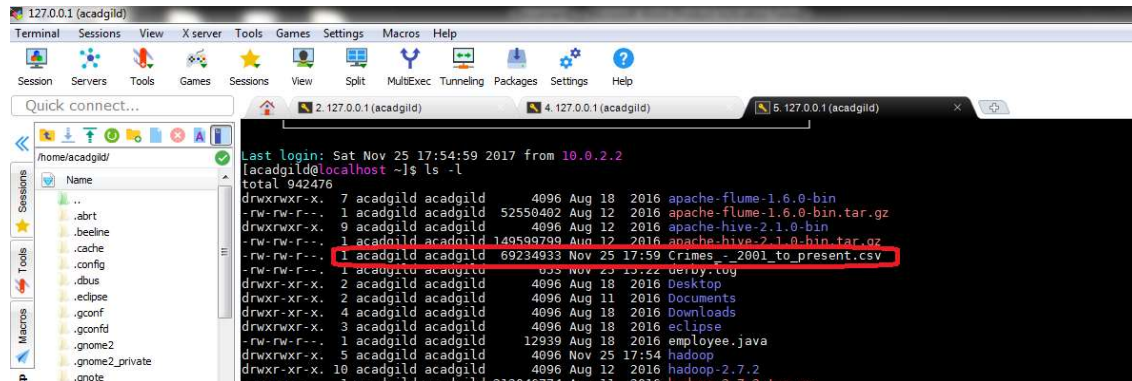


Project 1.1 – USA Crime Analysis

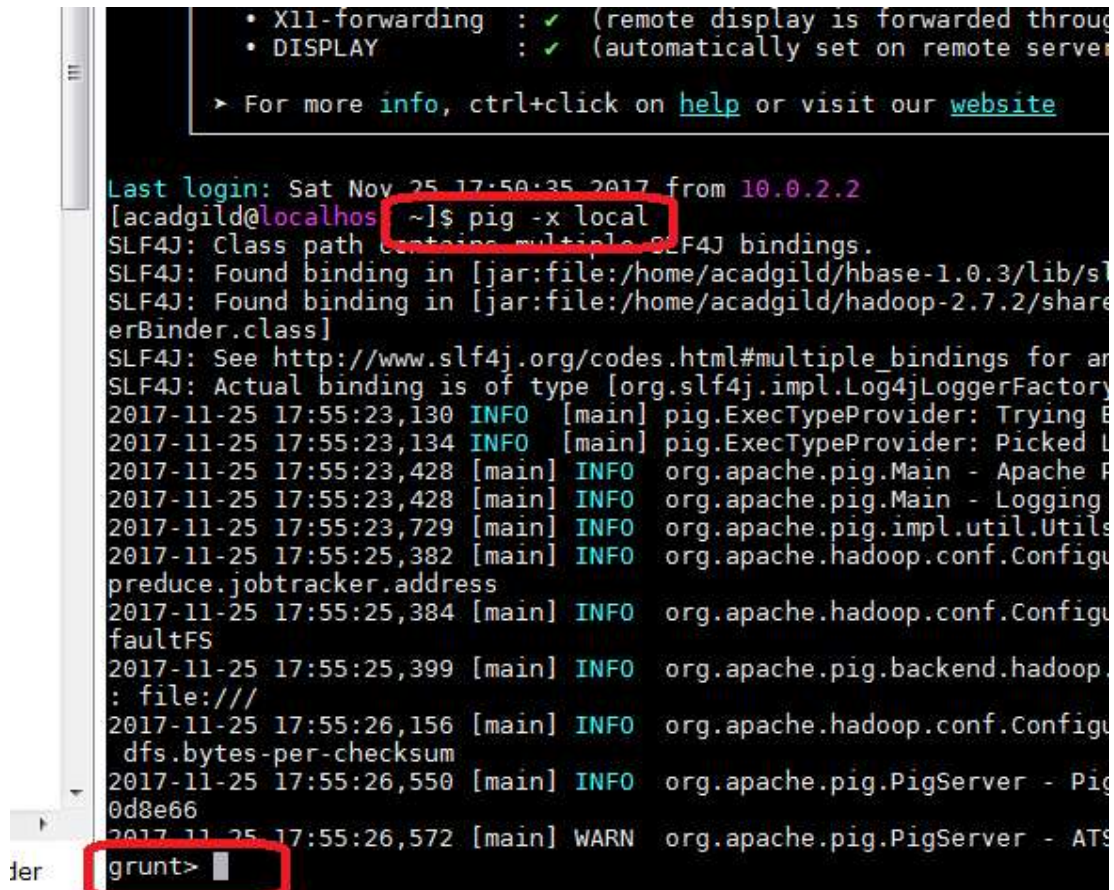
File uploaded



The screenshot shows a terminal window with a file upload summary. The file `Crimes - 2001 to present.csv` is highlighted with a red box. The terminal output includes the following table:

File Name	Size	Owner	Group	Mode	Permissions	Timestamp	Source
Crimes - 2001 to present.csv	69234933	acadgild	acadgild	-rw-rw-r--	1	Nov 25 17:59	10.0.2.2

Pig Started



The screenshot shows a terminal window with the output of a Pig execution. The command `pig -x local` is highlighted with a red box. The output includes the following log entries:

```
Last login: Sat Nov 25 17:50:35 2017 from 10.0.2.2
[acadgild@localhost ~]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12.jar:!/org.slf4j.impl.Log4jLoggerFactory.class]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12.jar:!/org.slf4j.impl.Log4jLoggerFactory.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-11-25 17:55:23,130 INFO [main] pig.ExecTypeProvider: Trying to load org.apache.pig.backend.hadoop.mapreduce.jobtracker.address
2017-11-25 17:55:23,134 INFO [main] pig.ExecTypeProvider: Picked org.apache.pig.backend.hadoop.mapreduce.jobtracker.address
2017-11-25 17:55:23,428 [main] INFO org.apache.pig.Main - Apache Pig
2017-11-25 17:55:23,428 [main] INFO org.apache.pig.Main - Logging
2017-11-25 17:55:23,729 [main] INFO org.apache.pig.impl.util.Util - Logging
2017-11-25 17:55:25,382 [main] INFO org.apache.hadoop.conf.Configuration - Logging
2017-11-25 17:55:25,384 [main] INFO org.apache.hadoop.conf.Configuration - Logging
2017-11-25 17:55:25,399 [main] INFO org.apache.pig.backend.hadoop.mapreduce.jobtracker.address
2017-11-25 17:55:26,156 [main] INFO org.apache.hadoop.conf.Configuration - Logging
2017-11-25 17:55:26,550 [main] INFO org.apache.pig.PigServer - Pig
2017-11-25 17:55:26,572 [main] WARN org.apache.pig.PigServer - AT
grunt>
```

Assignment 1.1.1

```
grunt> baseRel = LOAD 'home/acadgild/Crimes_-_2001_to_present.csv' using PigStorage(',') AS (ID:int,
Case_Number:int, Date:chararray, Block:chararray, IUCR:int, Primar_Type:chararray, Description:chararray,
Loc_Description:chararray, Arrest:chararray, Domestic:chararray, Beat:chararray, District:chararray,
Ward:chararray, Community_Are:chararray, FBIcode:int, X_Coord:chararray, Y_Coord:chararray,
Year:chararray, updated_on:chararray, latitude:chararray, longitude:chararray, location:chararray);
```

```
grunt> baseRel = LOAD 'home/acadgild/Crimes_-_2001_to_present.csv' using PigStorage(',') AS (ID:int, Case_Number:int, Date:chararray, Block:chararray, IUCR:int, Primar_Type:chararray, Description:chararray, Loc_Description:chararray, Arrest:chararray, Domestic:chararray, Beat:chararray, District:chararray, Ward:chararray, Community_Are:chararray, FBIcode:int, X_Coord:chararray, Y_Coord:chararray, Year:chararray, updated_on:chararray, latitude:chararray, longitude:chararray, location:chararray);
2017-11-25 18:39:20,884 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-25 18:39:20,885 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

