# Pasado, presente y futuro

POR **MARIO TASCÓN**

El autor introduce el Dossier con una alusión a diversos conceptos que guardan una estrecha relación con el *Big Data*, como *Business Intelligence*, minería de datos u *Open Data*. Asimismo, frente a las denominadas '3V' que constituyen la esencia del *Big Data* (volumen, variabilidad, velocidad), sugiere una cuarta V, la de la visualización. Finalmente, tras anticipar la previsible problemática en torno a la gestión de la privacidad de la información, concluye con una reflexión acerca de la dimensión del concepto *big* en un contexto de crecimiento exponencial de la información.

*Big Data* apareció el pasado año como uno de los términos de moda en todas las revistas de temática científica, sociológica o tecnológica, también en *blogs* y redes sociales e incluso ya ha dado el salto a las publicaciones económicas y empresariales y las de divulgación más popular.

¿Va a ser *Big Data* una etiqueta más que añadir a las múltiples modas que hemos ido viendo y consumiendo a lo largo de los últimos años en el panorama de Internet y los desarrollos digitales o es una tendencia de fondo que está afectando en su totalidad a la evolución de la Web? Esta es una de las principales preguntas a las que intentan responder los artículos de este Dossier que la revista TELOS pone en sus manos.

*Big Data* (o Macrodatos en castellano, si seguimos las recomendaciones de la Fundación del Español Urgente, Fundéu BBVA) es, sin la menor duda, uno de los campos más importantes de trabajo para los profesionales de las TIC. No hay área ni sector que no esté afectado por las implicaciones que este concepto está incorporando; cambian algunas herramientas, se modifican estrategias de análisis y patrones de medida.

*Big Data*. Aunque su nombre hace referencia a la cantidad -a la enorme cantidad, para ser más exactos- de datos, el tamaño y el número no son las únicas variables 'gigantes' que están implicadas.

Tradicionalmente, los principales conceptos agrupados que han definido este nombre han sido las denominadas '3 V': volumen, variabilidad y velocidad. Macrodatos es todo aquello que tiene que ver con grandes Volúmenes de información que se mueven o analizan a alta Velocidad y que pueden presentar una compleja Variabilidad en cuanto a la estructura de su composición. Siempre me ha parecido que debería añadirse una cuarta uve, la Visualización, ya que no solo forma también parte de ello, sino que muchas de las imágenes que nos traen a la memoria el trabajo con *Big Data* tienen que ver con estas nuevas formas de 'ver' estos datos.

Pero también es importante comprender que además de los datos estructurados, aquellos otros que provienen de fuentes de información conocidas y que, por tanto, son fáciles de medir y analizar a través de los sistemas tradicionales, empezamos a poder y querer manejar datos no estructurados: los que llegan de la Web, de las cámaras de los móviles y vídeos, redes sociales, sensores de las ciudades y edificios... La variedad de su origen, además de la rapidez con la que se incrementa su volumen, son algunos de los factores que habían dificultado su análisis hasta ahora. El nuevo *software* y los nuevos modelos permiten la incorporación a los estudios tanto de un tipo como de otro. Los avances en análisis semántico también permiten estructurar mínimamente parte de los textos escritos por personas de forma automática.

Este nuevo mundo está creando nuevos perfiles profesionales siendo el conocido como científico de datos el más citado. Los científicos de datos son profesionales con habilidades en matemáticas, estadística e ingeniería informática, que son capaces de extraer el máximo valor de los datos de la organización, cerrando la brecha entre las necesidades del negocio o la Administración y las Tecnologías de la Información.

## Antecedentes

En al ámbito empresarial y el mundo de los negocios, durante la última década del pasado siglo y los primeros años de este se hablaba de *Business Intelligenc*e (BI) para hacer referencia al conjunto de estrategias y herramientas que una empresa tenía a su disposición para poder analizar los datos de su organización. Con el BI se hacían previsiones y análisis.

*Big Data* también está emparentado con lo que se ha conocido como minería de datos, un campo de las Ciencias de la Computación que intenta descubrir patrones en grandes volúmenes de datos. La minería de datos (parte de BI), al igual que el *Big Data*, utiliza los métodos de la *Inteligencia Artificial* (IA) y la Estadística para analizar los patrones en las bases de datos con las que trabaja.

Permítanme que cite también la importancia del *Big Data* para el mundo de la prensa. En el periodismo se denomina específicamente periodismo de datos a aquel realizado con las herramientas de *Big Data*, cuyos antecedentes se hunden en el denominado periodismo de precisión o periodismo asistido por computadora (CAR, en sus siglas inglesas). Cuando en 1988 Bill Deadman recibió un premio Pulitzer por sus trabajos periodísticos con bases de datos, demostrando criterios racistas en la concesión de créditos para la adquisición de viviendas en Atlanta, posiblemente no imaginaba que inauguraba también en esa profesión

una línea de trabajo que hoy ha pasado a denominarse periodismo de datos. Trabajos como los de *The Guardian* o, en castellano, *La Nación de Argentina* se encuentran hoy a la vanguardia de la aplicación del análisis de bases de datos con aplicaciones en la prensa.

