

# Rapport détaillé – TP Pig avec Docker

## Prérequis techniques

- Windows 10 avec **WSL2** activé.
- **Docker Desktop** installé et configuré.
- **Git** installé pour cloner et gérer le projet si besoin.

---

## Étape 1 : Installation et vérification de Docker

1. Téléchargement et installation de **Docker Desktop** depuis [docker.com](https://docker.com).
2. Vérification de l'installation avec la commande :
3. `docker --version`

---

## Étape 2 : Téléchargement de l'image Docker Pig

Nous allons utiliser l'image **suhothayan/hadoop-spark-pig-hive:2.9.2**, qui contient :

- Hadoop
- Spark
- Pig
- Hive

Téléchargement :

```
docker pull suhothayan/hadoop-spark-pig-hive:2.9.2
```

---

## Étape 3 : Lancement du conteneur Docker

Exécution du conteneur avec les ports exposés :

```
docker run -it -p 50070:50070 -p 8088:8088 -p 8080:8080 suhothayan/hadoop-spark-pig-hive:2.9.2 bash
```

```
C:\Users\AGX\Desktop\Pig>docker run -it -p 50070:50070 -p 8088:8088 -p 8080:8080 suhothayan/hadoop-spark-pig-hive:2.9.2 bash
/etc/bootstrap.sh: line 9: /usr/local/spark/conf/spark-env.sh: Permission denied
/
* Starting OpenBSD Secure Shell server sshd                                     [ OK ]
Waiting for hdfs to exit from safemode
Safe mode is OFF
Started
root@07dde0c2a77:/#
```

Une fois à l'intérieur, nous disposons d'un terminal Linux avec Hadoop et Pig prêts à être utilisés.

---

## Étape 4 : Préparation des données

1. Création du fichier client.txt :

```
echo "Dupont;Pierre;1990-05-15;M;Ordinateur;1;800
Martin;Marie;1985-12-22;F;Smartphone;2;500
Bernard;Luc;1978-03-08;M;Tablette;1;300
Dubois;Sophie;1992-07-30;F;Ordinateur;1;800
Moreau;Jean;1980-11-14;M;Smartphone;3;750
Laurent;Claire;1995-02-18;F;Tablette;2;600
Simon;Philippe;1988-09-25;M;Ordinateur;2;1600
Michel;Julie;1983-06-11;F;Smartphone;1;250" > client.txt
```

```
root@07ddde0c2a77:/# echo "Dupont;Pierre;1990-05-15;M;Ordinateur;1;800
> Martin;Marie;1985-12-22;F;Smartphone;2;500
> Bernard;Luc;1978-03-08;M;Tablette;1;300
> Dubois;Sophie;1992-07-30;F;Ordinateur;1;800
> Moreau;Jean;1980-11-14;M;Smartphone;3;750
> Laurent;Claire;1995-02-18;F;Tablette;2;600
> Simon;Philippe;1988-09-25;M;Ordinateur;2;1600
> Michel;Julie;1983-06-11;F;Smartphone;1;250" > client.txt
```

2. Copier le fichier dans **HDFS** :
3. `hdfs dfs -put client.txt /client.txt`
4. Vérifier la présence du fichier :
5. `hdfs dfs -ls /`
6. pig ( lancer pig)

```
root@07ddde0c2a77:/# hdfs dfs -put client.txt /client.txt
root@07ddde0c2a77:/# hdfs dfs -ls /
Found 2 items
-rw-r--r--    1 root supergroup          345 2025-08-30 11:01 /client.txt
drwxr-xr-x   - root supergroup           0 2019-07-21 16:09 /user
root@07ddde0c2a77:/# pig
```

---

## Étape 5 : Lancement de Pig et traitement des données

### Question 1 : Combien de clients ont acheté des produits ?

```
clients = LOAD '/client.txt' USING PigStorage(';')
```

```
AS (nom:chararray, prenom:chararray, date_naissance:chararray,
```

sexe:chararray, produit:chararray, quantite:int, prix:int);

clients\_distincts = DISTINCT clients;

count\_clients = FOREACH (GROUP clients\_distincts ALL) GENERATE COUNT(clients\_distincts);

DUMP count\_clients;

✅ Résultat : 8 clients distincts.

```
HadoopVersion  PigVersion  UserId  StartedAt      FinishedAt      Features
2.9.2   0.17.0   root   2025-08-30 11:06:41  2025-08-30 11:06:42  GROUP_BY,DISTINCT

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime
edianReduceTime Alias  Feature Outputs
job_local1000002983_0001  1  1  n/a  n/a  n/a  n/a  n/a  n/a  n/a  clients
DISTINCT
job_local1413787181_0002  1  1  n/a  n/a  n/a  n/a  n/a  n/a  n/a  1-9,count_clients
GROUP_BY,COMBINER  file:/tmp/temp-736790497/tmp1480064867,

Input(s):
Successfully read 8 records from: "file:///client.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp-736790497/tmp1480064867"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1000002983_0001  ->  job_local1413787181_0002,
job_local1413787181_0002

2025-08-30 11:06:42,527 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:06:42,528 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:06:42,529 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:06:42,535 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:06:42,536 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:06:42,537 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:06:42,540 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success!
2025-08-30 11:06:42,543 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-08-30 11:06:42,543 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-08-30 11:06:42,555 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-08-30 11:06:42,555 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
(8)
grunt>
```

## Question 2 : Produits les plus vendus

produits\_groupes = GROUP clients BY produit;

ventes\_par\_produit = FOREACH produits\_groupes GENERATE group as produit,  
SUM(clients.quantite) as quantite\_totale;

ventes\_triees = ORDER ventes\_par\_produit BY quantite\_totale DESC;

DUMP ventes\_triees;

✅ Résultat attendu :

- Smartphone : 6 unités
- Ordinateur : 4 unités
- Tablette : 3 unités

```

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime
edianReductetime Alias  Feature Outputs
job_local733818384_0003 1 1 n/a n/a n/a n/a n/a n/a clients,produits_groupes,ventes_par_produit
GROUP BY,COMBINER
job_local1801413822_0004 1 1 n/a n/a n/a n/a n/a n/a ventes_triees SAMPLER
job_local918821747_0005 1 1 n/a n/a n/a n/a n/a n/a ventes_triees ORDER BY
file:/tmp/temp-736790497/tmp921036917,

Input(s):
Successfully read 8 records from: "file:///client.txt"

Output(s):
Successfully stored 3 records in: "file:/tmp/temp-736790497/tmp921036917"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local733818384_0003 -> job_local1801413822_0004,
job_local1801413822_0004 -> job_local918821747_0005,
job_local918821747_0005

2025-08-30 11:17:00,662 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,663 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,663 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,665 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,666 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,666 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,668 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,669 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,670 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:00,671 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-08-30 11:17:00,672 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-08-30 11:17:00,672 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-08-30 11:17:00,684 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-08-30 11:17:00,684 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Smartphone,6)
(Ordinateur,4)
(Tablette,3)
grunts>

```

### Question 3 : Montant total des ventes

ventes\_avec\_montant = FOREACH clients GENERATE nom, prenom, produit, quantite, prix, quantite \* prix as montant\_total;

montant\_global = FOREACH (GROUP ventes\_avec\_montant ALL) GENERATE SUM(ventes\_avec\_montant.montant\_total) as chiffre\_affaires;

DUMP montant\_global;

✅ Résultat attendu : 9 800 €

```

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime
edianReductetime Alias  Feature Outputs
job_local1998648276_0006 1 1 n/a n/a n/a n/a n/a n/a 1-38,clients,montant_global,ventes_avec_montant
GROUP BY,COMBINER
file:/tmp/temp-736790497/tmp-516986242,

Input(s):
Successfully read 8 records from: "file:///client.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp-736790497/tmp-516986242"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1998648276_0006

2025-08-30 11:17:49,838 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:49,839 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:49,839 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:17:49,840 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-08-30 11:17:49,841 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-08-30 11:17:49,841 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-08-30 11:17:49,860 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-08-30 11:17:49,860 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(9800)

