Lección 3.4 Lotes y Streaming :

Como ya avanzamos en la lección 3.2 (cuando hablábamos de las diferencias en la naturaleza de las fuentes de datos)...

- Las fuentes de datos presentan diferencias en ritmo en que se generan y en la forma en que se distribuyen
    - (Es uno de los aspectos a los que hace referencia la...) Característica de Velocidad de Big Data (5v's)
- Dos posibles escenario:
    - (Hasta ahora) En la mayoría de las fuentes, los datos se distribuyen bien empaquetados en archivos u otros formatos (ej. filas y columnas en consulta a BD)
    - (Pero cada vez más) En fuentes en las que el ritmo de generación es alto o muy alto, los datos se suelen distribuir de forma continua → **Datos en Streaming**
        - Ej: Datos de sensores, redes sociales,....
- Las particularidades de estos dos escenarios hacen necesario el uso de distintas técnicas y herramientas para la adquisición y el procesamiento de las fuentes de datos

1. Datos en lotes

(Introducir usando imágen de pila de libros o archivos)

Cuando los datos se distribuyen bien empaquetados también nos referimos a ellos como lotes

- **Lote (de datos)**: Datos **históricos** que se distribuyen de forma conjunta en un "paquete".
    - El paquete puede ser un archivo, un conjunto de archivos, múltiples filas de una tabla en una BD relacional…
    - ~~Los datos se almacenan en el "paquete" antes de su distribución~~
- ~~La gestión de este tipo de fuentes en Big Data es sencilla~~
- El proceso de carga de estos datos en un repositorio o almacén es conocido como **carga por lotes (batch loading)**
    - Carga directa o usando alguna herramienta de ETL (Extraction, Transformation and Load, en castellano, Extracción, Transformación y Carga)

2. Datos en Streaming

(Introducir usando imágen de río o cascada) Sin embargo, en otras ocasiones los datos se generan y distribuyen de forma continua, al igual que fluye el agua de un río. A este tipo de fuentes se las conoce como:

- **Streaming:** Fuentes donde los datos se generan y distribuyen de forma **continua**
    - Datos **actuales**: el periodo de tiempo que transcurre desde que se generan hasta que disponemos de los datos para su procesamiento es muy corto (del orden de minutos, segundos, milisegundos,...)
    - En el pasado el término streaming se aplicaba únicamente a la distribución de contenidos multimedia como al audio y el vídeo
    - (Sin embargo) En la actualidad, cada vez son más las fuentes que distribuyen datos estructurados, semi estructurados y nada estructurados, de forma continua.
- Ejemplos:
    - Streaming de redes sociales: API's streaming de Twitter
    - Streaming de datos de sensores: API's de Xively
        - Calidad del aire, consumo de energía,....
    - Logs de servidores web

Podemos sacar una segunda transparencia

3. Consideraciones

- **Ojo:** Los datos pueden generarse de forma continua pero almacenarse en lotes para su distribución.
- La adquisición de fuentes en streaming requiere de herramientas y técnicas especiales
    - Es frecuente que, para este tipo de fuentes en las que la Velocidad de generación es elevada, los requisitos de tiempo para el análisis también lo sean
    - (Sin embargo) El procesamiento y análisis en tiempo real aumenta los requerimientos hardware del sistema
- En módulo 5 estudiaremos las prácticas más actuales para gestionar ambos tipos de fuentes.