HFactory user guide – Hi!ckathon #5

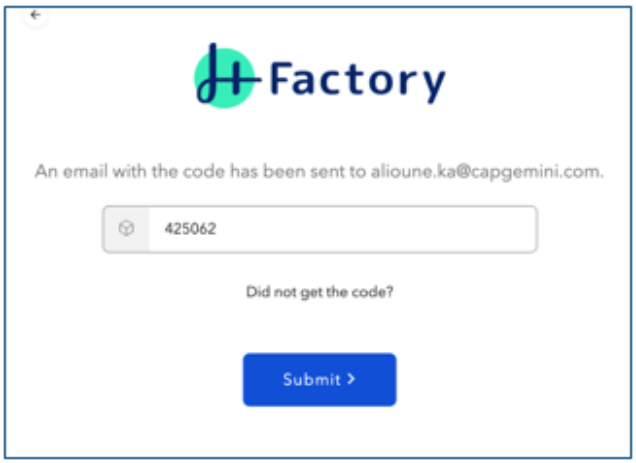
The Hi! PARIS center has decided to use HFactory, an all-in-one learning platform dedicated to Data Science and AI, for the Hi!ckathon #5.

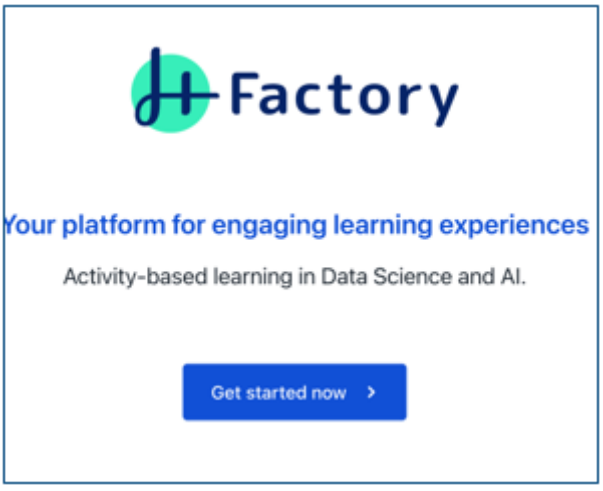
The learning platform provides:

- An **activity** section which includes a description of the datasets, use case, deliverables and overall timeline of the event, as well as information on your team/group.
- A **Data Science computing environment** where you can launch CPU/GPU instances compute instances and work with popular code editors (JupyterLab, Jupyter Notebook and VSCode).

I. Log-on to HFactory

Participants of the event received an email to log on to the "**Hi!ckathon #5 – Final Sprint**" activity on HFactory. If it isn't the case, reach out to a member of the Hi!ckathon #5 organizational team or send us a message on Discord.



