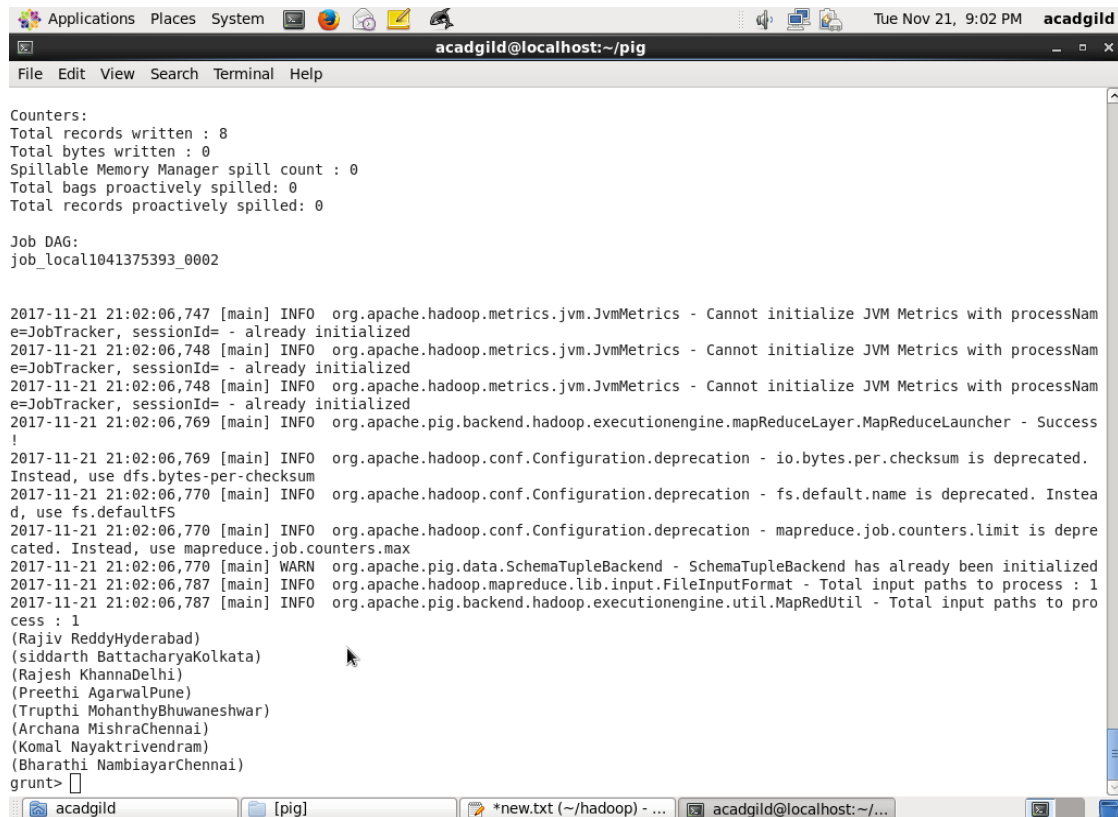# SESSION 4 ASSIGNMENT 2

Contents of the input dataset

```
001,Rajiv Reddy,21,9848022337,Hyderabad,89
002,siddarth Battacharya,22,9848022338, Bhuwaneshwar,78
003,Rajesh Khanna,22,9848022339,Delhi,90
004,Preethi Agarwal,21,9848022330,Pune,93
005,Trupthi Mohanthy,23,9848022336,Bhuwaneshwar,75
006,Archana Mishra,23,9848022335,Chennai,87
006,Archana Mishra,23,9848022335,Chennai,99
007,Komal Nayak,24,9848022334, Delhi,83
008,Bharathi Nambiayar,24,9848022333,Chennai,72
```

Load Command:

```
grunt> student_details = LOAD 'pig_data/student_details.txt' USING
PigStorage(',') as (id:int, name:chararray, age:int, phone:chararray,
city:chararray, gpa:int);
```

1)Concat:

```
grunt> student_name_concat = foreach student_details Generate CONCAT(name, city);
```

