ENSF 612

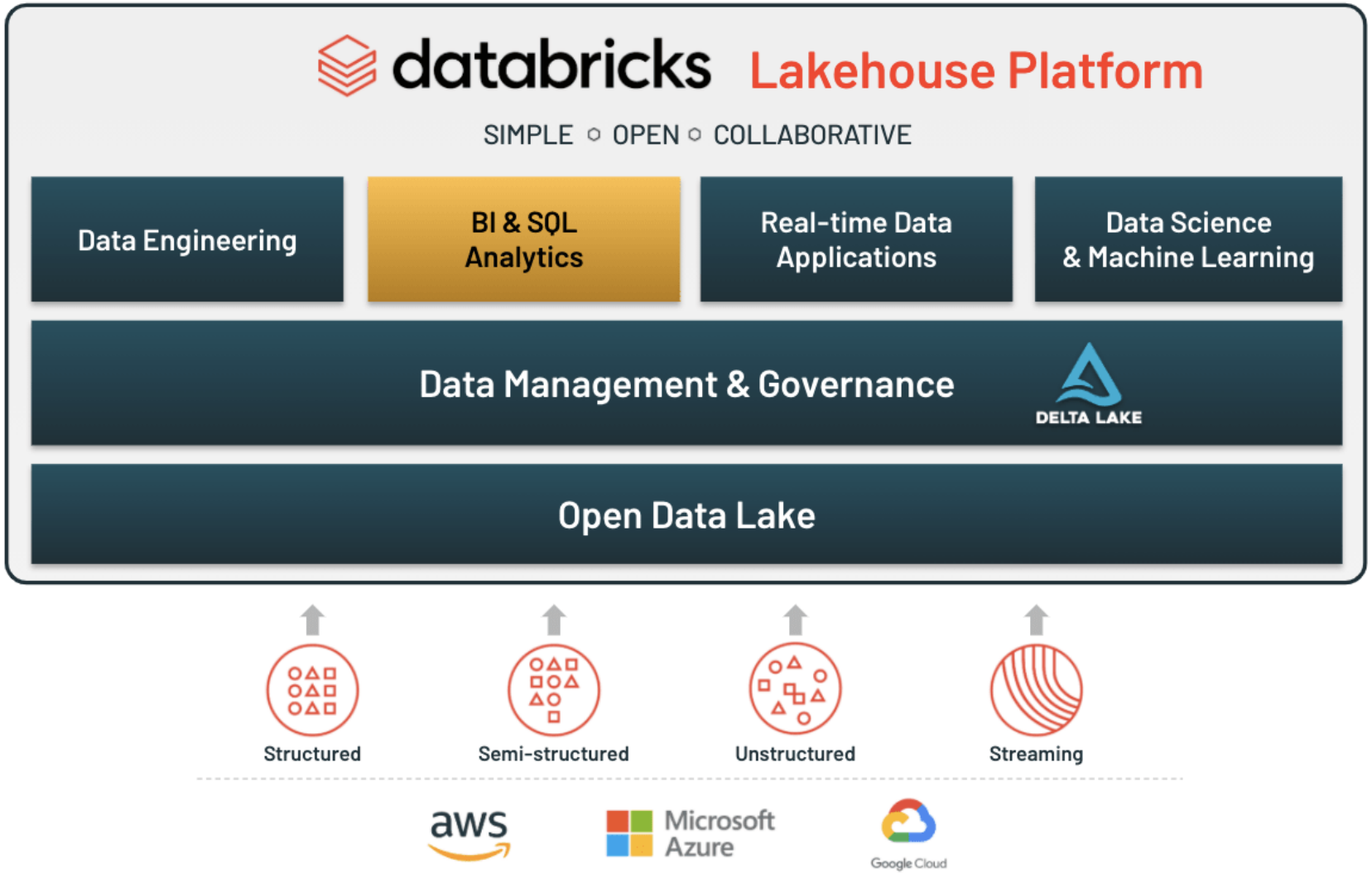
HOW TO USE DATABRICKS COMMUNITY EDITION

What is Databricks

- Databricks is a managed platform for running Apache Spark
  - You do not have to learn complex cluster management concepts.
  - No need to perform tedious maintenance tasks to take advantage of Spark.
- It's a point and click platform for those that prefer a user interface like data scientists or data analysts.

