Modulo 5

Proyecto: Portafolio ABP

Nombre: Marco Neira

Lección 1 – Arquitectura de datos

Objetivo

Diseñar un esquema arquitectónico robusto para la integración, almacenamiento y consumo de datos.

1. **Identificación de fuentes de datos**

Seleccionamos 3 fuentes estructuradas y 2 no estructuradas típicas en e-commerce:

Fuente de Datos	Tipo	Descripción
Base de datos de ventas (ERP)	Estructurada	Registra transacciones, productos, precios, fechas y clientes.
CRM (Customer Relationship Mgt)	Estructurada	Historial de interacciones, tickets de soporte, perfiles de clientes.
Inventario y logística	Estructurada	Información sobre stock, ubicación de productos y despachos.
Redes sociales (Twitter, IG)	No estructurada	Comentarios, menciones, hashtags relacionados con la marca.
Comportamiento en la web (logs)	No estructurada	Archivos de logs de navegación, clics, duración de sesiones, etc.

2. **Diseño de arquitectura escalable y modular (capas)**

Capa	Función
Ingesta	Extracción desde fuentes (APIs, bases de datos, FTP, etc.).
Integración	Limpieza, transformación y estandarización con ETL/ELT.
Almacenamiento	Organización por zonas: Raw, Trusted, Curated, DWH, Data Marts.
Calidad	Validación, control de duplicados, formatos y registros faltantes.
Consumo	Acceso por herramientas BI (Power BI, Tableau) y consultas analíticas.

3. **Principios incorporados**

Gobierno de datos: catalogación, políticas de acceso, trazabilidad.

Escalabilidad: uso de servicios en la nube (como AWS, Azure o GCP).

Flexibilidad: soporta tanto batch como streaming, estructurados y no estructurados.

4. **Diagrama arquitectónico**

Ingesta:

- API redes sociales → Landing zone

- Base de datos ventas / CRM → Conectores ETL
- Logs web → Servicio de streaming (ej. Kafka)

Integración:

- Transformación con Spark / Dataflow
- Enriquecimiento de datos (ej. agregación mensual)

Almacenamiento:

- Data Lake (Raw → Trusted → Curated)
- DWH en Redshift / BigQuery / Synapse
- Data Marts para áreas de negocio

Gobierno y calidad:

- Glue Data Catalog / Azure Purview
- Validación con reglas de calidad

Consumo:

- Power BI para dashboards de ventas y marketing
- Modelos ML en notebooks o servicios de predicción

Lección 2 – Enfoques para almacenamiento y gestión

Objetivo

Definir y justificar estrategias de almacenamiento y gobierno alineadas con la arquitectura diseñada.

1. Análisis del esquema arquitectónico de la Lección

El diseño propuesto se basa en una arquitectura **modular y escalable por capas**, que permite manejar fuentes estructuradas y no estructuradas, aplicar transformaciones y asegurar calidad, con el objetivo de ofrecer datos listos para análisis avanzado.

El esquema arquitectónico cumple con los principios de una arquitectura moderna para analítica empresarial: **es modular, escalable, gobernable y adaptable**, sentando las bases para decisiones informadas basadas en datos confiables.

2. Zonas del Data Lake y relación con DWH / Data Marts

Zona del Data Lake	Función principal
Raw	Almacenamiento en crudo. Datos tal como llegan de las fuentes.
Trusted	Datos validados y filtrados. Se eliminan errores obvios y se aplica schema.
Curated	Datos transformados, integrados y listos para análisis o carga en DWH.

Relación:

- El **Data Warehouse** se alimenta desde **Curated**, donde los datos ya están depurados y estructurados.
- Los **Data Marts** se derivan desde el DWH, personalizados según cada área de negocio (ventas, logística, etc.).

3. Tecnologías y servicios recomendado

Zona/Componente	Tecnología / Servicio sugerido
Raw	Amazon S3 / Azure Data Lake / Google Cloud Storage
Trusted	Glue Jobs / Dataflow / Synapse Pipelines
Curated	Apache Spark / Databricks / BigQuery
Data Warehouse	Amazon Redshift / Azure Synapse / Google BigQuery
Data Marts	Tablas particionadas en el DWH + vistas materializadas
Metadata catalog	AWS Glue Data Catalog / Azure Purview / GCP Data Catalog

4.Gobernanza y gestión de datos

Área	Práctica
Trazabilidad	Registro del linaje de los datos (de dónde vienen, cómo cambian, etc.).
Seguridad	Autenticación y autorización por roles. Cifrado en reposo y en tránsito.
Disponibilidad	Replicación geográfica, backups automáticos y arquitectura serverless.
Calidad	Controles automáticos por etapa. Alertas y dashboards de monitoreo.
Data Catalog	Definición clara de fuentes, campos, formatos y responsables.

Lección 3 – Calidad de los datos

Objetivo

Diseñar un plan de aseguramiento de calidad de datos integrado a la arquitectura general definida en lecciones anteriores.

1. **Revisión del flujo arquitectónico**

El flujo de datos parte desde múltiples fuentes (ventas, CRM, redes sociales, logs web), ingresa por conectores ETL o APIs, pasa por una capa de integración para limpieza y transformación, y luego entra al Data Lake segmentado en zonas (Raw → Trusted → Curated). Desde ahí, se alimenta el Data Warehouse y los Data Marts.

2. **Controles, métricas e indicadores de calidad**

Dimensión de Calidad	Métrica / Indicador
Completitud	% de registros sin valores nulos en campos críticos.
Validez	% de registros que cumplen con los formatos esperados.
Consistencia	Coincidencia entre fuentes cruzadas (ej. stock vs ventas).
Unicidad	Cantidad de registros duplicados por ID.
Trazabilidad	Disponibilidad de metadatos de origen y transformación.
Actualización	Frecuencia de actualización vs. SLA esperado.

3. **Diseño de proceso de monitoreo y remediación**

Proceso continuo:

1. **Captura automática de métricas** (por lotes o streaming).
2. **Dashboards de monitoreo** (ej. con Superset, Power BI o Grafana).
3. **Alertas automáticas** ante anomalías.
4. **Remediación**:
    - Corrección automática de errores simples (estandarización, imputación).
    - Escalamiento a Data Steward para casos complejos.
5. **Registro de acciones** (log de incidentes de calidad y resolución).

4. **Integración del plan de calidad en la arquitectura**

La calidad se integra como una **capa transversal**:

- Validaciones automáticas desde la **capa de integración**.
- Control y remediación continua en la zona **Trusted → Curated**.
- Data Catalog como herramienta clave para trazabilidad y documentación.
- Incorporación de métricas de calidad en cada flujo ETL.
- Paneles de monitoreo de calidad en la capa de consumo.

Lección 4 – Modelamiento multidimensional

Objetivo

Diseñar un modelo OLAP coherente con la arquitectura de datos y las estrategias de calidad implementadas.

1. Selección del área clave de negocio

Proponemos trabajar con el área de **Ventas**, ya que es:

- Crítica para el negocio e-commerce,
- Conectada a múltiples fuentes (ERP, inventario, campañas),
- Altamente analizable en términos de volumen, frecuencia y rentabilidad.

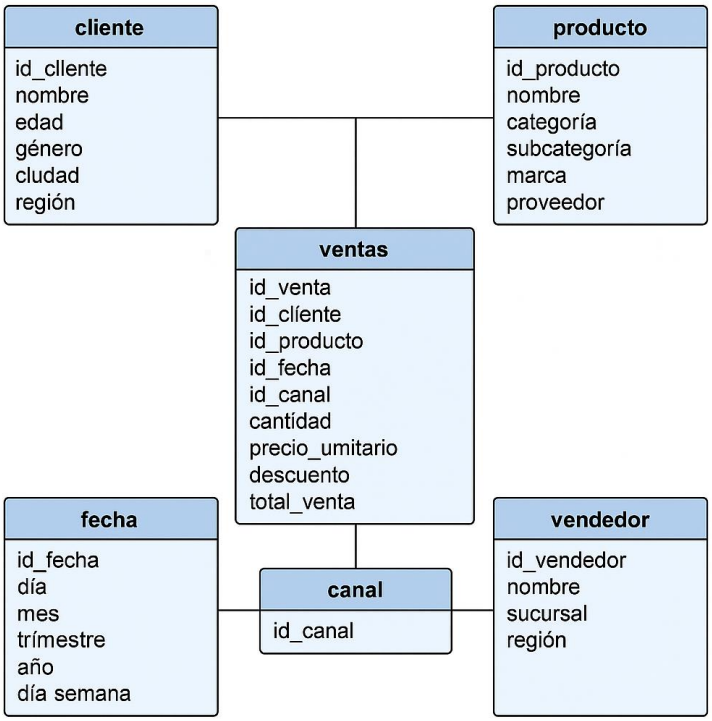
2. Diseño del modelo dimensional

Esquema tipo estrella ("Star Schema")

Es el más directo y fácil de interpretar por herramientas de BI. Consiste en una tabla de hechos central y múltiples dimensiones relacionadas.

Campo	Descripción
id_venta	Identificador único de la venta
id_cliente	FK hacia la dimensión cliente
id_producto	FK hacia la dimensión producto
id_fecha	FK hacia la dimensión fecha
id_canal	FK hacia la dimensión canal
id_vendedor	FK hacia la dimensión vendedor
cantidad	Unidades vendidas
precio_unitario	Precio por unidad
descuento	Descuento aplicado
total_venta	Monto final (calculado)

3. **Diagrama del modelo dimensional**



4. **Decisiones de modelado y criterios analíticos**

Se eligió el modelo **estrella** por su simplicidad y rendimiento en consultas agregadas.

Desnormalización aplicada en dimensiones (como producto y cliente) para facilitar el análisis.

Las **jerarquías temporales** (día → mes → trimestre → año) permitirán análisis por período.

Se evita modelar datos sensibles (ej. correo del cliente) en el modelo OLAP por privacidad.

Las medidas podrán agregarse fácilmente: sumas (total_venta), promedios (precio_unitario), conteo (id_venta), etc.