Installing Apache Spark

Via docker for Windows, MacOS and Linux

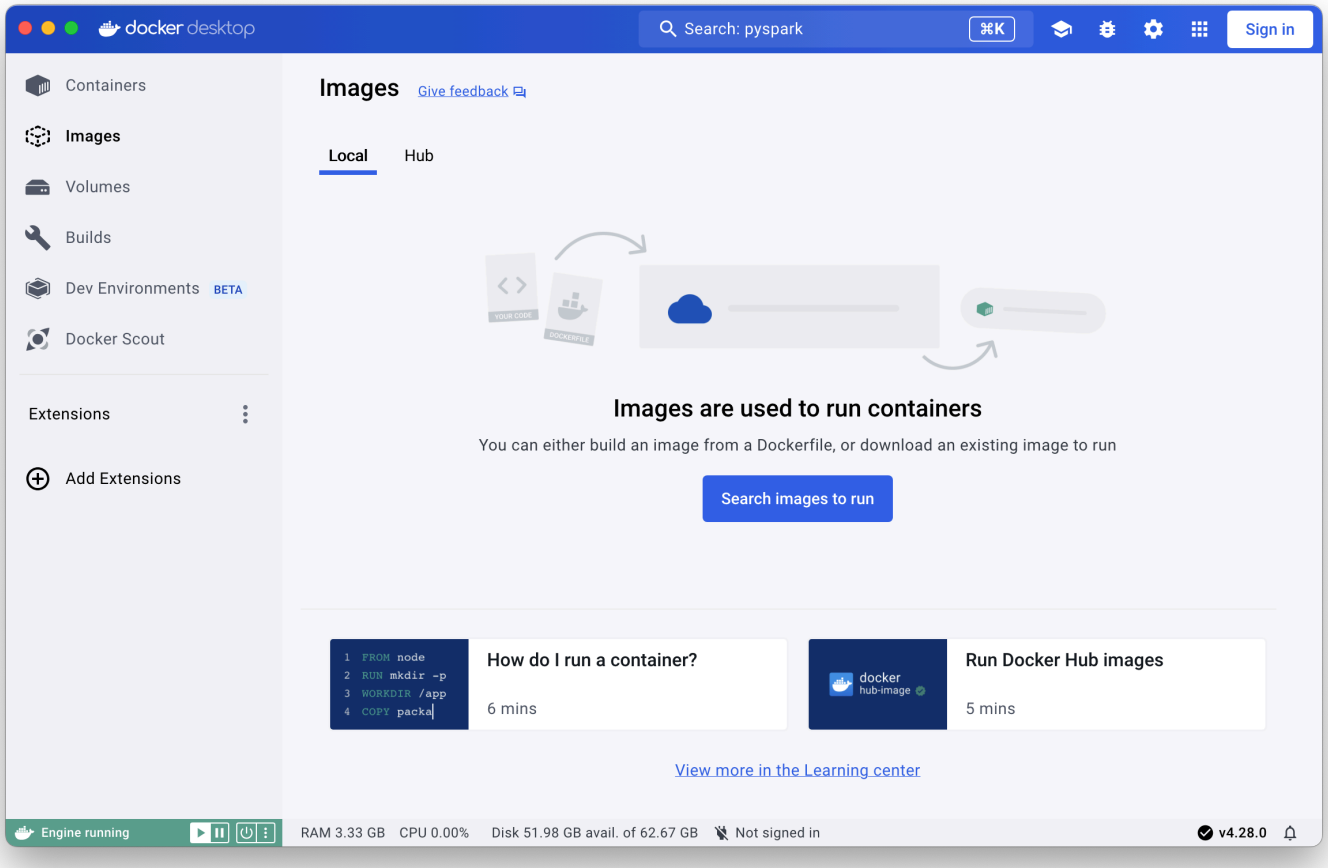
João Oliveira and Adriano Lopes 2024/2025

Installation for all operating systems

- To run Apache Spark in any operating system (windows, macOS and Linux) we recommend you to install the following:
  - Install docker desktop
  - Install visual studio code (VS code)
  - Install VS code Remote Development pack

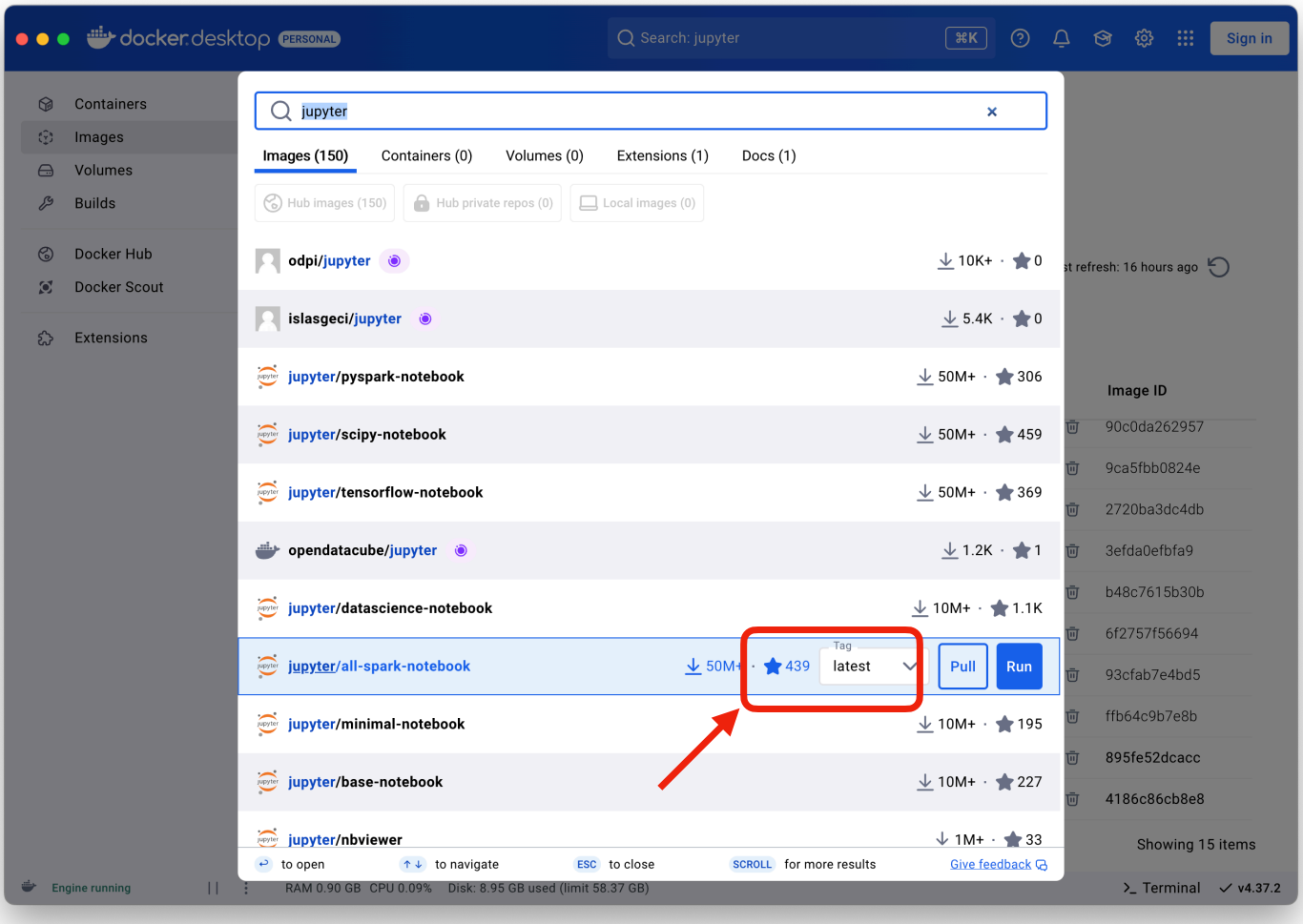
Install docker

- Go to docker desktop webpage, download the correct version for your operating system and install it (mac users must choose the correct version: Intel vs Apple Silicon chip)
- After correctly installed you should be able to open docker desktop



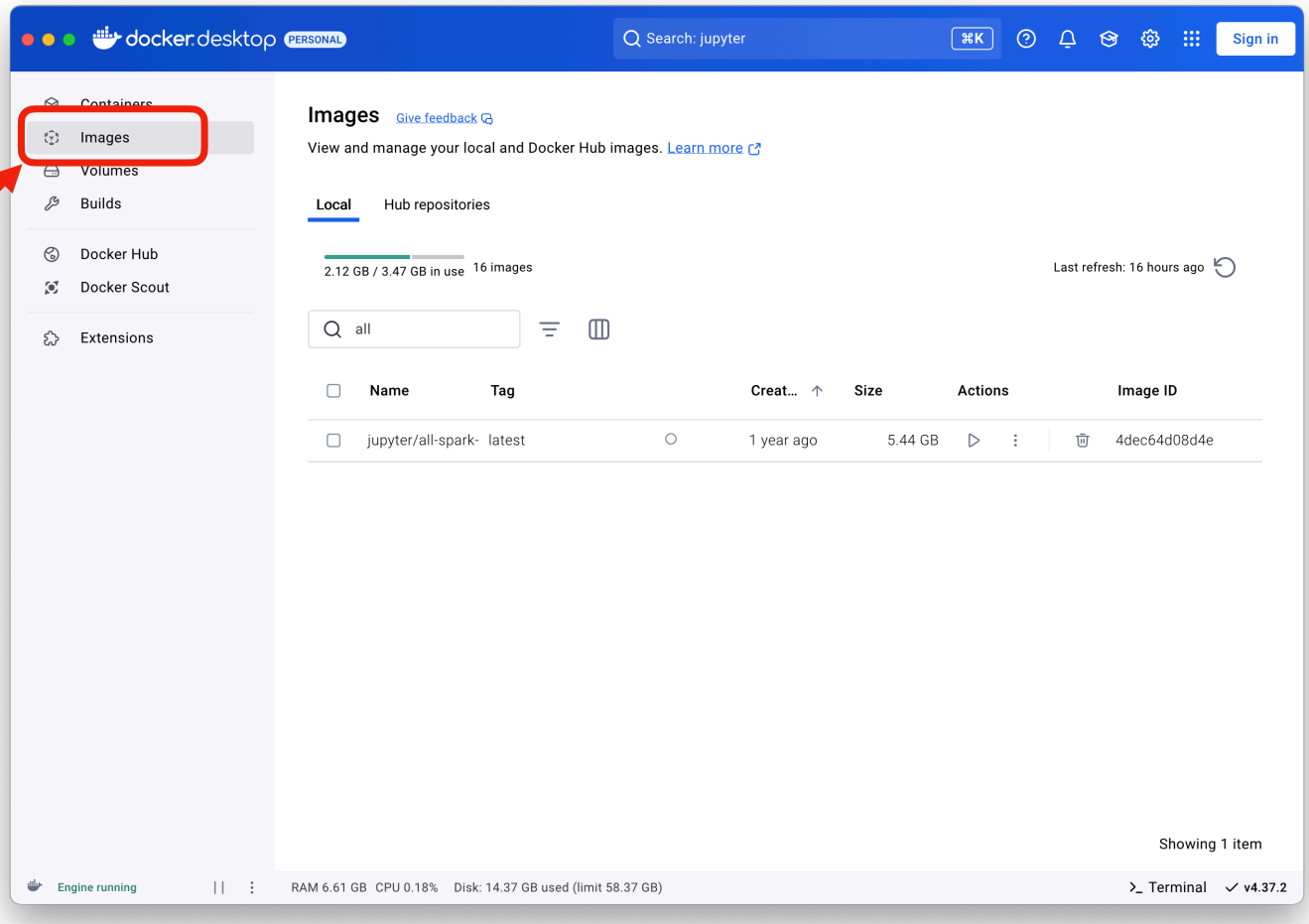
Download docker image

- On the top of you docker desktop search for `jupyter`
- Select jupyter/all-spark-notebook
- Choose the **latest** or **x86_64-latest** Tag and click Pull



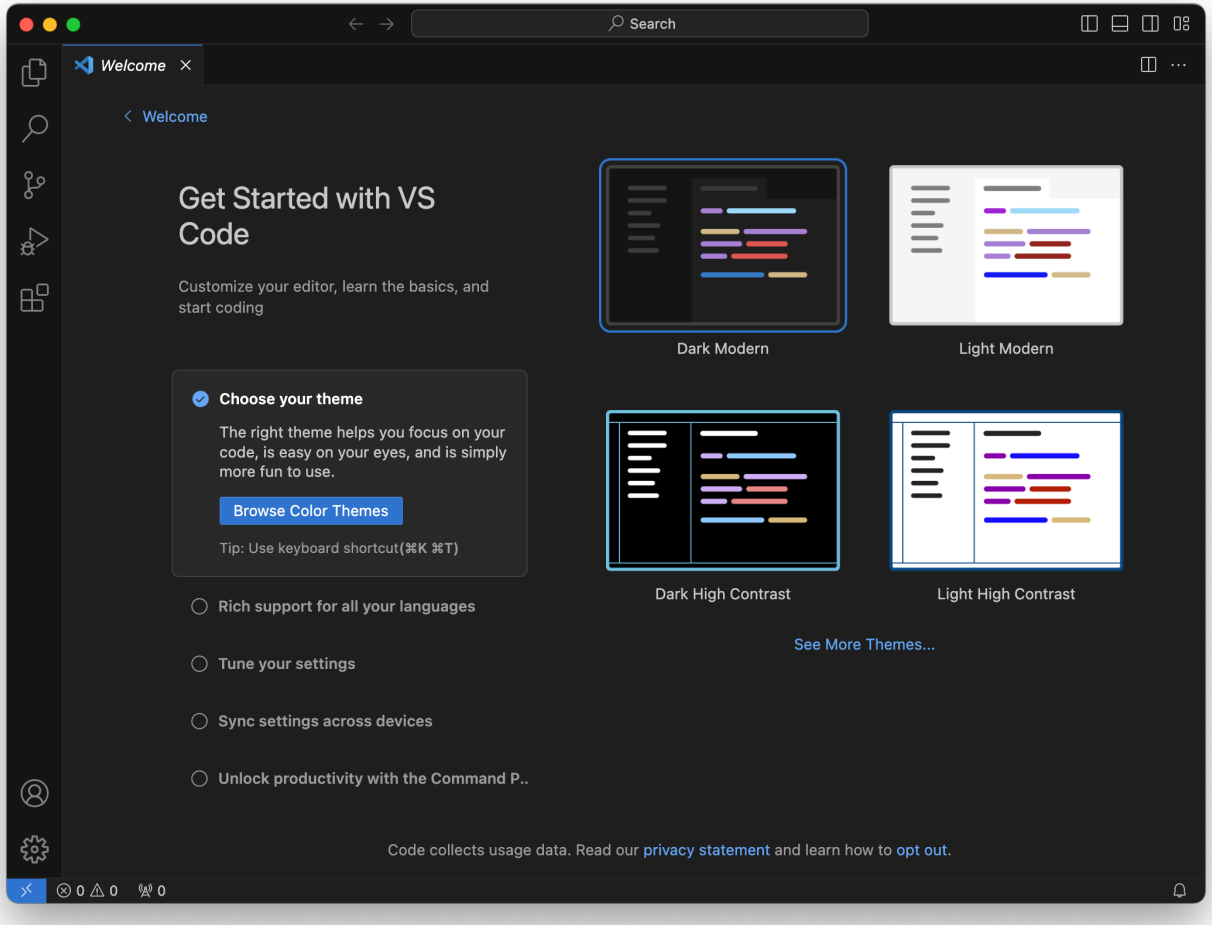
Confirm docker image

- Confirm that jupyer/all-spark-notebook image is under images tab



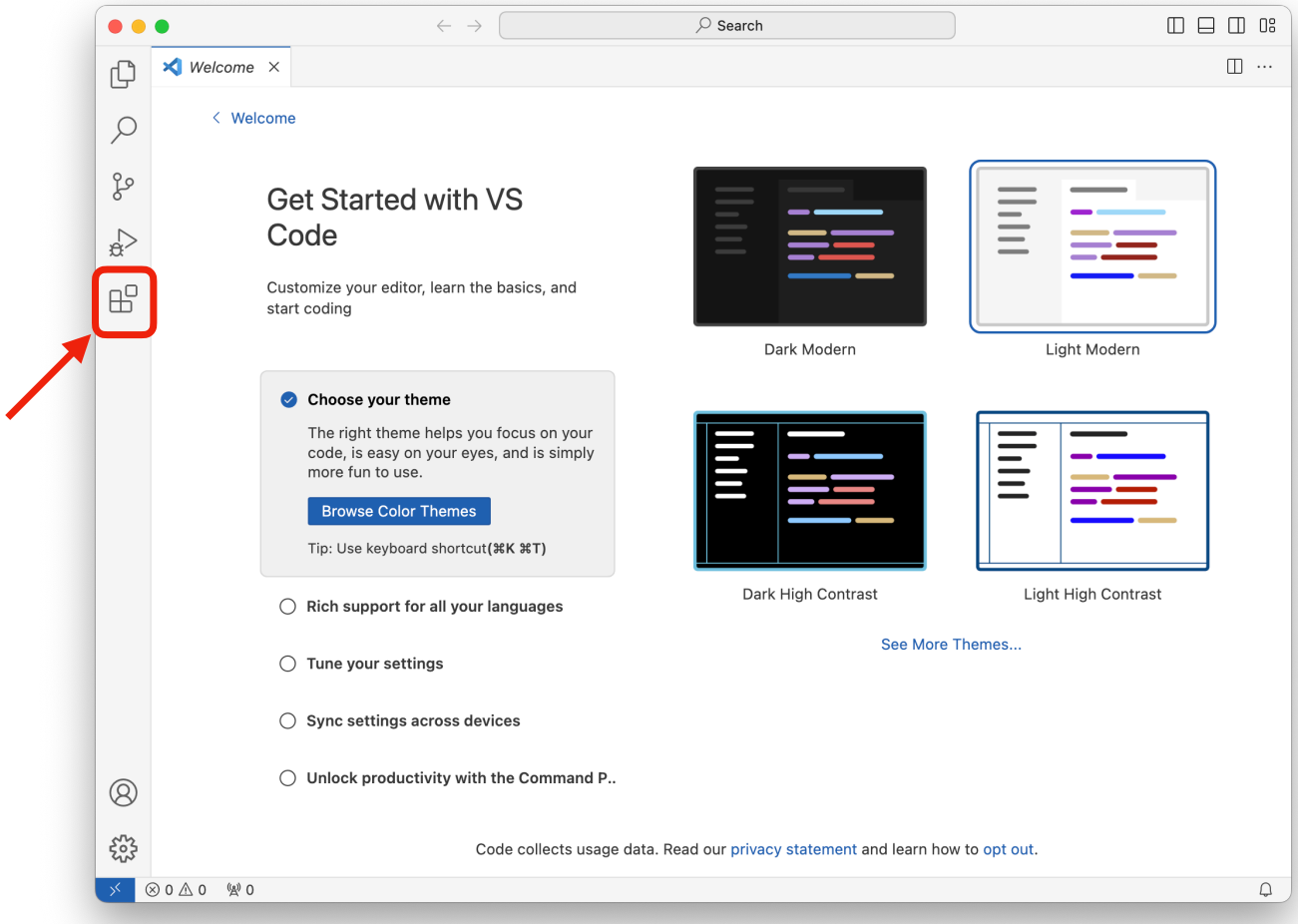
Install vscode

- Go to visual studio code download webpage and install the correct version for your operating system (mac users must choose the correct version: intel vs apple chip or universal)
- After correctly installed you should be able to open visual studio code



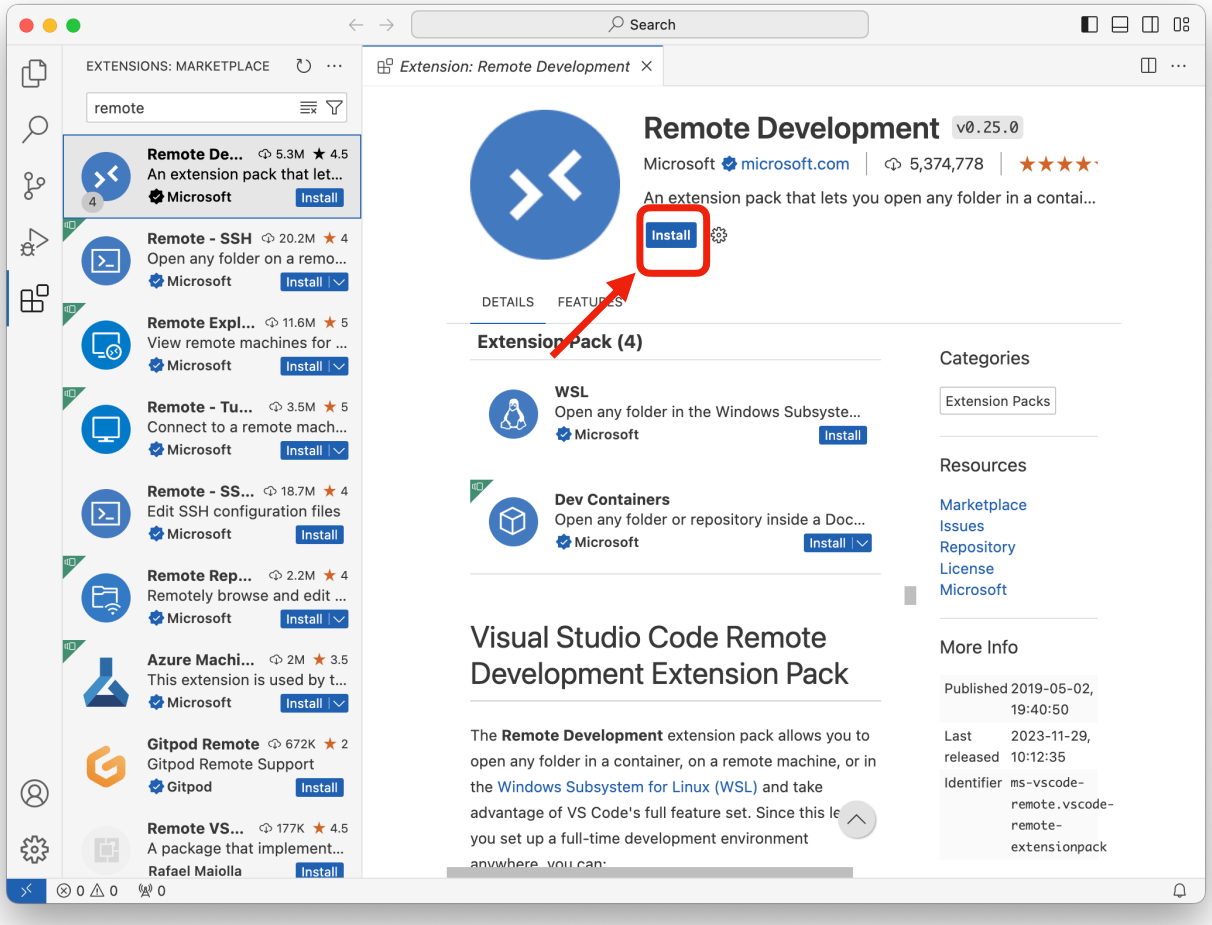
Install VS code Remote Development pack

- After you choose your theme, go the extension tab on the left and search for remote



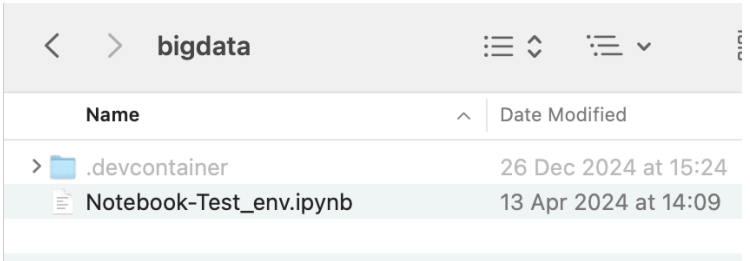
Install VS code Remote Development pack

- Choose Remote Development from Microsoft in the list of extensions and click Install



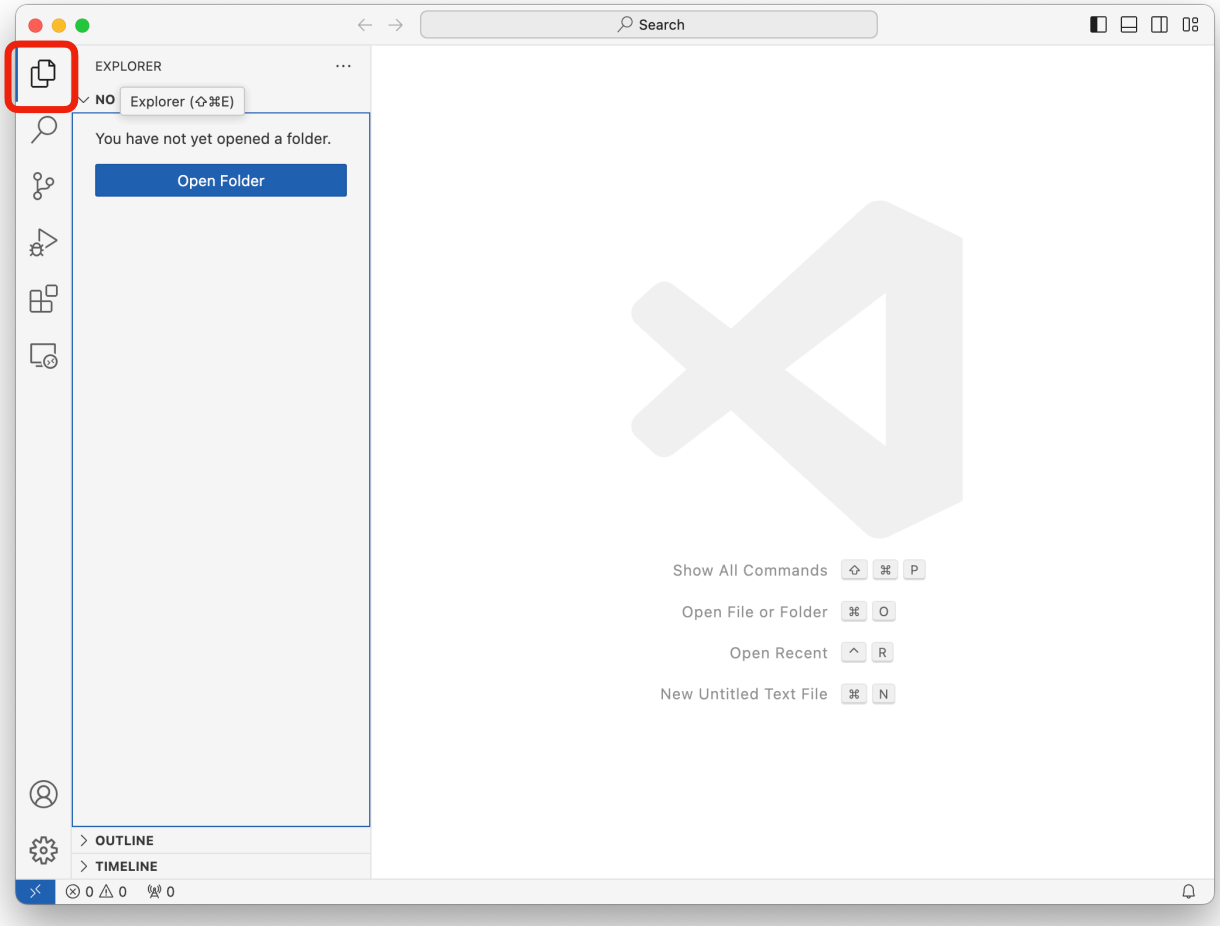
Create a local working directory

- Create a local directory on your computer for the classes
- Download VS code configuration archive (devcontainer.zip) from https://bigdata.iscte-iul.eu/common/devcontainer.zip
- Extract it inside your chosen directory. Ultimately you get a hidden directory .devcontainer alongside a notebook for testing purposes
  - Note for Mac users: to see hidden files (files that start with a dot .) in Finder, use the shortcut ⌘⇧. (Cmd + shift + dot) to enable/disable this feature
- Go back to VS code



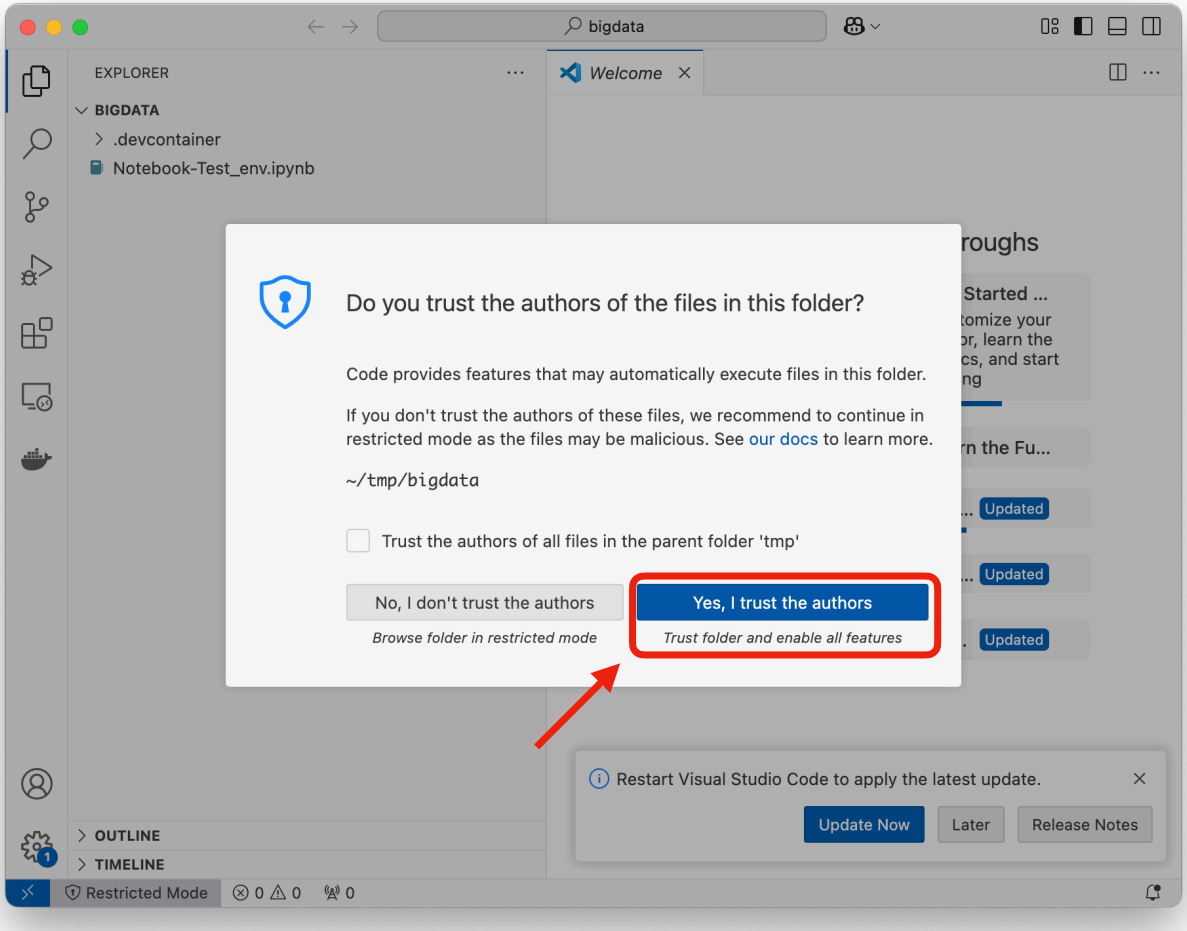
Open a folder in VS code

- On the left tab, in Explorer, click Open Folder
  - Alternatively on File Menu, click Open Folder
- Choose the folder containing the hidden directory .devcontainer previously mentioned



Open a folder in VS code

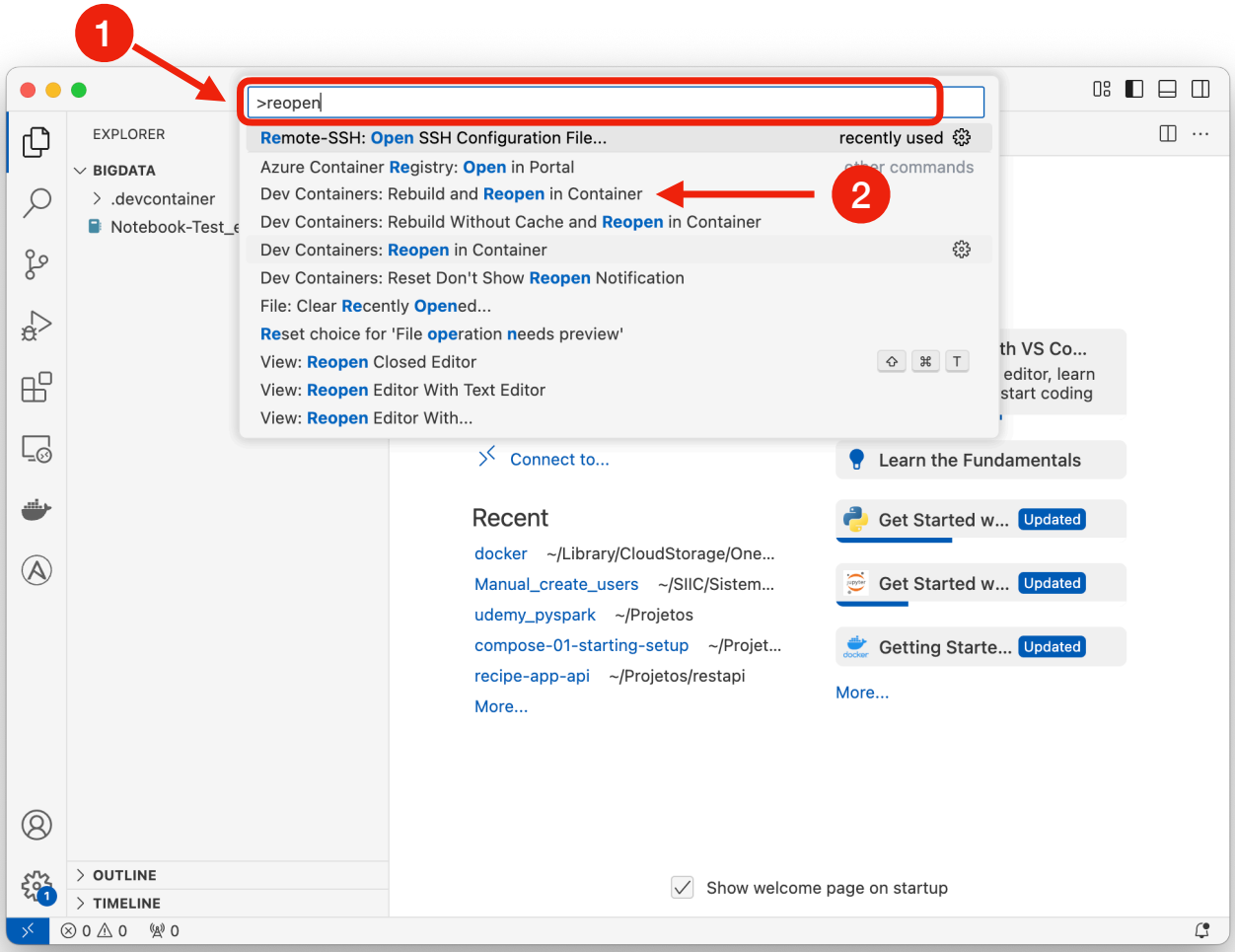
- Trust the folder's content, if that is required



