BIG DATA

Ignacio José MOREU MUNAIZ

IVIMOS en un mundo donde la interconexión digital ha permitido el almacenamiento masivo de datos. Desde hace mucho tiempo, estos solo servían para un uso inmediato, de tal modo que no se requería almacenarlos ni trabajar con ellos una vez finalizado el uso para el que se crearon. Este es el caso de una factura, que una vez se descontaba de existencias y se cobraba, se almacenaba en papel, exclusivamente para una posible inspección fiscal.

Con las tarjetas de crédito empezaron a aparecer datos digitales más voluminosos. El pago de un servicio ya no era en efectivo, sistema por el que no existía ningún dato almacenado de forma digital, y el pago virtual o por tarjeta suponía la necesaria anotación digital en un banco, y por consiguiente una información más voluminosa, pues con cada pago se almacenaba no solo este, sino también quien lo realizaba, en qué punto de venta, fecha y hora…

Posteriormente, la aparición de internet supuso un cambio en el almacenamiento de datos. En un principio, solo se podía navegar si uno conocía de antemano cuál era la dirección del sitio al que deseaba conectarse. Para resolver este problema, surgieron los buscadores. Podemos recordar los primeros, que lo hacían de una forma poco precisa, como fueron *Altavista, Yahoo, Lycos*… A finales del siglo pasado, apareció *Google*, obteniendo unos resultados impresionantes, y que prácticamente ha quedado como herramienta cuasi monopolística (ronda el 70 por 100 de las búsquedas en internet), obteniendo resultados grandiosos gracias a un algoritmo que indexaba todo el contenido de la *web,* de tal modo que casi dejó de ser necesaria la utilización de los metadatos.

Pero qué significa esto; puedo encontrar un documento por cada una de las palabras, acrónimos… que aparecen en dicho fichero. Anteriormente, los buscadores utilizaban temática y clasificación. Esta utilización de datos que

acompañaban al documento para proporcionar un índice de palabras clave, como pueden ser los antiguos introducidos al final de los libros, era el método principal para que pudiese ser localizado en la *web*. *Google* se atrevió a lo que nadie en esa época pensó que podría hacerse, indexar todo el fichero en inmensos índices que permiten localizar un documento con una simple palabra. Y no solo con esto, sino con información añadida por situación actual, temática de moda… de tal modo que un par de palabras permiten localizar un vídeo, documento o artículo que tratan sobre lo deseado por el usuario. Si en marzo 2018 se busca «gas espía», la primera noticia es la muerte del espía ruso en Inglaterra; si se pone simplemente «espía», este episodio será la octava referencia.

Este algoritmo requirió de una gran capacidad de almacenamiento para poder guardar enormes cantidades de datos que permitían localizar casi cualquier cosa desde cualquier sitio. Es necesario recordar que esto pudo ser posible gracias al abaratamiento del almacenamiento de los mismos. Desde los primeros ordenadores domésticos (Spectrum), con 16 Kb de memoria, hasta nuestros días, la capacidad de memoria se ha multiplicado por 100 millones.

En 1994 un ordenador de última generación disponía de un disco de 20 a 40 Mb, siendo su coste la mitad de un coche utilitario (medio millón de las antiguas pesetas). En 1998 los discos duros que se vendían eran ya de dos gigabytes (25 veces superiores en capacidad) a un precio bastante menor.

A partir de este crecimiento exponencial, se ha multiplicado la captación de datos, tanto en producción como en almacenamiento, lo que ha permitido hablar de diferentes conceptos en su análisis.

Con la popularización de internet en los años noventa, la velocidad de 56 Kbps mediante línea telefónica era insuficiente para visitar páginas *web*. En un año se aumenta la velocidad con el ADSL a 256 Kbps. Aparecen los primeros programas que van a manejar grandes cantidades de datos (*Gmail, YouTube, Google Earth, Facebook…*).

A esto se le suma la aparición de los primeros móviles (1994 IBM) residuales, pero que abrieron un campo que ha generado una cantidad ingente de datos. Ya no se necesita estar conectado al ordenador para proporcionarlos. Desde el metro, tren o coche se generan de forma ingente, desbordando toda cifra imaginable. Esto empieza a producirse en 2007 con el iPhone de Apple (Android apareció un año después), finalizando con las tabletas en 2010 (de la mano de Apple), a lo que se suman la cámara, el GPS, aplicaciones… incorporados al móvil.

Resumiendo, de los 7.500 millones de habitantes del planeta Tierra, la mitad está conectada a internet (1) (en el último año el número de usuarios

(1) *https://www.internetworldstats.com/stats.htm. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.*

creció un millón por día), un tercio utilizan redes sociales y el número de móviles supera el de la población mundial (2).

Facebook es la aplicación más popular y en Estados Unidos el 80 por 100 de los usuarios de internet lo son de *Facebook* (un cuarto de la población mundial). Le sigue *Instagram, Pinterest, LinkedIn, Twitter*… El número de búsquedas diarias en *Google* es de 100.000 millones.

Casi un tercio de los usuarios de *Facebook* tienen entre 25 y 34 años. Suben 300 millones de fotografías al día y pasan conectados 20 minutos de media diarios; tienen 155 «amigos», cada 20 minutos comparten un millón de *links*, mandan 20 millones de solicitudes de amistad y envían tres millones de mensajes.

Lo mismo se puede decir de aplicaciones como *Snapchat* (161 millones de usuarios activos en EE. UU. y Canadá, que comparten 400 millones de *snap* y 9.000 fotos por segundo), *Twitter* (328 millones), *Instagram* (en 2017 se compartieron 40.000 millones de fotografías y 4.000 millones de «me gusta» al día).

Todas las aplicaciones proporcionan además información adicional, como es la ubicación del usuario. La enorme cantidad de datos se almacena de forma temporal, ya no es información que se almacene durante largo plazo.

Existen otras fuentes de información que proporcionan millones de datos, desde controles de máquinas, conexiones de búsqueda por internet, cámaras de televisión en carreteras, datos meteorológicos, compras *on line*, fotografías y vídeos.

En resumen, la tecnología ha permitido procesar y almacenar una cantidad ingente de datos. Ahora se pretende analizar estos datos y explotar estos para obtener un beneficio añadido.

Existe multitud de vocabloss para un mismo concepto. El mundo comercial demanda cambiar el léxico para poder vender el mismo producto. Uno de los primeros términos fue el de *data business*. En el año 1977 John Tukey acuña el de *data analysis (Exploratory Datta Analysis)*. Y de este análisis de datos parte, a mi modo de ver, toda una teoría que va desarrollándose conforme las necesidades cambian y, sobre todo, cuando las empresas necesitan vender nuevas herramientas o, mejor dicho, versiones de las antiguas. Este análisis trata de depurar los datos, transformarlos y buscar conclusiones que sean de interés para la toma de decisiones.

De aquí aparece el vocablo Minería de Datos, que es lo mismo pero proporciona predicciones según la fuente que se lea. *Business Analytics*, que conforme se analiza el pasado permite pronosticar el futuro; Inteligencia Empresarial, para extraer conocimiento de los datos… En fin, un conjunto de términos que identifican lo mismo. Se trata de, a partir de grandes cantidades

(2) *https://www.internetworldstats.com/mobile.htm*.

de datos, buscar la lógica que nos ayude a predecir el futuro, a optimizar un proceso, con técnicas diferentes en función de cómo sean estos, de dónde provengan, qué se interprete con ellos.

Y entonces llegamos a los palabros de moda, Inteligencia Artificial (AI) y *Machine Learning* (ML), que hoy en día no pueden faltar si se habla de datos. Como bien dice su nombre, se trata de aprender conforme se utiliza el algoritmo. De forma resumida, se puede decir que ML pretende predecir el futuro a partir de lo que ya ocurrió en otras ocasiones. Pero qué es aprender. Desde el punto de vista del ML, está muy relacionado con determinados problemas y no de lo que podemos deducir por aprendizaje.

