Software to use: ER Studio/Navicat, Databricks, Alteryx, Snowflake, Power BI or Tableau (Pick your choice of tools for implementing the project.)

This is an end-to-end implementation of the BI project, starting with data analysis, understanding data properties, analyzing the data using Alteryx/Python ydata profile, documenting, creating a mapping document (Make sure to unify the columns from both data sets), staging the data, cleaning data, loading it into the integration schema, and generating reports based on the integration schema. Follow the Medallion architecture. Follow all standards on defining columns, facts, dimensions, standard mandatory/recommended columns to track source, load date and who loaded. Each row should be tradable to the source. Your code must contain validations at each step to move forward. If there are any rows dropped explain why they dropped before skipping them.

Business requirements are as follows:

As a Business Analyst, I should be able to -

- Get different types of professions for any given personnel (Primary and other professions if applicable)
- Generate a report to find personnel who have more professions (Primary and other professions, if applicable)
- Get the list of genres for a given title and identify a list of movies based on a given genre
- Find all moves that are released in a given year
- Track the movie length to generate metrics on movie length based on release year.
- List all adult and non-adult movies
- Generate all different languages in which a title is associated
- Get a list of regions in which a move is released
- Get the list of all directors and writers involved in making a movie so that I can identify the popular directors and writers
- Get the number of episodes associated with a season for a given title
- Find all different types of crew/cast, directors, writers, etc, for a given title
- Get the list of jobs involved in a given title
- Get a list of characters involved in a given title
- Find top-rated movies by year, by genre

As a Team, create visualizations to fulfill the business requirements. For example, the following are for reference purposes only. Come up with comparison and analysis reports that adds value to the project)

- Movie insights
- Analysis based on Crew members (Actor, Director, backstage resources, writers, etc)

- Movie trend analysis based on ratings
- Non-movie title analysis (Seasons and number of episodes comparison with ratings to understand what users like more)
- Region and Country based movie releases

Details of the project inventory are as follows:

- There are seven zip files which contain details about
    - All cast members and behind-the-scenes personnel details - name.basics.tsv.gz
    - All movie title details and its genres - title.basics.tsv.gz
    - For a given title, this file gives a list of title names in multiple languages - title.akas.tsv.gz
    - Gets all Directors and Writers (This includes a complete list of directors and writers involved in a given title) - title.crew.tsv.gz
    - This file Includes the series title, season number, and episode number - title.episode.tsv.gz
    - Principal credits are a set of the most important cast/crew credits for a title, with the selection and order determined by IMDb. Principal credits are often similar to top-billed cast - title.principals.tsv.gz
    - Movie ratings and their votes are in this file - title.ratings.tsv.gz
- Make sure to incorporate descriptions for Region and Language using the links below
- Document each file row counts and match them to after loading
    - name_basics
    - title_akas
    - title_crew
    - title_basics
    - title_episode
    - title_principals
    - title_ratings

Important links:

- The dataset can be downloaded from here (If you are having trouble downloading, contact TAs. They will provide one drive link): IMDb Non-Commercial DatasetsLinks to an external site.
  - https://datasets.imdbws.com/
- All region codes and their descriptions are available here: Country Codes For MoviesLinks to an external site.
- A list of language codes and their description is available here: List of ISO 639 language codesLinks to an external site. - https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes

- The mapping document template is here: Mapping_Template.xlsx Download Mapping_Template.xlsx

inks to an external site.Liks to an external site.

Deliverables:

- Create a mapping document to source and target and include the different transformations used to perform the changes.
- All source code must be on git, and we should see the team contributions (Create a new repo that the respective team members will use)
- Upload all the data profiling analysis document, data modeling design files, report files, pdf files, excel files, and ppt files, if any should be uploaded to Canvas