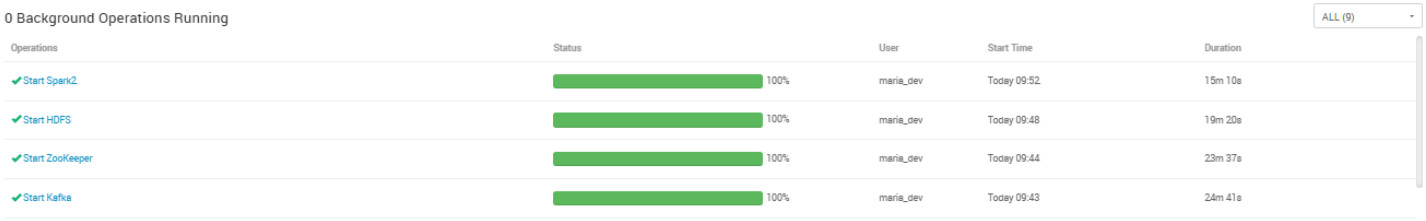
Projet : Traitement des données en temps réel

MEMBRES DU GROUPE :

TCHOUMBA Idriss Namen
MOWA Ethan
METSA TEDJOU Mélissa
NOUTSA FOTSO Franck
OUATTARA Lancina Tobin

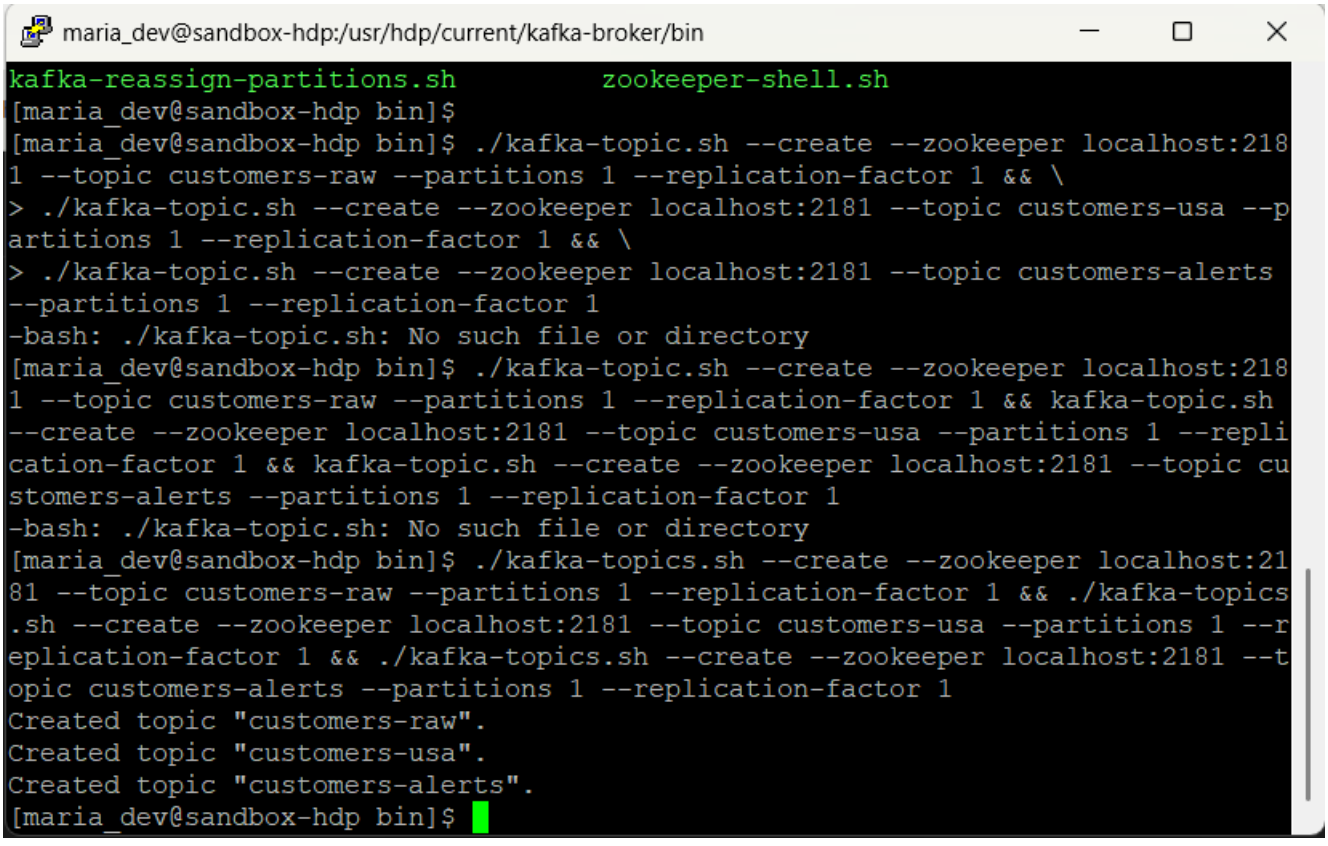
Etape 0 : Services

Préparation de l'infrastructure streaming sur HDP



Etape 1 : Création des topics

Action (via SSH / PuTTY) :



Etape 2 : Producer Python (S3 -> Kafka)

Exécution du producteur pour envoie de messages dans notre topic kafka (consumers-raw) à la suite de la récupération du fichier sur S3.

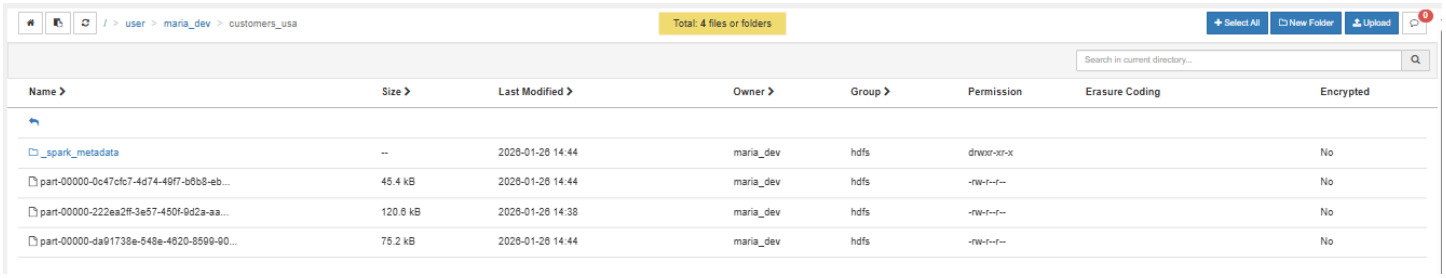
```
[maria_dev@sandbox-hdp ~]$ python python_producer.py
Tentative de lecture sur S3 (Mode Anonyme)...
/home/maria_dev/.local/lib/python2.7/site-packages/boto3/compat.py:86: PythonDeprecationW
arning: Boto3 will no longer support Python 2.7 starting July 15, 2021. To continue recei
ving service updates, bug fixes, and security updates please upgrade to Python 3.6 or lat
er. More information can be found here: https://aws.amazon.com/blogs/developer/announcing
-end-of-support-for-python-2-7-in-aws-sdk-for-python-and-aws-cli-v1/
  warnings.warn(warning, PythonDeprecationWarning)
Succes : Donnees recuperees depuis S3.
Termine : 1000 messages envoyes vers le topic 'customers-raw'.
```

Etape 3 : Spark Structured Streaming (Kafka -> Traitement)

En effectuant la commande suivante pour le consumer (avec un souci de versions à prendre en compte) :

