La creación de valor en las empresas a través del Big Data

Universitat de Barcelona

Autor: Aleix Galimany Suriol
Director: Jordi Bachs Ferrer
Departamento: Economía y Organización de Empresas
Enseñanza: Grado de Administración y Dirección de Empresas

Julio de 2014

datos
data
big
análisis
información
puede
empresas
pueden
forma
mediante
almacenamiento
ejemplo
ser
diferentes
web
solo
usuarios
obtener
redes
día
grandes
hacer
cada
tipo
facebook
sensores
nuevos
bases
sociales
cantidades
aprendizaje
hadoop
des
veces
valor
nuevas
google
poder
empresa
algoritmos
tiempo
patrones
obtención
proceso
tecnología
concepto
permite
mayor
tener
organizaciones
productos
podemos
desarrollo
nodos
actualmente
algoritmo
clientes
business
posibles
sistema
objetivo
capacidad
gran
real
hoy
the
memoria
masivos
además
cualquier
si
hace
uso
actual
hecho
muchas
of
cambio
procesamiento
salida
fuentes
herramientas
hdfs
entrada
compra
mapreduce
privacidad
cantidad
trabajo
campos
detección
volumen
todas
búsqueda
organización
parte
personas
costes
decir
comportamiento
transacciones
explotación
palabras
podríamos
medida
twitter
analizar
aplicación
acciones
vez
mundo
and
sistema
cliente
tecnologías
ventas
recogida
móviles
importante
protección
reconocimiento
probabilidad
proyecto
caso
mercados
llegar
coste
misma
software
variables
cosas
imágenes
adopción
futuros
ejemplos
importantes
base
tradicionales
velocidad
clave
junto
estructurados
conjunto
página
podría
distribuida
amazon
sino
así

Resumen y Palabras Clave

En el presente trabajo se pretende determinar, analizar y justificar la aportación que el Big data puede hacer en las empresas mediante la creación de valor, al tratarse de un concepto relativamente novedoso se definirá el Big data, atributos diferenciales, su propósito y aplicaciones relacionando conceptos del Big Data con las metodologías tradicionales de explotación de datos, como Big data puede convertirse en una fuente de ventajas competitivas, Big data con ejemplos prácticos, el escenario actual del Big data en las empresas, los nuevos desafíos y oportunidades que plantea y Big Data des de una perspectiva de ética y responsabilidad.

Big Data, Ventajas Competitivas, Datos, Información, Valor Añadido, Análisis, Innovación, Hadoop, Decisiones, Data Mining, BI

English:

Title: Creating Business Value through Big Data

Summary and Keywords:

The present project aims to identify, analyse and justify the contribution of the relatively new concept of the Big Data. First of all will define the Big data, differentiating attributes, purpose and applications, relating concepts of Big Data with traditional architectures and methodologies using data. Then will analyse which ways and through which concepts and technologies Big Data can become an important source of competitive advantages. Later will analyse the contribution of Big Data in profit of efficiency and innovative businesses with illustrative examples of startups companies; top technology companies like Google, Facebook, Twitter, Linkedin, Yahoo; as well as non-technology sectors such as Amazon or Rolls Royce, without forget to explain the basis of the technologies that currently supports the Big Data. Finally in the last part of the project a vision of the future challenges posed by Big Data to organizations and society in general will be provided. Also Big Data will be considered from the perspective of ethics and responsability of using data.

Big Data, Competitive Advantages, Data, Information, Value Added, Analysis, Innovation, Hadoop, Decisions, Data Mining, BI

Agradecimientos

Cariñosamente a Carmen por su paciencia en esos días duros.
A mi padre, aunque no estés aquí, por todas estas oportunidades que no pudieron llegar.

Índice

1. Introducción

En el Fórum Económico Mundial celebrado el 2012 en la ciudad Suiza de Davos se destacó el potencial del Big Data literalmente como "*A new class of economic asset, like currency or gold*"[1]. Big Data es el concepto que promete ser la siguiente gran tendencia "*The Next Tink*", y es que con la visión adecuada, los datos pueden utilizarse de forma inteligente por parte de las empresas e instituciones públicas y ser una gran fuente de información de las organizaciones y de necesidades potenciales de los consumidores, la utilización de Big Data puede proporcionar importantes ventajas competitivas para las empresas y las organizaciones, puede transformar los mercados, impulsar nuevos servicios e innovación.

Hoy en día poseemos datos (información) de infinidad de cosas: geoposicionamineto de personas, de los coches, de los carros de compra en un supermercado, rutas de los vehículos, frecuencias con las que acudimos a un sitio o hacemos determinadas cosas, localizaciones, fechas y términos de búsqueda en el buscador de Google, el significado y motivación de las búsquedas, nº de personas que introducen un/unos términos de búsqueda, blogs, wiki's, información publicada de las webs, flujos clicks, navegación en una página web, *twits*, estados emocionales, *tags*, compras habituales, precios, lecturas RFID y NFC, datos de mercados financieros, imágenes, documentos escaneados, amigos que uno tiene en Facebook, publicaciones, "me gusta" de Facebook, películas que vemos, libros que leemos, y que nos gustan, videos que reproducimos, canciones que escuchamos, programas de tv que visualizamos, personas a las que seguimos en las diferentes redes sociales, afinidades y relaciones, a quien llamamos, a qué hora, día y des de donde lo hacemos, edades, sexo, biometría, des de nuestras constantes vitales, ADN, reconocimiento facial, reconocimiento de voz, historial médico, tipo de escritura, la electricidad que consumimos y con qué patrón lo hacemos, que días, donde y a qué horas compramos determinados productos o servicios, datos meteorológicos, temperatura de una pieza de una máquina y sus vibraciones y cualquier otro dato que podamos llegar a recoger a través de sensores... Dejemos volar la imaginación con las posibilidades de recogida de datos, combinaciones y análisis de los mismos y posibles conclusiones.

[1] http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development [Un nuevo activo económico como el dinero o el oro]

2. Que es Big data

2.1 Definición y conceptos Generales

Actualmente no existe un consenso claro entre los diferentes autores sobre una definición de Big Data[2] generalmente aceptada, los autores aquí analizados definen el Big Data des de perspectivas diferentes, aunque todas ellas complementarias. Para el propósito de este trabajo y en un intento de hacer síntesis de entre las diferentes definiciones que dan varios autores y de una forma muy general podríamos definir el Big Data como "grandes volúmenes de datos" o "datos masivos", y todas las cosas que se pueden hacer con estos datos pero a gran escala. Pero Big Data no es solo datos: podríamos decir que el Big Data se reparte entre "los datos" y todos los procesos que se agrupan alrededor de estos datos, des de la recogida de los mismos, el almacenamiento y su posterior procesamiento o análisis, todo ello con objetivo de extraer valor de los mismos. Des de una perspectiva más conceptual se puede decir que el Big Data no es tan grande pensando en términos absolutos de una gran cantidad de datos (que los es), más bien es el concepto relativo de recoger casi el "todo" acerca algo y su posterior análisis.

De las distintas definiciones cabe destacar la interesante definición propuesta por la consultora Gartner[3] sobre las tres características que marcan la diferencia frente a otros conceptos como el *Datamining* o el *Business Intelligence*, estas son: Volumen (cantidad), Velocidad (en la creación y utilización) y Variedad (fuentes de datos heterogéneas), también llamadas "3V" o "V^3". Estas tres características constituyen la verdadera revolución del Big Data. Adicionalmente a las "3V", IBM[4] añade una cuarta y quinta característica del Big Data de notoria importancia: Valor (eficiente y rentable) y Veracidad (fiable con datos intrínsecamente imprecisos).

Siguiendo con la definición propuesta por Gartner y ampliada por IBM, analizamos que compone el Big Data:

Volumen: Hoy en día las empresas y el sector publico generan cantidades de datos imposibles de imaginar hace unos años, se ha pasado en poco tiempo de los Gigabytes a los Terabytes, a los Petabytes y recientemente a los Zettabytes (1.099.511.627.776 Gigabytes) y quizás en un futuro no muy lejano a los Yottabytes. Hacia el año 2020 se prevé alcanzar los 44 zettabytes de información almacenada en todo el mundo[5]. En 2008 solo Google

[2] *Big:* Grande. *Data:* Datos.
[3] http://www.gartner.com/technology/topics/big-data.jsp
[4] http://www.ibm.com/big-data/us/en/
[5] IDC The Digital Universe of Opportunities April 2014: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

procesaba 24 Petabytes de datos al día[6], la cantidad de datos almacenados crece de forma exponencial cada año.

Velocidad: La velocidad se ha impuesto en la obtención de los datos, internet ha contribuido en gran medida a ello, cada vez es más fácil obtener datos o mediciones con una frecuencia mayor, por ejemplo, hoy en día con el avance de los chips GPS y el alcance de las redes inalámbricas podemos medir la posición de un vehículo con acusada precisión cada pocos segundos y almacenarla con suma sencillez. De la misma manera la velocidad es sumamente importante en la transmisión y en el procesamiento y análisis de los datos, 5 minutos pueden llegar a ser excesivos para detectar un posible fraude en una transacción, otras veces el análisis de la información exige hacerse en tiempo real, Google sugiere palabras para nuestra búsqueda a medida que vamos escribiendo, no pasados 15 segundos. La no disponibilidad de la información en el momento oportuno no solo resta valor a la misma sino que inutiliza su aplicación.

Variedad: Big Data se nutre de fuentes de datos muy heterogéneas, recientemente la capacidad de análisis de datos se limitaba solo a los datos estructurados y almacenados en bases de datos relacionales[7], con Big Data se introduce la capacidad de combinar datos de orígenes distintos y de formatos no homogéneos ni predefinidos, esto aumenta exponencialmente las posibilidades de análisis y da juego a la creatividad. Por ejemplo en un análisis se puede combinar las trazas[8] de las estaciones emisoras de señal móvil con las *MAC address*[9] de los dispositivos móviles y los datos meteorológicos para localizar, medir y pronosticar el tráfico de vehículos en una ciudad sin necesidad de instalar ninguna aplicación en los dispositivos móviles.

Valor: Quizás la característica más importante y el propósito final de los datos, los datos en crudo carecen de valor, es el hecho de combinarlos, transformarlos y analizarlos lo que permite extraer el valor latente que hay en ellos, las organizaciones buscan la forma de convertir los datos en información de forma fiable, pero la explotación de los datos siempre debe ser coherente, pensada y alineada con la estrategia de la empresa para que pueda ser aprovechable, rentable y eficiente, en caso contrario se puede llegar a incurrir en unos costes sin crear valor para la organización.

Veracidad: Referida al nivel de fiabilidad asociada a los datos, obtener datos de calidad es indispensable para la correcta obtención de información, el concepto "*Garbage in, garbage*

[6] http://tech.fortune.cnn.com/2011/04/07/whats-12-petabytes-to-apple/
[7] Basadas en el modelo Relacional: Basado en el uso de tablas, registros y columnas
[8] Trazas: ficheros habitualmente de texto sin formato donde cada línea del documento suele constituir un registro de un evento de un dispositivo o proceso, son generadas normalmente en tiempo real a medida que se ejecuta una acción, la información no suele estar tabulada y normalmente suelen almacenarse en la memora interna de los dispositivos, no están enfocadas al tratamiento para la obtención de información, más bien a la trazabilidad y al depurado de las acciones, valores, errores, etc de un dispositivo o aplicación.
[9] MAC Address: Conjunto de números y letras que constituyen un identificador de dispositivo de red único a nivel mundial.

out" muy utilizado en el campo de la información significa que la entrada de datos sin sentido provoca la salida de información también sin sentido, este concepto junto con otro más novedoso, "*Garbage in, gospel out*" en donde se tiende a aceptar ciegamente los datos o información de salida informatizada aun que provenga de datos de pobres. En algunos tipos de datos se pueden aplicar métodos para "limpiarlos" pero a veces esto es imposible, no se puede eliminar la imprevisibilidad inherente a variables como la economía, condiciones climáticas, las futuras decisiones de compra de los clientes… La necesidad de asumir y tolerar esta incertidumbre es una de las características distintivas del Big Data.

Otro concepto que Big Data explota son los datos no estructurados, a diferencia de los datos estructurados que son especificados con detalle y tienen un esquema y estructura fijos pensados para ser almacenados en las tradicionales bases de datos relacionales, los datos no estructurados no están sujetos a un tipo o estructura fija predefinida y controlada, a diferencia de las tablas, los datos no estructurados se almacenan como "objetos" u "documentos" sin campos definidos, pueden ser des de fotografías, audio, video, documentos, imágenes, mails, mensajes instantáneos de por ejemplo WatsApp o Facebook Messenger, libros, trazas, artículos… La mayoría de datos e información que poseen las empresas suele estar diseminada por toda la organización y además es de tipo no estructurado.

