Getting started with Databricks and Spark

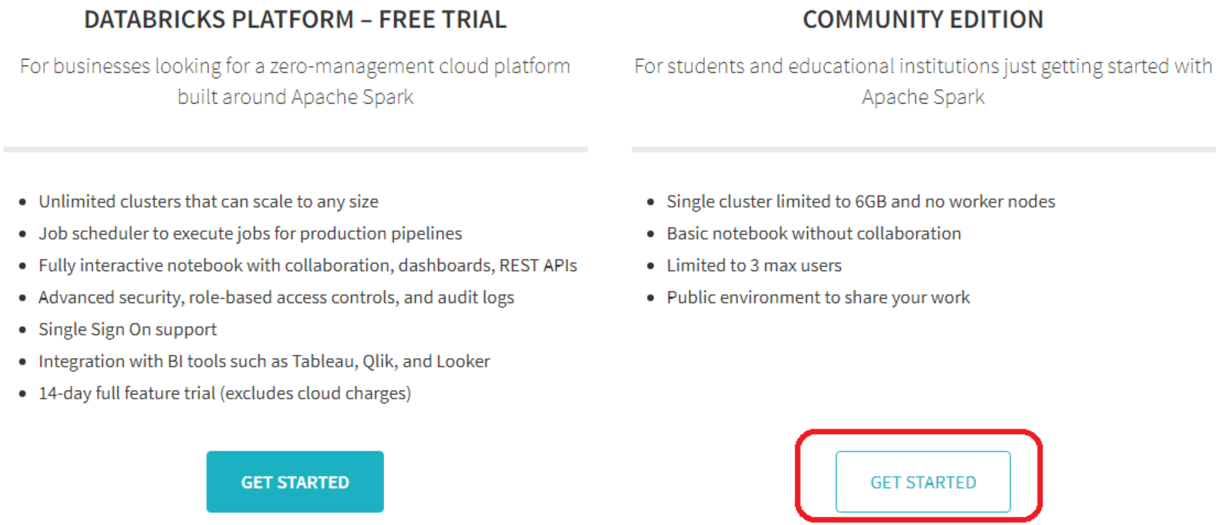
Ladle Patel

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)

COMMUNITY EDITION

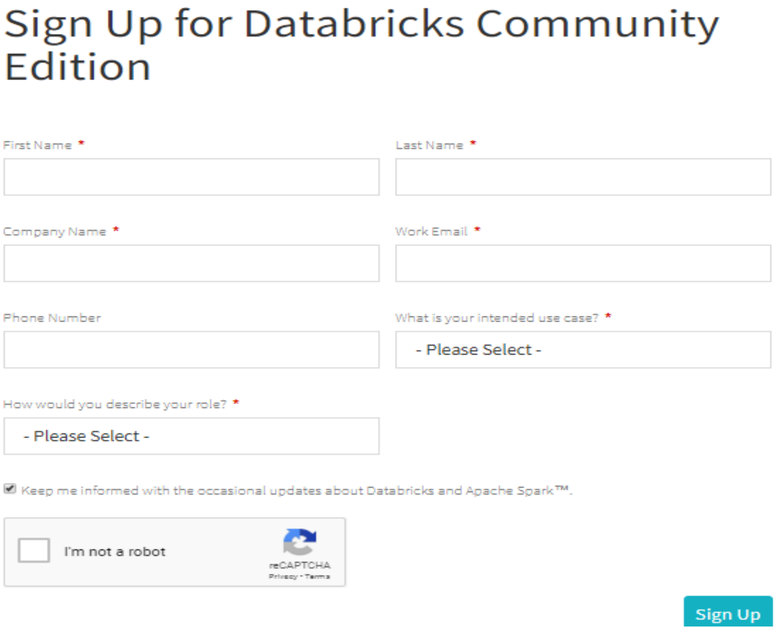
For students and educational institutions just getting started with Apache Spark

