**Project 1.1 USA Crime Analysis**

(014,228)

* **Registering the Piggy bank jar utility (UDF)**

grunt> REGISTER '/home/acadgild/Downloads/piggybank-0.15.0.jar';

2018-03-07 21:35:07,167 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.hours is deprecated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours

2018-03-07 21:35:07,167 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is deprecated. Instead, use mapreduce.jobtracker.heartbeats.in.second

2018-03-07 21:35:07,167 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion.poll.interval is deprecated. Instead, use mapreduce.client.completion.pollinterval

2018-03-07 21:35:07,167 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.tasks.sleeptime-before-sigkill is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebeforesigkill

grunt>



* **Loading Input crime file Data into crime\_data**

grunt> crime\_data = LOAD '/home/acadgild/Downloads/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,Domestic: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);

2018-03-07 21:36:38,433 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2018-03-07 21:36:38,433 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2018-03-07 21:36:38,433 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

grunt>



* **Problem 1: Number of cases investigated under each FBI code**

**// Grouping crime data by FBI Code**

grunt> group\_cases = GROUP crime\_data BY FBICode;

**// Generating group for group\_cases and countingFBICode**

grunt> cases\_investigated = FOREACH group\_cases GENERATE group,COUNT(crime\_data.FBICode);

**// Printing The Result**

grunt> dump group\_cases;

2018-03-07 21:51:57,264 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 21:51:57,265 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 21:51:57,268 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 21:51:57,273 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete

2018-03-07 21:51:57,273 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features

2.2.0 0.14.0 acadgild 2018-03-07 21:51:17 2018-03-07 21:51:57 GROUP\_BY

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs

job\_local780682936\_0003 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/temp-970368686/tmp503909403,

Input(s):

Successfully read 291268 records from: "/home/acadgild/Downloads/Crimes\_-\_2001\_to\_present.csv"

Output(s):

Successfully stored 27 records in: "file:/tmp/temp-970368686/tmp503909403"

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\_local780682936\_0003

2018-03-07 21:51:57,278 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 21:51:57,279 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 21:51:57,280 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 21:51:57,290 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING\_NON\_EXISTENT\_FIELD 21 time(s).

2018-03-07 21:51:57,290 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD\_DISCARDED\_TYPE\_CONVERSION\_FAILED 291267 time(s).

2018-03-07 21:51:57,290 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2018-03-07 21:51:57,291 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2018-03-07 21:51:57,291 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

2018-03-07 21:51:57,297 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2018-03-07 21:51:57,297 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized

2018-03-07 21:51:57,333 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2018-03-07 21:51:57,333 [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)

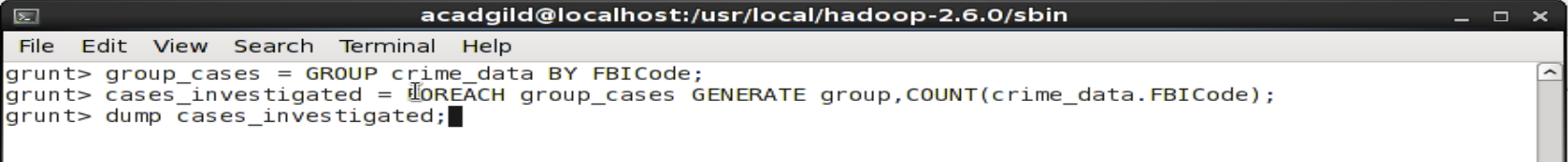
(08A,14167)

(08B,46938)