En el actual contexto tecnológico y social podemos llegar a recoger datos de casi todo, y muchas veces de forma pasiva para los usuarios, los flujos del cursor en una web, los clics, las coordenadas del GPS, los datos de los sensores… son recogidos muchas veces de forma inadvertida para los usuarios y con suma sencillez consiguen documentarlo casi todo sobre cualquier acontecimiento. Los teléfonos móviles que des de nuestros bolsillos nos acompañan a todas partes son un claro ejemplo, cuando salimos de casa para ir al trabajo o a la universidad los modernos smartphones con más o menos precisión pueden geolocalizarnos durante el recorrido por triangulación con los repetidores de señal móvil sin necesidad de GPS. Los poseedores de *smartphones*/*tablets* Android con el servicio de ubicación activado, o de iOS con aplicaciones de Google instaladas pueden sorprenderse e incluso quedarse perplejos al consultar el historial que se almacena sobre su ubicación en un día y hora concreta del año[10], pueden ver el recorrido que hicieron ese día y cuánto tiempo estuvieron en un lugar concreto, y así con todo un historial de sus movimientos que puede llegar hasta el año 2011 como se puede apreciar en la Ilustración 1. Algo similar sucede con la configuración de anuncios de Google donde se presentan algunos datos fruto de concusiones a las que llega Google con nuestros datos[11]. Con todo este lujo de detalles ahora no tenemos por qué ceñirnos a una muestra de los datos, podemos llegar a procesar y analizar la totalidad de datos sobre nuestras ubicaciones, esto nos trae a la luz detalles que con una granularidad superior pasarían inadvertidos o no serían recogidos al analizar

[10] Historial de ubicaciones de Google: https://maps.google.com/locationhistory/b/0
[11] Configuración de anuncios de Google: http://www.google.com/settings/ads

determinados subgrupos, y es que muchas veces los eventos verdaderamente relevantes o importantes suceden en los lugares insospechados que las muestras tomadas no consiguen captar. Poseer y analizar la totalidad de los datos nos permite tomar otra perspectiva y detectar conexiones y detalles que de otro modo no hubiesen salido a la luz, pensemos en el ejemplo de la geolocalización, si tomamos pequeñas muestras cada 5 o 10 minutos sobre el posicionamiento de un individuo podemos perdernos detalles verdaderamente relevantes sobre los comportamientos, y es que muchas veces los valores atípicos suelen ofrecer la información más interesante. A día de hoy con los datos masivos y las tecnologías de almacenamiento y proceso existentes, analizar el conjunto integro de datos en vez de usar el atajo de una muestra aleatoria está al alcance de cualquier organización, pero en cualquier caso debemos tener especial cuidado ya que cuando amentamos de escala el número de datos, suelen aparecer todavía más correlaciones espurias que nos pueden hacer parecer que ciertos fenómenos pueden estar conectados, cuando en la realidad esto puede no ser verdad.

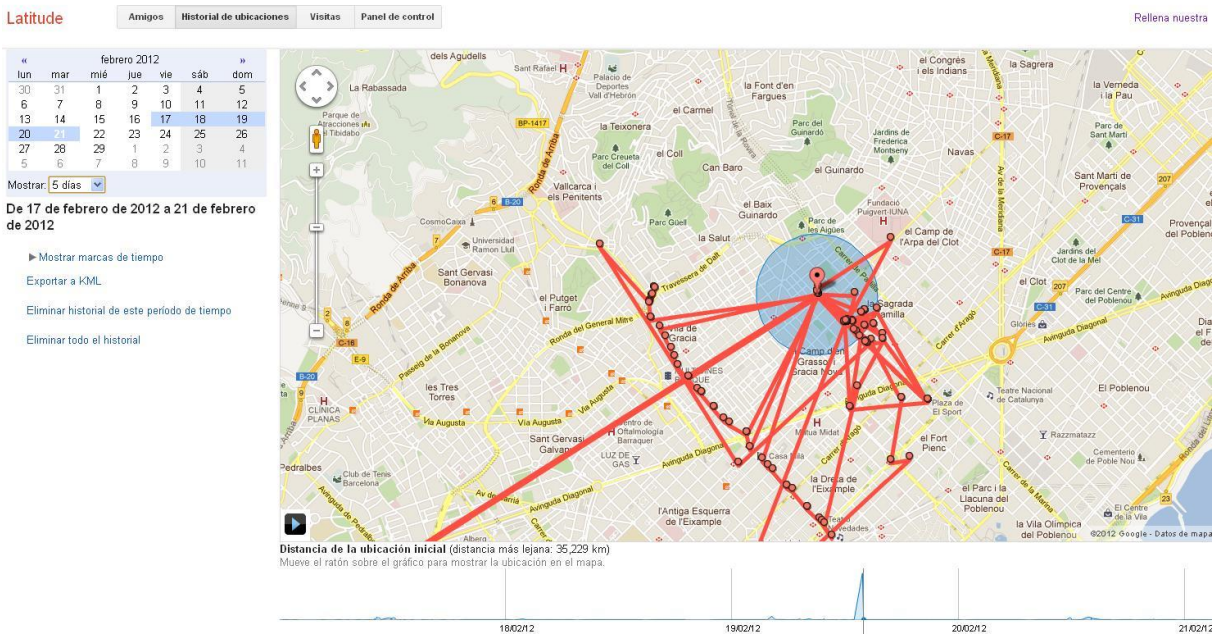

Ilustración 1

Otro concepto poderoso es la perspectiva en el uso de los datos, el mero hecho de disponer de la capacidad de almacenamiento y computación para amasar y analizar muchos datos no es garantía de éxito, sin la perspectiva sobre los posibles usos de los datos, como detalla Mayer “A veces las restricciones con las que vivimos, y que presumimos idénticas para todo, son, en realidad, únicamente funciones de la escala a la que operamos” esto nos lleva muchas veces a ignorar el valor y la capacidad de los datos que podemos tener delante, las organizaciones diseñan la recogida de datos de los procesos con una finalidad ligada al proceso, pero una perspectiva de datos más amplia nos puede desvelar posibilidades insospechadas para el propósito inicial de recogida de datos. Tener una mentalidad y una visión de datos masivos son claves para el aprovechamiento del Big Data.

En definitiva, Big Data implica también un cambio de mentalidad importante frente a visiones más tradicionales de los datos y su análisis, por ejemplo: La tolerancia al desorden y la imprecisión de los datos, la posibilidad de acercarse al análisis de un casi exhaustivo "todo" en vez de un subconjunto o una muestra representativa de este en beneficio de explorar el matiz, la reutilización de los datos…

2.2 Historia y evolución

La historia de Big Data empieza varios años antes del actual boom, cuando el término "Big Data" es por primera vez empleado en 1997 con relación al Big Data tal y como lo conocemos ahora, en un estudio de la NASA de dos investigadores, Michael Cox y David Ellsworth[12], para referirse a la generación de ingentes cantidades de información cuando simulaban en los supercomputadores de la época el flujo de aire alrededor de las aeronaves. En 2001 la consultora Gartner define el modelo "V3" en la publicación "3D Data Management: Controlling Data Volume, Velocity, and Variety"[13]. En 2004 Google crea MapReduce, un nuevo paradigma del procesamiento distribuido de grandes cantidades de datos, un año después, en 2005 Yahoo crea la solución Hadoop para su proyecto de motor de busqueda, basada en el funcionamiento de MapReduce de Google y Hadoop Distributed File System (HDFS) para el almacenamiento, posteriormente fue cedido con licencia Open Source[14] a la Apache Software Foundation para que esta comunidad continuase su desarrollo y lo distribuyese libremente con descarga gratuita, este último hecho supone el principio de la explosión del Big Data. Ocho años después de que Michael Cox y David Ellsworth nombrasen el Big Data como una cantidad ingente de datos. En 2004 con el surgimiento de la web 2.0 empieza el auge de los blogs y las redes sociales y es cuando el volumen de datos empieza a aumentar exponencialmente a medida que crecen los usuarios de estas y crece el rastro de datos que a su vez estos dejan, en la web, en las redes sociales, en transacciones, en geoposicionamiento… La creciente obtención de datos es tan grande que agota el potencial de las infraestructuras IT tradicionales. Según IBM el 90% de los datos del mundo actual han sido creados solo en los últimos dos años[15]. Posteriormente el abaratamiento de los sensores y su miniaturización así como la interconexión de los objetos están contribuyendo a lo que actualmente conocemos como el internet de las cosas otra fuente importante de datos para el Big Data.

[12] http://scn.sap.com/community/business-trends/blog/2013/11/20/big-data-to-the-rescue
[13] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
[14] Software desarrollado y distribuido libremente y con acceso al código fuente programado (Wikipedia).
[15] http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html

Y es que hace diez o quince años hubiese sido imposible el Big Data, ciertos cambios han impulsado, y están impulsando decisivamente el Big Data, la cantidad de datos e información ha aumentado tanto hasta alcanzar los límites de la tecnología clásica, hasta el punto que la información ya no cabía en la memoria que utilizan los servidores para procesarla, por lo que los ingenieros han tenido que modernizar las herramientas para poder procesarla, la capacidad de computación, el almacenamiento, la conectividad y la reducción de los costes, además de una nueva perspectiva sobre nuevos y posibles usos de los datos han impulsado el surgimiento de nuevas tecnologías de procesamiento como MapReduce o Hadoop y de almacenamiento como HDFS.

2.3 Entorno actual

Sin la evolución de la tecnología hasta tal y como la conocemos ahora Big Data no existiría, Gordon Earl Moore, cofundador de Intel anunció en 1965 la ley de Moore pronosticando que cada dos años se duplicaría el número de transistores en un circuito integrado[16] esta observación se ha venido cumpliendo de forma bastante aproximada, logrando aumentar la capacidad de proceso de forma exponencial así como la disminución también exponencial de los costes de producción de los circuitos integrados. Esto ha permitido afrontar la potencia de proceso necesaria para analizar las ingentes cantidades de datos del Big Data, el aumento de la velocidad y la capacidad de las memorias RAM[17] así como su abaratamiento también han sido determinantes como se verá más adelante, y sin lugar a dudas la capacidad de almacenaje ha jugado un papel muy importante para poder almacenar los grandes volúmenes de datos manejados por las plataformas Big Data. Pero el salto cualitativo ha venido de la mano del almacenamiento y procesamiento distribuido, juntando la capacidad de proceso de muchos ordenadores (servidores) con un hardware[18] de bajo coste se puede amasar una capacidad de proceso importante. La misma analogía se puede aplicar para los clústeres[19] HDFS. Otro concepto ya usado en los sistemas BI[20] y tecnológicamente relevante para Big Data, es la tecnología "en memoria" que permite el almacenaje de la información en la memoria RAM para aprovecharse de la mayor velocidad de esta frente a la de los discos duros (100.000 veces más rápida), en beneficio de un acceso a los datos más rápido, que a su vez revierte en una velocidad de procesamiento más rápida. *Cloud Computing* es otro concepto que de la mano de la virtualización y del *Software as a Service* (SaaS) facilita el

[16] http://en.wikipedia.org/wiki/Moore%27s_Law
[17] Memoria de acceso aleatorio: "memoria de trabajo para el sistema operativo y el software"
[18] Hardware según wikipedia: "las partes tangibles de un sistema informático" más vulgarmente "la electrónica" haciendo referencia a un equipo informático.
[19] Según Doctor Thomas Sterling: "una clase de arquitectura de computador paralelo que se basa en unir máquinas independientes integradas por medio de redes de interconexión, para obtener un sistema coordinado, capaz de procesar una carga"
[20] Business Intelligence

uso de Big Data en cualquier organización: *Cloud Computing*[21] de forma resumida, se puede decir que es la plataforma que aloja toda la infraestructura de una solución Big Data (almacenaje, proceso, gestión, seguridad…) pero con acceso des de internet. Añadámosle a la plataforma de Big Data el funcionamiento del *Cloud Computing* y la característica de ser proveído por otra organización, como si se tratase de un outsourcing en forma de servicio y tenemos el SaaS de Big Data.

Pensemos en el entorno social: Hoy en día cada uno de nosotros posee un móvil o smartphone y quizás una *tablet*, un portátil, un ordenador de sobremesa en casa y otro en el trabajo, tenemos un coche con ordenador de abordo[22] una Smart TV en casa… La tecnología está en todas partes, y se ha adoptado y asumido como algo natural, cada vez empezamos a utilizar más gadgets[23] tecnológicos y a más temprana edad, cada vez nos apoyamos más en la tecnología, nos aporta comodidad nos proporciona conectividad y seguridad… El desarrollo de la tecnología y el hecho de haberla adoptado en aplicaciones que nunca hubiésemos imaginado está permitiendo la recogida de datos e información de infinidad de cosas e interacciones que nutren las bases de datos de los Big Data. Cada vez hay más iniciativas de Big Data orientadas a los ciudadanos, des de informaciones del tránsito rodado en tiempo real, previsiones sobre el comportamiento de los precios de viajes en avión y de otros viene, etc.

En la actual situación económica donde el nivel de competitividad es más que vital, hallar nuevas fuentes de información y explotarlas para obtener información en el momento oportuno puede ser determinante en el apoyo de la toma de decisiones por parte de los directivos de las organizaciones. La reciente recesión ha obligado a las organizaciones a hacer cambios radicales, incluso a reinventarse, por lo que la capacidad de estudiar grandes cantidades de datos, tanto internos de la organización, como del entorno puede permitir conocer mejor el entorno y entender la situación actual de la empresa en el contexto y usar esta información en su beneficio. Además la reciente e incipiente recuperación económica no viene para traer calma, más bien al contrario, por lo que obtener el conocimiento de que se está produciendo en esta recuperación y como afecta a la empresa puede ser determinante para gestionar los cambios necesarios y posicionarse con ventaja en los escenarios futuros. Y no solo se limita a mejorar las organizaciones existentes: Big Data también ha sido impulsor de *startups* basadas en el uso o análisis de datos de formas innovadoras

[21] Según el NIST es "Un modelo que permite el acceso ubicuo, adaptado y bajo demanda en red a un conjunto compartido de recursos de computación configurables compartidos (redes, servidores, equipos de almacenamiento, aplicaciones y servicios) que pueden ser aprovisionados y liberados rápidamente con el mínimo esfuerzo de gestión o interacción con el proveedor de servicio"http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
[22] http://ww2.autoscout24.es/glosario/ordenador-de-a-bordo/183317/
[23] http://es.wikipedia.org/wiki/Gadget

El abaratamiento del hardware en general también ha contribuido a la adopción del Big Data, el coste de almacenamiento y de proceso ha caído drásticamente y también lo ha hecho el de las comunicaciones, de la misma manera el hecho de que las herramientas y plataformas para el análisis de grandes volúmenes de datos sean *Open Source* las hace asequibles y accesibles tanto para las grandes empresas, como para las pequeñas.