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

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 *

Company Name *

Work Email *

Phone Number

What is your intended use case? *
- Please Select -

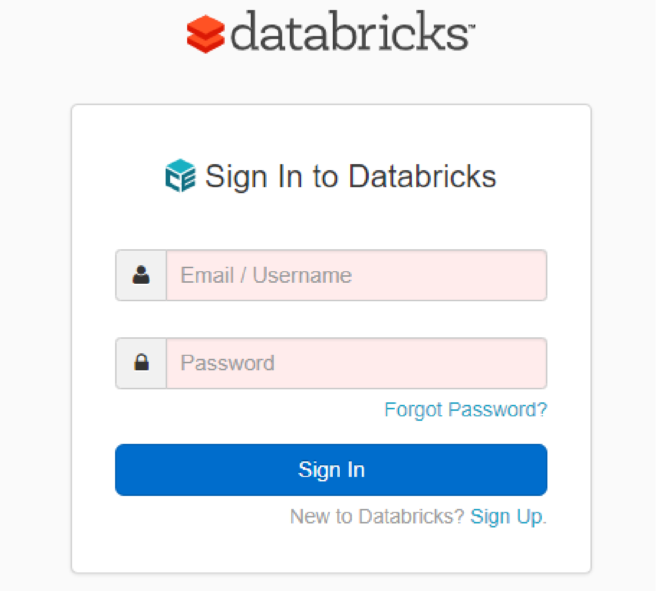
How would you describe your role? *
- Please Select -

☑ Keep me informed with the occasional updates about Databricks and Apache Spark™.

□ I'm not a robot

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



databricks

Sign In to Databricks

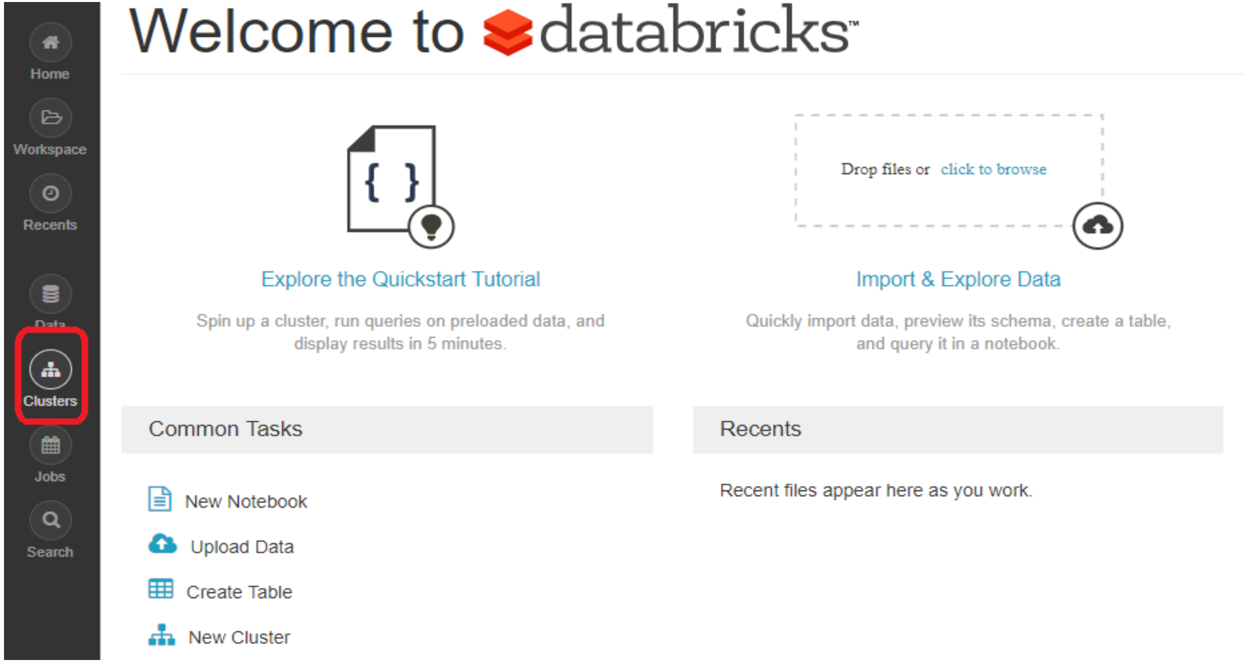
Email / Username

Password

New to Databricks? Sign Up.

