Group Project - Neo4j

of Systems and Methods for Big and Unstructured Data Course (SMBUD)

held by

Brambilla Marco

Tocchetti Andrea

Group 78

Pisante Giuseppe 10696936

Raffaelli Martina 10709893

Academic year 2024/2025

POLITECNICO
MILANO 1863

Contents

1 Introduction

The project aims to design and implement a database system to support the management of data related to the film industry. The database will include entities such as Person (actor, director, writer), Title (film, TV series), Episode, ratings, and genre. The goal is to create a comprehensive system that can store and query information about films, TV series, and the people involved in their production. The project will develop with Neo4j, to exploit the relations between the entities and find significant insights from the data, such as collaborations between directors and actors, trend of genres over time, most working actors and the most successful films.

2 Assumptions

The project is based on the following assumptions:

- Each person has a unique ID and a name, surname, and date of birth. Some could also have a death date, if passed away.
- Each person can be associated with multiple roles (actor, director, writer, archive footage, music department, producer)
- Each title has a unique identifier and a title type (film, TV series, shortfilm)
- Each title can have multiple episodes
- Each title can have multiple genres
- Each user can rate a title if and only if it has watched the film
- Each title has a unique rating, average rating from the users, and the number of votes
- Each person can be associated with multiple titles
- Each title can have multiple people associated with it

3 ER diagram

Figure 1: E-R Diagram

3.1 Entities

Starting from the considerations previously exposed regarding the implementation hypotheses, we have drawn an ER diagram (**Figure 1**) which includes 5 different entities and 7 many-to-many relationships described below in the logical model:

- **Person**(nconst, PrimaryName, BirthYear, DeathYear, PrimaryProfession, KnownForTitles)
- **Title**(tconst, PrimaryTitle, OriginalTitle, TitleType, StartYear, EndYear, RuntimeMinutes, Genres)
- **Episode**(tconst, ParentTconst, SeasonNumber, EpisodeNumber)
- **Ratings**(tconst, AverageRating, NumVotes)