Nina Odoux
MUCD

Fecha de entrega : 01/12/24

REPORTE ETL.

VIDEOJUEGOS

Fase I: Eleccion de Datos

ENLACE FUENTE : Video Games Sales Dataset
'https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset'

RELEVANCIA DEL DATASET

Este dataset es importante porque muestra cómo se venden los videojuegos en diferentes regiones y consolas como Xbox y PlayStation. Ayuda a entender qué juegos y géneros son más populares, y qué prefieren las personas en cada lugar.

Las empresas pueden usar esta información para decidir dónde lanzar sus juegos, qué géneros desarrollar y cómo enfocar sus campañas de marketing. También permite analizar si los juegos con mejores calificaciones venden más y cómo las opiniones de los jugadores y críticos influyen en las ventas.

Además, el dataset ayuda a estudiar patrones, como cuáles son los mejores momentos del año para lanzar un juego o cómo las clasificaciones por edades afectan el mercado. En general, es una herramienta valiosa para tomar decisiones estratégicas sobre productos, ventas y marketing en la industria de los videojuegos.

Fase II: Documentacion

Objetivos del Proyecto :

- Propósito del proyecto.

Este proyecto tiene como objetivo proporcionar una solución práctica para analizar las ventas de videojuegos a través de un proceso ETL. Al ofrecer datos estructurados y limpios, se busca que las empresas puedan tomar decisiones más informadas y adaptadas a sus necesidades comerciales.

El primer objetivo es ayudar a identificar los mercados clave analizando las ventas por región. Esto permite a las empresas enfocarse en áreas con mayor potencial de crecimiento y optimizar sus estrategias de distribución. También se busca entender el desempeño de los videojuegos según las plataformas, géneros y períodos de tiempo, lo que facilita la planificación de lanzamientos y la priorización de productos.

Otro objetivo es analizar la relación entre las calificaciones de críticos y usuarios con las ventas. Esto proporciona información sobre la importancia de la percepción del producto en el mercado y cómo puede influir en el éxito comercial. Además, el análisis de patrones de lanzamiento ayudará a identificar los momentos más estratégicos para presentar nuevos productos, maximizando el impacto en las ventas.

Este proyecto también permitirá detectar oportunidades en mercados menos explotados y generar reportes confiables para diseñar campañas de marketing más efectivas. Finalmente, la preparación de los datos para futuros análisis predictivos ayudará a anticipar tendencias y posicionar a la empresa de manera competitiva en la industria de los videojuegos.

- Objetivos específicos del proyecto

El propósito de este proyecto es diseñar y ejecutar un proceso ETL que permita analizar y comparar los datos de ventas de videojuegos en plataformas como PS4 y Xbox One. Este proceso tiene como objetivo principal organizar y limpiar los datos para facilitar su almacenamiento en un sistema estructurado, permitiendo al cliente manejar su información de manera más eficiente y accesible.

El análisis de estos datos busca identificar tendencias de mercado, patrones de ventas y factores clave que influyen en el éxito comercial de los videojuegos. Los resultados ofrecerán insights valiosos para desarrolladores, equipos de marketing y distribuidores, ayudándoles a tomar decisiones estratégicas basadas en datos confiables y organizados. Además, el sistema resultante permitirá realizar análisis futuros con mayor rapidez y precisión, maximizando el valor de la información almacenada.

FLUJO DE TRABAJO:

Vamos a ordenar los objetivos de cada capa del proyecto para que sea organizado el flujo de trabajo:

Extracción:

Cargamos los archivos CSV en ‘raw’ que contienen los datos de ventas de videojuegos, verificando que las columnas sean legibles y correctas con la separación adecuada

Creamos un Data Catalog para documentar los campos y las metadatos de las tablas para que sean entendibles las variables

Diseñamos un modelo conceptual para anticipar la organización de las tablas en la base de datos y garantizar una estructura lógica.

Transformación:

Realizamos una análisis de calidad de los datos para identificar inconsistencias, valores faltantes, duplicados y anomalías, y obtener un valor global de calidad de la tabla

Concluimos sobre los fallos e inconsistencias de la tabla, apuntando como planeamos arreglarlos.

Normalizamos si es necesario y limpiamos los datos, corrigiendo problemas como registros duplicados.

Diseñamos un modelo relacional y un modelo dimensional

Carga:

Almacenamos los datos transformados en una base de datos con una tabla de hechos

Los datos estarán listos para análisis estratégicos y su uso estratégico empresarial

ARQUITECTURA DE CARGA :

Teniendo en consideración que el Data Warehouse está diseñado para almacenar datos estructurados, parece ideal para el análisis histórico y reportes de ventas.

