**Netflix Data Normalization Documentation**

**Overview**

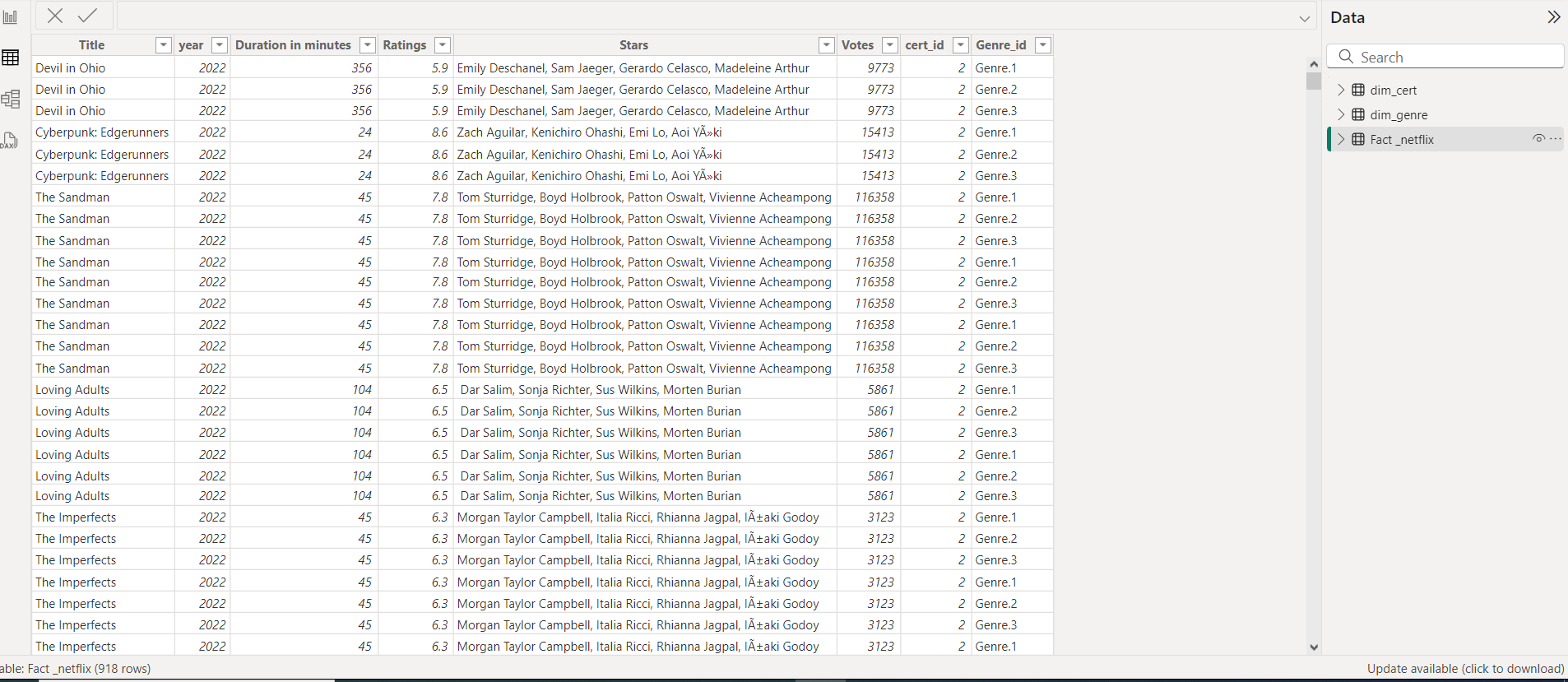
This documentation describes the normalization process applied to the Netflix dataset, which includes a fact table and two dimension tables. The normalization technique used in this dataset is **3rd Normal Form (3NF)**, which aims to eliminate redundancy and ensure data integrity.

**Tables and Schema**

**Fact Table: fact\_netflix**

This table contains the primary data about Netflix movies, including their titles, release years, durations, ratings, stars, votes, certification IDs, and genre IDs.

