The Course Project- Jasvitha Buggana – student59- js3225

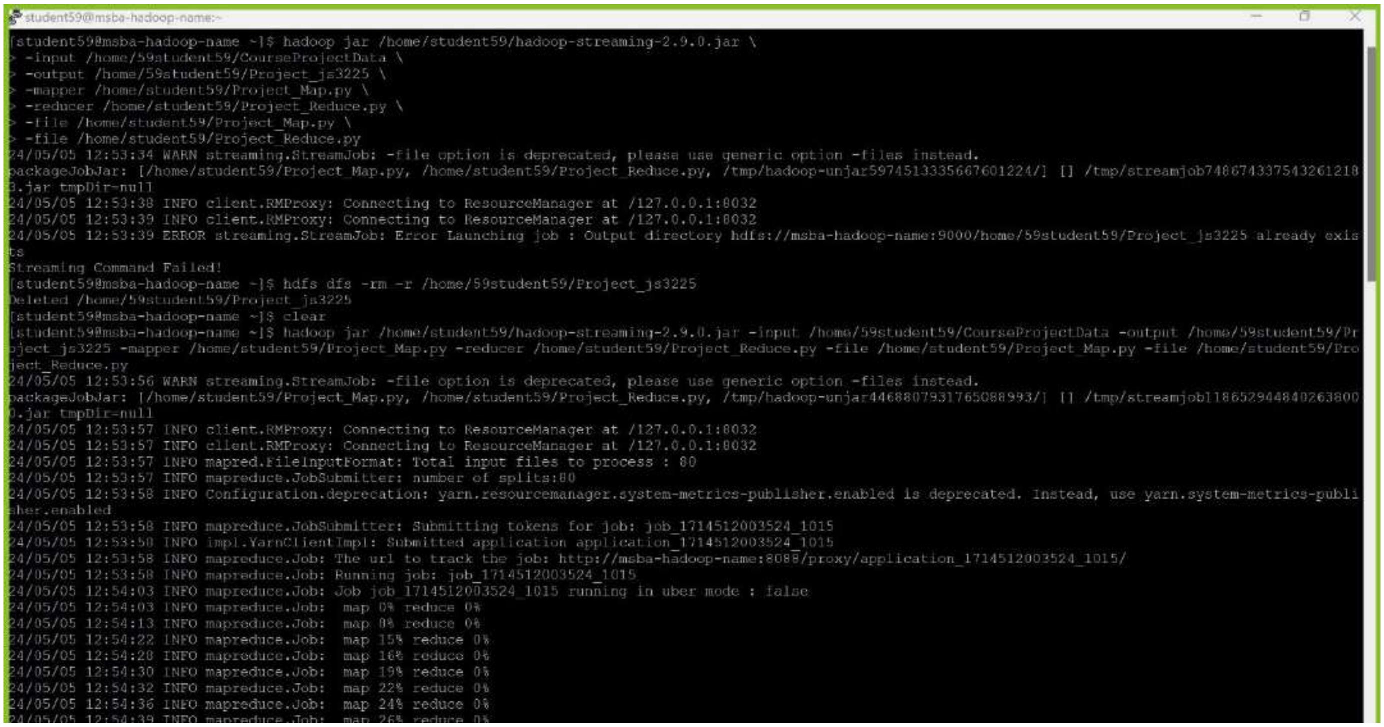
The course project includes 3 parts. The first part is to develop a Python application to retrieve Year and Temperature from original NCDC records (i.e., the dataset we are using for this class) and then write the Year and Temperature data into a text file. The second part is to load the text file into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

Project Steps and Outputs:

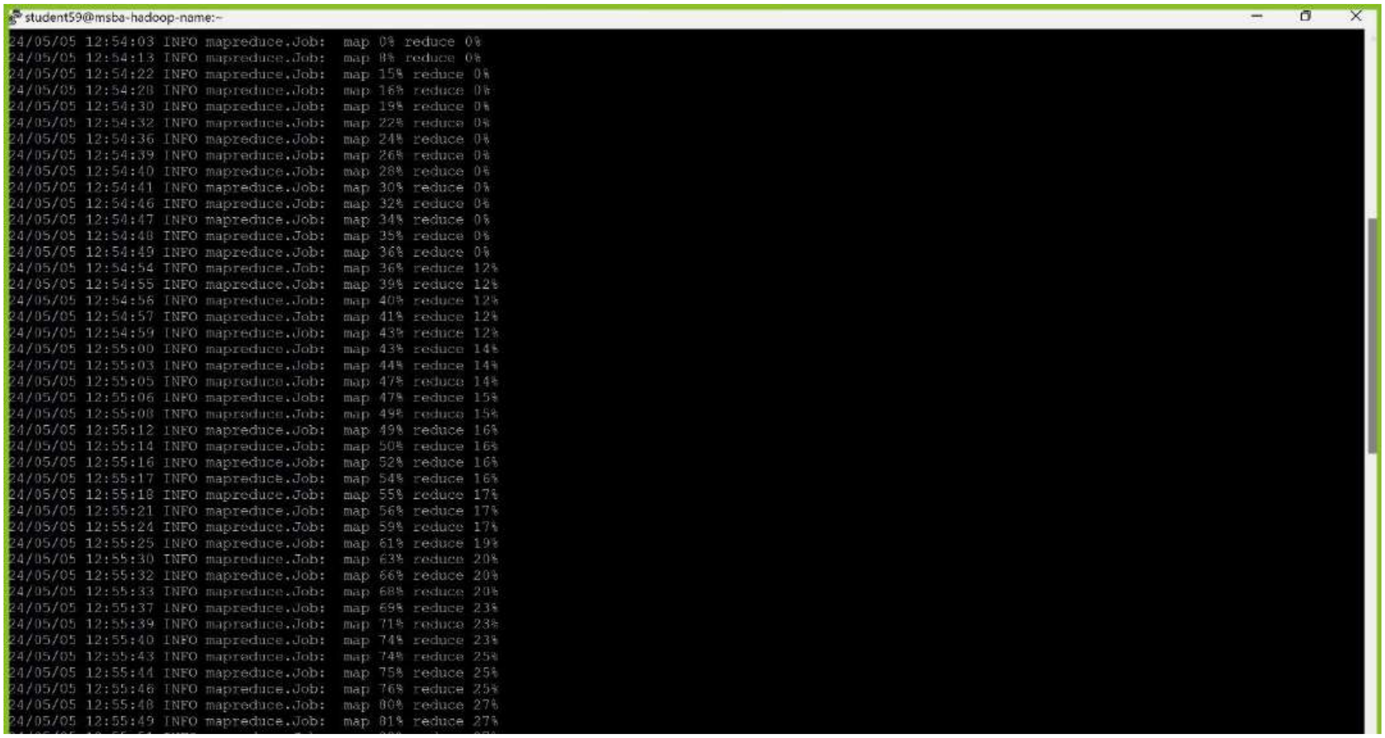
Created two Python files: Project_Map.py and Project_Reduce.py

Download CourseProjectData.zip and transfer the files using winscp.