Al evaluar las necesidades del proyecto, elegimos el Data Warehouse como la mejor opción. Ya que es perfecta para almacenar y analizar datos estructurados, como las que tenemos aquí con ventas de videojuegos organizadas por plataformas, géneros y regiones... Nos permitirá manejarlas para identificar tendencias y patrones de ventas, que son el foco principal del proyecto.

Los datos para el proyecto son estructurados y están en formato CSV, incluyendo campos como plataforma, ventas globales y regionales, fecha de lanzamiento y género del videojuego. El Data Warehouse permitirá organizar estos datos en tablas relacionales

Esta arquitectura facilita la visualización con herramientas tales como Power BI y Tableau, que es muy útil en el mundo comercial. Además, un Data Warehouse es escalable y puede manejar el crecimiento de datos con el tiempo, garantizando un rendimiento óptimo para reportes regulares. Aunque esta decisión dependerá también de los recursos del cliente.

• Justificación de la importancia de realizar un proceso ETL.

El proceso permite integrar nuestros datos de diferentes fuentes (.csv) y formatos (con distintas variables) en un único repositorio. Esta centralización facilita el manejo y el análisis de los datos, eliminando la necesidad de trabajar con información dispersa. Los datos crudos suelen contener errores, valores inconsistentes o elementos irrelevantes.

ETL garantiza la calidad de los datos al limpiar y transformar la información, logrando un conjunto de datos confiables y aptos para el análisis. El proceso ETL reduce significativamente el esfuerzo manual necesario para preparar los datos. Esta automatización permite a los analistas enfocarse en tareas de mayor valor, como interpretar resultados y generar estrategias basadas en los datos. Con datos organizados y de alta calidad, se pueden identificar patrones y tendencias con mayor precisión. Esto contribuye a una toma de decisiones basada en hechos concretos y no en suposiciones. El diseño del proceso ETL permite que sea escalable, lo que significa que puede adaptarse al crecimiento del volumen de datos sin necesidad de rediseñar desde cero. Esto asegura una gestión eficiente incluso ante un aumento significativo de información en el futuro. El proceso ETL es esencial para convertir datos crudos en información valiosa y útil, optimizando tanto el tiempo como los recursos en cualquier organización.

PORQUE UN PROCESO ETL ES ÚTIL

Un ETL es un proceso que toma los datos de diferentes lugares, los limpia y organiza, y los guarda en un sistema fácil de usar. Esto asegura que la información esté lista para analizar y tomar decisiones sin errores ni complicaciones. Para su negocio, un ETL ayuda a manejar mejor los datos de ventas de videojuegos, haciendo que sea más sencillo identificar tendencias y oportunidades. Con datos claros y organizados, podrán tomar decisiones más rápidas y efectivas para mejorar su estrategia y sus resultados.

Dataset:

• Enlace a los datasets utilizados:

Video Games Sales Dataset
‘https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset’

• Explicación de las fuentes de datos, acompañado de un Data Catalog
o Data Dictionary que describa las principales caracteriısticas de cada campo:

Origen	Nombre de Campo	Tipo de Dato	Descripción
PS4_GamesSales.csv	Game	Texto	Nombre del videojuego (PS4).
PS4_GamesSales.csv	Year	Entero	Año de lanzamiento del juego.
PS4_GamesSales.csv	Genre	Categoria/texto	Género del juego
PS4_GamesSales.csv	Publisher	Texto	Nombre de la empresa que publicó el juego.
PS4_GamesSales.csv	North America	Decimal	Ventas totales del juego en América del Norte (en millones de unidades).
PS4_GamesSales.csv	Europe	Decimal	Ventas totales del juego en Europa (en millones de unidades).
PS4_GamesSales.csv	Japan	Decimal	Ventas totales del juego en Japón (en millones de unidades).
PS4_GamesSales.csv	Rest of World	Decimal	Ventas totales en el resto del mundo (en millones de unidades).
PS4_GamesSales.csv	Global	Decimal	Ventas totales a nivel mundial (suma de las ventas por región, en millones).
Video_Games_Sales_as_at_22_Dec_2016.csv	Name	Texto	Nombre del videojuego.
Video_Games_Sales_as_at_22_Dec_2016.csv	Platform	Texto	Plataforma en la que se lanzó el juego (PS4, Xbox, PC, etc.).
Video_Games_Sales_as_at_22_Dec_2016.csv	Year_of_Release	Entero	Año de lanzamiento del juego.
Video_Games_Sales_as_at_22_Dec_2016.csv	Genre	Categoria/texto	Género del juego
Video_Games_Sales_as_at_22_Dec_2016.csv	Publisher	Texto	Nombre de la empresa que publicó el juego.
Video_Games_Sales_as_at_22_Dec_2016.csv	NA_Sales	Decimal	Ventas del juego en América del Norte (en millones de unidades).
Video_Games_Sales_as_at_22_Dec_2016.csv	EU_Sales	Decimal	Ventas del juego en Europa (en millones de unidades).
Video_Games_Sales_as_at_22_Dec_2016.csv	JP_Sales	Decimal	Ventas del juego en Japón (en millones de unidades).
Video_Games_Sales_as_at_22_Dec_2016.csv	Other_Sales	Decimal	Ventas del juego en otras regiones (en millones de unidades).
Video_Games_Sales_as_at_22_Dec_2016.csv	Global_Sales	Decimal	Ventas totales a nivel mundial (suma de las ventas por región, en millones).
Video_Games_Sales_as_at_22_Dec_2016.csv	Critic_Score	Entero	Puntaje promedio dado por los críticos al juego (normalmente de 0 a 100).
Video_Games_Sales_as_at_22_Dec_2016.csv	Critic_Count	Entero	Número de críticas hechas al juego por medios especializados.
Video_Games_Sales_as_at_22_Dec_2016.csv	User_Score	Decimal	Puntaje promedio dado por los usuarios al juego (normalmente de 0 a 10).
Video_Games_Sales_as_at_22_Dec_2016.csv	User_Count	Entero	Número de usuarios que puntuaron el juego.
Video_Games_Sales_as_at_22_Dec_2016.csv	Developer	Texto	Nombre del desarrollador del juego.
Video_Games_Sales_as_at_22_Dec_2016.csv	Rating	Categoria/texto	Clasificación del contenido del juego (E, T, M, etc.) segun el sistema americano
XboxOne_GameSales.csv	Pos	Entero	Posición del juego en el ranking de ventas.
XboxOne_GameSales.csv	Game	Texto	Nombre del videojuego
XboxOne_GameSales.csv	Year	Entero	Año de lanzamiento del juego.
XboxOne_GameSales.csv	Genre	Categoria/texto	Género del juego (acción, aventura, deportes, etc.).
XboxOne_GameSales.csv	Publisher	Texto	Nombre de la empresa que publicó el juego.
XboxOne_GameSales.csv	North America	Decimal	Ventas totales del juego en América del Norte (en millones de unidades).
XboxOne_GameSales.csv	Europe	Decimal	Ventas totales del juego en Europa (en millones de unidades).
XboxOne_GameSales.csv	Japan	Decimal	Ventas totales del juego en Japón (en millones de unidades).
XboxOne_GameSales.csv	Rest of World	Decimal	Ventas totales en el resto del mundo (en millones de unidades).
XboxOne_GameSales.csv	Global	Decimal	Ventas totales a nivel mundial (suma de las ventas por región, en millones).

- ➢ Para poder analizar, esta tabla de Excel está disponible en el contenido del proyecto bajo el nombre ‘Data Catalog’.

• Frecuencia de actualización de los datos.

La frecuencia de actualización es clave para la calidad del análisis, pero en este caso no está especificada en la fuente de KAGGLE. Esto genera una diferencia importante entre las tablas: con nuestra análisis hemos detectado que Video_Games_Sales_as_at_22_Dec_2016 refleja datos de ventas multi plataformas hasta el 22 de diciembre de 2016, mientras que PS4_GamesSales y XboxOne_GameSales incluye ventas uniplataforma hasta 2020.

Problemas detectados:
- Los datos de ventas entre las tablas no son directamente comparables por la diferencia de fechas.
- Los datos de 2016 pueden estar desactualizados para análisis recientes.

Propuesta de adaptación:
- Documentar las fechas de actualización de cada tabla para evitar confusiones, así podemos tener insights sobre las diferencias de conteos de ventas y des unidades porque tenemos distintas fechas de actualización. El usuario podrá decidir qué datos utilizar de manera sencilla con las consultas. Sin renovar información.
- Informaremos al cliente sobre las limitaciones temporales de los datos y la estructura del modelo en un archivo README: en la carpeta de LOAD.

Caracteŕısticas de los Datos:

• Descripción de los tipos de datos manejados:

En nuestro conjunto de datos, trabajamos con diferentes tipos de datos: textuales, temporales, numéricos y categoricos.

Cada tipo aporta insights valiosos sobre las ventas:

Datos textuales Incluyen información como el nombre de los videojuegos y el publisher. Estos datos son esenciales para identificar y analizar individualmente cada juego y su origen.

