“Big Data”

FBRTL23

Fernando Barranco Rodríguez

03 Enero 2017

Índice

Resumen

La naturaleza de la información hoy es diferente a la información en el pasado. Debido a la abundacia de sensores, micrófonos, cámaras, escáneres médicos, imágenes, etc. en nuestras vidas, los datos generados a partir de estos elementos serán dentro de poco el segmento más grande de toda la información disponible.

El uso de Big Data ha ayudado a los investigadores a descubrir cosas que les podrían haber tomado años en descubrir por si mismos sin el uso de estas herramientas, debido a la velocidad del análisis, es posible que el analista de datos pueda cambiar sus ideas basándose en el resultado obtenido y retrabajar el procedimiento una y otra vez hasta encontrar el verdadero valor al que se está tratando de llegar.

1. ¿Qué es Big Data?

Debido al gran avance que existe día con día en las tecnologías de información, las organizaciones se han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubrir y entender más allá de lo que sus herramientas tradicionales reportan sobre su información, al mismo tiempo que durante los últimos años el gran crecimiento de las aplicaciones disponibles en internet han sido parte importante en las decisiones de negocio de las empresas. El presente artículo tiene como propósito introducir al lector en el concepto de Big Data y describir algunas características de los componentes principales que lo constituyen.

1.1. Introducción

El primer cuestionamiento que posiblemente llegue a su mente en este momento es ¿Qué es Big Data y porqué se ha vuelto tan importante? pues bien, en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos.

1.2. ¿Cuánto es demasiada información para ser procesada utilizando Big Data?

Analicemos primeramente en términos de bytes:

- Gigabyte = 10e9 = 1,000,000,000
- Terabyte = 10e12 = 1,000,000,000,000
- Petabyte = 10e15 = 1,000,000,000,000,000

- Exabyte = 10e18 = 1,000,000,000,000,000,000

1.3. Las 3 "Vs" del Big Data

Además del gran **volumen** de información, esta existe en una gran **variedad** de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la **velocidad** de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data.

1.4. Tipos de Datos

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué información es la que se debe analizar?, sin embargo, el cuestionamiento debería estar enfocado hacia ¿qué problema es el que se está tratando de resolver?

Si bien sabemos que existe una amplia variedad de tipos de datos a analizar, una buena clasificación nos ayudaría a entender mejor su representación.

1. Web and Social Media: Incluye contenido web e información que es obtenida de las redes sociales como `Facebook, Twitter, LinkedIn`, etc, blogs.
2. Machine-to-Machine (M2M): M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.
3. Big Transaction Data: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc.

Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4. Biometrics: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.
5. Human Generated: Las personas generamos diversas cantidades de datos como la información que guarda un call center como llamadas a celular, correos, etc.

1.5. Componentes de una plataforma BD

Las organizaciones han atacado esta problemática desde diferentes ángulos. Todas esas montañas de información han generado un costo potencial al no descubrir el gran valor asociado. Desde luego, el ángulo correcto que actualmente tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto *Hadoop*.

Hadoop está inspirado en el proyecto de *Google File System(GFS)* y en el paradigma de programación MapReduce, el cual consiste en dividir en dos tareas para manipular los datos distribuidos a nodos de un clúster logrando un alto paralelismo en el procesamiento. *Hadoop* está compuesto de tres piezas: *Hadoop Distributed File System (HDFS), Hadoop MapReduce y Hadoop Common.*

2. Internet de las Cosas (IoT)

Internet de las cosas (en inglés, *Internet of things*, abreviado IoT) es un concepto que se refiere a la interconexión digital de objetos cotidianos con internet. Alternativamente, Internet de las cosas es el punto en el tiempo en el que se conectarían a internet más cosas u objetos que personas.

3. Ciencia de Datos

La Ciencia de datos es un campo interdisciplinario que involucra a los procesos y sistemas para extraer conocimiento o un mejor entendimiento de grandes volúmenes de datos en sus diferentes formas (estructurados o no estructurados) y formatos.
Es una continuación de algunos campos de análisis de datos como son: la minería de datos y la analítica predictiva.
La ciencia de datos es un nuevo paradigma sobre el cual los investigadores se apoyan de los sistemas y procesos que son muy diferentes a los utilizados en el pasado, como son modelos, ecuaciones, algoritmos, así como evaluación e interpretación de resultados.

3.1. Ciencia de Datos vs Business Intelligence y Big Data

El término Business Intelligence (BI) también se ha popularizado en nuestros tiempos e incluso, se ha llegado a utilizar de manera indiscriminada con el concepto de ciencia de datos para referirse al análisis de datos, pero en realidad existen diferencias abismales entre dichos conceptos, a continuación se presenta algunas de sus diferencias.

Ciencia de datos:

- Trabaja en datos incompletos.
- Los datos suelen estar desordenados.
- Analiza los datos para ver qué información obtiene.
- Grandes conjuntos de datos que es un desafío administrar.
- Los hallazgos impulsan decisiones sobre operaciones y productos.

Business intelligence (BI):

- Conjuntos de datos completos.
- Archivos de datos limpios.
- Informa lo que dicen los datos.
- Conjunto de datos manejable.
- Sus hallazgos miden el rendimiento pasado.

La sección 3 muestra conceptos básicos enfocados a Business Intelligence

3.2. Científico de Datos

Las personas que se dedican a la ciencia de datos se les conoce como científico de datos, de acuerdo con el proyecto Master in Data Science define al científico de datos como una mezcla de estadísticos, computólogos y pensadores creativos, con las siguientes habilidades:

(i) Recopilar, procesar y extraer valor de las diversas y extensas bases de datos.

(ii) Imaginación para comprender, visualizar y comunicar sus conclusiones a los no científicos de datos.

(iii) Capacidad para crear soluciones basadas en datos que aumentan los beneficios, reducen los costos.

(iv) Los científicos de datos trabajan en todas las industrias y hacen frente a los grandes proyectos de datos en todos los niveles.

[1]

[1]Documento elaborado por: Fernando Barranco Rodríguez