UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERIA Y ARQUITECTURA

ESCUELA DE INGENIERIA DE SISTEMAS INFORMATICOS

INGENIERÍA DE DATOS

CURSO DE ESPECIALIZACION DE INGENIERIA DE DATOS

DISEÑO DE UN MODELO DIMENSIONAL PARA SOPORTAR EL PROCESO DE NEGOCIOS DE STEAM.

PRESENTADO POR

ARIAS LÓPEZ, CARLOS AECIO

MOLINA GARCÍA, NESTOR ULISES

SÁENZ OSORIO, VÍCTOR MANUEL

PARA OPTAR AL TITULO DE:

INGENIERO DE SISTEMAS INFORMÁTICOS

CIUDAD UNIVERSITARIA, ENERO DE 2022

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSc. ROGER ARMANDO ARIAS ALVARADO

SECRETARIO GENERAL:

ING. FRANCISCO ANTONIO ALARCON SANDOVAL

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

PdH. EDGAR ARMANDO PEÑA FIGUEROA

SECRETARIO:

ING. JULIO ALBERTO PORTILLO

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. RUDY WILFREDO CHICAS VILLEGAS

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERIA Y ARQUITECTURA

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

Trabajo de Graduación previo a la opción al Grado de:

INGENIERO DE SISTEMAS INFORMÁTICOS

Título:

DISEÑO DE UN MODELO DIMENSIONAL PARA SOPORTAR EL PROCESO DE NEGOCIOS DE STEAM

Presentado por:

ARIAS LÓPEZ, CARLOS AECIO
NESTOR ULISES MOLINA GARCÍA
SÁENZ OSORIO, VÍCTOR MANUEL

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. MARLON ARMANDO MENJÍVAR MARTÍNEZ

SAN SALVADOR, ENERO 2022

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. MARLON ARMANDO MENJÍVAR MARTÍNEZ

Contenido

Tabla de contenidos

Tabla de figuras

INTRODUCCIÓN

El presente documento pretende mostrar un breve vistazo referente a la generación de un modelo dimensional para ayudar a un proceso del negocio. Por lo que veremos el proceso que se realiza al generar un modelo dimensional que apoya en la toma de decisiones que requiere una organización, para el caso de esta tesina es Steam; por lo que se muestra una breve introducción a la información más relevante, en donde se explica el modelo de negocios que posee Steam, permitiéndonos tener un enfoque más completo de la información a procesar.

Entre la información que se observará, se encuentran los resultados del data profiling que es el proceso de analizar la información mediante herramientas informáticas que nos ayudan a comprender la data que se posee.

Para realizar este proceso, se describen los diversos datasets que se tienen para generar el modelo dimensional y que nos permita solventar las diversas preguntas que necesitamos responder.

Podremos observar que se tiene también un diccionario de datos de los diversos campos que nos ayudarán a generar un modelo más adecuado a nuestras necesidades

Se muestra un diseño del modelo dimensional que nos ayudará a optimizar los procesos de la consulta de la data, definiendo las diversas estrategias de extracción, transformación y carga que se realizará de los diversos datasets

Otro de los elementos que se muestran son las Fact tables que se generarán a partir de la data de las diversas dimensiones conformadas que se definen en este documento

Se detallan las diversas dimensiones que se crearán para el almacenamiento de la data. Se definirán también las dimensiones Bridge que son aquellas que se utilizarán en este proceso de análisis de la información

OBJETIVOS

a. Objetivo General.

Generar un modelo dimensional que permita al usuario analizar la información recopilada de Steam, permitiendo responder diversas preguntas del proceso del negocio y aportando variedad de enfoques en la toma de decisiones.

b. **Objetivos Específicos.**

- Realizar un estudio de la plataforma de videojuegos Steam y su modelo de negocio, interacción de usuarios con la plataforma que permita establecer requerimientos analíticos para la toma de decisión
- Diseñar un modelo dimensional con el nivel de granularidad necesaria que permita adaptarse a las necesidades del negocio
- Generar una solución que permita procesar los datos de los datasets con el fin de proporcionar una solución que será ejecutada periódicamente de acuerdo a las necesidades del negocio
- Construir un adecuado proceso de ETL que permita una correcta transformación de las diferentes fuentes de datos
- Implementar una estructura de Cloud Computing que permitan el despliegue del modelo dimensional en el Data Warehouse
- Brindar estadísticas relevantes en el proceso de negocios de Steam, teniendo como meta la implementación de nuevos seguimientos o énfasis especial en algún tipo de producto que esté siendo utilizado en condiciones favorables para la compañía y así mejorar su toma de decisiones
- Crear visualizaciones de los procesos de negocio de Steam referente a ventas, reseñas y tiempos de juegos, con la finalidad de observar y comprender tendencias y comportamientos que los usuarios tienen respecto a algunos productos

CAPÍTULO 1: MARCO TEÓRICO.

1.1. Origen del objeto de investigación

Dentro del progreso de la informática como rama de la ingeniería, y en el marco de los contemporáneos y muy bien elaborados conceptos de servicios web, cloud computing, data lakes y demás tópicos involucrados en la disciplina de la ciencia de datos, nos encontramos con novedades que llevan la ingeniería de sistemas a una de sus bifurcaciones más novedosas y complejas, como es la creación y mantenimiento de data warehouses, estos encaminados a ser la herramienta principal en el contexto de toma de decisiones con el uso, de por medio de la inteligencia de negocios.

Si bien, el objeto de la investigación, principalmente es demostrar el ciclo de vida completo para la implementación de un datawarehouse, en el camino de creación nos encontramos con el uso de múltiples tecnologías y herramientas propias de los científicos de datos. En nuestro recorrido, que va desde la obtención de datos crudos, provenientes de microtransacciones diarias de la empresa que hemos decidido abordar e información que hemos obtenido en línea a través de archivos que consideramos valiosos de estudiar e interpretar para luego procesar toda esta información por medio de las tecnologías que muestran bondades y facilidades para el mantenimiento de estas grandes cantidades de datos, y que es importante recalcar, se realiza por medio de la infraestructura implementada, de la que se obtuvo conocimiento teórico a través de esta especialización y que lleva como meta, el mostrar en ambientes gráficos las resoluciones a las que como equipo llegamos luego de procesar la data recopilada, y que ejemplificamos, facilitan la toma de decisiones y aporta juicios objetivos de la manera en que el negocio ha sido desarrollado.

Dicho esto, encontramos como origen de esta investigación en primera instancia, el gran avance que este tipo de tecnologías tiene en estos años, este mismo tipo de tecnologías y recursos que ha ayudado enormemente al pleno desarrollo de muchas disciplinas de vanguardia, como las ciencias de la información e ingeniería de datos, consideramos como equipo, que no solo se debe a un éxito contemporáneo el uso y formación de profesionales en estas áreas, sino que se han encontrado muchos vacíos,

que cabe destacar que este tipo de disciplinas han logrado solventar a plenitud. Este curso por finalizar es una de las mayores evidencias, debido a que esta implementación a mediana escala, con condiciones normales del negocio es un ejemplo que este tipo de análisis/estudios, es totalmente necesario en estos tiempos modernos con tanto flujo de información constante.

1.2. Valoración sobre la estructura de la especialización

Si bien, como equipo, y con la cátedra en conjunto, conocimos el plan de estudio que íbamos a recorrer con antelación, es destacable resaltar el hecho que algunos conceptos teóricos son bastante resonados en el área de informática, pero que en un contexto práctico, no habíamos tenido la oportunidad de experimentar con estos, el mayor ejemplo de estos son los referentes a cloud computing o big data, si bien el razonamiento de su funcionamiento o el caso en que este tipo de información es aplicable a lo que conocemos, como estudiantes de ingeniería de sistemas, en nuestro último año no tuvimos la oportunidad de aplicar estos a nuestro plan de estudio, y es esta especialización la que nos ha llevado a modelar este tipo de infraestructuras a un ámbito totalmente práctico y en un escenario con condiciones típicas con características bastante realistas.

Es de resaltar que la preparación que se tuvo para la especialización, en parte se dio en paralelo con el transcurso de esta, debido a que estuvo estructurada en 4 capítulos y venían de menor a mayor complejidad, de menor a mayor conocimiento común y de menor a mayor calidad de los prerrequisitos técnicos envueltos. Este punto anterior es vital para el desarrollo del curso, debido a que fue esta misma estructura de amplios estudios teóricos la que permitió que llegáramos a los módulos 3 (de desarrollo en spark) y 4 (introducción a PowerBI), con una base robusta en conocimiento teórico sin necesidad de estudiar estos contenidos en materias o cursos anteriores.

Es por todas estas características discutidas, que encontramos la estructura y contenido de la especialización totalmente acertados, a esta altura de tener los contenidos cursados, de haber realizado prácticas en múltiples tecnologías, es donde

encontramos que este vacío de áreas informáticas locales necesita ser cubierto y con este tipo de especializaciones, que esperamos lograr este cometido.

1.3. Data warehousing

El data warehousing, que es uno de los pilares de toda la especialización para representar de una manera amplia todo lo que involucra, y según algunos elementos que los propios tópicos del curso toma, lo desglosaremos para su estudio en 3 subtemas.

1.3.1. Introducción a Data Warehousing

La traducción plana de Data Warehousing, y como normalmente lo solemos llamar, el almacenamiento de datos, consiste de manera breve en poseer un repositorio centralizado con información que se puede analizar para una mejora o seguimiento en la toma de decisiones del negocio. Lo común es que los datos fluyan desde una base transaccional hacia este repositorio centralizado, normalmente a un ritmo previamente establecido. Si bien, pasan muchos estados en el transcurso de este proceso que se acaba de describir, los hitos importantes para usuarios análiticos/gerenciales, ya sean estos científicos de datos o ingenieros de datos, es el estudiar estos resultados por medio de herramientas de inteligencia empresarial, las denominadas BI; clientes SQL u otro tipo de aplicaciones analíticas.

Actualmente este tipo de dashboards, que son capaces de generar informes de valor, ya sea predictivos o resolutivos que en las empresas han tomado mayor relevancia a la hora de tomar decisiones a futuro. Monitorear el desempeño de la empresa con estas herramientas es fundamental hoy en día, en estos tiempos donde el día a día se vive con tanto tráfico de información, es en este paso en donde los almacenes de datos son pieza vital, sino es que indispensable en su totalidad. No existe otra manera en la que

se puedan entregar este tipo de resultados en tan corto tiempo y con un extensivo nivel de análisis como el que estas herramientas proporcionan.

¿Qué beneficios presenta el implementar los almacenes de datos?

- Toma de decisiones fundamentadas
- Datos consolidados con orígenes diferentes
- Análisis de datos históricos
- Calidad, coherencia y precisión en la información resultante
- Separación del procesamiento de análisis de las bases transaccionales, lo que mejora el rendimiento de ambos sistemas

1.3.2. ¿Cómo se relacionan las bases de datos, el almacenamiento de datos y los lagos de datos?

Se trae a la mesa de estudio un concepto que ronda y tiene presencia en tópicos de almacenaje, pero que en la práctica no se aplica con propiedad, los lagos de datos.

Para conceptualizar el lago de datos es necesario volver a lo que conocemos con antelación, acerca del almacenamiento de datos (datawarehouse), comprendimos que es una estructura de grandes dimensiones capaz de soportar información proveniente de muchos orígenes, y que normalmente vienen de bases transaccionales, en cambio, un lago de datos es un almacén de datos sin procesar en formato nativo, listo para su uso en el momento que se considere debido, la principal diferencia es que en el lago de datos la información está en bruto y en un nivel muy plano, mientras que en el almacenamiento de datos, la información se guarda por ficheros o carpetas. Entonces, ¿qué beneficio nos trae implementar un lago de datos? El principal beneficio es la centralización de las fuentes de contenidos, que obviamente no tienen el mismo origen, y que, al consultar información a este nivel de la infraestructura, podríamos conocer el área de donde proviene cada información, ya que aún no está procesada. Queda un poco separada de toda esta información, las bases de datos transaccionales, pero

como ya sabemos, su rol sigue siendo estable, proveer de contexto e información del negocio a los ya mencionados lagos de datos y almacenamiento de datos.

Tabla 1.
Principales diferencias entre Datawarehouse - DataLakes

Características	**Datawarehouse**	**DataLakes**
Datos	Datos relacionales provenientes de sistemas transaccionales, bases de datos operativas y aplicaciones de línea de negocio	Todos los datos, incluidos los estructurados, los semiestructurados y los no estructurados
Esquema	Con frecuencia se diseña antes de la implementación del almacenamiento de datos, pero también se puede escribir al momento del análisis	Escrito al momento del análisis (esquema de lectura)
Precio / rendimiento	Resultados de búsqueda más rápidos con almacenamiento local	Resultados de búsqueda más rápidos con almacenamiento de bajo costo y desacoplamiento de la informática y el almacenamiento

Características	Datawarehouse	DataLakes
Calidad de los datos	Datos seleccionados detalladamente que funcionan como fuente certera	Cualquier dato que pueda estar seleccionado o no (es decir, datos no procesados)
Usuarios	Analistas empresariales, científicos de datos y desarrolladores de datos	Analistas empresariales (que usan datos seleccionados), científicos de datos, desarrolladores de datos, ingenieros de datos y arquitectos de datos
Análisis	Generación de informes en lotes, inteligencia empresarial y visualizaciones	Aprendizaje automático, análisis de exploración, descubrimiento de datos, streaming, análisis de operaciones, big data y generación de perfiles

1.4. Modelado dimensional

El modelado dimensional hace referencia al uso de tablas de hechos (fact tables) y dimensiones, para mantener un registro de datos históricos en el almacenamiento de datos, esto a manera de resumen. Es de mencionar que los modelos normalizados de relación de entidad (los modelos ER) están diseñados para eliminar la redundancia de datos y realizar rápidamente las operaciones de inserción, actualización, y para obtener los datos dentro de una base de datos transaccional en su mayoría.

En comparación de esto mencionado, los modelos dimensionales son estructuras desnormalizadas diseñadas para recuperar datos de un almacén de datos, están

optimizados para realizar la operación _select_ y se utilizan en el marco del diseño básico para constuir almacenes de datos altamente optimizados y funcionales.

1.4.1. Beneficios del modelado dimensional

Recuperación de datos más rápida:

El modelado dimensional fusiona las tablas en el propio modelo, lo que permite a los usuarios recuperar datos más rápidamente de diferentes fuentes de datos mediante la ejecución de consultas conjuntas. El esquema desnormalizado de un almacén de datos con modelado dimensional está optimizado para ejecutar consultas que únicamente se harían con ese modelo en específico, es decir, se construye a la medida de las necesidades; esto complementa en gran medida los objetivos de inteligencia empresarial de una organización.

Mejor comprensión de los procesos comerciales:

Los principios del modelado dimensional se basan en tablas de hechos y dimensiones, es esta estructura entidad-relación de un modelo dimensional la que permite presentar procesos comerciales complejos de una manera sumamente amigable de comprender para los analistas de datos.

Flexible ante los cambios:

Este mismo marco de modelado dimensional hace que el proceso de almacenamiento de datos sea extensible. El diseño se puede modificar fácilmente para incorporar nuevos requisitos comerciales o realizar ajustes en el repositorio central. Es posible la incorporación de nuevas entidades en el modelo, o se pueden cambiar las existentes para reflejar cambios de los procesos comerciales.

