Consumer Complaints

Hadoop can be applied with many projects of various domains: Data consolidation Machine learning Specialized analysis (domain specific) This project comes under Data consolidation

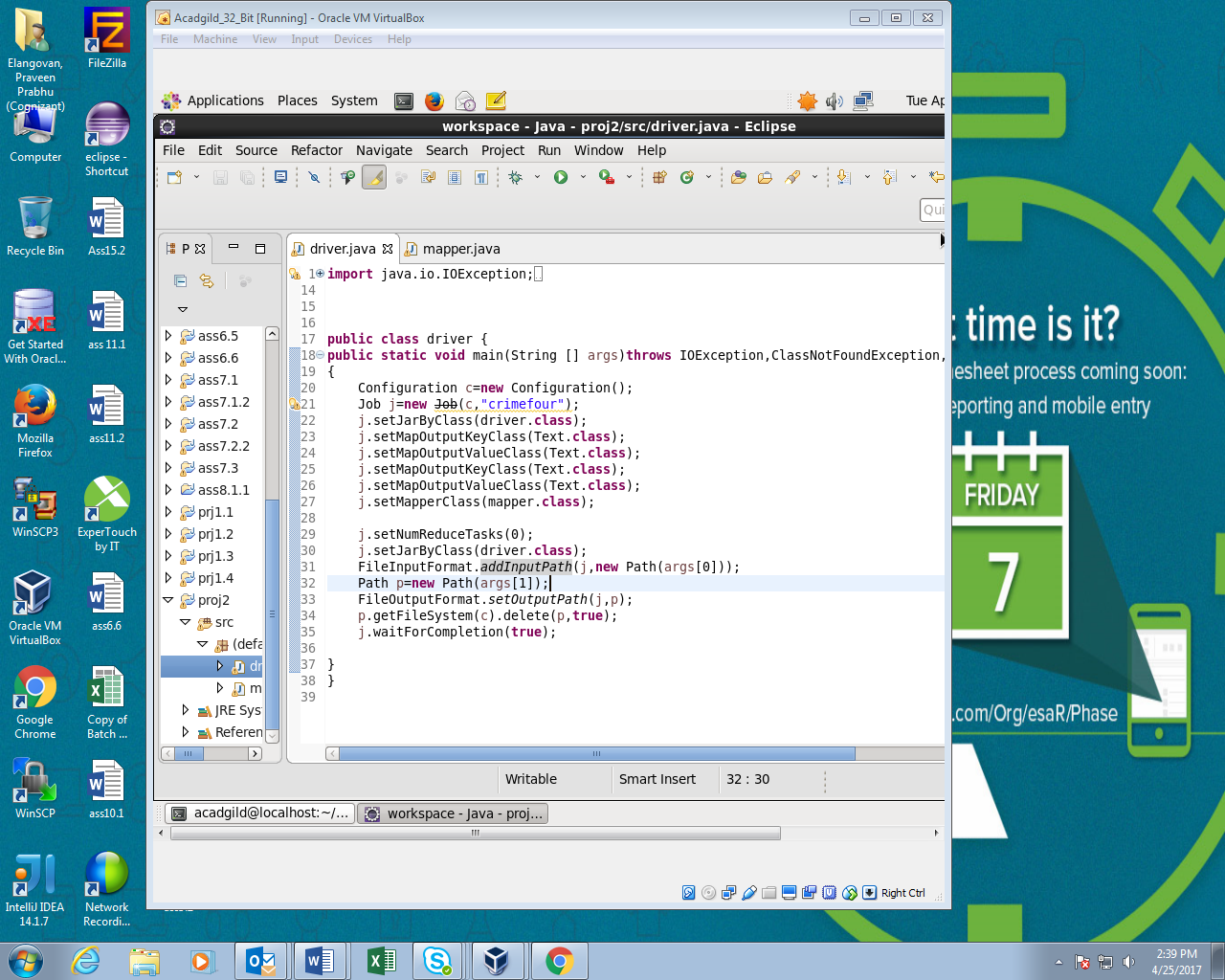
Data consolidation This is like "enterprise data hub" or "data lake." The idea is you have disparate data sources, and you want to perform analysis across them. This type of project consists of getting feeds from all the sources (either real time or as a batch) and transferring them into Hadoop. Sometimes this is the first step for becoming a “data-driven company”. Data lakes usually materialize as files on HDFS and tables in pig/Hive or Impala and then we can create simply pretty reports on the data contained in hdfs.

