

## Task 1.1:

### Word-Count in Pig:

I have put my file in /hadoopdata/pig/ directory

```
[acadgild@localhost ~]$ vi test.txt
[acadgild@localhost ~]$ cat test.txt
welcome to hadoop program.
This is acadgild. test!
This is morning session. Session lasts for 2 hrs.
Evening sessions might extend some times
[acadgild@localhost ~]$ hadoop fs -put test.txt /hadoopdata/pig
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
18/09/03 23:25:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop fs -ls /hadoopdata/pig
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
18/09/03 23:25:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r-- 1 acadgild supergroup      12 2018-03-15 12:02 /hadoopdata/pig/A.txt
-rw-r--r-- 1 acadgild supergroup      12 2018-03-15 12:02 /hadoopdata/pig/B.txt
-rw-r--r-- 1 acadgild supergroup    144 2018-09-03 23:25 /hadoopdata/pig/test.txt
[acadgild@localhost ~]$ hadoop fs -cat /hadoopdata/pig/test.txt
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
18/09/03 23:26:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
welcome to hadoop program.
This is acadgild. test!
This is morning session. Session lasts for 2 hrs.
Evening sessions might extend some times
```

### Loaded the data into A

```
grunt> A = load '/hadoopdata/pig/test.txt';
2018-09-03 23:27:33,889 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
y tried 5 time(s), retry policy is not yet implemented with execstack(maxRetries=10, sleepTime=1000 milliseconds)
2018-09-03 23:34:29,839 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-03 23:34:29,839 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-03 23:34:29,862 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-03 23:34:30,010 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-03 23:34:30,019 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(welcome to hadoop program.)
(This is acadgild. test!)
(This is morning session. Session lasts for 2 hrs.)
(Evening sessions might extend some times)
grunt> █
```

Tokenize function splits a string in a single tuple and returns a bag with a tuple for each chararray that results from the split. Flatten un-nests tuples as well as bags.

```
grunt> B = FOREACH A GENERATE flatten(TOKENIZE((chararray)$0)) as word;
grunt> █
```

```

!
2018-09-03 23:51:57,247 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2018-09-03 23:51:57,323 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-03 23:51:57,324 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(welcome)
(to)
(hadoop)
(program.)
(This)
(is)
(acadgild.)
(test!)
(This)
(is)
(morningning)
(session.)
(Session)
(lasts)
(for)
(2)
(hrs.)
(Evening)
(sessions)
(might)
(extend)
(some)
(times)
grunt> █

```

## Grouping by similar words

```

grunt> C = group B by word;
grunt> █ acadgild@localhost:~

```

## Generating count of each word

```

grunt> C = group B by word;
grunt> D = foreach C generate group, COUNT(B);
grunt> dump D; █

```

```

generate code.
2018-09-04 00:08:16,762 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 00:08:16,762 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(2,1)
(is,2)
(to,1)
(for,1)
(This,2)
(hrs.,1)
(some,1)
(lasts,1)
(might,1)
(test!,1)
(times,1)
(extend,1)
(hadoop,1)
(Evening,1)
(Session,1)
(welcome,1)
(program.,1)
(session.,1)
(sessions,1)
(acadgild.,1)
(morningning,1)
grunt> █

```

## Task1.2:

To start pig in local mode use command `pig -x local`

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ pig -x local  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/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]  
2018-09-04 01:32:11,837 INFO [main] pig.ExecTypeProvider: Trying ExecType : LOCAL  
2018-09-04 01:32:11,842 INFO [main] pig.ExecTypeProvider: Picked LOCAL as the ExecType  
2018-09-04 01:32:11,986 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49  
2018-09-04 01:32:11,986 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1536004931982.log  
2018-09-04 01:32:12,117 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found  
2018-09-04 01:32:12,978 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2018-09-04 01:32:12,979 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2018-09-04 01:32:12,986 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///   
2018-09-04 01:32:13,393 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2018-09-04 01:32:13,552 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-03fa5e93-f5f0-4fb6-8a91-ff9aaf900e2a  
2018-09-04 01:32:13,552 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false  
grunt> █
```

Loading data into emp\_details

```
grunt> emp_details = load 'employee_details.txt' using PigStorage(',') as (EmpID:int,Name:chararray,Salary:int,DepartmentID:int);  
2018-09-04 01:49:45,783 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2018-09-04 01:49:45,783 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
grunt> █ acadgild@localhost:~
```

To see the content of emp\_details use command dump emp\_details

