



Desarrollo de contenido

Unidad 2

Bases de Datos II

Ingeniería de *Software* y Datos

Unidad 2. Diseño de un *Data Warehouse*

Introducción *Unidad 2*

¡Hola de nuevo! Continuamos este proceso de aprendizaje en torno a las bases de datos. Ahora, en la segunda unidad, nos enfocaremos en el diseño de un *Data Warehouse*, una herramienta fundamental en el universo de la información y la toma de decisiones.

Imagina el *Data Warehouse* DWH como una especie de almacén donde guardamos y organizamos todos nuestros datos de manera ordenada, con toda la información de nuestro interés lista para ser explorada y analizada. En esta nueva unidad vamos a descubrir los secretos detrás de los elementos de un DWH, estudiaremos los diferentes tipos de *Data Warehouse* que existen y exploraremos la arquitectura que sustenta este universo de datos.

Te iremos guiando en cada paso del camino para hacer que este viaje sea tan emocionante como formativo. ¡Vamos a empezar!

Resultados de aprendizaje *Unidad 2*

- Reconoce los elementos fundamentales de un *Data Warehouse*, incluyendo sus componentes, tipos y arquitectura.
- Desarrolla una estrategia de implementación de arquitectura de *Data Warehouse* que optimice el rendimiento y la eficiencia operativa de una organización.
- Identifica patrones y tendencias en los datos de un *Data Warehouse* para responder preguntas empresariales específicas.

Introducción diseño de un *Data Warehouse*

En el mundo de la inteligencia de negocios hay diferentes elementos como *ETL*, *OLAP*, *reporting*, pero el principal es el almacén de datos, también conocido como *data warehouse*.

El *data warehouse* es, básicamente, el archivero maestro de la empresa. Su misión es proporcionar una visión global, clara e integrada de todos los datos relevantes de la organización. No importa cómo se utilicen después, este almacén los guarda con cualidades especiales: es estable, coherente, fiable y tiene un historial completo (Curto Díaz & Conesa Caralt, 2010).

Imagina un archivo que abarca toda la información generada por la organización y tiene datos desde hace mucho tiempo, fácilmente llegando a cantidades enormes (piensa en centenas de *terabytes*). Veamos a continuación algunas de sus características:

- **Orientado a un tema**

Agrupar la información alrededor de un tema central, como un capítulo especializado en un libro.

- **Integrado**

Recoge datos de todos los lugares posibles y los presenta de manera consistente, todo en un solo idioma, podríamos decir.

- **Variable en el tiempo**

No es solo una foto instantánea, sino que toma "fotos" de los datos basadas en fechas o eventos. Es similar a un álbum de fotos que muestra la evolución.

- **No volátil**

Nadie puede andar cambiando cosas aquí, solamente está para ser observado. Puedes imaginarlo como un museo de datos donde solo puedes mirar, mas no tocar.

Casi siempre este almacén gigante usa bases de datos relacionales, pero hay otras opciones interesantes como bases de datos orientadas a columnas o basadas en lógica asociativa. Podemos pensarlo como el cerebro de datos de la empresa, el cual mantiene todo organizado y listo para su análisis (Curto Díaz & Conesa Caralt, 2010).

Tema 1. Elementos de un DWH

En el mundo de las bases de datos, las bases operativas siguen reglas estrictas llamadas formas normales en su diseño. Pero, cuando hablamos de almacenes de datos, rompemos esas reglas.

La idea es que la información se presenta de una manera especial, llamada "desnormalizada". Suena complicado, ¿verdad? Sin embargo, es simplemente para hacer las consultas más rápidas y eficientes.

Entonces, ¿cómo lo logramos? Identificamos los procesos de negocio en nuestra organización, las diferentes vistas de esos procesos y las medidas cuantificables relacionadas

con ellos. En resumen, es como si estuviéramos organizando nuestro propio juego de LEGO para que sea más fácil y rápido construir lo que necesitamos.

Conceptos importantes relacionados con *Data Warehouse*

A continuación te explicaremos algunos conceptos importantes que debemos tener en cuenta cuando hablamos de un *Data Warehouse* según Josep Curto Díaz y Jordi Conesa Caralt (2010).

Tabla de hecho

Piensa la tabla de hecho como la representación, en el almacén de datos, de todos los movimientos que suceden en una organización. Por ejemplo: podríamos tener una tabla de hecho llamada *Ventas*, que registre todas las veces que alguien compra algo.

En las tablas de hecho podemos encontrar dos aspectos importantes.

- Medidas especiales: te dicen cuánto está sucediendo en la organización, cuántos movimientos se dan con los datos almacenados.
- Claves foráneas: esto es como enlazar partes de la historia, te lleva a otras tablas que tienen información más detallada. Es como tener una guía para entender cada detalle.

Existen diferentes tipos de tablas de hecho, te las explicamos a continuación.

- Imagina que estás viendo un partido de fútbol. La **Transaction Fact Table** sería como tener un registro detallado de cada momento emocionante durante el juego.

Aquí tendrías información específica sobre cada evento, quién hizo un gol (métrica importante), cuándo sucedió y otros detalles. Esta tabla te permite analizar el partido con un máximo detalle, así como revisar las estadísticas claves: la cantidad total de goles o quién fue el jugador más destacado.

En resumen, es como tener una grabación detallada de cada acción para comprender completamente lo que sucedió.

Factless Fact Table o Coverage Table

Vamos a visualizarlo de manera más sencilla. Piensa en una **Factless Fact Table o Coverage Table** como una lista de asistencia para un evento, como un acto benéfico.

En esta tabla, cada fila representa la asistencia de una persona al evento. A diferencia de otras tablas de hecho, no estás registrando medidas específicas (como el monto donado o la cantidad de productos comprados). Aquí el enfoque está en el hecho de que la persona asistió al acto benéfico.

A menudo, para facilitar las consultas, se pueden agregar contadores. Por ejemplo: podrías tener un contador que indique cuántas personas fueron en total al acto benéfico. Así, aunque no tengas métricas detalladas asociadas a cada persona, puedes analizar la cobertura del evento y obtener información útil mediante consultas SQL.

Periodic Snapshot Fact Table

Imagina una ***Periodic Snapshot Fact Table*** como un registro periódico que recopila información a intervalos regulares de tiempo. Esto es útil cuando deseas tener una visión de los datos en momentos específicos.

Considera, por ejemplo, el balance mensual de una empresa. En esta tabla tendrías datos acumulados recopilados al final de cada mes. A través de esta estructura, puedes analizar cómo evolucionan las métricas claves en períodos de tiempo definidos, lo que facilita la comprensión de tendencias y cambios a lo largo del tiempo.

Accumulating Snapshot Fact Table

Ahora imagina una ***Accumulating Snapshot Fact Table*** como un registro que sigue el ciclo completo de una actividad o proceso, desde su inicio hasta su fin. Esta tabla es ideal para actividades que tienen un proceso definido con múltiples etapas y dimensiones.

Supongamos que estamos revisando el proceso de matriculación de un estudiante. Esta tabla acumula datos durante todo el ciclo de vida del proceso, reemplazando y actualizando información relevante a medida que avanzas en las etapas del proceso, como la superación y recopilación de asignaturas. Es como seguir paso a paso el proceso completo de una actividad.

➤ **Dimensión:**

Imagina que ves una historia desde diferentes ángulos. Piensa en una "dimensión" como una vista especial para un proceso en tu organización, por ejemplo, cuando alguien realiza una compra. En esta vista, podríamos ver al cliente que compró, la fecha en que sucedió, el producto vendido, el cliente, el vendedor o incluso el canal de venta. Esto lo podemos comparar con la exploración de diferentes aspectos de una historia para entenderla mejor.

Vamos a conocer el mundo de las *Slowly Changing Dimensions (SCD)*, que son como una máquina del tiempo para tus datos. Supongamos que estás gestionando la evolución histórica de tus datos, ¿cómo lo haces? A continuación, te presentamos algunos enfoques:

- SCD Tipo 0: no se preocupa por cambios históricos, simplemente no los nota. La información permanece inmutable para siempre.
- SCD Tipo 1: ¿Quieres vivir en el presente? Este tipo sobrescribe el pasado con la nueva información. Es rápido y fácil de mantener; sin embargo, pierdes el rastro de la historia.
- SCD Tipo 2: aquí guardamos todas las versiones históricas. Cuando algo cambia creamos una nueva entrada y marcamos el cambio en el tiempo. Es como tener un álbum de fotos con cada versión de tu dimensión.
- SCD Tipo 3: similar al Tipo 2, pero en lugar de crear nuevas entradas, añadimos columnas para los valores antiguos y actuales. Un poco de espacio extra en la memoria puede contar toda la historia.
- SCD Tipo 4: llamado "tablas históricas", mantiene dos tablas: una con los datos actuales y otra con los antiguos o los cambios, al estilo "antes y después". Siempre puedes echar un vistazo al pasado.
- SCD Tipo 6/Híbrida: la fusión perfecta, combina enfoques de los tipos 1, 2 y 3. Imagina tener un registro de tipo 1, pero con poderes temporales. Es un poco complejo pero ayuda a reducir el tamaño de las consultas temporales.

¿Cómo quieres gestionar el paso del tiempo en tu almacén de datos? ¡Tú decides!

- Métrica: son los indicadores de un proceso de negocio, aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio.

Existen varios tipos de medidas que se diferencian según la información que recopilan y su función asociada (Building Talent, 2024).