- hdfs dfs -copyFromLocal /home/student59/CourseProjectData /home/59student59/CourseProjectData
- hadoop jar /home/student59/hadoop-streaming-2.9.0.jar \input /home/59student59/CourseProjectData \output /home/59student59/Project_js3225 \mapper /home/student59/Project_map.py \reducer /home/student59/Project_Reduce.py \file /home/student59/Project_map.py \file /home/student59/Project_Reduce.py

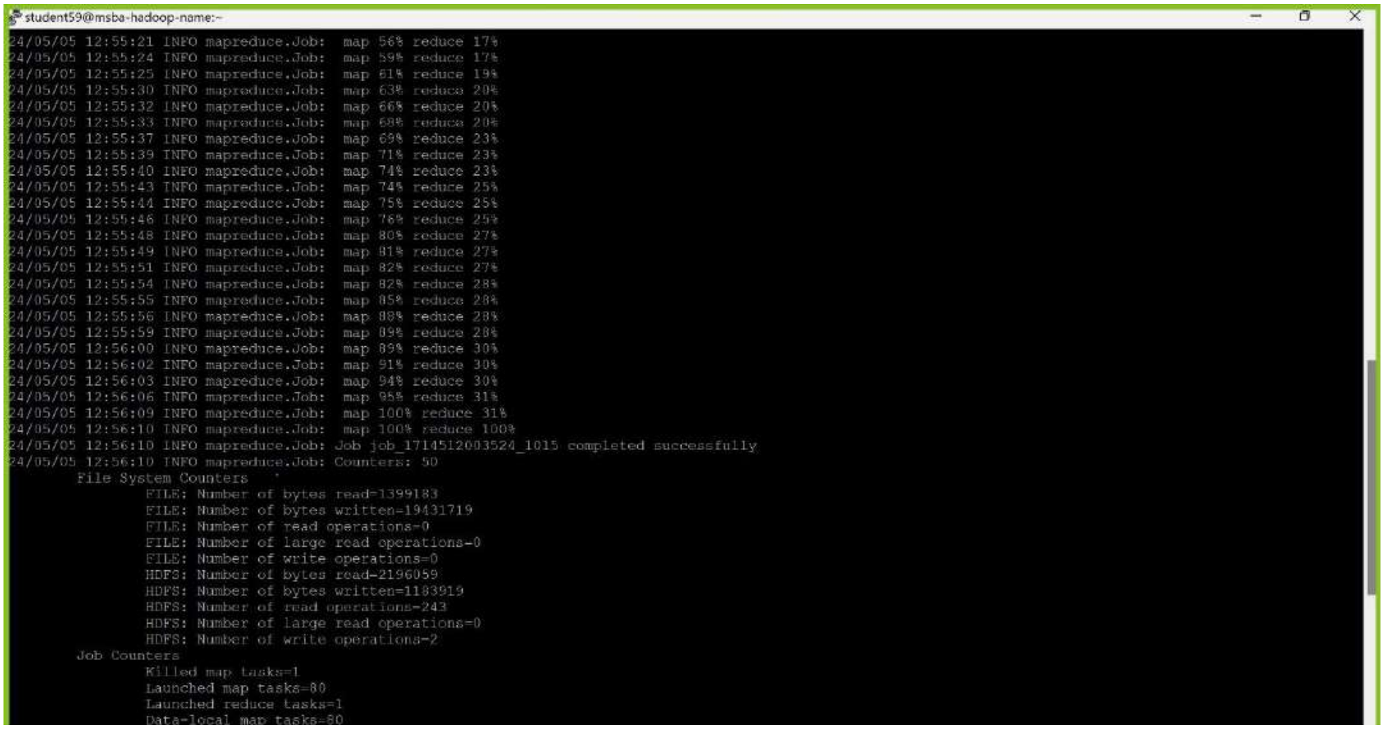


```
[student59@msba-hadoop-name ~]$ hadoop jar /home/student59/hadoop-streaming-2.9.0.jar \
> -input /home/59student59/CourseProjectData \
> -output /home/59student59/Project_js3225 \
> -mapper /home/student59/Project_Map.py \
> -reducer /home/student59/Project_Reduce.py \
> -file /home/student59/Project_Map.py \
> -file /home/student59/Project_Reduce.py
24/05/05 12:53:34 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/student59/Project_Map.py, /home/student59/Project_Reduce.py, /tmp/hadoop-unjar5974513335667601224/] [] /tmp/streamjob748674337543261218
3.jar tmpDir=null
24/05/05 12:53:38 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 12:53:39 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 12:53:39 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://msba-hadoop-name:9000/home/59student59/Project_js3225 already exis
ts
Streaming Command Failed!
[student59@msba-hadoop-name ~]$ hdfs dfs -rm -r /home/59student59/Project_js3225
Deleted /home/59student59/Project_js3225
[student59@msba-hadoop-name ~]$ clear
[student59@msba-hadoop-name ~]$ hadoop jar /home/student59/hadoop-streaming-2.9.0.jar -input /home/59student59/CourseProjectData -output /home/59student59/Pr
oject_js3225 -mapper /home/student59/Project_Map.py -reducer /home/student59/Project_Reduce.py -file /home/student59/Project_Map.py -file /home/student59/Pro
ject_Reduce.py
24/05/05 12:53:56 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/student59/Project_Map.py, /home/student59/Project_Reduce.py, /tmp/hadoop-unjar4468807931765088993/] [] /tmp/streamjob118652944840263800
0.jar tmpDir=null
24/05/05 12:53:57 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 12:53:57 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/05/05 12:53:57 INFO mapred.FileInputFormat: Total input files to process : 80
24/05/05 12:53:57 INFO mapreduce.JobSubmitter: number of splits:80
24/05/05 12:53:58 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publi
sher.enabled
24/05/05 12:53:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1714512003524_1015
24/05/05 12:53:58 INFO impl.YarnClientImpl: Submitted application application_1714512003524_1015
24/05/05 12:53:58 INFO mapreduce.Job: The url to track the job: http://msba-hadoop-name:8088/proxy/application_1714512003524_1015/
24/05/05 12:53:58 INFO mapreduce.Job: Running job: job_1714512003524_1015
24/05/05 12:54:03 INFO mapreduce.Job: Job job_1714512003524_1015 running in uber mode : false
24/05/05 12:54:03 INFO mapreduce.Job:  map 0% reduce 0%
24/05/05 12:54:13 INFO mapreduce.Job:  map 8% reduce 0%
24/05/05 12:54:22 INFO mapreduce.Job:  map 15% reduce 0%
24/05/05 12:54:28 INFO mapreduce.Job:  map 16% reduce 0%
24/05/05 12:54:30 INFO mapreduce.Job:  map 19% reduce 0%
24/05/05 12:54:32 INFO mapreduce.Job:  map 22% reduce 0%
24/05/05 12:54:36 INFO mapreduce.Job:  map 24% reduce 0%
24/05/05 12:54:39 INFO mapreduce.Job:  map 26% reduce 0%
```



student59@msba-hadoop-name:~

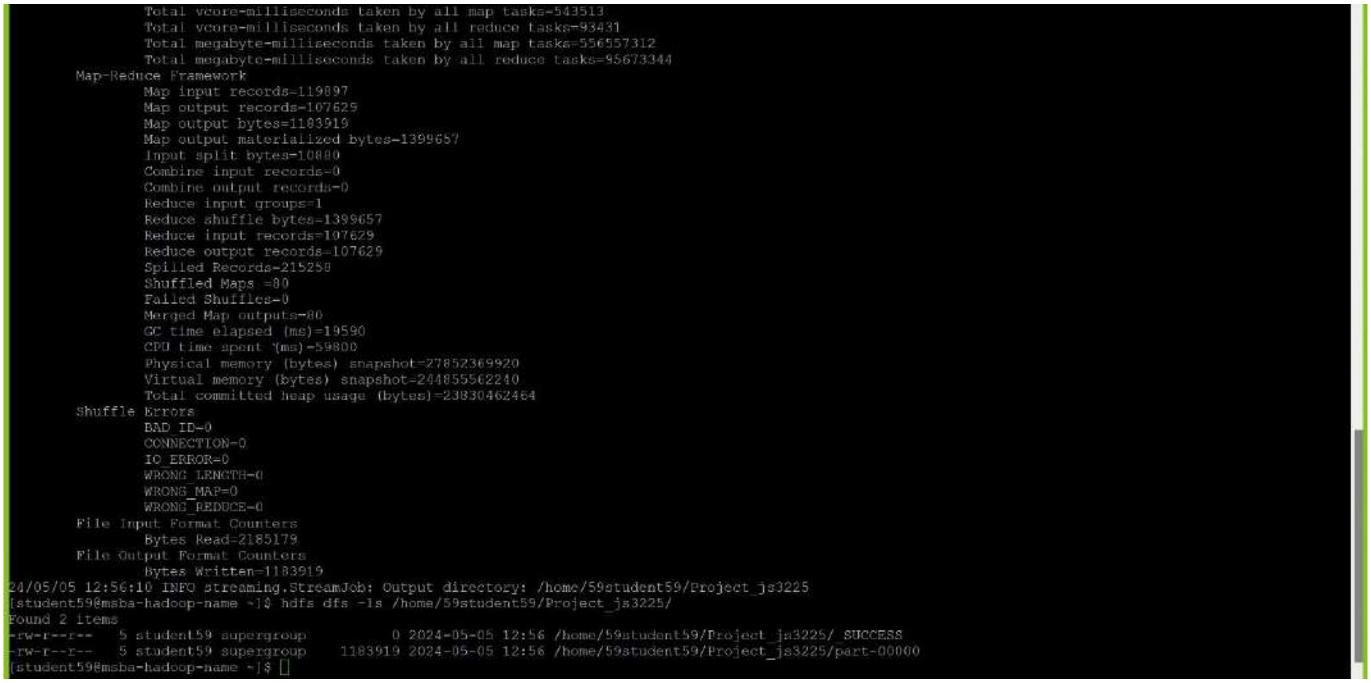
```
24/05/05 12:54:03 INFO mapreduce.Job:  map 0% reduce 0%
24/05/05 12:54:13 INFO mapreduce.Job:  map 8% reduce 0%
24/05/05 12:54:22 INFO mapreduce.Job:  map 15% reduce 0%
24/05/05 12:54:28 INFO mapreduce.Job:  map 16% reduce 0%
24/05/05 12:54:30 INFO mapreduce.Job:  map 19% reduce 0%
24/05/05 12:54:32 INFO mapreduce.Job:  map 22% reduce 0%
24/05/05 12:54:36 INFO mapreduce.Job:  map 24% reduce 0%
24/05/05 12:54:39 INFO mapreduce.Job:  map 26% reduce 0%
24/05/05 12:54:40 INFO mapreduce.Job:  map 28% reduce 0%
24/05/05 12:54:41 INFO mapreduce.Job:  map 30% reduce 0%
24/05/05 12:54:46 INFO mapreduce.Job:  map 32% reduce 0%
24/05/05 12:54:47 INFO mapreduce.Job:  map 34% reduce 0%
24/05/05 12:54:48 INFO mapreduce.Job:  map 35% reduce 0%
24/05/05 12:54:49 INFO mapreduce.Job:  map 36% reduce 0%
24/05/05 12:54:54 INFO mapreduce.Job:  map 36% reduce 12%
24/05/05 12:54:55 INFO mapreduce.Job:  map 39% reduce 12%
24/05/05 12:54:56 INFO mapreduce.Job:  map 40% reduce 12%
24/05/05 12:54:57 INFO mapreduce.Job:  map 41% reduce 12%
24/05/05 12:54:59 INFO mapreduce.Job:  map 43% reduce 12%
24/05/05 12:55:00 INFO mapreduce.Job:  map 43% reduce 14%
24/05/05 12:55:03 INFO mapreduce.Job:  map 44% reduce 14%
24/05/05 12:55:05 INFO mapreduce.Job:  map 47% reduce 14%
24/05/05 12:55:06 INFO mapreduce.Job:  map 47% reduce 15%
24/05/05 12:55:08 INFO mapreduce.Job:  map 49% reduce 15%
24/05/05 12:55:12 INFO mapreduce.Job:  map 49% reduce 16%
24/05/05 12:55:14 INFO mapreduce.Job:  map 50% reduce 16%
24/05/05 12:55:16 INFO mapreduce.Job:  map 52% reduce 16%
24/05/05 12:55:17 INFO mapreduce.Job:  map 54% reduce 16%
24/05/05 12:55:18 INFO mapreduce.Job:  map 55% reduce 17%
24/05/05 12:55:21 INFO mapreduce.Job:  map 56% reduce 17%
24/05/05 12:55:24 INFO mapreduce.Job:  map 59% reduce 17%
24/05/05 12:55:25 INFO mapreduce.Job:  map 61% reduce 19%
24/05/05 12:55:30 INFO mapreduce.Job:  map 63% reduce 20%
24/05/05 12:55:32 INFO mapreduce.Job:  map 66% reduce 20%
24/05/05 12:55:33 INFO mapreduce.Job:  map 68% reduce 20%
24/05/05 12:55:37 INFO mapreduce.Job:  map 69% reduce 23%
24/05/05 12:55:39 INFO mapreduce.Job:  map 71% reduce 23%
24/05/05 12:55:40 INFO mapreduce.Job:  map 74% reduce 23%
24/05/05 12:55:43 INFO mapreduce.Job:  map 74% reduce 25%
24/05/05 12:55:44 INFO mapreduce.Job:  map 75% reduce 25%
24/05/05 12:55:46 INFO mapreduce.Job:  map 76% reduce 25%
24/05/05 12:55:48 INFO mapreduce.Job:  map 80% reduce 27%
24/05/05 12:55:49 INFO mapreduce.Job:  map 81% reduce 27%
```



student59@msba-hadoop-name:~

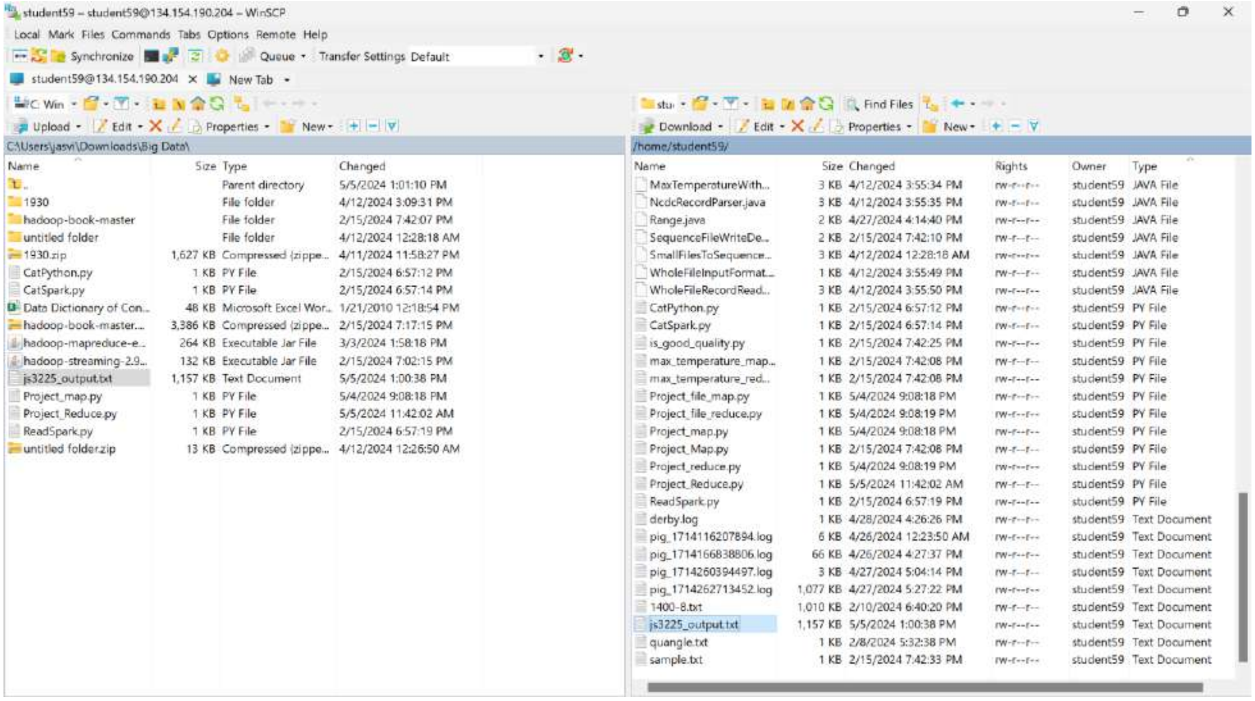
```
24/05/05 12:55:21 INFO mapreduce.Job:  map 56% reduce 17%
24/05/05 12:55:24 INFO mapreduce.Job:  map 59% reduce 17%
24/05/05 12:55:25 INFO mapreduce.Job:  map 61% reduce 19%
24/05/05 12:55:30 INFO mapreduce.Job:  map 63% reduce 20%
24/05/05 12:55:32 INFO mapreduce.Job:  map 66% reduce 20%
24/05/05 12:55:33 INFO mapreduce.Job:  map 68% reduce 20%
24/05/05 12:55:37 INFO mapreduce.Job:  map 69% reduce 23%
24/05/05 12:55:39 INFO mapreduce.Job:  map 71% reduce 23%
24/05/05 12:55:40 INFO mapreduce.Job:  map 74% reduce 23%
24/05/05 12:55:43 INFO mapreduce.Job:  map 74% reduce 25%
24/05/05 12:55:44 INFO mapreduce.Job:  map 75% reduce 25%
24/05/05 12:55:46 INFO mapreduce.Job:  map 76% reduce 25%
24/05/05 12:55:48 INFO mapreduce.Job:  map 80% reduce 27%
24/05/05 12:55:49 INFO mapreduce.Job:  map 81% reduce 27%
24/05/05 12:55:51 INFO mapreduce.Job:  map 82% reduce 27%
24/05/05 12:55:54 INFO mapreduce.Job:  map 82% reduce 28%
24/05/05 12:55:55 INFO mapreduce.Job:  map 85% reduce 28%
24/05/05 12:55:56 INFO mapreduce.Job:  map 88% reduce 28%
24/05/05 12:55:59 INFO mapreduce.Job:  map 89% reduce 28%
24/05/05 12:56:00 INFO mapreduce.Job:  map 89% reduce 30%
24/05/05 12:56:02 INFO mapreduce.Job:  map 91% reduce 30%
24/05/05 12:56:03 INFO mapreduce.Job:  map 94% reduce 30%
24/05/05 12:56:06 INFO mapreduce.Job:  map 95% reduce 31%
24/05/05 12:56:09 INFO mapreduce.Job:  map 100% reduce 31%
24/05/05 12:56:10 INFO mapreduce.Job:  map 100% reduce 100%
24/05/05 12:56:10 INFO mapreduce.Job: Job job_1714512003524_1015 completed successfully
24/05/05 12:56:10 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=1399183
                FILE: Number of bytes written=19431719
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2196059
                HDFS: Number of bytes written=1183919
                HDFS: Number of read operations=243
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=80
                Launched reduce tasks=1
                Data-local map tasks=80
```

