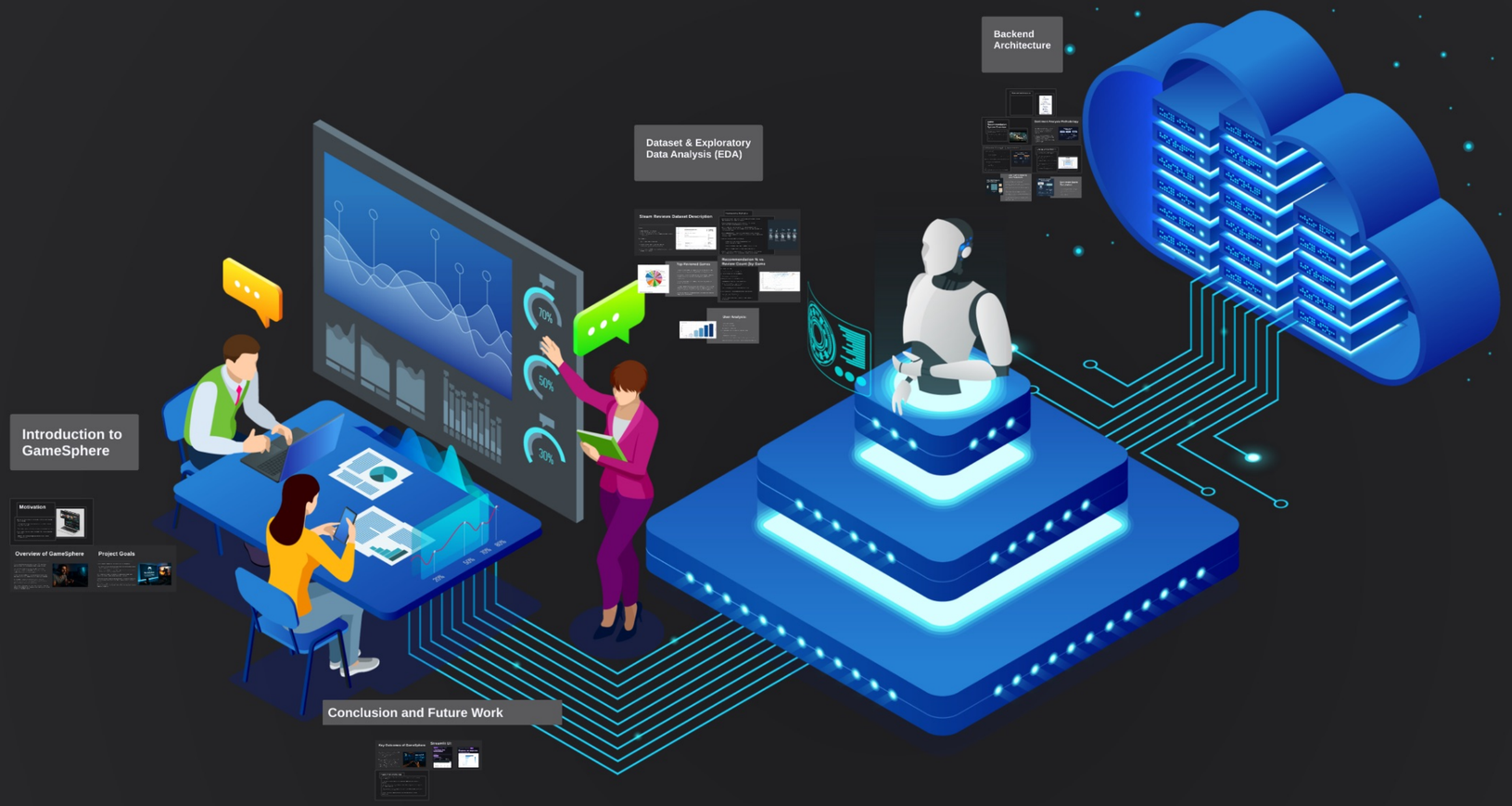


Group 11

GameSphere: Smart Gaming Analytics & Recommendations

Leveraging Big Data and Machine Learning to Enhance Game Discovery and User Engagement

Introduction to GameSphere

Motivation

- Steam contains millions of user reviews, making manual analysis overwhelming.
- GameSphere leverages advanced analytics to extract valuable insights from this data.
- Key insights include sentiment trends and game popularity metrics.
- Helps players discover games that match their preferences more effectively.
- Enhances the overall gaming experience with data-driven recommendations.

Project Goals

The main goal of GameSphere is to develop a dual-purpose system:

- 1. A game recommendation engine that suggests personalized game titles based on user activity and preferences.
- 2. An analytics dashboard to provide insights into user reviews, sentiment trends, game popularity, and player behavior.

Both components are fully integrated into an interactive Streamlit web application, offering a seamless and intuitive user experience.

The recommendation engine leverages collaborative filtering (ALS) to generate personalized suggestions, helping users discover games that align with their interests.

The dashboard enables real-time exploration of game trends, allowing users to filter and visualize data across various dimensions (such as playtime, sentiment, and review frequency).

Overview of GameSphere

It processes a large-scale dataset of over 8 GB, containing millions of user reviews from the Steam gaming platform.

The platform integrates advanced big data processing technologies (PySpark and HDFS) with machine learning models to generate actionable insights.

It offers personalized game recommendations based on user behavior, preferences, and collaborative filtering techniques.

GameSphere also provides in-depth gaming analytics, visualizing trends such as game popularity, user sentiment, playtime statistics, and review dynamics.

The system enables users to explore data through a clean and interactive dashboard built using Streamlit, making it accessible even to non-technical users.

