Técnicas de Estudio

**Big Data**

Ingeniería Informática

Universidad de Mendoza 2018

línea horizontal

# Prof. Alejandra Párraga

# Índice:

# Introducción………………………………………………………………..……………………………………………………………3

# Desarrollo………………………………………………………………………………………………………………………………...4

Definición…………………………………………………………………………….……………………………………….4

Características o “las 4 Vs”….…………………………..…………………………………………………………..5

Herramientas y tipos de datos en BigData……………………………………………………………………7

Técnicas de análisis de datos………………………………………………………………………………………..9

BigData y la privacidad………………………………………………………………………………………………..10

# Conclusión………………………………………………………………………………………………………………………………12

# Bibliografía……………………………………………………………………………………………………………………………..13

# Introducción

BigData es una herramienta poderosa que tiene como fin mejorar la toma de decisiones, aunque se descubrió que tiene un gran campo de aplicaciones.

Es un concepto que se refiere generalmente a volúmenes importantes de información cuyas necesidades de almacenamiento y procesamiento no pueden ser atendidas por las tecnologías tradicionales que son usadas en sistemas de información. Como por ejemplo en las bases de datos relacionales.

Tras su llegada se crearon y desarrollaron nuevas tecnologías, paradigmas de programación e incluso puestos de trabajo.

## Resultado de imagen para big data

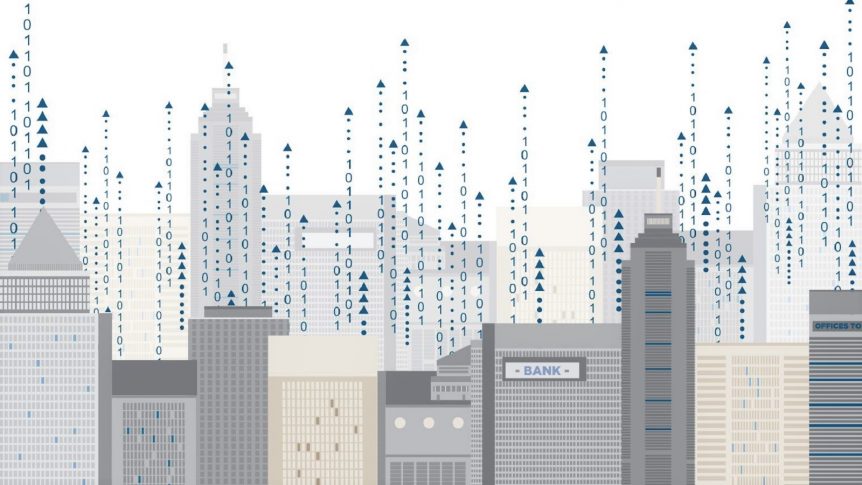
## Desarrollo

**Definición**

BigData y el análisis de datos es algo que está de moda hoy en día, pero tiene sus comienzos en 1930, cuando ya se llevaron a cabo la primera recolección y análisis de datos en esta escala.

BigData permite analizar información para generar métricas y en función de éstas, tomar decisiones que antes eran basadas en experiencia e intuición, pero ahora pueden llevarse a cabo tanto por sistemas como por personas, gracias a las grandes cantidades de datos que las máquinas pueden analizar con su gran poder de procesamiento. Estas revelan patrones y conexiones que mejoran las actividades humanas y también generan perfiles detallados de todos nosotros, incluyendo información que elegimos mantener privada.

La tendencia a manipular enormes cantidades de datos se debe a la necesidad en muchos casos de incluir dicha información para la creación de informes estadísticos y modelos predictivos utilizados en diversas materias, como los análisis de negocios, publicidades, prevención de enfermedades infecciosas, espionaje y seguimiento de la población o la lucha contra el crimen organizado. El superior poder de procesamiento ha facilitado esta tarea a lo largo de los años. Se estima que el mundo almacenó unos 5 [zettabytes](https://es.wikipedia.org/wiki/Zettabyte) de información en 2014. Y que, en los últimos 2 años, se generó el 90% de la información existente en la historia humana.

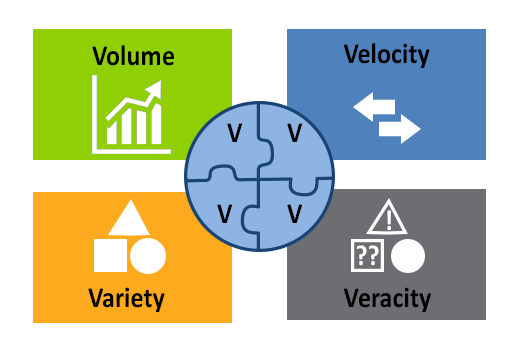
También se ha definido a BigData como datos lo suficientemente masivos como para poner en relieve cuestiones y preocupaciones en torno a la efectividad del anonimato y la privacidad. Esto permite generar un perfil virtual nuestro que, a pesar de no llevar nuestro nombre, posee información nuestra, como la ubicación o la fecha en la que estuvimos en ciertos lugares o información que desconocemos de nosotros mismos. Este tema será tratado más adelante.

**Características o “las 4 Vs”**

Volumen: hace referencia a la cantidad de datos generados y guardados, que pueden ser utilizados para la toma de decisiones. El volumen de los datos determina el valor y entendimiento potencial de la información que estos aportan. Los sistemas de BigData analizan petabytes de información en un periodo corto de tiempo.

Velocidad: es una característica crucial de BigData. La velocidad a la cual se generan, guardan y procesan los datos es fundamental para poder tomar decisiones rápidas, en “Streaming”, es decir, en tiempo real, y para cumplir las exigencias y desafíos de su análisis. Al BigData se lo define usualmente como “Datos en constante movimiento”.

La búsqueda de la disminución de la latencia, definida como el tiempo que transcurre entre que se crea el dato hasta que puede ser utilizado, es un tópico muy importante cuando se habla de BigData.



Variedad: refiere al tipo, naturaleza y fuente de los datos, estos pueden estar en forma estructurada, semi-estructurada o no estructurada. Sirven para ayudar a las personas a analizar el volumen datos y usar los resultados de forma eficaz. BigData utiliza textos, imágenes, audio, vídeo, entre otros tipos de datos. También completan pedazos perdidos a través de la fusión de dos o más datos. Una fuente importante de datos son las redes sociales y aplicaciones móviles.

Veracidad de la fuente: refiere a que la [calidad de datos](https://es.wikipedia.org/wiki/Calidad_de_datos) capturados puede variar mucho y así afectar a los resultados del análisis, por lo que en BigData es importante trabajar con un grado de incertidumbre. En BigData no existen los datos totalmente ciertos, ya que estos provienen de usuarios, los cuales no son completamente fiables y, además son muy diferentes entre sí.

Algunos autores incluyen una quinta “V”, llamada: Valor. La cual representa el potencial que poseen estos datos para la prosperidad de la empresa, el valor que estos datos poseen para el negocio.

Incluso otros llegan a definir hasta “7 Vs”, pero las mas importantes son las mencionadas anteriormente.

**Herramientas y tipos de datos en BigData**

Existen muchas herramientas para tratar con BigData. Algunos ejemplos incluyen Hadoop, [NoSQL](https://es.wikipedia.org/wiki/NoSQL), Cassandra, business intelligence, machine learning y [MapReduce](https://es.wikipedia.org/wiki/MapReduce). Estas herramientas tratan con algunos de los tres tipos de BigData:

Datos estructurados: son datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las [bases de datos](https://es.wikipedia.org/wiki/Bases_de_datos) relacionales y los [almacenes de datos](https://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos).

Datos no estructurados: son datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los [PDF](https://es.wikipedia.org/wiki/PDF), documentos multimedia, [correos electrónicos](https://es.wikipedia.org/wiki/E-mails) o documentos de texto.

Datos semi-estructurados: son datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semi-estructurados​ que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Como ejemplos tenemos los archivos tipo [hojas de cálculo](https://es.wikipedia.org/wiki/Hojas_de_c%C3%A1lculo), [HTML](https://es.wikipedia.org/wiki/HTML), [XML](https://es.wikipedia.org/wiki/XML) o [JSON](https://es.wikipedia.org/wiki/JSON).

¿De dónde provienen todos estos datos? Se cataloga la procedencia de los datos según las siguientes categorías:

Generados por las personas: el enviar correos electrónicos, mensajes por [WhatsApp](https://es.wikipedia.org/wiki/WhatsApp), publicar un estado en [Facebook](https://es.wikipedia.org/wiki/Facebook), twittear  o responder a una encuesta por la calle son cosas que hacemos a diario y que crean nuevos datos y [metadatos](https://es.wikipedia.org/wiki/Metadatos) que pueden ser analizados.

Transacciones de datos: la facturación, las llamadas o las [transacciones](https://es.wikipedia.org/wiki/Transacciones) entre cuentas generan información que tratada pueden ser datos relevantes. El principal ejemplo son las transacciones bancarias, lo que el usuario conoce como un ingreso de dinero, la [computación](https://es.wikipedia.org/wiki/Computaci%C3%B3n) lo interpretará como una acción llevada a cabo en una fecha y momento determinado, en un lugar concreto, entre unos usuarios registrados.

Marketing electrónico y web: Con la [web](https://es.wikipedia.org/wiki/Web) 2.0 se ha roto el paradigma webmaster - contenido - lector y los mismos usuarios se convierten en creadores de contenido gracias a su interacción con el sitio. Existen muchas herramientas de seguimiento utilizadas en su mayoría con fines de [mercadotecnia](https://es.wikipedia.org/wiki/Mercadotecnia) y [análisis de negocio](https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_de_negocio&action=edit&redlink=1). Los movimientos de ratón quedan grabados en [mapas de calor](https://es.wikipedia.org/w/index.php?title=Mapa_de_calor&action=edit&redlink=1) y queda registro de cuánto pasamos en cada página y cuando las visitamos. Esto da origen a los llamados SmartAds, los cuales utilizan a la metadata como su base.

Máquina a máquina: son las tecnologías que comparten datos con dispositivos medidores, [sensores](https://es.wikipedia.org/wiki/Sensor) de temperatura, de luz, de altura, de presión, de sonido, etc. Que transforman las magnitudes físicas o químicas en datos. Algunos ejemplos son los [GPS](https://es.wikipedia.org/wiki/GPS) o los sensores de signos vitales en la medicina.

Biométrica: datos que provienen de la seguridad, defensa y servicios de inteligencia. Son cantidades de datos generados por lectores biométricos como escáneres de retina, escáneres de [huellas digitales](https://es.wikipedia.org/wiki/Huella_digital), o lectores de cadenas de [ADN](https://es.wikipedia.org/wiki/ADN). El propósito de estos datos es proporcionar mecanismos de seguridad y suelen estar custodiados por los ministerios de defensa y departamentos de inteligencia. Un ejemplo de aplicación es el cruce de [ADN](https://es.wikipedia.org/wiki/ADN) entre una muestra de un crimen y una muestra en nuestra [base de datos](https://es.wikipedia.org/wiki/Base_de_datos).



**Técnicas de análisis de datos**

Asociación: permite encontrar relaciones entre diferentes variables. Bajo la premisa de causalidad, se pretende encontrar una predicción en el comportamiento de otras variables.

Minería de datos ([data mining](https://es.wikipedia.org/wiki/Data_Mining)): tiene como objetivo encontrar comportamientos predictivos. Engloba el conjunto de técnicas que combina métodos estadísticos y de [aprendizaje automático](https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico) con almacenamiento en bases de datos. Está estrechamente relacionada con los [modelos](https://es.wikipedia.org/wiki/Modelo_estad%C3%ADstico) utilizados para descubrir patrones en grandes cantidades de datos.

