

# Projet : Traitement des données en temps réel



MEMBRES DU GROUPE :

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

## Etape 0 : Services

### Préparation de l'infrastructure streaming sur HDP

0 Background Operations Running ALL (9)

Operations	Status	User	Start Time	Duration
✓ Start Spark2	<div><div></div></div> 100%	maria_dev	Today 09:52	15m 10s
✓ Start HDFS	<div><div></div></div> 100%	maria_dev	Today 09:48	19m 20s
✓ Start ZooKeeper	<div><div></div></div> 100%	maria_dev	Today 09:44	23m 37s
✓ Start Kafka	<div><div></div></div> 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]$
[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 cu
stomers-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 --t
opic 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 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.

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-00000-0c47cfc7-4d74-49f7-b8b8-eb...	45.4 kB	2020-01-26 14:44	maria_dev	hdfs	-rw-r--r--		No
part-00000-222ea2ff-3e57-450f-9d2a-aa...	120.6 kB	2020-01-26 14:38	maria_dev	hdfs	-rw-r--r--		No
part-00000-da91738e-548e-4820-8599-90...	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.

```
maria_dev@sandbox-hdp:~/usr/hdp/current/kafka-broker/bin$ kafka-console-consumer --zookeeper localhost:2181 --topic customers-usa --from-beginning
{"first_name": "Wilhelm", "last_name": "Foord", "age": "47", "email": "wfoord@limes.com", "location": "Euclides da Cunha", "id": "1"}
{"first_name": "Reggie", "last_name": "McIlhagga", "age": "50", "email": "rmcilhagga@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": "sfanning3@omnitel.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:~$ 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 recuperees depuis S3.
Termine : 1000 messages envoyes 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.network.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 12:24:55 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@24712f4b(/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 the 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@17fce285(/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@25a2a9f(/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-mc.s3.amazonaws.com/gettingstarted/customers.csv". Passage au mode local...
Succes : Donnees recuperees depuis le fichier local.
Termine : 1000 messages envoyes 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/bin^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": "rmcilhagga@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": "sfanning3@comniture.com", "location": "Dalupao", "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]$
```

- 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/customer_s_usa/_spark_metadata
-rw-r--r-- 1 maria_dev hdfs 46494 2026-01-26 13:44 /user/maria_dev/customer_s_usa/part-00000-0c47cfc7-4d74-49f7-b6b8-eb1lab0e1bd-c000.json
-rw-r--r-- 1 maria_dev hdfs 123508 2026-01-26 13:38 /user/maria_dev/customer_s_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/customer_s_usa/part-00000-da91738e-548e-4620-8599-904474a37729-c000.json
[maria_dev@sandbox-hdp bin]$
```

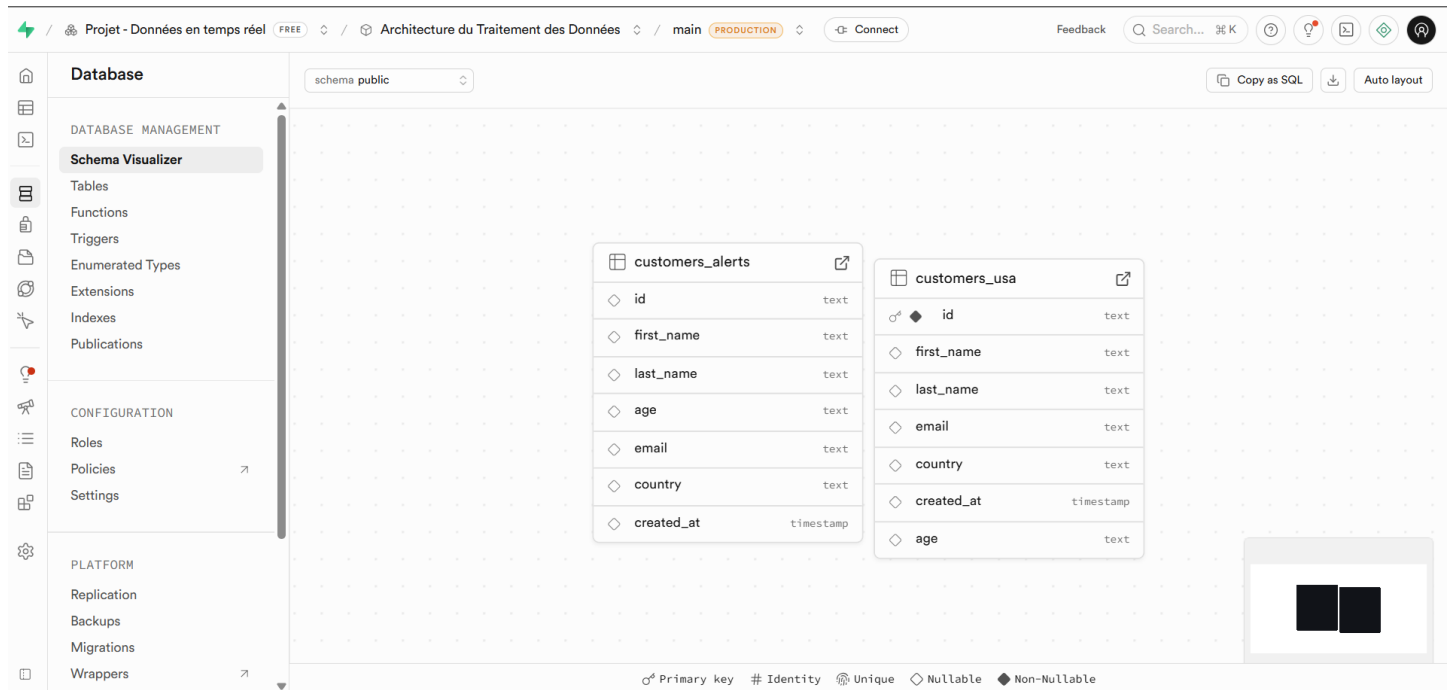
```
[maria_dev@sandbox-hdp bin]$ hdfs dfs -ls /user/maria_dev/customers_alerts
Found 4 items
drwxr-xr-x - maria_dev hdfs 0 2026-01-28 12:57 /user/maria_dev/customer_s_alerts/_spark_metadata
-rw-r--r-- 1 maria_dev hdfs 367524 2026-01-28 12:55 /user/maria_dev/customer_s_alerts/part-00000-3a546fd6-ceee-497a-a8eb-0b9542d0d389-c000.json
-rw-r--r-- 1 maria_dev hdfs 110628 2026-01-28 12:57 /user/maria_dev/customer_s_alerts/part-00000-48068fa3-7lee-465f-bfbf-12b2d37ff673-c000.json
-rw-r--r-- 1 maria_dev hdfs 11880 2026-01-28 12:57 /user/maria_dev/customer_s_alerts/part-00000-80b84163-a5f3-449b-acel-fadc2fcb4d03-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/tvugmxqzkwjuphokzai>



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

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

## 2. Visualisation : Power Bi

Power BI Mon espace de travail

Rechercher

Power Query

Obtenir les données

Choisir des données

Rechercher

Options d'affichage

- ☐ auth.refresh\_tokens
- ☐ auth.saml\_providers
- ☐ auth.saml\_relay\_states
- ☐ auth.schema\_migrations
- ☐ auth.sessions
- ☐ auth.sso\_domains
- ☐ auth.sso\_providers
- ☐ auth.users
- ☒ public.customers\_alerts
- ☒ public.customers\_usa
- ☐ realtime.messages
- ☐ realtime.schema\_migrations
- ☐ realtime.subscription
- ☐ storage.buckets

Sélectionner les tables associées

public.customers\_usa

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	Sinclare	Klafts	sklafts6d@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	Sigfrid	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	eforst8o@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@pmewswire.com	Altavista	30/01/2026 23:04:34	61
350	Penelope	Radeliffe	pradeliffe9p@wikipedia.org	Pittsburgh	30/01/2026 23:04:34	57
364	Simonette	Hindmoor	shindmoora3@theatlantic.com	Malaga	30/01/2026 23:04:34	43
388	Abraham	Tubritt	atubritt@blogger.com	Nazareth	30/01/2026 23:04:34	22
420	Catrina	Ritchings	critchingsbn@xrea.com	San Francisco	30/01/2026 23:04:34	63
442	Durand	Overington	doveringtonc9@flavors.me	Miami	30/01/2026 23:04:34	42
456	Sav	Sillman	ssillman@u3.com	Lorena	30/01/2026 23:04:34	80

Précédent

Créer un rapport

Transformer les données

CustomerStream RT

Rechercher

Power Query

Rechercher (Alt + Q)

Accueil Transformer Ajouter une colonne Afficher Aide

Requêtes [2]

- public customers\_alerts (2 étapes)
- public customers\_usa (Source, Navigation 1)

Paramètres de requête

Propriétés

Nom: public customers\_usa

Étapes appliquées

- Source
- Navigation 1

Source{[Schema = "public", Item = "customers\_usa"]}[Data]

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	Sinclare	Klafts	sklafts6d@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

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

Annuler Enregistrer

### 3. Lien Readme :

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