Data Analysis using Apache Pig in Hadoop

Dataset: There are three datasets:

- Employee_dataset.txt
- Department_dataset.txt
- Employee_bonus.txt

Employee Dataset: it consists of five columns (id, name, Title, Department id, salary). It has 10290 records separated by ‘,’.

Department Dataset: it consists of three columns (department id, department name, address). It has 16 records separated by ‘;’. Address is further divided into three columns (street, city, state).

Bonus Dataset: it consists of two columns (id, bonus amount). It has 45 records separated by ‘,’.

System and Resources details:

- OS - ubuntu 18.04.3 LTS
- Java - OpenJDK version 1.8.0_222
- Apache Hadoop - Hadoop 3.2.1
- Apache pig - Apache Pig version 0.17.0 (r1797386)

File Copied to HDFS:

```
hdfs dfs -put ./employee_dataset.txt hdfs://localhost:9000/Pig_data/

hdfs dfs -put ./employee_bonus.txt hdfs://localhost:9000/Pig_data/

hdfs dfs -put ./department_dataset.txt hdfs://localhost:9000/Pig_data/
```

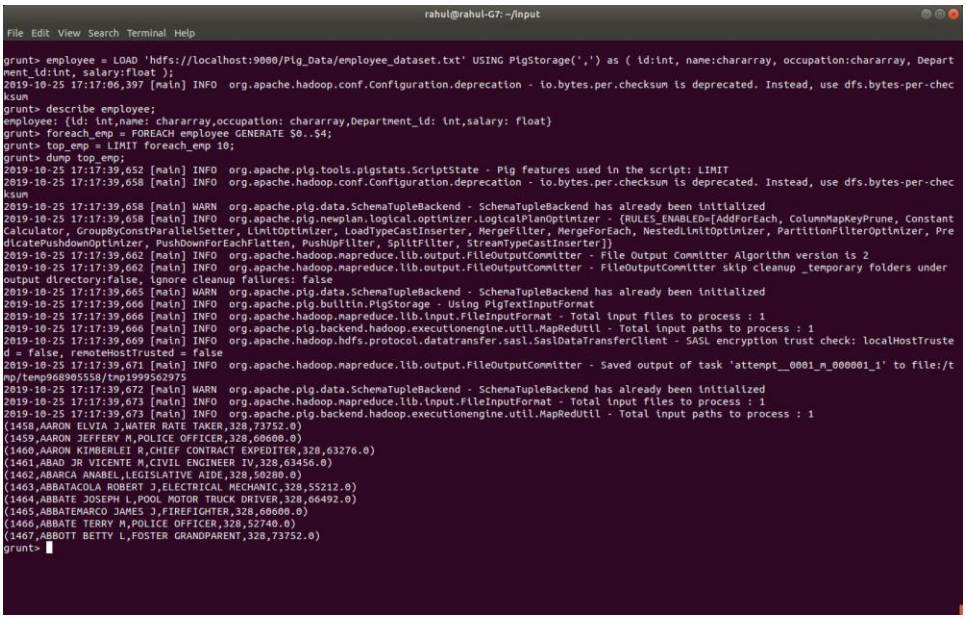
What to discover:

- Display 10 records of Each dataset by creating a relation with schema and describe those relations too.
- Join employee and department dataset to display Department Name and Department Address (as 3 separate columns - street, city and state) for each row in employee.
- Display all the employees who received bonus. The result should not display any employees who did not receive a bonus.