En cuanto al entorno legal, en España a pesar de la Ley Orgánica de protección de datos de carácter personal (LOPD) y la Ley de servicios de la sociedad de la información y del comercio electrónico, actualmente no hay un marco legal específico para Big Data. El derecho a la protección de los datos está exclusivamente constituido sobre el concepto de la privacidad, las leyes sobre privacidad regulan la cesión de los datos de carácter personal o de personas razonablemente identificables, mediante la imposición de obligaciones legales de notificación y consentimiento entre otras. Actualmente existe en el marco de la Comisión Europea un grupo de trabajo para el artículo 29 (WP29[24]) integrado por las Autoridades de Protección de Datos de los estados miembros y el Supervisor Europeo de Protección de Datos dedicado a promover y aconsejar a los estados sobre la protección de datos. En 2012 la Comisión Europea inició una reforma de la legislación comunitaria para la protección de los datos personales donde más adelante el WP29 introduzco el concepto Big Data y pronunció una serie de opiniones encaminadas a la protección de la privacidad en el marco del Big Data.

2.4 Datos, obtención y almacenamiento

Big Data es en parte resultado y evolución sobre la ancestral misión humana de medir, registrar y analizar el mundo. Hoy en día capturamos información de todo lo que nos rodea, este afán de recabar datos acerca todo es definido por Mayer con el término "Datificar", cuyo objetivo final es extraer el valor latente de estos datos. El primer salto cuantitativo en la *datificación* sucede en la década de los 90 con la digitalización de textos y continúa con la digitalización de música, videos e imágenes, hasta la actualidad donde se recogen ingentes cantidades de datos de forma pasiva sobre clics, *twit's* vibraciones de un motor... Con la Digitalización se consiguió que la información analógica fuese comprensible para los ordenadores facilitando su almacenamiento y procesamiento con un menor coste y una mayor rapidez. Este salto que supuso la digitalización no hubiese sido tal sin la tecnología de reconocimiento óptico de caracteres (OCR) mediante la cual se convierte una imagen sobre un texto tipográfico en datos con la identificación de los caracteres en la imagen. OCR permite transformar una imagen de un texto digitalizado que solo es comprensible y útil a

[24] http://ec.europa.eu/justice/data-protection/article-29/index_en.htm

los ojos de una persona, en datos comprensibles para los ordenadores, esto permitió que la información fuese indexable y analizable mediante algoritmos por ordenador lo que permitió hacer búsquedas en el texto. Google desarrollo el proyecto "*Google Print*" o "*Google Book Search"* en el cual se propuso involucrar editores universidades y bibliotecas para poder digitalizar y convertir en texto todos los libros del mundo[25]. La digitalización de libros y su posterior transformación en texto nutre las inmensas bases de datos de los Big Data pero a su vez OCR también se nutre de Big Data para la detección y corrección de errores (*Inteligent carácter recognition* o ICR) en la conversión de las imágenes a texto mediante inteligencia artificial (*Machine Learing*)[26].

Al auge de la datificación también han contribuido tecnologías como el GPS, como se ha comentado anteriormente la geolocalización ha permitido rastrear a la gente, obtener patrones de comportamiento, obtener cuantas veces por mes una familia va a comprar al supermercado, registrar rutas de distribución en las empresas de mensajería para optimizarlas, registrar el estado del tráfico en las ciudades… Otra gran fuente que contribuye hoy en día a la datificación son todas las interacciones que ocurren en las redes sociales. Esta tendencia de datificación ha dado lugar a las ingentes cantidades de datos (datos masivos) que pueden explotarse cada vez con una mayor profundidad para tener una nueva perspectiva de la realidad.

Pero otras veces la obtención de los datos no es tan visible como en los casos mencionados, a veces esta obtención se hace de forma inadvertida sin que los usuarios o poseedores de flujos de información sean conscientes del valor que hay en la reutilización de los datos. Un ejemplo son los datos sobre los términos de búsqueda que introducen los usuarios de Google, los coches de StreetView no solo recogían imágenes, el famosos *Recaptcha*[27] o las entidades emisoras de tarjetas como Visa o Mastercard, en otros casos, empresas importantes de prestación de servicios han establecido contratos con otras empresas para prestar servicios de *outsourcing* con el objetivo de interponerse de forma inadvertida entre el flujo de datos y poderlos obtener para su explotación sin que este fuese el propósito principal.

Con el fin de proporcionar una idea de posibles orígenes de datos para Big data y fomentar la generación de ideas sobre posibles usos, en los siguientes puntos se hará un análisis de algunos de ellos:

Datos de RFID y NFC: se refieren respectivamente a *Radio Frequency Identification* y *Neard Field Comunication,* son tecnologías de identificación mediante radiofrecuencia. RFID funciona mediante etiquetas, normalmente adhesivas colocadas en objetos (por ejemplo en paquetes de mensajería) estas etiquetas son escaneadas sin necesidad de contacto, a una

[25] http://www.google.com/googlebooks/about/history.html
[26] http://en.wikipedia.org/wiki/Optical_character_recognition
[27] https://www.google.com/recaptcha/intro/index.html#the-recaptcha-advantage

distancia de entre 10 y 30cm mediante un lector especifico el cual envía una señal de radiofrecuencia a la etiqueta y esta le devuelve la información de identificación, las etiquetas más habituales suelen ser de tipo pasivo, son las más baratas y no llevan ningún tipo de batería. El sistema RFID está siendo implementado en muchos sectores, des de etiquetado de productos, logística y control de stock automatizado, pagos, seguimiento e identificación de paquetes, control de accesos, automatización de procesos productivos… En muchos casos se pretende obtener datos de localización de los objetos, saber que objeto es y donde está para poder tener una trazabilidad y todo esto de forma desatendida, inalámbrica y a un coste muy pequeño por etiqueta. NFC va un paso más allá e incorpora la capacidad de comunicación bidireccional entre dos chips NFC de forma inalámbrica a un máximo de 10cm, pensada para un intercambio rápido pero de pocos datos, aunque más que una etiqueta RFID. NFC actualmente se utiliza para los pagos mediante tarjetas Contactless, control de accesos, transferencia y sincronización entre dispositivos.

Datos de la web: En 2005 Google examinó de la web en búsqueda de documentos traducidos para alimentar su sistema de *Machine Learing* del servicio *Google Translate*, utilizó 200 mil millones de palabras procedentes de documentos traducidos y publicados en la web de las Naciones Unidas[28], estas traducciones en varios idiomas y de excepcional calidad fueron recogidos y subministrados como input al sistema. La obtención de datos de las webs se hace mediante "arañas web" o "*Web Crawlers"* que consiste en programa o aplicación que navega de forma automatizada en la *World Wide Web*, la aplicación empieza examinando unas primeras webs proporcionadas. Examina la primera web y descarga y almacena una copia o "imagen" de la web en su base de datos, después el programa busca todos los enlaces que contiene la copia descargada y examina el primer enlace y descarga la web destino y así recursivamente repite el proceso con cada uno de los enlaces y destinos web a los que apuntan los enlaces web. De unas primeras webs proporcionadas, la aplicación va extendiéndose cada vez más a medida que va explorando nuevas webs construyendo además un grafo con las relaciones entre las webs y su contenido. La cantidad de datos obtenidos en este método es inmensa pero altamente desestructurada y heterogénea. Esta metodología es utilizada por Google para indexar las webs para su buscador o para otros propósitos como predecir acontecimientos futuros, la evolución de la bolsa e incluso "la conciencia global de nuestra civilización"[29]. La *Apache Software Foundation* lidera el proyecto de una araña web de código abierto Nutch[30] por lo que obtener datos de la web es posible con los conocimientos e infraestructura Big Data adecuada sin incurrir en excesivos costes.

Datos de Analítica Web: Provienen de analizar el tráfico de usuarios en una web de una organización, son tanto cuantitativos como cualitativos, esta analítica se basa en obtener observaciones de métricas y KPI. Por métricas podemos entender una medida cuantitativa

[28] http://blogoscoped.com/archive/2005-05-22-n83.html
[29] Global Consciousness Project Meaningful Correlations in Random Data: http://noosphere.princeton.edu/
[30] http://nutch.apache.org/#What+is+Apache+Nutch%3F

sobre un aspecto o estado de la página web, algunos ejemplos pueden ser: nº de visitantes únicos, nº de visitas, tiempo de navegación en una página concreta del sitio, y en el sitio, tasa de conversión (el porcentaje de resultados u objetivos cumplidos por número de visitantes), tasa de salida (como el porcentaje de visitas que ha abandonado el sitio des de una página concreta), tasa de rebote (como el porcentaje de vistas que ha visto una sola página de todo el sitio y ha abandonado rápidamente la página sin hacer ningún clic en ella), país o región de origen de las visitas, porcentaje de visitas originadas en buscadores (des de que buscador, des de enlaces patrocinados o no), "compromiso" entendido como la cuantificación de la fidelidad de los usuarios, o por ejemplo el hecho de que un visitante único nos visite cada día, que lo haga varias veces al día, que visite habitualmente ciertas secciones cuando entra en la web…. Los KPI (*Key Performance Indicator* o Indicadores de Rendimiento) son indicadores relevantes que permiten cuantificar o determinar el estado de cumplimiento de un determinado objetivo estratégico de la empresa, se dice que los KPIs deben cumplir el criterio SMART (Específicos, Medibles, Alcanzables, Relevantes y a Tiempo) a su vez, algunas métricas pueden ser simultáneamente KPI cuando son suficientemente relevantes para un objetivo, ejemplos sobre KPI en analítica web pueden ser: nº de formularios de compra completados por semana, importe promedio de venta, la tasa de conversión, el coste de campaña por venta…

Datos de las Redes Sociales: A Facebook, creada hace tan solo 10 años se suben más de diez millones de fotos nuevas cada hora, en Youtube se sube más de una hora de video cada segundo, cada día se escriben 500 millones de Tweets. La cantidad de datos que se generan en las redes sociales está creciendo cada día llegando incluso hasta la sobrecarga o exceso de información, por lo que en este tipo de datos es muy importante "separar la señal del ruido". Con los datos de redes sociales sucede en cierta medida lo mismo que en los datos de analítica web, hay una vertiente que se basa en métricas e indicadores KPI, pero los datos de redes sociales también permiten datificar interacciones con los stakeholders. Ejemplos de métricas en las redes sociales pueden ser nº de *followers* en Twitter, seguidores en Facebook, nº de publicaciones relacionadas con la empresa, *Topic Tends*, videos más vistos, procedencia de los seguidores o fans, dispositivo de acceso. Ejemplos de KPI podrían ser nº de retweets, nº de me gustas de Facebook que han acabado comprado un producto, nº de publicaciones compartidas… Otra vertiente puede ser analizar el contenido de las interacciones no como KPI's o métricas, sino pensando en relacionarlas con sentimientos, preferencias, hábitos, opiniones, tendencias…

Datos de telecomunicaciones: Con el auge de las comunicaciones tanto móviles como fijas las operadoras de telecomunicaciones han pasado a tener un papel protagonista en la obtención de datos para Big Data, des de las llamadas de voz, los SMS, a que destinos llamamos y quien nos llama, des de donde lo hacemos, a qué hora, con qué frecuencia, cuando hacemos uso de las comunicaciones de datos, que tipo de datos usamos

(navegación, Facebook, Twitter, juegos) y cuando los usamos, geoposicionamiento según los repetidores a los que estemos conectados, por donde nos movemos. Las operadoras pueden encontrar nuevos negocios con la venta o licenciamiento de datos. Telefónica por ejemplo ha creado la filial Telefónica Dynamic Insights dedicada a la explotación de los datos generados por sus redes móviles[31].

Datos de las Smart Cityes: Las ciudades inteligentes generan también grandes cantidades de datos, originadas en áreas como:

Seguridad Ciudadana: Geolocalización de coches de patrulla, bomberos, cámaras de video vigilancia, sensores de movilidad y tráfico, sensores de fuego, sensores de alertas, centros de alertas…

La Movilidad urbana: Sensores de tráfico, semáforos, geolocalización de los diferentes vehículos de transporte público y sus rutas, sensores de ocupación en el transporte público, información meteorológica, localización de aglomeraciones de personas mediante geoposicionamiento móvil por triangulación de repetidores, datos de redes sociales…

La Gestión de la energía: Datos del consumo de energía a partir de contadores inteligentes (*Smart meter*), sensores de iluminación, sensores de presencia de personas, iluminación inteligente.

La Gestión del agua: Datos sobre la calidad del agua, consumos, detección de fugas, cámaras de video vigilancia en las plantas de distribución y potabilización de agua.

La Gestión de los residuos urbanos: Sensores en los contenedores para medir el estado de capacidad en tiempo real, geoposicionamiento de camiones de recogida, datos sobre aglomeraciones de personas en eventos, alertas sobre estado de las calles o espacios públicos, información meteorológica.

El Análisis de los sentimientos ciudadanos: A través de las redes sociales se puede obtener datos para cuantificar la opinión de turistas o ciudadanos sobre aspectos de la ciudad.