- hdfs dfs -ls /home/59student59/Project_js3225/



```
                Total vcore-milliseconds taken by all map tasks=543513
                Total vcore-milliseconds taken by all reduce tasks=93431
                Total megabyte-milliseconds taken by all map tasks=556557312
                Total megabyte-milliseconds taken by all reduce tasks=95673344
        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=107629
                Spilled Records=215258
                Shuffled Maps =80
                Failed Shuffles=0
                Merged Map outputs=80
                GC time elapsed (ms)=19590
                CPU time spent (ms)=59800
                Physical memory (bytes) snapshot=27852369920
                Virtual memory (bytes) snapshot=244855562240
                Total committed heap usage (bytes)=23830462464
        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=1183919
24/05/05 12:56:10 INFO streaming.StreamJob: Output directory: /home/59student59/Project_js3225
[student59@msba-hadoop-name ~]$ hdfs dfs -ls /home/59student59/Project_js3225/
Found 2 items
-rw-r--r--   5 student59 supergroup          0 2024-05-05 12:56 /home/59student59/Project_js3225/_SUCCESS
-rw-r--r--   5 student59 supergroup    1183919 2024-05-05 12:56 /home/59student59/Project_js3225/part-00000
[student59@msba-hadoop-name ~]$ 
```

- hdfs dfs -copyToLocal /home/59student59/Project_js3225/part-00000 /home/student59/js3225_output.txt
- js_3225_output file will be generated.


student59 – student59@134.154.190.204 – WinSCP
Local Mark Files Commands Tabs Options Remote Help
Synchronize
Queue
Transfer Settings Default
student59@134.154.190.204
New Tab
C: Win
Upload
Edit
Properties
New
stu
Find Files
Download

C:\Users\jasvi\Downloads\Big Data\

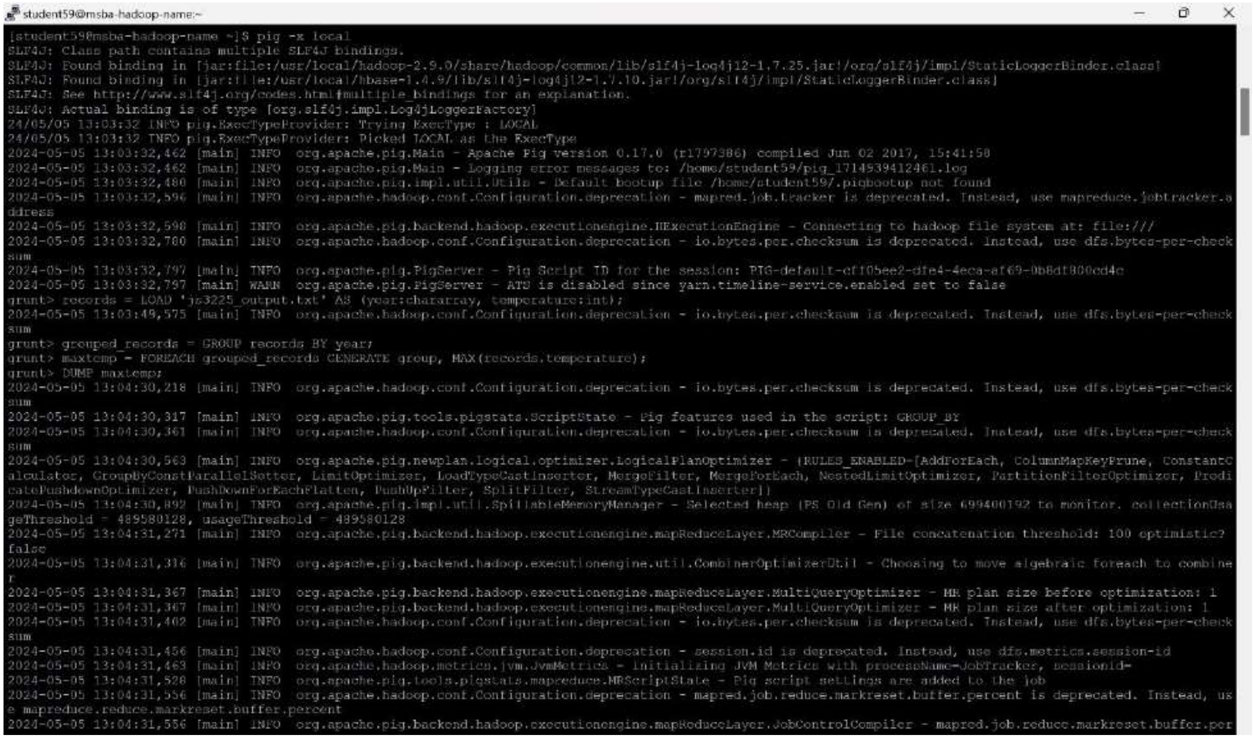
Name	Size	Type	Changed
..		Parent directory	5/5/2024 1:01:10 PM
1930		File folder	4/12/2024 3:09:31 PM
hadoop-book-master		File folder	2/15/2024 7:42:07 PM
untitled folder		File folder	4/12/2024 12:28:18 AM
1930.zip	1,627 KB	Compressed (zippe...	4/11/2024 11:58:27 PM
CatPython.py	1 KB	PY File	2/15/2024 6:57:12 PM
CatSpark.py	1 KB	PY File	2/15/2024 6:57:14 PM
Data Dictionary of Con...	48 KB	Microsoft Excel Wor...	1/21/2010 12:18:54 PM
hadoop-book-master....	3,386 KB	Compressed (zippe...	2/15/2024 7:17:15 PM
hadoop-mapreduce-e...	264 KB	Executable Jar File	3/3/2024 1:58:18 PM
hadoop-streaming-2.9...	132 KB	Executable Jar File	2/15/2024 7:02:15 PM
js3225_output.txt	1,157 KB	Text Document	5/5/2024 1:00:38 PM
Project_map.py	1 KB	PY File	5/4/2024 9:08:18 PM
Project_Reduce.py	1 KB	PY File	5/5/2024 11:42:02 AM
ReadSpark.py	1 KB	PY File	2/15/2024 6:57:19 PM
untitled folder.zip	13 KB	Compressed (zippe...	4/12/2024 12:26:50 AM

