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.

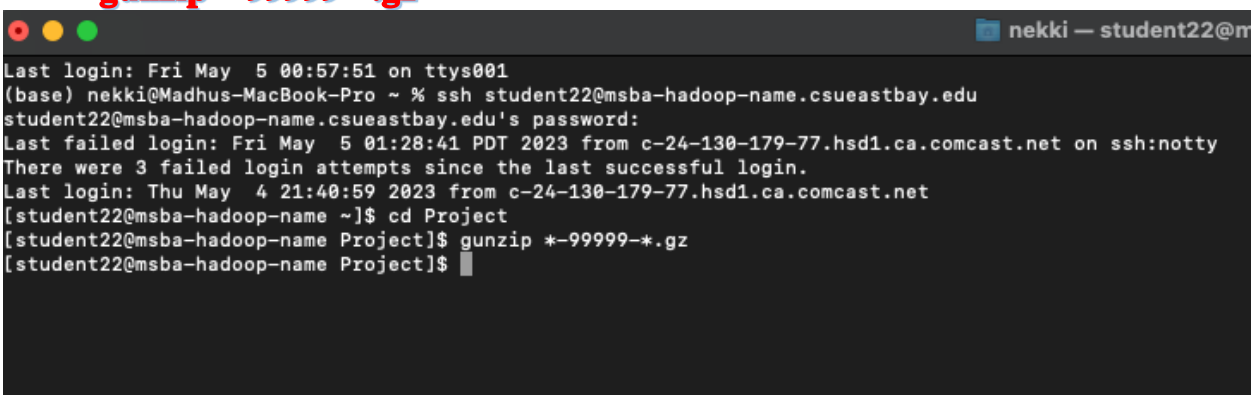
→ Connect to university Hadoop system:

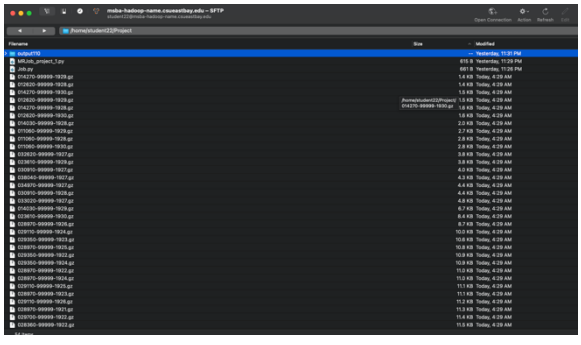
ssh student22@msba-hadoop-name.csueastbay.edu

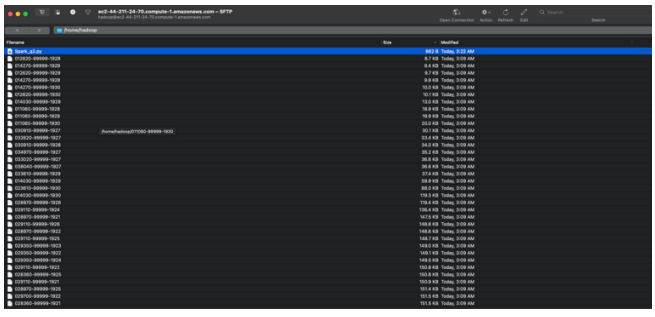
→ Copy the .gz files to cyber duck.

→ As they are zipped files, unzip the files:

gunzip *-99999-*.gz







→ 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

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

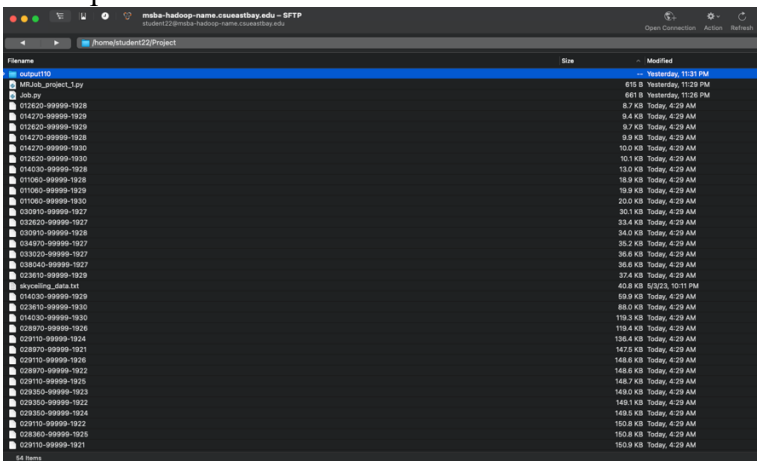
```python
from mrjob.job import MRJob

class MRMaxTemp(MRJob):
    def mapper(self, _, line):
        input = line.strip()
        (month, heights, quality) = (input[19:21], input[70:75], input[75:76])
        if(heights != '99999' and quality in ['0','1','4','5','9']):
            yield(month, heights)

    def reducer(self, month, heights):
        hmax = 0
        hmin = next(heights)
        for i in heights:
            if(i>hmax):
                hmax=i
            if(i<hmin):
                hmin=i
        hrange = int(hmax)-int(hmin)
        yield(month, hrange)

if __name__ == '__main__':
    MRMaxTemp.run()
```

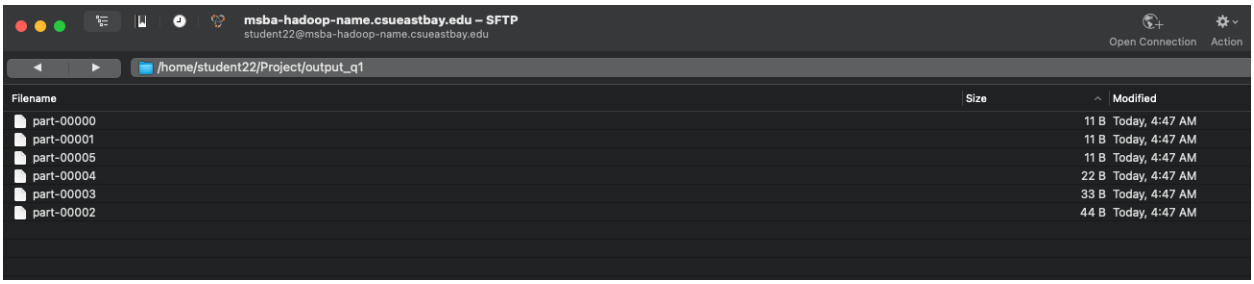
→ Running MRJob with all gun zipped NCDC files:

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:





→ 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:

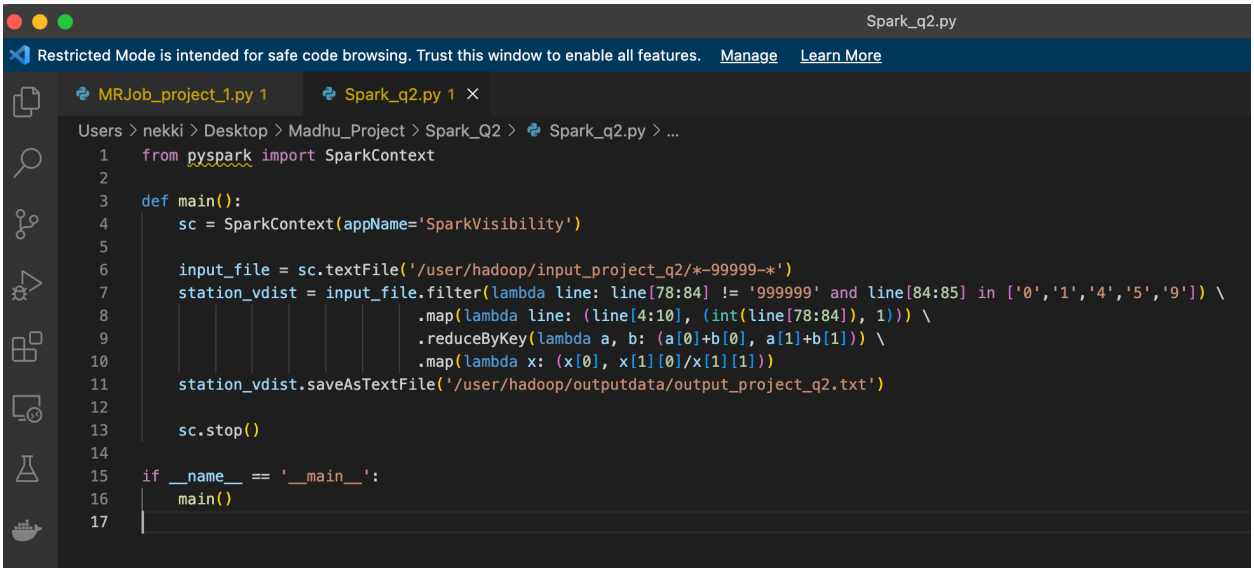
hdfs dfs -mkdir /user/hadoop/input_project_q2/

