Transfert Kafka Cassandra

Nous avons fait le choix d'utiliser des connecteurs et la base de données Cassandra car le KCQL Kafka Connect Query Language facilite la lecture des données depuis les topics.

1. **Mise en place de la base de données Cassandra**

- Téléchargement de Cassandra mise en place de la base de données dans des variable d'environnement
- Mode CQL pour créer un KEYSPACE et une table

```
CREATE KEYSPACE myspace WITH replication = {'class':'SimpleStrategy',
'replication_factor':1} ;
USE myspace;
CREATE TABLE infractions(
        summons_number int PRIMARY KEY,
        plate_id text,
        registration_state text,
        issue_date text,
        violation_code int,
        street_code1 int,
        house_number text,
        street_name text
);
```

2. **Configuration de Kafka Connect**

- Dans un premier temps on installe et configure Kafka Connect afin de lire les données du topic Kafka. La configuration telle que le nom du topics à écouter est indiqué dans le fichier connect-file-sink.properties.
- On lance l'instance de Kafka Connect dans le mode Distributed et pas Standalone. En effet il s'agit du mode qui correspond à notre architecture, nous serons amené à avoir plusieurs producers étant donné qu'il y aura plusieurs drones.
- Installation de Stream Reactor afin de pouvoir utiliser avro pour formater l'envoi des données.
- Cela nous amène à la mise en place de Cassandra Sink.
- Ainsi les données pourront être consultées depuis la base de données Cassandra.

3. **Statistiques des données dans Cassandra via Spark**

-
```
COPY infractions(summons_number , plate_id , registration_state , issue_date ,
violation_code , street_code1 , house_number , street_name )
FROM '/{chemin absolu}/MyCSV.csv' WITH header = true ;
```
- Dans un autre shell, aller dans le spark-2.1.0-bin-hadoop2.7, executer

```
bin/spark-shell --packages
com.datastax.spark:spark-cassandra-connector_2.11:2.0.0-M3 --conf
spark.cassandra.connection.host=10.0.2.15
```

- Pour récupérer les données de Cassandra depuis Spark il faut exécuter :

```
val infractions = spark.
    | read.
    | format("org.apache.spark.sql.cassandra").
    | options(Map( "table" -> "infractions", "keyspace" -> "myspace")).
    | load()
infractions.show()
```

- Requêtes SQL pour faire les statistiques

Nombre de contraventions par états

```
sql("SELECT registration_state, count(*) AS number FROM infractions GROUP BY registration_state ORDER BY number DESC").show()
```

```
+------------------+------+
|registration_state|number|
+------------------+------+
|                NY|    49|
|                NJ|    15|
|                CT|     6|
|                PA|     5|
|                99|     4|
|                IL|     3|
|                NC|     2|
|                MD|     2|
|                SC|     1|
|                LA|     1|
|                NH|     1|
|                NE|     1|
|                VA|     1|
|                ME|     1|
|                ID|     1|
|                MA|     1|
|                IN|     1|
|                OH|     1|
|                NB|     1|
+------------------+------+
```

Nombre de contraventions par rues dans l'état de New York

```
sql("SELECT street_name, count(*) AS number FROM infractions WHERE registration_state = 'NY' GROUP BY street_name ORDER BY number DESC").show()
```

```
+--------------------+------+
|         street_name|number|
+--------------------+------+
|                 CPW|     5|
|          LIBERTY ST|     2|
|760 BROADWAY BROOKLY|     2|
|              118 ST|     2|
|             W 175 ST|     2|
|          QUEENS BLVD|     1|
|             W 174 ST|     1|
|           UNION TPK|     1|
|             YORK AVE|     1|
|             W 177 ST|     1|
|      N.PORTLAND AVE|     1|
|        VICTORY BLVD|     1|
|          W 21 STREET|     1|
|  N/S WILLOUGHBY AVE|     1|
|          WEBSTER AVE|     1|
|          CONGRESS ST|     1|
|            W 27TH ST|     1|
|              PINE ST|     1|
|      E/S RIVER TER.|     1|
|             HYATT ST|     1|
+--------------------+------+
only showing top 20 rows
```