

Introducción a la Arquitectura de Datos

La arquitectura de datos es una disciplina clave dentro de la gestión de datos en las organizaciones. Consiste en diseñar y estructurar cómo se almacenan, acceden, procesan y protegen los datos para que cumplan con los objetivos estratégicos y operativos del negocio.

Este diseño no solo considera la tecnología utilizada, sino también los procesos, roles y responsabilidades asociados al ciclo de vida de los datos, actuando como un plano o mapa de alto nivel que establece las políticas, estándares y reglas necesarias para el uso efectivo, seguro y eficiente de los datos en toda la organización.

 **por Kibernetum Capacitación S.A.**



Preguntas de Activación de Contenidos



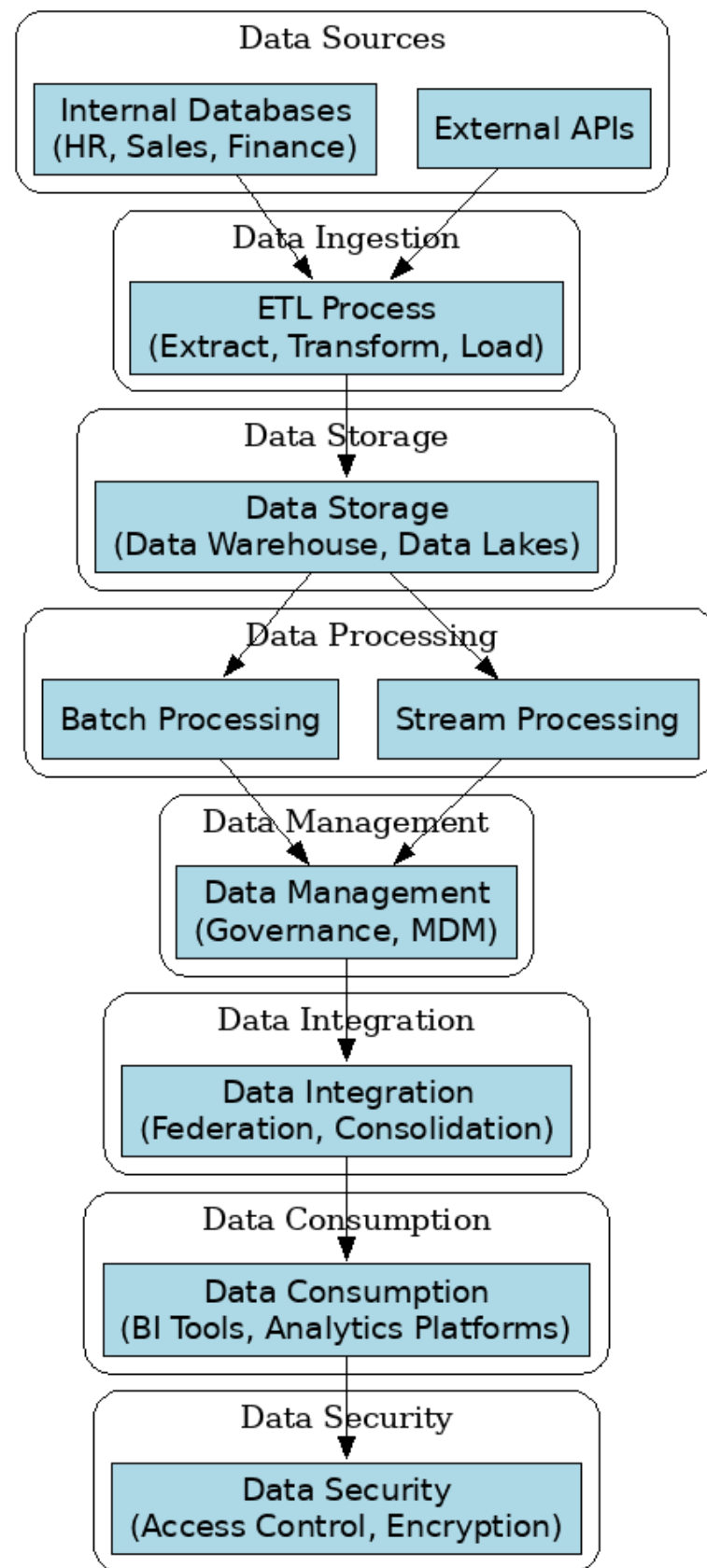
**¿Qué entiendes por "arquitectura de datos"?
¿Dónde crees que se aplica dentro de una organización?**



¿Qué tipos de datos crees que manejan las empresas actualmente y cómo los almacenan?



¿Has trabajado o interactuado con sistemas como un CRM, una base de datos o un dashboard de visualización? ¿Qué datos viste allí?



El Viaje de los Datos



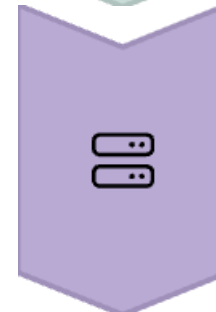
Fuentes de Datos

Todo parte aquí. Los datos pueden provenir de sistemas internos como Recursos Humanos, Finanzas o Ventas, así como de APIs externas.



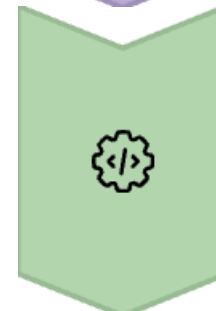
Ingesta de Datos

En esta etapa se realiza el famoso proceso ETL: Extraer, Transformar y Cargar los datos para su procesamiento eficiente.



Almacenamiento

Los datos se almacenan en sistemas como Data Warehouses o Data Lakes, dependiendo del tipo y volumen de información.



Procesamiento

Los datos se procesan ya sea en lotes (Batch) o en tiempo real (Stream), según la necesidad del negocio.

Continuación del Viaje de Datos



Gestión de Datos

Esta parte asegura que los datos estén bien organizados, actualizados y gobernados. Incluye prácticas como la gobernanza de datos y la gestión de datos maestros.



