Almacén del dato (*datawarehouse*)

Tipos y ciclo de vida del dato

CONTENIDO

Objetivos

- Conocer la arquitectura del almacén de datos (*data warehouse*).
- Conocer los tipos de bases de datos.
- Definir los procesos ETL y OLAP.

Introducción

Al momento de establecer la arquitectura que almacena los datos, se debe contemplar el proceso de implementación de ingesta de información, su automatización, así como su mantenimiento, por lo tanto, es necesario considerar lo siguiente:

- Cuáles son las necesidades asociadas al volumen y requerimientos de seguridad relacionadas con los datos que se manejan.
- Cuáles son los recursos con los que se cuenta tanto físicos como financieros.
- Qué opciones están disponibles en el mercado que satisfagan las necesidades.
- Los tiempos de implantación.
- Almacén y procesamiento local o en la nube.
- Si se cuenta con una fuente de información ya en uso, verificar la compatibilidad de la tecnología que usa con los recursos disponibles en el mercado, para evitar problemas de compatibilidad.

Elementos que integran el almacén de datos

Los conceptos básicos asociados a las estructurar usadas para el almacenamiento de los datos son los siguientes:

Data mart

Un *data mart* es una base de datos centrada en un ámbito que muchas veces es un segmento aislado de un almacén de datos de empresa. El subconjunto de datos contenido en un *data mart*. Los *data marts* aceleran los procesos comerciales al dar acceso a la información en un almacén de datos o un *data store* operativo en cuestión de días, y no meses o periodos más largos. Como un *data mart* tan solo contiene los datos aplicables a un ámbito comercial concreto, resulta una forma rentable de obtener información explotable rápidamente.

Data warehouse

El término proviene de la composición en inglés: *data* = datos y *warehouse* = almacén, por lo que la interpretación es directa a almacén de datos. Este término se acuñó especialmente para hablar de un almacén de datos diseñado para permitir las actividades de inteligencia de un negocio. Inmon define un *datawarehouse* como un conjunto de datos estructurados orientados por temas, integrados, variables con el tiempo y no volátiles empleados para tomar decisiones [1]. A continuación, se define cada una de estas características *(figura 1)*:

- **Orientados por temas:** colección de información relacionada y organizada alrededor de un tema central. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.
- **Integrados:** los datos se obtienen de fuentes diferentes, pero se deben aplicar técnicas de integración (agrupación) de los datos, para dar a estos una estructura de conjunto.
- **No volátiles:** quiere decir que los datos no van a cambiar con el tiempo una vez que se encuentran en el almacén. Los datos están para ser leídos, no modificado. La información ni se modifica ni se elimina.
- **Variables con el tiempo:** los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones. No siempre se cumple, ya que el diseño de recolección y almacén de datos es determinante para poder cumplir esta condición.

¿Por qué las empresas necesitan DWH?

Muchos usuarios empresariales se preguntan por qué es tan importante el almacenamiento de datos. La forma más sencilla de explicar esto es a través de los diversos beneficios que ofrece a los usuarios finales. Entre los beneficios se encuentran:

- Acceso mejorado del usuario final a una amplia variedad de datos empresariales.
- Mayor consistencia de datos.
- Documentación adicional de los datos.
- Costos informáticos potencialmente más bajos y mayor productividad.
- Proporcionar un lugar para combinar datos relacionados de fuentes independientes.
- Creación de una infraestructura informática que pueda soportar cambios en los sistemas informáticos y las estructuras comerciales.
- Empoderar a los usuarios finales para realizar consultas o informes *ad-hoc* sin afectar el rendimiento de los sistemas operativos.

La forma más habitual de almacenamiento de los datos son las bases de datos.