Step 4:

Creating New Cluster.



Welcome to databricks

Drop files or click to browse

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.

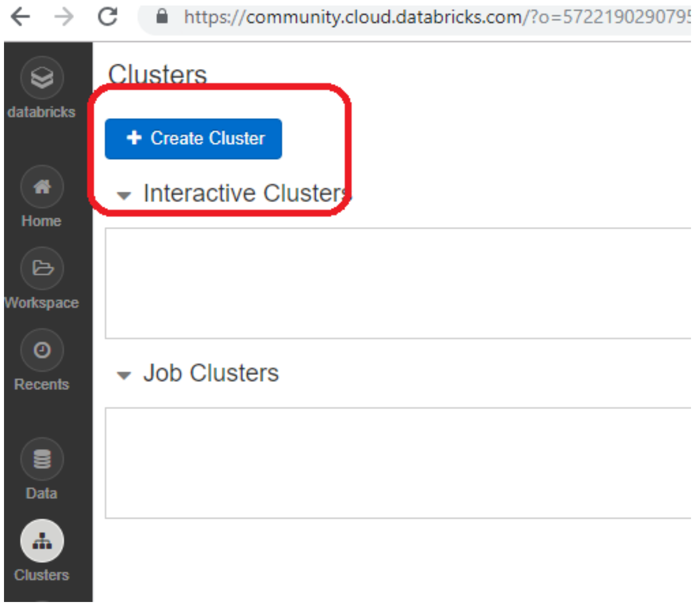
Common Tasks

- New Notebook
- Upload Data
- Create Table
- New Cluster

Recents

Recent files appear here as you work.

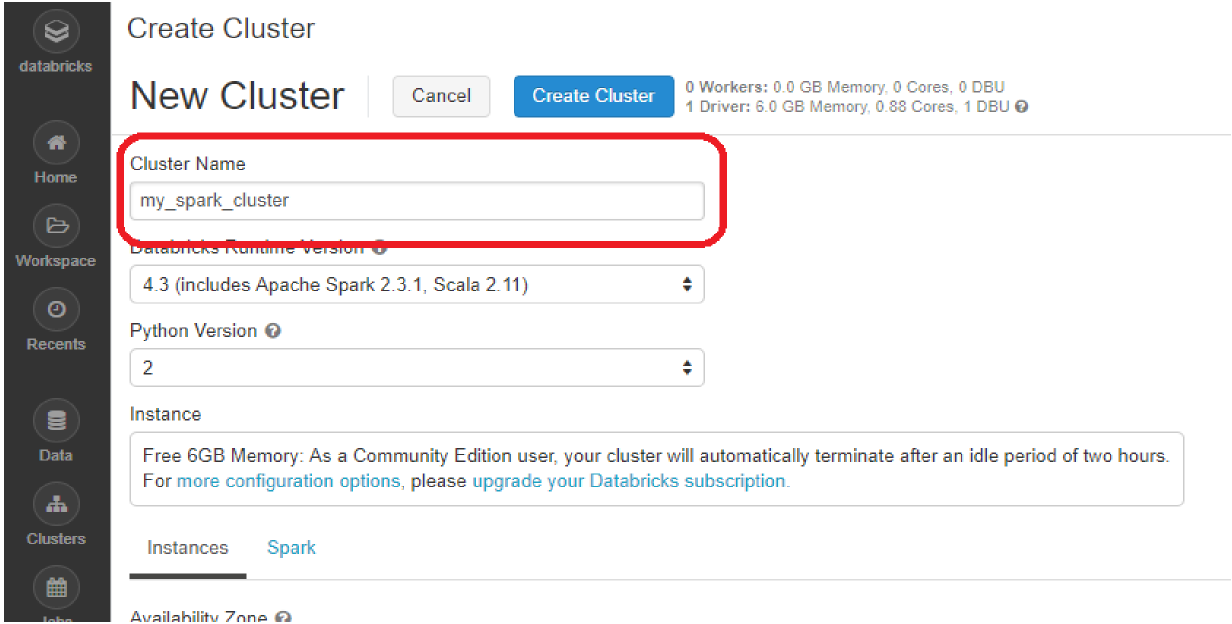
Step 4: Cont..



