

# Getting started with Databricks and Spark

Ladle Patel

# Step 1:

Go to the below URL and click on **GET STARTED** under Community Edition.

**NOTE: Please don't create FREE TRIAL account**

<https://databricks.com/try-databricks>

## DATABRICKS PLATFORM – FREE TRIAL

For businesses looking for a zero-management cloud platform built around Apache Spark

- Unlimited clusters that can scale to any size
- Job scheduler to execute jobs for production pipelines
- Fully interactive notebook with collaboration, dashboards, REST APIs
- Advanced security, role-based access controls, and audit logs
- Single Sign On support
- Integration with BI tools such as Tableau, Qlik, and Looker
- 14-day full feature trial (excludes cloud charges)

GET STARTED

## COMMUNITY EDITION

For students and educational institutions just getting started with Apache Spark

- Single cluster limited to 6GB and no worker nodes
- Basic notebook without collaboration
- Limited to 3 max users
- Public environment to share your work


GET STARTED

# Step 2:

Fill up the details and click on the Sign Up button.

NOTE: If you don't have (or) work in a company, then type NA in the Company Name field.

## Sign Up for Databricks Community Edition

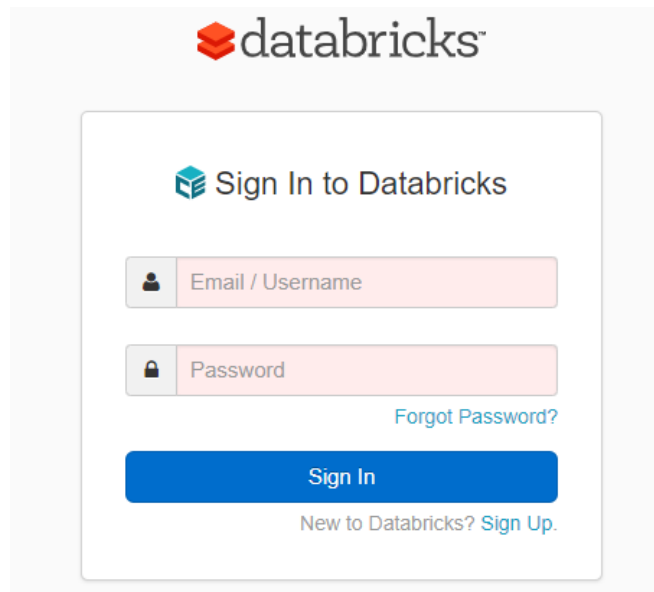
First Name *	Last Name *
<input type="text"/>	<input type="text"/>
Company Name *	Work Email *
<input type="text"/>	<input type="text"/>
Phone Number	What is your intended use case? *
<input type="text"/>	- Please Select -
How would you describe your role? *	
- Please Select -	
<input checked="" type="checkbox"/> Keep me informed with the occasional updates about Databricks and Apache Spark™.	
<div><input type="checkbox"/> I'm not a robot</div> <div> reCAPTCHA Privacy - Terms</div>	

