Operational Guide

Follow these **steps**:

1. Install the required libraries. They are also available in the file “requirements.txt”.
2. Configure “.env” file with your credentials for Giant Bomb API key, and Neo4j database. Otherwise, you can set values manually.
3. Run “.ipynb” files in sequence:
    a. Data_Acquisition_Characters_and_Games
    b. Data_Acquisition_Enemies
    c. Data_Acquisition_Bosses
    d. Data_Integration
    e. Data_Quality
    f. Data_Storage_GraphDB

Note that the data obtained via web scraping and the Giant Bomb API may differ from the data provided in this project, as the Mario Wiki and Giant Bomb pages are subject to updates.

Details about the project’s workspace

Jupyter notebooks handle the data processing pipeline, from acquisition to storage.

- *Data_Acquisition_Characters_and_Games.ipynb*: This notebook scrapes character information, and a comprehensive list of games with release dates and platforms from mariowiki.com. Then, it handles sales data from Kaggle.
- *Data_Acquisition_Enemies.ipynb*: Focuses on scraping data about common enemies that appear across the various Mario games.
- *Data_Acquisition_Bosses.ipynb*: Scrapes information about boss characters and the games in which they appear as bosses.
- *Data_Acquisition_API.ipynb*: Uses an external API (GiantBomb) to fetch additional data about general character relationships, to enrich the scraped datasets.

- *Data_Integration.ipynb*: This notebook is about the integration phase; it is responsible for cleaning, transforming, and merging the data from all the different sources (characters, bosses, enemies, API) into a unified dataset.
- *Data_Quality.ipynb*: Performs a data quality assessment.
- *Data_Storage_GraphDB.ipynb*: Takes the final, integrated data and loads it into a Neo4j graph database, creating nodes for characters and games, and relationships between them.

CSV data files store the raw, intermediate, and final datasets used and generated by the notebooks.

- *characters_df.csv*: Contains the list of all the plot's characters scraped from the wiki.
- *enemies_df.csv*: Stores the data for all scraped enemy characters.
- *bosses_df.csv*: Contains the data for all scraped boss characters.
- *games_data.csv*: A dataset of Mario games, including details like platform/console and release year.
- *sales_data.csv*: Contains sales data for many video games of the Mario franchise.
- *general_character_relationships.csv*: Stores character-to-character relationship data, sourced from the API.
- *merged_data.csv*: A comprehensive, merged dataset that combines information from characters, games, bosses, and enemies.
- *merged_data_API.csv*: An intermediate file containing data merged with information from the external API.
- *Video_Games.csv*: A general-purpose, potentially external, dataset about video games used as a source.

Optional Configuration Files for the security

- *.env*: An environment file to store configuration variables, such as API keys, securely.
- *.gitignore*: A standard Git file (in case of a future GitHub upload) that specifies which files and directories to exclude from version control.