
Universidad
Tecmilenio®

Universidad Tecmilenio

Consulta en Microsoft SQL Server®

Introducción al almacenamiento de datos

Semana 9

Introducción

En el día a día de las industrias y los mercados, la información toma cada vez más relevancia, volviéndose un factor determinante para su éxito.

Es fundamental considerar la opción de crear una base de datos para centralizar la información y así obtener fácil acceso y buen control.

La ingeniería de datos es vista como la columna vertebral de las actividades del análisis avanzado, inteligencia comercial y aprendizaje automático.

Explicación

Almacenamiento de datos: concebido como un repositorio centralizado en tablas de datos estructurados (tablas de base de datos).

Ventajas del uso de las bases de datos

- Recopilar la información de varias fuentes en un solo repositorio.
- Reducir los costos de generar una eficiente y sencilla captura de datos.
- Resguardar y tratar la información que históricamente fue recopilada.
- Buscar uniformidad en los datos para darles coherencia y precisión al unirlos.
- Utilizar métodos de análisis avanzado de datos para preparar la información.
- Procesos de análisis y recopilación de información más eficaces para que la toma de decisiones se vuelva más ágil.

Universidad Tecmilenio

Explicación

La arquitectura del almacenamiento de datos se trata de un sistema diseñado y estructurado que se conforma de varios niveles que interactúan con los datos de diferentes formas.

Niveles de un almacenamiento de datos:

Nivel inferior

La información se recopila, limpia y normaliza para realizar un proceso de extracción, transformación y carga (ETL).

Nivel intermedio

Se encuentra el motor que llevará a cabo el análisis, el cual es conocido como OLAP o servidor de procesamiento analítico en línea.

Nivel superior

Los datos procesados se presentan de forma visual gracias a la interfaz front-end.

Explicación

SQL Server logra destacarse como la mejor opción para implementar un sistema de gestión de base de datos relacional.

Microsoft ha creado un contenedor de información llamado **AdventureWorks** que permite la documentación para el manejo de diferentes versiones de bases de datos, de acuerdo con las necesidades del programador y de la empresa.

Las diferentes tecnologías que maneja Microsoft para el almacenamiento de datos son las siguientes (Microsoft, 2022a):

- SQL Server (todas las versiones compatibles).
- Azure SQL Database.
- Azure SQL Managed Instance.
- Azure Synapse Analytics.
- Analytics Platform System (PDW).

Explicación

Los lagos de datos pueden resguardar información sin procesar, provienen de múltiples fuentes no relacionales y admiten datos estructurados y semiestructurados. Asimismo, no se define su esquema hasta que los datos son leídos.

Explicación

Gracias a su naturaleza flexible y escalable, los lagos de datos se usan para realizar formas inteligentes de análisis de datos, como el aprendizaje automático (Microsoft, 2022b).

Rubro	Lago de datos	Almacén de lago de datos
Tipo	Estructurado, semiestructurado, no estructurado. Relacional, no relacional.	Estructurado, semiestructurado, no estructurado. Relacional, no relacional.
Esquema	Esquema al leer.	Esquema en lectura, esquema en escritura.
Formato	Sin procesar, sin filtrar, procesado, mantenido.	Archivos sin formato, sin filtrar, procesados, mantenidos, con formato delta.
Orígenes	Macrodatos, IoT, redes sociales, datos de *streaming*.	Macrodatos, IoT, redes sociales, datos de streaming, aplicación, negocios, datos transaccionales, informes por lotes.
Escalabilidad	Fácil escalado a bajo costo.	Fácil escalado a bajo costo.
Usuarios	Científicos de datos.	Analistas de negocios, ingenieros de datos, científicos de datos.
Casos de uso	Aprendizaje automático, análisis predictivo.	Informes principales, BI, aprendizaje automático, análisis predictivo.

Microsoft. (2022c). *¿Qué es Lago de datos?* Recuperado de https://azure.microsoft.com/es-mx/resources/cloud-computing-dictionary/what-is-a-data-lake/#get-started

Explicación

SQL Server contiene diferentes versiones y características que varían dependiendo de su costo, el cual puede cambiar debido a la edición, versión y a la configuración del servidor.

