Data Project Specification
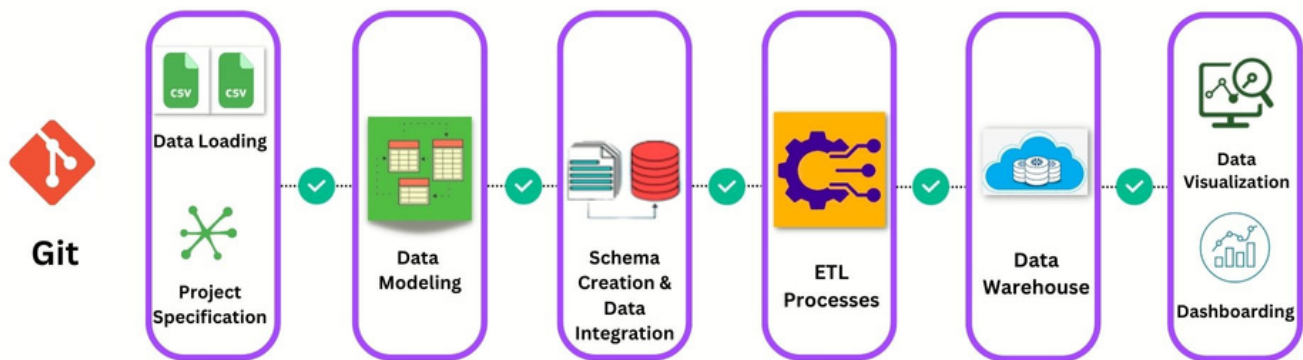
National School of Applied Science Al-Hociema

Provided by
DATA Club -
ENSAH

Project tasks:

- Clone the GitHub repository.
- Data Modeling (CDM & ER Model)
- Schema creation & Data Integration
- ETL process
- Data Warehousing
- Data Analysis using SQL queries
- Dashboarding

This pdf will be updated each week

Project description

This project is dedicated to data engineering students. The main idea is to provide you as a future data engineer with basic knowledge about databases and the process that the data get throw so we can use it to extract or extrapolate insights from it.

In this project, we will be working on DVD rental data representing the business processes of a DVD rental store. The data has been collected in CSV files and your job as a data engineer is to build and maintain the data system for the company so they can store it and do an analysis on it. So, we will develop a data warehouse that allows us to analyze data from a DVD rental store. We will start by creating a conceptual data model (CDM) and an entity-relationship (ER) model to represent the key entities and relationships in our rental data. We will then design and implement a relational database management system (RDBMS) using SQL to store our data.

Once we have our database in place, we will write an ETL (extract, transform, load) job to populate it with sample data. We will also explore the concept of data warehousing and design a star schema to optimize our data for analysis.

Finally, we will create interactive dashboards using popular data visualization tools to analyze our rental data. Along the way, we will practice time management and develop teamwork skills by working collaboratively in a small group.

What you will learn during this project:
- Time management
- Developing Teamwork skills
- Data modeling (CDM model, ER model)
- Relational database management systems
- SQL queries
- Write ETL job
- Data warehousing concept (star schema)
- Dashboarding & data analysis (techniques, tools)

Without any further talk, let's dive in.

Task 01: (Clone the Github repository)

The main goal of this task is to get familiar with **Git** and **GitHub** because **Git** is a powerful version control system that is widely used in software development and beyond. It allows you to keep track of changes to your codebase and collaborate with other developers effectively.

So, the first thing to do is to **download Git** on your computer.
Here is the link: https://git-scm.com/download/win
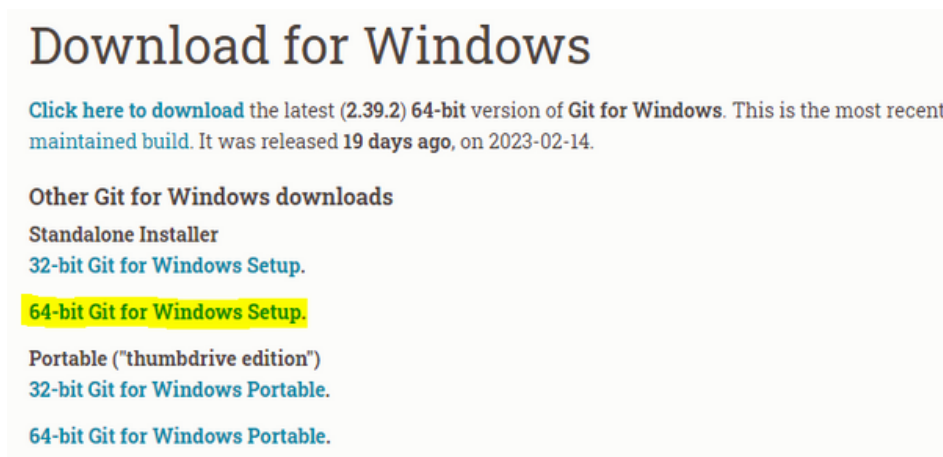
After installing Git on your computer, go ahead and open the **GitHub repository**:
 https://github.com/DATAClub23/Data-Engineering-Project-I

