# Factoría de Información Corporativa (CIF)

Una Factoría de Información Corporativa (CIF) es una arquitectura de datos que permite a las organizaciones transformar grandes volúmenes de datos dispersos en información estratégica, confiable y útil para la toma de decisiones. Funciona como una "fábrica de datos", donde se recolecta, limpia, integra, almacena y distribuye la información de forma sistemática y controlada.

En esta presentación, exploraremos los conceptos fundamentales del almacenamiento de datos, desde la arquitectura CIF hasta las implementaciones específicas como Data Lakes, Data Warehouses y Data Marts, analizando sus características, ventajas y casos de uso reales.

**por Kibernum Capacitación S.A.**

# PREGUNTAS DE ACTIVACIÓN

1) ¿Has escuchado antes los términos *Data Lake*, *Data Warehouse* o *Data Mart*? ¿Qué crees que los diferencia?
2) ¿Qué tipo de problemas podría tener una empresa si no organiza bien sus datos para el análisis?
3) En la sesión anterior hablamos de gobernanza y modelado. ¿Por qué crees que es importante tener claridad en los requerimientos del negocio antes de diseñar una solución de almacenamiento?

# ¿Qué es una Factoría de Información Información Corporativa?

### Centralización de datos

Consolida información proveniente de múltiples fuentes en un único entorno integrado, creando una fuente única de verdad para toda la organización.

### Transformación y calidad

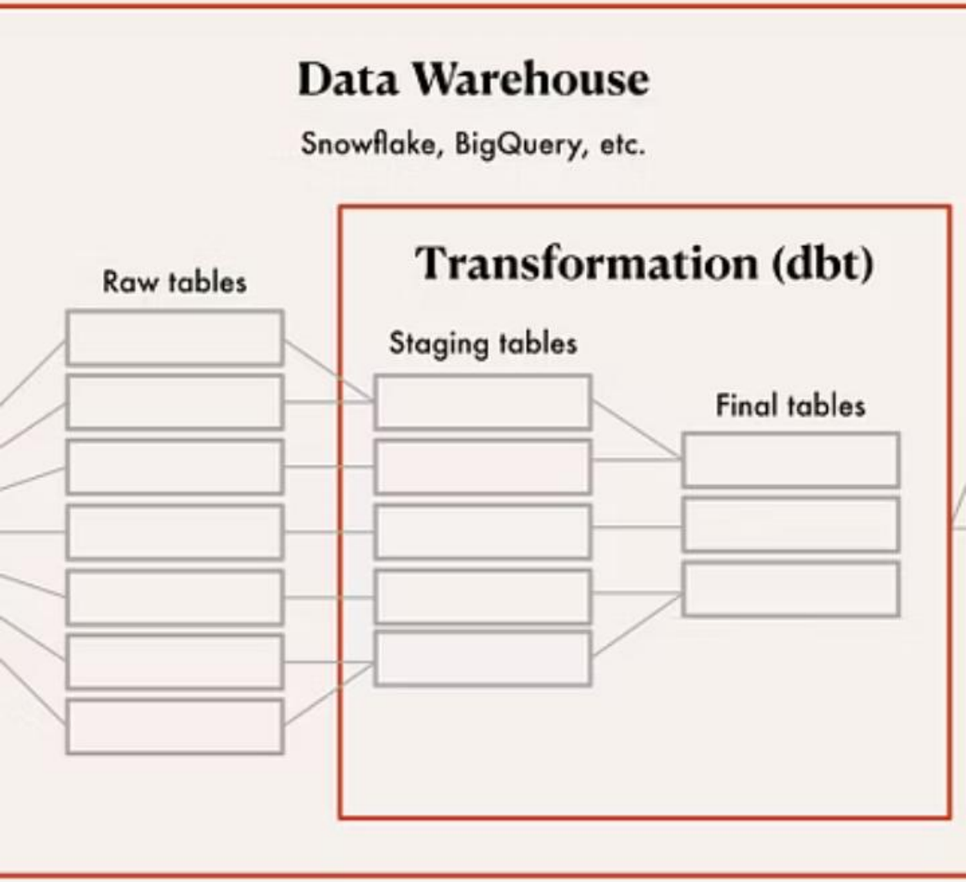
Los datos se transforman mediante procesos de limpieza, validación y enriquecimiento, asegurando que la información sea precisa, consistente y completa.

### Disponibilidad para análisis

Los datos transformados se almacenan en estructuras diseñadas para facilitar el análisis, como Data Marts y Data Warehouses, permitiendo consultas eficientes.

### Automatización de procesos

Permite automatizar tareas rutinarias como la carga de datos (ETL), la generación de reportes y el monitoreo de calidad, reduciendo errores humanos.




# Ecosistema moderno de arquitectura de datos

### Fuentes de datos

Sistemas dispersos como nube, bases de datos, ERP (SAP), archivos y aplicaciones que generan datos operacionales.

### Ingestión

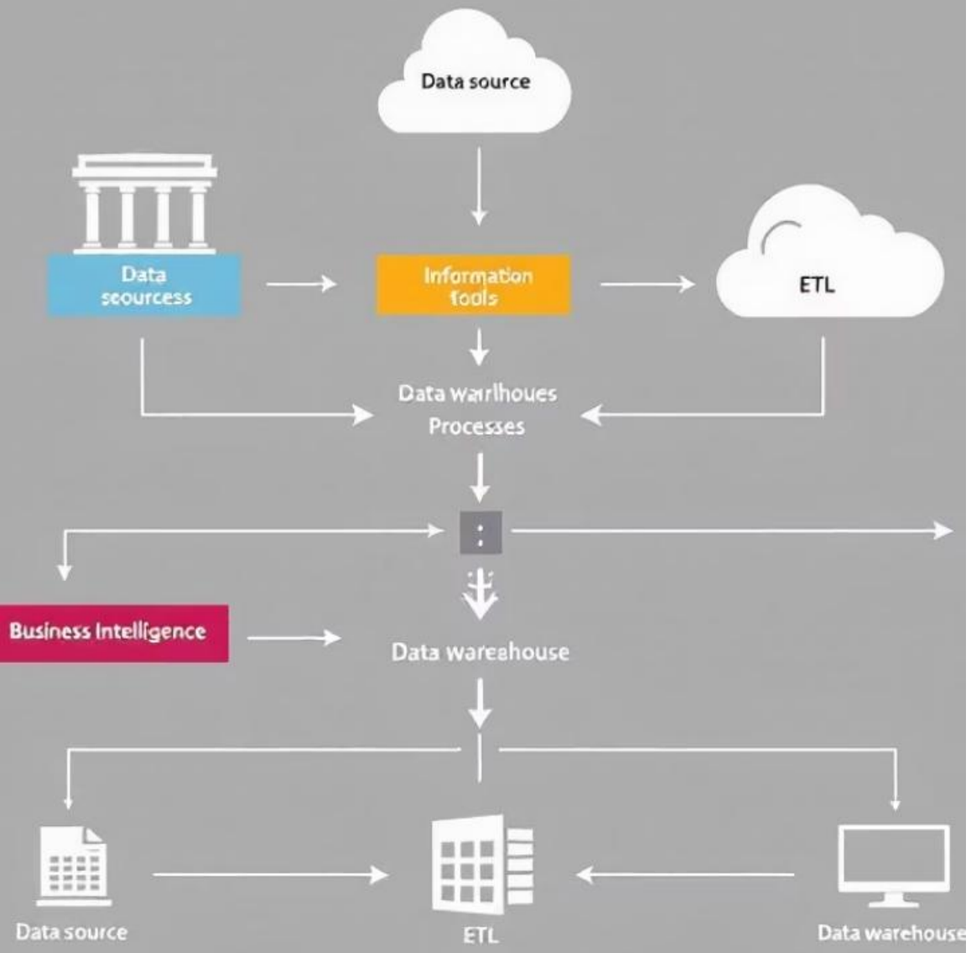
Herramientas como Fivetran, Stitch o Airbyte capturan y canalizan los datos desde las fuentes hacia el almacenamiento.

### Almacenamiento

Data Lake o Data Warehouse donde los datos se conservan de forma estructurada para su posterior procesamiento.

### Análisis y visualización

Herramientas como Power BI, Tableau que permiten explotar los datos para obtener insights y apoyar decisiones.




# Componentes principales de una CIF

### Fuentes de datos

Sistemas transaccionales, hojas de cálculo, aplicaciones externas, sensores IoT y otras fuentes que generan datos operacionales para la organización.

### Procesos ETL

Extracción, Transformación y Carga: procesos encargados de preparar los datos para su análisis, limpiándolos y estructurándolos adecuadamente.

### Almacenamiento estructurado

Data Warehouse corporativo que centraliza los datos transformados y los organiza para facilitar consultas analíticas complejas.

### Data Marts

Subconjuntos de datos orientados a áreas específicas del negocio como marketing, finanzas u operaciones, optimizados para consultas departamentales.

# Beneficios de implementar una CIF

### Mejor calidad de datos datos

Datos consistentes, precisos y confiables que permiten tomar decisiones basadas en información veraz y actualizada.

### Decisiones informadas informadas

Mayor capacidad para tomar decisiones basadas en evidencias, reduciendo la incertidumbre y aumentando la efectividad estratégica.

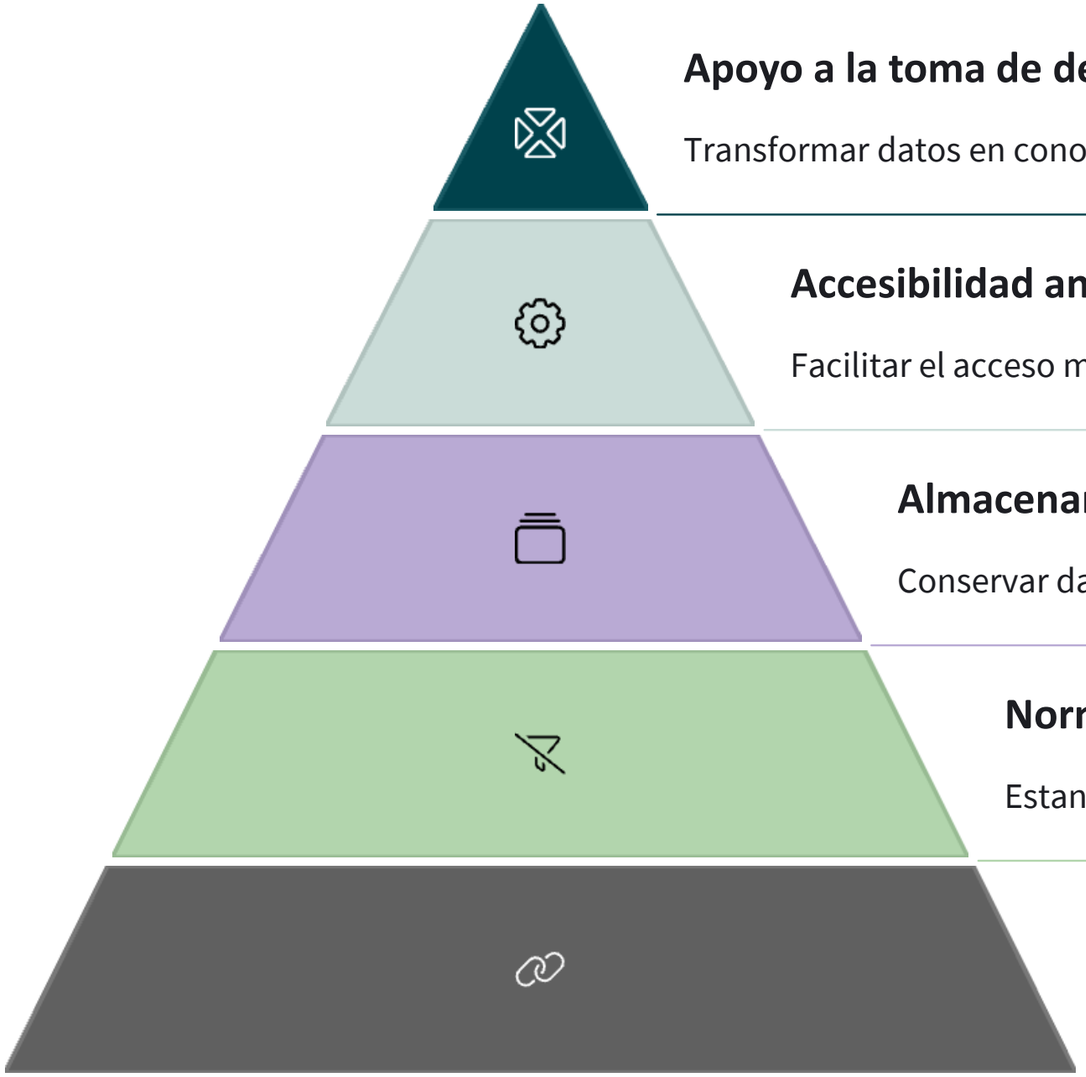
### Eficiencia operativa

Reducción del tiempo y esfuerzo en la preparación de reportes, automatizando procesos que antes requerían intervención manual.

### Autonomía analítica

Mayor independencia para las áreas de negocio en el análisis de información, sin necesidad de depender constantemente del departamento de TI.

# Objetivos de una CIF



**Apoyo a la toma de decisiones**

Transformar datos en conocimiento estratégico

**Accesibilidad analítica**

Facilitar el acceso mediante herramientas BI

**Almacenamiento histórico**

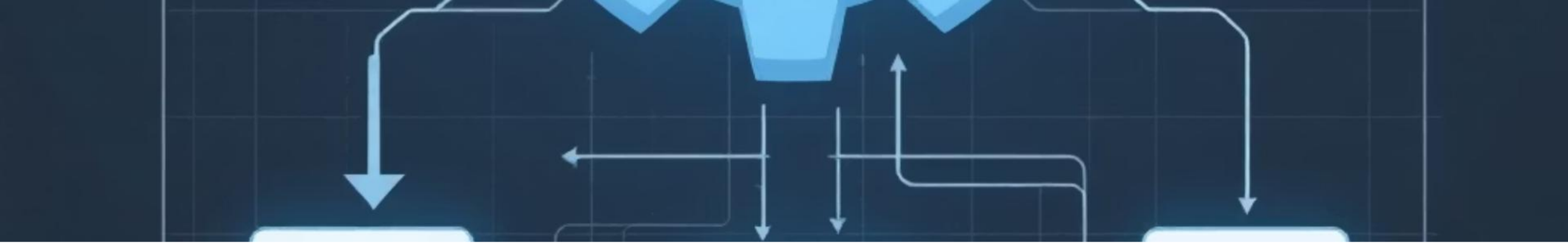
Conservar datos para análisis longitudinal

**Normalización de datos**

Estandarizar para garantizar coherencia

**Integración de fuentes**

Reunir datos de plataformas heterogéneas



# Principios fundamentales de una CIF

| Principio | Descripción |
| --- | --- |
| Centralización | Los datos se concentran en un entorno único que actúa como fuente de verdad. |
| Separación operativa/analítica | Se distingue entre los sistemas transaccionales (OLTP) y analíticos (OLAP). |
| Estandarización | Se aplican reglas uniformes de nombres, formatos y estructuras de datos. |
| Modularidad | La CIF se construye por capas, cada una con funciones específicas y bien definidas. |
| Escalabilidad | La arquitectura puede crecer con el tiempo sin perder rendimiento ni control. |
| Gobernanza de datos | Se aplican políticas, roles y responsabilidades claras para el manejo de datos. |

# Datos operacionales vs. datos analíticos

## Datos operacionales (OLTP)

Son los datos que generan los sistemas transaccionales (ERP, CRM, POS, etc.) y que reflejan operaciones en tiempo real como ventas, pagos o reservas.

- Alta frecuencia de actualización (segundos/minutos)
- Estructura cruda y detallada
- Orientados a registrar transacciones individuales
- Ejemplo: registro de una venta específica

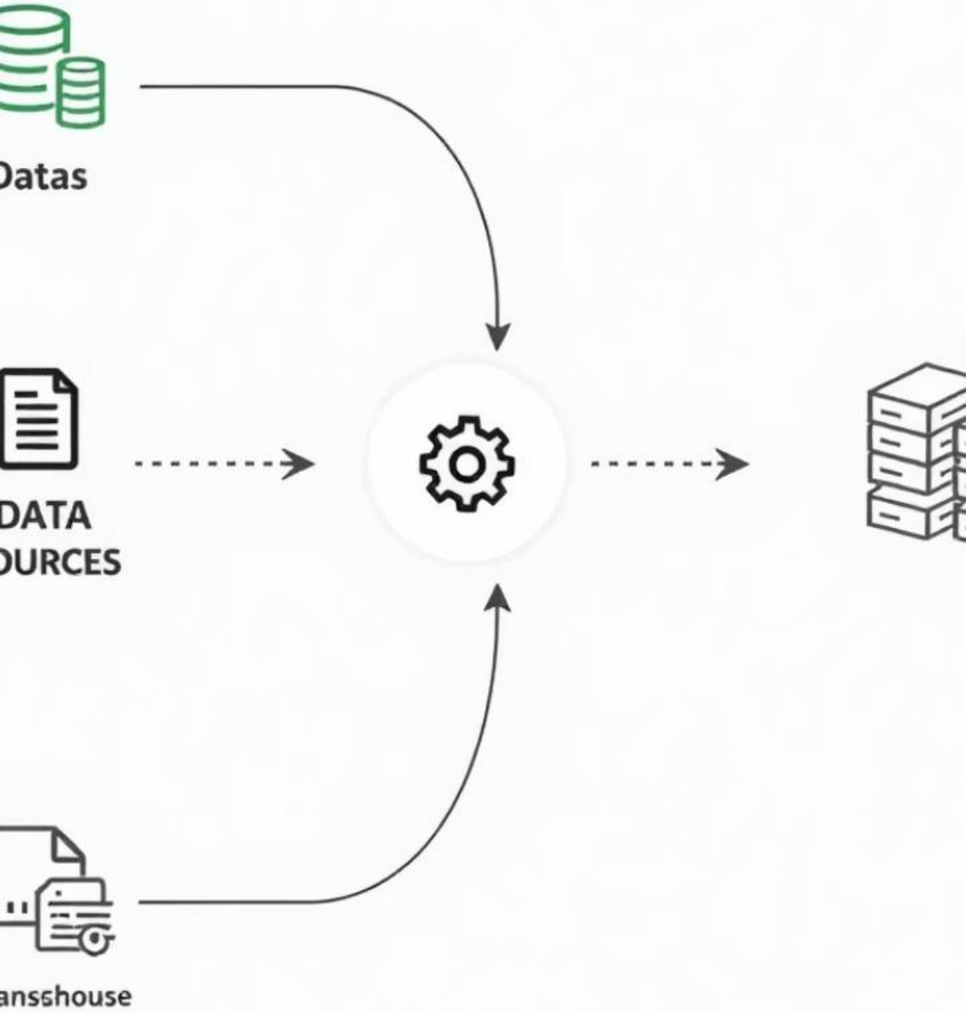
## Datos analíticos (OLAP)

Son los datos integrados, transformados y almacenados específicamente para análisis, reportes y toma de decisiones estratégicas.

- Actualización periódica (día/semana/mes)
- Estructura agrupada y resumida
- Orientados a analizar tendencias y patrones
- Ejemplo: ventas mensuales por región

# Comparativa: Datos operacionales operacionales vs. analíticos

| Característica | Datos operacionales (OLTP) | Datos analíticos (OLAP) |
|---|---|---|
| Origen | Sistemas transaccionales (ERP, CRM, POS, e-commerce) | Data Warehouse, Data Marts |
| Finalidad | Registrar operaciones en tiempo real | Analizar tendencias, comportamientos, KPIs |
| Estructura de archivos | Cruda, detallada, tabla por evento | Agrupada, resumida, estructurada por dimensiones |
| Ejemplo de dato | {"fecha": "2025-03-31", "cliente_id": 483, "total": 24990} | Ventas marzo 2025 = $893.500 distribuidas en 5 regiones |




# Flujo de datos en la CIF

## Extracción

Los datos se toman desde múltiples fuentes como bases de datos, archivos y APIs, manteniendo su formato original.

## Transformación

Los datos extraídos se limpian, filtran, combinan y validan para asegurar su calidad y coherencia.

## Carga

Los datos transformados se integran en estructuras de almacenamiento analítico como Data Warehouse y Data Marts.

## Análisis

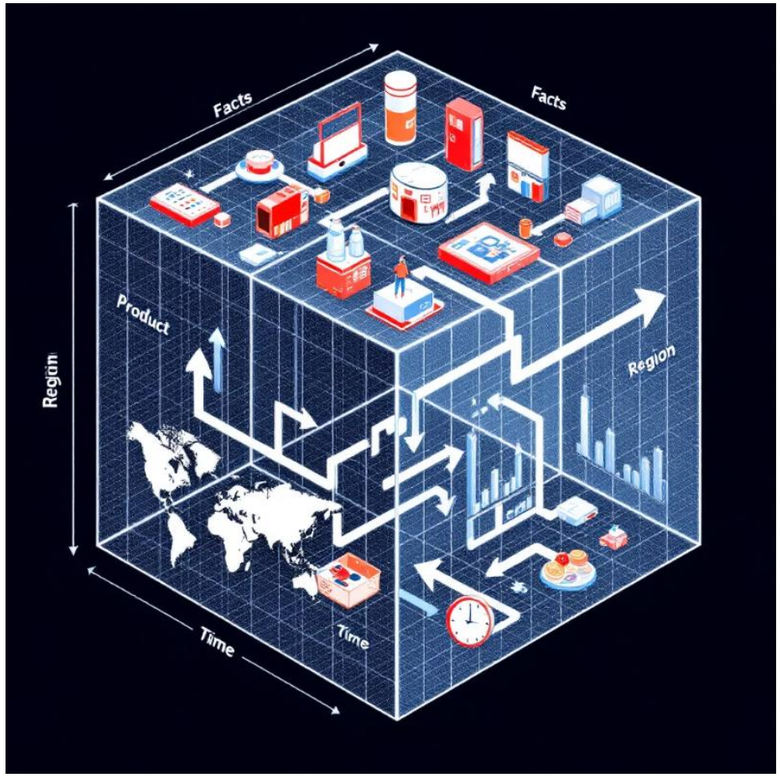
Los datos almacenados se utilizan para generar reportes, dashboards y análisis que apoyan la toma de decisiones.

# Modelos de datos en una CIF

## Modelo multidimensional

Basado en hechos (medidas numéricas) y dimensiones (categorías de análisis). Por ejemplo, ventas (hecho) analizadas por región, producto y tiempo (dimensiones).

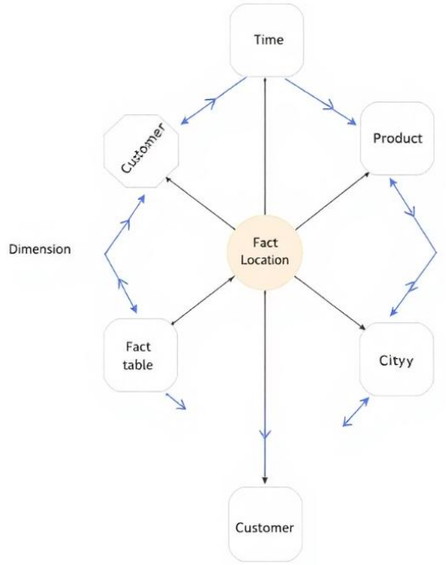
Este modelo facilita el análisis OLAP, permitiendo "navegar" por los datos desde diferentes perspectivas y niveles de detalle.




## Modelo en estrella o snowflake

Son los esquemas más comunes en el diseño de Data Warehouses. El modelo en estrella tiene una tabla central de hechos conectada a tablas de dimensiones.

El modelo snowflake es similar, pero con dimensiones normalizadas en múltiples tablas relacionadas, creando una estructura más compleja pero normalizada.



# Áreas funcionales de una CIF

### Staging area

Zona de recepción y carga temporal de datos crudos, donde se realiza la validación inicial antes de su procesamiento.

### Herramientas de acceso

Interfaces que permiten a los usuarios finales consumir la información mediante reportes, dashboards y consultas ad hoc.

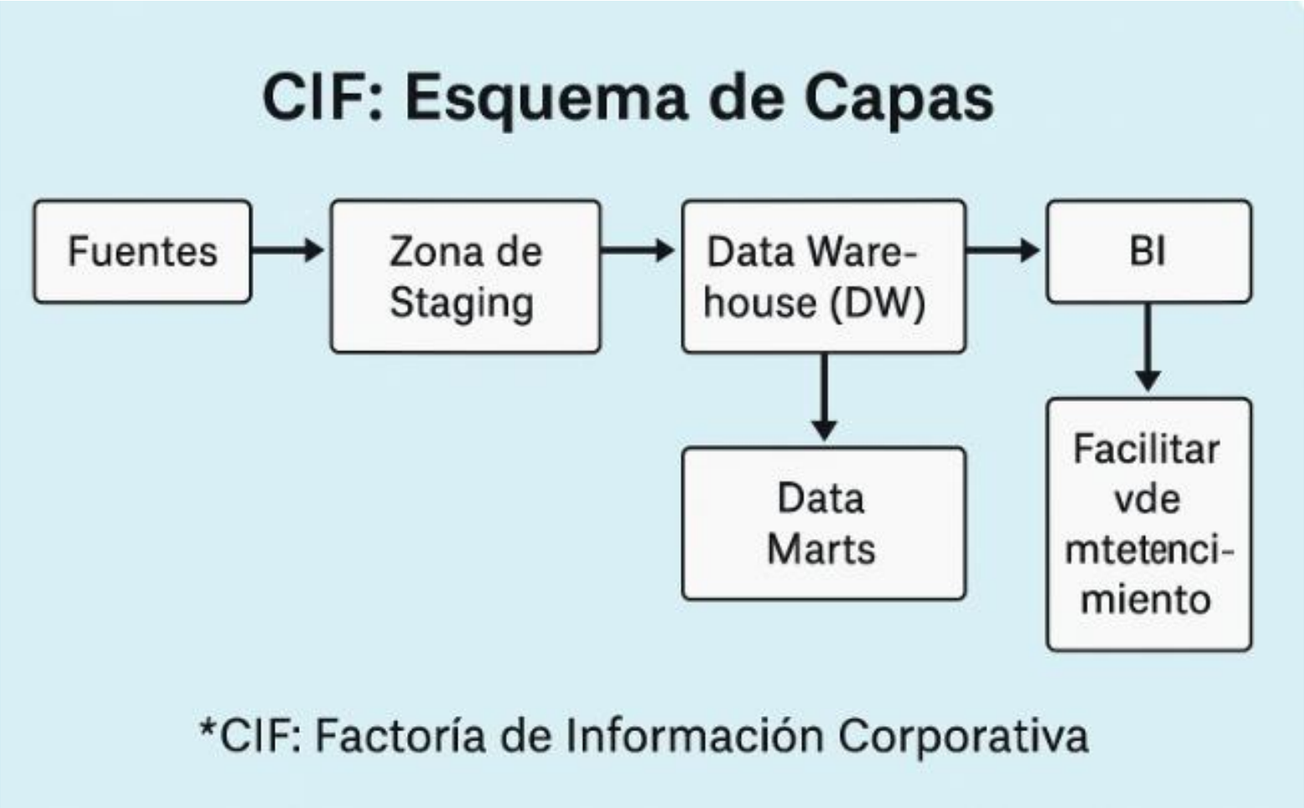

### Data warehouse corporativo corporativo

Almacenamiento central de datos consolidados y confiables, estructurados para análisis histórico y estratégico.

### Data marts

Subconjuntos de datos organizados por unidad de negocio (ventas, finanzas, marketing), optimizados para consultas específicas.

# Áreas funcionales de una CIF



# Capas de una CIF

1

## Capa Operacional

Captura datos desde sistemas fuente como ERP, CRM, sensores IoT

## Capa de Staging

Recibe datos en forma original y los prepara para procesamiento

## Capa de Data Warehouse

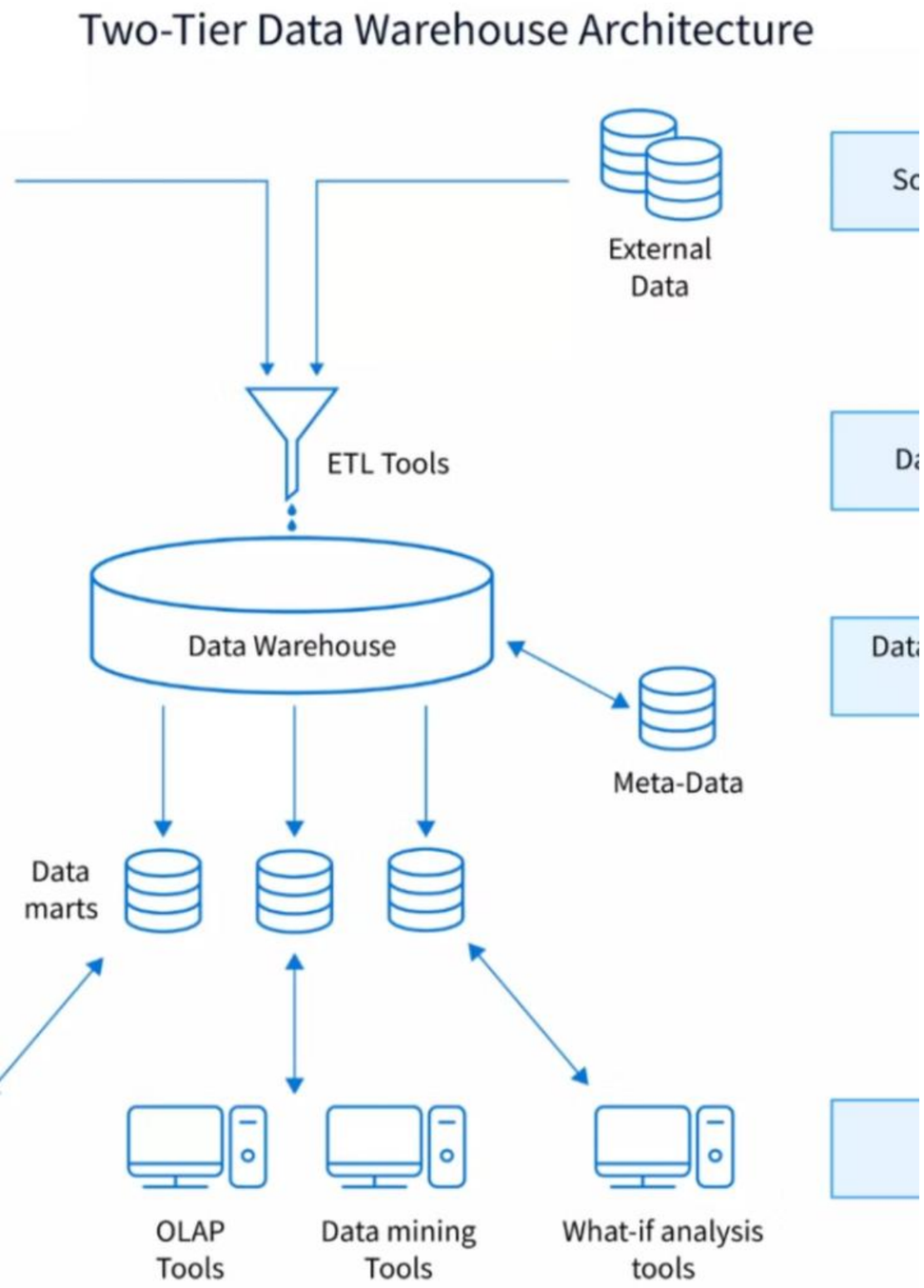
Almacena datos limpios y estructurados para análisis históricos

## Capa de Data Marts

Segmenta datos por áreas de negocio para consultas específicas

## Capa de Herramientas BI

Permite acceso, visualización y análisis de la información




# Representación visual de capas CIF

**1** **Fuentes de datos**

Sistemas operacionales que generan datos transaccionales como ventas, inventario, clientes y otras operaciones del negocio.

**2** **Integración y transformación**

Procesos ETL que extraen, transforman y cargan los datos, asegurando su calidad y consistencia para el análisis.

**3** **Almacenamiento analítico**

Repositorios estructurados como Data Warehouse y Data Marts que organizan la información para facilitar consultas complejas.

**4** **Presentación y análisis**

Herramientas de visualización y análisis que permiten a los usuarios finales explorar los datos y obtener insights valiosos.

# ¿Qué es un Data Lake?

Un Data Lake es un repositorio centralizado que permite almacenar grandes volúmenes de datos en su formato original, ya sean estructurados (tablas), semiestructurados (JSON, XML, CSV) o no estructurados (imágenes, audios, textos, logs, videos).

## Almacenamiento flexible

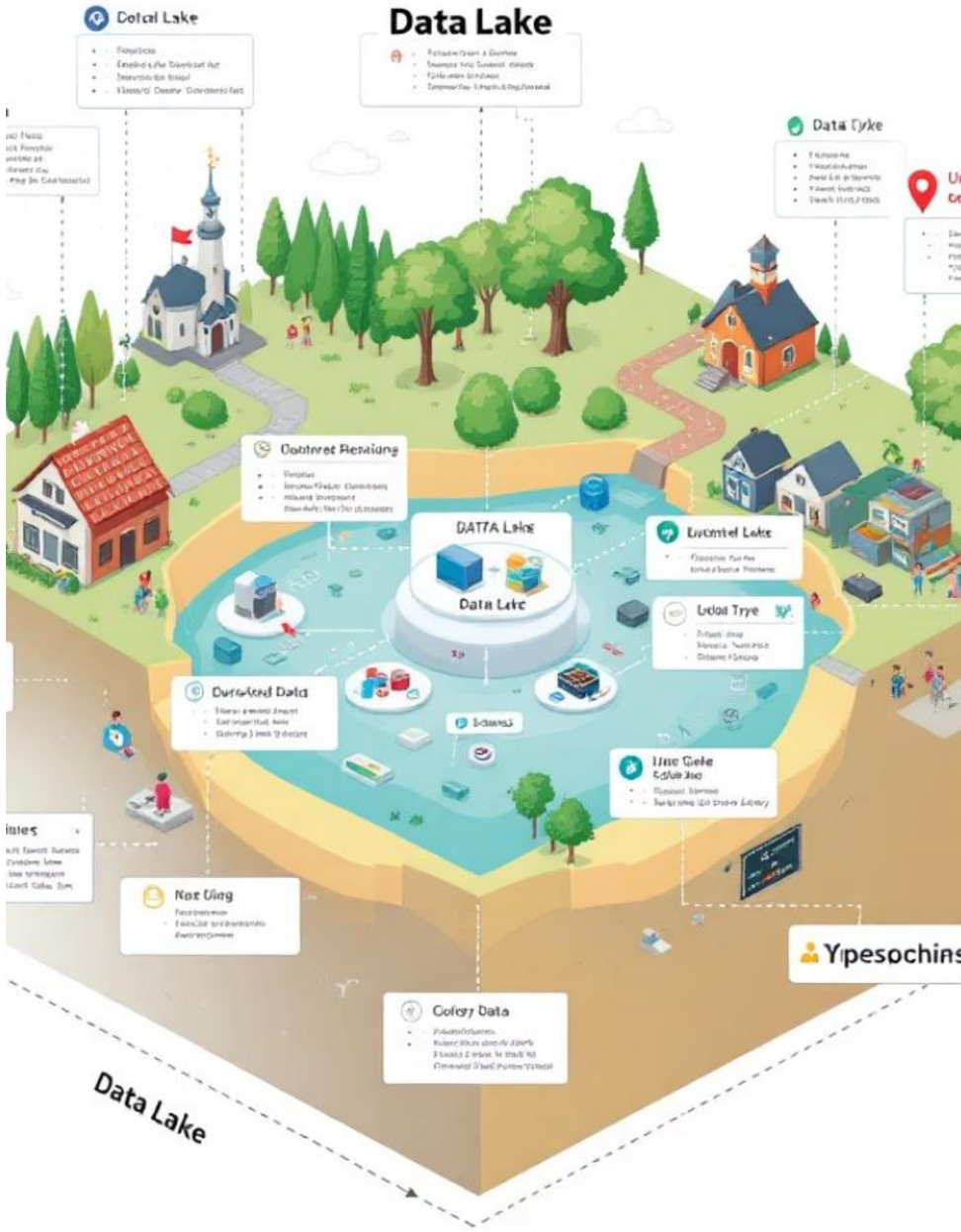
Acepta datos en su formato nativo, sin necesidad de transformación previa, permitiendo conservar toda la información original.