/home/student59/

Name	Size	Changed	Rights	Owner	Type
MaxTemperatureWith...	3 KB	4/12/2024 3:55:34 PM	rw-r--r--	student59	JAVA File
NcdcRecordParser.java	3 KB	4/12/2024 3:55:35 PM	rw-r--r--	student59	JAVA File
Range.java	2 KB	4/27/2024 4:14:40 PM	rw-r--r--	student59	JAVA File
SequenceFileWriteDe...	2 KB	2/15/2024 7:42:10 PM	rw-r--r--	student59	JAVA File
SmallFilesToSequence...	3 KB	4/12/2024 12:28:18 AM	rw-r--r--	student59	JAVA File
WholeFileInputFormat...	1 KB	4/12/2024 3:55:49 PM	rw-r--r--	student59	JAVA File
WholeFileRecordRead...	3 KB	4/12/2024 3:55:50 PM	rw-r--r--	student59	JAVA File
CatPython.py	1 KB	2/15/2024 6:57:12 PM	rw-r--r--	student59	PY File
CatSpark.py	1 KB	2/15/2024 6:57:14 PM	rw-r--r--	student59	PY File
is_good_quality.py	1 KB	2/15/2024 7:42:25 PM	rw-r--r--	student59	PY File
max_temperature_map...	1 KB	2/15/2024 7:42:08 PM	rw-r--r--	student59	PY File
max_temperature_red...	1 KB	2/15/2024 7:42:08 PM	rw-r--r--	student59	PY File
Project_file_map.py	1 KB	5/4/2024 9:08:18 PM	rw-r--r--	student59	PY File
Project_file_reduce.py	1 KB	5/4/2024 9:08:19 PM	rw-r--r--	student59	PY File
Project_map.py	1 KB	5/4/2024 9:08:18 PM	rw-r--r--	student59	PY File
Project_Map.py	1 KB	2/15/2024 7:42:08 PM	rw-r--r--	student59	PY File
Project_reduce.py	1 KB	5/4/2024 9:08:19 PM	rw-r--r--	student59	PY File
Project_Reduce.py	1 KB	5/5/2024 11:42:02 AM	rw-r--r--	student59	PY File
ReadSpark.py	1 KB	2/15/2024 6:57:19 PM	rw-r--r--	student59	PY File
derby.log	1 KB	4/28/2024 4:26:26 PM	rw-r--r--	student59	Text Document
pig_1714116207894.log	6 KB	4/26/2024 12:23:50 AM	rw-r--r--	student59	Text Document
pig_1714166838806.log	66 KB	4/26/2024 4:27:37 PM	rw-r--r--	student59	Text Document
pig_1714260394497.log	3 KB	4/27/2024 5:04:14 PM	rw-r--r--	student59	Text Document
pig_1714262713452.log	1,077 KB	4/27/2024 5:27:22 PM	rw-r--r--	student59	Text Document
1400-8.txt	1,010 KB	2/10/2024 6:40:20 PM	rw-r--r--	student59	Text Document
js3225_output.txt	1,157 KB	5/5/2024 1:00:38 PM	rw-r--r--	student59	Text Document
quangle.txt	1 KB	2/8/2024 5:32:38 PM	rw-r--r--	student59	Text Document
sample.txt	1 KB	2/15/2024 7:42:33 PM	rw-r--r--	student59	Text Document

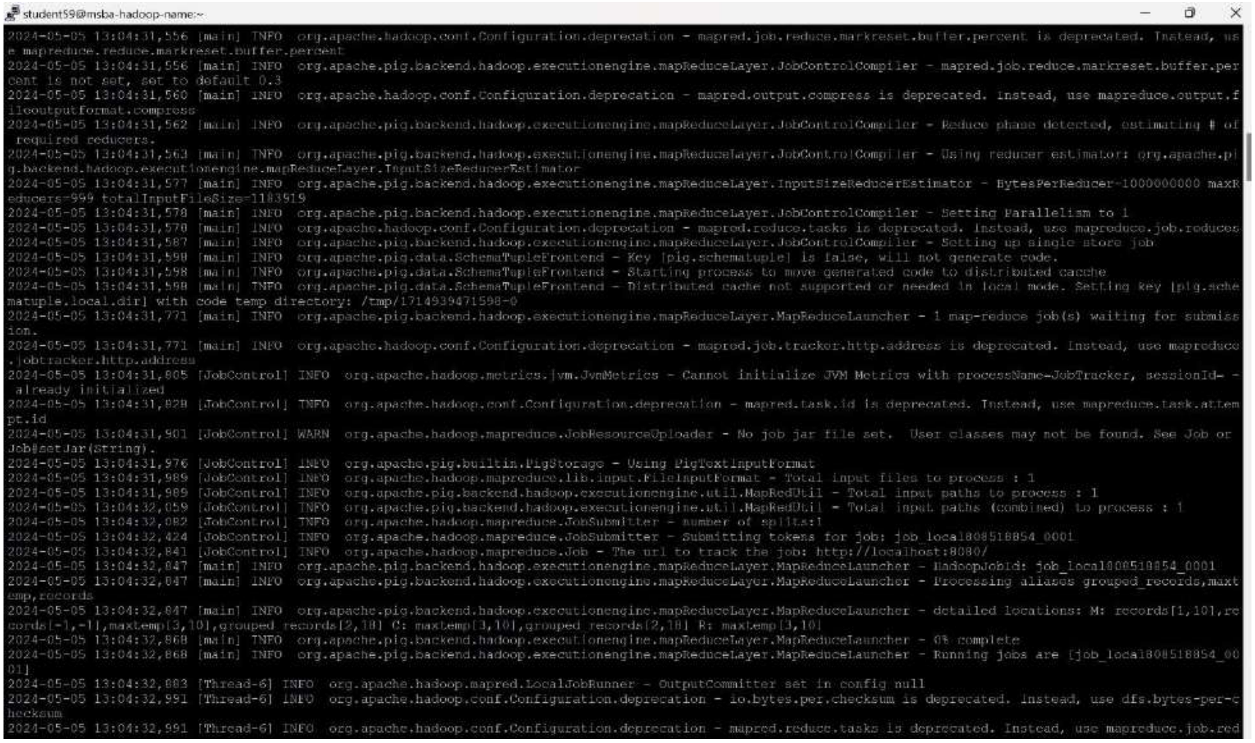
- **In Pig:** pig -x local
- records = LOAD 'js3225_output.txt' AS (year:chararray, temperature:int);
- grouped_records = GROUP records BY year;

- maxtemp = FOREACH grouped_records GENERATE group, MAX(records.temperature);
- DUMP maxtemp;

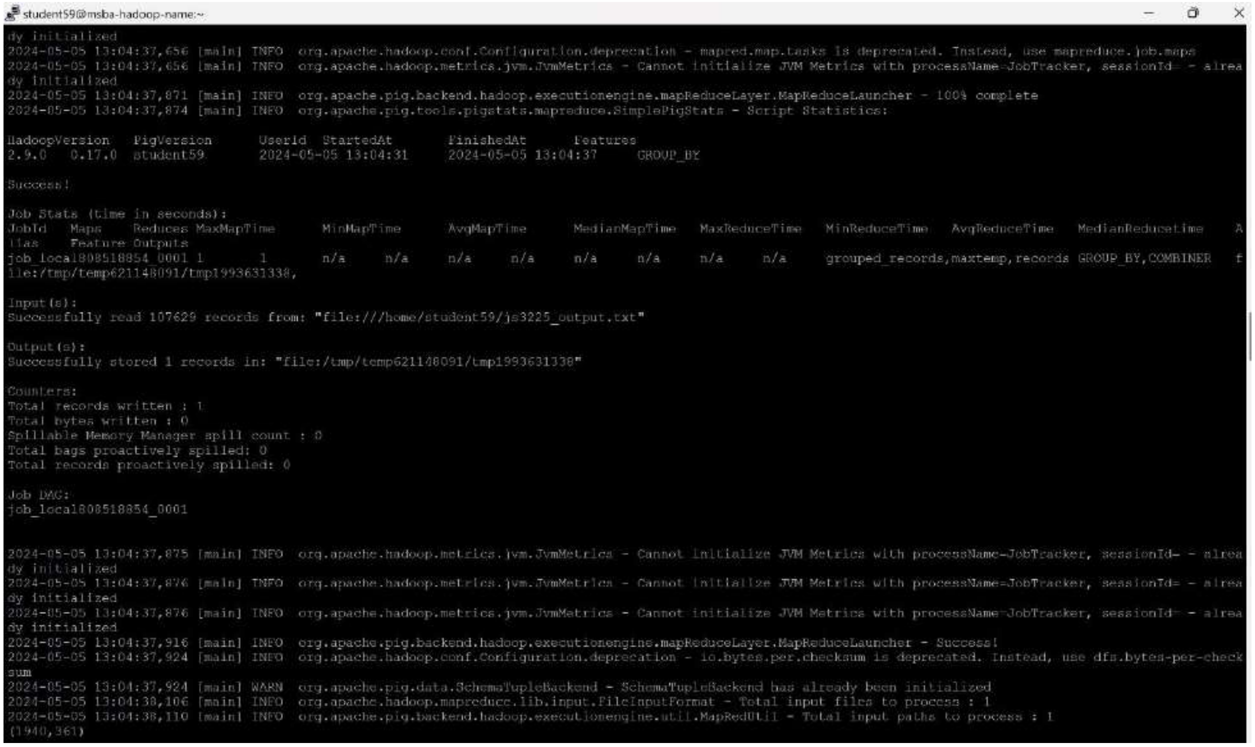


student59@msba-hadoop-name:~

```
[student59@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/StaticLoggerBinder.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 13:03:32 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
24/05/05 13:03:32 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-05-05 13:03:32,462 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-05-05 13:03:32,462 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/student59/pig_1714939412461.log
2024-05-05 13:03:32,480 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/student59/.pigbootup not found
2024-05-05 13:03:32,596 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.a
ddress
2024-05-05 13:03:32,598 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-05-05 13:03:32,780 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:03:32,797 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-cff05ee2-dfe4-4eca-af69-0b8df800cd4c
2024-05-05 13:03:32,797 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> records = LOAD 'js3225_output.txt' AS (year:chararray, temperature:int);
2024-05-05 13:03:49,575 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
grunt> grouped_records = GROUP records BY year;
grunt> maxtemp = FOREACH grouped_records GENERATE group, MAX(records.temperature);
grunt> DUMP maxtemp;
2024-05-05 13:04:30,218 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:04:30,317 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-05-05 13:04:30,361 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:04:30,563 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantC
alculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Predi
catePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-05-05 13:04:30,892 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsa
geThreshold = 489580128, usageThreshold = 489580128
2024-05-05 13:04:31,271 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic?
false
2024-05-05 13:04:31,316 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combine
r
2024-05-05 13:04:31,367 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-05-05 13:04:31,367 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-05-05 13:04:31,402 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:04:31,456 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - session.id is deprecated. Instead, use dfs.metrics.session-id
2024-05-05 13:04:31,463 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Initializing JVM Metrics with processName=JobTracker, sessionId=
2024-05-05 13:04:31,528 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2024-05-05 13:04:31,556 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, us
e mapreduce.reduce.markreset.buffer.percent
2024-05-05 13:04:31,556 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.per
```

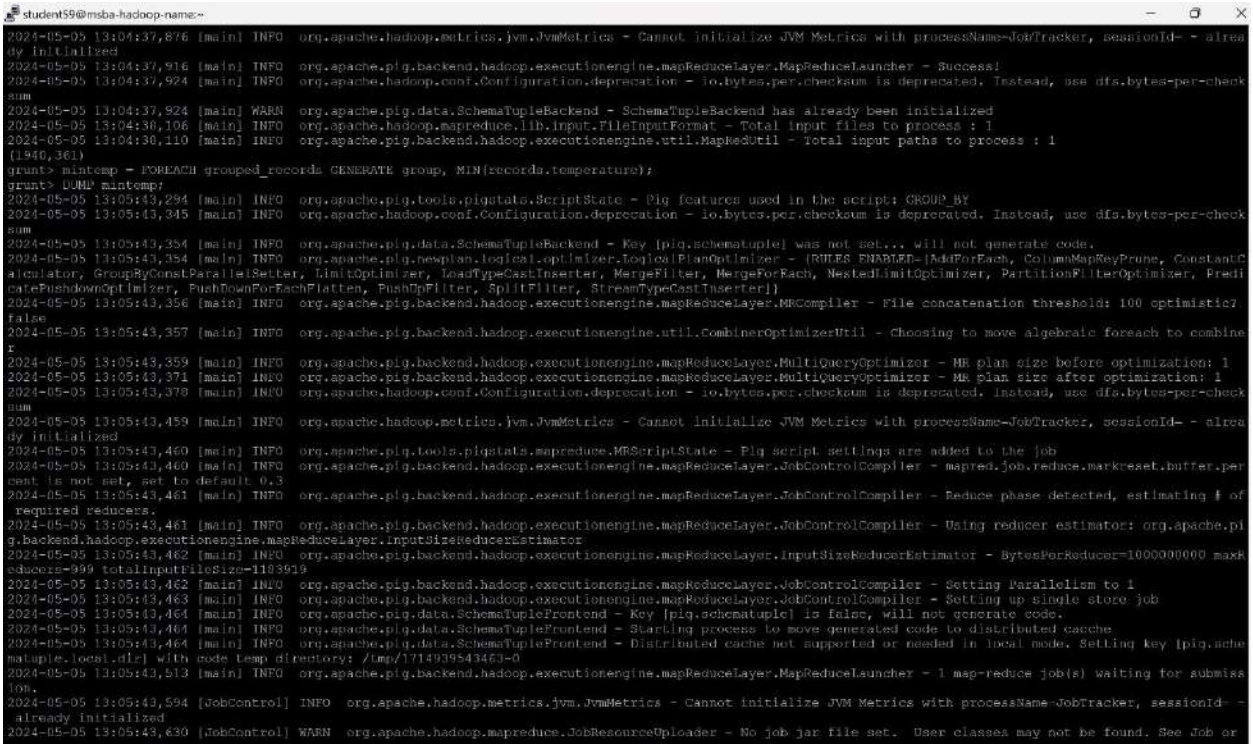


```
2024-05-05 13:04:31,556 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, us
e mapreduce.reduce.markreset.buffer.percent
2024-05-05 13:04:31,556 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.per
cent is not set, set to default 0.3
2024-05-05 13:04:31,560 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.f
ileoutputformat.compress
2024-05-05 13:04:31,562 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of
 required reducers.
2024-05-05 13:04:31,563 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pi
g.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-05-05 13:04:31,577 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxR
educers=999 totalInputFileSize=1183919
2024-05-05 13:04:31,578 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-05-05 13:04:31,578 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2024-05-05 13:04:31,587 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-05-05 13:04:31,598 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-05-05 13:04:31,598 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2024-05-05 13:04:31,598 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.sche
matuple.local.dir] with code temp directory: /tmp/1714939471598-0
2024-05-05 13:04:31,771 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submiss
ion.
2024-05-05 13:04:31,771 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce
.jobtracker.http.address
2024-05-05 13:04:31,805 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= -
 already initialized
2024-05-05 13:04:31,828 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attem
pt.id
2024-05-05 13:04:31,901 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or
Job#setJar(String).
2024-05-05 13:04:31,976 [JobControl] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2024-05-05 13:04:31,989 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-05-05 13:04:31,989 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2024-05-05 13:04:32,059 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2024-05-05 13:04:32,082 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2024-05-05 13:04:32,424 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local808518854_0001
2024-05-05 13:04:32,841 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2024-05-05 13:04:32,847 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local808518854_0001
2024-05-05 13:04:32,847 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases grouped_records,maxt
emp,records
2024-05-05 13:04:32,847 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: records[1,10],re
cords[-1,-1],maxtemp[3,10],grouped_records[2,18] C: maxtemp[3,10],grouped_records[2,18] R: maxtemp[3,10]
2024-05-05 13:04:32,868 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2024-05-05 13:04:32,868 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_local808518854_00
01]
2024-05-05 13:04:32,883 [Thread-6] INFO  org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2024-05-05 13:04:32,991 [Thread-6] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-c
hecksum
2024-05-05 13:04:32,991 [Thread-6] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.red
```