If you don't have a GitHub account, create one because you will need one sooner or later. It is extremely important.
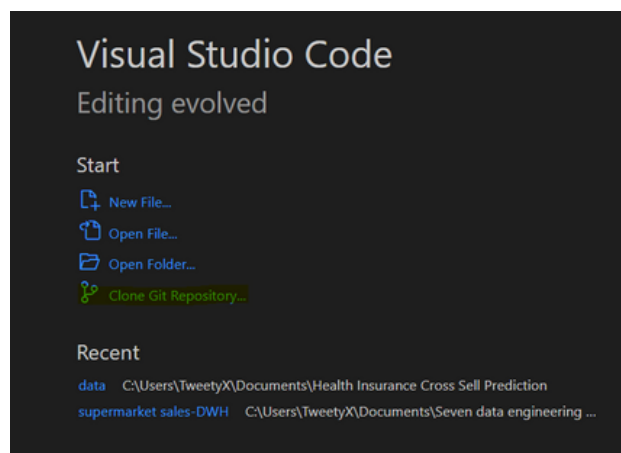
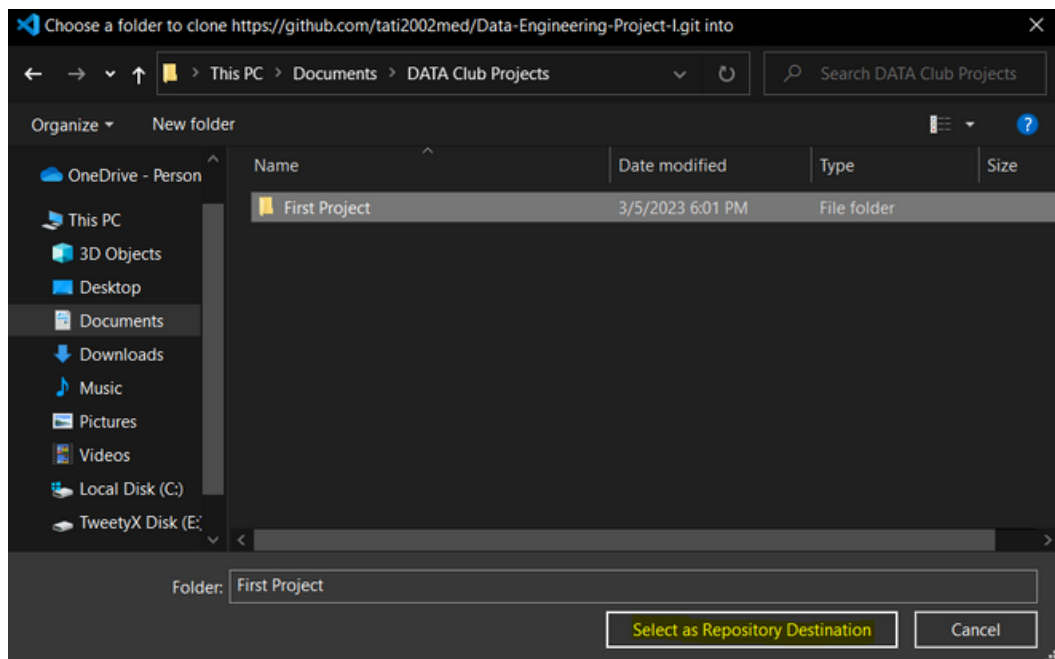Now, you need to fork the repository to your account to start working on it.
