NOVA IMS Information Management School

Analysing Big Data

Introduction to Databricks, Python & Spark

Maria Almeida
malmeida@novaims.unl.pt

Niclas Frederic Sturm
nsturm@novaims.unl.pt

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

About me – Maria

- ***Academic Path***
  - M.Sc. in Data Science and Engineering (Instituto Superior Técnico)
  - B.Sc. in Information Management (NOVA IMS)

- **Work**
  - Machine Learning Engineer @ Neuraspace (Space Traffic Management)

About me – Niclas

- ***Academic Path***
  - Ph.D. Student in *Information Management* (Nova IMS)
  - M.Sc. in Business Analytics (Nova SBE)
  - B.Sc. in Economics & Ancient History (University of Heidelberg)

- ***Research Interests***
  - Public Procurement
  - Network Science
  - Quantitative Methods in the Social Sciences and Humanities

Spark

APACHE Spark™

General Purpose Cluster Computing Framework for Big Data

Python

- Working in a notebook with Python (and soon PySpark)
- Using variables, operators, built in functions
- Controlling code flow with conditional statements and loops
- Using data types including lists, dictionaries, and tuples
- Defining and using both named functions and anonymous functions (lambda functions)

Databricks Community Edition

- Community: free

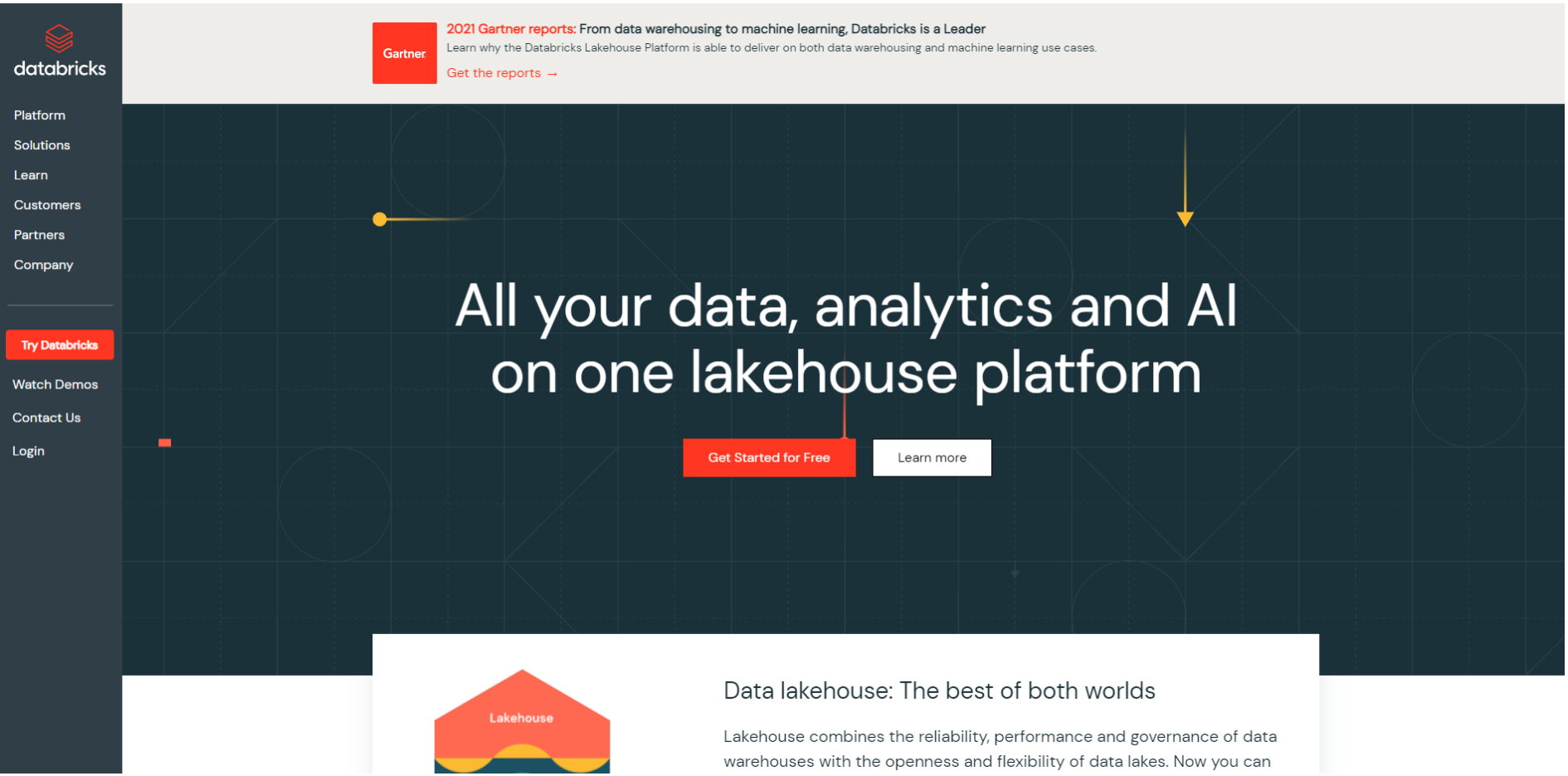
Communication

moodle

Goals for the class

1. **Creation of a Databricks account**
2. **Introduction to Databricks**
   - Overview of Databricks and its features
   - Comparison of Databricks and Jupyter Notebook
   - Benefits of using Databricks
3. **Uploading a File to Databricks**
4. **Introduction to Distributed Processing**
   - Databricks Cluster
5. **Introduction to Shell Commands**
   - Overview of shell commands and their importance in Big Data processing
   - How to use shell commands in Databricks

