Netflix Data Warehouse

Susan Hidalgo, Brenda Rojas, Cristopher Rodriguez, Melani Vargas

shidalgo70985@ufide.ac.cr
brojas60567@ufide.ac.cr

crodriguez00279@ufide.ac.cr

mvargas30424@ufide.ac.cr

Universidad Fidélitas, Ingeniería en Sistemas Computacionales, Heredia, Costa Rica

Data Warehouse

(SC-602)

Grupo #1

Resumen– Se quiere realizar una investigación a profundidad de la empresa Netflix para poder llegar a crear un Data Warehouse que logre satisfacer a la empresa y sus necesidades, tomando en cuenta datos proporcionados por el sitio Kaggle, donde se obtuvo información de los títulos que actualmente presenta la aplicaciones e información como sus géneros, fechas de lanzamiento, países entre otros.

Palabras Claves– Data Warehouse, Netflix, Kaggle, Streaming

Abstract– We want to carry out an in-depth investigation of the Netflix company in order to create a Data Warehouse that manages to satisfy the company and its needs, taking into account data provided by the Kaggle site, where information was obtained on the titles that the applications currently present and information such as their genres, release dates, countries among others.

Keywords– Data Warehouse, Netflix, Kaggle, Streaming

I. CAPÍTULO DE INTRODUCCIÓN

➢ INTRODUCCIÓN

Para nuestro proyecto de investigación del curso Data Warehouse hemos decidido basarnos en la famosa aplicación Netflix, la cual se basa en presentarle al comprador de la mensualidad de la aplicación una amplia gama de películas, series y shows televisivos donde se pueden descargar para ver sin internet, ver las películas o series en múltiples idiomas, entre otras muchísimas ventajas que nos ofrece el sitio.

En sus inicios Netflix era un centro de alquiler y venta de DVDs, pero sus fundadores Reed Hastings y Marc Randolph vieron un gran potencial en la venta de peliculas en linea, su nombre se debe a la unión de las palabras “internet” y “flicks” lo cual significa peliculas

Con el pasar de los años Netflix ha ganado un puesto en el mercado muy importante siendo prácticamente la aplicación de películas y series número 1 a nivel mundial, esto permitiéndoles no solo ser una plataforma de streaming sino que también les dio la oportunidad de empezar a grabar su propio contenido incluyendo series y películas, esto se debe también gracias a la pandemia ya que en el 2020 su uso se disparó de una manera estratosférica, actualmente cuenta con una base de datos muy amplia con diferentes países, fechas de lanzamientos, géneros de películas y series, lo que nos permite poder desarrollar diferentes gráficas, esquemas e investigaciones para poder profundizar más en una funcionalidad que logre satisfacer a la empresa.

➢ OBJETIVOS

Objetivo general:

Desarrollar un Data Warehouse funcional que satisfaga una necesidad empresarial específica.

Objetivos específicos:

- Implementar un proceso de extracción, transformación y carga de datos del dataset de Netflix disponible en Kaggle y otras fuentes relevantes.
- Diseñar un modelo multidimensional en el Data Warehouse que organice la información de manera específica.
- Integrar conexiones y herramientas de visualización de datos que faciliten un análisis basado en la toma de decisiones estratégicas.

➢ ANTECEDENTES

Al pasar los años la tecnología ha sido parte de cada innovación que se ha presentado en el mundo, esto también fue visto en el alquiler o renta de películas, ya que la fundación de Netflix se remonta a que uno de sus fundadores se atrasó en devolver una película a la tienda de videos y se le cobró una multa, debido a esto él decidió crear un sistema nuevo donde se pagará por mes y no existieran esos cobros o multas por atrasos de devolución, al pasar los años la venta de peliculas en linea a sido todo un

éxito permitiendo que la empresa de Netflix creciera y se convirtiera en lo que hoy en día conocemos, tantos años de trayectoria involucra muchísimos datos e información la cual podemos usar para dar el mejor funcionamiento y una experiencia de la más alta calidad a todos los usuarios.

