International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC)

IJISEBC, 01, II, 2015

Big Data: Estado de la cuestión

Big Data: State of the art

Antonio Paredes-Moreno[1]

[1] Departamento de Economía Financiera y Dirección de Operaciones. Universidad de Sevilla, España

aparedes@us.es

RESUMEN. Con el término Big Data se hace referencia normalmente a las colecciones de conjuntos de datos tan grandes y complejos que son muy difíciles de procesar con las herramientas que conocemos usadas por las aplicaciones en las bases de datos tradicionales. Este nuevo estado de cosas constituye un enorme desafío que incluye la captura,conservación, almacenamiento, búsqueda, intercambio, transferencia, análisis y visualización de los datos.
Los conjuntos de datos a que hacemos referencias están en constante crecimiento dado que las fuentes de las que emanan crecen en número cada día y además la velocidad de comunicación y capacidad de almacenamiento de las nuevas tecnologías cada vez es más alta. Ya no se habla de conjuntos de datos cuyo volumen viene expresado en Gigas o Terabytes sino más bien en Peta, Exa, o Zettabytes de datos.
Dado que Big Data indica grandes volúmenes a grandes velocidades, de datos muy variados es obvio que se necesitan nuevas formas de procesamiento a fin de que la información o conocimiento extraído pueda ser utilizado para una buena toma de decisiones.
En las páginas que siguen a continuación se intenta describir someramente cual es la situación presente de este reciente e interesante fenómeno que se ha dado en llamar "Big Data" o grandes datos.

ABSTRACT. The term Big Data generally refers to collections of sets of data so large and complex that are difficult to process with tools we know that the traditional database applications use. This new situation constitutes a huge challenge involving the capture, preservation, storage, search, exchange, transfer, analysis and visualization of data.
The data sets referenced are constantly growing becauseof the sources from which they emanate grown in number every day and also because the speed of the new communication and storagetechnologies is becoming increasingly high. We no longer speaks of datasets whose volume is expressed in gigabytes or terabytes but rather Peta, Exa, or Zettabytes data.
As Big Data indicates large volumes and high speeds communicationof diverse data, is obvious we need new ways and toolsto processing data in order that information or knowledge extracted can be used for good decision-making.
In the following pages we tray to describe briefly what is the present situation of this interesting phenomenon called "Big Data".

PALABRAS CLAVE: Big Data, Cloud Computing, Web and Social Media, Machine-to-Machine (M2M), Big Transaction Data, Hadoop.

KEYWORDS: Big Data, Cloud Computing, Web and Social Media, Machine-to-Machine (M2M), Big Transaction Data, Hadoop.

Paredes-Moreno, A. (2015). Big Data: Estado de la cuestión. International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC), Vol. 2, Num. 1, pp. 38-59. Consultado el [dd/mm/aaaa] en www.ijisebc.com

1. Big Data: Una primera aproximación

Cada día, la humanidad crea 1,5 trillones de bytes de datos. La imparable expansión de Internet, no sólo como un canal de información, sino como un instrumento al servicio de la gestión empresarial, explica en gran medida este incremento de datos. Pero a ello se unen otros fenómenos, como la explosión de las redes sociales, el desarrollo de la telefonía móvil (en particular de redes xG y smartphones con capacidades de conexión de datos impensables hace no tanto tiempo), el crecimiento de la producción y divulgación de materiales multimedia (foto y vídeo) por parte de usuarios particulares, la eclosión de medidores inteligentes (smartmetering) y el despliegue de dispositivos que transmiten información por radiofrecuencia.

Según el informe del año 2012 de Oracle [1], el 90% de los datos del planeta se ha generado en los últimos dos años y durante 2011 se rozaron los dos zettabytes (1 zettabyte = 1024 exabytes) de información en todo el mundo según el informe de 2011 sobre Big Data de McKinsey Global Institute [2], muchos de estos datos provienen de redes sociales (Facebook: más de 1.155 millones de usuarios al mes, 699 millones de personas se conectan cada día; 500 millones de seguidores de Twitter y cerca de 200 millones de blogs públicos); teléfonos móviles (7.000 millones en uso en todo el mundo); sistemas de telemedición; fotografías; vídeos; emails, etc. El conjunto de toda esta explosión de información recibe el nombre de Big Data y, por extensión, así también se denomina al conjunto de herramientas, técnicas y sistemas destinados a extraer todo su valor.

Big Data junto con Cloud Computing, son términos que están centrando el debate actual en el sector TI. Con el progresivo auge de smartphones, tablets y redes sociales, y una proporción cada vez mayor de procesos de negocios digitalizados, todos tenemos constancia de las cantidades enormes de datos que estas transacciones y comunicaciones producen.

Ahora bien, ¿Cómo pueden las empresas hacerse con este tesoro? ¿Cómo pueden convertir los datos propios y los que circulan por las redes y sistemas de información en un valor añadido para su negocio y en una ventaja competitiva?

No hay una respuesta única para estas preguntas. Cada organización, en función de su sector de actividad y de sus propias peculiaridades, deberá analizar qué uso puede hacer de este inmenso caudal de datos y cómo los puede aprovechar. Pero lo que sí hay es una clara respuesta tecnológica. Aprovechar el potencial de Big Data es una realidad perfectamente posible hoy en día, y a un coste razonable, gracias a los sistemas que las empresas ponen a disposición de sus clientes.

La combinación de datos masivos y su procesamiento con nuevas tecnologías ayudará a que las aseguradoras no tengan que pedir reconocimientos médicos a sus clientes para conocer su estado de salud, confirmará a un potencial comprador si un vehículo de segunda mano es de fiar a partir del color de su pintura y anticipará qué estudiantes flojean antes de un examen. El auge de internet y los teléfonos móviles inteligentes no sólo permiten rastrear datos inmensos y desvelar a quién se llama y a dónde se va, sino que con la tecnología se abre la puerta al desarrollo de sofisticados sistemas de predicción de mercado, que ya están ofreciendo resultados muy ajustados, frente a los obtenidos hasta ahora con tradicionales sondeos de opinión, no siempre tan acertados.

Big Data significa muchas cosas para muchas personas. El concepto aún no está claro, pero la investigación está comenzando a demostrar que el tema ya está entrando en el radar de muchas empresas, y algunas historias de éxito y a estallar para arriba aquí y allá. El estudio "Analytics: el uso del mundo real de Big Data" [3], realizado por IBM en colaboración con la Escuela de Negocios Saïd de la Universidad de Oxford, da una buena visión general del estado actual del uso de grandes volúmenes de datos. Otra referencia a la madurez que van adquiriendo tanto la tecnología "Cloud Computing" como "Big Data" la tenemos en Datacenter Dinamics [4].

2. Introdución: ¿Qué es Big Data?

La primera cuestión que posiblemente se nos presente en este momento es ¿Qué es Big Data y porqué se ha vuelto tan importante? pues bien, en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semiestructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis. Una buena explicación sobre Big Data la tenemos en [5].

Así pues, el concepto de Big Data se aplica a toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad específica, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos. Entonces ¿qué significa demasiada información de manera que sea posible ser procesada y analizada utilizando Big Data? Analicemos primeramente en términos de bytes:

Nombre	Cantidad de bytes	Equivalente
Bit	Unidad Básica	
Byte	8 Bit	
Kilobyte (KB)	1024	1024 bytes
Megabyte (MB)	1048576	1024 KB
Gigabyte (GB)	1073741824	1024 MB
Terabyte (TB)	1099511627776	1024 GB
Petabyte (PB)	1125899906842624	1024 TB
Exabyte (EB)	1152921504606846976	1024 PB
Zettabyte (ZB)	1180591620717411303424	1024 EB
Yottabyte (YB)	1208925819614629174706176	1024 ZB

Figura 1. Unidades de información (fuente: la Web).

Además del gran volumen de información, esta se produce y existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo suficientemente rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data.

Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho más vital cuando se usa en conjunto con la plataforma de Big Data. Pensemos en nuestras manos izquierda y derecha, cada una ofrece fortalezas individuales para cada tarea en especí fico. Por ejemplo, el jugador de beisbol sabe que una de sus manos es mejor para lanzar la pelota y la otra para atraparla; puede ser que cada mano intente hacer la actividad de la otra, mas sin embargo, el resultado no será el más óptimo.

3. ¿De dónde proviene la información?

Los seres humanos estamos creando y almacenando información constantemente y cada vez más, en cantidades astronómicas. Se podría decir que si todos los bits y bytes de datos del último año fueran guardados en CD's, se generaría una gran torre desde la Tierra hasta la Luna ida y vuelta.

Esta contribución a la acumulación masiva de datos la podemos encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto le añadimos transacciones financieras realizadas en línea o por dispositivos móviles, análisis de redes sociales (en Twitter son cerca de 12 Terabytes de tweets creados diariamente y Facebook almacena alrededor de 100 Petabytes de fotos y videos), ubicación geográfica mediante coordenadas GPS, en otras palabras, todas aquellas actividades que la mayoría de nosotros realizamos varias veces al día con nuestros smartphones. estamos hablando de que se generan alrededor de 2.5 quintillones de bytes diariamente en el mundo (1 quintillón = 1,000,000,000,000,000,000,000,000,000 bytes).

De acuerdo con un estudio realizado por Cisco, entre el 2011 y el 2016 la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78%, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta [6]. Las naciones unidas proyectan que la población mundial alcanzará los 7.5 billones para el 2016 de tal modo que habrá cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 Exabytes mensuales o 130 Exabytes anuales. Este volumen de tráfico previsto para 2016 equivale a 33 billones de DVDs anuales o 813 cuatrillones de mensajes de texto.

