

# Databricks


Apache Spark™ made a big step towards making big data simple by providing a unified framework for building data pipelines. Databricks takes this further by providing a zero-management cloud platform built around Spark that delivers

- 1) fully managed Spark clusters,
- 2) an interactive workspace for exploration and visualisation,
- 3) a production pipeline scheduler, and
- 4) a platform for powering your favorite Spark-based applications.

## Signing up in Databricks Community

This course need a Databrick notebook and cluster to run the queries. This workspace is provided by Databricks community. To make a free account click [here](#).

You need to fill out the sign up form. After giving you personal information, select the “intended use case” and “describe your role” as follow :

 Platform Solutions Customers Learn Partners Events Open Source Company 🔍 SUPPORT CONTACT LOG IN [TRY DATABRICKS](#)

## Sign Up for Databricks Community Edition

First Name \*

Last Name \*

Company Name \*

Work Email \*

Phone Number


What is your intended use case? \*

How would you describe your role? \*

Personal - Learning Spark

Student

☐ Keep me informed with the occasional update about Databricks and Apache Spark™.  
By clicking "Sign Up", you agree to the [Terms of Service](#) and the [Privacy Policy](#).

☐ I'm not a robot 

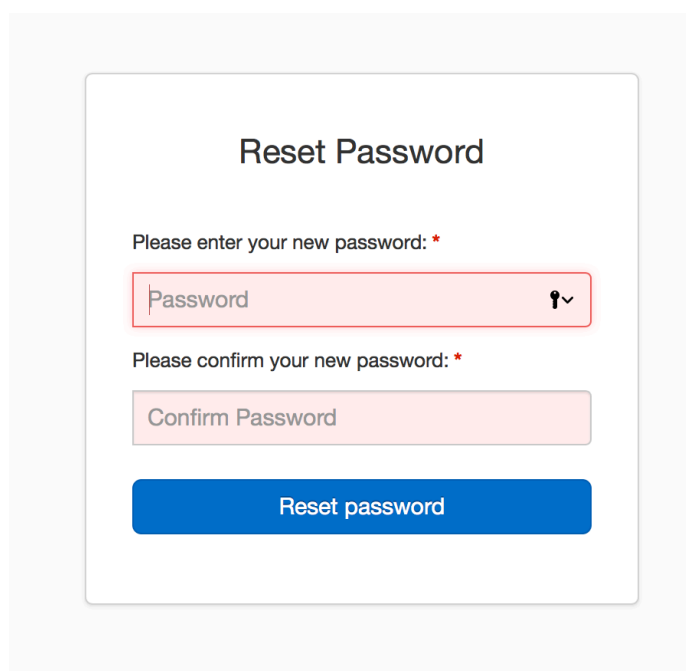
Sign Up

After successfully signing up, check you mail box of the corresponding email address.

## Time to check your email!

Thank you for signing up. Now it's time to validate your email address.  
Please check the email you provided for next steps.

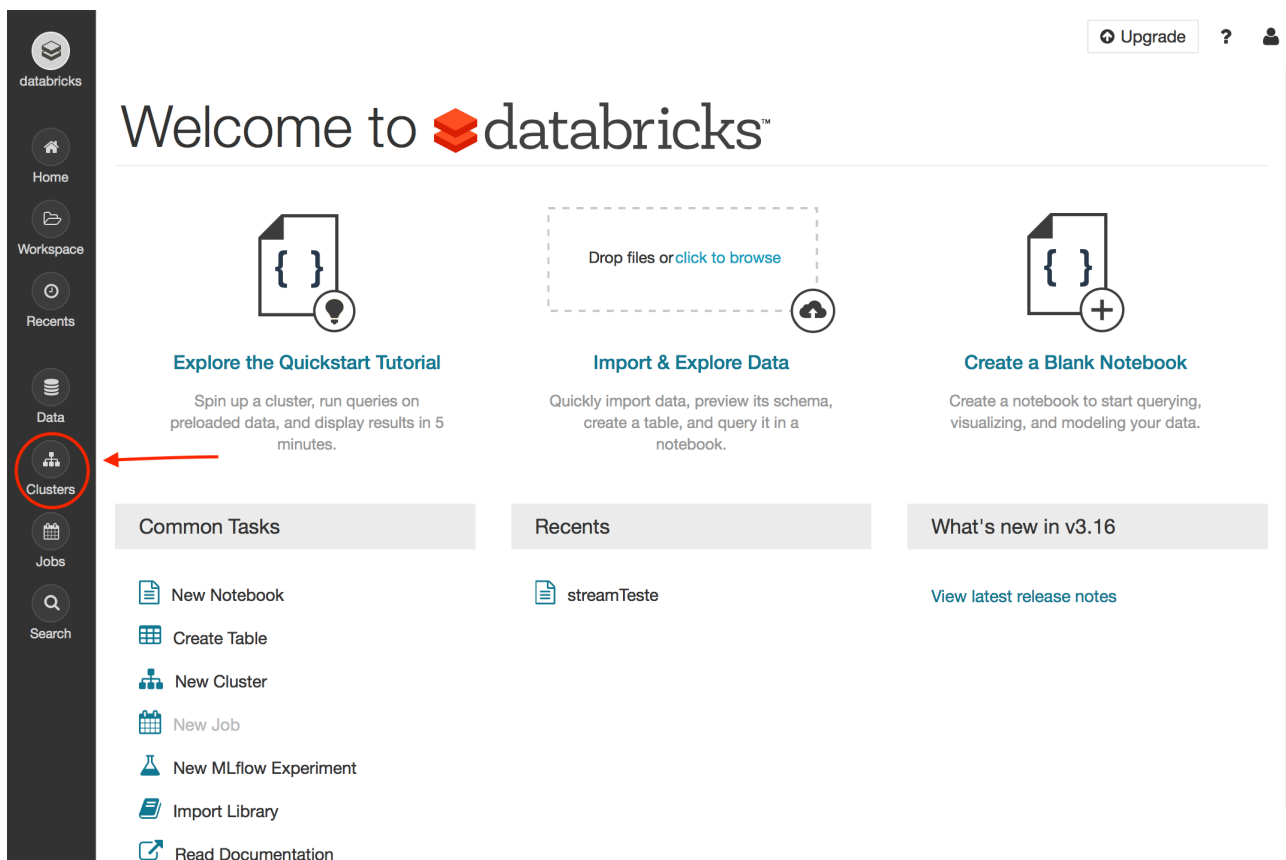
Click on the received link and set your password :



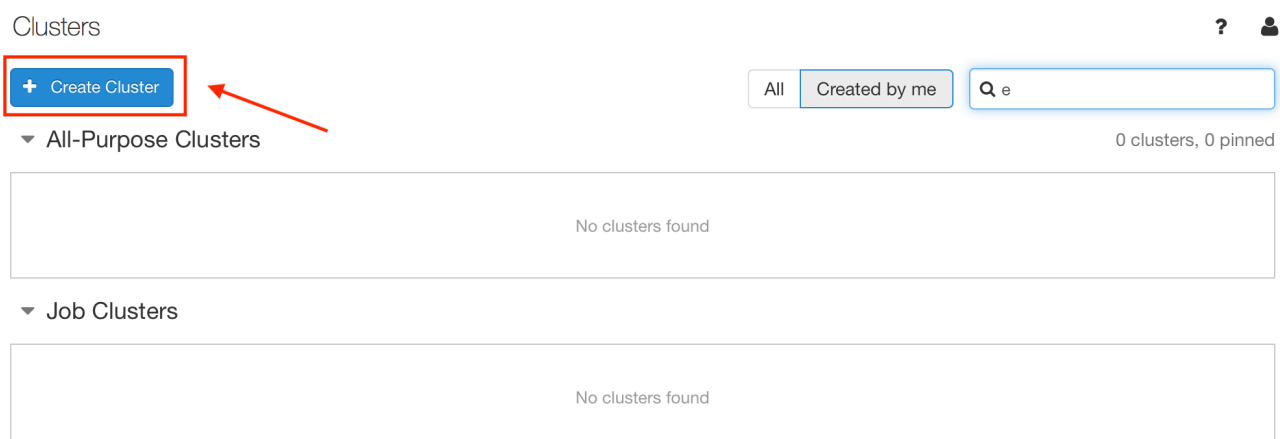
The image shows a 'Reset Password' form. It has a title 'Reset Password' at the top. Below the title, there are two input fields. The first field is labeled 'Please enter your new password: \*' and contains the text 'Password'. The second field is labeled 'Please confirm your new password: \*' and contains the text 'Confirm Password'. Below these fields is a blue button labeled 'Reset password'.

You will then directed to the Databricks community home page. Next step is to create a cluster. It is explained in the next page.

# Cluster Creation



Click on “Clusters” icon as represented in the image above. Then click on “Create Cluster” as mentioned below :



Following the next image below, give a name to your cluster and set the “DataBricks Runtime Version” to : 5.5 LTS (Scala 2.11, Spark 2.4.3)  
Keep the Python version to 3.

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Create Cluster

New Cluster

Cancel

Create Cluster

Workers: 0.0 GB Memory, 0 Cores, 0 DBU

Driver: 15.3 GB Memory, 2 Cores, 1 DBU

Cluster Name

myCluster

Databricks Runtime Version

Runtime: 5.5 LTS (Scala 2.11, Spark 2.4.3)

Python Version

3

Instance

Free 15GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For more configuration options, please [upgrade your Databricks subscription](#).

Instances

Spark

Availability Zone

us-west-2c

UI

JSON

Click on “Create Cluster”. It takes few minutes for the cluster to become available.

Clusters

+ Create Cluster

All Created by me

Filter

▼ All-Purpose Clusters

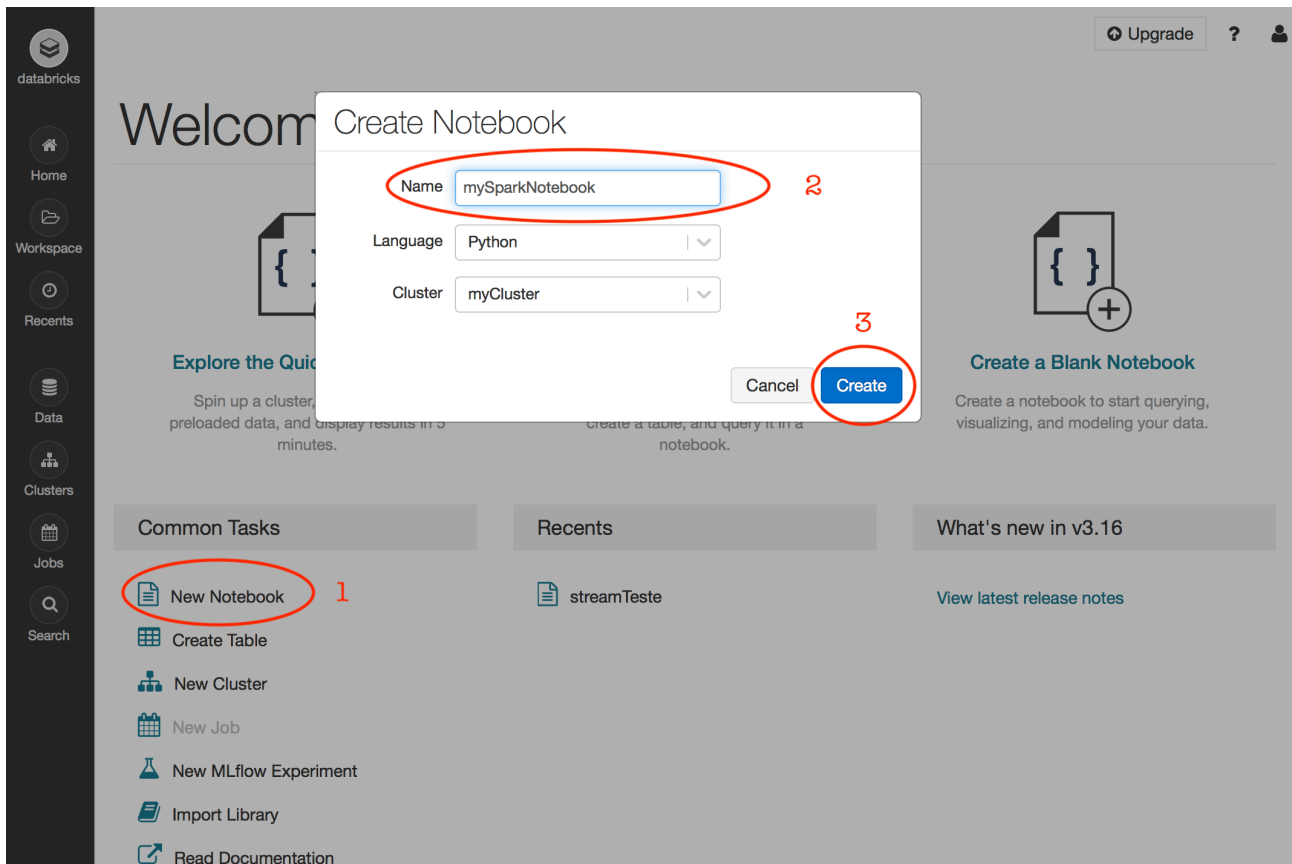
1 clusters, 0 pinned

Name	State	Nodes	Driver	Worker	Runtime	Creator	
myCluster	Running	1 (0 spot)	Community ..	Community ..	5.5 LTS (inclu...	farhang@nai..	0

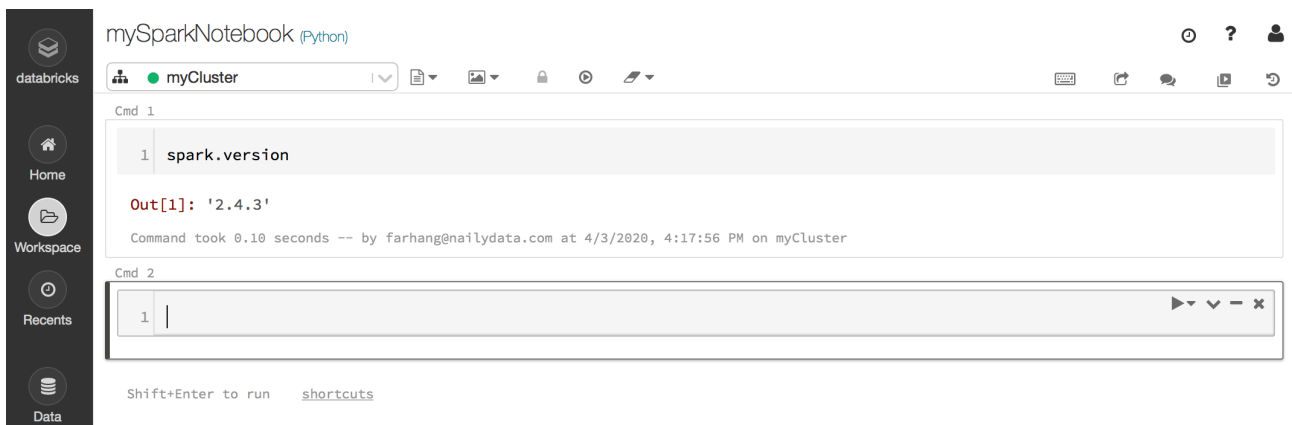
The green light shows that your cluster is available and working. On the menu bar click on databricks icon on top left to come back to the Databricks home page. You should then create a notebook as explained in the next page.

## Create a Notebook

Click on “New Notebook” as shown below and give a name to your notebook :

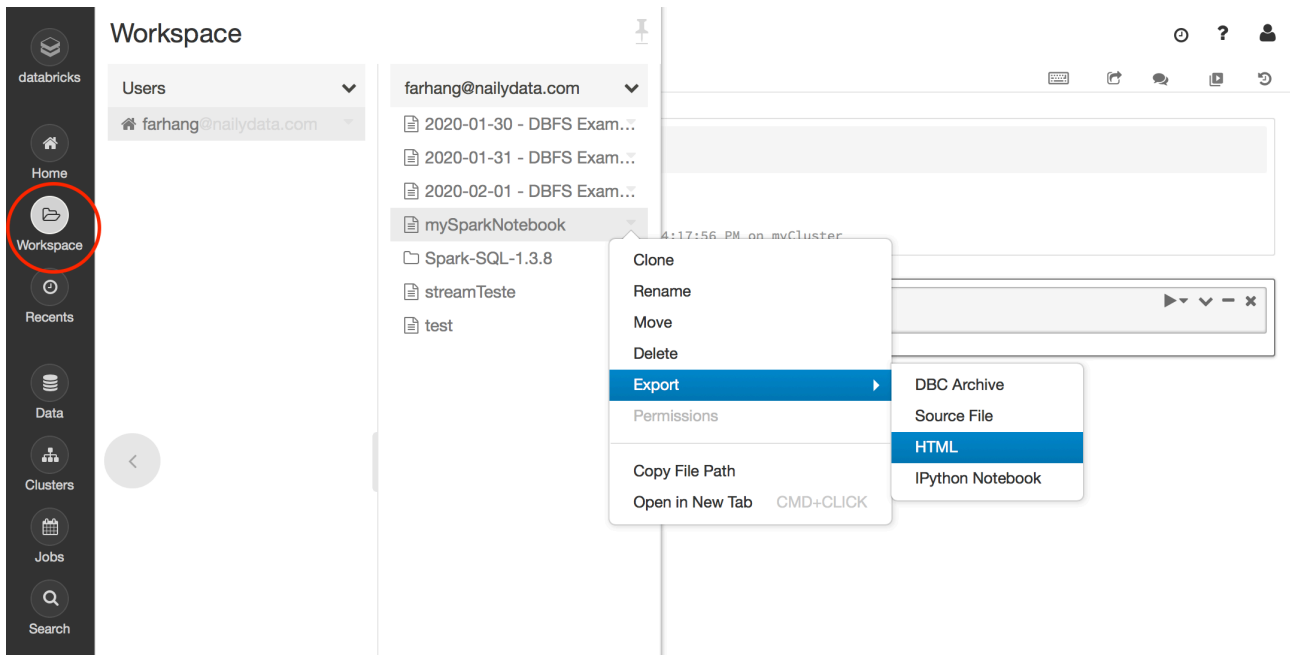


After clicking on “Create” you can start writing in the notebook :



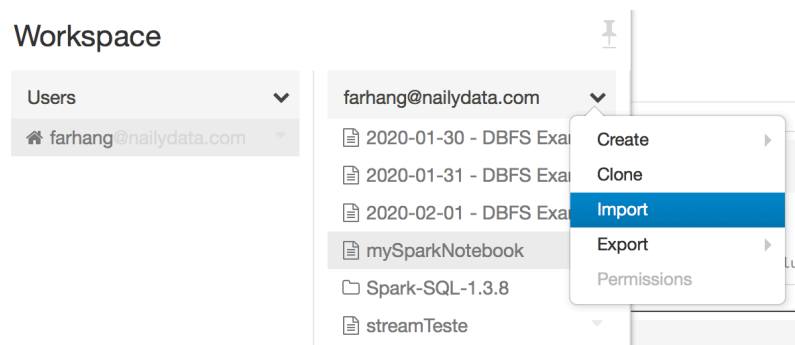
In the first cell write `spark.version` to see the Apache spark version of your notebook.

You can see the lists of your created notebooks by clicking on Workspace icon in the menu bar :



You will need to save your notebook at your local machine. Export your notebook in “DBC Archive” format if you want to reimport it later on Databricks. To send me your work, please export the notebook in HTML format.

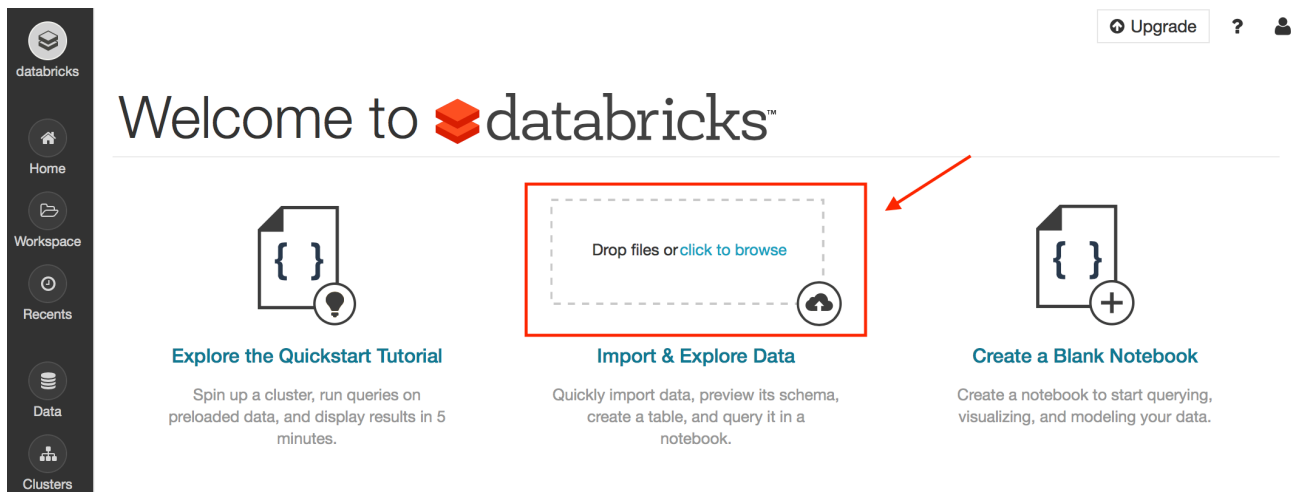
To import a notebook :



To know about how to upload data into the Databricks cluster see the next page.

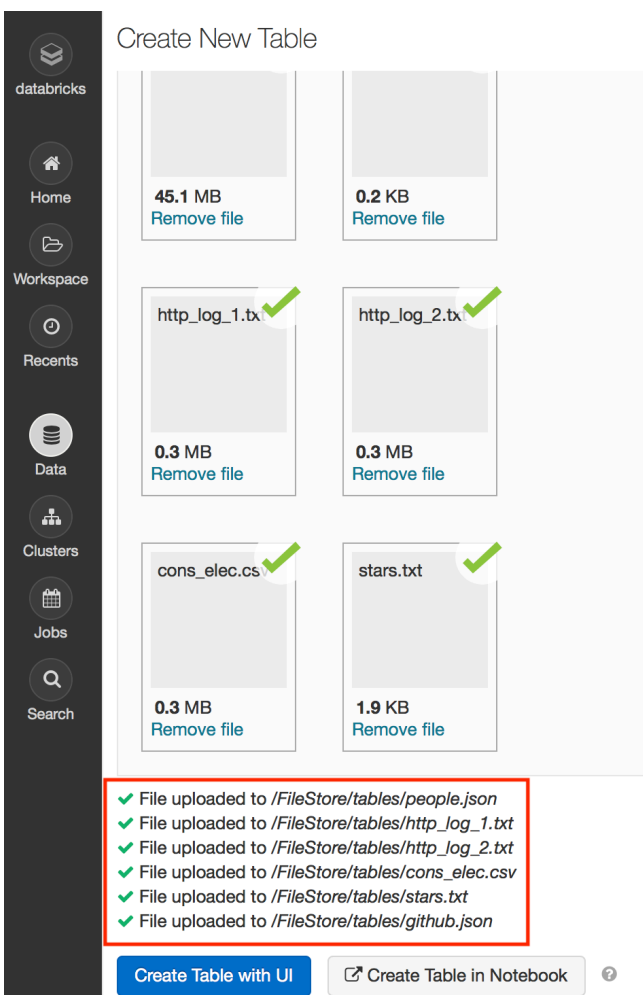
# Upload Data

To upload the data go to the Databricks community home page by clicking on databricks icon in the menu bar. Then click on “click to brows” menu :



“

Select your data file(s) from your local machine. It starts loading them. To read the files from your notebook, you should use the access path that is mentioned at the bottom of the page.



All your files will be saved on this directory in the cluster : **/FileStore/tables/**

Enjoy the adventure !