Datos de Sensores (M2M) internet de las cosas: Sin lugar a duda el abaratamiento y la miniaturización de los sensores está permitiendo su rápida proliferación, esto está produciendo un aluvión de datos sobre aspectos que hasta ahora no se podían recoger o no con el detalle experimentamos hoy en día, además estos sensores y dispositivos están conectados a internet permanentemente permitiendo la obtención y el análisis de los datos en tiempo real, la inteligencia e interactividad entre ellos. Ejemplos de orígenes de sensores pueden ser: sensores de geoposicionamiento, consumo de energía, sensores médicos,

[31] http://dynamicinsights.telefonica.com/479/about-us

sensores en edificios, sensores de automatización, monitorización industrial, sensores de diagnóstico, sensores meteorológicos, monitorización de objetos, móviles, sensores en smartphones, electrodomésticos, vehículos, instalaciones…

Datos de transacciones: Las empresas de *retail* están empezando a percatarse del potencial de los datos que se poseen y se recogen sobre las ventas, des de cuestiones relacionadas con la logística como los stocks, la localización de productos, volumen de las transacciones, productos comprados conjuntamente, perfil del comprador edad, sexo, país de origen, tipo de producto o gama comprada, donde gastan más dinero los compradores, canal de venta, satisfacción con la compra, devoluciones de compras, productos examinados o visitados… Otro ejemplo es el sistema interbancario de transferencias SWIFT, el hecho de posicionarse como intermediario en el flujo de datos de transferencias monetarias permite obtener grandes cantidades sobre transacciones, des de volumen de las transferencias, orígenes, destinos, nº de operaciones. Estos datos podrán correlacionarse con datos sobre la economía para poder hacer predicciones. Algo similar sucede con las transacciones de los bancos o los emisores de tarjetas, estos pueden recoger datos de lugares de compra, actividad económica del comercio donde se ha realizado la compra, tipo de tarjeta, hora de la compra, volumen de la compra, perfil de cliente, país de origen de la tarjeta, edad, sexo, poder adquisitivo, patrones de consumo. Las órdenes de los mercados bursátiles también son importantes fuentes de datos, de ellas se puede obtener datos sobre tipo de transacción, mercado, Símbolo *ticker*[32], datos económicos de la empresa, eventos sobre la empresa, tipo de activo, pay-out, valoración, rangos de cotización estático y dinámico, precio, indicadores bursátiles, volumen de transacciones, tipo de orden, fecha y hora, origen de la operación, estado de ejecución modificación o cancelación, tendencias en precios, volatilidad, ordenes ocultas, figuras de análisis técnico, *figuras doji*[33], eventos inusuales, ordenes puestas rápidamente y canceladas rápidamente para evaluar el mercado mediante sistemas de trading automático en alta frecuencia…

Datos de los gobiernos: Los gobiernos poseen ingentes cantidades de datos sobre los ciudadanos, empresas, organizaciones, turistas… A diferencia de las empresas, los estados tienen capacidad coercitiva para obligar a los ciudadanos y empresas a proporcionar sus datos. Esto les permite obtener un abanico de datos muy amplio sobre diferentes perspectivas. Un pequeño detalle de algunos de los datos que poseen los gobiernos podría ser: Todos los datos personales, datos catastrales, datos médicos, datos de la renta, datos sobre nuestros vehículos, datos sobre nuestros estudios, datos sobre las cuentas anuales de las empresas, datos sobre nuestros trabajos, propiedades que poseemos, litigios, infracciones o multas, viajes al extranjero, datos de transporte público, datos sobre economía nacional, datos sobre transacciones, datos sobre votaciones, datos sobre presupuestos, datos sobre mercados regulados…

[32] Identificador único de un activo en un mercado.
[33] http://stockcharts.com/school/doku.php?id=chart_school:chart_analysis:candlestick_pattern_

Los gobiernos también se están dando cuenta del valor de los datos que poseen, estos datos proporcionados por los ciudadanos u organizaciones deben ser gestionados en beneficio público. Hasta ahora eran explotados de forma interna por los gobiernos, pero ahora estos datos están siendo liberados de forma agregada para que cualquier pueda explotarlos como conocimiento, con iniciativas altruistas o empresariales para sacar el mayor rendimiento de estos. Estas iniciativas de datos abiertos u *open data*[34] están empezándose a promoverse por parte de gobiernos tanto estatales como locales, ejemplos de *open data* se pueden hallar en numerosas instituciones públicas de diferentes regiones: Estados Unidos (https://www.data.gov/), Reino Unido (http://data.gov.uk/), La Unión Europea (https://open-data.europa.eu/en/data) España (http://datos.gob.es/), Catalunya (http://dadesobertes.gencat.cat/es), Barcelona (http://opendata.bcn.cat/opendata/es/) o Terrassa (http://opendata.terrassa.cat/) por citar algunos ejemplos próximos.

Wearable's: El futuro está por descubrir, pero el auge de Wearable's[35] empieza a calar, dispositivos como las Google Glass[36], el reloj Galaxy Gear[37], zapatillas, camisetas y pulseras inteligentes permitirán nuevas formas de interacción y recopilación de datos en cualquier situación, cuando practicamos deportes, o monitorizar nuestras constantes en situaciones concretas, por ejemplo cuando estamos estresados.

Licenciamiento de datos: Aparte de las iniciativas *open data,* las organizaciones que por sus actividades recogen grandes cantidades de datos relacionados directamente o indirectamente con su negocio están en una posición de ventaja frente a otras que no disponen de grandes flujos de datos, sin embargo estas últimas (al igual que las primeras) pueden acceder a fuentes de datos ajenas mediante el licenciamiento de estos, existen empresas "supermercados de datos" dedicadas a proveer datos, por ejemplo algunas filiales de grandes empresas se dedican a comercializar datos agregados sobre transacciones, telecomunicaciones, logística[38]… permitiendo que cualquiera puede hacerse con datos agregados de compras con tarjetas, posicionamiento de usuarios, interacciones en redes sociales[39] entre otros. Pero también existen organizaciones y empresas que proveen datos de forma gratuita, Facebook o Twitter con ciertas restricciones permiten acceso gratuito a

[34] Joyanes: información general que puede ser utilizada libremente, reutilizada y redistribuida por cualquier persona y que puede incluir datos geográficos, estadísticos, meteorológicos, proyectos de investigación financiados con fondos públicos.
[35] Dispositivo electrónico de reducidas dimensiones y que se viste el usuario, también llamado "ropa electrónica" o "dispositivos corporales": prótesis inteligentes, relojes, pulseras, gafas, anillos, cascos, auriculares, monitores de constantes vitales...
[36] http://www.google.com/glass/start/what-it-does/
[37] http://www.samsung.com/es/consumer/mobile-phone/wearables/
[38] http://www.google.com/publicdata/directory?hl=en_US&dl=en_US#! , http://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1 , http://www.infochimps.com/ , http://datamarket.com/ , https://www.qunb.com/ , http://www.quandl.com/ , http://www.thinknum.com/, http://www.xdayta.com/
[39] https://datasift.com , http://gnip.com/ , http://www.spinn3r.com/

sus datos mediante API's[40] (https://developers.facebook.com/tools/ y https://dev.twitter.com/console)

Otros tipos de datos: Cualquier organización puede acumular grandes cantidades de datos por el mero hecho de desarrollar sus operaciones, estos pueden ser de tipos y orígenes muy distintos, Para Amazon podrían ser estadísticas de lecturas de libros, páginas que releen los usuarios, para Telefónica podrían ser el uso de ciertos puertos TCP[41] o la categorización de las incidencias de su servicio de atención al cliente, para Carrefour las rutas que siguen los consumidores en sus supermercados, para Volkswagen podrían ser datos de los robots en su cadena de montaje, para UPS la localización de sus flotas o los tiempos de entrega de paquetes, para King el tiempo que los usuarios pasan jugando al Candy Crush Saga en un nivel, para Bayer el tiempo de vida de un componente farmacéutico, para Torres la graduación de la uva, para el F.C. Barcelona podría ser el porcentaje de pases fallados por jugador y partido, para Abertis el nº de coches por hora que atraviesan un peaje, para CaixaBank el importe de las provisiones por morosidad, así con una infinidad de posibilidades para cada tipo de empresa u organización, la limitación podríamos decir que muchas veces solo la impone nuestra imaginación!

Otra piedra angular del Big Data es el almacenamiento, el surgimiento de tecnologías como el almacenamiento distribuido sin duda ha permitido el desarrollo de Big Data, las bases de datos relacionales han dado paso a las modernas bases de datos capaces de almacenar grandes cantidades de datos de tipo no estructurados y de forma distribuida, básicamente existen 4 tipos diferentes en función de las necesidades particulares de almacenamiento de los sistemas Big Data, de los cuales haremos una breve reseña: NoSQL, "En Memoria", MPP y en caché. Las primeras como su nombre indica, no (solo) utilizan el lenguaje de consultas SQL (que si es utilizado en las bases de datos de tipo relacional), utilizan el leguaje UnQL, están optimizadas para manejar grandes volúmenes de datos de forma distribuida en lugar de garantizar la consistencia en ellos como lo hacen las bases de datos relacionales, además no requieren de estructuras fijas como tablas, en este tipo de bases de datos la información se almacena en forma de:

- Clave-Valor: Cada valor almacenado tiene una clave relacionada, para simplificarlo mucho podríamos decir, que por ejemplo que una clave podría estar formada por el NIF de una persona + un atributo y esta calve estaría relacionada con un valor, de esta manera tenderíamos la siguiente estructura: Clave_Atributo="Valor":

[40] Aplication Programming Interface: Interfaz de Programación de Aplicaciones
[41] Transmission Control Protocol. Protocolo usado en Internet por ejemplo

12345678Z_Nombre=”Aleix” | 12345678Z_Apellido1=”Galimany” | 12345678Z_Apellido2=”Suriol”.

- Documentos: Para simplificar se podría decir que los documentos están dentro de un espacio llamado “contenedor” o “cajón”, dentro de cada cajón puede haber un documento con una estructura y tamaño diferente. Existe una clave que permite identificar y recuperar el documento que hay en el contenedor, en este símil el nº de cajón.
- Grafos: Utilizadas para almacenar datos que tienen tipología de grafo, la información se representa en forma de nodos con sus atributos, y las aristas que relacionan los nodos también con los atributos que las definen, permiten utilizar las teorías de grafos para trabajar con los datos. Un ejemplo muy simple podría ser una familia donde cada miembro sería un nodo con sus atributos (nombre, edad, sexo…) estos nodos estarían unidos mediante aristas que podrían representar la relación entre ellos (padre, hijo, abuelo, nieto…).

Las bases de datos NoSQL permiten su escalado de forma fácil al trabajar con los datos de forma distribuida (los datos se almacenan en diferentes nodos que pueden estar en diferentes ubicaciones y ser de hardware heterogéneo). Estas bases de datos son utilizadas por empresas que manejan grandes volúmenes de datos como Twitter, Facebook o Linkedin, con estructuras que van cambiando a lo largo del tiempo, además, las bases de datos NoSQL permiten utilizar las consultas de tipo MapReduce que se explicaran más adelante.

Las bases de datos “en memoria” en cambio están optimizadas para ejecutar operaciones con los datos a gran velocidad, aptas para procesamientos de datos en tiempo real. Basadas en el almacenamiento de los datos en la memoria RAM o en memorias flash o discos duros SSD para aprovechar su mayor velocidad frente a los tradicionales discos duros mecánicos. Dentro de las bases de datos “en memoria” existen dos subtipos: De memoria pura, los más rápidos, que se basan en cargar todo el modelo de datos a la memoria RAM antes de ejecutar cualquier consulta y *just-in-time* basadas en cargar a la memoria RAM solo la parte necesaria para una consulta en particular.

Por otro lado, los sistemas de bases de datos MPP como sus siglas indican (*Massive* Parallel Processing) se basan en el procesamiento paralelo masivo de datos. En este tipo de base de datos todo el conjunto de datos se trocea y se distribuye en piezas independientes con su correspondiente almacenamiento y unidad de proceso. Haciendo una analogía podríamos decir que los datos se distribuyen en muchos ordenadores personales, donde cada uno de ellos aloja una pequeña parte de todo el conjunto de datos. La clave de este sistema se basa en el enfoque “divide y vencerás” donde una consulta compleja sobre una gran cantidad de datos exigiría una gran capacidad de almacenamiento y proceso en un solo servidor, ahora

se puede trocear y ejecutar de forma distribuida en muchos servidores de bajo coste, logrando un mejor aprovechamiento de recursos y una disminución de los costes.

Las bases de datos en Caché son un subtipo de las de "en memoria", las cuales siguen la idea del almacenamiento en RAM pero limitándose a almacenar en memoria RAM solo consultas o conjuntos de datos usados repetitivamente, pero costosos de generar o agrupar y que previamente han sido generados con el objetivo de reutilizarlos en futuros procesamientos.

