
UNIVERSIDAD LIBRE®

# Introducción a Big Data

**Monografía diplomado Big Data**

**Autor**
**Oscar David Bernal Niño**

Ingeniería de Sistemas

8 de Agosto de 2017

Universidad libre de Colombia

Sede Bosque Popular

# Contenido

# Resumen

Dentro del contexto actual de divulgación y análisis de grandes volúmenes de información, el concepto de Big data cada vez toma más fuerza. Dentro de esta monografía se tratara de manera general el concepto de lo que es Big Data, qué características tiene, como se analiza la información y para que puede servir la implementación de un escenario de Big data dentro de diferentes contextos empresariales.

Teniendo en cuenta el párrafo anterior como premisa, esta monografía no tiene como objetivo puntualizar en un aspecto puntual de Big Data, si no abordar a grandes rasgos todo el concepto de Big data, y definir de manera puntual sus principales componentes, de tal manera que sirva como una introducción para aquellas personas interesadas en iniciar a investigar sobre este tema.

# Introducción

Actualmente, uno de los términos que más auge tiene en el ambiente empresarial es Big Data, este es un concepto que se ha vuelto importante para casi cualquier profesional, independiente su campo de acción. Si bien, puntualmente no se menciona de manera explícita, este concepto, está presente en cosas tan simples como la ruta que las aplicaciones de gps sugieren , las ofertas que llegan en base a las búsquedas de los usuarios en los portales de comercio electrónico, al igual que en temas más trascendentales, como la planeación de ciudades, toma de decisiones financieras , estudios médicos, y en general casi cualquier información que sea resultado del análisis de grades volúmenes de datos.

Esto se traduce en que a pesar que a la mayoría de las competencias de los profesionales encargados de analizar la información obtenida, no estén orientadas a conocer a nivel técnico como se define y se construye un solución de Big Data, cada día se requiere que los profesionales tengan conocimientos de cómo funciona y para qué sirve, ya que en realidad quienes pueden dar valor a una solución de Big Data son los profesionales que hacen uso de la información que da como resultado los diferentes procesos y análisis que se realizan.

Teniendo en cuenta esta premisa, esta monografía está orientada a un público general, más que a un público técnico y especializado en el tema. El objetivo, es que cualquiera que lea este documento entienda a grandes rasgos, el concepto, características y principales usos de Big Data. Por este motivo se utilizara lenguaje común, y se evitaran los conceptos técnicos que sean de interés exclusivo para los profesiones de la tecnología

# Desarrollo

Big Data

La primera impresión que puede dar un nombre simple como lo es Big Data, es tan ambiguo, que no se sabe realmente de que se está hablando, más teniendo en cuenta que es un concepto que puede ser visto desde varias perspectivas diferentes. Ahora el punto es saber que es Big Data.

Si se revisa el concepto desde el punto de vista del volumen de datos, se puede definir Big data como un modelo con la capacidad de almacenar grandes cantidades de información, previendo de manera anticipada el espacio de almacenamiento, y la capacidad de procesamiento de la información almacenada.

Ahora si se ve el concepto de Big Data desde la perspectiva de la velocidad, desde este punto de vista Big Data se define como los procesos necesarios para lidiar con grandes flujos de datos, de tal manera que se puedan obtener resultados en periodos relativamente cortos de tiempo. Básicamente es tener la capacidad de analizar y procesar información una vez llega a una compañía o proceso, en un tiempo que aporte valor al negocio o al proceso.

Otra perspectiva importante desde la cual se puede definir Big Data es por la variedad de los formatos y Datos. Esto se puede ver desde el punto de vista en donde se define a Big Data como un modelo capaz de procesar una gran variedad de datos en múltiples formatos y tipos de datos, como datos estructurados, semiestructurados y sin estructurar. De esta manera se puede ver a un modelo de Big Data, como un modelo flexible, capaz de procesar casi

cualquier tipo de información, siempre y cuando sea compatible con la definición del modelo de Big Data a utilizar.

Ahora se sigue con la veracidad de la información. Big Data revisado desde esta perspectiva es definido como un modelo capaz de identificar información sin valor dentro de todo el universo que se requiere analizar. Dentro de los análisis que se hacen en Big Data, existen datos que pueden ser significativos o que simplemente fueron capturados, pero no aportan valor alguno al análisis de la información. Cuando se evalúan en función de su veracidad, los datos pueden ser de dos tipos: Ruido, que son datos que no tienen valor alguno y Señal, datos que tienen valor y pueden ser utilizados para llegar a información de importancia para el proceso o negocio.

