Introducción a big data

Origen y definición

1.1 ¿Qué es big data y qué representa en realidad?

El término *big data* se ha puesto de moda en los últimos años -y no es para menos- ya que durante los inicios del siglo XXI se inició una revolución de datos a nivel mundial, aumentando no solo su tamaño sino también su complejidad. Esto llevó a que las necesidades de almacenamiento, limpieza, manipulación, uso, análisis y segmentación tomaran una especial relevancia, y la tecnología necesaria debía readaptarse a estos nuevos requerimientos tanto en hardware como en software.

Por estas mismas razones, el diccionario Oxford de la lengua inglesa agregó oficialmente la palabra “big data” como un término avalado para el idioma inglés en 2013, pero para que esto sucediera no solo fue necesario que este término ya fuera comúnmente utilizado, sino que su raíz, su comprensión y su uso fueron resultado de muchos años de maduración.

Este curso pretende adentrarse en el entendimiento de big data como una herramienta semántica para hablar del uso de grandes volúmenes de información que dan gran valor a su incorporación en el día a día de un profesional, y abordar también sus orígenes y su contexto. Así mismo es necesario desmitificar cualquier uso y acepción incorrecta, muy comunes hoy en día.

Es importante partir de una premisa fundamental: los datos existen, y en gran cantidad. La principal pregunta que se debe responder en un proceso educativo es: ¿cómo se hace para extraer y generar valor de los datos para resolver problemas reales? Esta es la pregunta a la que se intentará dar respuesta.

1.1.1 Definición de big data

El diccionario Oxford de lengua inglesa define “big data” como “Sets de datos extremadamente grandes que deben ser analizados computacionalmente para revelar patrones, tendencias y asociaciones, especialmente relacionadas al comportamiento humano y a las interacciones” (Oxford English Dictionary, 2013).

Esta definición tiene implicancias muy profundas que no solo son aplicables a la semántica de un concepto, sino que también otorgan un norte al cual apuntar cuando se trabaja sobre proyectos o programas de big data.

Para abordar en detalle los aspectos asociados a esta definición, se desglosará y se ampliará cada punto para darle un uso real y un contexto específico acorde:

 “Sets de datos extremadamente grandes”: se refiere a bases de datos estructuradas y no estructuradas con características muy particulares. Una de esas características es el volumen, es decir, big data en rangos superiores a 1TB (terabyte) de información por data set. Más adelante, se entrará en detalle sobre este punto y sus condicionantes.

 “Que deben ser analizados computacionalmente”: en esta parte de la definición, existe una gran verdad desde el punto de vista de la productividad. Si se tratara de analizar estos data sets extremadamente grandes de forma manual, se tardaría demasiado tiempo, probablemente no se podría llegar a las mismas conclusiones, ni extraer suficiente conocimiento de estos datos, sin mencionar que tampoco sería económicamente viable. Al realizar el análisis de datos con herramientas tecnológicas de software y de hardware preparados para esta tarea, no solo se logra mejorar la productividad, sino también se tiene la capacidad de generar valor donde antes no se tenía la habilidad de explotarlo.

 “Para revelar patrones, tendencias y asociaciones”: esta parte revela gran parte del conocimiento que es necesario incorporar a la hora de analizar datos: el conocimiento estadístico y matemático. Es ineludible que todo analista de datos debe conocer de estadística y de matemática, al menos lo suficiente para aplicarlo a resolver los problemas sobre los que necesita trabajar. Al trabajar con una sólida base estadística, es posible llegar a revelar este conocimiento anteriormente oculto a simple vista. Se utiliza la palabra “revelar” porque este conocimiento, estas tendencias, estos patrones, y demás elementos, ya existen dentro de los sets de datos con los que se trabaja, pero a simple vista no es tan sencillo reconocerlos. Es allí cuando es necesario aplicar técnicas de big data con una fuerte impronta de estadística para que dichas tendencias sean reconocibles y, lo más importante de todo, que sean aplicables. A este mismo nivel, también se habla de “asociaciones” puesto que una parte importante del análisis estadístico sobre los datos es también la capacidad de segmentar y de predecir. Trabajado en profundidad, puede generar un impacto sumamente fuerte en cualquier institución.

