**SESSION 21-24: PROJECT II**

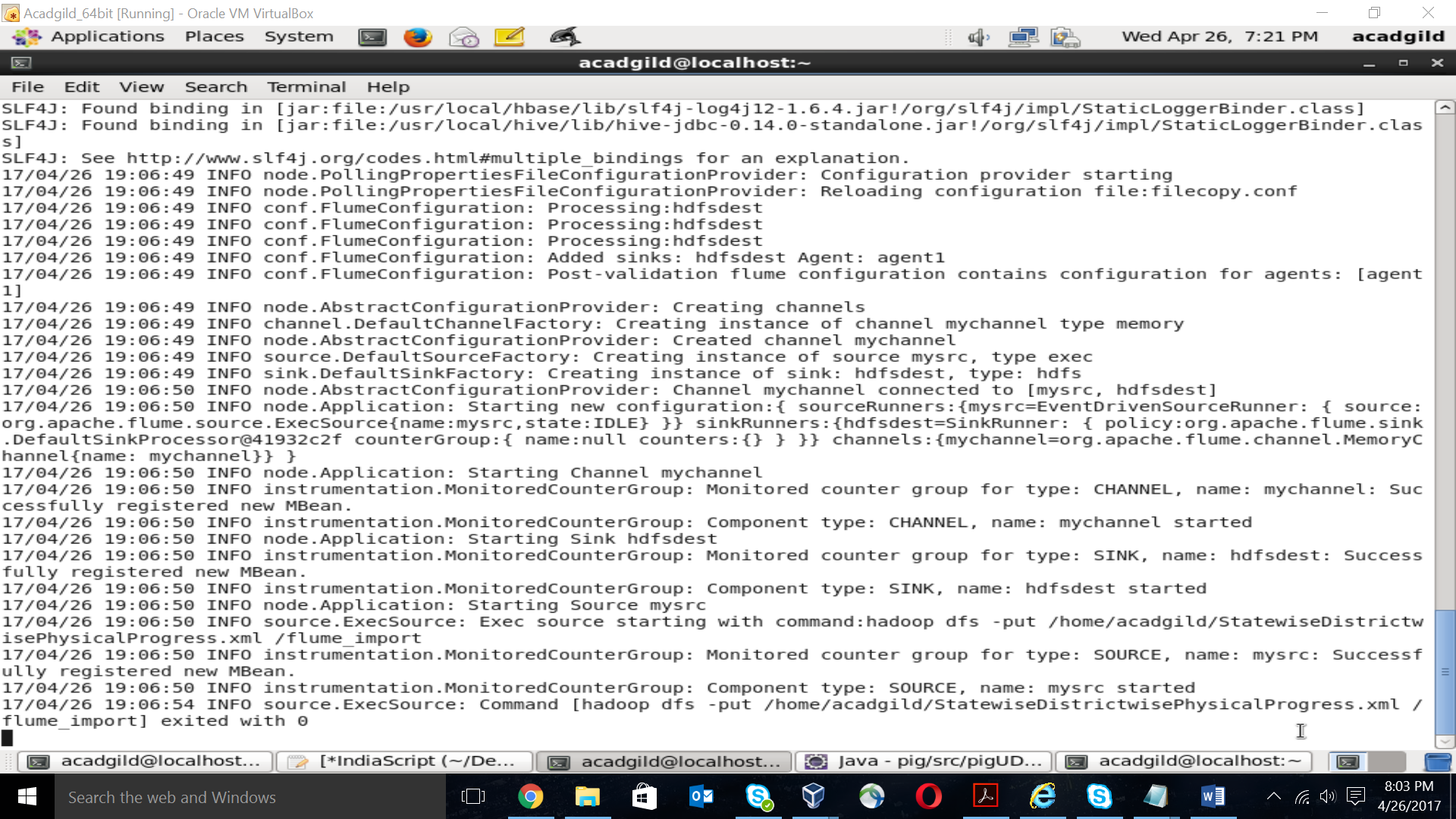
**State-Wise Development Analysis in India**

