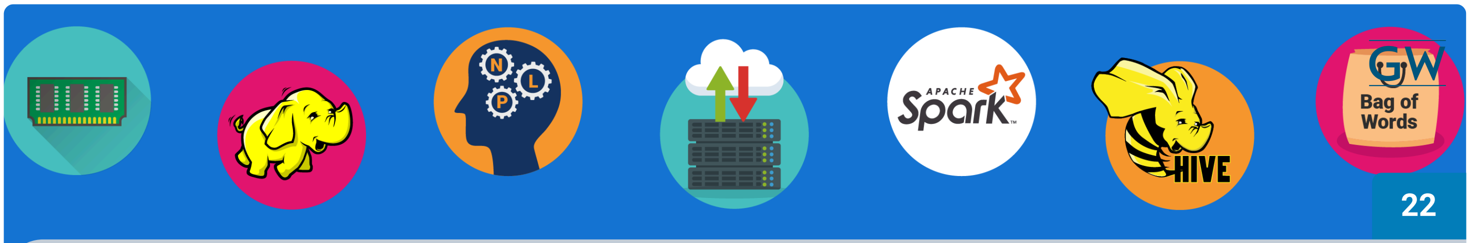


Getting Started

This section serves as a tutorial for you to set up the tools you need for this module: PySpark and Databricks.

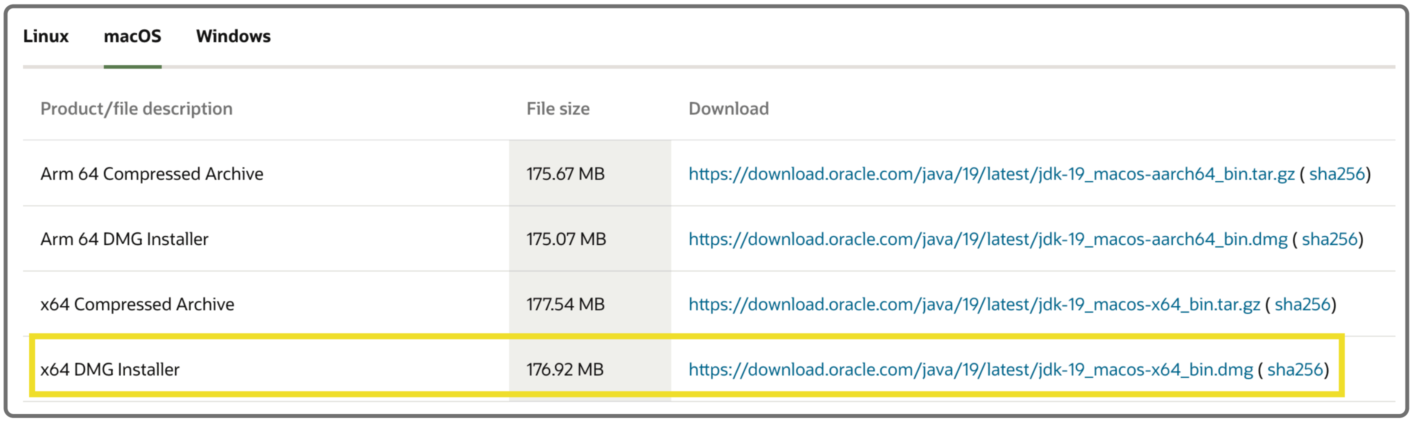
IMPORTANT

> Before installing new tools, open your terminal and make sure that your `dev` Conda environment is activated.

Installing and Setting up PySpark on macOS

Install Java

1. Java is required to run PySpark. Before you install Java, check to see if you have Java installed by running the following in the command line, `java -version`.
2. If Java is not installed, download the x64 Installer from the **Oracle website (https://www.oracle.com/java/technologies/downloads/#jdk20-mac)** .



Linux | **macOS** | Windows

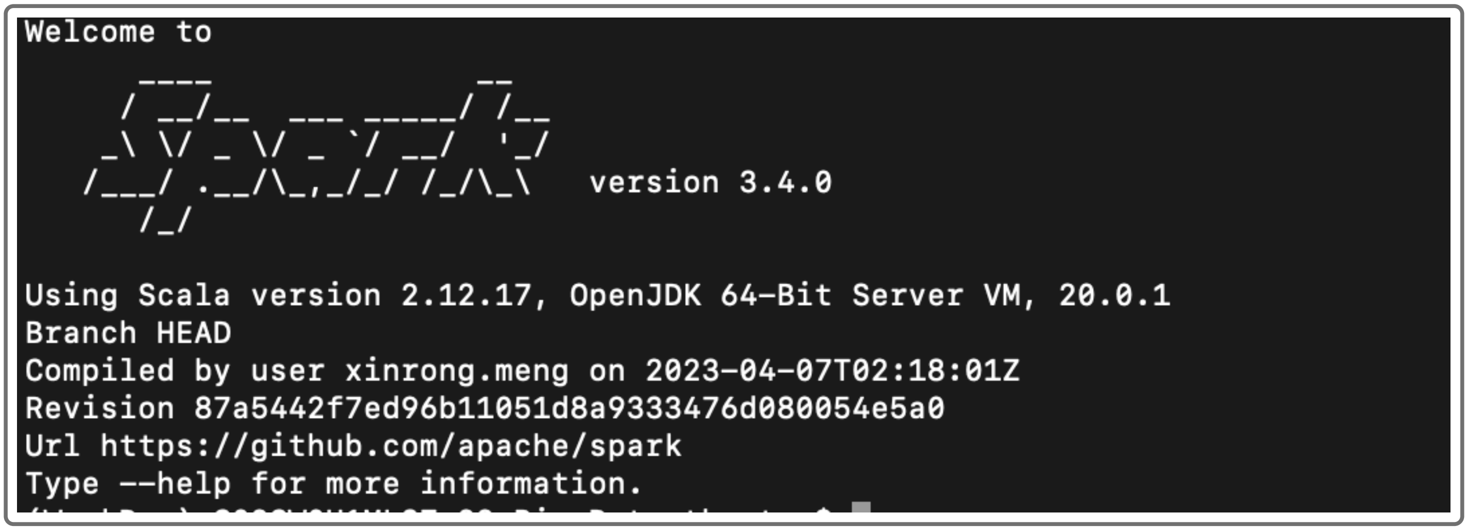
Product/file description	File size	Download
Arm 64 Compressed Archive	175.67 MB	https://download.oracle.com/java/19/latest/jdk-19_macos-aarch64_bin.tar.gz (sha256)
Arm 64 DMG Installer	175.07 MB	https://download.oracle.com/java/19/latest/jdk-19_macos-aarch64_bin.dmg (sha256)
x64 Compressed Archive	177.54 MB	https://download.oracle.com/java/19/latest/jdk-19_macos-x64_bin.tar.gz (sha256)
x64 DMG Installer	176.92 MB	https://download.oracle.com/java/19/latest/jdk-19_macos-x64_bin.dmg (sha256)

3. Or, you can use **Homebrew (https://formulae.brew.sh/formula/openjdk)** to install Java. On the terminal, type and run `brew install openjdk` to install Java.
4. After you install Java, you can check your installation by running, `java -version`.

Install PySpark

On the terminal, type and run `pip install pyspark==3.4.0`.

After you have installed PySpark, you can check your installation by running, `spark-submit --version` in the terminal. The output should be similar to the following image, just with updated version numbers:



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Scala version 2.12.17, OpenJDK 64-Bit Server VM, 20.0.1
Branch HEAD
Compiled by user xinrong.meng on 2023-04-07T02:18:01Z
Revision 87a5442f7ed96b11051d8a9333476d080054e5a0
Url https://github.com/apache/spark
Type --help for more information.
```

Install Findspark

On the terminal, type and run `conda install -c conda-forge findspark` to install Findspark.

- **Note:** Findspark adds a startup file to the current IPython profile so that the environment variables will be properly set and `pyspark` will be imported upon IPython startup.

Install PyArrow and Fastparquet

On the terminal, type and run `conda install -c conda-forge pyarrow` and `conda install -c conda-forge fastparquet`.

- **Note:** `pyarrow` and `fastparquet` will allow us to read and write parquet-format big data.

After you have installed `pyarrow` and `fastparquet`, you can check your installation by running; `conda list pyarrow` and `conda list fastparquet`.

Running PySpark in Jupyter Notebook

1. In your `dev` Conda environment, launch Jupyter notebook.
2. Select a new notebook with the `dev` kernel.
3. In the new notebook, type and run the following code:

```
# Import and initialize findspark
import findspark
findspark.init()
```

```
# Start Spark session
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Testing").getOrCreate()

