

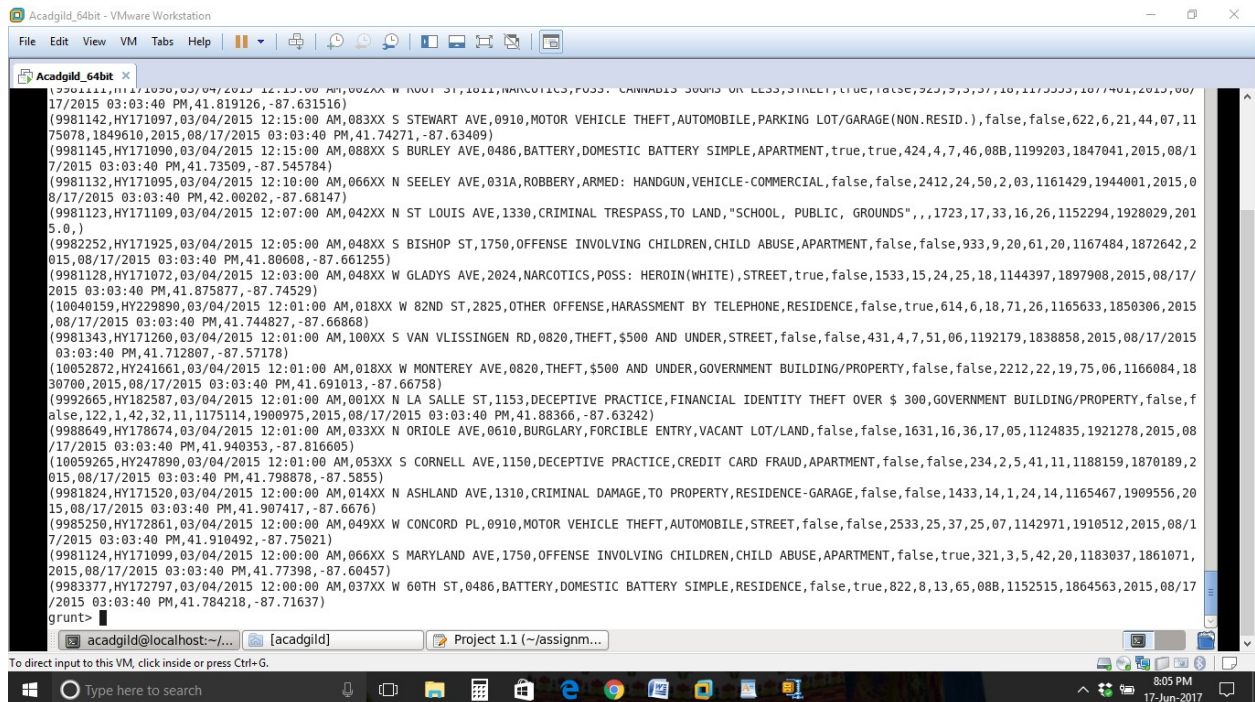
Project 1.1

USA Crime Analysis

Crime data from 2001 is being loaded in the name `crime_data_2001`

```
crime_data_2001 = load 'Crimes_-_2001_to_present.csv' USING PigStorage(',') AS (case_id:int,
case_number:chararray, case_date:chararray, block:chararray, IUCR:chararray,
primary_type:chararray, case_descr:chararray, loc_descr:chararray, arrest:chararray,
domestic:chararray, beat:int, district:int, ward:int, com_area:int, FBI_code:chararray,
x_coord:long, y_coord:long, year:int, update_on:chararray,lati:float, longi:float);
```

`dump crime_data_2001;`



The screenshot shows a VMware Workstation window titled "Acadgild_64bit - VMware Workstation". Inside the VM, a terminal window is open, displaying a large volume of crime data records. The records are formatted as CSV-like strings, including case IDs, case numbers, case dates, block numbers, IUCR codes, primary types, case descriptions, location descriptions, arrest status, domestic status, beats, districts, wards, community areas, FBI codes, coordinates, years, update dates, latitude, and longitude. The terminal window has a title bar "Acadgild_64bit" and a menu bar with "File", "Edit", "View", "VM", "Tabs", and "Help". The bottom of the window shows a taskbar with various application icons and a system clock indicating 8:05 PM on 17-Jun-2017.

