Kafka Streaming Data Processing Pipeline: NYC Taxi Trip Data

This Kafka pipeline is designed to simulate real-time processing of NYC Taxi Trip data using Python and Apache Kafka. The pipeline consists of two components: a Kafka producer and a Kafka consumer.

- ***Producer***: The producer reads trip records from a parquet file containing NYC taxi data. It sends one trip record at a time to Kafka topic named "nyc_taxi", to simulate real time streaming. To introduce delay for simulation, time.sleep() is used. Each message is serialized as JSON and includes timestamp fields converted to ISO format for consistency.
- ***Consumer***: The consumer subscribes to "nyc_taxi" topic and processes each incoming message. It enriches the raw data by:
  - Mapping pickup and drop-off location IDs to human readable zone names using a lookup table.
  - Trip duration is calculated in minutes.
  - Average trip speed in miles per hour is estimated.
  - Payment type codes are translated to descriptive labels.
  - A structured summary of each trip is generated which includes, route, duration, speed, fare and payment method.

This processing step transforms raw trip records into meaningful insights.

Real-World Applications

This pipeline mirrors real-world use cases in urban analytics and transportation systems such as:

- ***Real-time fleet monitoring***: To optimize taxi operations by tracking trip durations, speed and payment trends.
- ***Traffic and congestion analysis***: Utilizing trip speed and duration to identify slow zones or peak congestion periods.
- ***Anomaly detection***: Flag unusually long trips, high fares and uncommon payment types for further investigation.

Demo recording: NYC-TRIP-KAFKA-DEMO