Lab Setup for PySpark Trainings

Prepared by Y. Kanakaraju for CTS – August 2025

- Databricks Free edition is a free version of databricks which allows us to work with Serverless compute.

- We can use this for practicing **Spark SQL** and **Structured Streaming** modules of PySpark. As the free edition does not support creating standalone clusters (i.e. All-purpose compute instances) **we cannot use it for Spark Core API (RDD API)** and also any other work which requires access to a standalone cluster.

- The lab setup for PySpark, therefore, has two parts, which are described in this document.

    - Setting up Databricks Free Edition (about 50% of the course)
    - Setting Jupyter Notebooks for PySpark development. (about 50% of the course)

Setting up Databricks Free Edition

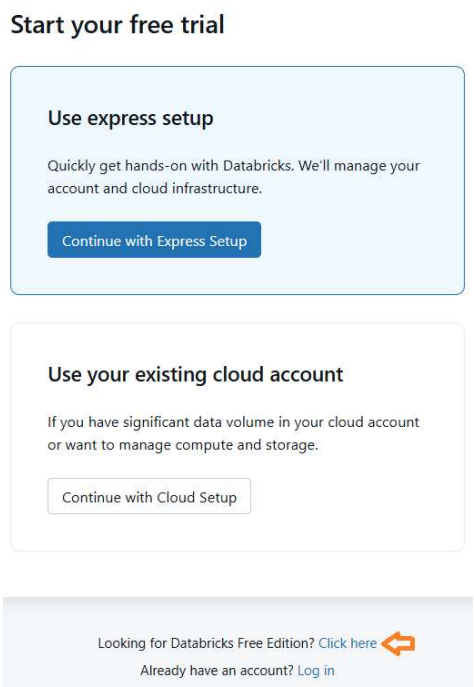
1. **Signup to Databricks Free Edition**

    a. You can sign-up to Databricks free edition **using any valid email id**.

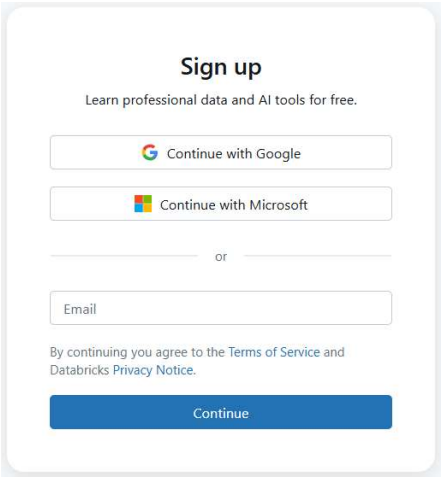
    b. Copy and paste the following link in a browser window.
    https://www.databricks.com/

    c. Click on "**Try Databricks**" button at the top-right corner.

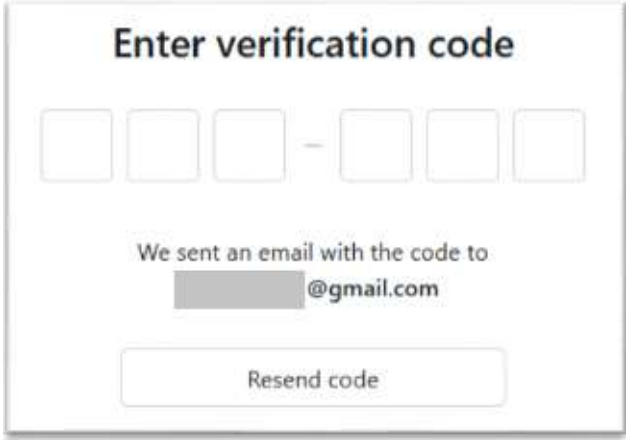
    d. Click on "**Click here**" link for the **Databricks Free Edition** near the bottom of the page.

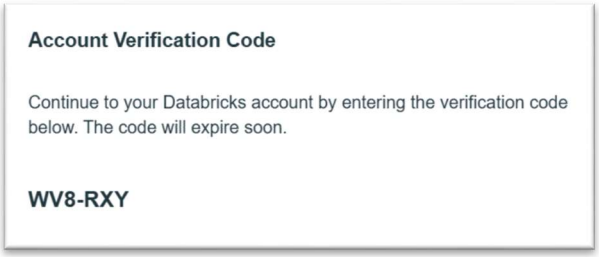


e. Provide a **valid email id** and click on **Continue** button. This is the email you use to login to the Databricks free edition.



f. You receive a verification code to the email. Paste the code to login to Databricks free edition.

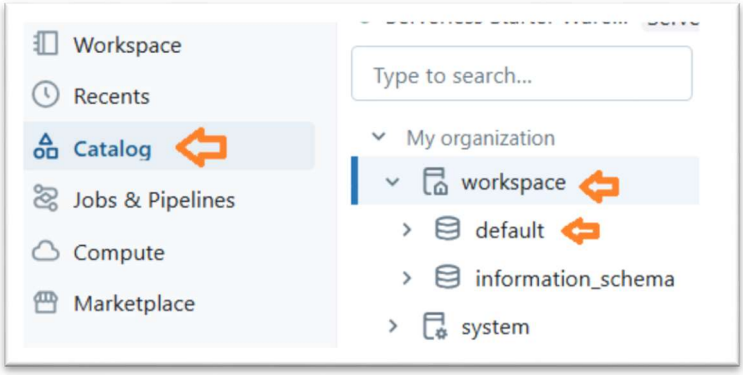




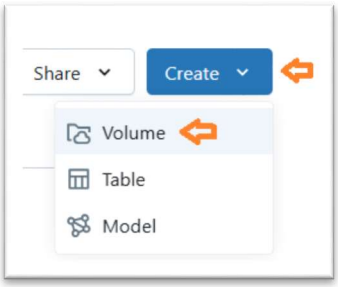
g. This will log you in to the Databricks Free Edition (for the first time). After this, for future logins you may use the next step.

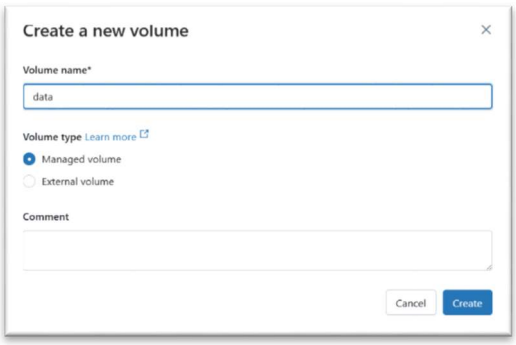
2. **Login to Databricks Free Edition**
   a. Once you sign-up, you can login to the Databricks free edition using the following link https://login.databricks.com/
   b. You receive a versification code for validation to the email. Complete the verification to login.

3. **Create a volume and upload your datasets to DBFS.**
   a. Login to the Databricks portal
   b. Click on the Catalog link from the left-side menu.
   c. You are by default connected to a catalog called "**workspace**". Select this workspace catalog. Select **default** database inside the catalog.

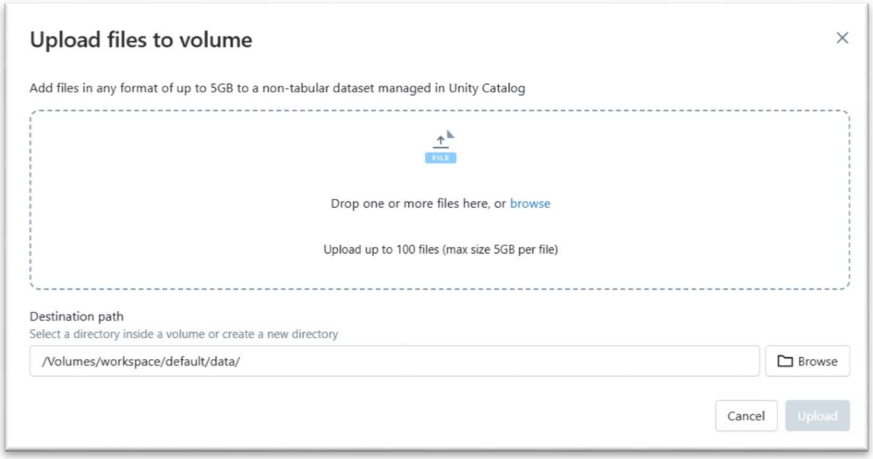


d. Click on "**Create**" button and select "**Volume**". Give your volume a name (such as *data*) and select "Managed volume option and click on **Create** button to create the volume.





e. To upload datasets to the volume, click on **Upload to this volume** button. This pops up a window. Drag and drop your files and folders into this window and click on **Upload** button.



Setting up Jupyter Notebooks for PySpark development

1. Install Java 8 or up

   - Check your current version of Java by running the following command at a terminal. It should show JDK 1.8.x or up.

     java -version

   - If you are running an older version of Java, then upgrade it by downloading and installing a suitable Java version from the following link. https://www.oracle.com/in/java/technologies/downloads/#jdk24-windows

2. Add JAVA_HOME environment variable

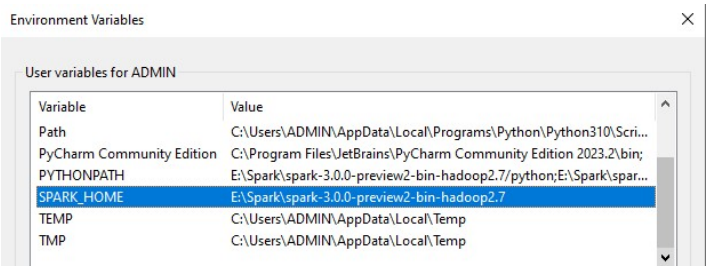
   - Go to 'Edit system environment variables' windows in your windows OS
   - Add **JAVA_HOME** environment variable with the path were Java is installed.

3. Download and extract Spark binaries.

    - URL: https://spark.apache.org/downloads.html
    - Choose the version
    - Download spark: <click on this link>
    - Go to the mirror site and download
    - Extract the downloaded file in a suitable folder

4. Add SPARK_HOME environment variable

    - Go to 'Edit system environment variables' windows
    - Add **SPARK_HOME** environment variable with the path were Spark is extracted.



5. Setup Hadoop **winutils** for windows

    - URL: https://github.com/cdarlint/winutils
    - Download the **winutils.exe** file
    - Copy it to the **bin** folder of your spark directory.

6. Add HADOOP_HOME environment variable

    - Go to 'Edit system environment variables' windows
    - Add **HADOOP_HOME** environment variable with the same path as that of SPARK_HOUME

7. Add the "bin" folders of the above to the **PATH environment variable**

    - %JAVA_HOME%\bin
    - %SPARK_HOME%\bin
    - %HADOOP_HOME%\bin

8. Open a command terminal and type the command **spark-shell.** This should launch the Spark Scala shell.

9. Download and install Python (3 or above) if you do not already have it.

10. Open a Python terminal and install the following tools – **findspark** and **Jupyter notebook**

    a. pip install findspark
    b. pip install Jupyter notebook

11. Open Jupyter notebook by opening a python terminal and typing the command **jupyter notebook**