GROUP 4

Steam Dataset 2025 : Multi-modal gaming Analytics

1. **Problem Statement** :

Player engagement and review behavior on Steam are influenced by many factors, but the relationships between game attributes and user sentiment are not clearly understood. This lack of insight makes it difficult to determine what truly drives game visibility, traction, and positive community response.

2. **Dataset Link** :

https://www.kaggle.com/datasets/crainbramp/steam-dataset-2025-multi-modal-gaming-analytics

3. **Objectives** :

1. Convert Steam data into insights by processing reviews and metadata to understand game quality, user sentiment, and market trends.
2. Build a scalable big-data pipeline using cloud storage and Spark for efficient ingestion, cleaning, and processing of the dataset.
3. Analyse user sentiment to track positive/negative trends, detect common issues, and extract feature requests from reviews.
4. Generate business insights to evaluate game success, developer performance, genre trends, pricing effects, and risk signals.
5. Create interactive dashboards (Power BI) for visualizing sentiment trends, game comparisons, genre popularity, and publisher scorecards.
6. Develop ML models for sentiment classification, success prediction, and trend forecasting using review and game metadata.
7. Enable automation and scalability with scheduled ETL jobs and support for future enhancements like real-time reviews, recommendations, and advanced NLP.

4. **Data Dictionary :**

Table Name	Columns	dtype
application_categories	appid	int64
	category_id	int64
application_developers	appid	int64
	developer_id	int64
application_genres	appid	int64
	genre_id	int64
application_platforms	appid	int64

	platform_id	int64
application_publishers	appid	int64
	publisher_id	int64
applications	about_the_game	object
	achievement_count	float64
	achievements	object
	appid	int64
	background	object
	base_app_id	float64
	combined_text	object
	content_descriptors	object
	created_at	object
	currency	object
	description_embedding	object
	detailed_description	object
	discount_percent	float64
	embedding_run_id	int64
	fetched_at	float64
	final_price	float64
	header_image	object
	initial_price	float64
	is_free	bool
	linux_requirements	object
	mac_requirements	object
	mat_achievement_count	float64
	mat_currency	object
	mat_discount_percent	float64
	mat_final_price	float64
	mat_initial_price	float64
	mat_pc_graphics_min	object
	mat_pc_graphics_rec	object
	mat_pc_memory_min	object
	mat_pc_memory_rec	object
	mat_pc_os_min	object
	mat_pc_os_rec	object
	mat_pc_processor_min	object

	mat_pc_processor_rec	object
	mat_supports_linux	bool
	mat_supports_mac	bool
	mat_supports_windows	bool
	metacritic_score	float64
	movies	object
	name	object
	name_from_applist	object
	package_groups	object
	pc_requirements	object
	price_overview	object
	ratings	object
	recommendations_total	float64
	release_date	object
	required_age	int64
	screenshots	object
	short_description	object
	steam_appid	int64
	success	bool
	supported_languages	object
	supports_linux	bool
	supports_mac	bool
	supports_windows	bool
	type	object
	updated_at	object
categories	id	int64
	name	object
developers	id	int64
	name	object
embedding_runs	created_at	object
	dimension	int64
	model_name	object
	normalized	bool
	notes	float64
	run_id	int64

genres	id	int64
	name	object
platforms	id	int64
	name	object
publishers	id	int64
	name	object
reviews	appid	int64
	author_last_played	int64
	author_num_games_owned	int64
	author_num_reviews	int64
	author_playtime_at_review	float64
	author_playtime_forever	int64
	author_playtime_last_two_weeks	int64
	author_steamid	int64
	comment_count	int64
	created_at	object
	embedding_run_id	int64
	language	object
	received_for_free	bool
	recommendationid	int64
	review_embedding	object
	review_text	object
	steam_purchase	bool
	timestamp_created	int64
	timestamp_updated	int64
	updated_at	object
	voted_up	bool
	votes_funny	int64
	votes_up	int64
	weighted_vote_score	float64
	written_during_early_access	bool

5. Key Performance Indicators:

1. Data Volume KPIs
   239K+ Games collected
   1M+ User Reviews processed
   13+ Normalized Tables in warehouse

2. Data Quality KPIs
   99% Schema Completeness
   <1% Missing Fields across key tables
   Daily Update Success Rate

3. Processing KPIs
   API Ingestion Speed: X records/sec
   ETL Pipeline Latency: < Y mins/batch
   Embedding Generation Speed: N items/sec

4. Database Performance KPIs
   Query Latency: <100 ms for metadata
   Vector Search Speed: <50 ms per query
   Index Hit Rate: >90%

5. Analytics KPIs
   Genre Coverage: 50+ genres
   Platform Coverage: Windows / macOS / Linux
   Sentiment Accuracy: >85%

6. User Interaction KPIs
   Search Response Time: <200 ms
   Recommendation Accuracy: Top-10 similarity >0.75 score
   Review Sentiment Trend: Positive vs negative %

7. System Reliability KPIs
   Uptime: 99%
   Pipeline Failure Rate: <1%

6. Architecture Diagram :