Big data: historia, definición, herramientas y aplicaciones en la industria

Mauricio Toro
Ingeniero de sistemas, Pontificia Universidad Javeriana, Cali, Colombia
Doctor en Informática, Université de Bordeaux, Bordeaux, Francia
Profesor del Departamento de Informática y Sistemas, Universidad EAFIT, Medellín, Colombia
mtorobe@eafit.edu.co

Henry Laniado
Licenciado en Física y Matemáticas,Universidad de Antioquia, Medellín, Colombia
Magister en Ingeniería Matemática, Universidad Carlos III de Madrid, Madrid, España
Doctor en Ingeniería Matemática, Universidad Carlos III de Madrid, Madrid, España
Profesor del Departamento de Ciencias Matemáticas, Universidad EAFIT, Medellín, Colombia
hlaniado@eafit.edu.co

Dan Ariely, profesor de psicología y economía de la Universidad de Duke, plantea que "El *big data* [en español, datos masivos] es como el sexo en la adolescencia: todo el mundo habla de ello, nadie sabe realmente cómo hacerlo, todos piensan que los demás lo están haciendo, así que todos dicen que también lo hacen…". En lo sigue de este escrito, vamos a relatar la historia de los datos masivos, brindar una definición, explicar algunas técnicas útiles para procesarlos, mencionar algunos contextos actuales de aplicación y algunas oportunidades de negocio.

Historia de los datos masivos

Las sociedades humanas pasaron de ser cazadoras y recolectoras a agricultoras y ganaderas durante el periodo neolítico, aproximadamente en el año 15 000 A.C. Naturalmente, apareció la necesidad de cuantificar. Inicialmente se utilizaron marcas en piedras y árboles. Otra forma de contar era, por ejemplo, almacenar en una bolsa una piedra por cada oveja que tuvieran: cuando mataban una, botaban una piedra de la bolsa; cuando compraban otra, agregaban una piedra a la bolsa. En nuestros días, con la aparición de los computadores electrónicos, se hizo posible almacenar muchos más datos y procesarlos de una manera mucho más rápida. Tomaría mucho tiempo efectuar los cálculos que se

realizan en la actualidad con el mecanismo de almacenar piedras en una bolsa o hacer marcas en un árbol; además, se necesitarían más árboles de los que existen en el planeta para lograrlo.

Según Eric Schmidt, director general (CEO) de Google, en 2010 "había cinco exabytes de información creados desde el amanecer de la civilización hasta el 2003, pero ahora la misma cantidad [de información] se crea cada día". Para entender esta cifra, recordemos que 1000 gigabytes equivalen a un 1 terabyte, 1000 terabytes corresponden a 1 petabyte y 1000 petabytes son 1 exabyte.

De los datos que hoy posee la humanidad, solo una pequeña fracción de ellos son información que las empresas tienen en forma de clientes, productos y ventas. Ya existen soluciones informáticas para este tipo de datos desde el siglo XX. Un porcentaje mayor, para el cual la tecnología actual es insuficiente para procesarlos con la velocidad que se requiere, es la información de las redes sociales compuesta por texto, video y fotos. La cantidad de cuentas en redes sociales en el presente excede la población mundial. Finalmente, están los datos producidos por sensores conectados a dispositivos que, a su vez, se conectan a Internet. En este último fragmento se encuentra una porción mayor de datos que los datos empresariables y los de las redes sociales juntos. Para procesarlo a la velocidad requerida, las herramientas con las que se cuenta también se quedan cortas.

Para darnos una idea de qué tan grandes son los datos que existen hoy en día, EMC Corporation e IDC encontraron en el año 2013 que el 90 % de los datos que tenía la humanidad se habían creado en los últimos dos años. En ese entonces, existían alrededor de 3000 exabytes y ellos pronosticaron que habrá alrededor de 40 000 más en 2020. Recordemos que 1 exabyte es un millar de millones de gigabytes. Desafortunadamente, solo alrededor del 0,5 % de esos datos se procesan. Es por esa razón que muchos gurús de la tecnología plantean que los datos serán el petróleo del siglo XXI. Es claro que esta es una gran oportunidad de negocio para las empresas colombianas.

Hablar de datos masivos implica alguna de estas características: gran tamaño, alta dimensión, alta frecuencia o una combinación de las anteriores. Como un ejemplo, datos de gran tamaño pueden ser la hora y lugar de las búsquedas del término *big data* en Google. De la misma forma, datos de alta dimensión pueden ser aquellos para los que tenemos muchos atributos para cada uno, por ejemplo, una tabla donde se muestren las películas que ha visto cada ser humano a lo largo de su vida. Finalmente, un ejemplo de datos de alta frecuencia son las mediciones de temperatura, humedad y presencia de

gases que realiza una empresa de manufactura varias veces por segundo, 24 horas al día y 7 días a la semana sin pausa.

Definición de datos masivos

De acuerdo con Francis X. Diebold, el término *big data* (datos masivos) se originó a partir de conversaciones en Silicon Graphics Inc. (SGI) a mediados de 1990. Comenzó a utilizarse de modo masivo en 2011 en muchos países. Actualmente, hace referencia a una cantidad de datos tal que supera la capacidad del software convencional para ser capturados, administrados y procesados en un tiempo razonable y. que por el momento, es difícil analizarlos utilizando herramientas tradicionales.