# Create a Spark DataFrame
df = spark.createDataFrame([
(0, "First row"),
(1, "Second row"),
(2, "Third row")
], ["ids", "rows"])

df.show()
```

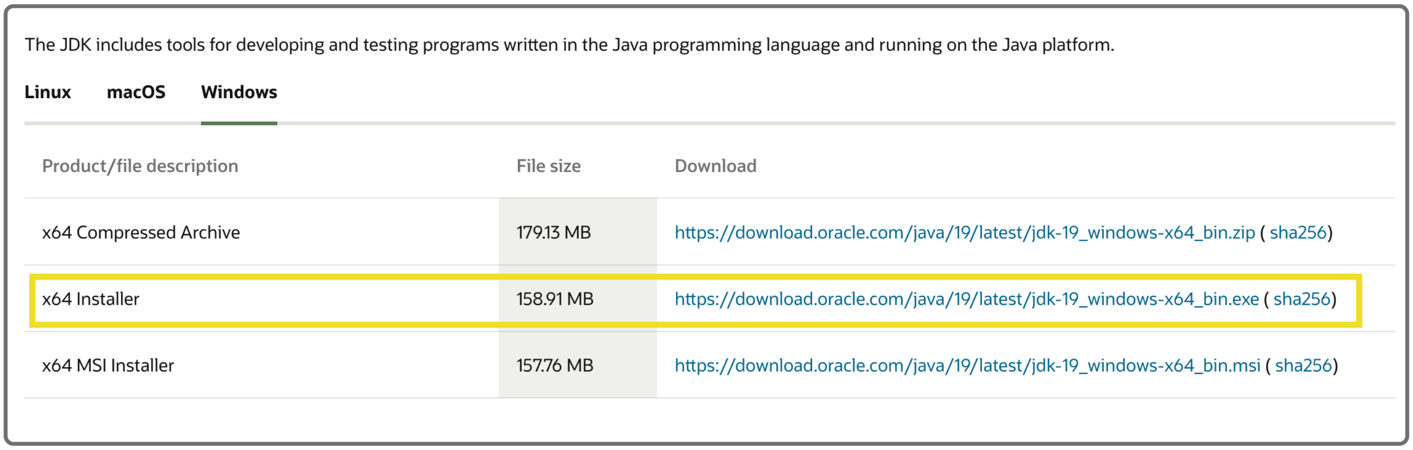
4. If your output looks like the following, congratulations!

ids	rows
0	First row
1	Second row
2	Third row

Installing and Setting up PySpark on Windows

Install Java

1. Before you install Java, check to see if you have Java installed by running the following in the command line, `java -version`.
2. If Java is not installed, download the x64 Installer from the **Oracle website (https://www.oracle.com/java/technologies/downloads/#jdk20-windows)** .



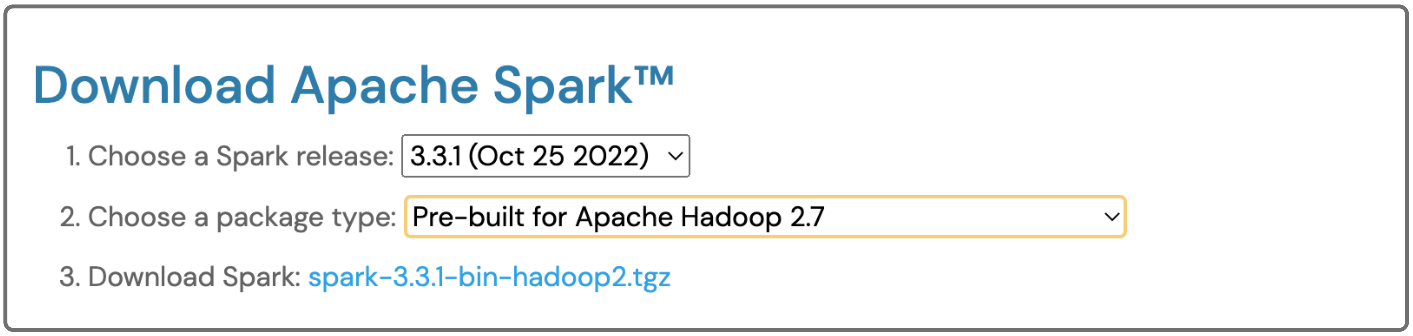
The JDK includes tools for developing and testing programs written in the Java programming language and running on the Java platform.

Product/file description	File size	Download
x64 Compressed Archive	179.13 MB	https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.zip (sha256)
x64 Installer	158.91 MB	https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.exe (sha256)
x64 MSI Installer	157.76 MB	https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.msi (sha256)

3. After you install Java, you can check your installation by running, `java -version`.

Download and Install PySpark

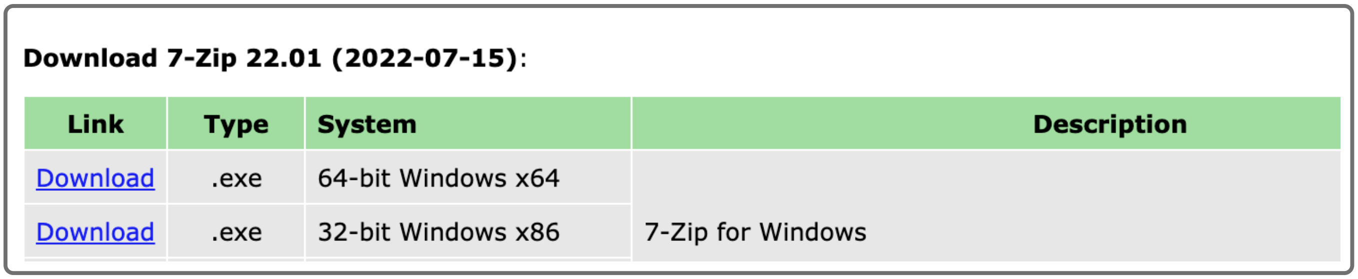
1. From the [Apache Spark (https://spark.apache.org/downloads.html)](https://spark.apache.org/downloads.html) distribution website, select the Spark 3.5.1 release and the Apache Hadoop 3.3 package.



Download Apache Spark™

1. Choose a Spark release: 3.3.1 (Oct 25 2022)
2. Choose a package type: Pre-built for Apache Hadoop 2.7
3. Download Spark: spark-3.3.1-bin-hadoop2.tgz

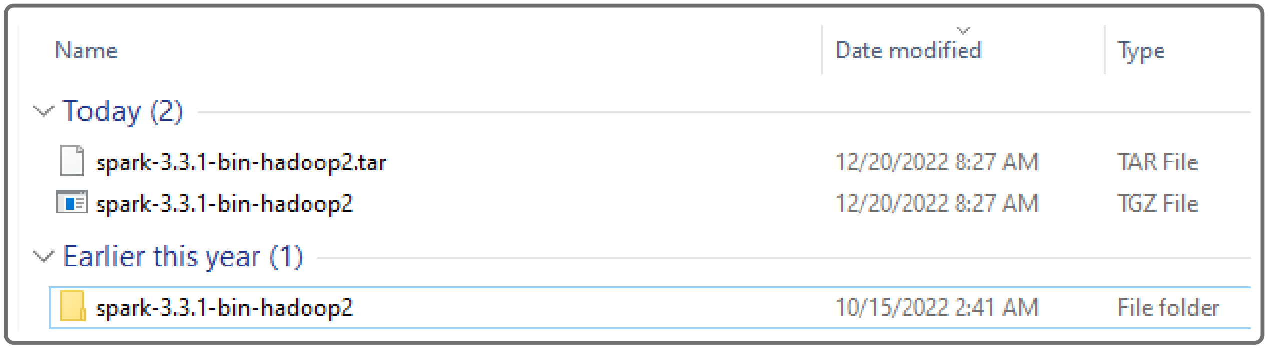
2. Apache Spark download is a `.tgz` file, which can be unpacked with [7-Zip (https://7-zip.org/download.html)](https://7-zip.org/download.html).
   - If you don't have 7-Zip, download and install the distribution for your system.