Sign Up

## Step 3:

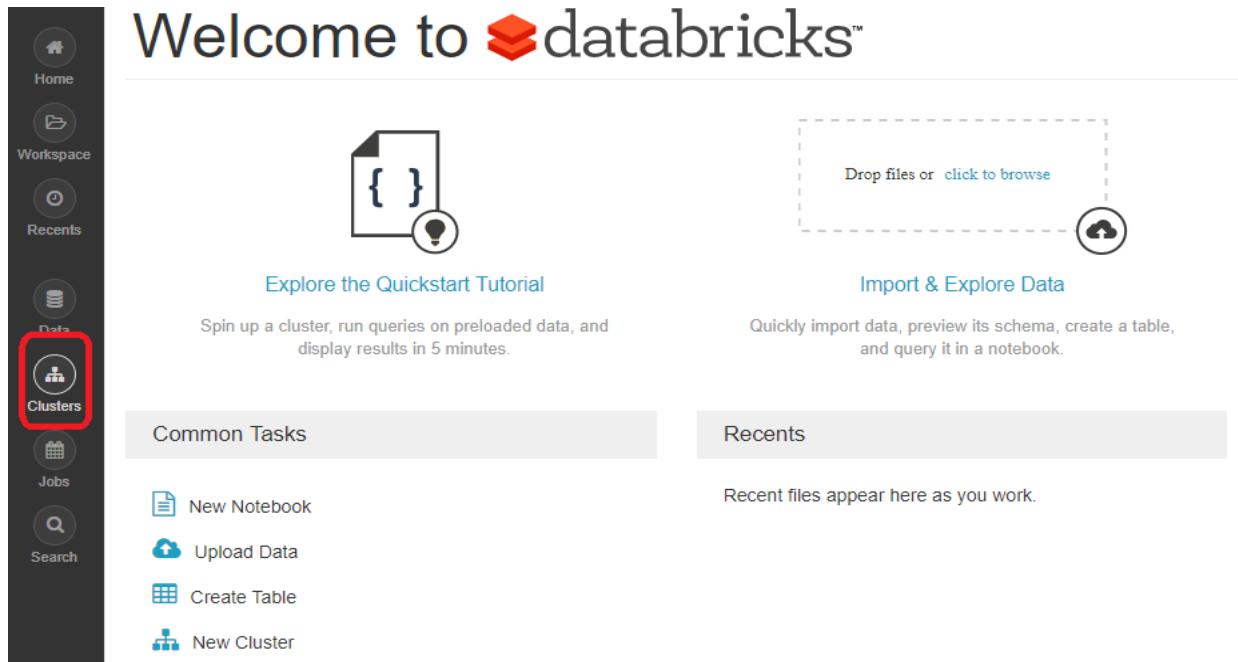
You will receive a link through an email to verify your account, click on the link to complete the account set up. Use below link to login to your account.

<https://community.cloud.databricks.com/login.html>

A screenshot of the Databricks login page. At the top is the Databricks logo. Below it is a white box with a rounded border containing the login form. Inside the box, there is a heading "Sign In to Databricks" with a small cube icon. Below the heading are two input fields: the first is labeled "Email / Username" with a person icon, and the second is labeled "Password" with a lock icon. To the right of the password field is a link "Forgot Password?". Below the input fields is a blue "Sign In" button. At the bottom of the box is a link "New to Databricks? Sign Up." data-bbox="270 417 610 965"/>


# Step 4:

Creating New Cluster.




The screenshot shows the Databricks home page. On the left is a dark sidebar with navigation icons and labels: Home, Workspace, Recents, Data, Clusters (highlighted with a red box), Jobs, and Search. The main content area has a large heading "Welcome to databricks™". Below it are two primary actions: "Explore the Quickstart Tutorial" (with a document icon) and "Import & Explore Data" (with a cloud upload icon). The "Import & Explore Data" section includes a dashed box with the text "Drop files or [click to browse](#)". At the bottom, there are two sections: "Common Tasks" with links for "New Notebook", "Upload Data", "Create Table", and "New Cluster"; and "Recents" with the text "Recent files appear here as you work."

## Welcome to databricks™





 [Explore the Quickstart Tutorial](#)

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

 [Import & Explore Data](#)

Quickly import data, preview its schema, create a table, and query it in a notebook.

**Common Tasks**

-  [New Notebook](#)
-  [Upload Data](#)
-  [Create Table](#)
-  [New Cluster](#)

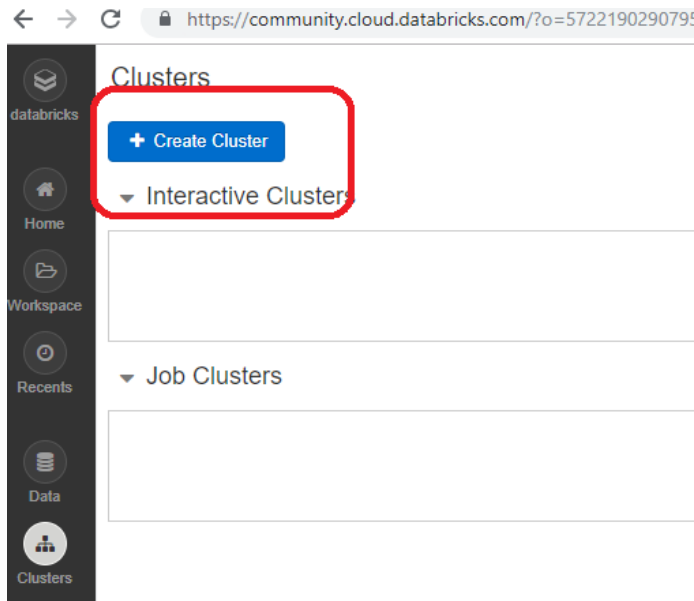
**Recents**

Recent files appear here as you work.



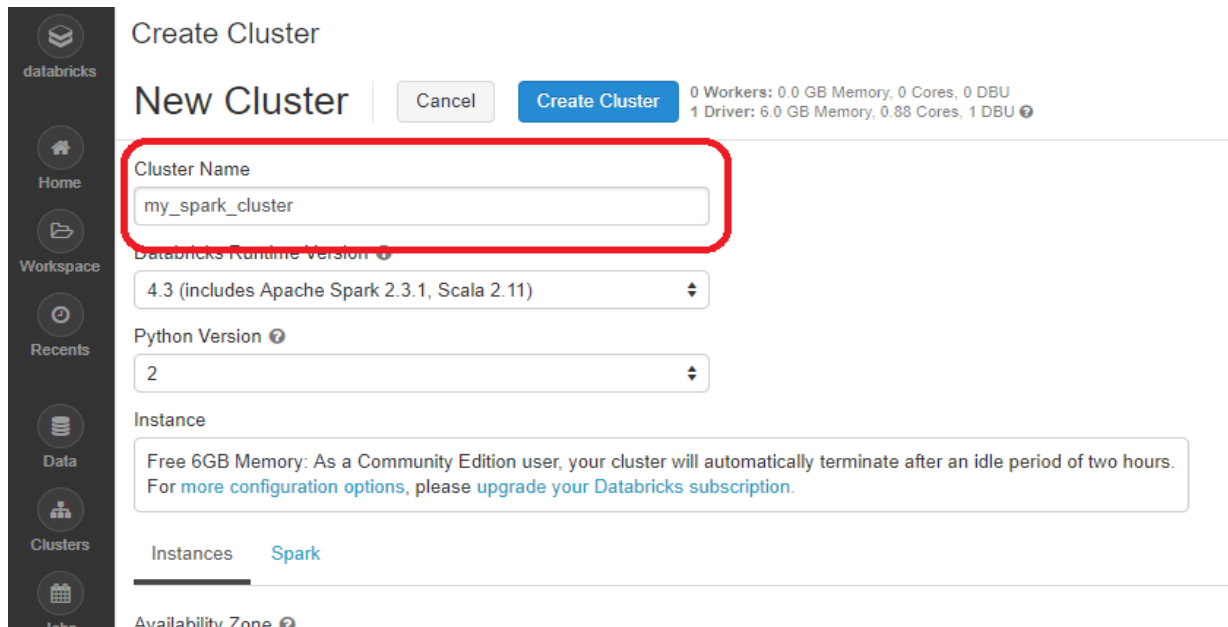
databricks

# Step 4: Cont..



# Step 5:

Enter cluster name



The screenshot shows the 'Create Cluster' page in the Databricks interface. A sidebar on the left contains navigation icons for Home, Workspace, Recents, Data, Clusters, and Jobs. The main content area is titled 'Create Cluster' and 'New Cluster'. It features a 'Cluster Name' input field with the text 'my\_spark\_cluster', which is highlighted by a red rectangle. To the right of the input field are 'Cancel' and 'Create Cluster' buttons. Below the input field are dropdown menus for 'Databricks Runtime Version' (set to 4.3) and 'Python Version' (set to 2). A section titled 'Instance' contains a message about the free 6GB memory limit for Community Edition users. At the bottom, there are tabs for 'Instances' and 'Spark', and a partially visible 'Availability Zone' section.

Create Cluster

New Cluster

Cancel Create Cluster

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU  
1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBU ⓘ

Cluster Name

my\_spark\_cluster

Databricks Runtime Version ⓘ

4.3 (includes Apache Spark 2.3.1, Scala 2.11) ⌵

Python Version ⓘ

2 ⌵

Instance

Free 6GB 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 ⓘ



databricks

# Step 6:



Upload data from your laptop to the cluster.

The screenshot displays the Databricks workspace interface. On the left is a dark sidebar with navigation icons and labels: Home, Workspace, Recents, Data, Clusters, Jobs, and Search. The main content area is divided into two columns. The left column features a document icon with curly braces and a lightbulb, titled 'Explore the Quickstart Tutorial', with the description 'Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.' The right column features a dashed box with the text 'Drop files or [click to browse](#)' and a cloud upload icon, titled 'Import & Explore Data', with the description 'Quickly import data, preview its schema, create a table, and query it in a notebook.' Below these are two sections: 'Common Tasks' and 'Recents'. The 'Common Tasks' section lists several actions: 'New Notebook', 'Upload Data' (highlighted with a red rectangle), 'Create Table', 'New Cluster', 'New Job', 'Import Library', and 'Read Documentation'. The 'Recents' section contains the text 'Recent files appear here as you work.'

Home  
Workspace  
Recents  
Data  
Clusters  
Jobs  
Search

Explore the Quickstart Tutorial  
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Import & Explore Data  
Quickly import data, preview its schema, create a table, and query it in a notebook.

Common Tasks

- New Notebook
- Upload Data
- Create Table
- New Cluster
- New Job
- Import Library
- Read Documentation

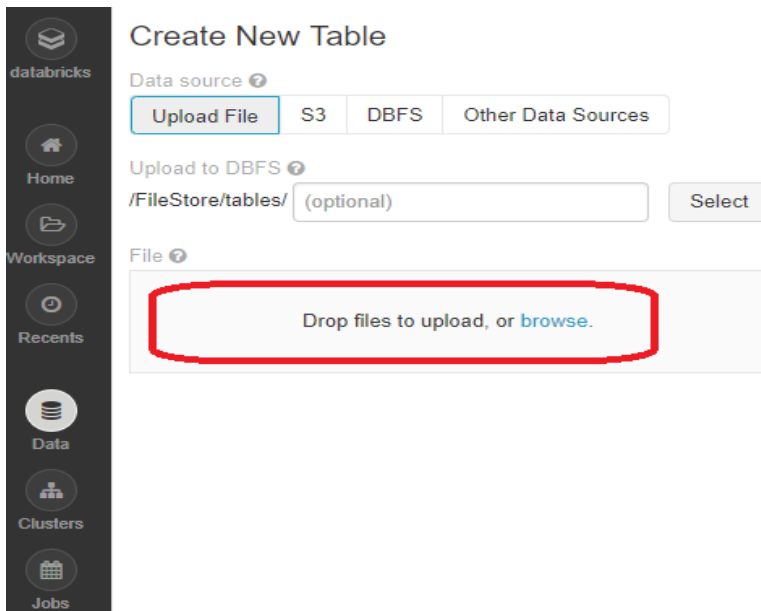
Recents

Recent files appear here as you work.



# Step 6: Cont..

Upload data from your laptop to the cluster.



**Create New Table**

Data source ?

**Upload File** S3 DBFS Other Data Sources

Upload to DBFS ?

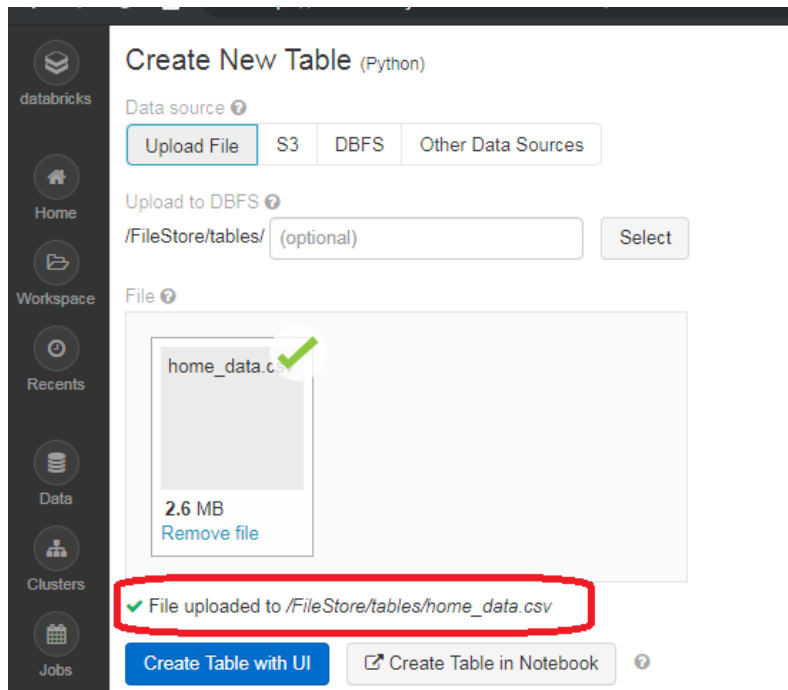
/FileStore/tables/ (optional) **Select**

File ?

