Lección 1: Arquitectura de Datos para E-commerce y Retail Digital

1. Identificación de Fuentes de Datos

Fuentes Estructuradas:

1. **Base de datos transaccional de ventas online**
    - Tipo: SQL (PostgreSQL/MySQL)
    - Datos: Órdenes de compra, detalles de productos, precios, métodos de pago, direcciones de envío
    - Volumen: ~10,000 transacciones diarias
    - Frecuencia de actualización: Tiempo real
2. **Sistema de inventario**
    - Tipo: SQL (SQL Server)
    - Datos: Stock de productos, ubicaciones, SKUs, proveedores, tiempos de reposición
    - Volumen: ~100,000 productos
    - Frecuencia de actualización: Diaria
3. **CRM**
    - Tipo: API/SQL
    - Datos: Perfiles de clientes, historiales de compra, segmentos, preferencias
    - Volumen: ~1 millón de clientes
    - Frecuencia de actualización: Diaria

Fuentes No Estructuradas:

1. **Registros de navegación web (clickstream)**
    - Tipo: Logs JSON/CSV
    - Datos: Rutas de navegación, tiempo en página, búsquedas, abandonos de carrito
    - Volumen: ~500GB diarios
    - Frecuencia de actualización: Tiempo real
2. **Redes sociales**
    - Tipo: APIs (Twitter, Facebook, Instagram)
    - Datos: Menciones, comentarios, sentimientos, engagement
    - Volumen: Variable (~10GB diarios)
    - Frecuencia de actualización: Casi tiempo real
3. **Comentarios y reseñas de productos**
    - Tipo: Texto no estructurado
    - Datos: Opiniones, valoraciones, fotos de usuarios
    - Volumen: ~5,000 reseñas diarias
    - Frecuencia de actualización: Tiempo real

2. Diseño Arquitectónico Basado en Capas

Capa de Ingesta

Responsable de capturar datos de las diversas fuentes y transportarlos a la capa de almacenamiento:

- **Ingesta batch**: Para cargas periódicas desde sistemas transaccionales
- **Ingesta en tiempo real**: Para capturar datos de clickstream, redes sociales y eventos en tiempo real

Capa de Almacenamiento

Organizada en tres zonas principales:

- **Zona Raw (Data Lake)**: Almacenamiento de datos en su forma original
- **Zona Trusted**: Datos validados, limpios y transformados
- **Zona Curated**: Datos agregados y preparados para análisis específicos

Capa de Procesamiento

Responsable de transformar, limpiar y enriquecer los datos:

- **Procesamiento batch**: Para transformaciones complejas y cargas pesadas
- **Procesamiento streaming**: Para análisis en tiempo real y detección de eventos

Capa de Gobierno y Calidad

Asegura que los datos cumplan con los estándares de calidad y seguridad:

- **Gobierno de datos**: Políticas, metadatos, linaje
- **Seguridad**: Acceso, encriptación, anonimización
- **Calidad**: Validación, monitoreo, alertas

Capa de Consumo

Proporciona acceso a los datos para diversos casos de uso:

- **Data Warehouse**: Para análisis estructurado
- **Data Marts**: Para necesidades específicas de departamentos
- **APIs de datos**: Para integración con aplicaciones
- **Herramientas de visualización**: Para dashboards y reportes

3. Principios de Gobierno, Escalabilidad y Flexibilidad

Principios de Gobierno

- **Propiedad de datos**: Asignación clara de responsabilidades sobre conjuntos de datos
- **Catalogación**: Registro centralizado de metadatos, definiciones y clasificaciones
- **Linaje**: Trazabilidad completa del origen y transformaciones de los datos
- **Políticas de retención**: Gestión del ciclo de vida de los datos
- **Cumplimiento normativo**: GDPR, PCI-DSS, normativas locales

Principios de Escalabilidad

- **Arquitectura distribuida**: Capacidad para escalar horizontalmente
- **Separación computación/almacenamiento**: Para escalar independientemente
- **Procesamiento elástico**: Ajuste automático según la carga
- **Particionamiento**: División lógica de datos para mejor rendimiento

Principios de Flexibilidad

- **Arquitectura modular**: Componentes independientes y sustituibles
- **Tecnologías cloud-native**: Aprovechamiento de servicios gestionados
- **Schema-on-read**: Flexibilidad para adaptarse a diferentes estructuras
- **Políglotas**: Soporte para diferentes tipos de almacenamiento según necesidades

4. Tecnologías Recomendadas (Cloud-Agnostic)

Ingesta

- **Batch**: Apache NiFi, Apache Airflow
- **Streaming**: Apache Kafka, Amazon Kinesis

Almacenamiento

- **Data Lake**: Amazon S3, Azure Data Lake Storage, MinIO
- **Data Warehouse**: Snowflake, Amazon Redshift, Google BigQuery
- **Data Marts**: Microsoft Analysis Services, Druid

Procesamiento

- **Batch**: Apache Spark, Apache Hive
- **Streaming**: Apache Flink, Apache Spark Streaming

Gobierno y Calidad

- **Gobierno**: Collibra, Apache Atlas
- **Calidad**: Great Expectations, Deequ

Consumo

- **Visualización**: Tableau, Power BI, Looker
- **APIs**: GraphQL, REST APIs

5. Justificación de Decisiones de Diseño

1. **Separación en capas**: Permite mantener la independencia entre componentes, facilitando actualizaciones y sustituciones sin afectar todo el sistema.

2. **Enfoque Data Lake + Data Warehouse**: Combina la flexibilidad del Data Lake para almacenar datos heterogéneos con la eficiencia analítica del Data Warehouse para consultas estructuradas.

3. **Arquitectura Lambda modificada**: Proporciona capacidades de procesamiento tanto batch como streaming, permitiendo análisis históricos y en tiempo real.

4. **Enfoque cloud-agnostic**: Evita el vendor lock-in y permite una migración más sencilla entre proveedores de nube si fuera necesario.

5. **Énfasis en gobierno de datos**: Fundamental para una empresa de e-commerce que maneja datos sensibles de clientes y transacciones financieras.

6. Consideraciones Específicas para E-commerce

1. **Gestión de picos de tráfico**: La arquitectura debe soportar incrementos sustanciales durante temporadas de ofertas (Black Friday, Navidad).

2. **Personalización en tiempo real**: Capacidad para analizar comportamiento y ofrecer recomendaciones personalizadas mientras el usuario navega.

3. **Análisis de abandono de carrito**: Detección y análisis de patrones de abandono para mejorar conversiones.

4. **Integración omnicanal**: Capacidad para integrar datos de tiendas físicas, web, apps móviles y redes sociales en una visión unificada del cliente.

5. **Seguridad y cumplimiento**: Protección de datos sensibles de tarjetas de crédito (PCI-DSS) e información personal (GDPR).