

Big Data

El reto de tratar de forma efectiva una ingente cantidad de información

Según el estudio “The Big Data Opportunity” Policy Exchange publicado en 2012, en este año la capacidad instalada de almacenamiento de información en el mundo alcanzó los 2´5 zetabytes. Se estima que, hasta 2020, su crecimiento anual será de un cincuenta por ciento, lo que llevaría a alcanzar los 100 zetabytes. El reto de hacer frente a un tratamiento efectivo de esta ingente cantidad de información afecta a todas las organizaciones, incluidas las Administraciones Públicas. En el presente artículo se analizan los retos y oportunidades a los que nos enfrentamos.



EMILIO GARCÍA
Jefe de Área de Cooperación en Tecnologías de la Información de la S.G. de Programas, Estudios y para el Impulso de la Administración Electrónica del Ministerio de Hacienda y AA.PP.

Lejos de toda duda, vivimos en la era de la información... Su almacenamiento, en formato electrónico, está comenzando a canibalizar la almacenada en formato papel, que lleva ya más de una década decreciendo.

Este crecimiento exponencial del volumen de información electrónica almacenado en el mundo (FIGURA 1) amenaza nuestra capacidad para procesarla. Las bases de datos tradicionales, y el procesamiento centralizado de la información, se muestran como técnicas insuficientes para hacer cara a este fenómeno, que ha originado el concepto de “Big Data”, conjuntos de datos demasiado grandes para su tratamiento con herramientas tradicionales de gestión de base de datos (1).

El paradigma Big Data es una consecuencia de nuestro tiempo, del desarrollo de más de una década de la Sociedad de la Información. Es quizás el fruto que más va a contribuir a que recojamos sus beneficios. Nuevas capacidades y tecnologías son necesarias para ello, también nuevos modelos organizativos, pero no explotar su potencial sería equivalente a no realizar el sprint final en una carrera de fondo.

Big Data nos plantea nuevos retos, especialmente en el ámbito de las Administraciones Públicas, dónde tanta información manejamos. Pero los beneficios que trataremos a continuación exigen ser asumidos como medio para conseguir mejores y más eficientes organizaciones, capaces de facilitar servicios más personalizados y efectivos.

Acotando el concepto

Siempre es difícil identificar el origen de los conceptos y paradigmas en un entorno rápidamente cambiante como el de las Tecnologías de la Información. “Big Data” no supone

una excepción a esta regla, siendo difícil identificar si surge como consecuencia o acompañante de otros conceptos como el “Open Data”. Sin embargo, la popularización del término viene, indudablemente, ligada al documento seminal del concepto publicado por McKinsey Global Institute (MGI) en Junio de 2011 (2), donde se define como “conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos”.

La definición de “Big Data” facilitada por el MGI tiene una orientación claramente tecnológica, que se percibe insuficiente en organizaciones que buscan la aplicación de la tecnología para un objetivo concreto. De igual modo, adolece de una caracterización clara del problema al que nos enfrentamos en términos diferenciales respecto a la situación previa.

Desde la presentación del término por el MGI han existido diversos intentos de acotación del concepto. Una de las aproximaciones más completas, y que usaremos como punto base de este artículo es la facilitada por Gartner (3). “Big data” son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones”.

Como puede apreciarse, en la definición facilitada, se realiza una caracterización de la información, los requerimientos tecnológicos que presenta y los resultados que obtendremos de superarlos. A lo largo de estas páginas trataremos de desglosar los mismos.

¿Un tipo distinto de información?

La caracterización de la información que realiza Gartner, identifica como la principal diferencia la forma en que nuestros sistemas habrán de ser