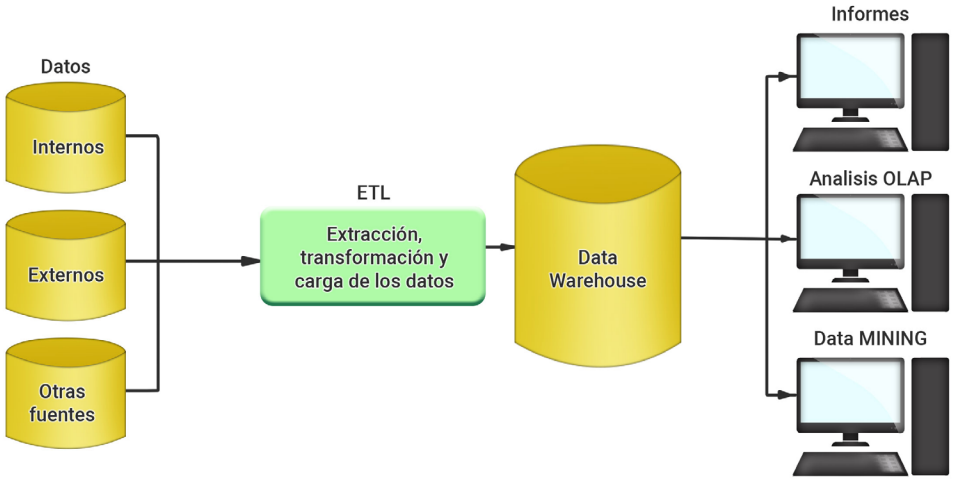

Figura 1. Arquitectura de data warehouse. Tomado de: https://www.areatecnologia.com/informatica/imagenes/data-warehouse-arquitectura.jpg

Data warehouse en la nube

El *data warehouse* están atravesando actualmente dos transformaciones muy importantes que tienen el potencial de impulsar niveles significativos de innovación empresarial:

- La primera área de transformación es el impulso para aumentar la agilidad general. La gran mayoría de los departamentos de TI están experimentando un rápido aumento de la demanda de datos. Los directivos quieren tener acceso a más y más datos históricos; mientras que, al mismo tiempo, los científicos de datos y los analistas de negocios están explorando formas de introducir nuevos flujos de datos en el almacén para enriquecer el análisis existente, así como impulsar nuevas áreas de análisis. Esta rápida expansión de los volúmenes y fuentes de datos significa que los equipos de TI necesitan invertir más tiempo y esfuerzo asegurando que el rendimiento de las consultas permanezca constante y necesitan proporcionar cada vez más entornos para equipos individuales para validar el valor comercial de los nuevos conjuntos de datos.
- La segunda área de transformación gira alrededor de la necesidad de mejorar el control de costes. Existe una creciente necesidad de hacer más con cada vez menos recursos, al mismo tiempo, que se garantiza que todos los datos sensibles y estratégicos estén completamente asegurados, a lo largo de todo el ciclo de vida, de la manera más rentable.

  En el almacenamiento en la nube existen tres ventajas principales para mover un *data warehouse* a la nube, que están directamente vinculados a los tres controladores clave enumerados anteriormente:

  - Mayor facilidad de consolidación y racionalización.
  - Monetización más rápida de los datos en la nube.
  - La nube ofrece mejor protección.

Bases de datos

Las bases de datos se pueden definir como una colección de datos o información estructurada almacenados comúnmente en un sistema electrónico. Las bases de datos, al tener gran cantidad de información ordenada, facilitan el uso y la manera de encontrar datos de forma rápida y sencilla.

Tipos de bases de datos

Relacionales

Entre las bases de datos más utilizada actualmente, se tiene la base de datos relacional. Una base de datos relacional es un tipo de base de datos que aplica un modelo de relación. Así, según esta definición de base de datos relacional, se trata de una base de datos que almacena y da acceso a puntos de datos relacionados entre sí. El modelo relacional es una forma intuitiva y directa de representar datos sin necesidad de jerarquizarlos (postulado por primera vez en 1970 por Edgar Frank Codd) [1].

