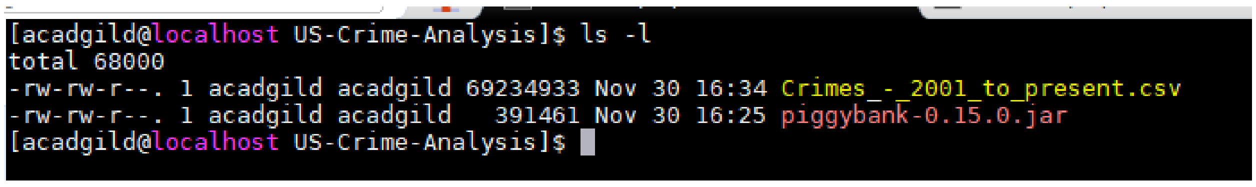
PROJECT-1.1

Project Description- This project is about doing some analysis on crimes taking place in various areas of USA.

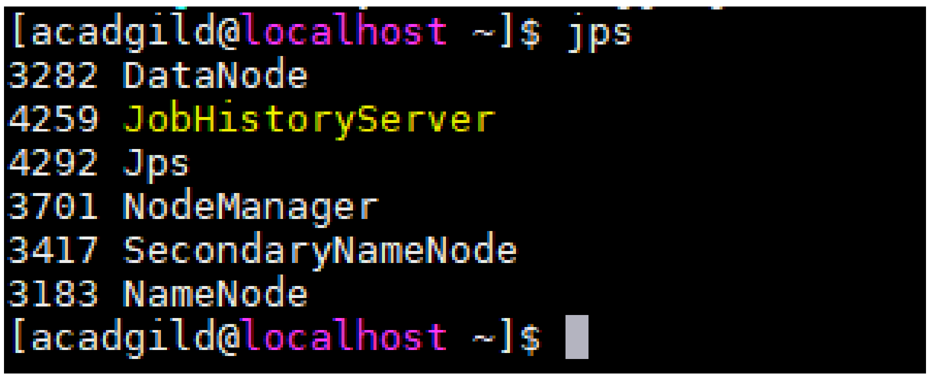
Pre-requisites- Hadoop Cluster should be installed on the machine.

We have to register **piggybank-0.15.0.jar** to do the operation.



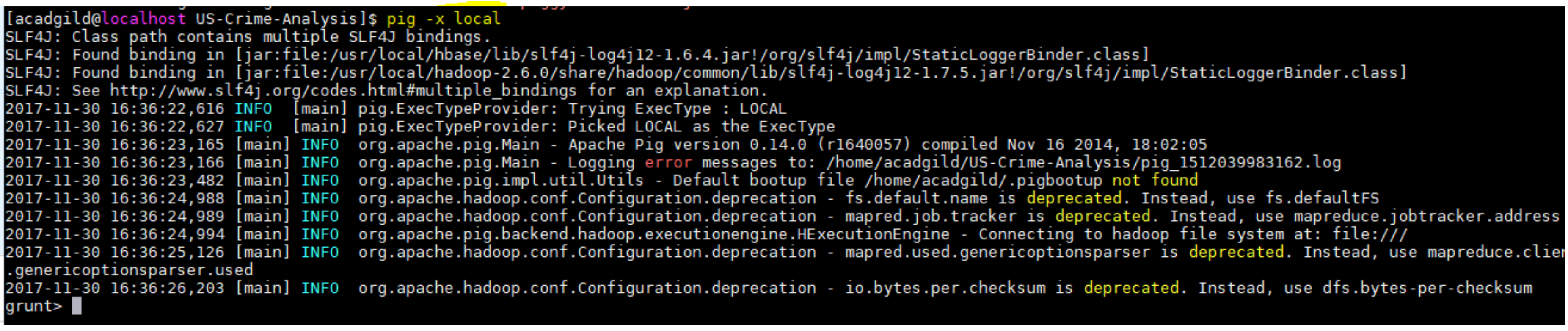
```
[acadgild@localhost US-Crime-Analysis]$ ls -l
total 68000
-rw-rw-r--. 1 acadgild acadgild 69234933 Nov 30 16:34 Crimes_-_2001_to_present.csv
-rw-rw-r--. 1 acadgild acadgild   391461 Nov 30 16:25 piggybank-0.15.0.jar
[acadgild@localhost US-Crime-Analysis]$ 
```

Since we are doing this using Pig so history server must be running. So we have started history Server along with other services-



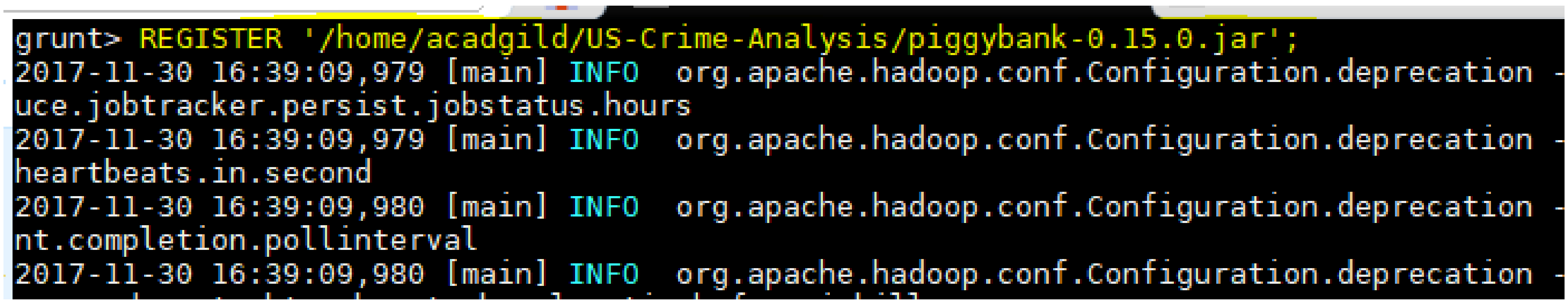
```
[acadgild@localhost ~]$ jps
3282 DataNode
4259 JobHistoryServer
4292 Jps
3701 NodeManager
3417 SecondaryNameNode
3183 NameNode
[acadgild@localhost ~]$ 
```

We are starting pig as shown below-



```
[acadgild@localhost US-Crime-Analysis]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-11-30 16:36:22,616 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-11-30 16:36:22,627 INFO  [main] pig.ExecTypeProvider: Picked LOCAL as the ExecType
2017-11-30 16:36:23,165 [main] INFO  org.apache.pig.Main - Apache Pig version 0.14.0 (r1640057) compiled Nov 16 2014, 18:02:05
2017-11-30 16:36:23,166 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/US-Crime-Analysis/pig_1512039983162.log
2017-11-30 16:36:23,482 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2017-11-30 16:36:24,988 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:36:24,989 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-30 16:36:24,994 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2017-11-30 16:36:25,126 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.used.genericoptionsparser is deprecated. Instead, use mapreduce.clie
.genericoptionsparser.used
2017-11-30 16:36:26,203 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> 
```

Before proceeding the **‘piggybank-0.15.0.jar’** must be registered in grunt shell using below command as shown below-



```
grunt> REGISTER '/home/acadgild/US-Crime-Analysis/piggybank-0.15.0.jar';
2017-11-30 16:39:09,979 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
uce.jobtracker.persist.jobstatus.hours
2017-11-30 16:39:09,979 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
heartbeats.in.second
2017-11-30 16:39:09,980 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
nt.completion.pollinterval
2017-11-30 16:39:09,980 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
```

<u>Task-1-</u> Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code-

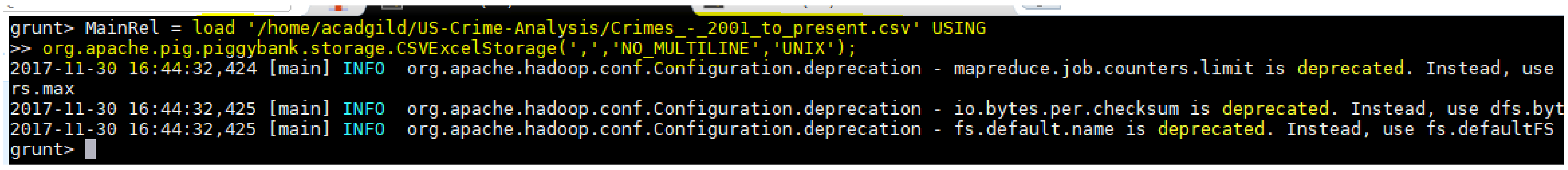
Below is the script used to accomplish the task-

- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**
- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$14 as FBI_Code;**
- **B = GROUP A BY FBI_Code;**
- **C = FOREACH B GENERATE group, COUNT(A.FBI_Code);**
- **Dump C;**

Now we will try to understand each relation one by one-

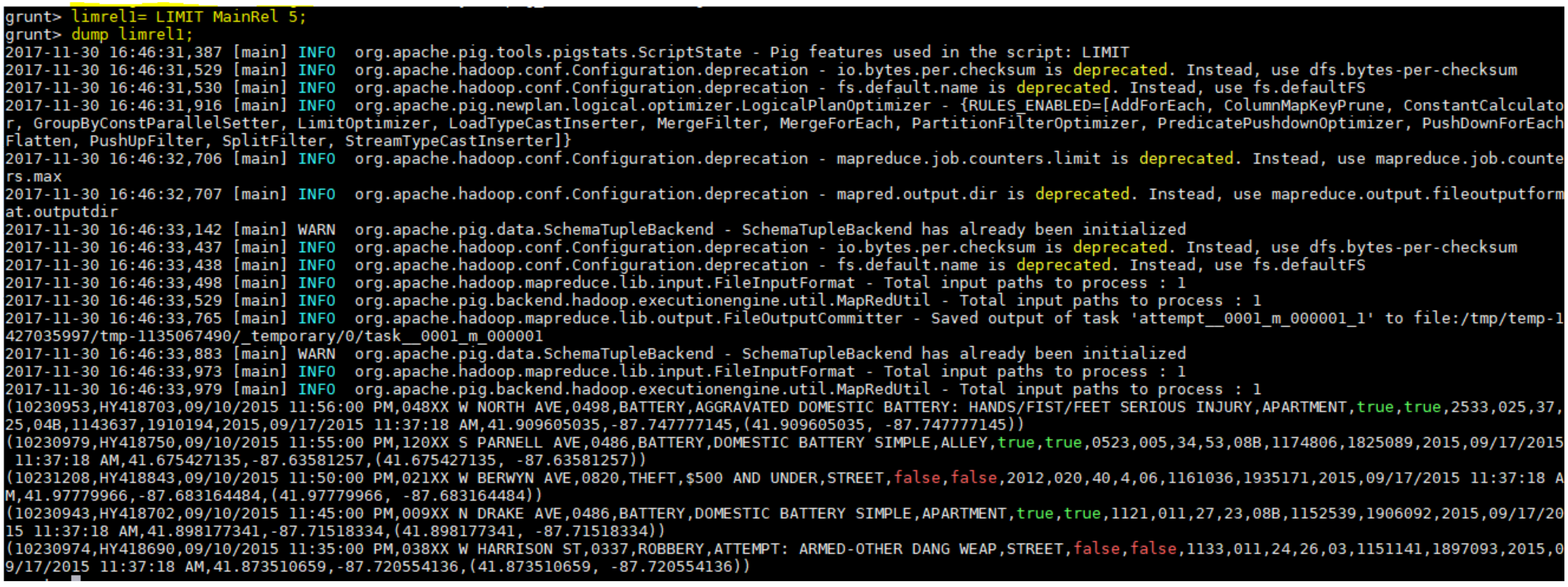
- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**

This one is basically used to Load a comma separated file to Pig Storage-



```
grunt> MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING
>> org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');
2017-11-30 16:44:32,424 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use
rs.max
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.byt
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

If we dump 5 rows of above relation we can very well see that. Below screenshot shows that data has been loaded to pig storage-

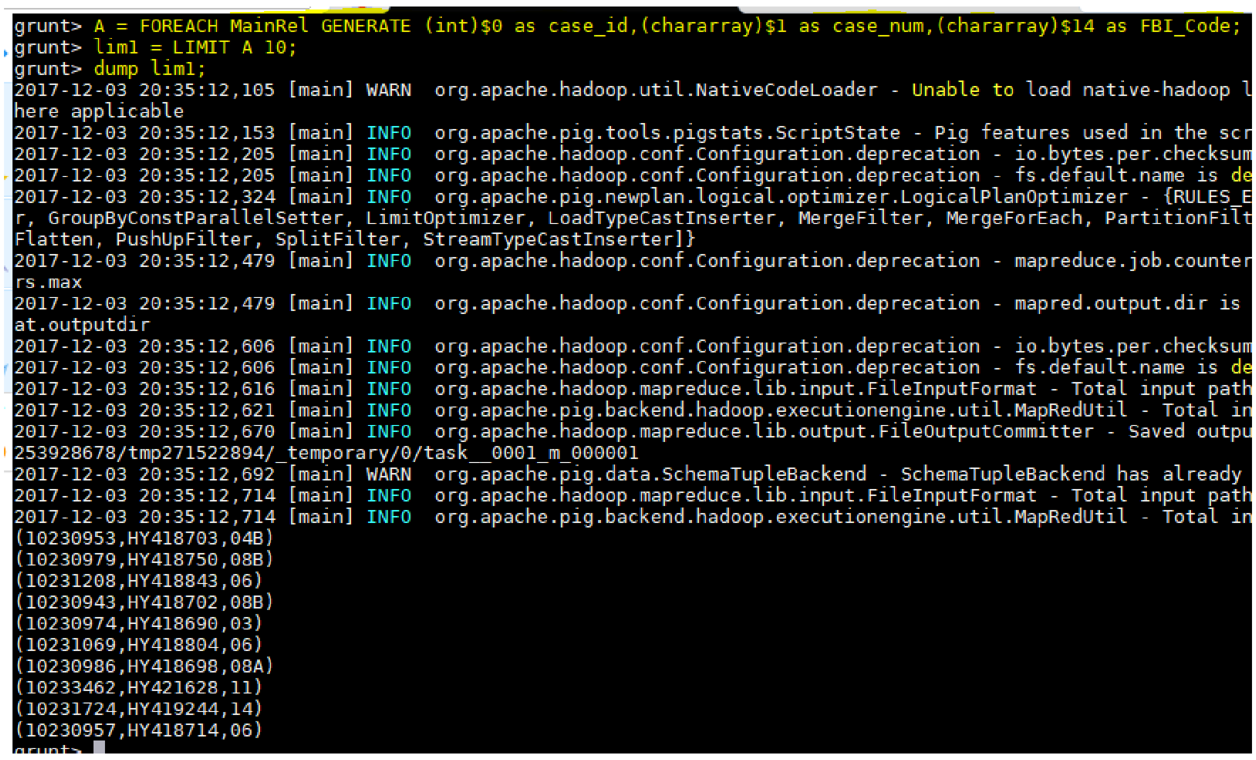


```
grunt> limrel1= LIMIT MainRel 5;
grunt> dump limrel1;
2017-11-30 16:46:31,387 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-11-30 16:46:31,529 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:31,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:31,916 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 16:46:32,706 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counte
rs.max
2017-11-30 16:46:32,707 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputform
at.outputdir
2017-11-30 16:46:33,142 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:33,438 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:33,498 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,529 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-11-30 16:46:33,765 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to file:/tmp/temp-1
427035997/tmp-1135067490/_temporary/0/task__0001_m_000001
2017-11-30 16:46:33,883 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,973 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,979 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10230953,HY418703,09/10/2015 11:56:00 PM,048XX W NORTH AVE,0498,BATTERY,AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY,APARTMENT,true,true,2533,025,37,
25,04B,1143637,1910194,2015,09/17/2015 11:37:18 AM,41.909605035,-87.747777145,(41.909605035, -87.747777145))
(10230979,HY418750,09/10/2015 11:55:00 PM,120XX S PARNELL AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,ALLEY,true,true,0523,005,34,53,08B,1174806,1825089,2015,09/17/2015
 11:37:18 AM,41.675427135,-87.63581257,(41.675427135, -87.63581257))
(10231208,HY418843,09/10/2015 11:50:00 PM,021XX W BERWYN AVE,0820,THEFT,$500 AND UNDER,STREET,false,false,2012,020,40,4,06,1161036,1935171,2015,09/17/2015 11:37:18 A
M,41.97779966,-87.683164484,(41.97779966, -87.683164484))
(10230943,HY418702,09/10/2015 11:45:00 PM,009XX N DRAKE AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,APARTMENT,true,true,1121,011,27,23,08B,1152539,1906092,2015,09/17/20
15 11:37:18 AM,41.898177341,-87.71518334,(41.898177341, -87.71518334))
(10230974,HY418690,09/10/2015 11:35:00 PM,038XX W HARRISON ST,0337,ROBBERY,ATTEMPT: ARMED-OTHER DANG WEAP,STREET,false,false,1133,011,24,26,03,1151141,1897093,2015,0
9/17/2015 11:37:18 AM,41.873510659,-87.720554136,(41.873510659, -87.720554136))
```

- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$14 as FBI_Code;**

The above script shows that we are extracting/generating some particular columns from the loaded file. The columns are case_id, case_num and FBI_Code.

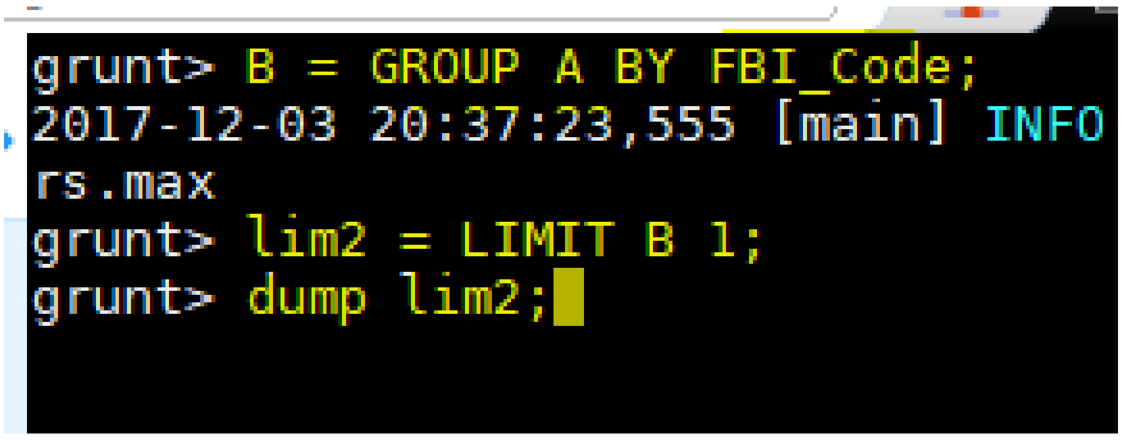
The same can be seen in below screenshot-