Databricks is a zero-management cloud platform that provides:

- Fully managed Spark clusters
- An interactive workspace for exploration and visualization
- A production focused pipeline scheduler
- A platform for powering Spark-based applications



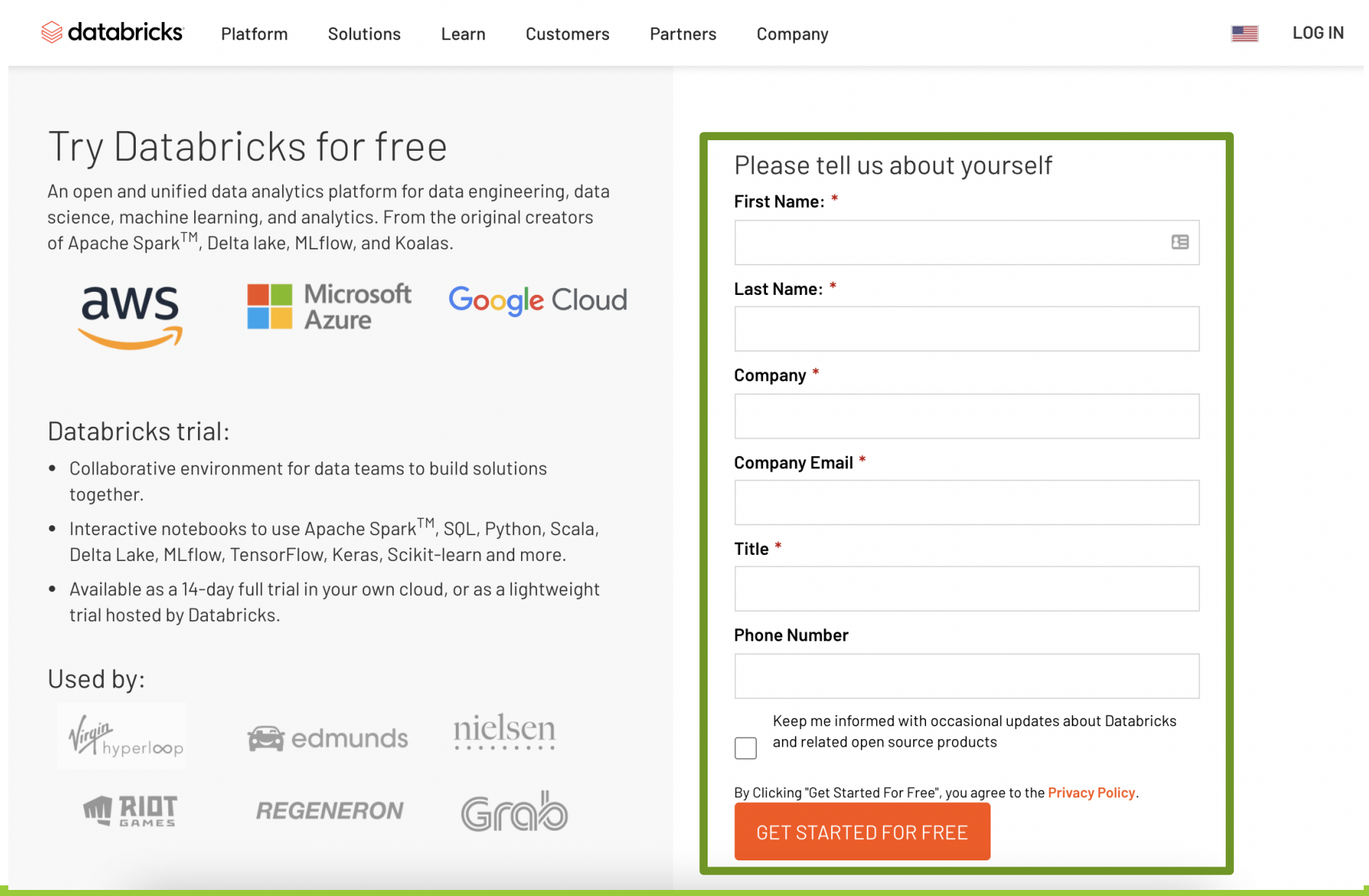
Source: https://databricks.com/blog/2020/11/12/announcing-the-launch-of-databricks-sql.html

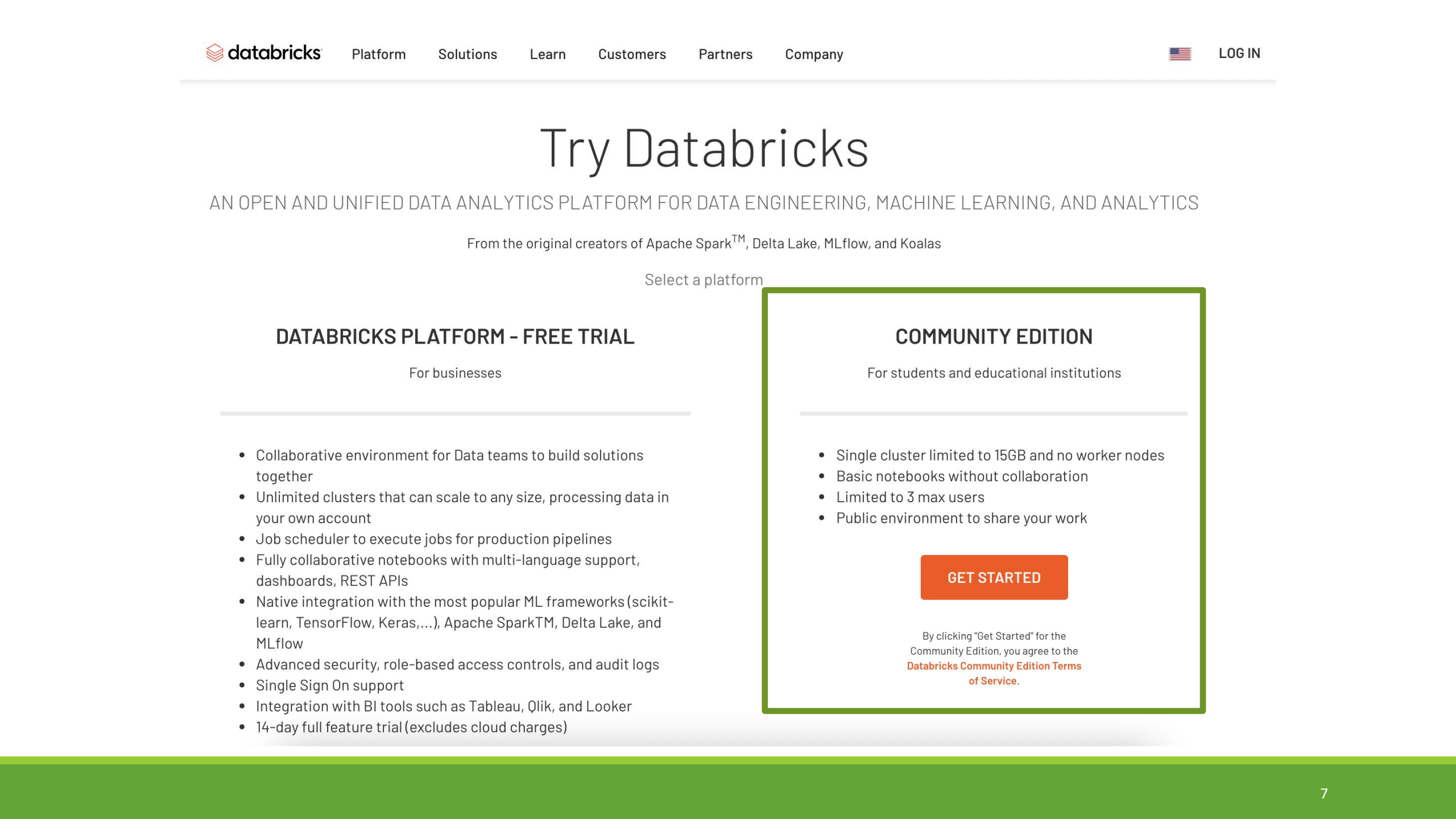
Databricks Terminology

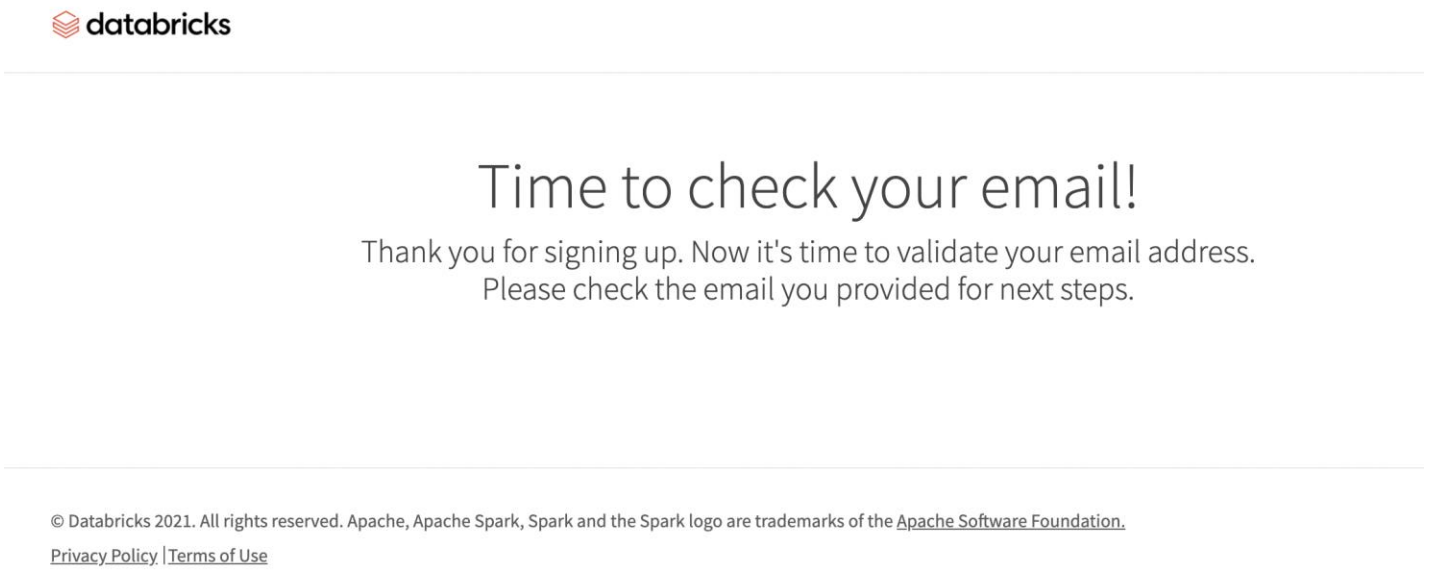
- Workspaces
  - Allow you to organize all the work on Databricks.
  - Like a folder structure in our computer.
  - It allows you to save notebooks and libraries and share them with other users.
- Clusters
  - Clusters are groups of computers that you treat as a single computer.
  - Clusters allow you to execute code from notebooks or libraries on set of data.
- Notebooks
  - Notebooks are a set of any number of cells that allow you to execute commands.
- Libraries
  - Libraries are packages or modules that provide additional functionality that you need to solve your business problems.
  - You can write and upload these manually or you may install them directly via package management utilities like pypi or maven.

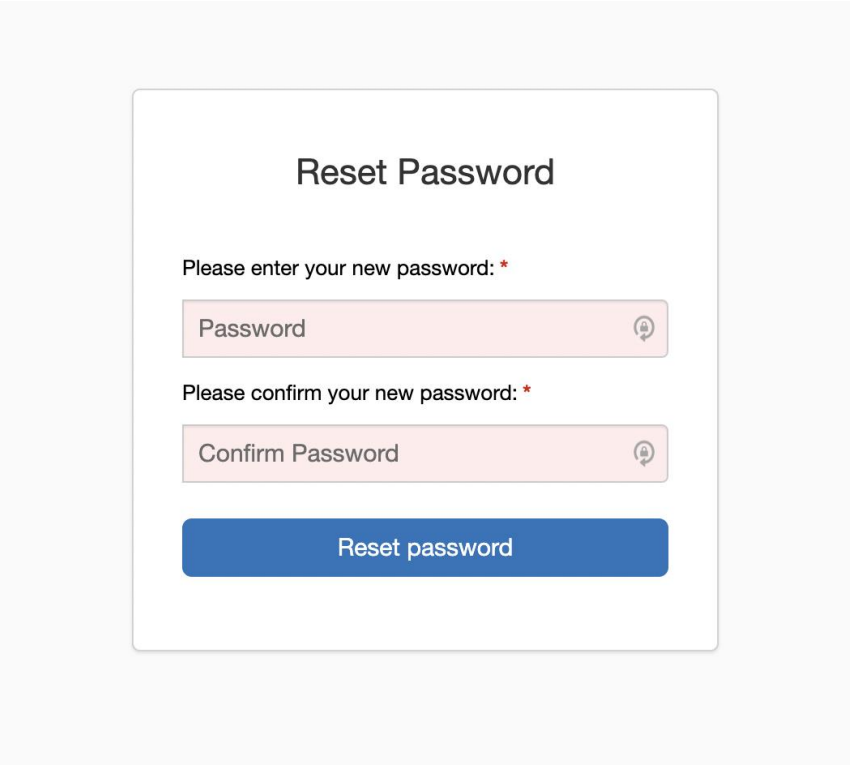
Let’s see a Demo!