Características

- Una base de datos relacional es, en esencia, un conjunto de tablas (o relaciones) formadas por filas (registros) y columnas (campos); así, cada registro (cada fila) tiene una ID única, denominada clave y las columnas de la tabla contienen los atributos de los datos.
- Evita la duplicidad de registros y a su vez garantiza la integridad referencial, es decir, si se elimina uno de los registros, la integridad de los registros restantes no será afectada.
- No pueden existir dos tablas con el mismo nombre y la relación entre una tabla padre y una tabla hija se lleva a cabo a través de claves primarias.

- Para poder almacenar, administrar, consultar y recuperar los datos guardados en la base de datos relacional es necesario emplear softwares específicos para gestión de bases de datos relacionales.

Ventajas

- Quizás la principal ventaja de la base de datos relacional reside en la sencillez del modelo relacional, que permite manejar grandes cantidades de datos con puntos de relación entre sí.
- Las bases de datos relacionales permiten mantener la uniformidad de los datos en todas las aplicaciones y copias de la propia base, denominadas instancias.
- Favorece la normalización al ser más comprensible y aplicable.
- No genera conflictos cuando varios usuarios o aplicaciones intentan acceder a los mismos datos en el mismo momento, pueden bloquear dicho acceso mientras los datos se están actualizando.
- Pueden concurrir varios usuarios o aplicaciones al mismo tiempo en la misma base de datos.

Desventajas

- Son deficientes a la hora de manejar datos gráficos, multimedia, CAD y sistemas de información geográfica, pues necesitan un soporte más dinámico.
- Tampoco permiten desarrollar tablas organizadas de formar jerárquica, es decir, no se puede crear un subfila, porque todas las filas están al mismo nivel jerárquico.
- Debido a que están distribuidas en tablas separadas, esto provoca un rendimiento negativo a la hora de hacer consultas y obtener la información deseada.

Estructura

La base de datos está dividida en dos secciones: el esquema y los datos. A través del esquema se define la estructura de la base de datos relacional, que almacena los siguientes datos:

1. **El nombre de cada tabla (o relación):** es el conjunto de tuplas que comparten los mismos atributos, es decir, un conjunto de filas y columnas.
2. **El nombre de cada columna (atributo o campo):** es un elemento etiquetado de una tupla.
3. El tipo de dato de cada columna.
4. La tabla a la que pertenece cada columna.

Esta sería la estructura básica de una tabla de una base de datos relacional *(figura 2)*.

Actualmente existen varios tipos de gestores de BDR, entre ellos, los más usados son:

- Oracle.
- MySQL.
- Microsoft SQL Server.
- PostgreSQL.
- DB2.

Finalmente, las relaciones que se pueden establecer entre los diferentes elementos de dos tablas en una base de datos relacional pueden ser de tres tipos:

- Relaciones uno a uno cuando se establecen entre una entidad de una tabla y otra entidad de otra tabla.
- Relaciones uno a varios cuando se establecen entre varias entidades de una tabla y una entidad de otra tabla.
- Relaciones varios a varios cuando se establecen entre varias entidades de cada una de las tablas.

No relacionales

Una base de datos no relacional es un sistema de almacenamiento de datos que cuenta con particularidades que las diferencian del otro gran grupo de bases de datos las relacionales.

Las bases de datos no relacionales son un sistema de almacenamiento de información que se caracteriza por no usar el lenguaje SQL para las consultas. Esto no significa que no puedan usar el lenguaje SQL, pero no lo hacen como herramienta de consulta, sino como apoyo. Por ello, también se les suele llamar NoSQL. Otra de sus principales características es que no trabajan con estructuras definidas.

Características

- La información no se almacena en tablas sino a través de documentos.
- Son bases de datos muy útiles para organizar y gestionar información no estructurada, o cuando no se tiene una noción clara de los datos a almacenar.
- Son altamente escalables y están diseñadas para soportar grandes volúmenes de datos.
- No utilizan el lenguaje SQL para consultas, aunque sí lo pueden usar como herramienta de apoyo.
- Es un sistema de almacenamiento de datos relativamente nuevo y, como tal, todavía no posee un sistema estandarizado.
- A diferencia de las no relacionales, no garantizan el cumplimiento de las cualidades ACID, esto es, atomicidad, consistencia, integridad y durabilidad.

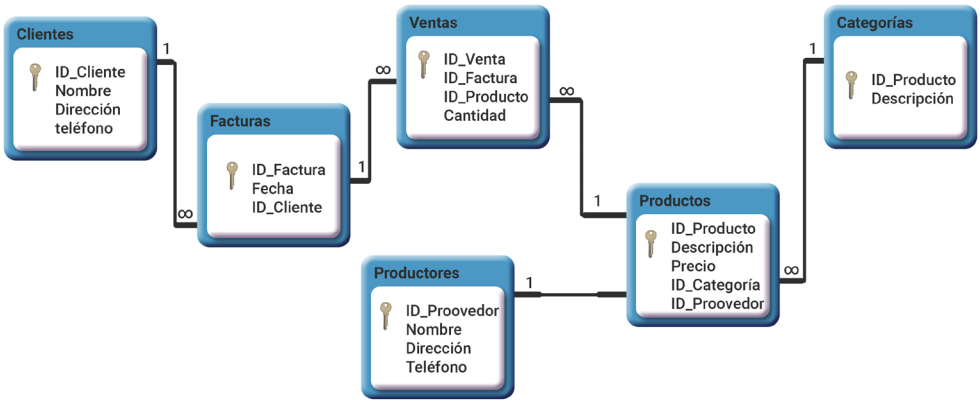

Figura 2. Ejemplo de base de datos relacional sistema de destino. Tomado de: WordPress, 2016, recuperado de: https://finanzastics2.files.wordpress.com/2016/07/aaaa.jpg.

Ventajas

- Son mucho más flexibles a la hora de crear esquemas de información, lo que las convierte en una solución ideal para el almacenamiento y gestión de datos no estructurados o semiestructurados.
- Ofrecen una mayor escalabilidad. Pueden soportar mayores volúmenes de datos y añadir mayor capacidad añadiendo nuevos módulos de *software*, sin necesidad de añadir nuevos servidores.
- Garantizan un alto rendimiento, ya que están diseñadas para trabajar con modelos de datos concretos y patrones de acceso específicos.
- Son muy funcionales, ya que cuentan con API exclusivas y proporcionan modelos de datos para trabajar con cada tipo de datos presentes en la base.

Desventajas

- No cumplen igual que las relacionales con las propiedades de atomicidad, consistencia, integridad y durabilidad.
- No son compatibles con determinadas consultas en lenguaje SQL.
- Carecen de un sistema estandarizado, ya que todavía son bases de datos relativamente nuevas.
- Muchos sistemas de gestión de bases de datos relacionales son de código abierto y tienen una gran comunidad detrás programando soluciones y nuevas funcionalidades. En el caso de las bases de datos no relacionales, este soporte es mucho más limitado.

Estructura

Una base de datos no relacional no requiere de tablas para el almacenamiento de la información. Normalmente la estructura de una base de datos relacional se basa en la organización de la información a través de documentos. Este tipo de data bases están pensadas para ofrecer mayor escalabilidad horizontal y no tienen identificadores que permitan establecer relaciones entre diferentes conjuntos de datos.

Cada uno de los documentos almacenados en la base de datos incluye todos los atributos del elemento, por lo que resultan muy útiles a la hora de guardar información poco estructurada o de la que no se tiene un esquema claro de inicio.

