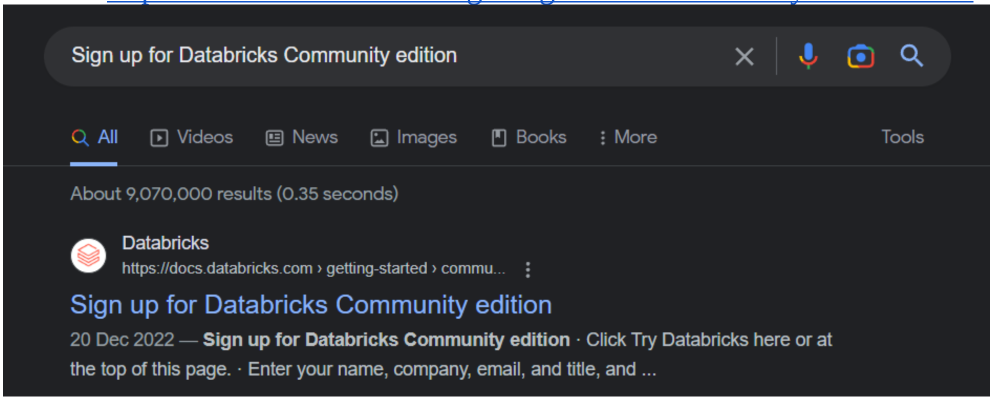
1 Databricks Guide

1.1 Databricks Community Edition Registration

To get started, we would be using the free version, databricks community edition and you would have to register for it.

1) Google "Sign up for Databricks Community edition" and click on the first link or click https://docs.databricks.com/getting-started/community-edition.html



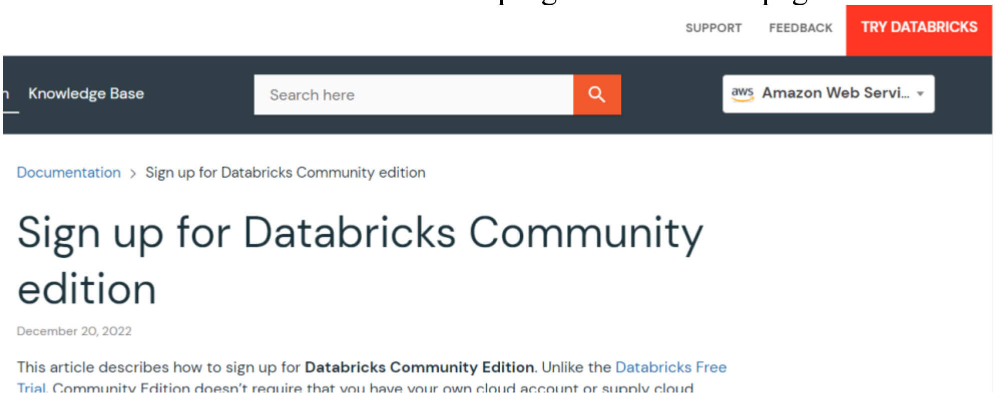
About 9,070,000 results (0.35 seconds)

Databricks
https://docs.databricks.com › getting-started › commu...

Sign up for Databricks Community edition

20 Dec 2022 — **Sign up for Databricks Community edition** · Click Try Databricks here or at the top of this page. · Enter your name, company, email, and title, and ...

2) click "TRY DATABRICKS" on the top right corner of the page



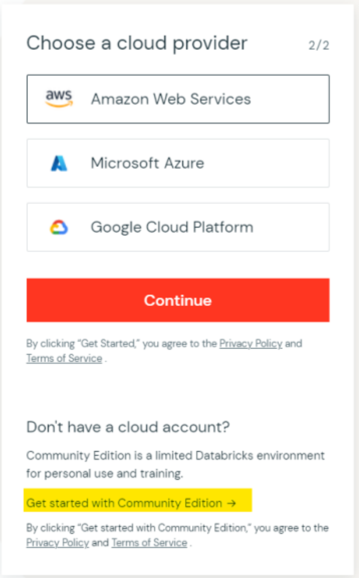
Documentation › Sign up for Databricks Community edition

Sign up for Databricks Community edition

December 20, 2022

This article describes how to sign up for **Databricks Community Edition**. Unlike the Databricks Free Trial, Community Edition doesn't require that you have your own cloud account or supply cloud

3) Fill in the necessary details and click continue to create an account
4) You will reach the page below and since we are using the Community Edition, please click on the portion highlighted below



Choose a cloud provider 2/2

Amazon Web Services

Microsoft Azure

Google Cloud Platform

Continue

By clicking "Get Started," you agree to the Privacy Policy and Terms of Service.