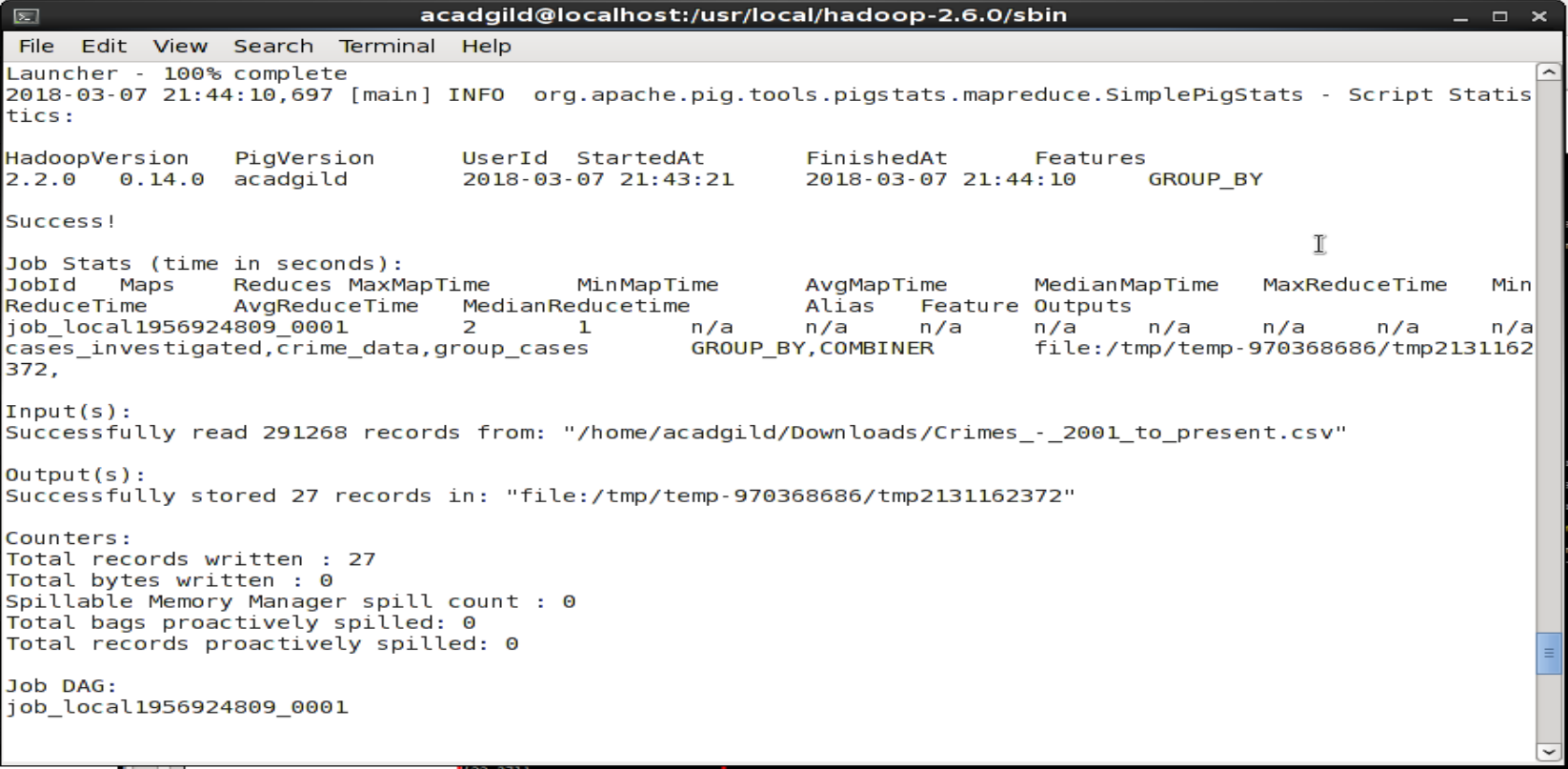
(,0)