- Métricas: son valores que registran el desarrollo o los resultados de una actividad, derivan directamente de la ejecución de la actividad de negocio (Curto Díaz, 2016, p14-15).

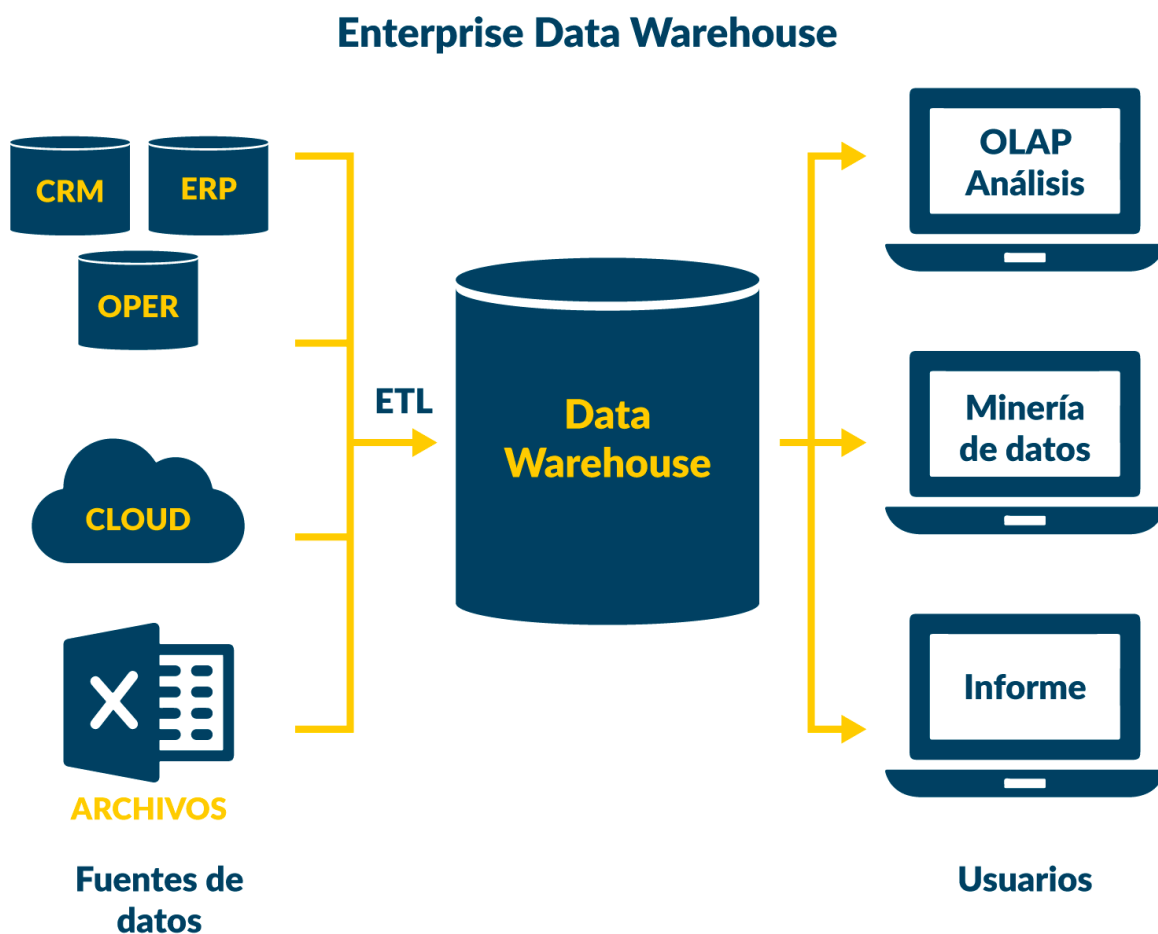
- Métricas de realización de actividad (*leading*): evalúan el progreso o cumplimiento de una actividad. Por ejemplo: la participación de una persona en un evento.
 - Métricas de resultado de una actividad (*lagging*): reflejan los logros o desempeño después de completar una actividad. Por ejemplo: la cantidad de puntos anotados por un jugador en un partido.
- Indicadores clave: son valores que se deben alcanzar y representan el nivel de cumplimiento de los objetivos. Estas medidas ofrecen información sobre el rendimiento de una actividad o el logro de una meta.
- *Key Performance Indicator* (KPI): indicadores clave de rendimiento. Van más allá de la eficacia, estableciendo valores que indican en qué rango óptimo de rendimiento deberíamos encontrarnos al alcanzar los objetivos. Son métricas del proceso.
 - *Key Goal Indicator* (KGI): indicadores de metas. Establecen mediciones para informar a la dirección general si un proceso TI (Tecnología de la Información) ha cumplido con sus requisitos de negocio, expresándose generalmente en términos de criterios de información.
 - Es esencial mencionar que también existen indicadores de rendimiento (que básicamente son los KPI). Son mediciones que nos dicen cómo le está yendo al proceso de TI para alcanzar la meta. Estos indicadores principales nos dicen si es posible o no lograr un objetivo, son muy buenos para evaluar las capacidades, prácticas y habilidades. Los indicadores más básicos sobre metas se transforman en indicadores de rendimiento cuando llegamos a niveles más altos (HubSpot, 2024).

