Pablo
Carreño
Nin de Cardona

Big Data. Análisis de grandes volúmenes de datos en organizaciones.

COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Big Data. Análisis de grandes volúmenes de datos en organizaciones.

Autor: Pablo Carreño Nin de Cardona
Director: María Jesús Giménez Abad

MADRID | abril de 2019

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Resumen y Abstract

Resumen

En este trabajo explicamos como obtienen valor las grandes organizaciones del Big Data. En primer lugar, introducimos el termino de Big Data con la definición más aceptada a nivel doctrinal. Tras esto, presentamos los tipos de datos existentes y las diferentes fuentes de datos existentes.

Tras esto, explicamos el concepto de las bases de datos, su función y los tipos de bases de datos con las que cuentan las empresas para procesar la información.

Tras esto presentamos el concepto de analítica de Big Data a la vez que mostramos las diferentes herramientas utilizadas por las organizaciones para analizar los grandes volúmenes de datos.

Finalmente comparamos las diferencias entre el Análisis de Big Data y el Business Intelligence, para poder analizar dos casos de éxito en negocios que han implementado el Big Data y extraer conclusiones acerca de si este permite obtener valor.

Palabras clave: Big Data, análisis de datos, bases de datos, Inteligencia de negocio y ventaja competitiva.

Abstract

In this essay we explain how the companies obtain value from the Big Data. Firstly, we define Big Data following the most accepted definition by experts. Next, we define the different types of data and the different sources of existing data.

Secondly, we explain the concept of database, the different types of existing databases and how the companies process the information.

Thirdly, we define Big Data Analytics while introducing the different resources used by companies to analyze large volumes of data.

Finally, we compare the difference between Big Data Analytics and Business Intelligence, in order to explain two business cases where the implementation of Big Data has been succesful.

Key words: Big Data, data analytics, databases, business intelligence and competitive advantage.

1. INTRODUCCIÓN

1.1. Objetivos

Nos encontramos ante un escenario en el que en cada instante se producen una gran cantidad de datos. No obstante, como es lógico estos datos sin ningún tratamiento ni análisis no tienen ninguna utilidad para las empresas. Por tanto, aquellas empresas que almacenen datos, no de manera indiferente, sino almacenando datos útiles y, que, además, tengan capacidad para analizar estos datos tomarán mejores decisiones que aquellas que no lo hagan, lo que repercutirá en un aumento de rentabilidad, ahorro de costes o incluso de tiempo.

El objetivo de este trabajo consiste en entender todo el proceso que hay desde la generación de los datos hasta la obtención de utilidad con los datos generados.

Podemos establecer como objetivo general: Entender como obtienen valor del Big Data las empresas.

Y como objetivos específicos:

- Conocer con precisión el concepto de Big Data. ¿Qué es Big Data? Así como los tipos de datos existentes.
- Conocer las fuentes de datos existentes.
- Identificar los métodos de recolección de datos existentes.
- Estudiar cómo se analizan esos datos para obtener valor.

1.2. Contextualización

Pese a que no definiremos detalladamente Big Data hasta el siguiente apartado, para entender la importancia de este, es suficiente con que por ahora entendamos Big Data como un concepto equivalente a datos.

Para entender la gran importancia del Big Data, hay que atender principalmente a dos grandes cuestiones. La primera de ellas es la gran cantidad de datos que se producen en el mundo, ya el 90% de los datos de los que disponemos hoy en día han sido creados en los últimos dos años y la tendencia actual es duplicar la velocidad de producción de datos cada dos años, (BSA, 2015) y la segunda cuestión es su incidencia en la llamada industria 4.0 o revolución industrial.

Se estima que para 2020 se habrán generado 40 zettabytes (EY, 2014). No obstante, estamos ante una cifra que seguirá aumentando de manera exponencial cada día. Para entender de la magnitud debemos detenernos en observar dos cuestiones:

- "Se han creado 5 exabytes desde el nacimiento de la civilización hasta 2003. Hoy en día esa información se crea cada dos días aproximadamente y este crecimiento no deja de aumentar" (Eric Smichdt, 2002, citado por Carlos Maté Jiménez en 2014)
- A fin de comprender la magnitud de semejante afirmación, hemos tomado como referencia una tabla realizada por el profesor de la Universidad Pontificia Comillas (ICAI-ICADE), Carlos Maté Jiménez en un artículo de revista (2014).

Tabla 1: Unidades básicas de información.

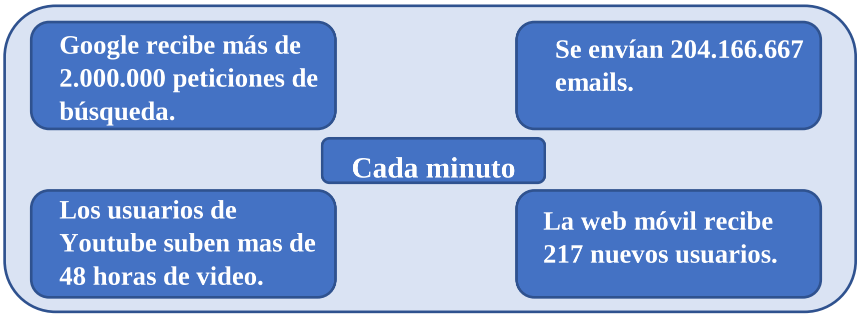
Nombre	Símbolo	Sistema Internacional
Byte	B	10^0
Kilobyte	KB	10^3
Megabyte	MB	10^4
Gigabyte	GB	10^9
Terabyte	TB	10^12
Petabyte	PB	10^15
Exabyte	EB	10^18
Zettabyte	ZB	10^23

Fuente: Adaptado de Carlos Maté Jiménez (2014)

De esta forma y apoyándonos en la tabla, podemos ver cómo aumenta la velocidad de creación de datos: 5 Exabytes hasta 2003 (Eric Smichdt, 2002), 6 Zettabytes hasta 2014 (EY, 2014) y alrededor de 40 Zettabytes en 2020 (EY, 2014)

Con esto, podemos entender la importancia que va a tener y está teniendo el Big Data actualmente, y que no es algo aislado, sino que va a seguir creciendo con el paso del tiempo gracias a la generación de datos y mayor capacidad para gestionarlos.

Figura 1: Datos generados cada minuto en internet.



Fuente: Adaptado de Domo (2017)

Hasta este momento, nos hemos centrado en un aspecto cuantitativo del Big Data como es la continua generación de datos y sus perspectivas a futuro.

A partir de aquí, y con ánimo de concluir con la contextualización vamos a analizar el impacto del Big Data en la revolución Industrial ante la que nos encontramos. De esta forma, podremos entender la importancia de este fenómeno y pasar a analizar qué papel van a jugar las empresas en relación con este.

La evolución histórica de la industria puede dividirse en 4 fases (Salvador Naya, 2018). La primera fase comienza en 1784, con la máquina de vapor, tras esto, la segunda fase la encontramos en 1870 con la producción en masa y la electricidad. En tercer lugar, en 1969 la electrónica da lugar a una nueva revolución. Finalmente, en la actualidad nos encontramos inmersos en una revolución industrial

Tabla 2: De la industria 1.0 a la industria 4.0.

1º Revolución Industrial	**2ª Revolución Industrial**	**3ª Revolución Industrial**	**4ª Revolución Industrial**
1784	1870	1969	Actualidad
Acontecimiento	**Acontecimiento**	**Acontecimiento**	**Acontecimiento**
Máquina de vapor y producción mecánica.	División del trabajo, electricidad y producción en masa.	Electrónica, IY y producción automatizada.	Nanotecnología, Internet de las cosas, Inteligencia Artificial y Big Data.

Fuente: Adaptado de Mckinsey Global Institute

Son 4 los pilares básicos de la industria 4.0 (Salvador Naya, 2018), y estos son desarrollados gracias a distintas herramientas como puede verse en la tabla.

Una de estas herramientas es el Big Data, en palabras de Salvador Naya, "Big Data, se refiere a la capacidad de recopilar, almacenar y analizar grandes cantidades de datos a través de sensores. La sensorización genera un volumen masivo de datos que será preciso analizar con detalle para su explotación futura" (Naya, 2018, p.6).

En el informe "Generación de talento Big Data en España" realizado por Cotec (2017, p.60) podemos ver que los sectores donde se prevé un mayor impacto del Big Data son: Telecomunicaciones (18%), finanzas (16%), sector público (15%), gran consumo (11%) y salud (10%).

Por tanto, podemos ver que el Big Data ha llegado para quedarse, y que como veremos en el siguiente apartado otorgará una ventaja competitiva a aquellas empresas que sean capaces de integrarlos en su organización.

1.3. Justificación

Por tanto, es necesario analizar en este punto si es rentable la inversión en Big Data, recurriendo a datos objetivos y opiniones de expertos acerca de este tema, ya que como hemos visto las empresas solo van a realizar una inversión en lo relacionado con Big Data si van a obtener rentabilidad de esta.

En primer lugar, podemos ver como BSA, The Software Alliance (3, 2015) afirma:

> *Hoy, el 90 % de los líderes de negocios citan a los datos como uno de los recursos clave y un factor distintivo fundamental para los negocios, a la par de recursos básicos como las tierras, la mano de obra y el capital.*

Por otro lado, López Zafra (Chief Data Officer) y Queralt (Chief Data Officer), profesores de CUNEF (2017) afirman que el negocio del Big Data ha alcanzado la cifra de 5.500 millones de euros, lo cual triplica los 2.000 millones de euros generados en 2013 (Citados en Valencia Plaza). Por tanto, consideran que las empresas deberán basar sus decisiones en el análisis de estos datos. En esta misma línea, se pronuncia Carmen Artigas (Citado por Cotec en 2017), cofundadora de Synergic Partners (Empresa pionera en Big Data):

> *El mercado de Big Data crece un 30% cada año en España, siete veces más que la inversión en tecnologías de la información tradicionales. Es una apuesta sólida para aumentar las fuentes de ingresos de las compañías, consolidar sus estrategias de personalización y fidelización de los clientes e impulsar su transformación digital.*

Siguiendo con los docentes, “los datos que almacenan las empresas en internet son su principal fuente de valor, no obstante, solo el 1% está procesado, por lo que podemos ver a simple vista que queda un nicho de un 99% por analizar” (López Zafra y Queralt, 2017)

Concluye el profesor López Zafra, afirmando que la utilización del Big Data incrementa la rentabilidad de dos de cada tres empresas europeas en un 15%, lo que justifica la creación de divisiones especializadas en este sector. (López Zafra, 2017)

El análisis basado en grandes volúmenes de datos optimiza la calidad de la producción, produce ahorro de energía y mejora los equipamientos de servicio. Por tanto, en la industria 4.0, la recopilación y evaluación de los datos procedentes de distintas fuentes (sensores, clientes…) serán básicas para apoyar la toma de decisiones en tiempo real. (BCG, 2015)

En términos de rentabilidad, se afirma que el aprovechamiento es transformador de industrias. Apoyándose en Erik Brynjolfsson director del MIT, "las empresas que adoptan decisiones basadas en datos logran entre un 5% y un 6% más de productividad y rentabilidad que aquellas homólogas o de su sector que no lo hacen" (2017).

