



# Ciencia y Tecnología **Big data:** procesar grandes volúmenes de información

**Imagínesse que se podría recolectar en una sola base de datos toda la información geosísmica del mundo, pero que a la vez se pudiese procesar en cuestión de segundos**

EVELIO MARTÍNEZ\* /COLABORADOR

La red internet, la computación en la nube, los dispositivos móviles, los datos científicos, entre otra información, ha servido para el desarrollo de la ciencia y tecnología y otras áreas del conocimiento en el mundo. En la actualidad se ha visto que la cantidad de información que se ha generado en forma digital es inmensa. Según la compañía IBM, sólo en los últimos dos años se ha generado 90 por ciento de la información total existente en el mundo. Por otro lado, la firma consultora Gartner, predice que la tasa de crecimiento anual de la información en el mundo es de 59 por ciento.

Cada día se generan más de 2.2 millones de terabytes de nuevos datos. Un terabyte equivale a un uno seguido por 12 ceros (1,000,000,000,000 bytes). Pero la cantidad de información que existe hoy en día, podría medirse ya en términos de petabytes y exabytes. Un petabyte es equivalente a un uno seguido de 15 ceros, mientras que un exabyte equivale a un uno

seguido por 18 ceros de bytes.

Para lograr la manipulación de tanta información ya existen nuevas propuestas tecnológicas. A esta nueva área dentro de las Tecnologías de la Información se le conoce como big data, o grandes datos en español.

## Qué es big data

“Son grandes y complejas colecciones de datos los cuales, debido a su enorme tamaño, es muy difícil su captura, curación, almacenamiento, búsqueda, compartición, análisis, visualización, administración, procesamiento, entre otras, utilizando las técnicas de software y base de datos tradicionales en un lapso de tiempo tolerable”.

Los grandes datos están compuestos de información estructurada y sin estructurar, éstos pueden estar en la red o fuera de la red. Diez por ciento de la información está estructurada dentro de bases de datos. Noventa por ciento de los grandes datos no están estructurados y corresponde a correos electrónicos, videos, tweets, comentarios en facebook, conversaciones de centros

de llamadas, videos de circuitos cerrados de TV, llamadas de teléfonos móviles, etc.

Según la firma consultora, Gartner, los grandes datos pueden ser definidos usando las tres “V” de la siguiente manera:

▶ **Volumen:** se refiere a los grandes volúmenes de información que se generan diariamente.

▶ **Velocidad:** se refiere a la velocidad de cómo los datos deben ser producidos y procesados para satisfacer las demandas.

▶ **Variedad:** se refiere a los diferentes tipos de formatos de información, ya sea estructurada y sin estructurar.

## Tecnologías para manipularlo

Muchos especialistas en el área han propuesto una variedad de tecnologías tanto de hardware como de software que podrían apoyar a la manipulación de los Grandes Datos, entre ellas, se encuentran las siguientes:

▶ Servidores económicos con un abundante espacio de almacenamiento

- ▶ Procesadores más rápidos
- ▶ Procesamiento en paralelo masivo de bases de datos
- ▶ Virtualización de servidores
- ▶ Enormes ambientes de malla (grid) de minería de datos
- ▶ Computación en la nube
- ▶ Bases de datos orientadas a columnas
- ▶ Bases de datos sin esquema
- ▶ Algoritmos genéticos
- ▶ Modelado predictivo
- ▶ Aprendizaje supervisado y no supervisado
- ▶ Aplicaciones como Hadoop, Hive, PIG, WibiData, Platfora y SkyTree, entre otras

## Conclusión

La manipulación de los grandes datos será en el futuro un gran reto para las organizaciones, pero también será un área oportunidad para otras.

Imagínesse compañías como Google, qué no podría hacer con toda la información que generan los usuarios con las búsquedas. Google es capaz de detectar epidemias con base en patrones de búsqueda de internautas en una región. Por otro

lado, imagínesse que se podría recolectar en una sola base de datos toda la información geosísmica del mundo. Pero que a la vez se pudiese procesar en cuestión de segundos. Se podrían detectar patrones para predecir movimientos de la Tierra que podrían, tal vez en un futuro cercano, la predicción de sismos y terremotos. Parece de ciencia ficción, pero todo esto se podría lograr con las nuevas tecnologías que se están desarrollando en el área de big data.

Como hemos mencionado, la forma tradicional para procesar la información de los grandes datos se está volviendo obsoleta, así que nuevas maneras de manipularla están desarrollándose, esto abrirá nuevos paradigmas y nuevo conocimiento en las ciencias de la computación enfocadas en el desarrollo de nuevos algoritmos y formas para almacenar, organizar, buscar, etc., la enorme información que se está generando diariamente.

\*El autor es profesor-investigador de la carrera de Lic. en Ciencias Computacionales de la Facultad de Ciencias, UABC. [evelio@uabc.edu.mx](mailto:evelio@uabc.edu.mx)