Tema 2. Tipos de DWH

Enterprise *Data Warehouse*

Almacena toda la información de una empresa y sirve para que cualquier parte pueda acceder a esos datos. Ayuda en la toma de decisiones y se centra en organizar y mostrar los datos de manera ordenada. También puede clasificar los datos según quién los esté usando y permite el acceso de acuerdo con las reglas internas de la empresa. En pocas palabras, es como una biblioteca central de información para toda la empresa (Gravitar, 2024).

Figura 1. Enterprise Data Warehouse



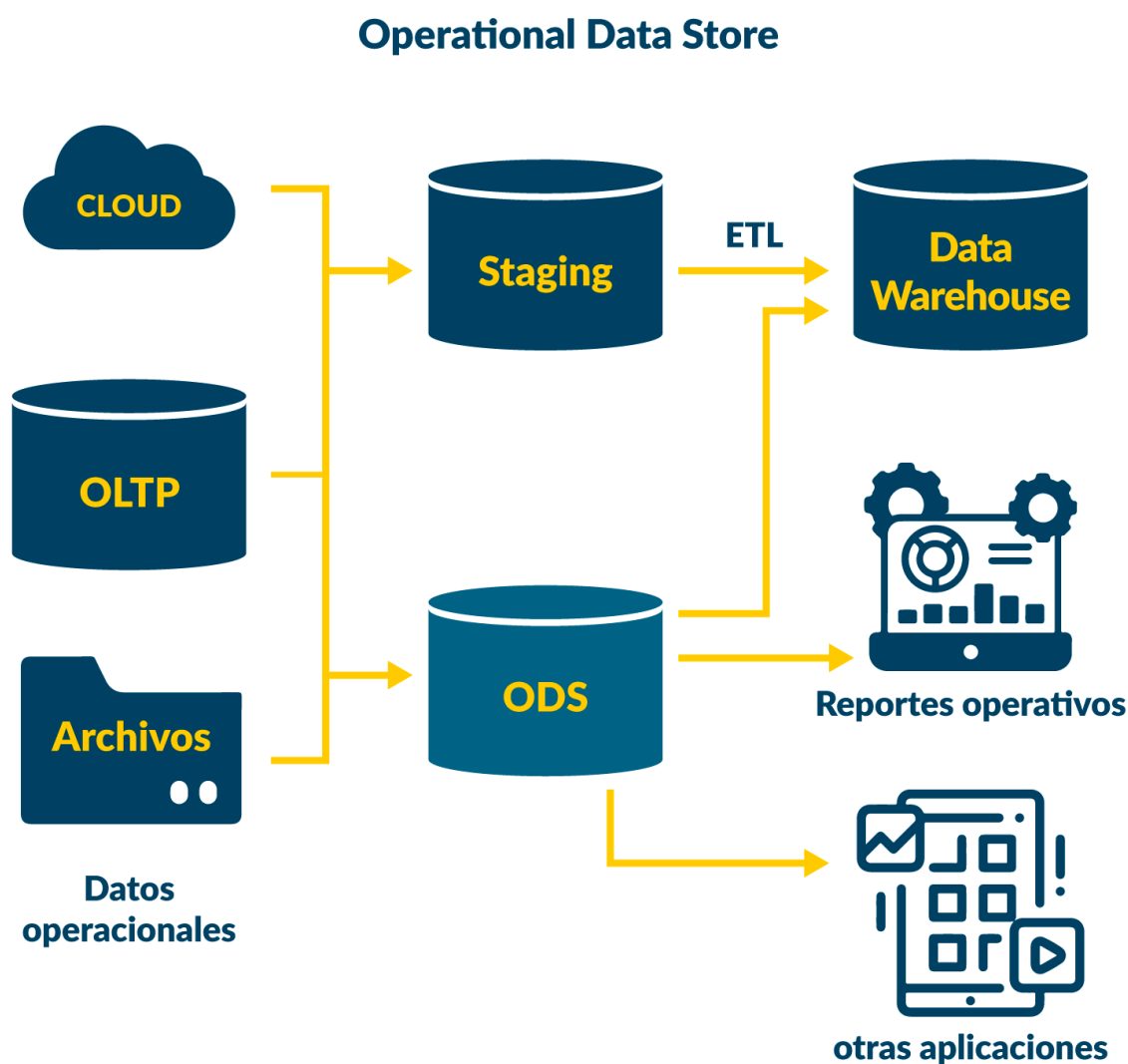
Fuente: adaptado de Gravitar (2024).

