***MAJOR PROJECT***

**1. Executive Summary**

1.1 Project Overview

To develop the System to analyze the log data (In XML format) of government progress of various development activities.

1.2 Purpose and Scope of this Specification

The purpose of this project is to capture the data for analyzing the progress of various activities.

In scope

The following requirement will be addressed in phase 1 of Project:

* Developing system to handle the incoming log feed and store the information in Hadoop

Cluster (Flume)

* Analyze the data and understand the progress
* Store the results in Hbase/RDBMS

Out of scope

We can use this data and visualization and get more insights

**2. Product/Service Description**

2.1 Assumptions

Log will be generated in XML format and stored in a server

2.2 Constraints

Describe any item that will constrain the design options, including

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

**3. Requirements**

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

Priority Definitions

The following definitions are intended as a guideline to prioritize requirements.

* 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:

Link:

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

Refer the below steps to understand the actual steps to create the

above project.

**Step 1:**

Copy dataset from local file system to HDFS using flume.

Note: use the conf file by downloading from below link.

Click here to download

Command:

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

**Step 2:**

Input file is in the XML format use Map reduce or pig to parse the data and get the results for

the below problem statements.

**4. Problem statement**

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

Export the results to mysql using sqoop

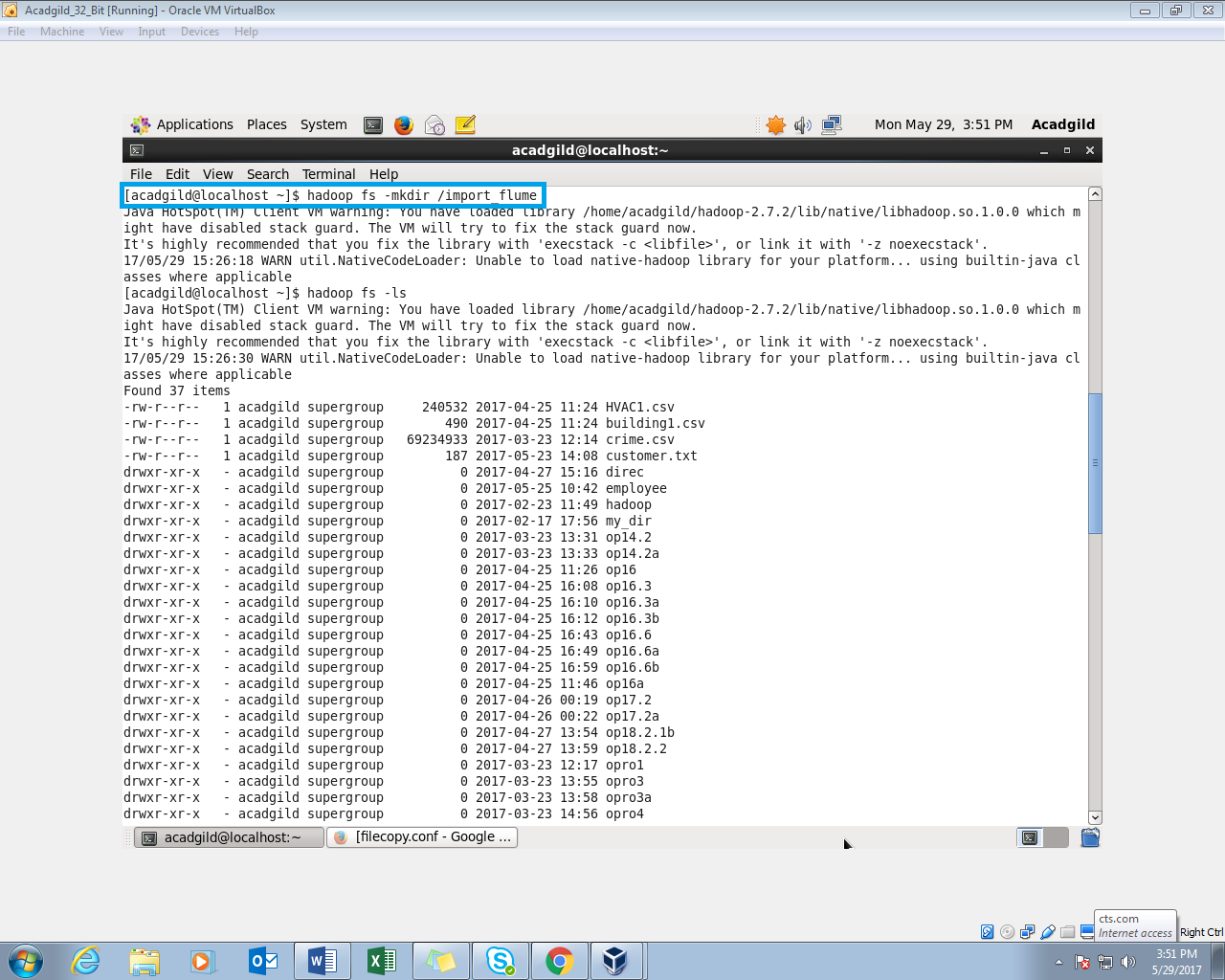
2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.

Export the results to MySQL using Sqoop

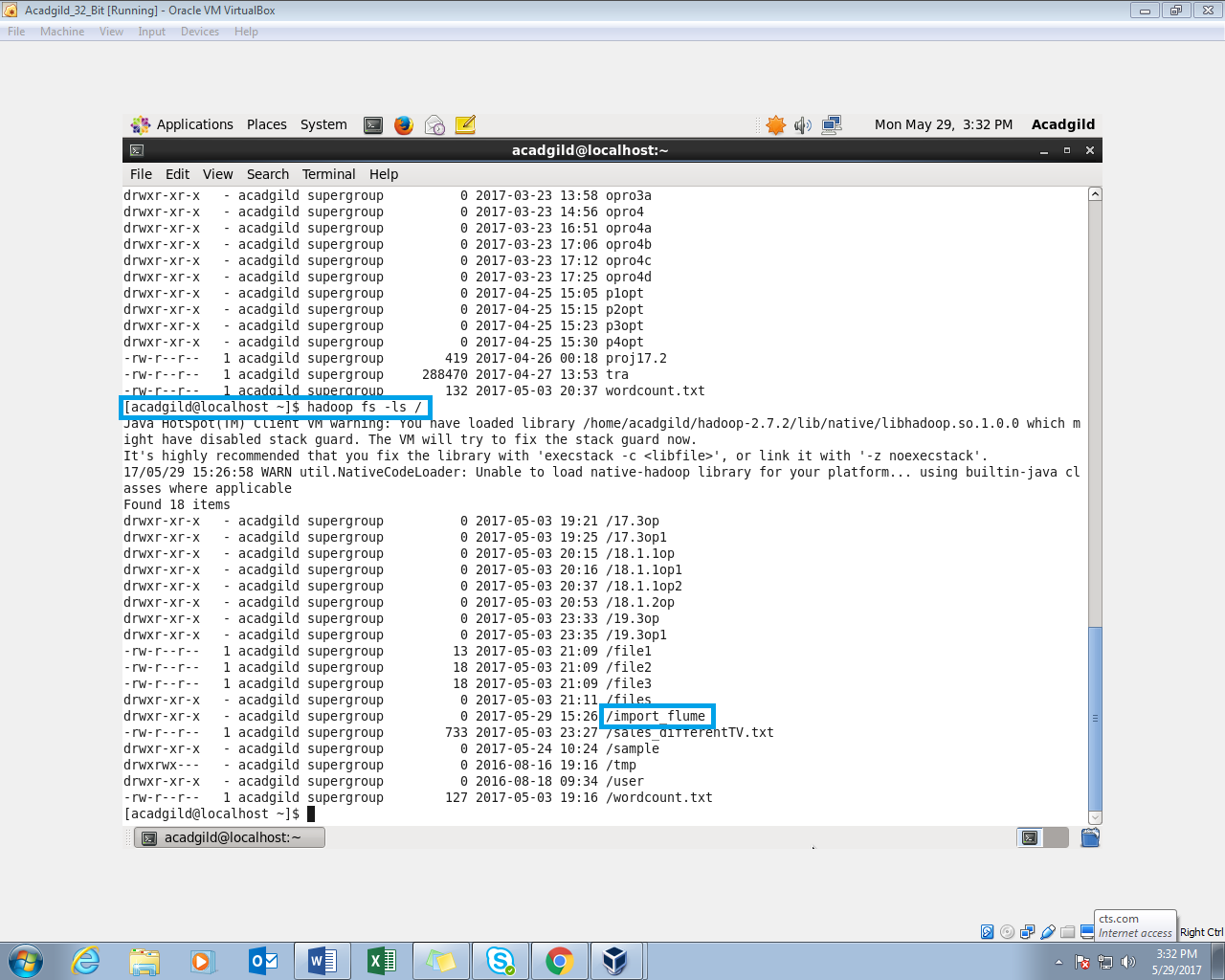
**SOLUTION :**

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

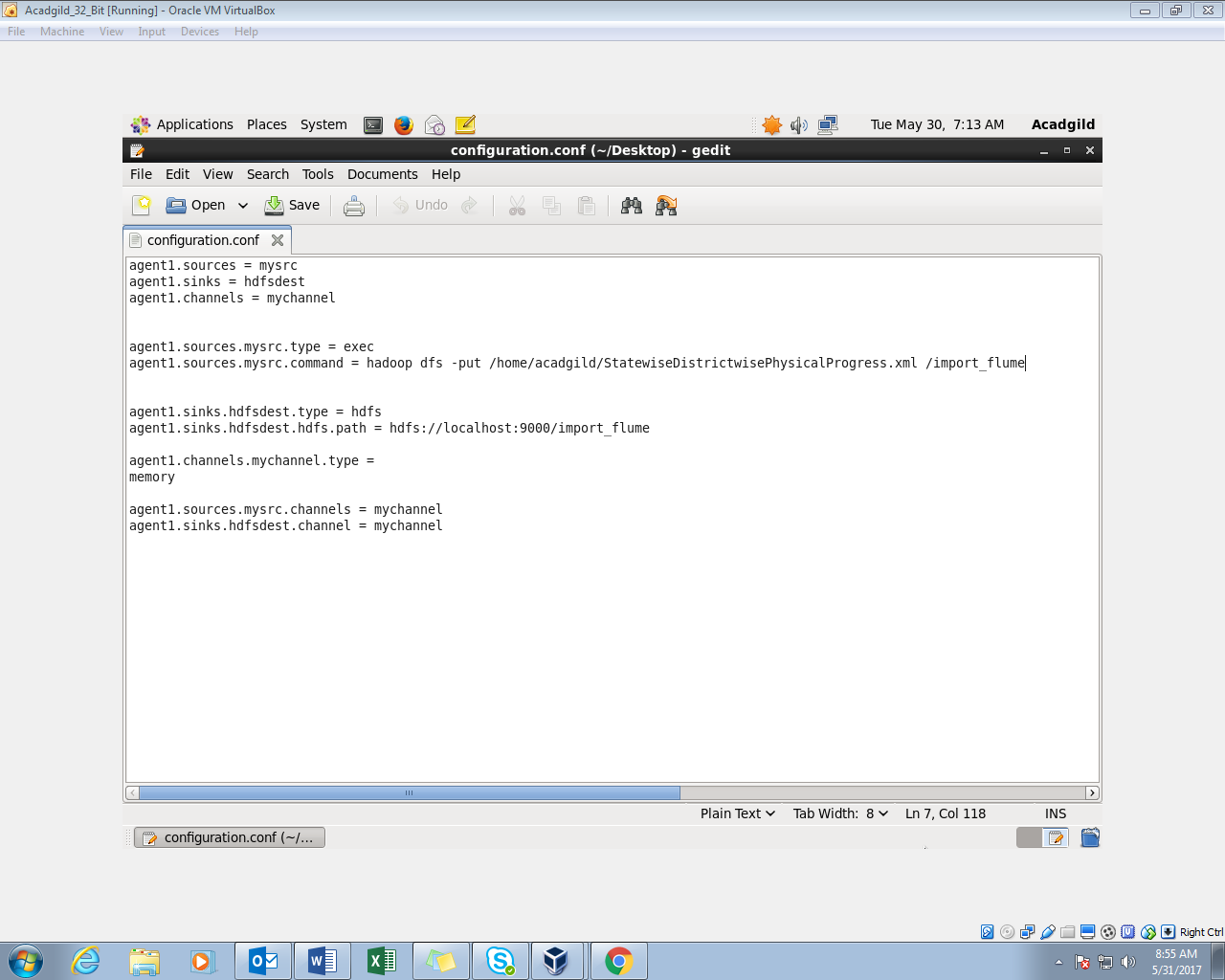
The dataset will copied to this directory for further analysing.

