# Refining and Visualizing Sentiment Data

David Hoyle
25 August 2013

## Introduction

### Summary

This tutorial describes how to refine raw Twitter data using the Hortonworks Data Platform, and how to analyze and visualize this refined sentiment data using the Power View feature in Microsoft Excel 2013.

### Sentiment Data

Sentiment data is unstructured data that represents opinions, emotions, and attitudes contained in sources such as social media posts, blogs, online product reviews, and customer support interactions.

### Potential Uses of Sentiment Data

Organizations use sentiment analysis to understand how the public feels about something at a particular moment in time, and also to track how those opinions change over time.

An enterprise may analyze sentiment about:

- A product – For example, does the target segment understand and appreciate messaging around a product launch? What products do visitors tend to buy together, and what are they most likely to buy in the future?
- A service – For example, a hotel or restaurant can look into its locations with particularly strong or poor service.
- Competitors – In what areas do people see our company as better than (or weaker than) our competition?
- Reputation – What does the public really think about our company? Is our reputation positive or negative?

In this tutorial, we will focus on a product launch. Specifically, we will look at public sentiment during the days leading up to and immediately following the recent release of the movie *Iron Man 3*.

How did the public feel about the debut, and how might that sentiment data have been used to better promote the movie's launch?

**Prerequisites:**

- Hortonworks Sandbox (installed and running)
- Hortonworks ODBC driver installed and configured
- Microsoft Excel 2013 Professional Plus
- Sentiment tutorial files

**Notes:**

- In this tutorial, the Hortonworks Sandbox is installed on an Oracle VirtualBox virtual machine (VM).
- Install the ODBC driver that matches the version of Excel you are using (32-bit or 64-bit).
- In this tutorial, we use the Power View feature in Excel 2013 to visualize the sentiment data. Power View is currently only available in Microsoft Office Professional Plus and Microsoft Office 365 Professional Plus.

**Overview**

To refine and visualize website sentiment data, we will:

1. Download and extract the sentiment tutorial files.
2. Load Twitter data into the Hortonworks Sandbox.
3. Copy a Hive script to the Sandbox.
4. Run the Hive script to refine the raw data.
5. Access the refined sentiment data with Excel.
6. Visualize the sentiment data using Excel Power View.

## Step 1 – Download and Extract the Sentiment Tutorial Files

1. You can download a set of sample Twitter data contained in a compressed (.zip) folder here:

   <the download link to SentimentFiles.zip>

   The Twitter data was obtained using Apache Flume. Flume can be used as a log aggregator, collecting log data from many diverse sources and moving it to a centralized data store. In this case, Flume was used to capture the Twitter stream data, which we can now load into the Hadoop Distributed File System (HFDS).

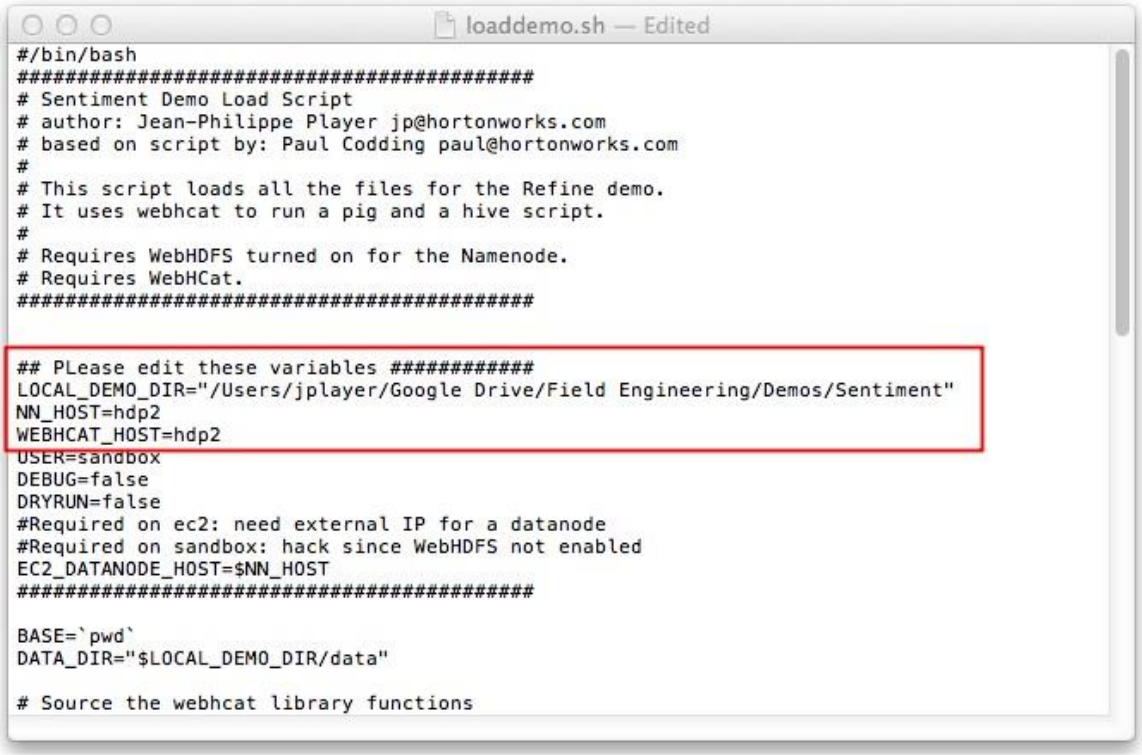
2. Save the SentimentFiles.zip file to your computer, then extract the files. You should see a SentimentFiles folder that contains a loaddemo.sh file.

## Step 2 – Load Twitter Data into the Hortonworks Sandbox

We will now load Twitter data into the Sandbox. The procedure is slightly different for Windows and Mac. This section describes the procedure on Mac OS X.

### Load Twitter Data into the Hortonworks Sandbox – Mac OS X

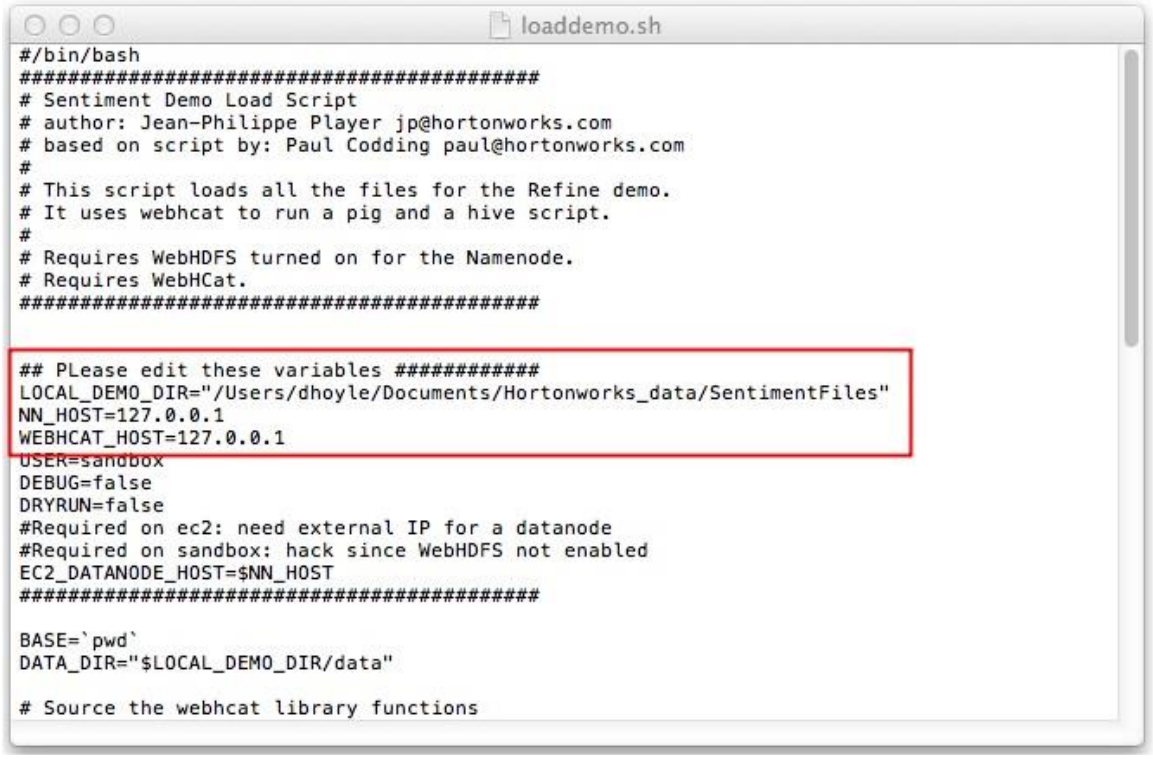
1. Open the loaddemo.sh file in the SentimentFiles folder using TextEdit in plain text mode. Near the beginning of the file, you will see an area of text labeled "Please edit these variables."




2. Edit the first three variables as follows:

   - LOCAL_DEMO_DIR – The path to the SentimentFiles folder.
   - NN_HOST – The Sandbox IP address.
   - WEBHCAT_HOST – The Sandbox IP address.

   Save your changes to the loaddemo.sh file.




3. Open a Terminal window and navigate to the SentimentFiles folder containing the loaddemo.sh file. Type in the following command, then press the Enter key:

   ```
   ./loaddemo.sh
   ```

4. A message about enabling WebHDFS appears in the Terminal window. Press any key to continue.

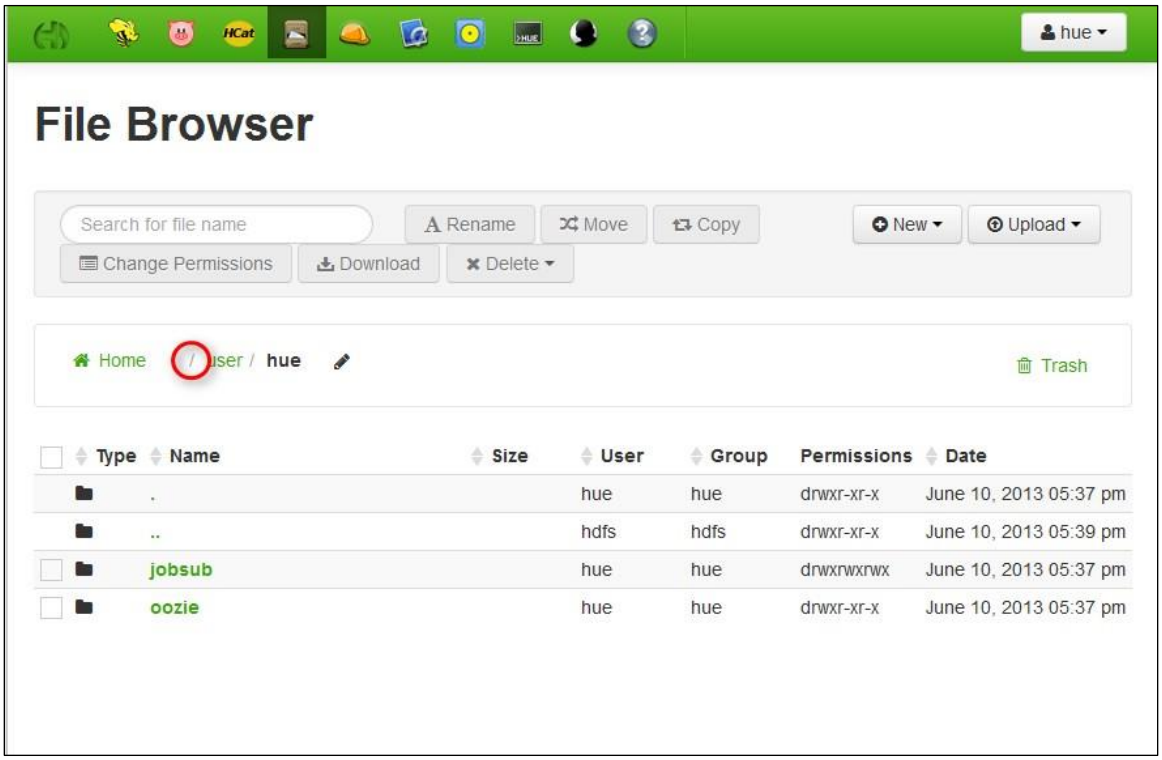
5. When the files have finished uploading to the Sandbox, a **Complete!** message appears at the command prompt.


Now that you have uploaded the Twitter sample data, you can refine the raw data for sentiment analysis.

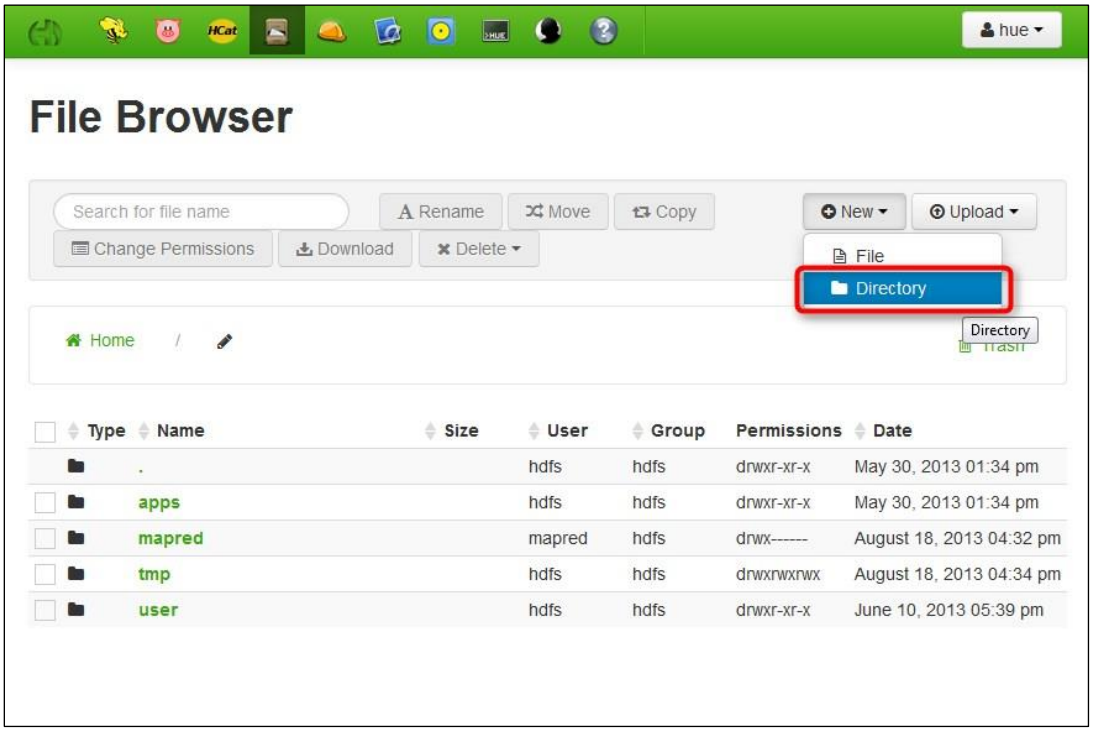
**Load Twitter Data into the Hortonworks Sandbox – Windows 7**

For Windows 7, we will manually upload the files using the Sandbox File Browser.

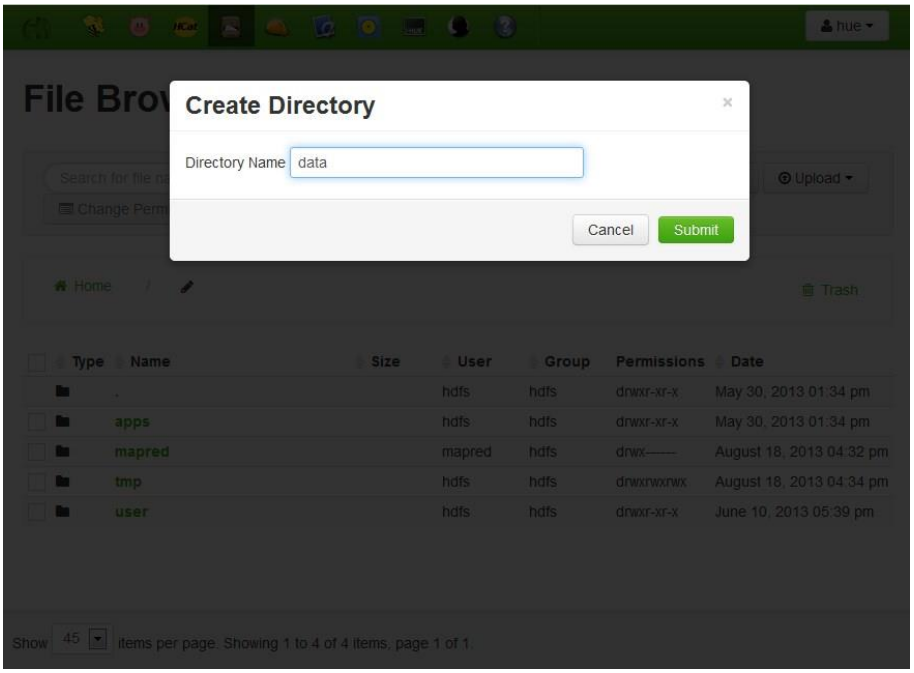
1. Open the Sandbox HUE and click the File Browser icon in the toolbar at the top of the page, then click the first slash to the right of **Home** (and to the left of "user") to navigate to the top level of the Sandbox file system.



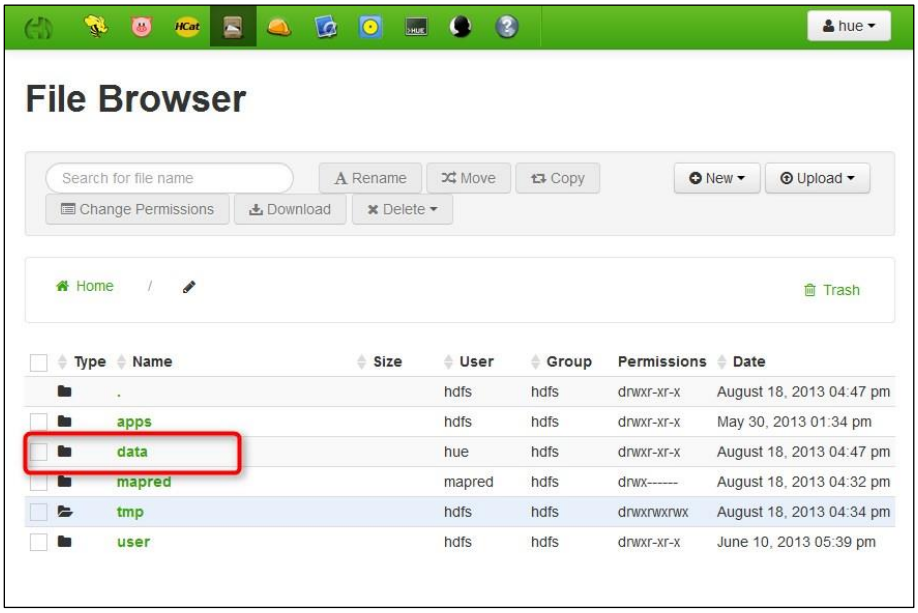
2. Click **New**, then select **Directory**.



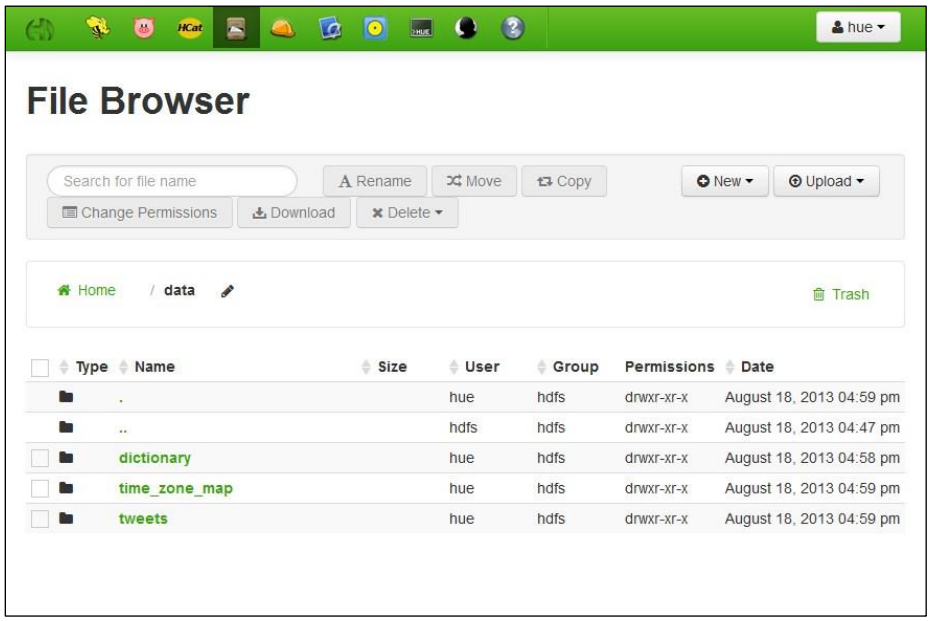

3. On the Create Directory pop-up, type "data" in the Directory Name box, then click **Submit**.



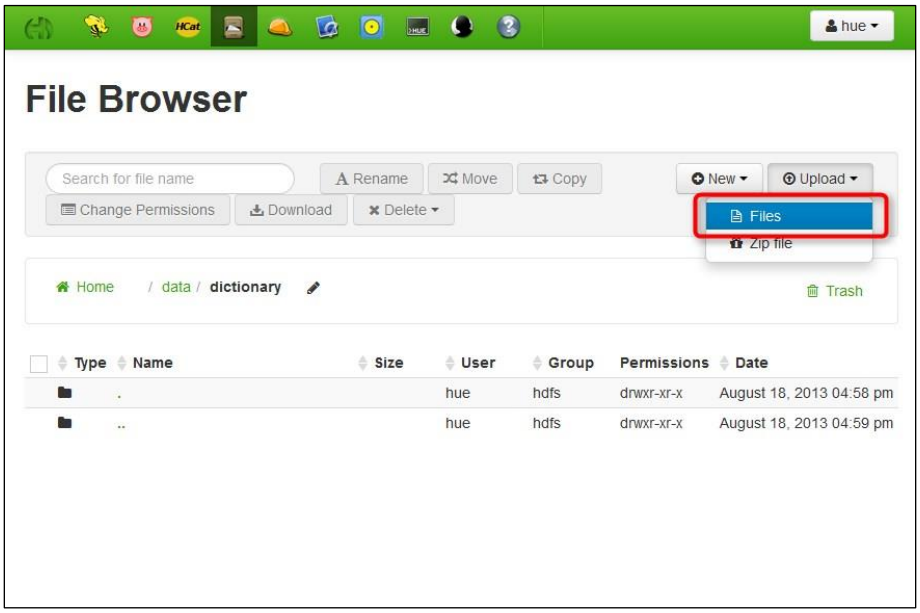
4. The "data" folder appears in the list of files.



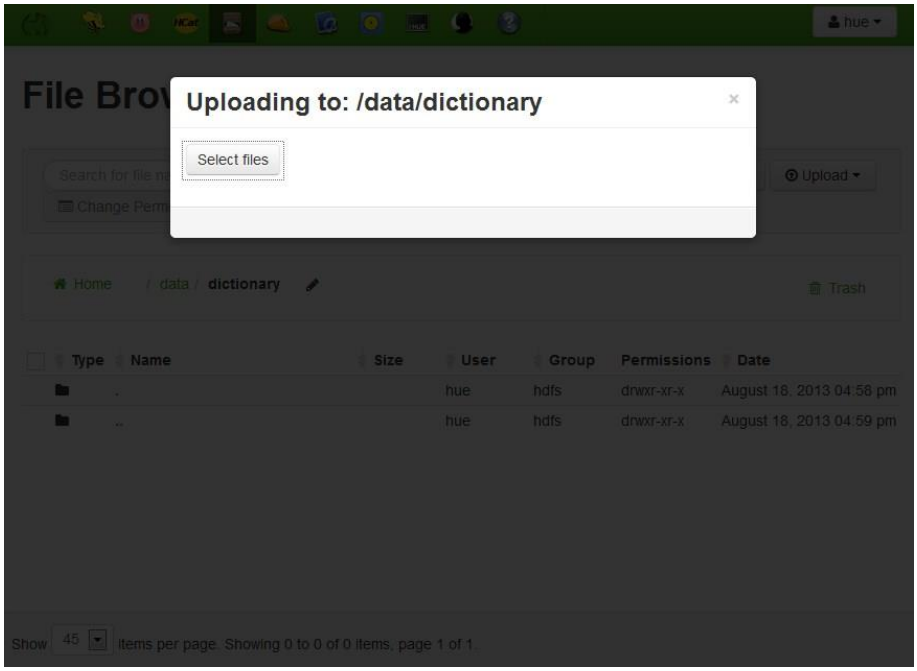

5. Click the "data" folder. In the "data" folder, use **New > Directory** to create three folders with the names "dictionary", "time_zone_map", and "tweets". You must use these exact names.



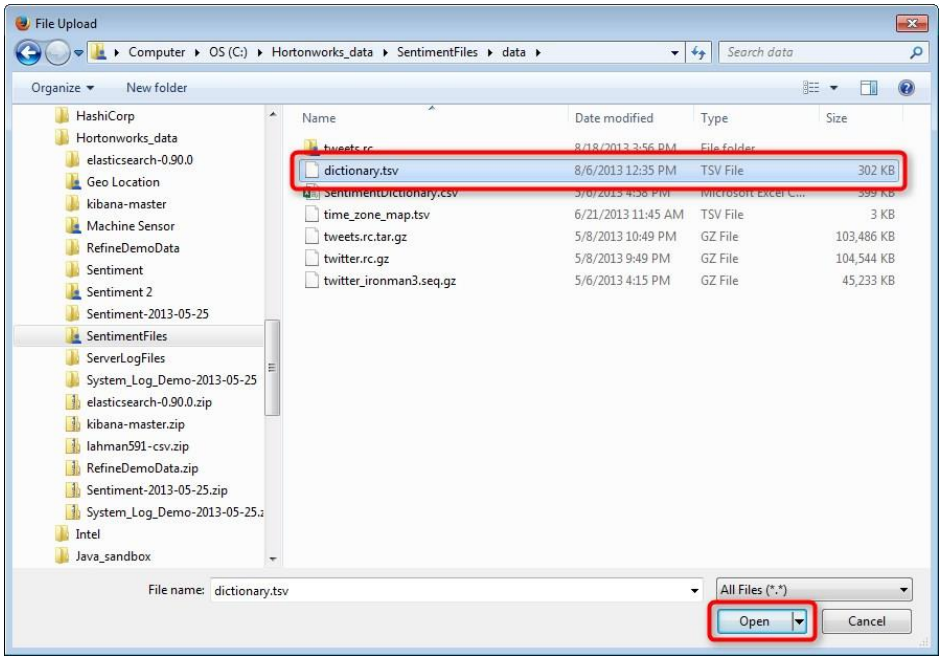
6. Click the "dictionary" folder. In the "dictionary" folder, click **Upload**, then select **Files**.



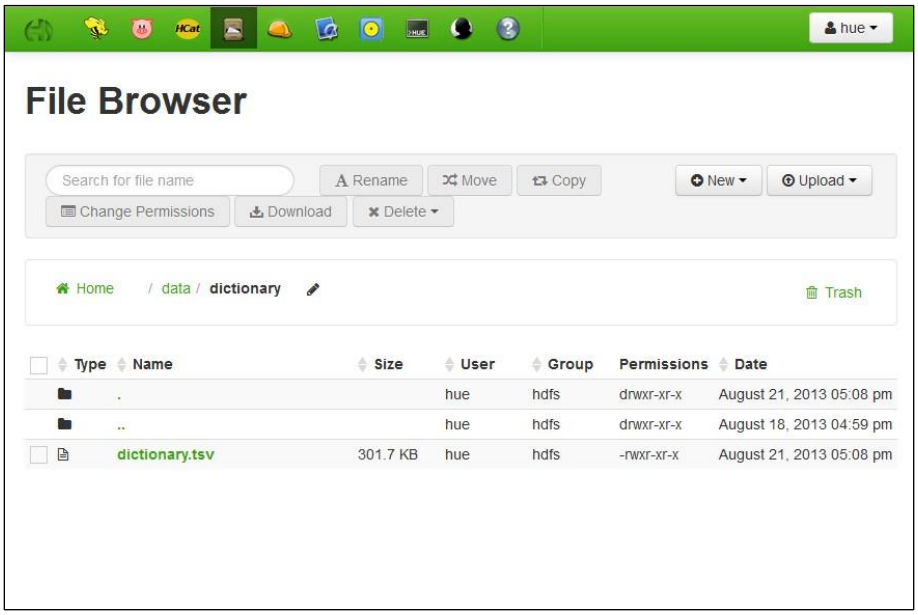

7. On the Uploading pop-up, click **Select files**.



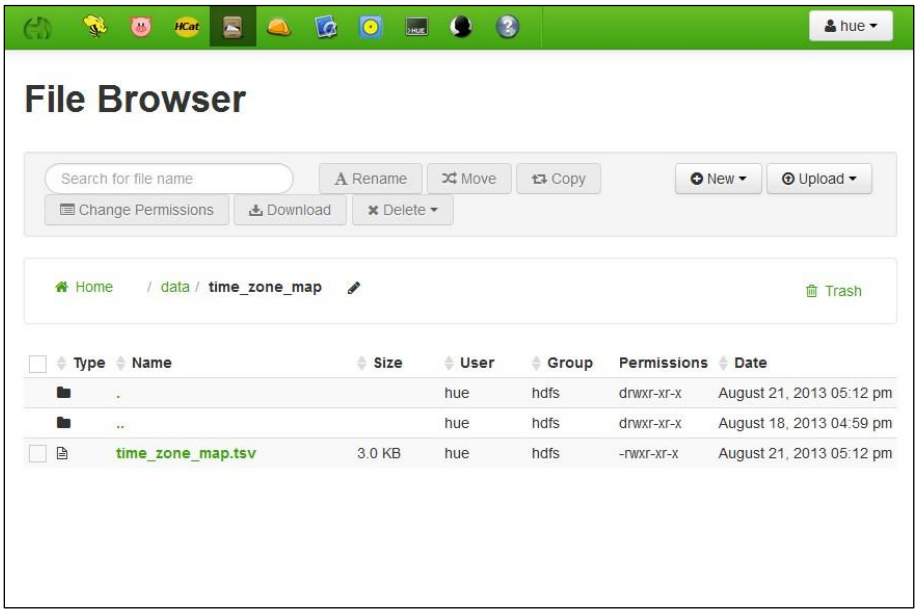

8. Use the File Upload dialog to browse to the "data" folder in the SentimentFiles folder you extracted previously. Select the dictionary.tsv file, then click **Open**.




9. The dictionary.tsv file appears in the /data/dictionary folder.

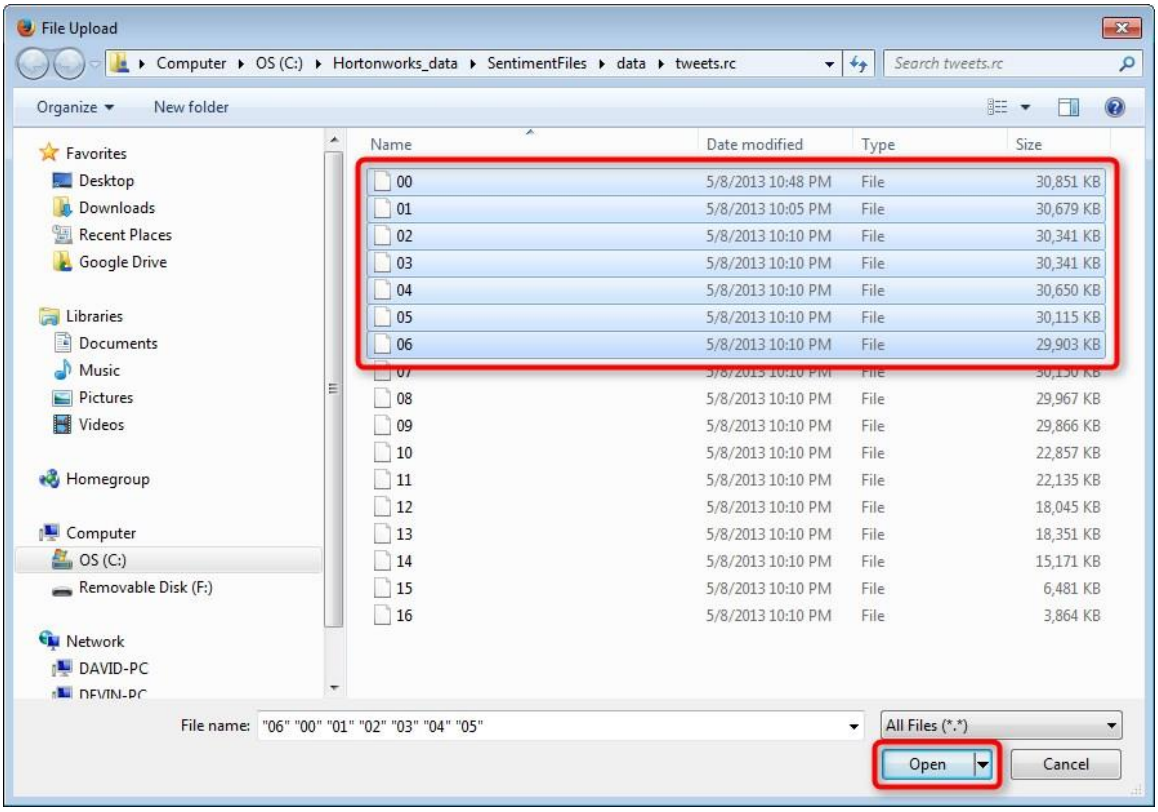


10. Navigate back up to the "data" folder. Use the same procedure to upload the time_zone_map.tsv file to the data/time_zone_map folder.

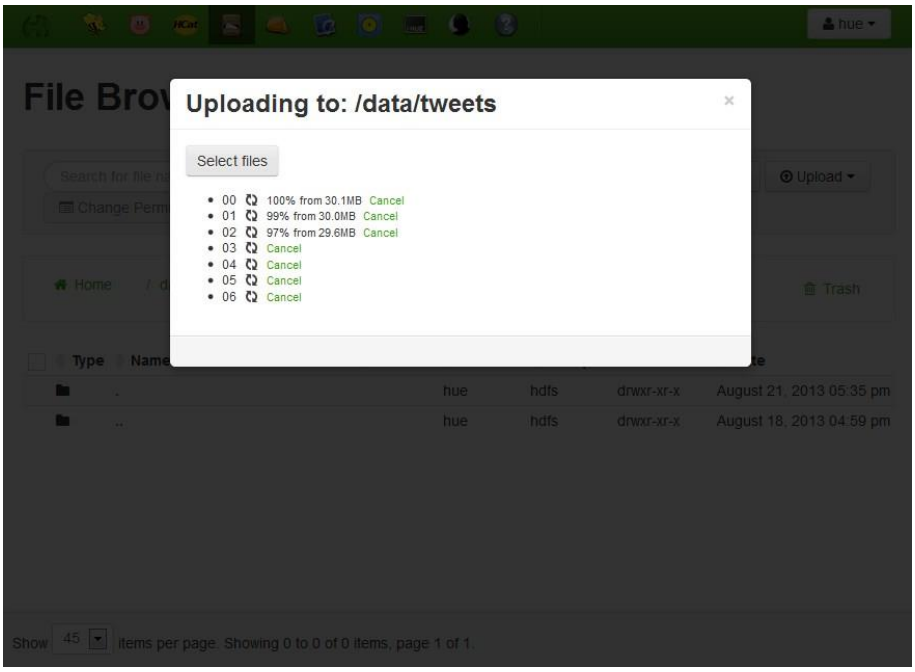



11. Navigate back up to the "data" folder, then click the "tweets" folder. Select **Upload > Files**, then click **Select files**.

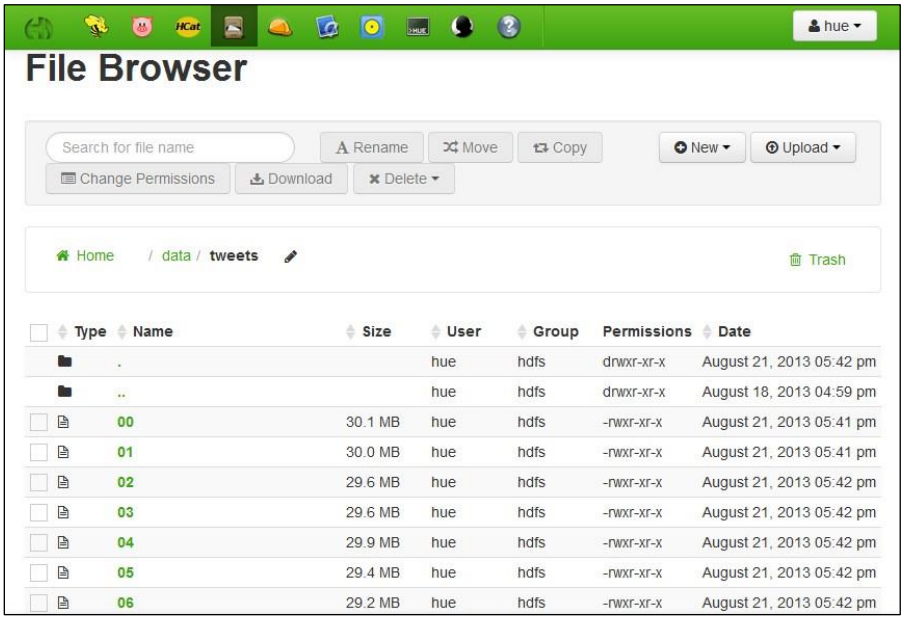
    Use the File Upload dialog to browse to the SentimentFiles\data\tweets.rc folder. Select the "00" file, then Shift-click the "06" file to select files 00-06. Click **Open** to upload the files.


File Upload
Computer ▸ OS (C:) ▸ Hortonworks_data ▸ SentimentFiles ▸ data ▸ tweets.rc
Search tweets.rc
Organize
New folder
Favorites
Desktop
Downloads
Recent Places
Google Drive
Libraries
Documents
Music
Pictures
Videos
Homegroup
Computer
OS (C:)
Removable Disk (F:)
Network
DAVID-PC
Name
Date modified
Type
Size
00 5/8/2013 10:48 PM File 30,851 KB
01 5/8/2013 10:05 PM File 30,679 KB
02 5/8/2013 10:10 PM File 30,341 KB
03 5/8/2013 10:10 PM File 30,341 KB
04 5/8/2013 10:10 PM File 30,650 KB
05 5/8/2013 10:10 PM File 30,115 KB
06 5/8/2013 10:10 PM File 29,903 KB
08 5/8/2013 10:10 PM File 29,967 KB
09 5/8/2013 10:10 PM File 29,866 KB
10 5/8/2013 10:10 PM File 22,857 KB
11 5/8/2013 10:10 PM File 22,135 KB
12 5/8/2013 10:10 PM File 18,045 KB
13 5/8/2013 10:10 PM File 18,351 KB
14 5/8/2013 10:10 PM File 15,171 KB
15 5/8/2013 10:10 PM File 6,481 KB
16 5/8/2013 10:10 PM File 3,864 KB
File name: "06" "00" "01" "02" "03" "04" "05"
All Files (*.*)
Open
Cancel

12. A progress indicator appears while the files are being uploaded.




13. When the upload is complete, files 00-06 are listed in the /data/tweets folder.



## Step 3 – Copy a Hive Script to the Sandbox

We will now use SCP to copy the hiveddl.sql file to the Sandbox. The procedure is slightly different for Windows and Mac, so both methods are described here.

### Copy the hiveddl.sql File to the Sandbox – Mac OS X

1. Open a Terminal window and navigate to "hive" subfolder in the SentimentFiles folder you extracted previously. Type in the following command, then press the Enter key:

   ```
   scp -P 2222 hiveddl.sql root@127.0.0.1:
   ```

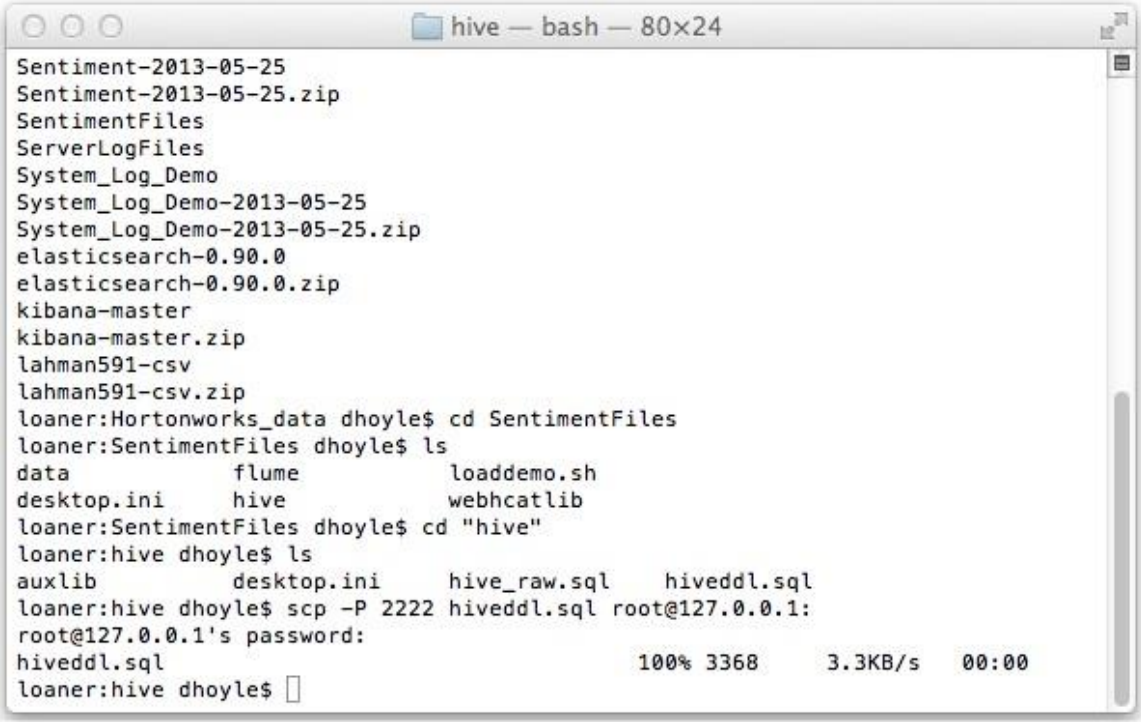
   **Notes:**

   - You must use an uppercase "P" for the "-P" in this command.
   - You may be prompted to validate the authenticity of the host. If so, type "yes" when prompted.

2. When prompted, type in the Sandbox password ("hadoop"), then Press Enter. This command will copy the hiveddl.sql file to the root folder on the Sandbox.

   When the file transfer is complete, a confirmation message will appear in the terminal window:

   ```
   hiveddl.sql                   100% 3368    3.3KB/s     00:00
   ```

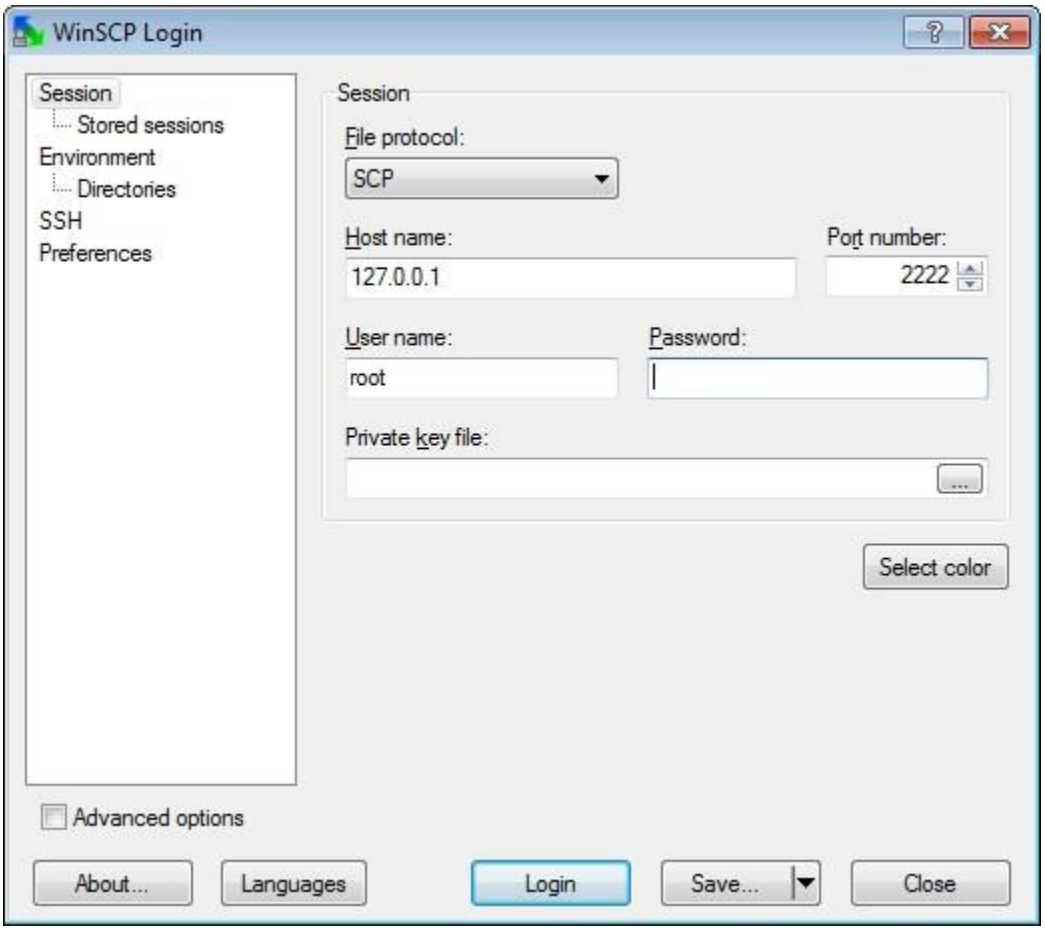


**Copy the hiveddl.sql File to the Sandbox – Windows 7**

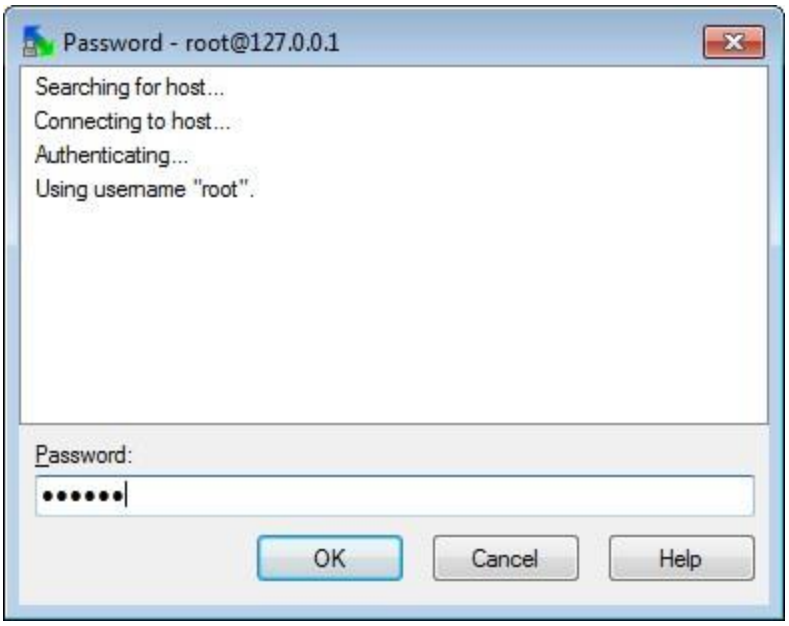
On Windows you must download and install the free WinSCP application.

1. Open WinSCP and type in the following settings, then click **Login**.

   - **Host name:** 127.0.0.1
   - **Port:** 2222
   - **User name:** root

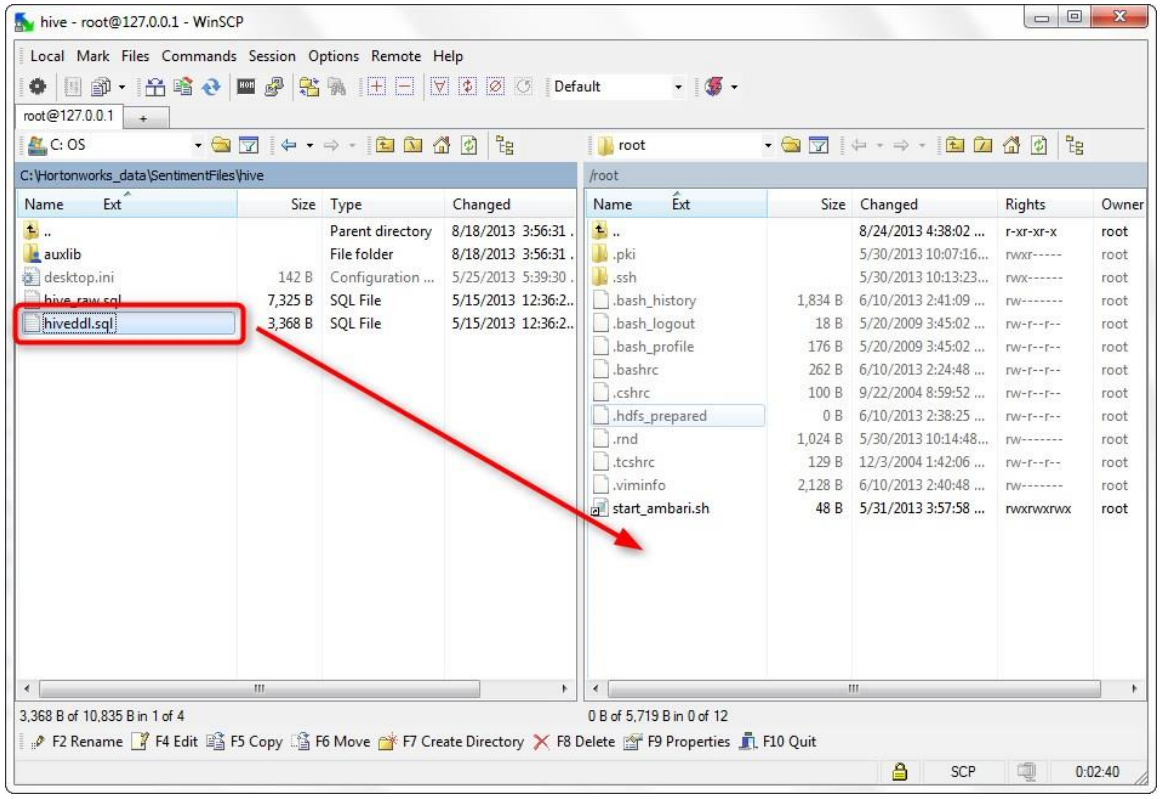



2. Type the Sandbox password ("hadoop") in the Password box, then click **OK**.

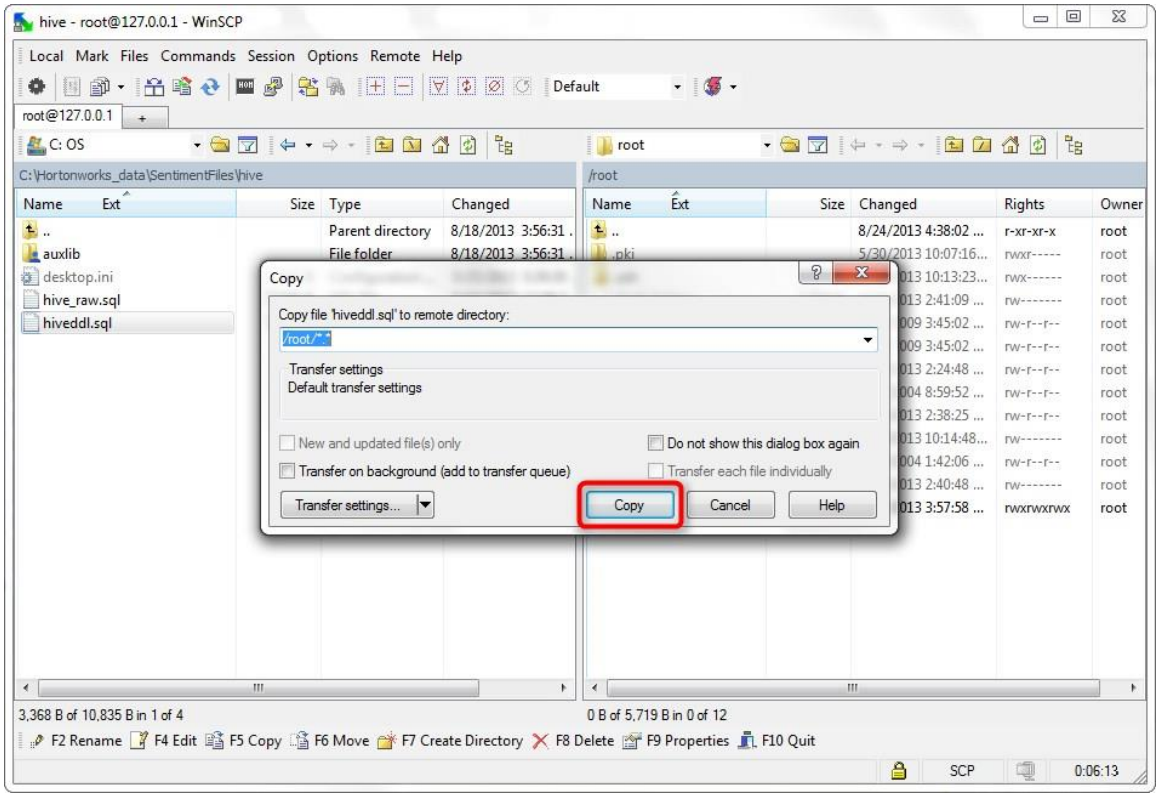



3. Use the WinSCP file browser to navigate to the SentimentFiles\hive folder in the left-hand pane, and to the Sandbox /root folder in the right-hand pane.

   Drag-and-drop the hiveddl.sql file from the SentimentFiles\hive folder to the /root folder on the Sandbox.



Click **Copy** on the Copy pop-up to confirm the file transfer.




## Step 4 – Run the Hive Script to Refine the Raw Data

1. With the Hortonworks Sandbox virtual machine (VM) command prompt window active, press the Alt and F5 keys, then log in to the Sandbox using the following user name and password:

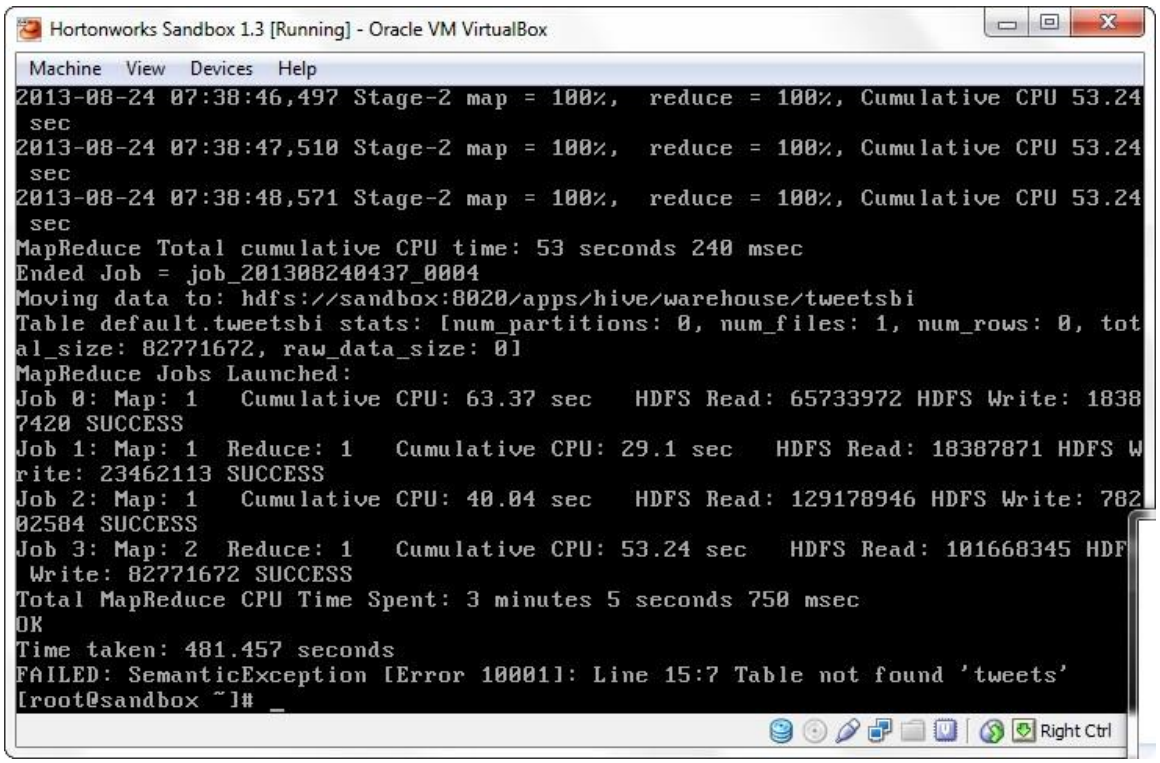
   Login: root
   Password: hadoop

   After you log in, the command prompt will appear with the prefix [root@sandbox ~]#:

2. At the command prompt, type in the following command, then press the Enter key:

   ```
   hive -f hiveddl.sql
   ```

Lines of text appear as the script runs a series of MapReduce jobs. It may take a few minutes for the script to finish running. When the script has finished running, the time taken is displayed, and the normal command prompt appears.

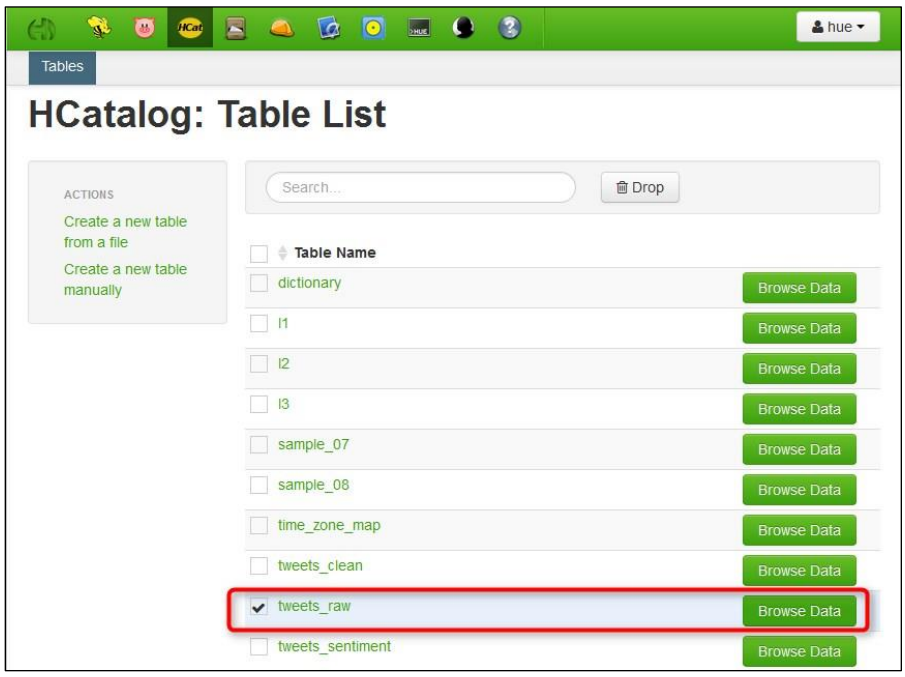



The hiveddl.sql script has performed the following steps to refine the data:

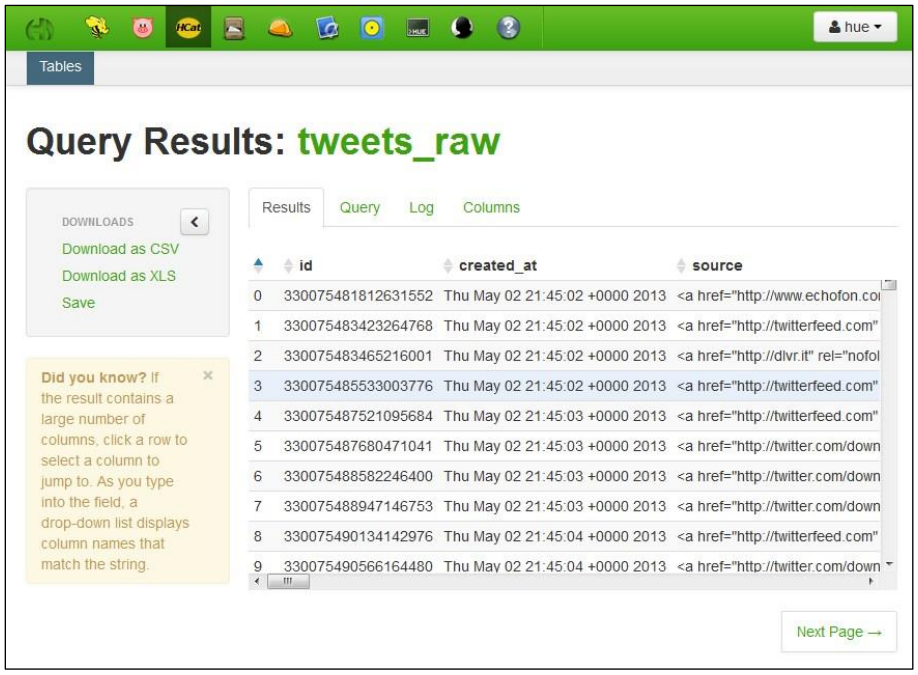
- Converted the raw Twitter data into a tabular format.
- Used the dictionary file to score the sentiment of each Tweet by the number of positive words compared to the number of negative words, and then assigned a positive, negative, or neutral sentiment value to each Tweet.
- Created a new table that includes the sentiment value for each Tweet.

Let's use HCatalog to take a quick look at the data.

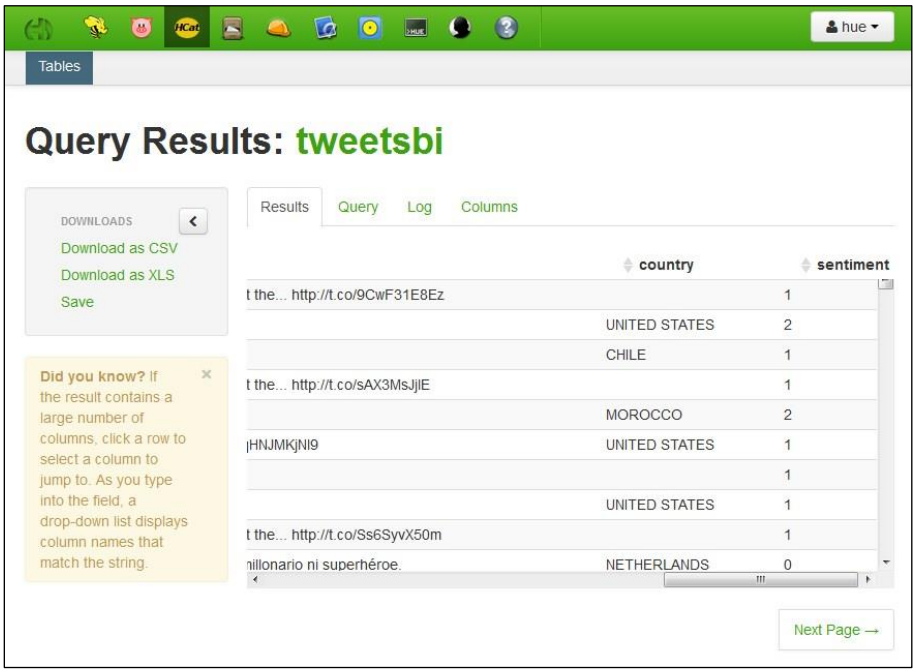
3. Open the Sandbox HUE user interface in a browser, then click **HCatalog** in the menu at the top of the page. Select the check box next to the "tweets_raw" table, then click **Browse Data**.




4. The "tweets_raw" table was created by the Hive script from the raw Twitter data. You should see columns with data for creation time, the number of re-tweets, the tweet content, the user name, and other Twitter data.



5. Click **HCatalog** in the menu at the top of the page, select the check box next to the "tweetsbi" table, and then click **Browse Data**. The "tweetsbi" table is the table created by the Hive script that includes a column with the sentiment value for each tweet.

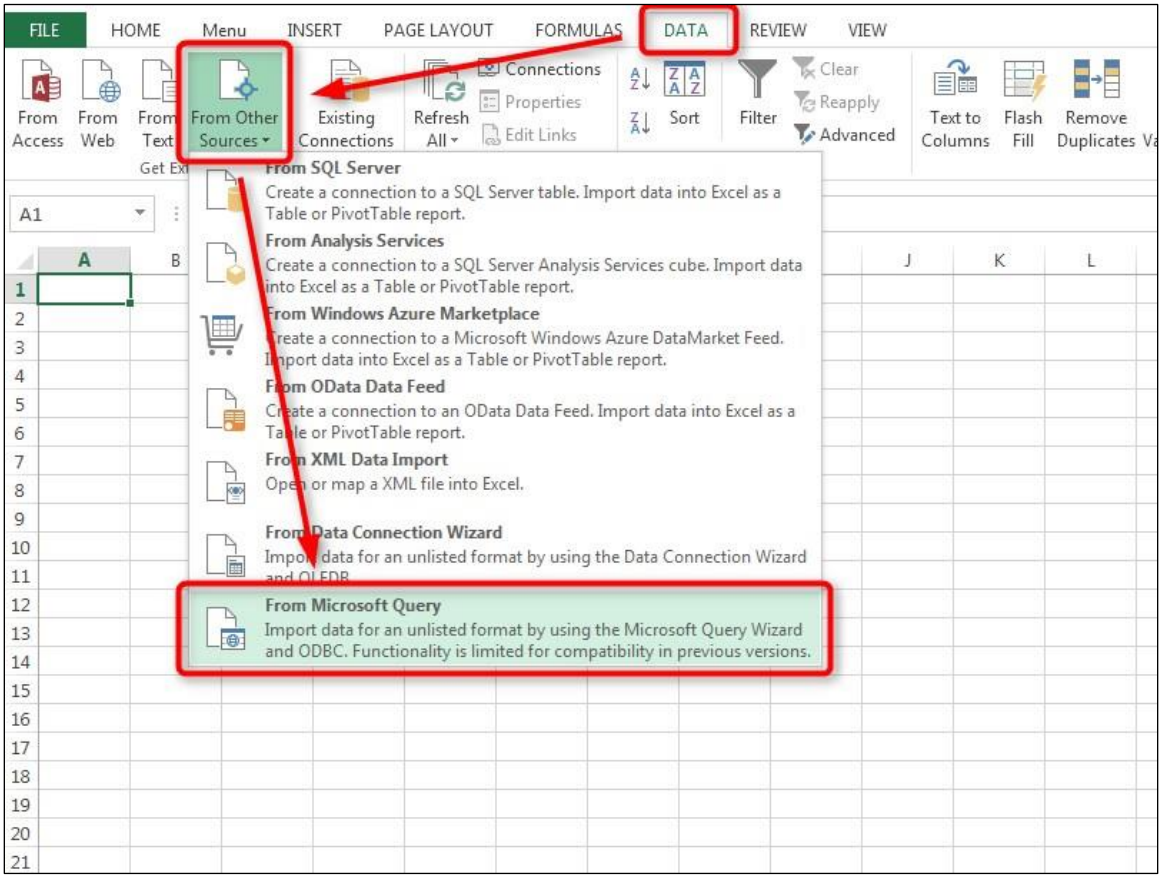



Now that we have refined Twitter data in a tabular format with sentiment ratings, we can access the data with Excel.

## Step 5 – Access the Refined Sentiment Data with Excel

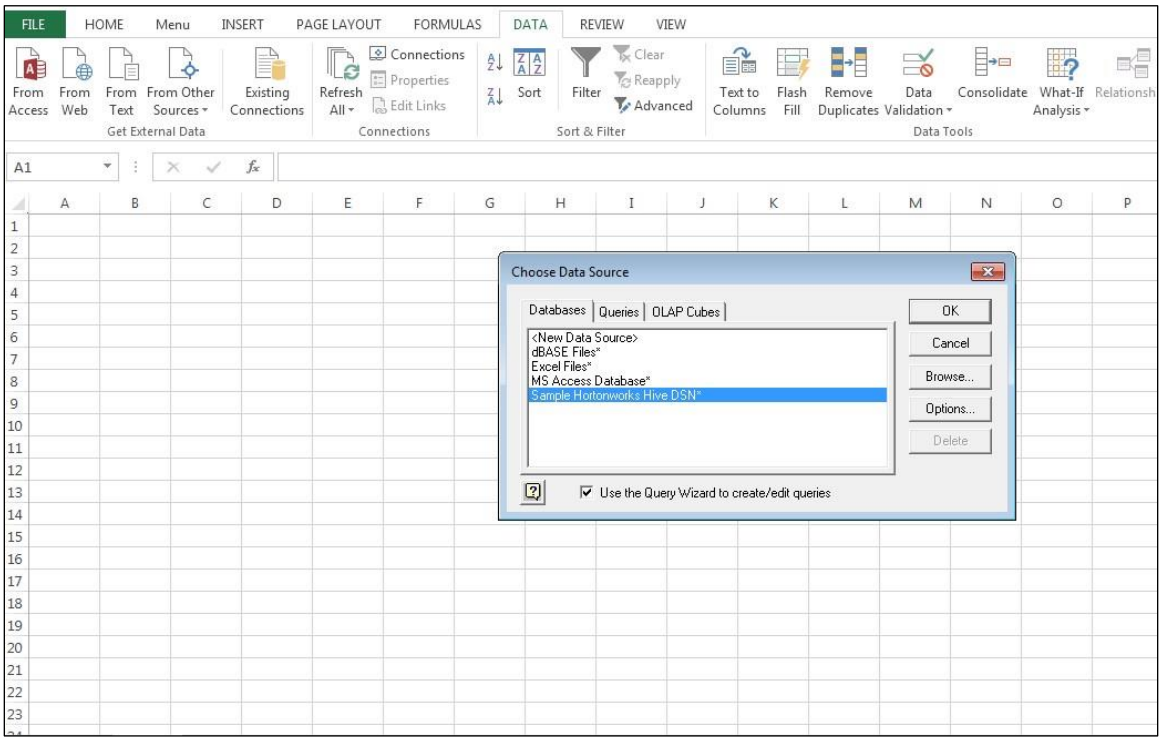
In this section, we use Excel Professional Plus 2013 to access the refined sentiment data.

1. In Windows, open a new Excel workbook, then select **Data > From Other Sources > From Microsoft Query**.

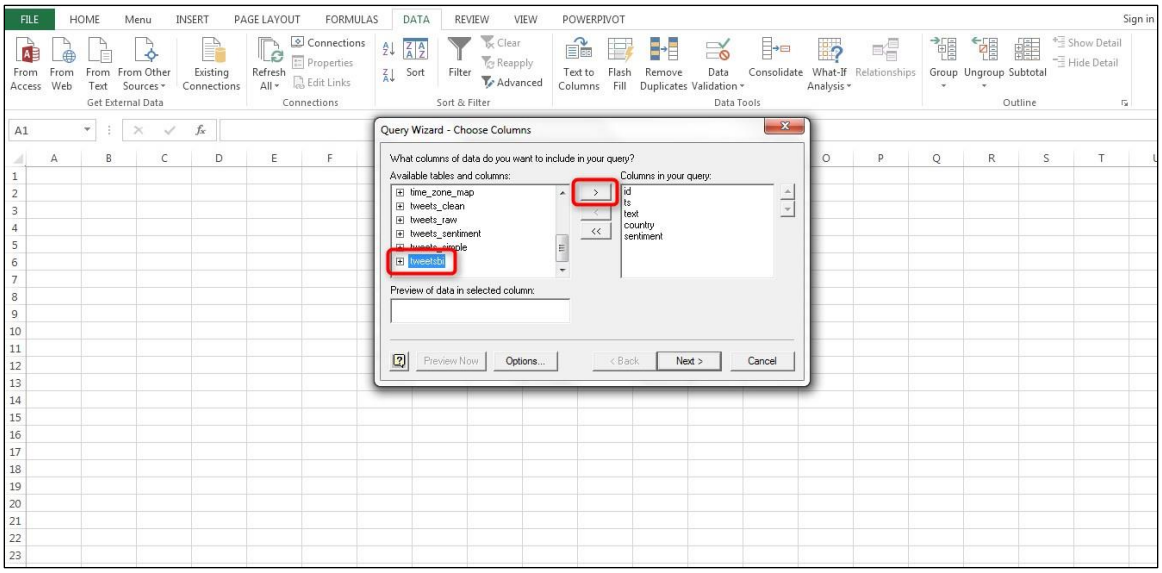


2. On the Choose Data Source pop-up, select the Hortonworks ODBC data source you installed previously, then click **OK**.

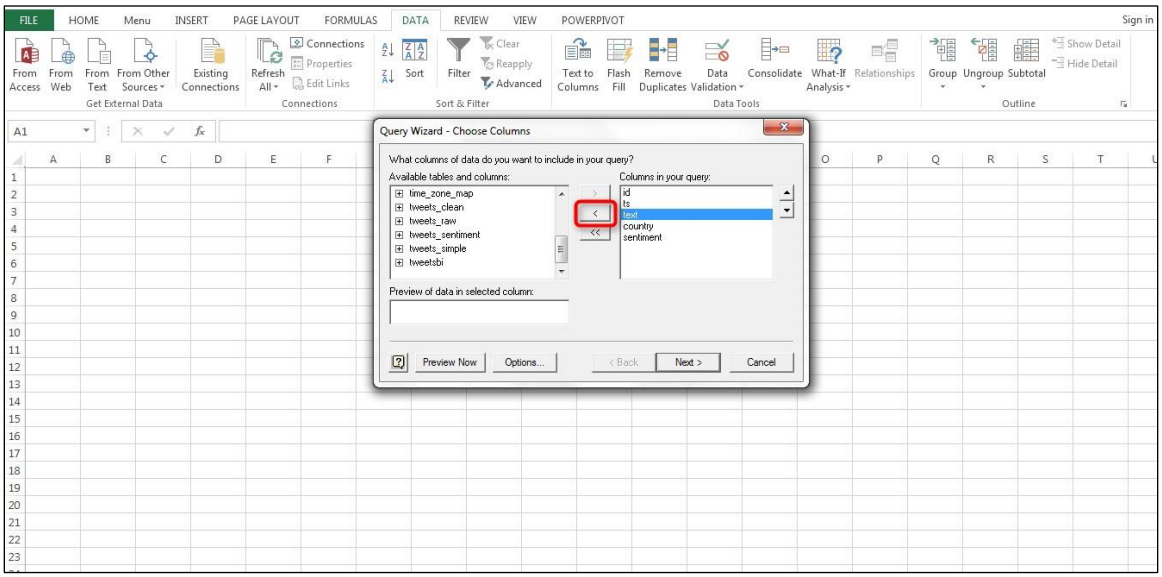
   The Hortonworks ODBC driver enables you to access Hortonworks data with Excel and other Business Intelligence (BI) applications that support ODBC.



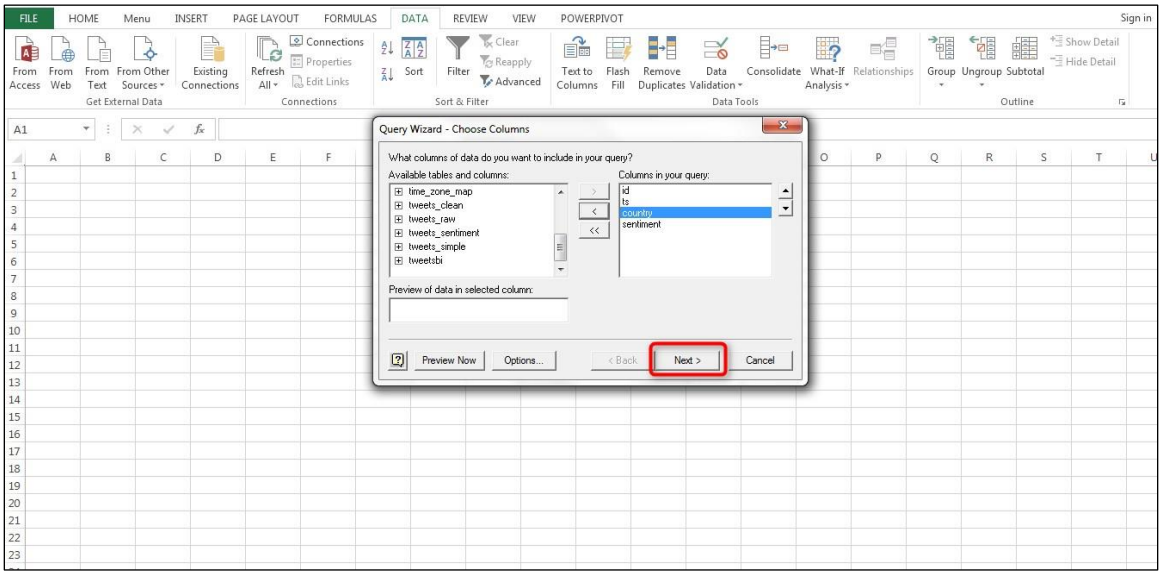
3. After the connection to the Sandbox is established, the Query Wizard appears. Select the "tweetsbi" table in the Available tables and columns box, then click the right arrow button to add the entire "tweetsbi" table to the query. Click **Next** to continue.



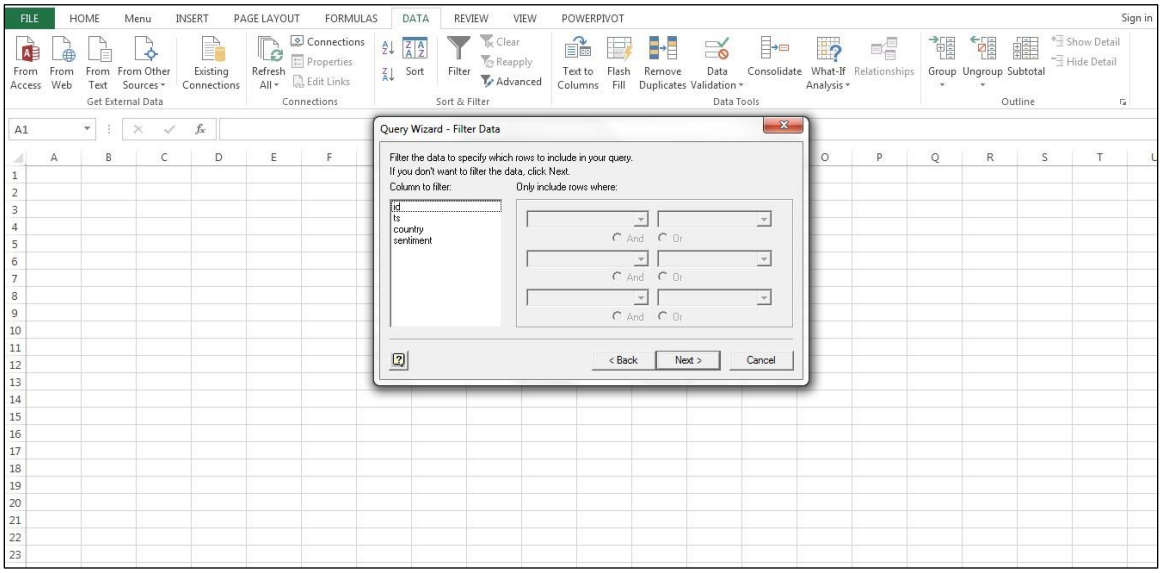

4. Select the "text" column in the "Columns in your query" box, then click the left arrow button to remove the text column.



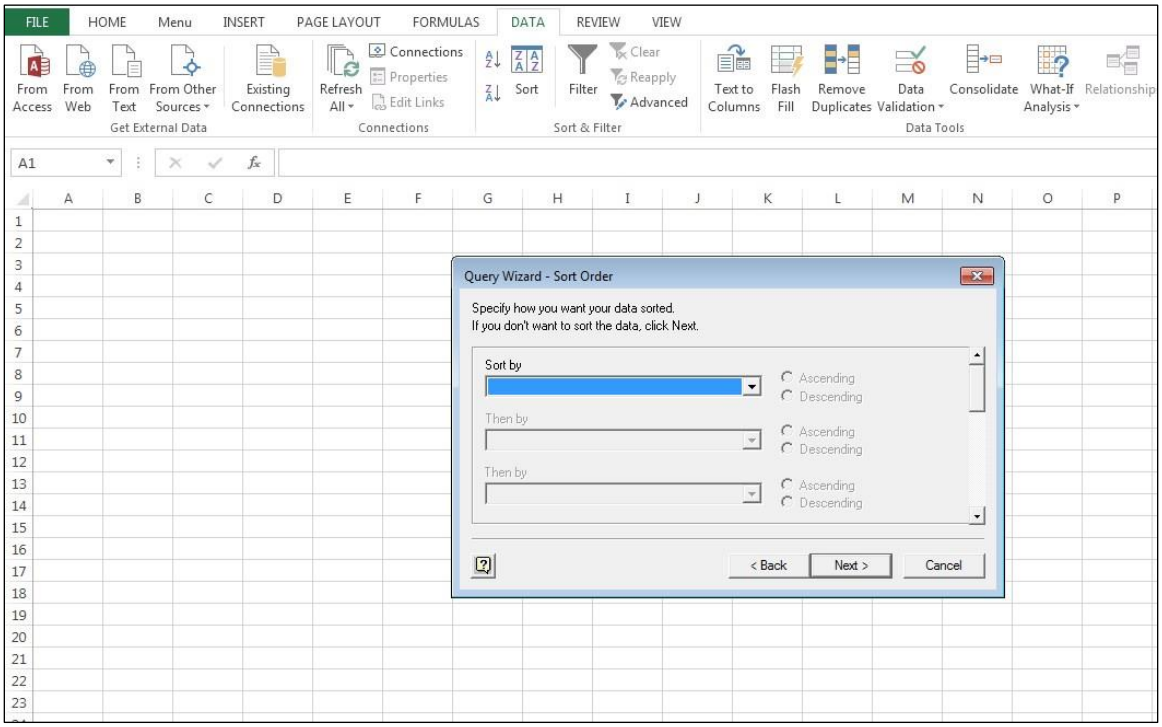
5. After the "text" column has been removed, click **Next** to continue.



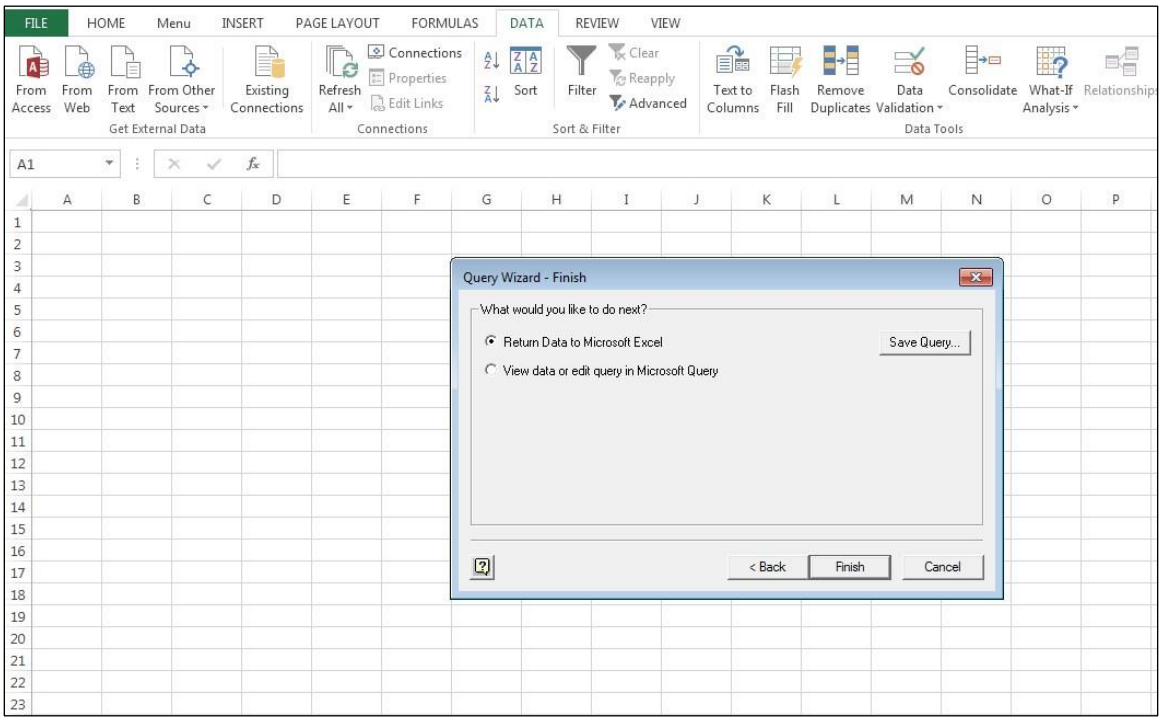

6. On the Filter Data screen, click **Next** to continue without filtering the data.



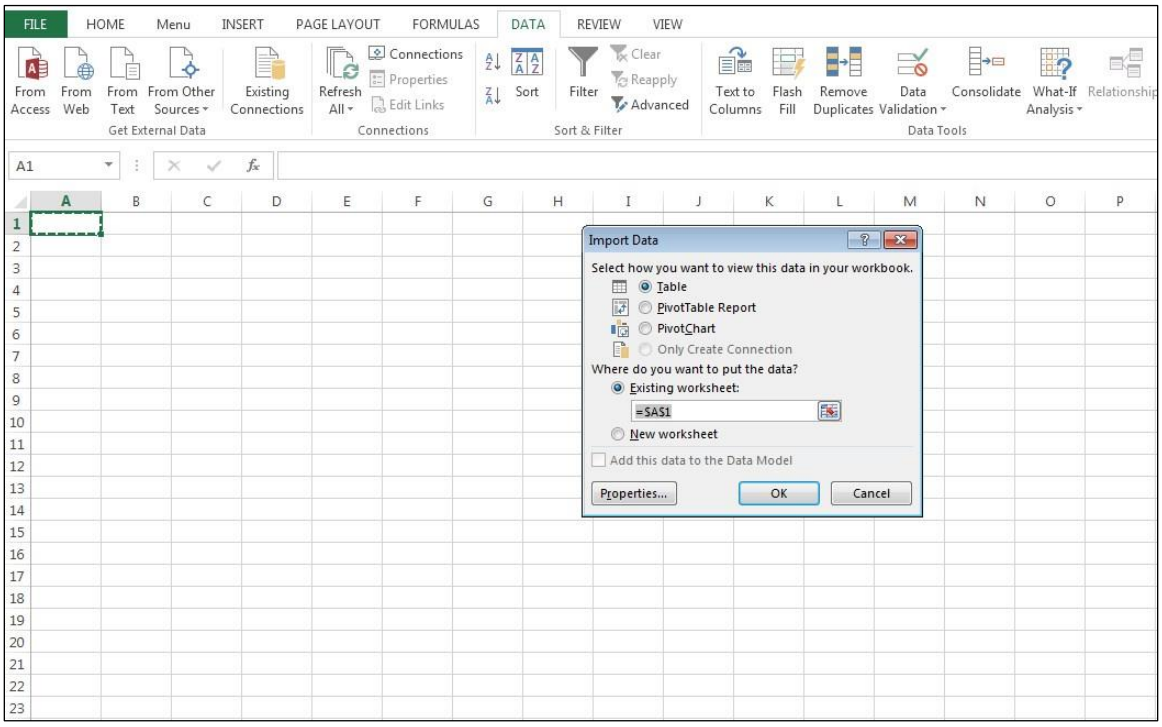
7. On the Sort Order screen, click **Next** to continue without setting a sort order.



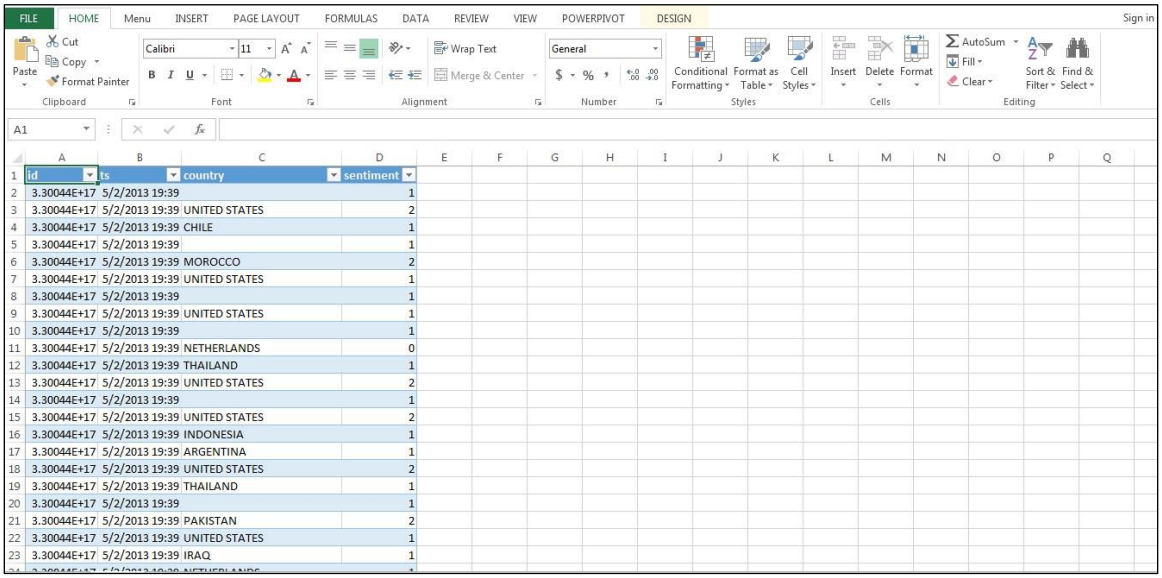

8. Click **Finish** on the Query Wizard Finish screen to retrieve the query data from the Sandbox and import it into Excel.



9. On the Import Data dialog box, click **OK** to accept the default settings and import the data as a table.




10. The imported query data appears in the Excel workbook.



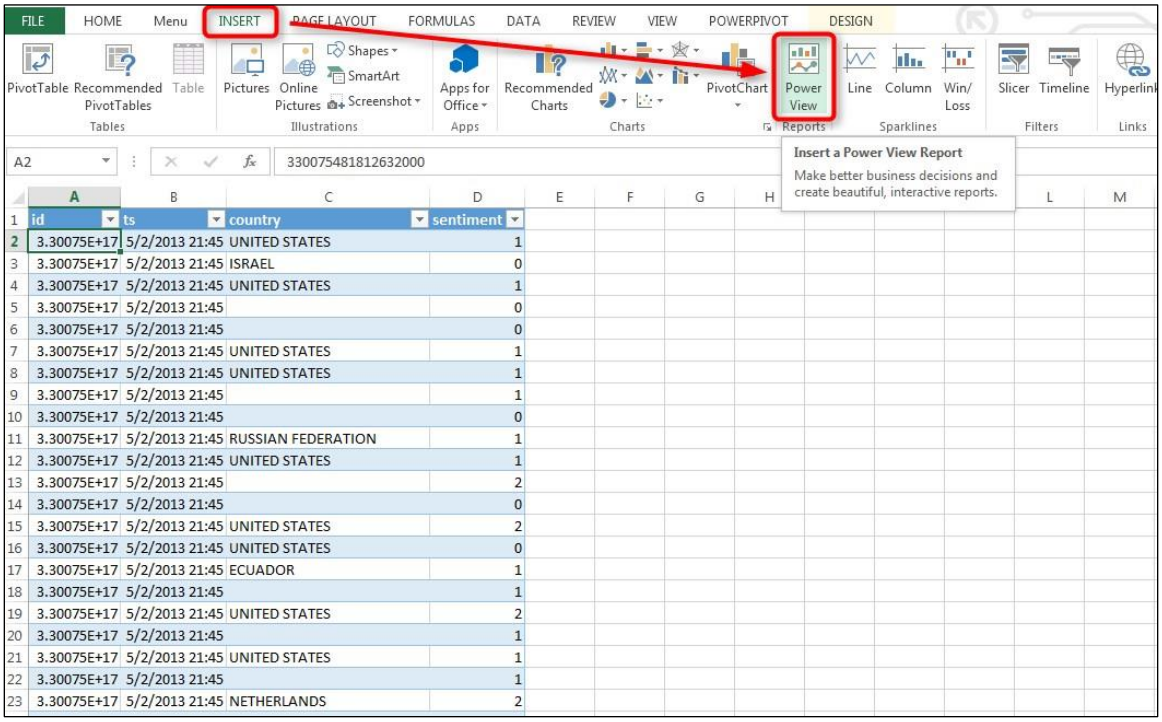
| id | ts | country | sentiment |
|---|---|---|---|
| 3.30044E+17 | 5/2/2013 19:39 | | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 2 |
| 3.30044E+17 | 5/2/2013 19:39 | CHILE | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | MOROCCO | 2 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | NETHERLANDS | 0 |
| 3.30044E+17 | 5/2/2013 19:39 | THAILAND | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 2 |
| 3.30044E+17 | 5/2/2013 19:39 | | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 2 |
| 3.30044E+17 | 5/2/2013 19:39 | INDONESIA | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | ARGENTINA | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 2 |
| 3.30044E+17 | 5/2/2013 19:39 | THAILAND | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | PAKISTAN | 2 |
| 3.30044E+17 | 5/2/2013 19:39 | UNITED STATES | 1 |
| 3.30044E+17 | 5/2/2013 19:39 | IRAQ | 1 |

Now that we have successfully imported the Twitter sentiment data into Microsoft Excel, we can use the Excel Power View feature to analyze and visualize the data.

## Step 6 – Visualize the Sentiment Data Using Excel Power View

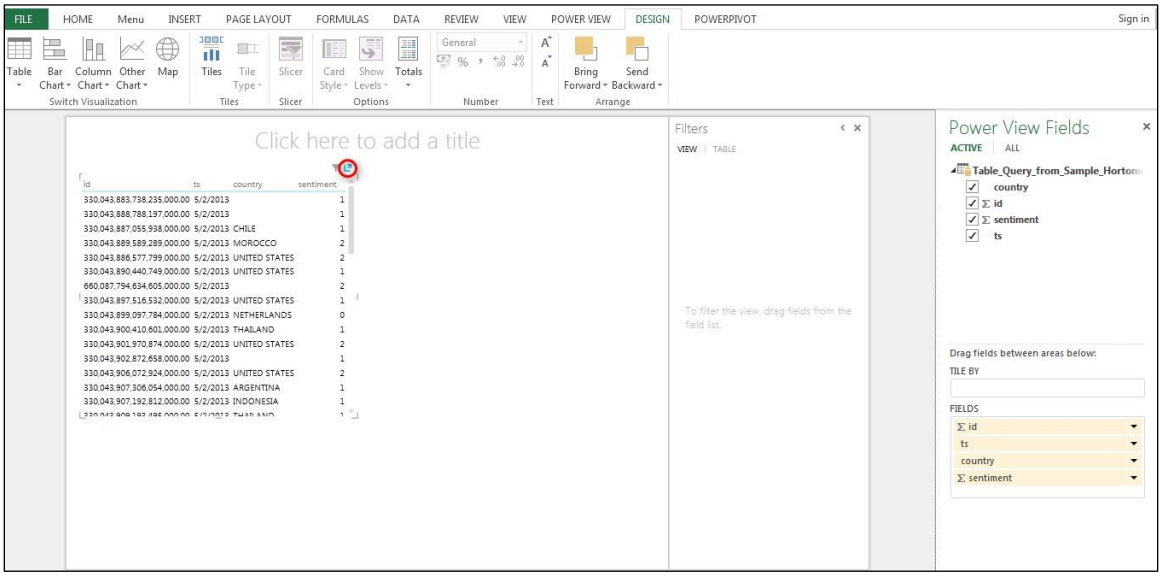
Data visualization can help you optimize your website and convert more visits into sales and revenue. In this section we will see how sentiment varies by country, and review the sentiment data for the United States.

1. In the Excel worksheet with the imported "tweetsbi" table, select **Insert > Power View** to open a new Power View report.

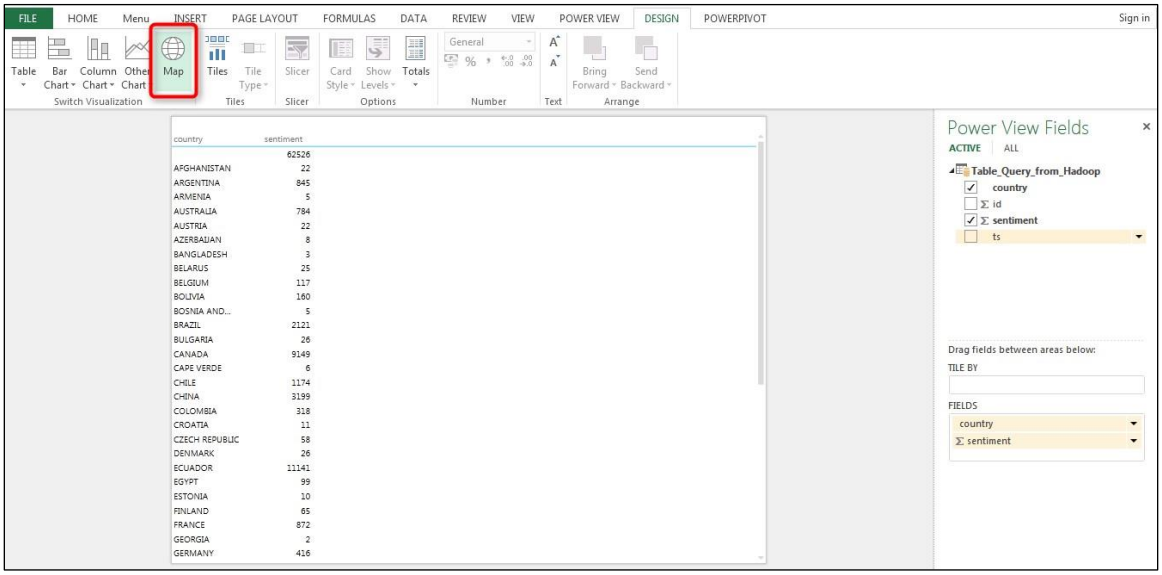



| | A | B | C | D |
|---|---|---|---|---|
| 1 | id | ts | country | sentiment |
| 2 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 1 |
| 3 | 3.30075E+17 | 5/2/2013 21:45 | ISRAEL | 0 |
| 4 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 1 |
| 5 | 3.30075E+17 | 5/2/2013 21:45 | | 0 |
| 6 | 3.30075E+17 | 5/2/2013 21:45 | | 0 |
| 7 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 1 |
| 8 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 1 |
| 9 | 3.30075E+17 | 5/2/2013 21:45 | | 1 |
| 10 | 3.30075E+17 | 5/2/2013 21:45 | | 0 |
| 11 | 3.30075E+17 | 5/2/2013 21:45 | RUSSIAN FEDERATION | 1 |
| 12 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 1 |
| 13 | 3.30075E+17 | 5/2/2013 21:45 | | 2 |
| 14 | 3.30075E+17 | 5/2/2013 21:45 | | 0 |
| 15 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 2 |
| 16 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 0 |
| 17 | 3.30075E+17 | 5/2/2013 21:45 | ECUADOR | 1 |
| 18 | 3.30075E+17 | 5/2/2013 21:45 | | 1 |
| 19 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 2 |
| 20 | 3.30075E+17 | 5/2/2013 21:45 | | 1 |
| 21 | 3.30075E+17 | 5/2/2013 21:45 | UNITED STATES | 1 |
| 22 | 3.30075E+17 | 5/2/2013 21:45 | | 1 |
| 23 | 3.30075E+17 | 5/2/2013 21:45 | NETHERLANDS | 2 |

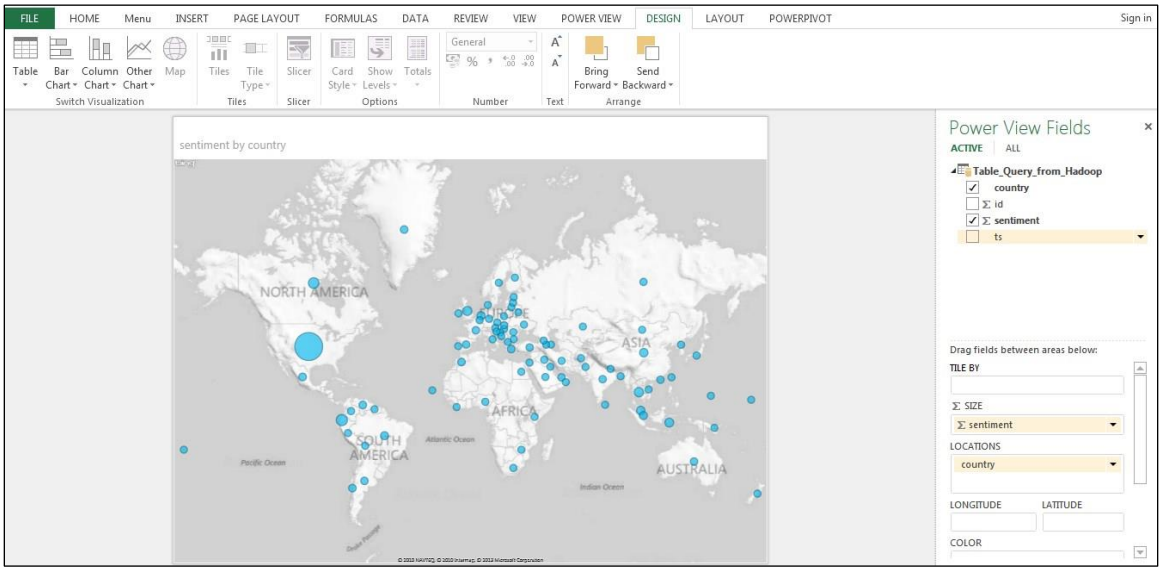
2. The Power View Fields area appears on the right side of the window, with the data table displayed on the left. Drag the handles or click the Pop Out icon to maximize the size of the data table.



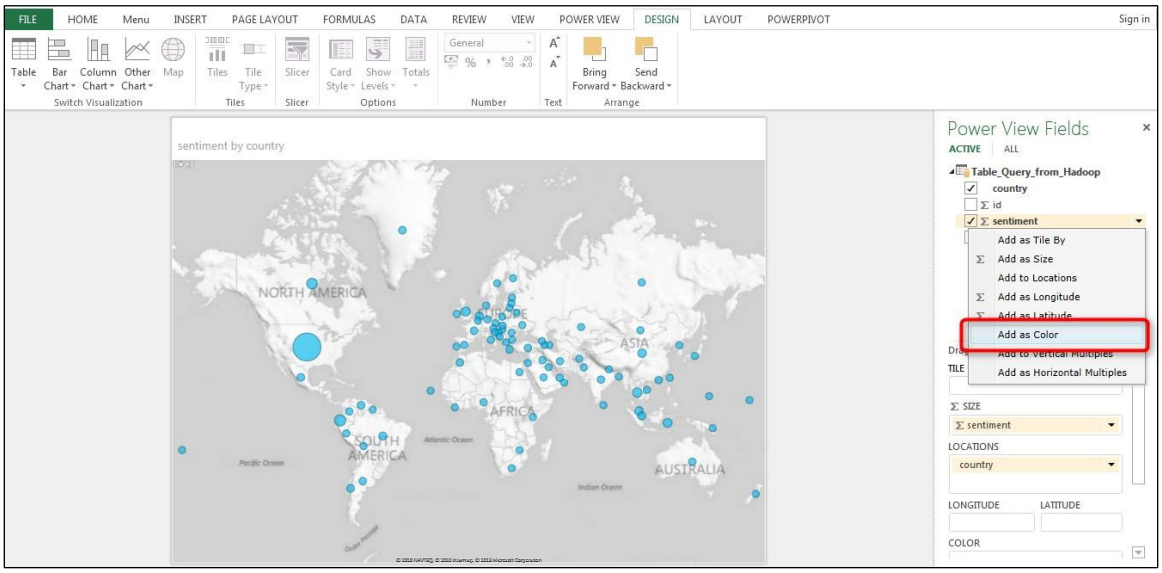

3. In the Power View Fields area, clear the checkboxes next to the **id** and **ts** fields, then click **Map** on the Design tab in the top menu.



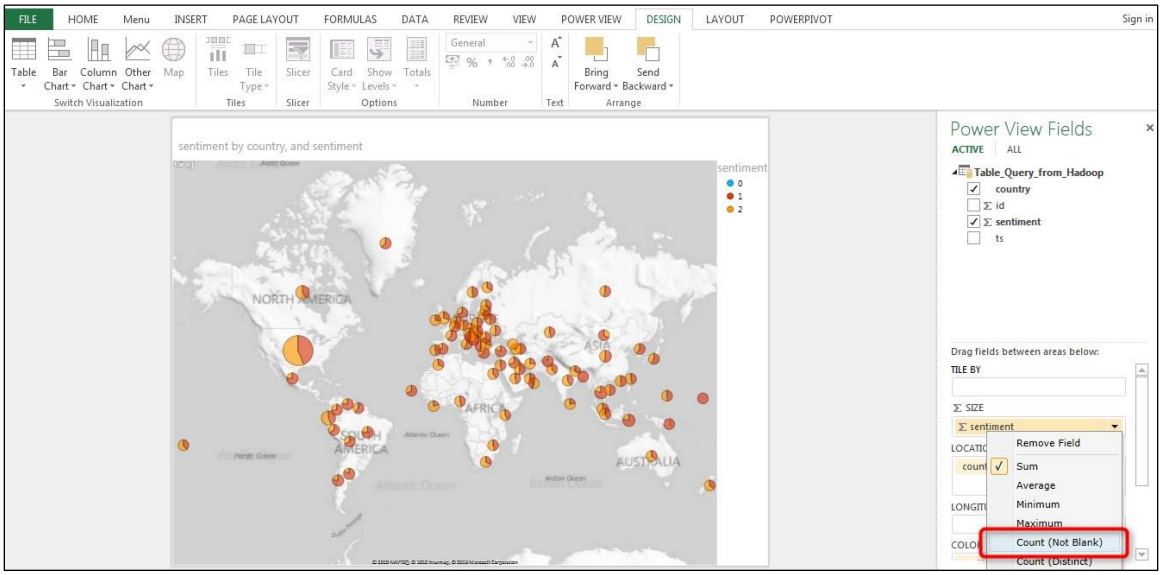
4. The map view displays a global view of the data.




5. Now let's display the sentiment data by color. In the Power View Fields area, click **sentiment**, then select **Add as Color**.

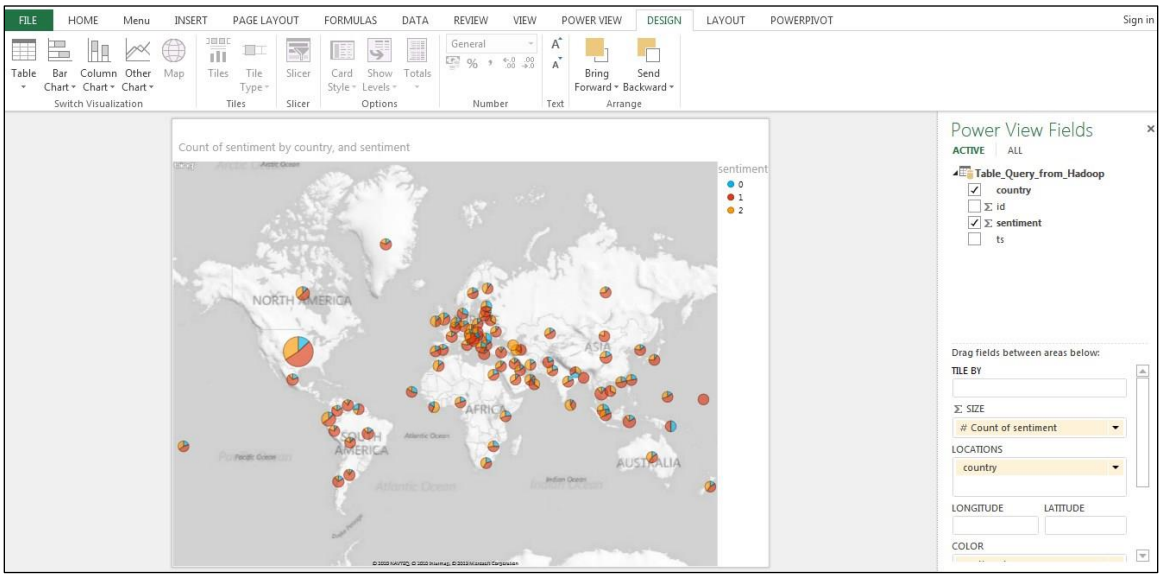


6. Under SIZE, click **sentiment**, then select **Count (Not Blank)**.

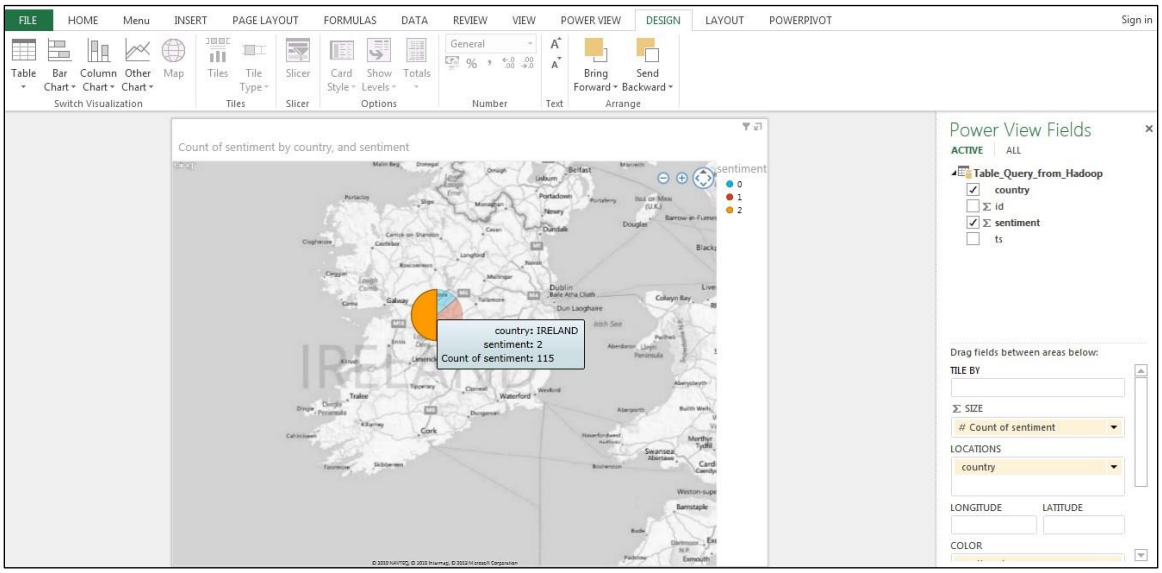



7. Now the map displays the sentiment data by color:

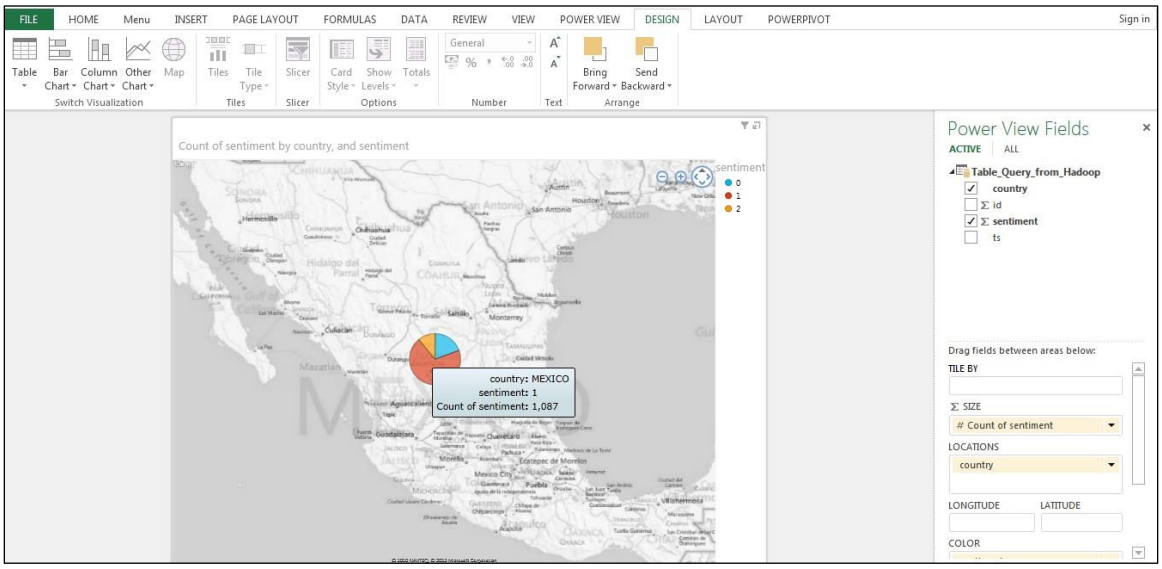
- Orange: positive
- Blue: negative
- Red: neutral



8. Use the map controls to zoom in on Ireland. About half of the tweets have a positive sentiment score, as indicated by the color orange.

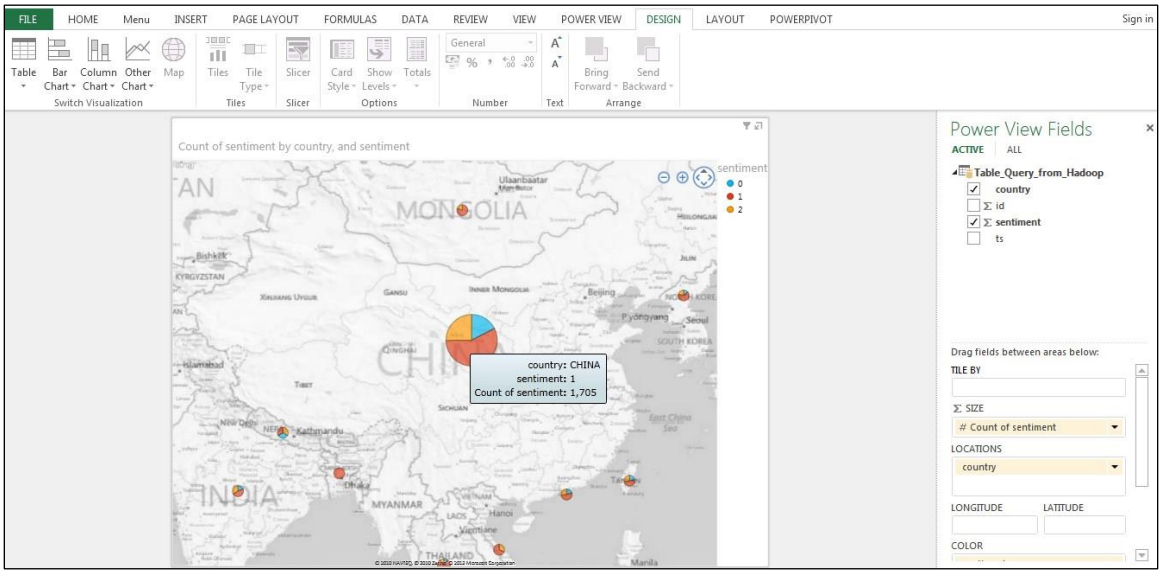



9. Now use the map controls to zoom in on Mexico. In Mexico, about one-fifth of the tweets expressed negative sentiment (shown in blue), and only a small portion of the tweets were positive. Most tweets from Mexico were neutral, as shown in red.



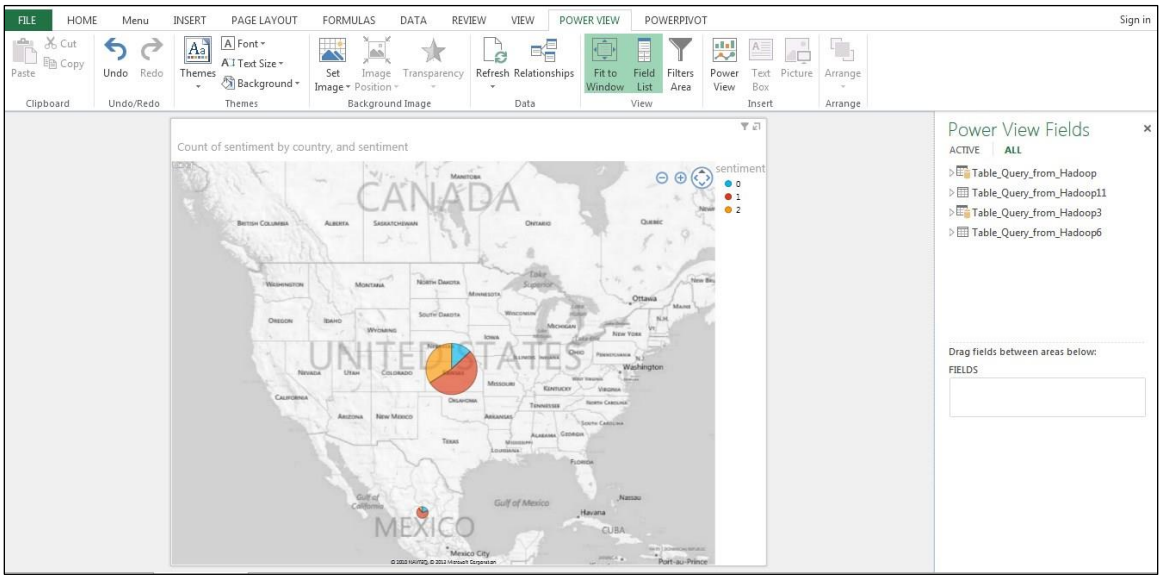
10. Next, use the map controls to zoom in on the sentiment data in China. Marvel studios and the Chinese studio DMG co-financed *Iron Man 3*, and the cast included a famous Chinese actress.

    We can see that the majority of tweets from China are neutral, with positive sentiment slightly outweighing negative sentiment.



11. The United States is the biggest market, so let's look at sentiment data there. The size of the United States pie chart indicates that a relatively large number of the total tweets come from the US.

    About half of the tweets in the US show neutral sentiment, with a relatively small amount of negative sentiment.




We've shown a visualization of Twitter sentiment data for the release of *Iron Man 3*. This information will be useful for planning marketing activities for any future *Iron Man* movie releases.