Ingeniería de Servidores
Big Data y TPCx-HS

Resumen

En este texto puedes incluir un resumen del documento. Este informa al lector sobre el contenido del texto, indicando el objetivo del mismo y qué se puede aprender de él.

Índice

1. Introducción a Big Data

Desde hace miles de siglos el ser humano ha investigado la manera de almacenar y recopilar información. Durante muchos siglos la escritura y la pintura eran los únicos mecanismos existentes. Posteriormente surgió la fotografía, los discos de vinilo... Sin embargo, poca información seguía ocupando mucho volumen físico. Gracias a los avances tecnológicos de las últimos décadas, hoy en día disponemos dispositivos electrónicos para el almacenamiento de datos binarios. Además, la evolución de estos dispositivos ha sido frenética. IBM comercializó el primer disco duro en 1956. Este constaba solamente de 5 Mega Bytes de capacidad [2] mientras que actualmente podemos utilizar discos duros con más de 1 Tera Byte.

La evolución en la capacidad de cómputo y procesamiento de los computadores también ha sido exponencial. El primer ordenador comercial se presentó en ....(Referencia). Este computador realizaba n cuentas por segundo. Actualmente hablamos de GHz cuando comparamos la velocidad del procesador de un ordenador. Además, es habitual utilizar ordenadores con 8 o más GB de memoria principal. Se han creado benchmarks.... (medio párrafo diciendo para que sirven y que son muy costosos computacionalmente).

Estas nuevas tecnologías han posibilitado que el almacenamiento de datos sea mucho más sencillo. Podemos guardar multitud de archivos multimedia en un dispositivo de unos centímetros y compartirlos con multitud de usuarios.

El mayor flujo de datos se produce gracias a Internet. Aunque es relativamente joven, se hizo público en 1993, actualmente existen más de mil millones de páginas webs [4]. Además, multitud de dispositivos electrónicos se conectan e interaccionan con Internet (lo que se denomina Internet de las cosas [5]). Los usuarios de estos dispositivos utilizan aplicaciones web y redes sociales, publicando textos y archivos multimedia.

Todo este cúmulo de tecnologías y actividades ha dado lugar a que hoy en día haya más de 10 Zeta Bytes de información almacenados (1 ZB = 10^{12} GB) (ver la Figura 1). De hecho, podemos observar que cada año se generan varios Zeta Bytes de información, el crecimiento es exponencial. 5 Exa Bytes (10^9 Giga Bytes) se habían almacenado hasta 2003. Actualmente, generamos esta cantidad de información en dos días [3].

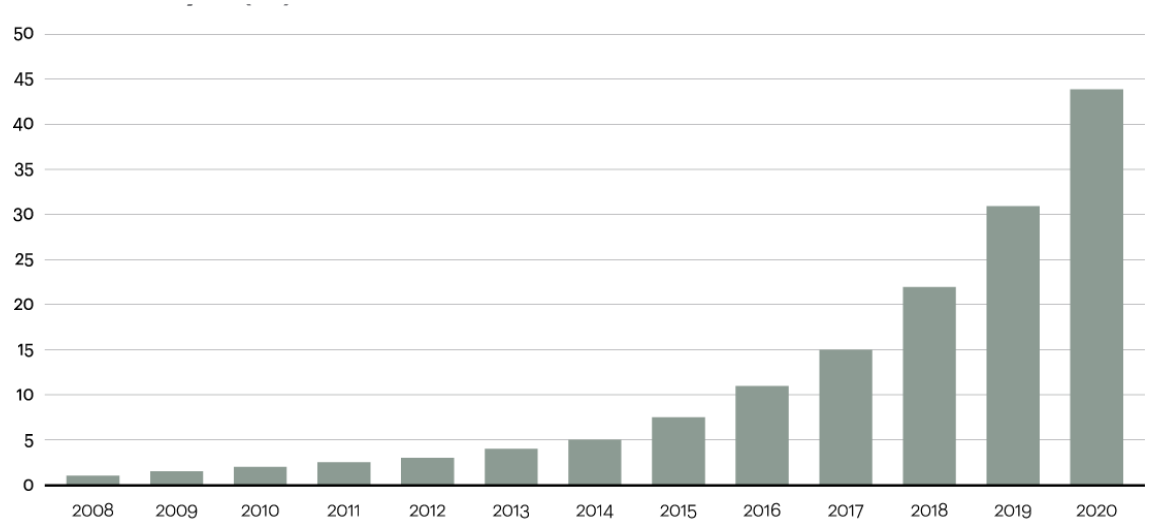

Figura 1: Evolución histórica del número de Zeta Bytes de información almacenados [1].

Esta cantidad de datos contiene conocimiento. Por ejemplo, todas las empresas almacenan información sobre sus clientes, la actividad y transacciones realizadas. Esta información necesita ser analizada en tiempo real para poder actuar en consecuencia. Aquella empresa que mejor conozca el mercado y actúa

rápidamente obtendrá mejores resultados. La ciencia de datos es la rama de la inteligencia artificial que se encarga de tratar y extraer conocimiento de los datos. Se han desarrollado múltiples algoritmos para ello (cita machine learning), que permiten resultados tan impactantes como el reconocimiento de voz o los sistemas de recomendación.

Sin embargo, a pesar de la evolución de los computadores en todas sus facetas, la cantidad de datos e información a procesar y almacenar crece incluso a mayor velocidad. Un ordenador normal no es capaz de tratar tantos GBs de datos. Es más, los algoritmos habituales de aprendizaje necesitan mucho tiempo de cómputo sobre estos conjuntos de datos masivos. Por tanto, necesitamos recurrir a servidores de altas prestaciones y procesamiento distribuido para poder aplicar técnicas de análisis de datos.

Los servidores de altas prestaciones llevan tiempo utilizándose en el ámbito de la inteligencia artificial (referencias). Se han desarrollado herramientas de cómputo en paralelo y distribuido, como OPEN MP y MPI (referencias), que permiten implementar algoritmos distribuidos sobre estos. Sin embargo, estas implementaciones dependen del servidor y son a bajo nivel. Una determinada implementación sobre un esquema hardware puede funcionar bien en determinado momento pero al año tendrá que ser capaz de trabajar con el doble de datos. Esto probablemente suponga la necesidad de ampliar el hardware y rehacer la implementación. Necesitamos pues nuevos paradigma de programación que permitan abstraer el desarrollo de software para plataformas distribuidas del hardware y proporcionen capacidad para el tratamiento de datos masivos.

- **Volumen.** El tamaño de la información a procesar es cada vez mayor, por ejemplo, facebook procesa cada día 500 TB de información. Este volumen de datos require técnicas específicas
- **Velocidad.** Los datos deben ser procesados rápidamente. Normalmente, el procesamiento de los datos debe ser incluso continuo.
- **Variedad.** Los datos a tratar provienen de una gran variedad de fuentes. Por tanto, las herramientas Big Data deben permitir procesar a la vez datos de diferentes características y tamaños.

2. Map Reduce

3. Spark y Flink

4. Benchmarks: TPCx-HS

5. Referencias

Referencias

[1] Hugo Evans y et. al. *Big Data and the Creative Destruction of Today's Business Models.* URL: http://www.atkearney.es/paper/-/asset_publisher/dVxv4Hz2h8bS/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192.

[2] Rex Farrance. «Timeline: 50 Years of Hard Drives». En: *PCWorld* (2016). URL: http://www.pcworld.com/article/127105/article.html.

[3] Seref Sagiroglu y Duygu Sinanc. «Big data: A review». En: *International Conference on. IEEE* (2013), págs. 42-47.

[4] Internet live stats. *Total number of Websites.* URL: http://www.internetlivestats.com/total-number-of-websites/.

[5] Mario Tascón y Arantza Coullaut. *Big data y el Internet de las cosas.* Catarata, 2016.