2.5 Procesamiento de datos y análisis de información

A grandes rasgos y de forma muy superficial el procesamiento de datos y análisis de información consiste en "aplicar las matemáticas" a enormes cantidades de datos para poder inferir probabilidades, pero el mero hecho de tener muchos datos y analizarlos no implica novedades si aplicamos las tradicionales metodologías que se viene utilizando en la empresa, Franks resume el cambio de paradigma del Big Data respecto al análisis de datos con una comparativa: si cambiamos nuestro viejo televisor por una Smart TV con 3D pero solo la usamos para ver la televisión por antena o cable solo obtendremos una mejora en la calidad de imagen, pero nos estaremos perdiendo una nueva manera de ver e interactuar con este nuevo televisor. Lo mismo sucede con el análisis de datos, dar el salto a nuevas tecnologías que permitan Big Data sin un cambio de paradigma o de mentalidad y unos objetivos claros de análisis de datos, puede llevarnos a ninguna parte. La puesta a disposición de los analistas de una infraestructura "*Data Sandbox*" ("campo de pruebas") permitiendo a los analistas poder disponer de los datos en una plataforma unificada y suficientemente flexible para permitir análisis exploratorios con datos propios o externos, puede ser un importante motor que permita desarrollar el potencial analítico de una organización, no debe de haber trabas para el análisis de los datos. El campo exploratorio debe ser amplio, la tecnología actual puede permitir chequear miles de millones de hipótesis a diferentes niveles de profundidad con una rapidez asombrosa, el análisis de los datos tampoco tiene que estar sujeto a escala a la que operamos, Mayer comenta que los zapateros de agua a diferencia de los humanos, pueden caminar por encima del agua sin problemas, para estos es algo natural, de la misma manera los peces están en contacto con el agua pero estos últimos desconocen que están mojados. Estas afirmaciones son fruto de las diferentes escalas con las que miramos las cosas. Presumir ciertas afirmaciones como validas en una escala puede vendarnos los ojos cuando operamos en una escala más grande como son los datos masivos. Citando a Franks podríamos decir que "el valor de un análisis está en mirar los datos de forma diferente".

Otro cambio que conlleva Big Data es la relajación en la constante búsqueda de exactitud, cuando trabajamos con los datos masivos obtener un gran nivel de detalle en los datos puede acarrear una pérdida de precisión en estos, que convenientemente tolerada puede

convertirse en algo positivo, a cambio de tolerar la relajación de errores permisibles obtenemos muchos más datos con detalle, recoger un gran volumen de datos hace que valga la pena renunciar a la exactitud estricta, sacrificamos la precisión de cada punto en beneficio de mayor amplitud y recibimos a cambio un grado de detalle que no habíamos podido observar de otro modo. También renunciamos a la exactitud en pro de frecuencia, de este modo percibimos un cambio que de otra forma se nos hubiera escapado. Muchas veces resulta mucho más provechoso tolerar el error en el análisis que prevenirlo. Es más, a veces incluso estamos dispuestos a sacrificar un poco de exactitud a cambio de descubrir la tendencia general.

Big Data puede proporcionar análisis sobre datos, pero este análisis no hay que confundirlo con reports, los reports son importantes e útiles pero tiene sus limitaciones, hay que diferenciar entre reporting y análisis, reporting (BI Business Inteligencie) se refiere a la ejecución de una serie de informes prediseñados, estos siempre son fruto de peticiones estructuradas de antemano por parte de los usuarios y por tanto son sumamente inflexibles. Si un usuario quiere saber el grado de cumplimiento de las ventas solo tiene que ejecutar un report de ventas por semana y obtiene la respuesta de inmediato, pero el análisis va más allá, un report da datos en respuesta a la cuestión planteada (el porcentaje de cumplimiento sobre el objetivo de ventas), en cambio un análisis da las respuestas necesarias (por qué no se ha alcanzado el objetivo de ventas), un análisis puede utilizar todo cuanto sea necesario para dar respuesta a una pregunta, el análisis se adapta totalmente a la cuestión que se aborda. Un análisis es un proceso interactivo y flexible que busca los datos necesarios para obtener una respuesta a un problema, analizando los datos, interpretando los resultados con el fin de obtener una recomendación para una acción. Un report pude ser el percusor de un análisis, por ejemplo, cuando un supervisor ve algo sumamente extraño en un report sobre calidad, este supervisor se dirigirá a los analistas preguntando por este suceso, en este momento se habrá iniciado un análisis para dar respuesta al porqué de la información extraña, de vuelta tendrá un análisis de la causa de la incoherencia, probablemente se habrá hallado un problema y una respuesta a este.

2.6 Técnicas de análisis y exploración de datos

Imaginemos que disponemos de grandes cantidades de datos, los tenemos almacenados en las bases de datos y tenemos a disposición las herramientas para iniciar un análisis sobre el conjunto de datos, podemos partir de hipótesis *a priori* donde intuimos que podemos hallar información relevante para el negocio, o podemos hallarnos ante un análisis exploratorio donde no hay expectativas *a priori* sobre posibles relaciones por ejemplo, en este punto es donde entra en juego el concepto de minería de datos o Data Mining que involucra campos como las matemáticas, estadística, inteligencia artificial y la administración o manipulación

de bases de datos. A grandes rasgos con la minería de datos se intenta clasificar o predecir en base a los datos. La diferencia entre la minería de datos tradicional aplicada sobre los data warehouse tradicionales y la minería de datos en Big Data radica en procedimientos similares pero sobre grandes cantidades de datos semi-estructurados o no estructurados, almacenados en bases de datos distribuidas y no relacionales (NoSLQ) y en su procesado distribuido basado en MapReduce.

Actualmente en la minería de datos existe un estándar procedimental llamado "Cross Industry Standard Process for Data Mining"[42] o CRISP-DM dividiendo el proceso de minería en 6 fases estructuradas de las siguiente manera: Comprensión de los objetivos y requisitos, comprensión de datos, preparación de datos, modelo, evaluación y despliegue. Este modelo según algunos autores queda obsoleto para los análisis exploratorios y características como el Volumen, la Velocidad y la Variedad del Big Data. Actualmente no hay un estándar procedimental pensado exclusivamente para el análisis de Big Data, pero algunos principios del CRISP-DM siguen siendo de aplicación.

En un intento de hacer accesible la comprensión de los conceptos y algoritmos clave explotados en el análisis de datos o minería de datos, así como para proporcionar algunos posibles detonantes de ideas para la generación de valor, se detalla a continuación un pequeño extracto de algunos con especial relevancia:

Filtrado colaborativo[43]: Como su nombre indica, basado en la colaboración de los usuarios, esta metodología de una forma muy rudimentaria podríamos decir que es una evolución del famoso "boca a boca": Si a mí me ha gustado, y tenemos gustos similares, te lo recomiendo. Bajo este concepto se trata de buscar patrones y utilizar estos patrones por ejemplo para ayudar a la gente a buscar cosas ¡¡que ni siquiera sabían que estaban buscando!! Además el filtrado colaborativo también se puede aplicar en sentido contrario, si a mí no me gusta esto, puede que a ti tampoco, este es un concepto usado también en los filtros de correo basura. Desarrollando más esta idea, imaginemos que los gustos o preferencias de un cliente no son preguntados explícitamente sino que son recogidos mediante datos sobre su comportamiento. El cliente no se percata, pero mediante su imperceptible rastro de datos y el filtrado colaborativo los sistemas son capaces de recomendar un producto relevante cuando el usuario va a comprar otro sin intervención de este. Quizás el mayor hándicap de este método radica en el concepto de similitud, llegar a una definición matemática de similitud que se acerque a la realidad.

[42] http://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
[43] Fuente: http://www.economist.com/node/3714044

Algoritmos de agrupamiento o Clustering: Ampliamente usados en los departamentos de marketing para hacer segmentaciones de mercados, para clasificar el perfil de riesgo de un cliente en un seguro, análisis de imágenes, reconocimiento facial, clasificación genética, análisis de redes sociales pertinencia a grupos, en motores de recomendación para clasificar gustos o preferencias similares. Existen diferentes clases de algoritmos de agrupamiento (*K-means* o k-grupos, *C-means* o agrupamiento difuso, *Hierachical clustering* o agrupación jerárquica) algunos de ellos se basan en medir la cercanía o similitud de los elementos a agrupar, en términos de distancia euclídea, buscando el agrupamiento optimo mediante iteraciones, otros se basan en el agrupamiento mediante densidad en el espacio euclídeo.

Algoritmos de aprendizaje de máquinas o *Machine Learning*: Es uno de los tipos de inteligencia artificial más poderosos, donde los algoritmos aprenden de los datos que procesan y analizan, a mayor cantidad de datos de entrada mejor será el desempeño del algoritmo, la retroalimentación de los datos de salida aumenta la eficacia de los sistemas. Es aplicado en muchos campos, como la detección de fraude, identificar infartos de corazón, control automatizado de accesos o permisos, anuncios personalizados, motores de recomendación, búsqueda de patrones, identificación de figuras o patrones en videos o imágenes, reconocimiento de voz, detección de spam, moderar un foro de discusión web[44], conducción automática de vehículos... Existen cuatro áreas diferentes de machine learning:

1- Aprendizaje supervisado: Para empezar el aprendizaje hay que suministrar al algoritmo datos de entrada junto con la salida que se espera del algoritmo, en función de estos datos subministrados a priori, el algoritmo se "entrena", posteriormente el algoritmo es liberado proporcionándole datos de entrada reales para que obtenga las salidas en base al aprendizaje inicial. A medida que el algoritmo procesa nuevos datos, va perfeccionándose con nuevos juicios basados en los datos existentes y nuevos.
2- Aprendizaje no supervisado: En el aprendizaje no supervisado no existe la fase inicial de aprendizaje donde se subministran datos de entrada junto con las salidas esperadas, en este tipo, el algoritmo tendrá que averiguar por sí mismo la salida des del principio, esto se hará mediante la "agrupación" de datos con patrones similares, la salida serán los diferentes patrones similares hallados. Un ejemplo famoso de aprendizaje no supervisado a gran escala es el que Google hizo con 16.000 CPU's analizando todos los videos de youtube, sin subministrar ningún patrón ni hipótesis previa, algo así como "póngase a mirar y a ver que encuentra…", resultó que uno de los patrones que el algoritmo aprendió a detectar con más fuerza en los videos fueron ¡¡caras de gatos!![45]

[44] Predict which new questions asked on Stack Overflow fòrum will be closed : http://www.kaggle.com/c/predict-closed-questions-on-stack-overflow
[45] http://googleblog.blogspot.com.es/2012/06/using-large-scale-brain-simulations-for.html http://research.google.com/archive/unsupervised_icml2012.pdf

3- Aprendizaje por refuerzo: Se basa en la toma de decisiones mediante el prueba-error, el objetivo es aprender a asignar acciones a las situaciones, con la finalidad de maximizar una recompensa numérica, pero de la misma forma que el aprendizaje no supervisado, no se inicializa el algoritmo diciéndole que acciones tiene que tomar, el algoritmo debe descubrir que acciones producen una mayor recompensa mediante el ensayo prueba-error. En desarrollos más completos de este algoritmo las acciones no solo pueden afectar a la recompensa de salida inmediata, sino que también pueden afectar a la siguiente situación, y por ende a todas las recompensas posteriores, por lo que da lugar a una nueva característica importante: "el arrepentimiento" de las maquinas. El prueba-error y las recompensas retrasadas son los rasgos distintivos más importantes de este aprendizaje. Este algoritmo se usa en investigación de operaciones y en la teoría de los juegos para explicar cómo surge el equilibrio en un juego competitivo bajo unas condiciones de racionabilidad limitada.
4- Aprendizaje profundo[46]: Se basa en la transformación de una observación en una nueva representación diferente a la original de esta, con el objetivo de que esta nueva representación pueda facilitar al algoritmo supervisado (o no), aprender patrones de relevancia. Actualmente los campos de investigación de este tipo de aprendizaje se basan en aprender a encontrar las formas de representación que son más propensas para el aprendizaje. Un ejemplo de transformación es la representación de una imagen en un vector de pixeles[47] para facilitar el reconocimiento de patrones. Algunas aplicaciones de este aprendizaje son la visión artificial, reconocimiento de voz e indexación de textos por relevancia.

Clasificadores Bayesianos ingenuos: Basado en el Teorema de Bayes sobre probabilidad condicional, citando Wikipedia, con Bayes: "… sabiendo la probabilidad de tener un dolor de cabeza dado que se tiene gripe, se podría saber (si se tiene algún dato más), la probabilidad de tener gripe si se tiene un dolor de cabeza, muestra este sencillo ejemplo la alta relevancia del teorema en cuestión para la ciencia en todas sus ramas."[48]. Imaginemos que queremos clasificar "el sentimiento" de un tweet sobre una empresa en positivo o negativo, en función de las palabras que contiene, para ello de entrada disponemos de una "bolsa de palabras" de entrenamiento, donde para cada una de las palabras hemos calculado la probabilidad de que sea positiva o negativa en relación al sentimiento, con estos datos podemos determinar la probabilidad de que un determinado tweet que contenga palabras clasificadas a priori sea positivo o negativo en relación a un sentimiento o criterio. Si este modelo lo juntamos con un algoritmo de aprendizaje de máquinas podemos retroalimentar el sistema con nuevas palabras clasificadas de forma automática. Este clasificador muy fácil de implementar, es

[46] http://research.microsoft.com/pubs/209355/NOW-Book-Revised-Feb2014-online.pdf
[47] La unidad de color más pequeña representable en una imagen, una imagen digital es una matriz de pixeles cada uno de ellos con un color.
[48] http://es.wikipedia.org/wiki/Teorema_de_Bayes

usado en el análisis de sentimientos como se ha ilustrado, en el filtrado de spam, clasificación de textos, diagnóstico médico y motores de recomendación.