Don't have a cloud account?

Community Edition is a limited Databricks environment for personal use and training.

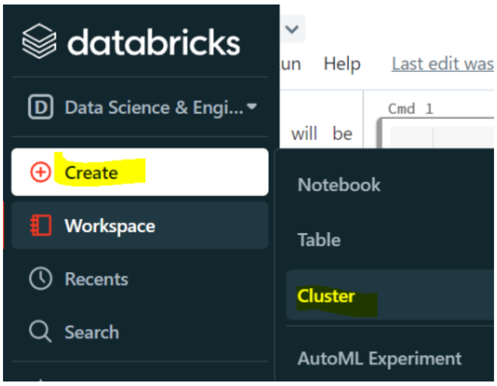
Get started with Community Edition →

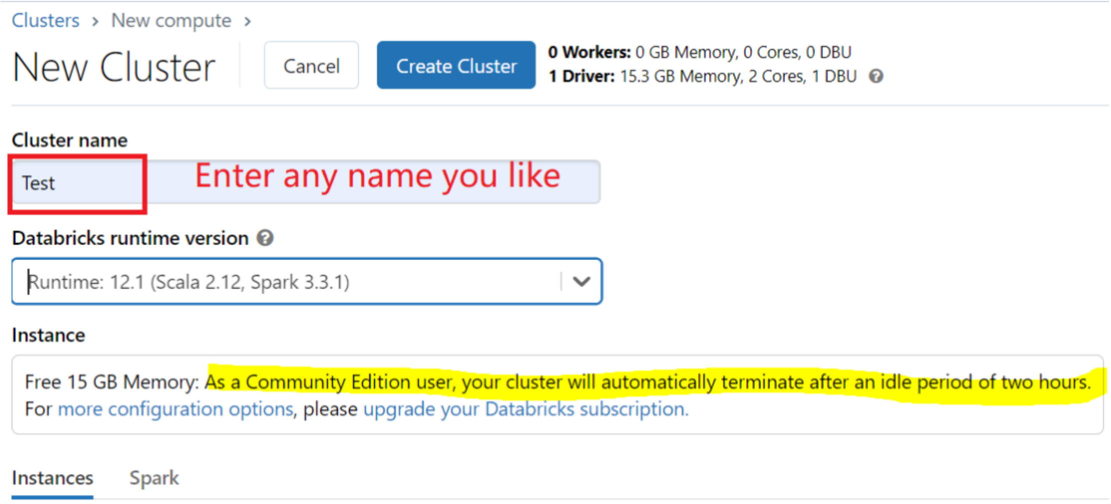
By clicking "Get started with Community Edition," you agree to the Privacy Policy and Terms of Service.

5) Validate your email address and login to databricks

1.2 Create a Cluster

You need to create a cluster to run your notebooks.





Clusters › New compute ›

New Cluster

Cancel

Create Cluster

0 Workers: 0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU

Cluster name

Test

Enter any name you like

Databricks runtime version

Runtime: 12.1 (Scala 2.12, Spark 3.3.1)

Instance

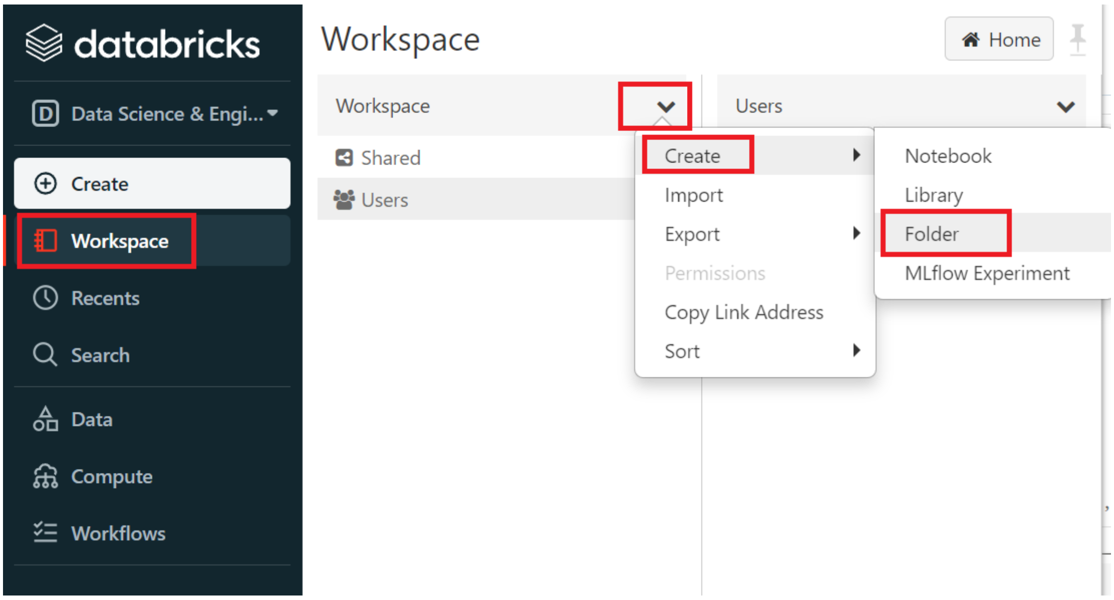
Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours.
For more configuration options, please upgrade your Databricks subscription.