Problem Statement

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

FBI code

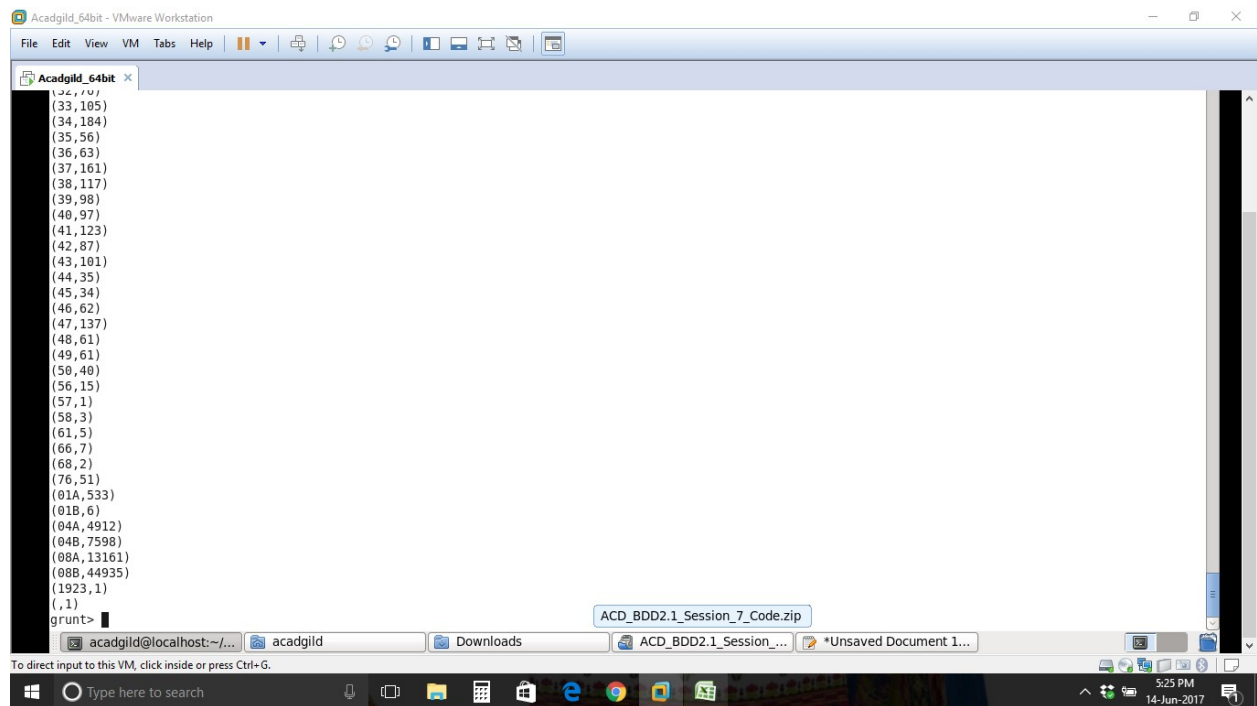
```
crime_by_fbicode = group crime_data_2001 by FBI_code;
```

```
dump crime_by_fbicode;
```

```
crime_count_fbicode = foreach crime_by_fbicode generate group as fbi_code,  
COUNT(crime_data_2001.case_id);
```

```
dump crime_count_fbicode
```