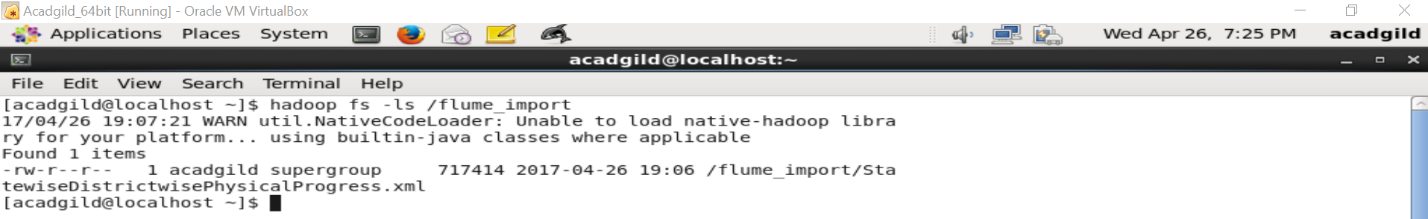
**Transferring Data into HDFS using Flume:**

Flume agent is setup in configuration file /home/acadgild/IndiaFlume.conf. The source is an exec type with command (hadoop fs –put /home/acadgild/StatewiseDistrictwisePhysicalProgress.xml /flume\_import).

flume-ng agent –n agent1 –f /home/acadgild/IndiaFlume.conf

The above command transfers the xml file to /flume\_import in HDFS.





**Processing Data through PIG:**

-- Common commands for all problems

REGISTER /usr/local/pig/lib/piggybank.jar

-- xml.jar contains the UDF for the second problem statement

REGISTER '/home/acadgild/xml.jar';

DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();

a = LOAD '/flume\_import/' USING org.apache.pig.piggybank.storage.XMLLoader('row') AS

(x:chararray);

data = FOREACH a GENERATE XPath(x,'row/District\_Name') AS district,

XPath(x,'row/Project\_Objectives\_IHHL\_BPL') AS obj, XPath(x,'row/Project\_Performance-

IHHL\_BPL') AS perf;