1.4.2. Elementos de un modelo de datos dimensionales

Tablas de hechos (Fact tables):
Almacenan la información numérica sobre medidas comerciales y llaves externas para las tablas dimensionales. Los hechos comerciales pueden ser aditivos, semi aditivos, o no aditivos.

Tablas de dimensiones:
Almacenan la información descriptiva sobre los hechos comerciales para ayudar a entender y analizar mejor los datos. Las tablas de dimensiones poseen tanto llaves primarias, como llaves externas. La llave primaria como ya se conocen, es una columna en las tablas de dimensiones que identifica registros únicos. Las llaves externas se usan normalmente para unir dos tablas, generalmente el enlace es entre tabla de hecho y dimensión.

1.4.3. Diseño de un modelo de datos dimensionales

Paso 1: Identificar los procesos comerciales

Antes de modelar los datos, los tipos de modelado dimensional deber ser apropiados para su modelo de datos. El proceso de modelado dimensional comienza con la identificación del proceso empresarial que desea rastrear.

Paso 2: Identificar hechos y dimensiones

Como ya lo mencionamos, la información en un modelo dimensional se clasifica en tablas de hechos y dimensiones, por lo tanto el siguiente paso es identificar los hechos comerciales que se desea medir y sus dimensiones asociadas.

Paso 3: Identificar los atributos de las dimensiones

Una vez se hayan identificado los hechos y las dimensiones del proceso empresarial que se ha escogido, el siguiente paso es identificar los atributos y crear una tabla dimensional separada para cada una de las dimensiones. Muy importante, cada registro de la tabla de dimensiones debe tener una clave única, esta clave se usará para identificar de forma única e individual los registros en la tabla de dimensiones y se usará como llave externa en la tabla de hechos, para hacer referencia a una dimensión en particular.

Paso 4: Definir la granularidad de los hechos comerciales

La granularidad se refiere al nivel de información que se almacena por cualquier tabla, hace referencia al tamaño de este nivel escogido, como ejemplo podemos mencionar “ventas registradas diariamente”, la granularidad en este caso es diaria. Es importante que las tablas de hechos en un modelo dimensional deben ser consistentes con la granularidad predefinida, y se recomienda fuertemente escoger la granularidad más atómica posible, de esta manera si se requiere la información a un nivel superior es posible tenerlo de igual manera.

1.5. Metodología para diseño de DataWarehouses

Las metodologías más discutidas y principales suelen ser las diseñadas por Ralph Kimball, Bill Inmon y Dan Linsted.

1.5.1. Multidimensional (Ralph Kimball)

Según especifica Kimball, un almacén de datos es la copia de los datos transaccionales específicamente estructurados para consultas analíticas e informes, con el fin de

apoyar en la toma de decisiones (Kimball & Ross, 2013). Con esta metodología se busca crear primero los datamarts (pequeños almacenes de datos específicos de un área dentro de la empresa) y se proporcionan de capacidades analíticas para procesos específicos y funcionales.

Principales ventajas:

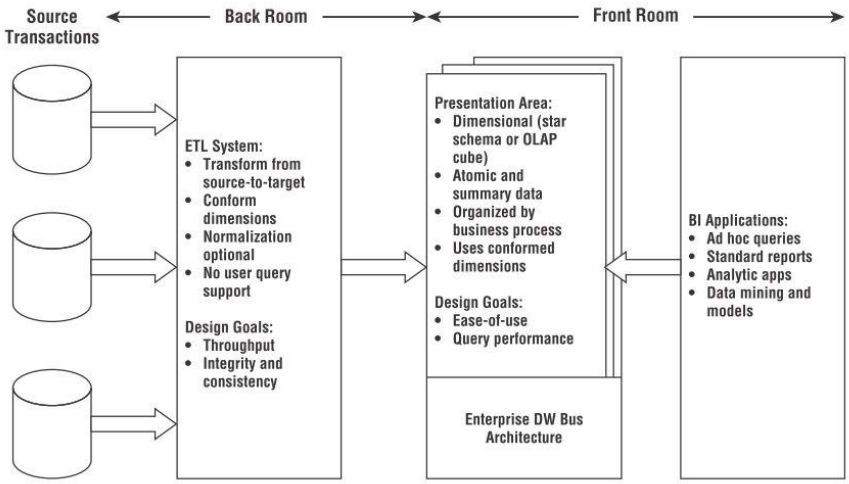
- No requiere un equipo muy grande de desarrolladores y arquitectos de datos para mantener el data warehouse (menor costo)
- Brinda funcionalidad y seguimiento de los indicadores operativos, orienta a cada data mart a brindar informes en cuanto a cada proceso de negocio
- La administración es más simplificada al estar concentrado en los procesos y áreas individuales
- La optimización de consultas es sencilla, predecible y controlable

Principales desventajas:

- Por su enfoque en procesos y áreas, puede no llegar a cubrir o manejar todos los requisitos o informes
- Constar de una menor flexibilidad de modificación

Figura 1.

Metodología multidimensional de Ralph Kimbal



Nota: La figura muestra como la arquitectura Kimball realiza el proceso de extracción, transformación y presentación de la data. *Adaptado de The Data Warehouse Toolkit (p. 55), por R Ralph Kimball and Margy Ross, 2013, Wiley*

1.5.2. Relacional (Bill Inmon)

Este es un diseño descendente, donde se construye primero el data warehouse y posteriormente los data marts, ubicando al data warehouse en el centro de la información corporativa, lo que asegura un marco lógico de los datos.

Principales ventajas:

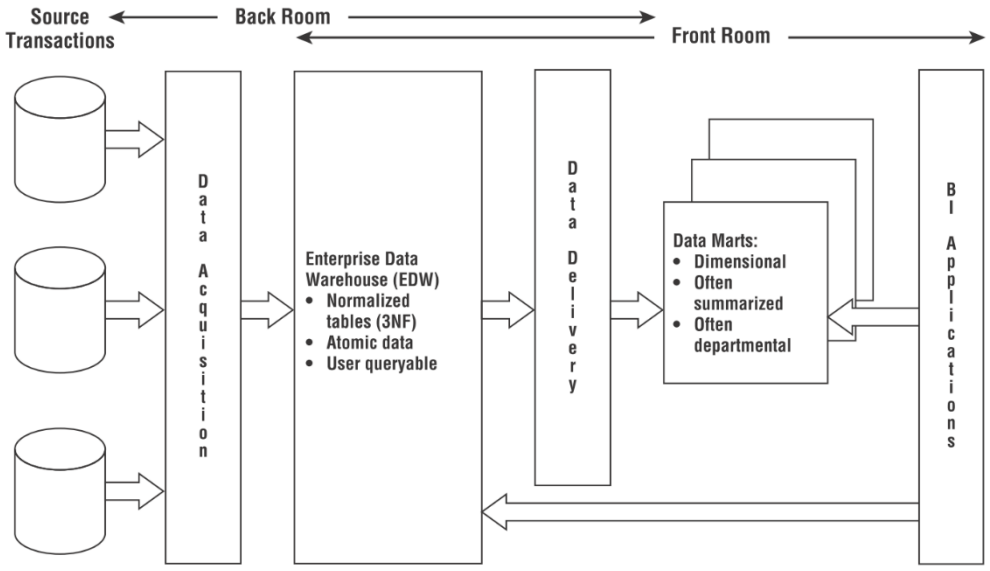
- El almacén de datos proporciona una única versión de la verdad, al ser el único origen de datos para los data marts
- Tiene una mayor facilidad de comprensión de los procesos empresariales para los usuarios

- Resulta más fácil y menos propenso al fracaso el proceso de ETL, debido a que la actualización de los datos y anomalías se evitan al contar con una redundancia muy baja
- Mayor flexibilidad a cambios por necesidades analíticas, de negocio o por fuente de datos

Principales desventajas:

- De mayor complejidad, se requieren recursos con mayor capacidad en modelado y almacenamiento
- Suele requerir de tiempos más largos para entrega de resultados

Figura 2.
Metodología de Bill Inmon



Nota: Este diagrama presenta la arquitectura de la Metodología de Bill Inmon. Adaptado de The Data Warehouse Toolkit (p. 64), por R Ralph Kimball and Margy Ross, 2013, Wiley

1.5.3. Data Vault (Dan Linstead)

Metodología híbrida que es usada principalmente cuando las empresas tienen un aumento exponencial constante de datos por lo que presentan problemas de rediseño y mantenimiento.

Esta metodología permite el almacenamiento y auditoría de la información histórica, carga paralela de datos y que, al contar con varios almacenes de datos se puede escalar sin tener que rediseñar por completo la solución. Además, proporciona flexibilidad, lo que resulta ideal para las organizaciones con un crecimiento exponencial constante.

Principales ventajas:

- Diseñado especializado para almacenar registros
- Automatiza con mayor facilidad los procesos ETL
- Fácil rastreo y auditoría de datos
- Permite varios sistemas de origen y relaciones con cambios frecuentes

Principales desventajas:

- Existe menor grado de especialización y documentación
- Puede llegar a requerir mayor esfuerzo, adaptación, y explotación de herramientas para el diseño de las capas que se requieran

1.6. Big data y cloud computing

Big data es un término que describe el gran volumen de datos – estructurados y no estructurados – que inundan una empresa todos los días. Pero no es la cantidad de

datos lo importante. Lo que importa es lo que las organizaciones hacen con los datos. El big data puede ser analizado para obtener resultados que conlleven a mejores decisiones y acciones de negocios estratégicas.

Big data también se refiere a los datos que son tan grandes, rápidos o complejos que es difícil o imposible procesarlos con los métodos tradicionales. El acto de acceder y almacenar grandes cantidades de información para la analítica ha existido desde hace mucho, pero el concepto de big data cobró impulso a principios de la década de los 2000, cuando Doug Laney articuló la definición actual de grandes datos como las 3 v.

Volumen: Las organizaciones recopilan datos de diversas fuentes, como transacciones comerciales, dispositivos inteligentes (IO), equipo industrial, vídeos, medios sociales y más. En el pasado su almacenamiento habría sido un problema, pero el almacenamiento más barato en plataformas como los data lakes y el Hadoop han aliviado la carga.

Velocidad: Con el crecimiento del Internet de las Cosas, los datos llegan a las empresas a una velocidad sin precedentes y deben ser manejados de manera oportuna. Las etiquetas RFID, los sensores y los medidores inteligentes están impulsando la necesidad de manejar estos torrentes de datos en tiempo casi real.

Variedad: Los datos se presentan en todo tipo de formatos: desde datos numéricos estructurados en bases de datos tradicionales hasta documentos de texto no estructurados, correos electrónicos, vídeos, audios, datos de teletipo y transacciones financieras.

Múltiples autores incluyen otras dos dimensiones, al hablar de big data.

Variabilidad: Además de las crecientes velocidades y variedades de datos, los flujos de datos son impredecibles, cambian a menudo y varían mucho. Es un reto, pero las empresas necesitan saber cuándo algo está de moda en los medios sociales, y cómo gestionar los picos de carga de datos diarios, estacionales y desencadenados por eventos.

Veracidad: La veracidad se refiere a la calidad de los datos. Debido a que los datos provienen de tantas fuentes diferentes, es difícil vincular, comparar, limpiar y transformar los datos a través de los sistemas. Las empresas necesitan conectar y correlacionar las relaciones, las jerarquías y los múltiples vínculos de datos. De lo contrario, sus datos pueden salirse de control rápidamente.

¿Por qué es importante el uso y estudio de big data?

La importancia de este no gira en torno a la cantidad de datos, sino en lo que se puede hacer con todos ellos. Se puede recopilar datos de cualquier fuente y analizarlos para encontrar respuestas que permiten reducir costos, reducir tiempos, desarrollar nuevos e innovadores productos, optimizar ofertas y tomar decisiones inteligentes.

1.7. Massive parallel programming con spark

Debido a limitaciones físicas, el procesador individual de las computadoras personales ha alcanzado en gran medida el techo superior para la velocidad con los diseños actuales, por lo tanto, los fabricantes de hardware agregaron más procesadores a las placas base (núcleos de CPU paralelos, que se ejecutan a la misma velocidad), pero es de mencionar que la mayoría de aplicaciones de software escritas en las últimas décadas no fueron escritas para procesamiento paralelo.

La recopilación de datos, junto con el auge de nuevas disciplinas como cloud computing y términos de la índole de big data han acaparado gran parte de los mercados transaccionales y de informática, todo esto en parte a que dispositivos relativamente baratos y de uso diario registran información, muchas veces continua, como es el caso de variables como la temperatura, sonido, velocidad. Para procesar toda esta información de manera más eficiente es que necesitamos paradigmas y formas de programación de vanguardia, como el MPP.

Un grupo de procesos informáticos es similar a un grupo de trabajadores, un equipo puede trabajar mejor y más eficientemente que un único trabajador, ellos juntan

recursos, esto significa que comparten información, desglosan las tareas y recopilan actualizaciones y resultados para obtener un único conjunto de resultados.

Para todos estos apartados, la computación en clúster y el procesamiento paralelo fue la respueta, es por eso que como producto de esto tenemos Apache Spark.

1.7.1. Apache Spark

Es un framework de computación en clúster, originalmente desarrollado por la Universidad de Carolina. El código base del proyecto Spark fue años más tarde donado a Apache, que se encarga de su mantenimiento desde entonces. Spark proporciona una interfaz para la programación de clusters completos con paralelismo de datos implicito y tolerancia a fallos.

Spark proporciona APIs para Java, Scala, Python y R, también proporciona un motor optimizado que soporta la ejecución de gráficos en general. De igual manera soporta un conjunto extenso y rico de herramientas de alto nivel, entra las que se incluyen Spark SQL (herramienta para el procesamiento de datos estructurados basada en SQL).

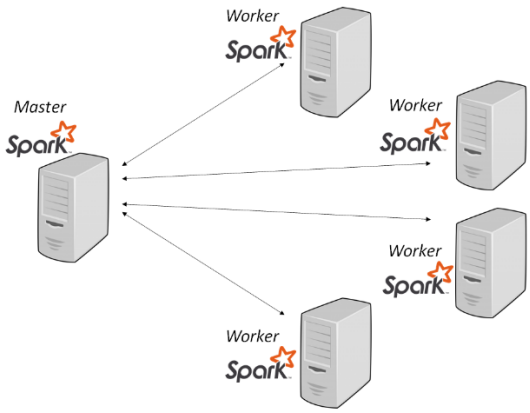
Spark tiene la base de su arquitectura en el llamado RDD (Resilient Distributed Dataset), que es un multiset de solo lectura de datos distribuidos a lo largo de un clúster de máquinas que se mantiene en un entorno tolerante a fallos.

1.7.2. Paralelismo de datos.

Es un paradigma de la programación concurrente, que subdivide cada conjunto de datos de entrada a un programa, de manera que, a cada procesador, se le asigna un subconjunto de estos datos, por lo tanto; cada procesador efectuará la misma secuencia de operaciones que los otros procesadores sobre su subconjunto asignado. En escenarios óptimos, esta manera de ejecutar las operaciones, resulta en una aceleración neta global de todo el cómputo de datos.

Es de mencionar que, el paralelismo de datos es un paradigma más que suficiente y adecuado para operaciones sobre vectores y matrices, dado que muchas de ellas consisten en aplicar la misma operación sobre cada uno de los elementos.

Figura 3.
Paralelismo en Spark



1.7.3. Lazy Evaluation.

Muchas veces denominada evaluación por necesidad, es una estrategia de evaluación que retrasa el cálculo de una expresión hasta que su valor sea necesario, y que también evitar repetir la evaluación en caso de ser necesaria en procesos posteriores. Esta compartición del cálculo puede reducir el tiempo de ejecución de ciertas funciones en forma exponencial, comparado con otros tipos de evaluación.

