**MÓDULO 5 - Modelos de datos**

**Objetivos de aprendizaje**

**Identificar y definir los principales modelos y esquemas de bases de datos**

**Evaluar qué modelo de diseño de base de datos y qué esquema son apropiados para los distintos datos**

**Debatir alternativas de modelos de datos que sean óptimas, eficaces y se ajusten a los requisitos de elaboración de informes teniendo en cuenta el tamaño y el crecimiento actual de los datos**

**Comenzando con bases de datos, esquemas y bases de datos**

**Modelado de datos, patrones de diseño y esquemas**

**Una base de datos OLTP es aquella que ha sido optimizada para el procesamiento de datos en lugar del análisis.**

Un tipo de sistema de destino es un data mart, que es una base de datos orientada a temas que puede ser un subconjunto de un almacén de datos más grande. Otra posibilidad es utilizar una base de datos OLAP o de procesamiento analítico en línea. Esta es una herramienta que ha sido optimizada para el análisis además del procesamiento y puede analizar datos de múltiples bases de datos.

Hay dos tipos de datos: no estructurados y estructurados. Los datos no estructurados no están organizados de ninguna manera fácilmente identificable. Los datos estructurados se han organizado en un determinado formato, como filas y columnas.

**Para crear el modelo de datos, los analistas de BI suelen utilizar lo que se conoce como un patrón de diseño. El patrón de diseño es una solución que utiliza medidas y hechos relevantes para crear un modelo que respalde las necesidades comerciales.**

**Obtener los hechos con modelos dimensionales**

Una base de datos relacional contiene una serie de tablas que se pueden conectar para formar relaciones. Estas relaciones se establecen utilizando claves primarias y foráneas. Veremos un ejemplo.

El ID de sucursal es la clave principal en la tabla de concesionarios de automóviles, pero es la clave externa en la tabla de detalles del producto. Esto conecta estas dos tablas directamente. Chasis es la clave principal en la tabla de detalles del producto y la clave externa en la tabla de piezas de reparación.

Una clave principal es un identificador en la base de datos que hace referencia a una columna en la que cada valor es único.

Una clave principal es un identificador en una base de datos que hace referencia a una columna o un grupo de columnas en el que cada fila identifica de forma única cada registro de la tabla. En esta base de datos tenemos claves primarias en cada tabla. ID de sucursal, chasis e ID de pieza.

**Una clave externa es un campo dentro de una tabla de base de datos que es una clave principal en otra tabla**. Las claves principales de cada tabla también aparecen como claves externas en otras tablas.

Básicamente, se puede usar una clave principal para imponer restricciones en la base de datos que aseguren que los datos en una columna específica sean únicos al identificar específicamente un registro en una tabla de base de datos relacional.

***Solo puede existir una clave principal en una tabla, pero una tabla puede tener muchas claves externas.***

**Un modelo dimensional es un tipo de modelo relacional que se ha optimizado para recuperar rápidamente datos de un almacén de datos. Los modelos dimensionales se pueden dividir en hechos para medir y dimensiones que agregan atributos para el contexto**.

*En un modelo dimensional, un hecho es una medida o métrica.* Por ejemplo, un número de ventas mensuales podría ser un hecho y una dimensión es una información que brinda más detalles y contexto con respecto a ese hecho. Es el quién, qué, dónde, cuándo, por qué y cómo. Entonces, si nuestro número de ventas mensuales es el hecho, entonces las dimensiones podrían ser información sobre cada venta, incluido el cliente, la ubicación de la sucursal y qué productos se vendieron. A continuación, consideremos los atributos*.*

*Un atributo es una característica o calidad de los datos utilizados para etiquetar las columnas de la tabla.* En los modelos dimensionales, los atributos funcionan de la misma manera.