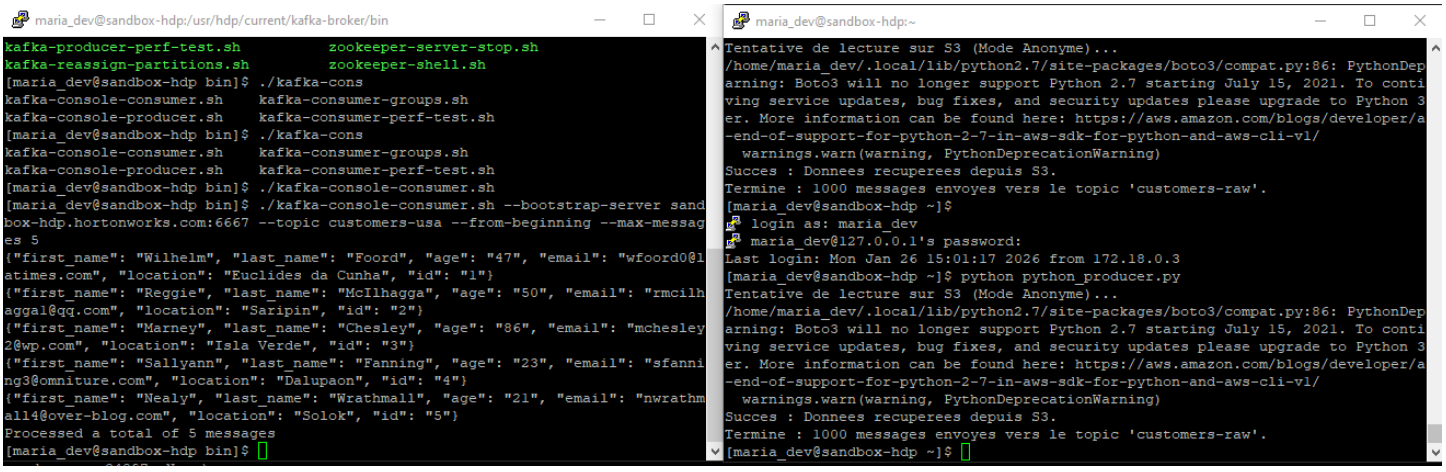
spark-submit --jars spark-sql-kafka-0-10_2.11-2.3.2.jar,kafka-clients-1.1.1.jar python_spark_job.py

Nous avons réussi à mettre en place un job PySpark pour le consommateur et écrire les clients USA vers HDFS.

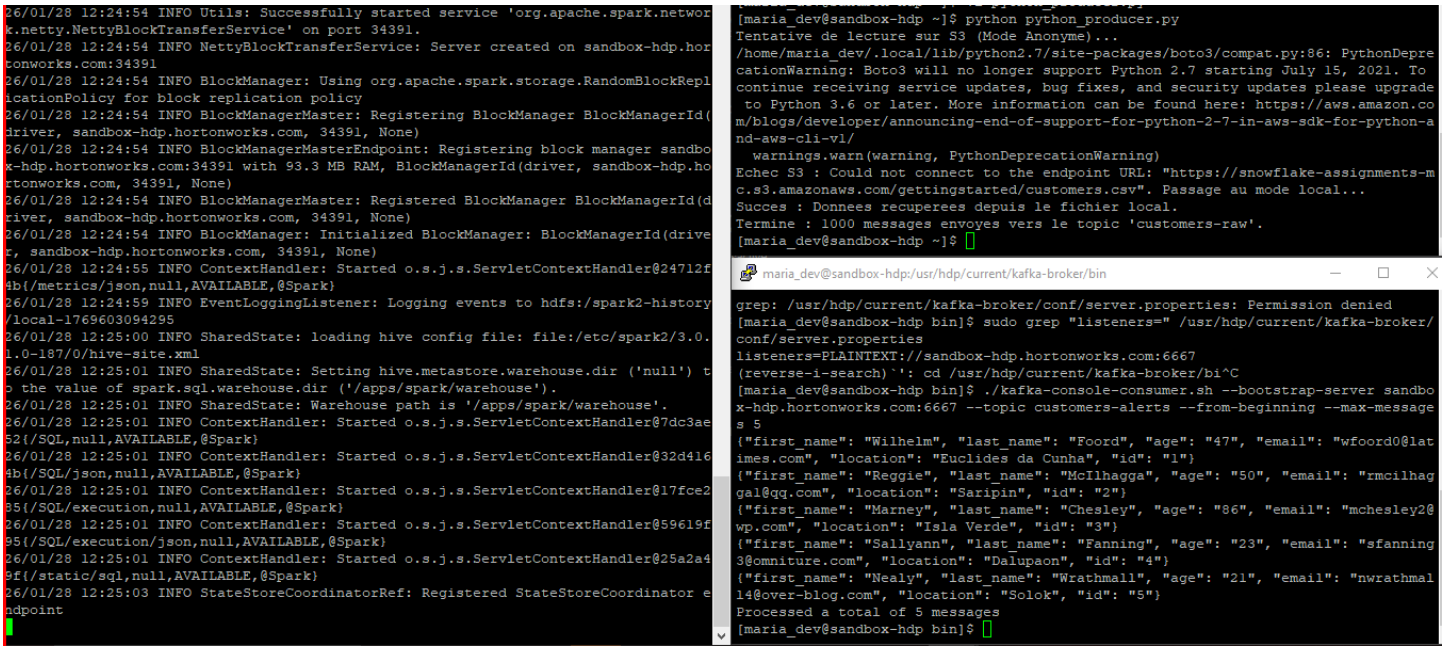


Etape 4 : Tests de bout en bout

- L'action éffectuée ici est le test entre le producteur vers customers-raw et consommer customers-usa avec kafka-console-consumer.



- L'action éffectuée ici est le test entre le producteur vers customers-raw et consommer customers-alerts avec kafka-console-consumer.