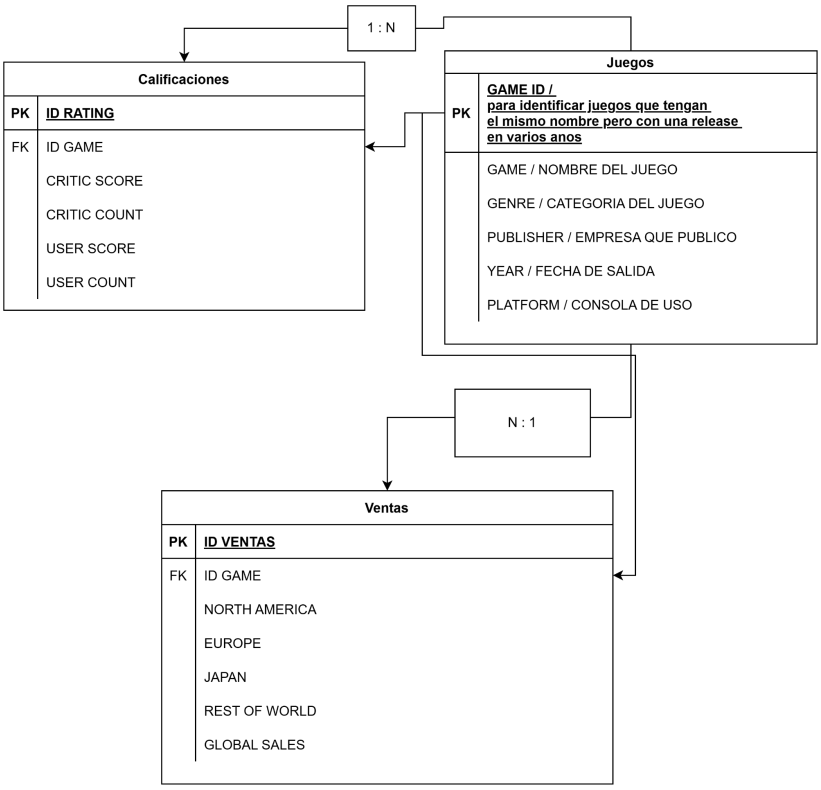
Datos categoricos permiten clasificar y organizar los datos según categorías como el género del juego. Esto facilita el analisis de las tendencias del mercado y las preferencias de los consumidores en diferentes generos.

Datos numericos comprenden las cifras de ventas en unidades, tanto para **Xbox One como para PS4 entre 2014 y 2020,** asi como las **ventas globales hasta el 22 de diciembre de 2016.** Estos datos cuantitativos son fundamentales para medir el rendimiento de ventas y comparar el exito entre plataformas y titulos.

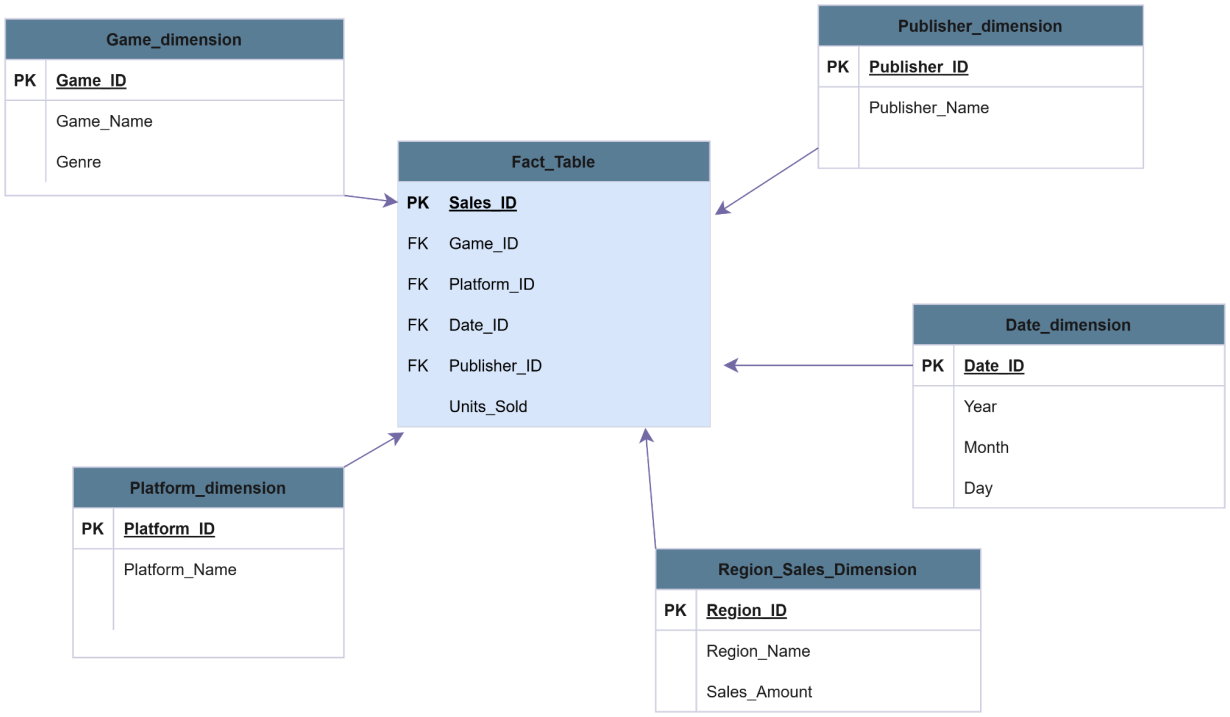
Datos temporales, aunque se representan como valores numericos (tipo 'float'), corresponden a fechas que son imprescindibles para saber la fecha de lanzamiento de cada version de cada juego.