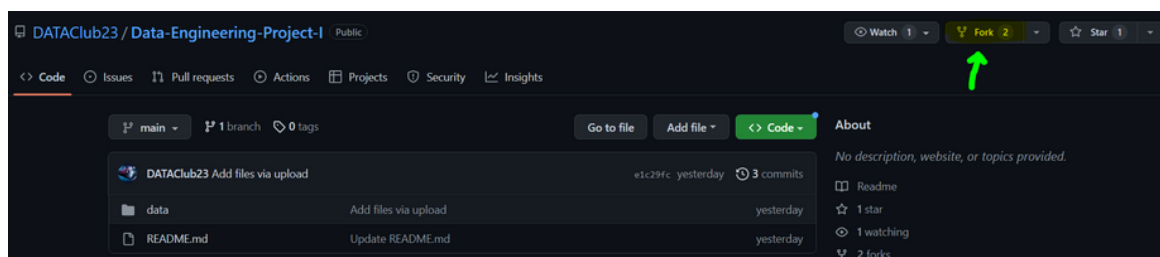
next, go to the repository that you have forked and copy the link to clone it in your **VS Code**.
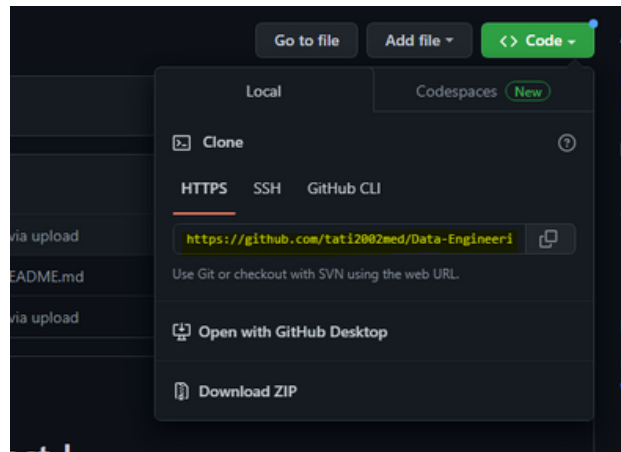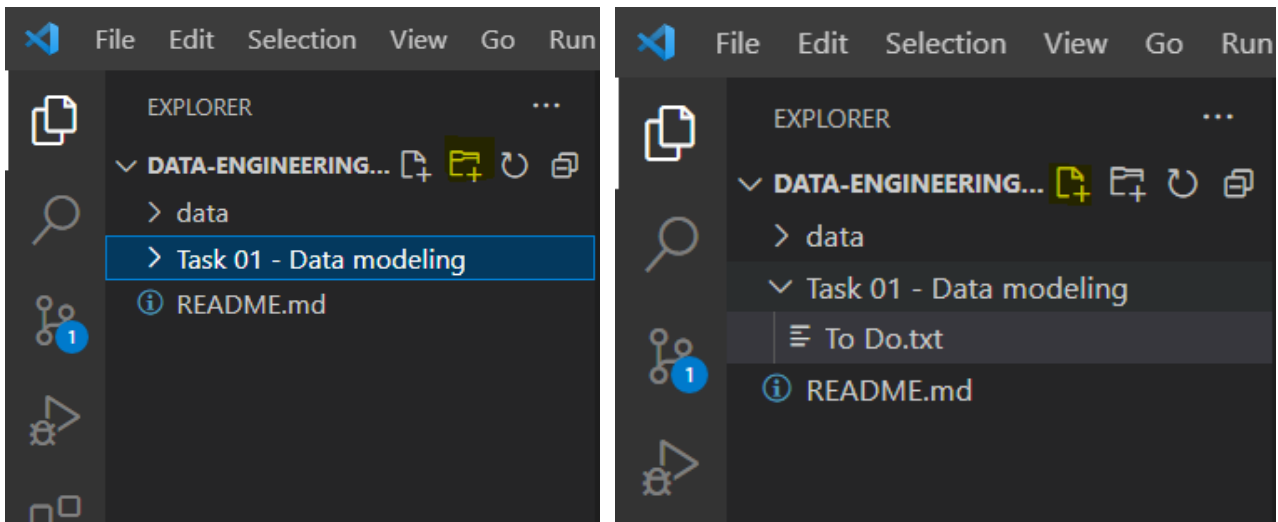
Open **VS Code** and click on **Clone Git Repository**:

Now, just **paste** the link & select the **destination folder** in your local machine:

Now, create a folder named: **Task 01 - Data modeling** & inside it create a To Do list TXT file to organize your work and schedule your tasks.
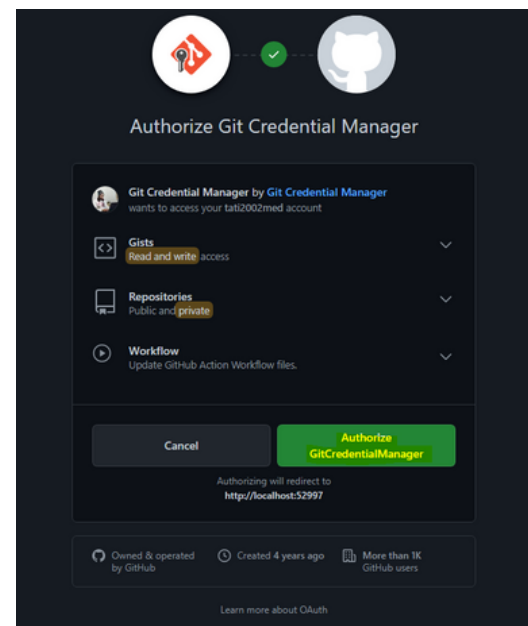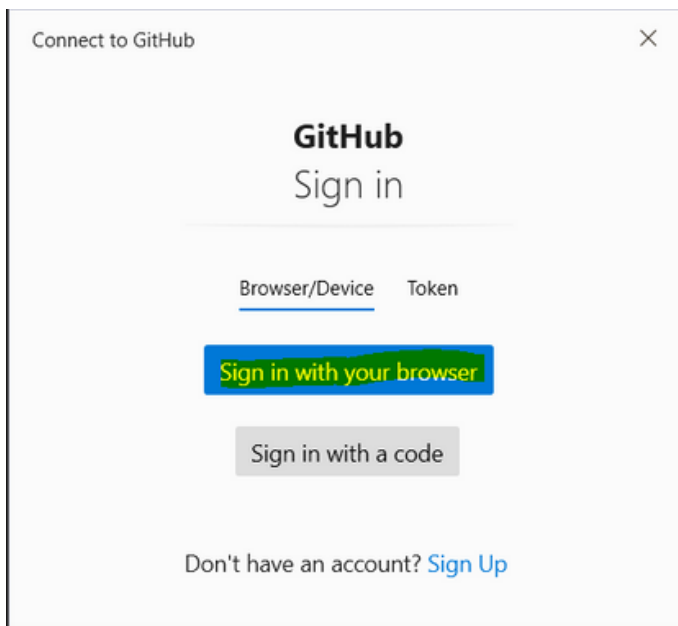

The most important thing is to control your work. After making the changes we need to commit them so our **GitHub repository gets updated automatically**. But, after that, we need to make some configurations. Go & open **Git Bash** and write the below commands (*Enter your GitHub email and username*):
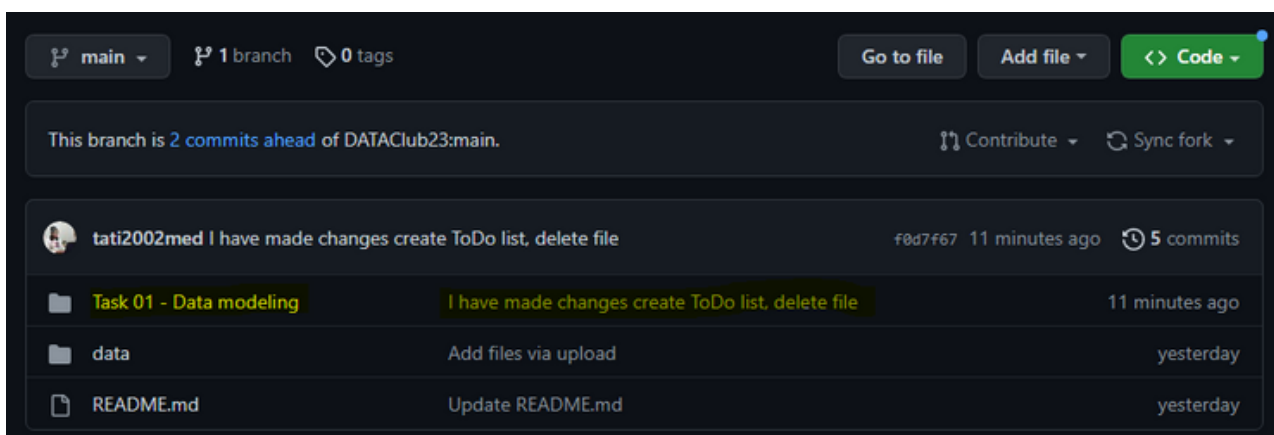

