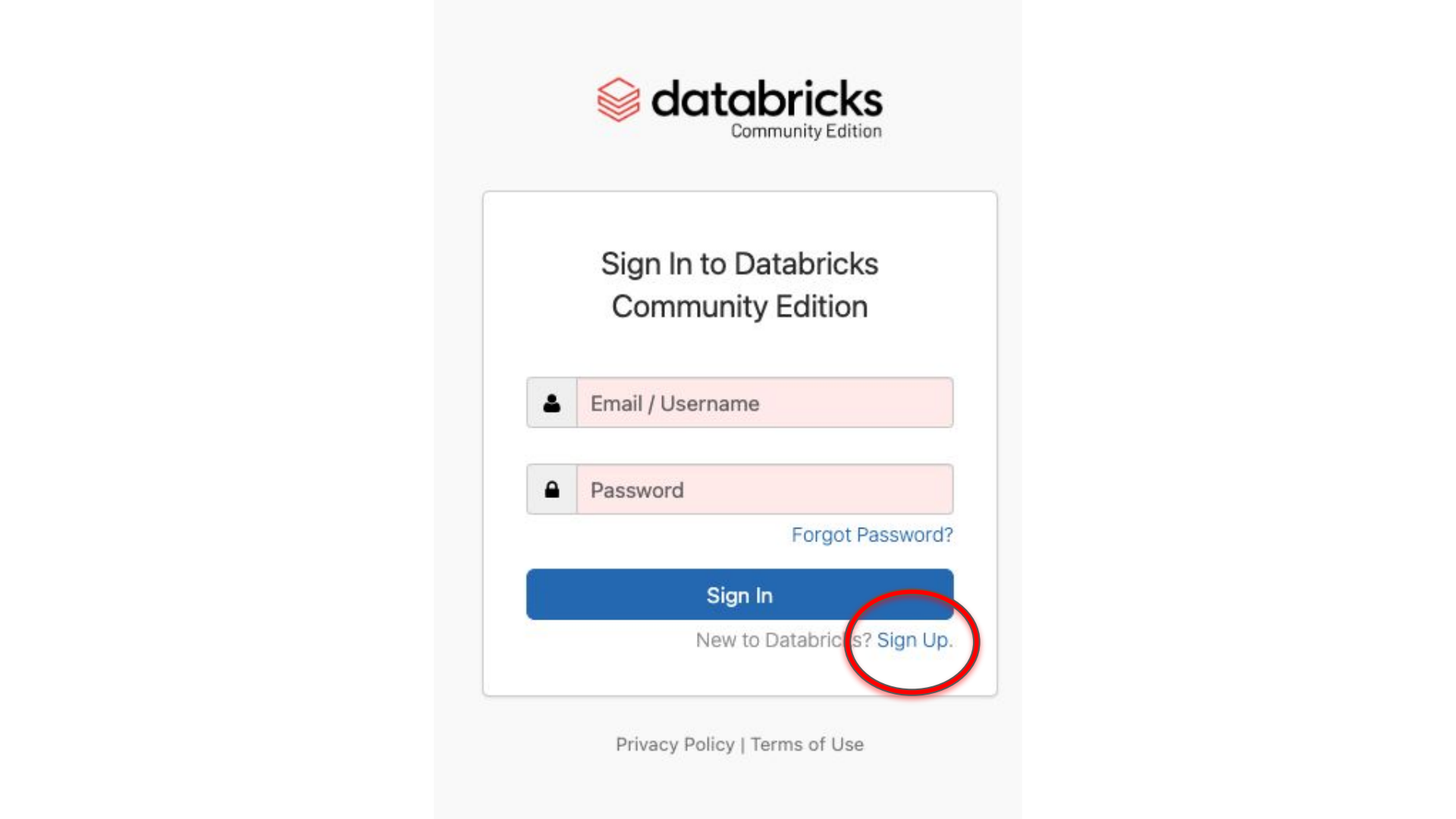
So ya wanna try Spark

Databricks

Databricks is an American enterprise software company founded by the creators of Apache Spark.[2] Databricks develops a web-based platform for working with Spark, that provides automated cluster management and IPython-style notebooks.

Step 1: Create an account (5 minutes)

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


databricks
Community Edition
Sign In to Databricks
Community Edition
Email / Username
Password
Forgot Password?
Sign In
New to Databricks? Sign Up.
Privacy Policy | Terms of Use

Try Databricks for free

An open and unified data analytics platform for data engineering, data science, machine learning, and analytics. From the original creators of Apache Spark™, Delta lake, MLflow, and Koalas.

aws Microsoft Azure Google Cloud

Databricks trial:

- Collaborative environment for data teams to build solutions together.
- Interactive notebooks to use Apache Spark™, SQL, Python, Scala, Delta Lake, MLflow, TensorFlow, Keras, Scikit-learn and more.
- Available as a 14-day full trial in your own cloud, or as a lightweight trial hosted by Databricks.

Used by:

Virgin hyperloop edmunds nielsen

RIOT GAMES REGENERON Grab

Please tell us about yourself

First Name: *

Chase

Last Name: *

Westhoff

Company *

Brigham Young University

Company Email *

chasemwesthoff@gmail.com

Title *

Student

Phone Number

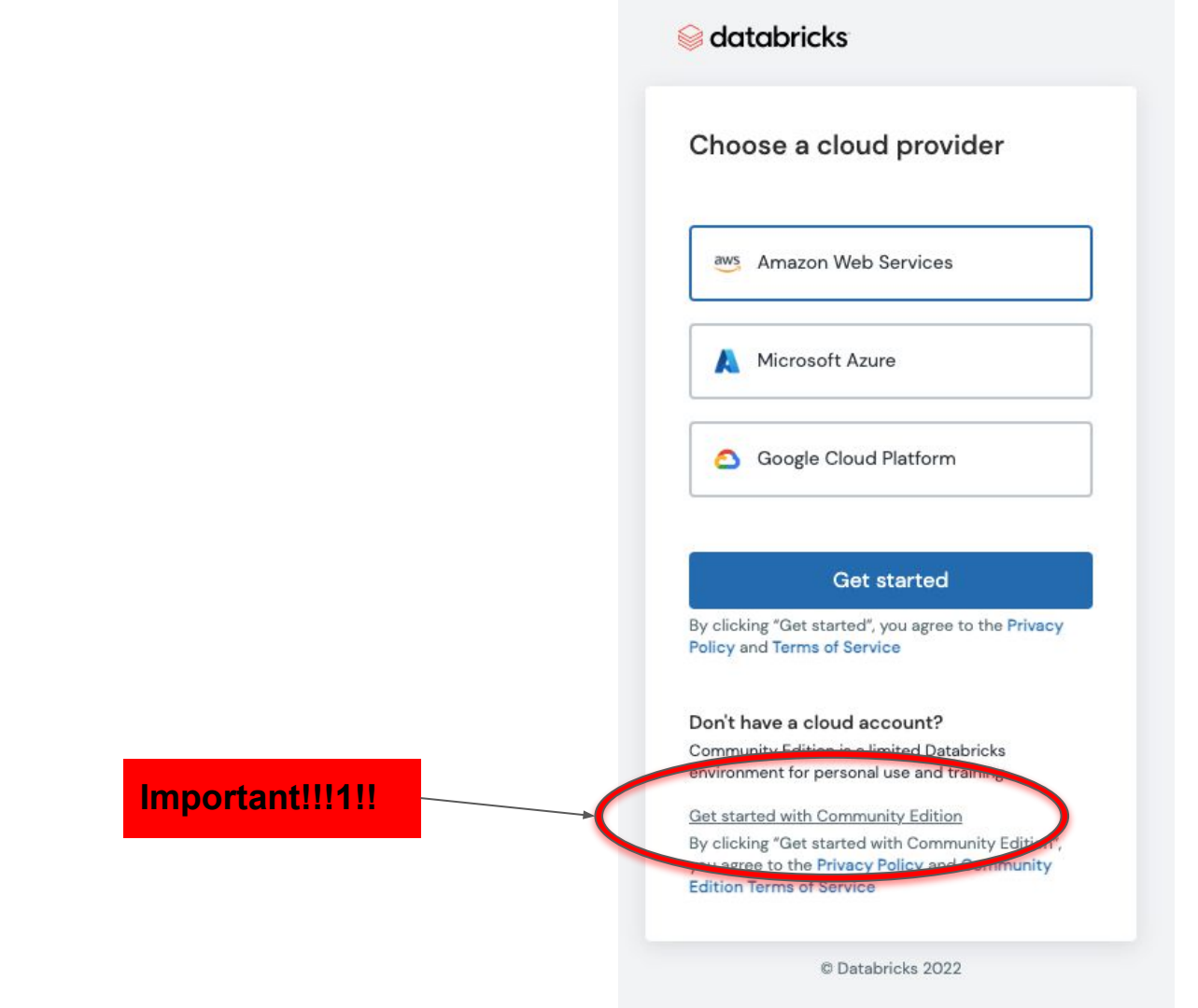
8013176655

Country: *

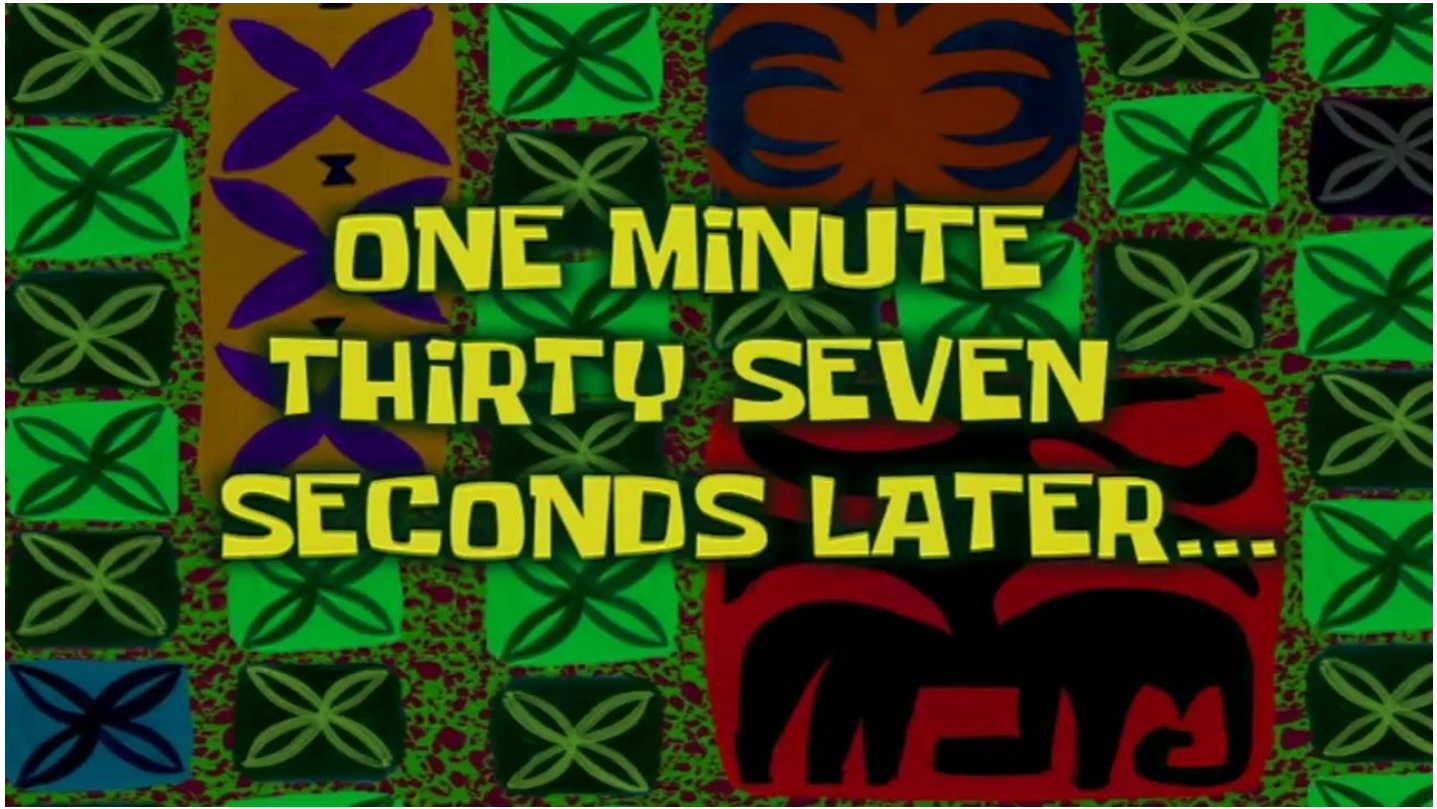
United States

By submitting, I agree to the processing of my personal data by Databricks in accordance with our Privacy Policy. I understand I can update my preferences at any time.

