COSC 6376: Grid/Cloud Computing

Instructor: Dr. Weidong Shi (Larry), Ph.D.

Homework 2

==========

Twitter Sentiment Analysis

==========

Project report

by

Venkata Yeshes Meka
PSID: 1141507
vmeka@uh.edu

Contents

1. Introduction:

Sentiment Analysis:

Sentiment analysis (also known as **opinion mining**) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. Sentiment analysis helps us determine the general observed mood of the person with respect to the identification keywords. A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level. It determines whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.

Sentiment analysis of tweets from the twitter (social networking micro blog) helps to determine the current trending topic. This is used for various promotional activities of the firms.

HBase:

HBase is an open source,non-relational, distributed database modeled afterGoogle's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoopproject and runs on top ofHDFS (Hadoop Distributed Filesystem), providing BigTable-like capabilities for Hadoop. That is, it provides afault-tolerant way of storing large quantities of sparse data. HBase is a column-oriented database management system that runs on top ofHDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java much like a typicalMapReduce application. HBase does support writing applications in Avro, REST, and Thrift. An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key. An HBase column represents an attribute of an object; for example, if the table is storing diagnostic logs from servers in your environment, where each row might be a log record, a typical column in such a table would be the timestamp of when the log record was written, or perhaps the server name where the record originated. In fact, HBase allows for many attributes to be grouped together into what are known as column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together.

2. Experimental Setup

Goal:

To learn how to use Hbase and a set of open source tools for collecting and analyzing big data from the Internet.

1. Understand the HBase design and implementation.
2. Build a classification model for the tweets from Twitter and understand the various algorithms which help build a hybrid classification model.
3. Analyze the effect of running these algorithms on a large dataset consisting of tweets and classify accordingly to analyze the sentiment of the users.

Tools Used:

1. Cloudera virtual machine with Hadoop eco-system installed which contains HBase.
2. Eclipse.
3. Java 1.6 or higher.

Dataset:

The dataset is a the collection of tweets gathered from user's twitter account using the twitter API.

The basic data-format of the dataset is as follows.

The tweets gathered from the twitter contain the keyword used to fetch the tweets, the timestamp and the tweet.
Eg: java:2013:12:06:20:31 1355106715946 1

3. Configuring the eco-system

a. Configuring HDFS of Cloudera virtual machine/ to run on local machine.

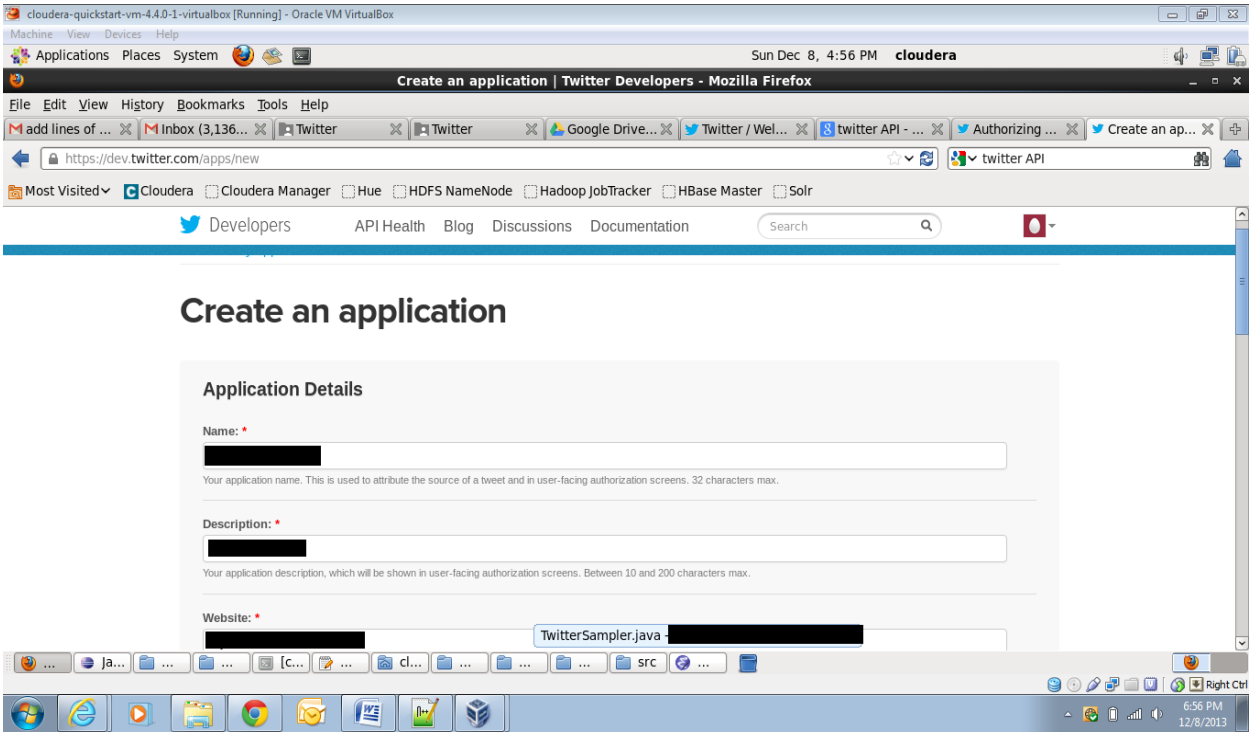
The Cloudera distribution comes with a graphical interface for the HDFS system with pre-installed Hadoop eco-system. The folder "Cloudera's-home" is the HDFS for the Hadoop eco –system installed on local machine.

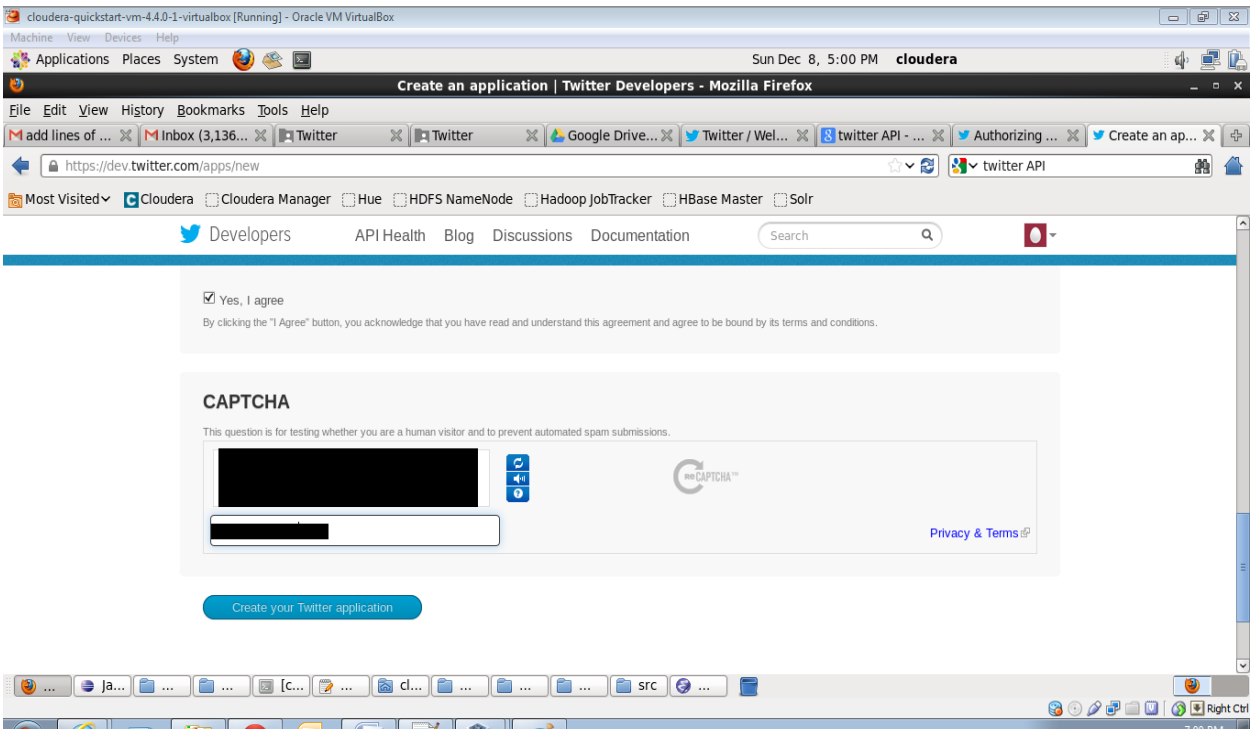
The Cloudera distribution comes with a management console for the Hadoop eco-system called "Cloudera Manager", which provides the graphical interface to control the tasks.