Se preguntará el lector que si ya existían en el periodismo, en el mundo de las empresas o el ámbito económico -entre otros- las herramientas y conceptos que hoy se agrupan bajo la denominación *Big Data*, el cambio de etiqueta pueda obedecer a una mera fórmula comercial en la que el *marketing* de empresas y consultoras reempaqueta y cobra de nuevo por un concepto que resucita al amparo de las modas. No faltando razón a la pregunta, la respuesta ha de ser negativa, ya que hay un cambio muy importante y es el que tiene que ver con las '3 V' citadas anteriormente: los volúmenes, velocidad y variabilidad de esos datos han crecido exponencialmente. Lo que antes eran unos números al alcance de un simple PC y una hoja de cálculo han pasado a ser ingentes cantidades que están almacenadas en 'la nube' a lo largo de granjas enteras de computadoras y que necesitan ser procesadas con programas especiales que permitan manejarlos con rapidez. Esta nube (*Cloud Computing*) es un nuevo modelo de prestación de servicios de computación, información y aplicaciones a través de Internet donde la mayoría del *software* se ejecuta en la propia Red. Cada vez con más frecuencia, las aplicaciones y *software* en la nube y el *Big Data* se dan la mano para poder desarrollarse juntos.

De la hoja Excel hemos pasado a Hadoop, un software que permite trabajar con miles de nodos distribuidos y con petabytes de información. ¿Qué es un petabyte, se preguntarán? Sirva contarles que si filmáramos toda (¡toda!) la vida de una persona longeva en alta definición, nos llegaría con medio petabyte; o que todo Facebook, con sus imágenes, vídeos, etc., ocupa 1,5 petabytes. Hemos saltado del mega al peta en apenas cinco años.

El crecimiento del volumen de datos que podemos manejar es exponencial, la velocidad también: esa es la principal diferencia con relación a las disciplinas predecesoras. Igual que las versiones del *software* van avanzando en unidades (1.0, 2.0...) las versiones del *Big Data*, si las hubiera, parecen avanzar a golpe de potencias, de exponentes (*Big Data*$^2$, *Big Data*$^3$), con todo lo que ello significa.

## Problemas con la privacidad

El trabajo con *Big Data* trae a la luz más problemáticas con un clásico digital: la privacidad. En un informe de la consultora internacional McKinsey se ponían en el punto de mira «las políticas relacionadas con la privacidad, seguridad, propiedad intelectual, e incluso con la responsabilidad». Son aspectos -advertía la compañía- que deberán ser abordados en breve para continuar con el desarrollo de los sistemas de *Big Data*.

El acceso a los datos críticos de las empresas es cada vez más una necesidad para poder integrar la información de múltiples fuentes de datos, a menudo de terceros, y poder analizarla, pero ese acceso raya en muchas ocasiones la frontera de lo privado.

«Hay que tener en cuenta los límites de las normativa. Ver si el usuario ha habilitado el permiso para obtener esa información o no. Asimismo, hay que trabajar mucho las

condiciones y términos de uso, ya que si no después nos encontraremos con un problema con el usuario», explicaba recientemente en un foro sobre regulación de datos.

Los expertos afirman que «se necesita un sistema que permita determinar los niveles de acceso dependiendo incluso de las edades». Además, también se necesita un mecanismo que «deje una huella para que se pueda disponer de esos datos y que al mismo tiempo esté todo relacionado con el cumplimiento de normativas, tanto internas como legales». No son dificultades menores que se añaden a las ya de por sí complejas del propio manejo de los datos.

Pero ¿cómo usar estos datos? Esta pregunta es una de las que mayor problemática está generando en torno a la puesta en práctica de las herramientas de *Big Data*, en concreto para la mejora de las ciudades gracias a un uso más inteligente de los datos en el entorno conocido como *Smart Cities* (ciudades inteligentes gracias al manejo de datos aplicados a una mejor gestión de sus infraestructuras) «se utilizarán para mejorar la 'inteligencia de la ciudad' y no tanto para determinar oportunidades de venta a personas concretas, a no ser que esas personas accedan a ello». ¿Son solo intenciones? ¿Puede el ciudadano estar tranquilo con el poder que cede cuando permite que sus datos pasen a poder ser estudiados con las poderosas herramientas que se usan en *Big Data*? Hay al menos que ar de este peligro, máxime cuando los legisladores, muy especialmente en nuestro país, ya han demostrado una sorprendente lentitud en otros casos relacionados con las nuevas tecnologías.

### *Open Data*

Aunque no es estrictamente *Big Data*, hemos de citar aquí la necesidad de que las empresas, pero sobre todo la Administración Pública y las universidades abran sus bases de datos y permitan su manejo de forma abierta a los ciudadanos o compañías que quieran hacer uso de esos números. Si los datos son el petróleo del siglo XXI, ese combustible no sirve de nada si no podemos extraerlo. Hoy sabemos de miles de bolsas 'subterráneas' de datos a los que no hay forma de acceder; y cuando esto puede hacerse, la forma de lo que nos encontramos no es la adecuada. Es obligación de nuevo de los legisladores establecer el marco en el que esto ha de darse. España, a la hora de escribir este artículo, es el único país de la Unión Europea que no cuenta con una Ley de transparencia, que no ha desarrollado una legislación que obligue a poner a disposición ciudadana aquellos datos que, no afectando a la privacidad y seguridad, los ciudadanos tienen derecho a poder conocer si los necesitan. Además han de poder acceder a ellos en las mejores condiciones de formatos y estructuras.

Esperemos que cuando se vuelva a preparar un nuevo Dossier sobre datos en esta revista, este artículo solo hable de un pasado ya superado. Por otra parte, quizás con el tiempo el adjetivo *big* deje de tener importancia, como publicaba recientemente la revista *Forbes*. Porque al ritmo de crecimiento de materia prima y herramientas, ¿a qué llamaremos *big* en el futuro?