The Course Project

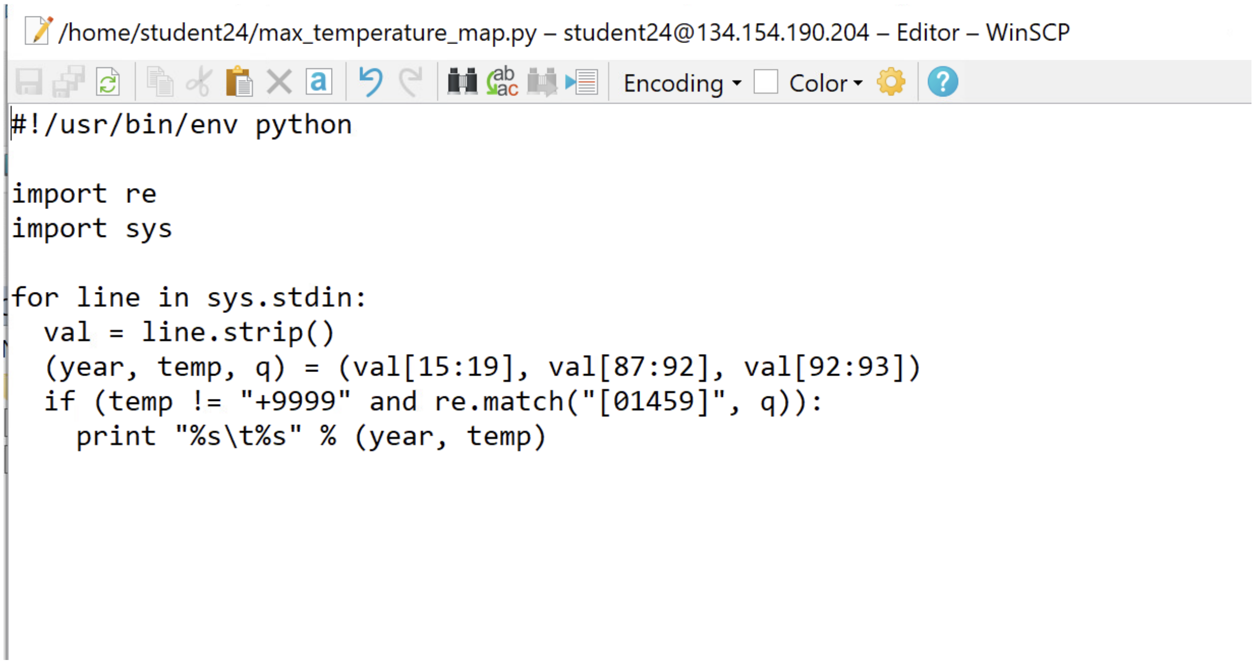
The course project includes 3 parts. The first part is to develop a Python application to retrieve Year and Temperature from original NCDC records (i.e., the dataset we are using for this class) and then write the Year and Temperature data into a text file. The second part is to load the text file into Pig and get the highest and lowest temperatures for each year. The third part is to load the text file into Hive and get the average temperature for each year.

You need to turn in 1) Python files (mapper and reducer), 2) the commands for executing the Python application in Hadoop, 3) the text file including Year and Temperature data created by you, 4) the screenshot of the text file being created, 5) the screenshot of the final Pig output showing the year and the highest and lowest temperatures, and 6) the screenshot of the final Hive output showing the year and average temperature.

The original dataset for this project is available on Canvas.

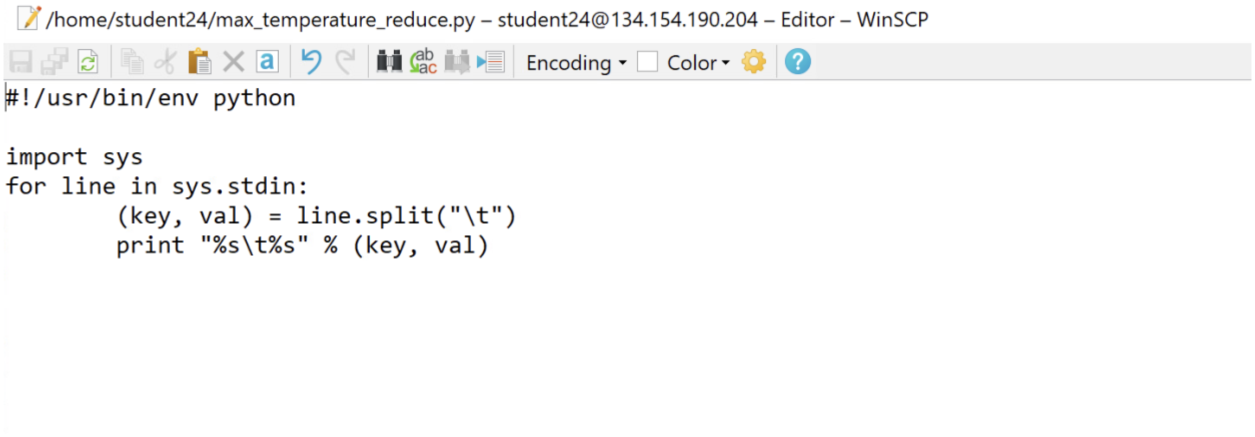
Python Files:

Mapper:


/home/student24/max_temperature_map.py – student24@134.154.190.204 – Editor – WinSCP
Encoding
Color

```
#!/usr/bin/env python

import re
import sys

for line in sys.stdin:
  val = line.strip()
  (year, temp, q) = (val[15:19], val[87:92], val[92:93])
  if (temp != "+9999" and re.match("[01459]", q)):
    print "%s\t%s" % (year, temp)
```

Reducer:


/home/student24/max_temperature_reduce.py – student24@134.154.190.204 – Editor – WinSCP
Encoding
Color

```
#!/usr/bin/env python

import sys
for line in sys.stdin:
        (key, val) = line.split("\t")
        print "%s\t%s" % (key, val)
```

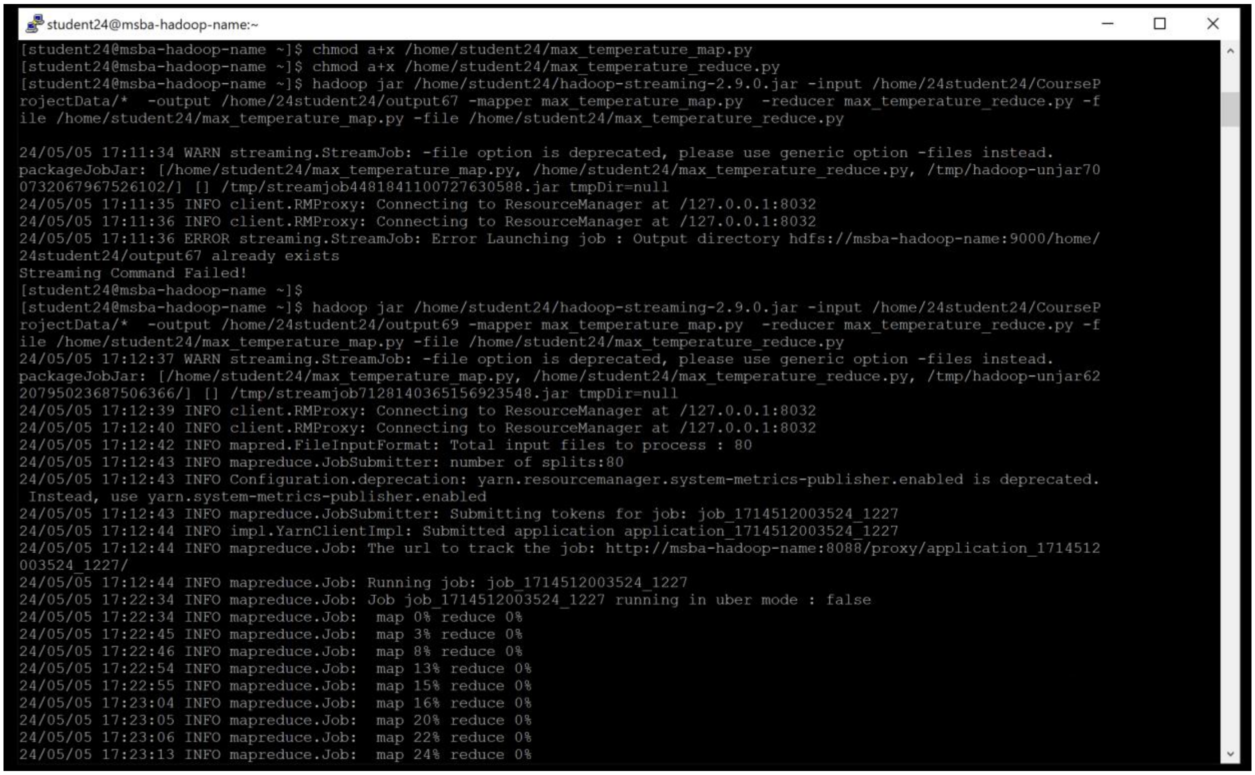
Running the Job:

```
chmod a+x /home/student24/max_temperature_map.py
```

