a. Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)

Below is the script used to derive top 5 employees with highest rating(names ordered in alphabetic manner)-

```
Pig -x local;
```

emp\_details = LOAD '/home/acadgild/pig/employee\_details.txt' USING PigStorage(',') AS (emp\_id:int, name:chararray, salary:int, rating:int);

```
grpd = GROUP emp_details BY rating;
top5 = FOREACH grpd {
    sorted = ORDER emp_details by name ASC;
    top1 = limit sorted 1;
    generate FLATTEN(top1);
```

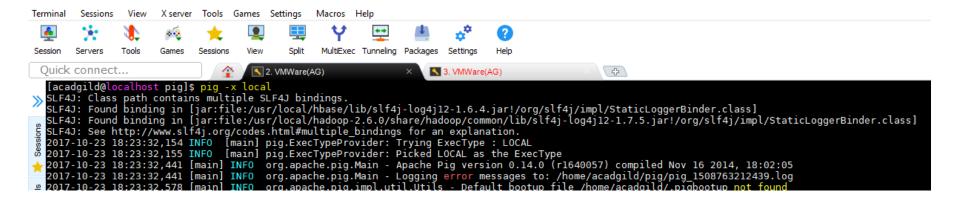
grpd2 = FOREACH top5 GENERATE emp\_id, name;

### dump grpd2;

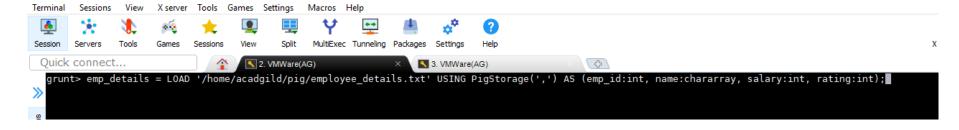
**}**;

Each and every relation has been explained below with its immediate corresponding output-

Start pig in local mode-



Load the emp\_details.txt file using PigStorage



Contents of file-

```
2017-10-23 18:33:26,020 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 18:33:26,085 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - 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) grunt>
```

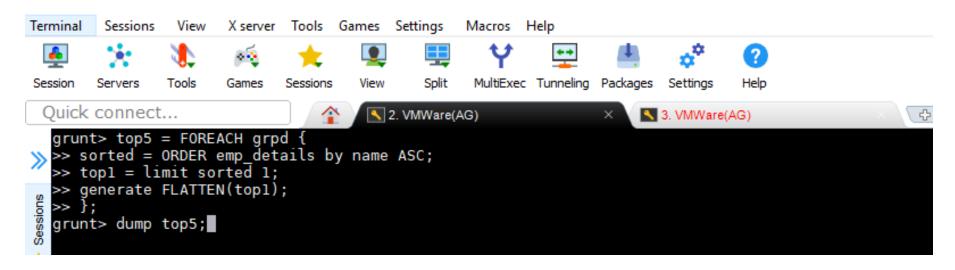
#### Group the data by rating-



### Result of GROUP-

```
2017-10-23 20:52:57,046 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 20:52:57,105 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 20:52:57,105 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (1,{(106,Aamir,25000,1),(101,Amitabh,20000,1),(113,Jubeen,1000,1),(111,Tushar,500,1)}) (2,{(114,Madhuri,2000,2),(112,Ajay,5000,2),(102,Shahrukh,10000,2),(107,Salman,17500,2)}) (3,{(103,Akshay,11000,3),(108,Ranbir,14000,3)}) (4,{(104,Anubhay,5000,4),(109,Katrina,1000,4)}) (5,{(110,Priyanka,2000,5),(105,Pawan,2500,5)})
```

Below script takes out the tuple from the bag arranged with name in alphabetic order-



## Dump results-

```
2017-10-23 20:55:47,617 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 20:55:47,716 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 20:55:47,716 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-10-23 20:55:47,
(106,Aamir,25000,1)
(112,Ajay,5000,2)
(103,Akshay,11000,3)
(104,Anubhav,5000,4)
(105,Pawan,2500,5)
grunt>
```

Extract emp\_id and name as per question-

```
++
                                                                                           a
                                          · •
Session
        Servers
                 Tools
                         Games
                                Sessions
                                                  Split
                                                        MultiExec Tunneling Packages Settings
 Quick connect...
                                                                                                               C C
                                          2. VMWare(AG)
                                                                            3. VMWare(AG)
  grunt> grpd2 = FOREACH top5 GENERATE emp_id, name;
>> grunt> dump grp2;
```

