BIG DATA

Fin del artículo

Venancio Guntiñas Rodríguez
vguntinas2@gmail.com

A las empresas actuales se les presentan nuevos problemas que antes no tenían. Tienen que manejar nuevos tipos de datos, se requieren mayores velocidades de procesamiento, nuevos enfoques de soluciones BI (Business Intelligence),... por estos motivos, han surgido diferentes propuestas para tratar estos problemas. Estas propuestas, no sustituyen a las Bases de Datos Relacionales ni a los procesos de creación de proyectos BI, sino que, son "herramientas" con las que se puede trabajar para resolver con mayor eficacia esos problemas.

Se llama Big Data, al proceso de recolectar grandes volúmenes de datos (del orden de petabytes y exabytes) no estructurados, para encontrar patrones o coincidencias entre ellos y determinar conductas o tendencias.

Hay que distinguir el BigData (BD) del Business Intelligence (BI). Con la evolución de la tecnología han surgido nuevos tipos de datos que no se pueden tratar con las tecnologías BI tradicionales. Se generan millones de datos no estructurados en muy poco tiempo, que se quieren analizar, pero no se pueden almacenar en bases de datos relacionales.

DATOS NO ESTRUCTURADOS

Se pueden considerar los siguientes tipos:

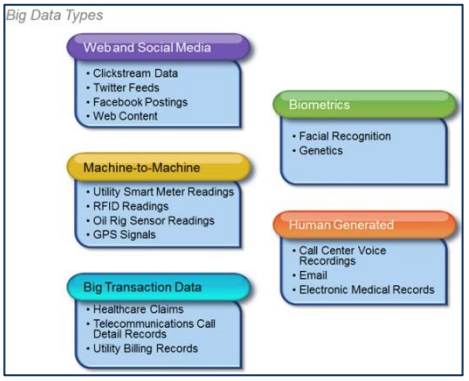
- **Datos no estructurados**: aquellos que no pueden almacenarse en bases de datos relacionales con estructuras predefinidas.
- **Datos semiestructurados**: datos que no se almacenan en bases de datos relacionales, pero presentan una organización interna que facilita su tratamiento. Por ejemplo: documentos XML y datos almacenados en bases de datos NoSQL (No solo SQL).
- **Datos no estructurados de tipo texto**: datos generados en las redes sociales, foros, e-mails, presentaciones Power Point, documentos Word,…
- **Datos no estructurados que no son texto**: archivos de imágenes jpeg, archivos de audio mp3, archivos de vídeo tipo flash,...

CARACTERÍSTICAS DE LOS DATOS NO ESTRUCTURADOS

Las principales características de los datos no estructurados son:

- **Volumen y crecimiento:** el volumen de datos y la tasa de crecimiento de los datos no estructurados es muy superior al de los datos estructurados. Por ejemplo, twitter genera 12 Terabytes de información cada día.

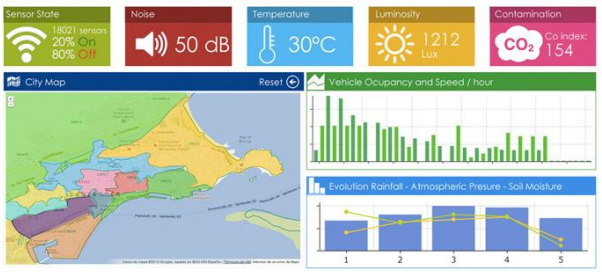
- **Orígenes de datos:** El origen de los datos es muy diverso: datos generados en redes sociales, datos generados en foros, e-mails, datos extraídos de la web, documentos internos de la compañía (word, pdf, ppt), datos de dispositivos móviles, audio, video, sistemas GPS, sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad, cambios químicos que sufre el aire. Las aplicaciones que analizan estos datos deben dar una respuesta rápida para lograr obtener la información correcta en el momento preciso.



