

# **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

0 Background Operations Running				
Operations	Status	User	Start Time	Duration
✓ Start Spark2	100%	maria_dev	Today 09:52	15m 10s
✓ Start HDFS	100%	maria_dev	Today 09:48	19m 20s
✓ Start ZooKeeper	100%	maria_dev	Today 09:44	23m 37s
✓ Start Kafka	100%	maria_dev	Today 09:43	24m 41s

## Etape 1 : Création des topics

Action (via SSH / PuTTY) :

```

maria_dev@sandbox-hdp:/usr/hdp/current/kafka-broker/bin
kafka-reassign-partitions.sh      zookeeper-shell.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-topic.sh --create --zookeeper localhost:2181 --topic customers-raw --partitions 1 --replication-factor 1 && \
> ./kafka-topic.sh --create --zookeeper localhost:2181 --topic customers-usa --partitions 1 --replication-factor 1 && \
> ./kafka-topic.sh --create --zookeeper localhost:2181 --topic customers-alerts --partitions 1 --replication-factor 1
-bash: ./kafka-topic.sh: No such file or directory
[maria_dev@sandbox-hdp bin]$ ./kafka-topic.sh --create --zookeeper localhost:2181 --topic customers-raw --partitions 1 --replication-factor 1 && kafka-topic.sh \
--create --zookeeper localhost:2181 --topic customers-usa --partitions 1 --replication-factor 1 && kafka-topic.sh \
--create --zookeeper localhost:2181 --topic customers-alerts --partitions 1 --replication-factor 1
-bash: ./kafka-topic.sh: No such file or directory
[maria_dev@sandbox-hdp bin]$ ./kafka-topics.sh --create --zookeeper localhost:2181 --topic customers-raw --partitions 1 --replication-factor 1 && ./kafka-topics.sh \
--create --zookeeper localhost:2181 --topic customers-usa --partitions 1 --replication-factor 1 && ./kafka-topics.sh \
--create --zookeeper localhost:2181 --topic customers-alerts --partitions 1 --replication-factor 1
Created topic "customers-raw".
Created topic "customers-usa".
Created topic "customers-alerts".
[maria_dev@sandbox-hdp bin]$ 

```

## Etape 2 : Producer Python (S3 -> Kafka)

Exécution du producteur pour envoi 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: PythonDeprecationWarning: Boto3 will no longer support Python 2.7 starting July 15, 2021. To continue receiving service updates, bug fixes, and security updates please upgrade to Python 3.6 or later. 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 recuperées depuis S3.
Termine : 1000 messages envoyés 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.

Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
_spark_metadata	--	2020-01-26 14:44	maria_dev	hdfs	drwxr-xr-x		No
part-00000-0c47cf07-4d74-49f7-b0b8-eb...	45.4 kB	2020-01-26 14:44	maria_dev	hdfs	-rw-r--r--		No
part-00000-222ea2ff-3e57-450f-0d2a-aa...	120.8 kB	2020-01-26 14:38	maria_dev	hdfs	-rw-r--r--		No
part-00000-da91738e-548e-4620-8599-90...	75.2 kB	2020-01-26 14:44	maria_dev	hdfs	-rw-r--r--		No