Instances Spark

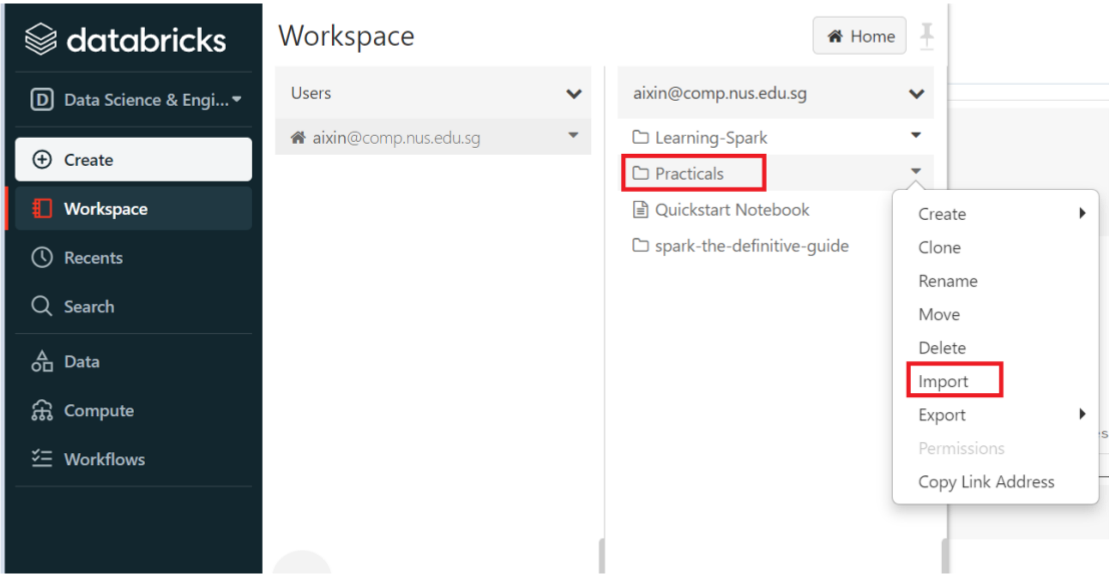
Note: the created cluster will automatically terminate after an idle period of two hours for the Community Edition. Hence you would need to re-create a new cluster to run your code if your cluster has terminated.