Operational Data Store (ODS)

También llamado ODS. Lo podemos pensar como un almacén de datos que se utiliza cuando las necesidades de informes de una organización no son cubiertas adecuadamente por el almacenamiento de datos y los sistemas OLTP.

En el ODS la información se actualiza en tiempo real o con una baja latencia de actualización, por lo que es útil para actividades cotidianas como almacenar registros operativos y transacciones de ventas. Podríamos decir que es una especie de base de datos temporal utilizada como un área lógica intermedia para un almacén de datos (Gravitar, 2024).

Figura 2. Operational Data Store



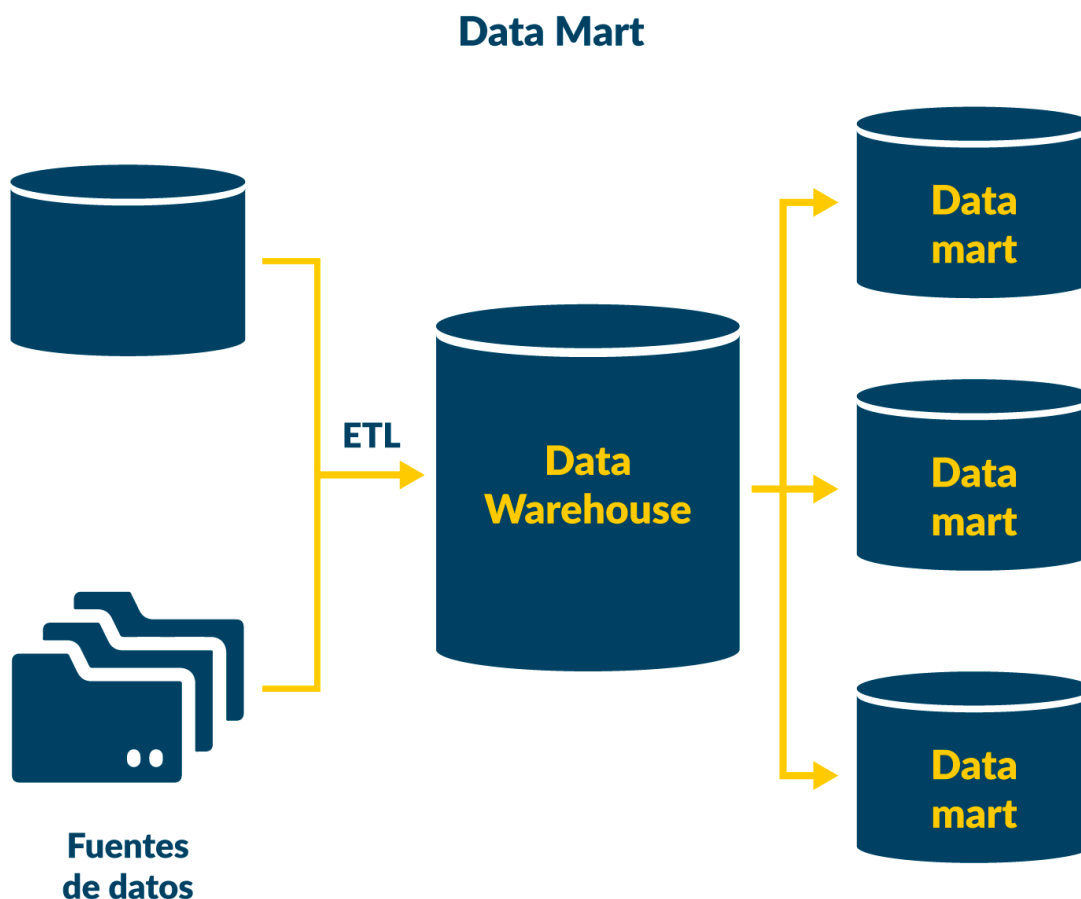
Fuente: adaptado de Gravitar, (2024).

Data Mart

Es una pequeña parte del almacenamiento de datos centrado en un área específica, también conocido como una base de información departamental. Se diseña especialmente para una línea de negocio particular, como ventas o finanzas. En un *Data Mart* independiente los datos pueden recopilarse directamente de las fuentes.

Este cumple diversas funciones como organizar la información para su posterior análisis, generar Indicadores Clave de Rendimiento (KPI), evaluar los objetivos del sector analizado, entre otros. La meta es llevar a cabo un análisis detallado sobre lo que está sucediendo en un negocio (Gravitar, 2024).