Final O/P-

```
2017-10-23 20:57:19,863 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 20:57:19,927 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 20:57:19,927 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (106,Aamir) (112,Ajay) (103,Akshay) (104,Anubhav) (105,Pawan) grunt> ■
```

b. Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference)

Below is the script used to finalize the result-

Pig -x local;

emp\_details = LOAD '/home/acadgild/pig/employee\_details.txt' USING PigStorage(',') AS (emp\_id:int, name:chararray, salary:int, rating:int);

odd\_emp = FILTER emp\_details BY emp\_id % 2 != 0;

order\_sal = ORDER odd\_emp BY salary DESC;

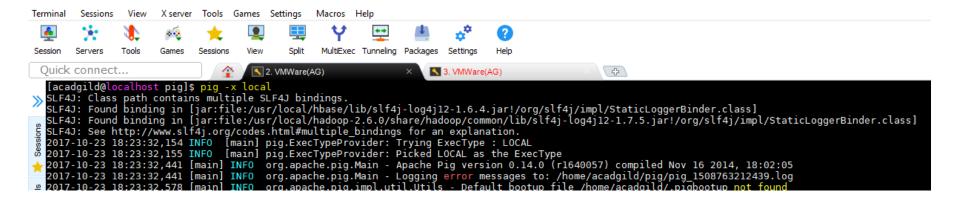
top3 = LIMIT order\_sal 3;

final\_res = FOREACH top3 GENERATE (emp\_id, name);

dump final\_res;

Each and every relation has been explained below with its immediate corresponding output-

Start pig in local mode-



Load the emp\_details.txt file using PigStorage

```
Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

Quick connect...

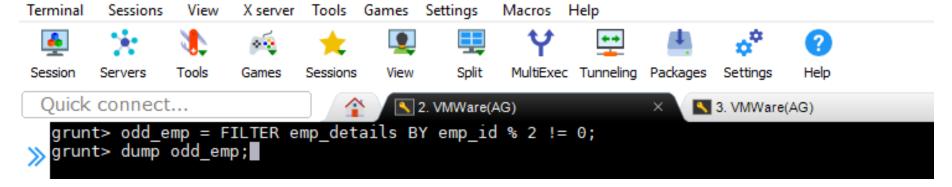
Quick connect...

The property of the property
```

### Contents of file-

```
2017-10-23 18:33:26,020 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 (101,Amitabh,20000,1) (102,Shahrukh,10000,2) (103,Akshay,11000,3) (104,Anubhay,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) (113,Jubeen,1000,1) (114,Madhuri,2000,2) grunt> ■
```

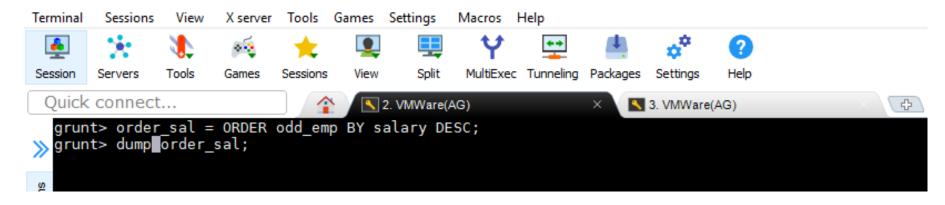
Filtering out the employees with odd emp id-



Records with odd emp ids-

```
2017-10-23 21:38:06,417 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 21:38:06,432 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 21:38:06,432 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (101,Amitabh,20000,1) (103,Akshay,11000,3) (105,Pawan,2500,5) (107,Salman,17500,2) (109,Katrina,1000,4) (111,Tushar,500,1) (113,Jubeen,1000,1) grunt> ■
```

Ordering the salary in descending manner-



Records with descending order of salary-

```
2017-10-23 21:40:08,864 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 21:40:08,889 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 (101,Amitabh,20000,1) (107,Salman,17500,2) (103,Akshay,11000,3) (105,Pawan,2500,5) (113,Jubeen,1000,1) (109,Katrina,1000,4) (111,Tushar,500,1) grunt> ■
```

