

## **PIG Project 3 : Problem Statement**

We have employee\_details and employee\_expenses files.

employee\_details.txt

employee\_expenses.txt

Use local mode while running Pig and

write Pig Latin script to get below results:

**(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)**

### **Input Commands:**

```
A0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_details.txt'
USING PigStorage(',') AS (empid: int, empname:chararray,
empsalary:int, emprating:int);
A1 = DISTINCT A0;
DESCRIBE A1;
ILLUSTRATE A1;
EXPLAIN A1;
DUMP A1;
A2 = ORDER A1 BY emprating asc, empname asc;
A3 = LIMIT A2 5;
STORE A3 INTO
'/home/cloudera/chhaya/PigProject3/Outputfile31.txt' using
PigStorage(',');
```

### **Output Screenshot:**

The screenshot shows a terminal window titled 'cloudera-quickstart-vm-5.13.0-0-virtualbox (Running) - Oracle VM VirtualBox'. The terminal output displays the execution of a Pig script. It starts with a timestamp '2018-06-17 21:23:25,594 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:'. Below this, a table-like structure shows job statistics: HadoopVersion (2.6.0-cdh5.13.0), PigVersion (0.12.0-cdh5.13.0), UserId (cloudera), StartedAt (2018-06-17 21:22:34), FinishedAt (2018-06-17 21:23:25), and Features (ORDER\_BY,DISTINCT,LIMIT). A 'Success!' message follows. Then, 'Job Stats (time in seconds):' is shown, followed by a list of jobs with their IDs, aliases, feature outputs, and A2 values. The input is specified as '/home/cloudera/chhaya/PigProject3/employee\_details.txt'. The output is specified as '/home/cloudera/chhaya/PigProject3/Outputfile31.txt'. A DAG (Directed Acyclic Graph) is also shown, illustrating the job dependencies. The terminal ends with a timestamp '2018-06-17 21:23:49,607 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!' and a 'grunt>' prompt. The taskbar at the bottom shows various application icons and the system clock indicating '21:24 17-06-2018'.

```
2018-06-17 21:23:25,594 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt      FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0 cloudera 2018-06-17 21:22:34 2018-06-17 21:23:25 ORDER_BY,DISTINCT,LIMIT

Success!

Job Stats (time in seconds):
JobId  Alias  Feature Outputs
job_local1032469601_0022  A2  /home/cloudera/chhaya/PigProject3/Outputfile31.txt,
job_local1214158075_0020  A2  SAMPLER
job_local44353334_0019  empdetails  DISTINCT
job_local637720181_0021  A2  ORDER_BY,COMBINER

Input(s):
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_details.txt"

Output(s):
Successfully stored records in: "/home/cloudera/chhaya/PigProject3/Outputfile31.txt"

Job DAG:
job_local44353334_0019 -> job_local1214158075_0020,
job_local1214158075_0020 -> job_local637720181_0021,
job_local637720181_0021 -> job_local1032469601_0022,
job_local1032469601_0022

2018-06-17 21:23:49,607 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

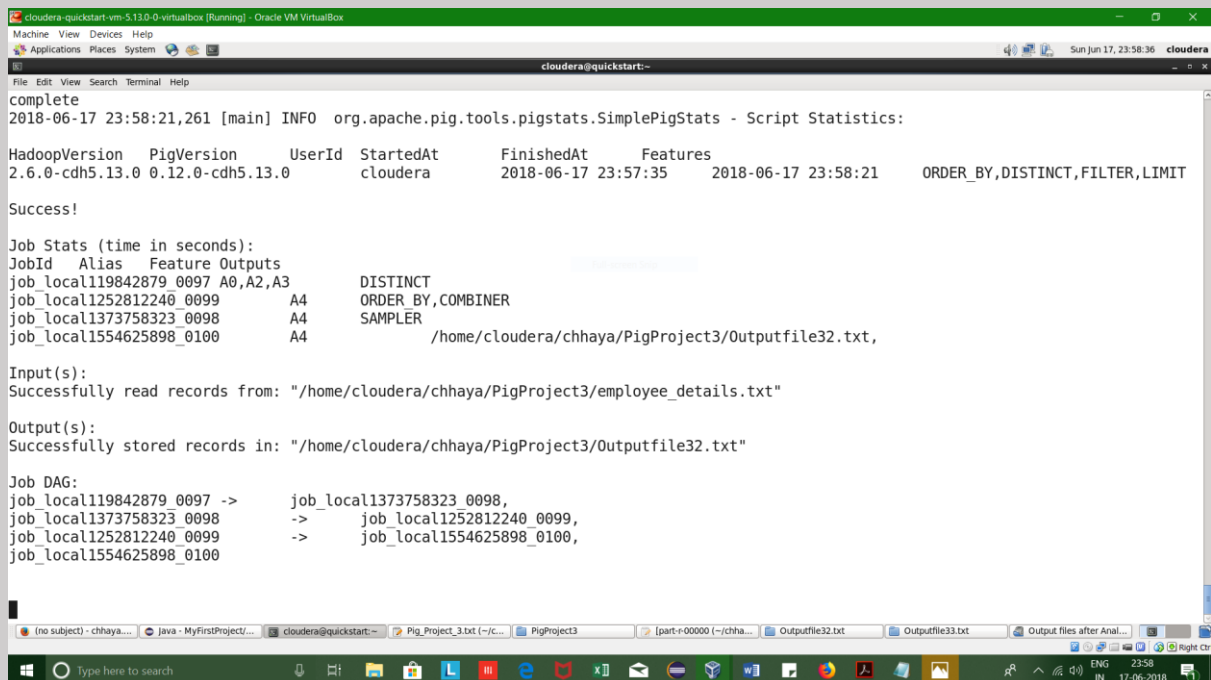
=====

**(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)**

### **Input Commands:**

