**CIS9440 Spring 2024**

**Homework 2**

**Homework Steps**

You are required to choose the previous data from homework 1. If you want to change your data, you are free to do so. However, you will have to start the whole homework 1 and it will not be graded. You are free to use any cloud provider. You are required to check the feedback from the professor.

**Transformation**

Once you have stored the data, the next steps would be to transform the data. Data should be transformed according to specific business rules. While transformation the data, you should consider the following.

* Unified date format YYYY-MM-DD
* Splitting the date into multiple unit (Year, Quarter, Month, Day, Hour, etc…)
* Removing NULL values if necessary
* Removing Duplicates rows if necessary
* Verify Data against data reference (currency, state, zipcode, county, NAICS, GICS, etc)
* Use the correct data type each new fact generated.
* Adding one or many columns
* Summing two or more columns
* Create a Data Mapping that will serve be incorporated into your data dictionary tools.
  + It should contain the name of the fields, their data type, their description, the source column and the destination column.

This is only a limited version of what you can do. There is more to that. Remember also to update your data dictionary.

You have the following options:

1. Use a transformation tool ETL tools to do the transformation.
2. Create scripts that does this transformation.
3. Git repository Updated.

**Deliverables 5pts**

* Scripts from previous steps updated.
* Data Mapping Created. Data dictionary updated.
* Data transformation project created on the cloud in case you are using Option One
* ETL fully created to push the data to DataWarehouse
* Git Repository Updated

**Modeling 1pts**

Once you have done the transformation, you will need to update 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. Your data warehouse should be in Redshift.

**Deliverables**

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

**Serving Data 5pts**

You will be using an online visualization tool to show the data that you have transform. You should apply all the visualization practices you have seen in all sessions. The following must be part of the Visual:

* A Filtering tool by date or by dimension: When you filter by date, all charts should change based on the filter
* A Pie Chart
* A Column Chart
* A Line Chart
* A Heat Map

As part of the service DATA as well, you will need to create a api that will generate a csv file that contains a summary of the data. This is optional.

**Deliverables**

* Git Repository Updated
* A link using AWS Quick Sight that connects to the Data Warehouse and shows the data.
* An API using python script that will generate data in a csv format which can be used by Data Analyst.
* A PowerBI or a Tableau Link
* 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.**