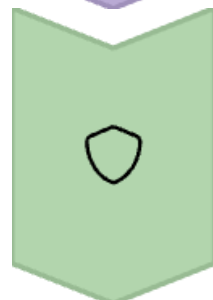
Integración de Datos

Los datos deben combinarse desde distintas fuentes para construir una visión unificada y útil mediante técnicas como federación o consolidación.



Consumo de Datos

Los usuarios finales acceden a los datos a través de herramientas como BI o plataformas analíticas para generar reportes y tomar decisiones.



Seguridad de Datos

Todo el sistema debe estar protegido con políticas de control de acceso y encriptación para mantener los datos seguros.

DATA JOURNEY

ON CONSUMPTION S



¿Qué es un Arquitecto de Datos?

Definición

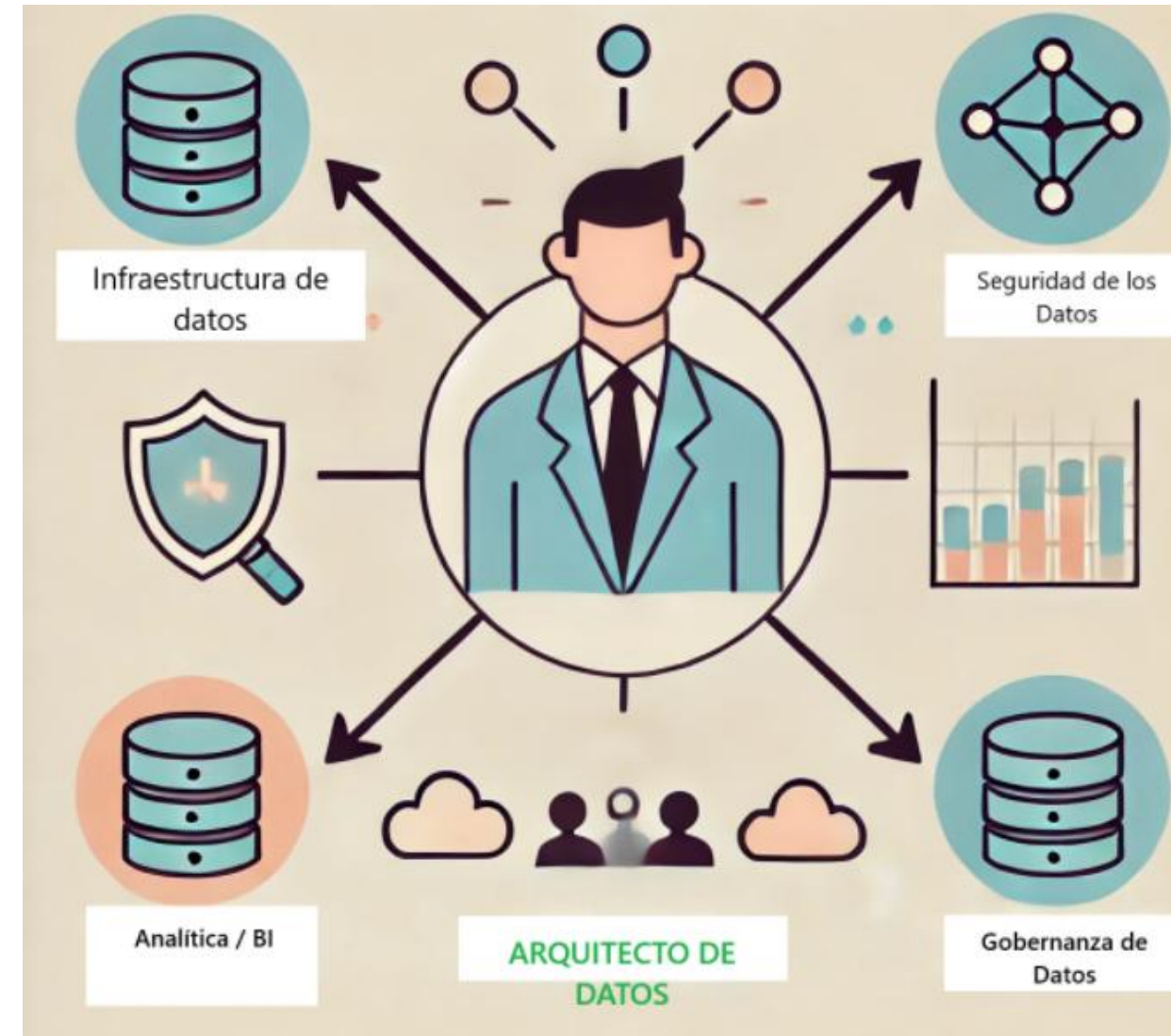
Un arquitecto de datos es como el urbanista de la información: diseña la ciudad (los sistemas de datos), define las rutas (los flujos de información), y garantiza que todo esté bien conectado, con buena seguridad y normas claras.

Rol Principal

Es el profesional encargado de diseñar, mantener y supervisar la arquitectura de datos en una organización, conectando el mundo técnico con el mundo del negocio.

Importancia

Sin un arquitecto de datos, las herramientas se implementan sin cohesión, los equipos duplican esfuerzos, los datos se fragmentan, y el análisis se vuelve lento e impreciso.





Funciones del Arquitecto de Datos

Función	Descripción práctica	Ejemplo
Visión Estratégica	Define cómo los datos pueden apoyar la toma de decisiones y la estrategia de negocio.	"Los datos de clientes deberían integrarse con los datos de ventas para predecir abandono."
Estándares Técnicos	Establece reglas para calidad, integración, interoperabilidad y seguridad.	"Usaremos formato JSON estandarizado para las APIs internas."
Gobernanza de Datos	Se asegura de que existan políticas claras sobre uso, propiedad y acceso a los datos.	"Cada departamento debe tener un Data Steward responsable de sus datos."
Escalabilidad	Garantiza que la arquitectura soporte el crecimiento de datos y usuarios.	"Esta plataforma en la nube puede crecer sin afectar el rendimiento."