• Definición de los Modelos de Datos

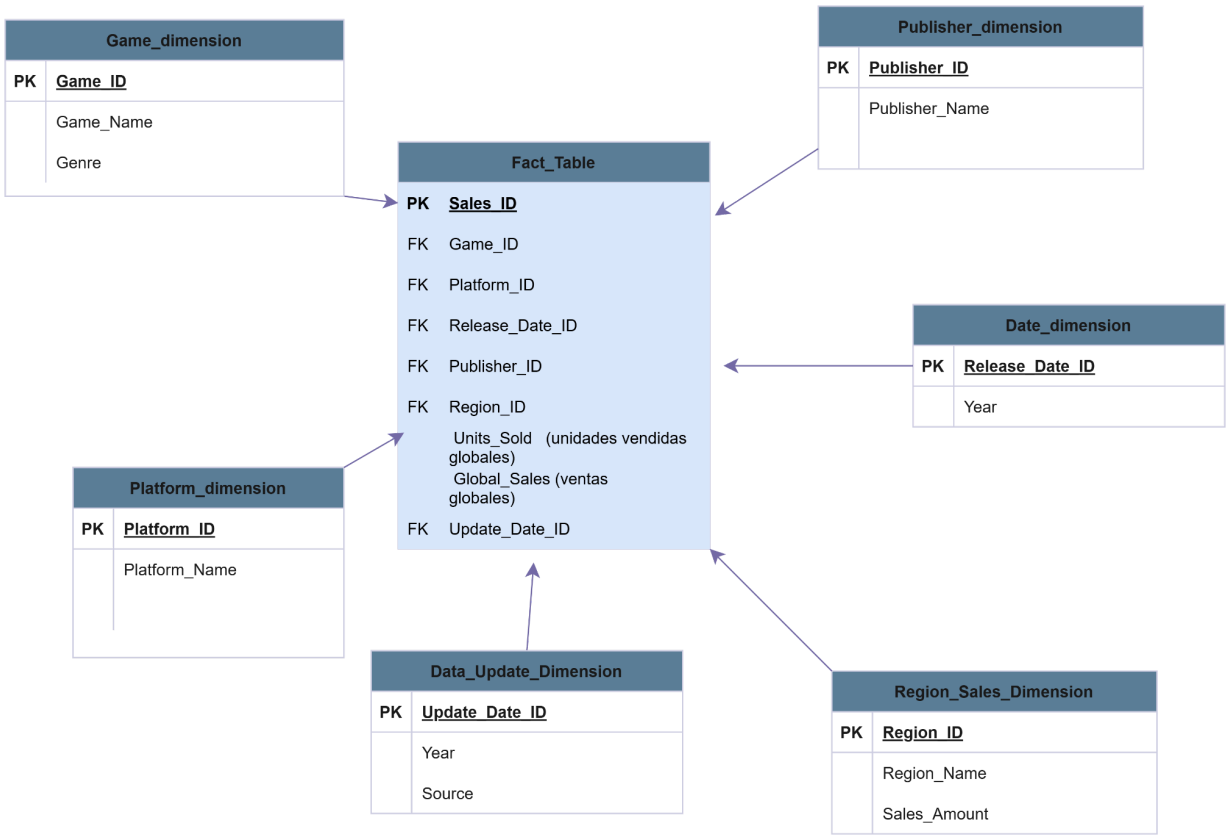
Modelo relacional



Modelo Dimensional Esquema Estrella 1



Modelo Dimensional Esquema Estrella 2,
ADAPTADO A PROBLEMAS DE ACTUALIZACIÓN :



EXPLICACIÓN DE LA ELECCIÓN:

Hemos elegido este modelo dimensional para nuestro objetivo por su capacidad para manejar datos de ventas de videojuegos de manera más eficiente y flexible (para la análisis y la visualización luego). Porque es bien estructurado con una única tabla de hechos central y dimensiones específicas (para una mejor comprensión y entendibilidad) y de esta manera asegura que sea fácil de usar y manejar para cualquier. Además es importante que sea escalable y preparado para análisis avanzados (por ejemplo tener la opción de añadir nuevas regiones…)

Calidad de los Datos:

Es decir encontrar el % de calidad de datos global para cada tabla:
• Evaluación de la calidad de los datos, señalando los problemas detectados.