```
grunt> A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$14 as FBI_Code;
grunt> lim1 = LIMIT A 10;
grunt> dump lim1;
2017-12-03 20:35:12,105 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop l
here applicable
2017-12-03 20:35:12,153 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the scr
2017-12-03 20:35:12,205 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum
2017-12-03 20:35:12,205 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
2017-12-03 20:35:12,324 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_E
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilt
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-03 20:35:12,479 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counter
rs.max
2017-12-03 20:35:12,479 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is 
at.outputdir
2017-12-03 20:35:12,606 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum
2017-12-03 20:35:12,606 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
2017-12-03 20:35:12,616 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input path
2017-12-03 20:35:12,621 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total in
2017-12-03 20:35:12,670 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved outpu
253928678/tmp271522894/_temporary/0/task__0001_m_000001
2017-12-03 20:35:12,692 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already 
2017-12-03 20:35:12,714 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input path
2017-12-03 20:35:12,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total in
(10230953,HY418703,04B)
(10230979,HY418750,08B)
(10231208,HY418843,06)
(10230943,HY418702,08B)
(10230974,HY418690,03)
(10231069,HY418804,06)
(10230986,HY418698,08A)
(10233462,HY421628,11)
(10231724,HY419244,14)
(10230957,HY418714,06)
grunt>
```

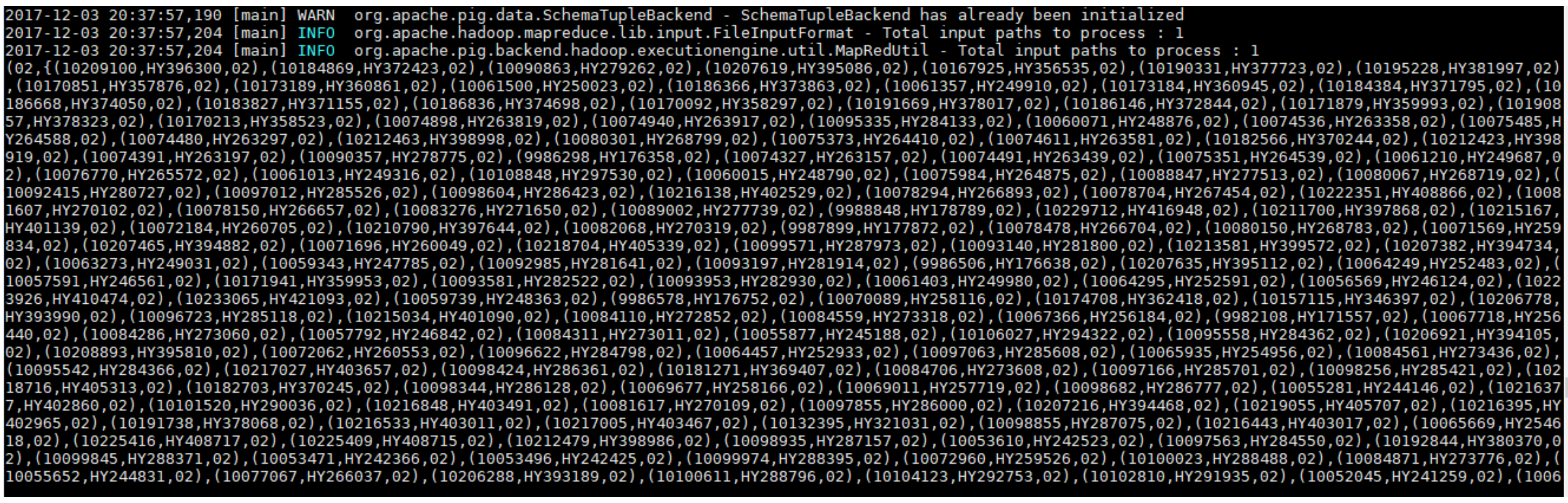
- **B = GROUP A BY FBI_Code;**

Now we are grouping the relation A by FBI_Code just to find the count of cases investigated under each FBI_Code going further-



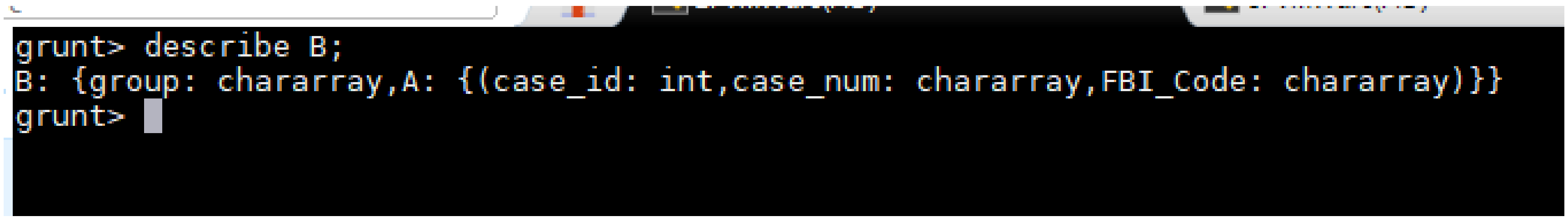
```
grunt> B = GROUP A BY FBI_Code;
2017-12-03 20:37:23,555 [main] INFO
rs.max
grunt> lim2 = LIMIT B 1;
grunt> dump lim2;
```

If we dump the relation B for grouped data we can see below group generated for the first FBI_Code 02-



```
2017-12-03 20:37:57,190 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-03 20:37:57,204 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-03 20:37:57,204 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(02,{(10209100,HY396300,02),(10184869,HY372423,02),(10090863,HY279262,02),(10207619,HY395086,02),(10167925,HY356535,02),(10190331,HY377723,02),(10195228,HY381997,02)
,(10170851,HY357876,02),(10173189,HY360861,02),(10061500,HY250023,02),(10186366,HY373863,02),(10061357,HY249910,02),(10173184,HY360945,02),(10184384,HY371795,02),(10
186668,HY374050,02),(10183827,HY371155,02),(10186836,HY374698,02),(10170092,HY358297,02),(10191669,HY378017,02),(10186146,HY372844,02),(10171879,HY359993,02),(101908
57,HY378323,02),(10170213,HY358523,02),(10074898,HY263819,02),(10074940,HY263917,02),(10095335,HY284133,02),(10060071,HY248876,02),(10074536,HY263358,02),(10075485,H
Y264588,02),(10074480,HY263297,02),(10212463,HY398998,02),(10080301,HY268799,02),(10075373,HY264410,02),(10074611,HY263581,02),(10182566,HY370244,02),(10212423,HY398
919,02),(10074391,HY263197,02),(10090357,HY278775,02),(9986298,HY176358,02),(10074327,HY263157,02),(10074491,HY263439,02),(10075351,HY264539,02),(10061210,HY249687,0
2),(10076770,HY265572,02),(10061013,HY249316,02),(10108848,HY297530,02),(10060015,HY248790,02),(10075984,HY264875,02),(10088847,HY277513,02),(10080067,HY268719,02),(
10092415,HY280727,02),(10097012,HY285526,02),(10098604,HY286423,02),(10216138,HY402529,02),(10078294,HY266893,02),(10078704,HY267454,02),(10222351,HY408866,02),(1008
1607,HY270102,02),(10078150,HY266657,02),(10083276,HY271650,02),(10089002,HY277739,02),(9988848,HY178789,02),(10229712,HY416948,02),(10211700,HY397868,02),(10215167,
HY401139,02),(10072184,HY260705,02),(10210790,HY397644,02),(10082068,HY270319,02),(9987899,HY177872,02),(10078478,HY266704,02),(10080150,HY268783,02),(10071569,HY259
834,02),(10207465,HY394882,02),(10071696,HY260049,02),(10218704,HY405339,02),(10099571,HY287973,02),(10093140,HY281800,02),(10213581,HY399572,02),(10207382,HY394734,
02),(10063273,HY249031,02),(10059343,HY247785,02),(10092985,HY281641,02),(10093197,HY281914,02),(9986506,HY176638,02),(10207635,HY395112,02),(10064249,HY252483,02),(
10057591,HY246561,02),(10171941,HY359953,02),(10093581,HY282522,02),(10093953,HY282930,02),(10061403,HY249980,02),(10064295,HY252591,02),(10056569,HY246124,02),(1022
3926,HY410474,02),(10233065,HY421093,02),(10059739,HY248363,02),(9986578,HY176752,02),(10070089,HY258116,02),(10174708,HY362418,02),(10157115,HY346397,02),(10206778,
HY393990,02),(10096723,HY285118,02),(10215034,HY401090,02),(10084110,HY272852,02),(10084559,HY273318,02),(10067366,HY256184,02),(9982108,HY171557,02),(10067718,HY256
440,02),(10084286,HY273060,02),(10057792,HY246842,02),(10084311,HY273011,02),(10055877,HY245188,02),(10106027,HY294322,02),(10095558,HY284362,02),(10206921,HY394105,
02),(10208893,HY395810,02),(10072062,HY260553,02),(10096622,HY284798,02),(10064457,HY252933,02),(10097063,HY285608,02),(10065935,HY254956,02),(10084561,HY273436,02),
(10095542,HY284366,02),(10217027,HY403657,02),(10098424,HY286361,02),(10181271,HY369407,02),(10084706,HY273608,02),(10097166,HY285701,02),(10098256,HY285421,02),(102
18716,HY405313,02),(10182703,HY370245,02),(10098344,HY286128,02),(10069677,HY258166,02),(10069011,HY257719,02),(10098682,HY286777,02),(10055281,HY244146,02),(1021637
7,HY402860,02),(10101520,HY290036,02),(10216848,HY403491,02),(10081617,HY270109,02),(10097855,HY286000,02),(10207216,HY394468,02),(10219055,HY405707,02),(10216395,HY
402965,02),(10191738,HY378068,02),(10216533,HY403011,02),(10217005,HY403467,02),(10132395,HY321031,02),(10098855,HY287075,02),(10216443,HY403017,02),(10065669,HY2546
18,02),(10225416,HY408717,02),(10225409,HY408715,02),(10212479,HY398986,02),(10098935,HY287157,02),(10053610,HY242523,02),(10097563,HY284550,02),(10192844,HY380370,0
2),(10099845,HY288371,02),(10053471,HY242366,02),(10053496,HY242425,02),(10099974,HY288395,02),(10072960,HY259526,02),(10100023,HY288488,02),(10084871,HY273776,02),(
10055652,HY244831,02),(10077067,HY266037,02),(10206288,HY393189,02),(10100611,HY288796,02),(10104123,HY292753,02),(10102810,HY291935,02),(10052045,HY241259,02),(1006
```

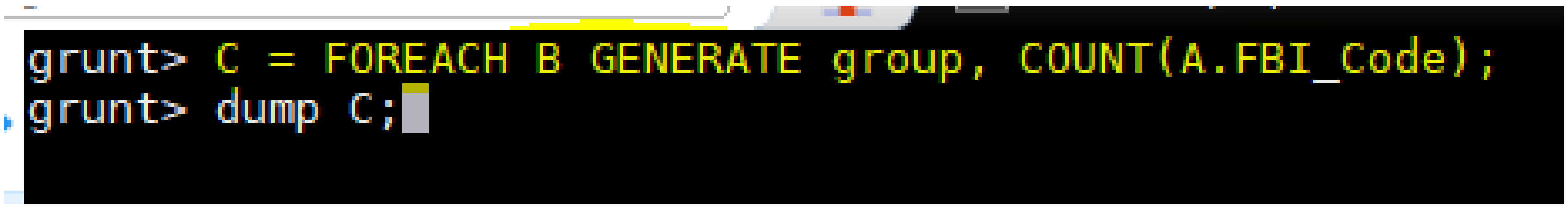
If we describe B relation we can very well see the schema of relation -



```
grunt> describe B;
B: {group: chararray,A: {(case_id: int,case_num: chararray,FBI_Code: chararray)}}
grunt>
```

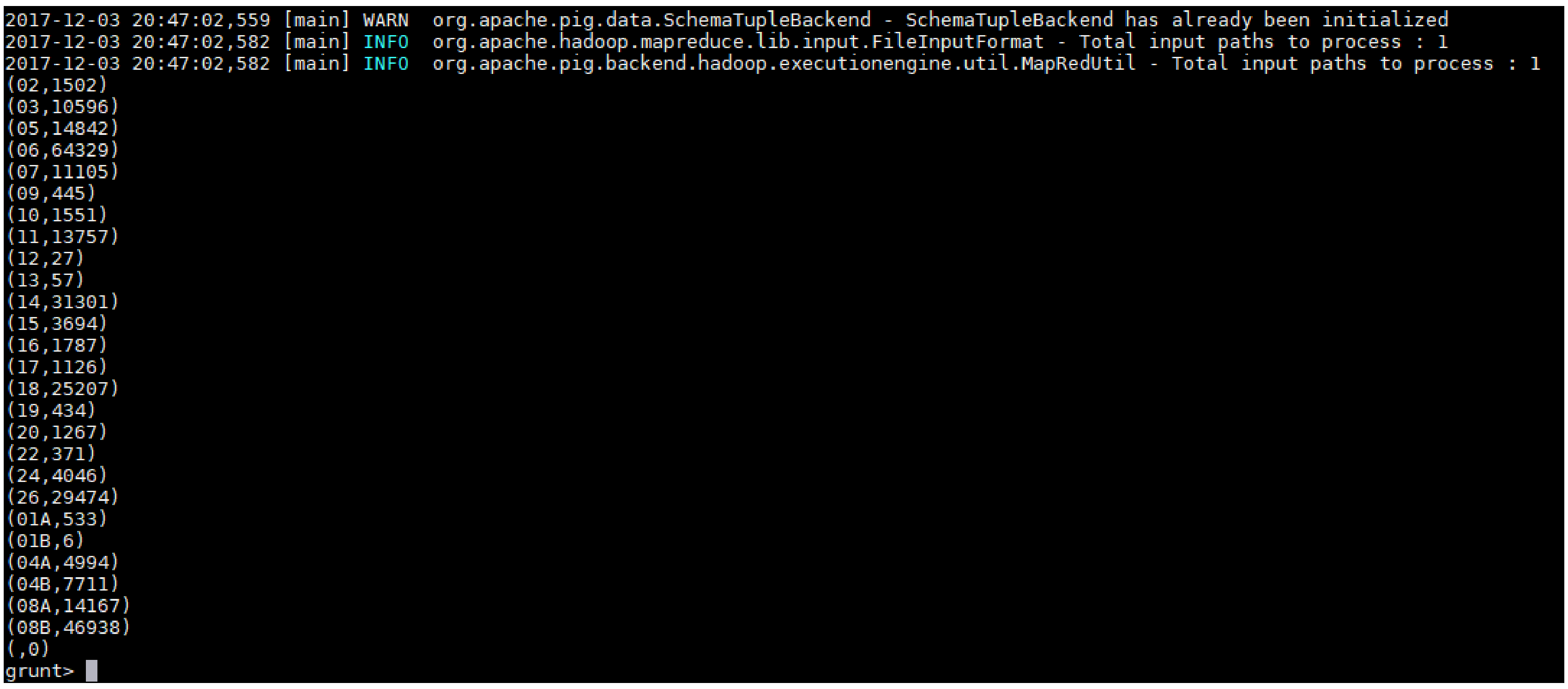
- **C = FOREACH B GENERATE group, COUNT(A.FBI_Code);**
- **Dump C;**

Now here we are generating count of each FBI_Code which was grouped in previous relation.