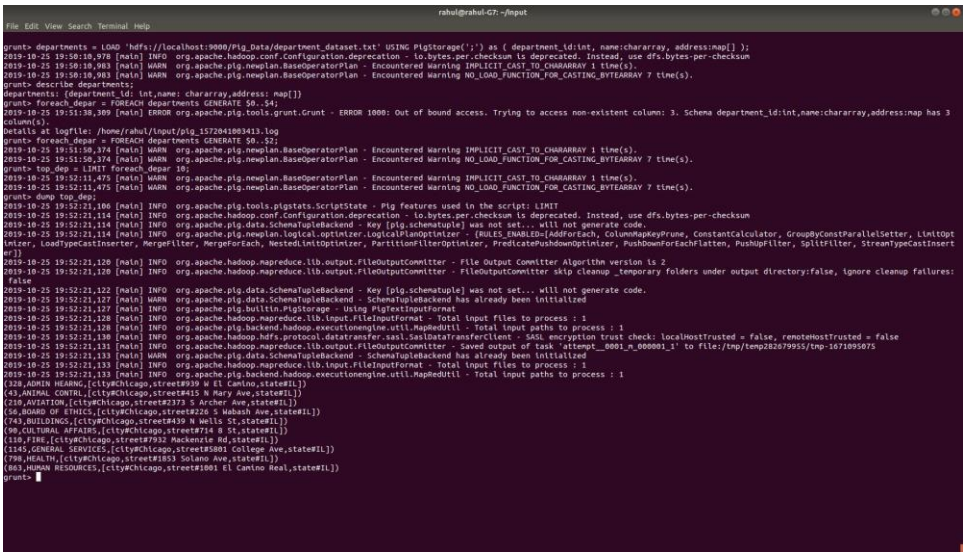
- Display the average bonus by department.
- Display the number of employees in each department using a nested FOREACH.

Screenshots:

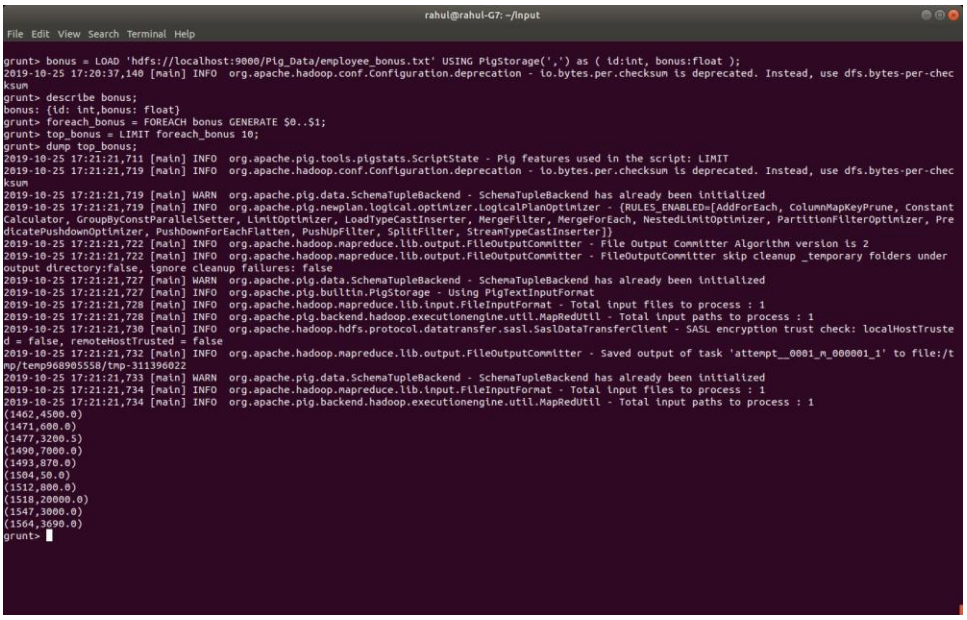
A. Creating a relation with employee dataset with schema and named it employee followed by describing that relation and displaying 10 records of that relation.