- **Almacenamiento:** No se puede utilizar una arquitectura relacional. Es necesario utilizar arquitecturas especiales llamadas "Big Data". En estas arquitecturas son fundamentales la **escalabilidad** y el **paralelismo**. En ciertos casos, es necesario almacenar los datos en la Nube. Para reducir costes de almacenamiento, se debe monitorizar la frecuencia de uso y la detección de datos inactivos.
- **Terminología e idiomas:** En el caso de datos no estructurados de tipo texto, se suele llamar a lo mismo de diferentes formas. Otra cuestión es el idioma en el que se ha generado la información tratada.
- **Seguridad:** Algunos datos no estructurados de tipo texto, pueden no ser seguros y el control de accesos a los mismos, es complejo debido a cuestiones de confidencialidad y a la difícil clasificación del dato.

ESCALABILIDAD Y PARALELISMO

Actualmente, la velocidad de generación de datos es muy elevada por lo que, se necesita un sistema con varias CPUs (sin límite) que puedan trabajar en paralelo y ahorrar tiempo siguiendo la técnica del "divide y vencerás". Ocurre que las bases de datos relacionales, no pueden distribuirse en nodos diferentes de manera transparente al usuario, de forma sencilla. Para conseguirlo, se necesita añadir CPUs y Memoria pero, en este caso, existe un límite.



EJEMPLO DE PROCESAMIENTO DE DATOS DE SENSORES EN UNA CIUDAD INTELIGENTE

Por otro lado, el modelo de bases de datos relacionales no soporta todos los problemas, por ejemplo: no se heredan objetos o no se pueden tener columnas variables según las filas.

Por estos motivos, se ha necesitado crear nuevas herramientas y sistemas que aporten una forma alternativa de abordar los problemas, para mejorar el procesamiento y análisis de datos.

CUESTIONES A CONSIDERAR EN EL TRATAMIENTO DE INFORMACIÓN NO ESTRUCTURADA

- **Crear una plataforma escalable (infraestructura y procesos)** que permita tratar grandes cantidades de datos. Es necesaria una capacidad de almacenamiento y una capacidad de proceso escalable. Teniendo en cuenta el coste económico de mantener plataformas escalables, hay que considerar como opción el **procesamiento en la Nube**. Desde el punto de vista de los procesos, en ocasiones es interesante utilizar in-memory analytics.
- **Añadir información/estructura complementaria a los datos no estructurados** que ayude a su tratamiento. Por ejemplo, en una colección de tweets de redes sociales puede ser interesante añadir campos tales como el idioma, la localización geográfica…

- **Crear conjuntos reducidos de datos que sean representativos**. Debido al gran volumen de datos, se deben crear muestras de datos que sean estadísticamente representativas. Este ha sido uno de los grandes problemas de Google al inicio de su actividad. Necesitaban procesar millones de millones de páginas para poder obtener el resultado de su PageRank de forma resumida, para ello, crearon el **algoritmo MapReduce** para resumir de forma sencilla todos esos datos.
- **Desarrollo de algoritmos**. Hay diferentes maneras de manejar la información no estructurada. Yahoo creó **Hadoop**, una base que envuelve el MapReduce y garantiza que se puedan ejecutar en nodos distribuidos, programas MapReduce realizados por los usuarios. Esta herramienta tiene un **sistema de archivos HDFS** que permite la distribución de trabajos a diferentes nodos que ejecutarán en paralelo el algoritmo de reducción MapReduce. El HDFS consigue hacer transparente o simplificar la creación de clusters de nodos que trabajan en paralelo como un solo nodo. Si se necesita más potencia, basta con añadir una nueva IP de un nodo en un archivo.

  Para procesos de text mining, puede utilizarse natural language processing combinado con redes neuronales. Otras técnicas como redes bayesianas permiten descubrir patrones sobre múltiples dimensiones. Son importantes también las técnicas de visualización de datos.
- **Procesos de depuración/limpiado de datos.** Debido al enorme volumen de los datos, hay que gestionar eficientemente el histórico de los datos. Detectar datos no usados o poco consultados para poder borrarlos y liberar espacio.

Inicio