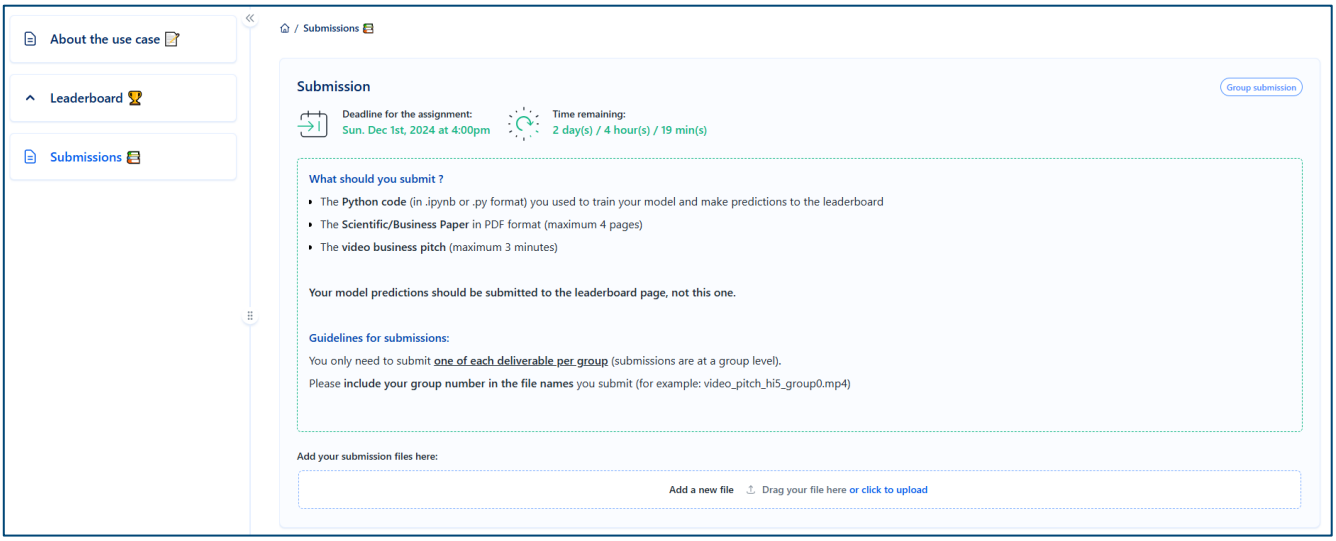


Make sure you log-on with the email you used during registration (or the email address you received the invitation to register).

II. Submissions

All of the submissions should be made on the **"Hickathon #5 – Final Sprint"** activity.

- Submit your **test predictions** to the **Leaderboard** section
- Submit all remaining files (**python code, scientific/business paper, video pitch**) to the **Submissions** section



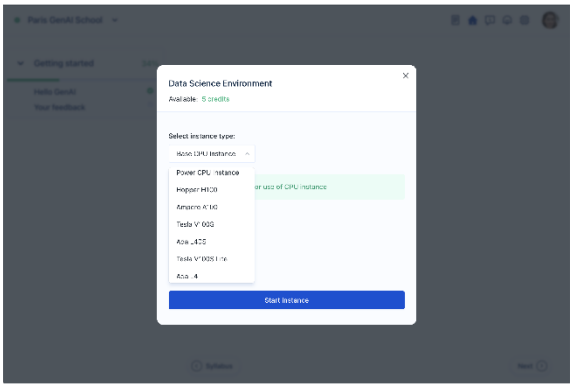
III. Access to compute instances

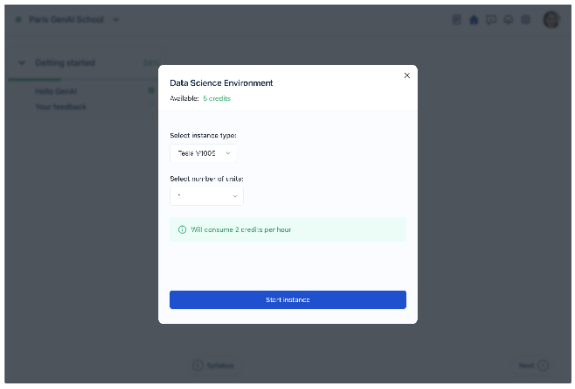
You have access to Cloud compute instances directly within HFactory. You can think of it as AWS Sagemaker or Google Colab fully integrated in the Learning platform.

1. Configuring an instance

Instances can be started from the **compute icon** in the top menu bar. These instances are operated on the OVHcloud AI training infrastructure.

Clicking on the compute icon will open a modal allowing you to **select and start an instance**.





You can launch CPU or GPU instances based on the number of credits you have.

- If you have **no compute credits** available, the modal will only offer you the option to start the **default 4-core CPU instance**.
- In case you have been **granted compute credits**, you will have a **choice among the various CPU and GPU instances available**. Selecting a specific instance type, will provide you with information on the number of compute credits used per hour.

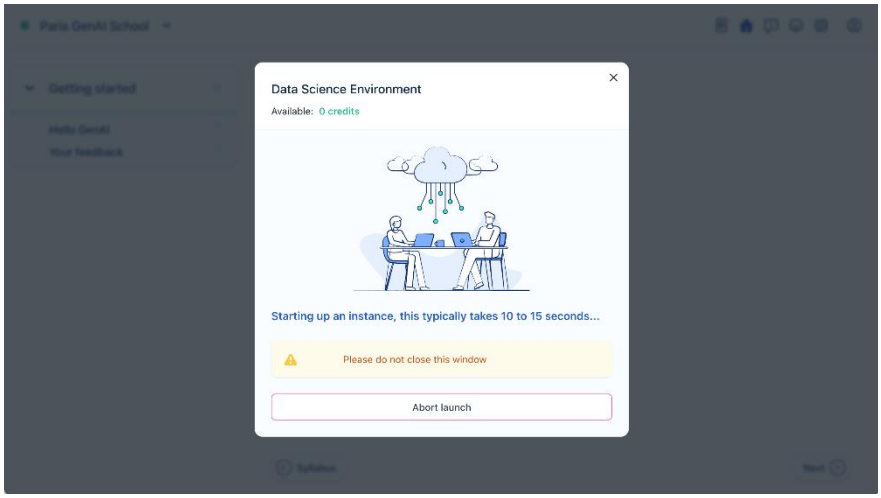
In the context of this hackathon, **we recommend that you start Tesla V100S Lite instances,** and as a **back-up choice if none is available Tesla V100S ones**. **Also, if you are programming and debugging, use a simple CPU instance to save credits.**