En este mundo, se entiende por aprendizaje cuando la máquina aprende mediante la experiencia. Y esto cómo se hace. Pues entrenando a un algoritmo con casos para que pueda predecir el siguiente, como en las redes neuronales. Difícil de explicar y bastante más de entender.

Pongamos el siguiente ejemplo: queremos adiestrar una red neuronal que pretenda localizar una cara a partir de unos parámetros biométricos (nariz, mentón, separación ojos…), que se introducen en una base de datos (BD). Para conseguir esto, se requiere entrenar a la red neuronal ajustando cada uno de los nodos para que localice la cara correcta. Después del ajuste con una gran BD, se espera que al dar los datos biométricos de un nuevo rostro, acierte con la identificación del individuo; en caso contrario se volvería a ajustar el valor de los nodos. Así, con cada acierto se afianza el valor de los nodos de la red y con cada fallo se modifican los valores de los mismos hasta llegar a un ajuste donde el sistema funcione.

Existen otros métodos de ML, como son los basados en técnicas estadísticas, donde el análisis de datos permite determinar parámetros estadísticos para poder hacer predicciones. *Data Mining* y *Machine Learning* están relacionados, y en ambos casos se utilizan análisis de datos.

Lo cierto es que toda fuente de datos requiere un proceso de depuración y transformación. Una vez corregidos, comprobados, se podrán correlacionar con otros. Esta primera fase supone un gran trabajo. De hecho, muchos expertos consideran que esta tarea supone el 80 por 100 del trabajo. Una vez que tenemos los datos depurados, se procederá a aplicar la técnica necesaria para obtener conocimiento. En función de qué datos sean, de dónde provengan, qué representen y qué se desea obtener, se aplicara una u otra.

Existen otros términos, pero todos tienen prácticamente la misma idea: KDD (*Knowledge Discovery in Databases*); otros para referirse más a las bases de datos que al análisis (*Data Warehouse, Cubos OLAP, OLTP…*) y todo para tratar los datos y sacar conocimiento de ellos. De este modo, es cuando se llega al *big data*.

Big data

Se trata de *big data* cuando los datos cumplen las llamadas «3V», esto es, cuando el conjunto de datos por su tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento y análisis mediante técnicas convencionales.

Como lo que gusta en complicar y poner palabros o hacer virtuosos juegos de palabras, ya se habla de las «5V», añadiendo a las anteriores la veracidad y el valor. En realidad, en un mundo donde los datos crecen a este ritmo, pedir veracidad es un brindis al sol. Y el valor que, como siempre, se presupone.

Tamaño

Dentro del *big data*, el tamaño que determina una BD no está definido, pero se puede considerar que va desde los 30-50 Terabytes hasta los Petabytes (10^{15} bytes). Y el mayor problema radica en el crecimiento más que en su tamaño, pues este puede ser tan grande que se haga inviable poder almacenar toda la información durante un período largo. Aunque los límites en la capacidad crezcan, también lo hacen los datos, de tal modo que nunca podrán almacenarse, e incluso cuando afiance el Internet de las Cosas, y es más que posible que los datos que se transmitan por internet crezcan en un grado o más que la capacidad de almacenamiento.

Complejidad

Muchas fuentes y tipos de datos complican la integración de estos. Se entiende por datos estructurados aquellos que se pueden guardar en tablas. Nos referimos a las BD relacionales, donde las tablas tienen relaciones entre ellas. Cuando los datos que se guardan se pueden parametrizar, es decir, la información se guarda con un conjunto finito de posibilidades, por ejemplo, el empleo (grado militar) es un dato parametrizado, existe un conjunto de empleos finito y fijo. Sin embargo, el nombre no es un conjunto finito y con cada individuo nuevo aparecerá un nombre nuevo. Esto implica que no es un parámetro, pero el nombre está estructurado, es decir, se almacena en un único campo, de tal modo que no tengo que buscar en el campo ciudad u observaciones para localizar el nombre de alguien.

