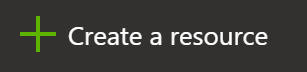
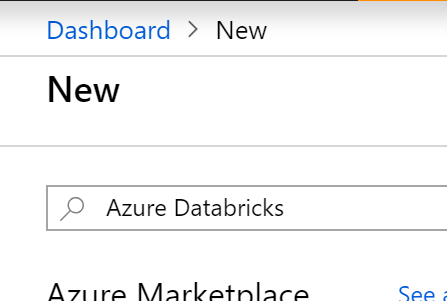
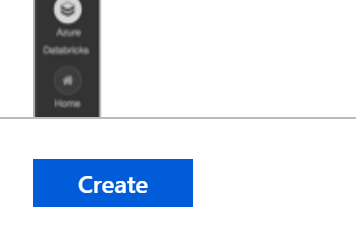
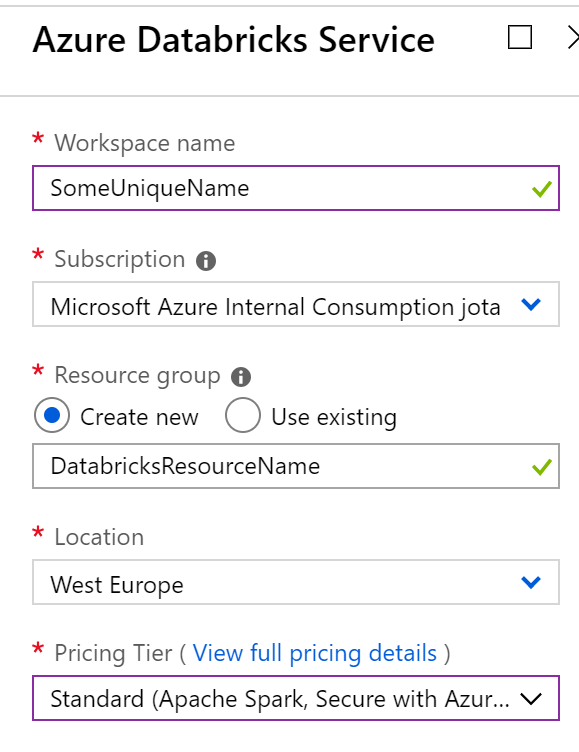
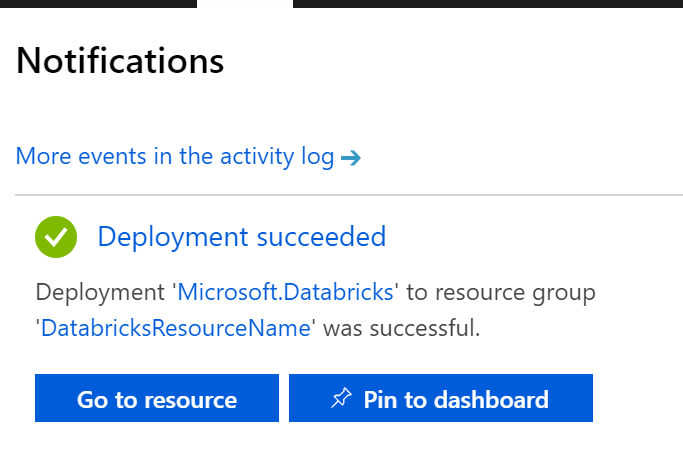
LearnAI – Hands-on with Databricks

Setup and basics

# Create an Azure Databricks workspace

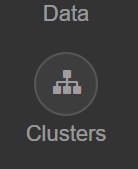
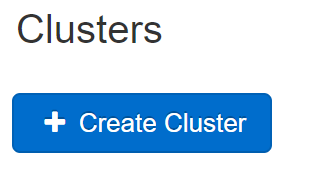
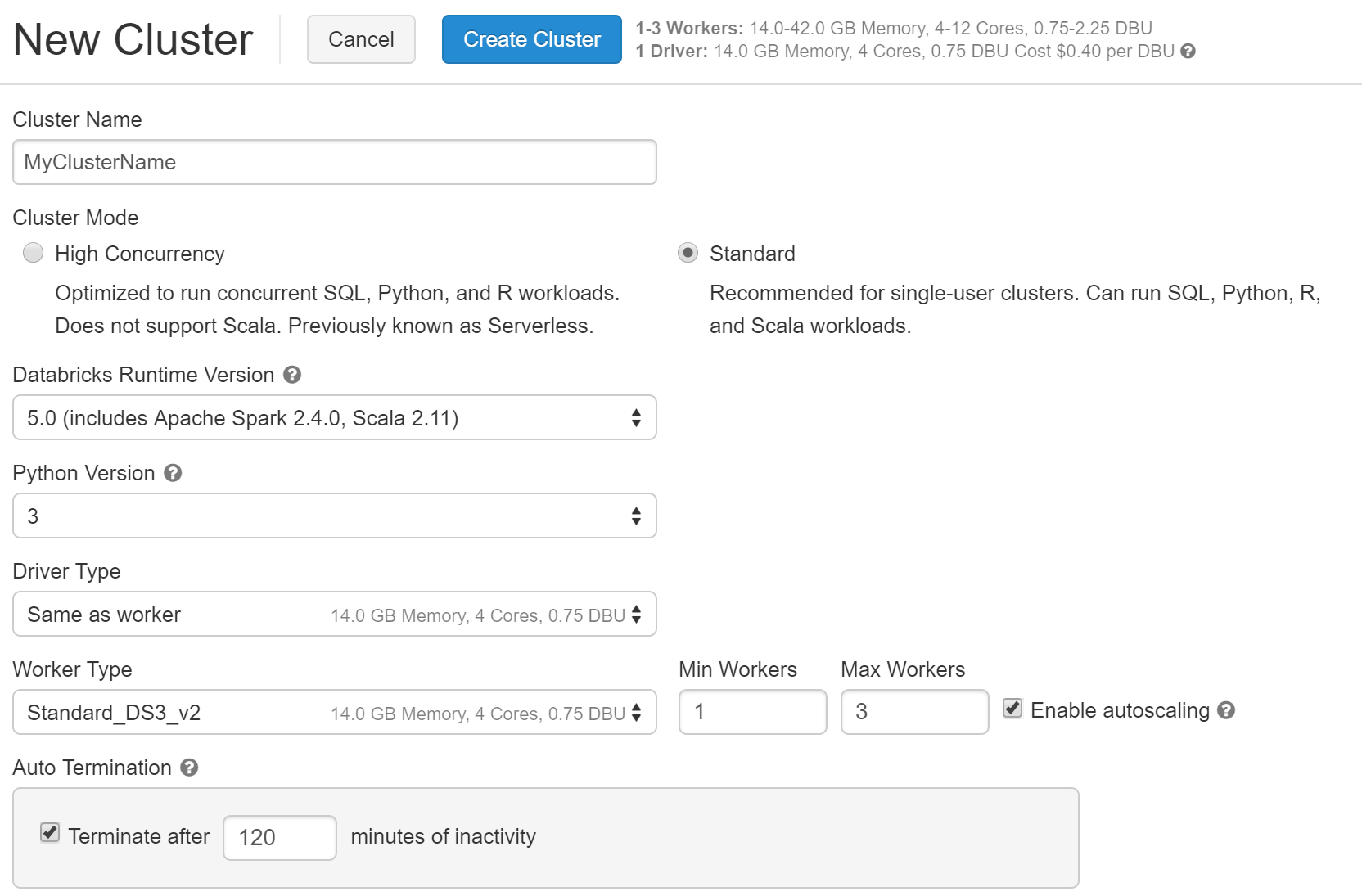
To use Azure Databricks, the first thing to do is to create a workspace. The Workspace is the special root folder for all of your organization’s Azure Databricks assets. The Workspace stores all your notebooks, libraries, and dashboards. By default, the Workspace and all its contents are available to 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.

Do the following to create your first workspace:

1. Open the azure Portal: portal.azure.com
2. Click Create a resource in the left column  
   
3. Write “databricks” in the search box and select Azure Databricks  
   
4. In the end of the information page that shows up, click Create:  
   
5. Fill in the basic data in the mini-form. Pick your own unique names and hit Create.  
   
6. Deployment of the Workspace should take about 2 minutes. When it’s over, head over to the Workspace, for example by clicking “Go to resource” in the notification:  
   
