Introduction to Airflow

COVID Automation 🦠

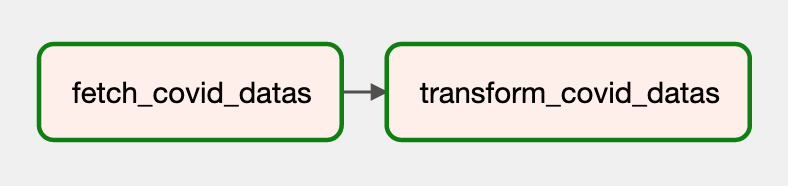
# **COVID Automation**

You are the first line of defence against the pandemic! You are a data-engineer for [Etalab](https://www.etalab.gouv.fr/), a branch of the French government that develops digital tools for french citizens.

You are asked to build a Airflow pipeline in order to automate data acquisition and data processing about COVID-19.

Your goal is to build a simple, yet useful, pipeline that will be used by the French government to monitor the COVID-19 pandemic.

The final DAG will look like this:



This DAG is responsible for the following tasks:

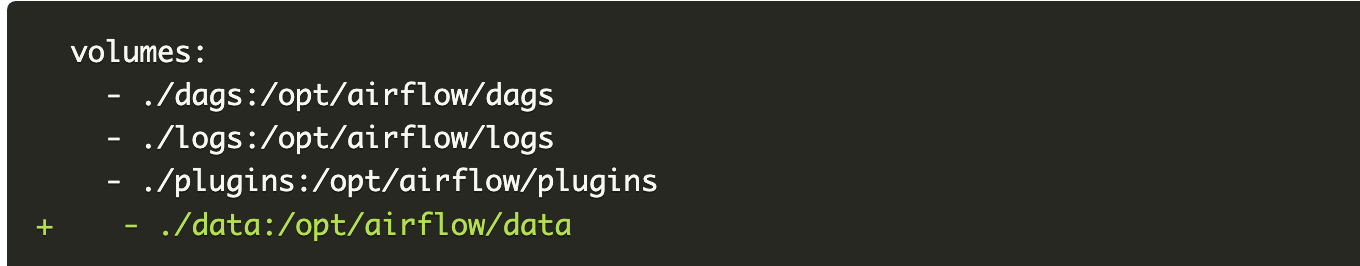
* Extract data from a URL (https://www.data.gouv.fr/fr/datasets/r/5c4e1452-3850-4b59-b11c-3dd51d7fb8b5) where you can download a CSV file with the latest COVID-19 data (which is updated every day). Dump the raw datas into a CSV file.
* Load the previous CSV, compute the *mean hospitalization in France per day* and export these statistic into a new CSV file.

## **Prepare the docker-compose.yaml**

Be sure you stoppped and clean all running Airflow containers. We need to restart a new one.

Create a new working directory and curl the docker-compose.yaml file.

You will need to add a new volume data:



Remember, you can set AIRFLOW\_\_CORE\_\_LOAD\_EXAMPLES to 'false' to disable the loading of the examples.

Create the data and dags folder if they don't exist.

Intialize the Airflow database and start the containers!

## **Explore the datas**

You can take a look at the raw datas before diving into the pipeline.

For example, create a new notebook and download the raw CSV file with:



And that's it!

You can have a full description of the dataset [here (in french)](https://www.data.gouv.fr/fr/datasets/synthese-des-indicateurs-de-suivi-de-lepidemie-covid-19/). Click on the link below to see a translation in english:

Click to expand!

## **First task**

The first task is to download the datas and dump them into a local CSV file in data for further transformations.

## **Second task**

The second task is to load the previous CSV file and compute the mean hospitalization in France per day.

The column of interest is: incid\_hosp.

The final CSV should look like this:

