**MAIN 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>

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.

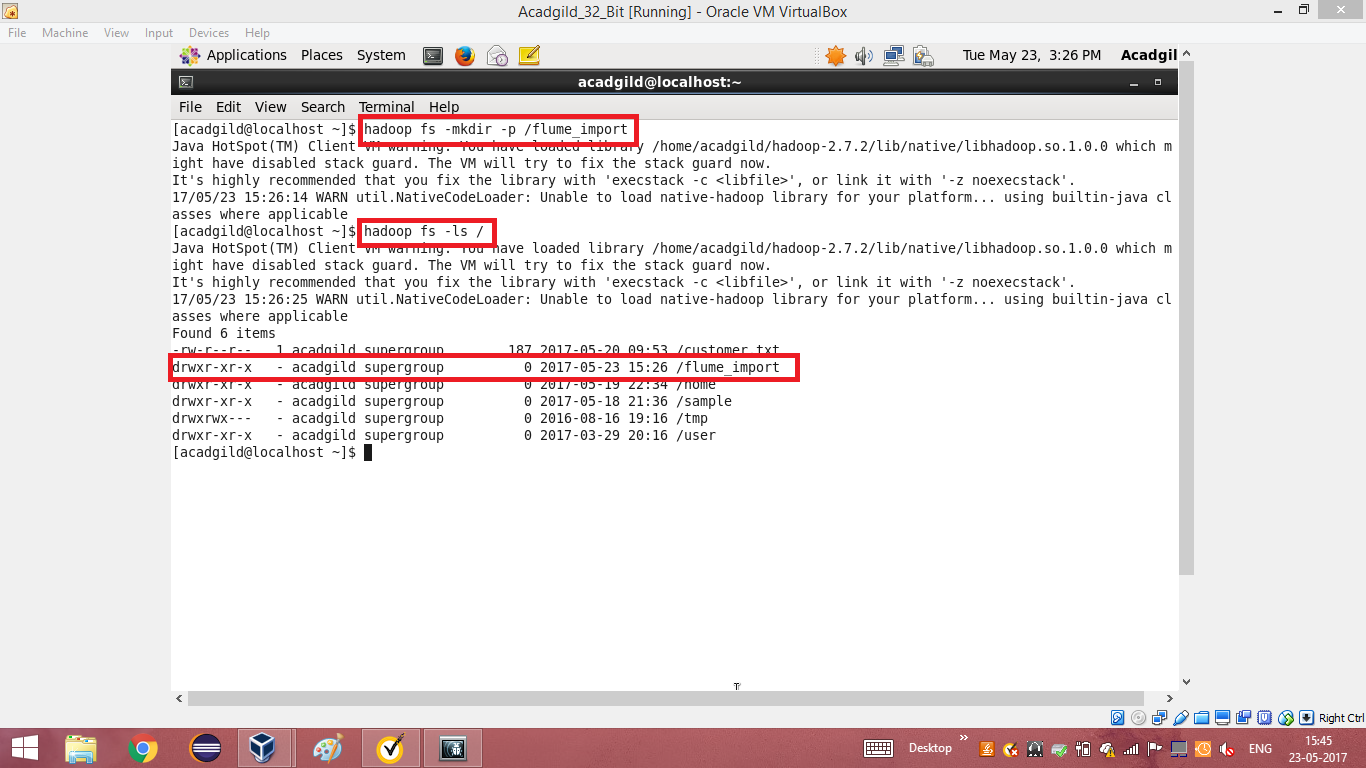
Export the results to MySQL using Sqoop.

**DATA INGESTION INTO HDFS USING FLUME:**

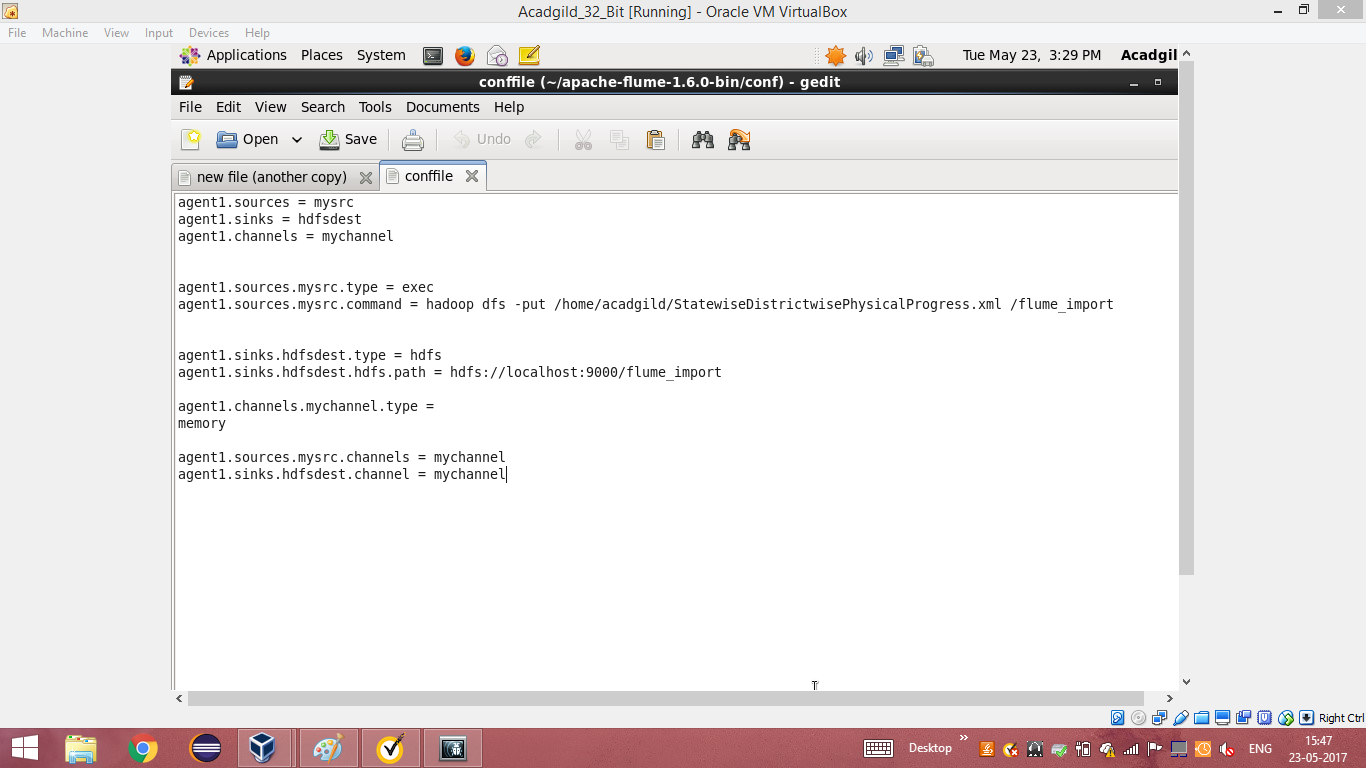
Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.

Flume is used to ingest data into the hdfs at very high speed.

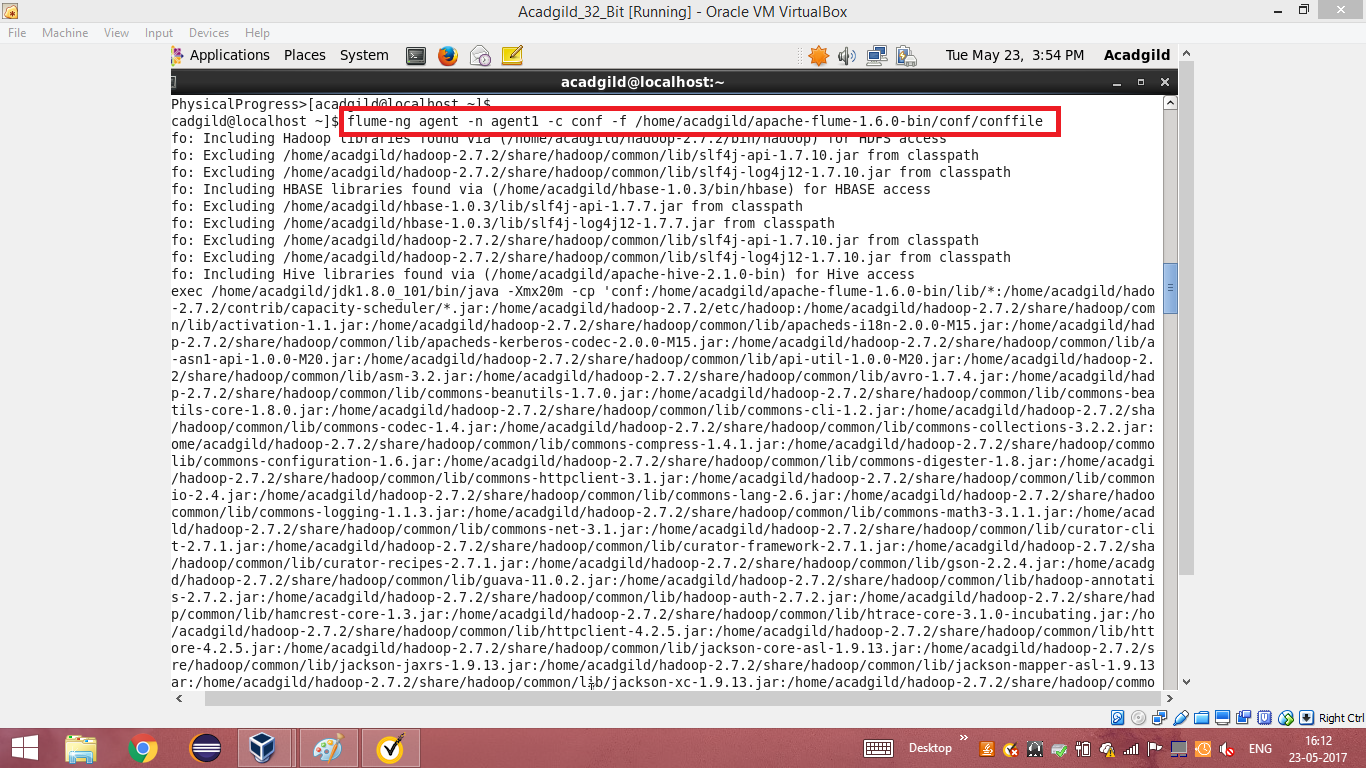
STEP 1: CREATING A DIRECTORY TO STORE THE DATASET IN HDFS



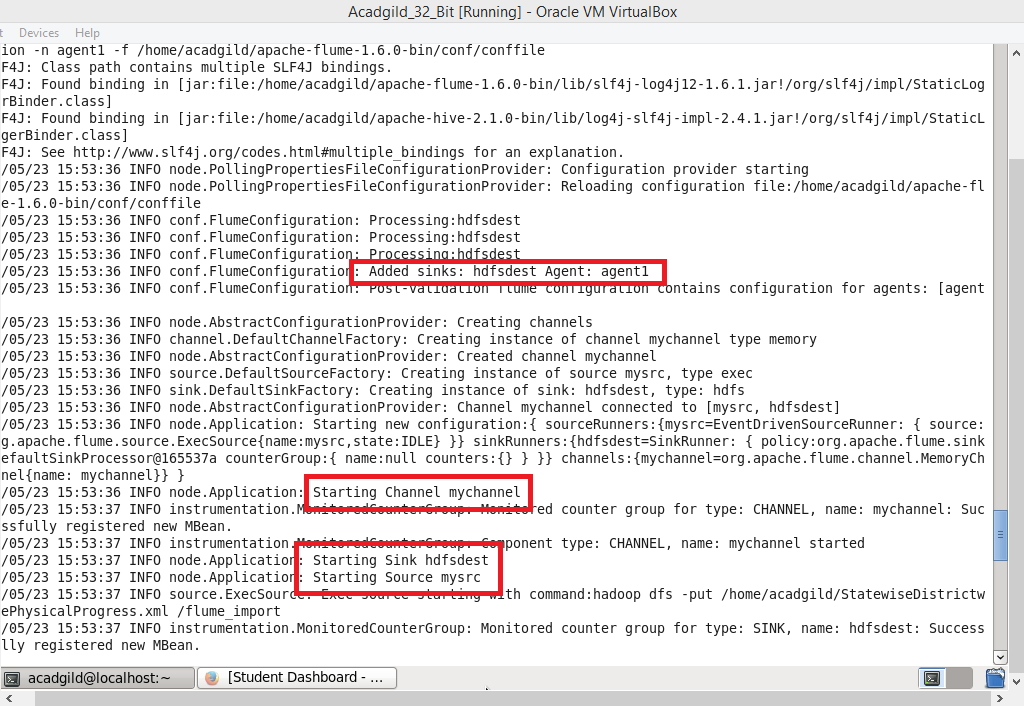
STEP 2: COPYING AND PASTING THE CONFIGURATION FILE INSIDE THE LIB FOLDER OF FLUME DIRECTORY.



