Description of the conception and implementation

Tools used:

* ETL (Extract, Transform and Load the Data)
  + Spark / Databricks
* Data Visualization
  + Tableau Desktop (free trial version)

The Databricks pipeline is used for merging all files and adding few modifications:

* column labels
* a new column with the name of the listener.

When the pipeline is performed, we download the new data in a csv file.

Then, we can open this file directly in Tableau or open the file in a PostgreSQL database and then establish the connection with Tableau.

The calculations of the KPIs are directly done in Tableau.

GitHub repository link: https://github.com/hdubouillon/Application\_Of\_Big\_Data

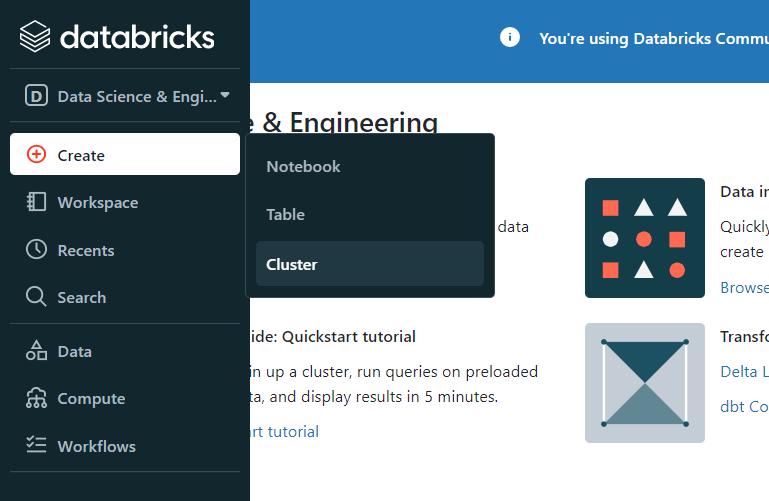
The GitHub repository include:

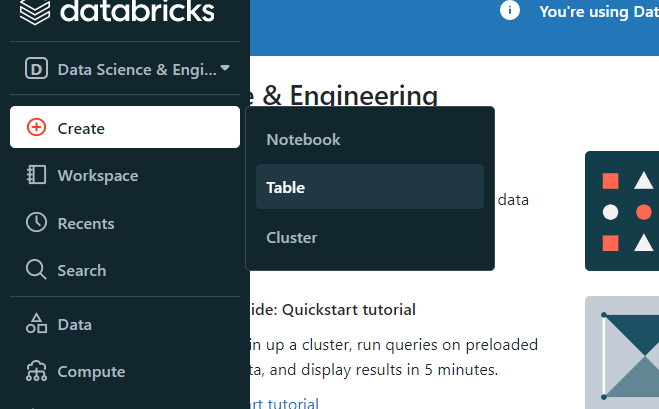
* this documentation
* 2 notebook files used for the pipelines in Databricks
* the Tableau file
* the full database csv file

The compilation process and the way to start and run the project

**ETL part on Databricks:**

Create a Databricks account, then a cluster and then a table.

****

****

Drag and drop all the files:

**Une image contenant texte, capture d’écran, ordinateur, intérieur

Description générée automatiquement**

Click on "create table in notebook":

**Une image contenant texte

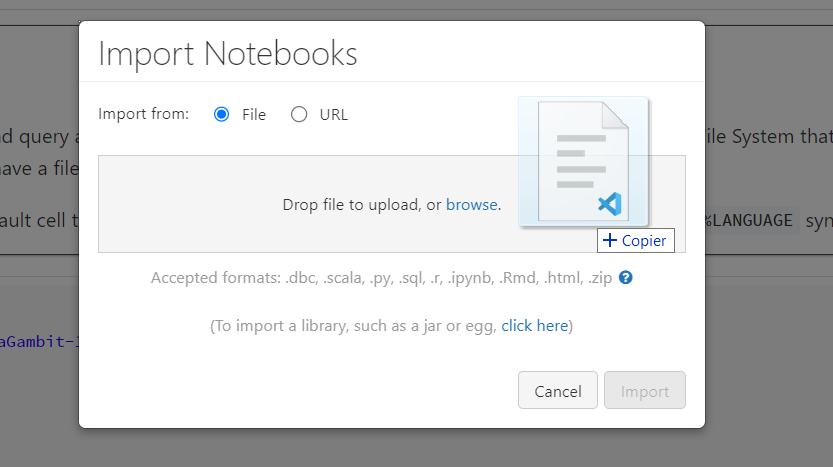
Description générée automatiquement**

Once the notebook is displayed, go to "file" then "import notebook":