Resumen de problemas encontrados en cada tabla:

PS4

Data_Quality_PS4			
Metric	**Description**	**Value**	**Observations**
Completitud	Media del porcentaje de datos no nulos para todas las columnas.	95.51	
Accuracy	Media del porcentaje de valores correctos para las métricas de Year y Global Sales(las unicas que podemo sverificar).	73.11	Incluye la precisión de Year y Global Sales.
Linaje	Porcentaje de rastreabilidad fiable en la tabla.	100	El linaje es completo gracias a la columna Game, que permite rastrear valores faltantes con fuentes externas.
Semantica	Porcentaje de filas con datos validos semánticamente (ventas no negativas, Publisher válido y Year dentro del rango logico).	79.79	Incluye verificaciones para ventas no negativas, Publisher válido y Year dentro del rango lógico (de 1950: decade de creacion del primer videojuego, hasta ahora).
Estructura	Promedio de validez estructural basado en columnas Year, Publisher y Genre.	59.93	Incluye validación de Year con un tipo DATE, Publisher con mayúscula o número inicial, y Genre comenzando con una mayúscula, los numerico ya estan todos en decimales (.2) con las normativas de la tabla.
Consistencia	Porcentaje de filas donde la suma de ventas regionales coincide con las ventas globales.	66.44	Verifica que las ventas en North America, Europe, Japan y Rest of World sumen correctamente a Global para que sea un total justo.
Moneda	Esta métrica no aplica aqui porque los valores no representan datos monetarios ni cuya medida puede cambiar a lo largo del tiempo		No es relevante calcular esta métrica para los datos actuales.
Puntualidad	Esta métrica no aplica aqui el dataset no trata de actualizaciones obligatorias o rangos temporales precisos de ventas		No es relevante calcular esta métrica para los datos actuales.
Identificabilidad	Porcentaje de filas únicas en funcion de Game y Year como indicatores.	99.9	Identifica duplicados considerando juegos con el mismo nombre y la misma fecha de lanzamiento.
Razonabilidad	Promedio de razonabilidad basado en consistencia de ventas y rangos razonables (0-50 millones).	83.37	Los valores fuera del rango 0-50 millones serán revisados durante el proceso de limpieza.
TOTAL	PROMEDIO DE CALIDAD GLOBAL BASADO EN VALORES DE METRICAS OBTENIDAS	82.26	

La completitud del dataset, aunque puede parecer faltar datos esenciales, no pone un verdadero problema porque en realidad son datos faltantes que no son obligatorios (Year y Publisher) y que se pueden encontrar con la presencia de variables de la fila tal que el nombre del juego que tiene una completitud del 100%. Podemos considerar a nivel empresarial usar una API para poder llenar de manera automatizada estos campos según el nombre del videojuego.

Lo que hemos podido notar en cuanto a la precisión de campos relacionados matemáticamente entre ellos (que el total de ventas ‘Global’ sea igual a la suma de las regiones alistadas). Es que el total no coincide, lo que puede querer decir que hay regiones que participan el total que no han sido recopiladas (como América Latina, o Asia en general)... En el modelo dimensional se considerará crear un nuevo total de ventas con los datos que tenemos para que sea más consistente.

La columna de Year se debe pasar a tipo Date para que se entienda mejor la tabla.

No parece tener duplicados, porque la única fila que es duplicada, simplemente lo es porque el año de lanzamiento aparece como NULL en ambas. Se supone que es una release o update del videojuego, y que no hace falta removar el duplicado.