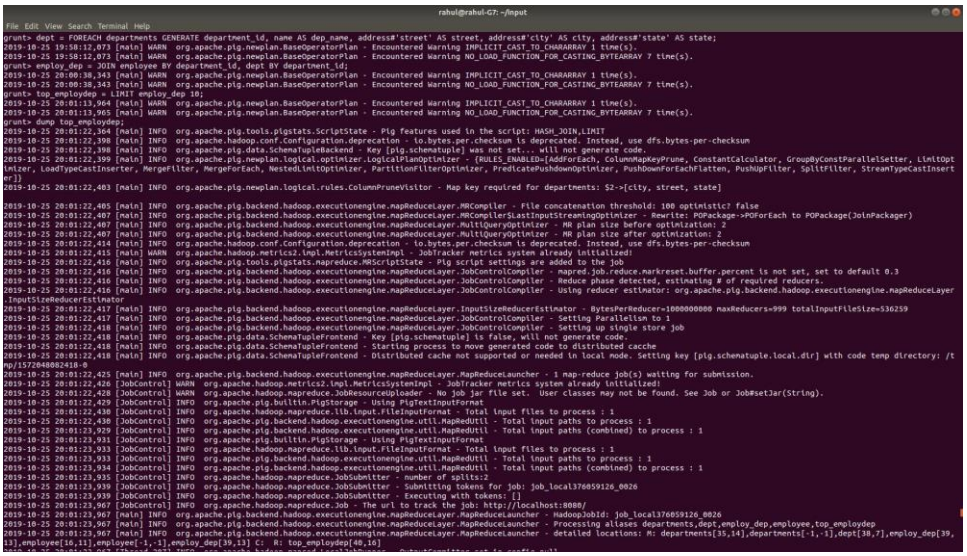
B. Creating a relation with department dataset with schema and named it departments followed by describing that relation and displaying 10 records of that relation.

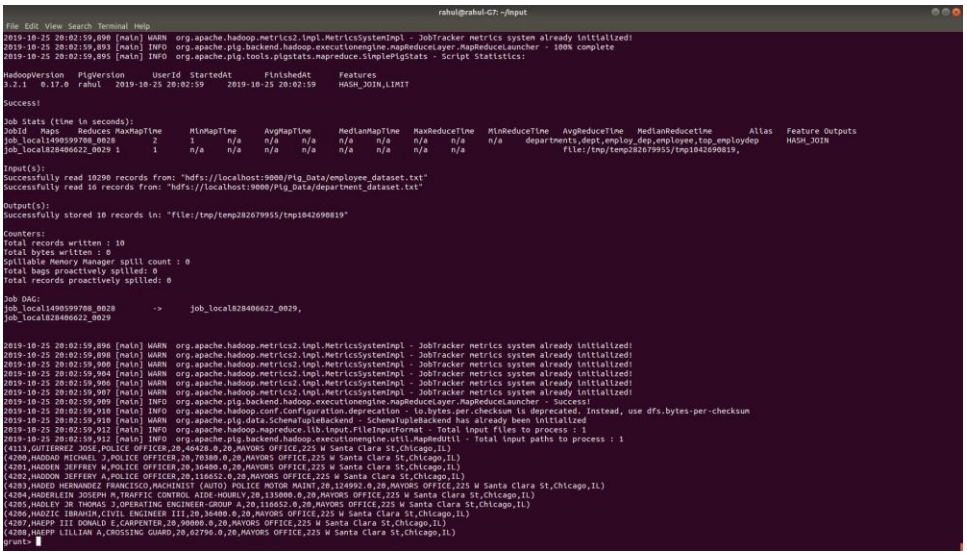


C. Creating a relation with employee bonus dataset with schema and named it bonus followed by describing that relation and displaying 10 records of that relation.