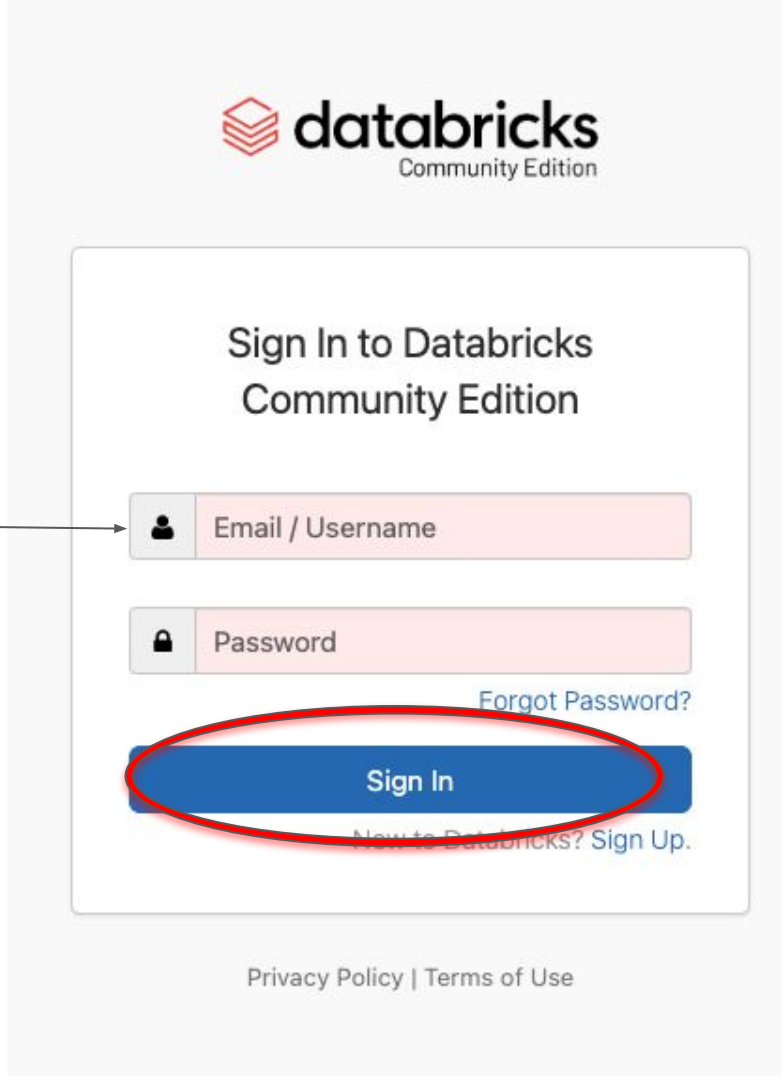
GET STARTED FOR FREE

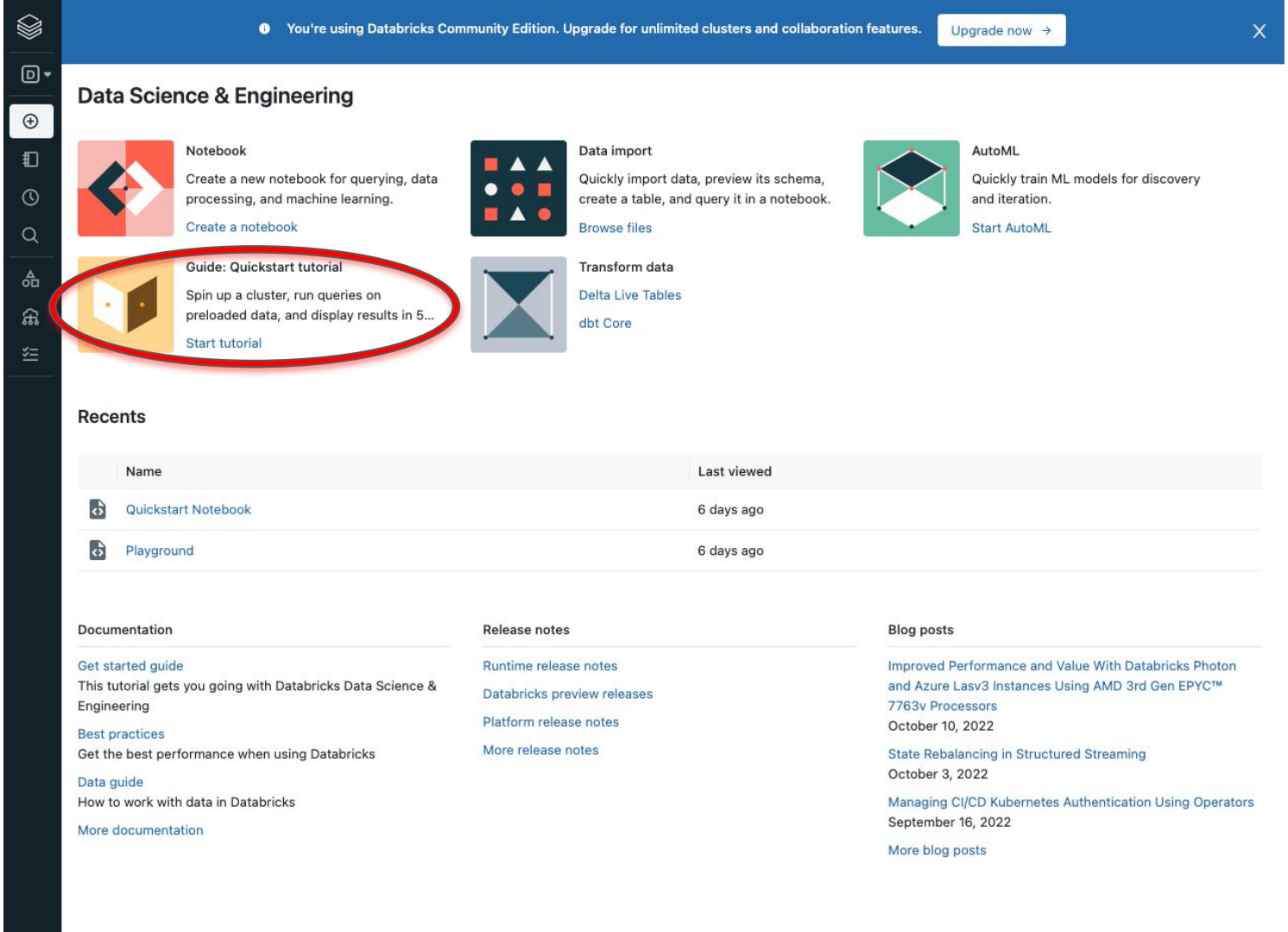

databricks
Choose a cloud provider
Amazon Web Services
Microsoft Azure
Google Cloud Platform
Get started
By clicking "Get started", you agree to the Privacy Policy and Terms of Service
Don't have a cloud account?
Get started with Community Edition
Edition Terms of Service
© Databricks 2022
Important!!!1!!

