1/part-00001
-rw-r--r--  5 student22 supergroup          44 2023-05-05 04:50 /home/22student22/Project/output_q1/part-00002
-rw-r--r--  5 student22 supergroup          33 2023-05-05 04:50 /home/22student22/Project/output_q1/part-00003
-rw-r--r--  5 student22 supergroup          22 2023-05-05 04:50 /home/22student22/Project/output_q1/part-00004
-rw-r--r--  5 student22 supergroup          11 2023-05-05 04:50 /home/22student22/Project/output_q1/part-00005
```

→ Displaying Final output as “month”      range of sky ceiling heights

```
[student22@msba-hadoop-name Project]$ hdfs dfs -cat /home/22student22/Project/output_q1/part-0000*
"01"    21985
"02"    21985
"03"    21985
"04"    21985
"05"    21985
"06"    21940
"07"    21940
"08"    21985
"09"    21985
"10"    21985
"11"    21985
"12"    21985
[student22@msba-hadoop-name Project]$
```

## 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).

Using **PySpark**, Calculated the average visibility distance(meters) for each USAF weather station ID from NCDC records.

→ Create AWS spark cluster:

→ Create input directory:

**hdfs dfs -mkdir /user/hadoop/input\_project\_q2/**

→ copying contents from local:

**hdfs dfs -copyFromLocal \*-99999-\* /user/hadoop/input\_project\_q2/**

```
[hadoop@ip-172-31-67-253 ~]$ gunzip *-99999-*.gz
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -mkdir /user/hadoop/input_project_q2/
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -copyFromLocal *-99999-* /user/hadoop/input_project_q2/
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -ls /user/hadoop/
Found 4 items
drwxr-xr-x - hadoop hdfsadmin 0 2023-05-05 09:55 /user/hadoop/.sparkStaging
drwxr-xr-x - hadoop hdfsadmin 0 2023-05-05 09:51 /user/hadoop/input
drwxr-xr-x - hadoop hdfsadmin 0 2023-05-05 10:12 /user/hadoop/input_project_q2
drwxr-xr-x - hadoop hdfsadmin 0 2023-05-05 09:55 /user/hadoop/outputdata
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -ls /user/hadoop/input_project_q2/
Found 50 items
-rw-r--r-- 1 hadoop hdfsadmin 18946 2023-05-05 10:12 /user/hadoop/input_project_q2/011060-99999-1928
-rw-r--r-- 1 hadoop hdfsadmin 19934 2023-05-05 10:12 /user/hadoop/input_project_q2/011060-99999-1929
-rw-r--r-- 1 hadoop hdfsadmin 19971 2023-05-05 10:12 /user/hadoop/input_project_q2/011060-99999-1930
-rw-r--r-- 1 hadoop hdfsadmin 8726 2023-05-05 10:12 /user/hadoop/input_project_q2/012620-99999-1928
-rw-r--r-- 1 hadoop hdfsadmin 9657 2023-05-05 10:12 /user/hadoop/input_project_q2/012620-99999-1929
-rw-r--r-- 1 hadoop hdfsadmin 10054 2023-05-05 10:12 /user/hadoop/input_project_q2/012620-99999-1930
-rw-r--r-- 1 hadoop hdfsadmin 13033 2023-05-05 10:12 /user/hadoop/input_project_q2/014030-99999-1928
-rw-r--r-- 1 hadoop hdfsadmin 59900 2023-05-05 10:12 /user/hadoop/input_project_q2/014030-99999-1929
-rw-r--r-- 1 hadoop hdfsadmin 119256 2023-05-05 10:12 /user/hadoop/input_project_q2/014030-99999-1930
-rw-r--r-- 1 hadoop hdfsadmin 9948 2023-05-05 10:12 /user/hadoop/input_project_q2/014270-99999-1928
-rw-r--r-- 1 hadoop hdfsadmin 9442 2023-05-05 10:12 /user/hadoop/input_project_q2/014270-99999-1929
-rw-r--r-- 1 hadoop hdfsadmin 9998 2023-05-05 10:12 /user/hadoop/input_project_q2/014270-99999-1930
-rw-r--r-- 1 hadoop hdfsadmin 37353 2023-05-05 10:12 /user/hadoop/input_project_q2/023610-99999-1929
-rw-r--r-- 1 hadoop hdfsadmin 87958 2023-05-05 10:12 /user/hadoop/input_project_q2/023610-99999-1930
-rw-r--r-- 1 hadoop hdfsadmin 151540 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1921
-rw-r--r-- 1 hadoop hdfsadmin 152226 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1922
-rw-r--r-- 1 hadoop hdfsadmin 151749 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1923
-rw-r--r-- 1 hadoop hdfsadmin 153031 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1924
-rw-r--r-- 1 hadoop hdfsadmin 150831 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1925
-rw-r--r-- 1 hadoop hdfsadmin 151838 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1926
-rw-r--r-- 1 hadoop hdfsadmin 147469 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1921
-rw-r--r-- 1 hadoop hdfsadmin 148648 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1922
-rw-r--r-- 1 hadoop hdfsadmin 151594 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1923
-rw-r--r-- 1 hadoop hdfsadmin 152011 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1924
-rw-r--r-- 1 hadoop hdfsadmin 151430 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1925
-rw-r--r-- 1 hadoop hdfsadmin 119414 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1926
-rw-r--r-- 1 hadoop hdfsadmin 150855 2023-05-05 10:12 /user/hadoop/input_project_q2/029110-99999-1921
```

## → Application code:

```
Spark_q2.py
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More
MRJob_project_1.py 1 Spark_q2.py 1 X
Users > nekki > Desktop > Madhu_Project > Spark_Q2 > Spark_q2.py > ...
1 from pyspark import SparkContext
2
3 def main():
4     sc = SparkContext(appName='SparkVisibility')
5
6     input_file = sc.textFile('/user/hadoop/input_project_q2/*-99999-*.')
7     station_vdist = input_file.filter(lambda line: line[78:84] != '999999' and line[84:85] in ['0','1','4','5','9']) \
8         .map(lambda line: (line[4:10], (int(line[78:84]), 1))) \
9         .reduceByKey(lambda a, b: (a[0]+b[0], a[1]+b[1])) \
10        .map(lambda x: (x[0], x[1][0]/x[1][1]))
11    station_vdist.saveAsTextFile('/user/hadoop/outputdata/output_project_q2.txt')
12
13    sc.stop()
14
15 if __name__ == '__main__':
16     main()
17
```

## → Running python file in spark:

**spark-submit --master yarn Spark\_q2.py**

```
[hadoop@ip-172-31-67-253 ~]$ spark-submit --master yarn Spark_q2.py
23/06/05 10:15:37 INFO SparkContext: Running Spark version 2.4.0-shm-2
23/06/05 10:15:37 INFO SparkContext: Submitted application: Spark_q2
23/06/05 10:15:37 INFO SecurityManager: Changing view acls to: hadoop
23/06/05 10:15:37 INFO SecurityManager: Changing modify acls to: hadoop
23/06/05 10:15:37 INFO SecurityManager: Changing view acls groups to:
23/06/05 10:15:37 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with view permissions: Set(
23/06/05 10:15:37 INFO SecurityManager: Changing modify acls groups to:
23/06/05 10:15:37 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with view permissions: Set(
23/06/05 10:15:37 INFO SparkEnv: Registering MapOutputTracker
23/06/05 10:15:37 INFO SparkEnv: Registering BlockManagerMaster
23/06/05 10:15:37 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
23/06/05 10:15:37 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
23/06/05 10:15:37 INFO DiskManager: Created local directory at /mnt/tmp/blockmgr-436dc377-114c-4866-9818-d47d5a804866
23/06/05 10:15:37 INFO MemoryStore: MemoryStore started with capacity 912.3 MB
23/06/05 10:15:37 INFO SparkEnv: Registering OutputCommitCoordinator
23/06/05 10:15:37 INFO SparkEnv: Registering OutputCommitCoordinator
23/06/05 10:15:38 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://ip-172-31-67-253.ec2.internal:4040
23/06/05 10:15:38 INFO UIUtils: Using 100 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
23/06/05 10:15:38 INFO RMProxy: Connecting to ResourceManager at ip-172-31-67-253.ec2.internal/172.31.67.253:8032
23/06/05 10:15:39 INFO Client: Requesting a new application from cluster with 1 NodeManagers
23/06/05 10:15:39 INFO Configuration: resource-types.xml not found
23/06/05 10:15:39 INFO ResourceUtils: Unable to find 'resource-types.xml'.
23/06/05 10:15:39 INFO ResourceUtils: Adding resource type - name = memory-mb, units = M, type = COUNTABLE
23/06/05 10:15:39 INFO ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
23/06/05 10:15:39 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
23/06/05 10:15:39 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
23/06/05 10:15:39 INFO Client: Setting up container launch context for our AM
23/06/05 10:15:39 INFO Client: Settling on the launch environment for our AM container
23/06/05 10:15:39 INFO Client: Preparing resources for our AM container
23/06/05 10:15:39 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