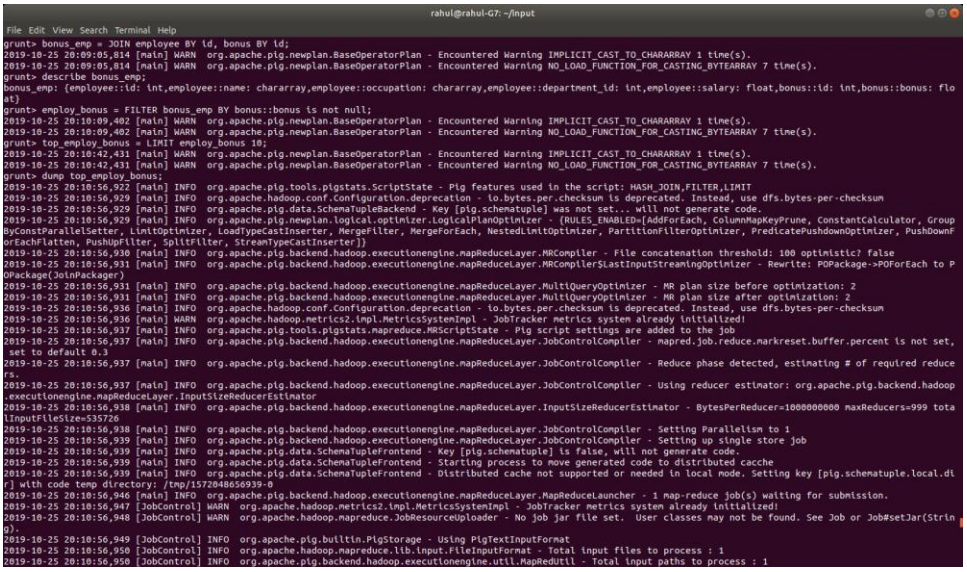


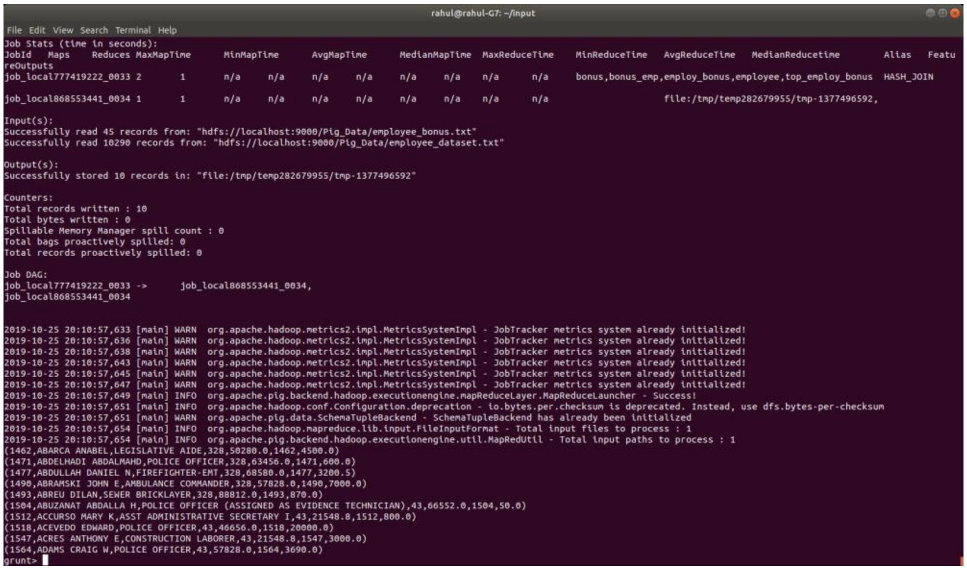
D. Joining employee and department dataset to display Department Name and Department Address (as 3 separate columns - street, city and state) for each row in employee. Displaying 10 of these records.



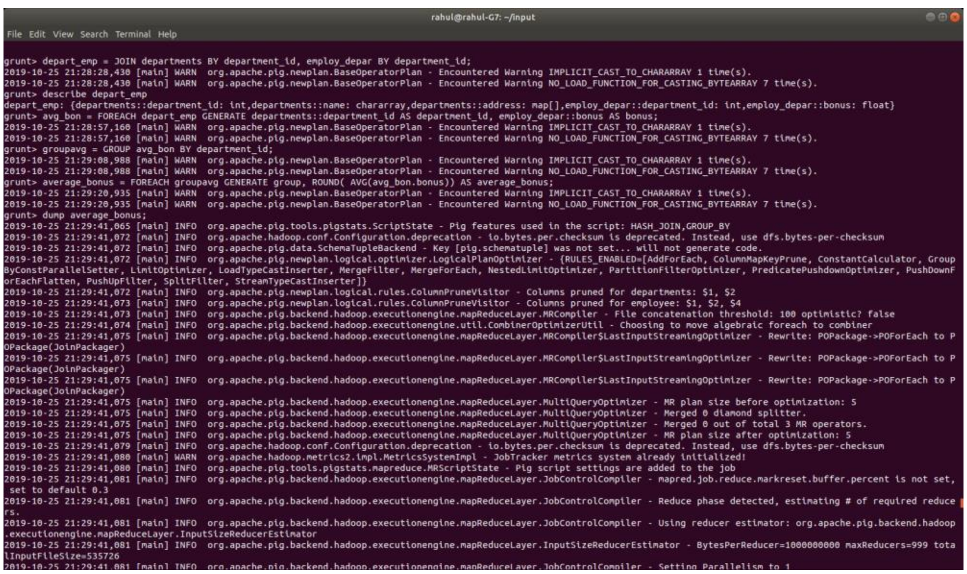


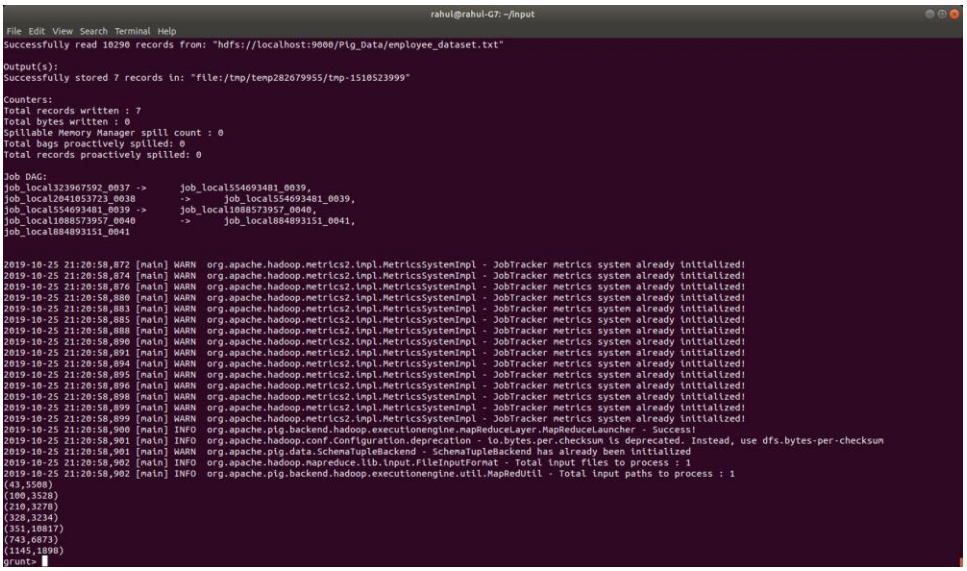
E. Displaying all the employees who received bonus. The result does not include employees who did not receive a bonus.



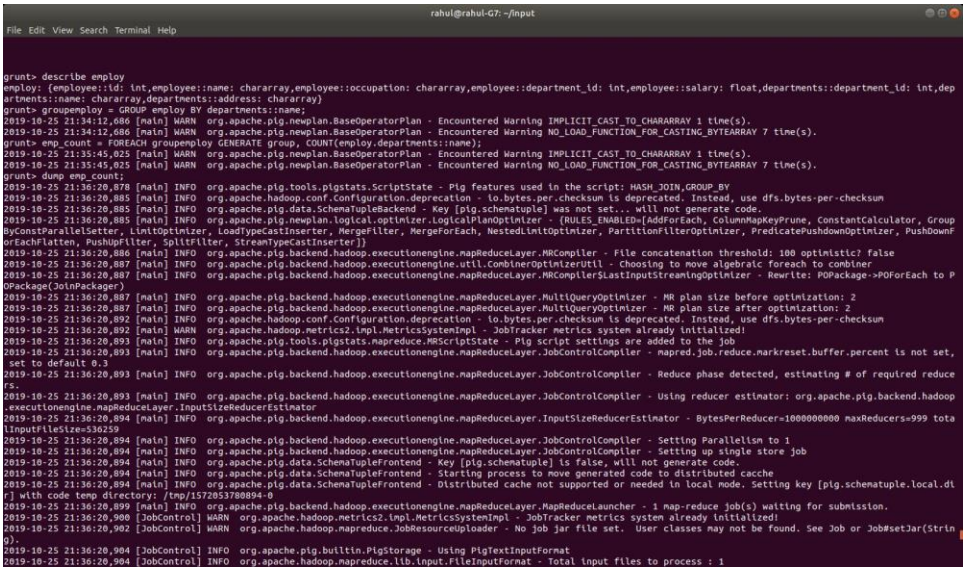


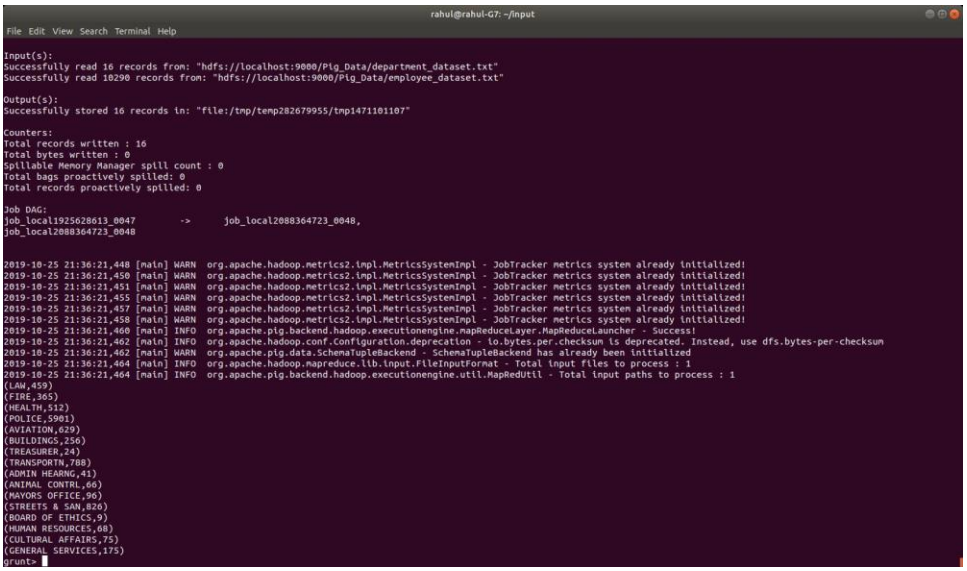
F. Displaying the average bonus by department.





G. Displaying the number of employees in each department using a nested FOREACH.




rahul@rahul-G7: ~/Input
File Edit View Search Terminal Help
Input(s):
Successfully read 16 records from: "hdfs://localhost:9000/Pig_Data/department_dataset.txt"
Successfully read 10290 records from: "hdfs://localhost:9000/Pig_Data/employee_dataset.txt"
Output(s):
Successfully stored 16 records in: "file:/tmp/temp282679955/tmp1471101107"
Counters:
Total records written : 16
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local1925628613_0047 -> job_local2088364723_0048,
job_local2088364723_0048
2019-10-25 21:36:21,448 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:36:21,450 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:36:21,451 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:36:21,455 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:36:21,457 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:36:21,458 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:36:21,460 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2019-10-25 21:36:21,462 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 21:36:21,462 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 21:36:21,464 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 21:36:21,464 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(LAW,459)
(FIRE,365)
(HEALTH,512)
(POLICE,5901)
(AVIATION,629)
(BUILDINGS,256)
(TREASURER,24)
(TRANSPORTN,788)
(ADMIN HEARNG,41)
(ANIMAL CONTRL,66)
(MAYORS OFFICE,96)
(STREETS & SAN,826)
(BOARD OF ETHICS,9)
(HUMAN RESOURCES,68)
(CULTURAL AFFAIRS,75)
(GENERAL SERVICES,175)
grunt>