The dataset consists of consumer complaints details. In the process consumer files the complaints to consumer forum and then consumer forum forwards the complaint to respective company.

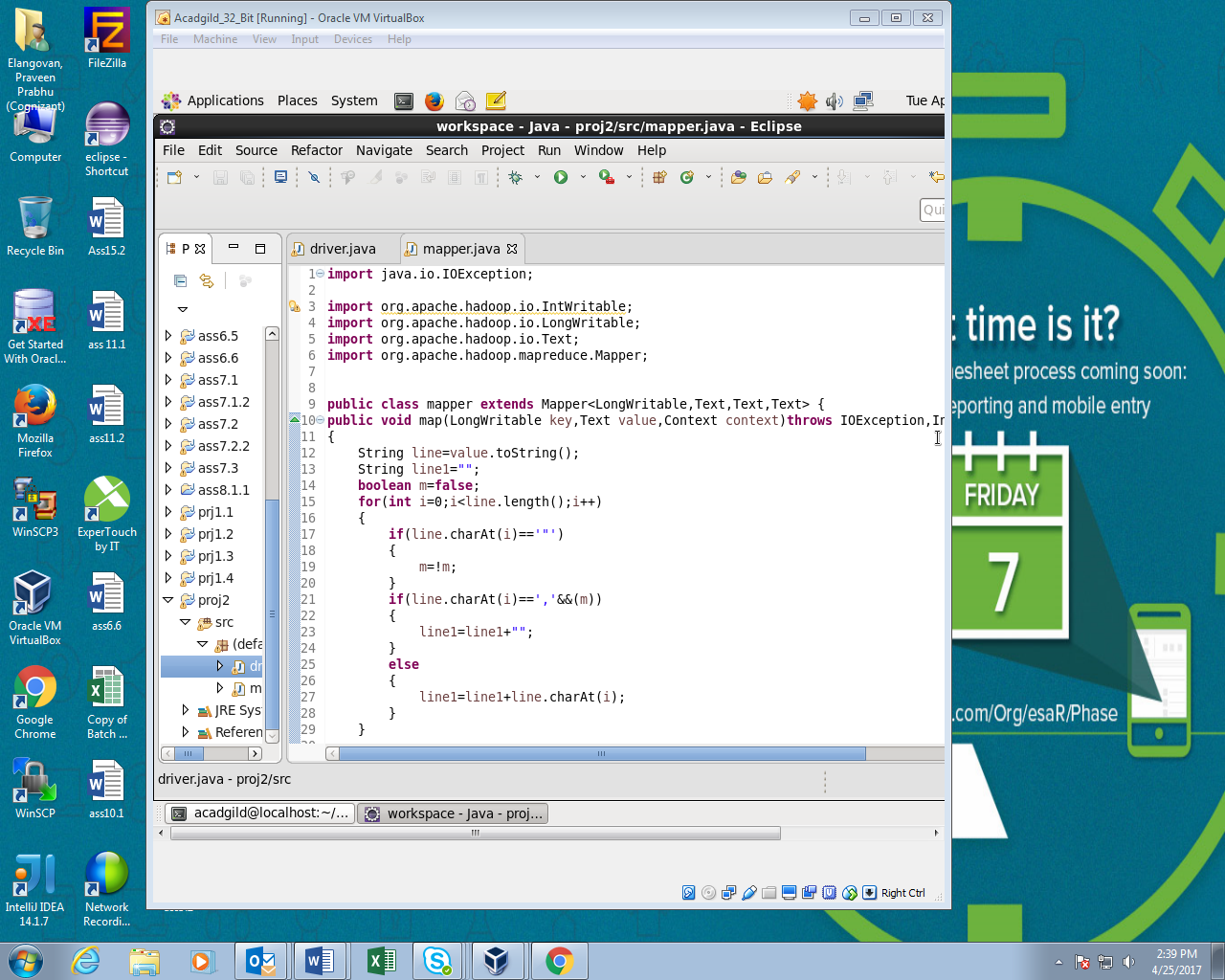
Sample record: 10/16/2015,Debt collection,"Other (phone, health club, etc.)",Cont'd attempts collect debt not owed,Debt was discharged in bankruptcy,,,"Convergent Resources, Inc.",OH,438XX,Web,10/16/2015,Closed with explanation,Yes,,1612132 This entire column "Other (phone, health club, etc.)" Should be product but, if we split this file based on comma then this column will be splitted into 3 columns which will result in wrong outputs. In order to tackle this we should remove the commas present only inside double quotes.