Por ultimo revisaremos el tema del valor de la información. Desde este punto de vista es un modelo, que tiene la capacidad de encontrar y analizar datos de utilidad para una organización. Ahora el motivo por el cual el valor de la información se expone en último lugar, es porque esta característica por sí sola no es nada, sin las demás expuestas.

Para dar más claridad al respecto de por qué el valor de la información no es diciente sin involucrar los demás conceptos expuestos, se va a dar un ejemplo de por qué el valor de la información depende directamente de las demás perspectivas expuestas anteriormente. El primer ejemplo que se va a utilizar es la relación directa con la veracidad de la información, ya que el valor está directamente relacionada con la característica de la veracidad, en la medida en que, entre más alta sea la fidelidad de los datos, mayor será la utilidad de los mismos para la empresa.

Ahora vamos se a exponer la relación con el tiempo de procesamiento, si bien puede parecer que el tiempo del procesamiento no puede afectar el valor de la

información, ya que la calidad de la misma no se está modificando, este resulta ser uno de los aspectos más significativos de la utilidad de la información, ya que el hecho de que la información se pueda procesar de manera rápida hace que esta sea relevante o de utilidad en procesos críticos y toma de decisiones, esto teniendo en cuenta que contar con información de calidad, es tan importante como contar con la información a tiempo.

De hecho las 5 perspectivas que se acaba de exponer de manera corta se conocen como las 5 v de Big Data, aunque en los últimos años se hablan de 7, en donde se agregan características como la visualización y la viabilidad.

En donde la visualización, se refiere como su nombre lo dice, a la manera como se presentaran los datos, para que sean legibles, de tal manera que cualquier persona pueda interpretarlos y entenderlos de forma sencilla sin tener un conocimiento técnico.

En el caso de la viabilidad, para el caso de Big Data, hablamos de la capacidad que tienen las compañías en generar un uso eficaz del gran volumen de datos que manejan.

Ahora después de exponer las diferentes perspectivas desde las cuales se puede definir el Big Data, que es Big Data?. En resumen Big Data es un modelo de procesamiento que permite la recolección y procesamiento de grandes cantidades de información, con el objetivo de encontrar patrones, información de utilidad, correlaciones, filtrar la información de manera óptima, de tal manera que los datos puedan ser procesados y analizados en la medida que lleguen al proceso, siempre y cuando los métodos convencionales no sean suficientes para analizar todo el volumen de información.

Teniendo como base el concepto general de Big Data , el paso siguiente es saber cómo funciona. Para saber cómo funciona se explicara de manera breve los conceptos referentes a análisis, analítica, minería de datos, machine learning e inteligencia de negocios (BI). Que en ultimas estos son los procesos que se ejecutan dentro de una solución de Big Data.

**Análisis**

Dentro del contexto de Big Data, el análisis es el proceso de la examinación de los datos, con el objetivo de hallar hechos, relaciones, patrones, explicaciones y tendencias. Básicamente el análisis dentro de Big Data busca encontrar información suficiente para argumentar y respaldar las decisiones sobre un negocio o proceso. Los análisis dentro de Big Data se pueden clasificar en dos tipos, cualitativo y cuantitativo.

Cuando hablamos de un análisis cuantitativo, nos referimos a las técnicas que tienen como objetivo cuantificar patrones y correlaciones hallados en los datos, esto implica que los resultados que se obtienen de estos análisis son utilizados para realizar comparaciones numéricas, dado que los resultados de este tipo de análisis son de naturaleza absoluta.

Ahora un análisis cualitativo esto orientado a describir cualidades de la información analizada en leguaje cotidiano. Este tipo de análisis, a diferencia de un análisis cuantitativo no se puede aplicar de forma general a todo un data set, sino que este tipo de análisis está orientado a analizar una muestra del data set a mayor profundidad de manera independiente. Esto teniendo en cuenta que debido al tipo de información que arroja una análisis cualitativo tampoco puede ser usado para realizar comparaciones estadísticas.

## Analítica

A diferencia del análisis la analítica se encarga de comprender la información que se está analizando, y en el contexto de Big Data, la analítica se enfoca a hallar patrones y correlaciones entre los datos. La analística, dentro de los procesos de big data, puede ser descriptiva, diagnostica, predictiva y prescriptiva.

Donde la analítica descriptiva es la que se encarga de comprender y responder preguntas sobre eventos que ya ocurrieron.