Drop files to upload, or [browse](#).

# Step 7:

Copy the path.



The screenshot shows the 'Create New Table (Python)' interface in Databricks. The left sidebar contains navigation icons for Home, Workspace, Recents, Data, Clusters, and Jobs. The main panel has the following elements:

- Create New Table (Python)**: Title of the interface.
- Data source**: A section with three tabs: 'Upload File' (selected), 'S3', and 'DBFS'.
- Upload to DBFS**: A section with a text input field containing '/FileStore/tables/' and '(optional)', followed by a 'Select' button.
- File**: A section showing a file upload confirmation. A green checkmark is next to the file name 'home\_data.csv'. Below the file name, it says '2.6 MB' and 'Remove file'.
- Confirmation message**: A green checkmark followed by the text 'File uploaded to /FileStore/tables/home\_data.csv', which is highlighted with a red rectangle.
- Buttons**: At the bottom, there are two buttons: 'Create Table with UI' and 'Create Table in Notebook'.



databricks

# Step 8:



## Create new notebook

The screenshot displays the Databricks workspace interface. On the left is a vertical sidebar with navigation icons and labels: Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main content area is divided into two columns. The left column features a 'Common Tasks' section with a list of actions: 'New Notebook' (highlighted with a red rectangle), 'Upload Data', 'Create Table', 'New Cluster', 'New Job', 'Import Library', and 'Read Documentation'. Above this list are two cards: 'Explore the Quickstart Tutorial' (with a code icon and a lightbulb) and 'Import & Explore Data' (with a cloud upload icon). The right column features a 'Recents' section with the text 'Recent files appear here as you work.' and a dashed box at the top containing the text 'Drop files or click to browse' and a cloud upload icon.

Home  
Workspace  
Recent  
Data  
Clusters  
Jobs  
Search

**Common Tasks**

- New Notebook**
- Upload Data
- Create Table
- New Cluster
- New Job
- Import Library
- Read Documentation

**Explore the Quickstart Tutorial**  
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

**Import & Explore Data**  
Quickly import data, preview its schema, create a table, and query it in a notebook.

**Recents**  
Recent files appear here as you work.

Drop files or [click to browse](#)



databricks



## Step 9:

Provide the name of the notebook and chose the language.

**Create Notebook**

Name








Language

Cluster

Recents

# Step 10:

## Read the data in spark

 Home  
 Workspace  
 Recents  
 Data  
 Clusters  
 Jobs  
 Search

Cmd 1

```
1 home_data = spark.read.csv('FileStore/tables/home_data.csv', header = True, inferSchema = True)
```

▶ (2) Spark Jobs

▶ home\_data: pyspark.sql.dataframe.DataFrame = [id: long, date: string ... 19 more fields]

Command took 0.89 seconds -- by lp.dataninja@gmail.com at 11/23/2018, 7:17:00 PM on l1l1

Cmd 2

```
1 display(home_data)
```

▶ (1) Spark Jobs

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_a
7129300520	20141013T000000	221900	3	1	1180	5650	1	0	0	3	7	1180
6414100192	20141209T000000	538000	3	2.25	2570	7242	2	0	0	3	7	2170
5631500400	20150225T000000	180000	2	1	770	10000	1	0	0	3	6	770
2487200875	20141209T000000	604000	4	3	1960	5000	1	0	0	5	7	1050
1954400510	20150218T000000	510000	3	2	1680	8080	1	0	0	3	8	1680
7237550310	20140512T000000	1225000	4	4.5	5420	101930	1	0	0	3	11	3890
1321400060	20140627T000000	257500	3	2.25	1715	6819	2	0	0	3	7	1715
2008000270	20150115T000000	291850	3	1.5	1060	9711	1	0	0	3	7	1060

# Further Reading

<https://docs.databricks.com/spark/latest/training/index.html>

<https://spark.apache.org/>

<https://github.com/databricks/Spark-The-Definitive-Guide>