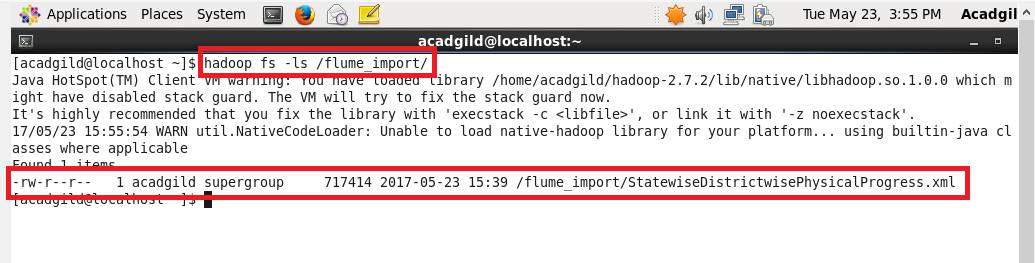
STEP 3: USING THE FLUME COMMAND TO INGEST THE DATASET( IN XML FORMAT) INTO THE HDFS.



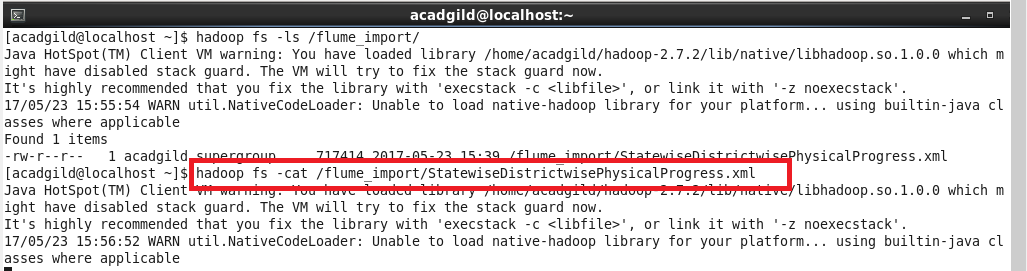
STEP 4: THE COMMAND HAS STARTED THE SINK, SOURCE AND THE CHANNEL



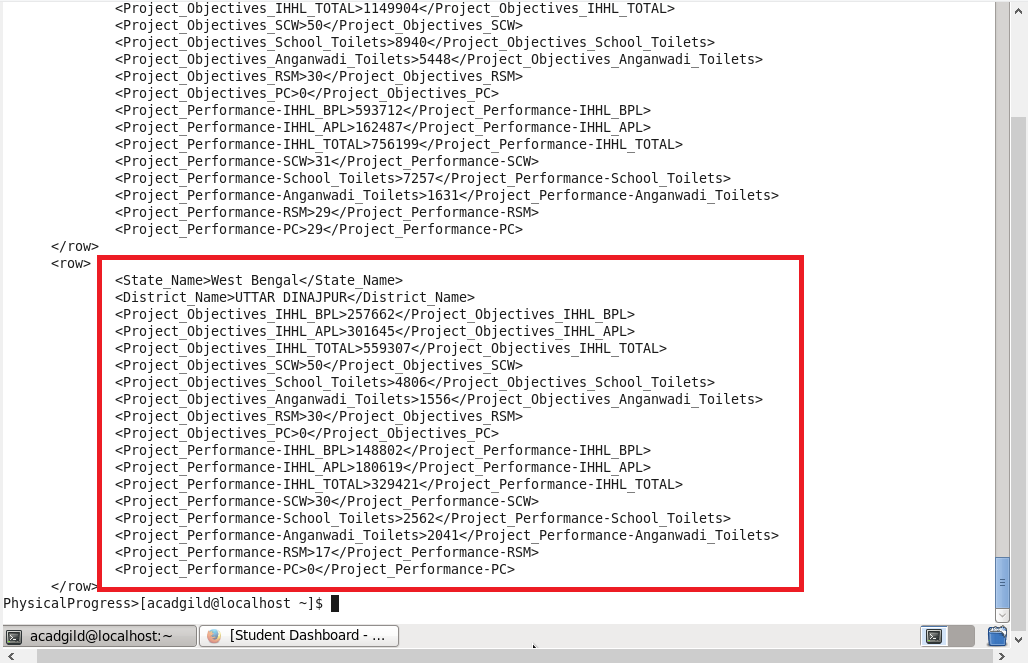
STEP 5: AFTER RUNNING THE FLUME COMMAND, THE DATA GETS STORED IN THE DESTINED PATH INSIDE THE HDFS



STEP 6: READING THE DATASET USING CAT COMMAND JUST TO CHECK WHETHER THE DATA HAS BEEN PROPERLY INGESTED.



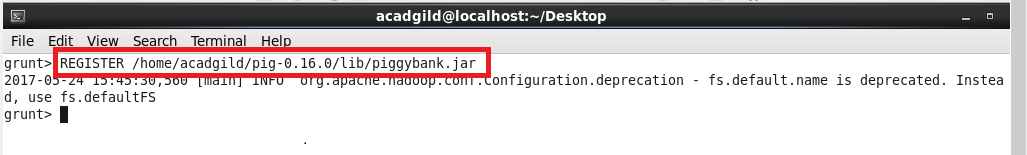
STEP 7: THIS ENSURES THAT THE DATA HAS BEEN PROPERLY INGESTED INTO THE HDFS.



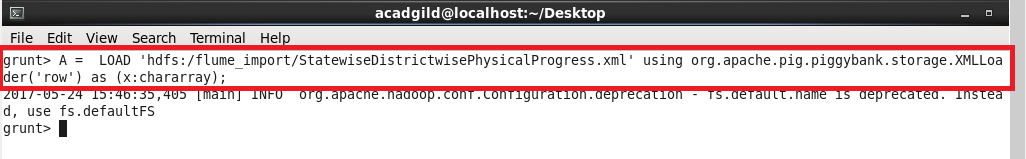
**PARSING THE DATA FROM XML FORMAT USING PIG:**

**I am running the pig with map reduce mode.**

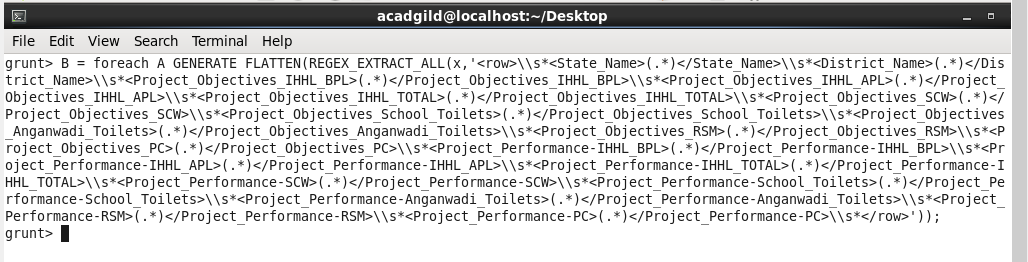
STEP 1: I AM USING THE XMLLOADER() IN PIGGY BANK UDF TO LOAD THE XML HENCE REGISTERING THE PIGGYBANK UDF WITH GRUNT SHELL.



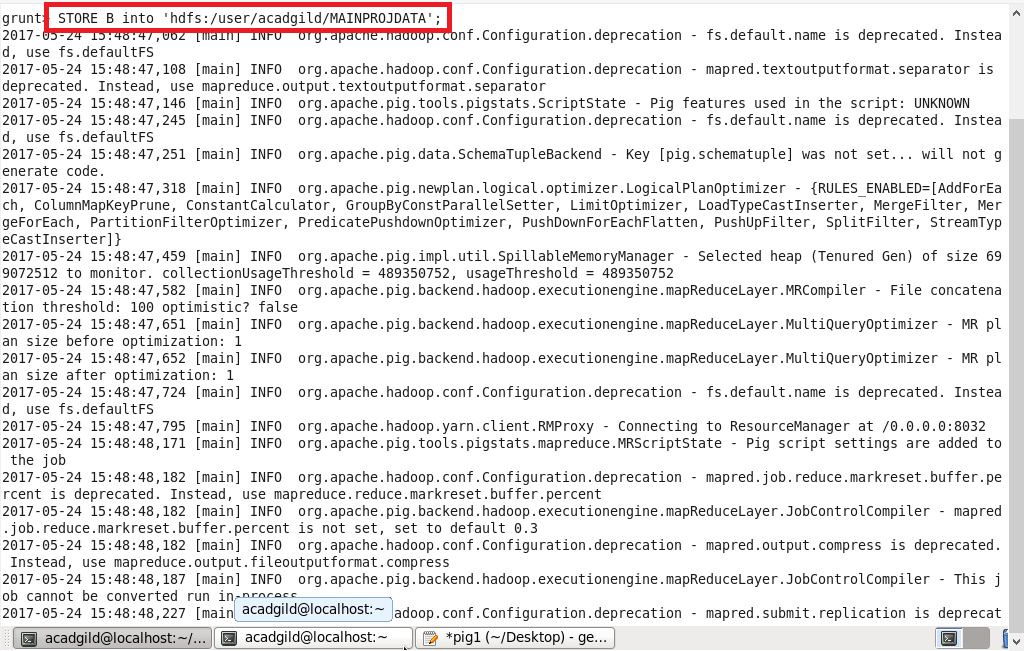
STEP 2: LOADING THE DATA IN XML FORMAT INTO THE RELATION A



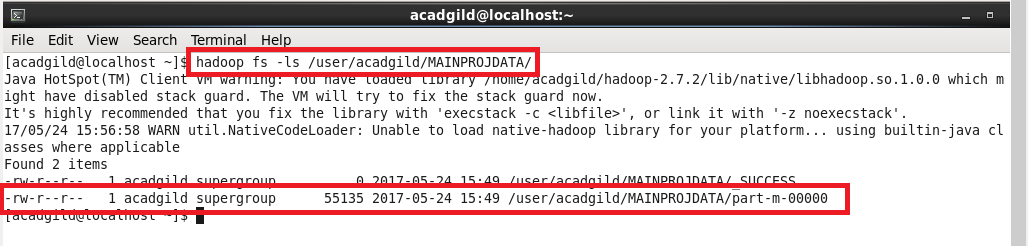
STEP 3: TO ELIMINATE THE TAGS, I AM USING REGEX\_EXTRACT\_ALL COMMAND. THIS COMMAND FETCHES ONLY THE RECORDS PRESENT IN BETWEEN THE TAGS. I’VE USED FLATTEN COMMAND TO UNNEST THE BAGS.



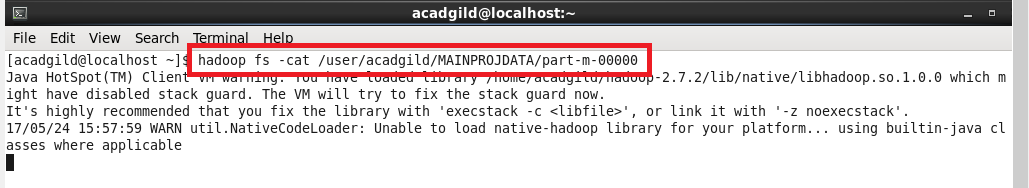
STEP 4: STORING THE FLATTENED FILE INTO HDFS USING STORE COMMAND.



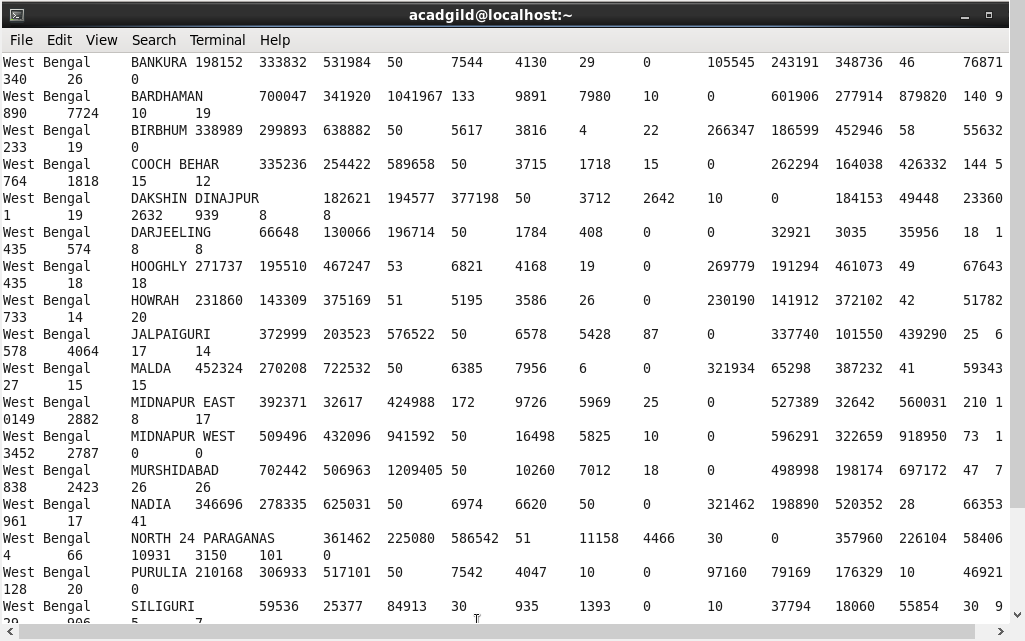
STEP 5: LISTING THE CONTENTS OF “MAINPROJDATA” USING “ls” COMMAND



STEP 6: READING THE PART-M-00000FILE USING CAT COMMAND JUST TO CHECK WHETHER THE PREVIOUS STEPS HAS BEEN DONE PROPERLY



STEP 7: THIS ENSURES THAT THE PREVIOUS STEPS WERE RIGHT AND WE GOT THE CORRECT FORMAT.