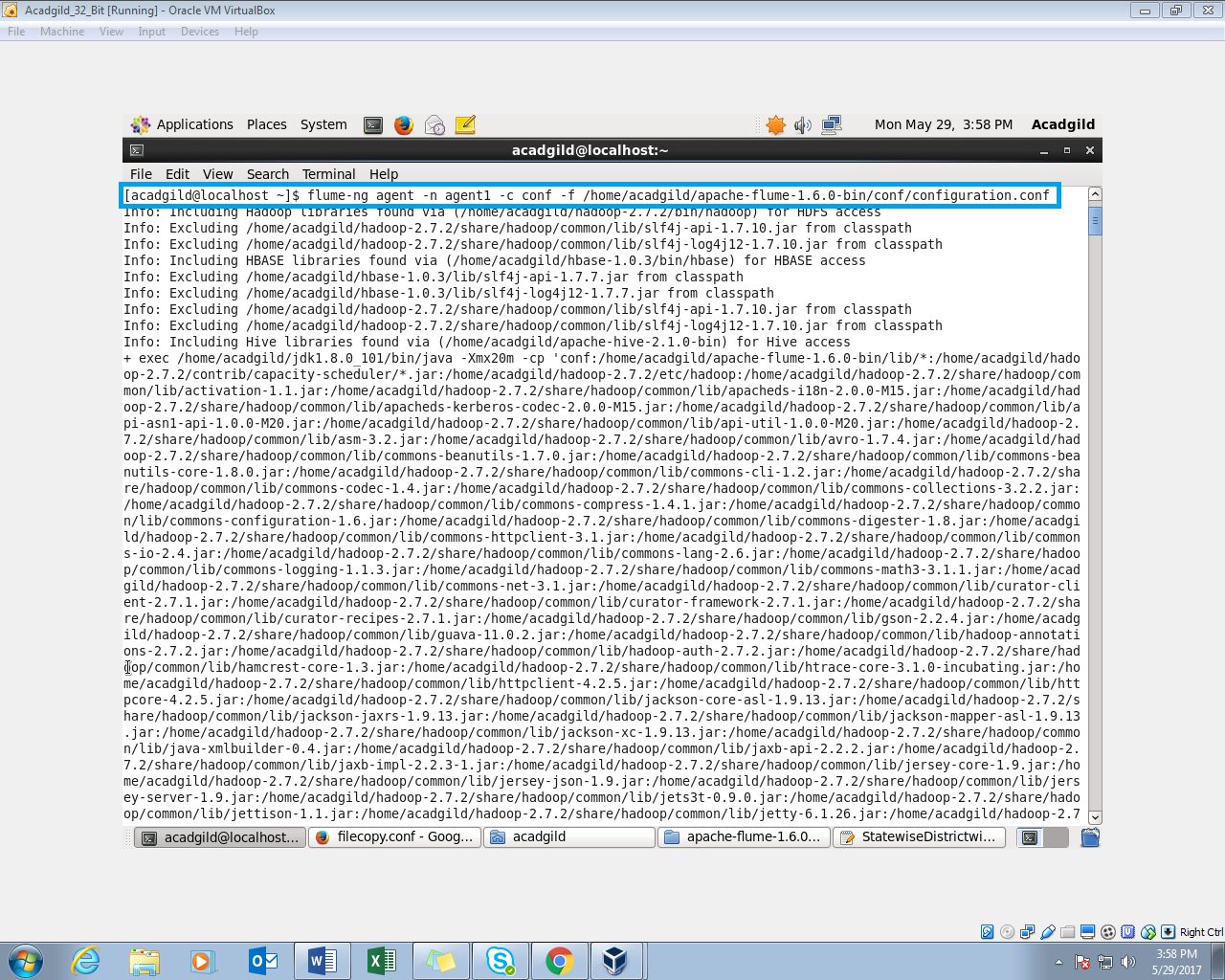


The Directory (import\_flume) is created in the HDFS:

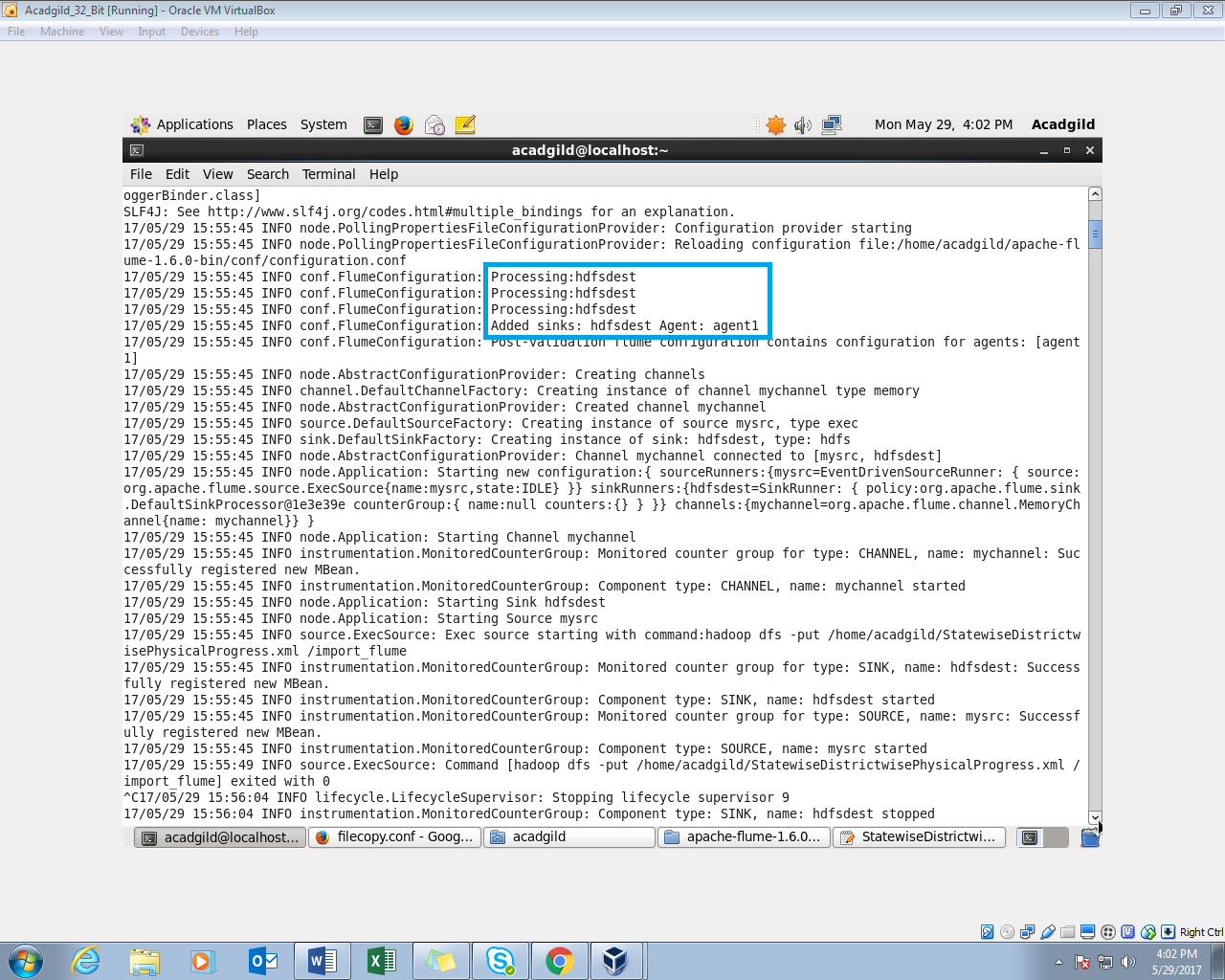


**Step1** : Create a ‘configuration.conf‘ file in the conf folder of the flume:

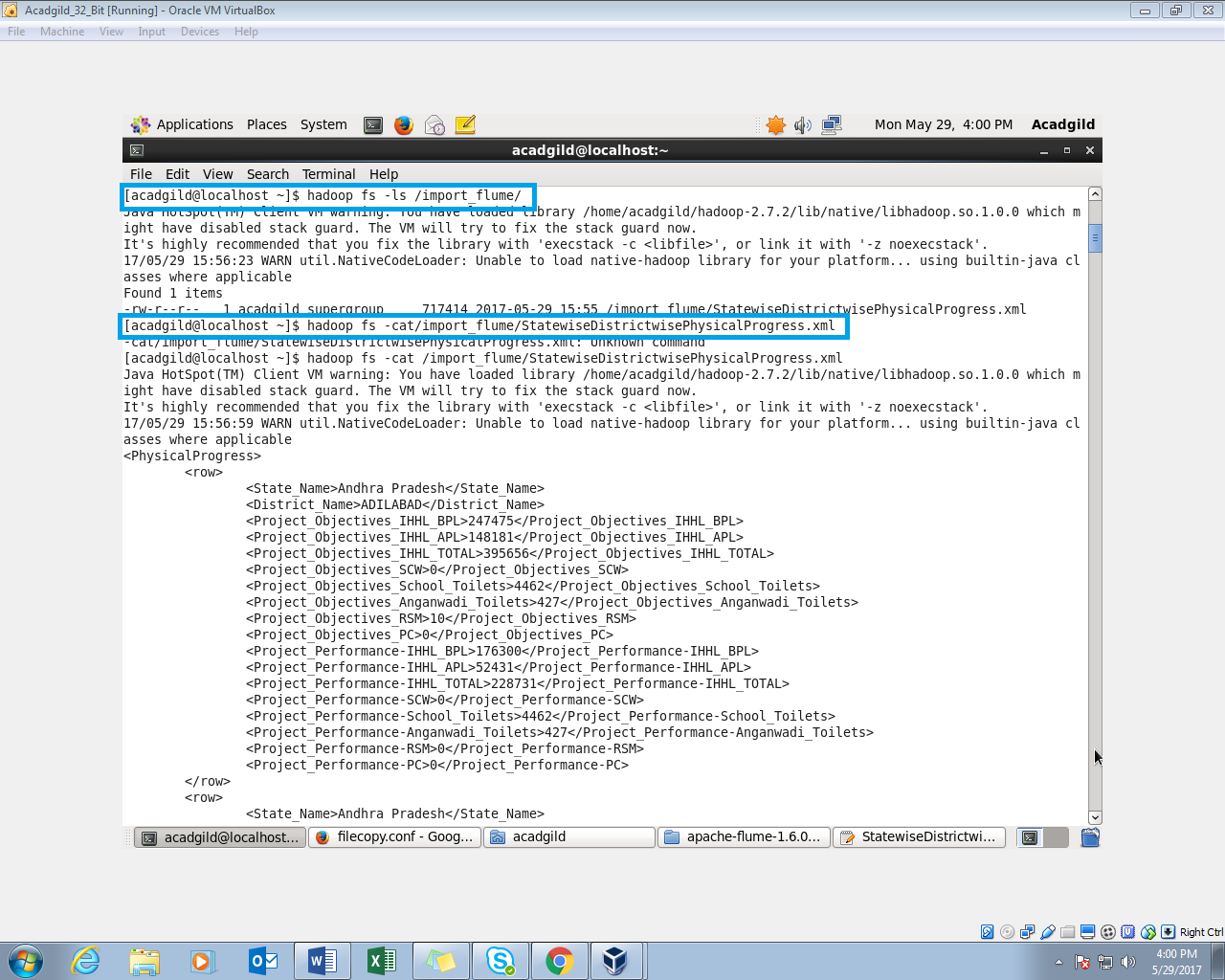


**Flume Command:** 

In the above screenshot, the name of the agent is agent1 and path to the configuration file is set.

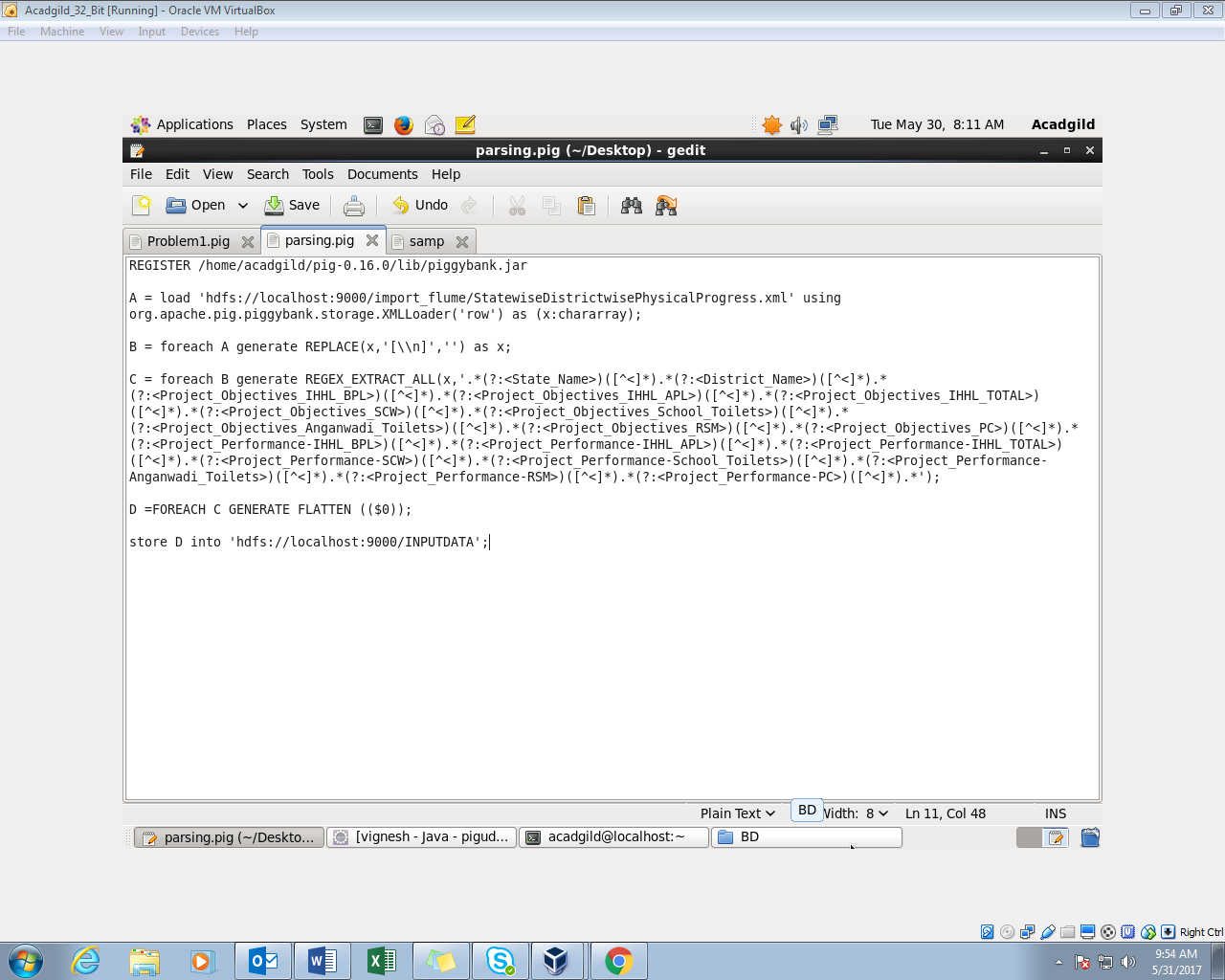


The import\_flume directory is created and the input dataset is copied into it:



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

Now the input data file is in xml file and it has to be converted to csv file. We will use a pig script to do it.

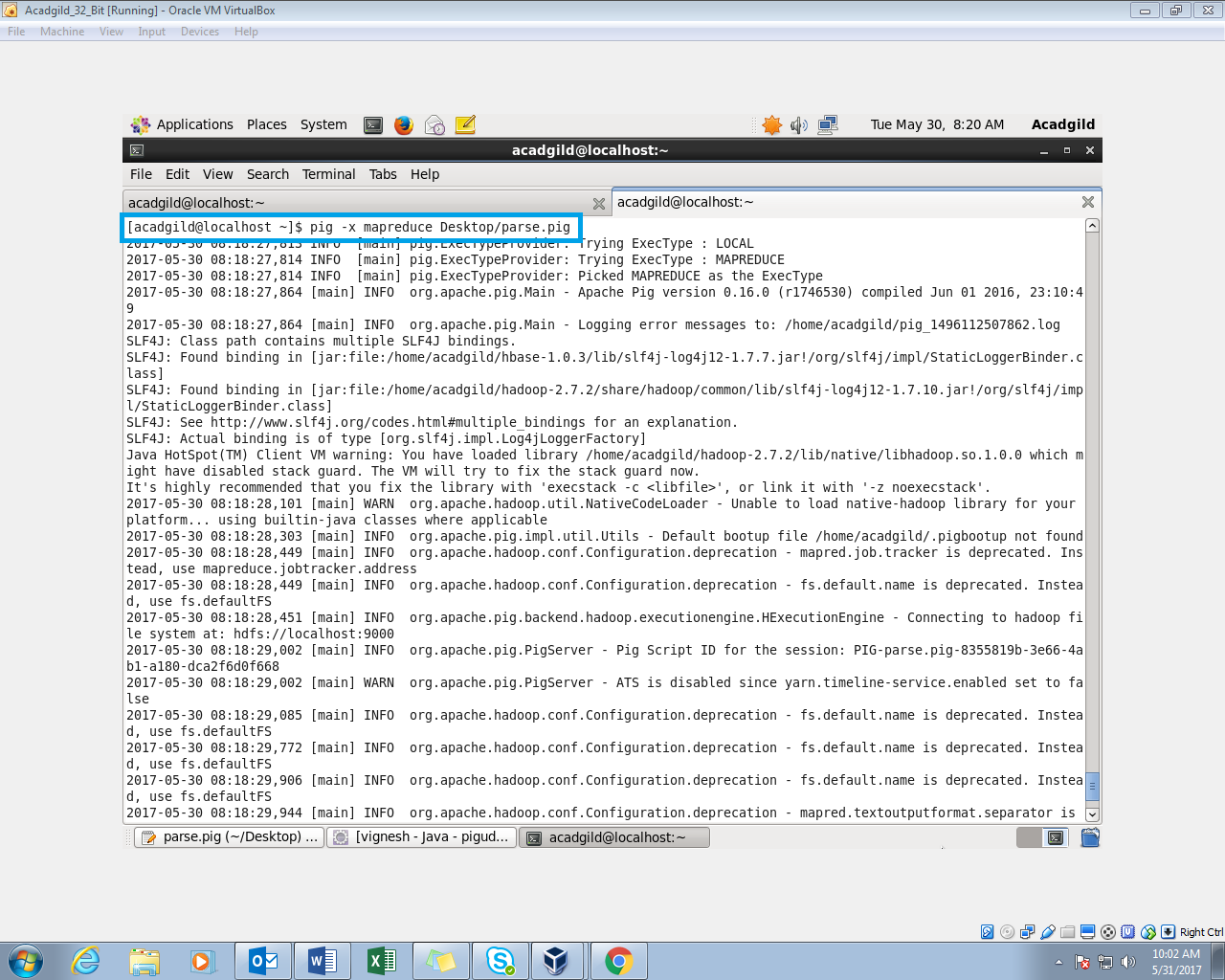


* Register the piggybank jar.
* Load the dataset from the hdfs using the XMLLoader using chararray. If there is a new line present in between the file, it will be removed.
* Use regex\_extract\_all function to capture values within tag.
* The Output will be in a bag and it is un-nested using flatten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

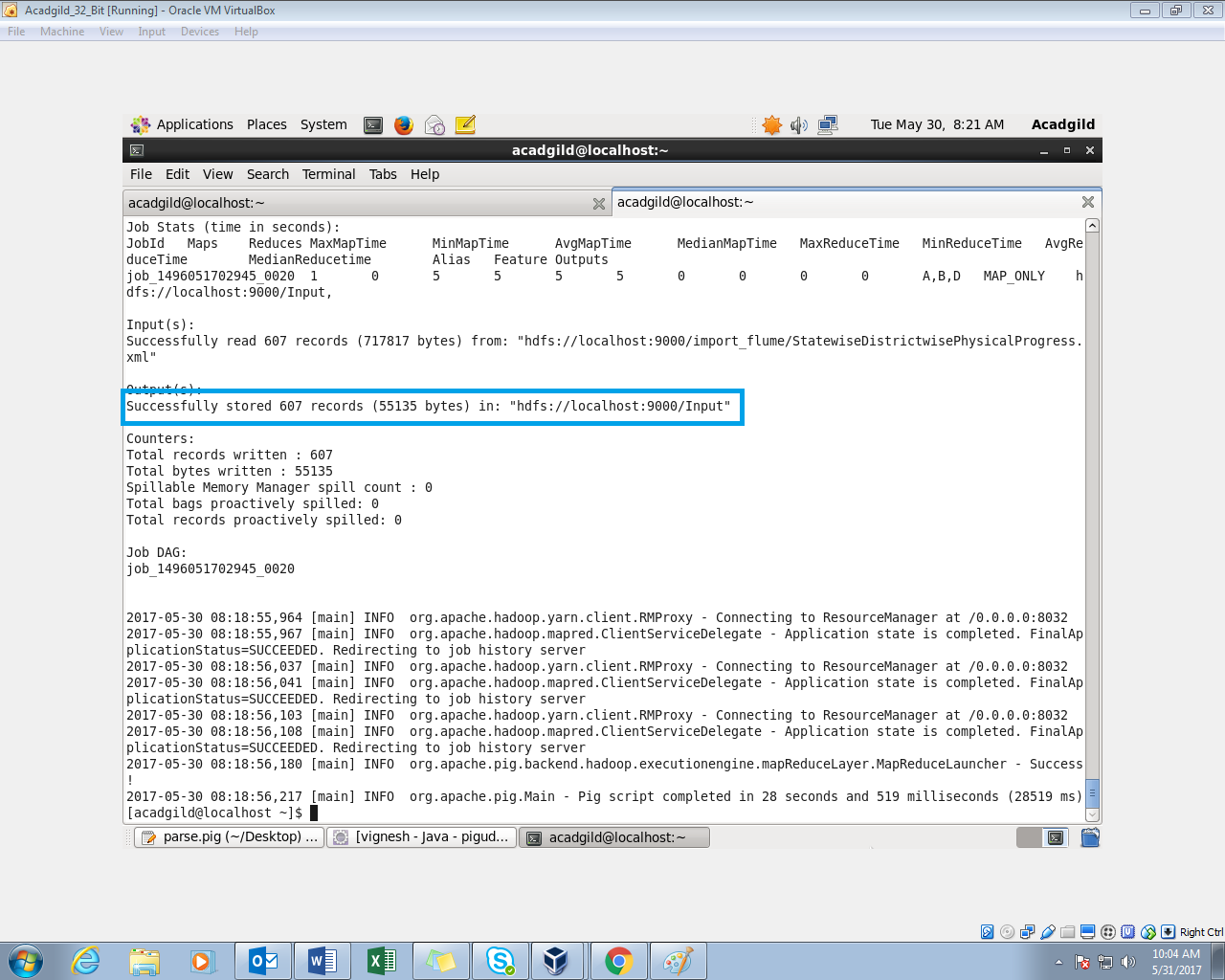
The above pig script is executed in the mapreduce mode. The Output of the data is stored in the ‘Input’ folder.

The name of the pigscript is parse.pig

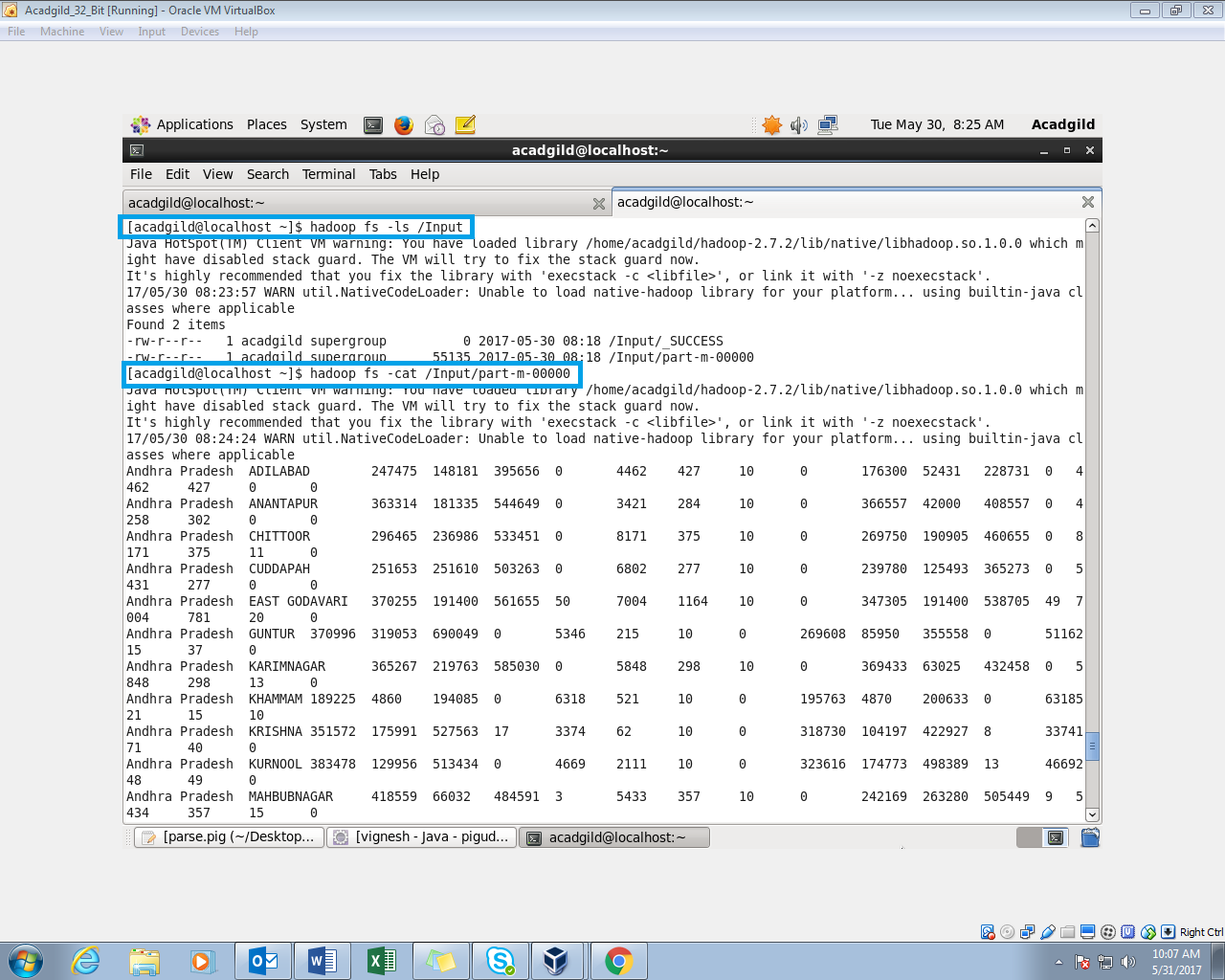
Executing the PigScript in mapreduce mode:



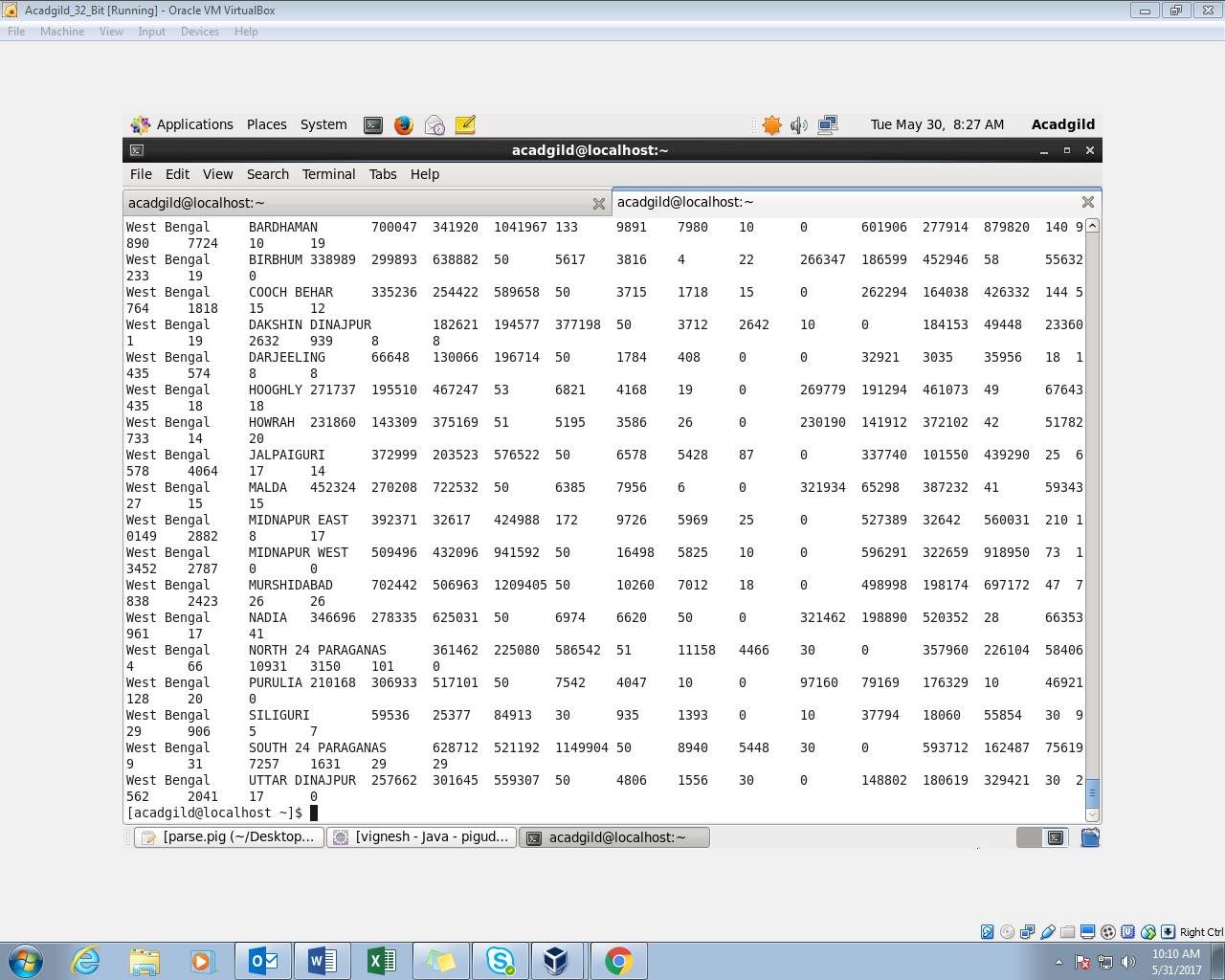
Output is successfully stored in the ‘Input’ folder:



Display the parsed input data file:



Sample data:



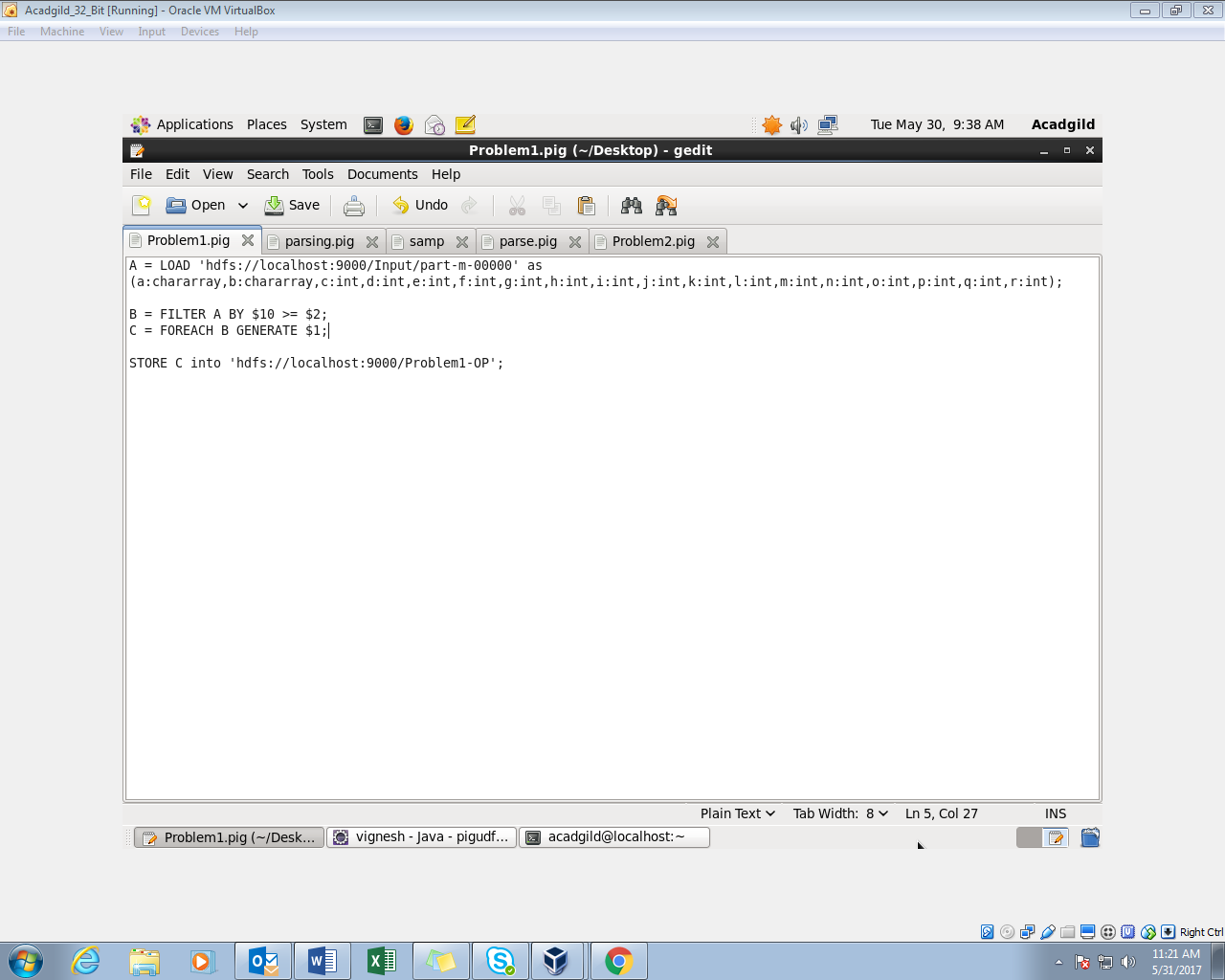
**Problem statement**

**Question 1:**

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

Export the results to MySql using sqoop

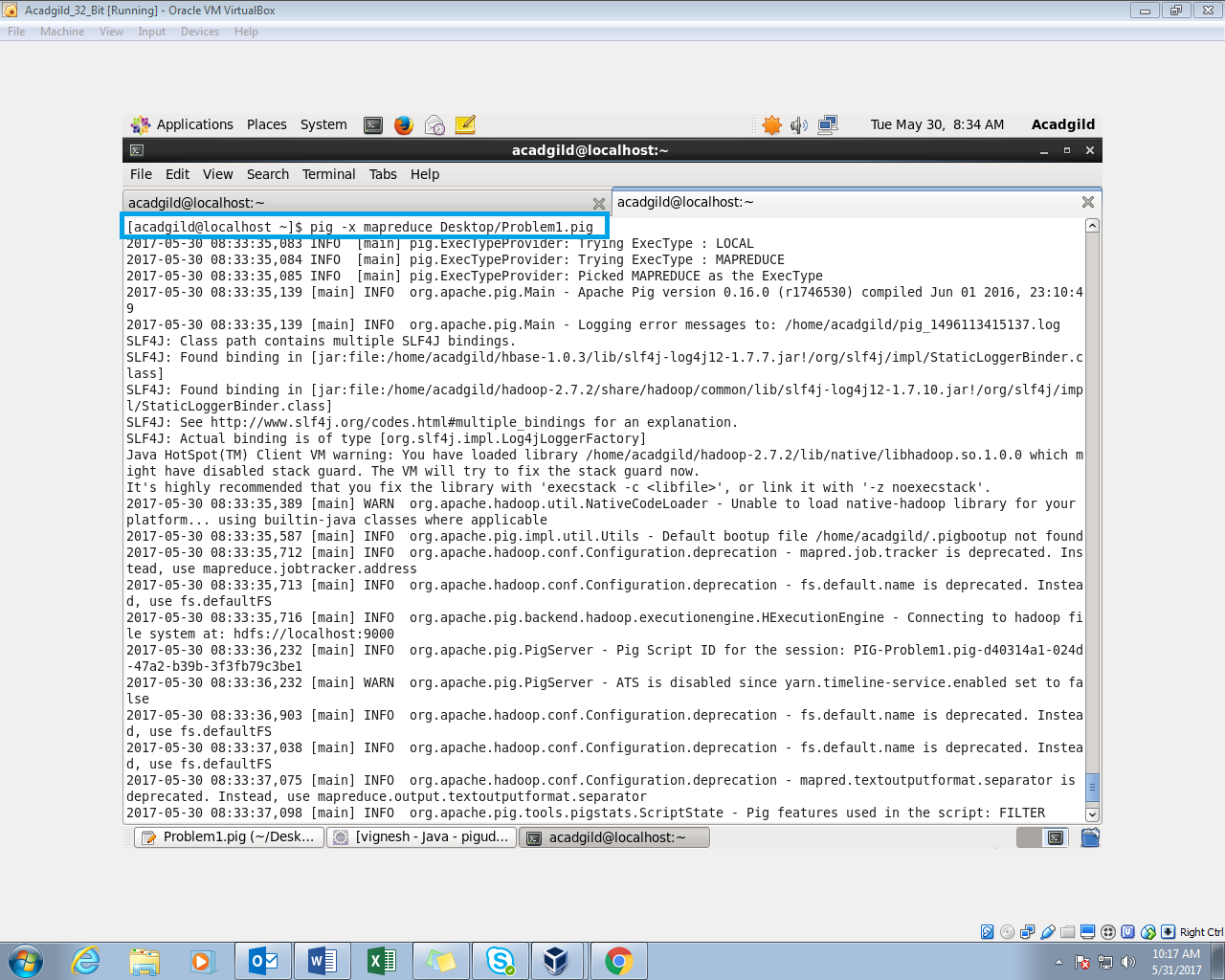
Pig Script:



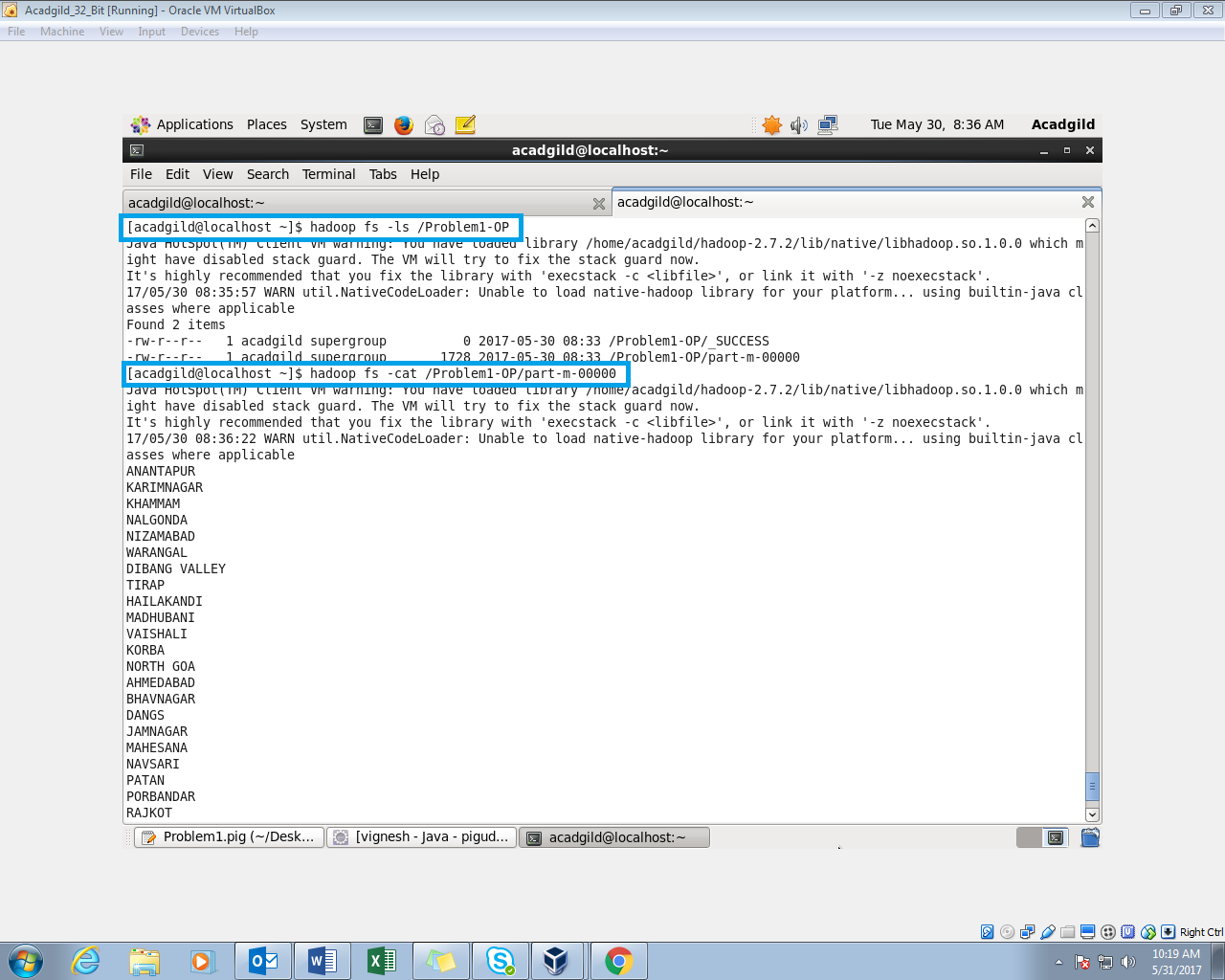
* Load the parsed data.
* We need the districts which achieved 100 percent objectives. 11th ($10) column is project performance and the 3rd ($2)column is project objective BPL. If the performance is higher than the objective, it is considered to be achieved 100 percent. We are using the filter command here.
* Since we need the districts we are only displaying the districts using the foreach generate command.
* Storing the output in the hdfs.

The name of the script is Problem1.pig.

Executing the Pig Script in MapReduce mode:



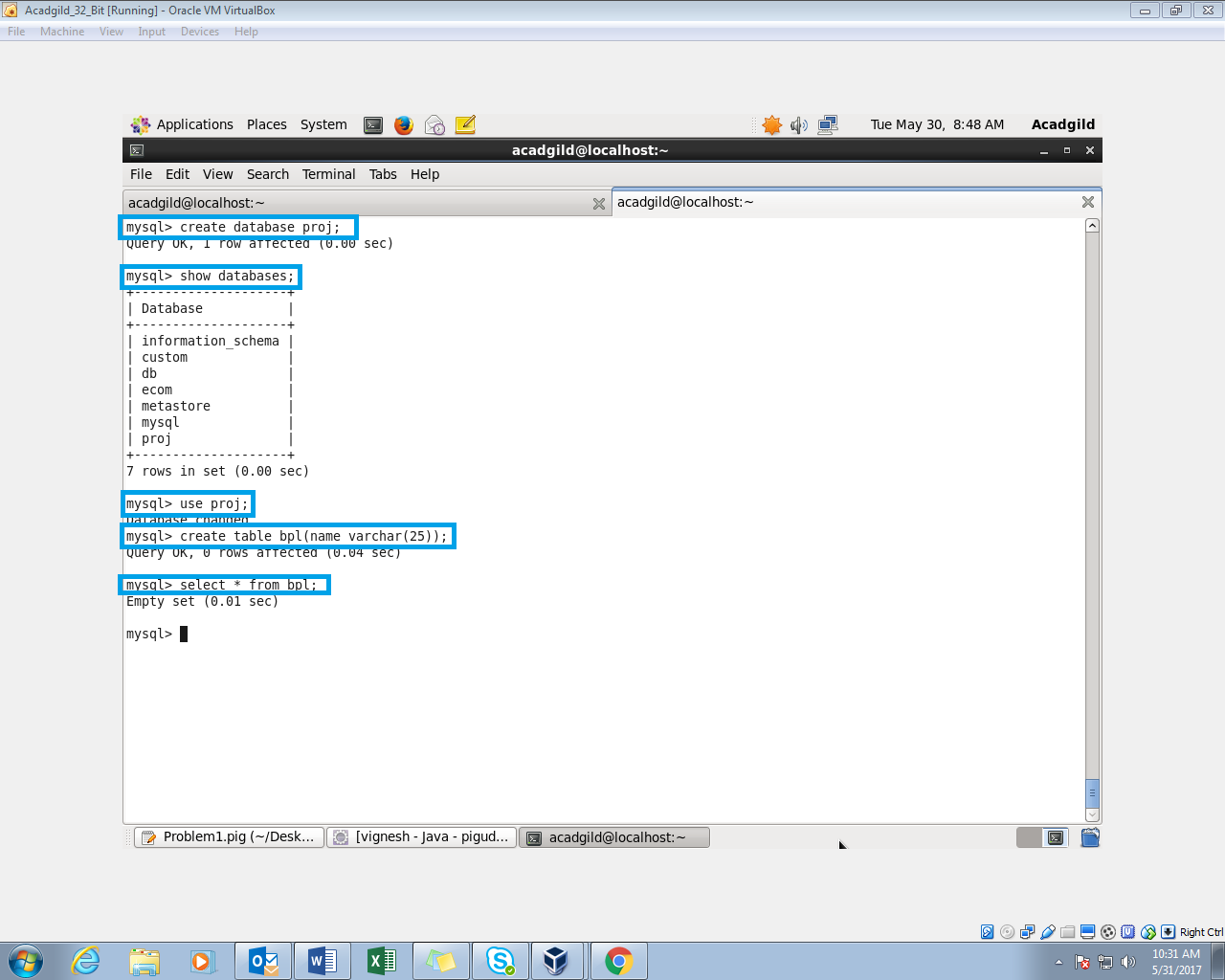
Output successfully stored in the hdfs:



Sample Output:

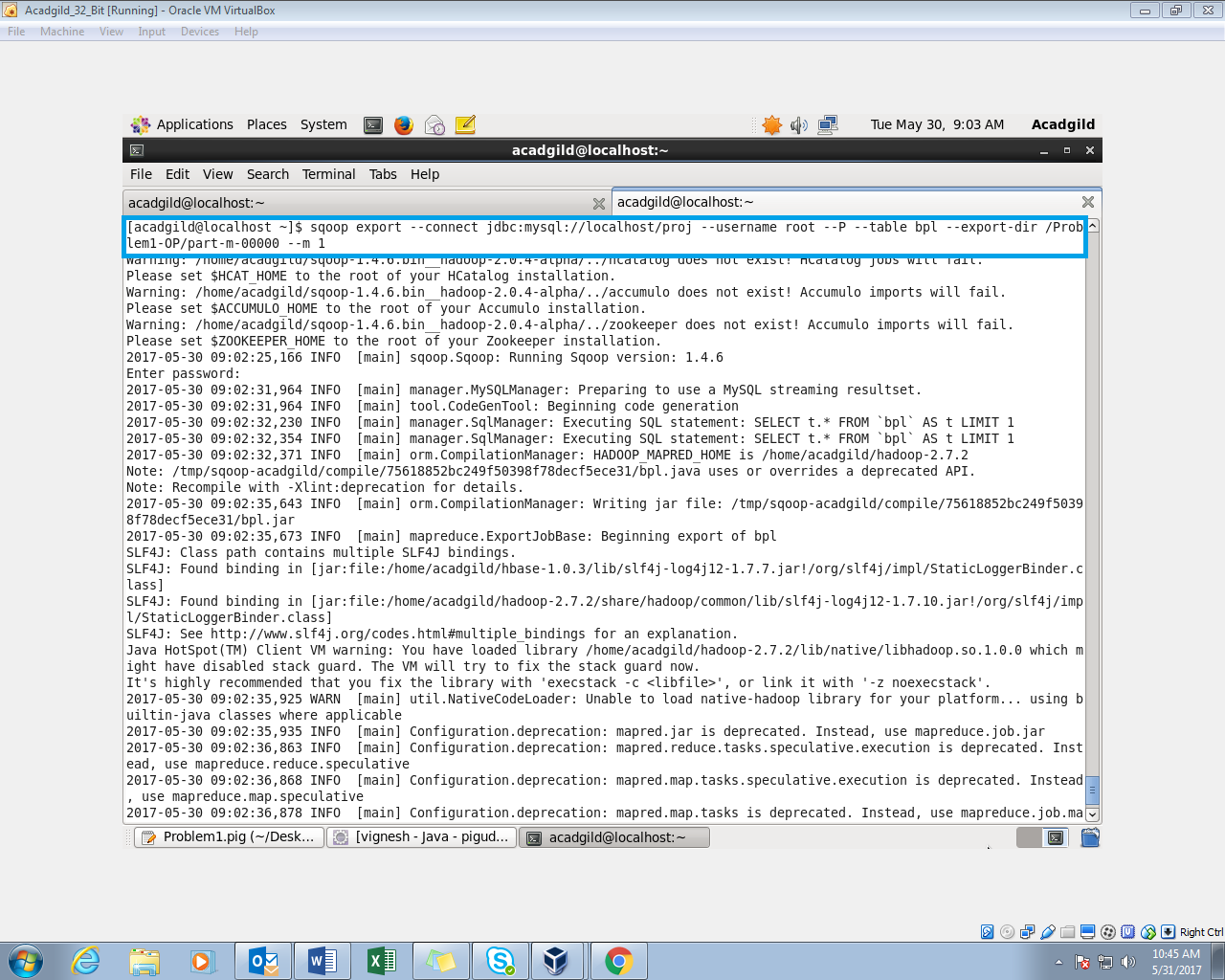


Exporting the Results to MySql:



MySql terminal is opened and the database,’proj’, is selected and an empty table named bpl is created. Before exporting the table will be empty.

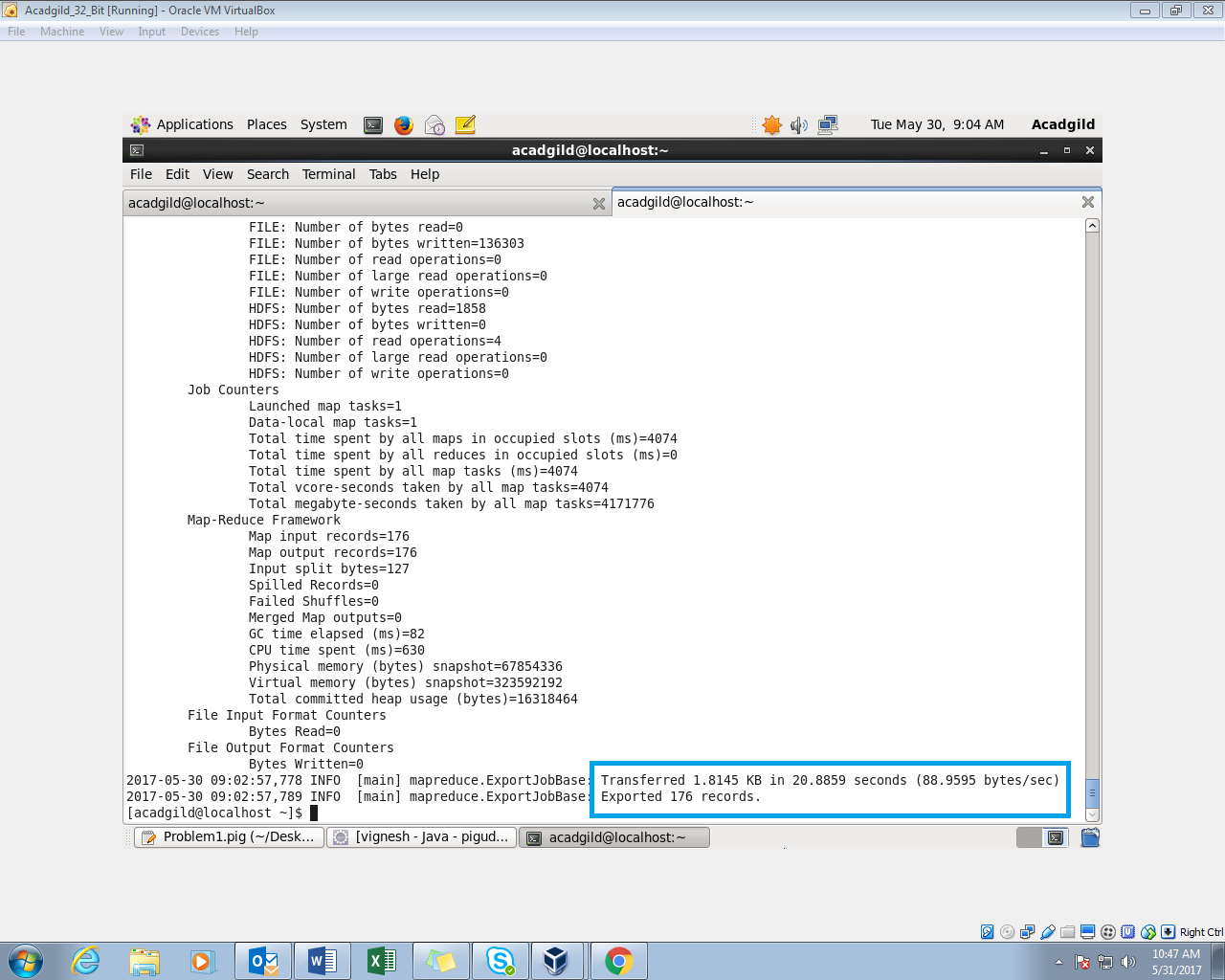
Sqoop command to export from hdfs to Mysql:



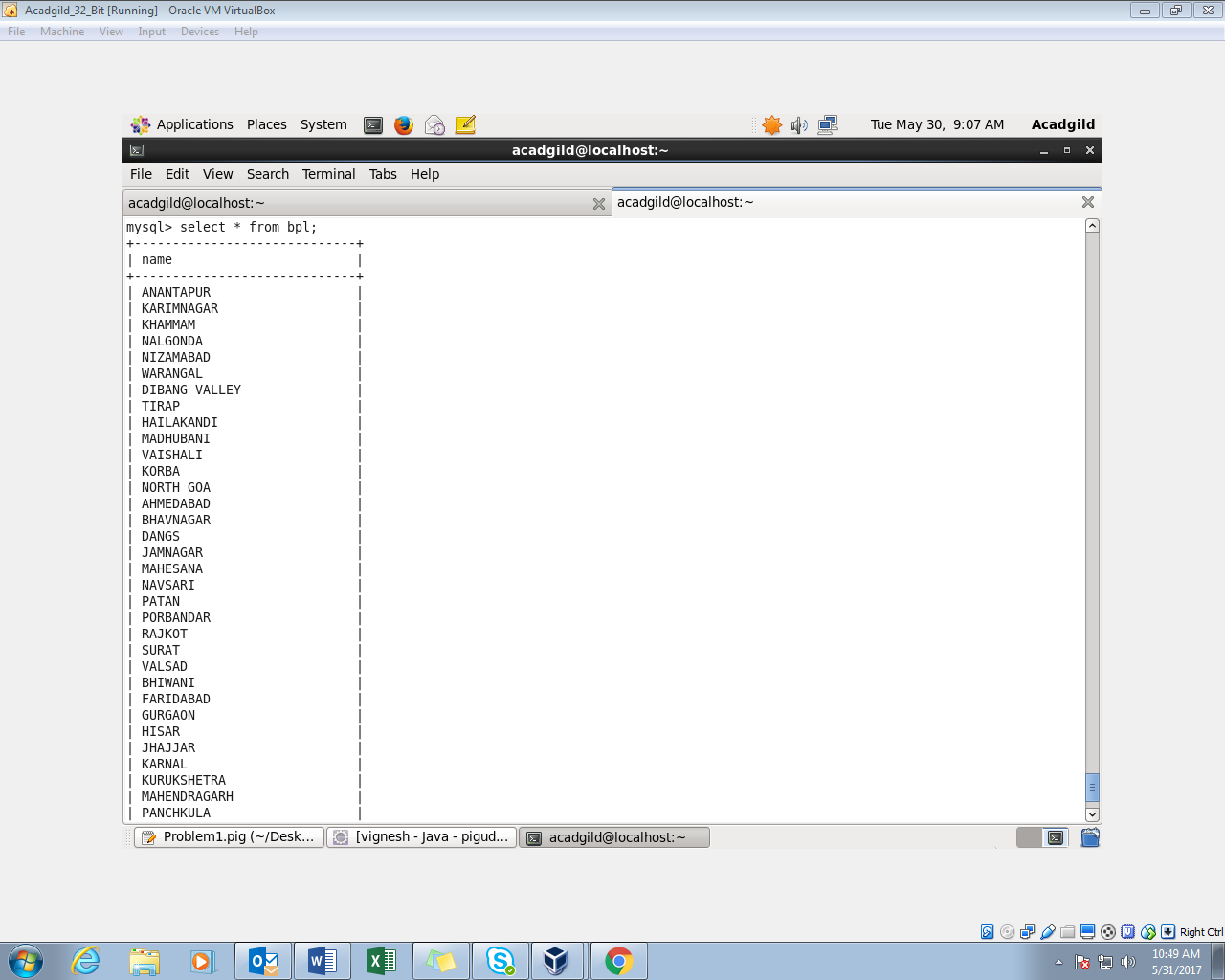
In the Command, we need to give the mysql database in which we created the table. The username and password (acadgild) to enter into the mysql are given in the command.

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

Successfully transferred the records to the mysql:



The contents of the ‘bpl’ table:



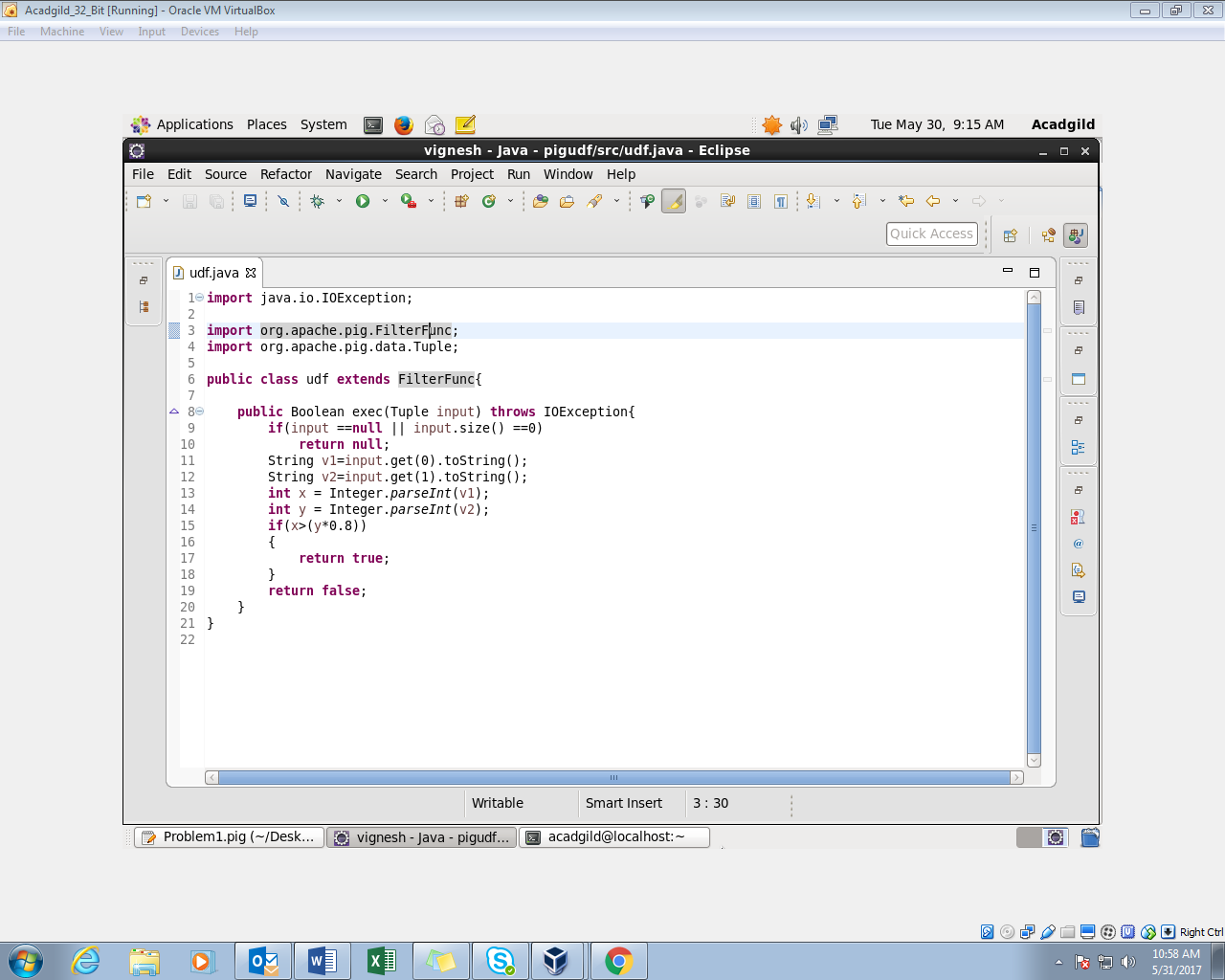
**Question-2:**

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

Export the results to MySQL using Sqoop

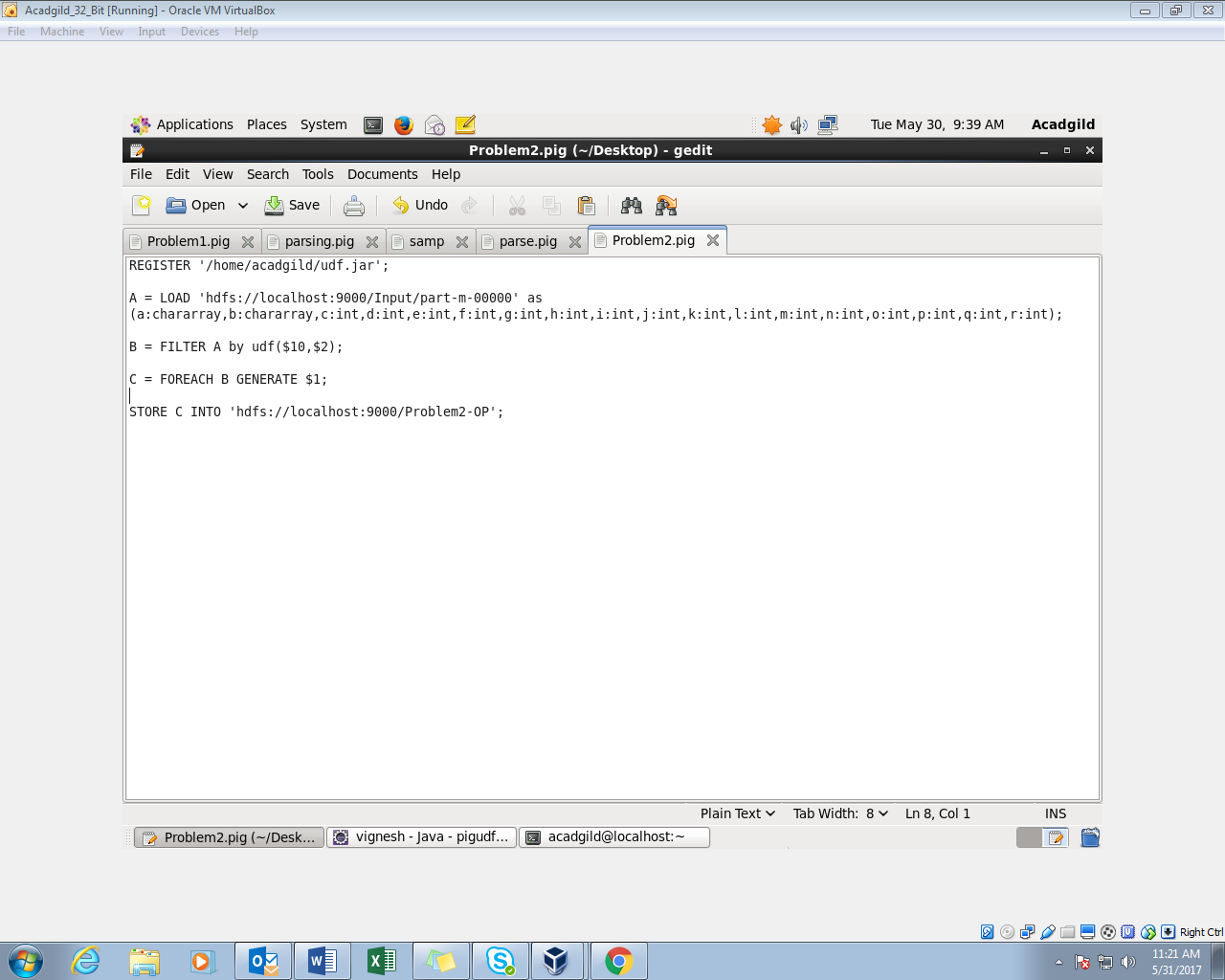
Creating a UDF to filter the districts:

Java Program:



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

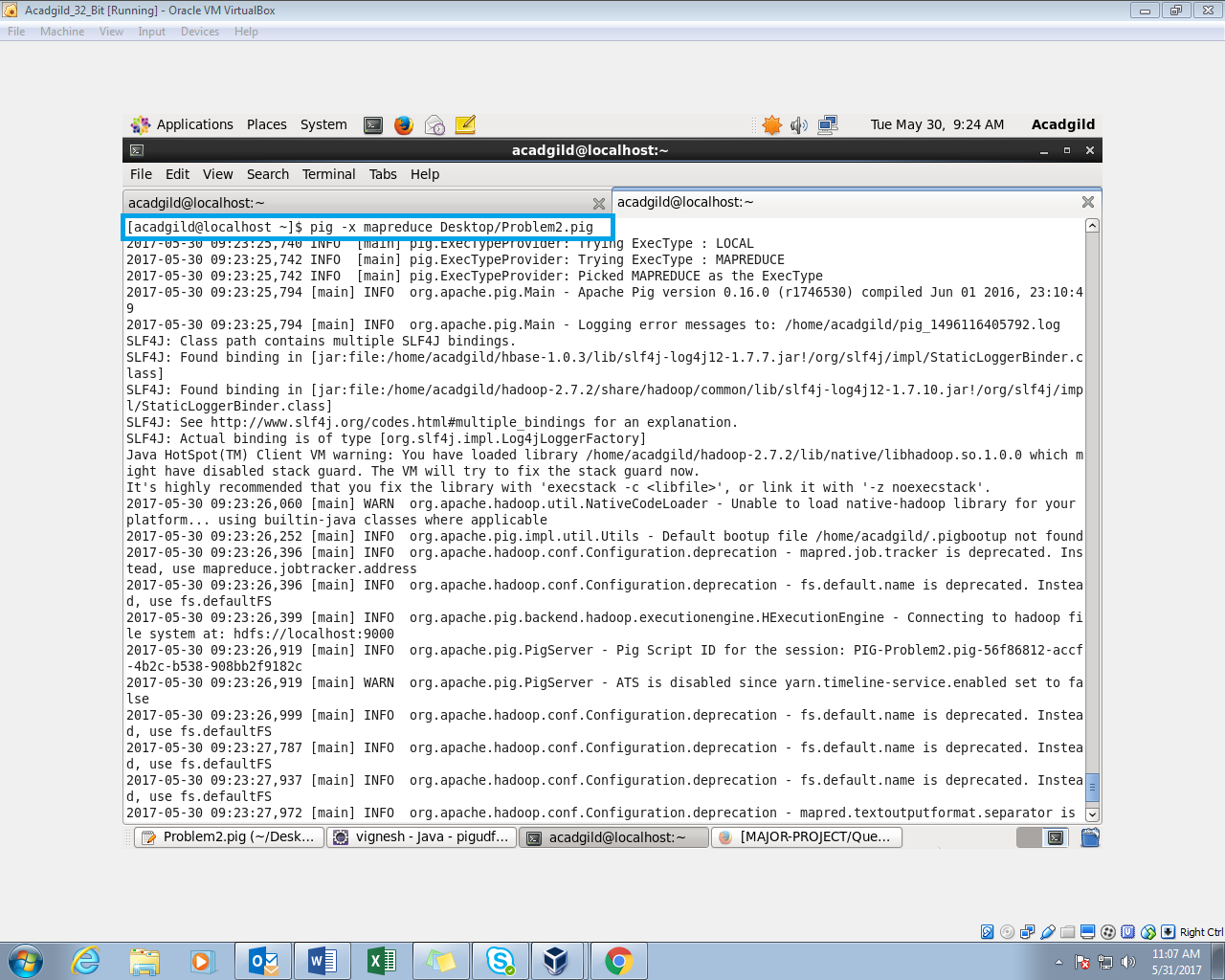
Pig Script:



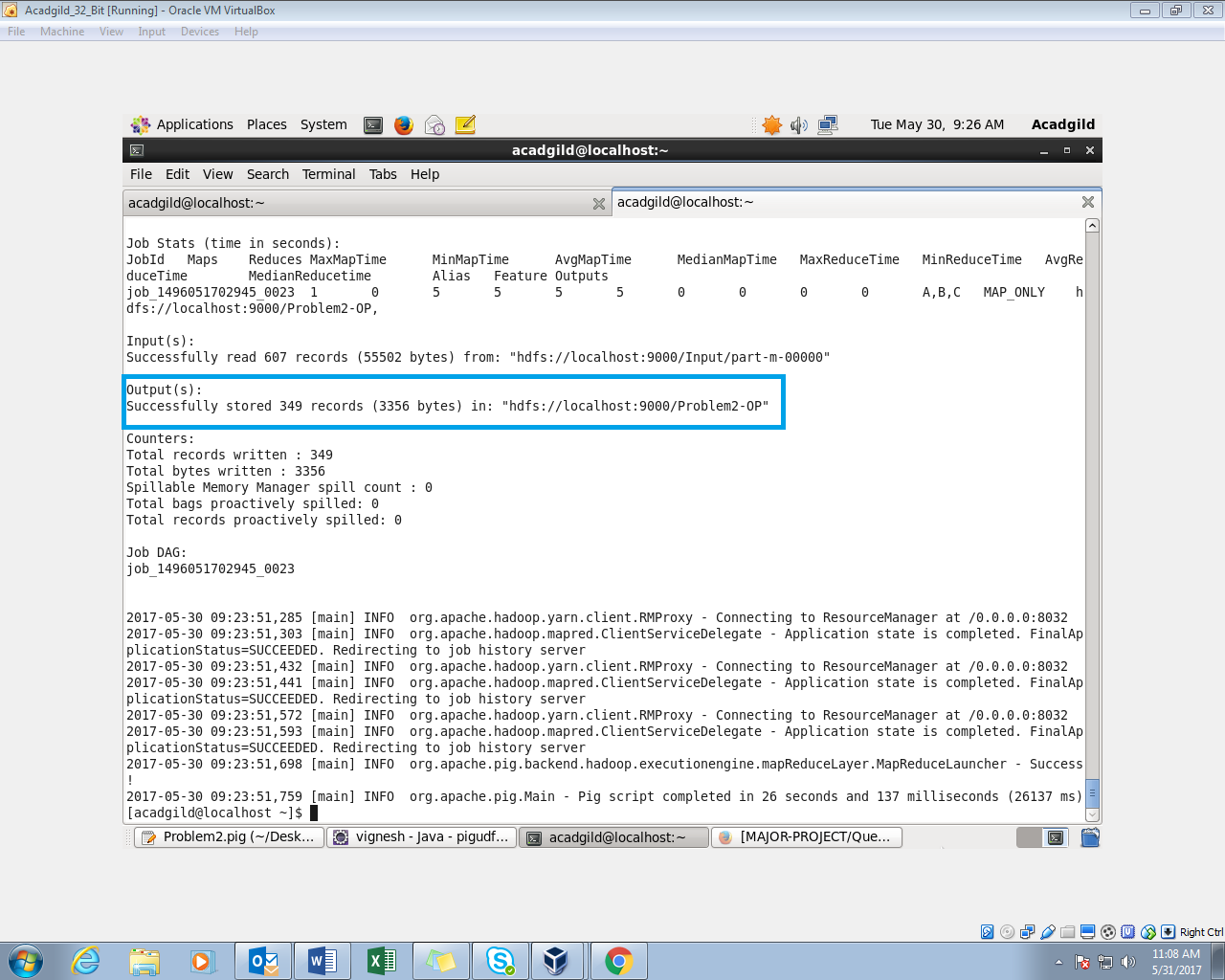
* The User defined function java program is exported to a jar file named udf, and this udf is registered.
* Load the input data from hdfs.
* Filter the loaded input by the user defined function. Here we have to use the class name (udf) to filter.
* Since we need only the districts, we are taking the districts name using foreach generate.
* Store the output in the hdfs.

The name of the pig script is Problem2.pig.

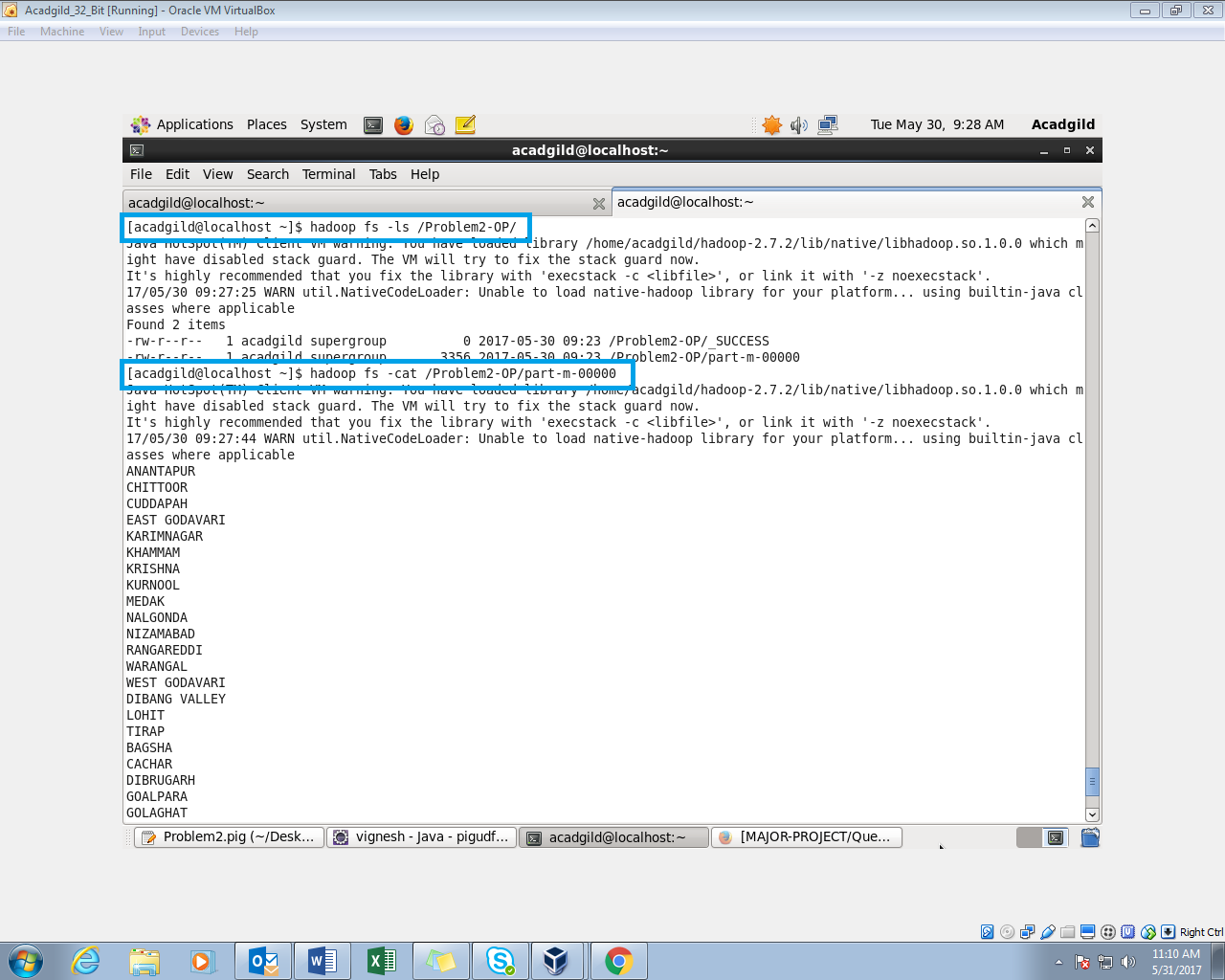
Execute the Pig Script in mapreduce mode:



Successfully stored the output in hdfs:



Check the HDFS directory:

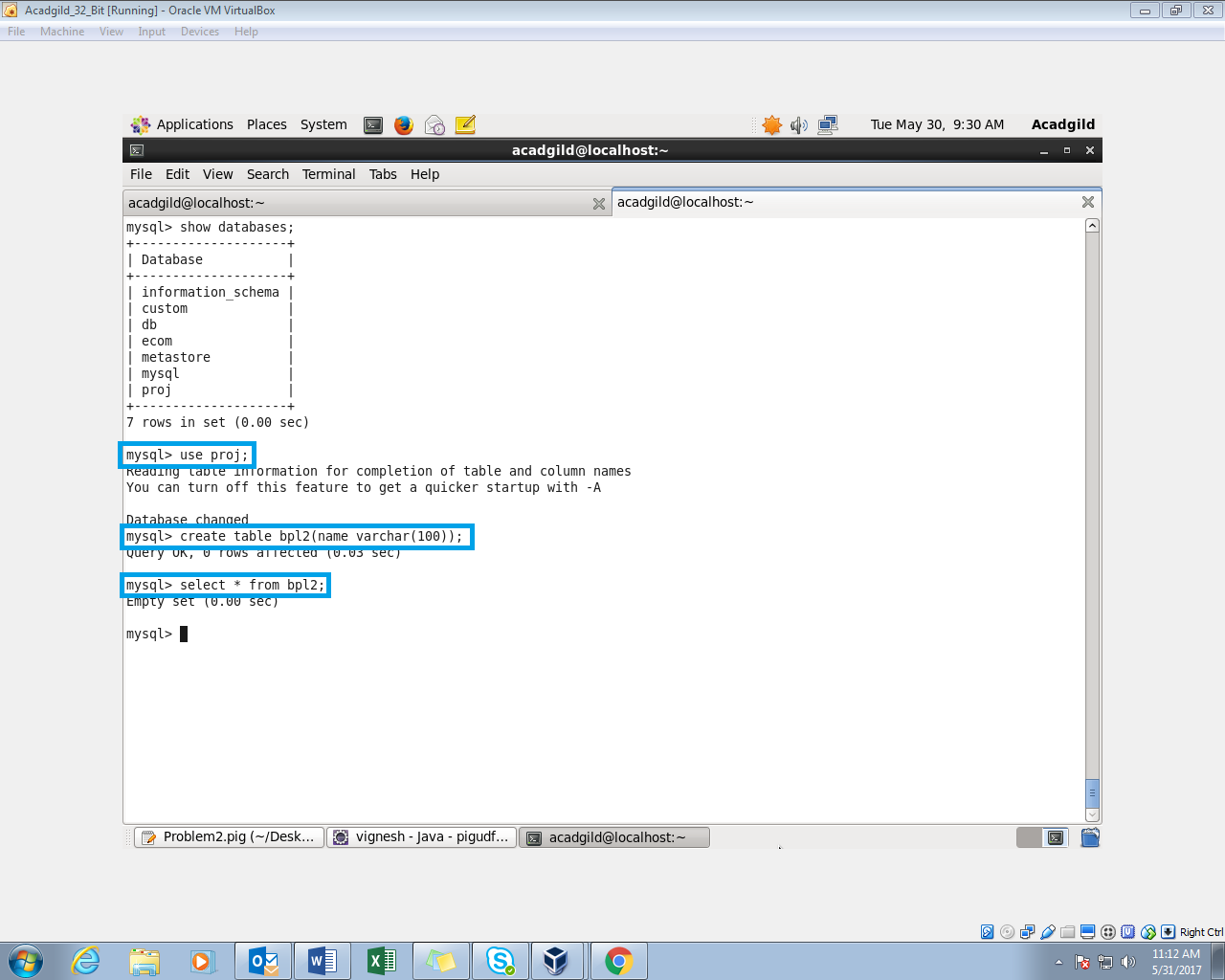


The pig script is successfully executed and the output is displayed using the cat command.

Sample Output:

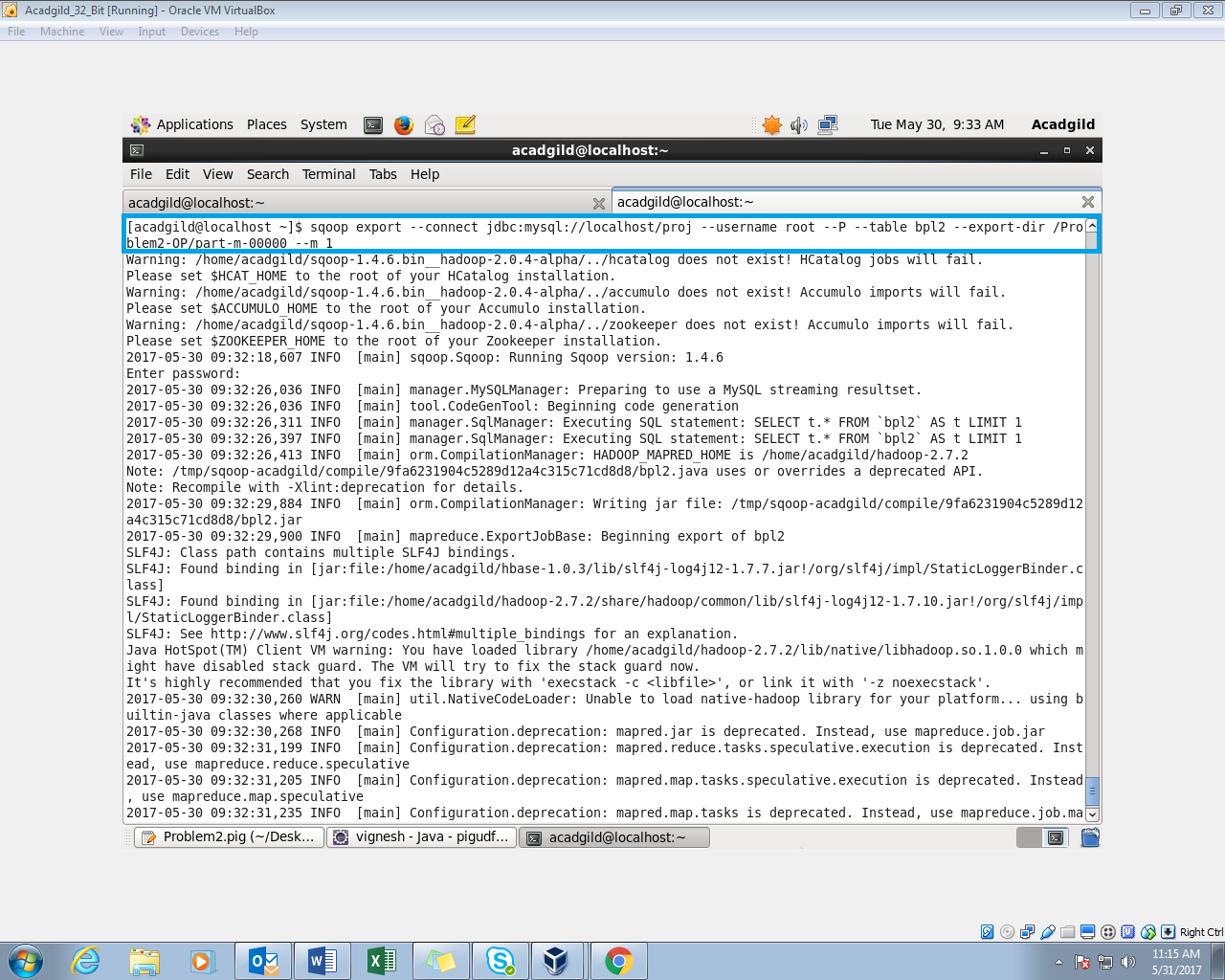


Exporting the output to mysql:



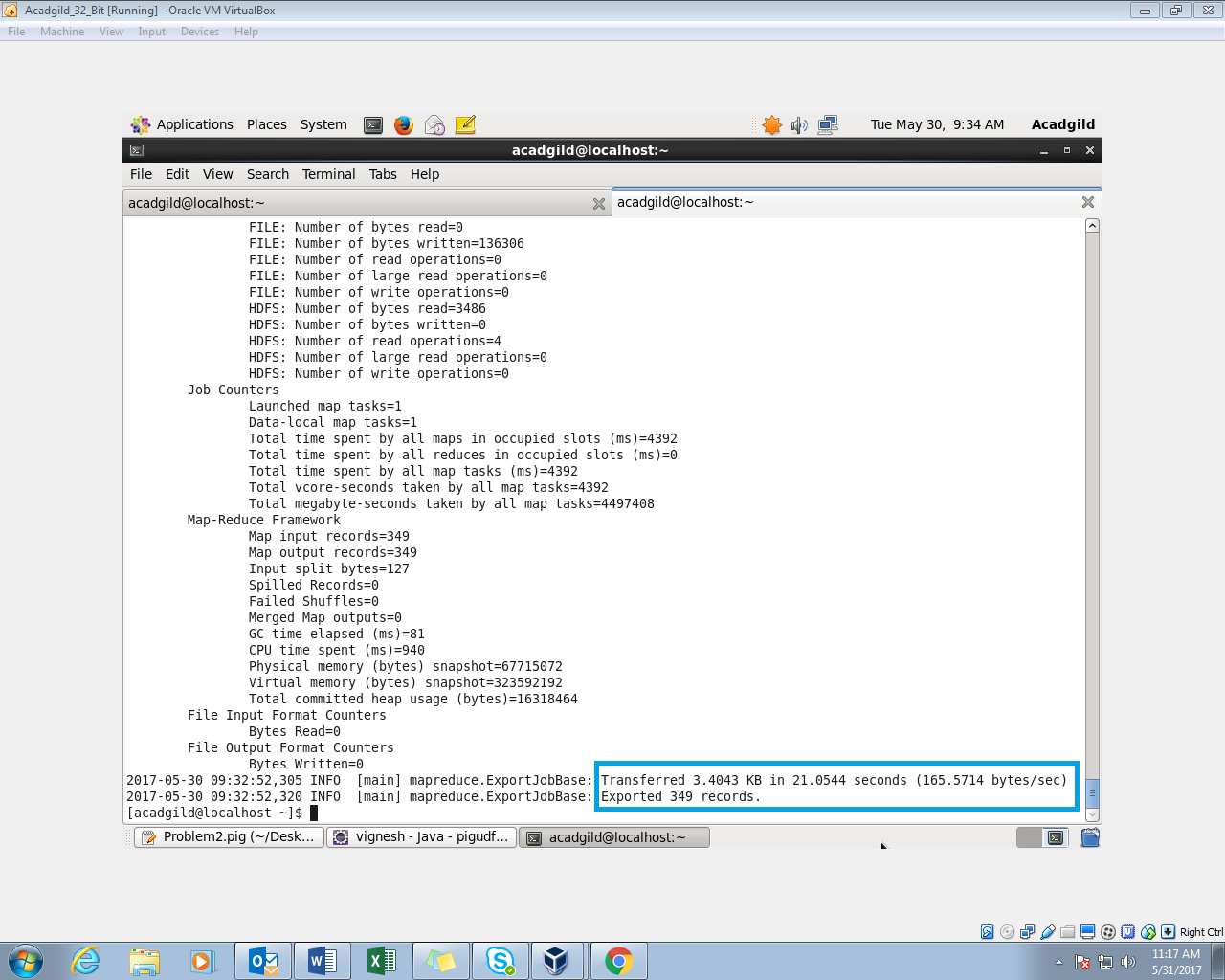
I am creating an empty table named blp2 in proj directory.

Sqoop Command to export the results to the mysql:



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

Successfully exported the results to mysql:



The Contents of the table is displayed:

