Intensive Data and Analytics
Summer Workshop for Accounting Courses and Programs

Data Wrangling in Spark with Python

Workshop Instructors:

- Scott Jensen, San José State University
- Esperanza Huerta, San José State University

Workshop website: http://www.sjsu.edu/people/scott.jensen/AnalyticsSummer2018

In this workshop we will be using Jupyter notebooks and Apache Spark to work with data from the USASpending.gov website. The dataset covers spending across Federal agencies for the fourth quarters of 2015 and 2017. We will then explore spending by a single agency, the General Services Administration (GSA), for 2015 through 2017. The data is hosted on Amazon S3 and you will access it directly from your Jupyter notebook, so there are no data files to upload for the workshop.

Jupyter and Spark are both open-source and also available through a number of vendors. For the workshop we will be using free web-based community accounts available through Databricks, a company founded by some of the initial developers of Spark at UC Berkeley's AMPLab. The community accounts are hosted by Databricks on Amazon (AWS) and are excellent for classroom use because students only need a web browser and an Internet connection.

Please sign up for a free Databricks account. Included on the following pages are instructions on how to sign up for a free Databricks account.

Prior to the workshop, a Jupyter notebook will be posted on the website listed at the top of these instructions. Jupyter notebooks are JSON documents and contain only the code and markdown (documentation) – not the data. This makes them lightweight and easy to share. The notebook for the workshop will be approximately 50K (so you could fit 20 of them on an old floppy disk).

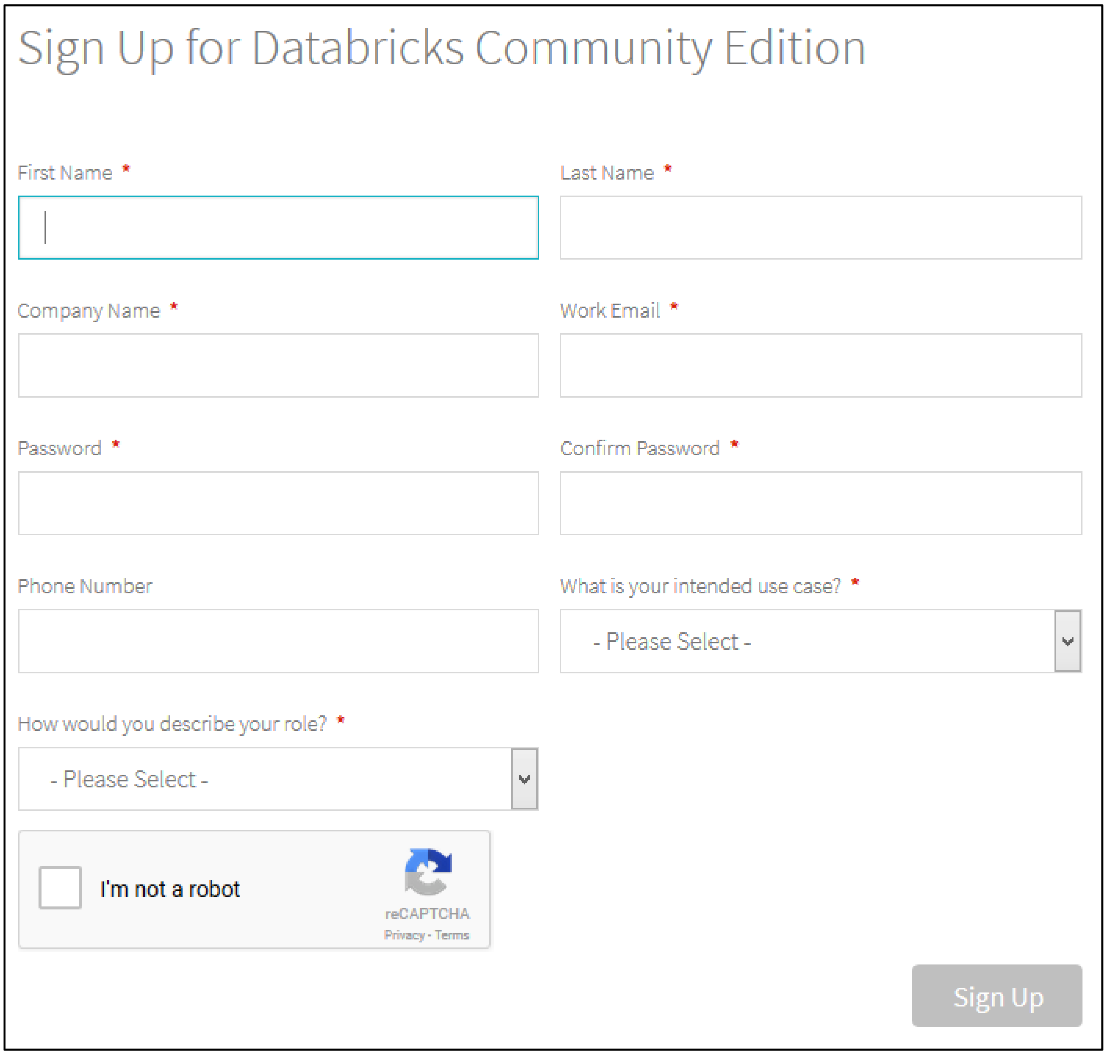
Prior to the workshop, a zipped copy of the notebook for the workshop will be uploaded to the website listed above along with instructions on how to load the notebook into your Databrcks account.