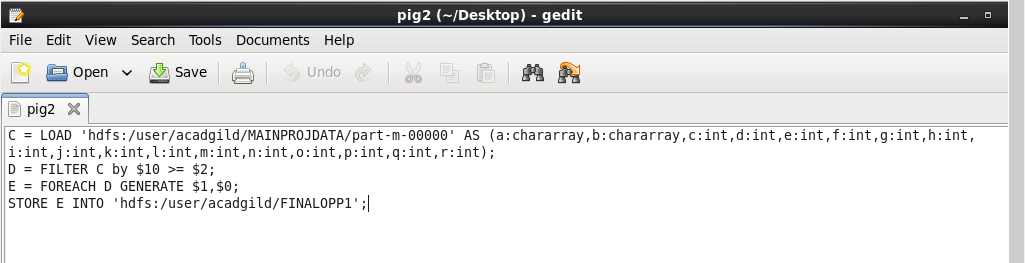


1. Find out the districts who achieved 100 percent objective in BPL cards.

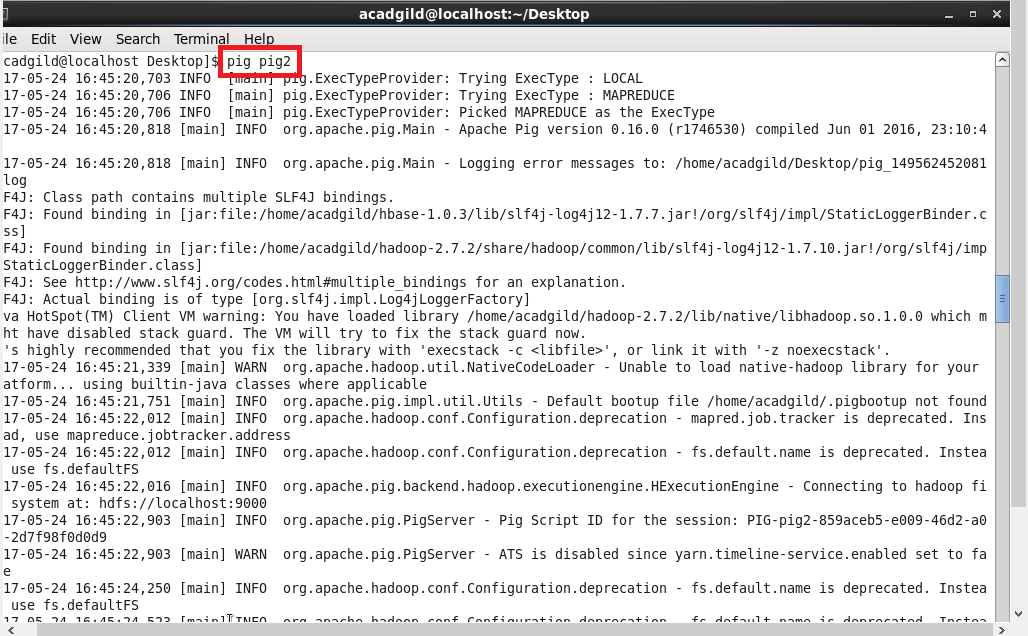
Export the results to mysql using sqoop.

STEP 1:

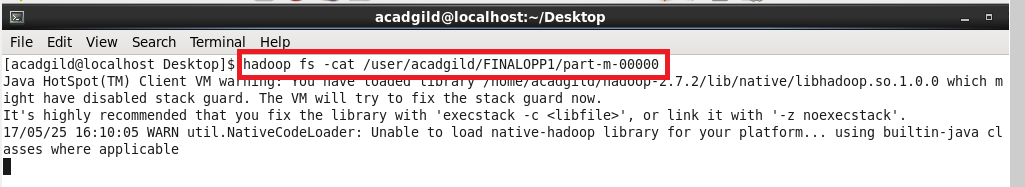
* Loading the dataset into relation C by specifying datatype with the columns.
* Filtering the dataset by specifying condition ( performance($10) >= objective ($2) )
* As we want only district, we are generating only districts and states by specifying the column $1 and $0.
* Storing the output in hdfs by specifying the path



STEP 2: RUNNING THE PIG SCRIPT IN MAPREDUCE MODE



STEP 3: READING THE OUTPUT USING CAT COMMAND. THIS IS JUST TO CHECK THE OUTPUT. EXPORTING STEP IS THE NEXT STEP.