Download 7-Zip 22.01 (2022-07-15):

Link	Type	System	Description
Download	.exe	64-bit Windows x64	
Download	.exe	32-bit Windows x86	7-Zip for Windows

3. Next, unpack the `.tgz` file to create the `.tar` file. Then, unpack the `.tar` file with 7-Zip to get the "spark-3.5.1-bin-hadoop3" folder.

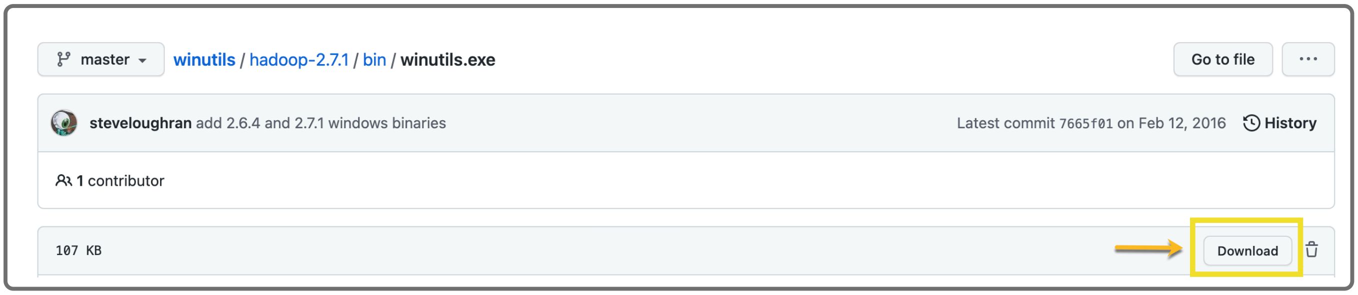


Name	Date modified	Type
Today (2)		
spark-3.3.1-bin-hadoop2.tar	12/20/2022 8:27 AM	TAR File
spark-3.3.1-bin-hadoop2	12/20/2022 8:27 AM	TGZ File
Earlier this year (1)		
spark-3.3.1-bin-hadoop2	10/15/2022 2:41 AM	File folder

4. Move the "spark-3.5.1-bin-hadoop3" folder into the `C:\Users` folder on your computer.

5. Download the Hadoop binary for Windows, `winutils.exe`, from Steve Loughran's **GitHub (https://github.com/steveloughran/winutils/)**.

    - Click on the "hadoop-3.0.0" version, since that is the Hadoop version we downloaded.
    - Open the "bin" folder.
    - Click on the `winutils.exe` file.
    - Click "Download" to download the `winutils.exe` file onto your computer.

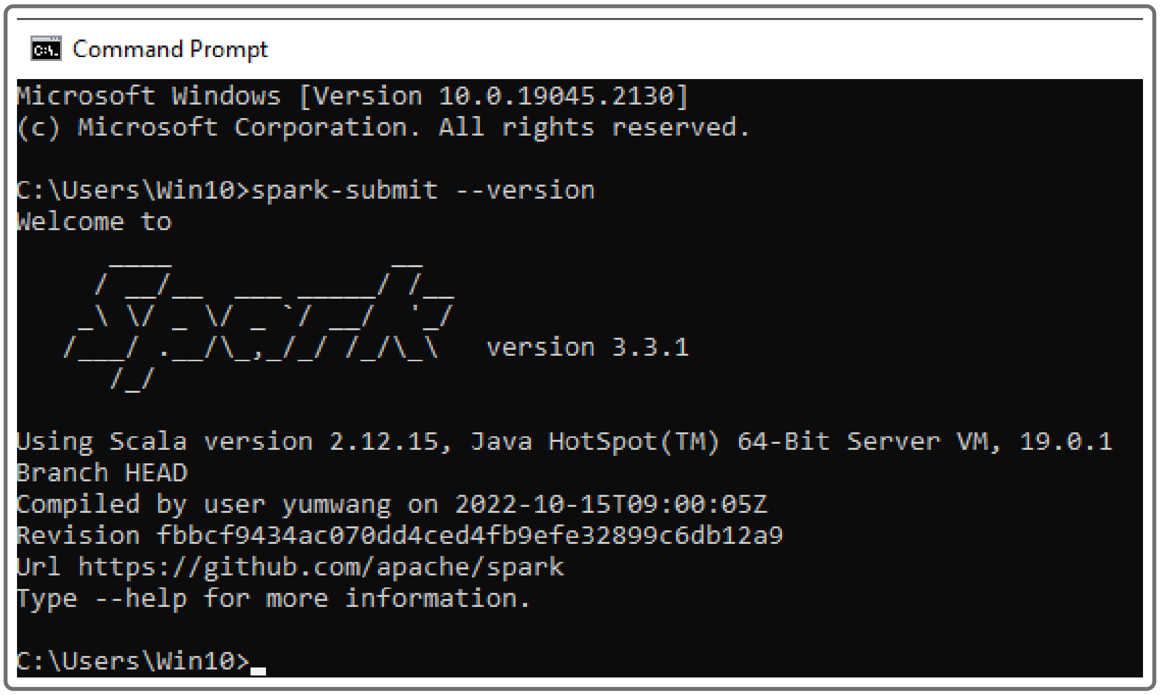


steveloughran add 2.6.4 and 2.7.1 windows binaries

Latest commit 7665f01 on Feb 12, 2016

6. Next, move the `winutils.exe` file into the "bin" folder of the "spark-3.5.1-bin-hadoop3" folder.

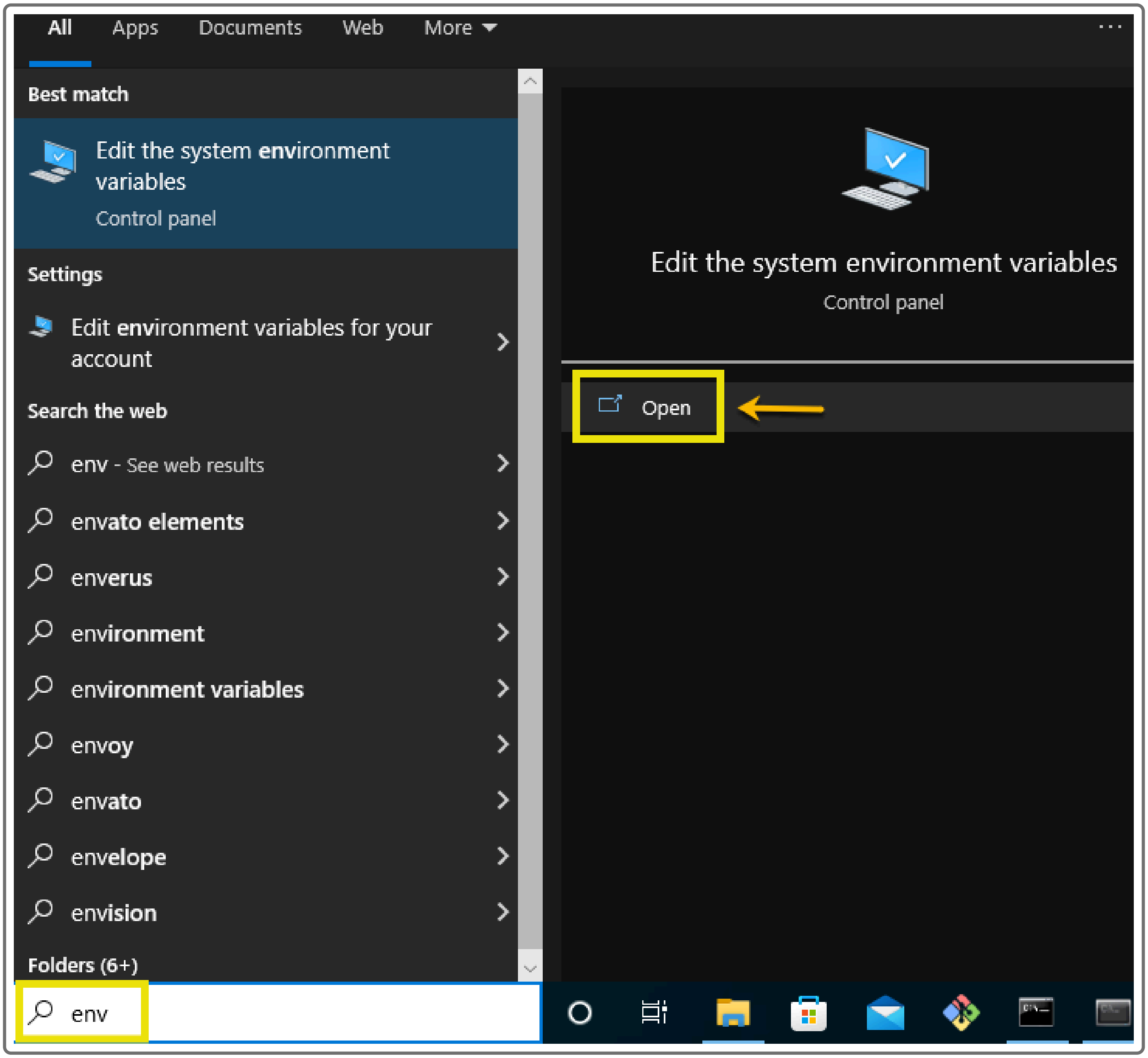
7. Check the installation of Spark by typing and running `spark-submit --version` in the command line. The output should be similar to the following image:



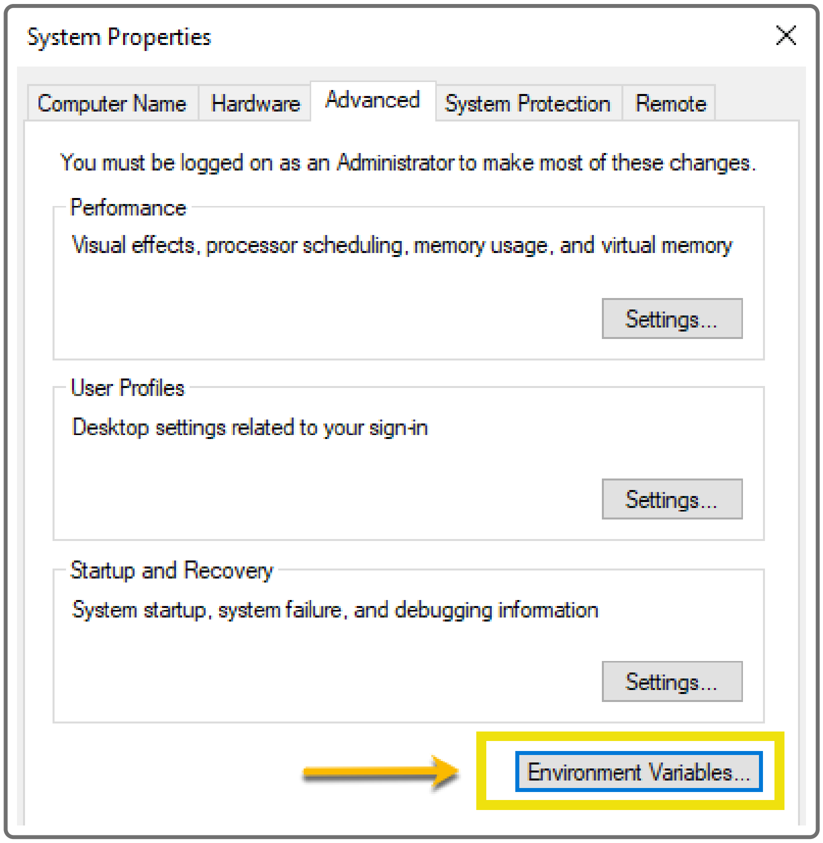
```
Microsoft Windows [Version 10.0.19045.2130]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Win10>spark-submit --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Scala version 2.12.15, Java HotSpot(TM) 64-Bit Server VM, 19.0.1
Branch HEAD
Compiled by user yumwang on 2022-10-15T09:00:05Z
Revision fbbcf9434ac070dd4ced4fb9efe32899c6db12a9
Url https://github.com/apache/spark
Type --help for more information.