Como hemos podido ver, aquellas empresas que integren el Big Data como parte de su toma de decisiones aumentarán su rentabilidad. El problema en este punto, que la única forma de beneficiarse de esta información es saber utilizarla, es decir, que influya de manera determinante en la toma de decisiones. De esta forma, se necesita que las organizaciones tengan procesos eficientes para convertir los grandes volúmenes de datos en ideas significativas. (Gandomi y Haider, 2014)

Por tanto, no cabe ninguna duda acerca de la rentabilidad del Big Data.

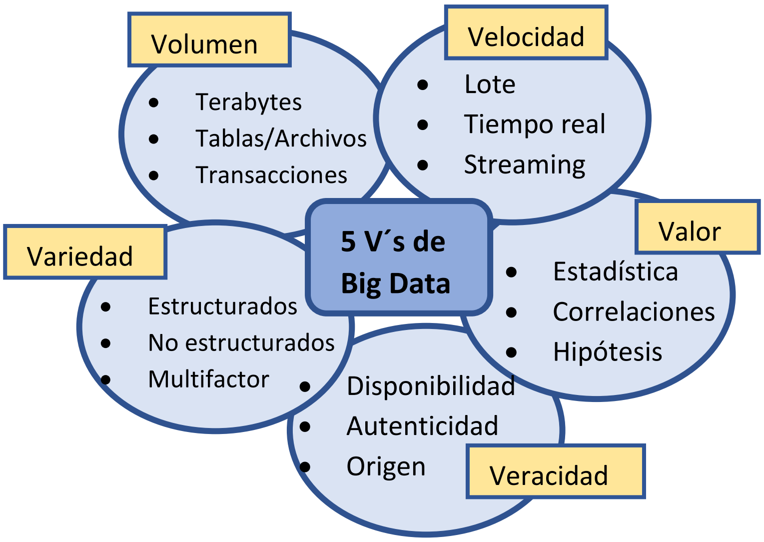
2. BIG DATA:

2.1. ¿Qué es el Big Data?:

Cuando hablamos de Big Data (Fundéu de BBVA opta por cualquiera de las siguientes expresiones: grandes datos, grandes volúmenes de datos o macrodatos) tenemos que partir de la idea de que no existe una definición única o aceptada plenamente, sino que, existen diferentes puntos de vista acerca de este concepto. Por tanto, en este TFG vamos a tratar de acercarnos al concepto de Big Data partiendo de las dos definiciones propuestas por la consultora Gartner, por ser aquellas más mediáticas y relevantes. La primera definición es la siguiente "Big Data excede el alcance de los entornos de hardware de uso común y herramientas de software para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios" (Adrian Merv, 2011) Un año más adelante se publicó la segunda definición que utilizaremos en este trabajo: "Big Data es un gran volumen de datos o información generada a gran velocidad y de gran variedad que exige formas de procesamiento de información innovadoras y rentables que permiten crear valor, tomar decisiones y automatizar procesos" (Gartner, 2012: Página web) Es importante destacar que ambas definiciones se centran más en las posibilidades que ofrecen los datos si son bien procesados y analizados, que en su dimensión cuantitativa.

No obstante, este término ha sido actualizado pasando de 3v´s (Velocidad, variedad y volumen) a 5v´s (Veracidad y valor). (Ishwarappa y Anuradha, 2015)

Figura 2: 5V´s de Big Data



Fuente: Adaptado de Ishwarappa y Anuradha (2015)

Ishwarappa y Anuradha (2015) analizan más detalladamente el significado de cada elemento que compone la definición de Big Data:

- **Volumen:** es el primer elemento en el que se piensa cuando hablamos de Big Data, debido a que se refiere al volumen como cantidad de datos. Temporalmente hablando, es el mayor desafío para las estructuras tecnológicas existentes. La gran mayoría de grandes empresas tienen una gran cantidad de datos almacenados de diferentes formas, pero no cuentan con la capacidad necesaria para procesarlos. Por ello, el Big Data Analytics (BAD) o análisis de Big Data, es tan importante, ya que proporciona la habilidad de procesar estos grandes volúmenes de datos. El volumen de datos producidos cada día no hace más que aumentar. Entre Twitter y Facebook generan 20 Terabytes al día, por tanto, no cabe duda del desafío que tienen las organizaciones frente a la gestión y análisis de esta ingente cantidad de volúmenes de datos existentes.
- **Velocidad:** la velocidad a la que los datos son creados y se generan a nuestro alrededor, así como, la velocidad en que las bases de datos de las empresas son capaces de almacenar, procesar y analizar a una velocidad similar a la de creación de los datos. Para que entendamos el fenómeno ante el que nos

encontramos, utilizaremos la empresa Wal-Mart´s. En 1999 la multinacional tenía un almacén de datos con 1.000 terabytes almacenados, construyendo en 2012 un almacén de más de 2,5 petabytes capaz de procesar alrededor de un millón de transacciones cada hora. Si acudimos a la contextualización del trabajo podremos entender con mayor exactitud la velocidad de creación de datos. Lo importante de la velocidad reside en los aspectos anteriores, ya que igual de importante que es la cantidad de datos que se generan cada minuto, es necesario que las empresas tengan capacidad para analizar esos datos en tiempo real. Para explicar esto utilizaremos el ejemplo del robo de un banco, en el cual los sistemas informáticos tienen que analizar si está existiendo un fraude o no, por lo que todo lo que sea tardar más de 5 minutos causaría un perjuicio gravísimo para el banco. En este escenario, es muy importante destacar el internet como la palanca para este crecimiento exponencial, ya que aquellas empresas punteras han estado destinando una gran cantidad de fondos para poder seguir el rastro digital de los internautas y en muchos casos consumidores, por la red. Este proceso ha dado lugar a dos grandes avances, en primer lugar, la existencia de tecnologías que permiten almacenar y analizar datos de manera casi ilimitada, así como, procesar datos indiferentemente de su tipo (texto, video, audio…), y, en segundo lugar, la evolución del análisis de los datos. (Cotec, 2017) De esta forma, vemos un gran cambio respecto a los límites que teníamos en el siglo anterior, lo que ha hecho posible la implantación de la inteligencia artificial en prácticamente todas las industrias

- **Variedad:** no todos los datos son iguales. Cuando se analizan datos es importante conocer a que categoría de datos pertenece, ya que, no se pueden procesar igual. Tenemos que distinguir entre los datos estructurados y los datos no estructurados, diferencia que explicaremos más detalladamente en un apartado del trabajo. El 90% de los datos existentes son no estructurados. Trabajar con ambos datos mezclados sin conocer de qué tipo son generan una gran complejidad a la hora de almacenar y analizar los datos. Por tanto, una aproximación muy sencilla a la importancia de la variedad es que esta desplaza los análisis tradicionales y exige la inclusión de datos semiestructurados y no estructurados, por lo que el éxito de la empresa dependerá en gran medida de su capacidad para integrar los diferentes tipos de datos existentes.

- **Veracidad:** cuando trabajamos con una gran cantidad de datos, generados y modificados a gran velocidad y que pueden ser estructurados y no estructurados, no es posible que no existan datos inútiles o falsos. La variedad de los datos captados cambia según la fiabilidad de la fuente, además, también repercute en la precisión del análisis y de las conclusiones extraídas. En otras palabras, si una empresa desea saber si lanzar un producto al mercado o no, debe tener claro que las respuestas obtenidas por los usuarios son fiables, es decir, provienen de fuentes autorizadas, ya que, si no, tomará una decisión en base a unos datos erróneos y por tanto será una decisión poco acertada, algo que se manifiesta por IBM cuando en la propia definición de Big Data comenta que 1 de cada 3 directivos no otorga fiabilidad a la información a la hora de tomar decisiones. Por tanto, este es uno de los principales retos del Big Data.
- **Valor:** es la característica más importante de los datos. Como hemos visto anteriormente, el potencial de los datos es espectacular, pero de nada sirve tener acceso a una gran cantidad de datos si somos incapaces de convertirlos en algo con valor. Es decir, la información no sirve de nada a las empresas si esta no les otorga una fuente de valor, por tanto, para que las empresas realicen la inversión en almacenes de datos y sistemas de procesamiento y análisis debe existir un retorno claro de esta inversión. Como consecuencia de esto nace Hadoop cuya función es el análisis de grandes volúmenes de datos.

2.2. ¿Qué tipos de datos existen?

Una vez hemos fijado con mayor precisión el concepto de Big Data, vamos a proceder a analizar los tipos de datos existentes, además de aclarar la diferencia entre lo que es Big Data y lo que son datos desde el punto de vista tradicional.

Cuando las empresas deciden llevar a cabo un proyecto de Big Data deben dar solución a una serie de cuestiones tales como: el origen de los datos, el volumen de información necesario para tomar una decisión, la información que aporta cada dato a mi negocio… Por tanto, es importante que la empresa reconozca las fuentes de datos existentes y el tratamiento que necesita cada dato.

Hemos de destacar que existen dos clasificaciones principales: Estructurados, no estructurados e híbridos y por origen. (Rayo, M., 2016)

Primera clasificación:

- **Estructurados:** componen esta categoría los datos que provienen de fuentes tradicionales. Son aquellos que poseen formato fijo, especificado detalladamente y cuya producción sigue un orden establecido. (Un ejemplo de formato típico sería: DNI o pasaporte: 8 números seguidos de una letra, la fecha de nacimiento (DD, MM, AA) o incluso la cuenta bancaria como serie de 20 números) (Joyanes, L., 2013)
    - **Creados:** se generan a través de los sistemas de maneras definidas previamente.
    - **Provocados:** necesitan de una acción simultanea que dé lugar a su creación, ya que no se generan automáticamente. Por ejemplo: la puntuación de una película en una web creada para valorar películas.
    - **Dirigido por transacciones:** mientras que los anteriores los crea una acción humana de manera directa, estos son un tipo de dato subyacente resultante de una acción previa. Son por ejemplo el ticket de compra cuando un usuario adquiere algo.
    - **Compilados:** agrupaciones de datos, es decir, son datos que de manera individual podrían ser de poca utilidad pero que la agrupación de todos ellos resulta realmente útil. Por ejemplo: el censo, el registro de recién nacidos…
    - **Experimentales:** son aquellos que las empresas generan para poder analizar si su idea de negocio es buena o no.
- **No estructurados:** a diferencia de las otras dos categorías, no contienen tipos predefinidos. Su almacenamiento se da sin estructura uniforme y no existe capacidad para controlar estos datos. Los ejemplos más claros son los videos, audios, fotos o datos de texto (SMS, WhatsApp, Correos electrónicos…) Estos datos suponen el 80% de los datos que poseen las empresas, siendo con diferencia aquellos que presentan una mayor dificultad en su análisis, por tanto, han dado lugar al nacimiento de herramientas como MapReduce, Hadoop o bases NoSQL que analizaremos más adelante.
    - **Capturados:** se crean partiendo de la conducta de un ser humano, es decir, si por ejemplo me pongo una pulsera que monitorice mi ritmo cardiaco, horas de sueño, ejercicio y kcal consumidas, estos datos serán capturados por la pulsera. Otro ejemplo puede ser los datos recogidos por el GPS de móvil acerca de nuestra posición.

