

Sentiment Analysis using Hadoop

Sponsored By Atlink Communications Inc

Instructor : Dr.Sadegh Davari

Mentors : Dilhar De Silva , Rishita Khalathkar

Team Members : Ankur Uprit Pinaki Ranjan Ghosh

Kiranmayi Ganti Srijha Reddy Gangidi

Capstone Project Group 1

What is Sentiment Analysis ?
Sentiment Analysis with Twitter
Classification of Data
Types of Sentiment Analysis
Introduction to the Project
What is Hadoop and HDFS ?
Structured and Unstructured Data



Ankur Uprit

Team Leader/ Application Developer

Capstone Project Group 1

Sentiment Analysis

➤ Sentiment analysis is the detection of **attitudes**

- Enduring, affectively colored beliefs, dispositions towards objects or persons

1. **Holder (source)** of attitude

2. **Target (aspect)** of attitude

3. **Type** of attitude

- From a set of types

- *Like, love, hate, value, desire, etc.*

- Or (more commonly) simple weighted **polarity**:

- *positive, negative, neutral, together with strength*

4. **Text** containing the attitude

- Sentence or entire document



Sentiment Analysis

(Cont...)

- 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
- The attitude may be his or her
 1. **Judgment**
 2. **Affective state** (that is to say, the emotional state of the author when writing)
 3. **Intended emotional communication** (that is to say, the emotional effect the author wishes to have on the reader)

Sentiment Analysis With Twitter

- twitter.com is a popular microblogging website
- Each tweet is 140 characters in length
- Tweets are frequently used to express a tweeter's emotion on a particular subject
- There are firms which poll twitter for analyzing sentiment on a particular topic
- The challenge is to gather all such relevant data, detect and summarize the overall sentiment on a topic

Classification Of Data

- Polarity classification – Positive

Negative Sentiment

- 3-way classification – Positive

Negative

Neutral



Types of sentiment analysis

- **Movie:** Is this review positive or negative?
- **Products:** What do people think about the new iPhone?
- **Public Sentiment:** How is consumer confidence? Is despair Increasing?
- **Politics:** What do people think about this candidate or issue?
- **Prediction:** Predict election outcomes or market trends from sentiment

Introduction to the project

Sentiment Analysis Using Hadoop & Hive

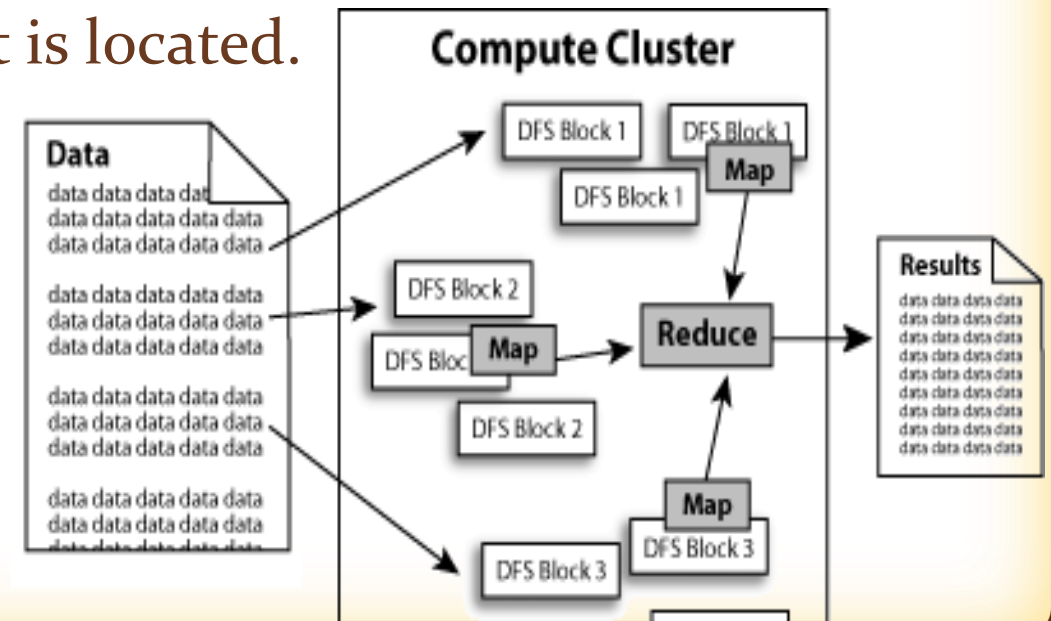


What is Hadoop and HDFS ?

- **Hadoop** : A Software Framework for Data Intensive Computing Applications
- Software platform that lets one easily write and run applications that process vast amounts of data. It includes:
 - MapReduce – offline computing engine
 - HDFS – Hadoop distributed file system
 - HBase (pre-alpha) – online data access
- Yahoo! is the biggest contributor

What does Hadoop do ?

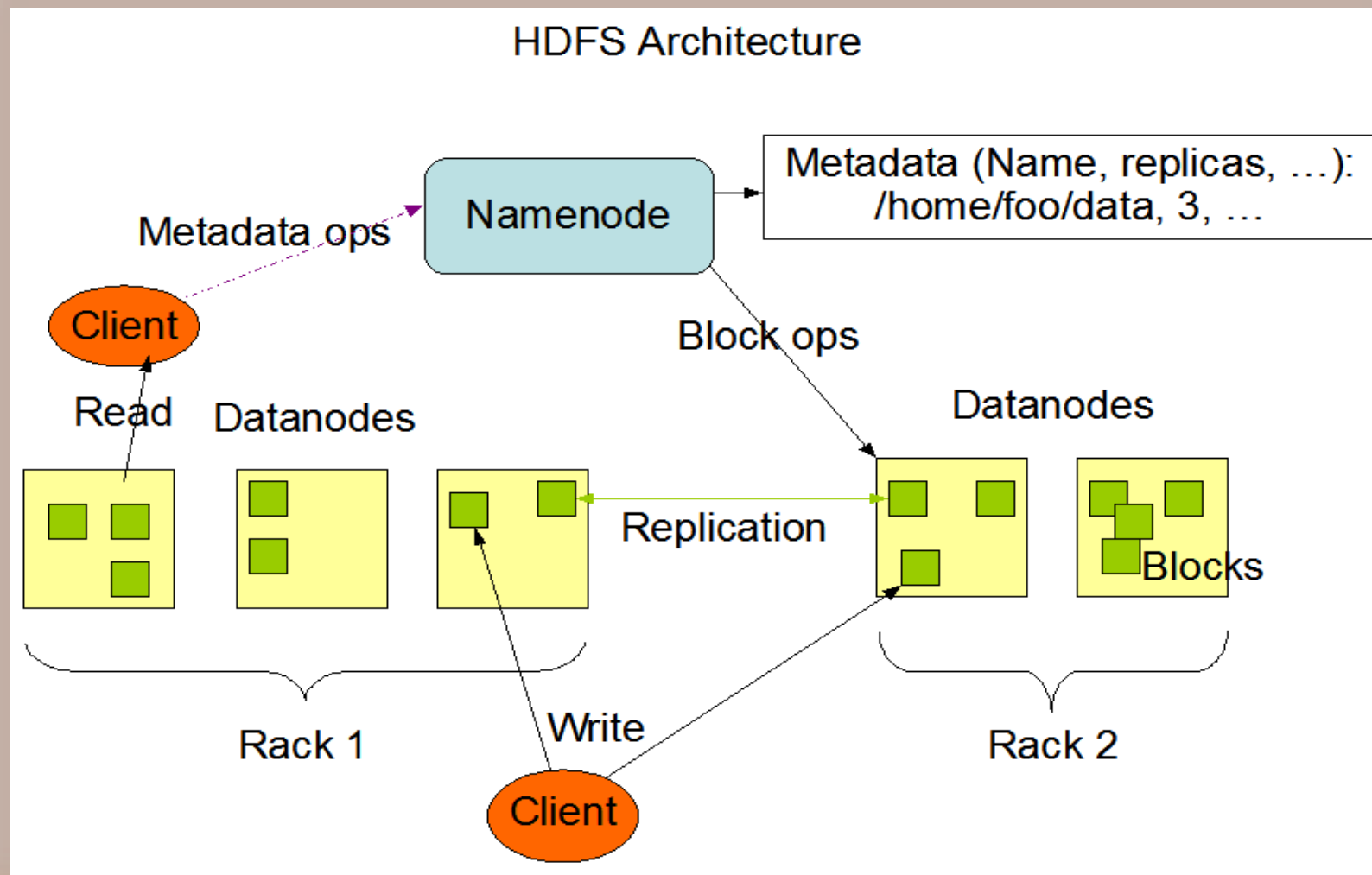
- Hadoop implements Google's MapReduce, using HDFS
 - MapReduce divides applications into many small blocks of work.
 - HDFS creates multiple replicas of data blocks for reliability, placing them on compute nodes around the cluster.
 - MapReduce can then process the data where it is located.
 - Hadoop's target is to run on clusters of the order of 10,000-nodes.
-
- The diagram shows a 'Data' document icon on the left containing the text 'data data data data data data data data data data data data data data data'. An arrow points from this icon to a box labeled 'Compute Cluster'. Inside the 'Compute Cluster' box, there are three boxes labeled 'DFS Block 1', one 'DFS Block 2' at the bottom, and another 'DFS Block 1' to its right. A 'Map' task box is positioned between the two 'DFS Block 1' boxes, with arrows pointing to each of them. Below the 'DFS Block 2' box, there is a partially visible box labeled 'DFS Block 3'.



HDFS - Hadoop Distributed File System

- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.
- It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant.
 - **Highly fault-tolerant** and is designed to be deployed on low-cost hardware.
 - **Provides high throughput access** to application data and is suitable for applications that have large data sets.
 - Relaxes a few POSIX requirements to enable streaming access to file system data.
 - Part of the Apache Hadoop Core project.
The project URL is <http://hadoop.apache.org/core/>.

HDFS Architecture



Sentiment Analysis Using Hadoop & Hive

- The twitter data is mostly unstructured
- Hadoop is the technology that is capable of dealing with such large unstructured data
- In this project, Hadoop Hive on Windows will be used to analyze data.
- This analysis will be shown with interactive visualizations using some powerful BI tools for Excel like Power View
- Finally, a real time case study will be used to create a report on how Sentiment Analysis can be implemented for a product
- What infrastructure, skills, technology would be most ideal and how it would help in improving the brand image/ quality of the product

Technologies Used

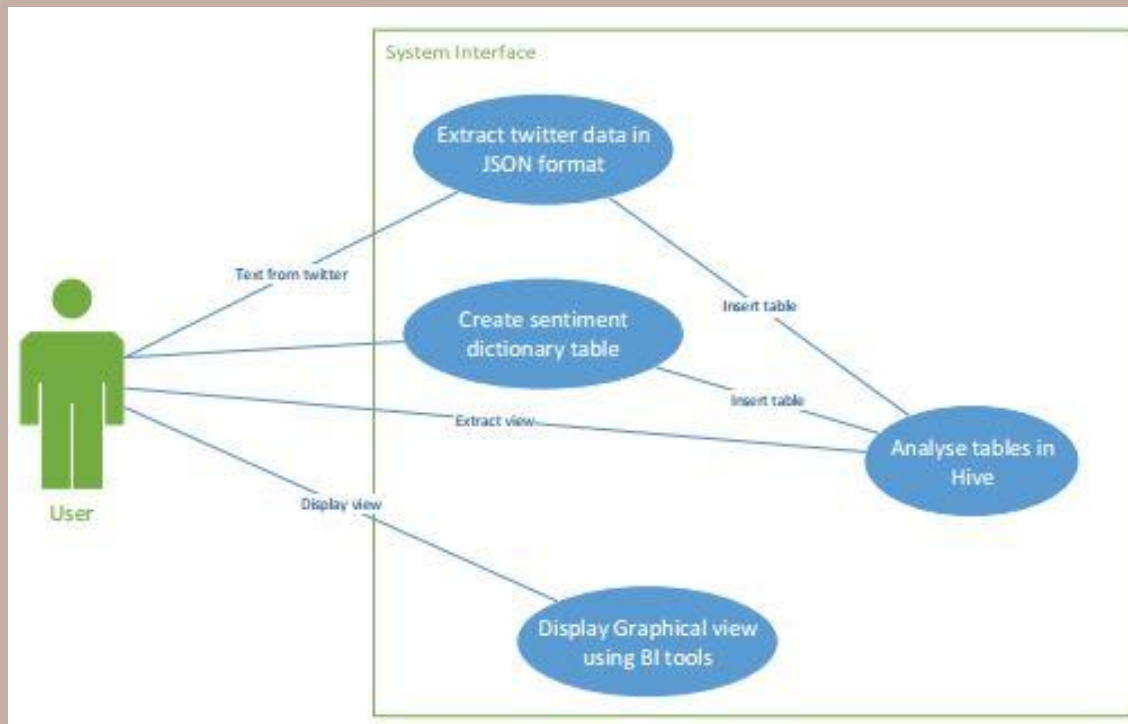
- HortonWorks Data Platform for Windows
- Hive and HiveQL
- BI tools for Excel

Research, Analysis and Design

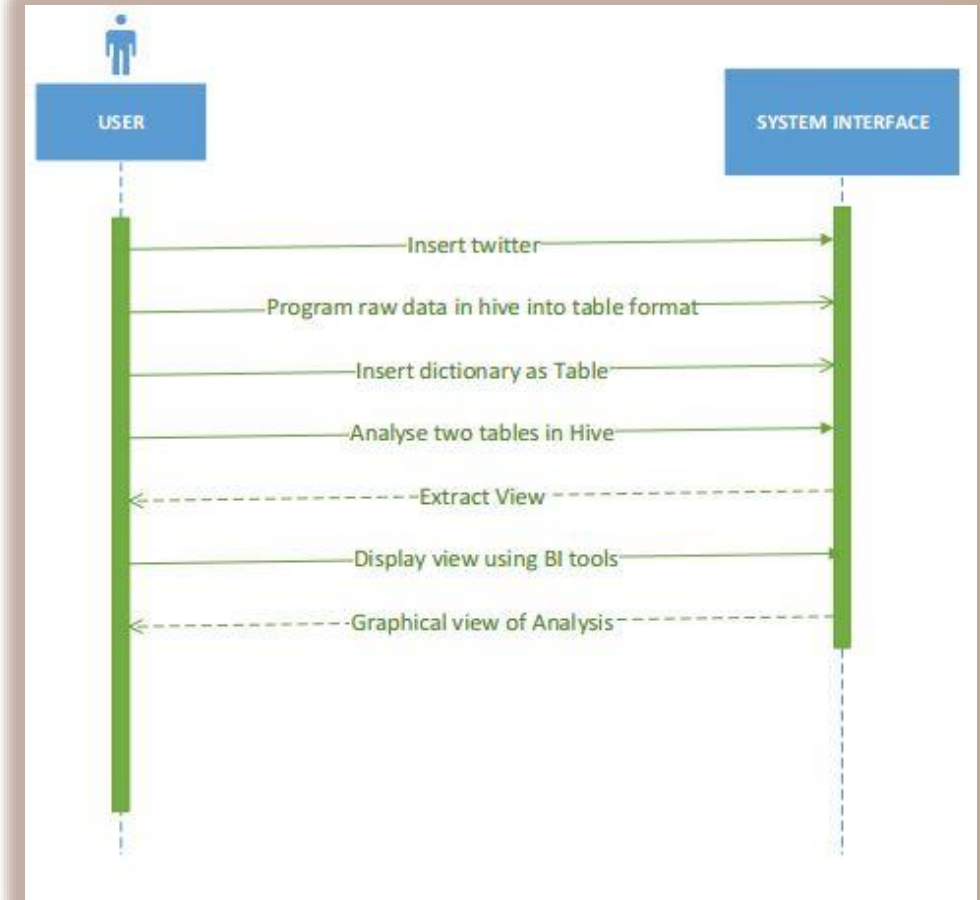
- We had carried out a detail analysis on existing solutions in the market within the project scope
- Followed tutorials on YouTube
- Analyze the raw data, learned about unstructured data. How its been used and managed

Requirements Specification

- Software Requirement Specification draft that includes a UML 2.0 use case, analysis and Sequence models



Use Case Diagram



Sequence Diagram

Design Specification

- Software Design Specification includes a UML 2.0 design model and a data model

Test and Deliver

- Product Tests specified with final and working version of the application with unit testing and system testing.

What Is Structured Data ?

- Data that resides in a fixed field within a record or file is called structured data including relational databases and spreadsheets
- Structured data first depends on creating a data model – a model of the types of business data that will be recorded and **how they will be stored, processed and accessed**
- Structured data has the advantage of being **easily entered, stored, queried and analyzed**
- At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data

What Is Unstructured Data ?

- Unstructured data, usually binary data that is proprietary, is that which has no identifiable internal structure
- Unstructured data is all those things that can't be so readily classified and fit into a neat box: **photos and graphic images, videos, streaming instrument data, webpages, pdf files, PowerPoint presentations, emails, blog entries, wikis and word processing documents**
- 80% of business-relevant information originates in unstructured form, primarily text

- ---

What is Hive ?
Why Hive ?
What is HiveQL?
HiveQL Operations?
What is Hortonworks Data Platform (HDP)?
HDP System Requirements
Setting HDP on Virtual Environment.

-



Pinaki Ranjan Ghosh
Application Developer / Designer

Hive

Large datasets stored in Hadoop's HDFS

Querying

Managing

Summarization

Analysis

- Tools to enable easy data extract/transform/load (ETL)
- A mechanism to impose structure on a variety of data formats
- Access to files stored either directly in **HDFS** or in other data storage systems
- Query execution via MapReduce

Hive

(Cont...)

Hive is a data-warehouseing infrastructure for Hadoop

Warehoused data

Easy to retrieve and Easy to manage.

The data are organized in three different formats in
HIVE

- **Tables:** They are very similar to RDBMS tables and contains rows and tables.
- **Partitions:** Hive tables can have more than one partition like subdirectories and file systems
- **Buckets:** Data may be divided into buckets which are stored as files in partition in the underlying file system.



HiveQL

- *HiveQL* is the Hive query language
- It is a SQL-like interface on top of Hadoop
- Hive converts queries written in HiveQL into MapReduce tasks that are then run across the **Hadoop cluster** to fetch the desired results
- Examples:
 1. Create TABLE sample_table (name String, age int);
 2. LOAD DATA LOCAL PATH 'input/mydata/data.txt' INTO TABLE mytable;
 3. Insert into birthday Select firstname, lastname, birthday from customers where birthday is NOT NULL;
 4. Select * from myTable;

HiveQL

Main Operations...

- Create and manage tables and partitions
- Support various Relational, Arithmetic and Logical Operators
- Evaluate functions
- Download the contents of a table to a local directory or result of queries to HDFS directory

ANALYZE TABLE

DESCRIBE COLUMN

DESCRIBE DATABASE

EXPORT TABLE

IMPORT TABLE

LOAD DATA

SHOW TABLE EXTENDED

SHOW INDEXES

SHOW COLUMNS

Hortonworks Data Platform (HDP)

- Hortonworks and Microsoft have partnered to bring the benefits of Apache Hadoop to Windows
- HDP provides an enterprise ready data platform that enables organizations to adopt a Modern Data Architecture and provide Hadoop data platform.
- With HDP for Windows, Hadoop is both **simple to install and manage**.
- **Familiar Tools on Hadoop** : The new offering enables the application of rich business intelligence (BI) tools such as **Microsoft Excel, PowerPivot for Excel and Power View** to pull actionable insights from not just big data but all of your enterprise data sources.



Hortonworks Data Platform (HDP) Types

HDP Sandbox

Runs on VirtualBox or VMWare

- Host Operating Systems: Windows 7, 8
- Virtual Machine : Virtual Box, VMWare or VMFusion

Automated (Ambari)

RHEL/CentOS/SLES (64-bit)

- Red Hat Enterprise Linux • CentOS • Oracle Linux • SUSE Linux Enterprise Server

Windows

Windows Server 2008 & 2012

- Windows Server 2008 R2 (64-bit) • Windows Server 2012 (64-bit)

HDP Minimum System Requirements



- Hosts:

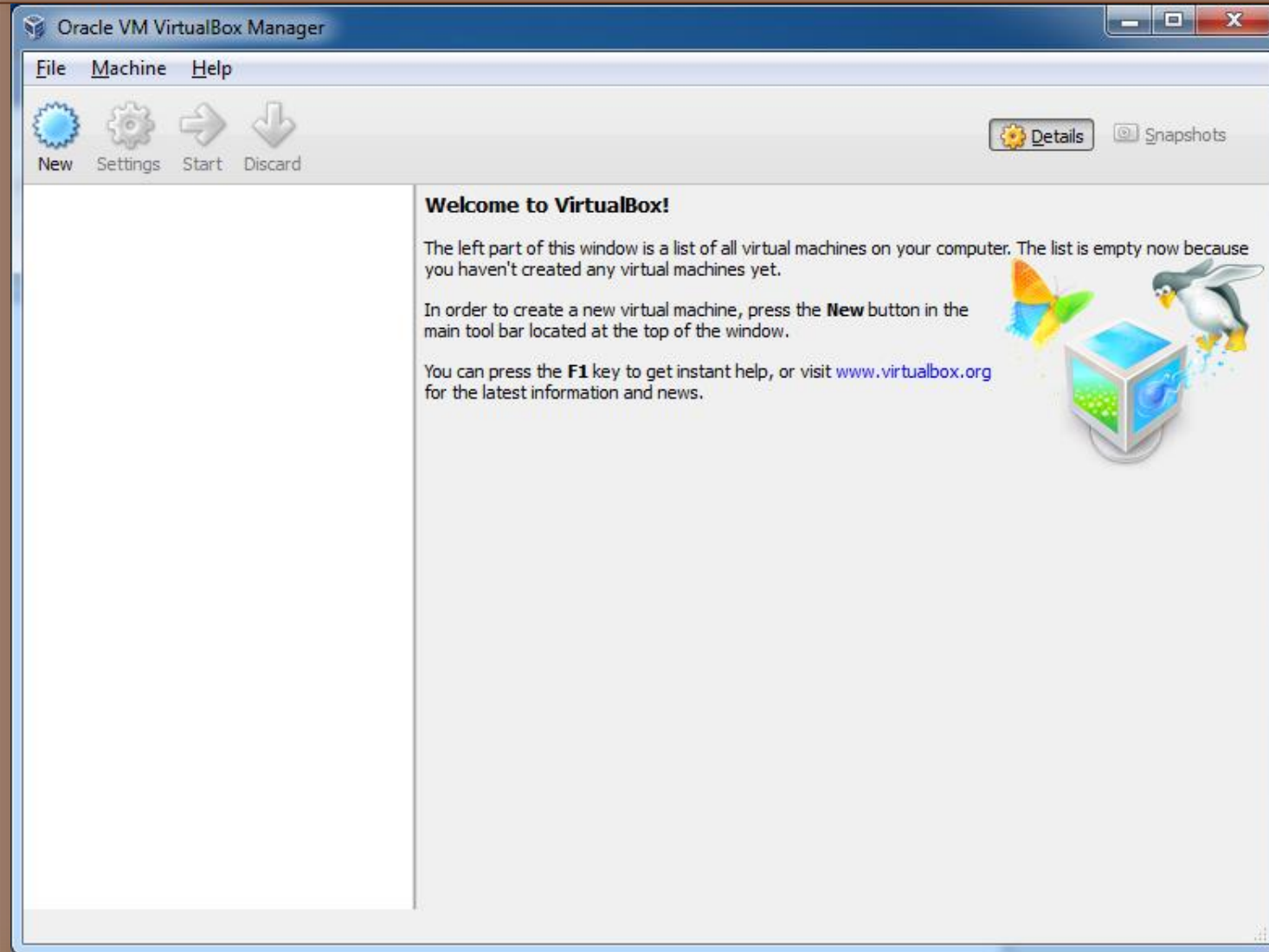
A 64-bit machine with a chip that supports virtualization.

A BIOS that has been set to enable virtualization support.

- Host Operating Systems : Windows 7, 8
- Supported Browsers: Internet Explorer , Google Chrome, Firefox
- At least 4 GB of RAM (Divide Total RAM by half between Host and Virtual Machine)
- Virtual Machine Environments: Oracle Virtual Box - version 4.2 or later, VMware, VMware Fusion, version 5.x (For Mac)

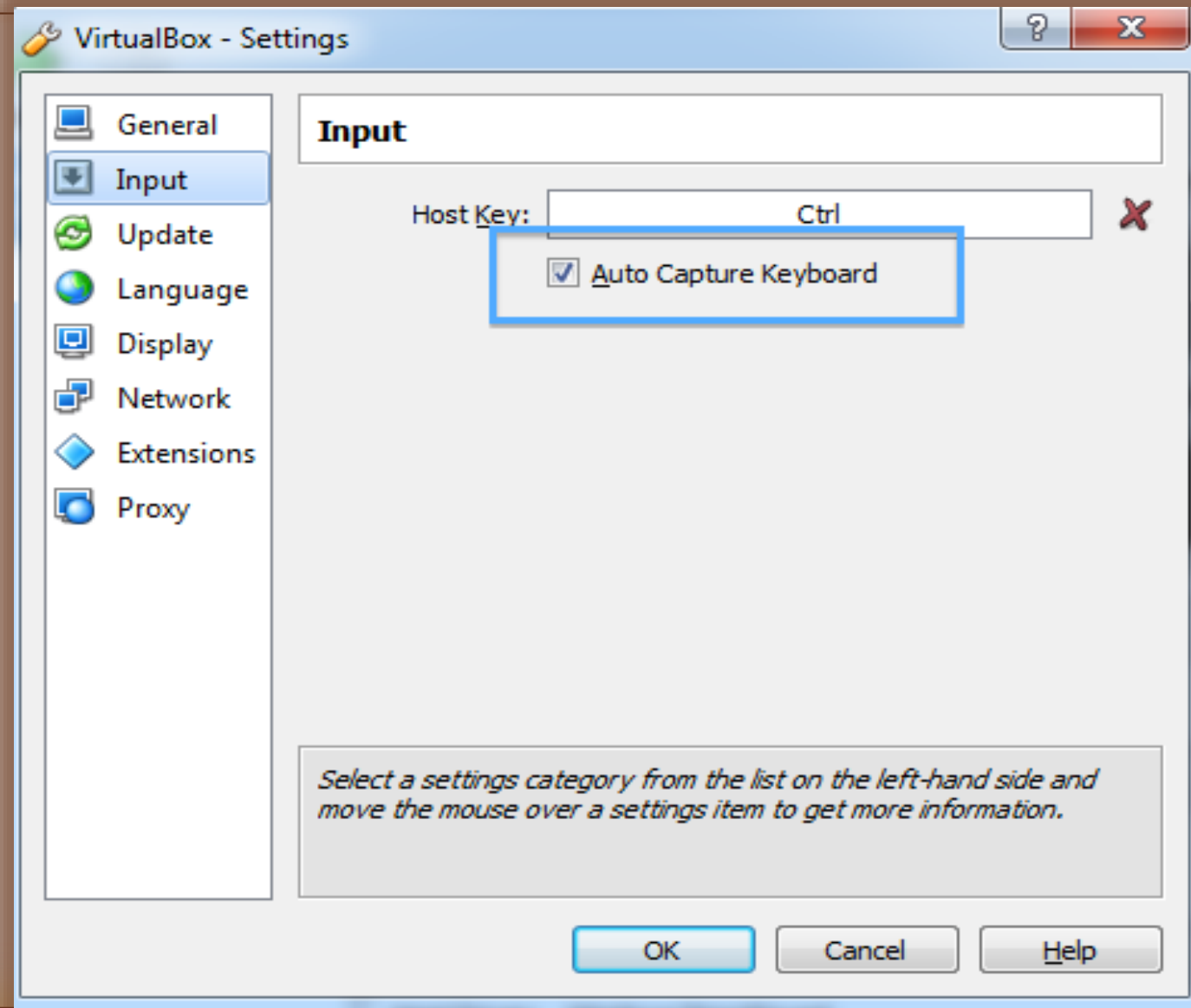


Setting up HDP inside Virtual Machine



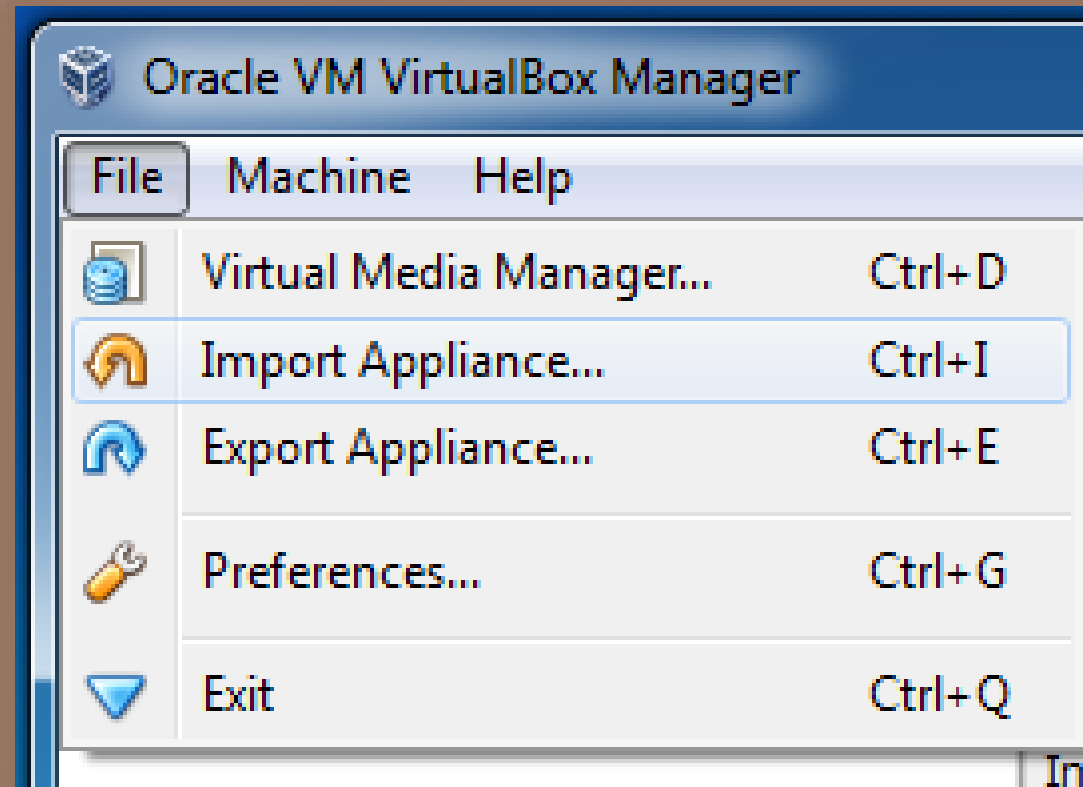
Setting up HDP inside Virtual Machine

(Cont...)



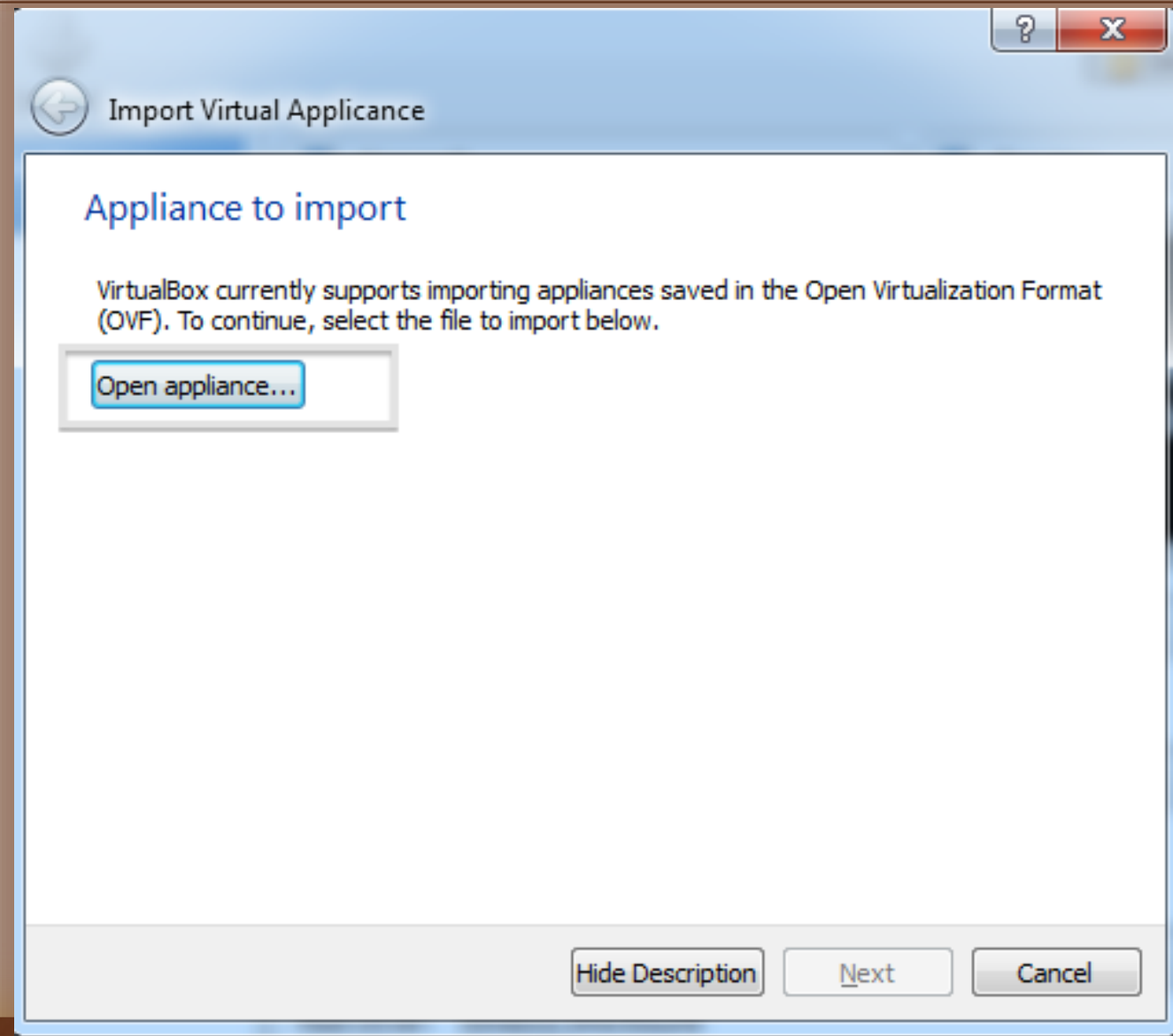
Setting up HDP inside Virtual Machine

(Cont...)



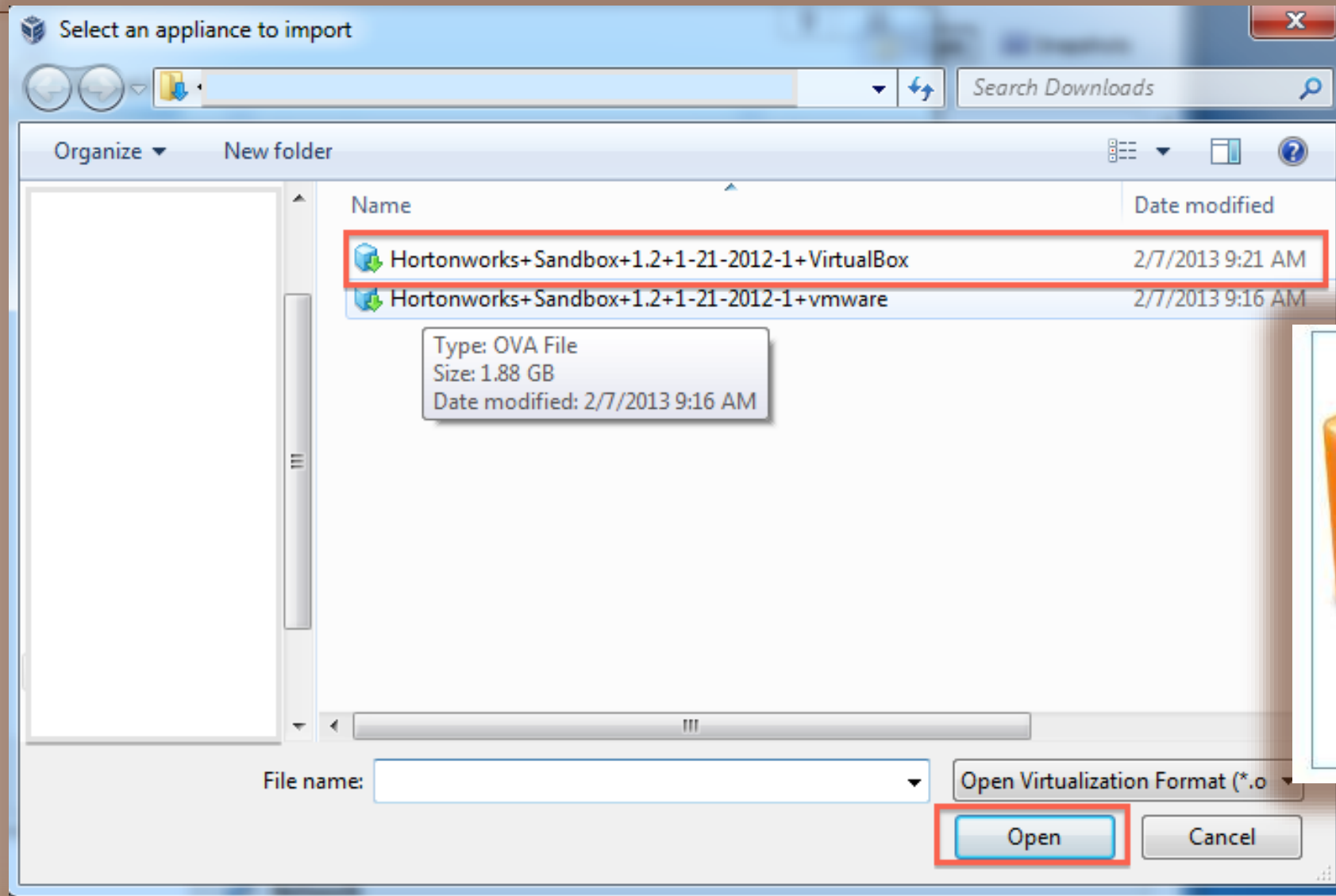
Setting up HDP inside Virtual Machine

(Cont...)



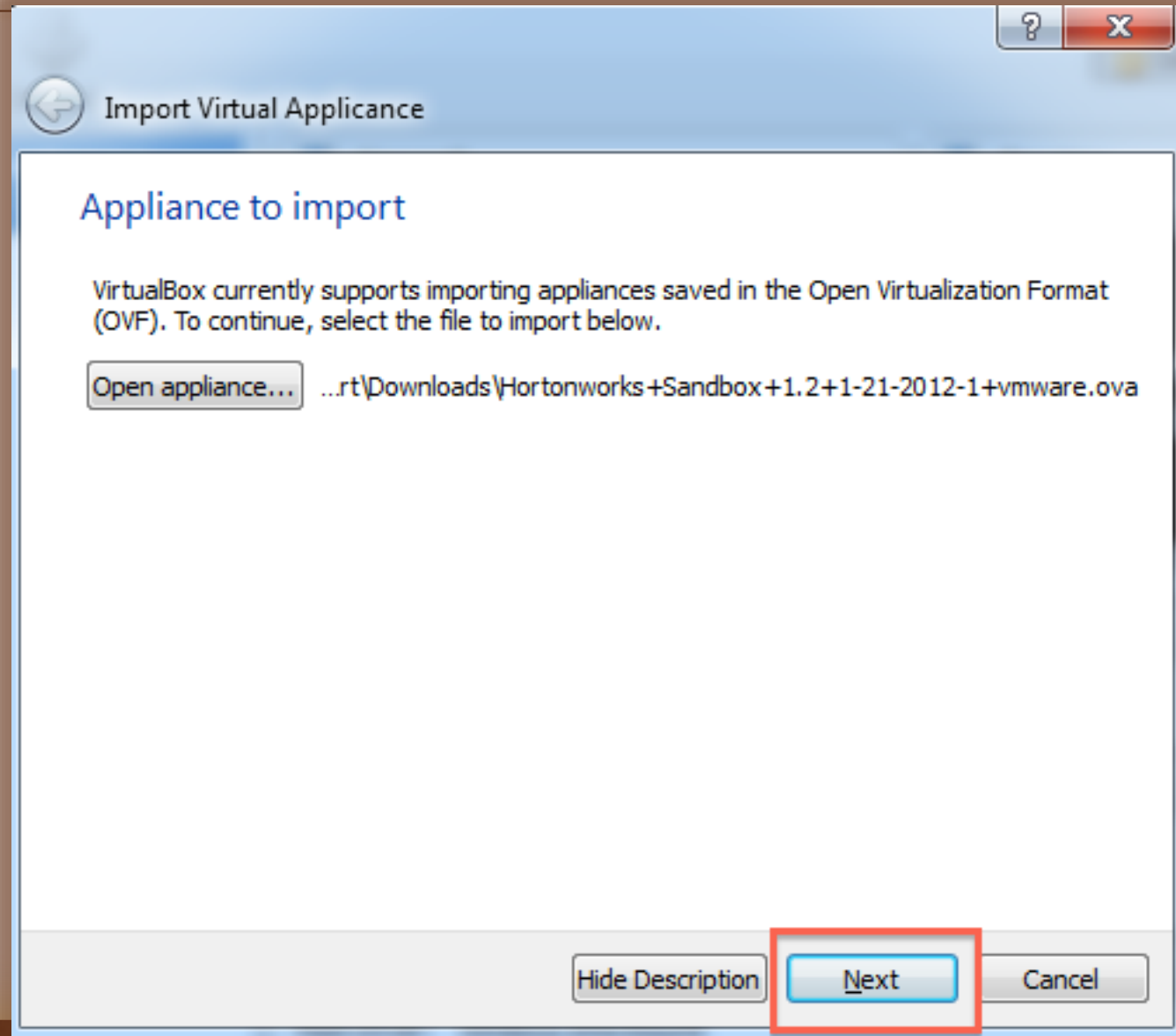
Setting up HDP inside Virtual Machine

(Cont...)



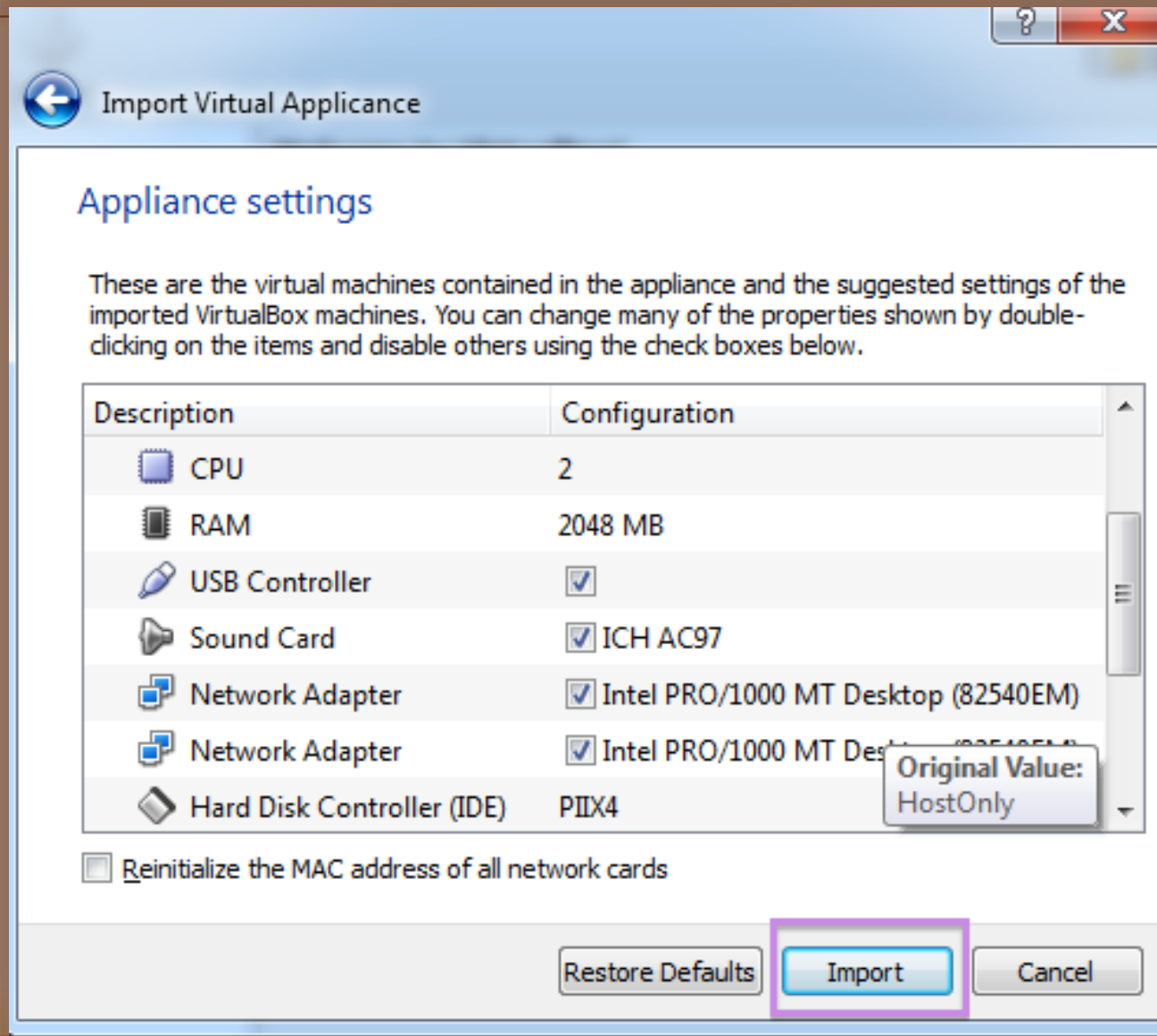
Setting up HDP inside Virtual Machine

(Cont...)



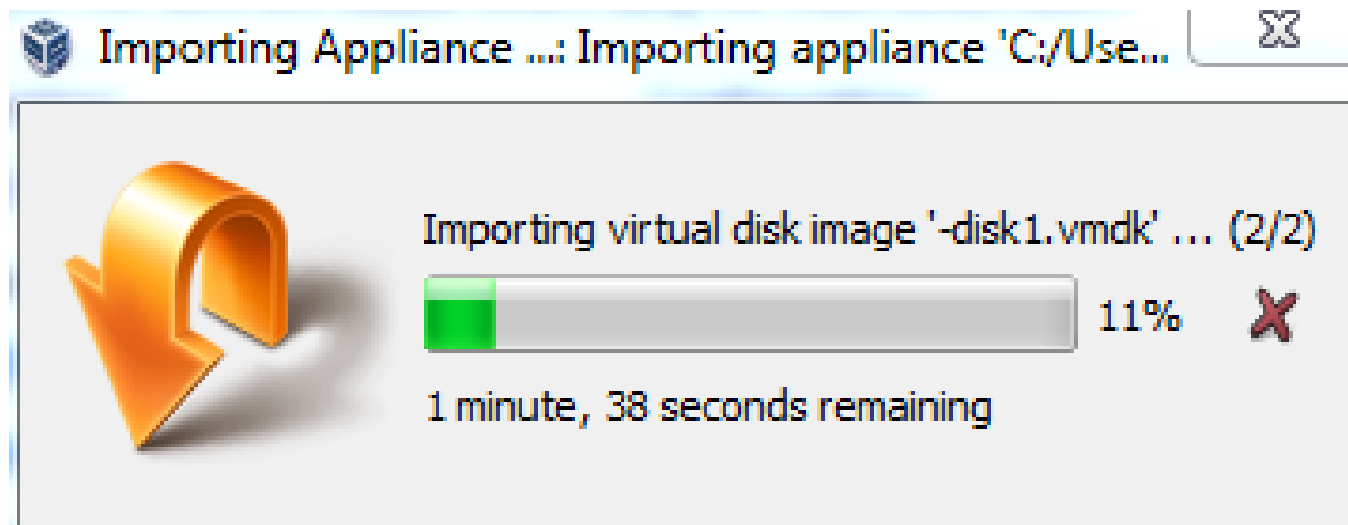
Setting up HDP inside Virtual Machine

(Cont...)



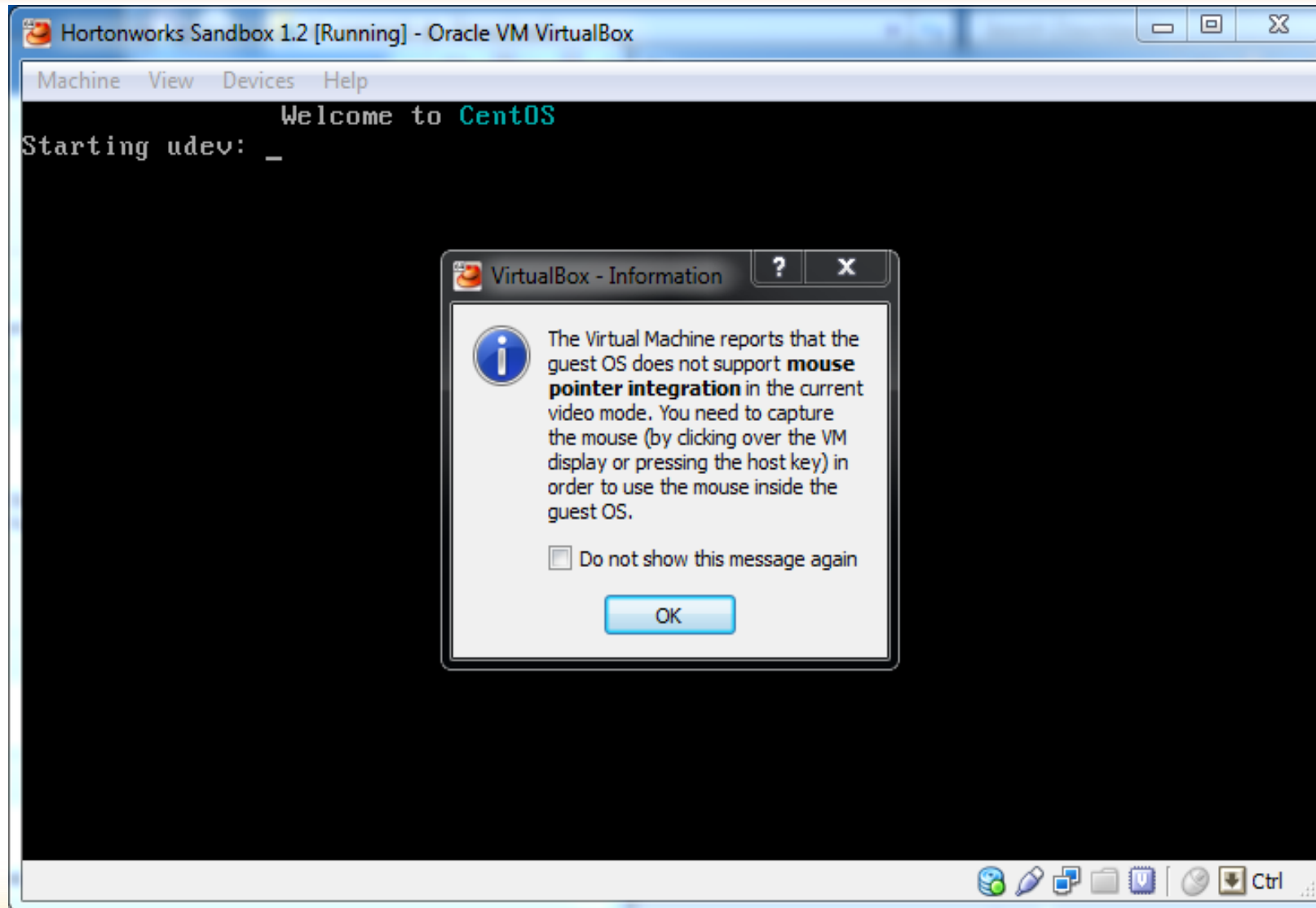
Setting up HDP inside Virtual Machine

(Cont...)

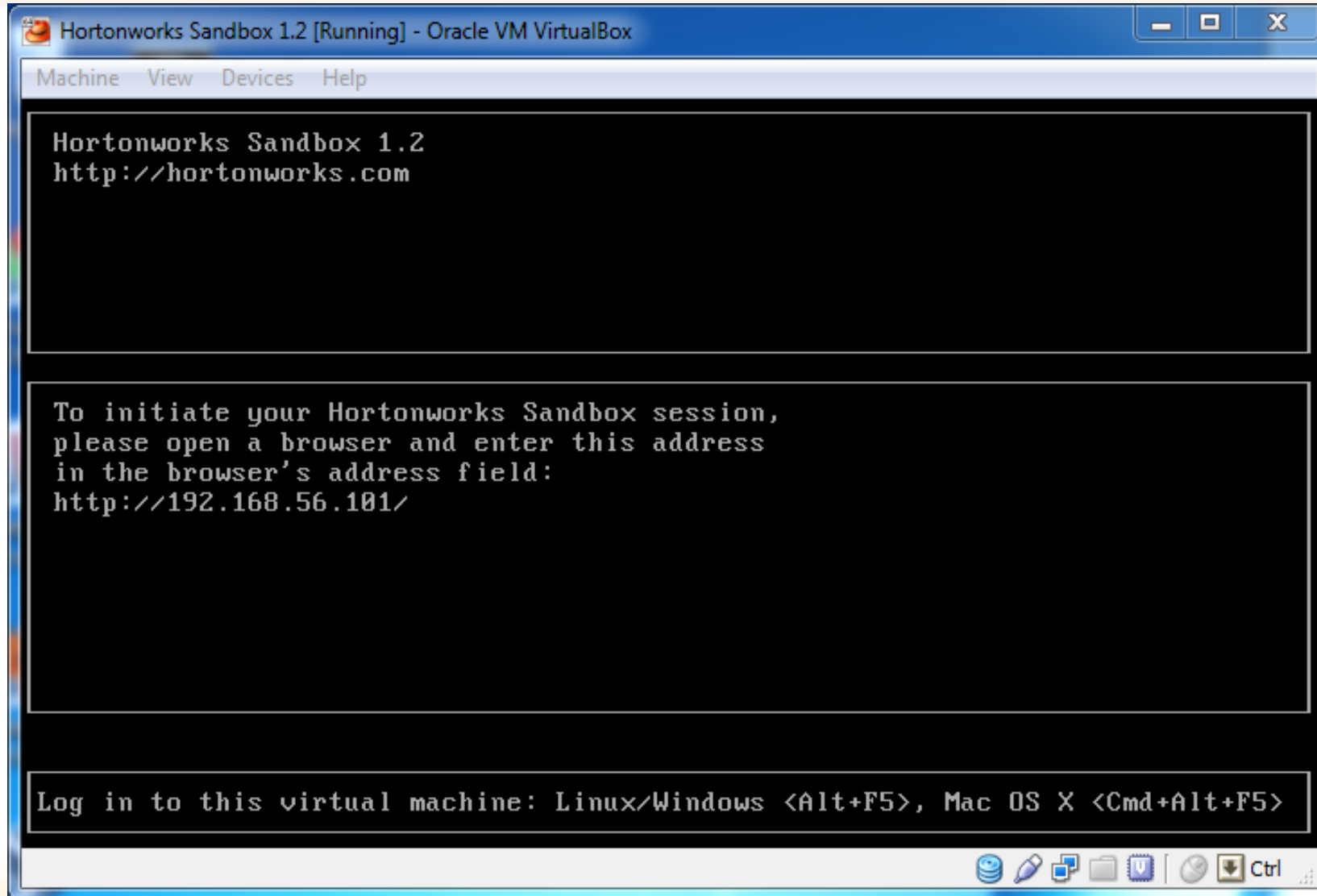


Setting up HDP inside Virtual Machine

(Cont...)



HDP Console Interface



HDP Web Interface at 127.0.0.1:8888



What is JSON file ?
What is Raw Data ?
What is JSON Serde file ?
How to load external data into Hive ?
from windows machine
What is Dictionary File ?



Kiranmayi Ganti
Application Developer / Maintenance

What is JSON file ?

- **JSON** (JavaScript Object Notation) is a lightweight data-interchange format
- It is easy for humans to read and write. It is easy for machines to parse and generate
- It is based on a subset of the JavaScript Programming Language

What is Raw Data ?

- Raw data is the data generated from twitter in JSON format using twitter API 1.1.
- The data has fields such as:
 - Name
 - Screen
 - Date time
 - Text
 - Hash tag
- These fields are generated when a user tweets or retweets .
- There are many other fields in the data for a particular record, which are not required for the analysis

Sample raw data

➤ {"filter_level":"medium","contributors":null,"text":"Really wanna see Iron Man 3 o-
o","geo":null,"retweeted":false,"in_reply_to_screen_name":null,"truncated":false,"lang":"en","entities
":{"symbols":[],"urls":[],"hashtags":[],"user_mentions":[]},"in_reply_to_status_id_str":null,"id":330064153
572163585,"source":"web","in_reply_to_user_id_str":null,"favorited":false,"in_reply_to_status_id":null,
"retweet_count":0,"created_at":"Thu May 02 21:00:01 +0000
2013","in_reply_to_user_id":null,"favorite_count":0,"id_str":"330064153572163585","place":null,"user":{"locat
ion":"Essex, UK.
","default_profile":false,"statuses_count":10702,"profile_background_tile":false,"lang":"en","profile_link_co
lor":"93A644","profile_banner_url":"https://sio.twimg.com/profile_banners/395521131/1363636228","id":395
521131,"following":null,"favourites_count":2963,"protected":false,"profile_text_color":"8D7916","description"
:"17. 6ft2.
http://ask.fm/Jayshaww","verified":false,"contributors_enabled":false,"profile_sidebar_border_color":"000
000","name":"Jay Shaw","profile_background_color":"B2DFDA","created_at":"Fri Oct 21 20:00:16 +0000
2011","default_profile_image":false,"followers_count":206,"profile_image_url_https":"https://sio.twimg.co
m/profile_images/3602472505/0a77b1f4a8ec3558e63dbdbb476a1d74_normal.jpeg","geo_enabled":false,"pro
file_background_image_url":"http://ao.twimg.com/profile_background_images/818049845/2f3e884115bbb
53b72285770b2847676.jpeg","profile_background_image_url_https":"https://sio.twimg.com/profile_backg
round_images/818049845/2f3e884115bbb53b72285770b2847676.jpeg","follow_request_sent":null,"url":"http
://Youtube.com/JaysRants","utc_offset":0,"time_zone":"Casablanca","notifications":null,"profile_use_back
ground_image":true,"friends_count":125,"profile_sidebar_fill_color":"215A90","screen_name":"Jayshaww","i
d_str":"395521131","profile_image_url":"http://ao.twimg.com/profile_images/3602472505/0a77b1f4a8ec3558
e63dbdbb476a1d74_normal.jpeg","listed_count":0,"is_translator":false},"coordinates":null}

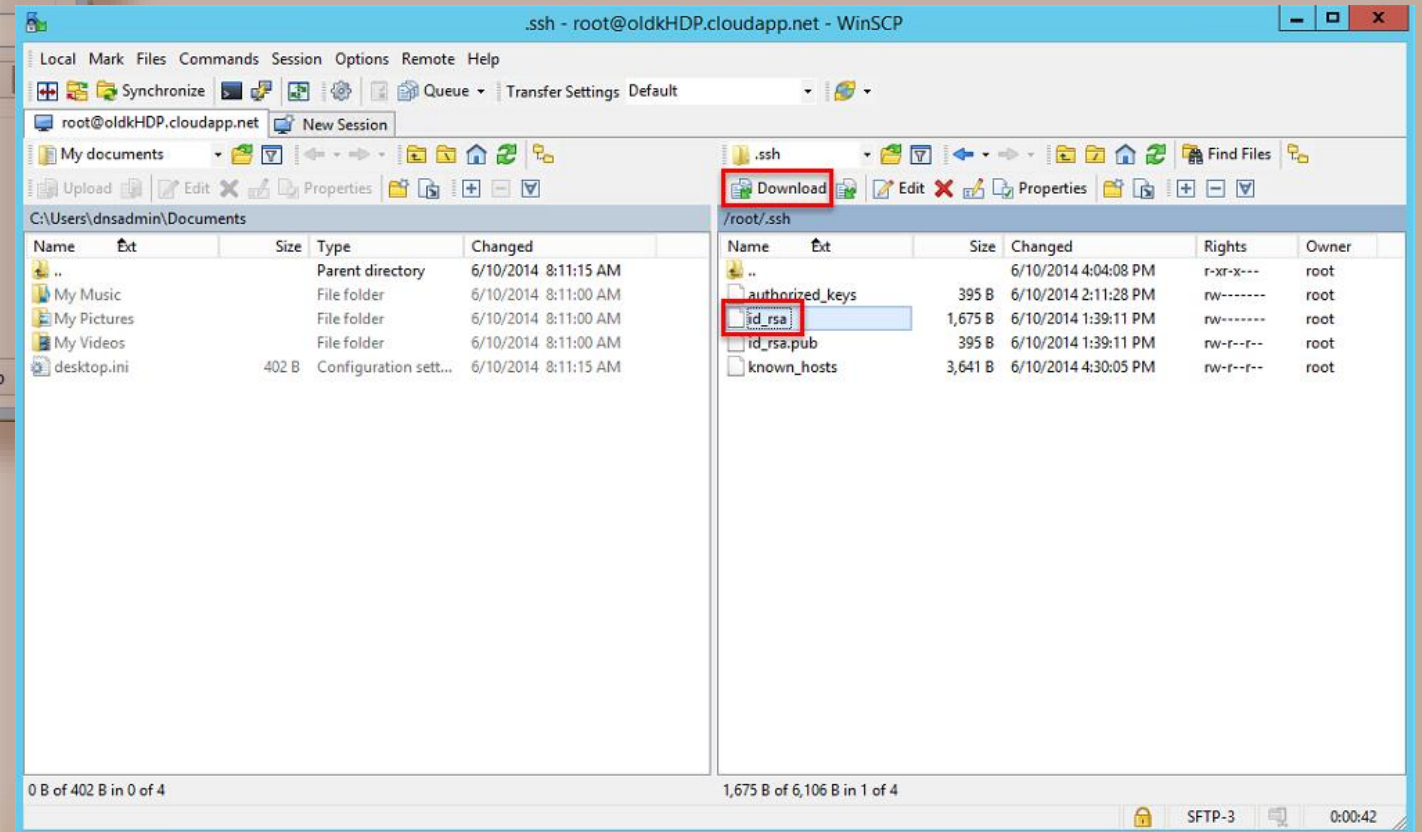
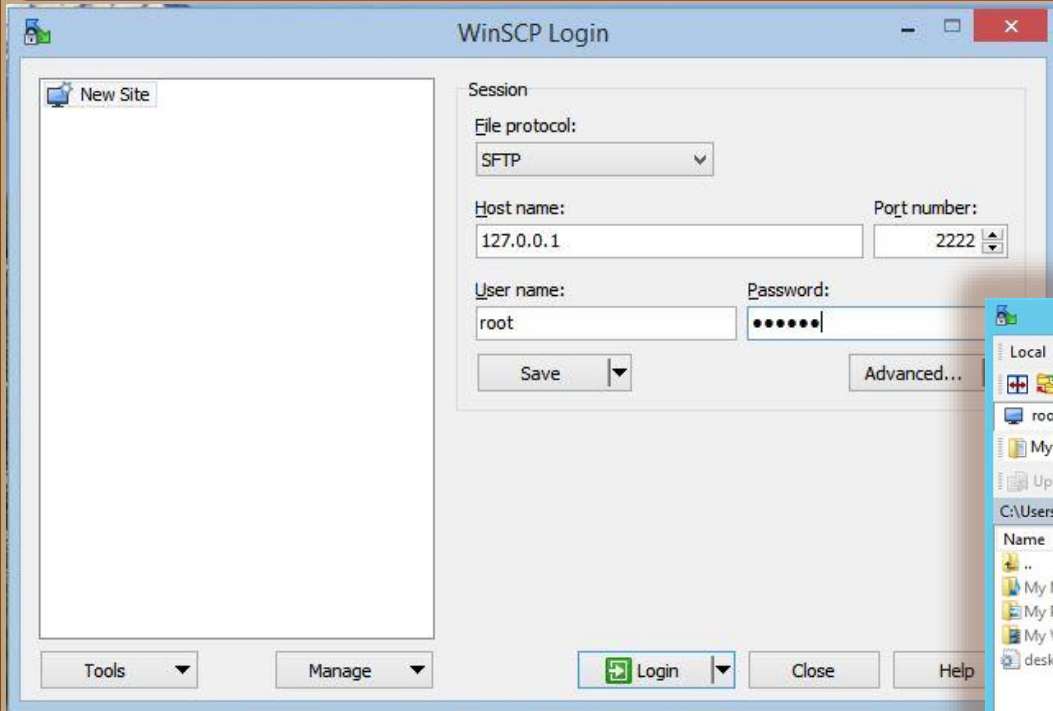
What is a JSON Serde File

- SerDe is short for Serializer/Deserializer.
- Hive uses the SerDe interface for IO.
- A SerDe allows Hive to read in data from a table, and write it back to HDFS in any custom format.
- Here we are using SerDe for row format.
- For JSON files, Amazon has provided a JSON SerDe.

Loading external data into Hive from Windows Machine

- Raw data and JSON SerDe files are the external data
- Hive uses external data and JSON SerDe file to load external tables
- These external files are transmitted from windows to Hadoop environment, using a win SCP recommended by Hortonworks
- It is a interface to access remote system from local machine, and store files and data from an external resource
- Here remote system is hortonworks sandbox and external resource is the external data

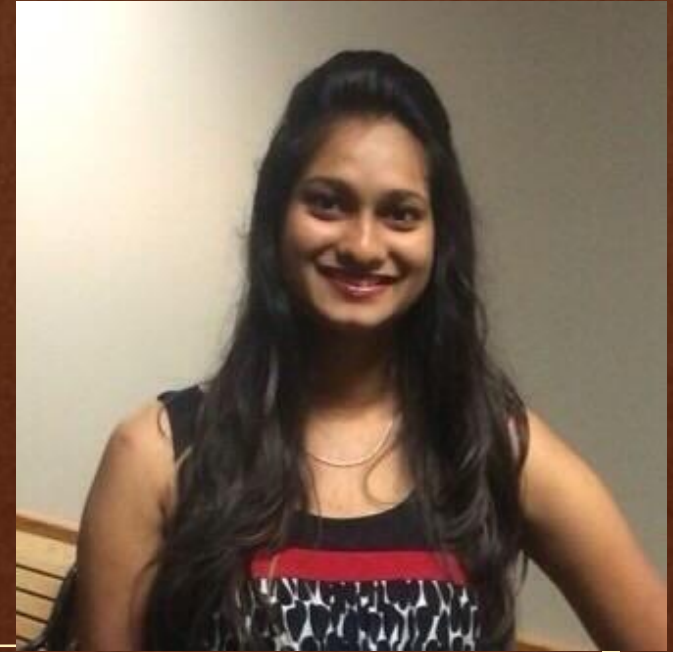
WinSCP Screen Shots



What is Dictionary File?

- It is text file with .tsv format.
- Data is arranged in three columns
- First column is the behavior of the word. A word can have weak subject or strong subject.
- Second column contains the word.
- Third column is the polarity of the word.
- Before every word, the polarity of each word is saved i.e. positive , negative or neutral.

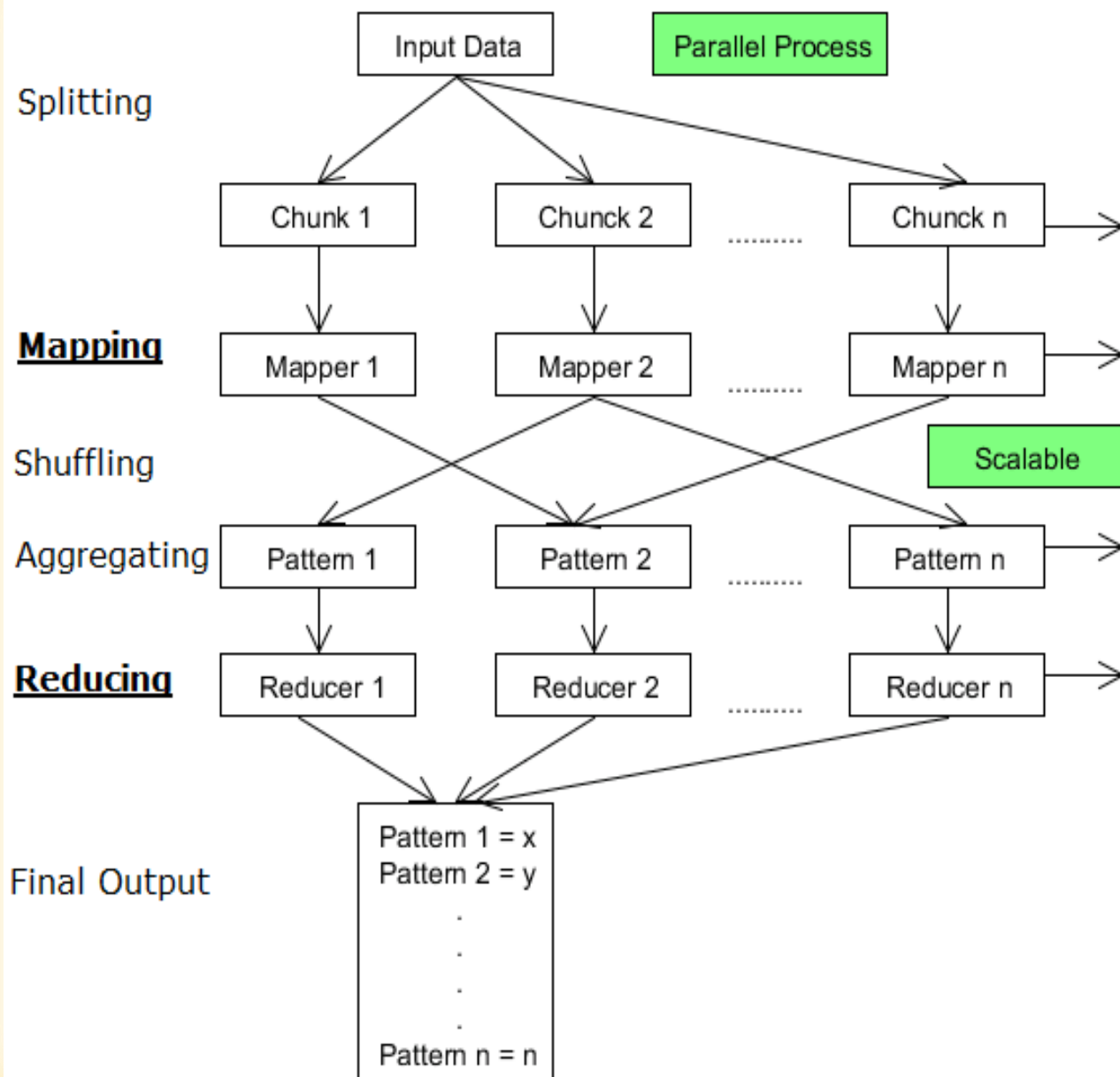
MAP and REDUCE functions in Hadoop
Division of Data Words
Business Intelligence Tools
How to connect HDP to MS-Excel ?
Power Query via CSV
Challenges and Overcomes



Srijha Reddy Gangidi
Application Developer / Tester

MAP and REDUCE functions in Hadoop

- **MapReduce** is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.
- A MapReduce program is composed of a **Map()** procedure that performs filtering and sorting
- A **Reduce()** procedure that performs a summary operation
- MapReduce can take advantage of locality of data, processing it on or near the storage assets in order to reduce the distance over which it must be transmitted.



- **"Map" step:** Each worker node applies the "map()" function to the local data, and writes the output to a temporary storage. A master node orchestrates that for redundant copies of input data, only one is processed.
- **"Shuffle" step:** Worker nodes redistribute data based on the output keys (produced by the "map()" function), such that all data belonging to one key is located on the same worker node.
- **"Reduce" step:** Worker nodes now process each group of output data, per key, in parallel.

Division of Positive, Negative and Neutral Data Words

- The identification of subjective opinion on text data involves the classification of text into **three categories** :

Positive, Negative and Neutral.

- Positive sentiment is measured in a similar way by looking for positive words not preceded by a negation.
- Similarly the negative sentiment is measured by looking for negative words.
- Neutral sentiment is measured by looking for positive words preceded by a negation or vice versa.

Business Intelligence (BI) Tools

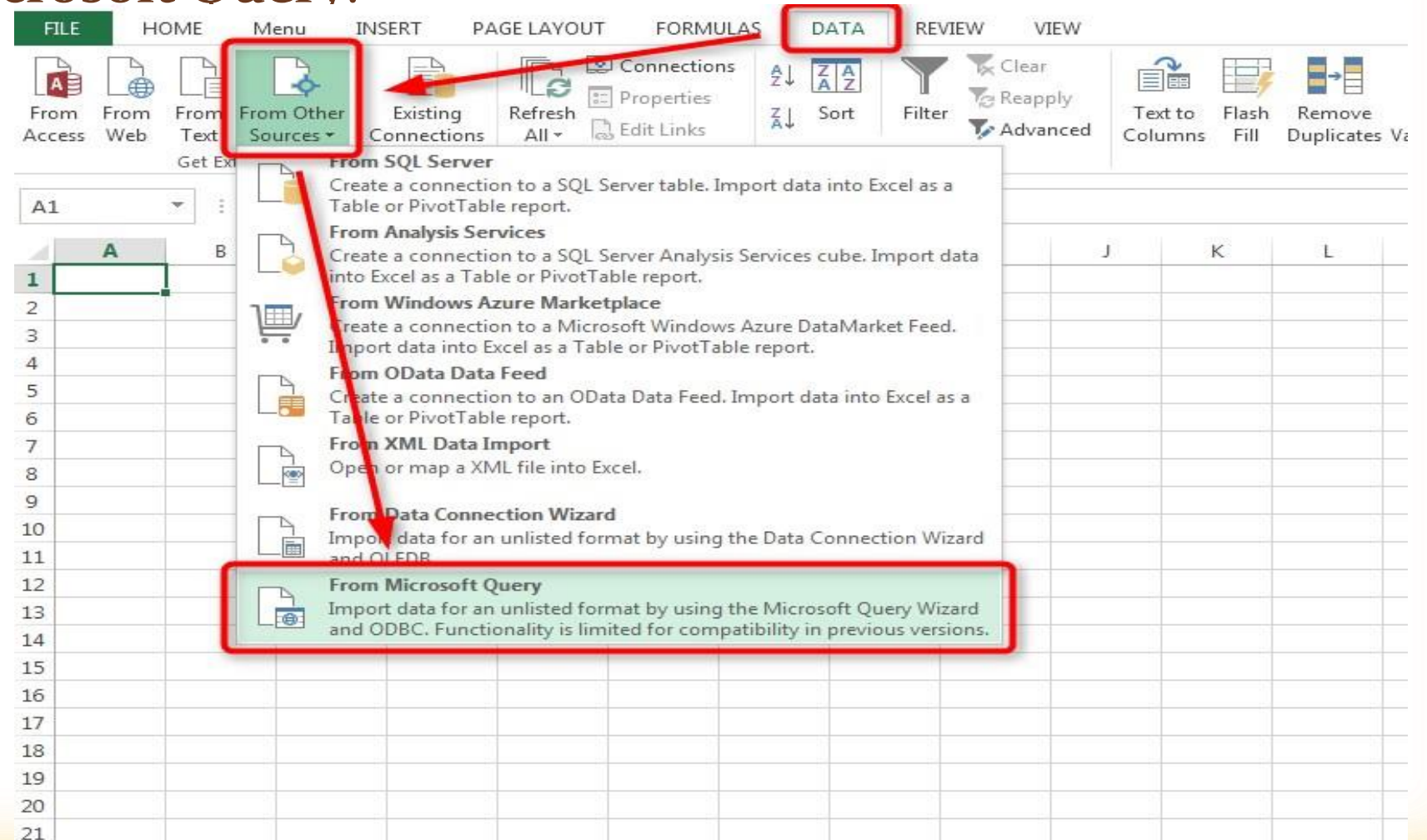
- **Business intelligence tools** are a type of application software designed to retrieve, analyze, transform and report data for **business intelligence**.
- The **tools** generally read data that have been previously stored in a data warehouse or data mart.
- The business intelligence (BI) represents the tools and systems that play a key role in the strategic planning process of the corporation. These systems allow a company to gather, store, access and analyze corporate data to aid in decision-making.

How to connect HDP to MS-Excel

- We use the **Power View** feature in **Excel 2013** to visualize the sentiment data. Other versions of Excel will work, but the visualizations will be limited to charts.
- Install the ODBC driver that matches the version of Excel you are using (32-bit or 64-bit).
- Connecting HDP to MS-Excel involves:
 - Accessing the refined sentiment data with Excel
 - Visualize the sentiment data using Excel Power View

Access the Refined Sentiment Data with Excel

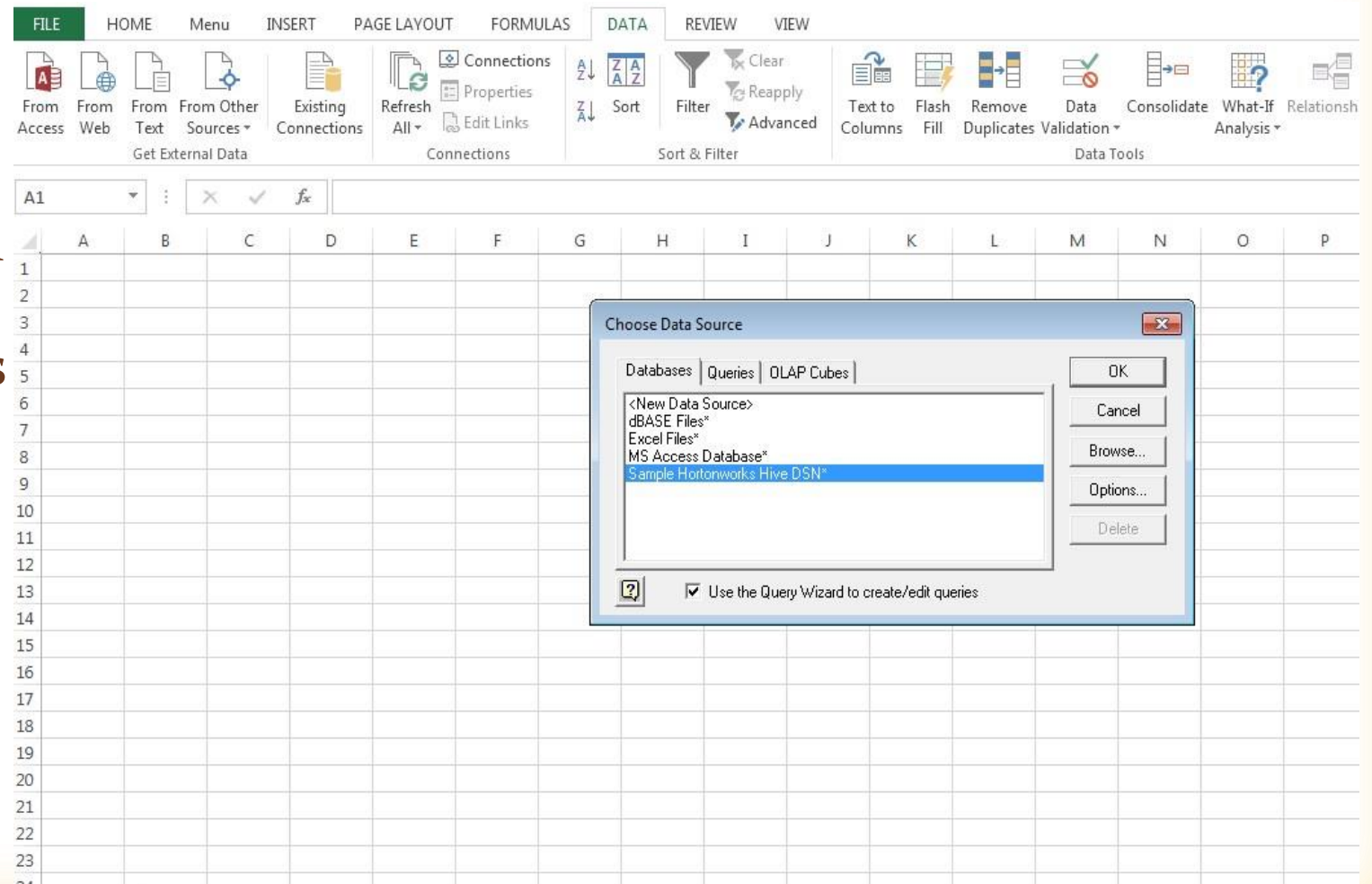
- In Windows, open a new Excel workbook, then select **Data > From Other Sources > From Microsoft Query**.



BI –Tools in Excel

(Cont..)

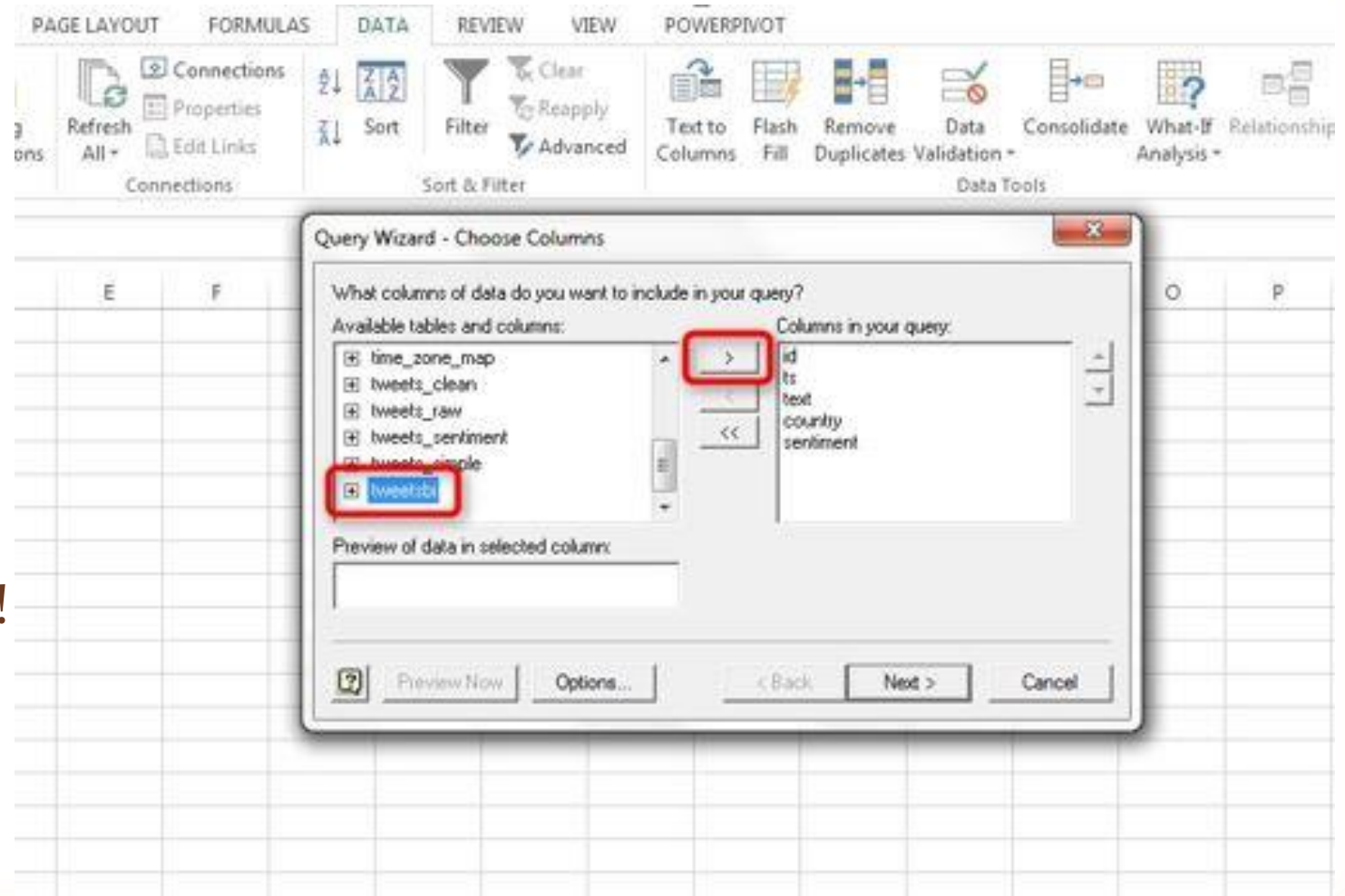
- On the Choose Data Source pop-up, select the Hortonworks ODBC data source you installed previously, then click **OK**.
- The Hortonworks ODBC driver enables you to access Hortonworks data with Excel and other Business Intelligence (BI) applications that support ODBC



BI –Tools in Excel

(Cont..)

- 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
- ODBC configuration ERROR!

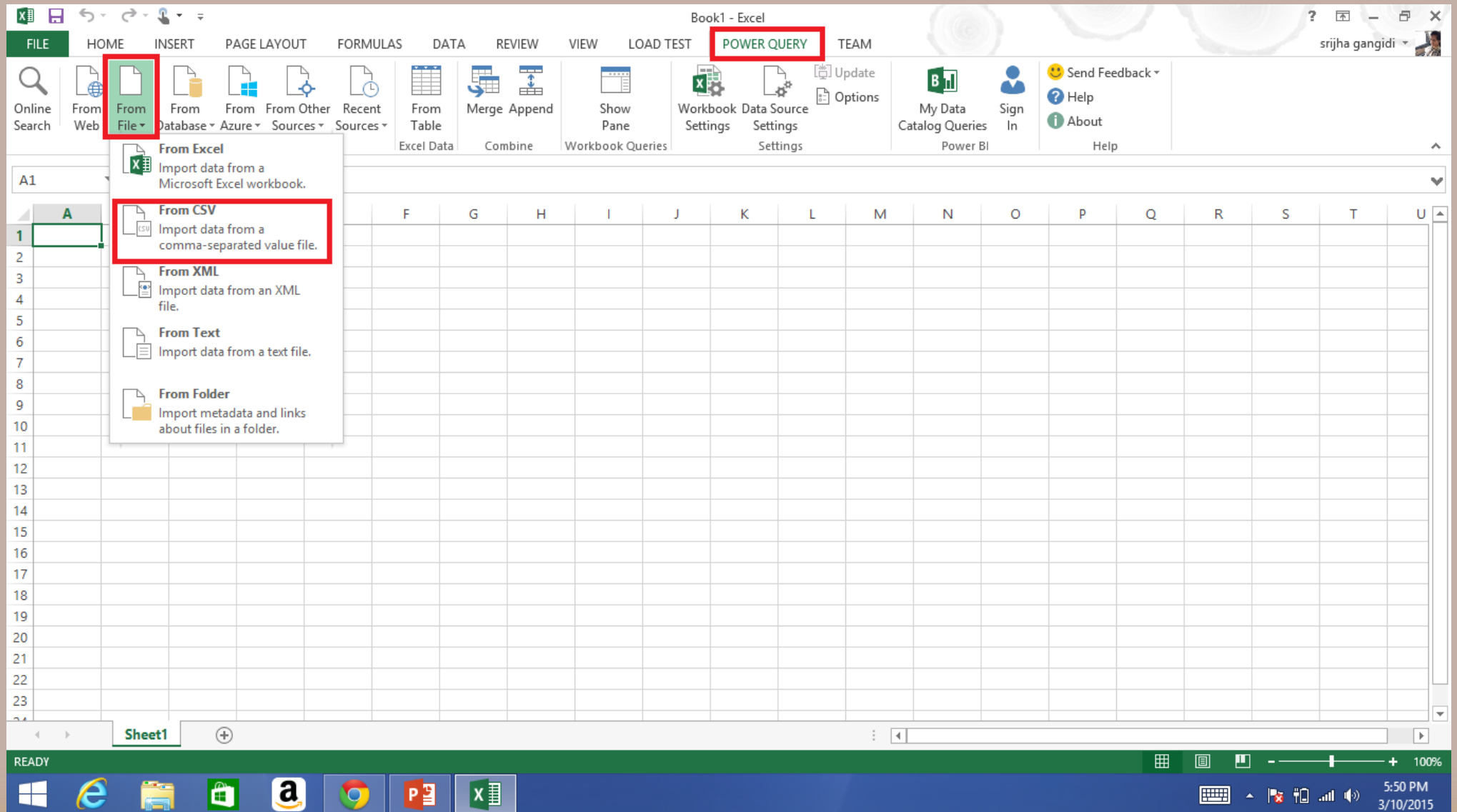


Power Query via CSV file

An alternative approach to BI –Tools in Excel

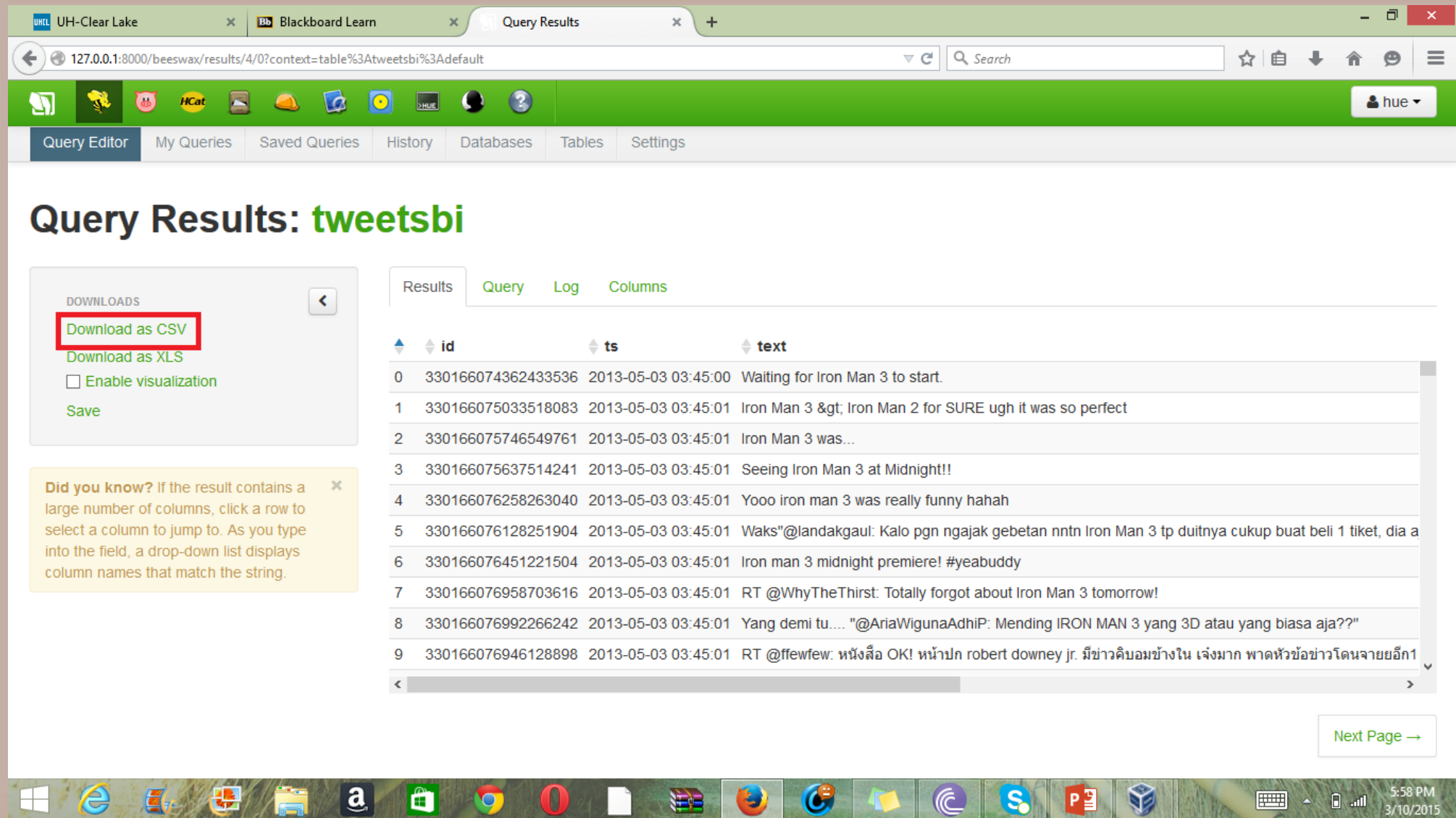
- Install power view and power query in MS Excel
- Export the table in CSV format from the web interface
- Open the table in Power Query and manage the table
- Load the manage table into excel worksheet
- Visualize it in Power view using Map view.

Power Query via CSV file – An alternative approach



Power Query via CSV file

(Cont...)



The screenshot shows a web browser window with the Beeswax Query Results interface. The browser tabs include 'UH Clear Lake', 'Blackboard Learn', and 'Query Results'. The address bar shows the URL '127.0.0.1:8000/beeswax/results/4/0?context=table%3Atweetsbi%3Adefault'. The interface has a green header bar with a navigation menu: 'Query Editor', 'My Queries', 'Saved Queries', 'History', 'Databases', 'Tables', and 'Settings'. The main content area is titled 'Query Results: tweetsbi'. On the left, there is a 'DOWNLOADS' section with a red box around 'Download as CSV', and other options like 'Download as XLS', 'Enable visualization', and 'Save'. Below this is a 'Did you know?' tip. The main area displays a table of results with columns 'id', 'ts', and 'text'. The table contains 10 rows of tweet data. At the bottom right, there is a 'Next Page' button. The Windows taskbar at the bottom shows various application icons and the system clock indicating 5:58 PM on 3/10/2015.

Query Results: tweetsbi

DOWNLOADS

- Download as CSV
- Download as XLS
- ☐ Enable visualization
- Save

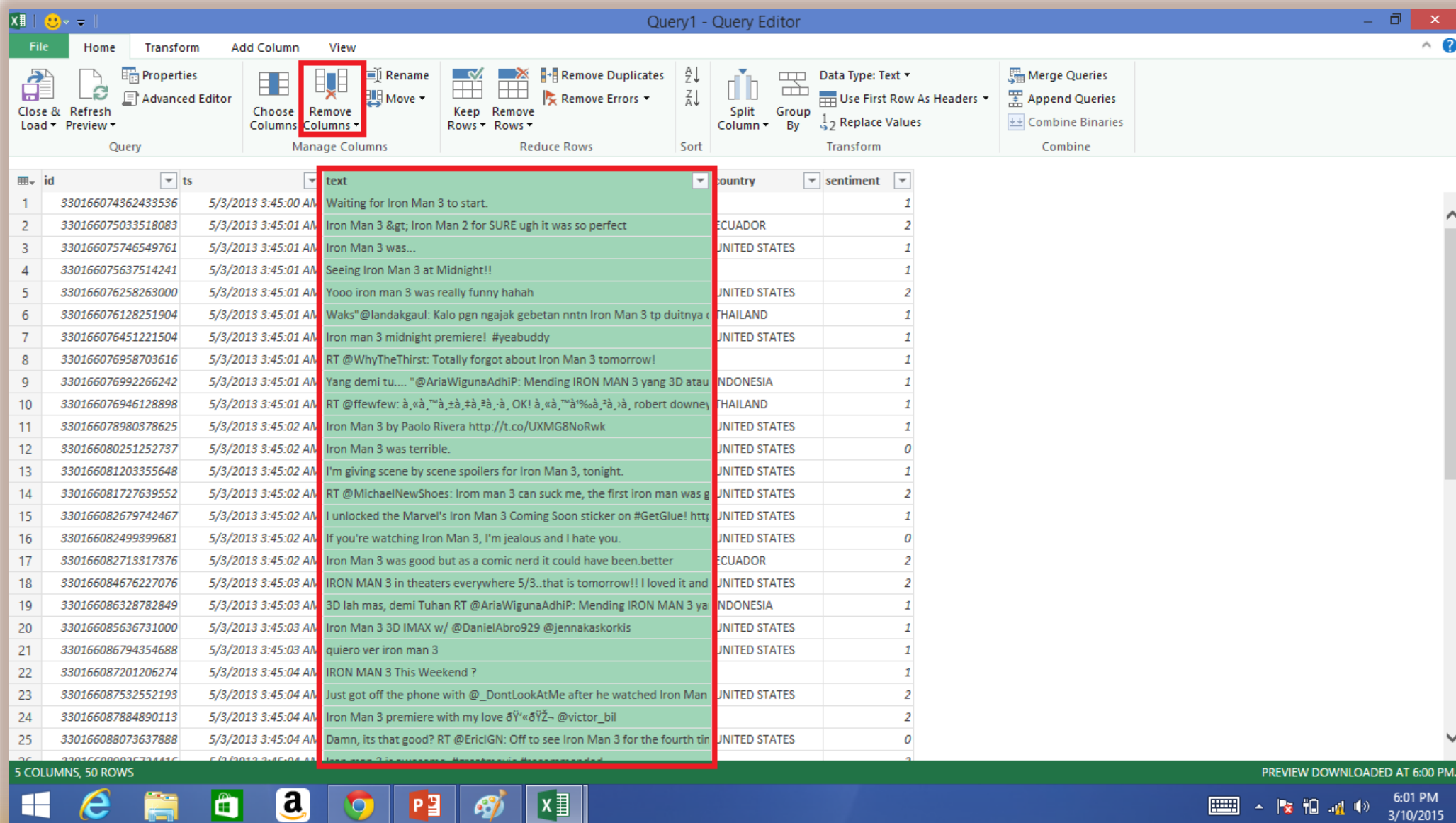
Did you know? If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that match the string.

Results Query Log Columns

	id	ts	text
0	330166074362433536	2013-05-03 03:45:00	Waiting for Iron Man 3 to start.
1	330166075033518083	2013-05-03 03:45:01	Iron Man 3 > Iron Man 2 for SURE ugh it was so perfect
2	330166075746549761	2013-05-03 03:45:01	Iron Man 3 was...
3	330166075637514241	2013-05-03 03:45:01	Seeing Iron Man 3 at Midnight!!
4	330166076258263040	2013-05-03 03:45:01	Yooo iron man 3 was really funny hahah
5	330166076128251904	2013-05-03 03:45:01	Waks"@landakgaul: Kalo pgn ngajak gebetan nntn Iron Man 3 tp duitnya cukup buat beli 1 tiket, dia a
6	330166076451221504	2013-05-03 03:45:01	Iron man 3 midnight premiere! #yeabuddy
7	330166076958703616	2013-05-03 03:45:01	RT @WhyTheThirst: Totally forgot about Iron Man 3 tomorrow!
8	330166076992266242	2013-05-03 03:45:01	Yang demi tu.... "@AriaWigunaAdhiP: Mending IRON MAN 3 yang 3D atau yang biasa aja??"
9	330166076946128898	2013-05-03 03:45:01	RT @fewfew: หนังสือ OK! หน้าปก robert downey jr. มีข้าวคืบอมข้างใน เจ๋งมาก พาดหัวข่าวโดนฉายอีก1

Next Page →

(Cont...)



Power Query via CSV file

(Cont...)

Query1 - Query Editor

File Home Transform Add Column View

Close & Load Refresh Preview Properties Advanced Editor

Choose Columns Remove Columns Rename Move

Keep Rows Remove Rows Remove Duplicates Remove Errors

Split Column Group By Data Type: Whole Number Use First Row As Headers Replace Values

Merge Queries Append Queries Combine Binaries Combine

	id	ts	country	sentiment
1	330166074362433536	5/3/2013 3:45:00 AM		1
2	330166075033518083	5/3/2013 3:45:01 AM	ECUADOR	2
3	330166075746549761	5/3/2013 3:45:01 AM	UNITED STATES	1
4	330166075637514241	5/3/2013 3:45:01 AM		1
5	330166076258263000	5/3/2013 3:45:01 AM	UNITED STATES	2
6	330166076128251904	5/3/2013 3:45:01 AM	THAILAND	1
7	330166076451221504	5/3/2013 3:45:01 AM	UNITED STATES	1
8	330166076958703616	5/3/2013 3:45:01 AM		1
9	330166076992266242	5/3/2013 3:45:01 AM	INDONESIA	1
10	330166076946128898	5/3/2013 3:45:01 AM	THAILAND	1
11	330166078980378625	5/3/2013 3:45:02 AM	UNITED STATES	1
12	330166080251252737	5/3/2013 3:45:02 AM	UNITED STATES	0
13	330166081203355648	5/3/2013 3:45:02 AM	UNITED STATES	1
14	330166081727639552	5/3/2013 3:45:02 AM	UNITED STATES	2
15	330166082679742467	5/3/2013 3:45:02 AM	UNITED STATES	1
16	330166082499399681	5/3/2013 3:45:02 AM	UNITED STATES	0
17	330166082713317376	5/3/2013 3:45:02 AM	ECUADOR	2
18	330166084676227076	5/3/2013 3:45:03 AM	UNITED STATES	2
19	330166086328782849	5/3/2013 3:45:03 AM	INDONESIA	1
20	330166085636731000	5/3/2013 3:45:03 AM	UNITED STATES	1
21	330166086794354688	5/3/2013 3:45:03 AM	UNITED STATES	1
22	330166087201206274	5/3/2013 3:45:04 AM		1
23	330166087532552193	5/3/2013 3:45:04 AM	UNITED STATES	2
24	330166087884890113	5/3/2013 3:45:04 AM		2
25	330166088073637888	5/3/2013 3:45:04 AM	UNITED STATES	0

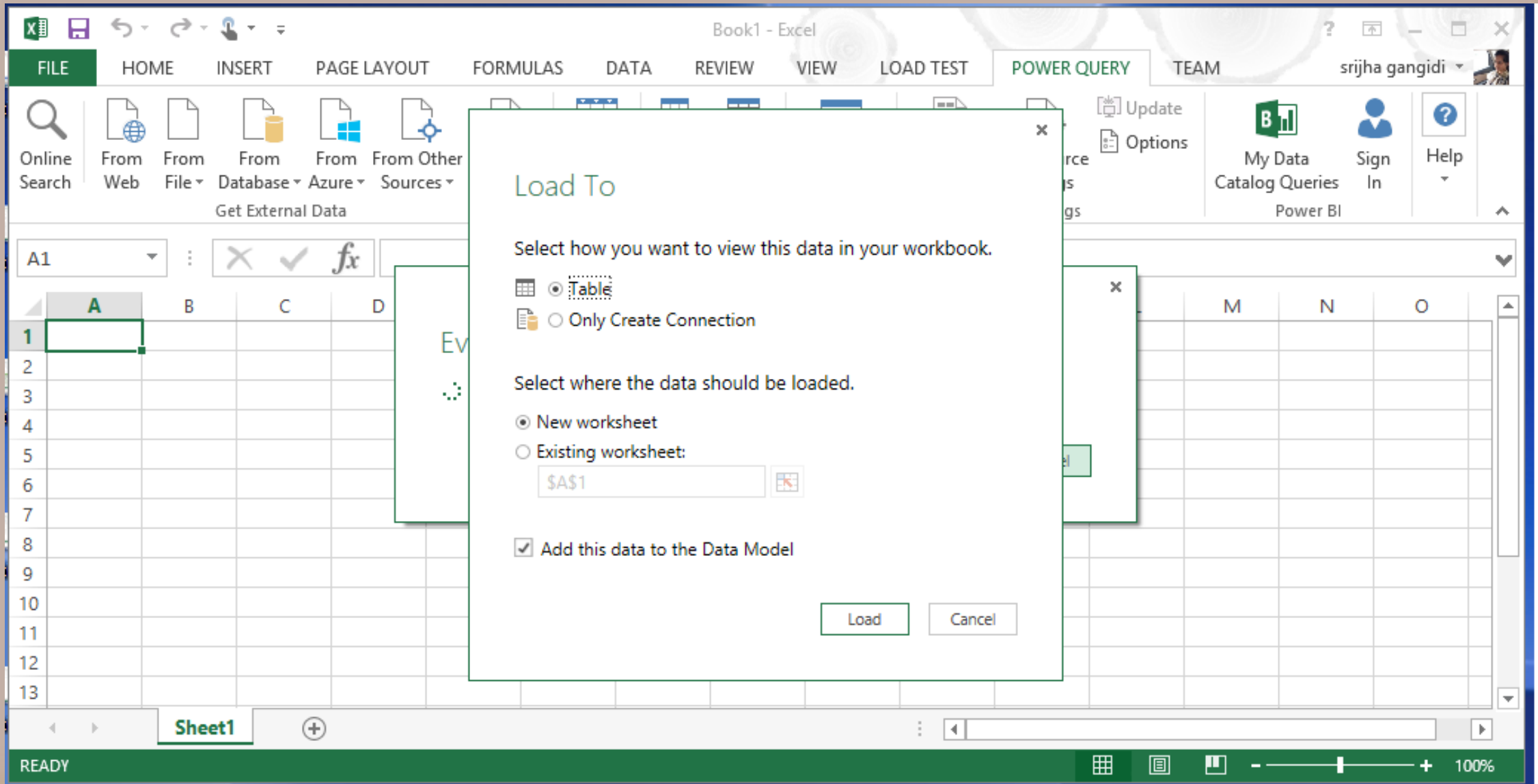
4 COLUMNS, 50 ROWS

PREVIEW DOWNLOADED AT 6:04 PM.

6:04 PM 3/10/2015

Power Query via CSV file

(Cont...)



Power Query via CSV file

(Cont...)

The screenshot shows the Microsoft Excel interface with the 'POWER QUERY' ribbon selected. The 'Power View' button is highlighted with a red box. A tooltip for 'Power View' is visible, stating: 'Insert a Power View Report. Make better business decisions and create beautiful, interactive reports.'

The data table in the worksheet is as follows:

	A	B	C	D	E	F	G	H	I
1	id	ts	country	sentiment					
2	3.30085E+17	5/2/2013 22:22	UNITED STATES	1					
3	3.30078E+17	5/2/2013 21:56	RUSSIAN FEDERATION	0					
4	3.30065E+17	5/2/2013 21:01	UNITED STATES	2					
5	3.30169E+17	5/3/2013 3:56	ECUADOR	2					
6	3.30142E+17	5/3/2013 2:07	CANADA	2					
7	3.30136E+17	5/3/2013 1:45	UNITED STATES	1					
8	3.30157E+17	5/3/2013 3:09		1					
9	3.30088E+17	5/2/2013 22:33	UNITED STATES	2					
10	3.30059E+17	5/2/2013 20:40	MOROCCO	0					
11	3.3016E+17	5/3/2013 3:20	UNITED STATES	1					
12	3.30086E+17	5/2/2013 22:28		0					
13	3.3014E+17	5/3/2013 1:59		2					
14	3.30062E+17	5/2/2013 20:51	NETHERLANDS	2					
15	3.30047E+17	5/2/2013 19:51	UNITED STATES	1					
16	3.30153E+17	5/3/2013 2:52	ECUADOR	1					
17	3.30154E+17	5/3/2013 2:56	UNITED STATES	0					
18	3.30071E+17	5/2/2013 21:26		1					
19	3.30047E+17	5/2/2013 19:52	GREECE	2					
20	3.30054E+17	5/2/2013 20:21	UNITED STATES	0					
21	3.3013E+17	5/3/2013 1:22	UNITED STATES	2					
22	3.30089E+17	5/2/2013 22:38	UNITED STATES	2					
23	3.30068E+17	5/2/2013 21:16	BRAZIL	2					
24	3.3016E+17	5/3/2013 3:20	UNITED STATES	1					

The status bar at the bottom shows: AVERAGE: 1.10041E+17 COUNT: 359376 SUM: 2.96593E+22. The taskbar at the bottom shows the Windows logo, Internet Explorer, File Explorer, Amazon, Google Chrome, Paint, Word, and Excel.

Power Query via CSV file

(Cont...)

Book1 - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW LOAD TEST POWER VIEW DESIGN POWER QUERY TEAM srijha gan...

Table Bar Column Other Map Tiles Tile Type Slicer Card Show Totals

Switch Visualization Tiles Slicer Options Number Text Arrange

Click here to add a title

id	ts	country	sentiment
220,042,882,728,225,000.00	5/2/2012		1
220,042,888,788,197,000.00	5/2/2012		1
220,042,887,055,928,000.00	5/2/2012	CHILE	1
220,042,889,529,289,000.00	5/2/2012	MOROCCO	2
220,042,888,577,799,000.00	5/2/2012	UNITED STATES	2
220,042,890,640,749,000.00	5/2/2012	UNITED STATES	1
660,087,794,824,605,000.00	5/2/2012		2
220,042,897,518,522,000.00	5/2/2012	UNITED STATES	1
220,042,899,097,784,000.00	5/2/2012	NETHERLANDS	0
220,042,900,410,601,000.00	5/2/2012	THAILAND	1
220,042,901,970,874,000.00	5/2/2012	UNITED STATES	2
220,042,902,872,658,000.00	5/2/2012		1
220,042,908,072,924,000.00	5/2/2012	UNITED STATES	2
220,042,907,206,054,000.00	5/2/2012	ARGENTINA	1
220,042,907,192,812,000.00	5/2/2012	INDONESIA	1
220,042,908,188,458,000.00	5/2/2012	THAILAND	1

Filters
VIEW TABLE

To filter the View, drag fields from the field list.

Power View Fields

ACTIVE | ALL

Query1

- ☒ country
- ☒ Σ id
- ☒ Σ sentiment
- ☒ ts

Drag fields between areas below:

TILE BY

FIELDS

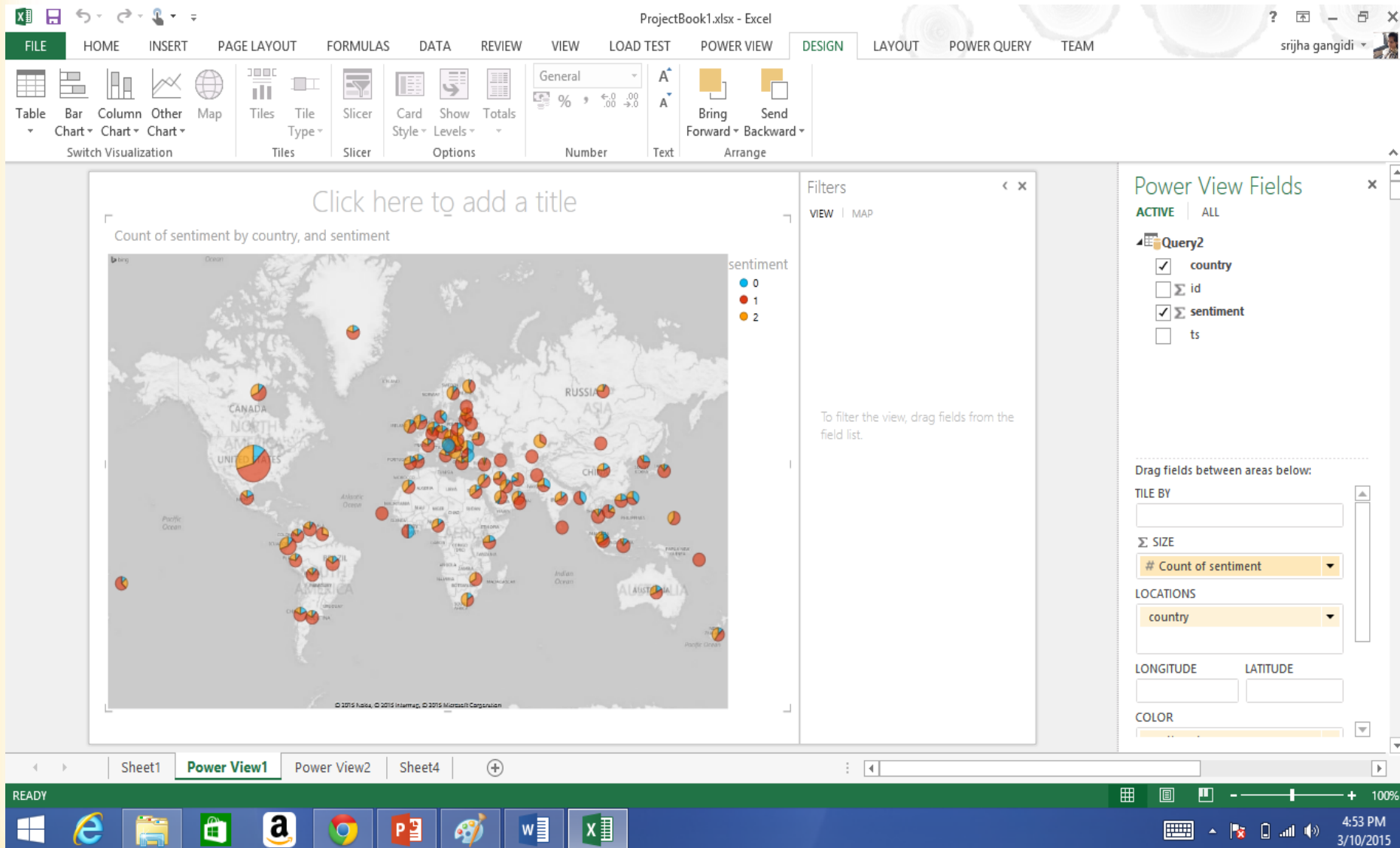
- Σ id
- ts
- country

Sheet1 Sheet2 Power View1

READY

100%

Map Display of Sentiment Data



Orange : Positive
Blue : Negative
Red : Neutral

Challenges and Overcomes

- Encountered issues while installing Hive and Hadoop Separately
 - Switched to HortonWorks Sandbox with preinstalled Hadoop and Hive as per [atlink](#).
- System got slow and got stuck upon installation of Hortonworks
 - Re-Divided Ram allocation equally between Windows and HDP
- Importing JSON file
 - ---- Implemented usage of WinSCP - A file transfer software to remote machine
- Hive & MapReduce jobs not configured
 - ---- Switched to Stable HDP 2.0 from HDP 2.2 with pre-configured Hive and MapReduce
- Currently facing the problem of ODBC Driver Configuration with Hortonworks

Sentiment Analysis using Hadoop

Sponsored By Atlink Communications Inc

Team Members : Ankur Uprit, Pinaki Ranjan Ghosh, Kiranmayi Ganti, Srijha Reddy Gangidi



Capstone Project Group 1

Instructor : Dr.Sadegh Davari

Mentors : Dilhar De Silva ,
Rishita Khalathkar