```
grunt> C = FOREACH B GENERATE group, COUNT(A.FBI_Code);
grunt> dump C;
```

Below screenshot shows the FBI_Code and count for the same-



```
2017-12-03 20:47:02,559 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-03 20:47:02,582 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-03 20:47:02,582 [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> 
```

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

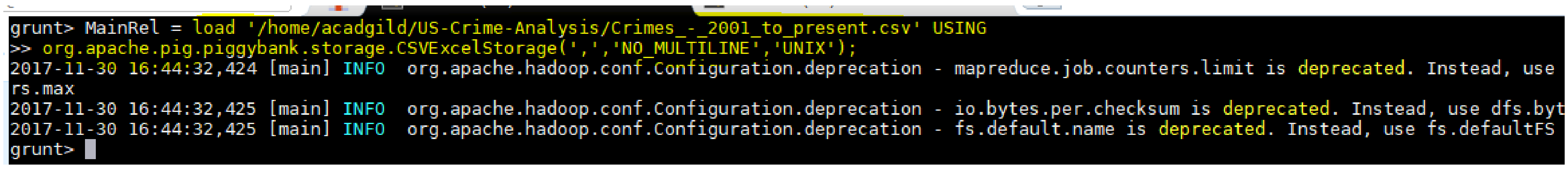
Below is the script used to find the number of cases investigated under FBI code 32-

- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**
- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$14 as FBI_Code;**
- **B = FILTER A BY FBI_Code == '32';**
- **Dump B;**

Now we will understand each relation one by one.

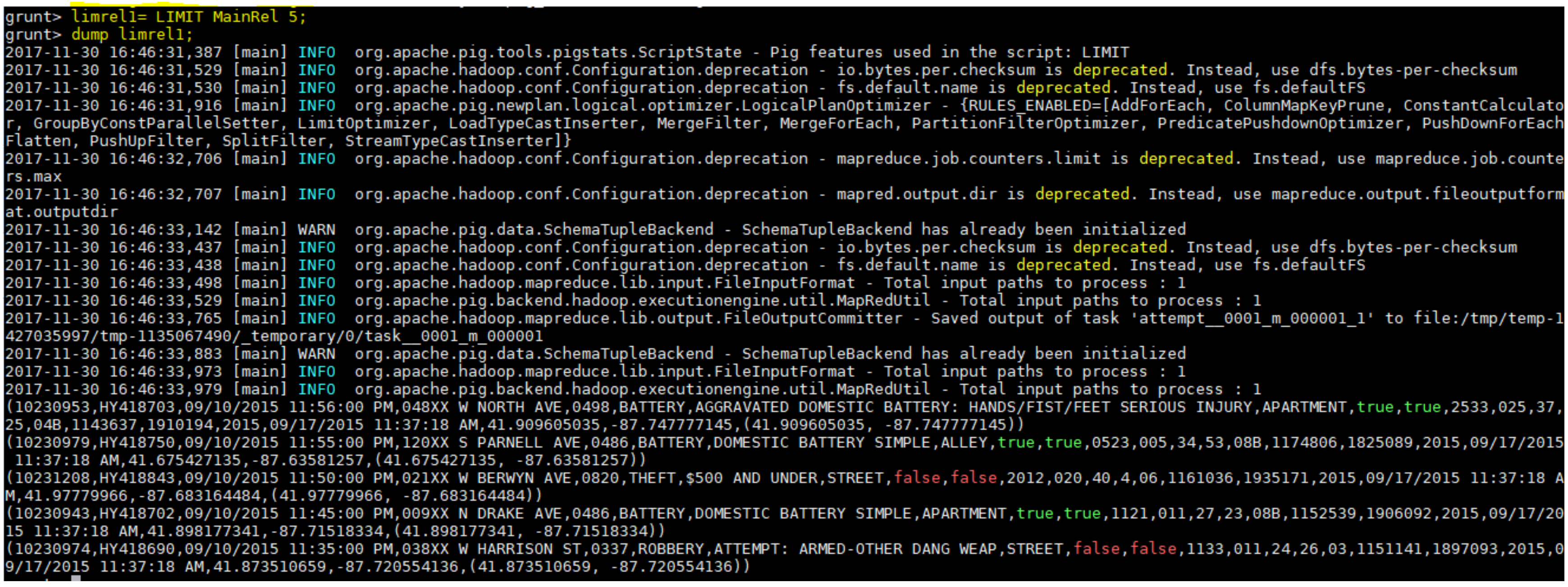
- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**

Here in above relation we are loading the comma separated file to Pig Storage-



```
grunt> MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING
>> org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');
2017-11-30 16:44:32,424 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use
rs.max
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.byt
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> 
```

Below screenshot shows that file has been loaded successfully

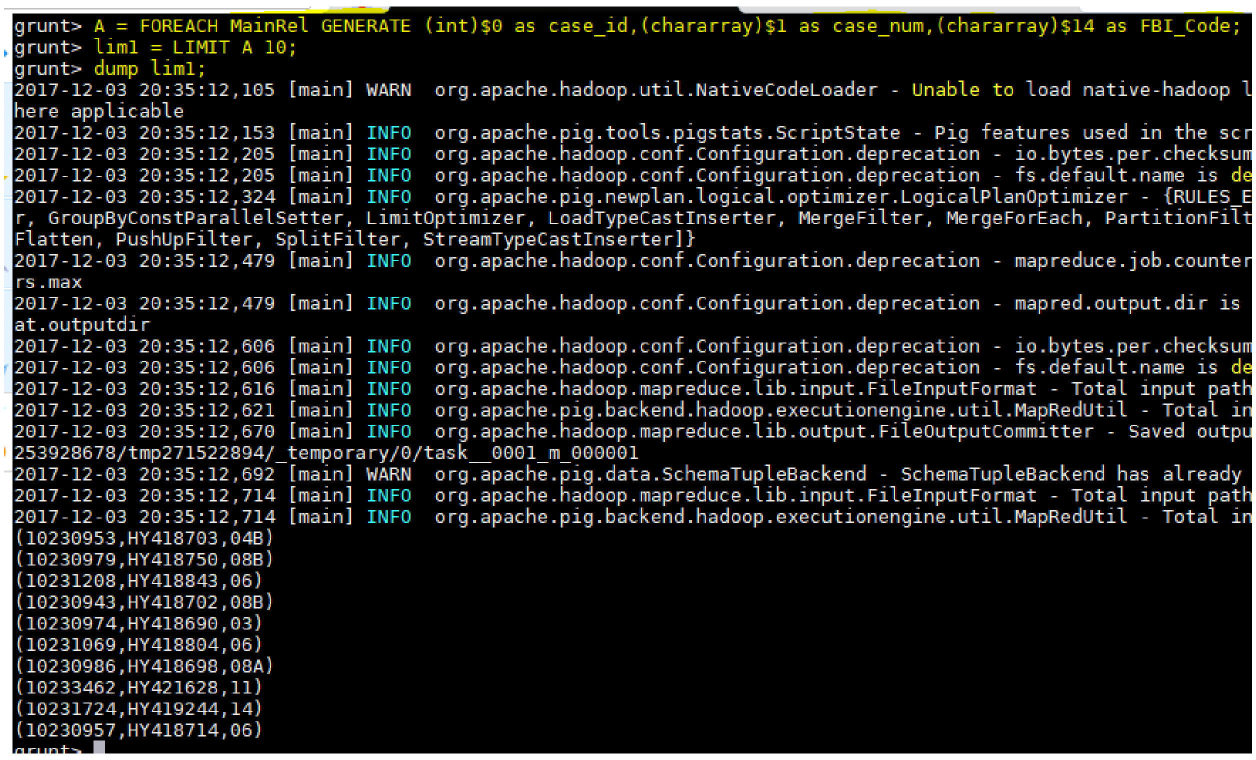


```
grunt> limrel1= LIMIT MainRel 5;
grunt> dump limrel1;
2017-11-30 16:46:31,387 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-11-30 16:46:31,529 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:31,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:31,916 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 16:46:32,706 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counte
rs.max
2017-11-30 16:46:32,707 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputform
at.outputdir
2017-11-30 16:46:33,142 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:33,438 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:33,498 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,529 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-11-30 16:46:33,765 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to file:/tmp/temp-1
427035997/tmp-1135067490/_temporary/0/task__0001_m_000001
2017-11-30 16:46:33,883 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,973 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,979 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10230953,HY418703,09/10/2015 11:56:00 PM,048XX W NORTH AVE,0498,BATTERY,AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY,APARTMENT,true,true,2533,025,37,
25,04B,1143637,1910194,2015,09/17/2015 11:37:18 AM,41.909605035,-87.747777145,(41.909605035, -87.747777145))
(10230979,HY418750,09/10/2015 11:55:00 PM,120XX S PARNELL AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,ALLEY,true,true,0523,005,34,53,08B,1174806,1825089,2015,09/17/2015
 11:37:18 AM,41.675427135,-87.63581257,(41.675427135, -87.63581257))
(10231208,HY418843,09/10/2015 11:50:00 PM,021XX W BERWYN AVE,0820,THEFT,$500 AND UNDER,STREET,false,false,2012,020,40,4,06,1161036,1935171,2015,09/17/2015 11:37:18 A
M,41.97779966,-87.683164484,(41.97779966, -87.683164484))
(10230943,HY418702,09/10/2015 11:45:00 PM,009XX N DRAKE AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,APARTMENT,true,true,1121,011,27,23,08B,1152539,1906092,2015,09/17/20
15 11:37:18 AM,41.898177341,-87.71518334,(41.898177341, -87.71518334))
(10230974,HY418690,09/10/2015 11:35:00 PM,038XX W HARRISON ST,0337,ROBBERY,ATTEMPT: ARMED-OTHER DANG WEAP,STREET,false,false,1133,011,24,26,03,1151141,1897093,2015,0
9/17/2015 11:37:18 AM,41.873510659,-87.720554136,(41.873510659, -87.720554136))
```

- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$14 as FBI_Code;**

In above script we are extracting/generating 3 columns from the file loaded. These 3 columns are case_id, case_num, FBI_Code.

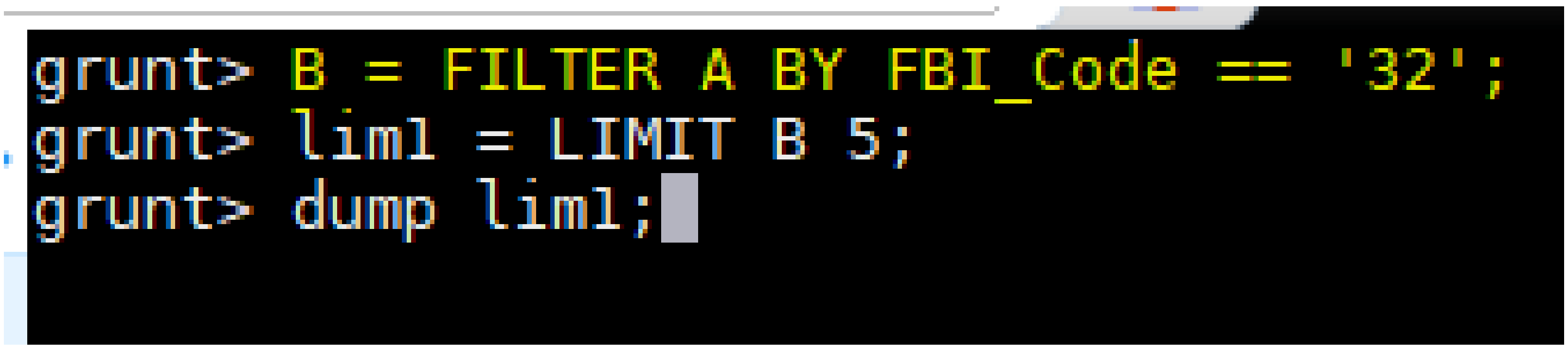
Below screenshot shows the same and also some dumped data for the same-



```
grunt> A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$14 as FBI_Code;
grunt> lim1 = LIMIT A 10;
grunt> dump lim1;
2017-12-03 20:35:12,105 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop l
here applicable
2017-12-03 20:35:12,153 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the scr
2017-12-03 20:35:12,205 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum
2017-12-03 20:35:12,205 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
2017-12-03 20:35:12,324 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_E
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilt
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-03 20:35:12,479 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counter
rs.max
2017-12-03 20:35:12,479 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is
at.outputdir
2017-12-03 20:35:12,606 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum
2017-12-03 20:35:12,606 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
2017-12-03 20:35:12,616 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input path
2017-12-03 20:35:12,621 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total in
2017-12-03 20:35:12,670 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved outpu
253928678/tmp271522894/_temporary/0/task__0001_m_000001
2017-12-03 20:35:12,692 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already
2017-12-03 20:35:12,714 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input path
2017-12-03 20:35:12,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total in
(10230953,HY418703,04B)
(10230979,HY418750,08B)
(10231208,HY418843,06)
(10230943,HY418702,08B)
(10230974,HY418690,03)
(10231069,HY418804,06)
(10230986,HY418698,08A)
(10233462,HY421628,11)
(10231724,HY419244,14)
(10230957,HY418714,06)
grunt>
```

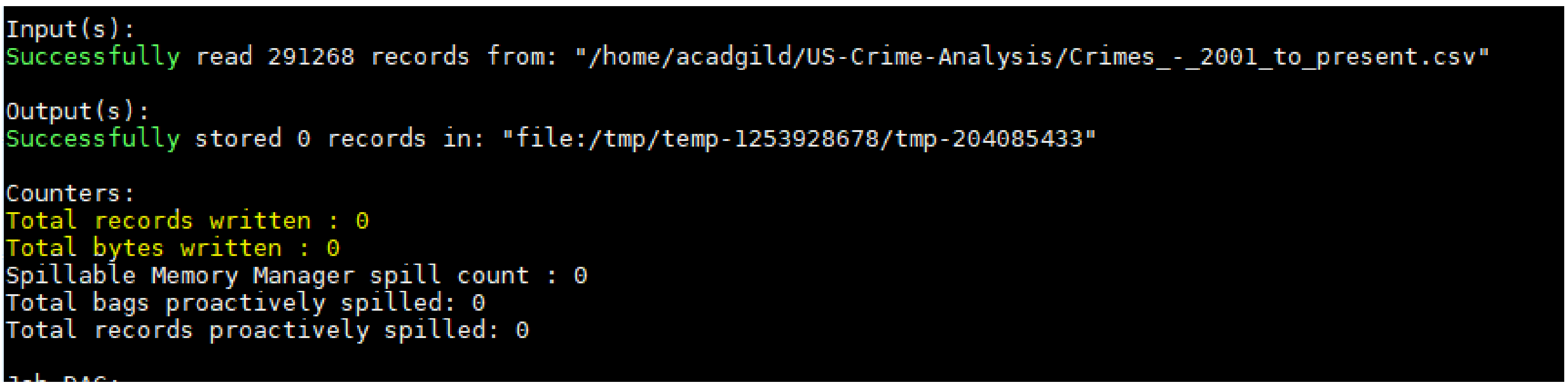
- **B = FILTER A BY FBI_Code == ‘32’;**

Now we will Filter above relation with FBI_Code 32.



```
grunt> B = FILTER A BY FBI_Code == '32';
grunt> lim1 = LIMIT B 5;
grunt> dump lim1;
```

Now after filtering with FBI_Code 32 we can find that the output of this relation is 0. Which means there are no records with FBI_Code =32 in the file as shown below-



```
Input(s):
Successfully read 291268 records from: "/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv"

Output(s):
Successfully stored 0 records in: "file:/tmp/temp-1253928678/tmp-204085433"

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
```

So the number of cases investigated under FBI Code 32 is 0.

P.S- After consulting with support I got to know that for Task-2 we have to calculate the number of cases investigated under Ward == '32'. So adding the solution for that as well-

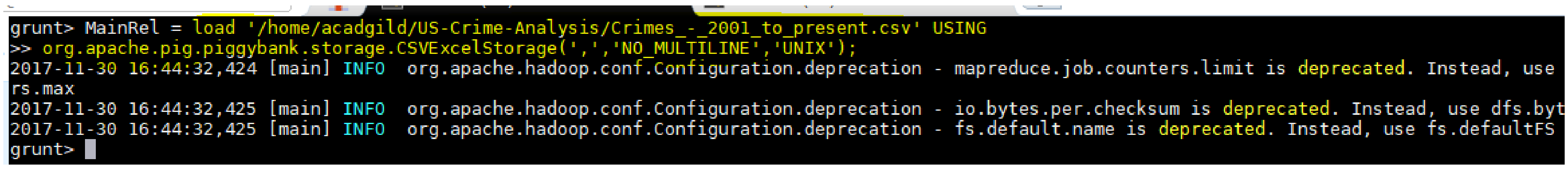
Below is the script used to find the number of cases investigated under Ward=32 –

- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**
- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num, (chararray)$12 as Ward;**
- **B = FILTER A BY Ward == '32';**
- **C = GROUP B by Ward;**
- **D = FOREACH C GENERATE group, COUNT(B.Ward);**
- **Dump D;**

Now let's go in details for each script one by one-

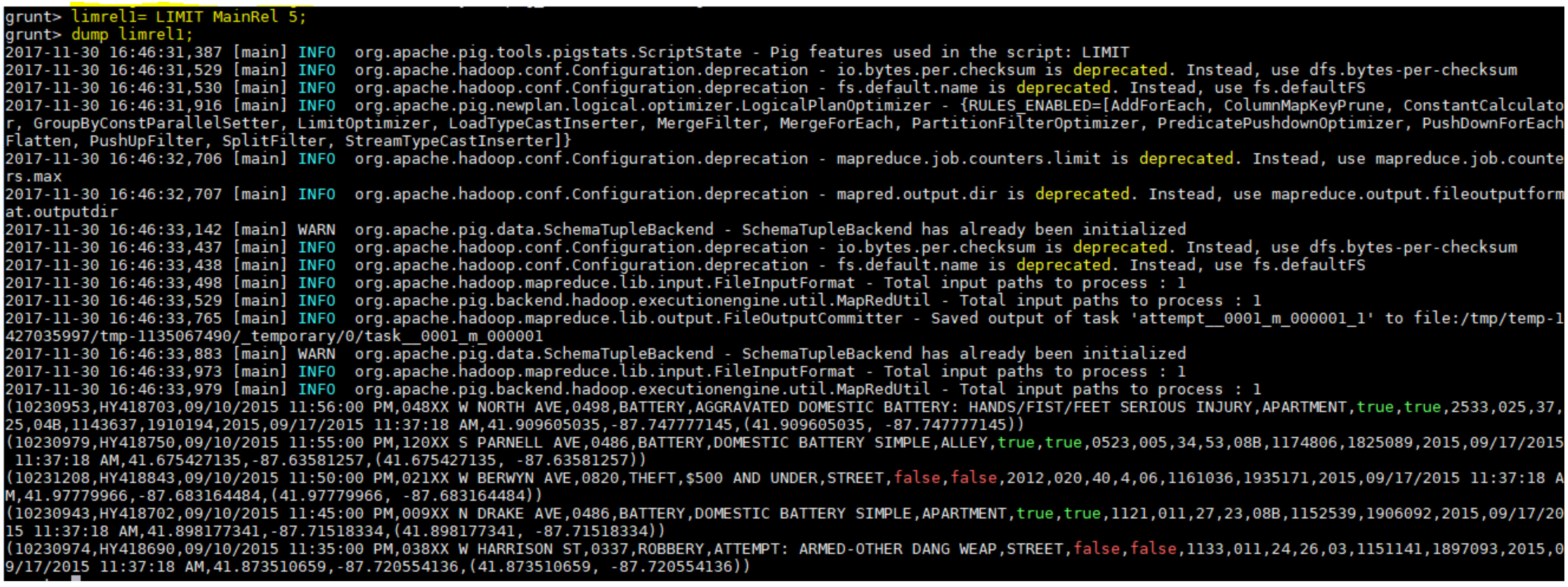
- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**

This one is basically used to Load a comma separated file to Pig Storage-