Extracting top 3 rows for top 3 employees with odd emp\_id-

```
Terminal
                    View
                            X server Tools Games Settings
                                                             Macros Help
          Sessions
Session
         Servers
                           Games
                                                       Split
                                                              MultiExec Tunneling Packages
                   Tools
                                    Sessions
                                              View
 Quick connect...
                                                2. VMWare(AG)
                                                                                      3. VMWare(AG)
   grunt> top3 = LIMIT order_sal 3;

>> grunt> dump top3;

Sessions
```

```
2017-10-23 21:42:05,323 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 21:42:05,363 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 21:42:05,363 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (101,Amitabh,20000,1) (107,Salman,17500,2) (103,Akshay,11000,3) grunt> ■
```

Generating emp\_id and emp\_name as per question-

```
Terminal
         Sessions
                   View
                         X server
                                 Tools
                                        Games
                                              Settings
                                                         Macros
                                                                Help
Session
                         Games
                                                          MultiExec Tunneling Packages
 Quick connect...
                                            2. VMWare(AG)
                                                                                                                  (4)
                                                                               3. VMWare(AG)
   grunt> final_res = FOREACH top3 GENERATE (emp_id, name);

    grunt> dump final_res;
```

## Final output-

```
2017-10-23 21:46:25,207 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 21:46:25,236 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 21:46:25,236 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 ((101,Amitabh)) ((107,Salman)) ((103,Akshay)) grunt>
```

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

Below is the script used to finalize the result-

Pig -x local;

emp\_details = LOAD '/home/acadgild/pig/employee\_details.txt' USING PigStorage(',') AS (emp\_id:int, name:chararray, salary:int, rating:int);

emp\_expense = LOAD '/home/acadgild/pig/employee\_expenses.txt' USING PigStorage('\t') AS (emp\_id:int, expense:int);

joined\_data = JOIN emp\_details by emp\_id, emp\_expense by emp\_id;

ordered\_data = ORDER joined\_data BY expense DESC, name ASC;

extract\_id\_name = FOREACH ordered\_data GENERATE \$0 AS emp\_id, \$1 AS name;

Final\_result = LIMIT extract\_id\_name 1;

dump Final\_result;

Each and every relation has been explained below with its immediate corresponding output-

Start pig in local mode-

```
Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Spit MultiExec Tunneling Packages Settings Help

Quick connect...

[acadgild@localhost pig]s pig -x local

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]

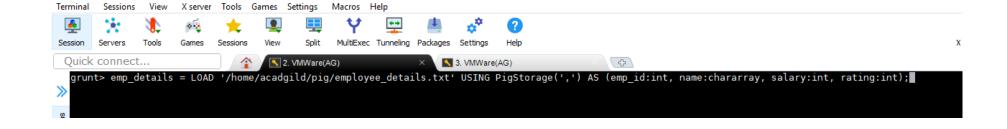
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.

SOUTY-10-23 18:23:32,154 INFO [main] pig.ExecTypeProvider: Trying ExecType: LOCAL

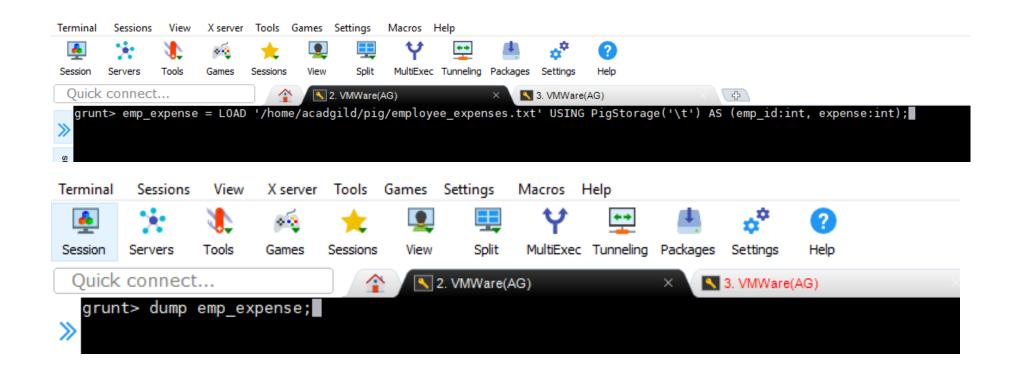
2017-10-23 18:23:32,441 [main] INFO org.apache.pig.Main - Apache Pig version 0.14.0 (r1640057) compiled Nov 16 2014, 18:02:05
2017-10-23 18:23:32,441 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/pig/pig_1508763212439.log
2017-10-23 18:23:32.578 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pig/bootup not found
```