Pero no solamente somos los seres humanos quienes contribuimos a este crecimiento enorme de información, existe también la comunicación denominada máquina a máquina (M2M machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante. Sensores digitales instalados en contenedores para determinar la ruta generada durante una entrega de algún paquete y que esta informa¬ción sea enviada a las compañías de transportación, sensores en medidores eléctricos para determinar el consumo de energía a intervalos regulares para que sea enviada esta información a las compañías del sector energético. Se estima que hay más de 30 millones de sensores interconectados en distintos sectores como automotriz, transportación, industrial, servicios, comercial, etc. y se espera que este número crezca en un 30% anualmente.

4. ¿Qué tipos de datos se deben analizar?

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué información es la que se debe analizar?, sin embargo, el cuestionamiento debería estar enfocado hacia ¿qué problema es el que se está tratando de resolver?

Para ello lo mejor es tener una buena clasificación de los tipos de datos según sus fuentes. Se muestran aquí algunas de estas categorías o tipos de datos según el estado actual de la tecnología:

- **Web and Social Media**: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc, blogs.
- **Machine-to-Machine (M2M)**: M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.
- **Big Transaction Data**: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.
- **Biometrics**: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.
- **Human Generated**: Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos elec-

trónicos, estudios médicos, etc.

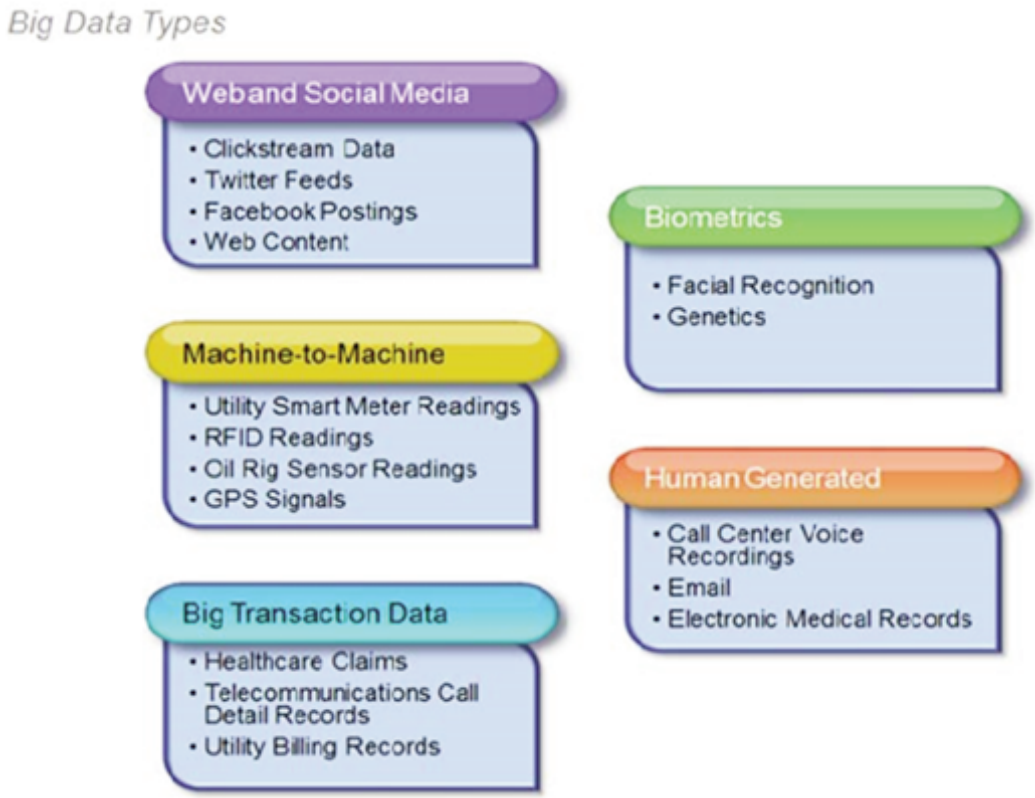

Figura 2. Big Data Types (fuente: la Web).

5. Caracteristicas de Big Data

Existe mucha confusión en torno al concepto de Big Data, confusión que comienza con la propia definición. No hay una determinada característica que predomine sobre el resto, sino que se da una división a la hora de describir el concepto de Big Data. Para unos se trata de grandes volúmenes de datos en crecimiento cada día, para otros se trata de nuevos tipos de datos y análisis o de los requisitos emergentes de un análisis de la información en tiempo aún más real

Siguiendo a Paul C. Zikopoulos, Chris Eato y otros [7], IBM en [3] e IBM en "Big Data and Analytics Hub" [8], se pueden caracterizar las dimensiones de Big Data con las llamadas cuatro V: Volumen, Variedad Velocidad y Veracidad (véase la figura 3).

- **Volumen**: La cantidad de datos. Al ser quizá la característica que se asocia con mayor frecuencia a Big Data, el volumen hace referencia a las cantidades masivas de datos que las organizaciones intentan aprovechar para mejorar la toma de decisiones en toda la empresa. Los volúmenes de datos continúan aumentado a un ritmo sin precedentes. No obstante, lo que constituye un volumen verdaderamente "alto" varía en función del sector e incluso de la ubicación geográfica y es más pequeño que los petabytes y zetabytes a los que a menudo se hace referencia. Algo más de la mitad de los encuestados consideran que conjuntos de datos de entre un terabyte y un petabyte ya son Big Data, mientras que otro 30% simplemente no sabía cuantificar este parámetro para su empresa. Aun así, todos ellos estaban de acuerdo en que sea lo que fuere que se considere un "volumen alto" hoy en día, mañana lo será más.
- **Variedad**: Diferentes tipos y fuentes de datos. La variedad tiene que ver con gestionar la complejidad

de múltiples tipos de datos, incluidos los datos estructurados, semiestructurados y no estructurados. Las organizaciones necesitan integrar y analizar datos de un complejo abanico de fuentes de información tanto tradicional como no tradicional procedentes tanto de dentro como de fuera de la empresa. Con la profusión de sensores, dispositivos inteligentes y tecnologías de colaboración social, los datos que se generan presentan innumerables formas entre las que se incluyen texto, datos web, tuits, datos de sensores, audio, vídeo, secuencias de clic, archivos de registro y mucho más.

- **Velocidad**: Los datos en movimiento. La velocidad a la que se crean, procesan y analizan los datos continúa aumentando. Contribuir a una mayor velocidad es la naturaleza en tiempo real de la creación de datos, así como la necesidad de incorporar datos en streaming a los procesos de negocio y la toma de decisiones. La velocidad afecta a la latencia: el tiempo de espera entre el momento en el que se crean los datos, el momento en el que se captan y el momento en el que están accesibles. Hoy en día, los datos se generan de forma continua a una velocidad a la que a los sistemas tradicionales les resulta imposible captarlos, almacenarlos y analizarlos. Para los procesos en los que el tiempo resulta fundamental, tales como la detección de fraude en tiempo real o el marketing "instantáneo" multicanal, ciertos tipos de datos deben analizarse en tiempo real para que resulten útiles para el negocio.
- **Veracidad**: La incertidumbre de los datos. La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. Esforzarse por conseguir unos datos de alta calidad es un requisito importante y un reto fundamental de Big Data, pero incluso los mejores métodos de limpieza de datos no pueden eliminar la imprevisibilidad inherente de algunos datos, como el tiempo, la economía o las futuras decisiones de compra de un cliente. La necesidad de reconocer y planificar la incertidumbre es una dimensión de Big Data que surge a medida que los directivos intentan comprender mejor el mundo incierto que les rodea (véase el recuadro "Veracidad, la cuarta V").

Algunos datos son intrínsecamente inciertos, por ejemplo, los sentimientos y la sinceridad de los seres humanos; los sensores GPS que rebotan entre los rascacielos de Manhattan; las condiciones climáticas; los factores económicos; y el futuro. A la hora de tratar con estos tipos de datos, ninguna limpieza de datos puede corregirlos. Aun así, y a pesar de la incertidumbre, los datos siguen conteniendo información valiosa. La necesidad de reconocer y abordar esta incertidumbre es una de las características distintivas de Big Data.

La incertidumbre se manifiesta en Big Data de muchas formas. Se encuentra en el escepticismo que rodea a los datos creados en entornos humanos como las redes sociales; en el desconocimiento de cómo se desarrollará el futuro y cómo las personas, la naturaleza o las fuerzas ocultas del mercado reaccionarán a la variabilidad del mundo que les rodea.

Un ejemplo de esta incertidumbre la encontramos en la producción energética: el tiempo es incierto, pero aun así una empresa de servicios públicos debe prever la producción. En muchos países las normativas exigen que una parte de la producción proceda de fuentes de energía renovables, pero ni el viento ni las nubes se pueden pronosticar con precisión. Entonces, ¿cómo puede planificarlo?

