**Lab Tutorial**

Created By:

Spriha Batar ([sbatar@calstatela.edu](mailto:sbatar@calstatela.edu)) Hasitha Reddy ([hpittu@calstatela.edu](mailto:hpittu@calstatela.edu))

Premdeep Parasa ([pparasa@calstatela.edu](mailto:pparasa@calstatela.edu)) Sumani Pidikiti ([spidiki@calstatela.edu](mailto:spidiki@calstatela.edu))

06/10/2016

**Data Analysis using Databricks Cluster and SQL**

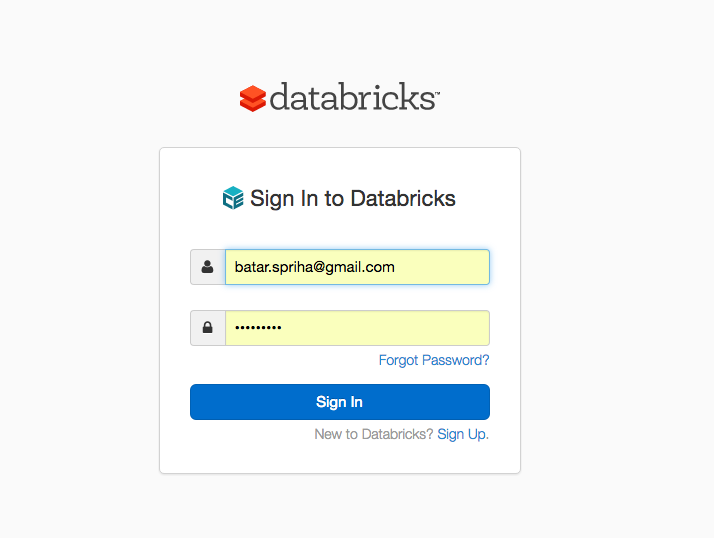
**Objectives**

In this hands-on lab, you will learn how to:

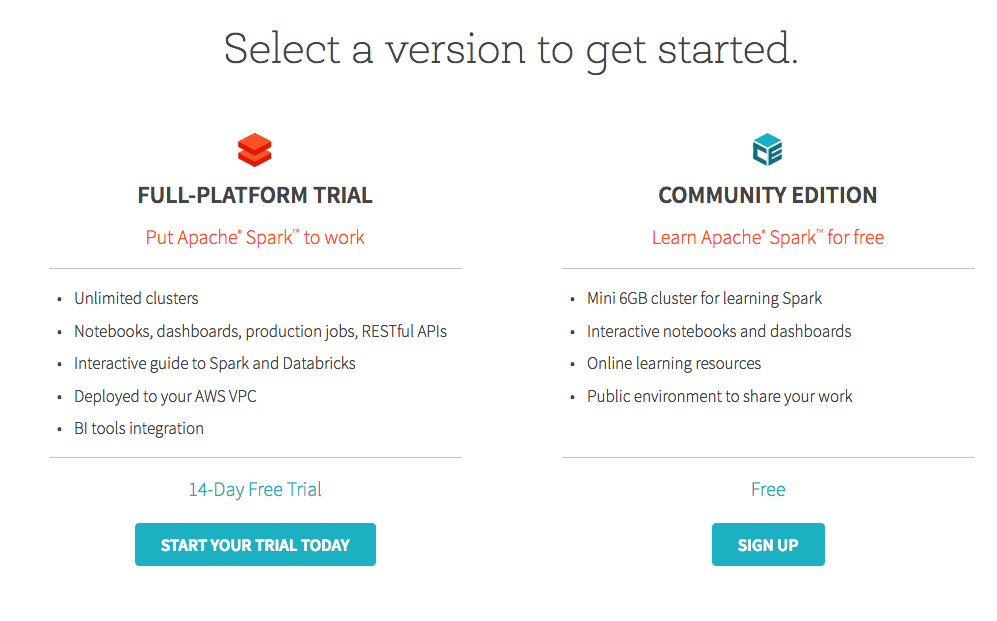
* Log in to Databricks account
* Create Databricks cluster
* Uploading data into tables and using it
* Use of SQL queries for Data Analysis
* Use of Microsoft Excel for Graphical Representation