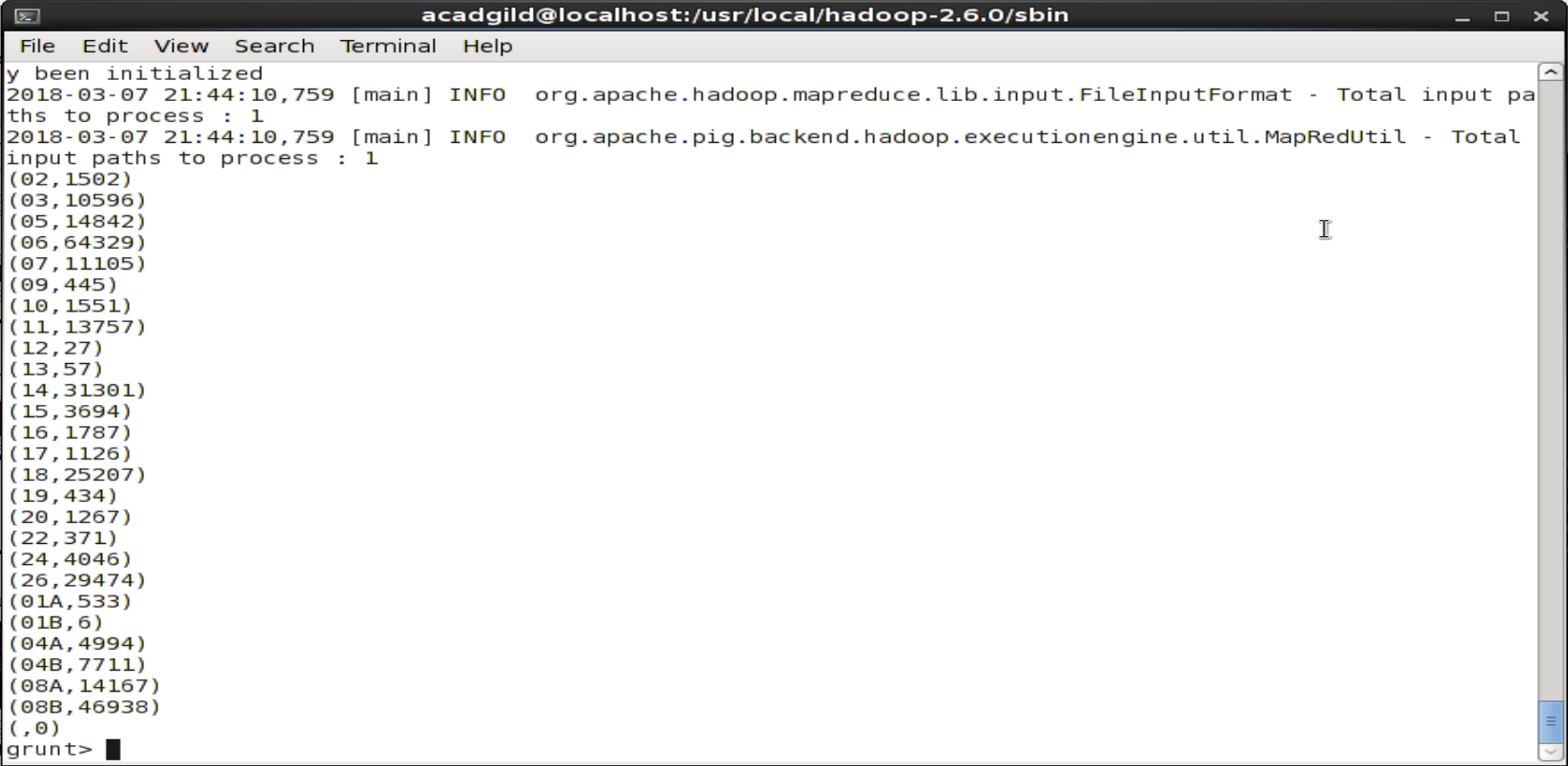
grunt>

**// Commands**



**// Problem 1 Outputs**





* **Problem 2: Number of cases investigated under FBI code 32**

**// Filtering crime data for FBICode 32**

grunt> filter\_data = FILTER crime\_data BY FBICode == '32';

**// Grouping filtered data for FBICode**

grunt> group\_filter\_crime\_data = GROUP filter\_data BY FBICode;

**// Generating group and counting the filtered data**

grunt> cases\_investigated\_FBI\_32 = FOREACH group\_filter\_data GENERATE group,COUNT(filter\_data);

**// Dumping the result**

grunt> dump cases\_investigated\_FBI\_32;

2018-03-07 22:00:14,149 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:00:14,149 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:00:14,149 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:00:14,158 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete

2018-03-07 22:00:14,158 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features

2.2.0 0.14.0 acadgild 2018-03-07 21:59:40 2018-03-07 22:00:14 GROUP\_BY,FILTER

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs

job\_local1378885033\_0004 2 1 n/a n/a n/a n/a n/a n/a n/a n/acases\_investigated\_FBI\_32,crime\_data,filter\_data,group\_filter\_data GROUP\_BY,COMBINER file:/tmp/temp-970368686/tmp-1377598463,

Input(s):

Successfully read 291268 records from: "/home/acadgild/Downloads/Crimes\_-\_2001\_to\_present.csv"

Output(s):

Successfully stored 0 records in: "file:/tmp/temp-970368686/tmp-1377598463"

Counters:

Total records written : 0

Total bytes written : 0

Spillable Memory Manager spill count : 0

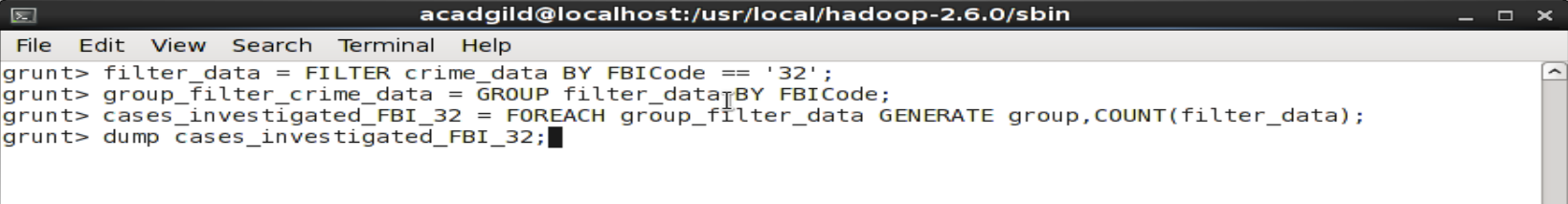
Total bags proactively spilled: 0

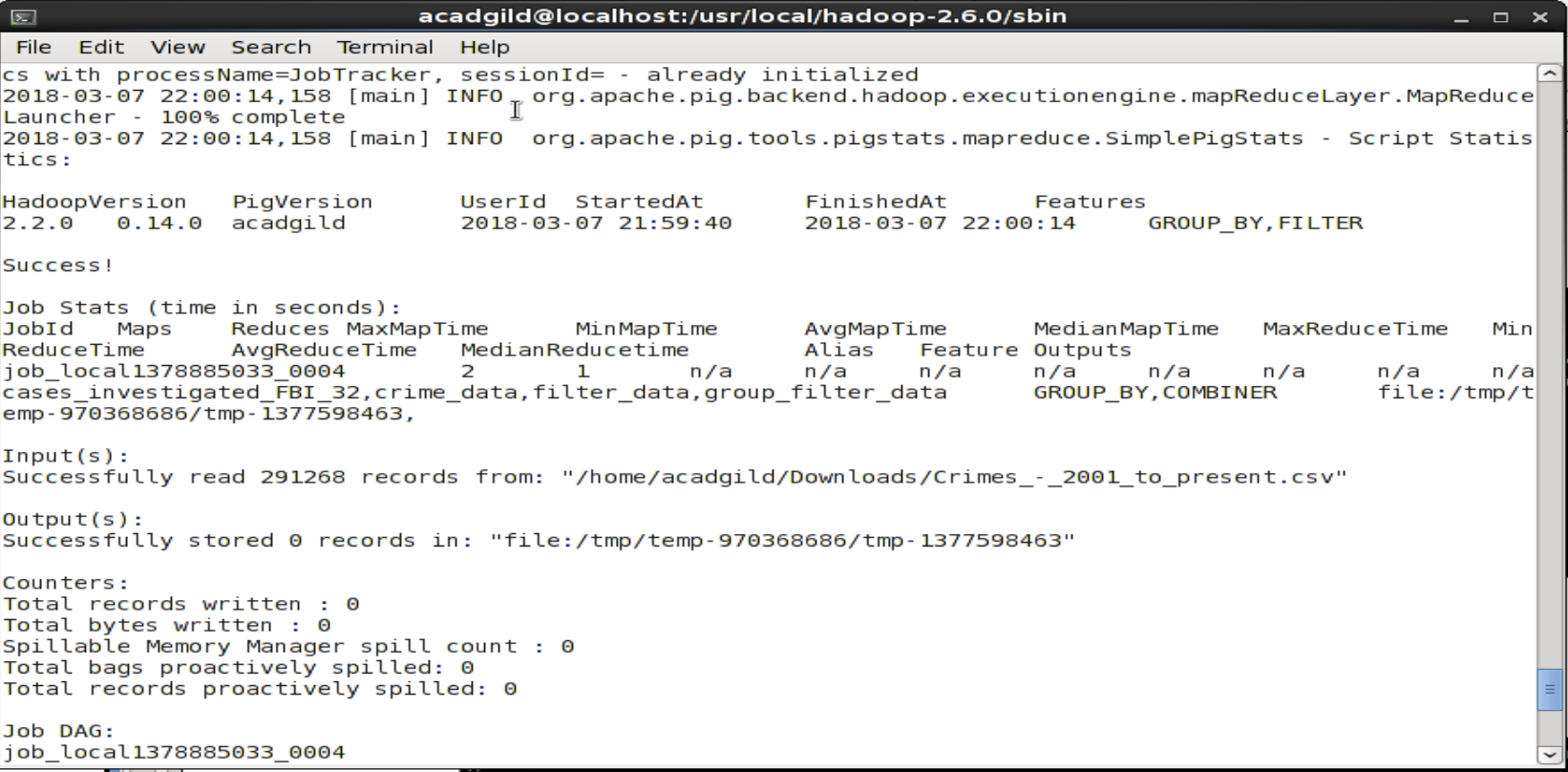
Total records proactively spilled: 0

Job DAG:

job\_local1378885033\_0004

**// Commands and Output**





* **Problem 3: Number of arrests in theft district wise**

**// Filtering data for Primary Type “theft” and arrest “true”**

grunt> filter\_data\_theft\_district = FILTER crime\_data BY Primary\_Type == 'THEFT' and Arrest == 'true';

**// Grouping the filtered data by District**

grunt> groupByDistrict = GROUP filter\_data\_theft\_district BY District;

**// Performing distinct**

grunt> distinct\_groupByDistrict = DISTINCT groupByDistrict;

**// Generating group and counting the filtered data**

grunt> arrestsTheftDistrictWise = FOREACH distinct\_groupByDistrict GENERATE group,COUNT(filter\_data\_theft\_district);

**// Dumping the Result**

grunt> dump arrestsTheftDistrictWise;

2018-03-07 22:23:51,182 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,184 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,184 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,191 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete

2018-03-07 22:23:51,198 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features

2.2.0 0.14.0 acadgild 2018-03-07 22:23:07 2018-03-07 22:23:51 GROUP\_BY,DISTINCT,FILTER

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs

job\_local1541404048\_0002 2 1 n/a n/a n/a n/a n/a n/a n/a n/a crime\_data,filter\_data\_theft\_district,groupByDistrict GROUP\_BY

job\_local376256229\_0003 1 1 n/a n/a n/a n/a n/a n/a n/a n/a arrestsTheftDistrictWise DISTINCT file:/tmp/temp34761927/tmp1902768983,

Input(s):

Successfully read 291268 records from: "/home/acadgild/Downloads/Crimes\_-\_2001\_to\_present.csv"

Output(s):

Successfully stored 22 records in: "file:/tmp/temp34761927/tmp1902768983"

Counters:

Total records written : 22

Total bytes written : 0

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job\_local1541404048\_0002 -> job\_local376256229\_0003,

job\_local376256229\_0003

2018-03-07 22:23:51,198 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,199 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,204 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,219 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,228 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,231 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:23:51,243 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING\_NON\_EXISTENT\_FIELD 21 time(s).