La analítica diagnostica, al igual que la descriptiva se basa en eventos que ya ocurrieron, pero diferencia de esta, su objetivo principal es explicar el origen de un fenómeno, más que describirlo.

La analítica predictiva, se centra en base a la información actual y a los diferentes patrones encontrados, en determinar el posible resultado de un fenómeno especifico.

Por último la analítica prescriptiva usa los resultados de la analítica predictiva, para sugerir acciones a seguir.

## Minería de Datos

Teniendo en cuenta que Big Data está orientado al análisis de grandes volúmenes de información, es necesario adoptar técnicas de análisis orientadas a explorar grandes bases de datos.

Este tipo de análisis permiten encontrar patrones en la información, en donde la minería de datos se define, simplemente como un conjunto de estrategias que

apoyadas en la tecnología nos permiten analizar bases de datos de grandes tamaños.
Como toda estrategia tiene un orden y unas características propias a los demás análisis, para el caso de la minería de datos las estrategias que tienen sus orígenes en la estadística y en la inteligencia artificial.

Dentro de las estadísticas de minería de datos encontramos redes neuronales, regresión lineal, árboles de decisión, modelos estadísticos Clustering, reglas de asociación, entre otras.

Por ultimo para finalizar este pequeño segmento de minería de datos es de utilidad mencionar que independientemente los procesos y estrategias para ejecutar la minería de datos, siempre se tienen una etapas definidas que permiten la correcta ejecución de un proceso de minería, estas etapas son: la determinación de los objetivos, el procesamiento de los datos, la determinación del modelo y el análisis de los resultados.

**Machine Learning**

Como tal machine learning puede ser definido como la disciplina o ámbito dentro de la inteligencia artificial, que se encarga de generar y desarrollar algoritmos y sistemas , que en base a la identificación de patrones y características permiten a un sistema aprender a procesar diferentes tipos de información, básicamente es la ciencia que se encarga de crear sistemas que aprenden de manera autónoma.

## Para que sirve Big Data

Ahora que se tiene un entendimiento de que es big data, y de que se compone. El siguiente paso es saber para que sirve, si es viable, cuando aplica y si es oportuno. En el mundo actual pensar en grandes volúmenes de información, es tan cotidiano, que en un minuto se están generando más de 8000 GB de información [1]. Esto indica que los mecanismos actuales de análisis de información, ya nos son los más indicados, Si, se tienen soluciones BI, pero estas requieren datos completamente estructurados, también se tienen las soluciones de minería de datos, pero por si sola no aporta valor a la información. Big Data surge, no como un modelo que intenta reemplazar todas estos modelos establecidos, si no como una solución que las integra, y las orquesta de tal manera que trabajen de manera conjunta.


1. Información generada en el mundo en GB en un minuto, tomado desde http://intal-interactivo.iadb.org/?p=804, Instituto para la integración de américa y el caribe [2]

Dentro de las aplicaciones más comunes del big data está el análisis de redes sociales. En el entorno actual y la evolución de la manera como la información es capturada y compartida por las personas, las redes sociales se han convertido en uno de los principales insumos de información, para cualquier análisis de los hábitos y preferencias de las personas, es fácil de evidenciar con las cifras que se mostraran a continuación

[1] (Instituto para la Integración de América Latina y el Caribe (INTAL), 2016)

- Para entrar en contexto, en marzo de 2016 la población mundial era de 7,4 mil millones (United Nations, 2016)[3]
- Internet tiene 3,17 mil millones de usuarios [4]
- Hay 2,3 mil millones de usuarios activos en redes sociales (Kemp, 2016)[5]
- El 91% de las marcas de retail usan dos o más canales de redes sociales (Morrison, 2016)[6]
- Los usuarios de internet tienen 5,54 cuentas en redes sociales de promedio (Mander, 2016)[7]
- Los usuarios de redes sociales crecieron 176 millones el año pasado (Redan, 2015)[8]
- Hay 1 millón de usuarios activos de redes sociales en móviles nuevos cada día. Es decir, 12 cada segundo (Redan, 2015) [9]
- Facebook Messenger y Whatsapp manejan 60 mil millones de mensajes diarios (Tynan, 2017)[10]
-

Ahora con esta información es más que lógico pensar que en el mundo actual, cualquier estudio que se desee realizar, para saber los gustos de la gente, las reacciones hacia un tema, posibles hábitos de consumo, validación de perfiles profesionales, y en general cualquier análisis que tenga que ver con el comportamiento humano a gran escala, va a ser un candidato muy válido para aplicar big data.

Ahora porque es un buen candidato para aplicar big data ¿? , en el caso de un análisis de redes sociales se pueden observar fácilmente varias características de

[3] (United Nations, 2016)
[4] (https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/, 2016)
[5] (Kemp, 2016)
[6] (Morrison, 2016)
[7] (Mander, 2016)
[8] (Redan, 2015)
[9] (Redan, 2015)
[10] (Tynan, 2017)

los modelos de big data, en donde para ilustrar de mejor manera vamos a dar un ejemplo de como las características de la información son buenos candidatos para aplicar big data:

- Volumen de datos: como ya se expuso previamente en el documento se va a hacer un análisis teniendo como insumo los mensajes que se envían por medio de Facebook y WhatsApp en la última semana, y se toma como ejemplo la cifra anteriormente expuesta de 60 mil millones de mensajes diarios, y lo multiplicaremos por 7 días, lo cual nos da como resultado un dataset de 420 mil millones de mensajes analizar, y si adicionalmente le damos un valor arbitrario de 0,5 mb por mensaje[11], esto nos deja con un volumen de información de 195,577 petabypte, lo cual ya no es un volumen que se pueda almacenar haciendo uso de los mecanismos convencionales. Esto indica que al menos por ahora un esquema convencional no va a poder alojar toda la información que tenemos para este dataset.
- Variedad de datos, si bien para el caso de los mensajes en las redes sociales ya tenemos una porción de la información de manera estructurada, como el origen, fecha de envió y el destinatario, como tal el contenido del mensaje no se encuentra estructurado, para este caso se hace necesario utilizar soluciones que nos permitan identificar si este mensaje habla de un tema específico, si son saludos rutinarios que no nos aportan valor , cuántos de estos tienen una correlación, que patrones repetitivos se pueden encontrar, todo depende del modelo y el objetivo que se tenga al generar el modelo de Big Data.
- Veracidad y valor de la información: para el caso de las redes sociales uno de los puntos críticos es darle valor a la información, aumentando la veracidad de esta, esto se logra generando algoritmos que de manera oportuna logren identificar el ruido dentro de la señal de datos, al español

[11] Este valor es aproximado, y solo se usa como una muestra.

normal, identificar información sin valor para el estudio que estemos realizando.

- Velocidad, por ultimo nos queda el hecho de que esta información debe ser analizada de tal manera, que los resultados sean entregados de manera oportuna. Para dar un ejemplo: vamos a utilizar el data set que tenemos actualmente, para predecir cuantas personas están interesadas en asistir a la visita del papa en la ciudad de Bogotá, vamos a partir de la premisa que estamos a una semana del evento, y tomamos los datos de la semana inmediatamente anterior, también sabemos que el tamaño del data set es de 195,577 petabytes. Para este caso la entrega de información debe ser casi inmediata, si tenemos una solución que tarda más de una en entregar resultados, la información que nos entregue el proceso ya no va a tener valor, por lo cual aquí entra claramente el aspecto de la velocidad del procesamiento.

Como tal se puede seguir desglosando este ejemplo hasta encontrar más razones y características que lo hacen un buen candidato para aplicar big data, y mirado desde varios puntos de vista, objetivos de estudio e intervalos de tiempo a analizar solo harían que se entrara en un ciclo sin fin, por lo cual se van a exponer diferentes contextos en donde puede aplicar big Data, en vez de seguir con solo una de sus aplicaciones.

Ahora se va a exponer el consumo en general desde el punto de vista del Big data. Cuantas veces cuando se busca algo tan simple como unos tenis nuevos, o los tiquetes de las próximas vacaciones, dentro de la publicidad que aparece dentro del correo, o de páginas completamente ajenas al tema se ven promociones y ofertas similares a las observadas. Incluso actualmente uno de los mejores ejemplos de esta tendencia al utilizar big data es Amazon, quien a parte de evaluar que productos se pueden vender mejor, ya están entrando en el campo de la analítica predictiva con aplicaciones que predicen los hábitos de

consumo de sus usuarios, lo que les permite generar anuncios personalizados, que ofrecen al usuario el producto que está buscando en un momento determinado (Sosa, 2016)[12]

Habitualmente las personas en general, constantemente están alimentado modelos de Big Data y aprovechando los mismos, en la mayoría de los casos sin darse cuenta de manera directa. Para este caso vamos a dar dos ejemplos, unos de movilidad, y otro de deportes y acondicionamiento físico.

Para el primer se utilizara las aplicaciones de movilidad y tráfico. Estas aplicaciones manejan el big data en tres aspectos importantes que son, la ruta optima, la velocidad de la ruta, y los puntos de interés. Estas aplicaciones constantemente están capturando las rutas más usadas por las personas, desvíos que toman de la ruta sugerida, puntos de demora, adicionalmente permiten al usuario dar su retroalimentación en tiempo real, lo que les permite capturar suficiente información para poder generar en tiempo real nuevas rutas, que de manera predictiva le van a indicar al usuario cual es la mejor ruta para poder seguir, en este punto es donde entrar los entornos de Big Data, ya que estos son los que permiten procesar la información de tal manera que se pueden identificar los patrones de comportamiento en tiempo real, y en base a la información capturada de la ciudad y de las preferencias del usuario, arrojar un resultado que aporte valor de uso. Hay que aclarar que cada vez estas aplicaciones se encuentran menos aisladas de otros modelos de big data, por lo cual no es raro, que si la persona ha estado haciendo búsquedas de productos específicos en la red, la aplicación de movilidad resalte en el mapa los sitios en donde puede encontrar el producto, y sugiera un cambio de ruta en función de aprovechar más la información del usuario. (Ottolini, 2015)[13]

[12] (Sosa, 2016)
[13] (Ottolini, 2015)

El segundo ejemplo se va a citar del Big Data en la vida cotidiana de las personas, son las aplicaciones de ejercicio, para este caso usaremos s health, el cual es un producto que incluye Samsung dentro de sus dispositivos, que utiliza los registros de entrenamiento de sus usuarios para analizarlos de manera masiva con toda la población, y de esta manera motivar a los usuarios, comparándolos contra sus contactos cercanos y grupo demográfico, generando retos y metas en base a los datos capturados de manera individual cursados con los datos obtenidos del grupo demográfico de la persona. (Samsung, 2016)[14]

[14] (Samsung, 2016)

# Conclusiones

En general este documento expuso de manera global y breve los principales conceptos de Big Data para dar una introducción al concepto y a sus usos, como se expuso al principio del documento el objetivo no es de puntualizar ni profundizar en cada uno de ellos, ni hacer un hincapié en los términos técnicos a utilizar. Si no tener un concepto general, que funcione a nivel general para cualquier tipo de lector.

# Bibliografía


Flach, P. (2012). *Machine Learning.* University of Bristol: Cambridge.

https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/. (07 de 09 de 2016). *The Statistics Portal*. Recuperado el 05 de 09 de 2016, de The Statistics Portal: https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Instituto para la Integración de América Latina y el Caribe (INTAL). (04 de 11 de 2016). *iadb*. Recuperado el 04 de 09 de 2017, de iadb: http://intal-interactivo.iadb.org/?p=804

Kemp, S. (2016). *Digital in 2016 .* We are Social.

Mander, J. (2016). Internet users have average of 5.54 social media accounts. *global web index*, 1.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill Science.

Morrison, K. (2016). 91% of Retail Brands Use Two or More Social Media Channels. *adweek*, 50.

*Number of internet users worldwide from 2005 to 2016 (in millions)*. (08 de 09 de 2016). Recuperado el 05 de 09 de 2017, de The Statistics Portal: https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Ottolini, M. (2015). How One Mobile App Uses Big Data To Detect Traffic. *CRN*, http://www.crn.com/news/applications-os/video/300078187/how-one-mobile-app-uses-big-data-to-detect-traffic.htm.

Redan, K. (09 de 07 de 2015). *Social Media Today.* Obtenido de Social Media Today: http://www.socialmediatoday.com/social-networks/kadie-regan/2015-08-10/10-amazing-social-media-growth-stats-2015

Samsung. (12 de 04 de 2016). *S Health and Samsung Digital Health SDK*. Obtenido de S Health and Samsung Digital Health SDK: http://developer.samsung.com/tech-insights/health/shealth-and-samsung-digital-health-sdk

Sosa, J. N. (2016). Cómo Amazon Usa Big Data para Predecir Tu Próxima Compra. *Capabilia*, 1.

Tynan, D. (27 de 06 de 2017). *Facebook's journey ‘only 1% done’ after surge in revenue, Zuckerberg says* . Obtenido de Facebook's journey ‘only 1% done’ after surge in revenue, Zuckerberg says : https://www.theguardian.com/technology/2016/jul/27/facebook-ad-sales-growth-quarterly-results

United Nations. (08 de 09 de 2016). *http://www.un.org*. Recuperado el 06 de 09 de 2017, de http://www.un.org: https://esa.un.org/unpd/wpp/DataQuery/