

Wrangling Big Data for Data-Driven Research Hands-on with Apache Spark and Jupyter Notebooks


Software Installation Instructions:

In this workshop we will be using the free community edition of the cloud-based 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>

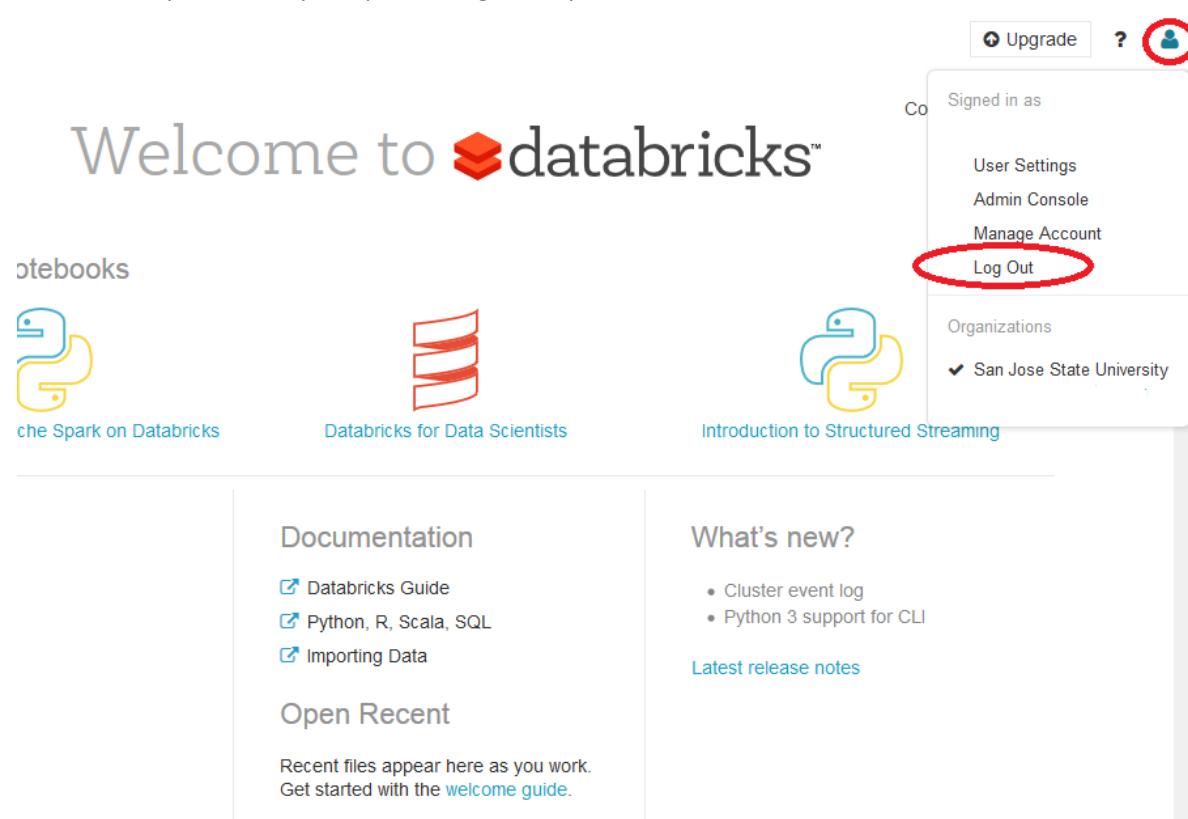
Sign Up for Databricks Community Edition

First Name *	Last Name *
<input type="text"/>	<input type="text"/>
Company Name *	Work Email *
<input type="text"/>	<input type="text"/>
Password *	Confirm Password *
<input type="password"/>	<input type="password"/>
Phone Number	What is your intended use case? *
<input type="text"/>	<input type="text" value="- Please Select -"/>
How would you describe your role? *	
<input type="text" value="- Please Select -"/>	
<input type="checkbox"/> I'm not a robot  reCAPTCHA Privacy - Terms	
<input type="button" value="Sign Up"/>	

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:



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).

For the workshop:

1. We will be using notebooks that have been prepopulated with code so that you don't need to start from scratch. Please check back a week before the workshop and download the notebooks: <http://www.sjsu.edu/people/scott.jensen/AoMBigDataWorkshop>
The notebooks will include instructions on how to upload them to your new account.
2. Please download the JSON version of Round 11 of the Yelp Dataset Challenge: <https://www.yelp.com/dataset>
You don't need to do anything further with the dataset at this point.