 “Especialmente relacionadas al comportamiento humano y a las interacciones”: aquí se expresa de forma demasiado simple el proceso. Al hablar de big data no solo se están evaluando formas en que los humanos interactúan, tanto presencial como virtualmente. Así mismo, es necesario entender la interacción entre distintos individuos y las variables que entran en juego para relacionarlos, darles un sentido y explicar una respuesta a la pregunta según el tipo de datos que se esté analizando. En este sentido, la definición oficial queda algo estrecha, ya que al hablar de comportamiento humano y de interacciones se está omitiendo una aplicación sumamente vital que se desarrolla con fuerza que es la interacción y el comportamiento no-humano de objetos y de aparatos que tengan la facultad de decidir por sí mismos. Todo lo relacionado a inteligencia artificial y a Internet de las cosas (IoT, por sus siglas en inglés) plantea escenarios donde los humanos no necesariamente deben tomar una decisión para que las interacciones o los comportamientos sean ejecutados. Para eso, se abordarán más adelante temas como estos desde el punto de vista de Machine Learning y el aprendizaje automático que dichos análisis “computacionales” pueden generar.

Estos nuevos sentidos encontrados en cada elemento de la definición permiten visualizar un panorama de lo que vendrá a partir de trabajar con técnicas, insumos y estructuras de big data. Usar big data para resolver problemas complejos no solo es posible, sino que también es perfectamente realizable y relativamente fácil de comenzar a aplicar si se cuenta con una serie de conocimientos específicos que permitan sacar provecho de esta disciplina. Para ello, también es necesario conocer qué es y qué no es, y el origen que le dio forma a todos estos elementos.

1.1.2 Lo que no es big data

El uso de este término se ha difundido tanto en los últimos años que incluso se ha desvirtuado su uso aplicándolo a temas que nada tienen que ver con lo que en realidad es una estructura, una técnica o un elemento de big data. Parte de estos malos usos del término son generados por el desconocimiento de lo que es y no es big data y esto mismo ha generado que el uso indiscriminado

de esta frase confunda al público en lo que realmente demanda este tipo de requerimientos.

De hecho, la revista Forbes publicó una gran columna de opinión llamada “5 mitos masivos sobre ‘big data’ que la mayoría de las personas creen y no deberían” (Marr, 2017) en donde se expresa perfectamente estos puntos. A continuación, se recuperarán algunos aportes para comentarlos.

- **Mito 1: todo el mundo lo está haciendo.**
  - **Realidad:** esto no es realmente cierto o, al menos, no aún. Como cualquier tendencia, muchos profesionales y compañías están llevando adelante esfuerzos de big data para resolver problemas complejos o tan solo para manejar información con distintos niveles de volumen y de capacidades. Pero como ocurre con todo lo que es tendencia o relativamente nuevo, no todo el mundo lo está trabajando y no siempre lo hacen de la misma manera. Para generar valor a partir del análisis de datos, no es obligatorio trabajar de esta manera, puesto que no todos los análisis de datos son big data. En gran medida, muchos profesionales trabajan con small data o medium data, puesto que trabajar con “big” es en sí mismo un esfuerzo importante que requiere también herramientas y capacidades tecnológicas particulares. Lo que es un mito es que cualquier análisis de datos puede ser llamado big data. De esta misma forma, el sentido de urgencia sobre no quedarse atrás puede ser un buen motivador para ponerse en tema, pero tampoco se debe apartar la vista del objetivo de negocio que se desea cumplir o el elemento práctico que se quiere generar para que no se convierta en un esfuerzo infructuoso o económicamente inviable. Es posible sacar mucho valor del análisis de datos, aun cuando no se llegue al nivel del big data.

- **Mito 2: big data es todo acerca del tamaño.**
  - **Realidad:** si por tamaño se entiende el volumen de datos implicados en los data sets trabajados, se está ante una visión muy parcial de lo que big data realmente representa. Esta es apenas una de las características del big data, las cuales se verán más adelante. Otros elementos como la velocidad de procesamiento, la variedad del tipo de datos o los elementos no estructurados también hacen del big data una herramienta poderosa y útil para revelar

patrones, tendencias y asociaciones. Incluso pensar en big data únicamente en términos del volumen, puede producir ineficiencias derivadas de datos que no son relevantes por su antigüedad, por su poco nivel de verificación o por su número limitado de fuentes, así como también incurrir en costos innecesarios relacionados a la captura, el almacenamiento y el procesamiento.

 Mito 3: big data dirá lo que pasará después.

o **Realidad:** esto tampoco es necesariamente cierto. La realidad es que los datos por sí solos no sirven para nada, sirven cuando se los utiliza para resolver un problema o responder a preguntas concretas, por lo que al hacer análisis predictivo es necesario tener una buena base estadística que amplíe el panorama hacia el futuro y no se quede solo con una proyección simple. Hacer predicciones se basa más en la extrapolación de lo sucedido en el pasado y, con esto, se trata de determinar lo que es más probable que ocurra. En análisis predictivo, se trabaja con probabilidades, no con certezas.

 Mito 4: trabajar con big data necesita grandes presupuestos.

o **Realidad:** si bien las grandes compañías y los gobiernos pueden estar invirtiendo fuertemente comprando infraestructura millonaria, comprando licenciamientos de software muy onerosos o contratando científicos de datos que suelen tener sueldos superiores a la media en cada país, esto no necesariamente implica que si no se cuenta con estos recursos se queda fuera del juego. Cada día se hace más barato trabajar con grandes cúmulos de datos. Aparecen más herramientas y más formas de almacenar y manipular los datos para hacerlos más baratos o, incluso, algunas de licenciamiento libre. Si bien esto no significa que se puede trabajar casi gratis con esta temática, es necesario pensar que es posible y alcanzable, y que el retorno de inversión suele suceder en un corto plazo si el proyecto se implementa apropiadamente, pensando en mejorar los resultados del negocio. Así mismo, es necesario considerar que mientras todavía se está pensando si es conveniente utilizar los datos disponibles para tomar decisiones, los competidores podrían ya estar ejecutando acciones de este tipo, y quedar relegado o fuera del mercado es aún más caro para cualquier compañía.

 Mito 5: big data es algo importante solo para las áreas de tecnología.

o **Realidad:** si se piensa en el inicio de las computadoras hace décadas, se recordará que en ese entonces las compañías que tenían alguna computadora la tenían dentro del área de sistemas, quienes eran los encargados de mantenerla. Posteriormente, a medida que las computadoras se volvieron más baratas y fáciles de utilizar, pudieron llegar a estar al alcance de toda la organización e incluso las personas en sus casas o hasta en sus bolsillos ya tienen computadoras que son parte de su vida cotidiana. El mismo principio aplica en el caso de big data, puesto que a medida que se vuelve cada vez más accesible, más aplicable y más económica, más personas y compañías la utilizarán como una herramienta cotidiana para la toma de decisiones. Esto ya está sucediendo y no tiene por qué ser algo privativo de las personas que saben de tecnología. Hoy en día, disciplinas como Analytics están siendo aprendidas y ejecutadas por profesionales de todo tipo para mejorar los resultados de su trabajo en cualquier área.

Así como estos, existe una gran cantidad de mitos alrededor del tema big data, por lo que es mejor desmentir las ideas erróneas y los conceptos confusos desde el principio.

Ahora que ya se ha definido lo que no es, se debe comenzar a entender lo que sí es, tal como se explica a continuación.

1.1.3 ¿En qué consiste big data?

Las técnicas de big data permiten mejorar los resultados de toda pregunta o problema que se quiera resolver, basado en el análisis de datos pertinentes para la premisa planteada. Como se ha desarrollado en las secciones anteriores, no todo es big data y tampoco es cien por ciento necesario utilizarlo para realizar un análisis de datos. Hay mucho que puede ser analizado incluso con small data, si estos datos son bien utilizados y explotados.

Utilizar big data comprende la aplicación de tres áreas muy particulares que son características intrínsecas de esta disciplina:

1. Volumen

2. Velocidad

3. Variedad

En base a estos tres puntos, se extiende una gran cantidad de puntos medios que hacen al buen uso de los datos para que sean útiles para la vida y los problemas reales. Estos puntos serán profundizados más adelante, pero al menos es necesario comenzar por comprender que cada parte del análisis de datos estará enmarcado en alguno de estos puntos.

Por esto mismo, la importancia del big data radica no solo en su volumen, en la velocidad de generación o de manipulación de los datos o en la variedad de la estructura -o no estructura- de los mismos, sino de la conjunción de todo esto en una misma situación, la cual es la generadora de las necesidades de ejecución de proyectos de big data.

Si bien no siempre se trabaja de esta manera, aún en proyectos con small data es posible utilizar técnicas de big data para influir en los resultados a obtener y, con ello, generar valor en aquello que se busca resolver.

De esta manera, es necesario entender de dónde viene todo este proceso ya que cada paso que se dio desde esta disciplina marcó un rumbo hasta el día de hoy y seguirá revolucionando la forma de analizar la información.

1.1.4 Origen del big data

Hablar de big data se volvió un tema de moda en los últimos años, aunque este tema ha sido un elemento relevante desde hace mucho tiempo. Desde que se comenzaron a realizar teorías de lo que se denominaba “La Explosión de los datos” en 1944 hasta lo que puede leerse en artículos de investigación o puede verse en ejecución de proyectos de datos en cualquier emprendimiento de este tipo, se hace notar que los avances se han ido reciclando constantemente para pulirse y para mejorarse en la forma de técnicas y de infraestructura más eficientes y también poco a poco más accesibles económicamente.

Para hacer una breve reseña de la historia del big data como concepto aplicable, se utilizará como base el compilado realizado por Ramesh Dontha

(2017) a través de una publicación en el sitio de especialidad *KDNuggets*, en donde se elabora una síntesis simple pero completa de los aspectos más relevantes y los hitos de esta disciplina que la convirtieron en lo que hoy se conoce.

Desde 1944, se comenzó a debatir sobre la importancia de la “Explosión de la Información” a partir de una serie de especulaciones realizadas en la Universidad de Wesleyan por Fremont Ryder donde predecía que en la Universidad de Yale, para el año 2040, existirían más de 200 millones de volúmenes debido a que en los siguientes años se realizarían tantas publicaciones o se generaría tanta información que se necesitaría de un gran espacio para almacenar tantos libros. Entonces, se advirtió la necesidad de encontrar un método más eficiente para hacer que todo ese conocimiento estuviera a disposición de la Humanidad.

Claramente, no fue necesario esperar hasta 2040 para ver cómo la “Explosión de los Datos” tomaba carácter propio y concreto.

Por su parte, en 1980, fue Charles Tilly en Inglaterra la primera persona conocida que utilizó el término “big data” en una publicación académica de prestigio, ya que desde Oxford se comenzó a hablar de todos los datos que se generaría en la “era de la computación”.

En los años ’90, se hizo más común el término gracias a distintos académicos que se apoyaban en la fuerza que habían adquirido las computadoras y los servidores a nivel mundial, aunque en esa época todavía era bastante costoso el almacenamiento de la información.

En 2001, Doug Laney comienza a utilizar el término “big data” y lo asocia a las famosas 3 V´s que mencionamos anteriormente (Volumen, Velocidad y Variedad), aunque fue hasta 2005 que Tom O´Reilly publica el libro *¿Qué es la Web 2.0?*, y es allí donde se comienza a utilizar el término big data de la forma en que se usa actualmente.

En paralelo a la difusión de este libro, también en 2005, se crea Hadoop por parte de ingenieros de Yahoo como una respuesta al uso que Google comenzó a hacer de MapReduce para la indexación ágil de millones de sitios web en su buscador.

En 2008, Google rompe la barrera de 20 petabytes procesados por día y se intensifica la batalla por la generación, la manipulación, el almacenamiento y el uso del big data con las acepciones actuales y las implicancias modernas.

En adelante, crece cada vez más la adopción de herramientas de Analytics, Business Intelligence y disciplinas similares que intensifican el uso de los datos de las compañías y de las instituciones para mejorar la toma de decisiones y para hacer más eficientes y económicamente viables sus esfuerzos de sistematización de la información.

Esto lleva a preguntarse si, en los próximos años, se logrará hacer uso de los datos para obtener conocimiento y para descubrir patrones y tendencias, y para segmentar elementos del pasado o para predecir el futuro, y cómo se hará.

Big Data
Desde los inicios hasta hoy

Memoria Virtual
(Fritz-Rudolf Güntsch)
Concepto desarrollado por este físico alemán como una idea que trataba el almacenamiento finito como infinito y permitía procesar datos sin las limitaciones de memoria del hardware.

1956

Ley bibliométrica del aumento exponencial (Derek Price)
Observa el aumento exponencial (x2 cada 15 años y x10 cada 50 años) del número de publicaciones científicas y artículos. Con esta ley explica que "cada avance [científico] genera una nueva serie de progresos a una tasa de natalidad razonablemente constante, de modo que el número de nacimientos es estrictamente proporcional al tamaño de la población de descubrimientos en cualquier momento".

1961

Reconocimiento de voz
(William C. Dersch)
Presenta "Shoebox" la primera máquina en comprender 16 palabras y 10 dígitos en inglés hablado mediante el uso de los datos disponibles en ese momento, y la capacidad de procesarlo de manera eficiente.

1962

La Era de la Automatización
Debido a la afluencia de información en los años 60, las organizaciones comienzan a diseñar, desarrollar e implementar sistemas de computación centralizados que les permiten automatizar sus sistemas de inventario.

1966

El auge de la comunicación bidireccional
El Censo de flujo de información de Japón, comenzó a rastrear el volumen de información. Con el número de palabras utilizadas como unidad de medida en los medios, concluyó que la demanda de comunicación unidireccional se había estancado. Sin embargo, aumentaba la demanda de comunicación bidireccional y más personalizada.

1975

Ley de Parkinson
(I A. Tjomsland)
En su charla, *"Where Do We Go From Here?"* dijo que la primera Ley de Parkinson se puede parafrasear para describir la industria: "Los datos se expanden hasta llenar el espacio disponible para el almacenamiento".

1980

La fundación de la World Wide Web
(Tim Berners-Lee)

1989

El primer informe de Bases de Datos
Crystal Reports crea la primera base de datos simple con Windows, lo que facilita a las empresas el trabajo. De esta forma, comprar más memoria incentiva el uso de técnicas de programación que usan la memoria de forma más intensiva.

1992

La World Wide Web explota
La década de 1990 fue un momento de crecimiento explosivo para la tecnología y los datos de Business Intelligence comenzaron a acumularse en forma de documentos de Microsoft Excel.

1995

El problema del Big Data
(Michael Cox & David Ellsworth)
El término "Big Data" fue utilizado por primera vez en un artículo de estos investigadores de la NASA donde afirman que "el aumento de los datos se estaba convirtiendo en un problema para los sistemas informáticos actuales". Esto también se conoce como el "problema del Big Data".

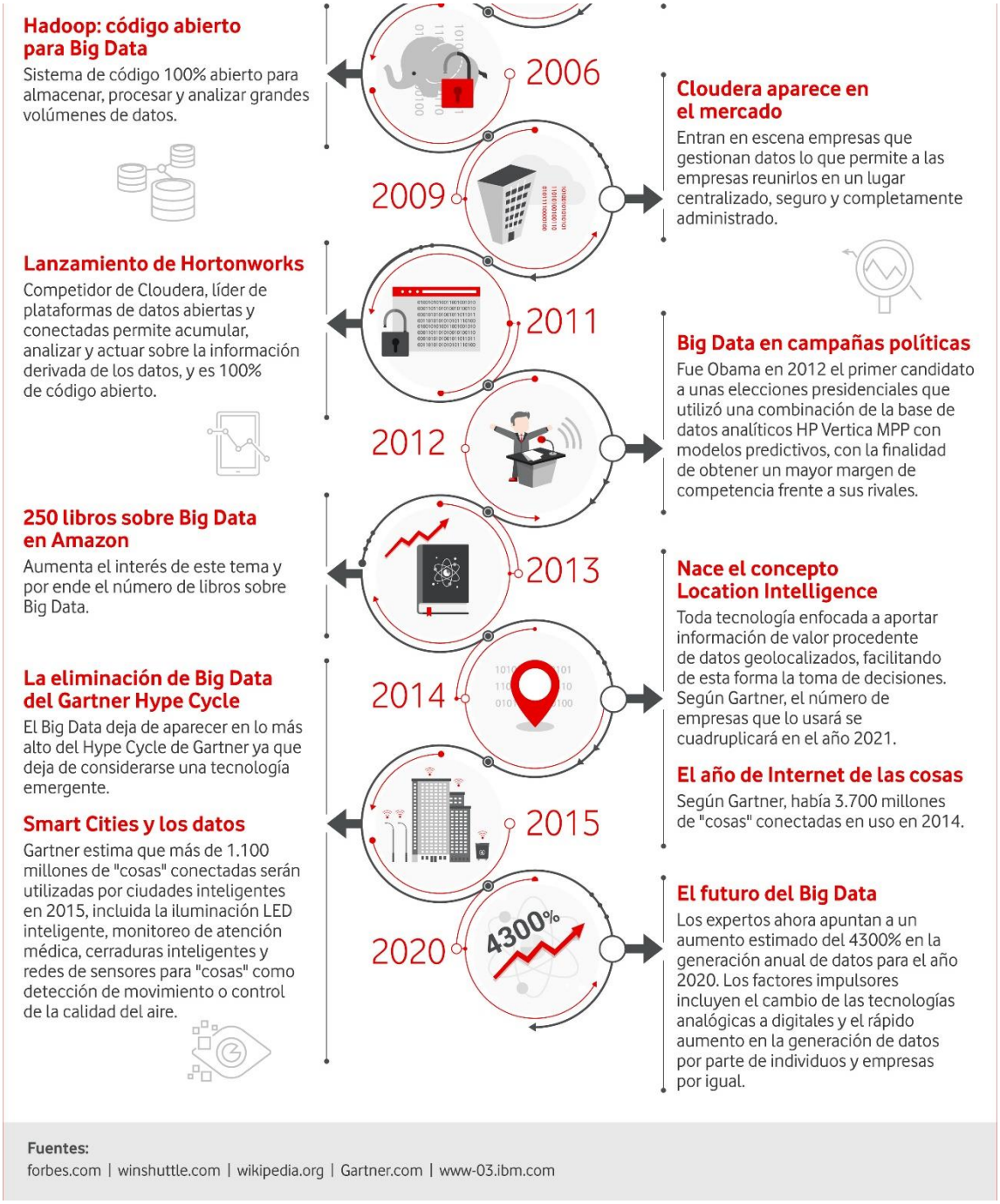
1997

Las tres V de Big Data de Gartner (Doug Laney)
Laney publicó publicó un trabajo de investigación titulado *3D Data Management: Controlling Data Volume, Velocity, and Variety* de donde se salieron las *"3V"* aceptadas del Big Data.

2001

Google publica GFS y MAPREDUCE
Google publica en 2003 y 2004 las publicaciones de GFS (Google FileSystem) y MapReduce que son los dos pilares fundamentales de Hadoop y de las tecnologías Big Data, que en 2006 incluyó Yahoo! en Hadoop.

2003



Conceptos clave

 Al hablar de big data se está hablando de “sets de datos extremadamente grandes que deben ser analizados computacionalmente para revelar patrones, tendencias y asociaciones, especialmente relacionadas al comportamiento humano y a las interacciones” (Oxford English Dictionary, 2013).

 Hay muchos mitos respecto al big data, pero lo real es que sí es posible entender las técnicas de manipulación y de análisis de datos para obtener un valor real y un retorno de inversión sobre los esfuerzos ejecutados.

 Big data se compone de velocidad, volumen y variedad del tipo de información a manipular.

Referencias Bibliográficas

Dontha, R. (2017). The Origins of Big Data. Recuperado de https://www.kdnuggets.com/2017/02/origins-big-data.html

Marr, B. (2017). 5 Massive 'Big data' Myths Most People Believe - But Shouldn´t. *Revista Forbes*. Recuperado de https://www.forbes.com/sites/bernardmarr/2017/09/26/5-massive-big-data-myths-most-people-believe-but-shouldnt/#1493352b2414

Oxford English Dictionary. (2013). Big data. Recuperado de https://en.oxforddictionaries.com/definition/big_data