Team Name: Gujjus

Project: Histopathology Data Ingestion

Data Area: Data Ingestion

Overview

The goal of this project is to create an SQL schema from histopathology data provided in CSV and JSON formats, clean the data, and then migrate it into SQL tables. The project involves a combination of data analysis, database design, data cleaning, and data migration skills.

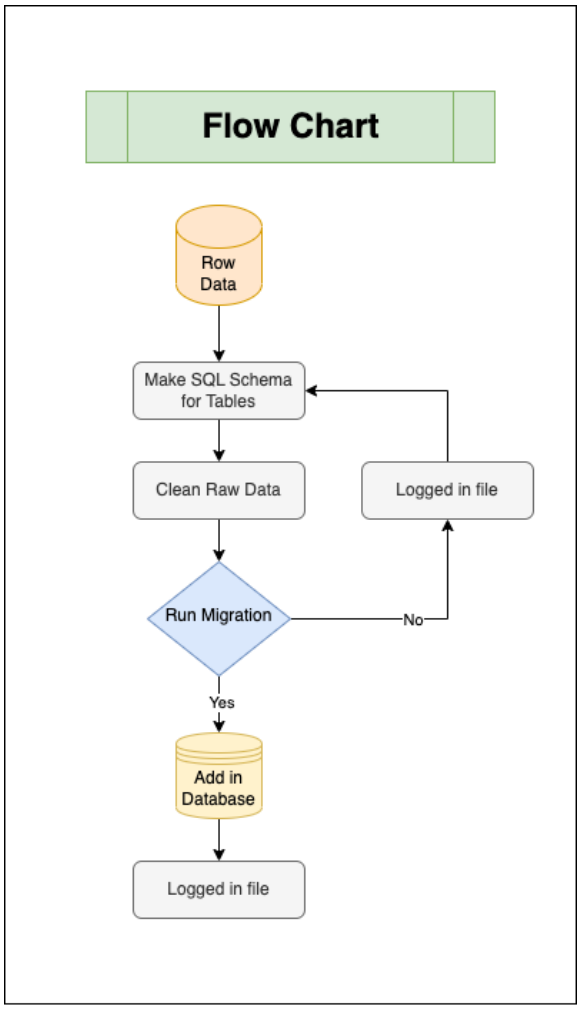
Requirements

1. SQL schema generation: The project has to generate an SQL schema for the data provided in the CSV and JSON files. It will be static, meaning that the schema will not change as the data is updated.
2. Cleaning data: The project has to clean the data before it can be inserted into the database. This involves preparing the CSV or JSON files so that the data is consistent, accurate, and ready for insertion into the SQL database.
3. Validating data: The project has to validate the data before it can be inserted into the database. This involves checking for errors, missing values, duplicates, and inconsistencies in the data.
4. Migration of data: The project has to migrate the cleaned data into the SQL database. This involves inserting the data into the appropriate tables in the database, according to the SQL schema we created earlier.
5. Scalability: The system should be scalable to accommodate large volumes of data.
6. Error handling: The system should have mechanisms for handling and reporting errors that occur during data ingestion.
7. Logging: The system should have mechanisms for logging the process of data ingestion, including any errors that may have occurred.

Technical Requirements

1. SQL schema generation: The project is using MySQL as the database management system. The schema will include tables, columns, and relationships between the tables. The SQL schema should be designed to efficiently store and retrieve data, while also ensuring data integrity and consistency.
2. Cleaning data: The project is using Python as the programming language for cleaning the data. The project is using Pandas library for data manipulation and cleaning.
3. Data migration script: The project is using Python as the programming language for migrating the data into the SQL database. Migration is usefull for adding CSV or JSON data into a MySQL database.
4. Error handling: The project is using Python as the programming language for handling errors. Errors are handled while cleaning the data and migrating the data into the SQL database.