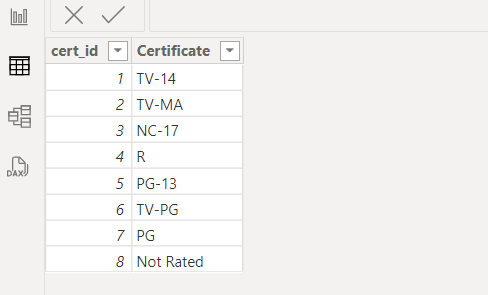
* **Columns:**
  + title: Title of the movie (String)
  + year: Release year of the movie (Integer)
  + duration: Duration of the movie in minutes (Integer)
  + ratings: Ratings of the movie (Float)
  + stars: Star rating of the movie (Float)
  + votes: Number of votes the movie received (Integer)
  + cert\_id: Certification ID (Foreign Key to dim\_cert table) (Integer)
  + genre\_id: Genre ID (Foreign Key to dim\_genre table) (Integer)



### Dimension Table 1: dim\_cert

This table contains certification details corresponding to the certification IDs used in the fact\_netflix table.

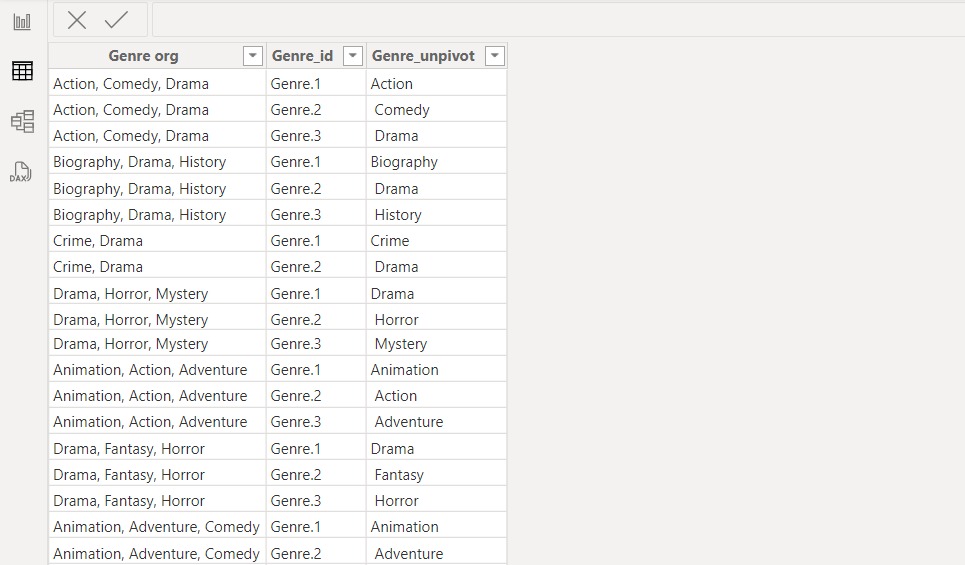
* **Columns:**
  + cert\_id: Certification ID (Primary Key) (Integer)
  + certificate: Certification type (e.g., PG, R) (String)



### Dimension Table 2: dim\_genre

This table contains genre details corresponding to the genre IDs used in the fact\_netflix table.

* **Columns:**
  + genre\_id: Genre ID (Primary Key) (Integer)
  + genre\_org: Original genre (String)
  + genre\_unpivot: Unpivoted genre for easier analysis (String)



## Relationships

* fact\_netflix → dim\_cert: Linked by cert\_id
* fact\_netflix → dim\_genre: Linked by genre\_id

## Normalization Technique: 3rd Normal Form (3NF)

The dataset is normalized using the Third Normal Form (3NF). The key aspects of 3NF normalization applied in this dataset include:

1. **Eliminating Redundancy**:
   * Data related to certifications and genres is stored in separate dimension tables (dim\_cert and dim\_genre). This reduces redundancy by ensuring that each piece of information is stored only once.
2. **Ensuring Data Integrity**:
   * Foreign key relationships are established between the fact table and the dimension tables. This maintains data integrity by ensuring that all certification and genre references in the fact table correspond to valid entries in the dimension tables.
3. **Eliminating Transitive Dependency**:
   * Non-key attributes are fully dependent on the primary key. For example, cert\_id in fact\_netflix refers to a single certification in dim\_cert, and genre\_id refers to a single genre in dim\_genre.