Dataset & Exploratory Data Analysis (EDA)

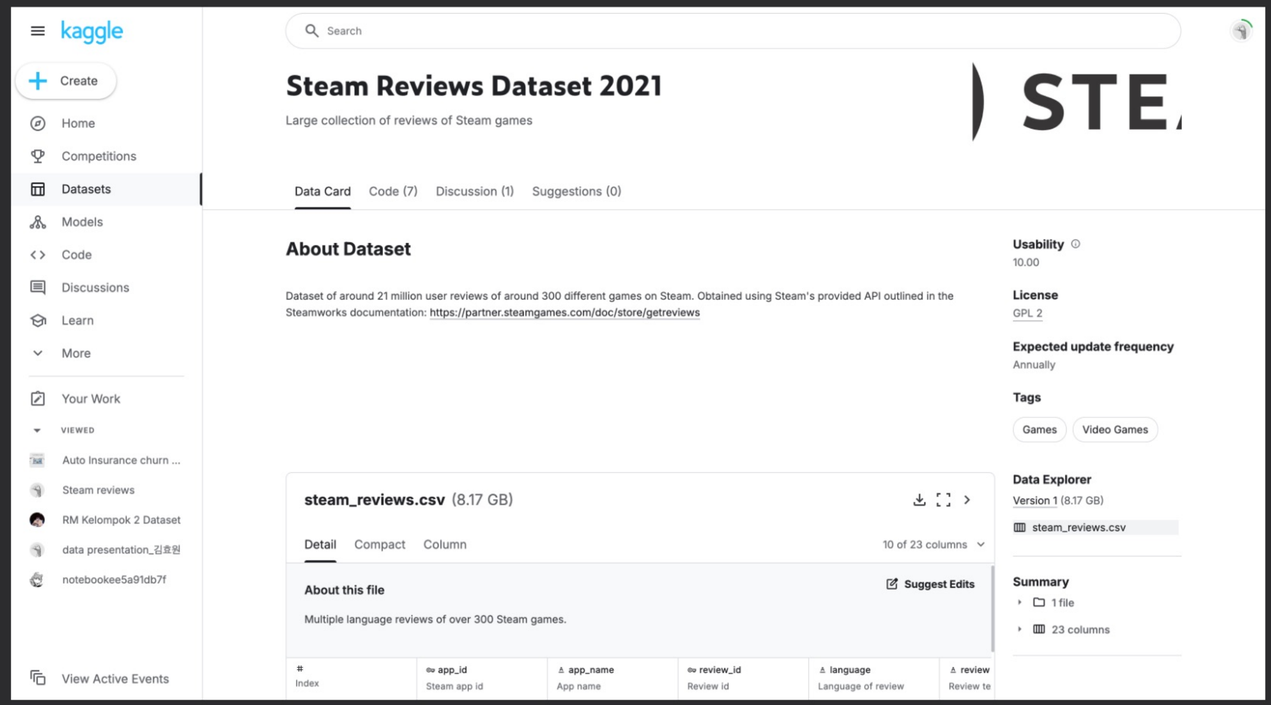
Steam Reviews Dataset Description

Dataset:

- Steam Reviews 2021 (Kaggle)
- Size: ~40+ million reviews (~8 GB)
- Shape : ~40+ million rows × 23 columns (40+ Million rows × 23 colums)

Key Features:

- app_id, app_name (Game Info)
- author_steamid, author_playtime_forever, author_num_games_owned (User Info)
- review, recommended, votes_helpful, timestamp_created, language (Review Info)



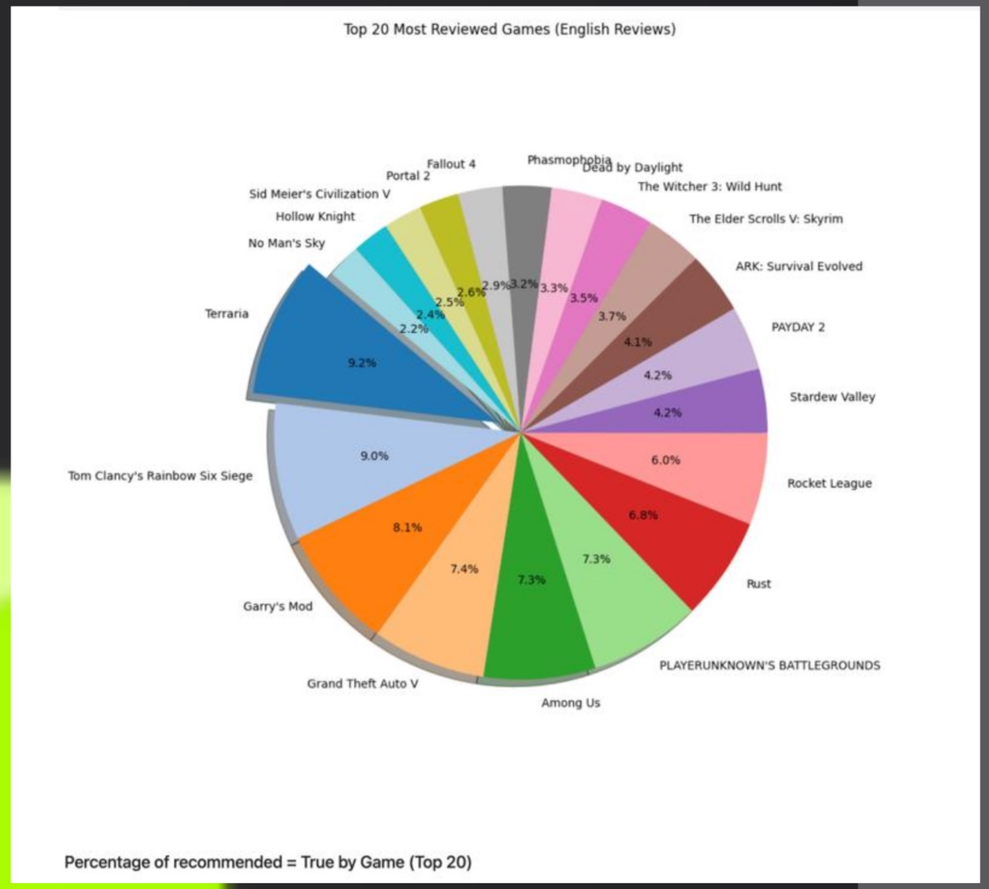
Preprocessing Highlights:

- A structured directory was set up in HDFS, organized into folders for raw data, processed data, models, and outputs.
- Column names were standardized for consistency (for example, author.steamid was renamed to author_steamid).
- Missing values were handled carefully by removing rows with nulls in essential fields such as app_id and review. Numeric nulls were replaced with zero where applicable.
- Text data was thoroughly cleaned by converting to lowercase, removing emojis, HTML tags, and special characters to ensure high-quality input for the sentiment model.

Data types were optimized for performance:

  - Integer types were used for fields like app_id and author_num_games_owned.
  - Floating-point types were used for author_playtime_forever.
  - Boolean types were applied to the recommended column.

- The fully processed dataset was saved in Parquet format for efficient storage and retrieval, located at /processed/steam_review_english.parquet.

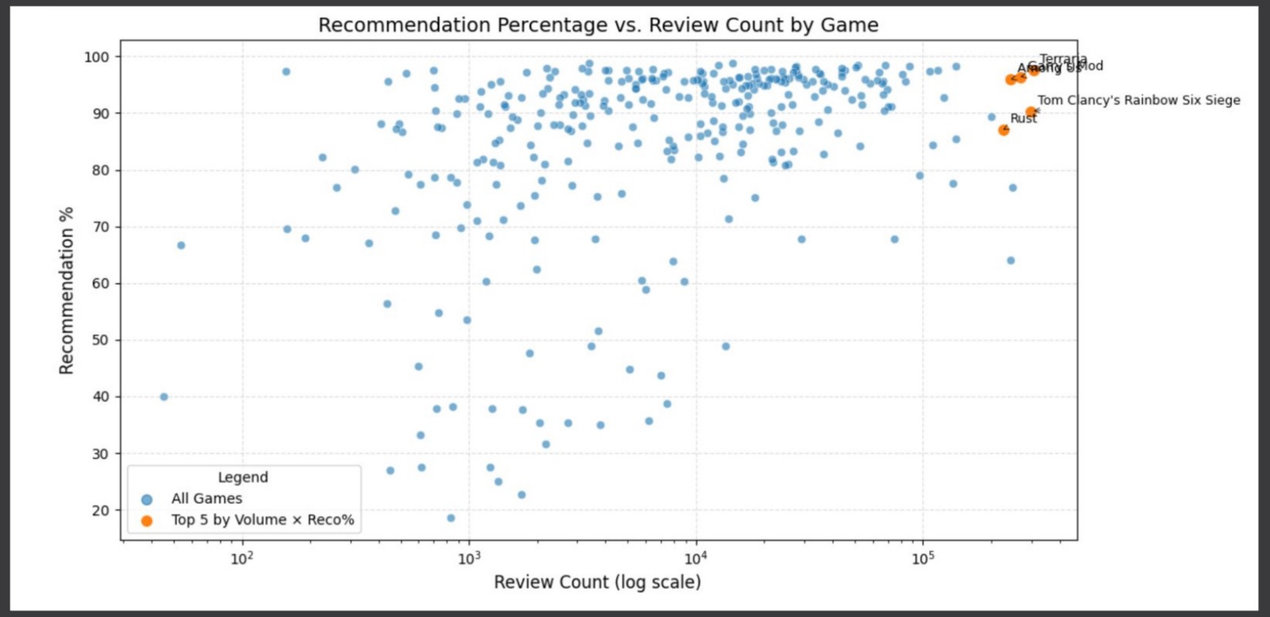


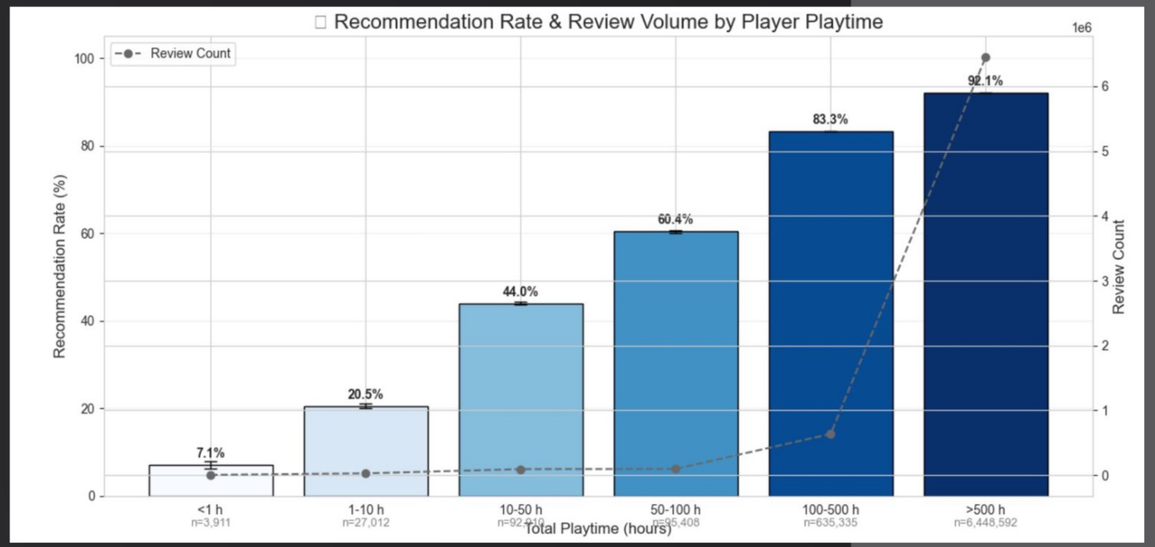
Top-Reviewed Games

- Terraria and Rainbow Six Siege dominate the most reviewed games, collectively representing 18.2% of all English reviews.
- Indie games (Terraria, Stardew Valley, Hollow Knight) together account for 15.8% of reviews, showing strong community engagement
- Survival games (Rust, ARK, Among Us) collectively make up 18.2% of the market
- The top 5 games alone represent over 40% of all English reviews
- Newer releases like Phasmophobia (3.2%) compete effectively against established franchises like The Elder Scrolls (3.7%)
- Multiplayer-focused titles generally receive more reviews than single-player experiences

Recommendation % vs. Review Count (by Game

- Each dot = one game:
- X-axis: Total number of reviews (log scale)
- Y-axis: % of reviews that are “recommended”
- Blue dots show overall distribution:
- Most games cluster between 60–95% rec-rate,
- but volume varies widely (from hundreds to millions)
- Orange dots highlight top 5 high-impact games:
  e.g., Terraria, Garry’s Mod, Tom Clancy’s...
- These games combine very high review counts and strong recommendation rates
- Top-right quadrant = strong community approval + massive reach
- Lower-right = popular but polarizing
  e.g., large player base but mixed opinions
- Helps spot standout games that are both widely played and highly recommended





User Analysis:

- Recommendation rate increases sharply with playtime
- <1h: only 7.1% recommend
- 1–10h: 20.5%, 10–50h: 44.0%
- 100–500h: 83.3%, >500h: 92.1%
- Review volume is heavily skewed toward high-hour players
- 500h group contributes ~6.4 million reviews
- <1h group has just ~3.9k reviews
- Veteran players dominate the dataset, heavily influencing overall sentiment
- Engagement appears tied to satisfaction – the more people play, the more they recommend

Backend Architecture

Flow and architecture:

Steam Reviews CSV:
- Source dataset with millions of game reviews.

Data Ingestion into HDFS:
- Uploads raw CSV files to HDFS (/data/raw directory).

Data Cleaning & Preprocessing (PySpark):
- Tasks include schema standardization, filtering, and text normalization.

Sentiment Analysis (DistilBERT via Hugging Face):
- Labels each review as POSITIVE or NEGATIVE with confidence scores.

Collaborative Filtering (Spark ALS):
- Builds a user-game interaction matrix to predict personalized recommendations.

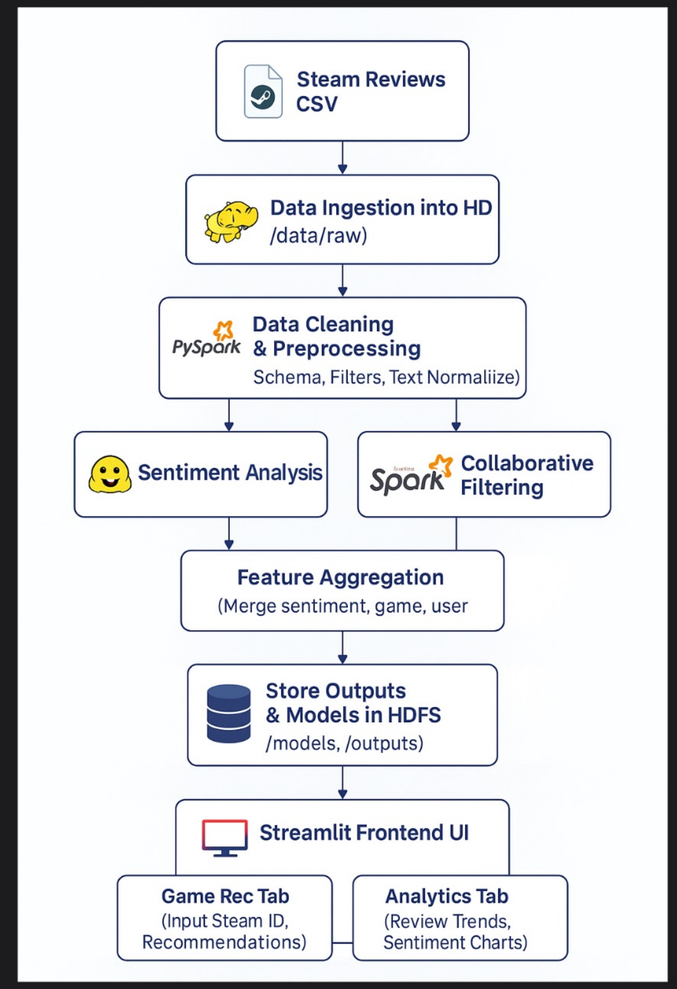
Feature Aggregation:
- Merges sentiment data, game info, and user data into a unified dataset.

Store Outputs & Models in HDFS:
- Saves processed data and trained models in /models and /outputs directories.

Streamlit Frontend UI:
- Interactive web app with two main tabs:

Game Rec Tab: Input Steam ID & view personalized game recommendations.

Analytics Tab: Explore review trends and sentiment visualizations.



Game Recommendation System Overview

GameSphere uses a recommendation system based on collaborative filtering.

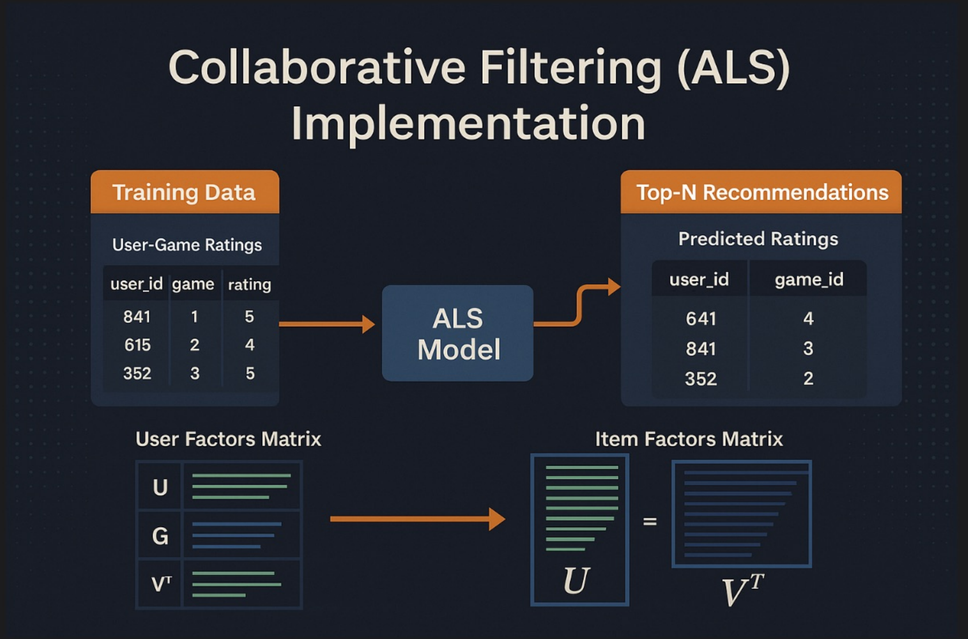
Provides personalized game suggestions for each user.