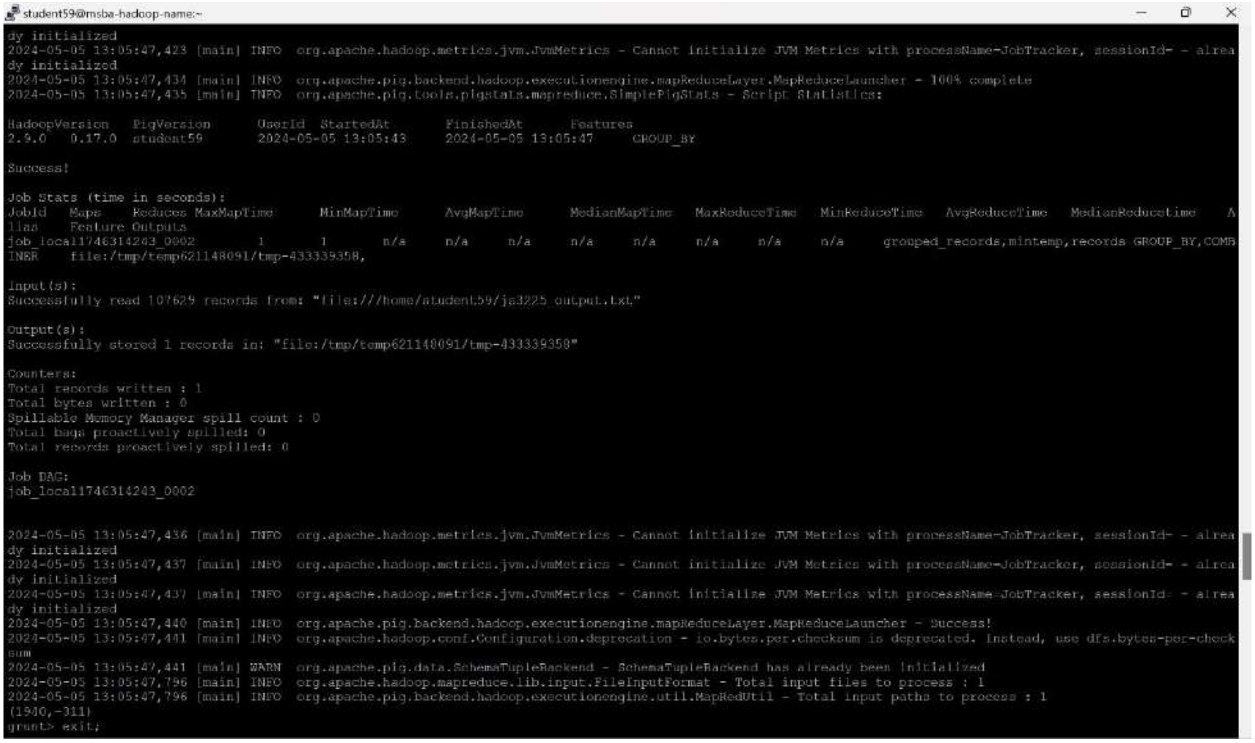


```
dy initialized
2024-05-05 13:04:37,656 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-05-05 13:04:37,656 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:04:37,871 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-05-05 13:04:37,874 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.9.0   0.17.0  student59       2024-05-05 13:04:31     2024-05-05 13:04:37     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime    A
lias    Feature Outputs
job_local808518854_0001 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     grouped_records,maxtemp,records GROUP_BY,COMBINER   f
ile:/tmp/temp621148091/tmp1993631338,

Input(s):
Successfully read 107629 records from: "file:///home/student59/js3225_output.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp621148091/tmp1993631338"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local808518854_0001


2024-05-05 13:04:37,875 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:04:37,876 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:04:37,876 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:04:37,916 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-05-05 13:04:37,924 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:04:37,924 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-05-05 13:04:38,106 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-05-05 13:04:38,110 [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;



```
2024-05-05 13:04:37,876 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:04:37,916 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-05-05 13:04:37,924 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:04:37,924 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-05-05 13:04:38,106 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-05-05 13:04:38,110 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1940,361)
grunt> mintemp = FOREACH grouped_records GENERATE group, MIN(records.temperature);
grunt> DUMP mintemp;
2024-05-05 13:05:43,294 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-05-05 13:05:43,345 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:05:43,354 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-05-05 13:05:43,354 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantC
alculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Predi
catePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-05-05 13:05:43,356 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic?
false
2024-05-05 13:05:43,357 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combine
r
2024-05-05 13:05:43,359 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-05-05 13:05:43,371 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-05-05 13:05:43,378 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:05:43,459 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:05:43,460 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2024-05-05 13:05:43,460 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.per
cent is not set, set to default 0.3
2024-05-05 13:05:43,461 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of
 required reducers.
2024-05-05 13:05:43,461 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pi
g.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-05-05 13:05:43,462 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxR
educers=999 totalInputFileSize=1183919
2024-05-05 13:05:43,462 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-05-05 13:05:43,463 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-05-05 13:05:43,464 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-05-05 13:05:43,464 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2024-05-05 13:05:43,464 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.sche
matuple.local.dir] with code temp directory: /tmp/1714939543463-0
2024-05-05 13:05:43,513 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submiss
ion.
2024-05-05 13:05:43,594 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= -
 already initialized
2024-05-05 13:05:43,630 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or
```



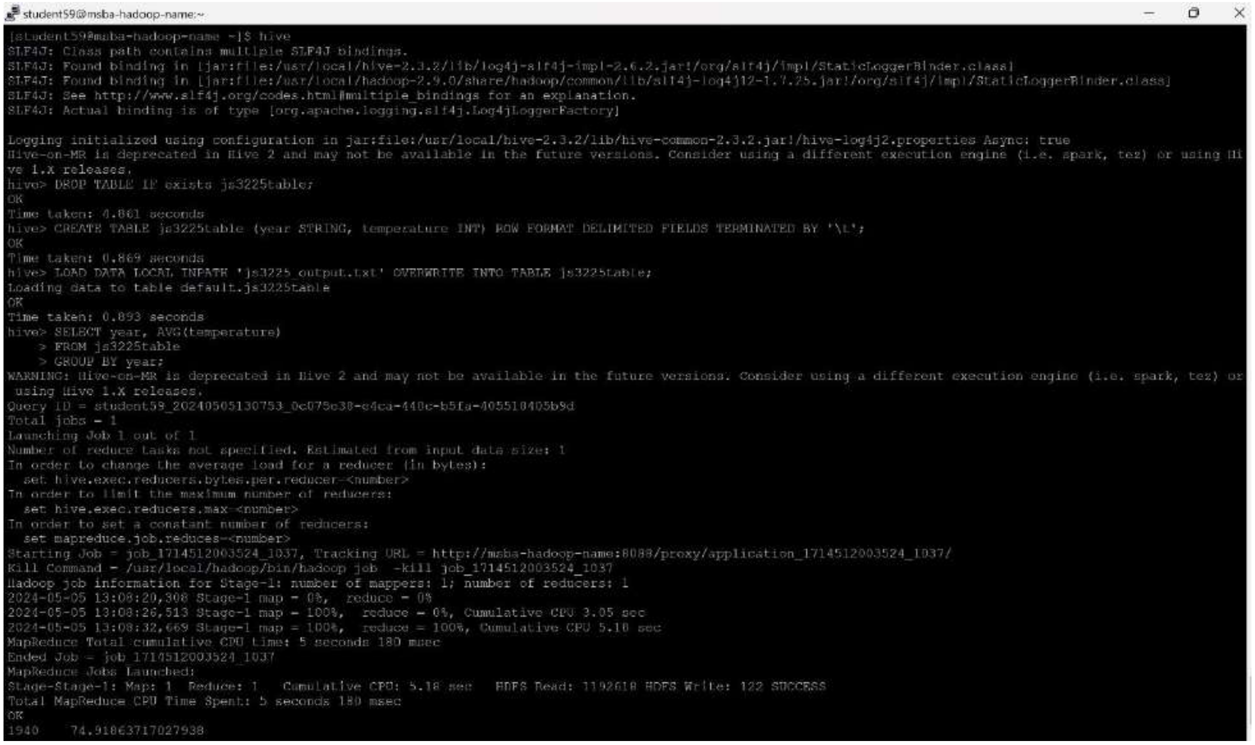
```
dy initialized
2024-05-05 13:05:47,423 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:05:47,434 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-05-05 13:05:47,435 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.9.0   0.17.0  student59       2024-05-05 13:05:43     2024-05-05 13:05:47     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime    A
lias    Feature Outputs
job_local1746314243_0002        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     grouped_records,mintemp,records GROUP_BY,COMB
INER    file:/tmp/temp621148091/tmp-433339358,

