Sentiment Analysis of Social Media Data

Using Hadoop

**Project Report**

**On**

**“Sentiment Analysis of Social Media Data Using Hadoop”**

****

**Submitted By: Project Mentor:**

1. Aabhas Sharma Manoj kumar Nama
2. Amit Kumar Singh Dept. of Computer Engineering

**Acknowledgement**

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

We are grateful to our project guide **Mr. Saurabh Ranjan Srivastav** the guidance, inspiration and constructive suggestions that helpful us in the preparation of this project.

We also thank our colleagues who have helped in successful completion of the project.

**Certificate**

This is to certify that this report of **Sentiment Analysis of Social Media Data Using Hadoop** embodies the original work done by **Aabhas Sharma, Amit Kumar Singh** this project submission as a minor project of the requirement for the Sentiment Analysis of Social Media Data using Hadoop Project of VII Semester, of the Computer Science branch.

Manoj Kumar Nama

(Reader)

Dept. of Computer Science

**Table of Contents**

ABSTRACT 5

**1. INTRODUCTION**

1.1 OPERATION ENVIRONMENT 6

**2. HADOOP ECOSYSTEM RELATED WORK**

2.1 HADOOP 7

2.2 HDFS 7

2.3 CLOUDERA’S SANDBOX 8

**3. PROPOSED WORK**

3.1 CLUSTER SETUP 9 3.2 DATA PRE-PROCESSING

A. DATA GATHERING 10 B. DATA CLEANING 10

C. LOADING DATA INTO HDFS 11 **4. SIMULATION**

4.1 transfer file from windows

to sandbox 12

4.2 remove ambiguities form the .txt

file using ‘sed’ command 13

4.3 Upload file into HDFS 13

* 1. Create table using hive/impala 14
  2. Hive v/s Impala query execution time 14
  3. To see the loaded table in Impala 15

4.7 query execution 15

**5. CONCLUSION & FUTURE SCOPE** 20

**6. REFERENCES** 21

**ABSTRACT**

Sentiment analysis of social media data using hadoop is a project which aims in performing sentiment analysis of the data which is available on the social media with the help of Hadoop. Social media site receives data in millions every day. This huge amount of data can be used for industrial or business purposes by organizing according to our requirement and processing. This management of data can positively impact different types of crucial objectives related to strategic management, governmental policies and reforms, product brand and retail management, customer relationship management, operational optimization etc. This project has many features like it will process the huge amount of data (in gigabytes) on a hadoop cluster in real times and with the help of Solr which is highly reliable, scalable and fault tolerant platform we can draw bar graph, pie chart for scaling our results. The obtained results may be helpful to the organization for future prediction.

1. **INTRODUCTION**

Today, the textual data on the internet is growing at a rapid pace. Different industries are trying to use this huge textual data for extracting the people’s views towards their products. Social media is a vital source of information in this case. It is impossible to manually analyze the large amount of data. This is where the need of automatic categorization becomes apparent (2). Subjective data is analyzed generally in this case. There are a large number of social media websites that enable users to contribute, modify and grade the content. Users have an opportunity to express their personal opinions about specific topics. The example of such websites include blogs, forums, product reviews sites, and social networks. An interesting application of Twitter analysis is focused on politics, e.g., gauging the sentiments of users regarding the current government’s policies, or the most favorable candidate (or political party) to win an upcoming election. The focus of our project is on the election data tweets of the United Kingdom held in the year 2010, 2011, 2012 &2014.

* 1. **OPERATION ENVIRONMENT**

|  |  |
| --- | --- |
| **PROCESSOR** | **INTEL CORE PROCESSOR FOR BETTER**  **PERFORMANCE** |
| **OPERATING SYSTEM** | **WINDOWS 8, WINDOWS 7** |
| **MEMORY** | **8 GB OR MORE** |
| **HARD DISK SPACE** | **MINIMUM 50GB FOR DATABASE USAGE** |
| **QUERY EXECUTION** | **HIVE SCRIPT, IMPALA** |
| **OTHER SOFTWARES** | **VMWARE, PUTTY,EVERNOTE** |

1. **Hadoop Eco-system Related Work**

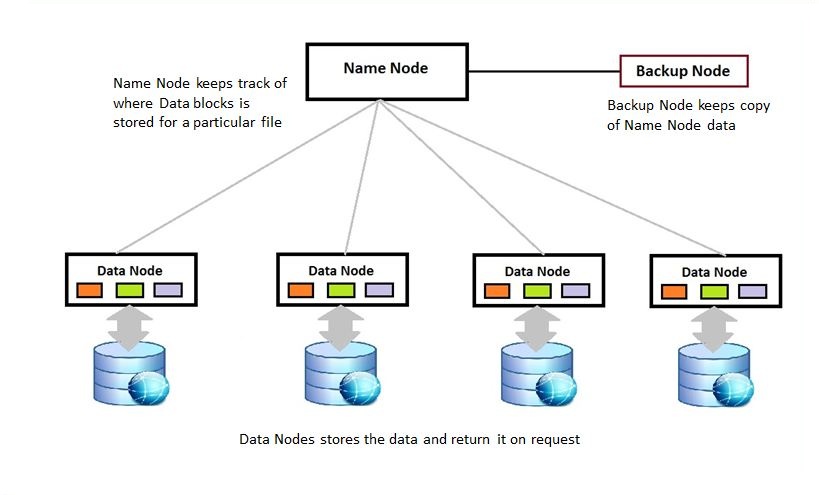
Sentiment analysis is most popular tend in today’s world. Lot of work has been done in this sector. Following are some approaches which are most popular in today’s world. There has been a lot of research in the area of Sentiment analysis. Bo Pang and Lee were the pioneers in this field. Current works in this area includes using a mathematical approach which uses a formula for the sentiment value depending on the proximity of the words with adjectives like ‘excellent’, ‘worse’, ‘bad’ etc.

Sentiment analysis has been done previously also in the 2008 US presidential elections. In that Topsy, which is a real time social network analytics provider, in a joint work with twitter, built the “Twitter Political Index” which presented a daily analysis of user’s sentiments for various topics related to elections. Later in 2012 US elections some modified approach was used which uses vender machines which are better than the old naïve Bayes.

In this paper, we are simply gauging the opinions of the users based on what data we have, although we already know that Twitter is commonly used across United Kingdom as in USA, Germany etc.

**2.1 Hadoop**

The Hadoop platform was designed to solve problems which had lot of data for processing. It uses the divide and rule methodology for processing. It is used to handle large and complex unstructured data which doesn’t fit into tables. Social media data being relatively unstructured can be best stored using Hadoop. Hadoop also finds a lot of applications in the field of online retailing, search engines, finance domain for risk analysis etc**.**



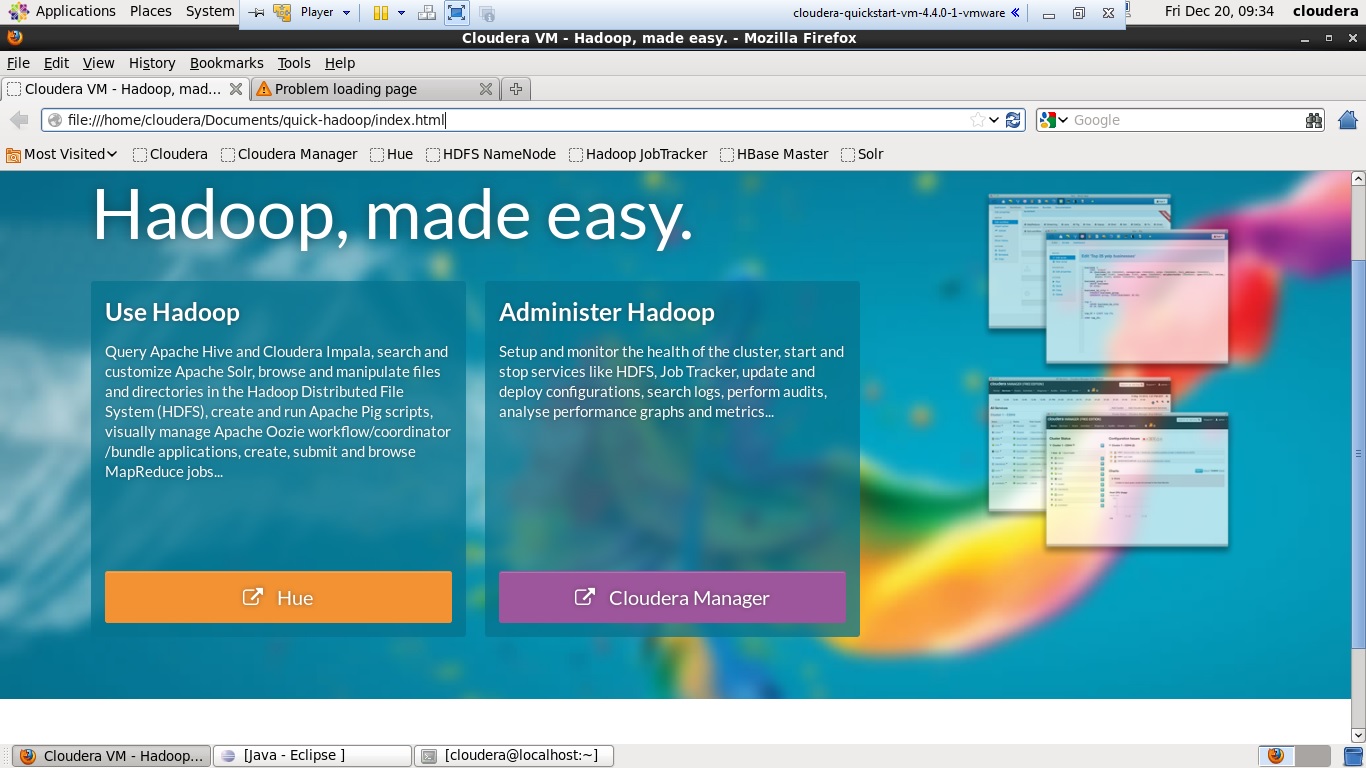
**2.2 HDFS**

Hadoop Distributed File System (HDFS) is a distributed file system which runs on commodity machines. It is highly fault tolerant and is designed for low cost machines. HDFS has a high throughput access to application and is suitable for applications with large amount of data. HDFS has a 1 master server architecture which has a single name node which regulates the file system access. Data nodes handle read and write requests from the file system’s clients. They also perform block creation, deletion, and replication upon instruction from the Name node. Replication of data in the file system adds to the data integrity and the robustness of the system.

Data replication is done for achieving fault tolerance. The large data cluster is stored as a sequence of blocks. Block size and the replication factor are configurable. Replication factor is set to 3 in our project which means 3 copies of the same data block will be maintained at a time in the cluster.

**2.3 Cloudera’s Sandbox**

Cloudera's open-source Apache Hadoop distribution, CDH (Cloudera Distribution Including Apache Hadoop), targets enterprise-class deployments of that technology. Cloudera says that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects (Apache Hive, Apache Avro, Apache HBase, and so on) that combine to form the Hadoop platform. Cloudera is also a sponsor of the Apache Software Foundation.



1. **Proposed Work**

**3.1 Cluster Setup**

In order to execute Cloudera Hadoop CDH4 VM in Windows, you need to download thequick start VM from here according to your VM version (i.e.VMware/Virtual Box/KVM). It requires a 64 bit host OS. This VM runs Centos 6.2 and includes CDH4.3, Cloudera Manager 4.6, Cloudera Impala 1.0.1 and Cloudera Search .9 Beta.

 We have used VMware version of Cloudera Quick start VM for running on Windows 8 64 bit host OS. Few points to ponder:

* This is a 64-bit VM, and requires a 64-bit host OS and a virtualization product that can support a64-bit guest OS.
* This VM uses 4 GB of total RAM. The total system memory required varies depending on the size of your data set and on the other processes that are running.
* The demo VM file is approximately 2 GB. Feel free to mirror internally or externally to minimize bandwidth usage.
* To use the VMware VM, you must use a player compatible with WorkStation 8.x or higher: Player 4.x or higher, ESXi 5.x or higher or Fusion 4.x or higher. Older versions of WorkStation can be used to create a new VM using the same virtual disk (VMDK file), but some features in VMware Tools won’t be available.
* After downloading the Cloudera VM, extract it & select the virtual machine configuration (.vmx) file.

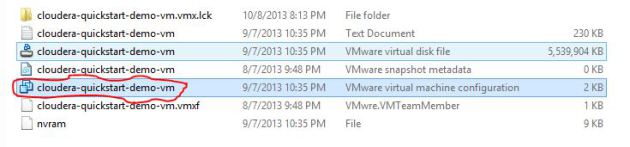


Fig.1 select vmx file.

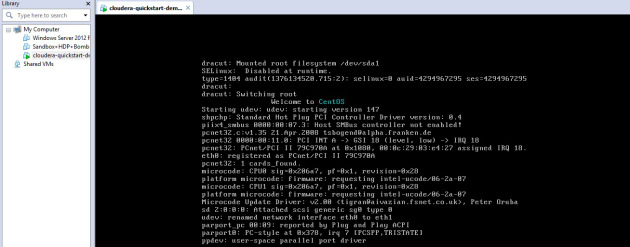


Fig.2 Open the .vmx file by VMware workstation & start the VM.

**3.2 Data Pre-processing**

The first step in big data analysis is data pre-processing   
 (data cleaning), which involves crucial steps for preparing the  
 data for analysis, e.g., data extraction, selection, reduction,  
 transformation etc. (3). Our data pre-processing steps are  
 described below.

1. **Data Gathering:**

Data was gathered using a website called *GitHub (*4), which allows you to fetch the data set on any topic. GitHub provide powerful collaboration, code review, and code management for open source and private projects. In GitHub the election data set of United Kingdom of the year 2010, 2011, 2012& 2014 was available which is used in this project for analyzing the sentiments.

1. **Data Cleaning:**

We downloaded the data sets as raw csv files and cleaned them by performing the following steps:

* Removed usernames, written as “@username”.
* rnames, written as “@username”  
   Removed all URLs.
* Removed the following punctuation marks: , " ‘? ! ; : #  
  $ % & ( ) \* + - / < > = [ ] \ ^ \_ { } | ~
* Removed all dates and timestamps of tweets from the  
  tweet content but stored it in a separate record against  
  the tweet.
* Replaced double and triple spaces between words  
  with single space.

After executing the aforementioned steps, we got a normalized  
version of the data. For instance, the tuple “2012 LOCAL KILLINGBECK [AND] SEACROFT CHAPMAN Harriet Liberal Democrats “223” 16905 @areef*”* was transformed to “2012 LOCAL KILLINGBECK AND SEACROFT CHAPMAN Harriet Liberal Democrats 223 16905*”* after cleaning.

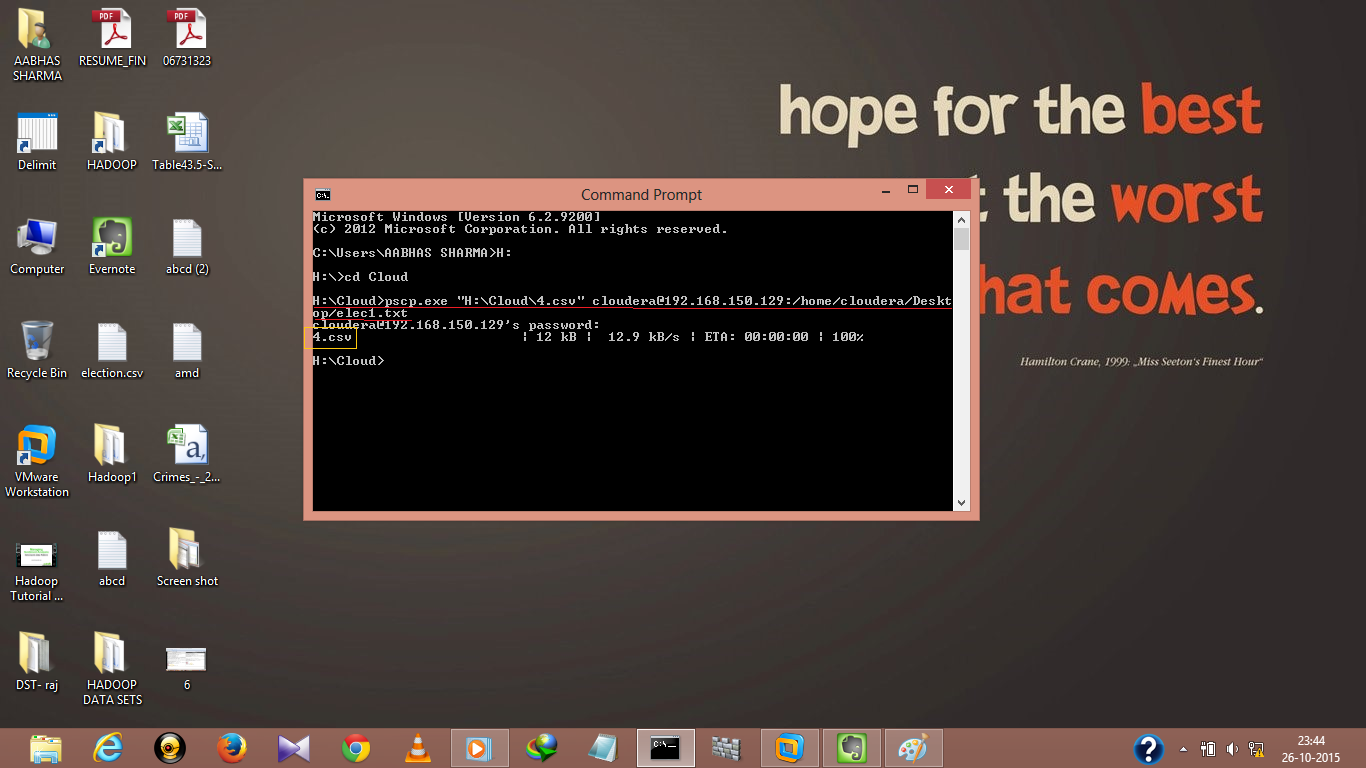
1. **Loading Data into HDFS:**

The data which is being cleaned is now loaded into HDFS. This is done in following steps:

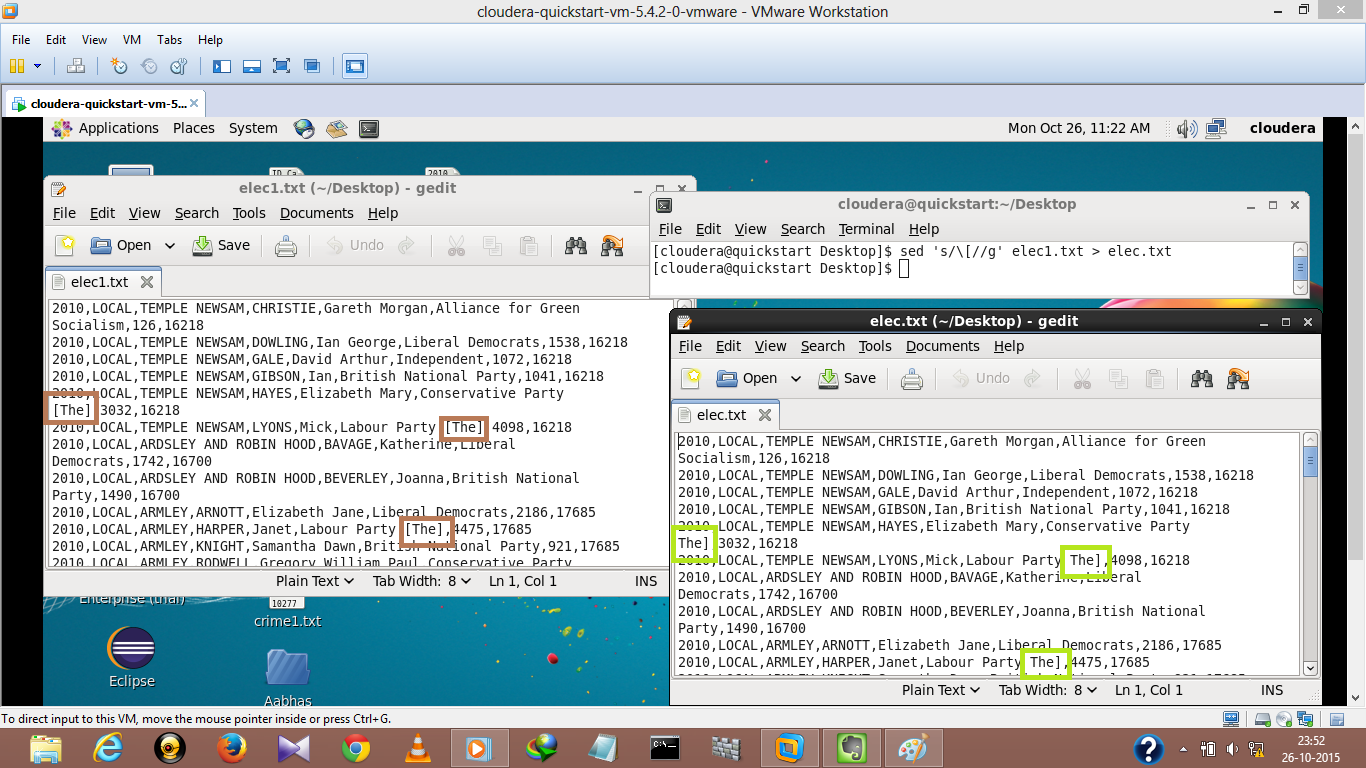
* First transfer the file which is in .csv format in windows to Cloudera’s Linux O.S. in txt format.
* Filter the .txt file in Linux by using ‘sed’ command.
* Upload this file into HDFS
* Create a table using hive script to put the data present in file into table.
* Apply the join operations on the tables to obtain the results.

1. **SIMULATION**

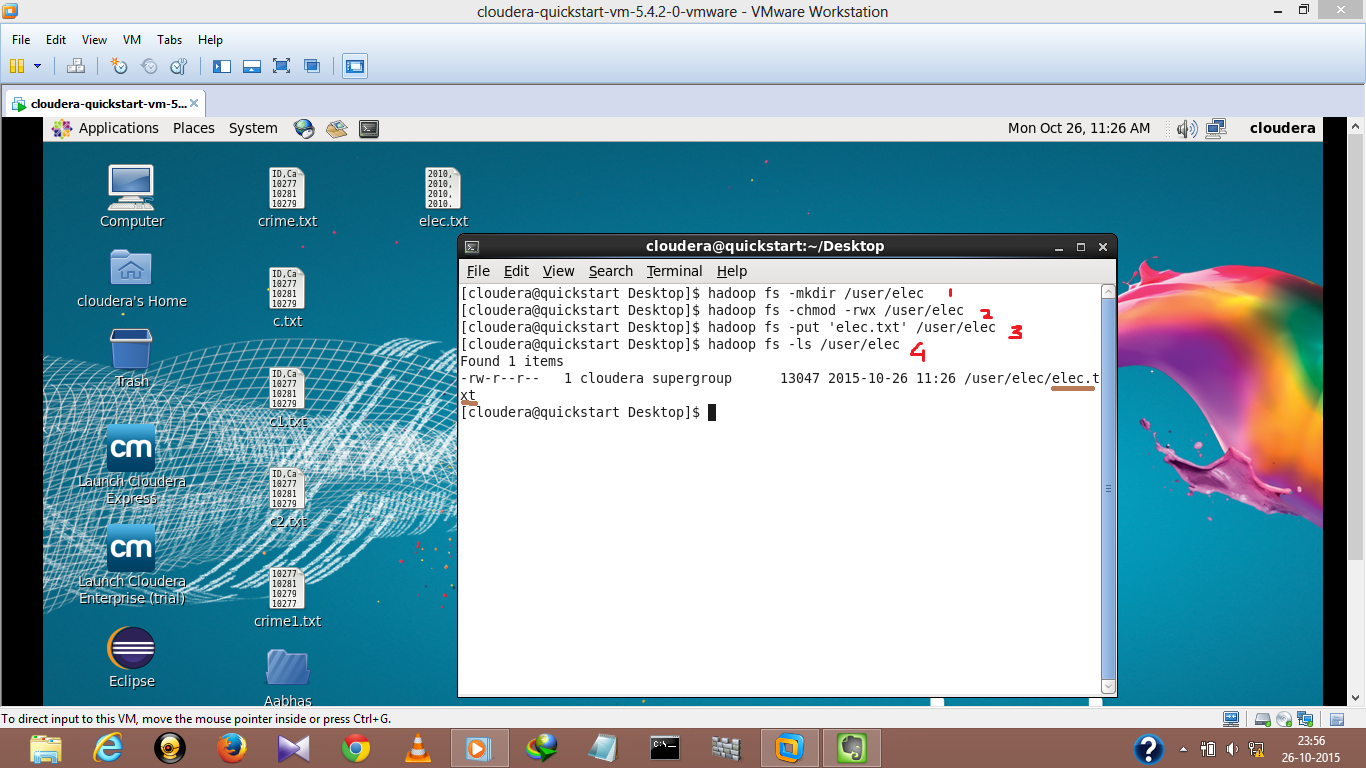
**4.1 transfer file from windows to sandbox**