```
A0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_details.txt'
USING PigStorage(',') AS (empid: int, empname:chararray,
empsalary:int, emprating:int);
A1 = DISTINCT A0;
A2 = FOREACH A1 GENERATE empid as id, empname as name,
empsalary as salary;
A3 = FILTER A2 by ((id % 2) != 0);
A4 = ORDER A3 BY salary desc, name asc;
A5 = LIMIT A4 3;
STORE A5 INTO
'/home/cloudera/chhaya/PigProject3/Outputfile32.txt' using
PigStorage(',');
```

## Output Screenshot:



```
complete
2018-06-17 23:58:21,261 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
2.6.0-cdh5.13.0  0.12.0-cdh5.13.0  cloudera  2018-06-17 23:57:35  2018-06-17 23:58:21  ORDER_BY,DISTINCT,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId  Alias  Feature  Outputs
job_local119842879_0097  A0,A2,A3  DISTINCT
job_local1252812240_0099  A4  ORDER_BY,COMBINER
job_local1373758323_0098  A4  SAMPLER
job_local1554625898_0100  A4  /home/cloudera/chhaya/PigProject3/Outputfile32.txt,

Input(s):
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_details.txt"

Output(s):
Successfully stored records in: "/home/cloudera/chhaya/PigProject3/Outputfile32.txt"

Job DAG:
job_local119842879_0097 -> job_local1373758323_0098,
job_local1373758323_0098 -> job_local1252812240_0099,
job_local1252812240_0099 -> job_local1554625898_0100,
job_local1554625898_0100
```

=====

**(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)**

## Input Commands:

```
B0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_expenses.txt'
USING PigStorage('\t') AS (empid: int, empexpenses: int);
A0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_details.txt'
USING PigStorage(',') AS (empid: int, empname:chararray,
empsalary:int, emprating:int);
A1 = DISTINCT A0;
A2 = FOREACH A1 GENERATE empid, empname ;
A2joinB0 = JOIN A2 BY empid, B0 BY empid;
DESCRIBE A2joinB0;
A2B1 = FOREACH A2joinB0 GENERATE A2::empid as id, A2::empname
as name, B0::empexpenses as exp;
A2B2 = GROUP A2B1 BY (id, name);
```

```

A3B3 = FOREACH A2B2 GENERATE group, SUM(A2B1.exp) as money;
A4B4 = FOREACH A3B3 GENERATE FLATTEN(group) as (id,name),
money;
A4B4 = ORDER A4B4 BY money DESC, name ASC;
A5B5 = LIMIT A4B4 1;
STORE A5B5 INTO
'/home/cloudera/chhaya/PigProject3/Outputfile33.txt' using
PigStorage(',');

```

## Output Screenshot:

```

Success!

Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1066118872_0102 A2B1,A2joinB0,B0 HASH_JOIN
job_local1562808921_0105 A4B4 ORDER_BY,COMBINER
job_local1733982743_0106 A4B4 /home/cloudera/chhaya/PigProject3/Outputfile33.txt,
job_local1754453811_0103 A2B2,A3B3,A4B4 GROUP_BY,COMBINER
job_local31110693_0104 A4B4 SAMPLER
job_local723043394_0101 A0,A2 DISTINCT

Input(s):
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_details.txt"
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_expenses.txt"

Output(s):
Successfully stored records in: "/home/cloudera/chhaya/PigProject3/Outputfile33.txt"

Job DAG:
job_local723043394_0101 -> job_local1066118872_0102,
job_local1066118872_0102 -> job_local1754453811_0103,
job_local1754453811_0103 -> job_local31110693_0104,
job_local31110693_0104 -> job_local1562808921_0105,
job_local1562808921_0105 -> job_local1733982743_0106,
job_local1733982743_0106

```

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

## Input Commands:

```

B0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_expenses.txt'
USING PigStorage('\t') AS (empid: int, empexpenses: int);
A0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_details.txt'

```

```

USING PigStorage(',') AS (empid: int, empname:chararray,
empsalary:int, emprating:int);
A1 = DISTINCT A0;
A2 = FOREACH A1 GENERATE empid, empname ;
A2joinB0 = JOIN A2 BY empid, B0 BY empid;
A2B1 = FOREACH A2joinB0 GENERATE A2::empid as id, A2::empname
as name;
A2B2 = DISTINCT A2B1;
DUMP A2B2;
STORE A2B2 INTO
'/home/cloudera/chhaya/PigProject3/Outputfile34.txt' using
PigStorage(',');

```

## Output Screenshot:

```

complete
2018-06-17 23:07:01,194 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
2.6.0-cdh5.13.0  0.12.0-cdh5.13.0  cloudera  2018-06-17 23:06:35  2018-06-17 23:07:01  HASH_JOIN,DISTINCT

Success!

Job Stats (time in seconds):
JobId  Alias  Feature  Outputs
job_local1324130770_0081  DISTINCT  /home/cloudera/chhaya/PigProject3/Outputfile34.txt,
job_local1989296975_0080  A2B1,A2joinB0,B0  HASH_JOIN
job_local1360763267_0079  A0,A2  DISTINCT

Input(s):
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_details.txt"
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_expenses.txt"

Output(s):
Successfully stored records in: "/home/cloudera/chhaya/PigProject3/Outputfile34.txt"

Job DAG:
job_local1360763267_0079 -> job_local1989296975_0080,
job_local1989296975_0080 -> job_local1324130770_0081,
job_local1324130770_0081

2018-06-17 23:07:19,204 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

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

## Input Commands:

```

B0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_expenses.txt'
USING PigStorage('\t') AS (empid: int, empexpenses: int);

```

```

A0 = LOAD
'/home/cloudera/chhaya/PigProject3/employee_details.txt'
USING PigStorage(',') AS (empid: int, empname:chararray,
empsalary:int, emprating:int);
A1 = DISTINCT A0;
A2 = FOREACH A1 GENERATE empid, empname ;
A2joinB0 = JOIN A2 BY empid LEFT OUTER, B0 BY empid;
A2B1 = FILTER A2joinB0 BY B0::empid is NULL;
A2B2 = FOREACH A2B1 GENERATE A2::empid as id, A2::empname as
name;
DUMP A2B2;
STORE A2B2 INTO
'/home/cloudera/chhaya/PigProject3/Outputfile35.txt' using
PigStorage(',') ;

```

## Output Screenshot:

```

110
2018-06-18 00:02:49,504 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-06-18 00:02:49,513 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be in
complete
2018-06-18 00:02:49,515 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion      UserId  StartedAt      FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0 cloudera 2018-06-18 00:02:29 2018-06-18 00:02:49 HASH_JOIN,DISTINCT,FILTER

Success!

Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1158081340_0109 A0,A2 DISTINCT
job_local744609754_0110 A2B1,A2B2,A2joinB0,B0 HASH_JOIN /home/cloudera/chhaya/PigProject3/Outputfile35.txt,

Input(s):
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_details.txt"
Successfully read records from: "/home/cloudera/chhaya/PigProject3/employee_expenses.txt"

Output(s):
Successfully stored records in: "/home/cloudera/chhaya/PigProject3/Outputfile35.txt"

Job DAG:
job_local1158081340_0109 -> job_local744609754_0110,
job_local744609754_0110

```

=====