Junior Data Engineer Case Study

# 1. Data Modeling/ SQL exercise (90 mins)

As a DE for the “supply” domain, you are asked to get timetables information about SNCF Intercity trains. Based on the information/ model which you make, data-analysts will be able to build reports.
Your main stakeholders are **data-analysts**.

> Important
> - For the database design you can use any tool (eg: SQL workbench, Google draw, etc)
> - Provide complete documentation about your tables/ models.

You will use the following information:

1. Read about, *Timetables of SNCF Intercities lines* (link)
2. Download the dataset (link) [Don’t forget to unzip it, to get the files]

There are following files in the dataset which you should focus on:

- routes.txt
- stop_times.txt
- stops.txt
- transfers.txt (file may just have headers which are enough for table design)
- trips.txt

# Exercise

1. Build a Relational database model from the dataset
    a. Add the data types to your tables
    b. Identify the PK, FK, UK for your tables
    c. Identify the relationship between tables & cardinality

2. Convert your Relational database model to Star Schema model
    a. Bonus : Add technical columns, Index, Partition, Cluster columns

3. SQL Exercise - Based on your Star Schema model, provide sample queries for data analysts
    a. Top 10 most popular routes in August 2022
    b. Routes with number of stops in descending order
    c. Number of trips with missed transfers in August 2022

# 2. Python exercise (60 mins)

In this exercise, the goal is to write a Python script which queries an API & stores the results in a CSV file.

Important:

- Please submit the code via GitHub repository (access to be shared with interviewers)
- Documentation for executing your code

We are going to use the public API for “Transport for The Netherlands” which provides information about OVAPI, country-wide public transport (API description)
We will use the Per Line endpoint

Base_url : http://v0.ovapi.nl/
Endpoint: /line/
Authorization: Not needed

#Exercise

1. Write a Python script which queries the endpoint & extract response.
    a. Store the response in a CSV file.

About your Script:

- Use Python to complete the assignment.
- No action should be needed before running your code. Otherwise, please provide documentation to initialize your project.
- The code should be written like it is executed every day.
- Your code is working whatever the date and the number of times we run the program.
- Your code is well documented and organized