UNIVERSIDAD NACIONAL DE ENTRE RÍOS

FACULTAD DE INGENIERÍA

CARRERA: Tecnicatura Universitaria en Procesamiento y Explotación de Datos

MATERIA: Base de Datos Multidimensionales

Nombre de la Actividad: Trabajo Integrador Final

Título: Data Warehouse para Epic Games: Transformando Datos en Acciones Estratégicas

Fecha de Entrega: 2/11/2023

Profesores: Walter Elias - Profesor Adjunto

Maximiliano Fernandez - JTP

Alumno:

Ruiz Diaz, Enzo

Data Warehouse para Epic Games: Transformando Datos en Acciones Estratégicas

Introducción

La empresa Epic Game tiene como producto una tienda de videojuegos que cada vez crece más en el mercado. Y para competir se centra en ofertas y promociones especiales para así atraer más clientes y que sea una plataforma más interesante para los desarrolladores y publicadores.

Situación problemática

En el dinámico mundo de los videojuegos, Epic Games ha emergido como un competidor serio en la industria con su plataforma de distribución digital. La creciente presencia de Epic Games en el mercado también ha captado la atención de desarrolladores y publicadores.
Para mantenerse a la vanguardia y seguir siendo competitivos se considera clave entender las preferencias de los jugadores y la crítica especializada. En este contexto surge la necesidad de analizar los datos de los juegos disponibles en la tienda, para así obtener insights valiosos para la toma de decisiones sobre estrategias futuras.

Planteamiento del problema

En este contexto, surge la siguiente pregunta de investigación: **¿Cuáles son los patrones de diseño para un data warehouse, que permita obtener los insights solicitados por Epic Games sobre los datos de su plataforma de distribución digital durante los años 2011-2022?.**

Este problema se presenta como un desafío fundamental para Epic Games, ya que requiere una profunda inmersión en los datos disponibles sobre los juegos en su plataforma. Analizar el rating promedio por género, publicador y desarrollador proporcionará una visión detallada de las áreas de fuerza y oportunidad, permitiendo a Epic Games adaptar su enfoque para satisfacer las expectativas de los jugadores y las demandas del mercado.

Objetivos generales

Obtener métricas básicas sobre el rating impuesto en cada juego, las cuales permitan a partir de ellas obtener diversas métricas derivadas, por género, desarrollador, publicador y empresa de crítica.Para que el sector pertinente pueda tomar decisiones basadas en datos.

Marco teórico

En el mundo dinámico de los videojuegos, Epic Games se enfrenta al desafío de mantenerse a la vanguardia de la industria. Para tomar decisiones estratégicas fundamentadas, es esencial comprender las preferencias y las tendencias del mercado. En este contexto, exploramos conceptos clave como Data Warehouse, Base de Datos Multidimensional, el modelo Kimball y tecnologías OLAP. Estas herramientas proporcionan el marco necesario para analizar datos históricos y actuales desde diversas perspectivas, permitiendo una comprensión profunda del mercado y orientando las estrategias futuras de Epic Games en su plataforma de distribución digital.

Estado del arte

Los data warehouse han evolucionado y experimentado diversos cambios desde su aparición.Esto sugiere que también han evolucionado en la industria de los videojuegos, adaptándose a las necesidades y desafíos específicos de este sector. Por ejemplo, para estrategias de Marketing y Publicidad, las empresas de videojuegos utilizan análisis de Data Warehouse para evaluar el impacto de las campañas de marketing y publicidad en las ventas y en el atractivo del juego. Esto permite ajustar las estrategias de marketing para llegar a audiencias más específicas y efectivas. También, en el análisis de sentimiento y opiniones en las redes sociales y los foros de juegos, que son fuentes ricas de datos para entender el sentimiento de los jugadores hacia los juegos. Herramientas de análisis de texto en Data Warehouses permiten analizar estas opiniones para entender mejor las críticas y percepciones de los jugadores sobre un juego específico o una franquicia.

Data Warehouse

Un data warehouse es una colección de datos orientado a temas, integrada, variante en el tiempo y no volátil, que respalda el proceso de toma de decisiones de una organización. Es

el núcleo de la inteligencia empresarial y está optimizado para el análisis y la presentación de datos.

Vamos a desglosar esta definición:

1. Orientado a temas: Los datos en un data warehouse se organizan en torno a temas o áreas específicas de interés para la organización. Por ejemplo, ventas, inventario, recursos humanos, etc.
2. Integrada: La integración se refiere a la consolidación de datos de múltiples fuentes en un solo repositorio. Los datos de diferentes sistemas y departamentos se combinan y se almacenan de manera coherente en el data warehouse.
3. Variante en el tiempo: Los datos en un data warehouse incluyen información histórica y están diseñados para rastrear los cambios en los datos a lo largo del tiempo. Esto permite el análisis de tendencias y patrones a lo largo de períodos temporales.
4. No volátil: Una vez que los datos se cargan en el data warehouse, se vuelven no volátiles, es decir, no se modifican. Esto garantiza la consistencia e integridad de los datos históricos para análisis retrospectivos.