Arquitecto vs. Ingeniero de Datos

Arquitecto de Datos

- Define el marco estratégico y tecnológico
- Diseña la arquitectura
- Se enfoca en estándares, interoperabilidad y gobernanza

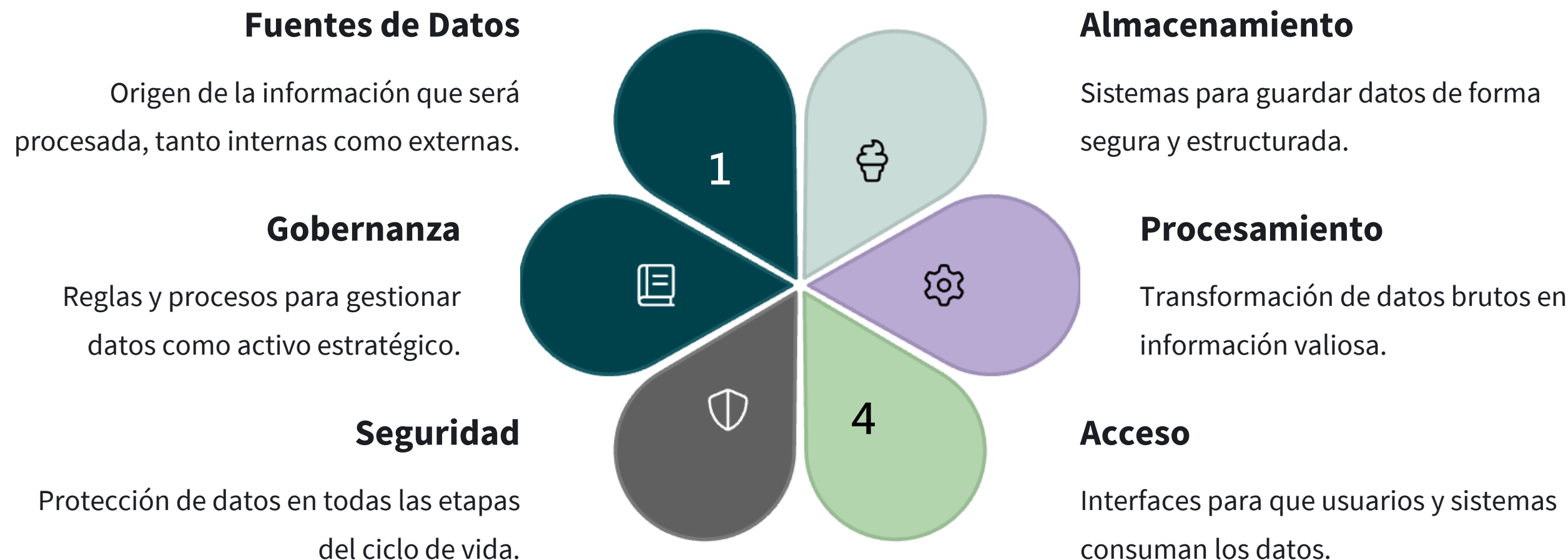
Piensa en el arquitecto como el diseñador del plano, quien establece la visión general y los estándares a seguir en la gestión de datos.

Ingeniero de Datos

- Implementa y opera los flujos de datos
- Construye pipelines y modelos
- Se enfoca en eficiencia, rendimiento y automatización

El ingeniero es quien convierte el plano en realidad, implementando las soluciones técnicas necesarias para que los datos fluyan correctamente.

Componentes de la Arquitectura de Datos





Fuentes de Datos Internas



Sistemas ERP

Gestionan procesos empresariales integrados como finanzas y logística. Ejemplos: SAP, Oracle ERP, Microsoft Dynamics, NetSuite.



Sistemas CRM

Administran interacciones con clientes. Ejemplos: Salesforce, HubSpot, Zoho CRM.



Bases de datos transaccionales

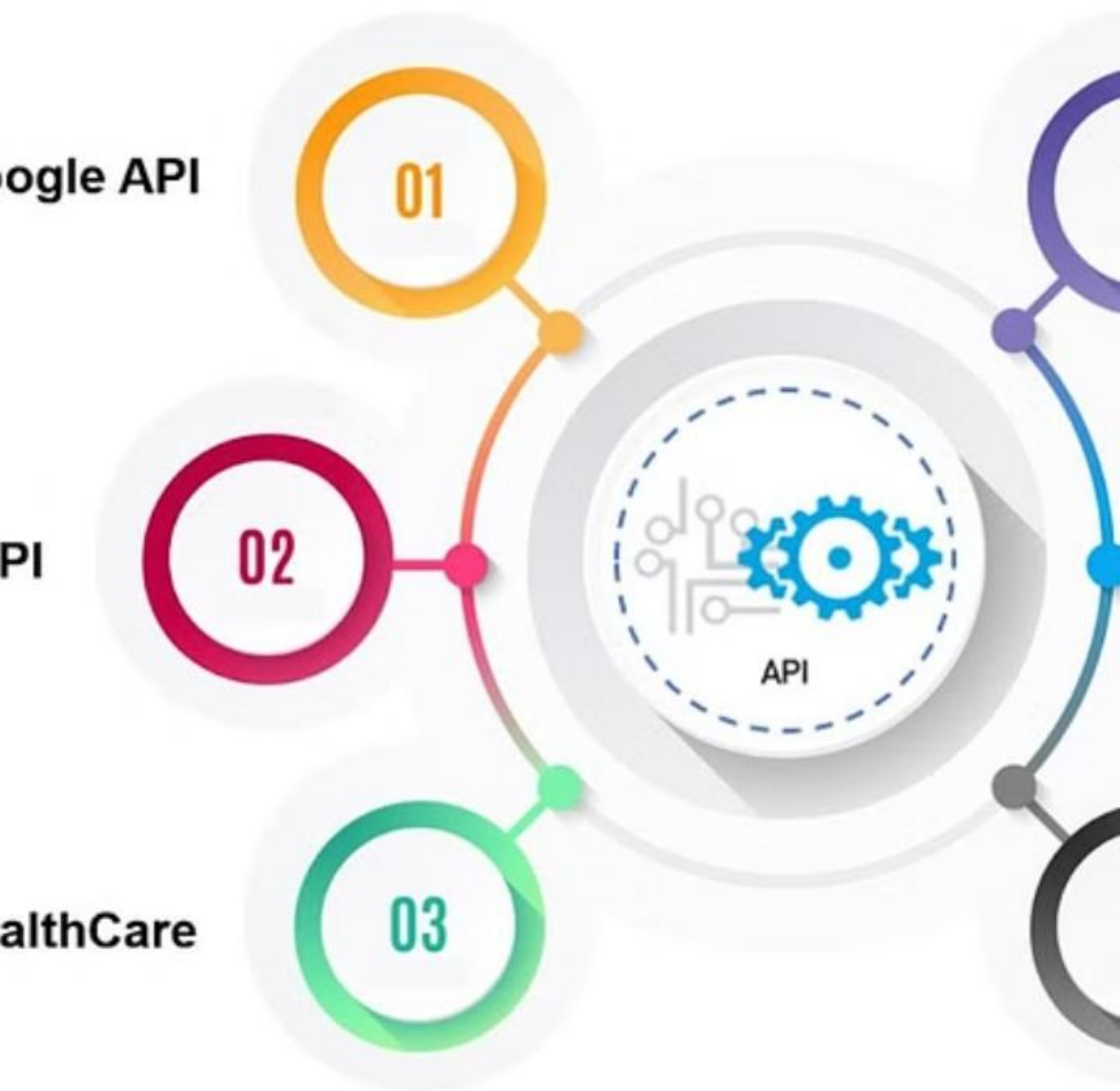
Almacenan y procesan operaciones en tiempo real. Ejemplos: Oracle Database, Microsoft SQL Server, PostgreSQL.