****

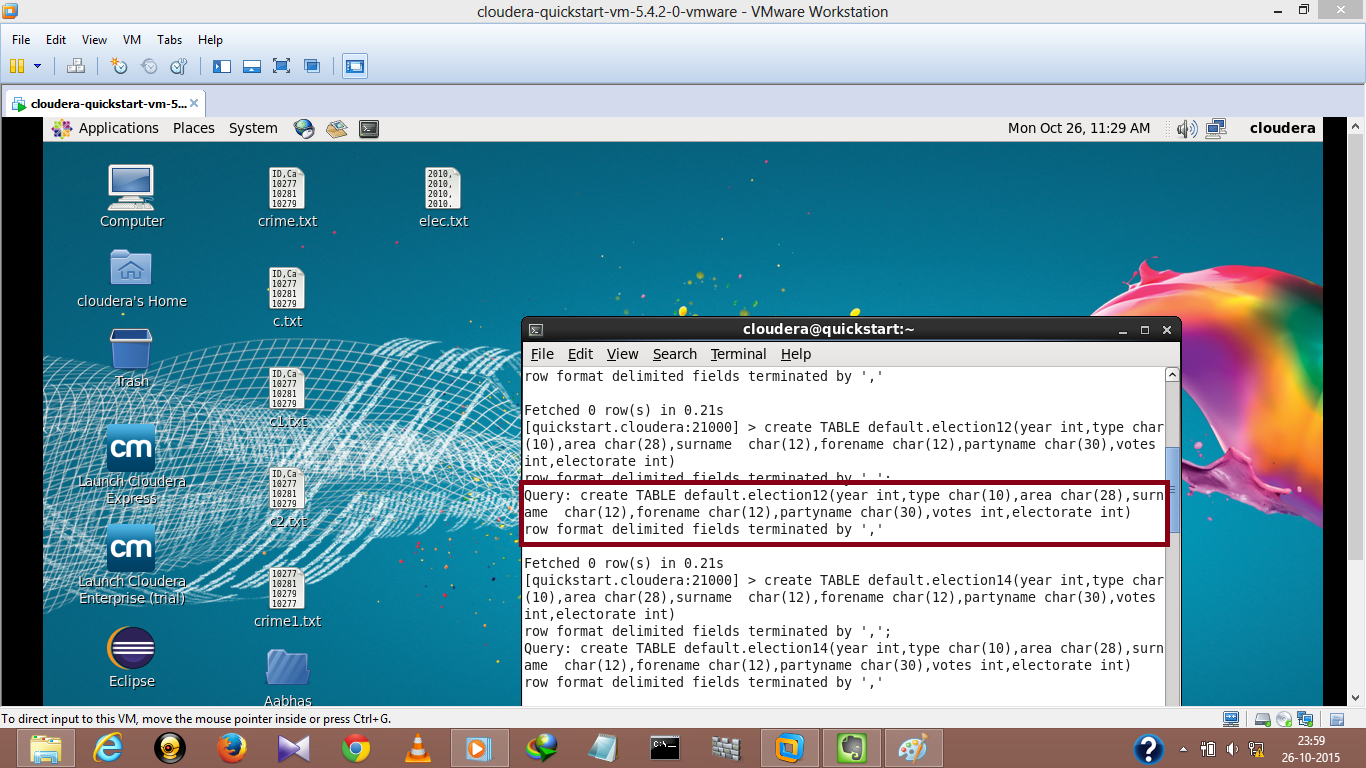
* 1. **remove ambiguities form the .txt file using ‘sed’ command**

****

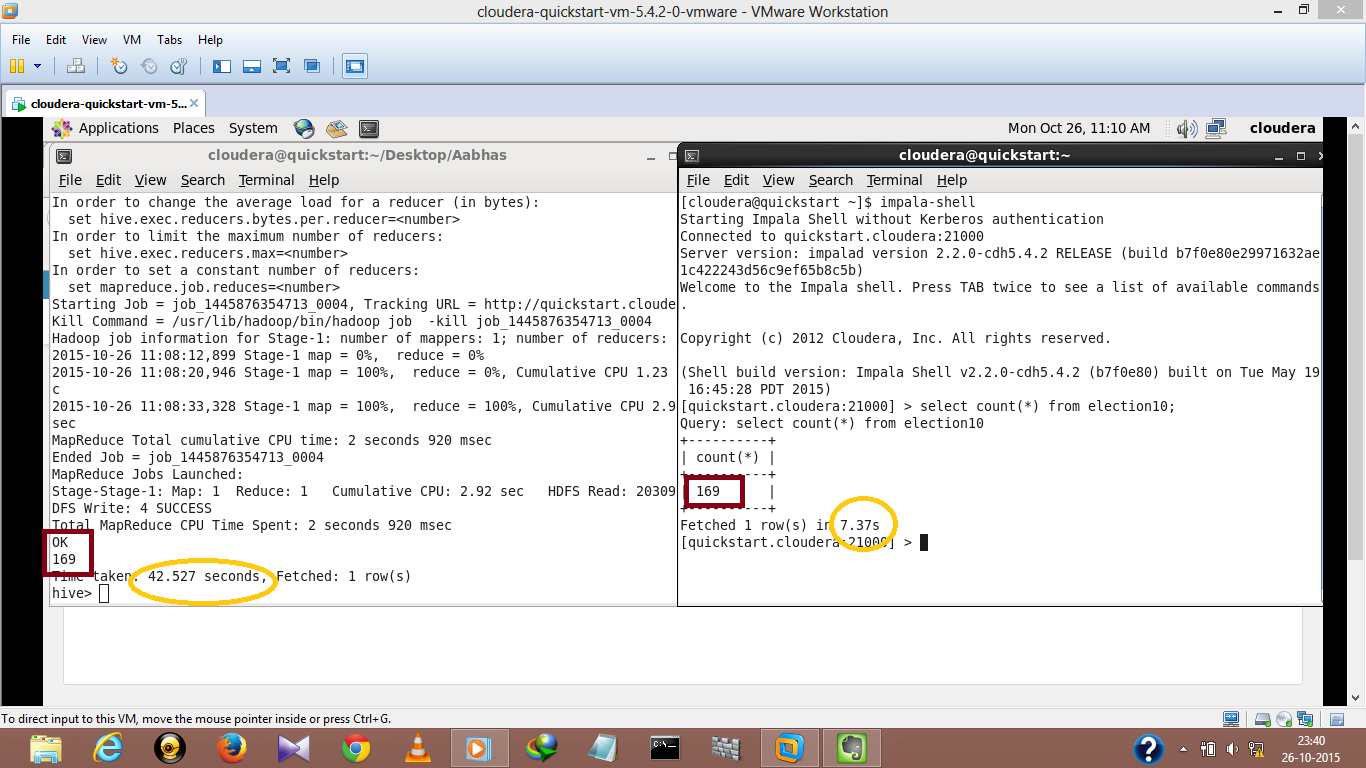
* 1. **Upload file into HDFS**

****

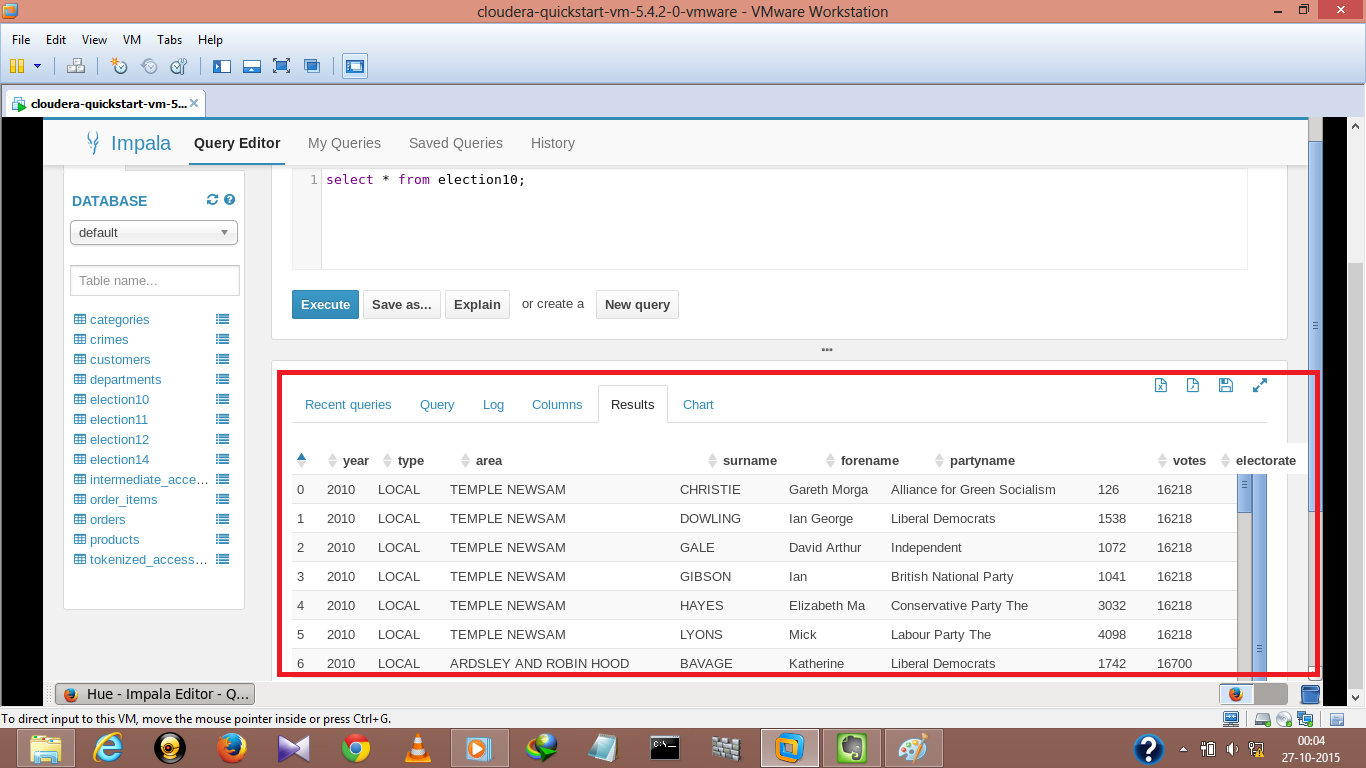
* 1. **Create table using hive/impala**

****

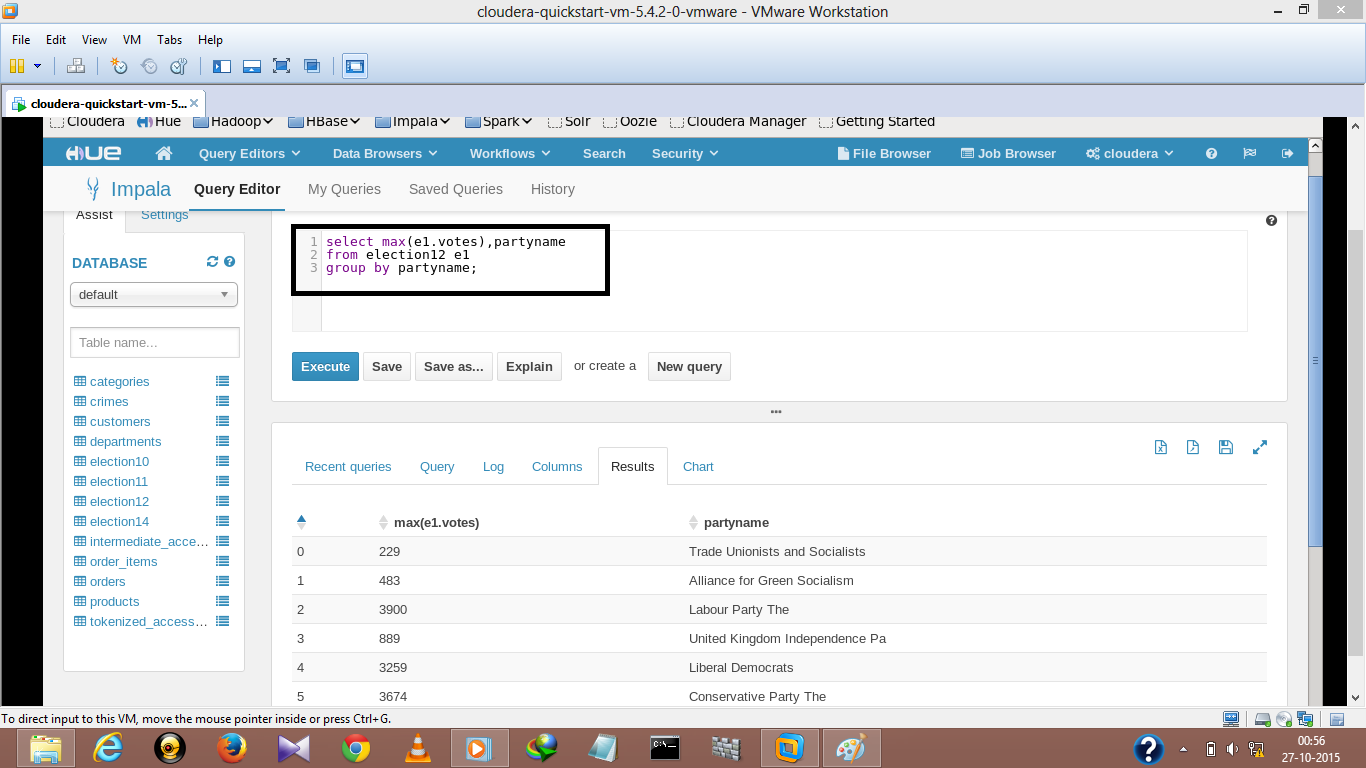
* 1. **Hive v/s Impala query execution time**

****

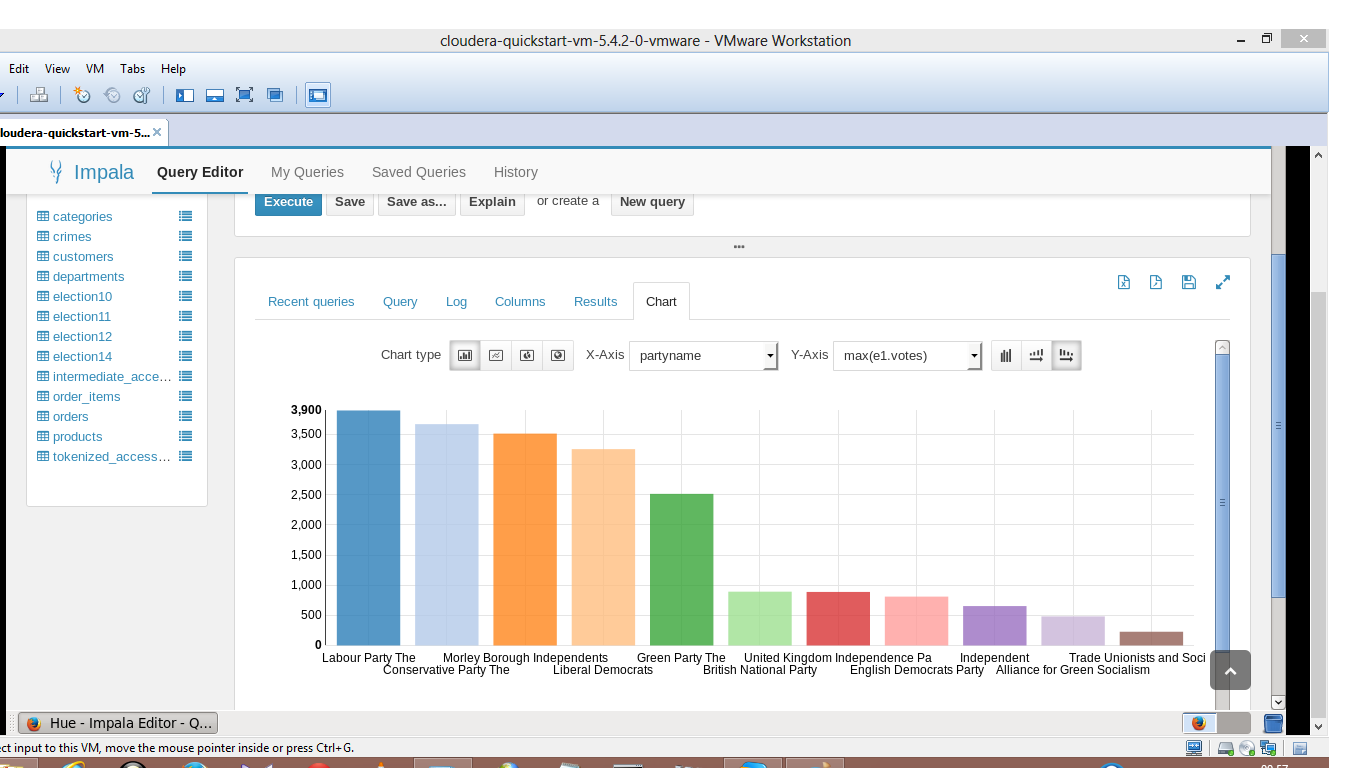
* 1. **To see the loaded table in Impala**

