Projet Architecture BigData & Hadoop

<u>Présenté par :</u>

Astrid Aurelien NKUMBE ENONGENE MSc 2 Info IA
Anne-Josée LOUIS MSc 2 Info IA

Ionis-STM
2024

Sommaire :

I. Ingestion et Normalisation

Sources de Données

1. Données de la base de données du site e-commerce contenant les détails sur les clients, commandes et livraisons.
2. Données de la base de données interne (ERP) contenant des détails sur les commandes et les livraisons et feedback des clients : réponses aux enquêtes de satisfaction client.
3. Données externes : des informations sur la météo et le trafic.

Script DDL pour MySQL

SOURCE 1

```
-- Création de la table pour la base de données e-commerce, uniquement la table customers
CREATE TABLE `CUSTOMERS` (
  `CUSTOMER_ID` INT(11) NOT NULL AUTO_INCREMENT,
  `CUSTOMER_NAME` VARCHAR(26) DEFAULT NULL,
  `GENDER` VARCHAR(26) DEFAULT NULL,
  `AGE` INT(11) DEFAULT NULL,
  `HOME_ADDRESS` VARCHAR(128) DEFAULT NULL,
  `ZIP_CODE` INT(11) DEFAULT NULL,
  `CITY` VARCHAR(26) DEFAULT NULL,
  `STATE` VARCHAR(128) DEFAULT NULL,
  `COUNTRY` VARCHAR(26) DEFAULT NULL,
  `DATE_NAISSANCE` DATE DEFAULT NULL,
  PRIMARY KEY (`CUSTOMER_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_general_ci;
```

SOURCE 2

```
-- Création de la table pour la base de données interne (ERP)
CREATE TABLE `erp_orders` (
  `order_id` INT NOT NULL AUTO_INCREMENT,
  `order_date` DATE NOT NULL,
  `expected_delivery_date` DATE NOT NULL,
  `actual_delivery_date` DATE,
  `delivery_status` VARCHAR(50),
  `delivery_cost` DECIMAL(10, 2),
  PRIMARY KEY (`order_id`)
);


-- Création de la table pour le feedback des clients

CREATE TABLE `customer_feedback` (
  `feedback_id` INT NOT NULL AUTO_INCREMENT,
  `order_id` INT NOT NULL,
  `satisfaction_rating` TINYINT NOT NULL,
  `comment` TEXT,
  PRIMARY KEY (`feedback_id`),
  FOREIGN KEY (`order_id`) REFERENCES `erp_orders`(`order_id`) ON DELETE CASCADE
);

-- Créer un index sur la colonne `satisfaction_rating`
CREATE INDEX idx_satisfaction ON customer_feedback (satisfaction_rating);
```

SOURCE 3

```
-- Création de la table pour les données externes (Conditions météorologiques et trafic)
CREATE TABLE `weather_data` (
  `data_id` INT NOT NULL AUTO_INCREMENT,
  `date` DATE NOT NULL,
  `weather_condition` VARCHAR(100),
  `traffic_intensity` VARCHAR(100),
  PRIMARY KEY (`data_id`)
);

-- Créer un index sur la colonne `date`
CREATE INDEX idx_date ON weather_data(date);
```

Dictionnaire de données

Table : CUSTOMERS

- CUSTOMER_ID (integer) : Identifiant unique du client.
- CUSTOMER_NAME (string) : Nom du client.
- GENDER (string) : Sexe du client.
- AGE (integer) : Âge du client.
- HOME_ADDRESS (string) : Adresse du domicile du client.
- ZIP_CODE (integer) : Code postal de l'adresse du client.
- CITY (string) : Ville de résidence du client.
- STATE (string) : État ou région de résidence du client.
- COUNTRY (string) : Pays de résidence du client.
- DATE_NAISSANCE (date) : Date de naissance du client.

Table : erp_orders

- order_id (integer) : Identifiant unique de la commande.
- order_date (date) : Date à laquelle la commande a été passée.
- expected_delivery_date (date) : Date à laquelle la livraison est prévue.
- actual_delivery_date (date) : Date réelle de la livraison.
- delivery_status (string) : Statut de la livraison (par exemple, en cours, livrée, retardée).
- delivery_cost (decimal) : Coût de la livraison en euros.

Table : external_data

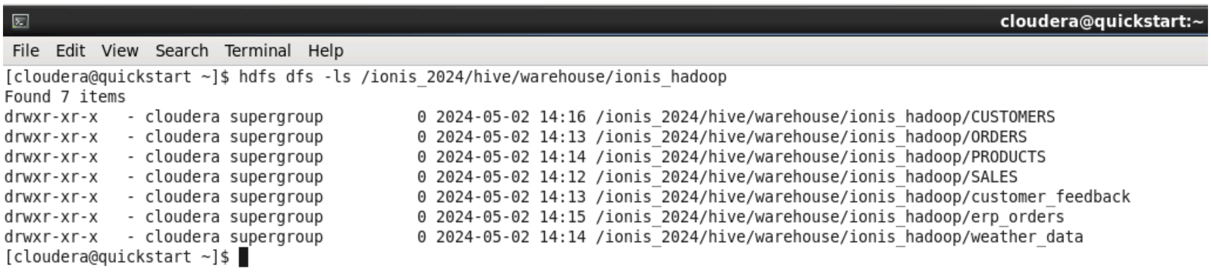
- data_id (integer) : Identifiant unique des données externes.
- date (date) : Date des données collectées.
- weather_condition (string) : Description des conditions météorologiques.
- traffic_intensity (string) : Description de l'intensité du trafic.

Table : customer_feedback

- feedback_id (integer) : Identifiant unique du retour client.
- order_id (integer) : Identifiant de la commande associée à ce retour.
- satisfaction_rating (integer) : Note de satisfaction du client (échelle de 1 à 5).
- comment (string) : Commentaire du client sur la commande.

Importation de la base de données sur HDFS

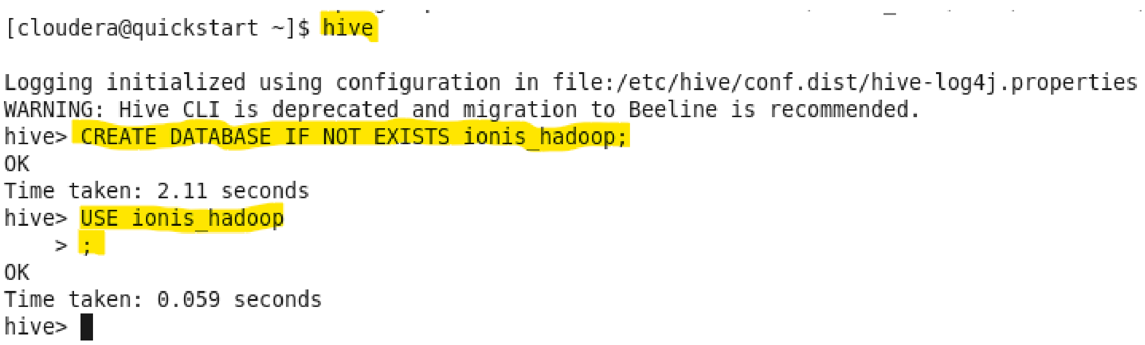
sqoop import-all-tables --connect jdbc:mysql://srv1048.hstgr.io/u682049460_ionis_hadoop --username u682049460_ionis --password “mB4U|H?j” --warehouse-dir /ionis_2024/hive/warehouse/ionis_hadoop -m 4



```
[cloudera@quickstart ~]$ hdfs dfs -ls /ionis_2024/hive/warehouse/ionis_hadoop
Found 7 items
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:16 /ionis_2024/hive/warehouse/ionis_hadoop/CUSTOMERS
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:13 /ionis_2024/hive/warehouse/ionis_hadoop/ORDERS
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:14 /ionis_2024/hive/warehouse/ionis_hadoop/PRODUCTS
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:12 /ionis_2024/hive/warehouse/ionis_hadoop/SALES
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:13 /ionis_2024/hive/warehouse/ionis_hadoop/customer_feedback
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:15 /ionis_2024/hive/warehouse/ionis_hadoop/erp_orders
drwxr-xr-x   - cloudera supergroup          0 2024-05-02 14:14 /ionis_2024/hive/warehouse/ionis_hadoop/weather_data
[cloudera@quickstart ~]$ 
```

Script Hive pour créer les tables au format Textfile

Initialisation de la base de données dans hive

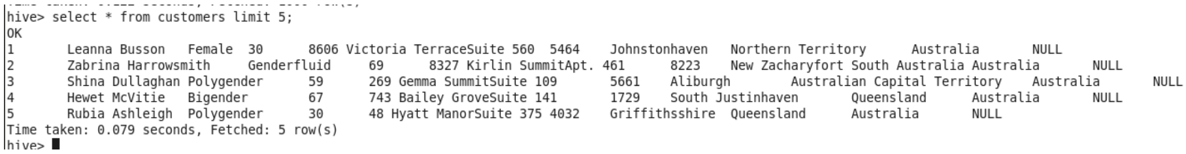


```
[cloudera@quickstart ~]$ hive

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 IF NOT EXISTS ionis_hadoop;
OK
Time taken: 2.11 seconds
hive> USE ionis_hadoop
    > ;
OK
Time taken: 0.059 seconds
hive> 
```

-- Création de la table customers venant la base de données e-commerces

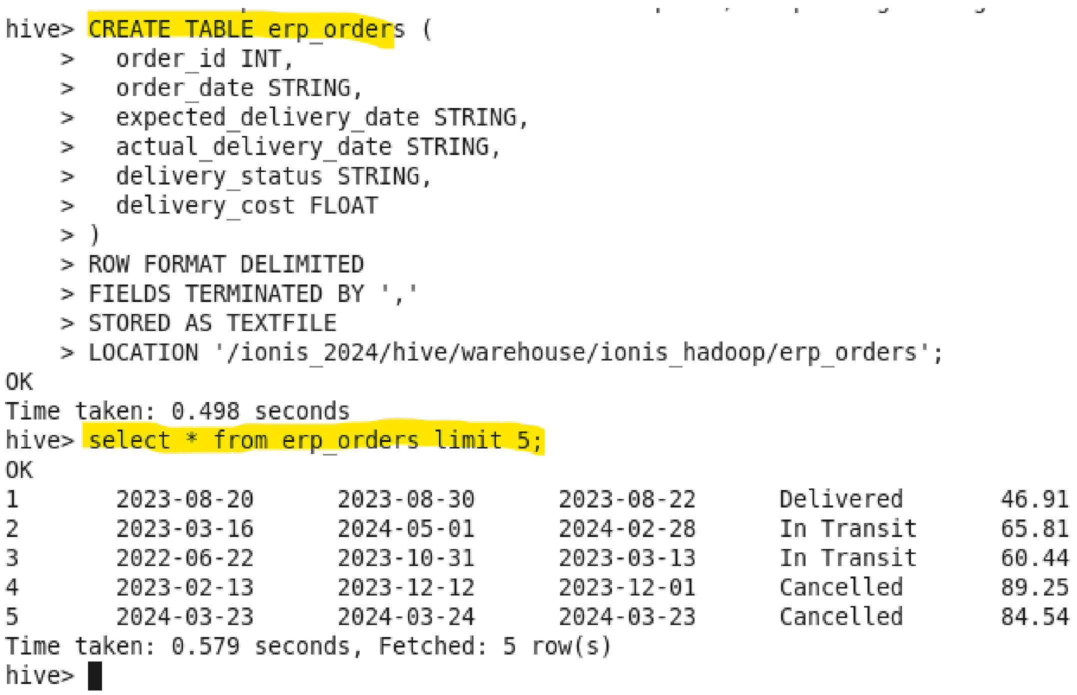
```
CREATE TABLE CUSTOMERS (
  CUSTOMER_ID int,
  CUSTOMER_NAME string,
  GENDER string,
  AGE int,
  HOME_ADDRESS string,
  ZIP_CODE int,
  CITY string,
  STATE string,
  COUNTRY string,
  DATE_NAISSANCE date
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/CUSTOMERS';
```



```
hive> select * from customers limit 5;
OK
1       Leanna Busson   Female  30      8606 Victoria TerraceSuite 560  5464    Johnstonhaven   Northern Territory      Australia       NULL
2       Zabrina Harrowsmith     Genderfluid     69      8327 Kirlin SummitApt. 461      8223    New Zacharyfort South Australia Australia       NULL
3       Shina Dullaghan Polygender      59      269 Gemma SummitSuite 109       5661    Aliburgh        Australian Capital Territory    Australia       NULL
4       Hewet McVitie   Bigender        67      743 Bailey GroveSuite 141       1729    South Justinhaven       Queensland      Australia       NULL
5       Rubia Ashleigh  Polygender      30      48 Hyatt ManorSuite 375 4032    Griffithsshire  Queensland      Australia       NULL
Time taken: 0.079 seconds, Fetched: 5 row(s)
hive> █
```

-- Création de la table pour la base de données interne (ERP)

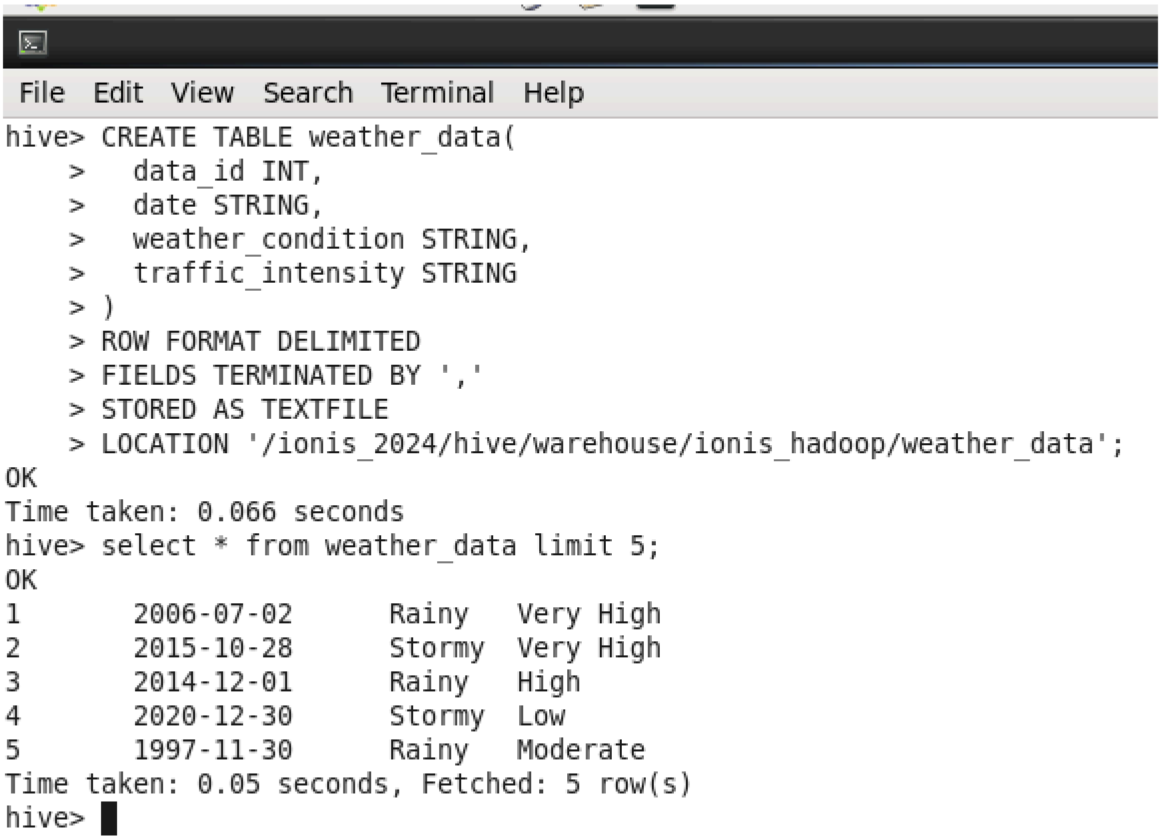
```
CREATE TABLE erp_orders (
  order_id INT,
  order_date STRING,
  expected_delivery_date STRING,
  actual_delivery_date STRING,
  delivery_status STRING,
  delivery_cost FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/erp_orders';
```



```
hive> CREATE TABLE erp_orders (
    >   order_id INT,
    >   order_date STRING,
    >   expected_delivery_date STRING,
    >   actual_delivery_date STRING,
    >   delivery_status STRING,
    >   delivery_cost FLOAT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/erp_orders';
OK
Time taken: 0.498 seconds
hive> select * from erp_orders limit 5;
OK
1       2023-08-20      2023-08-30      2023-08-22      Delivered       46.91
2       2023-03-16      2024-05-01      2024-02-28      In Transit      65.81
3       2022-06-22      2023-10-31      2023-03-13      In Transit      60.44
4       2023-02-13      2023-12-12      2023-12-01      Cancelled       89.25
5       2024-03-23      2024-03-24      2024-03-23      Cancelled       84.54
Time taken: 0.579 seconds, Fetched: 5 row(s)
hive> 
```

-- Création de la table pour les données externes (Conditions météorologiques et trafic)

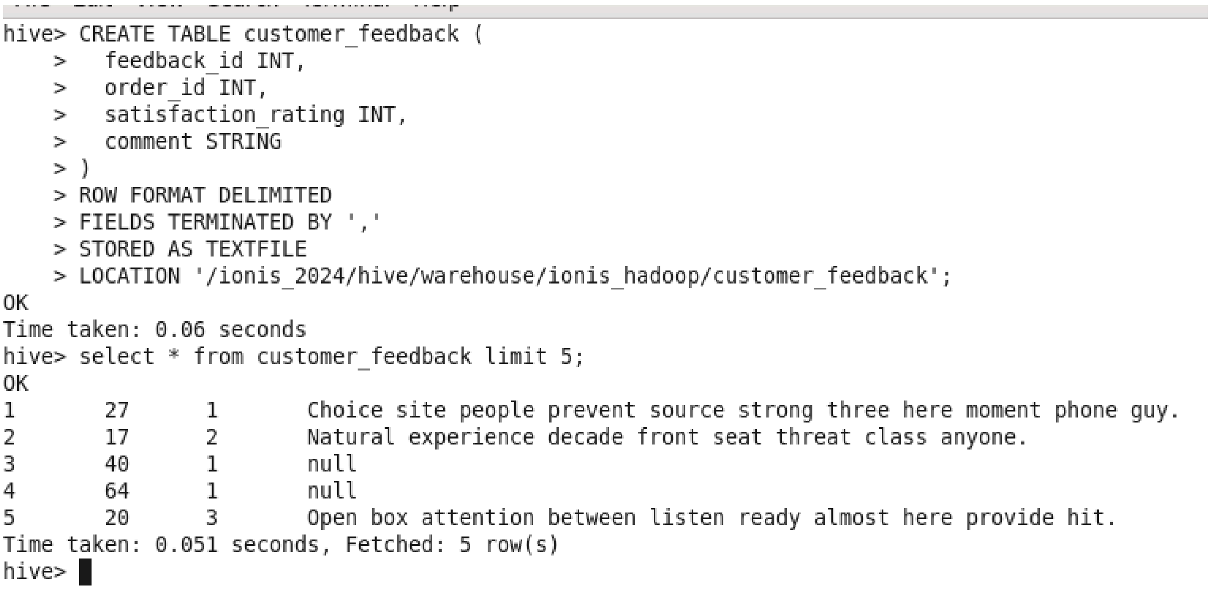
```
CREATE TABLE weather_data(
  data_id INT,
  date STRING,
  weather_condition STRING,
  traffic_intensity STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/weather_data';
```



```
hive> CREATE TABLE weather_data(
    >   data_id INT,
    >   date STRING,
    >   weather_condition STRING,
    >   traffic_intensity STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/weather_data';
OK
Time taken: 0.066 seconds
hive> select * from weather_data limit 5;
OK
1       2006-07-02      Rainy   Very High
2       2015-10-28      Stormy  Very High
3       2014-12-01      Rainy   High
4       2020-12-30      Stormy  Low
5       1997-11-30      Rainy   Moderate
Time taken: 0.05 seconds, Fetched: 5 row(s)
hive> 
```

-- Création de la table pour le feedback des clients

```
CREATE TABLE customer_feedback (
  feedback_id INT,
  order_id INT,
  satisfaction_rating INT,
  comment STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/customer_feedback';
```

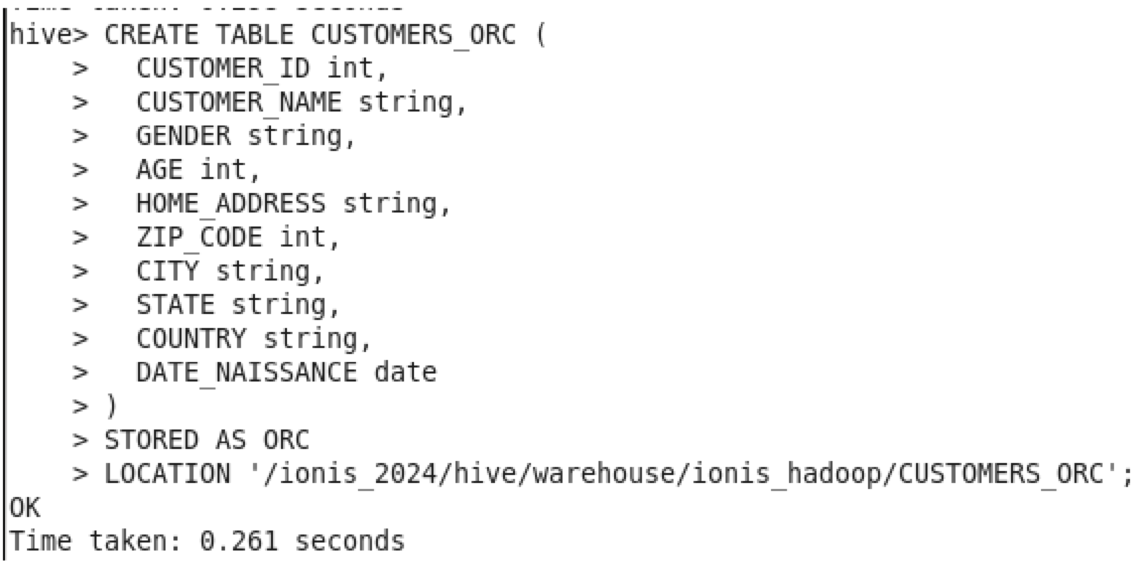


```
hive> CREATE TABLE customer_feedback (
    >   feedback_id INT,
    >   order_id INT,
    >   satisfaction_rating INT,
    >   comment STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/customer_feedback';
OK
Time taken: 0.06 seconds
hive> select * from customer_feedback limit 5;
OK
1       27      1       Choice site people prevent source strong three here moment phone guy.
2       17      2       Natural experience decade front seat threat class anyone.
3       40      1       null
4       64      1       null
5       20      3       Open box attention between listen ready almost here provide hit.
Time taken: 0.051 seconds, Fetched: 5 row(s)
hive> 
```

Script Hive pour créer les tables au format ORC

-- Création de la table customers venant la base de données e-commerces au format ORC

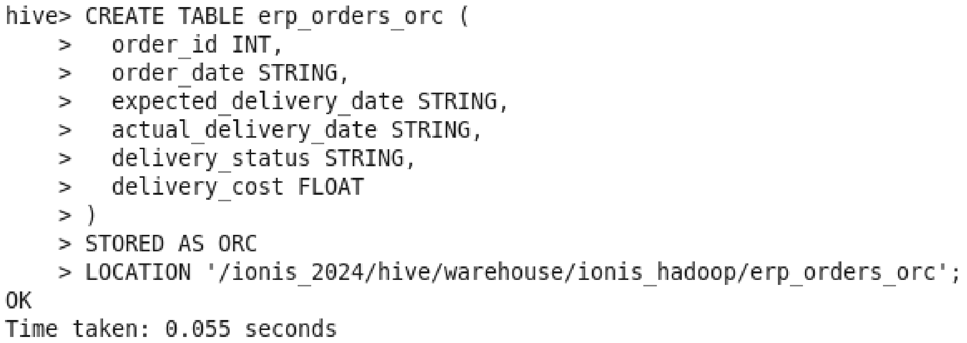
```
CREATE TABLE CUSTOMERS_ORC (
  CUSTOMER_ID int,
  CUSTOMER_NAME string,
  GENDER string,
  AGE int,
  HOME_ADDRESS string,
  ZIP_CODE int,
  CITY string,
  STATE string,
  COUNTRY string,
  DATE_NAISSANCE date
)
STORED AS ORC
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/CUSTOMERS_ORC';
```



```
hive> CREATE TABLE CUSTOMERS_ORC (
    >   CUSTOMER_ID int,
    >   CUSTOMER_NAME string,
    >   GENDER string,
    >   AGE int,
    >   HOME_ADDRESS string,
    >   ZIP_CODE int,
    >   CITY string,
    >   STATE string,
    >   COUNTRY string,
    >   DATE_NAISSANCE date
    > )
    > STORED AS ORC
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/CUSTOMERS_ORC';
OK
Time taken: 0.261 seconds
```

-- Création de la table pour la base de données interne (ERP) au format ORC

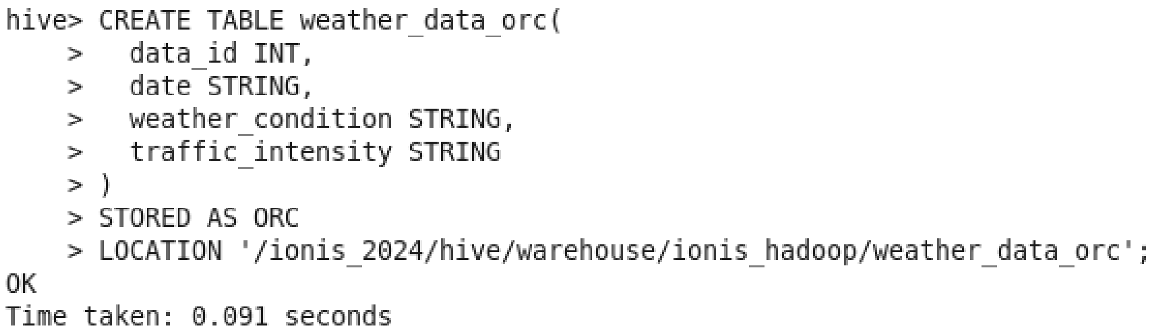
```
CREATE TABLE erp_orders_orc (
  order_id INT,
  order_date STRING,
  expected_delivery_date STRING,
  actual_delivery_date STRING,
  delivery_status STRING,
  delivery_cost FLOAT
)
STORED AS ORC
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/erp_orders_orc';
```



```
hive> CREATE TABLE erp_orders_orc (
    >   order_id INT,
    >   order_date STRING,
    >   expected_delivery_date STRING,
    >   actual_delivery_date STRING,
    >   delivery_status STRING,
    >   delivery_cost FLOAT
    > )
    > STORED AS ORC
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/erp_orders_orc';
OK
Time taken: 0.055 seconds
```

-- Création de la table pour les données externes (Conditions météorologiques et trafic) au format ORC

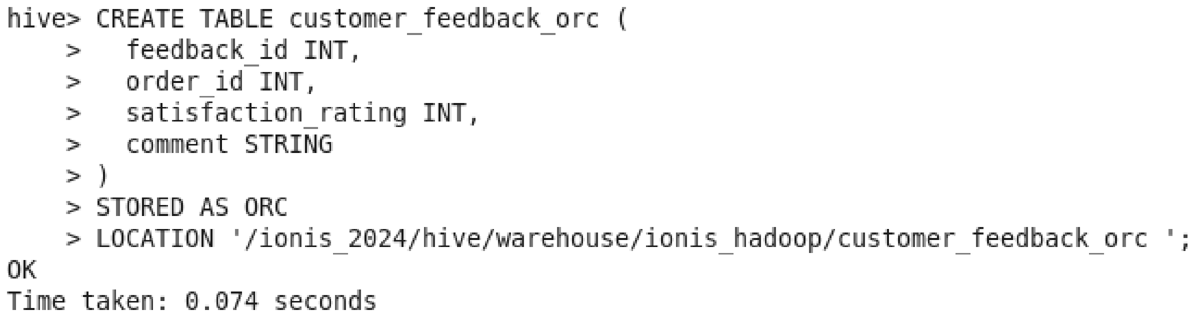
```
CREATE TABLE weather_data_orc (
  data_id INT,
  date STRING,
  weather_condition STRING,
 traffic_intensity STRING
)
STORED AS ORC
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/weather_data_orc';
```



```
hive> CREATE TABLE weather_data_orc(
    >   data_id INT,
    >   date STRING,
    >   weather_condition STRING,
    >   traffic_intensity STRING
    > )
    > STORED AS ORC
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/weather_data_orc';
OK
Time taken: 0.091 seconds
```

-- Création de la table pour le feedback des clients au format ORC

```
CREATE TABLE customer_feedback_orc (
 feedback_id INT,
 order_id INT,
 satisfaction_rating INT,
 comment STRING
)
STORED AS ORC
LOCATION
'/ionis_2024/hive/warehouse/ionis_hadoop/customer_feedback_orc';
```

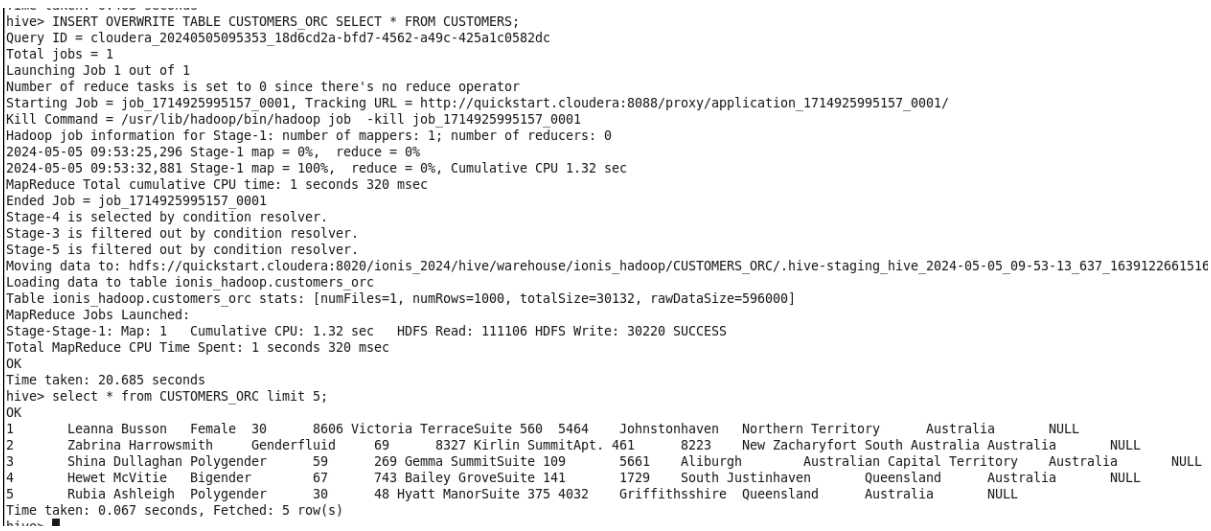


```
hive> CREATE TABLE customer_feedback_orc (
    >   feedback_id INT,
    >   order_id INT,
    >   satisfaction_rating INT,
    >   comment STRING
    > )
    > STORED AS ORC
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/customer_feedback_orc ';
OK
Time taken: 0.074 seconds
```

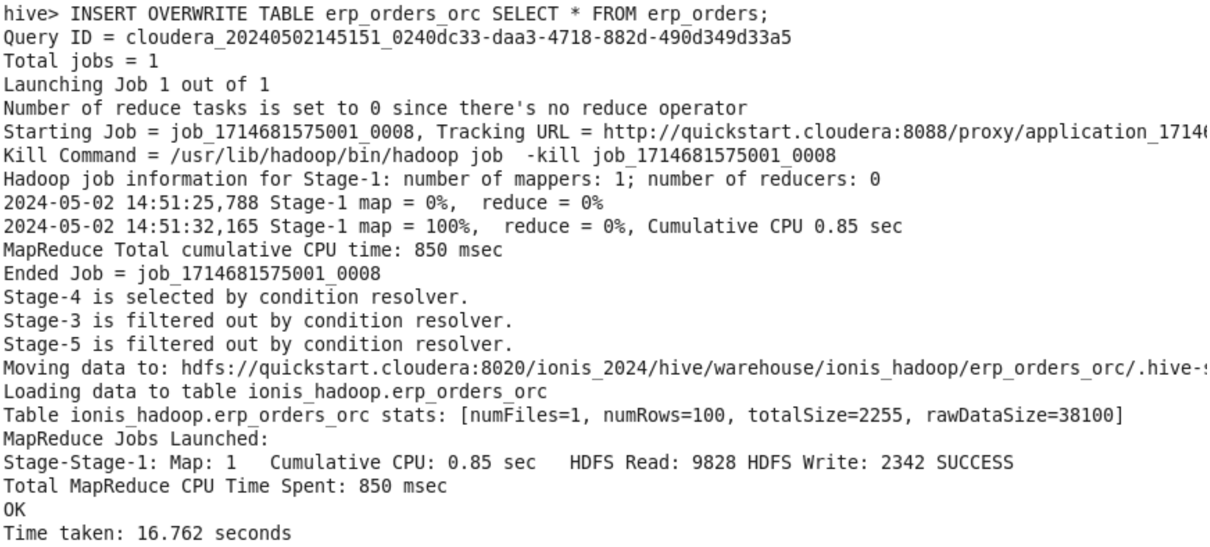
Script Hive pour transférer des données entre les tables TXT et ORC

```
INSERT OVERWRITE TABLE CUSTOMERS_ORC SELECT * FROM
CUSTOMERS;
INSERT OVERWRITE TABLE erp_orders_orc SELECT * FROM erp_orders;
INSERT OVERWRITE TABLE customer_feedback_orc SELECT * FROM
customer_feedback;
```

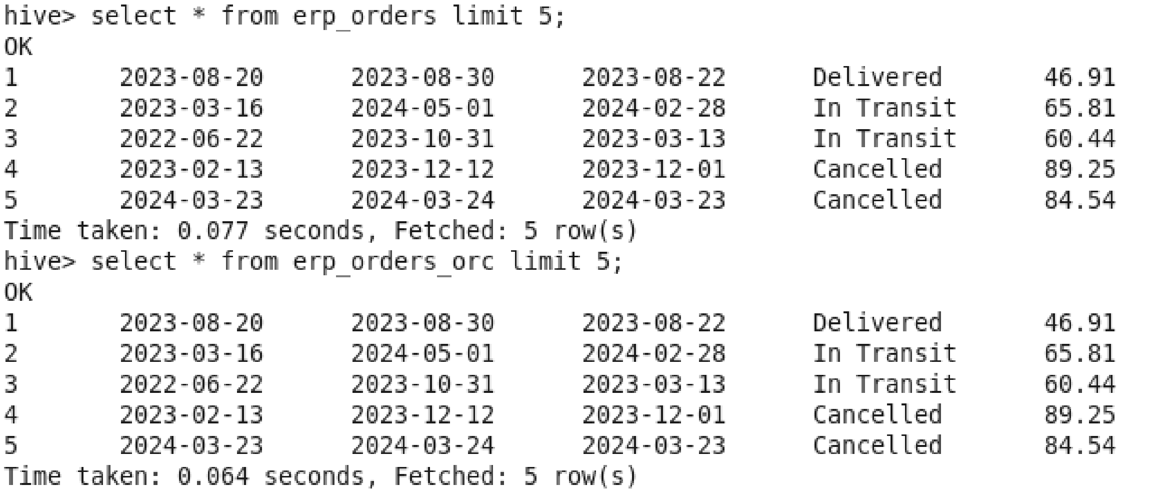
INSERT OVERWRITE TABLE weather_data_orc SELECT * FROM weather_data;



```
hive> INSERT OVERWRITE TABLE CUSTOMERS_ORC SELECT * FROM CUSTOMERS;
Query ID = cloudera_20240505095353_18d6cd2a-bfd7-4562-a49c-425a1c0582dc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1714925995157_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1714925995157_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1714925995157_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-05-05 09:53:25,296 Stage-1 map = 0%,  reduce = 0%
2024-05-05 09:53:32,881 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.32 sec
MapReduce Total cumulative CPU time: 1 seconds 320 msec
Ended Job = job_1714925995157_0001
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/ionis_2024/hive/warehouse/ionis_hadoop/CUSTOMERS_ORC/.hive-staging_hive_2024-05-05_09-53-13_637_1639122661516
Loading data to table ionis_hadoop.customers_orc
Table ionis_hadoop.customers_orc stats: [numFiles=1, numRows=1000, totalSize=30132, rawDataSize=596000]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.32 sec   HDFS Read: 111106 HDFS Write: 30220 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 320 msec
OK
Time taken: 20.685 seconds
hive> select * from CUSTOMERS_ORC limit 5;
OK
1       Leanna Busson   Female  30      8606 Victoria TerraceSuite 560  5464    Johnstonhaven   Northern Territory      Australia       NULL
2       Zabrina Harrowsmith     Genderfluid     69      8327 Kirlin SummitApt. 461      8223    New Zacharyfort South Australia Australia       NULL
3       Shina Dullaghan Polygender      59      269 Gemma SummitSuite 109       5661    Aliburgh        Australian Capital Territory    Australia       NULL
4       Hewet McVitie   Bigender        67      743 Bailey GroveSuite 141       1729    South Justinhaven       Queensland      Australia       NULL
5       Rubia Ashleigh  Polygender      30      48 Hyatt ManorSuite 375 4032    Griffithsshire  Queensland      Australia       NULL
Time taken: 0.067 seconds, Fetched: 5 row(s)
hive> █
```



```
hive> INSERT OVERWRITE TABLE erp_orders_orc SELECT * FROM erp_orders;
Query ID = cloudera_20240502145151_0240dc33-daa3-4718-882d-490d349d33a5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1714681575001_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_17146
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1714681575001_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-05-02 14:51:25,788 Stage-1 map = 0%,  reduce = 0%
2024-05-02 14:51:32,165 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.85 sec
MapReduce Total cumulative CPU time: 850 msec
Ended Job = job_1714681575001_0008
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/ionis_2024/hive/warehouse/ionis_hadoop/erp_orders_orc/.hive-
Loading data to table ionis_hadoop.erp_orders_orc
Table ionis_hadoop.erp_orders_orc stats: [numFiles=1, numRows=100, totalSize=2255, rawDataSize=38100]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 0.85 sec   HDFS Read: 9828 HDFS Write: 2342 SUCCESS
Total MapReduce CPU Time Spent: 850 msec
OK
Time taken: 16.762 seconds
```



```
hive> select * from erp_orders limit 5;
OK
1       2023-08-20      2023-08-30      2023-08-22      Delivered       46.91
2       2023-03-16      2024-05-01      2024-02-28      In Transit      65.81
3       2022-06-22      2023-10-31      2023-03-13      In Transit      60.44
4       2023-02-13      2023-12-12      2023-12-01      Cancelled       89.25
5       2024-03-23      2024-03-24      2024-03-23      Cancelled       84.54
Time taken: 0.077 seconds, Fetched: 5 row(s)
hive> select * from erp_orders_orc limit 5;
OK
1       2023-08-20      2023-08-30      2023-08-22      Delivered       46.91
2       2023-03-16      2024-05-01      2024-02-28      In Transit      65.81
3       2022-06-22      2023-10-31      2023-03-13      In Transit      60.44
4       2023-02-13      2023-12-12      2023-12-01      Cancelled       89.25
5       2024-03-23      2024-03-24      2024-03-23      Cancelled       84.54
Time taken: 0.064 seconds, Fetched: 5 row(s)
```

II. Enrichissement des données

Pour enrichir les données en croisant les différentes sources mentionnées, nous pouvons créer une nouvelle table qui regroupe des informations pertinentes de chaque source. Cela nous permettra d'avoir une vue consolidée qui peut être utile pour des analyses plus complexes, telles que l'impact des conditions météorologiques sur les livraisons d'ERP et la satisfaction client. Voici une proposition de schéma pour cette nouvelle table, ainsi que le code SQL pour créer cette table et l'interrogation correspondante pour peupler la table avec les données des tables existantes.

Nouveau Schéma de Table

Nous pouvons créer une table nommée **enhanced_order_insights_orc** qui inclut les éléments suivants :

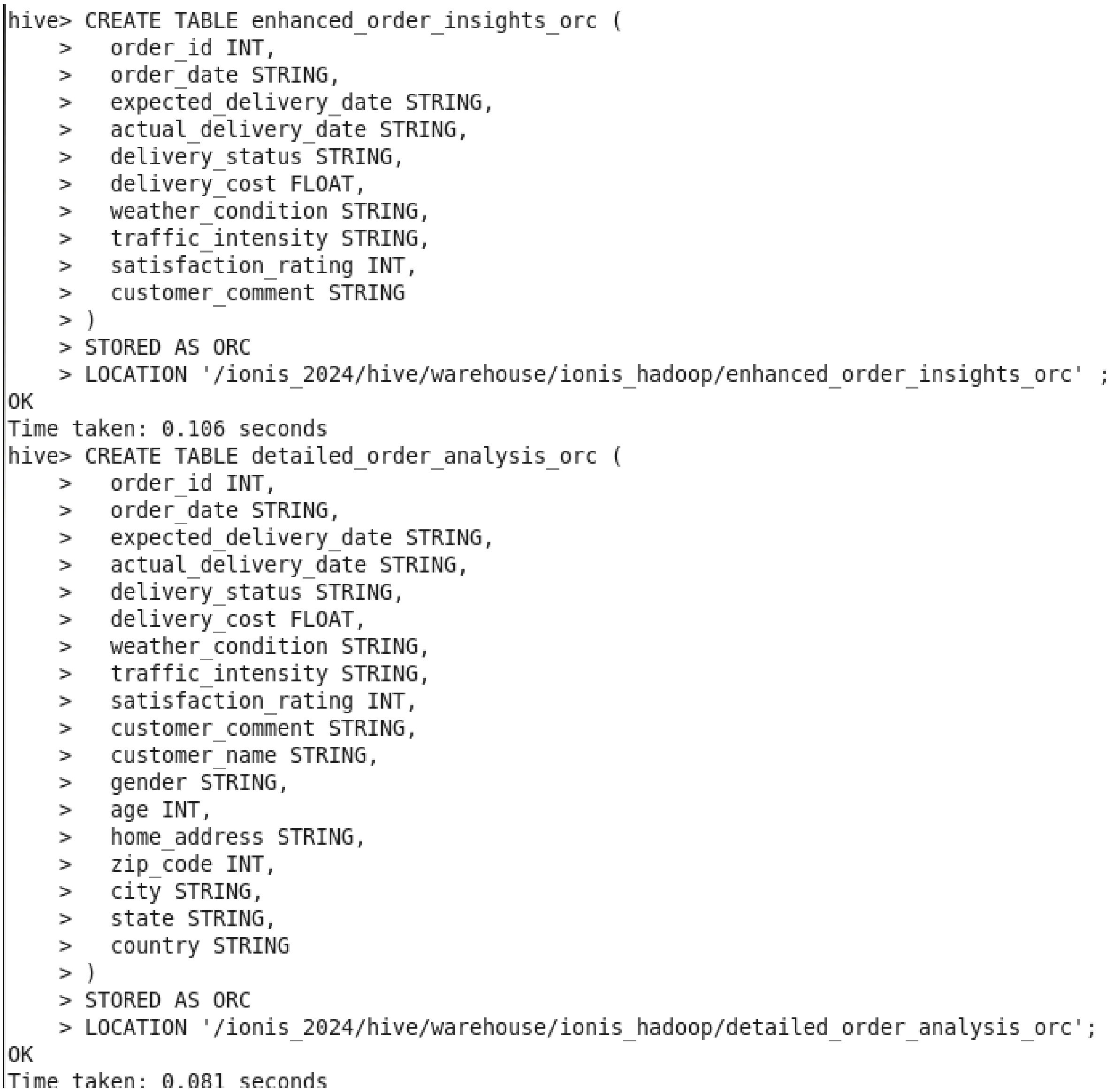
- Informations sur la commande (ID, dates, statut)
- Coût de livraison
- Condition météorologique le jour de la commande
- Intensité du trafic le jour de la commande
- Évaluation de la satisfaction du client et commentaire

```
CREATE TABLE enhanced_order_insights_orc (
  order_id INT,
  order_date STRING,
  expected_delivery_date STRING,
  actual_delivery_date STRING,
  delivery_status STRING,
  delivery_cost FLOAT,
  weather_condition STRING,
  traffic_intensity STRING,
  satisfaction_rating INT,
  customer_comment STRING
)
STORED AS ORC
LOCATION
'/ionis_2024/hive/warehouse/ionis_hadoop/enhanced_order_insights_orc' ;
```

Avec l'introduction de la nouvelle source de données sur les clients (CUSTOMERS_ORC), nous pouvons encore enrichir le schéma précédent en y intégrant des informations client pertinentes.

```
CREATE TABLE detailed_order_analysis_orc (
  order_id INT,
  order_date STRING,
  expected_delivery_date STRING,
  actual_delivery_date STRING,
  delivery_status STRING,
  delivery_cost FLOAT,
  weather_condition STRING,
  traffic_intensity STRING,
  satisfaction_rating INT,
  customer_comment STRING,
  customer_name STRING,
  gender STRING,
  age INT,
  home_address STRING,
  zip_code INT,
  city STRING,
  state STRING,
  country STRING
)
STORED AS ORC
LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/detailed_order_analysis_orc';
```

Ce schéma amélioré detailed_order_analysis_orc fournit une vue complète des interactions entre l'entreprise et le client, incluant les aspects logistiques et feedbacks associés à des données démographiques spécifiques, ce qui est crucial pour des analyses de marché approfondies.



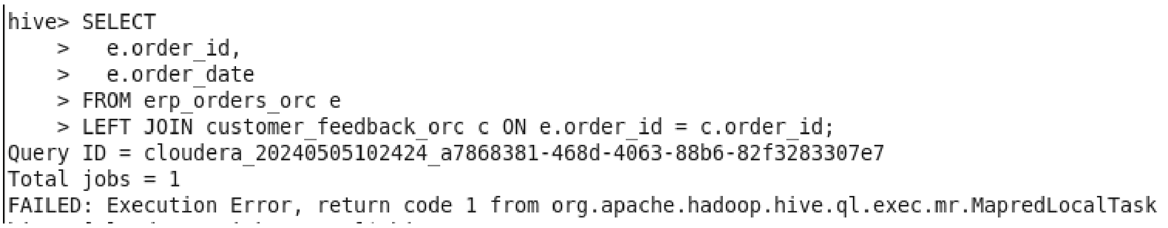
```
hive> CREATE TABLE enhanced_order_insights_orc (
    >   order_id INT,
    >   order_date STRING,
    >   expected_delivery_date STRING,
    >   actual_delivery_date STRING,
    >   delivery_status STRING,
    >   delivery_cost FLOAT,
    >   weather_condition STRING,
    >   traffic_intensity STRING,
    >   satisfaction_rating INT,
    >   customer_comment STRING
    > )
    > STORED AS ORC
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/enhanced_order_insights_orc' ;
OK
Time taken: 0.106 seconds
hive> CREATE TABLE detailed_order_analysis_orc (
    >   order_id INT,
    >   order_date STRING,
    >   expected_delivery_date STRING,
    >   actual_delivery_date STRING,
    >   delivery_status STRING,
    >   delivery_cost FLOAT,
    >   weather_condition STRING,
    >   traffic_intensity STRING,
    >   satisfaction_rating INT,
    >   customer_comment STRING,
    >   customer_name STRING,
    >   gender STRING,
    >   age INT,
    >   home_address STRING,
    >   zip_code INT,
    >   city STRING,
    >   state STRING,
    >   country STRING
    > )
    > STORED AS ORC
    > LOCATION '/ionis_2024/hive/warehouse/ionis_hadoop/detailed_order_analysis_orc';
OK
Time taken: 0.081 seconds
```

SQL pour Peupler nos Tables

```
INSERT INTO enhanced_order_insights_orc
SELECT
  e.order_id,
  e.order_date,
  e.expected_delivery_date,
  e.actual_delivery_date,
  e.delivery_status,
  e.delivery_cost,
  w.weather_condition,
  w.traffic_intensity,
  c.satisfaction_rating,
  c.comment AS customer_comment
FROM erp_orders_orc e
LEFT JOIN weather_data_orc w ON e.order_date = w.date
LEFT JOIN customer_feedback_orc c ON e.order_id = c.order_id;
```

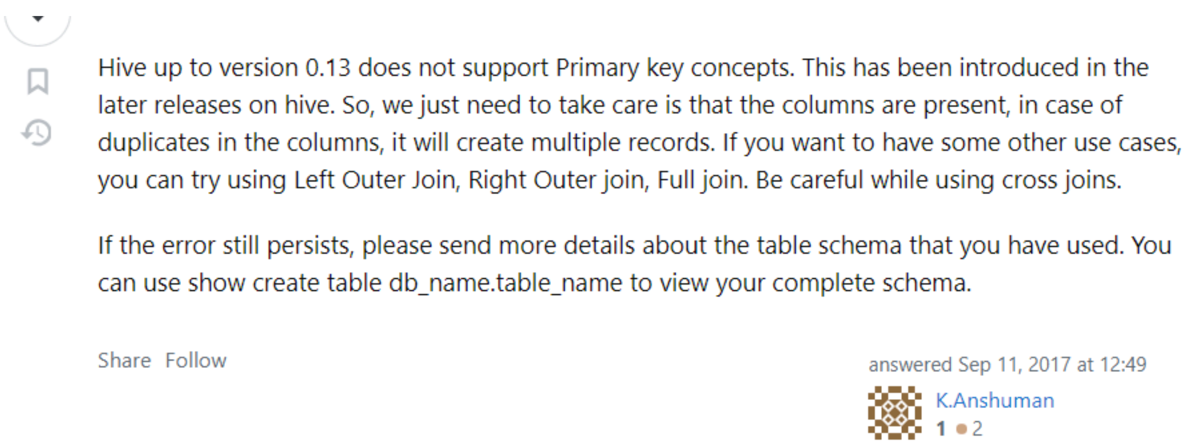
```
INSERT INTO detailed_order_analysis_orc
SELECT
  e.order_id,
  e.order_date,
  e.expected_delivery_date,
  e.actual_delivery_date,
  e.delivery_status,
  e.delivery_cost,
  w.weather_condition,
  w.traffic_intensity,
  c.satisfaction_rating,
  c.comment AS customer_comment,
  cu.customer_name,
  cu.gender,
  cu.age,
  cu.home_address,
  cu.zip_code,
  cu.city,
  cu.state,
  cu.country
FROM erp_orders_orc e
LEFT JOIN weather_data_orc w ON e.order_date = w.date
LEFT JOIN customer_feedback_orc c ON e.order_id = c.order_id
LEFT JOIN CUSTOMERS_ORC cu ON e.customer_id = cu.customer_id;
```

Premier problème avec les jointures



```
hive> SELECT
    >   e.order_id,
    >   e.order_date
    > FROM erp_orders_orc e
    > LEFT JOIN customer_feedback_orc c ON e.order_id = c.order_id;
Query ID = cloudera_20240505102424_a7868381-468d-4063-88b6-82f3283307e7
Total jobs = 1
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask
```

Recherche de la solution

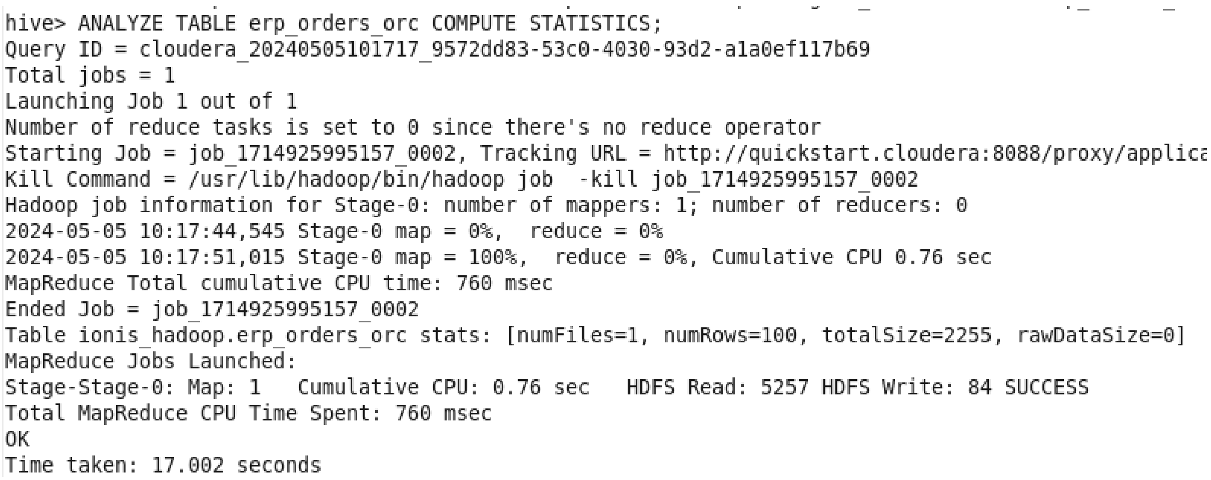


Hive up to version 0.13 does not support Primary key concepts. This has been introduced in the later releases on hive. So, we just need to take care is that the columns are present, in case of duplicates in the columns, it will create multiple records. If you want to have some other use cases, you can try using Left Outer Join, Right Outer join, Full join. Be careful while using cross joins.

If the error still persists, please send more details about the table schema that you have used. You can use show create table db_name.table_name to view your complete schema.

Pour Analyser les Statistiques de la Table

ANALYZE TABLE erp_orders_orc COMPUTE STATISTICS;
ANALYZE TABLE erp_orders_orc COMPUTE STATISTICS FOR COLUMNS order_date, delivery_status;

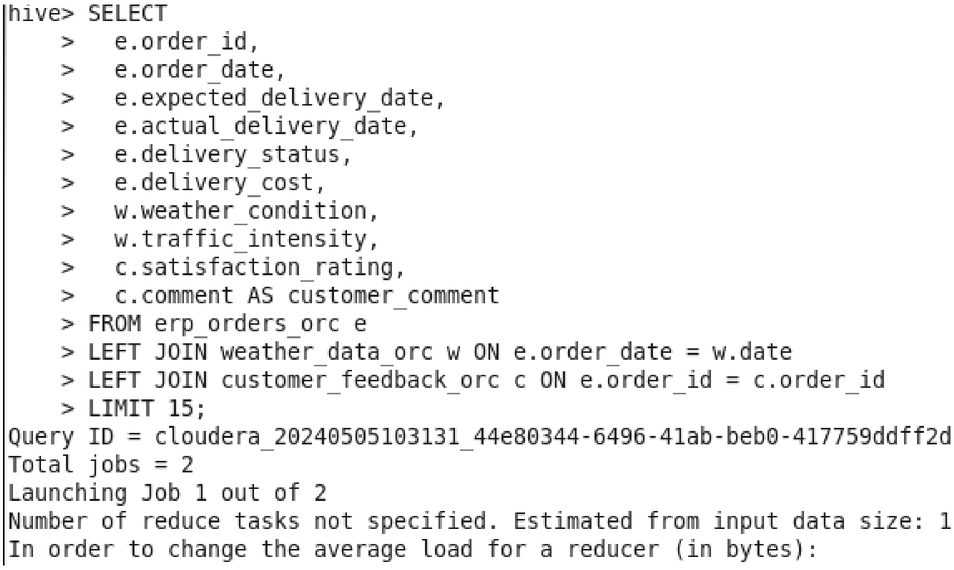


```
hive> ANALYZE TABLE erp_orders_orc COMPUTE STATISTICS;
Query ID = cloudera_20240505101717_9572dd83-53c0-4030-93d2-a1a0ef117b69
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1714925995157_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/applica
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1714925995157_0002
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2024-05-05 10:17:44,545 Stage-0 map = 0%,  reduce = 0%
2024-05-05 10:17:51,015 Stage-0 map = 100%,  reduce = 0%, Cumulative CPU 0.76 sec
MapReduce Total cumulative CPU time: 760 msec
Ended Job = job_1714925995157_0002
Table ionis_hadoop.erp_orders_orc stats: [numFiles=1, numRows=100, totalSize=2255, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1   Cumulative CPU: 0.76 sec   HDFS Read: 5257 HDFS Write: 84 SUCCESS
Total MapReduce CPU Time Spent: 760 msec
OK
Time taken: 17.002 seconds
```

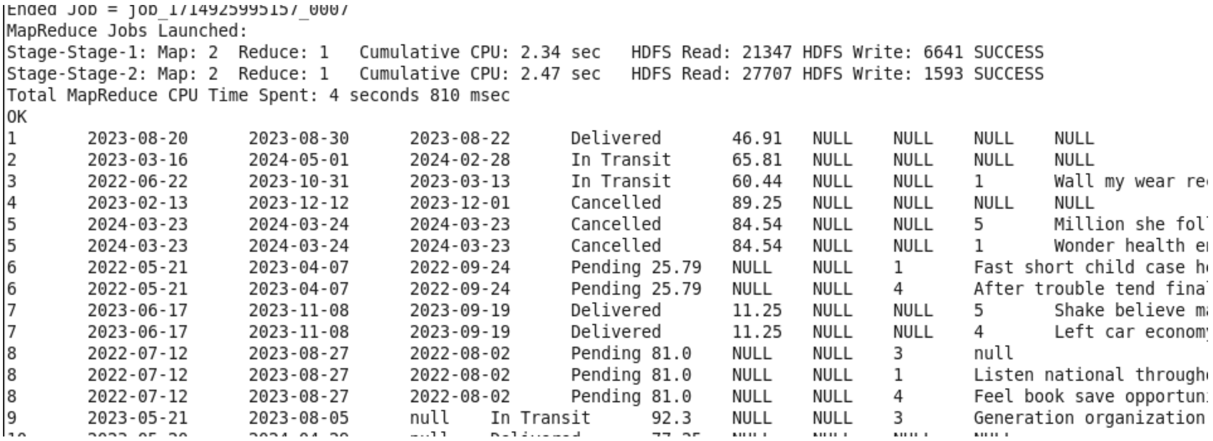
Solution aux problèmes de jointure

set hive.auto.convert.join=false;

Désactiver le paramètre hive.auto.convert.join=false empêche Hive de convertir automatiquement les jointures en jointures exécutées côté serveur de cartes, qui utilisent la mémoire pour améliorer la performance. Cette modification peut réduire les erreurs liées aux configurations inadéquates et aux estimations incorrectes de taille de données, mais elle pourrait aussi diminuer la performance des requêtes. Pour compenser, vous pourriez devoir ajuster d'autres configurations ou optimiser vos tables différemment, par exemple, en augmentant la mémoire allouée ou en utilisant des statistiques de table à jour.