Sensores IoT

Recopilan datos de dispositivos conectados como sensores industriales, wearables y sensores ambientales.

API INTEGRATI



Fuentes de Datos Externas



APIs de servicios públicos

Obtienen datos estandarizados de entidades externas como APIs de bancos, sistemas de transporte y servicios de pago (Stripe, PayPal).



Portales de datos abiertos

Acceden a datasets gubernamentales o institucionales como data.gov (EE.UU.), datos.gob.es (España) y World Bank Open Data.



Redes sociales

Recopilan interacciones y tendencias de usuarios en plataformas como Twitter (X), Facebook, LinkedIn e Instagram.



Proveedores de datos especializados

Suministran información sectorial como Bloomberg (datos financieros), AccuWeather (datos climáticos) y NielsenIQ (datos de mercados).

Almacenamiento de Datos

Relacional (SQL)

Ideal para datos estructurados y relaciones bien definidas. Muy útil para aplicaciones OLTP (procesamiento de transacciones en línea). Ejemplos: PostgreSQL, MySQL, Oracle.

No Relacional (NoSQL)

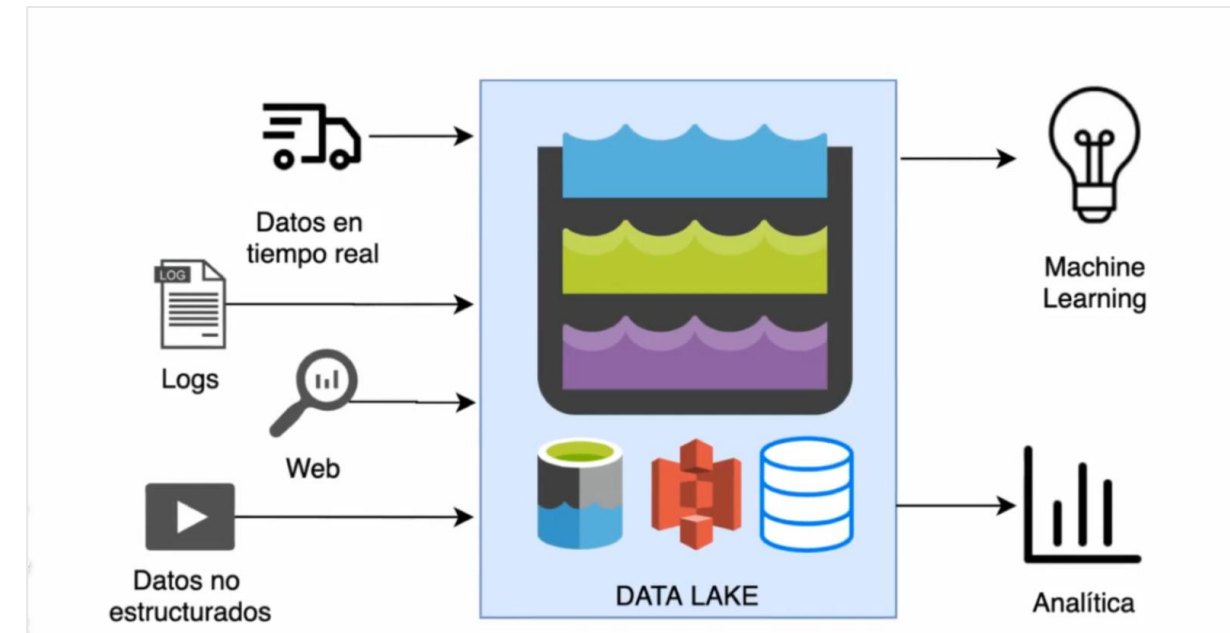
Adecuado para datos semiestructurados o que cambian rápidamente. Se adapta bien a aplicaciones web, logs o análisis en tiempo real. Ejemplos: MongoDB, Cassandra, Redis.

Data Lake

Espacio centralizado en la nube que permite guardar grandes volúmenes de datos en su forma original, sin necesidad de transformarlos previamente. Ideal para análisis exploratorios.

Data Warehouse

Sistemas diseñados específicamente para análisis de negocio (OLAP). Permiten consultas rápidas sobre grandes volúmenes de datos consolidados. Ejemplos: Snowflake, BigQuery.