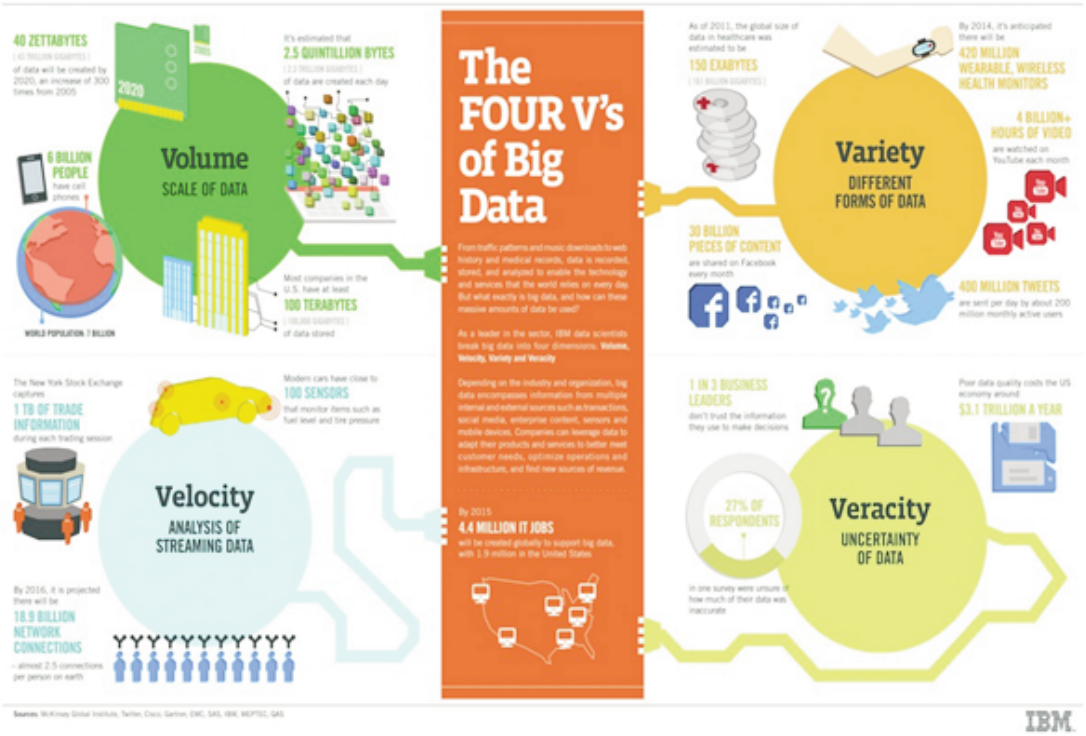

Figura 3. Las 4 Vs de Big Data (fuente: http://goo.gl/DRMrq9).

Para gestionar la incertidumbre los analistas han de crear un contexto en torno a los datos. Una forma de hacerlo es a través de la fusión de datos, donde la combinación de múltiples fuentes menos fiables da lugar a un punto de datos más preciso y útil, como comentarios sociales añadidos a la información acerca de una ubicación geoespacial. Otra forma de gestionar la incertidumbre es a través de las matemáticas avanzadas que la engloban, como sólidas técnicas de optimización y planteamientos de lógica difusa.

Por naturaleza, a los seres humanos no nos gusta la incertidumbre, pero ignorarla puede crear incluso más problemas que la propia incertidumbre. En la era de Big Data, los directivos necesitan abordar la dimensión de la incertidumbre de forma diferente. Deben reconocerla, aceptarla y determinar cómo aplicarla para su beneficio; la única certeza acerca de la incertidumbre es que no desaparecerá.

En definitiva, Big Data es una combinación de estas características que crea una oportunidad para que las empresas puedan obtener una ventaja competitiva en el actual mercado digitalizado. Permite a las empresas transformar la forma en la que interactúan con sus clientes y les prestan servicio, y posibilita la transformación de las mismas e incluso de sectores enteros. No todas las organizaciones adoptarán el mismo enfoque con respecto al desarrollo y la creación de sus capacidades de Big Data. Sin embargo, en todos los sectores existe la posibilidad de utilizar las nuevas tecnologías y analíticas de Big Data para mejorar la toma de decisiones y el rendimiento.

6. Desarrollo de Big Data

La mayor parte de las empresas se encuentra actualmente en las primeras fases del desarrollo de Big Data, la mayoría de ellas centradas en comprender los conceptos (24%) o definir una hoja de ruta relacionada con Big Data (47%). No obstante, el 28% de los encuestados trabaja en empresas de vanguardia en las que están

desarrollando pruebas de conceptos (POCs) o ya han implementado soluciones de Big Data a escala (véase la figura 4).

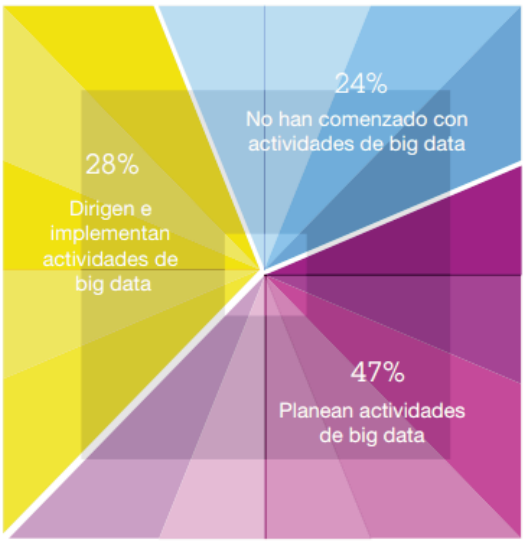

Figura 4. Fases de desarrollo de Big Data (fuente [3]).

En el estudio de IBM sobre el que se basan estas páginas [3], se extraen cinco conclusiones clave que reflejan algunas tendencias y conocimientos comunes e interesantes:

- En todos los sectores el negocio de Big Data está orientado en gran medida a abordar objetivos centrados en el cliente.
- La gestión escalable y extensible de la información es un requisito fundamental para el avance de Big Data.
- Las organizaciones están poniendo en marcha proyectos e implementaciones con fuentes de datos internas ya existentes o a las que han tenido acceso recientemente.
- Para que las empresas puedan obtener el máximo valor de Big Data son necesarias funcionalidades analíticas avanzadas, aunque a menudo carecen de ellas.
- A medida que la concienciación y la implicación de las empresas en Big Data crece se observa cómo surgen cuatro fases en el proceso de adopción de Big Data, que se describen más abajo.

6.1. Big Data centrado en el cliente

Aproximadamente la mitad de las empresas encuestadas identificaron los objetivos centrados en el cliente como la máxima prioridad de su empresa.

Las organizaciones están comprometidas con la mejora de la experiencia del cliente y con una mejor comprensión de las preferencias y el comportamiento de los mismos. Comprender al consumidor de hoy en día, mucho más "capacitado", también fue identificado como una prioridad de alto nivel tanto en la Encuesta global a directores de marketing (CMO) de 2011 como en la Encuesta global CEOs de 2012 [8].

Las empresas consideran que Big Data proporciona la capacidad para comprender y predecir mejor los comportamientos de los clientes y, al hacerlo, mejorar su experiencia. Transacciones, interacciones multicanal, redes sociales, datos sindicados a través de fuentes como las tarjetas de fidelidad y otra información relacionada con los clientes han aumentado la capacidad de las empresas para crear una imagen completa de las preferencias y demandas de los clientes: un objetivo de los departamentos de marketing, ventas y atención al cliente durante décadas.

A través de esta comprensión profunda, empresas de todo tipo encuentran nuevas formas de interactuar con sus clientes actuales y futuros. Este principio es aplicable al comercio minorista, pero también a las telecomunicaciones, la sanidad, el gobierno, la banca y las finanzas y al sector de productos al consumidor, donde usuarios finales y ciudadanos están involucrados en interacciones business-to-business (B2B) entre socios y proveedores.

De hecho, Big Data puede ser una carretera de doble sentido entre los clientes y las empresas: Por ejemplo, el Ford Focus eléctrico produce ingentes cantidades de datos mientras está siendo conducido y cuando está aparcado. Mientras se encuentra en movimiento el conductor recibe constantemente información actualizada acerca de la aceleración, la frenada, la carga de la batería y la ubicación del vehículo. Esto resulta útil para el conductor, pero esos mismos datos también llegan a los ingenieros de Ford, quienes reciben información acerca de los hábitos de conducción de los clientes, incluido cómo, cuándo y dónde cargan sus automóviles. Y mientras el vehículo se encuentra detenido continúa enviando datos acerca de la presión de los neumáticos y el sistema de batería al teléfono inteligente más cercano.

Big Data permite obtener una imagen más completa de las preferencias y demandas de los clientes; a través de esta profunda comprensión empresas de todo tipo encuentran nuevas formas de interactuar con sus clientes actuales y futuros.

De este escenario centrado en el cliente se derivan múltiples ventajas, ya que Big Data hace posibles nuevas y valiosas formas de colaboración. Los conductores reciben información útil cada segundo, mientras que los ingenieros en Detroit reúnen la información relativa al comportamiento al volante con el objetivo de extraer conocimientos acerca de los clientes y desarrollar mejoras para los productos. Y lo que es más, las empresas de servicios públicos y otros proveedores externos analizan millones de kilómetros de datos de conducción para decidir dónde ubicar nuevas estaciones de carga y cómo proteger las frágiles redes de servicio de las sobrecargas.

Empresas de todo el mundo son capaces de prestar un mejor servicio a sus clientes y de mejorar las operaciones gracias a Big Data. Empresas como

> **Mcleod Russel India Limited** han eliminado por completo el tiempo de inactividad de los sistemas en el comercio del té gracias a un seguimiento más preciso de las cosechas, la producción y el marketing de hasta 100 millones de kilos de té cada año.
>
> **Premier Healthcare Alliance** recurrió a funciones de intercambio de datos y analíticas avanzadas para mejorar los resultados de los pacientes y reducir al mismo tiempo su gasto en 2.850 millones de dólares.
>
> **Santam** mejoró la experiencia del cliente al implementar el análisis predictivo con el objetivo de reducir el fraude.

Además de los objetivos centrados en el cliente, también se abordan otros objetivos funcionales a través de las primeras aplicaciones de Big Data. La optimización operativa, por ejemplo, fue uno de los objetivos citados por el 18% de los encuestados, pero consiste principalmente en proyectos piloto. Otras aplicaciones de Big Data que se mencionaron con frecuencia incluyen la gestión financiera/de riesgos, la colaboración de los empleados y la habilitación de nuevos modelos de negocio.

