Projet Hadoop Big Data (Data Sciences)

Description du projet
Décembre 2023

Par Christophe GERMAIN

Description

Projet Big Data

Technologies

- Hadoop + python (HappyBase ...)
- Python Pandas
- Suggestions :

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt
```

Projet Big Data

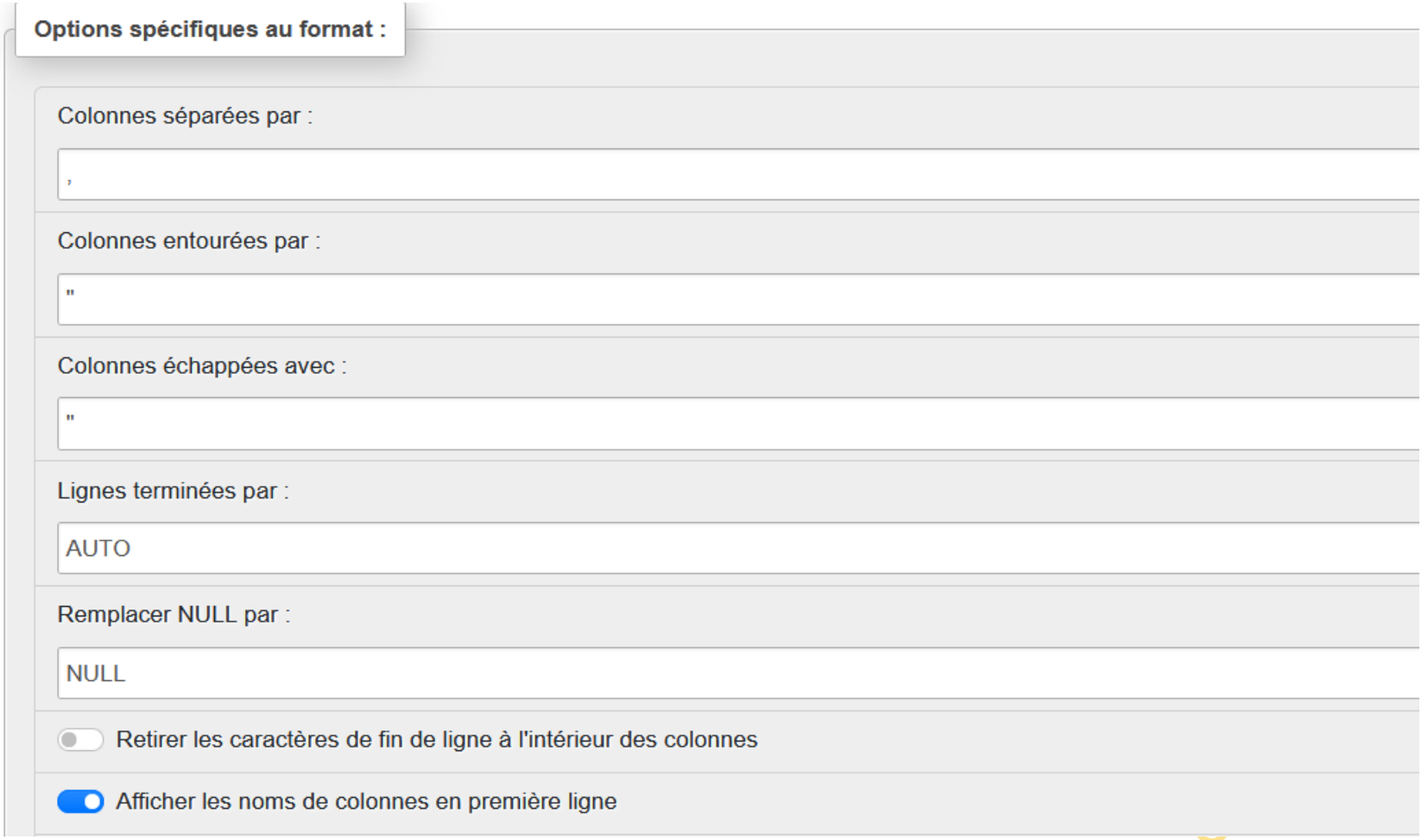
Objectifs :

- Le groupe doit livrer :
    - Un ensemble d'applications Big Data et Power BI
    - Un dossier comprenant :
        - L'analyse de la compréhension de la problématique
        - Des données qualifiées
        - Des procédures d'import des données
        - Des procédures de structuration
        - Des algorithmes d'analyse des données
        - Vos recommandations par rapport au déroulement du projet

Projet Big Data

Le projet :

- A partir du fichier csv : **dataw_fro.csv**
- Format du fichier :



Projet Big Data

Le projet (suite) :

- Entête du fichier :

	#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut
☐	1	**codcli**	int(11)			Non	*Aucun(e)*
☐	2	**genrecli**	varchar(8)	utf8mb4_general_ci		Oui	*NULL*
☐	3	**nomcli**	varchar(40)	utf8mb4_general_ci		Oui	*NULL*
☐	4	**prenomcli**	varchar(30)	utf8mb4_general_ci		Oui	*NULL*
☐	5	**cpcli**	varchar(5)	utf8mb4_general_ci		Oui	*NULL*
☐	6	**villecli**	varchar(50)	utf8mb4_general_ci		Oui	*NULL*
☐	7	**codcde**	int(11)			Non	*Aucun(e)*
☐	8	**datcde**	datetime			Oui	*NULL*
☐	9	**timbrecli**	float			Oui	*NULL*
☐	10	**timbrecde**	float			Oui	*NULL*
☐	11	**Nbcolis**	tinyint(4)			Oui	*NULL*
☐	12	**cheqcli**	float			Oui	*NULL*
☐	13	**barchive**	bit(1)			Oui	*NULL*
☐	14	**bstock**	bit(1)			Oui	*NULL*
☐	15	**codobj**	int(11)			Oui	*NULL*
☐	16	**qte**	smallint(6)			Oui	*NULL*

Projet Big Data

Le projet (suite) :

- Entête du fichier (suite) :

☐	17	**Colis**	int(11)	Oui	*NULL*
☐	18	**libobj**	varchar(50) utf8mb4_general_ci	Oui	*NULL*
☐	19	**Tailleobj**	varchar(50) utf8mb4_general_ci	Oui	*NULL*
☐	20	**Poidsobj**	double	Oui	*NULL*
☐	21	**points**	int(11)	Oui	*NULL*
☐	22	**indispobj**	bit(1)	Oui	*NULL*
☐	23	**libcondit**	varchar(50) utf8mb4_general_ci	Oui	*NULL*