### Contents of file-

```
org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized org.apache.hadoop.mapreduce.lib.input.FileInputFormat - 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,2500,0) (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) grunt>
```

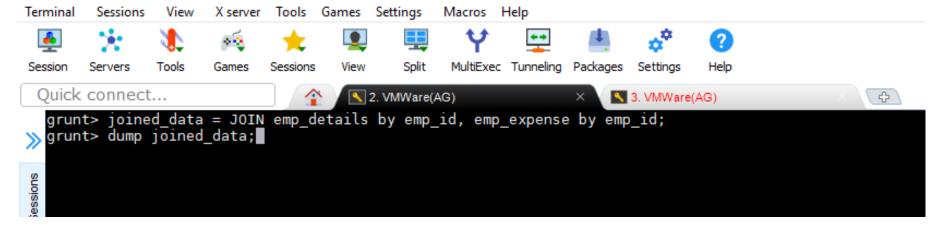
Load the emp\_expense.txt file using PigStorage



Contents of file-

```
2017-10-23 22:33:51,736 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 22:33:51,808 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 22:33:51,811 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (101,200) (102,100) (114,200) (119,200) (119,200) (105,100) (101,100) (101,100) (101,100) (101,100) (102,400) grunt>
```

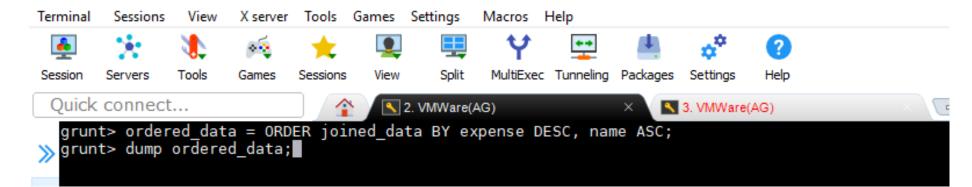
Join both files using JOIN relation-



#### Result after JOIN

```
2017-10-23 22:45:06,477 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 22:45:06,541 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 22:45:06,542 [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,10000,2,102,400) (104,Anubhav,5000,4,104,300) (105,Pawan,2500,5,105,100) (106,Priyanka,2000,5,110,400) (114,Madhuri,2000,2,114,200) grunt> ■
```

ORDER the JOIN output with expense in Descending order and name in Ascending order to take the name of employee in alphabetic manner with highest expense-



### Result

```
2017-10-23 23:52:58,966 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 23:52:59,020 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 23:52:59,020 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (110,Priyanka,2000,5,110,400) (102,Shahrukh,10000,2,102,400) (104,Anubhav,5000,4,104,300) (101,Amitabh,20000,1,101,200) (114,Madhuri,2000,2,114,200) (101,Amitabh,20000,1,101,100) (105,Pawan,2500,5,105,100) (102,Shahrukh,10000,2,102,100) grunt> ■
```

Extract only emp\_id and emp\_name as required -

```
Terminal
         Sessions
                  View
                         X server Tools Games Settings
                                                        Macros Help
                                          <u>.</u>
                                                                   ++
                                                                                            0
                          ٠.٠
Session
        Servers
                 Tools
                         Games
                                Sessions
                                                   Split
                                                         MultiExec Tunneling Packages Settings
                                                                                            Help
 Quick connect...
                                           2. VMWare(AG)
                                                                             3. VMWare(AG)
   grunt> extract_id_name = FOREACH ordered_data GENERATE $0 AS emp_id, $1 AS name;
>> grunt> dump extract_id_name;
```

```
2017-10-24 15:16:55,855 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-24 15:16:55,887 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-24 15:16:55,887 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (110,Priyanka) (102,Shahrukh) (104,Anubhav) (101,Amitabh) (114,Madhuri) (101,Amitabh) (114,Madhuri) (105,Pawan) (105,Pawan) (105,Shahrukh)
```