```
grunt> DESCRIBE baseRel;
baseRel: {ID: int,Case_Number: chararray,Date: chararray,Block: chararray,IUCR: int,Primar_Type: chararray,Description: chararray,Loc_Description: chararray,Arrest: chararray,Domestic: chararray,Beat: chararray,District: chararray,Ward: chararray,Community_Are: chararray,FBIcode: chararray,X_Coord: chararray,Y_Coord: chararray,Year: chararray,updated_on: chararray,latitude: chararray,longitude: chararray,location: chararray}
grunt>
```

```
grunt> DESCRIBE baseRel;
baseRel: {ID: int,Case_Number: chararray,Date: chararray,Block: chararray,IUCR: int,Primar_Type: chararray,Description: chararray,Loc_Description: chararray,Arrest: chararray,Domestic: chararray,Beat: chararray,District: chararray,Ward: chararray,Community_Are: chararray,FBIcode: chararray,X_Coord: chararray,Y_Coord: chararray,Year: chararray,updated_on: chararray,latitude: chararray,longitude: chararray,location: chararray}
grunt> selRel = FOREACH baseRel GENERATE Case_Number, FBIcode;
grunt> grpRel = GROUP selRel by FBIcode;
grunt> resRel = FOREACH grpRel GENERATE group, COUNT(selRel.Case_Number);
grunt> STORE resRel INTO '/home/acadgild/maxout' using PigStorage(',');
```

Results:

```

Counters:
Total records written : 71
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1190124156_0006

2017-11-25 19:46:43,819 [main] INFO org.apache
, sessionId= - already initialized
2017-11-25 19:46:43,979 [main] INFO org.apache
, sessionId= - already initialized
2017-11-25 19:46:44,049 [main] INFO org.apache
, sessionId= - already initialized
2017-11-25 19:46:44,184 [main] WARN org.apache
ACCESSING_NON_EXISTENT_FIELD 2 time(s).
2017-11-25 19:46:44,184 [main] INFO org.apache
2017-11-25 19:46:44,222 [main] INFO org.apache
dfs.bytes-per-checksum
2017-11-25 19:46:44,231 [main] INFO org.apache
faultFS
2017-11-25 19:46:44,232 [main] WARN org.apache
2017-11-25 19:46:45,803 [main] INFO org.apache
2017-11-25 19:46:45,855 [main] INFO org.apache
(1,172)
(2,362)
(3,266)
(4,154)
(5,107)

```

Data stored in Maxout directory

```

8,301
9,192
02,1480
03,10552
05,14735
06,62826
07,10520
09,437
10,1708
11,13637
12,79
13,151
14,31244
15,3780
16,1949
17,1165
18,24989
19,590
20,1435
21,293
22,483
23,77
24,4114
25,142
26,29009
27,175
28,385
29,196
30,115
31,93
32,78
"part-r-00000" 71L, 490C

```

Assignment 1.1.2

Filter by FBI Code == 35

```

grunt> DESCRIBE baseRel;
baseRel: {ID: int,Case_Number: chararray,Date: chararray,Block: chararray,IUCR: int,Primar_Type: chararray,Description: ch
cription: chararray,Arrest: chararray,Domestic: chararray,Beat: chararray,District: chararray,Ward: chararray,Community_Ar
ICode: chararray,X_Coord: chararray,Y_Coord: chararray,Year: chararray,updated_on: chararray,latitude: chararray,longitude
ation: chararray}
grunt> Describe selRel;
selRel: {Case_Number: chararray,FBIcode: chararray}
grunt> filRel = FILTER selRel BY FBIcode == '35';
grunt>

```

```

grunt> DESCRIBE grpRel;
2017-11-25 21:52:57,355 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.by
dfs.bytes-per-checksum
2017-11-25 21:52:57,360 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.de
faultFS
grpRel: {group: chararray,filRel: {(Case_Number: chararray,FBIcode: chararray)}}
grunt> resRel = FOREACH grpRel GENERATE group, COUNT(filRel.FBIcode);
grunt>

```



```

2017-11-25 21:59:3
2017-11-25 21:59:3
2017-11-25 21:59:3
2017-11-25 21:59:3
(35,56)
grunt>

```

Assignment 1.1.3

Calculate number of arrests in theft district wise

```

(9722099,ROBBERY,false,012)
(9722100,INTERFERENCE WITH PUBLIC OFFICER,true,008)
(9722101,BATTERY,false,008)
(9722069,NARCOTICS,true,011)
(9722074,NARCOTICS,true,011)
(9722050,ASSAULT,false,012)
(9722121,WEAPONS VIOLATION,true,006)
(9722082,THEFT,false,025)
(9722060,NARCOTICS,true,011)
(9723061,THEFT,true,001)
(9722058,NARCOTICS,true,007)
(9756410,THEFT,false,001)
(9722748,CRIMINAL DAMAGE,false,007)
(9850619,DECEPTIVE PRACTICE,false,001)
(9,,, )
grunt>
grunt> DESCRIBE selRel;
selRel: {ID: int,Primar_Type: chararray,Arrest: chararray,District: chararray}
grunt> filRel = FILTER selRel BY (Primar_Type == 'THEFT' and Arrest == 'true');
grunt>

```

```

grunt> grpRel = GROUP filRel by District;

```

```

grunt> resRel = FOREACH grpRel GENERATE group, COUNT(filRel.District);
grunt>

```

```

2017-11-25 22:37:32,262 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-25 22:37:32,262 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(001,1119)
(002,220)
(003,157)
(004,221)
(005,273)
(006,649)
(007,172)
(008,458)
(009,318)
(010,166)
(011,174)
(012,353)
(014,227)
(015,111)
(016,171)
(017,227)
(018,732)
(019,499)
(020,241)
(022,207)
(024,224)
(025,591)

```

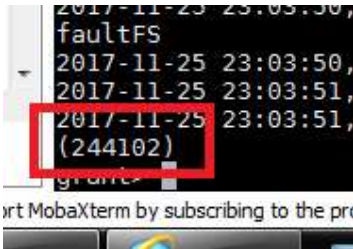
Assignment 1.1.4

Number of arrests done between Oct 2014 and Oct 2015

```

grunt> baseRel = LOAD 'Crimes_-_2001_to_present.csv' using PigStorage(',') AS (ID:int, Case_Number:chararray, Date:chararray, Block:chara
rray, IUCR:int, Primar_Type:chararray, Description:chararray, Loc_Description:chararray, Arrest:chararray, Domestic:chararray, Beat:chara
rray, District:chararray, Ward:chararray, Community_Are:chararray, FBIcode:chararray, X_Coord:chararray, Y_Coord:chararray, Year:chararra
y, updated_on:chararray, Latitude:chararray, Longitude:chararray, Location:chararray);
2017-11-25 22:51:29,920 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
dfs.bytes-per-checksum
2017-11-25 22:51:29,935 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2017-11-25 22:51:31,208 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
dfs.bytes-per-checksum
2017-11-25 22:51:31,212 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> DateGen = FOREACH baseRel GENERATE ToDate(SUBSTRING(Date,0,10), 'MM/dd/yyyy') as (dt:datetime);
grunt> FilterLDate = FILTER DateGen BY DaysBetween(dt,(datetime)ToDate('10/01/2014', 'MM/dd/yyyy')) >=(long)0;
grunt> FilterUDate = FILTER FilterLDate BY DaysBetween(dt,(datetime)ToDate('10/01/2015', 'MM/dd/yyyy')) <=(long)0;
grunt> GroupData = GROUP FilterUDate ALL;
grunt> resRel = FOREACH GroupData GENERATE COUNT(FilterUDate);
grunt>

```



```

2017-11-25 23:03:50,
faultFS
2017-11-25 23:03:50,
2017-11-25 23:03:51,
2017-11-25 23:03:51,
(244102)
grunt>

```

port MobaXterm by subscribing to the pro