Let's now just commit the changes that we did:
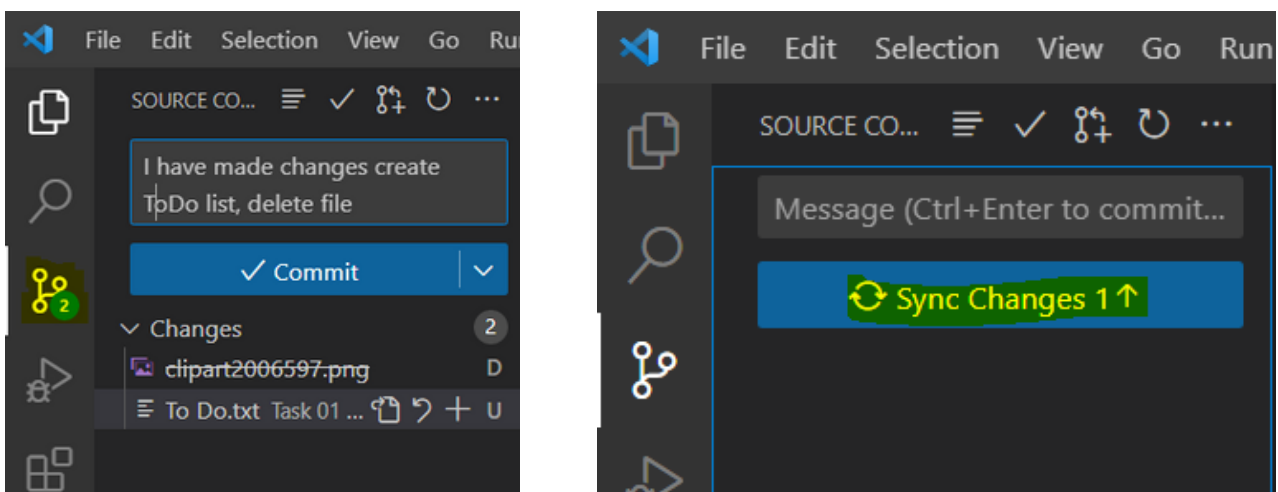
Sign in & Authentication:

The last step is to refresh your GitHub repo:

Everything is ready, Now let's start the first step in data modeling. CDM & ER models.

Task 02: Data Modeling (CDM & ER Model)

To create a conceptual data model for the DVD rental database, you can start by identifying the main entities and relationships within the system. The relationships between these entities can be identified based on the data attributes that they share. For example, the Customer entity has a relationship with the Rental entity, as each rental record is associated with a customer who rented the movie. Similarly, the Film entity has a relationship with the Actor entity, as each movie can have one or more actors. So read carefully.

Your manager has assigned you the task of creating a data model (CDM & ER model) for the company. The data has been gathered and stored in CSV files, and as a data engineer, your responsibility is to design and manage the data system that will enable the company to store this data. This entails analyzing the requirements, designing the conceptual data model, implementing the entity-relationship model, setting up a database management system, testing and validating the database, and maintaining the database to ensure it meets the company's changing business needs.

The database is comprised of several tables, each with a unique identifier and a set of columns that describe various aspects of the company's operations.

The "Actor" table contains information about the actors that have appeared in the movies. The table includes a unique identifier for each actor and their first and last names and the last_update column records the most recent date and time that the record was modified.

The "Film" table includes movie information in the rental company's collection. This table includes a unique identifier for each film, the title of the film, the description, the year it was released, the language, the rental duration, the rental rating, replacement cost, special features, full text, the last update, the length in minutes, and the rating. We should know which actors have appeared in which films, also the movies available for rental at each store and the rental company's physical stores.

The "Category" table includes information about the different categories that the movies in the rental company's collection belong to. This table includes a unique identifier for each category and the name of the category. We should know which films belong to which categories also when the last_update was made.

The "Rental" table includes information about each individual rental transaction. This table includes a unique identifier for each rental, the date the film was rented, we should be able to know the inventory item that was rented, the

customer who rented the film, the staff member who processed the rental, the date the film was returned, and the date and time when the rental information was last updated.

The "Payment" table includes information about each payment made by a customer for a rental. This table includes a unique identifier for each payment, as well as the customer and staff member associated with the payment, we should know the rental for which payment was made, the amount of the payment, and the date the payment was made (payment_date).

The "Staff" table stores detailed information about the employees who work at the DVD rental store. It includes the unique identifier for each employee, as well as their first and last names, physical addresses, email addresses, and the store location where they work. The active column indicates whether an employee is currently employed, and the username and password columns store login credentials for the staff member. Lastly, the last_update column records the most recent date and time that the record was modified and a picture.

The "Customer" table includes information about the customers who rent movies from the rental company. This table includes a unique identifier for each customer, we should know which store where the customer is registered in, their first and last names, email, address, city, country, create_date, and last_update.

The "Address" table holds data pertaining to the physical addresses of customers, stores, and staff members. This table includes a unique identifier for each address, the address line 1 and 2, the district or neighborhood, the city of the address, the postal code field that stores the postal code for the address, the phone number associated with the address, and the last_update column records the date and time that the address information was most recently modified.

The "City" table includes information about the cities where the rental company operates. This table includes a unique identifier for each city, the name of the city, and we should know the country of the city. Lastly, the last_update column records the most recent date and time that the record was modified.
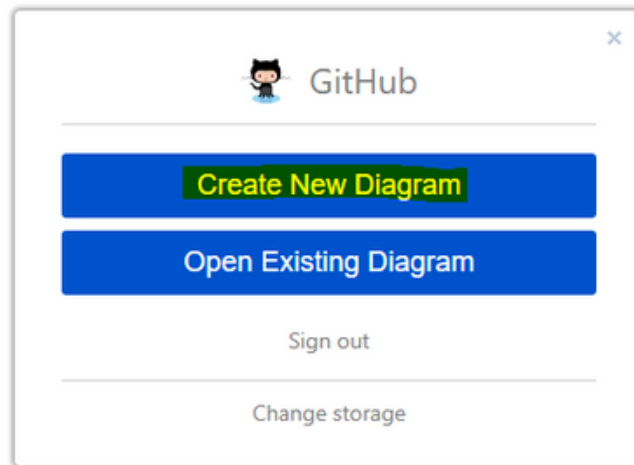
The "Country" table includes information about the countries where the rental company operates. This table includes a unique identifier for each country and the name of the country also the last_update.

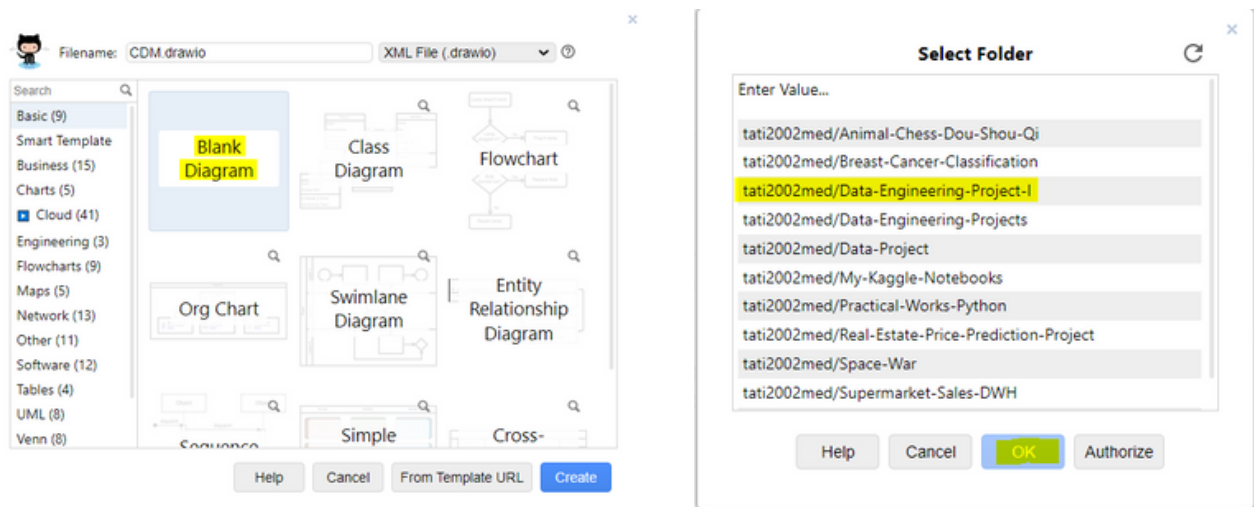**Once you finished the CDM model convert it to ER model.**

How to get the job done

Go to the link below:
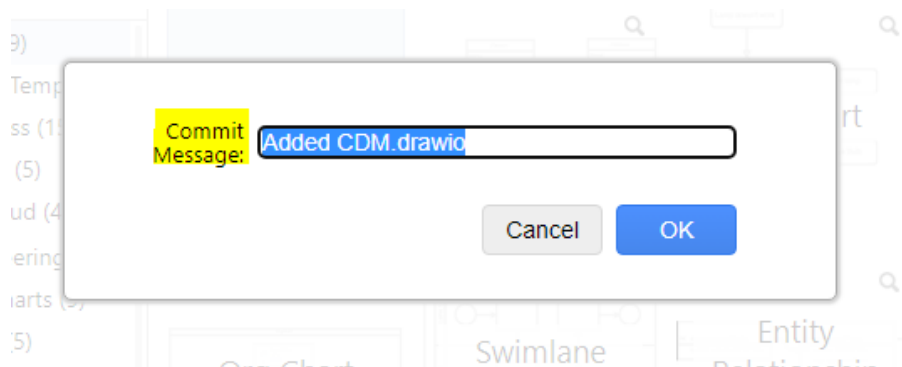https://app.diagrams.net/?mode=github

