Projet Hadoop Big Data et BI (Data Sciences)

Description du projet

Par Christophe GERMAIN

Description

Projet Big Data et BI

Technologies

- Hadoop + python (HappyBase …)
- PowerBI
- Suggestions :

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

Import happybase
```

Projet Big Data et BI

Le groupe doit livrer :

- Un ensemble d'applications Big Data et PowerBI
- 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

ProjetBigDataetBI

Le projet :

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

ProjetBigDataetBI

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*

ProjetBigDataetBI

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 et BI

LOT 0 (travail en amont)

- Le travail du Data Analyst est de comprendre et de nettoyer les données fournies par le client.
- Liste non exhaustive des actions à entreprendre :
  - Enlever les accents (traitement en ASCII sous Linux)
  - Données corrompues (date invalide)
  - Comprendre le Data mining (DataWarehouse csv -> gestion des doublons...)
  - ...

Projet Big Data et BI

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 HDFS + streaming jar) 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 meilleure commande :
   (1) **La somme des quantités la plus grande**
   (2) **Le plus grand nombre de « timbrecde »**
3. Exporter le résultat dans un fichier Excel.

Projet Big Data et BI

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) + Moyenne des quantités de chaque commande

       Avoir un PDF avec un graphe (PIE) (secteur par Ville)

ProjetBigData

LOT 3

(**De votre VM** : interroger votre VM LINUX sur le port 9090) (Ou local, mais plus lent)

1. Mettre en place une base NoSQL HBASE pour stocker le contenu du fichier CSV

2. Interroger la base de données NoSQL HBASE avec des scripts python.
   - La meilleure 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 timbrecde.

3. 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 matplotib exporté en pdf
   - Question 3 partie 1 du lot 3 en excel

ProjetBigData

LOT 4

(De votre poste local ou VM Window, interroger la Base NoSQL HBase de votre VM Linux, lors du LOT 3)

Mettre en œuvre des dashboards PowerBI récupérant les données depuis 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.
- Mise en place d'un Dashboard interactif

ProjetBigData

Liens :

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