Que los datos estén estructurados tiene muchas ventajas a la hora de buscar, calcular… Pongamos que tenemos en una BD un campo coste. Sabemos que sumándolo obtendremos el precio del conjunto de registros que consideremos, y siempre sabremos el coste de un registro. Si por el contrario,

buscamos en documentos, puede aparecer varias veces el coste de un mismo concepto, es más, puede haber varios documentos que hablan del mismo coste, o incluso que en un documento se desarrolle el coste en costes parciales o que existan documentos repetidos. A la hora de calcular el coste, tendremos que implementar algoritmos que determinen si son o no costes, si ya están imputados, si son añadidos. De tal modo que complicamos la obtención de datos.

Como ya se explicó al principio, la aparición de datos no estructurados (vídeos, fotos, audios) ha crecido exponencialmente y, llegados a este extremos, hay quien considera los documentos información cuasi estructurada, lo que no comparto en absoluto.

Los datos estructurados proporcionan un grado de veracidad que no aportan los no estructurados. Tengamos en cuenta que un registro que tiene un coste tiene también el código que identifica ese concepto, y en las BD estructuradas suele ser un código clave que impide que pueda estar duplicado.

Los que tratan de explicar qué es el *big data*, nos dicen que está formado por la mezcla de datos estructurados junto a otros no estructurados (radiofrecuencias, *web log*, sensores de equipos, redes sociales, búsqueda en internet, GPS, móviles). El mundo comercial propone utilizar BD estructuradas junto a aplicaciones ERP (*Enterprise Resource Planning*) o un CRM (*Customer Relationship Management*). Para entendernos, estas serían nuestras SIPERDEF, GALIA… Lo que no dice es que de las BD estructuradas se extrae la información principal, mientras que del resto el sistema tratará de buscar información para acompañar a la información estructurada.

Lo que pretende el sistema es buscar esta información difuminada para incorporarla a la estructurada para poder realizar un análisis de datos «tradicional», buscando patrones, tendencias, inferencias estadísticas…

De aquí salen herramientas como *Hadoop*, que es un *software* libre que permite el acceso a sistemas de archivos, que nos facilita leer de una forma distribuida y realizando el trabajo en los nodos que contiene la información. Lo complicado es que posibilita ejecutar su trabajo entre muchos ordenadores, como si se tratase de una única gran máquina. Pero hasta aquí es lo que es *Hadoop*; lo que lee y lo que busca, interpreta, convierta en información y está en conocimiento dependerá de la aplicación que se le ponga por encima. Si *Hadoop* es complejo, lo que le queda por hacer lo es más.

Velocidad de crecimiento de datos

El ritmo de producción de datos en los *big data* es tan grande que impide el proceso de captación, tratamiento y análisis pues, aunque incrementásemos la capacidad de procesamiento, la de crecimiento superaría al incremento de proceso.

Es muy importante que el sistema de recogida y depuración de datos esté automatizado y se acuda a sus fuentes con procesos independientes para captarlos y depurarlos. Proceso complejo es ir incorporando fuentes de información para que puedan ser «atacadas por los buscadores» y obtenerla con valor o datos concretos.

Uno de los aspectos de los que se habla en el *big data* es la gobernanza, palabro muy de moda, que según IBM es «una disciplina encargada de la orquestación de gente, procesos y tecnología que permite habilitar a una compañía a apalancar la información como un recurso de valor empresarial, encargada de mantener a los usuarios, auditores y reguladores satisfechos…». Es decir, todo y nada, sirve para datos, facturación… En resumen, es una frase más. Lo que está claro es que en un sistema donde todo es incierto por culpa de la complejidad, velocidad y tamaño, incluir más incertidumbre acerca de quién es el propietario de los datos, quién puede consultarlos y modificarlos se contrapone con un sistema donde el dato tiene una vida efímera y donde lo que se pretende es tomar el pulso a una ingente cantidad de estos, que quizá no podamos almacenar más allá de un mes.