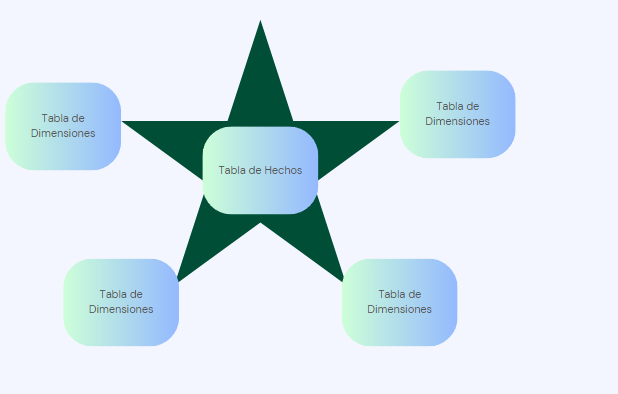
Un atributo es una característica o cualidad que se puede utilizar para describir una dimensión. Entonces, una dimensión proporciona información sobre un hecho y un atributo proporciona información sobre una dimensión.

**Una tabla de hechos contiene medidas o métricas relacionadas con un evento en particular**. **Esta es la tabla principal que contiene los hechos y su relación con las dimensiones. Básicamente, cada fila de la tabla de hechos representa un evento. Toda la tabla podría agregar varios eventos, como ventas en un día.**

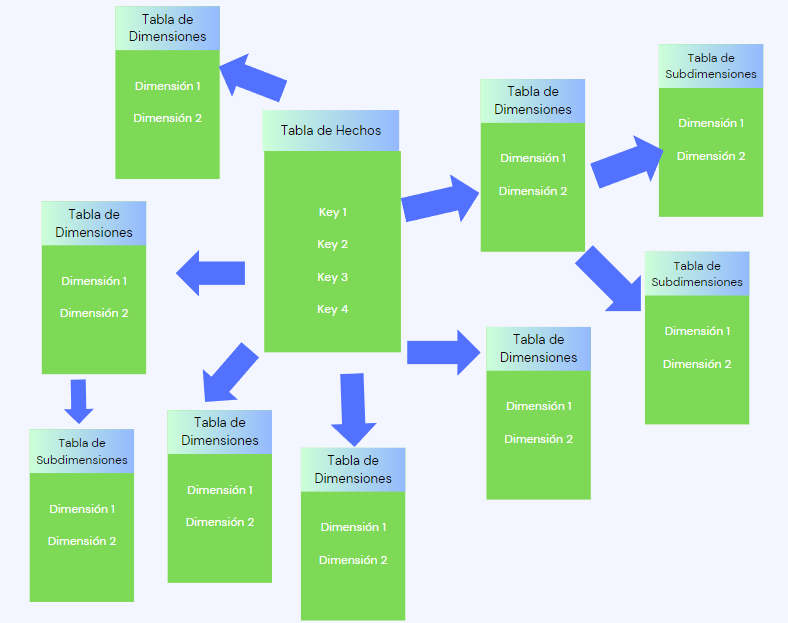
**Una tabla de dimensiones es donde se almacenan los atributos de las dimensiones de un hecho. Estas tablas se unen a la tabla de hechos apropiada utilizando la clave externa**. Esto da significado y contexto a los hechos. Así es como se conectan las tablas en el modelo dimensional.

**Modelos dimensionales con esquemas de estrella y copo de nieve**

Hay varios esquemas comunes que puede encontrar en la inteligencia empresarial, incluidos los esquemas de estrella, copo de nieve y desnormalizados o NoSQL.



Los esquemas de estrella y copo de nieve son algunas de las iteraciones más comunes de un modelo dimensional real en la práctica. *Un esquema en estrella es un esquema que consta de una tabla de hechos que hace referencia a cualquier número de tablas de dimensiones.* Como su nombre indica, este esquema tiene forma de estrella. Cada una de las tablas de dimensiones está conectada a la tabla de hechos en el centro. *Los esquemas en estrella están diseñados para monitorear datos en lugar de analizarlos.* De esta forma, permiten a los analistas procesar rápidamente los datos. Por lo tanto, son ideales para la entrega de información a gran escala y hacen que la salida sea más eficiente debido al número limitado de tablas y las relaciones directas claras.



