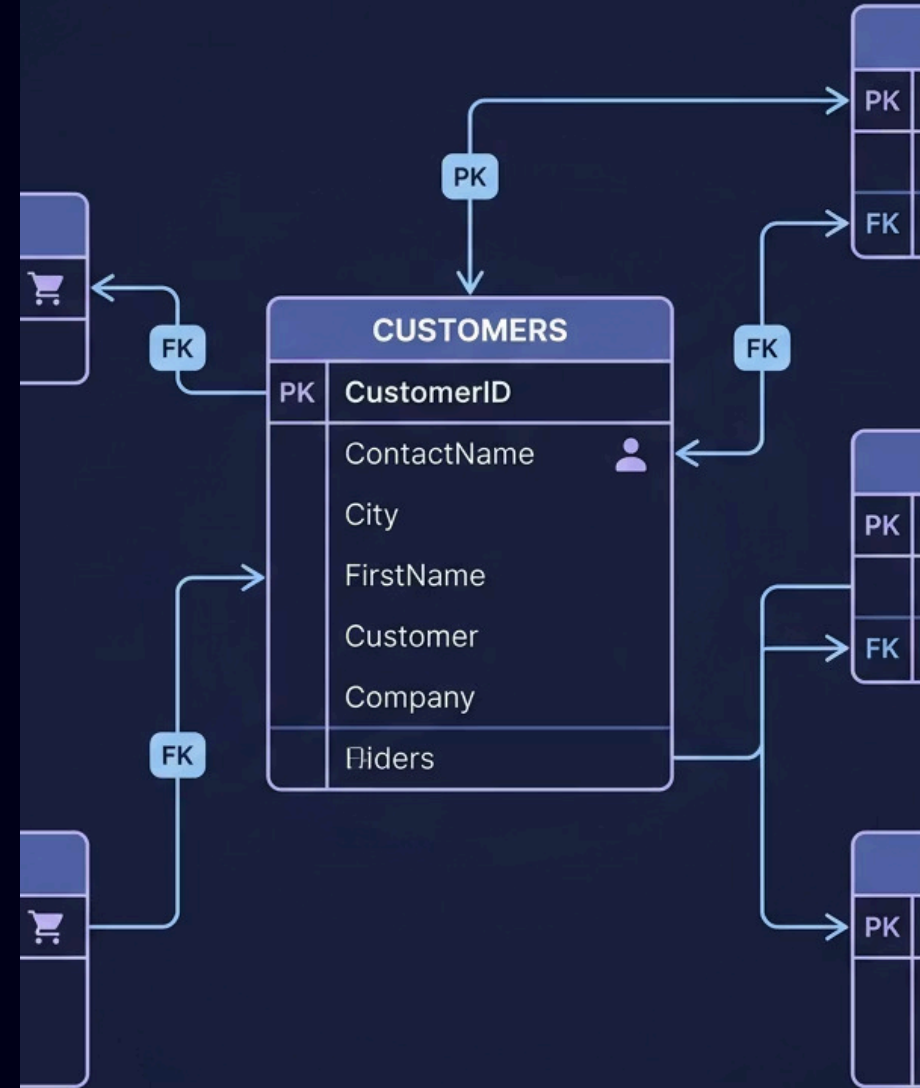




Modelado de datos: fundamentos y mejores prácticas

El modelado de datos es un paso crítico para hacer que los datos sean útiles para el negocio, aunque a menudo se pasa por alto. Una arquitectura de datos bien construida debe reflejar los objetivos y la lógica empresarial de la organización.



COMPARACIÓN DE ESTRUCTURAS DE DATOS



ESTRUCTURA DE DATOS ORGANIZADO

Acceso, Eficiencia, Coherencia



ESTRUCTURA DE DATOS DESORGANIZADO

Dificultad de Acceso, Ineficiencia

¿Qué es un modelo de datos?

Un modelo de datos representa la forma en que los **datos** se relacionan con el **mundo real**. Refleja cómo los datos deben estructurarse y estandarizarse para representar mejor los procesos, definiciones, flujos de trabajo y lógica de su organización.

Un buen modelo de datos captura cómo fluyen naturalmente la comunicación y el trabajo dentro de su organización. En contraste, un modelo de datos deficiente (o inexistente) es aleatorio, confuso e incoherente.

