**Streaming ETL Pipeline using Kafka**

This project is completed on Coursera’s cloud environment which is using Theia and Docker container.

**Scenario:**

Developed a data pipeline to analyze road traffic data from national highway toll plazas. As vehicles pass through, key data (vehicle ID, type, toll plaza ID, and timestamp) is streamed via Kafka. The pipeline ingests this real-time data and efficiently loads it into a structured database for analysis and reporting.

Download Kafka:

wget <https://archive.apache.org/dist/kafka/3.7.0/kafka_2.12-3.7.0.tgz>

Extract Kafka from zip file:

tar -xzf kafka\_2.12-3.7.0.tgz

Change path to go into kafka folder:

cd kafka\_2.12-3.7.0

Generate a cluster UUID that will uniquely identify the Kafka cluster, this new cluster id will be used by Kraft controller.

KAFKA\_CLUSTER\_ID="$(bin/kafka-storage.sh random-uuid)"

Kraft requires the log directories to be configures.

bin/kafka-storage.sh format -t $KAFKA\_CLUSTER\_ID -c config/kraft/server.properties

now Kraft is configured and we can start Kafka server as follows and after running command it will show message that kafka server started:

bin/kafka-server-start.sh config/kraft/server.properties

* Now start Mysql server, create database tolldata and table livetolldata:

create database tolldata;

use tolldata;

create table livetolldata(timestamp datetime,vehicle\_id int,vehicle\_type char(15),toll\_plaza\_id smallint);

exit;

* Now install python module Kafka-python, this module helps communicate with kafka server. It can be used to send and receive messages from kafka.

pip3 install kafka-python

* Install python module mysql-connector-python using pip:

pip3 install mysql-connector-python==8.0.31

* in toll\_traffic\_generator.py file set TOPIC to ‘toll’
* run the toll\_traffic\_generator.py file using command below:

python3 toll\_traffic\_generator.py

* Now for consumer, in streaming-data-reader.py file add your topic, database, username and password, then save file and run :

python3 streaming-data-reader.py

* Messages from producer our toll\_traffic\_generator.py will be read by consumer stream-data-reader.py and will be stored in mysql database table livetolldata.