6.2. Big Data depende de información escalable y extensible

La promesa de lograr un valor de negocio importante y cuantificable a partir de Big Data solo puede hacerse realidad si las empresas crean una base de información que respalde el volumen, la variedad y la velocidad de los datos de rápido crecimiento. En el estudio, las empresas afirmaron haber comenzado su viaje hacia Big

Data con una base de información integrada, escalable, extensible y segura. Cuatro fueron los componentes de la gestión de la información citados con mayor frecuencia como parte de las iniciativas de Big Data de los encuestados. La información integrada es un componente fundamental de cualquier esfuerzo analítico y es incluso más importante si hablamos de Big Data. Tal y como se apunta en el estudio realizado por el IBM Institute for Business Value en 2011 acerca de la analítica avanzada, los datos de una empresa han de estar disponibles y accesibles para las personas y sistemas que los necesitan (Ver figura 5).

Los dos siguientes componentes de la base de gestión de la información que se mencionan con mayor frecuencia en las iniciativas de Big Data son una infraestructura de almacenamiento escalable y un warehouse de gran capacidad. Ambos respaldan el rápido crecimiento de los datos, actuales y futuros, que llegan a la organización.

A primera vista, el hecho de añadir más capacidad de almacenamiento y uno o más servidores grandes puede parecer suficiente para respaldar el crecimiento de una base de gestión de la información. No obstante, es importante comprender que prever y configurar la infraestructura resulta clave para alcanzar el valor de negocio del caso de negocio pretendido. Las empresas han de plantearse cómo soportar de la mejor forma posible el vaivén de datos a fin de permitir a los usuarios acceder a los mismos cuando los necesiten y cómo analizar los datos teniendo en cuenta las limitaciones de tiempo de las empresas (ya sean días, horas, segundos o milisegundos). Este equilibrio de configuración y despliegue de servidores y almacenamiento tiene como resultado una infraestructura más optimizada

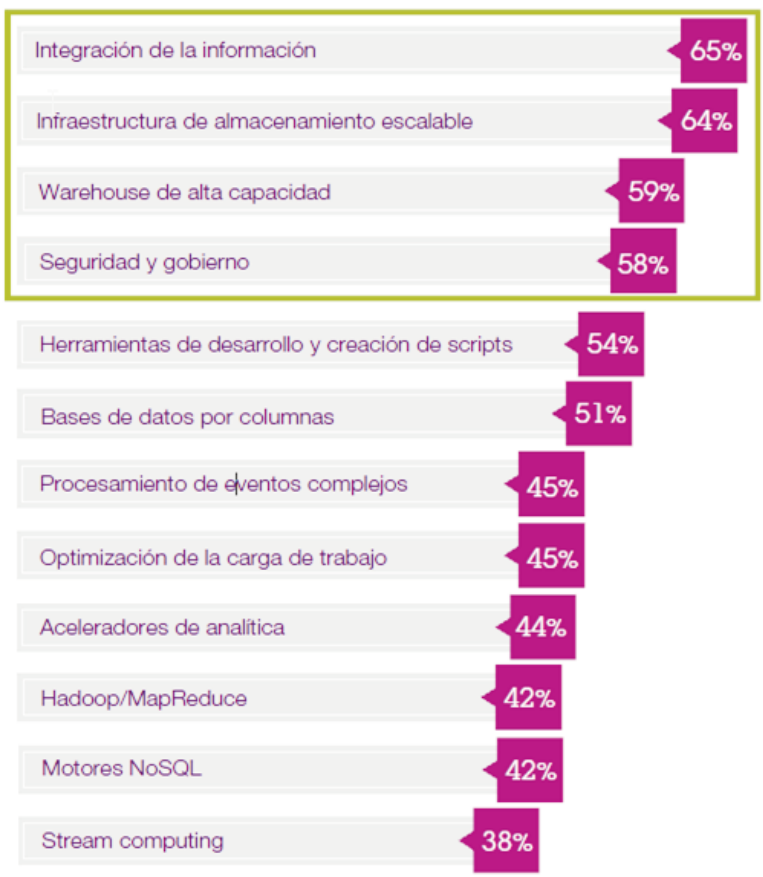

Figura 5. Infraestructura de Big Data. (fuente: [3]).

Estas tecnologías también son capaces de gestionar la creciente velocidad de los datos que llegan, y se almacenan, al hacer posible un movimiento coherente y automatizado de los datos en toda la empresa a medida que más personas necesitan tener acceso a tipos de información adicionales y diferentes. Tecnologías emergentes como la jerarquización y compresión de datos y los sistemas de archivos escalables, junto con bases de datos en memoria, hacen posible la gestión de cargas de trabajo mucho más grandes que los warehouses convencionales. Para muchas organizaciones, mejorar la capacidad para gestionar volúmenes de datos en crecimiento es la máxima prioridad de Big Data, seguida muy de cerca por la capacidad para abordar la creciente variedad de datos (un ejemplo de ello lo tenemos en Vestas Wind Systems. Ver 1.10.8 para una descripción más detallada).

El 58% de las empresas que afirman haber puesto ya en marcha iniciativas de Big Data cuenta con unos procesos de seguridad y gobierno sólidos. Si bien la seguridad y el gobierno han sido durante mucho tiempo un aspecto inherente al business intelligence, las nuevas consideraciones jurídicas, éticas y normativas de Big Data introducen nuevos riesgos y amplían el potencial de fallos públicos, tal y como hemos tenido la oportunidad de ver con algunas empresas que han perdido el control sobre los datos o los han utilizado de formas cuestionables.

Como resultado de ello, la seguridad de los datos, y especialmente la privacidad de los mismos, constituye una parte fundamental de la gestión de la información, tal y como afirman varios expertos en la materia y directivos empresariales. La seguridad y el gobierno serán todavía más importantes a medida que las empresas comiencen a utilizar nuevas fuentes de información, especialmente datos procedentes de redes sociales. Para complicar aún más la situación, las normativas sobre privacidad continúan evolucionando y pueden variar enormemente dependiendo del país.

"Existe la percepción de que la privacidad y la seguridad son aspectos fáciles, pero están muy regulados y se encuentran bajo un férreo control", señala un directivo del sector de las telecomunicaciones. Y no son solo las agencias gubernamentales las que ejercen este control, sino también los propios clientes.

Según este mismo directivo, "hay una serie de ámbitos nuevos, como pueden ser los datos la navegación web, donde existe una zona gris entre lo que es legal y lo que está bien. Una buena máxima a aplicar respecto a este tema podría ser la de considerar qué pensaría el cliente si la forma en la que tal empresa u organización utiliza sus datos apareciera reflejada en la página web de dicha empresa".

Para algunos de los directivos entrevistados, los costes de actualización de las infraestructuras constituían otra inquietud. Según afirmaron, la alta dirección exige un caso de negocio sólido y cuantificable, uno que defina las inversiones progresivas junto con las oportunidades para racionalizar y optimizar los costes de sus entornos de gestión de la información. Arquitecturas de menor coste, incluido el cloud computing, la externalización estratégica y la fijación de precios basada en el valor, fueron citadas como tácticas que están siendo desarrolladas en la actualidad. Aun así, otros han invertido en sus plataformas de información sobre la base de la convicción de que la oportunidad de negocio merecía el incremento de costes asociado.

6.3. Obtener conocimientos de fuentes internas nuevas o existentes

La mayor parte de los esfuerzos de Big Data están dirigidos a extraer y analizar datos internos. Más de la mitad de las empresas encuestadas afirmaron que la fuente principal de Big Data en sus empresas eran los datos internos. Esto sugiere que las empresas están siendo pragmáticas al adoptar Big Data y también que existe un tremendo valor por descubrir escondido en esos sistemas internos (ver figura 6).

Tal y como cabía esperar, los datos internos son los datos más desarrollados y mejor entendidos de las empresas. Estos se han recabado, integrado, estructurado y normalizado a lo largo de años de planificación de recursos empresariales, gestión de datos maestros, business intelligence y otras actividades relacionadas. Al aplicar la analítica, los datos internos obtenidos de las transacciones de los clientes, las interacciones, los even-

tos y los correos electrónicos pueden proporcionar conocimientos valiosos5, mayores ingresos a través de una mejor información. No obstante, en muchas empresas el tamaño y el alcance de sus datos internos (tales como datos detallados de transacciones y registros operativos) son ahora demasiado grandes o variados como para poder gestionarlos con los sistemas tradicionales.

Casi tres de cada cuatro empresas con iniciativas de Big Data en curso analizan datos procedentes de logs. Se trata de datos generados por máquinas/sensores que se utilizan para registrar detalles de funciones automatizadas llevadas a cabo en el marco de sistemas de información o empresariales, datos que han desbordado la capacidad de la que disponen muchos sistemas tradicionales para su almacenamiento y análisis. Como resultado de ello, muchos de estos datos se recaban pero no se analizan

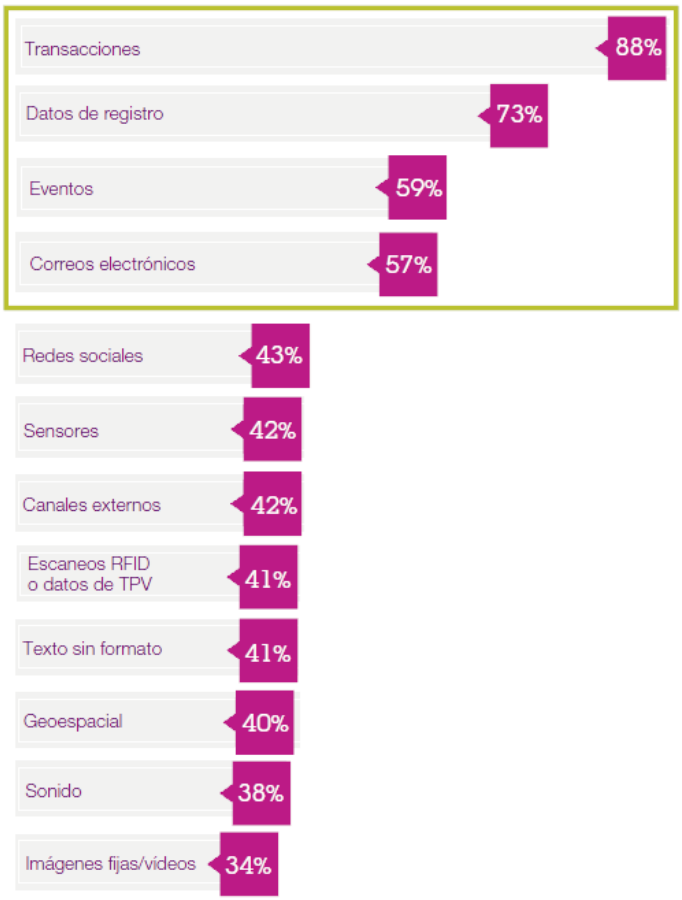

Figura 6. Fuentes de datos. (fuente [3]).

6.4. Big Data requiere funcionalidades analíticas sólidas

No obstante, Big Data no crea valor hasta que se utiliza para superar importantes retos empresariales. Esto requiere un acceso a más tipos de datos diferentes entre sí, así como sólidas funcionalidades analíticas que incluyen tanto herramientas de software como las habilidades necesarias para utilizarlas. Un análisis de las empresas inmersas en actividades de Big Data revela que comienzan con un sólido núcleo de funcionalidades analíticas diseñadas para abordar datos estructurados. A continuación, añaden capacidades para aprovechar la enorme cantidad de datos que llegan a la empresa, tanto datos semiestructurados (datos que se pueden con-

vertir a formatos de datos estándar) y no estructurados (datos en formatos no estándar).

Más del 75% de los encuestados con iniciativas de Big Data en curso señalaron que utilizan funcionalidades analíticas clave, tales como las consultas, la generación de informes y la extracción de datos para analizar Big Data, en tanto que más del 67% afirma que utiliza modelos predictivos. Comenzar con estas funcionalidades analíticas fundamentales es una forma pragmática de comenzar a interpretar y analizar Big Data, especialmente cuando están siendo almacenados en una base de datos relacional.

La necesidad de funciones de visualización de datos más avanzadas aumenta con la introducción de Big Data. A menudo los conjuntos de datos son demasiado grandes para que las empresas o los analistas de datos puedan visualizarlos y analizarlos con las herramientas tradicionales de generación de informes y extracción de datos. El 71% de las empresas con iniciativas de Big Data en curso depende de las habilidades de visualización de datos.

Las empresas inmersas en Big Data necesitan funciones cada vez más avanzadas para descubrir patrones en la inherente complejidad. Para lograrlo, se aplican modelos de optimización y analítica avanzada a fin de comprender mejor cómo transformar los procesos de negocio clave. Utilizan funciones de simulación para analizar las miles de variables disponibles con Big Data. Más del 50% de las iniciativas de Big Data en curso utilizan estas funciones de modelación avanzadas.

La mayor parte de las empresas centran la atención de sus primeras iniciativas de Big Data en analizar datos estructurados. Sin embargo, Big Data también genera la necesidad de analizar múltiples tipos de datos, incluida una gran variedad de datos que pueden ser completamente nuevos para muchas organizaciones. En más de la mitad de las iniciativas de Big Data en curso las empresas afirman utilizar funcionalidades avanzadas diseñadas para analizar texto en su estado natural, como pueden ser las transcripciones de las conversaciones de un centro de atención telefónica. Esta analítica incluye la capacidad para interpretar y comprender los matices del lenguaje, tales como los sentimientos, el argot y las intenciones.

Disponer de la capacidad para analizar datos no estructurados (por ejemplo, datos de una ubicación geoespacial, voz y vídeo) o en streaming sigue siendo un reto para la mayoría de las empresas. A medida que el hardware y el software de estos ámbitos evoluciona, las habilidades siguen siendo escasas. Menos del 25% de las empresas con iniciativas de Big Data en curso cuenta con las capacidades necesarias para analizar datos no estructurados, tales como voz y vídeo.

Adquirir o desarrollar estas capacidades analíticas y técnicas más avanzadas necesarias para el avance de Big Data se está convirtiendo en un importante reto para muchas empresas con iniciativas de Big Data en curso. Entre estas organizaciones, la falta de habilidades analíticas avanzadas constituye un gran obstáculo a la hora de obtener el máximo valor de Big Data.

6.5. Fases de adopción de Big Data

Con relación al nivel de actividades de Big Data existente en las empresas actualmente (2012), los resultados sugieren cuatro fases principales en el proceso de adopción y evolución de Big Data, junto con un continuo que denominado Educar, Explorar, Interactuar y Ejecutar (véase la figura 7).

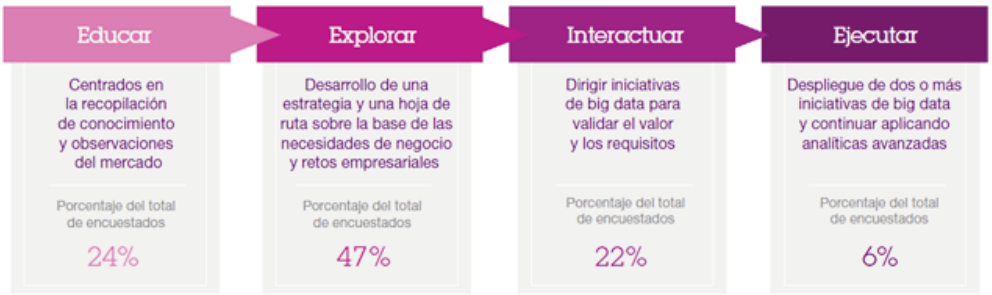

Figura 7. Fases de adopción de Big Data. (fuente [3]).

- **Educar**: crear una base de conocimiento. En la fase de educación la atención se centra en la concienciación y el desarrollo del conocimiento. Casi el 25% de las empresas encuestadas indica que aún no utiliza Big Data dentro de sus empresas. Si bien algunos siguen teniendo relativamente poca información acerca del concepto de Big Data, las personas entrevistadas sugieren que la mayoría de las empresas que se encuentran en esta fase están estudiando las posibles ventajas de las tecnologías y la analítica de Big Data e intentando comprender cómo puede ayudarles a abordar importantes oportunidades de negocio en sus propios sectores o mercados. En el seno de estas empresas son principalmente los empleados los encargados de recabar la información, a diferencia de los grupos de trabajo formales, y sus conocimientos aún no están siendo utilizados por la empresa. Como resultado de ello, los directivos empresariales aún no han comprendido totalmente ni abrazado el potencial de Big Data.

- **Explorar**: Definir el caso de negocio y la hoja de ruta. En esta fase la atención se centra en desarrollar la hoja de ruta de la empresa para el desarrollo de Big Data. Prácticamente la mitad de las empresas reconoce que tienen conversaciones formales en curso dentro de sus organizaciones acerca de cómo utilizar Big Data para abordar importantes retos empresariales. Los principales objetivos de estas empresas incluyen desarrollar un caso de negocio cuantificable y crear un proyecto de Big Data. La estrategia y la hoja de ruta tienen en cuenta los datos, la tecnología y las habilidades existentes y, a continuación, establecen dónde comenzar y cómo desarrollar un plan en consonancia con la estrategia de negocio de la empresa.

- **Interactuar**: Adoptar Big Data. En la fase de la interactuación las empresas comienzan a comprobar el valor de negocio de Big Data, así como a llevar a cabo una valoración de sus tecnologías y habilidades. Más de una de cada cinco empresas encuestadas está desarrollando en la actualidad POCs para validar los requisitos asociados a la implementación de iniciativas de Big Data, así como para articular los resultados esperados. Las empresas que se encuentran en este grupo están trabajando (dentro de un ámbito definido y limitado) para comprender y probar las tecnologías y habilidades necesarias para aprovechar nuevas fuentes de datos.

- **Ejecutar**: Implementar Big Data a escala. En la fase de ejecución, el nivel de operatividad e implementación de las funciones analíticas y de Big Data es mayor dentro de la empresa. No obstante, tan solo el 6% de las empresas encuestadas han implementado dos o más soluciones de Big Data a escala. Este escaso número de organizaciones en la fase de ejecución resulta coherente con las implementaciones que vemos en el mercado. Y, lo que es más importante, estas empresas líderes están aprovechando Big Data para transformar sus negocios, por lo que están obteniendo el máximo valor de sus activos de información. Con la tasa de adopción de Big Data aumentado rápidamente (tal y como demuestra el 22% de los encuestados en la fase de interactuación, ya sea con POCs o con proyectos piloto en curso), se espera que el porcentaje de empresas en esta fase se incremente en el futuro.

6.6. Obstáculos a Big Data

Los desafíos que obstaculizan la adopción de Big Data difieren a medida que las empresas avanzan a lo largo de cada una de las fases de adopción de Big Data. Sin embargo, los resultados muestran un reto sistemático, independientemente de la fase, que es la capacidad para articular un caso de negocio convincente. En cualquiera de las fases, las iniciativas de Big Data se someten a un escrutinio fiscal. El actual entorno económico global ha dejado a las empresas con un escaso apetito por nuevas inversiones en tecnología sin beneficios cuantificables, un requisito que, por supuesto, no es exclusivo de las iniciativas de Big Data. Después de implementar de forma satisfactoria los POCs, el principal desafío al que se enfrentan las empresas es encontrar las habilidades necesarias para que Big Data resulte operativo, incluídas las habilidades técnicas, analíticas y de gobierno.

6.7. Ejemplos de aplicación de Big Data

El análisis de datos para establecer nuevos modelos de negocio o definir estrategias comerciales será una de las mayores oportunidades para empresas e industrias en los próximos años, y cada vez más sectores se están dando cuenta. Viajar con Big Data puede convertirse en toda una experiencia a la hora de ofrecer nuevos servicios o productos, pero esta oportunidad dependerá de la capacidad de las empresas y los sectores económicos para adaptarse a uno de los activos más importantes hoy en día: la gestión de la información.

En este momento, existen varias prácticas de utilización de Big Data tanto en gigantes de la web como Google, Facebook o Linkedin como en compañías más tradicionales. Veamos a continuación de modo breve algunos ejemplos de sectores económicos o empresas donde se empieza a utilizar Big Data, así como los servicios a los se aplica:

Sector Público:

Servicios de inteligencia, defensa y protección (control de comunicaciones, vigilancia, intercepción de redes de telefonía, acumulación de todo tipo de datos); protección de la flota pesquera; vigilancia, seguridad y señalización y proyectos de Smart Cities, localizaciones por GPS, detección del fraude, control de presupuestos públicos, protección de la infraestructura pública, protección contra el maltrato, etc.

Sanidad:

Monitorización remota de pacientes, localización de emergencias y almacenamiento de historias clínicas, radiografías, escáneres y todo tipo de pruebas de forma Centralizada, Elaboración de estadísticas alrededor de incidencias de determinadas enfermedades por zonas concretas, acercamiento de la asistencia a domicilio, investigación clínica: estudios de medicamentos, ensayos clínicos, genoma humano, etc.

Retail-Gran Consumo:

Las prácticas de explotación de Big Data son el núcleo de su negocio desde hace muchos años por encima de las aplicaciones transaccionales: Control de la cadena de fabricación, análisis del ticket de compra, marketing personalizado y RFID (Identificación por Radio Frecuencia) en centros comerciales.

Telecomunicaciones:

Control de la red, venta de servicios de localización, servicios de publicidad asociados al patrón de llamadas o las aplicaciones descargadas, obtención de perfiles enriquecidos de consumidor y explotación de RFID para segmentar y personalizar ofertas, análisis de abandono , riesgo y fraude en clientes, satisfacción y lealtad de clientes, análisis de CDR (Call Data Record) o registro de llamadas), etc.

Utilities:

Interpretación de contadores inteligentes en todas las casas, control de la red comunicaciones, de tuberías, red del metro y proyectos de señalización de tramos de mantenimiento.

Sector turístico

Uno de los últimos sectores en subirse al tren del Big Data ha sido el turístico. Según un informe elaborado por la compañía Amadeus [9] y el artículo [10], la integración de la tecnología de análisis de datos en este sector supondrá todo un revulsivo que podría definir las directrices a seguir para superar los retos futuros del sector turístico. El estudio pone de manifiesto algunas de las prácticas más novedosas e interesantes, dentro de la tecnología Big Data, que se están llevando a cabo en el sector para establecer estrategias comerciales e impulsar un sector castigado por la coyuntura económica actual.

Un buen ejemplo de una compañía de turismo que obtiene ventajas competitivas a partir del uso de Big Data es British Airways. El objetivo de esta línea aérea es entender a sus clientes mejor que cualquier otra mediante su programa "Know me" (Conóceme), que analiza datos provenientes de decenas de millones de puntos de contacto para conocer al cliente. La empresa reconoce y recompensa la fidelidad de sus clientes, monitorea todo tipo de inconvenientes y los resuelve, y brinda a sus clientes ofertas personalizadas.

Otro ejemplo lo tenemos en la central de reservas online Kayak utiliza la tecnología Big Data para predecir el precio que tendrán los vuelos en un periodo de tiempo de entre 7 y 10 días, con el fin de ofrecer la mejor oferta de vuelos a precios competitivos para los usuarios habituales de esta plataforma.

Otro ejemplo del uso de tecnologías Big Data es el caso de aerolíneas como Air France-KLM que utiliza tecnología de Hadoop como base del sistema de gestión de ingresos de la compañía a nivel corporativo. Las ventajas de Big Data en la toma de decisiones y en la capacidad para anticiparse a las preferencias y hábitos de consumo de los clientes son claves para establecer servicios más diversificados y establecer relaciones más estrechas con los consumidores, gracias a la aplicación de nuevas estrategias en la gestión de clientes, beneficios y operaciones internas. Todo un desafío dentro de un sector tan sensible a los factores externos como el turístico.

El cualquier caso, el salto del sector de los viajes y el turismo al Big Data deberá superar desafíos y obstáculos coyunturales como la fragmentación de los datos a través de múltiples sistemas, las posibles fricciones por la coexistencia de arquitecturas de gestión de Big Data y arquitecturas tradicionales y la escasa oferta de profesionales especializados con perfil de científico de datos para la gestión y análisis de información.

Pero a pesar de estas dificultades iniciales, el marketing de datos ha experimentado un incremento del 227% a lo largo del primer semestre del 2013, según un estudio realizado por BlueKai. La encuesta realizada a directivos de marketing y anunciantes de todo el mundo revela que el 91% de los encuestados afirman que el uso y análisis de datos ocupó un lugar destacado en las estrategias de segmentación y focalización.

Otra de las conclusiones extraídas del estudio afirman que el 87% de los expertos confían en la importancia de los datos recopilados para contacto directo de los clientes, tales como formularios y tráfico web, como un activo importante para las empresas. Los datos más utilizados en la toma de decisiones de las estrategias de marketing alcanzan el 83% en referencia al sitio web, seguido de la del CRM y los datos de registro con un 79%, los datos de correo electrónico con un 72%, un 45% para la búsqueda de datos y un 28% en los datos del sitio o la aplicación móvil. Entre los servicios que ofrece Big Data en el sector turistico estan la optimización de precios, generación de ofertas personalizadas y el análisis de sentimientos.

Mercados financieros

Como afirman los expertos de BNY Mellon [11], "el crecimiento económico mundial probablemente sea más rápido. A cambio, las implicaciones resultantes para los mercados de capital globales son enormes. Construcción de infraestructuras, flujos de capital internacionales, el cambio de divisa, la diversificación de activos, la selección temática, la innovación de producto y, quizás lo más importante, las políticas económicas y financieras dependerán de los resultados de los nuevos métodos de Big Data".

Las implicaciones de esta revolución también pueden transformar totalmente la manera actual de interpretar los mercados financieros. "Varias de las docenas de interpretaciones de la tasa mensual de empleo en Estados Unidos son probablemente innecesarias" gracias al Big Data, según la gestora neoyorquina.

Por otra parte, la publicación de datos macro (PIB, inflación, PMIs), "puede volverse más certera y menos sorprendente". Desde BNY Mellon van todavía más lejos: "El Big Data reesculpirá la industria de gestión de activos", puesto que se utilizarán nuevos acercamientos a la información disponible en "búsqueda, análisis, distribución, trading y gestión del riesgo". Sus expertos creen que las nuevas herramientas permitirán abandonar definitivamente los aforismos bursátiles y permitirán analizar con todavía más detalle tanto los componentes fundamentales de las compañías y del crédito como la diferenciación temática.

También consideran que un procesamiento más rápido y eficiente incrementará la dificultad de los gestores para generar alfa; en un mercado donde los principales condicionantes de la volatilidad serán sólo movimientos irracionales o eventos geopolíticos no deseados, creen que se volverá cada vez más común la gestión pasiva vía indexación o ETF "en detrimento de la diferenciación temática".

Entre las aplicaciones de Big Data en el sector financiero se pueden citar los servicios de protección de marca, protección ante riesgos y fraude, servicios personalizados a clientes, búsqueda de patrones de uso de productos financieros, marketing personalizado, creación de servicios basados en la localización, etc.

6.8. Algunos casos de éxito en la aplicación de Big Data

- **Vestas Wind Systems**, fabricante de aerogeneradores danés[12] utiliza el análisis de los datos, entre los que se incluyen la temperatura, las precipitaciones, la velocidad del viento, la humedad y la presión atmosférica, para determinar la ubicación óptima de un aerogenerador. Gracias al uso de una solución de Big Data en un superordenador, y de una solución de modelado diseñada para aprovechar la información de un amplio conjunto de datos entre los que se incluyen datos estructurados y no estructurados, ahora la empresa puede ayudar a sus clientes a optimizar la ubicación del aerogenerador y, como resultado de ello, su rendimiento.

- **Automercados Plaza´s** [13] una cadena familiar de tiendas de alimentación de Venezuela, se dió cuenta de que disponía de más de seis terabytes de información sobre productos y clientes almacenada en diferentes sistemas y bases de datos. Al integrar la información de toda la empresa, la cadena de tiendas ha visto cómo sus ingresos aumentaban en aproximadamente un 30% y su rentabilidad anual se incrementaba en 7 millones de dólares. Por ejemplo, la empresa ha evitado pérdidas en aproximadamente el 35% de sus productos ahora que pueden programar reducciones de precio para vender productos perecederos antes de que se estropeen.

- **Netflix** [14] Tras las decisiones de Netflix se esconce un profundo análisis del comportamiento y los gustos de sus clientes. Netflix es un servicio para ver películas y series en streaming. Lo que tal vez no se sepa es que la serie House of Cards (en su versión norteamericana) es una producción de la propia Netflix, en la que se gastó 100 millones de dólares. Semejante cantidad es sin duda una inversión arriesgada. ¿Por qué Netflix se decidió a producir su propia serie, cuando por menos dinero podría haber comprado los derechos

de otras series de probado éxito? Sencillamente porque sabía que la serie sería un éxito. ¿Y cómo lo sabía? Gracias al Big Data. Netflix analiza cuántos espectadores han visto una serie completa, qué día y a qué hora ven un episodio, desde qué dispositivos, cuándo paran o aceleran la reproducción, y hasta qué hacen cuando llegan los títulos de crédito, para ver si el espectador quiere ver otro episodio cuando termina o cierra la aplicación. Hay más datos: Netflix pide a sus clientes que valoren su interés en diferentes géneros o películas que han visto. El algoritmo de recomendación les sugiere después títulos adaptados a sus gustos. Y parece que acierta bastante, ya que el 75% de lo que ven los usuarios procede de las recomendaciones. El procesamiento de toda esta información fue clave para que Netflix tomara una de las decisiones estratégicas más importantes de su historia: compitió con canales como HBO o AMC para hacerse con los derechos para EEUU de la serie inglesa House of Cards.

- **Smart Cities**. Big Data para las ciudades del siglo XXI. Los días 16 y 17 de abril de 2014 se celebró la cumbre "City & Big Data" en Singapur [15]. Esta cumbre en Singapur ha tenido un programa dedicado a varios de los puntos clave en la discusión sobre los usos de "Big Data" para las ciudades: el análisis predictivo, el manejo de la complejidad, seguridad y redes (cámaras, sensores), administración del desempeño global de una ciudad, formas de optimizar la infraestructura existente, las minas de datos de censos, métodos de visualización del espacio urbano y la llamada "city cloud": la "nube" que permite crear ventajas competitivas para una ciudad.

El desarrollo y uso de Sistemas de Información Geográfica (GIS -Geographic Information System) ha sido uno de los puntos fuertes mencionados en la sesión de apertura por el experto Peter Quek, a cargo del departamento que dirige las actividades de reestructuración de desarrollo en Singapur. El objetivo es maximizar las posibilidades de GIS para diseñar "comunidades vivibles", teniendo en cuenta dos condiciones: Singapur es uno de los países más densamente poblados del mundo y, por ende, la cantidad limitada de tierra obliga a destinar el suelo y uso de recursos con mínimos "márgenes de error".

De acuerdo con su presentación, el país está usando tecnología punta 3D para orientar a las personas que trabajan en planificación urbana para hacer simulaciones que permitan apreciar por adelantado los posibles impactos de las construcciones y hacer estudios más detallados sobre desarrollos urbanísticos propuestos (Más información en FutureGov Asia). Ejemplos del uso de estos sistemas de grandes datos los ofrecen también ciudades como Lyon, en donde la municipalidad se ha asociado con IBM para crear una plataforma que ayuda a los operadores responsables del tráfico a predecir congestiones y actuar para reducirlas (cambiando programación de tiempos de semáforos, por ejemplo).

En Boston, la Oficina de Nuevas Mecánicas Urbanas tiene un programa llamado "Adopte un hidrante" (toma de agua) mediante el cual se han localizado más de 13 mil hidrantes en toda la ciudad y se invita a la población a que adopte uno o varios para dejarlos al descubierto en caso de nevadas y tormentas. En los Estados Unidos, Seattle ha iniciado con Microsoft y Accenture un proyecto piloto para reducir el uso de energía eléctrica en un 25% mediante un programa que recoge y analiza datos sobre los equipos y funcionamiento de edificaciones en el centro de la ciudad, con el fin de establecer cuáles funcionan adecuadamente, cuáles no y cómo cambiarlos para un ahorro eficiente. Chicago es una de las ciudades del mundo prominentes en la creación y aprovechamiento de conjuntos de datos, con varios proyectos que incluyen grupos de investigación de la Universidad de Chicago. En noviembre 2013, como parte de la serie "Discoveries" de conferencias UChicago, se llevó a cabo una interesante discusión al respecto, moderada por el profesor Charlie Catlett, director del Centro Urbano de Computación y Datos (UrbanCCD, de UChicago y el Laboratorio Nacional Argonne, fundado en 2012). Se mostraron algunos ejemplos de casos en la ciudad, como por ejemplo estudios sobre las relaciones de llamadas a la línea 311 que informaban sobre ausencia o vandalismo con los contenedores de basura y aparición de ratas en distintos sectores de la ciudad. Esto ayudó al gobierno local a actuar rápidamente para prevenir un aumento del problema y cortar la cadena de zonas infestadas de roedores [16].

7. Una nueva y urgente profesión: Científico de Datos.

McKinsey Global Institute [17] relaciona el aumento del volumen de información con la demanda de expertos en extraer valor de los datos y asegura que en 2018 habrá un desfase de entre el 50 y 60% entre demanda de talento analítico para acometer proyectos y la oferta real de profesionales preparados para abordar tal tarea. Esto quiere decir que serán necesarios alrededor de 490.000 profesionales para diseñar estrategias Big Data en Estados Unidos pero que tan sólo habrá 300.000 para cubrir la demanda (Ver figura 8).

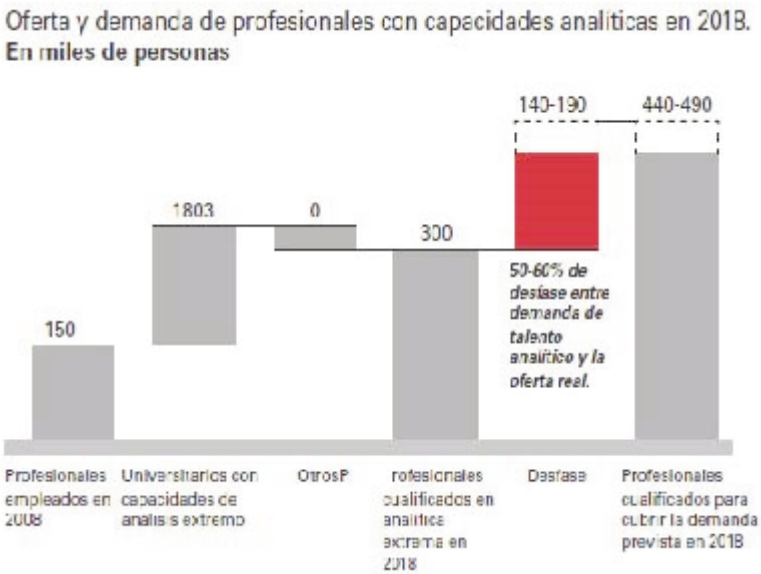

Figura 8. Perfiles demandados: estadísticos, científicos cuantitativos, analistas, managers con enfoque y experiencia cuantitativa y técnicos expertos en software y lenguajes de programación de análisis de datos (fuente: Oficina de Estadística de Empleo de Estados Unidos; Censo de EE.UU.; Dun & Bradstreet; McKinsey Global Institute).

Las capacidades del llamado científico de datos se concretarían en un perfil mixto que integre conocimientos tecnológicos con comprensión del negocio. El CIO (Chief Information Officer) ha acabado por estar vinculado a la parte más puramente tecnológica, lo que le convierte más en un CTO (Chief Technology Officer). Para entender la naturaleza y magnitud de Big Data se necesitaría una figura nueva: el Chief Data Officer capaz de entender la naturaleza de los datos y organizarlos y explotarlos para obtener un impacto positivo.

8. Recomendaciones

El análisis realizado por IBM de las conclusiones del estudio "Big Data @ Work Study"[3] ha proporcionado nuevos conocimientos acerca de cómo las empresas promueven sus iniciativas de Big Data en cada fase. Impulsadas por la necesidad de superar los retos empresariales, y a la vista de las tecnologías en desarrollo y de la naturaleza cambiante de los datos, las empresas están comenzado a estudiar más de cerca las posibles ventajas de Big Data. Para obtener más valor de Big Data se ofrece un amplio abanico de recomendaciones a las empresas a medida que avanzan en la implementación de Big Data.

- **Centrarse en el cliente**

Es fundamental que las empresas centren sus iniciativas de Big Data en ámbitos que puedan proporcionar el máximo valor para el negocio. Para muchos sectores, esto significará comenzar con una analítica de clientes que permita prestar un mejor servicio a los mismos como resultado de comprender verdaderamente sus necesidades y ser capaces de anticiparse a sus comportamientos futuros.

- **Desarrollar un proyecto de Big Data para toda la empresa**

Un proyecto abarca la visión, la estrategia y los requisitos de Big Data dentro de una empresa y resulta fundamental para armonizar las necesidades de los usuarios de negocio con la hoja de ruta de la implementación de TI. Crea una comprensión común de cómo la empresa pretende utilizar Big Data para mejorar sus objetivos de negocio. Un proyecto efectivo define el alcance de Big Data dentro de la empresa al identificar los retos empresariales clave a los que se aplicará, los requisitos de proceso de negocio que definen cómo se utilizarán esos datos masivos y la arquitectura que incluye los datos, las herramientas y el hardware necesarios para lograrlo.

- **Comenzar con los datos existentes para lograr resultados a corto plazo**

Para poder lograr resultados a corto plazo, al mismo tiempo que se crea el impulso y la experiencia para respaldar el programa de Big Data, resulta fundamental que las empresas adopten un enfoque pragmático. El lugar más lógico y rentable para comenzar a buscar estos nuevos conocimientos es dentro de la empresa. La mayor parte de las organizaciones desean hacer esto para aprovechar la información almacenada en repositorios existentes, a la vez que amplían su(s) data warehouse(s) para poder gestionar volúmenes y variedades de datos más grandes.

- **Desarrollar funcionalidades analíticas**

Las empresas tendrán que invertir en adquirir tanto herramientas como habilidades. Como parte de este proceso se espera que surjan nuevos roles y modelos de trayectorias profesionales para individuos con el equilibrio necesario de habilidades analíticas, funcionales y de TI. Centrar la atención en el desarrollo profesional y el avance de la trayectoria de los analistas internos, que ya están familiarizados con los retos y procesos de negocio únicos de la empresa, debería ser una prioridad para los directivos empresariales. Al mismo tiempo las universidades y los propios individuos, independientemente de su formación o especialidad, tienen la obligación de desarrollar sólidas habilidades analíticas.

- **Crear negocios sobre resultados cuantificables**

Desarrollar una estrategia de Big Data exhaustiva y viable, así como la posterior hoja de ruta requiere un caso de negocio sólido y cuantificable. Por lo tanto, es importante contar con la implicación y el respaldo de uno o más directivos empresariales a lo largo de todo el proceso. Igual de importante para lograr el éxito a largo plazo es una colaboración empresarial y de TI continua y sólida. Muchas empresas basan sus casos de negocio en las siguientes ventajas que se pueden derivar de Big Data:

- **Decisiones más inteligentes**: Aprovechar nuevas fuentes de datos para mejorar la calidad de la toma de decisiones.
- **Decisiones más rápidas**: Permitir una captura y análisis de datos en tiempo más real para respaldar la toma de decisiones en el "punto de impacto", por ejemplo cuando un cliente está navegando por su sitio web o al teléfono con un representante del servicio de atención al cliente.
- **Decisiones que marquen la diferencia**: Centrar las iniciativas de Big Data en ámbitos que proporcionen una verdadera diferenciación.

9. El futuro de Big Data

Presentamos las predicciones para Big Data en 2014 que realiza la editora del portal siliconangle.com, entre las que se encuentran (a) que el Big Data estará moldeado por la demanda de los usuarios para el data blending, (b) que el Big Data necesita convivir bien con otros, (c) que no se puede estar preparado para el hoy con herramientas de ayer; y (d) una rápida innovación por la comunidad de Big Data open source [18].

Predicción 1:

La "curva de potencia" Big Data en 2014 estará determinada por la demanda de los usuarios de negocios para la mezcla de datos. Clientes como Andrew Robbins de Paytronix y Andrea Dommers-Nilgen de TravelTainment, quien recientemente habló de sus proyectos en Pentaho en eventos en Nueva York y Londres, ambos provienen de la parte empresarial y están logrando las metas específicas para sus empresas mediante la mezcla de los grandes datos y datos relacionales. Se está aprovechando también los datos mezclados (blended) para obtener nuevos conocimientos con una vision del cliente, incluyendo la capacidad para analizar los patrones de comportamiento de los clientes y predecir la probabilidad de que estos puedan disfrutar de ofertas específicas.

Predicción 2:

Big Data tiene que jugar bien con los demás. Históricamente, los proyectos de grandes volúmenes de datos se han sentado en gran medida en los departamentos de TI debido a las habilidades técnicas necesarias. Los clientes deberán elegir entre las diversas tecnologías comerciales y de código abierto, incluyendo las distribuciones de Hadoop, bases de datos NoSQL, bases de datos de alta velocidad, plataformas de análisis, y muchas otras herramientas y plug-ins. Pero también deben tener en cuenta la infraestructura existente, incluidos los datos relacionales y los almacenes de datos.

El lado positivo de esta elección y de la diversidad es que, después de décadas de tiranía y "lock-in" impuesta por los proveedores de software de empresa, en adelante se desplazará aún mayor poder adquisitivo a los clientes. También significa que las TI estarán buscando herramientas Big Data para ayudar a implementar y administrar estas complejas arquitecturas. Incumbirá a los proveedores de tecnología Big Data la tarea de jugar bien con los demás y trabajar en pro de la compatibilidad. Después de todo, se trata de la capacidad de acceder y gestionar la información de múltiples fuentes que agreguen valor a la analítica Big Data.

Predicción 3:

Se verá una innovación aún más rápida de la comunidad de código abierto de Big Data. Nuevos proyectos de código abierto tales como como Hadoop 2.0 y YARN, harán la infraestructura Hadoop más interactiva. Nuevos proyectos de código abierto como STORM (protocolo de comunicaciones) funcionarán más en tiempo real, trabajarán bajo demanda sus análisis de la información en el ecosistema Big Data.

Desde que se anunció el primer conector nativo Hadoop en 2010, se ha estado trabajando con la misión de hacer la transición hacia arquitecturas Big Data más fácil y con menos riesgo, en el contexto de este ecosistema en expansión. En 2013, se han hecho algunos avances enormes en esta dirección. Esto permite a los departamentos de TI sentirse más inteligentes, más seguros y con más confianza en sus arquitecturas y abrir soluciones Big Data para los que se dedican a este negocio.

Predicción 4:

No se puede preparar el mañana con herramientas de ayer. Se sigue perfeccionando plataforma (Pentaho) para apoyar el futuro de la analítica. Se va a lanzar una nueva funcionalidad, mejoras y plug-ins para que sea aún más fácil y más rápido mover, mezclar y analizar fuentes relacionales y Big Data. Se esta planeando mejorar las capacidades de la capa adaptativa de datos y que sea más seguro y fácil para los clientes gestionar el flujo de datos. Por lo que respecta al análisis, se esta trabajando para simplificar la búsqueda de datos sobre la marcha, de todos los usuarios, y hacer que sea más fácil encontrar patrones y anomalías en la captura. En Pentaho Labs [19], van a seguir trabajando con los primeros usuarios para cocinar nuevas tecnologías que aporten cosas tales como datos predictivos de la máquina y análisis en tiempo real de la producción

10. Conclusión y trabajos futuros

Dado el condicionante sobre la extensión del presente documento, su contenido es necesariamente incompleto ya que no nos ha sido posible profundizar más en los temas descritos anteriormente, así como incluir otros ángulos de visión. El tema Big Data está generando una enorme cantidad de literatura, que como decimos no nos es posible abarcar.

Hay, sin embargo, un tema al que deberíamos dirigir nuestra mirada: se trata del estado actual de las diversas plataformas de software / hardware existentes o en proceso de desarrollo, que son las que harán posible capturar, limpiar, analizar y generar nuevos conocimientos de los grandes conjuntos de datos que se almacenan constantemente en sus diversas fuentes Big Data. Entre ellas podemos citar Pentaho, Talend Open Studio, Intel Data Platform, SAS the power to know, Soluciones de Big Data de IBM basadas en POWER8, etc. Casis todas ellas se basan en la infraestructura de software Hadoop de Apache. De momento, dada la extensión del presente documento no nos es posible adentrarnos en dicho campo, que se deja para un futuro

Cómo citar este artículo / How to cite this paper

Paredes-Moreno, A. (2015). Big Data: Estado de la cuestión. International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC), Vol. 2, Num. 1, pp. 38-59. Consultado el [dd/mm/aaaa] en www.ijisebc.com

Referencias

[1] Oracle, BIG DATA y su impacto en el negocio, Una aproximación al valor que el análisis extremo de datos aporta a las organizaciones, Mayo 2012.
[2] "Big Data: The next frontier for innovation, competition, and productivity", verlo en la dirección web http://goo.gl/ltgbb4 .
[3] "Analytics: el uso de Big Data en el mundo real; Cómo las empresas más innovado¬ras extraen valor de datos inciertos", IBM Institute for Business Value, Said Business School (University of Oxford Autores: Michael Schroeck, Rebecca Shockley, Dra. Janet Smart, Dolores Romero-Morales y Peter Tufano, 2012. Ver la dirección web http://goo.gl/Q7LHH5.
[4] Ver página web de Datacenter Dinamics en http://goo.gl/8XtlVa .
[5] Qué es Big Data? Véase http://goo.gl/QZZ4iv .
[6] Ver el informe de Cisco en http://goo.gl/jSozfq.
[7] "Understanding Big Data, Analytics for Enterprise Class Hadoop and Streaming Data, Paul Zikopoulos, Chris Eaton, Tom Deutsch, Dirk Deroos and George Lapis, Mc Graw Hill, New York, ISBN 978-0-07-179053, 2012.
[8] Encuestas globales de CEOs http://goo.gl/eGI5MZ.
[9] Véase la página de Amadeus en http://goo.gl/MgnIna .
[10] "La experiencia de viajar se transformará con Big Data" en Blog Think Big, Vease la página
http://blogthinkbig.com/viajar-con-big-data/
[11] "Cómo revolucionará el Big Data la gestión de activos", en Funds People, ver página http://goo.gl/ykPDsA .
[12] Véase la página de Vestas Systems en http://www.vestas.com/
[13] Véase la página de Automercados Plaza's en https://www.elplazas.com/
[14] "El Big Data está detrás del éxito de Netflix", en Baquia. Véase la página en http://goo.gl/vtBlCB
[15] "Big Data para las ciudades del siglo XXI", Lina María Aguirre, 29/04/2014, en http://blogs.lavanguardia.com/tecladomovil/?p=1799
[16] Para más información sobre el programa de UChicago ver "Chicago: City of Big Data" en http://goo.gl/Pwoj2u.
[17] "Big Data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute. Véase en la página http://goo.gl/Ua3D5Z.
[18] "2014 Technology Predictions Series: Pentaho on Big Data" by Suzanne Kattau. Véase en la página http://goo.gl/lsRmWh
[19] La página de puede consultarse en http://www.pentaho.com/labs