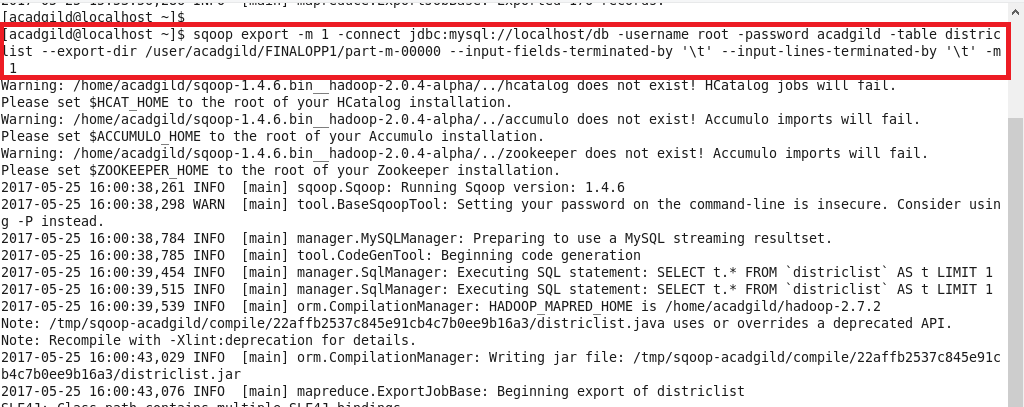
STEP 4: SAMPLE OUTPUT



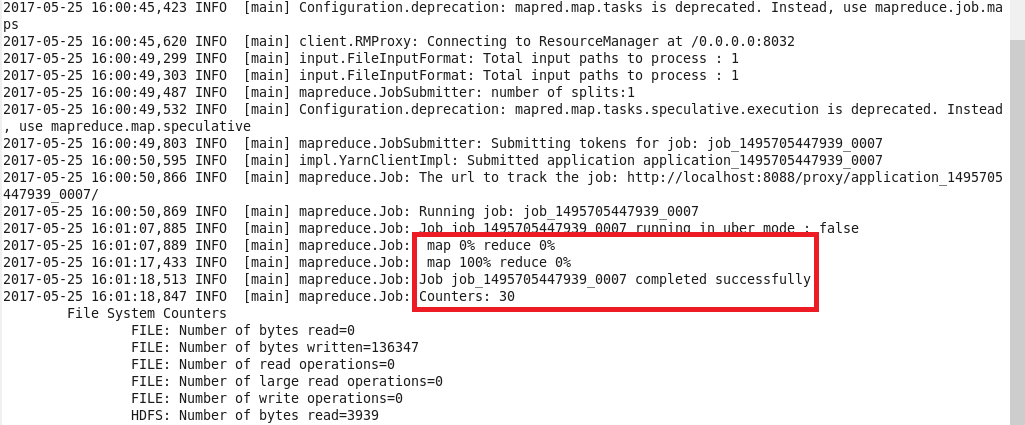
STEP 5: USING DB DATABASE AND CREATING A TABLE ‘DISTRICLIST’ IN MYSQL. USING SELECT COMMAND TO SHOW THAT THE TABLE IS EMPTY



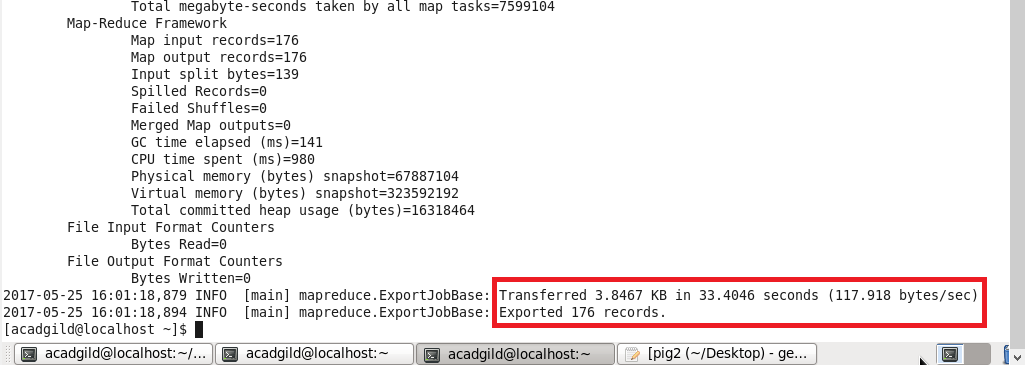
STEP 6: RUNNING SQOOP COMMAND BY SPECIFYING THE HDFS PATH , DATABASE NAME AND THE TABLE NAME.



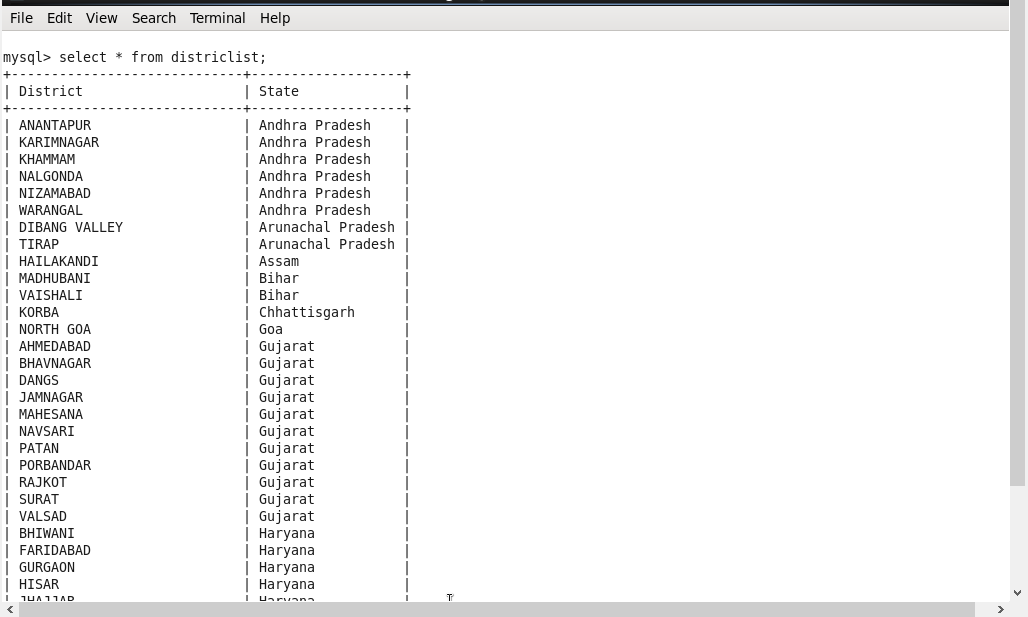
Map reduce framework is running



Successfully exported the output from hdfs to mysql



SAMPLE OUTPUT:

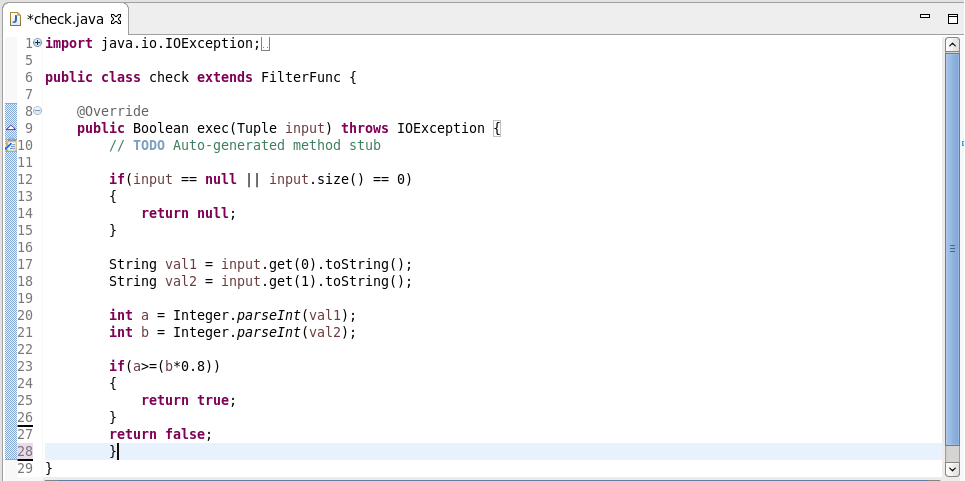


**I’ve attached the output with this document. Please check that.**

Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop.

