Lab 6: Apache Hive

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>
```

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.enforce.bucketing = true ;
No rows affected (0.017 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,
. . . . . . . . . . . . . . .> telephone 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,
mail STRING,
telephone 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  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20251101001121_7159d921-9325-4758-9fee-5e8fb8a451bd): CREATE TABLE clients ( client_id INT, nom STRING,
mail STRING,
telephone STRING )
```

```
0: jdbc:hive2://localhost:10000> CREATE TABLE reservations (
. . . . . . . . . . . . . . .>   reservation_id INT,
. . . . . . . . . . . . . . .>   client_id INT,
. . . . . . . . . . . . . . .>   hotel_id INT,
. . . . . . . . . . . . . . .>   date_fin DATE,
. . . . . . . . . . . . . . .>   prix_total DECIMAL(10,2)
. . . . . . . . . . . . . . .> )
. . . . . . . . . . . . . . .> PARTITIONED BY (date_debut STRING)
. . . . . . . . . . . . . . .> ROW FORMAT DELIMITED
. . . . . . . . . . . . . . .> FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20251101001605_462bbdd6-d464-418f-b94f-be5dd22e170f): CREATE TABLE reservations (
reservation_id INT,
```

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/cl
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
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>
```

Inserer 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

Chargement dans les tables

```
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0  /hive/data/warehouse/hotels_partitioned/.hive-stag
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
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
```

```
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-
01_00-46-04_980_1800327031563457623-2/-ext-10000        4        0        0       0       0
INFO  : Starting task [Stage-3:STATS] in serial mode    0        0        1       0       0
INFO  : Executing stats task-------------------------------------------------------------------
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       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>
```

5. Utilisation de requêtes simples

- List des clients

```
0: jdbc:hive2://localhost:10000> select * from clients
. . . . . . . . . . . . . . . . .> ;
INFO  : Compiling command(queryId=hive_20251101011441_dd540941-695b-4521-a0bc-fba5dcc47f35): select * from clients
INFO  : No Stats for default@clients, Columns: telephone, nom, client_id, email
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:clients.client_id, type:int, comment:null), FieldSchema(name:clients.nom, '
e:string, comment:null), FieldSchema(name:clients.email, type:string, comment:null), FieldSchema(name:clients.telephone, type:string, comment
ll)], properties:null)
INFO  : Completed compiling command(queryId=hive_20251101011441_dd540941-695b-4521-a0bc-fba5dcc47f35); Time taken: 2.798 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20251101011441_dd540941-695b-4521-a0bc-fba5dcc47f35): select * from clients
INFO  : Completed executing command(queryId=hive_20251101011441_dd540941-695b-4521-a0bc-fba5dcc47f35); Time taken: 0.006 seconds
+--------------------+--------------------+--------------------------------+--------------------+
| clients.client_id  |    clients.nom     |         clients.email          | clients.telephone  |
+--------------------+--------------------+--------------------------------+--------------------+
| NULL               | nom                | email                          | telephone          |
| 1                  | John Doe           | john.doe@example.com           | 1234567890         |
```

- List des hôtels à Paris

```
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  : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:hotels.hotel_id, type:int, comment:null), FieldSchema(name:hotels.nom, type:s
tring, comment:null), FieldSchema(name:hotels.etoiles, type:int, comment:null), FieldSchema(name:hotels.ville, type:string, comment:null)], prop
erties:null)
INFO  : Completed compiling command(queryId=hive_20251101011637_78b89243-37a3-40af-82b9-e98667d86d98); Time taken: 2.923 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20251101011637_78b89243-37a3-40af-82b9-e98667d86d98): SELECT * FROM hotels WHERE ville = 'Paris'
```

- List des 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
+------------------+-------------+------------------+------------+--------------------+--------------+-------------+--------------+
| 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

```
+--------------+--------------------+-------------------+
| 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

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

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 le Total des revenus générés par chaque hôtel

```
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 7.47 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20251101013102_ca94f81d-06eb-4e10-81a
+------------+--------------------+--------------+
| h.hotel_id |       h.nom        | revenu_total |
+------------+--------------------+--------------+
| 11         | Hotel de Ville     | 2500.00      |
| 23         | Oceanic Hotel      | 2400.00      |
| 24         | Hilltop Inn        | 2200.00      |
| 15         | Lakeview Resort    | 2100.00      |
| 25         | Cloud Nine         | 2000.00      |
| 5          | Luxury Palace      | 2000.00      |
| 22         | Golden Sands       | 1800.00      |
| 12         | Forest Escape      | 1800.00      |
| 19         | Horizon Hotel      | 1700.00      |
| 29         | Springfield Inn    | 1600.00      |
| 13         | Spa Paradise       | 1600.00      |
```

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

Revenus totaux par ville (partitionnée)

```
----------------------------------------
INFO  : Completed executing command(que
+----------+--------------------+
| h.ville  | revenu_par_ville   |
+----------+--------------------+
| 5        | 14500.00           |
| 4        | 12600.00           |
| 3        | 8500.00            |
| 2        | 5500.00            |
+----------+--------------------+
4 rows selected (17.159 seconds)
0: jdbc:hive2://localhost:10000>
```

Nombre total de réservations par client (bucketed)

```
----------------------------------------
INFO  : Completed executing command(query
+------------+------------------+
| client_id  | nb_reservations  |
+------------+------------------+
| 1          | 1                |
| 2          | 1                |
| 3          | 1                |
| 4          | 1                |
| 5          | 1                |
```

9. Nettoyage et suppression des données

- Supprimer les tables créés précédemment.

10. Script hql

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.

```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0 :/opt/hive/data/warehouse/hotel_booking.db/reserva
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 37.91 s
----------------------------------------------------------------------------------------------
```