Data Management and Ethics - Exercises

Axel Oehmichen
ao1011@imperial.ac.uk

Tuesday 8th June 2021

The task is to modify or add to existing Python code in a Jupyter and Spark environment to perform analysis and formatting on an Apache Web log file.

You will be supplied with the log file and a single code file containing instructions on how to complete each step.

Install docker:

https://docs.docker.com/desktop/

Docker desktop supports both MacOS and Windows.

NB for Mac users: if you have the M1 chip, select Mac with Apple Chip. If not (e.g. Intel users), please select Mac with Intel chip.

Download and start the docker container:

You need to download the necessary container with:

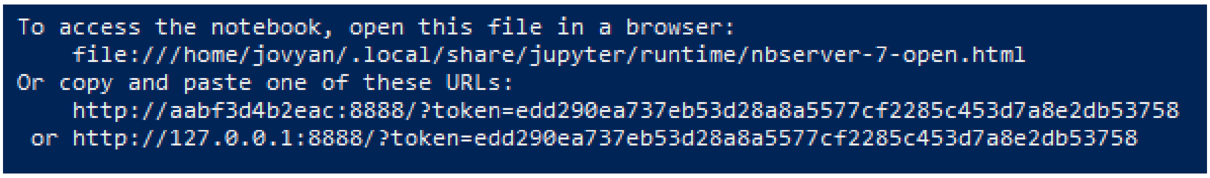
```
docker pull jupyter/pyspark-notebook
```

you can check the version with *docker images* once it is downloaded.

To start the container:

```
docker run -p 8888:8888 jupyter/pyspark-notebook
```

Once the container is ready, you should see something like this in the shell:



```
To access the notebook, open this file in a browser:
    file:///home/jovyan/.local/share/jupyter/runtime/nbserver-7-open.html
Or copy and paste one of these URLs:
    http://aabf3d4b2eac:8888/?token=edd290ea737eb53d28a8a5577cf2285c453d7a8e2db53758
 or http://127.0.0.1:8888/?token=edd290ea737eb53d28a8a5577cf2285c453d7a8e2db53758
```

Copy and paste any of the URLs into your favourite browser.

NB for Windows users: I recommend using PowerShell for those steps.

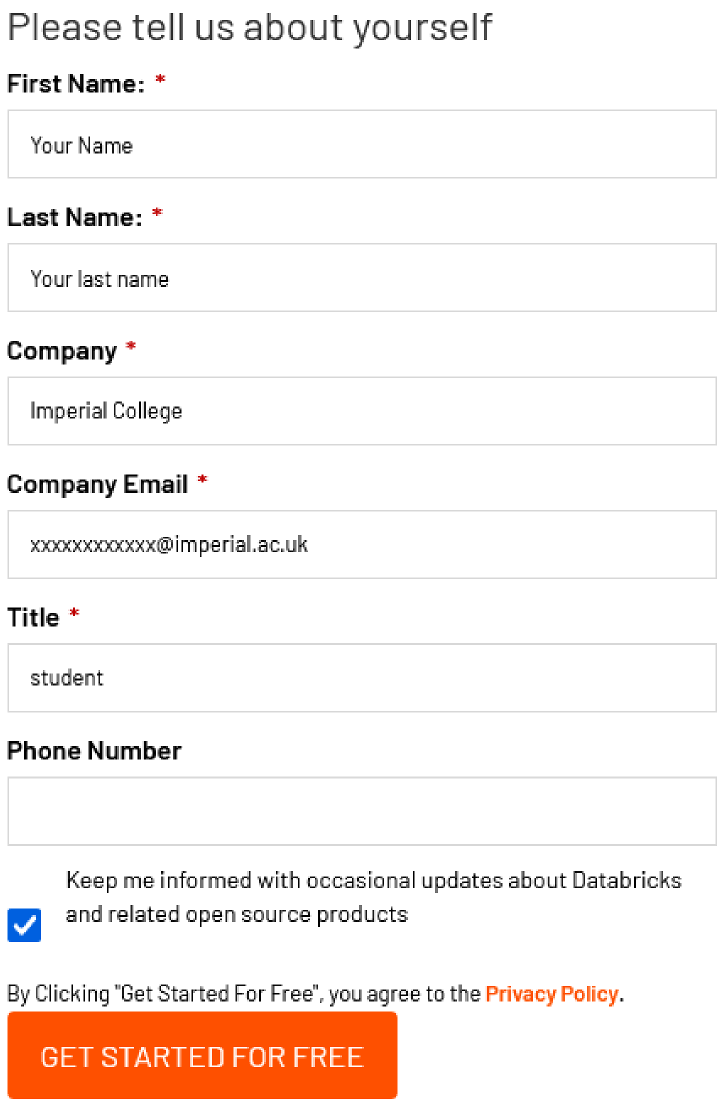
Databricks as an alternative to docker:

You can use an online platform for Spark, but the coursework will be using docker, so it is crucial to have the first option working.

This option allows you to test with a cluster of remote machines to give you some hands-on experience with a system closer to a production environment. The use of this platform is free, but the resource allocation is limited, and certain options are not available.

First, you need to register into the platform at:

https://databricks.com/try-databricks



Please tell us about yourself

First Name: *

Your Name

Last Name: *

Your last name

Company *

Imperial College

Company Email *

xxxxxxxxxxxx@imperial.ac.uk

Title *

student

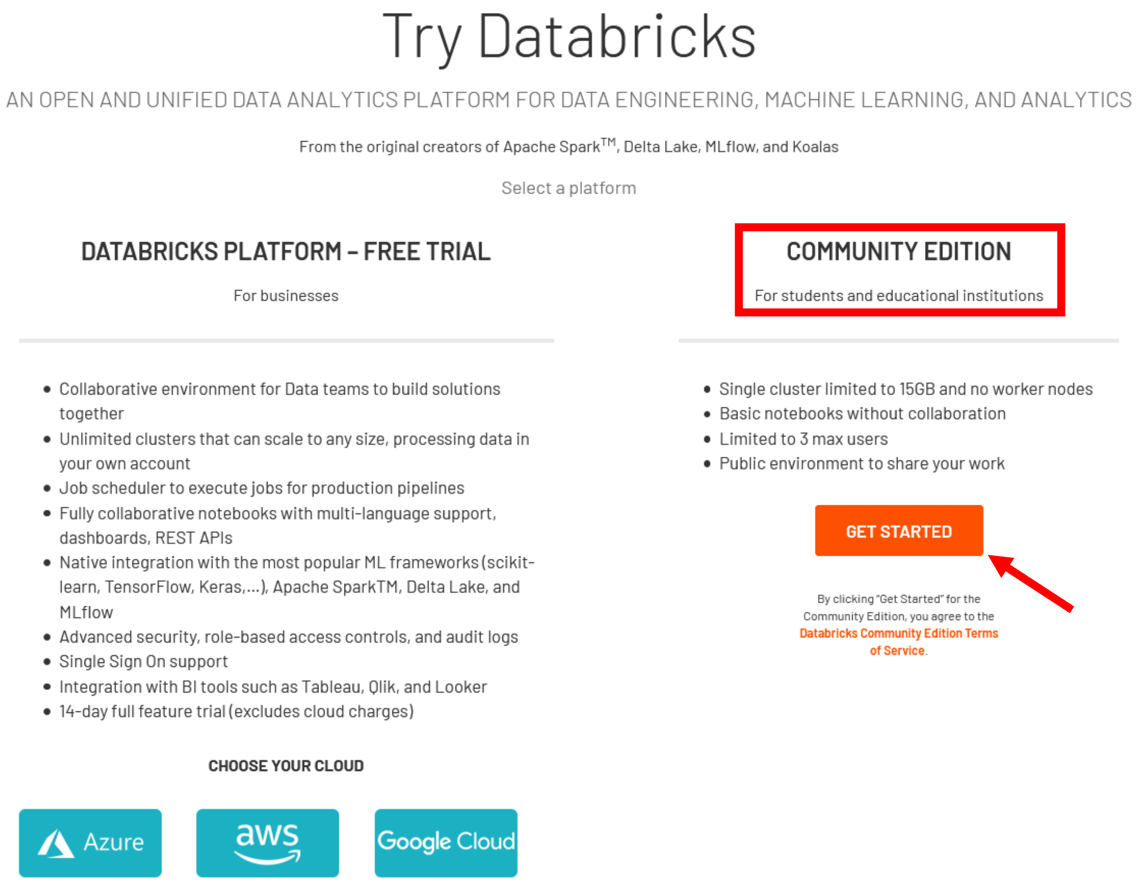
Phone Number

☑ 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.

You can register using your imperial email address.

The next page looks like this:



Try Databricks

AN OPEN AND UNIFIED DATA ANALYTICS PLATFORM FOR DATA ENGINEERING, MACHINE LEARNING, AND ANALYTICS

From the original creators of Apache Spark™, Delta Lake, MLflow, and Koalas

Select a platform

DATABRICKS PLATFORM – FREE TRIAL

For businesses

- Collaborative environment for Data teams to build solutions together
- Unlimited clusters that can scale to any size, processing data in your own account
- Job scheduler to execute jobs for production pipelines
- Fully collaborative notebooks with multi-language support, dashboards, REST APIs
- Native integration with the most popular ML frameworks (scikit-learn, TensorFlow, Keras,...), Apache SparkTM, Delta Lake, and MLflow
- Advanced security, role-based access controls, and audit logs
- Single Sign On support
- Integration with BI tools such as Tableau, Qlik, and Looker
- 14-day full feature trial (excludes cloud charges)

CHOOSE YOUR CLOUD

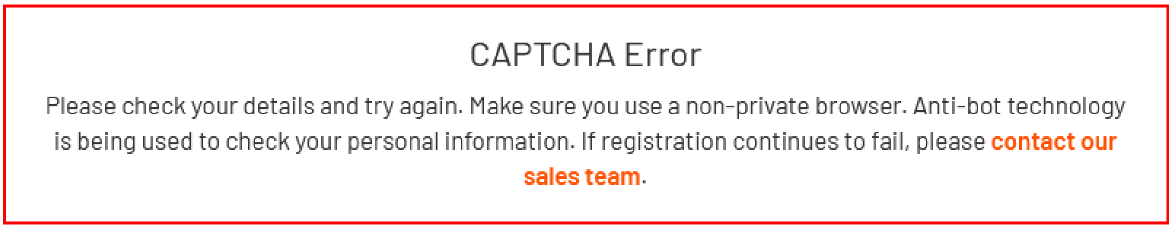
COMMUNITY EDITION

For students and educational institutions

- Single cluster limited to 15GB and no worker nodes
- Basic notebooks without collaboration
- Limited to 3 max users
- Public environment to share your work

By clicking "Get Started" for the Community Edition, you agree to the Databricks Community Edition Terms of Service.

Select the community edition. If you hit a time out or an error message such as:

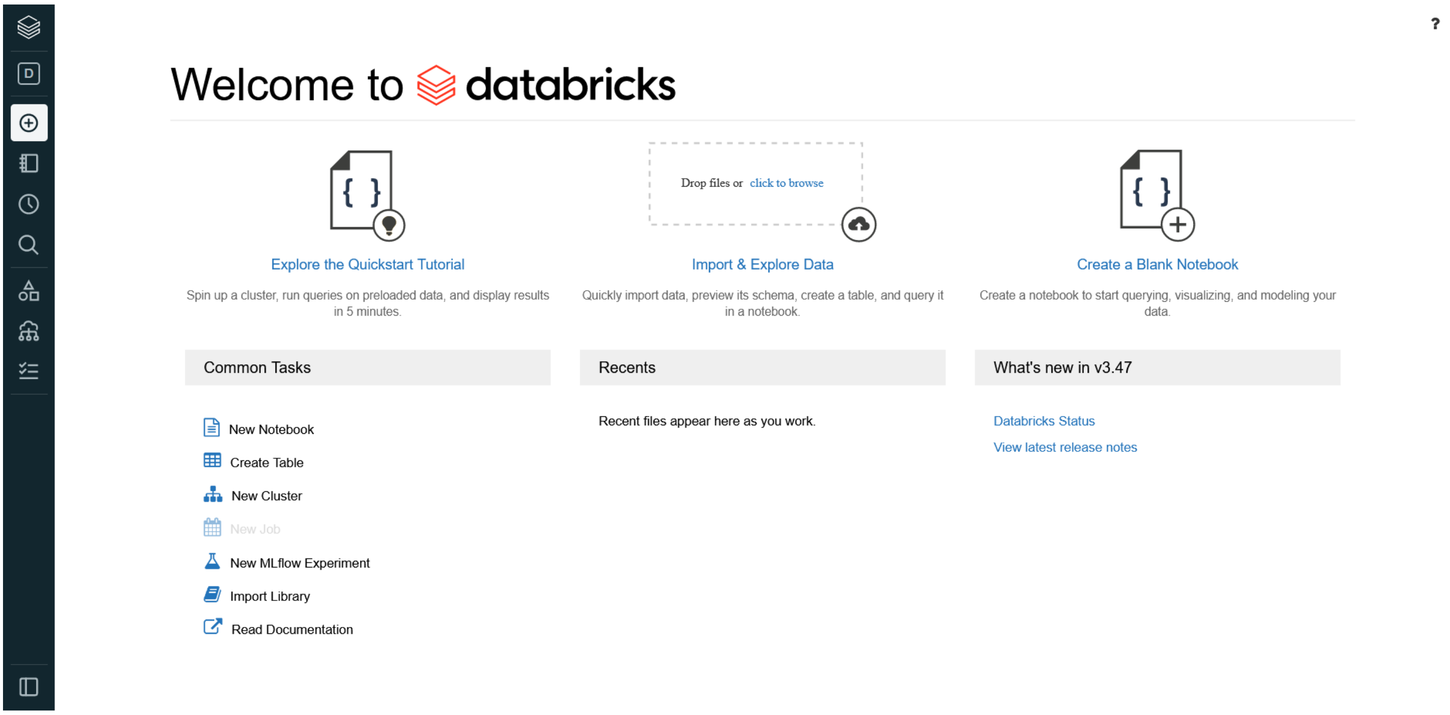


CAPTCHA Error

Please check your details and try again. Make sure you use a non-private browser. Anti-bot technology is being used to check your personal information. If registration continues to fail, please contact our sales team.

Please try again without any ad-block and not in private mode.

Once you have successfully registered, you arrive on the landing page:



Welcome to databricks

Explore the Quickstart Tutorial

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

Drop files or click to browse

Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.

Common Tasks

- New Notebook
- Create Table
- New Cluster
- New Job
- New MLflow Experiment
- Import Library
- Read Documentation

Recents

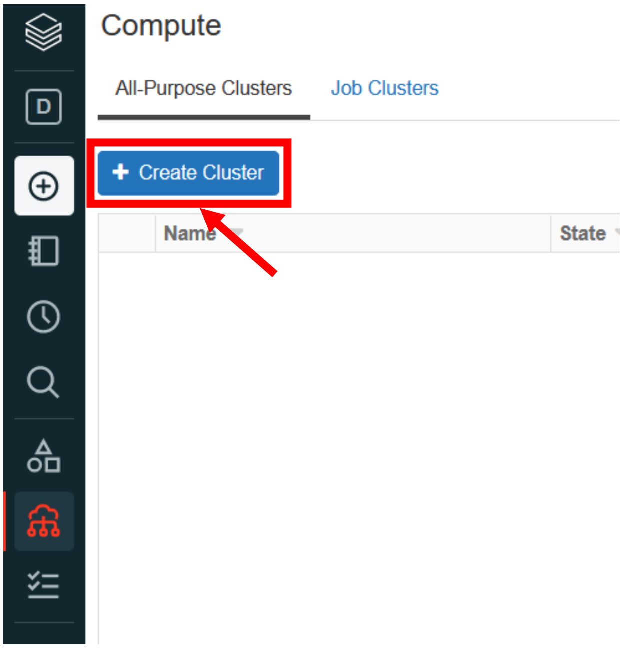
Recent files appear here as you work.

What's new in v3.47

Databricks Status

View latest release notes

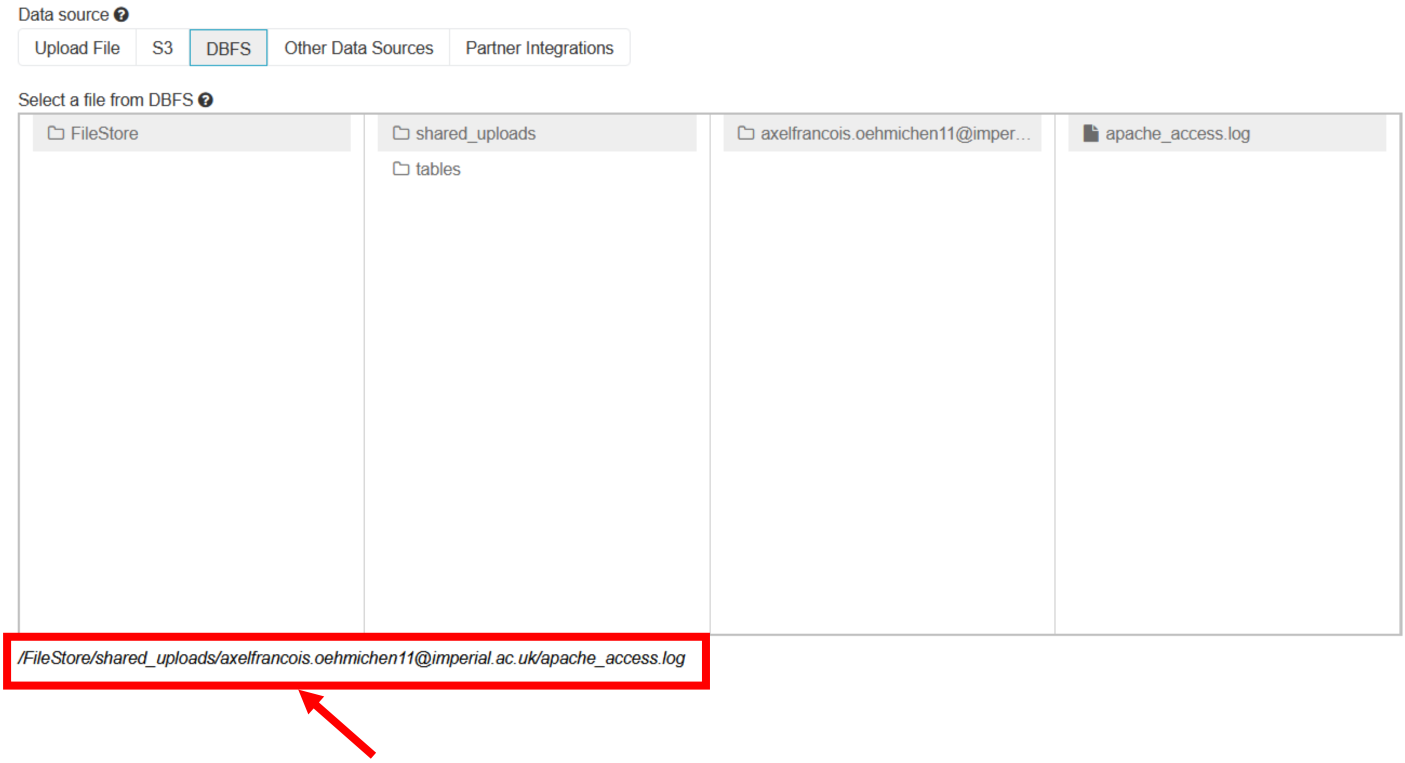
You will need to create a compute first:



Give the cluster any name you want and select "Create Cluster". Once the cluster is ready, you can go back to the home page, and we will upload both the data and the Jupyter Notebook.

Upload the data:

You can upload the log file by clicking “click to browse” or “Import & Explore Data”. At the end of the upload, a path where the file is located will be given to you. If you lose that information, you can find it by browsing your data source.

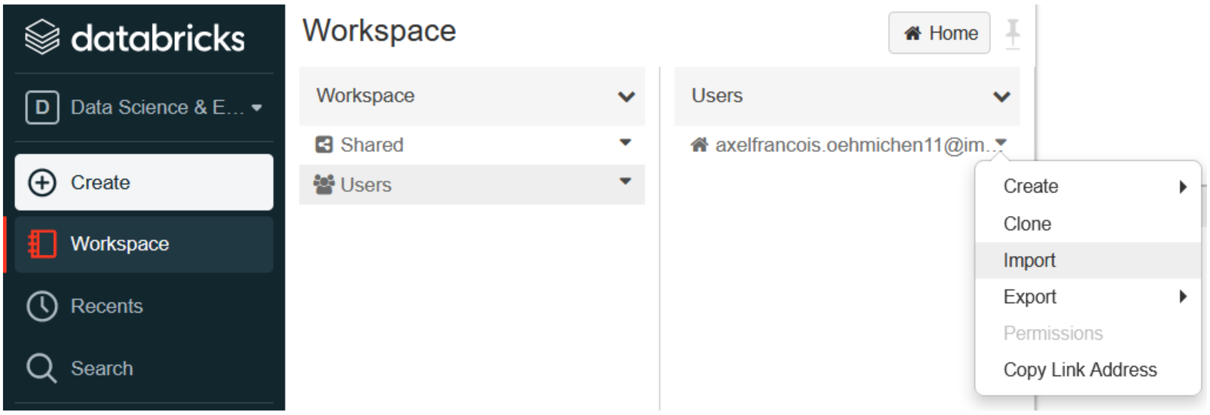


/FileStore/shared_uploads/axelfrancois.oehmichen11@imperial.ac.uk/apache_access.log

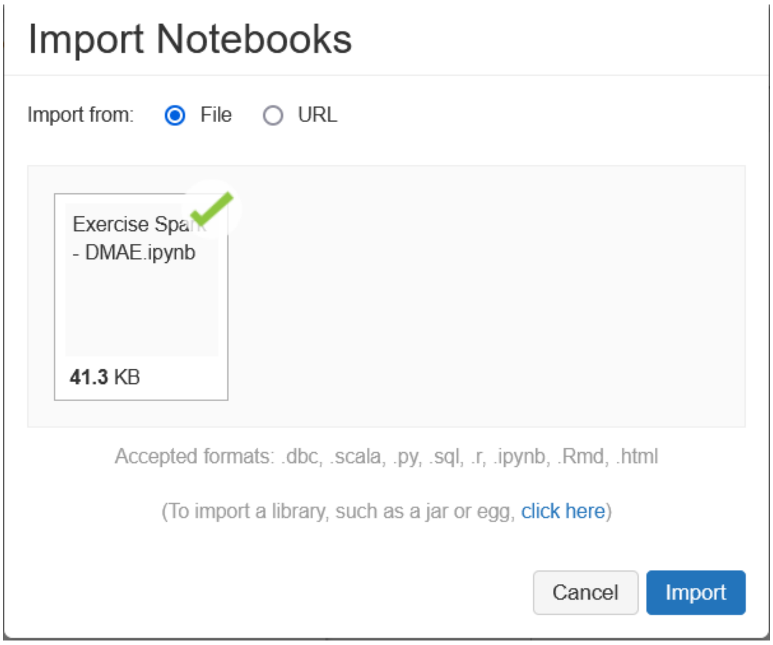
That path will specify the location of the log file in the notebook.

Upload the Notebook:

To upload the notebook, you need to go under your workspace and select the import option:



And



Import Notebooks

Import from: ● File ○ URL

Exercise Spark - DMAE.ipynb

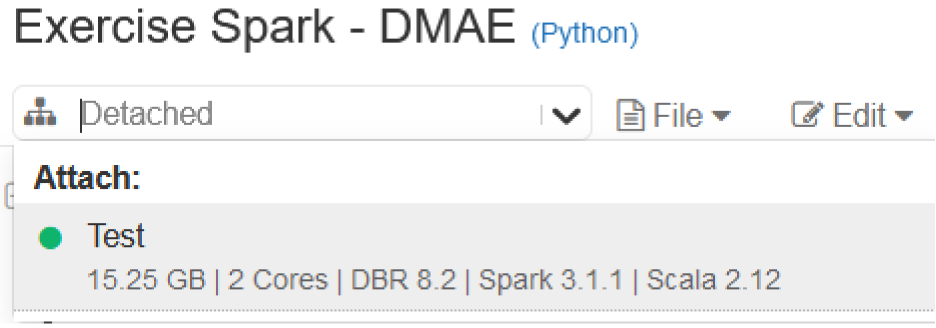
41.3 KB

Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html

(To import a library, such as a jar or egg, click here)

Once the file is imported, you will be transfer to the notebook.

In order to run the notebook, you need to select the cluster we created:



Exercise Spark - DMAE (Python)

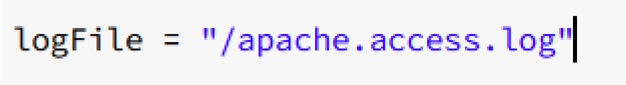
Detached

Attach:

● Test

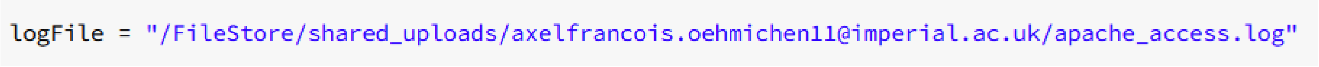
15.25 GB | 2 Cores | DBR 8.2 | Spark 3.1.1 | Scala 2.12

The final step is to replace the location of the log file by the path in DBS:



```
logFile = "/apache.access.log"
```

To



```
logFile = "/FileStore/shared_uploads/axelfrancois.oehmichen11@imperial.ac.uk/apache_access.log"
```