Beneficios de utilizar lazy evaluation:

- Evita cálculos innecesarios
- Capacidad de construir estructuras de datos potencialmente infinitas
- Capacidad de definir estructuras de control abstractas, en vez de operaciones primitivas

Normalmente, este método de evaluación es implementado encapsulando cada expresión en una función, que cuando sea computada devolverá el valor deseado, de esta manera cuando el resultado se necesite, la función recién creada será ejecutada para conseguirlo.

Este tipo de evaluación puede reducir el consumo de memoria de una aplicación, ya que los valores se crean solo cuando se necesitan, sin embargo, es difícil combinar con las operaciones típicas, debido a que el manejo de excepciones o las operaciones de entrada/salida de datos podrían quedar indeterminadas en funcionamiento.

1.8. PowerBI

Es un sistema predictivo inteligente y de gran apoyo, capaz de traducir datos ya sean simples o complejos, en gráficas, paneles o simplemente tablas cuantitativas. La gran capacidad gráfica que posee, lo ha convertido en una de las mejores herramientas de inteligencia de negocios. Es de mencionar que es gracias a PowerQuery (una de sus más grandes características), es posible extraer, transformar y cargar la información en diferentes maneras, no se trata de solamente de una herramienta visual, sino también de procesamiento.

La solución de análisis empresarial que emplea, es una tecnología basada en la nube, que permite unir diferentes fuentes de datos, analizarlos y presentar análisis de estos a través de informes y paneles, como ya se había mencionado. Con Power BI se tiene de manera fácil el acceso a datos dentro y fuera de la organización, casi en cualquier dispositivo, debido a que estos análisis pueden ser compartidos por diferentes usuarios dentro de cualquier organización.

1.8.1. Componentes.

- Power BI Desktop:
  Aplicación gratuita de escritorio para transformar, visualizar y crear informes, todo de manera local

- Power BI Service:
  Servicio online (SaaS) con funcionalidad similar a la aplicación de escritorio, con la diferencia que permite publicar informes y configurar la actualización de datos automáticamente para que el personal de la organización tenga los datos actualizados en tiempo real.

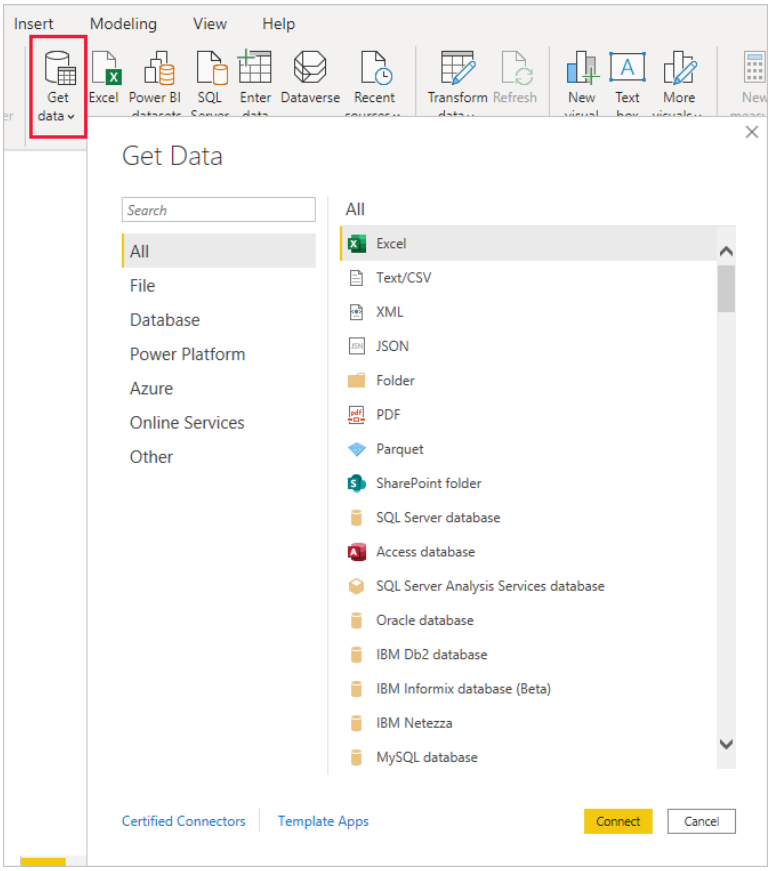
- Power BI Mobile:
  Aplicación móvil disponible para Windows, IOS y Android, con la misma funcionalidad de visualizar informes, y que se actualicen estos, en tiempo real.

1.8.2. Origen de datos

Power BI permite conectar a cientos de orígenes de datos en la nube o entorno local, creando informes con objetos integrados o creando objetos personalizados.

El acceso a los datos puede ser desde una tabla en un archivo Excel, hasta una conexión a Google Analytics, también permite conexión a base de datos on premise en la nube, como algunos servicios de Azure, lo cual facilita tener toda la información en una única visualización.

Figura 4.
Obtener origen de datos



Nota: Se puede observar que existen categorías para el origen de datos que se requiera, estos son desde archivos, bases de datos, hasta servicios online, como es el caso de algunos servicios web de Amazon.

1.8.3. ¿Por qué escoger Power BI?

Si bien, como equipo ya habíamos brindado parte de nuestra valoración de como están estructurados los contenidos para la especialización, entendemos perfectamente que Power BI como herramienta de visualización, es sumamente compleja y extensa en cuanto a posibilidades que brinda, es por eso que comprendemos el alcance que el escoger este tipo de softwares nos facilita.

En nuestro caso, no utilizaremos las bondades de Power BI para la creación de ETLs, pero si para representar visualmente los modelos dimensionales que hemos creado en nuestros escenarios. Uno de los puntos más importantes a considerar, y es de los principales del porque hacer uso de esta herramienta, es que, como aprendices en esta área de la ingeniería de datos, conocemos de la importancia que la inteligencia de negocios juega actualmente en las organizaciones y corporaciones grandes, es por esto que al interactuar desde ya con todo el proceso evolutivo de la creación de modelos dimensionales, y llegar hasta el punto de la representación gráfica, nos abre un gran espacio a muchas áreas más de estudio, que fácilmente podemos ir perfeccionando con este tipo de actividades.

CAPÍTULO 2: MARCO METODOLÓGICO.

2.1. Introducción a la lógica del negocio

Steam, que es una plataforma de distribución de contenido digital de videojuegos, fue desplegada como tal en septiembre del 2003, en sus primeras etapas solamente distribuía actualizaciones de juegos ya existentes, pero ha llegado a ampliarse al punto de incluso ofrecer juegos que no son de Valve (compañía que es la desarrolladora). Entre los principales servicios que Steam ofrece están la distribución de videojuegos, los servidores de emparejamiento, transmisiones de videos, características de redes sociales, implementación de grupos y listas de amigos, y la actualización automática de videojuegos.

Para poder disfrutar de los servicios que Steam ofrece, es necesario estar registrado mediante la creación de una cuenta gratuita, a la que se vinculan los videojuegos comprados por el jugador. Los juegos que se pueden usar en la aplicación de Steam, son todos aquellos que se ofrecen en su catálogo de compras y también aquellos que son comprados en tiendas físicas.

En el catálogo de compras hay productos bajo la modalidad de acceso anticipado que es donde permiten al jugador probar ciertos videojuegos que se encuentran en desarrollo, mejorando de esta manera la experiencia del usuario y retroalimentando con información relevante a las desarrolladoras sobre la aceptación del videojuego por parte de la comunidad. Otra de las modalidades es la Free to play que son todos aquellos videojuegos que el usuario puede jugar sin costos pero que posee mejorar su experiencia de usuario, al realizar compras o comúnmente llamadas microtransacciones.

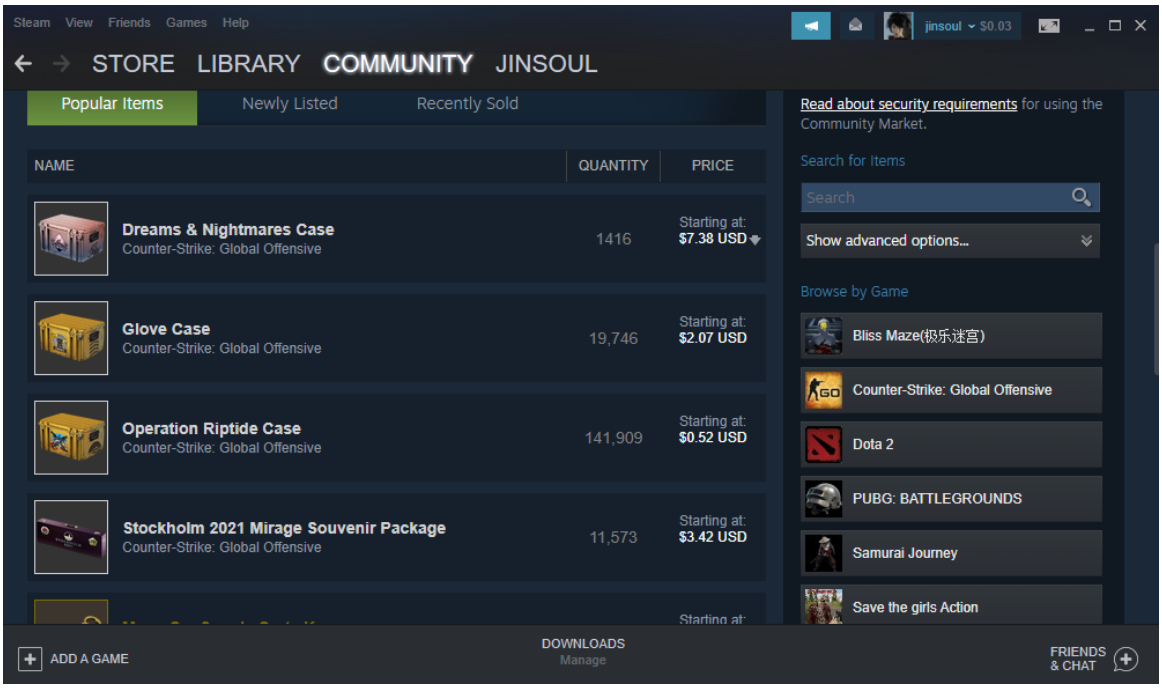
2.1.1. El mercado

El mercado de la comunidad de Steam es una sección de la plataforma que permite a los usuarios vender artículos de tipo estético correspondientes a un juego en específico.

Esta mecánica surge luego de que en juegos cuya desarrolladora es Valve (creadora de Steam) que son del género free to play, se implementara la capacidad de comprar artículos del corte de skins/tesoros/cofres a manera de microtransacciones, el usuario puede escoger entre mantener este tipo de ítems, o venderlos en el mercado a otros usuarios, importante recalcar que, al hacer uso del mercado, este agrego un impuesto extra a cada ítem que se pone a la venta.

¿Pero qué sucede con los juegos que no es posible adquirir este tipo de ítems, su acceso al mercado es limitado? No, existen algunos juegos del género “aventura”, como un ejemplo, en el que no es posible adquirir ningún tipo de skins o tesoro extra, esto, por la misma naturaleza de los juegos como tal, en estos casos Steam ofrece la posibilidad de adquirir cartas, estas cartas no están ligadas al juego y tampoco le agregan valor estético al mismo, pero ofrecen ventajas dentro de la plataforma de Steam, es decir, las recompensas de este tipo de colección se ven visualizadas en beneficios hacia Steam, y no al juego en sí; por lo que se observa que una de las principales de adquirir juegos vía Steam se obtienen muchas ventajas no solo para la fácil adquisición de objetos tipo estético a los juegos, sino bondades directas a la plataforma, con solo el hecho de usarla.

Figura 5.
Pestaña de comunidad en Steam



Nota: Captura del mercado de Steam, en su página principal.

La principal forma de ingresos que Steam posee, es la compra de los productos que se encuentran en su catálogo de videojuegos, estas compras se pueden hacer bajo muchísimos métodos de pago como tarjetas de crédito, PayPal, y se permite la inclusión de dinero a la billetera de Steam, a manera de mantenerlo como saldo a favor en la cuenta, esto en una amplia variedad de monedas. También el modelo de negocio que Steam mantiene, permite a las desarrolladoras subir sus títulos de videojuegos bajo ciertas normativas, de esta manera Steam se convierte en el intermediario con las diferentes tiendas internas que posean los dichos juegos

2.1.2. Bondades de seguridad.

Una de las principales ventajas que ofrece Steam, es que permite agregar un nivel de seguridad a las cuentas de sus usuarios, no solamente con el inicio de sesión por medio de las credenciales sino también a través de Steam Guard que es el proceso de enviar un código de acceso al correo o a la aplicación móvil cuando identifica que se está ingresando mediante un dispositivo desconocido.

Una característica importante de Steam es que nos permite acceder a datos de los videojuegos. Solo son accesibles por aquellos usuarios que han realizado compras mínimas en alguna tienda de Steam o en algunas de las tiendas internas de ciertos videojuegos.

Como toda organización que administra información de sus usuarios y por sus políticas de privacidad, la información que se comparte, tiene el consentimiento del usuario, respetando leyes tales como la ley de privacidad del consumidor (CCPA), que se encuentran vigentes en Estados Unidos en donde tienen centralizada la información.

Entre los datos que Steam comparte de sus videojuegos están las siguientes:

- Noticias para cada juego de Steam.
- Información de estadísticas globales por juego.
- Información sobre los usuarios de Steam siempre respetando la privacidad de los usuarios.
- Datos de los elementos del jugador.

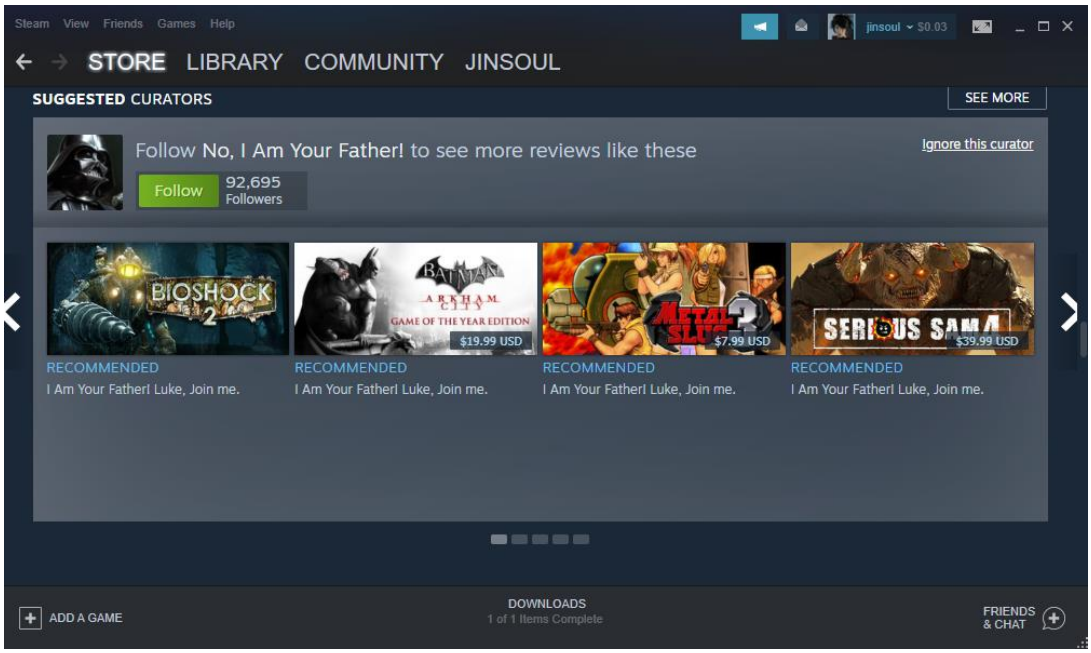
Para poder acceder a la información compartida por Steam, se debe solicitar una clave en donde el usuario podrá acceder y descargar la información en diferentes formatos de archivos.