➢ JUSTIFICACIÓN

El dataset de *Netflix Movies and TV shows* que se encuentra disponible en la página Kaggle, proporciona valiosa información acerca de los títulos que actualmente son más vistos en la plataforma, incluyendo detalles como género, país, fechas de lanzamiento, entre otros. Dado al gran volúmen e importancia de estos datos para la toma de decisiones estratégicas en la industria del entretenimiento, se plantea el desarrollo de un Data Warehouse funcional para el análisis y optimización del catálogo de contenido. A su vez, la solución de este proyecto facilitará la integración de diversas fuentes de datos y ayudará con la identificación de tendencias enfocadas en el consumo, popularidad de géneros y análisis de mercado, además la creación de un Data Warehouse optimizará la organización y acceso de datos, mejorando la eficiencia en la toma de decisiones, como la planificación de contenido basado en predicciones de futuras demandas.

REQUERIMIENTOS FUNCIONALES

- Fuentes de datos: Se utilizarán cinco datasets relacionados con Netflix, disponibles en la plataforma Kaggle. Estos incluyen datos sobre películas y series, reseñas de usuarios, comportamiento bursátil, y la base de usuarios de Netflix. Los datos estarán en formato CSV y serán transformados mediante Pentaho antes de ser almacenados en el Data Warehouse.
- *Componentes ETL:* El proceso ETL será gestionado por Pentaho, el cual se encargará de extraer los datos desde los archivos CSV, transformarlos y cargarlos en el SQL Server.

CARACTERÍSTICAS DE LA SOLUCIÓN

- Pentaho: Una herramienta que se utilizará para llevar a cabo los procesos ETL, es decir, extracción, transformación y carga. Se encargará de integrar los datasets encontrados,, limpiando y transformando los datos según sea necesario antes de almacenarlos en el Data Warehouse.
- SQL Server: Será el sistema de gestión de bases de datos donde se implementará el Data Warehouse. Aquí se crearán las tablas de hechos y dimensiones, organizadas bajo un modelo multidimensional como el esquema estrella, lo que permitirá almacenar y organizar la información de forma eficiente.
- Power BI: Se utilizará para generar reportes y visualizaciones a partir de los datos almacenados en el Data Warehouse, de modo que se pueda analizar el rendimiento de los contenidos de Netflix, identificar tendencias y realizar seguimientos de datos financieros.