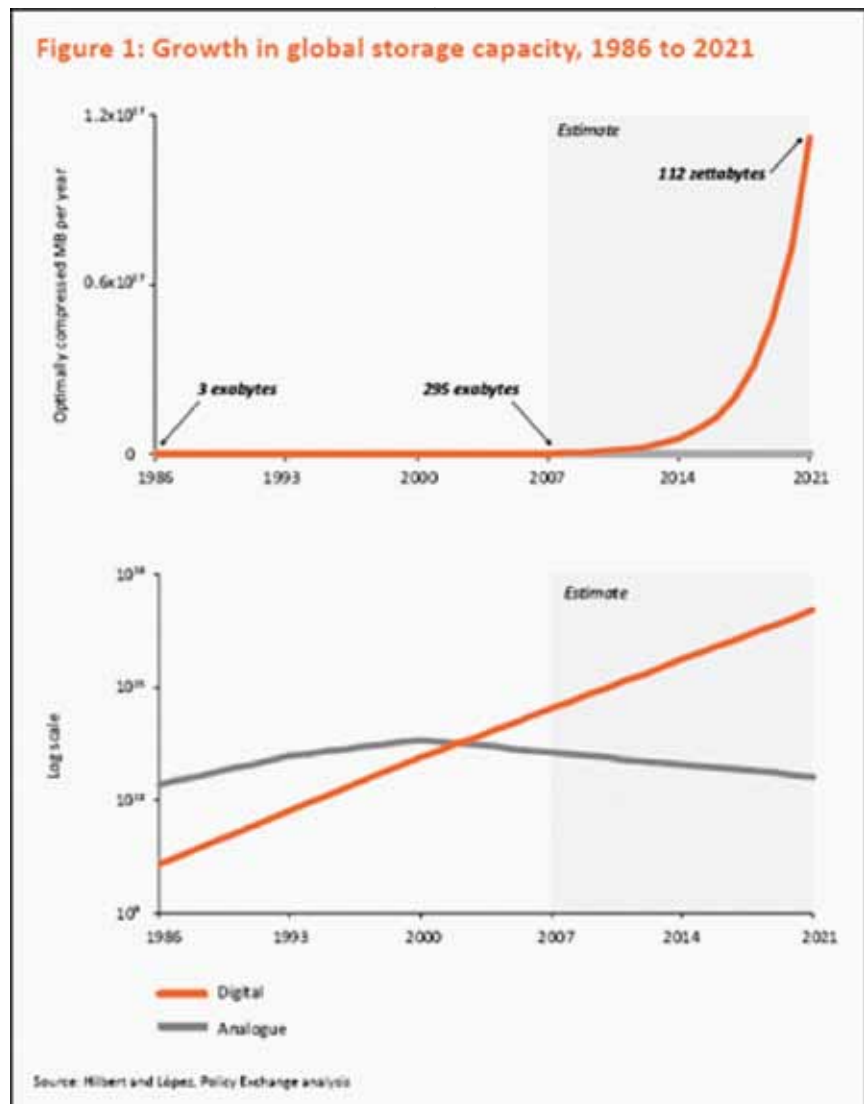


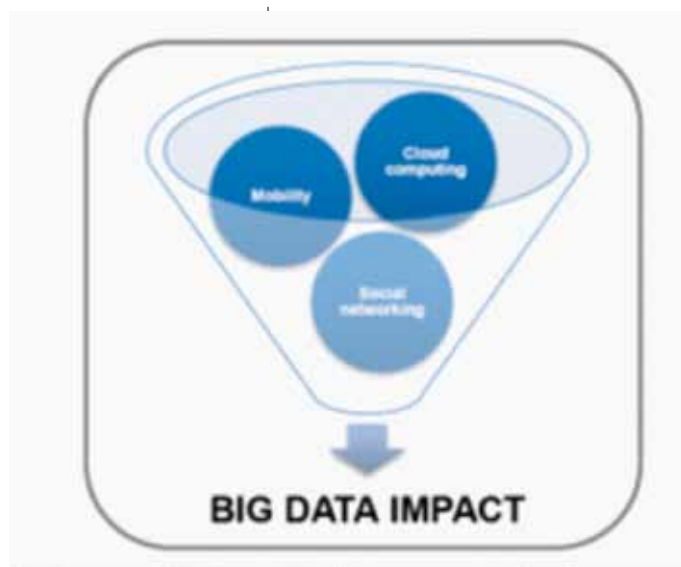
FIGURA 1. Evolución de la capacidad de almacenamiento

capaces de procesar la información respecto de la que, tradicionalmente, se venía utilizando. Son, sin embargo, todos ellos conceptos relativos que permiten dar estabilidad a la definición. Esta relativización es importante, ya que lo que hoy consideramos un alto volumen, gran velocidad o elevada variedad de formatos, resultará trivial el día de mañana cuando nuestros sistemas evolucionen de acuerdo a la Ley de Moore.