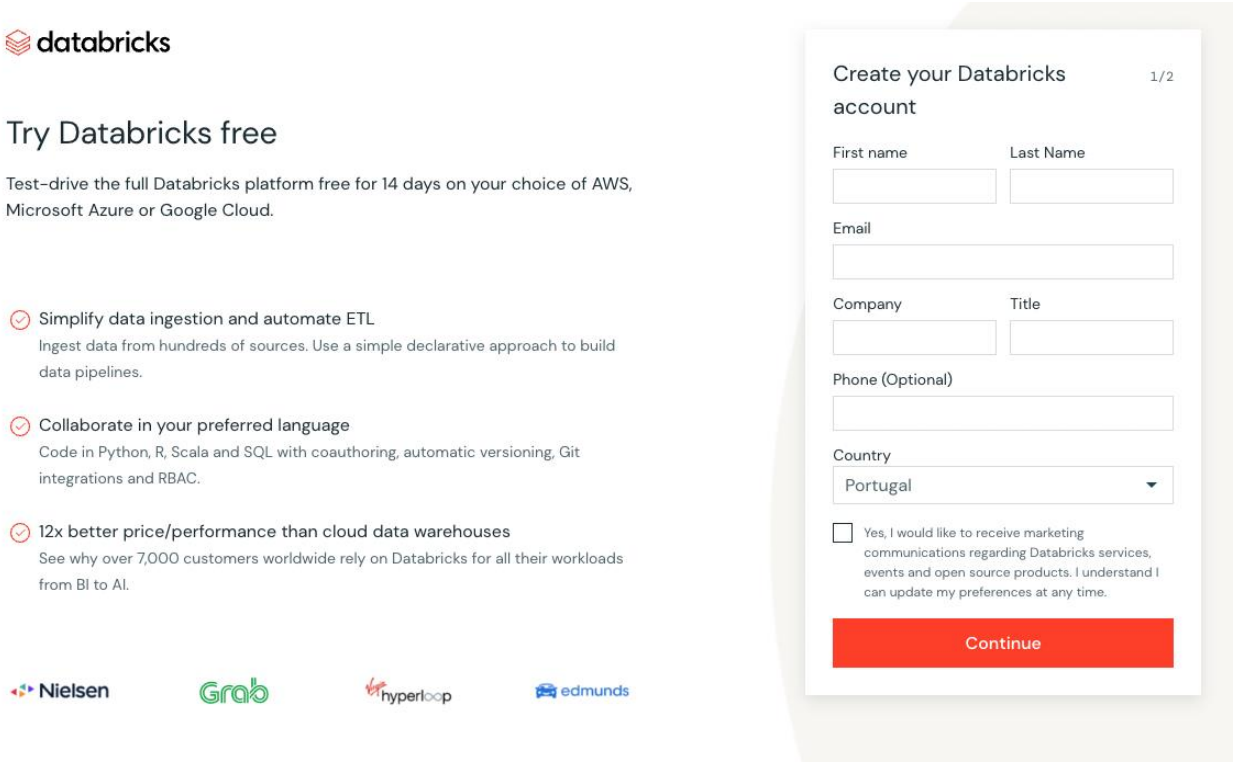
Databricks Community Edition



https://towardsdatascience.com/what-does-databricks-do-8a6c4ef9071b

Databricks Community Edition

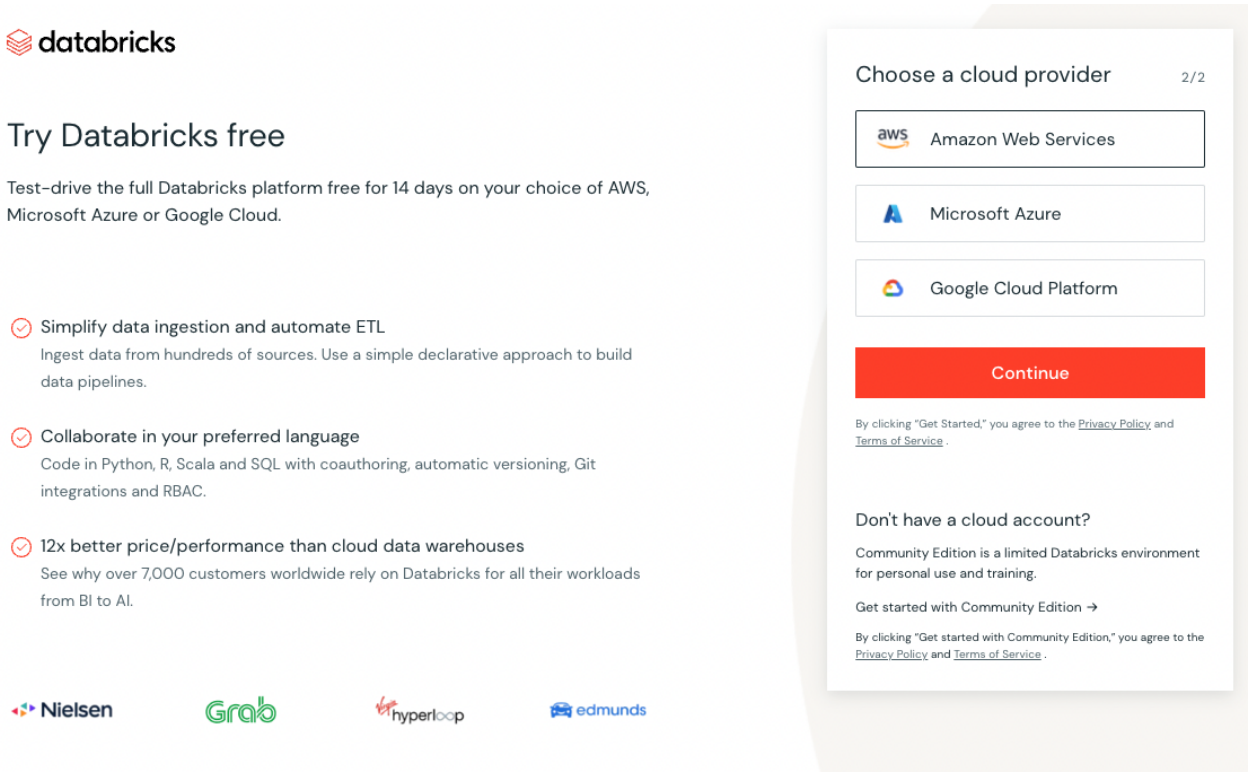
https://databricks.com/try-databricks



1. Fill in the registration form **(hint: use your personal email)**
2. Click "**Get Started for free**"
3. Solve the verification enigma
4. Choose "**Get started with Community Edition**" (in the bottom of the page)
5. When you receive the Welcome to Databricks email, click on the link to verify your email address
6. Reset your password. You are done!

Databricks Community Edition

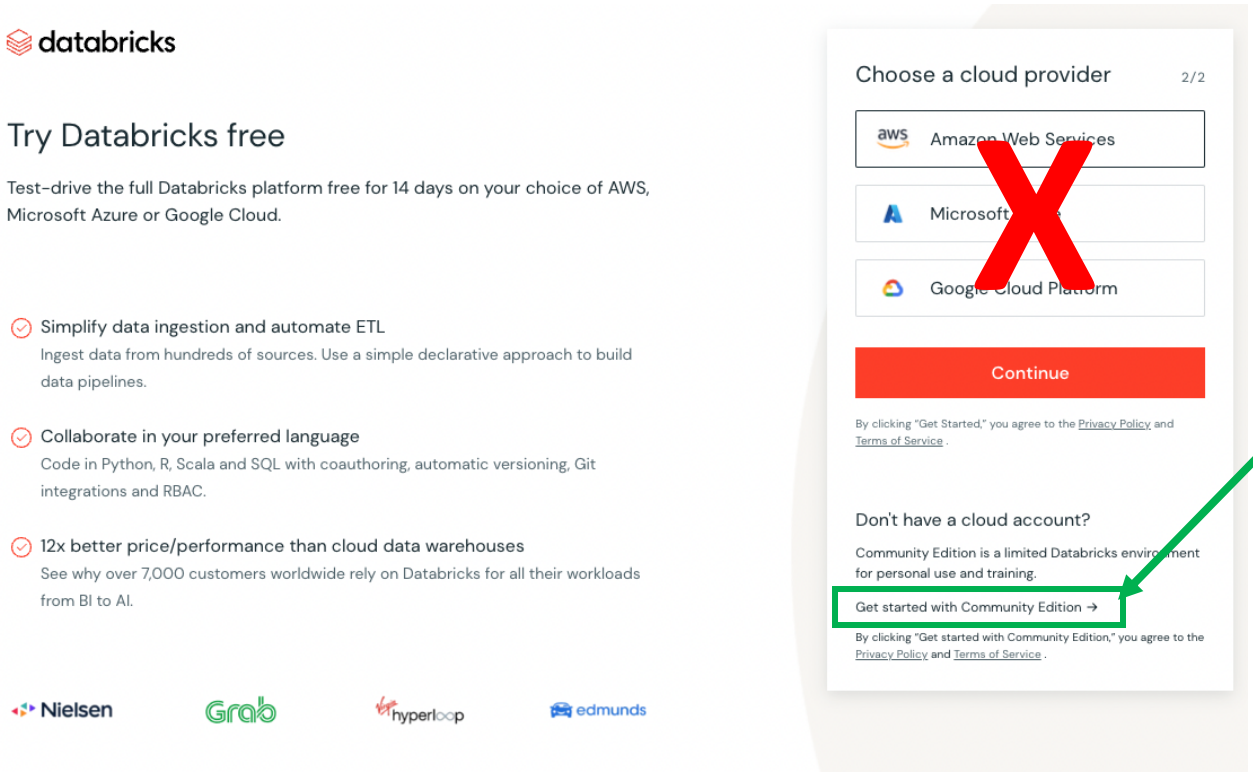
https://databricks.com/try-databricks



1. Fill in the registration form **(hint: use your personal email)**
2. Click "**Get Started for free**"
3. Solve the verification enigma
4. Choose "**Get started with Community Edition**" (in the bottom of the page)
5. When you receive the Welcome to Databricks email, click on the link to verify your email address
6. Reset your password. You are done!

Databricks Community Edition

https://databricks.com/try-databricks



1. Fill in the registration form **(hint: use your personal email)**
2. Click **"Get Started for free"**
3. Solve the verification enigma
4. Choose **"Get started with Community Edition"** (in the bottom of the page)
5. When you receive the Welcome to Databricks email, click on the link to verify your email address
6. Reset your password. You are done!

Goals for the class

1. ~~Creation of a Databricks account~~
2. **Introduction to Databricks**
   - Overview of Databricks and its features
   - Comparison of Databricks and Jupyter Notebook
   - Benefits of using Databricks
3. **Uploading a File to Databricks**
4. **Introduction to Distributed Processing**
   - Databricks Cluster
5. **Introduction to Shell Commands**
   - Overview of shell commands and their importance in Big Data processing
   - How to use shell commands in Databricks

Databricks Overview

- **Databricks**: web-based platform to work with Spark.

Databricks	Jupyter
Cloud-based platform	Local application
For large datasets	For small datasets
Distributed processing	Single-node processing
Offers cluster management, data pipelines, ML	No features

