# Scenario:

You have been tasked to work with data from Homelike that collects customer event data on the Homelike Platform. The task includes building a pipeline that uses raw tracking data, performs transformations, and loads the processed data into a DB for analysis. Additionally, there is a requirement to integrate some Machine Learning and Data Analytics tasks into the pipeline.

The assignment aims to test both your coding standards, analytics, problem solving abilities and to test your understanding of database structures. The assignment should take you 2 to 4 hours to complete.

# Tasks:

### 1. Data Understanding:

* Examine the provided sample raw data ([available here](https://drive.google.com/file/d/14w2r0VT6l1b2oLipWte9mLXkh2QMryte/view?usp=drive_link)).
* Identify potential data quality issues or inconsistencies.
* Document your observations and propose solutions to address any issues found.

### 2. ETL Pipeline Optimization:

* Develop a Python-based **class** that:
  + Extracts the data
  + Cleans the data
  + Exports the results to JSON
* You can structure your class however you like.
* Ensure that the pipeline efficiently handles large volumes of data and is scalable.
* You must submit the output JSON file as well as the source code.

### 3. Database Integration:

* Design a schema for storing the processed transaction data in a database.
* Modify the pipeline code to load the transformed data into the database’s tables.
* Ensure that appropriate data types and optimizations are applied for efficient querying.

### 4. Data Analysis / Science:

* Create a Jupyter Notebook that will run analysis on the data
* Choose a KPI and integrate a simple data analysis task from the provided data.
* Integrate a simple classification ML task with the provided data.
* Utilize Python libraries (e.g., Pandas, NumPy) for data manipulation and analysis.
* Document the process and rationale behind the chosen approach.

### 5. Documentation:

* Provide clear documentation for the pipeline code, including comments and explanations of key components.
* Prepare a brief report summarizing your findings, improvements made, and any remaining challenges or considerations.

### Deliverables:

* Data observation overview in a jupyter notebook.
* Python code uploaded to Github (make the repository private and provide access to luksfarris,ioanpapa1982)
  + along with the database schemas
  + and the jupyter notebook
* Documentation outlining observations, proposed solutions, and implementation details.

### Note:

* There is no right or wrong solution, there is only your solution.
* Make assumptions where details are not specified, and document them accordingly.
* Prioritize simplicity, efficiency, and maintainability in your solutions.