Ride-hailing services like Uber, Cabify, and Lyft rely on real-time data processing to match passengers with drivers efficiently, monitor ongoing trips, and optimize pricing strategies. These platforms generate vast amounts of streaming data, including ride requests, driver availability, trip statuses, and user feedback.

This project challenges students to **design and implement a real-time analytics pipeline** that can process and analyze streaming events in a ride-hailing system. The goal is to simulate a real-world scenario where massive amounts of data need to be ingested, processed, stored, and analyzed efficiently to extract meaningful insights.

Project Objectives

General Objective:

Develop a real-time analytics pipeline that can process ride-hailing service events and generate useful insights using modern stream processing technologies.

Specific Objectives:

Design and develop a Data Generator to simulate ride-hailing events, including passenger requests, driver availability, ride statuses, and system updates.
Implement an event ingestion pipeline using a message broker (e.g., Azure EventHub)..
Store and structure streaming data in appropriate storage solutions.
Perform real-time analytics on streaming data to generate business insights for ride-hailing services.
Develop and present a dashboard that visualizes key analytics in real time.

Milestone 1: Data Feed Generation

Design a **data generator** that simulates realistic events such as:
Passenger requests and cancellations.
Driver availability updates.
Ride statuses (Accepted, Ongoing, Completed).
Traffic conditions and surge pricing alerts.
Choose **two different data feeds** from the list above or propose alternative ideas.
Define an **AVRO schema** for the two data feeds and produce event data in **JSON and AVRO formats**.
Ensure schema design supports useful analytics and insights in later milestones.
Hint: Explore Kaggle and other dataset sharing sites to inform realistic data generation that supports relevant stream analytics scenarios

Script Features:

Realism and Variety: The scripts should generate data that mimics real driver and passenger behavior patterns.
Scalability: Simulate varying loads to test system scalability.

Configurability: Allow parameters such as active drivers, active riders, and pricing models to demonstrate various use cases (e.g., high demand period, low demand period, vehicle types, pricing differences...)

Tools for Synthetic Data Generation:

AVRO Serialization: fastavro for handling data serialization in AVRO format.
Python Datetime and Time libraries: Essential for creating realistic timestamps and durations. Explore libraries such as datetime, dateutil, arrow or the power of Pandas for time series functions....
Data Manipulation: Use numpy and pandas for calculations and structuring data.
Synthetic Data Generation: Use Faker or Mimesis for generating realistic data.
Generative AI: Consider using AI models to create complex, realistic datasets (e.g., synthetic user comments, interaction patterns).

DELIVERABLES

Upload your Milestone deliverables:

GitHub Repository: Invite the instructor to a private repository containing notebooks, scripts, event data, and project artifacts. Include the URL to your repository in this submission too.
Presentation Document: ATTACH A .PPTX DOCUMENT. NO OTHER FORMAT WILL BE ACCEPTED.
Data Feed Design (Milestone 1)