Databricks, Docker

Databricks is an analytics **platform** designed for big data analytics. It is focused on analytics and data science, providing a platform for data-related tasks. (A shared **playground** for data tasks.)
On the other hand, Docker is focused on **containerization**, allowing for consistent and portable execution of software applications. (A **box** that helps us pack and carry our applications so they can work anywhere.)
While Databricks can be used within Docker containers for deployment purposes, they are distinct technologies with different purposes.

DBFS, HDFS

Databricks File System (DBFS) is a distributed **file storage system** provided by Databricks **within their platform**, facilitating data storage, retrieval, and management for analytics and data processing tasks.
HDFS is a distributed file system designed for storing and managing large files **across a cluster of machines**.
While both systems serve the purpose of distributed file storage, DBFS is **specific** to the Databricks platform, while HDFS is more widely used in the broader **Hadoop ecosystem**.

Spark

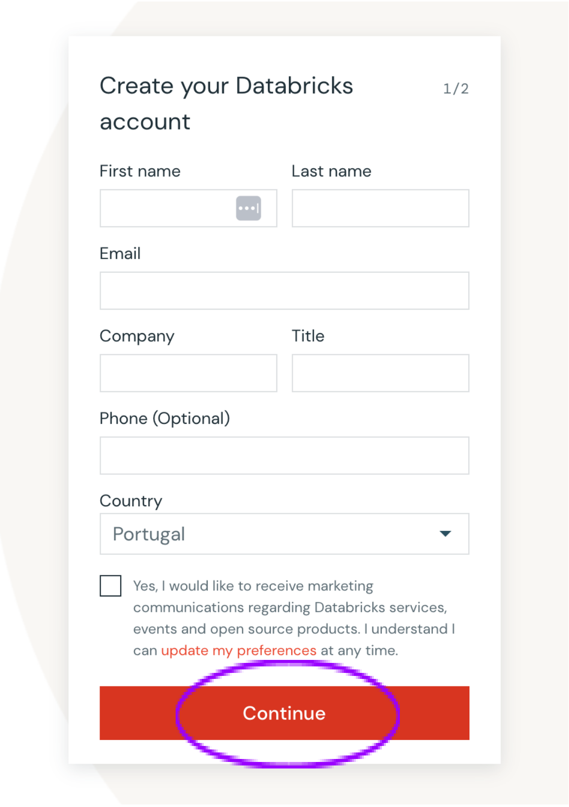
Apache Spark is a computing **framework** designed to engage with large volumes of data.
As our data is distributed across the HDFS, we seek an effective way to use it. We can use the MapReduce approach or Spark, a modern alternative. Spark revolves around the concept of Resilient Distributed Datasets (RDDs), prioritizing the loading of **data into memory** for faster calculations.

GETTING STARTED WITH DATABRICKS

- Sign up for a free community account at databricks (an alternative to Docker).
- Make sure to select the community edition.

https://databricks.com/try-databricks

#1



Create your Databricks account 1/2

First name

Last name

Email

Company

Title

Phone (Optional)

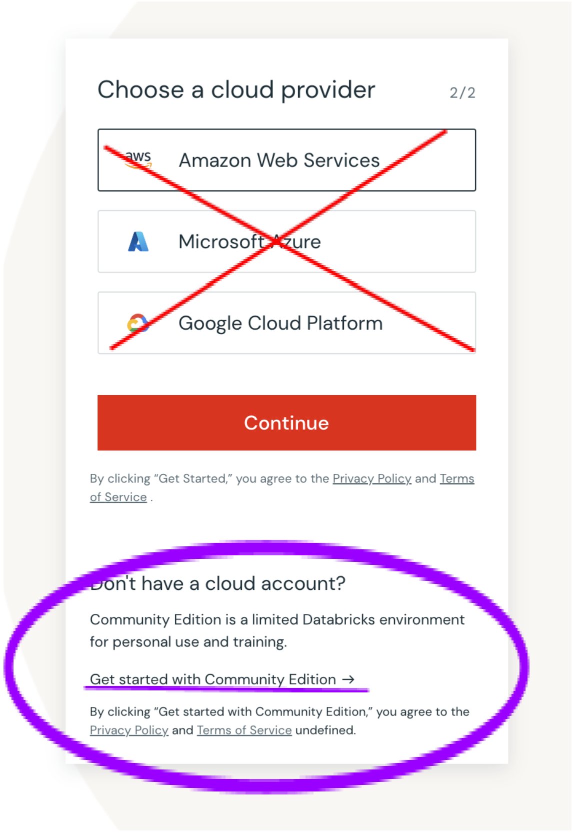
Country

Portugal

☐ Yes, I would like to receive marketing communications regarding Databricks services, events and open source products. I understand I can update my preferences at any time.

Continue

#2



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 .