La Importancia del modelado de datos

Coherencia empresarial

Garantiza que la lógica y las reglas del negocio se traduzcan correctamente en la capa de datos.

Resultados comerciales

Un buen modelo de datos debe correlacionarse con decisiones empresariales impactantes.

Definiciones claras

Ayuda a definir conceptos que pueden significar cosas diferentes para distintos departamentos (ej: ¿qué es un "cliente"?).

Evolución histórica del modelado de datos

1

Años 70-80

Técnicas de normalización desde los primeros días de los RDBMS.

2

Años 90

Desarrollo de técnicas de modelado para data warehousing.

3

2010-2015

Declive del modelado formal con el auge de data lakes 1.0, NoSQL y big data.

4

Actualidad

Regreso al modelado riguroso impulsado por la gobernanza de datos y la calidad de datos.

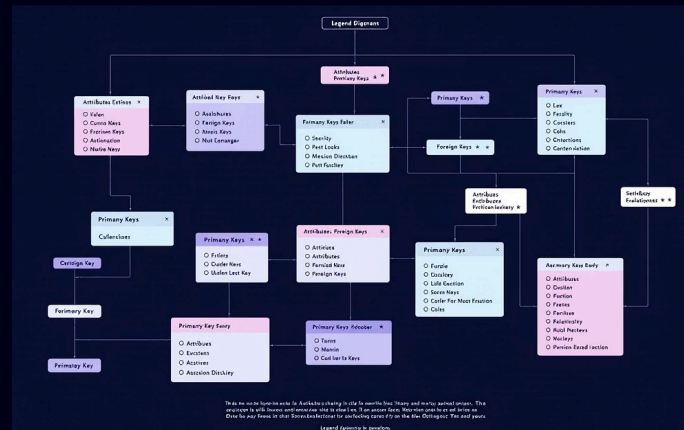
Tipos de modelos de datos

Al modelar datos, se avanza desde conceptos abstractos hasta implementaciones concretas a través de tres niveles principales:



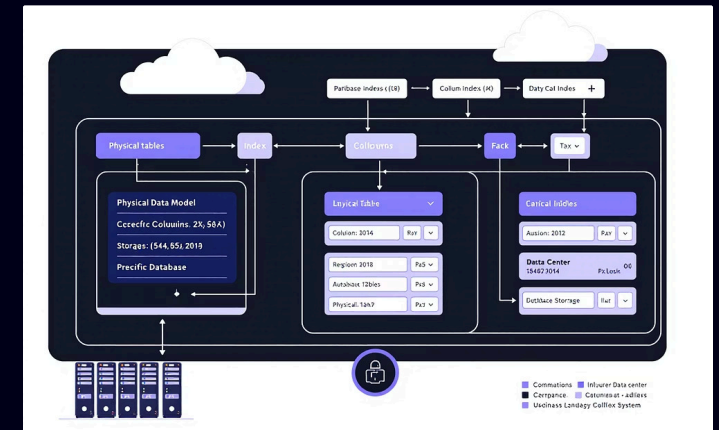
Modelo Conceptual

Contiene la lógica y las reglas de negocio, y describe los datos del sistema de una manera de alto nivel e independiente de la tecnología.



Modelo Lógico

Detalla cómo se implementará el modelo conceptual, añadiendo tipos de datos, claves primarias y foráneas, pero aún sin referencia a un sistema de base de datos específico.



Modelo Físico

Define cómo se implementará el modelo lógico en un sistema de base de datos específico, incluyendo detalles de almacenamiento y optimización para el rendimiento.

El modelo conceptual

El modelo conceptual contiene la lógica y reglas de negocio y describe las entidades y relaciones entre ellas.

Al crear un modelo conceptual, es útil visualizarlo en un diagrama de entidad-relación (ER), una herramienta estándar para visualizar las relaciones entre varias entidades en sus datos (pedidos, clientes, productos, etc.).