**Exercise 1: Log in to Databricks Account**

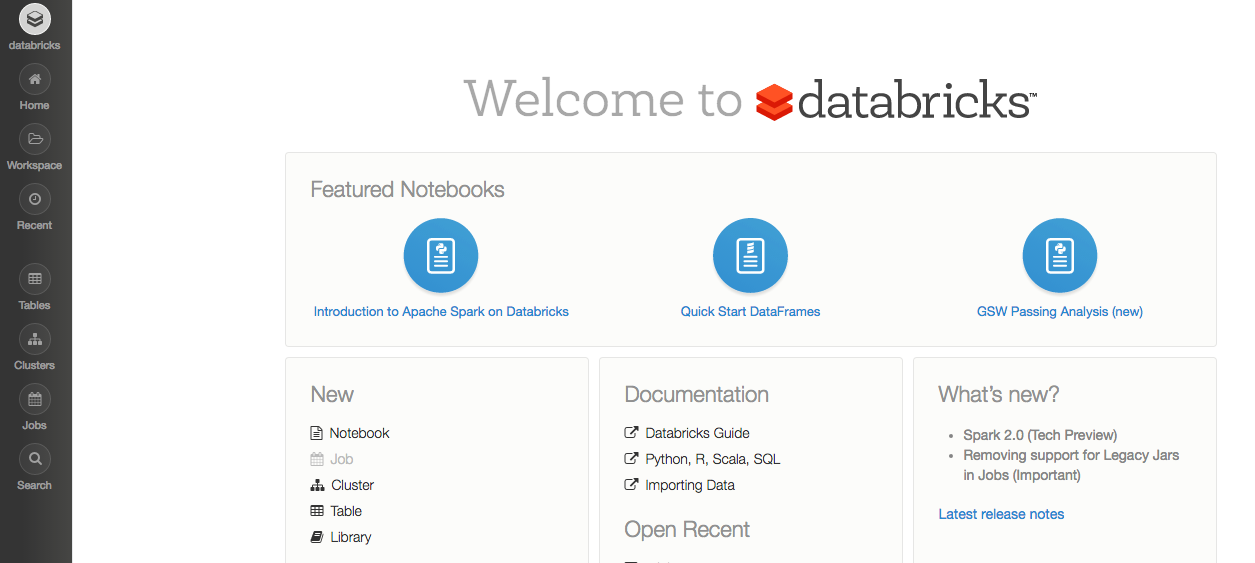
1. To open the Databricks website, click on the following link: <https://community.cloud.databricks.com>



2. Log in using your email id and password. If you don’t have an account, click Sign Up and create a new account. You can select 14-day free trail version.

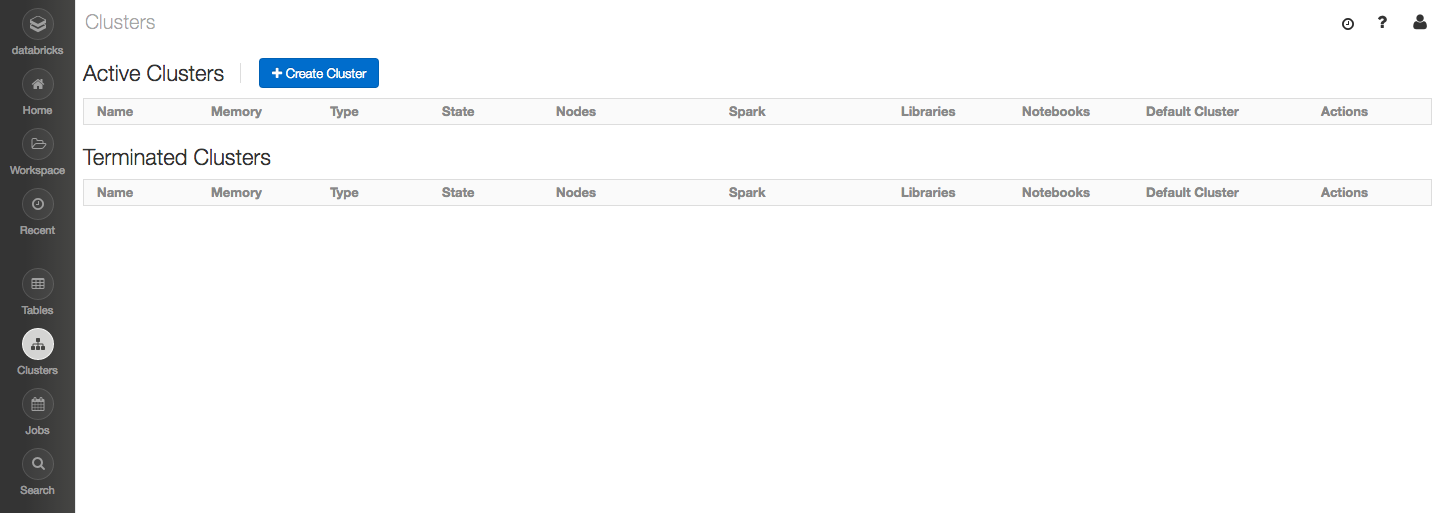


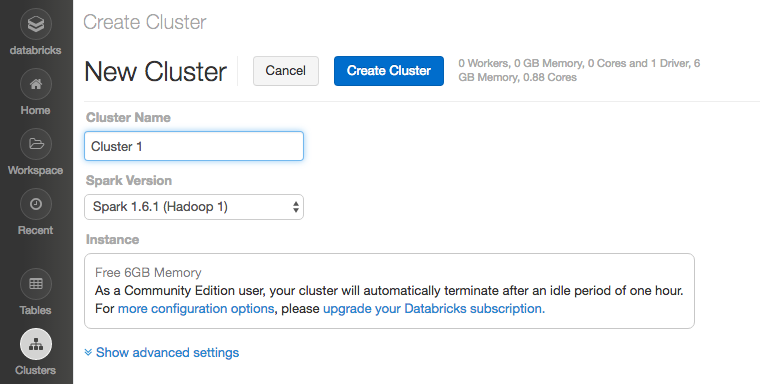
3. After signing in, you will see the Home page of Databricks.



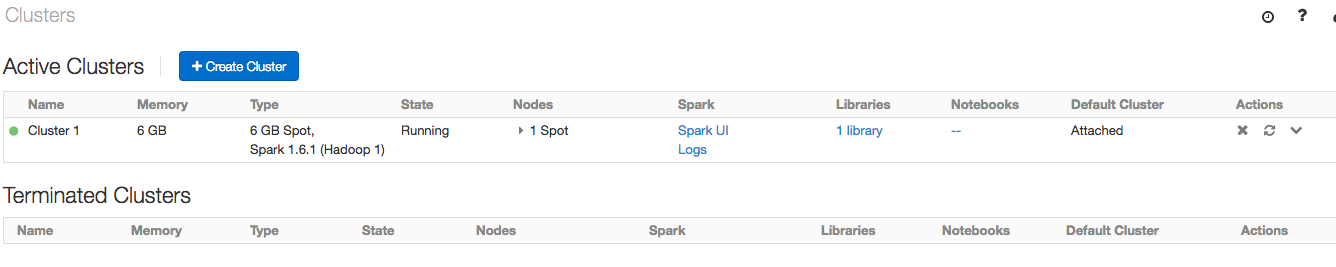
**Exercise 2: Creating a cluster**

1. Click on Cluster Icon present at the mid left corner of the Home Page. Click on New Cluster to create a new cluster. Give any name to the cluster you wish.





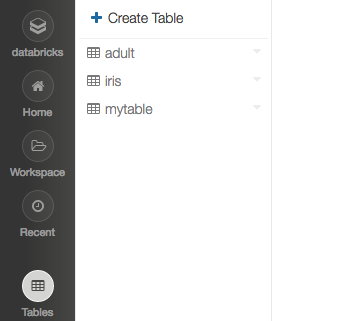
2. Once the cluster is created, you can view it under the cluster tab.



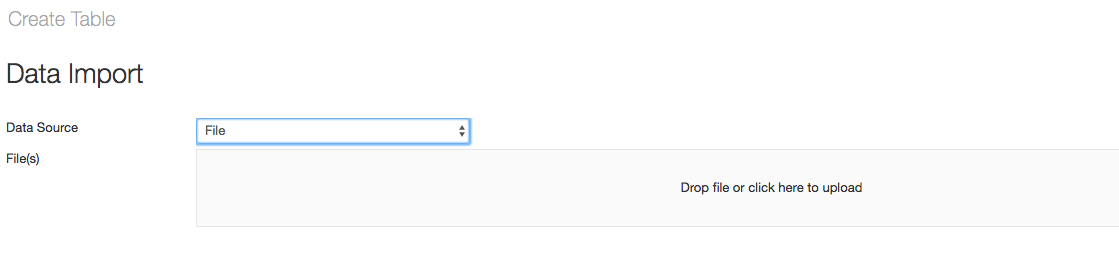
A Databricks cluster will automatically terminate after 1 hour.

**Exercise 3: Uploading data in tables**

1. Once a cluster is created, go back to home page and click on Tables icon. Then click on Create Table.



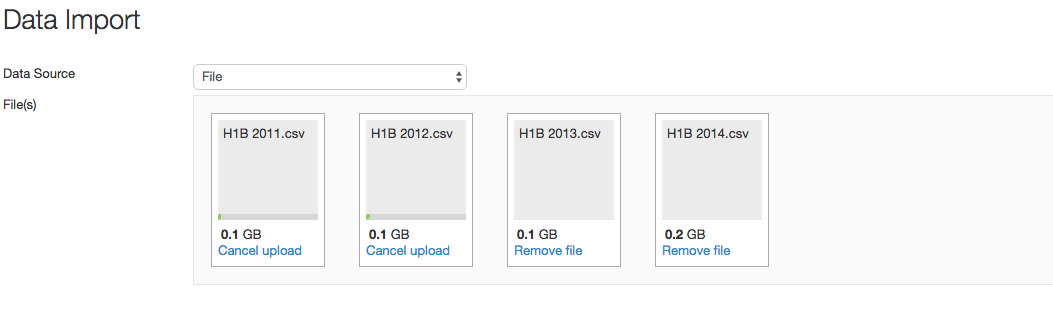
2. Drag and drop the excel files which you need to analyze. In this lab, we have analyzed H-1B Visa applications in four different files for four different years.



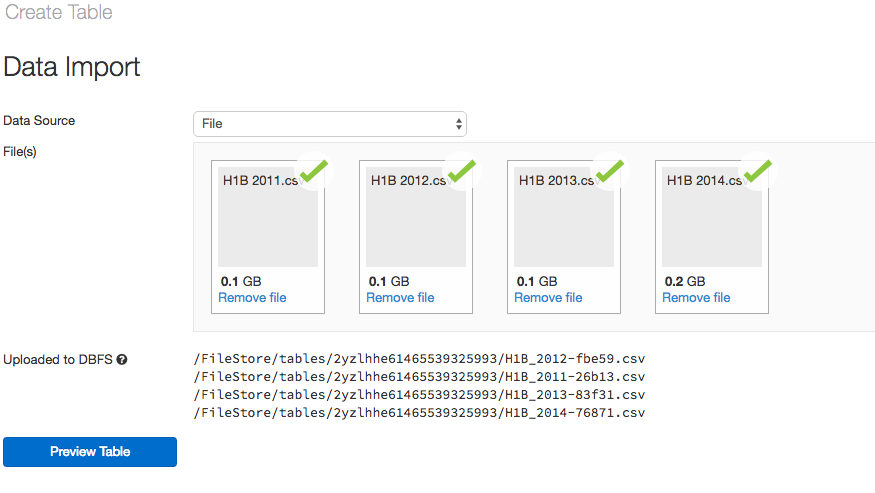
The data to be analyzed can be found at the following website:

<https://app.enigma.io/search/query/international%20student>

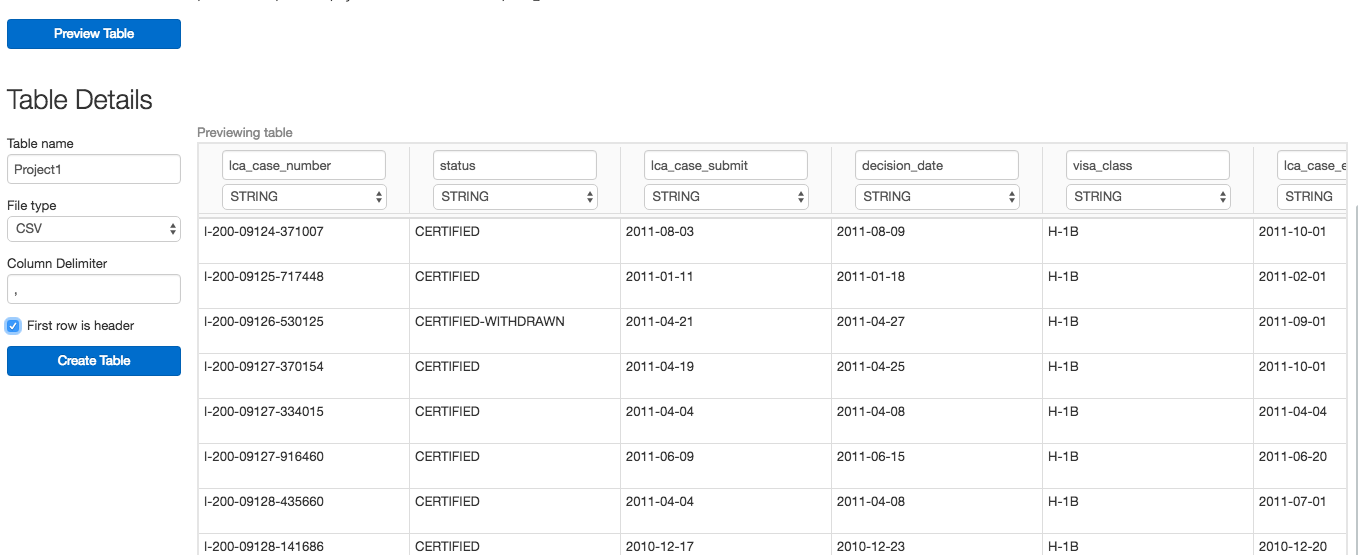
The data set that we used was around 600MB and in .csv format. To view or download the dataset, you must create an account on the website app.enigma.io



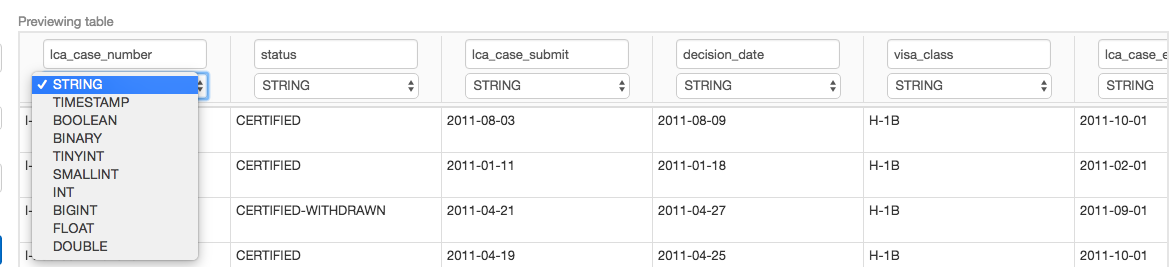
3. Once you have added all the files and the upload is complete, you will see a green tick mark on all the files uploaded. Also, the DBFS path for the data will be shown.



4. Click on Preview Table. Give any table to the table. The file type will be shown as .csv. Column Delimiter refers to the character used to distinguish between multiple columns. Also, check the icon for ‘First Row is header’. This will use the first row of the data as a header and not count it in queries.



You can also choose the data type of all column of the table here at this step.

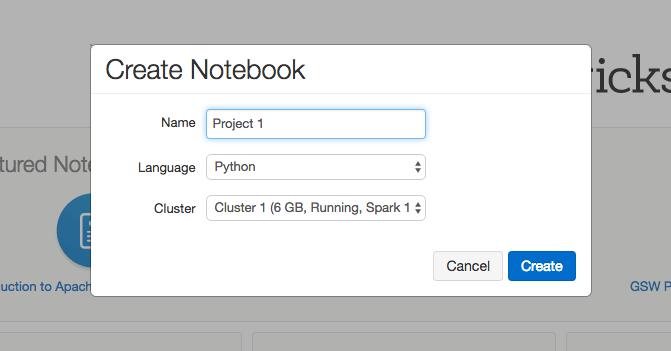


5. Click on Create Table. The whole schema of the table will be shown.

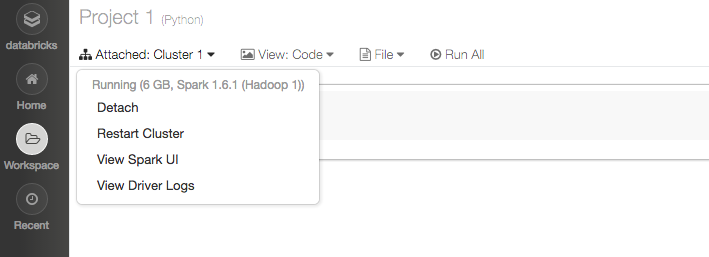


**Exercise 4: Using SQL queries to analyze data**

1. After data has been loaded into tables, go back to Databricks home page and click on new notebook. Give any name for the notebook and select the type. For this tutorial, we selected it as a python notebook.

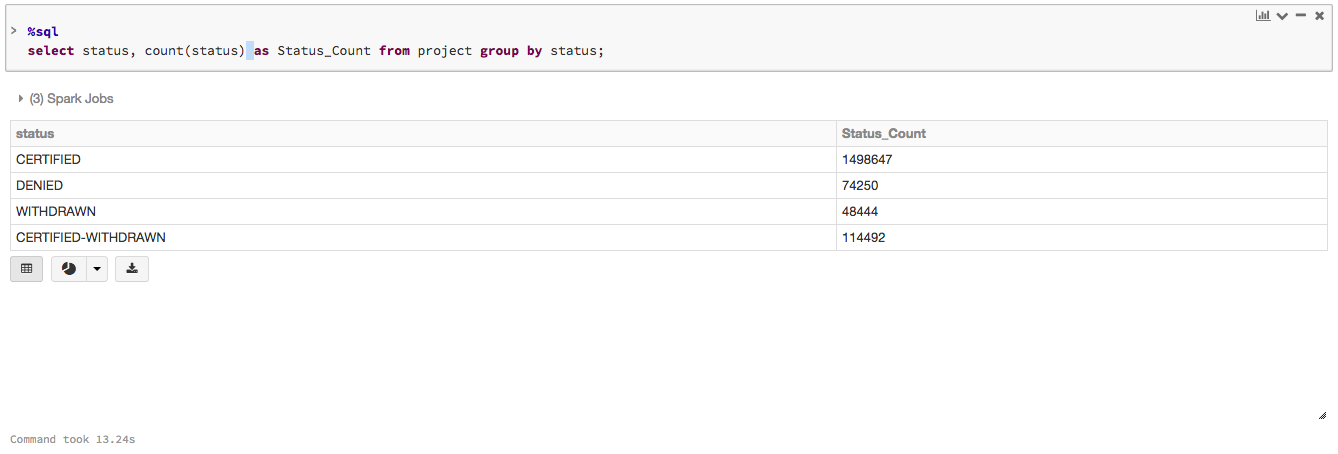


2. On the top left corner of the notebook, click on cluster icon and attach it to the above-created cluster.

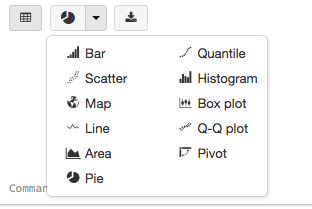


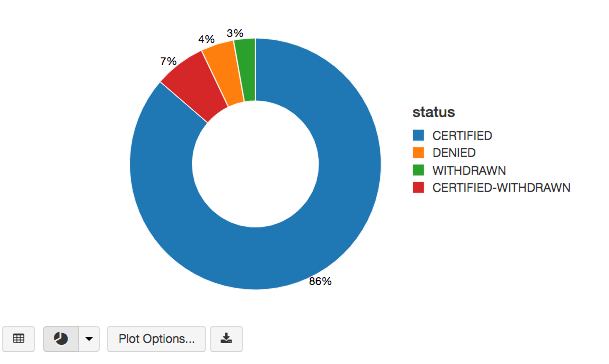
3. Now we are ready to start querying. To run SQL queries in a python notebook, write ‘%SQL’ before every query to be executed.

4. Press shift+enter to run the query. Once the query is executed, it will show the time taken by it to run and also the output of the query.

