# **Assignment 3**



**Last Updated:** 03-01-2024

# **Previous work**

i) Information is extracted from a web site by utilizing web scraping tools such as Beautiful Soup or Scrapy and structured the extracted data into a CSV file

ii) Extracted text from pdf using PyPDF2 and Grobid and structured the output into text files

iii) Uploaded the csv data and metadata in Snowflake database by using SQLAlchemy.

iv) Uploaded created csv and text files in WS S3 bucket by using python function.

# **Part 1:**

Current task is to work on the same data to make it clean and work on it further.

i) Designed Python class “URLClass” to represent the schema of each webpage.

ii) Designed MetaDataPDFClass and ContentPDFClass classes to represent the schema for the Grobid output metadata and extracted content of pdf files.

iii) Here data and schema validation was carried out using Pydantic 2.

iv) Built different test cases using Pytest for all classes to see how validation would succeed or fail.

v) A clean csv is created at this step.

# **Part 2:**

i) Loaded the clean data into Snowflake

ii) Created a summary table using DBT.  
 Added columns such as Level, Topic, Year, Number of articles, Min Length (Summary), Max Length (Summary), Min Length (Learning outcomes), Max Length (Learning outcomes)

iii) Materialized it to a new table.

iv) Created test cases to validate new columns.

# **Architectural Diagram**

### 

# **Research and Understanding**

· **Pydantic 2:** A Python library for data validation and schema definition, ensuring data adheres to specified constraints.

· **Pytest:** A popular testing framework in Python that allows for writing and executing test cases to verify code functionality.

· **DBT (Data Build Tool):** An open-source tool used to define and run data transformation workflows in various databases, including Snowflake.

# **Assumptions**

· **Integrity:** Data consistency and integrity are maintained throughout the process.

· **Reliability:** The data processing and upload functions perform flawlessly and efficiently.

· **Seamlessness:** Integration between the scraping tool, cloud storage, and database is smooth and well-documented.

### 

# **Discussion and Justification**

· **Pydantic 2:** Streamlines data validation and schema definition for the extracted and transformed data, guaranteeing data integrity and adherence to specified constraints.

· **Pytest:** Enables comprehensive unit testing of the Python code involved in web scraping, PDF processing, and data transformation, ensuring code reliability and functionality.

· **DBT (Data Build Tool):** Streamlines data transformation workflows within Snowflake, providing a centralized and modular approach to data manipulation and modeling.

### 

# **References**

·  [https://docs.getdbt.com/guides/snowflake?step=1Links to an external site.](https://docs.getdbt.com/guides/snowflake?step=1)

·  [https://courses.getdbt.com/courses/fundamentalsLinks to an external site.](https://courses.getdbt.com/courses/fundamentals)

·  [https://github.com/Coding-Crashkurse/Pydantic-v2-crashcourseLinks to an external site.](https://github.com/Coding-Crashkurse/Pydantic-v2-crashcourse)