Las fases de diseño de un data warehouse generalmente se dividen en tres fases principales:

1. Diseño conceptual: Consiste en identificar y describir los datos relevantes para el negocio, utilizando un modelo de alto nivel, como el modelo entidad/interrelación. En esta fase se deben definir las entidades, los atributos, las relaciones y las restricciones. El resultado es un esquema conceptual que refleja las necesidades de información de los usuarios.
2. Diseño lógico: Consiste en transformar el esquema conceptual en un esquema lógico, utilizando un modelo de datos específico, como el modelo relacional. En esta fase se deben definir las tablas, las columnas y las claves. También, normalizar las tablas para reducir redundancias y mejorar la integridad de los datos o, en ciertos casos, se puede desnormalizar para mejorar el rendimiento de las consultas. El resultado es un esquema de estrella, copo de nieve u otros según las necesidades de análisis.
3. Diseño físico:Se implementa el diseño lógico en una base de datos real,teniendo en cuenta los aspectos de rendimiento, seguridad y mantenimiento del data warehouse. También se ponen en implementación los procesos ETL para cargar datos desde diversas fuentes.

Dadas las necesidades de la empresa la cual necesita tomar decisiones basadas en datos se optó por diseñar un data warehouse, justificando con la teoría mencionada anteriormente

y agregando que lo que interesa es procesar, sobre un histórico que va de 2011 a 2022, la popularidad de distintos segmentos por lo que cumple con ser orientado a temas, integrada, variante en el tiempo y no volátil.

Base de datos multidimensional

Una base de datos multidimensional, también conocida como OLAP (Procesamiento Analítico en Línea), es una estructura de datos que permite el análisis desde múltiples dimensiones. Los datos se organizan en tablas de hechos y dimensiones, facilitando el análisis complejo desde diversas perspectivas. Los puntos clave de este concepto son:

- Organización en Dimensiones y Hechos: Los datos se dividen en dimensiones (como tiempo, producto y ubicación) y hechos (como ventas, ingresos). Esto facilita la comprensión de las relaciones entre distintos aspectos del negocio.
- Análisis Multidimensional: La estructura multidimensional permite el análisis desde varias perspectivas. Esto es crucial para entender patrones complejos y tomar decisiones informadas basadas en diferentes combinaciones de dimensiones.

Al unir estos conceptos, nuestro proyecto se beneficiará de la robustez y versatilidad de un data warehouse optimizado y una base de datos multidimensional, lo que nos permitirá analizar datos históricos y actuales desde diversas perspectivas, proporcionando así una comprensión profunda del mercado. La industria de los videojuegos se beneficia especialmente de las bases de datos multidimensionales, ya que permiten analizar el rendimiento de los juegos desde múltiples dimensiones como género, publicador y desarrollador. Esto proporciona información vital para la toma de decisiones estratégicas.

Modelo Kimball

El modelo de datos de Kimball, iniciado por Ralph Kimball, sigue un enfoque de abajo hacia arriba para el diseño de arquitectura de data warehouse en el que los data marts se forman primero en función de los requisitos comerciales. Este modelo se basa en el concepto de dimensiones y hechos, que son las dos principales estructuras de datos que se utilizan para almacenar y analizar la información. Las dimensiones son los atributos que describen las características de los datos, como el tiempo, el lugar, el producto, el cliente, etc. Los hechos son las medidas numéricas que representan el rendimiento del negocio, como las ventas, los costos, las ganancias, etc.

El modelo Kimball propone organizar las dimensiones y los hechos en esquemas de estrella o de copo de nieve, que son formas de representar los datos. Es útil para el análisis de datos porque permite a los usuarios acceder rápidamente a la información relevante y tomar decisiones informadas. Además, el modelo es fácil de mantener y actualizar, lo que lo hace ideal para empresas que necesitan analizar grandes cantidades de datos de manera constante.

La metodología Kimball también destaca por su enfoque colaborativo, ya que involucra a diferentes áreas de la empresa en el proceso de gestión de datos. De esta forma, se asegura una mayor comprensión de las necesidades y objetivos de cada área y se optimiza la toma de decisiones.Entre los beneficios de la metodología Kimball se encuentran una mejora en la calidad de los datos, una mayor eficiencia en la gestión de los mismos y una mayor capacidad para analizarlos y obtener insights relevantes para la empresa.