2.2. Cualidades principales

El principal servicio de Steam es el de permitir a sus usuarios descargar juegos, y otros softwares desde su biblioteca virtual hasta sus ordenadores. Los juegos que son integrados en Steam son almacenados en el disco duro como archivos únicos no comprimidos. Los archivos .gcf (que es el formato con el que Steam guarda los archivos en disco) hace que los juegos sean más portátiles, evita que los usuarios sobreescriban en forma accidental archivos importantes, permite la modificación de recursos en forma más sencilla y permite la validación de contenido para la búsqueda de errores. Existe la posibilidad de integrar juegos a Steam, que no sean adquiridos por

este medio, en este caso los archivos son tipo .ncf e indica la existencia de un directorio que contiene varios archivos sueltos en otro lugar del sistema.

Figura 6.
Vista principal del aplicativo de Steam



Nota: Captura de pantalla de la ventana de Steam (Valve corporation, 2022)

Steam brinda una gestión digital de derechos (DRM) mínima para todo el software que se distribuye en la plataforma, utilizando “Custom Executable Generation” para archivos ejecutables que son únicos para cada usuario, pero que permiten al usuario instalar el software en varios dispositivos a través de Steam, claro esto es debido a esto que el usuario debe usar Steam mientras está conectado a internet, para validar esta operación antes de iniciar cada juego.

Uno de los hitos importantes en el crecimiento de Steam, ocurrió en Septiembre del 2008, fecha en la que se añade soporte para Steam Cloud, un sistema de computación en la nube que permite guardar de forma automática juegos y otros archivos personalizados (como imágenes, capturas de pantalla), configuraciones; permitiendo a

los usuarios acceder a esta información dese cualquier otro ordenador en el que se tenga instalado el cliente de Steam.

Steam posee un método de protección en contra del robo de cuentas, esta funcionalidad denominada Steam Guard fue implementada en marzo del 2011, y utiliza el sistema de protección de identidad que ofrecen los procesadores Intel Core de segunda generación y placas bases compatibles para permitir al usuario asociar una cuenta a un ordenador en específico; una vez asociada, la actividad de esa cuenta en otros ordenadores debe ser aprobada por el usuario. Este método de protección posee la característica, opcional es de recalcar, que es la autenticación doble a través de un método de verificación, en estos escenarios es posible la capacidad de recibir el código de verificación tanto en correo electrónico o en la app de Steam para dispositivos móviles.

Es de recordar que además de todos los beneficios asociados a títulos de juegos y seguridad en el perfil de cada jugador, Steam cuenta con características de una red social, permitiendo a los usuarios identificar amigos, y unirse a grupos a través de la comunidad de Steam. Los usuarios pueden hacer uso tanto de chats de texto, como de servicios VoIP con otros usuarios, identificar que es lo que sus amigos y miembros de grupos están realizando dentro de la plataforma, unirse e invitar a más usuarios a partidas multijugador y participar en los foros especializados de cada juego.

2.2.1. Tienda

Steam posee una tienda en web, en la cual se adquieren (compran) los juegos de ordenador en manera digital, una vez comprados los juegos, la plataforma los asocia de manera automática y permanente a la cuenta de Steam del usuario, sin embargo, es posible el regalar juegos a otras cuentas luego de la adquisición. Steam vende sus productos en una amplia cantidad de monedas, dependiendo de la región en que el usuario se encuentre, para nuestra zona Steam se encuentra por defecto en dólares estadounidenses. Otra de las características que la tienda de Steam regula por defecto,

gracias a la ubicación geográfica es la cantidad de títulos disponibles para la compra, existe la posibilidad que existan títulos que no estén disponibles para todas las regiones.

La tienda de Steam brinda la posibilidad de añadir juegos que no hayan sido adquiridos por este medio a la plataforma, esto con la finalidad de que los usuarios puedan disfrutar la interfaz de acceso de Steam con estos títulos.

Unos de los hitos importantes en el desarrollo del negocio asociado a la tienda de Steam, fue que a mediados del año 2011, se comenzaron a ofertar juegos gratuitos, esto tras el desarrollo del soporte a microtransacciones para cada canal o juego, es de esta manera que los usuarios pueden utilizar juegos que son gratis, pero adquirir productos tales como accesorios, cosméticos, cartas, medallas, a través de la tienda.

2.2.2. Servicios/facilidades adicionales

Entre los diversos servicios que posee Steam están los siguientes:

- Big Picture:
  Esto se refiere a la adaptación de la ventana principal de Steam para ser utilizada en pantallas grandes, y optimizada para el uso de gamepads

- SteamWorks:
  Interfaz de programación libre que proporciona herramientas a los desarrolladores de videojuegos, dándoles las ventajas que tiene el cliente de Steam. Proporciona diversos beneficios al integrar los juegos con Steam, incluyendo redes de juego y un sistema de autentificación para los jugadores, tanto en multiplayer como en single player.

- Steam Greenlight:
  Sistema que se basa en la ayuda de la comunidad a la hora de escoger algunos de los nuevos videojuegos independientes que tendrán su lanzamiento en Steam. Los desarrolladores publican información, capturas de pantalla, y videos de sus juegos por publicar, todo con el fin de lograr apoyo anticipado.

- Steam Workshop:
  Es la manera que la plataforma emplea para que los usuarios descarguen contenido creado por otros usuarios. Workshop nace con la finalidad de distribuir objetos para un juego en específico, pero es debido a su éxito que se comienza a utilizar con todos los juegos que poseen contenido extra.

- Steam Direct:
  Posibilidad para publicar videojuegos independientes de manera directa. Los videojuegos se envían a un equipo supervisor de Steam, pagando una tarifa de $100, si el videojuego es capaz de generar $1,000 estos $100 de tarifa son devueltos y el juego se publica a la venta.

- Steam Link:
  Dispositivo diseñado para el hogar que permite jugar a cualquier juego de Steam desde cualquier rincón de la vivienda. El dispositivo se conecta a los televisores mediante cable HDMI y reconoce cualquier ordenador corriendo Steam en la red, reconociendo y permitiendo la jugabilidad de esta manera.

2.2.3. Plataformas compatibles y OS propios

Entre las diversas plataformas que posee Steam se encuentran las siguientes:

- Steam OS:
  Distribución derivada de Arch Linux, desarrollada completamente por Valve, que es el poseedor de Steam, y funciona como sistema operativo principal de la línea de videoconsolas Steam Machines.

- Steam Machine:
  Es una combinación entre videoconsola y ordenador personal que son manufacturadas y distribuidas por Valve, todas con las especificaciones

promedio trazadas para el uso y funcionamiento correcto de la mayoría de títulos triple AAA (juegos de alta gama). Estas consolas funcionan bajo Steam OS y son totalmente libres de modificar, esto con la finalidad de que sean las máquinas ideales en cuanto a hardware para tener una experiencia completa y plena con el uso de Steam y sus elementos.

- Steam Controller:
  Gamepad desarrollado en conjunto con las Steam Machine, que funciona tanto en ordenadores como en consolas ya existentes. Su diseño está hecho principalmente para el uso de Big Picture bajo Steam OS.

- Steam VR:
  Propuesta de hardware que permite utilizar los juegos digitales bajo las librerías de Steam que se encuentran en un ordenador personal, en un ambiente de realidad virtual.

2.3. Métodos de obtención de la información

Actualmente los procesos a analizar los podemos englobar en 3 grandes grupos, el primero referente a ventas de juegos, el segundo referente a reseñas y comentarios de juegos y el tercero tratándose de tiempos invertidos en los juegos. Para todos estos procesos existe información en forma de archivos .csv o .xlsx en sitios web y foros dedicados al estudio de esta información, por lo que como grupo para la implementación de nuestros modelos dimensionales hemos clasificado la información en 2 tipos.

- Archivos preexistentes
- Archivos por construir

2.3.1. Archivos preexistentes

Estas fuentes de datos fueron obtenidas de archivos .csv que circulan en internet, es decir, estos archivos ya estaban construidos con anterioridad y muy posiblemente ya utilizados en análisis como parte de estudios con bases de análisis de datos. Uno de los ejemplos de este tipo de archivos son los encontrados en sitios web como Kaggle, que son sitios especializados en machine learning y data science.

En nuestro caso, utilizaremos 3 datasets, cerca del 70% de toda la información es obtenida por estos medios, debido a que es información referente a videojuegos y reseñas/calificaciones que estos mismos tienen.

Figura 7.
Ejemplo de un dataset que se encuentra en kaggle

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	url	types	name	id	desc_snip	recent_re	all_reviev	release_d	developer	publisher	popular_t	game_det	languages	achievem	genre	game_des	mature_c	minimum	recc
2	https://st	app	DOOM	3.49E+08	Now includes all three premiu			########	Milestone	Milestone	Action	Captions a	English, Spanish, Chi		Adventure,Strategy,Action			SO: Windc	SO:
3	https://st	app	PLAYERUN	3.53E+08	PLAYERUN	Mixed		9/4/2019	CAPCOM (	CAPCOM (	Indie	Online Mu	Korean		Casual,Simulation			SO: Windc	SO:
4	https://st	app	BATTLETE(	1.09E+08	Take command of your own me			3/5/2015	Bugbyte L	Bugbyte L	Action	Shared/Sp	English, Spanish, Chi		Casual,Simulation			SO: Windc	SO:
5	https://st	app	DayZ	53875128	The post-soviet country of Che			########	Sylwester	Sylwester	Strategy	Shared/Sp	English, Spanish, Por		Adventure,Strategy,Animation			SO: Windc	SO:
6	https://st	app	EVE Onlin	56038151	EVE Online is a community-driv			########	Jutsu Gam	Jutsu Gam	Indie	Partial Co	English, Spanish, Chi		Adventure,Strategy,Animation			SO: Windc	SO:

Nota: Este es una captura de uno de los datasets que se utiliza en este proyecto

Uno de los principales beneficios de usar este tipo de información, es que la data ya viene ordenada y disminuye el proceso respecto a conseguir la información, lo que conlleva en el proceso lógico del análisis de datos, es realizar un profiling de datos.

2.3.2. Archivos por construir

Debido a la confidencialidad de alguna información que quisiéramos procesar para realizar un análisis, tales como regiones de uso de Steam, tiempo invertido en algunos juegos y precios de compra/venta de estos mismos, es muy probable que no llegáramos a encontrar puntualmente toda la información que deseábamos al principio

de nuestro análisis, pero al tratar con Steam como plataforma, encontramos entre sus bondades la capacidad de poder utilizar su API para la extracción de información.

También generamos la información del dataset de fecha desde enero del 2015 a enero del 2022, debido a que la data de todos nuestros datasets se encuentran en esos rangos. A partir de ello haciendo uso de Excel, se extrajo diferentes formatos de fechas tales como la fecha en formato inglés, así como columnas columnas que contenían solo el día, solo el mes, el año y así como los días festivos en base. Para los días festivos se consideró solo aquellos que poseen asueto en el país.

Otro dataset que se construyó contiene la información de países con sus respectivos valores internacionales como es el ISO 3166-1 y el ISO 3166-1 alfa-3, que son datos encontrados en algunos datasets.

Tanto el dataset de fecha como el de países, se exportaron a formato csv desde Excel

Métodos que utilizamos para la extracción de información

De entre la amplia cantidad de métodos que la API proporciona y que brindan información acerca de los usuarios y los tiempos que estos manejan en los juegos, hemos escogido dos.

- GetPlayerSummaries
- GetRecentlyPlayedGames

Los parámetros recibidos por ambos métodos son una lista de Steam id’s (identificadores únicos para cada jugador) y cada método responde con diferentes resultados, según sea el caso, y esto es en formato json.

Para el caso de GetPlayerSummaries, la información recibida como respuesta, en caso de que la información de los usuarios sea pública es: Steam id, persona name, profile url, avatar, avatar médium, avatarfull, personastate, communityvisibilitystate, profilestate, lastlogoff, commentpermission, numbergamesowned, numberreviews.

En el caso, en esos id’s se encuente información privada, pero que el usuario ha permitido su publicación, también se desplegará esta información.

Figura 8.
API GetPlayerSummaries

Nota: Se muestra los datos que Steam comparte como información privada (Valve corporation, 2022)

La única dificultad que tenemos en este punto es, encontrar n cantidad de id’s, debido a que la mayoría de las veces en los datasets que encontramos, los id’s no representa información que aporte valor a los análisis.

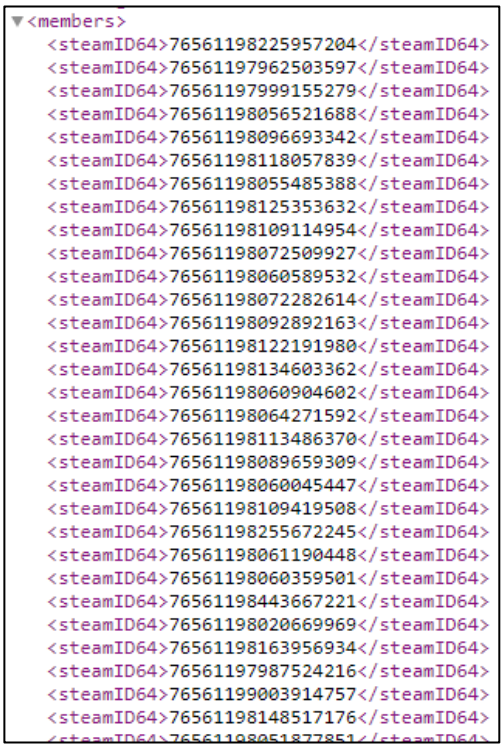
En Steam hay que recordar que existen grupos de las comunidades de jugadores. Estos son creados por los usuarios donde los miembros pueden compartir intereses, coordinar actividades de juego y organizar discusiones en foros específicos del grupo.

Por ello los grupos de Steam solucionan para nosotros este inconveniente de no tener id’s, debido a que es posible la extracción de la lista de miembros de una comunidad en formato xml, siempre y cuando el grupo sea público. Para realizar el proceso de extracción de id’s en la API, consideramos extraer al menos 100,000 id’s.

La estructura de una URL para conocer los miembros que forman parte de un grupo es la siguiente:

Steamcommunity.com/groups/**NOMBRE_DE_GRUPO**/memberslistxml/?xml=1

Figura 9
Resultados de la consulta en la API



Nota: Este resultado se genera al consultar id`s por comunidades de grupo de usuarios

Este es el proceso que realizamos para obtener los id’s a utilizar, para recuperar resultados de la API de Steam en los métodos con la información que necesitamos

Para la extracción y transformación de esta información a formato .csv, por conveniencia, utilizamos un pequeño script hecho en Python que realiza las consultas a la API, donde luego almacena la información en formato JSON para finalmente transformar el resultado a un archivo .csv, ya que decidimos estandarizar el formato para cada uno de los datasets a implementar.

Figura 10.
Segmento de código utilizado para realizar ETL desde la API

```
from urllib.request import urlopen

import json
import csv
import pandas as pd
from pandas.io.json import json_normalize

#ALMACENAR URLS
url = "https://api.steampowered.com/ISteamUser/GetPlayerSummaries/v0002/?key=CA486
url2 = "https://api.steampowered.com/ISteamUser/GetPlayerSummaries/v0002/?key=CA48
url3 = "https://api.steampowered.com/ISteamUser/GetPlayerSummaries/v0002/?key=CA48

#GUARDAR RESPUESTA DE CADA URL
response = urlopen(url)
response2 = urlopen(url2)
response3 = urlopen(url3)

#CONVERTIR A JSON CADA RESPUESTA
j1 = json.loads(response.read())
j2 = json.loads(response2.read())
j3 = json.loads(response3.read())

#GUARDAR ARCHIVOS
with open('j1','w') as file:
    json.dump(j1, file)

with open('j2','w') as file:
    json.dump(j2,file)

with open('j3','w') as file:
    json.dump(j3,file)

#convertir-----------------------------------------------

with open('j1') as file:
```

2.4. Descripción de los dataset y diccionario de datos del dataset

A continuación, se describirán los diferentes datasets que se utilizaron en la presente tesina.

Tabla 2.
Breve resumen de los datasets

Datasets	**Nombre estandarizado**	**Nombre original del archivo**	**Tamaño**
Primer Dataset de datos de videojuegos	DATASET1_fixed.csv	steam_games.csv	82 MB

Datasets	Nombre estandarizado	Nombre original del archivo	Tamaño
Segundo Dataset de ventas de videojuegos	DATASET2_fixed.csv	Steam_games.csv	2 MB
Tercer Dataset de reseñas de videojuegos	DATASET3_fixed.csv	steam_reviews.csv	8.17 GB
Cuarto Dataset “GetPlayerSummaries” (API)	DATASET4_fixed.csv		114.9 MB
Quinto Dataset “GetRecentlyPlayedGames” (API)	DATASET5_fixed.csv		10.2 MB
Sexto Dataset Fechas			
Septimo Dataset			

Primer Dataset de datos de videojuegos

Contenido:

Este dataset contiene información descriptiva de videojuegos encontrados en la tienda digital de Steam; cuenta con alrededor de 40,803 videojuegos registrados en la plataforma y tiene como origen uno de los archivos preexistentes.

Tabla 3.
Dataset de ventas de videojuegos

Atributo	Descripción
url	URL de un juego. Esta columna también contiene información adicional, como el ID del juego, el nombre del juego en texto plano
Types	Tipo de paquete – contiene información referente si es un videojuego individual o un paquete
Name	Nombre del videojuego registrado en la plataforma
desc_snippet	Breve descripción del videojuego
recent_reviews	Estadísticas de las reseñas recientes o de los últimos 30 días
all_reviews	Estadísticas de todas las reseñas hechas al videojuego
release_date	Fecha de lanzamiento del videojuego al mercado
Developer	Empresa desarrollador del videojuego
Publisher	Empresa editor o editores del videojuego
game_details	Detalles adicionales al género del videojuego
Languages	Idiomas admitidos por el videojuego
Achievements	Número de logros
Genre	Género (s) del videojuego
game_description	Descripción del videojuego
mature_content	Descripción de contenido para adultos en el videojuego
minimum_requirements	Especificaciones mínimas para el videojuego, este requiere segmentación de los campos relevantes de los requerimientos

Atributo	Descripción
recommended_requirements	Especificaciones recomendadas para el videojuego, este requiere segmentación de los campos relevantes de los requerimientos
original_price	Precio sin descuento del videojuego
discount_price	Precio con descuento del videojuego si lo tuviese

Notas: Adaptado de Steam Sale, por Jayant Jain, 2021, Kaggle (https://www.kaggle.com/trolukovich/Steam-games-complete-dataset)

Segundo Dataset de ventas de videojuegos

Contenido:

Este Dataset contiene información de las transacciones de ventas de videojuegos, tiene alrededor de 43,423 registros y también tiene como origen uno de los archivos preexistentes de información relacionada a ventas de Steam.

Tabla 4.
Dataset de ventas de videojuegos

Atributo	Descripción
#	Número de registro
name	Nombre del videojuego
id_juego	ID del juego dentro de Steam
Id_jugador	ID del jugador que realizó la transacción
rel_date	Fecha de transacción
orig_price	Precio original
discounted_price	Precio con descuento
discount%	Porcentaje de descuento

Notas: Adaptado de Steam games complete dataset, por Alexander Antonov, 2021, Kaggle (https://www.kaggle.com/xybervenom/Steam-sale)

Tercer Dataset de reseñas de videojuegos

Contenido:

Este Dataset contiene información de reseña de usuarios a los videojuegos registradas el año 2021, este dataset también tiene como origen uno de los archivos con información preexistente.

Tabla 5.
Dataset de reseñas de videojuegos

Atributo	Descripción
#	Número de registro
App_id	ID del videojuego
app_name	Nombre de la aplicación o videojuego
review_id	Id de la reseña
language	Lenguaje de la reseña
review	Contenido de la reseña
timestamp_created	Marca de tiempo de creación de la reseña
timestamp_updated	Revisar la marca de tiempo de la última actualización
recommended	Recomienda el juego
votes_helpful	La cantidad de otros usuarios que encontraron útil esta revisión
votes_funny	La cantidad de otros usuarios que encontraron divertida esta revisión
weighted_vote_score	Puntuación basada en el número de votos útiles
comment_count	Numero de comentarios de la reseña

Atributo	Descripción
Steam_purchase	El autor de la reseña que compro la aplicación o el videojuego
received_for_free	Si el autor de la reseña recibió la aplicación de forma gratuita
written_during_early_ac	Si la revisión se escribió durante el acceso anticipado
author.Steamid	ID de Steam del autor de la reseña
author.num_games_own	Número de juegos que posee el autor de la reseña
author.num_reviews	Número de reseñas de aplicaciones en el tiempo del autor
author.playtime_forever	Tiempo de uso de la aplicación o juego del autor de la reseña
author.playtime_last_tw	Tiempo de reproducción de la aplicación o juego del autor de la reseña en las últimas dos semanas
author.playtime_at_revi	Tiempo de reproducción de la aplicación o juego del autor de la reseña en el momento de la reseña

Notas: Adaptado de Steam Reviews Dataset 2021, por Marko M., 2021, Kaggle (https://www.kaggle.com/najzeko/Steam-reviews-2021)

Cuarto Dataset “GetPlayerSummaries” (información obtenida de la api de Steam)

Contenido:

Este dataset contiene la información de los perfiles de los jugadores, donde cada atributo está relacionado a su respectivo ID. Este Id se obtuvo de la comunidad de Steam

Tabla 6.
Dataset GetPlayerSummaries

Atributo	Descripción
Steamid	Usuario de Steam del jugador

Atributo	Descripción
Profileurl	URL completa del usuario
Lastlogoff	Última vez en línea del usuario, en formato unix
Timecreated	Fecha de creación del perfil
Gameextrainfo	Juego relevante en el perfil del usuario
gameid	Código del juego relevante en el perfil
countrycode	Código de país, si está configurado
continent	Continente al que pertenece el país del registro
commentpermission	Booleano que determina si se puede o no comentar en su perfil
num_games	Cantidad de juegos que posee dicho usuario
num_reviews	Cantidad de reseñas que ha publicado dicho usuario

Notas: Adaptado de Steam.API, 2022, Steam (http://api.Steampowered.com/ISteamUser/GetPlayerSummaries/v0002/?key=XXXXXXXXXXXXXXXXXXXXXXXXX&Steamids=76561197960435530)

Quinto Dataset “GetRecentlyPlayedGames” (información obtenida de la API de Steam)

Contenido:

Este dataset contiene información referente a las aplicaciones/juegos que cada jugador ha creado. Se usa el término crear debido a que es el usuario en calidad de jugador, genera estas métricas por cada juego. Estas métricas están asociadas a las fechas y horas invertidas en cada juego, es posible también conocer el sistema operativo en el que el usuario interactúa con Steam desde este dataset.

Tabla 7.
Dataset GetRecentlyPlayedGames

Atributo	Descripción
Steamid	Usuario de Steam del jugador
Gameplayed	Nombre del juego que se ejecutó
Appid	ID del juego que se ejecutó
Playtimeforever	Cantidad de tiempo, en minutos que se ejecutaron
Date_played	Fecha en la que se jugó
Plattform	Sistema operativo en el que se jugó

Notas: Adaptado de Steam.API, 2022, Steam (http://api.Steampowered.com/IPlayerService/GetRecentlyPlayedGames/v0001/?key=XXXXXXXXXXXXXXXXXX&Steamid=76561197960434622&format=json)

Sexto Dataset Fechas

Contenido:

Este dataset contiene información referente a las fechas. El rango que abarca es de enero de 2015 a enero de 2022.

Tabla 8. Dataset de fechas

Atributo	Descripción
Id	Contiene el valor sin guiones de la fecha en formato AAAAMMDD
Fecha1	Fecha en formato DD-MM-AAAA
Fecha2	Fecha en formato AAAA-MM-DD
Fecha3	Fecha en formato inglés, ejemplo "Ene 1, 2015"
Semana	Día de la semana
AA	Valor del año, ejemplo: 2015
MM	Valor numérico del mes, ejemplo para enero: 01
DD	Valor numérico del día, ejemplo de día: 5

Atributo	Descripción
NombreMM	Nombre del mes
NombreCortoMM	Nombre del mes en tres letras, ejemplo: ene
NombreDD	Nombre del día
NombreCortoDD	Nombre del día en tres letras
EsFestivo	Contiene el nombre de la festividad
DiaFestivo	Contiene el texto SI o NO de acuerdo a si es día festivo

Notas: Dataset generado con Excel

Séptimo Dataset CodigoPais

Contenido:

Este dataset contiene información referente a los códigos de país.

Tabla 9.
Dataset de CodigoPais

Atributo	Descripción
ISO_3166-10	Sistema de códigos de dos letras
País	Nombre del país
ISO_3166-12	Sistema de tres dígitos
ISO_3166-13	Sistema de códigos de tres letras

Notas: ISO 3166-1 como parte del estándar ISO 3166 proporciona códigos para los nombres de países y otras dependencias administrativas. (Mucattu Utils, s.f.)

2.5. Resultados de data Profiling

Para realizar el perfilado de los datos se utilizaron las herramientas DataCleaner que nos ayudó a tener un análisis general del dataset y sus métricas en los campos de interés. Los dataset por su volumen de datos utilizamos Excel, separando en columnas

independientes cada campo del dataset que se encontrada separado por comas y aplicando filtros para poder tener un panorama más específico de los campos de interés. Esto se realizó para ver los tipos de valores que los componen y el porcentaje de estos valores para reforzar lo mostrado por DataCleaner

La herramienta DataCleaner nos permite tener los mínimos y máximos caracteres en los campos de interés y tener una idea de la longitud de los atributos en las tablas del modelo dimensional a construir

De manera complementaria se utilizó la herramienta Power BI, nos permitió tener un análisis adicional de los datos, mostrando los errores en los datos, datos nulos, datos vacíos y duplicidad de datos en campos donde se requería datos únicos

Hallazgos encontrados por dataset:

Dataset de videojuegos (DATASET1_fixed)

- El atributo “id” que representa el identificador del juego se encontraron datos erróneos con valor “#N/A”.

- El atributo “id” que representa el identificador del juego se encontraron valores duplicados

- El atributo “reléase_date” que representa la fecha de publicación del videojuego contiene formatos de fechas con día, mes, año, también presenta fechas de publicación a futuro, lo cual no es un error sino que son videojuegos registrados en la plataforma pero aún no se han lanzado oficialmente y esa es la fecha estimada de publicación

- El atributo “lenguages” es un atributo multi valuado que representa los idiomas que el videojuego soporta y deberá ser separado en valores individuales en el ETL

- El atributo “genre” es un atributo multi valuado que representa los generos en los que el videojuego puede ser clasificado y deberá ser separado en valores individuales en el ETL

- Se encontró un alto grado de nulidad en el campo de descripción del videojuego, lo cual no afecta al modelo dimensional debido a que no usaremos ese atributo

- El atributo Publisher se encontró un porcentaje del 12% de nulidad esto es debido a que en ocasiones el publicador es el mismo que el desarrollador, esto será corregido en el ETL

- Los atributos de “minimum_requirements” y “recommended_requirements” representan los requerimientos mínimos y requerimientos recomendados, contienen información que deberá ser segmentada, estos se separarán en diferentes campos en las dimensiones para tener campos más atómicos y permitir un mejor análisis

- Limpieza en los encabezados de cada uno de los campos de requerimiento después de ser separados de esta forma tener el valor atómico “SO:” para sistema operativo, “RAM:” para la memoria RAM, “Graphics:” para tarjeta grafica, “Procesador:” para procesador

Dataset ventas de videojuegos (DATASET2_fixed)

- No se encontraron valores nulos en los registros de ventas en ninguno de sus campos

- La fecha de la transacción contiene día/mes /año y su formato día-mes-año, se unificará a un solo formato mediante ETL

- Los atributos precio, precio con descuento y porcentaje de descuento se encuentran en valores numéricos sin adición de caracteres

- El atributo “discounted_price” contiene el precio de venta con el descuento aplicado, los caso con valor “0.0” representan los juegos gratuitos

- El atributo “id juego” que representa el identificador del juego se encontraron datos erróneos con valor “#N/A”.

Dataset reseñas de videojuegos (DATASET3_fixed)

- No se encontraron valores nulos en los registros de reseñas en ninguno de sus campos

- El atributo “id_App” que representa el identificador del juego se encontraron datos erróneos con valor “#N/A”.

- La fecha de creación de la reseña y la fecha de su última actualización se encuentra en formato DD/MM/AA, esta deberá convertirse en el ETL a un formato DD-MM-AA para mantener el estándar con la dimensión fecha

- La recomendación se encuentra en true o false, esta se deberá convertir a un formato más textual para un mayor entendimiento de los usuarios analíticos

- El método de acceso al juego se encuentra en true o false. Estos datos indican si el juego fue comprado(steam_purchase), si fue acceso anticipado (written_during_early_access) o si fue gratis(received_for_free). Se convertirá en el ETL, a un formato más textual para un mayor entendimiento de los usuarios analíticos

- Los tiempos de juegos están en formato numérico y en unidades de tiempo en minutos. Estos campos son author.playtime_forev (tiempo de juego total del juego por el usuario), author.playtime_last_tw (tiempo de juego por el usuario en las últimas dos semanas), author.playtime_at_revie (tiempo de juego al momento de hacer la reseña)

Dataset tiempo de juego de videojuegos (DATASET4_fixed)

- El atributo “appid” que representa el identificador del juego se encontraron datos erróneos con valor “#N/A”.

- El atributo “date_played” representa la fecha de juego, se encuentra en formato DD/MM/AA, esta deberá convertirse en el ETL a un formato DD-MM-AA para mantener el estándar con la dimensión fecha

- No se encontraron valores nulos

Dataset de jugador (DATASET5_fixed)

- No se encontraron valores nulos

- El atributo “num_game” contiene el numero de juegos en la cuenta del jugador se encontraron valores erróneos con el valos “#N/A”

- El atributo “countrycode” contiene el país de origen del jugador, se encuentra en un formato conpuesto con los códigos de país de la forma “ISO 3166-12/ ISO 3166-12” esto se cambiará por el nombre de país

Dataset de fecha y código de país

Estos son dataset creados por el equipo de trabajo, por lo que todos los datos se encuentran limpios y listos para usar

CAPÍTULO 3: MARCO PROPOSITIVO.