Select the topmost row using LIMIT-

```
Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

Quick connect...

Quick connect...

Quick connect...

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

Quick connect...

The property of the property
```

## Final O/P

```
2017-10-24 15:22:14,238 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-24 15:22:14,265 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-24 15:22:14,265 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
grunt>
```

# d. <u>List of employees (employee id and employee name) having entries in employee expenses</u> file.

Below is the script used to finalize the result-

Pig -x local;

emp\_details = LOAD '/home/acadgild/pig/employee\_details.txt' USING PigStorage(',') AS (emp\_id:int, name:chararray, salary:int, rating:int);

emp\_expense = LOAD '/home/acadgild/pig/employee\_expenses.txt' USING PigStorage('\t') AS (emp\_id:int, expense:int);

joined\_data = JOIN emp\_details by emp\_id, emp\_expense by emp\_id;

extract\_id\_name = FOREACH joined\_data GENERATE \$0 AS emp\_id, \$1 AS name;

Final\_result = DISTINCT extract\_id\_name;

dump Final\_result;

Each and every relation has been explained below with its immediate corresponding output-

Start pig in local mode-

```
Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

Quick connect...

[acadgild@localhost pig]$ pig -x local

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]

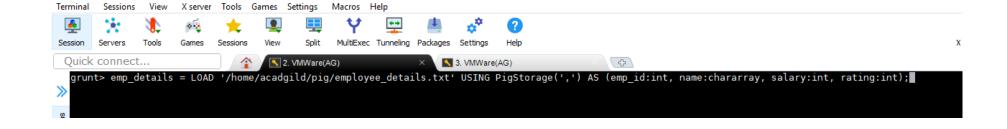
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.

2017-10-23 18:23:32,155 INFO [main] pig.ExecTypeProvider: Trying ExecType: LOCAL

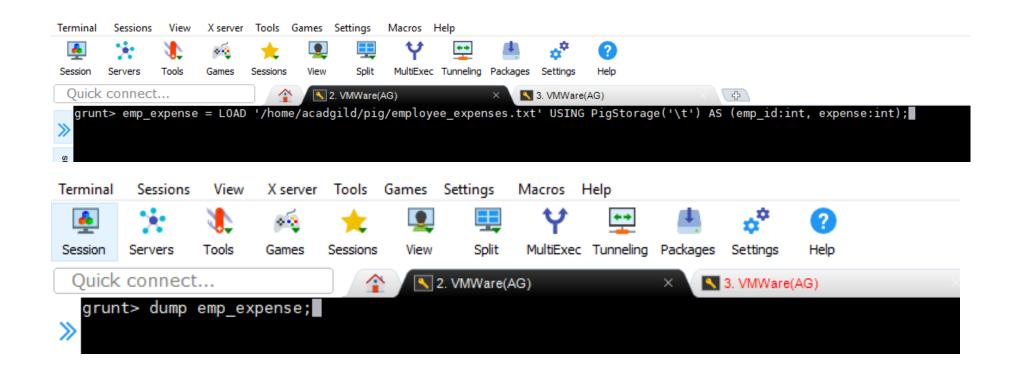
2017-10-23 18:23:32,441 [main] INFO org.apache.pig.Main - Apache Pig version 0.14.0 (r1640057) compiled Nov 16 2014, 18:02:05
2017-10-23 18:23:32,441 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/pig/pig_1508763212439.log
2017-10-23 18:23:32.578 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/.pig/pig_1508763212439.log
```



### Contents of file-

```
org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized org.apache.hadoop.mapreduce.lib.input.FileInputFormat - 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,2500,0) (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) grunt>
```

Load the emp\_expense.txt file using PigStorage



Contents of file-

```
2017-10-23 22:33:51,736 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 22:33:51,808 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 22:33:51,811 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (101,200) (102,100) (114,200) (119,200) (119,200) (105,100) (101,100) (101,100) (101,100) (101,100) (102,400) grunt>
```