II. DIAGRAMAS DE ENTIDAD-RELACIÓN DEL MODELADO DIMENSIONAL

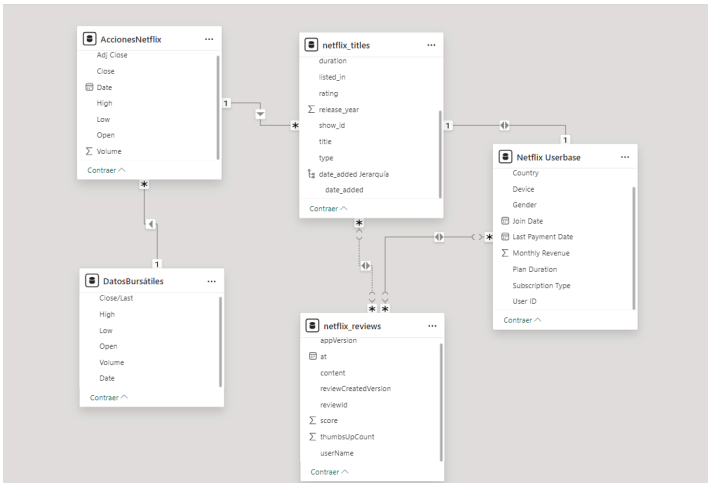

Fig 1 Modelado de Base de Datos de PowerBI

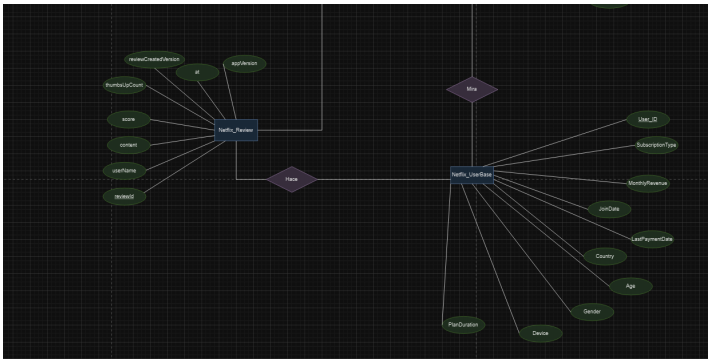

Fig 2 Diagrama de Entidad relación draw.io

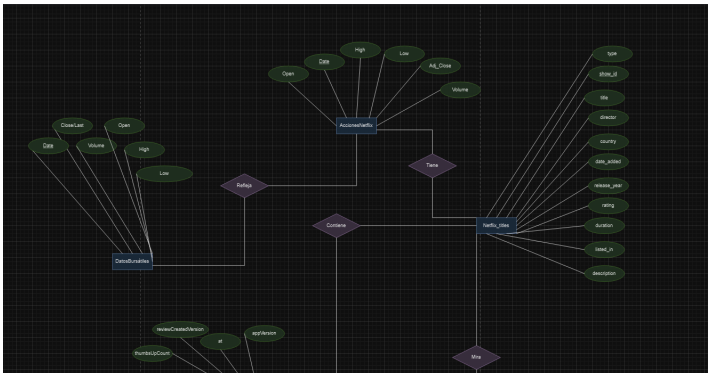

Fig 3 Diagrama de Entidad relación draw.io
https://drive.google.com/file/d/1x2qV4JiYZA8njhZ5y6_t5Vsxdqmzinhg/view?usp=sharing

III. DEFINICIÓN DE LAS FUENTES DE DATOS

Los datos utilizados en este proyecto provienen de la plataforma "Kaggle", que ofrece miles de datasets gratuitos para la práctica de análisis de datos. En nuestro caso, decidimos trabajar con datos relacionados con

Netflix, una de las plataformas de streaming de películas y series más reconocidas a nivel mundial, con millones de usuarios. Esto nos permitió acceder a una amplia cantidad de información disponible para el análisis.

Los archivos de datos que utilizamos en el proyecto son:

1. **NETFLIX Stock Data**
2. **Netflix Reviews [DAILY UPDATED]**
3. **Netflix Movies and TV Shows**
4. **Datos Bursátiles de Netflix**
5. **Netflix User Base**

Estos archivos contienen una variedad de información, como reseñas de usuarios, rankings de películas y series, los países donde Netflix tiene mayor audiencia, y datos sobre el comportamiento bursátil de la compañía. Todos estos datasets están en formato CSV (texto plano) y fueron abiertos en Excel para obtener una vista más clara y ordenada de su contenido.

El siguiente paso fue integrar los datos en Power BI, donde construimos un modelo de base de datos relacional. A partir de este modelo, generamos un diagrama entidad-relación (ERD) que representaba las conexiones entre las diferentes tablas y los datos.

En total, utilizamos cinco datasets con información diversa. Los datos se organizaron en tablas separadas, estableciendo las relaciones correspondientes a través de claves primarias (PK) y claves foráneas (FK), lo que nos permitió conectar adecuadamente los diferentes conjuntos de datos y realizar un análisis más profundo y visual de la información.

IV. REFERENCIAS BIBLIOGRÁFICAS

[1] A. González, "Ésta es la historia del origen de Netflix, la empresa que desplazó a Blockbuster", *Grupo Milenio*, 29-ago-2023. [En línea]. Disponible en: https://www.milenio.com/espectaculos/television/la-historia-detras-de-netflix-origenes-y-fundacion. [Consultado: 15-oct-2024].

[2] W. Guerra y M. R. Ichaso, "Cronología de Netflix: así se convirtió en el gigante del streaming", *CNN en Español*, 18-jul-2022. [En línea]. Disponible en: https://cnnespanol.cnn.com/2022/07/18/cronologia-netflix-asi-se-convirtio-gigante-streaming-orix/. [Consultado: 15-oct-2024].

[3] Abhishek. "NETFLIX Stock Data". Kaggle: Your Machine Learning and Data Science Community. Accedido el 16 de octubre de 2024. [En línea]. Disponible: https://www.kaggle.com/datasets/abhiram8/netflix-stock-data/code

[4] A. Kumar. "Netflix Reviews [DAILY UPDATED]". Kaggle: Your Machine Learning and Data Science Community. Accedido el 16 de octubre de 2024. [En línea]. Disponible: https://www.kaggle.com/datasets/ashishkumarak/netflix-reviews-playstore-daily-updated/data

[5] A. Kumar. "Netflix Reviews [DAILY UPDATED]". Kaggle: Your Machine Learning and Data Science Community. Accedido el 16 de octubre de 2024. [En línea]. Disponible: https://www.kaggle.com/datasets/ashishkumarak/netflix-reviews-playstore-daily-updated/data