Project 1:
USA Crime Analysis
(Using Pig)

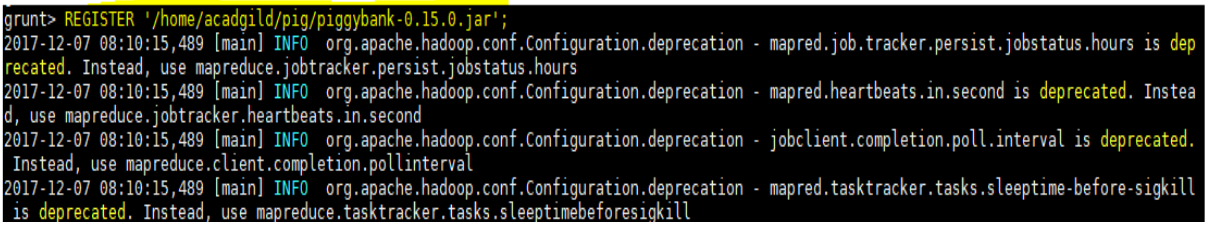
Contents

Problem Statement:

1. Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code.

Steps:

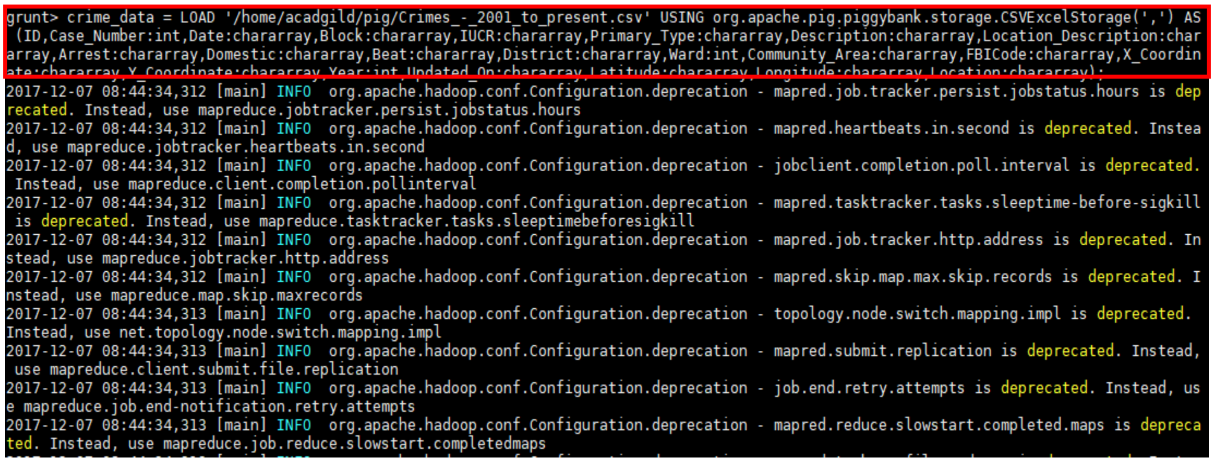
- Register the **piggybank-X.jar** to make use of the **CSVExcelStorage** functionality by using the beow query:
  REGISTER '/home/acadgild/pig/piggybank-0.15.0.jar';



```
grunt> REGISTER '/home/acadgild/pig/piggybank-0.15.0.jar';
2017-12-07 08:10:15,489 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.hours is dep
recated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2017-12-07 08:10:15,489 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is deprecated. Instea
d, use mapreduce.jobtracker.heartbeats.in.second
2017-12-07 08:10:15,489 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion.poll.interval is deprecated.
 Instead, use mapreduce.client.completion.pollinterval
2017-12-07 08:10:15,489 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.tasks.sleeptime-before-sigkill
 is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebeforesigkill
```