```
grunt> MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING
>> org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');
2017-11-30 16:44:32,424 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use
rs.max
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.byt
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> 
```

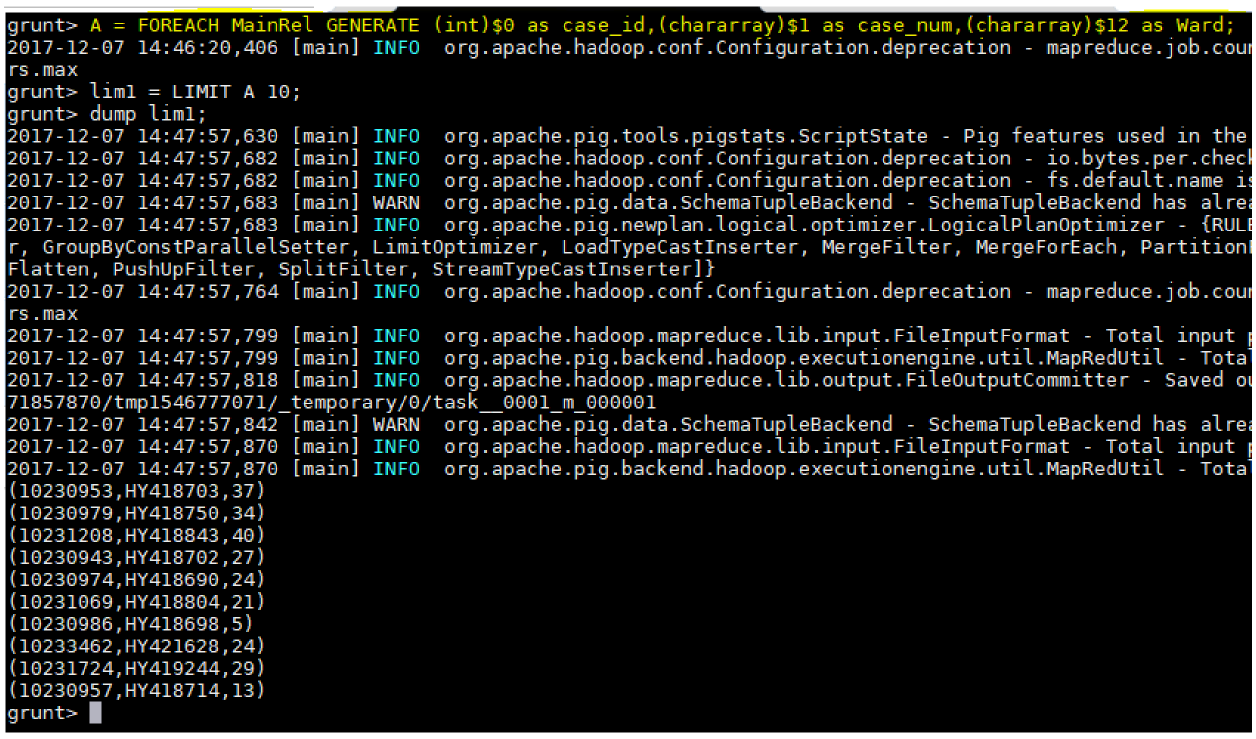
If we dump 5 rows of above relation we can very well see that. Below screenshot shows that data has been loaded to pig storage-



```
grunt> limrel1= LIMIT MainRel 5;
grunt> dump limrel1;
2017-11-30 16:46:31,387 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-11-30 16:46:31,529 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:31,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:31,916 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 16:46:32,706 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counte
rs.max
2017-11-30 16:46:32,707 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputform
at.outputdir
2017-11-30 16:46:33,142 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:33,438 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:33,498 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,529 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-11-30 16:46:33,765 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to file:/tmp/temp-1
427035997/tmp-1135067490/_temporary/0/task__0001_m_000001
2017-11-30 16:46:33,883 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,973 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,979 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10230953,HY418703,09/10/2015 11:56:00 PM,048XX W NORTH AVE,0498,BATTERY,AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY,APARTMENT,true,true,2533,025,37,
25,04B,1143637,1910194,2015,09/17/2015 11:37:18 AM,41.909605035,-87.747777145,(41.909605035, -87.747777145))
(10230979,HY418750,09/10/2015 11:55:00 PM,120XX S PARNELL AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,ALLEY,true,true,0523,005,34,53,08B,1174806,1825089,2015,09/17/2015
 11:37:18 AM,41.675427135,-87.63581257,(41.675427135, -87.63581257))
(10231208,HY418843,09/10/2015 11:50:00 PM,021XX W BERWYN AVE,0820,THEFT,$500 AND UNDER,STREET,false,false,2012,020,40,4,06,1161036,1935171,2015,09/17/2015 11:37:18 A
M,41.97779966,-87.683164484,(41.97779966, -87.683164484))
(10230943,HY418702,09/10/2015 11:45:00 PM,009XX N DRAKE AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,APARTMENT,true,true,1121,011,27,23,08B,1152539,1906092,2015,09/17/20
15 11:37:18 AM,41.898177341,-87.71518334,(41.898177341, -87.71518334))
(10230974,HY418690,09/10/2015 11:35:00 PM,038XX W HARRISON ST,0337,ROBBERY,ATTEMPT: ARMED-OTHER DANG WEAP,STREET,false,false,1133,011,24,26,03,1151141,1897093,2015,0
9/17/2015 11:37:18 AM,41.873510659,-87.720554136,(41.873510659, -87.720554136))
```

- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num, (chararray)$12 as Ward;**

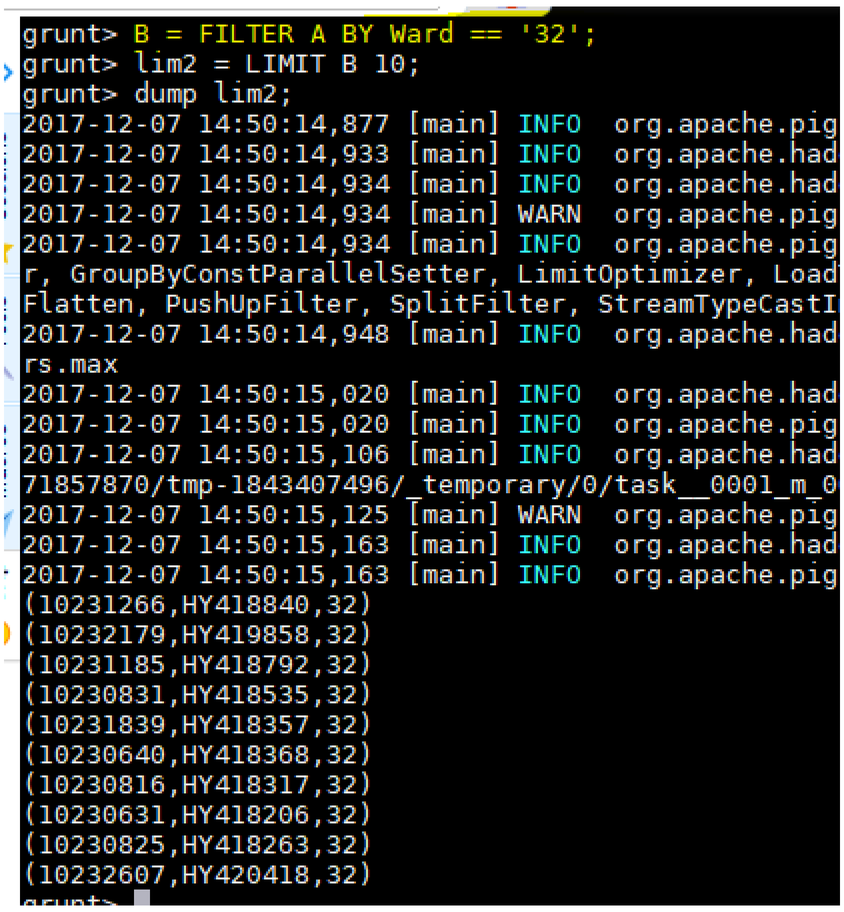
Here we are extracting/generating particular columns from the file loaded- case_id, case_num, Ward in this case-



```
grunt> A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$1 as case_num,(chararray)$12 as Ward;
2017-12-07 14:46:20,406 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.cou
rs.max
grunt> lim1 = LIMIT A 10;
grunt> dump lim1;
2017-12-07 14:47:57,630 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the
2017-12-07 14:47:57,682 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.chec
2017-12-07 14:47:57,682 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name i
2017-12-07 14:47:57,683 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alre
2017-12-07 14:47:57,683 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RUL
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, Partition
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-07 14:47:57,764 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.cou
rs.max
2017-12-07 14:47:57,799 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input
2017-12-07 14:47:57,799 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Tota
2017-12-07 14:47:57,818 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved o
71857870/tmp1546777071/_temporary/0/task__0001_m_000001
2017-12-07 14:47:57,842 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alre
2017-12-07 14:47:57,870 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input
2017-12-07 14:47:57,870 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Tota
(10230953,HY418703,37)
(10230979,HY418750,34)
(10231208,HY418843,40)
(10230943,HY418702,27)
(10230974,HY418690,24)
(10231069,HY418804,21)
(10230986,HY418698,5)
(10233462,HY421628,24)
(10231724,HY419244,29)
(10230957,HY418714,13)
grunt>
```

- **B = FILTER A BY Ward == '32';**

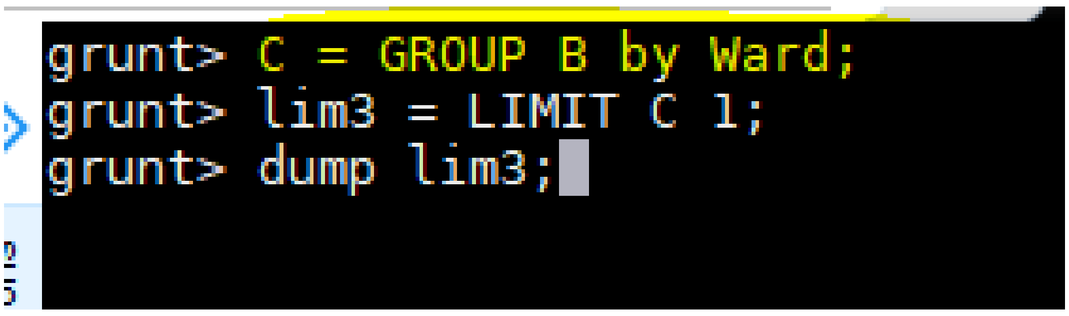
Now using above script we are filtering the relation A with Ward==32, since we are interested in calculating the crime investigated for ward 32 only.



```
grunt> B = FILTER A BY Ward == '32';
grunt> lim2 = LIMIT B 10;
grunt> dump lim2;
2017-12-07 14:50:14,877 [main] INFO  org.apache.pig
2017-12-07 14:50:14,933 [main] INFO  org.apache.had
2017-12-07 14:50:14,934 [main] INFO  org.apache.had
2017-12-07 14:50:14,934 [main] WARN  org.apache.pig
2017-12-07 14:50:14,934 [main] INFO  org.apache.pig
r, GroupByConstParallelSetter, LimitOptimizer, Load
Flatten, PushUpFilter, SplitFilter, StreamTypeCastI
2017-12-07 14:50:14,948 [main] INFO  org.apache.had
rs.max
2017-12-07 14:50:15,020 [main] INFO  org.apache.had
2017-12-07 14:50:15,020 [main] INFO  org.apache.pig
2017-12-07 14:50:15,106 [main] INFO  org.apache.had
71857870/tmp-1843407496/_temporary/0/task__0001_m_0
2017-12-07 14:50:15,125 [main] WARN  org.apache.pig
2017-12-07 14:50:15,163 [main] INFO  org.apache.had
2017-12-07 14:50:15,163 [main] INFO  org.apache.pig
(10231266,HY418840,32)
(10232179,HY419858,32)
(10231185,HY418792,32)
(10230831,HY418535,32)
(10231839,HY418357,32)
(10230640,HY418368,32)
(10230816,HY418317,32)
(10230631,HY418206,32)
(10230825,HY418263,32)
(10232607,HY420418,32)
grunt>
```

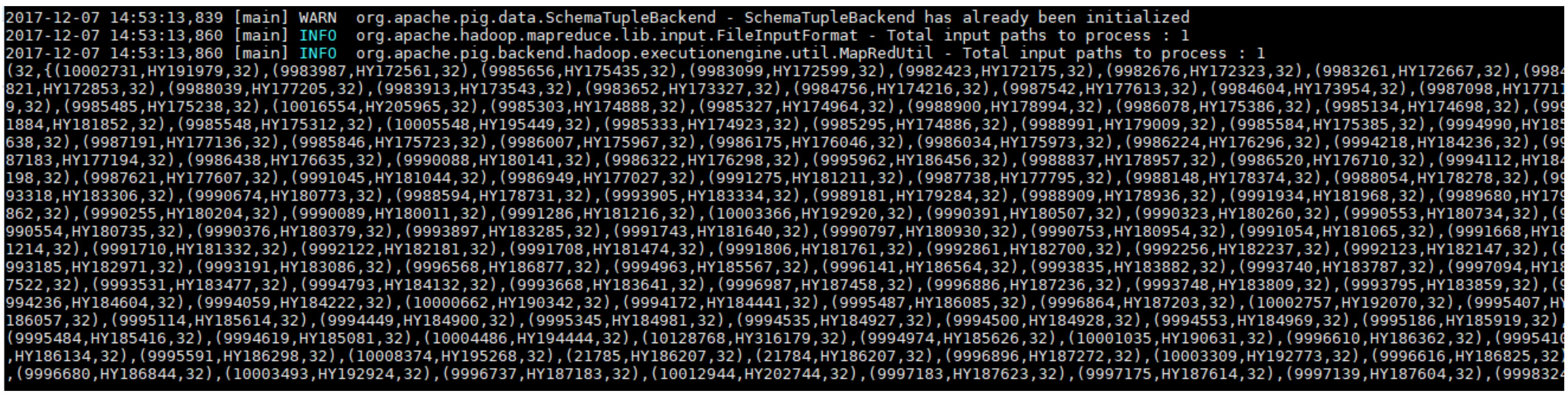
- **C = GROUP B by Ward;**

Here in above script we are grouping the relation B with ward in order to find the count of ward 32.