El modelo Kimball recomienda seguir un ciclo de vida para el desarrollo de data warehouse, que consiste en varias fases: planificación, análisis de requisitos, diseño, construcción, implementación y mantenimiento.

Teniendo como apoyo lo explicado anteriormente se decidió utilizar kimball y copo de nieve para el diseño del data warehouse. Se decanto por kimball principalmente por ser fácil de mantener y de actualizar ya que la naturaleza de los datos vienen de las diferentes críticas que un juego de la tienda de epic obtuvo en un día a lo largo de un año. Lo cual lleva a tener una gran cantidad de datos y es seguro que se agregaran más juegos. Por otro lado se optó por diagrama copo de nieve ya que se necesita un nivel de detalle en el análisis sobre los desarrolladores, publicadores, género y compañía de crítica.

OLAP (Procesamiento Analítico en Línea)

OLAP se refiere a un tipo de tecnología utilizada para analizar datos multidimensionales de manera interactiva. A diferencia de las bases de datos relacionales tradicionales que están diseñadas para manejar datos tabulares (bidimensionales), OLAP permite a los usuarios analizar datos desde múltiples dimensiones.

En OLAP, los datos se organizan en cubos multidimensionales, donde cada dimensión representa una característica o atributo específico de los datos (como tiempo, producto, ubicación, etc.). Los cubos OLAP permiten a los usuarios realizar análisis complejos y detallados, ya que pueden ver los datos desde diferentes perspectivas y realizar operaciones como cortar (slicing), cortar y seleccionar (dicing), subir (roll-up) y bajar (drill-down) para explorar los datos.

Para implementar sistemas OLAP existen varios enfoque pero nos centraremos en tres:

- MOLAP: Usa una base de datos multidimensional en la que la información se almacena multidimensionalmente, utiliza estructuras de índices especiales para permitir un rápido acceso a los datos multidimensionales.Es ideal para conjuntos de datos de tamaño moderado y proporciona respuestas interactivas para análisis en tiempo real.
- ROLAP: almacena los datos de un cubo OLAP en una base de datos relacional, como PostgreSQL, Oracle o SQL Server. En lugar de utilizar estructuras multidimensionales precalculadas, ROLAP genera consultas SQL complejas para recuperar datos multidimensionales en tiempo real.Entre sus virtudes encontramos que puede manejar grandes volúmenes de datos y es escalable,es flexible y puede adaptarse a modelos de datos complejos y permite el análisis detallado y la exploración ad hoc de datos.
- HOLAP: es una combinación de ROLAP y MOLAP almacena algunos datos en un formato multidimensional y otros en una base de datos relacional, esto permite a los usuarios elegir qué datos se almacenan de forma multidimensional o relacional proporcionando una combinación de rendimiento y flexibilidad.Con el costo de requerir una minuciosa configuración y complejidad a la hora de implementar.

Para el proyecto se eligió ROLAP debido a que puede manejar grandes volúmenes de datos y es escalable lo cual se ajusta de excelente manera sobre necesidades de nuestro proyecto. También nos permite el análisis detallado y la exploración ad hoc de datos lo cual quiere decir que permite investigar patrones, tendencias o relaciones en los datos que no han sido planificados de antemano, explorar los datos de manera flexible y creativa, utilizando herramientas analíticas para descubrir información relevante y obtener insights inmediatos.

Desarrollo

Diseño

En primer lugar, se analizó las fuentes de datos que van conformar el data warehouse. Estas se obtuvieron de Kaggle que es una plataforma de competencia de ciencia de datos y una comunidad en línea de científicos de datos y profesionales del aprendizaje automático de Google. La base original que se puede ver en la imagen 1 está compuesta por 6 archivos csv(games,necessary hardware,open critic .csv,social_networks.csv,tweets,twitter accounts) los cuales se originan de una base de datos relacional que ver en la imagen 1.

Gráfico 1
Diagrama de tablas base de datos relacional

Nota: La imagen se encuentra adjunta en los archivos del trabajo práctico como tabla_epci_games_original.png

Como partimos de la premisa de construir un data warehouse que sirva de base para llevar a cabo diferentes análisis sobre la popularidad de los juegos disponibles en la tienda de epic games se tomó la decisión de no incluir la información sobre el hardware necesario.

También, se tuvo que descartar los csv de social network, tweets y twitter accounts debido a que los ID de Twitter son todos 0. Lo mismo ocurre con la clave externa para los tweets. Por lo que no es posible vincular los tweets a la cuenta de twitter correspondiente.

Como se quiere obtener más detalle sobre la popularidad de los desarrolladores, publicadores y géneros se optó por normalizar la tabla games así mismo como interesa también la compañía a la cual el crítico pertenece se normaliza la tabla critic.

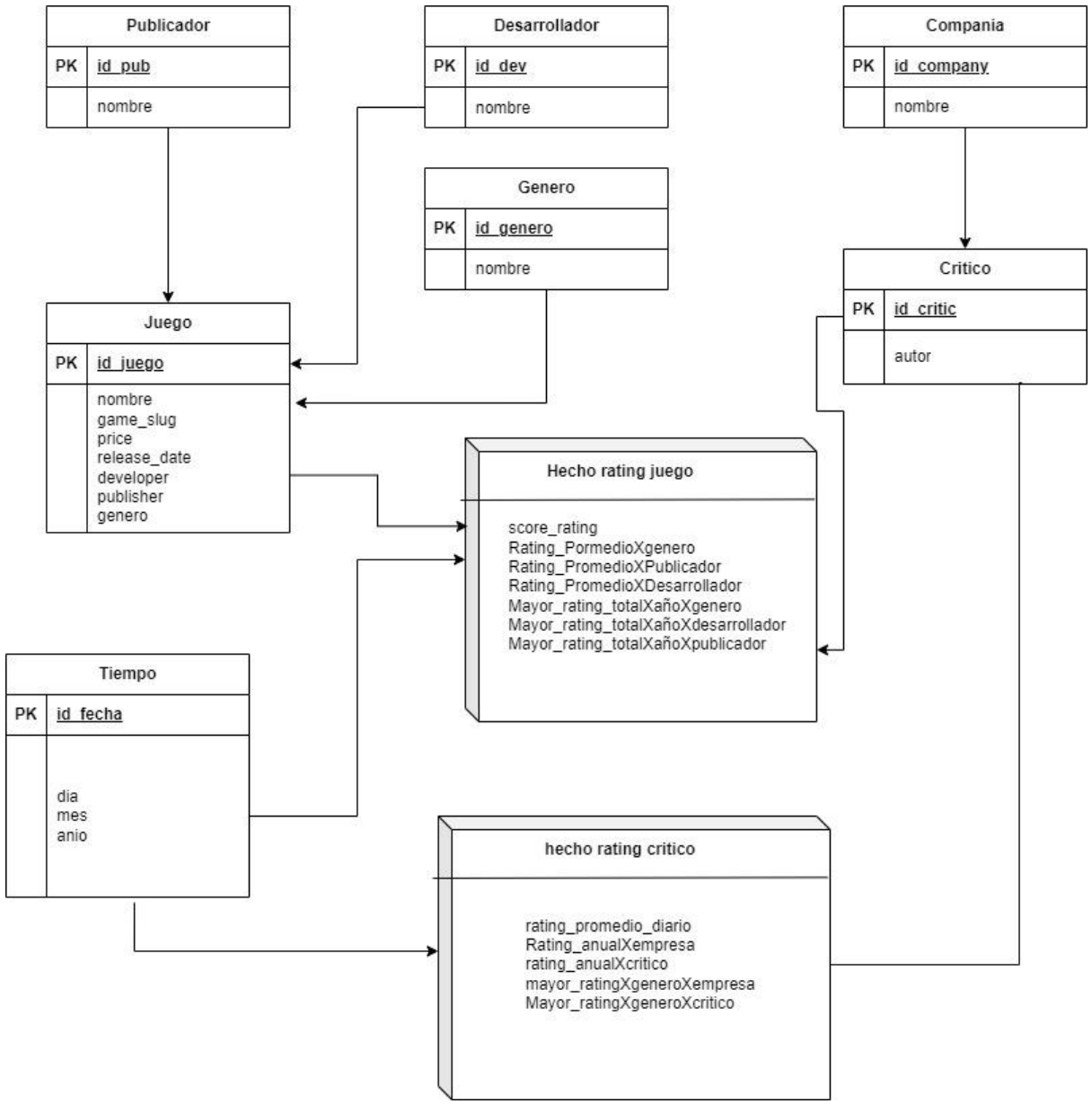
Dado lo anteriormente mencionado, las dimensiones serán las siguientes:

- Dimensión juego,representa juegos individuales, se normaliza en:
  - dimensión género, categoriza los juegos por su género
  - dimensión publicador, organiza los juegos por la empresa que los publica
  - dimensión desarrollador, organiza los juegos por la empresa que los desarrolla
- Dimensión crítico,representa críticos individuales , se normaliza en:
  - dimensión compañía, organiza los críticos por la compañia a la cual pertenecen
- Dimensión tiempo
- Hecho rating juego
  - Métrica: score rating para juegos individuales
- Hecho rating crítico
  - Métrica: rating promedio diario de críticos o compañías de críticos.

El diagrama multidimensional quedó de la siguiente manera.

Gráfico 2

Diagrama multidimensional



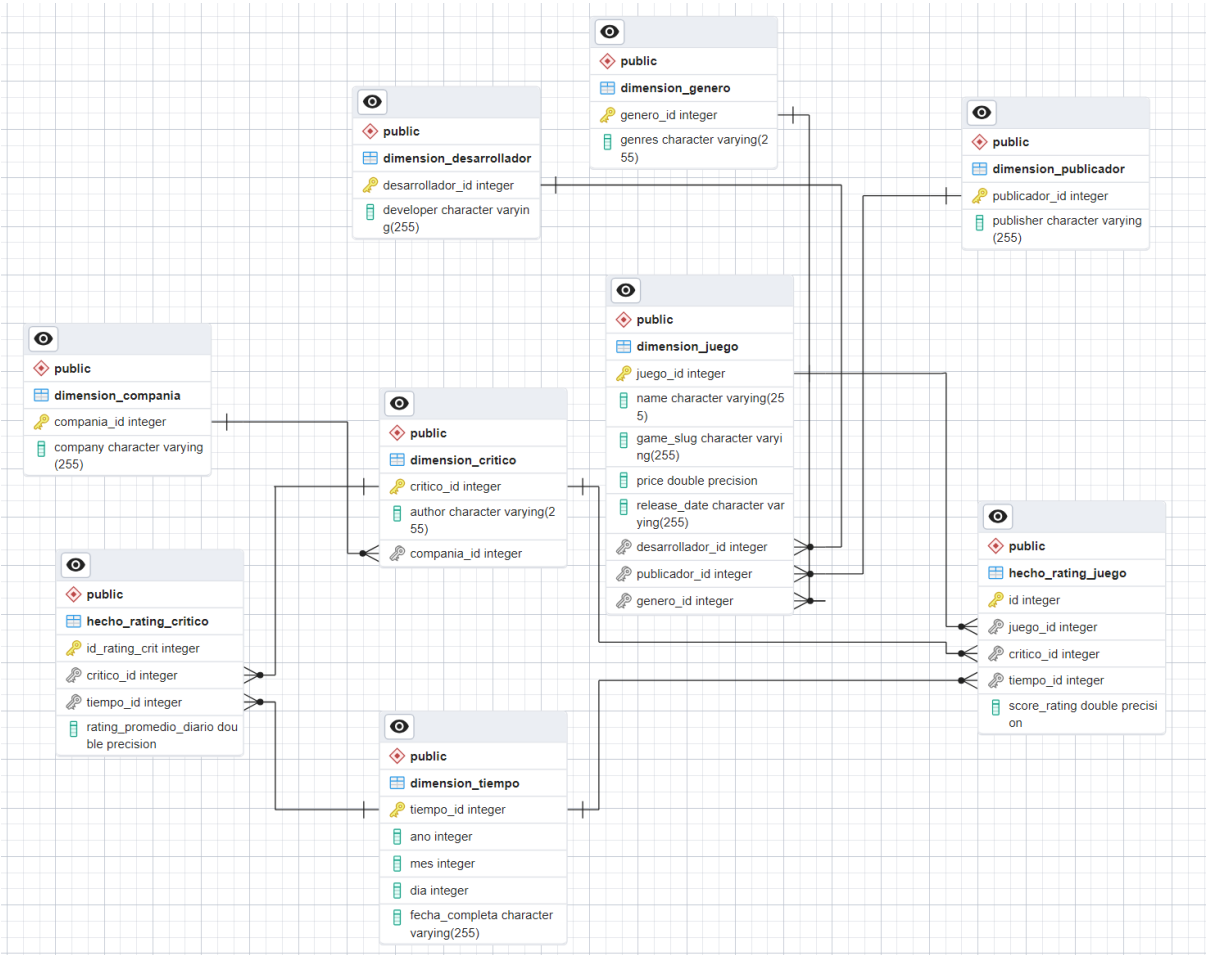
Nota: La imagen se encuentra adjunta en los archivos del trabajo práctico como diagrama_multidimensional.jpg.

En el diagrama podemos ver que hay dos tablas de hechos, la primera hecho rating juego tiene la métrica básica de score rating la cual es la suma de los rating que un juego tiene en un día sin importar el crítico. Esta medida permite conocer diferentes aristas sobre la popularidad del juego pudiendo segmentar en género, publicador o desarrollador. A su vez podemos ver, por ejemplo, para el año 2022 el género con mayor popularidad dentro de la crítica especializada.

La segunda tabla de hecho rating crítico, tiene una mirada similar pero centrada en los críticos y compañías. Su métrica básica es el rating promedio diario donde se suma en un día todo los rating que dio un crítico determinado el cual pertenece a una empresa determinada y luego se promedia. Esto permite ver por ejemplo qué empresas dan mejores rating por año o al revés qué empresas están dando peores rating.

Gráfico3

Diagrama de tablas del Data Warehouse



Nota: La imagen se encuentra adjunta en los archivos del trabajo práctico como diagrama_de_tablas.png

Implementación

Teniendo los diagramas con dimensiones y hechos definidos podemos proceder a codificar el DDL, como anteriormente mencionamos utilizaremos ROLAP y elegimos como SGBD postgresql (DDL). Una vez creada la base de datos en postgre podemos seguir con el proceso de ETL.

Para realizar el ETL, se utilizó el lenguaje de programación python con las librerías pandas, para manipular los archivos en formato data frame y sqlalchemy, una librería de ORM para

establecer conección a la base de datos y poder subir los datos a la misma. Todo se volcó en una notebook de jupyter.

ETL

- **Extracción de datos**: se cargó a python mediante pandas, y se analizó las columnas de los dos csv disponibles (games y open critic). Se decide quedar con las columnas relevantes para la base de datos. En el caso de games se descarta plataform (90% windows)y descrition(Largas cadenas de string) para quedarnos solo con 'id', 'name', 'game_slug', 'price', 'release_date', 'developer', 'publisher', 'genresy en el caso de open critic solo se descarta la columna top_critic (booleano) por falta de información sobre la misma.
- **Transformación:** Como solo interesan los juegos que tienen críticas y coinciden con los registros de data frame que contiene los juegos ,se realiza un merge (how='inner') para quedarnos con aquellos juegos que tengan críticas. También, se eliminan las filas que contengan celdas vacías en las columnas 'developer','rating','name','date','author','company', 'publisher' y 'genres'. Ya que son esenciales, quedando así 11 filas, no se cuentan los id que proporciona el dataset, y 12125 filas. Siendo cada fila una crítica a un juego.
- **Carga:** antes de realizar la carga dependiendo la dimensión se crea un nuevo data frame con un mapeo de los id de las dimensiones que necesite relacionarse, se puede ver más detallado en el código. Luego se utilizó una función para cargar los datos si estos no existen ya en la base de datos.

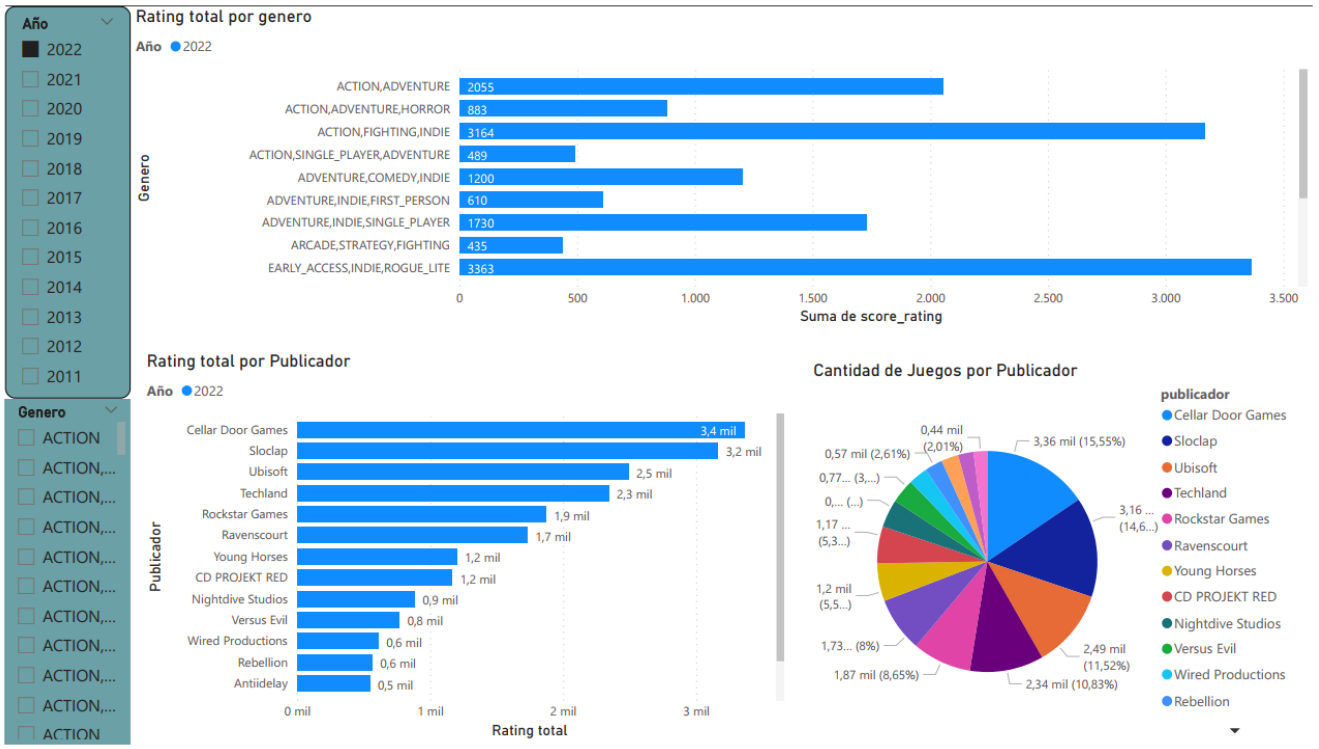
Se adjunta el archivo etl.ipynb el cual contiene el código correspondiente y su documentación.

Reporte

Para el reporte se utilizó el software Power BI, ya que permite diversos filtros dinámicos, calcular métricas derivadas de la métrica básica de forma sencilla y de ser necesario se puede codificar mediante el lenguaje DAX métricas personalizadas.

Gráfico 4

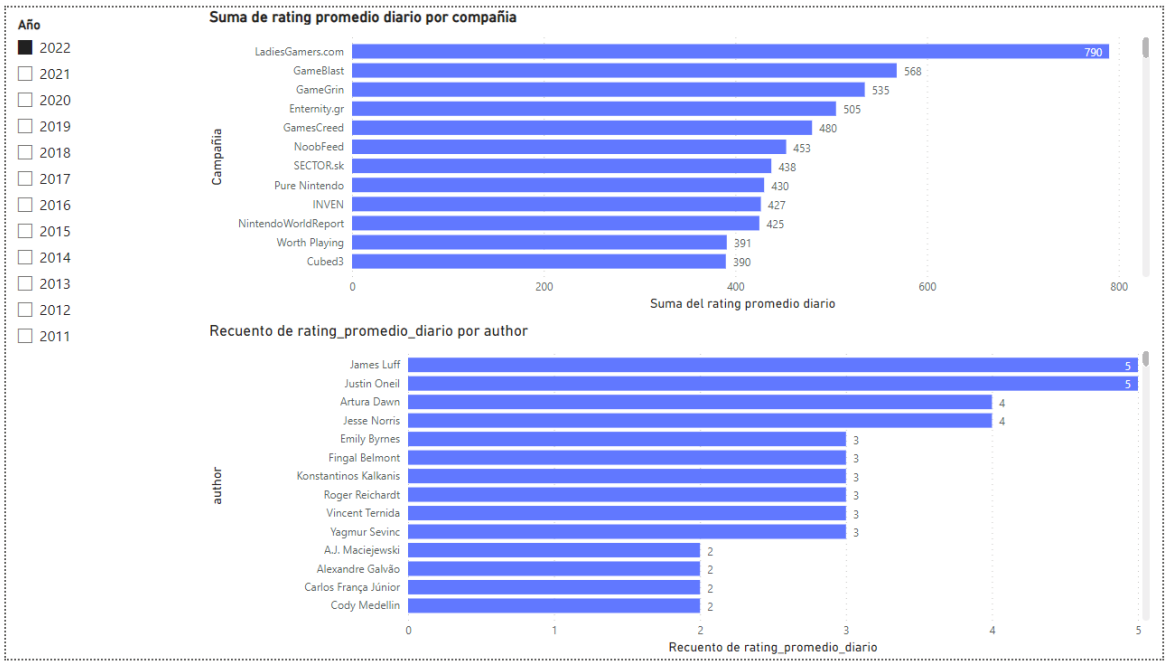
Reporte de popularidad por género y publicador



En el gráfico 4 podemos ver cuál fue el género más popular, cuál fue el publicador más popular y la cantidad de publicaciones. Esto lo podemos segmentar con un filtro interactivo por año, y para los gráficos de abajo referidos a los publicadores también lo podemos segmentar por género.

Gráfico 5

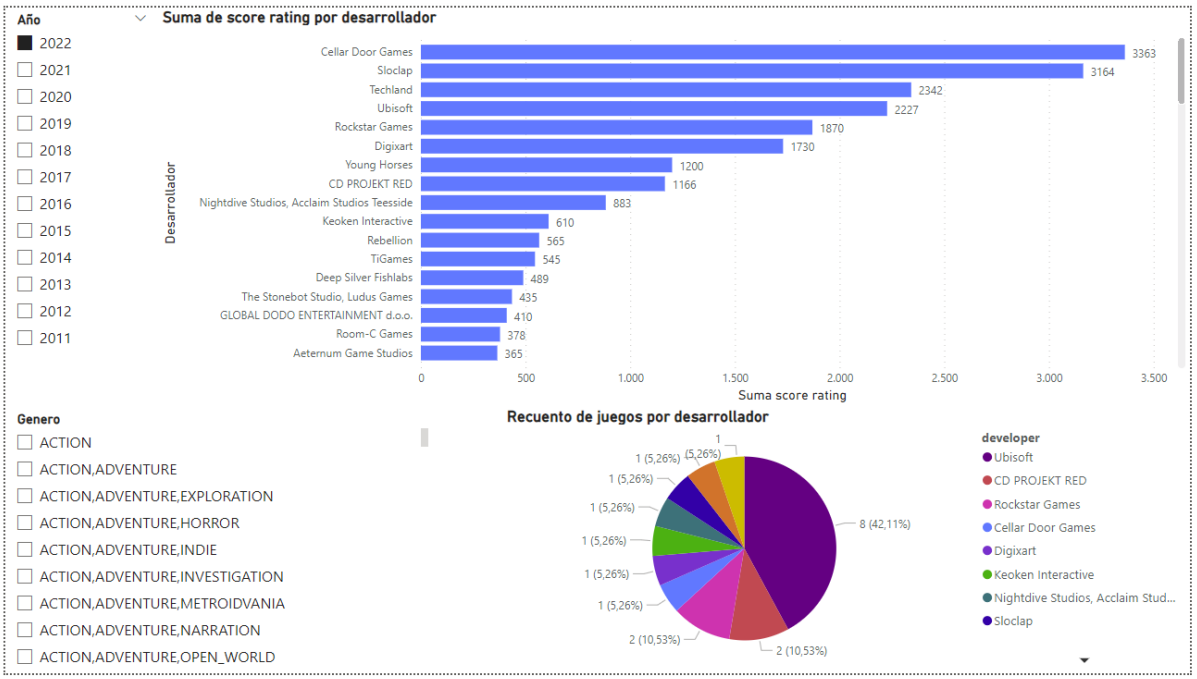
Reporte de empresas que tiene un rating promedio mayor por año y críticos con más críticas



En el gráfico 5 podemos observar para cada año cuales son los críticos con más críticas promedio diarias hechas y que empresas tiene un rating acumulado anual mayor .

Gráfico 6

Reporte de popularidad segmentado por desarrollador



En el gráfico 6 se observan los desarrolladores más populares por año y/o género. También, el recuento de juegos por desarrollador.

Plan de mantenimiento

Dada la naturaleza de los datos y las necesidades del negocio se propone un plan de mantenimiento mensual.

- **Cómo:** A través del ETL, se recalcula la vista
- **Cuándo:** Mensualmente a mes cerrado es decir durante los primeros 10 días del mes se actualizará al mes anterior al corriente.
- **Contexto:** Como no es el warehouse principal de la empresa y esta puede seguir en funcionamiento, el mantenimiento se realizará fuera de línea.

Conclusión

Después del desarrollo llevado adelante en el proyecto, queda aún más claro la importancia de tener un data warehouse ya que son una herramienta fundamental para el análisis de datos y la toma de decisiones basada en datos. No solo integran las fuentes de información sino que también al desarrollarlo basado en el modelo kimball obtenemos una solución más flexible que se adapta a las necesidades del negocio. Otra ventaja es la integración con diversas herramientas que facilitan y agilizan mucho el trabajo como es Power BI para reportes.

A nivel personal, me hubiera gustado poder integrar las tablas de redes sociales y tweets que por problemas en el id no se pudieron utilizar. También quiero resaltar que ahora comprendo mucho más la importancia de tener un warehouse y como con métricas básicas, a través de la integración con diversas herramientas, se puede obtener de manera sencilla información clave para una empresa que sin duda es un activo importante de cara a la estrategias comerciales.

Bibliografía

- Naeem, T. (2023, 27 de septiembre). Conceptos de Data Warehouse: enfoque de Kimball vs. Inmon. Astera. https://www.astera.com/es/tipo/blog/conceptos-de-almacén-de-datos/
- Fatima,Nida. (2023, 25 de octubre). ¿Qué es la arquitectura del almacén de datos?.Astera. https://www.astera.com/es/knowledge-center/data-warehouse-architecture/
- Gonzales, L. (2021, 6 de julio). La Metodología Kimball para Data Warehouses y BI exitosos - Explodat. Data Analytics. https://explodat.cl/Analytics/business-intelligence/la-metodologia-kimball-para-data-warehouses-y-bi-exitosos/
- Elias Walter. (2023). Diapositivas de clases. Bases de datos multidimensionales