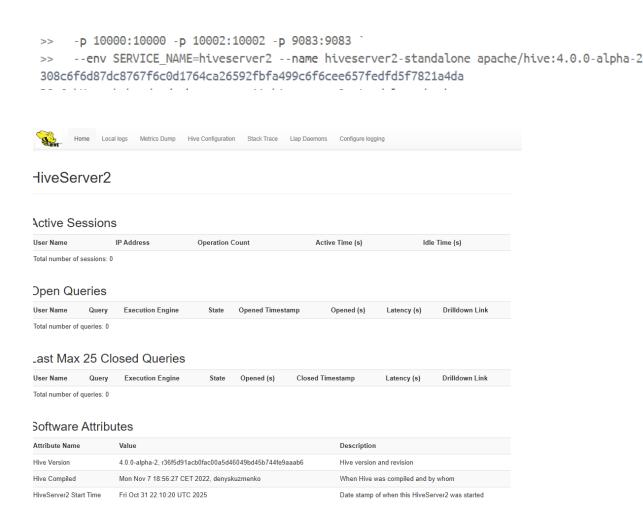
Lab 6: Apache Hive

I. Installation Apache Hive



II. Première utilisation beeline

```
hive@308c6f6d87dc:/opt/hive$ hadoop fs -ls
Found 16 items
                            4096 2023-05-02 18:52 data
drwxr-xr-x - root root
-rw-r--r-- 1 hive hive
                          19957 2025-10-31 22:09 derby.log
                            4096 2023-05-02 18:52 examples
drwxr-xr-x
           - root root
                            4096 2023-05-02 18:52 hcatalog
drwxr-xr-x
           - root root
                            4096 2023-05-02 18:52 jdbc
           - root root
drwxr-xr-x
           - root root
                          20480 2023-05-02 18:52 lib
drwxr-xr-x
drwxr-xr-x
           - hive hive
                           4096 2025-10-31 22:09 metastore db
drwx-wx-wx

    hive hive

                           4096 2025-10-31 22:09 scratch_dir
drwxr-xr-x
           - root root
                            4096 2023-05-02 18:52 scripts
hive@308c6f6d87dc:/opt/hive$
```

```
\verb|hive@308c6f6d87dc:/opt/hive$| cat /opt/hive/conf/hive-site.xml| \\
<?xml version="1.0" encoding="UTF-8"?>
  Licensed to the Apache Software Foundation (ASF) under one or more
  contributor license agreements. See the NOTICE file distributed with
  this work for additional information regarding copyright ownership.
  The ASF licenses this file to You under the Apache License, Version 2.0
    the "License"); you may not use this file except in compliance with
  the License. You may obtain a copy of the License at
      http://www.apache.org/licenses/LICENSE-2.0
  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS.
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License.
INFO . Starting task | Stage-0.DDL| til sertat m
INFO : Completed executing command(queryId=hiv
| database_name |
| default
1 row selected (9.959 seconds)
0: jdbc:hive2://localhost:10000>
```

III. Analyse de données de réservation d'hôtels

1. Créer la base de données

```
0: jdbc:hive2://localhost:10000> CREATE DATABASE hotel_booking;
INFO : Compiling command(queryId=hive_20251101000048_6d3dcb07-a2ce-42bd-92da-53c2dd8021a4): CREATE DATABASE hotel_booking
INFO : Semantic Analysis Completed (retrial = false)
INFO : Created Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20251101000048_6d3dcb07-a2ce-42bd-92da-53c2dd8021a4); Time taken: 0.192 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive 20251101000048 6d3dcb07-a2ce-42bd-92da-53c2dd8021a4): CREATE DATABASE hotel booking
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20251101000048_6d3dcb07-a2ce-42bd-92da-53c2dd8021a4); Time taken: 0.459 seconds
No rows affected (1.041 seconds)
0: jdbc:hive2://localhost:10000> USE hotel booking;
INFO : Compiling command(queryId=hive_20251101000056_06942fa2-1ffc-47e4-9dd2-40de24303ff8): USE hotel_booking
INFO : Semantic Analysis Completed (retrial = false)
INFO : Created Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20251101000056_06942fa2-1ffc-47e4-9dd2-40de24303ff8); Time taken: 0.194 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251101000056_06942fa2-1ffc-47e4-9dd2-40de24303ff8): USE hotel_booking
PS C:\Users\ahmed> docker exec -it hiveserver2-standalone bash -c "ls -l /opt/hive/data/warehouse || true"
total 4
drwxr-xr-x 2 hive hive 4096 Nov 1 00:00 hotel_booking.db
PS C:\Users\ahmed>
```

 Après CREATE DATABASE, Hive crée un dossier nommé hotel_booking.db dans le répertoire warehouse « /opt/hive/data/warehouse/hotel_booking.db ».

2. Créer les tables

```
0: jdbc:hive2://localhost:10000> set hive.exec.dynamic.partition=true;
No rows affected (1.176 seconds)
0: jdbc:hive2://localhost:10000> set hive.exec.dynamic.partition.mode=nonstrict;
No rows affected (0.029 seconds)
0: jdbc:hive2://localhost:10000> set hive.exec.max.dynamic.partitions=20000;
No rows affected (0.048 seconds)
0: jdbc:hive2://localhost:10000> set hive.exec.max.dynamic.partitions.pernode=20000;
No rows affected (0.025 seconds)
0: jdbc:hive2://localhost:10000> set hive.exec.max.dynamic.partitions.pernode=20000;
No rows affected (0.025 seconds)
0: jdbc:hive2://localhost:10000> cREATE TABLE clients ( client_id INT, nom STRING, email ST
...........> telephone STRING )
........> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
.......> STORED AS TEXTFILE;
```

Création de la table client

```
0: jdbc:hive2://localhost:10000> CREATE TABLE clients ( client_id INT, nom STRING, email STRING,
.....................> relephone STRING )
...............> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
.............> STORED AS TEXTFILE;
INFO : Compiling command(queryId=hive_20251101001121_7159d921-9325-4758-9fee-5e8fb8a451bd): CREATE TABLE clients ( client_id INT, nom STRING, netlephone STRING )
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO : Semantic Analysis Completed (retrial = false)
INFO : Created Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20251101001121_7159d921-9325-4758-9fee-5e8fb8a451bd); Time taken: 0.91 seconds
INFO : Executing command(queryId=hive_20251101001121_7159d921-9325-4758-9fee-5e8fb8a451bd): CREATE TABLE clients ( client_id INT, nom STRING, netlephone STRING )
```

De même pour la table des reservations et hotels

3. Charger les données dans les tables

Pour clients et hotel

```
0: jdbc:hive2://localhost:10000> LOAD DATA LOCAL INPATH '/shared_volume/hive_data/clients.txt' INTO TABLE clients;
INFO: Compiling \ command (queryId=hive\_20251101002601\_1740ae10-c93f-4b08-a4c4-fe7fd08d1724): \ LOAD \ DATA \ LOCAL \ INPATH \ '/shared_volume/hive\_data/classical-local \ LOCAL \ L
ients.txt' INTO TABLE clients
INFO : Semantic Analysis Completed (retrial = false)
INFO : Created Hive schema: Schema(fieldSchemas:null, properties:null)
INFO: Completed \ compiling \ command (queryId=hive\_20251101002601\_1740ae10-c93f-4b08-a4c4-fe7fd08d1724); \ Time \ taken: \ 0.837 \ seconds \ described by the compiling \ command \ compiling \ compiling \ command \ compiling \ compili
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251101002601_1740ae10-c93f-4b08-a4c4-fe7fd08d1724): LOAD DATA LOCAL INPATH '/shared_volume/hive_data/cl
ients.txt' INTO TABLE clients
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Loading data to table default.clients from file:/shared_volume/hive_data/clients.txt
INFO : Starting task [Stage-1:STATS] in serial mode
INFO : Executing stats task
INFO : Table default.clients stats: [numFiles=1. numRows=0. totalSize=1579. rawDataSize=0. numFilesErasureCoded=0]
INFO : Completed executing command(queryId=hive 20251101002601 1740ae10-c93f-4b08-a4c4-fe7fd08d1724); Time taken: 1.99 seconds
No rows affected (3.267 seconds)
0: jdbc:hive2://localhost:10000>
```

On insere les données dans une table de staging « raw reservations »

```
servations.txt' INTO TABLE raw_reservations

INFO : Starting task [Stage-0:MOVE] in serial mode

INFO : Loading data to table default.raw_reservations from file:/shared_volume/hive_data/reservations.txt

INFO : Starting task [Stage-1:STATS] in serial mode

INFO : Executing stats task

INFO : Table default.raw_reservations stats: [numFiles=1, numRows=0, totalSize=1228, rawDataSize=0, numFilesErasureCoded=0]

INFO : Completed executing command(queryId=hive_20251101003940_869e1a62-278a-49e8-aea4-cc7d2b0e4669); Time taken: 0.381 seconds

No rows affected (0.602 seconds)

0: jdbc:hive2://localhost:10000> SET hive.exec.dynamic.partition=true;

No rows affected (0.041 seconds)

0: jdbc:hive2://localhost:10000> SET hive.exec.dynamic.partition.mode=nonstrict;

No rows affected (0.021 seconds)
```

4. Utiliser des partitions et des buckets

Charger dans les tables

remplir hotels_partitioned à partir de la table hotels (dynamic partition) :

```
VERTICES
                                STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED
Reducer 2 .... container SUCCEEDED
                                          1
1
                                                      1
                                                               0
                                                                        0
                                                                                0
                                                  1
                                                               0
                                                                        0
                                                                               0
                                                                      0
Reducer 3 ..... container SUCCEEDED 1
                                                              0
                                                                              0
                                                                                       0
                               STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
       VERTICES MODE
Map 1 ..... container SUCCEEDED Reducer 2 .... container SUCCEEDED
                                                                               0
                                                                                      0 /hive/data/warehouse/hotels_partitioned/.hive-stag
Reducer 3 ..... container SUCCEEDED
                                                                               0
                                          1
VERTICES: 03/03 [======>>] 100% ELAPSED TIME: 7.44 s
INFO : Time taken to load dynamic partitions: 0.379 seconds INFO : Time taken for adding to write entity : 0.002 seconds
INFO : Starting task [Stage-3:STATS] in serial mode
```

remplir reservations_bucketed à partir d'une table de staging raw_reservations

```
INFO : Starting task [Stage-0:MOVE] in serial mode----
INFO : Loading data to table default.reservations_bucketed from file:/opt/hive/data/warehouse/reservations_bucketed/.hive-staging_hive_2025-11-
INFO : Executing stats task------
     VERTICES MODE
                         STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
                                   1
Map 1 ..... container SUCCEEDED

Reducer 2 ..... container SUCCEEDED
                                                    0
Reducer 2 ..... container
                         SUCCEEDED
Reducer 3 ..... container SUCCEEDED
                                   1
                                                                        0
VERTICES: 03/03 [======>>] 100% ELAPSED TIME: 10.03 s
INFO : Completed executing command(queryId=hive_20251101004604_41d7be87-4cca-415c-9fdb-fa33abd2c9ae); Time taken: 11.578 seconds
31 rows affected (14.54 seconds)
0: jdbc:hive2://localhost:10000>
DAM 4.33 CD CDI 137 E39 Dialy 33.13 CD wood (limit 1004 05 CD)
```

5. Utilisation de requêtes simples

Lister tous les clients

Lister tous les hôtels à Paris

INFO : Concurrency mode is disabled, not creating a lock manager

```
0: jdbc:hive2://localhost:10000 SELECT * FROM hotels WHERE ville = 'Paris';

INFO : Compiling command(queryId=hive_20251101011637_78b89243-37a3-40af-82b9-e98667d86d98): SELECT * FROM hotels WHERE ville = 'Paris'

INFO : No Stats for default@hotels, Columns: ville, etoiles, hotel_id, nom

INFO : Semantic Analysis Completed (retrial = false)

INFO : Completed Hive schema: Schema(fieldSchemas:[FieldSchema(name:hotels.hotel_id, type:int, comment:null), FieldSchema(name:hotels.nom, type:string, comment:null), FieldSchema(name:hotels.etoiles, type:int, comment:null), FieldSchema(name:hotels.ville, type:string, comment:null)

INFO : Completed compiling command(queryId=hive_20251101011637_78b89243-37a3-40af-82b9-e98667d86d98); Time taken: 2.923 seconds
```

Lister toutes les réservations avec les informations sur les hôtels et les clients

INFO : Completed executing command(queryId=hive_20251101011859_13b5ab6b-808d-4892-bb06-2274eaa9d3ed); Time taken: 6.434 seconds

INFO : Executina command(auervId=hive 20251101011637 78b89243-37a3-40af-82b9-e98667d86d98): SELECT * FROM hotels WHERE ville = 'Paris'

r.reservation_id	c.client_id	client_nom	h.hotel_id	hotel_nom	r.date_debut	r.date_fin	r.prix_total
23	23	Bruce Wayne	23	Oceanic Hotel	2024-12-10	2024-12-13	2400.00
12	12	David Miller	12	Forest Escape	2024-12-10	2024-12-12	1800.00
2	2	Sarah Connor	2	Beach Resort	2024-12-10	2024-12-15	800.00
1	1	John Doe	1	Grand Hotel	2024-12-01	2024-12-05	1500.00
6	6	Alex Dupont	6	Sea Breeze	2024-12-01	2024-12-03	1200.00
13	13	Laura Wilson	13	Spa Paradise	2024-12-01	2024-12-03	1600.00
7	7	Sophie Martin	7	Country Retreat	2024-12-15	2024-12-20	1000.00
26	26	Tony Stark	26	Urban Stay	2024-12-15	2024-12-18	1300.00
4	4	Lara Croft	4	Budget Inn	2024-12-15	2024-12-18	300.00
9	9	Emily Davis	9	City Center Lodge	2024-12-20	2024-12-22	700.00
27	27	Natasha Romanoff	27	Heritage Hotel	2024-12-20	2024-12-23	1400.00
3	3	James Bond	3	Mountain Lodge	2024-12-20	2024-12-25	600.00
16	16	Oscar Wilde	16	Château Bellevue	2024-12-02	2024-12-06	1300.00

6. Requêtes avec jointures

• Afficher le nombre de réservations par client

SELECT c.client_id, c.nom, COUNT(*) AS nb_reservations

FROM reservations r

JOIN clients c ON r.client_id = c.client_id

GROUP BY c.client_id, c.nom

ORDER BY nb_reservations DESC;

c.client_id	c.nom	nb_reservations
1	John Doe	1
2	Sarah Connor	1
3	James Bond	1
4	Lara Croft	1
5	Maria Gonzalez	1
6	Alex Dupont	1
7	Sophie Martin	1
8	Paul Durand	1
9	Emily Davis	1
10	Robert Brown	1
11	Alice Cooper	1
12	David Miller	1
13	Laura Wilson	1

Afficher les clients qui ont réservé plus que 2 nuitées

 $SELECT\ r.client_id,\ c.nom,\ SUM(datediff(CAST(r.date_fin\ AS\ DATE),\ CAST(r.date_debut\ AS\ DATE)))\ AS\ total_nuits$

FROM reservations r JOIN clients c ON r.client_id = c.client_id

GROUP BY r.client_id, c.nom HAVING total_nuits > 2;

r.client_id	c.nom	total_nuits	
+	+	-++	
1	John Doe	4	
2	Sarah Connor	5	
3	James Bond	5	
4	Lara Croft	3	
7	Sophie Martin	5	
8	Paul Durand	4	
10	Robert Brown	3	
11	Alice Cooper	5	
14	Chris Evans	3	
15	Megan Fox	3	
16	Oscar Wilde	4	
17	Rachel Green	4	
18	Ryan Gosling	4	
19	Emma Watson	4	

• Afficher les Hôtels réservés par chaque client

SELECT r.client_id, c.nom AS client_nom, concat_ws(',', collect_set(h.nom)) AS hotels_reserves

FROM reservations r

JOIN clients c ON r.client_id = c.client_id

JOIN hotels h ON r.hotel_id = h.hotel_id

GROUP BY r.client_id, c.nom;

+		++
r.client_id	client_nom	hotels_reserves
+		++
1	John Doe	Grand Hotel
2	Sarah Connor	Beach Resort
3	James Bond	Mountain Lodge
4	Lara Croft	Budget Inn
5	Maria Gonzalez	Luxury Palace
6	Alex Dupont	Sea Breeze
7	Sophie Martin	Country Retreat
8	Paul Durand	Riverside Hotel
9	Emily Davis	City Center Lodge
10	Robert Brown	Eco Stay
11	Alice Cooper	Hotel de Ville

• Afficher les noms des hôtels dans lesquels il y a plus qu'une réservation.

SELECT h.nom, COUNT(*) AS cnt

FROM reservations r

JOIN hotels h ON r.hotel_id = h.hotel_id

GROUP BY h.nom

HAVING cnt > 1;

• Afficher les noms des hôtels dans lesquels il y a pas de réservation.

SELECT h.nom

FROM hotels h

LEFT JOIN reservations r ON h.hotel_id = r.hotel_id

WHERE r.reservation_id IS NULL;

7. Requêtes imbriquées

Afficher les clients ayant réservé un hôtel avec plus de 4 étoiles

```
SELECT DISTINCT c.client_id, c.nom

FROM clients c

WHERE c.client_id IN ( SELECT r.client_id

FROM reservations r JOIN hotels h ON r.hotel_id = h.hotel_id

WHERE h.etoiles > 4);
```

Afficher le Total des revenus générés par chaque hôtel

```
SELECT h.hotel_id, h.nom, SUM(r.prix_total) AS revenu_total
FROM reservations r JOIN hotels h ON r.hotel_id = h.hotel_id
GROUP BY h.hotel_id, h.nom
ORDER BY revenu_total DESC;
```

8. Utilisation de fonctions d'agrégation avec partitions et buckets

Revenus totaux par ville (partitionnée)

```
SELECT h.ville, SUM(r.prix_total) AS revenu_par_ville

FROM reservations r JOIN hotels h ON r.hotel_id = h.hotel_id

GROUP BY h.ville

ORDER BY revenu_par_ville DESC;
```

SELECT client_id, COUNT(*) AS nb_reservations

Nombre total de réservations par client (bucketed)

9. Nettoyage et suppression des données

Supprimer les tables créés précédemment.

10. Script hal

Mis en place des scripts qui créent les tables et font l'affaire.

Pour le loading, la suite des mapreduce est mis en place comme processus de Hive.

De même pour l'execution du code de queries.hql.

HiveServer2

Active Sessions

User Name	IP Address	Operation Count	Active Time (s)	Idle Time (s)
anonymous	127.0.0.1	0	3849	758
scott	127.0.0.1	0	6827	6541

Total number of sessions: 2

Open Queries

User Name Query Execution Er	ne State	Opened Timestamp	Opened (s)	Latency (s)	Drilldown Link	
------------------------------	----------	------------------	------------	-------------	----------------	--

Total number of queries: 0

Last Max 25 Closed Queries

User Name	Query	Execution Engine	State	Opened (s)	Closed Timestamp	Latency (s)	Drilldown Link
scott	SELECT client_id, COUNT(*) AS nb_reservations FROM reservations_bucketed GROUP BY client_id ORDER BY nb_reservations DESC	tez	FINISHED	1	Sat Nov 01 01:58:15 UTC 2025	1	Drilldown
scott	SELECT h.ville, SUM(r.prix_total) AS revenu_par_ville FROM reservations r JOIN hotels h ON r.hotel_id = h.hotel_id GROUP BY h.ville ORDER BY revenu_par_ville DESC	tez	FINISHED	3	Sat Nov 01 01:58:13 UTC 2025	3	Drilldown
scott	SELECT h.hotel_id, h.nom, SUM(r.prix_total) AS revenu_total FROM reservations r JOIN hotels h ON r.hotel_id = h.hotel_id GROUP BY h.hotel_id, h.nom ORDER BY revenu_total DESC	tez	FINISHED	3	Sat Nov 01 01:58:09 UTC 2025	3	Drilldown
scott	SELECT DISTINCT c.client_id, c.nom FROM clients c WHERE c.client_id IN (SELECT r.client_id FROM reservations r JOIN hotels	tez	FINISHED	19	Sat Nov 01 01:58:05 UTC	18	Drilldown