GESTIÓN DE INFORMACIÓN

Lectura 1

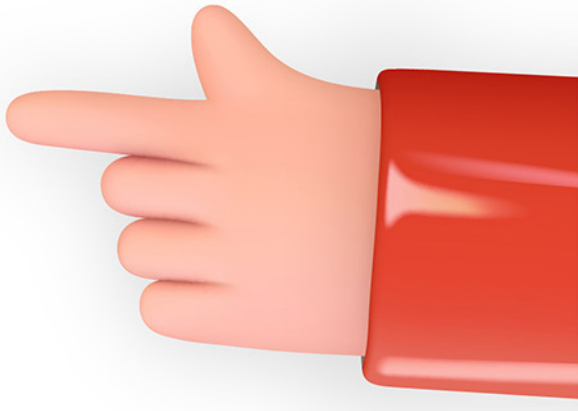
Entendiendo el Big Data

Lectura 1

DESBLOQUEADA

>>> Un paso más cerca de alcanzar tus objetivos

CONTENIDO



Entendiendo el Big Data

El Big Data se ha convertido en un concepto que a veces genera confusión. Para los expertos en tecnología tiene un significado claro. Sin embargo, en el mundo de los negocios, se ha utilizado de manera muy laxa para describir varios diferentes procesos relacionados con tecnología, desde el Big Data en sí hasta todo aquello que podemos hacer con los datos, como análisis y analítica de datos, ciencia de datos, aprendizaje automático, aprendizaje profundo e inteligencia artificial.

Durante este curso develaremos cada uno de los conceptos para entender claramente todas estas tecnologías, cuál es su papel en el mundo de los negocios y qué debemos de comprender claramente para así poder sacar plena ventaja de estas impactantes tecnologías en las empresas de hoy en día.

Para iniciar esta unidad descubriremos específicamente qué es el Big Data y empezaremos a ver más detalladamente cuáles son sus características y la importancia de cada una de ellas, así como comprender por qué el Big Data y todo lo relacionado con esta es de vital importancia para las empresas de esta era digital 4.0.

¿Qué es el Big Data?

En su mejor definición, Big Data es un campo de estudio que trata de analizar, extraer información sistemáticamente y trabajar con sets de datos que son demasiado grandes o complejos para lidiar con ellos basadas en aplicaciones de software para procesamiento de datos tradicionales.

Por ejemplo, uno de los métodos tradicionales de manejo de datos son las conocidas hojas de cálculo, como Excel o Google Spreadsheets. Estas siguen siendo muy útiles para guardar datos, pero tienen limitaciones importantes, como su tamaño (generalmente solo pueden manejar hasta 65 mil líneas de información), el tipo de datos que pueden manejar, que deben ser tabulares y estructurados, las limitadas formas de manipulación, manejo y análisis de datos y por supuesto, su increíblemente lento nivel de procesamiento.

Hoy en día, las bases de datos contienen millones, billones y hasta trillones de líneas, pueden manejar datos estructurados y no estructurados (como videos, imágenes, etc.), y pueden acceder a un sinfín de servicios y métodos variados para su manejo, manipulación y análisis.

Aparte del volumen y la variedad de datos, el Big Data también incluye el enfrentar varios retos de un mundo con un crecimiento exponencial en la cantidad de datos que recolecta. Algunos de los retos más importantes del Big Data son:

- Captura de datos.
- Almacenamiento de datos.
- Análisis de datos.
- Búsquedas eficientes.
- Compartir datos.
- Transferencia de datos.
- Visualización de datos.
- Actualización de datos.
- Privacidad de los datos.
- Abastecimiento de datos.

Cada uno de estos retos se desglosa en cientos de pequeños retos importantes de solucionar en el mundo actual. La realidad es que los datos son el petróleo del Siglo XXI, como lo menciona el Dr. James Bellini, especialista en Pensamiento Futuro; como el petróleo, los datos son valiosos, pero sin refinar no pueden ser utilizados para nada. El petróleo debe ser transformado en gas, plástico, químicos, etc. Para crear algo valioso que cree actividades generadoras de réditos, los datos también deben ser procesados y analizados para crear algo de valor.

Puesto que los datos son el petróleo del Siglo XXI, debemos aprender qué son, cómo se manejan y qué debemos hacer para crear procesos dentro de las empresas que nos permitan recolectarlos, manipularlos, analizarlos y refinarlos para crear productos y servicios de alto valor.

Las 5 V del Big Data

Cuando hablamos de datos, hablamos específicamente de información estructurada, semiestructurada y no estructurada, con características específicas. Estas características se conocen primordialmente como las 5 V.

Volumen

La primera característica del Big Data es el volumen. Simplemente se refiere a la cantidad de data existente.

El volumen en sí es la base del Big Data, ya que define el tamaño y la cantidad de datos que recolectamos.

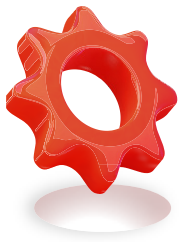
Si esta cantidad de datos es lo suficientemente grande, en tamaño y/o en cantidad, es cuando nos referimos específicamente a Big Data. En sí lo que se considera un volumen grande para ser considerado Big Data es difícil de definir, puesto que va cambiando con el tiempo, las tecnologías de recolección y almacenamiento, y el poder de procesamiento y más. En términos generales, podemos pensar que cuando hablamos de Big Data nos referimos a cantidades de millones de unidades de datos.

Una buena manera de entender el concepto de volumen dentro del Big Data es comprender un poco sobre las capacidades de los computadores de hoy en día. Una computadora laptop promedio tiene capacidades de almacenamiento entre los 512 gigabytes y 1 terabyte (1024 gigabytes). Para llenar uno de estos computadores con información tradicional, como lo serían filas y columnas de Excel, estaríamos hablando de varios o cientos de millones de filas, dependiendo de la cantidad de columnas, ya que cada fila puede ocupar entre unos pocos bytes hasta unos cuantos kilobytes. Pero, cuando hablamos de archivos más complejos como videos, un archivo de video en 4K ocupa aproximadamente 84 megabytes, y tomando en cuenta que en la plataforma YouTube se cargan, en promedio, 500 horas de video cada minuto —unos 30.000 minutos por minuto— (Statista Research Department, 2021), lo que equivale a 252 gigabytes por minuto. En otras palabras, una computadora promedio llegaría a capacidad en aproximadamente cuatro minutos. Eso es lo que realmente llamamos un alto volumen en Big Data.

Velocidad

La velocidad del Big Data se refiere a qué tan rápido la data que recolectamos se genera y qué tan rápido se mueve a través de nuestros sistemas. Es importante notar que las empresas, especialmente hoy en día, necesitan que sus datos fluyan con la mayor velocidad posible, para que esté disponible lo antes posible para poder tomar las mejores decisiones posibles para el negocio en cualquier instante.

Las empresas que utilizan Big Data tienen flujos grandes y continuos de datos que se crean constantemente y se envían a su destino lo antes posible. Los datos pueden fluir de fuentes tales como máquinas (por ejemplo, cualquier tipo de sensor), redes de trabajo, teléfonos inteligentes y redes sociales, entre otras. Estos datos deben ser almacenados, procesados y analizados lo más rápido posible, inclusive en tiempo real.



Un ejemplo del mundo de la medicina actual, hay dispositivos médicos, como los marcapasos inteligentes, que monitorean constantemente al paciente y recolectan datos cada segundo. Estos datos deben ser enviados rápidamente a su destino para ser analizados y poder descubrir si el paciente va camino a un episodio peligroso o si todo está bien en este momento.

Sin embargo, en algunos casos es importante disminuir la velocidad de los datos para estar seguros de que la empresa es capaz de procesarlos correctamente en vez de incrementar la velocidad y saturar los sistemas y las capacidades de la organización.

Para entender la velocidad, en la plataforma de búsqueda Google, se realizan unos 3.8 millones de búsquedas cada minuto. Google utiliza esta información para alimentar varios de sus sistemas de inteligencia artificial, en especial aquellos que realizan procesamiento de lenguaje natural, como el autocompletado de oraciones. Eso quiere decir que la velocidad de la información en el buscador de Google es de aproximadamente 63 333 búsquedas por segundo. Aquí es donde empezamos a ver que los números tanto de volumen como de velocidad se vuelven realmente alucinantes.

Variedad

La siguiente V del Big Data es la variedad. Esto se refiere a la diversidad de tipos de datos. Una empresa puede obtener sus datos de diferentes fuentes tanto dentro como fuera de la organización.