Join both files using JOIN relation-

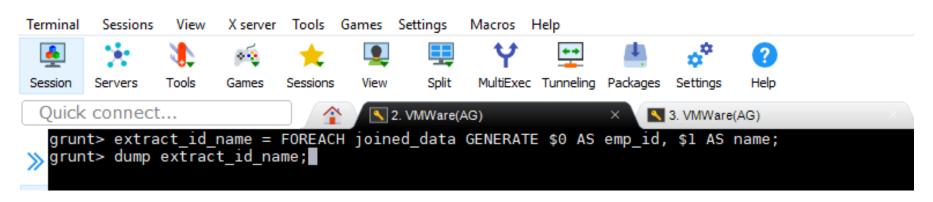
```
X server Tools Games Settings
                                                         Macros Help
Terminal
         Sessions
                   View
                                                                                              8
                                 Sessions
Session
        Servers
                  Tools
                          Games
                                           View
                                                    Split
                                                          MultiExec Tunneling Packages Settings
 Quick connect...
                                                                                                                   ( th
                                            2. VMWare(AG)
                                                                                3. VMWare(AG)
   grunt> joined_data = JOIN emp_details by emp_id, emp_expense by emp_id;

>> grunt> dump joined_data;
```

### Result after JOIN

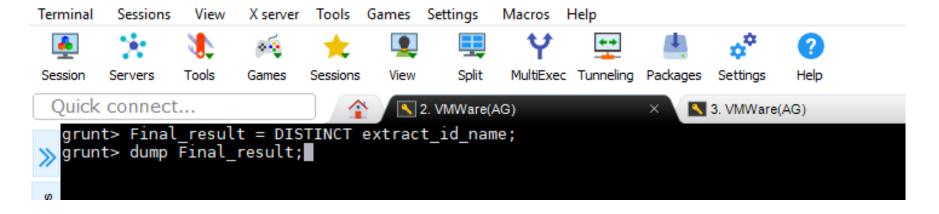
```
2017-10-23 22:45:06,477 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 22:45:06,541 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-23 22:45:06,542 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to proc (101,Amitabh,20000,1,101,200) (102,Shahrukh,10000,2,102,400) (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>
                                                                                                                  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
 grunt>
```

### Extract EMP\_ID and NAME as required-



```
2017-10-24 16:05:24,992 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(101,Amitabh)
(102,Shahrukh)
(102,Shahrukh)
(104, Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt>
```

# Take DISTINCT EMP\_IDs -



## FINAL O/P-

```
2017-10-24 16:06:33,764 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-24 16:06:33,798 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-24 16:06:33,798 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101, Amitabh)
(102, Shahrukh)
(102,5Nahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt>
```

# e. <u>List of employees (employee id and employee name) having no entry in employee expenses</u> file.

Below is the script used to finalize the result-

Pig -x local;

emp\_details = LOAD '/home/acadgild/pig/employee\_details.txt' USING PigStorage(',') AS (emp\_id:int, name:chararray, salary:int, rating:int);

emp\_expense = LOAD '/home/acadgild/pig/employee\_expenses.txt' USING PigStorage('\t') AS (emp\_id:int, expense:int);

joined\_data = JOIN emp\_details by emp\_id FULL, emp\_expense by emp\_id;

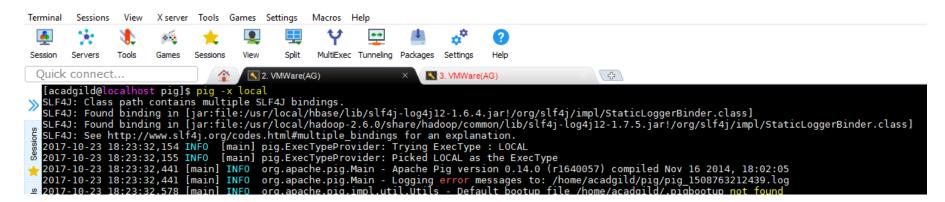
filter\_res = FILTER joined\_data BY emp\_expense::emp\_id IS NULL;

final\_result = FOREACH filter\_res GENERATE \$0, \$1;

dump final\_result;

Each and every relation has been explained below with its immediate corresponding output-

