Projet Hadoop Big Data (Data Sciences)

Description du projet
SEPTEMBRE 2024

Par Christophe GERMAIN

Description

Projet Big Data

Table des matières

I. Technologies

- Hadoop + python (HappyBase …)
- Python Pandas
- Suggestions :
  - import numpy as np
  - import pandas as pd
  - import matplotlib.pyplot as plt

Projet Big Data

II. Objectifs :

- Le groupe doit livrer :
- Un ensemble d'applications Big Data et Power BI / ELK
- 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

III. Le projet :

A. A partir du fichier csv :

dataw_fro.csv fourni dans le dossier du projet

B. Format du fichier :

Projet Big Data

	#	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

☐	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

IV. Le projet : LOT 1

- Contexte :

  1. Une Fromagerie (le client) a un datawarehouse depuis 2004 qui est représenté par le fichier csv fournit dans ce document.

  2. Créer des jobs (job.sh avec l'utilisation hadoop jar …) 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 2008 et 2012,
     Avec uniquement les départements 53, 61,75 et 28

  2. A partir du point 1 : Ressortir dans un tableau les 10 Clients les plus fidèles (Sommes du produit (des points et des quantités) sur l'ensemble des commandes par client) : récupérer les colonnes suivantes : Nom, Prénom, Ville, département du client et, nom de l'objet & la quantité commandée

  3. Exporter le résultat dans un fichier Excel et les 10 graphes (en pdf) par client avec le % de répartitions des objets commandés (produits commandés).

V. Lot2 : HBase et Moteur de Recherche :

A. Importation HBase :

Importer toutes les données du fichier csv dans HBase

1. Contraintes :

- Ne pas mettre dans le JSON des données à NULL,
- Date invalide interdite (enregistrement complet refusé),
- Ne pas importer l'année 2004 (enregistrement complet refusé)

B. Choisir Power BI ou ELK :

Créer un Dashboard afin de manière dynamique de :

- Voir la fidélité des clients sur selon l'intervalle de dates,
- Voir le nombre d'objets commandés par année et les 5 meilleurs (Palmarès),
- Voir les départements les plus représentatifs du programme de fidélité de la fromagerie (Palmarès).

Projet Big Data

VI. Liens :

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