Analysing MovieLens 20M dataset with polyglot persistence data model

Group 34

Andrew Lee

Short introduction of the data model name, purpose and motivation

The database that has been selected is called “MovieLens 20M”.

What is the data model about?

- Real world Emulation, the dataset depict multicomponent of interactions between user and movie content
- Holistic Data Mapping, systematic mapping with user preferences,ratings,links and tags.

Why is the data model needed?

- It encapsulates a wide array of attributes revolving user engagement and content personalisation.
- To concisely evaluate pros and cons of each selected database technologies in CAPS theory

Why did we select this data model?

- Rich volume of data as it contains over 20 million ratings, which will give valuable insights in user preferences
- Deployment of polyglot persistence model in a realistic scenario

Short introduction of the data sources with accessible source links

Where did you find the data?

- I found the dataset through grouplens, a research group in the department of computer science and engineering at University of Minnesota, they publish open source datasets that is easily accessible to the public at no given cost.

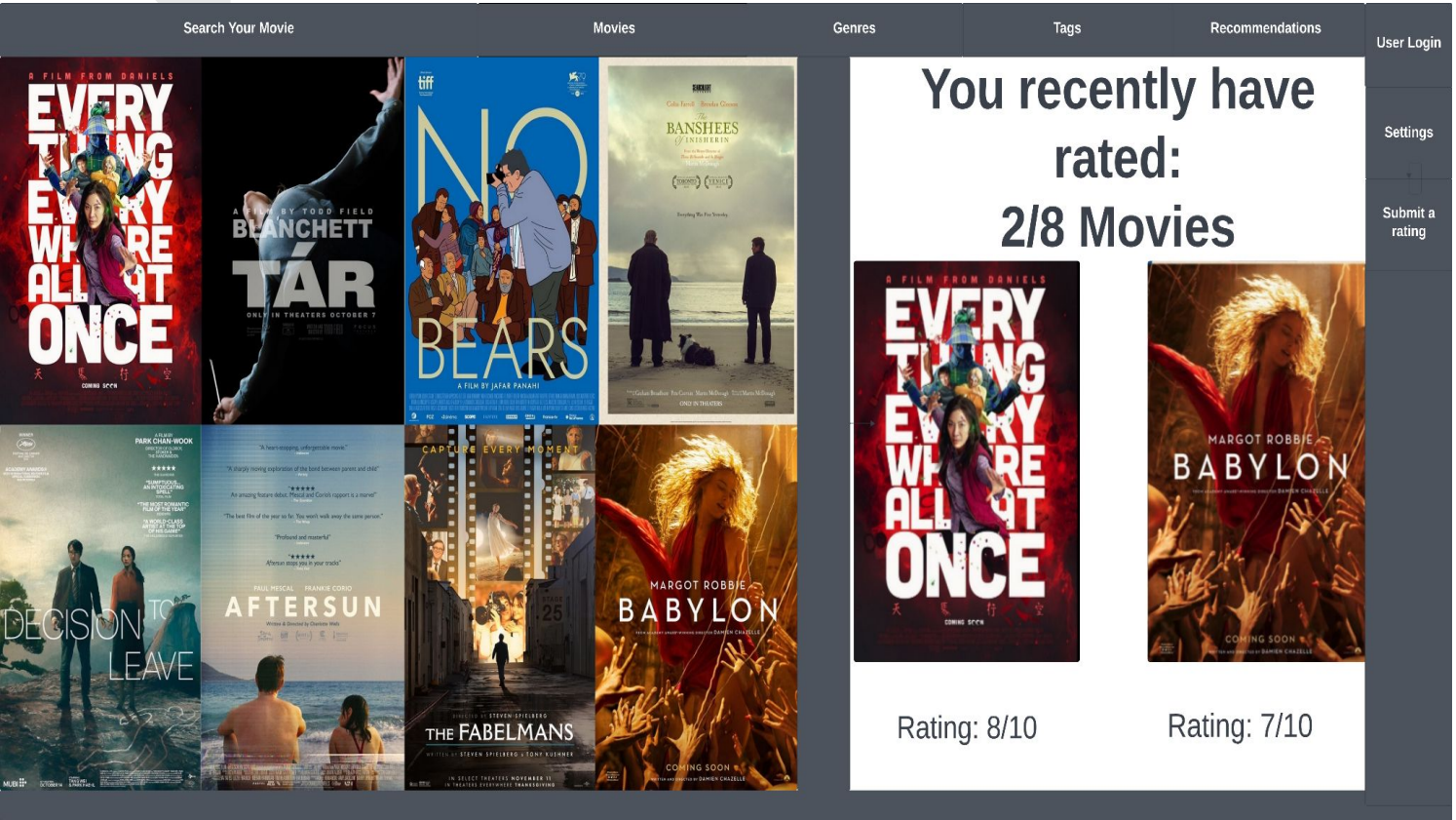
What is the link to your data source?

