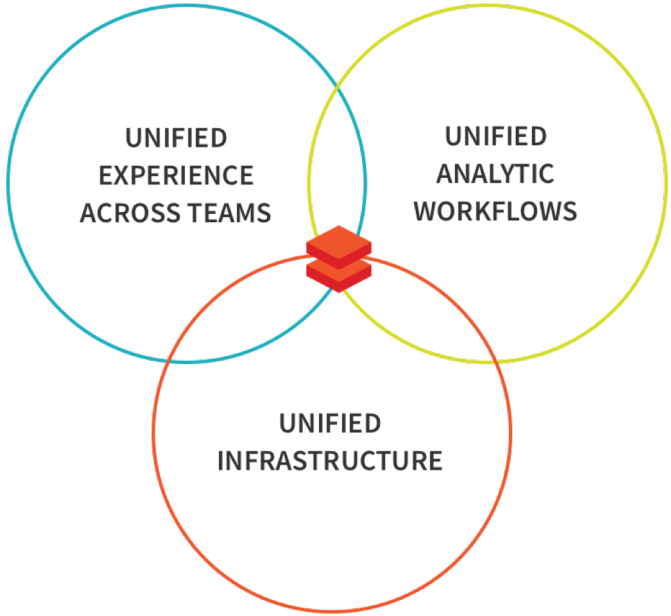
databricks®

Databricks

- A cloud-based managed platform for running Apache Spark.
- We don't have to learn cluster management and perform tedious maintenance tasks.
- It helps users to become more productive with Spark.
- It provides an easy to use user interface (UI) and a sophisticated API for Data Scientists and Analysts that want to automate aspects of their data workloads with automated jobs.
- Databricks is an implementation of Spark to help reduce complexity of setup and operation. On the other hand, Apache Spark is an open-source platform for distributed computing.



Databricks Community Editions

Databricks Community Edition is designed for developers, data scientists, data engineers and anyone who want to learn Spark.

- 6GB cluster
- Interactive notebooks and dashboards
- Public environment to share your work

Clusters

- Databricks clusters provide a unified platform for various use cases such as running production ETL pipelines, streaming analytics, ad-hoc analytics, and machine learning.
- Databricks has two types of clusters:
  - Interactive
    - Interactive clusters are used to analyze data collaboratively with interactive notebooks.
  - Job
    - Job clusters are used to run fast and robust automated workloads using the UI or API.

Workspace

- The Workspace is the special root folder for all of your organization's Databricks assets.
- The Workspace stores all your notebooks, libraries, and dashboards.
    - By default, the Workspace and all its contents are available to users, but each user also has a private home folder that is not shared.
- You can control who can view, edit, and run objects in the Workspace by enabling Workspace access control.