XBOX:

Data_Quality_Xbox			
Metric	Description	Value	Observation
Completitud	Promedio de completitud para cada variable	96.48	
Accuracy	Promedio de precision de la variable Year.	82.38	Incluye valores dentro del rango, es la unica metrica que podemos verificar con los recursos que tenemos
Linaje	Porcentaje de rastreabilidad fiable en la tabla.	100	El linaje es completo gracias a la columna Game, que permite rastrear valores faltantes con fuentes externas.
Semantica	Porcentaje de filas con datos validos semánticamente (ventas no negativas,y Publisher valido).	82.38	Verificaciones de ventas no negativas y Publisher válido.
Estructura	Promedio de validez estructural basado en las normas establecidas.	70.47	Validación de Year con tipo DATE, variables textuales con primer caracter en mayúscula o número inicial.
Consistencia	Porcentaje de filas donde las ventas Global coinciden con la suma de las ventas para cada region.	64.27	Verificación realizada sobre ventas globales y regionales para asegurar la coherencia de datos.
Moneda	Esta métrica no aplica aqui porque los valores no representan datos monetarios ni cuya medida puede cambiar a lo largo del tiempo		No es relevante calcular esta métrica para los datos actuales.
Puntualidad	Esta métrica no aplica aqui el dataset no trata de actualizaciones obligatorias o rangos temporales precisos de ventas		No es relevante calcular esta métrica para los datos actuales.
Identificabilidad	Porcentaje de registros unicos basados en Game y si Year es distinto.	100	Se considera duplicado si Game y Year son iguales en varias filas.
Razonabilidad	Promedio de razonabilidad basado en consistencia de ventas y rangos razonables (0-50 millones).	82.22	Los valores fuera del rango 0-50 millones serán revisados durante el proceso de limpieza.
TOTAL	PROMEDIO DE CALIDAD GLOBAL BASADO EN VALORES DE METRICAS OBTENIDAS	84.78	

El mayor problema es la falta de consistencia entre el total de ventas global y las ventas regionales; exactamente como la tabla de PS4, lo que refuerza la impresión de que algunas regiones no han sido recopiladas en los csv.

En cuanto al tipo, la columna Year tampoco es de tipo DATE.

La estructura tampoco parece muy completa, y sería imprescindible normalizar los valores.

SALES de diciembre 2016:

Data_Quality_22Dec2016			
Metric	Description	Value	Observation
Completitud	Porcentaje promedio de datos no faltantes en todas las columnas.	81.63	Se verifica la completitud general del dataset con el promedio de regirstros de cada variable.
Accuracy	Promedio de accuracy basado en ventas no negativas y precisas y rango lógico del año de lanzamiento.	89.41	Se verifica la precision de los datos de ventas y de Year.
Semantica	Promedio de validez basado en ventas no negativas, scorings en el rango y valores validos que existen en Rating.	97.87	Verificacion del sentido de las variables, de sus rangos, de acuerdo con el data catalog y la logica.
Estructura	Promedio de validez basado en tipo de datos y formato de columnas, ignorando valores NULL.	67.05	Los valores que no cumplen serán normalizados en el proceso de limpieza.
Moneda	No aplica porque no existen valores monetarios en este dataset.		Esta métrica no es relevante para este análisis.
Puntualidad	No aplica en este caso, ya que no hay campos que requieran actualizacion.		El dataset contiene datos de fecha de lanzamiento y no requiere evaluacion de puntualidad.
Linaje	Porcentaje de rastreabilidad fiable en la tabla.	100	El linaje es completo gracias a la columna Name, que permite rastrear valores faltantes con fuentes externas.
Consistencia	Porcentaje de registros donde las ventas globales coinciden con la suma de las ventas regionales.	59.24	Nos da insights sobre la consistencia de los datos de ventas, sobre el total global.
Identificabilidad	Porcentaje de registros unicos basados en Name, Year_of_Release y Platform cuando todas las columnas están completas.	99.99	La identificabilidad se asegura si los registros únicos basados en Name, Year_of_Release y Platform superan el 95%.
Razonabilidad	Promedio de razonabilidad basado en consistencia de ventas y puntajes en un rango establecido.	94.71	
TOTAL	Promedio de todas las metricas de calidad.	86.89	

El problema más importante es la presencia de datos fuera del rango lógico temporal, hay 3 registros fuera de 2016 en Year of Release.
De revisar el tipo de la columna de Year a DATE, y los requisitos de estructura con mayúsculas

MÉTRICAS NO EMPLEADAS :

En este proyecto no se incluyeron unas métricas de calidad debido a su falta de relevancia para el análisis específico del dataset.

La métrica de ‘Puntualidad’ se utiliza generalmente para evaluar si los datos están actualizados o alineados con un tiempo específico. Sin embargo, este dataset tiene un enfoque histórico y está basado en datos de ventas de 2016. Dado que no se requiere evaluar la frecuencia de actualización o vigencia de los datos, esta métrica no es aplicable en este caso.

Por otro lado, la métrica de ‘Moneda’ se emplea para evaluar la calidad de los datos financieros, como precios o transacciones monetarias. En este dataset, las columnas de ventas representan unidades vendidas y no valores monetarios, lo que hace innecesaria esta métrica. Por ello, se optó por no incluirla en el análisis de calidad.