```
2018-09-04 01:50:33,311 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized  
2018-09-04 01:50:33,724 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1  
2018-09-04 01:50:33,725 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1  
(101,Amitabh,20000,1)  
(102,Shahrukh,10000,2)  
(103,Akshay,11000,3)  
(104,Anubhav,5000,4)  
(105,Pawan,2500,5)  
(106,Aamir,25000,1)  
(107,Salman,17500,2)  
(108,Ranbir,14000,3)  
(109,Katrina,1000,4)  
(110,Priyanka,2000,5)  
(111,Tushar,500,1)  
(112,Ajay,5000,2)  
(113,Jubeen,1000,1)  
(114,Madhuri,2000,2)
```

Use command FOREACH emp\_details GENERATE (EmpID,Name,Salary,DepartmentID)

```
(114,Madhuri,2000,2)  
grunt> employee_details = FOREACH emp_details generate (EmpID,Name,Salary,DepartmentID); █
```

```

2018-09-04 01:55:26,531 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 01:55:26,902 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 01:55:26,903 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((101,Amitabh,20000,1))
((102,Shahrukh,10000,2))
((103,Akshay,11000,3))
((104,Anubhav,5000,4))
((105,Pawan,2500,5))
((106,Aamir,25000,1))
((107,Salman,17500,2))
((108,Ranbir,14000,3))
((109,Katrina,1000,4))
((110,Priyanka,2000,5))
((111,Tushar,500,1))
((112,Ajay,5000,2))
((113,Jubeen,1000,1))
((114,Madhuri,2000,2))

```

Loading employee\_expenses

```

-----
grunt> emp_expenses = load 'employee_expenses.txt' using PigStorage(',') as (EmpID,expense);
2018-09-04 02:07:37,446 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-09-04 02:07:37,448 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
acadmild@localhost:~

```

Use command FOREACH emp\_expenses GENERATE (EmpID,expense)

```

d, use fs.defaultFS
grunt> emp_exp = FOREACH emp_expenses generate (EmpID,expense);

```

Dump emp\_exp to see the content

```

2018-09-04 02:12:08,533 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 02:12:08,533 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((101 200,))
((102 100,))
((110 400,))
((114 200,))
((119 200,))
((105 100,))
((101 100,))
((104 300,))
((102 400,))

```

a)renamed departmentid to rating

```

((104 300,))
grunt> emp_with_rank = load 'employee_details.txt' using PigStorage(',') as (EmpID:int,Name:chararray,Salary:int,rating:int);
2018-09-04 02:48:36,218 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-09-04 02:48:36,218 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>

```

Use command order emp\_ranked = emp\_with\_rank by rating, Name.dump emp\_ranked

```

2018-09-04 02:59:32,566 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
cess : 1
(106,Aamir,25000,1)
(101,Amitabh,20000,1)
(113,Jubeen,1000,1)
(111,Tushar,500,1)
(114,Madhuri,2000,2)
(112,Ajay,5000,2)
(102,Shahrukh,10000,2)
(107,Salman,17500,2)
(103,Akshay,11000,3)
(108,Ranbir,14000,3)
(104,Anubhav,5000,4)
(109,Katrina,1000,4)
(110,Priyanka,2000,5)
(105,Pawan,2500,5)

```

Use command `emp_five=limit emp_ranked 5; dump emp_five;`

```

grunt> emp_five = limit emp_ranked 5;
grunt> dump emp_five;

```

```

e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,203 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,237 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,240 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,245 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,259 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,264 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,269 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,301 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,307 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,323 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:21:30,347 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-09-04 03:21:30,348 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-09-04 03:21:30,356 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-09-04 03:21:30,356 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 03:21:30,631 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 03:21:30,632 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(106,Aamir,25000,1)
(101,Amitabh,20000,1)
(113,Jubeen,1000,1)
(111,Tushar,500,1)
(112,Ajay,5000,2)

```

b)

```

(107,Salman,17500,2)
(101,Amitabh,20000,1)
grunt> emp_ordered_by_sal = order emp_with_rank by Salary,Name;
grunt> emp_odd = FILTER emp_ordered_by_sal by EmpID%2==1;
grunt> dump emp_odd;

```

```

e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,132 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,150 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,168 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,199 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,223 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,247 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,253 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-09-04 03:53:16,281 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-09-04 03:53:16,281 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-09-04 03:53:16,282 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-09-04 03:53:16,284 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 03:53:16,499 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 03:53:16,500 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(111,Tushar,500,1)
(113,Jubeen,1000,1)
(109,Katrina,1000,4)
(105,Pawan,2500,5)
(103,Akshay,11000,3)
(107,Salman,17500,2)
(101,Amitabh,20000,1)
grunt> █

```

**(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)**

**Ans)** grunt>employee\_expenses = LOAD 'hadoop/empLOYEE\_expenses.txt' USING PigStorage('\t') AS (EmpID:int,Expense:int);