- Load the **Crimes_-_2001_to_present.csv** to a **crime_data** variable using below query:
  crime_data = LOAD '/home/acadgild/pig/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',') AS (ID,Case_Number:int,Date:chararray,Block:chararray,IUCR:chararray,Primary_Type:chararray,Description:chararray,Location_Description:chararray,Arrest:chararray,Do

mestic:chararray,Beat:chararray,District:chararray,Ward:int,Community_Area:chararray,FBICode:chararray,X_Coordinate:chararray,Y_Coordinate:chararray,Year:int,Updated_On:chararray,Latitude:chararray,Longitude:chararray,Location:chararray);



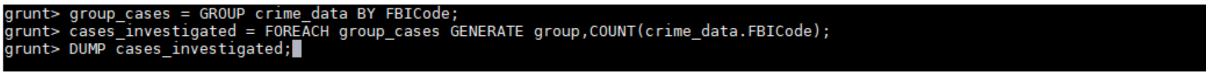
```
grunt> crime_data = LOAD '/home/acadgild/pig/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',') AS
 (ID,Case_Number:int,Date:chararray,Block:chararray,IUCR:chararray,Primary_Type:chararray,Description:chararray,Location_Description:char
array,Arrest:chararray,Domestic:chararray,Beat:chararray,District:chararray,Ward:int,Community_Area:chararray,FBICode:chararray,X_Coordin
ate:chararray,Y_Coordinate:chararray,Year:int,Updated_On:chararray,Latitude:chararray,Longitude:chararray,Location:chararray);
2017-12-07 08:44:34,312 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.hours is dep
recated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2017-12-07 08:44:34,312 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is deprecated. Instea
d, use mapreduce.jobtracker.heartbeats.in.second
2017-12-07 08:44:34,312 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion.poll.interval is deprecated.
 Instead, use mapreduce.client.completion.pollinterval
2017-12-07 08:44:34,312 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.tasks.sleeptime-before-sigkill
 is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebeforesigkill
2017-12-07 08:44:34,312 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. In
stead, use mapreduce.jobtracker.http.address
2017-12-07 08:44:34,312 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.map.max.skip.records is deprecated. I
nstead, use mapreduce.map.skip.maxrecords
2017-12-07 08:44:34,313 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - topology.node.switch.mapping.impl is deprecated.
Instead, use net.topology.node.switch.mapping.impl
2017-12-07 08:44:34,313 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead,
 use mapreduce.client.submit.file.replication
2017-12-07 08:44:34,313 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - job.end.retry.attempts is deprecated. Instead, us
e mapreduce.job.end-notification.retry.attempts
2017-12-07 08:44:34,313 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.slowstart.completed.maps is depreca
ted. Instead, use mapreduce.job.reduce.slowstart.completedmaps
```

- **Use the below query to group the crime_data by FBI code.**

  group_cases = GROUP crime_data BY FBICode;