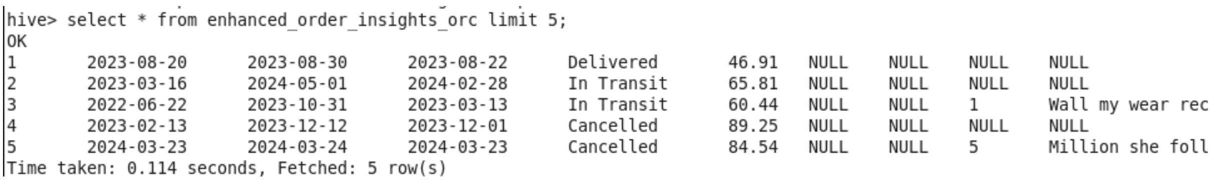


```
hive> SELECT
    >   e.order_id,
    >   e.order_date,
    >   e.expected_delivery_date,
    >   e.actual_delivery_date,
    >   e.delivery_status,
    >   e.delivery_cost,
    >   w.weather_condition,
    >   w.traffic_intensity,
    >   c.satisfaction_rating,
    >   c.comment AS customer_comment
    > FROM erp_orders_orc e
    > LEFT JOIN weather_data_orc w ON e.order_date = w.date
    > LEFT JOIN customer_feedback_orc c ON e.order_id = c.order_id
    > LIMIT 15;
Query ID = cloudera_20240505103131_44e80344-6496-41ab-beb0-417759ddff2d
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
```



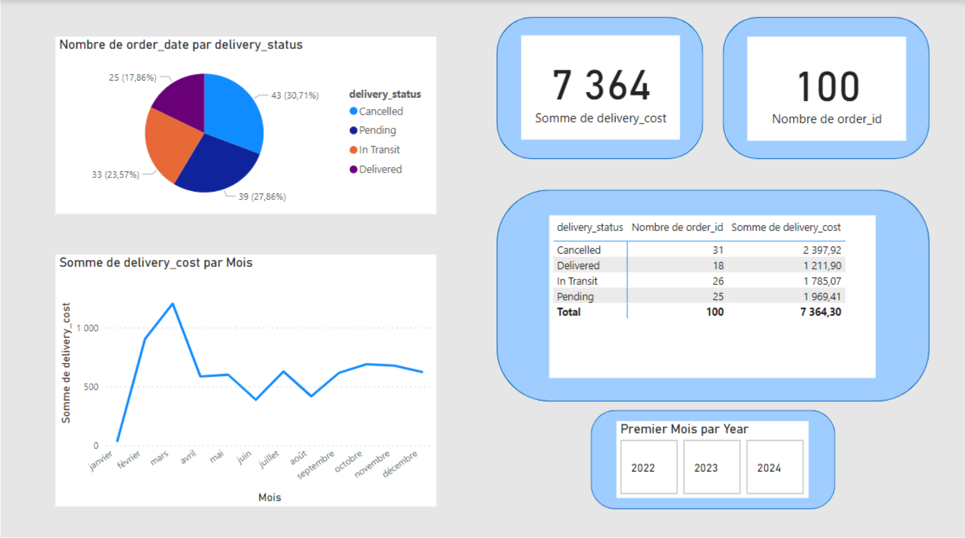
```
Ended Job = job_1714925995157_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 2.34 sec   HDFS Read: 21347 HDFS Write: 6641 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 1   Cumulative CPU: 2.47 sec   HDFS Read: 27707 HDFS Write: 1593 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 810 msec
OK
1       2023-08-20      2023-08-30      2023-08-22      Delivered       46.91   NULL    NULL    NULL    NULL
2       2023-03-16      2024-05-01      2024-02-28      In Transit      65.81   NULL    NULL    NULL    NULL
3       2022-06-22      2023-10-31      2023-03-13      In Transit      60.44   NULL    NULL    1       Wall my wear re
4       2023-02-13      2023-12-12      2023-12-01      Cancelled       89.25   NULL    NULL    NULL    NULL
5       2024-03-23      2024-03-24      2024-03-23      Cancelled       84.54   NULL    NULL    5       Million she fol
5       2024-03-23      2024-03-24      2024-03-23      Cancelled       84.54   NULL    NULL    1       Wonder health e
6       2022-05-21      2023-04-07      2022-09-24      Pending 25.79   NULL    NULL    1       Fast short child case h
6       2022-05-21      2023-04-07      2022-09-24      Pending 25.79   NULL    NULL    4       After trouble tend fina
7       2023-06-17      2023-11-08      2023-09-19      Delivered       11.25   NULL    NULL    5       Shake believe m
7       2023-06-17      2023-11-08      2023-09-19      Delivered       11.25   NULL    NULL    4       Left car econom
8       2022-07-12      2023-08-27      2022-08-02      Pending 81.0    NULL    NULL    3       null
8       2022-07-12      2023-08-27      2022-08-02      Pending 81.0    NULL    NULL    1       Listen national through
8       2022-07-12      2023-08-27      2022-08-02      Pending 81.0    NULL    NULL    4       Feel book save opportun
9       2023-05-21      2023-08-05      null    In Transit      92.3    NULL    NULL    3       Generation organization
```

Après avoir résolu le problème de jointure, nous avons pu peupler nos tables.



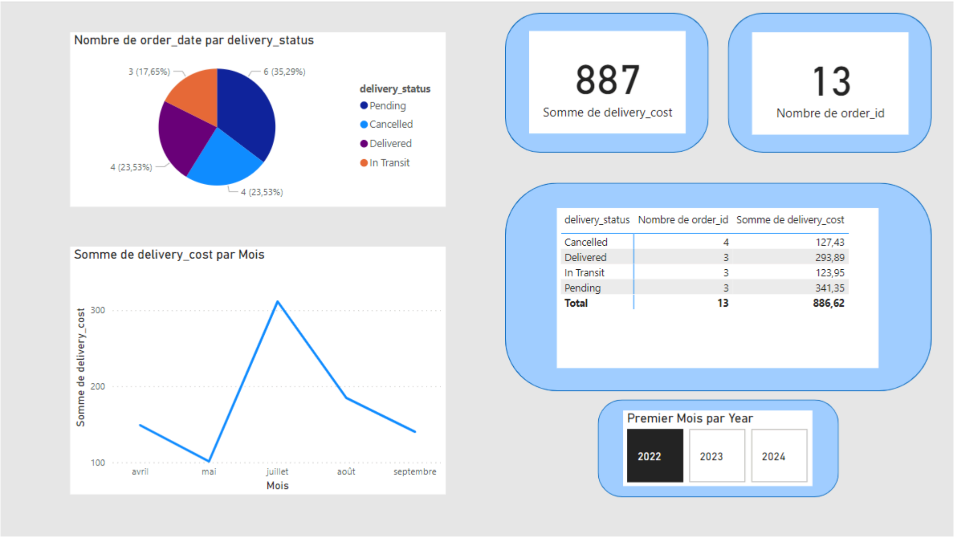
```
hive> select * from enhanced_order_insights_orc limit 5;
OK
1       2023-08-20      2023-08-30      2023-08-22      Delivered       46.91   NULL    NULL    NULL    NULL
2       2023-03-16      2024-05-01      2024-02-28      In Transit      65.81   NULL    NULL    NULL    NULL
3       2022-06-22      2023-10-31      2023-03-13      In Transit      60.44   NULL    NULL    1       Wall my wear rec
4       2023-02-13      2023-12-12      2023-12-01      Cancelled       89.25   NULL    NULL    NULL    NULL
5       2024-03-23      2024-03-24      2024-03-23      Cancelled       84.54   NULL    NULL    5       Million she foll
Time taken: 0.114 seconds, Fetched: 5 row(s)
```

III. Dashboard



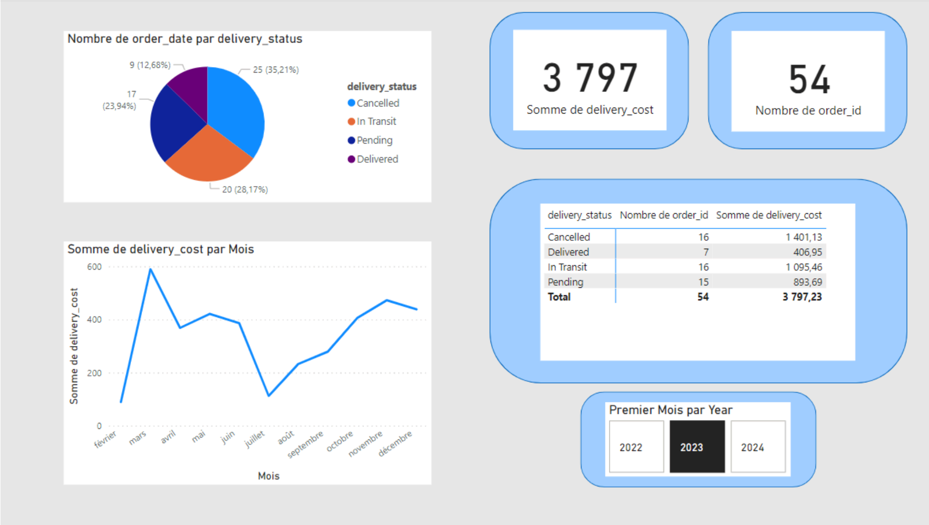
delivery_status	Nombre de order_id	Somme de delivery_cost
Cancelled	31	2 397,92
Delivered	18	1 211,90
In Transit	26	1 785,07
Pending	25	1 969,41
Total	**100**	**7 364,30**

Dashboards Global de notre base de données



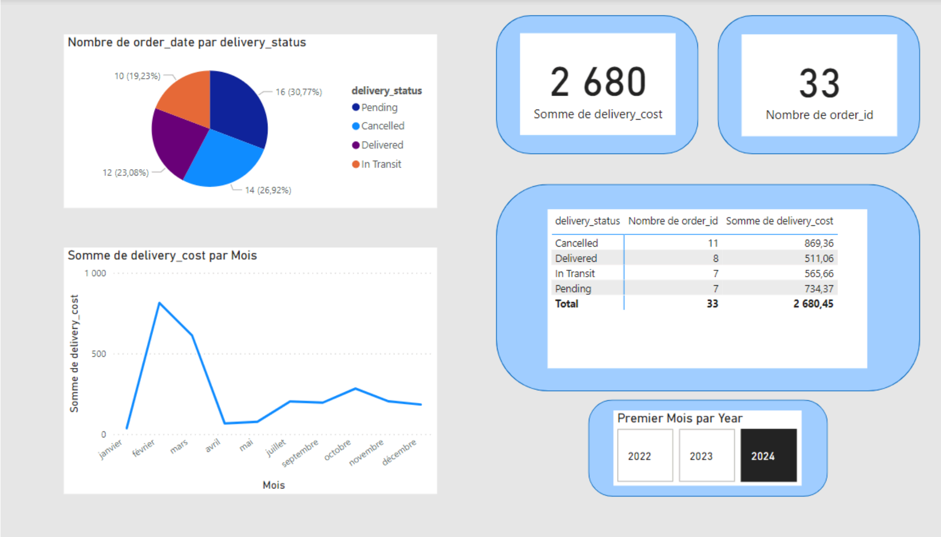
delivery_status	Nombre de order_id	Somme de delivery_cost
Cancelled	4	127,43
Delivered	3	293,89
In Transit	3	123,95
Pending	3	341,35
Total	**13**	**886,62**

Dashboards pour l’année 2022



delivery_status	Nombre de order_id	Somme de delivery_cost
Cancelled	16	1 401,13
Delivered	7	406,95
In Transit	16	1 095,46
Pending	15	893,69
Total	**54**	**3 797,23**

Dashboards pour l’année 2023

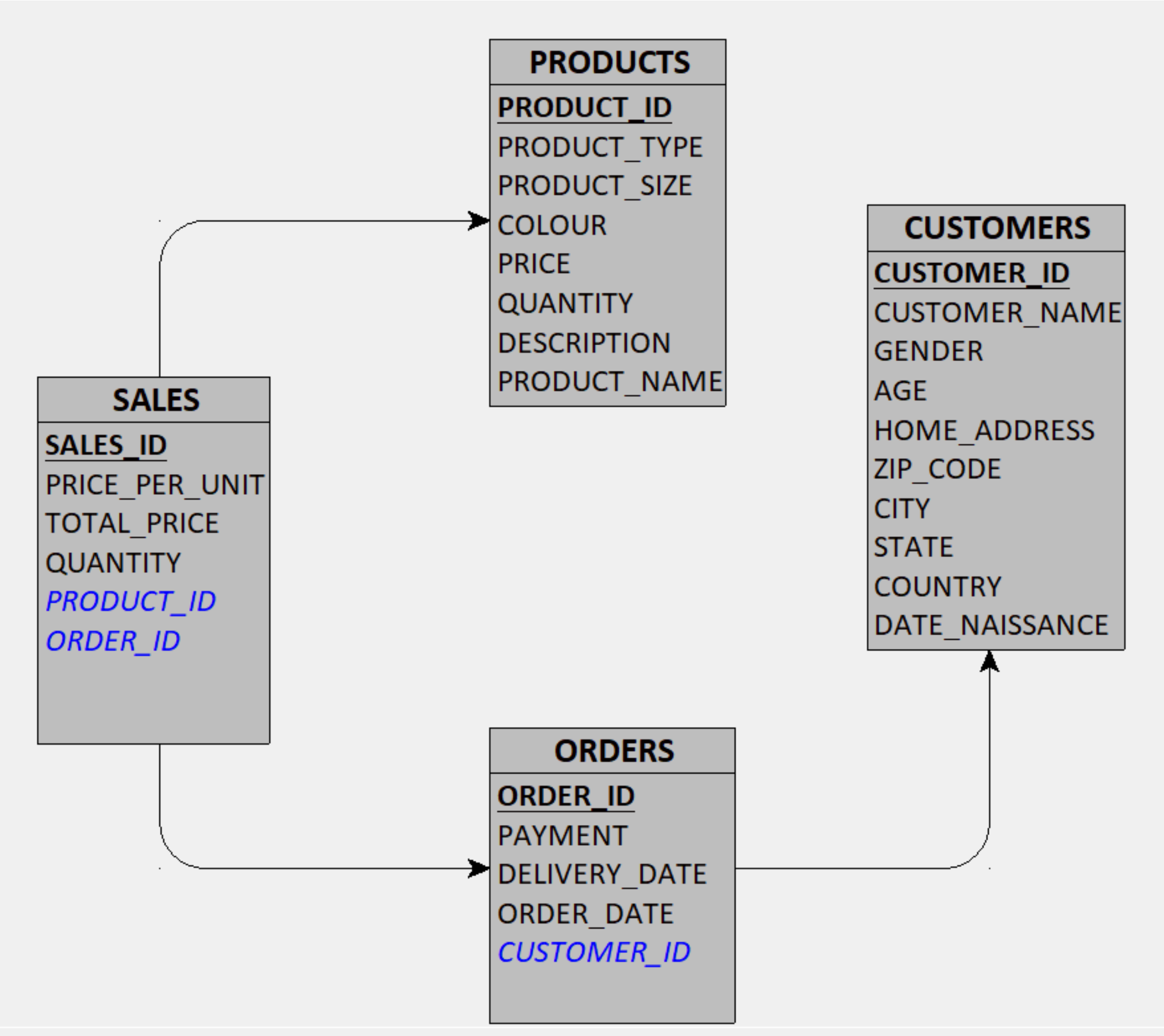


delivery_status	Nombre de order_id	Somme de delivery_cost
Cancelled	11	869,36
Delivered	8	511,06
In Transit	7	565,66
Pending	7	734,37
Total	**33**	**2 680,45**

Dashboards pour l’année 2024

IV. Annexe

Source de donnée 1 :



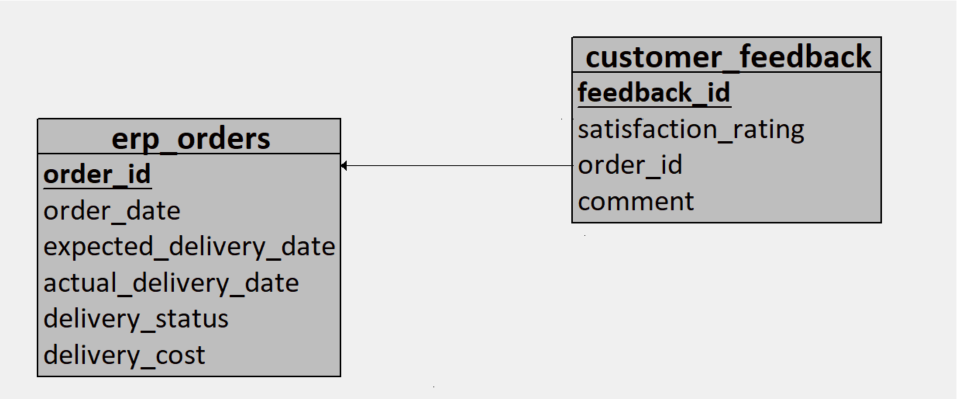
PRODUCTS
PRODUCT_ID
PRODUCT_TYPE
PRODUCT_SIZE
COLOUR
PRICE
QUANTITY
DESCRIPTION
PRODUCT_NAME

CUSTOMERS
CUSTOMER_ID
CUSTOMER_NAME
GENDER
AGE
HOME_ADDRESS
ZIP_CODE
CITY
STATE
COUNTRY
DATE_NAISSANCE

SALES
SALES_ID
PRICE_PER_UNIT
TOTAL_PRICE
QUANTITY
PRODUCT_ID
ORDER_ID

ORDERS
ORDER_ID
PAYMENT
DELIVERY_DATE
ORDER_DATE
CUSTOMER_ID

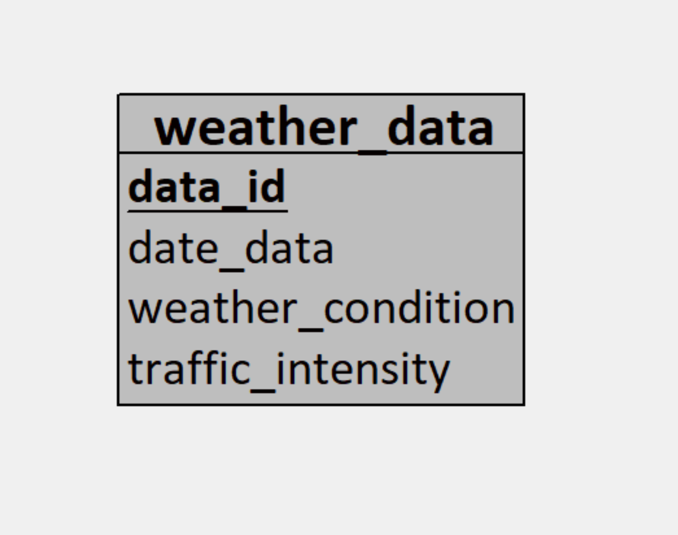
Source de donnée 2 :



erp_orders
order_id
order_date
expected_delivery_date
actual_delivery_date
delivery_status
delivery_cost

customer_feedback
feedback_id
satisfaction_rating
order_id
comment

Source de donnée 3 :



weather_data

data_id

date_data

weather_condition

traffic_intensity