7. To access your Databricks Workspace, now click “Launch Workspace”:  
   
8. Another browser window will open, and you’ll be automatically signed in to the cluster.

# Create an Azure Databricks cluster

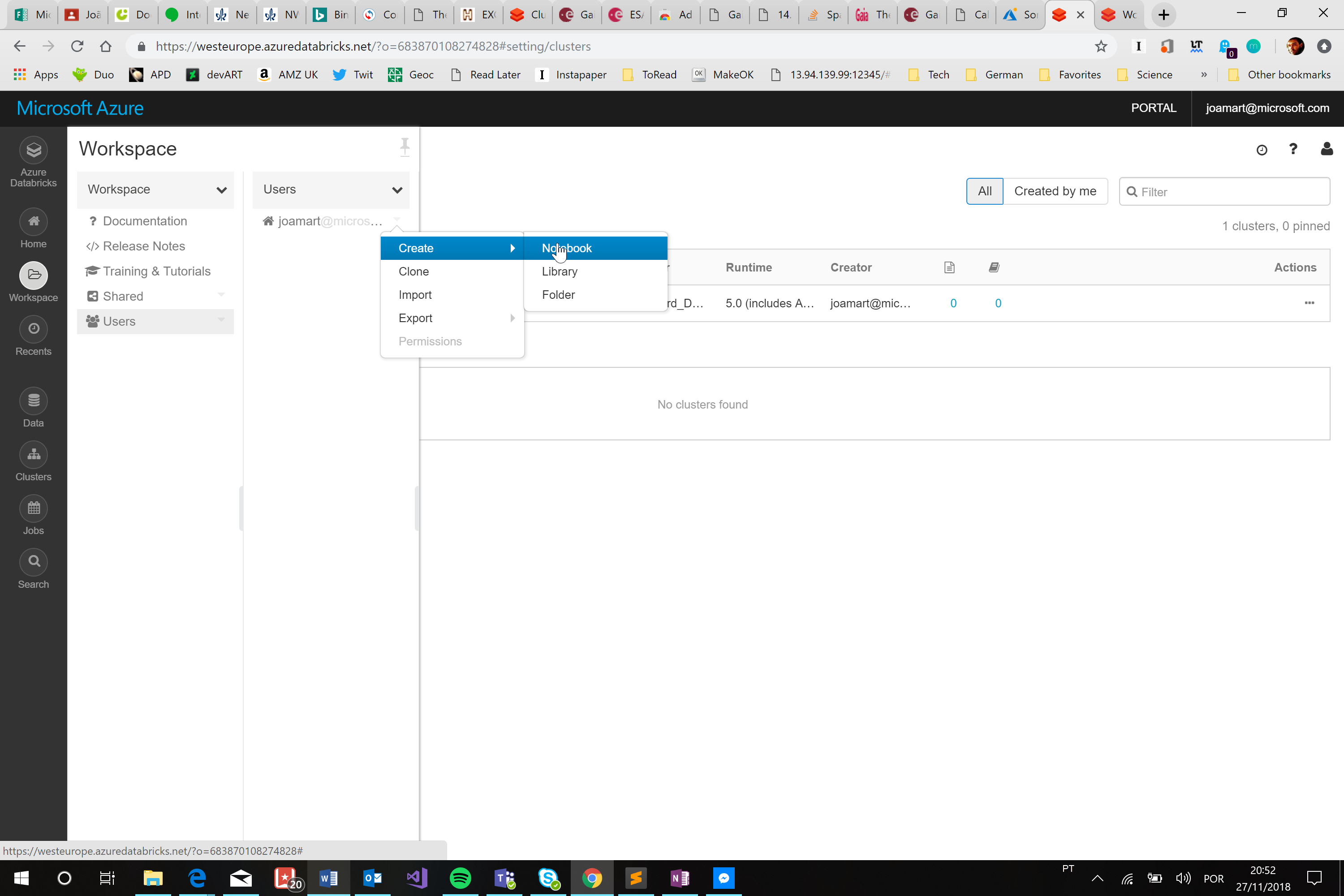
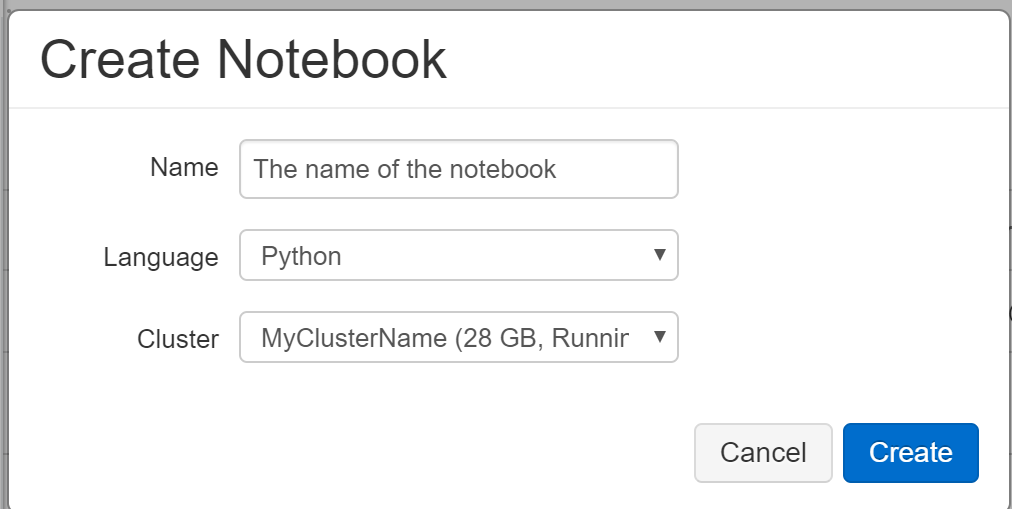
The workspace by itself allows you to store your artefacts but doesn’t include the any servers on which to run Spark. To do this you must create a cluster. To do this:

1. Access the Clusters option:  
   
2. Click Create Cluster:  
   
3. Fill in the form as in the following screenshot:  
     
   **Note: please pay attention to important fields which have different values than the defaults: Databricks Runtime Version, Python Version, Max Workers.**Also note that by default your cluster will be stopped after 120 minutes of inactivity, which avoids ramping up Databricks costs even if you forget to turn it off.
4. Creating a cluster like this takes 5 minutes. You’re now ready to create your first notebook.

# Create your first Databricks notebook

Databricks Notebooks are similar to Jupyter notebooks, but focused on offering a front-end to the Spark engine, which no longer needs to be programmed from the command line, and supporting multiple languages like R, Scala, Python or SQL.

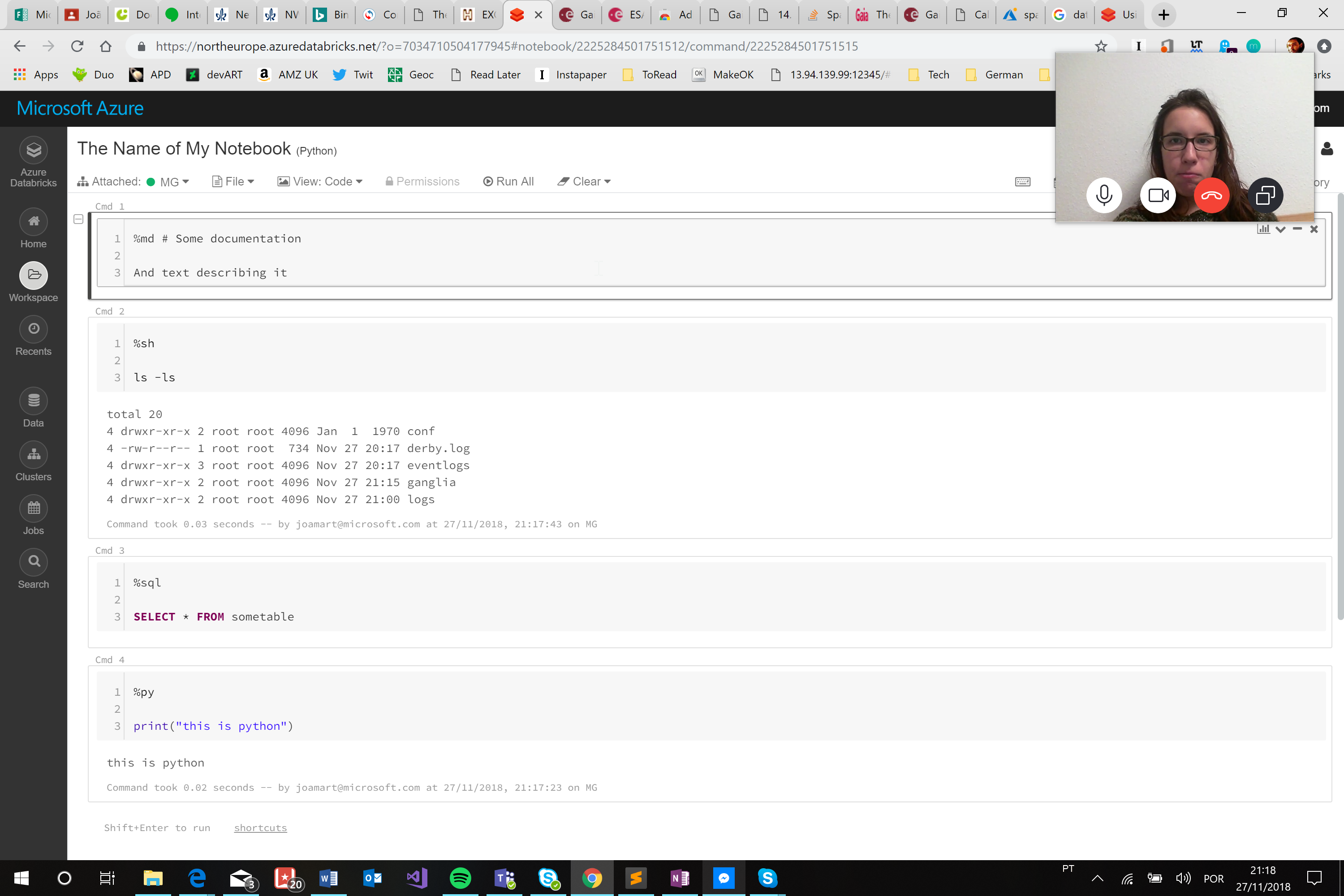
To create a notebook:

1. Click Workspace, then Users, then the little gray arrow next to your username, then Create > Notebook:  
   
2. Pick the name and default language and click Create  
   
3. The notebook is ready for use:  
   

Note: by default, any code you write in this notebook will be executed as Python. If you want to use other contents you can use the magic commands, by putting them right at the top of the cells, as shown in the following image. For this challenge the relevant magic commands are:

* **%py** - cell is in python language. As this is the default language of the Notebook we created, it can be omitted
* **%md** - cell is in markdown, a simplified HTML-like language used to write documentation (as in GitHub)
* **%sh** - cell includes Linux shell commands that execute directly in the underlying VM
* **%sql** - cell includes SQL commands.

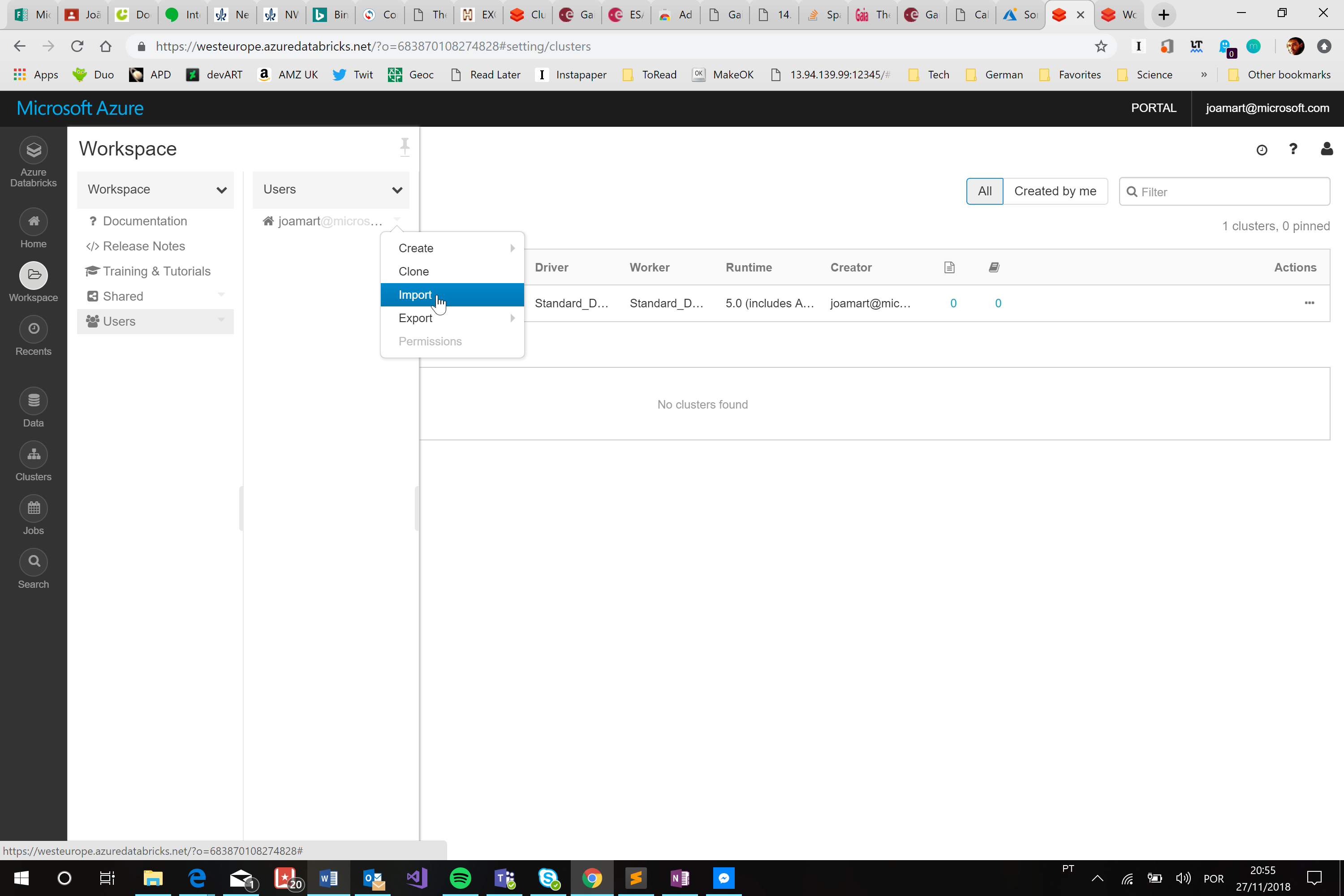
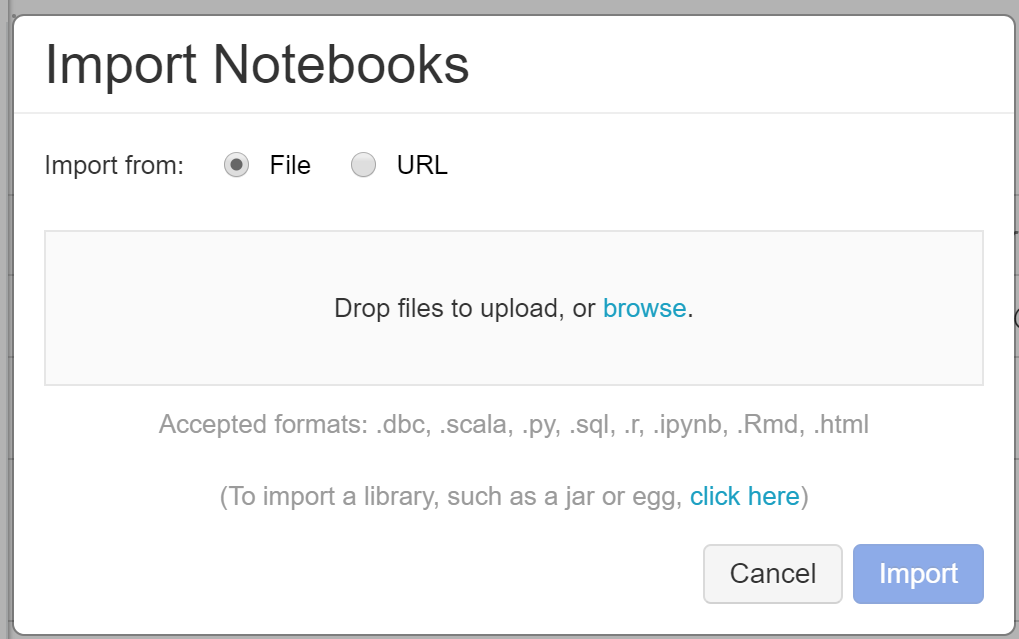
If it’s your first time using Databricks Notebooks, reading <https://docs.databricks.com/user-guide/notebooks/notebook-use.html> is highly recommended. Also note that Databricks Notebooks and Jupyter Notebooks are not exactly the same.



One additional note is that Databricks clusters come pre-installed with many popular libraries. In the case of Python, the list includes packages such as Pandas, numpy, scikit-learn matplotlib and ggplot. Full list here: <https://docs.databricks.com/release-notes/runtime/5.0.html#installed-python-libraries>

Finally, another relevant note is that Databricks Notebooks can be scheduled for regular execution, using the “Jobs” option in the left menu bar of the workspace.

# Import an existing notebook or collection of notebooks

1. To import an existing notebook or collection of notebooks, do similar steps to the previous, but select Import instead of Create:  
   
2. And in the following screen:  
   

The .dbc format is a Databricks archive, a package that can contain a folder of notebooks or a single notebook. A Databricks archive is a JAR file with extra metadata and has the extension “dbc”.

**For this course, download a DBC file from** [**https://aka.ms/UkLearnAI**](https://aka.ms/UkLearnAI) **and install it in your cluster.**