Start pig in local mode-



Load the emp\_details.txt file using PigStorage

```
Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

Quick connect...

Quick connect...

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

Quick connect...

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

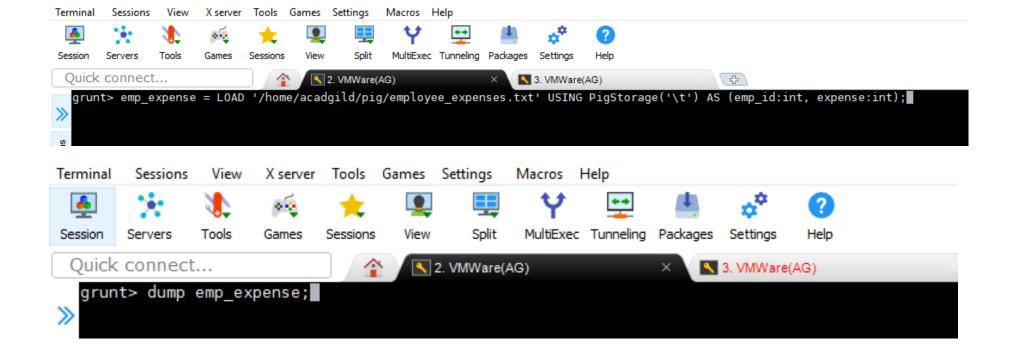
| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Help | X

| Session | Server | Tools | Games | Sessions | View | Split | MultiExec Tunneling | Packages | Settings | Packages | View | Split | View |
```

### Contents of file-

```
2017-10-23 18:33:26,020 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-23 18:33:26,085 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - 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) grunt>
```

Load the emp expense.txt file using PigStorage



Contents of file-

```
2017-10-23 22:33:51,736 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-23 22:33:51,808 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-23 22:33:51,811 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,200)
(102,100)
(114,200)
(114,200)
(119,200)
(105,100)
(101,100)
(101,300)
(104,300)
(102,400)
grunt> ■
```

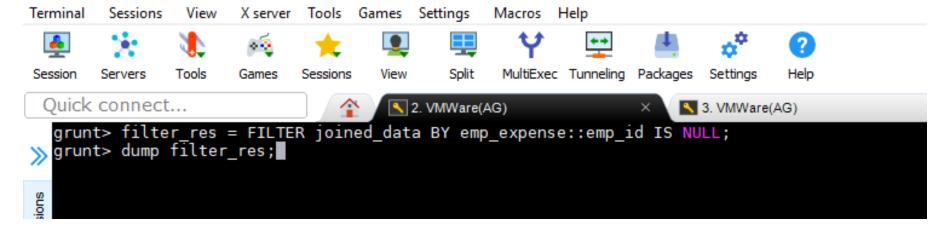
Do a FULL OUTER join on both files -

```
Terminal
                   View
                         X server Tools Games Settings
                                                        Macros Help
         Sessions
                                                                                              ?
                          ***
                                                   Split
Session
        Servers
                  Tools
                         Games
                                 Sessions
                                           View
                                                          MultiExec Tunneling Packages Settings
 Quick connect...
                                            2. VMWare(AG)
                                                                               3. VMWare(AG)
  grunt> joined_data = JOIN emp_details BY emp_id FULL, emp_expense by emp_id;
  grunt> dump joined_data;
```

Result of FULL OUTER JOIN-

```
| 13.mmx | 1
```

Check whether contents(emp\_id) from second file(emp\_expense) is NULL and FILTER them-



Result of Filtered out data which are not present in emp expense file-

```
rs.max
2017-10-24 16:42:41,996 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-24 16:42:42,026 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-24 16:42:42,026 [main] INFO org.apache.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,,)
```

Extract emp\_id aand emp\_name from first file as required

```
Terminal
          Sessions
                   View
                          X server Tools Games Settings
                                                           Macros Help
                                                                                                  a
 <u>...</u>
                            * 😍
Session
                                             View
                                                      Split
                                                            MultiExec Tunneling Packages Settings
                                                                                                 Help
         Servers
                  Tools
                          Games
                                  Sessions
 Quick connect...
                                              2. VMWare(AG)
                                                                                                                       (4)
                                                                                   3. VMWare(AG)
   grunt> final_result = FOREACH filter_res GENERATE $0, $1;

>> grunt> dump final_result;

ions
```

## Final O/P-

```
2017-10-24 16:45:21,036 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized 2017-10-24 16:45:21,062 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1 2017-10-24 16:45:21,062 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (103,Akshay) (106,Aamir) (107,Salman) (108,Ranbir) (109,Katrina) (111,Tushar) (112,Ajay) (113,Jubeen) grunt> ■
```