Figura 3. Data Mart



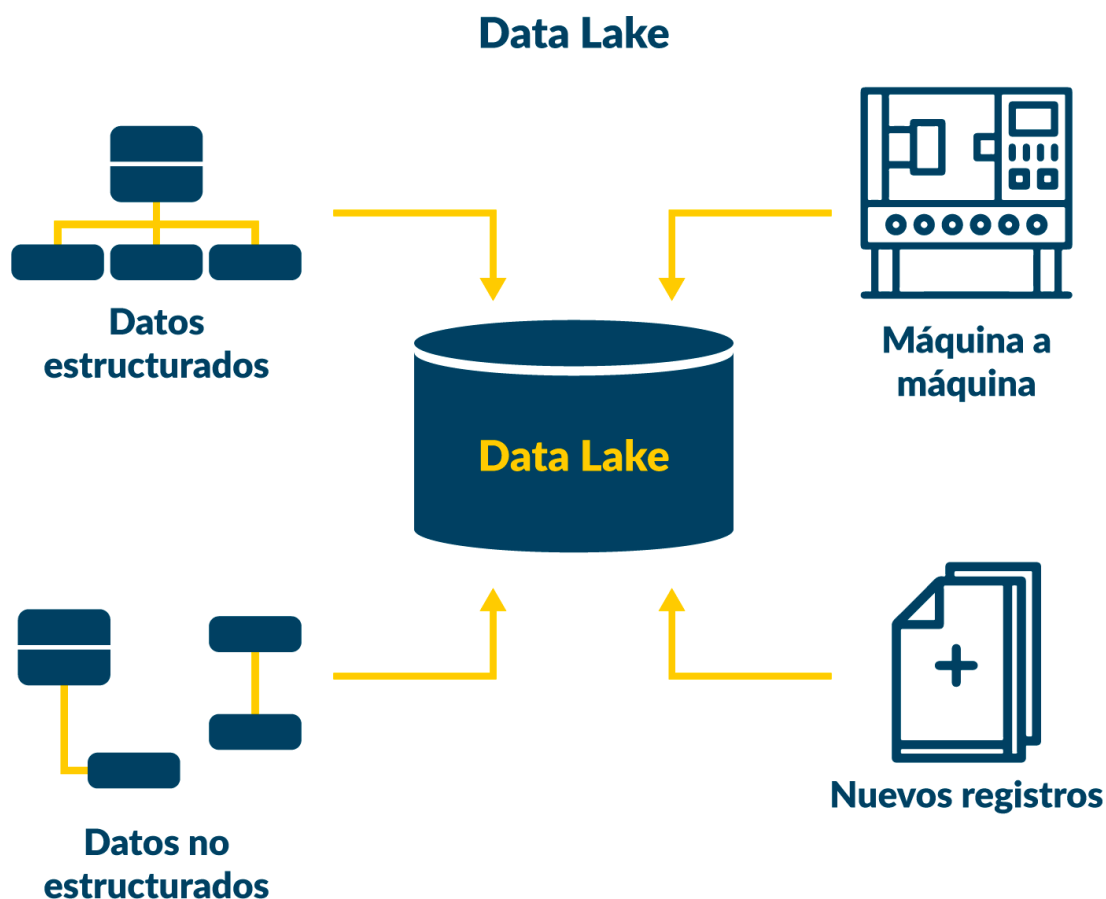
Fuente: adaptado de Gravitar, (2024).

Data Lake

Un *Data Lake* es como un "lago" que almacena una gran cantidad de datos, ya sean estructurados o sin estructurar, sin importar su tamaño. Permite guardar los datos como

son, sin necesidad de organizarlos previamente. A ello se suma que realiza diversos tipos de análisis, crea paneles de control y visualizaciones, además de ejecutar procesos de *big data*, análisis en tiempo real y de aprendizaje automático (*Machine Learning*) para facilitar la toma de decisiones (Gravitar, 2024).

Figura 4. Data Lake



Fuente: adaptado de Gravitar, (2024).