Number of Crimes under each FBI code



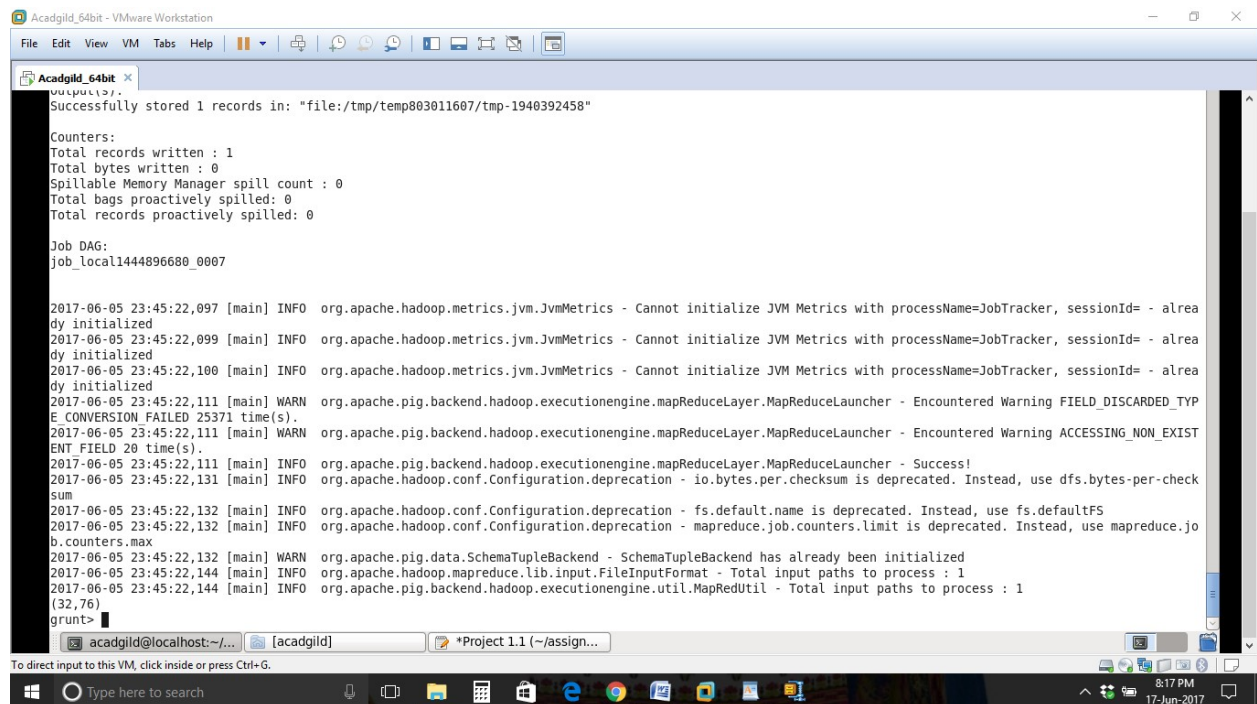
Problem Statement

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

crime_by_code_32 = filter crime_count_fbicode by fbi_code matches '32';

dump_crime_by_code_32;

Number of Crimes Under FBI_code 32



```
Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit x
Successfully stored 1 records in: "file:/tmp/temp803011607/tmp-1940392458"

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_local1444896680_0007

2017-06-05 23:45:22,097 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-05 23:45:22,099 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-05 23:45:22,100 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-05 23:45:22,111 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYP
E CONVERSION FAILED 25371 time(s).
2017-06-05 23:45:22,111 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXIST
ENT FIELD 20 time(s).
2017-06-05 23:45:22,111 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-05 23:45:22,131 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-06-05 23:45:22,132 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-05 23:45:22,132 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-06-05 23:45:22,132 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-05 23:45:22,144 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-05 23:45:22,144 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(32,76)
grunt>
```

Problem Statement

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

crime_arrest_district = foreach crime_data_2001 generate district as district , arrest as arrest;