1. Write a mapreduce program to remove commas present inside the double quotes.

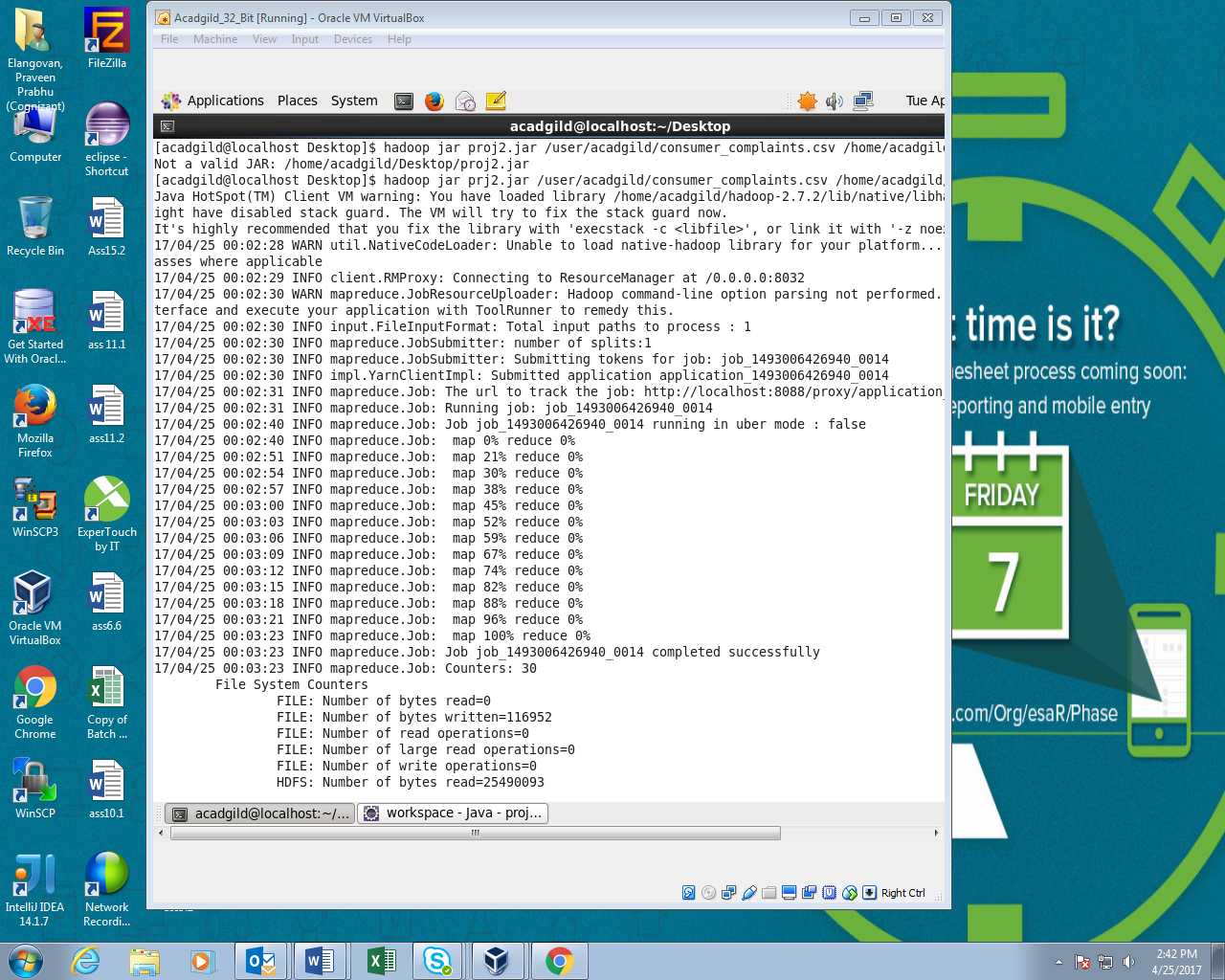
Driver:



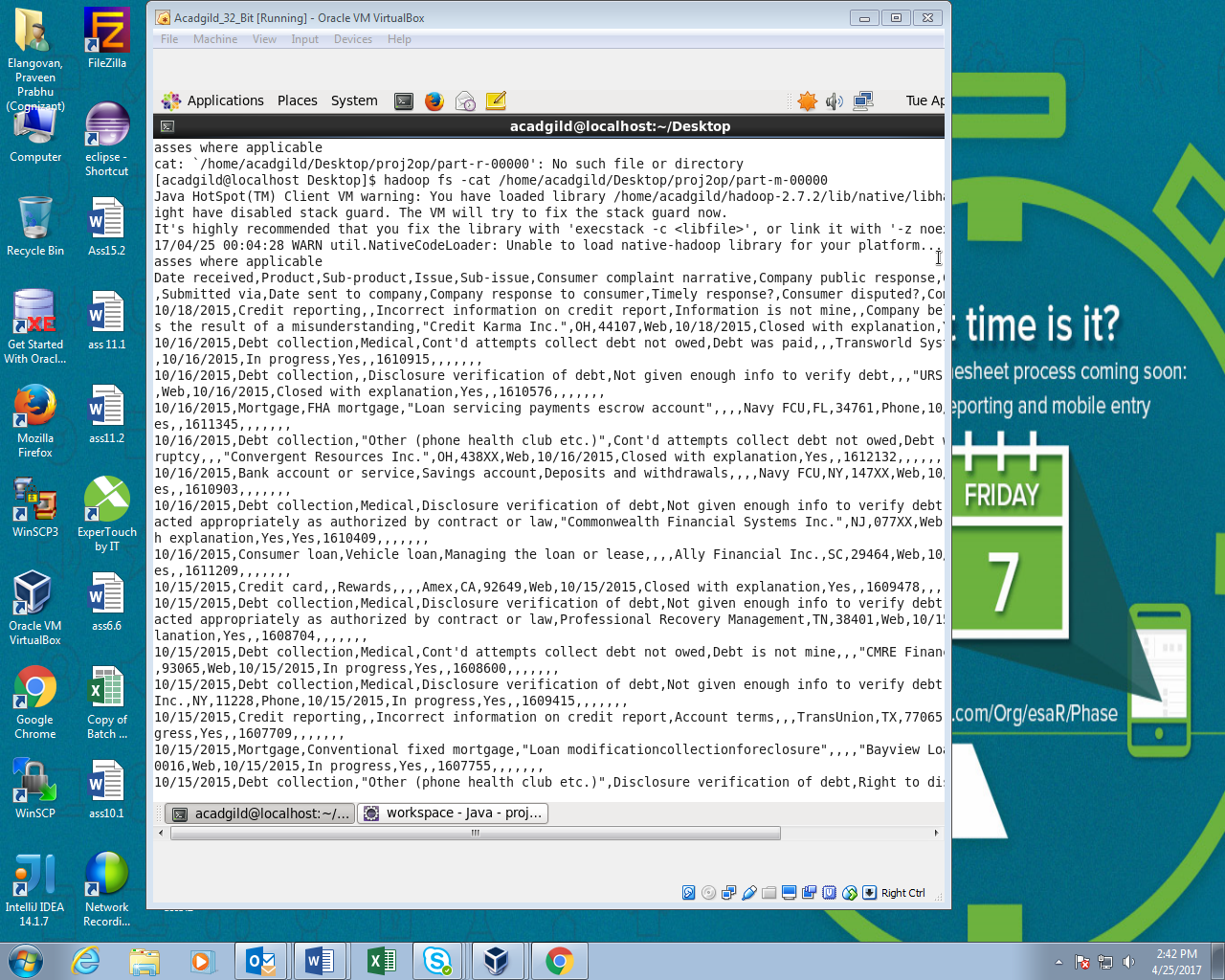
Mapper:



Running JAR file;



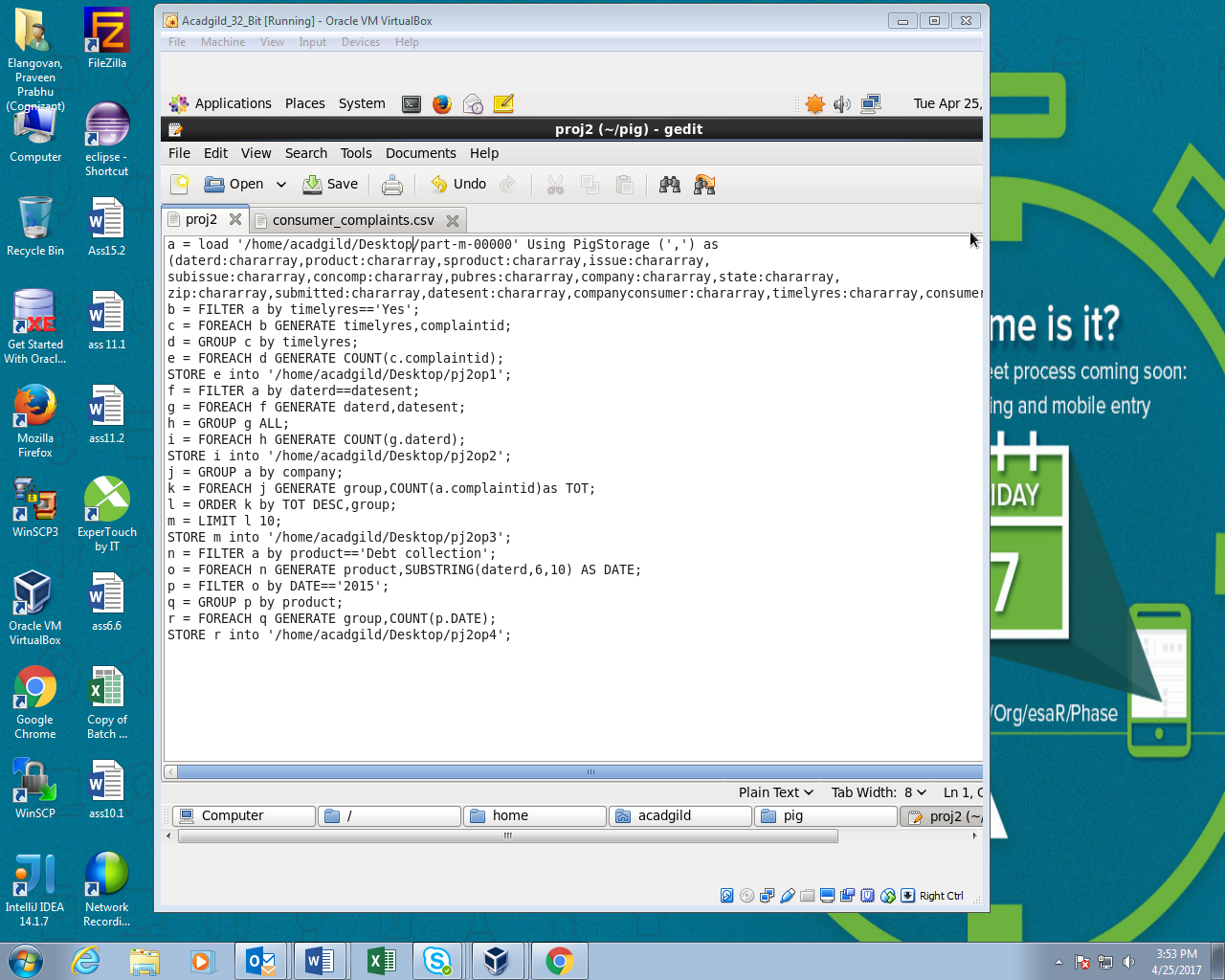
Output:



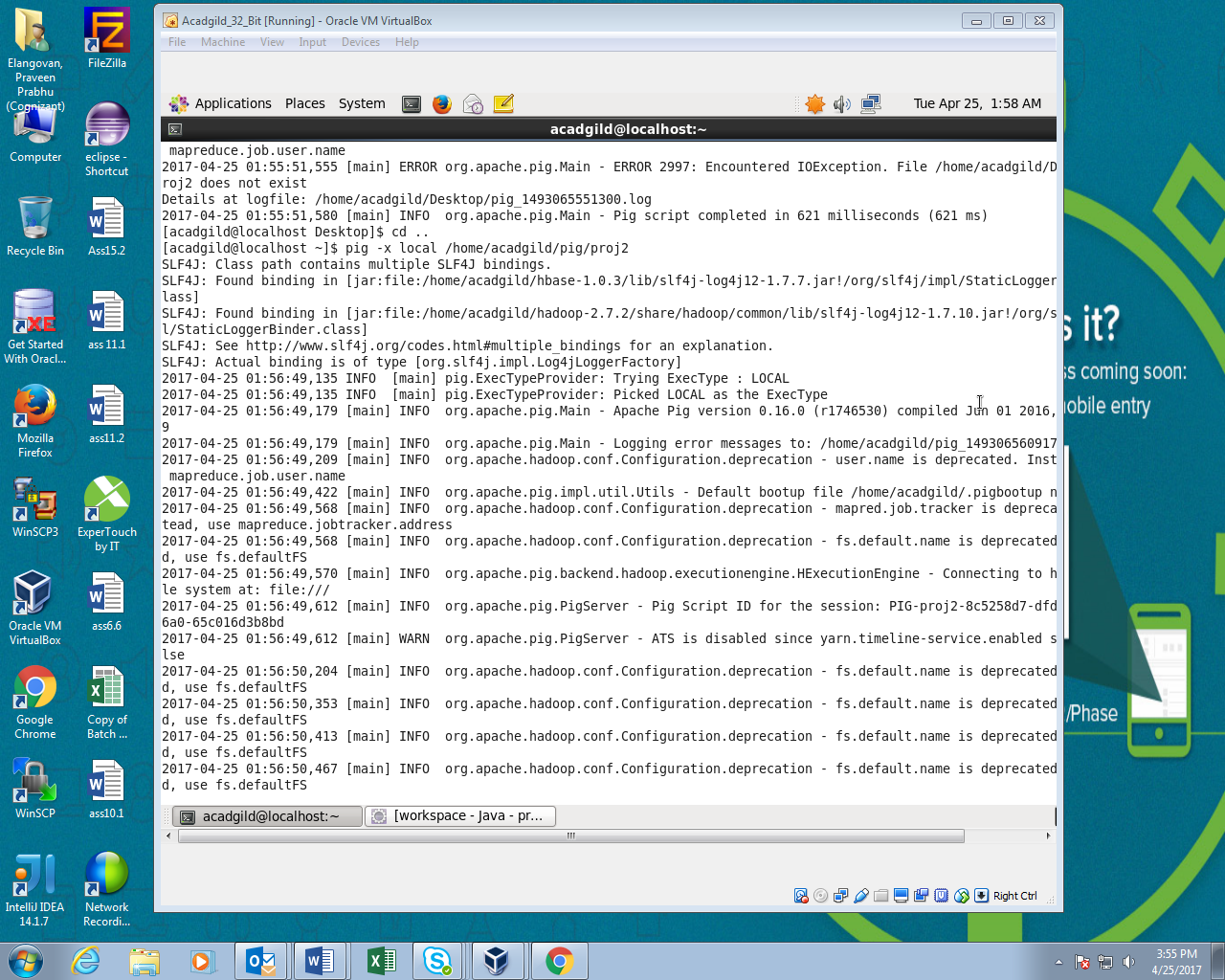
**USING PIG :**

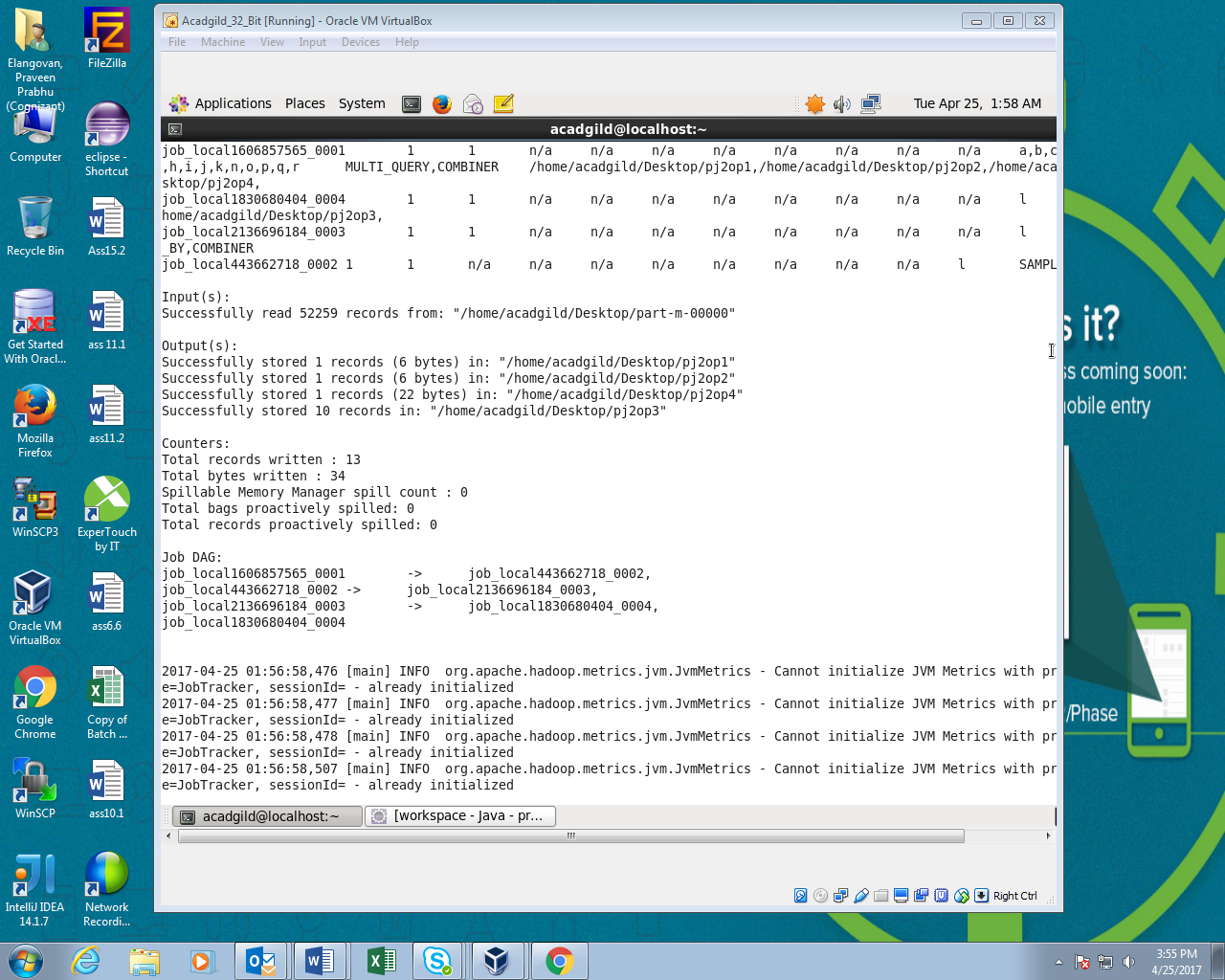
1. Write a pig script to find no of complaints which got timely response
2. Write a pig script to find no of complaints where consumer forum forwarded the complaint same day they received to respective company
3. Write a pig script to find list of companies topping in complaint chart (companies with maximum number of complaints)
4. Write a pig script to find no of complaints filed with product type has "Debt collection" for the year 2015

PIG command:



Running PIG script:





Output:

