Data Engineering - Assignment 01

There are two goals for the assignment this week. The first is to help you reflect on what data engineering is all about and how it may impact your work. For this, you will be having a mock conversation with a recent graduate from this program using ChatGPT. The second is to ensure you have the technical environment set up correctly between GitHub, your laptop, and Snowflake.

# A) What good is this course going to be?

Using the ChatGPT prompt below, have an AI conversation for at least 15 minutes about the topics in this course and explore how your work or interests will benefit from a better understanding of data engineering. (1) Submit your complete conversation transcript, (2) summarize 3 key insights you gained from this AI conversation, and (3) be prepared to discuss at least one of those during our next class. You may use any AI you wish for this conversation, but results may vary.

As you read the prompt, remember that ***"you"*** refers to **the AI** with whom you're having the conversation, and ***"I"*** refers to **you, the student**, as the one entering this prompt into ChatGPT. This is you, the student, explaining to ChatGPT what kind of conversation you want to have.

Suppose you are a recent graduate of a "Master of Data Analytics and Applications" from the Continuing and Professional Education program at Washington University in St. Louis. Your name is Joe Hoyt.

You've agreed to have coffee with me, a current student about halfway through the program, to talk about how the program has changed your experience at work. Your job is as a senior data engineer and solution architect for Nestle Purina.

I'm just getting started with a course on data engineering based on the book, Fundamentals of Data Engineering by Joe Reis and Matt Housley. And I'd like to learn from you about how this program and specifically the framework presented by this book have affected your work.

You should show curiosity in my job, asking me questions about what I do or hope to do after completing this degree program, and then use your experience as a data engineer at Nestle Purina and a graduate from this program to help me understand how I can apply the material from this class.

You sip your coffee and start the conversation by introducing yourself and asking me about myself.

# B) Setup your technical environment

The technical environment for this course will require you to use GitHub.com for managing coding assignments, Python for programming on your laptop, and a Snowflake AI Data Cloud student trial account for SQL programming and running data engineering workloads at scale.

## B.1) GitHub account and laptop client

Use the instructions at <https://github.com/edu/students> and use your wustl.edu email address to get access to special student benefits including private repositories for free. This is optional, but will allow you to keep your code protected. Note that you will need to submit some form of proof that you are a student such as your student ID (if it has a date on it) or an image of an enrollment document with your name and a date on it.

If you already have a GitHub account, you can use the process above, but only after you have added your wustl.edu email address to your profile information.

B.1.a) Once you have created your GitHub account, follow the instructions below to create a new repository.

1. Create a new repository named **data5035** at <https://github.com/new>
2. Provide a description such as "This is my WashU MDAA Data Engineering repository"
3. Make it private if you have signed up for student benefits or have a paid account
4. Check the "Add a README file" box
5. Add a .gitignore for Python
6. Choose License: None

B.1.b) Add **paulboal** as a collaborator on your repository so the instructor can review your work and provide feedback. These instructions explain how to add a collaborator if you aren’t familiar with this process: <https://docs.github.com/en/issues/planning-and-tracking-with-projects/managing-your-project/managing-access-to-your-projects>

B.1.c) Optionally install the Git command line client on your laptop following the instructions from <https://www.git-scm.com/downloads>. If you use Visual Studio Code for your IDE, you are not required to do this step because VS Code has a Git extension built in.

B.1.d) Configure security for your GitHub account with an SSH key following these instructions: <https://docs.github.com/en/authentication/connecting-to-github-with-ssh>. While this isn't required, GitHub recommends using SSH key authentication rather than HTTPS. It is also good to learn a little bit about SSH keys in the process.

## B.2) Snowflake AI Data Cloud Student Trial Account

Use the instructions at <https://signup.snowflake.com/?trial=student> to create a student Snowflake account. This will provide you with a 120-day Snowflake trial and $400 worth of Snowflake credit. This will be sufficient for our class, but Snowflake also allows students to create an unlimited number of trial accounts. You simply can't use any account for longer than 120 days.

## B.3) Python and IDE on your laptop

Install Python on your laptop using the instructions here: <https://www.anaconda.com/download>. Note that there is a "skip registration" option if you do not wish to provide your information to Anaconda.

Install Microsoft Visual Studio Code or another IDE to use for developing code and interacting with Git. VS Code is recommended because it has extensions for Git (built-in), Python, and Snowflake. After installing VS Code from here (<https://code.visualstudio.com/download>), install official extensions for Python and Snowflake.

## B.4) Create data5035 Snowflake user and objects.

Use the SQL script from GitHub here to create Snowflake objects we will use for this class. We create all these security roles, users, and objects to isolate your class work from other projects you might do, making it easier to clean up later.

<https://github.com/paulboal/data5035-private/blob/main/ex01/snowflake.sql>

Use VS Code or the Snowflake Snowsight web interface to update the password in this file and run the script. You will need this password for the dbt steps below as well.

**Remember to REMOVE the password from the file before committing it to your github repository!**

## B.4) Verify setup completeness

To verify that your setup is complete and everything is configured correctly, follow the instructions below. When finished, you will commit your changes to a new branch and push that to your GitHub repository.

You can use this tutorial from Snowflake for additional help: <https://quickstarts.snowflake.com/guide/data_teams_with_dbt_core>

To create a new virtual environment for this class…

pyenv virtualenv 3.11 data5035

pyenv activate data5035

or

conda create -n data5035 python=3.11

conda activate data5035

Then create a folder in repository called **ex01**, download the requirements.txt file from the class public github repository, place it in that folder and proceed…

pip install -r requirements.txt

dbt init

Provide the following input responses:

Enter a name for your project: **data5035\_ex01**

Which database would you like to use?

[1] snowflake

Enter a number: **1**

account: **[ORG.ACCOUNT FROM SNOWFLAKE]**

user (dev username): **data5035\_user**

[1] password

[2] keypair

[3] sso

Desired authentication type option (enter a number): **1**

password (dev password): **<\*\*\*>** *Use your own password here*

role (dev role): **data5035\_role**

warehouse (warehouse name): **data5035\_wh**

database (default database that dbt will build objects in): **data5035**

schema (default schema that dbt will build objects in): **public**

threads (1 or more) [1]: **1**

Check your setup using dbt debug:

**cd data5035\_ex01**

**dbt debug**

20:15:33 Running with dbt=1.9.1

20:15:33 dbt version: 1.9.1

20:15:33 python version: 3.12.3

…

20:15:34 Configuration:

20:15:34 profiles.yml file [OK found and valid]

20:15:34 dbt\_project.yml file [OK found and valid]

…

20:15:34 Connection:

20:15:34 account: rqb36878.us-east-1

20:15:34 user: data5035\_user

20:15:34 database: data5035

20:15:34 warehouse: data5035\_wh

20:15:34 role: data5035\_role

20:15:34 schema: public

…

20:15:34 Registered adapter: snowflake=1.9.0

20:15:35 Connection test: [OK connection ok]

20:15:35 All checks passed!

Add this new folder of files to your git repository, commit them, and push them to github.com.

# 

# Assignment Checklist

1. Have you submitted your AI conversation transcript and your three key insights? Are you prepared to talk about one of those insights during our next class?
2. Have you committed all of your changes to a new branch, pushed them to GitHub, and issued a pull request for the instructor to review?