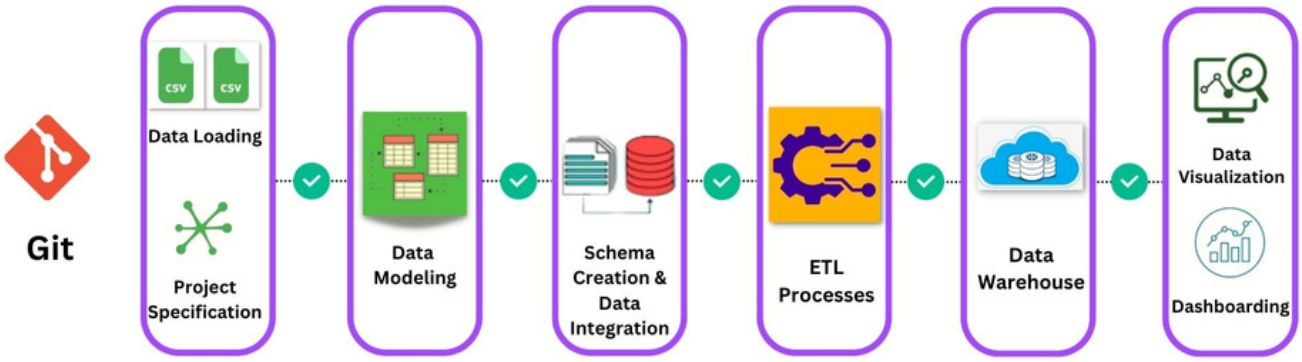
Data Project Specification

National School of Applied Science Al-Hociema

05-MAR | 2023



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:

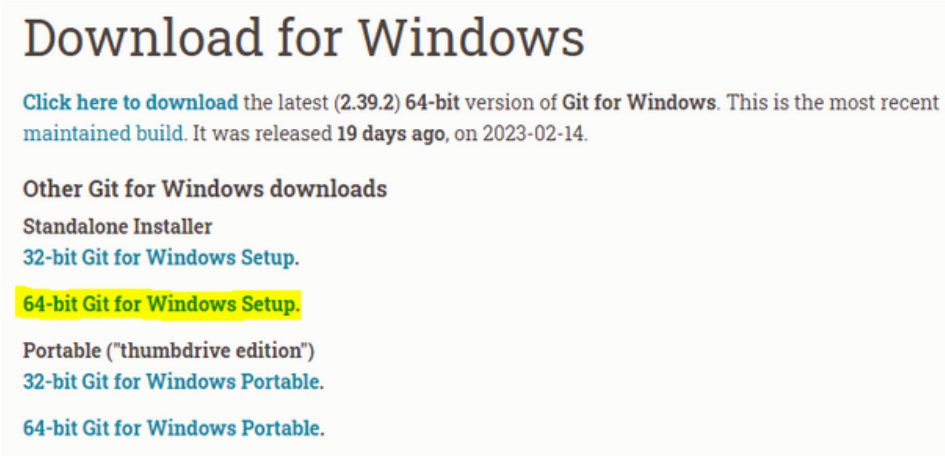
- Time management
- Developing Teamwork skills
- Data modeling (CDM model, ER model)
- Relational database management systems
- SQL queries
- Write ETL job
- Data warehousing concept (star schema)
- Dashboarding & data analysis (techniques, tools)

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.

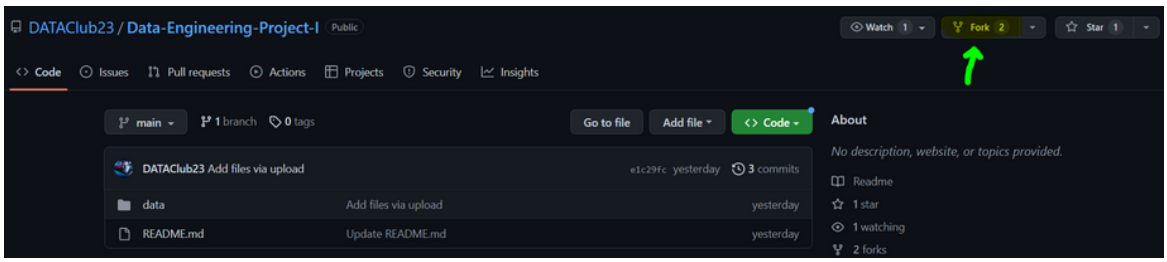
So, the first thing to do is to **download Git** on your computer.
Here is the link: https://git-scm.com/download/win



