Nombres: Rafael Cabrera, Laura Ortiz y Mariana Ramírez

Proyecto Final

1. Selección de la Problemática

Problemática

El objetivo de este proyecto es analizar la popularidad de diferentes videojuegos en tiempo real a través de los datos de publicaciones en Reddit. Específicamente, queremos identificar qué videojuegos están siendo mencionados con mayor frecuencia y visualizar esta información para comprender tendencias y patrones en la popularidad de los videojuegos.

Justificación

La industria de los videojuegos es una de las más dinámicas y de rápido crecimiento en el mundo del entretenimiento. Comprender qué juegos son más populares puede proporcionar información valiosa a desarrolladores, comercializadores y entusiastas de los videojuegos. Reddit es una plataforma relevante para estas discusiones, ya que alberga una gran comunidad de jugadores que comparten y comentan sobre sus experiencias y preferencias en tiempo real.

Fuente de Datos

Para este proyecto, utilizamos una base de datos que contiene aproximadamente 88.000 posts de Reddit y hacemos una simulación de una API desde nuestro servidor. Adicionalmente, utilizamos solo el contenido del post en nuestro modelo.

3. Metodología para Resolver la Problemática

Arquitectura General del Sistema:

1. **Extracción de Datos en Tiempo Real:** Utilizaremos nuestra API para obtener datos en tiempo real de publicaciones en reddits de videojuegos.
2. **Preprocesamiento de Datos:** Los datos serán preprocesados para limpiar texto, eliminar ruido y extraer los nombres de los videojuegos mencionados en cada publicación.
3. **Análisis de Datos:** Utilizaremos LLM para analizar los datos y generar modelos de popularidad.
4. **Visualización en AWS QuickSight:** Los resultados del análisis serán visualizados en PgAdmi al conectarnos en nuestro RDS postgres, para proporcionar una visión clara y comprensible de la popularidad de los videojuegos en tiempo real.

Pasos Específicos:

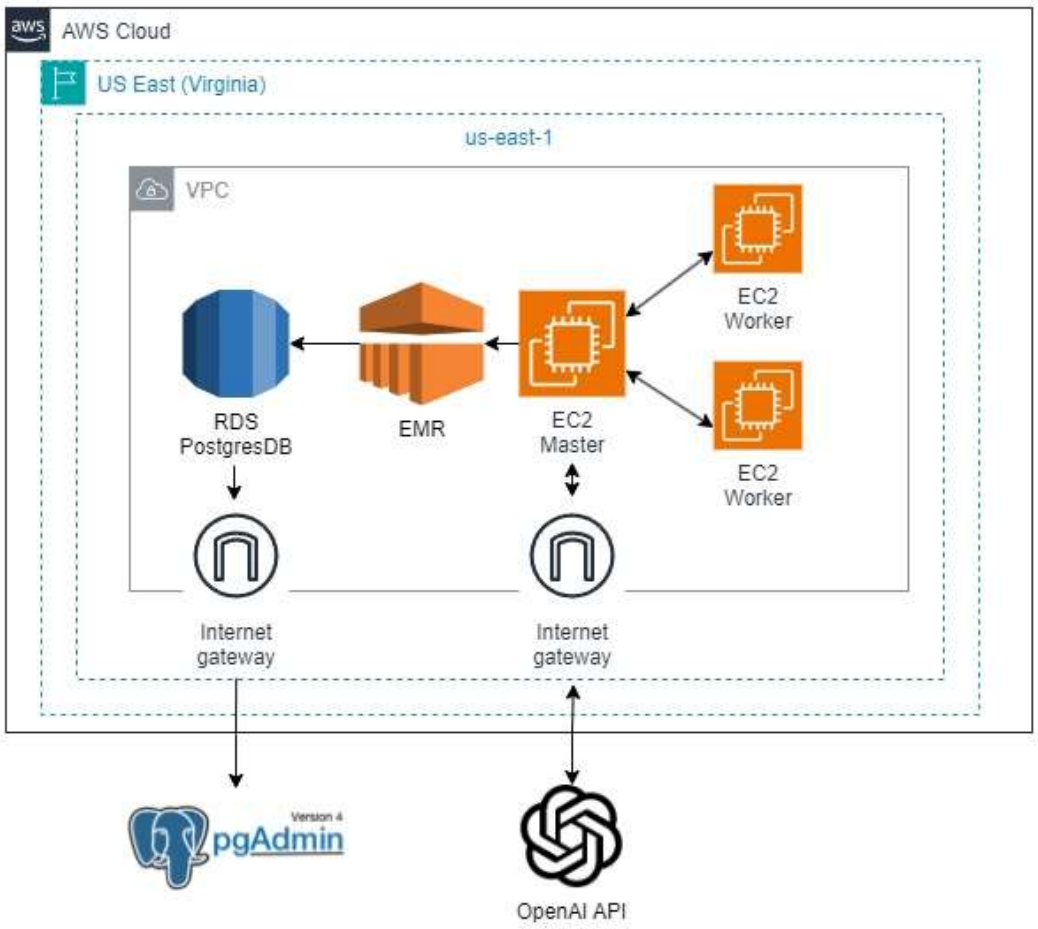
1. **Extracción de Datos:**
    - API de Reddit: Implementamos un script en Python para conectarse a la API de Reddit y extraer publicaciones de reddits relacionados con videojuegos.
    - Stream de Datos: Configurar un flujo continuo de datos para obtener publicaciones en tiempo real.

2. **Preprocesamiento de Datos**
    - Limpieza de Texto: Eliminar caracteres especiales, enlaces y cualquier otro ruido en el texto de las publicaciones.

3. **Análisis de los datos:**
    - Extracción de Nombres de Videojuegos: Hacer llamada a un LLM a través de un prompt para extraer los nombres de videojuegos mencionados.
    - Modelo de Popularidad: Construir un modelo para medir la popularidad de cada videojuego basado en la frecuencia y el contexto de las menciones.

4. **Visualización:**
    - Integración con AWS: Configurar la infraestructura en AWS para almacenar y procesar los datos, incluyendo el uso de Amazon RDS para ejecutar Spark.
    - Dashboards: Crear dashboards para visualizar la popularidad de los videojuegos en tiempo real. Incluimos gráfico de barra en pgAdmi.

Diagrama de la infraestructura de AWS:



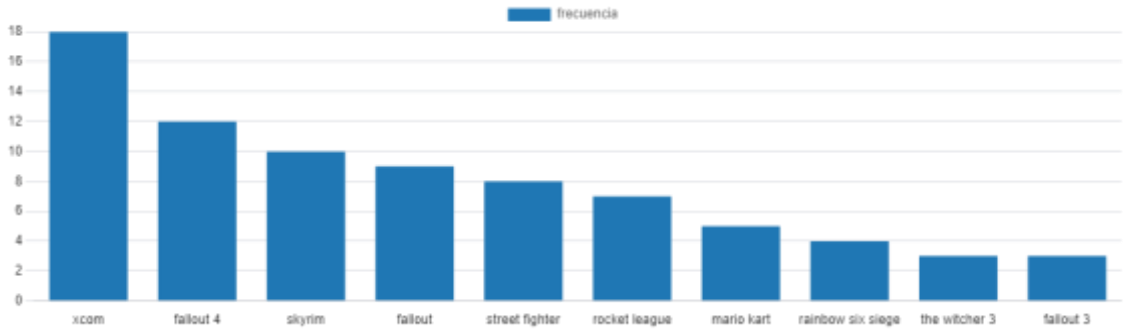
Resultados:
Aquí podemos ver el resultado del modelo, el cual saca los nombres de los videojuegos

mencionados. En este ejemplo podemos ver que hay 245 datos.

	id_juego_mencion [PK] integer	juego character varying (100)
268	743	runescape
269	744	mario party
270	745	counter strike
271	746	skyrim
272	747	far cry 4
273	748	project reality
274	749	rocket league
275	750	rocket league
276	751	wow
277	752	archer
278	753	street fighter
279	754	amnesia
280	755	hell yeah
281	756	xcom
282	757	xcom
283	758	street fighter 2

Total rows: 283 of 283 Query complete 00:00:00.912

Además, podemos ver la gráfica que muestra el top 10 de los videojuegos más mencionados.

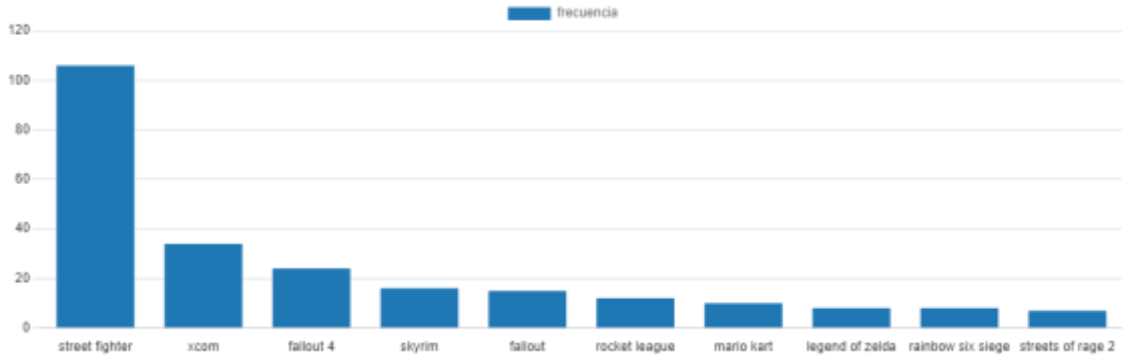


Ahora, veamos otro ejemplo:

Aquí podemos ver que tenemos 598 valores, después de cierto tiempo de recibir datos del RDS:

	id_juego_mencion [PK] integer	juego character varying (100)
584	1059	toad
585	1060	rainbow six siege
586	1061	street fighter
587	1062	street fighter v
588	1063	guilty gear strive
589	1064	super smash bros ultimate
590	1065	rainbow six siege
591	1066	gears of war
592	1067	rocket league
593	1068	streets of rage 2
594	1069	binary domain
595	1070	condemned criminal origins
596	1071	gunstar heroes
597	1072	renegade ops
598	1073	viking battle for asgard
Total rows: 598 of 598	Query complete 00:00:01.234	

Y el análisis de popularidad nos genera esto:



Conclusión:

Este proyecto aborda la problemática de analizar la popularidad de los videojuegos en tiempo real utilizando datos de Reddit. A través de una combinación de técnicas de análisis de datos con LLM y visualización, esperamos proporcionar una herramienta eficaz para entender las tendencias en la industria de los videojuegos. La metodología descrita asegura un enfoque estructurado y detallado para alcanzar los objetivos propuestos.