## Etape 4 : Tests de bout en bout

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

The screenshot shows two terminal windows side-by-side. The left window displays the execution of the producer script, showing the output of reading from S3 and sending 1000 messages to the 'customers-raw' topic. The right window shows the execution of the consumer script, which reads from the same topic and prints the received data to the console. Both windows include Python deprecation warnings about the use of Python 2.7.

```
[maria_dev@sandbox-hdp:/usr/hdp/current/kafka-broker/bin]$ kafka-producer-perf-test.sh
kafka-producer-reassign-partitions.sh      zookeeper-server-stop.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh      zookeeper-shell.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-console-producer.sh      kafka-consumer-groups.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-console-producer.sh      kafka-consumer-perf-test.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh      kafka-cons
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh      kafka-consumer-groups.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh      kafka-consumer-perf-test.sh
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh      [maria_dev@sandbox-hdp bin]$ --bootstrap-server sand
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh      box-hdp.hortonworks.com:6667 --topic customers-usa --from-beginning --max-messages 5
("first_name": "Wilhelm", "last_name": "Foord", "age": "47", "email": "wfoord@1atimes.com", "location": "Euclides da Cunha", "id": "1"}
("first_name": "Reggie", "last_name": "McIlhagga", "age": "50", "email": "rmcilhaggag@qq.com", "location": "Saripin", "id": "2"}
("first_name": "Marney", "last_name": "Chesley", "age": "86", "email": "mchesley2@wp.com", "location": "Isla Verde", "id": "3"}
("first_name": "Sallyann", "last_name": "Fanning", "age": "23", "email": "sfanning@omnitrite.com", "location": "Dalupaon", "id": "4"}
("first_name": "Nealy", "last_name": "Wrathmall", "age": "21", "email": "nwrathmall4@over-blog.com", "location": "Solok", "id": "5")
Processed a total of 5 messages
[maria_dev@sandbox-hdp bin]$ [maria_dev@sandbox-hdp ~]$ Tentative de lecture sur S3 (Mode Anonyme)...
/home/maria_dev/.local/lib/python2.7/site-packages/boto3/compat.py:86: PythonDeprecationWarning: Boto3 will no longer support Python 2.7 starting July 15, 2021. To continuing service updates, bug fixes, and security updates please upgrade to Python 3.6 or later. 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 recuperées depuis S3.
Termine : 1000 messages envoyés vers le topic 'customers-raw'.
[maria_dev@sandbox-hdp ~]$ [maria_dev@sandbox-hdp ~]$ login as: maria_dev
[maria_dev@127.0.0.1's password:
Last login: Mon Jan 26 15:01:17 2026 from 172.18.0.3
[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: PythonDeprecationWarning: Boto3 will no longer support Python 2.7 starting July 15, 2021. To continuing service updates, bug fixes, and security updates please upgrade to Python 3.6 or later. 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 recuperées depuis S3.
Termine : 1000 messages envoyés vers le topic 'customers-raw'.
[maria_dev@sandbox-hdp ~]$
```

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

```

26/01/28 12:24:54 INFO Utils: Successfully started service 'org.apache.spark.netwörk.netty.NettyBlockTransferService' on port 34391.
26/01/28 12:24:54 INFO NettyBlockTransferService: Server created on sandbox-hdp.hortonworks.com:34391
26/01/28 12:24:54 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
26/01/28 12:24:54 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, sandbox-hdp.hortonworks.com, 34391, None)
26/01/28 12:24:54 INFO BlockManagerMasterEndpoint: Registering block manager sandbox-hdp.hortonworks.com:34391 with 93.3 MB RAM, BlockManagerId(driver, sandbox-hdp.hortonworks.com, 34391, None)
26/01/28 12:24:54 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, sandbox-hdp.hortonworks.com, 34391, None)
26/01/28 12:24:54 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, sandbox-hdp.hortonworks.com, 34391, None)
26/01/28 12:24:55 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@24712fb{metrics/json,null,AVAILABLE,@Spark}
26/01/28 12:24:59 INFO EventLoggingListener: Logging events to hdfs://spark2-history/local-/1769603094295
26/01/28 12:25:00 INFO SharedState: loading hive config file: file:/etc/spark2/3.0.1-0-187/0/hive-site.xml
26/01/28 12:25:01 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to value of spark.sql.warehouse.dir ('/apps/spark/warehouse').
26/01/28 12:25:01 INFO SharedState: Warehouse path is '/apps/spark/warehouse'.
26/01/28 12:25:01 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7dc3ae52{/SQL,null,AVAILABLE,@Spark}
26/01/28 12:25:01 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@32d4164b{/SQL/json,null,AVAILABLE,@Spark}
26/01/28 12:25:01 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@17fce295{/SQL/execution,null,AVAILABLE,@Spark}
26/01/28 12:25:01 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@59619f95{/SQL/execution/json,null,AVAILABLE,@Spark}
26/01/28 12:25:01 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@25a2a9bf{/static/sql,null,AVAILABLE,@Spark}
26/01/28 12:25:03 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
[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: PythonDeprecationWarning: Boto3 will no longer support Python 2.7 starting July 15, 2021. To continue receiving service updates, bug fixes, and security updates please upgrade to Python 3.6 or later. 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)
Echec S3 : Could not connect to the endpoint URL: "https://snowflake-assignments-m.c.s3.amazonaws.com/gettingstarted/customers.csv". Passage au mode local...
Succès : Données récupérées depuis le fichier local.
Terminé : 1000 messages envoyés vers le topic 'customers-raw'.
[maria_dev@sandbox-hdp ~]$ 
[maria_dev@sandbox-hdp:usr/hdp/current/kafka-broker/bin]$ grep: /usr/hdp/current/kafka-broker/conf/server.properties: Permission denied
[maria_dev@sandbox-hdp bin]$ sudo grep "listeners" /usr/hdp/current/kafka-broker/conf/server.properties
listeners=PLAINTEXT://sandbox-hdp.hortonworks.com:6667
(reverse-i-search)`': cd /usr/hdp/current/kafka-broker/bi^C
[maria_dev@sandbox-hdp bin]$ ./kafka-console-consumer.sh --bootstrap-server sandbox-hdp.hortonworks.com:6667 --topic customers-alerts --from-beginning --max-messages 5
>{"first_name": "Wilhelm", "last_name": "Foord", "age": "47", "email": "wfoord0@latimes.com", "location": "Euclides da Cunha", "id": "1"}
>{"first_name": "Reggie", "last_name": "McIlhagga", "age": "50", "email": "rmcilhaggalog@gmail.com", "location": "Saripin", "id": "2"}
>{"first_name": "Marney", "last_name": "Chesley", "age": "86", "email": "mchesley20wp.com", "location": "Isla Verde", "id": "3"}
>{"first_name": "Sallyann", "last_name": "Fanning", "age": "23", "email": "sfanning3@ommiture.com", "location": "Daluqaon", "id": "4"}
>{"first_name": "Nealy", "last_name": "Wrathmall", "age": "21", "email": "nwrathmall1@over-blog.com", "location": "Solok", "id": "5"}
Processed a total of 5 messages
[maria_dev@sandbox-hdp bin]$ 
```

- 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-00000-0c47cf7-4d74-49f7-b6b8-eb1labe0elbd-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-acel-fadc2fcbb4d03-c000.json
[maria_dev@sandbox-hdp bin]$ 
```

## Etape 5 : Extension et Amélioration

### I. 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>

The screenshot shows the Supabase Schema Visualizer interface. On the left, there's a sidebar with various database management options like Database, Schema Visualizer (which is selected), Tables, Functions, Triggers, etc. The main area displays two tables: 'customers\_alerts' and 'customers\_usa'. Both tables have columns for id, first\_name, last\_name, age, email, country, and created\_at. The 'customers\_usa' table also includes a primary key column 'id'.

Table	Column	Type
customers_alerts	id	text
	first_name	text
	last_name	text
	age	text
	email	text
	country	text
	created_at	timestamp
customers_usa	id	text
	first_name	text
	last_name	text
	email	text
	country	text
	created_at	timestamp
	age	text

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

This screenshot shows the 'Database Tables' page in the Supabase interface. The sidebar on the left has 'Tables' selected under 'Schema Visualizer'. The main area lists the 'customers\_alerts' and 'customers\_usa' tables. Each table entry includes a preview icon, name, description, estimated rows, size, real-time status, and edit/copy icons.

NAME	DESCRIPTION	ROWS (ESTIMATED)	SIZE (ESTIMATED)	REALTIME ENABLED
customers_alerts	No description	951	128 kB	X
customers_usa	No description	49	32 kB	X

## 2. Visualisation : Power Bi

Power BI Mon espace de travail

Power Query

Obtenir les données

**Choisir des données**

Rechercher

Options d'affichage

- public.customers\_usa
- public.customers\_alerts
- public.customers\_usa
- realtime.messages
- realtime.schema\_migrations
- realtime.subscription
- storage.buckets

Rechercher

ID	first_name	last_name	email	country	created_at	age
116	Reece	Shapter	rshapter37@vkontakte.ru	Washington	30/01/2026 23:04:34	48
149	Gail	Rooke	grooke44@live.com	Chantilly	30/01/2026 23:04:34	44
176	Danya	Lippini	dlippini4v@unicef.org	Lyon	30/01/2026 23:04:34	60
184	Elia	Canepe	ecanepe53@qq.com	West End	30/01/2026 23:04:34	32
223	Joyan	Heigho	jheigho66@cornell.edu	Durham	30/01/2026 23:04:34	38
230	Sindare	Klaffs	sklaffs6d@umich.edu	Springfield	30/01/2026 23:04:34	44
271	Fidole	Malin	fmalin7i@home.pl	San Pedro	30/01/2026 23:04:34	69
287	Leia	Passy	lpassy7y@scribd.com	Tacna	30/01/2026 23:04:34	23
290	Jacynth	Burgher	jburgher81@oakley.com	Pasadena	30/01/2026 23:04:34	30
294	Sigrid	Oakeby	soakeby85@businessweek.com	Wilmington	30/01/2026 23:04:34	41
311	Reade	Dunseath	rdunseath8m@ocn.ne.jp	Una	30/01/2026 23:04:34	25
313	Egbert	Forst	efors8e@sphinn.com	Mobile	30/01/2026 23:04:34	21
331	Adelina	Yoodall	ayoodall96@biblegateway.com	Whittier	30/01/2026 23:04:34	53
335	Jeanne	Derks	jderks9a@prnewswire.com	Altavista	30/01/2026 23:04:34	61
350	Penelope	Radcliffe	pradcliffe8p@wikipedia.org	Pittsburgh	30/01/2026 23:04:34	57
364	Simonet	Hindmoor	shindmoor3@theatlantic.com	Malaga	30/01/2026 23:04:34	43
388	Abrahan	Tubritt	atubrittar@bloglines.com	Nazareth	30/01/2026 23:04:34	22
420	Catrina	Ritchings	critchingsbn@rea.com	San Francisco	30/01/2026 23:04:34	63
442	Durand	Overington	doverington9j@flavors.me	Miami	30/01/2026 23:04:34	42
456	Sav	Silliman	ssillimanrcn@w3.org	Lorena	30/01/2026 23:04:34	80

Rechercher les tables associées

Précédent

Créer un rapport

Transformer les données

CustomerStream RT

Power Query

Accueil Transformer Ajouter une colonne Afficher Aide

Rechercher

Rechercher (Alt + Q)

Requêtes [2]

public customers\_alerts

public customers\_usa

Source Navigation 1

Paramètres de requête

Propriétés Nom public customers\_usa

Étapes appliquées

Étape Annuler Enregistrer

Terminé en 2,42 s. Colonnes : 7 Lignes : 49

ID	first_name	last_name	email	country	created_at	age
1 116	Reece	Shapter	rshapter37@vkontakte.ru	Washington	30/01/2026 23:04:34	48
2 149	Gail	Rooke	grooke44@live.com	Chantilly	30/01/2026 23:04:34	44
3 176	Danya	Lippini	dlippini4v@unicef.org	Lyon	30/01/2026 23:04:34	60
4 184	Elia	Canepe	ecanepe53@qq.com	West End	30/01/2026 23:04:34	32
5 223	Joyan	Heigho	jheigho66@cornell.edu	Durham	30/01/2026 23:04:34	38
6 230	Sindare	Klaffs	sklaffs6d@umich.edu	Springfield	30/01/2026 23:04:34	44
7 271	Fidole	Malin	fmalin7i@home.pl	San Pedro	30/01/2026 23:04:34	69
8 287	Leia	Passy	lpassy7y@scribd.com	Tacna	30/01/2026 23:04:34	23

### 3. Lien Readme :

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