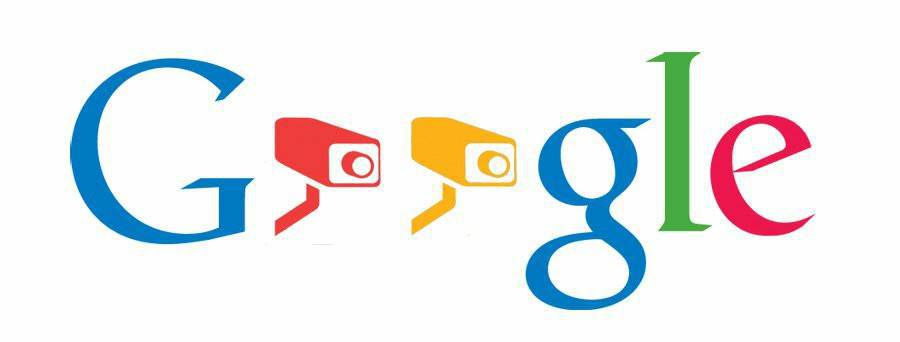
Agrupación ([clustering](https://es.wikipedia.org/wiki/Clustering)): tipo de [minería de datos](https://es.wikipedia.org/wiki/Data_Mining) que divide grandes grupos en grupos más pequeños de los cuales no conocíamos su parecido antes del análisis. El propósito es encontrar similitudes entre estos grupos, y el descubrimiento de nuevos, conociendo cuáles son las cualidades que lo definen. Es una metodología apropiada para encontrar relaciones entre resultados y hacer una evaluación preliminar de la estructura de los datos analizados. Existen diferentes técnicas y algoritmos de clusterización.

Análisis de texto (text analytics): gran parte de los datos generados por las personas son datos no estructurados, es decir, textos, correos electrónicos, búsquedas web, entre otros. Esta metodología permite extraer información de estos datos y así modelar temas y asuntos o predecir palabras.



**BigData y la privacidad**

Para tratar este tema es conveniente utilizar un ejemplo, el nuestro será Google.

Cuando utilizo Google para buscar algo en internet. ¿Google sabe quién soy? La respuesta aquí es fácil, sí, ya que tengo mi sesión iniciada. Pero ¿Y si no tengo mi sesión iniciada? Bueno, aquí entra en juego algo que ya mencionamos anteriormente y es nuestro perfil personal. Este es generado con nuestra información y es independiente de que tengamos o no nuestra sesión iniciada. Por lo tanto, en este caso Google también sabe quién soy.

¿Cómo hace eso?

Lo hace mediante la información que recopila de nuestro PC. Esta información puede ser el navegador que uso (Google Chrome, Mozilla Firefox, Safari, Edge, Internet Explorer, Opera), la versión del navegador, mi sistema operativo (Windows, Linux, Macintosh), mi versión de sistema operativo, mi dirección IP, mi resolución de pantalla, el horario en que suelo conectarme, el huso horario en el que estoy, mi ubicación, mis cookies, etc. Con toda esta información, Google genera un perfil mío y, por lo tanto, sabe quién soy independientemente de que decida iniciar sesión o no. A esto se le llama Fingerprint o huella digital.

Existen muchas páginas en internet que permiten saber tu fingerprint, otras dicen que tan únicas son las características de tu fingerprint y, por lo tanto, que tan rastreable eres. Mientras más raros sean estos datos, más fácil es reconocer al usuario. Esta es una herramienta muy usada para combatir el cibercrimen.

Sin embargo, algunas empresas explotan esta situación ofreciendo esta información a otras empresas, para que estas puedan incluir los datos en su BigData (por ejemplo, el caso de Facebook y Cambridge Analytica).

La aplicación de la fingerprint además de ser útil para la recolección de información para BigData sirve para detectar fraude, robo de identidad (muy útil en e-commerce), empresas de publicidad (Smart Ads).

Las Google Smart Ads funcionan con BigData, la información que Google utiliza para mostrar estas publicidades personalizadas esta disponible para el usuario y además, puede ser editada para que estas sean aún más personalizadas y precisas.

