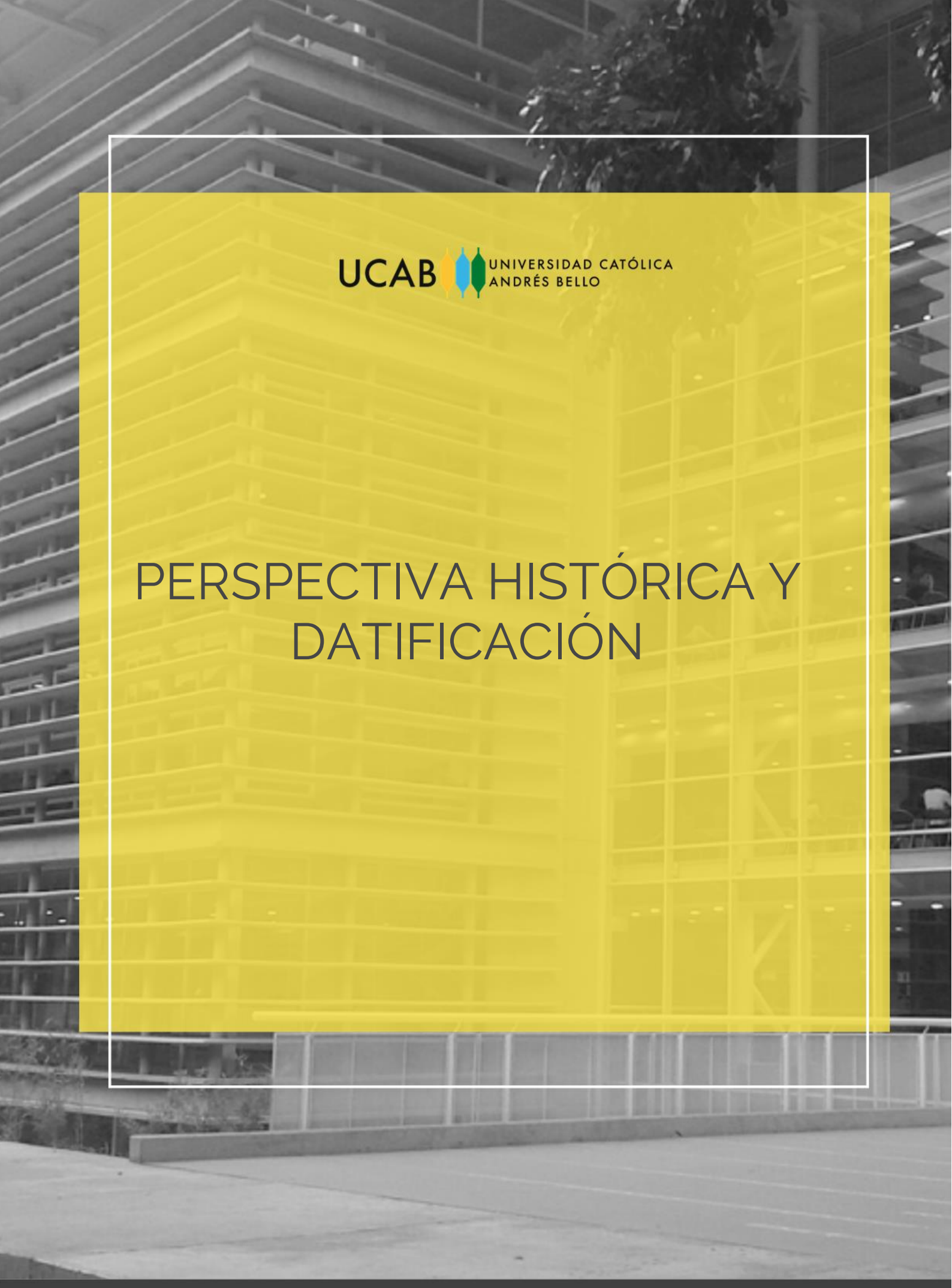


UCAB UNIVERSIDAD CATÓLICA ANDRÉS BELLO

# PERSPECTIVA HISTÓRICA Y DATIFICACIÓN

# Tabla de contenido

# Objetivo

- Conocer el origen de la gestión de grandes volúmenes de datos y comprender el término “ratificación” o “generación de datos”.

# Introducción

En gran medida, el término *big data* se emplea de forma generalizada desde finales de los años 90. Sin embargo, su historia es poco conocida, como sucede con las tendencias: cuando se hacen populares parece algo novedoso, pero realmente es el surgimiento de algo que ha estado madurando durante un tiempo.

# Orígenes de *big data*

Aunque hay expertos que sitúan la historia de *big data* en el paleolítico, relacionando el término con el primitivo interés de los seres humanos por obtener y procesar información, nos vamos a situar a partir de la década de los 50. Podemos decir que los desarrollos que influyeron en el *big data* son los siguientes:

En 1956, el físico Fritz-Rudolf Güntsch desarrolló el concepto de "memoria virtual", que trata el almacenamiento finito e infinito. De esta forma se podían procesar los datos sin las limitaciones de memoria de *hardware* que provocaban la partición del problema.

En 1958, el informático alemán Hans Peter Luhn definió el término "inteligencia de negocio".

En 1965 se proyectó el primer *data center* en Estados Unidos para almacenar más de 742 millones de declaraciones de impuestos y más de 175 millones de huellas dactilares en cintas magnéticas. Un año antes comenzaron a surgir voces que alertaron del problema de guardar la ingente cantidad de datos generada.

En 1970, con el surgimiento y uso masivo de las bases de datos relacionales por parte de las organizaciones que permitió separar los datos de las aplicaciones, y el desarrollo de las hojas de cálculo a partir de 1979, los analistas comenzaron a generar valor de los datos en las empresas. Sin embargo, los datos empezaron a estar en múltiples localizaciones y bajo diferentes formatos. Un primer acercamiento para resolver este problema fue por medio de reportes, pero estos en un principio eran muy planos.

Como solución a esto, hacia 1980 emergió el concepto de "almacén de datos" de la mano de Bill Inmon y Ralph Kimball, el cual permite centralizar y organizar la información. Ahora, con los datos organizados y centralizados, las necesidades de consumo de información por parte de las empresas aumentaron y nació toda un área de estudio denominada "Inteligencia de negocios", que reúne una serie de herramientas, arquitecturas y metodologías.

En el año 1989, Erik Larson utilizó por primera vez el término *big data* en un artículo sobre el marketing y habló de cómo se usarían los datos de los clientes en

los términos que actualmente conocemos. En torno a este año se empezaron a popularizar las herramientas de inteligencia de negocios.

En 1991 ocurrió la concepción de internet, a la postre, la gran revolución de la recolección, almacenamiento y análisis de datos. Tim Berners-Lee estableció las especificaciones de un sistema de red con interconexiones a nivel mundial accesible para todos en cualquier lugar.

En 1997 se registró el dominio ‹google.com›, un año antes de su lanzamiento, lo que inició el ascenso del motor de búsqueda al control y desarrollo de muchas otras innovaciones tecnológicas, incluidas las áreas de aprendizaje automático, *big data* y análisis.

En 1995 se construyó la primera supercomputadora que fue capaz de hacer tanto trabajo en un segundo de lo que puede hacer una calculadora operada por una sola persona en 30.000 años.

En 1996 la NASA se encontró con el problema de que no podía procesar una imagen del cielo nocturno en los supercomputadores de la época y, a raíz de ello, en 1997 los investigadores de la NASA, Michael Cox y David Ellsworth, publicaron un artículo donde afirmaban que el gran aumento de datos se estaba convirtiendo en un problema para los sistemas informáticos actuales. Esto se dio a conocer como el "problema del *big data*".

Basada en datos de 1999, la primera edición del influyente libro *How Much Information*, de Hal R. Varian y Peter Lyman (2000), intentó cuantificar la cantidad de información digital disponible en el mundo hasta la fecha.

En ese mismo año se hizo el primer uso del término *big data* en un trabajo académico: "Visually Exploring Gigabyte Datasets in Realtime" (1999). Asimismo, se empleó por primera vez el concepto *Internet of Things* (internet de las cosas) en una presentación de negocios de Kevin Ashton para Procter and Gamble.