C:\Users\Win10>
```

Set up the Environment Variables

1. Open the environment variables by typing "env " in the search box, then click "Open".



2. In the System Properties, open the "Environment Variables".



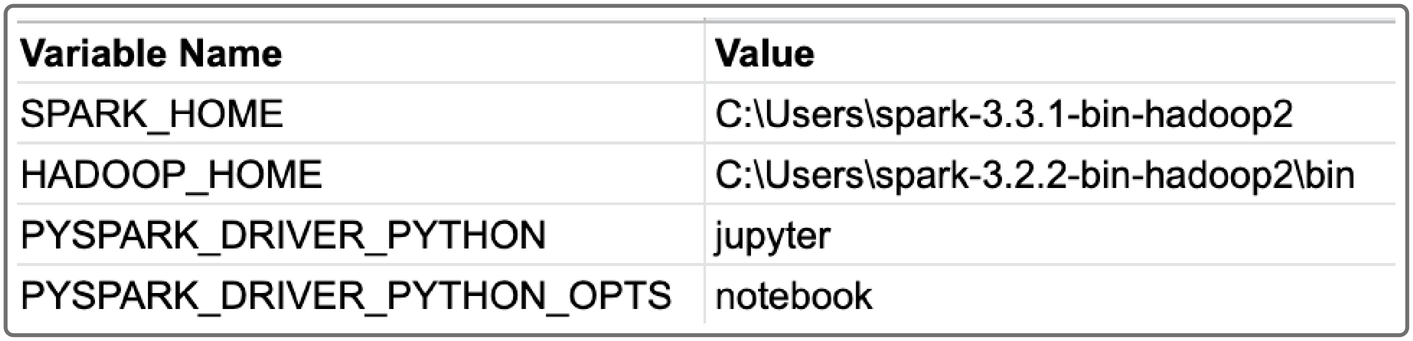
You must be logged on as an Administrator to make most of these changes.

Visual effects, processor scheduling, memory usage, and virtual memory

Desktop settings related to your sign-in

System startup, system failure, and debugging information

3. In the "User variables", create four new environment variables as follows:



Variable Name	Value
SPARK_HOME	C:\Users\spark-3.3.1-bin-hadoop2
HADOOP_HOME	C:\Users\spark-3.2.2-bin-hadoop2\bin
PYSPARK_DRIVER_PYTHON	jupyter
PYSPARK_DRIVER_PYTHON_OPTS	notebook

- **Note:** If you didn't move the "spark-3.5.1-bin-hadoop3" folder into the `C:\Users` folder, you'll have to add the new path as the value.

4. Save all your changes.

- **Note:** You may have to restart your computer to update the environment variables.

Install Findspark

Activate your `dev` Conda environment and then type and run `conda install -c conda-forge findspark` to install Findspark.

- **Note:** Findspark adds a startup file to the current IPython profile so that the environment variables will be properly set and `pyspark` will be imported upon IPython startup.

Install PyArrow and Fastparquet

On the terminal type, run `conda install -c conda-forge pyarrow` and `conda install -c conda-forge fastparquet`.

- **Note:** `pyarrow` and `fastparquet` will allow us to read and write parquet-format big data.

Running PySpark in Jupyter Notebook

1. Open the Anaconda prompt and activate your `dev` Conda environment, then launch Jupyter notebook.
2. Select a new notebook with the `dev` kernel.
3. In the new notebook, type and run the following code:

```
# Import and initialize findspark
import findspark
findspark.init()

# Start Spark session
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Testing").getOrCreate()

# Create a Spark DataFrame
df = spark.createDataFrame([
(0, "First row"),
(1, "Second row"),
(2, "Third row")
], ["ids", "rows"])