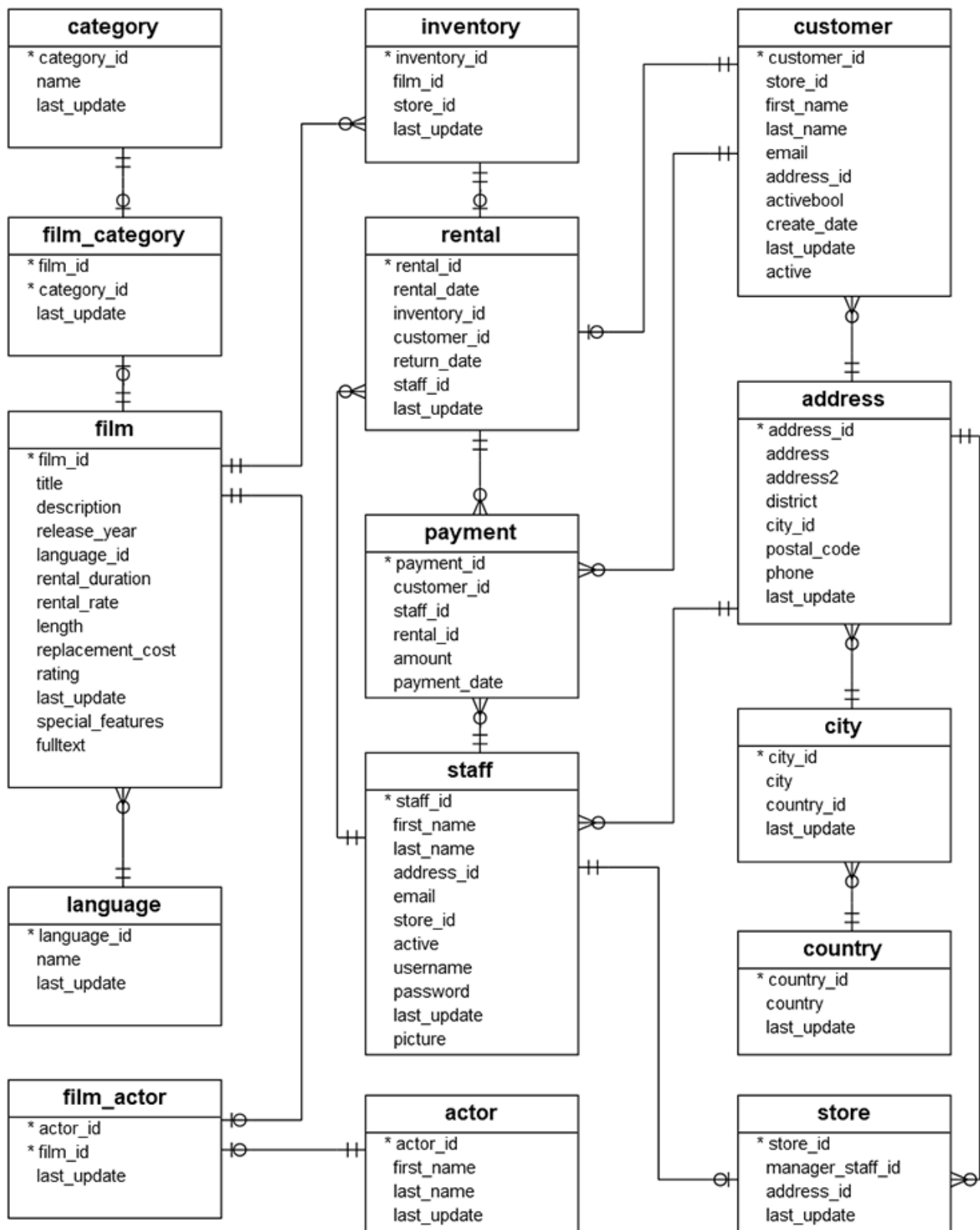Now, create a Blank Diagram & select the GitHub repository:

Commit:

Now, you can start making your CDM model. Do the same to ER model just change the Blank Diagram to Entity Relationship Diagram.
Don't forget to export your CDM & ER model as JPEG by clicking File > Export

Task 03: Schema creation & Data Integration (Python)
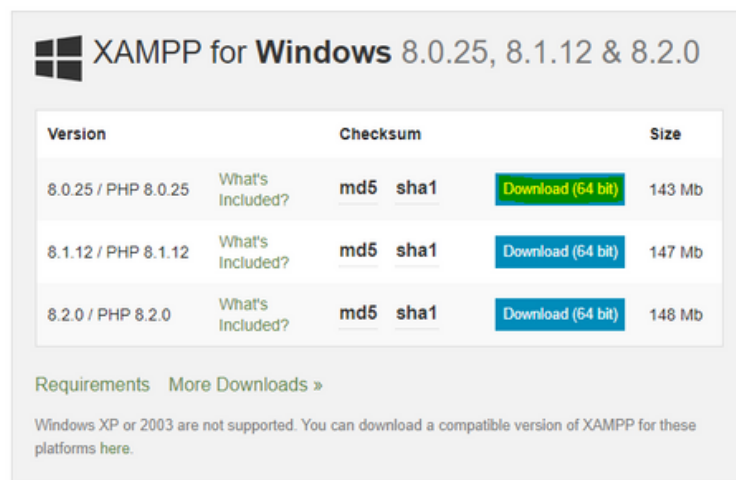
The ER model provided by **PostgreSQL**: (Solution)

Congratulations on finishing the past tasks!

