Curso de

Data Warehousing y Modelado OLAP

Edison Yepes

¿Qué es BI y Data Warehousing?

2021 This is What Happens In An Internet Minute

Creado por: LoliLewis OfficiallyChadd

¿Qué es BI?

"Habilidad para **entender** la interconexión de los hechos presentes como **medio para descubrir guías** que nos permitan llegar a metas futuras, lo que significa mejorar nuestra capacidad de hacer juicios **predictivos** para el acierto en la **toma de decisiones**."

[Hans Peter Luhn, 1958]

"BI usa datos de **ayer** y **hoy** para tomar **las mejores decisiones** acerca del **mañana**"

[Scheps, 2008]

"Habilidad para **transformar** los **datos** en información, y la **información** en **conocimiento**, de forma que se pueda optimizar el proceso de **toma de decisiones** en los negocios"

[Mora, 2007]

"Conjunto de **estrategias y herramientas** enfocadas a la administración y creación de **conocimiento** mediante el **análisis de datos** existentes en una organización o empresa."

[Howard Dresner (Grupo Gartner), 1989]

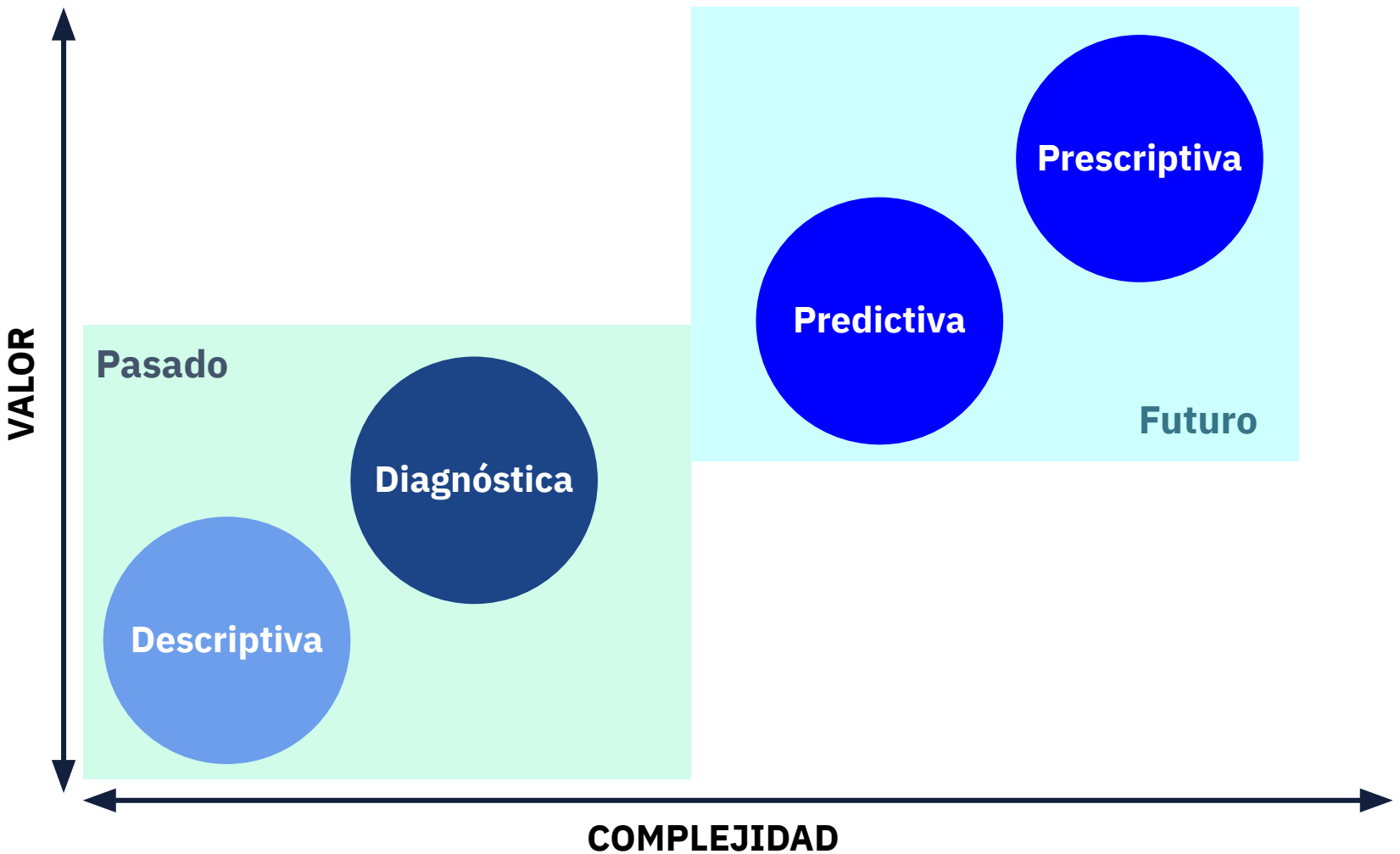
¿Qué es Data Warehousing?

“Es un proceso, no un producto, para ensamblar y administrar datos de diversas fuentes con el fin de obtener una visión única y detallada de una parte o de toda una empresa.”

Devlin, 2011

Niveles de analítica y jerarquía del conocimiento

Niveles de analítica



Fuente: https://gravitar.biz/bi/tipos-analitica/

Jerarquía del conocimiento

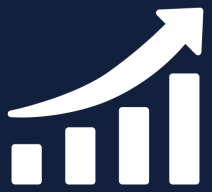
Pirámide D-I-K-W

Fuente: Hey,J.: The Data, Information, Knowledge, Wisdom Chaim: The Metaphorical Link

39°C

39°C de temperatura corporal en paciente

Si el paciente tiene 39°C de temperatura corporal significa que tiene una infección. El 60% de la población presenta síntomas.

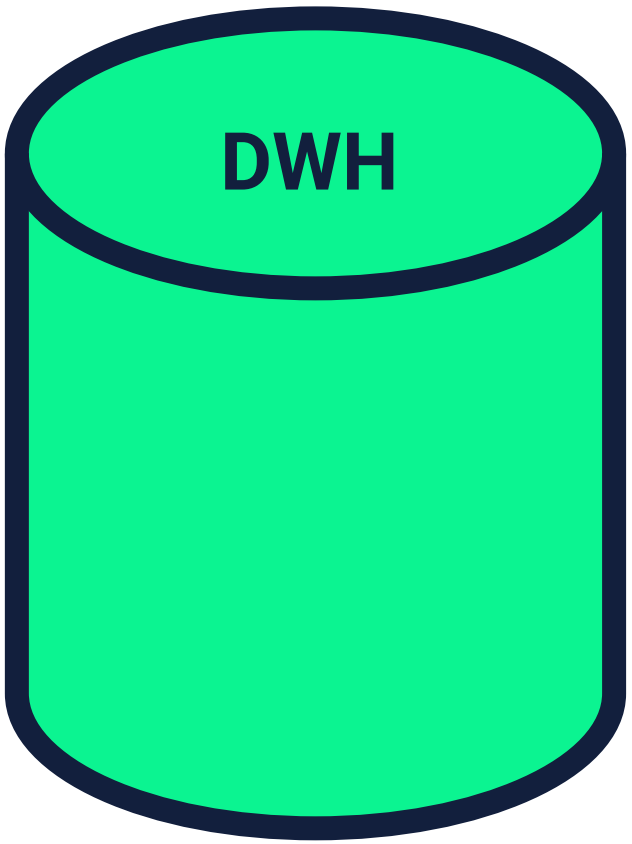


Los síntomas son asociados a un virus altamente transmisible y se deben tomar acciones específicas.

Data Warehouse, Data Mart, Dimensiones y Hechos

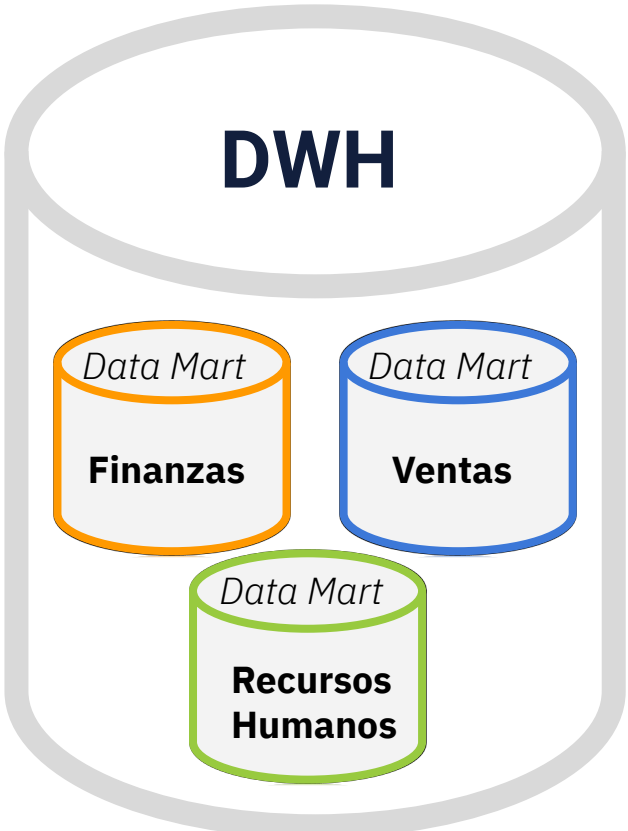
Data Warehouse (DWH)

- Base de datos que contiene información de muchas fuentes diferentes.
- Los informes creados a partir de un Data Warehouse son usados para tomar decisiones.



Data Mart

- Segmento del DWH orientado a un área específica del negocio.
- Contienen información sumarizada para el análisis en una unidad de la organización.



Dimensiones

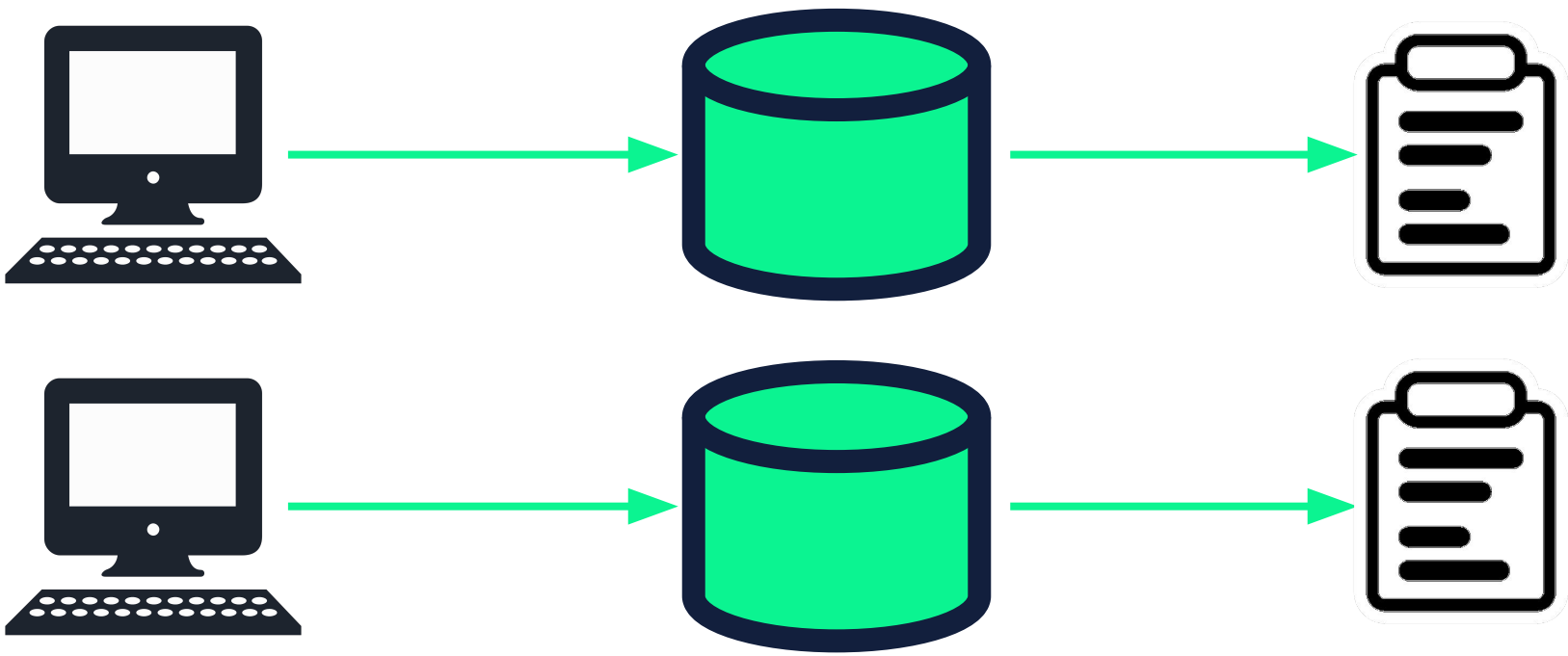
- Describen los procesos del negocio.
- Diferentes actores en los procesos del negocio.
- Entidades del negocio.