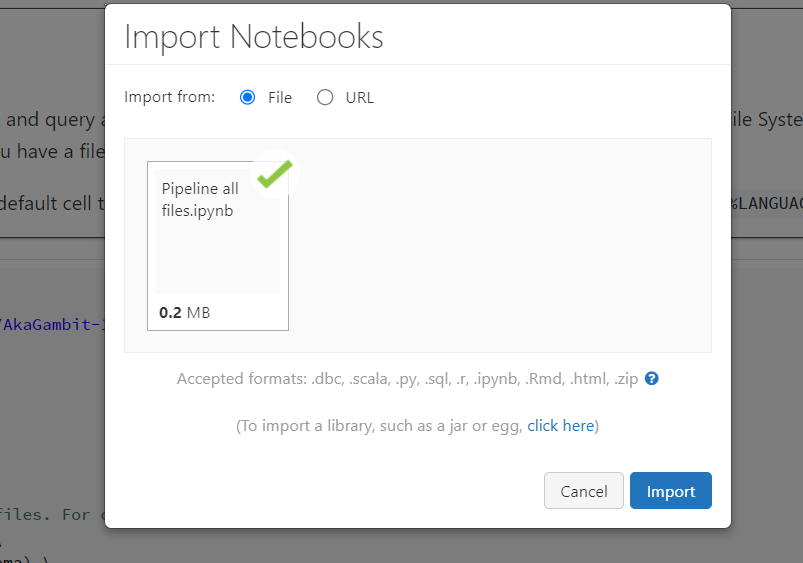
Une image contenant texte

Description générée automatiquement

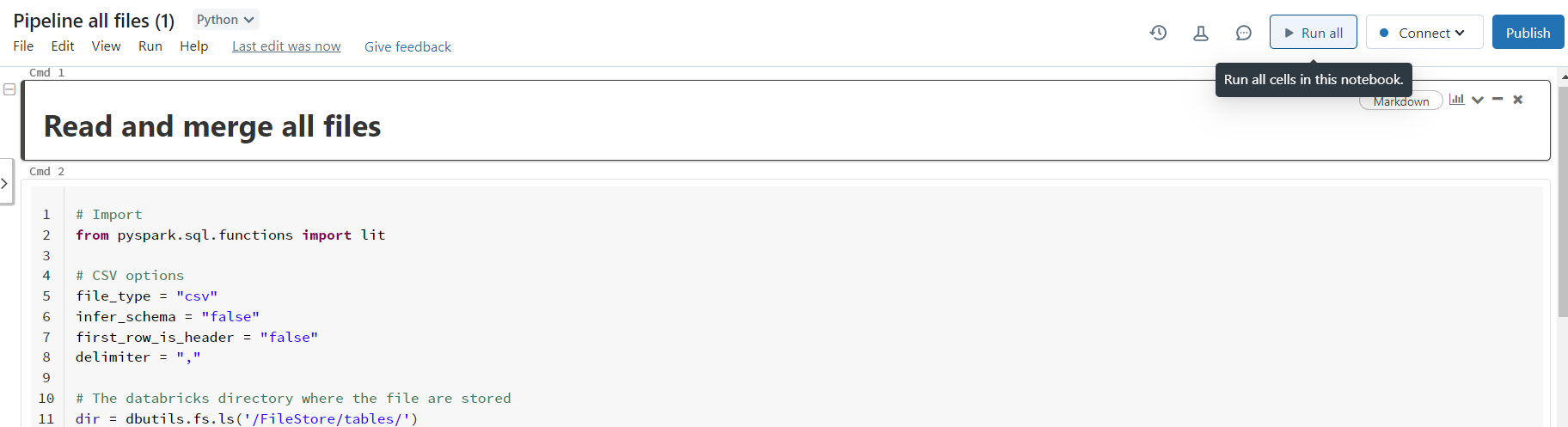
Drag the desired notebook ("Pipeline all files.ipynb" to create a table with all the files or "Pipeline new file.ipynb" to add a new file to an existing table):



Click on import:

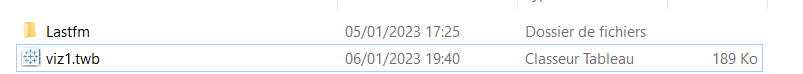


Once the notebook is open, click on "Run all":



Then download the entire table by clicking on "Download all rows...": Une image contenant table

Description générée automatiquement

**Open the Tableau file:**

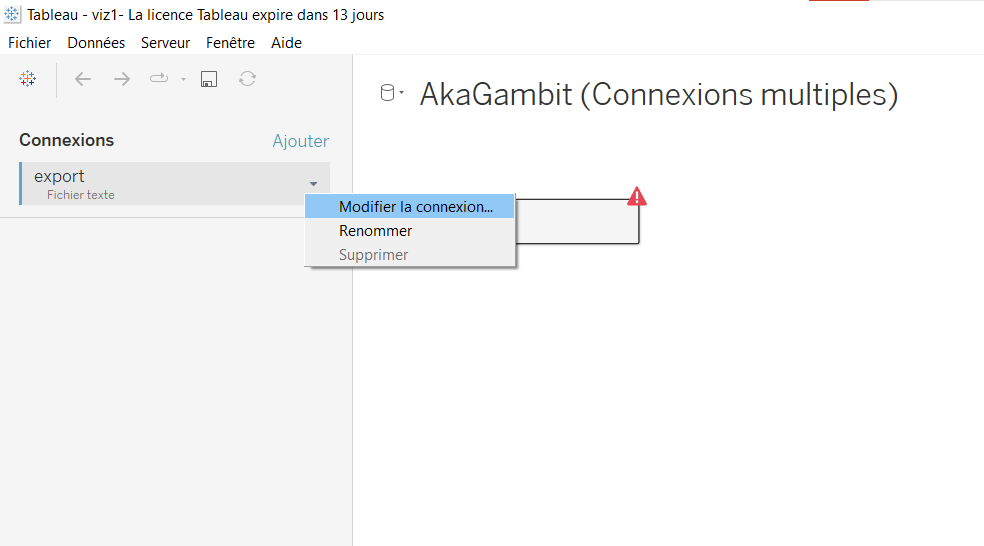
