## DATA ENGINEERING TASK

At this stage we would like to learn more about your analytical and technical skills. The task consists of two components:

* **Part 1: Design the data engineering** **approach for the provided task** (the data engineering task is described in the Task Description section below) – please provide your proposal for the approach in any form you choose (preferably pdf format, we suggest about 5 slides or 1-2 text document pages). Assessment will be based on the provided content, so the visual formatting is not important. This may include a short exploratory data analysis, but in general in part 1 we only expect the description of the approach, not the code to apply it.
* **Part 2: Provide a sample of your code in one of two forms**:
  + Option A: For part 2 you can provide code you have prepared in the past. This can be your private project, your university project, your public git repository etc. In other cases, please discuss with the recruiter. Ideally this should include:
    - Some basic data analysis or exploration,
    - SQL, PySpark or equivalent language code performing data formatting, cleaning, preparation, feature engineering or reshaping/wrangling,
    - Python, bash or similar language code to configure/setup some data infrastructure (for example code to setup AWS instance, load tables to database, set up cloud environment, run ETL process, workflow setup etc.)
    - \*If you have similar code you would like to present, but this code is performing something else, please discuss with recruiter.
  + Option B: If you prefer to not provide any code you prepared in the past, you can prepare code solving data engineering task (data engineering problem is described in Task Description section below).

We are interested mostly with general approach towards this problem. There is no need to spend a lot of time into preparing perfect solution (think of it in terms of hours rather than days). Feel free to suggest how you could make the solution better if you had more time for it in comments to your code or in your results summary. Code is the main deliverable here, so try to adhere to good programming practices and keep your code clean.

Provided code may be in any language of your choice or combination of few of them (preferably including some SQL, bash, Python). Please provide also short description of solution describing architecture discussing it advantages, related risks and potential extensions with codes that you used to arrive at the solution. Description part can be prepared as Markdown document, LaTeX, Word, PowerPoint etc. converted to pdf format. If for part 2 you decide to go with option B, you can create one report combining part 1 and part 2.

We hope you will enjoy solving this task.

## TASK DESCRIPTION

Imagine that in the company you are working for, the decision was made to build a model. The goal of the model would be to forecast weekly sales for the next 8 weeks. The business goal is to better understand demand which will allow to reduce waste and avoid stock outs, so the company can achieve better profits.

Your task is to prepare the dataset (link provided below) to build the model on and to set up the infrastructure to automatically prepare data for monthly model re-run:

* Prepare the provided data for modelling (joining provided tables, cleaning formats, creating additional features that might be useful). Try to include most of following additional features:
  + dataset should be aggregated to the level where each row represents sales in a specific store in a specific week (no split by department),
  + column “total sales” should be added,
  + column “total sales” from previous week should be added (in relation to current considered week),
  + minimum, maximum and sum of sales from previous 4 weeks (in relation to current considered week),
  + number of the current month (for example 2 instead of Feb),
  + columns with sales in each department should be added as separate columns (dept1\_sales, dept2\_sales … dept99\_sales),
  + all relevant columns from three datasets should be joined,
  + % markdown sales contributed to total sales (columns markdown1-5 divided by total sales) column should be added,
  + feel free to add additional columns, but it would be more important to add a small number of interesting features, then 100 similar columns,
* Design the solution of automating this code in a way such that:
  + Every week new csv file is provided (uploaded to a folder selected by you) containing data from most recent week,
  + Design a solution that would apply above operations to each such file automatically after file is uploaded and appended to the whole dataset,
* Part 1: Prepare description of the approach to conduct the task that includes:
  + what types of languages/statements/syntax/libraries you would use to conduct required task,
  + what infrastructure you would use in the automation part,
  + how would you set it up for the automation part,
  + potential limitations and risks of the solution,
* Part 2 (if you don’t want to provide another sample of your code):
  + please provide the most important parts of the codes required to achieve this (bash scripts, SQL or whatever is required)
  + some parts of the code (like potential configuration of AWS, database etc. can be described in words if preparing code would be difficult to do without provisioning actual resources/infrastructure),
  + feel free to make any assumptions (on the day of the month when data is uploaded, database/cloud company is working with etc.),

## DATA DESCRIPTION

You are provided with historical sales data for 45 stores located in different regions - each store contains several departments. The company also runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. There are 3 files – Stores, Features and Sales. Values in some columns might look weird, they might be rescaled for the dataset to be better anonymized.

Link: <https://www.kaggle.com/datasets/manjeetsingh/retaildataset> (use download button on the right, you would need to register on Kaggle.com, alternatively, we can share datasets with you)

### Stores

Anonymized information about 45 stores, indicating the type and size of the store.

### Features

Contains additional data related to the store, department, and regional activity for the given dates.

* Store - store number
* Date – week number
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011 and is not available for all stores all the time. Any missing value is marked with an NA
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

### Sales

Historical sales data, which covers 2010-02-05 to 2012-11-01.

* Store - store number
* Dept - department number
* Date – week number
* Weekly\_Sales - sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week