- **Generados:** a diferencia de los provocados, no se generan en una plataforma destinada a ello, sino que son comentarios en Instagram, videos que he visualizado en YouTube…

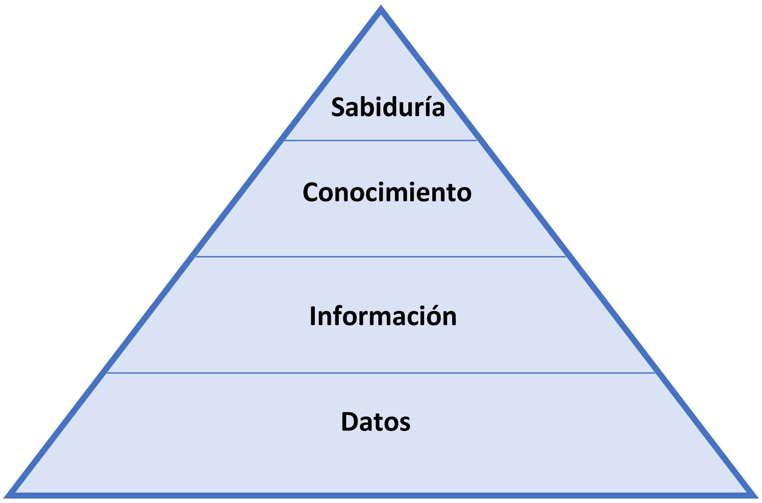
- **Híbridos o mixtos:** no siguen un patrón claramente comprensible (como si hacen los datos estructurados), a pesar de que, presentan un flujo claro y un formato definible. No existen formatos fijos como en los estructurados, pero si marcadores para separar los datos. En esta categoría destacamos registros Web logs de procedentes de conexiones a internet. Otro ejemplo son el texto HTML y XML. (Joyanes, L., 2013)

Segunda clasificación: Datos por origen. Se extiende una clasificación de 5 categorías, que pese a no ser las únicas son las más utilizadas.

- **RRSS y Páginas Web:**
    - Búsquedas en motores de búsqueda como Google o Bing.
    - Comentarios en Instagram, posts en Facebook o publicaciones en Instagram.
    - Información relativa a navegación del usuario.
    - Contenido de páginas web.
- **Maquinas interconectadas:**
    - Lectura RFID
    - GPS
    - Sensores colocados para ese uso como cámaras de tráfico, maquinas que controlan la gente que entra a Zara…
- **Transacciones:**
    - Llamadas móviles o mensajes de voz.
    - Registro de compra como la utilización de tarjeta, pago con el móvil, a través de la página web…
- **Biométricos:**
    - Sistemas de reconocimiento facial como el de Apple o Samsung.
    - ADN
- **Generados por personas:**
    - Las llamadas de atención al cliente que son grabadas, utilización del correo electrónico o la centralización de los datos médicos del paciente electrónicamente.

El proceso de obtención de la información a través de los datos se puede plasmar en la pirámide del conocimiento o DIKW, que relaciona cuatro componentes: La sabiduría, el conocimiento, la información y los datos. Es decir, esta pirámide muestra el proceso desde lo más pequeño, que es el dato a lo más grande que es la obtención de valor, de manera que para poder llegar a extraer valor de los datos primero hay que saber que buscas y por tanto que tipo de dato necesitas analizar, de ahí la importancia de la clasificación anterior.

Figura 3: Pirámide DIKW



Fuente: Adaptado de Rayo, M.

3. FUENTES DE DATOS:

3.1. Origen de las fuentes de datos

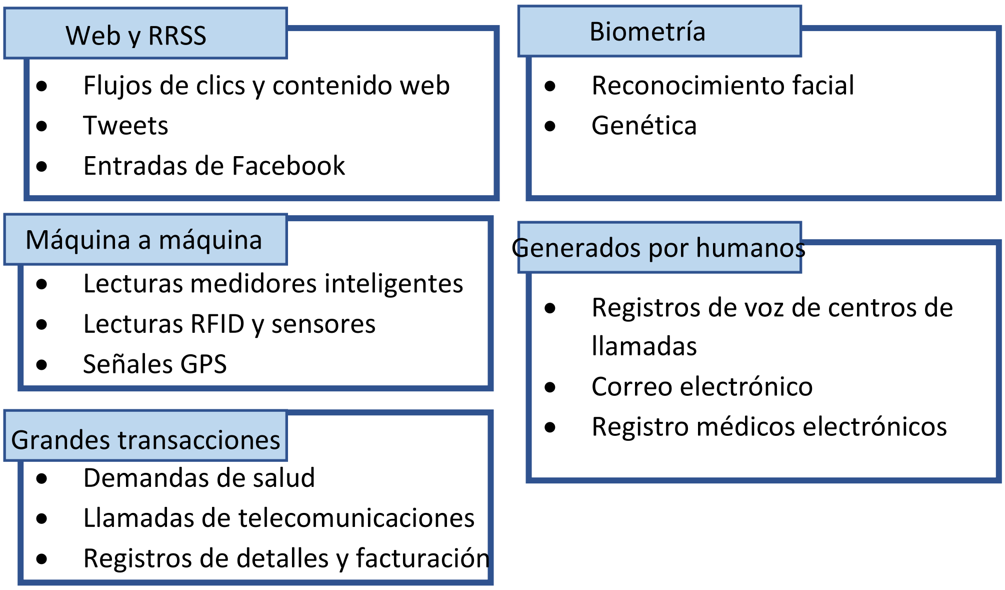
Hemos de tener claro que los datos no son algo de nueva creación, es decir, es algo que ha existido siempre y que las empresas han estado almacenando. Para que nos hagamos a la idea, en 2009 (Mckinsey, 2011) y apoyándonos en fuentes oficiales de Estados Unidos, podemos ver como ya una gran cantidad de compañías tenían almacenados una gran cantidad de datos.

Desglosando por sectores encontramos un almacenamiento de casi 1.000 petabytes en fabricación, 619 petabytes en banca, 850 petabytes gubernamentales y 715 en comunicación, medios y tecnología.

Con esto queremos dejar claro que en este apartado vamos a comentar fuentes de datos nuevas, pero que ya existían datos anteriores, conocidos como datos tradicionales, y que han sido y son a día de hoy de gran importancia para la empresa, además de ser responsables del surgimiento de las nuevas fuentes de datos. Una curiosidad respecto de estas fuentes de datos (tradicionales) es que muchas de ellas están empezando a resultar de utilidad ahora, pese a llevar años y años generándose, debido a las nuevas capacidades analíticas existentes.

3.2. Tipos de fuentes:

Figura 4: Tipos de Big Data



Fuente: Adaptado de Soares (2012)

Hemos escogido la clasificación realizada por Soares en 2012 por ser una de las más utilizadas en los artículos académicos. Como vemos esta clasificación primero realiza una clasificación en función del origen de los datos y tras esto realiza una precisión mayor de las fuentes de donde surgen. Por tanto, vamos a analizar detalladamente cada categoría para comprender con mayor precisión a que nos referimos exactamente con las fuentes de Big Data.

Web y Redes Sociales: en esta categoría incluimos contenido procedente de la web y de las redes sociales. En la categoría de redes sociales incluimos sitios como Facebook, Twitter, periódicos digitales (Marca, El Mundo…), marcadores sociales, enciclopedias virtuales como Wikipedia e incluso páginas de televisión. Los datos de este tipo son capturados, almacenados y distribuidos basándose en los flujos de clics, tweets, retweets, publicaciones, sistemas de gestión de contenidos (YouTube) o almacenamiento online como Dropbox. Los datos de la Web son imprescindibles para las empresas a la hora de tomar decisiones (Joyanes, L., 2013) Estos datos permiten obtener un conocimiento mucho más profundo de los usuarios, para ello nos podemos apoyar en la operativa desarrollada por Amazon consistente en realizar recomendaciones a sus usuarios basándose en compras anteriores, los dispositivos desde los que realiza la compra, los artículos en los que se muestra interesado, aunque finalmente no los compre, los videos que visualiza, búsquedas, lecturas de artículos e informes acerca del producto… De esta forma podemos ver que gracias a los datos web podemos ofrecer información muy personalizada que solo podríamos obtener realizando estudios personales, ya que se captura toda actuación de los usuarios en la web. Estos datos son de gran relevancia para las empresas, ya que, la analítica web proporciona los conocidos como Key Performance Indicators (KPI) sumamente útiles para la toma de decisiones.

El análisis de los datos de las redes sociales también otorga una gran ventaja a las empresas pues les permite cuantificar la influencia de un usuario en la red social. Es decir, un usuario que tenga mucho impacto en las redes sociales (influencer) puede generar una rentabilidad global bastante más grande que la que generaría como usuario individual debido a su influencia en el colectivo, lo que se materializa en el interés de las marcas por detectar y fidelizar a ese tipo de usuarios.

Maquina a Maquina: también conocido como internet de las cosas. Consiste en la conexión de varios dispositivos a través de redes inalámbricas o cableadas y teléfonos móviles a aplicaciones desarrolladas para recibir y gestionar esa información. Para ello se utilizan medidores, chips NFC[1] y etiquetas RFID[2] que recogen datos como la velocidad, la humedad o aquel cuya finalidad deba recoger y son traducidos en información útil por las aplicaciones desarrolladas.

Esto ha dado lugar al nacimiento del internet de las cosas, puesto que son objetos comunicados entre ellos que comparten información de la que se extraen conclusiones relevantes.

[1] NFC significa Near Field Communication. (Comunicación de campos cercanos) Es una estructura inalámbrica cuya finalidad es el intercambio de datos en distancias cortas. Esta tecnología sirve para conectar aquellos dispositivos móviles que poseen esta característica e intercambiar de manera veloz pequeñas cantidades de datos. No está destinada a un intercambio masivo, sino, a un intercambio veloz.

[2] RFID significa Radio Frecuency Identification. (Identificación por radiofrecuencia) Las etiquetas se colocan en objetos, de manera que cuando se leen con un lector estas retornan la información para la que son configuradas. Un ejemplo de aplicación son los pagos.

Grandes Transacciones de Datos: aúnan todo tipo de transacciones. Como vemos en la figura anterior pueden ser registros de llamadas, de telecomunicaciones o de facturación. Dan lugar a datos semiestructurados y no estructurados y permiten extraer información útil para las empresas. Por ejemplo, época del año en la que hay más bajas de una compañía telefónica a través de las llamadas o época de mayores compras a través de los datos generados por un terminal de venta.

Biometría: gracias a los avances tecnológicos este tipo de datos se han visto fuertemente incrementados. Estos datos son aquellos que permiten identificar a un ser humano basándose en sus características físicas. Los datos anatómicos son las huellas, el reconocimiento de iris y facial, ADN… es decir, los que son creados en base al aspecto físico. Por otro lado, la escritura y el registro de pulsaciones componen los datos de comportamiento. Pero los datos biométricos son generalmente clasificados en dos grupos, los procedentes de la genética y los que se encuadran en el reconocimiento facial. Un ejemplo del avance de la tecnología es la posibilidad que ofrecen hoy en día muchas marcas de teléfonos móviles de desbloquear el propio dispositivo mediante el reconocimiento facial o de la huella, algo difícil de imaginar años atrás.

Datos generados por humanos: el principal problema que presenta los datos originados por los seres humanos es el de la privacidad. Los datos pueden ser personales por lo que hacen necesaria la creación de una legislación que proteja el derecho de las personas a que estos datos sean privados, es decir, estén cifrados. Dentro de esta categoría encontramos correos electrónicos, notas de voz e incluso registros médicos electrónicos. Para el análisis de los correos electrónicos, SMS, e incluso publicaciones de texto en redes sociales se utilizan las aplicaciones de análisis de textos cuya función es la búsqueda de patrones dentro del texto para facilitar la toma de decisiones.

4. BASES DE DATOS:

Una vez hemos analizado los datos existentes y las diferentes fuentes de datos, es importante conocer las bases de datos existentes, ya que son el siguiente paso dentro del proceso de obtención de valor a partir de los datos por las empresas.

Tabla 3: Arquitectura de Big Data

Datos	Adquisición	Organización	Análisis	Decisión
Estructurados No Estructurados Semi-Estructurados Procedentes de las fuentes vistas anteriormente.	Se estudian aquellos datos procedentes de las diferentes bases de datos existentes. Ej: NoSQL	Con ayuda de las herramientas existentes se limpian, extraen, filtran, cargan y organizan los datos existentes.	Los grandes volúmenes de datos son analizados y se presentan en mapas conceptuales.	A partir de los datos extraídos se toman decisiones.

Fuente: Adaptado de Oracle (2016)

Como existen una gran variedad de bases de datos vamos a centrarnos en aquellas que son las más utilizadas en el día a día de aquellas empresas que han integrado el Big Data. Por tanto, debemos distinguir cuatro grandes categorías de bases de datos. (Joyanes, 2013)

4.1. Bases de datos relacionales

Es el modelo que ha permanecido en las décadas anteriores y que para procesos tradicionales sigue siendo de gran utilidad, además a diferencia del resto de modelos es fácilmente comprensible.

Las bases de datos relacionales, como bien indica su nombre, cumplen con el modelo relacional (Suarez, E. 2008). Estas bases de datos funcionan mediante tablas y esquemas. Por cada tabla, como es lógico, tenemos filas y columnas. Las columnas son el elemento o dato que se almacena y las filas componen los conjuntos de datos. De esta forma, tomando por ejemplo los datos de una persona, tendríamos una serie de columnas que responderían a DNI, edad, lugar de nacimiento, nombre… y las filas serían los datos concretos de cada columna. El software destinado a tratar con estas tablas es el Sistema de Gestión de Bases de Datos Relacionales (SGBDR). En este apartado vamos a dedicar unas líneas a las bases de datos transaccionales, que son aquellas que tienen poca utilización en la práctica, puesto que su utilidad es el envío de datos a grandes velocidades. Un ejemplo es una transferencia bancaria. El motivo de introducir estas bases de datos reside en que son muy utilizadas para conectar las bases de datos relacionales con Hadoop[3].

Las bases de datos relacionales no son las más eficientes a la hora de tratar con grandes volúmenes de datos, por lo que el panorama actual, caracterizado por millones de sensores, etiquetas y la naturaleza distribuida de los datos exige de nuevas bases capaces de hacer frente a esta situación. Hay versiones avanzadas de este tipo de bases como son las Massive Parallel Processing (MPP) que explicaremos posteriormente.

4.2. Bases de datos heredadas

Las bases de datos heredadas también se conocen como *Legacy*. Estas bases de datos son gestionadas por sistemas que no dependen de bases de datos relacionales. Son bases que, como cuyo nombre indica, han sido heredadas del pasado, es decir, contienen información de la empresa previa a las bases de datos modernas. Es por ello, que muchas no podrán ser directamente asociadas a Big Data, pero eso no excluye el 100% de las bases de datos heredadas existentes.

4.3. Bases de datos NoSQL

Estas bases de datos destacan porque los sistemas de gestión no utilizan el lenguaje SQL, de ahí su nombre Not Only Structured Query Language[4]. A diferencia de las bases de datos relacionales no necesitan tablas fijas y son escalables. Como hemos indicado anteriormente las bases de datos relacionales ya no son las únicas existentes ni las más utilizadas, si no que este tipo de bases de datos empiezan a ser una alternativa real (utilizada por Twitter, Netflix, Cisco…), por lo que una vez introducidas las bases de datos existentes dedicaremos un apartado a desarrollar estas bases de datos.

Una buena definición de este tipo de bases de datos nos la da Paul Williams (2012) que lo entiende como "una colección de diferentes tecnologías cuya característica común es que son no relacionales."

---

[3] Hadoop es un marco de trabajo de fuente abierta utilizados para el almacenamiento de datos y la ejecución de aplicaciones. Como permite almacenar de forma masiva todo tipo de datos es de gran importancia para el Big Data y hablaremos de el más adelante.

[4] SQL es un lenguaje de programación cuya finalidad es la recuperación de información de sistemas de gestión de bases relacionales. En español significa lenguaje de consulta estructurada. Las bases de datos NoSQL no utilizan este lenguaje.

La principal diferencia de este tipo de datos es la ausencia de tablas y relaciones entre los datos almacenados, es decir, mientras que en las relacionales se impone un esquema de tablas, en las NoSQL esto no es así, permitiendo el almacenamiento de datos en otros formatos como documentos, gráficos, clave-valor…

Estas bases de datos presentan una serie de características:

- A diferencia de las relacionales, permiten almacenar datos en grandes cantidades sin necesidad de que estos presenten estructura o esquema fijo.
- La posibilidad de añadir nodos hace que sean bases de datos escalables, lo cual permite a las empresas ampliar su capacidad con menor coste.
- Como el almacenamiento se da en memoria[5] (se explica de forma más precisa en el siguiente apartado) y no en el disco el acceso es muy rápido, no obstante, esto exige que las operaciones se realicen en más de un nodo para evitar la pérdida de información en caso de fallos.

Dentro de las bases de datos NoSQL encontramos principalmente cuatro grandes categorías (Acens):

- **Bases de datos clave-valor**: es el tipo de base de datos más simple y popularizado. El funcionamiento de estas bases de datos es sencillo, cada elemento (valor) tiene asociado una clave única. Es un proceso que permite rescatar la información de una manera muy rápida. Este es el principal motivo de utilización de estas bases de datos, ya que permiten dar respuesta a las exigencias de velocidad de los Big Data. La base de datos más popular es Casandra, una base creada por Facebook y escrita en Java algo que hace que pueda ser utilizada en todas las plataformas que cuenten con este. Es una base creada para soportar grandes cantidades de datos y ser utilizada desde la gran mayoría de lenguajes de programación. Organizaciones que utilizan esta base de datos son: IBM, Twitter, Facebook…
- **Bases de datos documentales:** estas bases de datos comprenden distintos formatos, tales como: Word, PDF o XML entre otros. Cada registro va unido a una clave y cada documento presenta su propia estructura no teniendo que

[5] El almacenamiento en memoria se refiere a que se da en la memoria RAM en vez a en el disco duro. La velocidad de acceso a la memoria RAM es mayor que la de acceso al disco duro, por tanto, son bases más veloces.

guardar relación con el resto. (Amazon AWS). Permiten recuperar la información asociada a un documento de manera veloz, pero son lentos a la hora de buscar elementos comunes entre documentos al carecer de índices. MongoDB es de las más utilizadas actualmente.

- **Bases de datos en grafos[6]:** como es evidente, se organiza la información en grafos. Son muy útiles a la hora de representar las relaciones existentes entre los datos almacenados. El proceso consiste en almacenar los datos en nodos y establecer las relaciones que guardan estos. Son de gran utilidad para operaciones propias de consulta de relaciones como se realiza en las redes sociales. Una de las más utilizadas es Neo4J.
- **Bases de datos orientadas a Bigtable** (Joyanes, 2014, 189): estas bases de datos se las debemos a Google, ya que fueron creadas para responder a su necesidad de buscador web. Son bases de datos cuyo almacenamiento es distribuido, en las que se guardan datos estructurados que permiten la escalabilidad a través de los servidores. Para que nos hagamos a la idea, estas bases de datos son un mapa disperso e indexado por tres variables, el tiempo, la clave fila y la clave columna. MapReduce[7] es el algoritmo utilizado para crear y modificar estas bases de datos. En este tipo de bases destaca Apache Hbase, una versión de código abierto de la propuesta por Google. Es una base escrita en Java y cuya finalidad es el análisis en tiempo real de grandes volúmenes de datos, por lo que es de gran importancia. Hadoop es un marco de trabajo que utiliza Hbase como base para el tratamiento de los datos. Además, los nodos Hbase permiten trabajar con MapReduce.

4.4. Bases de datos En Memoria

Las bases de datos en memoria almacenan toda la información en la memoria RAM, de forma que son bases de datos mucho más veloces, ya que la velocidad de acceso al disco duro es superior a la de acceso a la memoria RAM (IBM, 2014)

---

[6] Un grafo es un conjunto de objetos denominados nodos enlazados entre ellos para representar relaciones de tipo binario.

[7] Mapreduce: su principal objetivo es el que indica su nombre, ya que realiza dos tareas con los datos, por un lado los mapea (conversión de los datos en pares de clave-valor) y, por otro, los reduce alojándolos en nodos. La finalidad es que los datos puedan ser procesados en distintos ordenadores, lo que aumenta la velocidad de procesamiento de grandes volúmenes de datos.

Este Sistema de Gestión de Bases de Datos (SGBS) cuya especialidad es la utilización de la memoria del ordenador en vez de utilizar los dispositivos de almacenamiento como hacen el resto de las bases. La gran ventaja de este sistema es que ahorra una cantidad considerable de tiempo al no tener que acceder al disco a la hora de realizar operaciones.

De esta forma en estas bases de datos toda la información es enviada a la memoria principal, realizándose un procesamiento más veloz. Están siendo de gran utilidad en el día a día de la empresa, debido a su importancia en el proceso de análisis de Big Data.

Dentro de las bases de datos en memoria encontrados dos tipos principalmente (Joyanes 2013, 197):

- En memoria puras: aquellas donde el modelo se carga en la RAM.
- En memoria Just in Time (JIT): A diferencia de la anterior, en vez de cargar todos los datos solo carga aquellos necesarios para dar respuesta a la consulta realizada. Este modelo es el resultado de la combinación de la tecnología en memoria con las bases de datos en columnas.

Finalmente, pese a que no forma parte de la clasificación original es preciso destinar un apartado a los sistemas de bases de datos MPP para entender más en profundidad su funcionamiento.

4.5. Bases de datos MPP

Son bases de datos idóneas para el almacenamiento de grandes cantidades de datos. El funcionamiento de este tipo de bases de datos es el siguiente. Los datos son distribuidos en bloques independientes, que serán gestionados y almacenados también de manera independiente. Estos datos son distribuidos a pequeños conjuntos de la CPU, de forma que se replica una estructura de ordenadores pequeños que gestionan y almacenan pequeñas cantidades de información. La finalidad de esto es sencilla, cuando se realiza una consulta, en vez de tener que coger una gran masa de datos y buscar dentro de ella, el ordenador va a la parte de la CPU que almacena esa información, de forma que se realizan consultas más pequeñas de manera simultánea.

Por tanto, una definición más técnica de MPP sería (Joyanes, 2013, 181) “procesamiento coordinado de un programa por múltiples procesadores que trabajan en partes diferentes del programa y con cada uno de los procesadores utilizando su propio sistema operativo y memoria”

Las razones que explican el auge de esta base de datos son los siguientes:

- En primer lugar, como vimos cuando hablamos de las características de los datos, la velocidad y el volumen con que son creados hace necesario que existan bases de datos que puedan dar respuesta a estas necesidades.
- En segundo lugar, las empresas no pueden realizar desembolsos constantes de dinero para incluir tecnologías de bases de datos a la velocidad que surgen estos, por lo que se necesita un sistema escalable y fácilmente replicable según aumenta el volumen de datos con el mínimo coste posible.
- Finalmente, los usuarios exigen respuesta en el mismo momento que realizan su consulta y está puede ser a cualquier hora del día, por lo que la velocidad es una característica obligada.

4.6. Hadoop

Antes de adentrarnos en el análisis de datos realizado para obtener valor de estos es importante dedicarle un apartado a Hadoop, ya que es la herramienta más utilizada por las empresas para analizar los volúmenes de datos ante los que se enfrentan hoy en día. Como vimos anteriormente, Hadoop es un marco de trabajo de código abierto inspirado en The Google File System. El funcionamiento de Hadoop es sencillo, ya que se trata de un software presente en una gran cantidad de servidores interconectados, de manera que cuando los datos son cargados en Hadoop y este programa los aloja en los diferentes servidores interconectados entre sí, de esta forma se da respuesta al problema de almacenar una cantidad de datos que supera la capacidad de almacenamiento de una sola máquina.

Hadoop comprende dos elementos:

- Hadoop Distributed File System (HDFS): Sistema de almacenamiento distribuido en un conjunto de servidores distintos.
- Hadoop MapReduce: Es el núcleo central de este software y permite el procesamiento paralelo de esos archivos alojados en diferentes máquinas, para así poder analizar grandes cantidades de datos.

5. ANALÍTICA DE BIG DATA

Cuando hablamos de analítica de Big Data ocurre como con el propio concepto de Big Data, donde no hay una única definición, por lo que en aras de elegir aquella que goza de mayor reconocimiento hemos optado por la propuesta por la asociación Isaca en 2011: "Analítica de datos (Data Analytics) engloba aquellos procesos y actividades diseñadas para obtener y evaluar los datos, obteniendo de esta información útil".

Dentro del análisis de datos encontramos distintas categorías (Joyanes, 2011):

- Analítica de datos más propiamente de organizaciones. Es el proceso del análisis de los datos obtenidos de forma tradicional.
- Analítica web: nos referimos al análisis de los datos obtenidos del tráfico web de una página.
- Analítica social: aquellos datos provenientes de las redes sociales, blogs o la misma Wikipedia entre otros.
- Analítica móvil: como su nombre indica se trata del análisis de los datos generados por los móviles, ya sea los que reciben, envían e intercambian entre ellos.
- Analítica de Big Data: la centrada en grandes volúmenes de datos.

Por tanto, vemos que la analítica de Big Data es una categoría del análisis de datos.

5.1. Analítica de Big Data

La analítica de Big Data está compuesta de dos cosas, de analítica y de Big Data (Russom, 2011). La analítica de Big Data sirve a las empresas para analizar los cambios que se producen, de forma que esta herramienta provea a la organización de medios para afrontar los retos surgidos por este inmenso crecimiento de datos. Además, la analítica es un proceso que permite a las empresas descubrir mejores proveedores, nuevos clientes, observar la estacionalidad en su negocio…

Para Joyanes (243, 2013):

Analítica de Big Data es el proceso de examinar grandes cantidades de datos de una variedad de tipos para descubrir patrones ocultos, correlaciones *desconocidas y otra información útil. Dicha información puede proporcionar ventajas competitivas (...), brindar beneficios en los negocios (...) y un aumento en los ingresos.*

La analítica avanzada se compone de una serie de técnicas orientadas a obtener ventajas para la empresa que las implementa. Dentro de las diferentes técnicas y herramientas que engloba, hay que destacar la analítica predictiva, minería de datos, análisis estadísticos de los datos y programación utilizando lenguaje SQL. No obstante, el concepto de analítica avanzada está dejando paso al de analítica de descubrimiento o analítica exploratoria, ya que el usuario, es decir, el analista de la empresa, a partir de la existencia de un gran volumen de datos nuevos, está tratando de descubrir oportunidades y hechos existentes previamente en la empresa pero que no eran conocidos por esta.

El análisis de Big Data suele llevarse a cabo siempre a partir de las mismas herramientas de software, de entre las cuales podemos destacar: análisis estadístico de datos, minería de datos, de textos, web y social, análisis y modelado predictivo, visualización de datos y consultas en lenguaje SQL.

5.2. ¿Por qué juntar Big Data y análisis?

Existen una serie de razones que justifican la existencia de la analítica de Big Data y su mayor implementación en las empresas (Russom, 2011). Big Data, como ya hemos visto a lo largo del trabajo provee de una gran cantidad de datos a las empresas, además, las herramientas de análisis de datos, minería y análisis estadístico han sido optimizadas para responder ante estas grandes cantidades de datos, lo cual si seguimos la regla general podemos ver que cuanto mayor sea el numero de datos que poseemos, más revelador o preciso será el análisis o las respuestas que busquemos.

Por tanto, vemos como primer motivo de la creciente popularidad del Análisis de Big Data el hecho de que las herramientas y las bases de datos hayan sido optimizadas para su utilización con Big Data.

En segundo lugar, como vimos a lo largo de las bases de datos, el coste del almacenamiento ha experimentado una caída considerable, de manera que cada vez se tiene un almacenamiento mayor por un precio muy bajo

En tercer lugar, el análisis permite que las empresas den respuesta a preguntas que no han sido si quiera planteadas, aumentando su rentabilidad, permitiéndoles ahorrar costes o captando clientes, lo que hace que merezca la pena la inversión.

Como sabemos, las organizaciones dependen del Big Data para tomar las mejores decisiones, por tanto, existen una serie de herramientas destinadas al análisis de estos datos que deben ser contemplados por las empresas, de entre las cuales vamos a destacar los siguientes.

5.3. Herramientas de informes y visualización

Las herramientas de informes y visualización están siendo utilizadas para transferir los datos almacenados en las bases de datos en información relevante para la empresa, que genera informes para ser posteriormente analizados. (Mounica, 2016)

Estos informes son distribuidos de forma periódica entre los interesados y constituyen una de las herramientas de mayor utilización dentro de las compañías, debido a que dentro de las diferentes herramientas son de las más rápidas y sencillas. Por otro lado, existen situaciones donde se requieren herramientas más complejas.

En cuanto a las herramientas de visualización, no dejan de ser el medio donde se plasman de manera visual los informes, y dentro de las más utilizadas encontramos los cuadros de control (dashboards) y los cuadros de mando integral (balanced scorecard).

Los cuadros de control pueden recordar a los cuadros de mando de un coche y permiten plasmar los datos de los informes para ser comprendidos de forma más visual, presentando los datos con herramientas como gráficos o tablas y llegando a ser interactivos.

Los cuadros de mando integral son herramientas más complejas utilizadas para la implementación de las estrategias empresariales, ya que permiten enlazar las medidas con objetivos.

5.4. Análisis predictivo

Para entender porque puede ser útil el análisis predictivo hay que entender la lógica que subyace al modelo predictivo. El modelo predictivo tiene en cuenta los datos existentes, trata de dar respuesta a porqué se dan esos datos (un resultado concreto), monitoriza lo que está ocurriendo y trata de predecir lo que va a ocurrir en el futuro.

(Moro, 2014). Por tanto, lo que hace el análisis predictivo es extraer un modelo analítico que pretende dar respuesta a lo que va a ocurrir en el futuro. Por tanto, puede ser que el análisis predictivo se apoye en otras herramientas como la minería de datos para construir este modelo.

5.5. Minería de datos (Data Mining)

Data Mining puede ser definido como el proceso de búsqueda de información valiosa extrayéndola de una base de datos. Es decir, consiste en extraer modelos o patrones de los datos. (Fayyad, 1996)

Data Mining es utilizado principalmente con dos objetivos.

- La predicción de tendencias.
- Extracción de patrones no conocidos anteriormente. Esto se complementa con las herramientas de Business Inteligence, ya que permite predecir lo que sucederá. (Como vimos anteriormente es una herramienta del análisis predictivo)

Esta herramienta se apoya en diferentes ciencias como la estadística, la matemática, el aprendizaje de máquinas o incluso la inteligencia artificial con el objetivo de extraer de bases de datos información relevante para la empresa. Los patrones que pretende extraer pueden ser correlaciones, tendencias o modelos predictivos. Finalmente, debemos saber que existen herramientas más específicas según la fuente de datos como son la minería web, de textos o de sentimientos.

La ventaja principal de la minería de datos es que permite extraer datos antiguos presentes en bases de datos, donde no existía el conocimiento de esos datos.

La minería de datos puede presentar las conclusiones extraídas de diferentes formas:

- Detección de asociaciones, de forma que te recomienda productos en función de lo comprado.
- Clustering: dentro de los datos, busca aquellos elementos que guardan relación entre sí.
- Árboles de decisión.

- Series Temporales: combina magnitudes y tiempo para tratar de predecir comportamientos.

Dentro de la minería de datos, debemos reservar un apartado para hablar de la minería de sentimientos. El análisis de sentimientos o minería de sentimientos permite analizar la opinión expresada en un texto, red social o documento. (Zambrano et al, 2017) Es decir, nos permite conocer las reacciones positivas, negativas o neutras vertidas en palabras. Esto es importante debido a la proliferación del uso de las redes sociales y de la influencia de estas a la hora de tomar decisiones por las empresas. Las herramientas utilizadas para este análisis están dotadas de una gran capacidad de procesamiento ya que aquella frase que contiene palabras positivas puede ser negativa y viceversa.

5.6. Analítica Web

La analítica web, es definida por Avinash Kaushik (2012, citado por Joyanes, 2013, 160) como:

> *El análisis de datos cualitativos y cuantitativos de su sitio web y de la competencia, para impulsar una mejora continua de la experiencia online que tienen tanto los clientes habituales como los potenciales y que se traduce en unos resultados esperados, tanto online como offline.*

Se trata de unas herramientas de gran importancia dentro de la analítica de Big Data debido a que las páginas web son fuentes de creación de grandes volúmenes de datos. Por tanto, es importante que cuando se analizan estos datos se integren tanto los datos propios como los de la competencia para poder tener una visión holística.

En el análisis web se basa principalmente en analizar aquellos lugares en los que el usuario ha pinchado, es decir, el flujo de clics. Es este flujo el que representa utilidad para el analista ya que le permite conocer, almacenar y posteriormente analizar (con las herramientas de analítica web necesarias) el comportamiento del usuario en la web.

Existen dos componentes fundamentales a la hora de analizar los clics del usuario: las métricas y los indicadores de rendimiento clave (conocidos en inglés como KPI o Key Performance Indicators[8])

[8] En los diferentes artículos académicos, pese a que estén escritos en Castellano los Indicadores Clave de Rendimiento suelen ser llamados KPI o Key Performance Indicators.

Las métricas son medidas cuantitativas que permiten, como su nombre indica, medir diferentes aspectos de relevancia relacionados con una página web. (Wingu, 2016) Existen diferentes métricas, pero de manera generalizada, las que más se utilizan en el día a día son:

- Visitas: veces que se ha entrado a la web en un lapso temporal. No es de gran utilidad, ya que se utiliza normalmente visitante único.
- Visitante único: el mismo usuario puede acceder diferentes veces que solo se contabilizará como una visita. Permite hacerse una idea real del alcance de una página web.
- Tiempo en la web/sitio: útil para conocer cuanto está un visitante en la web y en que sitios exactos.
- Tasa de salida: muestra la página de la web desde donde se ha cerrado esta. Es muy útil para las páginas de comercio para conocer, por ejemplo, el numero de personas que cuando ve que existen gastos de envío deja de comprar y abandona la web.
- Compromiso: permite conocer lo fidelizados que están los usuarios, midiendo el número de veces que acede un visitante único.

Existen otras métricas, pero no resultan de gran importancia, por lo que conociendo estas es suficiente.

En cuanto a los Indicadores Clave de Rendimiento, hemos de indicar que se trata de una métrica, es decir, todos los Key Performance Indicators son métricas, pero no viceversa. ¿A qué se debe esto? A que los Indicadores Clave de Rendimiento están ligados a unos objetivos, por tanto, son aquellas métricas que nos permiten medir si estamos acercándonos o cumpliendo estos, de esta forma, aquellas métricas que no nos proporcionen información acerca del estado del negocio respecto de los objetivos no serán Key Performance Indicators. Un ejemplo sería el número de visitas únicas a un periódico online.

Una variedad de la analítica web que ha cobrado gran importancia con la creciente popularidad de los teléfonos inteligentes es la analítica web móvil. Esta analítica no es más que una variante de la analítica web para conocer como interacciona el usuario con la compañía a través de la página web diseñada para ser utilizada desde un dispositivo móvil.

5.7. Analítica social y Listening social

Son la consecuencia de la integración de los medios sociales en las organizaciones.

Social Listening es la práctica consistente en atender, escuchar y visualizar el comportamiento, las actitudes y razonamientos de los colectivos que deseas entender más allá de como se relacionan contigo o con tu marca para así entender sus influencias y extraer tendencias. (Rao, 2014)

Una vez realizado el proceso de escucha, es importante analizar la información que tienes, ya sea tras la realización de una campaña o de manera previa al lanzamiento de estas, para poder entender la aceptación o percepción que los clientes tienen de esta.

Realmente la escucha social se practica previa y durante la realización de la campaña o lanzamiento para extraer lo que el cliente piensa y siente hacía esta. El siguiente paso es el análisis de esta información.

El análisis social debe permitir a las empresas analizar el impacto de sus decisiones, pudiendo estas acceder a las interacciones con los clientes y extraer información valiosa acerca de la acción realizada, lo que el cliente esperaba o espera, debiendo ser capaz mediante este proceso de detectar oportunidades. Es decir, el análisis social comprende una serie de etapas, tales como: obtención de la información (datos obtenidos en la interacción con los clientes), análisis de los datos obtenidos, visualización de las posibles acciones y ejecución.

Un ejemplo de esta utilizando las redes sociales sería el lanzamiento de una campaña en Facebook y Twitter, la escucha social consistirá en acceder a los comentarios y conversaciones entre consumidores y tomar decisiones tras analizar esta.

Al igual que la analítica web existen métricas para la analítica social, son conceptos que han ganado popularidad con el auge de las redes sociales como: me gusta, trending topic[9], visualizaciones o seguidores.

[9] Se trata del tema del momento en una red social, es decir, de aquello que más se está hablando en un momento dado, por ejemplo, en Twitter.

5.8. Analítica M2M

Este proceso analítico engloba dos análisis, el que se da de los flujos de datos de máquinas interconectadas entre sí, y aquel más conocido como internet de las cosas.

Con el paso de los años es un número mayor de objetos los que presentan sensores que transmiten información, desde máquinas expendedoras a complejos sistemas presentes en aviones, por tanto, se utilizan herramientas para monitorizar los datos producidos por estos sensores y posteriormente analizarlos, extrayendo patrones o conclusiones.

6. BUSINESS INTELLIGENCE

En este apartado trataremos de definir lo que es Business Intelligence, que herramientas comprende y si existen diferencias o no entre Análisis de Big Data y el concepto de Business Intelligence. (También conocido como BI o Inteligencia de Negocio)

Por Business Intelligence (Sin Nexus) entendemos aquellas herramientas que permiten procesar los datos hasta convertirlos en información, para posteriormente convertir esta información en conocimiento, de manera que permita a la empresa una mejor elección a la hora de tomar decisiones.

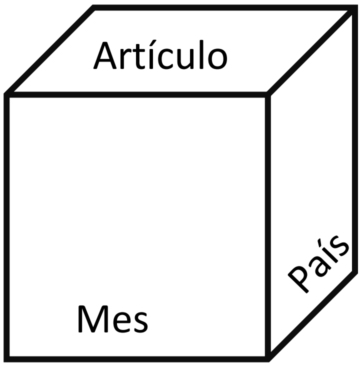
Si buscamos una información más técnica o precisa, definimos Business Intelligence como el proceso integrado por una serie de herramientas y tecnologías que permite agrupar, filtrar y convertir datos, desde los tipos de datos tradicionales y desestructurados a datos estructurados, facilitando su análisis y explotación, extrayendo información relevante para la compañía.

La Inteligencia de Negocio, siempre y cuando se realice de la manera adecuada constituye una fuente de ventaja competitiva para la empresa, ya que permite a la empresa obtener información detallada y fiable para hacer frente a la toma de decisiones o problemas a los que se enfrenta.

Por lo general, las aplicaciones de Business Intelligence suelen ser clasificadas dentro de tres grandes categorías: OLAP (On-line analytical processing) o análisis multidimensional, minería de datos (de la cual hemos hablado anteriormente y en breve entenderemos el porqué) y DSS (Decision Support System) o Sistemas de apoyo de las decisiones. (2013, Joyanes, p. 101):

- OLAP: como hemos visto se trata de un procesamiento analítico en línea, también conocido como análisis multidimensional que permite a las empresas conocer diferentes datos de forma multidimensional. Como veremos en la siguiente figura, se muestran unos datos determinados, si utilizásemos los sistemas utilizados por las empresas, veríamos que según vamos girando el cubo conocemos datos distintos.

Figura 5: Cubo de datos OLAP



Fuente: Elaboración propia

- Minería de datos: como vimos anteriormente es el proceso de búsqueda de información en una base de datos. Como si principal finalidad es la de detectar tendencias, para la Inteligencia de Negocio es útil de cara a entender lo que ha estado sucediendo en el pasado, que está sucediendo ahora y como tratar de anticipar o prever lo que va a suceder.
- Sistemas de apoyo a la decisión (Decision Support System): estas aplicaciones utilizan un conjunto de datos y modelos. Se utilizan principalmente para el análisis de problemas que no están estructurados, ya sean semiestructurados o no estructurados. Como todo modelo, se realiza una representación de la realidad simplificada para que sea más fácil de comprender y permite realizar el análisis presentando posibles resultados.

Por tanto, procederemos en el siguiente apartado a comparar la analítica de Big Data y Business Intelligence y a expresar si existen o no diferencia entre estos conceptos.

6.1. Diferencias entre Big Data Analytics y Business Intelligence

Para ilustrar de modo más visual antes de explicar las diferencias utilizaremos una tabla en la que compararemos los aspectos más relevantes en los que se diferencian ambos conceptos.

Tabla 4: Principales diferencias entre Big Data Analytics y Business Intelligence

	Big Data Analytics	Business Intelligence
Tipos de datos	Estructurados, semi-estructurados y no estructurados.	Estructurados, semi-estructurados y no estructurados.
Herramientas	Todas las que incluye la Inteligencia de Negocios, además de: Cuadros de Control, Cuadros de Mando Integral, Análisis Predictivo, Analítica Web, Analítica Social, Analítica de la Escucha Social y Analítica de Máquina a Máquina.	Minería de datos, OLAP y Sistemas de Apoyo a la Decisión.
Análisis	Gran capacidad de análisis de grandes cantidades de datos.	Los almacenes de datos son de menor cantidad. Menor capacidad de análisis.
Velocidad	Más reciente que Business Intelligence, por tanto, su velocidad de análisis es mayor debido a que incorpora mejores tecnologías.	Menor velocidad.
Escalabilidad	Gran escalabilidad a un coste inferior.	Peor escalabilidad a mayor coste.

Fuente: Adaptado de David López (2013)

Es cierto que a priori, sin un marco teórico que nos permita analizar punto por punto los dos conceptos, existe una tendencia a considerar que ambos términos son confusos o incluso lo mismo.

Como podemos ver en la tabla anterior, si miramos hacía el fondo, es decir, hacía la utilidad que tratan de aportar los dos conceptos, podemos ver que ambos son herramientas puestas a disposición de las empresas para una toma de decisiones más informada.

La realidad, es que Big Data Analytics es una evolución de Business Intelligence, pero una evolución natural, debido al surgimiento de un fenómeno como el Big Data se hace necesaria la existencia de una ciencia que permita procesar, analizar y almacenar esas grandes capacidades de datos para que las empresas sigan siendo capaces de obtener valor gracias a una mejor toma de decisiones. Por tanto, pese a que en ocasiones se puede llegar a vender el concepto de Business Intelligence como algo aislado, desde mi punto de vista no es cierto, sino que como hemos visto, es el precursor del Big Data Analytics y ha quedado superado por este.

7. CASOS DE ÉXITO EN LA IMPLEMENTACIÓN DEL BIG DATA

Para que nos hagamos una idea de lo importante que resulta el Big Data a las organizaciones, expondremos muy brevemente lo rentable que es este con un ejemplo.

En 2006, Netflix lanzó un importante concurso cuya finalidad era, mediante la utilización de la minería de datos y del aprendizaje automático, presentar un algoritmo para ser utilizado por Netflix y así poder mejorar el sistema de recomendación de películas aumentando, como mínimo, un 10% la precisión del sistema existente.

Lo diferente de este concurso es que el equipo, institución, empresa o particular que presentase este sistema, recibiría un premio de 1.000.000 de dólares. Finalmente, el premio fue entregado en 2009. Como curiosidad destacar que no ganó el equipo que presentó una solución que aumentaba un 10% la precisión, debido a que en ese su implementación era demasiado costosa y Netflix estaba cambiando del DVD al streaming[10], por lo que, el premio fue entregado al equipo que con un método mucho menos costoso conseguía un aumento del 8,4%.

Con este ejemplo podemos ver lo rentable que es el Big Data para las empresas, no obstante, ahora pasaremos a estudiar en profundidad dos casos de manera más desarrollada.

7.1. General Electric

GE factura 145 billones de dólares, y tiene una serie de negocios bastante diversificados, entre los que destacan la fabricación de motores de avión, las energías renovables o su presencia en el sector de la salud. General Electric se caracteriza entre otras cosas por su capacidad para realizar grandes apuestas, haciendo un enorme desembolso de dinero en mercados emergentes con un gran potencial, lo que significó la entrada de General Electric en el negocio de transformación digital, cuando observó que este mercado (Big Data) presentaba ese potencial del que hemos hablado.

En 2011 General Electric anunció la construcción de un centro de software global y los planes de inversión de 1 billón de dólares en software y conocimiento de para el análisis de Big Data. Para ello contrató a grandes expertos en la materia a los que encargó el desarrollo de ese software. Además, también contrató personal proveniente de Cisco Systems para dirigir el centro cercano a Silicon Valley, una zona de gran importancia al ser el epicentro de Start-Ups y compañías tecnológicas con mucho conocimiento en Big

[10] Streaming es la retransmisión de un contenido digital a través de un dispositivo electrónico. El producto no existe de manera física, se utiliza mientras se descarga. Hemos utilizado esta palabra por ser la común en este campo.

Data. Los planes del personal proveniente de Cisco eran la construcción de un centro de potenciación del Big Data y de Internet de las Cosas.

De esta forma, la idea de GE era tener un sitio donde poder almacenar todos los datos generados por los sensores y dispositivos digitales integrados en las máquinas fabricadas por General Electric, como pueden ser las turbinas de los aviones, los motores, locomotoras o incluso el equipamiento de los hospitales (Resonancias, TAC…) Estos flujos de datos que proceden de los sensores para ser tratados en el centro permiten a General Electric obtener una serie de ventajas como puede ser la mejor identificación de clientes, detección de problemas de manera previa a que sucedan, mejorar la eficiencia en repostaje de motores y locomotoras, e incluso realizar mejoras operacionales que sigan revirtiendo en un ahorro de millones de dólares. General Electric tenía como objetivo hacer de las máquinas objetos más inteligentes, que enviasen los datos adecuados a las personas correctas, para poder ser gestionadas en tiempo real.

Es importante en este momento hablar de las magnitudes de datos a las que nos enfrentamos, para que seamos capaces de entender lo que supone la integración de Big Data para General Electric.

Una turbina de gas genera aproximadamente 500 Gigabytes de datos al día, lo cual, si multiplicamos por las 12.000 turbinas poseídas por General Electric suponen sin duda alguna una gran cantidad de datos. Esto y para evitar dar números por cada elemento poseído por GE, puede ser replicado para motores de avión, parques eólicos, locomotoras… Ya que, si tenemos en cuenta que las plantas de energía y las turbinas de gas generan prácticamente un cuarto de la energía mundial, podemos ser capaces de entender ante el volumen de información que nos enfrentamos.

General Electric no ha dejado de invertir en este sector, como ejemplo, en el periodo de 2012-2016 invirtió un billón de dólares en su propio centro de software y análisis. Teniendo en cuenta las previsiones realizadas por la compañía estadounidense y partiendo de que se trata de la hipótesis conservadora, General Electric considera que puede conseguir en cinco sectores (aviación, energía, rail, healthcare y oil & gas) un ahorro de 300 billones de dólares en el periodo de 2015-2030. Si cogemos el sector de la aviación y tratamos de aterrizar estos datos, podemos ver que un ahorro de un 1% en carburante supone un ahorro de dos billones de dólares al año.

Por tanto, podemos ver que se trata de un fenómeno altamente rentable, ya que si sumamos las inversiones realizadas por General Electric y los retornos esperados (teniendo en cuenta que solo hemos cogido el ahorro que esperan obtener en 5 sectores), estos superan con creces a la inversión.

7.2. UPS (United Parcel Service)

UPS es una empresa de transporte de paquetes que opera en EEUU desde su sede de Atlanta, Georgia y en Europa, África y Oriente Medio desde su sede de Bienne, Suiza.

En este caso de estudio nos centraremos en Estados Unidos, donde un conductor de UPS se enfrenta a una media de 199 rutas diferentes para realizar sus 120 entregas diarias de media. Como podemos ver, existen multitud de posibilidades y una ruta equivocada supone mayor gasto de recursos y de tiempo. Como ocurría con el caso de General Electric, esto no es un hecho aislado, sino que, hay 55.000 conductores de UPS que deben entregar paquetes todos los días en Estados Unidos. Por tanto, de lo que vamos a hablar en este caso, es de como los conductores de UPS entregan cada día una media de 17 millones de paquetes a 9 millones de clientes, sin importar el contenido de este. La respuesta al como lo hacen no es otra que Big Data.

Partimos de otro número, que es el 1, ya que solo existe una única ruta que suponga el mayor ahorro en coste y recursos para realizar las operaciones diarias de cada conductor. Debido a esto, UPS ha gastado más de 35 millones de dólares en los últimos años. Rápidamente y para ver si se trata de algo rentable, hemos de saber que UPS tiene todos los días de media en EE. UU. 96.000 vehículos realizando entregas, por lo que una reducción de solo una milla en todos estos vehículos supondría un ahorro de 50 millones de dólares anuales. Un ahorro de un solo minuto supone 515.000 dólares de ahorro de gasolina y 14,6 millones de dólares de ahorro en costes operacionales.

Debido a la importancia del Big Data, UPS construyó una de las infraestructuras tecnológicas más complejas en relación con las empresas de paquetería, que cuenta con 10 IBM mainframes (ordenador de gran capacidad que realiza el procesamiento de grandes cantidades de datos) que operan las 24 horas del día 7 días a la semana. Esta sede es capaz de procesar más de 27 millones de instrucciones por segundo, lo que le permite rastrear alrededor de 16 millones de paquetes al día. Estos ordenadores, son los encargados de recolectar la información y distribuirla a los conductores de UPS en todos los países del mundo, así como, los pilotos de las aerolíneas. Esto significa que estos ordenadores se encargan de coordinar las operaciones globales de UPS, lo que en palabras de Randy Stashick (2014), presidente del departamento de ingeniería de UPS “Ha convertido a UPS de una empresa de camiones con tecnología a una empresa tecnológica con cambiones”.

La infraestructura de UPS es capaz de conocer donde se encuentra el paquete en cada momento, hacía donde se dirige y sí va a ser entregado a tiempo. Ocurre como hemos visto antes, esto multiplicado por los 17 millones de paquetes que se entregan al día supone un gran volumen de datos.

El proceso operacional realizado por UPS puede ser resumido en cuatro fases. Planificación, acción, comprobación y actuación en consecuencia.

Ups recopila los datos, analiza la situación y utiliza analítica predictiva para tratar de extraer patrones y predecir las necesidades futuras. Tras esto, basándose en las predicciones desarrollan un plan óptimo que les permita ser lo más eficiente posible. Este plan es ejecutado, mientras está siendo monitorizado, lo que les permite adaptar el plan en tiempo real. Finalmente, se cierra el circulo ya que, el plan es analizado y se vuelve a iniciar el proceso con el objetivo de seguir depurándolo hasta conseguir una gestión todavía más eficiente.

Este proceso se inició con la etiqueta PAL (Pre-Load Assist Label), que se imprime y se pega al paquete antes de que este sea entregado al transportista. Esta etiqueta permite al encargado de cargar el vehículo saber el orden y donde va a ser entregado, lo que permite cargar en función del orden. Pal es el primer componente de este proceso de entrega y genera los primeros datos. Otra herramienta de importancia es DIAD (Delivery Information Assistant Device), útil para el repartidor ya que configura el orden de reparto más eficiente e información personalizada que se haya pactado entre UPS y el cliente y que este no pudiese saber (repartidor) Estas herramientas están apoyadas por un sistema de predicción y planificación que genera las rutas de entrega diarias, analizando cada entrega para detectar posibles ineficiencias con el ánimo de corregirlas, es decir, se encuentra en constante revisión, lo que permite alterar el orden en tiempo real. El conjunto de sensores utilizados por UPS permite a estos conocer desde el dato más pequeño como puede ser si el conductor se ha abrochado el cinturón, hasta el tiempo que el vehículo ha estado en trayecto respecto al tiempo que ha estado parado mientras se entregaba el paquete.

Big Data ha permitido a UPS una reducción de 140 millones de kilómetros por año, es decir, gracias a estas herramientas se ha evitado realizar esa cantidad de kilómetros extra, además de 33 millones de litros de gasolina, una reducción del 95% de los tiempos de entrenamiento de los trabajadores o que el 100% de los conductores de abrochen el cinturón, con el consiguiente ahorro de dinero aparejado.

Por tanto, UPS sigue apostando por el Big Data a la hora de mejorar su proceso operativo, ya que el análisis de Big Data y la mejora de los negocios van de la mano.

Finalmente, destacar que UPS está desarrollando ORION (On-Road Integrated Optimization & Navigation), una herramienta que le permite conocer la mejor forma de entregar los paquetes mientras conoce en tiempo real las necesidades de los clientes. (Por ejemplo, alguien que no va a poder estar en casa cuando lo entreguen y dijo que si estaría)

ORION combina mas de 250 millones de direcciones y puntos de datos, así como las preferencias de los consumidores, para dar a los conductores instrucciones personalizadas. Entre las ventajas de ORION es el menor gasto de gasolina y la reducción de emisiones de dióxido de carbono. Además, debido a esto se están pudiendo ofrecer otros servicios como la devolución en el mismo día o la alteración de la hora o del punto de entrega

8. CONCLUSIONES

El objetivo del trabajo consistía en conocer el concepto de Big Data y todo lo relacionado con este, desde los tipos de datos existentes y las fuentes de producción, hasta el proceso que siguen para que las empresas puedan obtener valor.

Por otro lado, la idea tras introducir esto era ver como gracias al Big Data las empresas son capaces de obtener valor, entendiendo por valor el lograr un ahorro en costes o una mejora de la productividad o de la rentabilidad, que les permitiese situarse en mejor posición que sus competidores. En otras palabras, la idea podría entenderse como si el Big Data permitía a las empresas obtener una ventaja competitiva, por lo que para concluir realizaremos un VRIO[11] (Butler et Al, 2014).

[11] VRIO significa Value, Rarity, Imitability y Organization, y es el framework o marco de trabajo más utilizado por las empresas a la hora de analizar la estrategia para apreciar la existencia de una ventaja competitiva.

Con este marco de trabajo analizaremos la existencia de la ventaja competitiva para las dos organizaciones, y, por tanto, podremos extraer mejores conclusiones.

La V significa Value, que viene a ser si este recurso es valioso en dos sentidos. Por un lado, si nos permite explotar una oportunidad y colocarnos en una posición superior a la anterior y, por otro lado, si nos permite neutralizar una amenaza, ya sea del entorno o de los competidores. Como hemos visto, este recurso, el Big Data, nos permite explotar una oportunidad, ya que ante un panorama como el actual en el que la creación de datos se realiza de manera masiva, permite extraer aquellos datos de interés para nuestro negocio y tomar mejores decisiones, ahorrar costes o aumentar la rentabilidad, por lo que vemos que se trata de un recurso valioso.

La R significa Rarity, o, en otras palabras, si está al alcance de todos o solo de unos pocos. Es cierto que el Big Data como tal, es decir, las grandes cantidades de datos están al alcance de todos. Sin duda alguna el Big Data como tal está al alcance de cualquier persona, ya que la mayoría de los datos no se encuentran restringidos en su acceso, pero si es cierto, que el análisis de estos exige unas infraestructuras, software y profesionales de gran avance y complejidad que no son accesibles para todo el mundo, lo que convierte a esto en un recurso raro o escaso. (*scarcity* o escaso, es la otra forma en la que podemos encontrar el análisis la rareza de un recurso)

La I significa Imitability, es decir, si un competidor tratase de imitar esto, ¿sería fácil, poco costoso o rápido de hacer? Como hemos visto a lo largo de los ejemplos anteriores, no debemos hablar solo de la cantidad de millones de euros invertidos en Big Data por las dos organizaciones, ya que en el peor de los casos, cualquier organización de tamaño grande tendrá la disponibilidad de ese dinero, pero de lo que si debemos hablar es de que son medidas que se empezaron a implementar casi una década atrás, por tanto, puede que para una empresa grande no sea costoso pero si será imposible de imitar al tratarse de un proceso que autoaprendizaje de las propias empresas. Por poner un ejemplo, podemos ver que en el caso de UPS sus propias herramientas están siendo adelantadas por otra herramienta desarrollada por ellos (ORION) lo que nos hace darnos cuenta de que es un proceso continuo y lo convierte en algo muy costoso y complejo de imitar.

Finalmente, vemos que la O significa Organization, o lo que es lo mismo, si está la empresa organizada para explotar este recurso. Por matizar este significado, viene a tratar de analizar si toda la empresa está organizada para utilizar de la manera más eficiente posible este recurso que les permite obtener una ventaja competitiva. En General Electric no nos cabe duda, ya que se trata de una de las empresas punteras en el ámbito de la implementación del Big Data y el Internet de las Cosas, pero es que en el caso de UPS tampoco, ya que como vimos, ellos mismos consideran que han pasado de ser una empresa de transportes con tecnología a una empresa tecnológica que realiza transportes, permitiéndonos ver la importancia que la explotación del Big Data obtiene en la empresa.

Tabla 5: Análisis de una ventaja competitiva. Framework VRIO

¿Es el recurso o el activo?				**Implicaciones competitivas**	**Desempeño Económico**
Valioso	**Raro**	**Difícil de imitar**	**¿Está la organización organizada para ello?**		
NO				Desventaja	Por debajo de lo normal
SI	NO			Igualdad	Normal
SI	SI	NO		Ventaja temporal	Por encima de lo normal
SI	SI	SI	SI	Ventaja sostenible	Por encima de lo normal.

Fuente: Adaptado de Butler et Al (2014)

Por tanto, vemos que como se han cumplido las cuatro condiciones, nos encontramos ante una ventaja competitiva sostenible en el tiempo, lo que permitirá a estas empresas obtener posiciones de liderazgo en su sector, y, por tanto, aparte de conocer como obtienen valor las empresas del Big Data vemos la entidad del valor que obtienen estas.

8. BIBLIOGRAFÍA

- ¿Qué es Business Intelligence? *Sin Nexus White Paper*. Universidad Autónoma Metropolitana de México.
- Acens (2014) Bases de datos NoSQL. Qué son y tipos que nos podemos encontrar. *Acens Whitepapers.*
- Adrian Merv (2011) Big Data. *Teradata Magazine.* Recuperado de: http://nxtbook.com/nxtbooks/mspcomm/teradata_2011q1/index.php?startid=8#/40%3E
- Amazon. ¿Qué es una base de datos de documentos? *Amazon Web Services.* Disponible en: https://aws.amazon.com/es/nosql/document/
- Anguiano, J (2014, junio 30) Características y Tipos de Bases de Datos. *Developerworks IBM.* Recuperado de: https://www.ibm.com/developerworks/ssa/data/library/tipos_bases_de_datos/index.html
- Arias, N. (2015, noviembre 13) Big Data: una herramienta fundamental para aumentar las ventas. *Marketing4ecommerce* Recuperado de: https://marketing4ecommerce.net/bigdata-herramienta-fundamental-para-aumentar-ventas/
- Data Analytics a Practical Approach. *Isaca White Paper*, 2011.
- Butler et Al (2014) The VRIO Framework: Evaluating Competitive Resources and Capabilities. *Contemporary Strategic Management* 5, 174-177.
- Cabanillas, S et al (2017) Generación de talento Big Data en España. Fundación *COTEC para la innovación.*
- Derteano, A et al. (2014). Big Data en el sector financiero español. *EY*
- Engel, P et al. (2015) Industry 4.0: The future of productivity and Growth in manufacturing industries. *BCG.*
- Espinel, V (2015) ¿Por qué son tan importantes los datos? *BSA. The Alliance Software*.
- Fayyad, U. et al (1996) The KDD Process For Extracting Useful Knowledge From Volumes of Data. *Comunications Of The Acm.* 39 (11).
- Gandomi, A y Haider, M (abril de 2015) Beyond the hype: Big Data concepts, methods and analytics. *International Journal of Information Management*, 35(2),137-144.
- Hardoon, D y Shmueli, G, (2014) Introduction to Big Data. *NTNU.*

- Heller, P et al. (2016) An Enterprise Architect´s Guide To Big Data. *Oracle Enterprise Architecture White Paper.* March
- Ishwarappa y Anuradha, J. (2015) A Brief Introduction on Big Data 5V´s. Characteristics and Hadoop technology. *Procedia Computer Science* 48, 319 – 324.
- Joyanes, L. (2013) Big Data: Análisis de Grandes Volúmenes de Datos en Organizaciones. Marcombo.
- Labrinidis, A y Jagadish, H (2012) Challenges and opportunities with Big Data. *Proceedings of the VLDB Endowment*, 5, 2032-2033.
- López, D. (2013) Análisis de las posibilidades de uso de Big Data en las organizaciones. *Trabajo de fin de Máster*. Universidad de Cantabria.
- Manyika et al (2011) Big Data: The next frontier for innovation, competition and productivity. *McKinsey*.
- Maté Jiménez. C (2014). Big Data. Un nuevo paradigma de análisis de datos. *Anales de mecánica y electricidad*. (nov-dic) 10-16
- MathWorks (2016) El Concurso de Netflix Prize y Los Sistemas de Aprendizaje Automático en Producción: Visión Para Iniciado. Mathworks Whitepaper.
- Moro, E. (2014) Big Data y Análisis Predictivo. *Instituto de ingeniería del conocimiento.* UC3M
- Mounica, B. (2016) Efficient Reporting Tool For Business Intelligence With Increasing Volume of Data. *International Journal of Computer Science And Mobile Computing*. 5 (1) (enero) 86-90.
- Naya, S (2018). Nuevo paradigma de Big Data en la era de la industria 4.0. *Revista electrónica de terapia ocupacional Galicia, TOG* 27,4-9
- Randy Stashick (2014) Big Data Delivers Big Results at UPS. *UPS Pressroom.* Recuperado de: https://pressroom.ups.com/pressroom/ContentDetailsViewer.page?ConceptType=Speeches&id=1426415450350-355
- Rao, A. (2014) Social Listening: How Market Sensing Trumps Market Research. *Resilence. A Journal of Strategy and Risk.* PWC
- Rayo, M. (2016, mayo 17) Tipos de datos en Big Data: clasificación por categoría y por origen. *Big Data Fundations.* Recuperado de: https://www.bit.es/knowledge-center/tipos-de-datos-en-big-data/

- Russom, P. (2011) Big Data Analytics. *TDWI Best Practises Report.* TDWI Research. Fourth Quarter 2011
- Soares, S. (2003, junio 3) Not your type? Big Data Matchmaker On Five Data Types You Need To Explore Today. *Datadiversity.* Recuperado de: https://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/#
- Suarez, E. (2008) ¿Qué es una base de datos relacional? *Revista Universidad de Puerto Rico Humacao, 2008.*
- Tata Consultancy Services (2013) Big Data Case Study. *Tata Consultancy Services.* Recuperado de: https://sites.tcs.com/big-data-study/ge-big-data-case-study/
- Valencia Plaza (2017, marzo 30). El volumen de negocio del Big Data se triplicará en Europa hasta los 5.500 millones de € en 2018. Disponible en: https://valenciaplaza.com/el-volumen-de-negocio-del-big-data-se-triplicara-en-europa-hasta-los-5500-millones-en-2018
- Watson, H.J. (2014) Tutorial: Big Data Analytics: Concepts, technologies and applications. *Communications of the Association for Information Systems*, 34, art. 65.
- Williams, P (2012, octubre 11) The NoSQL Movement. What is it? *Dataversity.* Recuperado de: https://www.dataversity.net/the-nosql-movement-what-is-it/#
- Wingu (2016) Manual de Métricas. *Tecnologías sin Fines de Lucro.*
- Zambrano et Al (2017) Innovación Para el Análisis de Sentimientos en Texto, una revisión de la técnica actual aplicando metodologías de crowdsourcing. *Economía y desarrollo*, 158 (2)

-