TP n8 en Mise en œuvre des Framework d'IA et Big data

Installation Pig en utilisant Docker

Pig est un langage de traitement de données de haut niveau conçu pour travailler avec des ensembles de données volumineuses et complexes dans l'écosystème Hadoop.
Pig a été développé par Yahoo! pour simplifier le traitement de données massives en fournissant un langage de script intuitif et convivial.
Pig utilise une syntaxe de type SQL pour simplifier le processus de manipulation de données.
Au lieu de coder en Java, les utilisateurs peuvent écrire des scripts Pig pour extraire, transformer et analyser des données dans un environnement Hadoop.
Pig utilise une approche de traitement en pipeline où chaque opération de transformation de données est chaînée avec la suivante pour former une série de tâches qui seront exécutées en parallèle sur le cluster Hadoop. Cela permet un traitement rapide et efficace des données massives.
Les avantages de Pig comprennent sa facilité d'utilisation, sa flexibilité et sa capacité à gérer des données structurées et non structurées. Il permet également aux utilisateurs de travailler avec des données provenant de différentes sources, y compris HDFS, HBase, Cassandra et Amazon S3.
Pig est également compatible avec d'autres outils Hadoop tels que Hive, qui permet aux utilisateurs de traiter des données de manière interactive et SQL-like, et Spark, qui fournit des fonctions de traitement de données en mémoire.

Assurez-vous d'avoir Docker installé sur votre machine. Si ce n'est pas le cas, vous pouvez télécharger Docker à partir de ce lien : https://www.docker.com/get-started.

Exercice 1.

1. Ouvrir un terminal et exécuter la commande suivante pour télécharger l'image Docker de Pig :

```
C:\Users\ADmiN> docker pull suhothayan/hadoop-spark-pig-hive:2.9.2
```

Cette commande va télécharger l'image Docker de Pig à partir du registre Docker Hub. Cela peut prendre quelques minutes.

2. Lancer un conteneur Docker à partir de l'image téléchargée :

```
C:\Users\ADmiN>docker run -it -p 50075:50075 -p 8088:8088 -p 8085:8085 suhothayan/hadoop-spark-pig-hive:2.9.2 bash
```

Cette commande va lancer un conteneur Docker interactif à partir de l'image téléchargée. L'option "-p" permet de publier les ports nécessaires à Pig pour fonctionner. Vous pouvez remplacer les numéros de port par ceux que vous préférez, à condition qu'ils ne soient pas déjà utilisés sur votre machine.

3. Lancer un autre interpreteur de commande, et verifier l'installtion du docker « hadoop-spark-pig-hive:2.9.2 »

```
C:\Users\ADmiN>docker container ls
```

4. Creer et copier un fichier texte (input.txt) dans le conteneur « 9f549a293b57»

```
C:\Users\ADmiN>docker cp c:\TPPIG\input.txt 9f549a293b57:/tmp/
```

5. Créer un fichier de script nommé "**test.pig**" contenant le code suivant :

```
A = LOAD 'input.txt' AS (word:chararray);
B = GROUP A BY word;
C = FOREACH B GENERATE group, COUNT(A);
STORE C INTO 'output.txt';
```

Ce script charge un fichier d'entrée contenant des mots, groupe les mots en fonction de leur valeur, compte le nombre d'occurrences de chaque mot et stocke le résultat dans un fichier de sortie.

6. Copier ce fichier dans le conteneur « 9f549a293b57»
7. Démarrer le conteneur

8. Copier le fichier .txt de /tmp vers le dossier d'entrée de hdfs /user/root
9. Copier aussi le fichier test.pig de /tmp vers le dossier d'entrée de hdfs /user/root
10. Afficher le contenu du fichier test.pig
11. A partir de l'invite de commande, lancer l'interpréteur Pig :

C:\Users\ADmiN> **pig**

On obtiendra le shell pig: **grunt>**

12. Exécuter le script par la commande suivante dans l'interpréteur Pig :

grunt>exec test.pig

Pour quitter l'interpréteur en tapant la commande suivante : **quit**
Pour quitter le conteneur Docker en tapant la commande suivante : **exit**

Exercice 2.

Nous allons utiliser le jeu de données client qui contient des informations sur les clients d'une entreprise, telles que leur nom, prénom, date de naissance, sexe, produit acheté, quantité et prix. Nous allons travailler sur ce jeu de données pour illustrer les différentes fonctionnalités de Pig.

1. Copier le fichier texte (client.txt) dans le conteneur « 9f549a293b57»
2. Copier le fichier client.txt de /tmp vers le dossier d'entrée de hdfs /user/root
3. Charger le fichier client.txt

grunt> client = LOAD 'client.txt' USING PigStorage(',') AS (id:int, nom:chararray, prenom:chararray, datenaiss:datetime, sexe:chararray, produit:chararray, qte:int, prix:float);

4. Afficher la structure de l'objet client : DESCRIBE client ;
5. Editer le contenu de l'objet client : DUMP client ;
6. Obtenir le schéma et le plan d'exécution d'un bag (client)
   grunt> **ILLUSTRATE** client
7. Filtrer des tuples à partir du résultat de Load: Editer que les produits Livre
8. Faire un groupement de tuples par produit, et afficher la structure de l'objet cree :
9. Appliquer un traitement à chaque groupe (tuple de prod) pour avoir les produits les plus vendus
10. Stocker les resultats dans le dossier output
11. Quel est le montant total des ventes ?
12. Quel est le client qui a dépensé le plus ?
13. Quel est le produit le plus cher ?

Exercice 3:

1. Charger le fichier intitulé employe.csv dans le répertoire /user/root de votre conteneur, avec le contenu suivant:
   1201, ahmed, 25
   1202, salma, 58
   1203, amina, 39
   1204, ali, 23
   1205, mourad, 23
2. Editer le schéma du ficher
3. Lister les 3 premières lignes dans le fichier
4. Lister les noms des employés
5. Compter le nombre des employés par age
6. Filter les données par âge (age>23)
7. Grouper les données par âge et compter le nombre de personnes pour chaque âge