Caso Práctico Reflexivo: Gestión de Datos para una Plataforma de E-Commerce

Una empresa de e-commerce ha experimentado un crecimiento exponencial, generando grandes volúmenes de datos provenientes de diversas fuentes como transacciones, comportamiento de usuarios, opiniones de productos y datos logísticos. Este crecimiento ha presentado desafíos en el manejo, almacenamiento y análisis de estos datos para mejorar la toma de decisiones estratégicas y ofrecer mejores experiencias a los usuarios.

A continuación, se resuelve, de manera básica, el caso práctico, y al final se presentan preguntas abiertas para que el alumno reflexione y aplique los conceptos aprendidos.

Resolución de las Tareas Propuestas

1. **Almacenamiento de Datos:**
    - **Propuesta: Implementar una arquitectura híbrida combinando un Data Lake y un Data Warehouse:**
        - Data Lake: Utilizado para almacenar datos en bruto provenientes de múltiples fuentes, como logs de comportamiento en la web, datos de redes sociales y datos de sensores de logística. Este enfoque flexible permite la ingesta de datos no estructurados y en tiempo real.
        - Data Warehouse: Diseñado para almacenar datos transformados y estructurados, listos para generar informes y realizar análisis específicos, como métricas de ventas y comportamiento del cliente.
    - **Base de Datos Relacional vs. NoSQL:**
        - Relacionales para datos transaccionales estructurados como ventas e inventarios.
        - NoSQL (como MongoDB) para gestionar datos no estructurados, como comentarios de productos y preferencias de usuarios.

---

2. **Procesamiento de Datos:**
    - **Propuesta: Implementar un clúster distribuido basado en Hadoop para el almacenamiento escalable (HDFS) y usar Apache Spark para el procesamiento de datos en memoria:**
        - Hadoop HDFS: Asegura tolerancia a fallos y escalabilidad para almacenar logs masivos.

        - Apache Spark: Permite procesar datos rápidamente, tanto históricos como en tiempo real, ideal para entrenar modelos de Machine Learning para recomendaciones personalizadas.

---

3. **Proceso ETL:**
    - **Propuesta: Diseñar un flujo ETL para integrar datos de múltiples fuentes:**
        - Extracción: Obtener datos transaccionales de bases SQL y datos sociales mediante APIs de redes como Twitter.
        - Transformación: Limpiar valores faltantes, eliminar duplicados, estandarizar formatos (por ejemplo, fechas) y enriquecer los datos con etiquetas adicionales como categorías de producto.
        - Carga: Ingresar los datos transformados en el Data Warehouse para análisis y en el Data Lake para análisis exploratorio.

---

4. **Inteligencia de Negocios (BI):**
    - **Propuesta: Crear dashboards interactivos utilizando herramientas como Power BI o Tableau para visualizar métricas clave:**
        - Métricas recomendadas: Productos más vendidos, ingresos por categoría, tasas de conversión, análisis de tiempo de entrega, y segmentación de clientes según su comportamiento.

---

5. **Escalabilidad:**
    - **Propuesta: Adoptar una estrategia de escalabilidad combinada**:
        - Horizontal: Agregar nodos al clúster para manejar el incremento de datos y carga de trabajo.
        - Vertical: Mejorar recursos de los nodos existentes (memoria RAM, CPU) para mejorar el rendimiento del sistema.

Preguntas para Reflexión y Respuesta del Alumno

1. **Almacenamiento:**
   ¿Qué ventajas y desventajas ofrece la combinación de un Data Lake y un Data Warehouse para este escenario?
2. **Procesamiento:**
   ¿Cómo contribuyen tecnologías como Hadoop y Spark al análisis eficiente de grandes volúmenes de datos en esta plataforma?
3. **ETL:**
   ¿Por qué es importante limpiar y transformar los datos antes de cargarlos en el Data Warehouse?

4. **BI:**
   ¿Qué métricas incluirías en un dashboard de inteligencia empresarial para una plataforma de e-commerce y por qué?
5. **Escalabilidad:**
   ¿Qué retos podrías enfrentar al implementar escalabilidad horizontal en un clúster de procesamiento distribuido?
6. **Tecnologías NoSQL:**
   ¿Cuándo optarías por bases de datos NoSQL frente a bases relacionales en este caso práctico?
7. **Integración en tiempo real:**
   ¿Qué beneficios podría tener integrar herramientas como Apache Kafka para manejar datos en tiempo real en esta plataforma?
8. **Seguridad y Gobernanza:**
   ¿Qué estrategias implementarías para garantizar la seguridad y privacidad de los datos de clientes en esta infraestructura?
9. **Almacenamiento en la nube:**
   ¿Qué ventajas ofrece el almacenamiento en la nube frente a una infraestructura local para este caso?
10. **Impacto Ético:**
    ¿Cuáles son las principales implicaciones éticas relacionadas con la gestión y análisis de datos de clientes en plataformas de e-commerce?