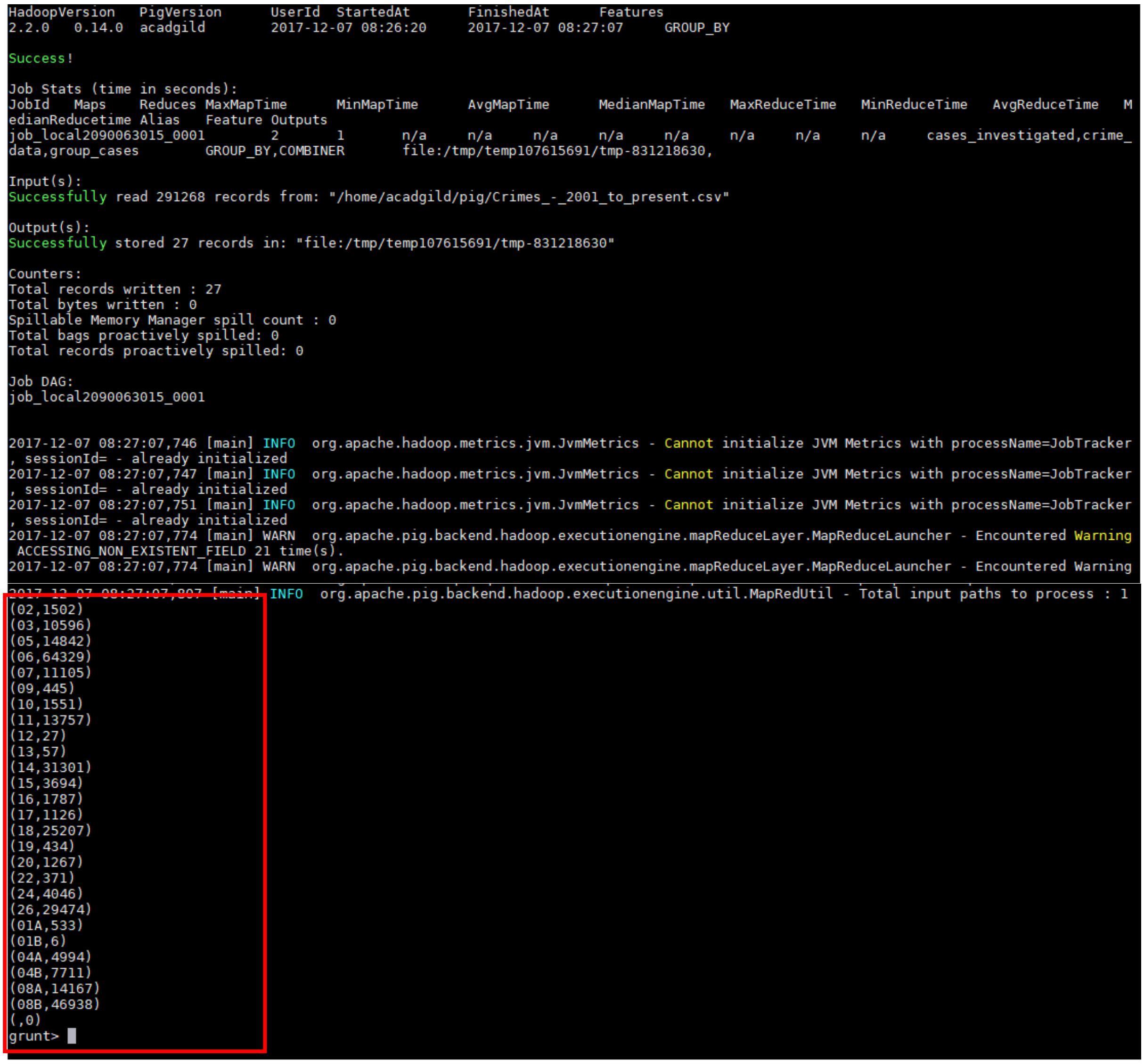
- **Now finally use below query to find the count of cases for each FBI code.**

  cases_investigated = FOREACH group_cases GENERATE group,COUNT(crime_data.FBICode);



```
grunt> group_cases = GROUP crime_data BY FBICode;
grunt> cases_investigated = FOREACH group_cases GENERATE group,COUNT(crime_data.FBICode);
grunt> DUMP cases_investigated;
```

<----------------------- Intermediate logs ------------------------------------>



```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2017-12-07 08:26:20     2017-12-07 08:27:07     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   M
edianReducetime Alias   Feature Outputs
job_local2090063015_0001        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     cases_investigated,crime_
data,group_cases        GROUP_BY,COMBINER       file:/tmp/temp107615691/tmp-831218630,

Input(s):
Successfully read 291268 records from: "/home/acadgild/pig/Crimes_-_2001_to_present.csv"

Output(s):
Successfully stored 27 records in: "file:/tmp/temp107615691/tmp-831218630"

Counters:
Total records written : 27
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2090063015_0001


2017-12-07 08:27:07,746 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2017-12-07 08:27:07,747 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2017-12-07 08:27:07,751 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2017-12-07 08:27:07,774 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning
 ACCESSING_NON_EXISTENT_FIELD 21 time(s).
2017-12-07 08:27:07,774 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning
2017-12-07 08:27:07,807 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(02,1502)
(03,10596)
(05,14842)
(06,64329)
(07,11105)
(09,445)
(10,1551)
(11,13757)
(12,27)
(13,57)
(14,31301)
(15,3694)
(16,1787)
(17,1126)
(18,25207)
(19,434)
(20,1267)
(22,371)
(24,4046)
(26,29474)
(01A,533)
(01B,6)
(04A,4994)
(04B,7711)
(08A,14167)
(08B,46938)
(,0)
grunt>
```

2. Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.

- **Use the below query to filter the data by FBI code 32**

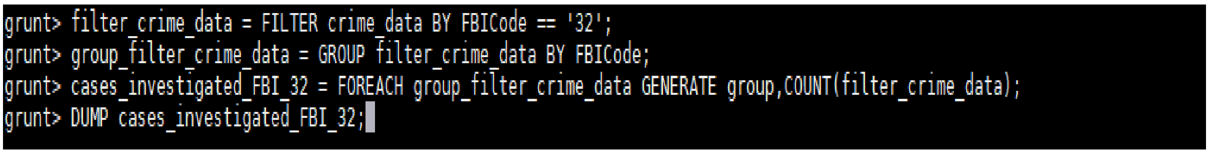
  filter_crime_data = FILTER crime_data BY FBICode == '32';

- **Group the above data by FBICode**

  group_filter_crime_data = GROUP filter_crime_data BY FBICode;

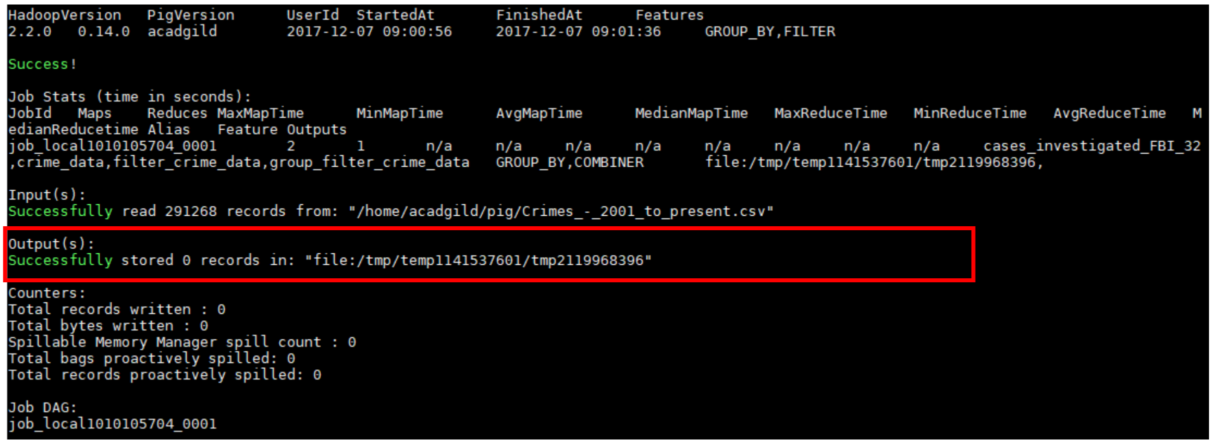
- **Now use the below query to find the count of cases investigated by FBI**

  cases_investigated_FBI_32 = FOREACH group_filter_crime_data GENERATE group,COUNT(filter_crime_data);



```
grunt> filter_crime_data = FILTER crime_data BY FBICode == '32';
grunt> group_filter_crime_data = GROUP filter_crime_data BY FBICode;
grunt> cases_investigated_FBI_32 = FOREACH group_filter_crime_data GENERATE group,COUNT(filter_crime_data);
grunt> DUMP cases_investigated_FBI_32;
```

We got the below output

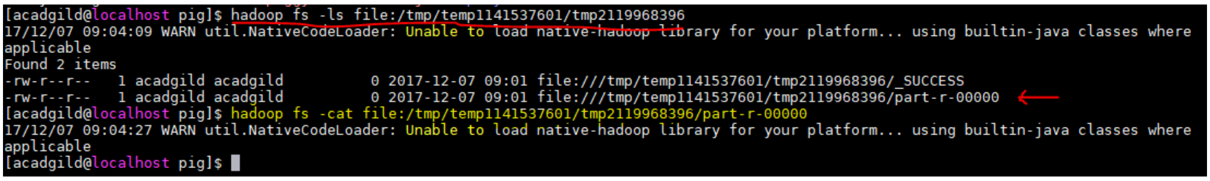


```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2017-12-07 09:00:56     2017-12-07 09:01:36     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   M
edianReducetime Alias   Feature Outputs
job_local1010105704_0001        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     cases_investigated_FBI_32
,crime_data,filter_crime_data,group_filter_crime_data   GROUP_BY,COMBINER       file:/tmp/temp1141537601/tmp2119968396,

Input(s):
Successfully read 291268 records from: "/home/acadgild/pig/Crimes_-_2001_to_present.csv"

Output(s):
Successfully stored 0 records in: "file:/tmp/temp1141537601/tmp2119968396"

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

Job DAG:
job_local1010105704_0001
```

- **Check the file in the hdfs using the below query to see the output:**

  hadoop fs -ls file:/tmp/temp1141537601/tmp2119968396

  hadoop fs -cat file:/tmp/temp1141537601/tmp2119968396/part-r-00000



```
[acadgild@localhost pig]$ hadoop fs -ls file:/tmp/temp1141537601/tmp2119968396
17/12/07 09:04:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
Found 2 items
-rw-r--r--   1 acadgild acadgild          0 2017-12-07 09:01 file:///tmp/temp1141537601/tmp2119968396/_SUCCESS
-rw-r--r--   1 acadgild acadgild          0 2017-12-07 09:01 file:///tmp/temp1141537601/tmp2119968396/part-r-00000
[acadgild@localhost pig]$ hadoop fs -cat file:/tmp/temp1141537601/tmp2119968396/part-r-00000
17/12/07 09:04:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
[acadgild@localhost pig]$
```

3. Write a MapReduce/Pig program to calculate the number of arrests in theft district wise.

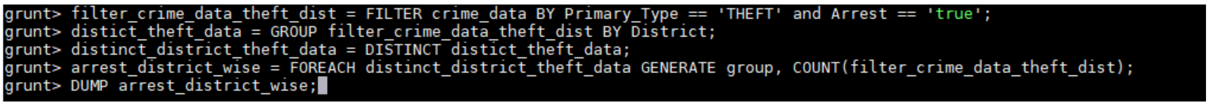
- **Filter the crime_data using the fields Primary_Type by 'THEFT' and Arrest by 'true'**

  filter_crime_data_theft_dist = FILTER crime_data BY Primary_Type == 'THEFT' and Arrest == 'true';
- **Group the filter_crime_data_theft_dist data from above query by District**

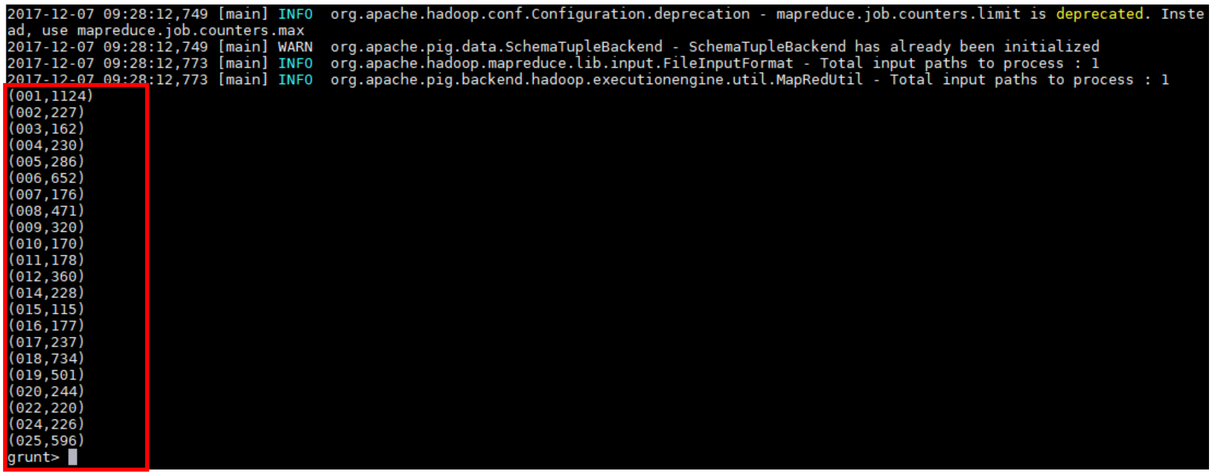
  distict_theft_data = GROUP filter_crime_data_theft_dist BY District;
- **Find the distinct distict_theft_data**

  distinct_district_theft_data = DISTINCT distict_theft_data;
- **Use the below query to find the count of arrests distict wise.**

  arrest_district_wise = FOREACH distinct_district_theft_data GENERATE group, COUNT(filter_crime_data_theft_dist);



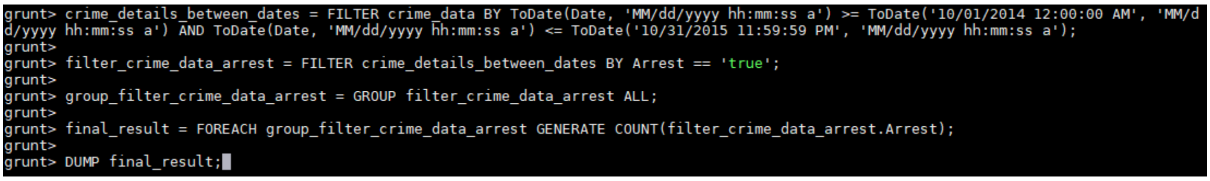
```
grunt> filter_crime_data_theft_dist = FILTER crime_data BY Primary_Type == 'THEFT' and Arrest == 'true';
grunt> distict_theft_data = GROUP filter_crime_data_theft_dist BY District;
grunt> distinct_district_theft_data = DISTINCT distict_theft_data;
grunt> arrest_district_wise = FOREACH distinct_district_theft_data GENERATE group, COUNT(filter_crime_data_theft_dist);
grunt> DUMP arrest_district_wise;
```



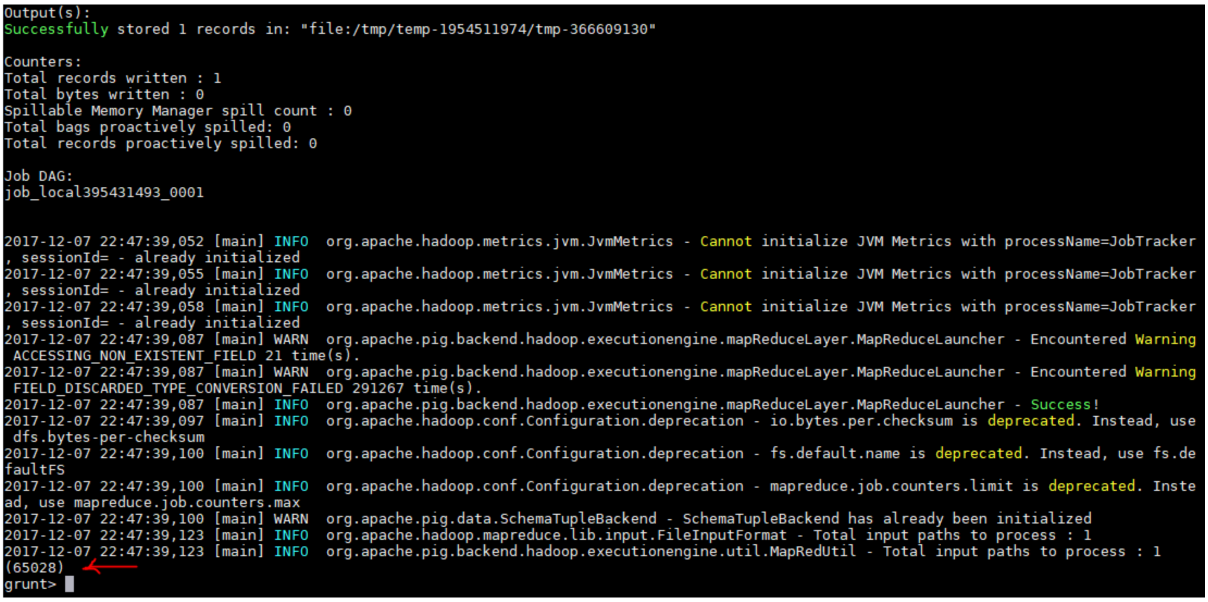
```
2017-12-07 09:28:12,749 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
ad, use mapreduce.job.counters.max
2017-12-07 09:28:12,749 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-07 09:28:12,773 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-07 09:28:12,773 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(001,1124)
(002,227)
(003,162)
(004,230)
(005,286)
(006,652)
(007,176)
(008,471)
(009,320)
(010,170)
(011,178)
(012,360)
(014,228)
(015,115)
(016,177)
(017,237)
(018,734)
(019,501)
(020,244)
(022,220)
(024,226)
(025,596)
grunt>
```

4. Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015.

- **Use below query to find the crime data between October 2014 and October 2015:**
  crime_details_between_dates = FILTER crime_data BY ToDate(Date, 'MM/dd/yyyy hh:mm:ss a') >= ToDate('10/01/2014 12:00:00 AM', 'MM/dd/yyyy hh:mm:ss a') AND ToDate(Date, 'MM/dd/yyyy hh:mm:ss a') <= ToDate('10/31/2015 11:59:59 PM', 'MM/dd/yyyy hh:mm:ss a');
- **Again filter the crime data for arrest made between October 2014 and October 2015**
  filter_crime_data_arrest = FILTER crime_details_between_dates BY Arrest == 'true';
- **Group the above result set with all**
  group_filter_crime_data_arrest = GROUP filter_crime_data_arrest ALL;
- **Finally find the count of arrests made from above group_filter_crime_data_arrest**
  final_result = FOREACH group_filter_crime_data_arrest GENERATE group, COUNT(filter_crime_data_arrest.Arrest);



```
grunt> crime_details_between_dates = FILTER crime_data BY ToDate(Date, 'MM/dd/yyyy hh:mm:ss a') >= ToDate('10/01/2014 12:00:00 AM', 'MM/d
d/yyyy hh:mm:ss a') AND ToDate(Date, 'MM/dd/yyyy hh:mm:ss a') <= ToDate('10/31/2015 11:59:59 PM', 'MM/dd/yyyy hh:mm:ss a');
grunt>
grunt> filter_crime_data_arrest = FILTER crime_details_between_dates BY Arrest == 'true';
grunt>
grunt> group_filter_crime_data_arrest = GROUP filter_crime_data_arrest ALL;
grunt>
grunt> final_result = FOREACH group_filter_crime_data_arrest GENERATE COUNT(filter_crime_data_arrest.Arrest);
grunt>
grunt> DUMP final_result;
```



```
Output(s):
Successfully stored 1 records in: "file:/tmp/temp-1954511974/tmp-366609130"

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_local395431493_0001


2017-12-07 22:47:39,052 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2017-12-07 22:47:39,055 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2017-12-07 22:47:39,058 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2017-12-07 22:47:39,087 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning
 ACCESSING_NON_EXISTENT_FIELD 21 time(s).
2017-12-07 22:47:39,087 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning
 FIELD_DISCARDED_TYPE_CONVERSION_FAILED 291267 time(s).
2017-12-07 22:47:39,087 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-07 22:47:39,097 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2017-12-07 22:47:39,100 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2017-12-07 22:47:39,100 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
ad, use mapreduce.job.counters.max
2017-12-07 22:47:39,100 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-07 22:47:39,123 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-07 22:47:39,123 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(65028)
grunt>
```