Procesamiento de Datos



ETL (Extract, Transform, Load)

Se extraen los datos de las fuentes, se transforman y se cargan a un destino como un Data Warehouse.



ELT (Extract, Load, Transform)

Primero se cargan los datos al destino y luego se transforman, aprovechando el poder computacional del sistema de destino.



Frameworks y herramientas

Apache Spark, Hadoop, dbt y Airflow para procesamiento distribuido, batch, transformaciones SQL y orquestación.

Procesamiento Batch vs Tiempo Real

Procesamiento Batch

Procesa grandes volúmenes de datos en bloques programados, ideal para análisis históricos y reportes periódicos.

- Eficiente para grandes volúmenes
- Menor costo computacional
- Ideal para reportes diarios/semanales
- Ejemplo: Apache Hadoop

Procesamiento en Tiempo Real

Procesa datos continuamente a medida que llegan, permitiendo análisis y respuestas inmediatas.

- Respuesta inmediata
- Mayor complejidad técnica
- Ideal para alertas y monitoreo
- Ejemplo: Apache Kafka + Spark Streaming

Acceso a Datos



Dashboards y Reportes

Herramientas como Power BI, Tableau o Looker permiten visualizar indicadores clave, gráficos y métricas de forma interactiva, facilitando la comprensión de los datos para la toma de decisiones.

Opening the CTE Name of the CTE

Parenthesis CTE as first step of query Parenthesis

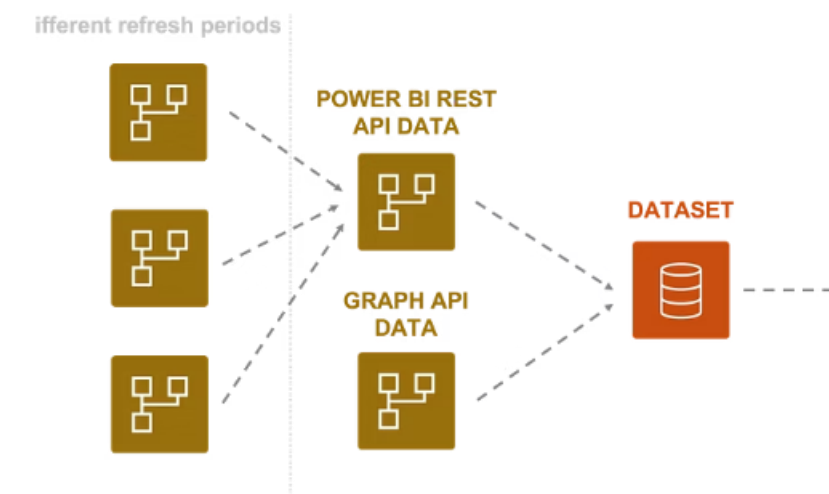
```
with value_per_order as
(
    select
        order_id
        ,sum(quantity * price) total_revenue_per_order
    from {{raw.e_commerce_sample.webshop_order_line}}
    group by order_id
)

select
    sum(total_revenue_per_order) as total_revenue
    ,avg(total_revenue_per_order) as average_revenue_per_order
    ,min(total_revenue_per_order) as smallest_order
    ,max(total_revenue_per_order) as largest_order
from value_per_order
```

Usual select statement CTE as location of table

Consultas Directas

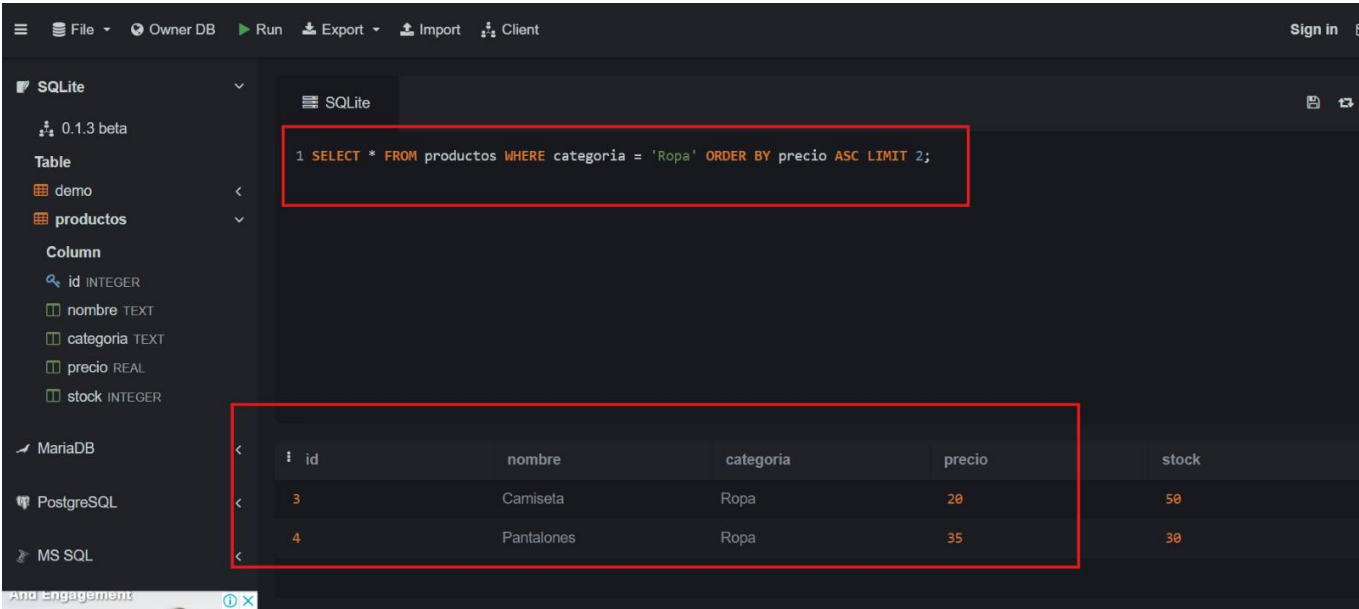
Los analistas pueden acceder directamente a los datos a través de SQL u otros lenguajes, dependiendo del sistema de almacenamiento, permitiendo análisis personalizados y profundos.