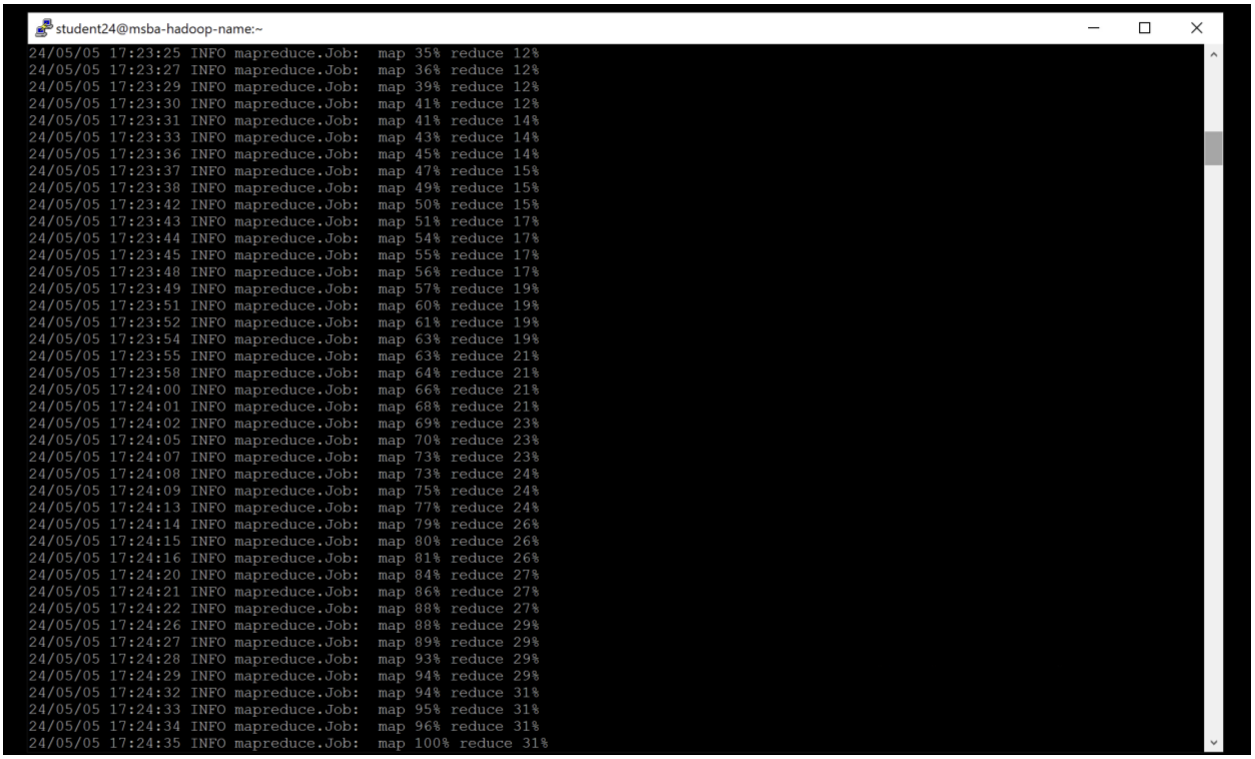
```
chmod a+x /home/student24/max_temperature_reduce.py
```

```
hadoop jar /home/student24/hadoop-streaming-2.9.0.jar -input /home/24student24/CourseProjectData/*  -output /home/24student24/output69 -mapper max_temperature_map.py  -reducer max_temperature_reduce.py -file /home/student24/max_temperature_map.py -file /home/student24/max_temperature_reduce.py
```

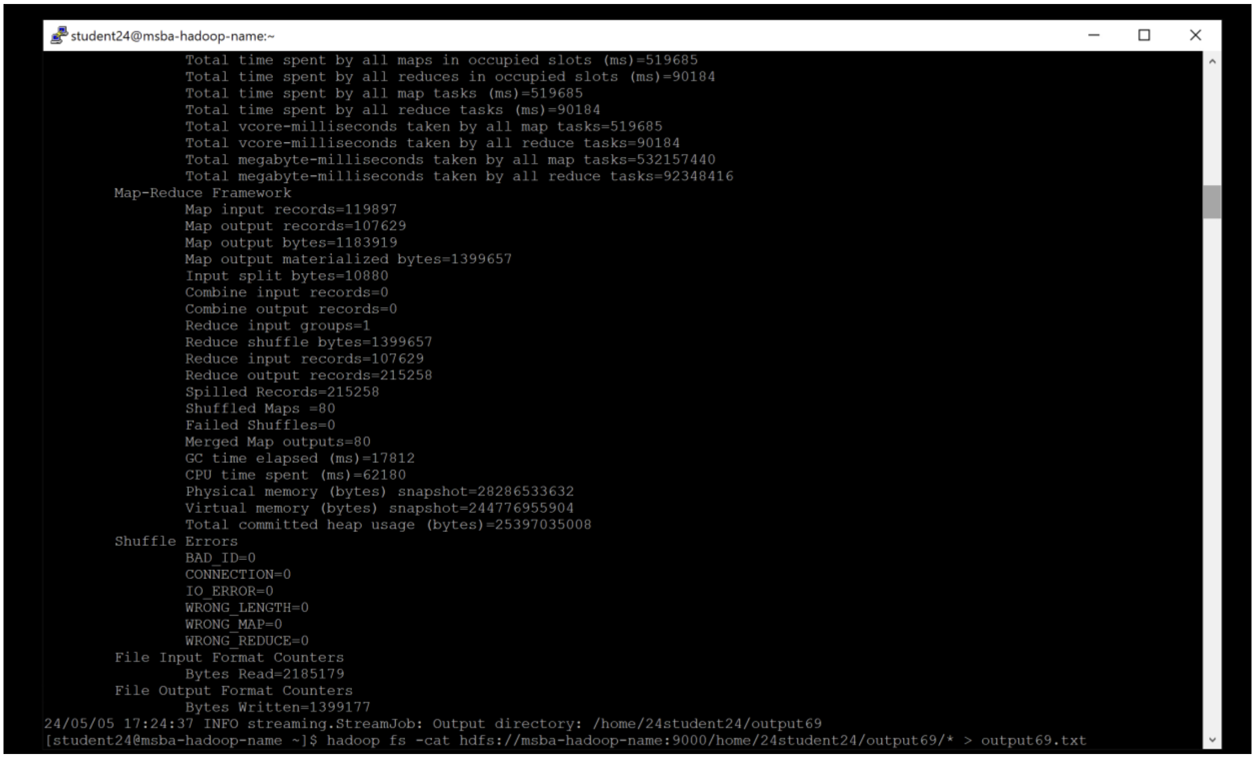
```
hadoop fs -cat hdfs://msba-hadoop-name:9000/home/24student24/output69/* > output69.txt
```



```
[student24@msba-hadoop-name ~]$ chmod a+x /home/student24/max_temperature_map.py
[student24@msba-hadoop-name ~]$ chmod a+x /home/student24/max_temperature_reduce.py
[student24@msba-hadoop-name ~]$ hadoop jar /home/student24/hadoop-streaming-2.9.0.jar -input /home/24student24/CourseP
rojectData/*  -output /home/24student24/output67 -mapper max_temperature_map.py  -reducer max_temperature_reduce.py -f
ile /home/student24/max_temperature_map.py -file /home/student24/max_temperature_reduce.py

24/05/05 17:11:34 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/student24/max_temperature_map.py, /home/student24/max_temperature_reduce.py, /tmp/hadoop-unjar70
07320679675261O2/] [] /tmp/streamjob4481841100727630588.jar tmpDir=null
24/05/05 17:11:35 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 17:11:36 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 17:11:36 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://msba-hadoop-name:9000/home/
24student24/output67 already exists
Streaming Command Failed!
[student24@msba-hadoop-name ~]$
[student24@msba-hadoop-name ~]$ hadoop jar /home/student24/hadoop-streaming-2.9.0.jar -input /home/24student24/CourseP
rojectData/*  -output /home/24student24/output69 -mapper max_temperature_map.py  -reducer max_temperature_reduce.py -f
ile /home/student24/max_temperature_map.py -file /home/student24/max_temperature_reduce.py
24/05/05 17:12:37 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/student24/max_temperature_map.py, /home/student24/max_temperature_reduce.py, /tmp/hadoop-unjar62
20795023687506366/] [] /tmp/streamjob7128140365156923548.jar tmpDir=null
24/05/05 17:12:39 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 17:12:40 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 17:12:42 INFO mapred.FileInputFormat: Total input files to process : 80
24/05/05 17:12:43 INFO mapreduce.JobSubmitter: number of splits:80
24/05/05 17:12:43 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
 Instead, use yarn.system-metrics-publisher.enabled
24/05/05 17:12:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1714512003524_1227
24/05/05 17:12:44 INFO impl.YarnClientImpl: Submitted application application_1714512003524_1227
24/05/05 17:12:44 INFO mapreduce.Job: The url to track the job: http://msba-hadoop-name:8088/proxy/application_1714512
003524_1227/
24/05/05 17:12:44 INFO mapreduce.Job: Running job: job_1714512003524_1227
24/05/05 17:22:34 INFO mapreduce.Job: Job job_1714512003524_1227 running in uber mode : false
24/05/05 17:22:34 INFO mapreduce.Job:  map 0% reduce 0%
24/05/05 17:22:45 INFO mapreduce.Job:  map 3% reduce 0%
24/05/05 17:22:46 INFO mapreduce.Job:  map 8% reduce 0%
24/05/05 17:22:54 INFO mapreduce.Job:  map 13% reduce 0%
24/05/05 17:22:55 INFO mapreduce.Job:  map 15% reduce 0%
24/05/05 17:23:04 INFO mapreduce.Job:  map 16% reduce 0%
24/05/05 17:23:05 INFO mapreduce.Job:  map 20% reduce 0%
24/05/05 17:23:06 INFO mapreduce.Job:  map 22% reduce 0%
24/05/05 17:23:13 INFO mapreduce.Job:  map 24% reduce 0%
```