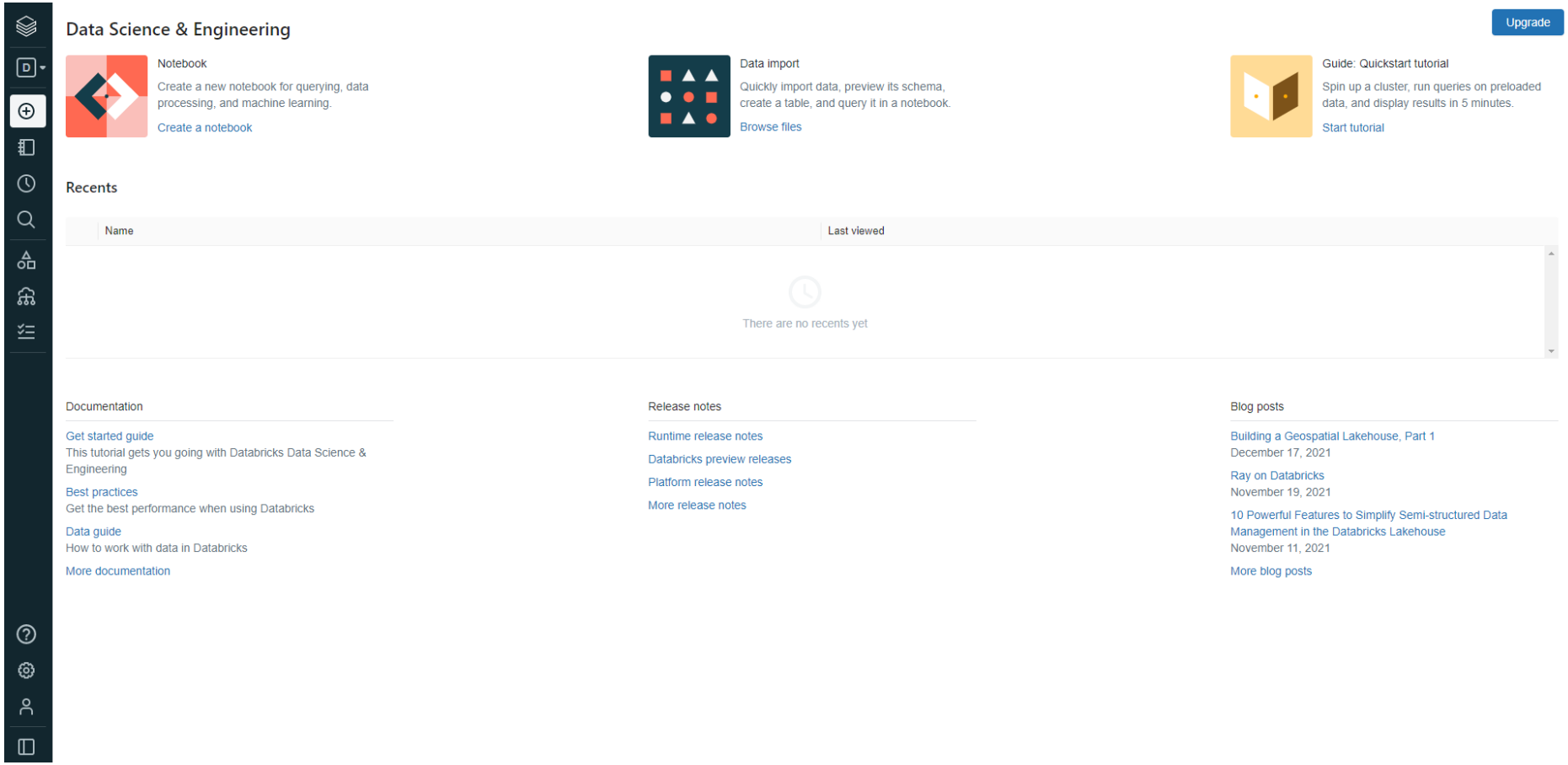
Main Benefits of using Databricks:

- Scalability
- Collaboration;
- Integration with other cloud services

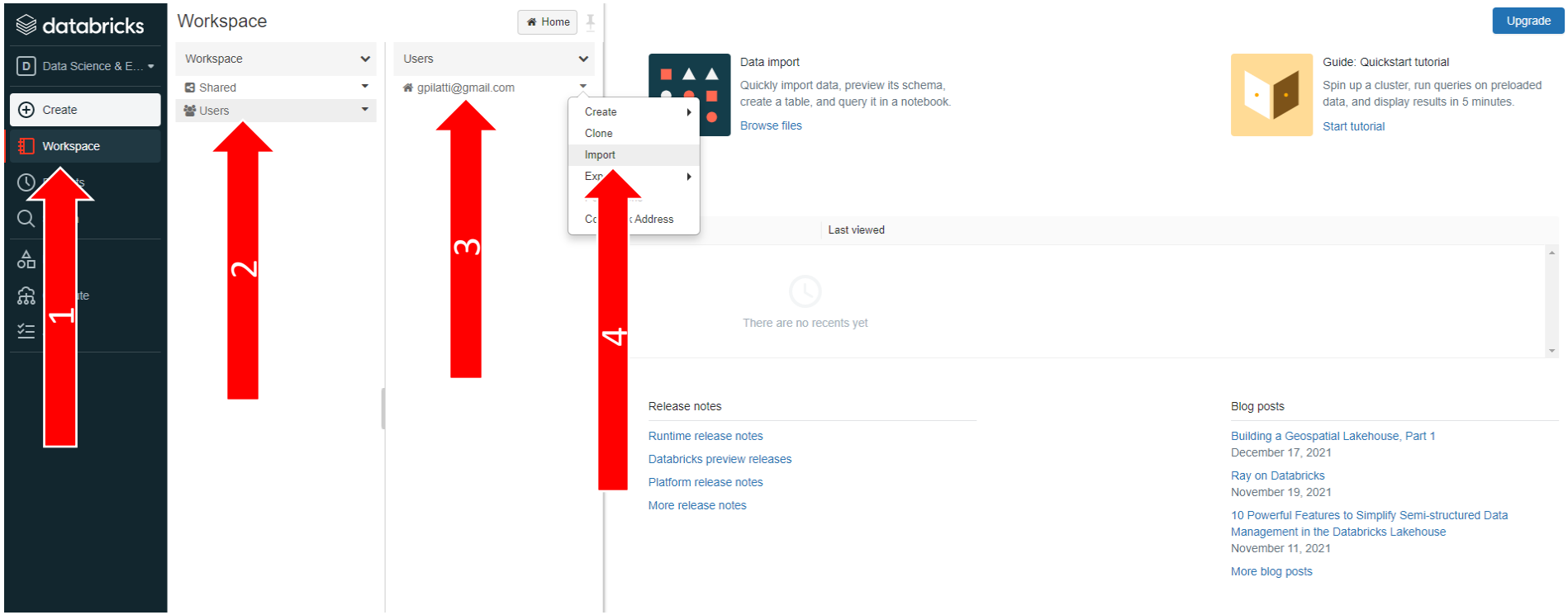
Goals for the class

1. ~~**Creation of a Databricks account**~~
2. ~~**Introduction to Databricks**~~
   - ~~Overview of Databricks and its features~~
   - ~~Comparison of Databricks and Jupyter Notebook~~
   - ~~Benefits of using Databricks~~
3. **Uploading a File to Databricks**
4. **Introduction to Distributed Processing**
   - Databricks Cluster
5. **Introduction to Shell Commands**
   - Overview of shell commands and their importance in Big Data processing
   - How to use shell commands in Databricks