Visit: https://databricks.com/try-databricks

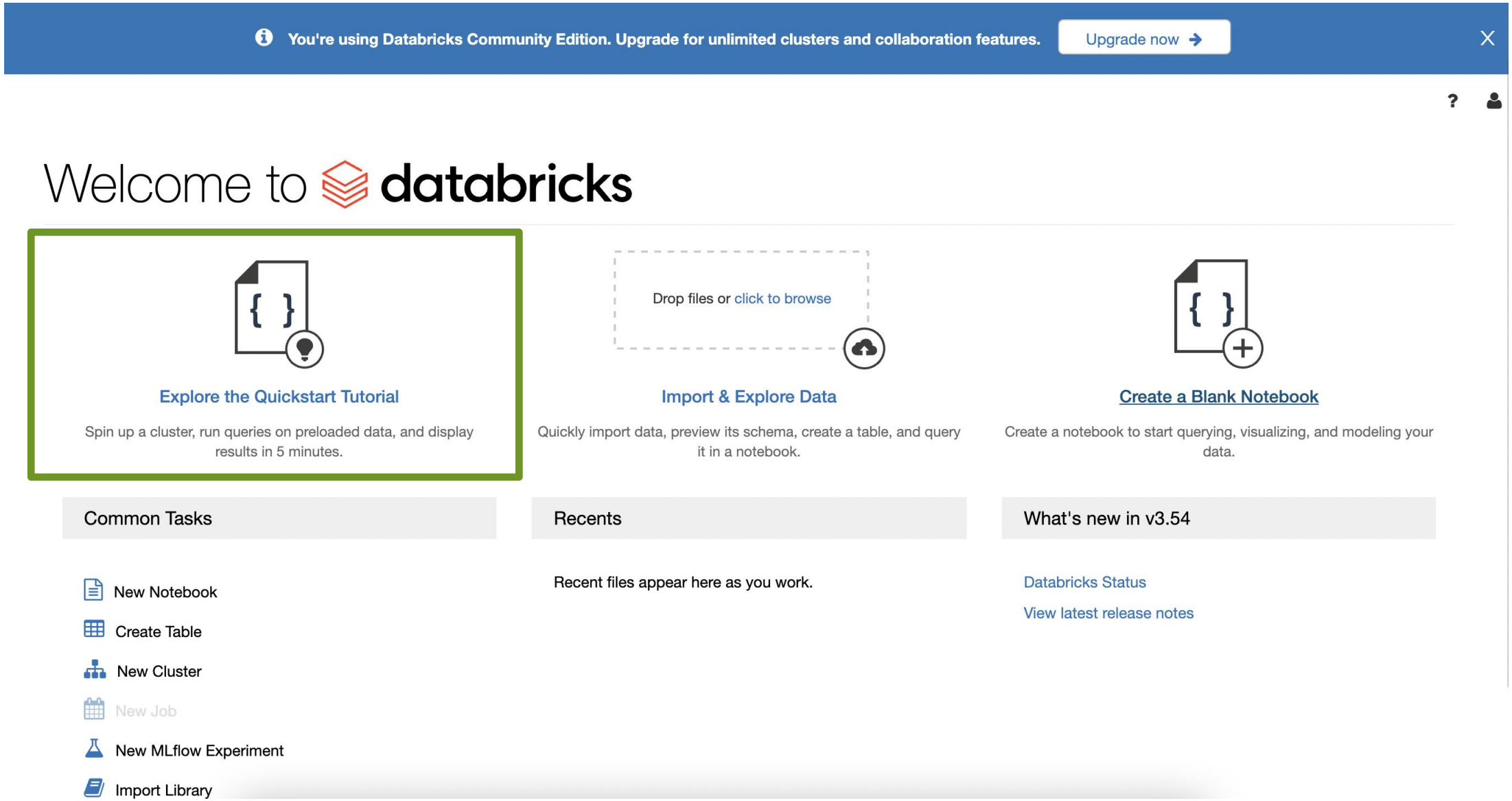


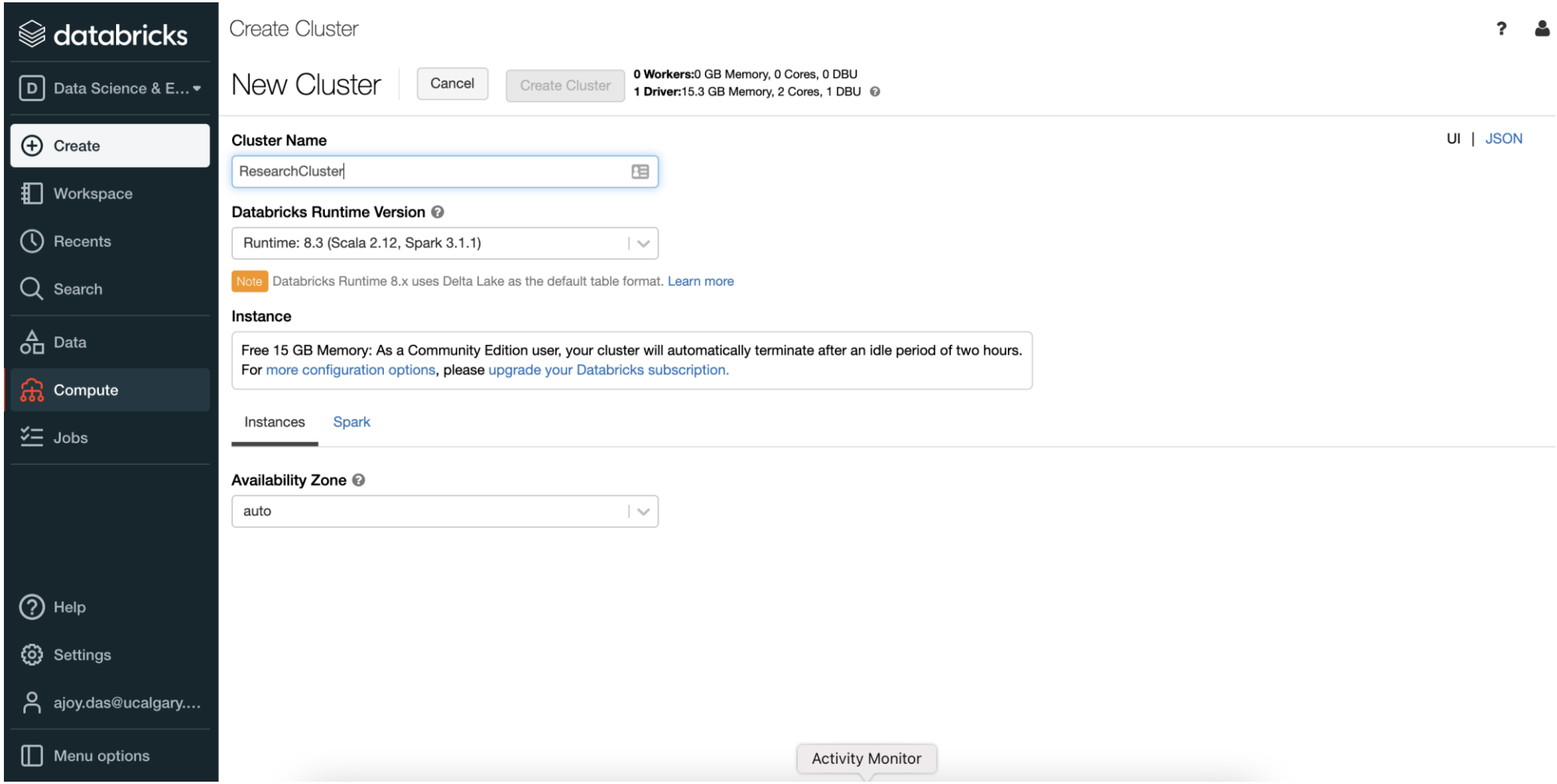

databricks
Platform
Solutions
Learn
Customers
Partners
Company
LOG IN
Try Databricks
AN OPEN AND UNIFIED DATA ANALYTICS PLATFORM FOR DATA ENGINEERING, MACHINE LEARNING, AND ANALYTICS
From the original creators of Apache Spark™, Delta Lake, MLflow, and Koalas
Select a platform
DATABRICKS PLATFORM - FREE TRIAL
For businesses
Collaborative environment for Data teams to build solutions together
Unlimited clusters that can scale to any size, processing data in your own account
Job scheduler to execute jobs for production pipelines
Fully collaborative notebooks with multi-language support, dashboards, REST APIs
Native integration with the most popular ML frameworks (scikit-learn, TensorFlow, Keras,...), Apache SparkTM, Delta Lake, and MLflow
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
Single cluster limited to 15GB and no worker nodes
Basic notebooks without collaboration
Limited to 3 max users
Public environment to share your work
GET STARTED
By clicking "Get Started" for the Community Edition, you agree to the Databricks Community Edition Terms of Service.

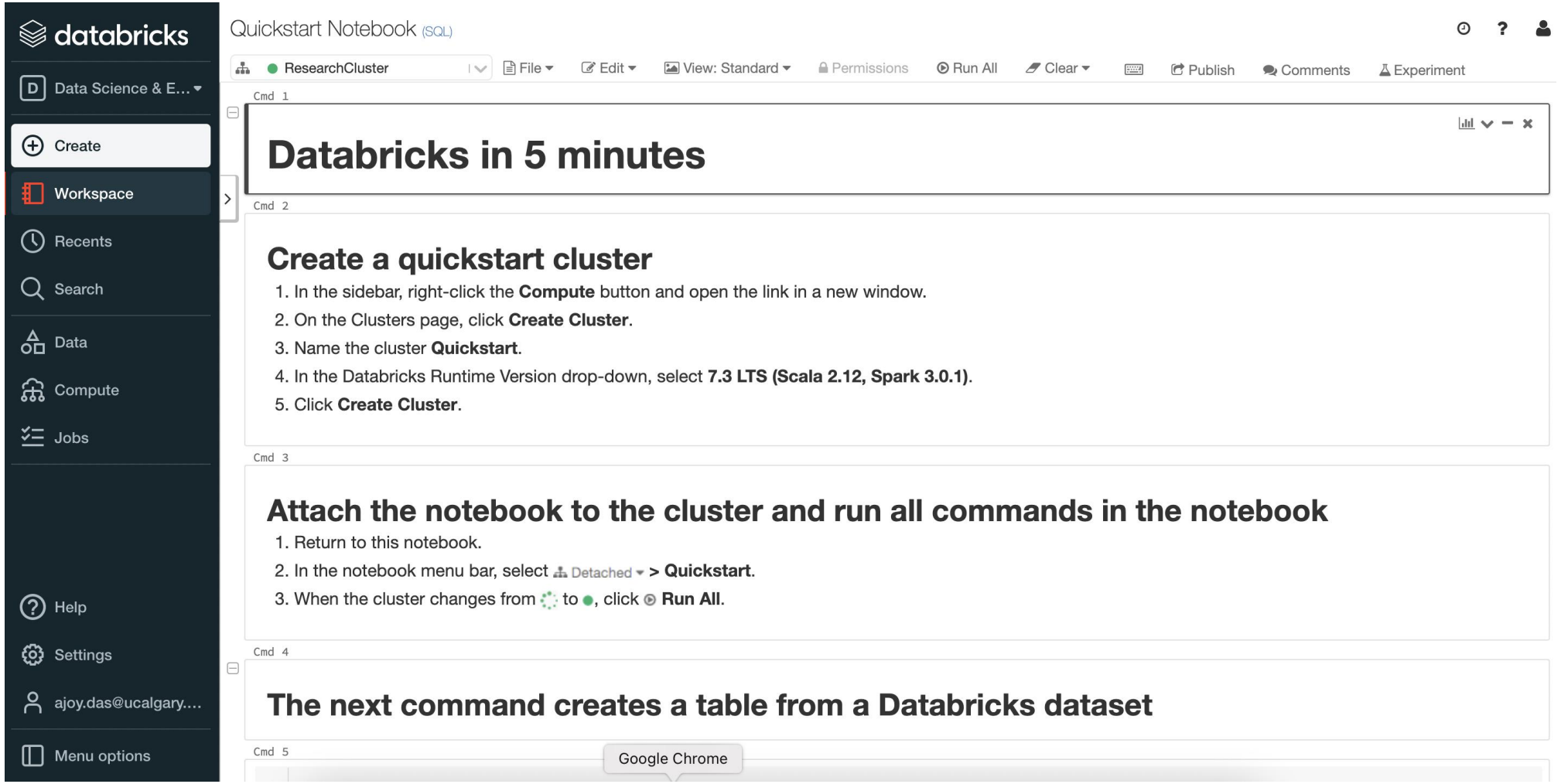




Visit: https://community.cloud.databricks.com/login.html


You're using Databricks Community Edition. Upgrade for unlimited clusters and collaboration features.
Upgrade now
Welcome to databricks
Explore the Quickstart Tutorial
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.
Drop files or click to browse
Import & Explore Data
Quickly import data, preview its schema, create a table, and query it in a notebook.
Create a Blank Notebook
Create a notebook to start querying, visualizing, and modeling your data.
Common Tasks
New Notebook
Create Table
New Cluster
New Job
New MLflow Experiment
Import Library
Recents
Recent files appear here as you work.
What's new in v3.54
Databricks Status
View latest release notes


databricks
Data Science & E...
Create
Workspace
Recents
Search
Data
Compute
Jobs
Help
Settings
ajoy.das@ucalgary....
Menu options
Create Cluster
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
ResearchCluster
UI | JSON
Databricks Runtime Version
Runtime: 8.3 (Scala 2.12, Spark 3.1.1)
Note Databricks Runtime 8.x uses Delta Lake as the default table format. Learn more
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
Availability Zone
auto
Activity Monitor


databricks
Data Science & E...
Create
Workspace
Recents
Search
Data
Compute
Jobs
Help
Settings
ajoy.das@ucalgary....
Menu options
Quickstart Notebook (SQL)
ResearchCluster
File
Edit
View: Standard
Permissions
Run All
Clear
Publish
Comments
Experiment
Cmd 1
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
Google Chrome

Resources

Demo Resources:

- https://docs.databricks.com/getting-started/quick-start.html
- https://docs.databricks.com/getting-started/spark/quick-start.html#notebook-qs

Documentations:

- Databricks Docs: https://docs.databricks.com/workspace-index.html
- Databricks User Manual: https://docs.databricks.com/user-guide/index.html
- Importing & Modifying Data: https://docs.databricks.com/data/data.html
- Community Edition FAQs: https://databricks.com/product/faq/community-edition

Thank You

ANY QUESTION?