La «gobernanza de datos» debe aplicarse a bases de datos estructuradas, en las que se almacena la información para explotación y donde deben existir políticas de seguridad y accesibilidad, pero no sobre lo que es incierto y descontrolado.

Según Boris Evelson, analista de Forrester Research Inc. en Cambridge, Massachusetts, «*Big data* es un área tan nueva que nadie ha desarrollado procedimientos y políticas de gobierno... Los datos no se pueden gobernar hasta que se modelan, pero no se pueden modelar tampoco hasta que se exploran [por los analistas de datos]».

Debemos entender el *big data* como algo en cierto modo inmanejable, de tal modo que lo que se pretende es ir analizando información de distintas fuentes, muchas veces repetidas, redundantes, copias y recopiadas, reenvidas una y mil veces, y que no podrá ser procesada por falta de capacidad; otra que requiere de una interpretación (como pueden ser el audio y el vídeo) para extraer información que pueda relacionarse posteriormente con la estructurada de una BD en otro punto distante, con políticas de seguridad.

Así, *big data* sería parecido a lo que se describe en la siguiente historia: una carretera, con muchas cámaras, va reconociendo matrículas; conforme realiza esta labor, una de las ellas se cruza con una BD de coches y aparece uno robado. El sistema detecta la posición de la cámara, hora de la toma y calcula cuál puede ser la siguiente cámara por la que pasará el vehículo, de tal modo que intentará reconocer la cara del conductor en la siguiente que esté en una ubicación más baja y que sea de más resolución. A continuación, se toman los parámetros biométricos de la cara y se cruzan con una BD que localiza si pertenecen a un delincuente buscado. Si esta búsqueda es positiva y localiza a un posible terrorista, todas las cámaras del resto de carreteras se ponen a

buscar la matrícula para detectar su posición, pasando esta información a la policía, cuyo coche más próximo será alertado para realizar la detención del vehículo sospechoso.

Este ejemplo será un sistema real dentro de no mucho tiempo. Desconozco si será un avance positivo o negativo pues, además de poder localizar al terrorista, también podrá detener a un padre que riñe a sus hijos cuando el *software* haya entendido que ha sido un poco agresivo. De hecho, hace poco se habló de un sistema de vigilancia masiva en China. Sin dudarlo, el *big data* ya está en funcionamiento.

Pero pisando tierra, si pensamos en las Fuerzas Armadas, más concretamente en la Armada, ¿qué bases de datos se pueden considerar *big data*? Descartando la Ciberdefensa, en la actualidad no hay nada que tenga las características del *big data*. Estamos alejados de esta necesidad. Pongamos por caso el problema de analizar los sensores de los buques. Supongamos que registramos 1.000 en cada buque y que obtenemos la medida cada segundo. Estaríamos hablando de 3.600 Kb de datos a la hora. Si estos requieren una media de 10 bytes (para seguir el sistema decimal más intuitivo), tendríamos 36 Mb de información a la hora. Algo que es más que factible de analizar, depurar y almacenar. Aunque se hablase de 30 buques simultáneamente, seguiríamos con un gigabyte de datos a la hora. Debemos entender que esto no es *big data*, sino datos estructurados con un tamaño grande, pero no más.

De hecho, lo que tenemos que obtener son datos para ir guardándolos y sacar conclusiones en un futuro. Algo que a día de hoy no tenemos, y nuestra realidad es que poseemos pocos datos almacenados. Por ejemplo, saber lo que ha costado un sistema determinado es prácticamente imposible. Y de conseguirlo no creo que sea utilizando *big data*, sino analizando documentos hasta llegar a una aproximación.

Pensar en un *big data* que analice nuestros documentos para obtener inteligencia no es imposible, pero los resultados no serán certeros. El desarrollo será costoso, y seguiremos sin datos. Lo que se debe pensar es en dejar de producir «papel» para almacenar datos. Esto nos permitirá en un par de años tener información y conocimiento del funcionamiento de la Armada.