pig_lab

Bouayaben bilal

```
PS C:\Users\lenovo\BigdataLabs> docker exec hadoop-
master bash -c "echo 'export HADOOP_CONF_DIR=/usr/l
ocal/hadoop/etc/hadoop' >> ~/.bashrc; source ~/.bas
hrc; /usr/local/pig/bin/pig -x local /shared_volume
/wordcount.pig"
HadoopVersion   PigVersion      UserId  StartedAt F
inishedAt       Features
3.2.0   0.17.0  root    2025-11-15 14:55:45     202
5-11-15 14:55:47        GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTimeA
vgMapTime       MedianMapTime   MaxReduceTime   Min
ReduceTime      AvgReduceTime   MedianReducetime  A
lias    Feature Outputs
job_local1673952241_0001        1       1       n/n
/a      n/a     n/a     n/a     n/a     n/a     n/D
,E,clean_w,lines,words  GROUP_BY,COMBINER       /sh
ared_volume/pig_out/WORD_COUNT,

Input(s):
Successfully read 3664 records from: "/shared_volum
e/alice.txt"

Output(s):
Successfully stored 2621 records in: "/shared_volum
```

```
Output(s):
Successfully stored 2621 records in: "/shared_volum
e/pig_out/WORD_COUNT"

Counters:
Total records written : 2621
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1673952241_0001
```

Cette étape démontre l'utilisation d'Apache PIG pour effectuer un comptage de mots (wordcount) sur le fichier "Alice au pays des merveilles". Le script PIG lit le fichier texte, tokenise chaque ligne en mots individuels, filtre les mots valides, les groupe par occurrence et compte leur fréquence.

- 3664 enregistrements lus depuis `/shared_volume/alice.txt`
- 2621 mots uniques identifiés
- Résultats stockés dans `/shared_volume/pig_out/WORD_COUNT/`
- Temps d'exécution : 3 secondes
- Job MapReduce : 1 Map task, 1 Reduce task avec Combiner activé

```
# 1. Charger le fichier texte
lines = LOAD '/shared_volume/alice.txt';

# 2. Tokeniser chaque ligne en mots (FLATTEN pour obtenir un mot par ligne)
words = FOREACH lines GENERATE FLATTEN(TOKENIZE((chararray)$0)) AS word;

# 3. Filtrer uniquement les mots alphanumériques (supprimer ponctuation)
clean_w = FILTER words BY word MATCHES '\\w+';

# 4. Grouper les mots identiques
D = GROUP clean_w BY word;

# 5. Compter les occurrences de chaque mot
E = FOREACH D GENERATE group, COUNT(clean_w);

# 6. Stocker les résultats
STORE E INTO '/shared_volume/pig_out/WORD_COUNT/';
```

```
PS C:\Users\lenovo\BigdataLabs> docker exec hadoop-
master bash -c "cat /shared_volume/pig_out/WORD_COU
NT/part-r-00000 | head -20"
A       8
C       1
D       1
I       273
M       1
O       1
V       1
X       1
a       609
c       1
e       1
p       1
AT      1
Ah      1
An      4
As      14
At      8
BE      1
Be      2
By      3
```

```
PS C:\Users\lenovo\BigdataLabs> docker exec hadoop-
master bash -c "hdfs dfs -ls /shared_volume/pigout/
avg_salary_dept/; hdfs dfs -cat /shared_volume/pigo
ut/avg_salary_dept/*"
Found 2 items
-rw-r--r--   2 root supergroup          0 2025-11-1
5 15:11 /shared_volume/pigout/avg_salary_dept/_SUCC
ESS
-rw-r--r--   2 root supergroup         48 2025-11-1
5 15:11 /shared_volume/pigout/avg_salary_dept/part-
r-00000
101     53200.0
102     57400.0
103     63000.0
104     62600.0
```

```
=== 2. Nombre employés par département ===
101     5
102     5
103     5
104     5

=== 3. Employés avec leurs départements ===
Dupont Jean     101
Martin Sophie   102
Leblanc Pierre  103
Durand Alice    104
Lemoine Lucie   101
```

```
=== 4. Employés salaire > 60000 ===
4       Durand Alice    Female  70000   104     Toulouse    Franc
e
7       Roux Camille    Female  62000   103     Lille       Belgi
que
10      Dubois Sara     Female  72000   102     Lille       Belgi
que
12      Giraud Nicolas  Male    61000   104     Bordeaux    Suiss
e
13      Pichon Aline    Female  66000   101     Nice        Suiss
e
15      Boucher Céline  Female  70000   103     Lyon        Suiss
e
16      Marchand Olivier        Male    68000   10Nantes    Franc
e
19      Caron Sophie    Female  75000   103     Bordeaux    Franc
e
```

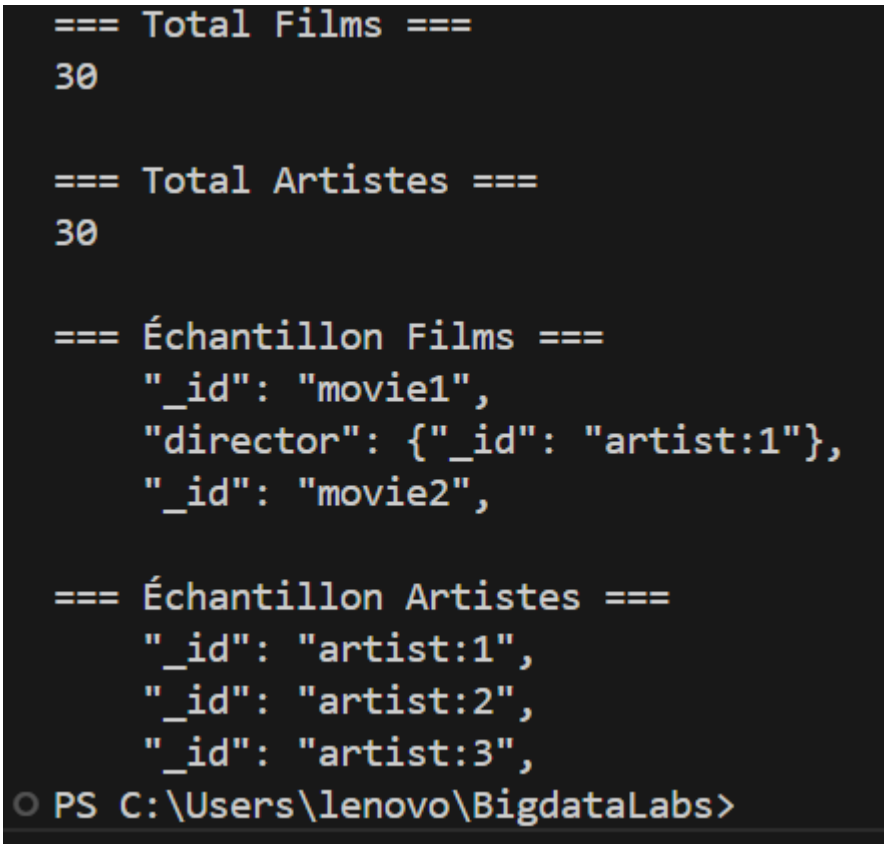
```
=== 5. Salaire max par département ===
101     66000
102     72000
103     75000
104     70000

=== 7. Total employés ===
20

=== 8. Employés de Paris ===
1

=== 9. Salaire total par ville ===
Lyon    130000
Nice    174000
Lille   187000
Paris   50000
Nantes  126000
Rennes  103000
Bordeaux        184000
Toulouse        172000
Marseille       55000
```

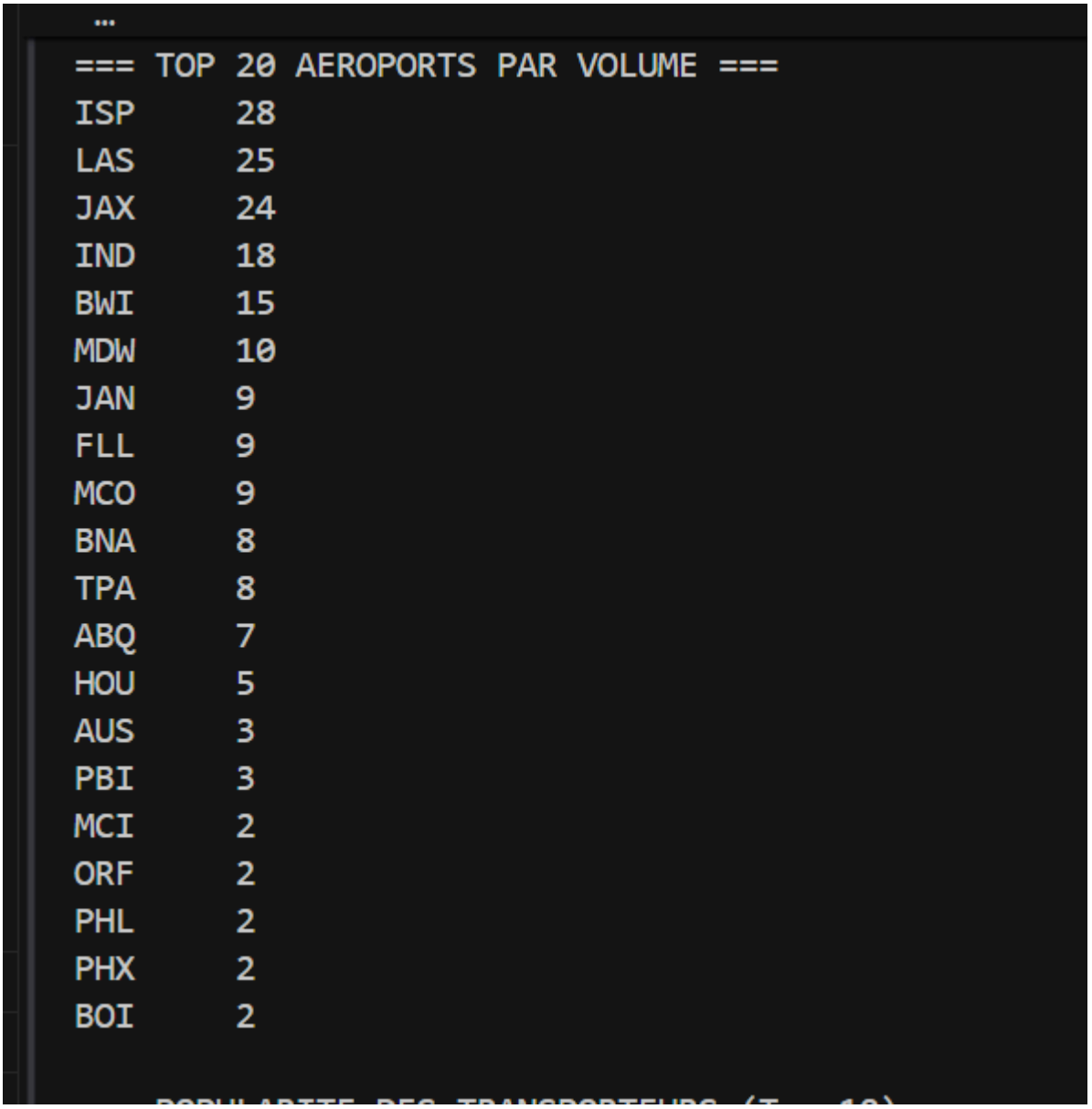
```
=== 10. Départements avec femmes ===
101     3
102     2
103     4
104     2
```

Conclusion:

L'analyse des 20 employés répartis dans 4 départements (101, 102, 103, 104) révèle que chaque département compte exactement 5 employés avec des salaires moyens respectifs de 53 200€, 57 400€, 63 000€ et 62 600€. Le département 103 présente le salaire maximum le plus élevé à 75 000€ (Caron Sophie), suivi du département 102 avec 72 000€ (Dubois Sara), du département 104 avec 70 000€ (Durand Alice) et du département 101 avec 66 000€ (Pichon Aline). Au total, 8 employés perçoivent un salaire supérieur à 60 000€, et 11 femmes sont réparties dans les 4 départements (3 au dept 101, 2 au dept 102, 4 au dept 103, et 2 au dept 104). L'analyse géographique montre que la ville de Lille concentre la masse salariale la plus importante avec 187 000€, suivie de Bordeaux (184 000€), Nice (174 000€) et Toulouse (172 000€), tandis que Paris ne compte qu'un seul employé avec

un salaire de 50 000€. Le script PIG a traité ces données en 2 jobs MapReduce en utilisant les fonctionnalités GROUP BY, FILTER et COMBINER pour optimiser les calculs, démontrant l'efficacité d'Apache PIG pour l'analyse de données structurées dans un environnement Hadoop distribué




...
=== TOP 20 AEROPORTS PAR VOLUME ===
ISP 28
LAS 25
JAX 24
IND 18
BWI 15
MDW 10
JAN 9
FLL 9
MCO 9
BNA 8
TPA 8
ABQ 7
HOU 5
AUS 3
PBI 3
MCI 2
ORF 2
PHL 2
PHX 2
BOI 2

```
=== POPULARITE DES TRANSPORTEURS (Top 10) ===
WN              100     757.35

=== RETARDS PAR HEURE ===
6       9       1       11.11111111111111
7       11      0       0.0
8       8       0       0.0
9       7       3       42.857142857142854
10      6       1       16.666666666666664
11      4       1       25.0
12      7       5       71.42857142857143
13      4       1       25.0
14      8       3       37.5
15      6       2       33.33333333333333
16      7       5       71.42857142857143
17      7       4       57.14285714285714
18      5       2       40.0
19      7       3       42.857142857142854
20      2       1       50.0
21      2       1       50.0
```

```
=== RETARDS PAR JOUR DE SEMAINE ===
4       100     33

=== RETARDS PAR MOIS ===
1       100     33      33.0
PS C:\Users\lenovo\BigdataLabs> docker exec hadoop-master bash -
c "echo '=== RETARDS PAR TRANSPORTEUR ==='; hdfs dfs -cat /share
d_volume/pigout/flights/carrier_delays/part-*; echo -e '\n=== TO
P 20 ROUTES LES PLUS FREQUENTES ==='; hdfs dfs -cat /shared_volu
me/pigout/flights/top_routes/part-* | head -20"
=== RETARDS PAR TRANSPORTEUR ===
WN      100     33      33.0
```

```
=== TOP 20 ROUTES LES PLUS FREQUENTES ===
ISP     BWI     7
LAS     ABQ     7
ISP     MCO     6
JAX     FLL     6
IND     MDW     4
LAS     BNA     4
ISP     MDW     4
JAN     HOU     4
JAX     BNA     4
ISP     PBI     3
JAX     TPA     3
ISP     TPA     3
JAX     BWI     3
IND     BWI     3
LAS     AUS     3
ISP     FLL     3
JAN     BWI     2
IND     MCO     2
IND     PHX     2
JAN     MDW     2
PS C:\Users\lenovo\BigdataLabs>
```