```
grunt> C = GROUP B by Ward;
grunt> lim3 = LIMIT C 1;
grunt> dump lim3;
```

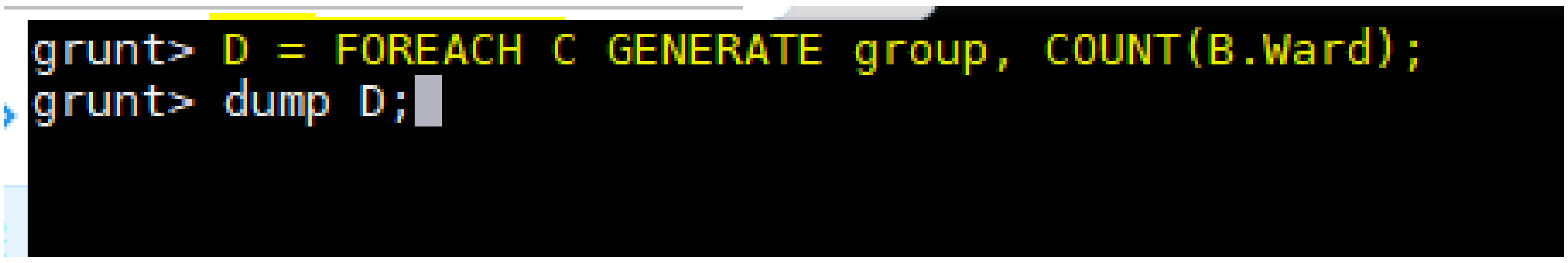
Same can be seen in below screenshot-



```
2017-12-07 14:53:13,839 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-07 14:53:13,860 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-07 14:53:13,860 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(32,{(10002731,HY191979,32),(9983987,HY172561,32),(9985656,HY175435,32),(9983099,HY172599,32),(9982423,HY172175,32),(9982676,HY172323,32),(9983261,HY172667,32),(9984
821,HY172853,32),(9988039,HY177205,32),(9983913,HY173543,32),(9983652,HY173327,32),(9984756,HY174216,32),(9987542,HY177613,32),(9984604,HY173954,32),(9987098,HY1771
9,32),(9985485,HY175238,32),(10016554,HY205965,32),(9985303,HY174888,32),(9985327,HY174964,32),(9988900,HY178994,32),(9986078,HY175386,32),(9985134,HY174698,32),(999
1884,HY181852,32),(9985548,HY175312,32),(10005548,HY195449,32),(9985333,HY174923,32),(9985295,HY174886,32),(9988991,HY179009,32),(9985584,HY175385,32),(9994990,HY185
638,32),(9987191,HY177136,32),(9985846,HY175723,32),(9986007,HY175967,32),(9986175,HY176046,32),(9986034,HY175973,32),(9986224,HY176296,32),(9994218,HY184236,32),(99
87183,HY177194,32),(9986438,HY176635,32),(9990088,HY180141,32),(9986322,HY176298,32),(9995962,HY186456,32),(9988837,HY178957,32),(9986520,HY176710,32),(9994112,HY184
198,32),(9987621,HY177607,32),(9991045,HY181044,32),(9986949,HY177027,32),(9991275,HY181211,32),(9987738,HY177795,32),(9988148,HY178374,32),(9988054,HY178278,32),(99
93318,HY183306,32),(9990674,HY180773,32),(9988594,HY178731,32),(9993905,HY183334,32),(9989181,HY179284,32),(9988909,HY178936,32),(9991934,HY181968,32),(9989680,HY179
862,32),(9990255,HY180204,32),(9990089,HY180011,32),(9991286,HY181216,32),(10003366,HY192920,32),(9990391,HY180507,32),(9990323,HY180260,32),(9990553,HY180734,32),(9
990554,HY180735,32),(9990376,HY180379,32),(9993897,HY183285,32),(9991743,HY181640,32),(9990797,HY180930,32),(9990753,HY180954,32),(9991054,HY181065,32),(9991668,HY18
1214,32),(9991710,HY181332,32),(9992122,HY182181,32),(9991708,HY181474,32),(9991806,HY181761,32),(9992861,HY182700,32),(9992256,HY182237,32),(9992123,HY182147,32),(9
993185,HY182971,32),(9993191,HY183086,32),(9996568,HY186877,32),(9994963,HY185567,32),(9996141,HY186564,32),(9993835,HY183882,32),(9993740,HY183787,32),(9997094,HY18
7522,32),(9993531,HY183477,32),(9994793,HY184132,32),(9993668,HY183641,32),(9996987,HY187458,32),(9996886,HY187236,32),(9993748,HY183809,32),(9993795,HY183859,32),(9
994236,HY184604,32),(9994059,HY184222,32),(10000662,HY190342,32),(9994172,HY184441,32),(9995487,HY186085,32),(9996864,HY187203,32),(10002757,HY192070,32),(9995407,HY
186057,32),(9995114,HY185614,32),(9994449,HY184900,32),(9995345,HY184981,32),(9994535,HY184927,32),(9994500,HY184928,32),(9994553,HY184969,32),(9995186,HY185919,32),
(9995484,HY185416,32),(9994619,HY185081,32),(10004486,HY194444,32),(10128768,HY316179,32),(9994974,HY185626,32),(10001035,HY190631,32),(9996610,HY186362,32),(9995410
,HY186134,32),(9995591,HY186298,32),(10008374,HY195268,32),(21785,HY186207,32),(21784,HY186207,32),(9996896,HY187272,32),(10003309,HY192773,32),(9996616,HY186825,32)
,(9996680,HY186844,32),(10003493,HY192924,32),(9996737,HY187183,32),(10012944,HY202744,32),(9997183,HY187623,32),(9997175,HY187614,32),(9997139,HY187604,32),(9998324
```

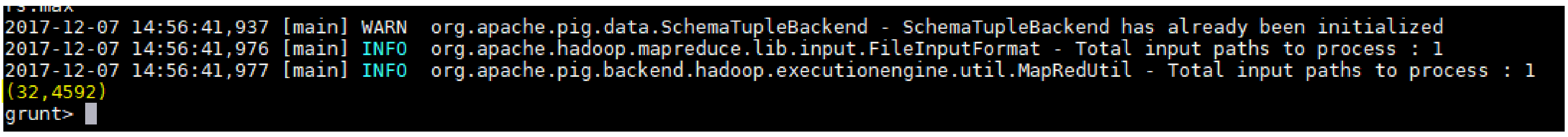
- ➢ **D = FOREACH C GENERATE group, COUNT(B.Ward);**
- ➢ **Dump D;**

Now here we are calculating the final count using above script. The COUNT(B.Ward) basically iterates through each group generated in relation C and counts the number of wards which is 32 in this case-



```
grunt> D = FOREACH C GENERATE group, COUNT(B.Ward);
grunt> dump D;
```

Below screenshot shows the count of cases investigated for ward-32 which **<u>4592</u>**-



```
2017-12-07 14:56:41,937 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-07 14:56:41,976 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-07 14:56:41,977 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(32,4592)
grunt>
```

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

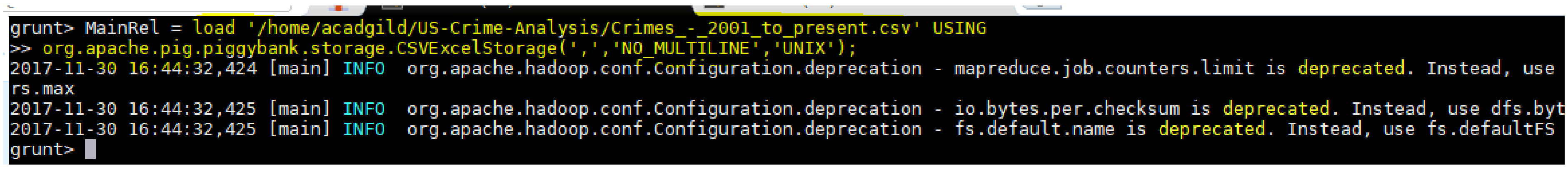
Below is the script used to calculate number of arrests in theft district wise-

- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**
- **B = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$5 as primary_type, (chararray)$8 as arrest_flag,(int)$11 as district;**
- **C = FILTER B BY primary_type == 'THEFT' AND arrest_flag == 'true';**
- **D = GROUP C by district;**
- **E = FOREACH D GENERATE group, COUNT(C.district);**
- **Dump E;**

Now let's get into detail of each script one by one-

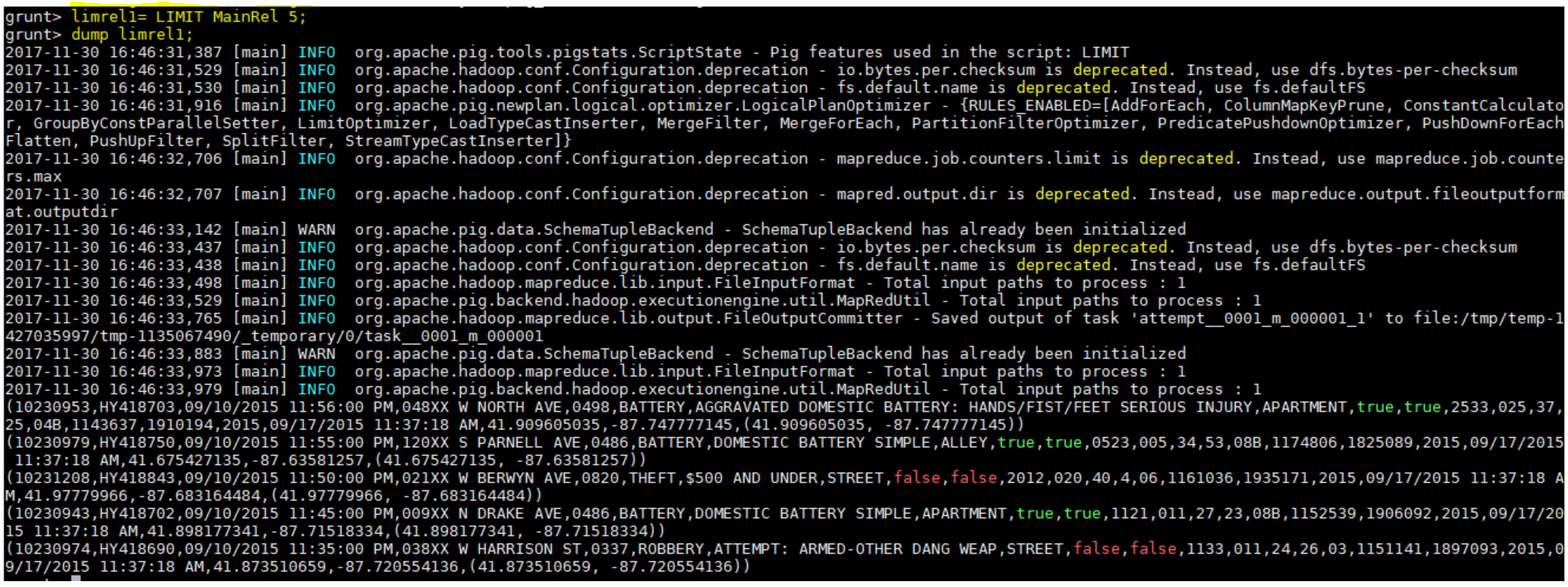
- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**

Here in above relation we are loading the comma separated file to Pig Storage-



```
grunt> MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING
>> org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');
2017-11-30 16:44:32,424 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use
rs.max
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.byt
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

Below screenshot shows that file has been loaded successfully

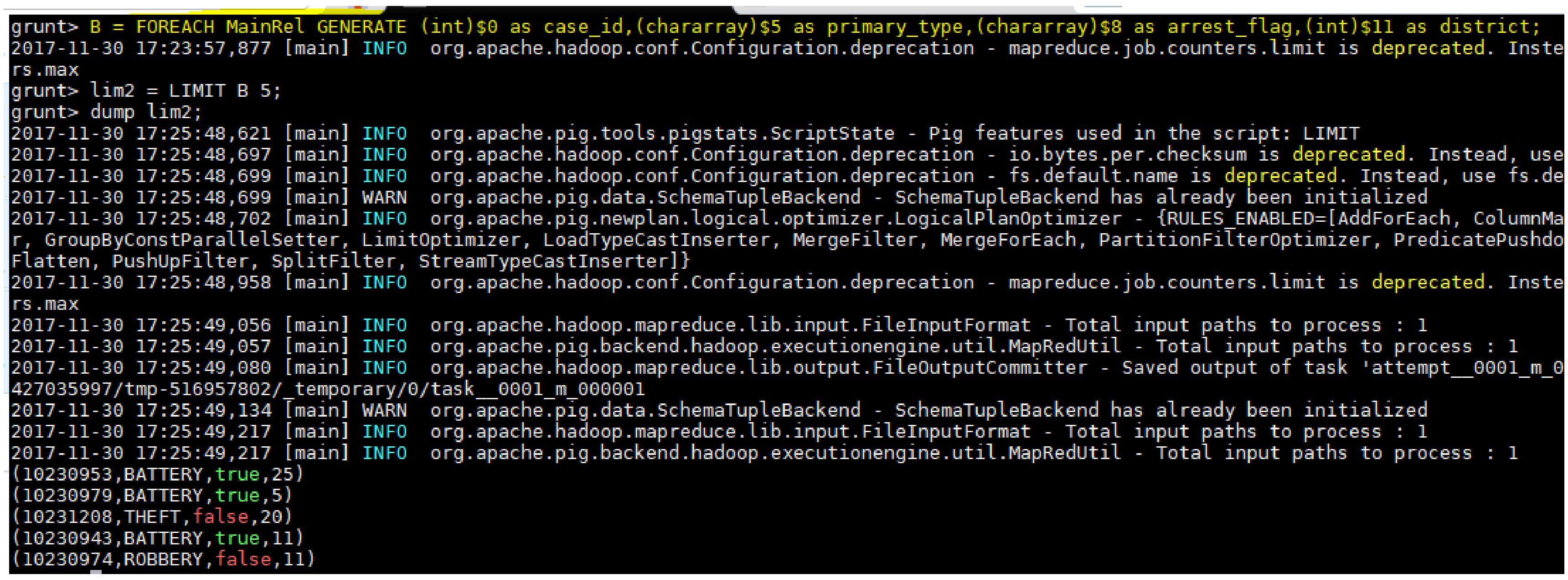


```
grunt> limrel1= LIMIT MainRel 5;
grunt> dump limrel1;
2017-11-30 16:46:31,387 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-11-30 16:46:31,529 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:31,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:31,916 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 16:46:32,706 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counte
rs.max
2017-11-30 16:46:32,707 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputform
at.outputdir
2017-11-30 16:46:33,142 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:33,438 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:33,498 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,529 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-11-30 16:46:33,765 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to file:/tmp/temp-1
427035997/tmp-1135067490/_temporary/0/task__0001_m_000001
2017-11-30 16:46:33,883 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,973 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,979 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10230953,HY418703,09/10/2015 11:56:00 PM,048XX W NORTH AVE,0498,BATTERY,AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY,APARTMENT,true,true,2533,025,37,
25,04B,1143637,1910194,2015,09/17/2015 11:37:18 AM,41.909605035,-87.747777145,(41.909605035, -87.747777145))
(10230979,HY418750,09/10/2015 11:55:00 PM,120XX S PARNELL AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,ALLEY,true,true,0523,005,34,53,08B,1174806,1825089,2015,09/17/2015
 11:37:18 AM,41.675427135,-87.63581257,(41.675427135, -87.63581257))
(10231208,HY418843,09/10/2015 11:50:00 PM,021XX W BERWYN AVE,0820,THEFT,$500 AND UNDER,STREET,false,false,2012,020,40,4,06,1161036,1935171,2015,09/17/2015 11:37:18 A
M,41.97779966,-87.683164484,(41.97779966, -87.683164484))
(10230943,HY418702,09/10/2015 11:45:00 PM,009XX N DRAKE AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,APARTMENT,true,true,1121,011,27,23,08B,1152539,1906092,2015,09/17/20
15 11:37:18 AM,41.898177341,-87.71518334,(41.898177341, -87.71518334))
(10230974,HY418690,09/10/2015 11:35:00 PM,038XX W HARRISON ST,0337,ROBBERY,ATTEMPT: ARMED-OTHER DANG WEAP,STREET,false,false,1133,011,24,26,03,1151141,1897093,2015,0
9/17/2015 11:37:18 AM,41.873510659,-87.720554136,(41.873510659, -87.720554136))
```

- **B = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$5 as primary_type, (chararray)$8 as arrest_flag,(int)$11 as district;**

The above script is being used to extract/generate some columns from the file loaded. Here we are extracting case_id, primary_type, arrest_flag and district.