Las principales versiones de SQL Server se presentan a continuación:

Actividad

Con base en lo descrito en el tema, reflexiona sobre las siguientes preguntas:

01 ¿Cuál es la importancia de tener una base de datos en una organización?

02 ¿En qué tipo de industria te imaginas que se podría implementar un proceso OLAP?

Universidad Tecmilenio

Cierre

Toda organización busca la mejor solución para sus necesidades de almacenamiento de datos. Para ello, **Microsoft** brinda soluciones de almacenamiento a través de su software **SQL server**, como una de las mejores opciones para el uso e instalación de **bases de datos**.

- La evolución de dicha tecnología ha dado paso a una sólida construcción de documentación que puede ser accesible y alcanzable a cualquier nivel.

Bibliografía

• Microsoft. (2022a). *Lección 1: Creación y consulta de objetos de base de datos*. Recuperado de https://docs.microsoft.com/es-es/sql/t-sql/lesson-1-creating-database-objects?view=sql-server-ver16

• Microsoft. (2022b). *¿Qué es el almacenamiento de datos?* Recuperado de https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-a-data-warehouse/

• Microsoft. (2022c). *¿Qué es Lago de datos?* Recuperado de https://azure.microsoft.com/es-mx/resources/cloud-computing-dictionary/what-is-a-data-lake/#get-started

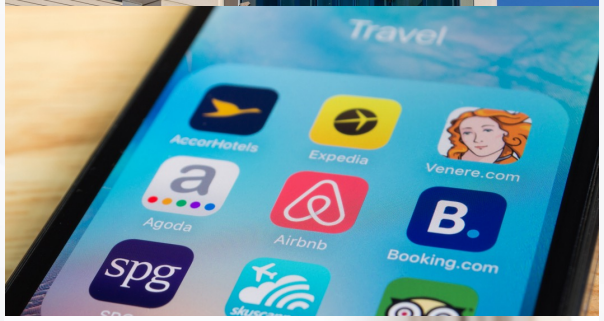
Universidad Tecmilenio

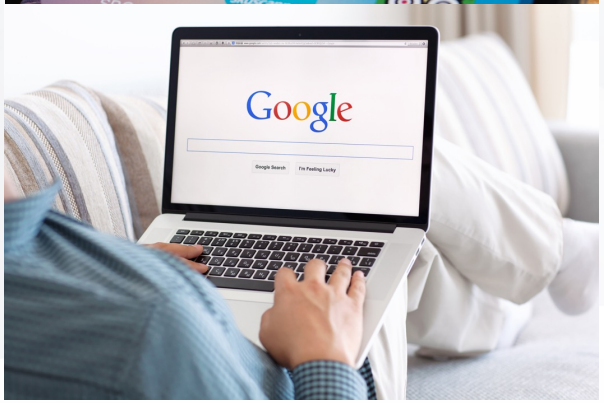
Consulta en Microsoft SQL Server®

Diseño e implementación de un almacén de datos

Semana 9

Introducción





Amazon, Airbnb, Google, Tesla, Starbucks, Facebook, Zara y Nike son empresas expertas en la relación de información de sus bases de datos, donde sus repositorios cuentan con un gran volumen de información que, al final, tienen datos que se relacionan entre sí.

¿Qué es lo que permite identificar el detalle de un elemento que se desea analizar para la toma de decisiones en un negocio?

Explicación

Descripción general del diseño del almacén de datos

El almacén de datos (*data warehouse* o DW) se convierte en un tipo de sistema que gestiona datos diseñados para dar soporte y habilitar las tareas de inteligencia empresarial o *business intelligence* (BI).

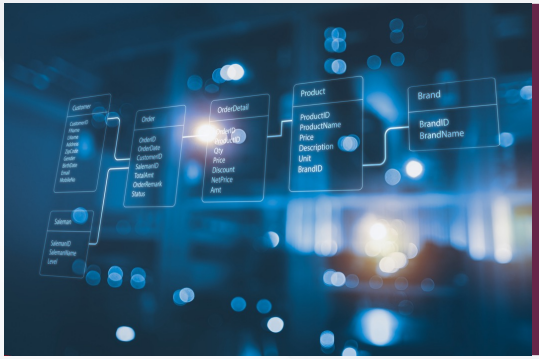
Existen elementos clave que deben contenerse en los repositorios con el propósito de hacerlos más eficientes.

