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

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 [3] mientras que actualmente podemos utilizar discos duros con más de 1 tera byte.

La capacidad de cómputo y procesamiento de los computadores también ha crecido de forma exponencial. El primer ordenador comercial se presentó en 1951 y se conoce como UNIVAC 1 [6]. Este computador realizaba n cuentas por segundo y tenía m kilo bytes de memoria principal o RAM. Actualmente utilizamos procesadores con más de 2 giga gercios de frecuencia de reloj, esto es, realizan hasta dos mil millones de operaciones por segundo. Además, es habitual utilizar ordenadores con 8 o más giga bytes de memoria principal, lo que permite trabajar con bastante información de forma eficiente.

Estas nuevas tecnologías han posibilitado que el almacenamiento de datos sea mucho más sencillo. Podemos guardar multitud de archivos en un dispositivo de unos centímetros y compartirlos con cualquier usuario. Además, el procesamiento de estos archivos e información es eficiente gracias a la capacidad de los computadores actuales.

El mayor flujo de datos es producido en Internet. Aunque es relativamente joven, se hizo público en 1993, actualmente existen más de mil millones de páginas webs [9]. Además, multitud de dispositivos electrónicos se conectan e interaccionan con Internet (lo que se denomina Internet de las cosas [10]). 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). La Figura 1 muestra la evolución histórica de la cantidad de información acumulada por el ser humano. Podemos observar que cada año se generan varios zeta bytes de información, el crecimiento es exponencial. Hasta 2003 se habían almacenado en total 5 exa bytes de información (1 EB = 10^9 GB). Actualmente, generamos esta cantidad de datos en dos días [8].

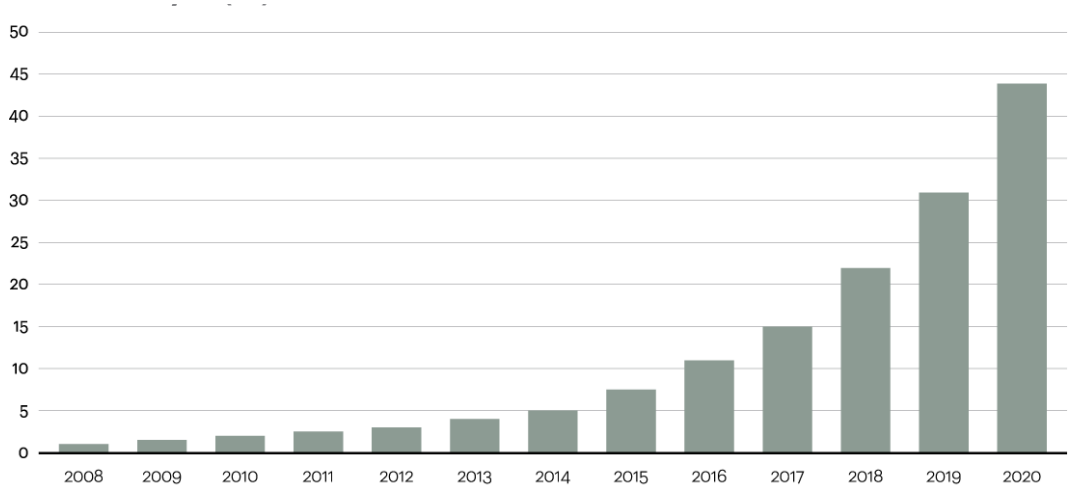

Figura 1: Evolución histórica del número de Zeta Bytes de información almacenados y predicción para los próximos años [2].

Toda esta cantidad de información necesita ser procesada y clasificada. Por ejemplo, encontramos empresas como Facebook y Google cuyos usuarios generan una gran cantidad de datos diariamente. Estos datos deben ser tratados en tiempo real para poder mantener los sistemas de recomendaciones asociados. En estos ejemplos la cantidad de datos a procesar supera con creces la capacidad de un computador usual. Se denominan conjuntos de datos masivos. Estos conjuntos de datos requieren nuevas tecnologías y algoritmos que permitan tratarlos eficientemente. El desarrollo de nuevas tecnologías, algoritmos y herramientas para tratar conjuntos de datos masivos es lo que se conoce como Big Data [4].

En este trabajo presentamos una introducción a Big Data y a las diferentes herramientas software existentes, destacando el papel de la ingeniería de servidores en este contexto. En particular, mostramos la importancia del desarrollo de nuevos benchmarks que permitan evaluar estas tecnologías de forma clara y objetiva.

El resto del texto se organiza como sigue. La Sección 2 contiene una mayor descripción del concepto de Big Data. En la Sección 3 introducimos el paradigma de programación map reduce y la tecnología que lo implementa, denominada Hadoop. En la Sección 4 describimos las nuevas tecnologías que han surgido para cohesionar la filosofía map reduce con el procesamiento iterativo. Estas se denominan Spark y Flink. En la Sección 5 explicamos uno de los benchmarks existentes para las tecnologías Big Data, denominado TPCx-HS. Por último, en la Sección 6 presentamos las conclusiones obtenidas.

2. Big Data

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. En muchas ocasiones un ordenador normal no es capaz de tratar tantos Giga Bytes de datos. Estos conjuntos de datos se denominan masivos. Además, en múltiples aplicaciones se necesita aplicar algoritmos sobre los conjuntos de datos recopilados, requiriendo mucho tiempo de cómputo en el caso de que estos sean de gran tamaño.

Big Data engloba el tratamiento de datos masivos desde el punto de vista tecnológico y algorítmico. En palabras de Michael J. Franklin, profesor de informática en la universidad de Berkley [11]:

“Un problema sobre datos entra en el ámbito de big data cuando la aplicación de las actuales tecnologías no permite al usuario obtener soluciones rápidas, efectivas en costo y de calidad”

En la literatura especializada se destacan las siguientes características de Big Data, que se denominan las 3 V’s de Big Data [8]:

- **Volumen.** El tamaño de los conjuntos de datos a procesar es cada vez mayor, por ejemplo, facebook procesa cada día 500 TB de información. Este volumen de datos requiere tecnologías específicas para que los servidores de altas prestaciones puedan manejar la información con éxito.
- **Velocidad.** Necesitamos herramientas que permitan procesar y analizar conjuntos de datos masivos en poco tiempo. Además, es habitual que el procesamiento de los datos deba ser incluso en tiempo real, esto es, los datos llegan al sistema de forma continua y este debe agregar la información de los mismos.
- **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. Es más, habitualmente encontramos datos de tres tipos: estructurados, semi estructurados y sin estructurar. Los datos estructurados son sencillos de clasificar. Sin embargo, los datos sin estructurar son aleatorios y difíciles de analizar. Por su parte, los datos semi estructurados requieren técnicas avanzadas para poder clasificarlos correctamente.

Algunos autores han extendido la definición hasta utilizar un total de 9 V's: veracidad, valor, viabilidad y visualización entre otras [12].

Los conjuntos de datos contienen conocimiento que necesitamos extraer. 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 (referencia). Se han desarrollado múltiples algoritmos para ello (referencia machine learning), que permiten resultados tan impactantes como el reconocimiento de voz o los sistemas de recomendación. Big Data Analytics [5].

3. Map Reduce. Hadoop

Por tanto, necesitamos recurrir a servidores de altas prestaciones y procesamiento distribuido para poder aplicar estos algoritmos.

Los servidores de altas prestaciones (alguna descripción). 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.

4. Spark y Flink

5. Benchmarks: TPCx-HS

La variedad de computadores es bastante heterogénea. Cada uno realiza de forma eficiente un determinado conjunto de operaciones a consta de presentar peores resultados en otros factores. Por tanto, la comparación entre diferentes modelos de computadores es compleja. Consecuentemente, se han creado benchmarks con el objetivo de aportar elementos de juicio con los que se discernir entre el uso de un computador u otro para una determinada aplicación. Habitualmente el benchmarking se define como la obtención de información útil mediante pruebas empíricas que ayude a una organización a mejorar sus procesos [1]. Sin embargo, el benchmarking de computadores es un proceso costoso computacionalmente. Gasta tanto energía como mucho tiempo de cómputo. Por tanto, se ha de reservar para cuestiones sean importantes y no para evaluar tareas simples [7].

6. Conclusión

Referencias

[1] Confederación Granadina de Empresarios. *¿Qué es el Benchmarking*. URL: `http://www.cge.es/portalcge/tecnologia/innovacion/4111benchmarking.aspx`.

[2] 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.
[3] Rex Farrance. «Timeline: 50 Years of Hard Drives». En: *PCWorld* (2016). URL: http://www.pcworld.com/article/127105/article.html.
[4] F. Herrera. *Inteligencia artificial, inteligencia computacional y Big Data*. Servicio de Publicaciones, Universidad de Jaén, 2014.
[5] Karthik Kambatla y col. «Trends in big data analytics». En: *Journal of Parallel and Distributed Computing* (2014), págs. 2561-2573.
[6] Computer History Museum. *Timeline of Computer History*. URL: http://www.computerhistory.org/timeline/1951/.
[7] Universidad Técnica José Peralta. *Ventajas, desventajas y causas posibles de fracasos del benchmarking*.
[8] Seref Sagiroglu y Duygu Sinanc. «Big data: A review». En: *International Conference on. IEEE* (2013), págs. 42-47.
[9] Internet live stats. *Total number of Websites*. URL: http://www.internetlivestats.com/total-number-of-websites/.
[10] Mario Tascón y Arantza Coullaut. *Big data y el Internet de las cosas*. Catarata, 2016.
[11] Steve Todd. *AMPed At UC Berkeley*. URL: http://stevetodd.typepad.com/my_weblog/2011/08/amped-at-uc-berkeley.html.
[12] P. Zikopoulos y C. Eaton. *Understanding big data: Analytics for enterprise class hadoop and streaming data.* McGraw-Hill Osborne Media, 2011.