The tweets are stored in the HBase of the Cloudera VM.

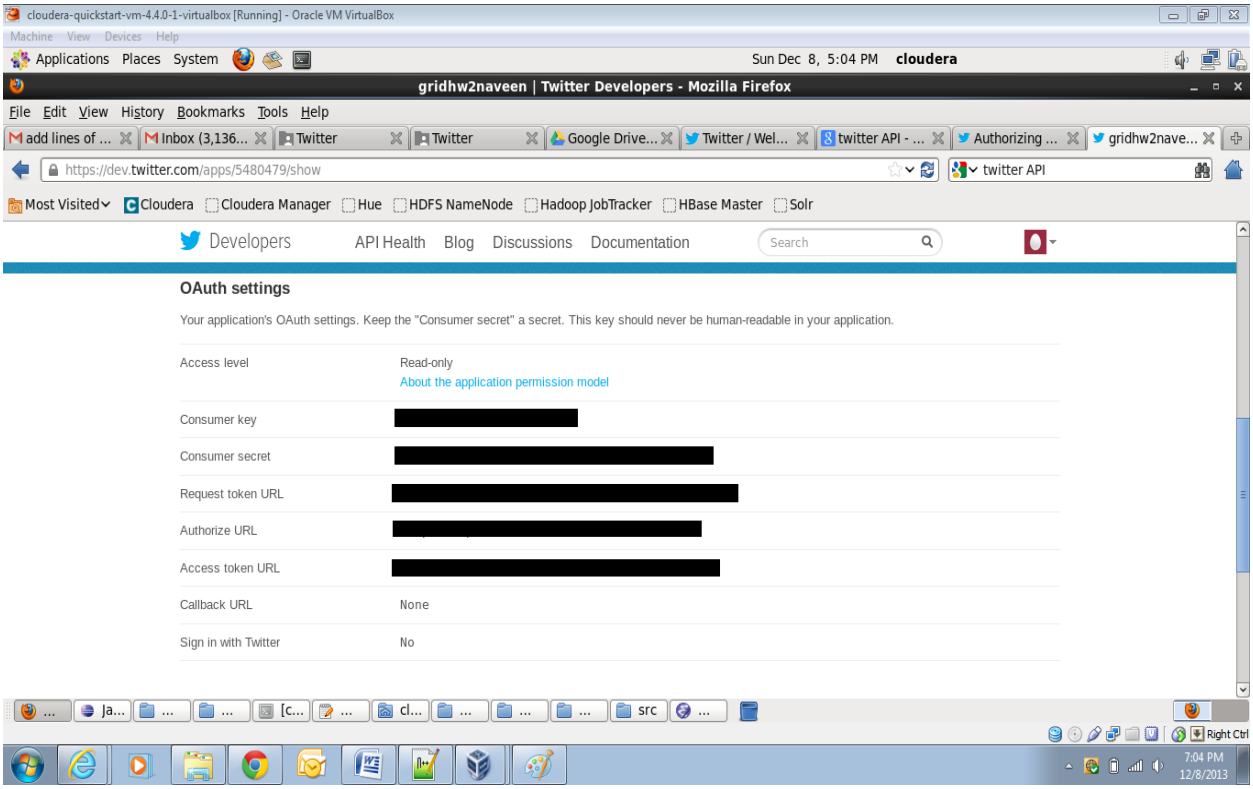
b. Setting up Twitter API developer's account:

1. Set up a conventional twitter account if you don't already have one and log in.
2. Go to https://dev.twitter.com/apps/new and fill the form on the page and agree to the terms and conditions.

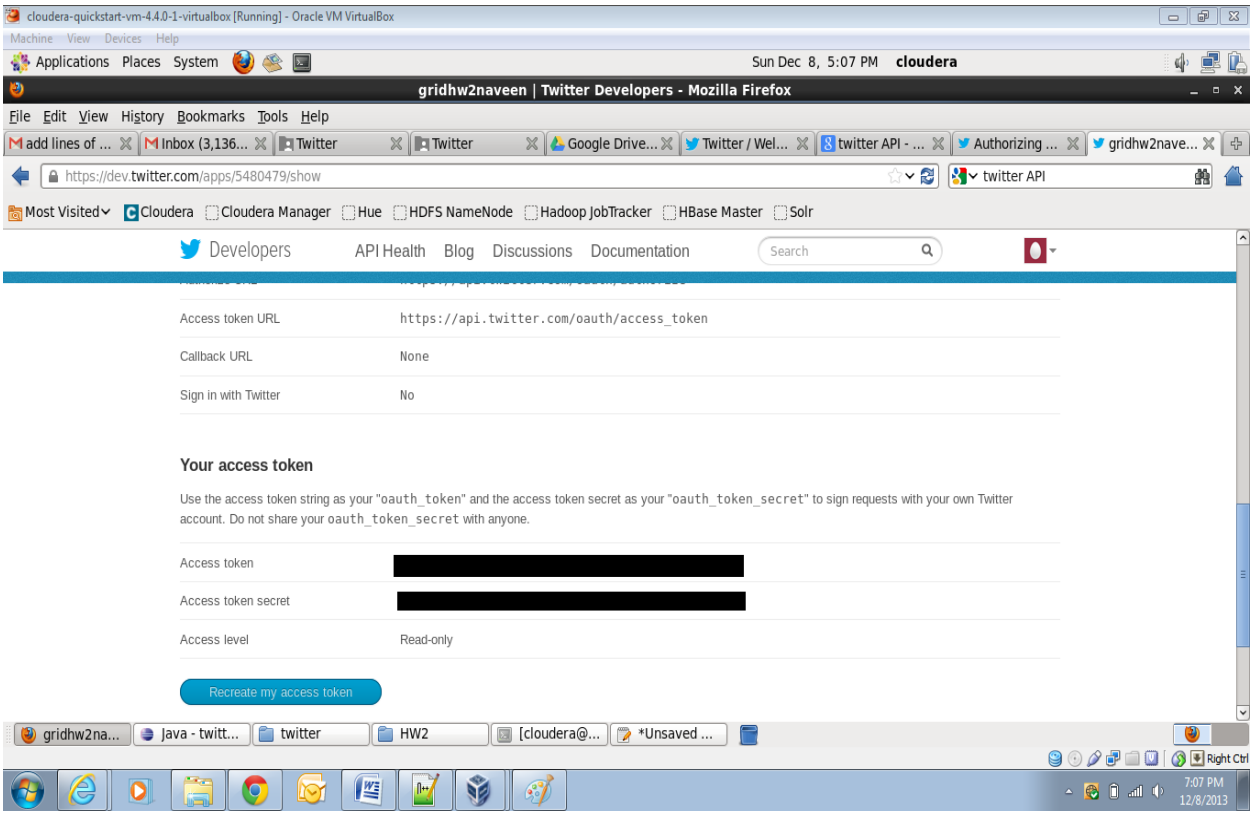

cloudera-quickstart-vm-4.4.0-1-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
Sun Dec 8, 4:56 PM cloudera
Create an application | Twitter Developers - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://dev.twitter.com/apps/new
twitter API
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
Developers API Health Blog Discussions Documentation
Search
Create an application
Application Details
Name: *
Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.
Description: *
Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.
Website: *
TwitterSampler.java -
6:56 PM
12/8/2013


cloudera-quickstart-vm-4.4.0-1-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
Sun Dec 8, 5:00 PM cloudera
Create an application | Twitter Developers - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://dev.twitter.com/apps/new
twitter API
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
Developers API Health Blog Discussions Documentation
Search
Yes, I agree
By clicking the "I Agree" button, you acknowledge that you have read and understand this agreement and agree to be bound by its terms and conditions.
CAPTCHA
This question is for testing whether you are a human visitor and to prevent automated spam submissions.
Privacy & Terms
Create your Twitter application

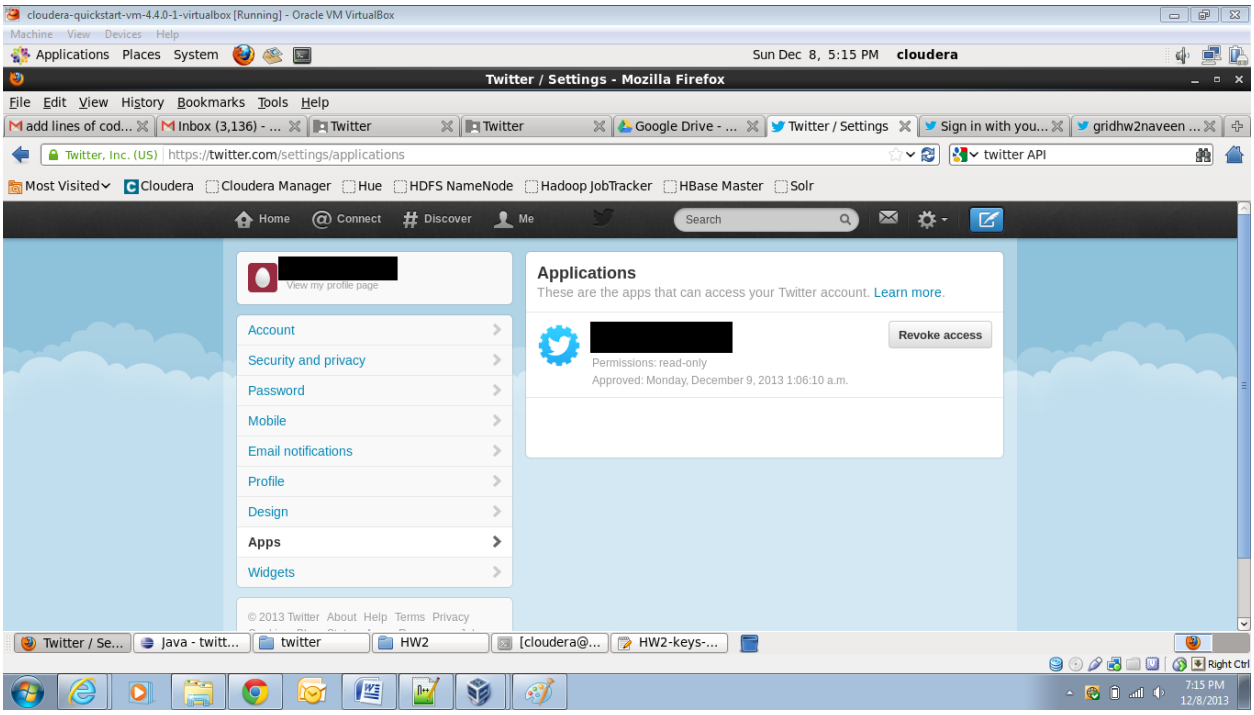
3. Make a note of the consumer key and the consumer secret key on the next page.



4. Also note your access token and the secret access token on the page following (3).



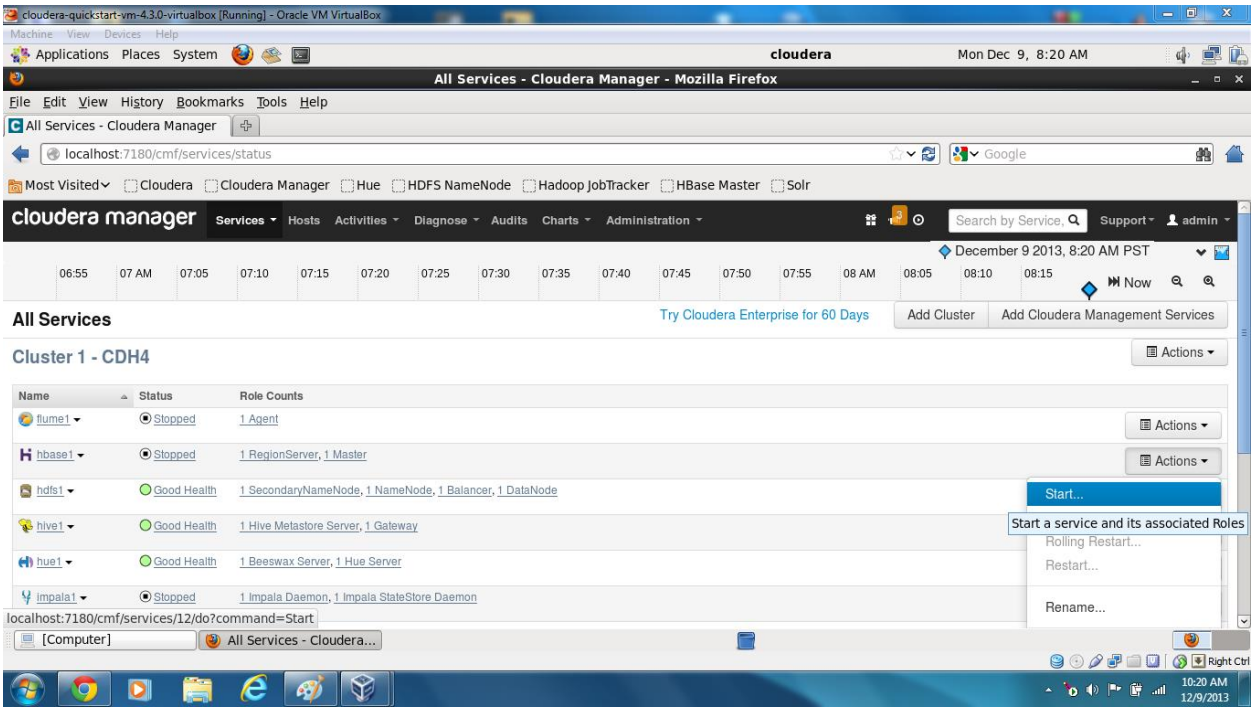
5. You can revoke the your twitter app's access to your account by:
    - Log in to your twitter account.
    - Click "Settings".
    - Click "Apps".



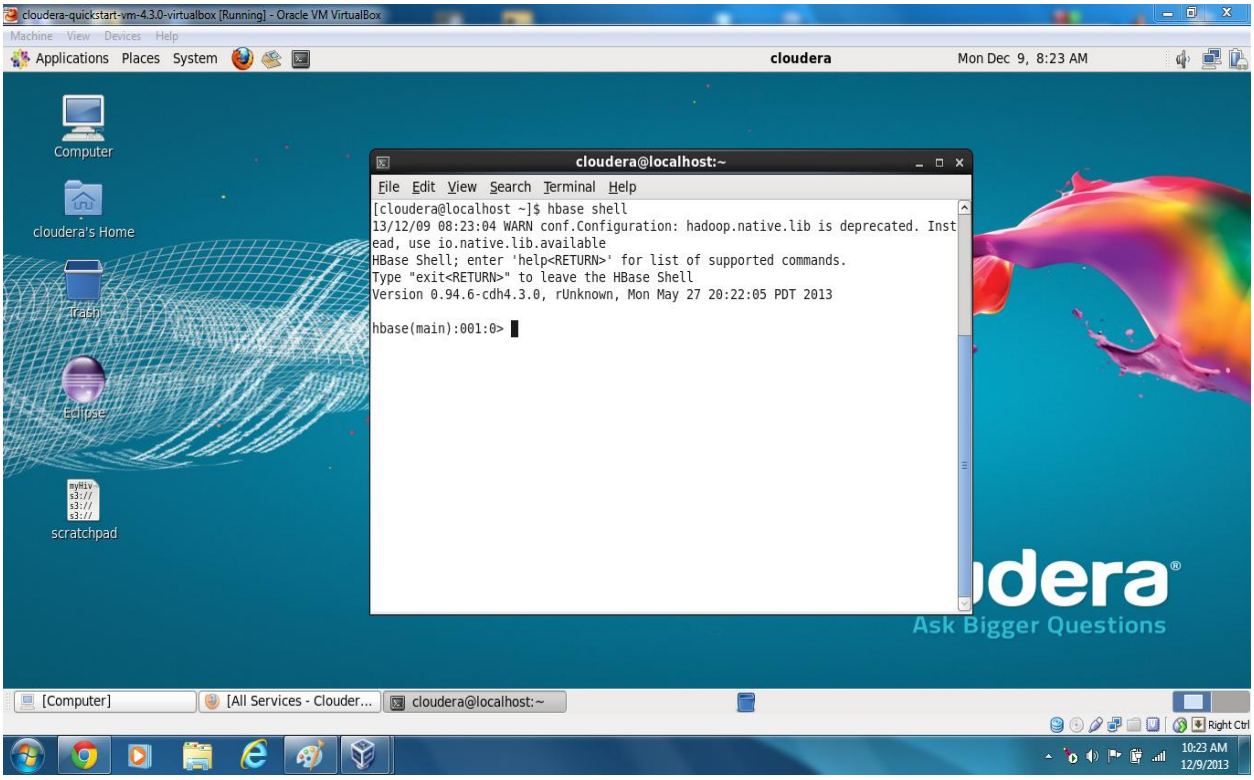
6. The four alphanumeric strings were recorded:
    - Consumer key.
    - Consumer secret key.
    - Access token.
    - Secret access token.
7. These will be used in our Java Application as it utilizes Twitter API v1.1 that supports OAuth which utilizes these keys.