Tipos de datos que almacenan

- **Clave - valor:** se trata de bases de datos no relacionales que almacenan la información en base a pares de clave valor. Es decir, cada clave sirve como un identificador único, y a cada una de ellas se le aplica un valor. Son especialmente usadas a la hora de almacenar datos de juegos, aplicaciones o aparatos que funcionan mediante el internet de las cosas (IoT).
- **Documentos:** en una base de datos relacional basada en documentos la información se representa como objetos o documentos JSON. Su principal ventaja es que los documentos son de naturaleza flexible, semiestructurada y jerárquica, lo que facilita a los desarrolladores las tareas de almacenamiento, gestión y consulta de datos. Es un modelo usado habitualmente en sistemas de administración de contenidos o para gestionar perfiles de usuarios.
- **Gráficos:** las bases de datos no relacionales basadas en gráficos están pensadas para crear relaciones y navegar por ellas. Las entidades de datos se almacenan mediante nodos y los bordes son los que crean las relaciones entre entidades. Con frecuencia las bases de datos gráficas se emplean en redes sociales, sistemas de detección o prevención de fraudes o sistemas de recomendaciones.
- **En memoria:** son bases de datos no relacionales diseñadas para ofrecer respuestas en milisegundos y soportar grandes picos de tráfico. Un ejemplo de bases de datos en memoria son las empleadas en tablas de clasificaciones de juegos o en herramientas para hacer análisis en tiempo real.

Base de datos no relacional *vs* relacional

- En las bases de datos relacionales la información se organiza de forma estructurada en tablas; en las no relacionales no es así.
- Una base de datos no relacional no usa el lenguaje SQL como lenguaje principal para sus consultas.
- Las bases de datos no relacionales se emplean, sobre todo, para almacenar datos no estructurados o semiestructurados.
- Una base de datos relacional no cumple con las propiedades ACID con la misma eficacia que una base de datos relacional.
- La escalabilidad es mayor en una base de datos no relacional, y también están preparadas para soportar mayor volumen de datos.
- Las bases de datos no relacionales o NoSQL también ofrecen una mayor flexibilidad y escalabilidad horizontal.
- A diferencia de las relacionales, las bases de datos no relacionales todavía no disponen de un lenguaje estandarizado (SQL).
- El soporte de la comunidad es mejor en el caso de las bases no relacionales. Extract, transform and load (ETL) son herramientas de extracción, transporte, transformación y carga de los datos.

Extract, transform and load (ETL) herramientas de extracción, transporte, transformación y carga de los datos

- **Extracción de datos:** es lo primero que hace una herramienta ETL. Se trata de obtener la información de las distintas fuentes de origen, tanto internas como externas. Después de la extracción de datos, tienen que ser transportados físicamente al sistema de destino o a un sistema intermedio para su posterior procesamiento y/o transformación.
- **Transformación:** es el filtrado, limpieza, depuración, homogeneización y agrupación de la información. Incluye la agrupación de los datos de las diferentes fuentes. La transformación se produce mediante el uso de reglas o tablas de consulta o mediante la combinación de los datos con otros datos.
- **Carga:** es el proceso de escribir los datos en la *data warehouse*. La fase de carga es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema de destino *(figura 3)*.

Las características clave de ETL son las siguientes:

- **Conectividad:** la herramienta ETL que elija debe conectarse a todas las fuentes de datos utilizadas por su empresa. Idealmente, debería tener conectores incorporados para todos sus sistemas necesarios, incluidas bases de datos, aplicaciones de ventas y *marketing*, formatos de archivo y más, lo que facilita la obtención de datos desde y hacia cualquier sistema.
- **Interfaz usuario:** una interfaz libre de errores y fácil de usar proporciona una experiencia consistente y confiable para manejar tareas relacionadas con datos. La configuración sencilla es un beneficio adicional que puede ayudarlo a dar vida a sus canalizaciones de datos en cuestión de minutos.
- **Escalabilidad:** a medida que el negocio crezca, las necesidades de datos también se expandirán. Por lo tanto, la herramienta ETL debe tener características de optimización del rendimiento, como la optimización *pushdown*, para satisfacer sus crecientes necesidades comerciales.
- **Manejo de errores:** la herramienta ETL debe ser capaz de manejar errores de manera eficiente, asegurando la consistencia y precisión de los datos. Además, debe ofrecer capacidades de transformación de datos fluidas y eficientes, asegurando una pérdida de datos cero.
- **Accesibilidad:** obtener datos en tiempo real se está volviendo imprescindible para las empresas que buscan obtener información oportuna. Una herramienta ETL debe poder acceder a los datos de las aplicaciones web en tiempo real para garantizar un tiempo de comprensión más rápido.