Entidades

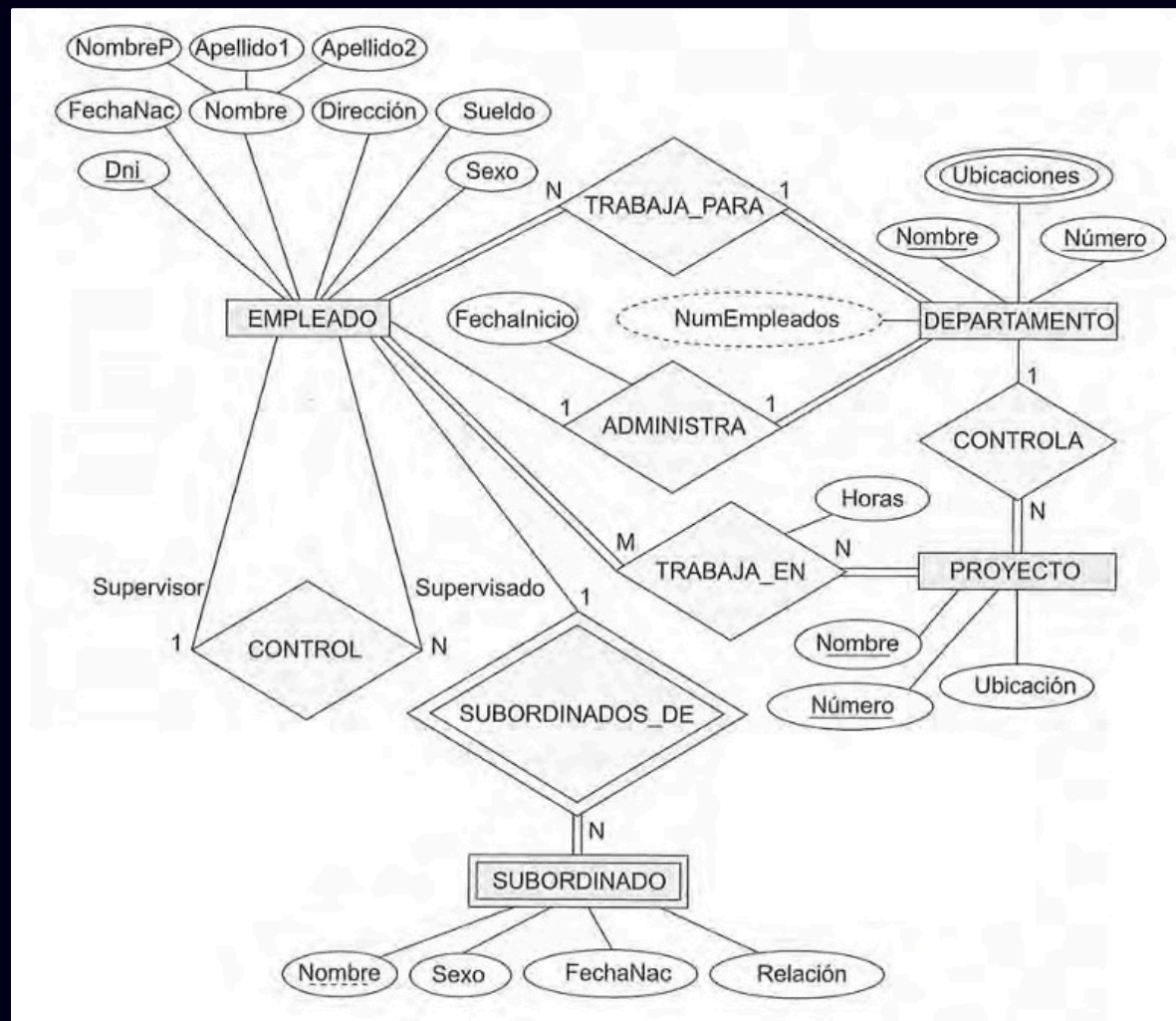
Los objetos o conceptos principales que se modelan, como clientes, productos, órdenes, etc.

Atributos

Las características o propiedades de cada entidad, como nombre, precio, fecha, etc.

Relaciones

Cómo se conectan las diferentes entidades, como "un cliente realiza varias órdenes" o "un producto pertenece a una categoría".



Cardinalidades

Describe **cuántas instancias de una entidad se relacionan con instancias de otra** en un modelo de datos.

Por ejemplo, un (1) departamento puede tener muchos (N) empleados.

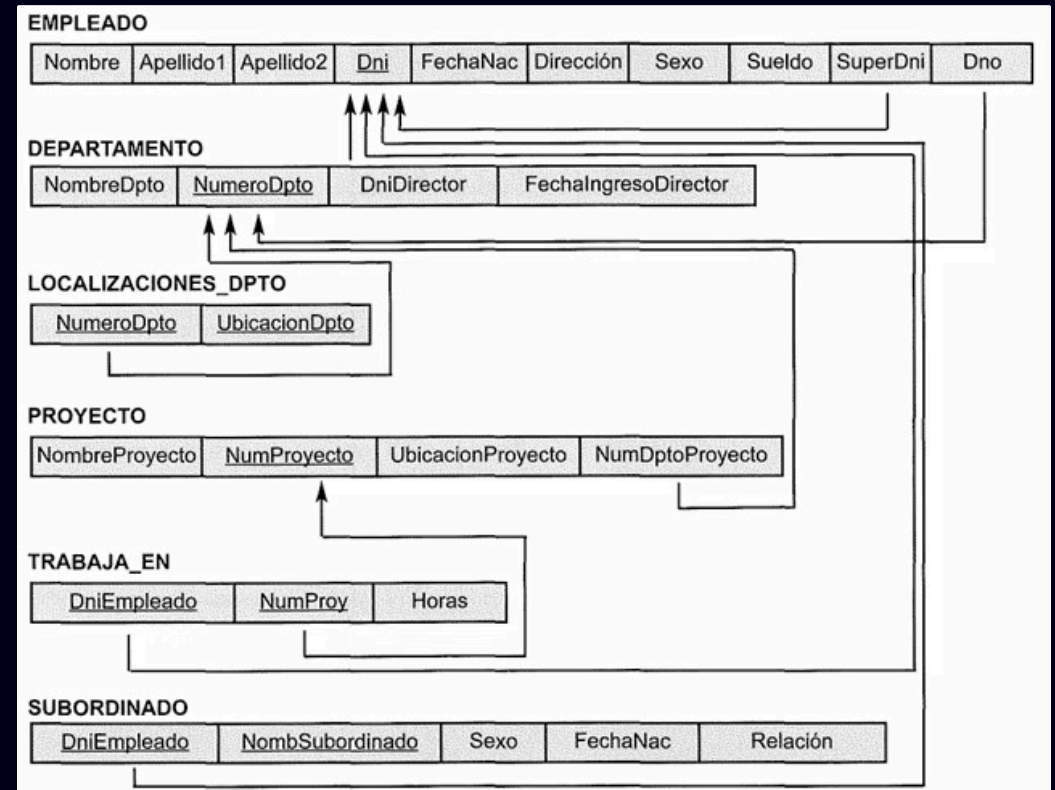
El modelo lógico

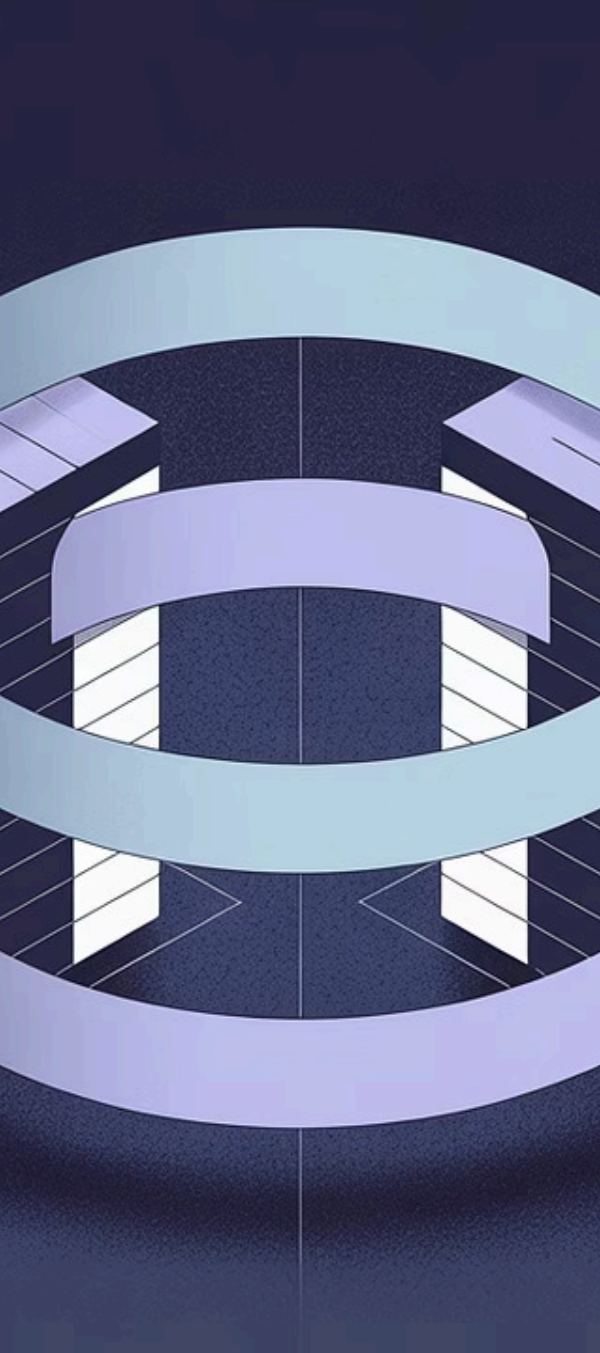
El modelo lógico profundiza en el modelo conceptual, añadiendo detalles prácticos para su implementación.

