# Introducción al Big Data

## Open Course Ware

Dagoberto Castellanos Nieves  Belén Melián Batista

Departamento de Ingeniería Informática
Universidad de La Laguna

# Contenidos

# Introducción

- El Big Data en una combinación de tecnologías de gestión de datos que han evolucionado en las últimas décadas.
- Permite a las compañías almacenar y manipular grances volúmenes de datos a la velocidad adecuada y en el momento oportuno.
- Big Data no es una solución aislada; es necesario hacer confluir una estrategia de negocio con una técnica para aprovechar esta tendencia tecnológica.
- Big Data es una combinación de los 50 años de evolución de la tecnología.
- Debemos conocer las tecnologías emergentes que usan las compañías (Hadoop, MapReduce, etc.)

# Introducción

- Las compañías han trabajado durante años para obtener información útil de sus clientes, productos y servicios.
- Algunos datos son estructurados y están almacenados en bases de datos.
- Sin embargo, otros, incluyendo documentos, imágenes y vídeos, son desestructurados.
- Además, las compañías tienen que considerar datos obtenidos de nuevas fuentes, como son los sensores, las redes sociales, las páginas web, etc.

# Evolución de los sistemas de gestión de datos

- La mayoría de las nuevas alternativas de gestión de datos se construyen sobre sus predecesoras.
- La gestión de datos se va adaptando a los avances tecnológicos del hardware, del almacenaciento, así como de los modelos de computación como el cloud computing.
- **Big Data se define** como cuaquier fuente de datos que tiene al menos tres características compartidas:
  1. **V**olúmenes de datos extremadamente grandes.
  2. **V**elocidad de los datos extremadamente alta.
  3. **V**ariedad de los datos extremadamente amplia.

# Estructuras de datos manejables I

- 1960s: datos almacenados en ficheros planos sin estructura.
- 1970s: invención del modelo de datos relacional y del sistema de gestión de bases de datos relacionales. Este modelo incluía un nivel de abstracción (uso de SQL) que permitía satisfacer demandas crecientes de negocio.
- Sin embargo, una demanda explosiva hizo poco económico almacenar el creciente volumen de datos. Además, el acceso era lento.
- Surge, por tanto, el modelo Entidad-Relación, que añadía abstracción adicional para incrementar la usabilidad de los datos.
- El mercado de las bases de datos relacionales explosiona y se mantiene vibrante aún en la actualidad.

# Estructuras de datos manejables II

- El data warehouse proporcionó la solución cuando el volumen de datos a manejar por las organizaciones comenzó a crecer sin control.
- 1990s: comercialización de los Data Warehouses.
- Sin embargo, cuando fue necesario manejar enormes volúmenes de datos no estructurados o semi-estructurados, el warehouse fue insuficiente.
- Cómo podían las compañías transformar sus enfoques de gestión tradicionales para manejar datos no estructurados?
- La solución no surgío de la noche a la mañana.
- Surgen los sistemas de gestión de bases de datos basadas en objetos.

# Web y gestión de contenidos

- La mayor parte de los datos disponibles en el mundo en la actualidad es no estructurada.
- En los años 80, los sistemas de gestión de contenidos de las empresas evolucionaron para proporcionar la capacidad de gestionar datos no estructurados; fundamentalmente documentos.
- En los años 90, con el auge de la web, las organizaciones pretendieron ir más allá y poder gestionar contenido web, imágenes, audio y video.

# Gestionando Big Data

- Es realmente nuevo el big data o es una evolución en la travesía de gestión de datos?
- La respuesta es que ambas cosas; el big data se ha construido sobre la evolución de las prácticas de gestión de datos de las últimas cinco décadas.
- La diferencia está en el coste; ahora es posible almacenar y manejar toda la información requerida por las compañías, y no sólo la más importante.
- Además, las mejoras en la velocidad y fiabilidad de la red han eliminado otras limitaciones físicas de ser capaz de gestionar cantidades masivas de datos a un ritmo aceptable.
- A esto debemos añadir el impacto de los cambios en el precio y sofisticación de la memoria de los ordenadores.

# Importancia del big data

- Si las compañías son capaces de analizar petabytes de datos con un rendimiento aceptable, para descubrir patrones y anomalías, los negocios pueden comenzar a tener sentido de los datos en nuevos modos.
- La ciencia, la investigación y las actividades de los gobiernos también se beneficiarán del salto al big data.
- Pensemos, por ejemplo, en la posibilidad de analizar el genoma humano o tratar con todos los datos astronómicos obtenidos en los observatorios para incrementar nuestra comprensión del mundo que nos rodea.
- Por tanto, el Big Data no está centrado únicamente en los negocios.

# Definición de big data

- Big data no es una única tecnología, sino una combinación de viejas y nuevas tecnologías que ayudan a las compañías a obtener una visión práctica.
- Big data es la capacidad de gestionar un gran volumen de datos dispares, a la velocidad adecuada y en el momento oportuno para permitir, en tiempo real, el análisis y la reacción.
- Tal como se indicó alteriormente, el big data se divide típicamente en tres características:
  1. ***V*olumen**: Cuántos datos?
  2. ***V*elocidad**: Cuánto de rápido se procesan esos datos?
  3. ***V*ariedad**: Los diversos tipos de datos.
- Aunque es conveniente simplificar el big data en estas tres *V*s, puede ser demasiado simplista. Por ello, surge una cuarta *V*:
  1. ***V*eracidad**: Cómo de exactos son los datos para alcanzar valor de negocio?