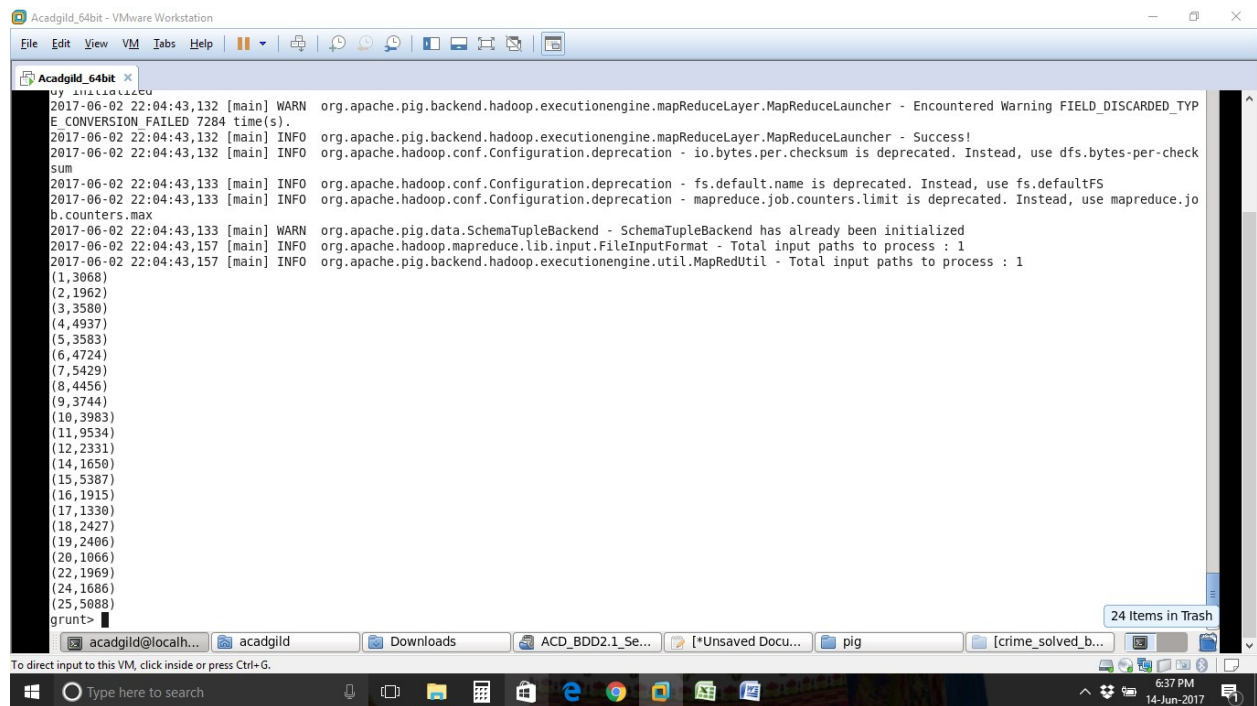
crime_arrest_done_district = filter crime_arrest_district by arrest matches 'true';

arrest_district_group = group crime_arrest_done_district by district;

arrest_count_by_district = foreach arrest_district_group generate group as district , COUNT(\$1) as no_of_arrest;

dump arrest_count_by_district;

Number of Arrests done district wise



```
Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit x
2017-06-02 22:04:43,132 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYP
E_CONVERSION_FAILED 7284 time(s).
2017-06-02 22:04:43,132 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-02 22:04:43,132 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-06-02 22:04:43,133 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-02 22:04:43,133 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job
b.counters.max
2017-06-02 22:04:43,133 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-02 22:04:43,157 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-02 22:04:43,157 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,3068)
(2,1962)
(3,3580)
(4,4937)
(5,3583)
(6,4724)
(7,5429)
(8,4456)
(9,3744)
(10,3983)
(11,9534)
(12,2331)
(14,1650)
(15,5387)
(16,1915)
(17,1330)
(18,2427)
(19,2406)
(20,1066)
(22,1969)
(24,1686)
(25,5088)
grunt>
acacgild@localh... acacgild Downloads ACD_BDD2.1_Se... [*]Unsaved Docu... pig [crime_solved_b...
To direct input to this VM, click inside or press Ctrl+G.
Type here to search 6:37 PM 14-Jun-2017
```