23/06/05 10:15:41 INFO Client: Uploading resource file:/mnt/tmp/spark-138ecce1-fd5e-4db7-990b-8f88fc6f729/_spark_libs_6378385683753864938.zip -> hdfs://ip-172-31-67-253.ec2.internal:8020/user/hadoop/sparkStaging/application_1683279664989_0002/_spark_libs_
3703864938.zip
23/06/05 10:15:42 INFO Client: Uploading resource file:/etc/hudi/conf/hudi-defaults.conf -> hdfs://ip-172-31-67-253.ec2.internal:8020/user/hadoop/sparkStaging/application_1683279664989_0002/hudi-defaults.conf
23/06/05 10:15:42 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-67-253.ec2.internal:8020/user/hadoop/sparkStaging/application_1683279664989_0002/pyspark.zip
23/06/05 10:15:42 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.7-src.zip -> hdfs://ip-172-31-67-253.ec2.internal:8020/user/hadoop/sparkStaging/application_1683279664989_0002/py4j-0.10.7-src.zip
23/06/05 10:15:42 INFO Client: Uploading resource file:/mnt/tmp/spark-138ecce1-fd5e-4db7-990b-8f88fc6f729/_spark_conf_65843460674945735.zip -> hdfs://ip-172-31-67-253.ec2.internal:8020/user/hadoop/sparkStaging/application_1683279664989_0002/_spark_conf_2
23/06/05 10:15:42 INFO SecurityManager: Changing view acls to: hadoop
23/06/05 10:15:42 INFO SecurityManager: Changing modify acls to: hadoop
23/06/05 10:15:42 INFO SecurityManager: Changing view acls groups to:
23/06/05 10:15:42 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with view permissions: Set(
23/06/05 10:15:42 INFO SecurityManager: Changing modify acls groups to:
23/06/05 10:15:42 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with view permissions: Set(
23/06/05 10:15:43 INFO SchedulerExtensionServices: Starting yarn extension services with app application_1683279664989_0002 and attemptId None
23/06/05 10:15:44 INFO Client: Application report for application_1683279664989_0002 (state: ACCEPTED)
23/06/05 10:15:44 INFO Client:
  client token: N/A
  diagnostics: N/A
  ApplicationMaster host: 172.31.67.253
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1683281743265
  final status: UNDEFINED
  tracking URL: http://ip-172-31-67-253.ec2.internal:20888/proxy/application_1683279664989_0002/
  user: hadoop
23/06/05 10:15:45 INFO Client: Application report for application_1683279664989_0002 (state: ACCEPTED)
23/06/05 10:15:46 INFO Client: Application report for application_1683279664989_0002 (state: ACCEPTED)
23/06/05 10:15:47 INFO YarnClientSchedulerBackend: Add WebUI Filter: org.apache.hadoop.yarn.server.webproxy.amfilter.AmFilter, Map(PROXY_HOSTS -> ip-172-31-67-253.ec2.internal, PROXY_URI_BASES -> http://ip-172-31-67-253.ec2.internal:20888/proxy/application_16
89_0002) /proxy/application_1683279664989_0002
23/06/05 10:15:47 INFO Client: Application report for application_1683279664989_0002 (state: RUNNING)
23/06/05 10:15:47 INFO Client:
  client token: N/A
  diagnostics: N/A
  ApplicationMaster host: 172.31.67.253
```

[illegible]

```
hdfs dfs -cat /user/hadoop/outputdata/output project q2.txt/part-000*
```

```
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -cat /user/hadoop/outputdata/output_project_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)
```



## 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.

→ Enter pig:

**pig -x local**

→ Load text file in to records variable:

```
records = LOAD 'Project/skyceiling_data.txt'  
As (USAF_ID: chararray, HEIGHTS: int);
```

[illegible]

→ Dumping records:

**dump records;**