APIs de Datos

Permiten que aplicaciones o servicios externos consuman datos automáticamente, fomentando la integración entre plataformas y sistemas para un ecosistema de datos conectado.

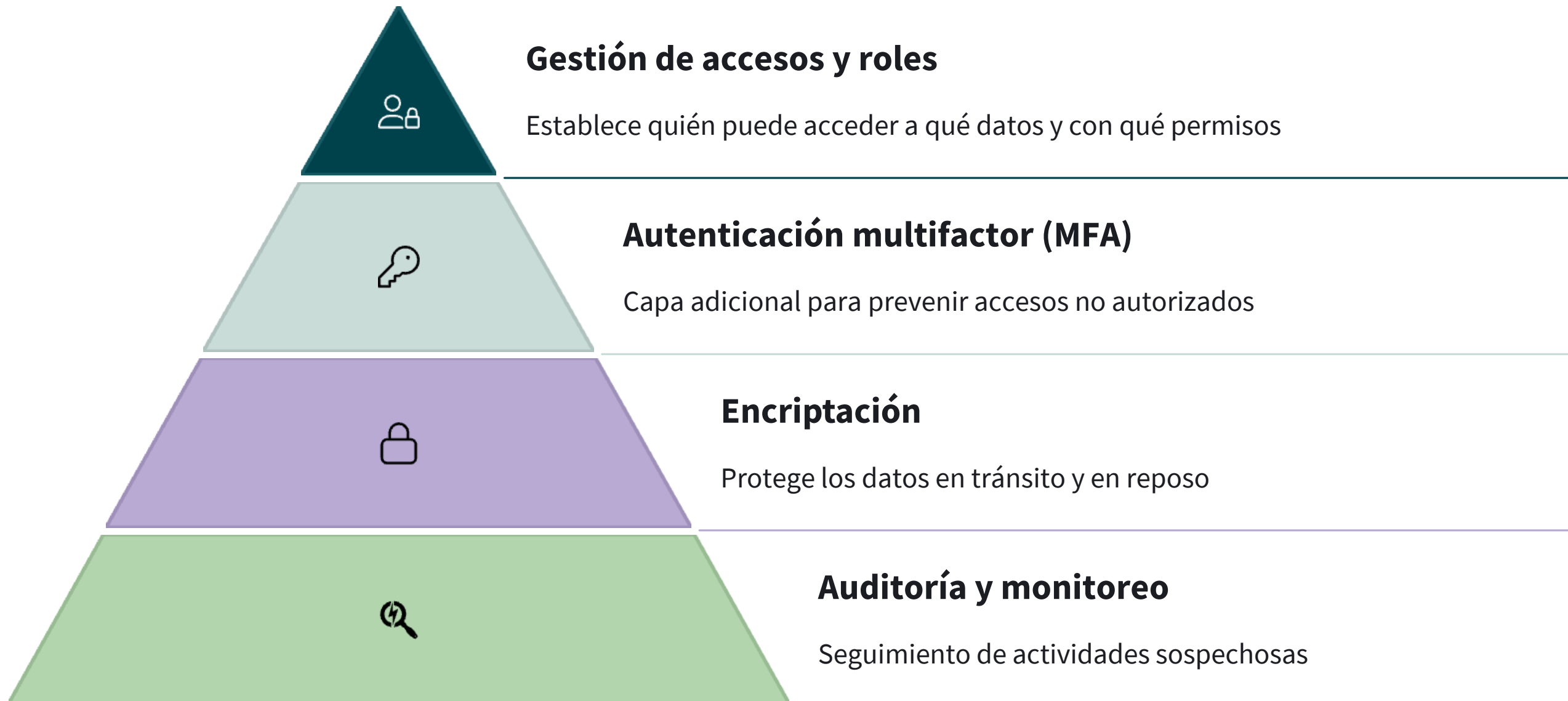
Ejemplos Acceso a Datos



Consulta SQL en SQLite extrayendo datos filtrados y ordenados de una tabla 'productos'

"Dashboard en Power BI con análisis de compras: \$4.13M en ventas, 499 transacciones, distribución por cliente/proveedor y rendimiento trimestral con filtros interactivos."

Seguridad de Datos



Gobernanza de Datos



Catálogo de datos

Repositorio centralizado que permite conocer qué datos existen, dónde están y cómo se utilizan, facilitando su descubrimiento y uso adecuado.



Diccionario de datos

Define los términos clave, estructuras y formatos para evitar ambigüedades y asegurar una comprensión común en toda la organización.



Políticas de calidad

Aseguran que los datos sean precisos, completos, actualizados y consistentes, estableciendo estándares y procesos de validación.



Gestión de roles

Define quiénes son los data stewards, custodios y propietarios del dato, y cuáles son sus responsabilidades en el ecosistema de datos.



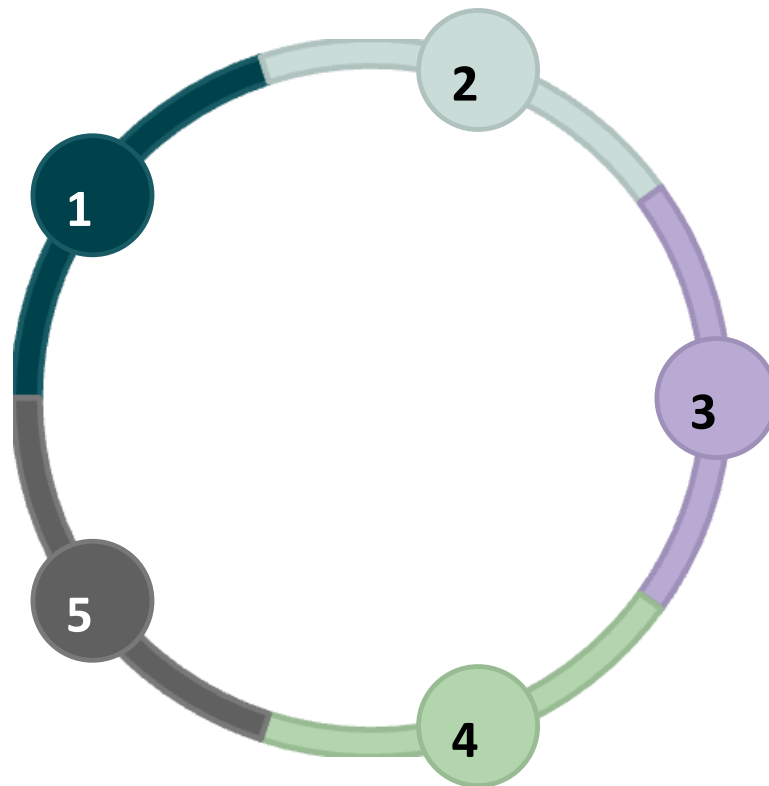
Marco de Referencia DAMA-DMBOK

Arquitectura de datos

Define cómo se estructuran y conectan los sistemas de datos dentro de una organización.

Modelado y diseño

Crea representaciones conceptuales, lógicas y físicas de los datos y sus relaciones.



Gobernanza de datos

Establece políticas, estándares y roles claros para la toma de decisiones relacionadas con los datos.

Calidad de datos

Asegura que los datos sean confiables, completos, exactos y relevantes para el negocio.

Seguridad de datos

Protege los datos frente a accesos no autorizados, pérdida o uso indebido.

Principios de Arquitectura de Datos



Centralización o federación

Control de datos según el tamaño y necesidades de la organización, equilibrando la gestión centralizada con la autonomía de las unidades de negocio.



Documentación clara

Información accesible sobre flujos y estructuras de datos, facilitando la comprensión y mantenimiento de los sistemas por todos los equipos.



Escalabilidad

Capacidad para crecer sin comprometer el rendimiento, adaptándose al aumento en volumen de datos y usuarios del sistema.



Flexibilidad

Adaptabilidad a nuevas tecnologías o requisitos del negocio, permitiendo evolucionar sin rediseños completos de la arquitectura.



Consideraciones Clave en Arquitectura

Escalabilidad

¿Soporta el crecimiento futuro en volumen y variedad de datos? La arquitectura debe poder expandirse sin rediseños completos, adaptándose al crecimiento de la organización.

Flexibilidad

¿Permite cambios sin rediseñar toda la solución? Los componentes deben poder actualizarse o reemplazarse individualmente sin afectar al sistema completo.

Interoperabilidad

¿Se integra con otras herramientas o plataformas? Los sistemas deben comunicarse eficientemente mediante estándares y protocolos comunes.

Seguridad y privacidad

¿Protege los datos sensibles y cumple con regulaciones? La protección debe ser inherente al diseño, no una capa adicional.

KEY CONSIDERATIONS IN DATA ARCHITECTURE DESIGN

SCALABILITY



FLEXIBILITY



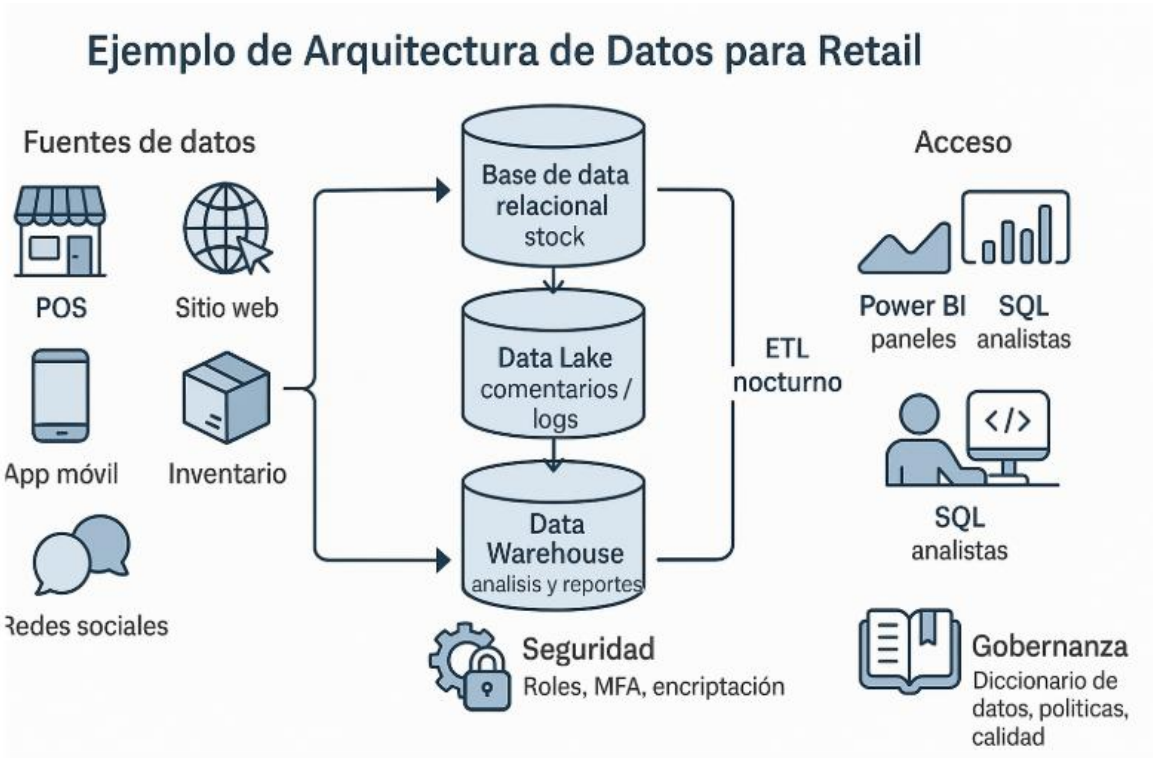
INTEROPERABILITY



SECURITY



Ejemplo: Arquitectura de Datos en Retail



Fuentes de datos

Punto de venta (POS), sitio web, aplicación móvil, inventario, redes sociales. Capturan transacciones, comportamiento de clientes e inventario en tiempo real.

2

Almacenamiento

Base de datos relacional para stock e inventario, Data Lake en la nube para datos semiestructurados, Data Warehouse para reportes gerenciales.



Procesamiento

Proceso ETL nocturno que integra datos de ventas, inventario y logística, permitiendo análisis completos del negocio.

4

Acceso

Paneles BI en Power BI para el área comercial y consultas SQL para los analistas, facilitando decisiones basadas en datos.

Seguridad y Gobernanza en el Ejemplo Retail

Seguridad

- Accesos controlados por rol (cajeros, gerentes, analistas)
- Autenticación multifactor para sistemas críticos
- Encriptación de datos sensibles de clientes
- Monitoreo continuo de accesos a bases de datos
- Cumplimiento con normativas de protección de datos

Gobernanza

- Diccionario de datos unificado para toda la empresa
- Políticas de acceso documentadas y auditadas
- Revisión semanal de calidad de datos de ventas e inventario
- Data stewards designados en cada departamento
- Procesos de resolución de discrepancias en datos

Ciclo de Vida del Dato

Captura

Definición de formatos y validaciones desde el origen para asegurar la calidad inicial de los datos.

Distribución

Acceso a usuarios y sistemas mediante capas de servicio como APIs y herramientas BI.



Almacenamiento

Elección del sistema adecuado según el tipo y uso del dato, optimizando acceso y seguridad.

Procesamiento

Aplicación de reglas de negocio, integración y transformación para generar valor.

Continuación del Ciclo de Vida

Consumo

Uso en análisis, reportes y modelos de IA para generar insights y valor para el negocio.

Eliminación

Borrado seguro cuando los datos ya no son necesarios, cumpliendo normativas.



Mantenimiento

Revisión, depuración y actualización continua para mantener la relevancia y calidad.

Archivado

Almacenamiento a largo plazo según políticas de retención y requisitos legales.

Representación Visual del Ciclo de Vida

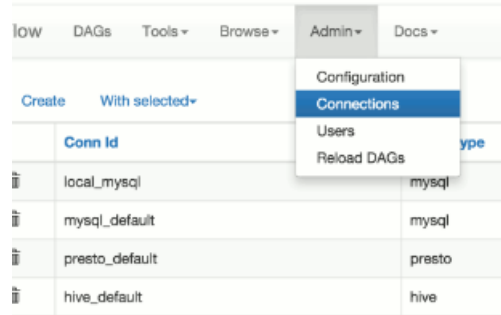


Tecnologías de Almacenamiento



Las tecnologías de almacenamiento de datos han evolucionado significativamente, ofreciendo soluciones especializadas para diferentes necesidades. Desde bases de datos relacionales como PostgreSQL hasta soluciones NoSQL como MongoDB, pasando por data lakes como Amazon S3 y data warehouses como Snowflake, BigQuery y Azure Synapse, cada tecnología tiene sus fortalezas específicas para distintos casos de uso.

Herramientas de Procesamiento y Análisis

A screenshot of the Apache Airflow Admin UI. The top navigation bar includes links for "Home", "DAGs", "Tools", "Browse", "Admin", and "Docs". The "Admin" dropdown menu is open, showing options: "Configuration", "Connections", "Users", and "Reload DAGs". Below the menu is a table of connections.

Conn Id	Type
local_mysql	mysql
mysql_default	mysql
presto_default	presto
hive_default	hive



El ecosistema de herramientas para procesamiento y análisis de datos es amplio y diverso. Para el procesamiento, Apache Spark permite computación distribuida, mientras que Airflow orquesta flujos de trabajo y dbt gestiona transformaciones SQL. Para visualización y análisis, Power BI, Tableau y Looker ofrecen potentes capacidades de business intelligence que transforman datos en insights accionables.

Enlaces de Interés

- Qué es la arquitectura de datos: <https://www.ibm.com/think/topics/data-architecture>
- **MARCO DE REFERENCIA DAMA-DMBOK:** [DAMA-DMBOK - Data Management Body of Knowledge](#)
- Qué es una API: [Que es una API](#)
- Apache Spark: [Apache Spark Quick Start](#)
- Apache Airflow: [Apache Airflow Documentation](#)
- Power BI: [Power BI Docs](#)
- Google Looker: [Google Looker Studio](#)
- **Video:** [ORACLE PL/SQL: Tipos de ROLES](#)
- **Video:** [Qué es MFA \(Autenticación Multifactor\) y cómo se usa](#)
- **Video:** [Bases de Datos Relacionales](#)
- **Video:** [¿Cómo funcionan las bases de datos NoSQL? ¡Explicado simplemente! \(ingles sub es\)](#)
- **Video:** [¿Qué es un Data Warehouse? - Álvaro Montero](#)
- **Video:** [¿Qué son los procesos ETL? Curso ETL con Pentaho Gratis \(Descripción\)](#)
- **Video:** [Data Security: Protect your critical data \(or else\)](#)
- **Video:** [AUDITORIA DE BASE DE DATOS](#)
- **Video:** [Gobernanza de datos explicada en 5 minutos](#)



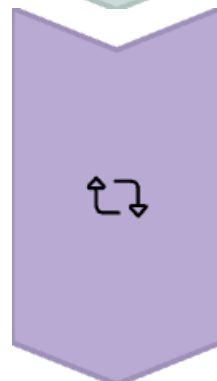
Preguntas de Reflexión Final



¿Cuál consideras que es el componente más crítico dentro de una arquitectura de datos moderna y por qué?



¿De qué forma la arquitectura de datos puede ayudar a que una empresa tome mejores decisiones?



¿Qué aspectos de la arquitectura de datos crees que deben actualizarse o adaptarse más frecuentemente en una organización en crecimiento?