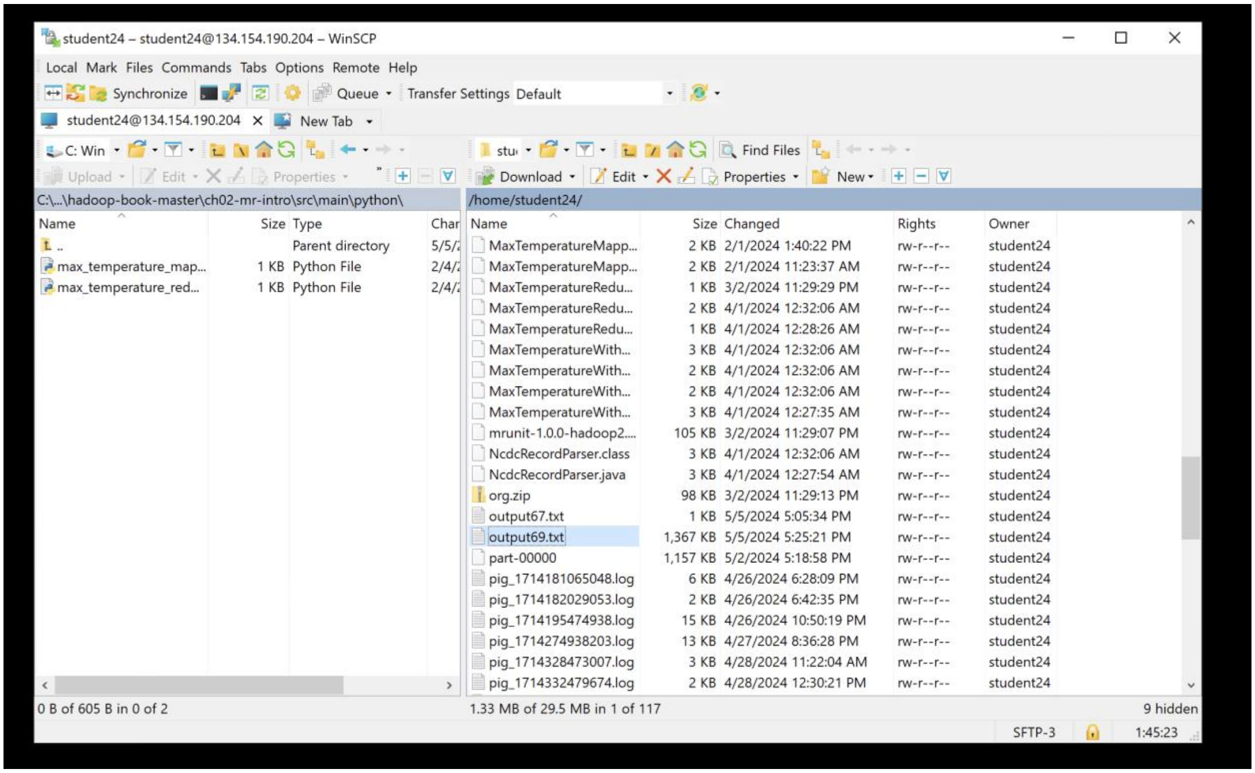


```
24/05/05 17:23:25 INFO mapreduce.Job:  map 35% reduce 12%
24/05/05 17:23:27 INFO mapreduce.Job:  map 36% reduce 12%
24/05/05 17:23:29 INFO mapreduce.Job:  map 39% reduce 12%
24/05/05 17:23:30 INFO mapreduce.Job:  map 41% reduce 12%
24/05/05 17:23:31 INFO mapreduce.Job:  map 41% reduce 14%
24/05/05 17:23:33 INFO mapreduce.Job:  map 43% reduce 14%
24/05/05 17:23:36 INFO mapreduce.Job:  map 45% reduce 14%
24/05/05 17:23:37 INFO mapreduce.Job:  map 47% reduce 15%
24/05/05 17:23:38 INFO mapreduce.Job:  map 49% reduce 15%
24/05/05 17:23:42 INFO mapreduce.Job:  map 50% reduce 15%
24/05/05 17:23:43 INFO mapreduce.Job:  map 51% reduce 17%
24/05/05 17:23:44 INFO mapreduce.Job:  map 54% reduce 17%
24/05/05 17:23:45 INFO mapreduce.Job:  map 55% reduce 17%
24/05/05 17:23:48 INFO mapreduce.Job:  map 56% reduce 17%
24/05/05 17:23:49 INFO mapreduce.Job:  map 57% reduce 19%
24/05/05 17:23:51 INFO mapreduce.Job:  map 60% reduce 19%
24/05/05 17:23:52 INFO mapreduce.Job:  map 61% reduce 19%
24/05/05 17:23:54 INFO mapreduce.Job:  map 63% reduce 19%
24/05/05 17:23:55 INFO mapreduce.Job:  map 63% reduce 21%
24/05/05 17:23:58 INFO mapreduce.Job:  map 64% reduce 21%
24/05/05 17:24:00 INFO mapreduce.Job:  map 66% reduce 21%
24/05/05 17:24:01 INFO mapreduce.Job:  map 68% reduce 21%
24/05/05 17:24:02 INFO mapreduce.Job:  map 69% reduce 23%
24/05/05 17:24:05 INFO mapreduce.Job:  map 70% reduce 23%
24/05/05 17:24:07 INFO mapreduce.Job:  map 73% reduce 23%
24/05/05 17:24:08 INFO mapreduce.Job:  map 73% reduce 24%
24/05/05 17:24:09 INFO mapreduce.Job:  map 75% reduce 24%
24/05/05 17:24:13 INFO mapreduce.Job:  map 77% reduce 24%
24/05/05 17:24:14 INFO mapreduce.Job:  map 79% reduce 26%
24/05/05 17:24:15 INFO mapreduce.Job:  map 80% reduce 26%
24/05/05 17:24:16 INFO mapreduce.Job:  map 81% reduce 26%
24/05/05 17:24:20 INFO mapreduce.Job:  map 84% reduce 27%
24/05/05 17:24:21 INFO mapreduce.Job:  map 86% reduce 27%
24/05/05 17:24:22 INFO mapreduce.Job:  map 88% reduce 27%
24/05/05 17:24:26 INFO mapreduce.Job:  map 88% reduce 29%
24/05/05 17:24:27 INFO mapreduce.Job:  map 89% reduce 29%
24/05/05 17:24:28 INFO mapreduce.Job:  map 93% reduce 29%
24/05/05 17:24:29 INFO mapreduce.Job:  map 94% reduce 29%
24/05/05 17:24:32 INFO mapreduce.Job:  map 94% reduce 31%
24/05/05 17:24:33 INFO mapreduce.Job:  map 95% reduce 31%
24/05/05 17:24:34 INFO mapreduce.Job:  map 96% reduce 31%
24/05/05 17:24:35 INFO mapreduce.Job:  map 100% reduce 31%
```



```
                Total time spent by all maps in occupied slots (ms)=519685
                Total time spent by all reduces in occupied slots (ms)=90184
                Total time spent by all map tasks (ms)=519685
                Total time spent by all reduce tasks (ms)=90184
                Total vcore-milliseconds taken by all map tasks=519685
                Total vcore-milliseconds taken by all reduce tasks=90184
                Total megabyte-milliseconds taken by all map tasks=532157440
                Total megabyte-milliseconds taken by all reduce tasks=92348416
        Map-Reduce Framework
                Map input records=119897
                Map output records=107629
                Map output bytes=1183919
                Map output materialized bytes=1399657
                Input split bytes=10880
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=1399657
                Reduce input records=107629
                Reduce output records=215258
                Spilled Records=215258
                Shuffled Maps =80
                Failed Shuffles=0
                Merged Map outputs=80
                GC time elapsed (ms)=17812
                CPU time spent (ms)=62180
                Physical memory (bytes) snapshot=28286533632
                Virtual memory (bytes) snapshot=244776955904
                Total committed heap usage (bytes)=25397035008
        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=2185179
        File Output Format Counters
                Bytes Written=1399177
24/05/05 17:24:37 INFO streaming.StreamJob: Output directory: /home/24student24/output69
[student24@msba-hadoop-name ~]$ hadoop fs -cat hdfs://msba-hadoop-name:9000/home/24student24/output69/* > output69.txt
```

