

TIC

BIG DATA:

una nueva
forma
de hacer
negocios



Una investigación de la *Sloan School of Management* del Instituto de Tecnología de Massachusetts demuestra que las empresas que toman decisiones basadas en datos aumentan su productividad el 5-6 %. El Big Data está cada vez más presente en las estrategias de las organizaciones como fuente de conocimiento; y la normalización trabaja para desarrollar normas que le doten de una nomenclatura de referencia, y garantice la correcta aplicación de procesos y herramientas en el tratamiento de los grandes volúmenes de datos disponibles.

Vidal Alonso y Alfonso López

Vocales
CTN 71/SC 9
Big Data

Universidad
Pontificia de
Salamanca

Olga Arranz

Universidad
Pontificia de
Salamanca

La rápida incorporación que el término Big Data ha tenido en la cultura empresarial es un hecho que pocos pueden cuestionar. La proliferación en la generación de datos en sus propios sistemas de almacenamiento o en *cloud*, con un incremento del 40 % al año, ha provocado un entusiasmo en el sector empresarial que busca cómo obtener un mayor rendimiento de este gran volumen de datos generados. Si bien, este entusiasmo se ve en ocasiones frenado por la constatación de la gran confusión que el término ha generado en los directivos, que suelen asociarlo a conceptos relacionados con grandes volúmenes de datos, análisis sociológicos, gestión de datos o aplicaciones comerciales. En este artículo se busca inculcar una claridad del concepto que ayude a su comprensión y permita una difusión creíble del término.

Aunque el primer documento académico relativo a Big Data fue publicado en el año 2003 por Francis X. Diebolt, el término ya fue empleado por primera vez a finales de los noventa por John Masley para advertir del significado que la rápida expansión del volumen de datos iba a suponer. En la actualidad cohabitan diferentes definiciones como referencia al término de Big Data. Una definición técnica, proporcionada por un referente del sector como IBM, lo define como *la convergencia digital entre los datos estructurados presentes en bases de datos y los datos no estructurados provenientes de las nuevas fuentes de información como redes sociales, dispositivos móviles, sensores RFID, smartphones o sistemas financieros*. Esta definición estaría apoyada por la teoría de las tres uves del Big Data, donde no sólo se considera el volumen de datos, sino también la variedad de las fuentes de

La industria española participa en la elaboración de los proyectos de normas internacionales a través del comité nacional de la Asociación Española de Normalización, UNE, CTN 71/SC 9 *Big Data*

información y la velocidad de captación de los datos.

Por otro lado, resulta también interesante considerar el concepto desde el punto de vista de las empresas consultoras. McKinsey, a través de su McKinsey Global Institute, define Big Data como *conjuntos de datos cuyo tamaño va más allá de las posibilidades de las típicas herramientas de bases de datos para capturar, almacenar y gestionar*; y la consultora IDC (International Data Corporation) considera que *Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad con el objeto de extraer valor económico de ellos*.

Independientemente de la definición que se adopte, las organizaciones evolucionan hacia una mayor atención de los datos recolectados, con el objeto de convertirlos en uno de sus principales activos para la toma de decisiones y el conocimiento de sus clientes. El objetivo que quieren alcanzar es la transformación de los datos en un conocimiento que pueda utilizarse en las decisiones empresariales, bien para segmentar o clasificar a los clientes y establecer acciones efectivas que incrementen su satisfacción, bien para mejorar la eficiencia de los recursos productivos disponibles. De acuerdo con una investigación de la *Sloan School of Management* del Instituto de Tecnología de Massachusetts (MIT), se demuestra que las empresas

que toman decisiones dirigidas por los datos tienen un aumento de productividad del 5-6 %. Esta tarea no está exenta de dificultades ya que, generalmente, los conjuntos de datos disponibles se encuentran desorganizados, haciendo necesario el establecimiento de un proceso de depuración para evitar la generación de incertidumbre.

