Lección 1.4 Problemática

1. Introducción:

- Como vimos en la lección anterior, Big Data proporciona un gran abanico de posibilidades a las organizaciones pero, al mismo tiempo, al mismo tiempo, sus características (recordemos las 5v's) plantean importantes problemas que los profesionales de IT de las organizaciones deben afrontar para lograr con éxito sus objetivos
- Encuestas recientes (TDWI, 2013) muestran como sólo un **12%** de las empresas que usan Big Data afirman conseguir un alto grado de éxito en sus objetivos de negocio, frente al **64%** que atribuyen al uso de Big Data un éxito moderado.
    - Un **24%** considera poco exitosa su aplicación.
    - El fracaso se debe a problemas como la incapacidad para gestionar la integración de las fuentes, pobre calidad de los datos, elección de la arquitectura incorrecta, mala dirección del equipo de IT (Infomation Technology) y la falta de personal con las habilidades adecuadas, aumentos de coste no previstos, gestión de los datos en Streaming y algunos otros más.

2. Problemas 1. Volumen:

- **Volumen**: El procesamiento y análisis de los enormes volúmenes es uno de los problemas más evidentes y antiguos.
- Sin embargo, la tecnología actual aporta soluciones, como Apache Hadoop y las bases de datos NoSQL, de bajo coste, escalables y que permiten analizar y procesar terabytes de datos en minutos, o incluso en segundos, sobre hardware comercial.
    - En el módulo 4 analizaremos en detalle esas y otras arquitecturas que dan soporte al Big Data.

3. Problemas 2. Integración, Calidad y Velocidad:

No menos importantes que el volumen son el resto de problemas planteados por las características o V's del Big Data:

- Integración de los datos (**Variedad**)

    - La combinación de fuentes de datos, internas y externas, es una de las formas de añadir valor a los datos originales y prepararlos para el análisis
    - Pero la heterogeneidad presente en la naturaleza de las fuentes de datos, hace que esta tarea requiera un gran esfuerzo humano, sobre todo cuando se maneja un gran número de fuentes.
        - Distintos modelos de datos (ej, distinto nombre de campo), distinto formato (ej, relacional, XML, texto), falta o inexistencia de metadatos (información para describir los datos)...
- Calidad de los datos (**Veracidad**)
    - La posibilidad de integrar fuentes externas con las fuentes internas de nuestra empresa, puede añadir un gran valor a los datos
    - Sin embargo, es difícil (en ocasiones imposible) comprobar la **precisión** de los datos contenido en las fuentes externas, pues su generación no depende de nosotros
        - Falta de datos, ruido, alteraciones…
- Gestión de los datos generados en tiempo real (**Velocidad**)
    - Ciertos datos pueden ver reducido su valor a medida que pasa el tiempo desde su generación (Oportunity)
    - En los últimos años cada vez hay más fuentes que se generan y distribuyen en tiempo real
        - Datos financieros, sensores, datos de transacciones de tarjetas de crédito,...
    - La gestión y análisis de estas fuentes en tiempo real supone acelerar procesos que antes podían llevar días, o incluso meses, realizarlos.
        - Requiere integración y comprobación de la calidad en tiempo real...
        - ...pero es frecuente reducir los requisitos de calidad de los datos resultantes para que sea viable.

4. Otros problemas:

Además de los problemas directamente relacionados con las características o V's que definen Big Data, existen otros problemas a tener muy en cuenta cuando se decide implementar Big Data en una organización

- La falta de personal con las **habilidades** adecuadas:
    - El **objetivo de este curso** es iniciar la formación de nuevos profesionales en Big Data
- Selección de la arquitectura idónea
    - ¿Qué base datos NoSQL es la más adecuada?

    - ¿Clúster local o uso de servicios y almacenamiento en la nube?
    - ¿Hadoop o BD's relacionales con características extendidas para Big Data?
- Otros problemas
    - Coexistencia con una infraestructura de almacén de datos (Data Warehouse) existente
    - Coste de implementación y mantenimiento
    - Pobre integración entre las herramientas Big Data existentes
    - Inmadurez en el manejo de los nuevos formatos de datos y fuentes
    - Falta de gobernanza de datos: Definición y cumplimiento de estándares de calidad, privacidad, seguridad, gestión de los metadatos ...
    - ...y algunos más

5. Resumen del módulo 1

Recapitulando, en este módulo hemos visto:

- Big Data surge por la evolución de las TIC y queda definido por la 5v's o características fundamentales
    - **Volumen, Variedad, Velocidad, Veracidad** y **Valor**
- El procesamiento y análisis de las nuevas fuentes Big Data permite la extracción de conocimiento implícito en los datos
    - Muy útil para la **toma de decisiones estratégicas** en una organización
    - Permite añadir valor a los datos internos de una organización con la integración fuentes externas, algunas de ellas de libre acceso (ej. Open Data)
- Sin embargo, las propias características del Big Data hacen que su aprovechamiento no sea, ni mucho menos, una tarea trivial
    - Principales problemas: Integración de los datos (**Variedad)**, comprobación de la calidad de los datos (**Veracidad**), gestión y análisis de las fuentes en tiempo real (**Velocidad**)
    - Las tecnologías y técnicas de procesamiento y análisis previamente existentes (ej. Data Warehouses sobre BD's relacionales) resultan insuficientes en Big Data
    - La aparición de **nuevas tecnologías** diseñadas para Big Data ha logrado **superar** en gran medida el problema de **Volumen**