Text File being created:



student24 – student24@134.154.190.204 – WinSCP

C:\...\hadoop-book-master\ch02-mr-intro\src\main\python\

Name	Size	Type	Char
..		Parent directory	5/5/2
max_temperature_map...	1 KB	Python File	2/4/2
max_temperature_red...	1 KB	Python File	2/4/2

0 B of 605 B in 0 of 2

/home/student24/

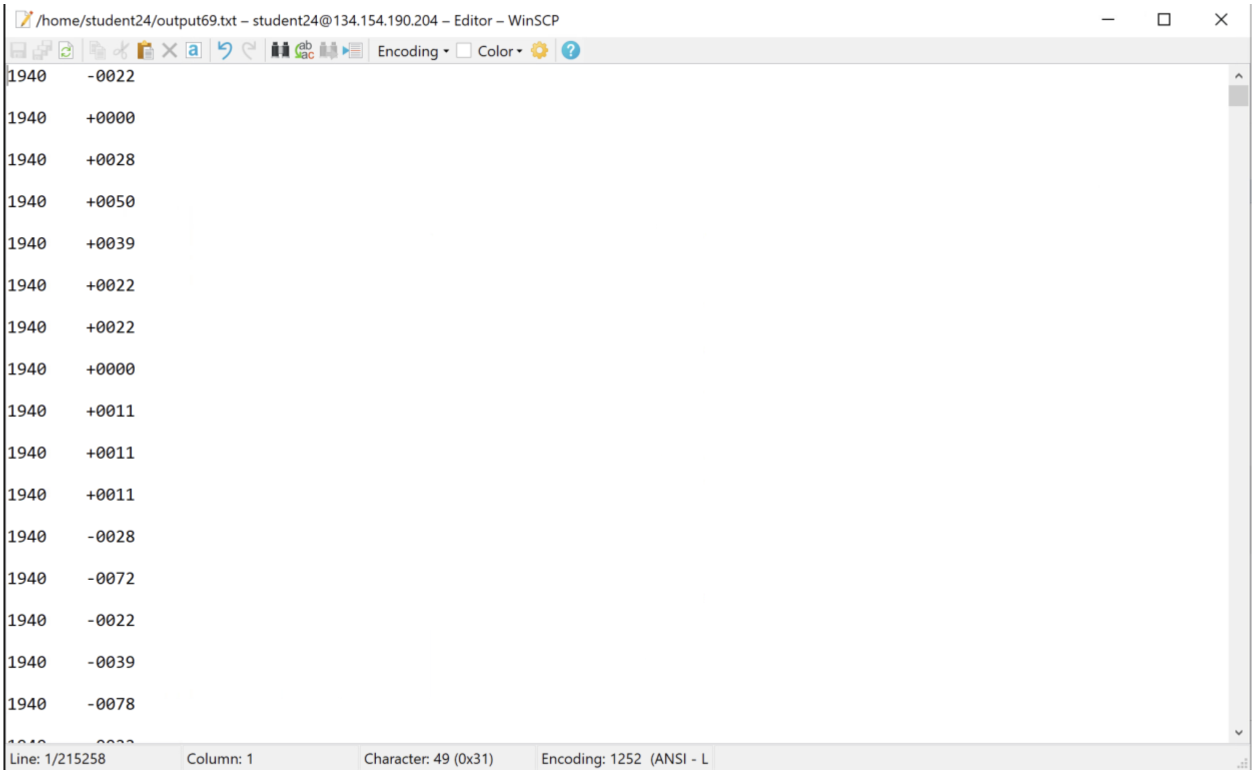
Name	Size	Changed	Rights	Owner
MaxTemperatureMapp...	2 KB	2/1/2024 1:40:22 PM	rw-r--r--	student24
MaxTemperatureMapp...	2 KB	2/1/2024 11:23:37 AM	rw-r--r--	student24
MaxTemperatureRedu...	1 KB	3/2/2024 11:29:29 PM	rw-r--r--	student24
MaxTemperatureRedu...	2 KB	4/1/2024 12:32:06 AM	rw-r--r--	student24
MaxTemperatureRedu...	1 KB	4/1/2024 12:28:26 AM	rw-r--r--	student24
MaxTemperatureWith...	3 KB	4/1/2024 12:32:06 AM	rw-r--r--	student24
MaxTemperatureWith...	2 KB	4/1/2024 12:32:06 AM	rw-r--r--	student24
MaxTemperatureWith...	2 KB	4/1/2024 12:32:06 AM	rw-r--r--	student24
MaxTemperatureWith...	3 KB	4/1/2024 12:27:35 AM	rw-r--r--	student24
mrunit-1.0.0-hadoop2....	105 KB	3/2/2024 11:29:07 PM	rw-r--r--	student24
NcdcRecordParser.class	3 KB	4/1/2024 12:32:06 AM	rw-r--r--	student24
NcdcRecordParser.java	3 KB	4/1/2024 12:27:54 AM	rw-r--r--	student24
org.zip	98 KB	3/2/2024 11:29:13 PM	rw-r--r--	student24
output67.txt	1 KB	5/5/2024 5:05:34 PM	rw-r--r--	student24
output69.txt	1,367 KB	5/5/2024 5:25:21 PM	rw-r--r--	student24
part-00000	1,157 KB	5/2/2024 5:18:58 PM	rw-r--r--	student24
pig_1714181065048.log	6 KB	4/26/2024 6:28:09 PM	rw-r--r--	student24
pig_1714182029053.log	2 KB	4/26/2024 6:42:35 PM	rw-r--r--	student24
pig_1714195474938.log	15 KB	4/26/2024 10:50:19 PM	rw-r--r--	student24
pig_1714274938203.log	13 KB	4/27/2024 8:36:28 PM	rw-r--r--	student24
pig_1714328473007.log	3 KB	4/28/2024 11:22:04 AM	rw-r--r--	student24
pig_1714332479674.log	2 KB	4/28/2024 12:30:21 PM	rw-r--r--	student24

1.33 MB of 29.5 MB in 1 of 117 — 9 hidden

SFTP-3 — 1:45:23

Local Mark Files Commands Tabs Options Remote Help
Synchronize
Queue
Transfer Settings Default
student24@134.154.190.204
New Tab
C: Win
Upload
Edit
Properties
stu
Find Files
Download
New

Text File Data:



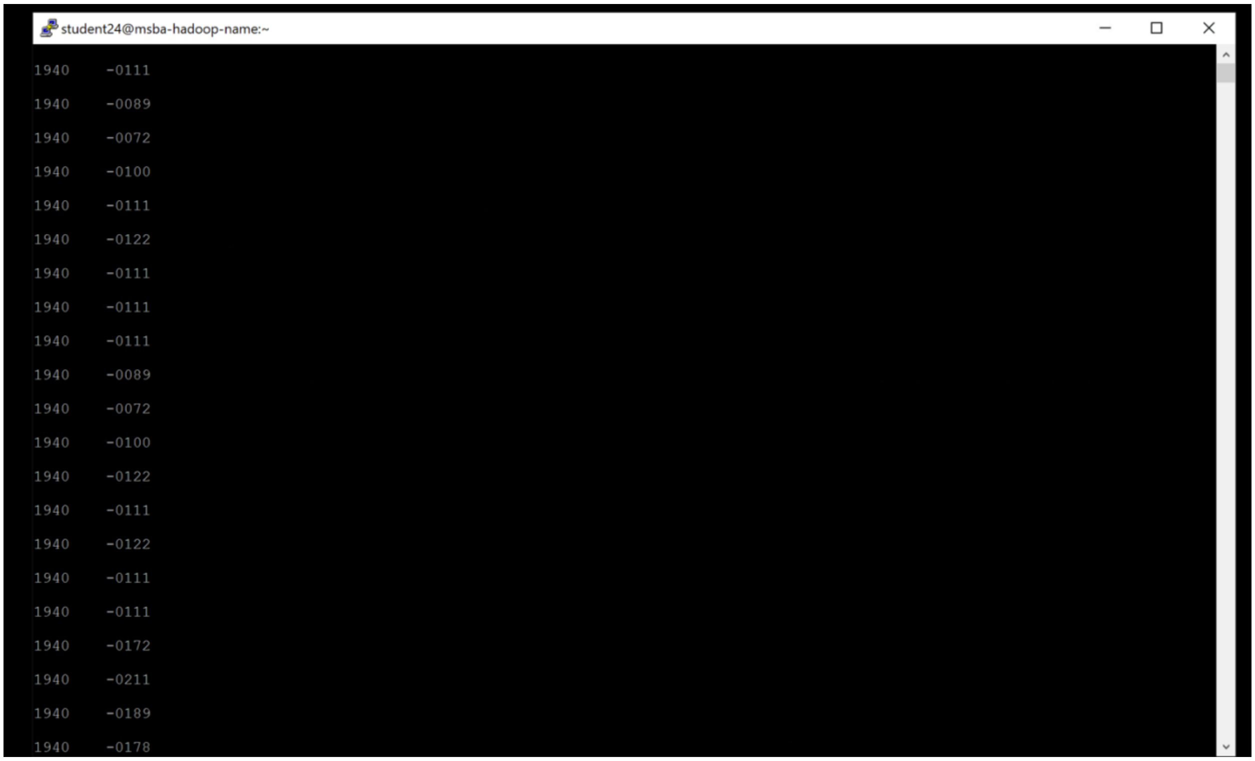
/home/student24/output69.txt – student24@134.154.190.204 – Editor – WinSCP

```
1940	-0022

1940	+0000

1940	+0028

1940	+0050

1940	+0039

1940	+0022

1940	+0022

1940	+0000

1940	+0011

1940	+0011

1940	+0011

1940	-0028

1940	-0072

1940	-0022

1940	-0039

1940	-0078
```

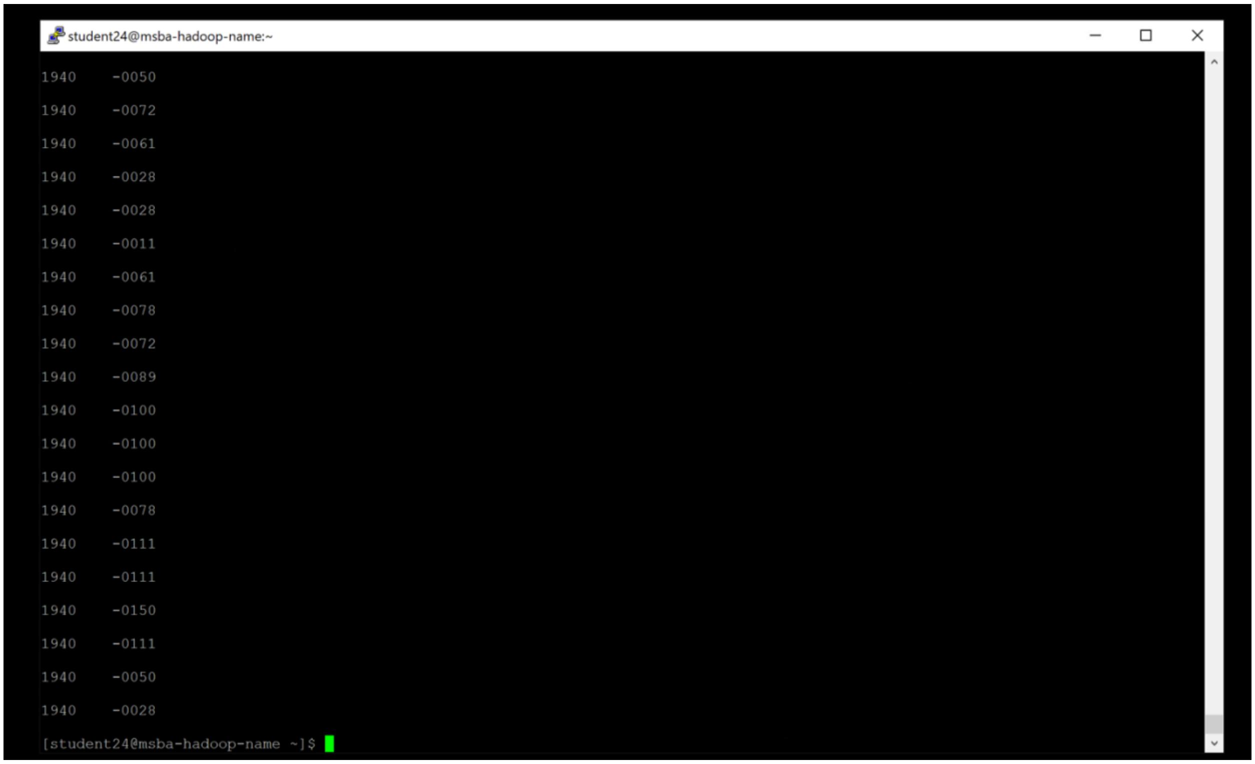
Line: 1/215258 Column: 1 Character: 49 (0x31) Encoding: 1252 (ANSI - L

Printing the text file:

```
hdfs dfs -cat /home/24student24/output69/part-00000
```



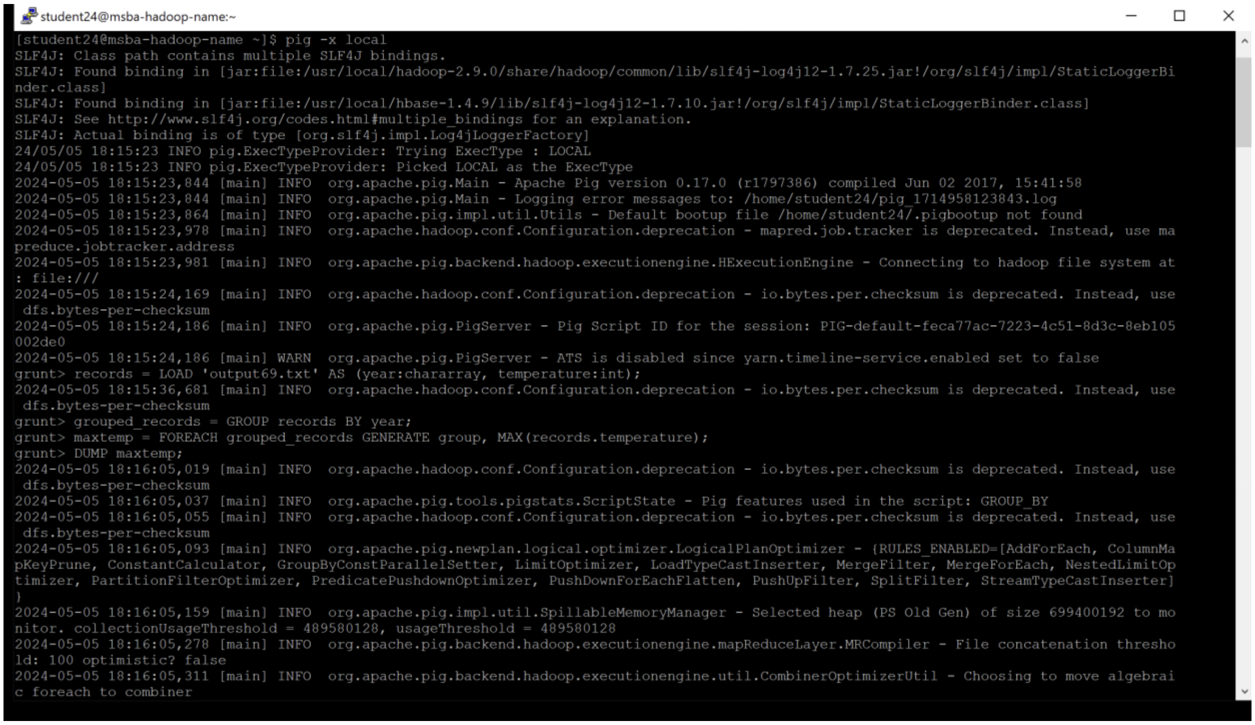
```
1940    -0111
1940    -0089
1940    -0072
1940    -0100
1940    -0111
1940    -0122
1940    -0111
1940    -0111
1940    -0111
1940    -0089
1940    -0072
1940    -0100
1940    -0122
1940    -0111
1940    -0122
1940    -0111
1940    -0111
1940    -0172
1940    -0211
1940    -0189
1940    -0178
```



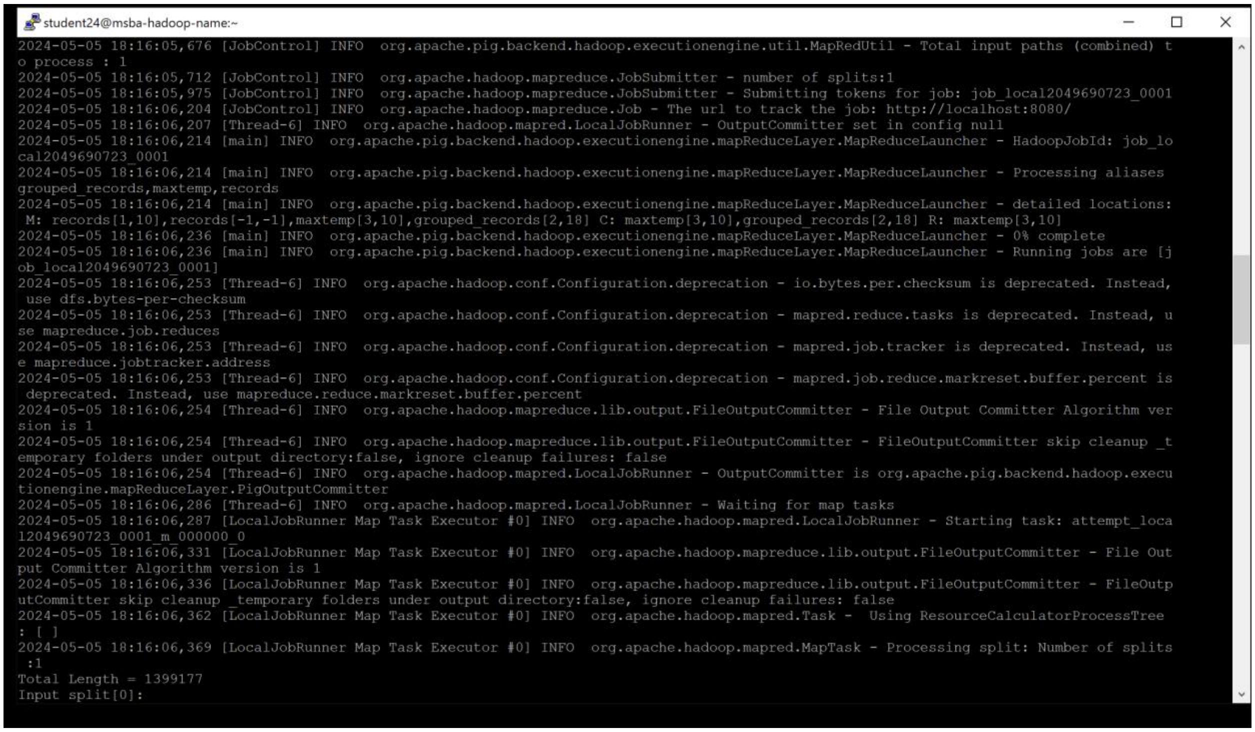
```
1940    -0050
1940    -0072
1940    -0061
1940    -0028
1940    -0028
1940    -0011
1940    -0061
1940    -0078
1940    -0072
1940    -0089
1940    -0100
1940    -0100
1940    -0100
1940    -0078
1940    -0111
1940    -0111
1940    -0150
1940    -0111
1940    -0050
1940    -0028
[student24@msba-hadoop-name ~]$
```

Running in PIG

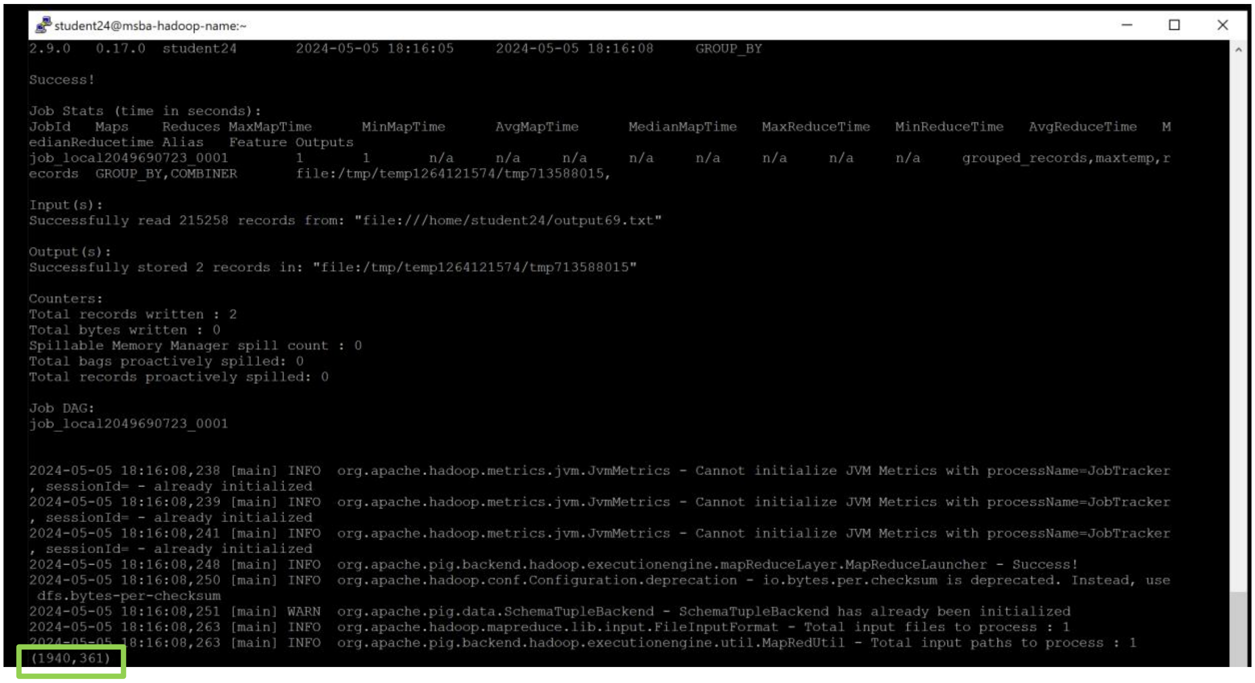
```
pig -x local

records = LOAD 'output69.txt' AS (year:chararray, temperature:int);

grouped_records = GROUP records BY year;

midtemp = FOREACH grouped_records GENERATE group, MAX(records.temperature);

DUMP maxtemp;
```



```
[student24@msba-hadoop-name ~]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
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/StaticLoggerBi
nder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.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.slf4j.impl.Log4jLoggerFactory]
24/05/05 18:15:23 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
24/05/05 18:15:23 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-05-05 18:15:23,844 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-05-05 18:15:23,844 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/student24/pig_1714958123843.log
2024-05-05 18:15:23,864 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/student24/.pigbootup not found
2024-05-05 18:15:23,978 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use ma
preduce.jobtracker.address
2024-05-05 18:15:23,981 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: file:///
2024-05-05 18:15:24,169 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2024-05-05 18:15:24,186 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-feca77ac-7223-4c51-8d3c-8eb105
002de0
2024-05-05 18:15:24,186 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> records = LOAD 'output69.txt' AS (year:chararray, temperature:int);
2024-05-05 18:15:36,681 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
grunt> grouped_records = GROUP records BY year;
grunt> maxtemp = FOREACH grouped_records GENERATE group, MAX(records.temperature);
grunt> DUMP maxtemp;
2024-05-05 18:16:05,019 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2024-05-05 18:16:05,037 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-05-05 18:16:05,055 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2024-05-05 18:16:05,093 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMa
pKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOp
timizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]
}
2024-05-05 18:16:05,159 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to mo
nitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-05-05 18:16:05,278 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresho
ld: 100 optimistic? false
2024-05-05 18:16:05,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebrai
c foreach to combiner
```