Tema 3. Arquitectura del DWH

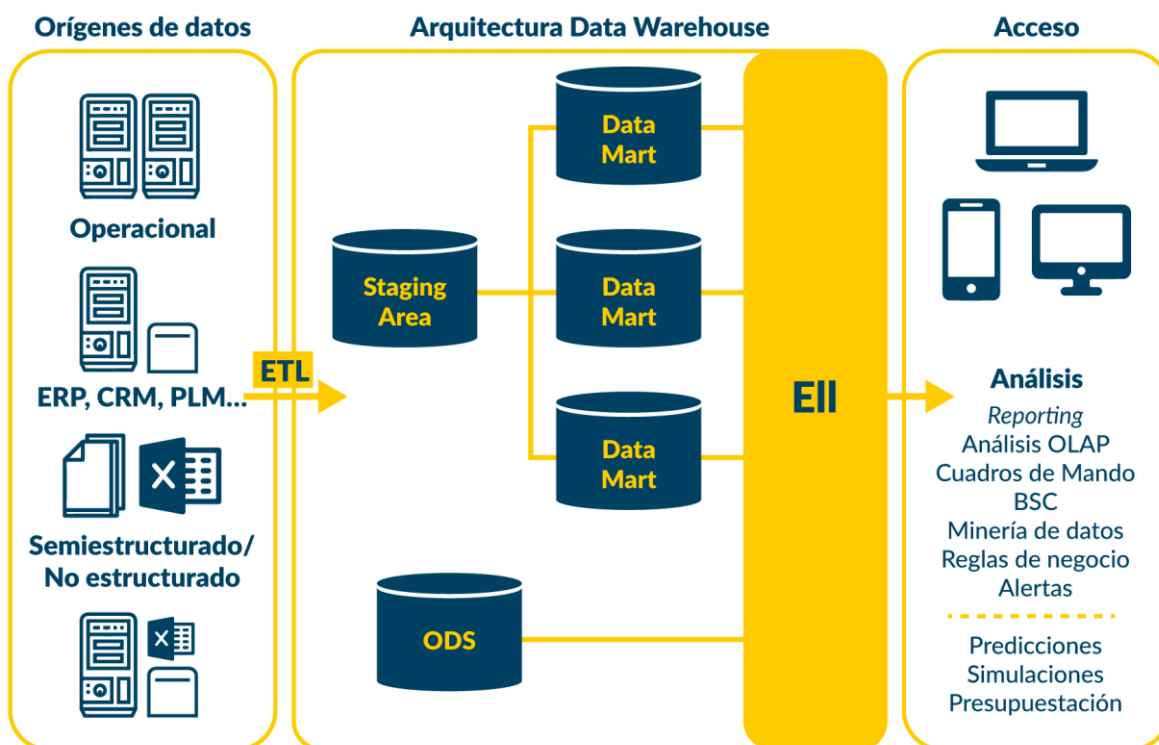
Existen, principalmente, tres enfoques en la arquitectura corporativa de un *Data Warehouse*.

- También conocida como **Arquitectura de Bus Empresarial (o Almacén de Datos Virtual/Federado)**

Esta estructura, también denominada MD (Arquitectura Multidimensional), se compone de *data marts* independientes federados, que pueden utilizar

una zona de preparación (*staging area*) si es necesario. La federación implica el uso de una herramienta EII (Integración de Información Empresarial) para realizar consultas como si se tratara de un único almacén de datos. En caso de necesidad, podría incluirse un ODS (Gravitar, 2024).

Figura 5. Arquitectura de Bus Empresarial (Almacén de Datos Virtual/Federado)

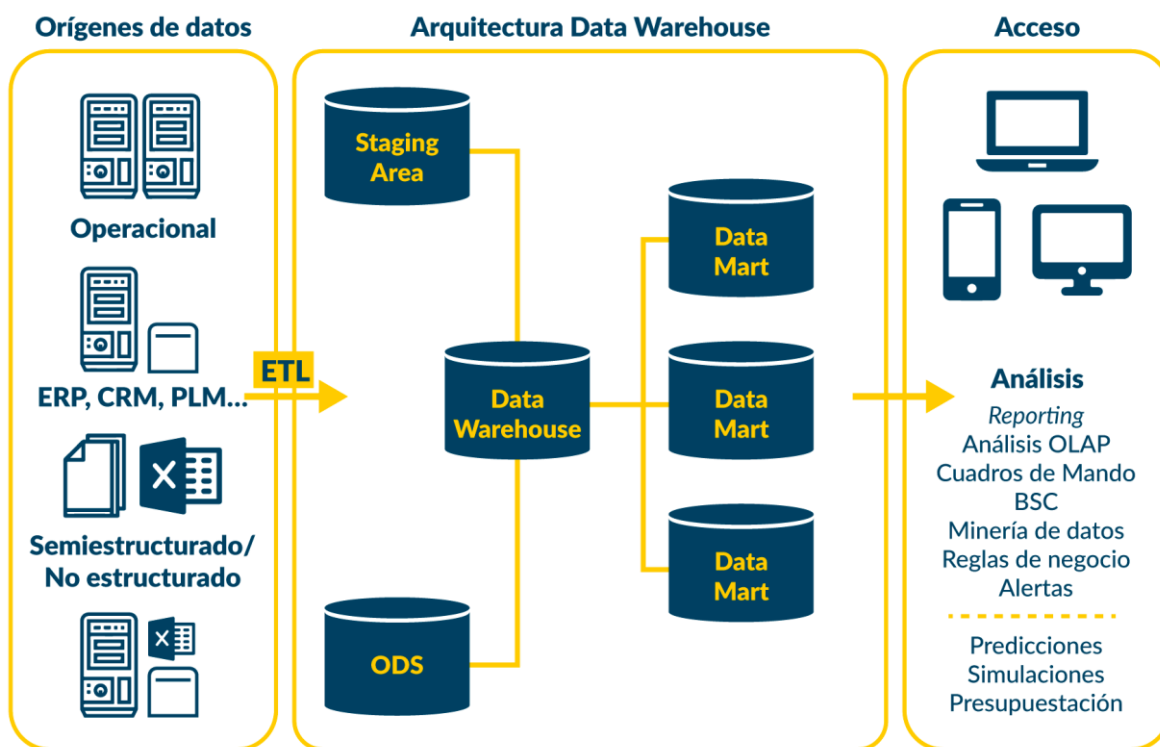


Fuente: adaptado de Gravitar, (2024).