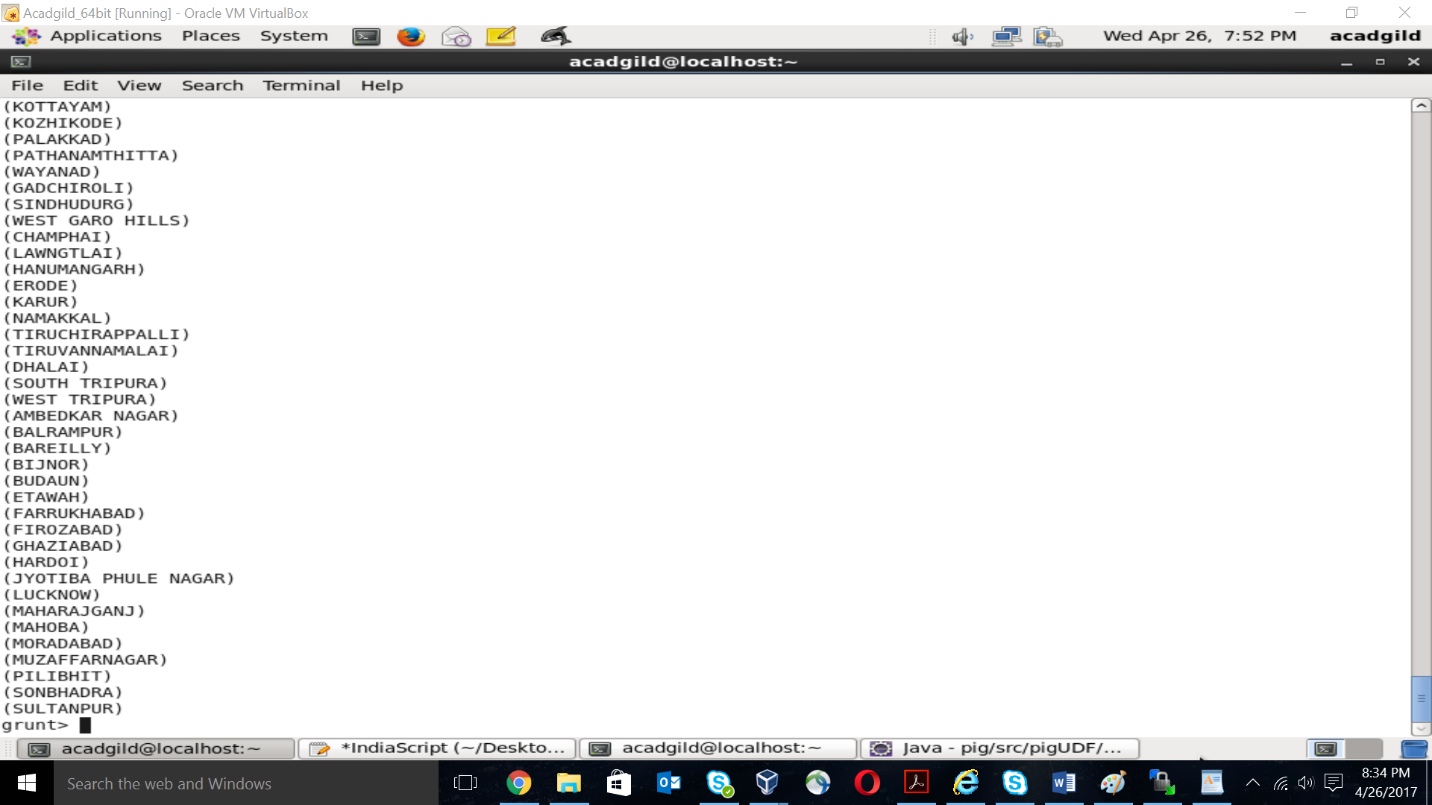
1. **Find out the districts who achieved 100% objective in BPL cards.**

filter\_data = FILTER data BY (obj == perf);

result = FOREACH filter\_data GENERATE district;

STORE result INTO ‘/output1/’ USING PigStorage(‘\t’);

dump result;



1. **Write a PIG UDF to filter the districts who have reached 80% of objectives of BPL cards.**

package pigUDF;

import java.io.IOException;

import org.apache.pig.FilterFunc;

import org.apache.pig.data.Tuple;

public class FilterUDF extends FilterFunc{

@Override

public Boolean exec(Tuple input) throws IOException {

try {

int val1 = (int) input.get(0);

int val2 = (int) input.get(1);

if(val2 >= (0.8\*val1))

return true;

else

return false;

}

catch(Exception e) {

throw new IOException(e);

}

}

}

**-- Pig grunt shell commands**

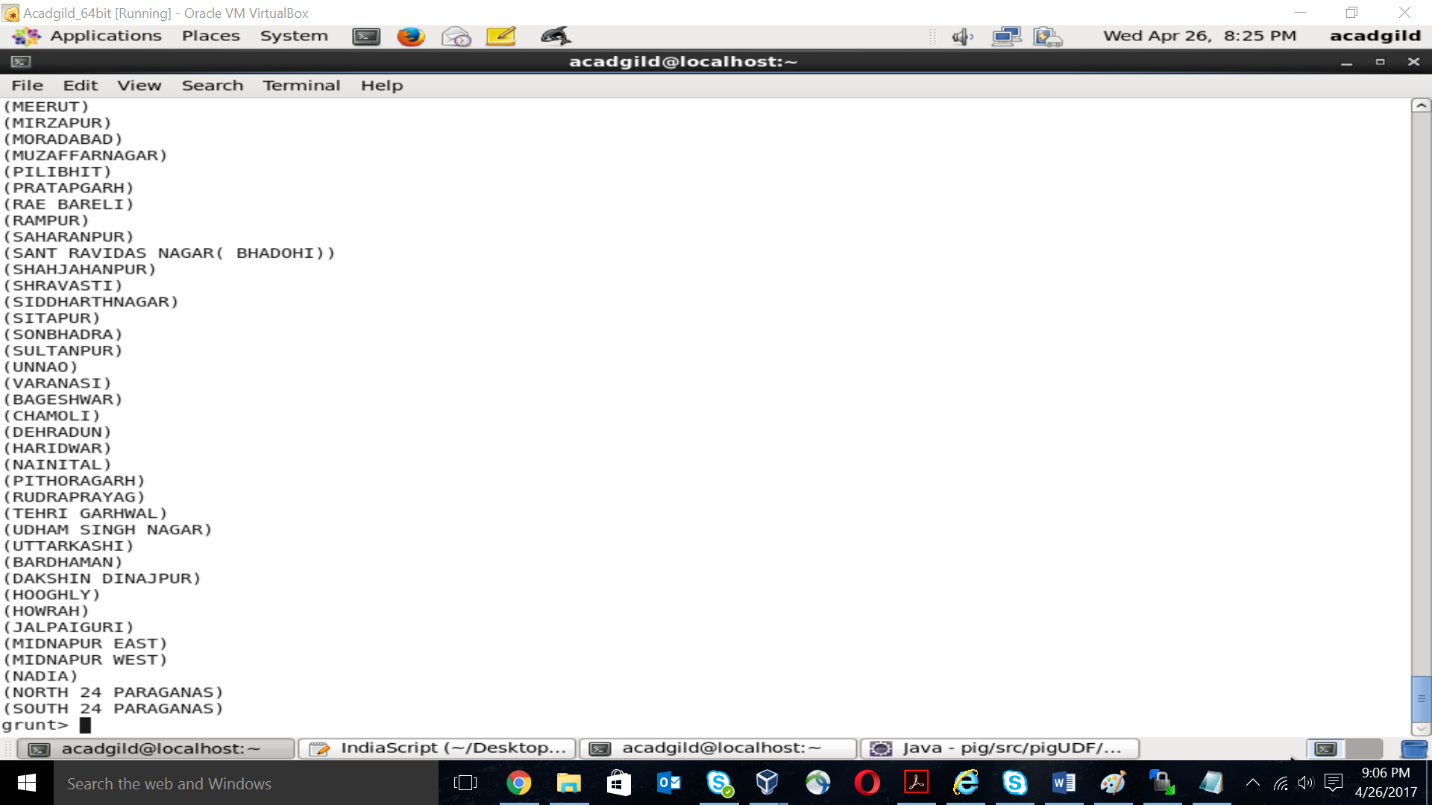
cast\_data = FOREACH data GENERATE district, (int)obj, (int)perf;

filter\_data = FILTER cast\_data BY (pigUDF.FilterUDF(obj,perf) == true);

result = FOREACH filter\_data GENERATE district;

STORE result INTO ‘/output2/’ USING PigStorage(‘\t’);

dump result;



**Exporting output files into RDBMS:**

The output of the above 2 queries were stored in 2 separate output directories in HDFS, namely /output1 and /output2, respectively.

In order to export these to RDBMS, we first need to create 2 tables in RDBMS which can obtain these outputs.

mysql –u root

mysql> create database IndiaProgress;

mysql> use IndiaProgress;

mysql> create table output1 (district varchar(50));

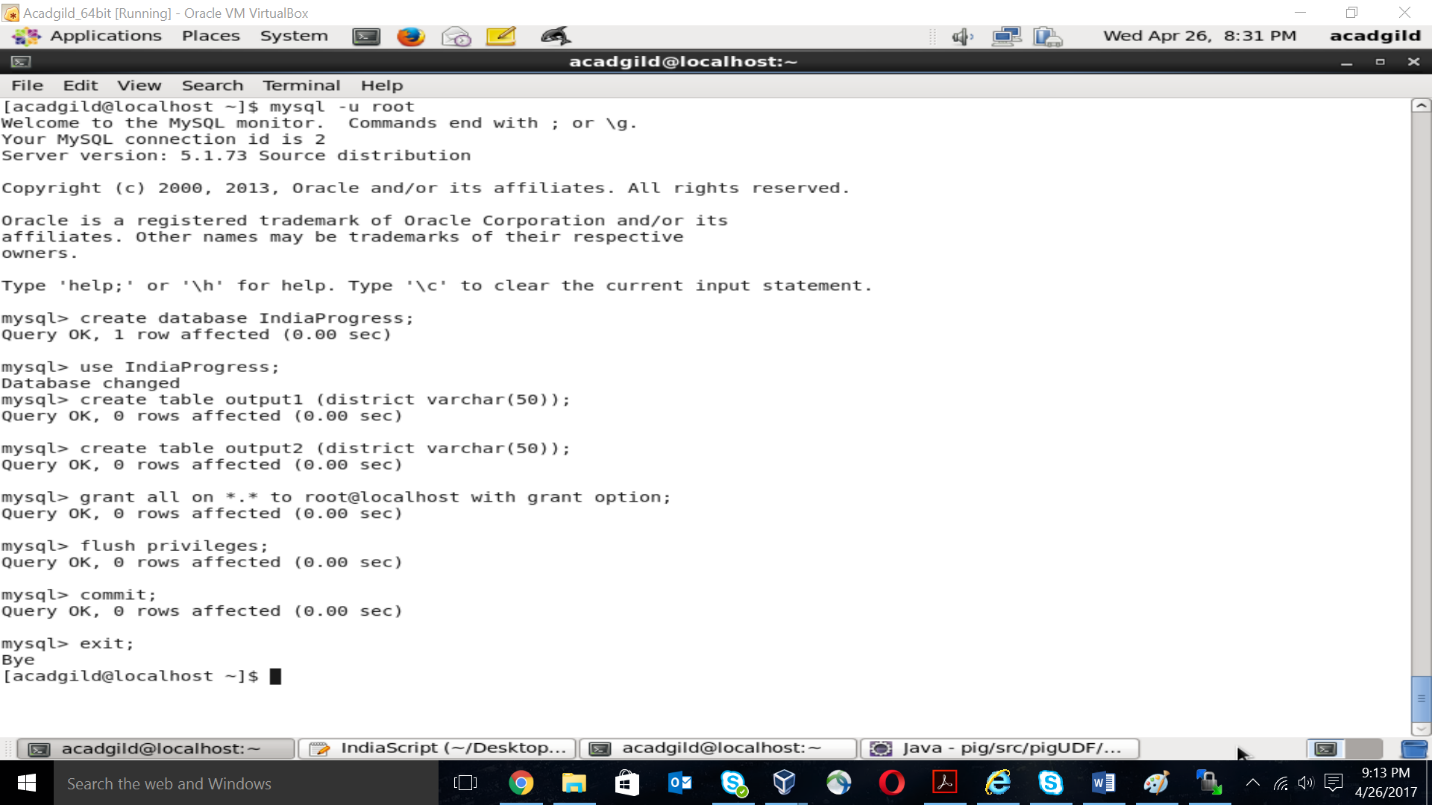
mysql> create table output2 (district varchar(50));

mysql> grant all on \*.\* to root@localhost with grant option;

mysql> flush privileges;

mysql> commit;

mysql> exit;



Now, we will transfer the outputs in HDFS to the above created tables in RDBMS using SQOOP.

sqoop export --connect jdbc:mysql://localhost/IndiaProgress --username 'root' --table 'output1' --export-dir '/output1/' --input-fields-terminated-by '\t' -m 1 --columns district;

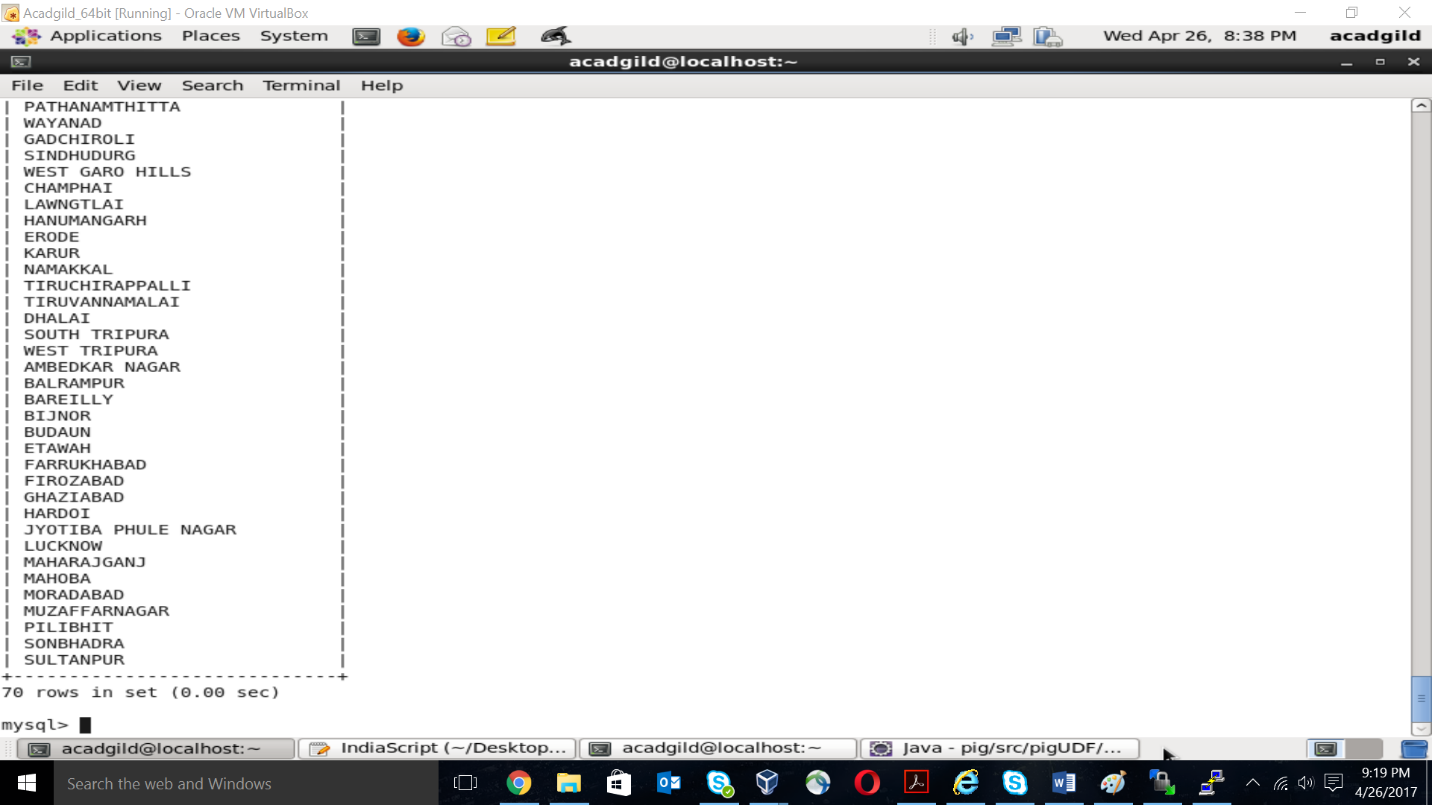
sqoop export --connect jdbc:mysql://localhost/IndiaProgress --username 'root' --table 'output2' --export-dir '/output2/' --input-fields-terminated-by '\t' -m 1 --columns district;

**Verification of output in RDBMS:**

mysql –u root

mysql> use IndiaProgress;

mysql> select \* from output1;



mysql> select \* from output2;