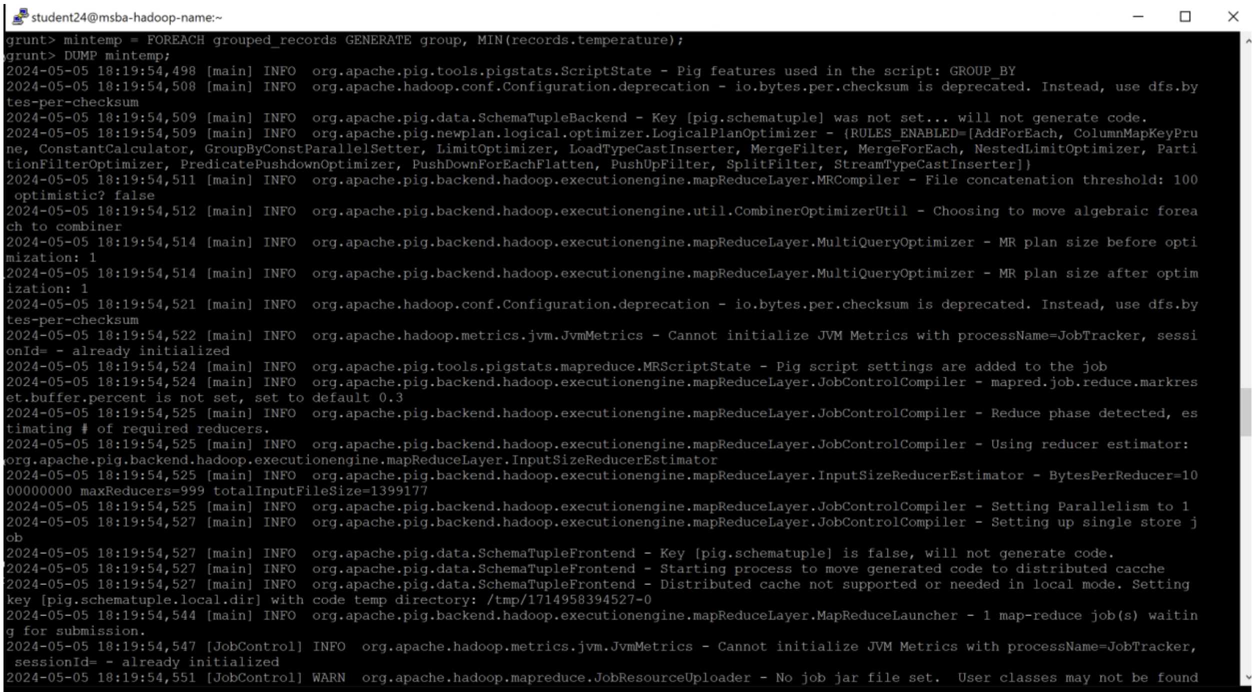
```
2024-05-05 18:16:05,676 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) t
o process : 1
2024-05-05 18:16:05,712 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2024-05-05 18:16:05,975 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local2049690723_0001
2024-05-05 18:16:06,204 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2024-05-05 18:16:06,207 [Thread-6] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2024-05-05 18:16:06,214 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_lo
cal2049690723_0001
2024-05-05 18:16:06,214 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases
grouped_records,maxtemp,records
2024-05-05 18:16:06,214 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations:
 M: records[1,10],records[-1,-1],maxtemp[3,10],grouped_records[2,18] C: maxtemp[3,10],grouped_records[2,18] R: maxtemp[3,10]
2024-05-05 18:16:06,236 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2024-05-05 18:16:06,236 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [j
ob_local2049690723_0001]
2024-05-05 18:16:06,253 [Thread-6] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead,
 use dfs.bytes-per-checksum
2024-05-05 18:16:06,253 [Thread-6] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, u
se mapreduce.job.reduces
2024-05-05 18:16:06,253 [Thread-6] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, us
e mapreduce.jobtracker.address
2024-05-05 18:16:06,253 [Thread-6] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is
 deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2024-05-05 18:16:06,254 [Thread-6] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm ver
sion is 1
2024-05-05 18:16:06,254 [Thread-6] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _t
emporary folders under output directory:false, ignore cleanup failures: false
2024-05-05 18:16:06,254 [Thread-6] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter is org.apache.pig.backend.hadoop.execu
tionengine.mapReduceLayer.PigOutputCommitter
2024-05-05 18:16:06,286 [Thread-6] INFO  org.apache.hadoop.mapred.LocalJobRunner - Waiting for map tasks
2024-05-05 18:16:06,287 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Starting task: attempt_loca
l2049690723_0001_m_000000_0
2024-05-05 18:16:06,331 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Out
put Committer Algorithm version is 1
2024-05-05 18:16:06,336 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutp
utCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-05-05 18:16:06,362 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task -  Using ResourceCalculatorProcessTree
: [ ]
2024-05-05 18:16:06,369 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.MapTask - Processing split: Number of splits
 :1
Total Length = 1399177
Input split[0]:
```



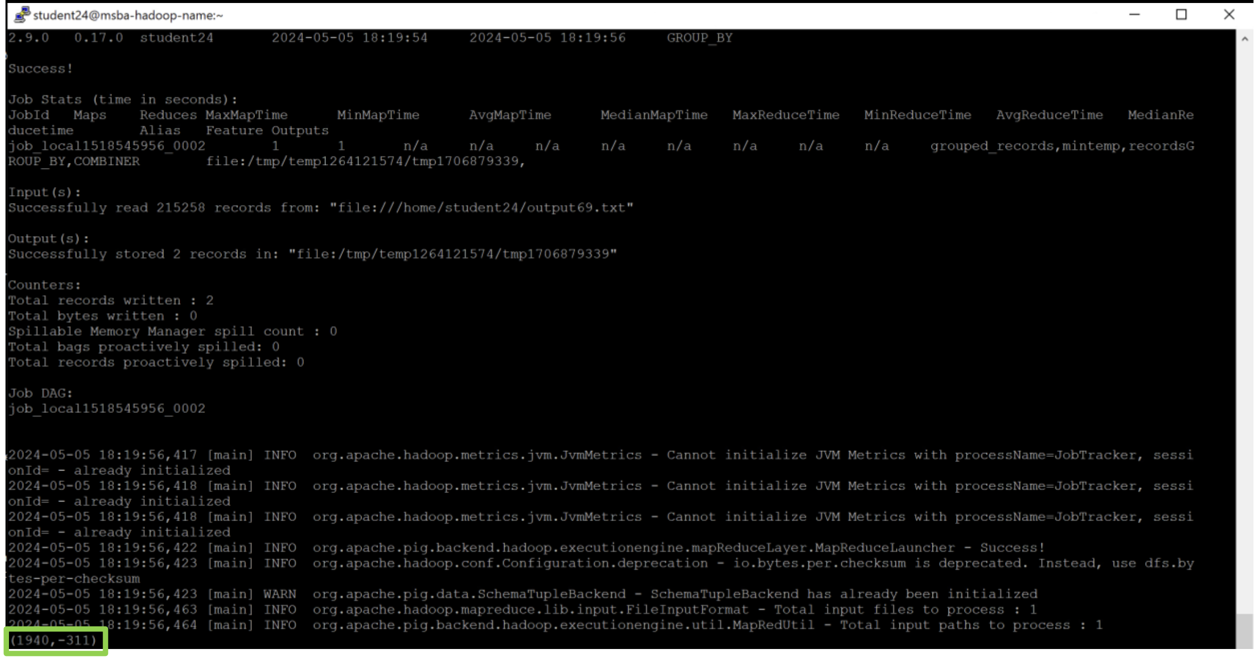
```
2.9.0   0.17.0  student24       2024-05-05 18:16:05     2024-05-05 18:16:08     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   M
edianReducetime Alias   Feature Outputs
job_local2049690723_0001        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     grouped_records,maxtemp,r
ecords  GROUP_BY,COMBINER       file:/tmp/temp1264121574/tmp713588015,

Input(s):
Successfully read 215258 records from: "file:///home/student24/output69.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp1264121574/tmp713588015"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2049690723_0001


2024-05-05 18:16:08,238 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2024-05-05 18:16:08,239 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2024-05-05 18:16:08,241 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2024-05-05 18:16:08,248 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-05-05 18:16:08,250 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2024-05-05 18:16:08,251 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-05-05 18:16:08,263 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-05-05 18:16:08,263 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1940,361)
```

mintemp = FOREACH grouped_records GENERATE group, MIN(records.temperature);

DUMP mintemp;