Luego tenemos esquemas de copo de nieve, que tienden a ser más complicados que los esquemas de estrella, pero el principio es el mismo. Un esquema de copo de nieve es una extensión de un esquema de estrella con dimensiones adicionales y a menudo, subdimensiones. Estas dimensiones y subdimensiones descomponen el esquema en tablas aún más específicas, creando un patrón de copo de nieve.

.

**Diseñar sistemas de bases de datos eficientes con esquemas**

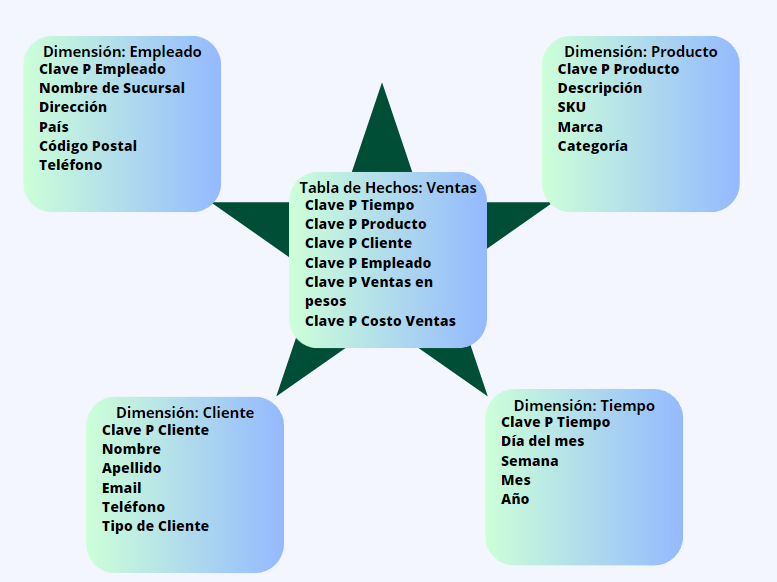
Tipos de esquemas

Estrella y copo de nieve

En los esquemas de copos de nieve, las tablas de dimensiones están normalizadas. **Esto divide los datos en tablas adicionales, lo que hace que los esquemas sean un poco más complejos.**

Un esquema en estrella es un esquema que consta de una o más tablas de hechos que hacen referencia a cualquier número de tablas de dimensiones. Como su nombre indica, este esquema tiene forma de estrella. Este tipo de esquema es ideal para la entrega de información a gran escala y hace que la salida de lectura sea más eficiente. También clasifica los atributos en hechos y atributos de dimensiones descriptivas (ID del producto, nombre del cliente, fecha de venta).

Aquí hay un ejemplo de un esquema de estrella:



En este ejemplo, esta empresa utiliza un esquema de estrella para realizar un seguimiento de la información de ventas dentro de sus tablas. Esto incluye:

Información al cliente

Información del Producto

El momento en que se realiza la venta.

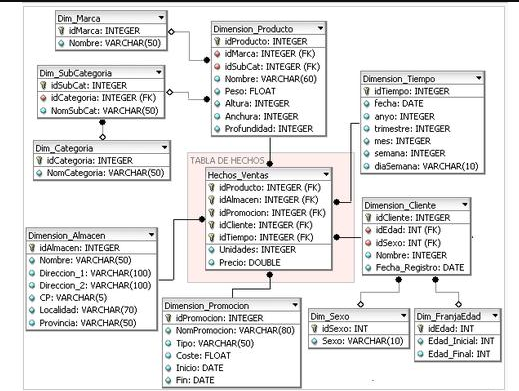
Información del empleado

Todas las tablas de dimensiones se vinculan a la tabla de hechos Ventas en el centro, lo que confirma que se trata de un esquema en estrella.

Un esquema de copo de nieve es una extensión de un esquema de estrella con dimensiones adicionales y, a menudo, subdimensiones. Estas dimensiones y subdimensiones crean un patrón de copos de nieve.

Los esquemas de copo de nieve son un tipo de organización diseñado para el procesamiento de datos ultrarrápido.

A continuación se muestra un ejemplo de un esquema de copo de nieve:



Se quiere diseñar un esquema de copo de nieve que contenga información sobre ventas. Se comienza en el centro con la tabla de hechos, que contiene:

ID PRODUCTO

ID ALMACÉN

ID CLIENTE

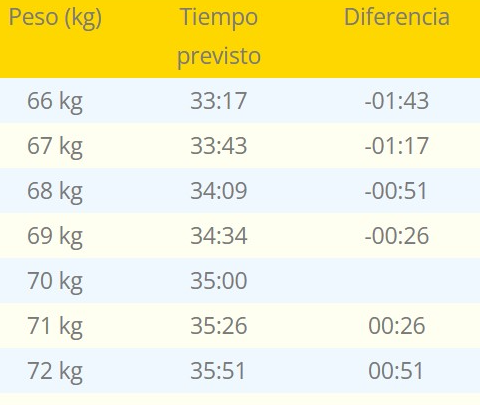
ID TIEMPO

Esta tabla de hechos se ramifica en varias tablas de dimensiones e incluso en subdimensiones. Las tablas de dimensiones desglosan múltiples detalles, como promociones y marcas,

Modelo plano

Los esquemas planos son sistemas de bases de datos extremadamente simples con una sola tabla en la que cada registro está representado por una sola fila de datos. Las filas están separadas por un delimitador, como una columna, para indicar las separaciones entre registros. Los modelos planos no son relacionales; no pueden capturar relaciones entre tablas o elementos de datos. Debido a esto, los modelos planos se usan con mayor frecuencia como una fuente potencial dentro de un sistema de datos para capturar datos menos complejos que no necesitan actualizarse.

Aquí hay una tabla plana de pesos y tiempo previsto para un ejercicio:



Estos datos no van a cambiar.

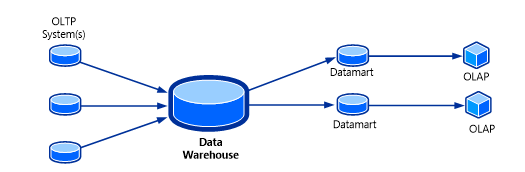
Como analista de BI, es posible que se encuentre con modelos planos en fuentes de datos que desee integrar en sus propios sistemas. Reconocer que estos no son ya modelos relacionales es útil al considerar la mejor manera de incorporar los datos en las tablas de destino.

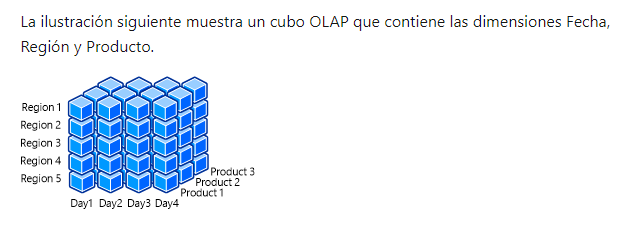
.



**Diferentes tipos de datos, diferentes bases de datos**

**Una migración de base de datos implica mover datos de una plataforma de origen a otra base de datos de destino. Durante una migración, los usuarios hacen la transición de los esquemas de base de datos actuales a un nuevo estado deseado. Esto podría implicar agregar tablas o columnas, dividir campos, eliminar elementos, cambiar tipos de datos u otras mejoras. El proceso de migración de la base de datos a menudo requiere numerosas fases e iteraciones, así como muchas pruebas.**

Analizaremos varios tipos de bases de datos, incluidas OLTP, OLAP, basadas en filas, en columnas, distribuidas, de alojamiento único, de almacenamiento separado y de cómputo y bases de datos combinadas. Las dos primeras tecnologías de bases de datos que se iban a explorar, los sistemas OLTP y OLAP, se basan en cómo se procesan los datos. 



*Las bases de datos OLTP administran la modificación de la base de datos y funcionan con el software tradicional del sistema de administración de bases de datos. Estos sistemas están diseñados para almacenar transacciones de manera efectiva y ayudar a garantizar la coherencia. Un ejemplo de una base de datos OLTP sería una librería en línea.* Si dos personas agregan el mismo libro a su carrito, pero solo hay una copia, la persona que complete el proceso de compra primero obtendrá el libro. Y el sistema OLTP garantiza que no se vendan más copias de las que hay en stock. Las bases de datos OLTP están optimizadas para leer, escribir y actualizar filas individuales de datos para garantizar que los procesos comerciales se desarrollen sin problemas.

. *Los sistemas OLAP extraen datos de múltiples fuentes al mismo tiempo para analizarlos y proporcionar información comercial clave*. Volviendo a nuestra librería en línea, un sistema OLAP podría extraer datos sobre compras de clientes de múltiples almacenes de datos, con el fin de crear páginas de inicio personalizadas para los clientes en función de sus preferencias. Los sistemas de bases de datos OLAP permiten a las organizaciones abordar sus necesidades analíticas a partir de una variedad de fuentes de datos.

## **OLAP vs OLTP**

|  |  |  |
| --- | --- | --- |
| **Tecnología de base de datos** | **Descripción** | **Uso** |
| OLAP | *Los sistemas de procesamiento analítico en línea* (OLAP) son bases de datos que se han optimizado principalmente para el **análisis**. | Proporcionar acceso de usuario a los datos de una variedad de sistemas de origen  Utilizado por BI y otros profesionales de datos para respaldar los procesos de toma de decisiones  Analizar datos de múltiples bases de datos.  Obtener información procesable de los datos entregados a las tablas de informes |
| OLTP | *Los sistemas de procesamiento de transacciones en línea* (OLTP) son bases de datos que se han optimizado para el **procesamiento de datos** en lugar del análisis. | Almacenar datos de transacciones  Utilizado por empleados de cara al cliente o aplicaciones de autoservicio del cliente  Lea, escriba y actualice filas individuales de datos  Actuar como sistemas de origen de los que se pueden extraer pipelines para su análisis. |

**Distribuido versus de alojamiento único**

|  |  |  |
| --- | --- | --- |
| **Tecnología de Bases de Datos** | **Descripción** | **Uso** |
| Distribuido | Las bases de datos distribuidas son colecciones de sistemas de datos distribuidos en múltiples ubicaciones físicas. | Se expande fácilmente para abordar las necesidades comerciales crecientes o de mayor escala.  Accedido desde diferentes redes  Más fácil de proteger que un sistema de base de datos de alojamiento único |
| Unidireccional | Las bases de datos de alojamiento único son bases de datos en las que todos los datos se almacenan en la misma ubicación física. | Los datos almacenados en una sola ubicación son más fáciles de acceder y coordinar entre equipos  Reduce la redundancia de datos  Más barato de mantener que los sistemas más grandes y complejos |

Almacenamiento y cómputo separados vs combinados

|  |  |  |
| --- | --- | --- |
| **Tecnología de Base de Datos** | **Descripción** | **Uso** |
| Almacenamiento y cómputo separados | Los sistemas informáticos y de almacenamiento separados son bases de datos en las que los datos menos relevantes se almacenan de forma remota y los datos relevantes se almacenan localmente para su análisis. | Ejecute consultas analíticas de manera más eficiente porque el sistema solo necesita procesar los datos más relevantes  Escale los recursos informáticos y los sistemas de almacenamiento por separado en función de las necesidades personalizadas de su organización |
| Computación y almacenamiento combinados | **Los sistemas combinados son sistemas de bases de datos que almacenan y analizan datos en el mismo lugar.** | Configuración tradicional que permite a los usuarios acceder a todos los datos posibles a la vez  Los recursos de almacenamiento y computación están vinculados, por lo que la administración de recursos es sencilla |

**Eligiendo la base de datos correcta**

**La forma de los datos**

*Un almacén de datos (Data Warehouse) es un tipo específico de base de datos que consolida datos de múltiples sistemas de origen para lograr consistencia, precisión y acceso eficiente a los datos.* Los data warehouse se utilizan para respaldar la toma de decisiones basada en datos. A menudo, estos sistemas son administrados por especialistas en almacenamiento de datos, pero los analistas de BI pueden ayudar a diseñarlos cuando se trata de diseñar un almacenamiento para datos.