En 2001, Doug Laney introdujo las tres V de los datos masivos: volumen, variedad y velocidad. Posteriormente, se agregaron otras dos características: veracidad y valor. Por otro lado, Mayer Schonberger plantea que el valor no reside en los datos, sino en la forma de correlacionarlos para descubrir patrones. También afirma que, para procesar datos masivos en un tiempo razonable, tenemos que estar dispuestos a tolerar una imprecisión mayor a la que aceptamos cuando procesamos datos de menor tamaño.

Algunas técnicas para procesar datos masivos

Para poder procesar datos de gran tamaño en un tiempo razonable, una de las técnicas que se utiliza es la reducción de tamaño mediante *wavelets*. Esto permite disminuir la magnitud de los datos. Aunque ocasiona que se pierda información, se ha encontrado que se pueden hacer predicciones con datos reducidos mediante *wavelets* con resultados aceptables para las preguntas de negocios que se tienen.

Para procesar datos de gran dimensión en un tiempo razonable, se emplea una técnica conocida como análisis de componentes principales, la cual posibilita aminorar las dimensiones de los datos. De igual manera, se han encontrado muchos casos de éxito en que se pueden hacer predicciones con datos cuya dimensión ha sido sometida a este proceso, con resultados aceptables para las preguntas de negocios que se tienen.

Finalmente, para hacer regresiones se usan técnicas como la regresión Lasso (inventada en 1996), la cual hace posible, por ejemplo, encontrar cuáles variables son significativas y cuáles no en el

contexto de una regresión multivariada, con un número enorme de dimensiones y una cantidad reducida de datos. Verbigracia, imaginemos que tenemos todas las calificaciones que una muestra de estudiantes obtuvo en un colegio; son más de 1000 por cada uno y queremos saber cuáles de ellas son significativas para predecir su desempeño en la universidad. La regresión Lasso es una técnica muy apropiada para aplicarse en este problema.

Aplicaciones en la industria de los datos masivos

El procesamiento de datos masivos tiene aplicaciones en todas las áreas del conocimiento, desde la literatura, la salud y la psicología hasta la biología, las matemáticas y la ingeniería. Para efectos de este escrito, nos vamos a centrar en las aplicaciones de la industria, en particular, en redes sociales y mercadeo.

Las redes sociales tales como Facebook, Twitter o Google+ se han convertido en los últimos años en los servicios digitales más utilizados en el mundo. Cerca de 1000 millones de usuarios interactúan intensivamente en ellas cada día. Esto hace que las redes sociales sean un recurso invaluable para quienes trabajan con publicidad, mercadeo o política, pues utilizarlas puede servir para recolectar información y para lanzar campañas. Analizarlas son una oportunidad de negocio importante en Colombia. Un reto importante es la identificación de los usuarios más influyentes[1], lo cual es importante para publicitar un producto, propagar un mensaje o mejorar la imagen de una compañía.

Otras oportunidades que tiene el uso de los datos masivos en el mercadeo son la segmentación automática de mercados y el encontrar correlaciones en los comportamientos de los consumidores que puedan ser utilizados para crear nuevas estrategias de mercadeo. Como un ejemplo de este último enfoque, dos empresas de comercio minorista que se han beneficiado son Wal-Mart y Target.

En la década de los años noventa, Wal-Mart encontró una correlación estadísticamente significativa entre la compra de cerveza y pañales. Descubrió que los consumidores eran en su mayoría varones entre 25 y 35 años. La compañía decidió colocar la cerveza frente a los pañales. Los resultados fueron espectaculares. No solo aumentaron las ventas en un 15 % tanto en cerveza como en pañales,

[1]A. Azcorra, L.F. Chiroque, R. Cuevas, et al, Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks", *Scientific Reports* 8 (2018).

sino que, además, hubo un cambio en los comportamientos de compra de los padres que compraban los pañales: al verlos, recordaban la falta de cerveza en casa.

Por su parte, en 2012 Target infirió estadísticamente el embarazo de una cliente adolescente, antes de que sus padres se dieran cuenta. ¿Cómo fue posible sacar conclusiones del comportamiento de sus clientes? Esta empresa encontró una forma de asignar una probabilidad al embarazo de una cliente a partir de información histórica, por ejemplo, hábitos de compras e información personal.

Datos masivos en Colombia

Existe un gran potencial para las empresas colombianas de mejorar sus ventas valiéndose de estrategias como las que Wal-Mart y Target utilizaron. Esta es una área fascinante y con un gran número de oportunidades de negocio. Sin embargo, para poder procesar datos masivos es necesario tener personas formadas en este tema. Pocas universidades colombianas ofrecen programas de posgrado en esta área. Existen tres en Bogotá —Maestría en Analítica para la Inteligencia de Negocios, Pontificia Universidad Javeriana); Maestría en Inteligencia Analítica para la Toma de Decisiones, Universidad de los Andes; y Maestría en Ingeniería y Analítica de Datos, Universidad Jorge Tadeo Lozano— y dos en Medellín —Especialización en Analítica, Universidad Nacional; Maestría en Ciencias de los Datos y la Analítica, Universidad EAFIT—. Además, dos universidades brindan cursos cortos: la Universidad Externado de Colombia (Big Data y Analítica de Datos) y la Universidad EAFIT (Analítica de Datos para la Toma de Decisiones Empresariales).