Análisis de Caso - Módulo 7: Big Data

1. 5V's de Big Data presentes en el caso

- Volumen: grandes cantidades de datos de redes sociales, apps, encuestas y compras.
- Velocidad: necesidad de procesar información en tiempo real para detectar tendencias.
- Variedad: múltiples fuentes heterogéneas (texto, clics, registros de compras, encuestas online).
- Veracidad: algunos datos pueden ser ruidosos o inconsistentes (ej. publicaciones en redes sociales).
- Valor: transformar los datos en recomendaciones personalizadas y decisiones estratégicas.

2. Arquitectura Big Data propuesta

Adquisición de datos:

- Apache Kafka o Flume para capturar datos en tiempo real.
- APIs de redes sociales, logs de aplicaciones móviles, integraciones con sistemas internos.

Almacenamiento distribuido:

- HDFS (Hadoop Distributed File System) o Amazon S3 para datos históricos y estructurados/semiestructurados.
- Bases NoSQL como MongoDB o Cassandra para datos de alta velocidad y flexibilidad de esquema.

Procesamiento distribuido:

- Apache Spark para procesamiento batch y streaming.
- Spark MLlib para algoritmos de machine learning.
- Hive o Presto para consultas SQL interactivas.

3. Beneficios esperados

1. Análisis en tiempo real para identificar tendencias de mercado.
2. Recomendaciones personalizadas que aumentan satisfacción y fidelización de clientes.
3. Mejor toma de decisiones estratégicas basada en datos.

4. Riesgos y medidas recomendadas

Riesgos/desafíos:

- Costos de infraestructura y complejidad de integración.
- Riesgo de pérdida o mal uso de datos sensibles.
- Problemas de calidad y consistencia en la ingesta desde múltiples fuentes.

Medidas recomendadas:

- Implementar controles de seguridad (cifrado, autenticación, control de accesos como Kerberos/IAM).
- Estrategia de Data Governance con políticas claras de calidad, propiedad y ciclo de vida de los datos.
- Validaciones automáticas de calidad de datos (profiling, detección de outliers, registros faltantes).