Databricks Community Edition

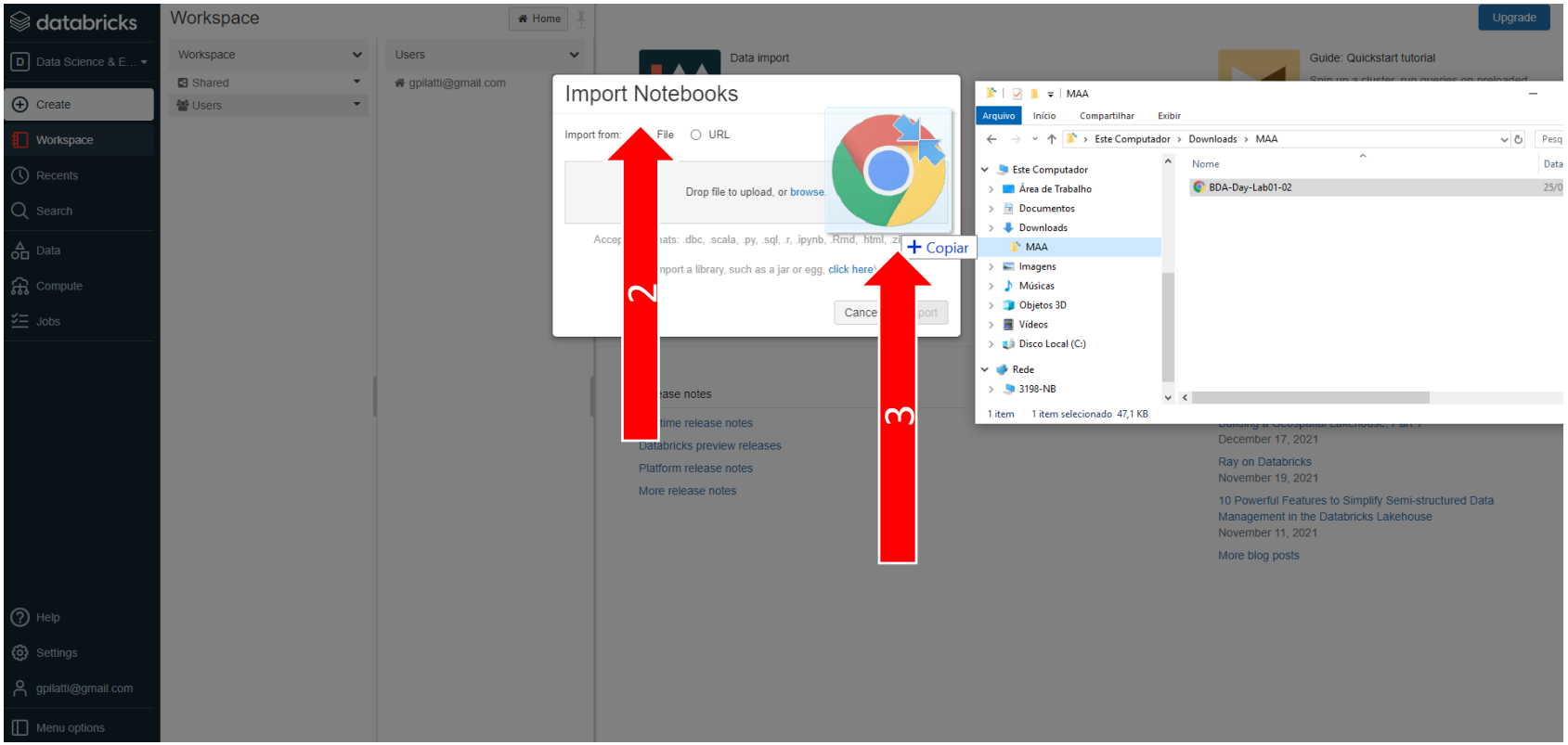


Databricks Community Edition



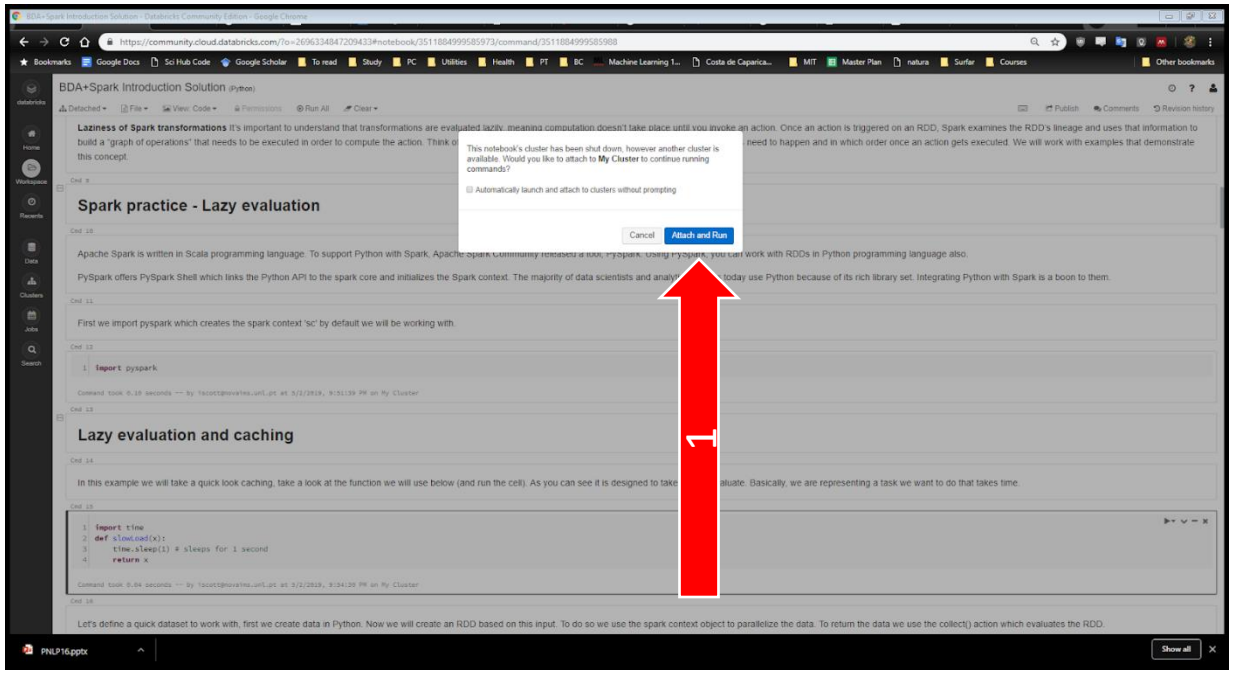
1. In Databricks press ‘Workspace’ on the left.
2. Click on 'Users'
3. Click on the small inverted triangle to the right of your email address
4. Select import

Databricks Community Edition



1. Click on the link to this week exercises on *Moodle* and download the notebook.
2. Choose Import from: File
3. Drag and drop the file you downloaded from *Moodle*
4. Click on Import

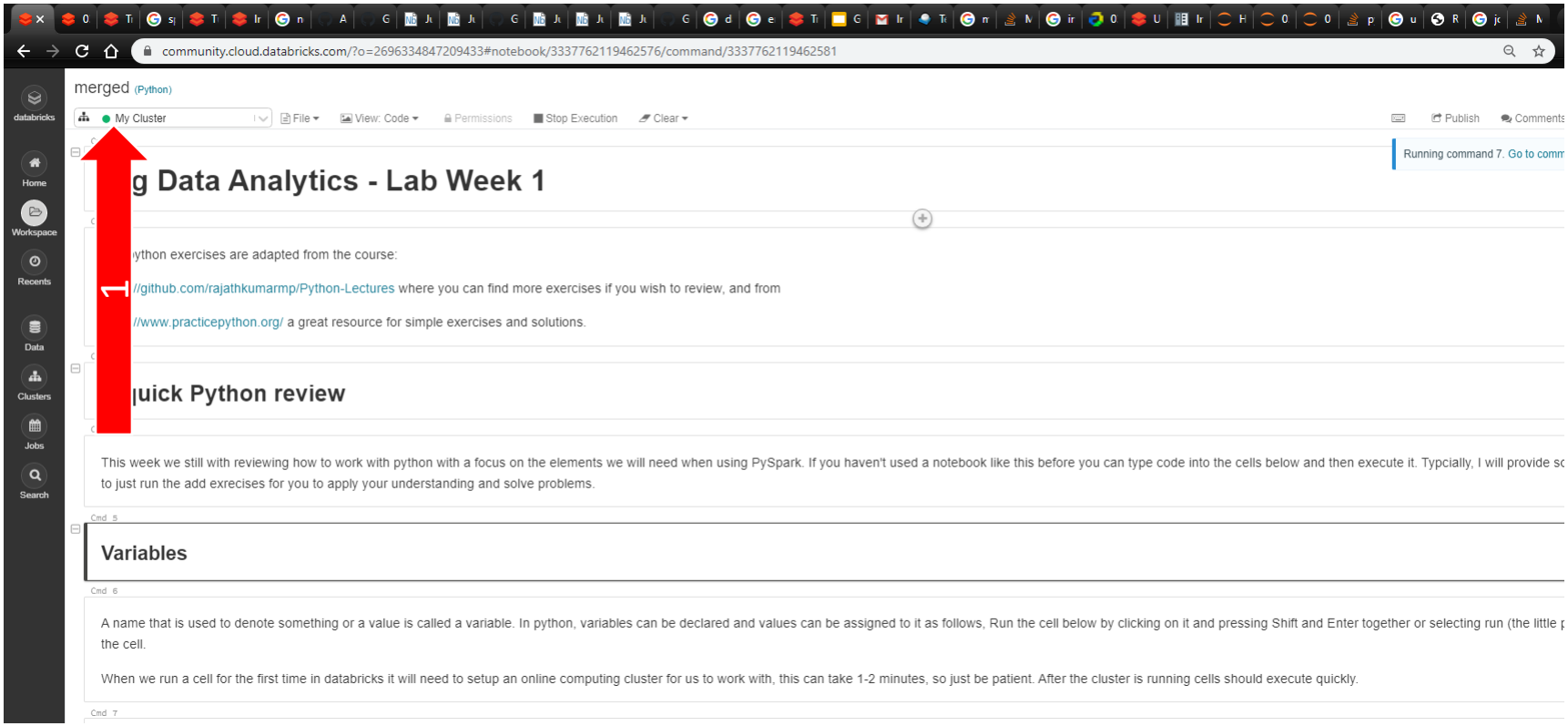
Databricks Community Edition



You should have a **notebook** filled with **exercises** to work with (the notebook will look slightly different to a Jupyter notebook in case that is what you are used to, you can still use Shift-Enter to run a cell).

You can now try to run a cell, you will see that you will be asked to **attach a cluster**, simply click "select resource" and "create resource".

Databricks Community Edition

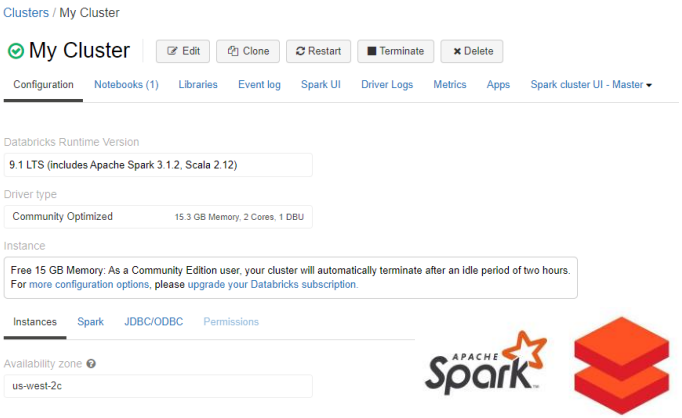


If you leave your notebook for a while it may **detach** from the **cluster**, giving you an error when you try to run cells. In this case you just need to reselect your cluster at the top left and the notebook will reattach to it.

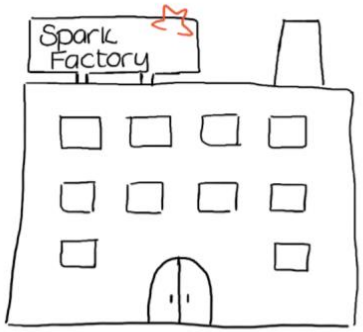
Goals for the class

1. ~~**Creation of a Databricks account**~~
2. ~~**Introduction to Databricks**~~
   - ~~Overview of Databricks and its features~~
   - ~~Comparison of Databricks and Jupyter Notebook~~
   - ~~Benefits of using Databricks~~
3. ~~**Uploading a File to Databricks**~~
4. **Introduction to Distributed Processing**
   - Databricks Cluster
5. **Introduction to Shell Commands**
   - Overview of shell commands and their importance in Big Data processing
   - How to use shell commands in Databricks

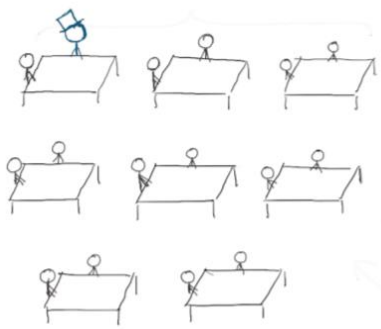
Databricks Community Edition - Cluster



Outside view



Under the hood



Cohesive unit

- **Set of computation resources** that **perform the data workloads** ran in Databricks (commands in notebooks, commands run from BI tools, ETLs, and so on).
- Consists of **multiple nodes** (individual machines) that operate on the workloads in parallel.
- There is **one driver node for every cluster**, which is the one that delegates tasks and oversees the execution of the specific workload.
- Spark uses **RAM** to **store data in memory** for fast processing, but it also uses **distributed memory** to **scale out** across multiple nodes in a cluster.
- Spark's in-memory computing model allows for **faster processing** of **large datasets**, but it also requires careful management of memory resources to avoid running out of memory.

Goals for the class

1. ~~**Creation of a Databricks account**~~
2. ~~**Introduction to Databricks**~~
   - ~~Overview of Databricks and its features~~
   - ~~Comparison of Databricks and Jupyter Notebook~~
   - ~~Benefits of using Databricks~~
3. ~~**Uploading a File to Databricks**~~
4. ~~**Introduction to Distributed Processing**~~
   - ~~Databricks Cluster~~
5. **Introduction to Shell Commands**
   - Overview of shell commands and their importance in Big Data processing
   - How to use shell commands in Databricks

Files, Shell, Bash, Scripting

One of the **original ways to use computers** to process data was developed for the Unix Environment and followed its philosophy. Shortly summarized as:

- Write programs to do one thing and do it well.
- Write programs to work together.
- Write programs to handle text streams because that is a universal interface.

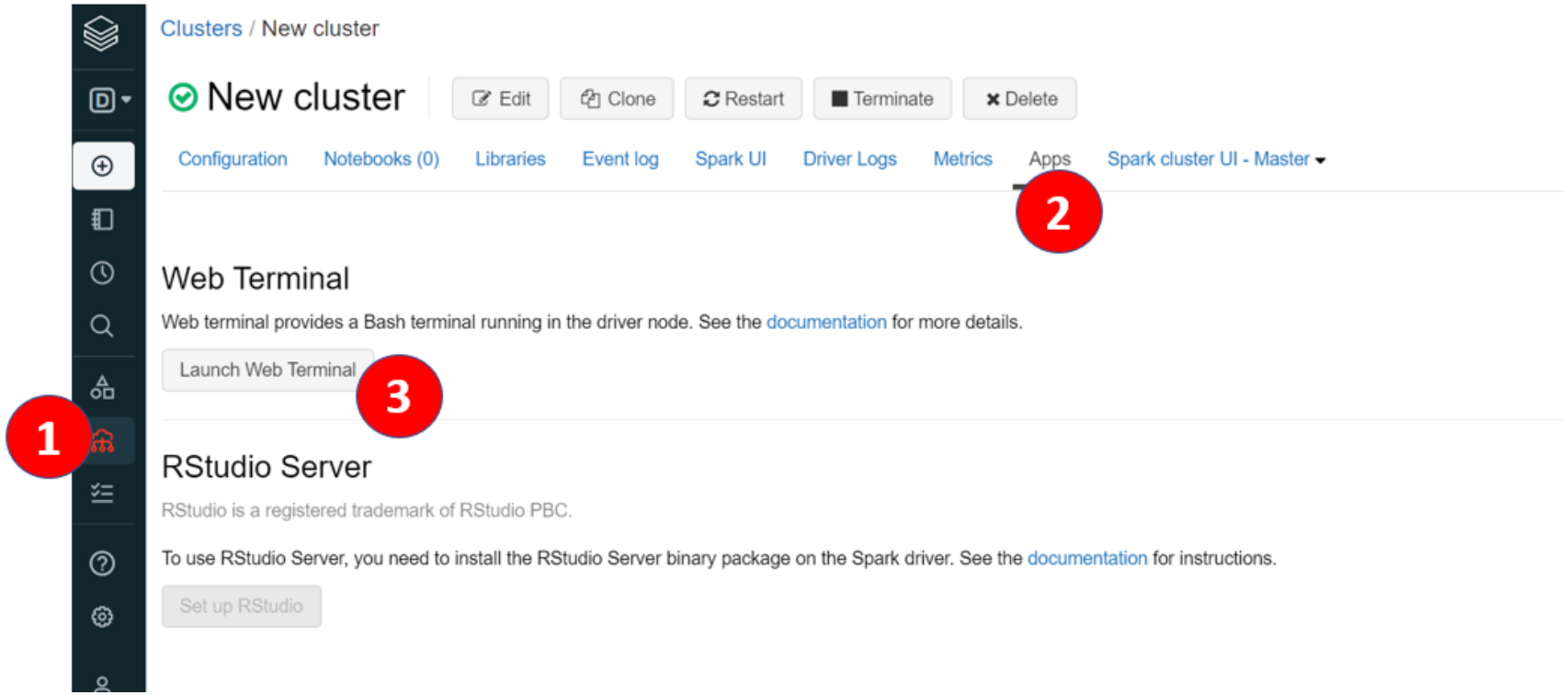
Most of the commands developed within the Unix ecosystem, more than 30 years ago, are still relevant today