*GPU credits will be given to participants of the Hi!ckathon **Saturday evening**. Until then, you will only have access to CPU instances.*

2. Launching an instance

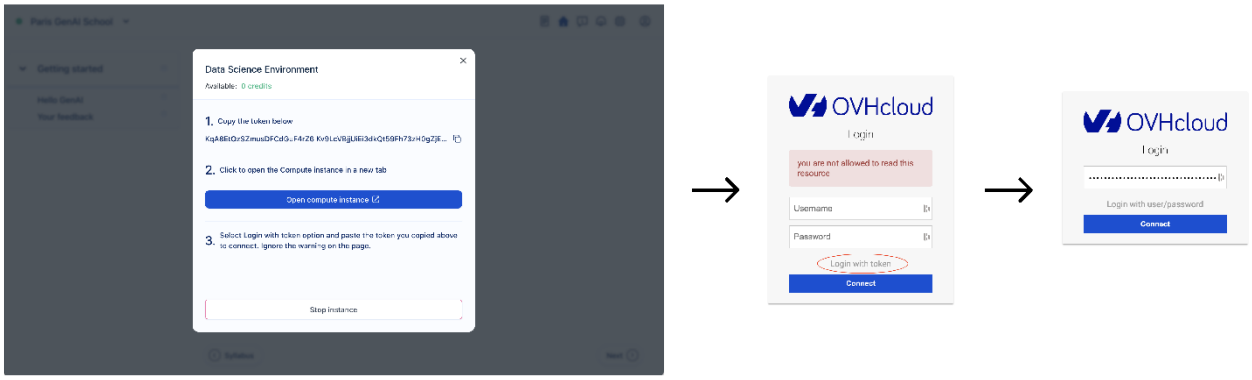
Launching an instance means starting a new machine in the Cloud, so this usually takes **around 15 seconds** for fresh new environments.

- This start up time can increase substantially (up to several minutes) if you are working with a high number of files and/or large files. The content of the attached volumes is synchronised each time you start a new instance, potentially slowing down the process.
- It is recommended to periodically clean your home directory of unnecessary files such as model checkpoints or dumps.



When your instance is available, you will need to **copy a token** and **paste it to the OVHcloud login window** after duly **selecting "login with token"** (this is not the default option).

This token is used to uniquely authenticate you and will not be asked on subsequent instances launch.



3. Stopping an instance

You can **reopen the instance modal at any time** by clicking on the compute icon from the top menu bar. If your instance is already running, it will offer you the option to either go back to the instance or to shut it down.

If you forget to switch off your instance after you have completed your work, it will **automatically shut down after a while**. The system indeed performs **background activity checks** and shuts down idle instances to limit unnecessary usage of compute infrastructure.

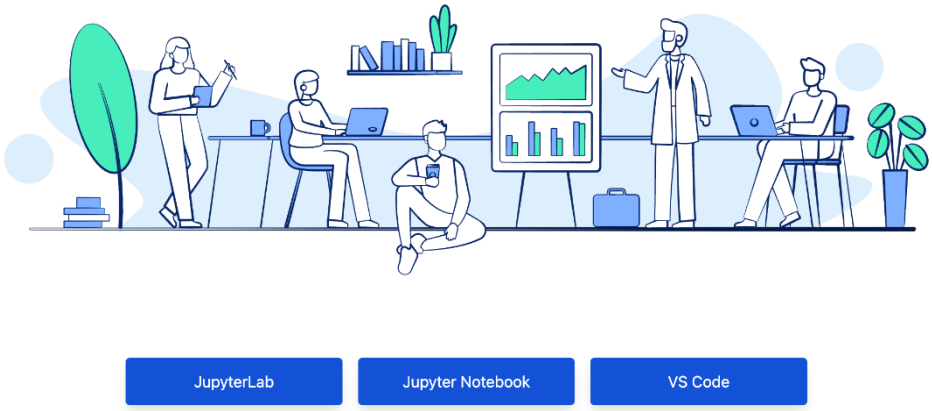
This process makes its **assessment based on browser activity**, **kernel activity** and for **GPU instances Cuda activity**.

- We cannot fully guarantee that the automatic closing of an instance will not interrupt a running script or process in exceptional cases.
- In this situation and all other cases, data on the instance is however persisted as part of the shutdown process, ensuring the absence of data loss.