## Escalabilidad masiva

Diseñado para crecer a petabytes de información, especialmente en entornos cloud con costos optimizados por volumen.

## Exploración avanzada

Base para análisis exploratorio, ciencia de datos y entrenamiento de modelos de machine learning e inteligencia artificial.



# Arquitectura de un Data Lake

### Raw Zone

Almacena los datos sin procesar, exactamente como llegan de la fuente, preservando su formato original y todos los detalles, sin filtrado ni transformación.

### Cleansed Zone

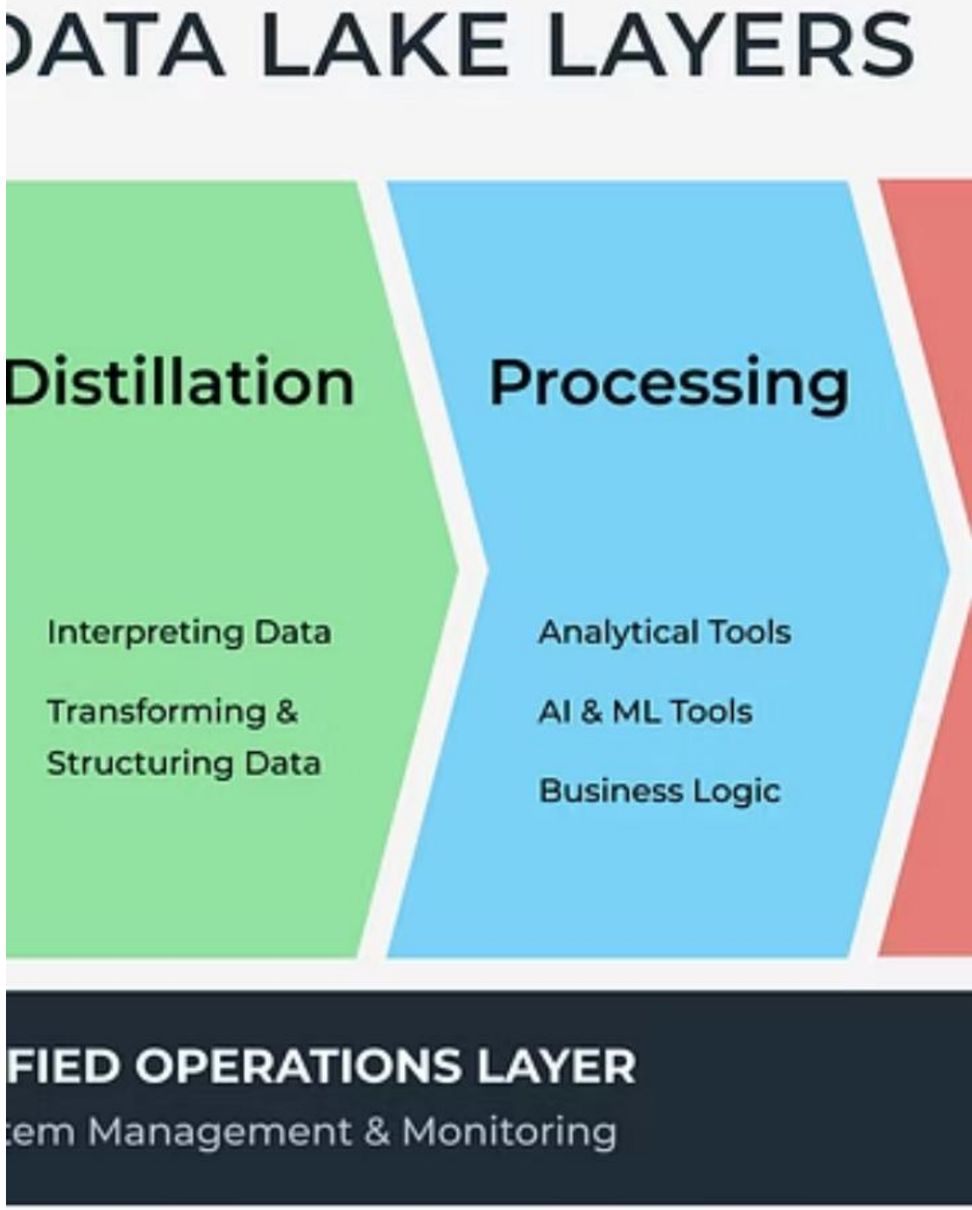
Contiene datos que han pasado por procesos de limpieza básica, eliminando errores, duplicados y valores anómalos para mejorar su calidad.

### Curated Zone

Alberga datos transformados y enriquecidos, organizados según modelos específicos y preparados para análisis o visualización.

### Analytics Zone

Ofrece datasets específicos optimizados para dashboards, modelos de IA o reportes, facilitando su consumo por aplicaciones analíticas.



# Componentes tecnológicos de un Data Lake

| Componente | Ejemplos comunes | Función |
|---|---|---|
| Almacenamiento | Amazon S3, Azure Data Lake Storage, Google Cloud Storage | Repositorio principal donde se guardan los datos en su formato nativo |
| Procesamiento | Apache Spark, Databricks, AWS Glue, Hadoop | Motores que transforman y analizan los datos almacenados |
| Consulta | Athena, Presto, Hive, Notebooks Jupyter | Herramientas para explorar y extraer información de los datos |
| Gobernanza | AWS Lake Formation, Apache Ranger, Amundsen, Glue Data Catalog | Control de acceso, catalogación y gestión de metadatos |

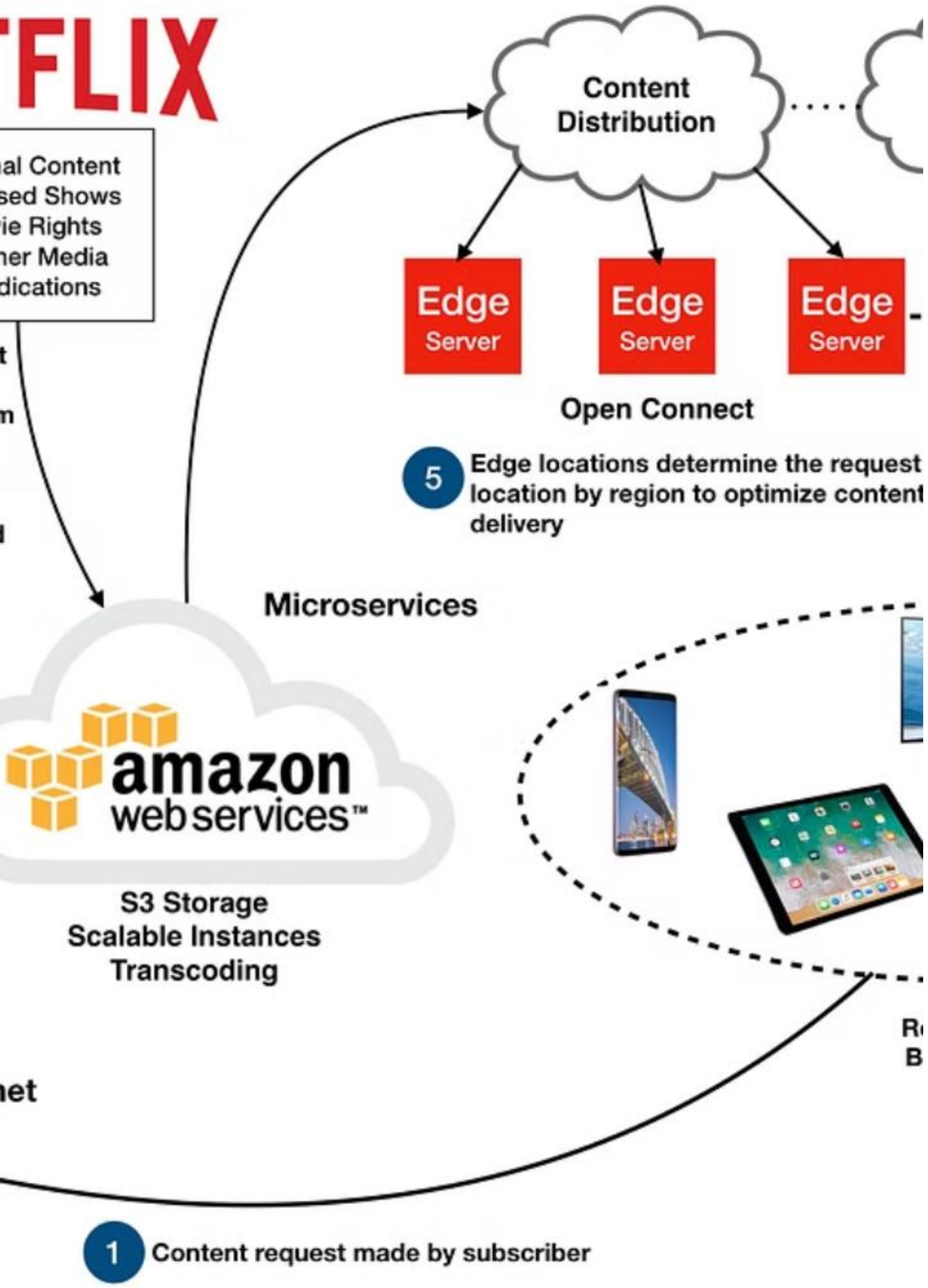
# Ventajas y desventajas del Data Lake

## Ventajas

- Bajo costo de almacenamiento, especialmente en la nube
- Flexibilidad para almacenar cualquier tipo de dato sin esquema previo
- Escalabilidad prácticamente ilimitada para grandes volúmenes
- Base ideal para proyectos de ciencia de datos y machine learning
- Preservación de datos en su formato original, sin pérdida de información

## Desventajas

- Riesgo de convertirse en un "Data Swamp" sin control ni organización
- Requiere habilidades técnicas avanzadas para su administración
- No es óptimo para reportes estructurados y consultas frecuentes
- Gobernanza y linaje de datos complejos si no se planifican adecuadamente
- Mayor tiempo de preparación para análisis comparado con datos estructurados




# Casos de uso: Data Lake en Netflix

### Almacenamiento de logs logs

Netflix almacena logs de navegación y reproducción en un Data Lake en AWS S3, capturando millones de eventos por segundo sobre el comportamiento de los usuarios en la plataforma.

### Entrenamiento de modelos modelos

Con esta información masiva, entrena modelos de recomendación personalizados para cada usuario, mejorando la experiencia y aumentando el tiempo de visualización.

### Tecnologías utilizadas

La arquitectura se basa en AWS S3 para almacenamiento, AWS Glue para procesamiento ETL y Amazon SageMaker para el entrenamiento de modelos predictivos.

# Casos de uso: Data Lake en IoT industrial industrial

### Sensores industriales

Siemens utiliza Data Lakes en Azure para almacenar datos de sensores en plantas industriales, capturando millones de mediciones por minuto.

### Mantenimiento predictivo

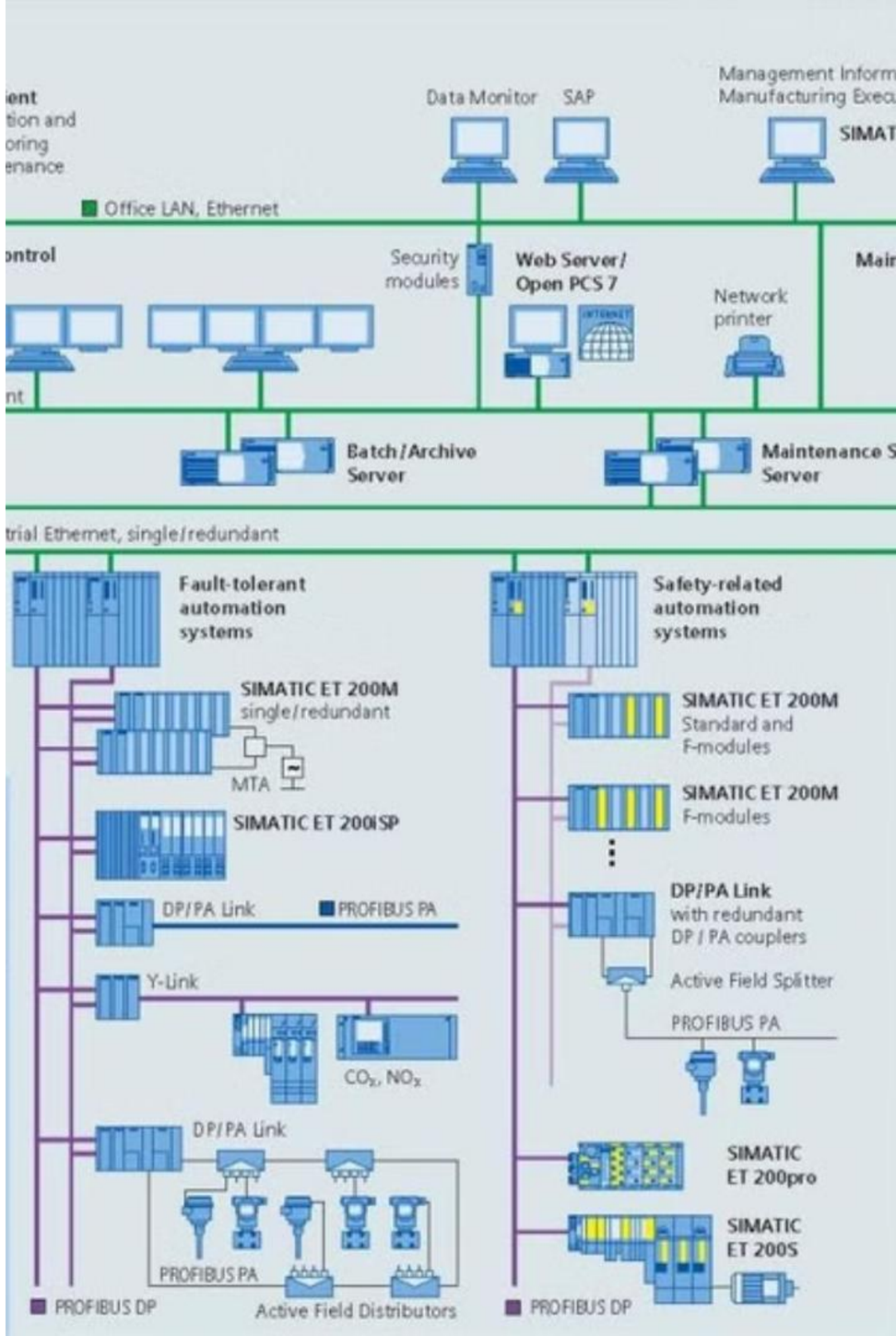
Los datos se analizan para predecir fallas en equipos antes de que ocurran, reduciendo tiempos de inactividad y costos de reparación.