Why start this way?

- Ease of execution of commands (no need to copy and paste every time)
- Powerful programming constructs

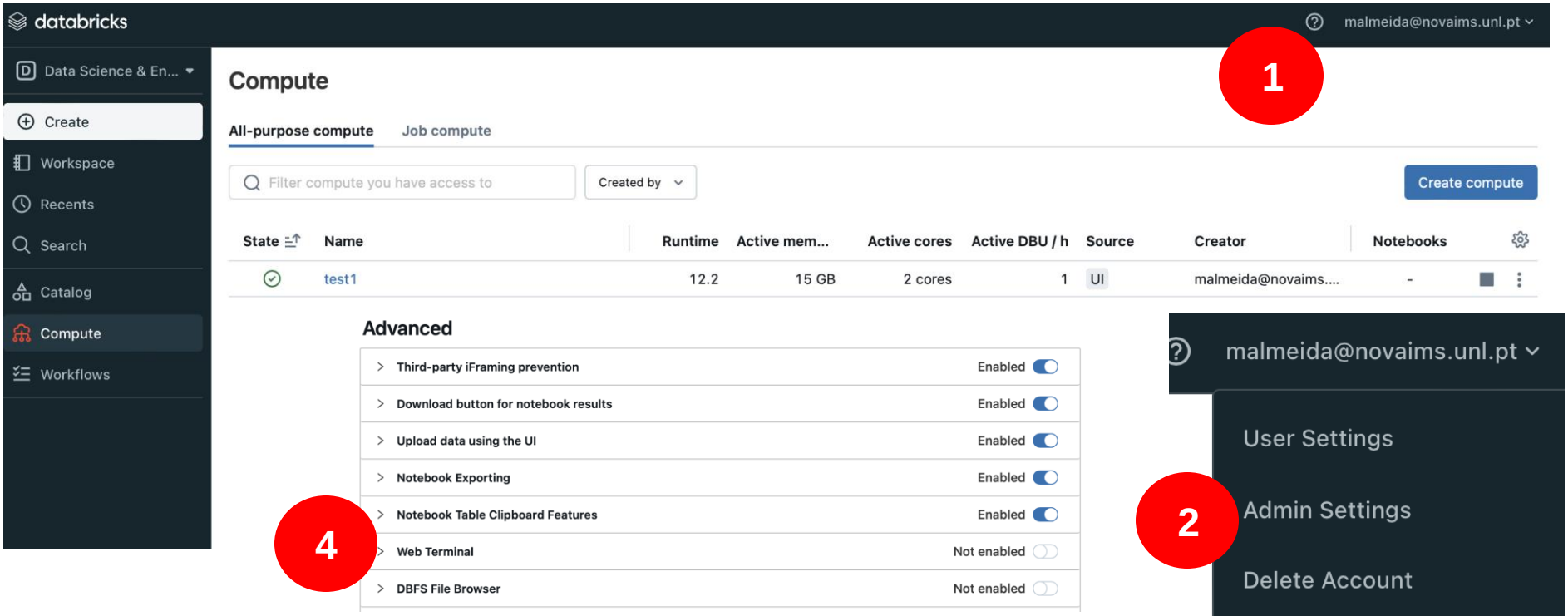
MORE INFO:
BDA 2023 Lab1 - Bash commands intro.pdf

Files, Shell, Bash, Scripting



If the Web Terminal is not enabled, follow these steps in the next page.

Files, Shell, Bash, Scripting



1. Go to the Settings
2. Click on Admin Settings.
3. Click the Workspace Settings tab.
4. In the Advanced section, click the Web Terminal toggle.
5. Refresh the page.

Shell commands

- Shell is an environment in which we can run our commands, programs, and shell scripts.
- There are different flavors of a shell, just as there are different flavors of operating systems.
- Each flavor of shell has its own set of recognized commands and functions.
- In this course we will use **Bash** ("Bourne Again Shell") shell as the main shell interpreter.

Tip: If the shell is running on your local filesystem:

- You can press the up arrow to cycle through previous commands
- When using windows, you can right-click to paste (instead of ctrl-v).

Shell commands

Practice and more information:

BDA 2024 Lab 1 – Bash commands intro

Python review (will not be covered in this class)

BDA-Lab1-p1 - Python review.html
BDA-Lab1-p2- Python review.html

End