Dichos elementos son los siguientes (Oracle, 2022):

- Una tabla relacional.
- Un paquete o alguna solución ETL.
- Un método estadístico para el análisis de información.
- Métodos de captura de datos.
- Elementos para la visualización y el análisis de información del negocio.
- De manera más compleja y, por tanto, opcional, se pueden introducir códigos y algoritmos.

Universidad Tecmilenio

Explicación



Las tablas en donde se deposita la información se clasifican de la siguiente forma:

Tabla de hechos

En ella se determinan todos los elementos a medir y/o analizar de manera global.

Tabla de dimensiones

Se realiza a partir de la tabla de hechos, dando un detalle exhaustivo al campo seleccionado correspondiente al elemento de la tabla de hechos.

Explicación

Características de las tablas de dimensiones

- Este tipo de tablas no cuenta con un límite de caracteres ni de elementos.
- Tienen una cantidad mínima y limitada de registros.
- Realizan un enlace con las tablas utilizando un campo clave.
- Los atributos que ofrece una tabla de dimensión otorgan datos acerca de las tablas de hechos.

Rivas, A. (2021). *Tabla de dimensiones*. Recuperado de https://www.muytecnologicos.com/diccionario-tecnologico/tabla-de-dimensiones

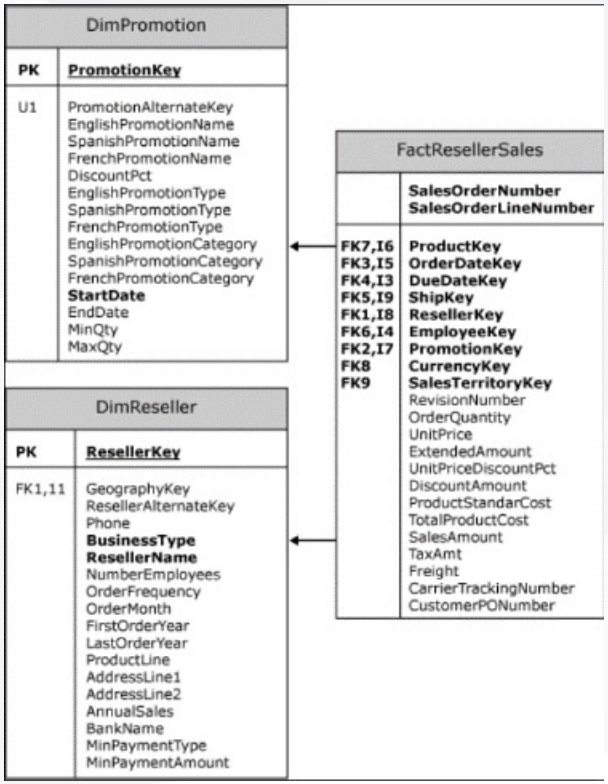
Explicación

Tipos de dimensiones que **SQL Server Analysis Services** pone a disposición del usuario para la elaboración de tablas de dimensión:

- **Regular**
- **Tiempo**
- **Organización**
- **Geografía**
- **Lista de materiales**
- **Cuentas**
- **Clientes**
- **Productos**
- **Guion**
- **Cuantitativo**
- **Utilidad**
- **Divisa**
- **Tarifas**
- **Canal**
- **Promoción**

Explicación

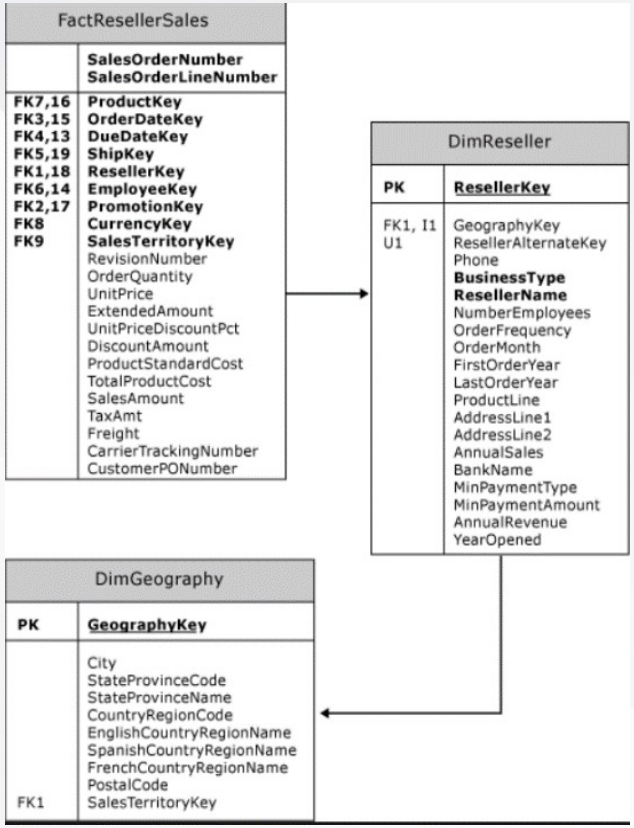
De acuerdo con Microsoft (2022), las tablas de dimensiones diseñadas en Microsoft SQL Server pueden clasificarse de la siguiente manera:



Dimensión basada en un diseño de **esquema en estrella**.

En donde cada dimensión se relaciona con un elemento de la tabla de hechos, mediante una relación de clave principal.

Explicación



Dimensión basada en un diseño de **esquema de copo de nieve**.

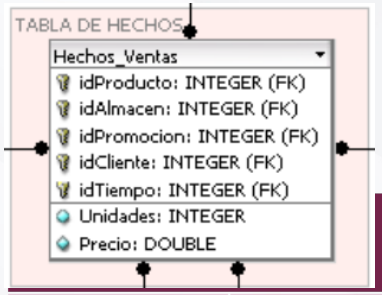
Se prefiere para definir atributos de una dimensión en varias tablas de hechos, con el objetivo de reducir espacio y aprovechar la redundancia para hacer referencia a una misma dimensión a partir de uno o más hechos.

Explicación

Diseño de tablas de hechos

La creación de tablas de hechos requiere de un estándar de nombres para orientar al usuario, mientras que el objetivo de la tabla de objetos tradicionales es que los datos se expongan directamente a los usuarios y es indispensable que los nombres sean intuitivos.

Ejemplo:



Secciones de una tabla de hechos		
Fundamental	Medidas del hecho.	Son valores cuantitativos que representan algún elemento del negocio.
	Referencias a dimensiones.	Contextualizan los hechos para determinar sus dimensiones y tener columnas que se relacionan con tablas de dimensión.
Opcional	Metadata.	Es información que describe o detalla a otros datos.

Actividad

Con base en lo descrito en el tema, reflexiona sobre las siguientes preguntas:

01 ¿Cuál es la diferencia entre una tabla de hechos y una de dimensiones?

02 Teniendo como contexto el sitio de Amazon para realizar compras, ¿qué tipo de tabla de dimensión se utiliza?

Cierre

Un almacén de datos brinda a las organizaciones la facilidad de contener toda la información relevante (actual e histórica) del negocio dentro de una misma base de datos.

Para generar **un buen almacén** de datos, se debe:

- Conocer a detalle los datos y el hecho sobre los cuales se va a trabajar.
- Tener un modelo dimensional diseñado, haciendo uso de las tablas de hechos y dimensiones como herramientas fundamentales.

Bibliografía

• Microsoft. (2022). *Dimensiones: introducción*. Recuperado de https://docs.microsoft.com/es-es/analysis-services/multidimensional-models-olap-logical-dimension-objects/dimensions-introduction?view=asallproducts-allversions

• Oracle. (2022). *¿Qué es un almacén de datos?* Recuperado de https://www.oracle.com/mx/database/what-is-a-data-warehouse/

• Rivas, A. (2021). *Tabla de dimensiones*. Recuperado de https://www.muytecnologicos.com/diccionario-tecnologico/tabla-de-dimensiones

Universidad Tecmilenio

Consulta en Microsoft SQL Server®

Creación de una solución de ETL con SQL Server Integration Services (SSIS)

Semana 9

Introducción

La empresa altruista “Sun for Fun” dedicada al mantenimiento de áreas de recreo en zonas costeras ha estado recopilando datos sobre las áreas afectadas.