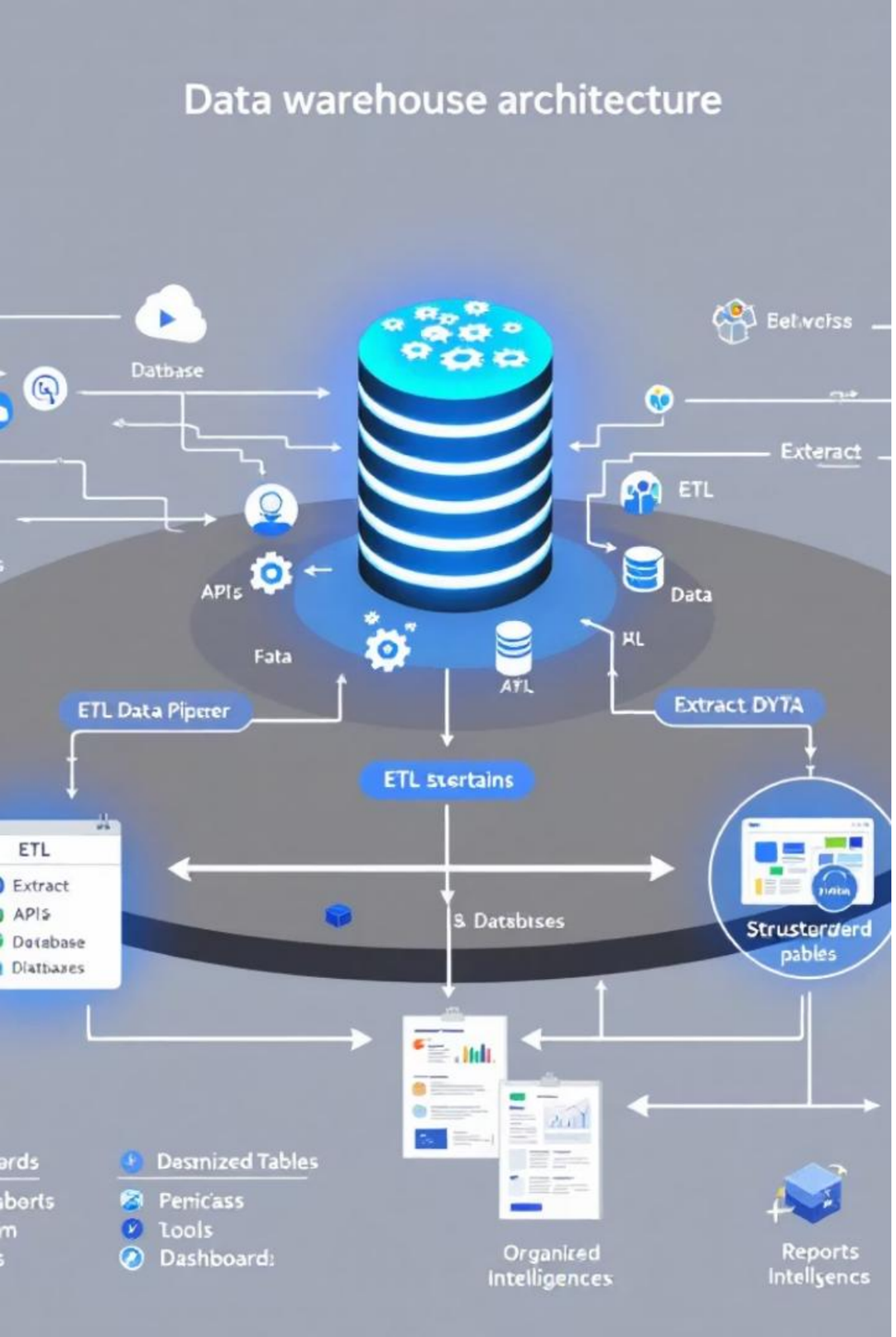
### Tecnología cloud

La solución se basa en Azure Data Lake Storage, Stream Analytics y Azure Machine Learning para procesar y analizar los datos en tiempo real.

### Resultados tangibles tangibles

Esta implementación ha permitido reducir hasta un 30% los costos de mantenimiento y aumentar la disponibilidad de los equipos industriales.






# ¿Qué es un Data Warehouse?

Un Data Warehouse (almacén de datos) es un repositorio centralizado y estructurado donde se almacenan datos transformados, limpios y organizados específicamente para el análisis y la toma de decisiones empresariales.

1

### Fuente única de verdad

Centraliza datos de múltiples sistemas en un repositorio confiable y consistente para toda la organización.

### Datos estructurados

Almacena información limpia, transformada y organizada según modelos dimensionales optimizados para consultas.

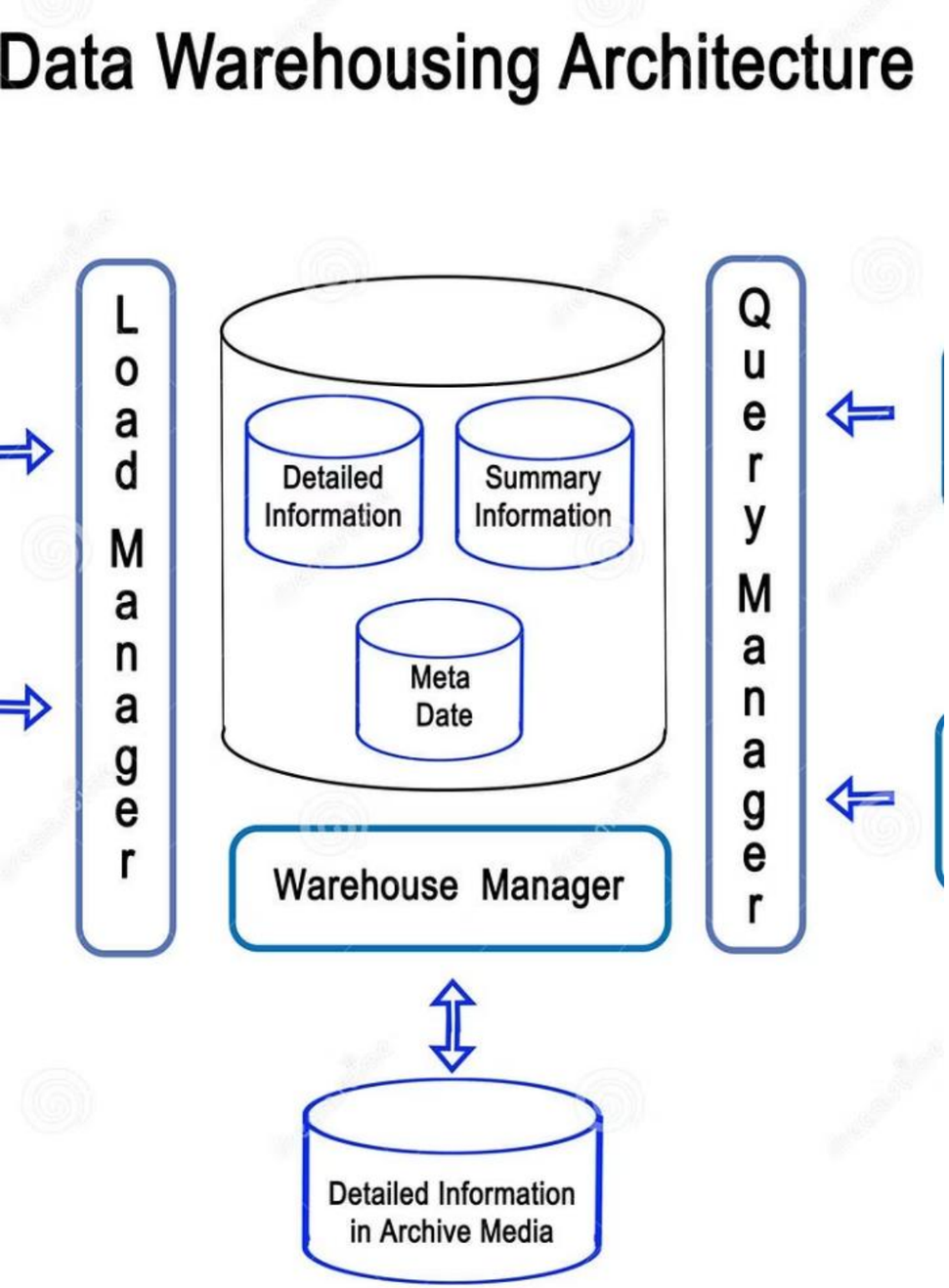
### Perspectiva histórica

Conserva datos históricos para análisis de tendencias y comparativas temporales, facilitando decisiones estratégicas.

4

### Orientado a análisis

Diseñado específicamente para consultas analíticas complejas, reportes y herramientas de Business Intelligence.




# Arquitectura de un Data Warehouse

### Integración de fuentes

Combina datos de múltiples orígenes como ERP, CRM, e-commerce y sistemas de recursos humanos en un repositorio unificado y coherente.

### Almacenamiento optimizado

Utiliza estructuras de almacenamiento por columnas (columnar storage) y técnicas de compresión para acelerar consultas analíticas complejas.

### Modelado dimensional

Organiza la información en esquemas tipo estrella o copo de nieve, con tablas de hechos y dimensiones que facilitan el análisis multidimensional.

### Temporalidad histórica

Conserva versiones y datos históricos para análisis longitudinal, permitiendo comparar rendimiento a lo largo del tiempo.

# Capas de un Data Warehouse

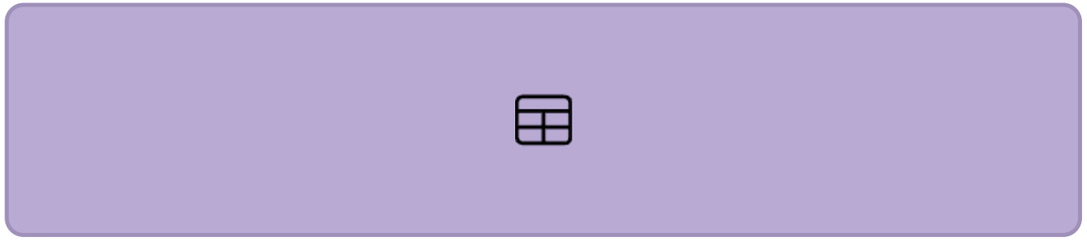
### ETL Layer

Extracción, transformación y carga de datos desde fuentes externas

2

### Enterprise DW

Almacenamiento principal de datos integrados y normalizados



### Presentation Layer

Vistas y estructuras optimizadas para consumo analítico

# Ventajas y desventajas del Data Warehouse

## Ventajas

- Datos consistentes, limpios y auditables para toda la organización
- Ideal para reportes repetitivos y estandarizados con alto rendimiento
- Soporte robusto para decisiones estratégicas basadas en datos históricos
- Compatibilidad nativa con herramientas de Business Intelligence
- Alta confiabilidad y seguridad para información crítica del negocio

## Desventajas

- Mayor complejidad técnica y costos de implementación inicial
- Requiere inversión significativa de tiempo y recursos especializados
- Menor flexibilidad ante cambios en las fuentes o requisitos de datos
- No es ideal para datos no estructurados como imágenes o texto libre
- Puede requerir licencias costosas para plataformas comerciales

# Caso de uso: Bradesco y su Data Warehouse

## 1

### Monitoreo de KPIs

Seguimiento de indicadores clave como ingresos, transacciones y eficiencia operativa por canal bancario

## 2

### Análisis financiero

Generación de informes consolidados, detección de riesgos y evaluación de rentabilidad por segmento

## 3

### Cumplimiento normativo

Consolidación de datos para reportes regulatorios requeridos por organismos financieros

## 4

### Cuadros de mando

Visualización de datos estratégicos en tiempo real para la toma de decisiones directivas

# ¿Qué es un Data Mart?

Un Data Mart es un subconjunto de un Data Warehouse que almacena datos relevantes para un área específica del negocio, como ventas, finanzas, marketing o recursos humanos, proporcionando una vista simplificada y optimizada para sus necesidades particulares.

### Especialización temática

Contiene datos específicos de una unidad de negocio, facilitando el acceso a información relevante para cada departamento.

### Rapidez de consulta

Ofrece tiempos de respuesta más bajos gracias a su menor volumen y complejidad, optimizando el rendimiento analítico.

### Autonomía departamental

Permite a los equipos de negocio acceder a su información sin depender constantemente del departamento de TI.

### Flexibilidad de diseño

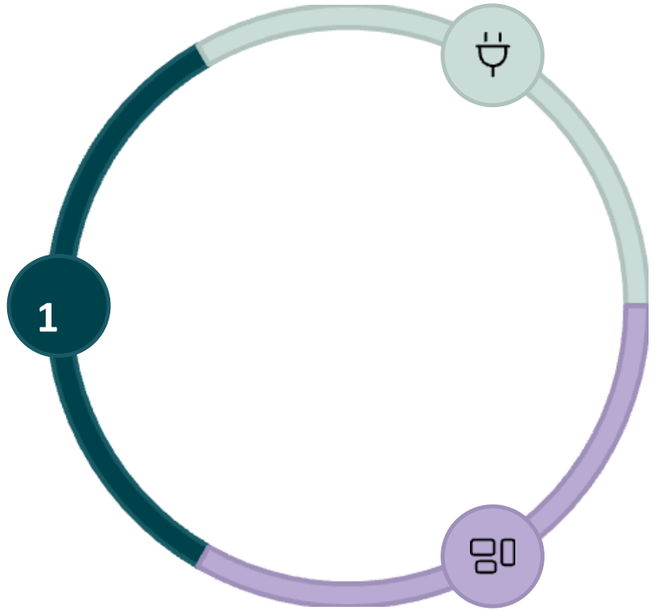
Puede implementarse como dependiente (alimentado desde el DW) o independiente (directamente desde fuentes).

# Tipos de Data Marts

### Data Marts dependientes

Se alimentan exclusivamente desde el Data Warehouse corporativo, garantizando consistencia con la fuente central de datos.

Ventaja: mantienen la integridad y coherencia con el resto de la organización.




### Data Marts independientes

Se alimentan directamente desde los sistemas fuente, sin depender de un Data Warehouse central.

Ventaja: mayor autonomía y rapidez de implementación para necesidades departamentales urgentes.

### Data Marts híbridos

Combinan datos del Data Warehouse corporativo con otras fuentes específicas del departamento.

Ventaja: equilibrio entre consistencia corporativa y flexibilidad departamental.

# Ventajas y desventajas de los Data Marts

## Ventajas

- Consultas más rápidas y eficientes por su volumen reducido
- Personalización por área o equipo según necesidades específicas
- Facilidad de uso para usuarios no técnicos del departamento
- Implementación más rápida que un Data Warehouse completo
- Mayor autonomía para los equipos departamentales

## Desventajas

- Riesgo de duplicidad en la lógica de negocio entre departamentos
- Posibles inconsistencias en definiciones entre diferentes áreas
- Menor flexibilidad cuando el análisis requiere datos de múltiples dominios
- Puede dificultar la gobernanza centralizada de datos
- Riesgo de crear "silos de datos" aislados en la organización

# Casos de uso de Data Marts


## Marketing

Métricas de campañas como CTR, tasa de conversión, engagement por red social y ROI, permitiendo optimizar estrategias y presupuestos publicitarios.

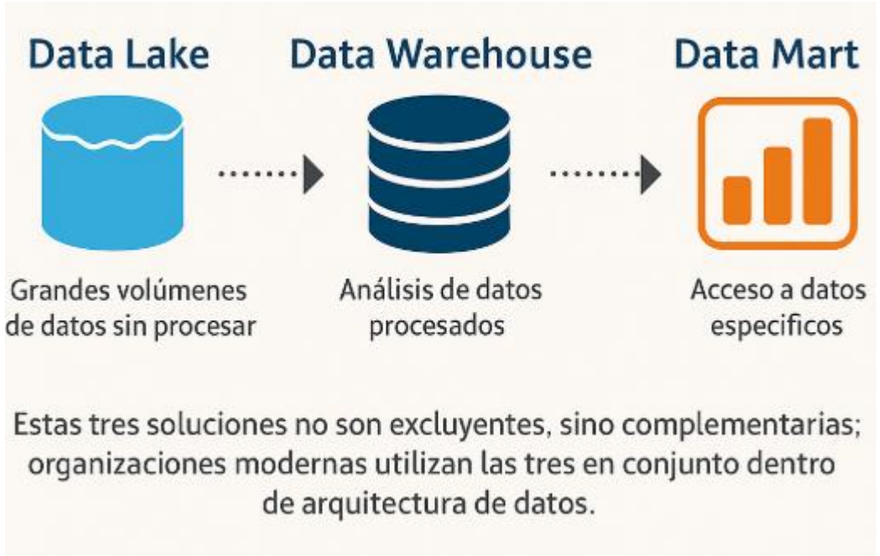

## Ventas

Análisis comparativo entre sucursales, desempeño por región, evolución mensual y cumplimiento de cuotas, facilitando la gestión comercial.


## Recursos Humanos

Dashboards con indicadores de rotación de personal, asistencia, evaluación de desempeño y costos laborales por departamento.

# ¿Cuándo usar cada opción?




| Criterio | Data Lake | Data Warehouse | Data Mart |
|---|---|---|---|
| Tipo de datos | Cualquier tipo (estructurado o no) | Estructurados y normalizados | Datos filtrados y específicos por área |
| Usuarios principales | Científicos de datos, ingenieros | Analistas, equipos BI, gerencia | Supervisores, áreas de negocio |
| Objetivo | Experimentación, exploración | Análisis histórico, KPIs | Reportes operativos y tácticos |
| Velocidad de consulta | Baja sin optimización | Alta (consultas optimizadas) | Muy alta (por volumen reducido) |

# Desafío: Análisis de Caso de Uso de Almacenamiento de Datos

**Objetivo:**

Explorar un caso de uso real en tu sector de interés (salud, retail, banca, transporte, educación, etc.) y analizar cómo se implementan soluciones de almacenamiento de datos, su propósito y las tecnologías involucradas.

**Instrucciones:**

Investiga y responde las siguientes preguntas con base en un caso real que puedas encontrar en artículos, blogs técnicos, estudios de caso o documentación oficial:

1) **¿Qué tipo de solución de almacenamiento se está utilizando?**
2) **¿Qué problema de negocio busca resolver?**
3) **¿Cómo se estructura el flujo de datos y qué tecnologías se mencionan?**

**Tiempo estimado:** 40 minutos.

**Modalidad:** Grupal.

Al finalizar, expondrás tus respuestas.

# Enlaces de Interés

- Centralización de datos: sepa lo que es y entienda su importancia
- ¿Qué es la transformación de datos?
- ¿Qué es ETL (Extraer, Transformar, Cargar)? Ingles sub es
- Kimball vs Inmon – Modelos de diseño de Data Warehouses
- Data warehouse_Módulo 2_La factoría de información corporativa.pdf
- *IBM (2024). La escalera de la IA: Desmitificar los desafíos de la IA.*
- AWS – ¿Qué es un Data Lake?
- IBM – Data Lake & AI Ladder (Video)
- Video: ¿Qué es un Data Warehouse? – IBM (YouTube)

# Preguntas para analizar

1. ¿Qué ventajas podría obtener una empresa al combinar distintas soluciones de almacenamiento (como Data Lake y Data Warehouse) en una arquitectura híbrida?

2. ¿Qué rol crees que juegan las áreas no técnicas (como marketing, ventas o finanzas) en la definición de una arquitectura de datos?

3. ¿Cómo influye la calidad de los datos almacenados en el éxito de las estrategias de inteligencia de negocio?