**CIS9440 Spring 2024**

**Homework 1**

**Data Requirements**

You are required to choose your own data for this homework.

* Your data should not come from Kaggle, or any main data source included in term project.
* Your data should not be the same as your team member in your term project.
* Your data must contain at least 10 columns and more than 7500 rows.
* Your data should not contain aggregate data. However, if there is no other way, you must consult with the instructor first.
* Your data should not be stock market data or bitcoin market data.
* You will be using the same data for all subsequent homework. So, it is important that you take your time to choose the right data. If you want to change data at a later stage, you are free to do so. However, you will have to start the whole homework 1 and it will not be graded.

**Data Sourcing**

Your first step with the project is to get familiar with the data. You need to understand how it is structured and most importantly find the data dictionary associated with it. If it is not there, then you will have to build a data dictionary. The latter should contain the name of the field, the description, the datatype and any constraints associated with the field.

You will need to source those data using one or more of the following methods:

* Web Scrapping
* Web API
* Connection to Database
* Connection to Data Store (Cloud Storage)

**Deliverables 3pts**

1. Link of all data sources
2. Explanation of the data (where does it come from)
3. Link that shows the data dictionary (excel, google sheets)
4. Github/AzureDevops/Jira account created
5. Scripts that gather these data
6. Git Repository Created
7. You script should be stored in a git repository that is accessible to all members of your team and the professor.

**Storage**

Your next step is to choose the appropriate data store for your data. Remember in the previous step, you had to source the data using a script or a specific tool. The data stores of choice are the following Database, Storage S3, MongoDB. Make sure the data are properly stored and not scattered. If need be, you will need also to mark the date the data was stored as well. It is recommend you watch the async videos.

**Deliverables 3pts**

* Storage of choice
* Data Stored in an orderly fashion in the storage.
* Scripts updated from the first deliverables. You will need to update those scripts to store to data into their specific storage.
* Git repository updated.

**Modeling**

Once you have done the storage, you will need to start the modeling of the data warehouse. Remember the Data Warehouse contains already two main aspects. A fact table and a dimension table. The fact table must have a surrogate key as well as each dimension table. Modeling can be done using any tools.

**Deliverables 4pts**

* Data Model documented showing the fact table and the dimension table.
* Scripts that create the Data Warehouse
* Scripts from previous steps updated.
* Data Warehouse accessible to everyone in the team and can’t be accessed through a client (DataGrip, DbSchema, SqlDBM)
* Git Repository Updated

Notes :

**All deliverables mut be in the Github/Azure Devops/ Jira Board (this is your final report) with all the tasks you have executed including a readme that explain the project.**