2018-03-07 22:23:51,243 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD\_DISCARDED\_TYPE\_CONVERSION\_FAILED 291267 time(s).

2018-03-07 22:23:51,243 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2018-03-07 22:23:51,243 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2018-03-07 22:23:51,253 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

2018-03-07 22:23:51,253 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2018-03-07 22:23:51,253 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized

2018-03-07 22:23:51,285 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2018-03-07 22:23:51,285 [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)

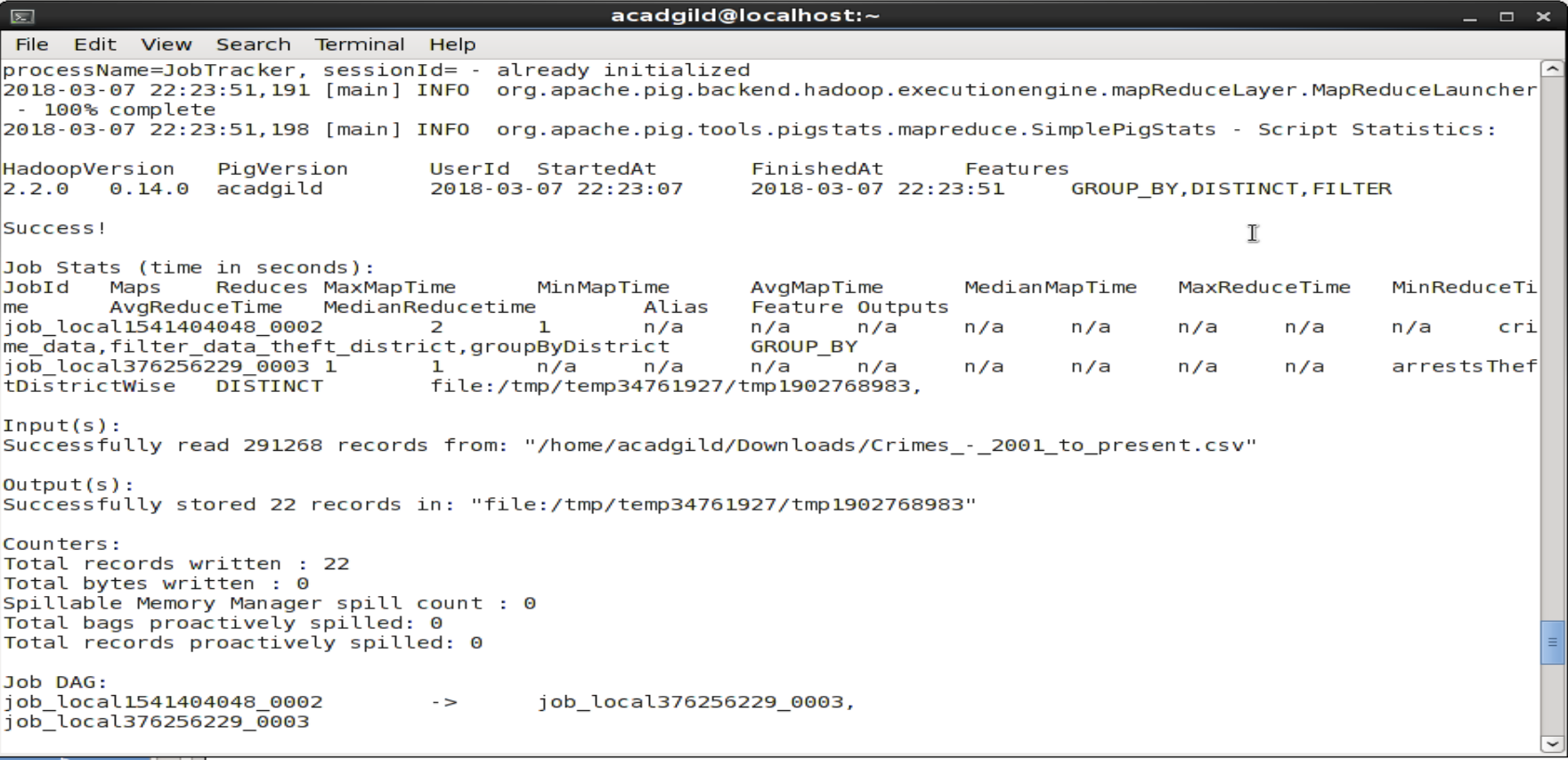
(022,220)

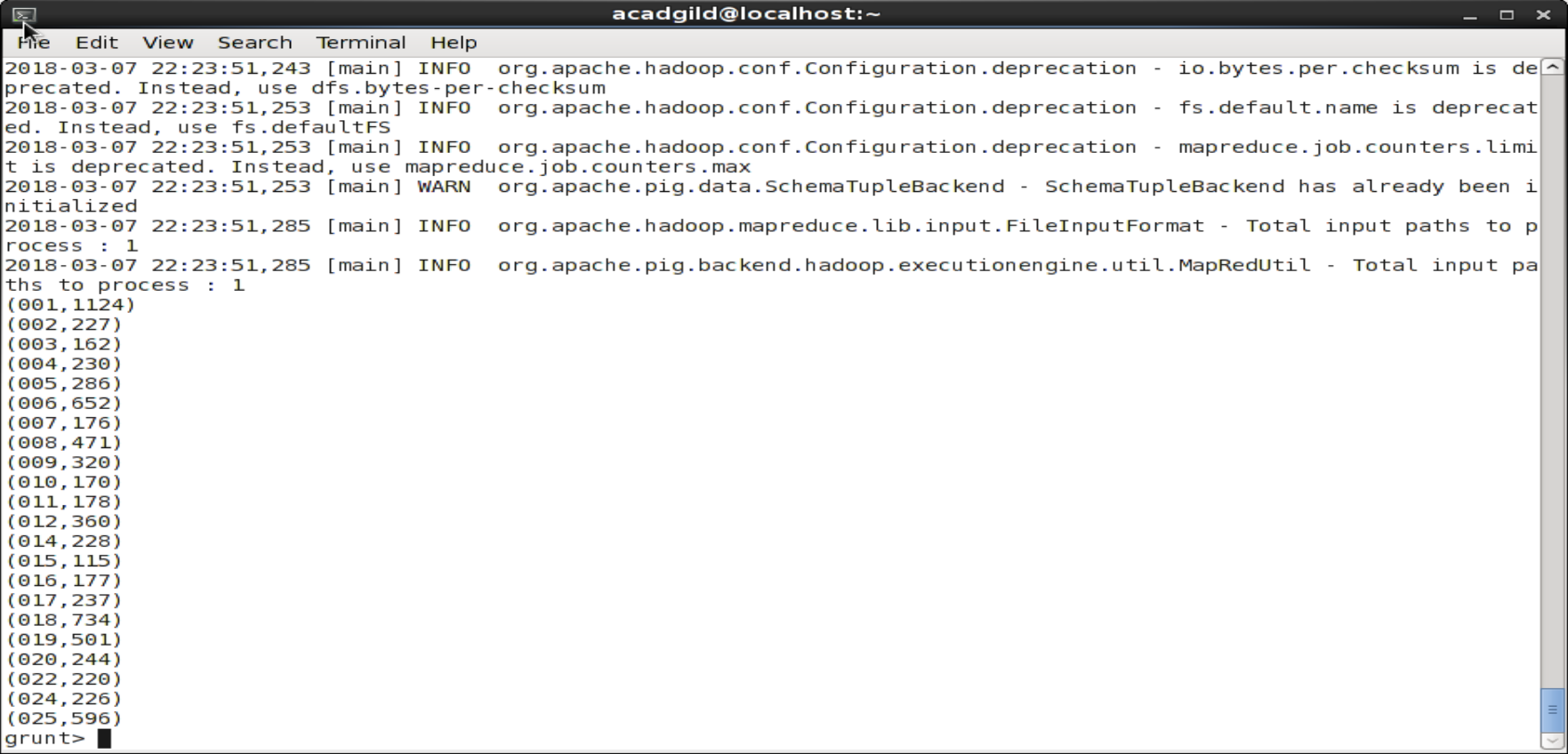
(024,226)

(025,596)

grunt>

**// Problem 3 Outputs**





* **Problem 4:** **Number of arrests done between October 2014 and October 2015**

**// Filtering the data for October 2014 and 2015**

grunt> crime\_data\_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');

grunt> filter\_data\_arrest = FILTER crime\_data\_between\_dates BY Arrest == 'true';

**// Grouping the filtered data**

grunt> group\_filter\_data\_arrest = GROUP filter\_data\_arrest ALL;

**// Generating group and counting the arrest**

grunt> arrestOctober1415 = FOREACH group\_filter\_data\_arrest GENERATE group, COUNT(filter\_data\_arrest.Arrest);

**// Dumping the result**

grunt> dump arrestOctober1415;

2018-03-07 22:32:39,756 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:32:39,760 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:32:39,768 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:32:39,783 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete

2018-03-07 22:32:39,783 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features

2.2.0 0.14.0 acadgild 2018-03-07 22:31:50 2018-03-07 22:32:39 GROUP\_BY,FILTER

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs

job\_local1840896008\_0004 2 1 n/a n/a n/a n/a n/a n/a n/a n/a arrestOctober1415,crime\_data,crime\_data\_between\_dates,group\_filter\_data\_arrest GROUP\_BY,COMBINER file:/tmp/temp34761927/tmp519874898,

Input(s):

Successfully read 291268 records from: "/home/acadgild/Downloads/Crimes\_-\_2001\_to\_present.csv"

Output(s):

Successfully stored 1 records in: "file:/tmp/temp34761927/tmp519874898"

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\_local1840896008\_0004

2018-03-07 22:32:39,784 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:32:39,789 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:32:39,789 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2018-03-07 22:32:39,803 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING\_NON\_EXISTENT\_FIELD 21 time(s).

2018-03-07 22:32:39,803 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD\_DISCARDED\_TYPE\_CONVERSION\_FAILED 291267 time(s).

2018-03-07 22:32:39,803 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2018-03-07 22:32:39,803 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2018-03-07 22:32:39,803 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

2018-03-07 22:32:39,803 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2018-03-07 22:32:39,803 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized

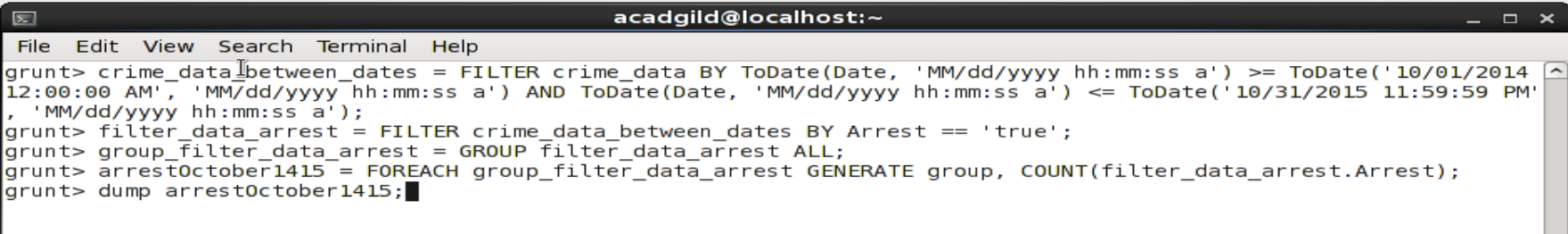
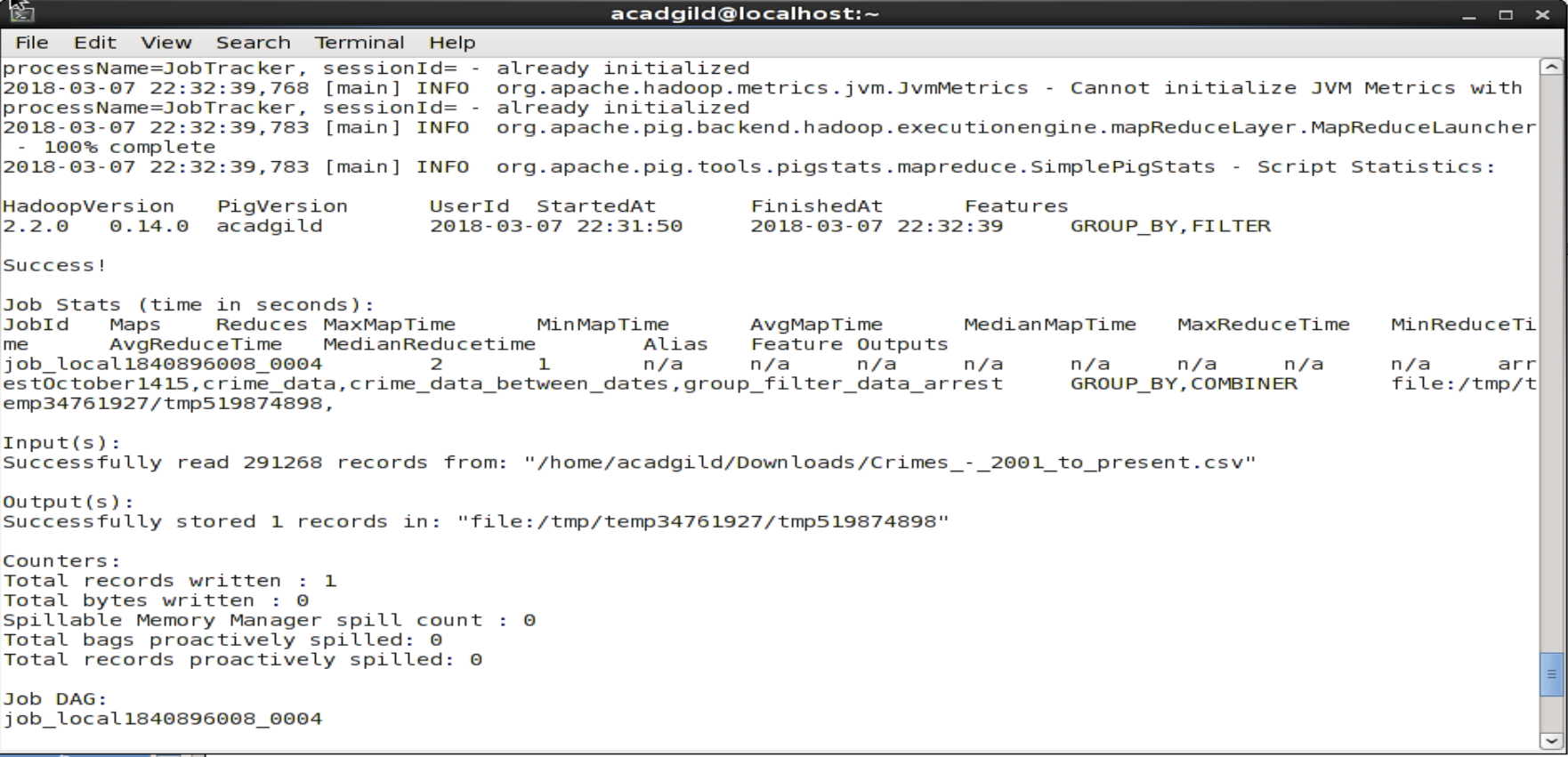
2018-03-07 22:32:39,854 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