Problem Statement

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

```
crime_by_date = foreach crime_data_2001 generate case_id as id ,
ToDate(case_date, 'MM/dd/yyyy hh:mm:ss aa') as date, arrest as arrest;

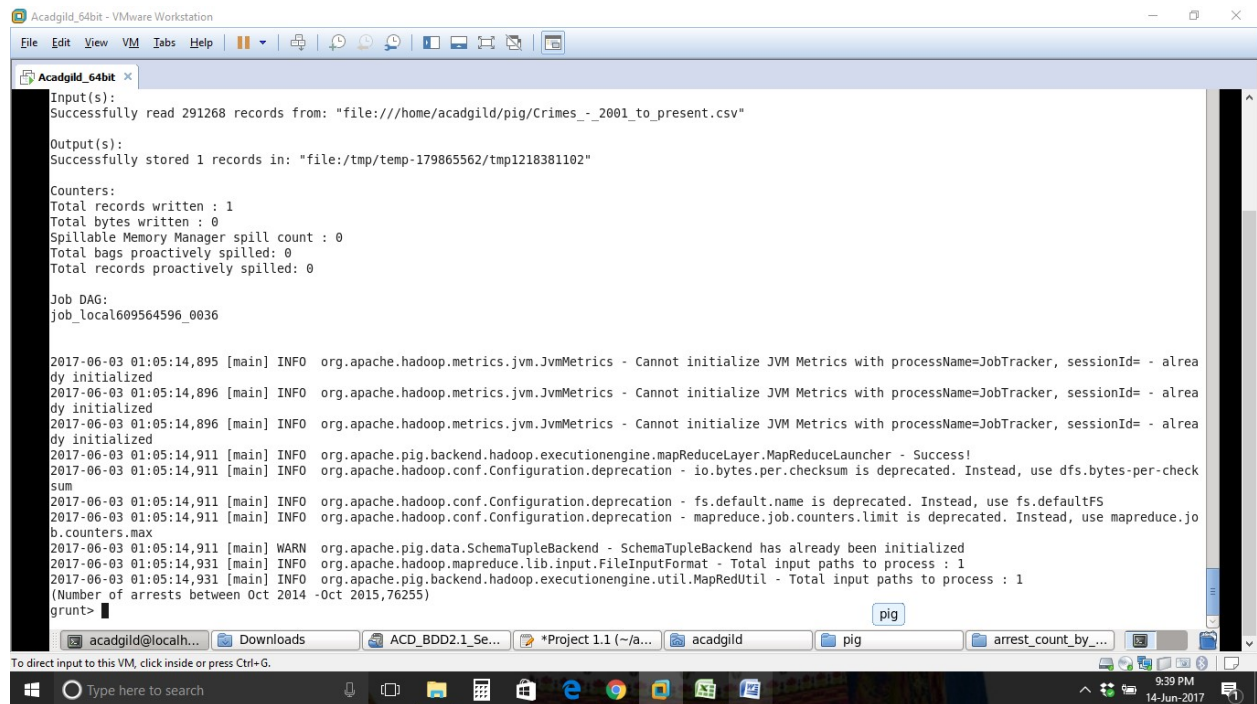
crime_by_date1 = filter crime_by_date by (GetYear(date)>=2014 and
GetMonth(date)>=10) or (GetYear(date)<=2015 and GetMonth(date)<=10);

crime_by_date2 = filter crime_by_date1 by arrest matches 'true';

crime_by_date3 = group crime_by_date2 all;

crime_by_date4 = foreach crime_by_date3 generate 'Number of arrests
between Oct 2014 -Oct 2015', COUNT(crime_by_date2.id);

dump crime_by_date4;
```



```
Acadgild_64bit - VMWare Workstation
File Edit View VM Tabs Help
Acadgild_64bit x
Input(s):
Successfully read 291268 records from: "file:///home/acadgild/pig/Crimes_-_2001_to_present.csv"
Output(s):
Successfully stored 1 records in: "file:///tmp/temp-179865562/tmp1218381102"
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_local609564596_0036
2017-06-03 01:05:14,895 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-03 01:05:14,896 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-03 01:05:14,896 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-03 01:05:14,911 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-03 01:05:14,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-06-03 01:05:14,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-03 01:05:14,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-06-03 01:05:14,911 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-03 01:05:14,931 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-03 01:05:14,931 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Number of arrests between Oct 2014 -Oct 2015,76255)
grunt>
pig
acadgild@localh... Downloads ACD_BDD2.1_Se... *Project 1.1 (~/a... acadgild pig arrest_count_by_...
To direct input to this VM, click inside or press Ctrl+G.
Type here to search 9:39 PM 14-Jun-2017
```