# Siglo XXI

En el 2001, Doug Laney, de la firma de analistas Gartner, acuñó las 3 V (volumen, variedad y velocidad) que definen las dimensiones y propiedades de los grandes volúmenes de datos. Las V encapsulan la verdadera definición de *big data* y marcan el comienzo de un nuevo período en el que puede verse como una característica dominante del siglo XXI.

En el 2006, Doug Couting y Max Cafarella crearon Hadoop, a partir de los desafíos que presentó su motor de búsqueda y los artículos liberados por Google, "Google File System" y "Map Reduce". Este pasó a formar parte de los proyectos de Apache en el 2008.

En el 2005 nació la Web 2.0, una web donde predomina el contenido creado por los usuarios, promoviendo así la colaboración y la interacción entre ellos, suponiendo una evolución trascendental en el uso del internet y la generación masiva de datos de parte de cualquier usuario, estableciendo la relación actual de que el 80 % de los datos generados son del tipo no estructurado o semiestructurado y solo el 20 % son del tipo estructurado.

En 2008, las CPU del mundo procesaron más de 9,57 *zettabytes* (o 9,57 billones de *gigabytes*) de datos, aproximadamente lo mismo que 12 *gigabytes* por persona. La producción mundial de nueva información alcanzó un estimado de 14,7 *exabytes*.

Entre 2009 y 2011 aparecieron empresas como Cloudera y Hortonworks. Ambas fueron concebidas con el objetivo de conseguir una mejor gestión de los datos. Estos servicios abren un mundo de posibilidades para las empresas.

En 2011 McKinsey informó que para 2018 Estados Unidos enfrentaría una escasez de talento analítico. Necesitaría entre 140.000 y 190.000 personas con habilidades analíticas profundas y 1,5 millones de analistas y gerentes adicionales con la capacidad de tomar decisiones precisas basadas en datos.

Además, Facebook lanzó el Open Compute Project para compartir especificaciones para centros de datos energéticamente eficientes. El objetivo de la iniciativa era lograr un aumento del 38 % en la eficiencia energética a un costo un 24 % menor.

En 2012, la administración Obama anuncia la Iniciativa de Investigación y Desarrollo de *Big Data* con un compromiso de $200 millones, citando la necesidad de mejorar la capacidad de extraer información valiosa de los datos y acelerar el ritmo del crecimiento STEM (ciencia, tecnología, ingeniería y matemáticas), la seguridad y la transformación del aprendizaje. Este se usa por primera vez en política para la campaña de Barack Obama, conociendo las opiniones de los votantes.

Harvard Business Review nombró al científico de datos como el trabajo más sexy del siglo XXI. A medida que más empresas reconocieron la necesidad de clasificar y obtener información a partir de datos no estructurados, la demanda de científicos de datos se disparó.

En 2013, el archivo de mensajes públicos de Twitter en la Biblioteca del Congreso de Estados Unidos llegó a los 170 billones de mensajes, creciendo a un ritmo de 500 millones al día. El mercado global de *big data* alcanzó los $10 mil millones.

En 2014 los móviles superaron a los computadores en accesos a internet. La conexión casi continua contribuyó a generar muchos más datos y a mejorar la conectividad con otros dispositivos.

En el 2017 los datos llegaron a las masas. Gracias a esto, hoy en día la gente controla sus patrones de descanso con pulseras, sabe en qué se gasta el dinero con aplicaciones móviles y se informa sobre la posesión de balón de su equipo de fútbol. Los datos están en todas partes y la población está ya predispuesta a usarlos.

Finalmente, en el 2020 Allied Market Research informó que el mercado de *big data* y análisis de negocios alcanzó los 193,14 mil millones de dólares en 2019, y estimó que incrementará a 420,98 mil millones de dólares para 2027, a una tasa de crecimiento anual compuesta del 10,9 %.

Podemos decir, entonces, que *big data* ya no es una tecnología emergente, sino que ya está instaurada.

# Datificación

La datificación supone la generación y utilización de datos, ya sean originados por las actividades personales, pertenecientes a los procesos de trabajo de una organización, o provenientes de sensores para la medición de parámetros como la contaminación, el tráfico, las temperaturas, los terremotos, entre otros.

