**Situación inicial**

**Unidad solicitante: Área de Analítica de Datos de una empresa de e-commerce y retail digital.**

La empresa ha experimentado un crecimiento acelerado y actualmente gestiona múltiples fuentes de datos (ventas online, comportamiento de usuarios, redes sociales, logística, inventario). La Dirección General solicita al equipo de datos el diseño de una arquitectura moderna y escalable que permita integrar, almacenar, gobernar y preparar estos datos para el análisis, garantizando su calidad y facilitando la toma de decisiones.

El desafío consiste en construir una arquitectura robusta que combine distintos niveles de almacenamiento (Data Lake, Data Warehouse, Data Mart), implemente estrategias de control de calidad y contemple un modelo multidimensional para el consumo analítico por parte de distintas áreas del negocio.

**Nuestro objetivo**

El objetivo principal del proyecto es diseñar, documentar y justificar una **arquitectura de datos integral**, que contemple la integración de fuentes heterogéneas, el almacenamiento y gobierno de los datos, la aplicación de criterios de calidad y la estructuración de un modelo dimensional preparado para su análisis.

Este diseño debe ser escalable, seguro y alineado con las necesidades analíticas y operativas de la organización.

**Requerimientos**

La arquitectura de datos a diseñar deberá cumplir con los siguientes requerimientos generales para garantizar su funcionalidad, calidad y alineación con las necesidades del negocio:

**Requerimientos generales:**

→ Integración de múltiples fuentes de datos estructurados y no estructurados.

→ Aplicación de los principios y buenas prácticas de arquitectura de datos.

→ Incorporación de modelos de almacenamiento diferenciados: Data Lake, Data Warehouse, Data Mart.

→ Implementación de un plan de aseguramiento y monitoreo de la calidad de los datos.

→ Construcción de un modelo multidimensional alineado a un área clave del negocio.

**Lección 1 – Arquitectura de datos**

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

**Tareas a desarrollar:**

1. **Fuentes de datos**

Identificar al menos tres fuentes de datos estructuradas y no estructuradas relevantes para la organización.

* **Fuentes de Datos Estructurados:**
  + **Sistema de Gestión de Pedidos (OMS):** Contiene información detallada sobre pedidos, clientes, productos, estados de envío y transacciones de pago.
  + **Sistema de Gestión de Relaciones con el Cliente (CRM):** Almacena datos de clientes, historial de interacciones, campañas de marketing y retroalimentación.
  + **Sistema de Gestión de Inventario (WMS):** Proporciona información sobre el stock de productos, ubicaciones de almacén y movimientos de inventario.
* **Fuentes de Datos No Estructurados:**
  + **Datos de Clics y Comportamiento del Usuario:** Registros de la actividad del usuario en el sitio web (páginas visitadas, clics, tiempo de permanencia).
  + **Datos de Redes Sociales:** Menciones de marca, comentarios, reseñas de productos y métricas de interacción en plataformas como Twitter o Instagram.
  + **Archivos de Log de Servidores Web:** Registros de acceso y errores que pueden ser analizados para identificar patrones de tráfico y problemas de rendimiento.

1. **Diseñar y documentar una arquitectura de datos escalable y modular, basada en capas: ingesta, integración, almacenamiento, calidad y consumo.**

* **Capa de Ingesta:** Se encarga de la recolección de datos desde las fuentes de origen. Se pueden usar herramientas de **ETL (Extract, Transform, Load)** o **ELT (Extract, Load, Transform)** para datos estructurados y de streaming para datos en tiempo real.
* **Capa de Almacenamiento:** Es el corazón de la arquitectura. Se debe considerar el uso de un **Data Lake** para almacenar datos en su formato original y un **Data Warehouse** para datos procesados y estructurados (en este caso esta capa la dejé antes porque considerando que se guardan datos crudos esto va antes de la integración donde se transforman los datos).
* **Capa de Integración y Calidad:** Aquí se aplican las transformaciones, se limpian los datos y se realizan validaciones para asegurar su calidad antes de ser movidos a la capa de consumo.
* **Capa de Consumo:** En esta capa, los datos están listos para ser utilizados por los usuarios finales. Se puede crear un **Data Mart** para un área de negocio específica (ej. marketing) o utilizar herramientas de **Business Intelligence (BI)** y **machine learning**.

1. **Incorporar principios de gobierno, escalabilidad y flexibilidad.**

* **Gobierno de datos:** Define políticas y procedimientos para la gestión de datos, incluyendo la seguridad, la privacidad y la propiedad. Esto asegura la trazabilidad y la responsabilidad sobre los datos.
* **Escalabilidad y flexibilidad:** La arquitectura debe ser capaz de crecer a medida que el volumen de datos aumenta. El uso de servicios en la nube (cloud computing) permite escalar recursos de forma dinámica y pagar solo por lo que se usa. Un enfoque de desacoplamiento de componentes permite la flexibilidad para cambiar tecnologías sin afectar todo el sistema. Ambos temas van muy de la mano, ya que sin flexibilidad es difícil considerar una escalabilidad.

1. **Elaborar un diagrama arquitectónico que represente las fuentes, las capas y el flujo de datos.**

Diagrama

El contenido generado por IA puede ser incorrecto.

**Lección 2: Enfoques para Almacenamiento y Gestión**

**Objetivo:** Definir y justificar las estrategias de almacenamiento y gobernanza de los datos, alineadas al diseño arquitectónico.

**1. Zonas de almacenamiento**

La arquitectura debe contemplar diferentes zonas de almacenamiento dentro del **Data Lake**:

* **Zona datos crudos:** Los datos se ingieren y se almacenan en su formato original, sin ninguna modificación. Es un repositorio de "datos originales" para su posterior procesamiento.
* **Zona datos limpios:** Los datos de la zona de datos crudos son limpiados, validados y estandarizados. Aquí se eliminan duplicados y se resuelven problemas de formato.
* **Zona datos enriquecidos:** Los datos de la zona de datos limpios se transforman y se enriquecen para casos de uso específicos, como el análisis de negocio. Aquí se pueden crear tablas agregadas y unificadas.

**2. Relación con Data Warehouse y Data Mart**

* **Data Lake vs. Data Warehouse:** El Data Lake almacena datos de forma económica y flexible, ideal para datos no estructurados y análisis exploratorio. El Data Warehouse, en cambio, almacena datos estructurados, optimizados para consultas y reportes de negocio, y se alimenta de la zona de datos limpios del Data Lake.
* **Data Warehouse vs. Data Mart:** Un Data Mart es una versión más pequeña y enfocada de un Data Warehouse, diseñada para las necesidades de un departamento específico (ej. ventas, marketing). Se alimenta del Data Warehouse para proporcionar información a medida.

**3. Gobernanza y gestión de datos**

* **Catálogo de datos:** Implementa un catálogo que describa los datos disponibles, su procedencia, su formato y su calidad. Esto facilita la búsqueda y el acceso a los datos por parte de los analistas.
* **Seguridad:** Define controles de acceso basados en roles para garantizar que solo los usuarios autorizados puedan acceder a la información sensible. Implementa el enmascaramiento de datos (data masking) para proteger la información personal.
* **Trazabilidad:** Monitorea y documenta el linaje de los datos, es decir, su recorrido desde la fuente hasta su uso final. Esto es crucial para la auditoría y la resolución de problemas.

**Lección 3: Calidad de los Datos**

Un plan de calidad de datos es esencial para asegurar que las decisiones de negocio se basen en información precisa y confiable.

**1. Métricas e indicadores**

Define métricas de calidad para cada etapa del flujo de datos:

* **Integridad:** Se revisa que los datos cumplan con las reglas definidas, por ej. el 100% de los pedidos tienen un ID de cliente.
* **Precisión:** Se revisar que los datos cumplen con la información que corresponde, por ej. el precio de un producto es el que realmente se registró en la transacción.
* **Consistencia:** Se revisa que todos los datos sean iguales en todas las fuentes por ej. el nombre de un cliente se escribe de la misma forma en el CRM y en el sistema de pedidos.
* **Vigencia:** Se revisa que los datos estén actualizados por ej. los registros de inventario reflejan el stock actual en tiempo real.
* **Unicidad:** Se revisa que no existan registros duplicados por ej. cada cliente tiene un único ID.

**2. Proceso de monitoreo y remediación**

* **Monitoreo:** Implementa un sistema automatizado que mida las métricas de calidad de datos de forma continua. Cuando una métrica no cumple con el umbral establecido, se genera una alerta.
* **Remediación:** Define un flujo de trabajo para corregir los problemas de calidad. Esto puede incluir la corrección manual de datos, la automatización de la limpieza de datos o la reingesta de datos desde la fuente.

**3. Integración en la arquitectura**

El plan de calidad debe estar integrado en la capa de integración de la arquitectura. Las validaciones de datos se deben ejecutar automáticamente antes de que los datos pasen a la zona de datos limpios o a la capa de consumo.

**Lección 4: Modelamiento Multidimensional**

El modelado multidimensional es la etapa final para preparar los datos para el análisis.

**1. Área de negocio clave**

Para esta parte del desarrollo seleccioné el área de ventas. Esto me permite hacer un modelo de datos específico y optimizado.

**2. Diseño del modelo dimensional**

Se puede usar un esquema estrella, que es ideal para la mayoría de los casos de análisis. Se compone de una tabla de hechos central y varias tablas de dimensiones que la rodean.

* **Tabla de Hechos (Fact Table):** Contiene métricas numéricas y claves foráneas que la conectan con las tablas de dimensiones. Para el análisis de ventas, la tabla de hechos podría llamarse Fact\_Ventas y contener datos como cantidad\_vendida, precio\_total, costo\_producto.
* **Tablas de Dimensiones (Dimension Tables):** Contienen atributos descriptivos que dan contexto a la tabla de hechos. Para el caso de ventas, se pueden incluir dimensiones como:
  + Dim\_Tiempo: año, trimestre, mes, día.
  + Dim\_Producto: nombre\_producto, categoría, subcategoría, marca.
  + Dim\_Cliente: nombre\_cliente, ciudad\_cliente, correo\_cliente, teléfono\_cliente.
  + Dim\_Tienda: nombre\_tienda, ciudad\_tienda, código\_tienda.

**3. Diagrama del modelo**

Interfaz de usuario gráfica, Diagrama

El contenido generado por IA puede ser incorrecto.

**4. Justificación de decisiones**

* **Desnormalización:** Un esquema estrella es un modelo desnormalizado que mejora el rendimiento de las consultas, ya que minimiza las uniones de tablas.
* **Criterios analíticos:** El diseño del modelo debe estar alineado con las preguntas de negocio que se quieran responder, como: "¿Qué producto se vendió más en el último trimestre?", "¿Qué región tuvo el mayor crecimiento de ventas?", etc. El modelo debe permitir la exploración de datos a diferentes niveles de granularidad.