2)Tokenize:

```
grunt> student_name_tokenize = foreach student_details  Generate TOKENIZE(name);
```



```
Counters:
Total records written : 8
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local369912192_0003


2017-11-21 21:03:09,449 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:03:09,450 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:03:09,450 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:03:09,453 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-21 21:03:09,458 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-21 21:03:09,458 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-21 21:03:09,458 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-21 21:03:09,459 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-21 21:03:09,475 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-21 21:03:09,475 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
({(Rajiv),(Reddy)})
({(siddarth),(Battacharya)})
({(Rajesh),(Khanna)})
({(Preethi),(Agarwal)})
({(Trupthi),(Mohanthy)})
({(Archana),(Mishra)})
({(Komal),(Nayak)})
({(Bharathi),(Nambiayar)})
grunt>
```

3)Sum:

```
grunt> student_gpa_group = Group student_details by city;

grunt> student_gpa_sum = foreach student_gpa_group Generate (student_details. city,
student_details.gpa),SUM(student_details.gpa);
```

acadgild@localhost:~/pig

File Edit View Search Terminal Help

```
Successfully stored 7 records in: "file:/tmp/temp-1841638081/tmp-68818700"

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_local310938594_0004


2017-11-21 21:44:06,550 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:44:06,551 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:44:06,554 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:44:06,560 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-21 21:44:06,560 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-21 21:44:06,565 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-21 21:44:06,565 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-21 21:44:06,565 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-21 21:44:06,597 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-21 21:44:06,597 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(Pune,{(4,Preethi Agarwal,21,9848022330,Pune,93)})
(Delhi,{(3,Rajesh Khanna,22,9848022339,Delhi,90)})
(Chennai,{(8,Bharathi Nambiayar,24,9848022333,Chennai,72),(6,Archana Mishra,23,9848022335,Chennai,87)})
(Kolkata,{(2,siddarth Battacharya,22,9848022338,Kolkata,78)})
(Hyderabad,{(1,Rajiv Reddy,21,9848022337,Hyderabad,89)})
(trivendram,{(7,Komal Nayak,24,9848022334,trivendram,83)})
(Bhuwaneshwar,{(5,Trupthi Mohanthy,23,9848022336,Bhuwaneshwar,75)})
grunt>
```