```

2023-05-03 22:48:19.128 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize
2023-05-03 22:48:19.129 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize
2023-05-03 22:48:19.130 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize
2023-05-03 22:48:19.141 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer
2023-05-03 22:48:19.144 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.p
2023-05-03 22:48:19.144 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend h
2023-05-03 22:48:19.157 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total
2023-05-03 22:48:19.157 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MappedDut
(11/66/ 22000)
(11/66/ 450)
(11/66/ 450)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 7500)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 3000)
(11/66/ 3000)
(11/66/ 3000)
(11/66/ 240)
(11/66/ 22000)
(11/66/ 1230)
(11/66/ 22000)
(11/66/ 780)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 1230)
(11/66/ 3000)
(11/66/ 3000)
(11/66/ 240)
(11/66/ 22000)
(11/66/ 7500)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)
(11/66/ 22000)

```

→ Grouping the records:  
**grouped\_records = GROUP records BY USAF\_ID;**

```

2023-05-03 22:47:08,632 [main] INFO org.apache.pig.tools.pigstats.ScriptStats - Pig features used in the script: GROUP BY
grunt> DUMP grouped_records;
2023-05-03 22:47:08,632 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes-per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-03 22:47:08,646 [main] INFO org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-05-03 22:47:08,646 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - (RULES_ENABLED={AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter,
10Optimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushFilter, SplitFilter, StreamTypeCastInspector})
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - File concatenation threshold: 100 optimistic? false
2023-05-03 22:47:08,659 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2023-05-03 22:47:08,659 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2023-05-03 22:47:08,669 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes-per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-03 22:47:08,676 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-03 22:47:08,676 [main] INFO org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-05-03 22:47:08,672 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapped.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2023-05-03 22:47:08,673 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2023-05-03 22:47:08,674 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reduce estimator: org.apache.pig.backend.hadoop.executionengine.mapRed
2023-05-03 22:47:08,676 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=40800
2023-05-03 22:47:08,677 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using Parallelism to 1
2023-05-03 22:47:08,677 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2023-05-03 22:47:08,677 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2023-05-03 22:47:08,676 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2023-05-03 22:47:08,676 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp direc
2023-05-03 22:47:08,676 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.WapHadoopLauncher - I mapped reduce job(s) waiting for submission.
2023-05-03 22:47:08,731 [JobControl] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-03 22:47:08,740 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or JobSetJar(String).
2023-05-03 22:47:08,743 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2023-05-03 22:47:08,743 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-05-03 22:47:08,743 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-05-03 22:47:08,744 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-05-03 22:47:08,747 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits: 1
2023-05-03 22:47:08,757 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local73239210_0002
2023-05-03 22:47:08,757 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job is http://localhost:8080/
2023-05-03 22:47:08,833 [Thread-11] INFO org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2023-05-03 22:47:08,820 [Thread-11] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes-per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-03 22:47:08,820 [Thread-11] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2023-05-03 22:47:08,820 [Thread-11] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-05-03 22:47:08,820 [Thread-11] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.reduce.shuffle.parallelcopies is deprecated. Instead, use mapreduce.reduce.markreset.buffe
2023-05-03 22:47:08,820 [Thread-11] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 1
2023-05-03 22:47:08,820 [Thread-11] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup_temporary folders under output directory: false, ignore clea
2023-05-03 22:47:08,821 [Thread-11] INFO org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter is org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigOutputCommitter
2023-05-03 22:47:08,821 [Thread-11] INFO org.apache.hadoop.mapred.LocalJobRunner - Waiting for map tasks
2023-05-03 22:47:08,835 [LocalJobRunnerMap Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Starting task: attempt_local73239210_0002_m_000000_0
2023-05-03 22:47:08,835 [LocalJobRunnerMap Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 1
2023-05-03 22:47:08,833 [LocalJobRunnerMap Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup_temporary folders under output di
2023-05-03 22:47:08,833 [LocalJobRunnerMap Task Executor #0] INFO org.apache.hadoop.mapred.Task - Using ResourceCalculatorProcessTree : [ ]
2023-05-03 22:47:08,835 [LocalJobRunnerMap Task Executor #0] INFO org.apache.hadoop.mapred.MapTask - Processing split: Number of splits 1
Total Length = 408000
Input split[0]:

```

## DUMP grouped\_records;

[illegible]



→ Finding range:

**USAF RANGE = FOREACH grouped\_records GENERATE group,  
MAX(records.HEIGHTS) – MIN(records.HEIGHTS):**

[illegible]**dump USAF RANGE;**

```

Input(s):
Successfully read 3391 records from: "file:///home/student22/Project/skycelling_data.txt"

Output(s):
Successfully stored 10 records in: "file:/tmp/temp-629790862/tmp-1307259655"

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_local152306330_0003

2023-05-03 22:53:03,980 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-03 22:53:03,981 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-03 22:53:03,982 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-05-03 22:53:03,984 [main] INFO org.apache.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - Success!
2023-05-03 22:53:03,985 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-05-03 22:53:03,985 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-05-03 22:53:03,994 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-05-03 22:53:03,994 [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)

```

## 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.

