

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.

```

File Edit View Search Terminal Help
rahu@rahu-G7: ~/input

grunt> employee = LOAD 'hdfs://localhost:9000/Pig_Data/employee_dataset.txt' USING PigStorage(',') as ( id:int, name:chararray, occupation:chararray, Department_id:int, salary:float );
2019-10-25 17:17:06.397 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe employee;
employee: {id: int,name: chararray,occupation: chararray,Department_id: int,salary: float}
grunt> foreach emp = FOREACH employee GENERATE $0..$4;
2019-10-25 17:17:39.662 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
grunt> dump top_emp;
2019-10-25 17:17:39.658 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 17:17:39.658 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 17:17:39.658 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PredicatePushdownOptimizer, PushdownForEachFilter, PushdownFilter, SplitFilter, StreamTypeCastInserter]}
2019-10-25 17:17:39.662 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2019-10-25 17:17:39.662 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2019-10-25 17:17:39.665 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 17:17:39.666 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - Total input files to process : 1
2019-10-25 17:17:39.666 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2019-10-25 17:17:39.669 [main] INFO org.apache.hadoop.hdfs.protocol.datatransfer.sasl.SaslDataTransferClient - SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2019-10-25 17:17:39.671 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_0001_m_000001_1' to file:/tmp/temp96898558/tmp1999562975
2019-10-25 17:17:39.672 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 17:17:39.673 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1458,AARON ELVIA J,WATER RATE TAKER,328,73752.0)
(1459,AARON JEFFERY M,POLICE OFFICER,328,60600.0)
(1460,AARON KIMBERLEY B,CHIEF CONTRACT EXPIRED,328,63276.0)
(1461,ABAD JR VICENTE A,CIVIL ENGINEER IV,328,63456.0)
(1462,ABARCA ANABEL,LEGISLATIVE AIDE,328,50280.0)
(1463,ABBATACOLA ROBERT J,ELECTRICAL MECHANIC,328,55212.0)
(1464,ABBATE JOSEPH L,POOL MOTOR TRUCK DRIVER,328,66492.0)
(1465,ABBATEMARCO JAMES J,FIRE FIGHTER,328,60600.0)
(1466,ABBATE TERRY M,POLICE OFFICER,328,52740.0)
(1467,ABBOTT BETTY L,FOSTER GRANDPARENT,328,73752.0)
grunt>

```

- 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.

```

File Edit View Search Terminal Help
rahu@rahu-G7: ~/input

grunt> departments = LOAD 'hdfs://localhost:9000/Pig_Data/department_dataset.txt' USING PigStorage(',') as ( department_id:int, name:chararray, address:map );
2019-10-25 19:50:18.978 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 19:50:18.983 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 19:50:18.983 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> describe departments;
departments: {department_id: int,name: chararray,address: map}
grunt> foreach depar = FOREACH departments GENERATE $0..$2;
2019-10-25 19:51:18.386 [main] ERROR org.apache.pig.tools.grunt.Grun - ERROR 1008: Out of bound access. Trying to access non-existent column: 3. Schema department_id:int,name:chararray,address:map has 3 column(s).