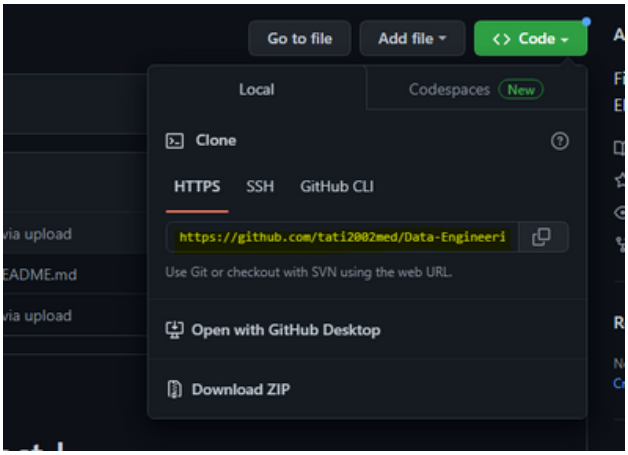
After installing Git on your computer, go ahead and open the **GitHub repository**:
https://github.com/DATAClub23/Data-Engineering-Project-I

If you don't have a GitHub account, create one because you will need one sooner or later. It is extremely important.

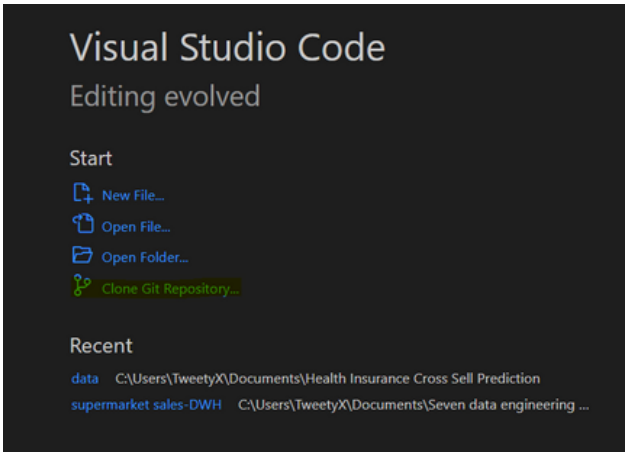
Now, you need to fork the repository to your account to start working on it.



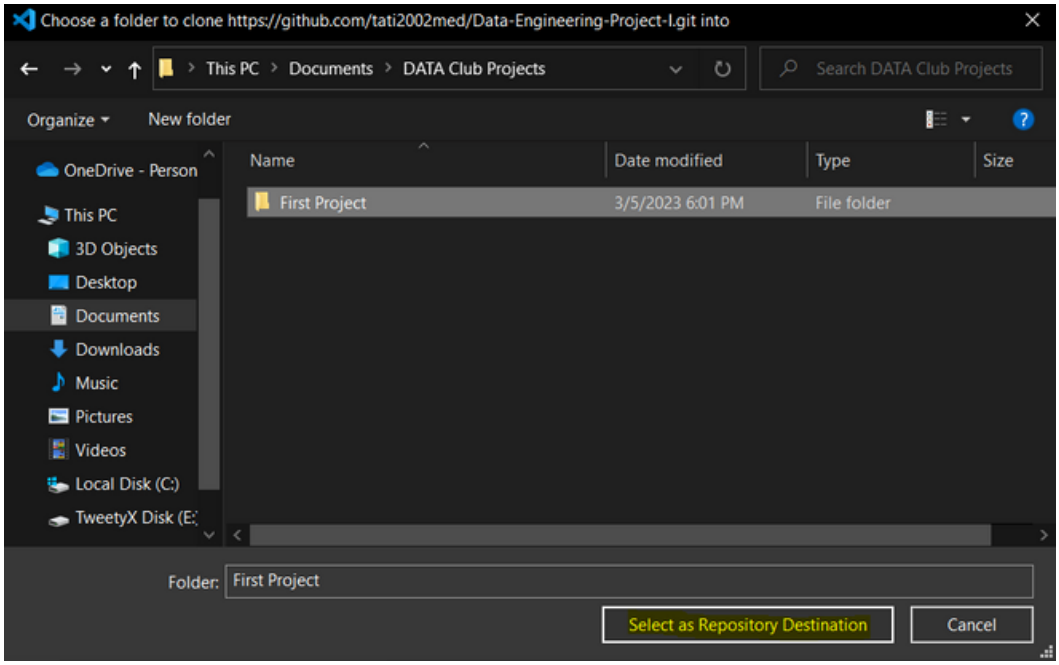
next, go to the repository that you have forked and copy the link to clone it in your **VS Code**.