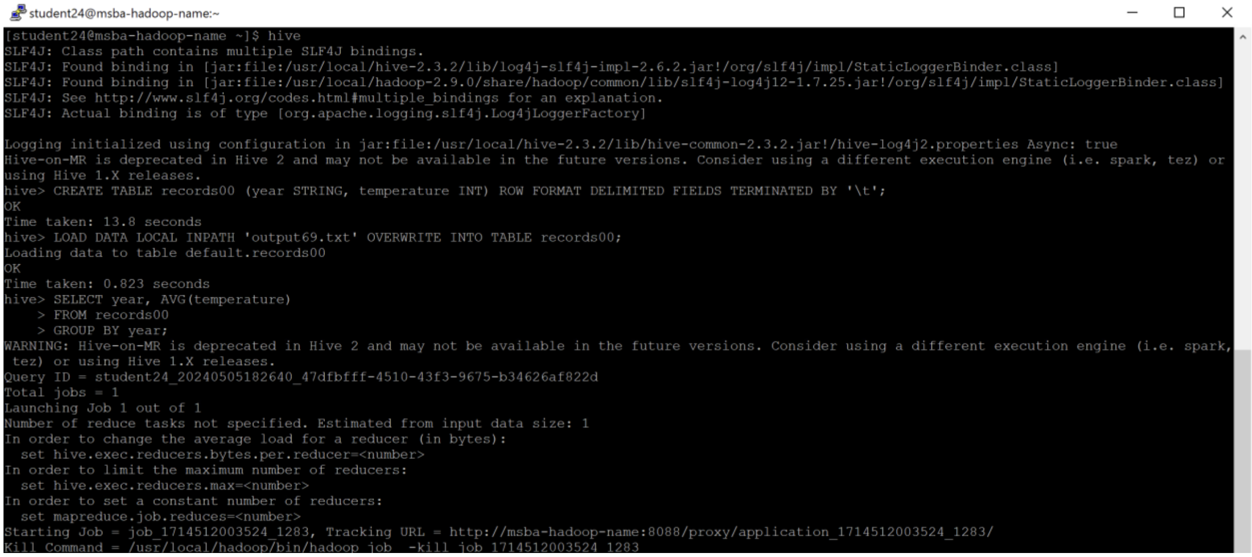
```
grunt> mintemp = FOREACH grouped_records GENERATE group, MIN(records.temperature);
grunt> DUMP mintemp;
2024-05-05 18:19:54,498 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-05-05 18:19:54,508 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.by
tes-per-checksum
2024-05-05 18:19:54,509 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-05-05 18:19:54,509 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPru
ne, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, Parti
tionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-05-05 18:19:54,511 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100
 optimistic? false
2024-05-05 18:19:54,512 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic forea
ch to combiner
2024-05-05 18:19:54,514 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before opti
mization: 1
2024-05-05 18:19:54,514 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optim
ization: 1
2024-05-05 18:19:54,521 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.by
tes-per-checksum
2024-05-05 18:19:54,522 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessi
onId= - already initialized
2024-05-05 18:19:54,524 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2024-05-05 18:19:54,524 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markres
et.buffer.percent is not set, set to default 0.3
2024-05-05 18:19:54,525 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, es
timating # of required reducers.
2024-05-05 18:19:54,525 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-05-05 18:19:54,525 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=10
00000000 maxReducers=999 totalInputFileSize=1399177
2024-05-05 18:19:54,525 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-05-05 18:19:54,527 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store j
ob
2024-05-05 18:19:54,527 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-05-05 18:19:54,527 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2024-05-05 18:19:54,527 [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 directory: /tmp/1714958394527-0
2024-05-05 18:19:54,544 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waitin
g for submission.
2024-05-05 18:19:54,547 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker,
 sessionId= - already initialized
2024-05-05 18:19:54,551 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found
```



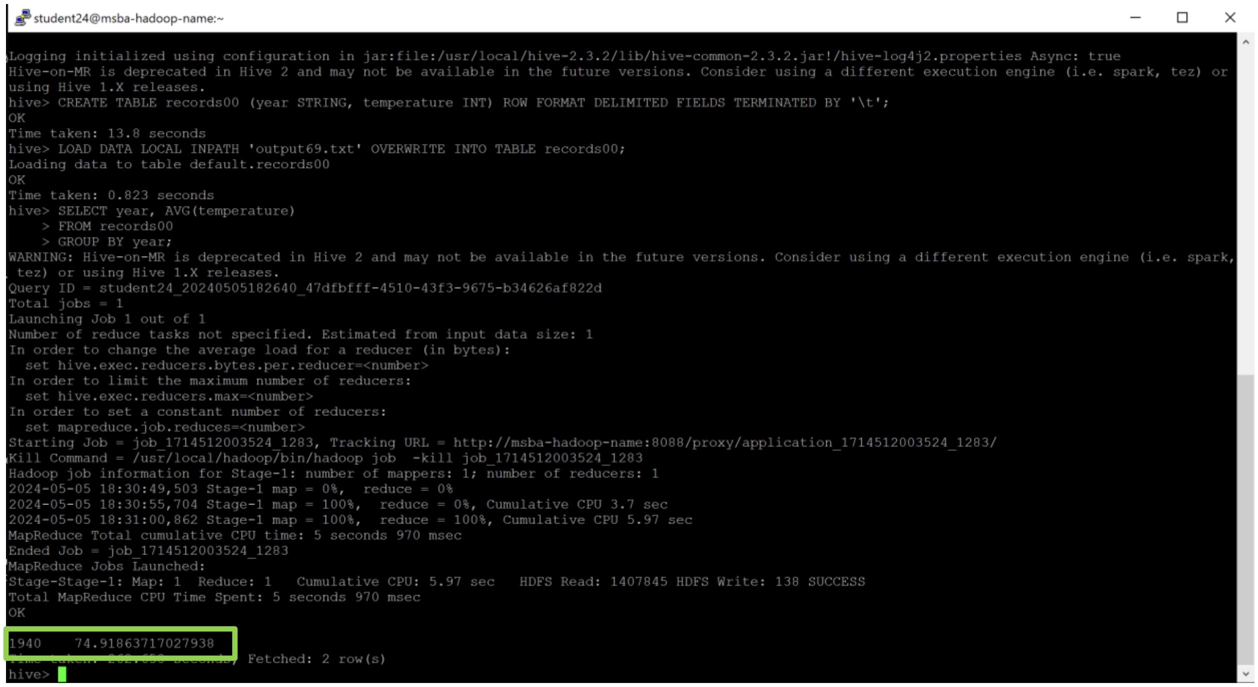
```
2.9.0   0.17.0  student24       2024-05-05 18:19:54     2024-05-05 18:19:56     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianRe
ducetime        Alias   Feature Outputs
job_local1518545956_0002        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     grouped_records,mintemp,recordsG
ROUP_BY,COMBINER        file:/tmp/temp1264121574/tmp1706879339,

Input(s):
Successfully read 215258 records from: "file:///home/student24/output69.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp1264121574/tmp1706879339"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1518545956_0002


2024-05-05 18:19:56,417 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessi
onId= - already initialized
2024-05-05 18:19:56,418 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessi
onId= - already initialized
2024-05-05 18:19:56,418 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessi
onId= - already initialized
2024-05-05 18:19:56,422 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-05-05 18:19:56,423 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.by
tes-per-checksum
2024-05-05 18:19:56,423 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-05-05 18:19:56,463 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-05-05 18:19:56,464 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1940,-311)
```

Running in Hive:

```
hive

CREATE TABLE records00 (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

LOAD DATA LOCAL INPATH 'output69.txt' OVERWRITE INTO TABLE records00;

SELECT year, AVG(temperature) FROM records00 GROUP BY year;
```



```
[student24@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> CREATE TABLE records00 (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 13.8 seconds
hive> LOAD DATA LOCAL INPATH 'output69.txt' OVERWRITE INTO TABLE records00;
Loading data to table default.records00
OK
Time taken: 0.823 seconds
hive> SELECT year, AVG(temperature)
    > FROM records00
    > GROUP BY year;
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 = student24_20240505182640_47dfbfff-4510-43f3-9675-b34626af822d
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_1714512003524_1283, Tracking URL = http://msba-hadoop-name:8088/proxy/application_1714512003524_1283/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1714512003524_1283
```



```
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> CREATE TABLE records00 (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 13.8 seconds
hive> LOAD DATA LOCAL INPATH 'output69.txt' OVERWRITE INTO TABLE records00;
Loading data to table default.records00
OK
Time taken: 0.823 seconds
hive> SELECT year, AVG(temperature)
    > FROM records00
    > GROUP BY year;
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 = student24_20240505182640_47dfbfff-4510-43f3-9675-b34626af822d
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_1714512003524_1283, Tracking URL = http://msba-hadoop-name:8088/proxy/application_1714512003524_1283/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1714512003524_1283
```

```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-05 18:30:49,503 Stage-1 map = 0%,  reduce = 0%
2024-05-05 18:30:55,704 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.7 sec
2024-05-05 18:31:00,862 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.97 sec
MapReduce Total cumulative CPU time: 5 seconds 970 msec
Ended Job = job_1714512003524_1283
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.97 sec   HDFS Read: 1407845 HDFS Write: 138 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 970 msec
OK

1940    74.91863717027938
Time taken: 262.658 seconds, Fetched: 2 row(s)
hive>
```