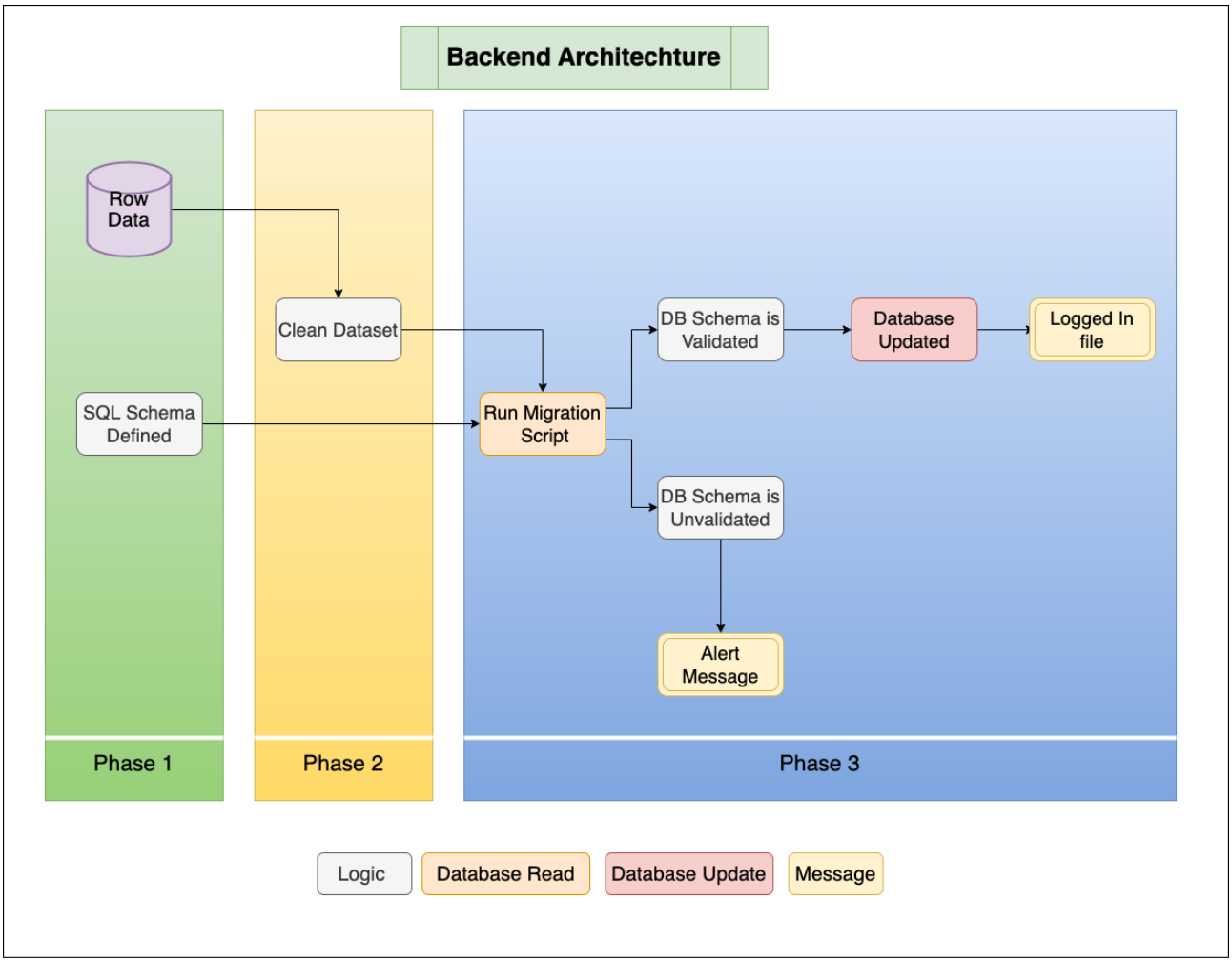
System Design



This flowchart is a graphical representation of a process that depicts the steps involved in completing a our project or task.

We have divided project into 3 phases.

1. SQL Schema Formation
2. Cleaning Data
3. Migration of data



Stage 1 (SQL Schema)

- The Data is provided in the form of CSV or JSON files.
- The first phase of our project is to create a SQL schema for the data.
- The schema will include tables, columns, and relationships between the tables.
- The SQL schema should be designed to efficiently store and retrieve data, while also ensuring data integrity and consistency.

Stage 2 (Cleaning Data)

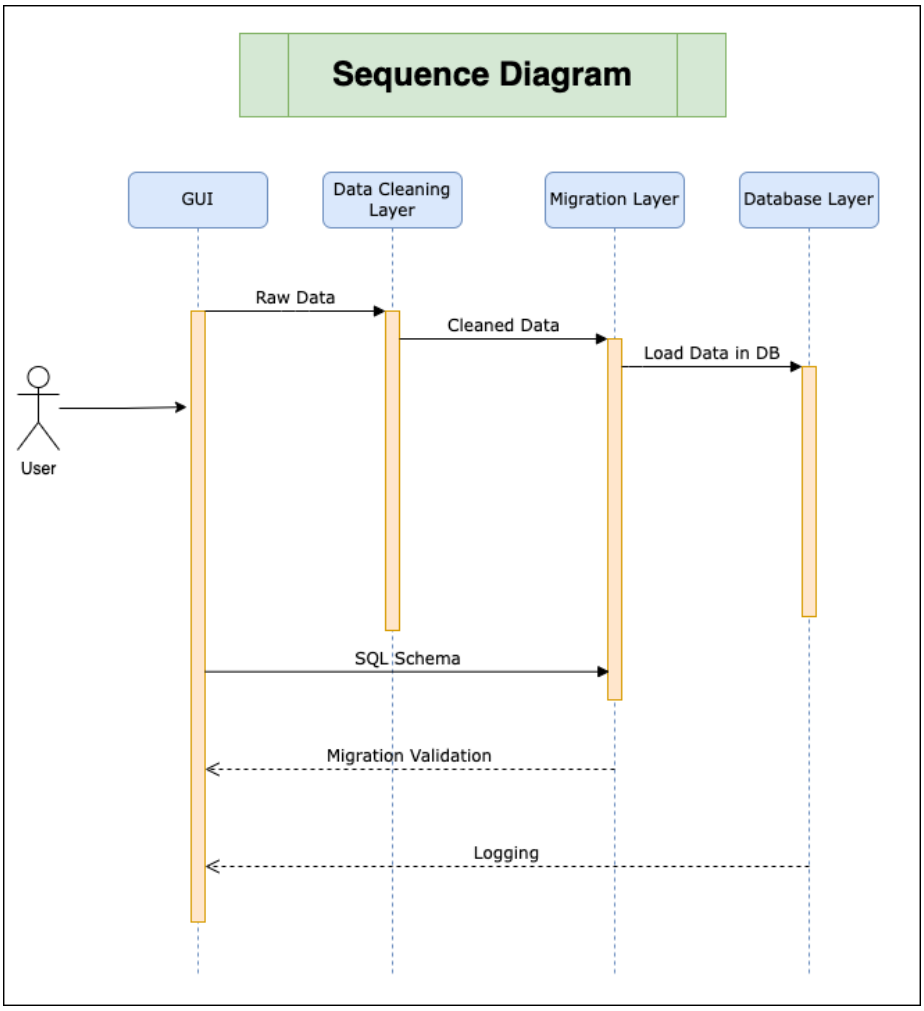
- The second phase of our project is to clean the data before it can be inserted into the database.
- This phase involves preparing the CSV or JSON files so that the data is consistent, accurate, and ready for insertion into the SQL database.
- During the cleaning process, we need to check for errors, missing values, duplicates, and inconsistencies in the data.

Stage 3 (Migration of data)

- The third and final phase of our project is to migrate the cleaned data into the SQL database. This involves inserting the data into the appropriate tables in the database, according to the SQL schema we created earlier.
- The migration process can be done using SQL commands or a variety of tools and programming languages, such as Python or Java.

- Once the data migration script executed, we will log the process and any errors that may have occurred.

Sequence Diagram



Team Members

- Jay Ghevariya (2020101070)
- Tirth Motka (2020101036)
- Urvish Pujara (2020101032)