```

(101,200)
(102,100)
(110,400)
(114,200)
(119,200)
(105,100)
(101,100)
(104,300)
(102,400)

```

grunt> new\_table = JOIN employee\_details BY EmpID, employee\_expenses BY EmpID;

```

(101,Amitabh,20000,1,101,100)
(101,Amitabh,20000,1,101,200)
(102,Shahrukh,10000,2,102,400)
(102,Shahrukh,10000,2,102,100)
(104,Anubhav,5000,4,104,300)
(105,Pawan,2500,5,105,100)
(110,Priyanka,2000,5,110,400)
(114,Madhuri,2000,2,114,200)

```

grunt>maxexpen\_employees = ORDER new\_table BY Expense DESC;

```

(110,Priyanka,2000,5,110,400)
(102,Shahrukh,10000,2,102,400)
(104,Anubhav,5000,4,104,300)

```

```
(114,Madhuri,2000,2,114,200)
(101,Amitabh,20000,1,101,200)
(105,Pawan,2500,5,105,100)
(102,Shahrukh,10000,2,102,100)
(101,Amitabh,20000,1,101,100)
```

```
grunt> final_output = FOREACH maxexpen_employees GENERATE employee_expenses::EmpID as
EmpID,employee_details::Name as Name;
```

```
(110,Priyanka)
(102,Shahrukh)
(104,Anubhav)
(114,Madhuri)
(101,Amitabh)
(105,Pawan)
(102,Shahrukh)
(101,Amitabh)
```

```
grunt> employee_details = LOAD 'hadoop/employee_details.txt' USING PigStorage(',') AS (EmpID:int,Name:chararray,Salary:int,Ra
ting:int);
2018-08-30 13:05:57,005 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> employee_expenses = LOAD 'hadoop/employee_expenses.txt' USING PigStorage('\t') AS (EmpID:int,Expense:int);
2018-08-30 13:06:27,501 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> new_table = JOIN employee_details BY EmpID, employee_expenses BY EmpID;
grunt> maxexpen_employees = ORDER new_table BY Expense DESC;
grunt> final_output = FOREACH maxexpen_employees GENERATE employee_expenses::EmpID as EmpID,employee_details::Name as Name;
grunt> █
```

**(d) List of employees (employee id and employee name) having entries in employee\_expenses file.**

```
Ans) grunt>new_relation = JOIN employee_details BY EmpID LEFT OUTER, employee_expenses BY
EmpID;
```

```
(101,Amitabh,20000,1,101,100)
(101,Amitabh,20000,1,101,200)
(102,Shahrukh,10000,2,102,400)
(102,Shahrukh,10000,2,102,100)
(103,Akshay,11000,3,,)
(104,Anubhav,5000,4,104,300)
(105,Pawan,2500,5,105,100)
(106,Aamir,25000,1,,)
(107,Salman,17500,2,,)
(108,Ranbir,14000,3,,)
(109,Katrina,1000,4,,)
(110,Priyanka,2000,5,110,400)
(111,Tushar,500,1,,)
(112,Ajay,5000,2,,)
(113,Jubeen,1000,1,,)
(114,Madhuri,2000,2,114,200)
```

```
grunt> final_result = FILTER new_relation BY employee_expenses::Expense is not null;
```

```
(101,Amitabh,20000,1,101,100)
(101,Amitabh,20000,1,101,200)
(102,Shahrukh,10000,2,102,400)
(102,Shahrukh,10000,2,102,100)
(104,Anubhav,5000,4,104,300)
(105,Pawan,2500,5,105,100)
(110,Priyanka,2000,5,110,400)
(114,Madhuri,2000,2,114,200)
```

```
2018-08-30 13:43:41,194 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh,20000,1,101,100)
(101,Amitabh,20000,1,101,200)
(102,Shahrukh,10000,2,102,400)
(102,Shahrukh,10000,2,102,100)
(104,Anubhav,5000,4,104,300)
(105,Pawan,2500,5,105,100)
(110,Priyanka,2000,5,110,400)
(114,Madhuri,2000,2,114,200)
grunt> █
```

**(e) List of employees (employee id and employee name) having no entry in employee\_expenses file.**

**Ans)** grunt> final\_result1 = FILTER new\_relation BY employee\_expenses::Expense is null;

```
(103,Akshay,11000,3,,)
(106,Aamir,25000,1,,)
(107,Salman,17500,2,,)
(108,Ranbir,14000,3,,)
(109,Katrina,1000,4,,)
(111,Tushar,500,1,,)
(112,Ajay,5000,2,,)
(113,Jubeen,1000,1,,)
```

