Hệ thống phân tích dữ liệu lớn khám phá Tri thức, xây dựng Dashboard trong Lĩnh vực Phim ảnh

NGUỒN DỮ LIỆU & ĐẶC TRƯNG

3 nguồn dữ liệu với đặc thù khác nhau (Structured, Semi-structured và Unstructured)

1. MovieLens

Link: https://grouplens.org/datasets/movielens/
Volume:

- 32,000,204 lượt đánh giá (ratings).
- 2,000,072 tags.
- 87,585 movies.
- 200,948 users.
- Thời gian: 28 năm (09/01/1995 – 12/10/2023).

Dữ liệu **structured**

Cấu trúc data

1. ratings.csv

Schema: userId, movieId, rating, timestamp.

2. movies.csv

Chứa thông tin metadata cơ bản của phim.
Schema: movieId, title, genres.

3. tags.csv

Chứa từ khóa do người dùng tự gán
Schema: userId, movieId, tag, timestamp.

4. links.csv (Bảng Ánh xạ - Mapping Table)

Schema: movieId (MovieLens), imdbId, tmdbId.
Đặc điểm:
imdbId: Mã định danh trên IMDb.

Nhận xét:
Thiếu thông tin tài chính & Ekip: Không có Đạo diễn, Diễn viên, Doanh thu. (Đây là lý do sử dụng thêm source crawl từ IMDb).
Title không chuẩn hóa 100%: Có cảnh báo "Errors and inconsistencies may exist". Cần cẩn trọng khi map bằng tên phim.

2. IDMb Non-Commercial Datasets

Link: https://developer.imdb.com/non-commercial-datasets/

Cung cấp ngữ cảnh để hiểu sâu hơn về bộ phim: đạo diễn, ai đóng, và thế giới đánh giá chung về nó như thế nào.

Định dạng: TSV (Tab-Separated Values).
Cập nhật: Hàng ngày (Daily), Hệ thống sẽ thiết kế pipeline chạy định kỳ để cập nhật thông tin mới nhất.

Schema data
A. title.basics.tsv.gz
Đây là bảng metadata chính của phim.
Trường quan trọng:

- tconst: Khóa chính để join.
- titleType: Loại phim
- primaryTitle & originalTitle: Tên phim.
- startYear: Năm sản xuất.
- runtimeMinutes: Thời lượng.
- genres: Thể loại (IMDb định nghĩa khác MovieLens).

B. title.ratings.tsv.gz
Trường quan trọng: averageRating, numVotes.
Cung cấp một "hệ quy chiếu" khác. Rating của user MovieLens cá nhân hơn so với Rating đại chúng trên IMDb.
numVotes phản ánh độ phổ biến toàn cầu của phim.

C. title.principals.tsv.gz & name.basics.tsv.gz

title.principals: Là bảng nối giữa Phim (tconst) và Người (nconst).
Chứa cột category: actor, actress, director.
Chứa cột ordering: Số càng nhỏ thì vai trò càng chính (Main cast).
name.basics: Từ nconst tra ra tên thật (primaryName) và năm sinh (birthYear).

3. Trang Rotten Tomatoes

Link: https://www.rottentomatoes.com/

Scrape dữ liệu movie, thu thập:

- Tomatometer Score (Critic Score): Điểm số (%) từ các nhà phê bình chuyên nghiệp.
  Giá trị: 0-100%.
  Phân loại: Rotten (<60%), Fresh (>=60%), Certified Fresh
- Audience Score (Popcornmeter): Điểm số (%) từ khán giả xác thực.

Insight: Sự chênh lệch giữa hai điểm số này là cơ sở cho phân tích "Phim gây tranh cãi" (Divisive Movies).

- MPAA Rating: Phân loại độ tuổi (R, PG-13).

- Box Office: doanh thu phim

4. Trang The Movie Database (TMDB)

Link: https://www.themoviedb.org/

API Documentation: https://developer.themoviedb.org/docs

Nguồn dữ liệu bổ sung hình ảnh và nội dung, giải quyết các hạn chế về hình ảnh và tóm tắt nội dung mà MovieLens hay IMDb Datasets không cung cấp.

