

INTRODUCCIÓN AL MODELADO DE DATOS

Fundamentos del modelado de datos moderno

Rubén Hermoso Díez



Hola!

Rubén Hermoso

Ingeniero de datos y analítica con experiencia en Business Intelligence, extracción y transformación de información y diseño de cuadros de mandos con diferentes herramientas de visualización.

rhermoso@ceste.com

[Rubén Hermoso Diez | LinkedIn](#)

Certificaciones



Empresas

hiberus

Deloitte.

GRUPO
AGORA

carreras

Formaciones

CAAR
Clúster de Automoción y Movilidad de Aragón

CLOUD
FORMACION

CESTE

VÉRTICE
eLEARNING INNOVATION

¿Qué necesito tener claro antes de empezar?

Cualquier duda es buena

- Se puede interrumpir la clase con dudas, preguntas... EN CUALQUIER MOMENTO
- Las explicaciones se repiten las veces que hagan falta
- Avisadme si me acelero
- Avisadme si algo no se entiende de primeras

● **¿Cómo organizamos el descanso?**



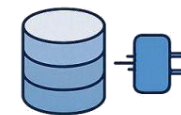
¿Qué tal nos manejamos?



¿Qué tal nos manejamos?

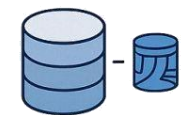


THE MODERN DATA STACK



DATA INGESTION

SQL + Connectors
Data Collection



DATA STORAGE

SQL + Data Lake/Data
Warehouse
Organized Storage



DATA PROCESSING

SQL + Big Data Processing
Engines
Data Transformation



DATA WAREHOUSING

SQL + Data Warehouse
Efficient Querying



DATA ORCHESTRATION

SQL + Workflow Tools
Automated Processes



DATA ANALYTICS & BI

SQL + BI Tools
Insight Generation

¿Qué vamos a aprender?

Introducción a los modelos de datos analíticos

- Evolución de las bases de datos
- Necesidad del modelo analítico
- ¿Qué es un datawarehouse?
- Enfoques para la creación de un datawarehouse

Kimball y el modelado dimensional

- ¿Qué es el modelo en estrella?
- Tablas de dimensiones y hechos
- Otros enfoques de modelado
- Creación de un modelo dimensional

Modelado dimensional avanzado

- Dimensiones SCD
- Tipos de tablas de hechos
- Casos especiales de modelado de tablas de dimensión
- Casos especiales de modelado de tablas de hechos

Data Vault

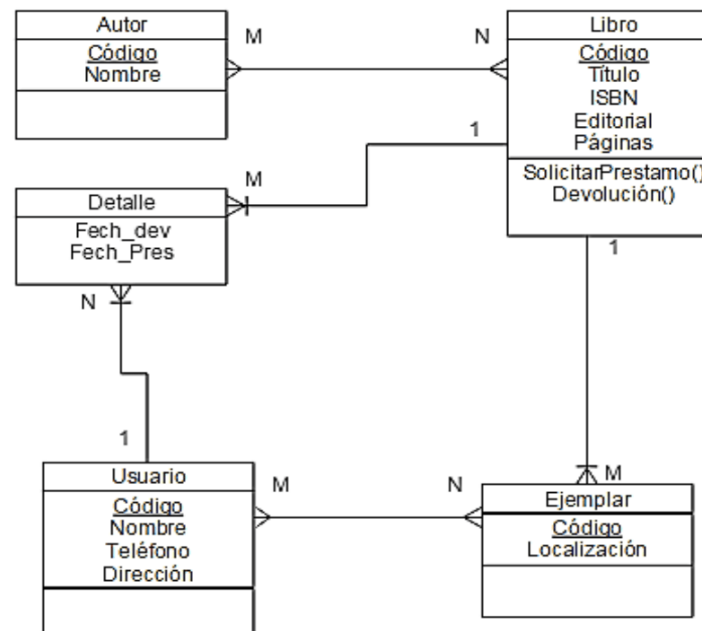
- Introducción a Data Vault
- Cómo se registra la información en Data Vault
- Hubs, satélites y links
- Pasos para la creación de un modelo de datos de tipo Data Vault

INICIACIÓN AL MODELADO DE DATOS

¿Qué es un modelo de datos?

Es la definición a nivel base de datos, de cómo la información está organizada, almacenada y conectada en el sistema de origen. Es equivalente a un **plano**, sobre cómo están los datos organizados

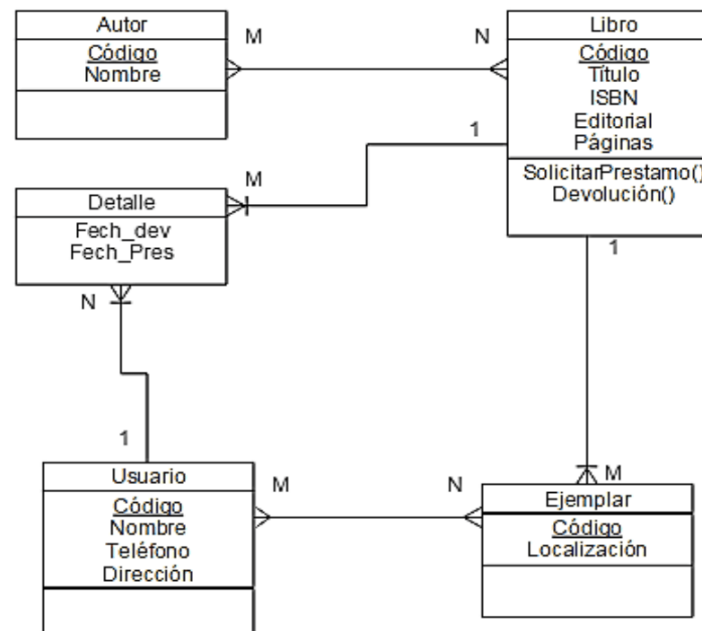
Al crear bases de datos, soluciones de datos o aplicaciones, es importante tener información sobre cómo funciona el modelo de datos para poder trabajar correctamente con él



¿Qué es un modelo de datos?

Es la definición a nivel base de datos, de cómo la información está organizada, almacenada y conectada en el sistema de origen. Es equivalente a un **plano**, sobre cómo están los datos organizados

Al crear bases de datos, soluciones de datos o aplicaciones, es importante tener información sobre cómo funciona el modelo de datos para poder trabajar correctamente con él



¿Qué es un modelo de datos?

En la era digital actual, el uso eficiente de los datos se ha convertido en un factor clave para la competitividad de cualquier organización. La cantidad y variedad de información disponible obligan a contar con estructuras que permitan organizar y relacionar adecuadamente los datos. En este sentido, los modelos de datos son esenciales, ya que permiten representar de forma lógica cómo se conecta la información entre sí, asegurando su **coherencia** y **facilitando** su **análisis**.

Gracias a un modelo de datos bien diseñado, es posible **integrar múltiples fuentes**, **evitar redundancias** y asegurar que la información esté disponible de forma fiable y comprensible para distintos perfiles dentro de la organización.

Los modelos de datos no solo ayudan a estructurar la información, sino que permiten convertir los datos en conocimiento útil para la toma de decisiones.



¿Se organizan los datos siempre de la misma forma?

SISTEMAS DE TRANSACCIONES (OLTP)

Sistemas diseñados para el manejo de datos, tanto para su almacenamiento como para su rápida recuperación.

Comúnmente asociados con los sistemas manejadores de base de datos. Estos sistemas suelen estar diseñados bajo la arquitectura cliente – servidor.

Modelo RELACIONAL

Operaciones CRUD



SISTEMAS DE ANÁLISIS (OLAP)

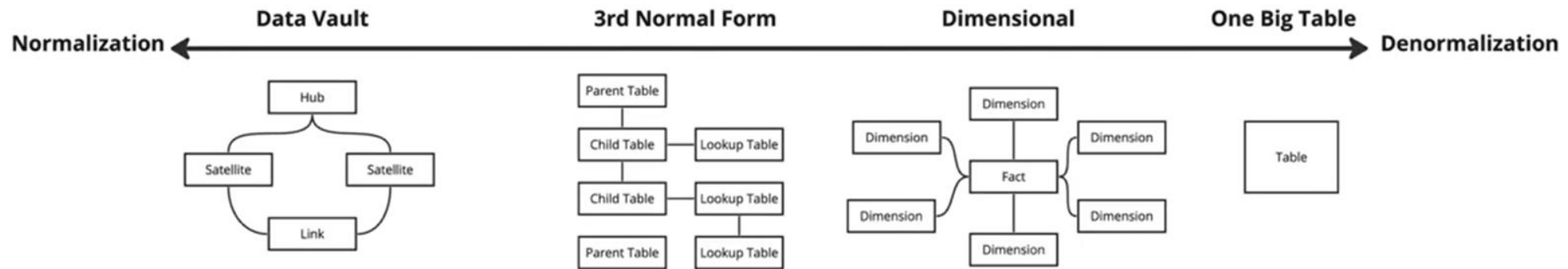
Es una categoría de tecnología de software, que permite a los analistas, gerentes y ejecutivos, obtener una mejor percepción de los datos, a través de un acceso rápido, consistente e interactivo, en una variedad de vistas posibles de la información que ha sido transformada, para reflejar la dimensionalidad real de la empresa de una manera que entienda el usuario.

Modelo DIMENSIONAL



Tipos de Modelado de Datos orientados a explotación del dato

Existen diferentes técnicas de modelado, y una forma de “clasificarlas” podría ser en función del nivel de normalización de los datos.



La **normalización** son un conjunto de reglas que se usan al diseñar y organizar una base de datos, que aseguran integridad en los datos y reducen redundancia. Básicamente, un modelo normalizado se organiza en tablas pequeñas y manejables, que son entidades independientes.

En cambio, la **desnormalización** implica añadir redundancia intencionada sobre una base de datos para mejorar el rendimiento en las lecturas o mejorar la recuperación de los datos

Ejemplo normalización: Tabla sin normalizar


Vamos a ver qué es la normalización, y lo vamos a explicar usando un ejemplo muy sencillo: una base de datos que guarda información sobre libros y sus autores.

Título del libro	Editorial	Editor Responsable	Autor 1	Autor 2
Bases de Datos	Alfa Ediciones	Marta López	Ana García	Pedro Torres
SQL Práctico	Beta Books	Juan Martín	Laura Fernández	(vacío)

Ejemplo normalización: 1FN

Aquí nos centramos en eliminar los grupos repetidos y asegurarnos de que cada celda tiene un único valor. Transformamos la tabla anterior así:

Título del libro	Editorial	Editor Responsable	Autor 1	Autor 2
Bases de Datos	Alfa Ediciones	Marta López	Ana García	Pedro Torres
SQL Práctico	Beta Books	Juan Martín	Laura Fernández	(vacío)



Título del libro	Editorial	Editor Responsable	Autor
Bases de Datos	Alfa Ediciones	Marta López	Ana García
Bases de Datos	Alfa Ediciones	Marta López	Pedro Torres
SQL Práctico	Beta Books	Juan Martín	Laura Fernández

Ejemplo normalización: 2FN

Ahora eliminamos los datos que **no dependen completamente de la clave primaria**.

En este caso, si usamos como clave “Título del libro + Autor”, vemos que la editorial o el editor no dependen del autor, solo del libro.

Título del libro	Editorial	Editor Responsable	Autor
Bases de Datos	Alfa Ediciones	Marta López	Ana García
Bases de Datos	Alfa Ediciones	Marta López	Pedro Torres
SQL Práctico	Beta Books	Juan Martín	Laura Fernández



Libros

Título del libro	Editorial	Editor Responsable
Bases de Datos	Alfa Ediciones	Marta López
SQL Práctico	Beta Books	Juan Martín

Autores de libros

Título del libro	Autor
Bases de Datos	Ana García
Bases de Datos	Pedro Torres
SQL Práctico	Laura Fernández

Ejemplo normalización: 3FN

Por último, eliminamos **las dependencias entre columnas que no son clave**.

Por ejemplo, en la tabla de libros, el “Editor Responsable” depende de la editorial, no del libro directamente.

Libros

Título del libro	Editorial	Editor Responsable
Bases de Datos	Alfa Ediciones	Marta López
SQL Práctico	Beta Books	Juan Martín

Autores de libros

Título del libro	Autor
Bases de Datos	Ana García
Bases de Datos	Pedro Torres
SQL Práctico	Laura Fernández



Editorial

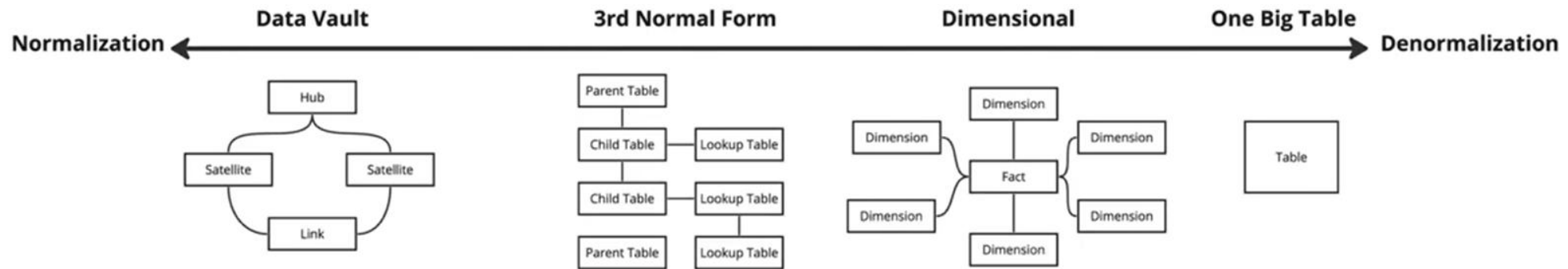
Editorial	Editor Responsable
Alfa Ediciones	Marta López
Beta Books	Juan Martín

Libros

Título del libro	Editorial
Bases de Datos	Alfa Ediciones
SQL Práctico	Beta Books

¿Normalización y desnormalización?

Existen diferentes técnicas de modelado, y una forma de “clasificarlas” podría ser en función del nivel de normalización de los datos.



La **normalización** son un conjunto de reglas que se usan al diseñar y organizar una base de datos, que aseguran integridad en los datos y reducen redundancia. Básicamente, un modelo normalizado se organiza en tablas pequeñas y manejables, que son entidades independientes. -> **OPTIMIZADO PARA ESCRITURA**

En cambio la **desnormalización** implica añadir redundancia intencionada sobre una base de datos para mejorar el rendimiento en las lecturas o mejorar la recuperación de los datos -> **OPTIMIZADO PARA ANÁLISIS**

Uno de los principales problemas cuando se desarrolla, es que se busca dar la solución al cliente muy rápidamente, y para ello, sacrificamos robustez y tiempo de diseño.

En ocasiones no nos detenemos a trabajar y elaborar un buen modelo de datos

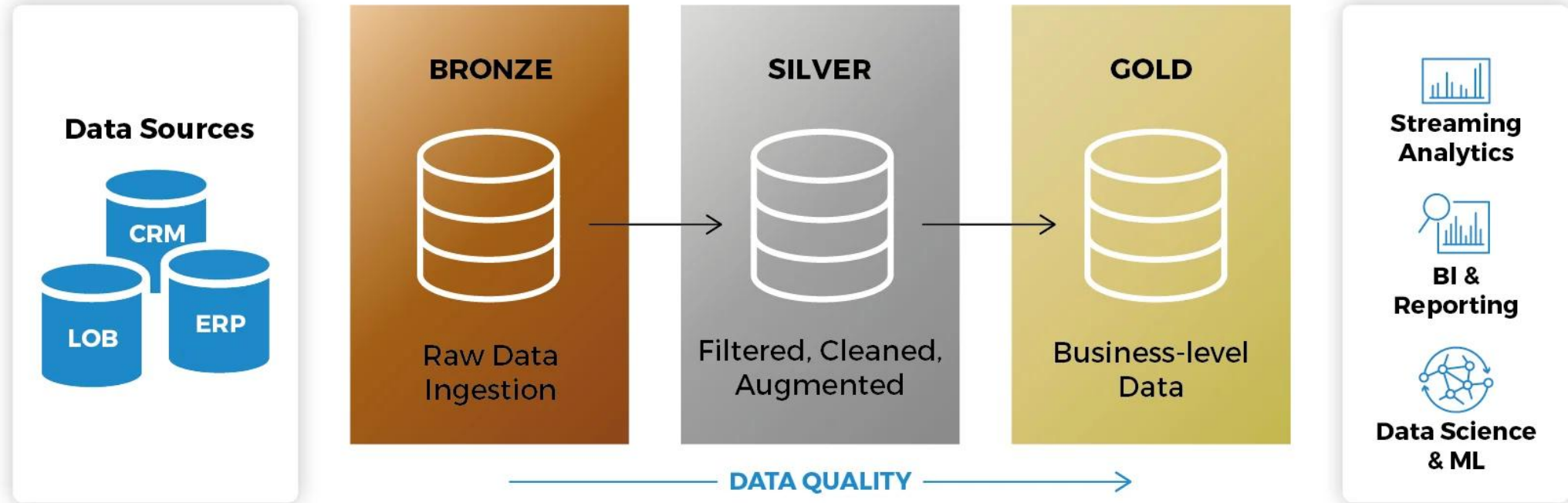
Características de cada tipo de modelo de datos

Modelo	Pros	Contras
Data Vault	- Altamente escalable	- Diseño e implementación complejos
	- Fácil integración de nuevas fuentes de datos sin afectar estructuras existentes	- Mayor uso de almacenamiento y sobrecarga de rendimiento
	- Ideal para trazabilidad	- Consultas complejas con joins y transformaciones
	- Buen rendimiento con cargas paralelas	
	- Actualizaciones incrementales eficientes	
3ª Forma Normal	- Minimiza redundancia y dependencia	- Consultas complejas para reporting
	- Flexible para actualizaciones sin impactar toda la estructura	- Optimizada para escritura, por lo que lectura más lenta
	- Eficiente en almacenamiento	- Diseño complejo con conjuntos grandes y cambiantes
	- Buen rendimiento en ciertos tipos de consultas	
	- Fácil de cargar datos	
Dimensional	- Consultas simples para usuarios finales	- Carga y actualización de datos más compleja
	- Fácil de entender	- No gestiona bien problemas de calidad de datos
	- Datos pueden agregarse previamente para mejorar rendimiento	- Difícil modificar partes establecidas
	- Mejor rendimiento para analítica	- No ideal para analítica en tiempo real
	- Menos objetos y facilita fuente única de verdad	
	- Bueno para análisis por lotes	
One Big Table	- Muy fácil de leer (todo en una sola tabla)	- Escalabilidad limitada con más volumen y complejidad
	- Buen rendimiento en consultas	- Alta redundancia de datos y código
	- Fácil de mantener	- Mal rendimiento si toda la información que necesitas no está en la tabla
	- Muy flexible	

Es muy habitual combinar diferentes modelos o enfoques de dato

Bronze	Silver	Gold	Platinum
Relational	Data Vault	Dimensional	One Big Table
Raw data stored in a relational format	Layer for adaptability and auditability	Data separated into facts & dimensions	Views built on top of the dimensional model for ease of consumption

Es muy habitual combinar diferentes modelos o enfoques de dato



**Fundación
iberCaja** 