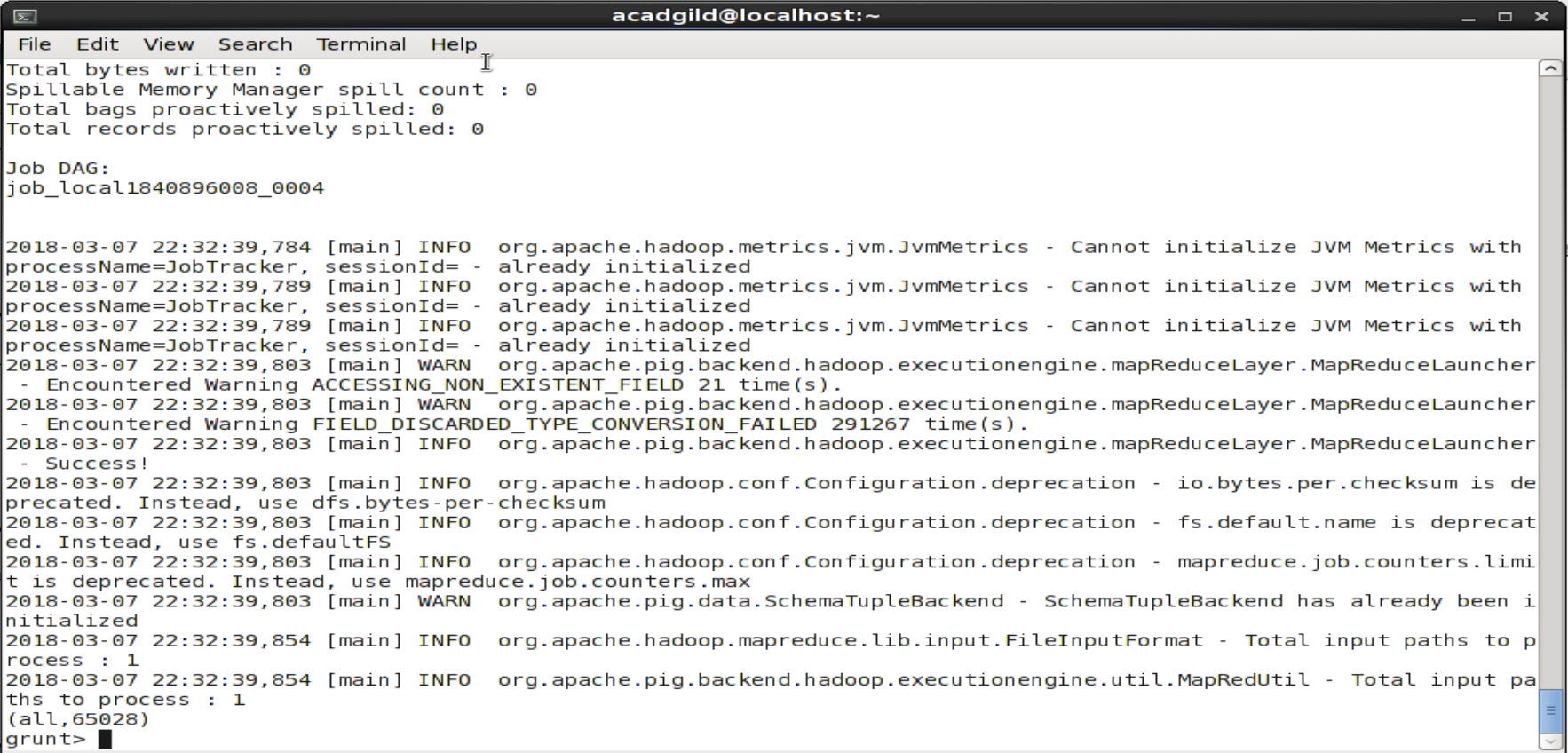
2018-03-07 22:32:39,854 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(all,65028)

grunt>

**// Commands and Outputs**

1

5,115)

(016,177)

(017

(018,734)

(019,501)

(020,244)

(022,220)