Lab 5: Apache PIG

- installation pig

```
C:\Users\hp\OneDrive\Documents\hadoop_java\hadoop_docker>docker exec -it hadoop-master bash
root@hadoop-master:~# wget https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
--2025-11-15 13:55:24--  https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 230606579 (220M) [application/x-gzip]
Saving to: 'pig-0.17.0.tar.gz'

pig-0.17.0.tar.gz                    42%[===============================>                                          ]  93.25M   142KB/s    in 11m 50s

2025-11-15 14:07:19 (134 KB/s) - Connection closed at byte 97779712. Retrying.

--2025-11-15 14:07:20--  (try: 2)  https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 206 Partial Content
Length: 230606579 (220M), 132826867 (127M) remaining [application/x-gzip]
Saving to: 'pig-0.17.0.tar.gz'

pig-0.17.0.tar.gz                   100%[+++++++++++++++++++++++++++++++=========================================>] 219.92M   572KB/s    in 15m 30s

2025-11-15 14:23:27 (140 KB/s) - 'pig-0.17.0.tar.gz' saved [230606579/230606579]
```

- création du fichier alice

```
root@hadoop-master:~# cat > /shared_volume/alice.txt << 'EOF'
> Alice was beginning to get very tired of sitting by her sister on the bank
> and of having nothing to do once or twice she had peeped into the book
> her sister was reading but it had no pictures or conversations in it
> and what is the use of a book thought Alice without pictures or conversati
ons
> So she was considering in her own mind as well as she could
> for the hot day made her feel very sleepy and stupid
> whether the pleasure of making a daisy chain would be worth the trouble
> of getting up and picking the daisies when suddenly a White Rabbit
> with pink eyes ran close by her
> EOF
root@hadoop-master:~# pig -version
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/
slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.25.
jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanat
ion.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
root@hadoop-master:~# pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/
slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.25.
jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanat
ion.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2025-11-15 14:51:38,452 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-11-15 14:51:38,452 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecT
```

- Ouvrir pig en local

```
root@hadoop-master:~# pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7
.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2025-11-15 14:51:38,452 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-11-15 14:51:38,452 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2025-11-15 14:51:38,530 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r179738
6) compiled Jun 02 2017, 15:41:58
2025-11-15 14:51:38,531 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/p
ig_1763218298528.log
2025-11-15 14:51:38,574 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /ro
ot/.pigbootup not found
2025-11-15 14:51:38,762 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapre
d.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-11-15 14:51:38,765 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecution
Engine - Connecting to hadoop file system at: file:///
2025-11-15 14:51:38,915 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.by
tes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-15 14:51:38,939 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session
: PIG-default-46367ffa-da93-44ed-bb93-07e905633fd1
2025-11-15 14:51:38,939 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.ti
meline-service.enabled set to false
```

- Chargement de fichier :

```
grunt> lines = LOAD '/shared_volume/alice.txt';
2025-11-15 14:53:04,122 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.by
tes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt>
```

- Traitement du fichier :

```
grunt> lines = LOAD '/shared_volume/alice.txt';
2025-11-15 14:53:04,122 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-chec
ksum
grunt> words = FOREACH lines GENERATE FLATTEN(TOKENIZE((chararray)$0)) AS word;
2025-11-15 14:54:21,268 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUs
ageThreshold = 489580128, usageThreshold = 489580128
grunt> clean_w = FILTER words BY word MATCHES '\\w+';
grunt> D = GROUP clean_w BY word;
grunt> E = FOREACH D GENERATE group AS word, COUNT(clean_w) AS count;
grunt> DUMP E;
```

- Affichage du résultat de wordcount :

```
2025-11-15 14:55:05,429 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-11-15 14:55:05,440 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-11-15 14:55:05,441 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process :
(a,3)
(So,1)
(as,2)
(be,1)
(by,2)
(do,1)
(in,2)
(is,1)
(it,2)
(no,1)
(of,5)
(on,1)
(or,3)
(to,2)
(up,1)
(and,4)
(but,1)
(day,1)
(for,1)
(get,1)
(had,2)
(her,5)
(hot,1)
(own,1)
(ran,1)
(she,3)
(the,7)
(use,1)
(was,3)
(bank,1)
(book,2)
(eyes,1)
(feel,1)
(into,1)
(made,1)
(mind,1)
(once,1)
(pink,1)
```

- Le storage du résultat :

```
grunt> STORE E INTO '/shared_volume/pig_out/WORD_COUNT/';
2025-11-15 15:01:03,668 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.by
tes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-15 15:01:03,847 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapre
d.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.se
parator
2025-11-15 15:01:03,904 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
 used in the script: GROUP_BY,FILTER
2025-11-15 15:01:03,945 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.by
tes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-15 15:01:03,946 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schema
tuple] was not set... will not generate code.
2025-11-15 15:01:03,946 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOpti
mizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParall
elSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimiz
er, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilte
r, SplitFilter, StreamTypeCastInserter]}
2025-11-15 15:01:03,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceL
ayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2025-11-15 15:01:03,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.Combi
nerOptimizerUtil - Choosing to move algebraic foreach to combiner
2025-11-15 15:01:03,965 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceL
ayer.MultiQueryOptimizer - MR plan size before optimization: 1
2025-11-15 15:01:03,965 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceL
ayer.MultiQueryOptimizer - MR plan size after optimization: 1
2025-11-15 15:01:03,989 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.by
tes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-15 15:01:03,992 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobT
```

- Le storage réussi

```
 Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.0   0.17.0  root    2025-11-15 15:01:04     2025-11-15 15:01:06     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxRe
duceTime        MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outpu
ts
job_local1828263358_0002        1       1       n/a     n/a     n/a     n/a     n/a     n/a n
/a      n/a     D,E,clean_w,lines,words GROUP_BY,COMBINER       /shared_volume/pig_out/WORD_C
OUNT,

Input(s):
Successfully read 9 records from: "/shared_volume/alice.txt"

Output(s):
Successfully stored 76 records in: "/shared_volume/pig_out/WORD_COUNT"

Counters:
Total records written : 76
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1828263358_0002


2025-11-15 15:01:06,593 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobT
racker metrics system already initialized!
2025-11-15 15:01:06,600 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobT
racker metrics system already initialized!
2025-11-15 15:01:06,605 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobT
racker metrics system already initialized!
2025-11-15 15:01:06,619 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceL
ayer.MapReduceLauncher - Success!
```

- voir les résultats :

```
root@hadoop-master:~# ls -la /shared_volume/pig_out/WORD_COUNT/
total 4
drwxr-xr-x 1 root root 4096 Nov 15 15:01 .
drwxr-xr-x 1 root root 4096 Nov 15 15:01 ..
-rw-r--r-- 1 root root    8 Nov 15 15:01 ._SUCCESS.crc
-rw-r--r-- 1 root root   16 Nov 15 15:01 .part-r-00000.crc
-rw-r--r-- 1 root root    0 Nov 15 15:01 _SUCCESS
-rw-r--r-- 1 root root  574 Nov 15 15:01 part-r-00000
```

- Affichage du contenu :

```
root@hadoop-master:~# cat /shared_volume/pig_out/WORD_COUNT/part-*
a       3
So      1
as      2
be      1
by      2
do      1
in      2
is      1
it      2
no      1
of      5
on      1
or      3
to      2
up      1
and     4
but     1
day     1
for     1
get     1
had     2
her     5
hot     1
own     1
ran     1
she     3
the     7
use     1
was     3
bank    1
book    2
eyes    1
feel    1
into    1
made    1
mind    1
once    1
pink    1
very    2
```

- Fichier employee :

```
root@hadoop-master:~# cat > /shared_volume/employees.txt << 'EOF'
> 1,Dupont,Jean,10,Paris,55000
> 2,Martin,Marie,20,Lyon,62000
> 3,Bernard,Sophie,10,Paris,48000
> 4,Dubois,Pierre,30,Marseille,71000
> 5,Thomas,Julie,20,Lyon,58000
> 6,Robert,Michel,10,Paris,52000
> 7,Petit,Claire,30,Marseille,65000
> 8,Durand,Paul,20,Lyon,59000
> 9,Leroy,Anne,10,Paris,53000
> 10,Moreau,Luc,30,Marseille,68000
> 11,Simon,Camille,20,Lyon,61000
> 12,Laurent,Emma,10,Paris,56000
> 13,Lefebvre,Alice,30,Marseille,72000
> 14,Michel,Lucie,20,Lyon,60000
> 15,Garcia,Marc,10,Paris,54000
> EOF
```

- Fichier département :

```
root@hadoop-master:~# cat > /shared_volume/departements.txt << 'EOF'
> 10,Informatique
> 20,Marketing
> 30,Finance
> 40,RH
> EOF
```

- On copie les fichiers dans hdfs :

```
root@hadoop-master:~# hdfs dfs -mkdir -p /user/root/input
root@hadoop-master:~# hdfs dfs -put /shared_volume/employees.txt /user/root/input/
root@hadoop-master:~# hdfs dfs -put /shared_volume/departements.txt /user/root/input/
root@hadoop-master:~# hdfs dfs -ls /user/root/input/
Found 2 items
-rw-r--r--   2 root supergroup         46 2025-11-15 15:06 /user/root/input/departements.txt
-rw-r--r--   2 root supergroup        467 2025-11-15 15:06 /user/root/input/employees.txt
root@hadoop-master:~# hdfs dfs -cat /user/root/input/employees.txt | head -5
1,Dupont,Jean,10,Paris,55000
2,Martin,Marie,20,Lyon,62000
3,Bernard,Sophie,10,Paris,48000
4,Dubois,Pierre,30,Marseille,71000
5,Thomas,Julie,20,Lyon,58000
root@hadoop-master:~#
```

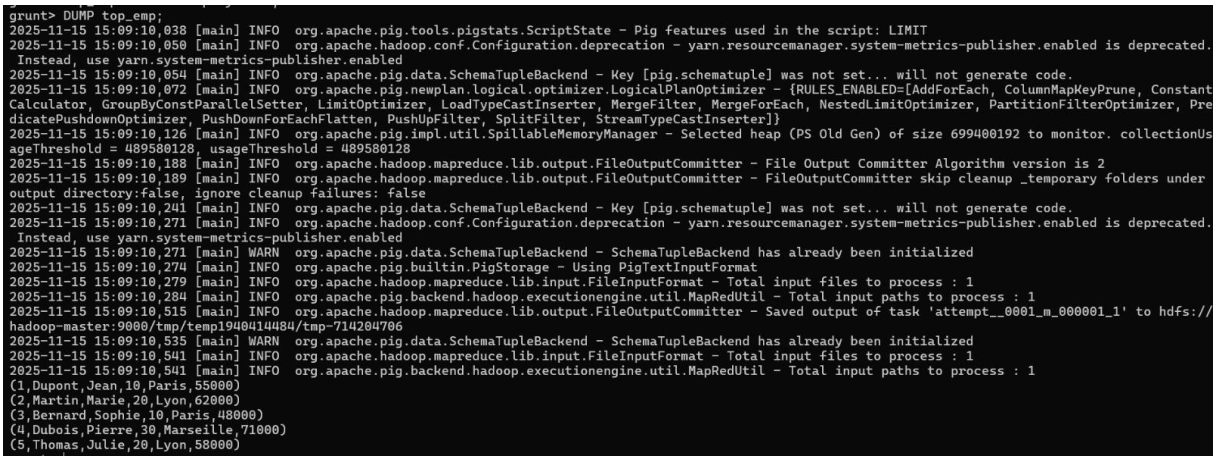
- Le chargement dans pig :

```
grunt> employees = LOAD '/user/root/input/employees.txt' USING PigStorage(',')
>>     AS (id:int, nom:chararray, prenom:chararray, depno:int, region:chararray, salaire:int)
;
2025-11-15 15:07:35,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.
resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metr
ics-publisher.enabled
grunt> departements = LOAD '/user/root/input/departements.txt' USING PigStorage(',')
>>     AS (depno:int, nom_dep:chararray);
2025-11-15 15:07:45,969 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.
resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metr
ics-publisher.enabled
```

- On verifier le chargement :

```
grunt> DESCRIBE employees;
employees: {id: int,nom: chararray,prenom: chararray,depno: int,region: chararray,salaire: in
t}
grunt> DESCRIBE departements;
departements: {depno: int,nom_dep: chararray}
```

- affichage de quelque ligne :



- J’ai utilisé un script pour tous le traitement car si on sort de pig on doit refaire voici le script :

```
cat > /shared_volume/analyse_employees.pig << 'EOF'
-- Charger les données
employees = LOAD '/user/root/input/employees.txt' USING
PigStorage(',')
   AS (id:int, nom:chararray, prenom:chararray, depno:int,
region:chararray, salaire:int);

departements = LOAD '/user/root/input/departements.txt' USING
PigStorage(',')
   AS (depno:int, nom_dep:chararray);

-- Question 1: Salaire moyen par département
grp_dep = GROUP employees BY depno;
salaire_moyen = FOREACH grp_dep GENERATE
   group AS depno,
   AVG(employees.salaire) AS salaire_moyen;
STORE salaire_moyen INTO '/user/root/pigout/salaire_moyen';

-- Question 2: Nombre d'employés par département
nb_emp_dep = FOREACH grp_dep GENERATE
   group AS depno,
   COUNT(employees) AS nb_employes;
STORE nb_emp_dep INTO '/user/root/pigout/nb_employes_dep';
```

```
-- Question 3: Employés avec leurs départements
emp_dep = JOIN employees BY depno, departements BY depno;
emp_dep_list = FOREACH emp_dep GENERATE
   employees::id AS id,
   employees::nom AS nom,
   employees::prenom AS prenom,
   departements::nom_dep AS departement,
   employees::region AS region,
   employees::salaire AS salaire;
STORE emp_dep_list INTO '/user/root/pigout/emp_avec_dep';

-- Question 4: Employés avec salaire > 60000
emp_hauts_salaires = FILTER employees BY salaire > 60000;
STORE emp_hauts_salaires INTO '/user/root/pigout/hauts_salaires';

-- Question 5: Département avec le salaire le plus élevé
salaire_max_dep = FOREACH grp_dep GENERATE
   group AS depno,
   MAX(employees.salaire) AS salaire_max;
salaire_max_dep_order = ORDER salaire_max_dep BY salaire_max
DESC;
top_dept = LIMIT salaire_max_dep_order 1;
STORE top_dept INTO '/user/root/pigout/top_departement';

-- Question 6: Départements sans employés
all_deps = FOREACH departements GENERATE depno;
deps_avec_emp = FOREACH grp_dep GENERATE group AS depno;
dep_compare = COGROUP all_deps BY depno, deps_avec_emp BY
depno;
dep_sans_emp = FILTER dep_compare BY IsEmpty(deps_avec_emp);
resultat_dep_vides = FOREACH dep_sans_emp GENERATE group AS
depno;
STORE resultat_dep_vides INTO
'/user/root/pigout/dep_sans_employes';
```

```
-- Question 7: Nombre total d'employés
grp_all = GROUP employees ALL;
total_emp = FOREACH grp_all GENERATE COUNT(employees) AS
total;
STORE total_emp INTO '/user/root/pigout/total_employes';

-- Question 8: Employés de Paris
emp_paris = FILTER employees BY region == 'Paris';
STORE emp_paris INTO '/user/root/pigout/emp_paris';

-- Question 9: Salaire total par ville
grp_ville = GROUP employees BY region;
salaire_ville = FOREACH grp_ville GENERATE
   group AS ville,
   SUM(employees.salaire) AS salaire_total;
STORE salaire_ville INTO '/user/root/pigout/salaire_par_ville';

-- Question 10: Départements avec des femmes employées
emp_femmes = FILTER employees BY
   (prenom == 'Marie' OR prenom == 'Sophie' OR prenom == 'Julie' OR
    prenom == 'Claire' OR prenom == 'Anne' OR prenom == 'Camille'
OR
    prenom == 'Emma' OR prenom == 'Lucie' OR prenom == 'Alice');
emp_femmes_dep = JOIN emp_femmes BY depno, departements BY
depno;
grp_femmes = GROUP emp_femmes_dep BY
departements::nom_dep;
deps_avec_femmes = FOREACH grp_femmes GENERATE
   group AS departement,
   COUNT(emp_femmes_dep) AS nb_femmes;
STORE deps_avec_femmes INTO
'/user/root/pigout/employes_femmes';
EOF
```

- L’exécution du script :

```
root@hadoop-master:~# pig /shared_volume/analyse_employees.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/
slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.25.
jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanat
ion.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2025-11-15 15:14:33,470 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-11-15 15:14:33,472 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDU
CE
2025-11-15 15:14:33,472 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the E
2025-11-15 15:17:51,591 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,604 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,604 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,609 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,632 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,632 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,635 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,648 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,648 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,651 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,662 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,662 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,665 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,684 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,685 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,689 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,706 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,706 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,708 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,723 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,723 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,725 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,744 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,744 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,747 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,762 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:17:51,763 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:17:51,767 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:17:51,783 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-11-15 15:17:51,801 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 18 seconds and 360 milliseconds (198360 ms)
```

- Tous les dossiers de résultats :

```
root@hadoop-master:~# hdfs dfs -ls /user/root/pigout/
Found 10 items
drwxr-xr-x   - root supergroup          0 2025-11-15 15:17 /user/root/pigout
/dep_sans_employes
drwxr-xr-x   - root supergroup          0 2025-11-15 15:15 /user/root/pigout
/emp_avec_dep
drwxr-xr-x   - root supergroup          0 2025-11-15 15:14 /user/root/pigout
/emp_paris
drwxr-xr-x   - root supergroup          0 2025-11-15 15:16 /user/root/pigout
/employes_femmes
drwxr-xr-x   - root supergroup          0 2025-11-15 15:14 /user/root/pigout
/hauts_salaires
drwxr-xr-x   - root supergroup          0 2025-11-15 15:15 /user/root/pigout
/nb_employes_dep
drwxr-xr-x   - root supergroup          0 2025-11-15 15:15 /user/root/pigout
/salaire_moyen
drwxr-xr-x   - root supergroup          0 2025-11-15 15:14 /user/root/pigout
/salaire_par_ville
drwxr-xr-x   - root supergroup          0 2025-11-15 15:17 /user/root/pigout
/top_departement
drwxr-xr-x   - root supergroup          0 2025-11-15 15:14 /user/root/pigout
/total_employes
```

- Les résultats de chaque requêtes :

```
=== Nombre d'employés par département ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/nb_employes_dep/part-*
10      6
20      5
30      4
```

```
=== Salaire moyen par département ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/salaire_moyen/part-*
10      53000.0
20      60000.0
30      69000.0
```

```
=== Employés avec leurs départements ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/emp_avec_dep/part-* |
head -10
1       Dupont  Jean    Informatique    Paris   55000
12      Laurent Emma    Informatique    Paris   56000
9       Leroy   Anne    Informatique    Paris   53000
15      Garcia  Marc    Informatique    Paris   54000
6       Robert  Michel  Informatique    Paris   52000
3       Bernard Sophie  Informatique    Paris   48000
8       Durand  Paul    Marketing       Lyon    59000
2       Martin  Marie   Marketing       Lyon    62000
14      Michel  Lucie   Marketing       Lyon    60000
5       Thomas  Julie   Marketing       Lyon    58000
```

```
=== Employés avec salaire > 60000 ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/hauts_salaires/part-*
2       Martin  Marie   20      Lyon    62000
4       Dubois  Pierre  30      Marseille       71000
7       Petit   Claire  30      Marseille       65000
10      Moreau  Luc     30      Marseille       68000
11      Simon   Camille 20      Lyon    61000
13      Lefebvre        Alice   30      Marseille       72000
```

```
=== Département avec salaire max ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/top_departement/part-*
30      72000
```

```
=== Départements sans employés ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/dep_sans_employes/part
-*
40
```

```
=== Départements avec femmes ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/employes_femmes/part-*
Finance 2
Marketing       4
Informatique    3
```

```
=== Total employés ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/total_employes/part-*
15
```

```
=== Salaire total par ville ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/salaire_par_ville/part
-*
Lyon    300000
Paris   318000
Marseille       276000
```

```
=== Employés de Paris ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/emp_paris/part-*
1       Dupont  Jean    10      Paris   55000
3       Bernard Sophie  10      Paris   48000
6       Robert  Michel  10      Paris   52000
9       Leroy   Anne    10      Paris   53000
12      Laurent Emma    10      Paris   56000
15      Garcia  Marc    10      Paris   54000
```

- Vérifier que les films bien créés :

- J’ai trouvé des problème avec le json donc j’ai fait un script pour la conversion en csv

```
root@hadoop-master:~# python3 /shared_volume/json_to_csv.py
Conversion terminée!
root@hadoop-master:~# cat /shared_volume/films.csv
movie:1,Vertigo,1958,drama,US,artist:3
movie:2,Psycho,1960,thriller,US,artist:3
movie:3,Rear Window,1954,thriller,US,artist:3
movie:10,Blade Runner,1982,sci-fi,US,artist:20
movie:12,Alien,1979,sci-fi,US,artist:20
movie:34,Le retour du Jedi,1983,sci-fi,US,artist:25
movie:50,Amelie,2001,romance,FR,artist:30
movie:51,La Haine,1995,drama,FR,artist:32
root@hadoop-master:~# cat /shared_volume/films_actors.csv
movie:1,artist:15,John Ferguson
movie:1,artist:16,Madeleine Elster
movie:2,artist:17,Norman Bates
movie:2,artist:18,Marion Crane
movie:3,artist:15,L.B. Jefferies
movie:3,artist:19,Lisa Fremont
movie:10,artist:24,Deckard
movie:10,artist:21,Rachael
movie:12,artist:22,Ripley
movie:34,artist:24,Han Solo
movie:34,artist:23,Luke Skywalker
movie:50,artist:31,Amelie Poulain
movie:51,artist:33,Vinz
root@hadoop-master:~# cat /shared_volume/artists.csv
artist:3,Hitchcock,Alfred,1899
artist:15,Stewart,James,1908
artist:16,Novak,Kim,1933
artist:17,Perkins,Anthony,1932
artist:18,Leigh,Janet,1927
artist:19,Kelly,Grace,1929
artist:20,Scott,Ridley,1937
artist:21,Young,Sean,1959
artist:22,Weaver,Sigourney,1949
artist:23,Hamill,Mark,1951
artist:24,Ford,Harrison,1942
artist:25,Marquand,Richard,1937
artist:30,Jeunet,Jean-Pierre,1953
artist:31,Tautou,Audrey,1976
artist:32,Kassovitz,Mathieu,1967
```

- Le script pour le traitement :
  cat > /shared_volume/analyse_films_csv.pig << 'EOF'
  -- Charger les CSV
  films = LOAD '/user/root/input/films.csv' USING PigStorage(',')
  AS (filmid:chararray, title:chararray, year:int, genre:chararray, country:chararray, directorid:chararray);

  artists = LOAD '/user/root/input/artists.csv' USING PigStorage(',')
  AS (artistid:chararray, lastname:chararray, firstname:chararray, birthdate:chararray);

```
filmsactors = LOAD '/user/root/input/films_actors.csv' USING
PigStorage(',')
   AS (filmid:chararray, actorid:chararray, role:chararray);

-- Filtrer les films américains
filmsUSA = FILTER films BY country == 'US';

-- Question 1: Films américains par année
byYear = GROUP filmsUSA BY year;
resultYear = FOREACH byYear GENERATE
   group AS annee,
   COUNT(filmsUSA) AS nbfilms,
   filmsUSA.title AS titres;
STORE resultYear INTO '/user/root/pigout/films_par_annee';

-- Question 2: Films américains par réalisateur
byDirector = GROUP filmsUSA BY directorid;
resultDirector = FOREACH byDirector GENERATE
   group AS directeurid,
   COUNT(filmsUSA) AS nbfilms,
   filmsUSA.title AS titres;
STORE resultDirector INTO '/user/root/pigout/films_par_realisateur';

-- Question 3: Triplets (film, acteur, role) pour films US
usaactors = JOIN filmsUSA BY filmid, filmsactors BY filmid;
triplets = FOREACH usaactors GENERATE
   filmsUSA::filmid AS filmid,
   filmsactors::actorid AS actorid,
   filmsactors::role AS role;
STORE triplets INTO '/user/root/pigout/films_acteurs_roles';

-- Question 4: Films-acteurs avec infos complètes
moviesactors = JOIN triplets BY actorid, artists BY artistid;
fullactors = FOREACH moviesactors GENERATE
   triplets::filmid AS filmid,
```

```
    artists::artistid AS actorid,
    artists::firstname AS prenom,
    artists::lastname AS nom,
    artists::birthdate AS datenaissance,
    triplets::role AS role;
STORE fullactors INTO '/user/root/pigout/films_acteurs_complet';

-- Question 5: Films complets avec tous les acteurs
fullmovies = COGROUP filmsUSA BY filmid, fullactors BY filmid;
completefilms = FOREACH fullmovies GENERATE
    group AS filmid,
    FLATTEN(filmsUSA.title) AS titre,
    FLATTEN(filmsUSA.year) AS annee,
    FLATTEN(filmsUSA.genre) AS genre,
    fullactors AS acteurs;
STORE completefilms INTO '/user/root/pigout/films_complets';

-- Question 6: Acteurs-Réalisateurs
filmsdirected = FOREACH filmsUSA GENERATE
    directorid AS artistid,
    filmid,
    title;

filmsplayed = JOIN filmsactors BY filmid, filmsUSA BY filmid;
filmsplayedfinal = FOREACH filmsplayed GENERATE
    filmsactors::actorid AS artistid,
    filmsUSA::filmid AS filmid,
    filmsUSA::title AS titre,
    filmsactors::role AS role;

acteursreals = COGROUP filmsdirected BY artistid, filmsplayedfinal
BY artistid;
resultat = FOREACH acteursreals GENERATE
    group AS artistid,
    filmsdirected AS filmsdiriges,
```

filmsplayedfinal AS filmsjoues;
STORE resultat INTO '/user/root/pigout/ActeursRealisateurs';
EOF

```
root@hadoop-master:~# pig /shared_volume/analyse_films_csv.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2025-11-15 15:45:20,035 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-11-15 15:45:20,037 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-11-15 15:45:20,037 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-11-15 15:45:20,068 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-11-15 15:45:20,068 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1763221520067.log
2025-11-15 15:45:20,262 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2025-11-15 15:45:20,321 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.
address
2025-11-15 15:45:20,321 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://hadoop-mas
ter:9000/
2025-11-15 15:45:20,593 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-analyse_films_csv.pig-599860a9-53f9-48de-9a59-89730434666b
6
2025-11-15 15:47:42,446 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,446 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,448 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,474 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,474 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,477 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,491 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,491 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,494 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,509 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,509 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,512 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,537 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,538 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,540 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,554 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,554 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,557 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,572 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,573 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,577 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,613 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,613 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,617 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,633 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 15:47:42,634 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 15:47:42,636 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 15:47:42,648 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-11-15 15:47:42,664 [main] INFO  org.apache.pig.Main - Pig script completed in 2 minutes, 22 seconds and 653 milliseconds (142653 ms)
```

- Affichage des résultats :

```
=== Films par année ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/films_par_annee/part-*
 2>/dev/null
1954    1       {(Rear Window)}
1958    1       {(Vertigo)}
1960    1       {(Psycho)}
1979    1       {(Alien)}
1982    1       {(Blade Runner)}
1983    1       {(Le retour du Jedi)}
```

```
=== Films par réalisateur ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/films_par_realisateur/
part-* 2>/dev/null
artist:3        3       {(Rear Window),(Psycho),(Vertigo)}
artist:20       2       {(Alien),(Blade Runner)}
artist:25       1       {(Le retour du Jedi)}
```

```
=== Triplets film-acteur-role ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/films_acteurs_roles/pa
rt-* 2>/dev/null | head -10
movie:1 artist:15       John Ferguson
movie:1 artist:16       Madeleine Elster
movie:2 artist:17       Norman Bates
movie:2 artist:18       Marion Crane
movie:3 artist:19       Lisa Fremont
movie:3 artist:15       L.B. Jefferies
movie:10        artist:24       Deckard
movie:10        artist:21       Rachael
movie:12        artist:22       Ripley
movie:34        artist:24       Han Solo
```

```
=== Acteurs-Réalisateurs ===
root@hadoop-master:~# hdfs dfs -cat /user/root/pigout/ActeursRealisateurs/pa
rt-* 2>/dev/null
artist:3        {(artist:3,movie:3,Rear Window),(artist:3,movie:2,Psycho),(a
rtist:3,movie:1,Vertigo)}       {}
artist:15       {}      {(artist:15,movie:3,Rear Window,L.B. Jefferies),(art
ist:15,movie:1,Vertigo,John Ferguson)}
artist:16       {}      {(artist:16,movie:1,Vertigo,Madeleine Elster)}
artist:17       {}      {(artist:17,movie:2,Psycho,Norman Bates)}
artist:18       {}      {(artist:18,movie:2,Psycho,Marion Crane)}
artist:19       {}      {(artist:19,movie:3,Rear Window,Lisa Fremont)}
artist:20       {(artist:20,movie:12,Alien),(artist:20,movie:10,Blade Runner
)}      {}
artist:21       {}      {(artist:21,movie:10,Blade Runner,Rachael)}
artist:22       {}      {(artist:22,movie:12,Alien,Ripley)}
artist:23       {}      {(artist:23,movie:34,Le retour du Jedi,Luke Skywalke
r)}
artist:24       {}      {(artist:24,movie:34,Le retour du Jedi,Han Solo),(ar
tist:24,movie:10,Blade Runner,Deckard)}
artist:25       {(artist:25,movie:34,Le retour du Jedi)}        {}
```

- Pour les données de vols j'ai essayé la dataset du lab mais c'est très volumineux donc j'ai utilisé un échantillon :

```
root@hadoop-master:~# cd /shared_volume
root@hadoop-master:/shared_volume# wget http://stat-computing.org/dataexpo/2
009/2008.csv.bz2
--2025-11-15 15:52:58--  http://stat-computing.org/dataexpo/2009/2008.csv.bz
2
Resolving stat-computing.org (stat-computing.org)... 172.67.209.204, 104.21.
23.78
Connecting to stat-computing.org (stat-computing.org)|172.67.209.204|:80...
connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: https://gamingcy.com [following]
--2025-11-15 15:52:59--  https://gamingcy.com/
Resolving gamingcy.com (gamingcy.com)... 5.161.231.17
Connecting to gamingcy.com (gamingcy.com)|5.161.231.17|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 197517 (193K) [text/html]
Saving to: '2008.csv.bz2'

2008.csv.bz2       100%[================>] 192.89K   239KB/s    in 0.8s

2025-11-15 15:53:01 (239 KB/s) - '2008.csv.bz2' saved [197517/197517]
```

- Le script de traitement :

```
cat > /shared_volume/analyse_flights.pig << 'EOF'
-- Charger les données de vols
flights = LOAD '/user/root/input/flights/*.csv' USING PigStorage(',') AS (
    Year:int,
    Month:int,
    DayofMonth:int,
    DayOfWeek:int,
    DepTime:chararray,
    CRSDepTime:chararray,
    ArrTime:chararray,
    CRSArrTime:chararray,
    UniqueCarrier:chararray,
    FlightNum:chararray,
    TailNum:chararray,
    ActualElapsedTime:chararray,
    CRSElapsedTime:chararray,
    AirTime:chararray,
    ArrDelay:int,
    DepDelay:int,
    Origin:chararray,
    Dest:chararray,
    Distance:int,
    TaxiIn:chararray,
    TaxiOut:chararray,
    Cancelled:int,
    CancellationCode:chararray,
    Diverted:int,
    CarrierDelay:chararray,
    WeatherDelay:chararray,
    NASDelay:chararray,
    SecurityDelay:chararray,
    LateAircraftDelay:chararray
);

-- Filtrer l'en-tête
```

```
flightsclean = FILTER flights BY Year IS NOT NULL AND Year > 1900;

-- ========================================
-- Question 1: TOP 20 AÉROPORTS PAR VOLUME
-- ========================================

-- Vols sortants
volssortants = GROUP flightsclean BY Origin;
nbsortants = FOREACH volssortants GENERATE
   group AS aeroport,
   COUNT(flightsclean) AS sortants;

-- Vols entrants
volsentrants = GROUP flightsclean BY Dest;
nbentrants = FOREACH volsentrants GENERATE
   group AS aeroport,
   COUNT(flightsclean) AS entrants;

-- Joindre et calculer le total
aeroportsstats = JOIN nbsortants BY aeroport FULL OUTER,
nbentrants BY aeroport;
aeroportstotal = FOREACH aeroportsstats GENERATE
   (nbsortants::aeroport IS NOT NULL ? nbsortants::aeroport :
nbentrants::aeroport) AS aeroport,
   (nbsortants::sortants IS NOT NULL ? nbsortants::sortants : 0L) AS
sortants,
   (nbentrants::entrants IS NOT NULL ? nbentrants::entrants : 0L) AS
entrants,
   (nbsortants::sortants IS NOT NULL ? nbsortants::sortants : 0L) +
   (nbentrants::entrants IS NOT NULL ? nbentrants::entrants : 0L) AS
total;

-- TOP 20
aeroportssorted = ORDER aeroportstotal BY total DESC;
top20 = LIMIT aeroportssorted 20;
```

```
STORE top20 INTO '/user/root/pigout/top20_aeroports';

-- ========================================
-- Question 2: POPULARITÉ DES TRANSPORTEURS
-- ========================================

-- Volume par transporteur et année
carrieryear = GROUP flightsclean BY (UniqueCarrier, Year);
carriervolume = FOREACH carrieryear GENERATE
   FLATTEN(group) AS (carrier, year),
   COUNT(flightsclean) AS nbvols;
STORE carriervolume INTO '/user/root/pigout/carriers_volume';

-- Volume médian par transporteur
carrierall = GROUP carriervolume BY carrier;
carriermedian = FOREACH carrierall GENERATE
   group AS carrier,
   AVG(carriervolume.nbvols) AS volumemedian;
carriermedianorder = ORDER carriermedian BY volumemedian DESC;
STORE carriermedianorder INTO '/user/root/pigout/carriers_median';

-- ========================================
-- Question 3: PROPORTION VOLS RETARDÉS
-- ========================================

-- Marquer les vols retardés (> 15 min)
volsretard = FOREACH flightsclean GENERATE
   Year, Month,
   (DepDelay IS NOT NULL AND DepDelay > 15 ? 1 : 0) AS retarde;

-- Par année
retardannee = GROUP volsretard BY Year;
propannee = FOREACH retardannee GENERATE
   group AS annee,
   SUM(volsretard.retarde) AS nbretardes,
```

```
   COUNT(volsretard) AS total,
   (double)SUM(volsretard.retarde) / (double)COUNT(volsretard) AS
proportion;
STORE propannee INTO '/user/root/pigout/retards_par_annee';

-- Par mois
retardmois = GROUP volsretard BY (Year, Month);
propmois = FOREACH retardmois GENERATE
   FLATTEN(group) AS (annee, mois),
   SUM(volsretard.retarde) AS nbretardes,
   COUNT(volsretard) AS total,
   (double)SUM(volsretard.retarde) / (double)COUNT(volsretard) AS
proportion;
STORE propmois INTO '/user/root/pigout/retards_par_mois';

-- ========================================
-- Question 4: RETARDS PAR TRANSPORTEUR
-- ========================================

volscarrierretard = FOREACH flightsclean GENERATE
   UniqueCarrier, Year,
   (DepDelay IS NOT NULL AND DepDelay > 15 ? 1 : 0) AS retarde;

retardcarrier = GROUP volscarrierretard BY UniqueCarrier;
propcarrier = FOREACH retardcarrier GENERATE
   group AS carrier,
   SUM(volscarrierretard.retarde) AS nbretardes,
   COUNT(volscarrierretard) AS total,
   (double)SUM(volscarrierretard.retarde) /
(double)COUNT(volscarrierretard) AS proportion;

propcarriersorted = ORDER propcarrier BY proportion DESC;
STORE propcarriersorted INTO '/user/root/pigout/retards_par_carrier';

-- ========================================
```

```
-- Question 5: ITINÉRAIRES PLUS FRÉQUENTÉS
-- ========================================

-- Créer des paires non ordonnées
routes = FOREACH flightsclean GENERATE
  (Origin < Dest ? CONCAT(CONCAT(Origin, '-'), Dest) :
CONCAT(CONCAT(Dest, '-'), Origin)) AS route;

routesgrouped = GROUP routes BY route;
routesfreq = FOREACH routesgrouped GENERATE
  group AS route,
  COUNT(routes) AS nbvols;

routessorted = ORDER routesfreq BY nbvols DESC;
toproutes = LIMIT routessorted 50;
STORE toproutes INTO '/user/root/pigout/top_routes';
EOF
```

- L’exécution du script avec succès :

```
root@hadoop-master:/shared_volume# pig /shared_volume/analyse_flights.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2025-11-15 15:59:07,223 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-11-15 15:59:07,226 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-11-15 15:59:07,226 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-11-15 15:59:07,278 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-11-15 15:59:07,278 [main] INFO  org.apache.pig.Main - Logging error messages to: /shared_volume/pig_1763222347277.log
2025-11-15 15:59:07,572 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2025-11-15 15:59:07,654 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.
address
2025-11-15 15:59:07,655 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://hadoop-mas
ter:9000/
2025-11-15 15:59:08,043 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-analyse_flights.pig-68e56248-a7d7-4e1f-a0ed-948d7e3cabe4
2025-11-15 15:59:08,247 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: localhost:8188
2025-11-15 15:59:08,559 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2025-11-15 15:59:08,589 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enabled
2025-11-15 15:59:09,485 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enabled
2025-11-15 15:59:09,716 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enabled
2025-11-15 16:03:20,482 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,482 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,484 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,496 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,496 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,498 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,508 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,508 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,511 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,532 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,532 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,534 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,546 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,546 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,548 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,558 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,558 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,560 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,577 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,578 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,579 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,591 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at hadoop-master/172.18.0.3:8032
2025-11-15 16:03:20,592 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at localhost/127.0.0.1:10200
2025-11-15 16:03:20,593 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
recting to job history server
2025-11-15 16:03:20,604 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TY
PE_CONVERSION_FAILED 9 time(s).
2025-11-15 16:03:20,604 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-11-15 16:03:20,618 [main] INFO  org.apache.pig.Main - Pig script completed in 4 minutes, 13 seconds and 453 milliseconds (253453 ms)
```

- Les résultats :

```
=== TOP 20 AÉROPORTS ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/top20_aer
oports/part-*
LAX     3       9       12
IND     9       0       9
BWI     0       9       9
LAS     0       8       8
ORD     3       3       6
ATL     3       3       6
FLL     3       3       6
SEA     3       0       3
BOS     3       0       3
SFO     0       3       3
DEN     3       0       3
DFW     0       3       3
EWR     3       0       3
PHX     3       0       3
HNL     3       0       3
IAH     0       3       3
JFK     3       0       3
MIA     3       0       3
IAD     2       0       2
TPA     0       2       2
```

```
=== VOLUME TRANSPORTEURS PAR ANNÉE ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/carriers_
volume/part-* | head -20
AA      2008    3
AS      2008    3
B6      2008    3
CO      2008    3
DL      2008    3
F9      2008    3
HA      2008    3
NK      2008    3
UA      2008    3
US      2008    3
WN      2008    14
YV      2008    3
```

```
=== VOLUME MÉDIAN TRANSPORTEURS ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/carriers_
median/part-*
WN      14.0
YV      3.0
US      3.0
UA      3.0
NK      3.0
HA      3.0
F9      3.0
DL      3.0
CO      3.0
B6      3.0
AS      3.0
AA      3.0
```

```
=== RETARDS PAR ANNÉE ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/retards_p
ar_annee/part-*
2008    4       47      0.0851063829787234
```

```
=== RETARDS PAR MOIS (échantillon) ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/retards_p
ar_mois/part-* | head -12
2008    1       4       14      0.2857142857142857
2008    2       0       3       0.0
2008    3       0       3       0.0
2008    4       0       3       0.0
2008    5       0       3       0.0
2008    6       0       3       0.0
2008    7       0       3       0.0
2008    8       0       3       0.0
2008    9       0       3       0.0
2008    10      0       3       0.0
2008    11      0       3       0.0
2008    12      0       3       0.0
```

```
=== RETARDS PAR TRANSPORTEUR ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/retards_p
ar_carrier/part-*
WN      4       14      0.2857142857142857
YV      0       3       0.0
US      0       3       0.0
UA      0       3       0.0
NK      0       3       0.0
HA      0       3       0.0
F9      0       3       0.0
DL      0       3       0.0
CO      0       3       0.0
B6      0       3       0.0
AS      0       3       0.0
AA      0       3       0.0
```

```
=== TOP 50 ROUTES ===
root@hadoop-master:/shared_volume# hdfs dfs -cat /user/root/pigout/top_route
s/part-* | head -20
BWI-IND 6
EWR-IAH 3
ATL-BWI 3
ATL-JFK 3
BOS-FLL 3
DEN-LAS 3
DFW-ORD 3
FLL-ORD 3
HNL-LAX 3
LAS-PHX 3
LAX-MIA 3
LAX-SEA 3
LAX-SFO 3
IAD-TPA 2
IND-LAS 2
IND-JAX 1
```