

# 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 (HDFS).

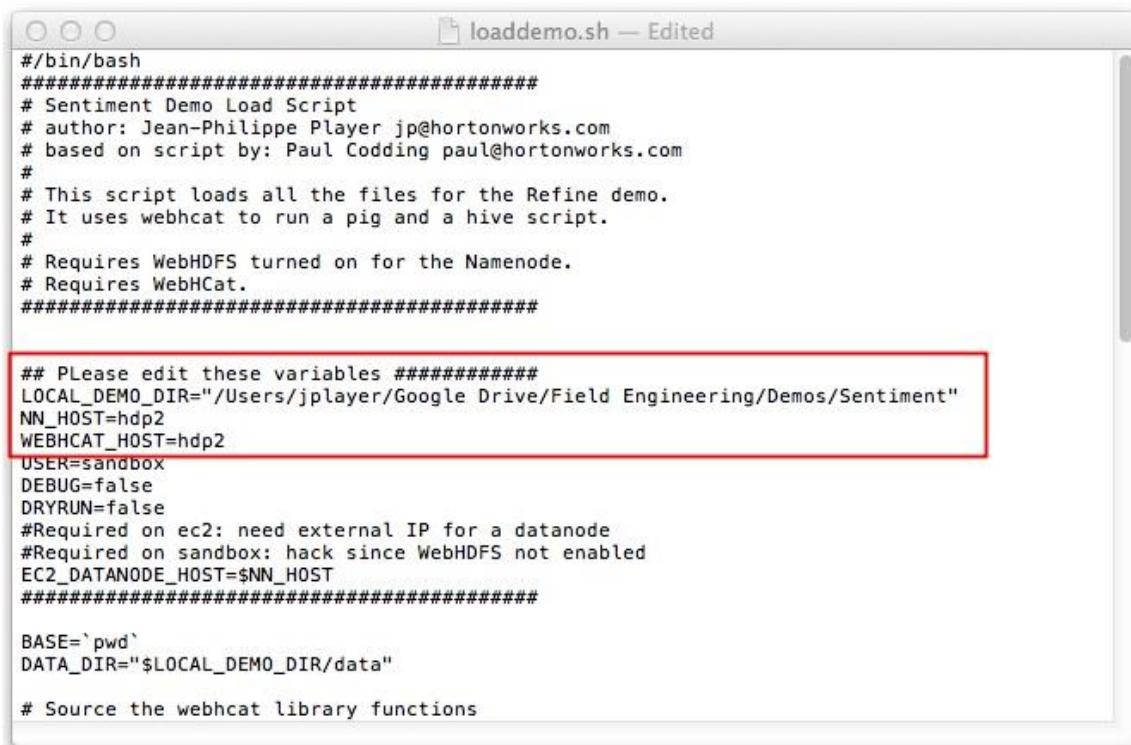
2. Save the SentimentFiles.zip file to your computer, then extract the files. You should see a SentimentFiles folder that contains a loaddemo.sh file.

## 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 usingTextEdit in plain text mode. Near the beginning of the file, you will see an area of text labeled "Please edit these variables."



```
#!/bin/bash
#####
# Sentiment Demo Load Script
# author: Jean-Philippe Player jp@hortonworks.com
# based on script by: Paul Coddington paul@hortonworks.com
#
# This script loads all the files for the Refine demo.
# It uses webhcatt to run a pig and a hive script.
#
# Requires WebHDFS turned on for the Namenode.
# Requires WebHCat.
#####

## Please edit these variables #####
LOCAL_DEMO_DIR="/Users/jplayer/Google Drive/Field Engineering/Demos/Sentiment"
NN_HOST=hdp2
WEBHCAT_HOST=hdp2
USER=sandbox
DEBUG=false
DRYRUN=false
#Required on ec2: need external IP for a datanode
#Required on sandbox: hack since WebHDFS not enabled
EC2_DATANODE_HOST=$NN_HOST
#####

BASE=`pwd`
DATA_DIR="$LOCAL_DEMO_DIR/data"

# Source the webhcatt library functions
```

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.

```
#!/bin/bash
#####
# Sentiment Demo Load Script
# author: Jean-Philippe Player jp@hortonworks.com
# based on script by: Paul Codding paul@hortonworks.com
#
# This script loads all the files for the Refine demo.
# It uses webhcatt to run a pig and a hive script.
#
# Requires WebHDFS turned on for the Namenode.
# Requires WebHCat.
#####

## Please edit these variables #####
LOCAL_DEMO_DIR="/Users/dhoyle/Documents/Hortonworks_data/SentimentFiles"
NN_HOST=127.0.0.1
WEBHCAT_HOST=127.0.0.1
USER=sandbox
DEBUG=false
DRYRUN=false
#Required on ec2: need external IP for a datanode
#Required on sandbox: hack since WebHDFS not enabled
EC2_DATANODE_HOST=$NN_HOST
#####

BASE=`pwd`
DATA_DIR="$LOCAL_DEMO_DIR/data"

# Source the webhcatt library functions
```

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.

```
desktop.ini      hive          webhcattlib
loaner:SentimentFiles dhoyle$ ./loaddemo.sh
Edit this script to configure WebHDFS location
Current configuration:
  LOCAL_DEMO_DIR="/Users/dhoyle/Documents/Hortonworks_data/SentimentFiles"
  NN_HOST=127.0.0.1
  WEBHCAT_HOST=127.0.0.1
  USER=sandbox
  DEBUG=false
  DRYRUN=false
  EC2_DATANODE_HOST=127.0.0.1

You need WebHDFS enabled for this script to work.
If not, enter a value for EC2_DATANODE_HOST as a workaround.
Press a key to continue or Ctrl-c to quit.

Uploading dictionary.tsv
curl -s -i -X PUT curl -s -S -X PUT -T dictionary.tsv "http://127.0.0.1:50075/webhdfs/v1/data/dictionary/dictionary.tsv?op=CREATE&user.name=hdfs&overwrite=false"
Uploading time_zone_map.tsv
curl -s -i -X PUT curl -s -S -X PUT -T time_zone_map.tsv "http://127.0.0.1:50075/webhdfs/v1/data/time_zone_map/time_zone_map.tsv?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 00
curl -s -i -X PUT curl -s -S -X PUT -T 00 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/00?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 01
curl -s -i -X PUT curl -s -S -X PUT -T 01 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/01?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 02
curl -s -i -X PUT curl -s -S -X PUT -T 02 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/02?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 03
curl -s -i -X PUT curl -s -S -X PUT -T 03 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/03?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 04
curl -s -i -X PUT curl -s -S -X PUT -T 04 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/04?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 05
curl -s -i -X PUT curl -s -S -X PUT -T 05 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/05?op=CREATE&user.name=hdfs&overwrite=false"
Uploading 06
curl -s -i -X PUT curl -s -S -X PUT -T 06 "http://127.0.0.1:50075/webhdfs/v1/data/tweets/06?op=CREATE&user.name=hdfs&overwrite=false"
Uploading hiveddl.sql
curl -s -i -X PUT curl -s -S -X PUT -T hiveddl.sql "http://127.0.0.1:50075/webhdfs/v1/user/sandbox/script/hiveddl.sql?op=CREATE&user.name=hdfs&overwrite=false"

Complete!
loaner:SentimentFiles dhoyle$
```

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

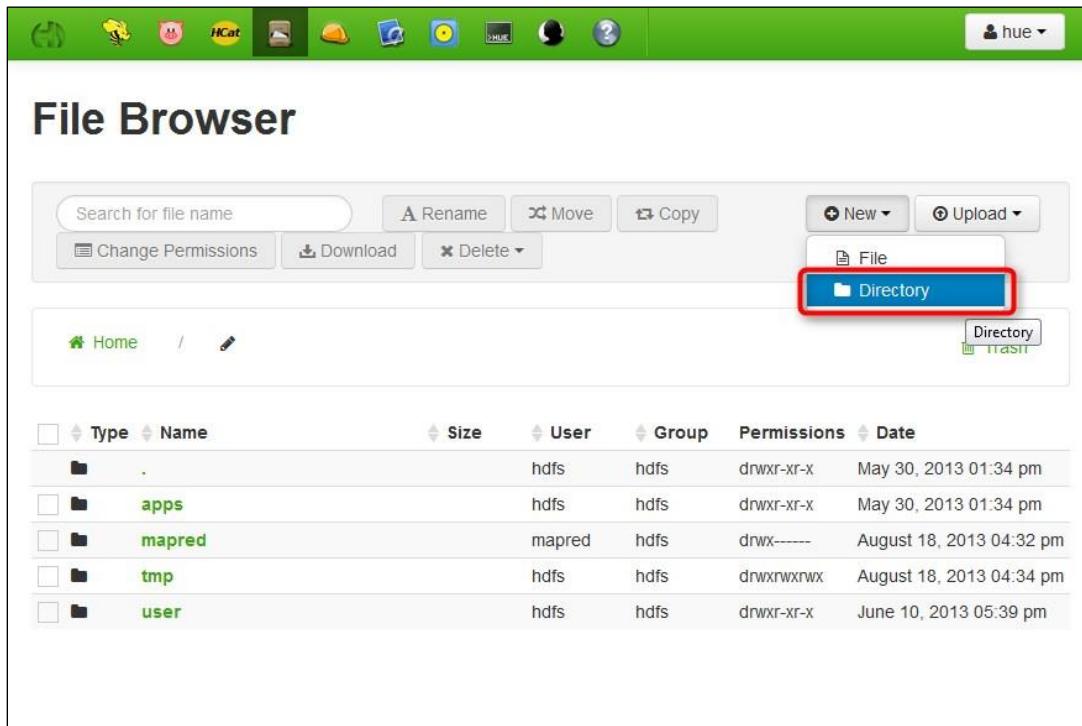
For Windows 7, we will manually upload the files using the Sandbox File Browser.

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.

The screenshot shows the Hortonworks Sandbox File Browser. The URL bar displays the path '/user/hue'. The main area is a table listing directory contents. The columns are: Type, Name, Size, User, Group, Permissions, and Date. The data is as follows:

Type	Name	Size	User	Group	Permissions	Date
Folder	.		hue	hue	drwxr-xr-x	June 10, 2013 05:37 pm
Folder	..		hdfs	hdfs	drwxr-xr-x	June 10, 2013 05:39 pm
Folder	jobsub		hue	hue	drwxrwxrwx	June 10, 2013 05:37 pm
Folder	oozie		hue	hue	drwxr-xr-x	June 10, 2013 05:37 pm

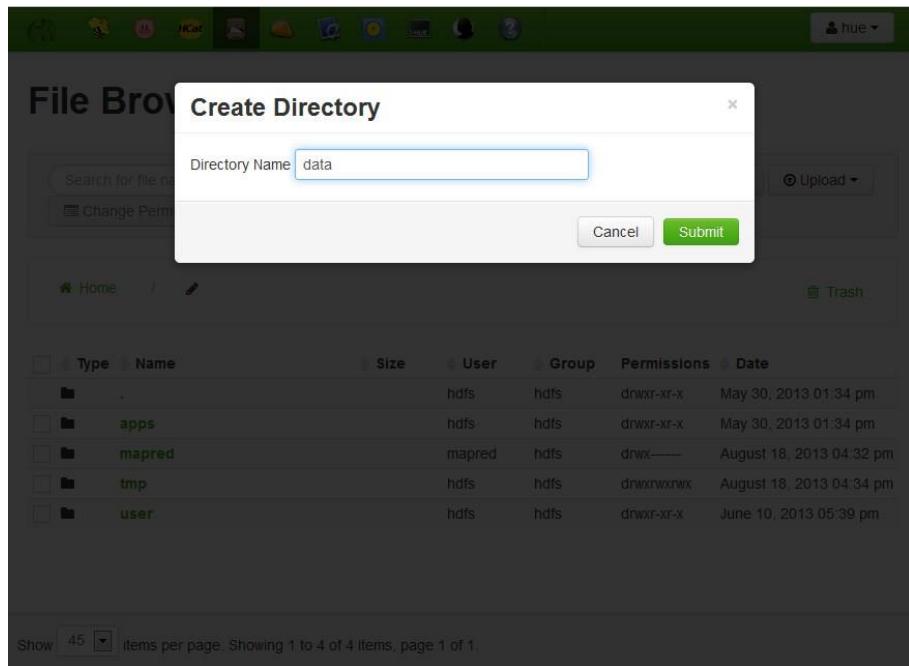
2. Click **New**, then select **Directory**.



The screenshot shows the Hue File Browser interface. At the top, there's a toolbar with various icons. Below it is a header bar with the title "File Browser". On the right side of the header is a user icon and a dropdown menu labeled "hue". Underneath the header is a toolbar with several buttons: "Search for file name", "A Rename", "Move", "Copy", "New", "Upload", "Change Permissions", "Download", and "Delete". The "New" button has a dropdown arrow pointing down. A red box highlights the "Directory" option in this dropdown menu. Below the toolbar is a breadcrumb navigation bar showing "Home /". To the right of the breadcrumb is a small trash bin icon. The main content area is a table listing files and directories. The columns are: Type, Name, Size, User, Group, Permissions, and Date. The table contains the following data:

Type	Name	Size	User	Group	Permissions	Date
Folder	.		hdfs	hdfs	drwxr-xr-x	May 30, 2013 01:34 pm
Folder	apps		hdfs	hdfs	drwxr-xr-x	May 30, 2013 01:34 pm
Folder	mapred		mapred	hdfs	drwx-----	August 18, 2013 04:32 pm
Folder	tmp		hdfs	hdfs	drwxrwxrwx	August 18, 2013 04:34 pm
Folder	user		hdfs	hdfs	drwxr-xr-x	June 10, 2013 05:39 pm

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



The screenshot shows the Hue File Browser interface with a "Create Directory" dialog box overlaid. The dialog box has a title "Create Directory" and a "Directory Name" input field containing "data". At the bottom of the dialog are two buttons: "Cancel" and "Submit", with "Submit" being green. In the background, the main file browser table is visible, showing the same directory structure as the previous screenshot. The table columns are: Type, Name, Size, User, Group, Permissions, and Date. The data in the table is identical to the one in the first screenshot.

4. The “data” folder appears in the list of files.

The screenshot shows the Hue File Browser interface. At the top, there's a toolbar with various icons and a user dropdown labeled "hue". Below the toolbar is a header bar with the title "File Browser". Underneath is a search bar and a set of buttons for "Rename", "Move", "Copy", "New", "Upload", "Change Permissions", "Download", and "Delete". The main area is a table listing files and directories:

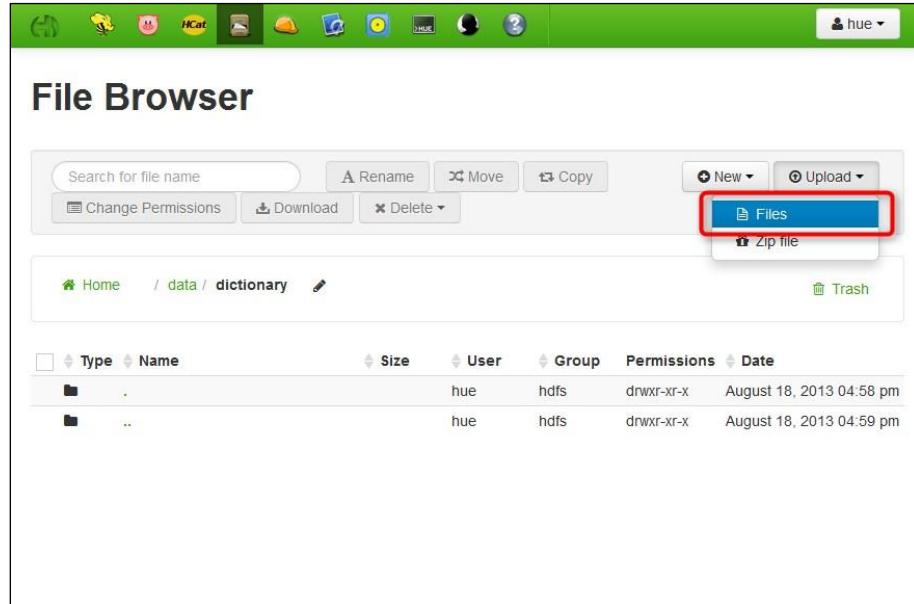
	Type	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	Folder	.		hdfs	hdfs	drwxr-xr-x	August 18, 2013 04:47 pm
<input type="checkbox"/>	Folder	apps		hdfs	hdfs	drwxr-xr-x	May 30, 2013 01:34 pm
<input type="checkbox"/>	Folder	data		hue	hdfs	drwxr-xr-x	August 18, 2013 04:47 pm
<input type="checkbox"/>	Folder	mapred		mapred	hdfs	drwx-----	August 18, 2013 04:32 pm
<input type="checkbox"/>	Folder	tmp		hdfs	hdfs	drwxrwxrwx	August 18, 2013 04:34 pm
<input type="checkbox"/>	Folder	user		hdfs	hdfs	drwxr-xr-x	June 10, 2013 05:39 pm

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.

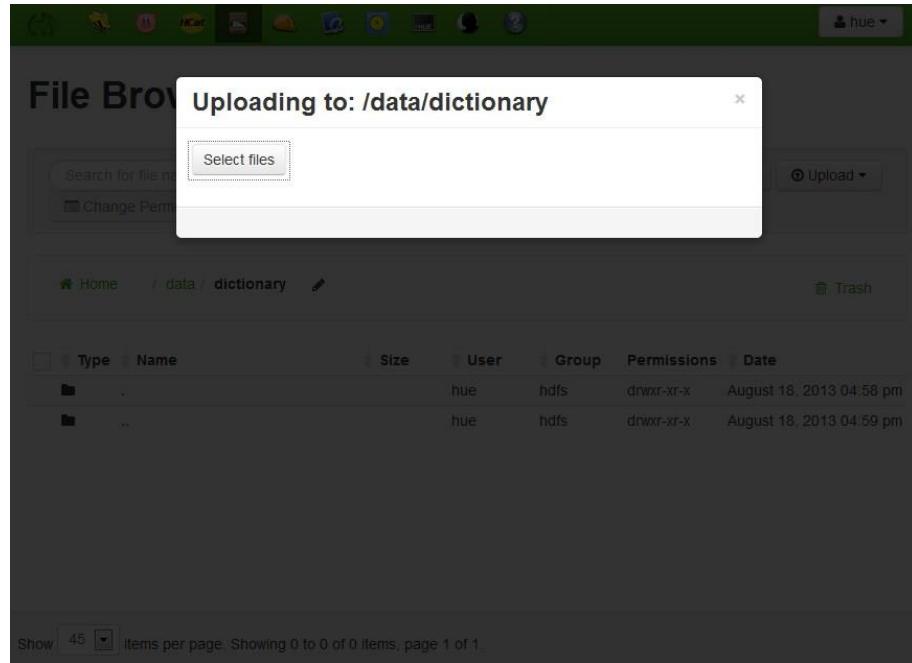
The screenshot shows the Hue File Browser interface, similar to the previous one, but now the "data" folder is selected. The path in the header is "Home / data". The table lists the contents of the "data" folder:

	Type	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	Folder	.		hue	hdfs	drwxr-xr-x	August 18, 2013 04:59 pm
<input type="checkbox"/>	Folder	..		hdfs	hdfs	drwxr-xr-x	August 18, 2013 04:47 pm
<input type="checkbox"/>	Folder	dictionary		hue	hdfs	drwxr-xr-x	August 18, 2013 04:58 pm
<input type="checkbox"/>	Folder	time_zone_map		hue	hdfs	drwxr-xr-x	August 18, 2013 04:59 pm
<input type="checkbox"/>	Folder	tweets		hue	hdfs	drwxr-xr-x	August 18, 2013 04:59 pm

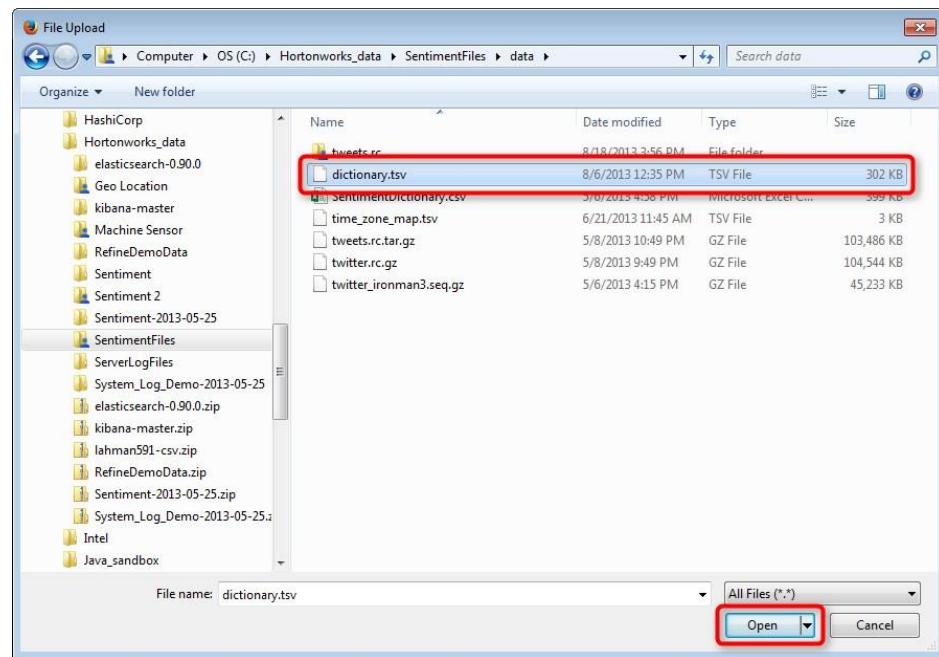
6. Click the “dictionary” folder. In the “dictionary” folder, click **Upload**, then select **Files**.



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.

A screenshot of the Hue File Browser interface. The title bar says 'File Browser'. The top navigation bar includes icons for Home, HDFS, HCat, and Hue. The main area shows a breadcrumb trail: Home / data / dictionary. There are buttons for Search, Rename, Move, Copy, New, Upload, Change Permissions, Download, and Delete. Below this is a table listing files:

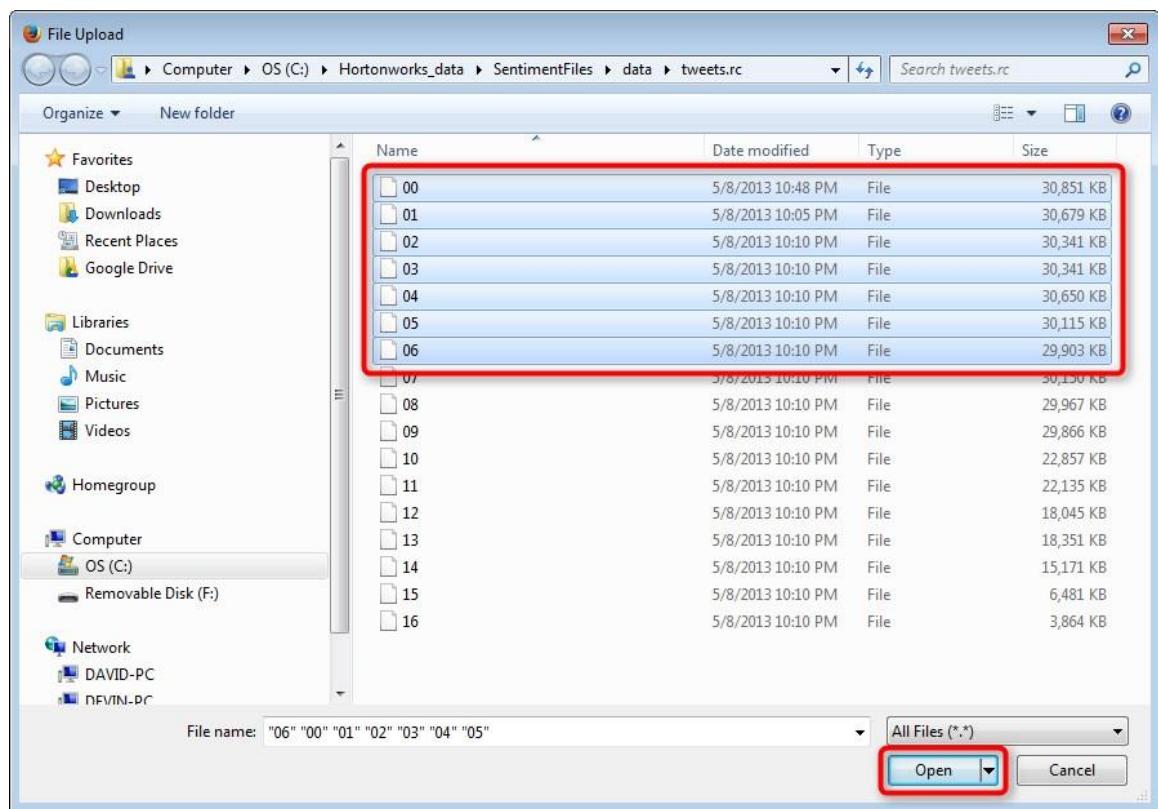
Type	Name	Size	User	Group	Permissions	Date
..	..		hue	hdfs	drwxr-xr-x	August 18, 2013 04:59 pm
File	dictionary.tsv	301.7 KB	hue	hdfs	-rwxr-xr-x	August 21, 2013 05:08 pm

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.

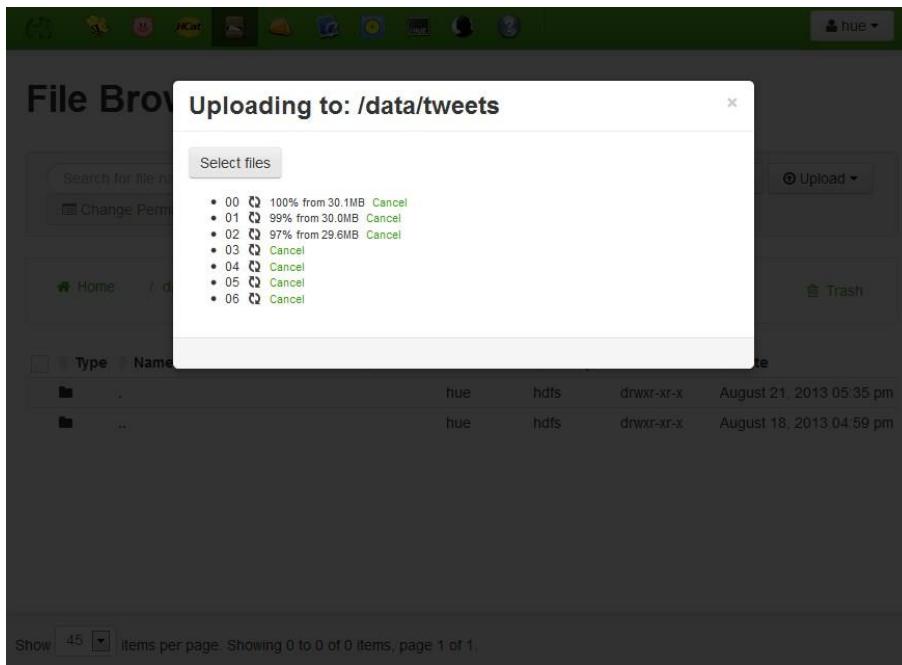
Type	Name	Size	User	Group	Permissions	Date
..	.		hue	hdfs	drwxr-xr-x	August 21, 2013 05:12 pm
..	..		hue	hdfs	drwxr-xr-x	August 18, 2013 04:59 pm
File	time_zone_map.tsv	3.0 KB	hue	hdfs	-rwxr-xr-x	August 21, 2013 05:12 pm

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

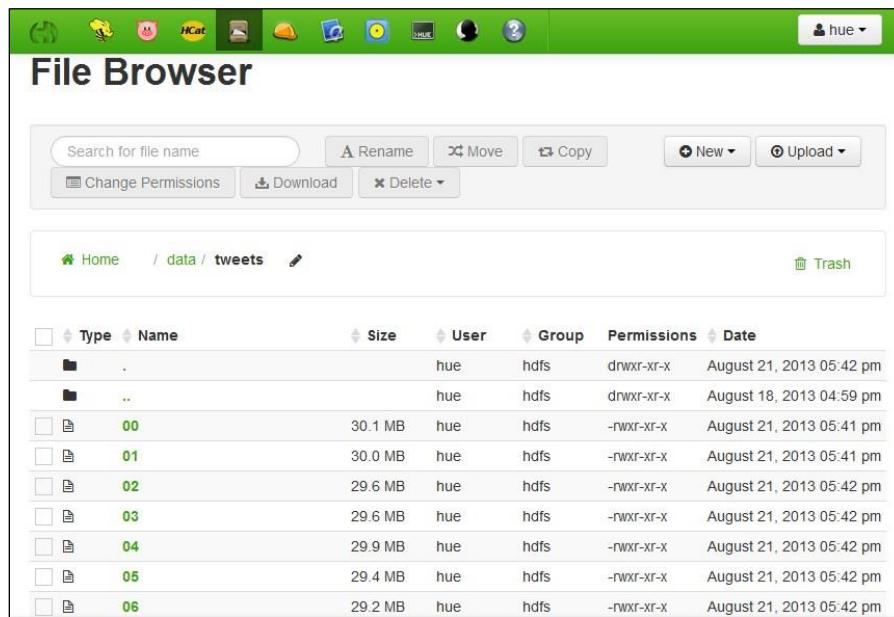
Use the File Upload dialog to browse to the SentimentFiles\data\tweets.rc folder. Select the “00” file, then Shift-click the “06” file to select files 00-06. Click **Open** to upload the files.



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:

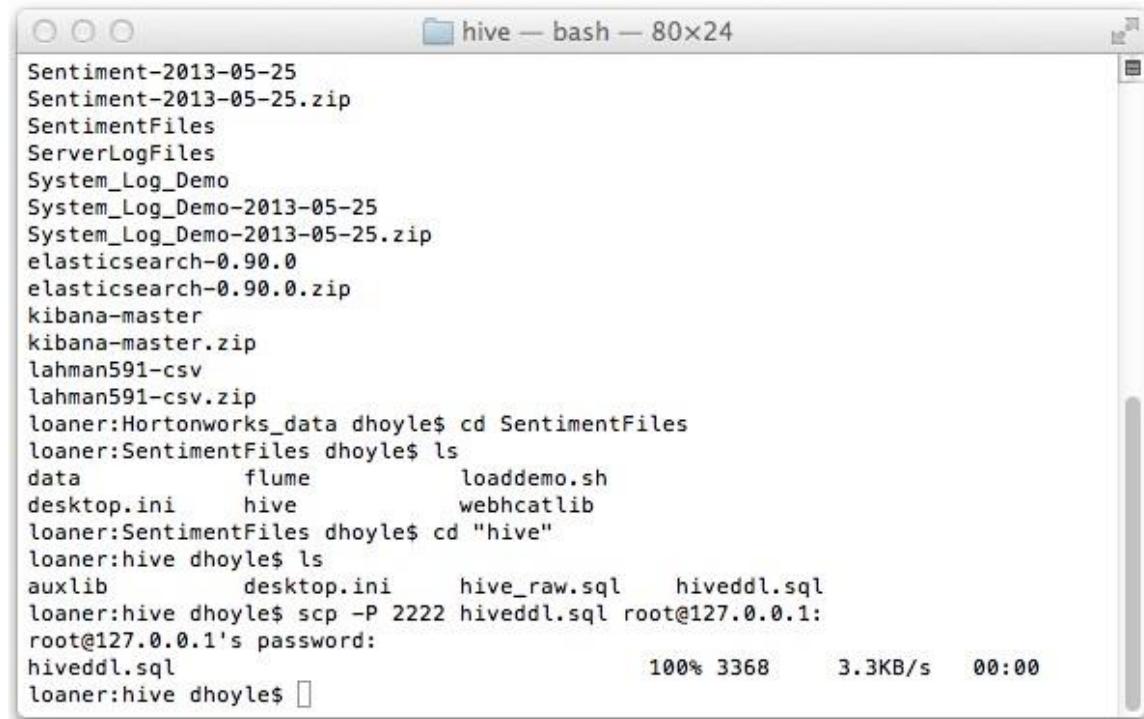
```
scp -P 2222 hiveddl.sql root@127.0.0.1:
```

#### 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
```



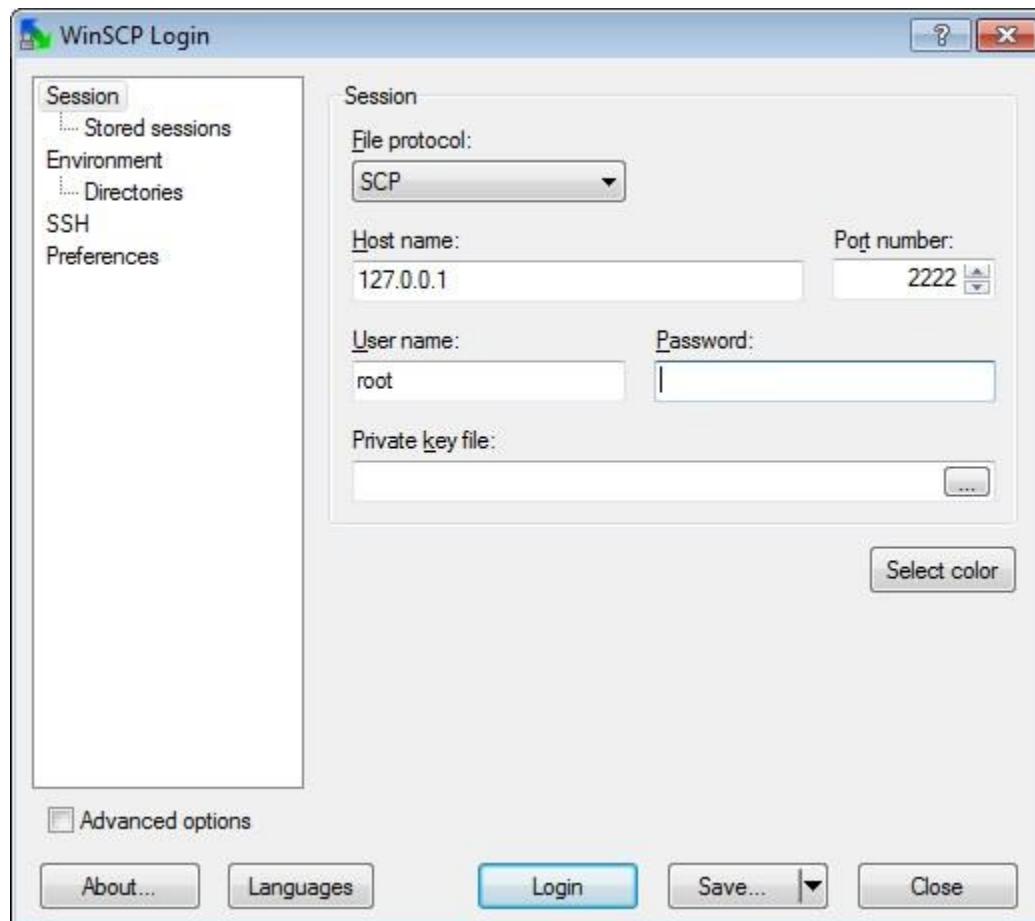
```
Sentiment-2013-05-25
Sentiment-2013-05-25.zip
SentimentFiles
ServerLogFiles
System_Log_Demo
System_Log_Demo-2013-05-25
System_Log_Demo-2013-05-25.zip
elasticsearch-0.90.0
elasticsearch-0.90.0.zip
kibana-master
kibana-master.zip
lahman591-csv
lahman591-csv.zip
loaner:Hortonworks_data dhoyles$ cd SentimentFiles
loaner:SentimentFiles dhoyles$ ls
data          flume          loaddemo.sh
desktop.ini   hive           webhcattlib
loaner:SentimentFiles dhoyles$ cd "hive"
loaner:hive dhoyles$ ls
auxlib        desktop.ini    hive_raw.sql    hiveddl.sql
loaner:hive dhoyles$ scp -P 2222 hiveddl.sql root@127.0.0.1:
root@127.0.0.1's password:
hiveddl.sql          100% 3368    3.3KB/s   00:00
loaner:hive dhoyles$
```

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

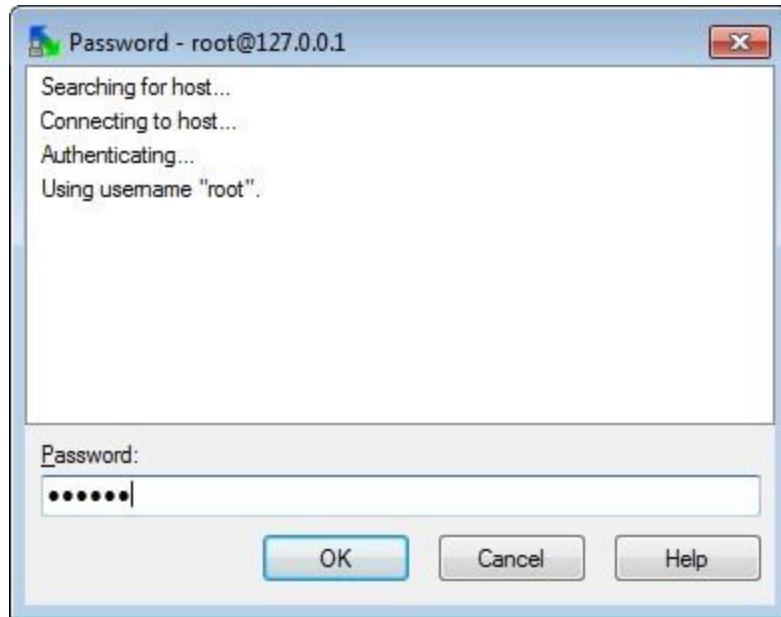
On Windows you must download and install the free [WinSCP](#) application.

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

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

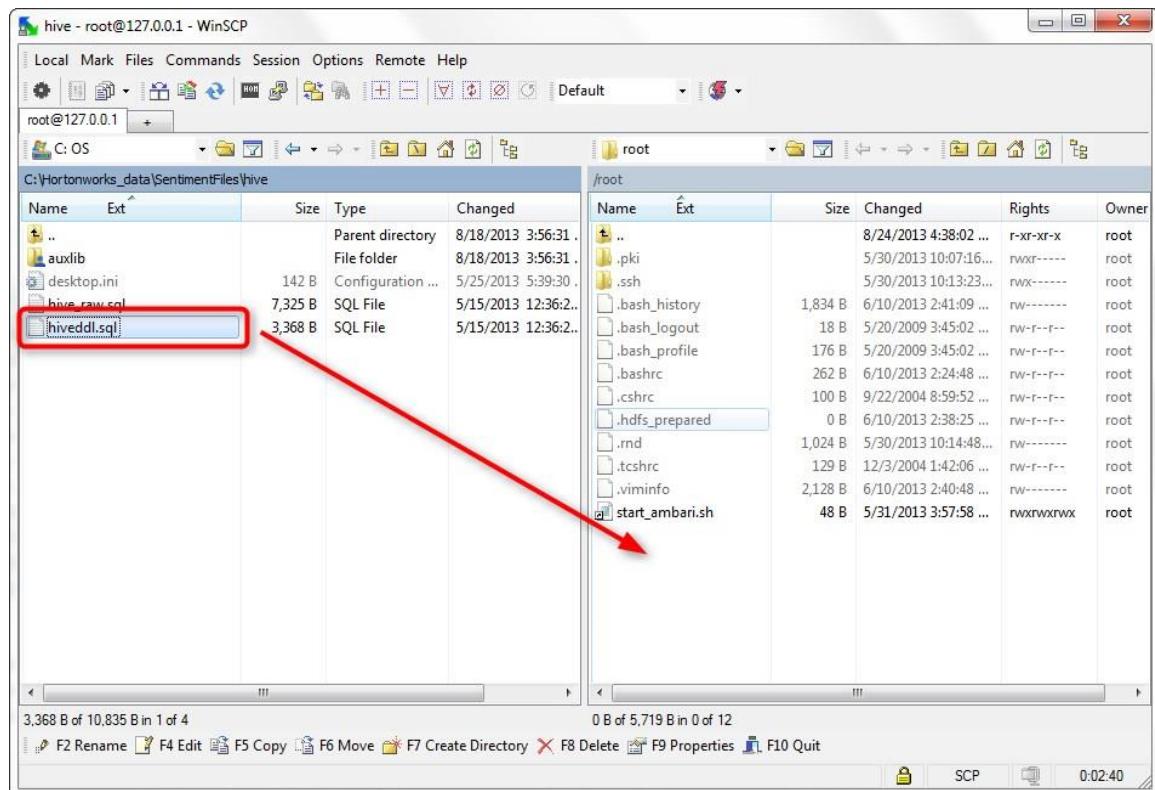


2. Type the Sandbox password ("hadoop") in the Password box, then click **OK**.

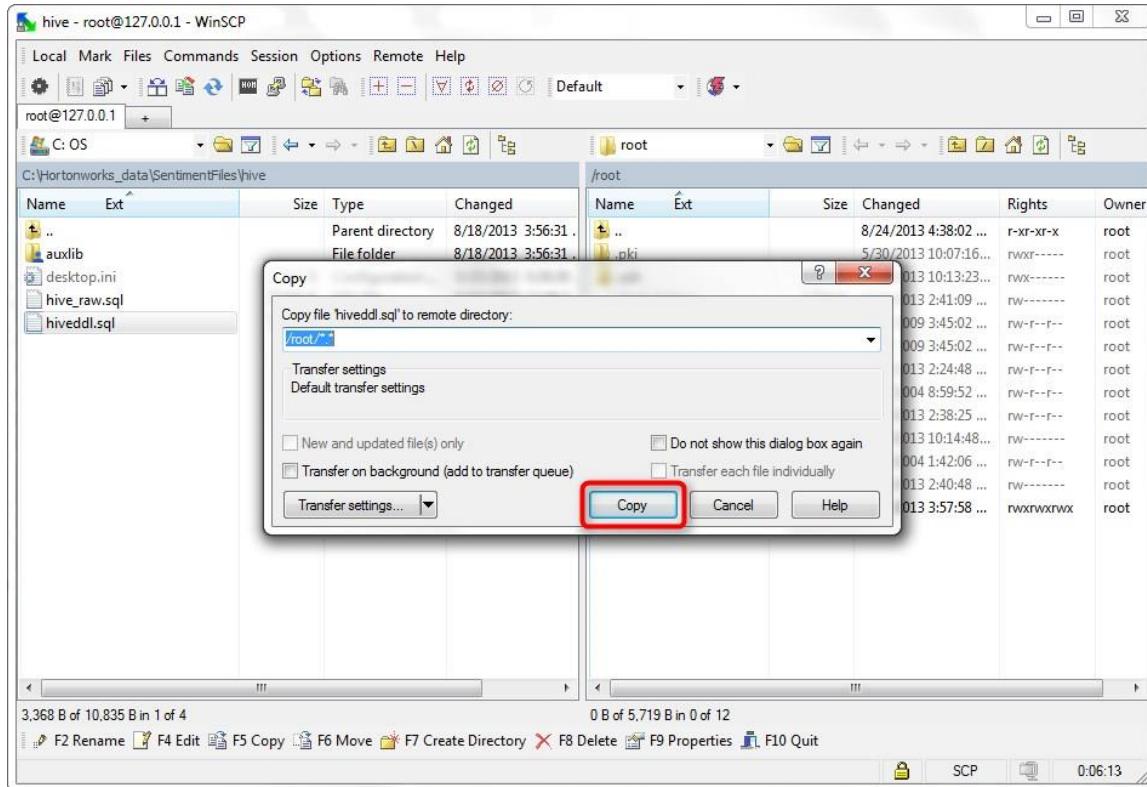


3. Use the WinSCP file browser to navigate to the SentimentFiles\hive folder in the left-hand pane, and to the Sandbox /root folder in the right-hand pane.

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

- 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 ~]#:

- 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.

```
Hortonworks Sandbox 1.3 [Running] - Oracle VM VirtualBox
Machine View Devices Help
2013-08-24 07:38:46,497 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 53.24
sec
2013-08-24 07:38:47,510 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 53.24
sec
2013-08-24 07:38:48,571 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 53.24
sec
MapReduce Total cumulative CPU time: 53 seconds 240 msec
Ended Job = job_201308240437_0004
Moving data to: hdfs://sandbox:8020/apps/hive/warehouse/tweetsbi
Table default.tweetsbi stats: [num_partitions: 0, num_files: 1, num_rows: 0, total_size: 82771672, raw_data_size: 0]
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 63.37 sec   HDFS Read: 65733972 HDFS Write: 1838
7420 SUCCESS
Job 1: Map: 1   Reduce: 1   Cumulative CPU: 29.1 sec   HDFS Read: 18387871 HDFS W
rite: 23462113 SUCCESS
Job 2: Map: 1   Cumulative CPU: 40.04 sec   HDFS Read: 129178946 HDFS Write: 782
02584 SUCCESS
Job 3: Map: 2   Reduce: 1   Cumulative CPU: 53.24 sec   HDFS Read: 101668345 HDF
Write: 82771672 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 5 seconds 750 msec
OK
Time taken: 481.457 seconds
FAILED: SemanticException [Error 10001]: Line 15:7 Table not found 'tweets'
[root@sandbox ~]# _
```

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**.

The screenshot shows the HCatalog: Table List page in the Hue interface. On the left, there's a sidebar with actions like 'Create a new table from a file' and 'Create a new table manually'. The main area lists tables with checkboxes and 'Browse Data' buttons. The 'tweets\_raw' table is checked and highlighted with a red box.

Table Name	Browse Data
dictionary	<button>Browse Data</button>
I1	<button>Browse Data</button>
I2	<button>Browse Data</button>
I3	<button>Browse Data</button>
sample_07	<button>Browse Data</button>
sample_08	<button>Browse Data</button>
time_zone_map	<button>Browse Data</button>
tweets_clean	<button>Browse Data</button>
<input checked="" type="checkbox"/> tweets_raw	<button>Browse Data</button>
tweets_sentiment	<button>Browse Data</button>

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.

The screenshot shows the Query Results page for the 'tweets\_raw' table. It displays a list of tweets with columns 'id', 'created\_at', and 'source'. A tooltip provides information on how to use the search feature.

id	created_at	source
0	Thu May 02 21:45:02 +0000 2013	<a href="http://www.echofon.com"
1	Thu May 02 21:45:02 +0000 2013	<a href="http://twitterfeed.com"
2	Thu May 02 21:45:02 +0000 2013	<a href="http://dlvr.it" rel="nofollow"
3	Thu May 02 21:45:02 +0000 2013	<a href="http://twitterfeed.com"
4	Thu May 02 21:45:03 +0000 2013	<a href="http://twitterfeed.com"
5	Thu May 02 21:45:03 +0000 2013	<a href="http://twitter.com/down"
6	Thu May 02 21:45:03 +0000 2013	<a href="http://twitter.com/down"
7	Thu May 02 21:45:03 +0000 2013	<a href="http://twitter.com/down"
8	Thu May 02 21:45:04 +0000 2013	<a href="http://twitterfeed.com"
9	Thu May 02 21:45:04 +0000 2013	<a href="http://twitter.com/down"

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.

The screenshot shows the Hue interface with the following details:

- Top Bar:** Includes icons for HDFS, HIVE, HCAT, and other tools, along with a user dropdown labeled "hue".
- Left Sidebar:** Labeled "Tables" and shows a "DOWNLOADS" section with options to "Download as CSV", "Download as XLS", and "Save".
- Central Area:** Title "Query Results: tweetsbi". Below it, a "Did you know?" tip: "If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that match the string." A "Results" tab is selected, showing a table with columns "country" and "sentiment".
- Data Table:**

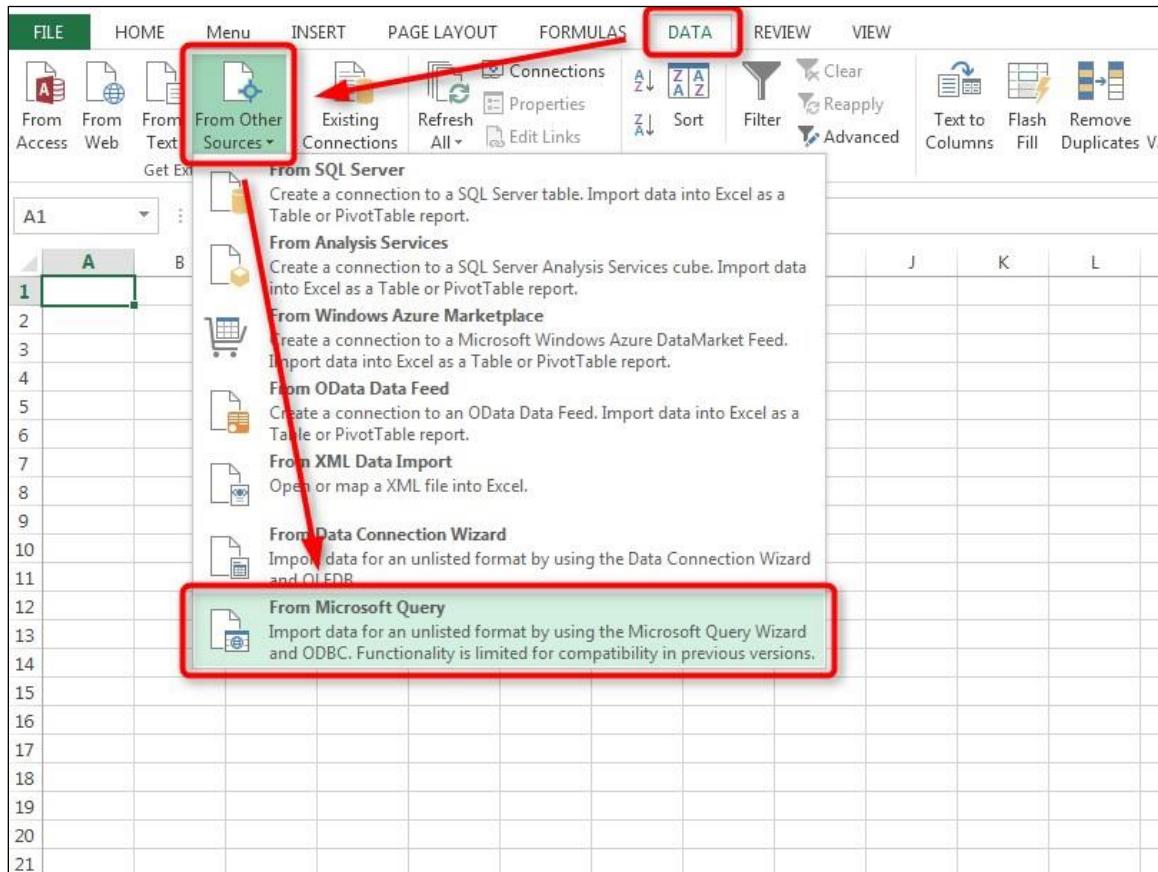
	country	sentiment
t the... http://t.co/9CwF31E8Ez	UNITED STATES	1
	CHILE	2
t the... http://t.co/sAX3MsJiE	MOROCCO	1
jHNJMkjN19	UNITED STATES	2
	UNITED STATES	1
t the... http://t.co/Ss6SyvX50m	NETHERLANDS	1
nillonario ni superhéroe.		0
- Bottom Right:** "Next Page →" button.

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

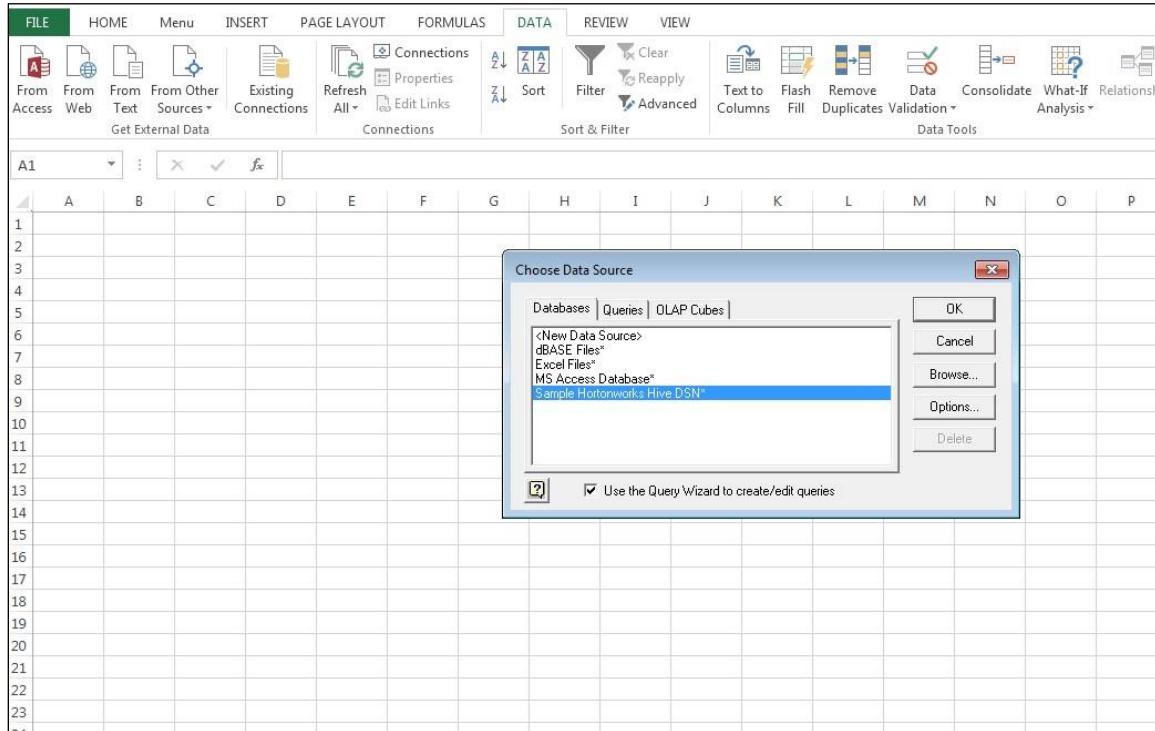
In this section, we use Excel Professional Plus 2013 to access the refined sentiment data.

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

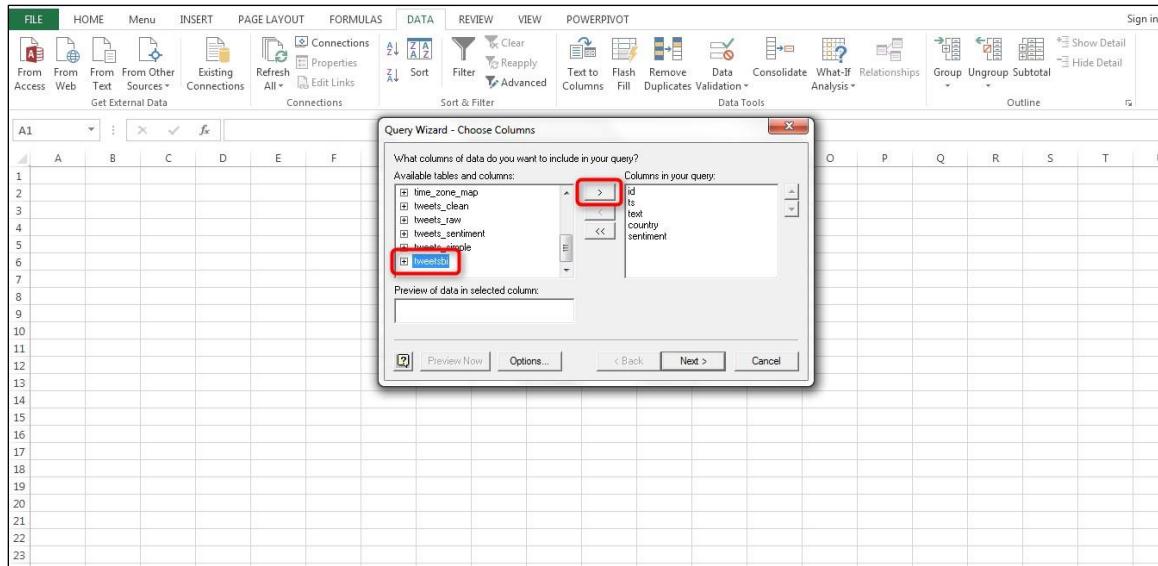


2. On the Choose Data Source pop-up, select the Hortonworks ODBC data source you installed previously, then click **OK**.

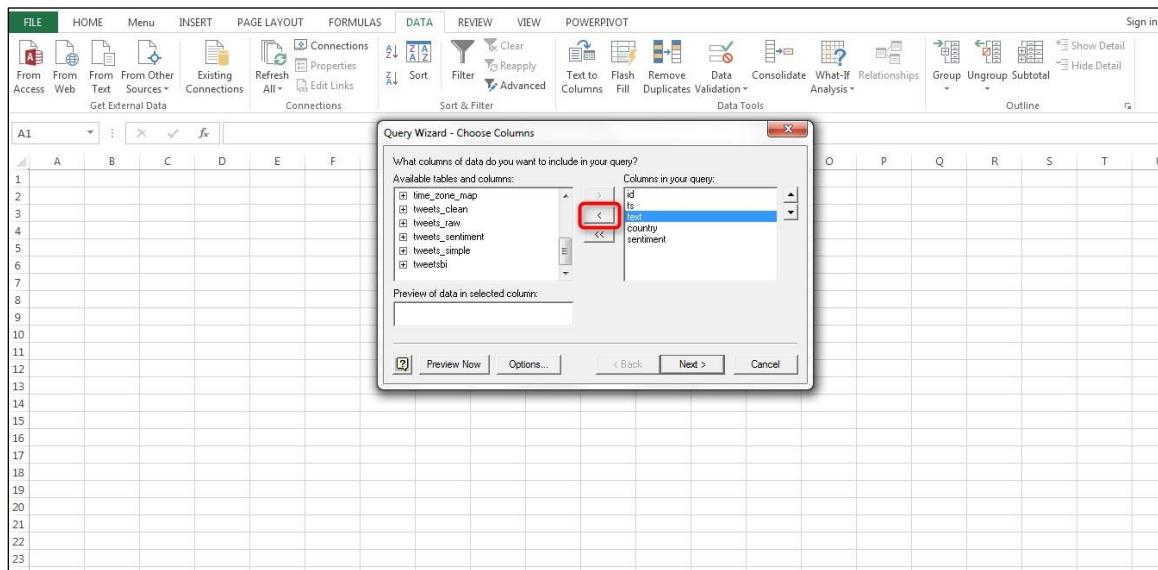
The Hortonworks ODBC driver enables you to access Hortonworks data with Excel and other Business Intelligence (BI) applications that support ODBC.



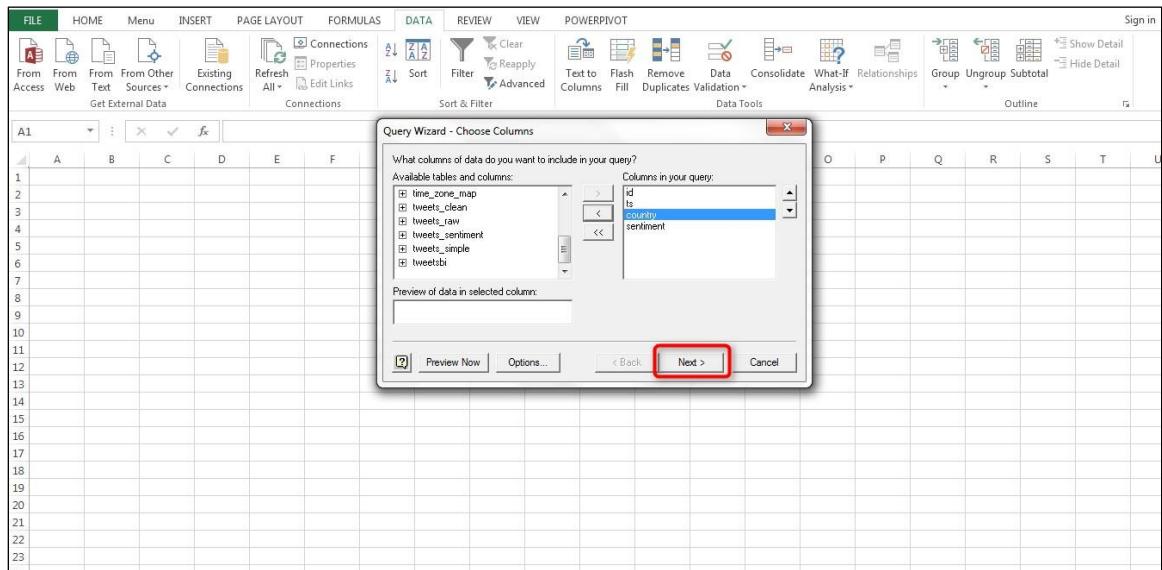
3. After the connection to the Sandbox is established, the Query Wizard appears. Select the “tweetsbi” table in the Available tables and columns box, then click the right arrow button to add the entire “tweetsbi” table to the query. Click **Next** to continue.



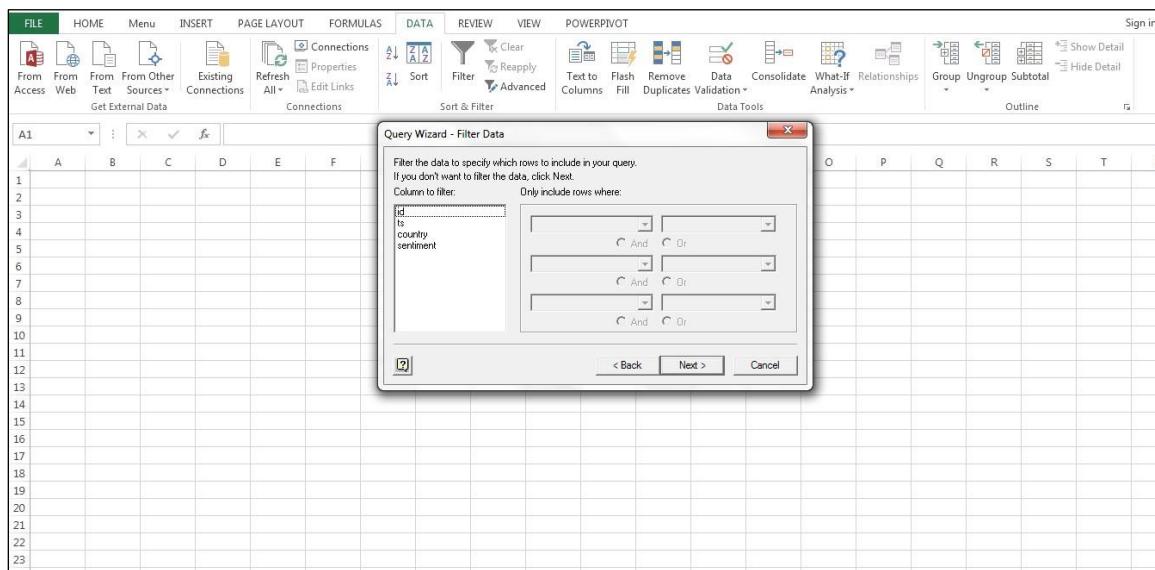
4. Select the “text” column in the “Columns in your query” box, then click the left arrow button to remove the text column.



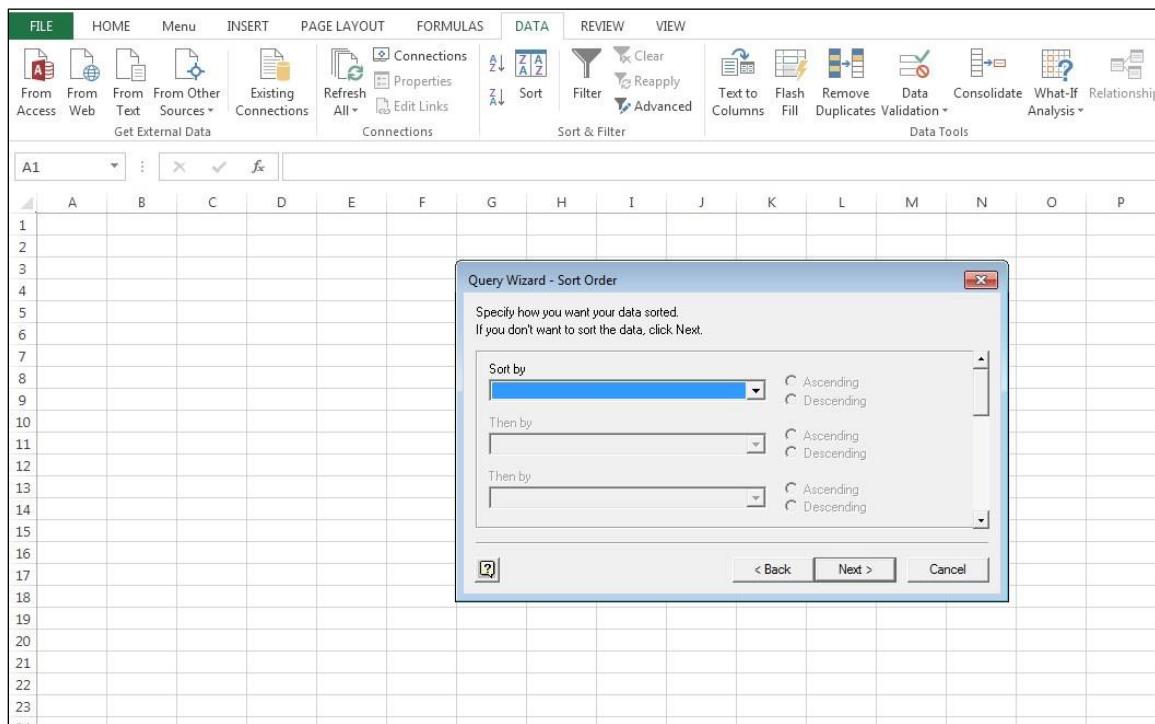
5. After the “text” column has been removed, click **Next** to continue.



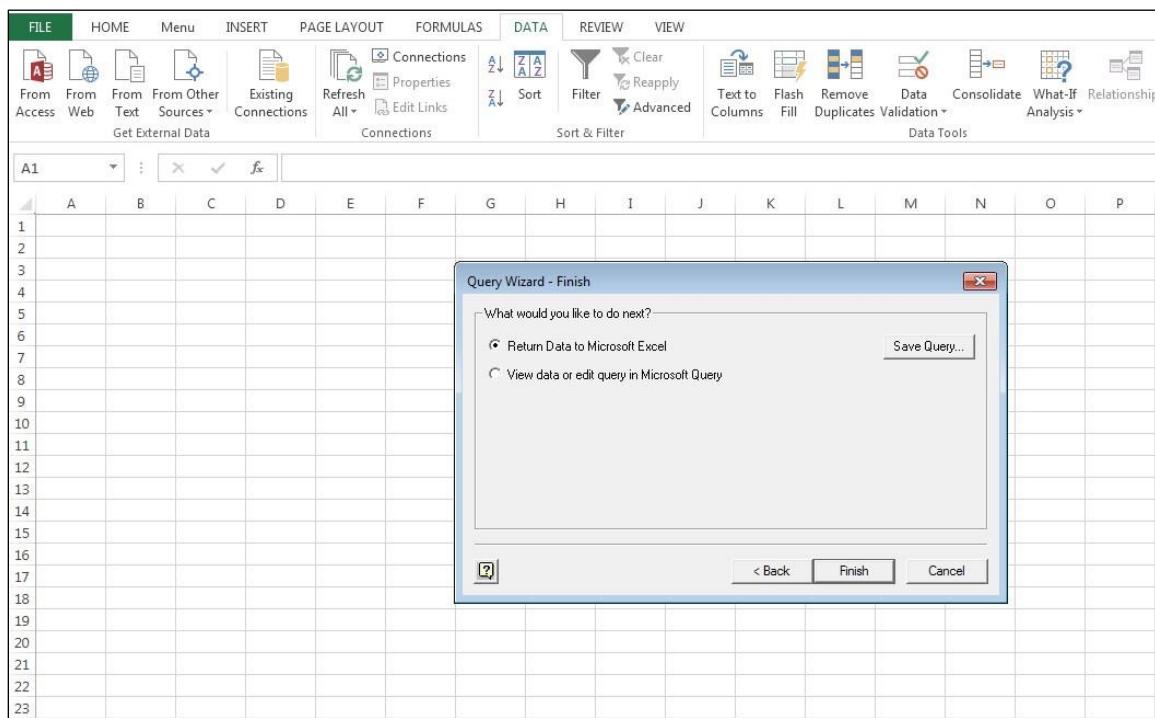
6. On the Filter Data screen, click **Next** to continue without filtering the data.



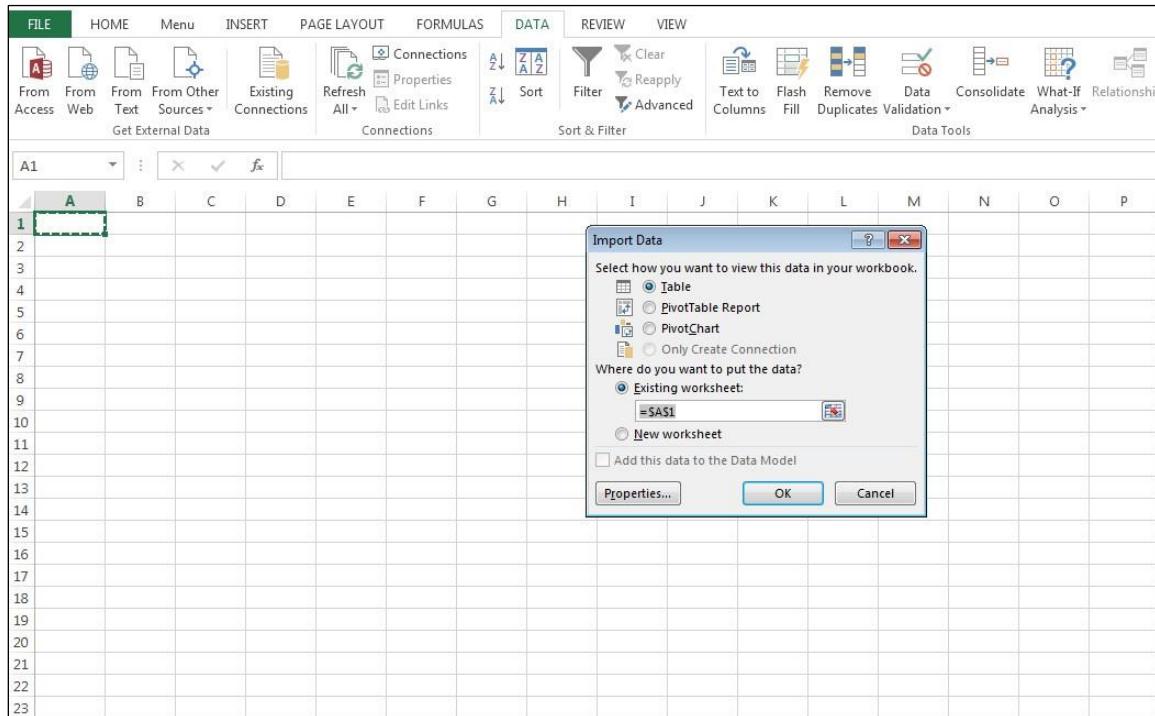
7. On the Sort Order screen, click **Next** to continue without setting a sort order.



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.

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>	<b>H</b>	<b>I</b>	<b>J</b>	<b>K</b>	<b>L</b>	<b>M</b>	<b>N</b>	<b>O</b>	<b>P</b>	<b>Q</b>
1	<b>id</b>	<b>ts</b>	<b>country</b>	<b>sentiment</b>													
2	3.30044E+17	5/2/2013 19:39		1													
3	3.30044E+17	5/2/2013 19:39	UNITED STATES	2													
4	3.30044E+17	5/2/2013 19:39	CHILE	1													
5	3.30044E+17	5/2/2013 19:39		1													
6	3.30044E+17	5/2/2013 19:39	MOROCCO	2													
7	3.30044E+17	5/2/2013 19:39	UNITED STATES	1													
8	3.30044E+17	5/2/2013 19:39		1													
9	3.30044E+17	5/2/2013 19:39	UNITED STATES	1													
10	3.30044E+17	5/2/2013 19:39		1													
11	3.30044E+17	5/2/2013 19:39	NETHERLANDS	0													
12	3.30044E+17	5/2/2013 19:39	THAILAND	1													
13	3.30044E+17	5/2/2013 19:39	UNITED STATES	2													
14	3.30044E+17	5/2/2013 19:39		1													
15	3.30044E+17	5/2/2013 19:39	UNITED STATES	2													
16	3.30044E+17	5/2/2013 19:39	INDONESIA	1													
17	3.30044E+17	5/2/2013 19:39	ARGENTINA	1													
18	3.30044E+17	5/2/2013 19:39	UNITED STATES	2													
19	3.30044E+17	5/2/2013 19:39	THAILAND	1													
20	3.30044E+17	5/2/2013 19:39		1													
21	3.30044E+17	5/2/2013 19:39	PAKISTAN	2													
22	3.30044E+17	5/2/2013 19:39	UNITED STATES	1													
23	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.

The screenshot shows the Microsoft Excel ribbon with the 'INSERT' tab selected. A red box highlights the 'Power View' icon in the 'REPORTS' section. Below the ribbon, a table titled 'tweetsbi' is displayed with columns: id, ts, country, and sentiment. The data in the table includes rows for various tweets from different countries, with sentiment values ranging from 0 to 2.

	A	B	C	D	E	F	G	H		
1	id	ts	country	sentiment					L	M
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						

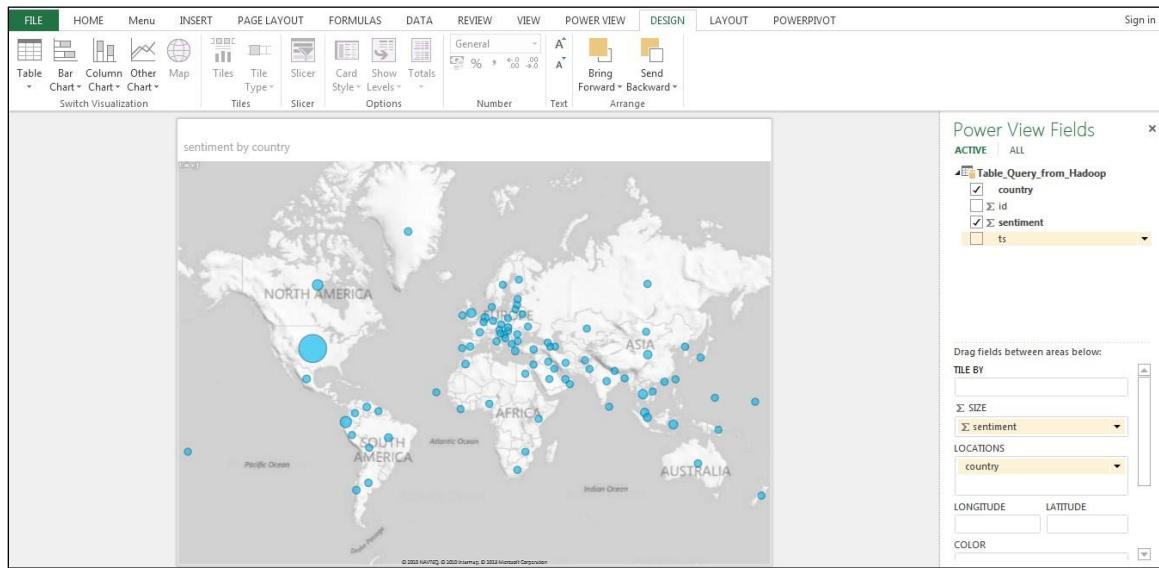
2. The Power View Fields area appears on the right side of the window, with the data table displayed on the left. Drag the handles or click the Pop Out icon to maximize the size of the data table.

This screenshot shows the Microsoft Power BI interface. On the left, a data table is displayed with columns: id, ts, country, and sentiment. A red circle highlights the 'Pop Out' icon in the top right corner of the table area. On the right, the 'Power View Fields' pane is open, showing the 'ACTIVE' section with a table named 'Table\_Query\_from\_Sample\_Horton'. The fields selected are country, id, sentiment, and ts. Below this, there are sections for 'Filters', 'Drag fields between areas below: TILE BY', and 'FIELDS' where id, ts, country, and sentiment are listed.

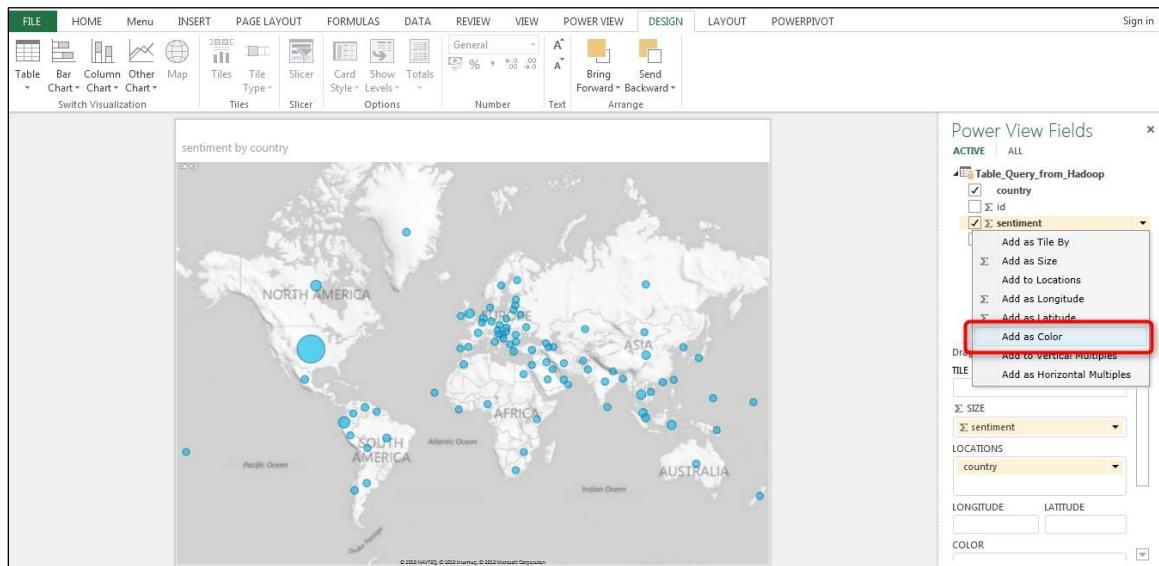
3. In the Power View Fields area, clear the checkboxes next to the **id** and **ts** fields, then click **Map** on the Design tab in the top menu.

This screenshot shows the Microsoft Power BI interface after step 3. The 'Power View Fields' pane now shows the table 'Table\_Query\_from\_Hadoop' with the checkboxes for id and ts cleared. The FIELDS section still lists id, ts, country, and sentiment. The data table on the left shows a list of countries with their corresponding sentiment values. A red box highlights the 'Map' icon in the top menu bar.

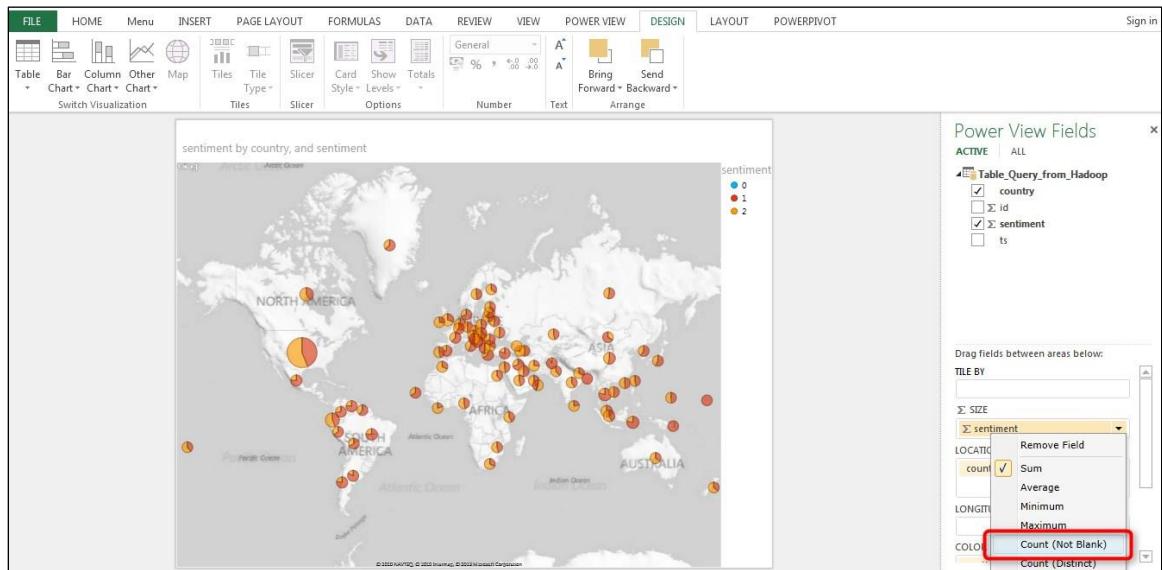
4. The map view displays a global view of the data.



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

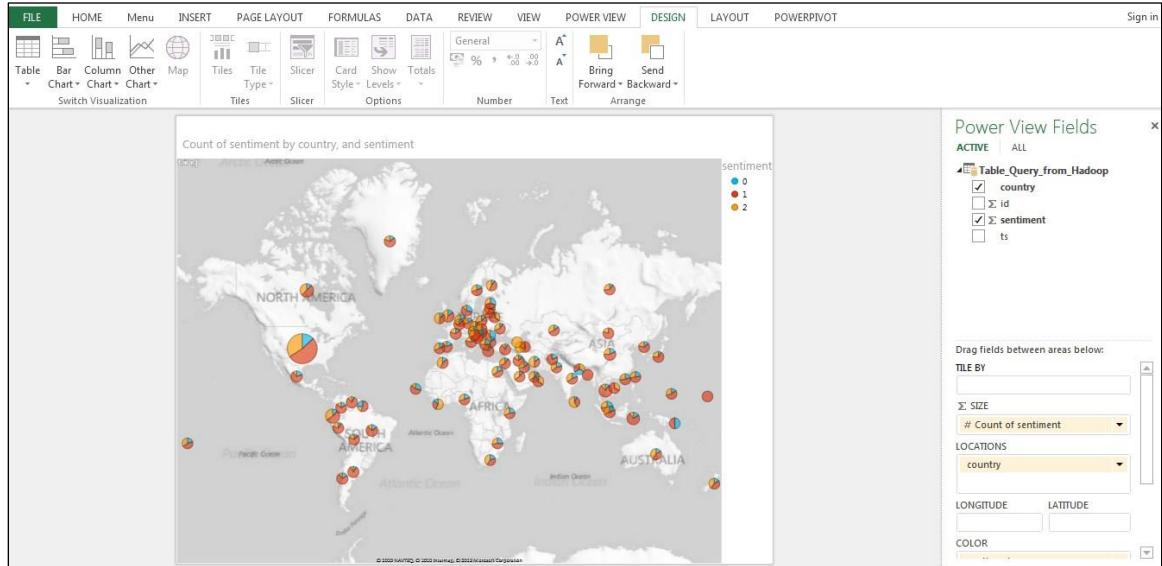


6. Under SIZE, click **sentiment**, then select **Count (Not Blank)**.

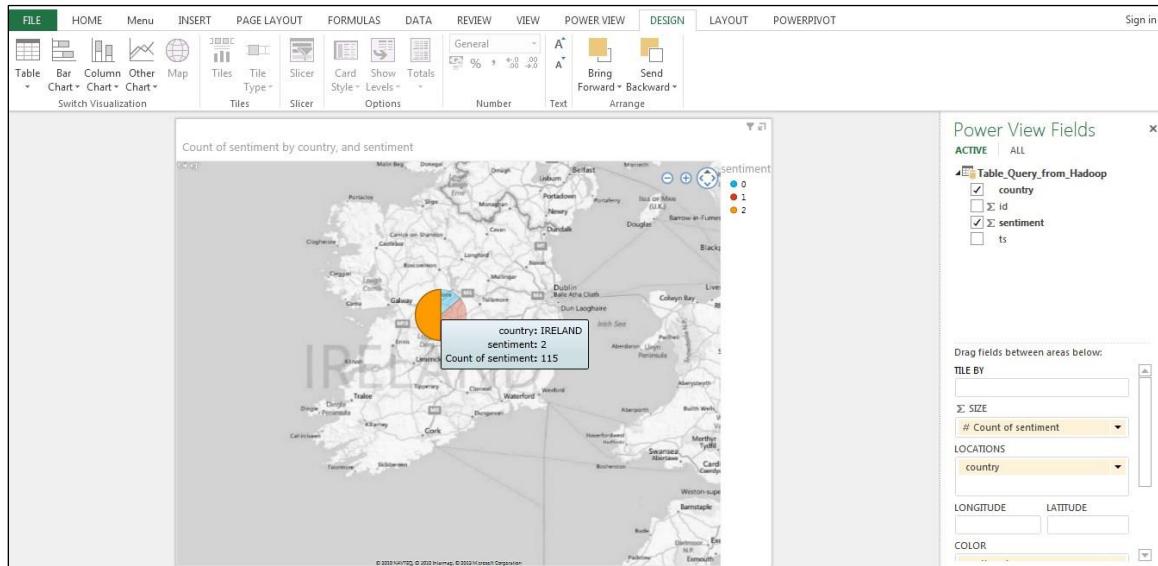


7. Now the map displays the sentiment data by color:

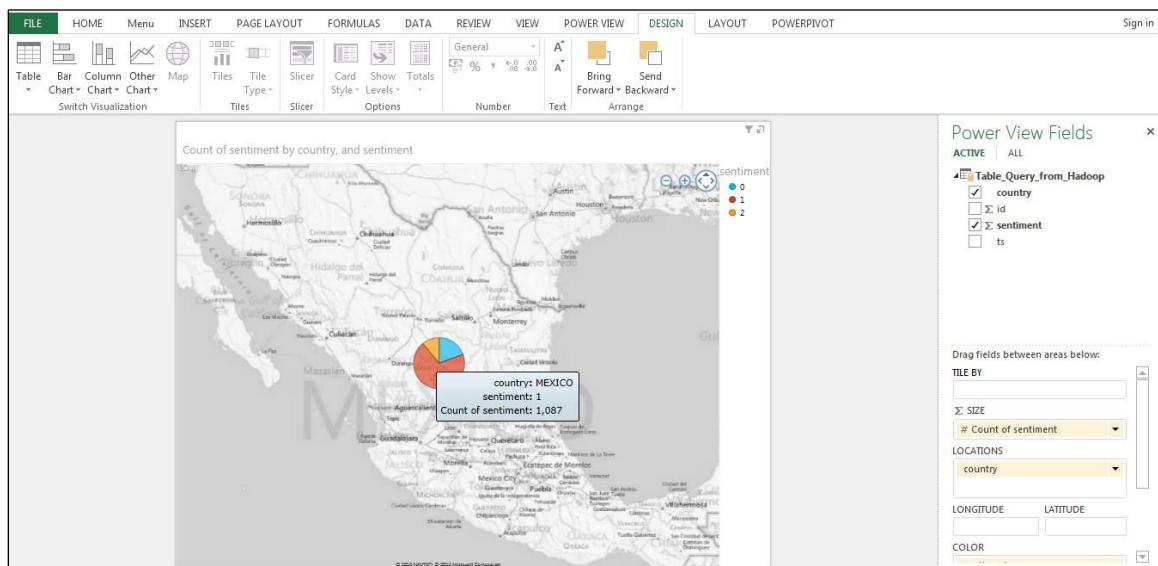
- Orange: positive
- Blue: negative
- Red: neutral



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.

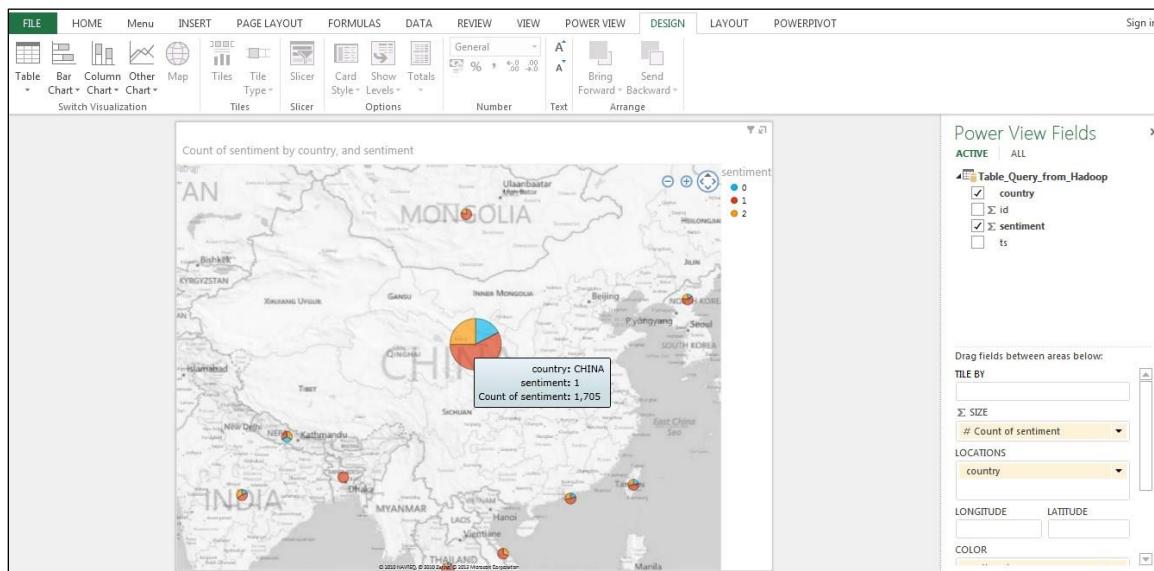


9. Now use the map controls to zoom in on Mexico. In Mexico, about one-fifth of the tweets expressed negative sentiment (shown in blue), and only a small portion of the tweets were positive. Most tweets from Mexico were neutral, as shown in red.



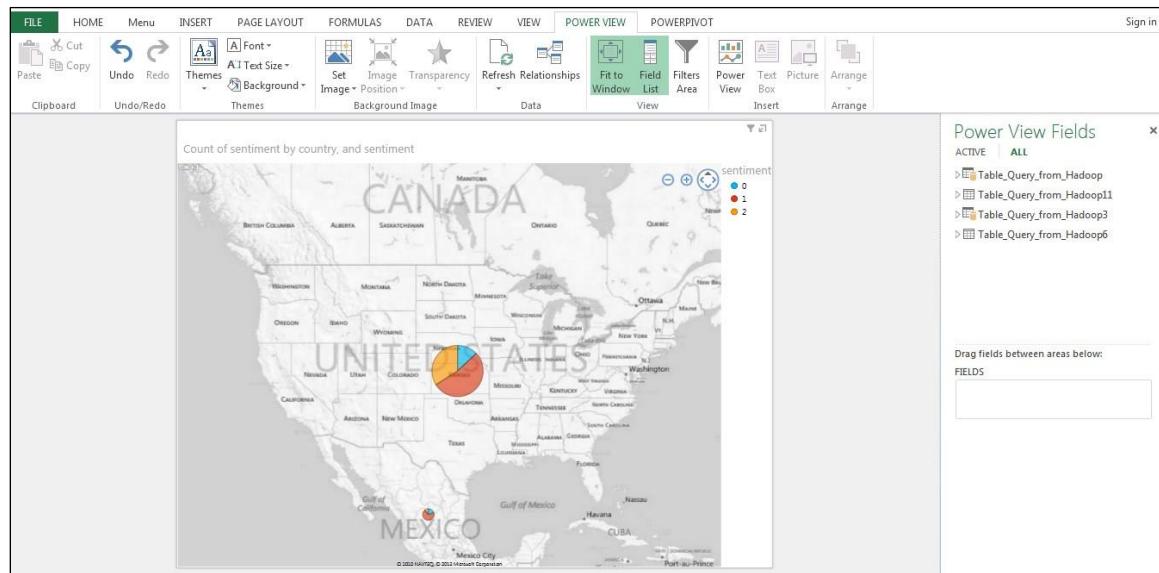
10. Next, use the map controls to zoom in on the sentiment data in China. Marvel studios and the Chinese studio DMG co-financed *Iron Man 3*, and the cast included a famous Chinese actress.

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.