

Université Assane SECK Ziguinchor

UFR : Sciences et Technologies

Département Informatique



Année 2022-2023

Master 2 Génie Logiciel

Big Data

TP PIG

TP PIG : Introduction à Pig pour le traitement de données massives

Objectif: Apprendre à utiliser Pig pour effectuer des opérations de traitement de données sur un jeu de données simple.

1. Conditions préalables

1.1. Installation du cluster Hadoop

Apache Pig est une plateforme construite sur Hadoop. Il est nécessaire d'installer d'abord hadoop sur votre PC.






2. Téléchargement d'Apache Pig

- Pour télécharger Apache Pig, vous devez vous rendre sur le lien suivant : <https://downloads.apache.org/pig/>

Pig Releases










Please make sure you're downloading from [a nearby mirror site](#), not from www.apache.org.

Older releases are available from the [archives](#).

Name	Last modified	Size	Description
 Parent Directory		-	
 latest/	2022-06-17 12:56	-	
 pig-0.16.0/	2022-06-17 12:56	-	
 pig-0.17.0/	2022-06-17 12:56	-	
 KEYS	2017-06-19 08:12	11K	

Si vous recherchez la dernière version, accédez au répertoire « dernière », puis téléchargez le fichier `pig-x.xx.x.tar.gz`

Index of /pig/pig-0.17.0

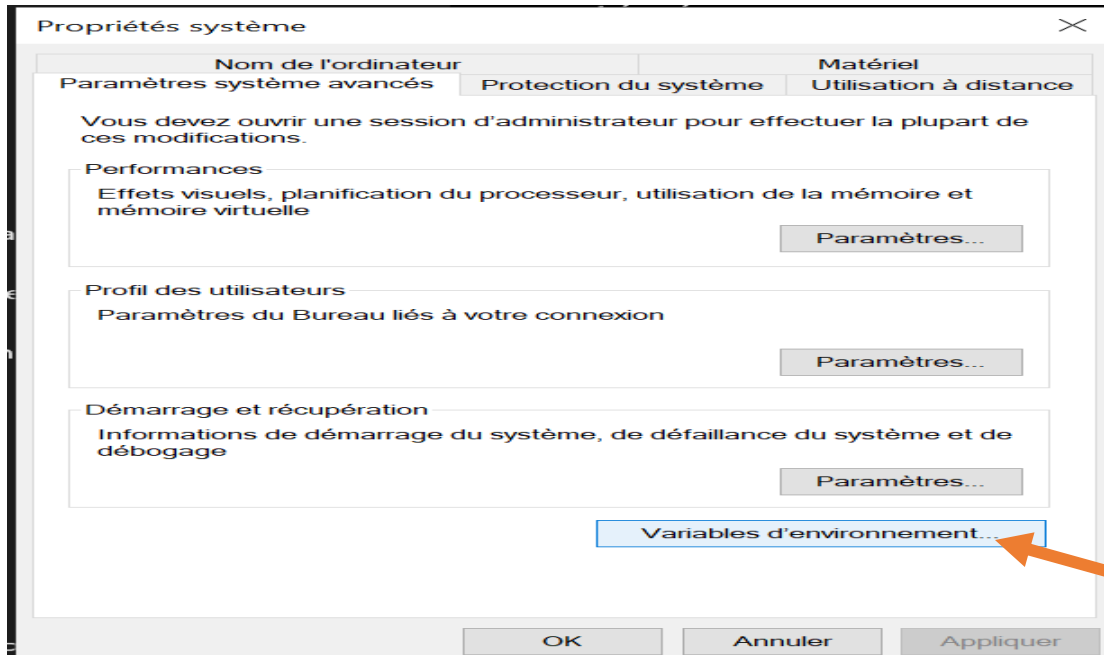
Name	Last modified	Size	Description
 Parent Directory		-	
 README.txt	2017-06-16 18:10	1.4K	
 RELEASE_NOTES.txt	2017-06-16 18:10	1.9K	
 pig-0.17.0-src.tar.gz	2017-06-16 18:11	15M	
 pig-0.17.0-src.tar.gz.asc	2017-06-16 18:11	488	
 pig-0.17.0-src.tar.gz.md5	2017-06-16 18:11	56	
 pig-0.17.0.tar.gz	2017-06-16 18:10	220M	
 pig-0.17.0.tar.gz.asc	2017-06-16 18:11	488	
 pig-0.17.0.tar.gz.md5	2017-06-16 18:11	52	

Une fois le fichier téléchargé, nous devons l'extraire le dossier Pig dans un répertoire tel que « C:\tp_pig»

3. Définition des variables d'environnement

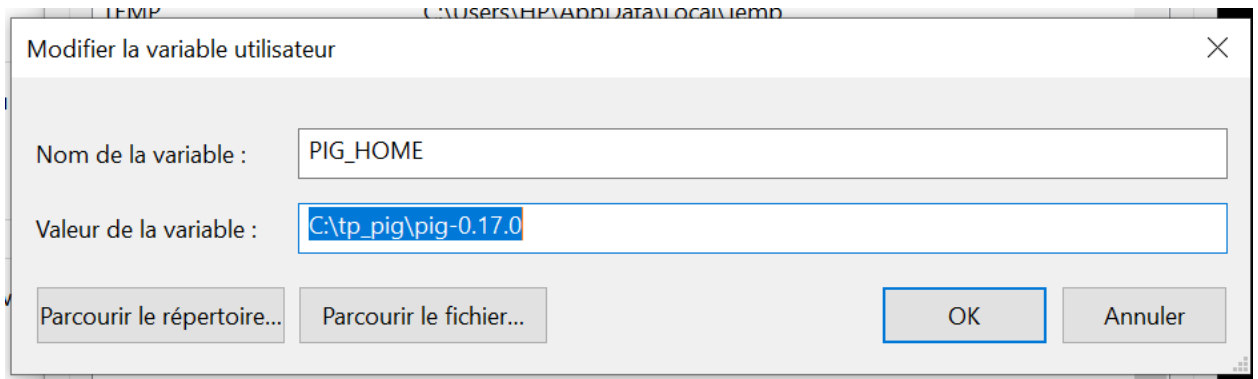
Après avoir extrait les archives, nous devons aller dans Panneau de configuration > Système et sécurité > Système. Cliquez ensuite sur « Paramètres système avancés.

Dans la boîte de dialogue des paramètres système avancés, cliquez sur le bouton « Variables d'environnement»

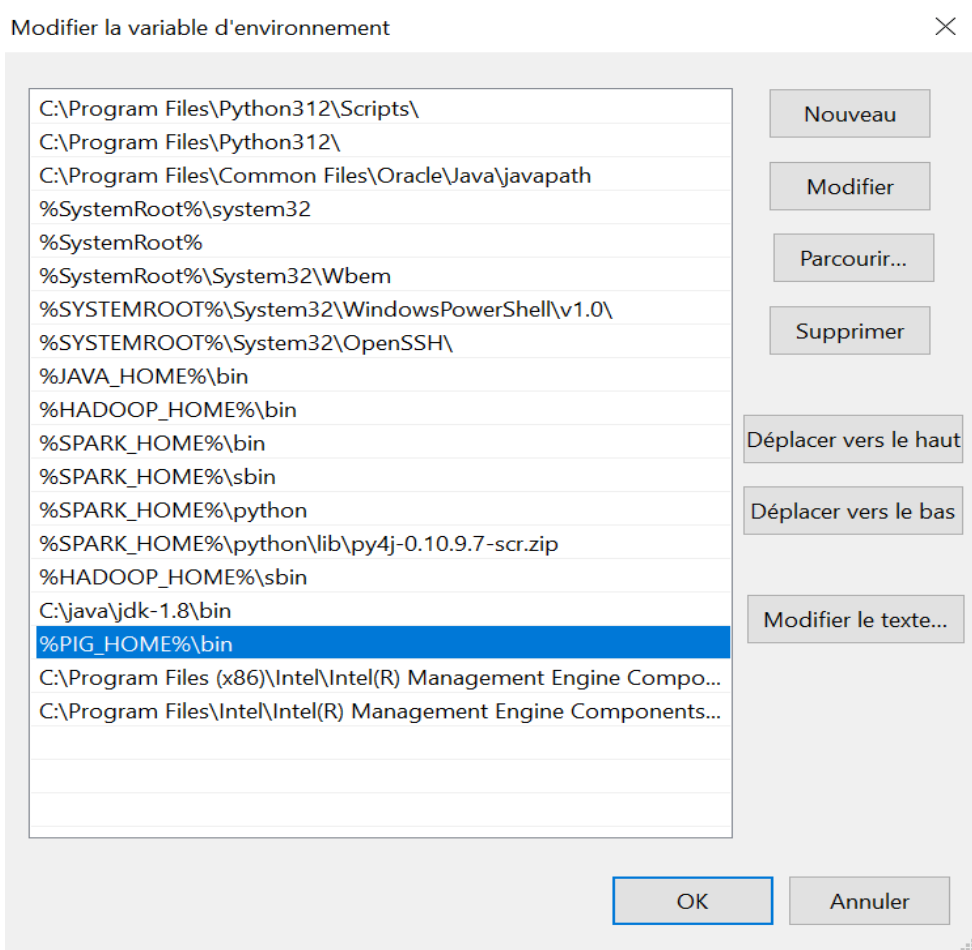


Nous devons maintenant ajouter les variables utilisateur suivantes :

- PIG_HOME : "PIG_HOME : "E:\hadoop-env\pig-0.17.0"



- %PIG_HOME%\bin



4. Démarrage d'Apache Pig

Après avoir défini les variables d'environnement, essayons d'exécuter Apache Pig.

Remarque : les services Hadoop doivent être en cours d'exécution

Ouvrez une invite de commande en tant qu'administrateur et exécutez la commande suivante : « **pig -version** ». Si tout se passe bien on obtient la version de pig installée

```
C:\Windows\system32>pig -version
La syntaxe du nom de fichier, de répertoire ou de volume est incorrecte.
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58

C:\Windows\system32>
```

Par contre si vous recevez une exception, nous devons éditer le fichier pig.cmd situé dans le répertoire « pig-0.17.0\bin » en changeant la valeur HADOOP_BIN_PATH de « %HADOOP_HOME%\bin » à « %HADOOP_HOME%\libexec ».

```
9
0  setlocal enabledelayedexpansion
1
2  set HADOOP_BIN_PATH=%HADOOP_HOME%\libexec
3
4  set hadoop-config-script=%HADOOP_BIN_PATH%\hadoop-config.cmd
5  call %hadoop-config-script%
6
```

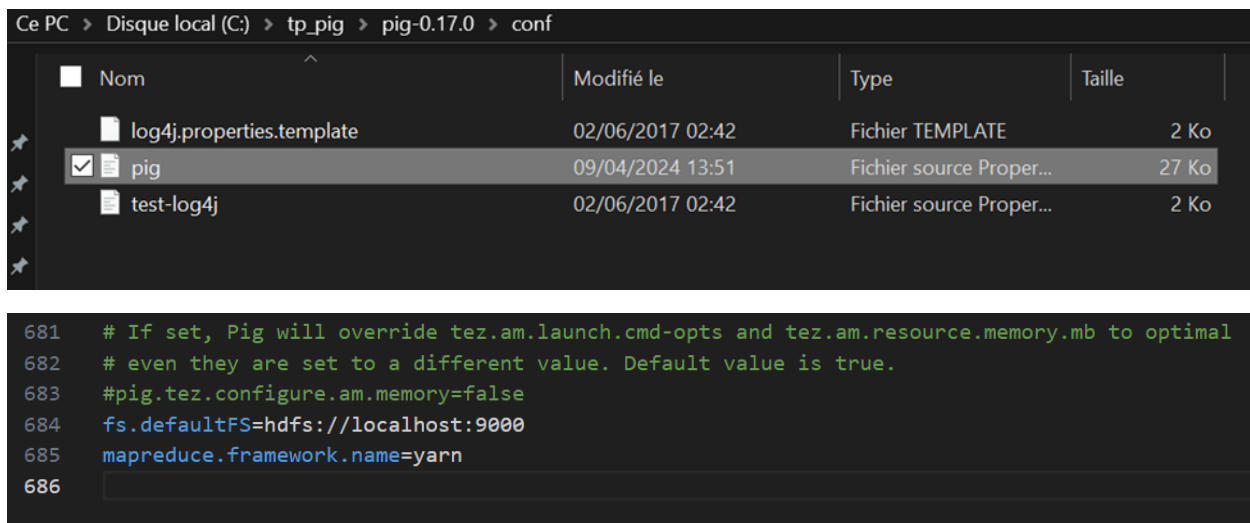
Après cela nous pouvons exécuter à nouveau la commande « pig -version »

Maintenant exécuton la commande « **pig** » qui va nous amener le shell de pig pour pouvoir y exécuter des fichiers avec l'extension « .pig »

```
C:\Windows\system32>pig
La syntaxe du nom de fichier, de répertoire ou de volume est incorrecte.
2024-04-09 04:59:11,781 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-04-09 04:59:11,785 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-04-09 04:59:11,785 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-04-09 04:59:12,243 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-04-09 04:59:12,243 [main] INFO org.apache.pig.Main - Logging error messages to: C:\Users\HP\Downloads\hadoop\logs\pig_1712678352220.log
2024-04-09 04:59:12,301 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\HP/.pigbootup not found
2024-04-09 04:59:12,929 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-04-09 04:59:12,929 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-04-09 04:59:13,917 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-e4859b45-7842-45b8-8e05-ce27b947b0af
2024-04-09 04:59:13,917 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> _
```

5. Configuration de Pig

Pour configurer Pig pour qu'il fonctionne avec votre installation Hadoop il faut modifier le fichier **pig.properties** (généralement situé dans le répertoire **conf** de Pig) en y ajoutant les propriétés suivantes :



Ces configurations indiquent à Pig d'utiliser HDFS comme système de fichiers par défaut et de soumettre les jobs à YARN.

6. Exécution de Pig

Pour exécuter Pig sur votre cluster Hadoop, il faut :

- Lancez Hadoop (assurez-vous que tous les services Hadoop nécessaires sont en cours d'exécution, y compris HDFS et YARN) en utilisant les commandes « start-dfs.cmd » et « start-yarn.cmd ».
- Utilisez la commande « **pig** » pour démarrer l'interface Pig interactif ou exécuter des scripts Pig.

Nous allons considérer un **Scénario d' Analyse de données de vente en ligne et écrire un script qui va nous montrer quelques opérations qu'on peut réaliser avec pig sur un jeu de données :**

Supposons que nous ayons un ensemble de données représentant les transactions de vente en ligne. Nous allons utiliser Apache Pig pour effectuer quelques opérations d'analyse sur cet ensemble de données.

Étape 1 : Chargement des données

Nous avons créer un fichier **sales_data.csv** contenant des données de vente en ligne

user_id	product_id	category	price	quantity	timestamp
101	P001	Electronics	1200	1	2024-03-15 10:30:00
102	P002	Clothing	500	2	2024-03-15 11:45:00
103	P003	Home	800	1	2024-03-16 09:15:00
101	P004	Electronics	1500	1	2024-03-16 12:00:00
104	P005	Books	300	3	2024-03-17 14:20:00
102	P006	Clothing	700	1	2024-03-18 16:00:00

Étape 2 : Écriture du script Pig

Créez un script Pig nommé « **tpPig.pig** » pour analyser les données de vente.

```

-- Charger les données depuis le fichier CSV
sales = LOAD 'hdfs://localhost:9000/user/HP/sales_data.csv' USING PigStorage(',') AS (
    user_id:chararray,
    product_id:chararray,
    category:chararray,
    price:float,
    quantity:int,
    timestamp:chararray
);

-- Filtrer les ventes dans la catégorie "Electronics"
electronics_sales = FILTER sales BY category == 'Electronics';

-- Calculer le montant total des ventes par utilisateur
user_sales = FOREACH (GROUP electronics_sales BY user_id) {
    total_sales = SUM(electronics_sales.price * electronics_sales.quantity);
    GENERATE group AS user_id, total_sales AS total_sales;
};

-- Classer les utilisateurs en fonction du montant total des ventes (descendant)
sorted_users = ORDER user_sales BY total_sales DESC;

-- Définir le chemin de sortie dans HDFS (ne pas utiliser un chemin local)
output_path = 'hdfs://localhost:9000/user/HP/output';

-- Stocker les résultats dans un fichier de sortie sur HDFS
STORE sorted_users INTO output_path USING PigStorage(',');

-- Afficher un message de confirmation
DESCRIBE sorted_users;

```

Dans ce script :

- Nous chargeons les données à partir du fichier CSV en spécifiant les délimiteurs et les types de colonnes.
- Ensuite, nous regroupons les données par catégorie de produit à l'aide de l'opérateur **GROUP**.
- Nous calculons le chiffre d'affaires total pour chaque catégorie en multipliant le prix par la quantité, puis en sommant ces valeurs.
- Enfin, nous affichons les résultats avec **DUMP**.

Étape 3 : Exécution du script Pig

Pour exécuter ce script sur votre environnement Apache Pig, on utilise la commande suivante dans le terminal : **pig -f nom_script.pig**

Résultat attendu :

Après avoir exécuté le script, vous devriez voir une sortie qui ressemble à ceci :

```
(Electronics, 2700.0)
(Clothing, 1900.0)
(Home, 800.0)
(Books, 900.0)
```

Remarques finales

Ce scénario simple illustre comment vous pouvez utiliser Apache Pig pour effectuer des analyses de base sur des données structurées. En fonction de vos besoins, vous pouvez étendre ce script en ajoutant d'autres opérations Pig telles que le filtrage, le tri, ou même des jointures avec d'autres ensembles de données. En expérimentant avec des scénarios similaires et en explorant les fonctionnalités avancées d'Apache Pig, vous pourrez construire des applications plus complexes et utiles pour le traitement de données distribuées.