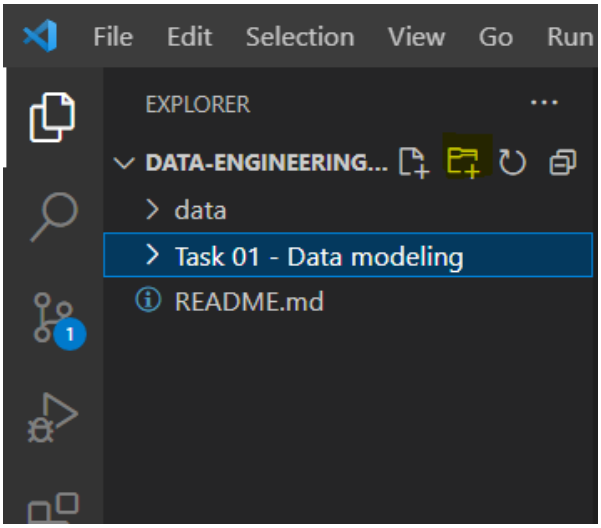
Open **VS Code** and click on **Clone Git Repository**:

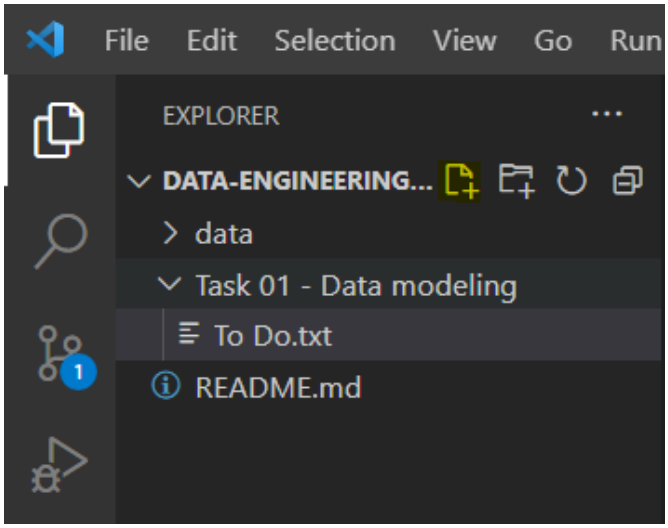


Now, just **paste** the link & select the **destination folder** in your local machine:



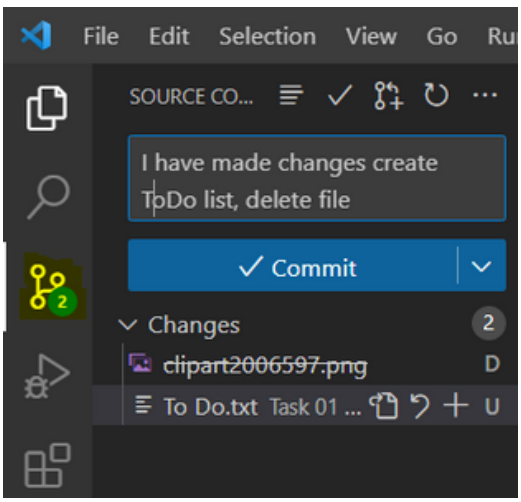
Now, create a folder named: **Task 01 - Data modeling** & inside it create a To Do list TXT file to organize your work and schedule your tasks.

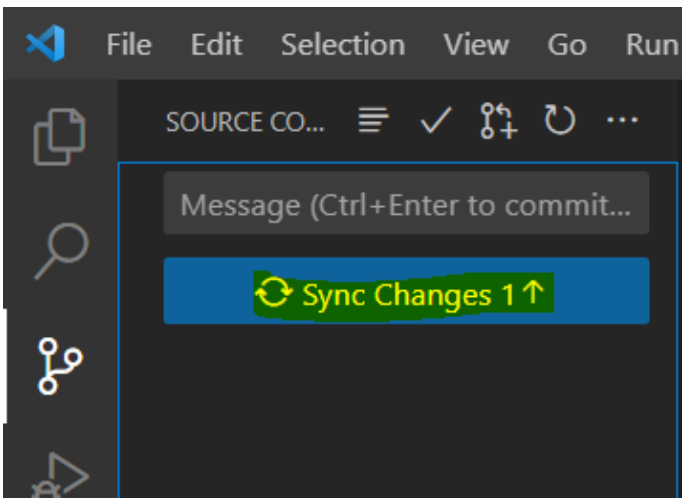




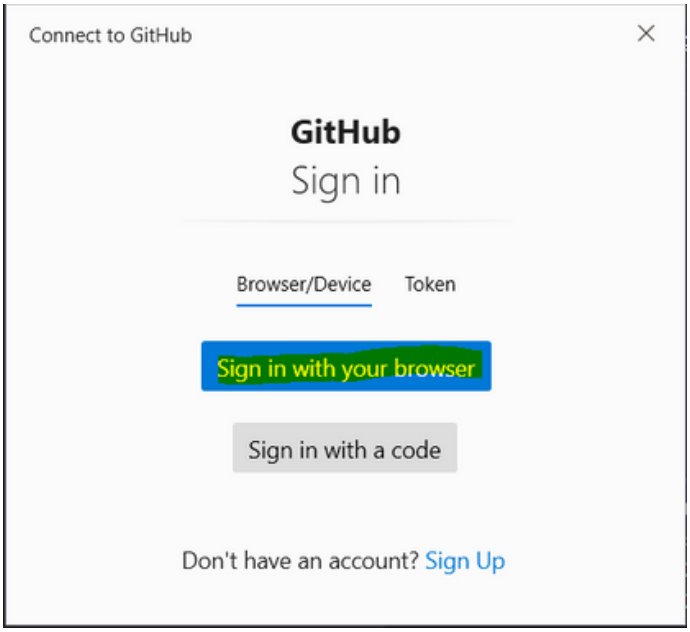
The most important thing is to control your work. After making the changes we need to commit them so our **GitHub repository gets updated automatically**. But, after that, we need to make some configurations. Go & open **Git Bash** and write the below commands (*Enter your GitHub email and username*):

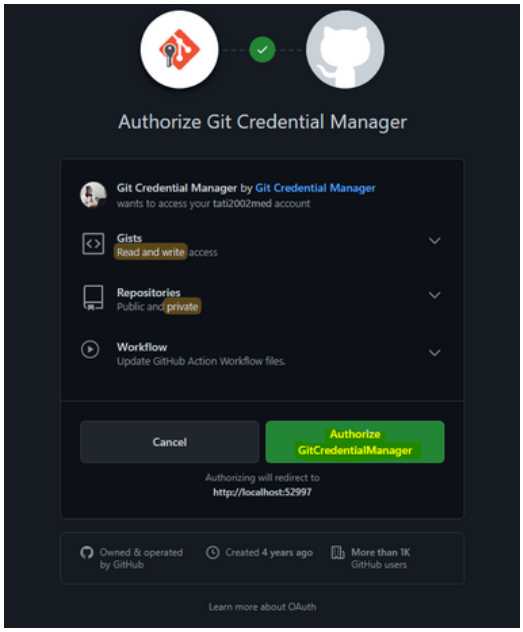
Let's now just commit the changes that we did:



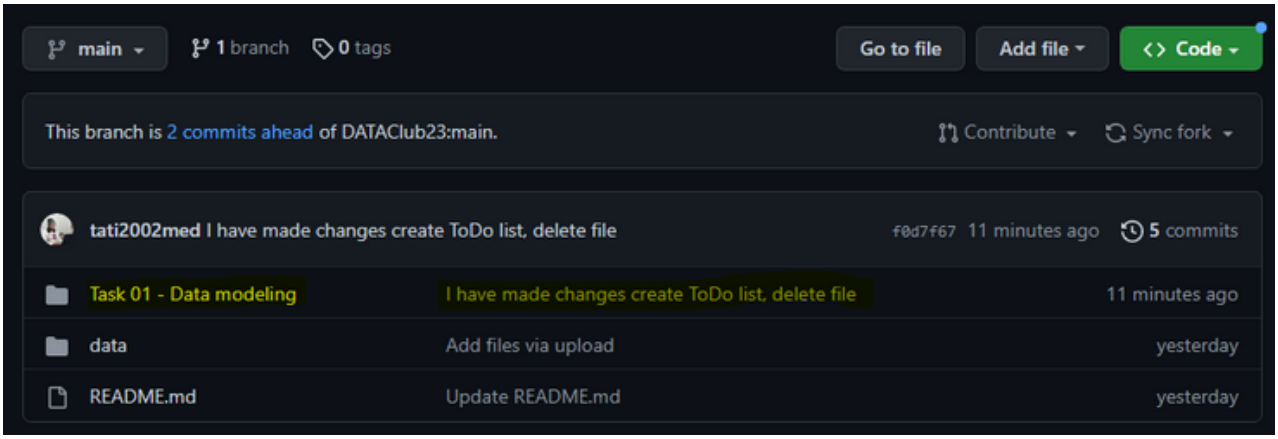


Sign in & Authentication:





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.