https://community.cloud.databricks.com/?o=572219029079

Step 5:

Enter cluster name



Create Cluster

New 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

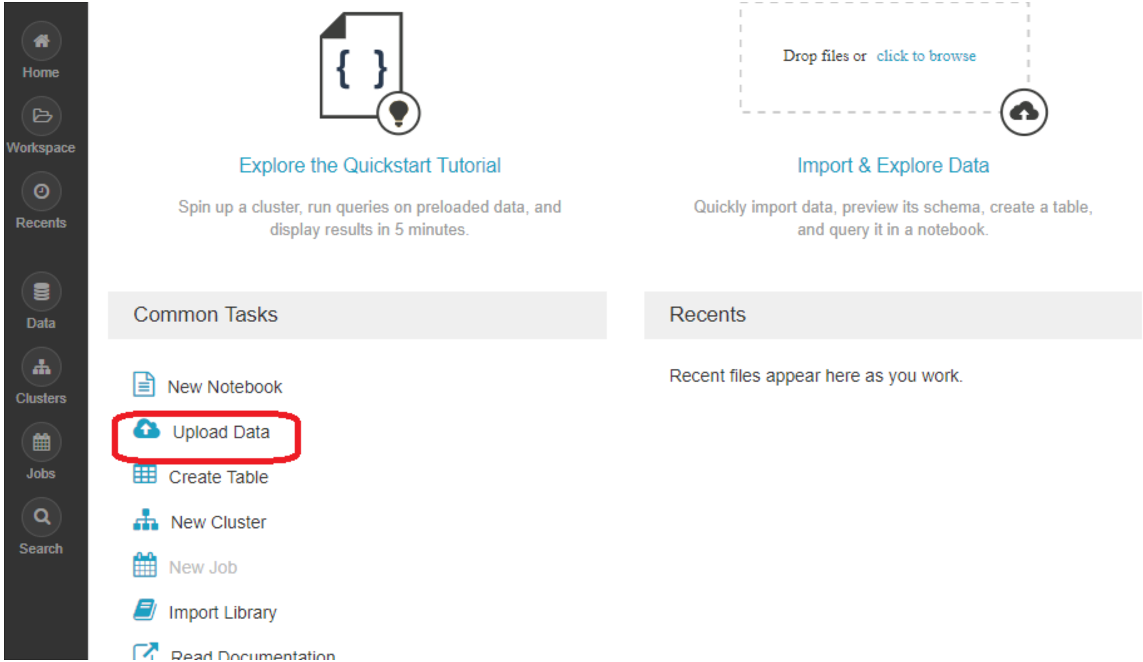
Instance

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.

Availability Zone

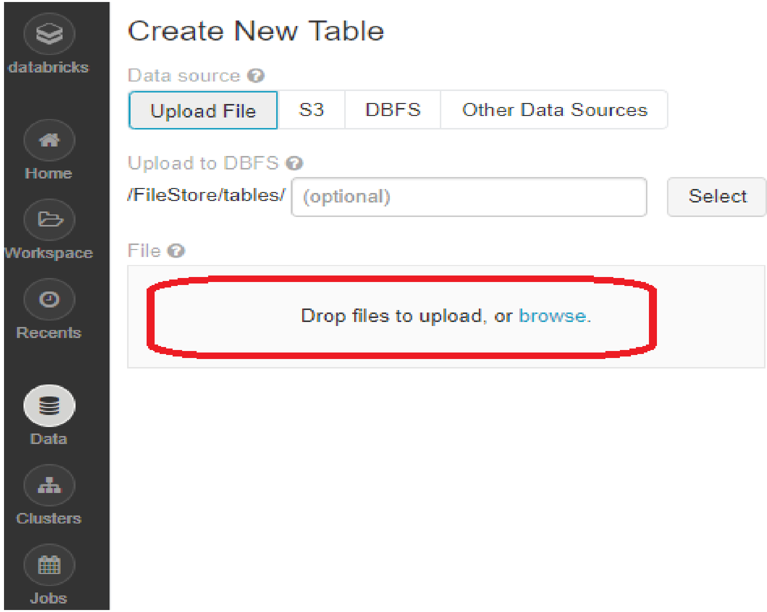
Step 6:

Upload data from your laptop to the cluster.


Home
Workspace
Recents
Data
Clusters
Jobs
Search
Drop files or click to browse
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

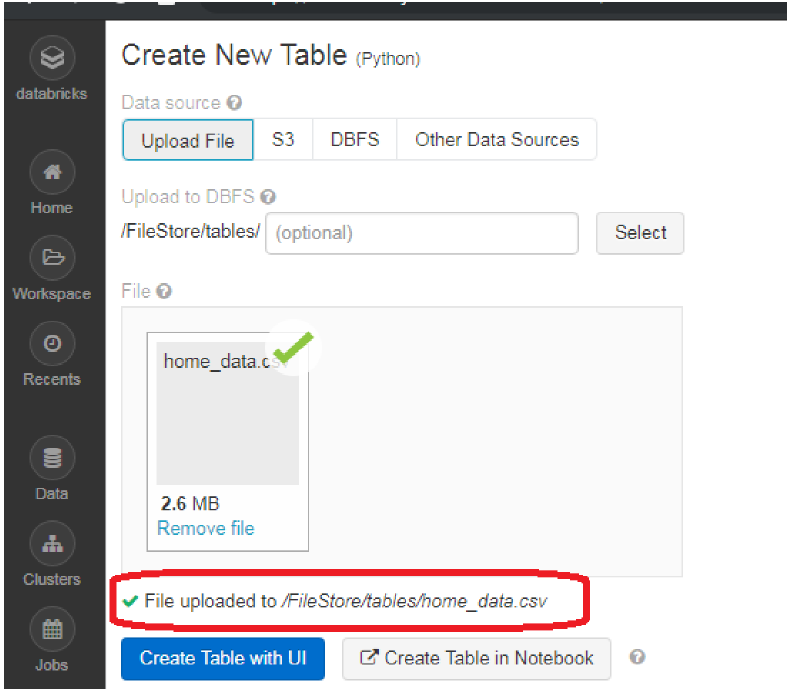
/FileStore/tables/ (optional) Select

File

Drop files to upload, or browse.

Step 7:

Copy the path.



Create New Table (Python)

Data source

Upload File | S3 | DBFS | Other Data Sources

Upload to DBFS

/FileStore/tables/ (optional) Select

File

home_data.csv

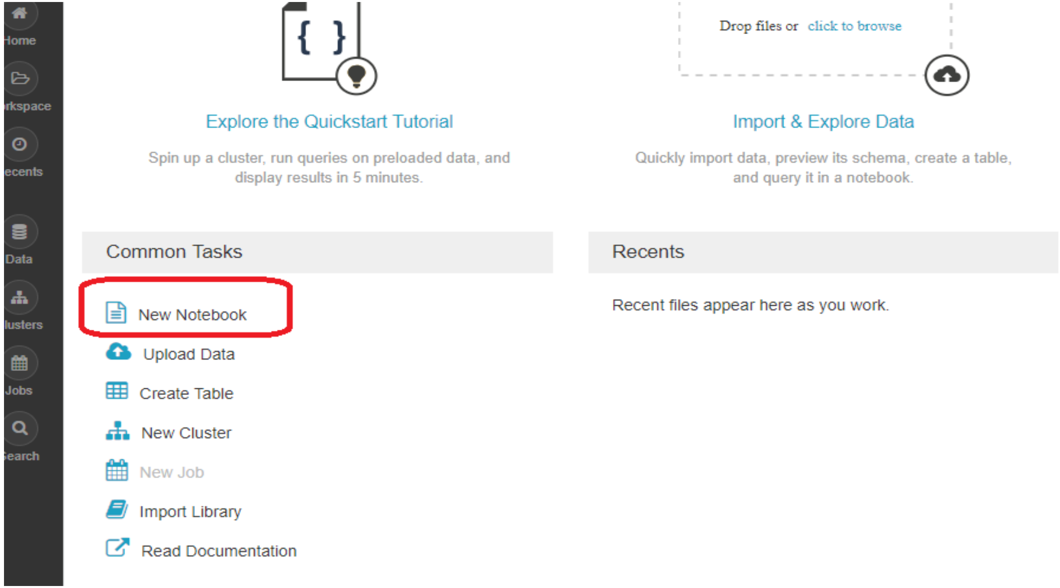
2.6 MB

Remove file

✓ File uploaded to */FileStore/tables/home_data.csv*

Step 8:

Create new notebook



Drop files or click to browse

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

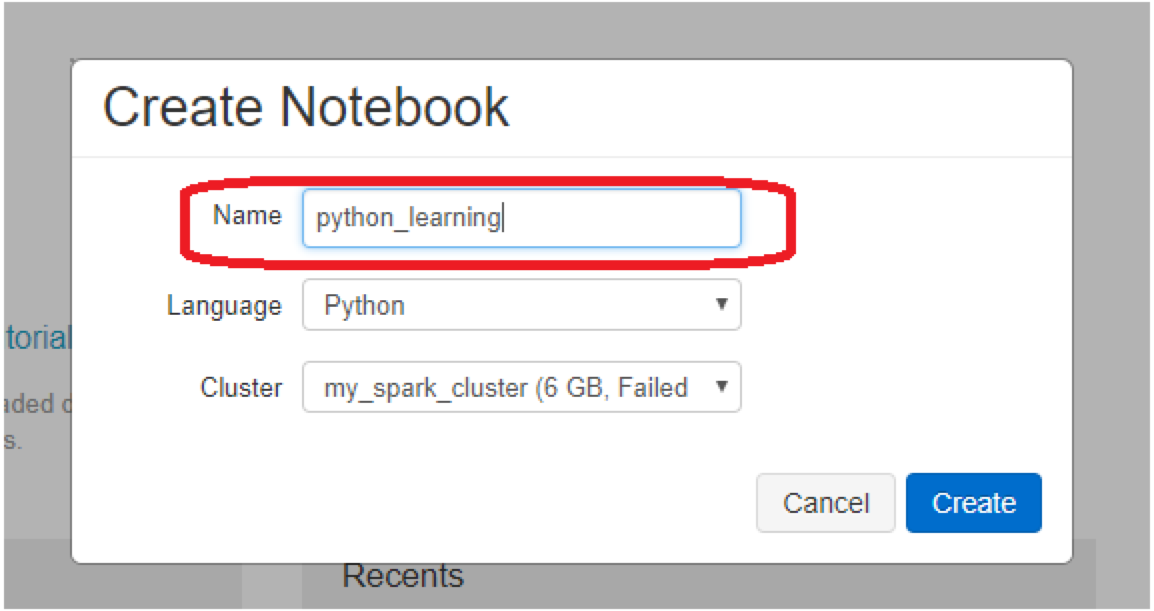
- New Notebook
- Upload Data
- Create Table
- New Cluster
- New Job
- Import Library
- Read Documentation

Recents

Recent files appear here as you work.

Step 9:

Provide the name of the notebook and chose the language.



Create Notebook

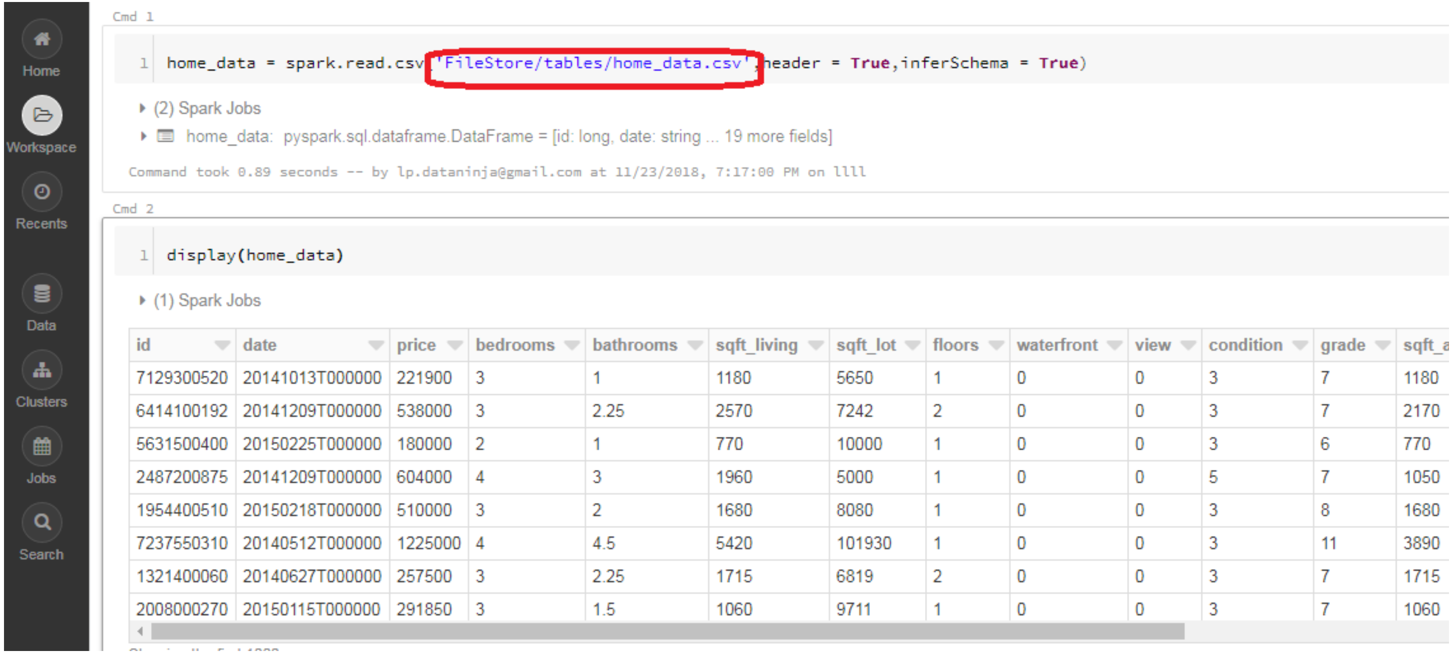
Name: python_learning

Language: Python

Cluster: my_spark_cluster (6 GB, Failed

Step 10:

Read the data in spark



Cmd 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 llll
```

Cmd 2

```
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

https://databricks.com/sparkaisummit/north-america/sessions?eventName=Summit%202018

https://databricks.com/sparkaisummit/north-america/sessions

https://www.youtube.com/user/TheApacheSpark/playlists