3.1 Definición del proyecto

Para el desarrollo de este Data Warehouse se identificaron los procesos de negocio dentro de la plataforma Steam que representan un alto grado de interés por la información que brindan para la toma de decisión

Por lo que se definió un proyecto dirigido a esas áreas interesadas en el comportamiento de los videojuegos como producto dentro de la plataforma Steam

3.2 Alcances y justificación del proyecto

El alcance de este proyecto es el desarrollo de un datawarehouse, el cual mantendrá información de análisis sobre la venta de videojuegos, la percepción de los usuarios de los videojuegos mediante reseñas y el tiempo de uso de los videojuegos, con la finalidad de presentar información útil para el soporte de decisiones

3.3 Definición de requerimientos de negocio

Para la definición de los requerimientos el equipo de trabajo realizo ciertas actividades que ayudaron a determinar las necesidades de los usuarios finales del datawarehouse las cuales son:

- Definir los usuarios finales, para esta actividad se decidió que el modelo serio orientado a suplir las necesidades de personas o entidades que interactúan con la plataforma con fines de lucro como pueden ser desarrolladoras o publicadoras de videojuegos, creadores de contenido, distribuidores de servicio, entre otros interesados en la interacción de los usuarios con los videojuegos como producto
- Estudiar el modelo de negocio de la plataforma Steam, con el fin de entender la necesidad de los usuarios
- Estudiar y analizar la información que se tiene a la disposición para desarrollar el modelo dimensional del datawarehouse

Tomando en cuenta los puntos anteriores se plantearon los requerimientos que se mencionan a continuación:

Ventas de videojuegos

- Obtener montos totales de ventas en función de año, mes o día
- Visualizar el monto de descuento otorgado en la venta de videojuegos
- Identificar periodos con mayor flujo de ventas, esto puede ser meses con mayores ventas, también determinar si se producen mayores ventas días de semanas, fines de semana o días feriados
- Determinar categoría de videojuegos más vendidas
- Conocer cantidad de ventas de acuerdo requerimientos de videojuego: procesador, RAM, almacenamiento, gráficos
- Determinar Idiomas en los videojuegos más vendidos
- Determinar las ventas por región de acuerdo a la procedencia del usuario

Reseñas de usuarios

- Presentar Videojuegos con mayor número de reseñas
- Determinar puntuación dadas en reseñas por genero
- Realizar recuento de reseñas positivas o negativas (recomendar o no videojuego) por género o videojuego
- Identificar el tiempo de uso antes de realizar la reseña y tiempo total de uso del videojuego
- Identificar videojuegos con mayor número de reseñas, comentarios, votos útiles y votos divertidos en reseñas
- Determinar la cantidad de reseñas por idioma
- Identificar regiones con mayor número de reseñas, de manera general, positivas o negativas
- Determinar cantidad de reseñas según su método de acceso a un videojuego en específico, este puede ser comprado, acceso anticipado o de forma gratuita
- Identificar periodos con mayor flujo de reseñas

Tiempo de juego

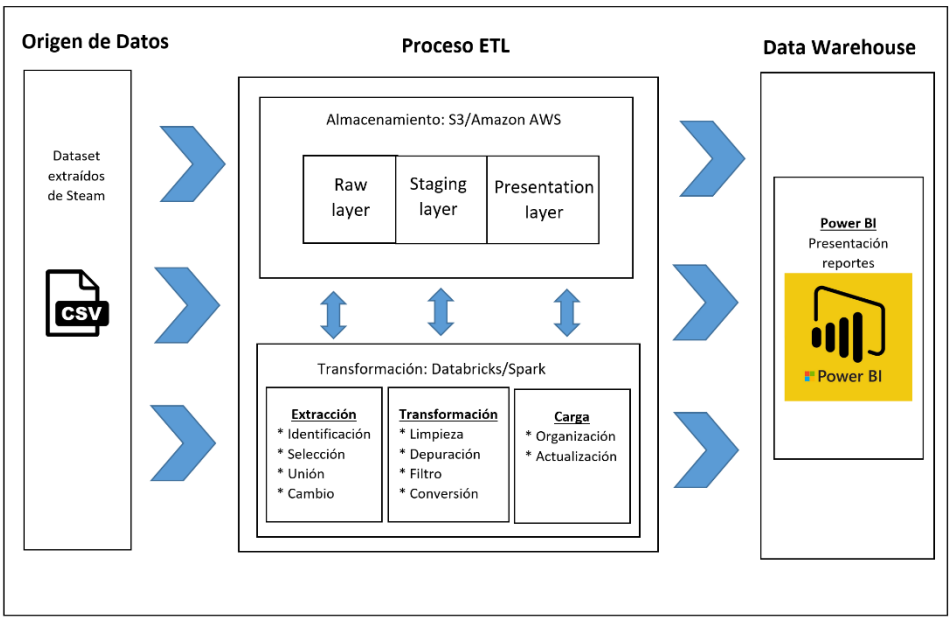
- Identificar promedio de horas diarias que los jugadores dedican a los videojuegos
- Determinar periodos con mayor flujo de jugadores con el fin de planificar el aumento de requerimientos de servidores en los videojuegos
- Visualizar tiempos promedio de uso de videojuegos, por periodo de tiempo: día, mes, trimestre, año, o día feriado
- Visualizar tiempos de uso en videojuegos por región
- Determinar tiempos de uso en horas de juego por categorías
- Comparar flujo de uso de un videojuego entre su uso máximo registra y su mínimo
- Determinar regiones con mayor flujo de tiempo de juego
- Identificar sistemas operativos con mayor tiempo de juego
- Determinar idiomas de juego con mayor tiempo de uso

3.4 Diseño técnico de la arquitectura

La fuente de datos será tomada de la plataforma Steam en dataset formato CSV, estas fuentes de datos serán almacenados en S3/AWS en un Bucket que estará dividido en 3 particiones raw later, staging layer y presentation layer, los datos iniciales sin transformar serán colocados en la partición staging layer donde posteriormente serán tomados desde Databricks y se harán las debidas transformaciones con el motor de Spark, las capas staging layer y presentation layer guardaran data semi transformada y completamente transformada, los datos ya procesados se colocaran en Databricks con el motor de Redshift para ser alojada, de donde será tomada por Power BI para la creación de los reportes de presentación

Figura 11.

Arquitectura de la solución



Nota: Este diagrama expresa visualmente los pasos a seguir para realizar la transformación de la data que se encuentra en los datasets

3.5 Selección de productos y plataformas de trabajo

Para el desarrollo del datawarehouse con la arquitectura planteada se seleccionaron plataformas y herramientas de desarrollo que cumplieran las características necesarias y al alcance del equipo de trabajo

Capa de Almacenamiento (Raw layer)

Para el almacenamiento se utilizó la plataforma de Amazon AWS con su servicio Amazon Simple Storage Service (Amazon S3), el cual es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento, permitiendo organizar los datos y configurar los controles de acceso precisos con objeto de satisfacer requisitos empresariales, organizativos y de conformidad específica con la herramienta AWS Identity and Access Management (IAM).

Capa de procesamiento – ETL (Staging layer)

Para la limpieza y transformación de los datos se utilizaron los servicios en la nube de la plataforma Databricks basada en Spark, las transformaciones se realizaron el lenguaje Scala

Capa de Presentación (Presentation layer)

Para la presentación de los datos y reportes analíticos se utilizó Power BI en su versión desktop, que permite unir diferentes fuentes de datos y presentarlos mediante informes y paneles, permitiendo presentar los modelos dimensionales del datawarehouse y las consultas analíticas de este.

3.6 Estrategias ETL

Para el modelo dimensional planteado en este trabajo se tiene dos escenarios posibles con los cuales se establecerán las estrategias de ETL

- El primero es el caso puntual de este trabajo donde se obtuvo la Data con la cual se trabajó se le aplico el proceso de ETL y se construyó los modelos dimensionales que conforman el Data warehouse, por lo que siguiendo la teoría de Slowly changing dimensions, se tomaran las dimensiones como SCD Tipo 0 donde los valores son estáticos y no cambian.
  Aquí se implementó una carga Full de los datos,
  - aplicando los procesos de transformación de los datos y llenado de dimensiones
  - luego el llenado de dimensiones estáticas como dimensión Fecha que no requiere ETL
  - carga full en fact tables con su respectivo ETL

- El segundo caso es la puesta del Data warehouse en producción y su mantenimiento en el tiempo, para este caso recomendamos la extracción de datos por región, en horarios donde cada una de las regiones tengan poco flujo de actividad preferiblemente de madrugada con el fin de no impactar en el rendimiento de los sistemas transaccionales al momento de extracción de datos, en este caso se implementarán las dimensiones como SCD Tipo 2 donde al existir cambios en los valores se adiciona una nueva fila en la dimensión manteniendo la integridad de los datos históricos
  En este caso se realizarán las cargas incrementales en dimensiones y fact table

3.7 Limitaciones operativas

I. Espacio de almacenamiento para el buen funcionamiento del servidor
   Al implementar herramientas tipo PAAS tales como AWS S3 y Databricks, no se requiere un almacenamiento para esos aplicativos porque se está contratando un producto

II. El espacio requerido para almacenar los datasets
   Para efectos de esta tesina se tiene 8.38 GB de espacio requerido para almacenar la data. Los cálculos se encuentran en el anexo 3

III. Datos de procesados en las diversas capas Staging
   Estimando que cuando haremos el procesamiento y la limpieza de los diversos datasets, eliminaremos información, se considerará que debemos tener disponibles 16GB que es aproximadamente el doble del espacio de almacenamiento de los datasets.

IV. Datos finales totales
   A manera global debemos de tener como mínimo 24.38 GB, para procesar la data que poseemos almacenada en los servidores.

3.8 Modelo Dimensional y mapeo de datos

Los modelos dimensionales que conforman el datawarehouse fueron realizados siguiendo el esquema estrella que permite mejorar el rendimiento en las consultas y la optimización en los tiempos de respuesta.

En el modelo cuenta con las siguientes Fact table

- FactTableVentas
- FactTableReseña
- FactTableTiempoJuego

Se analizaron las siguientes dimensiones

- DimensionVideoJuego
- DimensionGenero
- DimensionIdioma
- DimensionJugador
- DimensionFecha
- DimensionContextoReseña

Se usaron dimensiones tipo Bridge para agrupar valores entre dimensiones con múltiples valores

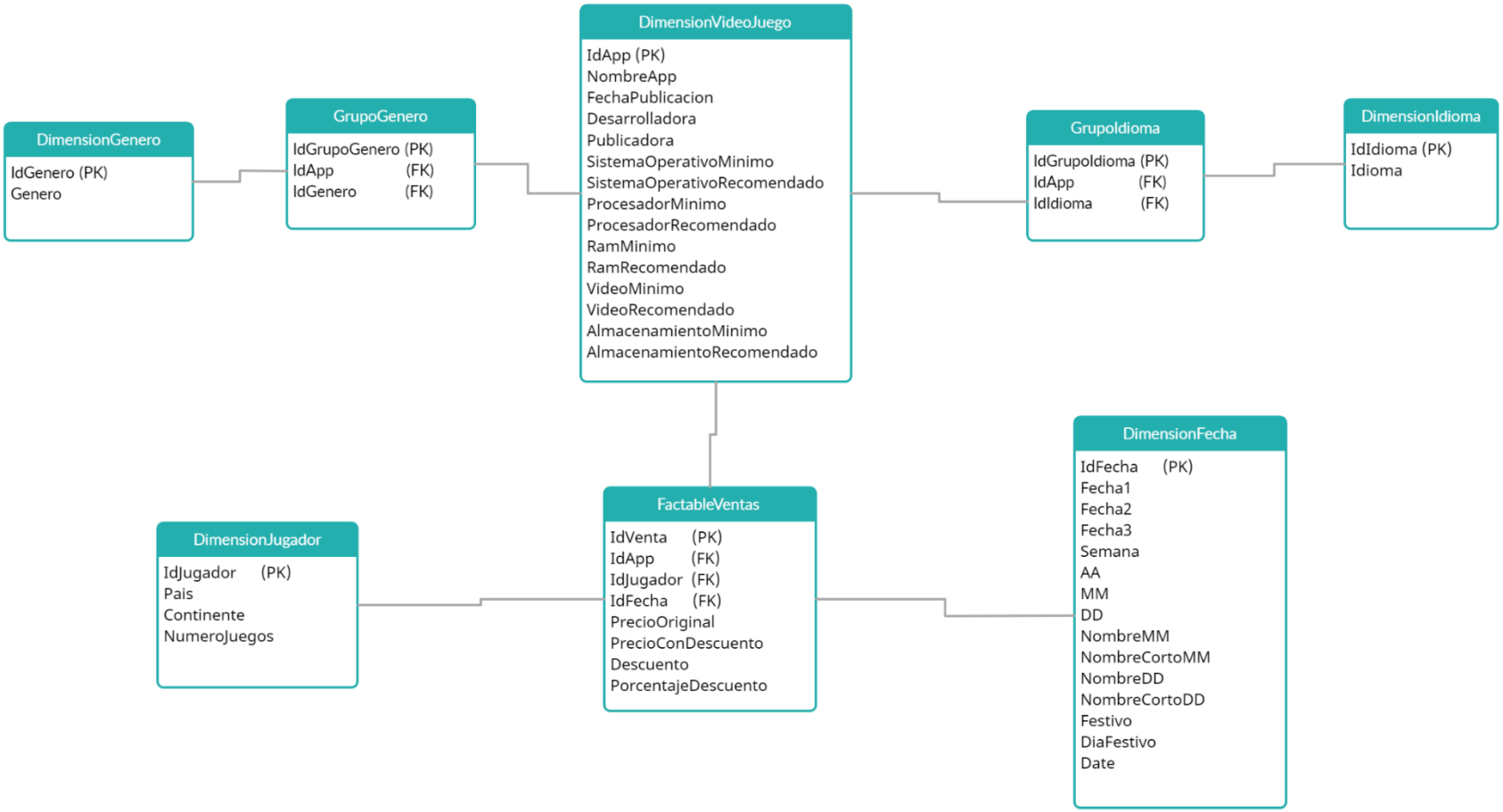
- GrupoGenero
- GrupoIdioma

A continuación, se presenta los diagramas de los modelos dimensionales junto a su descripción y mapeo de datos

3.8.1 Modelo dimensional Venta de Video juegos

Figura 12.

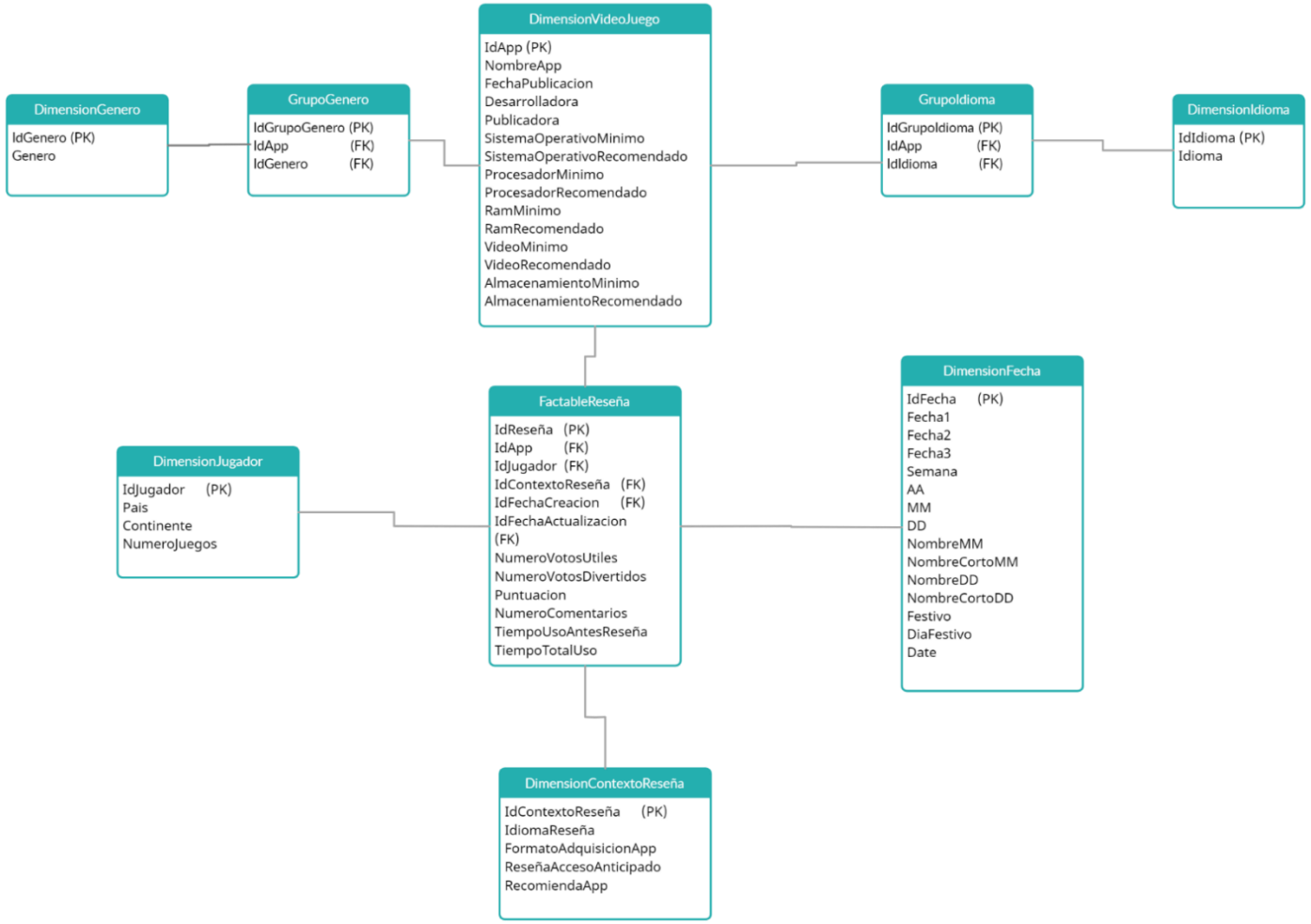
Ventas de videojuegos



Nota: La granularidad de cada fila en la Fact table, representa la transacción de Venta de videojuegos de manera individual

Figura 13.

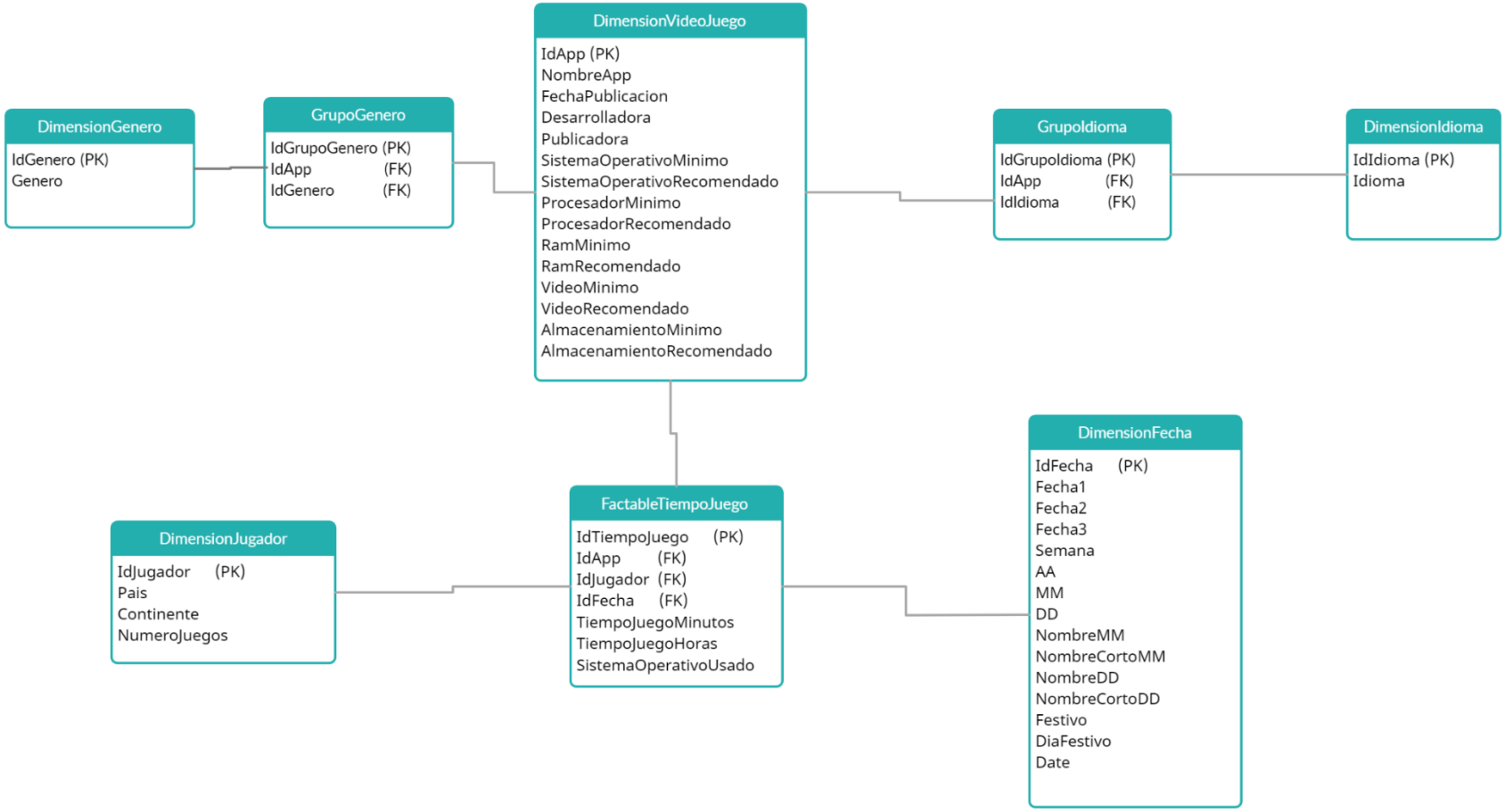
Reseña de videojuegos



Nota: Cada fila en la Fact table representa el Ingreso de una reseña individual por jugador para un videojuego especifico

Figura 14.

Modelo dimensional Tiempo de Video juegos



Nota: Cada fila en la fact table representa las horas jugadas para un cierto día, para un juego en específico.

3.8.2 Fact Table Ventas

Nombre de la tabla: Fact tableVentas

Tipo de tabla: Fact table / Transaccional

Tabla 10.

Descripción de la Fact table Ventas

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
IdVenta	Numérico (integer)	Clave primaria de ventas, identifica la transacción de ventas	Primaria	
IdApp	Alfanumérico (string)	Clave foránea de Videojuegos, identifica videojuego vendido en la transacción	Foránea	ETL de DATASET2_fixed
IdJugador	Numérico (long)	Clave foránea de jugador, identifica al jugador que realizo la compra	Foránea	ETL de DATASET2_fixed
IdFecha	Numérico (integer)	Clave foránea de fecha, identifica la fecha de la transacción	Foránea	ETL de DATASET2_fixed
PrecioOriginal	Numérico (doublé)	Precio normal del videojuego en la tienda sin descuento		ETL de DATASET2_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
PrecioConDescuento	Numérico (doublé)	Precio con descuento del videojuego, de no existente se refleja el precio normal del juego; también representa el precio de venta en la transacción		ETL de DATASET2_fixed
Descuento	Numérico (doublé)	Descuento que se otorga en la transacción respecto al precio normal de venta		Campo calculado
PorcentajeDescuento	Numérico (doublé)	Porcentaje de descuento dado respecto al precio normal		ETL de DATASET2_fixed

Nota: Contiene información de la transacción de venta de videojuegos su nivel de detalle es de venta de videojuego individual por cada registro

3.8.3 Factable Tiempo de Juego

Nombre de tabla: FactableTiempoJuego

Tipo de tabla: Factable / Transaccional

Tabla 11.

Descripción de la Fact table tiempo de juego

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdTiempoJuego	Numérico (long)	Clave primaria de la fact table, identifica la actividad del tiempo de juego	Primaria	
IdApp	Alfanumérico (string)	Clave foránea de Videojuegos, identifica videojuego utilizado durante la actividad registrada	Foránea	ETL de DATASET5_fixed
IdJugador	Numérico (long)	Clave foránea de jugador, identifica al jugador que jugo en la actividad registrada	Foránea	ETL de DATASET5_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
IdFecha	Numérico (integer)	Clave foránea de fecha, identifica la fecha de la actividad de juego registrada	Foránea	ETL de DATASET5_fixed
TiempoJuegoMinutos	Numérico (integer)	Representa el conteo total de minutos que el jugador realizó la actividad de juego en la fecha registrada		ETL de DATASET5_fixed
TiempoJuegoHoras	Numérico (doublé)	Representa el conteo total de horas que el jugador realizó. La actividad de juego en la fecha registrada		ETL de DATASET5_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
SistemaOperativoUsado	Alfanumérico (string)	Sistema operativo utilizado por el jugador en la actividad de juego		ETL de DATASET5_fixed

Nota: Contiene información del tiempo de juego de los usuarios, su nivel de detalle es tiempo de juego diario de cada usuario por juego

3.8.4 Fact Table Reseña

Nombre de tabla: FactableReseña

Tipo de tabla: Factable / Transaccional

Tabla 12.

Descripción de la Fact table reseña

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdReseña	Numérico (integer)	Clave primaria de la fact table, identifica la reseña al videojuego	Primaria	
IdApp	Alfanumérico (string)	Clave foránea de Videojuegos, identifica videojuego al que se refiere la reseña	Foránea	ETL de DATASET3_fixed
IdJugador	Numérico (long)	Clave foránea de jugador, identifica al jugador que realiza la reseña	Foránea	ETL de DATASET3_fixed
IdContextoReseña	Numérico (long)	Clave foránea, que identifica el contexto que da información adicional de la reseña	Foránea	

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
IdFechaCreacion	Numérico (integer)	Clave foránea de fecha, identifica la fecha de creación de la reseña	Foránea	ETL de DATASET3_fixed
IdFechaActualizacion	Numérico (integer)	Clave foránea de fecha, identifica la fecha de última actualización de la reseña, por defecto es la fecha de creación de la reseña	Foránea	ETL de DATASET3_fixed
NumeroVotosUtiles	Numérico (integer)	Representa el conteo total de votos útiles dados a la reseña por el resto de usuarios de Steam		ETL de DATASET3_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
NumeroVotosDivertidos	Numérico (integer)	Representa el conteo total de votos divertidos dados a la reseña por el resto de usuarios de Steam		ETL de DATASET3_fixed
Puntuacion	Numérico (integer)	Representa el conteo total de puntuación dados a la reseña por el resto de usuarios de Steam		ETL de DATASET3_fixed
NumeroComentarios	Numérico (integer)	Representa el conteo total de comentarios en la reseña		ETL de DATASET3_fixed
TiempoUsoAntesReseña	Numérico (integer)	Conteo total de minutos que el jugador uso el videojuego antes de realizar la reseña		ETL de DATASET3_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
TiempoTotalUso	Numérico (integer)	Conteo total de minutos que el jugador uso el videojuego hasta la fecha de la captura de los datos		ETL de DATASET3_fixed

Nota: Contiene información de reseñas realizadas a videojuegos por los usuarios, su nivel de detalle es reseña de jugador por videojuego

3.8.5 Dimensión Videojuego

Nombre de tabla: DimensionVideoJuego

Tipo de tabla: Dimensión

Tabla 13.

Dimensión Videojuego

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdApp	Alfanumérico (string)	Clave primaria de la dimensión , identifica al videojuego	Primaria	ETL de DATASET1_fixed
NombreApp	Alfanumérico (string)	Nombre comercial del videojuego registrado en Steam		ETL de DATASET1_fixed
FechaPublicacion	Alfanumérico (string)	Fecha de publicación o lanzamiento al mercado del videojuego		ETL de DATASET1_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
Desarrolladora	Alfanumérico (string)	Empresa desarrolladora del videojuego		ETL de DATASET1_fixed
Publicadora	Alfanumérico (string)	Empresa publicadora del videojuego		ETL de DATASET1_fixed
SistemaOperativoMinimo	Alfanumérico (string)	Sistema operativo mínimo para el funcionamiento del videojuego		ETL de DATASET1_fixed
SistemaOperativoRecomendado	Alfanumérico (string)	Sistema operativo recomendado para el funcionamiento del videojuego		ETL de DATASET1_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
ProcesadorMinimo	Alfanumérico (string)	Procesador mínimo para el funcionamiento del videojuego		ETL de DATASET1_fixed
ProcesadorRecomendado	Alfanumérico (string)	Procesador recomendado para el funcionamiento del videojuego		ETL de DATASET1_fixed
RamMinimo	Alfanumérico (string)	Memoria RAM mínima para el funcionamiento del videojuego		ETL de DATASET1_fixed
RamRecomendado	Alfanumérico (string)	Memoria RAM recomendada para el funcionamiento del videojuego		ETL de DATASET1_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
VideoMinimo	Alfanumérico (string)	Tarjeta de video mínima para el funcionamiento del videojuego		ETL de DATASET1_fixed
VideoRecomendado	Alfanumérico (string)	Tarjeta de video recomendada para el funcionamiento del videojuego		ETL de DATASET1_fixed
AlmacenamientoMinimo	Alfanumérico (string)	Espacio en disco duro mínimo para el funcionamiento del videojuego		ETL de DATASET1_fixed

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
AlmacenamientoRecomendado	Alfanumérico (string)	Espacio en disco duro recomendado para el funcionamiento del videojuego		ETL de DATASET1_fixed

Nota: Contiene información de los videojuegos que se encuentran en la plataforma Steam

3.8.6 Dimensión Genero

Nombre de tabla: DimensionGenero

Tipo de tabla: Dimensión

Tabla 14.

Dimensión de Generos

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdGenero	Numérico (long)	Clave primaria de la dimensión , identifica a un género de videojuego	Primaria	
Genero	Alfanumérico (string)	Nombre de clasificación de un género de videojuego		ETL de DATASET1_fixed

Nota: Contiene los géneros en la clasificación de temáticas que pueden clasificarse los videojuegos

3.8.7 Dimensión Idioma

Nombre de tabla: DimensionIdioma

Tipo de tabla: Dimensión

Tabla *15*.