El reto de la variedad consiste en la forma en que se estandariza, almacena y distribuye la data luego de su captura.

Existen tres grandes formatos relativos a la data, estos pueden ser:

- Estructurados: la forma más común en la que capturamos y guardamos datos. Aquí nos referimos generalmente a tablas tradicionales. La forma usual de almacenar estos datos es mediante Bases de Datos Relacionales que se manejan a través de SQL —Structured Query Language (Lenguaje de Consulta Estructurado) —.

- Semiestructurados: estructuras más recientes que nos dan mayor flexibilidad en la captura y almacenamiento de datos. Estructuras basadas en diccionarios, como, por ejemplo, JSON y XML, así como otros modelos interesantes como columnas variables y otros. Existen varios sistemas de almacenamiento de datos de esta índole, como MongoDB y Cassandra. Generalmente se los conoce como sistemas NoSQL.

- No estructurados: aquí nos encontramos un sinfín de tipos de data, sin estructuras específicas y que poseen sus propios retos para el almacenamiento, distribución y análisis. Incluyen videos, audios, grafos y demás.

Para manejar la variedad de datos en el Big Data, hoy en día se está migrando de los modelos más tradicionales de los almacenes de datos (Data Warehouses) a modelos como los lagos de datos (Data Lakes).

Las plataformas más avanzadas, como Google, Facebook, Amazon, Microsoft y similares, utilizan lagos de datos altamente avanzados que les permiten recibir cualquier tipo de información, procesarla, almacenarla, transformarla y alimentarla a otros sistemas en tan solo instantes. Pensemos por ejemplo en todos los sistemas de búsqueda de Google para entender un poco como todos estos sistemas se integran. Si decidimos hacer una búsqueda con el asistente de Google utilizando nuestra voz, lo primero que sucede es una app que capta la voz y la transforma en un archivo de audio; luego este archivo de audio se transfiere a otra app que transforma el archivo de audio en texto, luego a otra que reconfigura el texto para intentar entender la búsqueda, luego a otra app que realiza la búsqueda y devuelve la información, otra app que transforma el texto encontrado a un archive

de audio y finalmente una app que reproduce el archivo de audio. Así que vemos como la variedad de data no es solo una realidad existente, sino una que debemos comprender para que nuestros sistemas puedan integrarse unos con otros eficientemente.

Veracidad

La cuarta V del Big Data es la veracidad. Este es un reto que requiere ir mucho más allá de lo tecnológico ya que se refiere a la calidad y precisión de los datos que recolectamos. La data puede tener piezas faltantes, puede ser poco precisa y por ende ser incapaz de proveer valor a la organización. La veracidad, en resumen, se refiere a qué tanto podemos confiar en los datos.

La data puede ser a veces desordenada y difícil de utilizar. Una base de datos grande con partes incompletas puede generar problemas a la hora de analizar los datos, o peor aún, generar revelaciones falsas.

Por ejemplo, si en el sector médico tenemos datos faltantes en relación con un nuevo medicamento que está en prueba, podemos poner en riesgo la vida de los pacientes.

La veracidad y el valor son los que finalmente definirán la calidad de los insights que logremos descubrir de nuestra data.

Valor

El fin del Big Data es poder generar insights de alto valor para las organizaciones, por ello, la última V del Big Data es el valor. El poder generar valor de los datos es el requerimiento principal de cualquier proyecto e inversión en Big Data; mientras mayor sea el valor generado, mayor será la inversión futura en más y más datos. Como decíamos al principio, la data es el petróleo del Siglo XXI, y debemos ser capaces de transformarla en algo de utilidad. Productos o servicios nuevos, mejoras en procesos existentes que generen ahorros o ventajas competitivas, en fin, lograr hacer algo útil y valioso con los datos que generamos y recolectamos.

Aquí vemos que la pericia de las personas toma un poco más de precedencia, ya que varias organizaciones pueden tener los mismos datos, utilizar los mismos sistemas, pero lograr sacar valores muy diferentes a la misma.

Un buen ejemplo de esto es la empresa Zillow, que utilizó datos abiertos de cientos de miles de propiedades en los Estados Unidos, y mediante su equipo de ciencia de datos y utilizando modelos de aprendizaje automático, lograron crear uno de los mejores sistemas de predicción de precio de venta para propiedades, dándoles una ventaja competitiva enorme en el mercado.

El significado del Big Data para la economía mundial

El mundo del Big Data sigue cambiando y transformándose de manera estrepitosamente rápida. Por ello, se ha propuesto en algunos casos agregar un V adicional, y en otros dos más: variabilidad y visualización.

La variabilidad se refiere a la capacidad de los datos de cambiar su significado e interpretaciones en relación con el contexto o momento histórico en que se analicen. De la misma forma, podemos encontrar en ellos una variabilidad basada en los modelos, herramientas y tecnologías específicas que se utilicen para analizarlos y que están en constante evolución y cambio. A continuación, se desarrolla ejemplo para comprender esto, supongamos que todos los días vamos a una cafetería, en esta cafetería venden seis tipos de café (esto sería variedad), siempre pedimos exactamente el mismo tipo de café, pero según quién sea el barista del momento, cómo está la calidad del agua en una época determinada o cuál máquina de café se utilice para hacer nuestra taza específica de café, este sabe ligeramente diferente, esto es justamente a lo que nos referimos con variabilidad.

La visualización se refiere al potencial e importancia de contar con sistemas y herramientas que nos permitan visualizar de la mejor manera posible los datos, lo que nos permite entenderlos de mejor manera y logra crear mejor insights y así generar un mayor valor en los mismos.

Aunque estas dos V son también importantes, muchos especialistas las consideran parte del proceso de análisis del Big Data y no una característica intrínseca del mismo.

Ya en este punto podemos empezar a comprender el Big Data y sus características. La realidad es que esto nos afecta a todos hoy en día, queramos o no, lo aceptemos o no. ¿Es posible que una empresa funcione actualmente sin utilizar herramientas como el correo electrónico o las videollamadas? ¿Puede nuestra empresa funcionar sin utilizar servicios de empresas como Amazon, Microsoft, Facebook o Google? Es altamente probable que esto sea imposible.

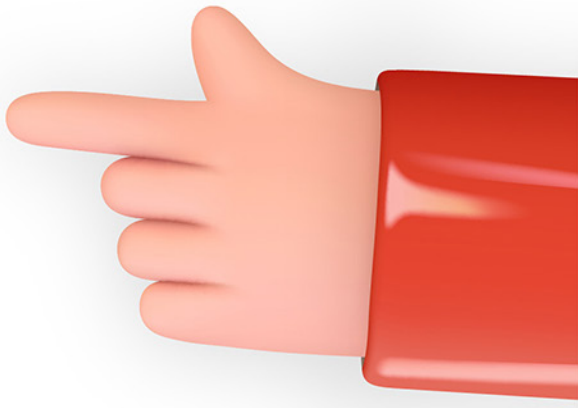
Para entender lo increíblemente interconectados que estamos con la tecnología hoy en día, en Julio del 2016, la empresa Google tuvo un corte en sus servicios que tardó aproximadamente cinco minutos. Para Google, esto significó la pérdida de 545.000$ dólares en ingresos (en otras palabras, en ese momento Google recibía unos 108 000$ en ingresos cada minuto). Muchas organizaciones, desde instituciones públicas hasta empresas privadas, sufrieron este corte de servicios. Aunque no tenemos en este momento una manera objetiva y sólida de entender cuáles fueron las pérdidas reales en la economía mundial, especialistas estiman que debe de ubicarse en al menos unos pocos billones de dólares que se perdieron en esos tan solo cinco minutos.

Tomando en cuenta que, según el reporte de Canalys sobre el mercado de los servicios web en el Q1 del 2021, Google únicamente controla el 7 % del mercado actual (AWS tiene el 32 % y Microsoft Azure el 19 %).

Se puede imaginar que, si un pequeño corte de Google es capaz de generar pérdidas billonarias en la economía mundial, la realidad es que nuestra economía mundial está ya inexorablemente entrelazada con la tecnología y, en el centro, se encuentran los servicios y aplicaciones del Big Data.

Las organizaciones de hoy deben de centrarse en la data, o prepararse para desaparecer.

BIBLIOGRAFÍA



Analytic Insights. (2021). *The impact of Big Data in Agriculture*. https://www.analyticsinsight.net/the-impact-of-big-data-in-agriculture/

Butcher, J. (2021). *Data is the new oil of the 21st century*. S4RB. https://blog.s4rb.com/data-is-the-oil-of-the-21st-century

Canalys. (2021). *Global cloud services market surges by US$10 billion in Q4 2020*. https://www.canalys.com/newsroom/global-cloud-market-q4-2020

Casamichana, M. (2019). *5 empresas que usan Big Data y han conseguido los mejores resultados*. https://business-intelligence.grupobit.net/blog/empresas-que-usan-big-data-y-han-conseguido-los-mejores-resultados

CFI. (s.f.). *Big Data in Finance*. https://corporatefinanceinstitute.com/resources/knowledge/other/big-data-in-finance/

Doz, Y. (2017). *The Strategic Decisions That Caused Nokia's Failure*. Insead. https://knowledge.insead.edu/strategy/the-strategic-decisions-that-caused-nokias-failure-7766

Durcevic, S. (2021). *18 Examples of Big Data Analytics in Healthcare That Can Save People*. DataPine. https://www.datapine.com/blog/big-data-examples-in-healthcare/

Gillis, A. S. (2021). *Las 5 V's del Big Data*. TechTarget. https://searchdatamanagement.techtarget.com/definition/5-Vs-of-big-data

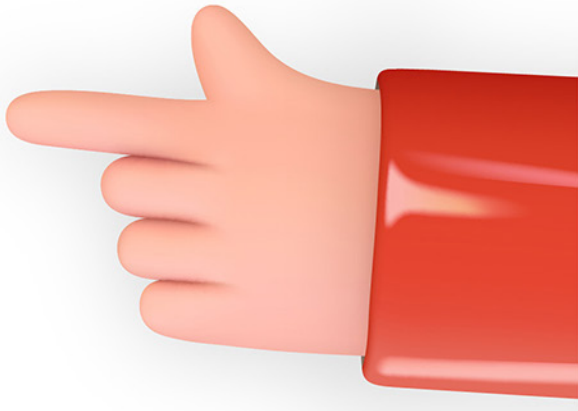
Ilchenko, V. (2020). *How Big Data is Boosting the Food Industry: The Best Examples*. Byteant. https://www.byteant.com/

Kahn Academy. (s.f.). *Crecimiento Exponencial*. https://es.khanacademy.org/science/ap-biology/ecology-ap/population-ecology-ap/a/exponential-logistic-growth

Kauflin, J. y Stoller, K. (2019). *First, Fire All the Brokers: How Lemonade, A Millennial-Loved Fintech Unicorn, Is Disrupting the Insurance Business*. Forbes. https://www.forbes.com/sites/jeffkauflin/2019/05/02/lemonade-fintech-insurance-unicorn/?sh=7384baa16cde

Lebied, M. (2017). *5 Examples of How Big Data in Logistics Can Transform the Supply Chain*. DataPine. https://www.datapine.com/blog/how-big-data-logistics-transform-supply-chain/

BIBLIOGRAFÍA



Marr, B. (2016). *Big data en la Práctica*. Teell Editorial, S.L.

OECD. (2020). *The Impact of Big Data and Artificial Intelligence (AI) in the Insurance Sector*. https://www.oecd.org/pensions/impact-big-data-ai-in-the-insurance-sector.htm

Pérez. A. (2018). *¿Cómo ayuda el Big Data a las empresas? OBS Business School*. https://www.obsbusiness.school/blog/como-ayuda-el-big-data-las-empresas

Statista Research Department. (2021). *Hours of video uploaded to YouTube every minute as of May 2019*. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

Zohu, K., Fu, C. y Yang, S. (2016). *Big data driven smart energy management: From big data to big insights*. ScienceDirect, 56, 215-225. https://www.sciencedirect.com/science/article/abs/pii/S1364032115013179#:~:text=Big%20data%20analytics%20can%20provide,the%20same%20time%20for%20us

Bibliografía complementaria

Mayer-Schönberger, V. y Cukier, K. (2013). *Big data: La revolución de los datos masivos*. Turner Publicaciones S.L.

Mui, C. (2012). *How Kodak Failed*. Forbes. https://www.forbes.com/sites/chunkamui/2012/01/18/how-kodak-failed/?sh=6f05d2e6f27a

Schmarzo, B. (2013). *Big Data: Understanding How Data Powers Big Business*. Wiley.