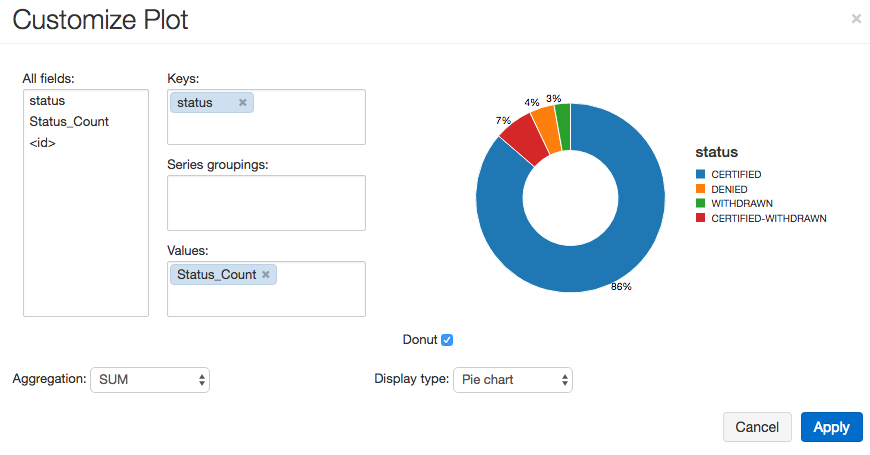


5. Click on the small graph icon to see different types of visualization available.



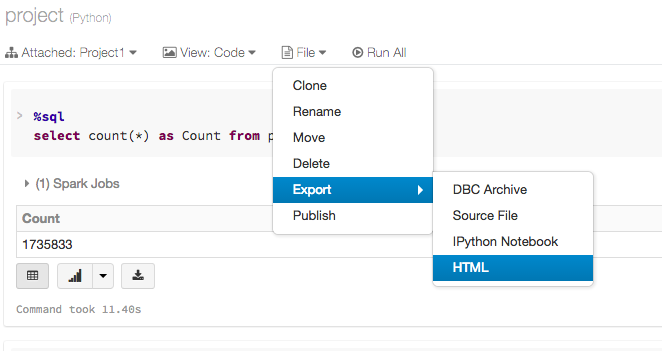


6. For this query, we chose a pie chart. You can click on plot options and adjust the x/y axis values.



7. Similarly all the queries can be run in the same notebook and different visualizations can be performed.

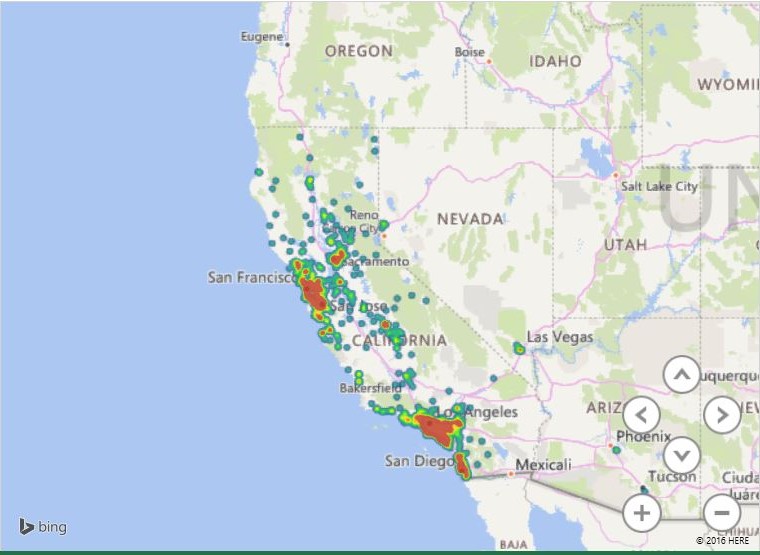
8. Once all the queries are executed go to the file option present on top of the notebook. Select Export from the drop down menu and choose HTML to save the whole notebook as an HTML file.



**Exercise 5: Using Microsoft Excel for Heat Map**

1. In the notebook, run the query to extract postal codes for different states of USA. Download the result in CSV format.

2. Open results in Microsoft Excel in .xlsx format.

3. Go to View in excel. Select 3D Maps. Then click on Open 3D Maps. In the Data tab, select visualization as Heat Map.

Multiple queries can be executed using Databricks and SQL language. Once the desired product is achieved, you must terminate the cluster or it will automatically get terminated after 1 hour.