- https://grouplens.org/datasets/movielens

Is your data source easily accessible (no account login/google drive share etc.)

- The data source does not require any form of logins or google drive, as it is downloadable to the public through the domain website.

Mockup of User Interface



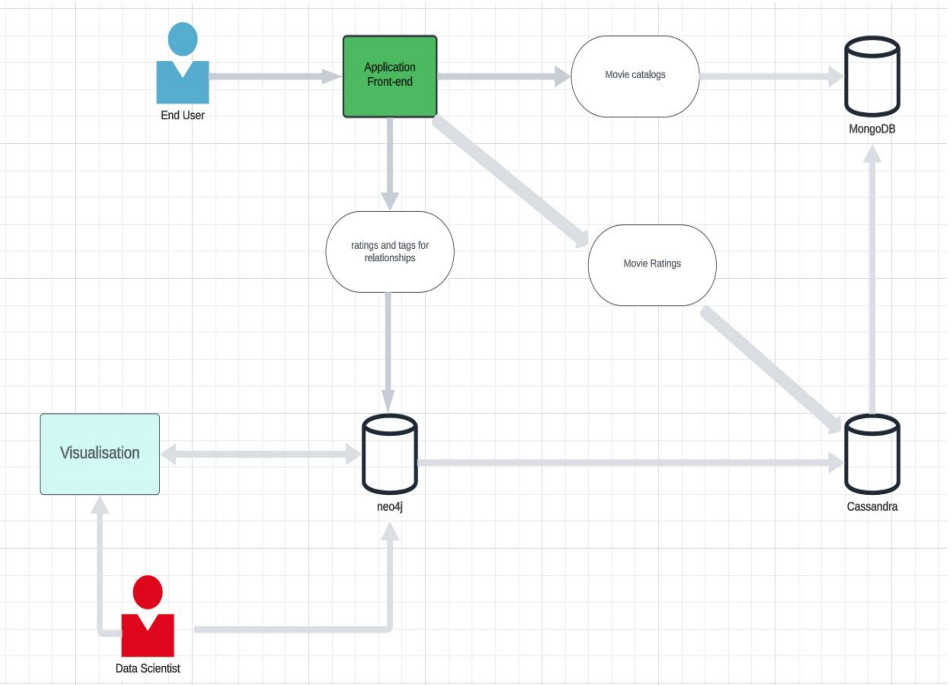
Landing page:

- Searchbar
- Movies
- Genres
- Tags
- Recommendations
- User Login
- Settings
- Rating

Design of Polyglot persistence Data Model with Transaction Management

How will you manage transactions?

A high level overview of polyglot deployment:



Legend:

mongoDB

cassandra

neo4j

Short justifications on why specific databases were used to store specific data

Neo4j
- Atomicity
- Consistency
- Isolation
- Durability
- Graph Relations
-

Cassandra

- It is basically available
- Eventual consistency
- Soft State

MongoDB
- Atomicity
- Consistency
- Isolation
- Durability

MongoDB
- will be used to store static data

Cassandra
- is used to handle large write operations

Neo4j will
- then be used to create recommendations

User will be
- Recommended Movies based on previous ratings
- Similar movies
- Popular movies
- Expandable to other forms of data analysis

Any references to support your justifications

Data Consistency Models: ACID vs. BASE Explained

https://neo4j.com/blog/acid-vs-base-consistency-models-explained/

Week 12 – NoSQL Transactions (continued) and Polyglot Persistence

https://lms.monash.edu/mod/resource/view.php?id=11900090