Procesamiento analítico en línea (OLAP)

El término proviene de la sigla en inglés OLAP que significa *on-line analytical processing* o procesamiento analítico en línea. Es el método más utilizado para analizar y evaluar los datos de la *data warehouse* en línea. Permite a los gerentes y analistas obtener una idea de la información. Así pues, para los programas OLAP un tiempo de respuesta es una medida de su eficacia. Este modelo de procesamiento de la información permite al usuario extraer y ver con facilidad y de forma selectiva los datos desde diferentes puntos de vista. Recursos de análisis numérico y grafico (*data mining*) implementación de los procesos de ingesta de información, su automatización, así como su mantenimiento *(figura 4)*.

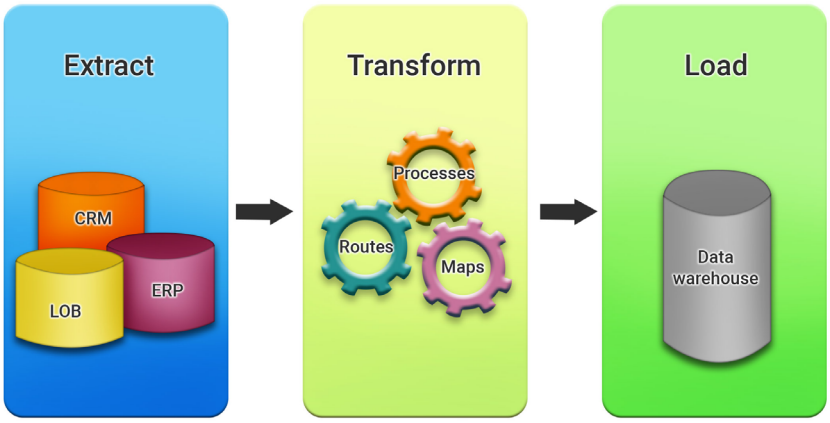

Figura 3. Modelo elemental de ETL. Tomado de: Astera, 2020, recuperado de: https://www.astera.com/wp-content/uploads/2020/04/etl.png

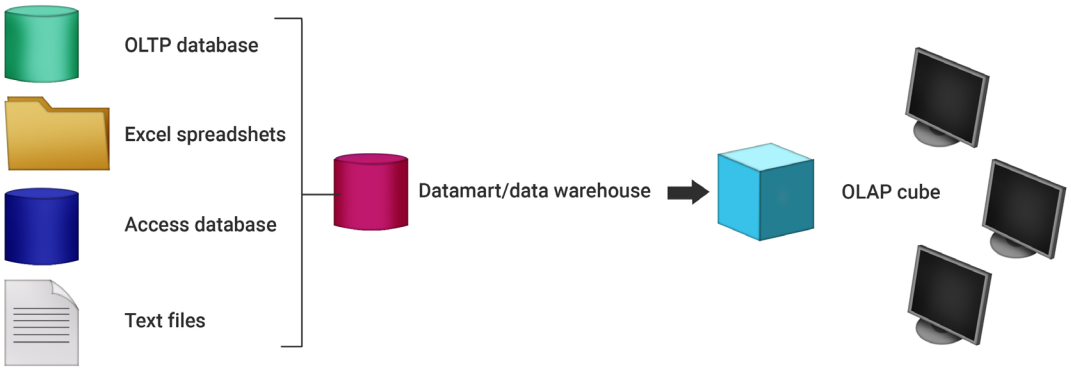

Figura 4. Esquema general de las etapas de la condición de datos. Tomado de: Galaktika, 2018, recuperado de: https://galaktika-soft.com/wp-content/uploads/2018/01/oltp.jpg

Bibliografía

[1] Grupo PowerData, "Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos", Powerdata.es. [En línea]. Disponible en: https://www.powerdata.es/data-warehouse. [Accedido: 12-oct-2021].