Course 5 – Assignment 1

Stream data into pipeline in near-real-time using Spark

Objective

The objective of this assignment is to be able to have data in a platform to run streaming data pipeline. In this assignment, we learn how to

- produce the content of a CSV file to a Kafka topic,
- consume messages from a Kafka topic

Data set

The data set is the one that you analyzed in Course 1 and done ETL in Course 4 and it is STM GTFS data.

Problem statement

We get the information of STM every day and need to run an ETL pipeline to enrich data for reporting and analysis purpose in real-time. Data is split in two

1. A set of tables that build dimension (batch style)
2. Trips that needed to be enriched for analysis and reporting (streaming)

In order to be able to run streaming analysis with a platform such as Spark Streaming, we need to have the records in a streaming platform such as Kafka.

Schema

Trip

Field Name	Data Type
trip_id	String
service_id	String
route_id	Integer
trip_headsign	String
wheelchair_accessible	Integer

Enriched Trip

Field Name	Data Type
trip_id	String
service_id	String
route_id	Integer
trip_headsign	String
date	String
exception_type	Integer
route_long_name	String
route_color	String
wheelchair_accessible	Boolean

Assignment requirements

Data pipeline installation
Create a Kafka topics called `trip` and `enriched_trip` • Replication factor 1 • Number of partitions 3
Extract data from STM
Download the data set of STM GTFS from http://stm.info/sites/default/files/gtfs/gtfs_stm.zip
Have trips.txt file in your local machine accessible
Data pipeline
Produce trips.txt file to Kafka using kafka-console-producer. Each line is one message.
Consume the `trip` topic into your application
Parse each record polled from Kafka into a Trip object
Instantiate an object of `EnrichedTrip` for each message (only populate trip data and leave the rest empty e.g., null)
Convert each `EnrichedTrip` to CSV format and produce it to `enriched_trip` topic

Bonus

- Leverage key to proper partition the data. In `enriched_trip` topic, you can use the combination of `route_id` and `service_id` as the key.