Lab 07 – Apache Pig

Full name: Chloe Tee Rouyi

Student ID: 0354731

Tasks:

1. Open terminal

2. Check running services: **sudo jps**

```
[cloudera@quickstart ~]$ sudo jps
7413 HistoryServer
6917 RunJar
8328
6739 ThriftServer
7377 Bootstrap
8257 Bootstrap
5285 DataNode
6840 RunJar
6601 RESTServer
5871 NodeManager
5732 Bootstrap
5604 SecondaryNameNode
5212 QuorumPeerMain
5386 JournalNode
7994 Bootstrap
8285
6191 ResourceManager
5472 NameNode
5787 JobHistoryServer
16648 Jps
[cloudera@quickstart ~]$
```

3. Run Pig in local mode **pig -x local**

after a while grunt shell will be appeared. Grunt is Pig's interactive shell. The prompt will be “grunt>”

```
[cloudera@quickstart ~]$ pig -x local
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2024-11-19 18:31:06,528 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.10.0 (rexported) compiled Jan 20 2017, 12:05:43
2024-11-19 18:31:06,528 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1732069866495.log
2024-11-19 18:31:06,547 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2024-11-19 18:31:06,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-19 18:31:06,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-11-19 18:31:06,839 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-11-19 18:31:07,263 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-11-19 18:31:07,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-19 18:31:07,443 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess

2024-11-19 18:31:08,098 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-11-19 18:31:08,197 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-19 18:31:08,198 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-11-19 18:31:08,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-11-19 18:31:08,316 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-19 18:31:08,317 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-11-19 18:31:08,318 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-11-19 18:31:08,376 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-19 18:31:08,378 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-11-19 18:31:08,379 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt>
```

4. Short Research: What is Pig in Local mode?

In this mode, Pig accesses files stored on the local file system. Data processing happens on the local machine. This mode is generally used for testing locally and speeding up development.

5. Execute the following commands and write your understanding about what each command does based on the previous practices:

   a. grunt> **ls**

```
grunt> ls
file:/home/cloudera/.vboxclient-draganddrop.pid<r 1>    5
file:/home/cloudera/Desktop     <dir>
file:/home/cloudera/.mozilla    <dir>
file:/home/cloudera/kerberos<r 1>       5007
file:/home/cloudera/Templates   <dir>
file:/home/cloudera/.esd_auth<r 1>      16
file:/home/cloudera/Videos      <dir>
file:/home/cloudera/Downloads   <dir>
file:/home/cloudera/parcels<r 1>        4228
file:/home/cloudera/WordCount.jar<r 1>  3349
file:/home/cloudera/.gtk-bookmarks<r 1> 152
file:/home/cloudera/.gnome2_private     <dir>
file:/home/cloudera/.vboxclient-clipboard.pid<r 1>      5
file:/home/cloudera/workspace   <dir>
file:/home/cloudera/Pictures    <dir>
file:/home/cloudera/.fontconfig <dir>
file:/home/cloudera/.xsession-errors<r 1>       1733
file:/home/cloudera/.pulse-cookie<r 1>  256
file:/home/cloudera/.pulse      <dir>
file:/home/cloudera/Processfile2.txt<r 1>       167
file:/home/cloudera/enterprise-deployment.json<r 1>     53655
file:/home/cloudera/.dbus       <dir>
file:/home/cloudera/.config     <dir>
file:/home/cloudera/.gconf      <dir>
file:/home/cloudera/.bash_logout<r 1>   18
file:/home/cloudera/.xsession-errors.old<r 1>   2470
file:/home/cloudera/.gnome2     <dir>
file:/home/cloudera/.bash_history<r 1>  1203
file:/home/cloudera/.bash_profile<r 1>  176
file:/home/cloudera/express-deployment.json<r 1>        50515
file:/home/cloudera/.gconfd     <dir>
file:/home/cloudera/.local      <dir>
file:/home/cloudera/Music       <dir>
file:/home/cloudera/.gstreamer-0.10     <dir>
file:/home/cloudera/eclipse     <dir>
file:/home/cloudera/.cache      <dir>
file:/home/cloudera/.bashrc<r 1>        176
file:/home/cloudera/Documents   <dir>
file:/home/cloudera/.vboxclient-display.pid<r 1>        5
file:/home/cloudera/Public      <dir>
file:/home/cloudera/lib <dir>
file:/home/cloudera/cloudera-manager<r 1>       5387
file:/home/cloudera/.ICEauthority<r 1>  620
file:/home/cloudera/cm_api.py<r 1>      9964
file:/home/cloudera/.vboxclient-seamless.pid<r 1>       5
file:/home/cloudera/.gvfs       <dir>
file:/home/cloudera/Processfile.txt<r 1>        66
file:/home/cloudera/.nautilus   <dir>
file:/home/cloudera/.pig_history<r 1>   3
```