El reto que presenta el “Big Data” no es, aun siendo importante, tanto las tres V’s que popularmente le caracteriza (velocidad, volumen, varie-

dad), sino la complejidad de tratarlas. Esta complejidad viene derivada de la multiplicidad de fuentes de información que maneja cualquier organización, caracterizadas por estas 3 V. A las tradicionales fuentes internas, ya complejas de combinar en las grandes organizaciones, se suman las provenientes del exterior en forma de Redes Sociales, actuando como factor multiplicador las, cada vez más ubicuas, tecnologías de movilidad. Es decir, el mayor problema no es tanto el tamaño de los conjuntos de datos, sino la complejidad derivada de la variedad de fuentes. »

FIGURA 2.
Impacto del
Big Data



En definitiva, podemos identificar, cuando saltamos de un entorno tradicional de manejo de la información a un entorno Big Data, los siguientes parámetros

- Volumen: Las capacidades de nuestros sistemas tradicionales de almacenamiento y procesamiento se ven sobrepasadas por la cantidad de datos a tratar
- Velocidad: Parte o toda la información la recibimos en tiempo real, sin poder identificar un patrón en el ritmo de recepción de la misma
- Variedad: No podemos establecer una tipología o estructura de la información, cuyo formato y estructura desconocemos hasta el momento de su tratamiento
- Complejidad: Carencia de relación aparente entre las fuentes de información sobre la que hemos de aplicar nuevas capacidades de análisis, de naturaleza distribuida y proveniente de varios actores

La diferenciación anteriormente expuesta queda resumida en la **TABLA 1**.

Es también importante resaltar que, la información “Big Data”, no queda eximida de someterse a controles más tradicionales antes de ofrecerse a los consumidores de la

información. La calidad de la información, su caducidad, clasificación y cumplimiento con las legislaciones diversas aplicables ha de ser mantenida dentro del entorno de velocidad, variedad, volumen y complejidad. Sólo, manteniendo estos controles en el nuevo entorno, será posible extraer todo el valor de la información para las áreas de negocio.

Mantener la calidad de la información resulta crítico en el entorno Big Data. Una mayor disponibilidad de información, tanto en cantidad como en fuentes, no supone necesariamente una mayor utilidad, sólo cierta si mantenemos un balance positivo entre cantidad y utilidad de la información. En un símil con el mundo de la transmisión de la información, podemos hablar de la necesidad de aumentar la señal sin que ello signifique el aumento del ruido.

La velocidad que caracteriza al Big Data nos presenta también el reto de mantener el valor temporal de la información.

Por último, las limitaciones de la legislación son más difíciles de mantener en este entorno. De un lado, la necesidad de contemplar las restricciones derivadas del mantenimiento de la privacidad. El temor a su que-

branto ha llevado a algunos críticos a hablar de Big Brother, en lugar de Big Data, apareciendo ciertamente nuevos dilemas éticos a raíz del aumento de las capacidades de tratamiento de la información. Su utilización para la prevención del crimen o la aplicación en el ámbito de la investigación en la salud, origina nuevas consideraciones a contemplar en cuanto a que la privacidad de los datos personales están o no supeditados a su valor para la comunidad. De otro lado, la variedad de fuentes, y en muchos casos la ausencia de capacidad de control sobre el contenido de la información que nos facilitan, puede llevarnos en su tratamiento a “elipses” no deseadas de las leyes de propiedad intelectual.

El valor para el negocio

La explotación del valor del Big Data no resulta trivial. Esta dificultad, ha llevado a los analistas a concluir que, durante 2015, hasta el 85% de las organizaciones del ranking Fortune 500 no serán capaces de ello (4). Como suele suceder ante nuevos paradigmas, ello se deberá en gran medida a un desconocimiento relativo de cuál es el valor que podemos extraer del mismo.

La información sólo es valiosa en la medida en que sirve para mejorar el negocio. Ello, como ya se indicó anteriormente, implica la necesidad de mantener la calidad en el entorno de disponibilidad de información extrema que supone el Big Data. La mejora del negocio será, tanto mayor, en cuanto más completa sea la información de la que disponemos, pudiendo en caso contrario ser completada con más. De no ser así, no obtendremos la mejora de conocimiento que nos conduzca también a la del proceso de toma de decisiones en la organización, resultando, en tal caso, los recursos dedicados al Big Data carentes de valor.

El tratamiento del Big Data origina

Datos tradicionales		Big Data
Gigabytes/Terabytes	Volumen	Petabytes/Exabytes
Batch	Velocidad	Streaming
Estructurada	Variabilidad	Semiestructurada/desestructurada
Modelo de datos estable		Carencia de Esquemas
Centralizados	Complejidad	Distribuidos
Relación conocida		Ausencia de relación

TABLA 1. Diferencia entre entornos de manejo de la información

información de mayor valor añadido. Ésta sólo será valiosa en la medida en que la organización confíe en su corrección. Es por ello necesario presentarla sin rupturas en la cadena histórica de análisis. El modelo de información resultante ha de proporcionarse sólido en la perspectiva histórica (hindsight), mostrándose válido para entender el pasado; en la perspectiva presente (insight), siendo concluyente para comprender la situación actual; proyectándose hacia el futuro (foresight), facilitando modelos predictivos creíbles. En la medida en que seamos capaces de preservar la continuidad entre las tres perspectivas, mantendremos la confianza y el valor para el negocio de los resultados de la aplicación del paradigma Big Data.

