# PROJECT 2

The goals of this project are for all of you to work together in order to successfully transform raw data into meaningful pieces of information.

Therefore it is a **MUST** for all of you to collaborate using Git. Gitlab would be preferable, but you can also use Github and if you use Azure services then Azure DevOps is also acceptable. Your git commits and how you have collaborated with each other will be evaluated.

You will be evaluated both as a group and individually. If you work and complete the bonus points, you will be awarded extra points. If you do not, you will not be penalized.

The format in the medallion architecture will be CSV (if done locally. Bronze can contain the raw data in their original format).

All the scripts need to be saved in Git.

Finally all decisions you have made in regards to the code, architecture or in any other aspect of the project needs to be explained.

## Exercise 1

Create medallion architectures for the files:

<https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset>

<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>

The architecture consists of the 3 layers:

* Bronze
* Silver
* Gold

They can be separate folders. Use pandas, pyspark and polars (1 library for each dataset. You can also use a combination of them and use other libraries as you see fit). If you see the same file in different formats, use only one of the files. You can extract all of the files into a single folder named Bronze.

You need to

* Clean and transform the data from bronze to silver. You decide on what is needed to transform the datasets. You can drop rows and columns or add columns from the files as you see fit. In addition to this other transformations may apply (how to deal with NULL values, whether to drop those rows or fill them with an arbitrary value). Also check if you think data needs to be normalized in these files and if joins may be useful.
* Create 1 or 2 gold files for each silver file that you have produced

You will be evaluated on the choices that you make as a group.

**Bonus**: Create the same medallion architecture using Azure Data Factory (If you can configure databricks that is even better! And you can use only PySpark for all 3 datasets).

## Exercise 2

Create a Data warehouse in MySQL. Star schema or Snowflake. The data warehouse will have 2 Layers (stage and hist) of tables (at minimum. You may add an additional layer called raw which may hold the initial data from the raw files). You can decide if the layers in different databases (database raw, stage, hist) or different schemas within the same database ([raw],[stage],[hist]).

The dataset that will be used for the data warehouse is:

<https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews>

You need to study and understand the data beforehand.

Your tasks are:

* The data entering the stage should be normalized and cleaned (This can be done using a store procedure from raw to stage or using a python script on the files before their insertion into the stage)
* Decide which tables are dimensional and which tables are fact tables
* Stage tables should be transferred to the hist layer using store procedures
* Dimensional tables should be type 2 slowly changing dimensions
* Create 3 triggers on after update, after insert and after delete on any tables in the data warehouse
* Create 3 views on the hist layer of the data warehouse that filter data
* Create 3 views on the hist layer of the data warehouse that aggregate data
* Create 3 different store procedures that perform a particular task in the data warehouse

**Bonus**: Create the same data warehouse using Azure SQL Server.

## Exercise 3

Create a MongoDB database with the following dataset:

<https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam>

Each of the files represents a collection. The creation and insertion of these documents to the database should be done via a python script.

You need to also create

* 6 views in which a filtering is performed
* 6 views in which an aggregation is performed
* Indexes on 6 different collections that would make queries more optimal

Explain your choices.

## Exercise 4

Create reports from the gold layer files using Matplotlib and/or seaborn

<https://matplotlib.org/>

<https://seaborn.pydata.org/>

Create the following charts:

* 3 Line Plot
* 3 Bar Charts
* 3 Pie Charts
* 3 other chart types

For each gold file that has been created by the medallion architecture in

**Bonus:** Create the same reports also using **Streamlit:** <https://streamlit.io/>