Databases and Tables

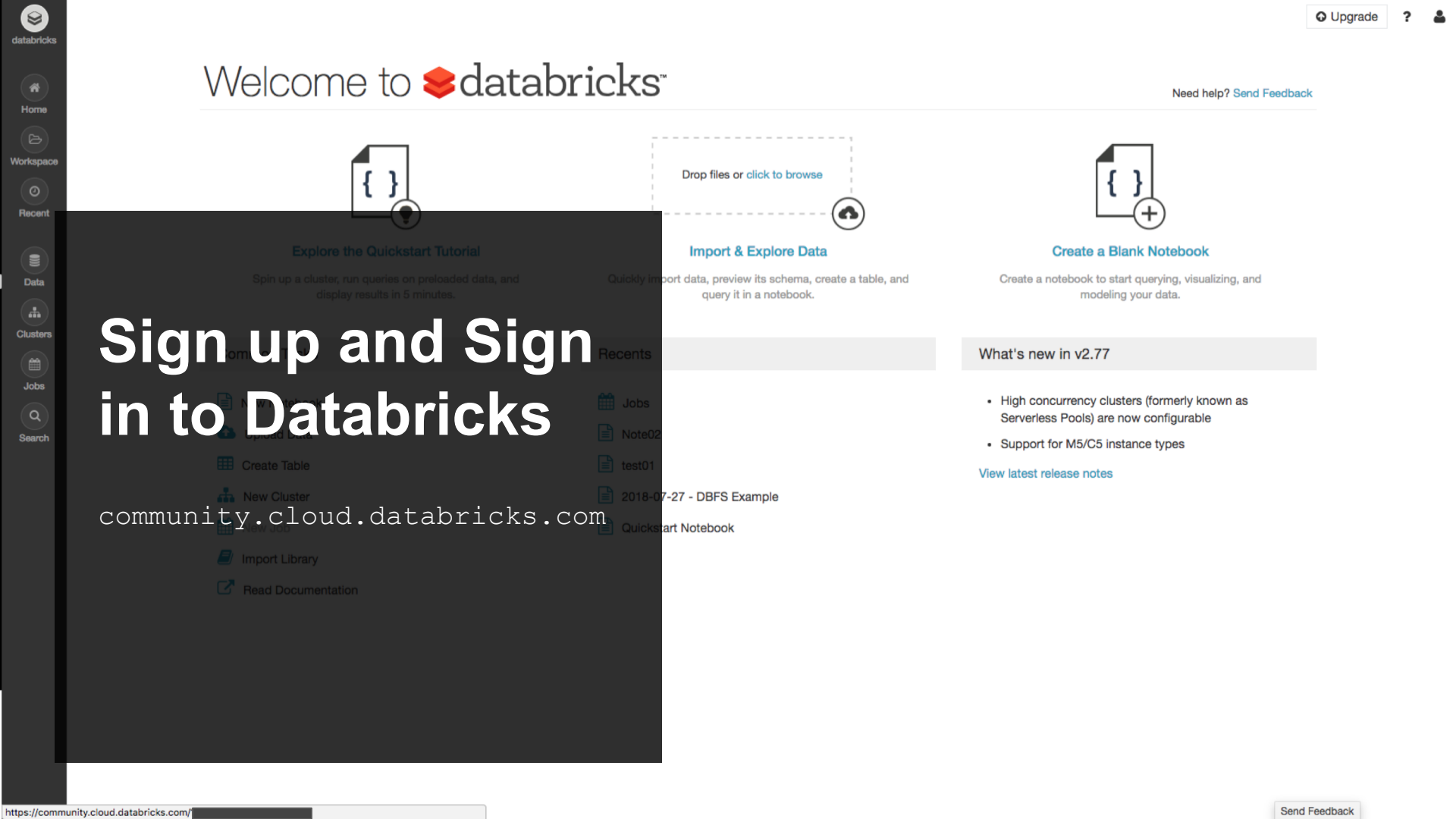
- A Databricks database is a collection of tables.
- A Databricks table is a collection of structured data.
  - Tables are equivalent to Apache Spark DataFrames. This means that you can cache, filter, and perform any operations supported by DataFrames on tables. You can query tables with Spark APIs and Spark SQL.
- There are two types of tables:
  - Global
    - A global table is available across all clusters. Databricks registers global tables to the Hive metastore.
  - Local
    - A local table is not accessible from other clusters and is not registered in the Hive metastore. This is also known as a temporary table or a view.

Libraries

- To make third-party or locally-built code available to execution environments running on your clusters, you create a library. Libraries can be written in Python, Java, Scala, and R.
- To allow a library to be shared by all users in a Workspace, create the library in the Shared folder. To make it available to a single user, create the library in the user folder.

Jobs

- A job is a way of running a notebook or JAR either immediately or on a scheduled basis.
- You can create and run jobs using the UI, the CLI, and by invoking the Jobs API. Similarly, you can monitor job run results in the UI, using the CLI, by querying the API, and through email alerts.
- It does not exist on the community edition.


databricks
Home
Workspace
Recent
Data
Clusters
Jobs
Search
Upgrade
Welcome to databricks
Need help? Send Feedback
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
Create a Blank Notebook
Create a notebook to start querying, visualizing, and modeling your data.
Recents
Jobs
Note02
test01
2018-07-27 - DBFS Example
Quickstart Notebook
Create Table
New Cluster
Import Library
Read Documentation
What's new in v2.77
High concurrency clusters (formerly known as Serverless Pools) are now configurable
Support for M5/C5 instance types
View latest release notes
Sign up and Sign in to Databricks
community.cloud.databricks.com
Send Feedback