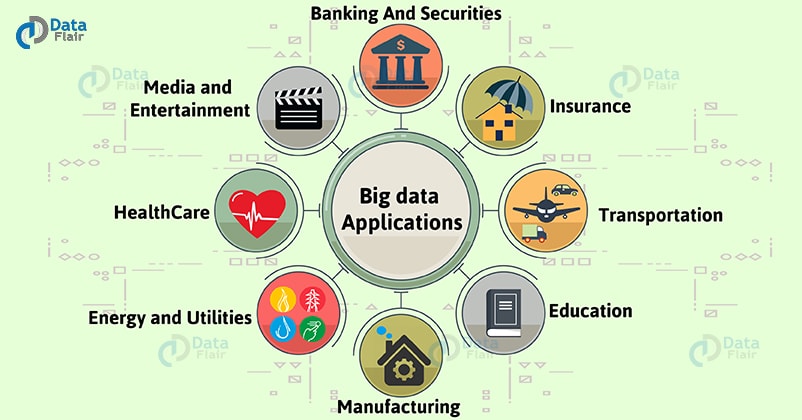
### 

### Conclusión

¿Hará el uso de BigData que nuestra privacidad se vuelva obsoleta? O traerá transparencia, responsabilidad y progreso con la recaudación y análisis de datos y la creación de patrones para entendernos mejor.

BigData tiene sus puntos a favor y en contra como todo, en el primer caso puede ser mostrándonos anuncios que sean más interesantes para nosotros o autos inteligentes que nos ayuden a evitar accidentes. Llamar a emergencias si nos encontramos en una, dispositivos “wearables” o implantables que nos ayuden a monitorear nuestra salud y notificar a nuestro doctor en el caso de que sea necesario. Pero, por otra parte, puede llevar a un gran problema en nuestra privacidad. Está claro que al generar toda esta cantidad de datos (donde vamos, que compramos, con quien nos comunicamos, qué miramos, cuando nos ejercitamos o dormimos y mucho más), somos muy vulnerables a cierto tipo de exposiciones. Es evidente que información tan detallada en las manos de instituciones financieras, empresas y gobiernos pueden afectar nuestro día a día, como por ejemplo nuestras relaciones, y hasta incluso a la hora de conseguir un trabajo. Y que, en manos de personas con malas intenciones, traería grandes consecuencias.

Muchos nos quejamos y debatimos sobre el tema de la privacidad en internet, pero no mucho se ha hecho para mejorar la situación actual.

Sin embargo, BigData demostró ser una tecnología muy potente y beneficiosa tanto para empresas como para usuarios.

**Bibliografía**

* Aden. Big Data y analítica para la toma de decisiones. YouTube (25 abr. 2018). <https://www.youtube.com/watch?v=Kb-ZpiGMIW4>
* DNews en Español. ¿Sabes qué es BIG DATA?. YouTube (12 oct. 2017). <https://www.youtube.com/watch?v=Ju2oDsHAL-o>
* Maligno Alonso. [2014] Big Data y Privacidad por Chema Alonso. YouTube (16 ene. 2016). <https://www.youtube.com/watch?v=_oeqjepFkEY>
* John P. Stevens. Why you need metadata for Big Data success. Data Science Central (6 abr. 2016).

<https://www.datasciencecentral.com/profiles/blogs/why-you-need-metadata-for-big-data-success>

* SupraPixel. GDPR | ¿Por qué todos actualizan ahora sus políticas de privacidad?. YouTube (29 may. 2018). <https://www.youtube.com/watch?v=1ZMtoMcQ9KM>
* Anónimo. Macrodatos. Wikipedia (5 jun. 2018).

<https://es.wikipedia.org/wiki/Macrodatos>

* World Economic Forum. What is big data?. YouTube (7 mar. 2016). <https://www.youtube.com/watch?v=eVSfJhssXUA>
* TEDx Talks. El internet de las cosas y Big Data | Theodore Hope | TEDxPuraVidaSalon. YouTube (15 nov. 2016).

<https://www.youtube.com/watch?v=xVqniTnMfQE>

* TEDx Talks. ¿Querés predecir el futuro? Usá datos | Nicolás Loeff | TEDxMontevideo. YouTube (21 jul. 2015).

<https://www.youtube.com/watch?v=1iqh1B1OZAg>

* [Techquickie](https://www.youtube.com/channel/UC0vBXGSyV14uvJ4hECDOl0Q). Big Data as Fast As Possible. YouTube (Aug 3, 2014).

<https://www.youtube.com/watch?v=CyWg5jBYIj0>

* Jeffrey Pomerantz. METADATA. The MIT essential knowledge series (2015).