Signing up for a free Databricks account:

In this workshop we will be using the free community edition of the cloud-based Jupyter and Spark platform from Databricks.

Signing up for the Community Edition of Databricks

Anyone can sign up for a free account at the following URL: https://databricks.com/ce



Sign Up for Databricks Community Edition

First Name *

Last Name *

Company Name *

Work Email *

Password *

Confirm Password *

Phone Number

What is your intended use case? *

- Please Select -

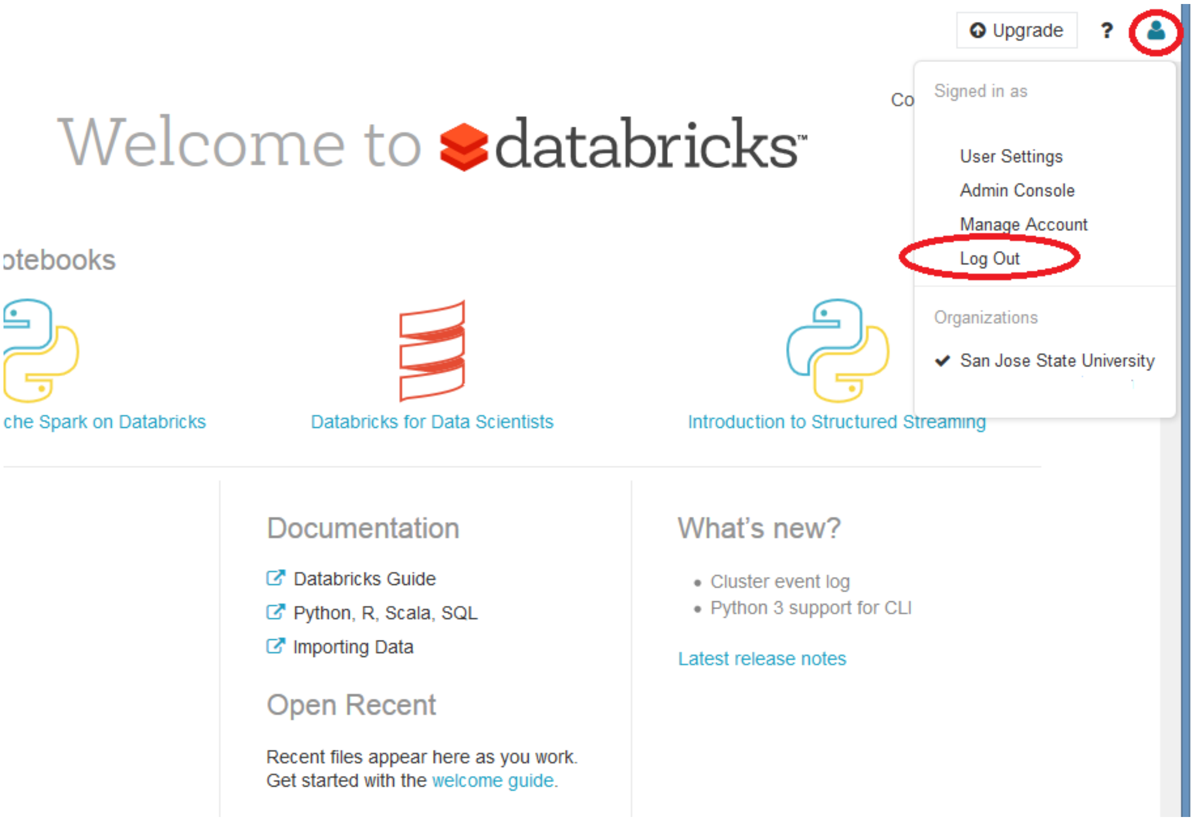
How would you describe your role? *

- Please Select -

After you fill in the above screen and click the Sign Up button, you will be prompted to accept their terms.

Click the agree button at the bottom of that screen.

An email will be sent to the email address you specified when you signed up. When you click on the link in that email, you will be prompted to sign into your Databricks account:



Recent files appear here as you work.
Get started with the welcome guide.

Feel free to use your new account or poke around in it. Keep in mind that in the community edition you are limited to 6GB of data storage.

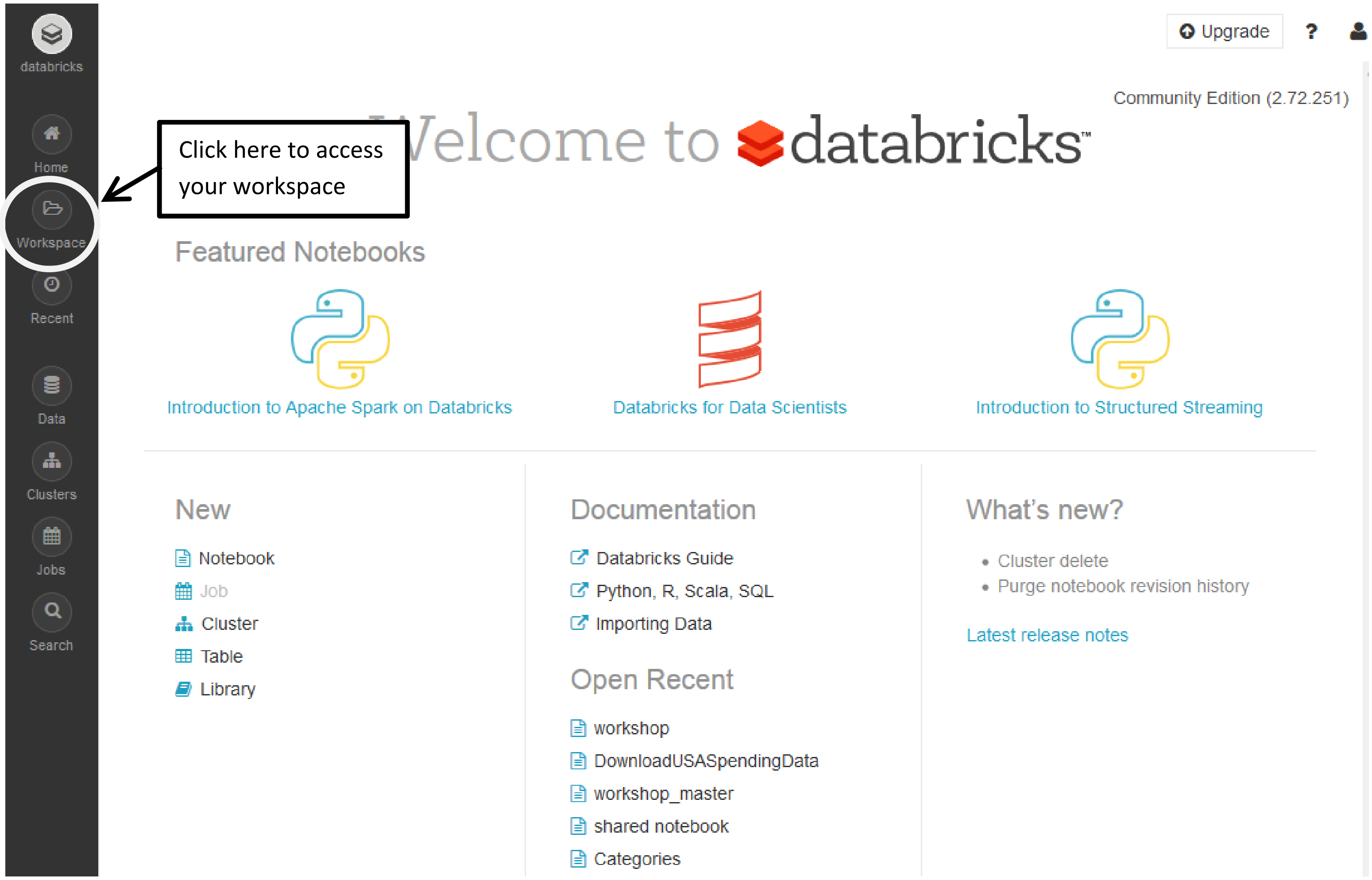
To exit your account, just click on the icon in the upper right-hand corner of the screen and select "Log Out" from the menu that appears (the icon and Log Out option are both circled in the screen capture above).

Importing the Notebook

We will be using a notebook that has been prepopulated with code so that you don't need to start from scratch. Please download the notebook from the conference DropBox folder or from the following website: http://www.sjsu.edu/people/scott.jensen/AnalyticsSummer2018

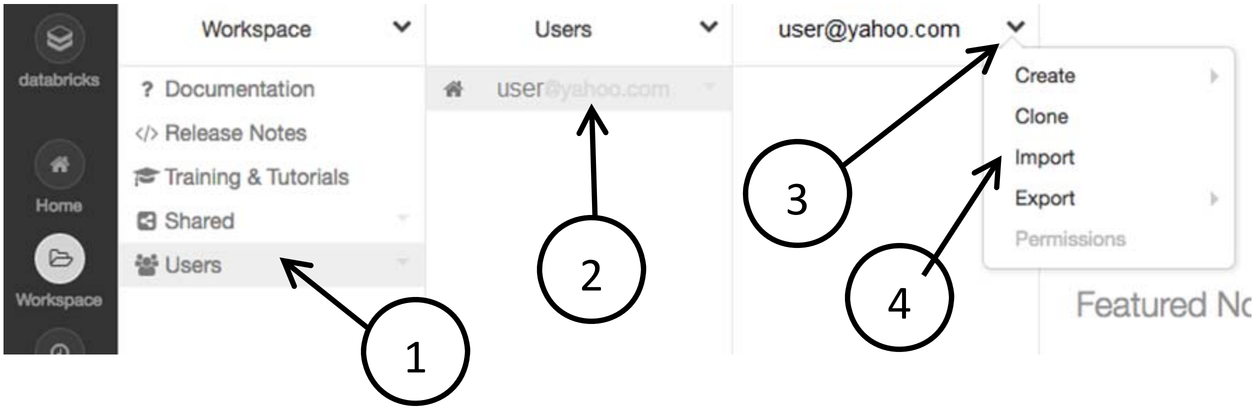
To import the notebook, first log into your new Databricks community account. If you have not yet created an account, see the instructions above. If you have already created an account but have logged out, log back into your account: https://community.cloud.databricks.com/login.html

When you first log into your account, you will be at the screen shown below. When you are in your account, you can always return to this screen by clicking the "databricks" icon at the top of the toolbar on the left side of the screen.



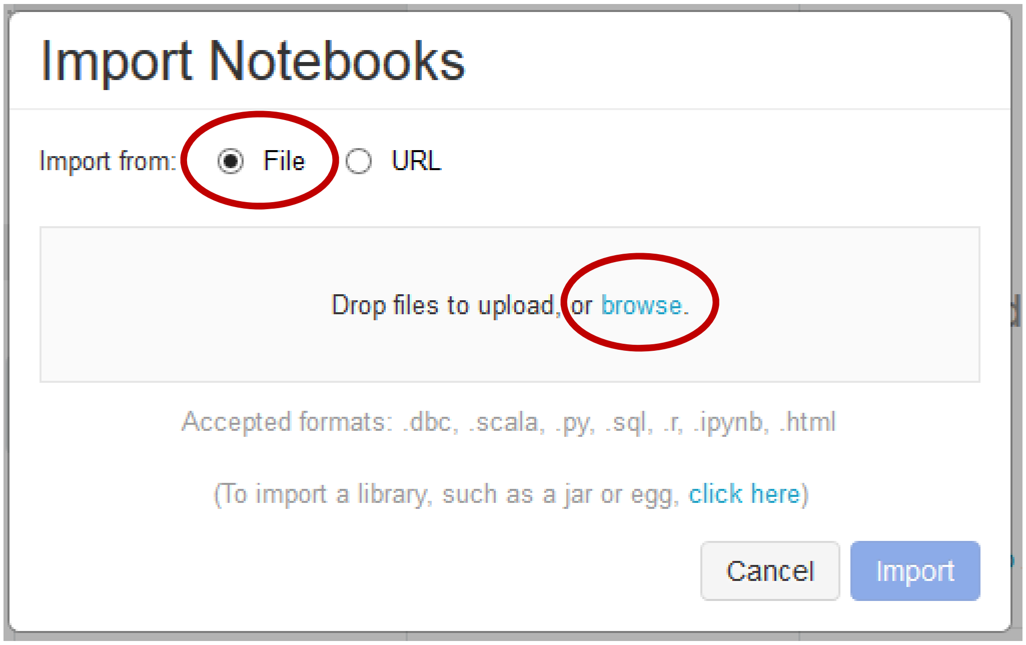
Click here to access your workspace

Click on the Workspace icon in the menu. This will display a workspace list as shown below.



1. Click on the user option in the Workspace list. This will show a new column with a list of the users in your account – currently that is just you, so you should see your email address listed (the email address you used to sign up and log in).
2. Click on your email address. If you had created any notebooks at this time, you would see them in the new column under your email.
3. To import the workshop.ipynb notebook, click on the down arrow next to your email address.
4. Select "Import".

This will display the following Import dialog:


Import Notebooks

Import from: ● File ○ URL

Drop files to upload, or browse.

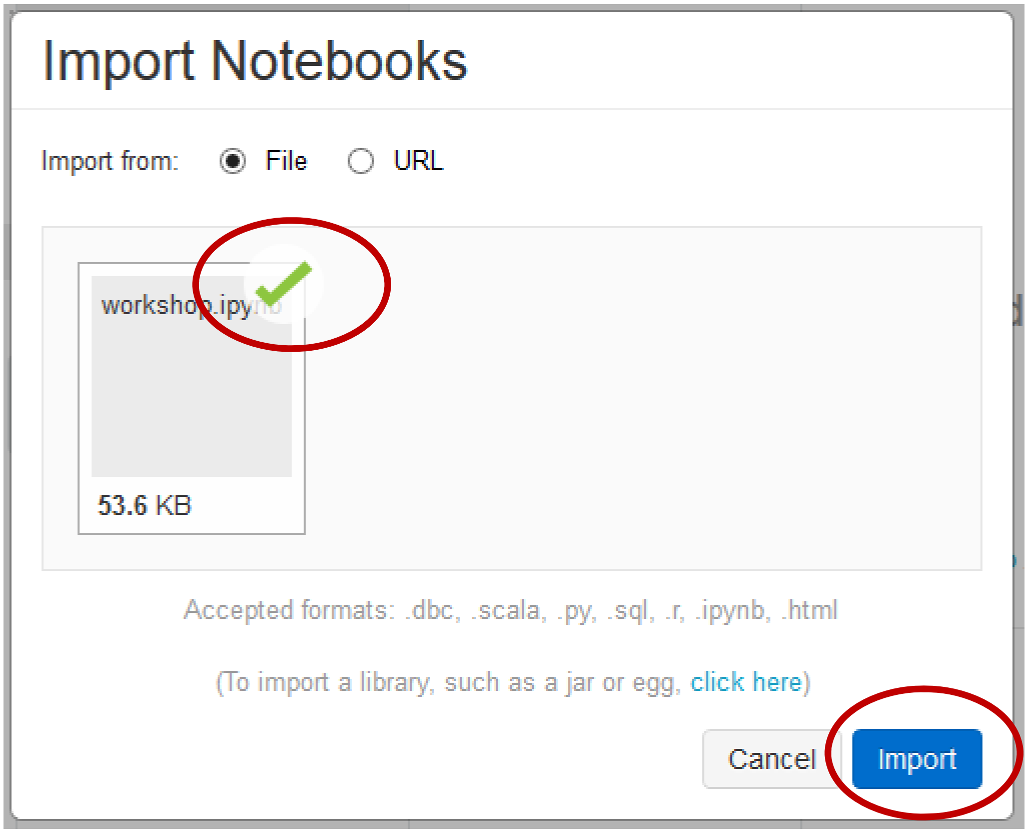
Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .html

(To import a library, such as a jar or egg, click here)

Be sure the "File" option is selected for the "Import from" prompt, and then click the link to browse for the file to upload.

In the File Upload dialog that appears, find where you previously had downloaded the file named workshop.ipynb from either the DropBox folder for the workshop session or from the webpage describing the workshop (see the link at the start of these instructions).

After you select the file, an Import dialog will read the file. This is relatively quick because the file is only a JSON file with code and some images. It should be 53.6KB. When a green checkmark appears, click the Import button at the bottom of the dialog.



Import Notebooks

Import from: ● File ○ URL

workshop.ipynb ✓

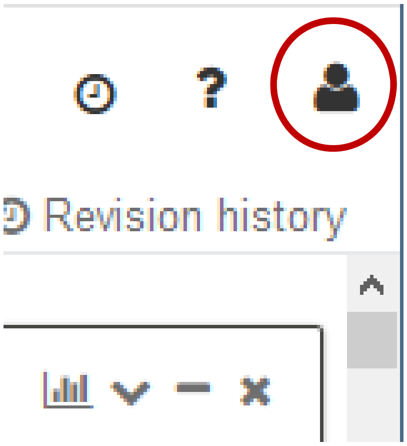
53.6 KB

Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .html

(To import a library, such as a jar or egg, click here)

When the import finishes, you will be in your notebook. You are ready for the workshop!

To log out of your account, click on the little silhouette of a person in the upper right-hand corner and select “Log Out”.



?
Revision history