Creating the HBase Table (on Cloudera VM):

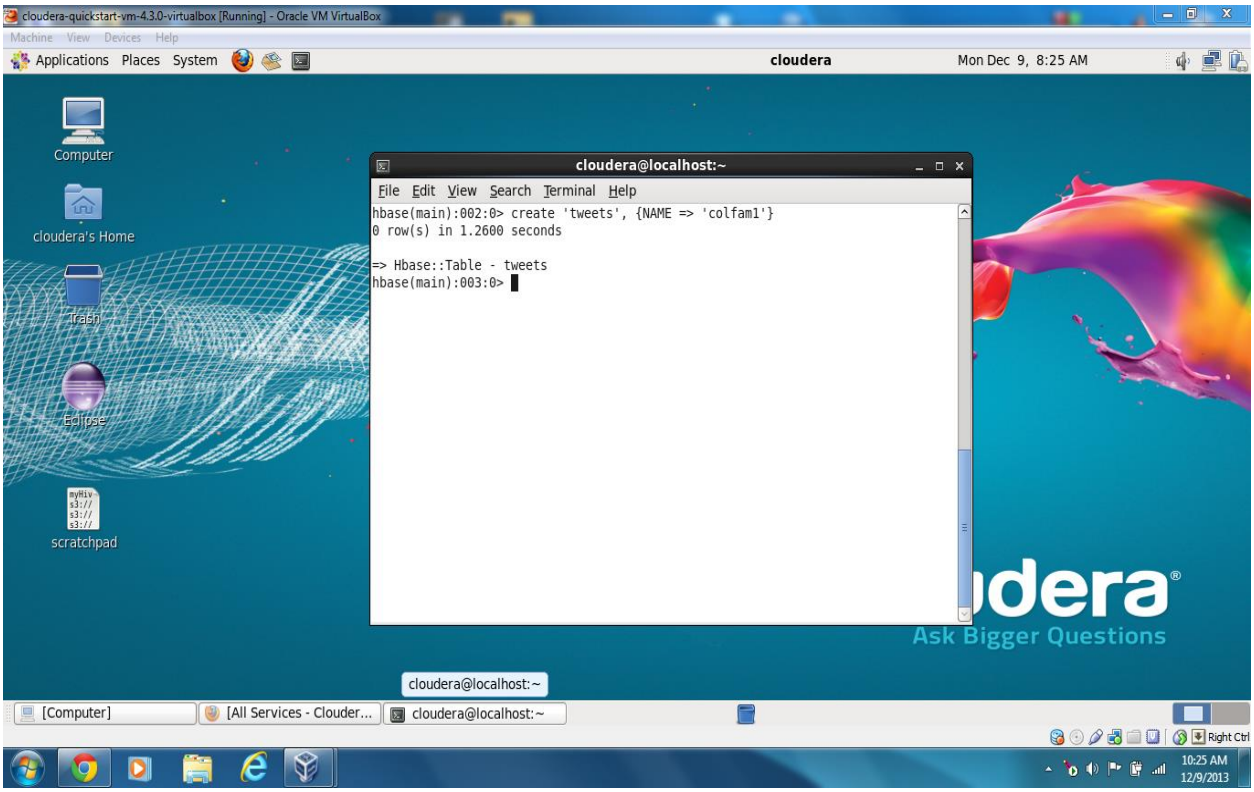
1. Start the HBase master server.



2. Start HBase shell in a BASH terminal.



3. Execute the query to create the table.



To view the data in the table, use the following command.

Scan <table name>

Scan “tweets” gives the following output.

csk:2013:12:06:23:22 colfam1:NEUTRAL 1355116976663 1
csk:2013:12:06:23:22 colfam1:POSITIVE 1355116976663 1
csk:2013:12:06:23:23 colfam1:NEGATIVE 1355116997208 0
csk:2013:12:06:23:23 colfam1:NEUTRAL 1355116997208 1
csk:2013:12:06:23:23 colfam1:POSITIVE 1355116997208 1
sachin:2013:12:06:19:43 colfam1:NEGATIVE 1355103780812 0
sachin:2013:12:06:19:43 colfam1:NEUTRAL 1355103780812 0
sachin:2013:12:06:19:43 colfam1:POSITIVE 1355103780812 1

4. Program Sequence

The task to classify the tweets in order to perform sentiment analysis can be carried out in three stages.

a. Collecting the tweets from the twitter using the twitter API:

Twitter streaming agent:

a. I used the twitter4J library to "stream" public tweets to classify and store in the database.
b. Twitter4J is an unofficial Java library for the Twitter API.
c. Twitter4J allows easily integration of a Java application with the Twitter service. It has built in OAuth support.
d. When integrated allows the creation of instances of theTwitterStream class which accepts several arguments
e. Twitter account username.
f. Twitter account password.
g. Directive i.e. "-f" for filter and "-s" for sampling.
h. Filtering keywords when using "-f" directive.
i. This library is a self-contained jar file which allows for easy integration with the java code.

HBase connector

a. It is used to write the classified tweets into HBase.
b. We used a simple database with a single column family called "colfam1" in a table called tweets.
c. The client automatically "sorts" the tweets into 3 row families; NEUTRAL, POSITIVE AND NEGATIVE.
d. One thing that should be noted is that the HBase Master is running before the code is run and that the HBase table is created prior to code execution.

b. Performing the sentiment analysis and storing in the HBase:

a. I used an open source third party library to perform sentiment analysis which relies on Weka classifiers.
b. The library uses an ensemble of 3 classifiers namely the J48 decision tree, voted perceptron and a Bayesian classifiers.
c. The library comes in built in text base and a preprocessor, making it relatively easy to use out of the box.

d. The library also comes with prebuilt models thus allowing the user to skip the training phase entirely.

c. Retrieving the data stored in the HBase and displaying in the webpage:
   a. This component is used to read the data from the HBase table and then display them to the user.
   b. The output of the read is in XML format.
   c. This component also parses this XML to extract the actual tweet.
   d. Please note that the HBase rest service must be running prior to execution of the JavaScript code.
   e. To start the HBase service use the command "*hbase rest start*" on the bash shell.

5. Challenges

The main challenges I faced during the phase of development are:

1. ***HBase – JavaScript connection complexity:***
   The problem of fetching the data from the HBase and then posting it to the JavaScript file is a hassle since, the data from takes time to load if there are large number of tweets which are constantly classified.
   Strategy: I used an intermediate XML file to process the data being read from the HBase.

2. ***Computational complexity***
   Generally, the problem of working with large datasets is the computational complexity. Since the given dataset of collects the tweets from the Twitter website, the data to be classified can be huge and the runtime may vary depending on the keyword used for fetching.
   Strategy: The algorithms have to be optimized in order to achieve the analytic results. I used less common/ trending keyword initially to determine if everything works correctly. Then I analyzed on most trending topics.

3. ***Processing power***
   The major concern is the processing power since these BigData tasks demand high processing power to process the data. This means the availability of clusters which can be specifically used to run Hadoop MapReduce jobs by the HBase master.

4. ***Twitter API requests restriction:***

   The problem with Twitter API is that there is a restriction on the numbers of tweets that can be fetched from the account in a given timeframe. This harnesses the speed of the application. Sometime it also leads to erroneous results or even the fetch is blocked.

6. References

[1] www.code.google.com
[2] https://dev.twitter.com/apps/new
[3] http://en.wikipedia.org/wiki/Sentiment_analysis
[4] http://money.cnn.com/2012/08/23/technology/twitter-api-tumblr/