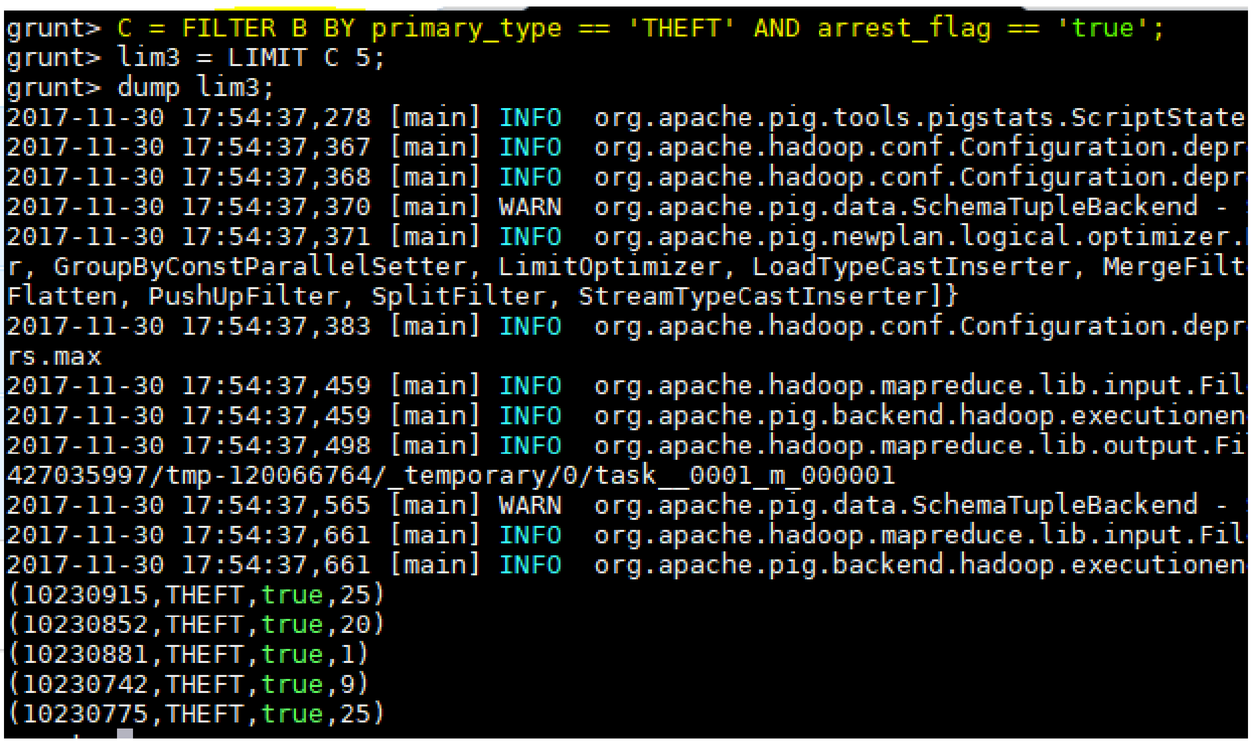
The same can be seen in below screenshot-



```
grunt> B = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$5 as primary_type,(chararray)$8 as arrest_flag,(int)$11 as district;
2017-11-30 17:23:57,877 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
rs.max
grunt> lim2 = LIMIT B 5;
grunt> dump lim2;
2017-11-30 17:25:48,621 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-11-30 17:25:48,697 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
2017-11-30 17:25:48,699 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
2017-11-30 17:25:48,699 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 17:25:48,702 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMa
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdo
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 17:25:48,958 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
rs.max
2017-11-30 17:25:49,056 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 17:25:49,057 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-11-30 17:25:49,080 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_0
427035997/tmp-516957802/_temporary/0/task__0001_m_000001
2017-11-30 17:25:49,134 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 17:25:49,217 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 17:25:49,217 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10230953,BATTERY,true,25)
(10230979,BATTERY,true,5)
(10231208,THEFT,false,20)
(10230943,BATTERY,true,11)
(10230974,ROBBERY,false,11)
```

- **C = FILTER B BY primary_type == 'THEFT' AND arrest_flag == 'true';**

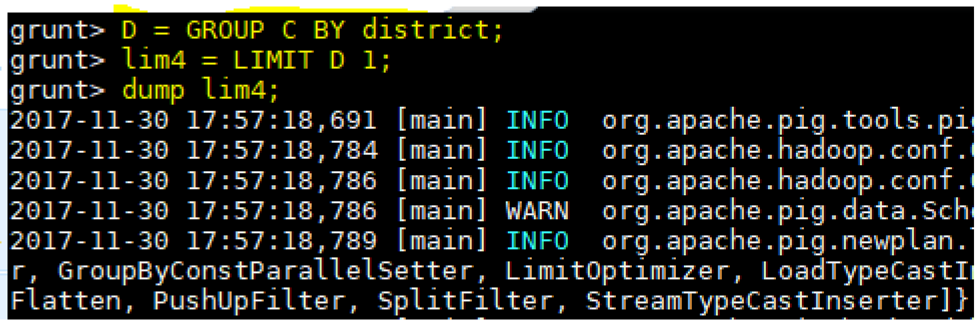
  Here in above script we are filtering the records with primary type as "theft" and arrest_flag as "true" because we are interested in calculating the number of arrests only for theft. Same can be seen in below screenshot-



```
grunt> C = FILTER B BY primary_type == 'THEFT' AND arrest_flag == 'true';
grunt> lim3 = LIMIT C 5;
grunt> dump lim3;
2017-11-30 17:54:37,278 [main] INFO  org.apache.pig.tools.pigstats.ScriptState
2017-11-30 17:54:37,367 [main] INFO  org.apache.hadoop.conf.Configuration.depr
2017-11-30 17:54:37,368 [main] INFO  org.apache.hadoop.conf.Configuration.depr
2017-11-30 17:54:37,370 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
2017-11-30 17:54:37,371 [main] INFO  org.apache.pig.newplan.logical.optimizer.
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilt
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 17:54:37,383 [main] INFO  org.apache.hadoop.conf.Configuration.depr
rs.max
2017-11-30 17:54:37,459 [main] INFO  org.apache.hadoop.mapreduce.lib.input.Fil
2017-11-30 17:54:37,459 [main] INFO  org.apache.pig.backend.hadoop.executionen
2017-11-30 17:54:37,498 [main] INFO  org.apache.hadoop.mapreduce.lib.output.Fi
427035997/tmp-120066764/_temporary/0/task__0001_m_000001
2017-11-30 17:54:37,565 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
2017-11-30 17:54:37,661 [main] INFO  org.apache.hadoop.mapreduce.lib.input.Fil
2017-11-30 17:54:37,661 [main] INFO  org.apache.pig.backend.hadoop.executionen
(10230915,THEFT,true,25)
(10230852,THEFT,true,20)
(10230881,THEFT,true,1)
(10230742,THEFT,true,9)
(10230775,THEFT,true,25)
```

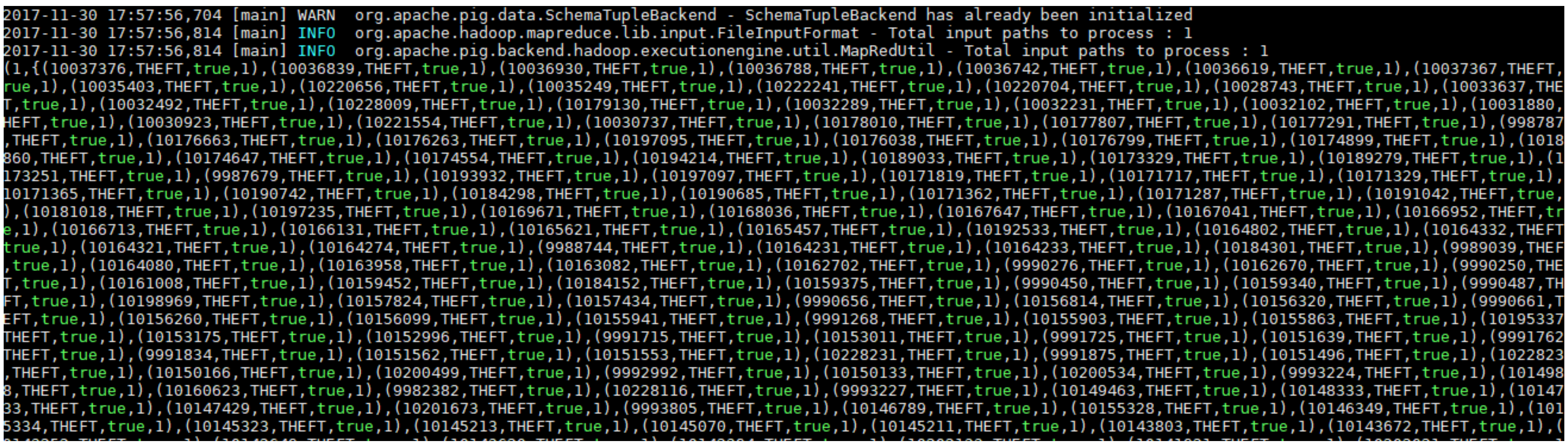
- **D = GROUP C by district;**

Now as we have to find the number of arrests for theft district wise we are grouping the above relation with district-



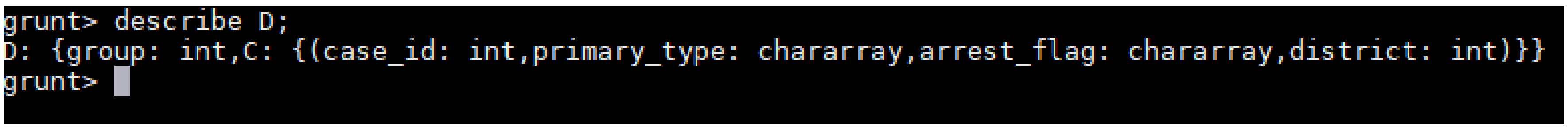
```
grunt> D = GROUP C BY district;
grunt> lim4 = LIMIT D 1;
grunt> dump lim4;
2017-11-30 17:57:18,691 [main] INFO  org.apache.pig.tools.pi
2017-11-30 17:57:18,784 [main] INFO  org.apache.hadoop.conf.
2017-11-30 17:57:18,786 [main] INFO  org.apache.hadoop.conf.
2017-11-30 17:57:18,786 [main] WARN  org.apache.pig.data.Sch
2017-11-30 17:57:18,789 [main] INFO  org.apache.pig.newplan.
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastI
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
```

Below screenshot shows one sample grouped data for district 1-



```
2017-11-30 17:57:56,704 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 17:57:56,814 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 17:57:56,814 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,{(10037376,THEFT,true,1),(10036839,THEFT,true,1),(10036930,THEFT,true,1),(10036788,THEFT,true,1),(10036742,THEFT,true,1),(10036619,THEFT,true,1),(10037367,THEFT,
rue,1),(10035403,THEFT,true,1),(10220656,THEFT,true,1),(10035249,THEFT,true,1),(10222241,THEFT,true,1),(10220704,THEFT,true,1),(10028743,THEFT,true,1),(10033637,THE
T,true,1),(10032492,THEFT,true,1),(10228009,THEFT,true,1),(10179130,THEFT,true,1),(10032289,THEFT,true,1),(10032231,THEFT,true,1),(10032102,THEFT,true,1),(10031880,
HEFT,true,1),(10030923,THEFT,true,1),(10221554,THEFT,true,1),(10030737,THEFT,true,1),(10178010,THEFT,true,1),(10177807,THEFT,true,1),(10177291,THEFT,true,1),(998787
,THEFT,true,1),(10176663,THEFT,true,1),(10176263,THEFT,true,1),(10197095,THEFT,true,1),(10176038,THEFT,true,1),(10176799,THEFT,true,1),(10174899,THEFT,true,1),(1018
860,THEFT,true,1),(10174647,THEFT,true,1),(10174554,THEFT,true,1),(10194214,THEFT,true,1),(10189033,THEFT,true,1),(10173329,THEFT,true,1),(10189279,THEFT,true,1),(1
173251,THEFT,true,1),(9987679,THEFT,true,1),(10193932,THEFT,true,1),(10197097,THEFT,true,1),(10171819,THEFT,true,1),(10171717,THEFT,true,1),(10171329,THEFT,true,1),
10171365,THEFT,true,1),(10190742,THEFT,true,1),(10184298,THEFT,true,1),(10190685,THEFT,true,1),(10171362,THEFT,true,1),(10171287,THEFT,true,1),(10191042,THEFT,true,
),(10181018,THEFT,true,1),(10197235,THEFT,true,1),(10169671,THEFT,true,1),(10168036,THEFT,true,1),(10167647,THEFT,true,1),(10167041,THEFT,true,1),(10166952,THEFT,tr
e,1),(10166713,THEFT,true,1),(10166131,THEFT,true,1),(10165621,THEFT,true,1),(10165457,THEFT,true,1),(10192533,THEFT,true,1),(10164802,THEFT,true,1),(10164332,THEFT
true,1),(10164321,THEFT,true,1),(10164274,THEFT,true,1),(9988744,THEFT,true,1),(10164231,THEFT,true,1),(10164233,THEFT,true,1),(10184301,THEFT,true,1),(9989039,THEF
,true,1),(10164080,THEFT,true,1),(10163958,THEFT,true,1),(10163082,THEFT,true,1),(10162702,THEFT,true,1),(9990276,THEFT,true,1),(10162670,THEFT,true,1),(9990250,THE
T,true,1),(10161008,THEFT,true,1),(10159452,THEFT,true,1),(10184152,THEFT,true,1),(10159375,THEFT,true,1),(9990450,THEFT,true,1),(10159340,THEFT,true,1),(9990487,TH
FT,true,1),(10198969,THEFT,true,1),(10157824,THEFT,true,1),(10157434,THEFT,true,1),(9990656,THEFT,true,1),(10156814,THEFT,true,1),(10156320,THEFT,true,1),(9990661,T
EFT,true,1),(10156260,THEFT,true,1),(10156099,THEFT,true,1),(10155941,THEFT,true,1),(9991268,THEFT,true,1),(10155903,THEFT,true,1),(10155863,THEFT,true,1),(10195337
THEFT,true,1),(10153175,THEFT,true,1),(10152996,THEFT,true,1),(9991715,THEFT,true,1),(10153011,THEFT,true,1),(9991725,THEFT,true,1),(10151639,THEFT,true,1),(9991762
THEFT,true,1),(9991834,THEFT,true,1),(10151562,THEFT,true,1),(10151553,THEFT,true,1),(10228231,THEFT,true,1),(9991875,THEFT,true,1),(10151496,THEFT,true,1),(1022823
,THEFT,true,1),(10150166,THEFT,true,1),(10200499,THEFT,true,1),(9992992,THEFT,true,1),(10150133,THEFT,true,1),(10200534,THEFT,true,1),(9993224,THEFT,true,1),(101498
8,THEFT,true,1),(10160623,THEFT,true,1),(9982382,THEFT,true,1),(10228116,THEFT,true,1),(9993227,THEFT,true,1),(10149463,THEFT,true,1),(10148333,THEFT,true,1),(10147
33,THEFT,true,1),(10147429,THEFT,true,1),(10201673,THEFT,true,1),(9993805,THEFT,true,1),(10146789,THEFT,true,1),(10155328,THEFT,true,1),(10146349,THEFT,true,1),(101
5334,THEFT,true,1),(10145323,THEFT,true,1),(10145213,THEFT,true,1),(10145070,THEFT,true,1),(10145211,THEFT,true,1),(10143803,THEFT,true,1),(10143672,THEFT,true,1),(
```

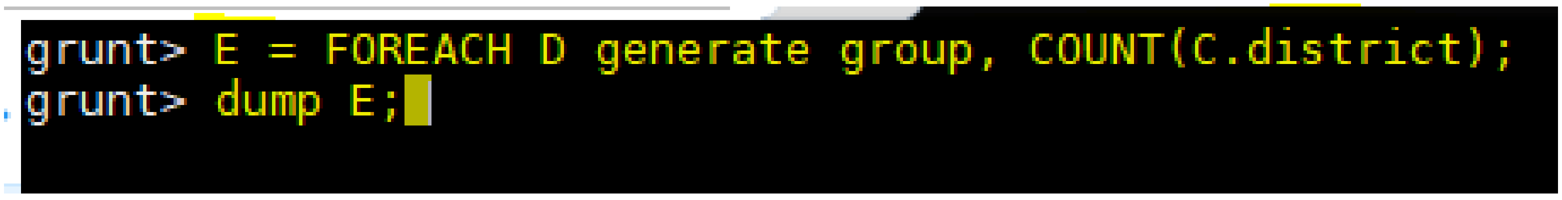
If we describe the above relation D we get below schema-



```
grunt> describe D;
D: {group: int,C: {(case_id: int,primary_type: chararray,arrest_flag: chararray,district: int)}}
grunt>
```

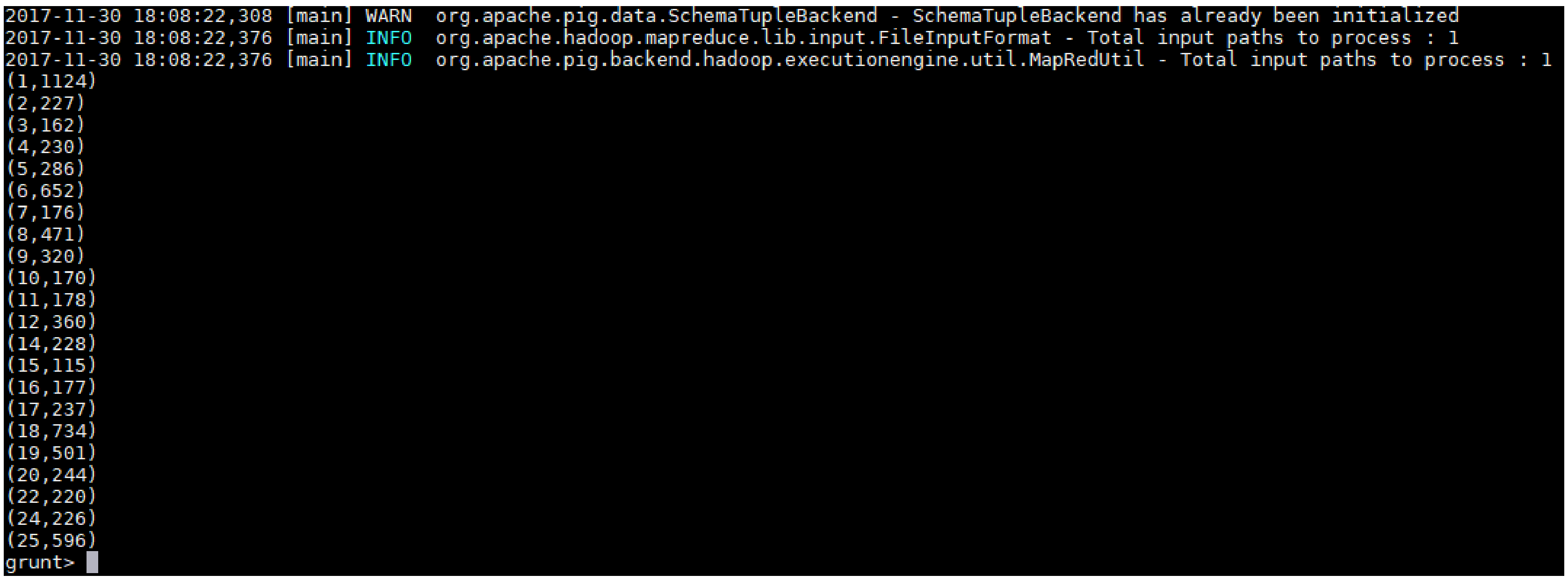
- **E = FOREACH D GENERATE group, COUNT(C.district);**

  Now we can very well calculate the count of each district from the relation C as it has been grouped by district only.



```
grunt> E = FOREACH D generate group, COUNT(C.district);
grunt> dump E;
```

Below screenshot shows the final result. It contains the district number and its count for each district having arrests done for theft-



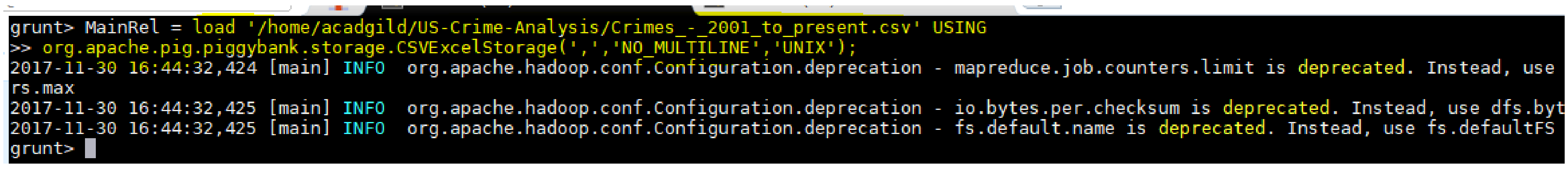
```
2017-11-30 18:08:22,308 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 18:08:22,376 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 18:08:22,376 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,1124)
(2,227)
(3,162)
(4,230)
(5,286)
(6,652)
(7,176)
(8,471)
(9,320)
(10,170)
(11,178)
(12,360)
(14,228)
(15,115)
(16,177)
(17,237)
(18,734)
(19,501)
(20,244)
(22,220)
(24,226)
(25,596)
grunt>
```

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

Below is the script used to find number of arrests done between October 2014 and October 2015

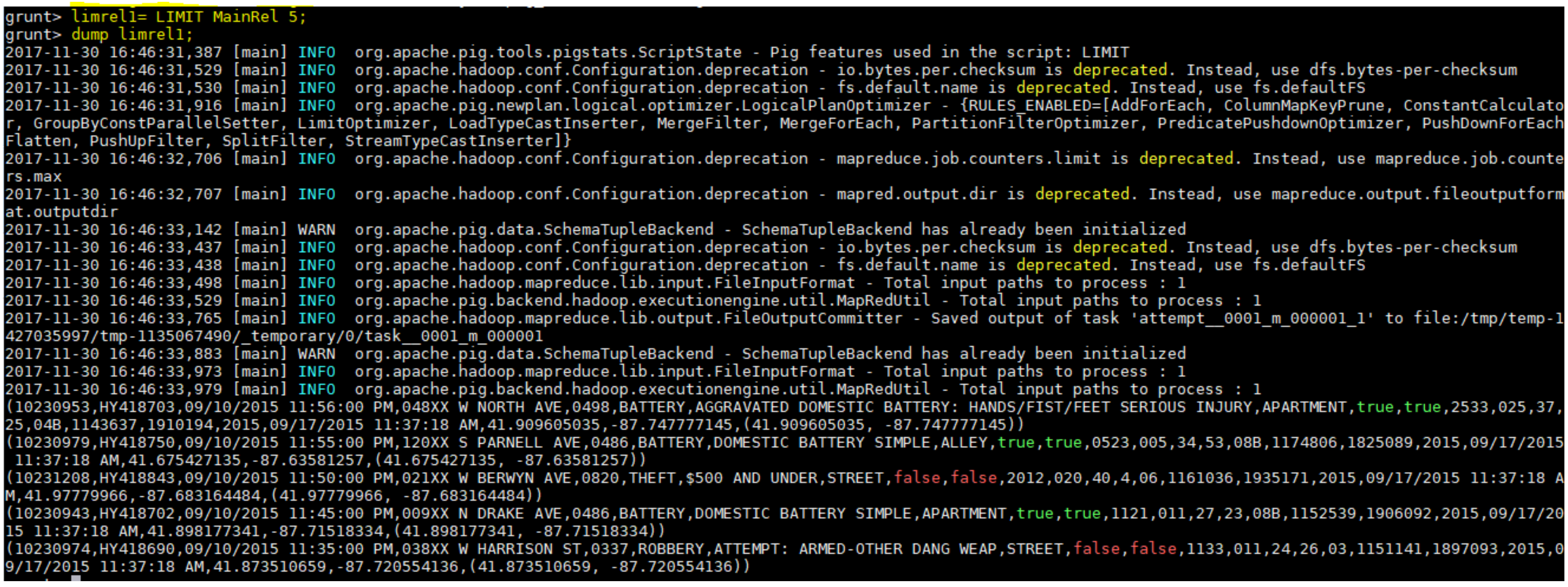
- **MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING**
- **org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');**
- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$8 as arrest_flag, (chararray)$18 as updated_on;**
- **B = FOREACH A GENERATE case_id, arrest_flag, ToDate(updated_on, 'MM/dd/yyyy hh:mm:ss aa') as date1;**
- **C = FILTER B BY arrest_flag=='true' AND date1 >= (datetime)Todate('10/01/2014', 'MM/dd/yyyy') AND date1 <= ToDate('10/31/2015', 'MM/dd/yyyy');**
- **D = GROUP C BY arrest_flag;**
- **E = FOREACH D GENERATE group, COUNT(C.arrest_flag);**
- **Dump E;**

Here in above relation we are loading the comma separated file to Pig Storage-



```
grunt> MainRel = load '/home/acadgild/US-Crime-Analysis/Crimes_-_2001_to_present.csv' USING
>> org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX');
2017-11-30 16:44:32,424 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use
rs.max
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.byt
2017-11-30 16:44:32,425 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> 
```

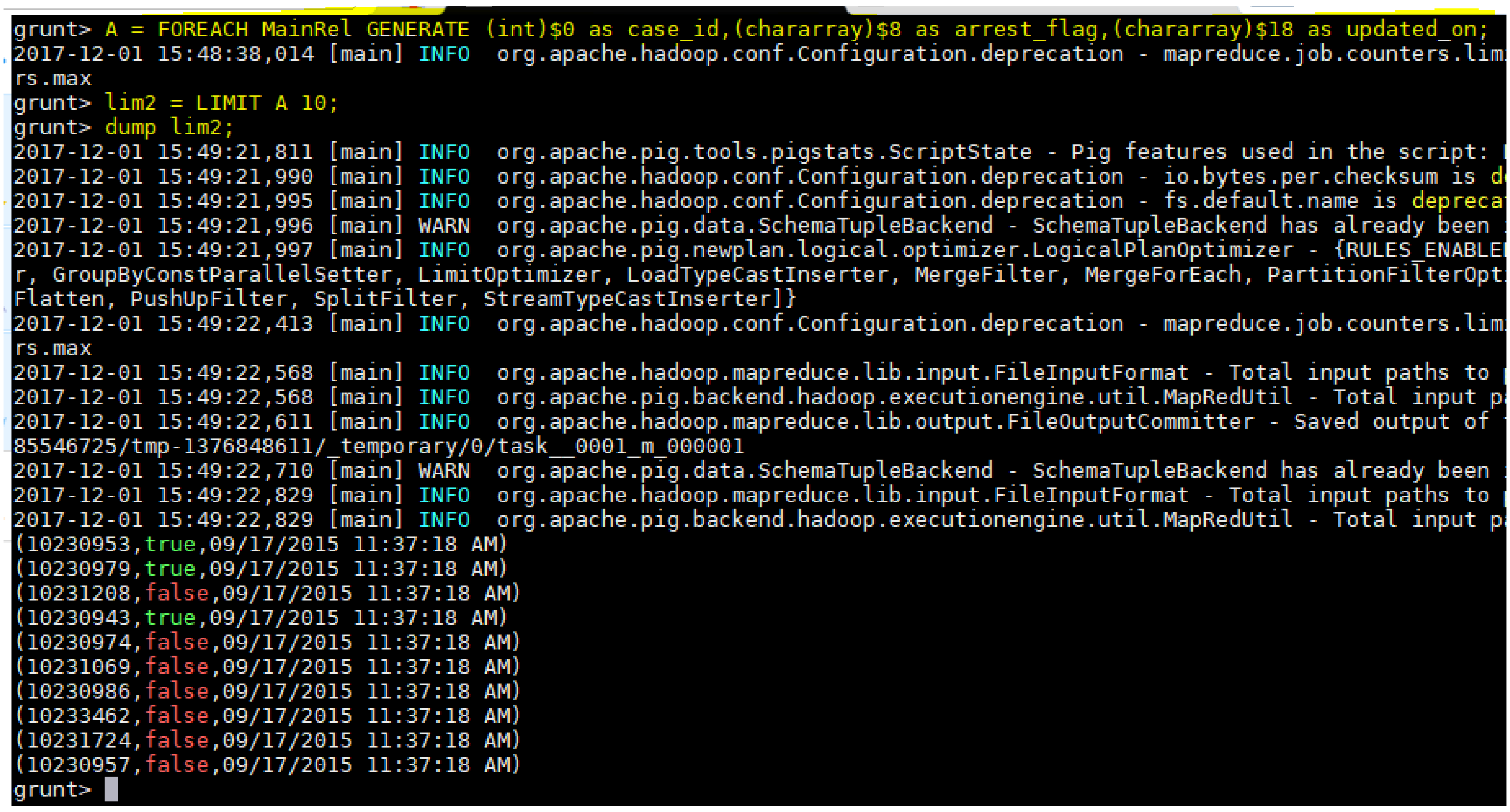
Below screenshot shows that file has been loaded successfully



```
grunt> limrel1= LIMIT MainRel 5;
grunt> dump limrel1;
2017-11-30 16:46:31,387 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-11-30 16:46:31,529 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:31,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:31,916 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-30 16:46:32,706 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counte
rs.max
2017-11-30 16:46:32,707 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputform
at.outputdir
2017-11-30 16:46:33,142 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-30 16:46:33,438 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-30 16:46:33,498 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,529 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-11-30 16:46:33,765 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to file:/tmp/temp-1
427035997/tmp-1135067490/_temporary/0/task__0001_m_000001
2017-11-30 16:46:33,883 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-30 16:46:33,973 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-30 16:46:33,979 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10230953,HY418703,09/10/2015 11:56:00 PM,048XX W NORTH AVE,0498,BATTERY,AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY,APARTMENT,true,true,2533,025,37,
25,04B,1143637,1910194,2015,09/17/2015 11:37:18 AM,41.909605035,-87.747777145,(41.909605035, -87.747777145))
(10230979,HY418750,09/10/2015 11:55:00 PM,120XX S PARNELL AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,ALLEY,true,true,0523,005,34,53,08B,1174806,1825089,2015,09/17/2015
 11:37:18 AM,41.675427135,-87.63581257,(41.675427135, -87.63581257))
(10231208,HY418843,09/10/2015 11:50:00 PM,021XX W BERWYN AVE,0820,THEFT,$500 AND UNDER,STREET,false,false,2012,020,40,4,06,1161036,1935171,2015,09/17/2015 11:37:18 A
M,41.97779966,-87.683164484,(41.97779966, -87.683164484))
(10230943,HY418702,09/10/2015 11:45:00 PM,009XX N DRAKE AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,APARTMENT,true,true,1121,011,27,23,08B,1152539,1906092,2015,09/17/20
15 11:37:18 AM,41.898177341,-87.71518334,(41.898177341, -87.71518334))
(10230974,HY418690,09/10/2015 11:35:00 PM,038XX W HARRISON ST,0337,ROBBERY,ATTEMPT: ARMED-OTHER DANG WEAP,STREET,false,false,1133,011,24,26,03,1151141,1897093,2015,0
9/17/2015 11:37:18 AM,41.873510659,-87.720554136,(41.873510659, -87.720554136))
```

- **A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$8 as arrest_flag, (chararray)$18 as updated_on;**

Now here in above script we are extracting/generating some columns from the loaded relation. These columns are case_id, arrest_flag and updated_on. The field updated_on is for date but we are loading it as chararray as of now.

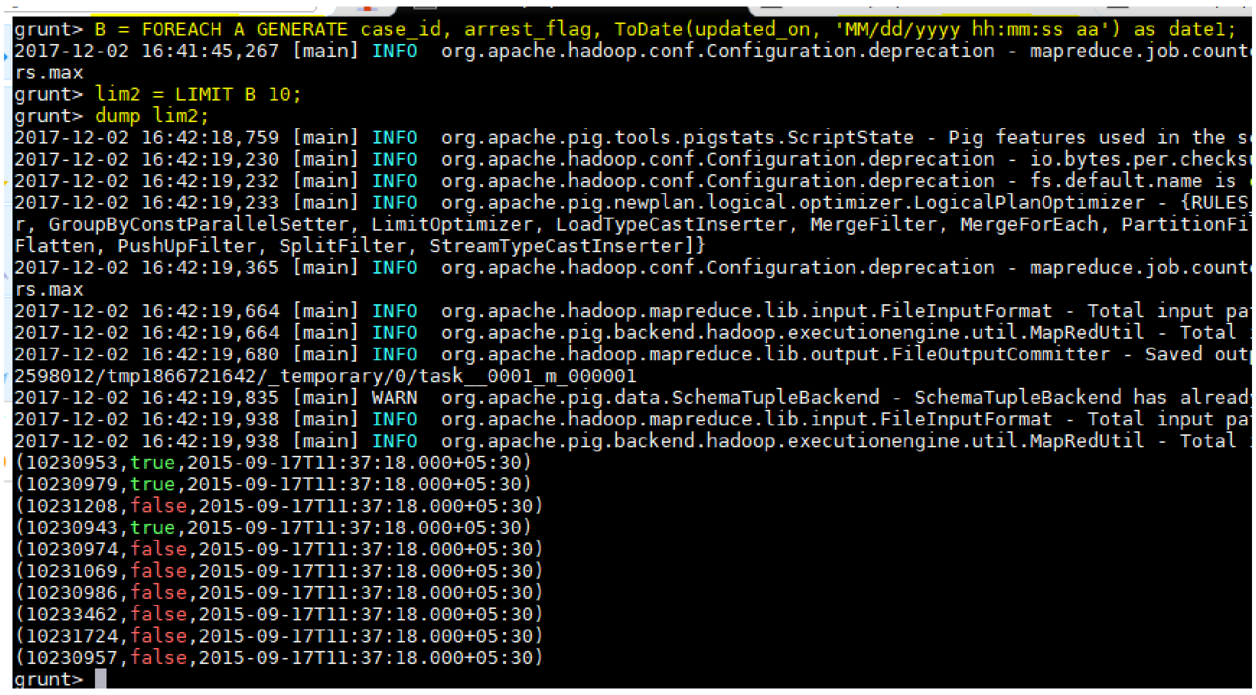


```
grunt> A = FOREACH MainRel GENERATE (int)$0 as case_id,(chararray)$8 as arrest_flag,(chararray)$18 as updated_on;
2017-12-01 15:48:38,014 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.lim
rs.max
grunt> lim2 = LIMIT A 10;
grunt> dump lim2;
2017-12-01 15:49:21,811 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script:
2017-12-01 15:49:21,990 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is d
2017-12-01 15:49:21,995 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depreca
2017-12-01 15:49:21,996 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been
2017-12-01 15:49:21,997 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLE
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOpt
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-01 15:49:22,413 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.lim
rs.max
2017-12-01 15:49:22,568 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
2017-12-01 15:49:22,568 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input p
2017-12-01 15:49:22,611 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of
85546725/tmp-1376848611/_temporary/0/task__0001_m_000001
2017-12-01 15:49:22,710 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been
2017-12-01 15:49:22,829 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
2017-12-01 15:49:22,829 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input p
(10230953,true,09/17/2015 11:37:18 AM)
(10230979,true,09/17/2015 11:37:18 AM)
(10231208,false,09/17/2015 11:37:18 AM)
(10230943,true,09/17/2015 11:37:18 AM)
(10230974,false,09/17/2015 11:37:18 AM)
(10231069,false,09/17/2015 11:37:18 AM)
(10230986,false,09/17/2015 11:37:18 AM)
(10233462,false,09/17/2015 11:37:18 AM)
(10231724,false,09/17/2015 11:37:18 AM)
(10230957,false,09/17/2015 11:37:18 AM)
grunt>
```

- **B = FOREACH A GENERATE case_id, arrest_flag, ToDate(updated_on, ‘MM/dd/yyyy hh:mm:ss aa’) as date1;**

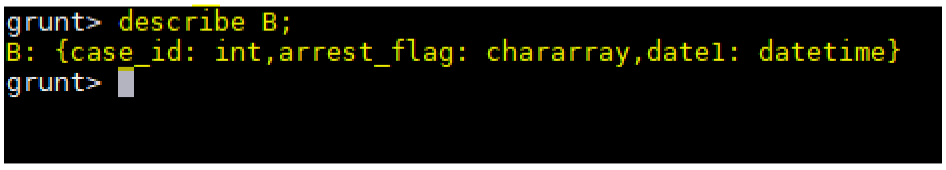
Here in above script we are again generating same columns but we are converting the field updated_on as ToDate and formatting it using ‘MM/dd/yyyy hh:mm:ss aa’ and giving it name as date1.

Same can be seen below once we dump this relation-



```
grunt> B = FOREACH A GENERATE case_id, arrest_flag, ToDate(updated_on, 'MM/dd/yyyy hh:mm:ss aa') as date1;
2017-12-02 16:41:45,267 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.count
rs.max
grunt> lim2 = LIMIT B 10;
grunt> dump lim2;
2017-12-02 16:42:18,759 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the s
2017-12-02 16:42:19,230 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checks
2017-12-02 16:42:19,232 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
2017-12-02 16:42:19,233 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFi
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-02 16:42:19,365 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.count
rs.max
2017-12-02 16:42:19,664 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input pa
2017-12-02 16:42:19,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
2017-12-02 16:42:19,680 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved out
2598012/tmp1866721642/_temporary/0/task__0001_m_000001
2017-12-02 16:42:19,835 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alread
2017-12-02 16:42:19,938 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input pa
2017-12-02 16:42:19,938 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
(10230953,true,2015-09-17T11:37:18.000+05:30)
(10230979,true,2015-09-17T11:37:18.000+05:30)
(10231208,false,2015-09-17T11:37:18.000+05:30)
(10230943,true,2015-09-17T11:37:18.000+05:30)
(10230974,false,2015-09-17T11:37:18.000+05:30)
(10231069,false,2015-09-17T11:37:18.000+05:30)
(10230986,false,2015-09-17T11:37:18.000+05:30)
(10233462,false,2015-09-17T11:37:18.000+05:30)
(10231724,false,2015-09-17T11:37:18.000+05:30)
(10230957,false,2015-09-17T11:37:18.000+05:30)
grunt>
```

If we describe above relation we can very well see that date1 field is of datetime type-

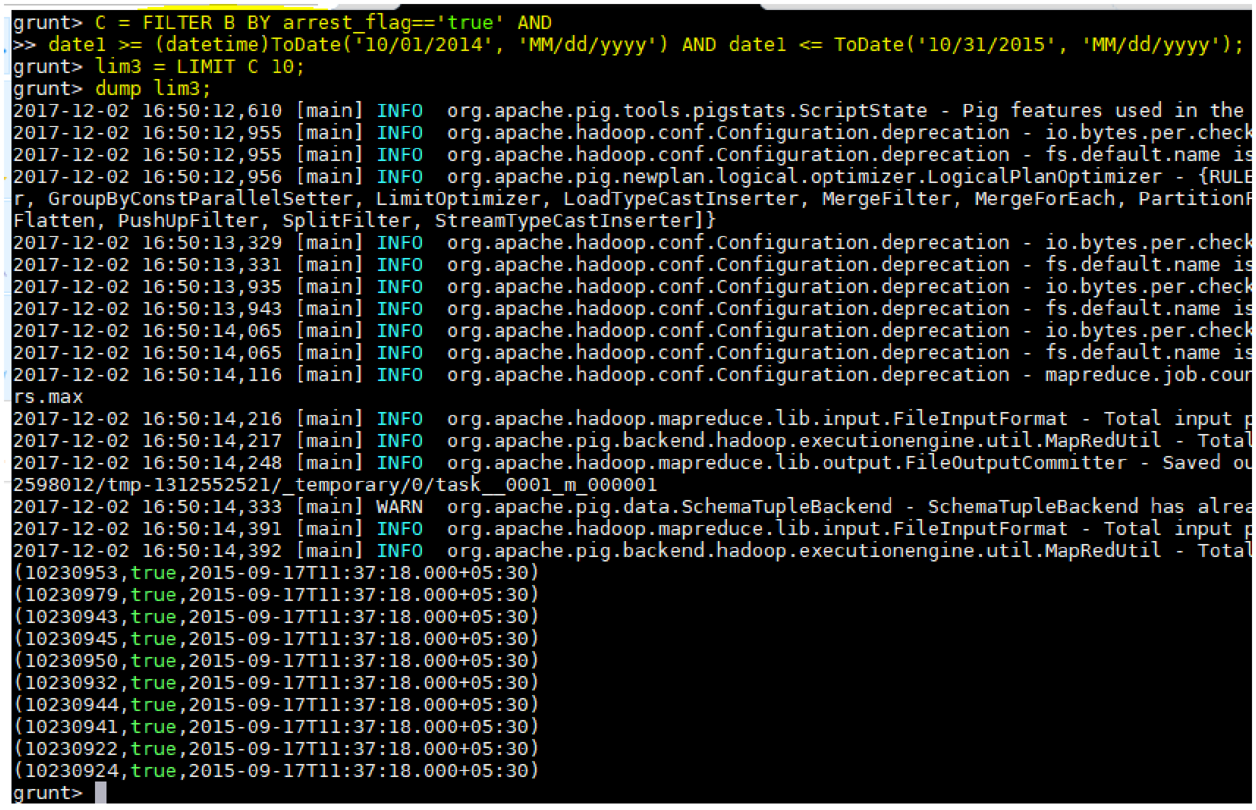


```
grunt> describe B;
B: {case_id: int,arrest_flag: chararray,date1: datetime}
grunt>
```

- **C = FILTER B BY arrest_flag=='true' AND date1 >= (datetime)Todate('10/01/2014', 'MM/dd/yyyy') AND date1 <= ToDate('10/31/2015', 'MM/dd/yyyy');**

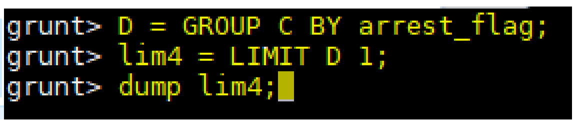
Now in order to find count of arrest done in between October 2014 and October 2015 we are filtering above B relation with arrest_flag == "true" and date in between 10/01/2014 and 10/31/2015

Below screenshot shows same and some sample dumped data-

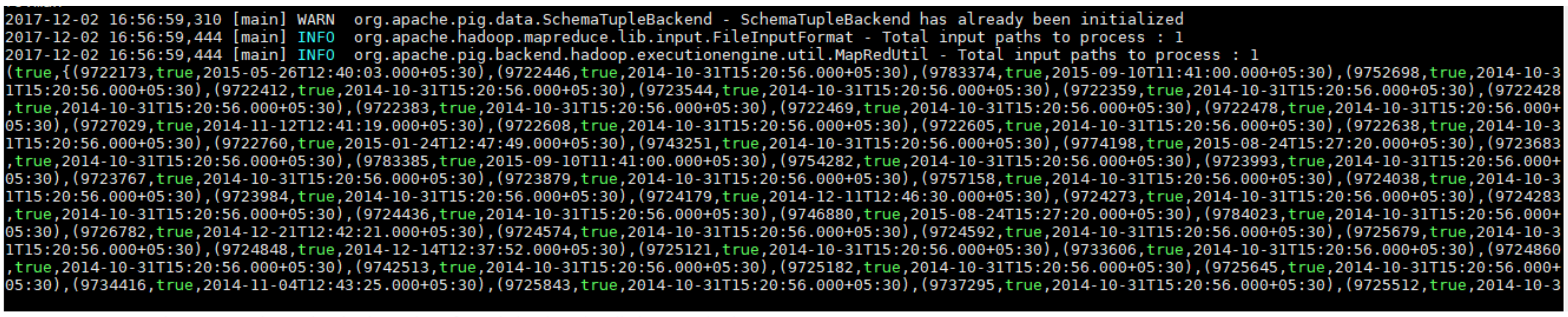


```
grunt> C = FILTER B BY arrest_flag=='true' AND
>> date1 >= (datetime)ToDate('10/01/2014', 'MM/dd/yyyy') AND date1 <= ToDate('10/31/2015', 'MM/dd/yyyy');
grunt> lim3 = LIMIT C 10;
grunt> dump lim3;
2017-12-02 16:50:12,610 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the
2017-12-02 16:50:12,955 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
2017-12-02 16:50:12,955 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
2017-12-02 16:50:12,956 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULE
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionF
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-02 16:50:13,329 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
2017-12-02 16:50:13,331 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
2017-12-02 16:50:13,935 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
2017-12-02 16:50:13,943 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
2017-12-02 16:50:14,065 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
2017-12-02 16:50:14,065 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
2017-12-02 16:50:14,116 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.coun
rs.max
2017-12-02 16:50:14,216 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input p
2017-12-02 16:50:14,217 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
2017-12-02 16:50:14,248 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved ou
2598012/tmp-1312552521/_temporary/0/task__0001_m_000001
2017-12-02 16:50:14,333 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alrea
2017-12-02 16:50:14,391 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input p
2017-12-02 16:50:14,392 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
(10230953,true,2015-09-17T11:37:18.000+05:30)
(10230979,true,2015-09-17T11:37:18.000+05:30)
(10230943,true,2015-09-17T11:37:18.000+05:30)
(10230945,true,2015-09-17T11:37:18.000+05:30)
(10230950,true,2015-09-17T11:37:18.000+05:30)
(10230932,true,2015-09-17T11:37:18.000+05:30)
(10230944,true,2015-09-17T11:37:18.000+05:30)
(10230941,true,2015-09-17T11:37:18.000+05:30)
(10230922,true,2015-09-17T11:37:18.000+05:30)
(10230924,true,2015-09-17T11:37:18.000+05:30)
grunt>
```

Now in order to find number of arrests done we will group the relation C with arrest_flag which has been already filtered with "true"

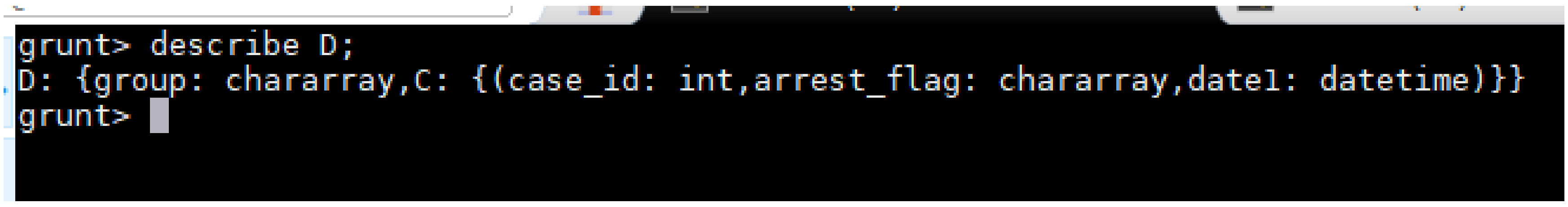


```
grunt> D = GROUP C BY arrest_flag;
grunt> lim4 = LIMIT D 1;
grunt> dump lim4;
```



```
2017-12-02 16:56:59,310 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-02 16:56:59,444 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-02 16:56:59,444 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(true,{(9722173,true,2015-05-26T12:40:03.000+05:30),(9722446,true,2014-10-31T15:20:56.000+05:30),(9783374,true,2015-09-10T11:41:00.000+05:30),(9752698,true,2014-10-3
1T15:20:56.000+05:30),(9722412,true,2014-10-31T15:20:56.000+05:30),(9723544,true,2014-10-31T15:20:56.000+05:30),(9722359,true,2014-10-31T15:20:56.000+05:30),(9722428
,true,2014-10-31T15:20:56.000+05:30),(9722383,true,2014-10-31T15:20:56.000+05:30),(9722469,true,2014-10-31T15:20:56.000+05:30),(9722478,true,2014-10-31T15:20:56.000+
05:30),(9727029,true,2014-11-12T12:41:19.000+05:30),(9722608,true,2014-10-31T15:20:56.000+05:30),(9722605,true,2014-10-31T15:20:56.000+05:30),(9722638,true,2014-10-3
1T15:20:56.000+05:30),(9722760,true,2015-01-24T12:47:49.000+05:30),(9743251,true,2014-10-31T15:20:56.000+05:30),(9774198,true,2015-08-24T15:27:20.000+05:30),(9723683
,true,2014-10-31T15:20:56.000+05:30),(9783385,true,2015-09-10T11:41:00.000+05:30),(9754282,true,2014-10-31T15:20:56.000+05:30),(9723993,true,2014-10-31T15:20:56.000+
05:30),(9723767,true,2014-10-31T15:20:56.000+05:30),(9723879,true,2014-10-31T15:20:56.000+05:30),(9757158,true,2014-10-31T15:20:56.000+05:30),(9724038,true,2014-10-3
1T15:20:56.000+05:30),(9723984,true,2014-10-31T15:20:56.000+05:30),(9724179,true,2014-12-11T12:46:30.000+05:30),(9724273,true,2014-10-31T15:20:56.000+05:30),(9724283
,true,2014-10-31T15:20:56.000+05:30),(9724436,true,2014-10-31T15:20:56.000+05:30),(9746880,true,2015-08-24T15:27:20.000+05:30),(9784023,true,2014-10-31T15:20:56.000+
05:30),(9726782,true,2014-12-21T12:42:21.000+05:30),(9724574,true,2014-10-31T15:20:56.000+05:30),(9724592,true,2014-10-31T15:20:56.000+05:30),(9725679,true,2014-10-3
1T15:20:56.000+05:30),(9724848,true,2014-12-14T12:37:52.000+05:30),(9725121,true,2014-10-31T15:20:56.000+05:30),(9733606,true,2014-10-31T15:20:56.000+05:30),(9724860
,true,2014-10-31T15:20:56.000+05:30),(9742513,true,2014-10-31T15:20:56.000+05:30),(9725182,true,2014-10-31T15:20:56.000+05:30),(9725645,true,2014-10-31T15:20:56.000+
05:30),(9734416,true,2014-11-04T12:43:25.000+05:30),(9725843,true,2014-10-31T15:20:56.000+05:30),(9737295,true,2014-10-31T15:20:56.000+05:30),(9725512,true,2014-10-3
```

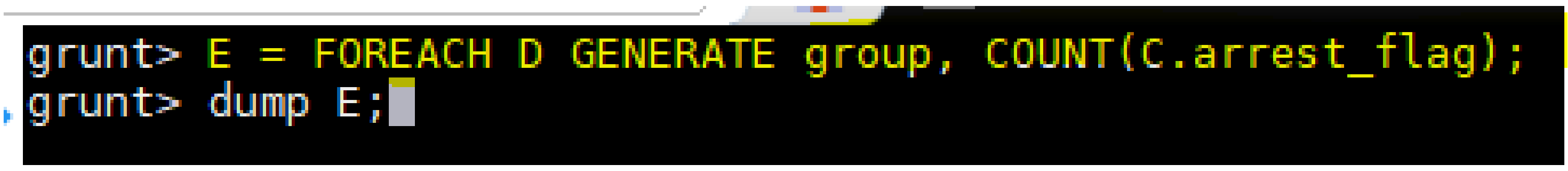
Below screenshot shows the schema for relation D



```
grunt> describe D;
D: {group: chararray,C: {(case_id: int,arrest_flag: chararray,date1: datetime)}}
grunt>
```

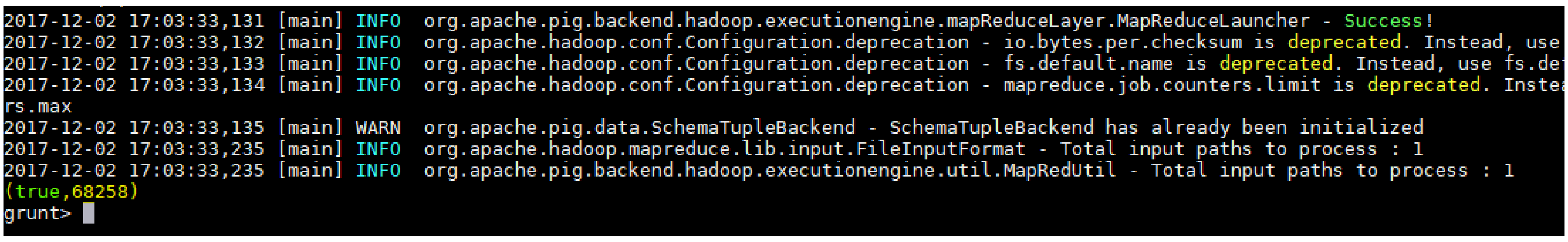
- **E = FOREACH D GENERATE group, COUNT(C.arrest_flag);**

Now we are generating count of arrest_flag which was grouped in relation D-



```
grunt> E = FOREACH D GENERATE group, COUNT(C.arrest_flag);
grunt> dump E;
```

Below screenshot shows the final result of the number of arrests done between October 2014 and October 2015 which is 68258-



```
2017-12-02 17:03:33,131 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-02 17:03:33,132 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
2017-12-02 17:03:33,133 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
2017-12-02 17:03:33,134 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instea
rs.max
2017-12-02 17:03:33,135 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-02 17:03:33,235 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-02 17:03:33,235 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(true,68258)
grunt>
```