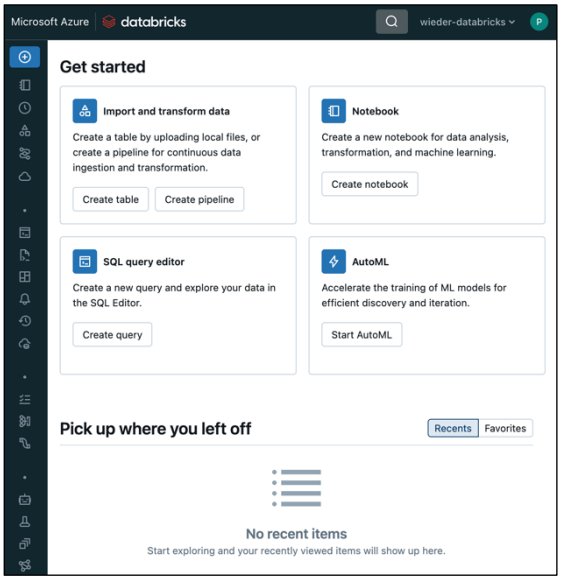
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Data Science Infrastructures – Exercise 07 (DSI E07 ST 24)

07/07/2023 / Philipp Wieder / philipp.wieder@gwdg.de
Related lecture: Data Analytics Platforms II (DSI L08)
Points overall for this exercise: 5 points



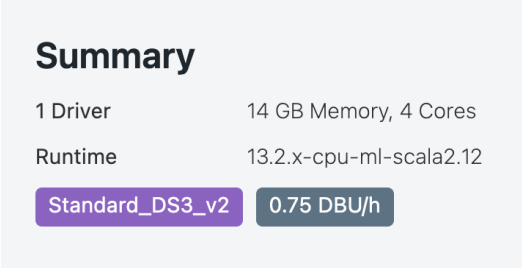
General Prerequisites

As we are using Databricks[1] to try out a few things with this "next generation" analytics platform. To do this, the first thing you need to do is to go to the course-specific URL https://adb-1277333736011036.16.azuredatabricks.net at Azure and start doing the assignment.

Logging in you will see something like in the screenshot to the right. This is the Databricks home page, the entry point to manage your data science projects, your configurations and other things.

In a first step we need to set up a cluster. **Please note that creating a cluster generates actual cost, which have to be paid by GWDG. So please use the resources in a sensible manner and stop the cluster in case they are not in use (e.g. while reading the assignments 😉)**. Click on "New" →"Cluster" and use the following config parameters:

- Policy: "Personal Compute"
- Access mode: "Single user"
- Node type: "Standard_DS3_v2"
- Terminate after: Keep it short, best something like 30 minutes.



Clicking "Create cluster" generates and **starts the cluster directly**. So please terminate it in case you are not using it directly. You can start/terminate the cluster via the "Compute" menu item using "Start" or "Terminate" (fyi: "Terminate" means stopping; you can also "Delete" a cluster).

You can play around with the parameters and you will see varying cost depending on the setting (Databricks uses its own cost units DBU/h). Please do not start a cluster other than the one configured as shown above to keep cost on a moderate level. You can start the cluster by clicking on

[1] https://www.databricks.com/

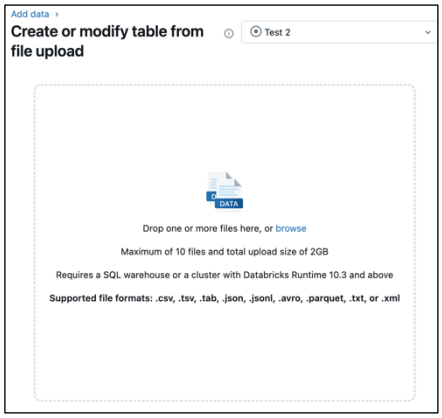
"Start" in the workspace's top right, but this is not necessary now. Stopping the cluster is done by clicking on "Terminate", which does only stop the execution, but allows you to restart it again.

Assignment I – Getting data into Databricks

We start with importing some example basic data. Go there via "Data Engineering" → "Data Ingest" (you can select it from the left menu) and select "Create or modify tabel". There are also other ways to get data into Databricks (like e.g. selecting "New" → "Data" from the menu at the left).

Points: 2

Assignment



You now see the upload window as shown in the screenshot. Take the file *diamonds.csv* from Stud.IP and upload it to Databricks. It contains data about diamonds, their size in carat, their quality and many more features. Make sure that your cluster is actually assigned (drop down at the top right) and do NOT use some Warehouse, which is also offered by Databricks. With "Create table" you finally generate the table. Make a screenshot of the table and provide it as part of the solution. Also use the "Data Explorer" from the left menu to have a closer look at the data and submit a screenshot of the "Columns" as part of the solution. Do not forget to terminate the cluster if it is not in use …

You know have data in the internal Databricks "Lakehouse". In case you are interested to understand how this is implemented you can find the information in the respective Azure documentation[2].

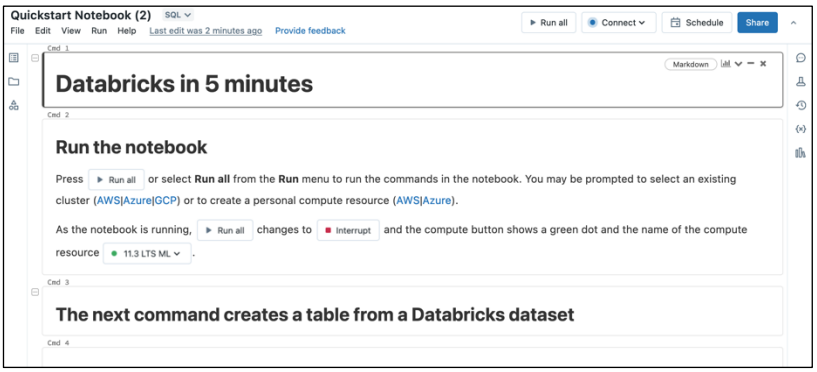
[2] https://learn.microsoft.com/en-us/azure/databricks/lakehouse/data-objects#create-table-ui

Assignment II – Working with Data

Now import a notebook via "New" → "Notebook" → "File" → "Import notebook …". You get the notebook from the following URL: https://www.databricks.com/notebooks/gcp-qs-notebook.html. This will open a pre-filled notebook as you can see at the screenshot. Connect the notebook to your cluster (which will then start again) and you are ready to go.

Assignment (3 points)

The notebook uses the same data as we have imported before, but drops the table in case it exists. You can, of course, also create a second instance of the same data.



Do the following:

1. Run the notebook by executing "Run all". You see messages within the notebook indicating that there is some action on the Spark cluster.
2. Have a look at the different commands. Even without knowing SQL and/or Python, it becomes quickly clear what is going on.
3. Under "The next command manipulates the data and displays the results" there is some Python code. Select the feature "cut" instead of "color" and display the data accordingly. Hand in a screenshot of the changed code and of the result.
4. Use the left menu to go to the cluster and have a look at the Spark UI. Provide screenshots of the timeline and of the last 5 to 10 jobs that have been executed.

Please delete the cluster once you are done.