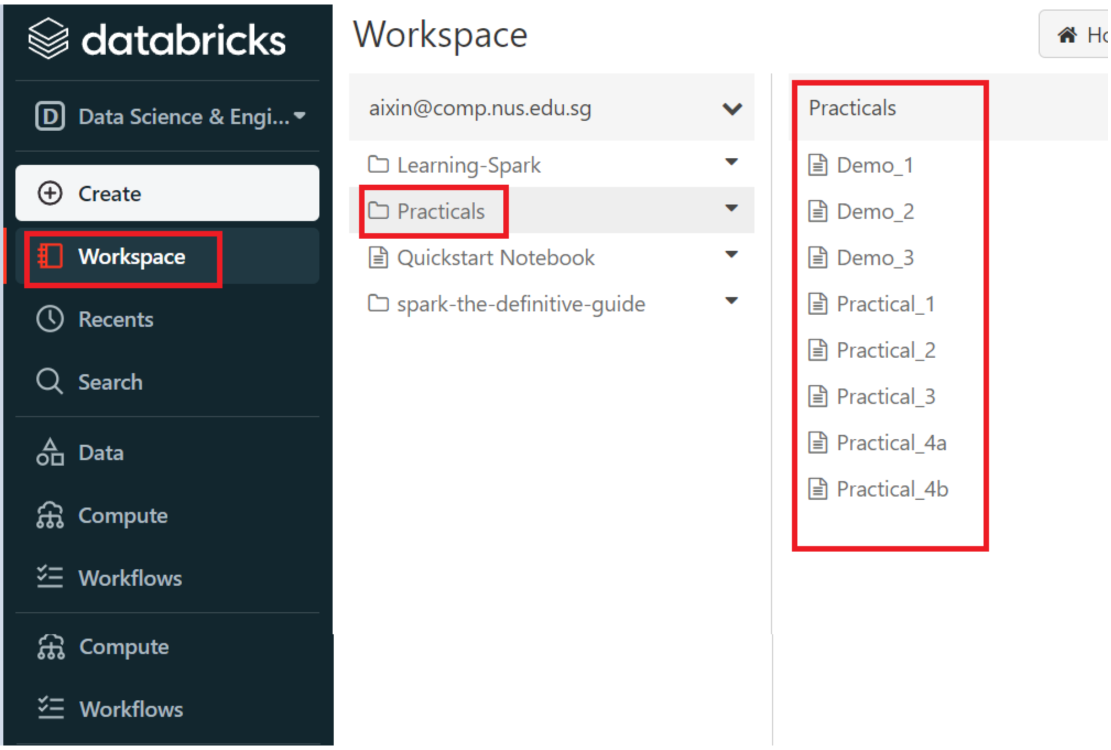
1.3 Upload Notebooks

First, create a folder, e.g. Practicals (or any other name you like).

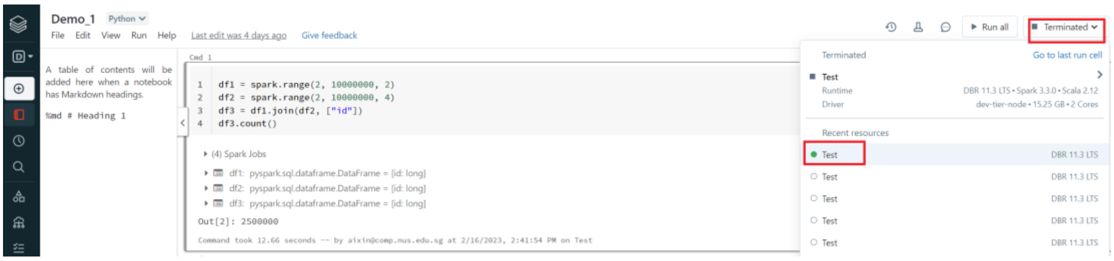


Then import all the notebooks into this folder.





Open Demo_1 and attach the notebook to your active cluster, e.g. Test.



A table of contents will be added here when a notebook has Markdown headings.

```
%md # Heading 1
```

```
df1 = spark.range(2, 10000000, 2)
df2 = spark.range(2, 10000000, 4)
df3 = df1.join(df2, ["id"])
df3.count()
```

```
▸ (4) Spark Jobs
▸ df1: pyspark.sql.dataframe.DataFrame = [id: long]
▸ df2: pyspark.sql.dataframe.DataFrame = [id: long]
▸ df3: pyspark.sql.dataframe.DataFrame = [id: long]
Out[2]: 2500000
Command took 12.66 seconds -- by aixin@comp.nus.edu.sg at 2/16/2023, 2:41:54 PM on Test
```

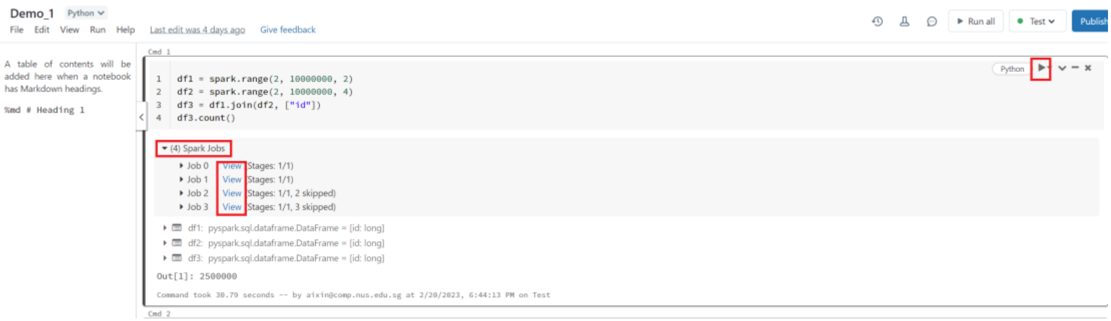
Terminated

■ Test	
Runtime	DBR 11.3 LTS • Spark 3.3.0 • Scala 2.12
Driver	dev-tier-node • 15.25 GB • 2 Cores

Recent resources

● Test	DBR 11.3 LTS
○ Test	DBR 11.3 LTS
○ Test	DBR 11.3 LTS
○ Test	DBR 11.3 LTS
○ Test	DBR 11.3 LTS

Now you can start run the notebook cell by cell. You can also access Spark UI for each Job by click “View”.

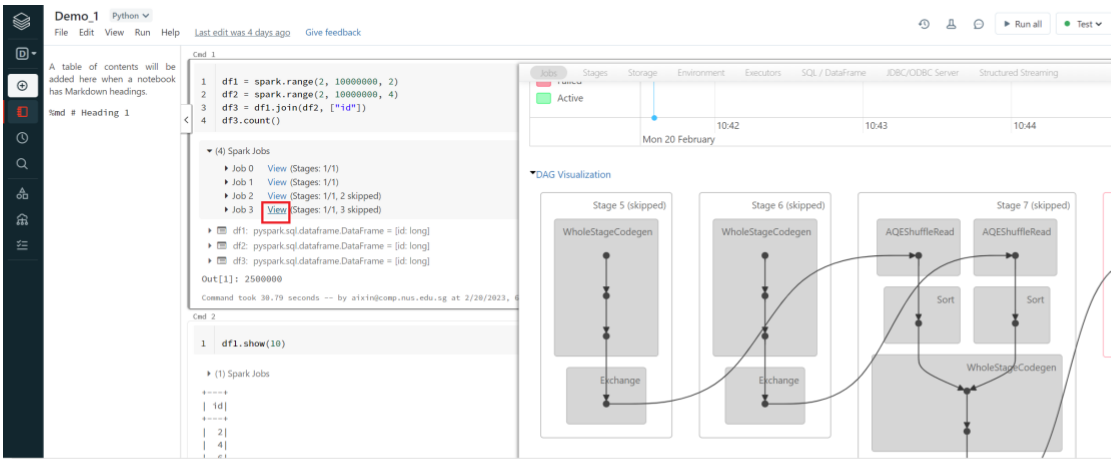


A table of contents will be added here when a notebook has Markdown headings.

```
%md # Heading 1
```

```
df1 = spark.range(2, 10000000, 2)
df2 = spark.range(2, 10000000, 4)
df3 = df1.join(df2, ["id"])
df3.count()
```

```
▾ (4) Spark Jobs
    ▸ Job 0  View (Stages: 1/1)
    ▸ Job 1  View (Stages: 1/1)
    ▸ Job 2  View (Stages: 1/1, 2 skipped)
    ▸ Job 3  View (Stages: 1/1, 3 skipped)
▸ df1: pyspark.sql.dataframe.DataFrame = [id: long]
▸ df2: pyspark.sql.dataframe.DataFrame = [id: long]
▸ df3: pyspark.sql.dataframe.DataFrame = [id: long]
Out[1]: 2500000
Command took 30.79 seconds -- by aixin@comp.nus.edu.sg at 2/20/2023, 6:44:13 PM on Test
```



A table of contents will be added here when a notebook has Markdown headings.

```
%md # Heading 1
```

Cmd 1

```python
df1 = spark.range(2, 10000000, 2)
df2 = spark.range(2, 10000000, 4)
df3 = df1.join(df2, ["id"])
df3.count()
```

(4) Spark Jobs

- Job 0 View (Stages: 1/1)
- Job 1 View (Stages: 1/1)
- Job 2 View (Stages: 1/1, 2 skipped)
- Job 3 View (Stages: 1/1, 3 skipped)

```
df1: pyspark.sql.dataframe.DataFrame = [id: long]
df2: pyspark.sql.dataframe.DataFrame = [id: long]
df3: pyspark.sql.dataframe.DataFrame = [id: long]
Out[1]: 2500000
Command took 38.79 seconds -- by aixin@comp.nus.edu.sg at 2/20/2023, 6
```

Cmd 2

```python
df1.show(10)
```

(1) Spark Jobs

```
+---+
| id|
+---+
|  2|
|  4|
```