Una Introducción al Big Data

Leonardo H. Añez Vladimirovna*

Universidad Autónoma Gabriél René Moreno,
Facultad de Ingeniería en Ciencias de la Computación y Telecomunicaciones,
Santa Cruz de la Sierra, Bolivia

23 de agosto de 2019

1. ¿Que es el Big Data?

Big Data no es un tipo de aplicación específico, sino más bien una tendencia, o incluso una colección de tendencias, que toma múltiples tipos de aplicaciones. Por esto diremos que *Big Data* se refiere a grandes conjuntos de datos complejos, tanto estructurados como no estructurados, en los que las técnicas y / o algoritmos de procesamiento tradicionales no pueden operar. Su objetivo es revelar patrones ocultos y ha llevado a una evolución de un paradigma científico basado en modelos a un paradigma científico basado en datos. Y esto es algo que se puede apreciar en el día a día si nos ponemos a pensar en toda la información que hay en el Internet, en cada sistema y entorno. Tenemos ejemplos como los siguientes:

- Google procesa 20 PB por día
- Wayback Machine tiene 3 PB + 100 TB / mes
- Facebook tiene 2.5 PB de datos de usuario + 15 TB / día
- eBay tiene 6.5 PB de datos de usuario + 50 TB / día
- El Gran Colisionador de Hadrones (LHC) del CERN genera 15 PB al año

1.1. Dimensiones

1.1.1. Volumen

Los datos actuales existentes están en petabytes, que ya es problemático; se predice que en los próximos años es aumentar a zettabytes (ZB) [39]. Esto se debe a un mayor uso del móvil. dispositivos y redes sociales principalmente.

1.1.2. Velocidad

Se refiere tanto a la velocidad a la que se encuentran los datos capturado y la tasa de flujo de datos. Aumentado la fiabilidad de los datos en vivo causa desafíos para análisis tradicional ya que los datos son demasiado grandes y continuamente en movimiento.

1.1.3. Variedad

Como los datos recopilados no son específicos categoría o de una sola fuente, hay numerosos formatos de datos sin procesar, obtenidos de web, textos, sensores, correos electrónicos, etc. que son estructurado o no estructurado Esta gran cantidad hace que los viejos métodos analíticos tradicionales fallen gestionando big data.

1.1.4. Veracidad

La ambigüedad dentro de los datos es la principal enfocarse en esta dimensión, típicamente por ruido y anormalidades dentro de los datos.

*Correo Electrónico: toborochi98@outlook.com

1.2. ¿Cómo lidiar con Big Data?

Existen dos categorías. La primera recoge datos que incluyen informaciones sobre clientes o fechas y se agrupan en tablas aunque, existen otros tipos de datos como imágenes, vídeos o audios que se clasifican como datos no agrupados porque no se pueden clasificar con este método.

La segunda recoge datos más detallados y que están relacionados con actividades comerciales como pueden ser las valoraciones, encuestas, registros de ventas, etc. Los últimos están relacionados con las interacciones sociales como por ejemplo, los datos que aportan las redes sociales.

1.2.1. Clasificación

- **Estructurado**

  Cualquier dato que se pueda almacenar, acceder y procesar en forma de formato fijo se denomina datos 'estructurados'. A lo largo del tiempo, el talento en ciencias de la computación ha logrado un mayor éxito en el desarrollo de técnicas para trabajar con este tipo de datos (donde el formato es bien conocido de antemano) y también para obtener valor de él. Sin embargo, hoy en día, estamos previendo problemas cuando el tamaño de dichos datos aumenta en gran medida, los tamaños típicos están en la rabia de múltiples zettabytes.

- **No Estructurado**

  Cualquier dato con forma o estructura desconocida se clasifica como dato no estructurado. Además de que el tamaño es enorme, los datos no estructurados plantean múltiples desafíos en términos de su procesamiento para obtener valor de ellos. Un ejemplo típico de datos no estructurados es una fuente de datos heterogénea que contiene una combinación de archivos de texto simples, imágenes, videos, etc. Hoy en día las organizaciones tienen una gran cantidad de datos disponibles, pero desafortunadamente, no saben cómo obtener valor de ellos desde entonces. Estos datos están en su forma cruda o formato no estructurado.

- **Medianamente Estructurado**

  Los datos semiestructurados pueden contener ambas formas de datos. Podemos ver datos semiestructurados como una forma estructurada, pero en realidad no está definida con, por ejemplo: Una definición de tabla en DBMS relacional. Un ejemplo de datos semiestructurados son los datos representados en un archivo XML.

1.3. Importancia

La importancia del big data no gira en torno a cuántos datos tiene usted, sino qué hace con ellos. Puede tomar datos de cualquier fuente y analizarlos para hallar respuestas que hagan posibles 1) reducciones de costos, 2) reducciones de tiempo, 3) desarrollo de nuevos productos y soluciones optimizadas, y 4) toma de decisiones inteligente. Cuando se combina el big data con analítica poderosa, se pueden realizar tareas relacionadas con negocios, tales como:

- Determinar las causas de origen de fallos, problemas y defectos casi en tiempo real.
- Generar cupones en el punto de venta basados en los hábitos de compra del cliente.
- Recalcular portafolios de riesgo completos en minutos.
- Detectar conducta fraudulenta antes de que afecte a su organización.

1.3.1. Utilizacion

En el mundo de los negocios, el Big Data proporciona información a las empresas. Les permiten fijar de forma más estratégica sus objetivos, centrando sus acciones en utilizar y sacar partido a las nuevas oportunidades que pueden aparecer entre estos datos e incluso, abandonar aquellos objetivos o estrategias que les resulten más problemáticas. Uno de los campos donde es más útil este sistema es el marketing. El análisis de más variedad de datos permite conocer mejor los gustos o deseos de los clientes y poder ofrecer así productos nuevos que satisfagan sus necesidades.

Principalmente

Para una mejor comprensión de lo que se considera Big Data estos pueden ser principales casos de uso de la tecnología Big Data y que, a día de hoy, otras tecnologías más tradicionales no tienen capacidad de proceso para afrontar dicho caso de uso:

- Web Search Engine (como Google, Baidu, o Bing) que son capaces de ubicar los documentos más populares y relevantes que se encuentra en Internet y que incluyen una o varias palabras por las que se realiza la búsqueda, para después ordenarlos de acuerdo a una serie de criterios (como por ejemplo el número de referencias a dicho documento desde otros documentos).
- Recommendation Systems que a partir de las preferencias personales de un usuario (deducidas a partir de las compras que ha realizado o de aquellos ítems que ha visitado en un portal de eCommerce) o de las preferencias de usuarios "similares" (calculado como los usuarios que les gustan cosas similares), proponen nuevos productos que un usuario quisiera comprar. Un ejemplo lo podéis encontrar en el portal de eCommerce por excelencia como es Amazon.
- Clickstream Analysis que utiliza los datos que producen los usuarios cuando navegan por Internet con el fin de segmentar a los usuarios y entender sus preferencias. Las agencias de medios digitales también pueden analizar los flujos de clicks e impresiones de publicidad para ofrecer anuncios más eficaces.
- Log Processing que analiza un número masivo de registros generados por aplicaciones web y móviles, de tal forma que ayuda a las empresas a convertir petabytes de datos desestructurados o semi-estructurados en información útil acerca de sus aplicaciones o usuarios.