Hay algunas cosas importantes que el analista de BI considerará: necesidades comerciales, la forma y el volumen de los datos y qué modelo seguirá el almacén de datos*.*

*Las necesidades comerciales son las preguntas que la organización quiere responder o los problemas que quieren resolver.*

A continuación, exploremos la forma y el volumen de los datos del sistema de origen. Por lo general, la forma de los datos se refiere a las filas y columnas de las tablas dentro del almacén y cómo se distribuyen. El volumen de datos actual y futuro también cambia la forma en que se diseña el almacén y el modelo que seguirá el almacén incluye todas las herramientas y restricciones del sistema, como la propia base de datos y cualquier herramienta de análisis que se incorporará al sistema.

La información de estas tablas brinda más contexto a las tablas de hechos que registran los procesos y eventos comerciales.

Comprender la lógica detrás del diseño del almacén de datos ayudará a desarrollar procesos y sistemas efectivos, se trabajará más con esquemas de bases de datos y se aprenderá cómo los datos se extraen del almacén desde otras fuentes.

**Cuatro elementos clave de los esquemas de bases de datos**

Aquí hay una lista de verificación de los elementos comunes que debe incluir un esquema de base de datos:

* **Los datos relevantes: el esquema describe cómo se modelan y dan forma a los datos dentro de la base de datos y debe abarcar todos los datos que se describen.**
* **Nombres y tipos de datos para cada columna: incluir nombres y tipos de datos para cada columna en cada tabla dentro de la base de datos.**
* **Formato consistente: Garantizar un formato consistente en todas las entradas de datos. Cada entrada es una instancia del esquema, por lo que debe ser coherente.**
* **Claves únicas: el esquema debe usar claves únicas para cada entrada dentro de la base de datos. Estas claves crean conexiones entre las tablas y permiten a los usuarios combinar datos relevantes de toda la base de datos.**

Conclusiones clave

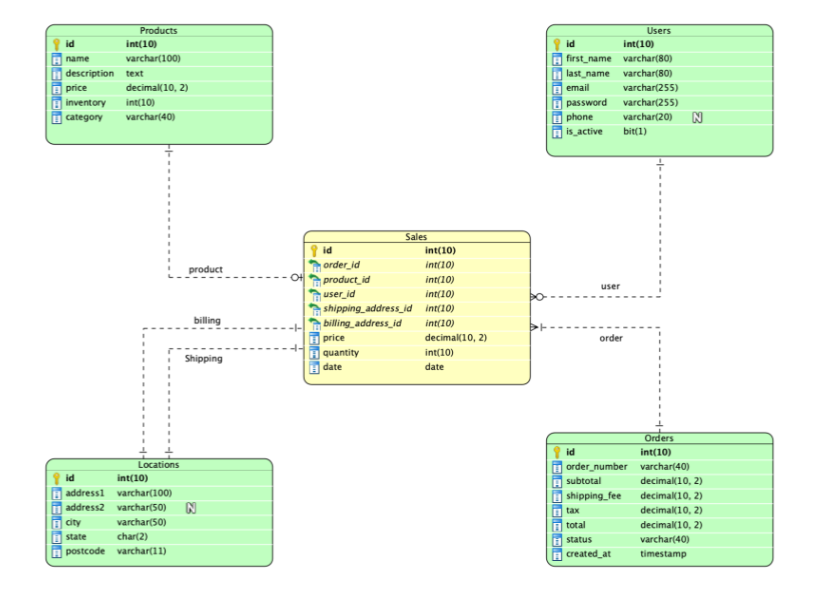
A medida que recibe más datos o cambian las necesidades comerciales, es posible que las bases de datos y los esquemas también deban cambiar. La optimización de la base de datos es un proceso iterativo, lo que significa que es posible que se deba verificar el esquema varias veces durante la vida útil de la base de datos.

**Revisar un esquema de base de datos ACTIVIDAD**

Esquema de base de datos

Conexión Electrónica está por implementar venta online para su nueva línea de productos de oficina en casa. Si tiene éxito, los responsables de la toma de decisiones de la empresa también planean poner el resto de sus productos en línea. La empresa contrató a Mia, una ingeniera sénior de BI, para ayudar a diseñar su almacén de datos. La base de datos necesitaba almacenar datos de pedidos para análisis e informes, y el gerente de ventas necesitaba generar informes rápidamente para rastrear las ventas para poder determinar el éxito del sitio.