df.show()
```

4. If your output looks like the following, congratulations you are all set!

ids	rows
0	First row
1	Second row
2	Third row

Creating a Databricks Account.

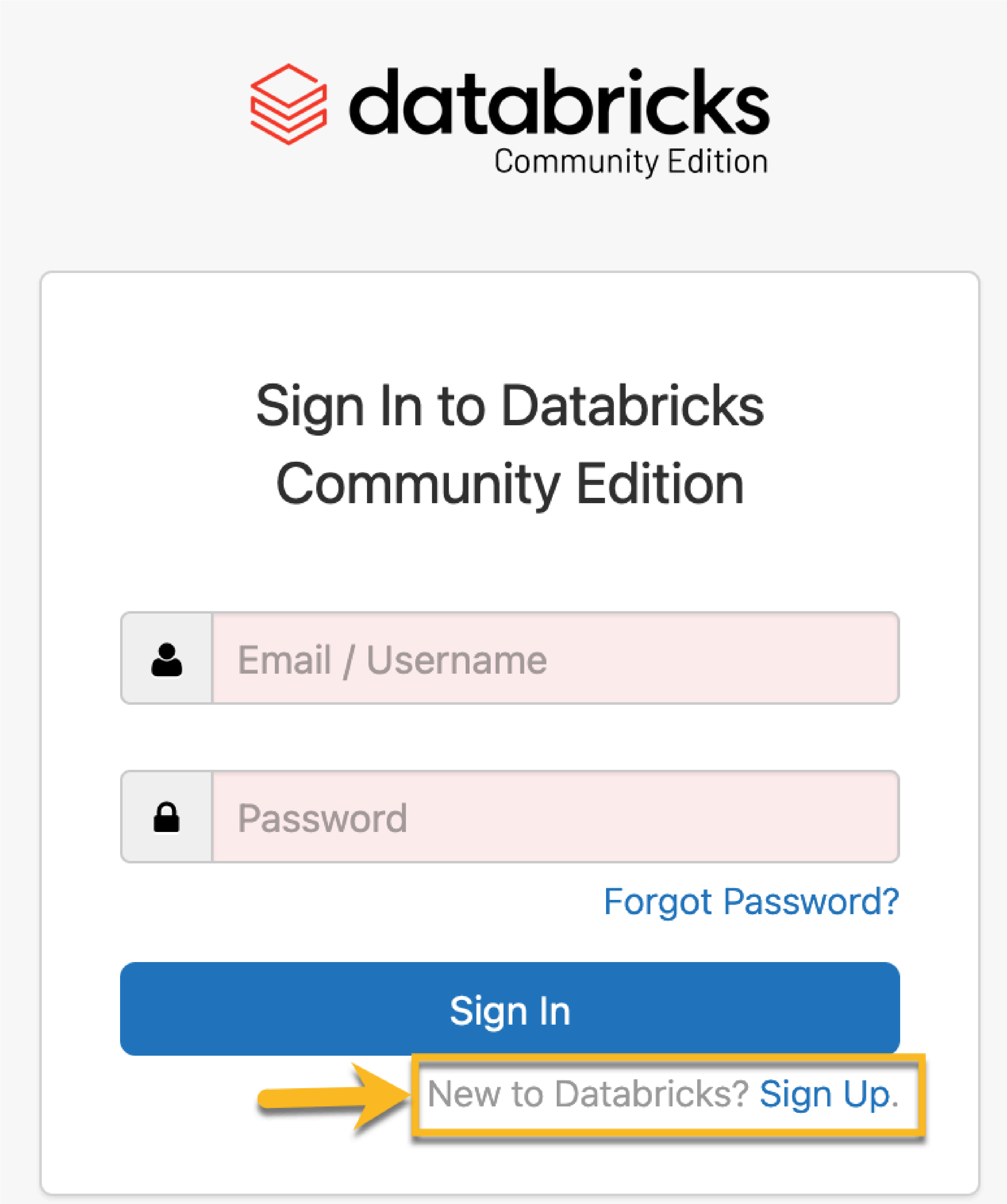
IMPORTANT

> The Databricks Community Edition is good for 14-days. We suggest that you to create an account the day before you use Databricks in the course.

This guide reviews the steps for creating a Databricks Community Edition account and using Databricks.

Create an Account

1. Go to the **Databricks Community Edition site** **(https://community.cloud.databricks.com/login.html)** and click “Sign Up”.



databricks
Community Edition

Sign In to Databricks Community Edition

Email / Username

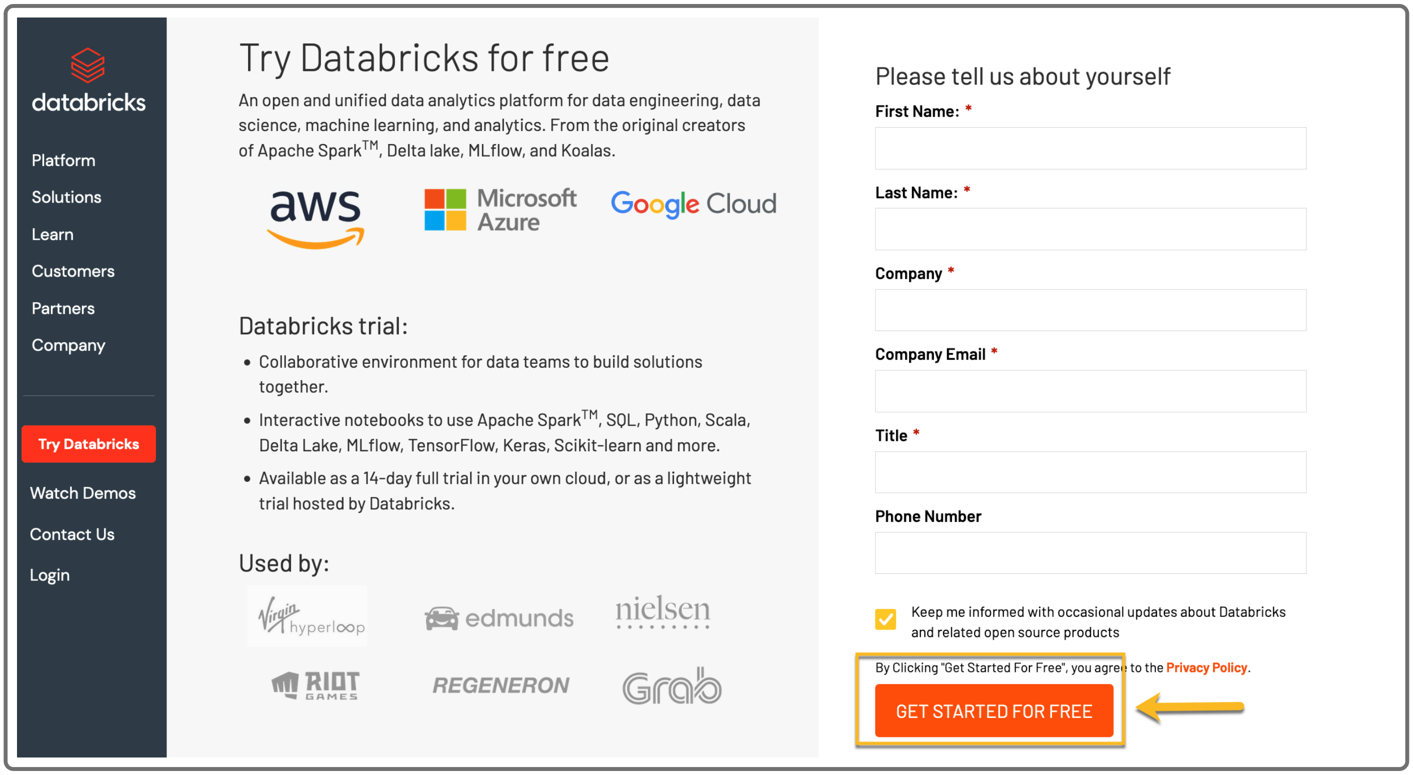
Password

Forgot Password?

Sign In

New to Databricks? Sign Up.

2. On the next page, fill out the required information and click "Get Started For Free."



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^{TM}, Delta lake, MLflow, and Koalas.

Databricks trial:

- Collaborative environment for data teams to build solutions together.
- Interactive notebooks to use Apache Spark^{TM}, 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:

Please tell us about yourself

First Name: *

Last Name: *

Company *

Company Email *

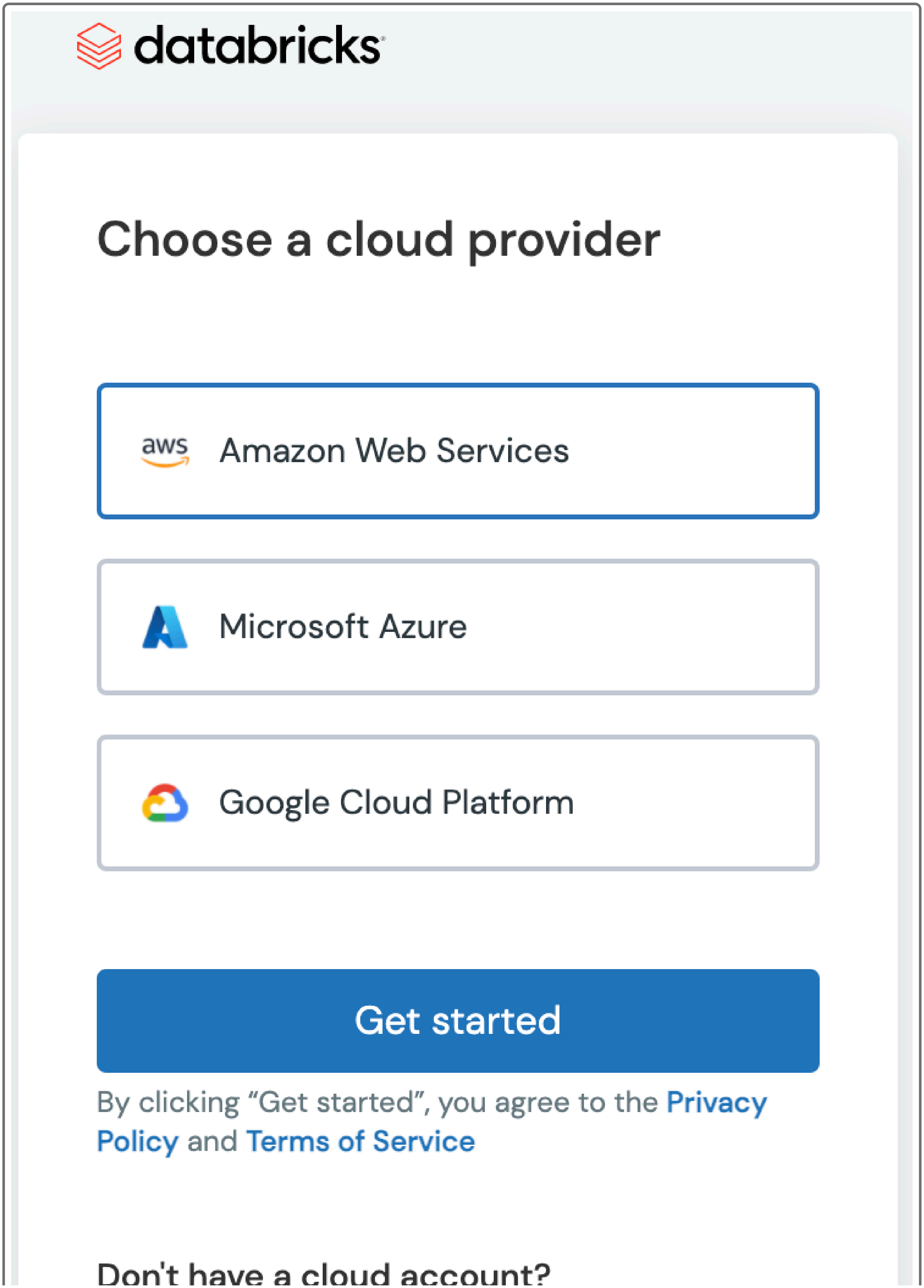
Title *

Phone Number

☑ Keep me informed with occasional updates about Databricks and related open source products

By Clicking "Get Started For Free", you agree to the **Privacy Policy**.

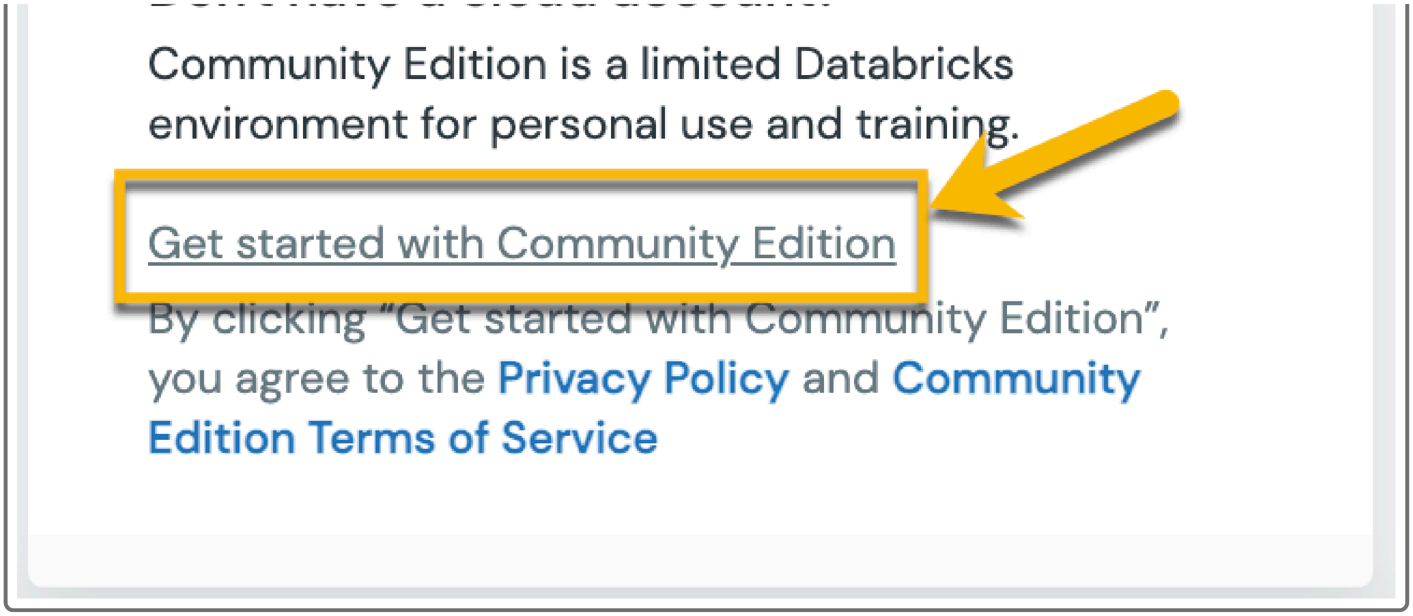
3. You will be redirected to sign up for the standard Databricks account. Do NOT click any of the cloud provider options. To use the Community Edition, click "Get started with Community Edition."



Choose a cloud provider

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

Don't have a cloud account?

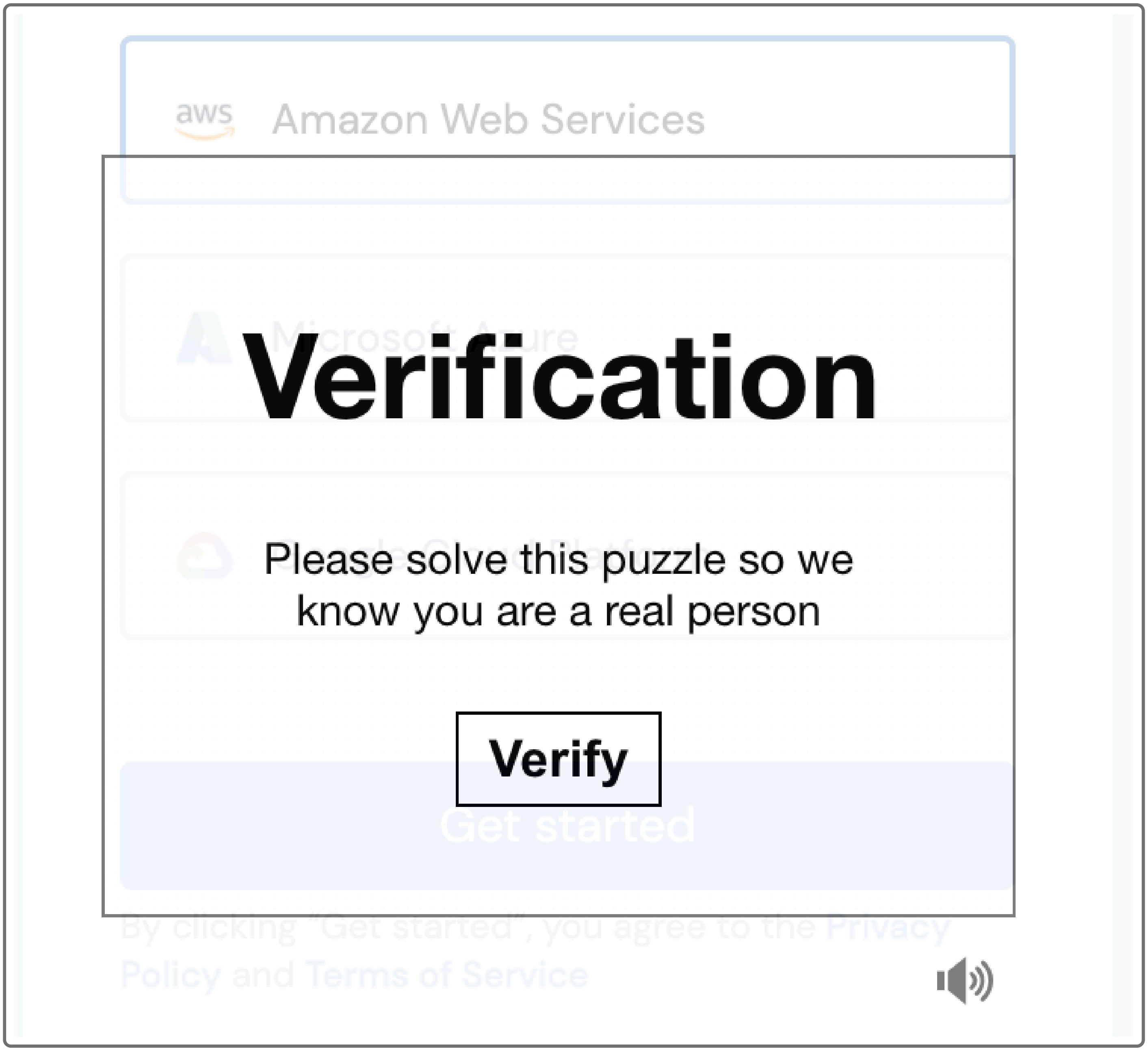


Community Edition is a limited Databricks environment for personal use and training.

Get started with Community Edition

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

4. Follow the onscreen prompts to verify your account.



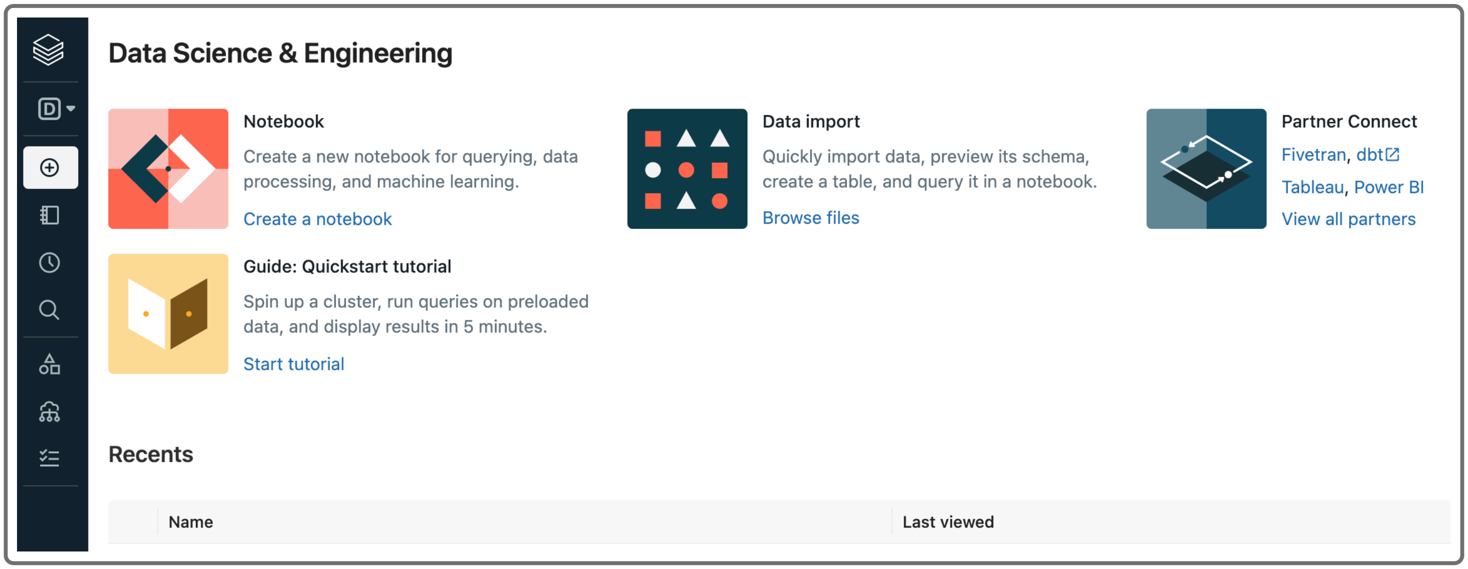
Please solve this puzzle so we know you are a real person

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

5. When prompted, check your email and click the link to verify your account and reset your password. Once you reset your password, you can log into the Community Edition.

Navigate the Community Edition

When you log into your Databricks Community Edition account, you'll see the Data Science and Engineering landing page:



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

Partner Connect

Fivetran, dbt

Tableau, Power BI