Une image contenant texte

Description générée automatiquement

Connect the database either with the csv file or with the PostgreSQL

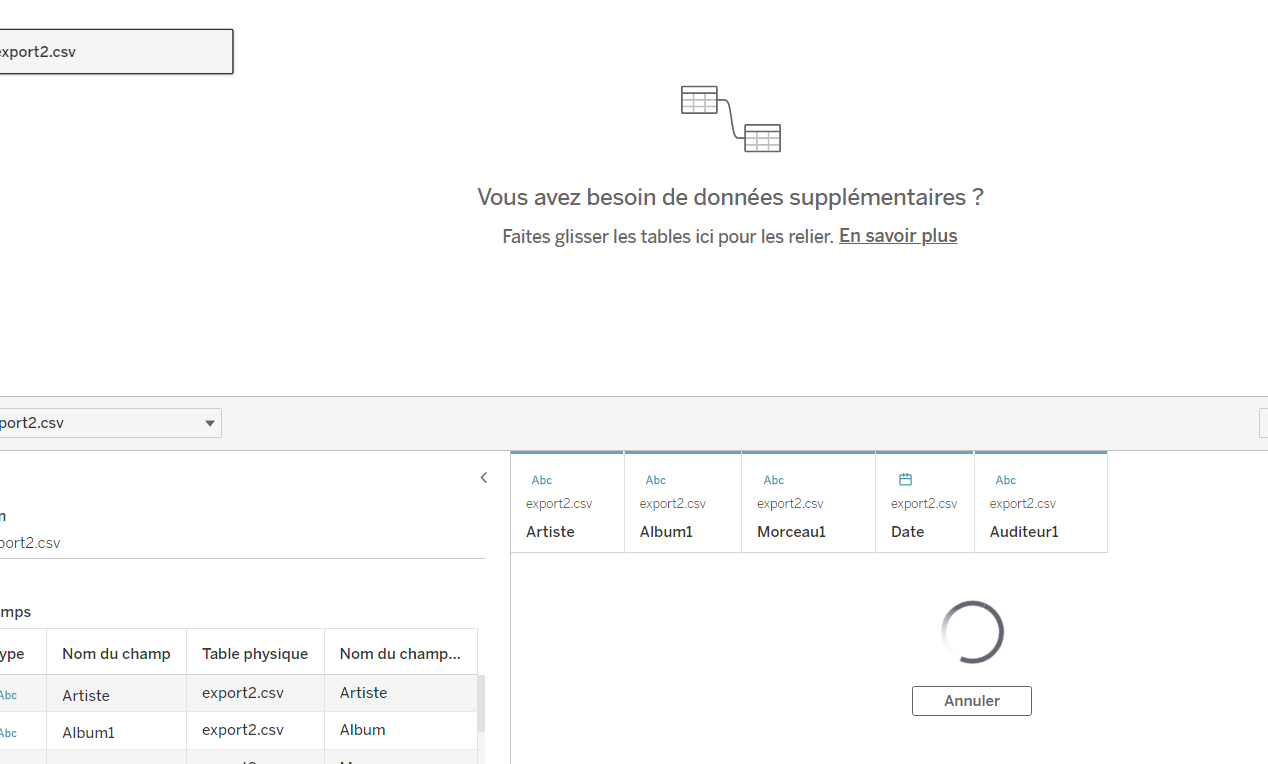
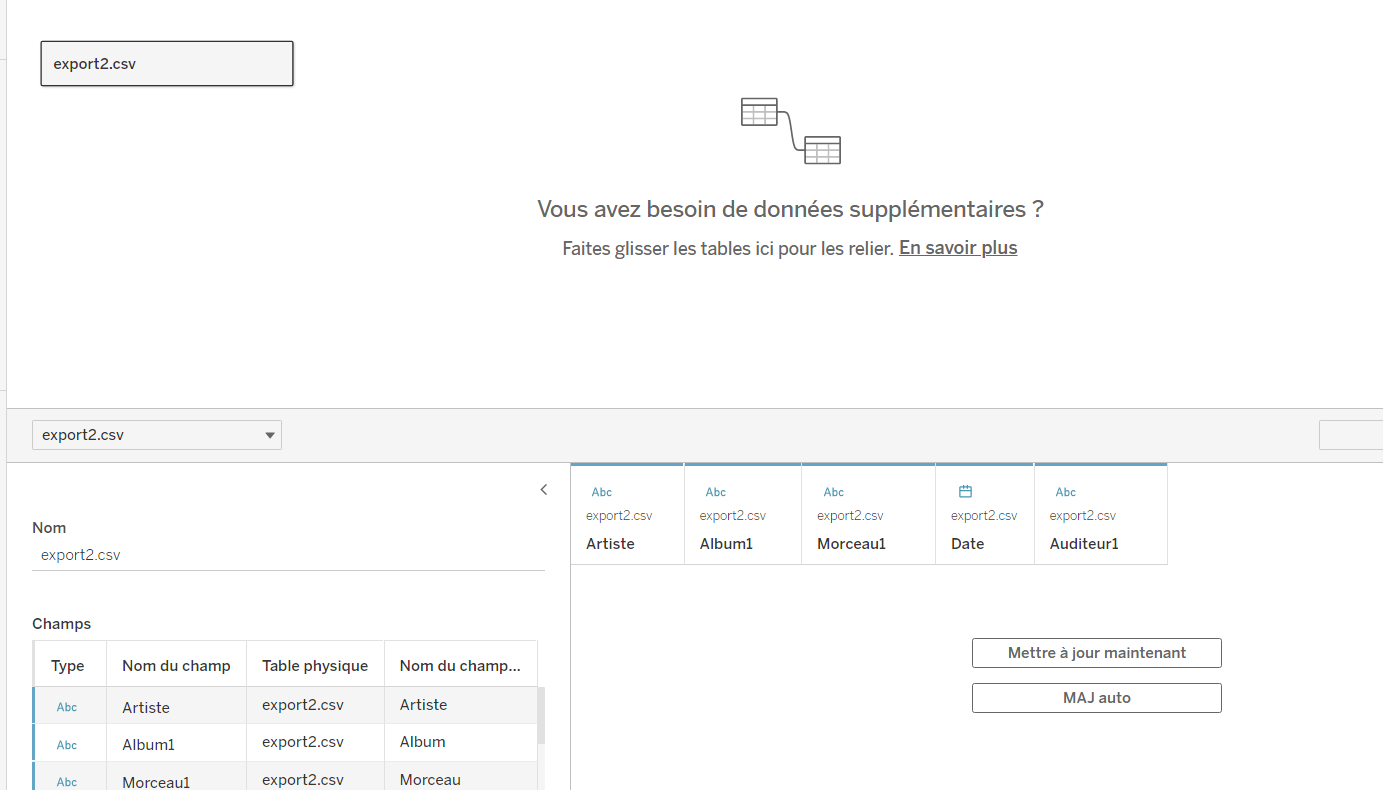
Option 1: csv:

Une image contenant texte

Description générée automatiquement

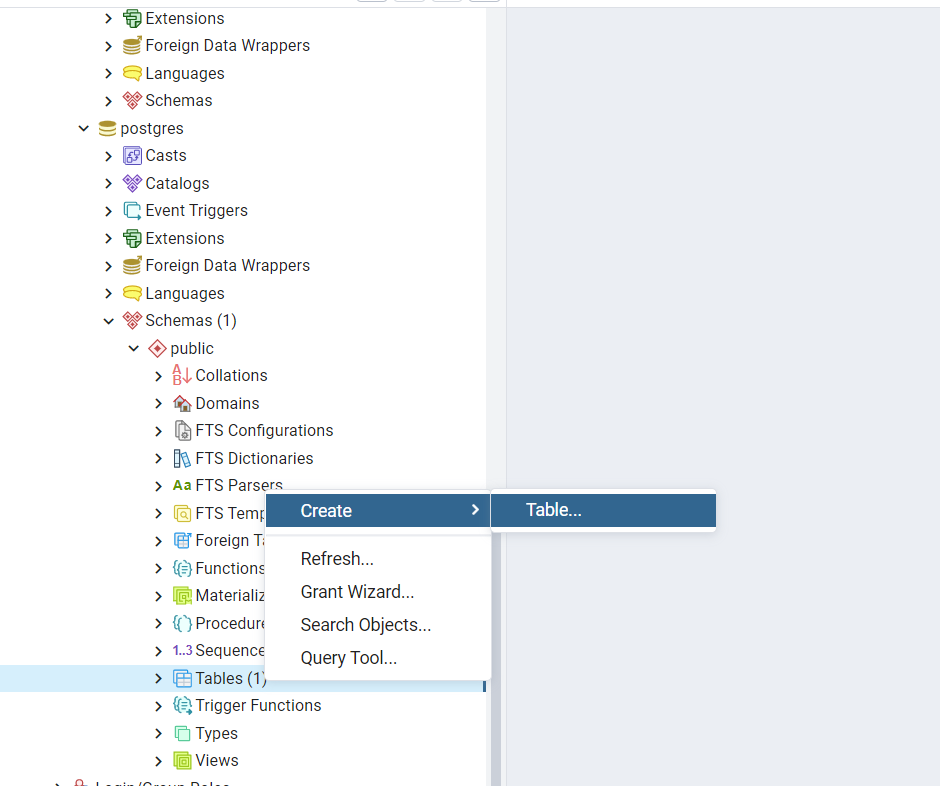
Une image contenant texte

Description générée automatiquement

Click on « mettre à jour maintenant »:Une image contenant texte, capture d’écran, intérieur

Description générée automatiquement

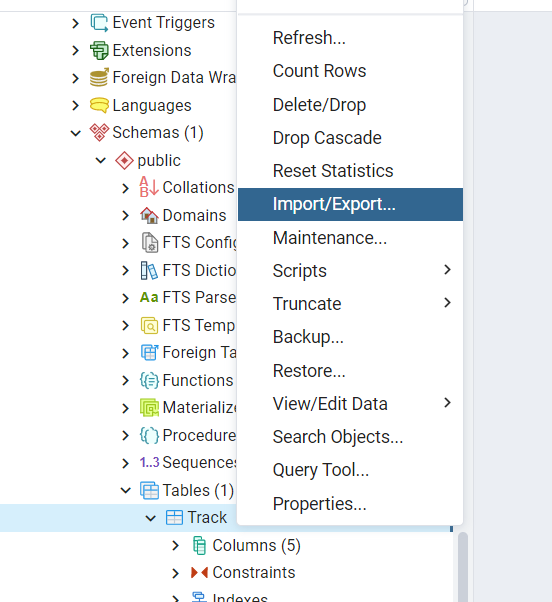
The file is ready to be used.