→ 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 hdfsadmingroup          0 2023-05-05 09:55 /user/hadoop/.sparkStaging
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 09:51 /user/hadoop/input
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 10:12 /user/hadoop/input_project_q2
drwxr-xr-x   - hadoop hdfsadmingroup          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 hdfsadmingroup      18946 2023-05-05 10:12 /user/hadoop/input_project_q2/011060-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup      19934 2023-05-05 10:12 /user/hadoop/input_project_q2/011060-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup      19971 2023-05-05 10:12 /user/hadoop/input_project_q2/011060-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup       8726 2023-05-05 10:12 /user/hadoop/input_project_q2/012620-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup       9657 2023-05-05 10:12 /user/hadoop/input_project_q2/012620-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup      10054 2023-05-05 10:12 /user/hadoop/input_project_q2/012620-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup      13033 2023-05-05 10:12 /user/hadoop/input_project_q2/014030-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup      59900 2023-05-05 10:12 /user/hadoop/input_project_q2/014030-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup     119256 2023-05-05 10:12 /user/hadoop/input_project_q2/014030-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup       9948 2023-05-05 10:12 /user/hadoop/input_project_q2/014270-99999-1928
-rw-r--r--   1 hadoop hdfsadmingroup       9442 2023-05-05 10:12 /user/hadoop/input_project_q2/014270-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup       9998 2023-05-05 10:12 /user/hadoop/input_project_q2/014270-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup      37353 2023-05-05 10:12 /user/hadoop/input_project_q2/023610-99999-1929
-rw-r--r--   1 hadoop hdfsadmingroup      87958 2023-05-05 10:12 /user/hadoop/input_project_q2/023610-99999-1930
-rw-r--r--   1 hadoop hdfsadmingroup     151540 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     152226 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     151749 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     153031 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     150831 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     151838 2023-05-05 10:12 /user/hadoop/input_project_q2/028360-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup     147469 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1921
-rw-r--r--   1 hadoop hdfsadmingroup     148648 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1922
-rw-r--r--   1 hadoop hdfsadmingroup     151594 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1923
-rw-r--r--   1 hadoop hdfsadmingroup     152011 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1924
-rw-r--r--   1 hadoop hdfsadmingroup     151430 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1925
-rw-r--r--   1 hadoop hdfsadmingroup     119414 2023-05-05 10:12 /user/hadoop/input_project_q2/028970-99999-1926
-rw-r--r--   1 hadoop hdfsadmingroup     150855 2023-05-05 10:12 /user/hadoop/input_project_q2/029110-99999-1921
```

→ Application code:



→ Running python file in spark:

spark-submit --master yarn Spark_q2.py

→ Output file is created:

```
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -ls /user/hadoop/
Found 4 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/.sparkStaging
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 09:51 /user/hadoop/input
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 10:12 /user/hadoop/input_project_q2
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -ls /user/hadoop/outputdata/
Found 2 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 09:55 /user/hadoop/outputdata/205.txt
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt
[hadoop@ip-172-31-67-253 ~]$ hdfs dfs -ls /user/hadoop/outputdata/output_project_q2.txt/
Found 51 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00002
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00003
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00004
-rw-r--r--   1 hadoop hdfsadmingroup         47 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00005
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00006
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00007
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00008
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00009
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00010
-rw-r--r--   1 hadoop hdfsadmingroup         29 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00011
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00012
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00013
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00014
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00015
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00016
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00017
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00018
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00019
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00020
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00021
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00022
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00023
-rw-r--r--   1 hadoop hdfsadmingroup         30 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00024
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00025
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00026
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00027
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00028
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00029
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00030
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00031
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00032
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00033
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00034
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00035
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00036
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00037
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00038
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00039
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00040
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00041
-rw-r--r--   1 hadoop hdfsadmingroup         16 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00042
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00043
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00044
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00045
-rw-r--r--   1 hadoop hdfsadmingroup         31 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00046
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00047
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00048
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-05-05 10:15 /user/hadoop/outputdata/output_project_q2.txt/part-00049
```

→ Displaying contents in output file:

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.

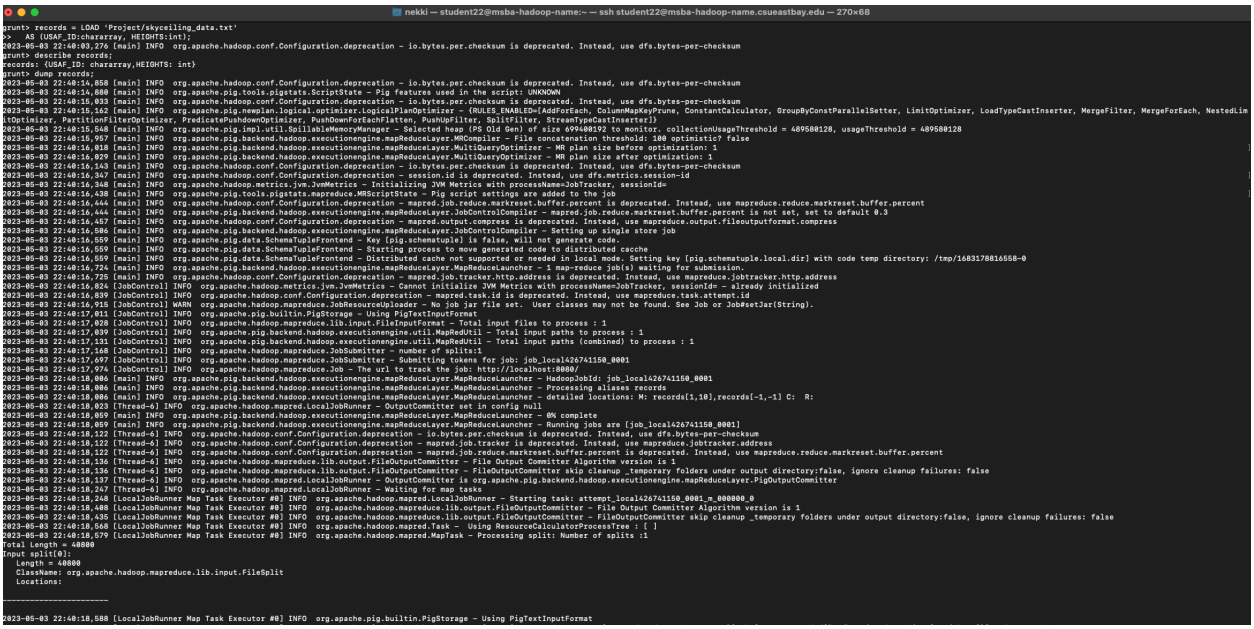
→ Enter pig:

pig -x local

→ Load text file in to records variable:

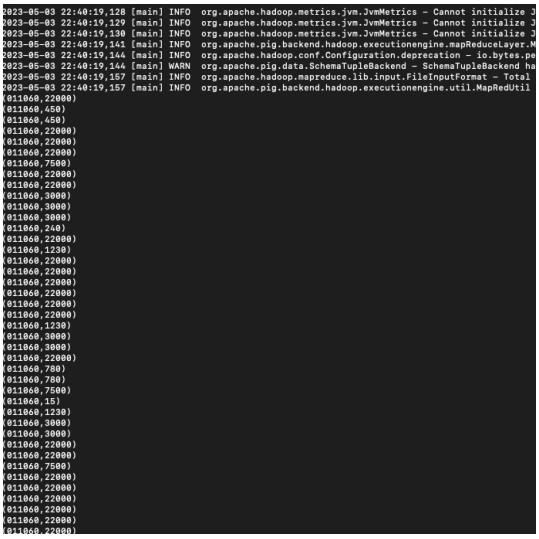
records = LOAD 'Project/skyceiling_data.txt'

As (USAF_ID: chararray, HEIGHTS: int);



→ Dumping records:

dump records;



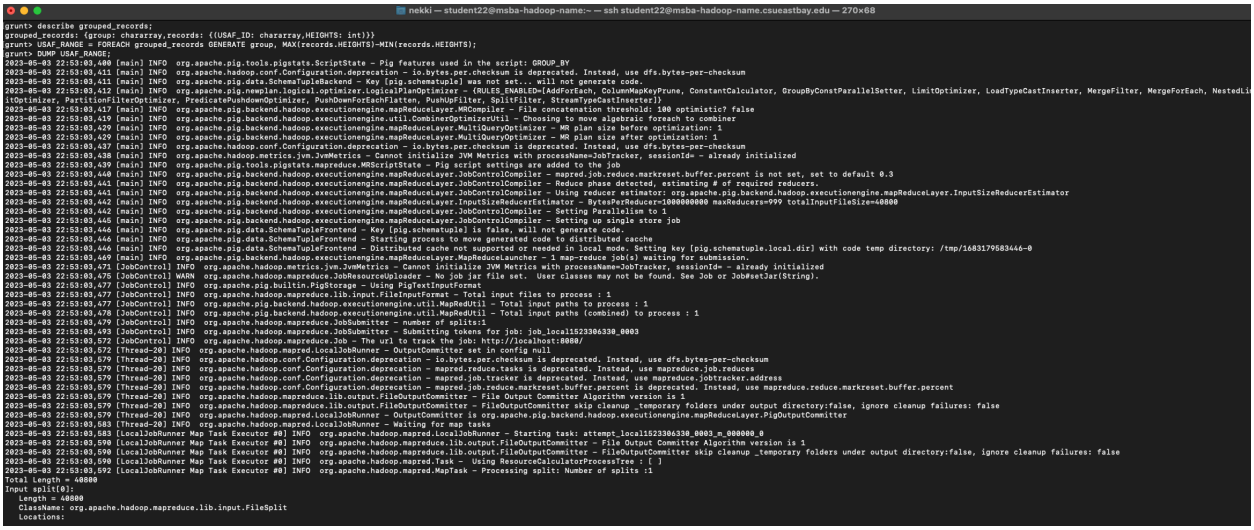
→ Grouping the records:

grouped_records = GROUP records BY USAF_ID;

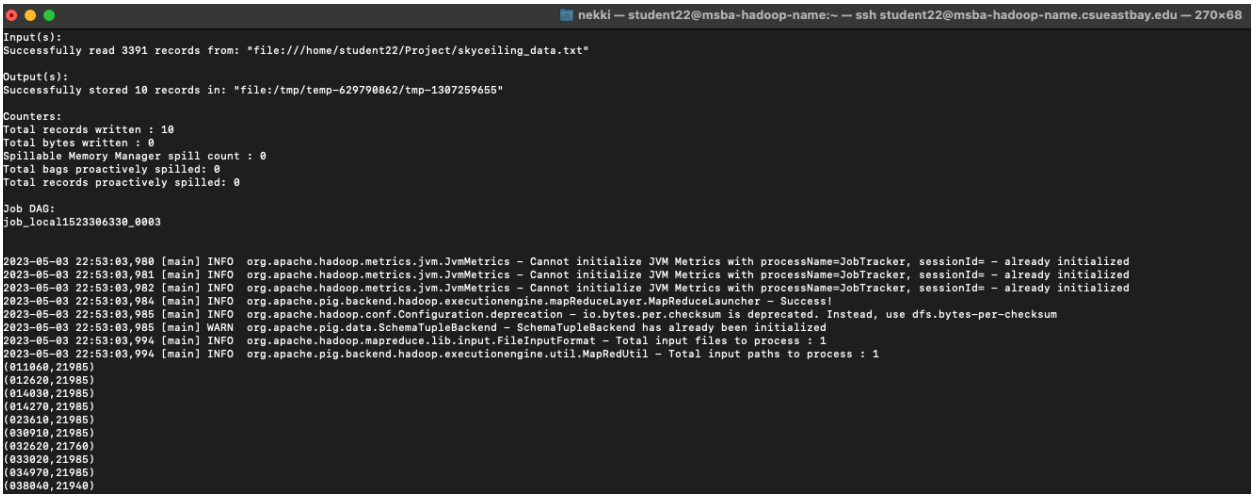
DUMP grouped _records;

→ Finding range:

USAF RANGE = FOREACH grouped_records GENERATE group, MAX(records.HEIGTHS) – MIN(records.HEIGHTS):



dump USAF_RANGE;



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;

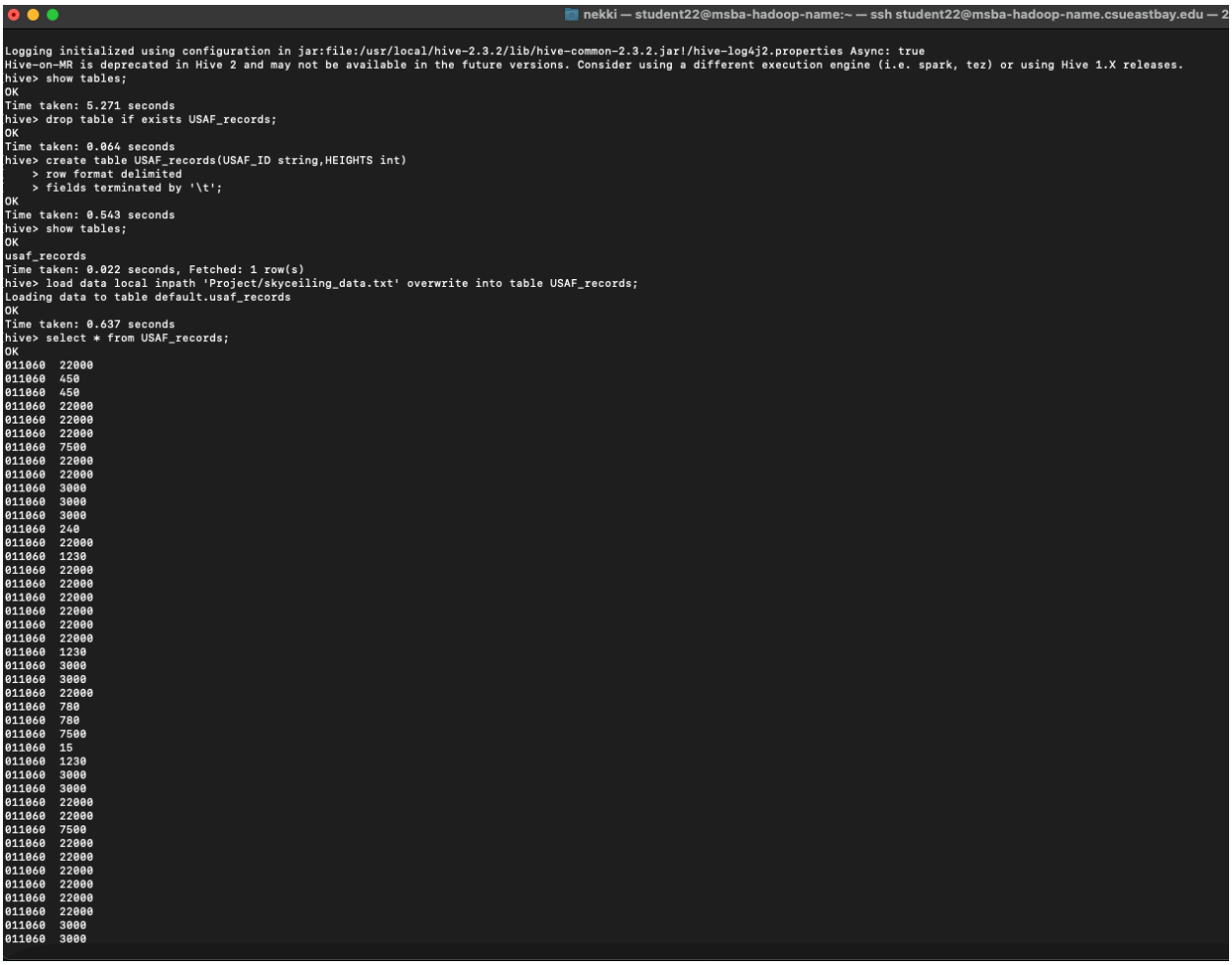
```
[student22@msba-hadoop-name ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive-2.3.2/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.9.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.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;



select USAF_ID, avg (HEIGHTS)
from USAF_records
group by USAF_ID;