



Projet Hadoop Big Data (Data Sciences)

Description du projet
JANVIER 2024

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 : [dataaw_fro.csv](#)
- Format du fichier :

Options spécifiques au format :

Colonnes séparées par :

,

Colonnes entourées par :

"

Colonnes échappées avec :

"

Lignes terminées par :

AUTO

Remplacer NULL par :

NULL

☐ Retirer les caractères de fin de ligne à l'intérieur des colonnes

☒ Afficher les noms de colonnes en première ligne

Projet Big Data

Le projet (suite) :

- Entête du fichier :

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

Projet Big Data

Le projet (suite) :

- Entête du fichier (suite) :

<input type="checkbox"/>	17	Colis	int(11)	Oui	NULL
<input type="checkbox"/>	18	libobj	varchar(50) utf8mb4_general_ci	Oui	NULL
<input type="checkbox"/>	19	Tailleobj	varchar(50) utf8mb4_general_ci	Oui	NULL
<input type="checkbox"/>	20	Poidsobj	double	Oui	NULL
<input type="checkbox"/>	21	points	int(11)	Oui	NULL
<input type="checkbox"/>	22	indispobj	bit(1)	Oui	NULL
<input type="checkbox"/>	23	libcondit	varchar(50) utf8mb4_general_ci	Oui	NULL
<input type="checkbox"/>	24	prixcond	double	Oui	NULL
<input type="checkbox"/>	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

(**De votre poste local** : interroger votre VM LINUX sur le port 9090 (port privé de votre VM => prendre votre endpoint public correspondant)

1. Mettre en place une base NoSQL HBASE pour stocker le contenu du fichier CSV afin d'interroger ce Data Warehouse avec des scripts python.
 - La meilleur commande de Nantes de l'année 2020.
 - Le nombre total de commandes effectuées entre 2010 et 2015, réparties par année
 - Le nom, le prénom, le nombre de commande et la somme des quantités d'objets du client qui a eu le plus de frais de timbrece.
2. Créer un programme python (avec Panda) pour créer des graphes en pdf et des tableaux Excel et csv de votre importation dans HBase :
 - Question 1 partie 1 du lot 3 en csv
 - Question 2 partie 1 du lot 3 en barplot matplotlib exporté en pdf
 - Question 3 partie 1 du lot 3 en excel

Projet Big Data

Le projet (suite) : LOT 4

(De votre poste local, importer dans le HBase de votre VM Linux)

(De votre VM Windows, utiliser Power BI)

- Mettre en œuvre un moteur de recherche avec Power BI pour interroger le Data Warehouse HBase.
 - Pour répondre au Lot 1 et Lot 2 au niveau des résultats avec les graphes,
 - Vous avez carte blanche pour créer d'autres graphes , d'autres types de requêtes avec par exemple de la géolocalisation etc.
 - Mise en place d'un Dashboard interactif

Projet Big Data

Liens :

- [Python_Complet](#)
- https://pandas.pydata.org/docs/getting_started/index.html