After you have finished the ER model, it's time to implement it into a database. We have chosen **MySQL** as our RDBMS.

First of all, let's set the work environment. We'll be using XAMPP to make things easier. To download XAMPP hit the link below:
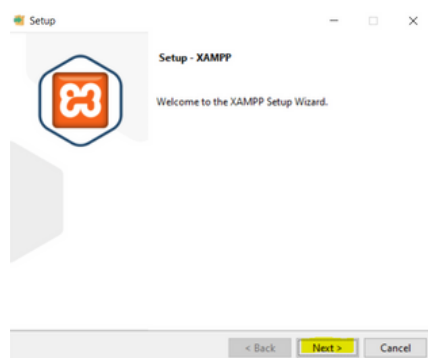https://www.apachefriends.org/download.html

Just download the latest version
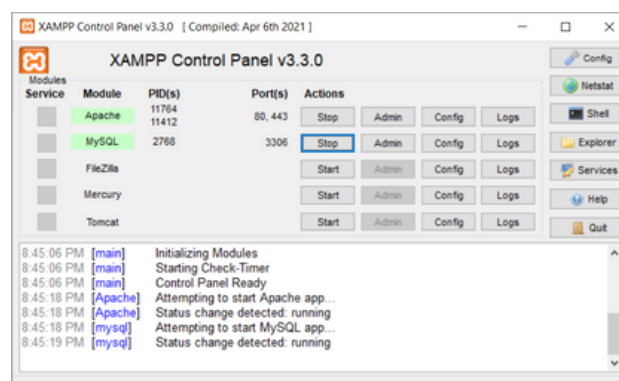

Installing:
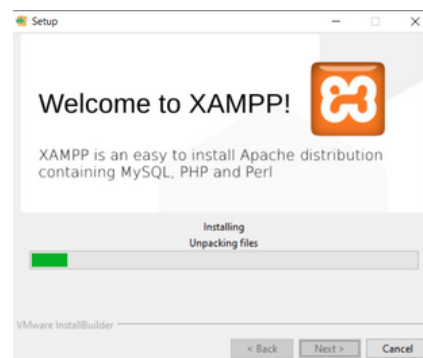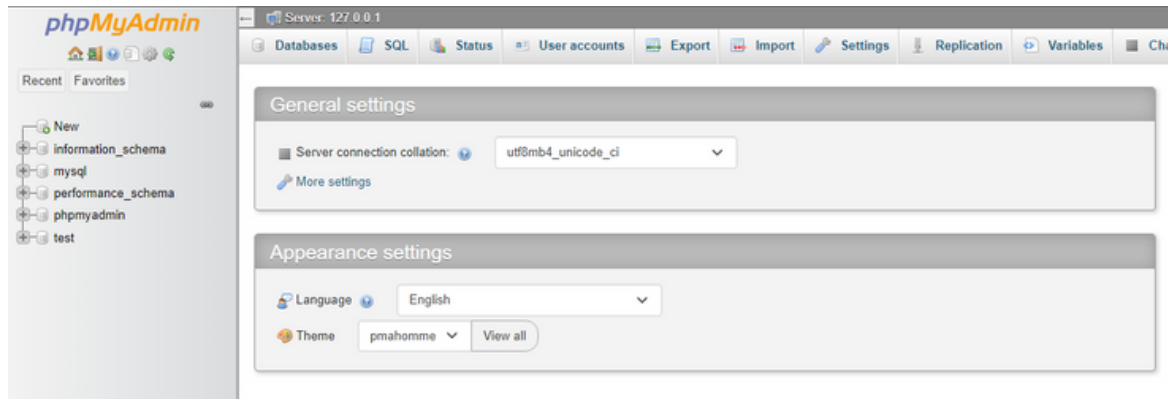

Start Apache & MySQL and that hit the Admin button of MySQL service:

This is the PhpMyAdmin:

Here is an example of how to do the task. **But feel free to use your method and your problem-solving skills**:

https://github.com/tati2002med/Schema-creation-data-integration-Example

The data will be in a folder called "data", your task is to create the database from the ER model provided by PostgreSQL and integrate the data into it.