Estas decisiones aseguran que el análisis se centre únicamente en métricas relevantes para la naturaleza de los datos y los objetivos del proyecto.

Limpieza de los Datos:

• Descripción´on de los métodos y ́ecnicas utilizadas para la limpieza y preparación de los datos.

Hemos trabajado en la limpieza y normalización de tres conjuntos de datos de ventas de videojuegos: "Video_Games_Sales_as_at_22_Dec_2016", "XboxOne_GameSales" y "PS4_GamesSales".

En el primer conjunto Video_Games_Sales_as_at_22_Dec_2016, corregimos la inconsistencia entre la columna "Global_Sales" y la suma de las ventas regionales, asegurando que el total global sea exactamente la suma de "NA_Sales", "EU_Sales", "JP_Sales" y "Other_Sales", lo que refleja con precisión las ventas en cada región.

Creamos una nueva columna "Year" en formato de fecha "YYYY-01-01" para facilitar análisis temporales y eliminamos la antigua columna de años.

Normalizamos los valores textuales en las columnas "Name", "Platform", "Genre", "Publisher", "Developer" y "Rating" para que todos los valores comienzan con mayúscula, mejorando la consistencia y el formato.

En "XboxOne_GameSales", creamos la columna "Year Release" en formato de fecha, reemplazando la columna "Year" original.
Y aplicamos capitalización inicial a las columnas "Game", "Genre" y "Publisher".
Eliminamos registros duplicados basados en "Game", "Year Release", "Genre" y "Publisher".

Para **"PS4_GamesSales"**, realizamos cambios similares: creamos la columna "Year Release" en formato de fecha, normalizamos los valores textuales y eliminamos duplicados. Estos ajustes mejoran la consistencia de los datos, permiten análisis temporales más precisos y aseguran que los cálculos y reportes sean confiables, optimizando la calidad y evitando redundancias en los datos.

- **Para todas** : Reemplazamos valores nulos en las ventas regionales por cero para garantizar cálculos, para las textuales ponemos un NULL. Hemos adaptado el cálculo de columnas de totales globales, a la real suma de cada región y resto del mundo para que los datos sean correctos.Suponiendo que el fallo era en el cálculo (aunque se considera que el fallo puede provenir de otra causa)

Problemas y Próximos Pasos:

Durante las fases de extracción y transformación de datos, enfrentamos desafíos como la inconsistencia y actualización desigual de información proveniente de distintas fuentes y fechas, lo que generó discrepancias en cifras de ventas y detalles de videojuegos.

También lidiamos con valores nulos y datos faltantes en columnas textuales y numéricas, lo que dificulta un análisis completo.

La normalización de formatos y estándares entre las diferentes fuentes fue esencial para unificar el esquema y garantizar una integración en el modelo dimensional.

Además, fue necesario detectar y eliminar duplicados para evitar redundancias y asegurar que cada videojuego estuviera representado de manera única. (un paso de amelioration podría ser de revisar la estructura de registros textuales para normalizar más allá que el primero carácter; la estructura global con cada palabra que empecé con una mayúscula)

Como propuesta de mejora, sugerimos también integrar una API para completar datos faltantes en columnas textuales, aprovechando el nombre del videojuego para obtener información adicional como género, editor y fecha de lanzamiento, lo que enriquecería nuestro modelo.

Fase III: Codificación

- Tres Scripts se adjuntan en la carpeta del proyecto bajo estos nombres para recopilar pasos de nuestro ETL:

- 01_EXTRACTION
- 02_TRANSFORMATION
- 03_LOAD

• Dataset Final:

- En cada etapa del proyecto, (representadas con carpetas distintas), ponemos los CSV actualizados según la etapa; en la etapa final de LOAD,aparecen las tablas del modelo dimensional en formato CSV. Representando la estructura en el sistema de almacenamiento final.

CONCLUSION :

Este proyecto ETL nos permitió organizar y limpiar datos de ventas de videojuegos para crear un sistema estructurado y fácil de analizar. Resolvimos problemas como inconsistencias, valores nulos y datos duplicados.

El modelo dimensional que diseñamos facilita el análisis y la visualización, ayudando a las empresas a tomar decisiones estratégicas basadas en datos confiables. Aunque hubo desafíos, como la desactualización de algunos datos, propusimos soluciones para mejorar la precisión en futuros análisis.

En resumen, este proyecto cumple con los objetivos de centralizar y mejorar los datos, brindando herramientas útiles para el análisis y la toma de decisiones en la industria de los videojuegos para los usuarios.