Ciclo de vida

La ejecución del proceso de Big Data tiene lugar a lo largo de un ciclo de vida donde se selecciona un subconjunto de datos, o data set, y se procede a su procesamiento en sucesivas etapas. Se desarrolla en cinco fases:

Adquisición de datos: los datos que se emplean en Big Data provienen de actividades cotidianas interesantes para la sociedad. Pueden captarse y almacenarse a través de sensores, simulaciones o transacciones que efectúen los ciudadanos. La frecuente diversidad de estos datos hace necesario que sean filtrados y organizados de forma correcta para su posterior procesamiento. Esta organización cataloga las diferentes fuentes de datos en dos tipos:

- Estructurados: Bases de Datos, Sensores, Datos empresariales.
- No estructurados: Web, Blogs, Social media, Video, Audio.

Extracción y Limpieza de Información: con relativa frecuencia, la información recolectada es una información que excede a la que realmente se necesita. Por ello, es preciso efectuar una extracción de la información ►►

TIC

► que permita disponer únicamente de la información necesaria y que ésta sea almacenada de forma estructurada. Además, estos conjuntos de datos suelen contener ausencias de información o detalles inútiles. Esto conlleva a que, muchas veces, el conocimiento encerrado en ellos sea incompleto, lo que obliga a realizar una limpieza del conjunto de datos. Esta limpieza se hace especialmente necesaria si el conjunto de datos proviene de la Web. Esta fase de extracción y limpieza también se conoce con el nombre de reducción de datos, y es un proceso complejo que debe tener la habilidad de no eliminar información que pudiera ser relevante para el proceso de Big Data.

Integración y Conversión de Datos: con frecuencia, la información recolectada proviene de múltiples fuentes heterogéneas, videos, imágenes, audios, tuits, que tienen su propio formato de almacenamiento, provocando que el formato de la muestra sea inadecuado al necesario para afrontar el análisis.

Para conseguir integrar y convertir los datos de la muestra, se emplean una serie de técnicas de transformación que estructuran y convierten los datos a un formato útil para los algoritmos que hay que emplear en el análisis. La complejidad y duración de esta fase de conversión e integración depende, sobremedida, de cómo ha sido diseñada la base de datos; y si se ha considerado un almacenamiento de datos universal que pueda utilizarse en diferentes aplicaciones.

Análisis y Modelización: los métodos que se emplean para analizar y modelizar el proceso de Big Data difieren de los métodos tradicionales estadísticos que se suelen usar cuando se dispone de una muestra pequeña. En este caso, la muestra de Big



La Unión Europea estima que a finales de 2017 más del 30 % de las principales empresas europeas tendrán incorporado procesos de Big Data en su funcionamiento empresarial

Data se caracteriza por ser dinámica, heterogénea, interrelacionada y con cierta presencia de ruido; pero, aun así, es una muestra cuyo análisis proporciona información de gran valor al trabajar los algoritmos con un mayor volumen de información, y obtener patrones y conocimiento hasta ahora desconocidos. Tecnologías como Hadoop o NOSQL son ya referentes en esta fase, al igual que la analítica de sensores o el análisis predictivo.

Interpretación: esta fase recibe los resultados obtenidos en el análisis y requiere de una interpretación por parte de la persona que tiene que tomar las decisiones. En esta interpretación se descartan las reglas carentes de valor y se conservan las tendencias o asociaciones que puedan ser de mayor utilidad en su toma de decisiones. Actualmente, esta etapa interpretativa

se ejecuta por un grupo de personas, quienes debaten la importancia de los resultados alcanzados para evitar que se descarte algún resultado relevante.

Beneficios y riesgos empresariales

Los principales gobiernos mundiales se hacen eco de la importancia que Big Data tendrá en un futuro próximo en el ámbito empresarial, en aras de generar los mayores beneficios posibles. La Unión Europea espera que para 2020 la aplicación de Big Data represente el equivalente a un 1,9 % de su crecimiento anual, es decir, 206.000 millones de euros, y que tenga un impacto positivo en el empleo generando más de 3 millones de nuevos puestos de trabajo. Igualmente, estima que a finales de 2017 más del 30 % de las principales empresas europeas tendrán incorporado procesos de Big Data en su funcionamiento empresarial; y que el 70 % restante habrán incorporado algún tipo de herramientas que procesen los datos almacenados para evitar estar fuera



de una economía basada en el conocimiento. (Ver tabla 1)

Si bien Big Data aportará sustanciosos beneficios, no es una tecnología exenta de riesgos. Riesgos empresariales que deben ser abordados por las empresas a la hora de considerar su implantación. Uno de ellos proviene de la escasez de profesionales existentes para abordar las necesidades de analistas en gestión e interpretación de datos. Sólo en Estados Unidos está previsto que alcance las 180.000 personas en el plazo de dos años. Otra fuente de riesgos está presente en la ética de los datos. La vulnerabilidad de los datos, cuando provienen de redes sociales, hace necesario abordar aspectos de privacidad, transparencia o confidencialidad, obligando a la realización de un inventario exhaustivo de todas las fuentes de datos que se incluyen en los análisis. La posible subsanación de este riesgo lleva aparejada la necesidad de una estricta política de datos que defina el alcance de los datos, establezca un sistema de aseguramiento de la calidad de los datos e identifique criterios para evaluar la integridad y oportunidad de los datos.

Tendencias futuras

Aunque resulta difícil realizar una predicción de las tendencias, numerosos

expertos coinciden en señalar algunos de los sectores donde el Big Data presenta mayores perspectivas:

- *Smart Cities*: los comportamientos de los habitantes de las ciudades proporcionarán información suficiente para diseñar actuaciones basadas en patrones de su comportamiento.
- *Mercado de Datos*: la disposición y almacenamiento de datos empresariales dará lugar a la creación de un nuevo mercado de datos donde éstos serán una fuente nueva de valor en las transacciones comerciales. Su correcta configuración, estructuración, calidad y relevancia establecerá los diferentes criterios de valoración de los datos en el nuevo mercado.
- *Deep learning*: la combinación de los datos de consumo de la sociedad con mecanismos propios de inteligencia artificial llevará a un aprendizaje de los hábitos y costumbres de los ciudadanos, dando lugar a una personalización masiva del mercado. Esto podrá utilizarse en estrategias de marketing basadas en experiencias anteriores.

En este contexto, y dada la importancia que se estima que alcanzará Big Data, sería conveniente disponer de una normativa reguladora que dote de una nomenclatura de referencia común para el sector, y que garantice la correcta aplicación de procesos y herramientas en el tratamiento de los grandes volúmenes de datos disponibles.

La Organización Internacional de Normalización (ISO) y la Comisión Electrotécnica Internacional (IEC) han creado el grupo de trabajo conjunto ISO/IEC JTC 1/WG 9 *Big Data*. Su objetivo es elaborar los primeros estándares horizontales internacionales que unifiquen la diversidad de especificaciones que existen en la actualidad. Los proyectos de normas

REFERENCIAS:

- European Big Data Value Strategic Research & Innovation Agenda (2014). Recuperado de: http://www.nessi-europe.com/Files/Private/EuropeanBigDataValuePartnership_SRIA_v099%20v4.pdf
- Big Data: Impacts & Benefits (2013). ISACA white paper. Recuperado de: <http://www.isaca.org/Big-Data-WP>

LOS DATOS

Tabla 1
■ **Estimación del potencial económico del Big Data por sectores**



internacionales en los que se está trabajando en estos momentos son:

- ISO/IEC 20546 *Information Technology - Big Data - Definition and Vocabulary*
- Serie ISO/IEC 20547 *Information technology - Big Data reference architecture*

La industria española participa en la elaboración de estos proyectos a través del comité nacional de la Asociación Española de Normalización, UNE, CTN 71/SC 9 *Big Data*. A través de este órgano de normalización se hacen llegar los intereses de las empresas, universidades y centros de investigación españoles a estos foros internacionales, evitando quedar fuera de la regulación que se está realizando en otros países sobre Big Data y poder integrarse con mayor facilidad en el mercado global del futuro. ▶