Lección 3.2 . Naturaleza de las fuentes de datos Big Data :

1. Heterogeneidad

En la lección anterior ya vimos la gran variedad de fuentes de datos que tienen cabida en un sistema de BI - Big Data. Precisamente, a ello se debe....

- La característica de Variedad (una de las 5v's) hace referencia a la fuerte **heterogeneidad** que suelen presentar las fuentes de datos disponibles en un contexto de Big Data
- La heterogeneidad se manifiesta en diversos aspectos como:
    - **Nivel de Estructura:** Información bien estructurada, semi estructurada o no estructurada.
    - **Modelo de datos**: El tipo de esquema usado para representar los datos manejados en nuestro sistema de información. (Ej. Relacional, Multidimensional...)
    - **Convenciones de sintaxis:** Ej. "Calle Alcalá, nº 32" o "C\Alcalá, 32 "
    - **Convenciones de semántica:** Ej. campo para almacenar el teléfono de un cliente: Teléfono, Telefono, Tlfno, Tel…
    - **Unidades de medida:** Ej. Distancia en kilómetros, metros, millas,...
- Las fuentes de datos, aún conteniendo información sobre el mismo dominio, suelen presentar una alta heterogeneidad.

2. Formatos

Además, la heterogeneidad también se manifiesta en la existencia de un...

- Amplio abanico de formatos en los que se distribuyen los datos: texto, vídeo, audio, imagen, relacional (conjuntos de filas y columnas), pares clave-valor, Excel, CSV, XML, JSON,...
- En algunos casos, la misma información se distribuye en más de un formato
    - **Interoperabilidad**: Facilita su uso en distintas aplicaciones (ej. ETL, formatos soportado por una BD, aplicaciones de BI)
- Frecuentemente, la estructura de la información determina o limita el número de formatos a usar.

    - Ej. Hoja de Cálculo: Excel, CSV,....
    - Ej. Vídeo: Información sin estructura. Formatos AVI, MP4, MPEG…
- El intercambio de información a través de la red ha fomentado la aparición de formatos ligeros para el intercambio de datos: (entre ellos destacan…)
    - XML: eXtensible Markup Language ('lenguaje de marcas extensible')
    - JSON: JavaScript Object Notation

3. Distribución y velocidad de generación

- En cada fuente de datos la información se genera a una velocidad o ritmo que suele ser distinto
    - Datos económicos **anuales** de un país (ej. PIB)
    - Resumen de ventas de una organización: trimestral, mensual, diario...
    - Información financiera: segundos
- La característica de Velocidad (5v's) en Big Data hace referencia a:
    - El aumento del ritmo al que se genera la información y del número de fuentes en las que la información se distribuye de forma continua, en tiempo real.
    - Los requisitos temporales respecto de la disponibilidad de los datos para su uso o análisis.
        - Algunos datos pierden valor a medida que pasa el tiempo (ej. datos financieros, sensores de seguridad…)
- Los datos de una fuente pueden distribuirse bien empaquetados (ej. archivos, conjuntos de filas y columnas,...) o de forma continua, a medida que se van generando
    - En la lección 4 analizaremos en detalle ambos tipos de fuentes