UCAB UNIVERSIDAD CATÓLICA ANDRÉS BELLO

CONCEPTO DE *BIG DATA*, LAS VS DE *BIG DATA* Y LOS ROLES EN *BIG DATA*

Tabla de contenido

Objetivo

- Unificar y comprender el concepto de *big data*, usando sus características asociadas a las Vs de *big data* y a los roles de un equipo de *big data*.

Introducción

Desde el surgimiento del término *big data*, han sido varias las miradas y formas de definirlo. En este tema revisaremos tanto sus distintas definiciones como las famosas Vs de *big data*, las que dieron origen al concepto y las que han ido sumándose a lo largo del tiempo como forma de adaptarse a las nuevas necesidades.

También estudiaremos las características que debe tener un equipo empresarial que quiera llevar adelante un proyecto de *big data*.

Concepto de *big data*

Desde que se acuñó el término de *big data* (en español, macro datos) se han propuesto varias definiciones, las cuales revisaremos para llegar a una que abarque los aspectos más importantes (Ward y Barker, 2013).

Entre las definiciones más citadas se encuentra la incluída en un informe de Meta, ahora Gartner, de 2001. El informe de Gartner no menciona la frase "*big data*" y es anterior a la tendencia actual. Sin embargo, desde entonces este ha sido adoptado como una definición clave. Gartner propuso una definición triple que abarca las "tres V": volumen, velocidad, variedad. Esta consiste en "...un gran volumen, velocidad o variedad de información que demanda formas costeables e innovadoras de procesamiento que permitan ideas extendidas, toma de decisiones y automatización de procesos" (Gartner, 2013, s.p.).

Como es común en toda la literatura sobre *big data*, la evidencia presentada en la definición de Gartner es completamente anecdótica; no hay una cuantificación numérica de *big data*. Desde entonces, esta definición ha sido reiterada por NIST y Gartner en 2012, ampliada por IBM y otros para incluir las limitaciones de las bases de datos tradicionales en torno al modelo relacional, y una cuarta V, la veracidad, que implica preguntas de confianza e incertidumbre con respecto a los datos y al resultado del análisis de los mismos.

Oracle evita emplear Vs en su concepción de *big data*. En cambio, sostiene que los macrodatos (*big data*) son la derivación de valor de la toma de decisiones comerciales impulsada por bases de datos relacionales tradicionales, aumentada con nuevas fuentes de datos no estructurados. Oracle, por lo tanto, da una perspectiva que trata sobre la inclusión de fuentes de datos adicionales, no relacionales, para aumentar las operaciones existentes. En particular, y tal vez como era de esperar, la definición de Oracle se centra en la infraestructura. (Ward y Barker, 2013, pág. 1).

Si bien esta definición se aplica algo más fácilmente que otras, también carece de cuantificación. Según la definición de Oracle, no está claro exactamente cuándo se aplica el término *big data*, sino que proporciona un medio para "saberlo cuando lo vea".

Intel es una de las pocas organizaciones que proporciona cifras concretas en su literatura. Esta vincula *big data* a organizaciones que "...generan una media de 300

terabytes (TB) de datos por semana". En lugar de proporcionar una definición de acuerdo a las organizaciones antes mencionadas, Intel describe el *big data* mediante la cuantificación de las experiencias de sus socios comerciales, indicando que las organizaciones encuestadas manejan ampliamente datos no estructurados y semiestructurados y ponen énfasis en realizar análisis sobre sus datos, que se producen a una velocidad de hasta 500 TB por semana. (Ward y Barker, 2013, pág. 1).

Microsoft proporciona una definición notablemente concisa:

> *Big data* es el término que se utiliza cada vez más para describir el proceso de aplicación de poder de cómputo serio, lo último en aprendizaje automático e inteligencia artificial, a conjuntos de información enormemente masivos y, a menudo, muy complejos. (Ward y Barker, 2013, p.2).

Se puede obtener una definición, o al menos una indicación, de tecnologías relacionadas mediante una investigación de términos vinculados. Google Trends proporciona los siguientes términos, en relación con *big data*, de mayor a menor frecuencia: análisis de datos, Hadoop, NoSQL, Google, IBM y Oracle. A partir de estos términos, resultan evidentes una serie de tendencias. En primer lugar, los macrodatos están intrínsecamente ligados con el análisis de datos y el descubrimiento del significado de los datos. En segundo lugar, está claro que hay una serie de tecnologías relacionadas a las que se alude en la definición de Microsoft, a saber, NoSQL y Apache Hadoop. Finalmente, es evidente que hay una serie de organizaciones, específicamente organizaciones industriales, que están asociadas con *big data*.

A pesar del rango y las diferencias que existen dentro de cada una de las definiciones antes mencionadas, existen algunos puntos de similitud. En particular, todas las definiciones hacen al menos una de las siguientes afirmaciones:

- Tamaño: el volumen de los conjuntos de datos es un factor crítico.
- Complejidad: la estructura, el comportamiento y las permutaciones de los conjuntos de datos son un factor crítico.

– Tecnologías: las herramientas y técnicas que se utilizan para procesar un conjunto de datos considerable o complejo es un elemento esencial..

Todas las definiciones analizadas aquí abarcan al menos uno de estos factores; la mayoría abarca dos. Por lo tanto, una extrapolación de estos factores postularía lo siguiente: *big data* es un término que describe el almacenamiento y análisis de conjuntos de datos grandes o complejos utilizando una serie de técnicas que incluyen, entre otras: NoSQL, MapReduce, procesamiento paralelo y distribuido y aprendizaje automático.

Una definición que engloba estos aspectos que hemos destacado, propuesta por Mauro, Greco, y Grimaldi (2014) p.17, es la siguiente: "*Big Data* representa los activos de información caracterizados por un volumen, velocidad y variedad tan altos que requieren tecnología y métodos analíticos específicos para su transformación en valor".

Las Vs de *big data*

Desde la definición de Gartner, basada en las 3 Vs, pasando por la de IBM que incluye una cuarta, en la actualidad se habla de 10 y hasta de 12 Vs. A continuación, trataremos las 8 que se consideran más relevantes (TDWI, 2017).

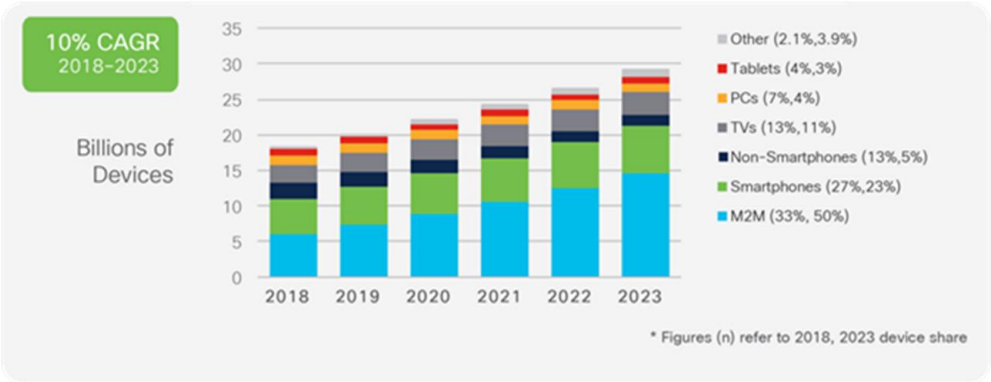
- El volumen se refiere a la cantidad de datos que se generan por unidad de tiempo en nuestro entorno, proveniente de diversas fuentes, especialmente de fuentes externas a la empresa, como redes sociales, datos móviles e IoT, como se puede observar en los gráficos tomados de la página de Cisco.

Es la característica más asociada al *big data*, ya que hace referencia a las cantidades masivas de datos que se almacenan con la finalidad de procesarlos para transformarlos en acciones.

Crecimiento de trafico de datos móviles para 2022

Fuente: Crecimiento de tráfico de datos móviles para 2022. CISCO (2020)

Crecimiento global de dispositivos y conexiones



Fuente: Crecimiento global de dispositivos y conexiones. CISCO (2020)

- La velocidad se refiere a los datos en movimiento por las constantes interconexiones que realizamos, es decir, a la rapidez en la que son creados, almacenados y procesados en tiempo real.

  Un ejemplo de esto es la velocidad con la que se generan datos en redes sociales, tal como se muestra en esta infografía.

Fuente: 2021. This is what happens in an internet minute. Extraída de Lori Lewis (2021)

Para los procesos en los que el tiempo resulta fundamental, tales como la detección de fraude en una transacción bancaria o la monitorización de un evento en redes sociales, estos tipos de datos deben estudiarse en tiempo real para que resulten útiles para el negocio y se consigan conclusiones efectivas.