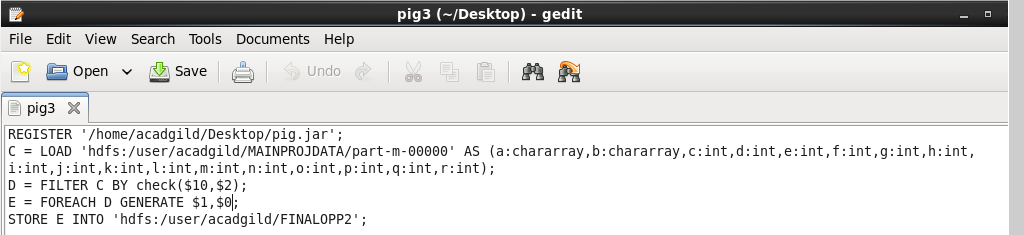
STEP 1: WRITING USER DEFINED FUNCTION TO FILTER THE DATASET.

FILTERING THE DATASET BY SPECIFYING CONDITION ( PERFORMANCE >= 0.8 \* OBJECTIVE )

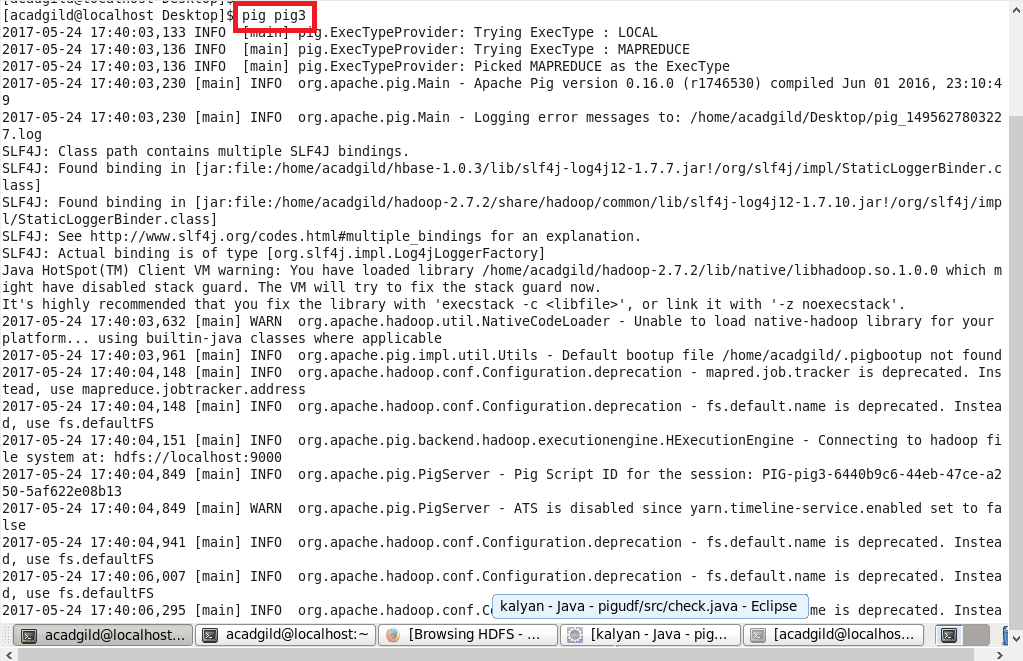


STEP 2:

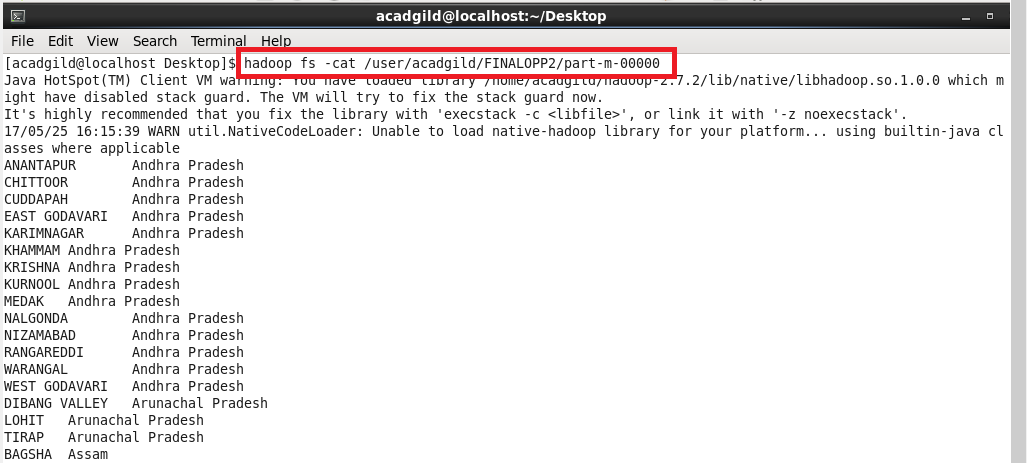
* Loading the dataset into relation C by specifying datatype with the columns.
* check function filters the dataset as we have already specified condition in the udf
* As we want only district and state, we are generating only districts by specifying the columns $1 and $0.
* Storing the output in hdfs by specifying the path.



STEP 3: RUNNING THE PIG SCRIPT IN MAPREDUCE MODE



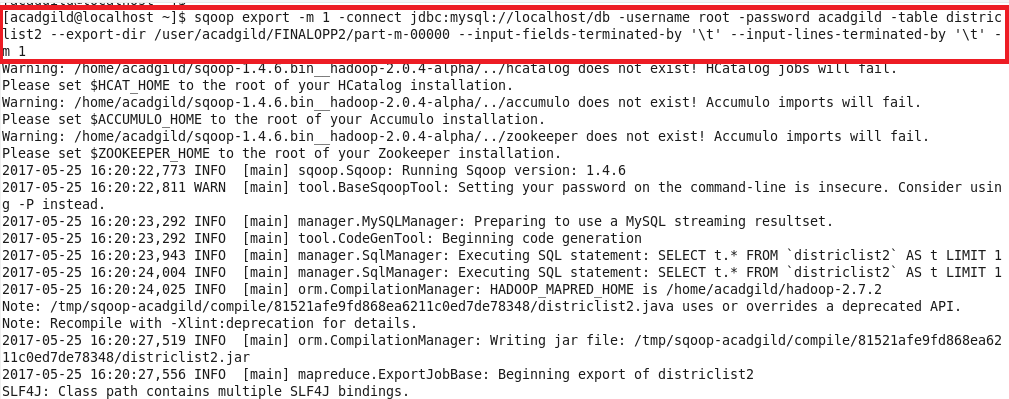
STEP 4: READING THE CONTENTS OF THE OUTPUT JUST TO CHECK WHETHER THE CODE IS CORRECT



