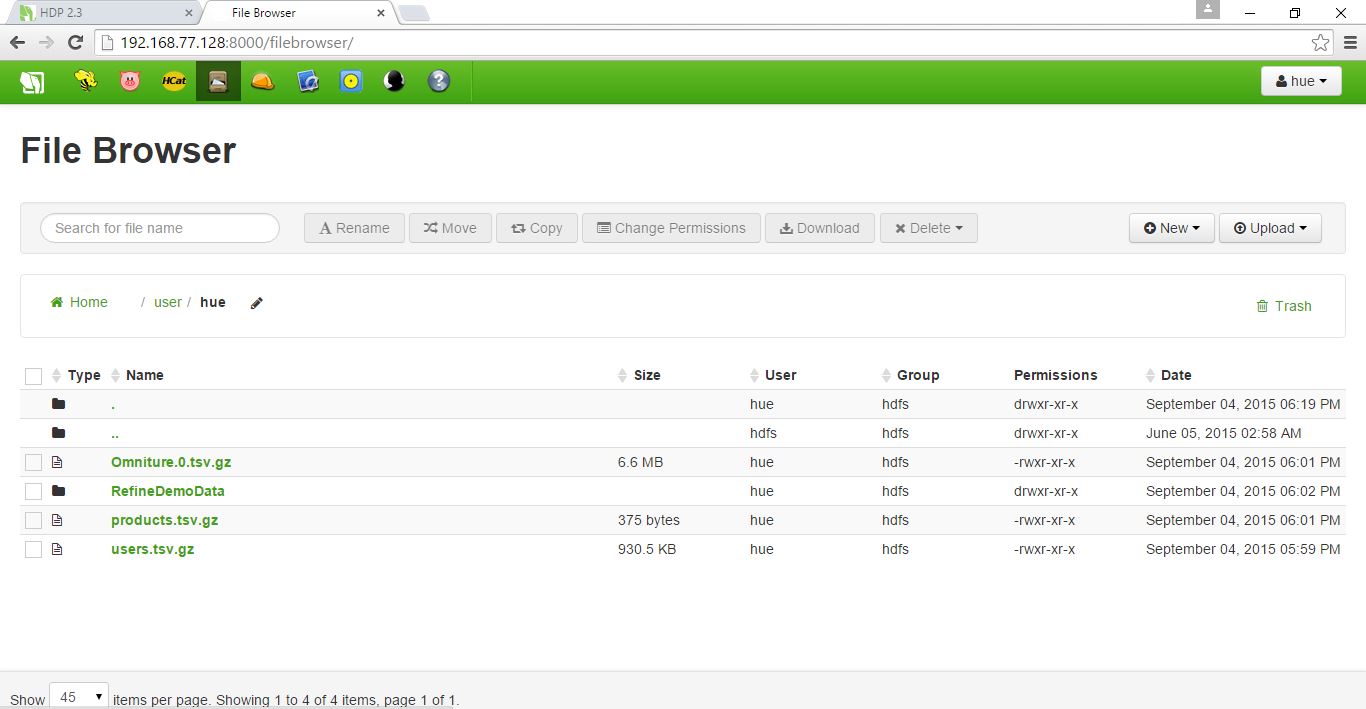
Visualizing website clickstream data

Firstly we need some clickstream data to kick start our project. Thanks to GitHub! Here’s a summary of the data we’re working with:

* Omniture logs\* – website log files containing information such as URL, timestamp, IP address, geocoded IP address, and user ID (SWID).
* users\* – CRM user data listing SWIDs (Software User IDs) along with date of birth and gender.
* products\* – CMS data that maps product categories to website URLs.

Next, let’s open our Hortonwork Sandbox. OMG! It is so damn slow..

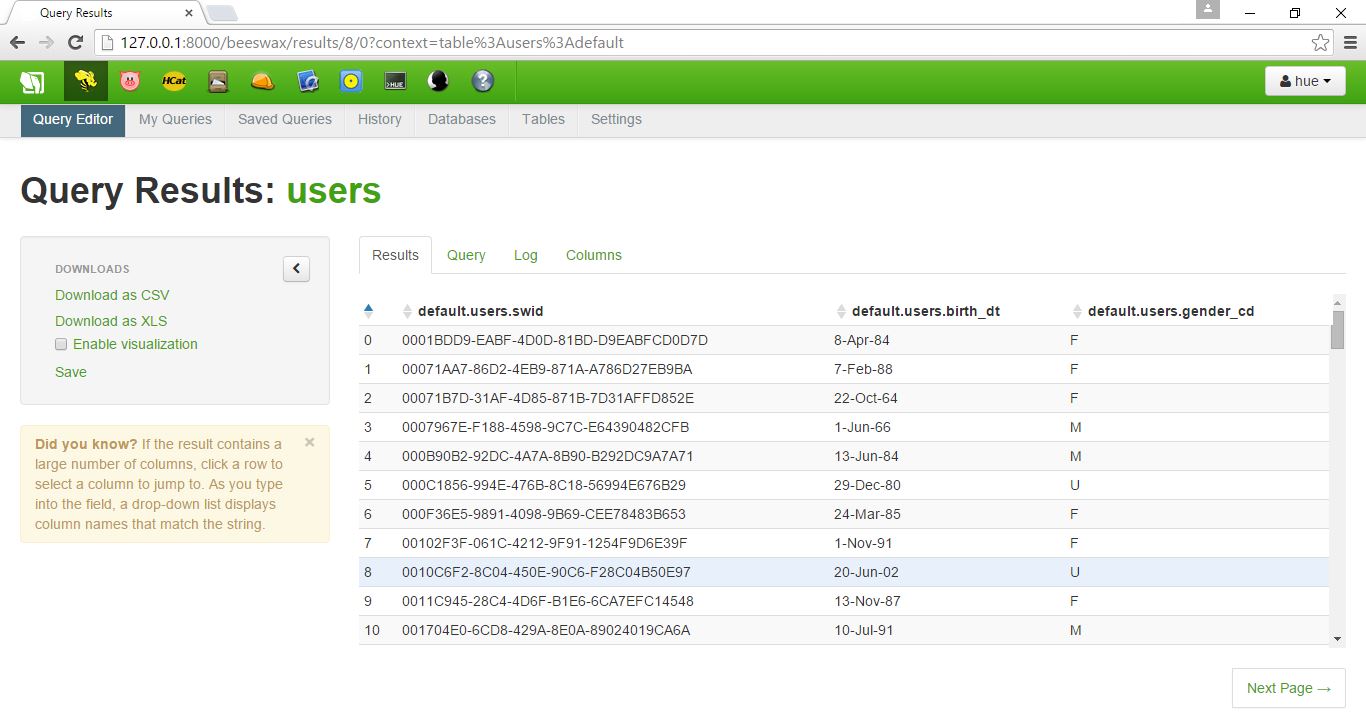
Once it opens, login and load the clickstream data to the HDFS. Now you can see your files in the File Browser:



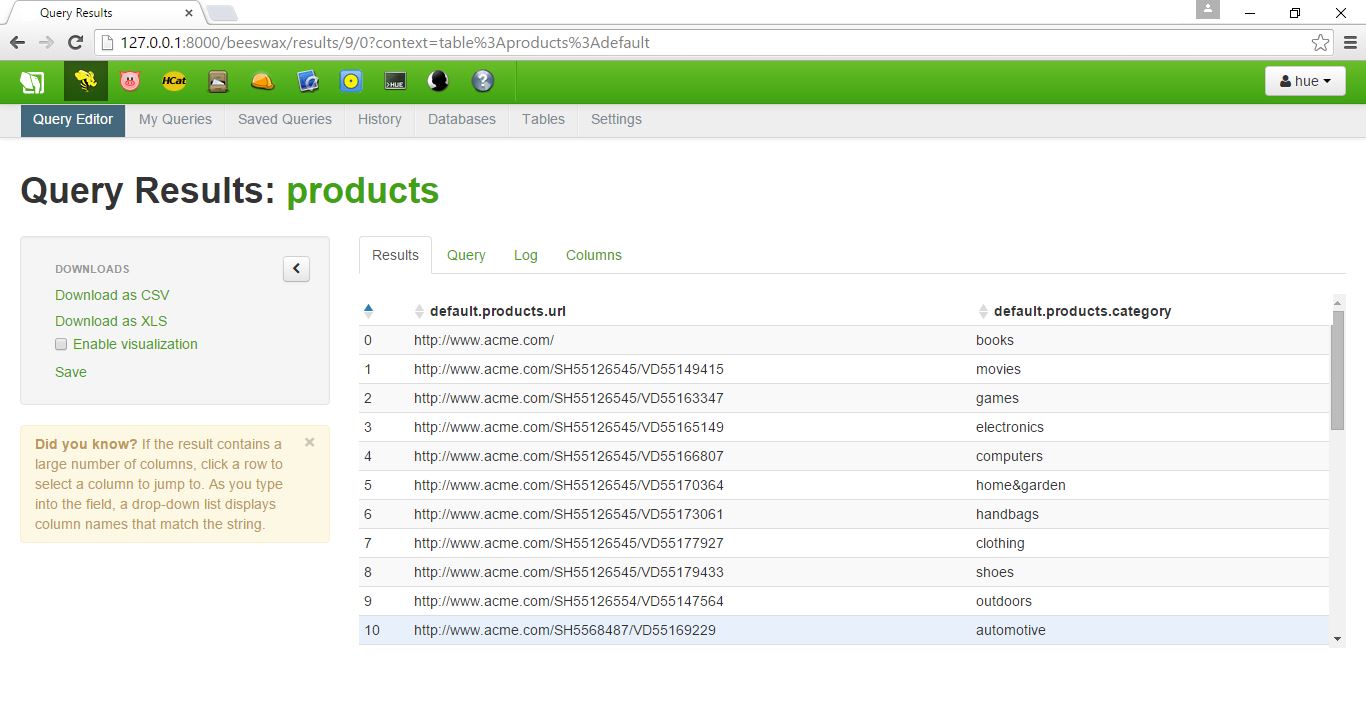
The Omniture logs\* file consists of 5 days of clickstream data with almost 4 million rows!! That’s huge!!

Click the HCat icon in the toolbar at the top of the page. We don’t have any tables created prior. So, let’s create the tables: **omniturelogs**, **users** and **products**. You can click **Browse data** option to see the data.

The users table consists of the SWID, birth date, and gender columns:

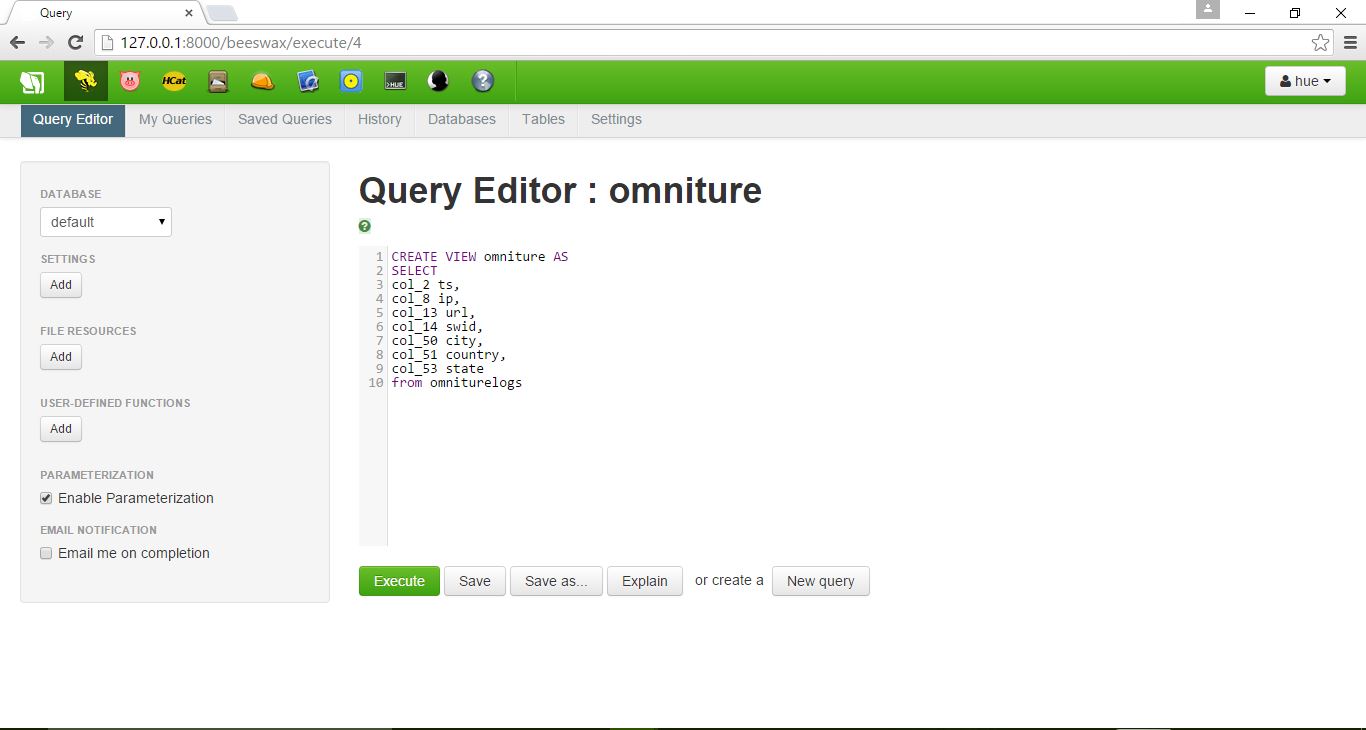


The data in the products table maps product categories to website URLs:



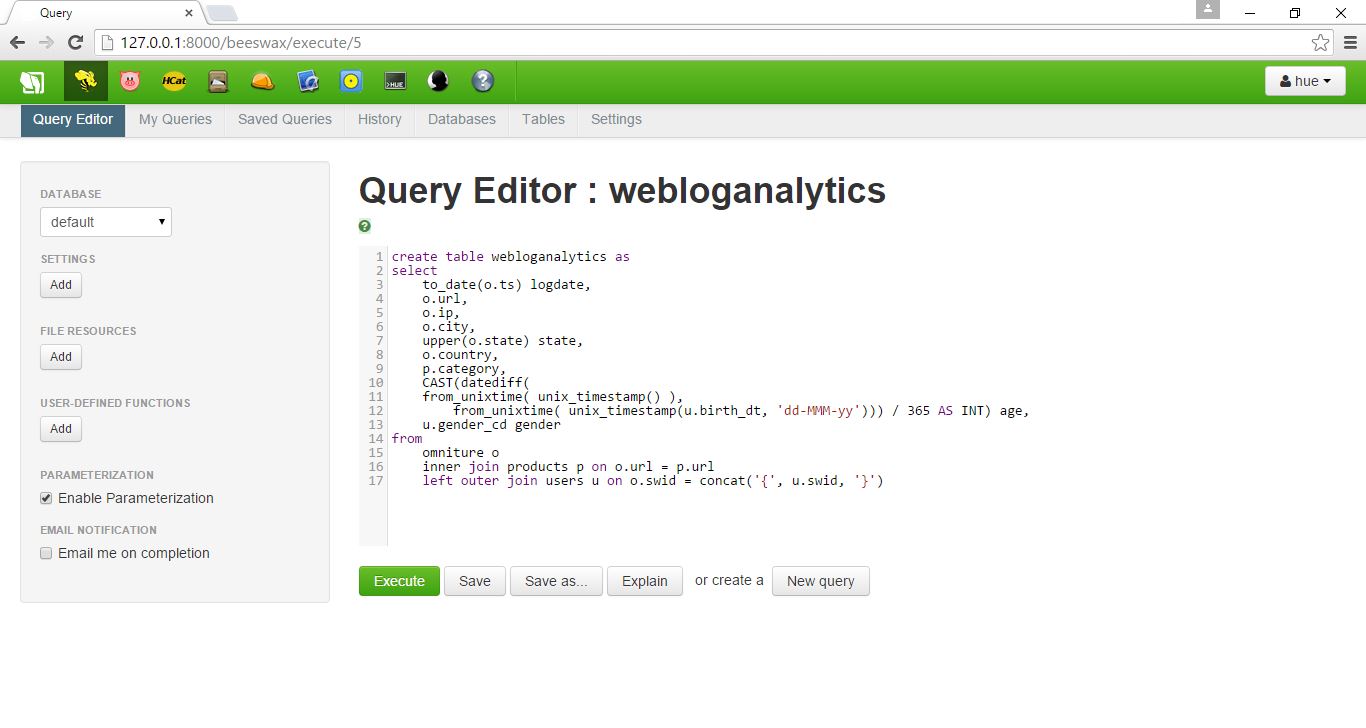
Now that we have all the data needed, we need to sort the data and pull up only the required fields. We need to write a bunch of **hive** queries to get this done:

First, we used a Hive script to generate an “omniture” view that contained a subset of the data in the omniturelogs table:

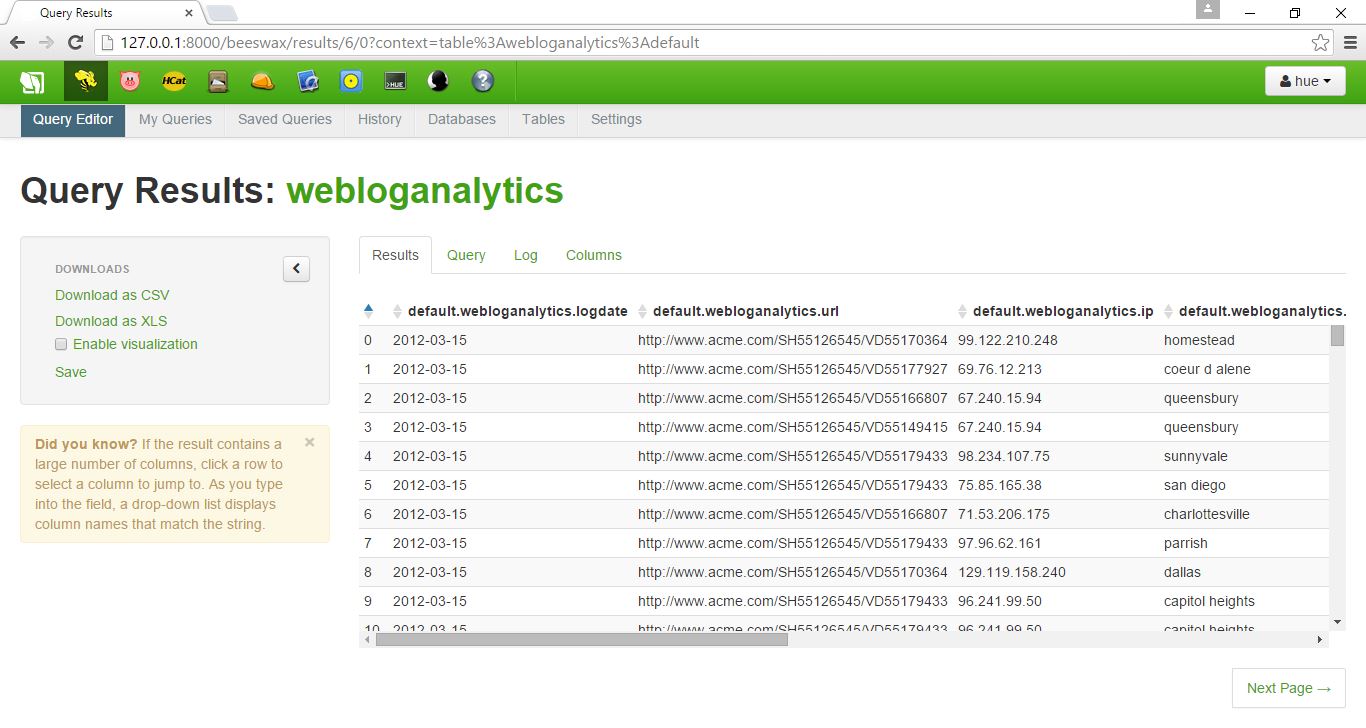


To view the omniture data in Hive, click **Table**, then click **Browse Data.**

Moving on, we created a “webloganalytics” script to join the omniture website log data to the CRM data (registered users) and CMS data (products). This Hive query executed a join to create a unified dataset across our data sources:

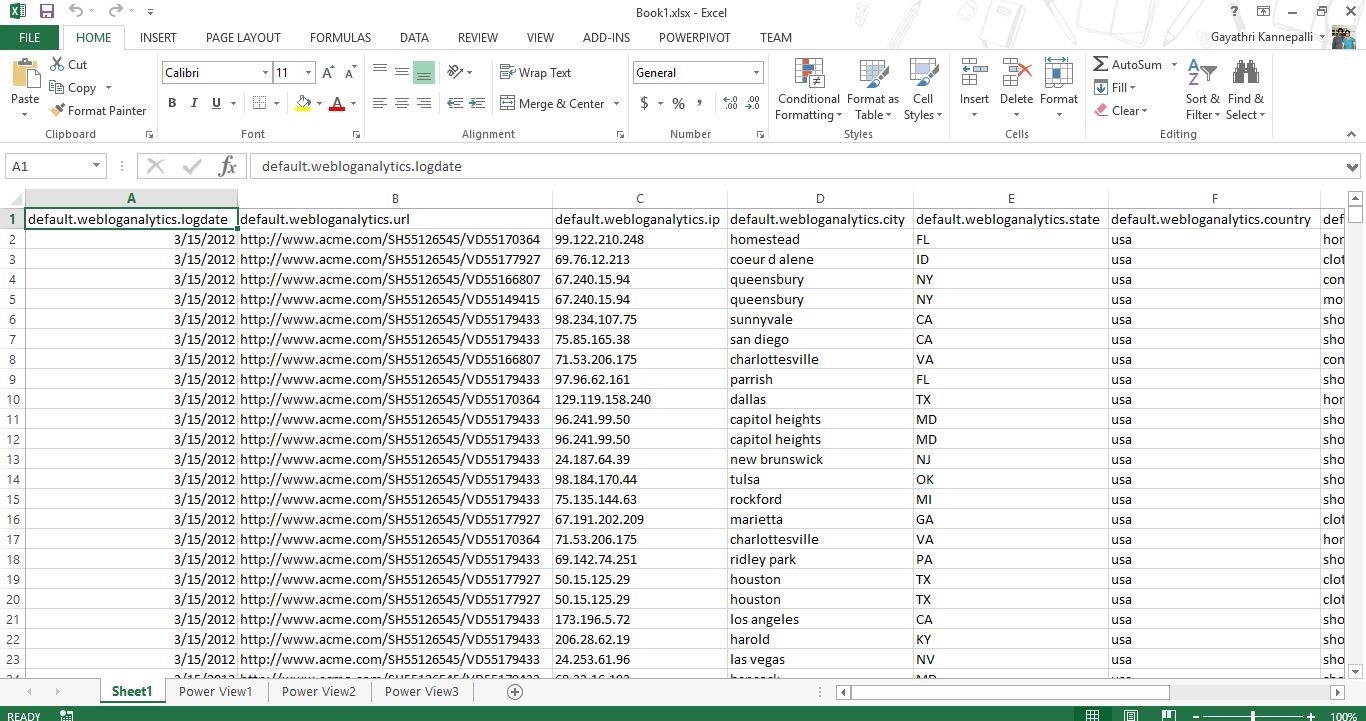


You can browse the data once the query is executed and the data is loaded into the table:



Now is the time to visualize this stuff on our Excel spreadsheet to view exciting results. I have tried for more than hour to install and connect the ODBC connector to the Excel so that we can directly import the data using the option – From Microsoft Query. My Bad! I couldn’t do it. So, I’ve simply downloaded the CSV format file of the webloganalytics table. [You can see the option available in the picture above.]

Now we can import the data by opening Excel workbook, selecting **DATA** option in the menu and select **FROM TEXT** option. It prompts you to browse for a file. Select the downloaded .csv file. The data gets dumped into the Excel sheet:



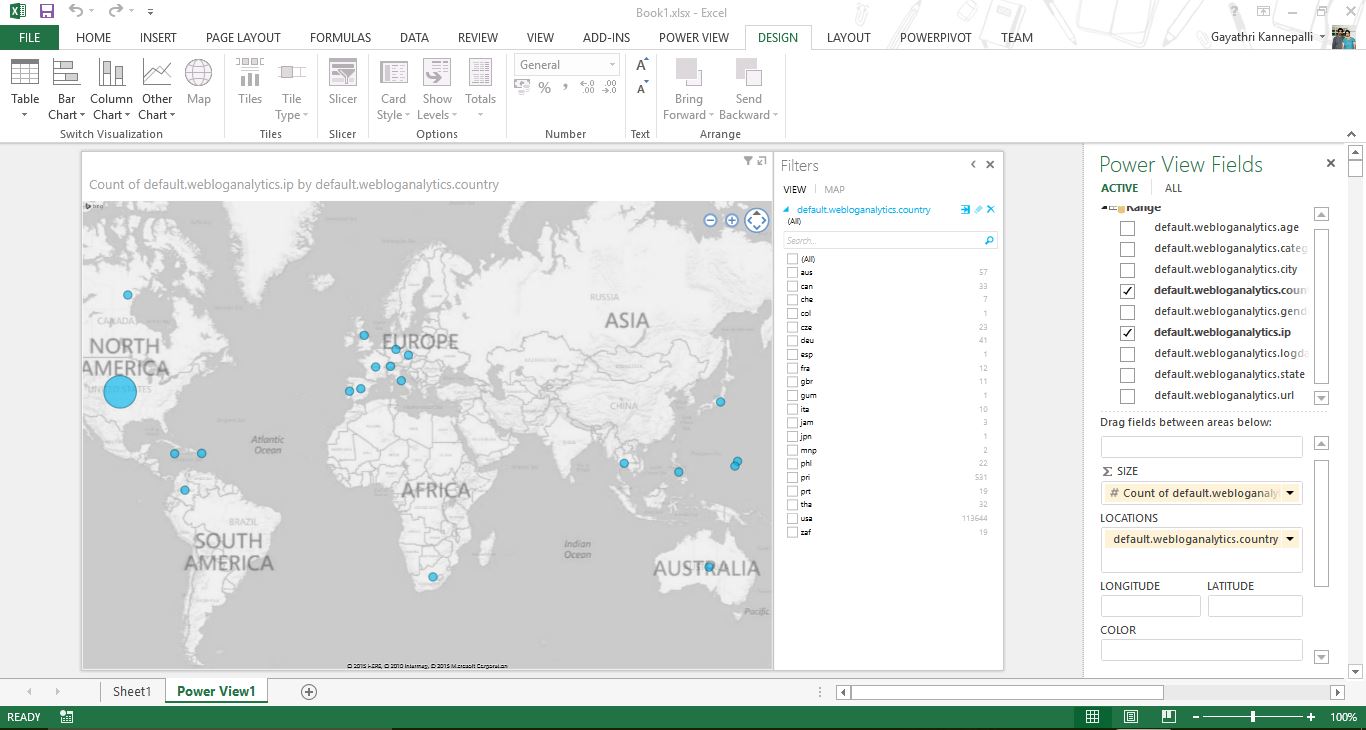
Data visualization can help you optimize your website and convert more visits into sales and revenue.

Here, we will:

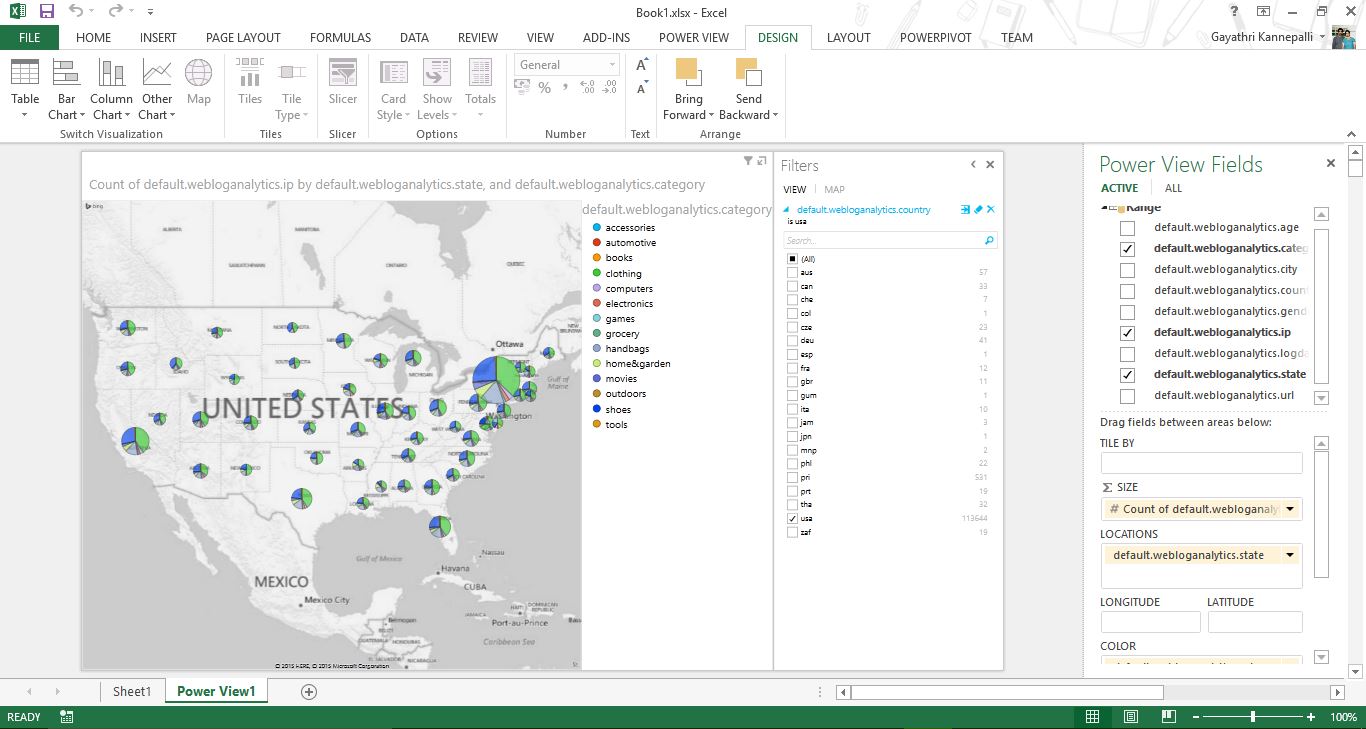
* Analyze the clickstream data by location
* Filter the data by product category
* Graph the website user data by age and gender
* Pick a target customer segment
* Identify a few web pages with the highest bounce rates

In the Excel workbook with the imported webloganalytics data, select Insert > Power View to open a new Power View report.

* Let’s start by taking a look at the countries of origin of our website visitors.
* In the Power View Fields area, leave the **country** checkbox selected, and clear all of the other checkboxes. The data table will update to reflect the selections.
* On the Design tab in the top menu, click **Map**. The map view displays a global view of the data.



* Now let’s take a look at a count of IP address by state.
* First, drag the **ip** field into the SIZE box.
* Drag **country** from the Power View Fields area into the Filters area, then select the **usa** checkbox.
* Next, drag **state** into the LOCATIONS box.
* Remove the **country** field from the LOCATIONS box by clicking the down-arrow and then **Remove Field**.
* Use the map controls to zoom in on the United States. Move the pointer over each state to display the IP count for that state.
* Our dataset includes product data, so we can display the product categories viewed by website visitors in each state.
* To display product categories in the map by color, drag the **category** field into the COLOR box. The map displays the product categories by color for each state.
* Move the pointer over each state to display detailed category information. We can see that the largest number of page hits were for clothing, followed by shoes.



Now let’s look at the movies data by age and gender so we can optimize our content for these customers. Select **Insert > Power View** to open a new Power View report.

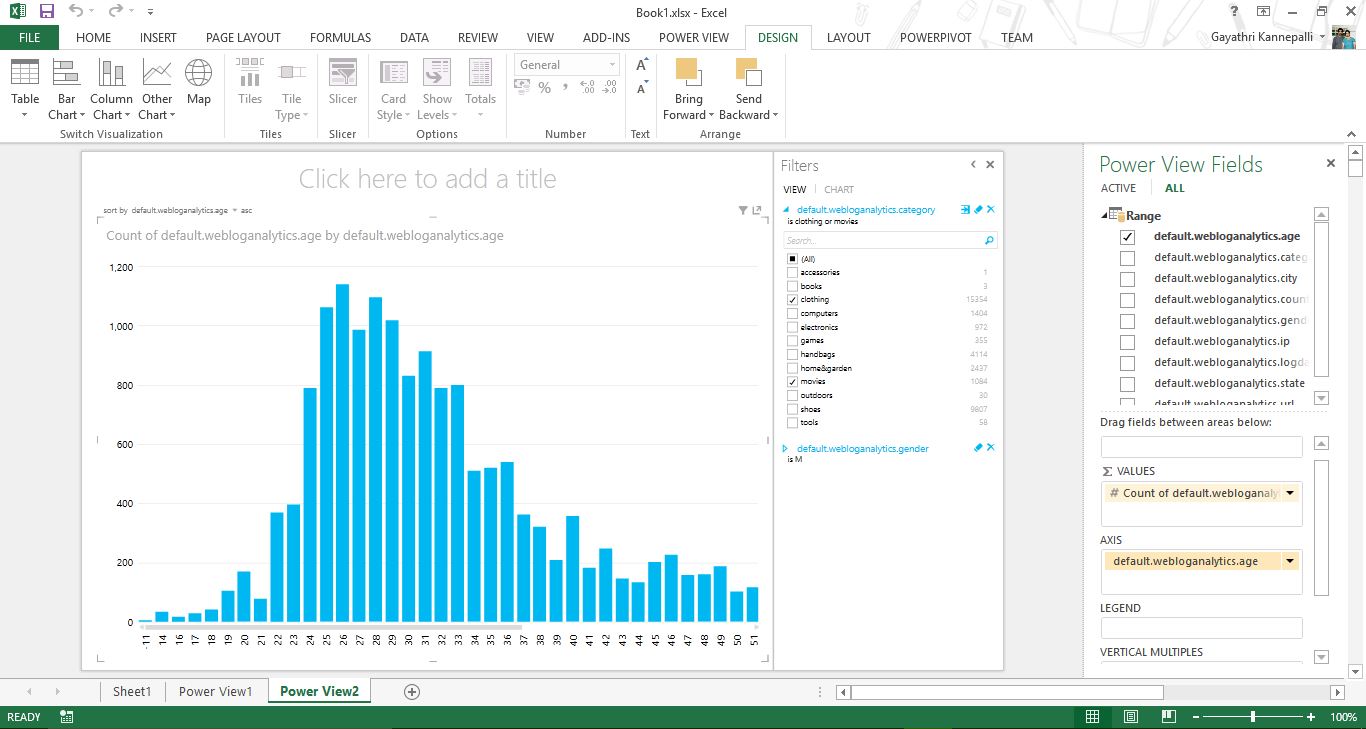
To set up the data, set the following fields and filters:

* In the Power View Fields area, select ip and age. All of the other fields should be unselected.
* Drag category from the Power View Fields area into the Filters area, then select the movies checkbox.
* Drag gender from the Power View Fields area into the Filters area, then select the M (male) checkbox.

After setting these fields and filters, select Column Chart > Clustered Column in the top menu.

To finish setting up the chart, drag **age** into the AXIS box.

Also, remove **ip** from the AXIS box by clicking the down-arrow and then **Remove Field**. The chart shows that the majority of men shopping for movies on our website are between the ages of 22 and 30. With this information, we can optimize our content for this market segment.



Let’s assume that our data includes information about website pages (URLs) with high bounce rates.

A page is considered to have a high bounce rate if it is the last page a user visited before leaving the website.

By filtering this URL data by our target age group, we can find out exactly which website pages we should optimize for this market segment.

Select **Insert > Power View** to open a new Power View report.

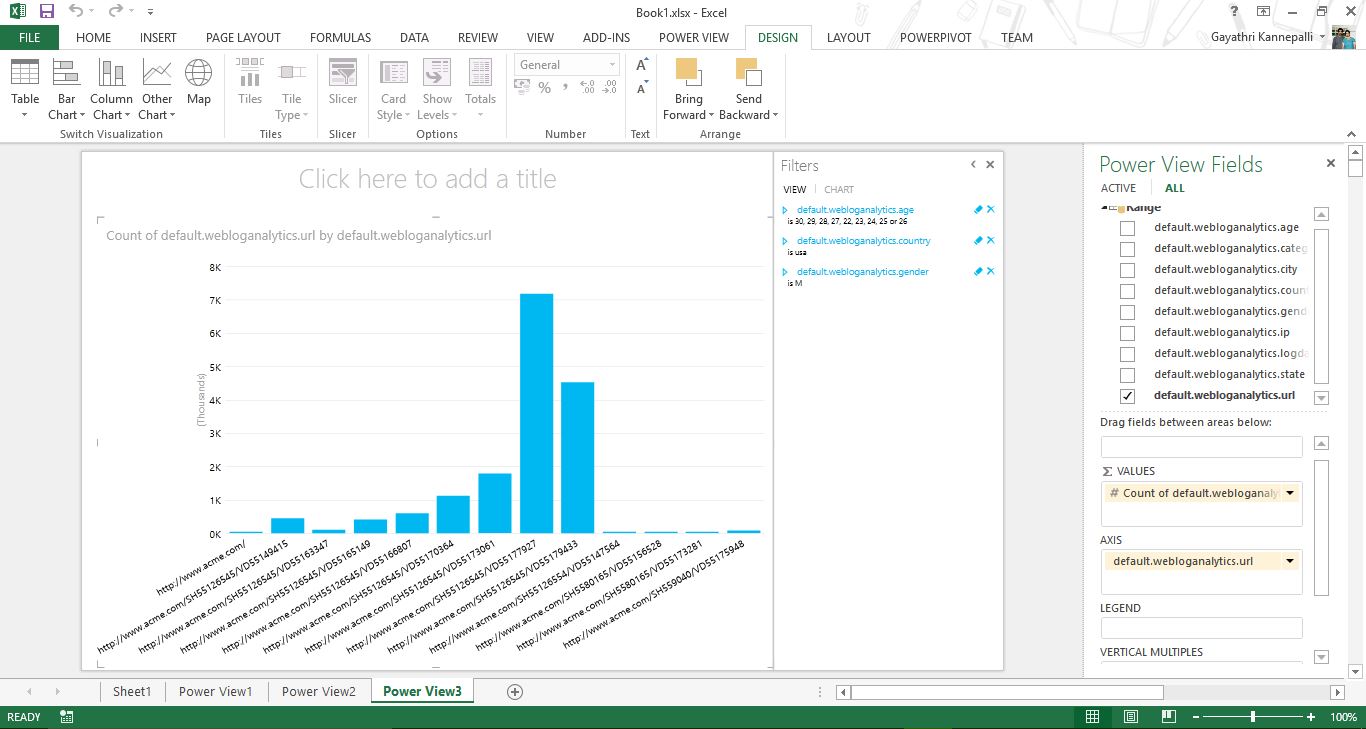
To set up the data, set the following fields and filters:

* Drag age from the Power View Fields area into the Filters area, then drag the sliders to set the age range from 22 to 30.
* Drag gender from the Power View Fields area into the Filters area, then select the M (male) checkbox.
* Drag country from the Power View Fields area into the Filters area, then select the usa checkbox.
* In the Power View Fields area, select url. All of the other fields should be unselected.
* In the Power View Fields area, move the pointer over url, click the down-arrow, and then select Add to Table as Count.

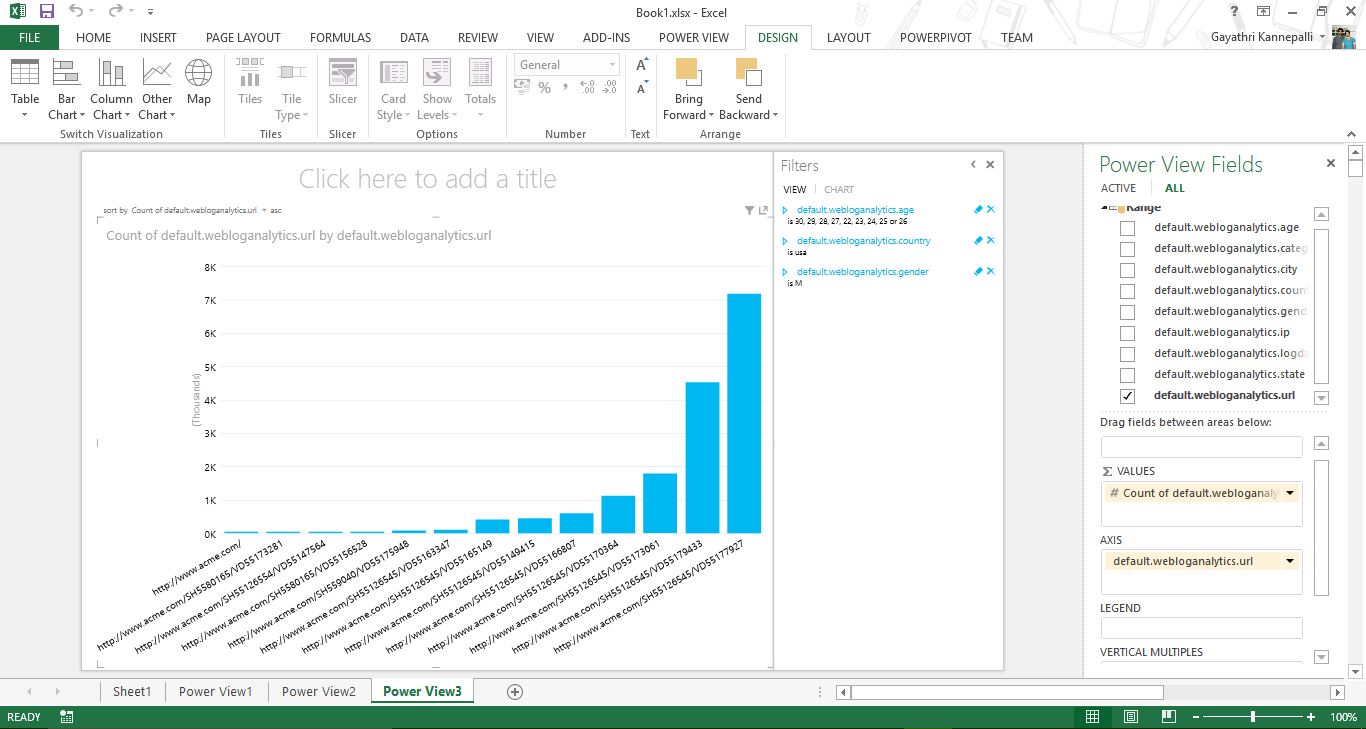
After setting these fields and filters, select **Column Chart > Clustered Column**in the top menu.

The chart shows that we should focus on optimizing four of our website pages for the market segment of men between the ages of 22 and 30.

Now we can redesign these four pages and test the new designs based on our target demographic, thereby reducing the bounce rate and increasing customer retention and sales.



Note: You can use the controls in the upper left corner of the map to sort by Count of URL in ascending or descending order.



That’s it!! We can still play around with the data, which can be used for basket analysis, A/B testing, personalized product recommendations, and other sales optimization activities. (Basket analysis and A/B testing are the terms that I’ve heard from you! I felt, we could apply the same here.)