View all partners

Guide: Quickstart tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Start tutorial

Recents

Name	Last viewed

On the landing page, you can choose from four options:

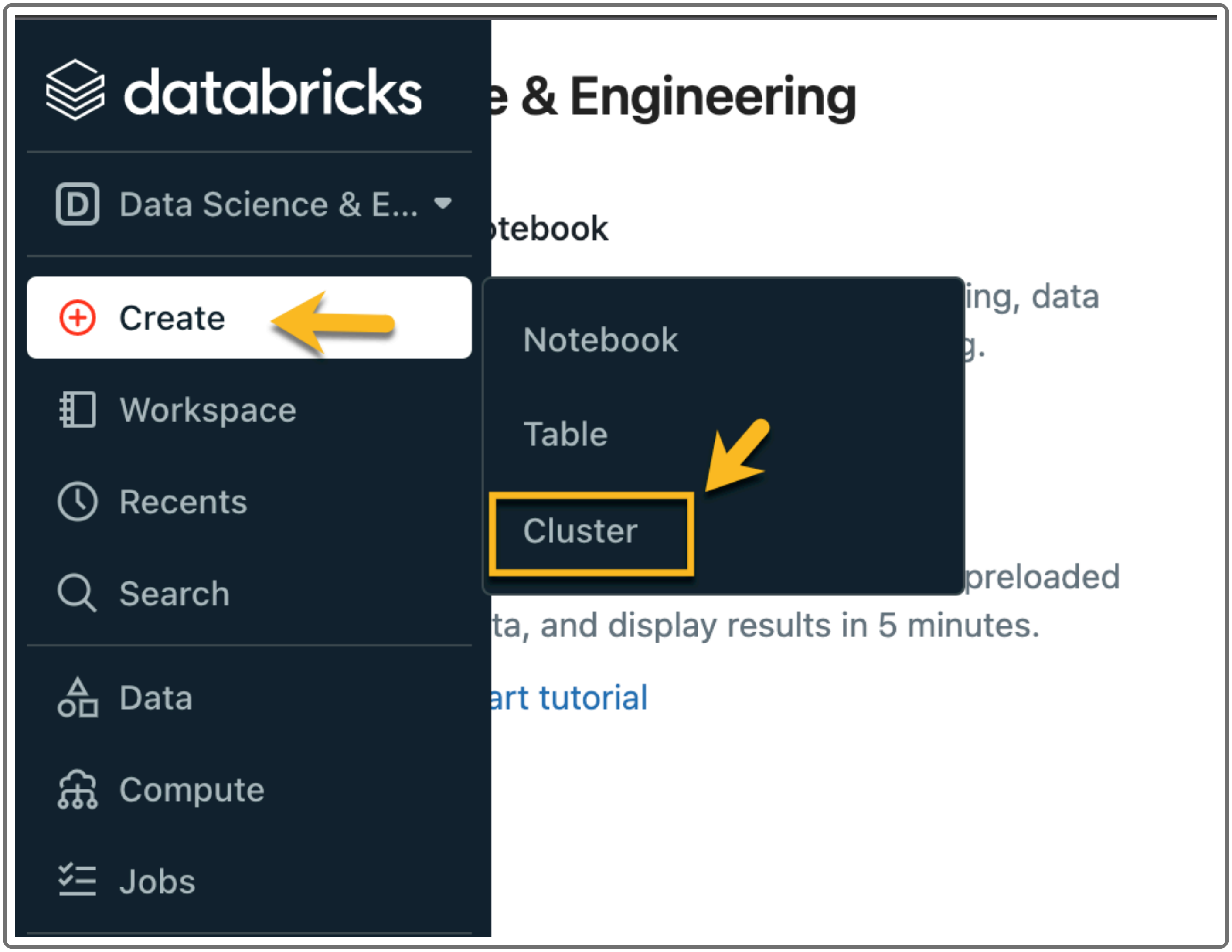
1. A quick start tutorial to help you create a cluster, attach a notebook to your cluster, create a table for a dataset, query the table using SQL, create a table and a graph, and create a DataFrame.
2. Create a new notebook, such as a Jupyter notebook.
3. Import data.
4. Connect to external software, like Tableau, Power BI, and more. **Note:** The Community Edition does not allow connections to external software.

You can use the quick start tutorial to familiarize yourself Databricks, or proceed to the following steps to start using Databricks.

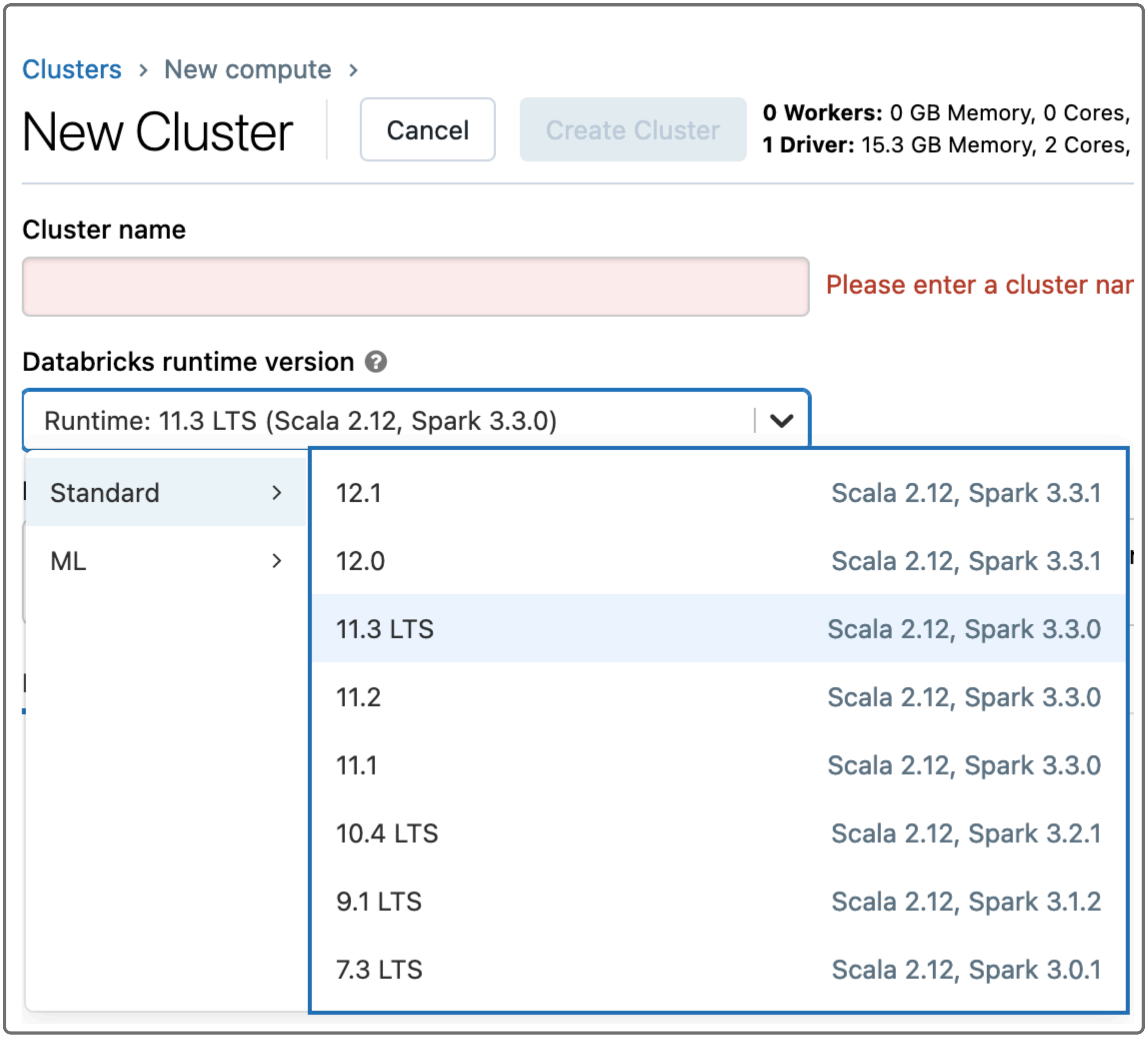
Using Databricks

Follow these steps to get started using Databricks.

1. Before you create a notebook, you have to create a cluster. On the navigation pane on the left side of the landing page, click "+" and select Cluster.



2. Use the default runtime settings, `11.3 LTS (Scala 2.12, Spark 3.3.0)`, or select an alternate version.


Clusters › New compute ›
Cancel
Create Cluster
Standard ›
ML ›
12.1 Scala 2.12, Spark 3.3.1
12.0 Scala 2.12, Spark 3.3.1
11.3 LTS Scala 2.12, Spark 3.3.0
11.2 Scala 2.12, Spark 3.3.0
11.1 Scala 2.12, Spark 3.3.0
