ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

# Atelier 4

Big Data

**4IIR14**


**Réalisé par :**

Amayou Aya

El ibrahimi Jawad


ANNEE UNIVERSITAIRE : 2024-2025

# Objectif

L'objectif de cet atelier est de manipuler et interroger des données tabulaires en utilisant **Hive**, un moteur d'entrepôt de données basé sur Hadoop. À travers une série d'exercices pratiques, l'étudiant apprend à :

- Créer des bases de données et des tables internes ou externes avec Hive.
- Charger et interroger des fichiers CSV via le langage **HQL (Hive Query Language)**.
- Comprendre la différence entre table interne et table externe.
- Observer le comportement de Hive selon les types de requêtes (avec ou sans déclenchement de job MapReduce).

## Hive: Atelier 4

1. Dans un terminal, lancer la commande: **hive**
2. Créer la BD *analyse*: **hive>** **CREATE DATABASE** *analyse*
3. Lister le contenu du dossier HDFS : */user/hive/warehouse*
4. Utiliser de la BD *analyse* : **hive>** **Use** *analyse*
5. Créer la table *vol1* (*year, month, day, fl, dep, arr, distance*)

```
hive> CREATE TABLE vol1
      ( year INT, month INT, day INT, fl STRING, dep STRING, arr STRING, distance INT )
      ROW FORMAT DELIMITED FIELDS TERMINATED BY '\;'
      STORED AS TEXTFILE;
```

6. Afficher la liste des tables de la BD courante.
7. Consulter le **Metastore** pour avoir le schéma de la table **vol1**: **hive>** **DESCRIBE** *vol1* ;
8. Charger le fichier **local** **vol.csv** daNs la table *vol1* en utilisant la commande *LOAD*

```
hive> LOAD DATA LOCAL INPATH '/home/cloudera/hive_lab/vol.csv' INTO TABLE vol1;
```

9. Consulter de la table:

```
hive> SELECT year, dep, COUNT(fl)
      FROM vol1
      GROUP BY dep, year;
```

**N.B:** *Remarquez les jobs Map-Reduce crées.*

**1------>7**

```
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create database analyse ;
OK
Time taken: 9.284 seconds
hive> dfs -ls
    > ;
Found 14 items
-rw-r--r--   1 cloudera cloudera      23392 2025-05-27 14:36 entre
drwxr-xr-x   - cloudera cloudera          0 2025-05-30 07:39 hive_lab
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 05:41 in
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 04:59 input
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 06:21 out
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 06:36 outp
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 06:40 outpp
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 14:16 outppp
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 05:34 output
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 05:26 output2
drwxr-xr-x   - cloudera cloudera          0 2025-05-30 05:34 pair
drwxr-xr-x   - cloudera cloudera          0 2025-05-30 05:21 soortie
drwxr-xr-x   - cloudera cloudera          0 2025-05-30 05:29 soortiee
drwxr-xr-x   - cloudera cloudera          0 2025-05-27 14:32 sort
hive> dfs -ls /user/hive/warehouse ;
Found 1 items
drwxrwxrwx   - cloudera supergroup          0 2025-05-30 07:47 /user/hive/warehouse/analyse.db
hive> use analyse ;
OK
Time taken: 0.417 seconds
hive> create table vol1
    > (year INT , month INT , day INT , fl STRING ,dep STRING , arr STRING , distance INT)
    > row format delimited fields terminated by '\;'
    > stored as textfile;
OK
Time taken: 4.558 seconds
hive> describe vol1 ;
OK
year                    int
month                   int
day                     int
fl                      string
dep                     string
arr                     string
distance                int
Time taken: 3.037 seconds, Fetched: 7 row(s)
hive>
```

8-

```
hive> load data local inpath '/home/cloudera/hive_lab/vol.csv' into table vol1 ;
Loading data to table analyse.vol1
Table analyse.vol1 stats: [numFiles=1, totalSize=2924]
OK
Time taken: 3.488 seconds
```

9-

```
hive> select year , dep , count(fl) from vol1 group by dep , year ;
Query ID = cloudera_20250530080606_7f0f28be-7507-4385-822b-a4f6346972c8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1748345518227_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/applic
ation_1748345518227_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1748345518227_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-05-30 08:07:25,405 Stage-1 map = 0%,  reduce = 0%
2025-05-30 08:07:55,897 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.5 sec
2025-05-30 08:08:12,594 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.87 sec
MapReduce Total cumulative CPU time: 7 seconds 870 msec
Ended Job = job_1748345518227_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.87 sec   HDFS Read: 11526 HDFS Write: 70 SUC
ESS
Total MapReduce CPU Time Spent: 7 seconds 870 msec
OK
2008    IAD     1
2008    IND     17
2008    ISP     28
2008    JAN     9
2008    JAX     23
2008    LAS     22
Time taken: 101.461 seconds, Fetched: 6 row(s)
```

## Hive: Atelier 4

10. Créer la table **externe** ***vol2*** (*year, month, day, fl, dep, arr, distance*) en indiquant son dossier HDFS de données qu'il faut créer, par exemple: ***/user/cloudera/hive/data/db***

```
CREATE EXTERNAL TABLE IF NOT EXISTS vol2
( year INT, month INT, day INT, fl STRING, dep STRING, arr STRING, distance INT )
COMMENT 'table des vols'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\;'
LOCATION '/user/cloudera/hive/data/db' ;
```

11. Copier le fichier local **vol.csv** dans le dossier HDFS: ***/user/cloudera/hive/data/db***

***Hive>*** dfs -put /home/cloudera/hive_lab/**vol.csv** /user/cloudera/hive/data/db;

12. Effectuer une requête HQL sur la table ***vol2***.

Select * from vol2;

13. Charger le fichier local **vol.csv** dans la table ***vol2*** avec **LOAD** et sans l'option **overwrite**.

***Hive>*** LOAD DATA LOCAL INPATH '/home/cloudera/hive_lab/vol.csv' INTO TABLE vol2;

10.

```
hive> dfs -mkdir -p /user/cloudera/hive/data/db ;
hive> dfs -ls /user/cloudera/hive/data/ ;
Found 1 items
drwxr-xr-x   - cloudera cloudera          0 2025-05-30 08:50 /user/cloudera/hive/data/db
hive> CREATE EXTERNAL TABLE IF NOT EXISTS vol2 (
    >     year INT,
    >     month INT,
    >     day INT,
    >     fl STRING,
    >     dep STRING,
    >     arr STRING,
    >     distance INT
    > )
    > COMMENT 'table des vols'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ';'
    > LOCATION '/user/cloudera/hive/data/db';
```

**11 ET 12**

```
Time taken: 0.272 seconds
hive> dfs -put /home/cloudera/hive_lab/vol.csv /user/cloudera/hive/data/db ;
hive> select * from vol2 ;
OK
2008    1       3       N772SW  IAD     TPA     810
2008    1       3       N428WN  IND     BWI     515
2008    1       3       N612SW  IND     BWI     515
2008    1       3       N464WN  IND     BWI     515
2008    1       3       N726SW  IND     JAX     688
2008    1       3       N763SW  IND     LAS     1591
2008    1       3       N428WN  IND     LAS     1591
2008    1       3       N689SW  IND     MCI     451
2008    1       3       N648SW  IND     MCI     451
2008    1       3       N690SW  IND     MCO     828
2008    1       3       N334SW  IND     MCO     828
2008    1       3       N476WN  IND     MDW     162
2008    1       3       N765SW  IND     MDW     162
2008    1       3       N420WN  IND     MDW     162
2008    1       3       N263WN  IND     MDW     162
2008    1       3       N286WN  IND     PHX     1489
2008    1       3       N778SW  IND     PHX     1489
2008    1       3       N674AA  IND     TPA     838
2008    1       3       N643SW  ISP     BWI     220
2008    1       3       N497WN  ISP     BWI     220
2008    1       3       N724SW  ISP     BWI     220
2008    1       3       N786SW  ISP     BWI     220
2008    1       3       N714CB  ISP     BWI     220
2008    1       3       N222WN  ISP     BWI     220
2008    1       3       N394SW  ISP     BWI     220
2008    1       3       N215WN  ISP     FLL     1093
2008    1       3       N243WN  ISP     FLL     1093
2008    1       3       N454WN  ISP     FLL     1093
2008    1       3       N712SW  ISP     LAS     2283
2008    1       3       N798SW  ISP     MCO     972
2008    1       3       N736SA  ISP     MCO     972
2008    1       3       N795SW  ISP     MCO     972
2008    1       3       N247WN  ISP     MCO     972
2008    1       3       N707SA  ISP     MCO     972
2008    1       3       N443WN  ISP     MCO     972
2008    1       3       N753SW  ISP     MDW     765
2008    1       3       N779SW  ISP     MDW     765
2008    1       3       N704SW  ISP     MDW     765
2008    1       3       N709SW  ISP     MDW     765
2008    1       3       N459WN  ISP     PBI     1052
2008    1       3       N621SW  ISP     PBI     1052
2008    1       3       N206WN  ISP     PBI     1052
```

13.

```
hive> select * from vol2 ;
OK
2008    1       3       N772SW  IAD     TPA     810
2008    1       3       N428WN  IND     BWI     515
2008    1       3       N612SW  IND     BWI     515
2008    1       3       N464WN  IND     BWI     515
2008    1       3       N726SW  IND     JAX     688
2008    1       3       N763SW  IND     LAS     1591
2008    1       3       N428WN  IND     LAS     1591
2008    1       3       N689SW  IND     MCI     451
2008    1       3       N648SW  IND     MCI     451
2008    1       3       N690SW  IND     MCO     828
2008    1       3       N334SW  IND     MCO     828
2008    1       3       N476WN  IND     MDW     162
2008    1       3       N765SW  IND     MDW     162
2008    1       3       N420WN  IND     MDW     162
2008    1       3       N263WN  IND     MDW     162
2008    1       3       N286WN  IND     PHX     1489
2008    1       3       N778SW  IND     PHX     1489
2008    1       3       N674AA  IND     TPA     838
2008    1       3       N643SW  ISP     BWI     220
2008    1       3       N497WN  ISP     BWI     220
2008    1       3       N724SW  ISP     BWI     220
2008    1       3       N786SW  ISP     BWI     220
2008    1       3       N714CB  ISP     BWI     220
2008    1       3       N222WN  ISP     BWI     220
2008    1       3       N394SW  ISP     BWI     220
2008    1       3       N215WN  ISP     FLL     1093
2008    1       3       N243WN  ISP     FLL     1093
```

## Hive: Atelier 4

14. Lister le contenu du dossier HDFS: **/user/cloudera/hive/data/db**

```
hive> dfs -ls /user/cloudera/hive/data/db;
```

15. Afficher les métadonnées **détaillées** de la table ***vol2***.

```
hive> desc formatted vol2;
```

16. Exécuter séparément les deux requêtes:

```
SELECT   *  FROM vol2;

SELECT      year, dep, COUNT(fl)
FROM        vol1
GROUP BY  dep, year;
```

**Quelle est la différence entre les deux lors de leur exécution?**

La requête SELECT year, dep, COUNT(fl) FROM vol1 GROUP BY dep, year déclenche l'exécution d'un job MapReduce, car elle nécessite une phase d'agrégation distribuée, tandis que la requête SELECT * FROM vol2 peut être exécutée sans générer de job MapReduce, car elle se contente de lire et retourner les données.

14

```
hive> dfs -ls /user/cloudera/hive/data/db;
Found 2 items
-rw-r--r--   1 cloudera cloudera       2924 2025-05-30 08:57 /user/cloudera/hive/data/db/vol.csv
-rwxr-xr-x   1 cloudera cloudera       2924 2025-05-30 08:58 /user/cloudera/hive/data/db/vol_copy_1.csv
hive>
```

15

```
hive> desc formatted vol2;
OK
# col_name              data_type               comment

year                    int
month                   int
day                     int
fl                      string
dep                     string
arr                     string
distance                int

# Detailed Table Information
Database:               analyse
Owner:                  cloudera
CreateTime:             Fri May 30 08:54:15 PDT 2025
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/cloudera/hive/data/db
Table Type:             EXTERNAL_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   true
        EXTERNAL                TRUE
        comment                 table des vols
        numFiles                2
        numRows                 0
        rawDataSize             0
        totalSize               5848
        transient_lastDdlTime   1748620726

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             ;
        serialization.format    ;
Time taken: 0.719 seconds, Fetched: 40 row(s)
hive>
```

16.

```
hive> SELECT
    > year, dep, COUNT(fl)
    > FROM
    > vol1
    > GROUP BY dep, year;
Query ID = cloudera_20250530095151_6f265839-5920-44f1-93b2-c813fda058a3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1748345518227_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1748345
518227_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1748345518227_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-05-30 09:52:04,870 Stage-1 map = 0%,  reduce = 0%
2025-05-30 09:52:43,408 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.71 sec
2025-05-30 09:53:03,161 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.61 sec
MapReduce Total cumulative CPU time: 9 seconds 610 msec
Ended Job = job_1748345518227_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.61 sec   HDFS Read: 11596 HDFS Write: 70 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 610 msec
OK
2008    IAD     1
2008    IND     17
2008    ISP     28
2008    JAN     9
2008    JAX     23
2008    LAS     22
Time taken: 106.299 seconds, Fetched: 6 row(s)
```

17.

```
hive> CREATE TABLE vol3
    > ( year INT, month INT, day INT, fl STRING, dep STRING, arr STRING, distance INT )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\;'
    > STORED AS TEXTFILE;
OK
Time taken: 0.372 seconds
hive> INSERT INTO TABLE vol3 SELECT * FROm vol2;
Query ID = cloudera_20250530095555_a4e100d9-9667-4672-bcf2-d9066d00016c
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1748345518227_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1748345
518227_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1748345518227_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-05-30 09:55:21,590 Stage-1 map = 0%,  reduce = 0%
2025-05-30 09:55:35,638 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.13 sec
MapReduce Total cumulative CPU time: 5 seconds 130 msec
Ended Job = job_1748345518227_0013
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/analyse.db/vol3/.hive-staging_hive_2025-05-
30_09-55-05_982_5528657245929350171-1/-ext-10000
Loading data to table analyse.vol3
Table analyse.vol3 stats: [numFiles=1, numRows=200, totalSize=5648, rawDataSize=5448]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 5.13 sec   HDFS Read: 10158 HDFS Write: 5719 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 130 msec
OK
Time taken: 33.041 seconds
hive>
```

```
hive> SELECT * FROM vol2;
OK
2008    1       3       N772SW  IAD     TPA     810
2008    1       3       N428WN  IND     BWI     515
2008    1       3       N612SW  IND     BWI     515
2008    1       3       N464WN  IND     BWI     515
2008    1       3       N726SW  IND     JAX     688
2008    1       3       N763SW  IND     LAS     1591
2008    1       3       N428WN  IND     LAS     1591
2008    1       3       N689SW  IND     MCI     451
2008    1       3       N648SW  IND     MCI     451
2008    1       3       N690SW  IND     MCO     828
2008    1       3       N334SW  IND     MCO     828
2008    1       3       N476WN  IND     MDW     162
2008    1       3       N765SW  IND     MDW     162
2008    1       3       N420WN  IND     MDW     162
2008    1       3       N263WN  IND     MDW     162
2008    1       3       N286WN  IND     PHX     1489
2008    1       3       N778SW  IND     PHX     1489
2008    1       3       N674AA  IND     TPA     838
2008    1       3       N643SW  ISP     BWI     220
2008    1       3       N497WN  ISP     BWI     220
2008    1       3       N724SW  ISP     BWI     220
2008    1       3       N786SW  ISP     BWI     220
2008    1       3       N714CB  ISP     BWI     220
2008    1       3       N222WN  ISP     BWI     220
2008    1       3       N394SW  ISP     BWI     220
2008    1       3       N215WN  ISP     FLL     1093
2008    1       3       N243WN  ISP     FLL     1093
2008    1       3       N454WN  ISP     FLL     1093
2008    1       3       N712SW  ISP     LAS     2283
2008    1       3       N798SW  ISP     MCO     972
2008    1       3       N736SA  ISP     MCO     972
2008    1       3       N795SW  ISP     MCO     972
2008    1       3       N247WN  ISP     MCO     972
2008    1       3       N707SA  ISP     MCO     972
2008    1       3       N443WN  ISP     MCO     972
2008    1       3       N753SW  ISP     MDW     765
2008    1       3       N779SW  ISP     MDW     765
2008    1       3       N704SW  ISP     MDW     765
2008    1       3       N709SW  ISP     MDW     765
2008    1       3       N459WN  ISP     PBI     1052
2008    1       3       N621SW  ISP     PBI     1052
2008    1       3       N206WN  ISP     PBI     1052
2008    1       3       N280WN  ISP     RSW     1101
2008    1       3       N241WN  ISP     TPA     1034
2008    1       3       N200WN  ISP     TPA     1034
2008    1       3       N459WN  ISP     TPA     1034
```

# Conclusion

À la fin de cet atelier, les étudiants ont acquis une maîtrise de base de **Hive** pour le traitement des données en mode distribué. Ils ont appris à :

- Gérer des structures tabulaires dans un environnement Big Data.
- Charger, explorer et analyser des données à l'aide de requêtes HQL.
- Comprendre le fonctionnement du Metastore et l'impact du type de table sur le comportement des requêtes.
- Différencier les traitements nécessitant un job MapReduce de ceux exécutés localement.

Ces compétences sont essentielles pour toute démarche d'analyse de données volumineuses dans un contexte Hadoop, et constituent une base solide pour des traitements analytiques plus avancés.