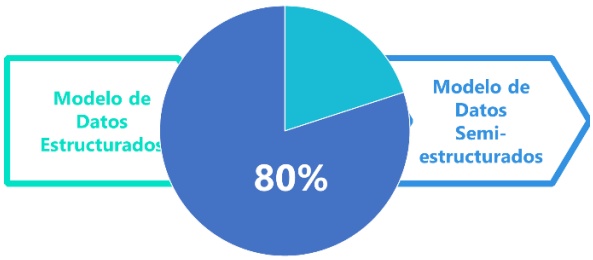
- La variedad se refiere a las formas, tipos y fuentes en las que se registran los datos. Estos datos pueden ser:

    - Datos estructurados, fáciles de gestionar como son las bases de datos u hojas de cálculo.
    - Datos no estructurados, entre los que se incluyen documentos de texto, correos electrónicos, datos de sensores, audios, vídeos o imágenes que tenemos en nuestro dispositivo móvil, hasta publicaciones en nuestros perfiles de redes sociales, artículos que leemos en blogs, las secuencias

de click que hacemos en una misma página, formularios de registro e infinidad de acciones más que realizamos desde nuestro *Smartphone*, *Tablet* y ordenador.

- Datos semiestructurados que tienen cierta estructura pero que es muy flexible, como es el caso de los metadatos o los servicios web publicados en formatos como XML o Json.

Fuente: Tipos de datos del big data. Extraída de GoConqr (s.f.)



Algunas fuentes han determinado que en la actualidad el 80 % de los datos generados son no semiestructurados y solo el 20 % son estructurados.

- Cuando hablamos de veracidad nos referimos a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información recibida y de las fuentes. Se refiere al sesgo, ruido y/o posible alteración de los datos.

Imagina un conjunto de datos estadísticos sobre lo que la gente compra en los restaurantes y los precios de estos artículos en los últimos cinco años. Puedes preguntar: ¿Quién creó la fuente? ¿Qué metodología siguieron para recopilar los datos? ¿Solo se incluyeron ciertas cocinas o ciertos tipos de restaurantes? ¿Los creadores de datos resumieron la información? ¿Esta ha sido editada o modificada por alguien más?

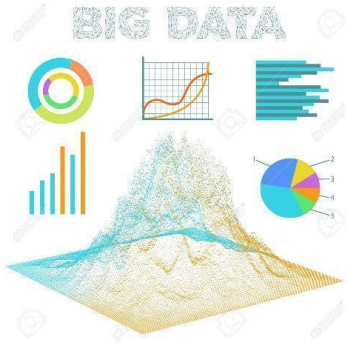
Es necesario invertir tiempo si se desea conseguir datos de calidad, aplicando soluciones y métodos para eliminar datos imprevisibles que puedan surgir.

La necesidad de explorar y planificar la incertidumbre es un reto para el *big data* que está a la orden del día en las compañías dedicadas al análisis de datos.

- La variabilidad es diferente a la variedad; se refiere a datos cuyo significado cambia constantemente. Muchas veces las organizaciones necesitan desarrollar programas sofisticados para poder comprender el contexto en ellos y decodificar su significado exacto (O'Reilly, 2021).

Por ejemplo, una palabra o frase puede tener significados diferentes de acuerdo al contexto social o cultural, lo que sucede mucho en redes sociales cuando damos nuestra opinión con respecto a un producto.

- Otra característica de los macrodatos es lo difícil que es visualizarlos.

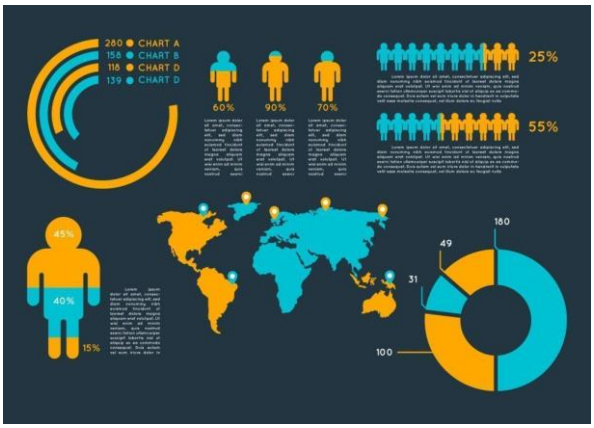


Fuente: *Big data*.
Extraída de 123RF (s.f.)

Las herramientas de visualización de *big data* actuales enfrentan desafíos técnicos debido a las limitaciones de la tecnología en memoria y a la escasa escalabilidad, funcionalidad y tiempo de respuesta. No se pueden interpretar adecuadamente los gráficos tradicionales cuando intentan trazar mil millones de puntos de datos, por lo que se hacen necesarias diferentes formas de representar datos, como la agrupación de datos o el uso de mapas, rayos de sol, coordenadas paralelas, diagramas de redes circulares o árboles de conos.