10.4 LTS Scala 2.12, Spark 3.2.1
9.1 LTS Scala 2.12, Spark 3.1.2
7.3 LTS Scala 2.12, Spark 3.0.1

New Cluster

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

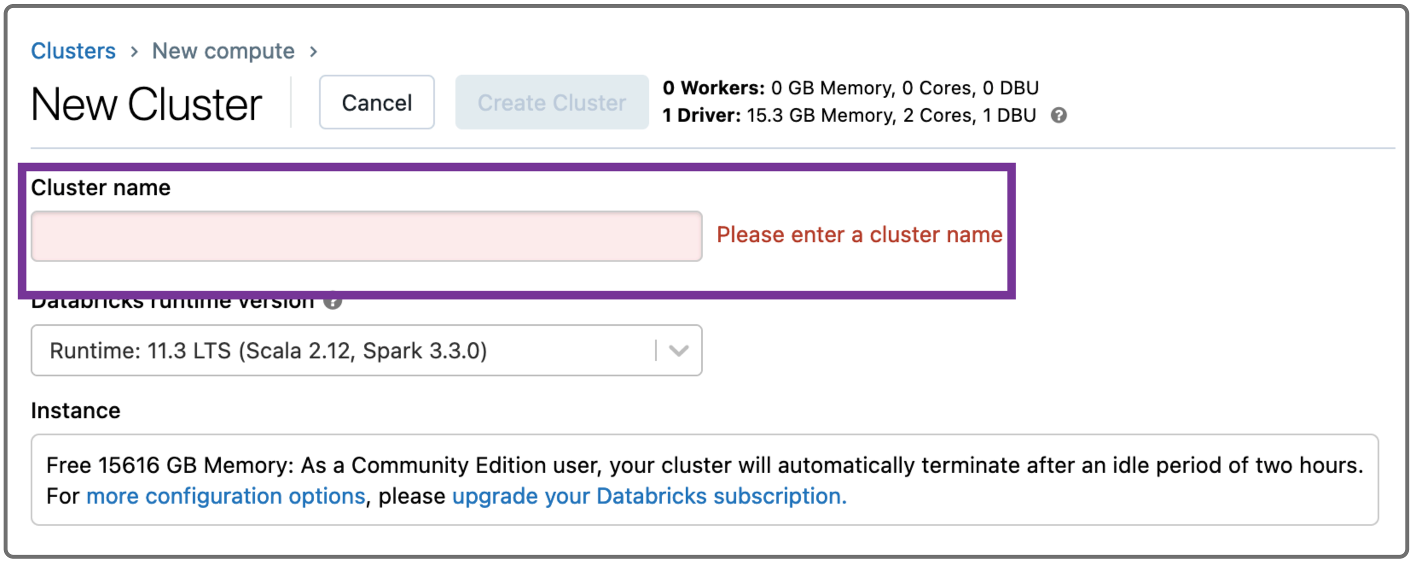
Cluster name

Please enter a cluster nar

Databricks runtime version

Runtime: 11.3 LTS (Scala 2.12, Spark 3.3.0)

3. Enter a name for your cluster.



Clusters › New compute ›

New Cluster

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

Cluster name

Please enter a cluster name

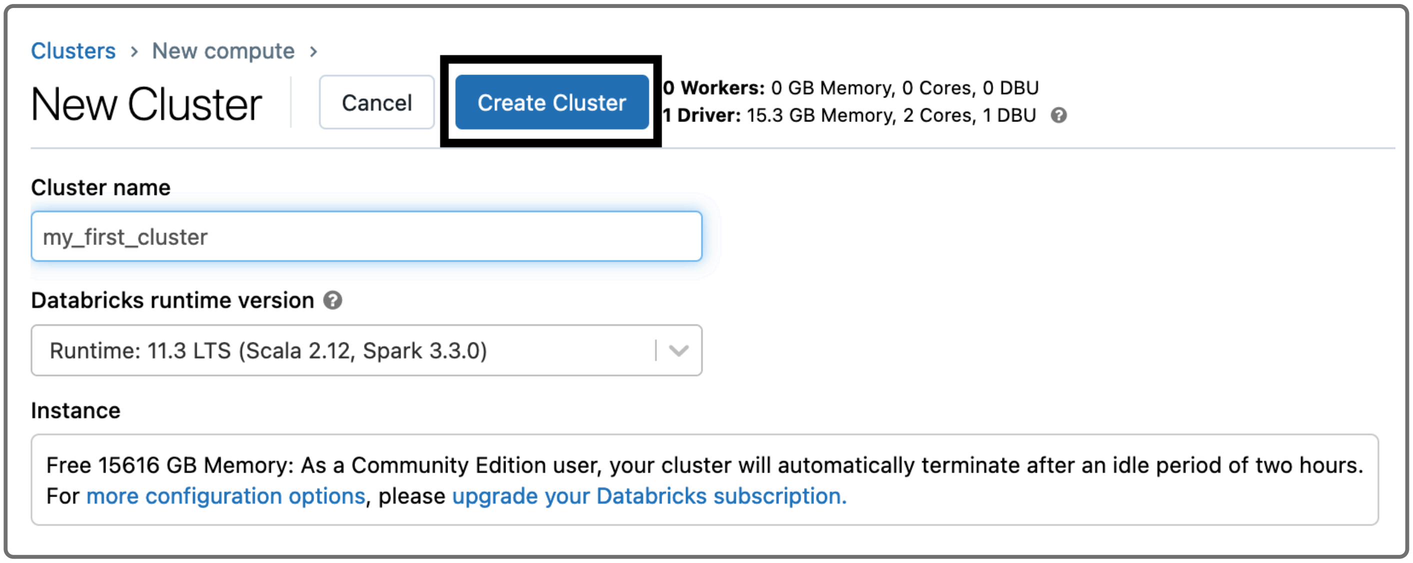
Databricks runtime version

Runtime: 11.3 LTS (Scala 2.12, Spark 3.3.0)

Instance

Free 15616 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.

4. Click the "Create Cluster" button at the top of the "Create Cluster" page.



Clusters › New compute ›

New Cluster

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

Cluster name

my_first_cluster

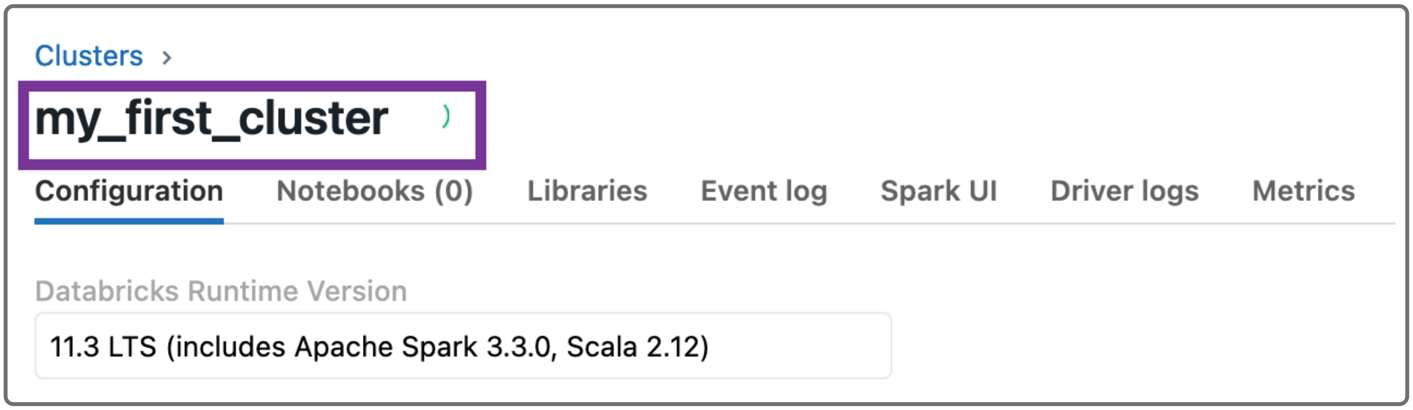
Databricks runtime version

Runtime: 11.3 LTS (Scala 2.12, Spark 3.3.0)

Instance

Free 15616 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.

5. After clicking "Create Cluster", a progress circle icon will spin while the cluster is being created. This may take a few minutes.



Clusters ›

my_first_cluster

Databricks Runtime Version

11.3 LTS (includes Apache Spark 3.3.0, Scala 2.12)

You’re now ready to use Databricks!