Dimensión de Idiomas

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
IdIdioma	Numérico (long)	Clave primaria de la dimensión, identifica al idioma	Primaria	
Idioma	Alfanumérico (string)	Nombre del idioma		ETL de DATASET1_fixed

Nota: Contiene los idiomas que pueden ser asignados a un videojuego

3.8.8 Dimensión Jugador

Nombre de tabla: DimensionJugador

Tipo de tabla: Dimensión

Tabla 16.

Dimensión de jugador

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdJugador	Numérico (long)	Clave primaria de la dimensión , identifica al jugador	Primaria	ETL de DATASET4_fixed
Pais	Alfanumérico (string)	País al cual pertenece el jugador		ETL de DATASET4_fixed y CodigoPais
Continente	Alfanumérico (string)	Continente al cual pertenece el jugador		ETL de DATASET4_fixed
NumeroJuegos	Numérico (Integer)	Numero de videojuegos registrados en la cuenta del jugador		ETL de DATASET4_fixed

Nota: Contiene la información del jugador que hace uso de la plataforma Steam para las transacciones estudiadas

3.8.9 Dimensión Fecha

Nombre de tabla: DimensionFecha

Tipo de tabla: Dimensión

Tabla 17.

Dimensión de fecha

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdFecha	Numérico (Integer)	Clave primaria de la dimensión , identifica la fecha	Primaria	Dataset Fecha
Fecha1	Alfanumérico (string)	Primer formato de fecha (día-mes-año)		Dataset Fecha
Fecha2	Alfanumérico (string)	Segundo formato de fecha (año-mes-día)		Dataset Fecha
Fecha3	Alfanumérico (string)	Tercer formato de fecha (mes abreviado día, año)		Dataset Fecha
Semana	Numérico (Integer)	Numero de semana del mes de la fecha		Dataset Fecha
AA	Numérico (Integer)	Año de la fecha en formato numérico		Dataset Fecha
MM	Numérico (Integer)	Mes de la fecha en formato numérico		Dataset Fecha
DD	Numérico (Integer)	Día de la fecha en formato numérico		Dataset Fecha
NombreMM	Alfanumérico (string)	Nombre del mes de la fecha		Dataset Fecha

Nombre Campos	Tipos de dato	Descripción	Clave	Fuente de datos
NombreCortoMM	Alfanumérico (string)	Nombre abreviado del mes de la fecha		Dataset Fecha
NombreDD	Alfanumérico (string)	Nombre del día de la fecha		Dataset Fecha
NombreCortoDD	Alfanumérico (string)	Nombre abreviado del día de la fecha		Dataset Fecha
Festivo	Alfanumérico (string)	Nombre del día festivo		Dataset Fecha
DiaFestivo	Alfanumérico (string)	Indica si el dia es festivo con los valores “SI” o “NO”		Dataset Fecha
Date	Fecha (Date)	Fecha en el formato año-mes-día en el tipo de dato Date		Dataset Fecha

Notas: Contiene información de fechas desde el año 2015 al año 2022, aquí se registran diferentes formatos de fecha e información relevante de ellas

3.8.10 Dimensión Contexto Reseña

Nombre de tabla: DimensionContextoReseña

Tipo de tabla: Dimensión

Tabla 18.

Dimensión de contexto de reseña

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdContextoReseña	Numérico (long)	Clave primaria que identifica el contexto de una reseña en especifico	Primaria	
Idioma	Alfanumérico (string)	Idioma de la reseña		ETL de DATASET3_fixed
FormatoAdquisicionApp	Alfanumérico (string)	Formato de adquisición del videojuego reseñado con los valores “Comprado” o “Gratuito”		ETL de DATASET3_fixed
ReseñaAccesoAnticipado	Alfanumérico (string)	Identifica si la reseña se da en un juego con acceso anticipado con los valores “SI” o “NO”		ETL de DATASET3_fixed
RecomiendaApp	Alfanumérico (string)	Identifica si en la reseña el jugador recomienda el videojuego con los valores “SI” o “NO”		ETL de DATASET3_fixed

Nota: Contiene información que da contexto a la transacción de la reseña a un videojuego

3.8.11 Bridge Grupo Genero

Nombre de tabla: GrupoGenero

Tipo de tabla: Bridge

Tabla 19.

Dimensión de contexto de reseña

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdGrupoGenero	Numérico (long)	Clave primaria que identifica las combinaciones de juego y los géneros que se le asignan	Primaria	
IdGenero	Numérico (long)	Identificador de un género de videojuego	Foránea	DimensionGenero
IdApp	Alfanumérico (string)	Identificador de un videojuego	Foránea	DimensionVideoJuego

Notas: Contiene las claves para identificar los diferentes generos en los que se puede clasificar un videojuego

3.8.12 Bridge Grupo Idioma

Nombre de tabla: GrupoIdioma

Tipo de tabla: Bridge

Tabla 20.

Bridge Grupo Idioma

Nombre Campos	**Tipos de dato**	**Descripción**	**Clave**	**Fuente de datos**
IdGrupoIdioma	Numérico (long)	Clave primaria que identifica las combinaciones de juego y los idiomas que este soporta	Primaria	
IdIdioma	Numérico(long)	Identificador de un idioma de videojuego	Foránea	DimensionIdioma
IdApp	Alfanumérico (string)	Identificador de un videojuego	Foránea	DimensionVideoJuego

Notas: Contiene las claves para identificar los diferentes idiomas en los que se puede clasificar un videojuego

3.9 Datawarehouse busmatrix

De acuerdo a la metodología de Kimball, “la matriz de bus de un Datawarehouse es la herramienta esencial para diseñar y comunicar la arquitectura del bus del almacén de datos empresarial” (Ross, 2013, p. 52). Por ello haciendo uso de esa herramienta en la tabla 18 y 19, se muestran las diferentes fact tables con sus respectivas dimensiones conformadas y sus stakeholders

Tabla 21.

Matriz de bus para las diferentes fact tables

Procesos del negocio (Fact table)	**DimensionVideojuego**	**DimensionJugador**	**DimensionFecha**	**DimensionContextoReseña**	**DimensionIdioma**	**DimensionGenero**	**GrupoIdioma**	**GrupoGenero**
FactableVentas	X	X	X		X	X	X	X
FactableReseña	X	X	X	X	X	X	X	X
FactableTiempoJuego	X	X	X		X	X	X	X

Nota: En la tabla se especifican las dimensiones conformadas que usa cada fact table de los diversos procesos del negocio

Tabla 22.

Matriz de bus con los StakeHolders de las fact tables

Procesos del negocio	Planeación	Mercadeo	Marketing	Operaciones	Finanzas
Ventas	X		X		X
Reseñas		X			
Horas de juego	X	X	X	X	

Nota: En la tabla se especifican los procesos del negocio que usarán las fact tables

3.10 Visualización de los datos

Para la visualización de datos el equipo de trabajo creo unos dashboard con la herramienta power BI, en ella tratamos de representar la información cuantitativa con diferentes gráficos que permitirán resolver dudas relacionadas al proceso del negocio que se está analizando.
Por ello se generó un dashbord para cada uno de las fact tables que contienen la información lo más limpia posible

Dashbord de venta de juegos:

En este reporte se presentan visualizaciones interactivas que permiten visualizar las ventas por región, montos de ingresos en ventas y número de ventas en periodos, idiomas o géneros más vendidos, ventas según requerimientos, ingresos totales en ventas, descuento total otorgado en ventas, también se presentan filtros por periodo o días festivos (ver figura 13)

Figura 15.
Dashbord ventas de videojuegos

Reseñas de Videojuegos

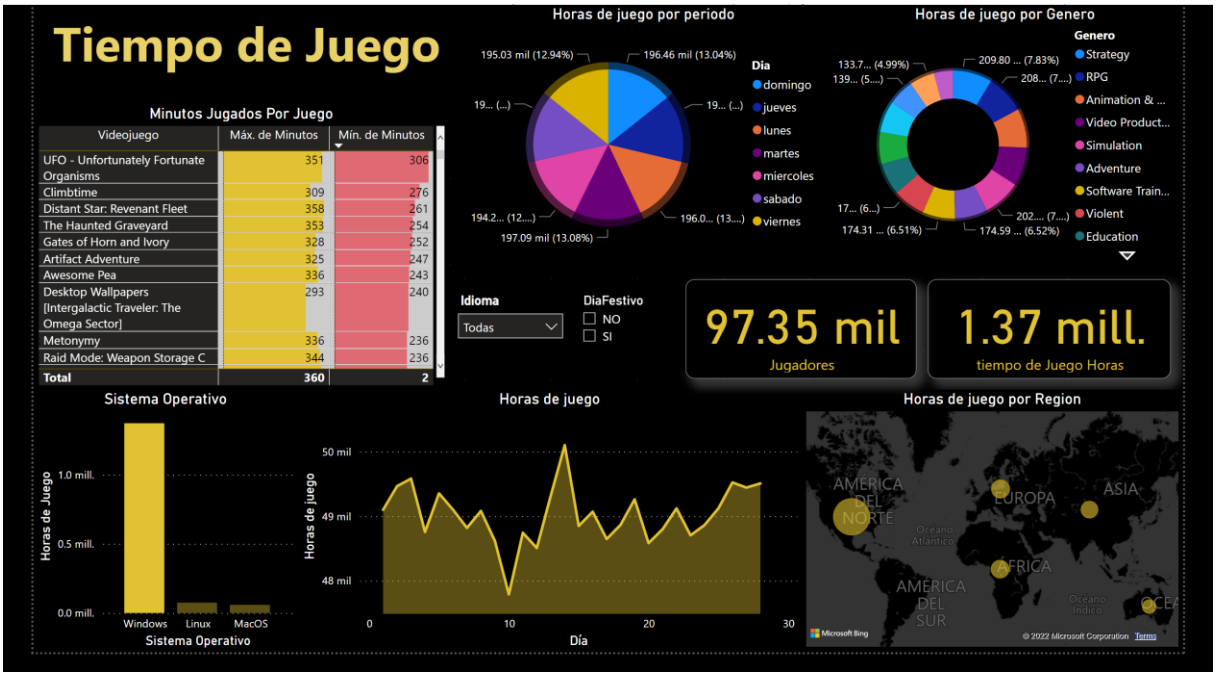
En este informe se presentan tarjetas que presentan sumatorias en número de reseñas, comentarios, votos a reseñas y horas de juego las cuales interactúan con el resto de visualizaciones, donde podemos ver el flujo de reseñas en el tiempo pudiendo seleccionar periodos como año, trimestre, mes, día, también flujo de reseñas por región, por videojuego, por categoría, puntuación dada a los videojuegos y su proporción al género que pertenecen, se añaden filtros como idioma en que se redactó la reseña, si es un positiva o no en la cual se recomiende el videojuego, el tipo de adquisición del juego de parte del autor de la reseña si fue acceso anticipado, comprado o gratuito

Figura 16.
Dashbord de reseñas de videojuegos

Tiempo de juego:

Este informe nos permite visualizar los tiempos de uso de los jugadores para los diferentes videojuegos, sus máximo o mínimo tiempo registrado para su uso, los géneros más jugados, los sistemas operativos más usados por los usuarios según horas de juego, tiempo de juego según región, periodos con mayor flujo de juego, el idioma en juegos más usado así como valores totales como número de jugadores y tiempos de juego en horas, la interacción entre las diferentes visualizaciones nos permitirán filtrar y obtener más contexto útiles para el análisis

Figura 17.
Dashbord de tiempos de juegos



CONCLUSIONES Y RECOMENDACIONES.

4.1 Conclusiones

Se analizaron los procesos de negocio de la plataforma de videojuegos Steam y se identificaron las principales transacciones que los usuarios de esta plataforma realizan en ella, los datos que estas transacciones generan proporcionan una gran capacidad de análisis que nos brindan información útil para la toma de decisión con respecto a los creadores de productos así como de las desarrolladoras o publicadoras para sus futuros juegos o la mejora de los existentes juegos, basados en la tendencia de los usuarios y su aceptación hacia los videojuegos de esta plataforma.

La presentación de indicadores mediante gráficos, ayuda a los usuarios del negocio tener un mayor conocimiento de la situación actual del negocio, permitiendo tomar mejores decisiones justificando dichas acciones con información congruentes.

El Data warehouse representa una excelente opción cuando se cuenta con grandes volúmenes de datos y se quiere hacer uso del valioso activo, que estos datos representan dentro de toda organización permitiendo tener una gran capacidad de organización y rápido acceso a los datos producidos por los sistemas transaccionales, esto nos ayuda a unificar la información de la organización bajo reglas de negocio definidas que permitan una gran flexibilidad de análisis sin impactar en los sistemas transaccionales

Con el Data Warehouse se incrementa la confiabilidad de la información, ya que en su construcción los datos se someten a un proceso de limpieza y transformación que garantizan la integridad de los datos con el nivel de detalle o granularidad más bajo posible que permite una gran flexibilidad en el análisis. El resultado de procesar la información nos genera un modelo dimensional que es accesible para los usuarios analíticos mediante aplicaciones como Power BI que es la utilizada en este proyecto para la presentación de informes

Una decisión acertada es el uso de las dimensiones conformadas, que básicamente significa que ocuparemos las mismas dimensiones en las diferentes fact tables, incrementando la dificultad a la hora de integrar todos los dataset pero disminuyendo los posibles problemas que pueden surgir como tener información en silos datos.

4.2 Recomendaciones

Con el creciente aumento de información generada en los sistemas transaccionales es importante planificar correctamente la periodicidad en la implementación de los ETL, generalmente se realizan a media noche cuando el uso de los sistemas transaccionales es menor, pero se deben tomar en cuenta las variables puntuales de cada organización y buscar el momento en el cual el impacto en el rendimiento de los sistemas transaccionales sea el menor posible

La seguridad siempre será un aspecto delicado y de gran prioridad, al implementar un Data warehouse con almacenamiento y computo en la nube, se deben incorporar políticas de control de acceso con la asignación de un acceso mínimo o con menos privilegios. cada usuario debe tener un acceso mínimo requerido para hacer su trabajo. El principio del menor privilegio incluye la limitación de los recursos y aplicaciones accesibles por el usuario, así como el acceso en tiempo permitido, permitiendo una mejor gestión de la seguridad, recursos limitados y el presupuesto

Se debe manejar el nivel de detalle o granularidad más bajo posible en cada modelo de negocio implementado en el Data warehouse, para tener una mayor flexibilidad en el análisis y poder soportar o facilitar futuros requerimientos analíticos que se le puedan solicitar

Es recomendable llevara cabo un estudio del negocio y de todas las transacciones involucradas en su modelo de negocios, así como de la variedad de fuentes de datos que lo respaldan, para garantizar un mejor análisis y una mayor calidad en la integración de los datos

BIBLIOGRAFÍA

Databricks Ink. (n.d.). *Databricks Pricing*.

Gupta, S., & Giri, V. (2018). *Practical Enterprise Data Lake Insights.* Bangalore, Karnataka, India: Apress.

Kimball , R., & Ross, M. (2013). *The Data Warehouse Toolkit.* Indianapolis, Indiana: Wiley.

Menjivar, M. (2021). Material de clases IDT115. El Salvador.

Mucattu Utils. (n.d.). *Código de países según ISO 3166-1*. Retrieved from http://utils.mucattu.com/iso_3166-1.html

Valve corporation. (2022, enero). *Steam*. Retrieved from https://store.steampowered.com/

Valve corporation. (2022). *Valve corporation*. Retrieved from https://www.valvesoftware.com/