Según Mayer-Schönberger y Cukier (2013), "datificar" un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado. Ahora bien, esto es algo muy diferente de la digitalización, proceso en el que se convierte la información analógica en los unos y ceros del código binario para que los ordenadores puedan manejarla.

Para diferenciar la digitalización de la datificación vamos a tomar un ejemplo en el que se han producido ambos:

En 2004, Google anunció un proyecto de digitalización de cuantos libros pudiera y —en la medida posible en el marco de las leyes sobre propiedad intelectual— permitir a cualquier persona del mundo acceder a esos libros por internet y realizar búsquedas gratuitas en ellos. Para lograr esto se asoció con algunas de las mayores y más prestigiosas bibliotecas universitarias del mundo y puso a punto unas máquinas de escanear que pasaban automáticamente las páginas, de manera que el escaneado de millones de libros fuese al tiempo factible y económicamente viable.

Primero, Google digitalizó el texto: todas y cada una de las páginas fueron escaneadas y guardadas en archivos de imagen digital de alta resolución que se almacenaron en los servidores de la empresa. Cada página había sido transformada en una copia digital que podría ser fácilmente recuperada a través de la red por cualquier persona. Sin embargo, para recuperar esa información hacía falta o bien saber qué libro la contenía, o bien leer mucho hasta dar con el pasaje correcto. No se podían buscar unas palabras determinadas en el texto, ni analizarlo, porque el texto no había sido "datificado". Google disponía solo de unas imágenes que los seres humanos podían convertir en información útil únicamente leyéndolas.

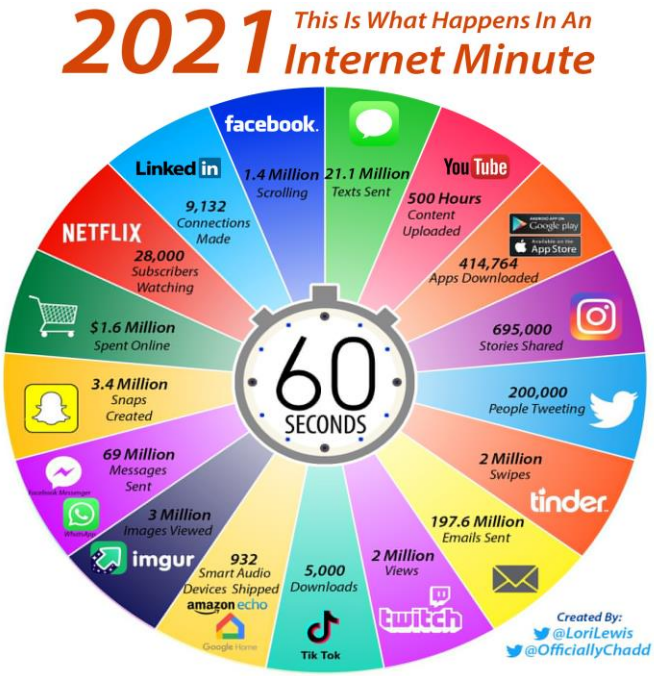
Aunque esto habría supuesto una gran herramienta de todas maneras —una biblioteca de Alejandría digital, más exhaustiva que ninguna otra antes—, Google quería más. La compañía comprendía que la información encerraba un valor que

únicamente se haría evidente una vez esta fuera datificada. Así que empleó un programa de reconocimiento óptico de caracteres que podía tomar una imagen digital e identificar las letras, palabras y párrafos presentes en ella. El resultado fue un texto datificado en lugar de una imagen digitalizada de una página.

Ahora, la información de la página podría ser utilizada no solo por lectores humanos, sino también por los computadores, que podían procesarla, y por los algoritmos, que podían analizarla. La datificación hizo que pudiera indexarse el texto y que, por consiguiente, pudieran hacerse búsquedas en él. Además, permitió un flujo inacabable de análisis textual: actualmente podemos descubrir cuándo se utilizaron por primera vez determinadas palabras o frases, o cuándo se volvieron populares, conocimientos que arrojan nueva luz sobre la diseminación de las ideas y la evolución del pensamiento humano a través de los siglos, y en muchos idiomas.

Entonces la "datificación" se trata de tomar un proceso o actividad, que antes era invisible, y convertirla en datos que puedan ser procesados por computadores.