La aplicabilidad del análisis realizado permite a las organizaciones obtener el valor del Big Data profundizando en los objetivos del tratamiento de la información:

- Mayor personalización: El aumento de la granularidad de la información permite desplazar el área de conclusiones del nivel macro al micro. Ello es importante, por ejemplo, en la prestación de cualquier servicio final al cliente, ya que facilita una mayor personalización del mismo.

- Ampliación de la capacidad de resolución de problemas: El mayor volumen de datos junto con nuevos algoritmos permite descubrir patrones hasta ahora ocultos en los datos. Nuevas correlaciones que facilitan la resolución de problemas antes irresolubles por falta de capacidad.

- Aprendizaje: La combinación de fuentes de información, especialmente las externas a las organizaciones, permite establecer nuevos indicadores de rendimiento de las mismas. El seguimiento de la evolución de éstos aporta más información sobre cómo mejorar las capacidades.

El valor para el negocio del “Big Data” depende, en gran medida, de las características del mismo. No obstante, existen áreas transversales tradicionales de amplio espectro donde son de aplicación:

- Motores de recomendaciones basados en cantidades masivas de información de clientes, tales como los utilizados en redes sociales y profesionales o en los vendedores online

- Análisis de sentimientos, a partir de los comentarios de un tema o marca realizados en redes sociales

- Modelización del riesgo, que permita el diseño de acciones basado en datos históricos de larga serie y minimizando así las posibilidades de fracaso»

“Las Administraciones Públicas son uno de los sectores de negocio que más invierte en infraestructuras de almacenamiento a nivel mundial. Se espera, que en 2016, constituyan el 20% de los 43,7 billones de dólares de volumen de negocio global”

	Mejora de eficiencia operativa		Reducción de fraudes y errores		Incremento impuestos			
Base para análisis (billones)	170,8		145,2		225,9			
Alcance posible	20%	25%	1%	3%	5%	10%	Potencial impacto en la Administración española	
Ahorros potenciales	15%	20%	30%	40%	10%	20%		
Total (billones)	5,1	8,5	0,4	1,7	1,1	4,5	6,7	14,8
	Min	Max	Min	Max	Min	Max	Min	Max

TABLA 2. Estimación del potencial impacto del Big Data en la Administración pública Española



- Detección de fraude, a partir de análisis más profundo de transacciones y datos contextuales de las mismas y quien las realiza
- Marketing e influencia social, incor-

- porando información de tendencias y preferencias de mayor granularidad de un conjunto mayor de potenciales clientes
- Optimización y seguridad de redes

de telecomunicaciones, considerando los datos de una mayor cantidad de sondas dentro de las herramientas de análisis de redes y detección de intrusiones

Todo el valor del Big Data dependerá, no obstante, de que sepamos hacer uso de las tecnologías adecuadas y, sobre todo, adquiramos para la organización las capacidades necesarias.

Big Data y Administración Pública

Las Administraciones Públicas son unas de las áreas de negocio donde se estima que el concepto “Big Data” tendrá mayor aplicabilidad. Tres razones existen para ello. De un lado, las Administraciones Públicas son uno de los sectores de negocio que más invierte en infraestructuras de almacenamiento a nivel mundial. Es más, se espera que la tendencia continúe, de tal modo que en el año 2016 las Administraciones Públicas constituyan un 20% de los 43,7 billones (5) de dólares de volumen de negocio global que será para entonces en este mercado.

De otro lado, las Administraciones Públicas son, probablemente, el sector de negocio con mayor volumen de “Dark Data”. Este tipo de información agrupa a aquellos datos generados en el desempeño de la actividad habitual de negocio, que si bien no se analizan o procesan, son almacenados por cuestiones regulatorias o preventivas.

Finalmente, existen amplias sinergias entre el concepto de “Big Data” y políticas públicas ya impulsadas en un gran número de Administraciones. Tanto el impulso de la reutilización de la información del sector público, como la promoción del gobierno abierto, son dos tendencias con puntos de contacto con el “Big data”.

Ejemplos potenciales de aplicación del concepto de los grandes datos en las Administraciones Públicas podrían ser los tres siguientes:

- Gestión del rendimiento en tiempo real, facilitando a los gestores políticos datos múltiples de acciones automatizadas y no automatizadas que suceden en la Administración Pública. Un ejemplo podría ser el desarrollo de una campaña concreta de recepción de escritos ciudadanos por diversos medios, como la anual campaña del IRPF
- Lucha contra el fraude fiscal, combinando los datos de declaraciones de impuestos con datos administrativos de otra índole (por ejemplo, pensiones o empleo) o provenientes del mundo privado (por ejemplo, información capturada en webs de ventas de activos inmobiliarios)
- Salud, realizando un análisis conjunto de historiales médicos, tratados de forma anónima, a fin de identificar tendencias y patrones en enfermedades y epidemias
- Personalización de servicios, realizando a los ciudadanos una oferta proactiva de la Administración Electrónica a partir de los datos disponibles, aunque esto requeriría para su plena explotación algún tipo de reforma en el marco legal de protección de datos personales

El estudio seminal de McKinsey mencionado varias veces a lo largo del presente artículo, identifica tres grandes áreas de impacto económico para las Administraciones Públicas derivados de la aplicación del con-

cepto “Big Data”. Ahorros derivados de una mayor eficiencia, de las reducciones de fraudes y errores, y las mejoras en la recolección de impuestos. La consultora estima entre, aproximadamente, 150 y 300 billones de Euros los beneficios económicos anuales para la Unión Europea.

El impacto económico, en el caso de España, en estos mismos ámbitos lo calculamos siguiendo dos aproximaciones. Una primera estimando que el potencial impacto económico para España sería proporcional al peso de su PIB en la Unión Europea. De acuerdo a los datos del Eurostat (6), en 2011 el PIB nominal de España es aproximadamente un 8,5% del PIB de la Unión Europea. Consecuentemente, el impacto económico sobre la Administración española sería, aproximadamente, de entre 12,75 billones y 25,5 billones de Euros.

Una segunda aproximación puede realizarse aplicando los mismos multiplicadores que McKinsey en su estudio de los datos macroeconómicos españoles, observando las pautas que han sido seguidas en informes dentro del Reino Unido (7). Ello nos lleva a una estimación más modesta del potencial impacto del Big Data en la Administración pública Española, de entre los 6,7 billones y los 14,8 billones de Euros.

La necesidad de crear nuevas infraestructuras y adquirir nuevas capacidades tecnológicas y de análisis, hace del “Big Data”, un objeto natural de un Centro de Competencia específico. Ello conllevaría la ventaja de evitar duplicidad de esfuerzo en una materia novedosa y, además, es necesario dado que el mayor potencial valor de un proyecto Big data es directamente proporcional a la diversidad de las fuentes de datos utilizadas.

Conclusiones

El paradigma Big Data es una consecuencia de nuestro tiempo, del desa-

rollo de más de una década de la Sociedad de la Información. Es quizás el fruto que más va a contribuir a que recojamos sus beneficios. Nuevas capacidades y tecnologías son necesarias para ello, también nuevos modelos organizativos, pero no explotar su potencial sería equivalente a no realizar el sprint final en una carrera de fondo.

Big Data nos plantea nuevos retos, especialmente en el ámbito de las Administraciones Públicas, donde tanta información manejamos. Pero los beneficios anunciados exigen que los asumamos como medio de conseguir mejores y más eficientes organizaciones que sean capaces de facilitar servicios más personalizados y efectivos. *

NOTAS

- (1) Definición de “Big Data” facilitada en “The Big Data opportunity”, Policy Exchange, 2012
- (2) “Big Data: The next frontier for innovation, competition and opportunity”, McKinsey Global Institute, 2011.
- (3) “The importance of ‘Big Data’: A definition”, Gartner, Junio 2012.
- (4) “From Data to Decision: Delivering value from ‘Big Data’”, Gartner, Marzo 2012.
- (5) En todo el artículo, se toma como equivalencia al billón la interpretación anglosajona de 1.000 millones.
- (6) Ver http://epp.eurostat.ec.europa.eu/portal/page/portal/national_accounts/data/main_tables.
- (7) “The Big Data opportunity”, Policy Exchange, Junio 2012.
- (8) En todo el artículo, se toma como equivalencia al billón la interpretación anglosajona de 1.000 millones