En primer lugar se define el tipo de base de datos: relacional, documental, grafos, etc.

Luego se traducen las entidades y relaciones del modelo conceptual.

Para un modelo relacional se definen las tablas y claves primarias y foráneas, esenciales para establecer relaciones y asegurar la integridad de los datos sin depender de un sistema de base de datos específico.





Normalización

La normalización es una práctica de modelado de datos que impone un control estricto sobre las relaciones de tablas y columnas dentro de una base de datos. El objetivo es eliminar la redundancia de datos y garantizar la integridad referencial.

Fue introducida por Edgar Codd en los años 70 con cuatro objetivos principales: eliminar dependencias indeseables, reducir la necesidad de reestructuración, hacer el modelo más informativo y neutralizar las estadísticas de consulta.

Formas Normales

1 Desnormalizado

El estado desnormalizado es aquel en el que los datos no han sido sometidos a ningún proceso de normalización. Se caracteriza por la presencia de datos anidados, duplicados y redundantes dentro de una misma estructura, lo que puede llevar a inconsistencias y dificultades en el mantenimiento, pero a veces se utiliza para optimizar el rendimiento de lectura en ciertos casos de uso.

3 Segunda Forma Normal (2NF)

La Segunda Forma Normal exige que la tabla cumpla con los requisitos de 1NF y, además, que todos los atributos no clave dependan completamente de la clave primaria completa. Esto significa que si la clave primaria es compuesta (formada por varias columnas), ningún atributo no clave puede depender solo de una parte de esa clave. Se eliminan las dependencias parciales, a menudo dividiendo la tabla en dos o más tablas relacionadas.

2 Primera Forma Normal (1NF)

Para cumplir con 1NF, cada columna en una tabla debe ser atómica, es decir, contener un solo valor y no grupos repetitivos. Además, cada tabla debe tener una clave primaria única que identifique de forma inequívoca cada fila. Esto elimina la repetición de grupos dentro de una misma fila y asegura que cada campo representa un hecho elemental.

4 Tercera Forma Normal (3NF)

Para estar en 3NF, una tabla debe satisfacer los requisitos de 2NF y, adicionalmente, no debe tener dependencias transitivas. Una dependencia transitiva ocurre cuando un atributo no clave depende de otro atributo no clave, en lugar de depender directamente de la clave primaria. El objetivo es que cada atributo no clave de la tabla dependa únicamente de la clave primaria, eliminando así redundancias y mejorando la integridad de los datos.

Consideraciones finales

Participación de stakeholders

El modelado de datos exitoso involucra a los stakeholders del negocio desde el inicio del proceso.

Definiciones de negocio

Los ingenieros necesitan obtener definiciones y objetivos comerciales para los datos.

Esfuerzo continuo

El modelado de datos debe ser una tarea recurrente e iterativa.