Un segundo caso muy cercano de la datificación es más personal: nuestras relaciones, experiencias y estados de ánimo. La idea de la datificación constituye el interés central de muchas de las compañías de medios sociales de la red. Las plataformas de redes sociales no nos ofrecen solo una forma de localizar y mantener el contacto con amigos y colegas: también toman elementos intangibles de nuestra vida diaria y los transforman en datos que pueden usarse para hacer cosas nuevas. Facebook datificó las relaciones; siempre existieron y constituyeron información, pero nunca fueron definidas formalmente como datos hasta el "grafo social" de Facebook. Twitter permitió la datificación de los sentimientos al crear una forma fácil de que la gente anotase y compartiese sus pensamientos inconexos, que previamente se perdían en las brumas del tiempo. LinkedIn datificó nuestras experiencias profesionales pasadas, convirtiendo esa información en predicciones acerca de nuestro presente y futuro: a quién conocemos, o qué trabajo puede interesarnos. Seguidamente, puedes ver la famosa infografía publicada en Twitter por @lorilewis sobre lo que pasa en 60 segundos en las más famosas aplicaciones de redes sociales en internet.




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

Otro ejemplo de datificación es el internet de las cosas convirtiendo en datos los parámetros de luz, temperatura, entre otros, de una casa, una ciudad o tomando los datos de geolocalización o biométricos de una persona con dispositivos para tal fin.

La datificación tiene mucho que enseñarnos acerca de cómo funciona nuestro cuerpo. Dos profesores del Instituto de Investigación Tecnológica de Georgia, Robert Delano y Brian Parise, están poniendo a punto otra app llamada iTrem que utiliza el acelerómetro del teléfono para monitorizar los tremores corporales de una persona en busca de la detección de Parkinson y otras enfermedades neurológicas.

## Procedencia de los datos

Entonces, ¿de dónde proceden los datos? Existen datos que ya el negocio genera internamente en sus operaciones diarias o en su interrelación con los clientes; la mayoría se almacena en sus bases de datos, pero otros proceden de los procesos de negocio, así como de correos, presentaciones o documentos. De estos podemos decir que son generados y controlados por el negocio. Pero cada vez más datos provienen de fuentes ajenas, tales como medios sociales, portales web, nuestras actividades y conversaciones, del internet de las cosas, datos biométricos, los cuales generan los usuarios y manejan sistemas externos a la empresa.

# Cierre

Todo lo mencionado anteriormente hace que surjan una serie de preguntas:

- ¿En qué porcentaje estamos aprovechando los datos?
- ¿Estamos preparados para afrontar los retos de la ratificación?
- ¿Contamos con la infraestructura de *hardware* y *software*?
- ¿Contamos con personal especializado?
- ¿Qué conocimiento se genera a partir de los datos de interés?
- ¿Extrema información relevante que nos de ventajas competitivas?
- ¿Somos eficientes en la extracción de ese conocimiento?

Actualmente, estamos viviendo la cuarta revolución industrial, donde se considera vital la tecnología y la transformación de la información en conocimiento y nos está llevando a la transformación de nuestro entorno tal y como lo conocemos.

Este movimiento ha generado grandes cambios, tanto en particulares, con el aumento del uso de dispositivos y aplicaciones, como en empresas y organizaciones, que se encuentran en un proceso de adaptación al nuevo entorno tecnológico. En este paradigma, la gestión de grandes volúmenes de datos cobra una gran importancia.

Muchos expertos denominan al *big data* como el nuevo petróleo y cada vez más empresas están adoptando y avanzando lentamente hacia una cultura basada en datos.

# Referencias

Bryson, S., Kenwright, D., Cox, M., Ellsworth, D. y Haimes, R. (1999). Visually Exploring Gigabyte Datasets in Real Time. Communications of the ACM, 42(8), pp. 82-90. https://m-cacm.acm.org/magazines/1999/8/7840-visually-exploring-gigabyte-data-sets-in-real-time/fulltext?mobile=true

Mayer-Schönberger, V. y Cukier, K. (2013). Big Data, La revolución de los datos masivos. Turner Publicaciones.

Varian, H. y Lyman, P. (2000). How much information? University of California at Berkeley.

## Referencias de las imágenes

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