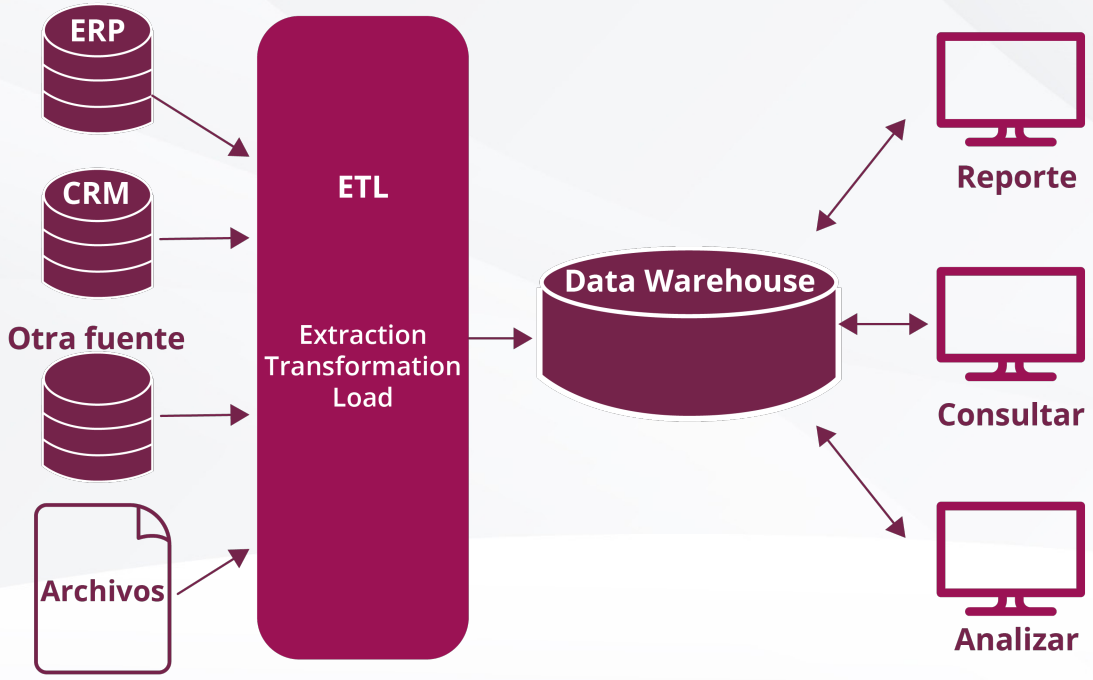
Se te ha contratado para elaborar una base de datos y te das cuenta de que la información recabada es demasiada, por lo que investigas la manera más eficiente de subir la información al repositorio que crearás. Después de un momento de investigación, encuentras que la solución óptima es utilizar el **proceso ETL**.

Explicación

Introducción a ETL con SSIS

La información es uno de los recursos más valiosos para las organizaciones, sin embargo, su obtención provenie de diferentes fuentes.

> Se puede hacer uso del proceso extraer, transformar y cargar, o ETL, por sus siglas en inglés (*extraction*, *transformation* y *load*), el cual es una canalización de datos que se usa para recopilar datos de varios orígenes (Microsoft, 2022a).