Input(s):
Successfully read 107629 records from: "file:///home/student59/js3225_output.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp621148091/tmp-433339358"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1746314243_0002


2024-05-05 13:05:47,436 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:05:47,437 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:05:47,437 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2024-05-05 13:05:47,440 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-05-05 13:05:47,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2024-05-05 13:05:47,441 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-05-05 13:05:47,796 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-05-05 13:05:47,796 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1940,-311)
grunt> exit;
```

- **In hive:**
- DROP TABLE IF exists js3225table;
- CREATE TABLE js3225table (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

- LOAD DATA LOCAL INPATH 'js3225_output.txt' OVERWRITE INTO TABLE js3225table;
- SELECT year, AVG(temperature)

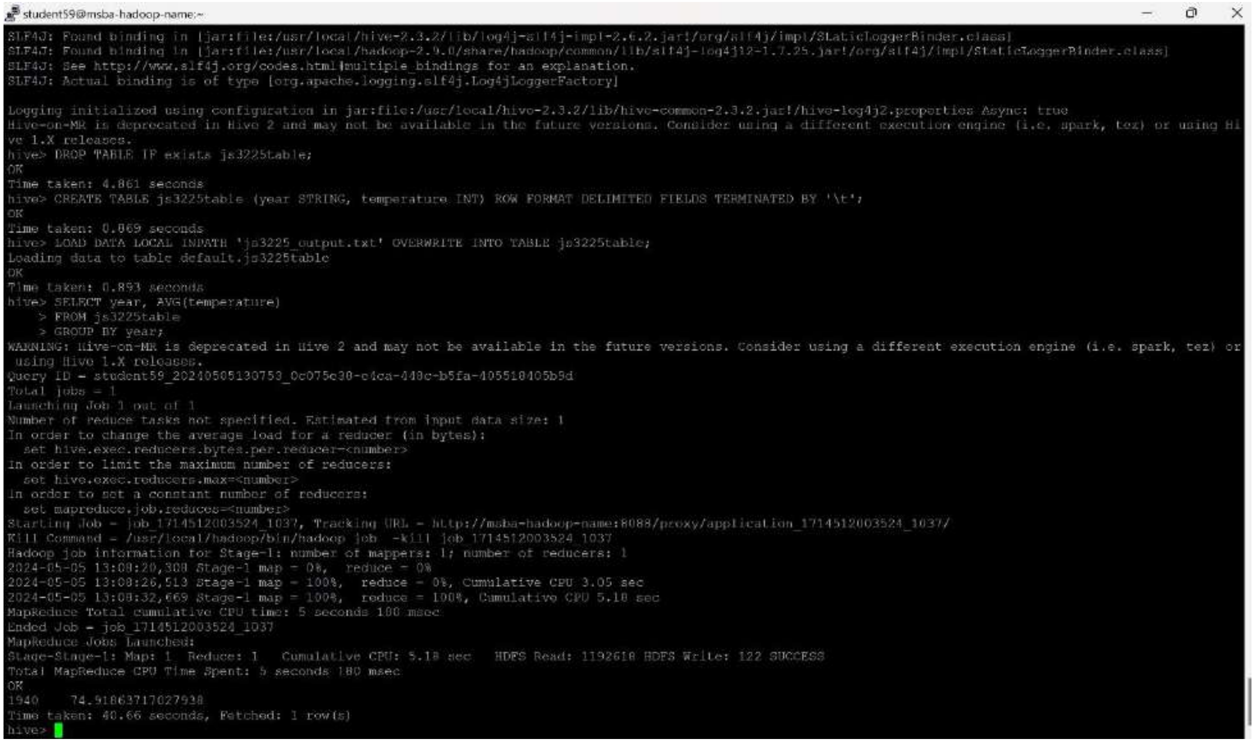
  FROM js3225table

  GROUP BY year;



student59@msba-hadoop-name:~

```
[student59@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 Hi
ve 1.X releases.
hive> DROP TABLE IF exists js3225table;
OK
Time taken: 4.861 seconds
hive> CREATE TABLE js3225table (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.869 seconds
hive> LOAD DATA LOCAL INPATH 'js3225_output.txt' OVERWRITE INTO TABLE js3225table;
Loading data to table default.js3225table
OK
Time taken: 0.893 seconds
hive> SELECT year, AVG(temperature)
    > FROM js3225table
    > 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 = student59_20240505130753_0c075c38-e4ca-440c-b5fa-405518405b9d
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_1037, Tracking URL = http://msba-hadoop-name:8088/proxy/application_1714512003524_1037/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1714512003524_1037
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-05 13:08:20,308 Stage-1 map = 0%,  reduce = 0%
2024-05-05 13:08:26,513 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
2024-05-05 13:08:32,669 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.18 sec
MapReduce Total cumulative CPU time: 5 seconds 180 msec
Ended Job = job_1714512003524_1037
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.18 sec   HDFS Read: 1192618 HDFS Write: 122 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 180 msec
OK
1940    74.91863717027938
```



```
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 Hi
ve 1.X releases.
hive> DROP TABLE IF exists js3225table;
OK
Time taken: 4.861 seconds
hive> CREATE TABLE js3225table (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.869 seconds
hive> LOAD DATA LOCAL INPATH 'js3225_output.txt' OVERWRITE INTO TABLE js3225table;
Loading data to table default.js3225table
OK
Time taken: 0.893 seconds
hive> SELECT year, AVG(temperature)
    > FROM js3225table
    > 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 = student59_20240505130753_0c075c38-c4ca-448c-b5fa-405518405b9d
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_1037, Tracking URL = http://msba-hadoop-name:8088/proxy/application_1714512003524_1037/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1714512003524_1037
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-05 13:08:20,308 Stage-1 map = 0%,  reduce = 0%
2024-05-05 13:08:26,513 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
2024-05-05 13:08:32,669 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.18 sec
MapReduce Total cumulative CPU time: 5 seconds 180 msec
Ended Job = job_1714512003524_1037
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.18 sec   HDFS Read: 1192618 HDFS Write: 122 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 180 msec
OK
1940    74.91863717027938
Time taken: 40.66 seconds, Fetched: 1 row(s)
hive> 
```

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.