**A picture containing text

Description automatically generated**

**A01 – REPORT.**

# Description and Relevance

Bus routes are planned to take into consideration traffic patterns, patterns of movement etc. Over a period, these assumptions change. Traffic in one part of the city may increase while traffic in another part of the city may decrease. Also, the distribution of traffic throughout the day may change. All this implies that bus schedules need constant updating to reflect the current conditions.

One way to approach this problem is to consider every one-hour window and percentage the ratio of times buses has been late at every stop.

The routes and one-hour windows where we have the highest percentage of buses reporting late will be the prime candidates for replanning.

This dataset allows us to compute this information.

In this problem, we consider a bus late when it is late more than a threshold (set at 5 minutes by default).

Also, we only consider those routes where buses have been late by at least a minimum number in a month, currently set at 100. This is done to reduce noise.

Finally, we want to only display the routes for which the percentage of times they were late is bigger than a threshold, and sort them in descending order.

# Novelty

This problem is different from the four other problems in the assignment.

In the first problem, we take spatial rectangle based on the latitude and longitude. Then for every hour, within this rectangle, we calculate how often congestion is reported.

In the second problem, we want to calculate the timetable of a physical vehicle, as a pair of lineID and stationID, for a given day.

In the third problem, we compute the station that has the highest number of buses stopping at it so that we can use that station for best reach of advertisements.

In the fourth problem, we compute the distance travelled by individual vehicles so that we can find out when to send them for service.

This new problem is different from all others because we compute how many times a route has reported delays at stations, and then aggregate them by route and hour. This is not done for any of the given problems.

# Technical Discussion

To do this task, we must do the following steps:

1. Load the dataset
2. Filter by the month to reduce the dataset
3. Discard any rows where the bus is not stopped at a station
4. Calculate the number of times buses stops at a station for every route and hour
   1. We must exclude consecutive entries where a bus continues to stop at a station, and only take one of those entries
5. Calculate the number of times a bus stops at a station and is also late for every route and hour
   1. Again, we must exclude consecutive entries where the bus continues to stop at the same station and only take in one of those entries
   2. Discard any entries where the number of times a bus is late is less than the specified threshold
6. Perform an inner join of the results from Step 4 and 5 on the route and hour and calculate the percentage of times
7. Discard any rows where the percentage calculated in step 6 is less than the threshold
8. Sort in descending order and display