****

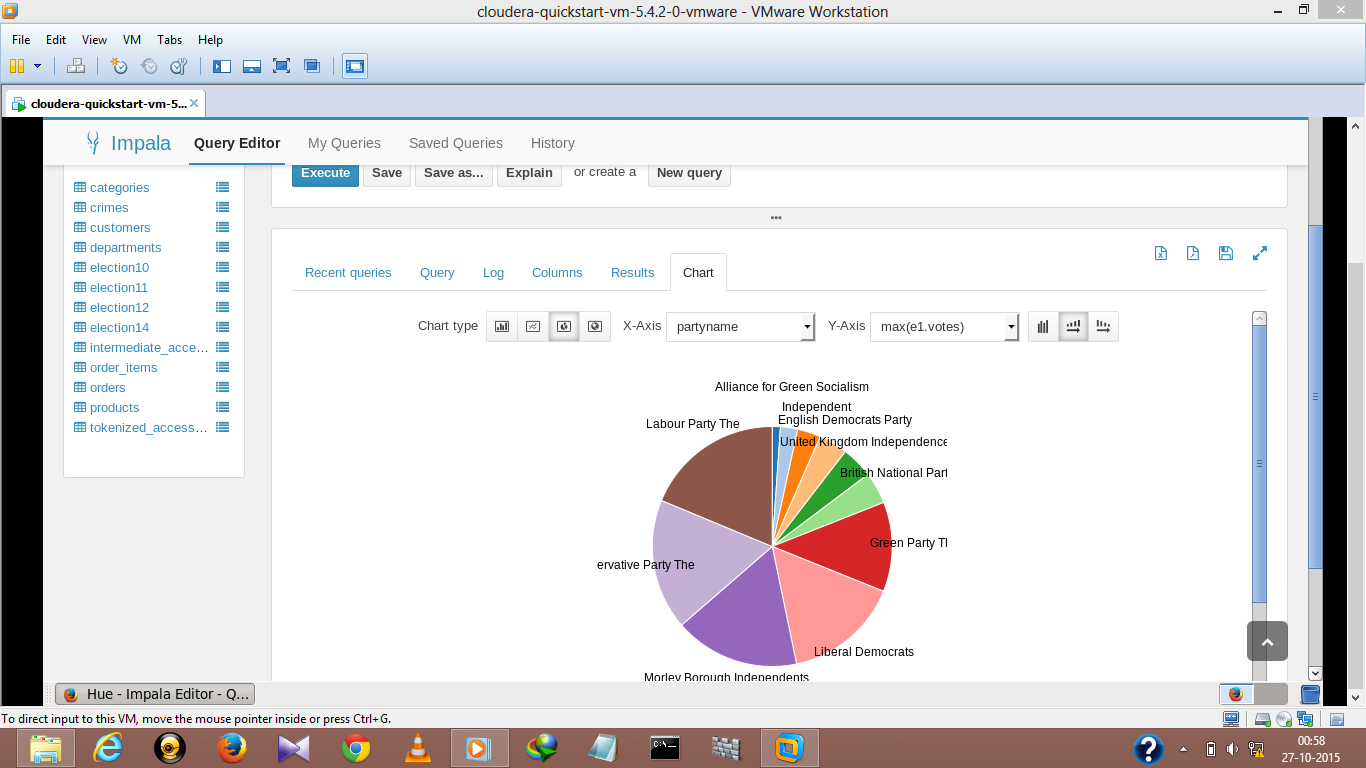
* 1. **query execution**
  2. **on single table**

****

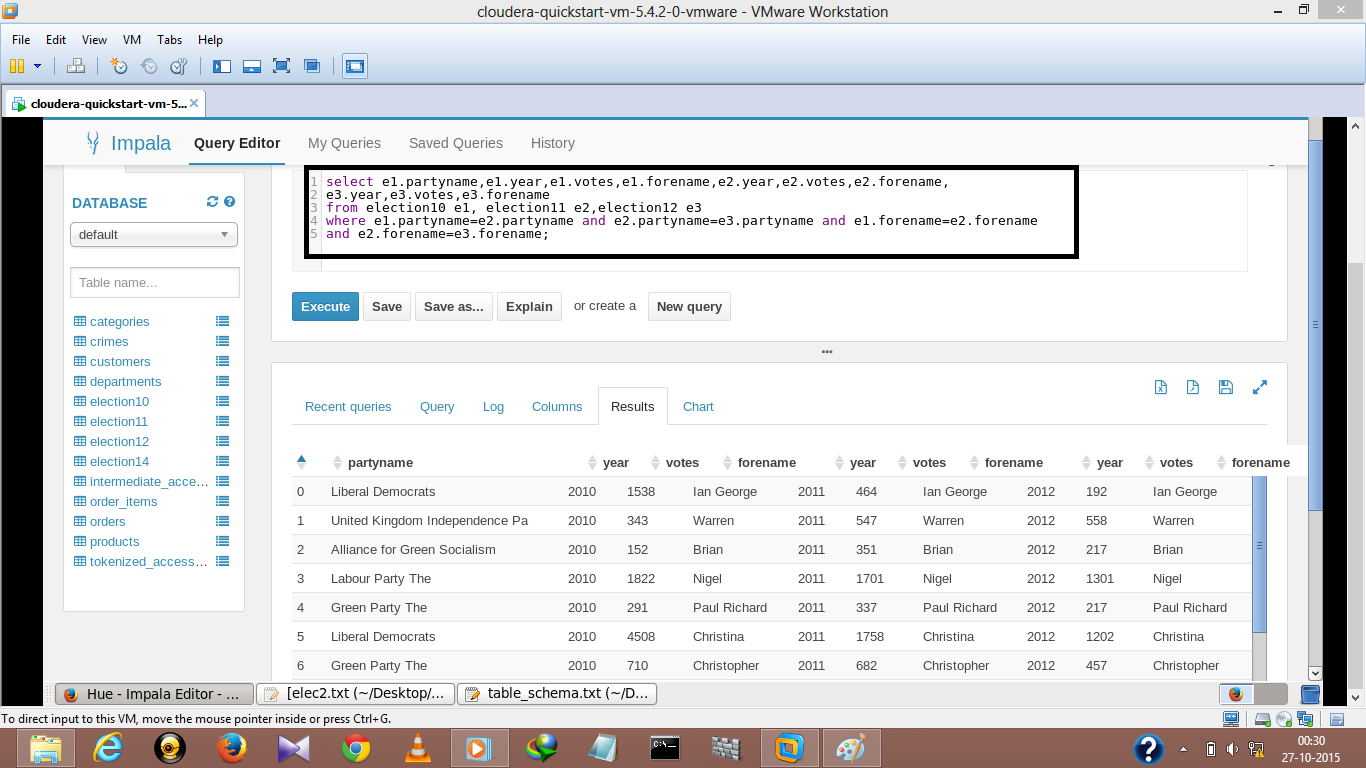
**Bar-Graph**

****

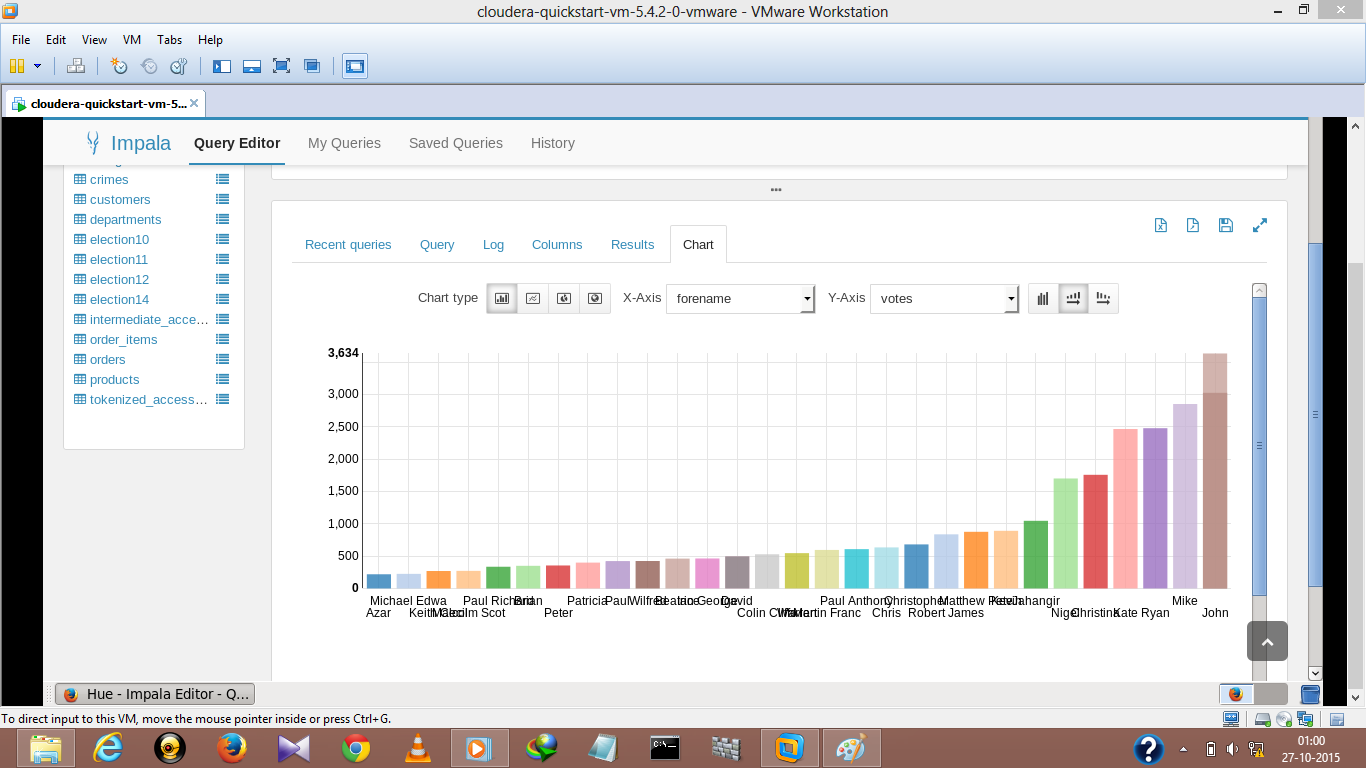
**Pie-chart**

****

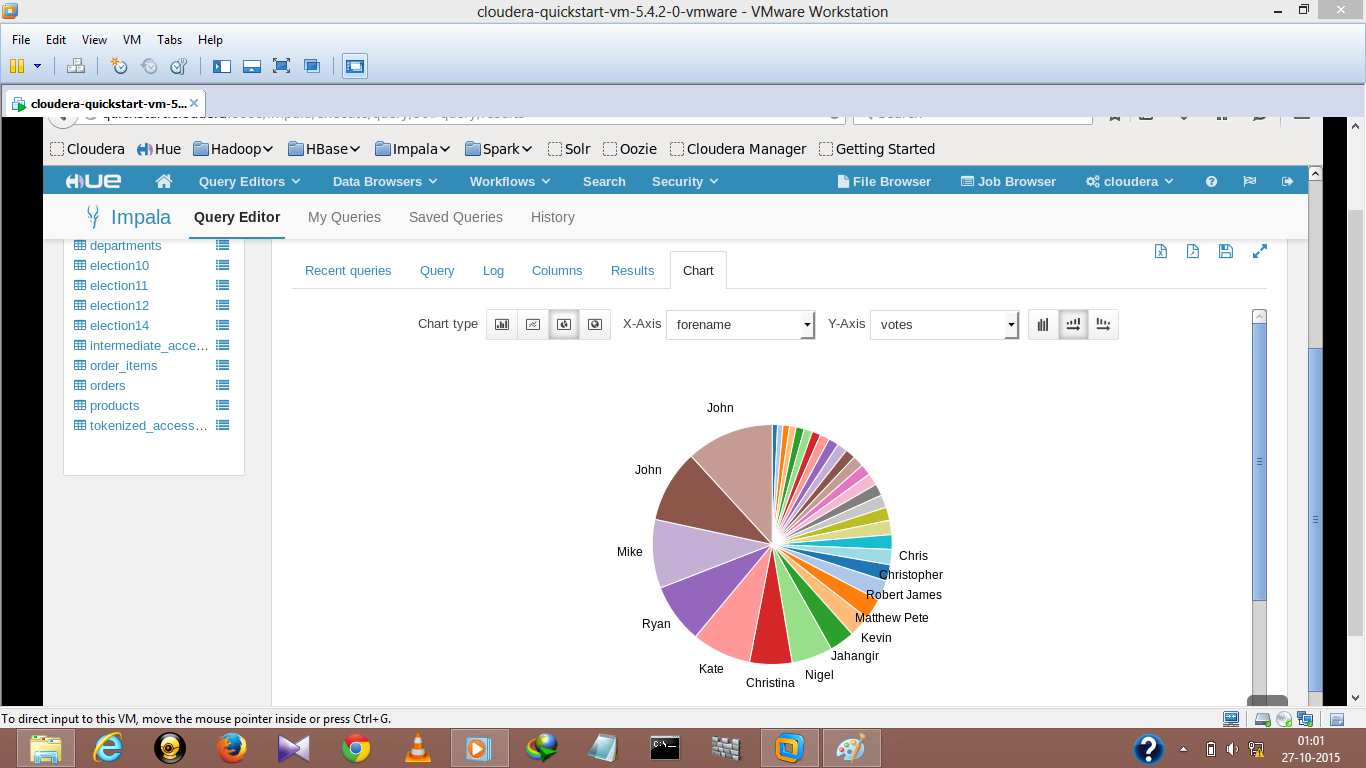
* 1. **on multiple tables using join operations**

****

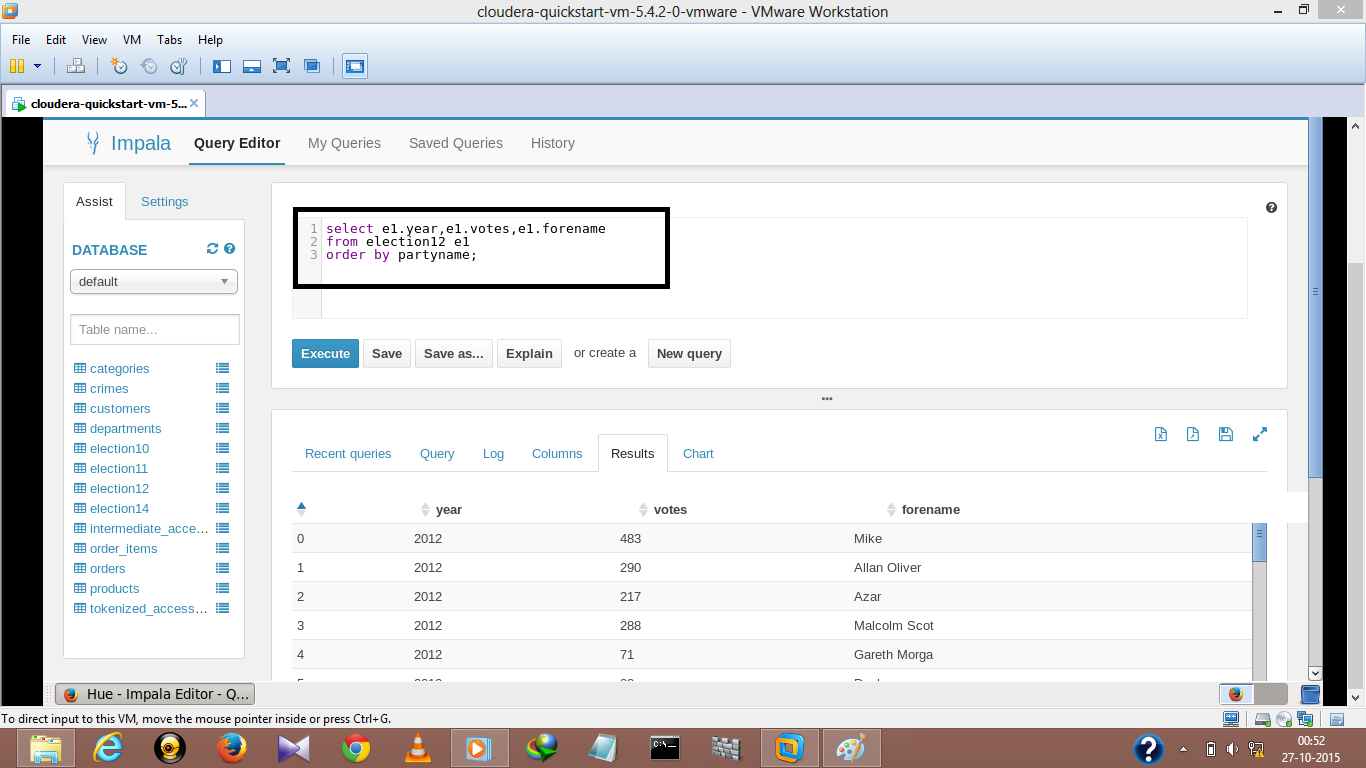
**Bar-Graph**

****

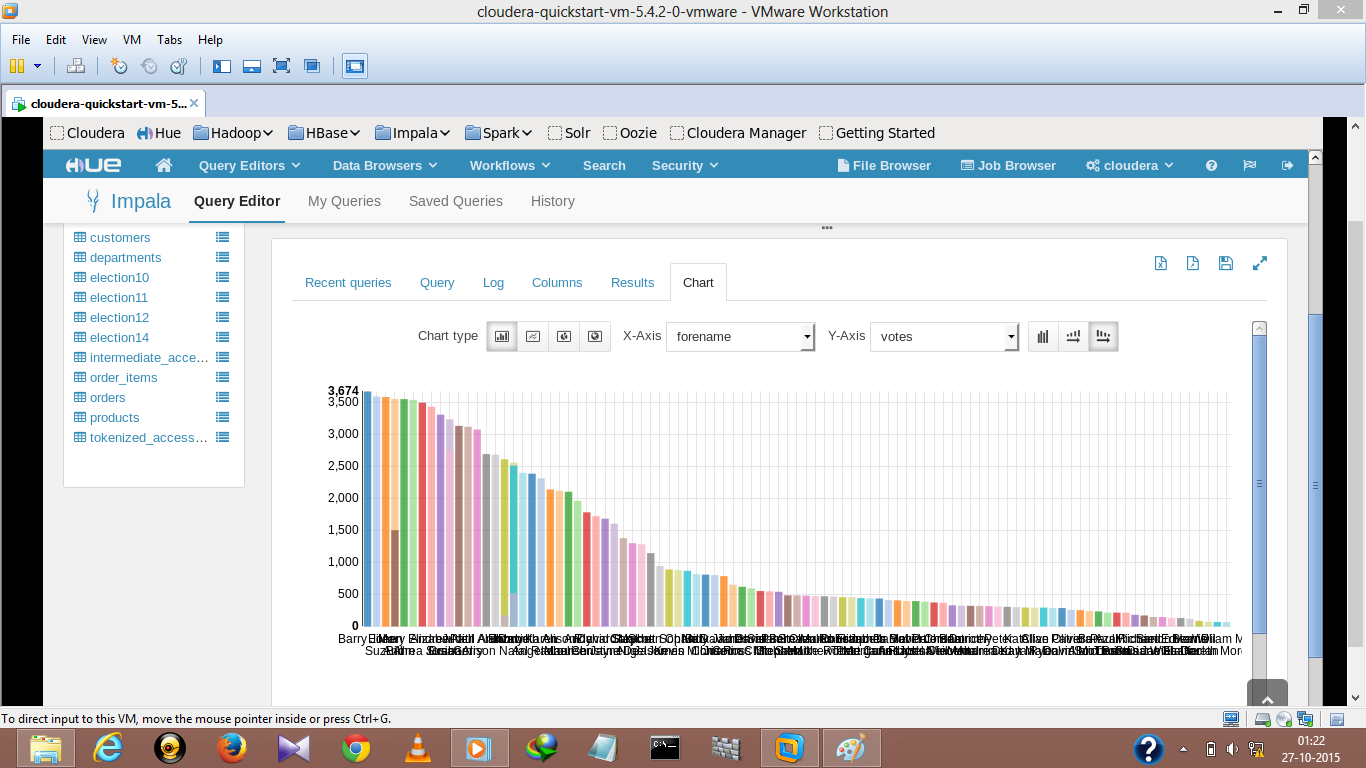
**Pie-Chart**

****

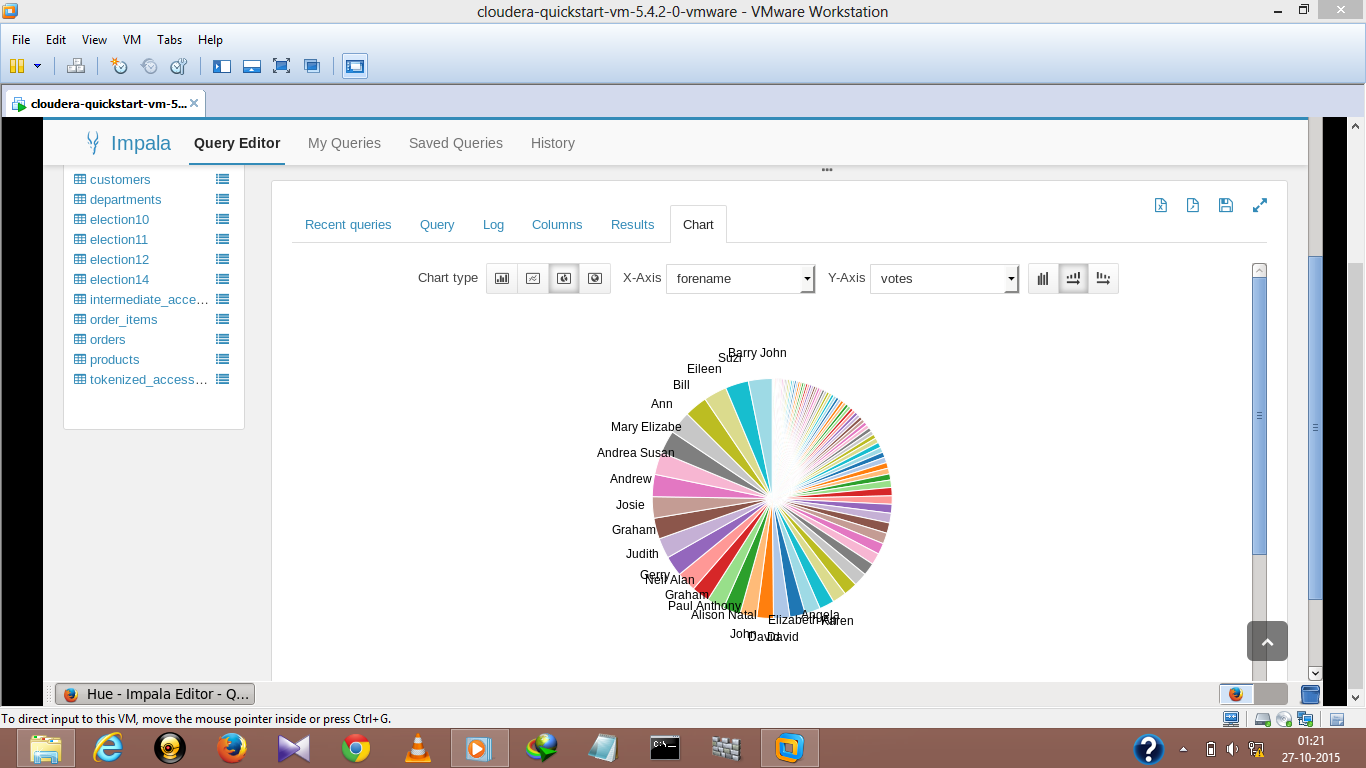
* 1. **Using ‘Order By’ clause for analysis**

****

**Bar-Graph**

****

**Pie-Chart**

****

1. **CONCLUSION & FUTURE SCOPE**

Sentiment analysis is a very wide branch for research. We have covered some of the important aspects. We plan ahead to improve our algorithm used for determining the sentiment value. Also the project as of now can also be expanded to other social media platform usages like movie reviews (IMDB reviews), personal blogs. Emoticons and the use of hash tags for the sentiment evaluation is a very important inference related to sentiment analysis of social media data. This project does not use emoticons & hash tags to determine the context of the tweet is not done.

At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also

depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret.

1. **REFERENCES**
2. <http://www.google.com/>
3. Bo Pang, Lillian Lee, “Opinion Mining and Sentiment Analysis”.
4. Murphy Choy, Michelle L.F. Cheong, Ma Nang Laik, and Koo Ping Shung, “A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction”, Journal CoRR, Vol. abs/1108.5520, 2011
5. <https://github.com/datasets/>
6. Dr. Tariq Mahmood, Tasmiyah Iqbal, Farnaz Amin, Wajeeta Lohanna, Atika Mustafa *“Mining Twitter Big Data to Predict 2013 Pakistan* *Election Winner”* Department of Computer Science National University of Computer & Emerging Sciences Karachi, Pakistan.
7. <https://en.wikipedia.org/wiki/Cloudera>
8. <https://anindita9.wordpress.com/2013/08/10/how-to-execute-cloudera-hadoopcdh4-quickstart-vmcentos-6-2-in-windows-with-vmware-workstation/>
9. <http://www.ijcsit.com/docs/Volume%205/vol5issue03/ijcsit2014050393.pdf>