Analyzing Movie Ratings

A Data Engineering Approach

DSCI – Mid term Project

Team Members

Rachana Chinthanippu
(Team Leader & Data Engineer)

Selva Vigneshwar Amuthan
(Data Architect)

Nuthan Kishore Maddineni
(Python Developer)

Kaleswara Manikanta Daddanala
(Data Analyst)

THE PROBLEM

- Fragmented movie rating sources (e.g., IMDb, Rotten Tomatoes) and inconsistent rating scales hinder efficient decision-making and analysis. Users face scattered data, confusion, and time-consuming research, highlighting the need for a unified platform for aggregated, normalized movie ratings and reviews.

CHALLENGES

Inconsistent Data Formats

- Ratings data across platforms (IMDb, Rotten Tomatoes, etc.) often vary in format and scale, complicating aggregation and analysis.

Large Volume of Data

- Handling and processing a massive volume of movie ratings data can lead to performance bottlenecks, especially with real-time analysis needs.

Data Quality Issues

- The dataset may contain incomplete or erroneous records, such as duplicate entries, missing values, or incorrect ratings, affecting the reliability of insights.

SOLUTION

Standardization and ETL Processes

★ Develop robust ETL (Extract, Transform, Load) pipelines to standardize data formats and scales, ensuring consistency across sources for accurate analysis.

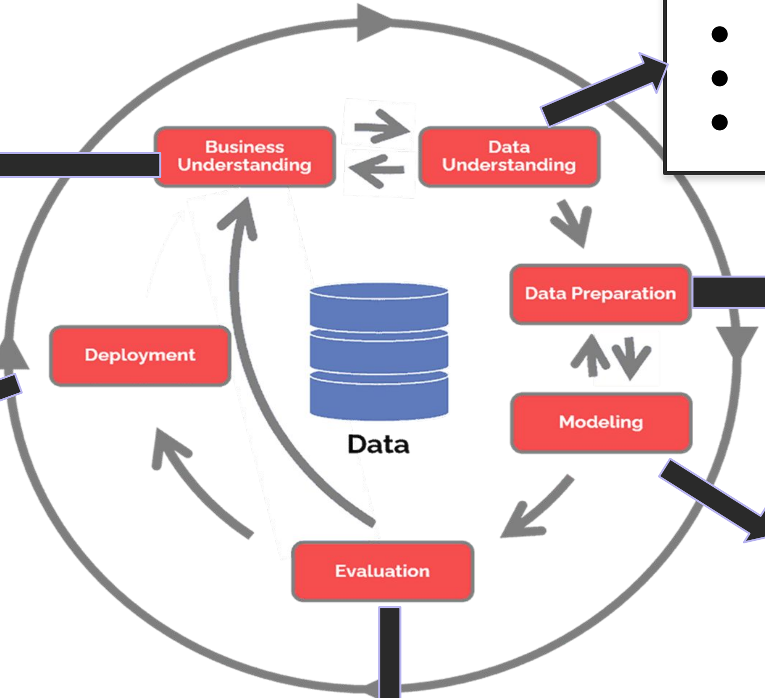
Big Data Technologies

★ Utilize big data processing frameworks to efficiently handle and analyze large datasets, ensuring scalability and performance.

Data Cleansing and Deduplication

★ Implement data cleansing procedures to correct or remove inaccurate records and apply deduplication strategies to eliminate redundancies, enhancing data quality for analysis.

CRISP - DM



- Determine business objectives
- Assess situation
- Determine Data mining goals
- Produce object plan

- Collect initial data
- Describe data
- Explore data
- Verify data quality

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

- Select modeling techniques
- Generate test design
- Build model or Assess model

- Assess the model
- Review the steps

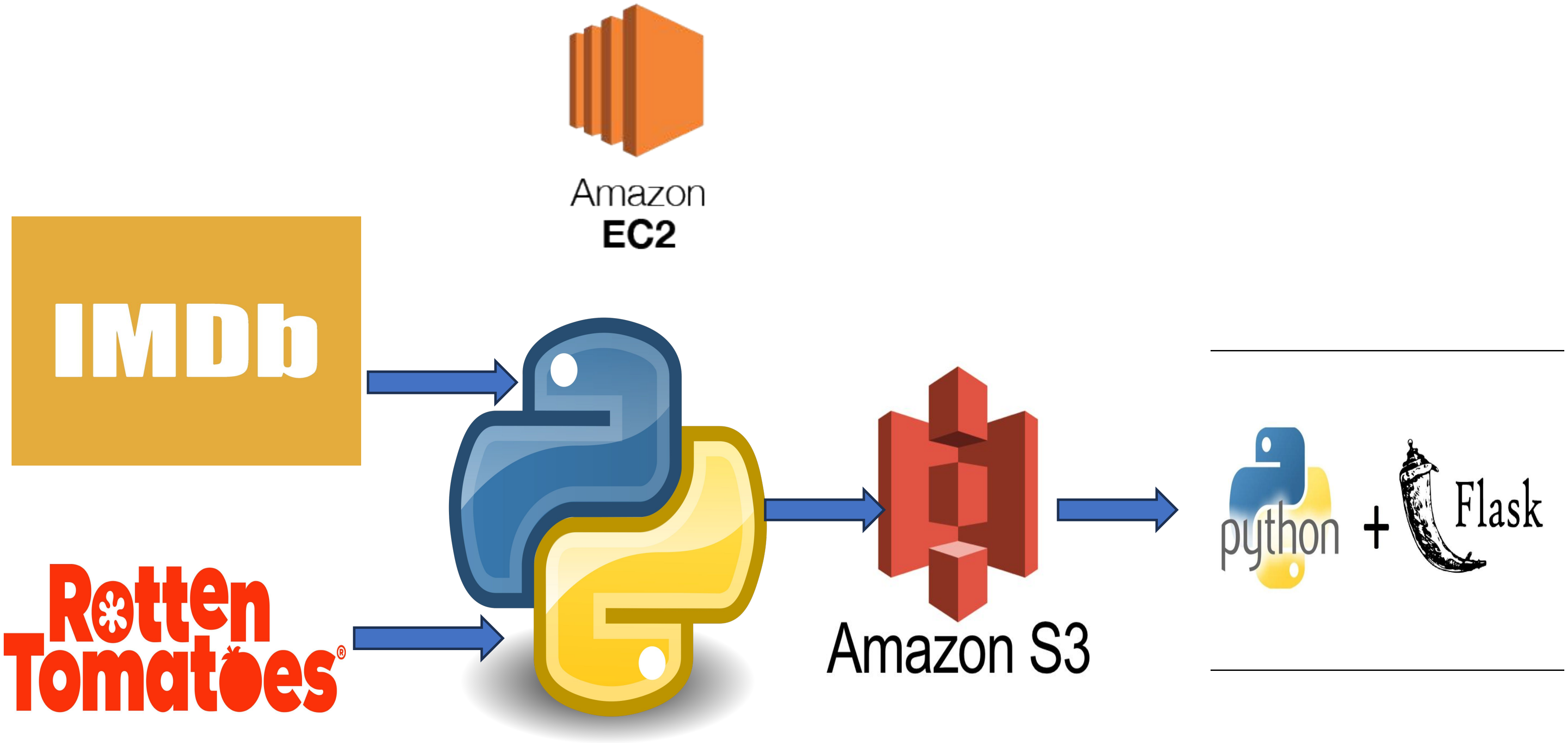
- Plan deployment
- Plan monitoring
And maintenance
- Produce final report
- Review project

Data Sources

- **IMDB**
- **Rotten Tomatoes**

We will get data from these two websites through web scraping.

We will create a DataFrame with columns **Movie name**, **Release year**,**rating from different platforms.**



Web Scraping & Data Processing

Flask Application

THANK YOU!