**1. Executive Summary**

*1.1Overview*

*To develop the System to analyze the log data (In XML format) of government progress there are some various development tools with facilities.*

*1.2 Purpose and Scope*

*To capture the data for analyzing the progress of various activities is the purpose of this project and they are viewed in two manner they are:*

***In scope:***

*The main requirements are explained in phase 1 of the Project:*

* *Analyze the data and understand the progress*
* *Store the results in Hbase/RDBMS*
* *Developing system: This is used to handle the incoming log feed and store the information in Hadoop*

*Cluster.*

***Out of scope:***

*We can use this data and also used for visualization and get more insights*

**2. Product/Service Description**

*2.1* ***Assumptions:***

*Log will be created in XML format and stored in the server*

* 1. ***Constraints:***

*It describes any item that will constrain the design options*

* *This system may not be used for searching currently. But it will be used for analysis and saving the relevant information.*
* *System will be using Hbase as a database*

**3. Requirements:**

* *Pig,MapReduce job for parsing the XML data.*
* *Create the Sqoop job to store the data in database*
* *Create Pig scripts,MapReduce jobs to analyze the data*
* *The jobs in flume which will format the data and place the data to HDFS*

***Priority Definitions:***

*The following definitions are intended as a guideline.*

* *Priority 1 – Create FLUME job for fetching log files from spool directory the data*
* *Priority 2 – MapReduce/pig job to preprocess*

***Download the dataset using the below link****:*

*https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing*

*Refer the below steps for better understanding to create the project.*

**Step 1:**

*Copy dataset* ***from local file system to HDFS using flume.***

***Command****:*

*flume-agent agent –n agent1 –c conf –f <path to filecopy.conf>*

**Step 2:**

*Input file is in the XML format to parse the data and get the results by using Map reduce or pig.*

***4. Problem statement***

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

*Export the results to mysql using sqoop*

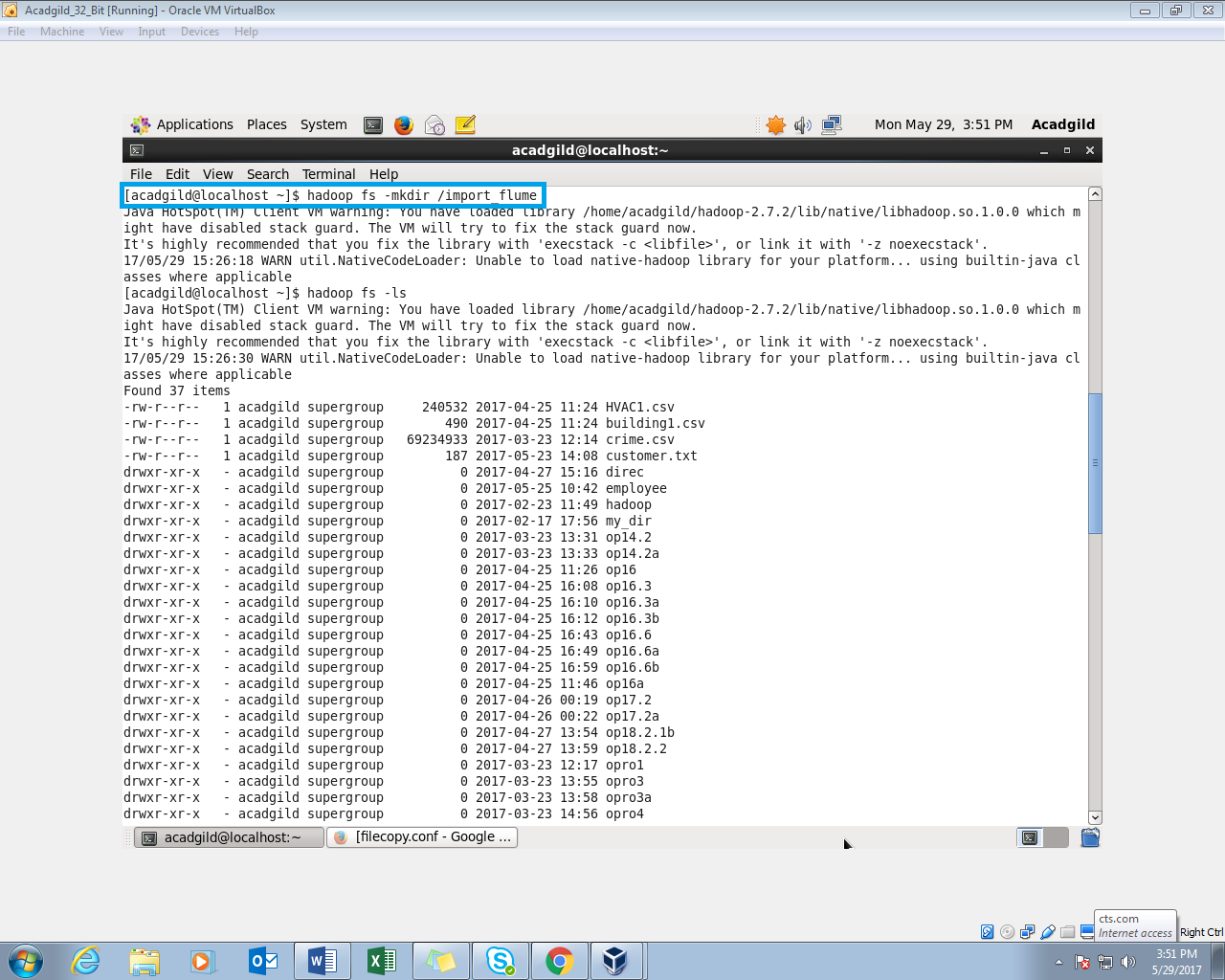
*2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.*

*Export the results to MySQL using Sqoop*

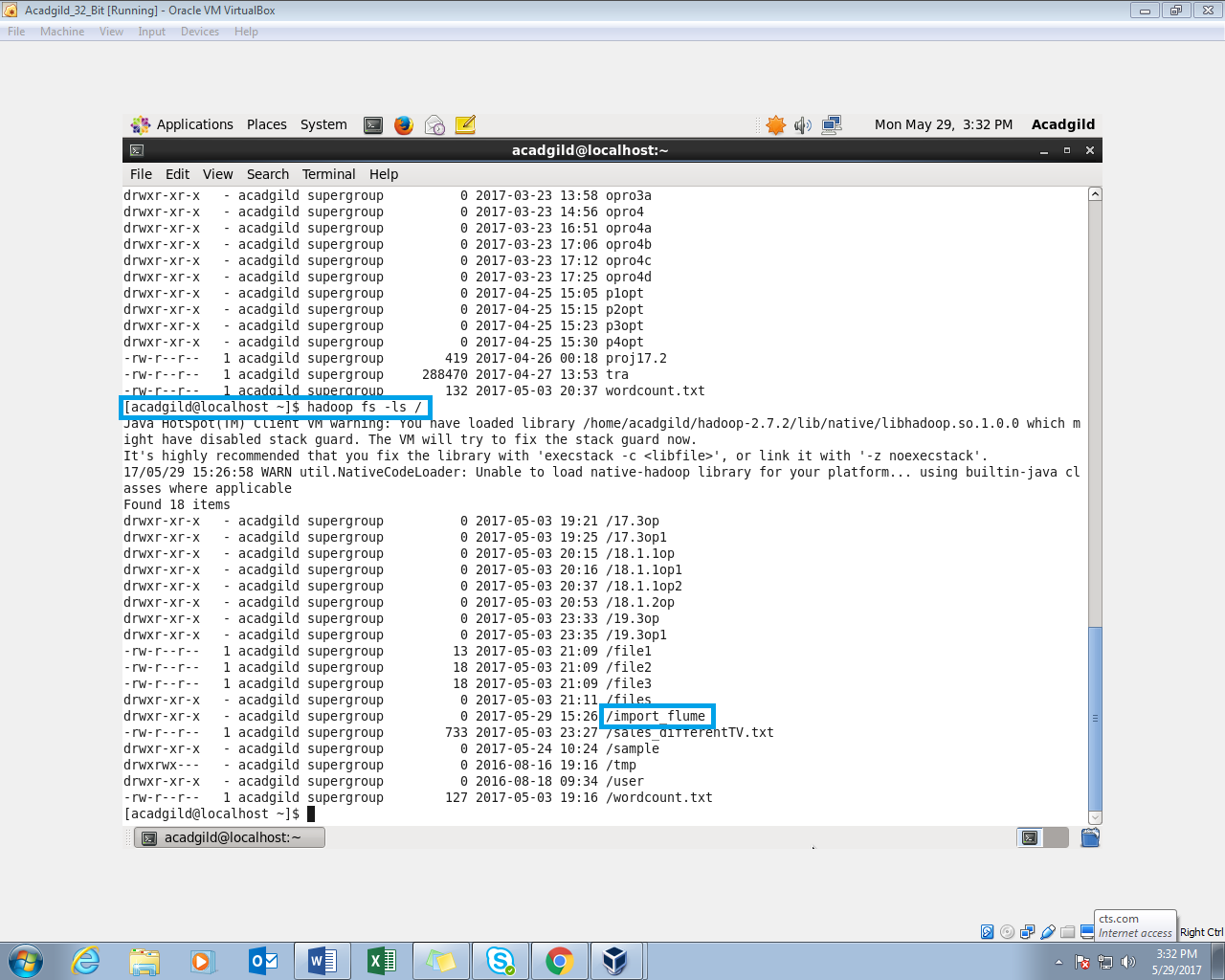
**SOLUTION :**

*Create a directory named ‘import\_flume’ using mkdir command:*

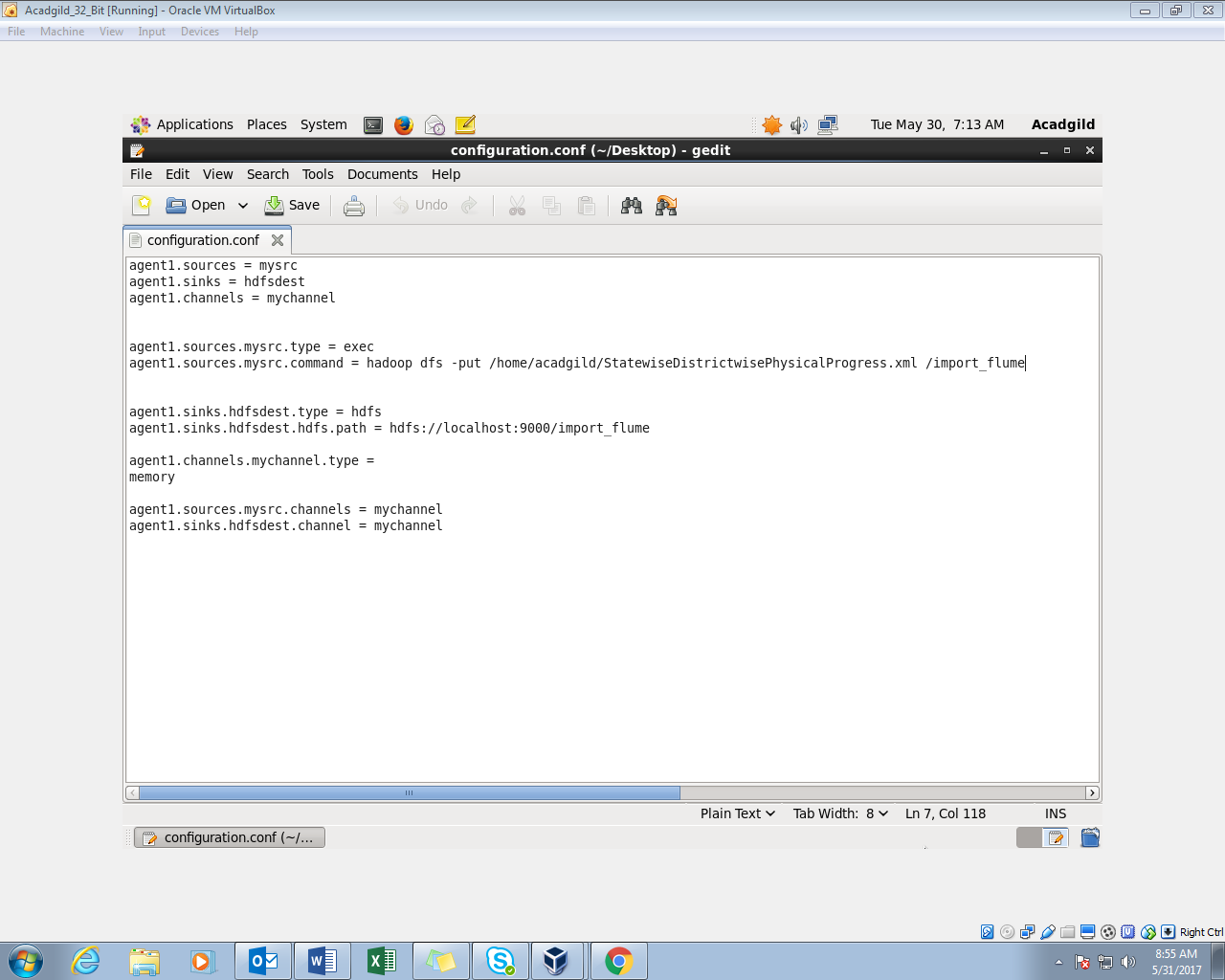
*The dataset will be copied to this directory*.