Email validation stuff…

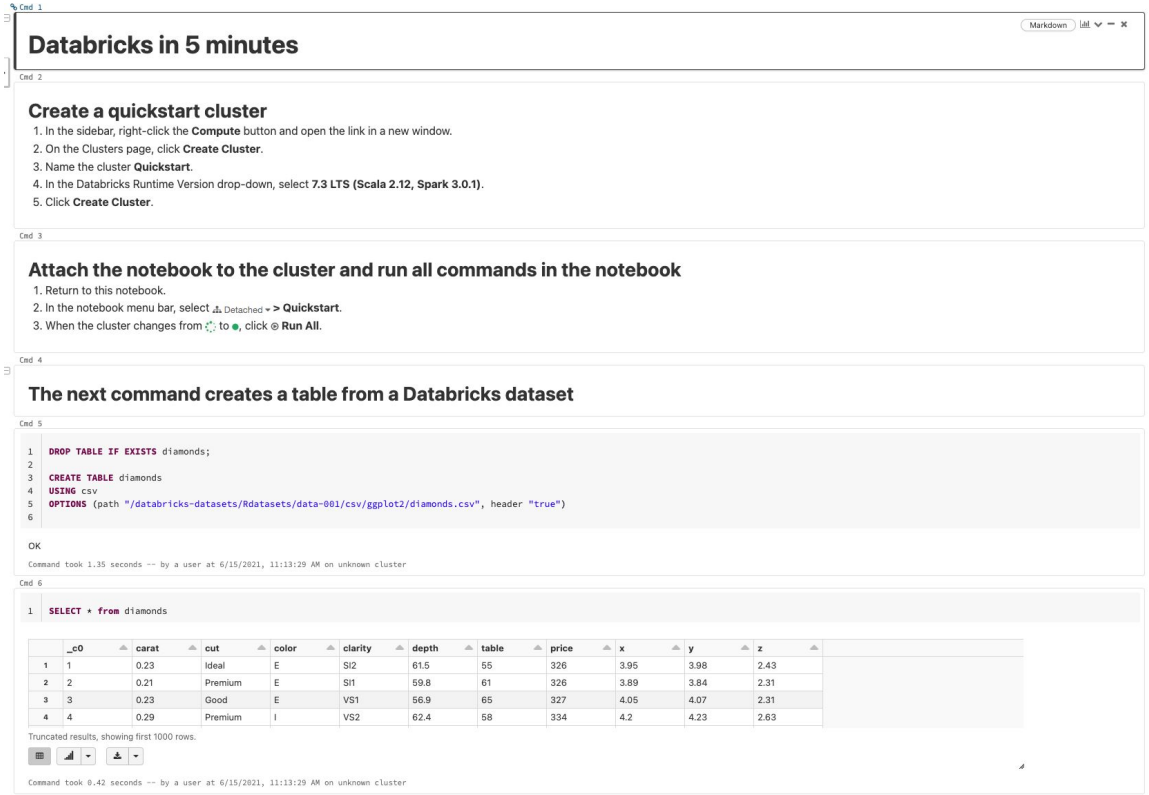


Email is case sensitive for some reason??




You're using Databricks Community Edition. Upgrade for unlimited clusters and collaboration features.
Upgrade now →
Data Science & Engineering
Notebook
Create a new notebook for querying, data processing, and machine learning.
Create a notebook
Data import
Quickly import data, preview its schema, create a table, and query it in a notebook.
Browse files
AutoML
Quickly train ML models for discovery and iteration.
Start AutoML
Guide: Quickstart tutorial
Spin up a cluster, run queries on preloaded data, and display results in 5...
Start tutorial
Transform data
Delta Live Tables
dbt Core
Recents
Name
Last viewed
Quickstart Notebook
6 days ago
Playground
6 days ago
Documentation
Get started guide
This tutorial gets you going with Databricks Data Science & Engineering
Best practices
Get the best performance when using Databricks
Data guide
How to work with data in Databricks
More documentation
Release notes
Runtime release notes
Databricks preview releases
Platform release notes
More release notes
Blog posts
Improved Performance and Value With Databricks Photon and Azure Lasv3 Instances Using AMD 3rd Gen EPYC™ 7763v Processors
October 10, 2022
State Rebalancing in Structured Streaming
October 3, 2022
Managing CI/CD Kubernetes Authentication Using Operators
September 16, 2022
More blog posts

Go through the tutorial! -5 minutes (or more??)



Cmd 1

Markdown

Databricks in 5 minutes

Cmd 2

Create a quickstart cluster

1. In the sidebar, right-click the **Compute** button and open the link in a new window.
2. On the Clusters page, click **Create Cluster**.
3. Name the cluster **Quickstart**.
4. In the Databricks Runtime Version drop-down, select **7.3 LTS (Scala 2.12, Spark 3.0.1)**.
5. Click **Create Cluster**.

Cmd 3

Attach the notebook to the cluster and run all commands in the notebook

1. Return to this notebook.
2. In the notebook menu bar, select Detached > **Quickstart**.
3. When the cluster changes from to , click **Run All**.

Cmd 4

The next command creates a table from a Databricks dataset

Cmd 5

```
1  DROP TABLE IF EXISTS diamonds;
2
3  CREATE TABLE diamonds
4  USING csv
5  OPTIONS (path "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv", header "true")
6
```

OK

Command took 1.35 seconds -- by a user at 6/15/2021, 11:13:29 AM on unknown cluster

Cmd 6

```
1  SELECT * from diamonds
```

	_c0	carat	cut	color	clarity	depth	table	price	x	y	z
1	1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63

Truncated results, showing first 1000 rows.

Command took 0.42 seconds -- by a user at 6/15/2021, 11:13:29 AM on unknown cluster

Problem Set: Video Game Sales

1. Read in the data from a .csv file

https://www.kaggle.com/datasets/gregorut/videogamesales

2. Write code to find the best selling Xbox 360 Game in Japan

3. Write code to find the console that sold the most games overall.

4. What was the best year of sales for platforming games?

5. Find the total EU sales for each publisher, sort by highest to lowest

6. Find something interesting in the data and share it with the class!