```
2018-08-30 13:46:30,070 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Akshay,11000,3,,)
(106,Aamir,25000,1,,)
(107,Salman,17500,2,,)
(108,Ranbir,14000,3,,)
(109,Katrina,1000,4,,)
(111,Tushar,500,1,,)
(112,Ajay,5000,2,,)
(113,Jubeen,1000,1,,)
grunt> █
```

**Task 1.3:**

Show databases; command displays all the databases available



```

Time taken: 0.154 seconds, Fetched: 1 row(s)
hive> show databases;
OK
default
Time taken: 0.154 seconds, Fetched: 1 row(s)
hive> █

```

Task2.1) Create database custom ; creates database with name custom. use custom; command uses custom database

```

hive> show databases;
OK
default
Time taken: 0.154 seconds, Fetched: 1 row(s)
hive> create database custom;
OK
Time taken: 1.616 seconds
hive> show databases;
OK
custom
default
Time taken: 0.383 seconds, Fetched: 2 row(s)
hive> use custom;
OK
Time taken: 0.105 seconds
hive> █

```

Create table command creates table with specified fields

```

hive> create table temperature_data
> (
>   full_date string,
>   zip_code int,
>   temperature int
> )
> row format delimited
> fields terminated by ',';
OK
Time taken: 1.53 seconds █
hive> show tables;
OK
temperature_data
Time taken: 0.431 seconds, Fetched: 1 row(s)
hive> █

```

```

File Edit View Search Terminal Help
[acadgild@localhost ~]$ cat temperature_dataset.txt
10-01-1990,123112,10
14-02-1991,283901,11
10-03-1990,381920,15
10-01-1991,302918,22
12-02-1990,384902,9
10-01-1991,123112,11
14-02-1990,283901,12
10-03-1991,381920,16
10-01-1990,302918,23
12-02-1991,384902,10
10-01-1993,123112,11
14-02-1994,283901,12
10-03-1993,381920,16
10-01-1994,302918,23
12-02-1991,384902,10
10-01-1991,123112,11
14-02-1990,283901,12
10-03-1991,381920,16
10-01-1990,302918,23
12-02-1991,384902,10[acadgild@localhost ~]$ █

```

```

hive> load data local inpath 'temperature_dataset.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 3.755 seconds
hive> select * from temperature_data;
OK
10-01-1990      123112  10
14-02-1991      283901  11
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902   9
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
10-01-1993      123112  11
14-02-1994      283901  12
10-03-1993      381920  16
10-01-1994      302918  23
12-02-1991      384902  10
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 6.231 seconds, Fetched: 20 row(s)
. █

```

Data for zipcode greater than 300000 and zipcode less than 399999

```

Time taken: 0.231 seconds, Fetched: 20 row(s)
hive> select full_date,temperature from temperature_data where zip_code>300000 and zip_code<399999;
OK
10-03-1990      15
10-01-1991      22
12-02-1990       9
10-03-1991      16
10-01-1990      23
12-02-1991      10
10-03-1993      16
10-01-1994      23
12-02-1991      10
10-03-1991      16
10-01-1990      23
12-02-1991      10
Time taken: 2.466 seconds, Fetched: 12 row(s)
hive> █

```

## Maximum temperature corresponding to every year

```

hive> select year,MAX(t1.temperature) as temperature from (select substring(full_date,7,4) year,temperature from temperature_
data) t1 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 execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180904062900_382f5d73-9ae8-4011-bf5b-ba2cee5f4b38
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_1535993336114_0006, Tracking URL = http://localhost:8088/proxy/application_1535993336114_0006/
Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1535993336114_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-04 06:29:46,058 Stage-1 map = 0%, reduce = 0%
2018-09-04 06:30:17,424 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 7.72 sec
2018-09-04 06:30:46,797 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 11.7 sec
2018-09-04 06:30:49,737 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 12.91 sec
MapReduce Total cumulative CPU time: 12 seconds 910 msec
Ended Job = job_1535993336114_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 12.91 sec HDFS Read: 9348 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 910 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 112.405 seconds, Fetched: 4 row(s)

```

## Maximum temperature those years who have more than 2 entries

```

Time taken: 112.405 seconds, Fetched: 4 row(s)
hive> select year,MAX(t1.temperature) as temperature from (select substring(full_date,7,4) year,temperature from temperature_
data) t1 group by year having count(t1.year)>2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180904063317_0090765b-1d26-4b63-a01e-ac50abb9acae
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_1535993336114_0007, Tracking URL = http://localhost:8088/proxy/application_1535993336114_0007/
Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1535993336114_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-04 06:33:59,034 Stage-1 map = 0%, reduce = 0%
2018-09-04 06:34:27,110 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.75 sec
2018-09-04 06:34:55,098 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.36 sec
2018-09-04 06:35:01,774 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 14.34 sec
MapReduce Total cumulative CPU time: 14 seconds 340 msec
Ended Job = job_1535993336114_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 14.34 sec HDFS Read: 10511 HDFS Write: 127 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 340 msec
OK
1990 23
1991 22
Time taken: 107.393 seconds, Fetched: 2 row(s)

```

## Creating view with previous command data

```

: full_date, zip_code, temperature)
hive> create view temperature_data_vw as select year,MAX(t1.temperature) as temperature from (select substring(full_date,7,4)
year,temperature from temperature_data) t1 group by year having count(t1.year)>2;
OK
Time taken: 0.687 seconds
hive select * from temperature_data_vw

```

```

hive> select * from temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180904063945_5978822e-da69-4469-96e4-28fb61792ca9
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_1535993336114_0008, Tracking URL = http://localhost:8088/proxy/application_1535993336114_0008/
Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1535993336114_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-04 06:40:28,151 Stage-1 map = 0%, reduce = 0%
2018-09-04 06:40:59,290 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.83 sec
2018-09-04 06:41:33,436 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.99 sec
2018-09-04 06:41:37,504 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.67 sec
MapReduce Total cumulative CPU time: 13 seconds 670 msec
Ended Job = job_1535993336114_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 13.67 sec HDFS Read: 10566 HDFS Write: 127 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 670 msec
OK
1990 23
1991 22
Time taken: 114.721 seconds, Fetched: 2 row(s)

```

## Exporting data from view to local file with | delimited

```

12-02-1991,384902,10[acadgild@localhost ~]$ mkdir hivedata
[acadgild@localhost ~]$ ls
apache-flume-1.6.0-bin          jdk-8u101-linux-i586.tar.gz  pig_1536004931982.log
apache-flume-1.6.0-bin.tar.gz  jhg                           pig_1536005533611.log
apache-hive-2.1.0-bin          max-temp.txt                  pig_1536005723779.log
apache-hive-2.1.0-bin.tar.gz  max-temp.txt                  pig_1536013112458.log
derby.log                     metastore_db                   pig_1536013197361.log
Desktop                       metastore_db.tmp              Public
Documents                     Music                          Softwares
Downloads                     orderBySal                    sqoop-1.4.6.bin__hadoop-2.0.4-alpha
eclipse                       orderBySal.pig                sqoop-1.4.6.bin__hadoop-2.0.4-alpha.tar.gz
employee_details.txt           Pictures                       temperature_dataset.txt
employee_expenses.txt         pig-0.16.0                    Templates
employee.java                 pig-0.16.0.tar.gz             test1.txt
hadoop                        pig_1470979104717.log         testappend.txt
hadoop-2.7.2                  pig_1471462105724.log        test.txt
hadoop-2.7.2.tar.gz           pig_1521175291666.log        Videos
hbase-1.0.3                   pig_1521175425511.log        wordcount.pig
hbase-1.0.3-bin.tar.gz        pig_1535996536994.log        Wordcount.pig
hivedata                     pig_1535999974342.log        workspace
jdk1.8.0_101                  pig_1536000427628.log
[acadgild@localhost ~]$ ls hivedata
[acadgild@localhost ~]$

```

```

hive> insert overwrite local directory '/home/acadgild/hivedata/output' row format delimited
> fields terminated by '|'
> select * from temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180904064845_bfb4e000-2ffd-440a-94e3-dd18857438cf
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_1535993336114_0009, Tracking URL = http://localhost:8088/proxy/application_1535993336114_0009/
Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1535993336114_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-04 06:49:32,434 Stage-1 map = 0%, reduce = 0%
2018-09-04 06:50:00,637 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.87 sec
2018-09-04 06:50:28,358 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.97 sec
2018-09-04 06:50:33,766 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.92 sec
MapReduce Total cumulative CPU time: 13 seconds 920 msec
Ended Job = job_1535993336114_0009
Moving data to local directory /home/acadgild/hivedata/output
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 13.92 sec HDFS Read: 10271 HDFS Write: 16 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 920 msec
OK
Time taken: 111.359 seconds

```

```
[acadgild@localhost ~]$ cd hivedata
[acadgild@localhost hivedata]$ cd output
[acadgild@localhost output]$ ls
000000_0
[acadgild@localhost output]$ cat 000000_0
1990|23
1991|22
[acadgild@localhost output]$ █
```