INFO 6210
Data Management and Database Design
[Section 3]

# Physical Data Model and Social Media

## Assignment 2



Authors

Purvang Jayesh Thakkar 001387983
Ira Abhijeet Pantbalekundri 001423854

# Introduction

The domain chosen for this Assignment is on Movies.

Cinema is an extremely popular source of entertainment worldwide. Numerous movies are produced each year and people watch these in large numbers. Cinema impacts our life both positively and negatively. Just as everything else in this world, cinema also has positive as well as negative impact on our life.

## Data Sources:

We have gathered data from 3 sources:

1. IMDB Page Web Scraping
2. Social Media-Twitter (Using twitter API)
3. Social Media-Instagram(Using Instaloader API)

# Part-1 Conceptual Model

- Domain

  The assignment is based on Movies.

- Conceptual models (entities) for a tweet/post, a Social Media user, a person, and a company.

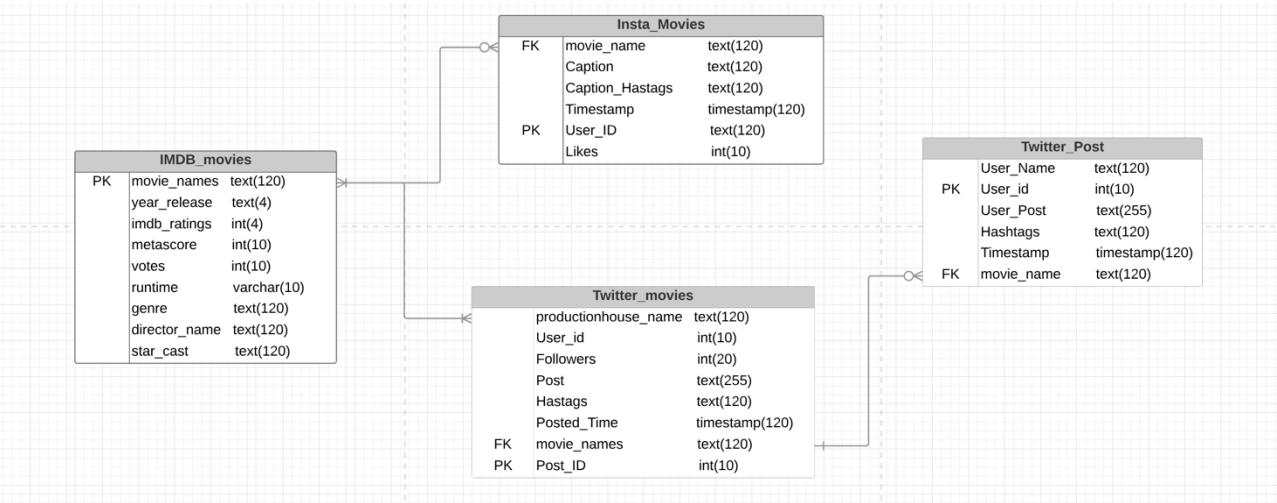
We have four entities overall in the entire assignment, first entity is the IMDB_movies which we have scraped from IMDB webpage. Attributes include Movie names, imdb rating, metascore, director name, star cast, year of release etc. This entity is are Producers as we are fetching the director names from this.

Primary key used for this entity is Movie Name as it is unique in our scenario.

Second entity taken is from Instagram posts which are our Consumers, where in movie goers, comment and review about the movies they watch and post about the same.
Attributes will only contain information like their User Id, caption they posted on their post, Caption hashtags in the post and timestamp.

Third entity taken is from Twitter which is our Company, where in we take the different production houses under whom the movies are made, and we extract the tweets of these production houses and get their tweets and other related information.
Attributes include Production house, production tweets, timestamp, user id and name of the movies.

Fourth entity is again taken from twitter where in whatever movie names we extract from the production houses tweets we pass them as hashtags, and get movie reviews from movie goers and extract their tweets.



Part -1

7 Questions to be answered:

1. What are the ranges, data types and format of all of the attributes in your entities?
   Instagram:
   - Movie_name: text(120)
   - Caption : text (120)
   - Caption_Hashtag : text (120)
   - Timestamp : timestamp(120)
   - User_id : text(120)
   - Likes: int(10)

   Twitter Movies:
   - Productionhouse_name : text (120)
   - User_id : int(10)
   - Followers : int(10)
   - Post : text(255)
   - Hashtags : text (120)
   - Posted_Time: timestamp (120)
   - Movie_name: text (120)
   - Post_ID: int(10)

   Twitter_Post:
   - User_name: text (120)
   - User_id: int(10)
   - User_post: text(255)
   - Hashtag: text (120)
   - Timestamp:timestamp(120)
   - Movie_name: text (120)

2. When should you use an entity versus attribute?

   If an attribute is multivalued/ composite, it can be used as an entity.

3. When should you use an entity or relationship, and placement of attributes?

   A relationship is used to join two tables. At times there could be a self join where a relationship is there with the same entity itself.

   Attributes could be associated to both entities and relationship.

   Eg: Twitter is joined with IMDB based on movie name as both twitter post about the movie and IMDB deals with the name of the movie
   Each table already has its own set of attributes defining something about the post in twitter and ratings in IMDB

4.How did you choose your keys? Which are unique?

We have chosen Movie Name as our primary key in Movies table, User ID as primary key in Instagram, Twitter_Post has User ID as primay and Twitter_Movies has Post_ID as the primary key.

Eg: Twitter is joined with IMDB based on movie name as both twitter post about the movie and IMDB deals with the name of the movie
Each table already has its own set of attributes defining something about the post in twitter and ratings in IMDB

5.Did you model hierarchies using the “ISA” design element? Why or why not?

No, we did not model the hierarchy using ISA design element. But further down it could have been modeled as an Actor can be a director or a Producer and vice versa.
Currently, we have not modeled it using ISA.

6.Were there design alternatives? What are their trade-offs: entity vs. attribute, entity vs. relationship, binary vs. ternary relationships?

Yes. The design alternative could have been to have a movie ID associated with all the entities, and it would have been easier to project it and relate it together. Currently, it is been related using the Movie Name as the common relation between all entities, but movie name does not tend to be unique for all the Entities.

7.Where are you going to find real-world data to populate your model?

Part of my data is from IMDB Website and we have scraped from it.
We have used two Social Media accounts, one is Instagram where we are fetching data for Consumers and Twitter for Consumers as well as Company.
Instagram API: Instaloader
Twitter API: Twitter

# Contributions

Purvang worked on Instagram API and fetching data from it.
Ira worked on fetching data from the Web Source from IMDB website and Twitter API.
Purvang carried out the creating and querying in SQL using SQLite.
Ira designed and developed the Conceptual Database Model.
Both the authors worked on Use Cases.
Both the authors discussed the results and contributed to the final report.

# References


[1] Prof.Nik Bear Brown https://github.com/nikbearbrown/INFO_6210
[2] https://datatofish.com/create-database-python-using-sqlite3/
[3] Python Tutorials: https://www.tutorialspoint.com/python/
[4] https://stackoverflow.com/questions/34040402/sql-query-for-knowing-the-popular-hashtag-from-a-column-that-has-a-list-of-hash
[5] https://www.w3schools.com/sql/sql_like.asp
[6] https://www.lucidchart.com
[7] https://en.wikipedia.org/wiki/Conceptual_schema
[8] https://www.w3schools.com/sql/sql_count_avg_sum.asp
[9] https://instaloader.github.io/
[10] https://www.e-education.psu.edu/geog485/node/141
[11] https://en.wikipedia.org/wiki/Conceptual_schema

# License