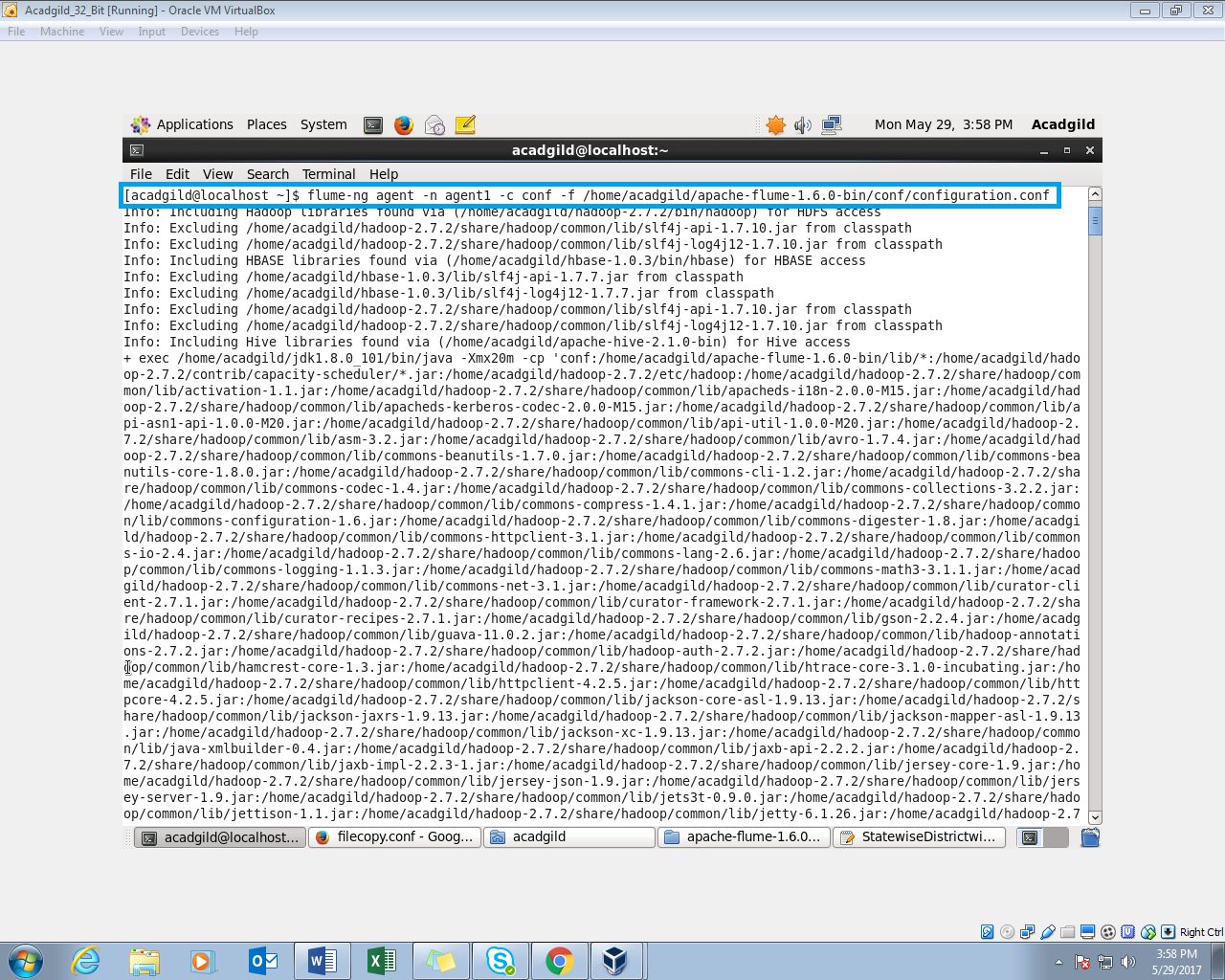


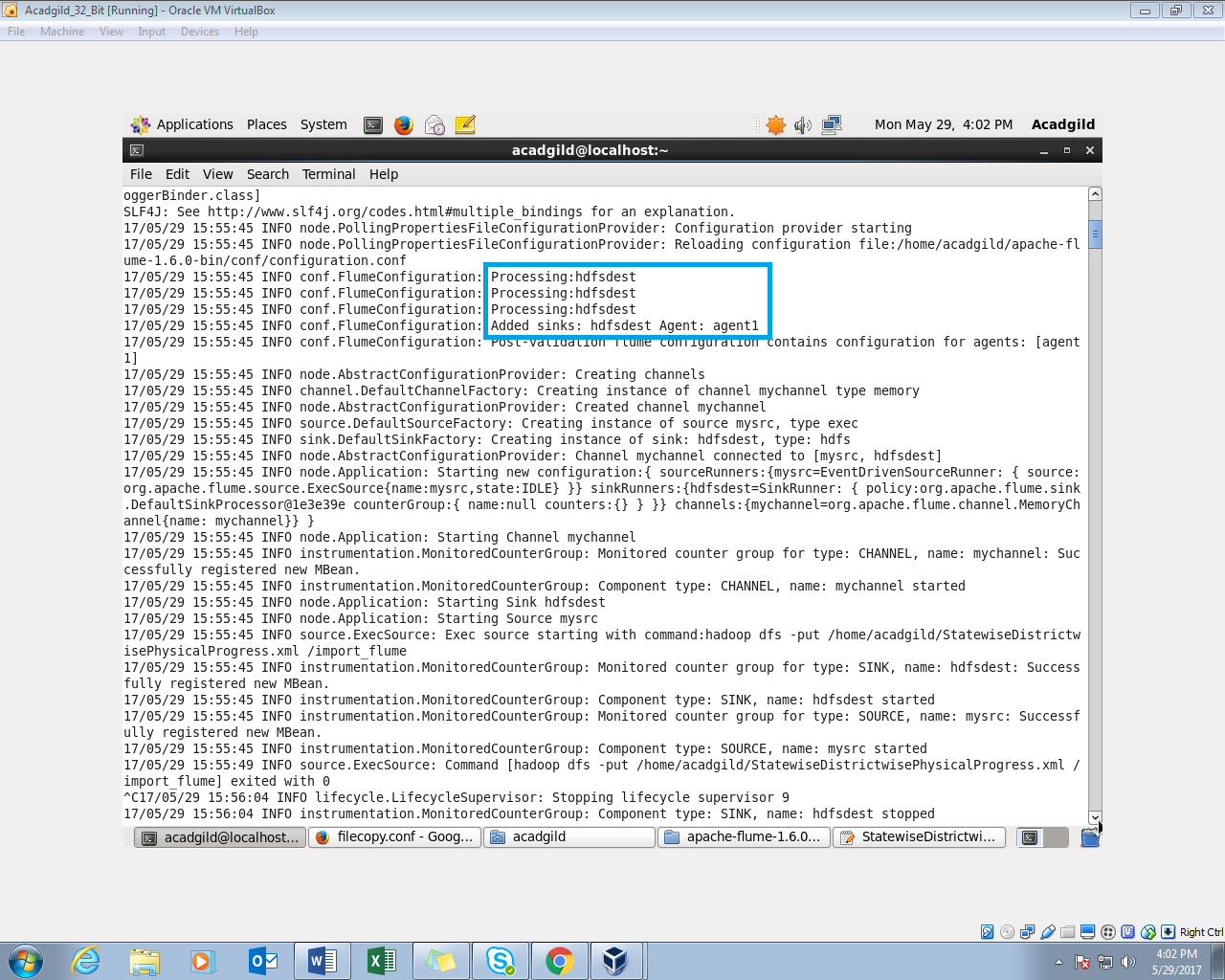
*The Directory (import\_flume) is created in the HDFS:*



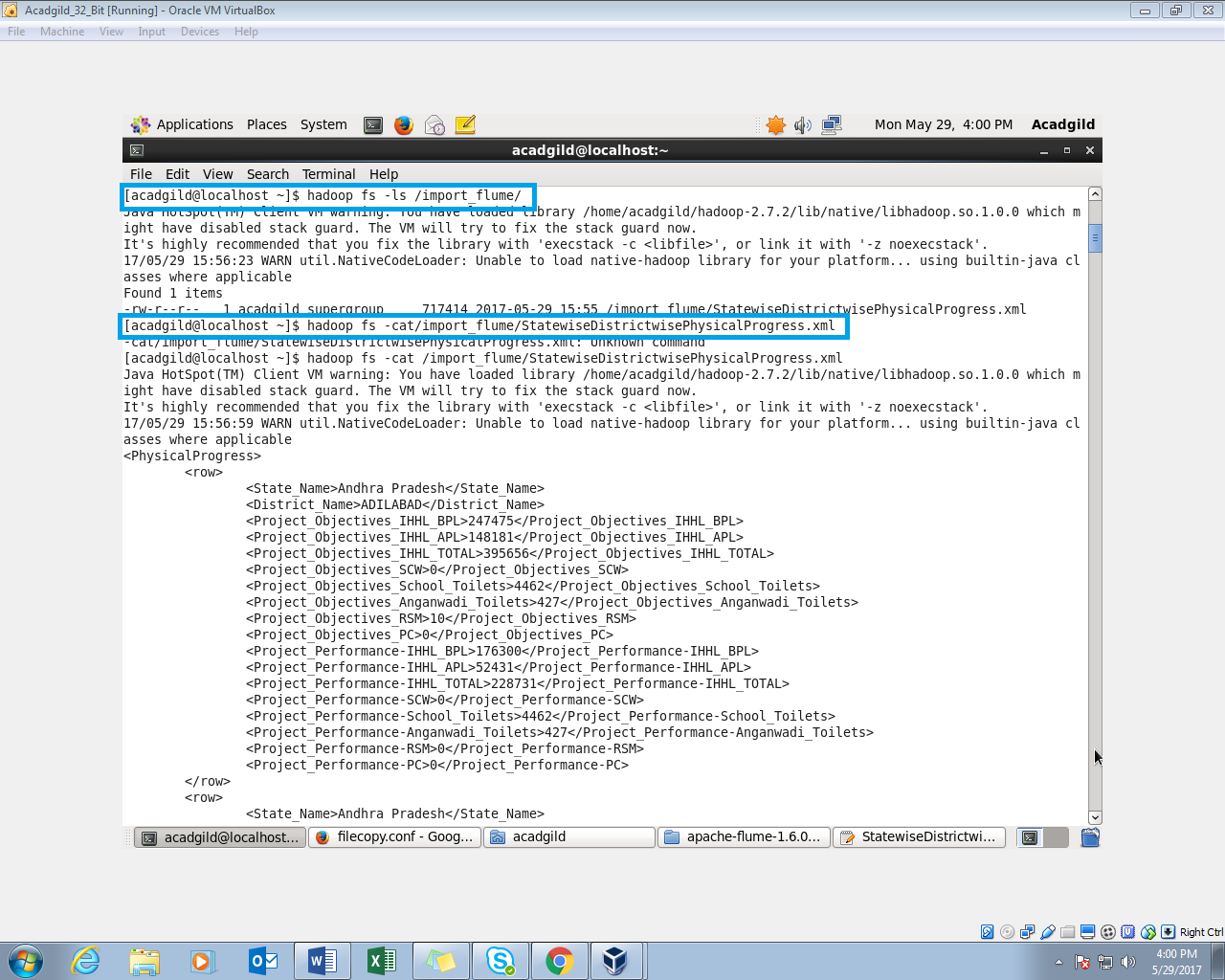
**Step1** : *Create a ‘configuration.conf‘ file in the conf folder of the flume*:



**Flume Command:** 

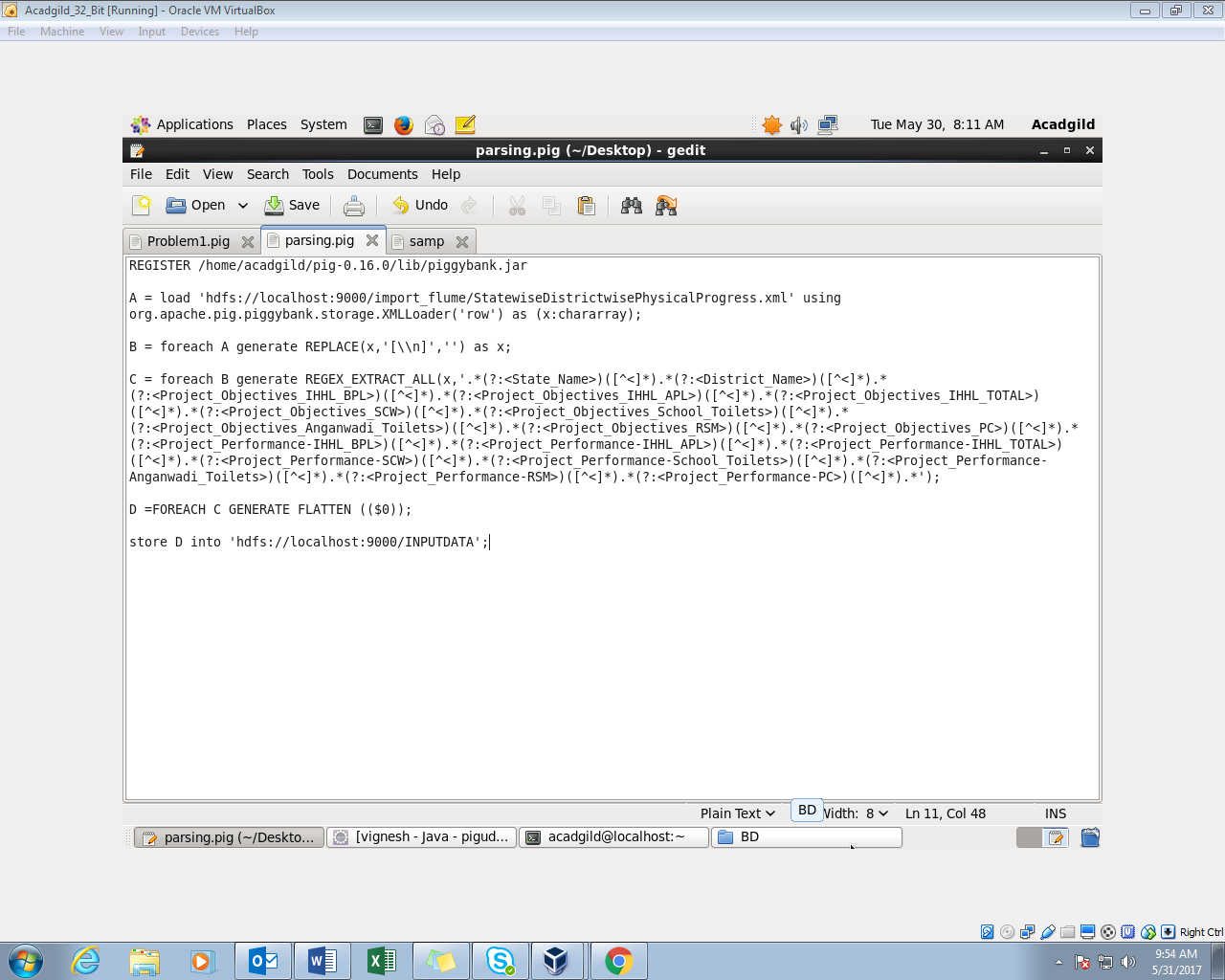
*the name of the agent is taken as agent1* 

*The import flume directory is created and the dataset is copied into it:*



**Step2 :** *Parsing the input XML file:*

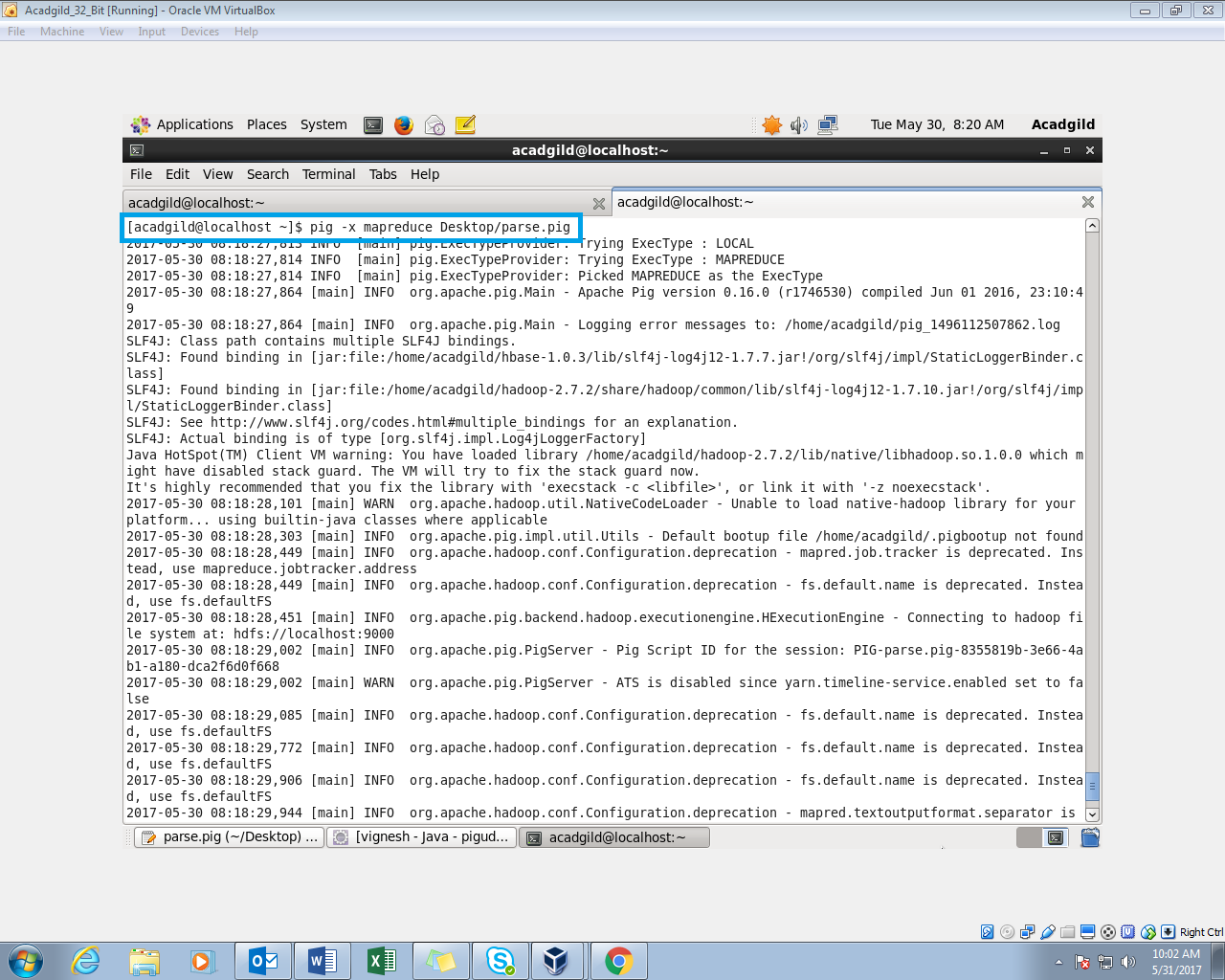
*Now the input data file is in xml file is converted into csv file. We used a pig script to complete this process.*



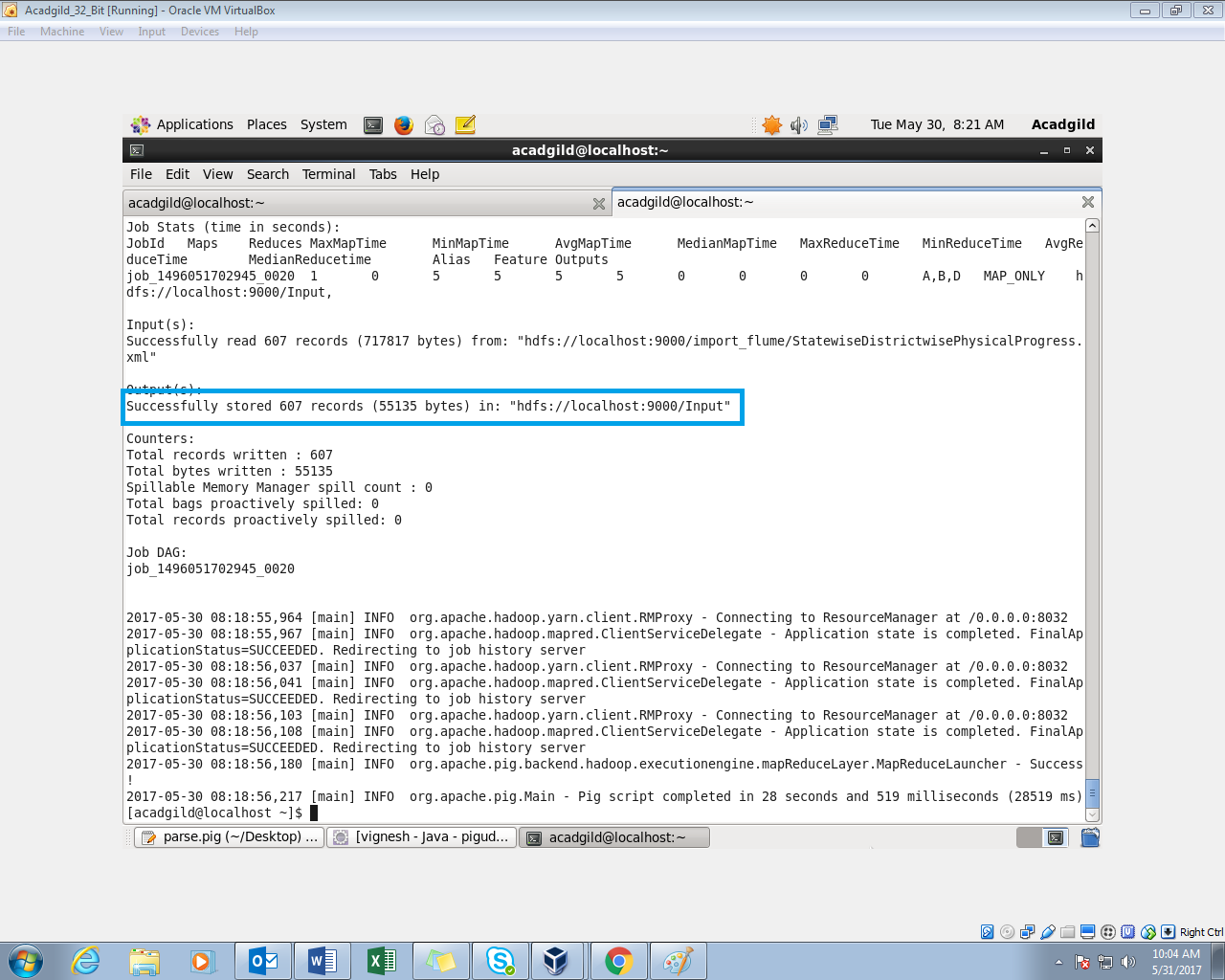
* *Register the piggybank jar.*
* *Load the dataset from the hdfs using the XMLLoader using chararray.*
* *Use regex\_extract\_all function to capture values within tag.*
* *The Output will be in a bag and it is un-nested using flatten command.*
* *Store the output into hdfs for further process.*

*The above pig script is carried out in the mapreduce mode. The Output of the data is stored in ‘Input’ folder.*

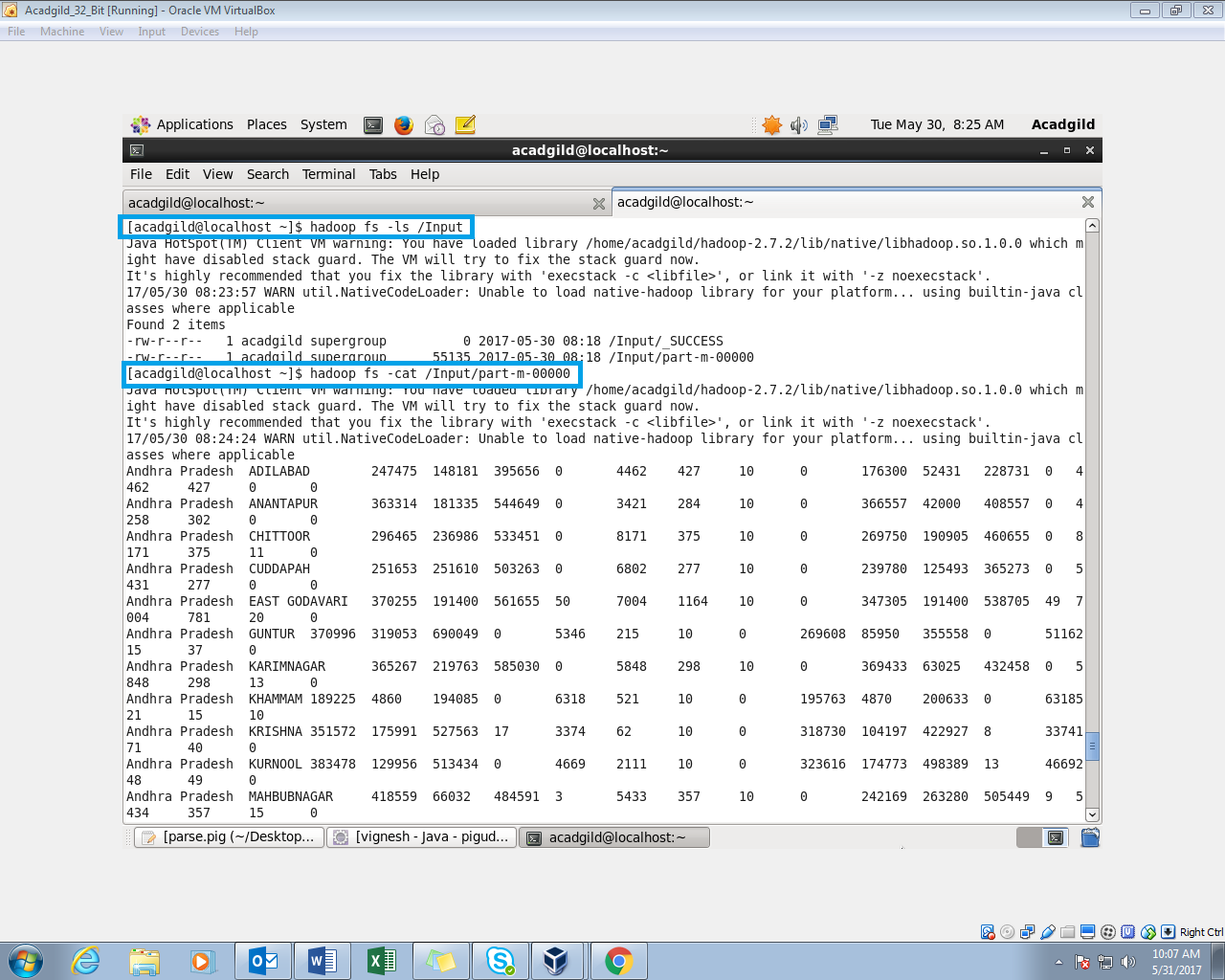
***Executing the PigScript in mapreduce mode****:*



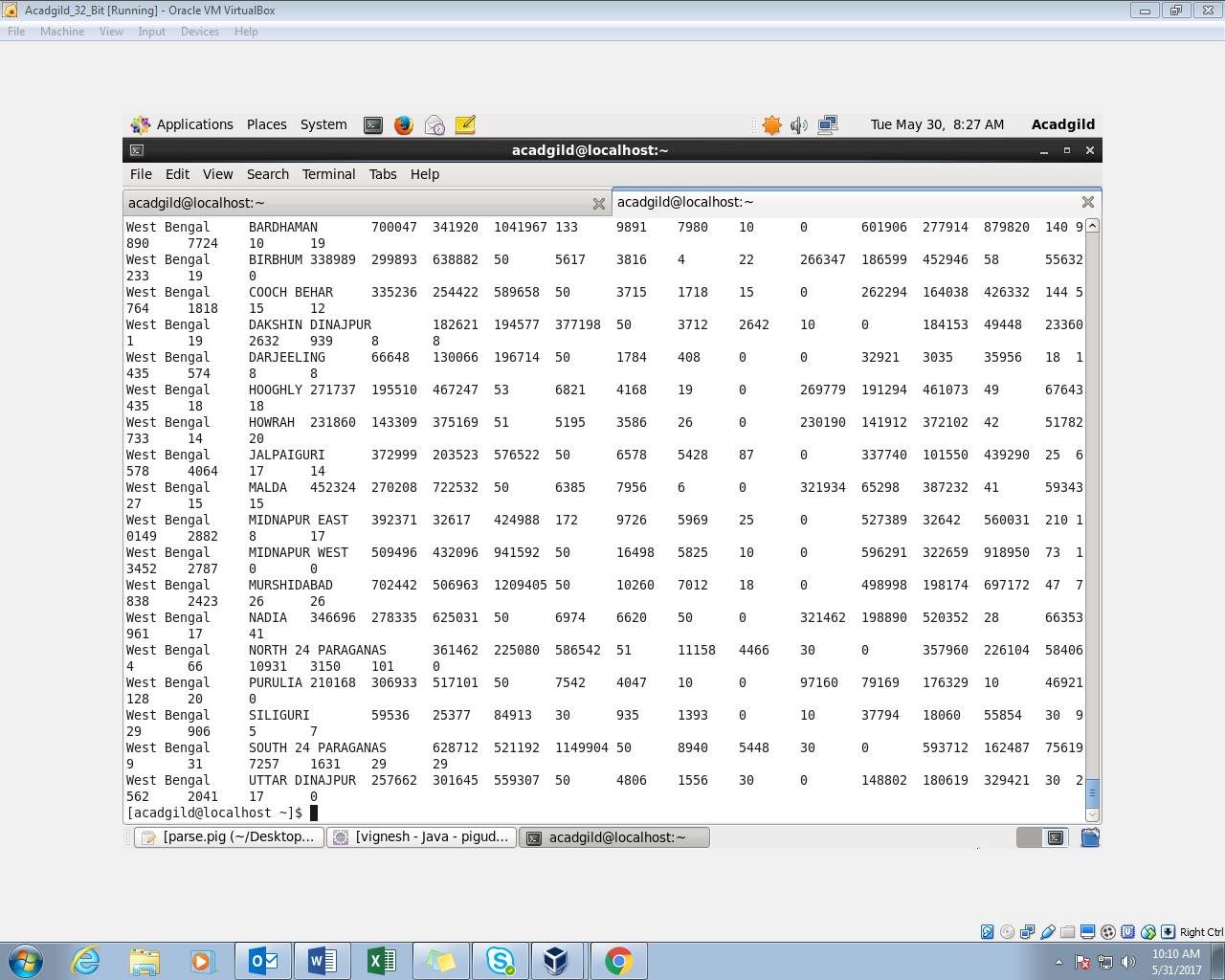
*Output is stored in the ‘Input’ folder*:



*Display the parsed input data file*:



*Sample data*:



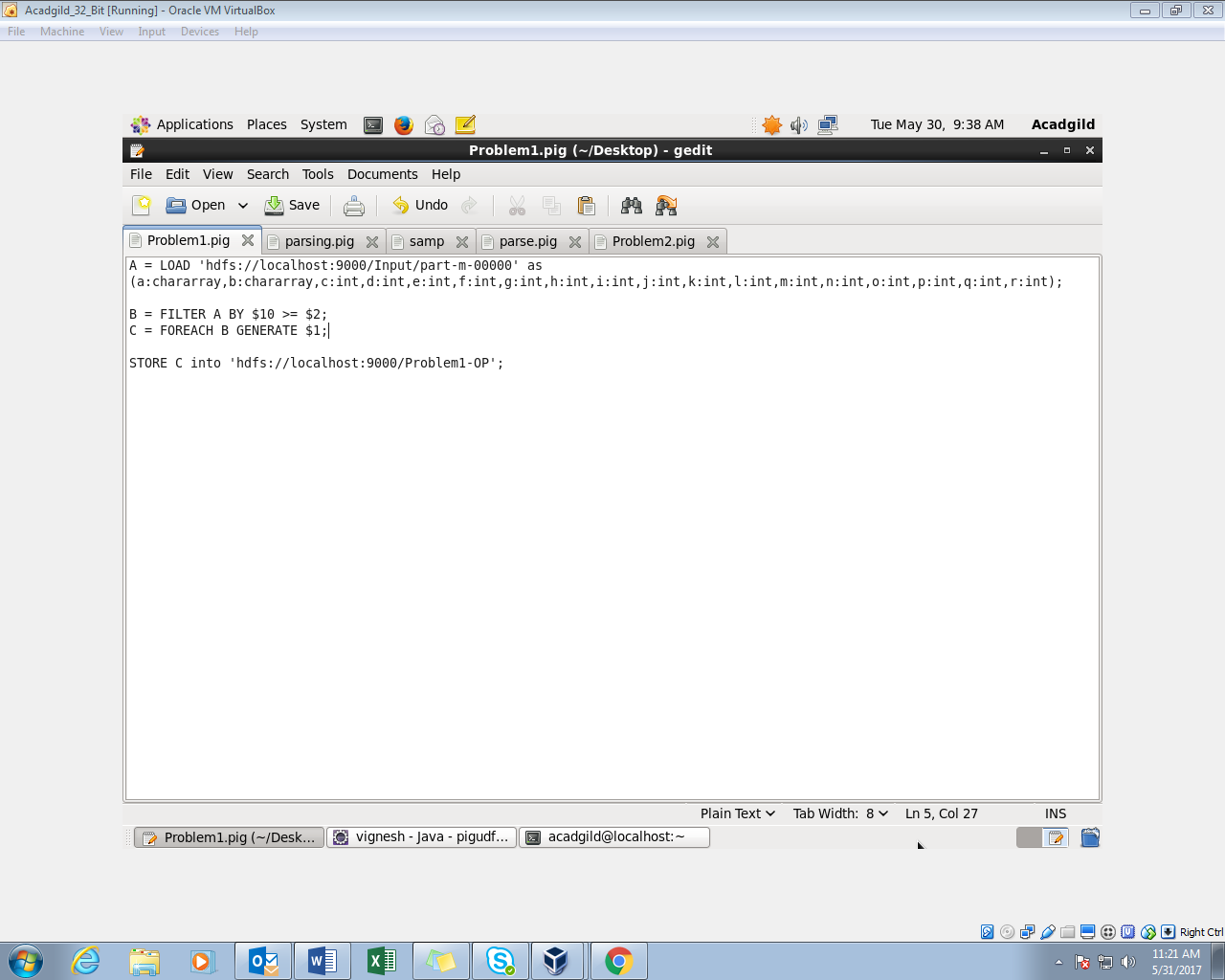
**Problem statement**

**Question 1:**

*Find out the districts who achieved 100 percent objective in BPL cards*

*Export the results to MySql using sqoop*

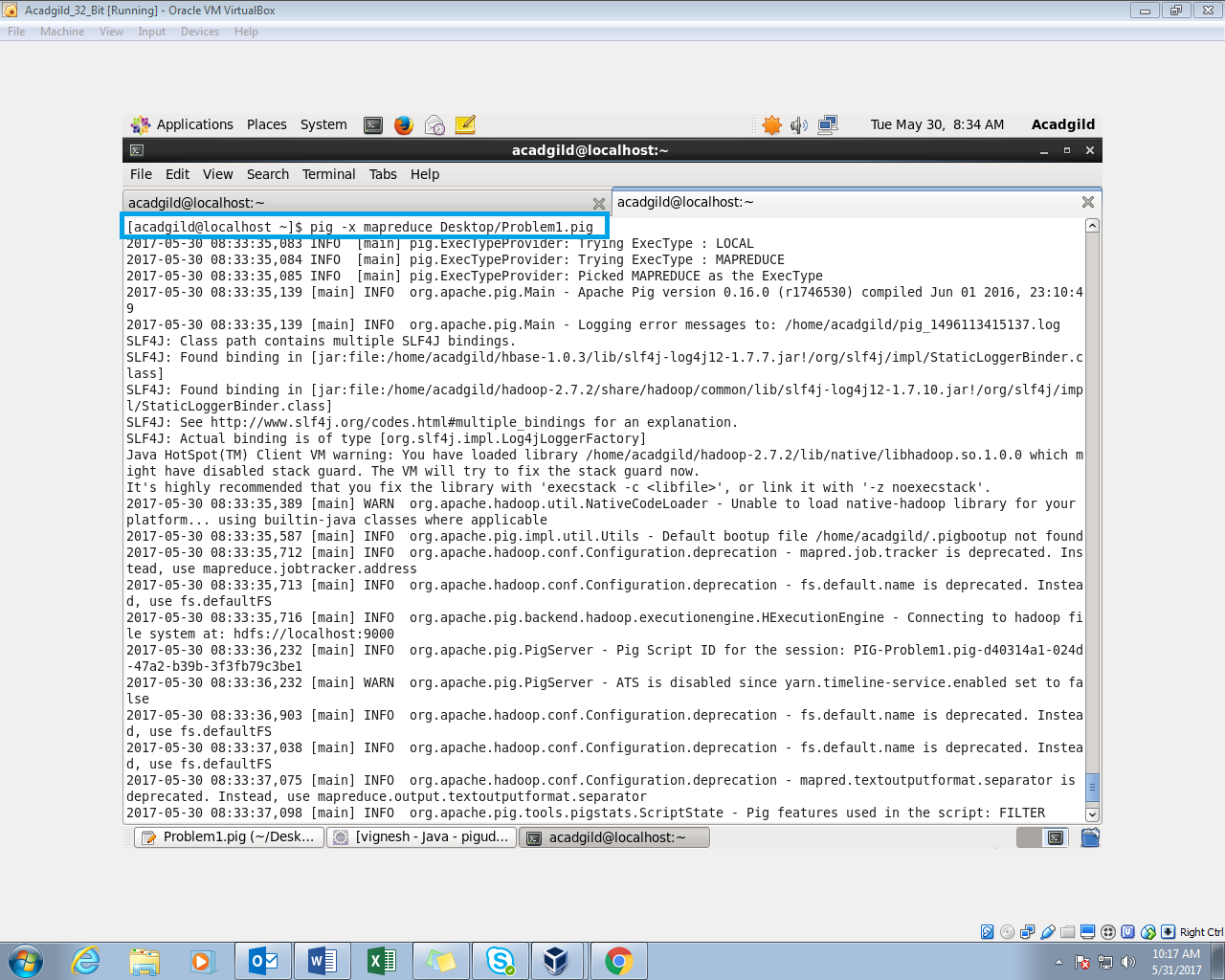
*Pig Script:*



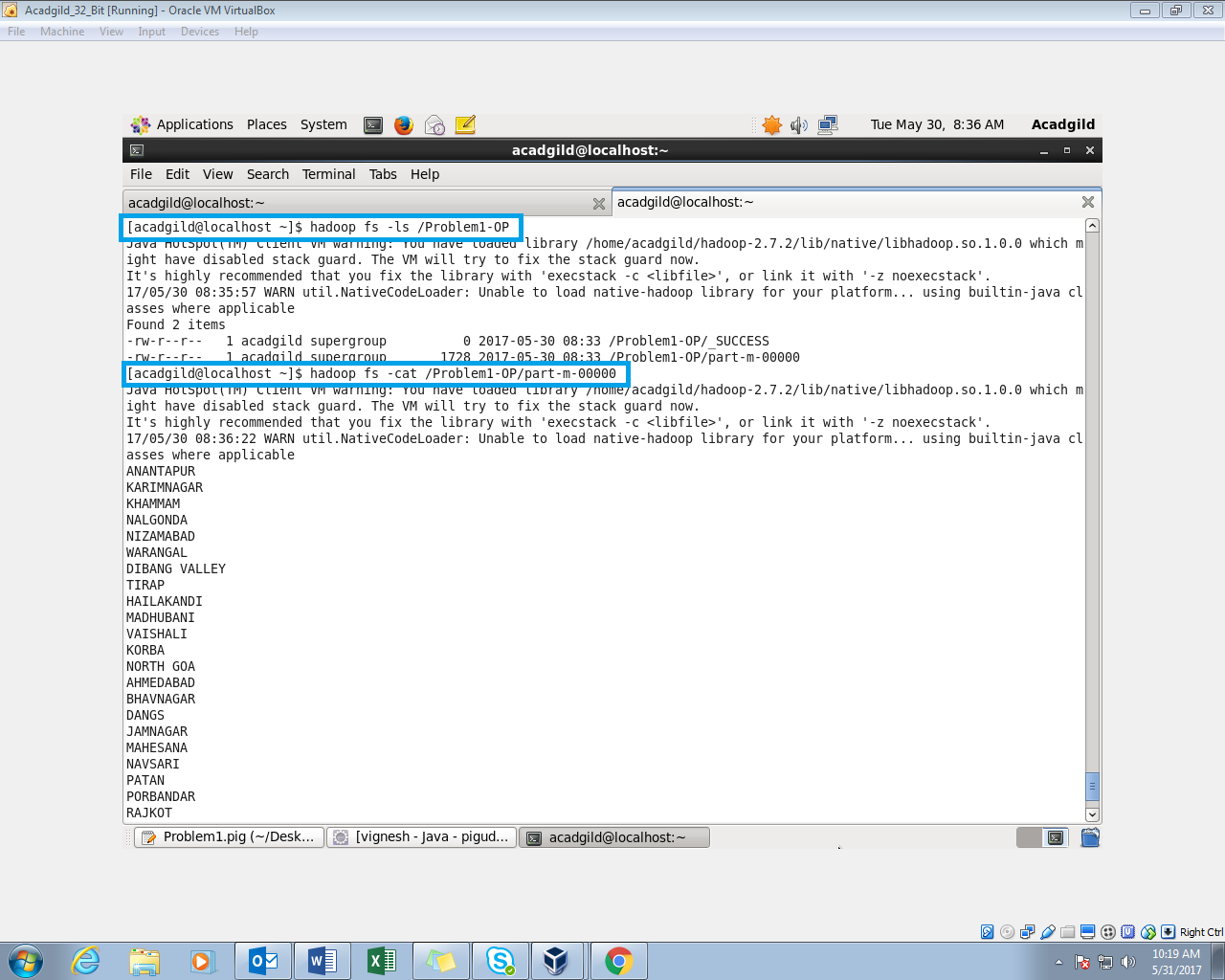
* *Load the parsed data.*
* *We need the districts which achieved 100 percent objectives. 11th column is project performance and the 3rd column is project objective BPL. If the performance is higher than the objective, it is considered to be achieved 100 percent. We are using the filter command to sort out.*
* *As of now we need the districts so we are displaying the districts only by using the foreach generate command.*
* *Storing the output in the hdfs.*

*The name of the script is Problem1.pig*.

*proceeding the Pig Script in MapReduce* :



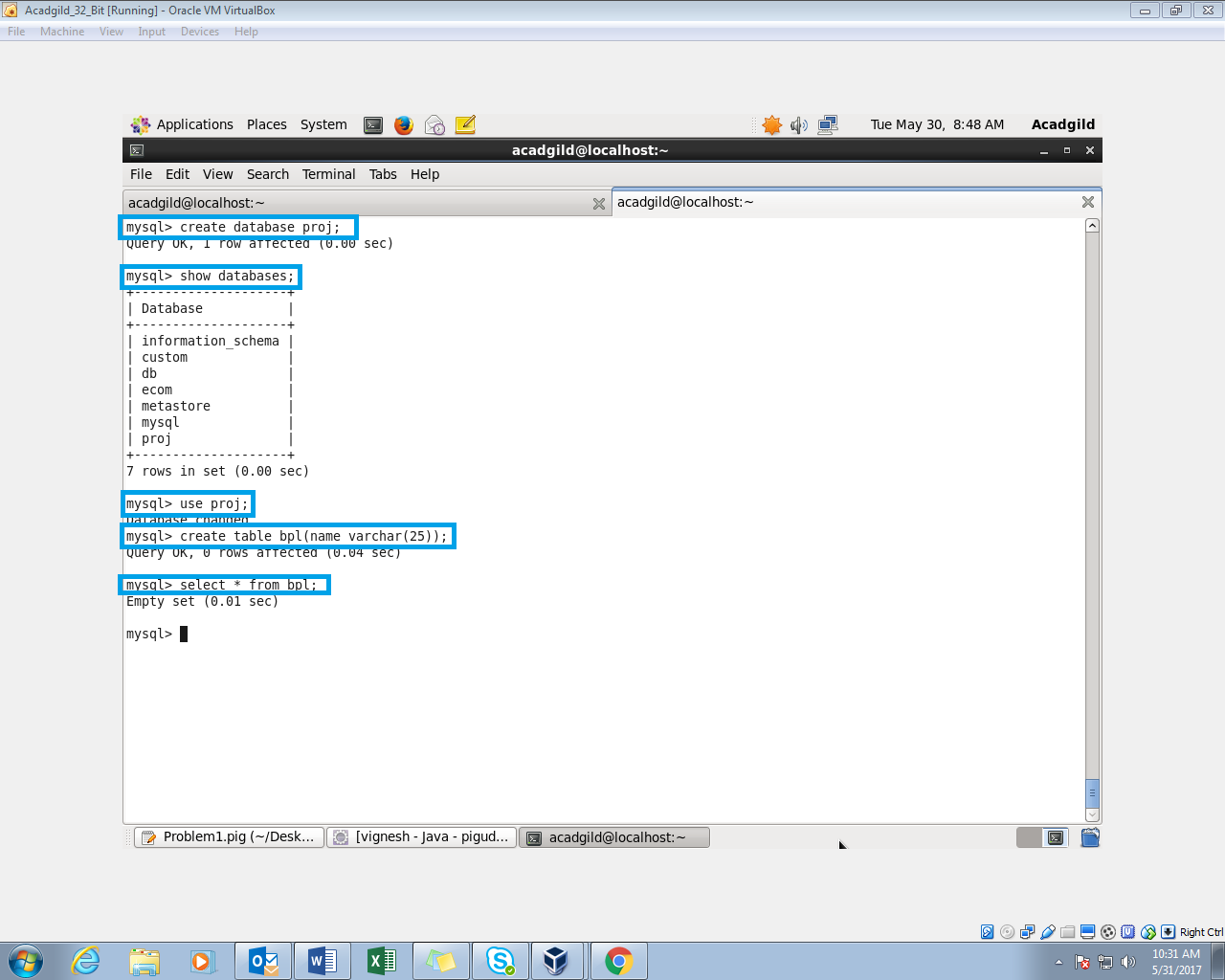
*Output stored in the hdfs*:



*Sample Output*:

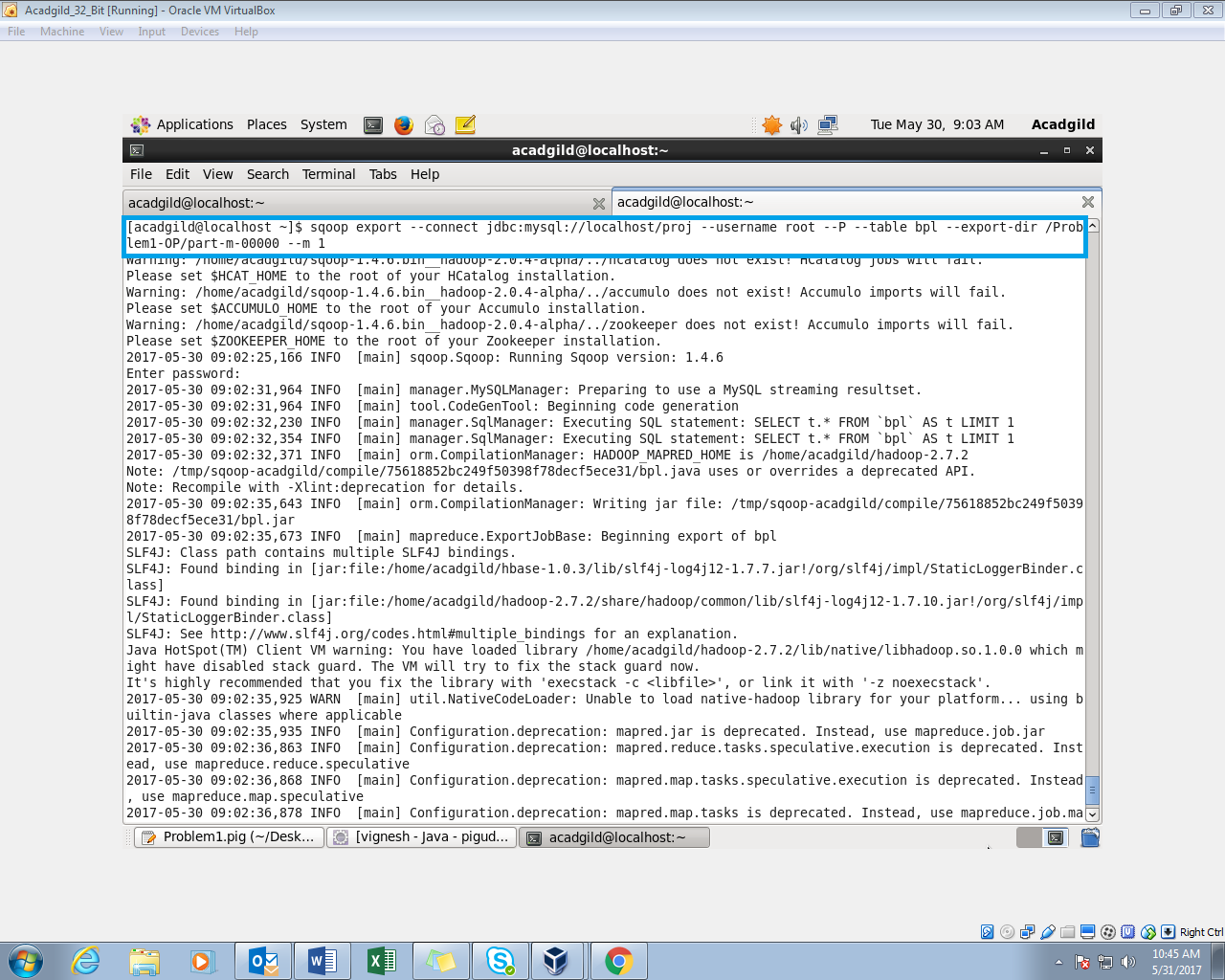


*Exporting the Results to MySql*:



*MySql terminal is opened and we are creating a database and we use it in that we are created the table and we kept it empty before exporting.*

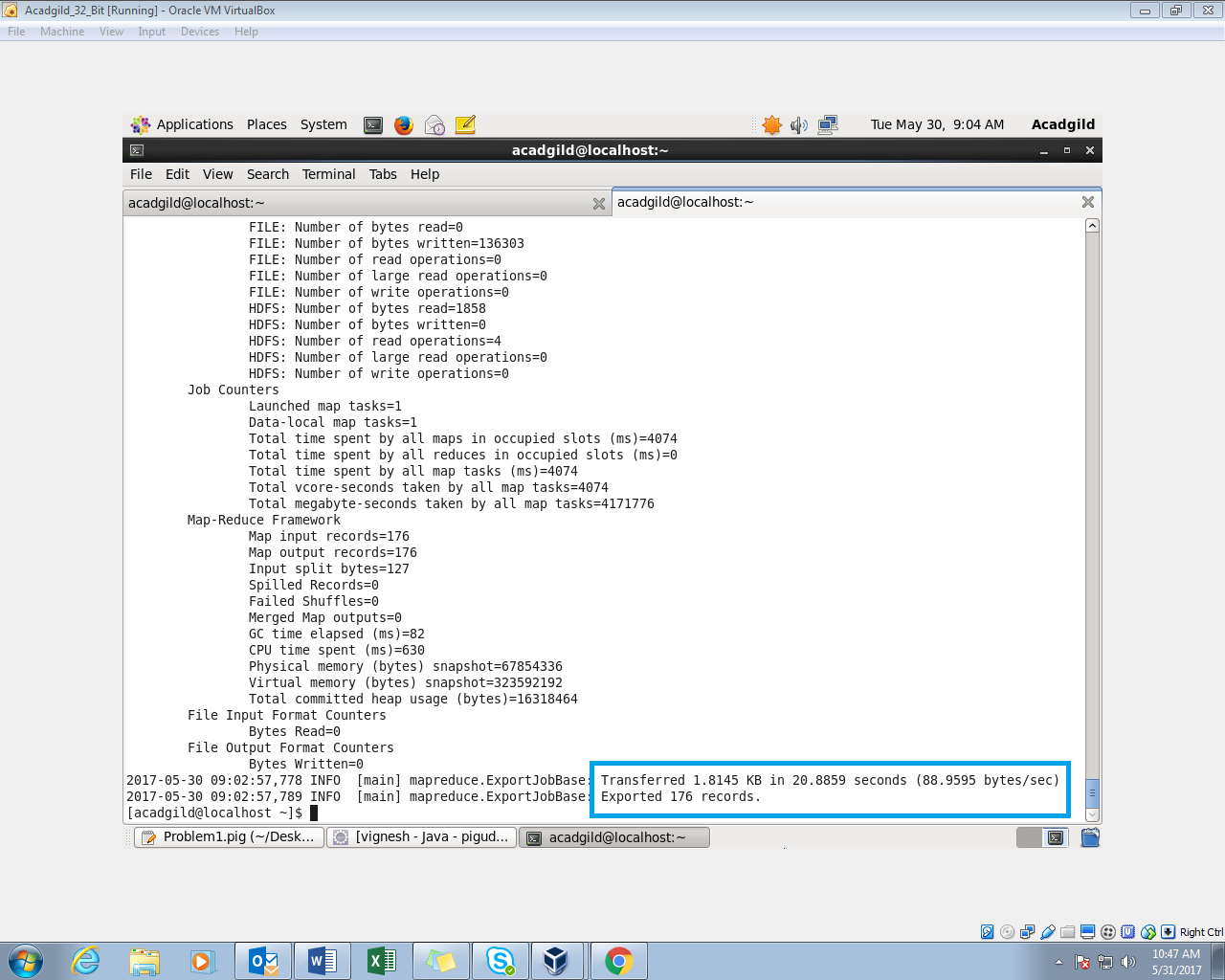
***Sqoop command to export from hdfs to Mysql***:



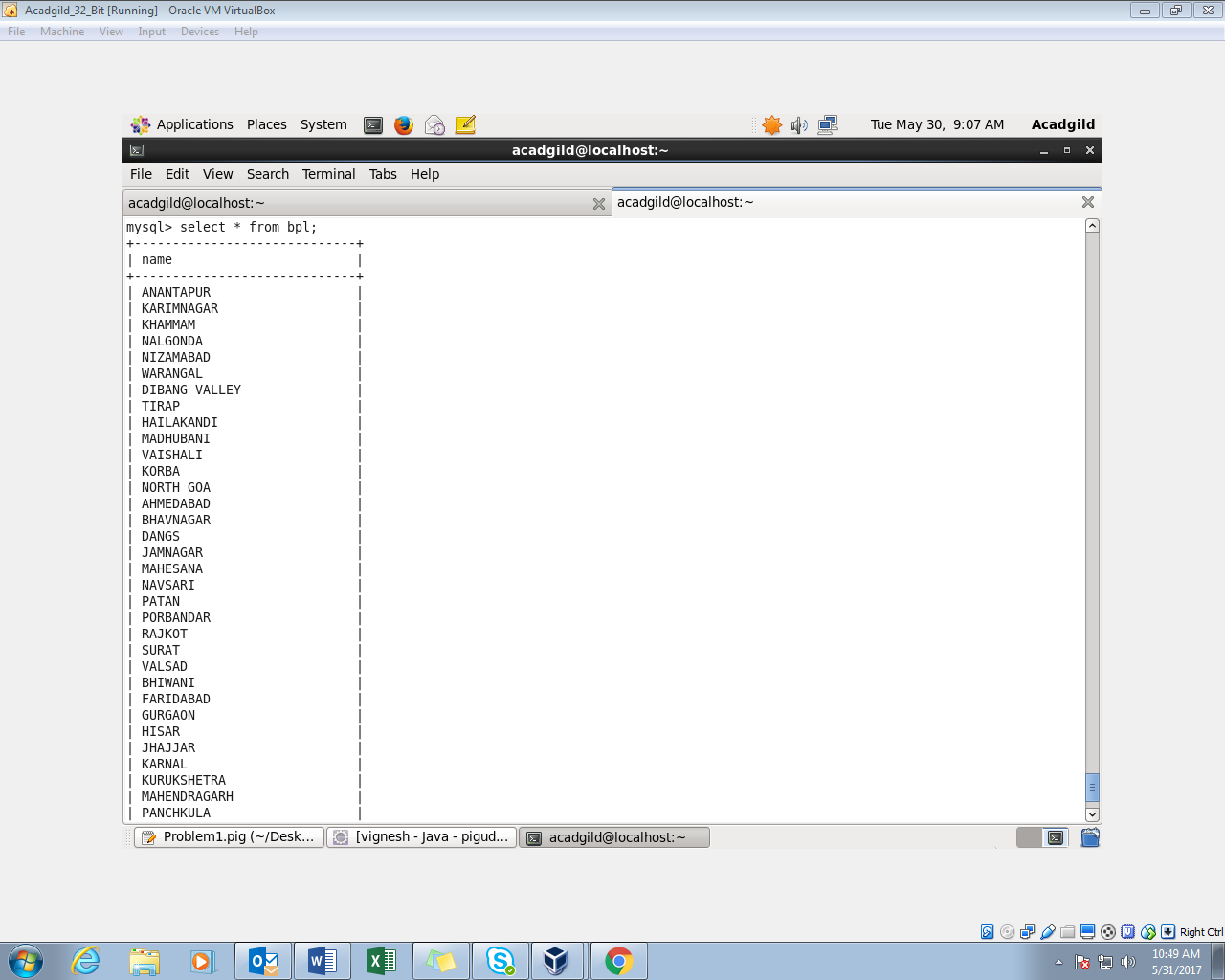
*In this Command, we need to give the mysql database in which we created the table. The username and password (acadgild) to enter into the mysql*

*The hdfs output directory path and table name also given in the command. When the command is executed the sqoop will export the results from hdfs to the table in mysql.*

*It transferred the records to the mysql*:



*The contents of the table*:



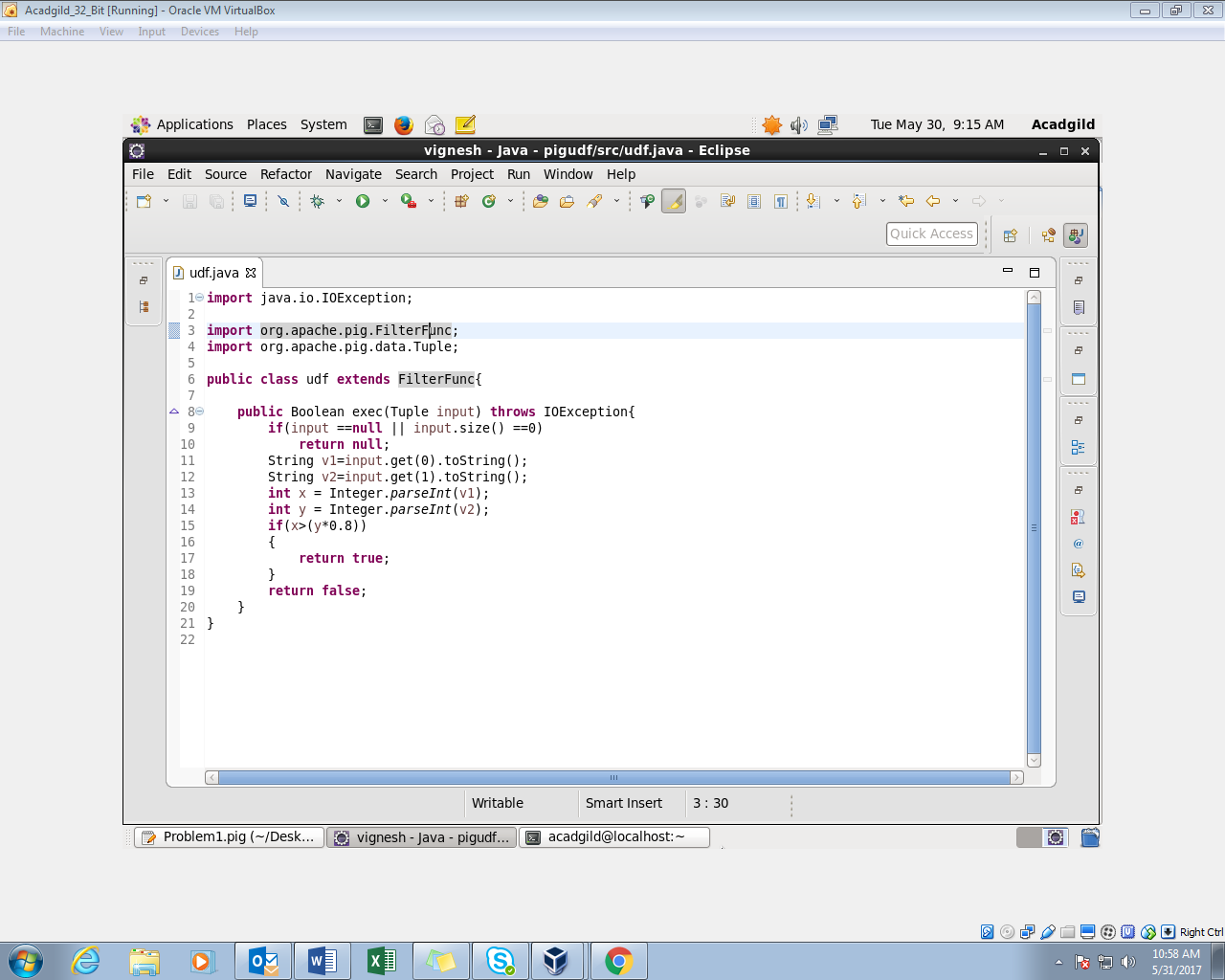
**Question-2:**

*Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.*

*Export the results to MySQL using Sqoop*

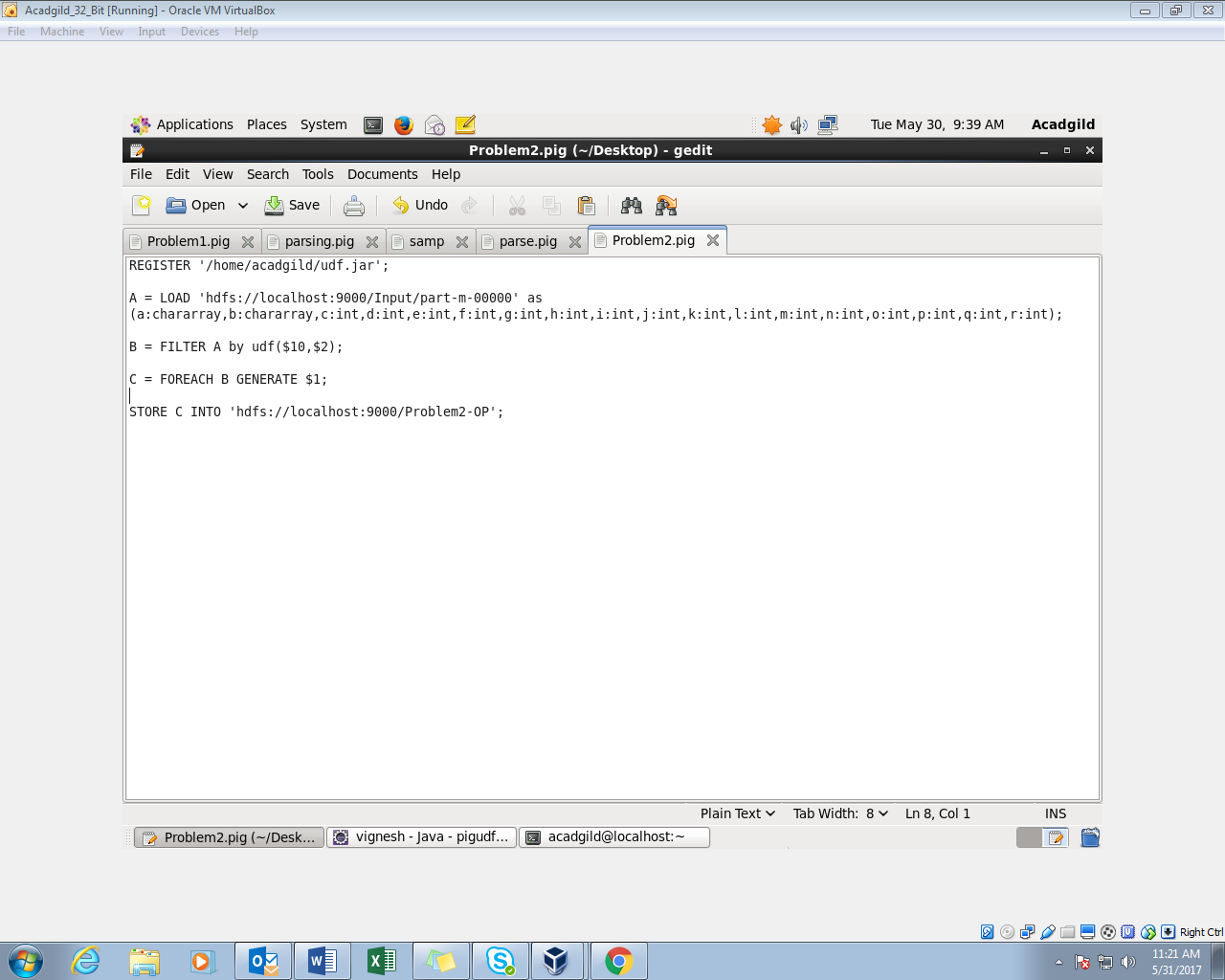
*Creating a UDF to filter the districts:*

*Program*:



* + *we are filtering the districts which have reached 80% of objectives.*
  + *The java class extends the FilterFunc. The return type of the class should be in Boolean.*
  + *If the project performance is higher than 80% of the project objectives, then Boolean value of true is returned.*
  + *We export the program into a jar file. This jar file will be registered in the pigscript. Then we are using the udf to filter*

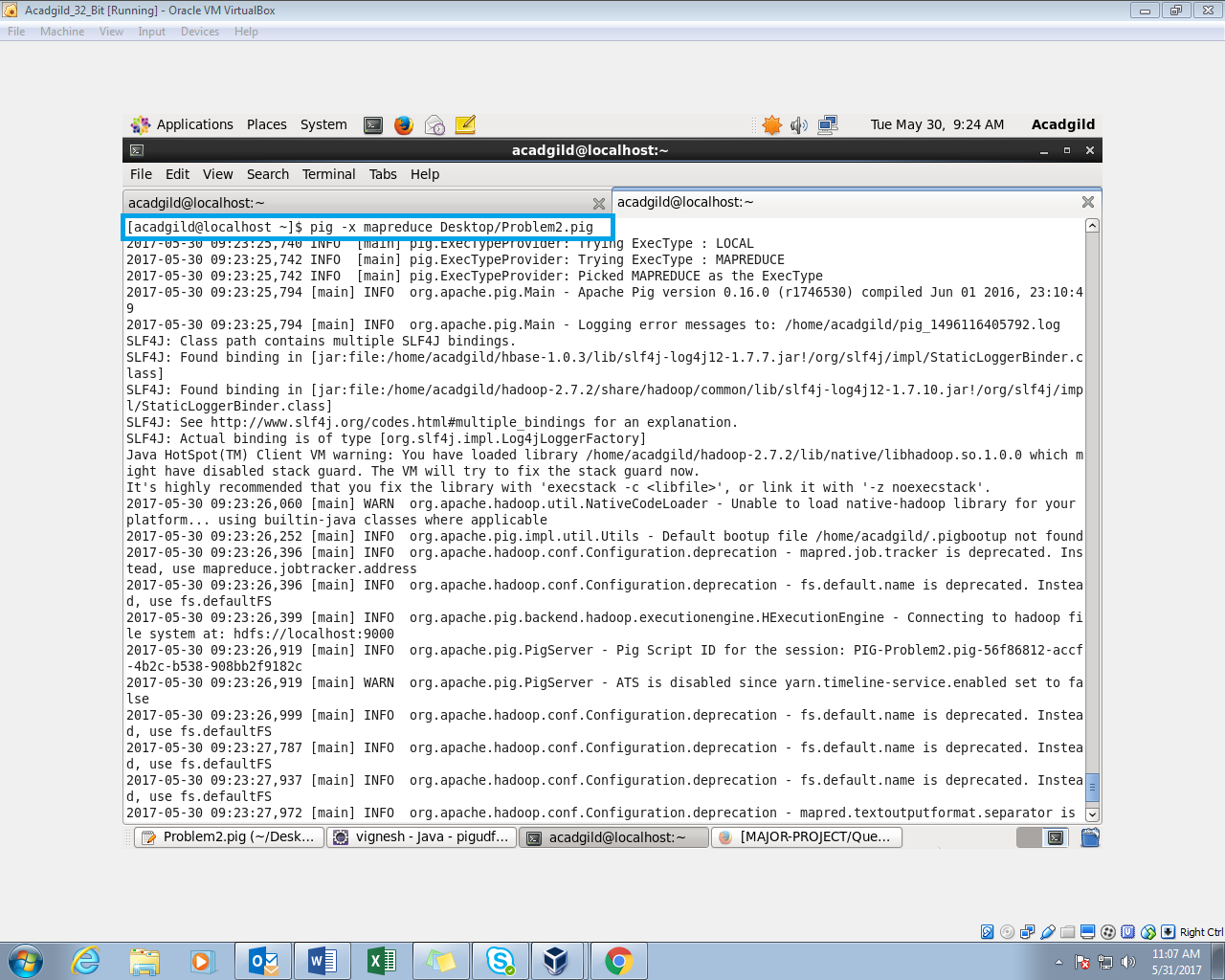
Pig Script:



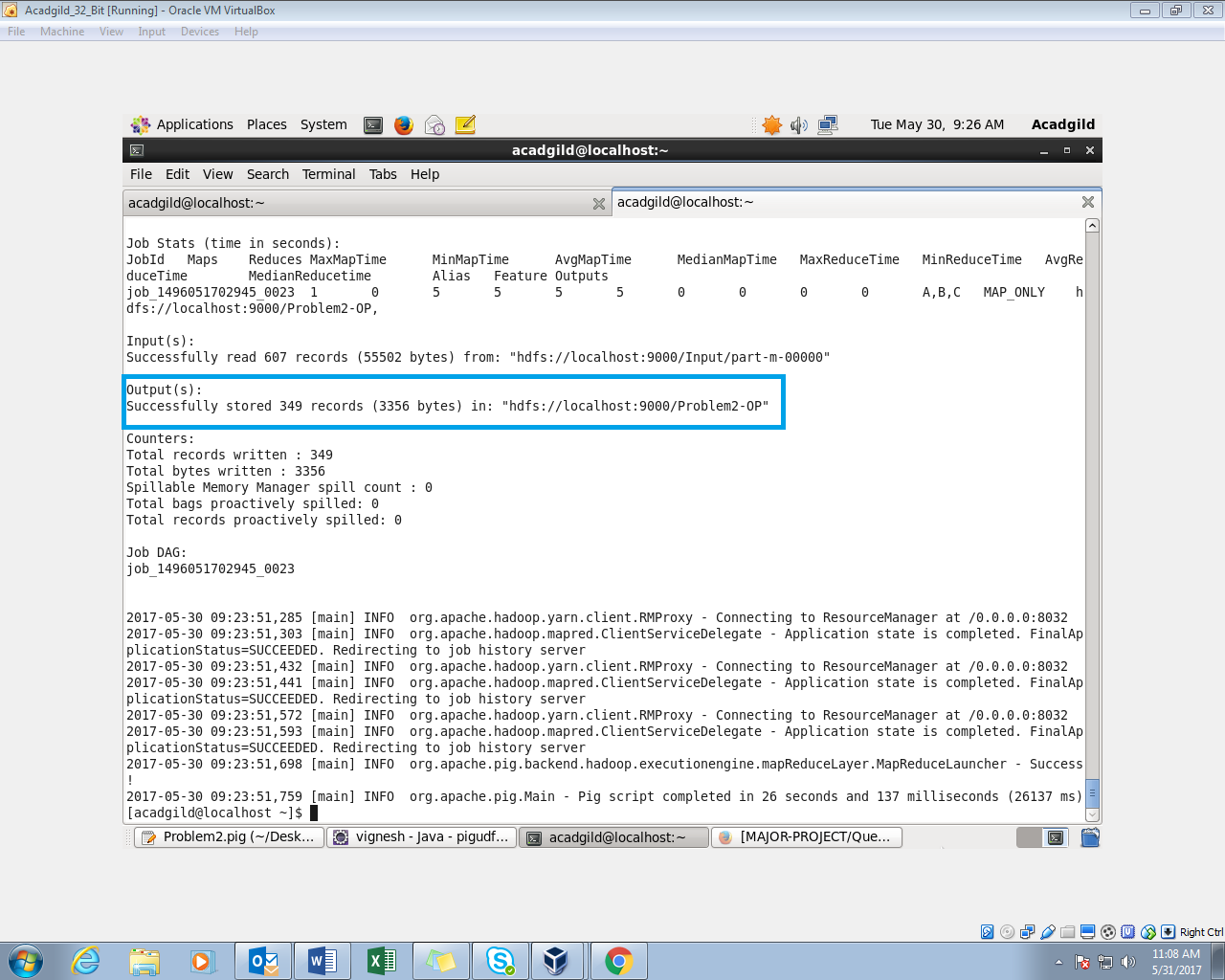
* *The User defined java program is exported to jar file named udf, and it is registered.*
* *Load the input data from hdfs.*
* *Filter the loaded input by the user defined function using udf for filtering process*
* *we are taking the districts name using foreach generate.*
* *Store the output in the hdfs.*

*.*

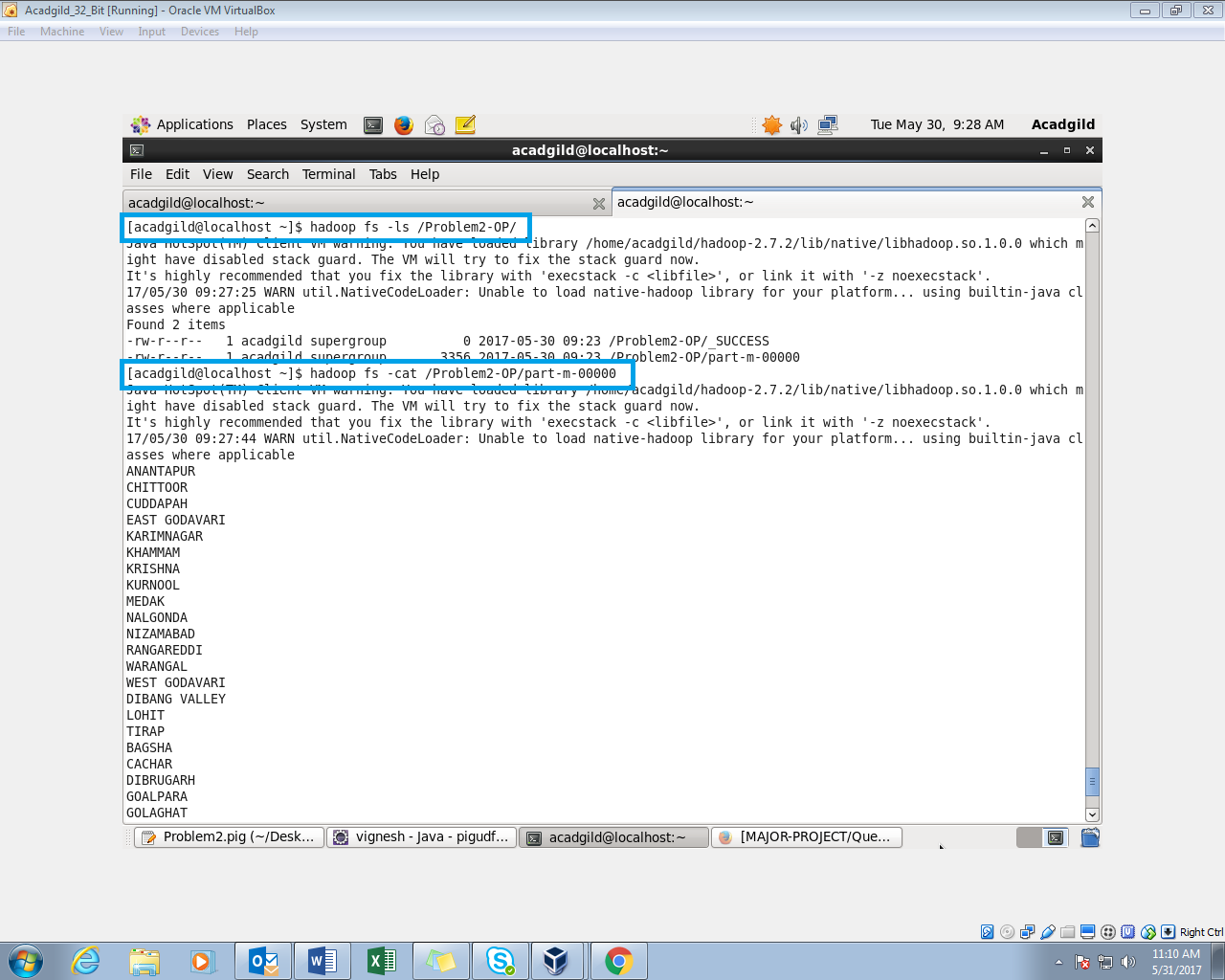
*proceed the Pig Script in mapreduce mode*:



*we stored the output in hdfs*:



*Check the HDFS directory*:

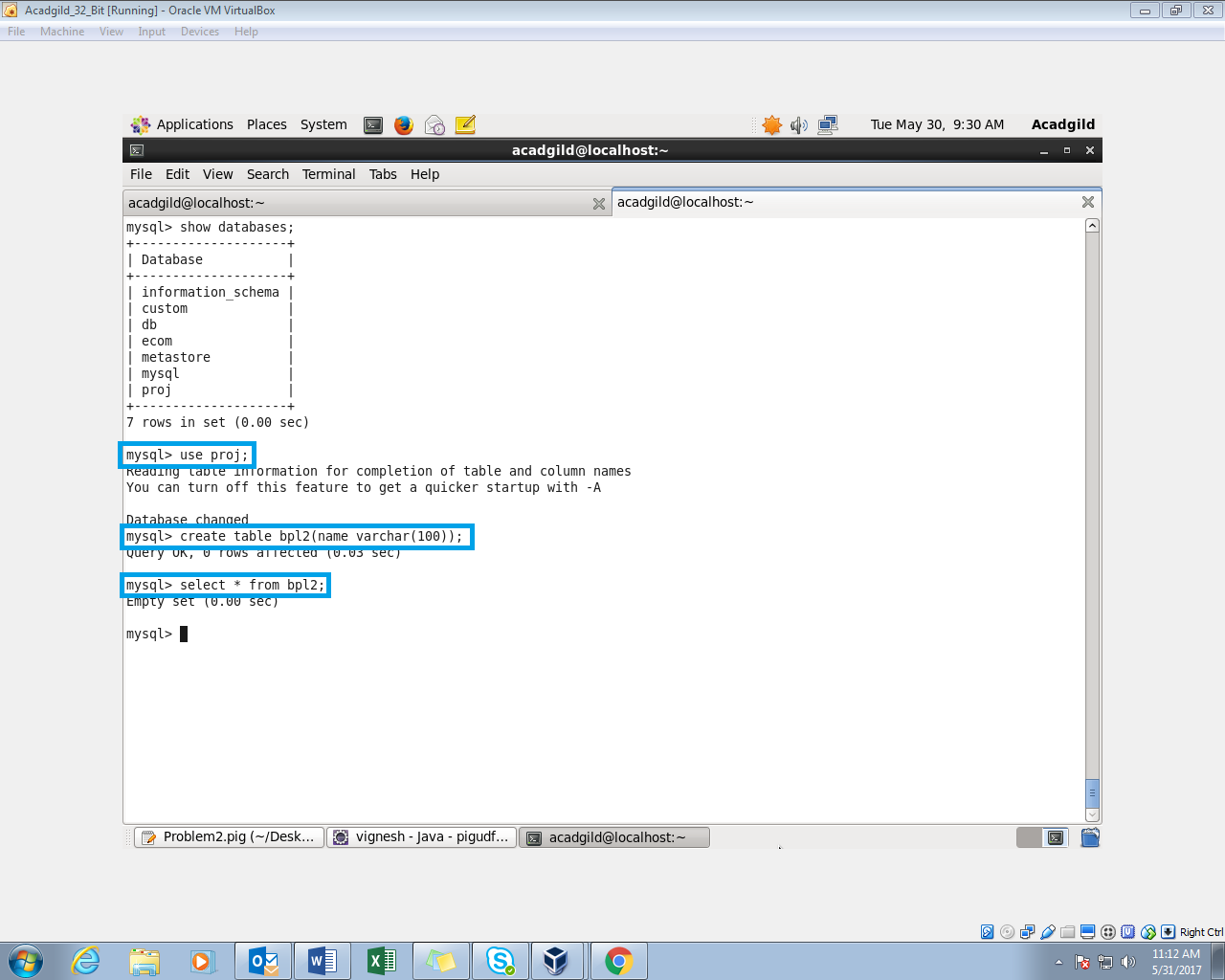


*The pig script is executed and it is displayed using the cat command.*

*Sample Output*:

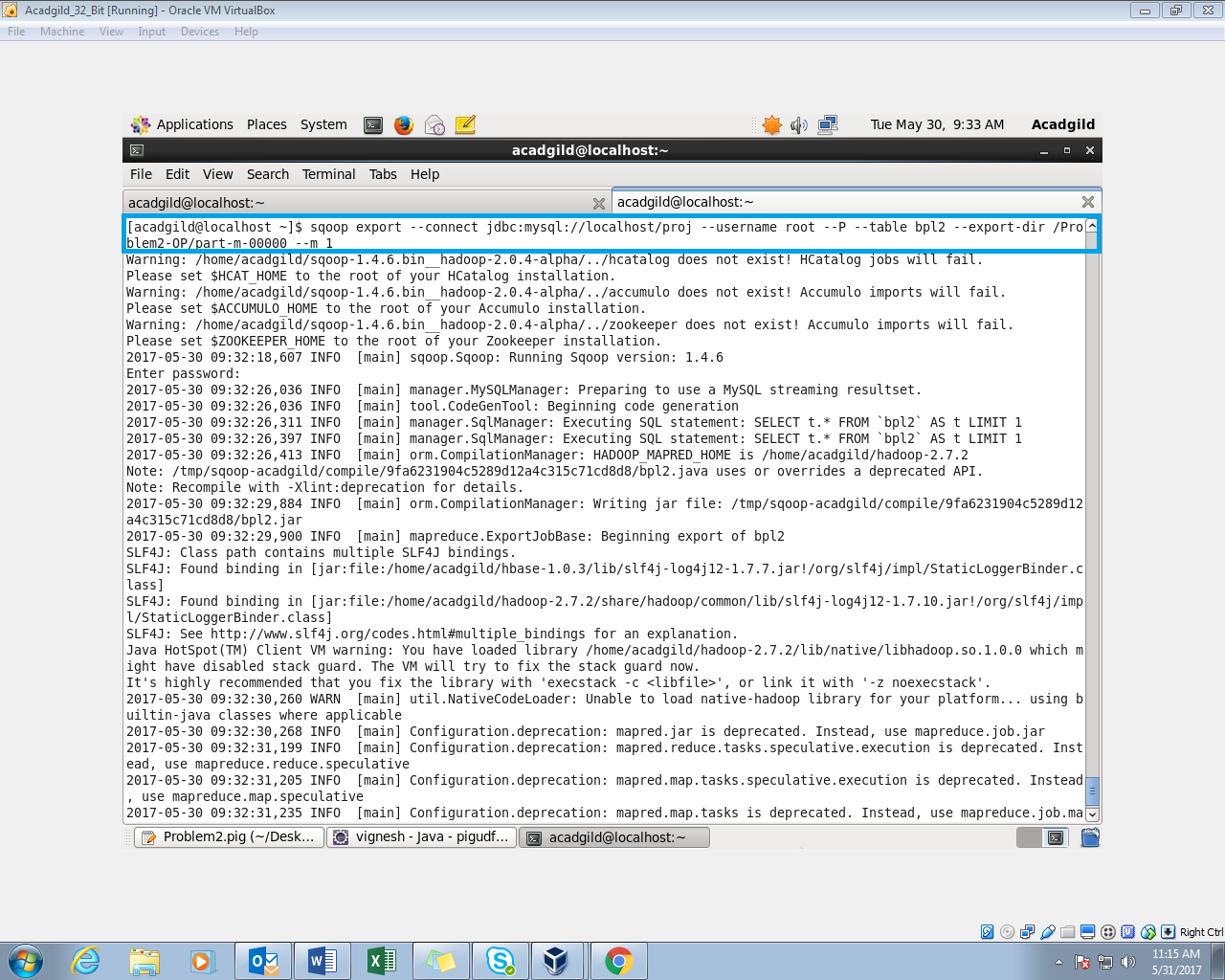


*Exporting the output to mysql*:



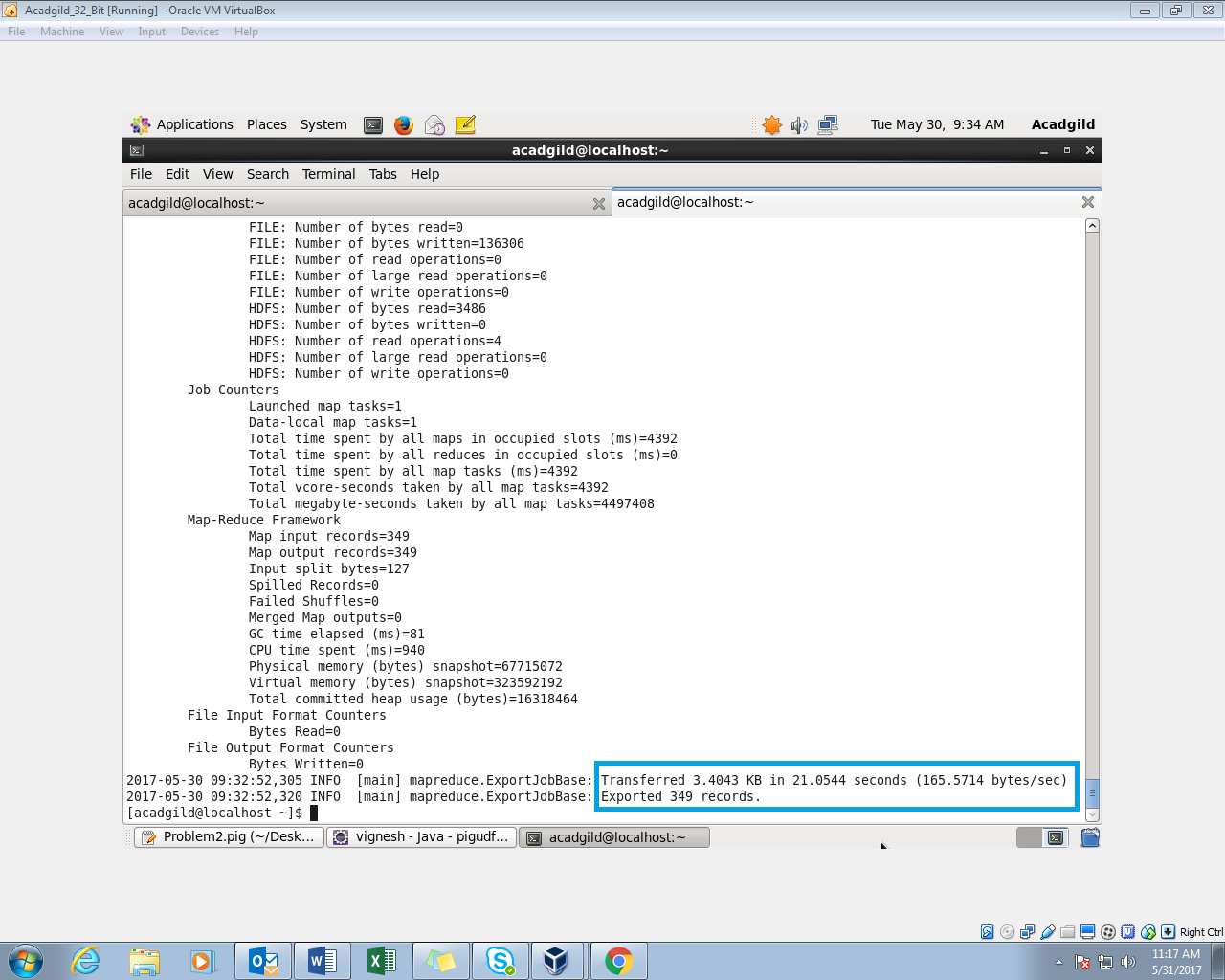
.

*Sqoop Command to export the results to the mysql*:



*The hdfs directory path and the table names are provided in the command.*

*exported the results to mysql*:



*The Contents of the table is displayed*:

