Movie Recommendation System

Phoebe Chen, Tim Deng

Background

“To create a system that can continuously take in multiple data sources on movie characteristics & reviews and provide most relevant output based on user searches, in addition to providing recommendations based on user profile and feedback.”

- A one stop shop for users looking for movie information and recommendations
- Information include screening times and locations, basic movie background (cast, length, director, etc.)
    - Users can click on links to purchase tickets or to stream
- Users will be prompted to create a user profile
- As more user-specific data is collected, movie recommendations can be made based on user profile, feedback, and past searches

Interface Prototype

General movie information

Movie trailer

Film safety rating and movie genre users can search for

Links to interviews, theaters, and streaming services

User reviews for more detailed recommendations

Data Sources

The New York Times | Rotten Tomatoes | IMDb

Procurement

- Create NYT developers account to obtain API key
- Scrape Rotten Tomatoes from website and Kaggle
- Extract IMDb from website
- Shows content from streaming services
- Scrape data from stream services (e.g. Netflix, Hulu etc.) to obtain available movies & links
- Scrape data from local theaters to identify and provide links to movie times

System

- Continuously retrieve updated datasets from sources
- Update movie ratings in existing dataset
- Add to existing dataset as more movies come out

Future Source

- User-created profiles
- Collect user feedback on recommendations as user base grows
- Collect past searches

Data Infrastructure Comparisons

	Use	Data Sources	Size	Setup Time	Data Held
Data Warehouses	Keep data available in structured format	Multiple internal & external sources	Petabytes	Years	Raw data, metadata, summary data
Data Lakes	Keep data stores; transaction-oriented	From databases, data warehouses, data marts	Petabytes	Months ~ Years	Any type of data structure & any format
Data Marts	Pick data for use of specific line of business	Relatively few sources	< 100GB	3-6 Months	Summary data

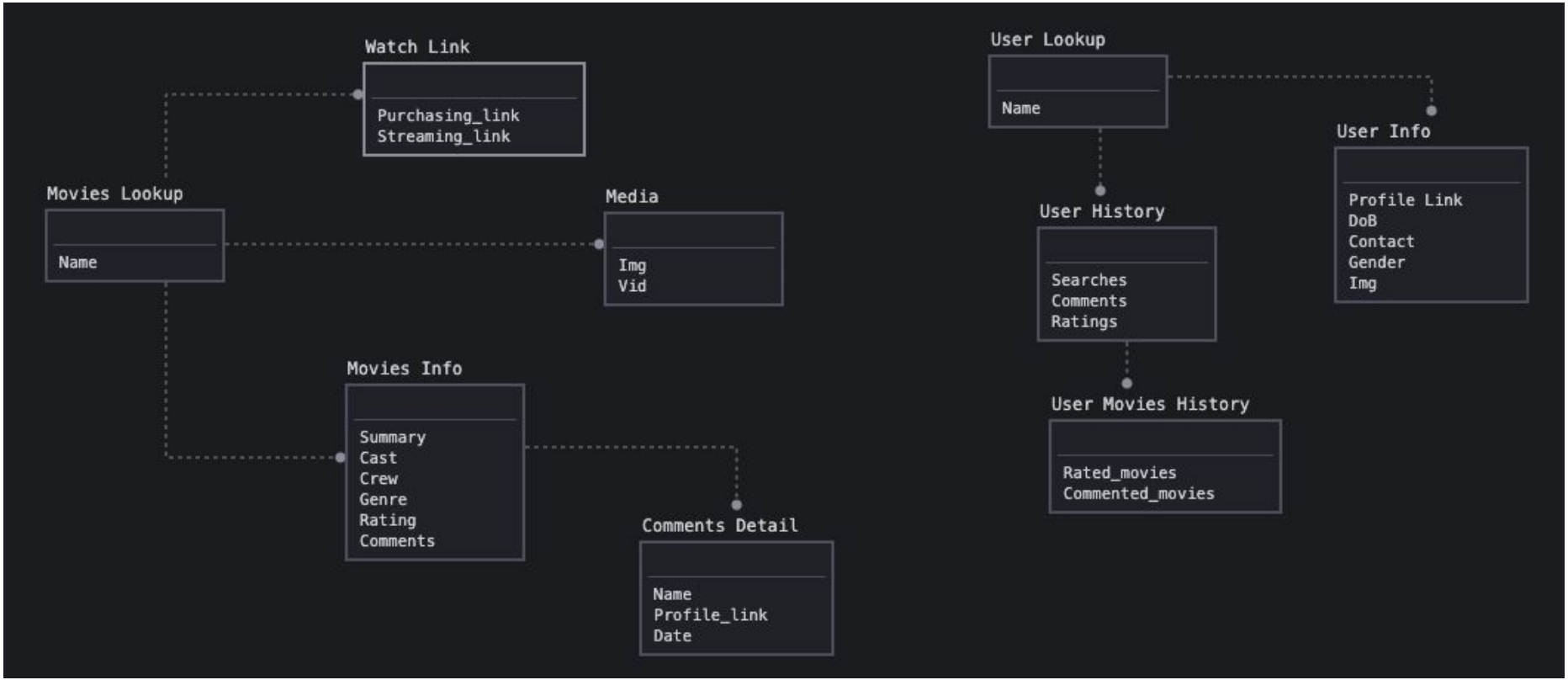
NoSQL Database Comparisons

	Primary Database Model	MapReduce	Application Scenarios	Advantages	Clients
ElasticSearch	Search Engine	EH-Hadoop Connector	Search engine, text search, spell checker, log analytics	Fast full-text search speed, text analysis capability	Shopify, Instacart, Udemy, Robinhood
Neo4j	Graph Database	No	Recommendation engine, identity & access management, graph-based search, graph analytics	Requires less hardware, platform agnostic, first graph ML for enterprise	eBay, WalmartUBS, Cisco
Cassandra	Wide Column Store	Yes	Fraud detection, recommendation engine, product catalogs	Scalability and performance, used by 40% of the Fortune 100, operational simplicity	Netflix, Uber, Instagram, Reddit
MongoDB	Document Store	Yes	Single view, catalogs, gaming, payment processing	Document storage, highly scalable, consistent developer experience	Forbes, Toyota, KPMG, EA

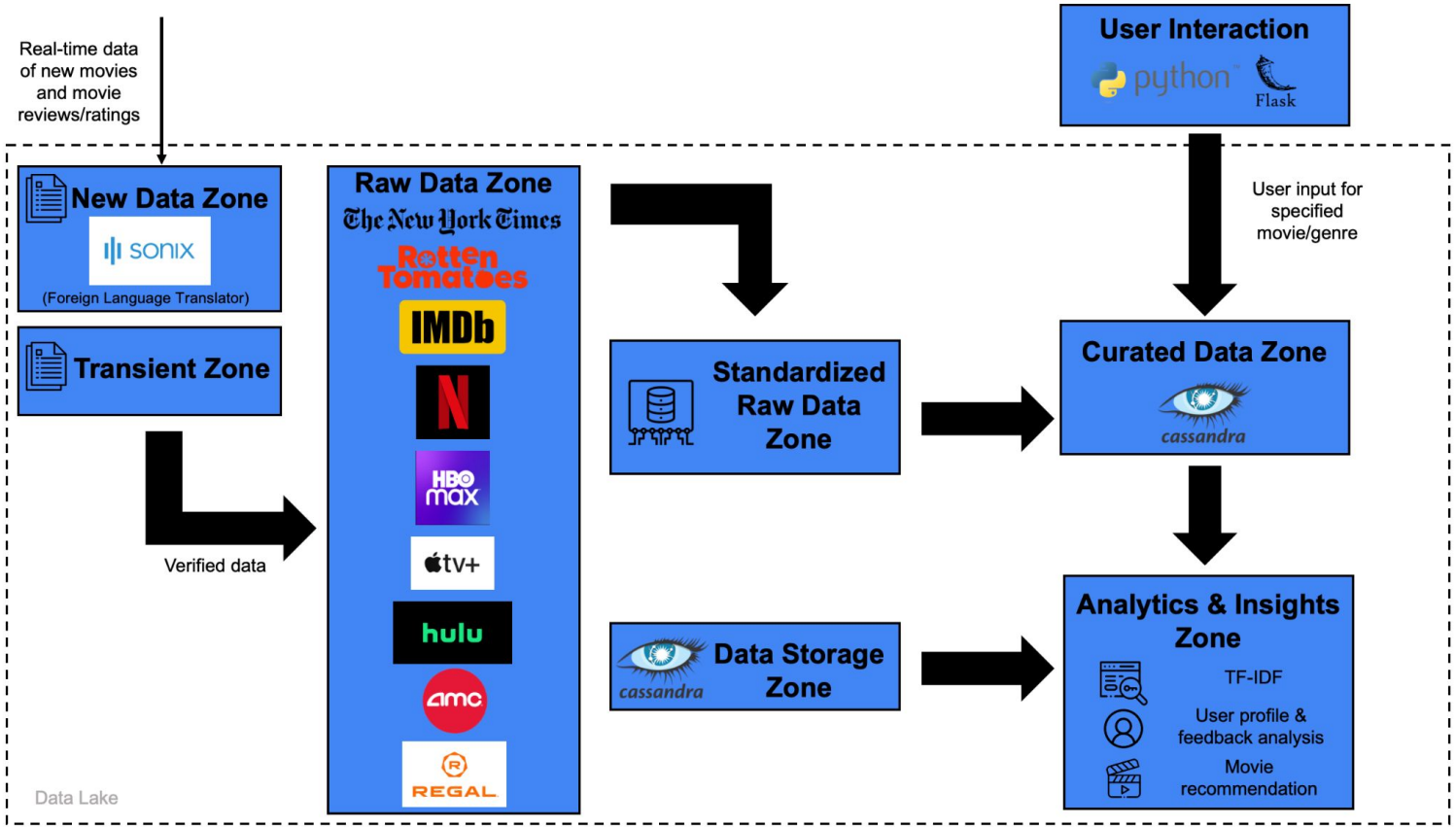
Design Choices & Selected Technologies

- Data Lake
  - Has to store text, pictures, & videos
  - No common schema across data sources
  - Supports advanced analytics
  - Highly scalable
- Cassandra Database
  - Proven success across enterprises
  - Homogenous environment
    - Required components for operations are built-in
  - Easy-to-integrate core applications
  - Data reliability
    - Stores data, makes copies and stores them in other locations
- Flask Application
  - Input: User profile, user searches of genre/characteristics
  - Output: Positively-rated movies within search context, recommended movies
  - TF-IDF: To identify important keywords in movie reviews to use as recommendation basis to other users
  - Will become feedback loop as users provide feedback & ratings on movies

Schema Diagram



Data Lake Architecture



Data Governance

- Data Security
  - Terms of service to inform users collection of their names, emails, search history etc.
  - No user data will be sold to third parties
- Data Scalability
  - Data Lake’s ability to incorporate large amount of data
  - Continuously incorporate multiple data sources over time
- Data Integrity
  - System will be made to continuously retrieve updated databases to ensure data accuracy
  - Data Lake’s ability to store real-time data
- Data Availability
  - Available data on social media sites e.g. NYT, IMDb etc.

Cost Implications

Cassandra	• Data engineer: $200,000/year • Support: $90,000/year • Recruiting: $58,000/year • Equipment: $30,000/year • Others: $5000/year
Data Lake	• Data storage: $0.022/GB • Data retrieval: $0.0004/1000 requests
Others: Marketing, Labor, Legal, Web development etc.	
Time to set up Data Lake & Cassandra	
Time to Increase Scalability	

Evaluation Criteria

- Data loading & retrieval time
- Accuracy of output generation based on user input
- Data volume & reliability of user feedback on movies
- Consumer satisfaction on output relevance
- Scalability of database systems
- Computational cost & time for modeling

Future Recommendations

- Incorporate user ratings and other user metrics to improve development of recommendation engine
- Develop foreign language translation capabilities to include foreign movies and reviews in our dataset using tools such as Sonix
- Expand existing features to include TV shows

THANKS!

Any questions?

Works Cited

Amazon Web Services, Inc. (n.d.). *Amazon S3 pricing*. aws. https://aws.amazon.com/s3/pricing/?nc=sn&loc=4.

Amazon.com, Inc. (n.d.). *Fast & And Furious 9 The Fast Saga Movie Poster 2 Sided Original Final 27x40*. Amazon. https://www.amazon.com/FURIOUS-MOVIE-POSTER-Sided-ORIGINAL/dp/B092T77NFX.

Apple Inc. (n.d.). *Apple Tv Plus Logo*. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Apple_TV_Plus_Logo.svg.

Cassandra System Properties. Cassandra system properties. (n.d.). https://db-engines.com/en/system/Cassandra.

FilmIsNow Movie Bloopers & Extras. (2021). *F9: The Fast Saga | Cast & Filmmakers Interview (Part 1)*. Youtube. Youtube. https://www.youtube.com/results?search_query=f9+interview.

Hbo Max logo. (n.d.). HBO Max. play.hbomax.com.

Instaclustr Pty Ltd . (n.d.). *The true cost of Do It Yourself Cassandra Implementations*. Instaclustr. https://www.instaclustr.com/the-true-cost-of-do-it-yourself-cassandra-implementations/.

New York Times. (n.d.). *Movie Reviews API*. New York Times Developer. https://developer.nytimes.com/docs/movie-reviews-api/1/overview.

Rotten Tomatoes movies and critic reviews dataset. Kaggle. (n.d.). https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/code.

Sonix, Inc. (n.d.). Sonix. https://sonix.ai/.

Universal Pictures India. (2021). *Fast & Furious 9 - Official Telugu Trailer 2 (Universal Pictures) Hd*. Youtube. Youtube. https://www.youtube.com/results?search_query=F9+official+trailer.

Welch, C. (2016). *Netflix logo*. The Verge. Vox Media, LLC. https://www.theverge.com/2016/6/20/11979948/netflix-new-icon-logo