details at logFile: /home/rahu/input/pig_1572841803413.log
grunt> foreach depar = FOREACH departments GENERATE $0..$2;
2019-10-25 19:51:18.374 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 19:51:18.374 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> top_dep = LIMIT foreach_depar 10;
2019-10-25 19:52:11.471 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 19:52:11.475 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> dump top_dep;
2019-10-25 19:52:12.106 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2019-10-25 19:52:12.114 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 19:52:12.114 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2019-10-25 19:52:12.114 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PredicatePushdownOptimizer, PushdownForEachFilter, PushdownFilter, SplitFilter, StreamTypeCastInserter]}
2019-10-25 19:52:12.126 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2019-10-25 19:52:12.126 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2019-10-25 19:52:12.127 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2019-10-25 19:52:12.127 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 19:52:12.127 [main] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-10-25 19:52:12.128 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 19:52:12.128 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2019-10-25 19:52:12.133 [main] INFO org.apache.hadoop.hdfs.protocol.datatransfer.sasl.SaslDataTransferClient - SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2019-10-25 19:52:12.133 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_0001_m_000001_1' to file:/tmp/temp282879955/tmp-1671095075
2019-10-25 19:52:12.133 [main] INFO org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 19:52:12.133 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 19:52:12.133 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(328,ADMIN HEARING,{city=Chicago,street#000 W El Camino,street#01})
(43,ADMIN CONTIN,{city=Chicago,street#45 N Mary Ave,street#1})
(119,AVIATION,{city=Chicago,street#373 S Archer Ave,street#1})
(36,BORROW OF ETHICS,{city=Chicago,street#26 S Mahan Ave,street#1})
(143,BUILDING SERVICES,{city=Chicago,street#49 N Wells St,street#1})
(99,CULTURAL AFFAIRS,{city=Chicago,street#714 N St,street#1})
(110,FIRE,{city=Chicago,street#302 Mackenzie Rd,street#1})
(114,GENERAL SERVICES,{city=Chicago,street#801 College Ave,street#1})
(798,HEALTH,{city=Chicago,street#1851 Solana Ave,street#1})
(603,HUMAN RESOURCES,{city=Chicago,street#1801 El Camino Real,street#1})
grunt>

```

- 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.

```
File Edit View Search Terminal Help
rahu@rahu-G7: ~/input

grunt> bonus = LOAD 'hdfs://localhost:9000/Pig_Data/employee_bonus.txt' USING PigStorage(',') as ( id:int, bonus:float );
2019-10-25 17:20:37,146 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

grunt> describe bonus;
bonus: {id: int, bonus: float}

grunt> foreach_bonus = FOREACH bonus GENERATE $0.$1;
grunt> top_bonus = LIMIT foreach_bonus 10;
grunt> dump top_bonus;
2019-10-25 17:21:21,711 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2019-10-25 17:21:21,722 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2019-10-25 17:21:21,719 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 17:21:21,719 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - (RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCaster, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushdownForEachFlatten, PushupFilter, SplitFilter, StreamTypeCasterInserter])
2019-10-25 17:21:21,722 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2019-10-25 17:21:21,722 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2019-10-25 17:21:21,727 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 17:21:21,727 [main] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-10-25 17:21:21,728 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 17:21:21,728 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2019-10-25 17:21:21,730 [main] INFO org.apache.hadoop.hdfs.protocol.datatransfer.sasl.SaslDataTransferClient - SASL encryption trust check: localhost:truste
d = false, remote:trusted = false
2019-10-25 17:21:21,732 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_0001_n_000001_1' to file:/t
mp/tempt96898558/tmp-311396022
2019-10-25 17:21:21,733 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 17:21:21,733 [main] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-10-25 17:21:21,734 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1462,4500.0)
(1471,600.0)
(1477,3200.5)
(1490,7000.0)
(1493,870.0)
(1504,50.0)
(1512,800.0)
(1518,10000.0)
(1547,3000.0)
(1564,3600.0)
grunt> █
```

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.

```
File Edit View Search Terminal Help
rahu@rahu-G7: ~/input

grunt> dept = FOREACH departments GENERATE department_id, name AS dep_name, address'street' AS street, address'city' AS city, address'state' AS state;
2019-10-25 19:18:12,073 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
grunt> employ_dep = JOIN employee BY department_id, dept BY department_id;
2019-10-25 20:00:30,343 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
grunt> top_employdep = LIMIT employ_dep 10;
2019-10-25 20:01:13,964 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
grunt> dump top_employdep;
2019-10-25 20:01:13,964 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
2019-10-25 20:01:12,364 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: WARN,JOIN,LIMIT
2019-10-25 20:01:12,398 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 20:01:12,398 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2019-10-25 20:01:12,399 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - (RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCasterInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushdownForEachFlatten, PushupFilter, SplitFilter, StreamTypeCasterIn
serter])
2019-10-25 20:01:12,403 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Map key required for departments: $2->[city, street, state]
2019-10-25 20:01:12,405 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MRCompiler - File concatenation threshold: 100 optimistic: false
2019-10-25 20:01:12,407 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MRCompiler$LastInputStreamingOptimizer - Rewriter: POForEach to POForEach to POForEach (JoinPacker)
2019-10-25 20:01:12,407 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MultiQueryOptimizer - MR plan size before optimization: 2
2019-10-25 20:01:12,407 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MultiQueryOptimizer - MR plan size after optimization: 2
2019-10-25 20:01:12,414 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 20:01:12,414 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized
2019-10-25 20:01:12,416 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2019-10-25 20:01:12,416 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.JobControlCompiler - mapped job.reduce.markreset.buffer.percent is not set, set to default 0.3
2019-10-25 20:01:12,416 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2019-10-25 20:01:12,416 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapreducelayer
.InputSizeReducerEstimator
2019-10-25 20:01:12,417 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.InputSizeReducerEstimator - BytesPerReducer:1000000000 maxReducers:999 totalInputFileLsize:336259
2019-10-25 20:01:12,417 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.JobControlCompiler - Setting Parallelism to 1
2019-10-25 20:01:12,418 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.JobControlCompiler - Setting up single store job
2019-10-25 20:01:12,418 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] is false, will not generate code.
2019-10-25 20:01:12,418 [main] INFO org.apache.pig.data.SchemaTupleBackend - Starting process to move generated code to distributed cache
2019-10-25 20:01:12,418 [main] INFO org.apache.pig.data.SchemaTupleBackend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /t
mp/157284002418-8
2019-10-25 20:01:12,425 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopDeuceLauncher - 1 map-reduce job(s) waiting for submission.
2019-10-25 20:01:12,426 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized
2019-10-25 20:01:12,426 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or JobsetJar(String).
2019-10-25 20:01:12,426 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-10-25 20:01:12,430 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 20:01:12,430 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2019-10-25 20:01:12,930 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2019-10-25 20:01:12,933 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-10-25 20:01:12,933 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 20:01:12,933 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2019-10-25 20:01:12,934 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2019-10-25 20:01:12,934 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:2
2019-10-25 20:01:12,939 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local376859126_0026
2019-10-25 20:01:12,939 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: [1]
2019-10-25 20:01:12,967 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2019-10-25 20:01:12,967 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopDeuceLauncher - HadoopJobId: job_local376859126_0026
2019-10-25 20:01:12,967 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopDeuceLauncher - Processing aliases departments,dept,employ_dep,employ_top,employdep
2019-10-25 20:01:12,967 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.HadoopDeuceLauncher - Detailed locations: R: departments[35,14],departments[1,1],dept[30,7],employ_dep[39,
11],employdep[16,11],employee[1,1],employ_dep[39,11] C: R: top_employdep[49,10]
grunt> █
```



```
File Edit View Search Terminal Help
Job Stats (time in seconds):
JobID  Maps  Reduces MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  Featu
reOutputs
job_local777419222_0033 2  1  n/a  n/a  n/a  n/a  n/a  n/a  bonus_bonus_emp,employ_bonus,employee_top_employ_bonus  HASH_JOIN
job_local868553441_0034 1  1  n/a  n/a  n/a  n/a  n/a  n/a  file:/tmp/temp282679955/tmp-1377496592,

Input(s):
Successfully read 45 records from: "hdfs://localhost:9000/Pig_Data/employee_bonus.txt"
Successfully read 10290 records from: "hdfs://localhost:9000/Pig_Data/employee_dataset.txt"

Output(s):
Successfully stored 10 records in: "file:/tmp/temp282679955/tmp-1377496592"

Counters:
Total records written : 10
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local777419222_0033 -> job_local868553441_0034,
job_local868553441_0034

2019-10-25 20:10:57,633 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 20:10:57,636 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 20:10:57,638 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 20:10:57,643 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 20:10:57,645 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 20:10:57,647 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 20:10:57,649 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - Success!
2019-10-25 20:10:57,651 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 20:10:57,651 [main] INFO org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-10-25 20:10:57,654 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 20:10:57,654 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
(1462,AMARCA ANABEL,LEGISLATIVE AIDE,328,50920.0,1462,4590.0)
(1471,ABDELHADI ABDALLAH,POLICE OFFICER,328,63456.0,1471,680.0)
(1477,ABDULLAH DANIEL N,FIREFIGHTER-EMT,328,60580.0,1477,3280.5)
(1490,ABRANKST JOHN E,AMBULANCE COMMANDER,328,57828.0,1490,7000.0)
(1493,ABREU DILAN,SEWER BRICKLAYER,328,88812.0,1493,870.0)
(1504,ABRUZANAT ABDALLA H,POLICE OFFICER (ASSIGNED AS EVIDENCE TECHNICIAN),43,66552.0,1504,50.0)
(1512,ACUROSS MARY K,ASST ADMINISTRATIVE SECRETARY I,43,21548.0,1512,880.0)
(1518,ACEVEDO EDWARD,POLICE OFFICER,43,46656.0,1518,20800.0)
(1547,ACRES ANTHONY E,CONSTRUCTION LABORER,43,21548.0,1547,3000.0)
(1564,ADAMS CRAIG W,POLICE OFFICER,43,57828.0,1564,3690.0)
run: 1
```

F. Displaying the average bonus by department.

```
File Edit View Search Terminal Help

grunt> depart_emp = JOIN departments BY department_id, employ_depar BY department_id;
2019-10-25 21:28:28,430 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 21:28:28,430 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> describe depart_emp
depart_emp (departments:department_id: int,departments:name: chararray,departments:address: map[],employ_depar:department_id: int,employ_depar:bonus: float)
grunt> avg_bon = FOREACH depart_emp GENERATE departments:department_id AS department_id, employ_depar:bonus AS bonus;
2019-10-25 21:28:57,160 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 21:28:57,160 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> groupavg = GROUP avg_bon BY department_id;
2019-10-25 21:29:08,988 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 21:29:08,988 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> average_bonus = FOREACH groupavg GENERATE group, ROUND( AVG(avg_bon.bonus)) AS average_bonus;
2019-10-25 21:29:20,935 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 21:29:20,935 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grunt> dump average_bonus;
2019-10-25 21:29:41,065 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP,BY
2019-10-25 21:29:41,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 21:29:41,072 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2019-10-25 21:29:41,072 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES_ENABLED=[addrforeach, ColumnKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedJoinOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownOrEachFlatten, PushupFilter, SplitFilter, StreamTypeCastInserter]]
2019-10-25 21:29:41,072 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for departments: $1, $2
2019-10-25 21:29:41,073 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for employe: $1, $2, $4
2019-10-25 21:29:41,073 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCompiler - File concatenation threshold: 100 optimistic? false
2019-10-25 21:29:41,074 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2019-10-25 21:29:41,075 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCompilerLastInputStreamOptimizer - Rewrite: POForEach->POForEach to P
OPackage(JoinPacker)
2019-10-25 21:29:41,075 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCompilerLastInputStreamOptimizer - Rewrite: POPackage->POForEach to P
OPackage(JoinPacker)
2019-10-25 21:29:41,075 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCompilerLastInputStreamOptimizer - Rewrite: POPackage->POForEach to P
OPackage(JoinPacker)
2019-10-25 21:29:41,075 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - MR plan size before optimization: 5
2019-10-25 21:29:41,075 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - Merged 0 diamond splitter.
2019-10-25 21:29:41,075 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - Merged 0 out of total 3 MR operators.
2019-10-25 21:29:41,079 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - MR plan size after optimization: 5
2019-10-25 21:29:41,079 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 21:29:41,080 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:29:41,080 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2019-10-25 21:29:41,081 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - napped.job.reduce.markreset.buffer.percent is not set,
set to default 0.3
2019-10-25 21:29:41,081 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Reduce phase detected, estimating # of required reduce
rs.
2019-10-25 21:29:41,081 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop
.executionengine.mapreduce_layer.InputSizeReducerEstimator
2019-10-25 21:29:41,081 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 tota
lInputFileSize=535726
2019-10-25 21:29:41,081 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Setting Parallelism to 1
```



```
File Edit View Search Terminal Help
Successfully read 10290 records from: "hdfs://localhost:9000/Plg_Data/employee_dataset.txt"
Output(s):
Successfully stored 7 records in: "file:/tmp/202679955/tmp-1510523999"
Counters:
Total records written : 7
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local323967592_0037 -> job_local554693481_0039,
job_local2041053723_0038 -> job_local554693481_0039,
job_local554693481_0039 -> job_local1088573957_0040,
job_local1088573957_0040 -> job_local1084893151_0041,
job_local1084893151_0041
2019-10-25 21:20:58.872 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.874 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.876 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.880 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.883 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.885 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.888 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.890 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.891 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.894 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.895 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.896 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.898 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.899 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.900 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:20:58.901 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - To bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 21:20:58.901 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-10-25 21:20:58.902 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
(43,3500)
(100,3528)
(210,3278)
(320,3234)
(351,10017)
(743,6873)
(1145,1098)
grun>
```

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

```
File Edit View Search Terminal Help
grun> describe employ
employ: (employee: {id: int, employee: {name: chararray, employee: {occupation: chararray, employee: {department_id: int, employee: {salary: float, departments: {department_id: int, dep
artments: {name: chararray, departments: {address: chararray}}}}}}}}
grun> groupemploy = GROUP employ BY departments: {name}
2019-10-25 21:13:42.686 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
grun> emp_count = FOREACH groupemploy GENERATE group, COUNT(employ.departments: {name});
2019-10-25 21:13:45.025 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-10-25 21:13:45.025 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning NO_LOAD_FUNCTION_FOR_CASTING_BYTEARRAY 7 time(s).
grun> dump emp_count;
2019-10-25 21:13:20.078 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN, GROUP_BY
2019-10-25 21:13:20.085 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - To bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 21:13:20.085 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2019-10-25 21:13:20.085 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - RULES_ENABLED[addrforeach, ColumnMapKeyPrune, ConstantCalculator, Group
ByConstParallelFilter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownF
orEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]
2019-10-25 21:13:20.086 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCompiler - File concatenation threshold: 100 optimistic? false
2019-10-25 21:13:20.087 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2019-10-25 21:13:20.087 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MRCompilerLastInputStreamingOptimizer - Rewrite: POPackage->POForEach to P
OPackage(JoinPackage)
2019-10-25 21:13:20.087 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - MR plan size before optimization: 2
2019-10-25 21:13:20.087 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - MR plan size after optimization: 2
2019-10-25 21:13:20.092 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - To bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-10-25 21:13:20.092 [main] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:13:20.093 [main] INFO org.apache.pig.tools.pigstats.MapReduceScriptState - Pig script settings are added to the job
2019-10-25 21:13:20.093 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - mapred.job.reduce.mapresnet.buffer.percent is not set,
get to default 0.3
2019-10-25 21:13:20.093 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Reduce phase detected, estimating # of required reduce
rs.
2019-10-25 21:13:20.093 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop
.executionengine.mapreduce_layer.InputSizeReducerEstimator
2019-10-25 21:13:20.094 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 tota
l input files=36259
2019-10-25 21:13:20.094 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Setting Parallelism to 1
2019-10-25 21:13:20.094 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Setting up single store job
2019-10-25 21:13:20.094 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2019-10-25 21:13:20.094 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2019-10-25 21:13:20.094 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.di
st] with code temp directory: /tmp/157205370004-0
2019-10-25 21:13:20.099 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2019-10-25 21:13:20.900 [JobControl] WARN org.apache.hadoop.metrics2Impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2019-10-25 21:13:20.902 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or JobSetJar(Strin
g).
2019-10-25 21:13:20.904 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-10-25 21:13:20.904 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
```

```
File Edit View Search Terminal Help
rahu@rahu-G7: ~/input

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/tmp147110107"

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.MapReduceUtil - Total input paths to process : 1
(LAW,459)
(FIRE,365)
(HEALTH,512)
(POLICE,5001)
(AVIATION,629)
(BUILDINGS,256)
(TREASURER,24)
(TRANSPORT,788)
(ADMIN HEARING,41)
(ANIMAL CONTROL,66)
(MAYORS OFFICE,90)
(STREETS & SAN,826)
(BOARD OF ETHICS,9)
(HUMAN RESOURCES,68)
(CULTURAL AFFAIRS,75)
(GENERAL SERVICES,175)
run: |
```