Hechos

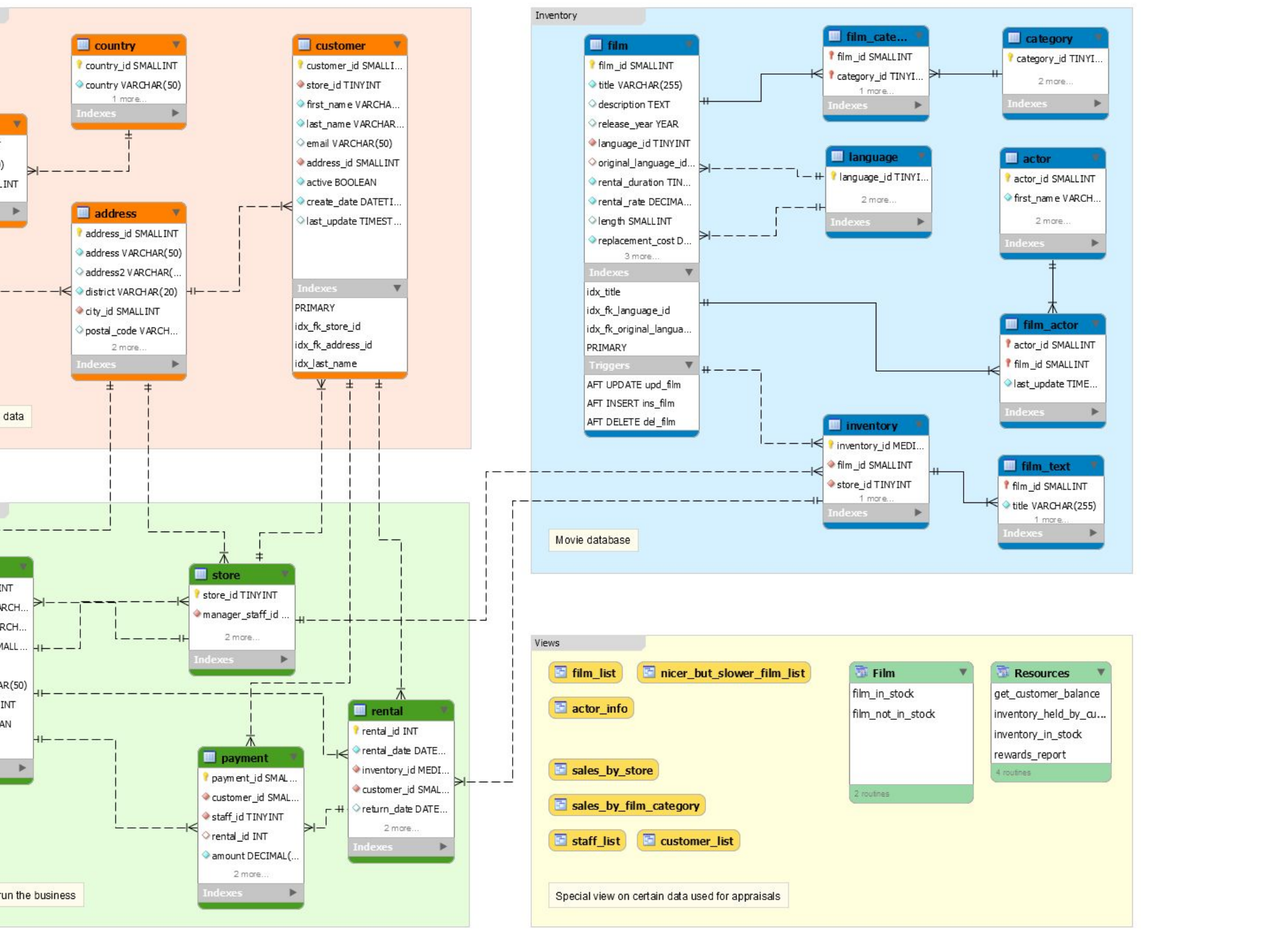
- Es la información cuantitativa de un proceso de negocio.
- Se denominan **Medidas** o **Métricas.**

OLTP vs. OLAP

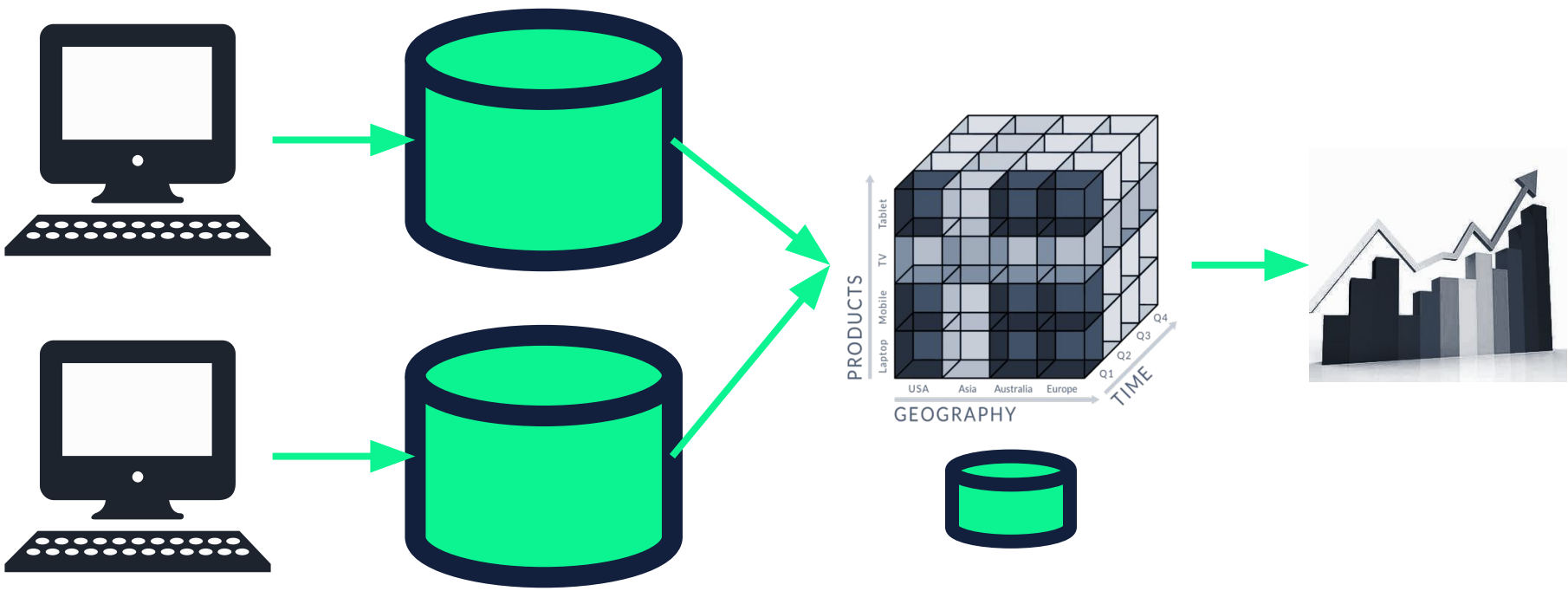
OLTP



Procesamiento de Transacciones En Línea


country
country_id SMALLINT
country VARCHAR(50)
1 more...
Indexes
customer
customer_id SMALLI...
store_id TINYINT
first_name VARCHA...
last_name VARCHAR...
email VARCHAR(50)
address_id SMALLINT
active BOOLEAN
create_date DATETI...
last_update TIMEST...
Indexes
PRIMARY
idx_fk_store_id
idx_fk_address_id
idx_last_name
address
address_id SMALLINT
address VARCHAR(50)
address2 VARCHAR(...
district VARCHAR(20)
city_id SMALLINT
postal_code VARCH...
2 more...
Indexes
data
Inventory
film
film_id SMALLINT
title VARCHAR(255)
description TEXT
release_year YEAR
language_id TINYINT
original_language_id...
rental_duration TIN...
rental_rate DECIMA...
length SMALLINT
replacement_cost D...
3 more...
Indexes
idx_title
idx_fk_language_id
idx_fk_original_langua...
PRIMARY
Triggers
AFT UPDATE upd_film
AFT INSERT ins_film
AFT DELETE del_film
film_cate...
film_id SMALLINT
category_id TINYI...
1 more...
Indexes
category
category_id TINYI...
2 more...
Indexes
language
language_id TINYI...
2 more...
Indexes
actor
actor_id SMALLINT
first_name VARCH...
2 more...
Indexes
film_actor
actor_id SMALLINT
film_id SMALLINT
last_update TIME...
Indexes
inventory
inventory_id MEDI...
film_id SMALLINT
store_id TINYINT
1 more...
Indexes
film_text
film_id SMALLINT
title VARCHAR(255)
1 more...
Indexes
Movie database
store
store_id TINYINT
manager_staff_id ...
2 more...
Indexes
rental
rental_id INT
rental_date DATE...
inventory_id MEDI...
customer_id SMAL...
return_date DATE...
2 more...
Indexes
payment
payment_id SMAL...
customer_id SMAL...
staff_id TINYINT
rental_id INT
amount DECIMAL(...
2 more...
Indexes
run the business
Views
film_list
nicer_but_slower_film_list
actor_info
sales_by_store
sales_by_film_category
staff_list
customer_list
Film
film_in_stock
film_not_in_stock
2 routines
Resources
get_customer_balance
inventory_held_by_cu...
inventory_in_stock
rewards_report
4 routines
Special view on certain data used for appraisals

OLAP



Procesamiento Analítico En Línea

OLAP

Fuente: https://i1.wp.com/blog.auriboxtraining.com/wp-content/uploads/2017/08/modelo-dimensional-1.png

OLTP vs. OLAP

Sistema Operacional (OLTP)

- ¿La factura ### fue cancelada?
- ¿Cliente que compró el producto X el día de hoy?

Sistema de Bodega de Datos (DW)

- ¿Producto más vendido en el año 2022 por línea de producto?
- ¿Comparación de las ventas vs. el presupuesto, mes a mes, por tienda?

OLTP


MySQL


PostgreSQL


ORACLE

OLAP


amazon
REDSHIFT


Azure Synapse Analytics

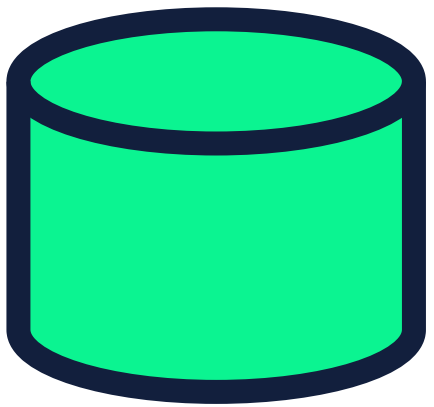

Google BigQuery

OLTP	OLAP
Diseñado para soportar las transacciones comerciales.	Diseñado para apoyar el proceso de toma de decisiones.
Data es volátil.	Data NO es volátil.
Data detallada.	Datos resumidos.
Modelado E-R.	Modelado dimensional.
Procesamiento de transacciones.	Procesamiento analítico.
Alta concurrencia.	Baja concurrencia.

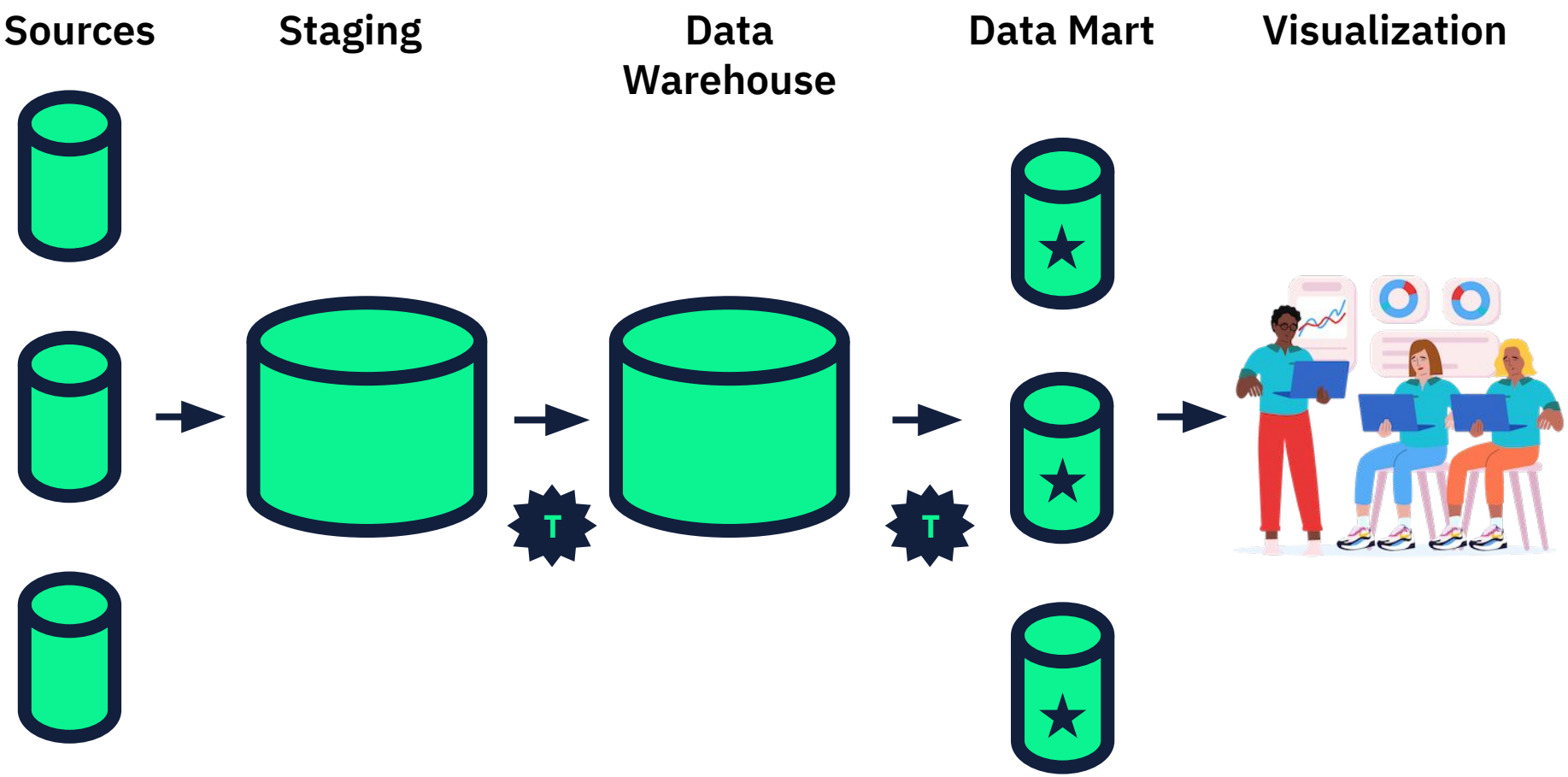
Metodologías de Data Warehouse

Metodologías de DWH

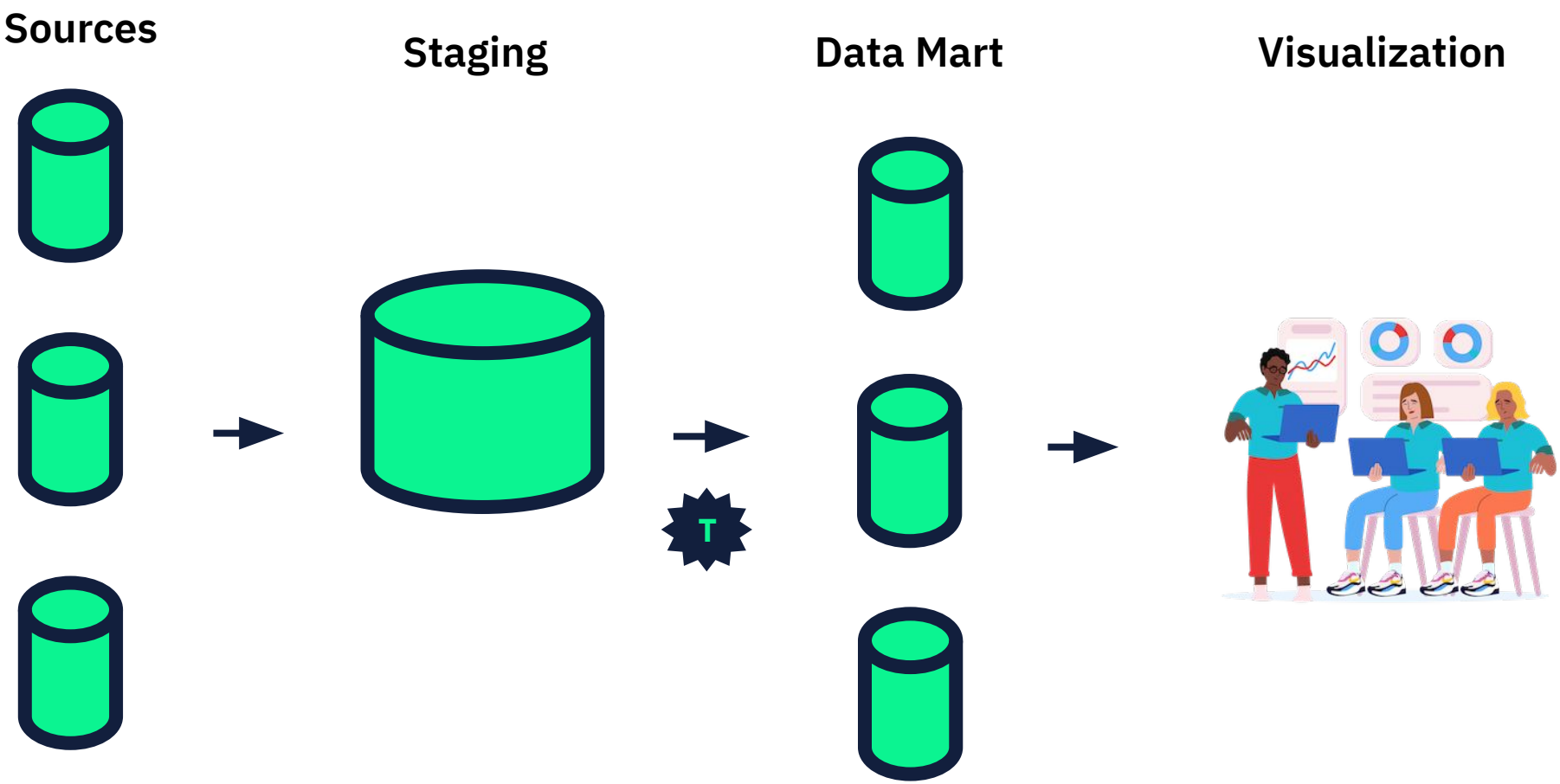
- Sources
- Dimensional Models
- Visualization



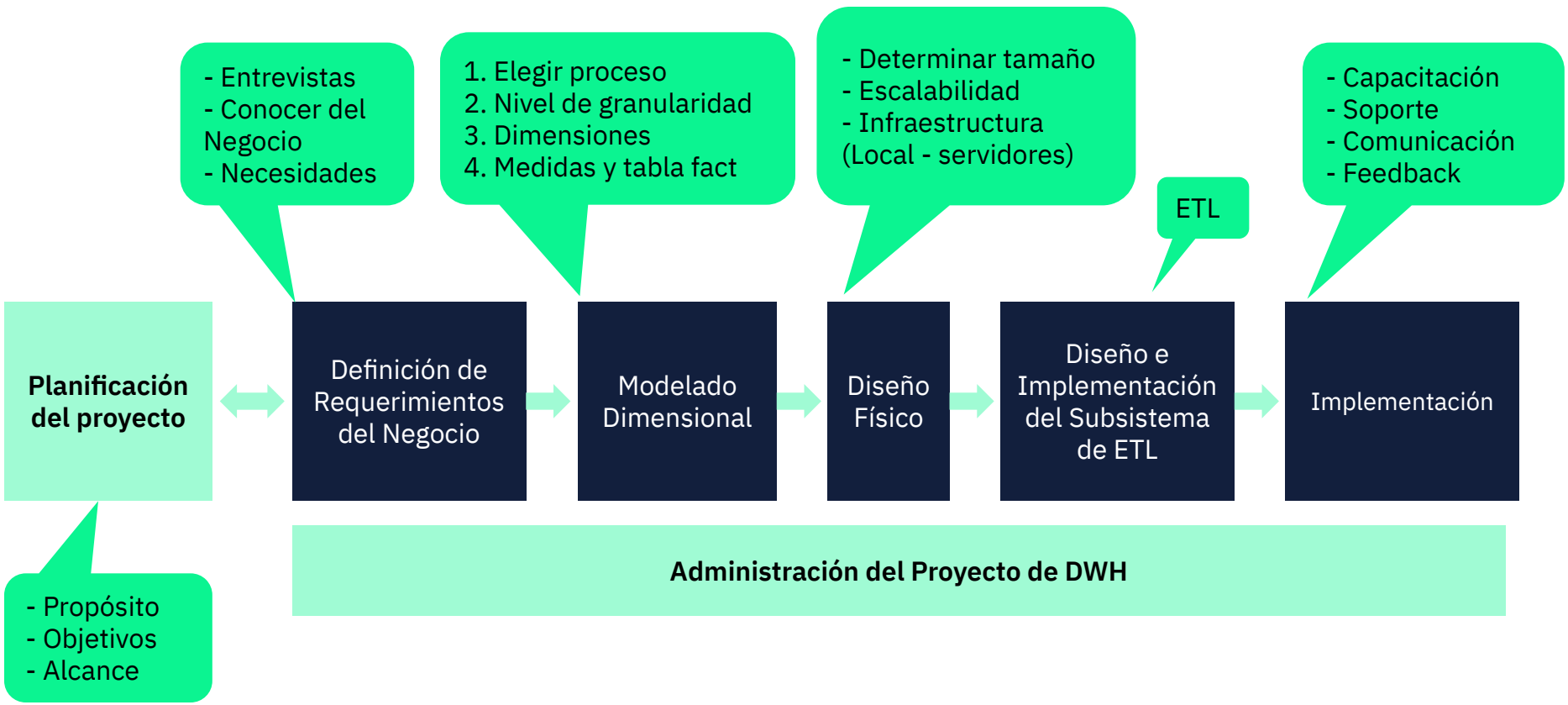
Bill Inmon


Sources
Staging
Data Warehouse
Data Mart
Visualization
T
T

Ralph Kimball

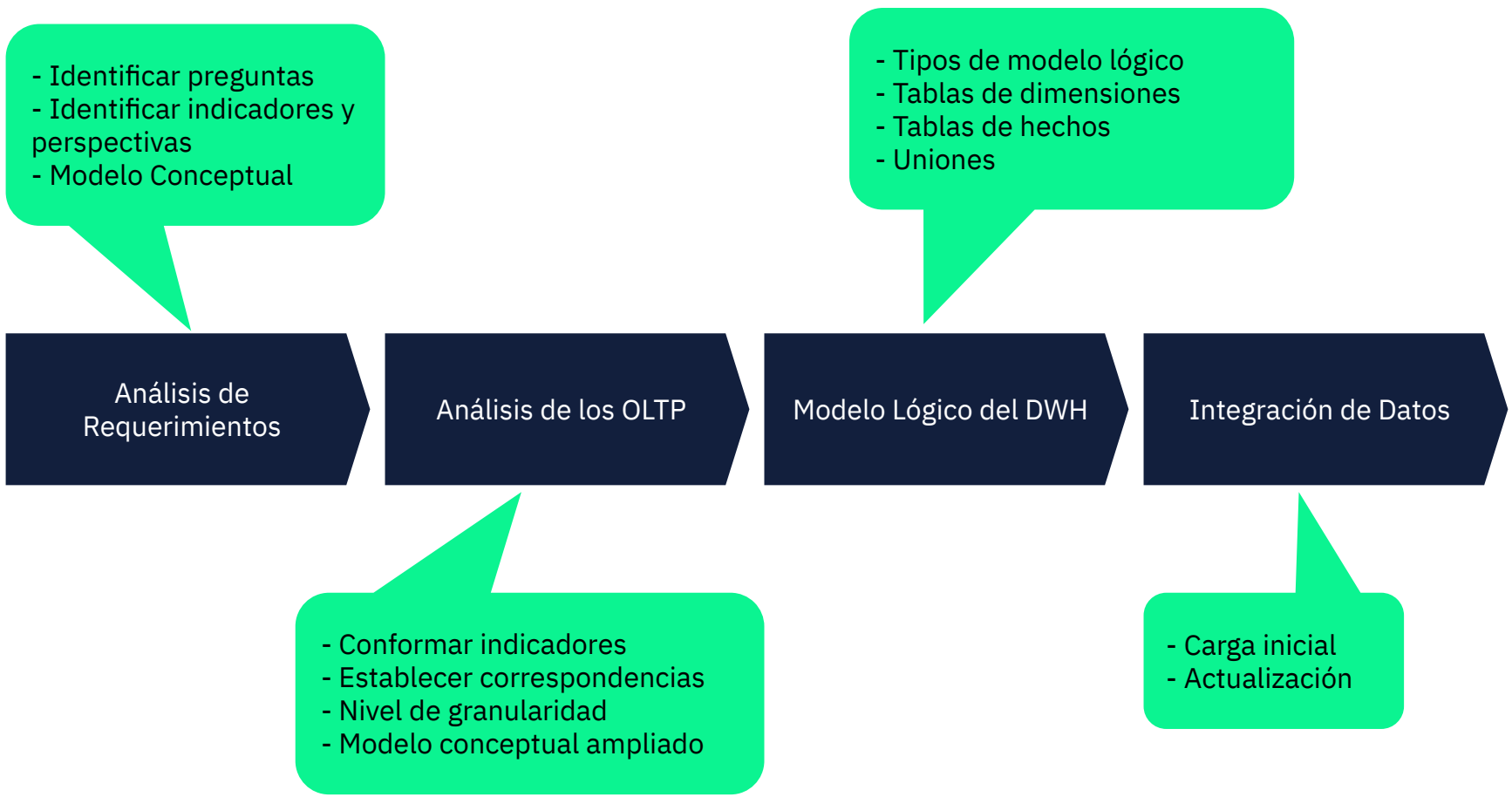


Ralph Kimball - fases



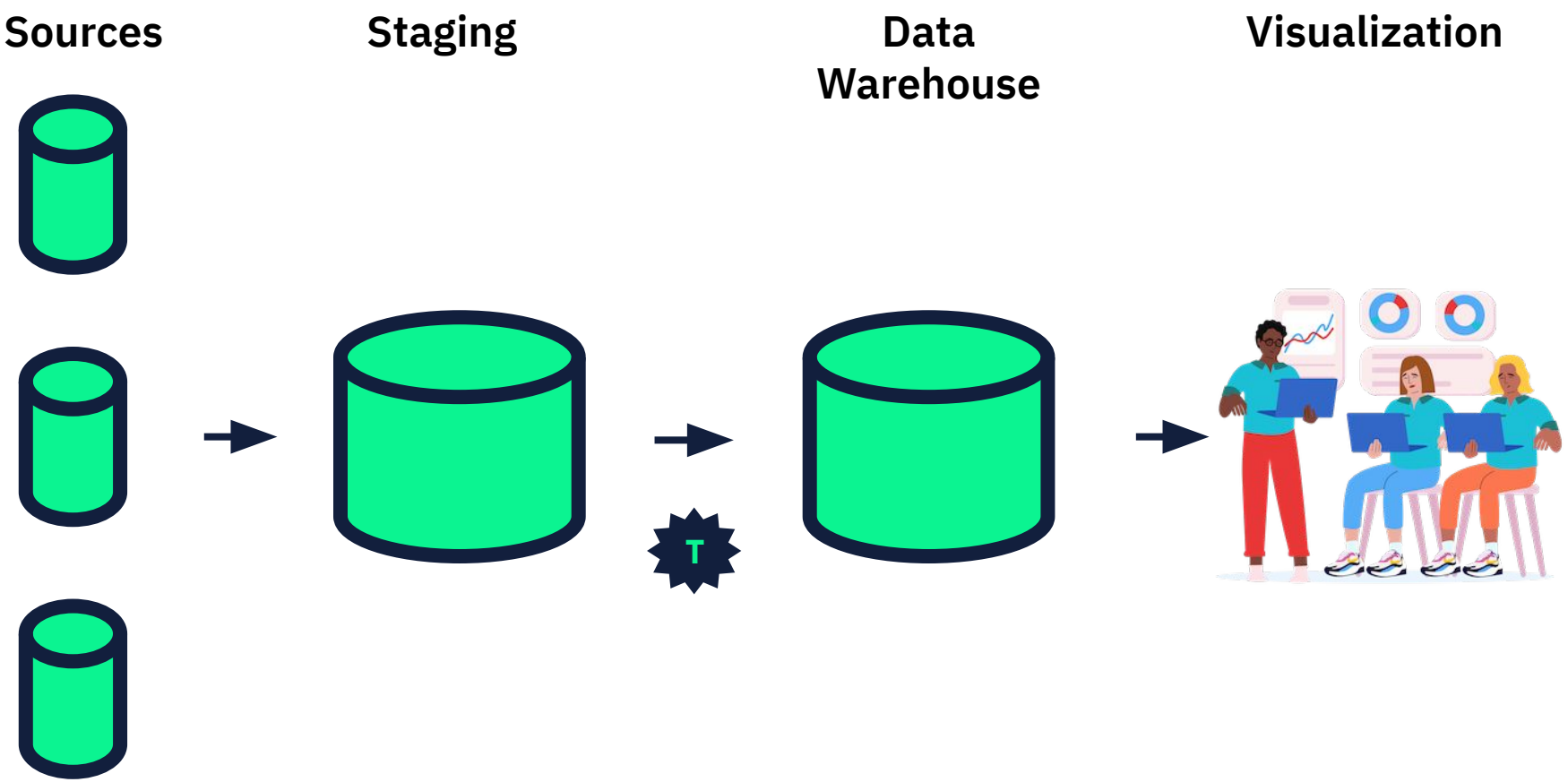
Fuente: https://docplayer.es/docs-images/40/689468/images/page_12.jpg

Hefesto

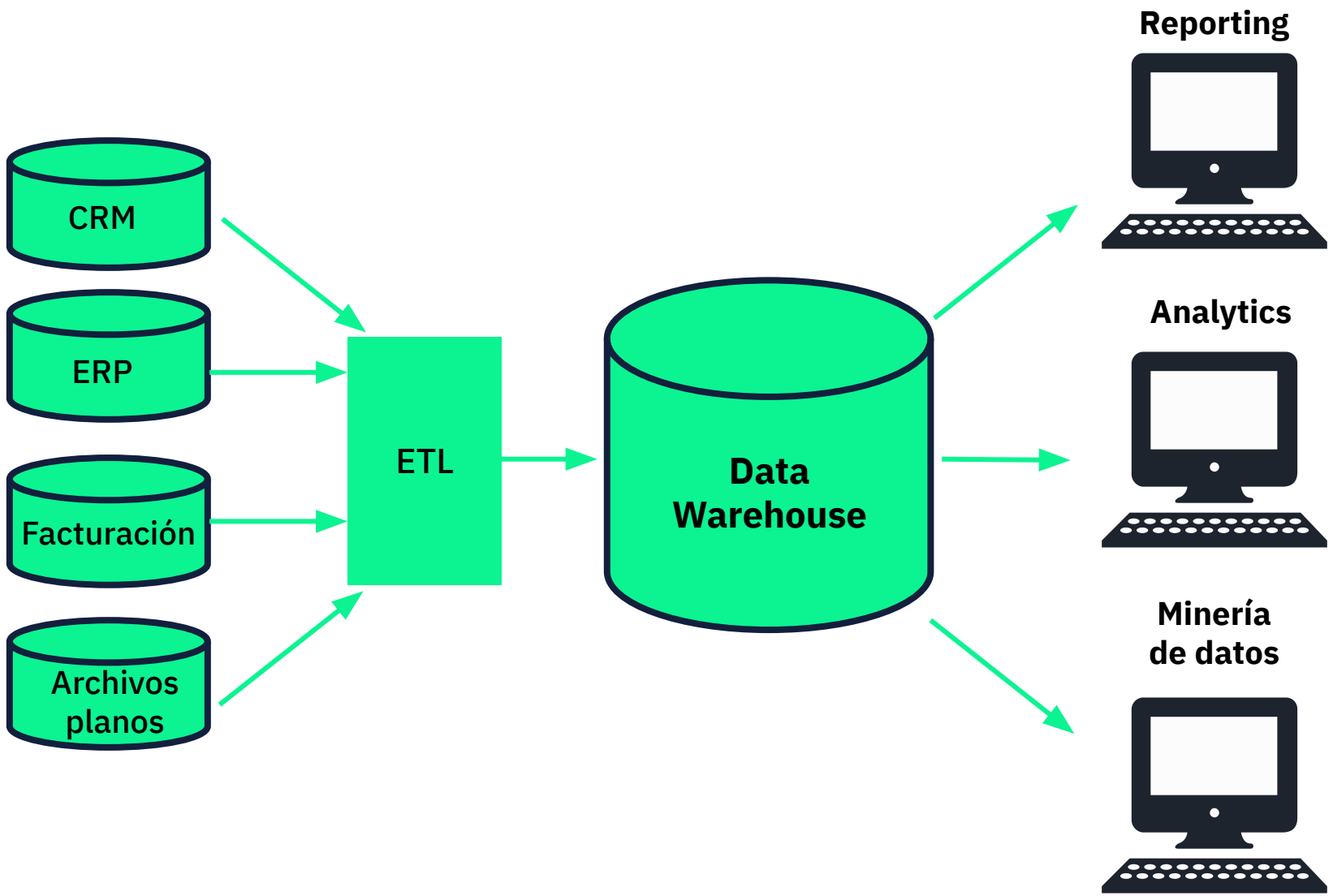


Fuente: https://docplayer.es/docs-images/40/689468/images/page_12.jpg

Personalización

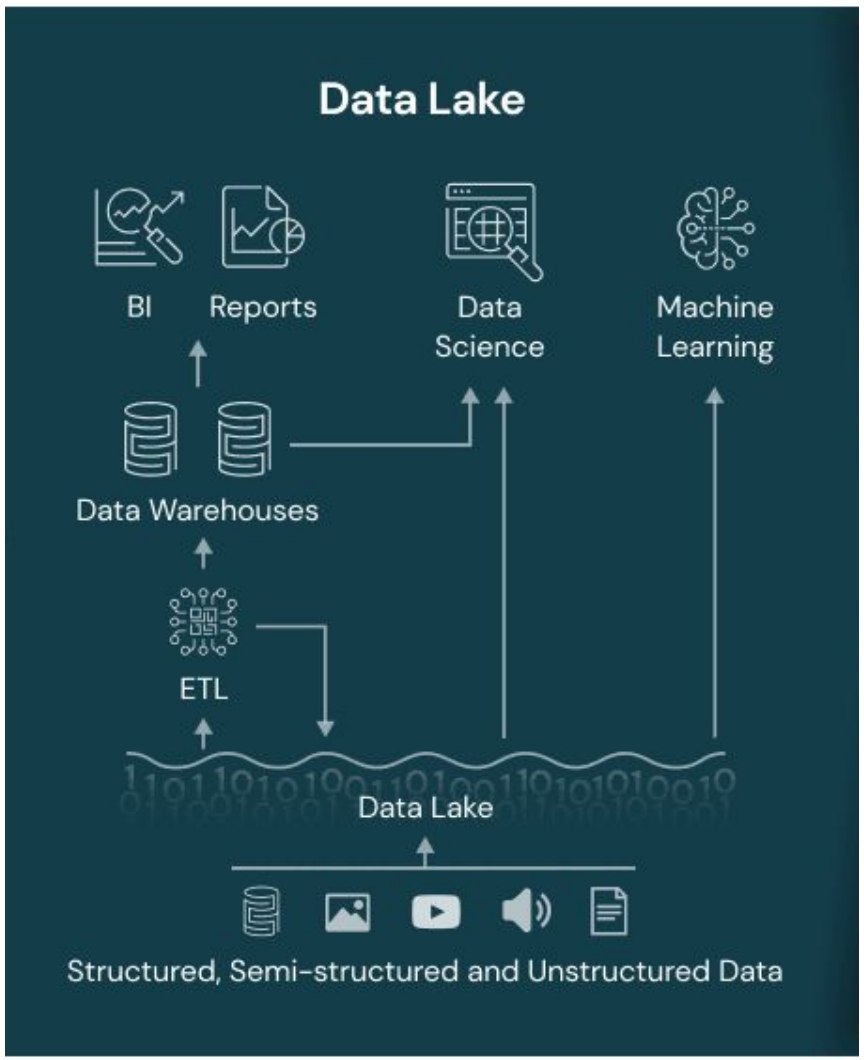


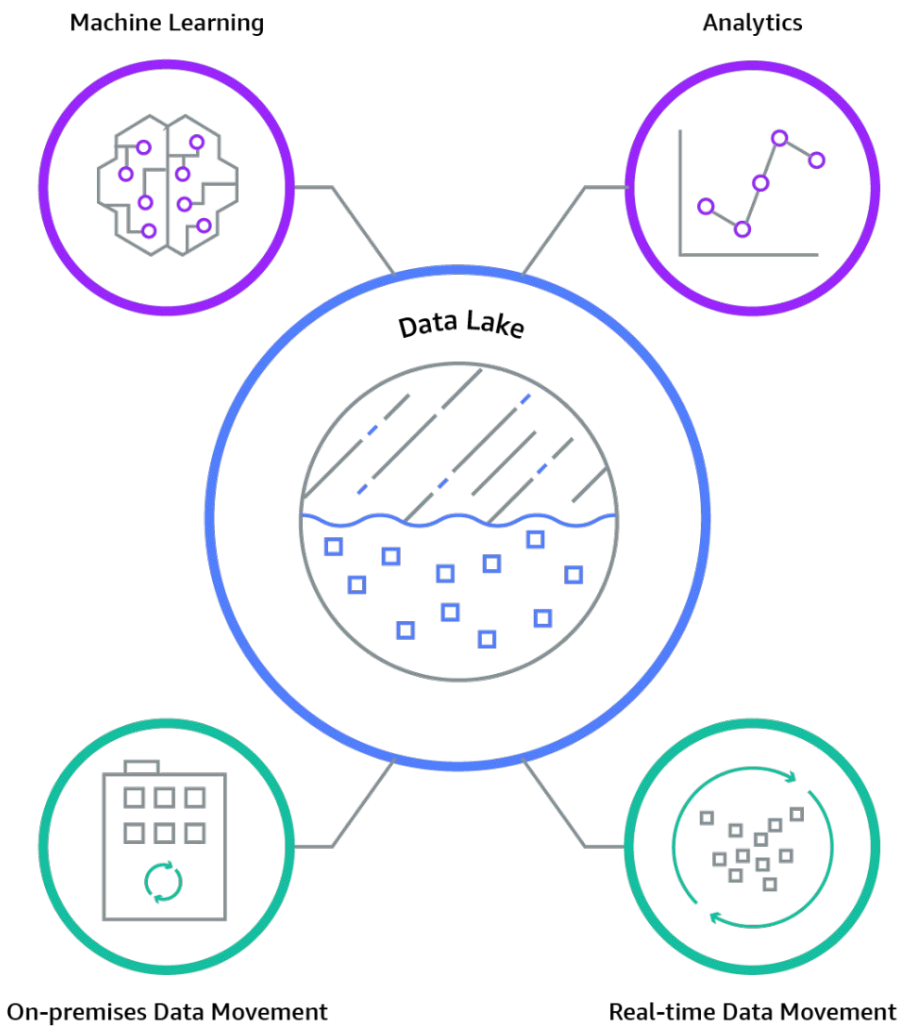
Data Warehouse, Data Lake y Data Lakehouse: ¿cuál utilizar?



Fuente: https://www.erp-information.com

Data Lake





Fuente: https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Data Warehouse vs. Data Lake

Características	Data Warehouse	Data Lake
Data	Optimizado para analizar datos relacionados de fuentes como BD transaccionales, operativas y aplicaciones de negocio.	Datos no relacionados de fuentes como web sites, redes sociales, dispositivos IoT, aplicaciones móviles.
Schema	La estructura de datos se define antes de la implementación para optimizar las consultas (schema-on-write).	Almacena información sin la definición de una estructura de datos. Permite implementar sin conocer aún las preguntas de negocio (schema-on-read).
Data Quality	Los datos se limpian, enriquecen y transforman para que puedan actuar como la "única fuente de verdad".	Cualquier dato que pudo o no pasar por un proceso de limpieza y transformación (raw data).
Users	Analistas de negocio.	Científicos de datos, ingenieros de datos, y analistas de datos (usando información limpia).
Analytics	Reportes, tableros de control y BI.	Machine Learning, análisis predictivos, data discovery y profiling.

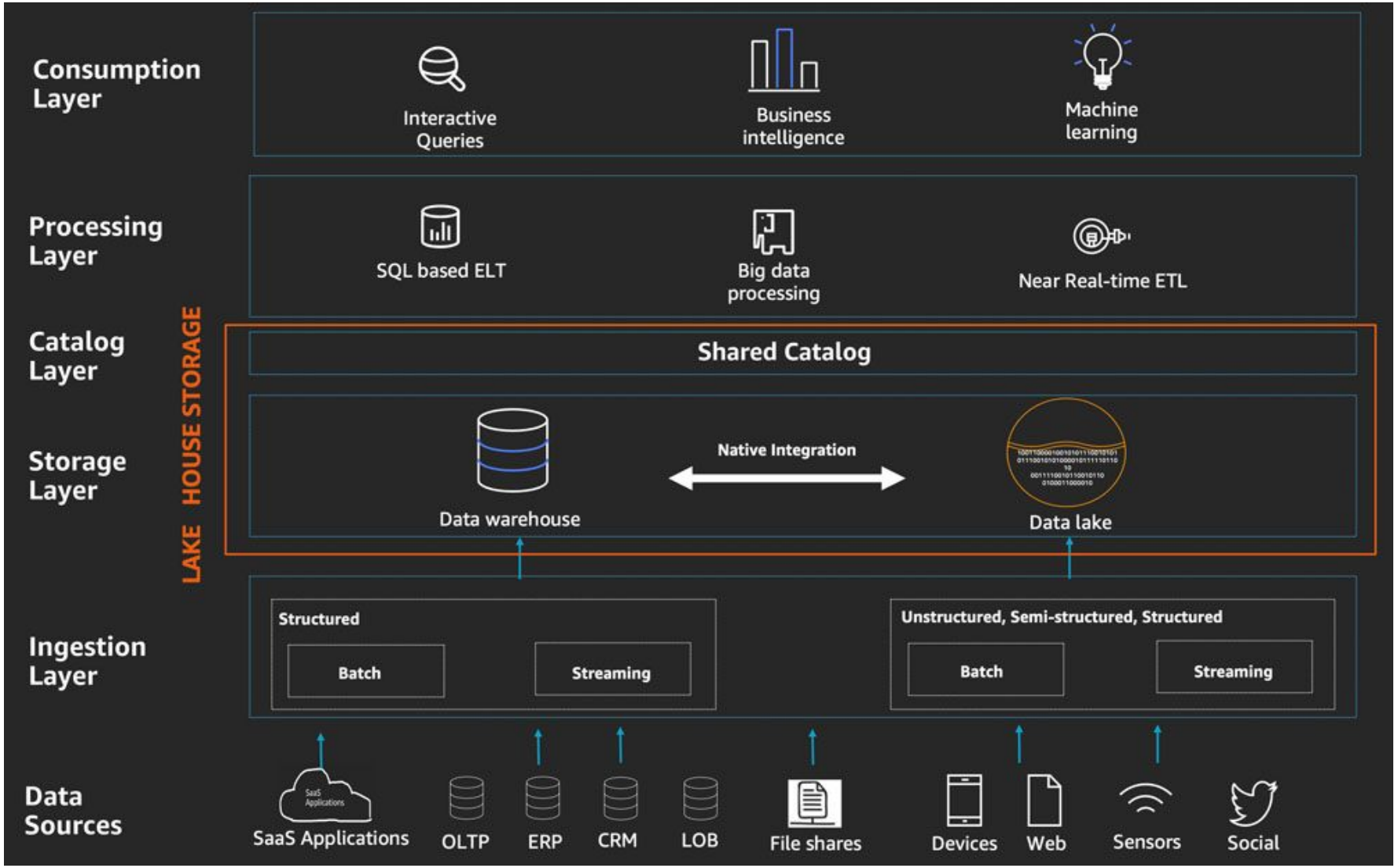
¿Data Lakehouse?

Data Lakehouse

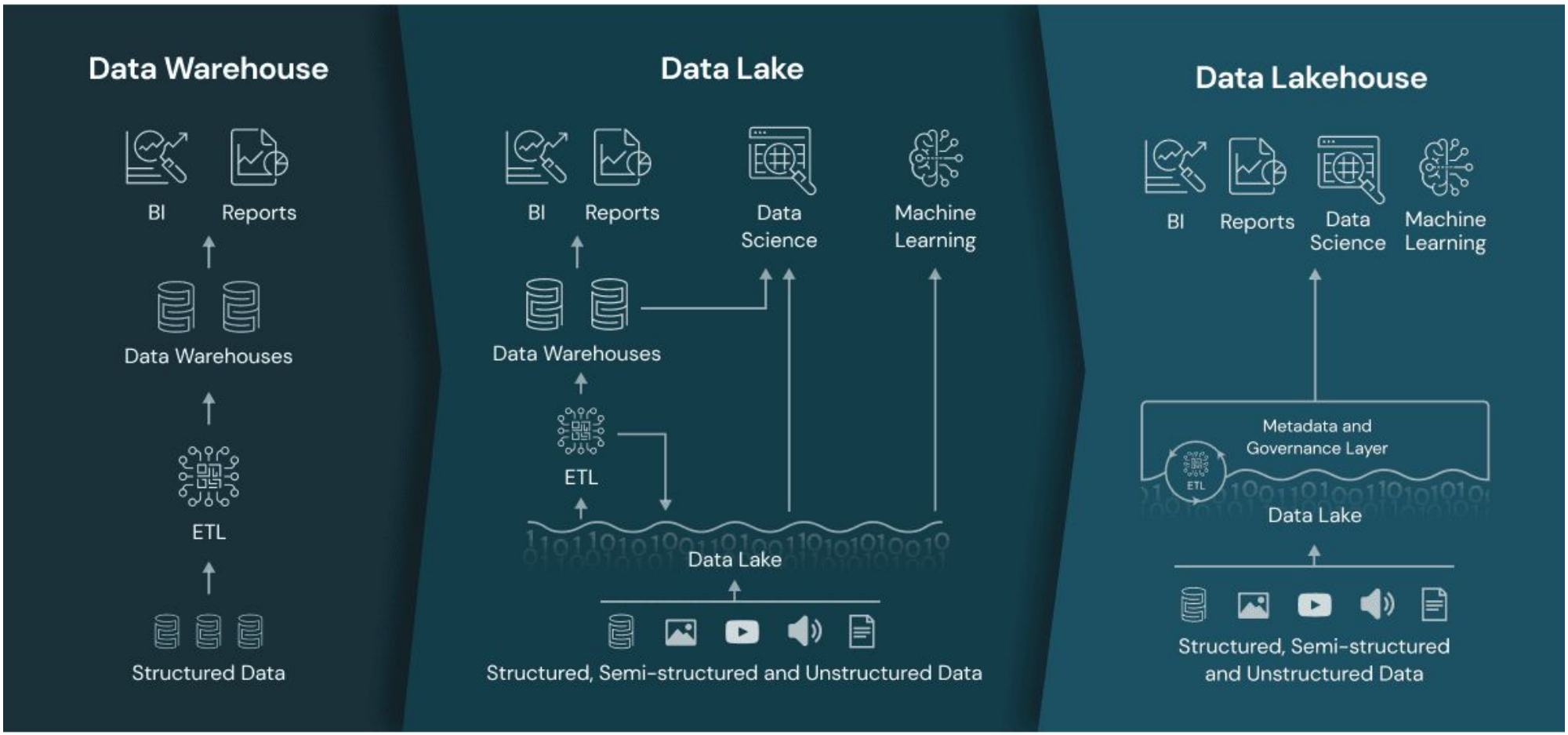
Data Warehouse + Data Lake = Data Lakehouse

https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/

Data Lakehouse



Fuente: https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws

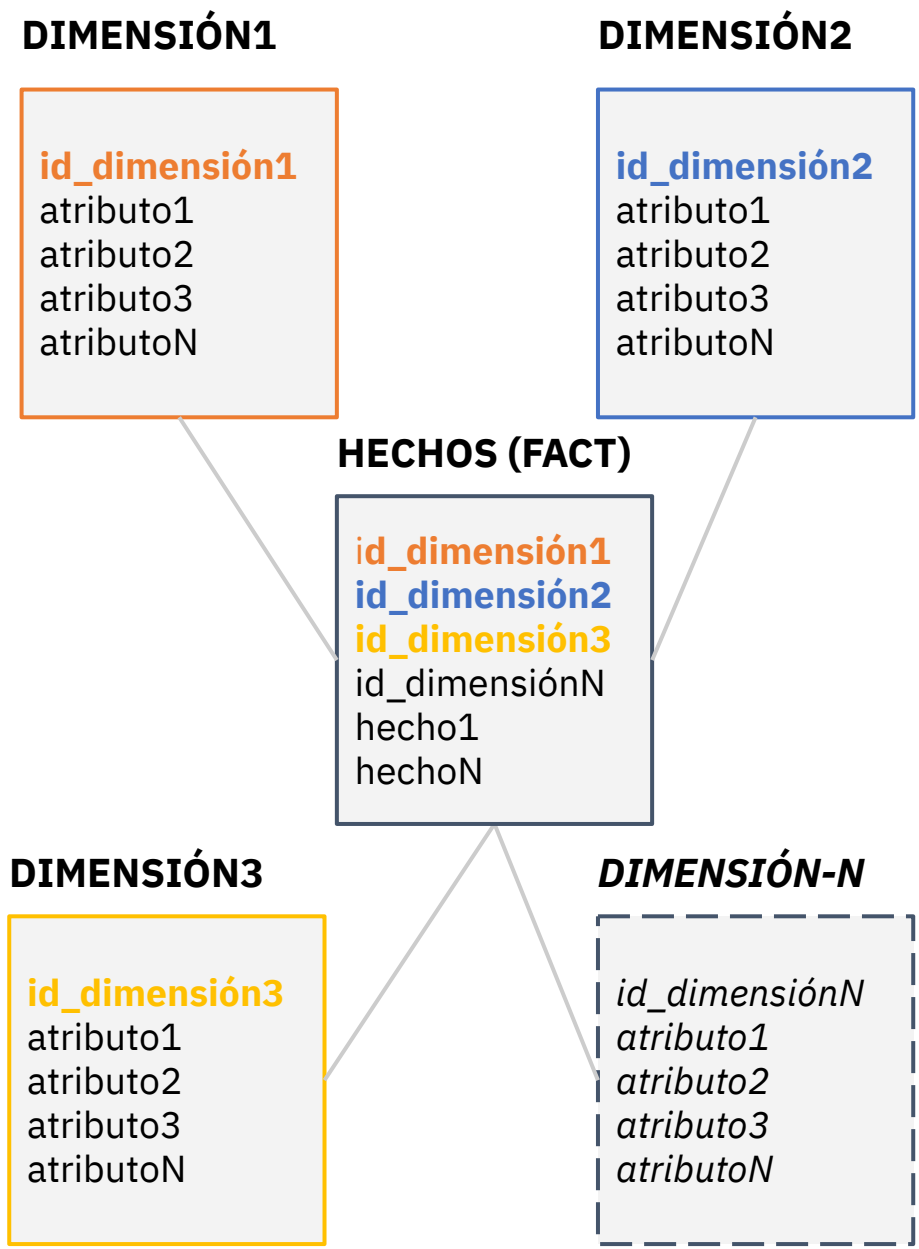


Fuente: https://www.databricks.com/wp-content/uploads/2020/01/data-lakehouse-new.png

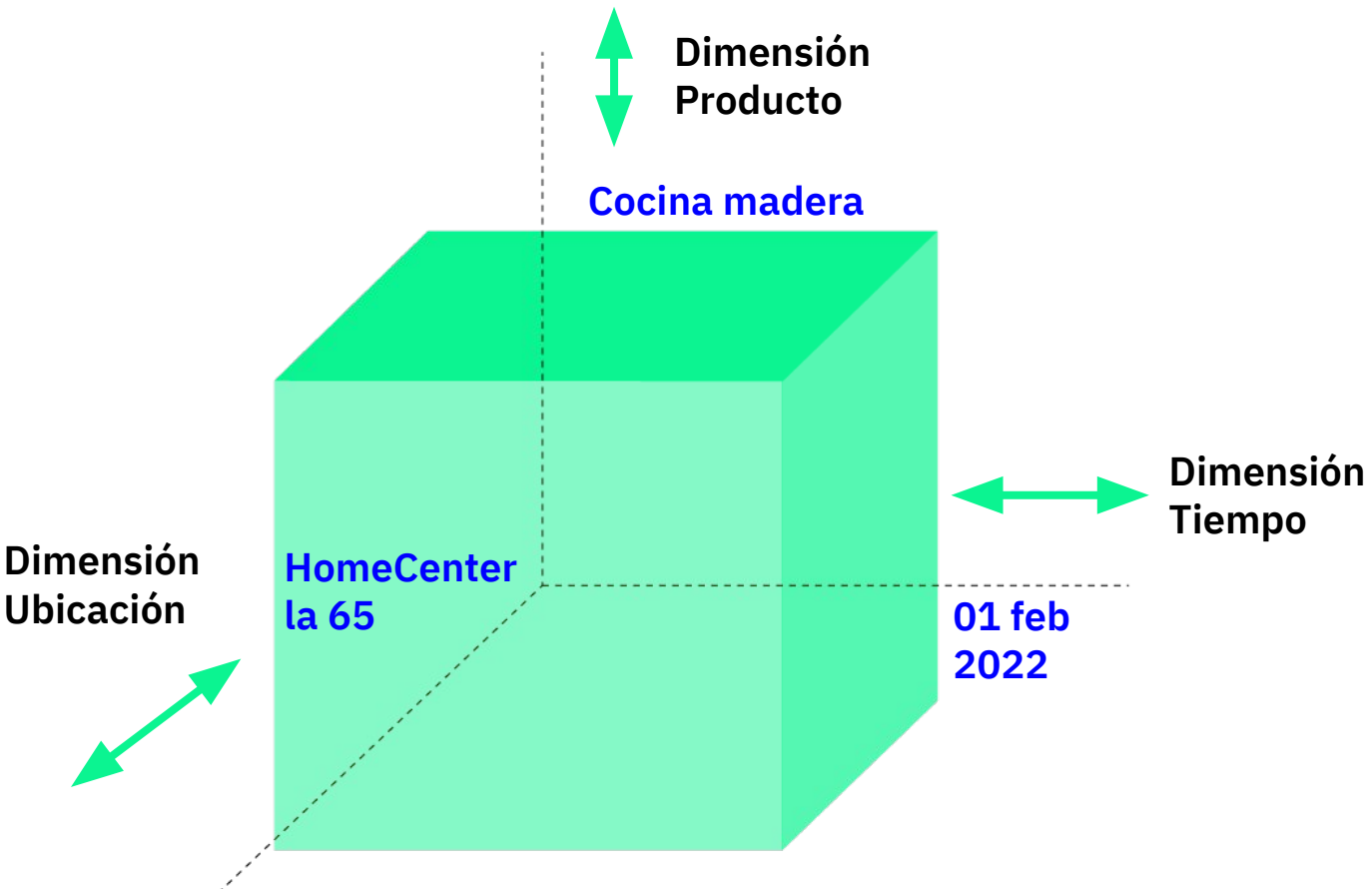
Tipos de esquemas dimensionales

Modelo multidimensional

Un evento objeto de análisis (hecho) que es evaluado bajo diferentes perspectivas o puntos de vista (dimensiones).

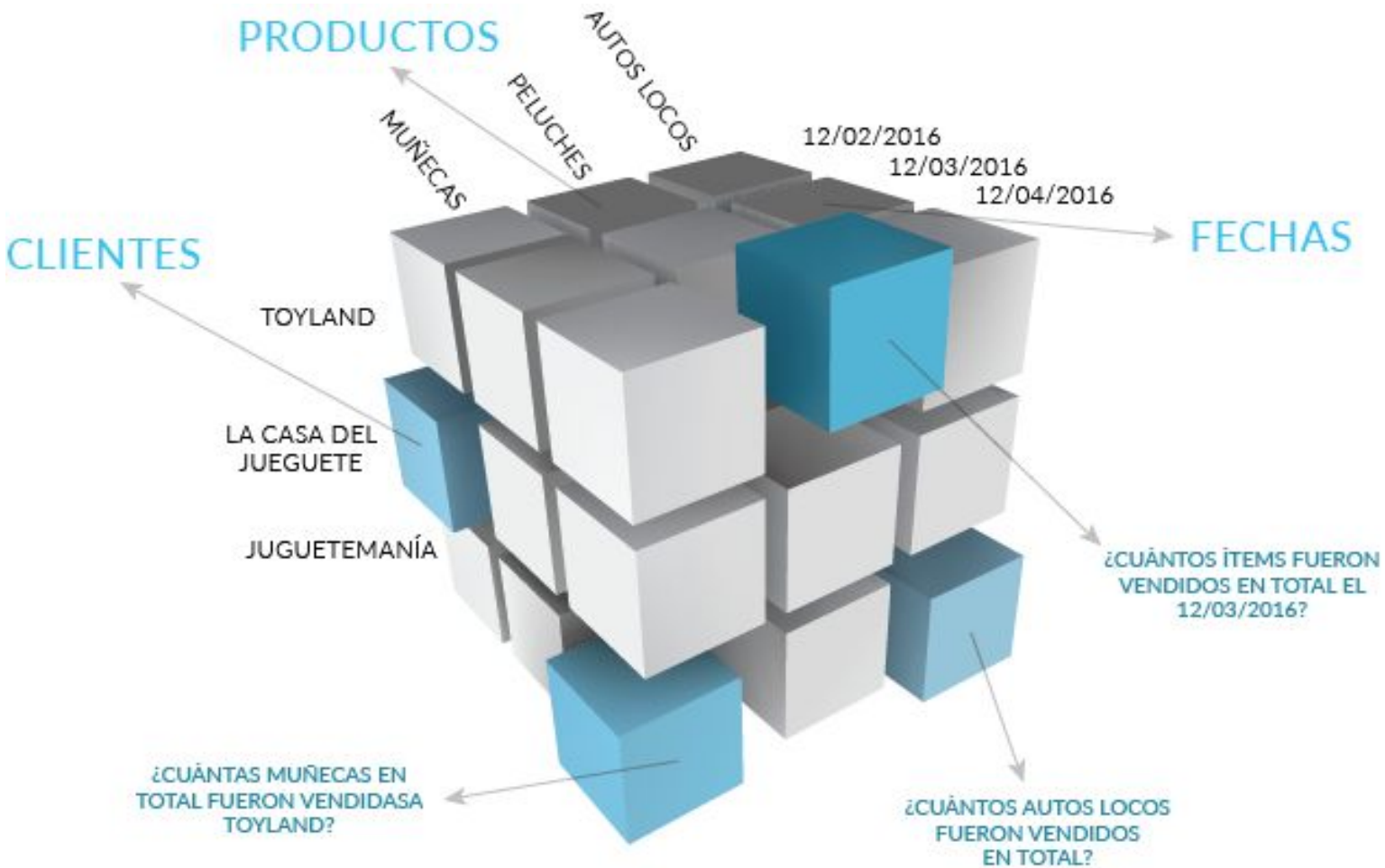


Cubo



Conceptualmente, el modelo de DWH
se puede representar mediante un cubo.

Cubo



https://i.imgur.com/Pz1wxud.png

Esquema estrella

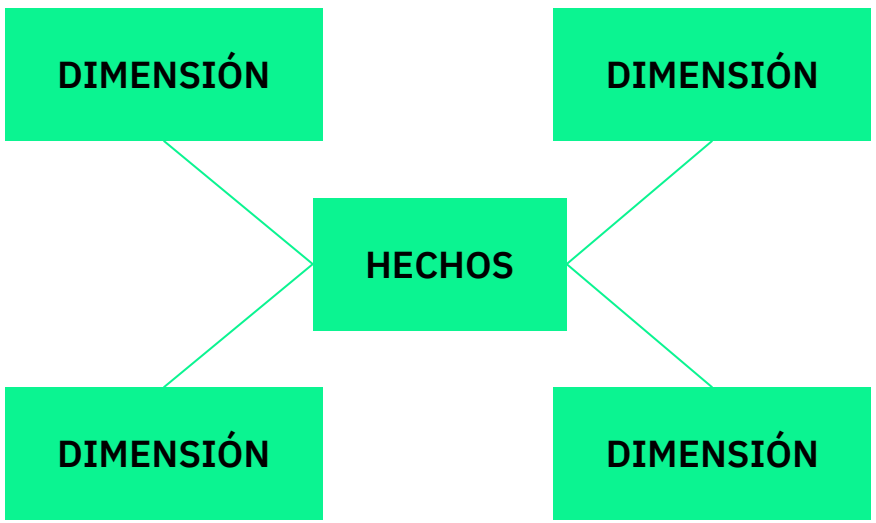
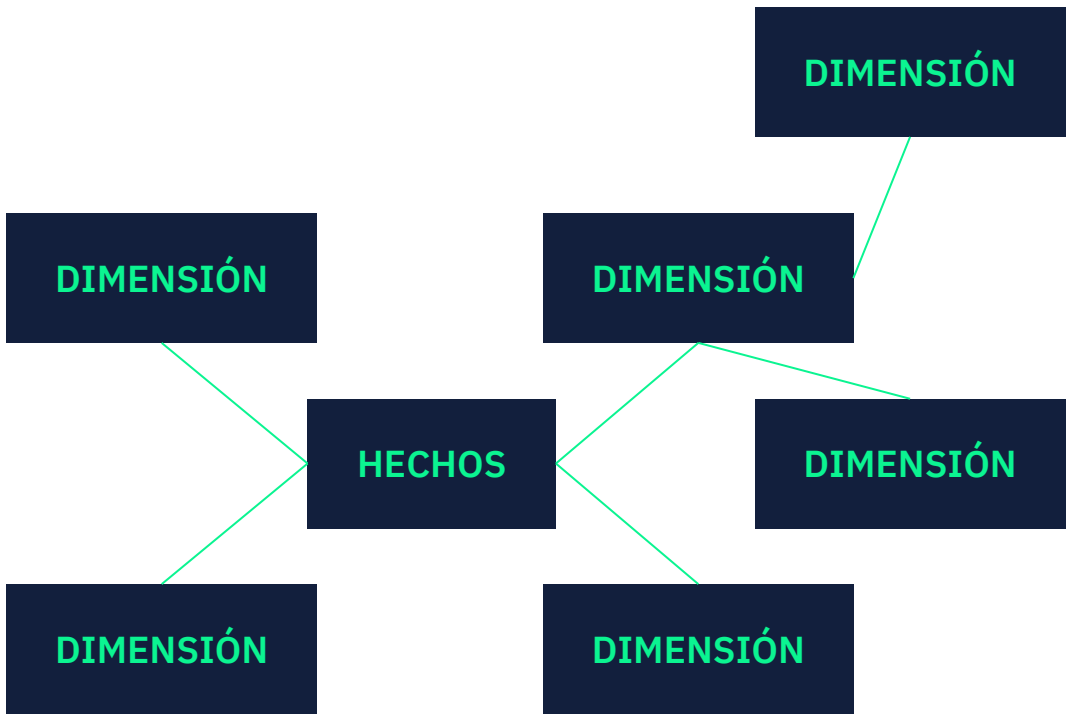


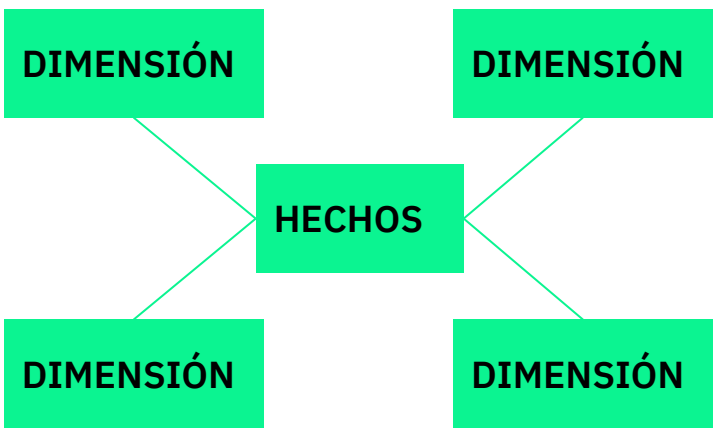
Tabla central (hechos) que se relaciona con dimensiones

Esquema copo de nieve

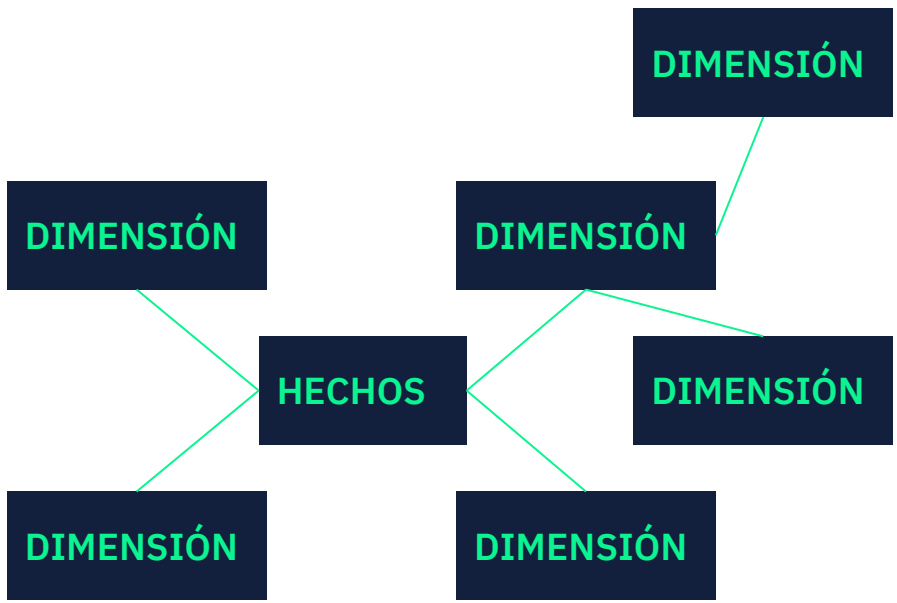


Dimensiones normalizadas que se relacionan con otras dimensiones

Tipos de esquemas



Estrella



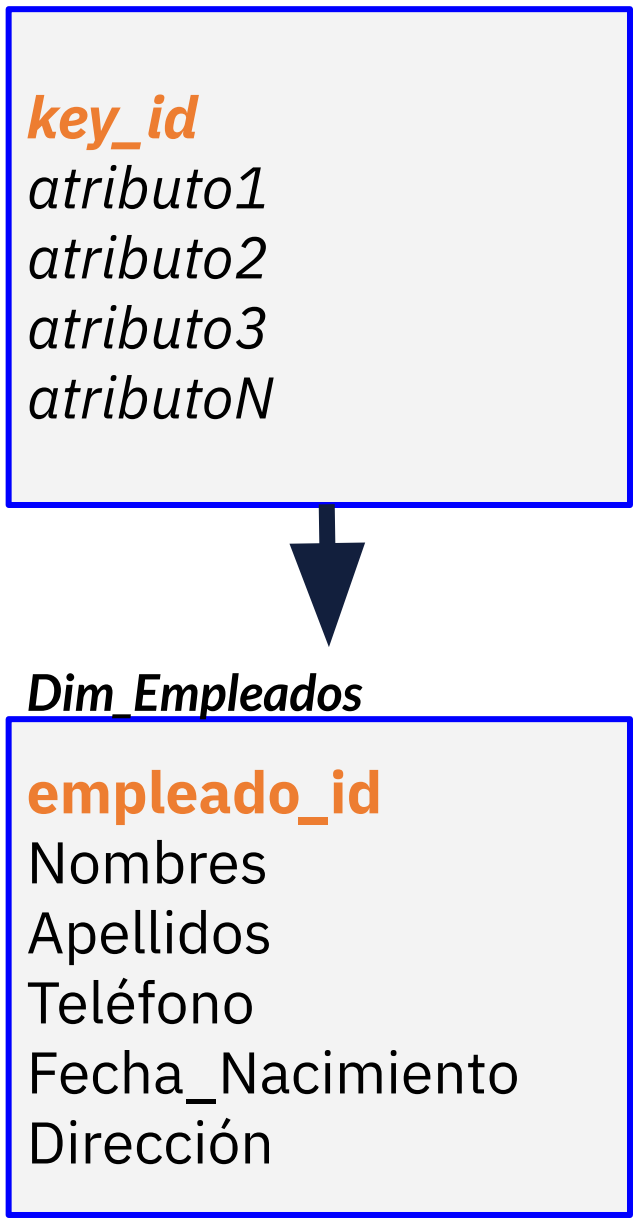
Copo de nieve

Dimensiones lentamente cambiantes

Dimensión

Los diferentes actores en los procesos del negocio.

- ¿Quién?
- ¿Qué?
- ¿Cómo?
- ¿Cuándo?
- ¿Dónde?



⚙️ Atributos

- **Jerárquicos**

  Permiten ir de lo general a lo particular. Consolidar y desagregar. Por ej: país.
- **Descriptivos**
  Información relevante. Netamente descriptiva. Por ej: dirección, teléfono, talla, clima.

⚙️ Atributos

- **De control**
  Datos de auditoría. No pertenecen al conocimiento del negocio. Por ej: fecha de carga.

↗ Tipos de SDC

- **Tipo 1**

  Sobreescribir el atributo actualizado.
- **Tipo 2**

  Agrega un nuevo registro con el cambio.
- **Tipo 3**

  Agrega un nuevo atributo “anterior”.

Dimensión lentamente cambiante tipo 1

SCD tipo 1 - Reemplaza

Id_Estudiante	Nombre Completo	Facultad
EST12345	Pepito Perez	Mercadeo

Cambio de facultad a **Ingeniería**

Id_Estudiante	Cod_Estudiante	Nombre Completo	Facultad
1	EST12345	Pepito Perez	Ingeniería

Dimensión lentamente cambiante tipo 2

SCD tipo 2 - Agrega fila

Id_Estudiante	Nombre Completo	Facultad
EST12345	Pepito Perez	Mercadeo

Cambio de facultad a **Ingeniería**

Id_Estudiante	Cod_Estudiante	Nombre Completo	Facultad	Start_date	End_date
1	EST12345	Pepito Perez	Mercadeo	01/01/2020	**01/01/2023**
2	EST12345	Pepito Perez	**Ingeniería**	**02/01/2023**	***31/12/9999***

Dimensión lentamente cambiante tipo 3

SCD tipo 3 - Agrega atributo

Id_Estudiante	Nombre Completo	Facultad
EST12345	Pepito Perez	Mercadeo

Cambio de facultad a **Ingeniería**

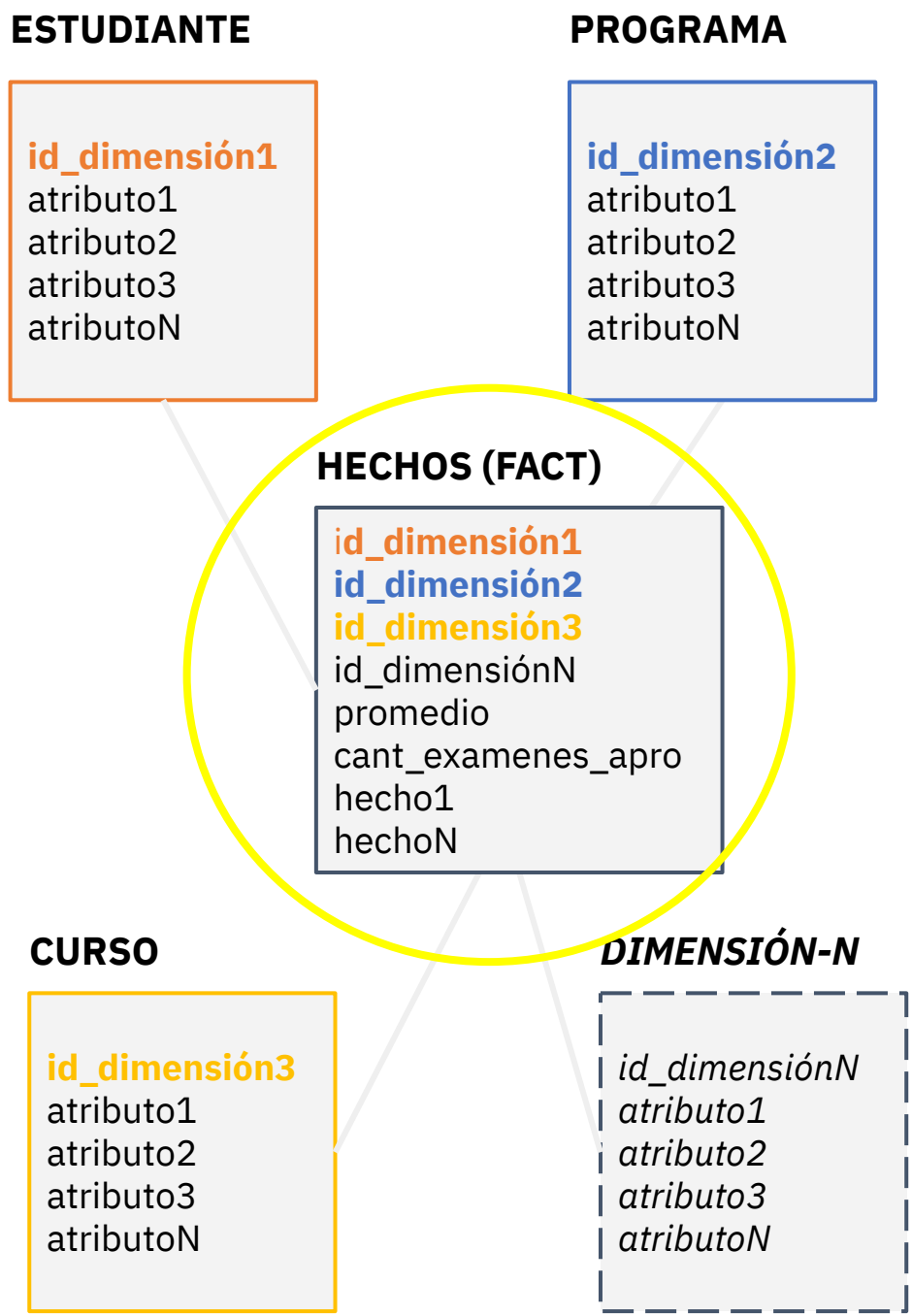
Id_Estudiante	Cod_Estudiante	Nombre Completo	Facultad_old	Facultad_new
1	EST12345	Pepito Perez	Mercadeo	Ingeniería

Solución reto anterior

Tabla de hechos

Hechos (Fact)

- Contienen información cuantitativa de un proceso de negocio.
- **Medidas - Métricas.**
- Contiene claves foráneas (llaves de las dimensiones).



Configuración del Setup

- Postgres:
- Redshift:
- Pentaho:
- BD AdventureWorks: https://wiki.postgresql.org/wiki/Sample_Databases

Identificación de dimensiones y métricas

Preguntas del negocio

El trabajo consiste en hacer preguntas, todas las que se puedan, y hacer frente a la falta de respuestas precisas con una cierta humildad.

Arthur Miller

Preguntas de Negocio

- **Unidades vendidas de cada producto por cliente en un tiempo determinado.**

- **Cantidad de contrataciones por área en un país específico.**

Ejemplo

Organización:
Cadena de supermercados.

Actividad objeto de análisis:
Ventas de productos.

Información registrada sobre una venta:
Del **producto** “crema dental” se han vendido en el **almacén** “Almacén nro.1” el **día** 2/2/2030, 5 **unidades** por un **valor** de $20.

Ejercicio

- ¿**Cuánto** ha sido en cantidades y valores, los descuentos y las ventas netas (venta-descuento), para cada mes y día?
- ¿**Cuánto** ha crecido o disminuido las ventas netas al corte del mes de Marzo del año 2013 para cada vendedor?
- ¿**Cuál** es el producto más vendido a corte del día? Por categoría.
- ¿**Quién** es el cliente que más unidades ha comprado en el último año?

- ¿**Cuánto** ha sido en cantidades y valores, los descuentos y las ventas netas (venta-descuento), para cada mes y día?
- ¿**Cuánto** ha crecido o disminuido las ventas netas al corte del mes de Marzo del año 2013 para cada vendedor?
- ¿**Cuál** es el producto más vendido a corte del día? Por categoría.
- ¿**Quién** es el cliente que más unidades ha comprado en el último año?

Tu reto

- ¿**Dónde** geográficamente se ubican las mayores ventas netas del mes de Diciembre del año 2014?

- ¿**Quién** es el vendedor que más ventas realizó por semestre, teniendo en cuenta el nivel del cargo en el momento de la venta?

- ¿**Dónde** geográficamente, se ubican las mayores ventas netas del mes de Diciembre del año 2014?

- ¿**Quién** es el vendedor que más ventas realizó por semestre, teniendo en cuenta el nivel del cargo en el momento de la venta?

Diseño de modelo

Reglas de negocio

- Crear un campo con el nombre completo del cliente.

- El campo de observación del producto es demasiado largo. Recortar a los 100 primeros caracteres.

- Si un vendedor tiene personas a cargo, marcarlo como beneficiario del bono.

Ejercicio

Documento de mapeo

ETL

Identificar transformaciones

v, h

1, 0

varón, hembra

V, H

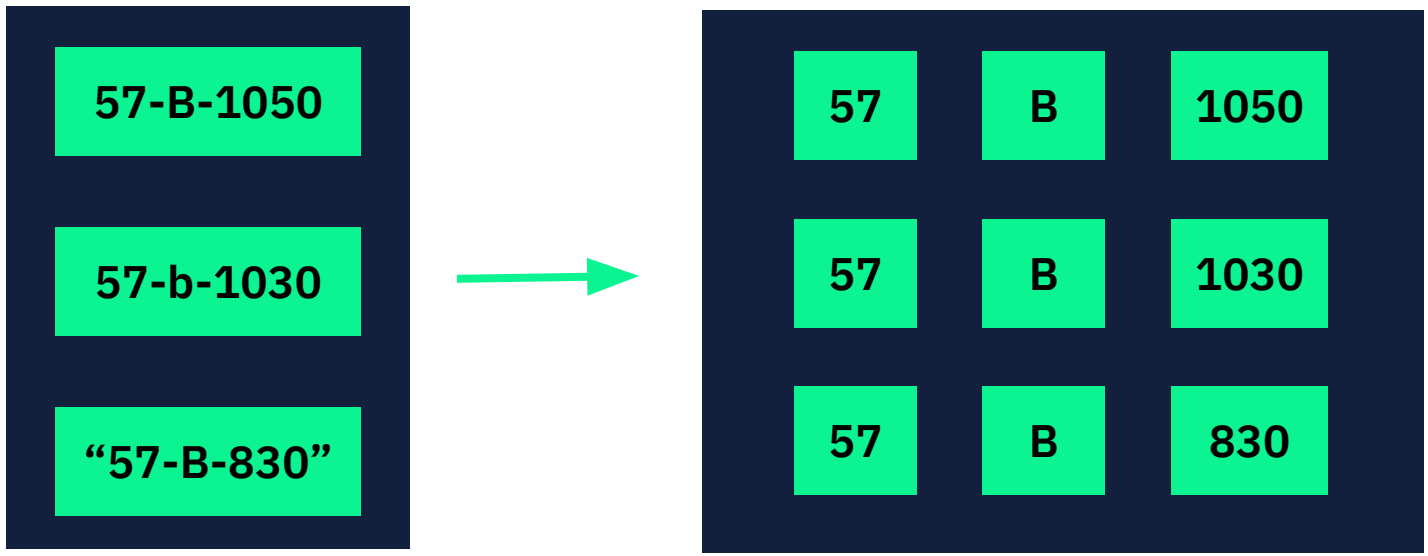
---

dd/mm/yyyy

yyyy-mm-dd

mm-yyyy

dd/mm/yyyy



Código de producto = 57B1050

código país

zona de inventario

código de producto

Tu reto

Creación del modelo físico

ETL

Tu reto

Extracción

ETL - Dimensiones

Transformación

ETL - Dimensiones

Carga

ETL - Dimensiones

Tu reto

Extracción

ETL - Hechos

Transformación

ETL - Hechos

Carga

ETL - Hechos

Orquestar ETL

Ejercicio para el estudiante

Reflexiones y cierre

- Repaso
- Dificultades
- Siguientes pasos: Curso de BI con Power BI

- Enviar el proyecto
- Autoevaluarse en los comentarios
- Evaluar este curso

- Redes
- Tchau