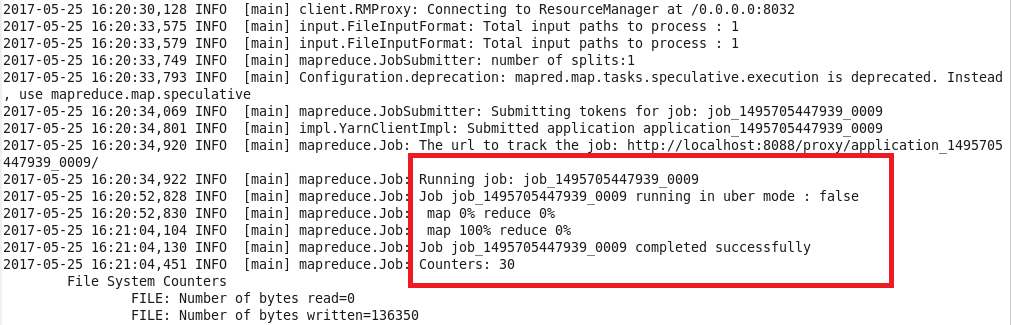
STEP 5: USING DB DATABASE AND CREATING A TABLE ‘DISTRICLIST2’ IN MYSQL. USING SELECT COMMAND TO SHOW THAT THE TABLE IS EMPTY



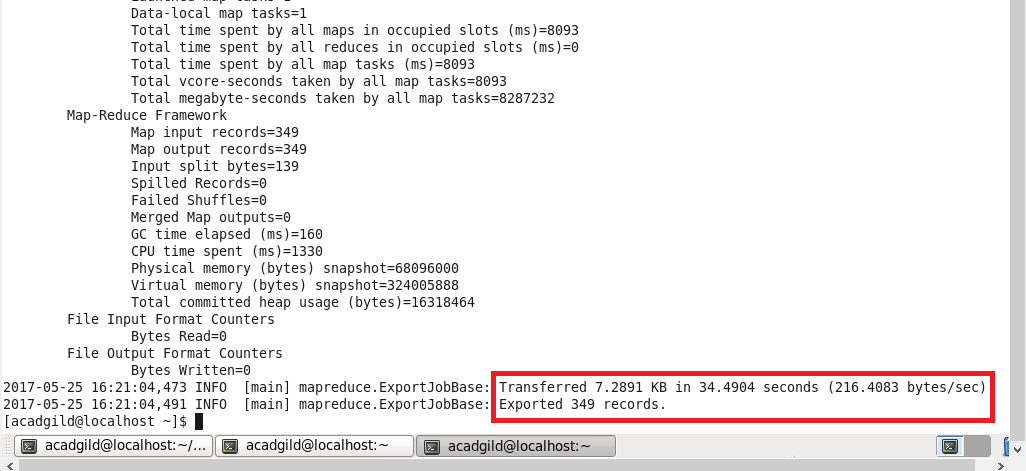
STEP 6: RUNNING SQOOP COMMAND BY SPECIFYING THE HDFS PATH , DATABASE NAME AND THE TABLE NAME.



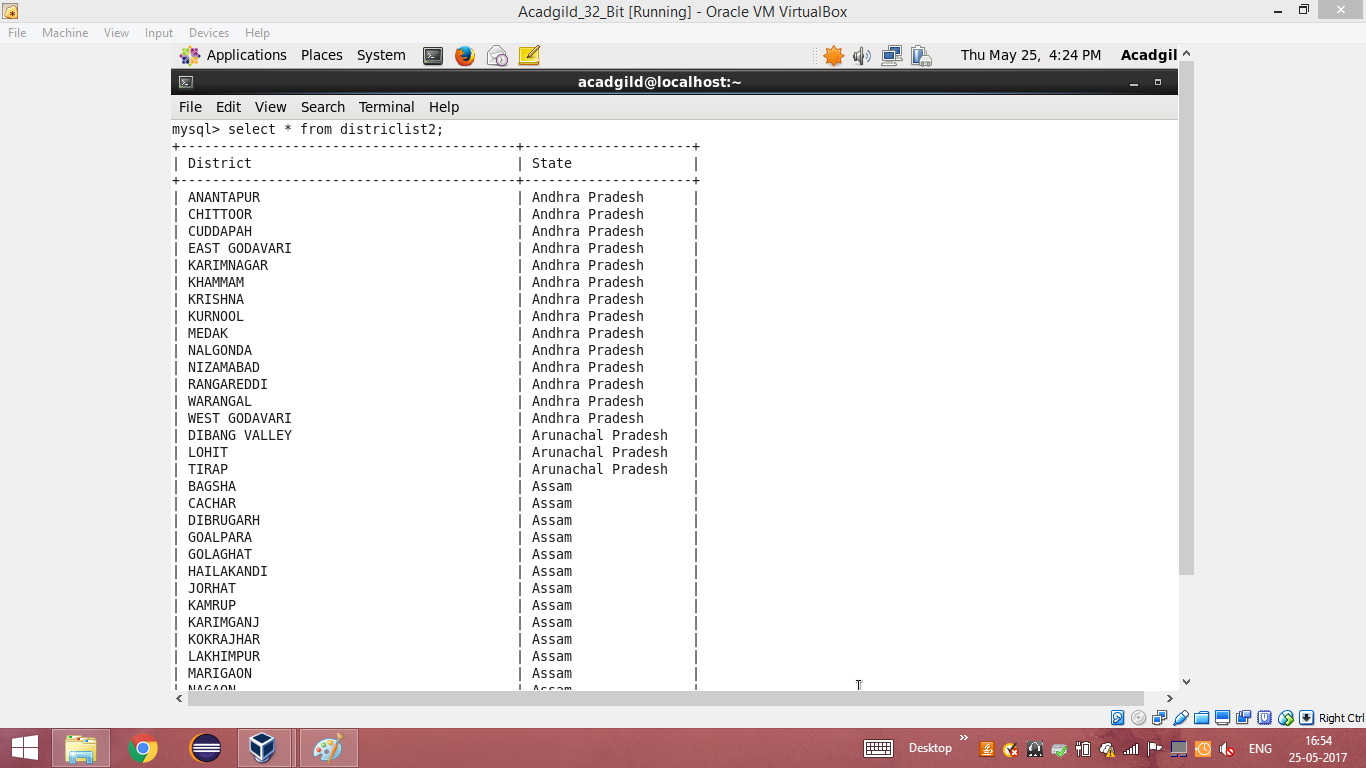
Map reduce framework is running



Successfully exported the output from hdfs to mysql



STEP 7: SAMPLE OUTPUT



**I’ve attached the output with this document. Please check that.**