acadgild@localhost:~/pig

File Edit View Search Terminal Help

```
Successfully stored 7 records in: "file:/tmp/temp-1841638081/tmp-1258890408"

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_local1743563254_0005


2017-11-21 21:47:46,416 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:47:46,422 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:47:46,423 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-21 21:47:46,429 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-21 21:47:46,430 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-21 21:47:46,430 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-21 21:47:46,430 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-21 21:47:46,430 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-21 21:47:46,448 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-21 21:47:46,448 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(({(Pune)},{(93)}),93)
(({(Delhi)},{(90)}),90)
(({(Chennai),(Chennai)},{(72),(87)}),159)
(({(Kolkata)},{(78)}),78)
(({(Hyderabad)},{(89)}),89)
(({(trivendram)},{(83)}),83)
(({(Bhuwaneshwar)},{(75)}),75)
grunt>
```
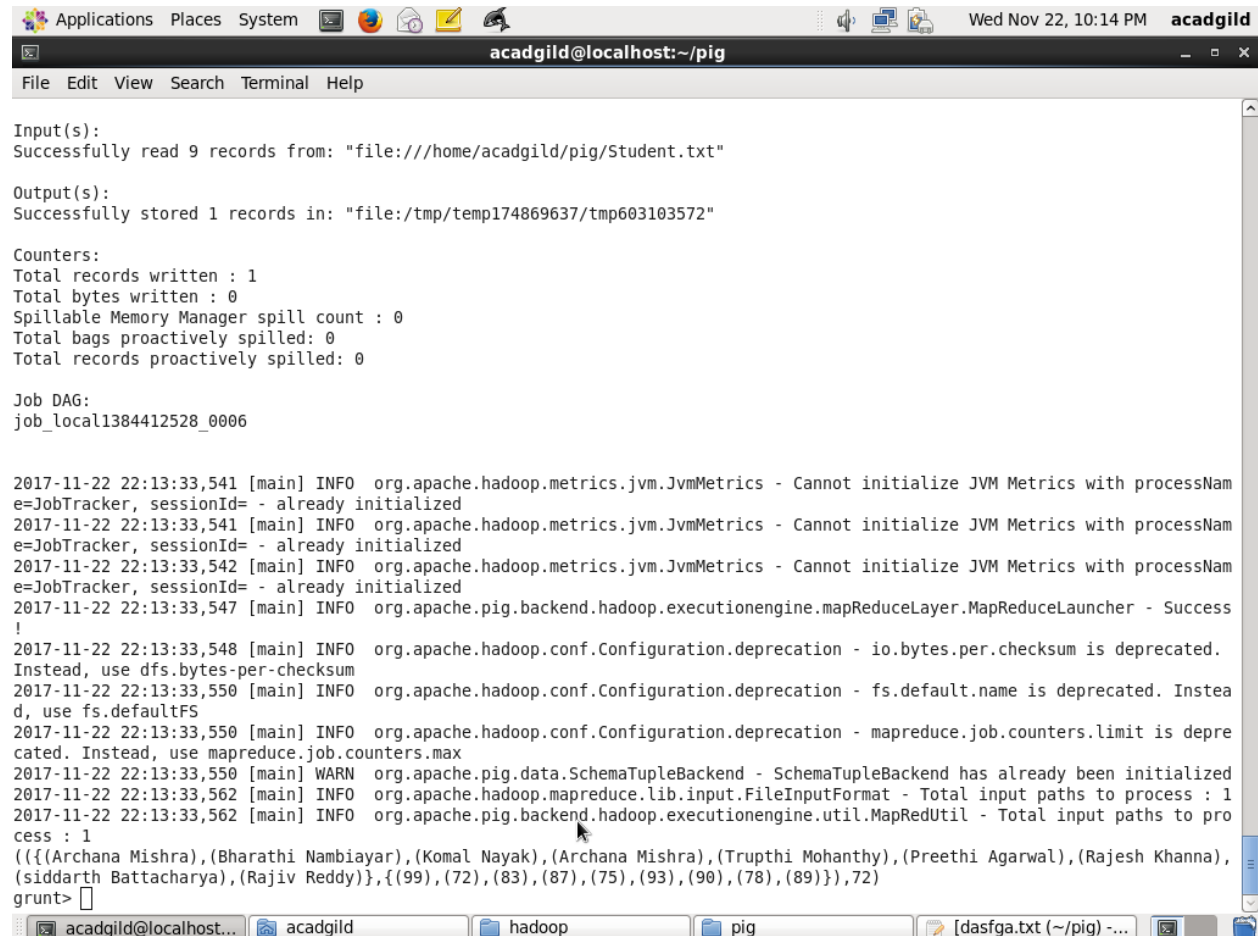
4)Min:

```
grunt> student_group_all = Group student_details All;

grunt> student_gpa_min = foreach student_group_all  Generate
(student_details.firstname, student_details.gpa), MIN(student_details.gpa);
```


5)Max:

```
grunt> student_group_all = Group student_details All;

grunt> student_gpa_max = foreach student_group_all  Generate
(student_details.firstname, student_details.gpa), MAX(student_details.gpa);
```

6)Limit:

```
grunt> limit_data = LIMIT student_details 4;
```

7)Store:

```
grunt> STORE student INTO ' hdfs://localhost:9000/pig_Output/ ' USING PigStorage
(',');
```

The screenshot shows a terminal window titled "acadgild@localhost:~/pig" with the following content:

```
duceTime        MedianReducetime      Alias   Feature Outputs
job_local1357044580_0008      1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     limit_data,st
udent_details
job_local1740695811_0009      1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     student_detai
ls              file:///home/acadgild/pig/pig_Output.txt,

Input(s):
Successfully read 4 records from: "file:///home/acadgild/pig/Student.txt"

Output(s):
Successfully stored 4 records in: "file:///home/acadgild/pig/pig_Output.txt"

Counters:
Total records written : 4
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1357044580_0008        ->      job_local1740695811_0009,
job_local1740695811_0009


2017-11-22 22:17:23,739 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:17:23,740 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:17:23,745 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:17:23,751 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:17:23,751 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:17:23,751 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:17:23,766 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
grunt>
```
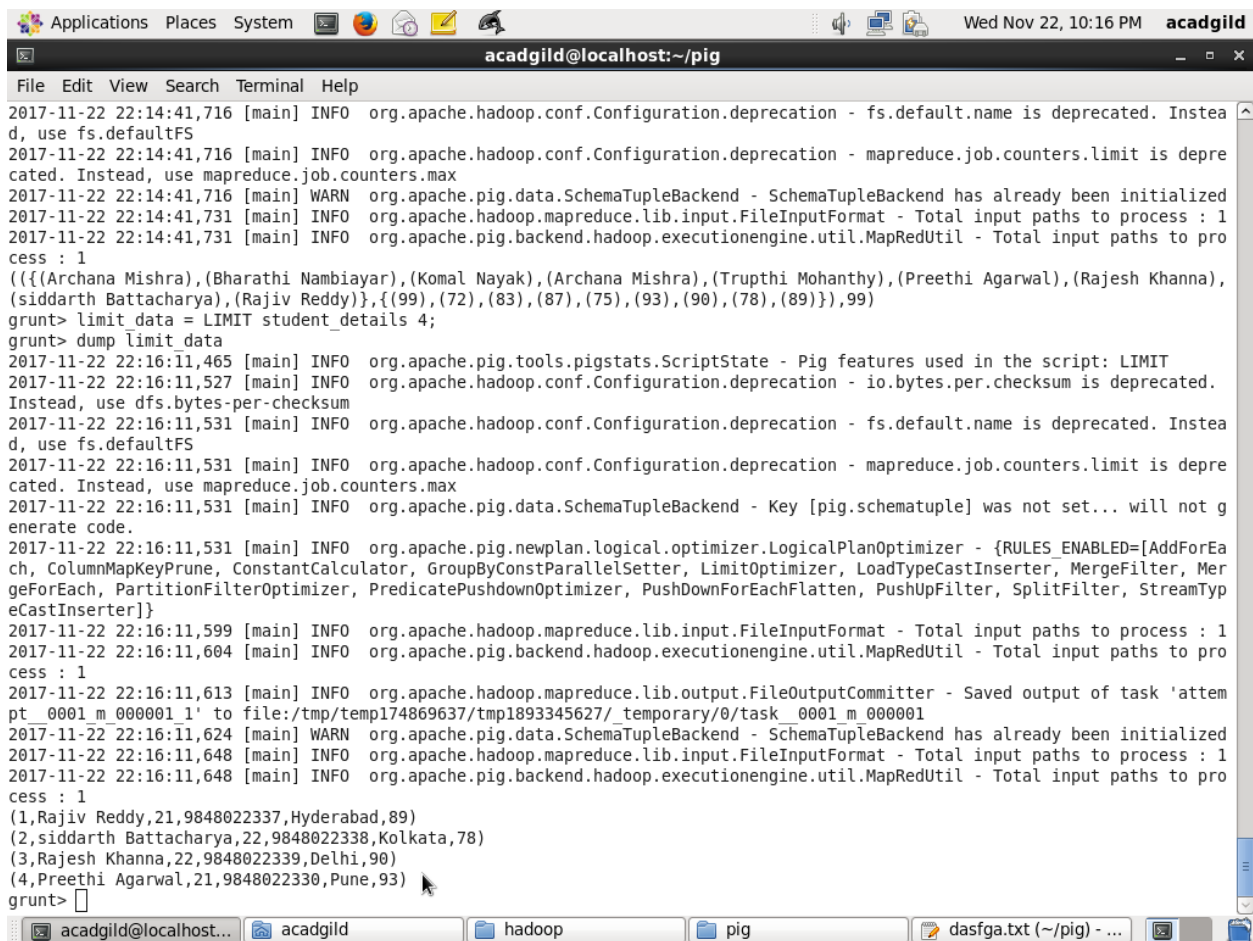
8)Distinct:

```
grunt> distinct_data = DISTINCT student_details;
```
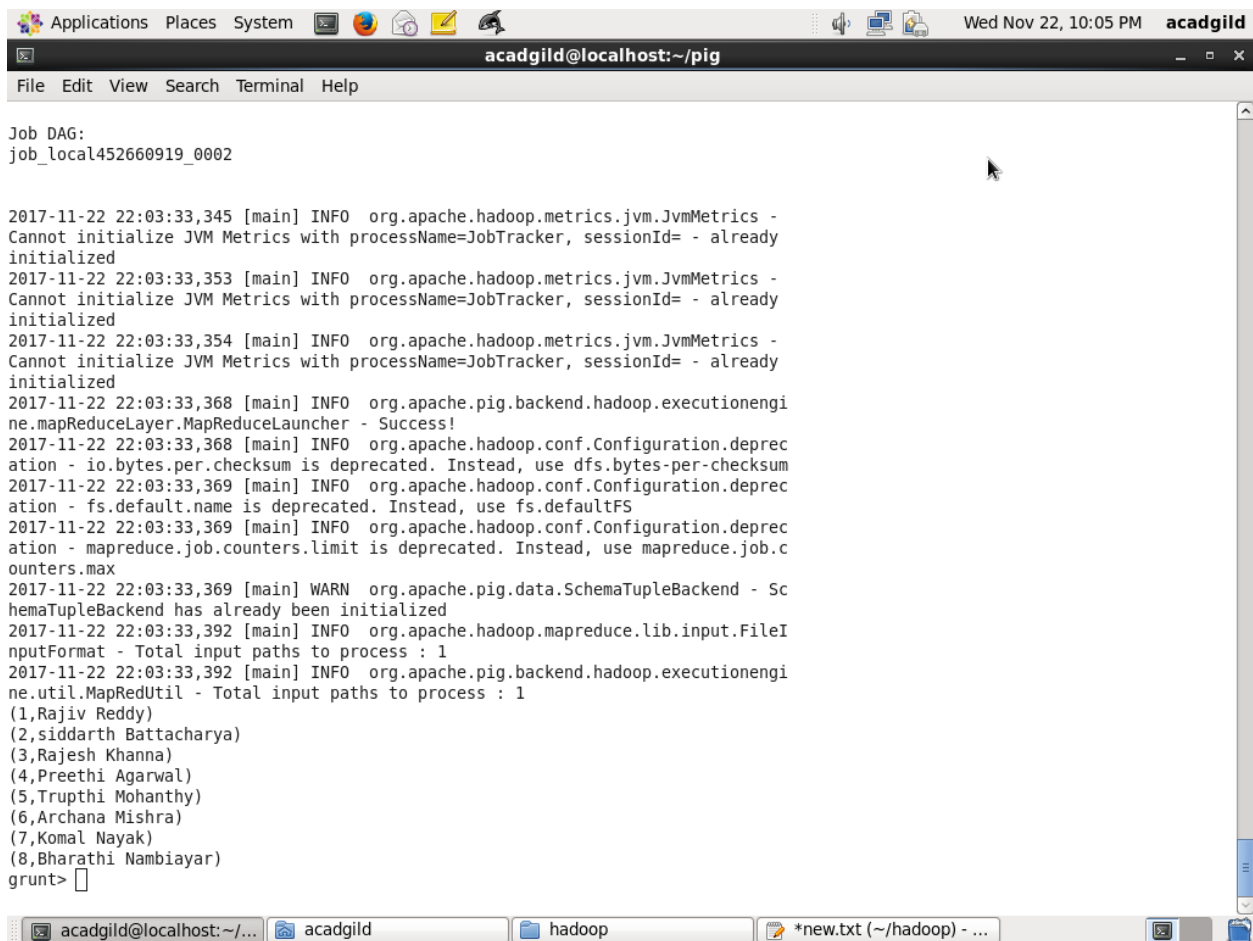
```
Counters:
Total records written : 9
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1799825510_0010


2017-11-22 22:24:25,012 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:24:25,012 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:24:25,013 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:24:25,020 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-22 22:24:25,020 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-22 22:24:25,020 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-22 22:24:25,020 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-22 22:24:25,020 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-22 22:24:25,046 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-22 22:24:25,046 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(1,Rajiv Reddy,21,9848022337,Hyderabad,89)
(2,siddarth Battacharya,22,9848022338,Kolkata,78)
(3,Rajesh Khanna,22,9848022339,Delhi,90)
(4,Preethi Agarwal,21,9848022330,Pune,93)
(5,Trupthi Mohanthy,23,9848022336,Bhuwaneshwar,75)
(6,Archana Mishra,23,9848022335,Chennai,87)
(6,Archana Mishra,23,9848022335,Chennai,99)
(7,Komal Nayak,24,9848022334,trivendram,83)
(8,Bharathi Nambiayar,24,9848022333,Chennai,72)
grunt>
```

9)Flatten:

```
grunt> student_faltten = FOREACH student_details GENERATE id, flatten(name);
```

```
Job DAG:
job_local452660919_0002


2017-11-22 22:03:33,345 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-11-22 22:03:33,353 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-11-22 22:03:33,354 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-11-22 22:03:33,368 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2017-11-22 22:03:33,368 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-22 22:03:33,369 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-22 22:03:33,369 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.c
ounters.max
2017-11-22 22:03:33,369 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2017-11-22 22:03:33,392 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-11-22 22:03:33,392 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv Reddy)
(2,siddarth Battacharya)
(3,Rajesh Khanna)
(4,Preethi Agarwal)
(5,Trupthi Mohanthy)
(6,Archana Mishra)
(7,Komal Nayak)
(8,Bharathi Nambiayar)
grunt>
```

10) ISEMPTY:

Contents of the input dataset

STUDENT_1:

```
001,Rajiv Reddy,21,9848022337,Hyderabad,89
002,siddarth Battacharya,22,9848022338, Bhuwaneshwar,78
003,Rajesh Khanna,22,9848022339,Delhi,90
004,Preethi Agarwal,21,9848022330,Pune,93
005,Trupthi Mohanthy,23,9848022336,Bhuwaneshwar,75
006,Archana Mishra,23,9848022335,Chennai,87
006,Archana Mishra,23,9848022335,Chennai,99
007,Komal Nayak,24,9848022334, Delhi,83
008,Bharathi Nambiayar,24,9848022333,Chennai,72
```

STUDENT_2

```
001,Rajiv Reddy,21,9848022337,Hyderabad,89
002,siddarth Battacharya,22,9848022338, Bhuwaneshwar,78
003,Rajesh Khanna,22,9848022339,Delhi,90
004,Preethi Agarwal,21,9848022330,Pune,93
005,Trupthi Mohanthy,23,9848022336,Bhuwaneshwar,75
006,Archana Mishra,23,9848022335,Chennai,87
006,Archana Mishra,23,9848022335,Chennai,99
007,Komal Nayak,24,9848022334, Delhi,83
008,Bharathi Nambiayar,42,9848022333,Chennai,72
```

```
grunt> student_details = LOAD 'STUDENT_1.txt' USING PigStorage(',') as
(id:int, name:chararray, age:int, phone:chararray, city:chararray, gpa:int);

grunt> student_details_1 = LOAD 'STUDENT_2.txt' USING PigStorage(',') as
(id:int, name:chararray, age:int, phone:chararray, city:chararray, gpa:int);

grunt> cogroup_data = COGROUP student_details by age, student_details_1 by
age;

grunt> isempty_data = filter cogroup_data by IsEmpty(student_details);
```

🖥   **acadgild@localhost:~/pig**   _ □ ✕

File   Edit   View   Search   Terminal   Help

```
Input(s):
Successfully read 8 records from: "file:///home/acadgild/pig/Student.txt"
Successfully read 8 records from: "file:///home/acadgild/pig/stu.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp174869637/tmp174792832"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1445198437_0023


2017-11-22 22:45:58,251 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:45:58,251 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:45:58,251 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-22 22:45:58,255 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-22 22:45:58,255 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-22 22:45:58,256 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-22 22:45:58,261 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-22 22:45:58,261 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-22 22:45:58,278 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-22 22:45:58,278 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(45,{},{(8,Bharathi Nambiayar,45,9848022333,Chennai,72)})
grunt> 
```

🖥 acadgild@localhost:~/...   🏠 acadgild   📁 hadoop   📁 pig   🖥 🗔