Motorq Data Science Assignment 2025

In this task, you will be provided with a dataset that contains various anomalies and issues which we commonly encounter in real-world vehicle telemetry data. Your objective is to clean the dataset by addressing the issues described below. You are expected to manipulate and transform the data as required for each of these tasks.

## Dataset description

Each record in the dataset is a JSON object containing vehicle telemetry data. Following are the relevant fields present in the JSON object

1. VIN: This is the Vehicle Identification Number that uniquely identifies a vehicle
2. statusTimestamp: This is the timestamp at which all the other parameters in the JSON were recorded for the given vehicle
3. Each parameter is stored as an item in the statusData List
   1. attribute: Name of the parameter recorded
   2. value: Value of the parameter recorded
4. pollerFetchTime: This is the timestamp at which we received the record containing all the data from the vehicle

## Data Cleaning Tasks

1. **Out of order data**

Motorq receives the messages as a stream of data. Ideally, all of these messages would be fetched one by one in order of their statusTimestamp. However, we sometimes receive a message with an older statusTimestamp *after* we have already received a different message with a newer statusTimestamp.

Impact of these out of order messages cold be severe, in some cases if the system is not designed properly, it could lead to down time. Identify and delete these out of order messages i.e. the messages which have an older statusTimestamp than any other message that has already been fetched before the time this message is fetched.

Example, if you receive a message with statusTimestamp of 2025-07-01 13:50:42.654, you should drop all the messages which are polled after this with statusTimestamp less than 2025-07-01 13:50:42.654

1. **Duplicate/Conflicting Messages**In cases where we have multiple messages with the same statusTimestamp, only accept the one that was fetched first and delete the rest.

Impact of this could be conflicting messages, when you want to check for last known message, you will end up with multiple messages with same timestamp

Example, if you receive a message with statusTimestamp of 2025-07-01 13:50:42.654, you should drop all the messages which are polled after this with same statusTimestamp i.e. 2025-07-01 13:50:42.654

1. **Odometer Anomalies**The odometer values are expected to be non-decreasing since the distance driven by a vehicle can only stay the same or increase over time. Identify and delete the records where the odometer value is decreasing.

If you complete these tasks, please reach out to us for a follow-up.