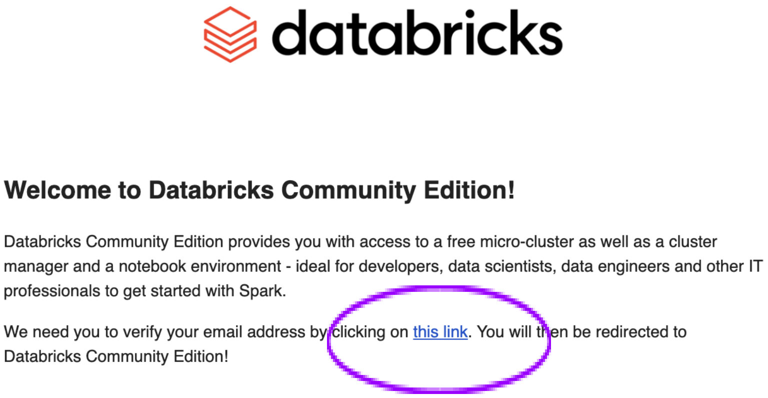
Don't have a cloud account?

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 undefined.

#3



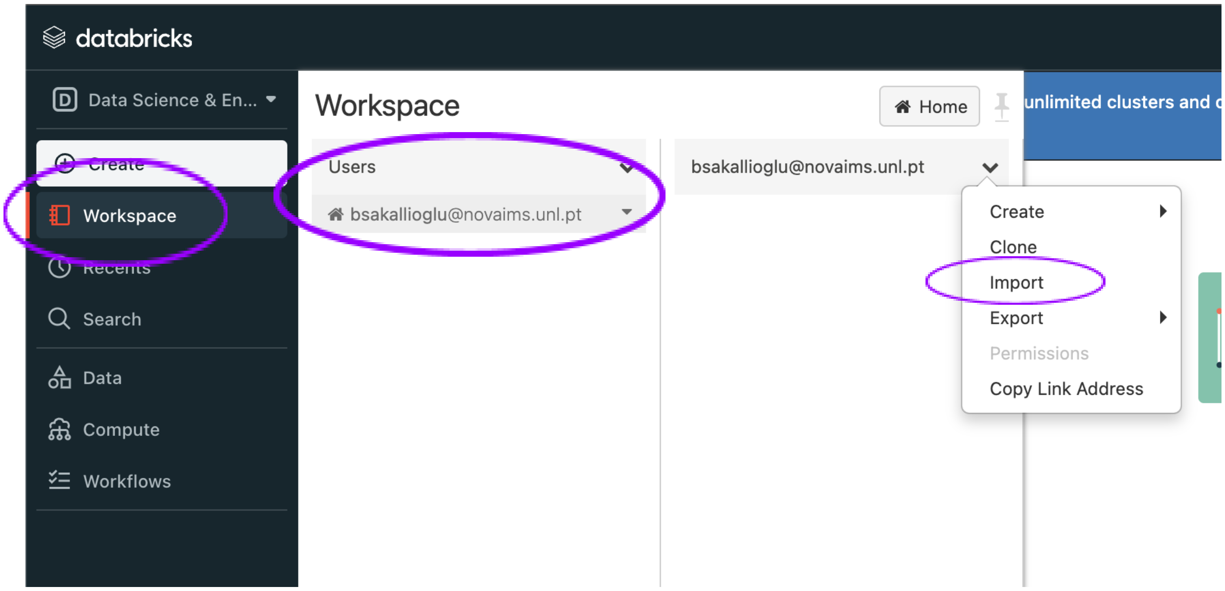
databricks

Welcome to Databricks Community Edition!

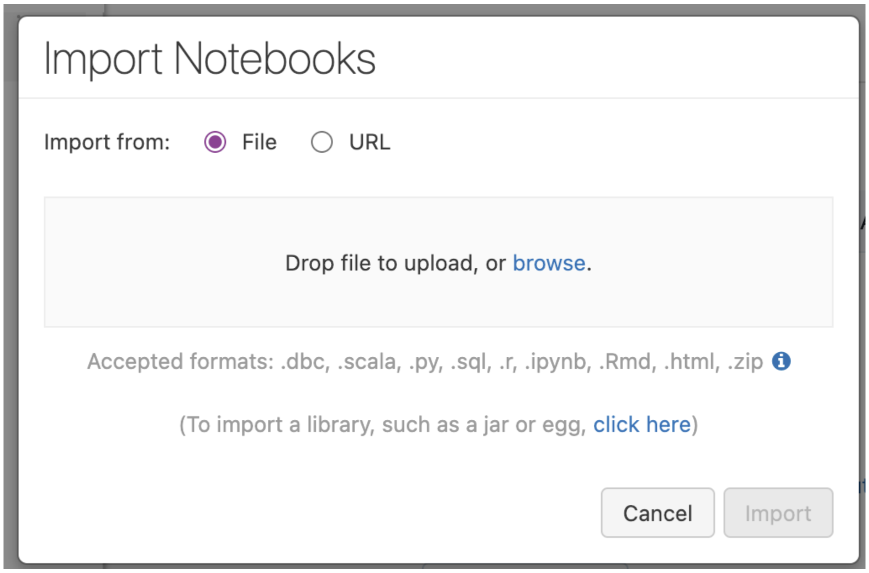
Databricks Community Edition provides you with access to a free micro-cluster as well as a cluster manager and a notebook environment - ideal for developers, data scientists, data engineers and other IT professionals to get started with Spark.

We need you to verify your email address by clicking on this link. You will then be redirected to Databricks Community Edition!

#4



#5



Drop file to upload, or browse.

Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html, .zip

(To import a library, such as a jar or egg, click here)

You can choose our *spark_rdd_introduction.ipynb* file!