☐	24	**prixcond**	double	Oui	*NULL*
☐	25	**puobj**	double	Oui	*NULL*

Projet Big Data

Le projet (suite) : LOT 1

- Contexte :
  - Une Fromagerie (le client) a un datawarehouse depuis 2004 qui est représenté par le fichier csv fournit dans ce document.
  - Créer des jobs pour limiter le flux d'information (Mapper-Reducer) pour obtenir uniquement les informations voulues pour répondre au besoin du client décrit ci-dessous :
  - Le client désire les statistiques suivantes :
    1. Filtrer les données selon les critères suivants :
       Entre 2006 et 2010,
       Avec uniquement les départements 53, 61 et 28
    2. A partir du point 1 : Ressortir dans un tableau des 100 meilleures commandes avec la ville, la somme des quantités des articles et la valeur de « timbrecde » (la notion de meilleures commandes : la somme des quantités la plus grande ainsi que le plus grand nombre de « timbrecde »)
    3. Exporter le résultat dans un fichier Excel.

Projet Big Data

Le projet (suite) : LOT 2

- Contexte :
  - (Comme le LOT 1)
  - Le client désire les statistiques suivantes :
    1. Filtrer les données selon les critères suivants :
       Entre 2011 et 2016,
       Avec uniquement les départements 22, 49 et 53
    2. A partir du point 1 : Ressortir de façon aléatoire de 5% des 100 meilleures commandes avec la ville, la somme des quantités des articles sans « timbrecli » (le timbrecli non renseigné ou à 0) avec la moyenne des quantités de chaque commande)
       Avoir un PDF avec un graphe (PIE) (par Ville)
    3. Exporter le résultat dans un fichier Excel.

Projet Big Data

Le projet (suite) : LOT 3

- Mettre en place une base NoSQL HBASE pour stocker le contenu du fichier CSV et de mettre en œuvre un moteur de recherche avec Power BI pour interroger ce Data Warehouse.
    - Pour répondre au Lot 1 et Lot 2 au niveau des résultats avec les graphes,
    - Mise en place d'un Dashboard interactif

Projet Big Data

Liens :

- **Python_Complet**
- **https://pandas.pydata.org/docs/getting_started/index.html**