b. grunt> **cat mysales.txt**

```
grunt> cat mysales.txt
2024-11-19 18:35:50,089 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 2997: Encountered IOException. Directory mys
ales.txt does not exist.
Details at logfile: /home/cloudera/pig_1732069866495.log
```

Error: file does not exist

c. grunt> **quit**

```
grunt> quit
[cloudera@quickstart ~]$
```

6. Run Pig in MapReduce (HDFS) mode

a. $ **pig**

OR

b. $ **pig -x mapreduce**

```
[cloudera@quickstart ~]$ pig -x mapreduce
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2024-11-19 18:37:27,516 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.10.0 (rexported) compiled Jan 20
 2017, 12:05:43
2024-11-19 18:37:27,516 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1732070247494.log
2024-11-19 18:37:27,534 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/cloudera/.pigbootup not fo
und
2024-11-19 18:37:28,382 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:28,382 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:28,382 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop
 file system at: hdfs://quickstart.cloudera:8020
2024-11-19 18:37:29,777 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:29,777 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-re
duce job tracker at: localhost:8021
2024-11-19 18:37:29,783 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:29,861 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:29,862 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.


Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:30,113 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:30,113 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:30,178 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:30,179 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:30,278 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:30,279 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:30,345 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:30,345 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:30,511 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:30,512 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 18:37:30,571 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 18:37:30,574 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
grunt>
```

7. Short Research: What is Pig in MapReduce (HDFS) mode?

MapReduce mode - In this mode, Pig loads and processes the data stored on HDFS. Pig Latin statements invoke a MapReduce job to perform the processing. It is the recommended mode in a production environment.

8. What are the differences between Pig in local mode ad Pig in MapReduce (HDFS) mode?

Both MapReduce mode and local mode seem same to the user but the difference is the way they execute.

1) **MapReduce mode:**

In MapReduce mode, Pig script is executed on Hadoop cluster. The Pig scripts are converted into MapReduce jobs and then executed on Hadoop cluster (hdfs)

2) **Local mode:**

In this mode, Pig script runs on a Single machine without the need of Hadoop cluster or hdfs. Local mode is used for development purpose to see how the script would behave in an actual environment.

9. Execute the following commands and write your understanding about what each command does based on the previous practices:

a. grunt>**mkdir /user/myinput**

```
grunt> mkdir /user/myinput
```

b. grunt> **ls /user/myinput**

```
grunt> ls /user
hdfs://quickstart.cloudera:8020/user/cloudera   <dir>
hdfs://quickstart.cloudera:8020/user/history    <dir>
hdfs://quickstart.cloudera:8020/user/hive       <dir>
hdfs://quickstart.cloudera:8020/user/hue        <dir>
hdfs://quickstart.cloudera:8020/user/jenkins    <dir>
hdfs://quickstart.cloudera:8020/user/myinput    <dir>
hdfs://quickstart.cloudera:8020/user/oozie      <dir>
hdfs://quickstart.cloudera:8020/user/root       <dir>
hdfs://quickstart.cloudera:8020/user/spark      <dir>
```

c. grunt> **quit**

```
grunt> quit
[cloudera@quickstart ~]$
```

d. Use the commands that you learned for HDFS and create a file with name **mysales.txt** in "**/user/myinput/**". Use the following content and add it to the create file "**/user/myinput/mysales.txt**". Write all the steps.

Content of the file "**/user/myinput/mysales.txt**" should be:

E2001,400,4000.00
E2004,300,3000.30
E2011,500,5500.55
E2012,200,2000.20
E2001,100,500.50
E2011,600,7000.70

```
[cloudera@quickstart ~]$ cat > /home/cloudera/mysales.txt

E2001,400,4000.00
E2004,300,3000.30
E2011,500,5500.55
E2012,200,2000.20
E2001,100,500.50
E2011,600,7000.70
[cloudera@quickstart ~]$ cat /home/cloudera/mysales.txt

E2001,400,4000.00
E2004,300,3000.30
E2011,500,5500.55
E2012,200,2000.20
E2001,100,500.50
E2011,600,7000.70
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/mysales.txt /user/myinput/
[cloudera@quickstart ~]$ hdfs dfs -cat /user/myinput/mysales.txt

E2001,400,4000.00
E2004,300,3000.30
E2011,500,5500.55
E2012,200,2000.20
E2001,100,500.50
E2011,600,7000.70
```

e. $ **pig**

```
grunt> cat /user/myinput/mysales.txt

E2001,400,4000.00
E2004,300,3000.30
E2011,500,5500.55
E2012,200,2000.20
E2001,100,500.50
E2011,600,7000.70
grunt>
```

f. grunt> **cat /user/myinput/mysales.txt**

```
grunt> cat /user/myinput/mysales.txt

E2001,400,4000.00
E2004,300,3000.30
E2011,500,5500.55
E2012,200,2000.20
E2001,100,500.50
E2011,600,7000.70
```

Create a Pig table (relation):

10. To create Table 'employee' from mysales.txt

    a. grunt> **employee = LOAD 'hdfs://quickstart.cloudera:8020/user/myinput' USING PigStorage(',') AS (eid:chararray,sales:int,comm:double);**

```
grunt> employee = LOAD 'hdfs://quickstart.cloudera:8020/user/myinput'
>> USING PigStorage(',') AS (eid:chararray,sales:int,comm:double);
```

11. To view the content of employee

    a. grunt> **DUMP employee**

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0001  1       0       7       7       7       7       n/a     n/a     n/a     n/a     employee        MA
P_ONLY  hdfs://quickstart.cloudera:8020/tmp/temp-137711490/tmp1759285407,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 7 records (150 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-137711490/tmp1759285407"

Counters:
Total records written : 7
Total bytes written : 150
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0001


2024-11-19 20:08:16,234 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Enco
untered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2024-11-19 20:08:16,234 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
2024-11-19 20:08:16,238 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 20:08:16,238 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 20:08:16,239 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will no
t generate code.
2024-11-19 20:08:16,259 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2024-11-19 20:08:16,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
(,,)
(E2001,400,4000.0)
(E2004,300,3000.3)
(E2011,500,5500.55)
(E2012,200,2000.2)
(E2001,100,500.5)
(E2011,600,7000.7)
```

12. To view the structure of employee

a. grunt> **DESCRIBE employee**

```
grunt> DESCRIBE employee
employee: {eid: chararray,sales: int,comm: double}
```

13. To read data from MapReduce result (tab delimited)

a. grunt> **employee2 = LOAD 'hdfs://quickstart.cloudera:8020/outputfolder/part-r-00000' USING PigStorage('\t') AS (eid:chararray,comm:double);**

```
grunt> employee2 = LOAD
>> 'hdfs://quickstart.cloudera:8020/outputfolder/part-r-00000' USING
>> PigStorage('\t') AS (eid:chararray,comm:double);
```

b. grunt> **DESCRIBE employee2**

```
grunt> DESCRIBE employee2
employee2: {eid: chararray,comm: double}
```

c. grunt> **DUMP employee2**

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0002  1       0       8       8       8       8       n/a     n/a     n/a     n/a     employee2       MA
P_ONLY  hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp217276531,

Input(s):
Successfully read 8 records (426 bytes) from: "hdfs://quickstart.cloudera:8020/outputfolder/part-r-00000"

Output(s):
Successfully stored 8 records (154 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp217276531"

Counters:
Total records written : 8
Total bytes written : 154
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0002


2024-11-19 20:14:24,721 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
2024-11-19 20:14:24,725 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 20:14:24,725 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 20:14:24,725 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will no
t generate code.
2024-11-19 20:14:24,769 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2024-11-19 20:14:24,769 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
(COUNT,2.0)
(HI,7.0)
(WORD,2.0)
(chloe,1.0)
(hi,2.0)
(is,1.0)
(my,1.0)
(name,1.0)
```

Selected Pig table operations

14. To see transactions with sales quantity greater than 400

    a. grunt> **highSales = FILTER employee BY sales > 400;**

```
grunt> highSales = FILTER employee BY sales > 400;
grunt>
```

    b. grunt> **DUMP highSales**

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0003  1       0       7       7       7       7       n/a     n/a     n/a     n/a     employee,highSales
MAP_ONLY        hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp-1862497999,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 2 records (48 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp-1862497999"

Counters:
Total records written : 2
Total bytes written : 48
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0003

2024-11-19 20:25:49,576 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Enco
untered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2024-11-19 20:25:49,576 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
2024-11-19 20:25:49,577 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 20:25:49,577 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 20:25:49,577 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will no
t generate code.
2024-11-19 20:25:49,601 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2024-11-19 20:25:49,601 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
(E2011,500,5500.55)
(E2011,600,7000.7)
```

15. To sort transactions based on highest commission

a. grunt> **highestComm = ORDER employee BY comm DESC;**

```
grunt> highestComm = ORDER employee BY comm DESC;
```

b. grunt> **DUMP highComm**

```
grunt> DUMP highComm
2024-11-19 20:27:45,195 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1003: Unable to find an operator for alias h
ighComm
Details at logfile: /home/cloudera/pig_1732075937612.log
```

Error: highComm is not the correct name for the table, the correct command should be DUMP highestComm

```
grunt> DUMP highestComm
2024-11-19 20:27:58,184 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY
2024-11-19 20:27:58,185 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddFo


Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0004  1       0       7       7       7       7       n/a     n/a     n/a     n/a     employee        MA
P_ONLY
job_1730968617936_0005  1       1       7       7       7       7       8       8       8       8       highestComm     SA
MPLER
job_1730968617936_0006  1       1       7       7       7       7       8       8       8       8       highestComm     OR
DER_BY  hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp1096089914,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 7 records (150 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp1096089914"

Counters:
Total records written : 7
Total bytes written : 150
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0


Job DAG:
job_1730968617936_0004  ->      job_1730968617936_0005,
job_1730968617936_0005  ->      job_1730968617936_0006,
job_1730968617936_0006


2024-11-19 20:30:01,539 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
(E2011,600,7000.7)
(E2011,500,5500.55)
(E2001,400,4000.0)
(E2004,300,3000.3)
(E2012,200,2000.2)
(E2001,100,500.5)
(,,)
grunt>
```

Note: For Mathematical functions, group the table

16. To find total overall sales quantity

a. grunt> **employeeGR = GROUP employee ALL;**

```
grunt> employeeGR = GROUP employee ALL;
grunt>
```

b. grunt> **sumSales = foreach employeeGR GENERATE (employee.eid), SUM(employee.sales);**

```
grunt> sumSales = foreach employeeGR GENERATE (employee.eid),SUM(employee.sales);
grunt>
```

c. grunt> **DUMP sumSales**

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0007  1       1       7       7       7       7       8       8       8       8       employee,employeeG
R,sumSales      GROUP_BY        hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp564517988,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 1 records (65 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp564517988"

Counters:
Total records written : 1
Total bytes written : 65
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0007


2024-11-19 20:35:34,050 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Enco
untered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2024-11-19 20:35:34,050 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
2024-11-19 20:35:34,055 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 20:35:34,055 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 20:35:34,056 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will no
t generate code.
2024-11-19 20:35:34,068 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2024-11-19 20:35:34,068 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
({(E2011),(E2001),(E2012),(E2011),(E2004),(E2001),()},2100)
```

17. To find total commission for each employee

a. grunt> **employeeEID = GROUP employee BY eid;**

```
grunt> employeeEID = GROUP employee BY eid;
grunt>
```

b. grunt> **totComm = foreach employeeEID GENERATE (employee.eid), SUM(employee.comm);**

```
grunt> totComm = foreach employeeEID GENERATE (employee.eid), SUM(employee.comm);
grunt>
```

c. grunt> **DUMP totComm**

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0009  1       1       6       6       6       6       8       8       8       8       employee,employeeE
ID,totComm      GROUP_BY        hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp-1329502530,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 5 records (123 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1397961678/tmp-1329502530"

Counters:
Total records written : 5
Total bytes written : 123
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0009

2024-11-19 20:41:40,455 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Enco
untered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2024-11-19 20:41:40,455 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
2024-11-19 20:41:40,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 20:41:40,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2024-11-19 20:41:40,456 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will no
t generate code.
2024-11-19 20:41:40,465 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2024-11-19 20:41:40,465 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
({(E2001),(E2001)},4500.5)
({(E2004)},3000.3)
({(E2011),(E2011)},12501.25)
({(E2012)},2000.2)
({()},)
```

Saving Data

18. To see all executed commands

    a. grunt> **history**

```
grunt> history
1   employee2 = LOAD

'hdfs://quickstart.cloudera:8020/outputfolder/part-r-00000' USING

PigStorage('\t') AS (eid:chararray,comm:double);
2   employee = LOAD 'hdfs://quickstart.cloudera:8020/user/myinput' USING PigStorage(',') AS (eid:chararray,sales:int,comm:
double);
3   highSales = FILTER employee BY sales > 400;
4   highestComm = ORDER employee BY comm DESC;
5   employeeGR = GROUP employee ALL;
6   employeeGR = GROUP employee ALL;
7   sumSales = foreach employeeGR GENERATE (employee.eid),SUM(employee.sales);
8   employeeEID = GROUP employee BY eid;
9   totComm = foreach employeeEID GENERATE (employee.eid), SUM(employee.comm);
```

19. To save result (HDFS) in CSV (comma delimited format)

    a. grunt> **STORE totComm INTO 'hdfs://localhost:9000/user/mypigoutput' using PigStorage(',');**

    The result is stored in folder user/mypigoutput (automatically created)

```
grunt> STORE totComm INTO 'hdfs://localhost:9000/user/mypigoutput' using PigStorage(',');
2024-11-19 20:43:32,409 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-11-19 20:43:32,410 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddFo
rEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer
, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, Spli
tFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-11-19 20:43:32,414 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator
is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2024-11-19 20:43:32,427 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 6000:
```

Error: the port 9000 is incorrect, the correct port is 8020, therefore the correct command should be **STORE totComm INTO 'hdfs://localhost:8020/user/mypigoutput' using PigStorage(',');**

```
grunt> STORE totComm INTO 'hdfs://localhost:8020/user/mypigoutput' using PigStorage(',');
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0010  1       1       7       7       7       7       8       8       8       8       employee,employeeE
ID,totComm      GROUP_BY        hdfs://localhost:8020/user/mypigoutput,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 5 records (92 bytes) in: "hdfs://localhost:8020/user/mypigoutput"

Counters:
Total records written : 5
Total bytes written : 92
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0010

2024-11-19 20:44:37,466 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Enco
untered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2024-11-19 20:44:37,466 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
```

20. To save result (LOCAL) in Tab delimited format

a. grunt> **STORE totComm INTO 'mypigoutput' using PigStorage('\t');** The result is stored in home subfolder mypigoutput (automatically created) **Running Script (Saving commands)**

```
grunt> STORE totComm INTO 'mypigoutput' using PigStorage('\t');

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0011  1       1       7       7       7       8       8       8       employee,employeeE
ID,totComm      GROUP_BY        hdfs://quickstart.cloudera:8020/user/cloudera/mypigoutput,

Input(s):
Successfully read 7 records (494 bytes) from: "hdfs://quickstart.cloudera:8020/user/myinput"

Output(s):
Successfully stored 5 records (92 bytes) in: "hdfs://quickstart.cloudera:8020/user/cloudera/mypigoutput"

Counters:
Total records written : 5
Total bytes written : 92
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0011

2024-11-19 20:49:06,785 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Enco
untered Warning ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2024-11-19 20:49:06,785 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
ess!
```

To save commands as script Create script file named myscript.pig Enter the following as content:

```
 GNU nano 2.0.9                         File: myscript.pig

employee = LOAD 'hdfs://localhost:8020/user/myinput' USING PigStorage(',') AS (eid:chararray,sales:int,comm:double);
DESCRIBE employee;
```

21. If using Pig in HDFS mode:

a. grunt>**employee = LOAD 'hdfs://localhost:9000/user/myinput2' USING PigStorage(',') AS (eid:chararray,sales:int,comm:double);**

b. grunt>**DESCRIBE employee;**

22. If using Pig in LOCAL mode:

a. grunt>**employee = LOAD 'mysales.txt' USING PigStorage(',') AS (eid:chararray,sales:int,comm:double);**

b. grunt>**DESCRIBE employee;**

23. To execute myscript.pig in /user/myinput folder using Pig in HDFS mode:

a. grunt> **RUN hdfs://localhost:9000/user/myinput/myscript.pig;**

This doesn't work as RUN syntax should be RUN /path/to/myscript.pig

b. grunt> **DUMP employee**

24. To execute myscript.pig in home folder using Pig in LOCAL mode:

a. grunt> **RUN myscript.pig;**

```
grunt> RUN myscript.pig
2024-11-19 21:21:26,079 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Ins
tead, use fs.defaultFS
2024-11-19 21:21:26,080 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
grunt> employee = LOAD 'hdfs://localhost:8020/user/myinput' USING PigStorage(',') AS (eid:chararray,sales:int,comm:double)
;
grunt> DESCRIBE employee;
employee: {eid: chararray,sales: int,comm: double}
grunt>
```

b. grunt> **DUMP employee**

```
grunt> DUMP employee
```

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   Av
gReduceTime     MedianReducetime        Alias   Feature Outputs
job_1730968617936_0012  1       0       8       8       8       8       n/a     n/a     n/a     n/a     employee        MA
P_ONLY  hdfs://quickstart.cloudera:8020/tmp/temp-1962885059/tmp1026856895,

Input(s):
Successfully read 7 records (484 bytes) from: "hdfs://localhost:8020/user/myinput"

Output(s):
Successfully stored 7 records (150 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1962885059/tmp1026856895"

Counters:
Total records written : 7
Total bytes written : 150
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1730968617936_0012
```

25. Save and submit your Lab documents in PDF with the following filename format. Submit your report via the submission link for this available on MyTIMeS.

Filename format: Name_ID_Lab07