Universidad Tecmilenio

Explicación

Fases del proceso ETL

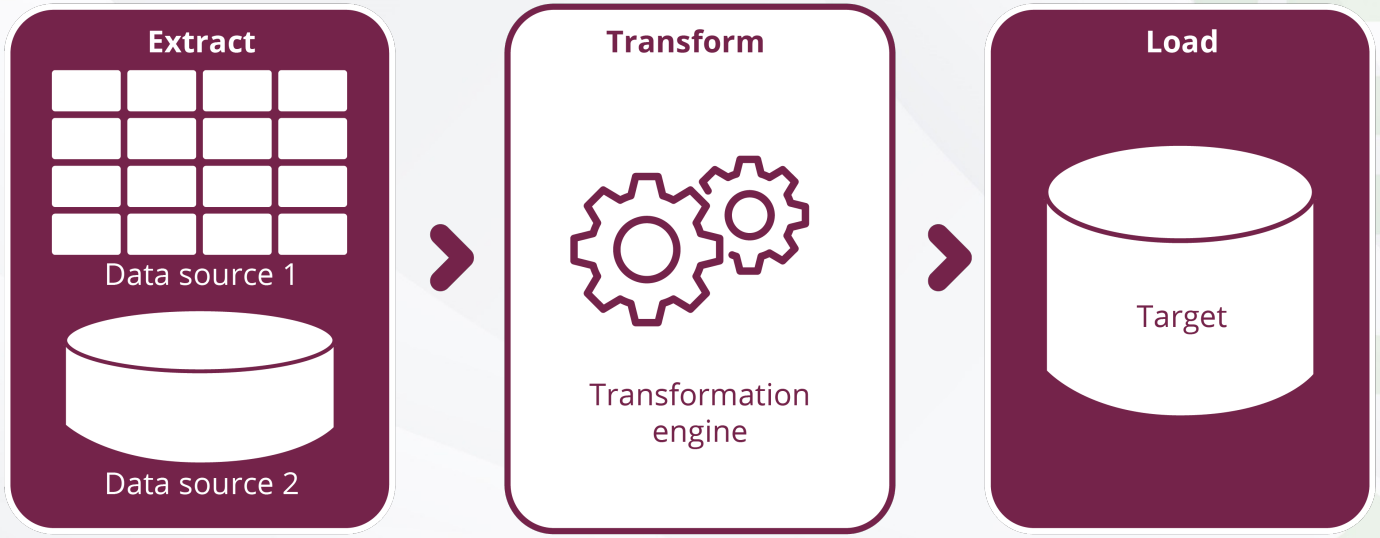


Figura 1. Fases del proceso ETL (extracción, transformación y carga).

Explicación

Exploración de datos de origen

Al explorar los datos en el diseñador de vistas del origen de datos, puedes ver el contenido seleccionado de cada columna de datos de una tabla, vista o consulta con nombre (Microsoft, 2022b).

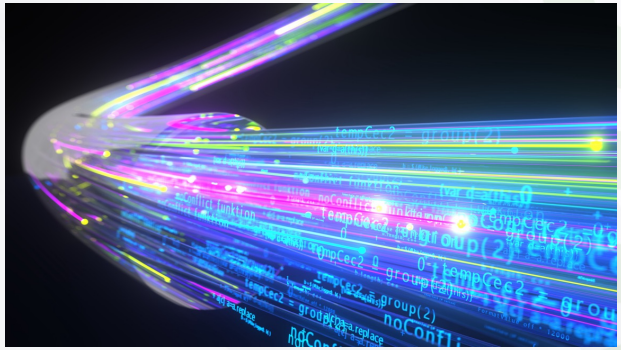
Diferentes fuentes de datos que pueden ser extraídas, transformadas y almacenadas en nuestra base de datos de SQL Server Integration Services:

Origen de datos

- Excel y MS Access
- Oracle
- XML y servicios web
- Archivos planos
- ODBC
- Teradata
- Otras fuentes heterogéneas

Explicación

Implementación del flujo de datos



El flujo de datos de **SSIS** se implementa como una canalización lógica, donde los datos fluyen desde una o más fuentes.

La tarea de flujo de datos se utiliza para transferir datos de un origen a un destino.

Se debe recordar que el origen y el destino pueden ser de varios tipos.

SSIS realiza las transformaciones en la memoria, por lo que es mucho más rápido que leer y escribir los datos en una unidad.

Las transformaciones permiten realizar cambios en la información conforme esta es reubicada de un origen a un destino.

Actividad

Con base en lo descrito en el tema, reflexiona sobre las siguientes preguntas:

01 ¿Qué tipos de archivos crees que utiliza una institución para el resguardo de su información?

02 ¿En qué situación utilizarías un proceso de ETL?

Como se puede apreciar, el proceso ETL (extraer, transformar y cargar) facilita el proceso de obtención de datos contenidos en diversas fuentes.

Gracias a la plataforma SQL Server **Integration Services (SSIS),** la transformación e integración de datos se vuelve más simple, pues logra cumplir con los principales objetivos del proceso ETL.

Bibliografía

• Microsoft. (2022a). *Extracción, transformación y carga de datos (ETL)*. Recuperado de https://docs.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

• Microsoft. (2022b). *Explorar datos en una vista del origen de datos (Analysis Services)*. Recuperado de https://docs.microsoft.com/es-es/analysis-services/multidimensional-models/explore-data-in-a-data-source-view-analysis-services?view=asallproducts-allversions