Đặc điểm:

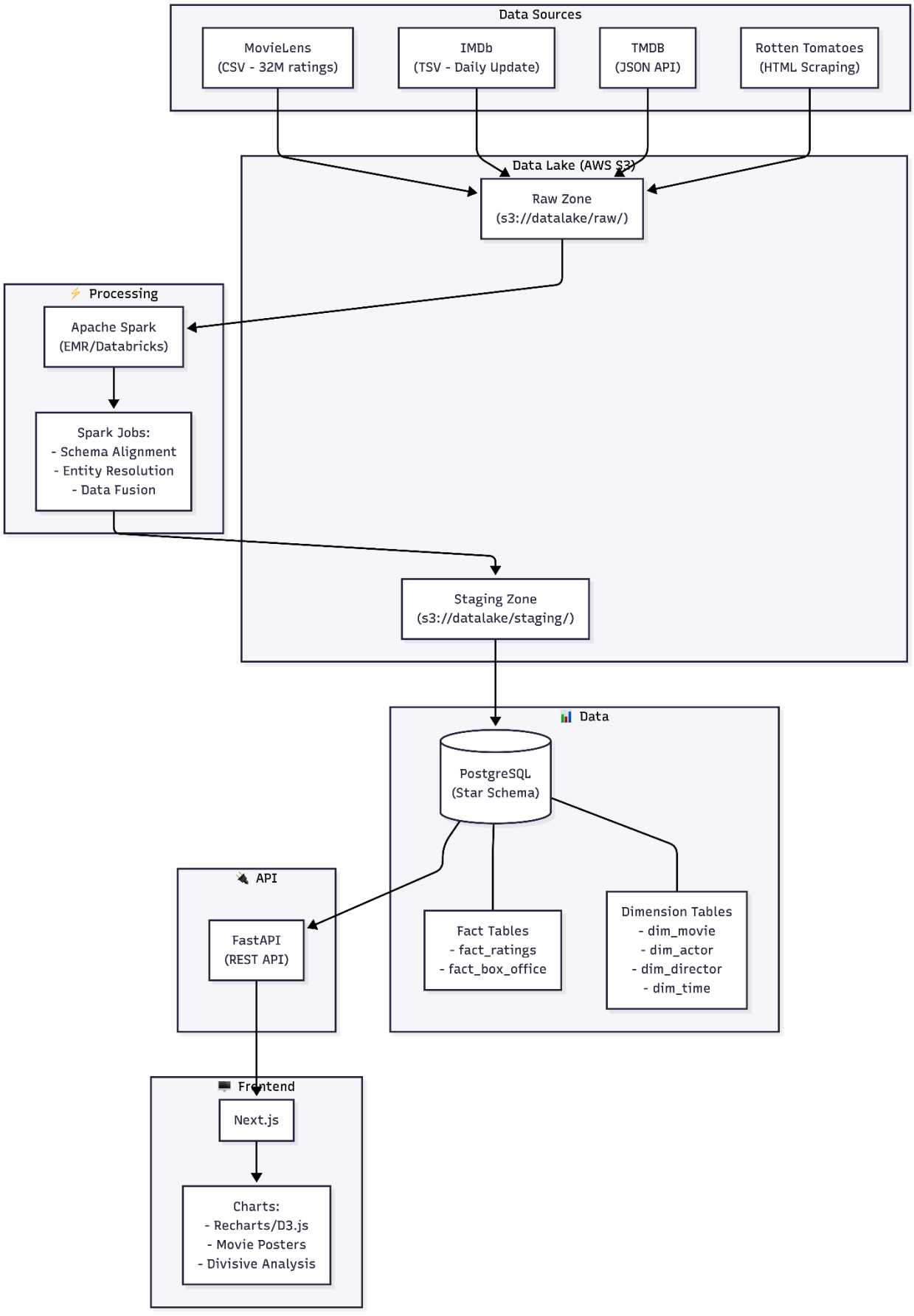
- Loại dữ liệu: Semi-structured (JSON response từ API).
- Cơ chế tích hợp: Sử dụng tmdbId từ file links.csv của MovieLens để query API

Tổng kết

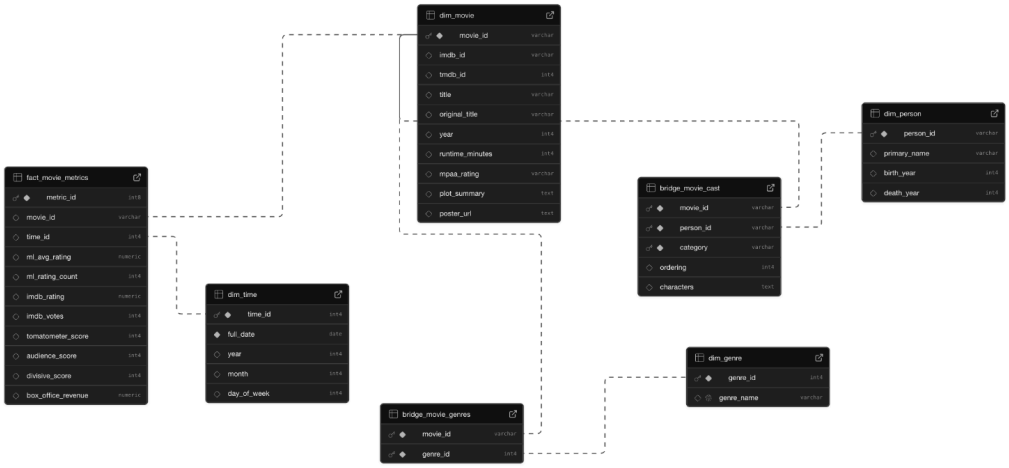
Nguồn	Loại dữ liệu	Cách lấy	Vai trò
MovieLens	Structured	Download (định kỳ)	Core: Hành vi người dùng, ID gốc
IDMb	Structured	Download (định kỳ)	Knowledge Graph: Metadata, Mạng lưới quan hệ (Cast/Crew)
Rotten Tomatoes	Unstructured (HTML)	Scrape	Insight: Chất lượng chuyên môn, Sentiment, Đối sánh thị hiếu.
TMDB	Semi-structured (JSON)	API	Hình ảnh, tóm tắt (plot)

- Hành vi (MovieLens)
- Kiến thức (IMDb)
- Thị hiếu & Chất lượng (Rotten Tomatoes)
- Hình ảnh & Ngữ cảnh (TMDB).

Kiến trúc hệ thống


Data Sources
MovieLens
(CSV - 32M ratings)
IMDb
(TSV - Daily Update)
TMDB
(JSON API)
Rotten Tomatoes
(HTML Scraping)
Data Lake (AWS S3)
Raw Zone
(s3://datalake/raw/)
⚡ Processing
Apache Spark
(EMR/Databricks)
Spark Jobs:
- Schema Alignment
- Entity Resolution
- Data Fusion
Staging Zone
(s3://datalake/staging/)
Data
PostgreSQL
(Star Schema)
API
FastAPI
(REST API)
Fact Tables
- fact_ratings
- fact_box_office
Dimension Tables
- dim_movie
- dim_actor
- dim_director
- dim_time
Frontend
Next.js
Charts:
- Recharts/D3.js
- Movie Posters
- Divisive Analysis

Mediated Schema



1. Bảng Dimension (Chiều dữ liệu)

1.1. Bảng dim_movie (Thông tin phim)

Tên trường	Kiểu dữ liệu	Khóa	Mô tả	Nguồn dữ liệu (Source Mapping)
movie_id	VARCHAR(20)	PK	Khóa chính nội bộ (Surrogate Key).	Tạo mới hoặc dùng MovieLens ID
movielens_id	INT		ID gốc từ MovieLens.	`movielens.movies.movieId`

imdb_id	VARCHAR(20)	ID gốc từ IMDb.	`movielens.links.imdbId`
tmdb_id	INT	ID gốc từ TMDB.	`movielens.links.tmdbId`
title	VARCHAR(500)	Tên chính của phim.	Ưu tiên `imdb.title.basics.primaryTitle`
original_title	VARCHAR(500)	Tên gốc của phim.	`imdb.title.basics.originalTitle`
year	INT	Năm phát hành.	`imdb.title.basics.startYear`
runtime_minutes	INT	Thời lượng phim (phút).	`imdb.title.basics.runtimeMinutes`
mpaa_rating	VARCHAR(10)	Phân loại độ tuổi (G, PG, R...).	`rottentomatoes.mpaa_rating`
plot_summary	TEXT	Tóm tắt nội dung phim.	`tmdb_api.overview`
poster_url	TEXT	Link ảnh poster.	`tmdb_api.poster_path`

1.2. Bảng dim_person

Tên trường	Kiểu dữ liệu	Khóa	Mô tả	Nguồn dữ liệu (Source Mapping)
person_id	VARCHAR(20)	PK	Mã định danh người (IMDb nconst).	`imdb.name.basics.nconst`
primary_name	VARCHAR(255)		Tên đầy đủ.	`imdb.name.basics.primaryName`
birth_year	INT		Năm sinh.	`imdb.name.basics.birthYear`
death_year	INT		Năm mất (nếu có).	`imdb.name.basics.deathYear`
primary_profession	VARCHAR(255)		Nghề nghiệp chính.	`imdb.name.basics.primaryProfession`

1.3. Bảng dim_genre

Tên trường	Kiểu dữ liệu	Khóa	Mô tả	Nguồn dữ liệu (Source Mapping)
genre_id	SERIAL	PK	Mã thể loại tự tăng.	Hệ thống tự sinh

genre_name	VARCHAR(50)	Tên thể loại (Action, Horror...).	`imdb.title.basics.genres`

1.4. Bảng dim_time

Tên trường	Kiểu dữ liệu	Khóa	Mô tả	Nguồn dữ liệu (Source Mapping)
time_id	INT	PK	Mã thời gian (Format: YYYYMMDD).	Hệ thống tự sinh
full_date	DATE		Ngày đầy đủ.	Generated
year	INT		Năm.	Generated
quarter	INT		Quý trong năm.	Generated
month	INT		Tháng.	Generated
day_of_week	INT		Thứ trong tuần.	Generated

---

2. Bảng Fact (Bảng dữ liệu sự kiện/Thống kê)

2.1. Bảng fact_movie_metrics

Tên trường	Kiểu dữ liệu	Khóa	Mô tả	Nguồn dữ liệu (Source Mapping)
metric_id	BIGSERIAL	PK	Khóa chính bảng Fact.	Hệ thống tự sinh

movie_id	VARCHAR(20)	FK	Tham chiếu `dim_movie`.	
time_id	INT	FK	Ngày thu thập dữ liệu (Snapshot Date).	Tham chiếu `dim_time`
ml_avg_rating	DECIMAL(3,2)		Điểm trung bình User MovieLens.	Aggregate từ `movielens.ratings`
ml_rating_count	INT		Tổng số lượt vote MovieLens.	Count từ `movielens.ratings`
imdb_rating	DECIMAL(3,1)		Điểm IMDb (Scale 10).	`imdb.title.ratings.averageRating`
imdb_votes	INT		Số lượt vote IMDb (độ phổ biến).	`imdb.title.ratings.numVotes`
tomatometer	INT		Điểm nhà phê bình (Scale 100).	`rottentomatoes.tomatometer_score`
audience_score	INT		Điểm khán giả RT (Scale 100).	`rottentomatoes.audience_score`
divisive_score	INT		Độ tranh cãi (ABS(Critic - Audience)).	Tính toán: `ABS(tomatometer - audience_score)`
box_office	DECIMAL(15,2)		Doanh thu phòng vé (USD).	`rottentomatoes.box_office`

3. Bảng Bridge (Bảng cầu nối Many-to-Many)

3.1. Bảng bridge_movie_cast

Liên kết Phim và Người, xác định vai trò cụ thể.

Tên trường	Kiểu dữ liệu	Khóa	Mô tả	Nguồn dữ liệu (Source Mapping)
movie_id	VARCHAR(20)	PK, FK	Tham chiếu `dim_movie`.	`imdb.title.principals.tconst`
person_id	VARCHAR(20)	PK, FK	Tham chiếu `dim_person`.	`imdb.title.principals.nconst`
category	VARCHAR(50)		Vai trò (actor, director, actress).	`imdb.title.principals.category`
ordering	INT		Thứ tự xuất hiện (1 = Main role).	`imdb.title.principals.ordering`
job	VARCHAR(255)		Chi tiết công việc (nếu có).	`imdb.title.principals.job`
characters	TEXT		Tên nhân vật trong phim.	`imdb.title.principals.characters`

3.2. Bảng bridge_movie_genres

Tên trường Kiểu dữ liệu Khóa Mô tả Nguồn dữ liệu (Source Mapping)

movie_id VARCHAR(20) PK, FK Tham chiếu dim_movie.

genre_id INT PK, FK Tham chiếu dim_genre.