b. **Corporate Information Factory** (o *Enterprise Data Warehouse*)

La Fábrica de Información Corporativa (o Almacén de Datos Empresarial) implica una arquitectura que consta de un almacén de datos corporativo y *data marts* (o incluso cubos OLAP) que dependen de él. El acceso a los datos se lleva a cabo a través de los *data marts* o el ODS, en caso de existir, pero nunca directamente al propio almacén de datos. Podría incluirse una zona de preparación en caso de ser necesaria (Gravitar, 2024).

Figura 6. Corporate Information Factory



Fuente: adaptado de Gravitir, (2024).

c. Enterprise Data Warehouse 2.0

La versión 2.0 del Almacén de Datos Empresarial implica una revisión de la metodología de Bill Inmon que incorpora toda la experiencia de las últimas dos décadas. La diferencia clave radica en la segregación de la información según su antigüedad y su clasificación según su uso. Se caracteriza por abarcar tanto la información estructurada como la no estructurada, centrándose en el objetivo de satisfacer todas las necesidades actuales del negocio. El siguiente gráfico ilustra una arquitectura completa (Gravitir, 2024).

Has culminado la segunda unidad del curso, estudiaste los elementos del *Data Warehouse*, los tipos que existen y la arquitectura que sostiene el universo de los datos. En la próxima unidad nos enfocaremos en el diseño de procesos ETL (Extracción, Transformación y Carga).

¡Sigue adelante en tu aprendizaje!



Lecturas y Material Complementario

A continuación te recomendamos las siguientes lecturas para ampliar sobre los temas tratados en la unidad.

Título: Introducción al Business Intelligence

Autores: Jordi Conesa Caralt (coord.) Josep Curto Díaz

URL: https://cursos.yura.website/wp-content/uploads/2020/03/Introduccion_al_Business_Intelligence.pdf

Título: Nociones, consideraciones y ventajas de la inteligencia de negocios BI

Autores: Hans Zamora Carrillo, Norberto Novoa Torres y Davián Ricardo Bermúdez Huérfano

DOI: <https://doi.org/10.14483/2322939X.15592>



Referencias bibliográficas

Auribox Training, (15 de junio de 2017). ¿Qué es un Data Warehouse? | Business Intelligence. [Video]. YouTube. <https://www.youtube.com/watch?v=jFsRdTcljeU>

Building Talent, (2024). 10 conceptos para entender el Business Intelligence | Blog IL3 - UB. <https://www.il3.ub.edu/blog/10-conceptos-basicos-para-entender-el-business-intelligence-del-presente/>

Curto Díaz, Josep., & Conesa Caralt, Jordi. (2010). Introducción al Business Intelligence. Editorial UOC. https://cursos.yura.website/wp-content/uploads/2020/03/Introduccion_al_Business_Intelligence.pdf

El Dataverso, (23 de febrero de 2023). Modelos Copo de Nieve y Estrella, ¿cuál es el camino? [Video]. YouTube. <https://www.youtube.com/watch?v=EWlhiklfNbl>

Gravitar, (2024). Tipos de Data Warehouse | Business Intelligence, Data Warehouse, Monterrey, México: Gravitar. <https://gravitar.biz/datawarehouse/data-warehouse-tipos/>

HubSpot (2024). Data warehouse: guía sencilla para entender qué es un almacén de datos. <https://blog.hubspot.es/marketing/data-warehouse>

Smartbase Group (5 nov 2015) Modelos Estrella Vs Copo de Nieve: Dimensiones. [Video]. YouTube. https://www.youtube.com/watch?v=lkJTj_XgOC7U

Zamora Carillo, H., Novoa, N. y Bermúdez, D. (2019). Nociones, consideraciones y ventajas de la inteligencia de negocios BI. *Revista vínculos*, 16(2), 280-287. (dic. 2019), 280–287. DOI: <https://doi.org/10.14483/2322939X.15592>



IU Digital de Antioquia

INSTITUCIÓN UNIVERSITARIA
DIGITAL DE ANTIOQUIA



Esta licencia permite a otros distribuir, remezclar, retocar, y crear a partir de esta obra de manera no comercial y, a pesar que sus nuevas obras deben siempre mencionar a la **IU Digital** y mantenerse sin fines comerciales, no están obligados a licenciar obras derivadas bajo las mismas condiciones.