1. ¿QUÉ ES BIG DATA?

Big Data son volúmenes masivos de datos que generan las empresas diariamente. En el caso de DataShop, hablamos de millones de registros que no pueden procesarse con herramientas tradicionales. La idea es convertir esos datos crudos en información útil que ayude al negocio a tomar mejores decisiones.

2. LAS 5 V DEL BIG DATA

Volumen: DataShop maneja millones de registros diarios. Una base de datos normal no puede procesarlos.

Velocidad: Los datos llegan constantemente en tiempo real. Los clics y compras suceden ahora, no después. Necesitan procesarse rápidamente.

Variedad: DataShop recibe diferentes tipos de datos: logs de clics, compras estructuradas, búsquedas de texto y valoraciones. Cada uno en formato diferente.

Veracidad: No todos los datos son exactos. Hay información duplicada o incompleta que hay que limpiar antes de usarla.

Valor: Lo importante es qué sacamos de todo esto. Para DataShop significa personalizar experiencias, vender más y detectar fraudes.

3. ¿POR QUÉ MYSQL O SQL SERVER NO SON SUFICIENTES?

Las bases de datos relacionales no escalan horizontalmente (no pueden distribuir datos entre múltiples máquinas fácilmente), no manejan variedad de formatos (los logs y textos no encajan bien en tablas rígidas), son lentas para analizar millones de registros, y requieren cambios complicados en el esquema cuando quieres agregar nuevos tipos de datos.

4. DOS TECNOLOGÍAS DE BIG DATA

- **Hadoop**
- **Apache Spark**

5. PARA QUÉ SE UTILIZAN

Hadoop: Distribuye datos enormes entre múltiples servidores para procesarlos en paralelo. DataShop lo usaría para almacenar seguramente millones de registros y procesar análisis históricos grandes.

Apache Spark: Procesa datos rápidamente en memoria y puede hacer procesamiento en tiempo real, machine learning y análisis complejos. DataShop lo usaría para analizar clics en tiempo real, entrenar modelos de recomendación y detectar fraudes.

6. ¿QUÉ TIPO DE DATOS MANEJA DATASHOP?

- **Estructurados (20%):** Los registros de compra con estructura clara (ID cliente, producto, fecha, precio).
- **Semiestructurados (30%):** Los logs de clics y búsquedas, que tienen cierta estructura pero son flexibles.
- **No estructurados (50%):** Las valoraciones de clientes en texto libre, que requieren análisis especiales para extraer información