A continuación se muestra un diagrama del esquema de la base de datos sales\_warehouse diseñada por Mia. Contiene diferentes símbolos y conectores que representan dos piezas importantes de información: las tablas principales dentro del sistema y las relaciones entre estas tablas.



El esquema de la base de datos sales\_warehouse contiene cinco tablas: Ventas, Productos, Usuarios, Ubicaciones y Pedidos, que están conectadas mediante claves. Las tablas contienen de cinco a ocho columnas (o atributos) que varían en tipo de datos. Los tipos de datos incluyen varchar o char (o carácter), entero, decimal, fecha, texto (o cadena), marca de tiempo, bit y otros tipos según el sistema de base de datos elegido.

**Revisar el esquema de la base de datos**

Para comprender el esquema de una base de datos, es útil comprender el propósito de usar ciertos tipos de datos y las relaciones entre los campos. Las respuestas a las siguientes preguntas justifican por qué Mia diseñó el esquema de Conexión Electrónica de esta manera:

¿Qué tipo de esquema de base de datos es este? ¿Por qué se seleccionó este tipo de base de datos?

Mia diseñó la base de datos con un esquema de estrella porque Conexión Electrónica está usando esta base de datos para informes y análisis. Los beneficios del esquema en estrella incluyen consultas más simples, lógica de informes comerciales simplificada, ganancias en el rendimiento de las consultas y agregaciones rápidas.

¿Qué convenciones de nomenclatura se utilizan para las tablas y los campos? ¿Hay algún beneficio de usar estas convenciones de nomenclatura?

Este esquema utiliza una convención de nomenclatura de mayúsculas y minúsculas. En el caso de snake case, los guiones bajos reemplazan los espacios y la primera letra de cada palabra está en minúscula. El uso de una convención de nomenclatura ayuda a mantener la coherencia y mejora la legibilidad de la base de datos. Dado que el caso de snake case para tablas y campos es un estándar de la industria, Mia lo usó en la base de datos.

¿Cuál es el propósito de usar los campos decimales en los elementos de datos?

Para los campos relacionados con el dinero, existen errores potenciales al calcular precios, impuestos y tarifas. Es posible que tenga valores que sean técnicamente imposibles, como un valor de $0,001, cuando el valor más pequeño de la moneda local es un centavo o $0,01. Para mantener los valores consistentes y evitar errores acumulados, Mia usó un tipo de datos decimal (10,2), que solo mantiene los dos últimos dígitos después del punto decimal.

Nota: Otros valores numéricos, como el tipo de cambio y las cantidades, pueden necesitar lugares decimales adicionales para minimizar las diferencias de redondeo en los cálculos. Además, otros tipos de datos pueden ser más adecuados para otros campos. Para rastrear cuándo se crea un pedido (created\_at), puede usar un tipo de datos de marca de tiempo. Para otros campos con varios tamaños de texto, puede usar varchar.

¿Cuál es el propósito de cada clave externa y principal en la base de datos?

Mia diseñó la tabla Ventas con un ID de clave principal e incluyó claves externas en las otras tablas para hacer referencia a las claves principales. Las claves foráneas deben ser del mismo tipo de datos que sus correspondientes claves primarias. Las claves primarias identifican de manera única y precisa un registro en una tabla, y las claves externas establecen referencias de integridad desde esa clave principal a los registros en otras tablas.

|  |  |
| --- | --- |
| **Tabla de Ventas Clave primaria y foránea** | **Tabla Asociada** |
| **pedido\_id** | **Tabla Pedidos** |
| **producto\_id** | **Tabla Producto** |
| **usuario\_id** | **Tabla Usuario** |
| **dirección\_envío\_id** | **Tabla Localidades** |
| **dirección\_facturación\_id** | **Tabla Localidades** |

Actividad

Explicar los beneficios y potenciales inconvenientes del diseño en estrella en una base de datos