Option 2: PostgreSQL:

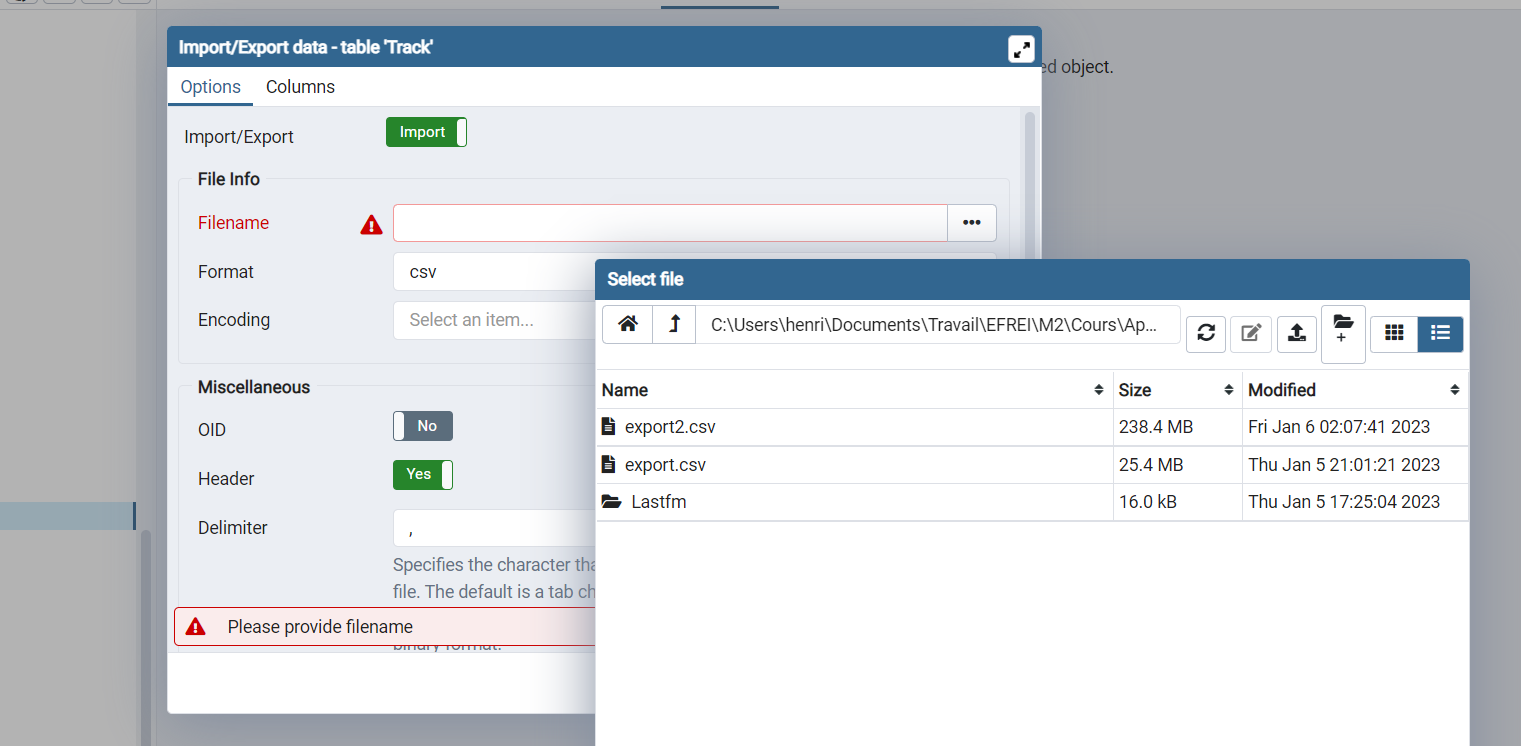
Une image contenant table

Description générée automatiquement

