Pig Assignment 5.1

(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)

```
emp_details = load '/home/vignesh/Desktop/PIG_EXAMPLES/employee_details.txt' using PigStorage(',')  as (id:int, name:chararray, salary:int, rating:int);

ordered_details = ORDER emp_details by rating DESC, name ASC;

limited_ordered_details = LIMIT ordered_details 5;

final = FOREACH limited_ordered_details GENERATE  id,name, rating;

dump final;
```

(105,Pawan,5)
(110,Priyanka,5)
(104,Anubhav,4)
(109,Katrina,4)
(103,Akshay,3)

```
grunt> emp_details = load '/home/vignesh/Desktop/PIG_EXAMPLES/employee_details.txt' using PigStorage(',') as (id:int,name:chararray,salary:int,rating
:int);
2017-12-13 16:43:05,702 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2017-12-13 16:43:05,703 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> ordered_details = ORDER emp_details by rating DESC, name ASC;
grunt> limited_ordered_details = LIMIT ordered_details 5;
grunt> final = FOREACH limited_ordered_details GENERATE id,name,rating;
grunt> dump final;
2017-12-13 16:52:36,902 [main]
2017-12-13 16:54:01,510 [main]
2017-12-13 16:54:01,511 [main]
(105,Pawan,5)
(110,Priyanka,5)
(104,Anubhav,4)
(109,Katrina,4)
(103,Akshay,3)
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)

```
salary_details = FOREACH emp_details GENERATE id, name, salary;

oddNumberSalary = FILTER salary_details BY (id %2 !=0) ;

orderedSalary = ORDER oddNumberSalary by salary Desc,name ASC;

limitedSalary = LIMIT orderedSalary 3;

dump limitedSalary;
```

```
grunt> emp_details = load '/home/vignesh/Desktop/PIG_EXAMPLES/employee_details.txt' using PigStorage(',') as (id:int,name:chararray,salary:int,rating
:int);
2017-12-13 19:50:15,393 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2017-12-13 19:50:15,394 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> salary_details = FOREACH emp_details GENERATE id, name, salary;
grunt> oddNumberSalary = FILTER salary_details BY (id %2 !=0) ;
grunt> orderedSalary = ORDER oddNumberSalary by salary Desc,name ASC;
grunt> limitedSalary = LIMIT orderedSalary 3;
grunt> dump limitedSalary;
(101,Amitabh,20000)
(107,Salman,17500)
(103,Akshay,11000)
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)

emp_details = load '/home/vignesh/Desktop/PIG_EXAMPLES/employee_details.txt' using PigStorage(',') as (id:int, name:chararray, salary:int, rating:int);

expense = LOAD 'employee_expenses.txt' USING PigStorage('\t') AS (id:int, expenses:int);

join_table = join emp_details by id, expense by id;

join_table_expense = FOREACH join_table GENERATE $0 as id,$1 as name,$5 as expenses;

order_expense = ORDER join_table_expense by expenses Desc,name ASC;
limited_expense = limit order_expense 1;

dump limited_expense;

```
grunt> emp_details = load '/home/vignesh/Desktop/PIG_EXAMPLES/employee_details.txt' using PigStorage(',') as (id:int,name:chararray,salary:int,rating
:int);
2017-12-13 20:09:49,123 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2017-12-13 20:09:49,123 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> expense = LOAD '/home/vignesh/Desktop/PIG_EXAMPLES/employee_expenses.txt' USING PigStorage('\t') AS (id:int, expenses:int);
2017-12-13 20:09:56,543 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2017-12-13 20:09:56,543 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> join_table = join emp_details by id, expense by id;
2017-12-13 20:09:56,810 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. coll
ectionUsageThreshold = 489350752, usageThreshold = 489350752
grunt> join_table_expense = FOREACH join_table GENERATE $0 as id,$1 as name,$5 as expenses;
grunt> order_expense = ORDER join_table_expense by expenses Desc,name ASC;
grunt> limited_expense = limit order_expense 1;
grunt> dump limited_expense;
(110,Priyanka,400)
grunt>
```

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

```
emp_expense = join emp_details by id, expense by id;
empInExpense = foreach emp_expense generate $0,$1;
distinctEmpInExpense = distinct empInExpense;
dump distinctEmpInExpense;
```

```
grunt> empInExpense = foreach join_table generate $0,$1;
grunt> distinctEmpInExpense = distinct empInExpense;
grunt> dump distinctEmpInExpense;
2017-12-13 20:41:32,021 [main] INFO  org.apache.pig.tools.pigstats.Script
2017-12-13 20:41:34,
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt>
```

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

```
emp_expense_full = join emp_details by id FULL OUTER, expense by id;

empNotInExpense_filter = filter emp_expense_full by $4 is null and $5 is null;

empNotInExpense = foreach empNotInExpense_filter generate $0,$1;

distnctEmpNotInExpense = distinct empNotInExpense;

dump distnctEmpNotInExpense;
```

```
grunt> emp_details = load '/home/vignesh/Desktop/PIG_EXAMPLES/employee_details.txt' using PigStorage(',') as (id:int,name:chararray,salary:int,rating
:int);
2017-12-13 20:49:42,793 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2017-12-13 20:49:42,794 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> expense = LOAD '/home/vignesh/Desktop/PIG_EXAMPLES/employee_expenses.txt' USING PigStorage('\t') AS (id:int, expenses:int);
2017-12-13 20:50:11,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2017-12-13 20:50:11,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> emp_expense_full = join emp_details by id FULL OUTER, expense by id;
2017-12-13 20:50:11,708 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. coll
ectionUsageThreshold = 489350752, usageThreshold = 489350752
grunt> empNotInExpense_filter = filter emp_expense_full by $4 is null and $5 is null;
grunt> empNotInExpense = foreach empNotInExpense_filter generate $0,$1;
grunt> distnctEmpNotInExpense = distinct empNotInExpense;
grunt> dump distnctEmpNotInExpense;
```

```
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
```