Leverages an implicit ratings matrix created from user reviews.

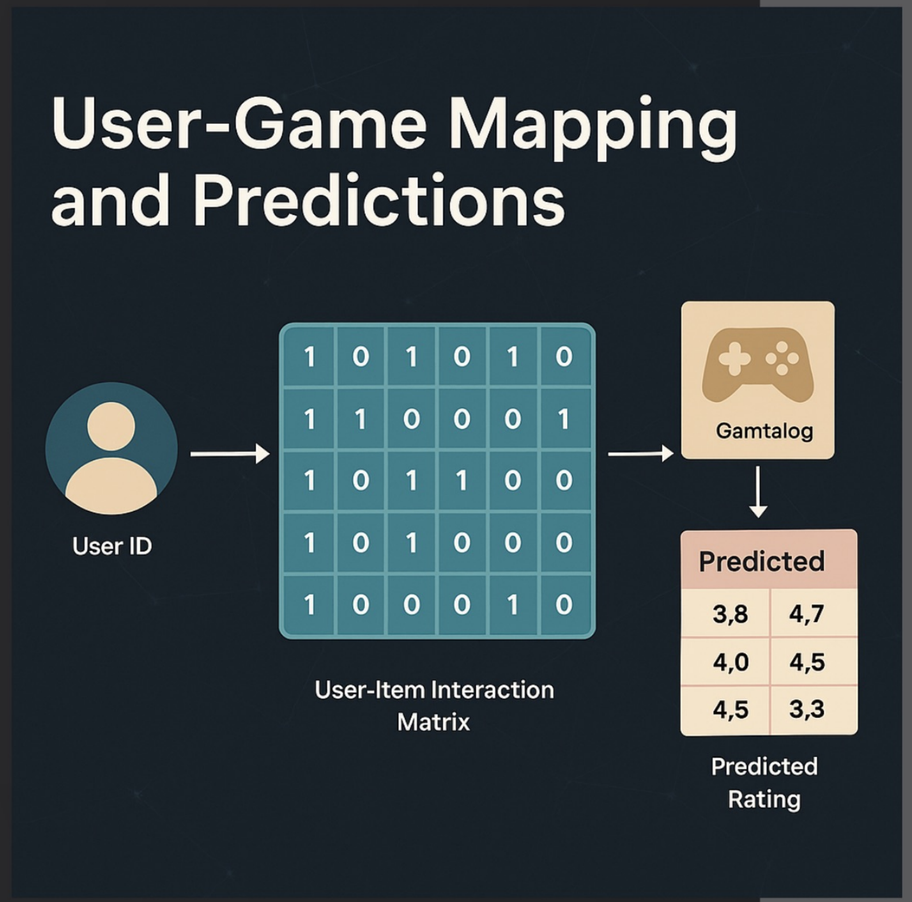
Enhances the relevance and accuracy of game recommendations.

Collaborative Filtering (ALS) Implementation

- The recommendation engine is powered by the Alternating Least Squares (ALS) algorithm.
- Steam's 'recommended' field is converted into numeric ratings:
  - 5 -Positive recommendation
  - 1 -Negative recommendation.
- Games with fewer than 200 reviews are removed to ensure meaningful data.
- Outlier users (abnormally high/low playtime) are filtered using the IQR method to maintain data integrity.
- The ALS model is trained with key parameters:
  - userCol: author_index
  - itemCol: app_index
  - ratingCol: rating.
- Hyperparameters (rank, regParam, maxIter) are fine-tuned for optimal performance.
- Output: Top 5 game recommendations per user are stored in HDFS for app integration.



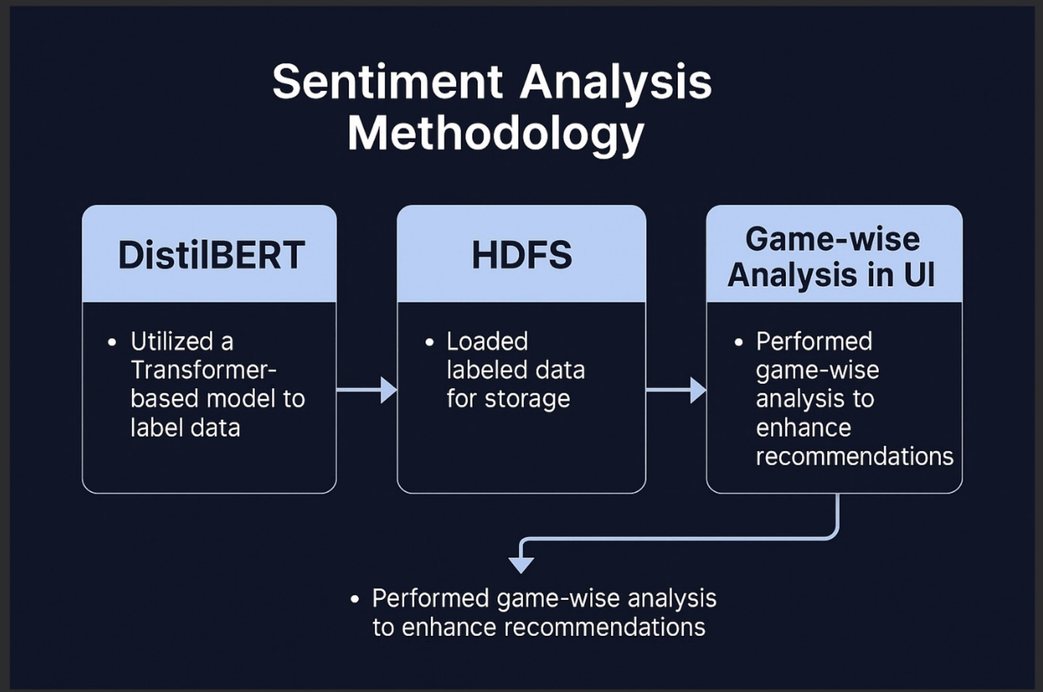
User-Game Mapping and Predictions



- Each user and game is mapped to numeric indices (via StringIndexer) for ALS matrix factorization.
- This mapping enables the creation of a user-item interaction matrix capturing user-game interactions at scale.
- The indexed dataset allows efficient computation and faster predictions for millions of users.
- ALS identifies latent features that connect user preferences with game attributes, boosting recommendation accuracy.
- The system generates personalized, data-driven suggestions that reflect true user interests.
- Final recommendations are seamlessly integrated into the GameSphere app, allowing users to explore their personalized game list in real-time.

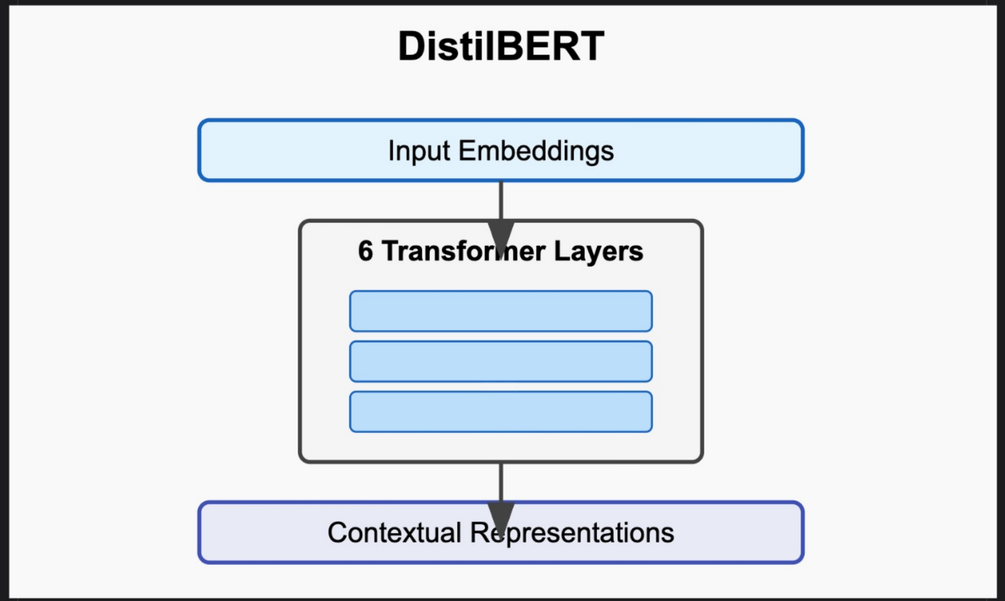
Sentiment Analysis Methodology

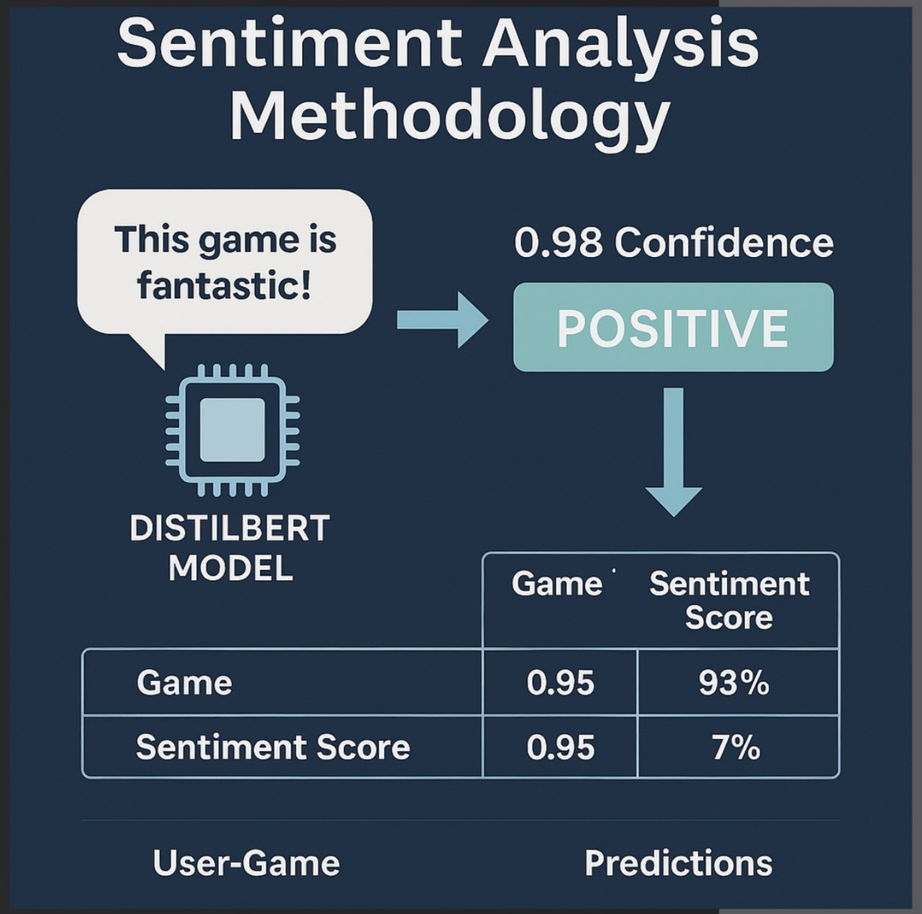
- GameSphere implements a robust sentiment analysis pipeline that classifies reviews as positive or negative.
- This dual approach enhances user recommendations by incorporating sentiment scores into the recommendation framework, adding depth to insights.



Choice of DistilBERT:

- Initially tested Spark MLlib models for sentiment analysis but results were limited in accuracy.
- Switched to DistilBERT, a Transformer-based model, to better capture nuanced emotions in reviews.
- DistilBERT provided highly accurate sentiment labeling (positive/negative) with confidence scores.
- Sentiment-labeled data was stored in HDFS for scalable processing.
- Integrated sentiment data into GameSphere to enable game-wise sentiment analysis.
- Results are visualized in the app UI, adding deeper context to each game's recommendation.
- This approach enhanced both the insight quality and user trust in recommendations.



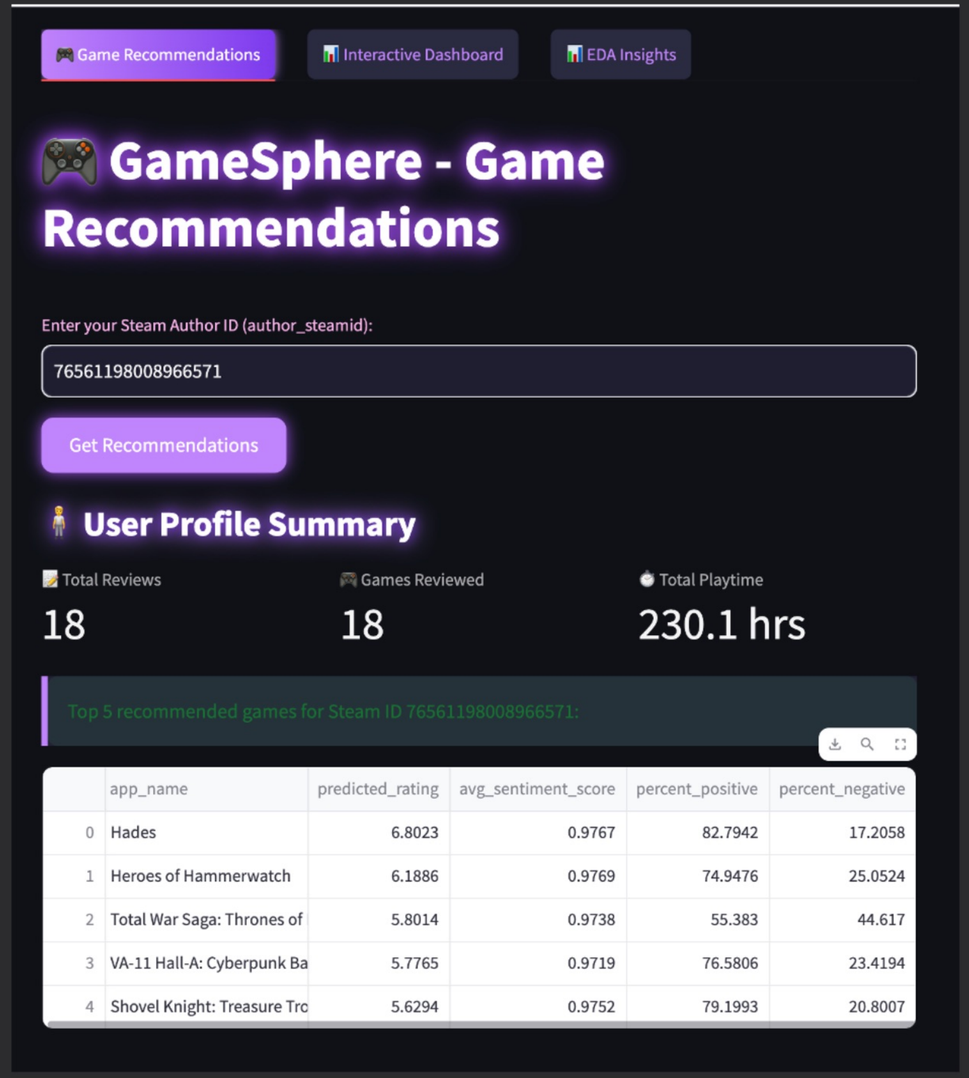


Sentiment Score Calculation

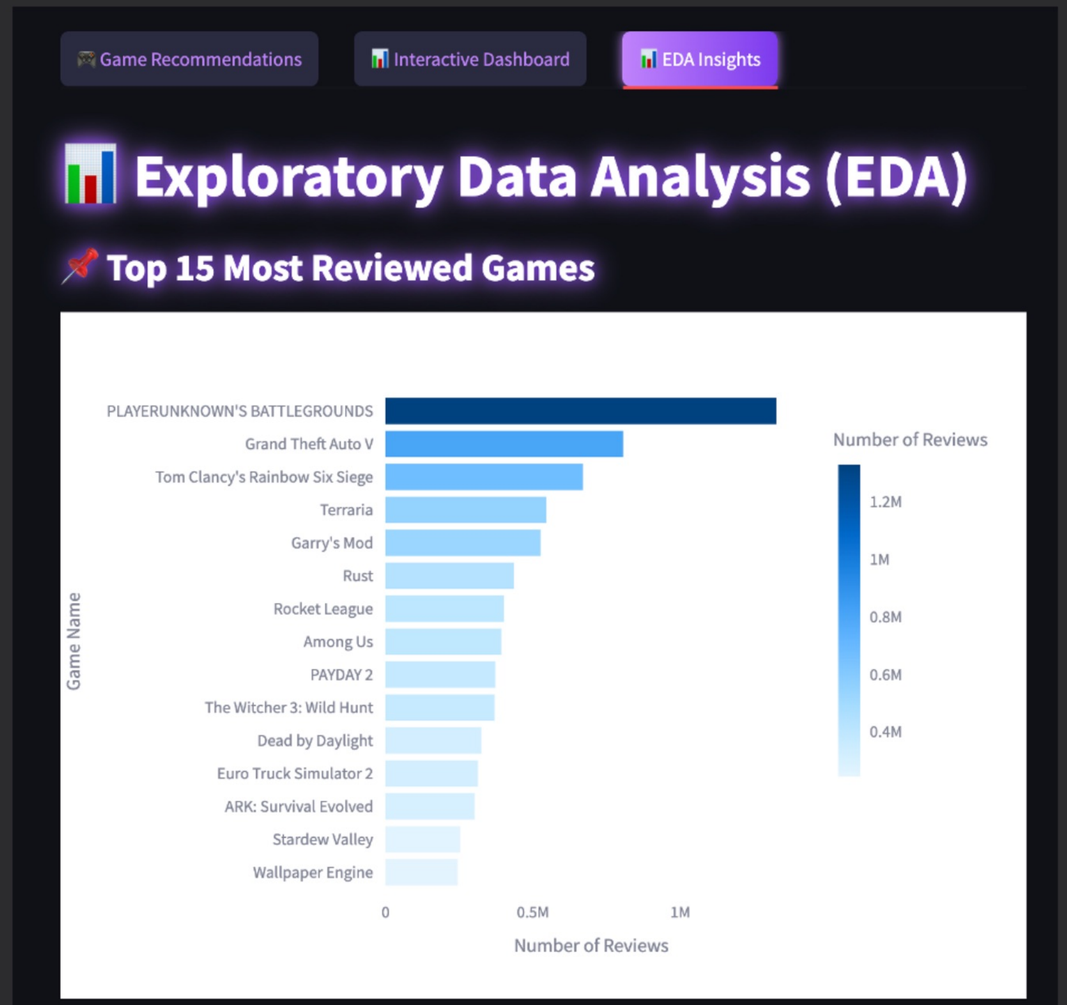
- Sentiment scores are calculated for each review using the DistilBERT model, generating confidence metrics alongside binary sentiment labels.
- These scores are aggregated at the game level, providing insights into overall player sentiment and game popularity.

Conclusion and Future Work

Streamlit UI:



	app_name	predicted_rating	avg_sentiment_score	percent_positive	percent_negative
0	Hades	6.8023	0.9767	82.7942	17.2058
1	Heroes of Hammerwatch	6.1886	0.9769	74.9476	25.0524
2	Total War Saga: Thrones of	5.8014	0.9738	55.383	44.617
3	VA-11 Hall-A: Cyberpunk Ba	5.7765	0.9719	76.5806	23.4194
4	Shovel Knight: Treasure Trc	5.6294	0.9752	79.1993	20.8007



Key Outcomes of GameSphere

- GameSphere effectively transformed 40+ million raw Steam reviews into actionable insights and recommendations.
- The synergy of sentiment analysis with collaborative filtering enhanced recommendation relevance, guiding players toward suitable game choices and providing gamers with a clear understanding of community sentiment.

Future Enhancements Ideas

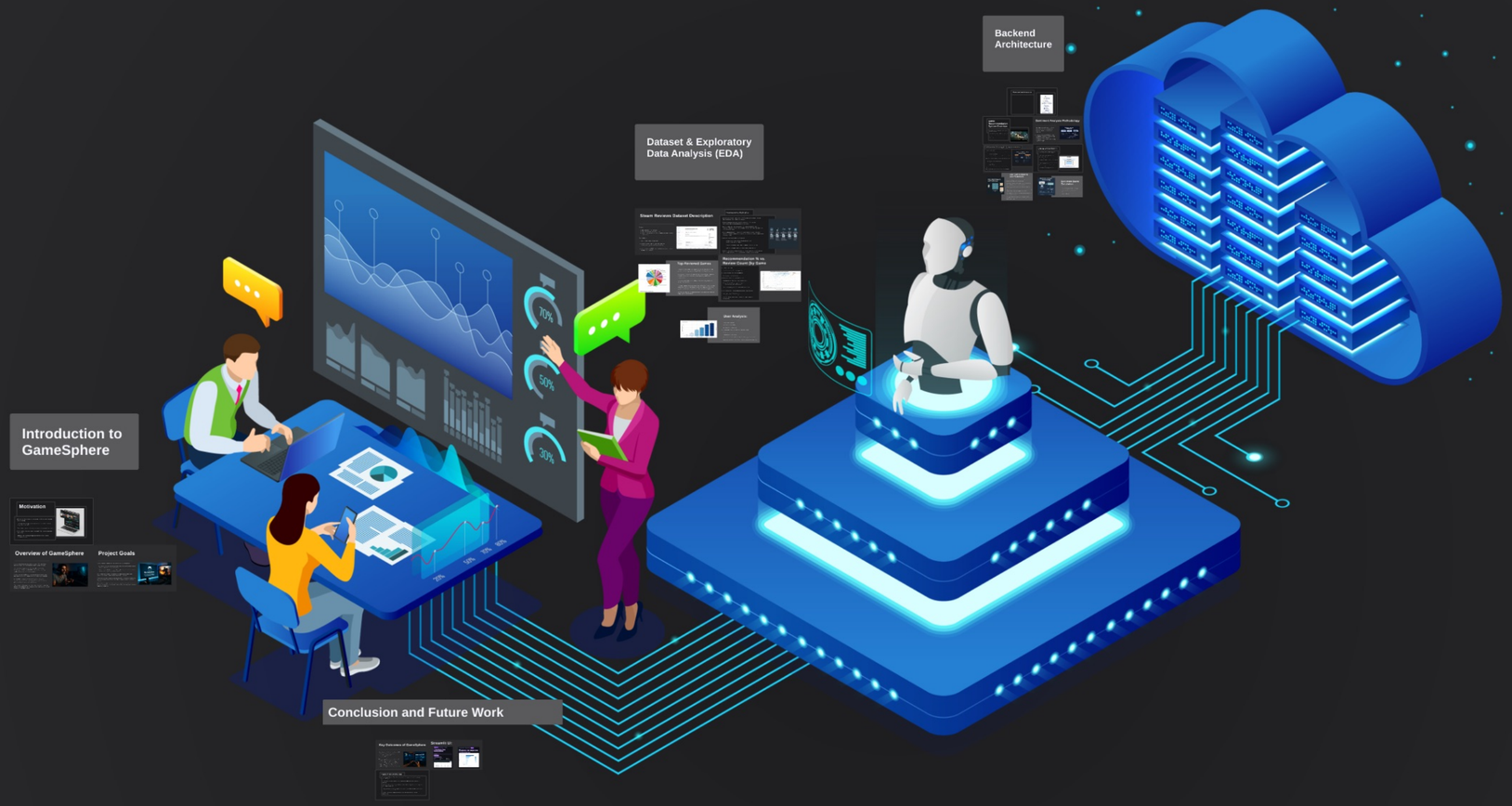
• Multilingual Support: Expand sentiment analysis to support reviews in languages beyond English.

• User-Level Insights: Enable tracking of personalized sentiment and behavioral patterns.

• Real-Time Processing: Integrate Kafka and Spark Streaming for live review ingestion and dashboard updates.

• Game Metadata Fusion: Incorporate game genre, release dates, and pricing for more context-aware recommendations.

• Mobile Optimization: Adapt dashboard UI for mobile platforms to enhance accessibility.



Group 11

GameSphere: Smart Gaming Analytics & Recommendations

Leveraging Big Data and Machine Learning to Enhance Game Discovery and User Engagement