```
mv metastore_db metastore_db.tmp  
schematool -initSchema -dbType derby
```

```
[student22@msba-hadoop-name ~]$ mv metastore_db metastore_db.tmp  
[student22@msba-hadoop-name ~]$ schematool -initSchema -dbType derby  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/hive-2.3.2/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.9.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]  
Metastore connection URL: jdbc:derby:;databaseName=metastore_db;create=true  
Metastore Connection Driver : org.apache.derby.jdbc.EmbeddedDriver  
Metastore connection User: APP  
Starting metastore schema initialization to 2.3.0  
Initialization script hive-schema-2.3.0.derby.sql  
Initialization script completed  
schemaTool completed  
[student22@msba-hadoop-name ~]$
```

```
$ hive  
show tables;
```

```
[student22@msba-hadoop-name ~]$ hive  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/hive-2.3.2/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.9.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]  
  
Logging initialized using configuration in jar:file:/usr/local/hive-2.3.2/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true  
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.  
hive> show tables;  
OK  
Time taken: 5.271 seconds  
hive>
```

```
drop table if exists USAF_records;
```

```
create table USAF_records (USAF_ID string, HEIGHTS int)
```

```
> row format delimited
```

```
>fields terminated by 'It';
```

```
load data local inpath 'Project/skyceiling_data.txt' overwrite into table  
USAF_records;
```

```
select * from USAF_records;
```

```
nekki — student22@msba-hadoop-name:~ — ssh student22@msba-hadoop-name.csueastbay.edu — 2

Logging initialized using configuration in jar:file:/usr/local/hive-2.3.2/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
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.
hive> show tables;
OK
Time taken: 5.271 seconds
hive> drop table if exists USAF_records;
OK
Time taken: 0.064 seconds
hive> create table USAF_records(USAF_ID string, HEIGHTS int)
  > row format delimited
  > fields terminated by '\t';
OK
Time taken: 0.543 seconds
hive> show tables;
OK
usaf_records
Time taken: 0.022 seconds, Fetched: 1 row(s)
hive> load data local inpath 'Project/skyceiling_data.txt' overwrite into table USAF_records;
Loading data to table default.usaf_records
OK
Time taken: 0.637 seconds
hive> select * from USAF_records;
OK
011060 22000
011060 450
011060 450
011060 22000
011060 22000
011060 22000
011060 7500
011060 22000
011060 22000
011060 3000
011060 3000
011060 3000
011060 240
011060 22000
011060 1230
011060 22000
011060 22000
011060 22000
011060 22000
011060 1230
011060 1230
011060 3000
011060 3000
011060 22000
011060 780
011060 780
011060 7500
011060 15
011060 1230
011060 3000
011060 3000
011060 22000
011060 22000
011060 7500
011060 22000
011060 22000
011060 22000
011060 22000
011060 22000
011060 3000
011060 3000
```

**select USAF\_ID, avg (HEIGHTS)  
from USAF\_records  
group by USAF\_ID;**

```
hive> select USAF_ID, avg(HEIGHTS)
  > from USAF_records
  > group by USAF_ID;
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 = student22_20230503234949_e0377a22-5e38-40bd-823d-a99a41ae57aa
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_3199, Tracking URL = http://msba-hadoop-name:8088/proxy/application_1669743171306_3199/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1669743171306_3199
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-03 23:49:56,638 Stage-1 map = 0%, reduce = 0%
2023-05-03 23:50:01,846 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.1 sec
2023-05-03 23:50:09,006 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.11 sec
MapReduce Total cumulative CPU time: 4 seconds 110 msec
Ended Job = job_1669743171306_3199
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.11 sec HDFS Read: 49473 HDFS Write: 460 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 110 msec
OK
011060 10047.007377049181
012620 8079.468085106383
014030 9564.349282296651
014270 7448.4051724137935
023610 11340.55266579974
030910 11135.77380952391
032620 8904.336734693070
033020 8474.796511627907
034970 10304.42857142857
038040 10921.725888324872
Time taken: 19.443 seconds, Fetched: 10 row(s)
hive>
```