```

#### Question 4 : Client qui a dépensé le plus

depenses\_clients = FOREACH clients GENERATE nom, prenom, quantite \* prix as  
montant\_depense;

clients\_groupes = GROUP depenses\_clients BY (nom, prenom);

total\_par\_client = FOREACH clients\_groupes GENERATE group.nom as nom, group.prenom as  
prenom, SUM(depenses\_clients.montant\_depense) as total\_depense;

clients\_tries = ORDER total\_par\_client BY total\_depense DESC;

top\_client = LIMIT clients\_tries 1;

DUMP top\_client;

✓ Résultat attendu : **Simon Philippe – 3200 €**

```
Success!

Job Stats (time in seconds):
JobId  Maps   Reduces MaxMapTime
edianReduceTime Alias   Feature Outputs
job_local1487450536_0007 1      1      n/a      n/a      n/a      n/a      n/a      n/a      n/a      n/a      clients,clients_groupes,depenses_clients,total_par_client
ROUP_BY,COMBINER
job_local284485603_0010 1      1      n/a      n/a      n/a      n/a      n/a      n/a      n/a      n/a      clients_tries
file:/tmp/temp-736790497/tmp1454037906,
job_local289191294_0008 1      1      n/a      n/a      n/a      n/a      n/a      n/a      n/a      n/a      clients_tries SAMPLER
job_local992797406_0009 1      1      n/a      n/a      n/a      n/a      n/a      n/a      n/a      n/a      clients_tries ORDER_BY,COMBINER

Input(s):
Successfully read 8 records from: "file:///client.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp-736790497/tmp1454037906"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1487450536_0007 -> job_local289191294_0008,
job_local289191294_0008 -> job_local992797406_0009,
job_local992797406_0009 -> job_local284485603_0010,
job_local284485603_0010

2025-08-30 11:19:28,768 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,769 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,769 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,770 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,770 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,772 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,772 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,772 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,773 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,774 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,774 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,775 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:19:28,775 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-08-30 11:19:28,776 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-08-30 11:19:28,776 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-08-30 11:19:28,788 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-08-30 11:19:28,788 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Simon, Philippe, 3200)
```

#### Question 5 : Produit le plus cher

produits\_groupes = GROUP clients BY produit;

prix\_par\_produit = FOREACH produits\_groupes GENERATE group as produit, MAX(clients.prix)  
as prix\_max;

produits\_tries = ORDER prix\_par\_produit BY prix\_max DESC;

produit\_plus\_cher = LIMIT produits\_tries 1;

DUMP produit\_plus\_cher;

✓ Résultat attendu : **Ordinateur – 1600 €**

```
Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime
edianReduceTime Alias Feature Outputs
job_local1329726892_0014 1 1 n/a n/a n/a n/a n/a n/a n/a produits_tries
File:/tmp/temp-736790497/tmp-45802337,
job_local1592512231_0011 1 1 n/a n/a n/a n/a n/a n/a n/a clients,prix_par_produit,produits_groupes
ROUP_BY,COMBINER
job_local1649793064_0012 1 1 n/a n/a n/a n/a n/a n/a n/a produits_tries
AMPLER
job_local170127598_0013 1 1 n/a n/a n/a n/a n/a n/a n/a produits_tries ORDER_BY,COMBINER

Input(s):
Successfully read 8 records from: "file:///client.txt"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp-736790497/tmp-45802337"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1592512231_0011 -> job_local1649793064_0012,
job_local1649793064_0012 -> job_local170127598_0013,
job_local170127598_0013 -> job_local1329726892_0014,
job_local1329726892_0014

2025-08-30 11:20:29,023 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,024 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,024 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,026 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,026 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,027 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,028 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,029 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,030 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,031 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,032 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,032 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2025-08-30 11:20:29,033 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - Success!
2025-08-30 11:20:29,034 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated, instead, use dfs.bytes-per-checksum
2025-08-30 11:20:29,034 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-08-30 11:20:29,046 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-08-30 11:20:29,046 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Ordinateur,1600)
```

---

## Étape 6 : Sortie de Pig

Pour quitter Pig :

quit;

---

## Étape 7 : Sortie du conteneur Docker

Pour sortir du conteneur :

Exit

**REALISER PAR :**  
**YASSIR EL GHRISSI**