IV. Using compute instances

1. Choice of IDEs

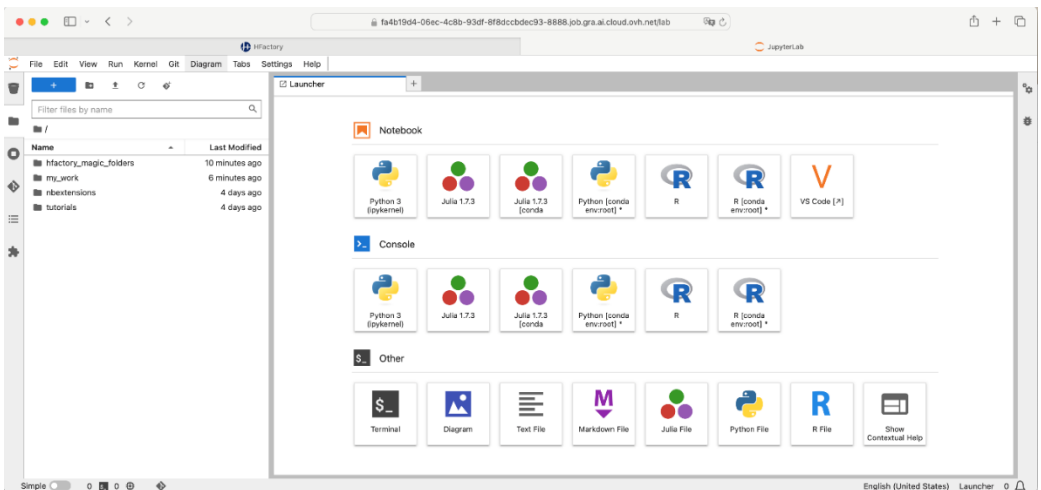
Within your compute instance, you will be able to choose the coding environment of your preference between **JupyterLab**, classic **Jupyter notebooks** and **VSCode**.



2. Personal storage

The full content of your home directory is persisted between sessions.

However to make it easier to structure you work, you will find that a folder called my_work has been automatically at the root.



As mentioned earlier, we also recommend making period checks of the content of your home directory and removing unnecessary files to keep instance start / stop time as low as possible.

To this end, we recommend using the **ls -la** command at root from the terminal: it will give you a clear picture of large files / directories as well as of any directory containing a high number of files.

3. Mounted shared volume

All of the resources for the Hi!ckathon #5 are in the **/hfactory_magic_folders** directory.

The datasets of the Hi!ckathon #5 competition are in the **~/hfactory_magic_folders/water_shortage_prediction** folder

As a participant, you only have **read-only** to the files the organisers will have placed there (trying to create or delete files there will throw back a permission error).

Please copy any file you want to modify inside /water_shortage_prediction in /shared_storage or anywhere you want inside your home directory (**my_work**) and execute them, using either JupyterLab, Jupyter Notebooks or VS Code.

For python **.py** files:

- From Jupyter, go to a terminal and run **python <python file>** from the relevant directory.
- From VS Code, run it using the triangle on the top right part of the screen. If you want to see the plot, run it with the Interactive mode (scroll the menu below the triangle)

4. Available libraries and packages

HFactory contains all the classic Data Science and AI packages and libraries (numpy, pandas, scikit-learn, ...). You do not need to reinstall them.

- You can get the list of installed packages by running in the terminal the **pip freeze** command.
- Additionally, you can perform the installation of any further package as long as that installation does not require root access.

Tools like Tensorboard, Streamlit, Shapash or Label Studio can easily be used with HFactory. After you install / launch them, you just need to make a copy of the URL of your instance in a new tab and then make two small modifications to access their web interface.

- **Step 1:** Replace the 8888 by whatever default port is used by the application (for example 8501 by Streamlit)

- **Step 2:** Remove everything after cloud.ovh.net from the original URL. It should look like https://6cc9a089-f938-48be-9efc-ac0c069461dc-8888.job.gra.training.ai.cloud.ovh.net/lab/tree/my_work if you are inside the my_work subdirectory from JupyterLab)

So in the case of Streamlit, the URL for the webapp will look like *https://6cc9a089-f938-48be-9efc-ac0c069461dc-8501.job.gra.training.ai.cloud.ovh.net*.

5. Checking GPU processes

When using a GPU-enabled instance, a useful command to list running GPU processes is nvidia-smi . For a real-time refresh of this info every two seconds, you can use watch **nvidia-smi .**