- Vérification faite au niveau HDFS pour customers-usa

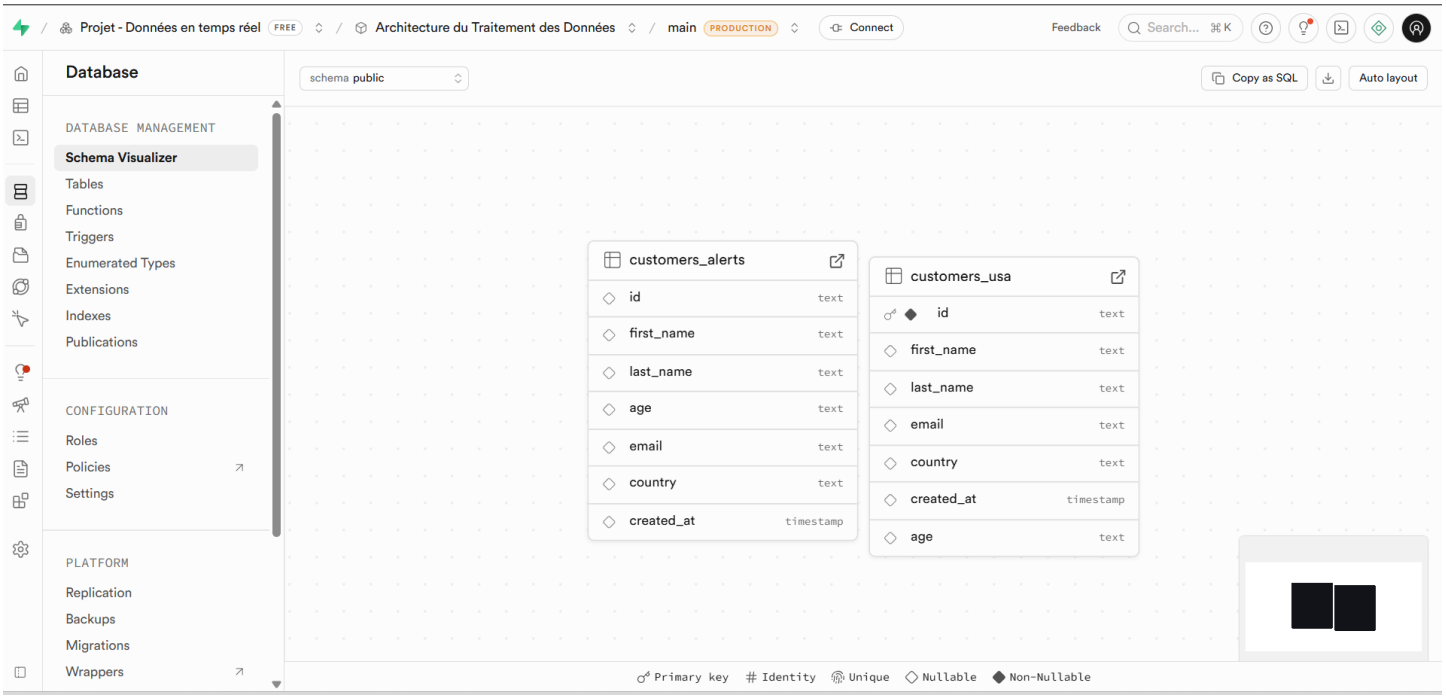
```
[maria_dev@sandbox-hdp bin]$ hdfs dfs -ls /user/maria_dev/customers_usa
Found 4 items
drwxr-xr-x   - maria_dev hdfs          0 2026-01-26 13:44 /user/maria_dev/customers_usa/_spark_metadata
-rw-r--r--   1 maria_dev hdfs      46494 2026-01-26 13:44 /user/maria_dev/customers_usa/part-00000-0c47cfc7-4d74-49f7-b6b8-ebllabe0elbd-c000.json
-rw-r--r--   1 maria_dev hdfs     123508 2026-01-26 13:38 /user/maria_dev/customers_usa/part-00000-222ea2ff-3e57-450f-9d2a-aa9bce2a8d7f-c000.json
-rw-r--r--   1 maria_dev hdfs      77014 2026-01-26 13:44 /user/maria_dev/customers_usa/part-00000-da91738e-548e-4620-8599-904474a37729-c000.json
[maria_dev@sandbox-hdp bin]$
```

```
[maria_dev@sandbox-hdp bin]$ hdfs dfs -ls /user/maria_dev/customers_alerts
Found 4 items
drwxr-xr-x   - maria_dev hdfs          0 2026-01-28 12:57 /user/maria_dev/customers_alerts/_spark_metadata
-rw-r--r--   1 maria_dev hdfs     367524 2026-01-28 12:55 /user/maria_dev/customers_alerts/part-00000-3a546fd6-ceee-497a-a8eb-0b9542d0d389-c000.json
-rw-r--r--   1 maria_dev hdfs     110628 2026-01-28 12:57 /user/maria_dev/customers_alerts/part-00000-48068fa3-71ee-465f-bfbf-12b2d37ff673-c000.json
-rw-r--r--   1 maria_dev hdfs      11880 2026-01-28 12:57 /user/maria_dev/customers_alerts/part-00000-80b84163-a5f3-449b-ace1-fadc2fcb4d03-c000.json
[maria_dev@sandbox-hdp bin]$
```

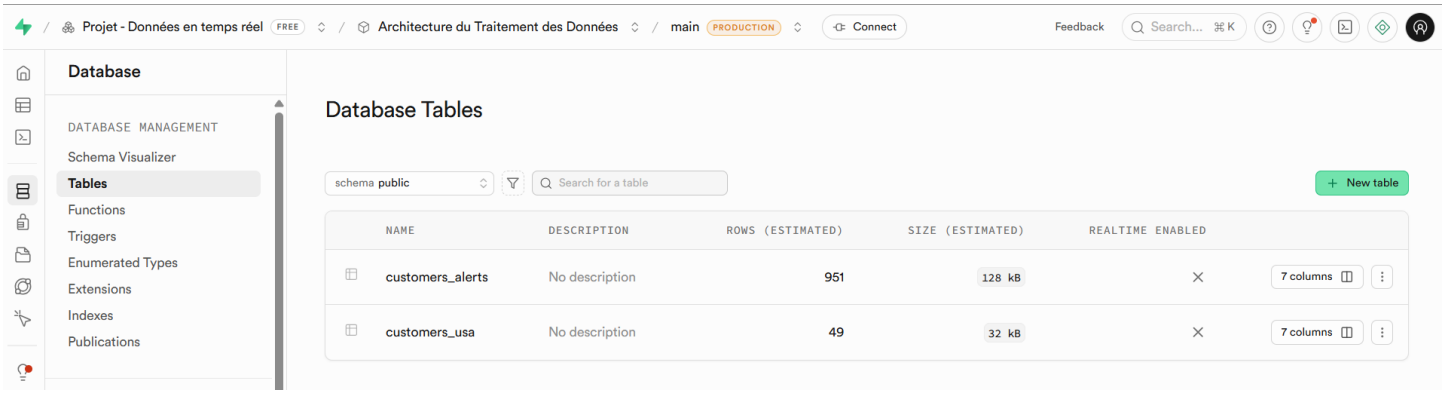
Etape 5 : Extension et Amélioration

1. Stockage dans une vraie base de données

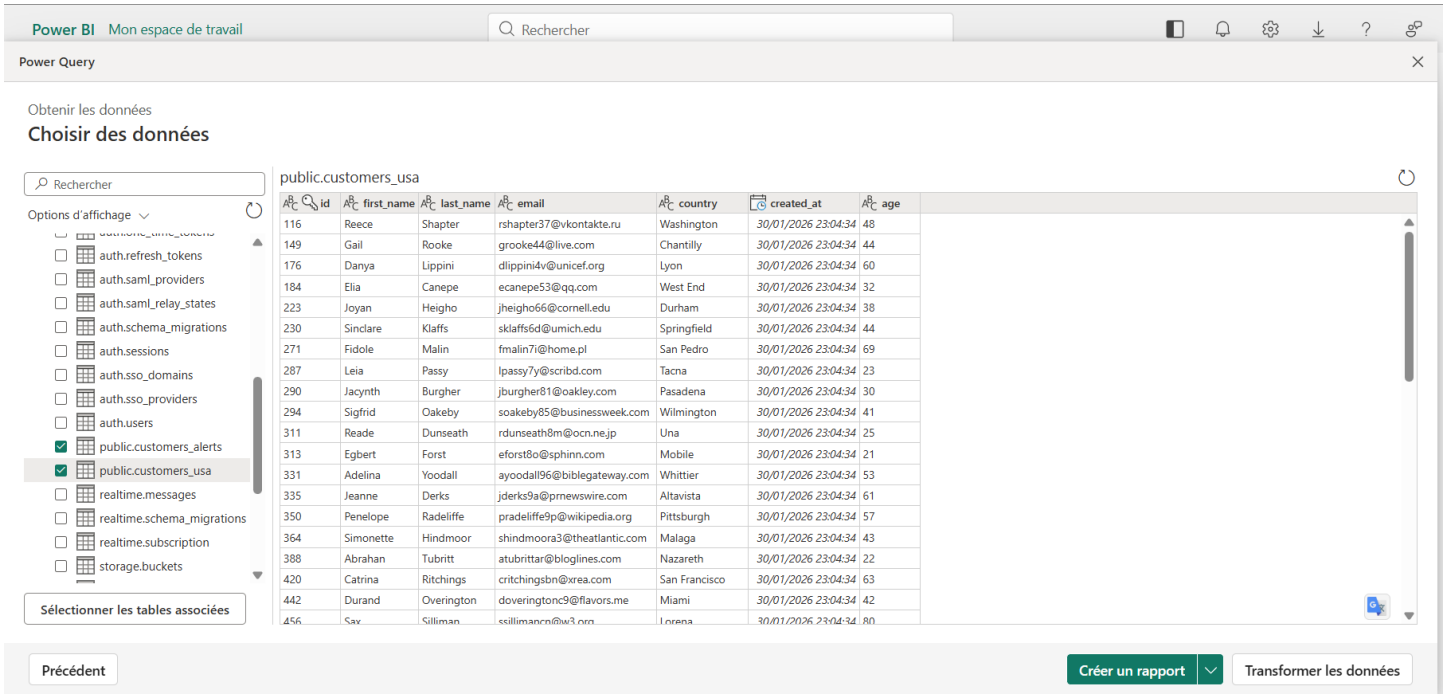
Comme choix de notre base donnée, Supabase a été choisi comme base de données car il repose sur PostgreSQL, une base relationnelle robuste et très utilisée en production.
Lien de base de données : https://supabase.com/dashboard/org/tvugxmxqzkwjuphokzai

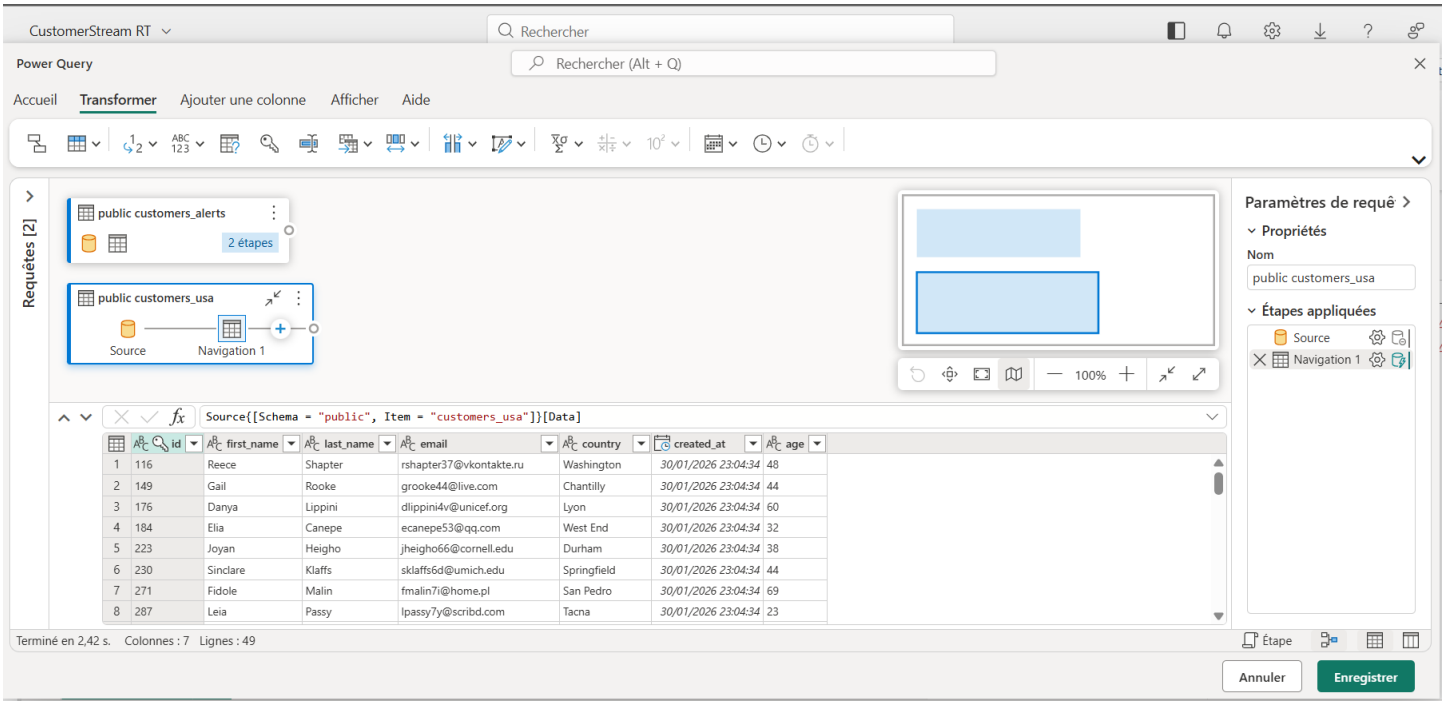


L'image suivante illustre le stockage des données dans les tables :



2. Visualisation : Power Bi





3. Lien Readme :

https://github.com/Ouatson/Projet-Traitement-de-donn-es-en-temps-r-el.git