Análisis de grafos[49]: Con algoritmos de análisis de grafos buscamos hacer una representación de los datos y sus relaciones para facilitar la detección de relaciones y/o dependencias, hallar los influyentes nodos centrales de una red, identificar las conexiones entre nodos, identificar comunidades. Tal y como se ha comentado anteriormente, las bases de datos orientadas a grafos están especialmente diseñadas para operar con datos que presentan estructura de grafo complexa. Algunos de los algoritmos se basan en posicionar los nodos en espacios de 2 o 3 dimensiones intentando que los arcos o aristas tengan longitudes similares y se crucen lo mínimo posible en la representación, basándose en asignar "fuerzas" atractivas y repulsivas. Algunos campos de aplicación de este tipo de algoritmos son: la representación de redes de telecomunicaciones, análisis de la diseminación de rumores, identificación de patrones y relaciones en el análisis de las redes sociales, problemas logísticos, planificación de rutas, detección de fraude, diseño de chips, relaciones entre factores de decisión…

Métodos de previsión: Conjunto de técnicas que basándose en datos históricos de una serie estimamos los valores futuros de esta. La metodología de tipo cuantitativa se divide en dos tipos:

1- Paramétricos: Como medias móviles, medias simples, alisado exponencial, medianas estacionales, dobles medianas móviles, alisado exponencial de Holt, descomposición, alisado exponencial de Holt-Winters.
2- No paramétricos: Metodología Box-Jenkins. Estimamos los valores futuros de una serie mediante datos históricos de esta combinando modelos autoregresivos y/o de medias móviles, no estacionarios (ARIMA) y estacionarios (SARIMA).

Esta metodología es muy utilizada para hacer previsiones de series debido a la sencillez de algunos de los métodos y sus buenos resultados. Por ser de las más conocidas no se explicará la base de su funcionamiento. Algunos posibles usos son: predicciones de series económicas por ejemplo del PIB, previsiones sobre aprovisionamientos en la cadena de producción, previsiones de ventas, previsiones meteorológicas…

Random forests[50]: Es un algoritmo de clasificación utilizado para medir la importancia de variables cuantitativas, basado en aplicar arboles de decisión a los datos. Se basa en: Crear un subconjunto aleatorio de aproximadamente 2/3 del total de los datos. La creación de los árboles se hace con un proceso iterativo, se escoge aleatoriamente *n* variables de entre todas las predictoras; para formar el primer nodo se prueban todas las *n* variables con el

[49] https://infocus.emc.com/william_schmarzo/how-can-graph-analytics-uncover-valuable-insights-about-data/ http://en.wikipedia.org/wiki/Force-directed_graph_drawing
[50] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro y http://randomforest2013.blogspot.com.es/2013/05/randomforest-definicion-random-forests.html

subconjunto de datos, la variable que consigue una mejor división (según el criterio de salida buscado) en el nodo es la que se utiliza, en el siguiente nodo se vuelve a elegir aleatoriamente *n* variables entre todas las predictoras y se hace lo mismo. Siguiendo este método se crean diferentes árboles. Los datos de entrada se introducen en todos los arboles creados dando diferentes resultados sobre el criterio de salida para cada uno de los árboles, el resultado del algoritmo es el resultado de los diferentes árboles promediado o promedio ponderado, o en caso de variables dicotómicas un mayoría de votos de los diferentes árboles (cada árbol daría un voto en un sentido u otro según el criterio fijado). Este algoritmo puede manejar miles de variables de entrada y es de los algoritmos de aprendizaje más precisos. Es usado para filtrar variables relevantes, clasificar datos, hacer predicciones.

Algoritmos genéticos: Como se puede intuir de su nombre, están inspiradas en la base genética de la evolución biológica. Se basan en aplicar los conceptos de mutaciones y combinaciones genéticas. A una población de individuos (entrada al algoritmo), se le somete a unas acciones aleatorias que los alteran de forma y manera que en la salida del algoritmo solo salen los mejores individuos en función de un criterio de selección fijado. Es uno de los tipos de inteligencia artificial más prometedores y aplicado en campos como diseño de piezas, optimización logística, resolución de equilibrios en teoría de juegos, análisis lingüístico y lenguaje natural, previsión de comportamiento en mercados financieros, planificación de procesos, construcción de horarios y turnos…

Redes neuronales: Basadas el aprendizaje profundo, son algoritmos inspirados en la arquitectura de proceso en paralelo de los cerebros humanos emulada mediante nodos de proceso. La red consiste en un primer conjunto de nodos de entrada, a continuación de estos nodos diferentes capas de nodos intermedias y ocultas, y un último conjunto de nodos de salida. Cada nodo tiene un "peso" en el conjunto del proceso o decisión. Los datos son proporcionados a los nodos de entrada y mediante ensayo de prueba-error el algoritmo ajusta iterativamente los "pesos" de cada nodo hasta alcanzar un criterio de parada definido. Este tipo de algoritmo es conocido como caja negra y usado en aplicaciones como predicciones de mercados financieros, meteorológicas, reconocimiento facial o de patrones en imágenes, detección de fraude, clasificación de proteínas, reconocimiento de texto manuscrito, detección de explosivos…

3. Obtención de ventajas competitivas a través de analítica Big data

3.1 Ganancias de eficiencia y productos innovadores

Big Data puede traer tanto importantes ganancias de eficiencia a las empresas como nuevos productos, a priori el procesamiento de datos nos lleva a obtener información y a mejorar la toma de decisiones, esta información junto con la experiencia nos llevara a obtener nuevos conocimientos que podrán utilizarse para mejorar procesos existentes y reducir costes, implementar nuevos procesos, descubrir nuevos componentes, productos o variables relevantes en la producción, también nos puede permitir tener un conocimiento más profundo de cómo es y cómo se comporta la empresa internamente, como es el entorno en el que se mueve y cómo influye este a la empresa, conocer cómo interactúan los stakeholders, como se comportan frente los productos, cuáles son sus expectativas, cuáles son sus sugerencias, como podemos satisfacerlos más y mejor. Big Data puede convertirse en la piedra filosofal de algunas empresas, obtener algún tipo de ventaja competitiva con los datos puede posicionar a la empresa que lo haga en una posición privilegiada frente a los actores del mercado. Los descubrimientos producidos por el Big Data pueden desencadenar la transformación de la empresa, de sus productos e incluso del mercado. La creatividad, la innovación y el talento en el tratamiento de los datos así como una clara mentalidad de datos pueden ayudar a las empresas a encontrar verdaderos tesoros en sus propios datos, o con la combinación de fuentes de datos externos. Cualquier empresa puede subirse al carro del Big Data, se puede empezar por recoger grandes cantidades de datos de la producción por ejemplo, y con las herramientas necesarias iniciar un análisis exploratorio en búsqueda de información relevante, cualquier tipo de información recogida puede ser relevante, muchas veces los desprecios de datos pueden contener información valiosa pero este valor solo emerge con su análisis y las acciones que se tomen, no por los datos en sí. También se puede explotar las posibles relaciones y combinaciones con datos proveniente de las redes sociales, estas pueden inundarnos de detalles sobre las opiniones, gustos preferencias, necesidades de clientes, podemos correlacionar la opinión o sentimiento que despierta una marca con la cotización de esta en los mercados bursátiles. Recomendar productos proactivamente y en tiempo real mientras los clientes están navegando o comprando en nuestra web…

Un dato que revela el potencial del Big Data por ejemplo en el campo de las decisiones, es el que ofrece el economista de la MIT[51], Erik Brynjolfsson: las empresas que “toman decisiones dirigidas por los datos” disfrutan de un cinco a un seis por ciento de Impulso en la productividad[52]

[51] Instituto Tecnológico de Massachusetts
[52] https://www.technologyfirst.org/magazine-articles/109-august-2012/760-john-swalwell-sms-protech.html

En los siguientes puntos se presentará una "pequeña cata" que puede despertar la curiosidad o abrir la mente a nuevas posibilidades de análisis de los datos y aplicaciones de estos en las organizaciones. Algunos ejemplos de análisis y posibles ventajas ya obtenidas y en uso en algunos campos son:

Motores de recomendación: Amazon empezó a utilizar recomendaciones de libros basadas en Big Data y en sustitución de un equipo de críticos editores, el éxito fue rotundo y se patentó[53]. De la misma forma actúa Google con los anuncios recomendados AdSense[54], Twitter con sus sugerencias, Facebook con los amigos que podrías conocer, Linkedin con los contactos sugeridos. Estas recomendaciones se basan en análisis del perfil de los usuarios y su comportamiento para hacerse "una idea" de lo que quieren. La técnica utilizada es el Filtrado Colaborativo, mediante modelos de co-ocurrencia y algoritmos de *machine learning* se pueden hacer exitosas recomendaciones en tiempo real no influenciadas por los sesgos que podría tener cualquier experto de carne y hueso.

Análisis de campañas de marketing: A medida que los departamentos de marketing tienen más información sobre las ventas pueden identificar y segmentar con una mayor granularidad a los clientes objetivos y así personalizar aún más los contenidos. La puesta a disposición de datos como el flujo en las webs, los clics, las preferencias en Facebook, las codificaciones detalladas de llamadas, los metadatos de los tweets… permiten hallar nuevos conocimientos sobre los patrones de compra y comportamiento de los clientes. Los algoritmos de agrupamiento o análisis clúster son los responsables de esta segmentación y claros beneficiados de los datos masivos.

Fidelización de clientes: Muchas veces un aumento del número de productos comprados por cliente puede indicar un mayor grado de fidelización y una menor probabilidad de perder el cliente, es por esto que muchas empresas se esfuerzan para mejorar las ventas cruzadas y el *up-selling*[55], pero muchas veces el análisis de las ventas entre diferentes líneas de negocio choca contra problemas como la heterogeneidad de criterios y datos. Big Data es capaz de tratar de forma unificada estos datos que de otra manera no podrían ser analizados permitiendo hacer un análisis a gran escala e identificar patrones que se correlacionen con la pérdida de clientes o con su mayor fidelización y concentrar los esfuerzos en iniciativas eficaces.

Análisis Social: Hace unos años los científicos de datos de Facebook descubrieron que si un amigo comparte una foto, o hace clic en un "me gusta" en concreto y esto es visto por sus amigos esto anima a que los amigos de este actúen de la misma manera, esto propició por ejemplo que Facebook modificase la página de noticias dando una mayor visibilidad de las acciones de nuestros amigos, esto dio lugar a un circulo virtuoso de nuevas contribuciones.

[53] http://www.google.com/patents/US7113917?hl=es#v=onepage&q=&f=false
[54] http://es.wikipedia.org/wiki/Google_AdSense
[55] Ofrecer un producto de más valor, más completo o más enfocado a un necesidad.

Este tipo de patrones de comportamiento serían muy difíciles de hallar sin los medios que proporciona Big Data, en el análisis social se pueden buscar por ejemplo los usuarios que presentan mayor influencia sobre los demás, esto puede ayudar a las empresas a verificar si estos clientes se corresponden o no con los clientes que acaparan más volumen de compras o más variedad y que atributos diferenciales pueden tener.

Monitorización de redes: Las redes inteligentes son infraestructuras susceptibles de ser analizadas mediante Big Data, las redes de telecomunicaciones, transporte, energía, gas generan grandes cantidades de datos que se pueden beneficiar de forma interna para las empresas al ser explotados proactivamente, detectando por ejemplo posibles cuellos de botella antes de que tengan efectos para los usuarios. Pero también de forma externa a la empresa, se puede identificar patrones de consumo de electricidad, telefonía, extrapolar poder adquisitivo con consumo, obtener geoposicionamiento de usuarios, correlacionar el tráfico de las calles, carreteras y autopistas con el PIB de una región, saber qué locales de ocio están de moda en una ciudad, etc.

Análisis de mercados financieros: La búsqueda de indicadores para replicar las cotizaciones es uno de los campos exploratorios del Big Data, relacionar los sentimientos sobre una empresa o sobre sus acciones y correlacionarlo con su cotización es algo posible con la actual cantidad de datos. Mediante algoritmos de aprendizaje y clasificadores Bayesianos se puede hacer un análisis de palabras clave con el propósito de clasificar unos datos (información) dentro de unos criterios. Si además a este modelo le añadimos la posibilidad de analizar datos en tiempo real de las redes sociales como Twitter, entonces tenemos la posibilidad seguir la relación con las cotizaciones de unas acciones en tiempo real. Este tipo de análisis es utilizado por grandes bancos para conocer la opinión pública sobre su empresa, sobre un mercado o sobre la economía.

Análisis Predictivo: Las entidades financieras utilizan complejos algoritmos de correlación y probabilidad para predecir los futuros cambios en los mercados mediante datos actuales e históricos de los mercados, pero el inmenso volumen de datos hace que este análisis solo pueda resolverse mediante tecnología Big Data: Big Data permite que estos análisis se puedan hacer con suma rapidez o incluso en tiempo real y además a un coste relativamente bajo.

Análisis del riesgo: En el actual entorno cada día más globalizado, hacer una gestión proactiva y continuada del riesgo es clave para el éxito empresarial, para ello se analizan grandes cantidades de datos acerca los factores de riesgo, pero cada vez se exige que este análisis se haga con datos heterogéneos de diferentes departamentos o empresas de un holding y además que se pueda hacer sobre la marcha para poder aplicar las medidas oportunas antes que la competencia. Un caso de ejemplo analizado en una noticia del diario

expansión[56] es el de la empresa "Big Data Scoring"[57,58] que evalúa la solvencia de un cliente en función de datos de las redes sociales, datos que pueden ir des del tiempo que pasan los usuarios consultando Facebook, la titulación universitaria de nuestra red de amigos, nuestros "me gustas" o preferencias indicadas. También mencionan en la noticia que otra compañía cruza los datos entre las diferentes redes sociales buscando alertas sobre datos incoherentes, mencionan, por ejemplo, que un usuario informe en su perfil de la red social Linkedin que está graduado en Harvard pero sin embargo ninguno de los amigos de su red ha estudiado en Harvard puede ser posible pero es poco probable.

Detección de fraude: El análisis de correlaciones entre múltiples fuentes sin relación puede traer a la luz actividades fraudulentas antes que los métodos de detección tradicionales, esto puede suceder cuando relacionamos datos de compras en los TPV[59], con comportamiento en la página web del banco o en el smartphone y en las redes sociales, junto con operaciones en cajeros, transacciones con otros bancos o transferencias internacionales SWIFT. Este contexto de datos masivos mejora la detección del fraude de forma notoria.

Investigación y desarrollo: Cualquier empresa con una dotación importante de personal dedicado a investigación y desarrollo utiliza o puede estar en condiciones de utilizar el potencial de los datos masivos para ayudar a desarrollar nuevos productos o mejorar los existentes. Empresas farmacéuticas, fabricantes de software, empresas químicas, banca, empresas de biotecnología pueden reutilizar los datos históricos y presentes de sus procesos para obtener mejoras de eficiencia en muchos campos.

Análisis de Juegos: La telemetría que se genera cuando jugamos a los videojuegos es similar al análisis de datos sobre una web, la telemetría permite capturar todas las acciones de un jugador y su comportamiento ante los sucesos de un juego, estos datos permiten modificar la conducta del juego y adaptarla a la del jugador personalizando atributos como la dificultad o el comportamiento de una escena, ofrecer la compra de "vidas". El conocimiento sobre el comportamiento de los jugadores mediante la telemetría puede llegar a transformar la industria del videojuego a un nuevo escenario donde los juegos se personalizaran a medida de los usuarios.

Trading "sin escrúpulos": Recientes casos muestran que mediante análisis donde se correlaciona contabilidad y el seguimiento de posiciones de trading y el uso de sistemas de trading automatizado se puede desvelar información valiosa para especulación maliciosa que hasta ahora no se podría descubrir con las herramientas tradicionales que no permiten las fuentes de datos heterogéneas y análisis en tiempo real de patrones. Esta información

[56] http://www.presscuttingservice.com/noticias/20140407_EXPANSION_P48_hous1449.pdf o http://www.expansion.com/2014/04/07/empresas/digitech/1396898958.html
[57] http://www.bigdatascoring.com/home/about/
[58] http://www.actibva.com/magazine/productos-financieros/scoring-el-programa-que-aprueba-tu-credito
[59] Terminales Punto de Venta o Datafonos, permiten realizar pagos con tarjeta.

usada de forma ilícita puede contribuir a generar grandes pérdidas a particulares, arruinar empresas, bancos y desestabilizar países.

Otros campos: Big data también consigue ganancias de eficiencia en sectores como la sanidad, la agricultura, televisión interactiva, industria automovilística, aviación, genética y telefonía o la traducción automatizada de textos, este último es un campo que ha experimentado grandes cambios con Big Data, a finales de los 80 los investigadores de IBM desarrollaron una idea novedosa en la creación de un traductor automático: en vez de intentar introducir en un algoritmo las reglas lingüísticas explicitas junto con un diccionario, decidieron permitir que el algoritmo emplease la probabilidad estadística para calcular la palabra más adecuada en una traducción. Google fue más allá y empleando la idea de los datos masivos alimentó su sistema de traducción con todo el contenido global de internet, el resultado es actualmente usado por millones de usuarios en el mundo y mejorado por todos los usuarios mediante filtrado colaborativo. Otro ejemplo es el del fabricante de motores a reacción Rolls Royce, este transformó su negocio con el Big Data, pero no solo con el análisis de los procesos de producción, sino que además introdujo la monitorización continua de todo parque de motores a reacción que tiene hoy en día funcionando en los aviones, esto por ejemplo les permitió detectar problemas de forma preventiva antes de que se produzca una avería.

3.2 Negocios innovadores: Big Data, Big Opportunities

Big Data puede, y de hecho está siendo la precursora de gran cantidad de startups que se dedican a explotar los datos de forma innovadora o combinando datos de forma que la información resultante sea un producto innovador, con la perspectiva adecuada los datos pueden dar origen a nuevos negocios, y es que el valor de los datos no radica tanto en los datos en si como en los usos que les podamos dar, en cualquier caso, poseer los datos necesarios puede ayudar a formular nuevas aplicaciones o análisis que quizás serían más difíciles de obtener si no se poseen los datos necesarios, en este aspecto las iniciativas Open Data pueden tener un impacto positivo en el tejido empresarial, son bastantes los ejemplos de empresas que empiezan su andadura combinando datos libres o un mezcla de datos libres y licenciados, www.flyontime.us combina datos meteorológicos abiertos de USA con información abierta también, sobre los vuelos para predecir si un vuelo se demorará. www.feedzai.com utiliza datos en tiempo real y algoritmos de machine learning para prevenir fraudes y actualmente vende sus servicios a SAP, www.ayasdi.com utiliza Big Data para resolver problemas complejos como explorar nuevas fuentes de energía, ayudar en la búsqueda de curas para el cáncer, priceonomics.com proporciona una guía de precios para casi cualquier producto basándose en la recogida de datos de la World Wide Web. Pero no solo el análisis y la información de los datos son nuevas formas de negocio basadas en Big

Data, empresas como www.memsql.com proporciona tecnología de bases de datos en memoria orientadas a las consultas en tiempo real, datasift.com actúa como un mercado de datos, proporciona acceso y licenciamiento a fuentes de datos como Facebook, Twitter, WordPress, Instagram, IMDb, YouTube, Wikipedia…, www.mu-sigma.com ofrece ingeniería de datos y software como servicio para ayudar a la toma de decisiones. www.kaggle.com es una plataforma de crowdsourcing[60] donde empresas llevan a concurso de equipos expertos públicos análisis o desarrollos relacionados con Big Data, Instituciones y empresas importantes como la NASA, Facebook, Microsoft, Ford o el proyecto Atlas de búsqueda del Boson de Higgs.

Un producto innovador fruto del Big Data y que recientemente ha patentado Amazon es el concepto de la compra anticipada[61] basado en llevar las recomendaciones y las predicciones quizás demasiado lejos para los tiempos que corren, la idea de Amazon es usar los datos que recoge de los clientes referentes a compras, recomendaciones, productos deseados… para predecir lo que quieren los clientes y enviar los productos automáticamente, el concepto rompedor es que Amazon anticiparía los envíos ¡¡sin que los clientes hayan hecho la compra todavía!!

Otros aspectos que pueden originar nuevos negocios relacionados con Big Data, pueden ser por ejemplo las necesidades de formación sobre Big Data y la consultoría especializada, o el sector público demandando soluciones para las SmartCityes.

Como rasgos en común hallados en las startups basadas en Big Data econtramos: fundadores con experiencia en minería de datos o visionarios sobre aplicaciones de los datos, o docentes con relación con el mundo de la minería de datos, un claro aprovechamiento de la incipiente incursión del Big Data como innovación, el hecho de que buena parte del software necesario para Big Data sea open source además de las posibilidades de reutilización de hardware hacen que los costes de entrada no sean especialmente elevados, por ultimo hay bastantes casos la financiación de las empresas proveniente de fondos de inversión especializados en empresas tecnológicas.

3.3 Hadoop

Hadoop es un proyecto de la *Apache Software Foundation* que tiene por objetivo desarrollar un marco de trabajo de código abierto y de bajo coste para el procesado de forma distribuida de grandes cantidades de datos mediante clústeres de servidores, como se ha comentado anteriormente, Yahoo! creó Hadoop basándose en el paradigma MapReduce ideado por Google y posteriormente lo cedió a Apache. Hadoop es la unión de diferentes

[60] Definido según Wikipedia como colaboración abierta o externalización abierta de tareas. Consistente en externalizar un proyecto o tareas a una comunidad o un grupo de expertos a través de una convocatoria abierta.
[61] http://www.forbes.com/sites/onmarketing/2014/01/28/why-amazons-anticipatory-shipping-is-pure-genius/

tecnologías desarrolladas por Apache en código abierto como: Hadoop Distributed File System (HDFS) y MapReduce. Aunque Hadoop es liberado en código abierto por Apache, existen diferentes distribuciones modificadas por fabricantes de software que incluyen complementos adicionales y soporte, pero todo ello ya de pago.

HDFS es un sistema de archivos[62] que se basa en un almacenamiento de los archivos de forma distribuida, en otras palabras la información de un documento por ejemplo no está almacenada en una única unidad y replicada en otras unidades, sino que se reparte en partes que son almacenadas en diferentes unidades no necesariamente localizadas en un mismo lugar, esto permite aprovechar el almacenamiento de muchas unidades diferentes como si fuesen una sola unidad permitiendo almacenar archivos de gran tamaño. Además de que HDFS permite utilizar hardware común de bajo coste. Pero HDFS ni Hadoop son por si solas una base de datos! el sistema de base de datos que elijamos se ejecutará sobre HDFS como si fuese una "capa" superpuesta a este, aprovechando las funcionalidades que proporciona este sistema de almacenamiento, Hadoop vendría a complementar un Data Warehouse.

MapReduce es un marco de trabajo de Hadoop que permite la ejecución de tareas de forma distribuida y su procesado en paralelo siguiendo el mismo paradigma que HDFS. MapReduce es el núcleo de Hadoop, está basado en el principio "divide y vencerás", las tareas a ejecutar se dividen de forma y manera que puedan ejecutarse como subprocesos de la tarea principal en servidores diferentes, de características heterogéneas y el ubicaciones diferentes. Esto permite un ahorro de costes importante al poder reaprovechar hardware existente en las empresas. El funcionamiento de MapReduce se divide en dos subprocesos, "map" y "reduce". "map" es el subproceso que se ejecuta de forma distribuida en cada uno de los nodos, simplificando podríamos decir que "map" es el proceso que ejecuta la consulta en cada nodo, pidiendo el subconjunto que seseamos de todo el conjunto de información. La información de salida de "map" es tratada por "shuffle&sort" para ordenar y agrupar los datos según el criterio que exija la consulta. Finalmente el proceso "reduce" recibe los datos de salida de "shuffle&sort" y en un nodo único aplicará la última transformación o agrupamiento que exija la consulta. En la Ilustración 2 se puede apreciar un esquema del funcionamiento.

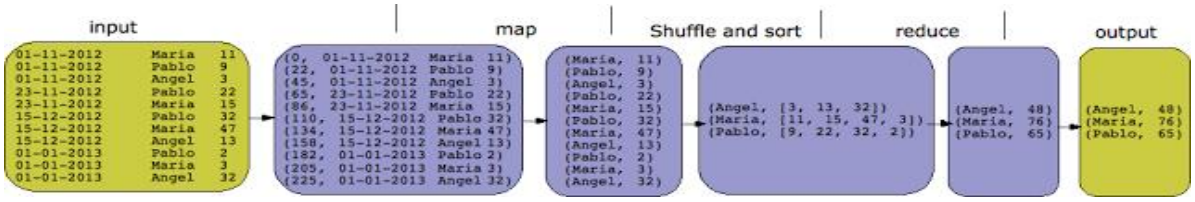

Ilustración 2 [63]

[62] Componente de un sistema operativo encargado de administrar y asignar el espacio de una unidad de almacenamiento, administrar el acceso de los datos tanto para lectura como para escritura. En los entornos domésticos son familiares sistemas como FAT y NTFS por ejemplo.

[63] http://hadoopontheroad.blogspot.com.es/2013/02/mapreduce-ejemplo-teorico.html

Hoy en día Hadoop ya se provee como software as a service (SaaS) por lo que cualquier empresa puede hacerse con una plataforma para el análisis de datos masivos sin necesidad de hacer una inversión en infraestructura, de la misma manera que cualquier outsourcing, se puede contratar bajo demanda, permitiendo flexibilizar el coste de la obtención de valor a las necesidades concretas y/o puntuales de la empresa. Algunos de los proveedores de Hadoop como SaaS son: Amazon[64], Cloudera[65] y Microsoft[66], existen también soluciones donde la empresa cliente alquila la infraestructura como PaaS (*Plataform as a service*) para que la empresa pueda disponer y administrar su plataforma Hadoop de forma más independiente.

Hoy en día son muchas la empresas se benefician de Big Data con Hadoop[67], algunas de ellas son: Amazon, Telefónica, Ebay, Google, Yahoo, Facebook, Twitter, Salesforce, The New York Times, HP, BBVA, PayPal, ING…

3.4 Estado actual de adopción en las empresas

Hasta ahora la mayoría de grandes empresas disponen de toda la cadena de explotación de datos, des de los data warehouse, las aplicaciones de data mining, las herramientas business intelligence, reporting y cuadros de mandos, en los mejores casos estas infraestructuras habían y están ayudado a entender las compras de sus clientes, los patrones de comportamiento de estos en los diferentes canales, la optimización del packaging y de los precios de venta, la mejora de la imagen de marca, en el peor de los casos las empresas habían invertido excesivamente como para poder recuperar las inversiones en almacenamiento, analítica y visualización. Muchas de estas empresas después de haber hecho importantes inversiones en sus data warehouse y en toda la cadena de explotación de datos no quieren invertir en remplazar tecnologías que funcionan bien, para solucionar las dependencias y costumbres interiorizadas. Por lo que normalmente se intenta optar por una arquitectura hibrida que convine lo mejor de las tradicionales cadenas de explotación de datos con las nuevas tecnologías de los datos masivos, ya que ambas arquitecturas pueden funcionar de forma simultánea manteniendo la carga de trabajo habitual de los sistemas: el uso de herramientas de BI, la generación por ejemplo de reports de ventas o cuadros de mando… Big Data puede compenetrarse con los almacenes de datos de la tecnología existente, esto es vital, ya que sin modificar nada las bases de datos existentes se irían nutriéndose de datos nuevos generados por la empresa, estas bases de datos heredadas podrán ser explotadas por la tecnología Big Data (obviamente con ciertas limitaciones y un menor rendimiento) junto con los datos de las nuevas bases de datos NoSQL. Como se

[64] http://aws.amazon.com/es/elasticmapreduce/
[65] http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html
[66] http://azure.microsoft.com/es-es/services/hdinsight/
[67] http://wiki.apache.org/hadoop/PoweredBy

puede apreciar, esta arquitectura logra juntar las ventajas de las arquitecturas tradicionales con las de Big Data, logrando anular el impacto negativo en la empresa motivado por un cambio de tecnología, consiguiendo además incorporar aditivamente todo el potencial de Big Data. Pero esta arquitectura no solo está pensada y es interesante para las grandes empresas, sino que también es un enfoque recomendable para pequeñas y medianas empresas que poseen ya las infraestructuras de datos clásicos y quieren empezar su andadura en Big Data, el hecho de comenzar a construir la arquitectura Big Data apoyándose por ejemplo en los existentes data warehouse permite acometer la entrada en el mundo del Big Data de una forma más manejable, con un riesgo y una inversión menor.

Sobre el modelo de despliegue de Big Data, para las grandes empresas o corporaciones con infraestructura y recursos suficientes se recomienda desplegar Big Data como una plataforma de nube privada, esto permite beneficiarse de las ventajas de conectividad y transparencia para los usuarios sin renunciar a la flexibilidad que supone poder administrar y construir la plataforma propia con total libertad y adaptabilidad. Para las pymes sucede lo contrario, una solución Big Data como SaaS en la nube evita la inversión en infraestructura, viéndose claramente beneficiada por la flexibilidad de la solución de pago por uso.

Del interesante estudio de IBM “Analytics el uso de Big Data en el mundo real” sobre la adopción del Big data en las empresas cabe destacar conclusiones interesantes: La mayoría de empresas que adoptan o planean adoptar Big Data se encuentran en las fases de desarrollo del proyecto Big Data, aproximadamente solo un tercio están explotando Big Data. De estos datos se desprende que la adopción del Big Data en comparación con las metodologías y herramientas de análisis clásico de datos, todavía es incipiente. Las empresas están en el camino de la adopción, pero todavía queda recorrido. En cuanto a los objetivos que se plantean las empresas, un 50% aproximadamente orientan Big Data al cliente, solo un 18% y 14% están orientados a la optimización de la operativa y a nuevos modelos empresariales respectivamente. Del estudio se desprende que los proyectos de Big Data actuales se nutren en gran parte de datos de transacciones, datos de registros y correos, el uso de las redes sociales, el geoposicionamiento y los sensores se están explotando pero en menor medida todavía. En cuanto a la explotación analítica o data mining una acaparadora mayoría se destinan a la generación de informes, tan solo los algoritmos de visualización de datos y el modelado predictivo se están aplicando de forma bastante habitual. También es remarcable y comprensible en el actual contexto, que entre los condicionantes revelados para la adopción de Big Data no solo se busque un ROI positivo sino que también se somete el proyecto a un riguroso escrutinio fiscal.

4. Big data para el mañana

Big Data todavía está en un estado inicial de adopción, las grandes empresas y las empresas de tipo tecnológico llevan el liderazgo en la adopción del Big Data, pero aún queda mucho recorrido para consolidar Big Data en las empresas, a día de hoy hay bastantes áreas relacionadas con el Big Data en un estado exploratorio, por lo que aún podemos esperar nuevas innovaciones tanto en el almacenamiento como en tratamiento y explotación de los datos, de la misma manera que un iceberg, del Big Data solo estamos viendo la pequeña parte que sobresale por encima del agua, la gran mayoría de datos que se poseen todavía están sin explotar, a estos se les añadirán la geolocalización, los sensores integrados, el internet de las cosas, la posibilidad de integrar una SIM en todos los dispositivos, los *wearables*, las ciudades inteligentes… todos estos elementos pueden marcar los nuevos pasos para el Big Data. Las tecnologías evolucionaran, probablemente surgirán nuevas tecnologías que vendrán a mejorar, complementar e incluso substituir a la omnipresente Hadoop. Los datos masivos están, y van a seguir transformando muchos aspectos de nuestra sociedad, obligando a cambios de valores, nuevas profesiones, nuevas formas de pensar, nuevos empleos, nuevos negocios y nuevas legislaciones. Los gobiernos tendrán que abordar retos como gestionar y explotar sus datos, regular la privacidad o el valor de los datos, también habrá que pensar en los datos como un activo intangible para las empresas que pueda reflejarse en los balances de estas.

Pero no solo eso, Big Data también impulsa un cambio muy importante en el enfoque, tradicionalmente el análisis de datos en la empresa se iniciaba en los usuarios con conocimiento de negocio, estos formulaban las preguntas que tenían que responder los datos y los departamentos de IT estructuraban los datos para poder hacer los análisis y dar respuestas a las preguntas, los análisis solían ser estructurados y repetidos periódicamente. Con el enfoque de Big Data se pasa a un enfoque de análisis exploratorio, los departamentos de IT proveen tecnología y las herramientas para que los usuarios con conocimiento de negocio exploren de forma creativa que preguntas se pueden responder.

5. Ética y responsabilidad

El auge en el uso de tecnologías Big Data está planteando muchas cuestiones acerca la privacidad, la propiedad intelectual, la libertad de expresión. El uso del Big Data puede terminar socavando la privacidad de los usuarios, Google sabe que buscamos a menudo, que buscamos ocasionalmente, que paginas visitamos, monitoriza y recoge datos sobre nuestros hábitos de navegación a internet, probablemente sepa donde vivimos sin que se lo hayamos dicho expresamente, Twitter sabe que pensamos, Amazon y Ebay que compramos, TripAdvisor a donde viajamos, Facebook sabe nuestros gustos y como nos sentimos, Linkedin donde trabajamos y que hemos estudiado, Endomondo que deportes practicamos y donde, los emisores de tarjetas de crédito y los bancos saben que compramos y cuando lo hacemos. Hoy en día compartimos nuestros gustos, preferencias, viajes, lecturas, deportes estados de ánimo… y todo ello gratuitamente y gustosamente, añadamos que muchos de los datos que hoy en día se generan incorporan información personal! Así pues con este nivel de detalle no puede ser muy difícil socavar la privacidad de las personas, no hace falta ni tener la intención, puede que alguna empresa lo acabe haciendo de forma no intencionada, mediante algún tipo de análisis en busca de otro tipo de información. También cabría preguntarse qué pasaría si nuestros datos caen en manos de organizaciones con propósitos maliciosos, podrían convertirse en un instrumento de represión para los usuarios, los organismos dedicados al espionaje de bien seguro que también están aprovechando Big Data para sus propósitos, y quizás llegando más allá de un propósito de seguridad… Como se ha detallado, algunas de las funcionalidades de Big Data son hacer pronósticos y clasificaciones, la policía basándose en Big Data y de forma preventiva, podría llegar a investigar o detener personas en base a pronósticos sobre futuros actos de estas! De momento en algunos países ya se están empezando a usar algoritmos para determinar donde y cuando patrullar.

La normativa actual tampoco va al rebufo de la tecnología, más bien la sigue des de algo lejos, con cierto retraso, en España la normativa nos permite el tratamiento de datos en caso de estar anonimizados[68] ya que no se les aplica la LOPD por no ser de carácter personal. En caso de ser datos de carácter personal obliga a recoger el consentimiento explícito de los usuarios para el tratamiento de los datos, que además no podrá ser desproporcionado ni exceder las finalidades para los que fueron obtenidos. El primer escenario puede ser vulnerable ya que puede que los datos no apunten a una persona concreta, pero cuando los datos son procesados y relacionados con otros datos pueden conducir a un individuo concreto, el segundo fácilmente también, hoy en día, ¿cuantas personas al aceptar políticas de privacidad y tratamiento de los datos se leen los términos y condiciones de privacidad? ¿y los comprenden?

[68] Consiste en eliminar los rasgos de identidad personal en los datos, por ejemplo, nombres, NIF, dirección, fecha de nacimiento

Mayer menciona algunas ideas sobre cómo mejorar la protección de la privacidad, por ejemplo mediante auditorias, no solo de datos como exige la LOPD, sino de los análisis, obligar por ley a las organizaciones a informar de los usos y análisis y propósitos de datos, facilitar una mayor transparencia en la cesión de los datos, hacer cumplir a las organizaciones unas formalidades para clarificar los usos a la los que van a ser sometidos los datos

Pero no solo esto: ¿cuál es el valor de estos datos que estamos cediendo? ¿Cuáles serán los futuros usos de los datos? ¿Se puede dar carta blanca al análisis de los datos que hemos cedido? ¿Permiso de análisis de datos al por mayor?

El grupo del artículo 29 de la Comisión Europea (WP29) ya señala que una protección de notificación y consentimiento puede ser insuficiente cuando los datos se convierten en una importante fuente de ingresos para empresas como Google, Facebook y Twitter entre otras. El WP29 considera que debería regularse el derecho de la propiedad de los datos realizando un enfoque de los datos como un tipo de patrimonio más. El WP29 también hace referencia a que a menudo, no es la información recogida en sí misma sensible, sino más bien, las inferencias que se extraen de ella y por tanto las compañías deberían informar también de la política o los criterios de decisión en el uso de datos.

La futura regulación debe proteger a los usuarios pero también debe dar margen y legitimidad hacer posibles los análisis y para fomentar el desarrollo del Big Data. Encontrar el equilibrio no se prevé fácil.

6. Conclusiones

Después de concluir este proyecto parece más que claro, que Big Data es una fuente importante de valor para las empresas, aun siendo una tendencia que empiezan a adoptar las empresas, numerosos casos reales avalan la idea de que Big Data es un percusor de nuevas innovaciones y por tanto de ventajas competitivas que no solo transforman las empresas y sus productos, sino que son capaces de crear y transformar mercados. Con la debida cautela generada por episodios como el de las ".com", la sociedad en general puede verse beneficiada y transformada a un nuevo escenario como el que trajo en su día la computación. Ante toda esta aparente generosidad del Big Data, se deberá estar muy vigilante siempre a aspectos como la privacidad y el valor patrimonial de los datos que compartimos y generamos cada instante. Cohesionar el Big Data con los objetivos estratégicos de las empresas, aplicar un mentalidad de datos y poner un énfasis en el análisis, más que en los datos son claves para la generación de valor a través del Big Data.

8. Biografía

- [V. Mayer, 2013] Viktor Mayer Schönberger, "Big data", Turner Publicaciones 2013.
- [L. Joyanes, 2014] Luis Joyanes Aguilar, "Big Data", Alfaomega-Marcombo 2014.
- [F. J. Ohlhorst, 2013] Frank J. Ohlhorst, "Big Data Analytics", Wiley & SAS Business Series 2013.
- [B. Franks, 2012] Bill Franks, "Taming the Big Data Tidal Wave", Wiley & SAS Business Series. 2012.
- [John W. Foreman, 2013] John W. Foreman, "Data Smart: Using Data Science to Transform Information into Insight", John Wiley & Sons 2013.
- [T. Dunning & E. Fideman, 2014] Ted Dunning & Ellen Frideman, "Practical Machine Learning: Innovations in Recomendation", O'Reilly Media 2014.
- [T. H. Davenport, 2014] Thomas H. Davenport, "Big Data at work", Harvard Business Press Books, 2014.
- [El IBM Institute for Business Value & Saïd Business School Oxford University] "Analytics: el uso de big data en el mundo real" 2013. Publicado en http://www-05.ibm.com/services/es/bcs/pdf/Big_Data_ES.PDF
- [R. L. Villars et al, 2011] Richard L. Villars, Carl W. Olofson, Matthew Eastwood, "Big Data White Paper", International Data Corporation (IDC) & AMD 2011. Publicado en http://sites.amd.com/us/Documents/IDC_AMD_Big_Data_Whitepaper.pdf
- [Cebr, 2012] Center for Economics and Business Research Ltd (Cebr), "Data equity, Unloking the value of big data", SAS UK 2012. Pubicado en http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf
- [Talend, 2013] "Four Key Pillars To A Big Data Management Solution" White Paper publicado en https://info.talend.com/4pillarsbigdata.html?type=productspage

8.1 Enlaces Web

- Big Data @CSAIL Massachusetts Institute of Technology: http://bigdata.csail.mit.edu
- Wikipedia: http://en.wikipedia.org

- Gartner: http://www.gartner.com
- CIO España: http://www.ciospain.es
- Information Systems Audit and Control Association: http://www.isaca.org
- BI-Latino.com: http://www.bi-spain.com
- LinkedIn (Big Data Chanel): https://www.linkedin.com/channels/big_data
- Apache Hadoop: http://hadoop.apache.org/
- Forbes: http://www.forbes.com
- Telefónica: A un clic de las TIC: http://www.aunclicdelastic.com
- SAS: http://www.sas.com

9. Anejos

9.1 Cronograma

Primera versión:

	Febrero	Marzo	Abril	Mayo	Junio	Julio
Elección tema						
PPE y Índice TFG						
Borrador TFG						
Desarrollo						
Entrevista caso de aplicación						
Entrega						
Presentación y defensa						

*Duración de las fases orientativa sujeta a posteriores ajustes durante la elaboración

Revisión de la planificación a 18/05/2014:

	Febrero	Marzo	Abril	Mayo	Junio	Julio
Elección tema						
PPE y Índice TFG						
Borrador TFG						
Desarrollo						
Entrevista caso de aplicación						
Entrega						
Presentación y defensa						

*Duración de las fases orientativa sujeta a posteriores ajustes durante la elaboración

Definitivo:

	Febrero	Marzo	Abril	Mayo	Junio	Julio
Elección tema						
PPE y Índice TFG						
Borrador TFG						
Desarrollo						
Entrega						
Presentación y defensa						