Reopen folder in container

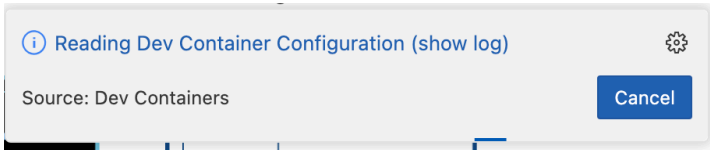
- Open the command palette from the menu's or use a keyboard shortcut:
  - macOS: ⌘⇧P (Cmd+Shift+P)
  - Windows: CTRL⇧P (Ctrl+Shift+P)
- Type reopen and then choose:

  Dev Containers: Rebuild and Reopen in Container

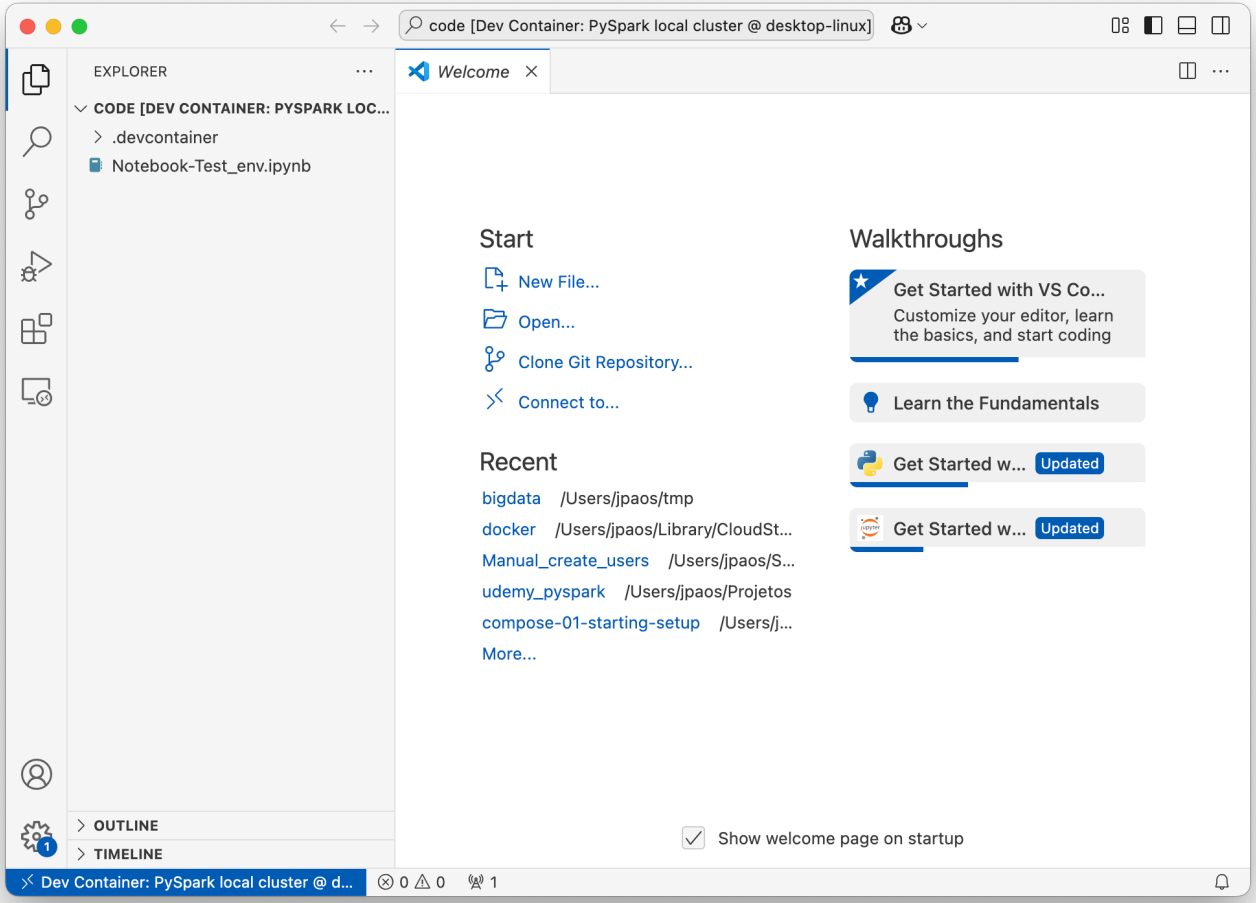


Reopen folder in container

- Wait until everything installs
- You can watch the progress by clicking on Show Log on the bottom right dialog box:

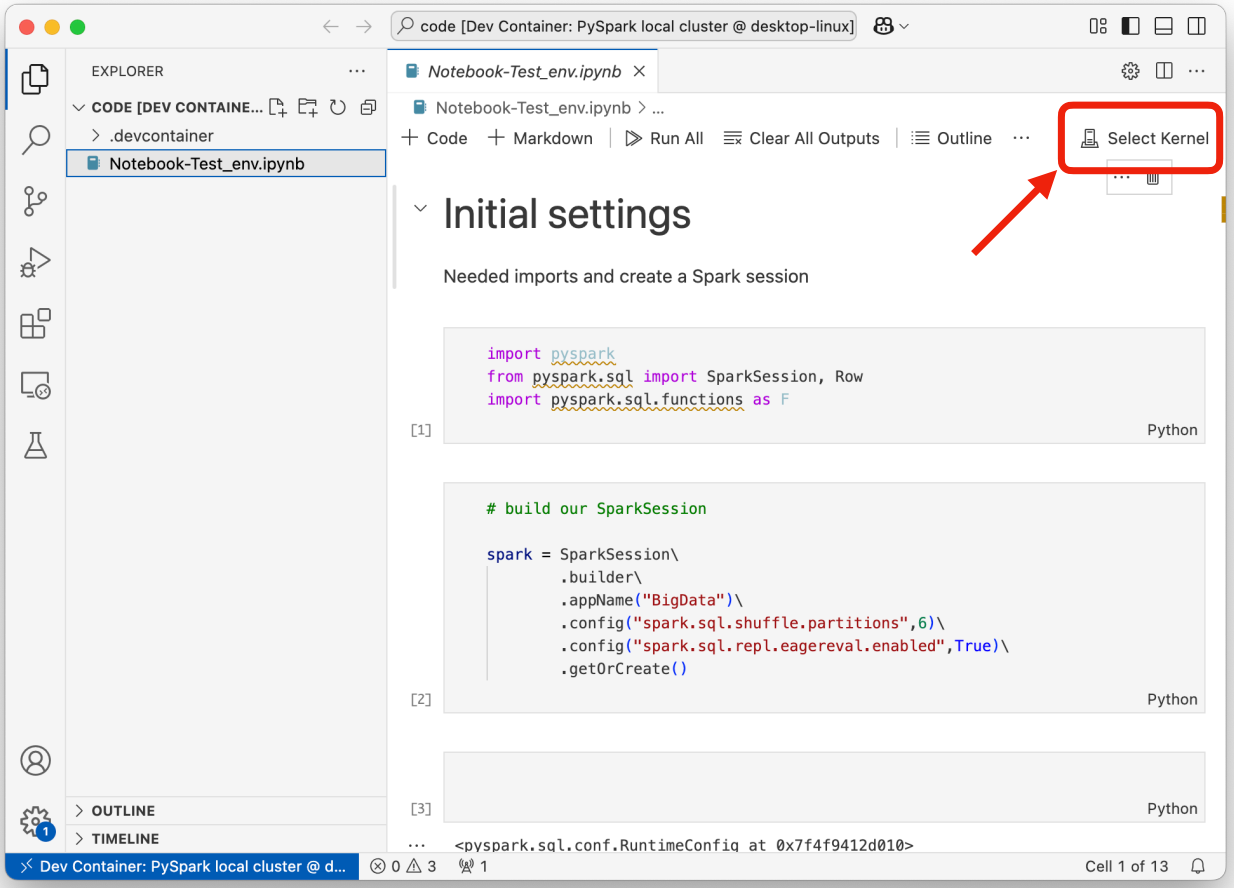


- At the end you should see a screen similar to the one on the right



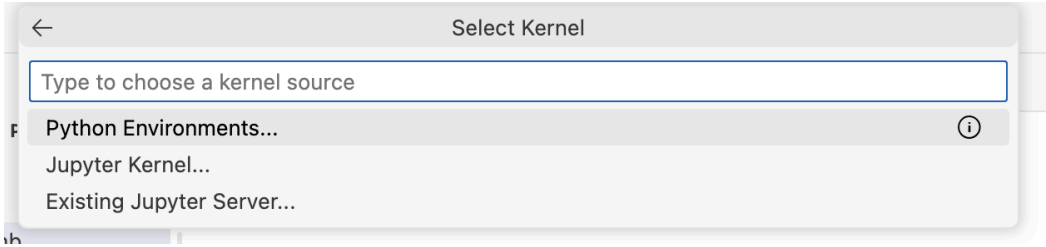
Test installation

- On the explorer click on the file Notebook-Test_env.ipynb
- On the top right of the notebook, click on Select Kernel (you should wait at most 30s for everything to be ready)

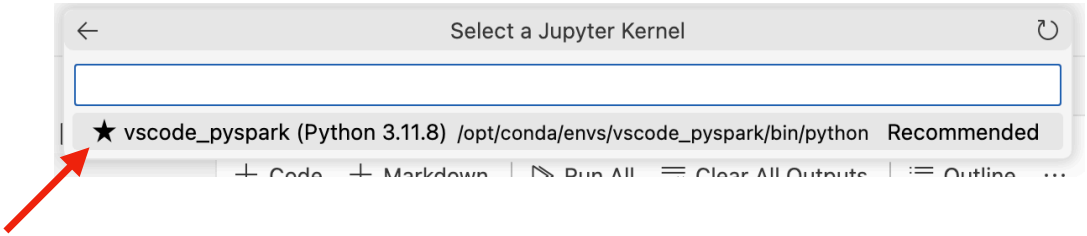


Test installation

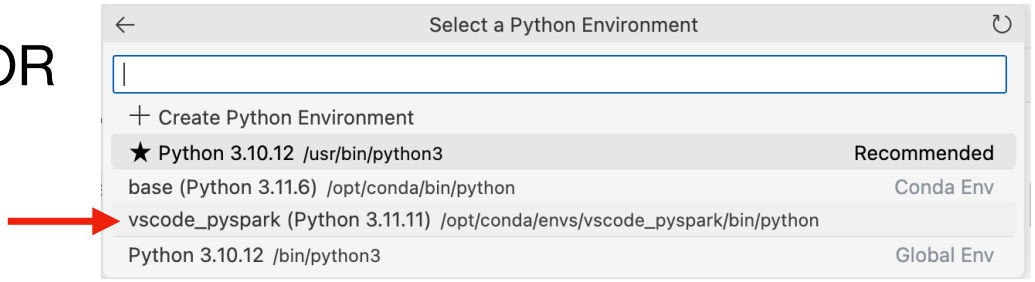
- Choose Jupyter Kernel…



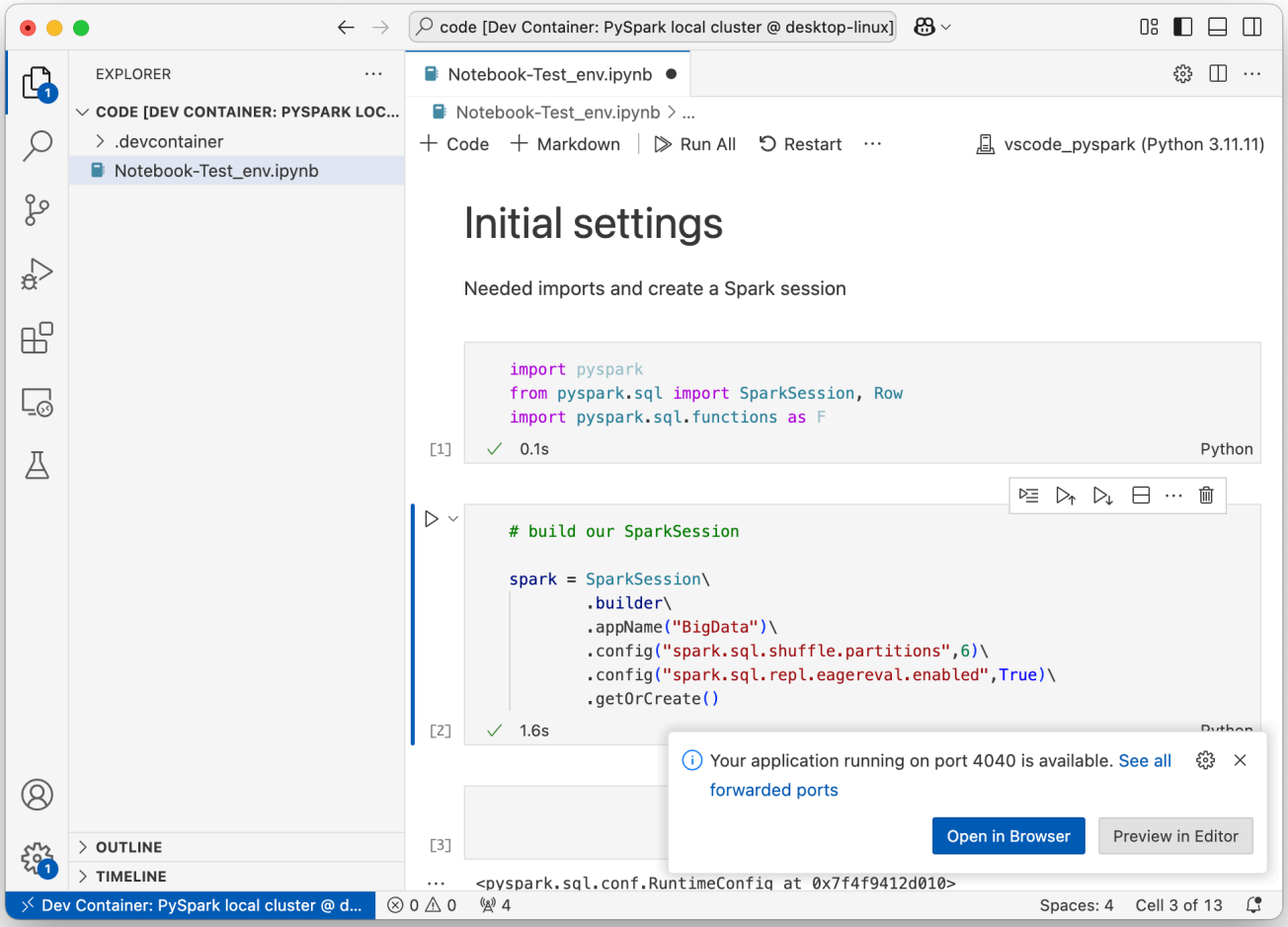
- Choose **vscode_pyspak** (depending on your installation)



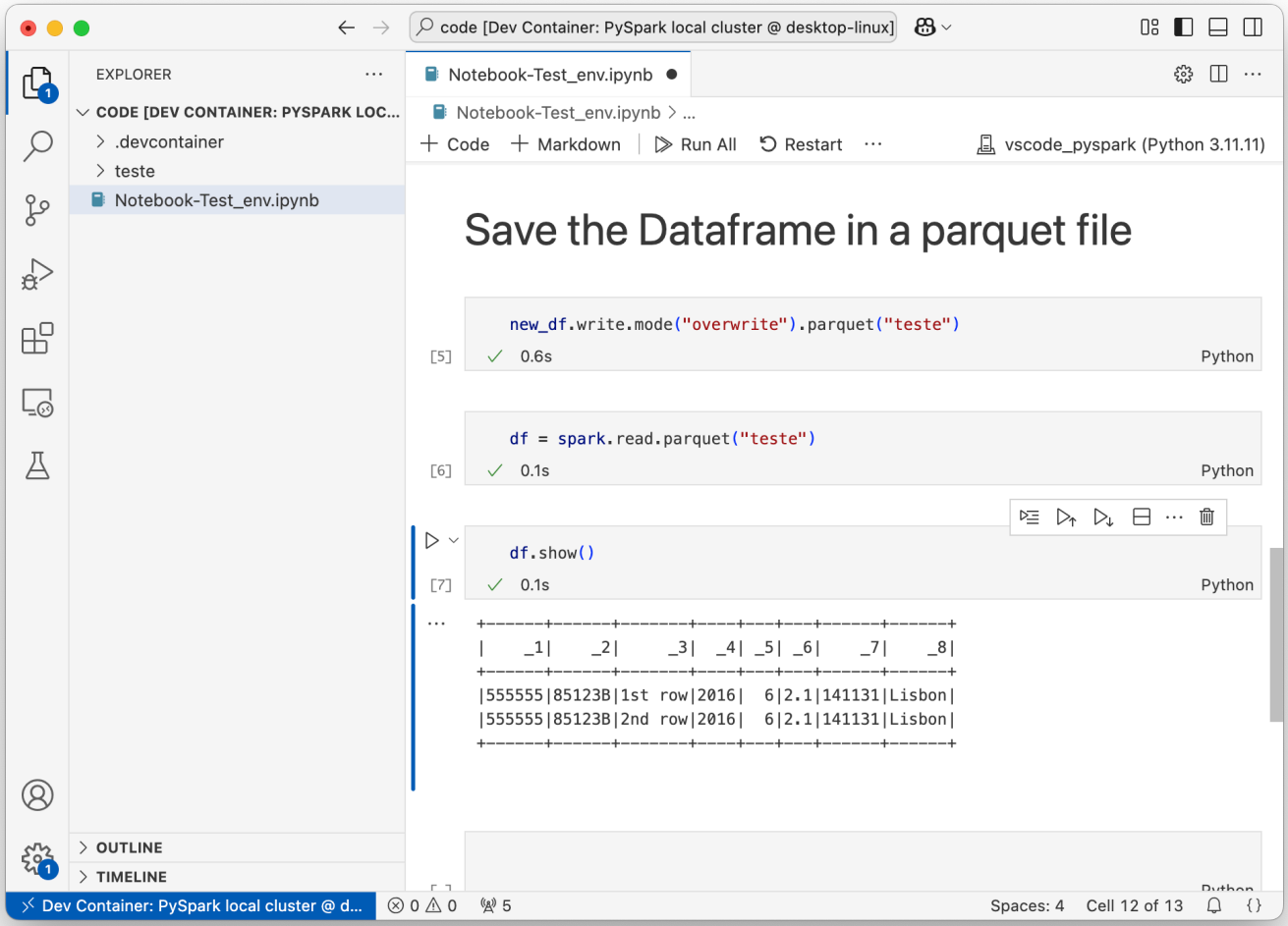
OR



Run the notebook



Run all the cells until the end



Test is done