Al combinar esto con la multitud de variables que resultan de la variedad y velocidad de los datos, y las relaciones complejas entre ellos, se evidencia que desarrollar una visualización significativa no es fácil.

Por otro lado, se tienen las infografías que requieren una interpretación previa, ya que son estáticas, pero fácilmente interpretables como la que se muestra acá:



Fuente: Visualización del *big data*. Extraída de Daniel Fernández (2019)

- La volatilidad se refiere al tiempo que deben conservarse los datos antes de que se consideren irrelevantes, históricos o ya no sean útiles.

Antes del *big data* se tendía a almacenar datos indefinidamente, debido a que su pequeño volumen apenas suponía gastos. Incluso podía mantenerse en la base de datos en vivo sin causar problemas de rendimiento.

Sin embargo, debido a la velocidad y el volumen en *big data,* su volatilidad debe considerarse cuidadosamente. Ahora hay que establecer reglas para la disponibilidad y la vigencia de estos datos, así como para garantizar una recuperación rápida de la información cuando sea necesario.

- Por último, y no por eso menos importante, tenemos el valor, el cual se puede ver como el elemento final del *big data*. Cada usuario debe comprender que la organización necesita generar valor después de que se realizan esfuerzos y se gastan recursos en las V mencionadas anteriormente. Esta normalmente se convierte en una acción o decisión que genera una ventaja o ganancias para la organización (o sus usuarios), como crear un nuevo producto, debido a que se conoce mejor al cliente o mejora los existentes, detecta relaciones que permiten establecer un nuevo tratamiento para una enfermedad, conoce tendencias de mercado o la propagación de un virus como es el caso actual.

Roles de un equipo de *big data*

Cuando buscas información acerca de las habilidades que debe tener un científico de datos, nos encontramos con algo como esto:

Fuente: Ciencia de datos: el trabajo del siglo XXI. Extraída de Marketing Distillery (s.f.)

Lo anterior hace que nos preguntemos si una sola persona puede tener experticia en todas estas áreas, por lo que se concluye que un proyecto de *big data* debe contar con un equipo de trabajo multidisciplinario con diferentes habilidades. Entre los roles que podemos encontrar en este tipo de equipos se tienen (TodoBI, 2019):

1. Ingeniero de datos, también llamado ingeniero de *big data*.
   - **Rol**: desarrollar, construir, probar y mantener arquitecturas tales como bases de datos y sistemas de procesamiento a gran escala.
   - **Habilidades y talentos:**
     - Sistemas de bases de datos (SQL y NoSQL)
     - Modelado de datos y herramientas ETL
     - Apis de datos
     - Soluciones de inteligencia de negocios
     - Hadoop (instalación, configuración y mantenimiento).
   - **Lenguajes y programas:** SQL, Hive, Pig, R, Python, Java, C++, entre otros.

2. Arquitecto de datos.
   - **Rol:** crear tuberías de datos para integrar, centralizar, proteger y mantener fuentes de datos.
   - **Habilidades y talentos:**
     - Soluciones de almacenes de datos
     - Conocimiento en profundidad de arquitectura de bases de datos y de sistemas de *big data*
     - Conocimiento y comprensión completa de las tecnologías del ecosistema de *big data* y de sus casos de uso
     - Análisis de requisitos técnicos y del sistema y selección de tecnología
     - Administración de todas las bases de datos, sus objetos y datos en la plataforma de *big data*
     - Gestión del ciclo de vida de la solución de *big data*
     - Trabajo en conjunto con el ingeniero de *big data*.
   - **Lenguajes y programas:** SQL, XML, Hive, Pig, Spark y demás tecnologías de *big data.*

3. Administrador de base de datos.
   - **Rol:** asegurar que las bases de datos para todos los usuarios relevantes trabajen adecuadamente y mantenerla segura.
   - **Habilidades y talentos:**
     - Respaldo y recuperación
     - Modelado y diseño de base de datos
     - Computación distribuida (hadoop)
     - Sistemas de bases de datos (SQL y NoSQL)
     - Seguridad de datos
     - Conocimientos de ERP.
   - **Lenguajes:** SQL, Java, Json, XML, Python.

4. Analista de datos.
   - **Rol:** recopila, procesa y realiza análisis de datos estadísticos.
   - **Habilidades y talentos:**
     - Manejo de hojas de cálculo (p.e. Excel)
     - Sistemas de bases de datos (SQL y NoSQL)
     - Comunicación y visualización
     - Matemáticas, estadística, *machine learning*.
   - **Lenguajes:** R, Python, HTML, Javascript, SQL, Json.

5. Gerente de data y analítica / *Chief Data Officer* (CDO).
   - **Rol:** maneja a un equipo de analistas y científicos de datos.
   - **Habilidades y talentos:**
     - Sistemas de bases de datos (SQL y NoSQL)
     - Liderazgo y manejo de proyectos
     - Comunicación interpersonal
     - Minería de datos y modelado predictivo.
   - **Lenguajes:** SQL, R, SAS, Python, Java.

6. Científico de datos.
   - **Rol:** limpia, masajea, organiza y evalúa los grandes volúmenes de datos.
   - **Habilidades y talentos:**
     - Computación distribuida (hadoop)
     - Modelado predictivo
     - Comunicación y visualización
     - Matemáticas, estadística, *machine learning.*
   - **Lenguajes:** R, Python, SQL, Hive, Pig, Spark.

7. Chief Data Officer.
   - **Rol:** encargado de asegurar el gobierno del dato y de definir y comunicar la estrategia alrededor de los mismos.
   - **Habilidades y talentos:**
     - Gobierno del dato: normativas y regulaciones
     - Calidad de datos, herramientas ETL
     - Herramientas para calidad y gobierno del dato en la plataforma de *big data*
     - Comunicación y liderazgo
     - Visualización de datos.

Cierre

En este contenido se abordó cómo diferentes organizaciones definen o dan su concepto de *big data*, lo que nos permite seleccionar los elementos más relevantes y proponer una definición que los abarque. Estas definiciones parten de caracterizar los grandes volúmenes de datos a través de las Vs, las cuales comenzaron siendo 3 y han ido ampliándose. También estudiamos las 8 Vs que se consideran las más relevantes. Luego se buscaron cuáles deben ser las habilidades y talentos que debe tener un científico de datos para concluir que un equipo de *big data* no depende de una sola persona sino de un equipo entero, dentro del cual describimos los roles más importantes.

Referencias

Firican, G. (8 de febrero de 2017). *The 10 Vs of Big Data*. TDWI. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Gartner. (Abril de 2013). *Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V" s.* Obtenido de https://blogs.gartner.com/svetlana-sicular/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

Mauro, A. D., Greco, M., y Grimaldi, M. (2014). *What is Big Data? A Consensual Definition and a Review of Key Research Topics* [Presentación en papel]. 4th International Conference on Integrated Information, Madrid, España.

Meta Group (2001). *3d Data Management: Controlling Data Volume, Velocity and Variety*. https://idoc.pub/documents/3d-data-management-controlling-data-volume-velocity-and-variety-546g5mg3ywn8

O'Reilly (2021). *Variability of data*. https://www.oreilly.com/library/view/big-data-analytics/9781788628846/63e2e44b-66af-41e5-892d-7339de89c7d7.xhtml

TodoBI (2019). *Tipos de roles en Analytics (Business Intelligence, Big Data)*. https://todobi.com/tipos-de-roles-en-analytics-business/

Ward, J. y Barker, A. (2013). *Undefined By Data: A Survey of Big Data Definitions.* University of St Andrews. https://doi.org/10.48550/arXiv.1309.5821

Referencias de las imágenes

123RF (s.f.). Big data [Imagen]. Disponible en: https://es.123rf.com/photo_86737457_visualizaci%C3%B3n-de-big-data-

diagramas-de-barras-y-gr%C3%A1ficos-de-l%C3%ADneas-an%C3%A1lisis-de-informaci%C3%B3n-dise%C3%B1o-de-infograf%C3%ADas-de-d.html

CISCO (2020). Crecimiento global de dispositivos y conexiones [Imagen]. Disponible en: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.htm

CISCO (2020). Crecimiento de tráfico de datos móviles para 2022 [Imagen]. Disponible en: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.htm

Fernández, D. (2019). Visualización del big data [Imagen]. Disponible en: https://www.fdezdaniel.com/informando/visualizacion-del-big-data/

GoConqr (s.f.). Tipos de datos del big data [Imagen]. Disponible en: https://cdn.goconqr.com/en/p/31188374?dont_count=true&frame=true&fs=true

Lewis, L. (2021). 2021: This is What Happens in an Internet Minute [Imagen]. Disponible en: https://twitter.com/lorilewis/status/1382346076271837184

Marketing Distillery (s.f.). Ciencia de datos: el trabajo del siglo XXI [Imagen]. Disponible en: https://www.evaluandoerp.com/los-cientificos-datos-demanda-actual-previsiones-futuras/