

**Team Members**

Sashank Machiraju –  
Pritika Purushothaman – 002796817 – purushothaman.pr@northeastern.edu  
Manimanya Reddy – 002763582 – jukantivenkat.m@northeastern.edu  
Neha Reddy -

**Dataset**

IMDB Movie Ratings  
  
**I. Perform Data Profiling with Alteryx**A diagram of a computer network

Description automatically generated with medium confidenceA line of a person's face

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedA screenshot of a data

Description automatically generatedA screenshot of a computer

Description automatically generated  
  
Alteryx profiling on files-   
1. IMDB SQL tables   
2. IMDB TSV files  
  
- Alteryx detailed profiling document, available.  
  
**Explanation -**   
  
- The Alteryx workflows involve data profiling on TSV files containing information from the IMDB database, specifically focusing on six tables. In the first table, "IMDB NAME BASICS," individual identifiers, names, birth and death years, professions, and known titles are stored, each with specific data constraints.   
  
- The second table, "IMDB TITLE AKAS," contains details about movie titles, including language, region, and release information.   
  
- The third table, "IMDB TITLE BASICS," provides information about movies, such as title type, adult content, release years, and genres. These tables are interconnected through unique identifiers, ensuring data integrity and relational structure.

- The fourth table, "IMDB TITLE CREW," lists the directors and writers associated with each movie, with creation timestamps.   
  
- The fifth table, "IMDB TITLE PRINCIPALS," includes details about individuals involved in a movie, such as their roles, order of appearance, and associated characters.   
  
- Lastly, the sixth table, "IMDB TITLE RATINGS," involves ratings assigned to movies, with director and writer information. The data in these tables is crucial for understanding the relationships between movies, individuals, and their respective roles in the filmmaking process.

- Additionally, the workflows involve analyzing revenue data, which is spread across nine separate CSV files. Each file has 11 columns with specific data constraints. The revenue data includes movie identifiers, titles, dates, ranks, gross earnings, percentages, theater information, total gross, and duration. Data validation checks reveal issues such as trailing spaces and null values, providing insights into data quality and potential cleaning requirements.  
  
- The comprehensive overview of these Alteryx workflows and associated TSV files underscores the complexity and richness of the IMDB dataset, offering valuable insights for data analysis and exploration.  
  
**II. Load Data into Staging Tables with Talend**

**Staging Jobs –**

|  |  |
| --- | --- |
| S.no | Job name |
| 1 | Actors\_combined |
| 2 | Combined\_titles |
| 3. | Converting\_runtime\_int |
| 4. | Json\_staging\_names |
| 5. | Json\_staging\_titles |
| 6. | Sql\_scripts\_staging |
| 7. | Tsv\_file\_staging |

**DIM\_Loading Jobs –**

|  |  |
| --- | --- |
| S.no | Job Name |
| 1. | Date\_dim\_loading |
| 2. | Directors\_dim |
| 3. | Genres\_dim |
| 4. | Location\_dim |
| 5. | Movie\_dim\_Scd |
| 6. | Person\_dim\_scd |
| 7. | Principals\_dim |
| 8, | Writers\_dim |

**Fact\_Loading Jobs –**

|  |  |
| --- | --- |
| S.no | Job Name |
| 1. | Fact\_movie\_accomplishments |
| 2. | Fact\_movie\_revenue |

**Bridge\_Loading Jobs –**

|  |  |
| --- | --- |
| S.no | Job Name |
| 1. | Genre\_movie\_bridge |
| 2. | Region\_bridge\_loading |

**III. Dimensional Modelling -**

A computer screen shot of a computer

Description automatically generated

**IV. Mapping of source and target mapping**

In the below diagram columns colour in each staging tables corresponds to the colour of the stag table references

A screenshot of a spreadsheet

Description automatically generated

Categorised into 4 kinds   
1. Movie -   
2. Revenue/accomplishments -   
3. Workflow  
4. Person-   
  
  
  
  
  
  
  
  
  
  
**Transformations -**

Documented File available

A white sheet with text

Description automatically generated  
  
  
  
**Data Visualization -**

**PowerBI Dashboard Link**

**Tableau Dashboard Link**