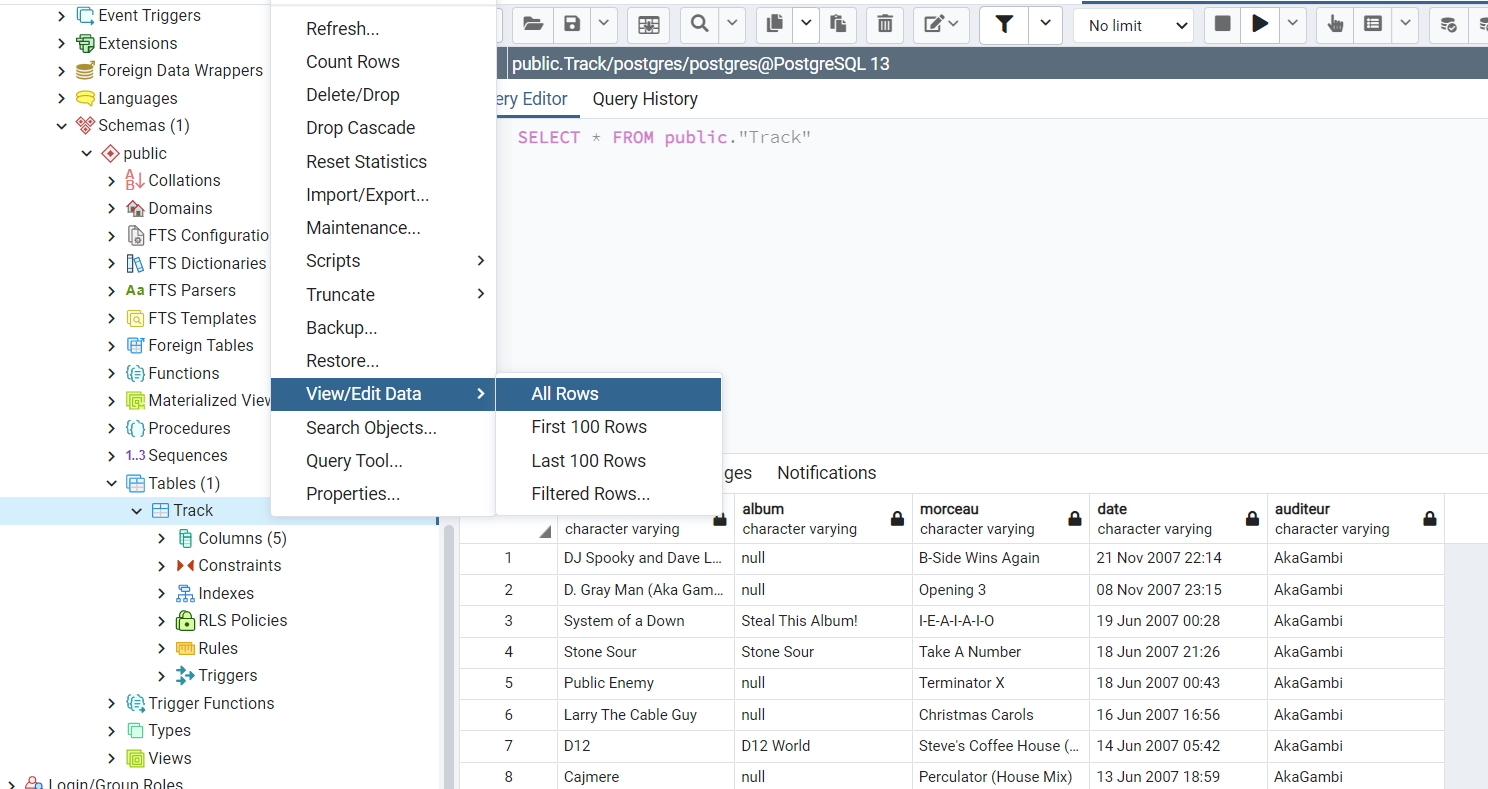
Right click on the table name, then click on export import:



Select the csv file with the options on the screenshot:



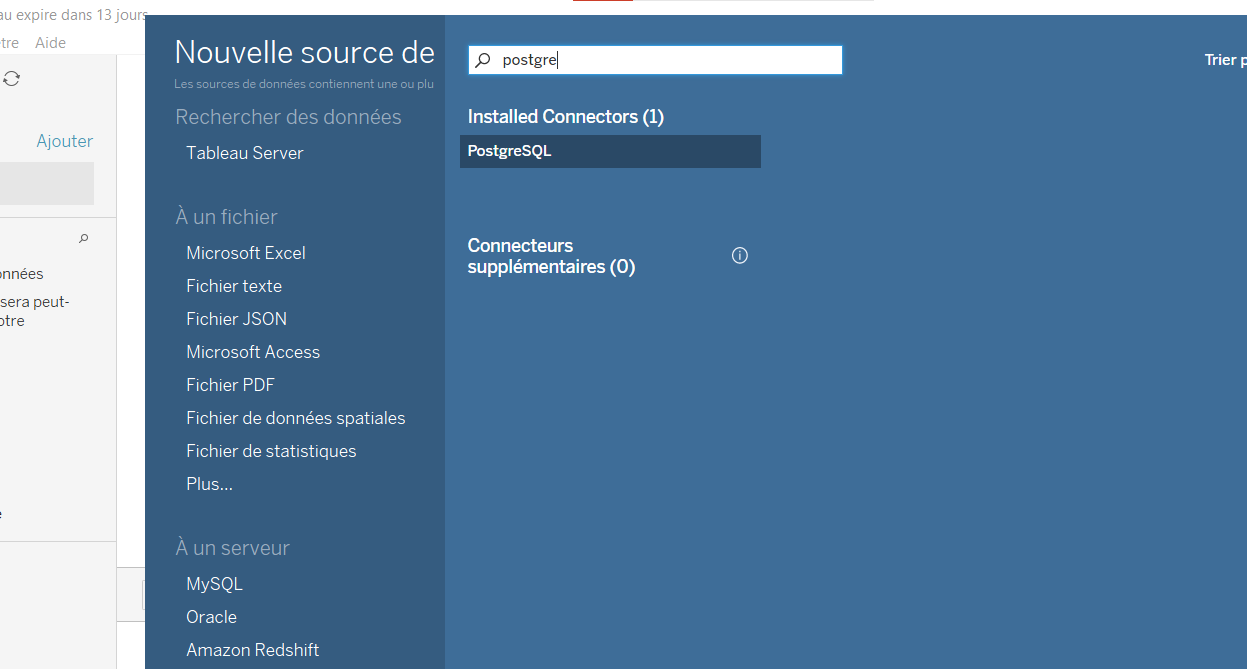
Check that the rows of the csv have been added:



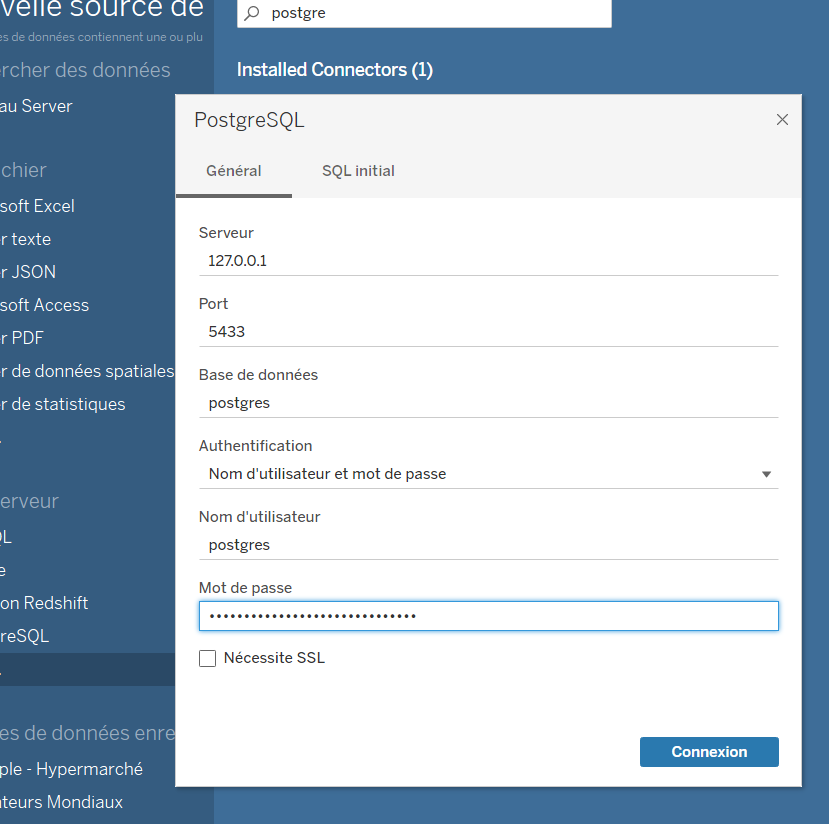
Return to Table and click on "New Data Source":

Une image contenant texte

Description générée automatiquement



Enter the PostgreSQL server information (port, table name, password...):

**Dataviz part on Tableau:**

Final version of the dashboard on Tableau:

