



Facultad de Ciencias Económicas y Empresariales

# **BIG DATA: ANÁLISIS DE GRANDES VOLÚMENES DE DATOS EN ORGANIZACIONES**

Autor: Clara Ranney Martínez  
Director: María Jesús Giménez Abad



## **RESUMEN**

Este trabajo pretende analizar el estado del arte y la problemática que presenta el Big Data, el nuevo paradigma de la analítica de grandes volúmenes de datos de la era digital, aplicado a las organizaciones.

Se analizan sus aspectos más característicos y relevantes, su evolución y la problemática que presenta para las empresas desde diferentes ángulos, en base a una amplia revisión bibliográfica y estudio de ejemplos ilustrativos.

Se concluye que el Big Data es una oportunidad para las empresas ya que es capaz de dar respuesta a preguntas que, de otro modo, no se habrían planteado.

La dificultad de un proyecto de Big Data reside no solo en la necesidad de disponer de recursos económicos, técnicos y humanos adecuados, sino en el cambio que exige a las empresas en el modo en el que se aborda la toma de decisiones. Por ello es indispensable, para su buen fin, asegurar la adecuada coordinación entre áreas implicadas y la esponsorización desde los niveles más altos de la organización.

## **ABSTRACT**

This work aims to analyze the state of art and the problems presented by Big Data, the new paradigm of the analytics of large volumes of data from the digital era, applied to organizations.

The most relevant and characteristic aspects are analyzed, its evolution and the problems it presents to companies from different angles, based on an extensive literature review and study of illustrative examples.

It is concluded that Big Data is an opportunity for companies since it can answer questions that, otherwise, would not have been raised.

The difficulty of a Big Data project lies not only in the need to have adequate economic, technical and human resources, but in the change it requires from companies in the way in which decision-making is approached. For this reason, it is essential, for its good purpose, to ensure adequate coordination between the areas involved and sponsorship from the highest levels of the organization.

## **PALABRAS CLAVE**

Big Data, Analytics, Sistema de toma de decisiones, predicción, ventaja competitiva, datos, volumen, calidad, modelos, algoritmos, coordinación.

## **KEYWORDS**

Big Data, Analytics, Decision making system, prediction, competitive advantage, data, volume, quality, models, algorithms, coordination.

## ÍNDICE

<b>1. Introducción</b> .....	<b>7</b>
<b>2. Contexto actual</b> .....	<b>10</b>
2.1 Economía digital y sociedad.....	10
2.2 Tendencias.....	12
2.3 Economía digital en España.....	13
2.4 Escenarios futuros .....	14
<b>3. Big Data: qué es / qué no es</b> .....	<b>16</b>
3.1 Big Data vs. Machine Learning e Inteligencia Artificial.....	16
3.2 Big Data vs. Business Intelligence .....	19
3.3 Definición de Big Data: las 5V's y más.....	20
<b>4. Retos inherentes a la naturaleza del Big Data</b> .....	<b>25</b>
4.1 Tipología de datos.....	25
4.2 Almacenamiento de datos.....	26
4.3 Calidad de los datos.....	31
4.4 Marco legal.....	32
4.5 Ciberseguridad.....	34
<b>5. El Big Data en las organizaciones</b> .....	<b>36</b>
5.1 Contexto en España.....	36
5.2 Impacto en las organizaciones.....	37
5.3 Transformación de una empresa hacia Big Data .....	38
5.3.1 Tipología de utilización de datos.....	39
5.3.2 Etapas del viaje empresarial hacia Big Data.....	40
5.3.3 Niveles de valor de Data Analytics .....	42
<b>6. Retos y riesgos en la implantación de Big Data en la empresa</b> .....	<b>46</b>
6.1 Retos y riesgos inherentes a la naturaleza del Big Data.....	46
6.1.1 Asegurar la calidad del dato .....	46
6.1.2 Asegurar el cumplimiento del marco legal (GDRP) .....	50
6.2 Retos y riesgos de tipo organizativo .....	52
<b>7. Modelos de negocio basados en el dato</b> .....	<b>57</b>
<b>8. Ciclo del Business Data</b> .....	<b>61</b>

<b>9. Modelos y algoritmos</b> .....	<b>65</b>
9.1 Modelos estadísticos.....	66
9.2 Modelos predictivos .....	68
9.3 Algoritmos .....	71
<b>10. Proyectos de Big Data</b> .....	<b>73</b>
<b>11. Caso de éxito</b> .....	<b>76</b>
<b>12. Conclusiones</b> .....	<b>80</b>
<b>13. Bibliografía</b> .....	<b>81</b>

## 1. INTRODUCCION

La actual fase de desarrollo de la Economía Digital, o Nueva Economía, nos sitúa en un momento disruptivo en el que la generación y análisis de datos ha adquirido una importancia única en la historia: “El dato es el nuevo petróleo”<sup>1</sup>

Como ejemplo, cada minuto se envían más de 187 millones de correos electrónicos, 18 millones de mensajes de texto, 38 millones de mensajes de WhatsApp. Se realizan cerca de 4 millones de búsquedas en Google y se suben 300 horas de vídeos nuevos a YouTube. Se publican al día más de 690 millones de tuits (Mittum, 2018) y se comparten 38 millones de mensajes “me gusta” en Instagram (Instagram, diciembre 2018). En media hora, el motor de un avión que vuela de Londres a Nueva York genera 10 terabytes de datos (MacKinnon, 2013).

Por eso, no es extraño que desde que en 2013 Kenneth Culkier y Viktor Schönberger en su ensayo “La revolución de los datos masivos” acuñaran el término “Big Data”, este se haya convertido en uno de los más utilizados a todos los niveles: académico, empresarial, medios de comunicación e incluso en análisis de resultados deportivos. Pero, quizás precisamente por esta sobreutilización del término, muchas preguntas de enorme relevancia no parece que tengan aún una respuesta clara.

¿Es “Big Data” sólo una tendencia más o va a suponer una transformación real en la forma de trabajar de las compañías? ¿Por qué surge precisamente ahora y no en otro momento? ¿Qué lugar ocupa en la Economía Digital?

Y otra aún más básica, pero no por ello menos relevante, ¿Qué es realmente “Big Data”? Y, por contraposición, ¿Qué no es “Big Data”? ¿Es una solución o un problema para las empresas? En definitiva, qué puede aportar y qué no cabe esperar del Big Data.

---

<sup>1</sup> Ascolta, A. (2018, 12 octubre). Ann Winblad: the woman who built Silicon Valley. Recuperado 3 mayo, 2019, de <https://www.morningfuture.com/en/article/2018/07/13/annwinblad-silicon-valley-venture-capitalist-investor-women-startup/365/>

A lo largo de este trabajo se va a tratar de dar respuesta a todas estas preguntas, desde un punto de vista de las organizaciones empresariales.

La importancia del Big Data para las empresas reside en el hecho de ser capaz de responder a preguntas que ni las propias empresas sabían que tenían que hacerse. Si siempre se ha considerado la información como un activo relevante de la empresa, con el Big Data se convierte en la clave competitiva, ya que permite obtener datos de distintas fuentes, almacenarlos y convertirlos en información accionable para descubrir patrones y tendencias clave para el negocio. Según un reciente estudio de IDC Research, las empresas que implanten estos modelos y que trabajen con ellos, tienen 5 veces más posibilidades de superar sus expectativas de negocio.

Y, sin embargo, solo el 15 % de las empresas consigue llevar su proyecto de Big Data a producción (Gartner, octubre 2016). En España, según la encuesta del Instituto Nacional de Estadística sobre uso de Tecnologías de Información y Comunicaciones (TIC)<sup>2</sup>, solo el 9% de las empresas de más de 10 trabajadores hizo algún uso de Big Data en el 2017.

Parece que el reconocimiento de la importancia del Big Data va en línea con la dificultad de llevar a cabo su implantación con éxito. Y eso a pesar de que, según este mismo estudio, ya solo un 10% de los proyectos de Big Data son considerados como de IT y la gran mayoría empiezan a liderarse desde las Direcciones ejecutivas de los departamentos implicados, esponsorizados por la Alta Dirección de la empresa.

El objetivo de este trabajo es conocer los principios y características de Big Data y analizar los retos y oportunidades que presenta para las organizaciones en el momento actual:

---

<sup>2</sup> Instituto Nacional de Estadística. (2018). *Encuesta sobre el uso de Tecnologías de la Información y las Comunicaciones (TIC) y del comercio electrónico en las empresas Año 2017*. Recuperado de [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176743&menu=ultiDatos&idp=1254735576799](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176743&menu=ultiDatos&idp=1254735576799)

- Definir la ubicación y aportación de Big Data en el contexto de la Economía Digital.

- Analizar los aspectos más relevantes relacionados con Big Data, su evolución, y la problemática asociada desde diferentes ángulos.

- Definir los contextos más propicios para poner en marcha un proyecto de Big Data y los retos de negocio que nos puede ayudar a resolver.

- Establecer los requisitos previos a cumplir antes de lanzarse a su desarrollo, así como los desafíos que tendrá que enfrentar la organización.

- Definir los puntos críticos en el desarrollo del proyecto y las opciones que tienen las empresas para superarlos, según el estado del arte tecnológico y las experiencias de organizaciones que lo han logrado.

- Presentar ejemplos y analizar resultados de casos reales de compañías, enfatizando las aportaciones de Big Data no previstas inicialmente.

El enfoque metodológico se hará tanto desde una perspectiva teórica, con revisión bibliográfica exhaustiva, como práctica, en base a ejemplos ilustrativos y a la participación de expertos que han estado involucrados en proyectos de Big Data. En la realización del trabajo se utilizará:

- Bibliografía sobre el tema, informes de consultoría, estudios de mercado.

- Entrevistas con expertos de distintos sectores, que hayan estado involucrados en un proyecto de Big Data desde distintos departamentos y niveles organizativos y que transmitan su experiencia, con independencia del grado de éxito final alcanzado.

- Documentación pública o confidencial, con los permisos oportunos para su uso académico.

## 2. CONTEXTO ACTUAL

### 2.1 Economía Digital & Sociedad

La Economía digital (ED), Nueva economía, Economía en Internet o Economía Web, se sustenta en los avances que la sociedad es capaz de conseguir a través del uso de las llamadas tecnologías de la información y la comunicación (TIC), las cuales proporcionan los instrumentos necesarios para el procesamiento, la administración y distribución de la información por medio de distintos soportes o dispositivos tecnológicos (ordenadores, teléfonos móviles, tablets, Smart TV, consolas de videojuegos, etc)

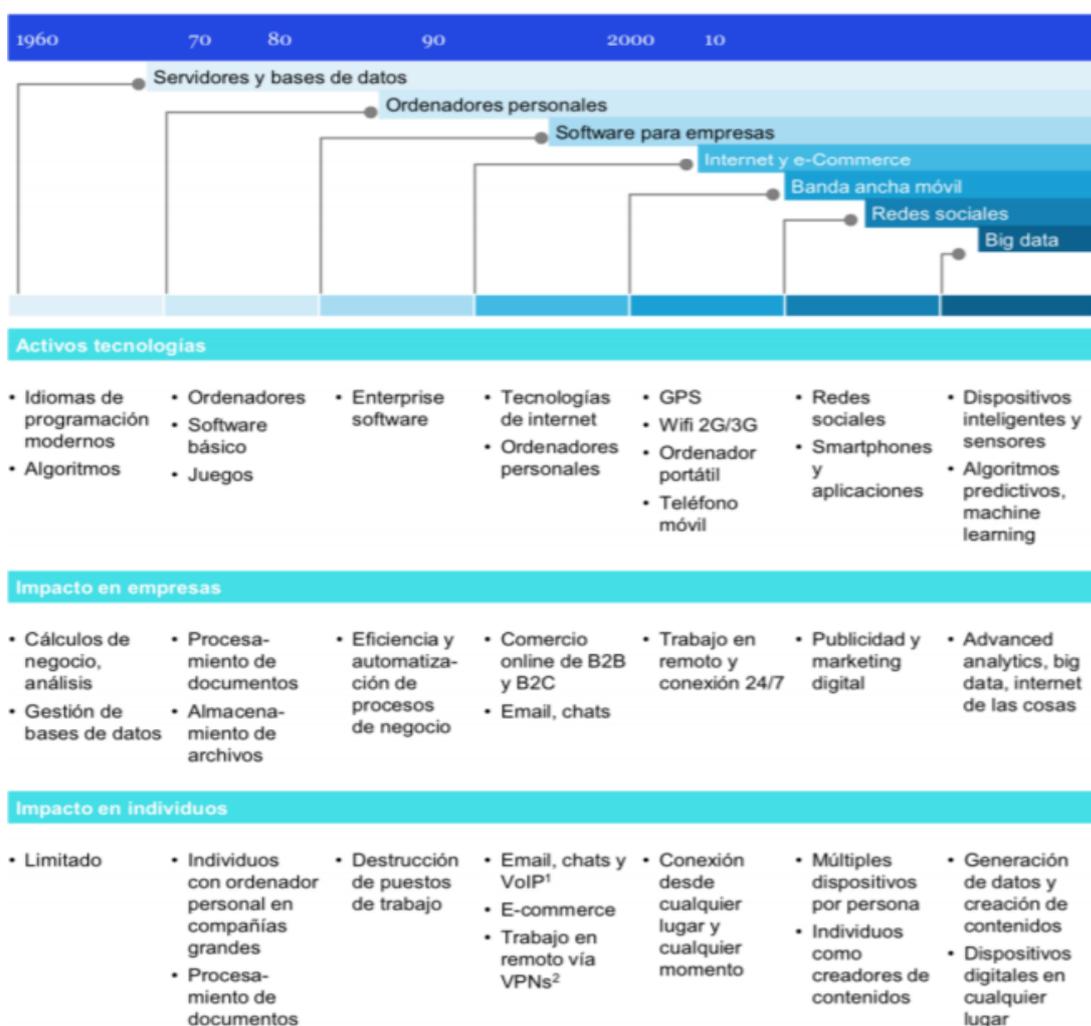
Las TIC favorecen el uso de la información de una manera más rápida, directa y barata, lo que facilita los intercambios de todo tipo y, en especial, los intercambios comerciales. La ED incluye tres tipos de componentes: las infraestructuras de negocios (tecnología, hardware, software, telecomunicaciones), el negocio electrónico o e-business (aplicaciones informáticas y plataformas on-line) y el comercio electrónico o e-commerce (negocios a través de Internet).

En el ámbito de las empresas, la ED permite abaratar los factores de producción (capital y trabajo), disponer de un menor volumen de existencias, reducir los tiempos de llegada de los productos y servicios a los mercados, disminuir los costes de transacción, ampliar el ámbito de actuación (mercados más amplios y extensos) y reducir los costes de acceso al mercado. Los beneficios para las empresas son considerables, ya que pueden incrementar su capacidad competitiva y de diversificación en los mercados, que se convierten en mercados globalizados, lo que también beneficia a los consumidores que obtienen un acceso más amplio, rápido y barato a bienes y servicios. Sus efectos están llegando a todos los sectores económicos: comercio, finanzas, transportes, hostelería, turismo, medios de comunicación...

La implantación de la ED es innegable, ya representa más del 20% del crecimiento del PIB a nivel mundial. La Estrategia Europea 2020 contempla como iniciativa la Agenda Digital para Europa. En España, según el estudio «*Digital disruption: the growth multiplier*» elaborado por la compañía Accenture, su impacto total en el 2015 fue de 231 mil millones de dólares, lo que representa el

19,4% del PIB; y se espera que este año llegue a alcanzar entre el 22% y el 24%. Sus efectos son patentes, tanto en la economía convencional o clásica, industria tradicional y sector servicios, como en las nuevas formas de economía basadas en las tecnologías. Se trata, por tanto, de un fenómeno económico ya consolidado, con fuerte implantación a nivel mundial y con enormes efectos sociales.<sup>3</sup>

La ED ha sido modelada por olas sucesivas de innovación. La rapidez de la implantación de la innovación tecnológica viene impulsada por la proliferación y expansión de plataformas digitales. La fase actual de digitalización se centra en conectividad, plataformas, software y datos.



Fuente: McKinsey & Company, 2017

<sup>3</sup> Accenture. (2018). *Digital disruption: the growth multiplier*. Recuperado de <https://www.accenture.com/acnmedia/PDF-14/Accenture-Strategy-Digital-Disruption-Growth-Multiplier-Brazil.pdf>

## 2.2 Tendencias

Según un reciente estudio (Gartner, 2017) habrá 20 mil millones de cosas conectadas a Internet en 2020. Estas "cosas" no son dispositivos de propósito general, como los teléfonos inteligentes o los PC, sino objetos de funciones específicas, como máquinas expendedoras, motores a reacción, automóviles conectados, etc. Al aumentar los dispositivos conectados, aumentan los datos procedentes de los mismos, lo que ha llevado a un gran esfuerzo por desarrollar las capacidades y tecnologías de almacenamiento y procesamiento a altas velocidades, produciéndose un increíble abaratamiento de ambas capacidades.<sup>4</sup>

“Los datos son el nuevo petróleo” es la frase con la que Ann Winblad, empresaria e inversionista de software, resume la importancia de los datos en esta era digital. Cada día se generan conexiones, datos y procesos que van más allá de nuestro conocimiento. Las relaciones que tenemos entre nosotros, con las máquinas, y las máquinas entre sí son cada vez más sofisticadas.

De su teoría se pueden resaltar tres aspectos. Primero, explica cómo los datos se convierten rápidamente en bienes sin los que el éxito en cualquier negocio es imposible. En segundo lugar, los datos ahora son el combustible que impulsa la nueva economía digital. Esta información es un material valiosísimo para empresas de todo tipo, no solo las de comercio electrónico y marketing. Finalmente, su expresión «los datos son el nuevo petróleo» nos da a entender que son un bien valioso y muy cotizado que, si no se protege, podría generar conflictos.<sup>5</sup>

En conclusión, la generalización en el análisis de grandes volúmenes de datos (Big Data) y su capacidad para generar negocios es definitivamente la característica más definitoria de la próxima, ya actual, ola de digitalización.

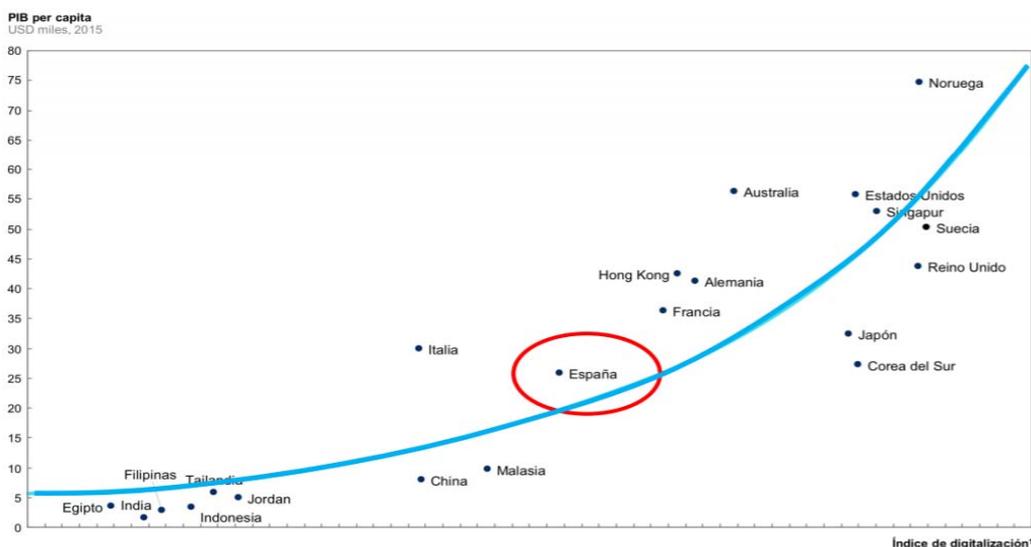
---

<sup>4</sup> Gartner. (2018). *Informe la digitalización de la economía* (Leading the IoT). Recuperado de [https://www.gartner.com/imagesrv/books/iot/iotEbook\\_digital.pdf](https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf)

<sup>5</sup> Ascolta, A. (2018, 12 octubre). Ann Winblad: the woman who built Silicon Valley. Recuperado 3 mayo, 2019, de <https://www.morningfuture.com/en/article/2018/07/13/annwinblad-silicon-valley-venture-capitalist-investor-women-startup/365/>

## 2.3 Economía digital en España

Como puede observarse en el cuadro adjunto, la digitalización y el PIB per cápita están fuertemente correlacionados.



Fuente: McKinsey & Company, 2017

España no llega al nivel de los principales países europeos como Suecia, Noruega, Reino Unido, Francia o Alemania, ni al nivel de Estados Unidos. Aun así, según el estudio, se encuentra a nivel de la media europea en la mayoría de las métricas digitales: individuos que han creado una página web, empresas comprando servicios cloud, empresas usando redes sociales, suscripciones de banda ancha, individuos participando en redes sociales o accediendo a internet a través del móvil regularmente, hogares con acceso a internet, compañías con conexión de banda ancha y cobertura 3G.

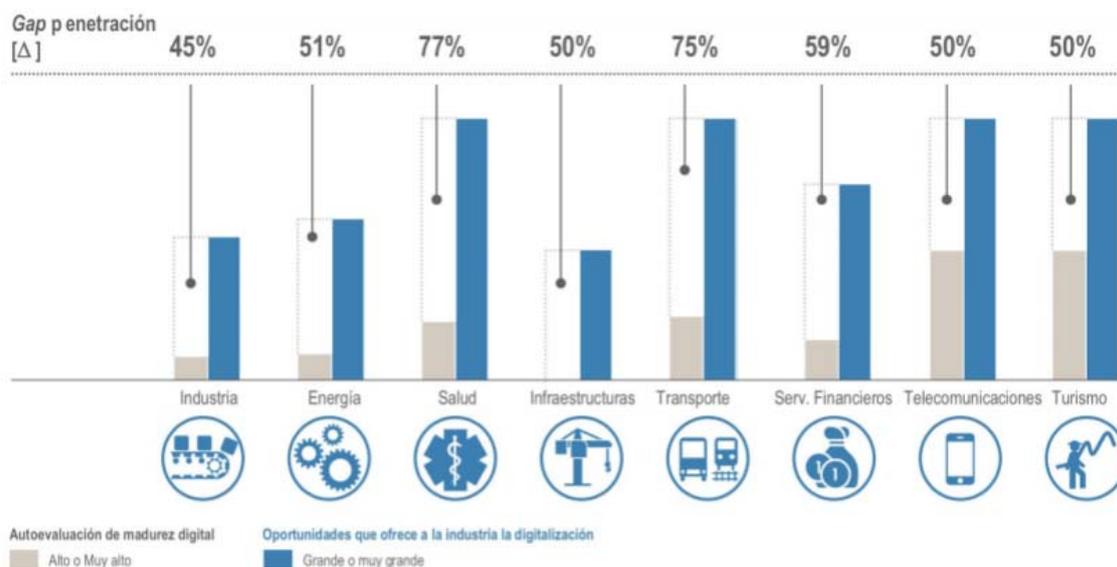
Según la Comisión Europea, DESI (*Digital Economy and Society Index*)<sup>6</sup>, España se ha convertido en un referente en despliegue de redes ultrarrápidas: en 2011 se encontraba por debajo de la media comunitaria en cobertura de fibra, mientras que, en 2017, tiene la red de fibra más extensa de Europa con 33,3 millones de accesos. Sin embargo, se encuentra muy por debajo de otros países en inversiones en capital riesgo, el porcentaje de facturación por comercio electrónico, y el uso de negocio digital, donde según el Informe global de

<sup>6</sup> European Commission. (2018, 14 mayo). Digital Economy and Society Index 2018 Report - Digital Single Market. Recuperado 6 abril, 2019, de <https://ec.europa.eu/digital-single-market/en/news/digital-economy-and-society-index-2018-report>

Tecnologías de la información elaborado por el Fondo Monetario Internacional, ocupa el puesto número 45.

## 2.4 Escenarios futuros

La siguiente imagen nos muestra el gap de penetración digital en España por sectores, usando como fuente el último Informe Global de Tecnologías de la Información del Fondo Monetario Internacional (Ronald Berger).



Fuente: Fondo Monetario Internacional (Ronald Berger)

Según este informe, los 10 pilares por seguir para conseguir la digitalización en España en los próximos años y lograr explotar al máximo las oportunidades que ofrece la digitalización a las diferentes industrias son: <sup>7</sup>

Las instituciones públicas:

- 1, Reducir las barreras regulatorias y la burocracia para promover el *Digital Single Market*.
2. Desarrollar el ecosistema digital a través de financiación y soporte a empresas.
3. Impulsar el capital humano digital y atraer y retener el talento digital.
4. Incrementar los flujos de datos abiertos.

<sup>7</sup> Consejo Económico y Social España. (2018). *Informe la digitalización de la economía*. Recuperado de <http://www.ces.es/documents/10180/4509980/Inf0317.pdf>

5. Continuar digitalizando los procesos de la administración.

Empresas:

6. Definir una estrategia digital integral y utilizar modelos de innovación abiertos.

7. Proporcionar herramientas digitales a los empleados y enseñarles a utilizarlas.

8. Educar a los clientes en digital e incentivar su transición.

9. Invertir en infraestructura de sistemas y en ciberseguridad.

Así pues, el fenómeno de Big Data como nuevo paradigma dentro de la economía digital, no es un fenómeno puntual, sectorial o transitorio, sino que forma parte intrínseca de la nueva economía y ha venido para quedarse. Aunque con una penetración aun limitada en el tejido empresarial, tal como se ha comentado en la introducción, cabe pensar que se irá desarrollando en la medida en que consiga ir cerrando el gap existente de penetración digital en los distintos sectores.

### 3. BIG DATA: QUÉ ES/ QUÉ NO ES

#### 3.1. Big Data vs. Machine Learning e Inteligencia Artificial

La utilización de aplicaciones prácticas de las nuevas tecnologías de Machine Learning (ML), Inteligencia Artificial (AI) y Big Data presentan una tendencia ascendente en todos los dominios de negocios y en la vida cotidiana. A menudo se usan estos tres términos como sinónimos, pero AI, ML y Big Data son, en realidad, muy distintos entre sí, a pesar de que los tres se refieran a tecnologías basadas en datos.<sup>8</sup>

La Inteligencia Artificial se define como el área de la informática que enfatiza la creación de máquinas inteligentes que funcionan y reaccionan como los humanos. La AI es común hoy en día y se usa de alguna manera en casi todas las principales plataformas de Internet. Aunque nos encontramos en el comienzo de la era de la AI, ya se usa en todas partes, desde en juegos simples como el Pac-Man hasta en coches totalmente autónomos.<sup>9</sup> Todos tenemos asistentes de inteligencia artificial en nuestros teléfonos como el Asistente de Google, Siri de Apple, Bixby de Samsung, Alexa de Amazon... Las aplicaciones de la IA van más allá de la imaginación.

El Machine Learning (ML) es el método para hacer que los ordenadores aprendan y piensen como lo hacen los humanos, por observación. La razón por la que el aprendizaje automático se está volviendo tan popular es porque para que funcione, no se tiene que saber cómo resolver un problema en particular, únicamente la naturaleza de un problema y su respuesta ideal, después ML funciona para resolver el problema en función de la cantidad de problemas similares con las respuestas correctas que se le hayan proporcionado. El aprendizaje generalmente se almacena como una pieza de software llamada modelo de aprendizaje automático. Estos modelos están entrenados con grandes cantidades de datos etiquetados que indican lo que representan. Por ejemplo: si está entrenando un modelo ML para identificar números con una cámara, entonces los datos de entrenamiento consisten en imágenes del número

---

<sup>8</sup> Ico. (2018). *Big data, artificial intelligence, machine learning and data protection*. Recuperado de <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf>

<sup>9</sup> BBVA. (2018, 22 febrero). ¿Qué es la inteligencia artificial? Recuperado 7 abril, 2019, de <https://www.bbva.com/es/que-es-la-inteligencia-artificial-2/>

junto con una etiqueta que indica qué número está presente en cada imagen dada. El ML implica ejecutar algoritmos sobre estos datos. Estos algoritmos reconocen patrones y tendencias en ellos y aprenden a predecir las etiquetas de los nuevos datos. De esta manera, los modelos de ML pueden hacer predicciones en función de lo que han aprendido de los datos con los que fueron entrenados.<sup>10</sup>

No existe unanimidad en la definición de Big Data, aunque si un cierto consenso en la fuerza disruptiva que suponen los grandes volúmenes de datos y la necesidad de su captura, almacenamiento y análisis.

La consultora tecnológica IDC considera la siguiente definición: "Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar, gestionar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad con el objeto de extraer valor económico de ellos".

Gartner, Inc. define Big Data en términos similares: "Big Data son activos de información de gran volumen, alta velocidad y gran variedad que exigen formas innovadoras y rentables de procesamiento de información para una mejor visión y toma de decisiones ".<sup>11</sup>

La definición de Big Data puede variar según las características de las empresas, para algunas prima el volumen, para otras la velocidad o la variabilidad de las fuentes.

Durante muchos años, los sistemas distribuidos (dos o más máquinas trabajando conjuntamente para alcanzar un mismo resultado) han permitido procesar cantidades relativamente grandes de datos a costa de una programación compleja ad-hoc. No existían en el mercado productos de referencia que permitiera la adopción de esta arquitectura fuera del ámbito académico o de investigación. Este escenario motivó la aparición de las tecnologías Big Data que permiten la gestión de cantidades masivas de

---

<sup>10</sup> Peter Jeffcock. (2018, 11 julio). What's the Difference Between AI, Machine Learning, and Deep Learning? Recuperado 7 abril, 2019, de <https://blogs.oracle.com/bigdata/difference-ai-machine-learning-deep-learning>

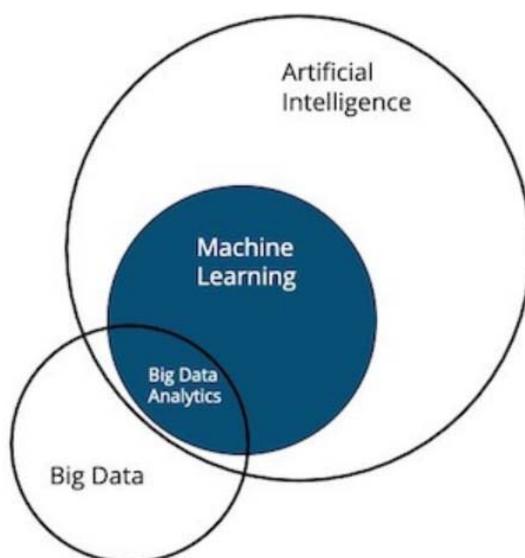
<sup>11</sup> Gartner IT Glossary. (2016, 19 diciembre). What Is Big Data? - Big Data. Recuperado 8 abril, 2019, de <https://www.gartner.com/it-glossary/big-data/>

datos. Desde el punto de vista de la arquitectura, Big Data no es más que un tipo de sistema distribuido compuesto por un conjunto de hardware de igual o incluso menor capacidad que los sistemas distribuidos tradicionales, y una capa de software capaz de optimizar esta arquitectura al máximo. De esta forma se pueden procesar y analizar grandes cantidades de datos de una manera más simple y tolerante a fallos.

Big Data implica grandes desafíos:

1. El tamaño del “Big data”, que supera la capacidad de almacenamiento y procesamiento de las bases de datos relacionales.
2. El gran porcentaje de los datos son no estructurados.
3. El análisis de los datos en tiempo real.

La inteligencia artificial es en realidad un concepto amplio que actúa como un paraguas bajo el cual vive el ML. Aunque parece similar, el ML tiene criterios completamente diferentes para llevar a cabo tareas. Se permite que las máquinas accedan a los datos y dejamos que aprendan por sí mismas, y los resultados se analizan más a fondo.



Big data puede definirse como la recopilación y el almacenamiento de estos datos para su uso futuro por parte de modelos de aprendizaje automático.

*Fuente: Oracle*

Podemos observar, pues, que AI, ML y Big Data son interdependientes, lo que significa que no pueden existir uno sin el otro. Si uno es sacado de la imagen, los otros pierden significado, o tal vez ni siquiera existieran.

### 3.2. Big Data vs. Business Intelligence

La importancia del análisis de los datos en los procesos corporativos no es reciente. Durante años las soluciones de Business Intelligence (BI) han permitido transformar los datos en información para optimizar los procesos de toma de decisiones empresariales. Las metodologías BI están basadas en centralizar la información de forma estructurada, utilizando bases de datos relacionales convencionales en distintos almacenes de datos o silos. El análisis de estos datos se realiza en modo offline y de forma focalizada en determinados ámbitos (ventas, calidad, marketing, producción, etc.) para responder preguntas concretas sobre situaciones del presente.

En cambio, Big Data, como un nuevo paradigma de almacenamiento, procesamiento, gestión, y análisis de datos, tiene como objetivo localizar patrones y tendencias que nos permitan extraer el conocimiento oculto en volúmenes masivos de datos, que se generan a gran velocidad, no necesariamente estructurados y almacenados en sistemas distribuidos.<sup>12</sup>

En la siguiente tabla se resumen las principales diferencias entre Big Data y Business Intelligence

	Business Intelligence	Big Data
<b>Foco</b>	Informes, KPIs, tendencias	Patrones, correlaciones, modelos
<b>Proceso</b>	Estático, comparativo	Exploración, experimentación, visualización
<b>Datos</b>	Planificado, estructurados, crece lentamente	'On the fly', según necesidad, no estructurados, crece por segundos
<b>Análisis</b>	Retrospectivo, descriptivo	Predictivo, prescriptivo
<b>Arquitectura</b>	Centralizada	Distribuida

*Fuente: Elaboración propia*

Big Data no sustituye, sino que amplía las posibilidades del BI, ya que permite realizar análisis con volumen de datos muy superiores y de todas las

---

<sup>12</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

tipologías (Ver Tipología de Datos), en menos tiempo, así como aplicar patrones y algoritmos de Machine Learning de una forma más eficiente.

### **3.3 Definición de BIG DATA: Las 5V's y mas**

Las primeras definiciones de Big Data hacían referencia e hincapié, sobre todo, en el gran volumen de datos a tratar. Sin embargo, las definiciones han ido evolucionando en el tiempo con el fin de resaltar no sólo “la gran colección de datos” y tener en cuenta las características propias de los mismos.

Así, varios autores James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, definen Big Data como un “set de datos cuyo tamaño está más allá de las habilidades típicas del software de datos tradicionales para capturarla, almacenarla, administrarla y analizarla”.<sup>13</sup> Gartner, la define como “activos de información de alto volumen, alta velocidad y variedad que requieren formas innovadoras y eficientes de procesamiento que habiliten su mayor comprensión y una toma de decisiones superior”<sup>14</sup>. IDC 2012 (Vesset et al., 2012) se refiere al Big Data como “la nueva generación de tecnologías y arquitecturas de TICs que están diseñadas para extraer valor económico desde grandes bases de datos en una amplia variedad de ellos a través de habilitar la captura en alta velocidad, el descubrimiento y el análisis de los mismos”.

Desde que en 2001 Doug Laney propuso tres características que definían el Big Data: Volumen, Variedad y Velocidad de los datos, se han ido agregando variadas ‘Vs’ para describir con mayor detalle su complejidad: Veracidad, Valor, Volatilidad, Variabilidad y Validez.

#### VOLUMEN

El volumen es la característica más propia y diferencial asociada al Big Data, ya que hace referencia a las cantidades masivas de datos que se almacenan con la finalidad de procesar dicha información, transformando los datos en información lista para tomar decisiones y realizar acciones.

---

<sup>13</sup> Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., & Roxburgh, C. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*

<sup>14</sup> Gartner IT Glossary. (2016, 19 diciembre). What Is Big Data? - Big Data. Recuperado 8 abril, 2019, de <https://www.gartner.com/it-glossary/big-data/>

Cada vez estamos más conectados al mundo 2.0 por lo que generamos más y más datos. Para aquellas empresas que están en el mundo digital la cantidad de datos generados es aún mayor. La siguiente imagen ilustra todo lo que ocurre en internet en 1 segundo: se crean 11 cuentas en Twitter, se mandan 5700 tweets, se descargan 634 aplicaciones en dispositivos de Apple y 1236 en Android, se compran 51 artículos en Amazon por un valor de 2.359 dólares, se suben a YouTube 2 horas de video, etc. Aproximadamente, en 1 segundo, se transfieren 22.574 GB de datos a través de internet.



*Fuente: The Internet in Real-Time interactive GIF*

El tamaño utilizado para determinar si un conjunto de datos se considera Big Data no está firmemente definido y es cambiante en el tiempo. La mayoría de los analistas y profesionales actualmente se refieren a conjuntos de datos que van desde 30-50 Terabytes a varios petabytes.<sup>15</sup>

En el Sistema Internacional de Unidades un carácter se almacena en un byte de información, un kilobyte (kB) equivale a 1000 bytes o caracteres. Un megabyte (MB) son 100 kB, un gigabyte (GB) son 1000 MB, un terabyte (TB) son 1000 GB y un petabyte (PB) son 1000 TB. Un libro de 700 páginas suele ocupar

<sup>15</sup> Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.

una media de 500kB, por lo que un petabyte serían más o menos 2.000 millones de libros de 700 páginas.<sup>16</sup>

No obstante, no se debe olvidar que el valor real no se encuentra en los grandes volúmenes de datos, sino en lo que se puede hacer con ellos.

## VARIEDAD

Una de las principales diferencias entre el Big Data y los sistemas tradicionales de almacenes de datos es la necesidad de almacenar y analizar datos en distintos formatos, tipos y estructuras de forma conjunta. Una aplicación sencilla puede generar y almacenar muchos tipos de datos. (Ver tipología de Datos y su almacenamiento)

## VELOCIDAD

Nos referimos a la velocidad a la que se generan los datos y la velocidad a la que deben analizarse y aplicarse. La proliferación de dispositivos digitales, como teléfonos inteligentes, sensores, ordenadores, tabletas, etc. ha llevado a una tasa de creación de datos sin precedentes y está impulsando una creciente necesidad de análisis en tiempo real y planificación basada en la evidencia.

Los datos que emanan de dispositivos móviles y fluyen a través de aplicaciones móviles producen torrentes de información que se pueden utilizar para generar ofertas personalizadas en tiempo real para los clientes diarios. Estos datos proporcionan información sólida sobre los clientes, como la ubicación geoespacial, la demografía y los patrones de compra pasados, que se pueden analizar en tiempo real para crear un valor real para el cliente.<sup>17</sup>

Por lo tanto, la velocidad hace referencia no solo a la generación rápida de datos sino también a la necesidad de su análisis en tiempo real dada su naturaleza cambiante, es decir, su volatilidad.

---

<sup>16</sup> IBM. (s.f.). Knowledge Center. Recuperado 15 mayo, 2019, de [https://www.ibm.com/support/knowledgecenter/es/SSNE44\\_5.2.4/com.ibm.tpc\\_V524.doc/fqz0\\_r\\_unit\\_s\\_measurement\\_data.html](https://www.ibm.com/support/knowledgecenter/es/SSNE44_5.2.4/com.ibm.tpc_V524.doc/fqz0_r_unit_s_measurement_data.html)

<sup>17</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

## VERACIDAD

El volumen y variedad de los datos de cada día es más amplio, lo que implica un mayor reto en la veracidad de los datos es importante contar con un equipo imparcial que asegure en los datos: confianza, autenticidad, disponibilidad, responsabilidad, origen y reputación.

Datos provenientes de aplicaciones, tarjetas de fidelidad o puntos de venta son muy confiables. En el caso de comentarios en redes sociales y datos menos confiables, el gran volumen de datos permite contrarrestar la desconfianza con una muestra suficientemente grande de datos para validar su veracidad.

De la veracidad de los datos recogidos y analizados depende en gran manera la validez de los mismos para un entorno de trabajo.

## VALOR

El valor se refiere a la extracción de conocimiento oculto a partir del análisis de grandes volúmenes de datos, y a la capacidad de transformar este conocimiento en mejoras en la toma de decisiones, reducción de riesgos y detección de nuevas oportunidades. El Valor potencial imbuido en los datos hace que 'Big Data' puede considerarse un activo estratégico en sí mismo, el cual debe ser monitoreado, administrado, protegido y hasta incluso comercializado.

Un ejemplo de cómo Big Data genera valor es la aplicación web del MIT Media Lab, en el Instituto de Massachusetts, que permite a los usuarios visualizar la economía mundial componiendo una narrativa visual sobre los países y los productos que intercambian. Genera millones de visualizaciones 'onthefly', apoyado por la herramienta de visualización D3plus.

<http://atlas.media.mit.edu/>

La compañía SAS introdujo la Variabilidad y la Complejidad como dos dimensiones adicionales de Big Data. La variabilidad se refiere a la variación en las tasas de flujo de datos. A menudo, la velocidad del Big Data no es consistente, presenta picos y mínimos periódicos. La complejidad hace referencia al hecho de que los grandes datos se generan a través de una gran

cantidad de fuentes. Esto impone un desafío crítico: la necesidad de conectar, combinar, limpiar y transformar los datos recibidos de diferentes fuentes.<sup>18</sup>

En definitiva, Big data no puede considerarse una colección inmensa de datos de una misma fuente, que no se actualiza con rapidez o que presentan la misma estructura y valor equivalente. Para poder hablar de Big Data, es preciso que se den las características que lo cualifican más allá de su cuantificación, pues son éstas de las que se derivan las oportunidades que lleva implícitas.

---

<sup>18</sup> SAS. (s.f.). Big Data, mejor Marketing. Recuperado 1 junio, 2019, de [https://www.sas.com/es\\_mx/insights/big-data/big-data-marketing.html](https://www.sas.com/es_mx/insights/big-data/big-data-marketing.html)

## 4. RETOS INHERENTES A LA NATURALEZA DEL BIG DATA

El volumen y la variedad en los datos, así como su velocidad de generación y transformación, implican retos muy relevantes en cuanto a estructuración y almacenamiento de los mismos.

Por otra parte, la veracidad de los mismos nos enfrenta a la problemática de la calidad del dato. Por último, no hay que olvidar dos aspectos transversales como son la ciberseguridad y los aspectos legales y éticos relacionados con la propiedad del dato en sus distintas fases de captación, almacenaje y utilización.

### 4.1. Tipología de Datos

Existen, básicamente, tres tipos de datos en base a su estructura. De la variedad de los mismos y el peso de cada tipología depende en gran medida su riqueza y por lo tanto las oportunidades que puedan derivarse de su análisis.<sup>19</sup>

#### DATOS ESTRUCTURADOS

Todos aquellos con longitud y forma definida. El 20% de los datos de una organización son estructurados<sup>20</sup> y suelen almacenarse en data warehouse o data marts. Corresponden a datos de finanzas, ventas, almacén, etc. Son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos, normalmente se consultan a través de lenguaje SQL. Algunos ejemplos son: fecha de nacimiento (DD/MM/AA), DNI, número de cuenta bancaria...

#### DATOS SEMIESTRUCTURADOS

Tienen un esquema definido al igual que los estructurados, pero a diferencia de estos el esquema es flexible, tienen etiquetas y otros elementos marcadores que permiten separar los elementos del dato. Corresponden a formatos del tipo EDI, HTML y XML. Un ejemplo son los registros *Web log* compuestos por distintas piezas de información que cumplen distintos propósitos.

---

<sup>19</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

<sup>20</sup> Grupo PowerData. (s.f.-c). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad.

Recuperado 7 marzo, 2019, de <https://www.powerdata.es/big-data>

## DATOS NO ESTRUCTURADOS

Opuesto a los estructurados, carecen de formato fijo. Representan el 80% de los datos de una organización. Se almacenan como documentos u objetos y apenas se tiene control sobre ellos, por lo que son los más difíciles de dominar para los analistas. Corresponden a datos de tipo imágenes de satélite, datos científicos, redes sociales, dispositivos móviles, etc. <sup>21</sup>

Por ejemplo, la compañía Starbucks tiene 24000 locales en mundo, 150000 empleados, 36 millones de seguidores en Facebook, 13 millones de usuarios en la aplicación móvil, millones de comentarios en redes sociales, miles de visitas a la página web, tarjetas de fidelidad entre otros tipos de datos. Los datos estructurados de Starbucks son la información que proviene de los puntos de venta y del club de fidelidad, los datos no estructurados son los que provienen de la página web, y los datos no estructurados son fotos y comentarios en redes sociales.

La existencia de un gran número de datos no estructurados o semiestructurados presenta un reto importante para hacerlos manejables y compatibles con el resto

La adopción de tecnologías que satisfagan esa necesidad no es una cuestión sencilla y supone un nuevo contexto con utilización, en muchos casos, de sistemas híbridos. Si el Business Intelligence se centra en analizar la causa de un resultado basándose en fuentes estructuradas de información, las tecnologías que sacan partido de los grandes datos se basan en la analítica predictiva. (Ver Modelos predictivos)

### **4.2. Almacenamiento de datos**

El principio más básico de las bases de datos es que las decisiones que se toman a la hora de diseñar y seleccionar su forma de almacenamiento tienen

---

<sup>21</sup> Telefónica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

una enorme repercusión en su eficiencia, el coste, la seguridad del sistema, el mantenimiento y la velocidad y facilidad de consulta.

El almacenamiento y procesamiento de datos ha sido una de las tareas que se les han encargado a los ordenadores prácticamente desde su aparición. La información se almacenaba en cintas, y cada registro y todos sus componentes básicos, llamados atributos, se añadía de forma manual. Pero ni empresas, instituciones o administración hacían mucho más que almacenar porque su procesamiento era costoso en tiempo y esfuerzo.

Afortunadamente, el almacenamiento de datos ha sufrido una gran transformación y se ha abaratado enormemente. Hace dos décadas hacía falta una máquina del tamaño de un frigorífico y con un peso de 800 kilos para almacenar un gigabyte de datos. Hoy en día, hay teléfonos móviles de más 200 gigabytes de capacidad de almacenamiento.<sup>22</sup>

El precio de los dispositivos de almacenamiento también ha bajado. En esas dos décadas, el coste de almacenar un gigabyte ha pasado de más de 700 euros a cuatro o cinco céntimos.

Además, la velocidad de procesamiento ha aumentado enormemente con la aparición, a principios de los años 2000, de la computación paralela masiva. En vez de procesar tareas de una en una, ahora los ordenadores pueden procesar una gran cantidad de tareas en paralelo, es decir, todas a la vez. Así es como Google, Facebook o Amazon han sido capaces de construir sus servicios.

No sólo el hardware ha aumentado su velocidad. También ha sido decisiva la aparición de software inteligente que permite sacar partido de la capacidad de procesamiento paralelo, analizando grandes cantidades de datos en muy poco tiempo. Un detalle importante es que este software inteligente puede analizar, no sólo datos almacenados y estáticos, sino también datos volátiles que se analizan en tiempo real a la vez que se producen.

---

<sup>22</sup> Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.

## **Bases de datos**

Hasta la llegada de Internet, cualquier base de datos relacional era suficiente para almacenar grandes cantidades de datos, podían tratar sin ningún tipo de problemas tablas que contenían millones de filas.

### Bases de datos relacionales

Las bases de datos relacionales aparecieron por primera vez en 1970, en un artículo escrito por el programador y antiguo piloto de la RAF Edgar Codd. Su objetivo era solucionar el problema de la calidad de las bases de datos.

Codd proponía dividir los datos en estructuras con forma de tabla (relaciones) que se combinan entre sí para obtener la información. Los registros pasan a llamarse filas, y los atributos pasan a llamarse columnas. El problema de la redundancia se resolvía dividiendo tablas grandes en varias más pequeñas, que no incluían aquellos datos que no se pueden duplicar, evitando errores y el desperdicio de recursos. Adicionalmente, Codd propone separar el cómo se graban los datos (nivel físico) del nivel lógico o conceptual. Así, el usuario no tiene que preocuparse de la distribución de la información, lo hace un programa especializado llamado “sistema gestor de bases de datos”<sup>23</sup>. La compañía Oracle desarrollo un lenguaje de consultas de bases de datos relacionales denominado SQL, hoy en día, este lenguaje continúa siendo el estándar. Las bases de datos relacionales fueron la primera gran revolución en el mundo de los datos, pero con el avance de la tecnología surgieron una serie de problemas.

En primer lugar, la imposibilidad de las bases de datos de admitir el uso de lenguajes de programación orientados a objetivos, como C++ o Java, que permitían representar de forma natural estructuras de datos complejas y aparecieron en 1990. Este problema, el de combinar dos lenguajes o tecnologías diferentes en el mismo sistema se conoce como el “problema de impedancia”. Este problema se trató de resolver creando un nuevo modelo de bases de datos, las orientadas a objetivos. Su éxito fue escaso.

---

<sup>23</sup> Codd, E. (1970). A Relational Model of Data for Large Shared Data Banks. Recuperado 6 marzo, 2019, de <https://www.seas.upenn.edu/%7Eezives/03f/cis550/codd.pdf>

En segundo lugar, la llegada de Internet volvió a suponer un problema de impedancia, esta vez entre la complejidad y variedad de una página web y la simplicidad y homogeneidad de las tablas relacionales.

Por último, la generación de nuevos datos crecía a un ritmo mucho mayor que la capacidad de consumir y procesar esos datos. En 1980, el científico I.A. Tjomsland decía que “la gente continúa acumulando datos sin medida. El coste que conlleva mantener datos obsoletos parece menos evidente que el coste que el coste que puede conllevar la pérdida de datos potencialmente útiles”.<sup>24</sup>

### Big Data y NoSQL

En 1990, Peter J. Denning en su artículo “Saving All the Bits”, insiste en la imposibilidad de guardar todos los datos y aporta una solución: que los ordenadores sean capaces de encontrar patrones comunes en los datos, de procesarlos para obtener lo que se desea y que a ser posible puedan procesar datos de forma rápida en tiempo real.<sup>25</sup>

En 1997, Michael Cox y David Ellsworth utilizan por primera vez el término Big Data para referirse a sistemas que requieren características que sobrepasan las posibilidades de las bases de datos relacionales estándar.

Todas las bases de datos relacionales son a su vez bases de datos NoSQL. Sin embargo, NoSQL se refiere en general a cualquier base de datos alternativa al modelo relacional, aunque no tenga las características del Big Data. Un ejemplo sería la base de datos Redis, especialmente útil con bases de datos de pequeñas.

La principal razón de la definición por negación asociada al término NoSQL es que hay una gran cantidad de bases de datos no relacionales, cada una con sus propias características. Dentro del ecosistema NoSQL, que rompe con la estructuralidad, podemos distinguir cuatro bases de datos distintas: las

---

<sup>24</sup> Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.

<sup>25</sup> Denning, P. J. (1990). Saving All the Bits. Recuperado 7 abril, 2019, de <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19910023503.pdf>

orientadas a documento, las bases de datos clave-valor, las orientadas a columna y bases de datos orientadas a grafo.

La llegada de grandes volúmenes de datos ha supuesto también la aparición de lenguajes específicos, como R, creado en 1995 en la Universidad de Auckland (Nueva Zelanda) o Python. Estos lenguajes incluyen versiones especializadas en minería de datos y para entornos Big Data.

El pronóstico de la mayor parte de los expertos es que, en los próximos años, seguirá incrementándose la tendencia a la baja de las bases de datos relacionales.<sup>26</sup>

### **Cómo se almacenan los datos**

Habida cuenta de que las bases de datos tradicionales se diseñaron para albergar y trabajar con datos estructurados, resultan inadecuados para asumir un crecimiento masivo de los mismos, muy a menudo adoptando la forma no estructurada. Son, en definitiva, una herramienta que no puede dar respuesta a los datos masivos heterogéneos que se engloban bajo el término de Big Data.

Cuando la cantidad de datos a almacenar creció exponencialmente y ya no cabía en los discos, y los discos cada vez más grandes y con más memoria eran muy caros y a menudo requerían de personal dedicado a ellos, se popularizó la idea de los clústeres de ordenadores: una gran cantidad de ordenadores pequeños conectados entre sí trabajando de forma conjunta. Los ordenadores no tienen ni teclado ni pantalla, pueden ser de gama baja y se les conoce como nodos de clústeres, están conectados para compartir información.

Esta idea es la auténtica revolución detrás del Big Data. Se trata de un modelo escalable, añadir más ordenadores no conlleva grandes inversiones ni supone notables pérdidas de eficiencia. También ofrecen velocidad, todos los ordenadores trabajan en paralelo sobre un fragmento de la base de datos.

Las compañías agrupan sus clústeres en los centros de procesamientos de datos. Suelen encontrarse en lugares apartados, suponen una gran cantidad

---

<sup>26</sup> Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.

de consumo eléctrico, no solo por la energía necesaria para hacer funcionar los ordenadores, sino también por la corriente eléctrica necesaria para mantenerlos refrigerados.

Propietarios de centros de datos, permiten a las empresas alojar sus datos en sus centros a cambio de alquiler, esto se llama “alojamiento en la nube”.<sup>27</sup>

Los Data Lakes (Lagos de Datos) son plataformas para la gestión de datos corporativos que permiten realizar diferentes tipos de análisis sobre datos en su formato original. La idea es aprovisionar una vez y acceder múltiples veces para analizar, independientemente del tipo de dato (estructurado y no estructurado).

Los Datawarehouse son un repositorio de datos corporativos que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.

Los Data Mart son un repositorio de datos diseñado para atender necesidades específicas de un departamento. Muchas veces se confunde con Datawarehouse pero la realidad es que un Data Mart es un subconjunto del Datawarehouse.<sup>28</sup>

A este respecto, NoSQL, Hadoop y su rico ecosistema se han revelado como soluciones especialmente útiles por ser escalable, su bajo coste, eficiencia y fiabilidad, y lo mismo cabe decir de los recursos y servicios ofrecidos en el marco del cloud computing.

### **4.3 Calidad de los datos**

La calidad del conocimiento extraído de los datos dependerá de la calidad de los mismos. Pero aún más, ¿de qué sirve analizar y sacar conclusiones que apoyen la toma de decisiones si la calidad de datos es deficiente? Una deficiente calidad de datos puede llevar a decisiones erróneas con su consiguiente coste empresarial.<sup>29</sup>

---

<sup>27</sup> Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.

<sup>28</sup> Telefónica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

<sup>29</sup> Grupo PowerData. (s.f.). Calidad de Datos. Cómo impulsar tu negocio con los datos. Recuperado 12 abril, 2019, de <https://www.powerdata.es/calidad-de-datos>

Para garantizar la calidad del conocimiento debemos garantizar la calidad del dato. No existen estandarizaciones, ni una talla única en lo que se refiere a data quality. Se trata de asegurar, en función del contexto de análisis a realizar, una serie de facetas del dato:

**ACCESIBILIDAD:** Un analista tiene que tener los datos accesibles, facilitando permisos y también poniendo a su disposición herramientas que conviertan los datos en usables y analizables.

**CONSISTENCIA:** Existe acuerdo en los datos. Sólo una fuente de datos debe ser considerada fuente maestra. Cuando existan conflictos entre datos se desecharán aquellos hasta el entendimiento del motivo de desacuerdo.

**COHERENCIA:** Los datos se pueden combinar con otros de manera precisa. La coherencia será impulsada por el conjunto de claves que unen los datos en las diferentes partes de la base de datos.

**CONFIABILIDAD Y OPORTUNIDAD:** Los datos están completos (tienes todos los datos que deberías esperar) y a la vez son precisos (los datos proporcionan información correcta. Existe un margen de tiempo razonable entre la recogida de datos y la disponibilidad o liberación del análisis.

**DEFINICION:** Cada campo individual de datos tiene que tener un significado bien definido y sin ambigüedad.

**RELEVANCIA:** Los datos tienen que mantener relación con el análisis que se lleva acabo.

**PRECISION:** Los valores representan el verdadero valor de los mismos.

**COMPLETITUD:** No puede haber datos faltantes. Ni parcial no totalmente.

#### **4.4 Marco legal**

El Reglamento General de Protección de Datos (GDPR) (Reglamento 2016/679) es un reglamento por el que el Parlamento Europeo, el Consejo de la

Unión Europea y la Comisión Europea refuerzan y unifican la protección de datos para todos los individuos dentro de la Unión Europea (UE).<sup>30</sup>

A partir del 25 de mayo de 2018, las empresas que operen en Europa, en cualquier sector y con independencia de su país de origen, están obligadas a ser transparentes sobre la forma de recopilar, guardar y procesar datos personales. También se ocupa de la exportación de dichos datos personales fuera de la UE.

El objetivo principal es dar control a los ciudadanos y residentes sobre sus datos personales y simplificar el entorno regulador de los negocios internacionales unificando la regulación dentro de la UE. La ambiciosa norma incluye temas como la privacidad por diseño, el derecho al olvido, la notificación de cualquier incidente que implique fugas de datos sensibles o el consentimiento expreso para el uso de información personal.

Si un consumidor no desea recibir comunicaciones, puede registrarse en la lista de Robinson de la Federación de Comercio Electrónico y Marketing Directo (FECEMD), ahora llamada Asociación Española de la Economía Digital. Los usuarios a quienes se solicitan datos personales (por ejemplo, número de teléfono móvil) deben ser informados adecuadamente sobre la existencia de un archivo de datos personales, su propósito y el destinatario de la información, así como sus derechos de acceso, rectificación o cancelación.

Son muchos los que opinan que” el dato no es el petróleo, sino la confianza”, por lo que resulta indispensable, en palabras de Roberto Viola, responsable de la agenda digital en la Comisión Europea, “regular la propiedad y el intercambio de los datos porque si no somos responsables en este tema, toda la economía digital colapsaría”.<sup>31</sup>

El GDPR impacta directamente en el almacenamiento, procesamiento, acceso, transferencia y divulgación de los registros de datos de un individuo y afecta a cualquier organización a nivel mundial que procese datos personales de

---

<sup>30</sup>BBVA Fintech e innovación. (2018, 25 abril). Big Data: Ejemplos reales del uso. Recuperado 15 abril, 2019, de <https://www.bbva.com/es/ejemplos-reales-uso-big-data/>

<sup>31</sup> Iglesias Fraga, A. (2018, 28 septiembre). Innovadores | Los datos no son el nuevo petróleo: es la confianza. Recuperado 10 abril, 2019, de <https://innovadores.larazon.es/es/not/los-datos-no-son-el-nuevo-petroleo-es-la-confianza>

personas de la Unión Europea. Adicionalmente, el GDPR se debe aplicar a todos los datos, ya sean proporcionados voluntariamente o recopilados por sistemas automatizados. Esto incluye datos personales almacenados y utilizados en Data Lakes y plataformas de Big Data.

El Big Data complica el proceso de mantener el cumplimiento del GDPR así como otras normas de protección de datos de empresas. El volumen de datos es enorme y éstos son transformados en diferentes formas y usados de distintas maneras.

#### 4.5 Ciberseguridad

Uno de los principales problemas del Big Data es la falta de medidas de seguridad a la hora de almacenar, transferir y/o tratar esos datos. La diversidad de fuentes, la cantidad y su variabilidad, incrementan y complican la seguridad de los procesos. Algunos de los problemas más frecuentes son:<sup>32</sup>

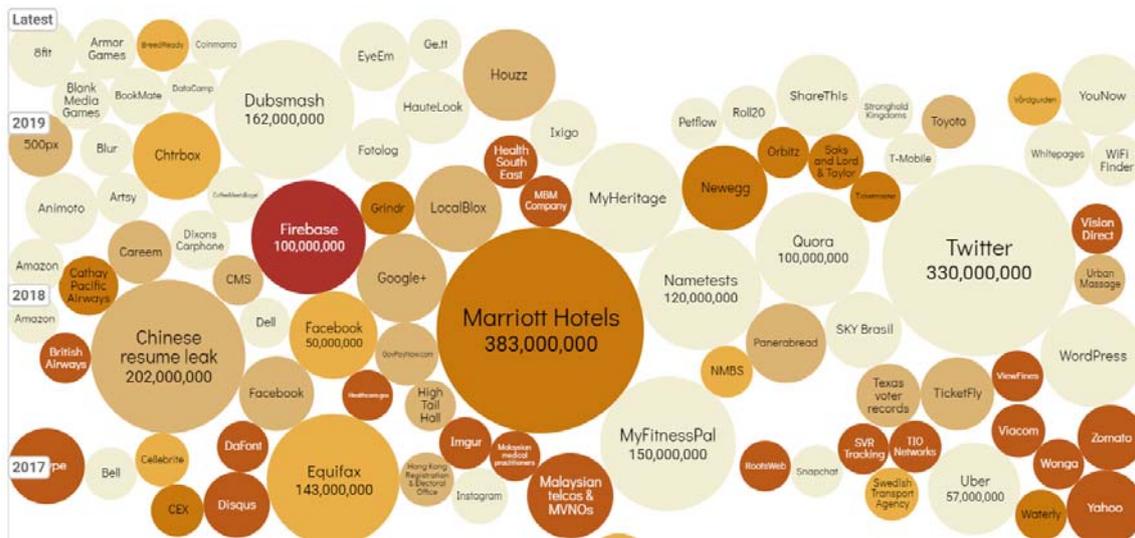
- **Software o proyecto diseñado con fallos de seguridad.** Este es uno de los errores más habituales, probablemente provocado por la falta de experiencia de estas plataformas en un sector tan novedoso como el Big Data. El ejemplo paradigmático es la principal empresa tecnológica de nuestros tiempos, Google, que tuvo que cerrar su red social, Google +, al exponer datos de 500.000 usuarios de forma pública.
- **Falta de protección de información personal.** Este factor es clave para que una compañía pueda usar los datos de terceros con total libertad. (Ver apartado legal). Siguiendo con Google, un error de seguridad que permitió a desarrolladores de aplicaciones de terceros acceder a información personal de 52,5 millones de perfiles de usuarios.
- **Pérdida o robo de datos.** Este es el problema más habitual del Big Data, prácticamente todos los meses aparece un nuevo caso en el que, por ejemplo, se han robado miles de bases de datos de correos en una red social.

---

Euroforum. (2019, 14 enero). Seguridad y Big Data. ¿Son compatibles? | Euroforum. Recuperado 30 abril, 2019, de <https://www.euroforum.es/blog/ciberseguridad-y-big-data-son-compatibles/>

<sup>32</sup>

También es muy común la pérdida de series de datos por parte de las empresas que los gestionan.



Fuente: Databreaches.net, IDTheftCentre and media reports

En el gráfico puede apreciarse qué empresas han sufrido los problemas de seguridad más famosos a nivel mundial relacionados con la pérdida o el robo de información en Big Data.

Aunque las amenazas online han aumentado y están a la orden del día, Malware, Ransomware, Phishing..., la solución parece encontrarse en el propio Big Data ya que, gracias a los propios datos, es posible anticipar los ataques e incluso crear modelos predictivos que ayuden a detectar los puntos más sensibles del sistema informático. Hay que tener en cuenta que el factor tiempo es vital cuando se produce un ciberataque y, si se consigue identificar con rapidez, se puede atajar de una forma mucho más eficaz.

## 5. EL BIG DATA EN LAS ORGANIZACIONES

Como ya se ha comentado, la utilidad del Big Data en las empresas reside en la capacidad que el análisis masivo de datos proporciona para dar respuesta, en base a las tendencias y patrones observados, a preguntas que ni siquiera sabían que se tenían que hacer, lo que mejora en conjunto su capacidad para la toma de decisiones y, en consecuencia, su situación competitiva.

La cadena de valor del Big Data proporciona un marco que examina cómo reunir datos dispersos de forma organizada y crear información de valor que pueda servir para la toma de decisiones a nivel empresarial.



*Fuente: Telefónica Digital*

Big Data es, por tanto, un proceso nuevo, diferente a lo hecho hasta ahora, complejo y con importantes zonas de incertidumbre. Precisamente en esas áreas imprecisas y desconocidas que el Big Data hace aflorar es donde se encuentran las oportunidades.

### 5.1. Contexto en España

Tal como se indicaba en el contexto de la Economía Digital, todos los sectores y mercados se ven afectados por la nueva economía y en concreto por la tendencia actual que enfatiza el análisis de datos. En este sentido, cualquier empresa podría beneficiarse del desarrollo de Big Data.

En España, los últimos datos publicados por el Instituto Nacional de Estadística (INE) en la "Encuesta sobre el uso de Tecnologías de la Información y las Comunicaciones (TIC) y del Comercio Electrónico en la empresa", se observa que solo el 11,2% de las empresas de 10 o más empleados realizaron

análisis de Big Data en 2017. Este porcentaje es 2,4 puntos superior al del año anterior. Si se analiza por Comunidades, destacan Madrid Cataluña y Galicia. Se puede concluir que existe aún un gran camino por explorar y recorrer en este tema.<sup>33</sup>

Por otra parte, no hay que olvidar que el tejido empresarial español se compone básicamente de autónomos y pequeñas y medianas empresa (PYMES) existiendo solo un 0,2% de grandes empresas, más de 250 empleados, (Ministerio de Empleo y Seguridad Social, enero 2018)<sup>34</sup>.

Aunque el Big Data no es algo que manejen únicamente las empresas de Silicon Valley y cualquier industria puede beneficiarse de los datos, incluso las compañías más tradicionales del mundo, no es menos cierto que para pymes y autónomos el reto aun mayor, pues no solamente les faltan los recursos económicos para realizar ciertas inversiones, sino que también se enfrentan a la falta de información, de recursos humanos formados o de infraestructura para sumarse al reto del Big Data.

## **5.2. Impacto en las organizaciones**

El valor real que el Big Data genera en las empresas se refleja en 3 áreas diferenciadas:<sup>35</sup>

### **1. Captación de nuevos ingresos.**

Facilita la creación de nuevos productos para satisfacer necesidades de clientes no detectadas con anterioridad.

Permite personalizar las campañas de Marketing para lograr una mayor efectividad.

Mejora la experiencia de usuario.

Personalización de productos y servicios.

---

<sup>33</sup> INE. (s.f.). Encuesta sobre el uso de Tecnologías de la Información y las Comunicaciones (TIC) y del Comercio Electrónico en la empresa. Recuperado 12 abril, 2019, de [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C)

<sup>34</sup> Ministerio de Empleo y Seguridad Social (MEySS). (2018, enero). Cifras Pyme. Recuperado 30 abril, 2019, de <http://www.ipyme.org/es-ES/ApWeb/EstadisticasPYME/Documents/CifrasPYME-enero2018.pdf><https://www.euroforum.es/blog/ciberseguridad-y-big-data-son-compatibles/>

<sup>35</sup> Telefónica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

Mejora de la propensión a la compra de un producto.

2. Eficiencia operativa: reducción de costes e incremento de eficiencias.

Crea o mejora modelos de predicción de averías.

Optimiza la eficiencia de los equipos de mantenimiento.

Mejora de los acuerdos de niveles de servicio (ANS o SLAs) con clientes.

Optimiza de rutas logísticas, ubicación de tiendas, localización de antenas (reducción de CAPEX), gasolineras...

3. Prevención y predicción de riesgos

Reduce de la fuga de clientes (*churn rate*)

Refina el scoring de riesgo de impago.

Detecta situaciones de descuentos anómalos en áreas de compras.

### **5.3 Transformación de una empresa hacia BIG DATA**

La adopción del Big Data en las empresas es progresiva, dada la complejidad y novedad que supone su incorporación en los procesos habituales. De todas las “V” características, es el valor de los datos y su potencial explotación la que está en la base del interés por el Big Data. Parece obvio, además, que según las empresas van avanzando a convertirse en compañías data-driven, aumenta precisamente la generación de valor basada en los datos.

El peso del dato en las compañías varía según la industria, hay ecosistemas donde el dato tiene, ya de entrada, un gran protagonismo, motivado por cercanía tecnológica, disponibilidad de datos diferenciales, o sencillamente, necesidad de reinventarse. En otros entornos el dato no es estratégico, sino táctico.

### **5.3.1 Tipologías de utilización del dato**

En orden incremental de impacto:

#### **MEJORA DE PROCESOS**

El dato obtenido se utiliza para determinar los problemas o ineficiencias que tiene un proceso, y de manera iterativa, se mejora el mismo.

#### **MEJORA DE LA CALIDAD DEL PRODUCTO**

El dato obtenido se utiliza para determinar los problemas o ineficiencias que tiene un producto, y de manera iterativa, se realizan los cambios en el mismo.

#### **NUEVO PRODUCTO**

Los datos abren la oportunidad de un producto o línea de producto que la compañía no tenía previamente, diversificando el portfolio.

#### **NUEVO MODELO DE NEGOCIO**

La compañía cambia su posición en la cadena de valor. Pasa a tener un modelo diferente, centrado en el dato. El dato pasa a ser un activo fundamental en la compañía.

Una empresa de pallets, cuya actividad se basaba en alquilar pallets a otras empresas que los utilizaban para mover sus mercancías, decidió apostar por el Big Data para tratar de averiguar dónde y qué clientes perdían sus pallets. Se colocaron sensores en el 5% de los pallets, solo en España había 300.000 pallets sensorizados que enviaban su posición cada 30 minutos. Asumiendo que el 50% de los pallets se movían 8 horas al día, se obtenían 2,4 millones de posiciones por día. Aplicando distintos modelos a los datos obtenidos la empresa logró:

- Mejora de procesos: detectó aquellas rutas y áreas donde los pallets se extraviaban, evitando las fugas de pallets y el daño que estas producían a la marca.
- Mejora de la calidad: capacidad para garantizar la disponibilidad de pallets en ubicaciones específicas y para estimar como los distintos tipos de entregas afectan a la vida útil de los pallets.

- Nuevo producto: creación de un servicio premium para clientes específicos que requieren monitoreo en tiempo real o casi en tiempo real de sus mercancías.
- Nuevo modelo de negocio: sobre la base de datos históricos de los movimientos de bienes, el PIB de los países se puede estimar con varios meses de anticipación. Esta información puede ser ofrecida al gobierno.<sup>36</sup>

### **5.3.2. Etapas del viaje empresarial hacia el Big Data**

Cada empresa aborda de distinta manera y velocidad la adopción del Big Data en base a numerosos factores, entre otros, aspectos económicos y de infraestructuras, disponibilidad de recursos humanos o compromiso de la dirección.

Se explican a continuación, de forma general, las principales fases de este proceso que concluye en la transformación de la organización en una compañía data- driven.

Una compañía data-driven es aquella que provisiona, procesa y disponibiliza datos veraces y a tiempo para ganar efectividad, visión y competitividad.

Estas organizaciones se caracterizan por una cultura colaborativa, abierta, democrática e inquieta. La toma de decisiones está basada en hechos, disponer de datos es imprescindible para cualquier iniciativa y en cualquier proyecto se testean todos los datos. Existe un uso entendido de los datos, los usuarios son autónomos para conocer, acceder y cruzar los datos, las opiniones y las discusiones se basan siempre en ellos y las herramientas de análisis y preparación son siempre colaborativas. Además de un equipo analítico existe un CDO (Chieff Data Officer), dependiente del CEO, que garantiza la calidad y gestión de los datos. La tecnología es escalable, flexible, móvil e híbrida; las nuevas tecnologías no son nuevas barreras.

---

<sup>36</sup> Basado en el material del Diploma de Fundamentos en Business Analytics

## Fases en la transformación de la empresa



### 1. STATUS QUO

Hay una total ignorancia sobre qué es y qué puede aportar Big Data. Pocas fuentes de datos reconocidas. No hay datos en formato accionable. No existen métricas para la toma de decisiones de negocio.

### 2. EXPLORACION

La compañía está interesada en usar Big Data en su negocio, pero las fuentes de datos y su calidad son casi totalmente desconocidas. Además, no cuentan con infraestructura, capacidades, conocimientos ni procesos para usarla. Necesitan hacer pruebas de concepto (PoC), con soporte de consultoras, para validar la utilidad de esta nueva tecnología y venderla internamente.

### 3. TRANSFORMACION

La organización decide invertir en Big Data. Se necesita tecnología e infraestructura. Esta es una de las fases más complejas y normalmente requiere el apoyo de una consultora experta para generar capacidades y apoyar la transformación. Es decir, existe la necesidad de un programa de transformación tecnológica, cultural y de negocio.

### 4. DATA-DRIVEN

Gestión del cliente basada en Big Data. La predicción va sobre la reacción, es la nueva cultura. Se necesita enriquecer la información con fuentes adicionales para obtener más valor de las decisiones. Existe un modelo implementado para la generación de insights de negocio centradas en los datos. Surgen nuevos modelos de negocio. Existe una necesidad de soporte consultivo para la integración de datos externos.<sup>37</sup>

---

<sup>37</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

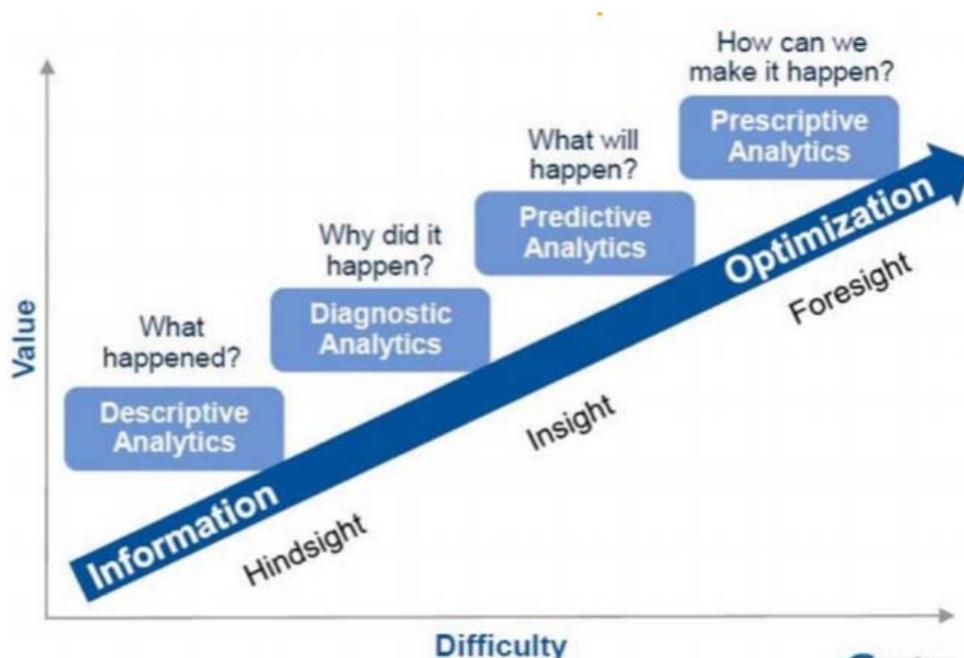
### 5.3.3 Niveles de valor de Data Analytics

En paralelo, cabe hablar de los distintos niveles de utilización del dato dentro de las organizaciones.

El dato por si solo no tiene valor si no se canaliza hacia acciones y destinatarios concretos: toma de decisiones, aprovechamiento del conocimiento de los clientes, proveedores, productos, usuarios, stakeholders, procesos, operaciones, etc.

El objetivo de los negocios basados en el dato es el de conseguir un nuevo producto, proceso o posicionamiento a partir de los datos accesibles por la empresa, a través de un procesado de esta información. Estas oportunidades no son evidentes.

La consultora Gartner distingue entre 4 niveles incrementales de valor y dificultad del análisis de datos dentro de empresas: <sup>38</sup>



Fuente: Gartner Analytic Ascendancy Model (Gartner, March 2012)

#### DESCRIPTIVE ANALYTICS

Los análisis que resumen los datos sin procesar y los convierten en algo que se puede interpretar por los humanos. Describen el pasado. El pasado se

<sup>38</sup> Gartner. (2018). *Analytics*. Retrieved from Gartner: <https://www.gartner.com/it-glossary/analytics/>

refiere a cualquier punto del tiempo en el que un evento haya ocurrido, ya sea hace un minuto o hace un año. Es decir, responde a la pregunta ¿Qué ha pasado?

La gran mayoría de las estadísticas que utilizamos se encuentran en esta categoría: aritmética básica como sumas, promedios, cambios porcentuales. Por lo general, los datos subyacentes son un recuento o agregado de una columna de datos filtrada a la que se aplican las matemáticas básicas.

Las estadísticas descriptivas son útiles para mostrar cosas como el stock total en inventario, el promedio de dólares gastados por cliente y el cambio anual en las ventas.

Los ejemplos comunes de análisis descriptivos son informes que brindan información histórica con respecto a la producción, finanzas, operaciones, ventas, finanzas, inventario y clientes de la compañía.

### DIAGNOSTIC ANALYTICS

Utilizado para descubrir o para determinar por qué sucedió algo. Es decir, responde a la pregunta ¿Por qué ha pasado?

El objetivo es detectar correlaciones para detectar relaciones de causalidad entre magnitudes. No se debe mezclar correlación y causalidad.

Para una campaña de marketing en redes sociales, se puede utilizar este análisis para evaluar el número de publicaciones, menciones, seguidores, admiradores, vistas de página, reseñas, etc. Puede haber miles de menciones en línea que se pueden resumir en una sola vista para ver lo que funcionó en campañas pasadas y lo que no.

### PREDICTIVE ANALYTICS

El análisis predictivo tiene sus raíces en la capacidad de "predecir" lo que podría suceder. Estas analíticas son sobre la comprensión del futuro. El análisis predictivo proporciona estimaciones sobre la probabilidad de un resultado futuro. Combinan los datos históricos encontrados en los sistemas ERP, CRM, HR y POS para identificar patrones en los datos y aplicar modelos estadísticos y

algoritmos para capturar relaciones entre varios conjuntos de datos. Responde a la pregunta ¿Qué va a ocurrir?

El análisis predictivo se puede utilizar en toda la organización, desde la previsión del comportamiento del cliente y los patrones de compra hasta la identificación de tendencias en las actividades de ventas. También ayudan a pronosticar la demanda de insumos de la cadena de suministro, las operaciones y el inventario.

Una aplicación común con la que la mayoría de las personas está familiarizada es el uso de análisis predictivo para crear un credit score. Estas puntuaciones son utilizadas por los servicios financieros para determinar la probabilidad de que los clientes realicen pagos de crédito futuros a tiempo.

El concepto de un modelo que describe el comportamiento de la realidad o un fenómeno aparece en este nivel de análisis.

### PRESCRIPTIVE ANALYTICS

Da una serie de diferentes acciones posibles y guía hacia una solución. El objetivo de estos análisis es proporcionar asesoramiento. Tratan de cuantificar el efecto de las decisiones futuras para asesorar sobre los posibles resultados antes de que se tomen las decisiones. Estos análisis van más allá del análisis descriptivo y predictivo al recomendar uno o más cursos de acción posibles. Esencialmente analizan múltiples futuros. Responden a la pregunta ¿Cómo puedo hacer que ocurra?

Estos análisis requieren una definición exhaustiva de los KPI a optimizar. Las grandes empresas están utilizando con éxito el análisis prescriptivo para optimizar la producción, la programación y el inventario en la cadena de suministro para asegurarse de que están entregando los productos correctos en el momento adecuado y optimizando la experiencia del cliente.

El siguiente gráfico elaborado por la compañía McKinsey & Company ilustra gráficamente cómo los distintos niveles de análisis de datos aportan una ventaja competitiva mayor cuanto mayor es el esfuerzo realizado por el equipo analítico.

La gestión de los datos, es decir, los datos en crudo, los datos limpios y el reporting (informes y consultas ad hoc), requieren de poco esfuerzo y apenas aportan una ventaja competitiva.

La analítica descriptiva, es decir, el filtrado de datos, las alertas, las técnicas de clustering, el pronóstico de tendencias y los análisis estadísticos, son proyectos laboriosos que aportan una mayor ventaja competitiva.

La analítica avanzada, es decir, el análisis predictivo, los modelados de simulación y el análisis prescriptivo, son proyectos muy difíciles y laboriosos que aportan una ventaja competitiva excepcional.



Fuente: McKinsey & Company

## **6. RETOS Y RIESGOS EN LA IMPLANTACION DE BIG DATA EN LA EMPRESA**

Como se ha comentado anteriormente, solo el 15% de las empresas consigue llevar a producción su proyecto de Big Data, incluso contando con la esponsorización de la alta dirección (Gartner, octubre 2016).

Un proceso de adopción como el que requiere la implantación de Big Data en la empresa estresa los recursos de la misma desde muchos puntos de vista. No son solo los recursos económicos y técnicos los que se ponen a prueba sino también los recursos humanos y organizativos.

Y todo ello no es exento de riesgos. Se detallan a continuación los más comunes, así como la vía para encontrar posibles soluciones.

### **6.1 Retos y riesgos inherentes a la naturaleza del Big Data**

Como ya se ha comentado, Big Data supone un reto relevante en cuanto a la disposición, estructura y almacenamiento de datos. La lentitud en la disponibilización de datos, que pueden en ocasiones encontrarse en distintos silos en la organización, resultando difícil y costoso disponibilizarlos a tiempo para poder ser utilizados.

Por otra parte, la calidad de los mismos, la ciberseguridad y el cumplimiento de la legislación vigente son aspectos que ninguna empresa puede descuidar.<sup>39</sup>

#### **6.1.1 Asegurar la calidad del dato**

La calidad de los datos afecta de múltiples maneras al negocio de la empresa<sup>40</sup>. Los datos de mala calidad implican:

---

<sup>39</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

<sup>40</sup> Samaniego, J. (2018b, 18 mayo). Big Data: cómo se pasa de cantidades ingentes de datos a información valiosa que ayude en la toma de decisiones. Recuperado 13 abril, 2019, de

<https://hablemosdeempresas.com/grandes-empresas/big-data-smart-dat>

- Procesos de negocio ineficientes por información duplicada o errónea.
- Incapacidad de detectar los mejores clientes y proveedores.
- Imposibilidad de tomar decisiones de negocio sólidas.
- Iniciativas de negocio y marketing fallidas.
- Menor satisfacción del cliente

Mientras que los datos de buena calidad:

- Reducen gastos en distintos departamentos.
- Optimizan la cadena de suministros y reducen el coste de materiales.
- Suponen mejores decisiones de negocio.
- Ayudan a tener una visión consistente y global de la empresa.
- Detectan mejores clientes.
- Incrementan valor de los datos como activo.
- Mejoran la imagen corporativa

En muchas ocasiones no existe consciencia directa del impacto de la inadecuada calidad de los datos. Sin embargo, existen síntomas que nos pueden indicar esta situación:<sup>41</sup>

- Desconfianza en el proceso de toma de decisiones debido a la falta de confianza en los datos.
- Iniciativas de negocio que no alcanzan objetivos previstos. (CRMs, Campañas de marketing).
- No es posible conocer al cliente. La relación con el cliente se ve afectada: disminuye su satisfacción.
- Aumenta la complejidad para detectar fraudes, sobrepagos, etc. No se identifican duplicados, unidades familiares (households), relaciones entre empresas, etc.
- Problemas derivados del no cumplimiento normativo. La calidad de datos es uno de los pilares fundamentales de conformidad con las diferentes normativas y regulaciones (ej. GDPR).

---

<sup>41</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

Algunas cifras relativas a la calidad de los datos: <sup>42</sup>

El 94% de las empresas considera la calidad de los datos muy importante a la hora de realizar una estrategia comercial y de marketing.

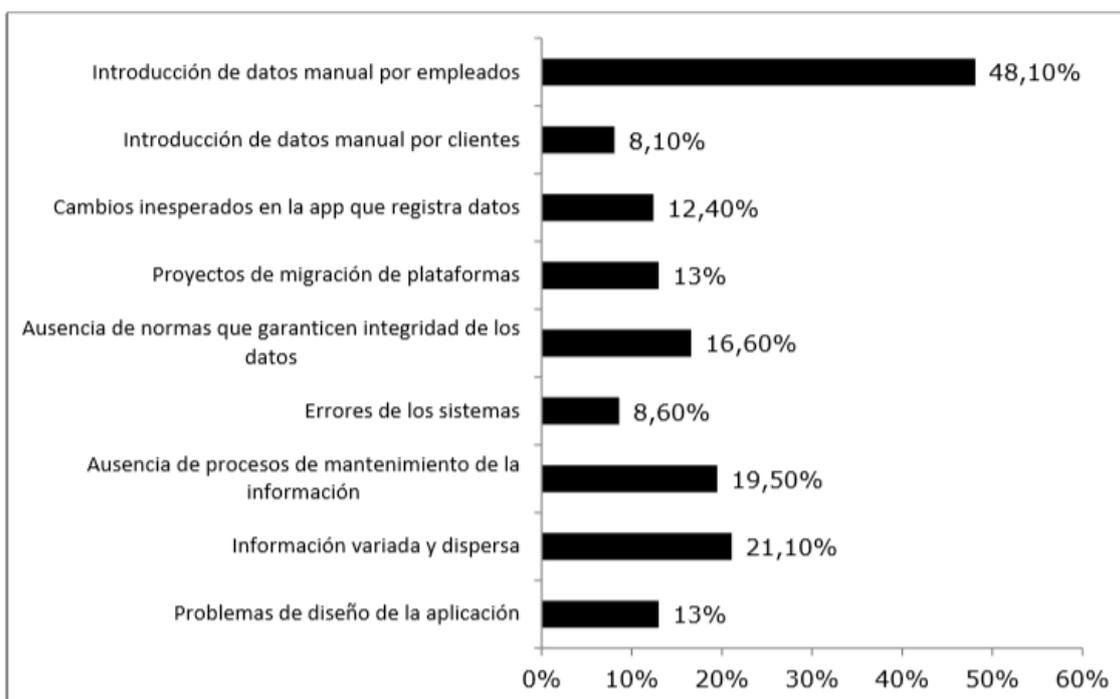
El 74% opina que la calidad de los datos afecta de manera notable a la toma de decisiones de la empresa.

El 10-20% de los ingresos totales es el coste estimado de una mala calidad de datos en la empresa.

El 84% de las empresas señala que una buena calidad de datos incrementaría los beneficios de su empresa.

El 70,8% de las empresas piensa que la falta de datos puede afectar al cumplimiento normativo (GDPR, Basilea...)

La siguiente figura muestra las principales causas por las que se cometen errores en los datos y estos son de mala calidad.



*Fuente: Telefónica Synergic Partners*

<sup>42</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

## Democratización del dato y su calidad

La calidad del dato no debe contemplarse como responsabilidad exclusiva de los analistas e ingenieros, sino que debe ser considerada una responsabilidad de toda la empresa.<sup>43</sup>

Con la finalidad de democratizar el dato en una compañía es necesario trabajar en los ejes personas, procesos y tecnologías de una organización para cambiar la forma en que los datos son adquiridos, gestionados, mantenidos y transformados en información.



*Fuente: Telefónica Synergic Partners*

Para ello es clave que la organización disponga de un sistema de Gobernanza o Gobierno de Datos:

“Sistema que define las responsabilidades y deberes de cualquier proceso relacionado con los datos, en base a unas políticas existentes, las cuales describen quién puede hacer qué, saber qué datos y en qué circunstancias”. (The Data Governance Institute).

“Orquestación formal de procesos, personas y tecnología para permitir que una organización convierta sus datos en un archivo estratégico.” (The MDM Institute).

<sup>43</sup> IT-NOVA. (s.f.). La democratización de los datos. Recuperado 11 marzo, 2019, de <https://it-nova.co/es/es/la-democratizacion-de-los-datos/>

“Estructura Organizativa que crea y promueve Políticas y Procedimientos de datos para uso por negocio y por tecnología, a través de toda la Organización” (TDWI).

Los propósitos que persigue el Gobierno de Datos (Data Governance)<sup>44</sup> son:

- Asegurar que los datos son siempre fiables y válidos en cada contexto empresarial.
- Mantener su calidad a lo largo del tiempo.
- Asegurar que existen mecanismos de control sobre quién puede hacer qué con los datos en cada momento.
- Garantizar la seguridad, privacidad y custodia de datos.

Todo ello con el objetivo de apalancar los datos como un activo corporativo de gran valor empresarial.

Existen en el mercado herramientas informáticas que ayudan a la optimización y consecución de mejores resultados con el Data Governance.

Por último, se necesitan implementar auditorias recurrentes para dar visibilidad y responsabilidad en cada paso del proceso, que es lo que permite a la TI "gobernar" los datos en lugar de simplemente establecer políticas y controles de acceso esperando simplemente que las cosas vayan bien.

### **6.1.2 Asegurar el cumplimiento del marco legal (GDPR)<sup>45</sup>**

El GDPR tiene un Impacto significativo para las organizaciones y su forma de manejar los datos, con sanciones potencialmente muy grandes para aquellas empresas que sufran una violación, pudiendo llegar a suponer un 4% de los ingresos globales, o 20 millones de euros (lo que sea mayor). Esta es la multa máxima que puede imponerse por las infracciones más graves.

El artículo 12 de la GDPR (Ver punto 4) requiere que se comuniquen información sobre el procesamiento de datos personales de manera que sea:

---

<sup>44</sup> Grupo PowerData. (s.f.-d). Desmitificando el Data Governance: Qué, cuándo, dónde y por qué. Recuperado 5 mayo, 2019, de <https://www.powerdata.es/data-governance>

<sup>45</sup> Grupo PowerData. (s.f.-e). GDPR: Lo que debes saber sobre el reglamento general de protección de datos. Recuperado 5 mayo, 2019, de <https://www.powerdata.es/gdpr-proteccion-datos>

concisa, transparente, en un lenguaje claro y sencillo, inteligible, de fácil acceso y gratuita. Las compañías necesitan cubrir los siguientes puntos de información en sus comunicaciones:

1. Indicar quién es el controlador de datos a cargo de decidir qué datos personales se recogen.

2. Información de contacto para el controlador de datos.

3. Indicar si se utiliza datos personales para tomar decisiones automatizadas, como puede ser la puntuación crediticia

4. Informar a los usuarios de los 8 derechos bajo el GDPR: derecho a ser informado, derecho de acceso, derecho de rectificación, derecho de borrado, derecho a restringir el procesamiento de datos, derecho a la portabilidad de los datos, derecho a objetar y derechos de decisión automática y perfilado.

5. Indicar si es obligatorio el suministro de datos personales.

6. Indicar si se transfieren datos a nivel internacional.

7. Base jurídica para el tratamiento de los datos.<sup>46</sup>

Un ejemplo práctico de la necesidad de transparencia a la hora de usar datos de los individuos y el contragolpe que puede tener lugar si una empresa no es lo suficientemente transparente es el caso de la red social LinkedIn. En junio de 2015, la empresa acordó pagar 13 millones de dólares para satisfacer una demanda colectiva por enviar invitaciones de correo electrónico a las listas de contactos de los usuarios. Es importante ser claro con los clientes, con los datos que se recopilan y como se pretenden usar, ahora LinkedIn permite seleccionar que contactos recibirán invitaciones automáticas y correos de seguimiento.<sup>47</sup>

---

<sup>46</sup> EU. (2018, 5 octubre). General Data Protection Regulation (GDPR). Recuperado 7 marzo, 2019, de <https://eugdpr.org/the-regulation/>

<sup>47</sup> Marr, B. (2015). *Big Data en la práctica, cómo 45 empresas exitosas han utilizado análisis de Big Data para ofrecer resultados extraordinarios*. Madrid, España: Teell Editorial, S.L..

Con el fin de conseguir un adecuado tratamiento y cumplimiento del GDPR, conviene implementar un proyecto de Master Data Management de acuerdo a las directrices del Gobierno de Datos.

Se trataría de asegurar que los datos personales se almacenan en una única fuente confiable de datos, evitando duplicidades y asegurando que los procesos sean correctos y actualizados. Se establecería así una representación exacta y completa de cada cliente, independiente de canales de entrada, departamentos y líneas de negocio.

Asimismo, en este ámbito prácticamente infinito incluso con el apoyo de la tecnología, es importante analizar el riesgo legal de las actividades de procesamiento de datos, identificando aquellas de mayor riesgo y que necesitaran de una mayor supervisión y control.

También es esencial conocer en detalle los derechos de los clientes y racionalizar las solicitudes de acceso a sus datos con procedimientos eficientes para solicitudes de “rectificación y borrado” así como restricciones de procesamiento

Por último, el preciso saber cuándo notificar una infracción a la autoridad de protección de datos (no todas las infracciones lo requieren) y revisar los procedimientos para asegura el poder cumplir los plazos estrictos que suele exigir dicha autoridad (unas 72 horas).

## **6.2 Retos y riesgos de tipo organizativo**

Las operaciones analíticas requieren en muchos casos de la contratación de nuevos empleados, personas adecuadas con las habilidades adecuadas. Según una encuesta reciente publicada por la empresa Gartner, más de la mitad de los negocios consideran que el principal obstáculo para llevar a cabo un análisis de Big Data es la dificultad de encontrar los perfiles profesionales adecuados. En España, el 65% de los profesionales admite no estar familiarizados con tecnologías como Big Data, ML O BI.<sup>48</sup>

---

<sup>48</sup> IT Digital Media Group. (2018, 25 abril). El 65% de los profesionales desconoce tecnologías como Big Data o IoT. Recuperado 4 mayo, 2019, de <https://tecnologiaparatuempresa.ituser.es/estrategias/2018/04/el-65-de-los-profesionales-desconoce-tecnologias-como-big-data-o-iot>

La contratación de nuevos empleados suele ser vista con recelo por los actuales, sobre todo en el área de IT. Sin embargo, resulta fundamental asegurarse la colaboración de todos ellos puesto que sólo así, con su involucración, pueden comprender la razón de la utilización de la analítica de datos y la empresa conseguir la máxima rentabilidad del Big Data.

Para una empresa que quiera transformarse hacia Big Data existe una inherente necesidad de perfiles con una alta especialización para trabajar en los diferentes puntos de la cadena de valor. Se considera que, por vía de mínimos, son necesarios los siguientes perfiles: <sup>49</sup>

### BUSINESS ANALYST

Perfil encargado de generar la visión del negocio desde la información obtenida en proyectos Big Data.

Entre sus roles destacan la identificación de necesidad de negocio, el análisis de la información y la elaboración de insights aplicados al negocio.

El *business analyst* debe contar con habilidades estadísticas y matemáticas, conocimientos de negocio y de programación general.

Entre las herramientas más usadas por los *business analyst* destaca el análisis DAFO, que profundiza en las debilidades, las amenazas, las oportunidades y las fortalezas de la compañía en cuestión.

### DATA SCIENTIST

Perfil encargado de la extracción de valor y conocimiento de los datos.

Entre los roles del *data scientist* destacan el análisis de la información, la elaboración de modelos predictivos y el reporting de resultados.

Los *data scientist* deben tener habilidades estadísticas y matemáticas, y altos conocimientos sobre programación y entornos de desarrollo.

---

<sup>49</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

Cuentan con herramientas como Hadoop (un “ecosistema” donde distintos componentes colaboran para almacenar y procesar Big Data), hive y lenguajes de programación.

### BIG DATA ENGINEER

Perfil responsable del flujo de información de extremo a extremo desde un punto de vista técnico.

Entre los roles del *big data engineer* destacan la responsabilidad sobre el flujo del dato de extremo a extremo, la integración de las distintas fuentes de información y el desarrollo de los procesos de transferencia.

Los *big data engineer* cuentan con habilidades de desarrollo de ETLs, el desarrollo de software, la programación y el control de herramientas de ingesta.

Cuentan con herramientas como Hadoop, el lenguaje de programación Python y diversos productos ofrecidos por la compañía IBM.

### BIG DATA ARCHITECT

Perfil encargado del diseño, despliegue y gestión de la arquitectura Big Data.

Entre los roles del *big data architect* se encuentran el diseño de la arquitectura, el despliegue de los componentes y la resolución de incidencias en la plataforma.

Los *big data architect* cuentan con habilidades de gestión de bases de datos, ecosistemas Hadoop, lenguajes de programación y sistemas Unix.

Se apoyan en herramientas como Hadoop, diversos lenguajes de programación y los sistemas Unix.

### DATA VISUALIZATION EXPERT

Perfil responsable del diseño y creación de visualización de datos de gran impacto.

Entre los roles del *data visualization expert* destacan la presentación de los datos, la supervisión de las herramientas de visualización y el apoyo a el analista o los analistas.

Los *data visualization expert* cuentan con importantes habilidades en técnicas de visualización, lenguajes de programación y herramientas de visualización avanzadas.

Algunas de las herramientas más populares con las que cuentan son QlikView, D3plus o Tableau.

Para facilitar el proceso de conseguir talento muchas empresas acuden a donde se encuentra dicho talento, creando colaboraciones con universidades, o acudiendo al mercado de Big Data.

### **El mercado de Big Data**

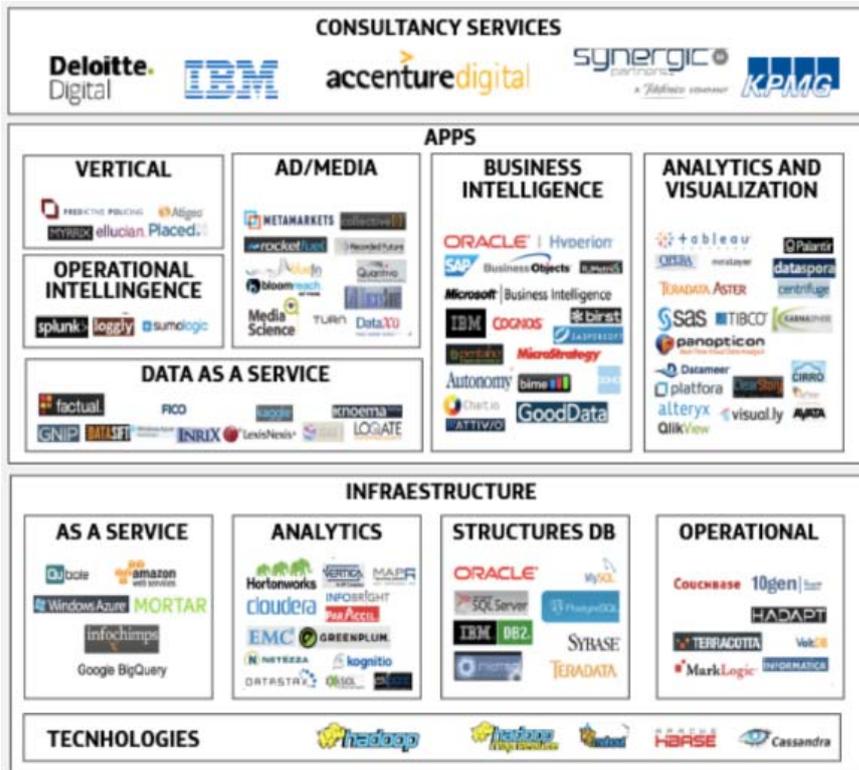
El mercado del Big data es incipiente, grande, de rápido crecimiento (crece 6 veces más rápido que el mercado de IT tradicional), atractivo y fragmentado.

Se compone de gran número de tecnologías que abarcan las distintas funcionalidades dentro una arquitectura Big Data cuyo crecimiento ha sido exponencial en los últimos años. Una división del mercado podría ser: tecnologías, infraestructura, apps y servicios profesionales de consultoría. Cada día entran nuevas compañías en el mercado y se transforma en un mercado fragmentado sin lideres claramente definidos. Según Wikibon la distribución por categoría en el año 2021 será: 30% servicios profesionales, 40% software y 30% hardware.

La analítica de datos, en especial el Big Data, es muy difícil, ya que implica trabajar con grandes y complejos conjuntos de datos en continuo cambio. Muchas compañías adquieren valiosas aplicaciones de analítica de datos y necesitan ayuda para sacarle provecho, ya que hay que crear algoritmos, una estructura hardware para almacenar los datos, ejecutar el análisis y anunciar los resultados. Esta enorme brecha entre lo que los negocios son capaces de concebir y lo que son capaces de construir ha potenciado la aparición de negocios que ofrecen “*data-as-a-service*” (DAAS). Microsoft lidera este mercado

con su producto *Analytics Platform System*, que se vende como una solución que ofrece “Big Data on the box” (listo para su utilización).<sup>50</sup>

La imagen muestra algunos de los principales *players* del mercado del Big Data.<sup>51</sup>



Fuente: Telefónica Synergic Partners

Otro obstáculo que vencer, especialmente al inicio, es conseguir la colaboración de los empleados. Es fundamental lograr su involucración para que comprendan la razón por la que se utiliza la analítica de datos y, de esta manera, consigan sacar el máximo partido a las tendencias que se van descubriendo y ayudar a construir una organización data-driven.

<sup>50</sup> Marr, B. (2015). *Big Data en la práctica, cómo 45 empresas exitosas han utilizado análisis de Big Data para ofrecer resultados extraordinarios*. Madrid, España: Teell Editorial, S.L..

<sup>51</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

## **7. MODELOS DE NEGOCIO BASADOS EN EL DATO**

Un modelo de negocio es la manera en la que una compañía articula las propuestas de valor que ofrece a sus clientes, es decir, el valor creado para los usuarios gracias a las tecnologías que se usan o los productos que se comercializan. Distinguimos entre tres opciones de modelo de negocio basados en el dato: <sup>52</sup>

### **ANALITICA COMO PRODUCTO**

Compañías que ofrecen modelos de procesado y servicios de consultoría para que sus clientes puedan extraer todo el valor de sus datos.

Utilizan tecnología propia o externa para procesar información y crear entornos de ejecución y modelos para procesar información.

### **DATO COMO PRODUCTO**

Compañías que ofrecen acceso y agregación de datos de diferentes tipologías y fuentes.

Utilizan datos internos y externos de uso y comportamiento de usuarios para lograr conjuntos de datos procesados a medida para sus clientes.

### **PRODUCTO BASADO EN EL DATO**

Compañías que ofrecen productos o procesos diferenciales para el cliente.

Utilizan datos internos y externos de uso del producto, así como datos de los usuarios para lograr productos o procesos mejorados en tipología, calidad o novedad.

Para definir o cambiar el modelo de negocio y orientarlo hacia el dato hay que plantearse una serie de cuestiones estratégicas.

En primer lugar, hay que analizar los diferentes puntos de la operación donde se captura o requieren datos relevantes y, para cada uno de estos puntos, determinar una serie de preguntas estratégicas. El objetivo es evidenciar el

---

<sup>52</sup> Basado en el material del Diploma de Fundamentos en Business Analytics

margen de mejora u optimización a través de un procesado del dato que se maneja o impacta ese punto.

Se analiza la disponibilidad del dato para generar margen de mejora en los procesos actuales. A continuación, se analizan los datos generados y su posible monetización de manera directa o bien con un procesado intermedio para anonimizar. En ocasiones, el cliente no es muy evidente.<sup>53</sup>

Por último, se evalúa la utilización del dato en todo el conjunto de la operativa de la compañía.

Adicionalmente, la consultora McKinsey Digital; McKinsey Strategy and Corporate Finance Practice identifica 10 preguntas que toda compañía debe hacerse y sobre las que debe reflexionar para lograr una estrategia exitosa en el mundo digital:<sup>54</sup>

- ¿Tu estrategia anticipa los efectos que la digitalización va a tener en tu ecosistema de negocio?
- ¿Tu estrategia proactivamente da forma al ecosistema para conseguir crecer y diferenciarse?
- ¿Tu estrategia construye un modelo de negocio digital desde cero, incluso si canibaliza ciertas partes de tu negocio actual?
- ¿Tu estrategia contempla oportunidades en otras industrias proyectando tu propuesta de valor en nuevos espacios?
- ¿Tu estrategia mantiene partes de tu negocio actual que no tienen potencial en un futuro digital?
- ¿Tu estrategia refleja la velocidad e incertidumbre de la era digital?
- ¿Tu estrategia reconoce explícitamente el impacto potencial de las nuevas tecnologías y sus aplicaciones?
- ¿Tu estrategia se asegura de que los mejores de tu equipo están totalmente comprometidos con la ejecución de la estrategia digital?

---

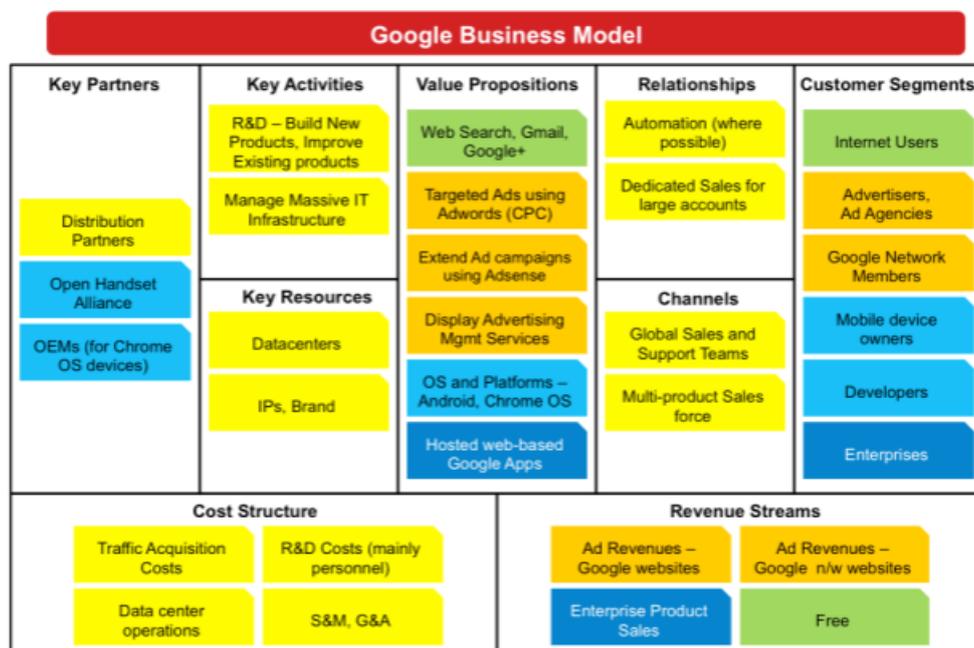
<sup>53</sup> Basado en el material del Diploma de Fundamentos en Business Analytics

<sup>54</sup> McKinsey & Company. (s.f.). Changing focus amid digital disruption. Recuperado 15 mayo, 2019, de <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/how-we-help-clients/changing-focus-amid-digital-disruption>

- ¿Tu estrategia se traduce en una relocalización de los recursos financieros y del tiempo de gestión en construir nuevas capacidades digitales?
- ¿Tu estrategia utiliza el horizonte temporal y las métricas de rendimiento para determinar si tu estrategia digital va a vencer al mercado?

Google es un gran responsable de la introducción del Big Data en nuestro día a día. Al utilizar Google como motor de búsqueda estamos trabajando con Big Data, ya que se estima que el índice de Google entre los que se buscan los resultados es aproximadamente de 100 petabytes, es decir 100 millones de gigabytes. Pero Google ha continuado expandiendo su negocio apoyándose en la sólida tecnología del Big Data que les hizo famosos, creando buscadores web, correo electrónico, sistemas operativos para teléfonos móviles y la red de publicidad en línea más grande del mundo.<sup>55</sup>

La siguiente imagen detalla el modelo de negocio de Google siguiendo el modelo Canvas.



Fuente: Material del Diploma de Fundamentos de Business Analytics

<sup>55</sup> Marr, B. (2015). *Big Data en la práctica, cómo 45 empresas exitosas han utilizado análisis de Big Data para ofrecer resultados extraordinarios*. Madrid, España: Teell Editorial, S.L..

Se puede decir que el reto más relevante de los modelos de negocio basados en el dato, una vez asumidos y resueltos los riesgos y retos en el viaje de adopción del Big data, es conseguir la monetización del dato, esto es, la disposición a pagar por él.

Este reto es habitual en las disrupciones o nuevas tecnologías que llegan al mercado. Las compañías necesitan ver casos de éxito claro y ROI basados en analítica.

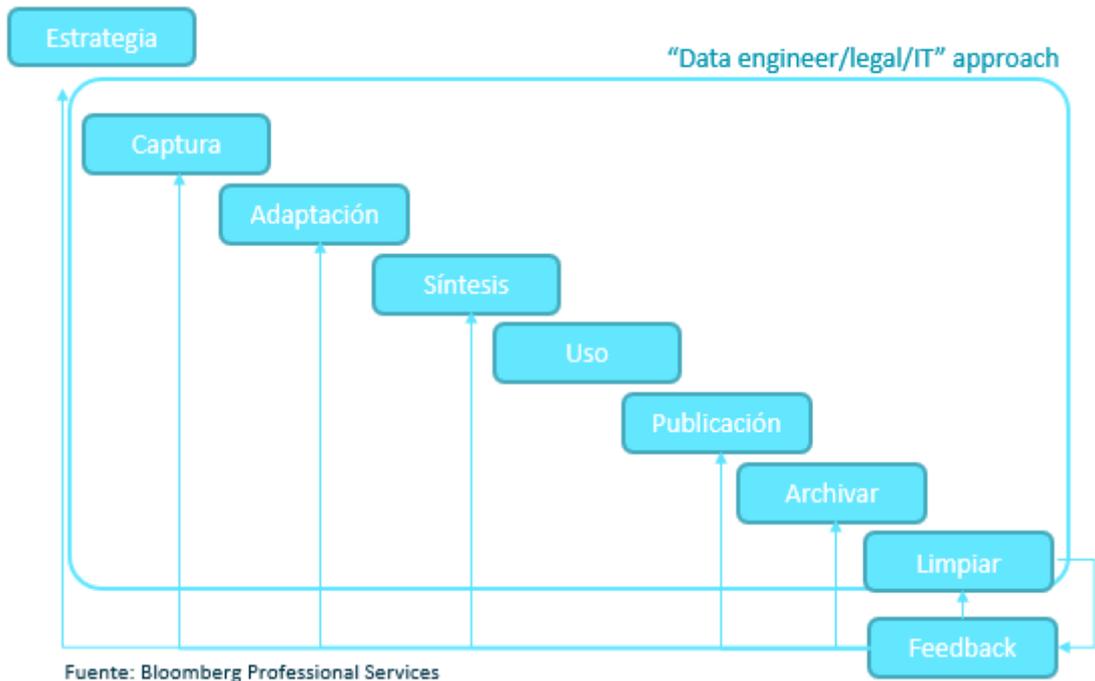
Se soluciona a mayor o menor velocidad en función de evidencias de mejora por parte de la competencia.<sup>56</sup>

---

<sup>56</sup> Basado en el material del Diploma en Fundamentos de Business Analytics

## 8. CICLO DEL BUSINESS DATA

Los datos pasan por distintas fases durante su viaje dentro de las empresas, basándonos en el modelo de los Servicios Profesionales de Bloomberg, distinguimos entre:



*Fuente: Bloomberg Professional Services*

### 1. ESTRATEGIA

La estrategia es “un plan de acción diseñado para lograr un objetivo a largo plazo o general”. No es ni la visión, no los objetivos, ni las tácticas.

“La estrategia sin tácticas es la ruta más lenta hacia la victoria. Las tácticas sin estrategia son el ruido antes de la derrota” Sun Tzu.

Para elaborar la estrategia es imprescindible hacerse determinadas preguntas:

- ¿Por qué estamos mirando los datos? Especificar si buscamos ofrecer analítica como producto, datos como producto o productos basados en datos.
- ¿Cómo estamos impactando en los negocios? Nuevos ingresos, nuevas geografías, nuevos productos, mayores eficiencias, menor tiempo de

comercialización, mejores ubicaciones y decisiones, mitigación de riesgos...

- ¿Cuáles son los tiempos hasta alcanzar el punto de equilibrio y los pasos para llegar hasta ahí?

Un paso muy importante para la empresa consiste en decidir cuáles son los datos necesarios, además del hecho de tener que recopilarlos.

## 2. CAPTURA

Los analistas necesitarán siempre los datos correctos, recogidos de la manera correcta, en el formato correcto, en el lugar correcto y en el momento adecuado. Si cualquiera de estos aspectos falta o desaparece, tendrán limitaciones en las preguntas que podrán contestar y en el tipo o calidad de las ideas que podrán derivar de los datos.

Debemos considerar todos los aspectos legales y políticos en cada parte de la captura de datos.

Adquisición de datos: los datos ya existen y se generan en algún lugar dentro o fuera de la empresa.

Entrada de datos: los datos aún no existen y se deben completar de forma manual o automática desde los dispositivos.

## 3. ADAPTACION

Los datos deben adaptarse en términos de modelo de datos o incluso en la plataforma en la que se almacenan: movimiento, integración, limpieza, enriquecimiento, captura de datos modificados, extracción de procesos de transformación y carga.

## 4. SINTESIS

Los datos procedentes de telemetría básica o avanzada y de soluciones hiper especializadas no suelen requerir almacenar los datos en bruto, sino almacenar una abstracción de alto nivel de los datos, después de una fase de preprocesamiento.

## 5. USO

Todos los datos deben proporcionar información procesable, o al menos mitigar una debilidad.

¡Incluso si no se toma una acción específica, el valor generado por el procesamiento de datos será positivo!

## 6. PUBLICACION

Fase opcional que se debe considerar si los datos se van a transferir a través de las subsidiarias de la compañía, o si se van a compartir con socios u otros terceros. Los aspectos legales deben de tenerse en cuenta antes de proceder a la publicación.

Si se incluye a terceras partes, típicamente se firmarán acuerdos de confidencialidad y otros acuerdos contractuales específicos apropiados a la situación. Esto típicamente impulsa una internalización de estos departamentos.

## 7. ARCHIVAR

Una vez que se utilizan los datos, estos necesitan ser almacenados. Esta fase no es tanto una fase técnica sino un punto de decisión entre valor y coste. En muchas situaciones, simplemente no sabemos cuánto tiempo tiene sentido almacenar algunos datos, ya que en el futuro podríamos descubrir que necesitamos retroceder en el tiempo para comprender algo.

## 8. LIMPIAR

En relación con la fase anterior, una vez que se han almacenado los datos, o una vez finalizado el período de almacenamiento, los datos deben ser depurados.

Se deben considerar los procesos y mecanismos para hacerlo sin bloquear bases de datos y sistemas.

## 9. FEEDBACK

Cada fase necesita ser monitoreada, o debemos plantearnos las siguientes cuestiones periódicamente:

- ¿Está funcionando correctamente?

- ¿Estamos obteniendo los resultados de manera oportuna?
- ¿Tenemos problemas con otros sistemas que consumen o interactúan con sistemas de procesamiento de datos? ¿Estos problemas se repiten a tiempo?
- ¿Son todas las fases rentables? ¿Se pueden optimizar?<sup>57</sup>

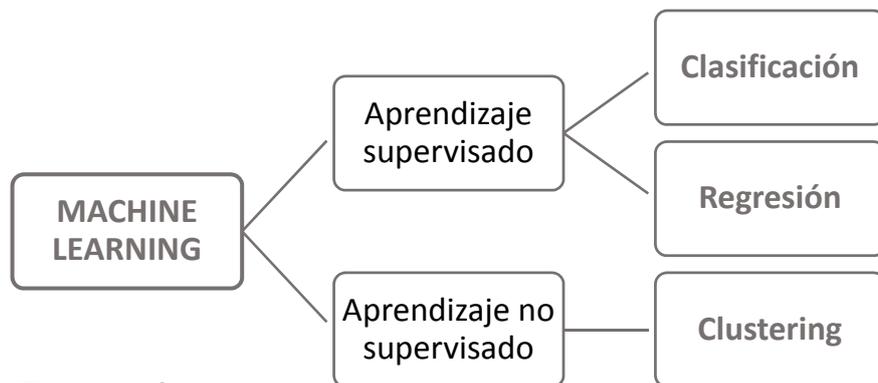
---

<sup>57</sup> Basado en el material del Diploma en Fundamentos de Business Analytics

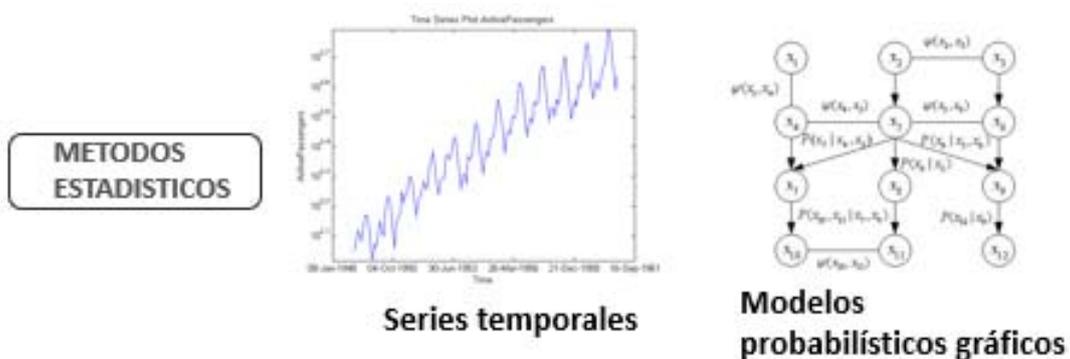
## 9. MODELOS Y ALGORITMOS

En Analytics, la predicción es la extracción de información de los conjuntos de datos existentes para determinar patrones y construir modelos para pronosticar resultados y tendencias. El proceso predictivo siempre implica incertidumbre.

Distinguimos entre métodos estadísticos y modelos creados con Machine Learning. Los métodos estadísticos son las series temporales y los modelos gráficos probabilísticos. El Machine Learning puede ser supervisado o no supervisado. El aprendizaje supervisado desarrolla un modelo predictivo basado en los datos de entrada y salida, distinguimos entre modelos de clasificación y modelos de regresión. El aprendizaje no supervisado agrupa e interpreta los datos basados solo en datos de entrada utilizando técnicas de clustering.



Fuente: Elaboración propia



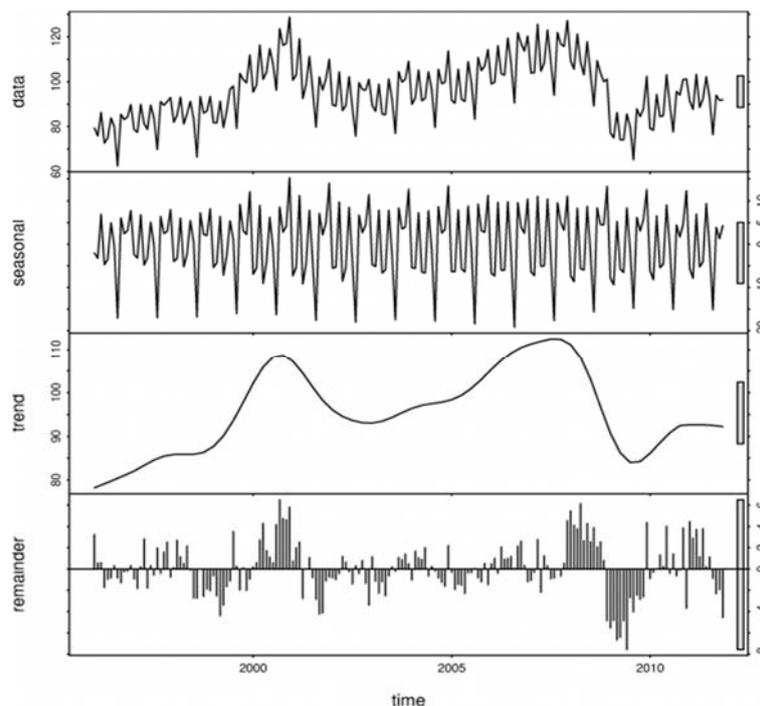
Fuente: Diploma Fundamentos de Business Analytics

## 9.1 Modelos Estadísticos

### Series temporales

El análisis y pronóstico de las series temporales es un campo activo de investigación. La serie histórica está descompuesta en componentes de tendencia, estacionales y aleatorios.

Las series temporales son muy útiles para pronosticar las ventas o la demanda de servicios públicos.



*Fuente: Diploma de Fundamentos de Business Analytics*

Si asumimos una descomposición aditiva, entonces podemos escribir:

$Y_t = S_t + T_t + R_t$ , donde  $Y_t$  son los datos,  $S_t$  es el componente estacional,  $T_t$  es el componente de tendencia y  $R_t$  es el componente restante, todo en el periodo  $t$ .

Alternativamente, una descomposición multiplicativa se escribiría como:

$$Y_t = S_t \times T_t \times R_t$$

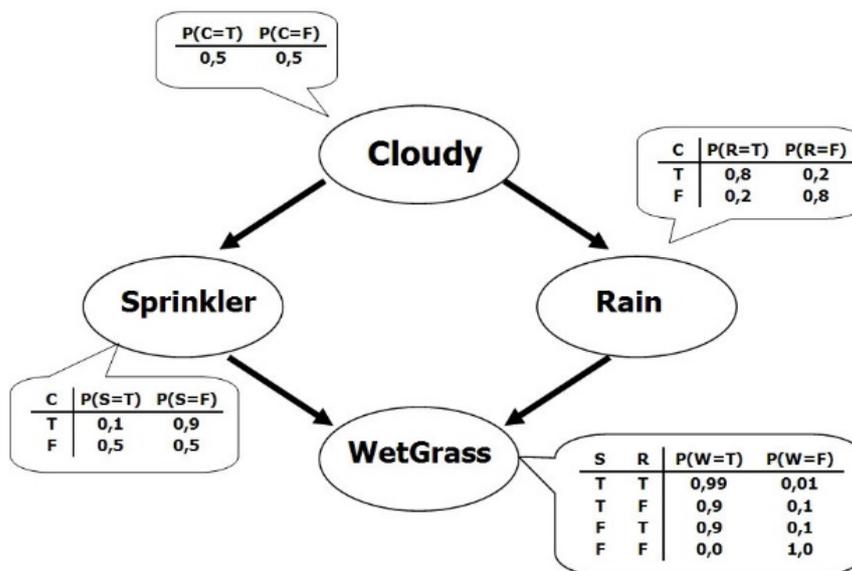
Las descomposiciones multiplicativas son comunes con las series temporales económicas.<sup>58</sup>

Utilizando la imagen superior como ejemplo, los tres componentes se muestran por separado en los tres paneles inferiores. Estos componentes se pueden agregar para reconstruir los datos que se muestran en el panel superior. Las barras grises a la derecha de cada panel muestran las escalas relativas de los componentes.

### Modelos gráficos probabilísticos

Los modelos gráficos probabilísticos son representaciones de dependencias entre variables aleatorias. Los valores pasados de probabilidad permiten calcular los resultados presentes.

Las redes bayesianas, definidas por Judea Pearl en 1985, son los modelos gráficos probabilísticos más comunes. Son útiles para predecir la probabilidad de fallo de los componentes de aviones y otras maquinarias y, por lo tanto, el personal de mantenimiento puede reemplazarlos antes de que fallen. Otro campo de aplicación es el diagnóstico clínico.<sup>59</sup>



Fuente: *Diploma de Fundamentos de Business Analytics*

<sup>58</sup> Hyndman, R. J., & Athanasopoulos, J. (2018, abril). *Forecasting: Principles and Practice*. Recuperado 12 marzo, 2019, de <https://otexts.com/fpp2/>

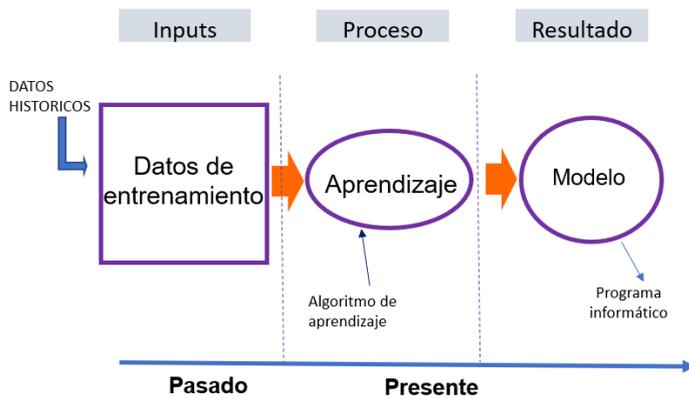
<sup>59</sup> Díez, F. J. (2014, 24 noviembre). *Introducción a los Modelos Gráficos Probabilistas*. Recuperado 15 abril, 2019, de <http://www.ia.uned.es/~fjdiez/libros/intro-mgp.pdf>

## 9.2 Modelos predictivos

Los modelos predictivos no funcionan por causalidad sino por correlación, ya que puede ser un predictor preciso.

Lo que se hace en Machine Learning es encontrar una fórmula matemática que, cuando se aplica a una colección de insumos que llamamos "datos de entrenamiento", produce las salidas deseadas.

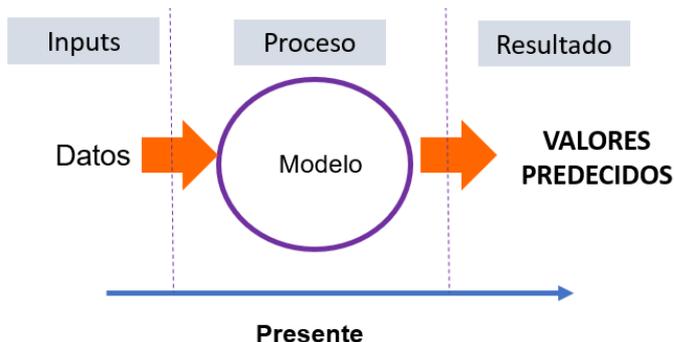
Construcción del modelo predictivo:



*Fuente: Diploma de Fundamentos de Business Analytics*

Esta fórmula matemática también genera los resultados correctos para otros insumos, distintos de los datos de entrenamiento, con la condición de que dichos insumos provengan de una distribución estadística similar o similar a la de la que se obtuvieron los datos de entrenamiento.<sup>60</sup>

Proceso de predicción:



*Fuente: Diploma de fundamentos de Business Analytics*

<sup>60</sup> Burkov, A. (s.f.). The Hundred-Page Machine Learning Book. Recuperado 15 abril, 2019, de <http://themlbook.com>

Utilizamos modelos predictivos cuando no se conoce la relación matemática entre las distintas variables, cuando existe un patrón y cuando tenemos un conjunto de datos lo suficientemente grande para aprender. Nuestro conjunto de datos históricos cuenta con diferentes características y conocemos el resultado que queremos aprender a predecir (etiqueta) para cada caso. Nuestro objetivo es predecir la etiqueta de nuevos casos con el mínimo error posible. Si el resultado es una predicción categórica se llama clasificación. Si el resultado es numérico, entonces se llama regresión.<sup>61</sup>

Se utiliza entre un 70% y un 80% de los datos del conjunto de datos inicial para entrenar el modelo; y entre un 20% y un 30% de los datos para probar el modelo y evaluar como de bueno es.

Los modelos predictivos son una tarea matemática y computacional difícil. Un problema de predicción del mundo real puede comenzar con cientos de características y se deben encontrar los hiperplanos óptimos. La definición de óptimo depende de la estrategia que elija, la precisión del modelo solo se puede evaluar después de hacer muchas predicciones nuevas. Un buen modelo requiere un equilibrio de complejidad y flexibilidad. Si el modelo es demasiado simple, las predicciones fallarán, pero si lo hacemos demasiado complejo, estamos sobre ajustando.

Para evaluar la calidad de los modelos existen una serie de indicadores:  
62

La precisión o *accuracy* es la proporción entre las predicciones correctas que ha hecho el modelo y el total de predicciones. resulta práctico por su facilidad de cálculo, otras veces es necesario profundizar un poco más y tener en cuenta los tipos de predicciones correctas e incorrectas que realiza el clasificador.

La matriz de confusión de un problema de *clases* es una *matriz nxn* en la que las filas se nombran según las clases reales y las columnas, según las clases previstas por el modelo. Sirve para mostrar de forma explícita cuándo una clase

---

<sup>61</sup> Basado en el material del diploma en Fundamentos de Business Analytics

<sup>62</sup> Burkov, A. (s.f.). The Hundred-Page Machine Learning Book. Recuperado 15 abril, 2019, de <http://themlbook.com>

es confundida con otra. Por eso, permite trabajar de forma separada con distintos tipos de error.

	N (modelo)	S (modelo)
n (real)	Negativos Reales	Falsos Positivos
p (real)	Falsos Negativos	Positivos reales

*Fuente: The Hundred-Page Machine Learning Book*

De esta forma, la diagonal principal contiene la suma de todas las predicciones correctas. La otra diagonal refleja los errores del clasificador: los falsos positivos y los falsos negativos.

El problema radica en que al medir la precisión del algoritmo de esa forma no distinguimos entre los errores de tipo falso positivo y falso negativo, como si ambos tuvieran la misma importancia, por ejemplo:

Falso positivo o “Error tipo I” El paciente no tiene cáncer, pero el algoritmo ha diagnosticado que sí lo padece. Se realizarán pruebas diagnósticas adicionales que acabarán descartando el diagnóstico. Tendrá un coste económico y un impacto emocional sobre el paciente, pero no se traducirá en riesgo vital.

Falso negativo o “Error tipo II”: El paciente sí tiene cáncer, pero el algoritmo predice que no. Este error del algoritmo se traduce en una falta de detección temprana de la enfermedad. El paciente no recibirá tratamiento a tiempo y esto, indudablemente reduce sus posibilidades de superar la enfermedad.

Dada la importancia que tiene discriminar en cada caso concreto los distintos tipos de error que pueden resultar de la aplicación de algoritmo, entendemos mejor la necesidad de trabajar con diferentes métricas:

La Exactitud se refiere a lo cerca que está el resultado de una medición del valor verdadero. Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

La sensibilidad, (o Tasa de Verdaderos Positivos), y la especificidad (o Tasa de Verdaderos Negativos) son dos valores que nos indican la capacidad de nuestro estimador para discriminar los casos positivos, de los negativos. La sensibilidad es la fracción de verdaderos positivos, mientras que la especificidad, es la fracción de verdaderos negativos.

La conveniencia de usar una métrica otra como medida del estimador dependerá de cada caso en particular y, en concreto, del “coste” asociado a cada error de clasificación del algoritmo.

### **9.3 Algoritmos de aprendizaje**

Los algoritmos son fórmulas matemáticas y procesos estadísticos utilizados para analizar datos.

Actualmente, hay algoritmos capaces de identificar a las personas que aparecen en una foto y buscar en internet otras fotos distintas de esas mismas personas. Otros algoritmos comprenden las palabras cuando hablamos, las transforman en texto escrito y analizan contenido significado y emociones, es decir, saben identificar si decimos cosas buenas o malas.<sup>63</sup>

Cada día aparecen nuevos algoritmos avanzados que ayudan a comprender el mundo y predecir el futuro. La habilidad de los algoritmos de aprender (ML) y tomar decisiones de forma independiente (AI) trae grandes desarrollos y oportunidades que evolucionan muy rápidamente.

Uno de los algoritmos más conocidos y utilizados es el Edge-rank. Es el algoritmo creado por Facebook que juega un papel fundamental a la hora de que los contenidos que publiquemos en nuestras páginas de fans lleguen a la mayor parte de gente posible, ya que es el encargado de decidir qué es lo más interesante para el usuario en cada momento. La fórmula matemática está basada en los factores que influyen en la publicación de las actualizaciones:

El primer factor es el grado de afinidad entre el usuario y el creador del contenido, para lo que se tienen en cuenta, entre otras cosas, el grado de

---

<sup>63</sup> Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.

interactuación, por lo que se entiende que a mayor interacción mayor nivel de afinidad.

La relevancia del contenido publicado es el segundo factor a tener en cuenta, ya que Facebook entiende que cuanto mayor es la interacción de los usuarios con el contenido, de mejor calidad es, para lo que va asignando puntuaciones distintas según el número de “Me Gusta”, comentarios o veces que se ha compartido el contenido.

El tiempo transcurrido desde que se publicó el contenido es el tercer factor de la fórmula del algoritmo, y dice que cuanto mayor es el tiempo desde que se publicó menor es la importancia de éste.<sup>64</sup>

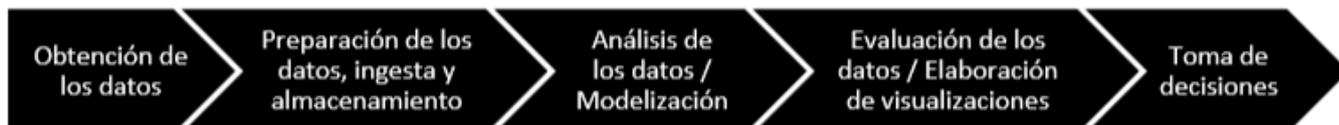
Otros algoritmos a menudo usados en modelos de clasificación y en modelos de regresión son el perceptrón, el decision tree, el random forest o el Support Machine Vector.

---

<sup>64</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

## 10. PROYECTOS DE BIG DATA

Los principales retos a afrontar en cuanto a las capacidades en Big Data están asociados a la cadena de procesos y sistemas que se encuentran involucrados en la extracción del conocimiento de los datos; desde la recolección, ingesta y almacenamiento de grandes volúmenes de datos hasta las técnicas empleadas para conseguirlo; procesamiento, analítica (minería de datos, análisis predictivo) y visualización de datos.



En todos los proyectos la comunicación es clave, pero en proyectos de Big Data es especialmente importante ser innovador para comunicar los conceptos. Cada proyecto debe ser considerado como un trampolín para el siguiente, o como una oportunidad para evangelizar enfoques centrados en datos en la empresa.

En dependencia de la tipología de proyecto en la que trabajar y el rol en la compañía que presenta la necesidad nos podemos encontrar con la siguiente clasificación de proyectos de Big Data:<sup>65</sup>

### PROYECTO ARQUITECTURA BIG DATA

El reto de estos proyectos es incorporar tecnologías Big Data y disciplinas de analítica avanzada para mejorar el conocimiento del jefe y la eficiencia operativa a corto/medio y largo plazo.

Enfoque de la solución:

- Evidenciar y realizar un análisis de la arquitectura actual.
- Realizar la definición de una nueva arquitectura en base a los posibles casos de uso.
- Establecer un roadmap de adopción (líneas estratégicas, planes de acción y calendario).

<sup>65</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

## PROYECTO ANALÍTICO

El reto de estos proyectos es Diseñar campañas y productos personalizados, además de mejorar la eficiencia de las interacciones con el cliente para aumentar la satisfacción y ratio de conversión *click to sell*.

Enfoque de la solución:

- Realización de trabajos sobre la plataforma de Big Data y empleo de algoritmos de machine learning para la clusterización de clientes y posterior segmentación de campañas.
- Identificación y aprovisionamiento de fuentes al Data Lake: campañas de captación, web, encuestas, BB.DD. externas, redes sociales, etc.
- Realización de análisis exploratorios (clasificación, patrones, segmentaciones) iteración, refinamiento y mejora de modelos; incorporación de nuevas variables (emails, info. del canal de venta, etc..) y segmentaciones.
- Informe de resultados.

Para aquellos proyectos cuyo objetivo es el desarrollo de nuevos productos o la actualización de productos existentes, existen diversas metodologías que se pueden seguir: Lean, AGILE, Kanban, XP, Scrum, waterfall...

Agile Analytics se diferencia del resto de técnicas porque va directamente a buscar respuestas a preguntas planteadas directamente por el negocio en sí, no se dedica a acumular una gran cantidad de números que son innecesarios en muchas ocasiones.

El modo de funcionamiento es un planteamiento previo de un número determinado de cuestiones para saber las herramientas de análisis precisas que vamos a necesitar. Algunos de los planteamientos previos que se realizan son: elementos que necesitamos medir, qué mediremos, como realizaremos esta medición, por qué tenemos esta necesidad de medición, etc.<sup>66</sup>

Su objetivo es recoger los datos imprescindibles y que sean estrictamente necesarios considerando previamente los objetivos del negocio. Esta

---

<sup>66</sup> Deusto Formación. (s.f.). ¿Qué es agile analytics en Big Data? Recuperado 7 mayo, 2019, de <https://www.deustoformacion.com/blog/gestion-empresas/que-es-agile-analytics-big-data>

metodología de trabajo es inversa al resto de soluciones Big Data que se focalizan en recopilar una gran cantidad de datos. Agile Analytics, trabaja con pocos datos, pero siempre intentando que estos datos puedan tener el mayor impacto posible con los resultados del negocio. La filosofía de esta técnica es que cuantos menos datos se traten más positivo será el resultado y lo obtendremos con mayor eficacia, la información se trabaja con mayor inteligencia.

Esta metodología nace en 2001 en la industria del desarrollo de 'software', cuando las compañías de este sector comprendieron que la forma tradicional de trabajo retrasaba mucho la entrega del producto final. Los CEO de las principales empresas de 'software' se reunieron en Utah, donde pusieron en común las mejores prácticas de cada compañía y crearon el "Manifiesto Agile".<sup>67</sup>

### PROYECTO DE VISUALIZACIÓN

El reto de estos proyectos es mostrar el producto del almacenamiento, análisis de la información y resultados estadísticos con una visualización amigable que emplea gráficos o mapas que son un material atractivo, entretenido y simplificado siendo mucho más fácil de entender para el cliente

Enfoque de la solución:

- Realizar la identificación de fuentes de datos internas/externas.
- Realizar el aprovisionamiento del Data Lake para la creación de análisis e informes.
- Presentación de resultados y conclusiones mediante una visualización.

"Una imagen vale más que mil palabras".<sup>68</sup>

---

<sup>67</sup> Tena, M., & BBVA. (2018, 28 noviembre). ¿Qué es la metodología 'agile'? Recuperado 7 mayo, 2019, de <https://www.bbva.com/es/metodologia-agile-la-revolucion-las-formas-trabajo/>

<sup>68</sup> Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía

## 11. CASO DE EXITO

Existen numerosas empresas en el mercado que han implementado de manera exitosa el Big Data en sus estrategias de negocio. Algunas de ellas han llegado a convertirse en organizaciones data-driven donde todas las decisiones de negocio se renuevan continuamente en función de los insights generados por el uso del Big Data en sus procesos. A continuación, presentamos un proyecto de Big Data aplicado al Marketing Digital.

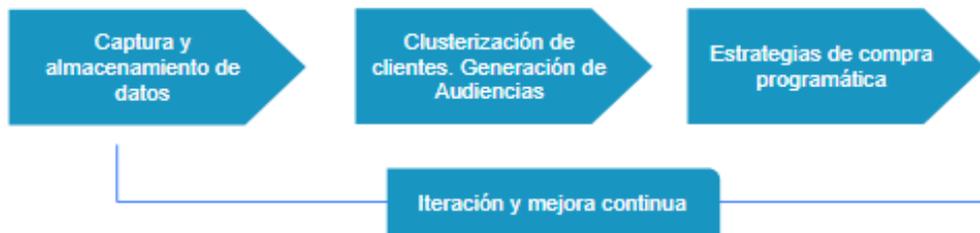
El caso que vamos a presentar se refiere a una empresa energética española, cotizada en el Ibex35, que no se menciona por motivos de confidencialidad, cuyos objetivos de negocio a conseguir eran:

- Incrementar el volumen total de ventas realizadas en Canales Digitales.
- Mejorar el % de nuevos clientes en el total de ventas.
- Mejorar el CPA (coste unitario por venta) en las ventas digitales.
- Conocer la contribución de los diferentes soportes a la venta (contribución versus last click).

Para conseguir estos tres objetivos simultáneamente, se plantearon pasar de un esquema de publicidad online tradicional basado fundamentalmente en display a un esquema basado en datos y compra programática.

En un esquema de display tradicional los resultados obtenidos dependen de los soportes y las creatividades. Las impresiones se lanzan sin tener en cuenta las características de los clientes que las visualizan. Los soportes más visitados impactan a más audiencia, pero no necesariamente es público objetivo.

En un esquema basado en datos y compra programática se dirigieron únicamente a clientes con algunas características predeterminadas en aquellos soportes que éstos visitan por lo que los impactos son mucho más efectivos.



*Fuente: Empresa energética*

El proyecto consistió en lanzar acciones de comunicación dirigidas a dos públicos objetivo:

1. Usuarios que no han terminado de contratar, pero han interactuado con el ecosistema digital de la empresa (Retargeting). De esta forma, se recuperan clientes que han comenzado la contratación, pero no la han terminado.
2. Usuarios ajenos al ecosistema digital de la empresa que por diversos motivos pueden estar interesados en contratar (Prospecting).

Se trata de acciones iterativas y en mejora continua.

Los principales focos del proyecto fueron:

- La recopilación de datos en una plataforma DMP (Data Management Platform), identificando tanto orígenes como fuentes de información, lo que permitió conseguir un elevado número de registros y grandes volúmenes de información.
- El análisis exhaustivo de los datos de comportamiento en el ecosistema digital de los compradores y no compradores, lo que permitió inferir audiencias con tendencia a la contratación basándose en datos de navegación y de comportamiento en internet. Esto hubiera sido imposible sin las técnicas de Big Data.
- Pago por impresión mediante subasta mientras se carga la página (Real Time Bidding). El máximo valor de la puja determina si se muestra el anuncio con optimización continua de estrategias digitales para cada

audiencia (pujas, frecuencias, pruebas desde diferentes DSPs para garantizar la eficiencia).

Se sigue exhaustivamente el comportamiento de los diferentes clústeres definidos modificándolos de forma continua e infiriendo nuevos.

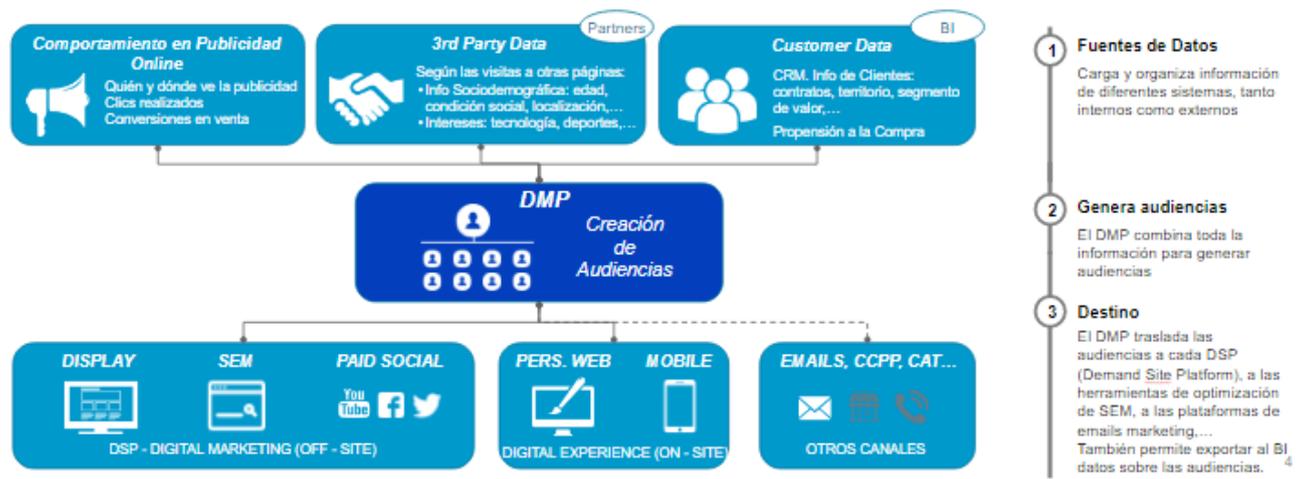
**Gest. Audiencias**

**PRINCIPALES INDICADORES**

		CLUSTERS	ANUNCIOS MOSTRADOS	AUDIENCIA RECUPERADA / GENERADA	PUNOS MEDIAS (CPM)	FRECUENCIA DE IMPACTOS
RETARGETING	Audiencia segmentada por el punto donde abandonó el proceso de contratación	Formulario de Contratación	876.133	1.357	5,27€	40
		Fichas de Producto	7.627.643	11.051	5,29€	36
	Audiencia segmentada por su fuente de tráfico de origen al entrar en el ecosistema	Home	4.569.926	3.993	3,22€	24
		Tráfico Orgánico	6.396.641	14.436	3,45€	20
		Buscadores	4.342.599	7.591	3,16€	14
PROSPECTING	Generación de nueva audiencia que no ha accedido al ecosistema de Endesa	SEO	839.784	960	3,58€	14
		Custom affinity	6.802.219	4.344	0,90€	6
		Third Party Data	3.823.595	2.121	2,71€	6
		Lookalike	16.775.103	10.690	0,84€	6
		Prospecting RON	306.630.857	73.000	0,24€	6
		Whitelist	102.771.261	38.961	0,89€	6
		Data Mobile	1.454.599	6.749	3,67€	6
		Keywords	12.353.111	8.034	1,45€	6
VENTA CRUZADA	Utilización del DMP para realizar venta cruzada de productos	Venta Cruzada	7.433.399	3.295	2,45€	12

Fuente: Empresa energética

Los principales elementos del proyecto se resumen en el siguiente esquema:



Fuente: Empresa energética

Los resultados del proyecto para la energética han sido en un año:

- Incremento de las ventas globales en un 12%.
- Incremento en 13 puntos porcentuales de los nuevos clientes en las ventas total.

- Mejora de un 24% del CPA.

Adicionalmente, desde el punto de vista cualitativo:

- Se identifican de forma continua nuevas audiencias a las que impactar basadas en el comportamiento en el ecosistema digital e imposibles de identificar sin estas técnicas.
- Se ha mejorado notablemente la información disponible pudiendo orientar mejor la inversión hacia soportes más rentables.

## 12. CONCLUSIONES

Big Data es el nuevo paradigma de la analítica de datos, llamado a transformar el modo en el que las organizaciones se enfrentan a la toma de decisiones.

La “grandeza” de Big Data, su aspecto más diferencial e innovador, consiste en que da respuesta a preguntas que las empresas ni siquiera sabían que debían de plantearse. En consecuencia, los mejores resultados son los que se obtiene sin buscarlos. Esto es debido a que no se apoya en el análisis de las inmensas cantidades de datos disponibles en un determinado momento, sino en algoritmos y modelos predictivos capaces que describir situaciones futuras.

Su implantación en las empresas, de cualquier tipo y sector, es un proceso complejo, no exento de riesgos, que tensa las organizaciones en distintos aspectos como son el económico (necesidad de disponer de recursos suficientes, no únicamente de IT) y organizativos.

En el creciente y pujante mercado de Big Data, una empresa que desee llevar a cabo un proyecto y que no cuente internamente con las capacidades necesarias, puede proveerse de los recursos técnicos necesarios, tanto de software como de hardware, así como reclutar personas capacitadas con los perfiles necesarios.

Antes de comenzar un proyecto de Big Data, hay que tener claro el objetivo que se persigue y asegurarse de tener los recursos técnicos y humanos que se precisan. Para su ejecución hay que disponer de equipos transversales que involucren a gran parte de la organización, ya que afectará a numerosas áreas de la misma.

Al exigir un cambio importante en la forma y el fondo de “hacer las cosas” y de tomar decisiones e involucrar e impactar en distintos departamentos, es necesario que el proyecto este esponsorizado por la alta dirección.

Antes de comenzar un proyecto de Big Data es necesario conocer el objetivo que se persigue, disponer del personal adecuado que sepa traducir la complejidad de los resultados técnicos encontrados a las distintas áreas involucradas y establecer los mecanismos de coordinación necesarios.

Por último, conviene recordar que no se trata de que todas las decisiones de las personas se hagan a partir exclusivamente de datos sino, más bien, que todas las decisiones de las personas deben ser avaladas y estar informadas por datos y su análisis. No se pretende que los datos eliminen por completo el juicio personal, sino que ayuden de forma eficiente a una mejor toma de decisiones, apoyada en las posibilidades de la tecnología de la era digital, que permita a las organizaciones a servir mejor a sus clientes y adquirir ventajas competitivas en sus mercados.

### 13. BIBLIOGRAFIA

- Accenture. (2018). *Digital disruption: the growth multiplier*. Recuperado de [https://www.accenture.com/\\_acnmedia/PDF-14/Accenture-Strategy-Digital-Disruption-Growth-Multiplier-Brazil.pdf](https://www.accenture.com/_acnmedia/PDF-14/Accenture-Strategy-Digital-Disruption-Growth-Multiplier-Brazil.pdf)
- Ascolta, A. (2018, 12 octubre). Ann Winblad: the woman who built Silicon Valley. Recuperado 3 mayo, 2019, de <https://www.morningfuture.com/en/article/2018/07/13/annwinblad-silicon-valley-venture-capitalist-investor-women-startup/365/>
- BBVA Fintech e innovación. (2018, 25 abril). Big Data: Ejemplos reales del uso. Recuperado 15 abril, 2019, de <https://www.bbva.com/es/ejemplos-reales-uso-big-data/>
- BBVA. (2018, 22 febrero). ¿Qué es la inteligencia artificial? Recuperado 7 abril, 2019, de <https://www.bbva.com/es/que-es-la-inteligencia-artificial-2/>
- Burkov, A. (s.f.). *The Hundred-Page Machine Learning Book*. Recuperado 15 abril, 2019, de <http://themlbook.com>
- Caballero, R., & Martín, E. (2015). *Las Bases de Big Data*. Madrid, España: Catarata.
- Codd, E. (1970). A Relational Model of Data for Large Shared Data Banks. Recuperado 6 marzo, 2019, de <https://www.seas.upenn.edu/%7Ezives/03f/cis550/codd.pdf>
- Consejo Económico y Social España. (2018). *Informe la digitalización de la economía*. Recuperado de <http://www.ces.es/documents/10180/4509980/Inf0317.pdf>
- Deusto Formación. (s.f.). ¿Qué es agile analytics en Big Data? Recuperado 7 mayo, 2019, de <https://www.deustoformacion.com/blog/gestion-empresas/que-es-agile-analytics-big-data>
- Diploma de Fundamentos de Business Analytics, material de clase (2019)
- EU. (2018, 5 octubre). General Data Protection Regulation (GDPR). Recuperado 7 marzo, 2019, de <https://eugdpr.org/the-regulation/>
- Euroforum. (2019, 14 enero). Seguridad y Big Data. ¿Son compatibles? | Euroforum. Recuperado 30 abril, 2019, de <https://www.euroforum.es/blog/ciberseguridad-y-big-data-son-compatibles/>
- European Commission. (2018, 14 mayo). Digital Economy and Society Index 2018 Report - Digital Single Market. Recuperado 6 abril, 2019, de <https://ec.europa.eu/digital-single-market/en/news/digital-economy-and-society-index-2018-report>
- Gartner IT Glossary. (2016, 19 diciembre). What Is Big Data? - Big Data. Recuperado 8 abril, 2019, de <https://www.gartner.com/it-glossary/big-data/>
- Gartner. (2018). *Informe la digitalización de la economía* (Leading the IoT). Recuperado de [https://www.gartner.com/imagesrv/books/iot/iotEbook\\_digital.pdf](https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf)
- Gartner. (2018). *Analytics*. Retrieved from Gartner: <https://www.gartner.com/it-glossary/analytics/>
- Grupo PowerData. (s.f.-a). Calidad de Datos. Cómo impulsar tu negocio con los datos.. Recuperado 12 abril, 2019, de <https://www.powerdata.es/calidad-de-datos>

- Grupo PowerData. (s.f.-b). Seguridad de datos: En qué consiste y qué es importante en tu empresa. Recuperado 26 abril, 2019, de <https://www.powerdata.es/seguridad-de-datos>
- Grupo PowerData. (s.f.-c). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Recuperado 7 marzo, 2019, de <https://www.powerdata.es/big-data>
- Grupo PowerData. (s.f.-d). Desmitificando el Data Governance: Qué, cuándo, dónde y por qué. Recuperado 5 mayo, 2019, de <https://www.powerdata.es/data-governance>
- Grupo PowerData. (s.f.-e). GDPR: Lo que debes saber sobre el reglamento general de protección de datos. Recuperado 5 mayo, 2019, de <https://www.powerdata.es/gdpr-proteccion-datos>
- IBM. (s.f.). Knowledge Center. Recuperado 15 mayo, 2019, de [https://www.ibm.com/support/knowledgecenter/es/SSNE44\\_5.2.4/com.ibm.tpc.V524.doc/fqz0\\_r\\_units\\_measurement\\_data.html](https://www.ibm.com/support/knowledgecenter/es/SSNE44_5.2.4/com.ibm.tpc.V524.doc/fqz0_r_units_measurement_data.html)
- Ico. (2018). *Big data, artificial intelligence, machine learning and data protection*. Recuperado de <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf>
- Iglesias Fraga, A. (2018, 28 septiembre). Innovadores | Los datos no son el nuevo petróleo: es la confianza. Recuperado 10 abril, 2019, de <https://innovadores.larazon.es/es/not/los-datos-no-son-el-nuevo-petroleo-es-la-confianza>
- INE. (s.f.). Encuesta sobre el uso de Tecnologías de la Información y las Comunicaciones (TIC) y del Comercio Electrónico en la empresa. Recuperado 12 abril, 2019, de [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C)
- Instituto Nacional de Estadística. (2018). *Encuesta sobre el uso de Tecnologías de la Información y las Comunicaciones (TIC) y del comercio electrónico en las empresas Año 2017*. Recuperado de [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176743&menu=ultiDatos&idp=1254735576799](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176743&menu=ultiDatos&idp=1254735576799)
- IT Digital Media Group. (2018, 25 abril). El 65% de los profesionales desconoce tecnologías como Big Data o IoT. Recuperado 4 mayo, 2019, de <https://tecnologiaparatuempresa.ituser.es/estrategias/2018/04/el-65-de-los-profesionales-desconoce-tecnologias-como-big-data-o-iot>
- IT-NOVA. (s.f.). La democratización de los datos. Recuperado 11 marzo, 2019, de <https://it-nova.co/es/es/la-democratizacion-de-los-datos/>
- Denning, P. J. (1990). Saving All the Bits. Recuperado 7 abril, 2019, de <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19910023503.pdf>
- Díez, F. J. (2014, 24 noviembre). Introducción a los Modelos Gráficos Probabilistas. Recuperado 15 abril, 2019, de <http://www.ia.uned.es/~fjdiez/libros/intro-mgp.pdf>
- Hyndman, R. J., & Athanasopoulos, J. (2018, abril). *Forecasting: Principles and Practice*. Recuperado 12 marzo, 2019, de <https://otexts.com/fpp2/>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., & Roxburgh, C. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*.

- Marr, B. (2015). *Big Data en la práctica, cómo 45 empresas exitosas han utilizado análisis de Big Data para ofrecer resultados extraordinarios*. Madrid, España: Teell Editorial, S.L.
- McKinsey & Company. (s.f.). Changing focus amid digital disruption. Recuperado 15 mayo, 2019, de <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/how-we-help-clients/changing-focus-amid-digital-disruption>
- Ministerio de Empleo y Seguridad Social ( MEySS). (2018, enero). Cifras Pyme. Recuperado 30 abril, 2019, de <http://www.ipyme.org/es-ES/ApWeb/EstadisticasPYME/Documents/CifrasPYME-enero2018.pdf><https://www.euroforum.es/blog/ciberseguridad-y-big-data-son-compatibles/>
- Peter Jeffcock. (2018, 11 julio). What's the Difference Between AI, Machine Learning, and Deep Learning? Recuperado 7 abril, 2019, de <https://blogs.oracle.com/bigdata/difference-ai-machine-learning-deep-learning>
- Samaniego, J. (2018a, 18 mayo). Big Data: cómo se pasa de cantidades ingentes de datos a información valiosa que ayude en la toma de decisiones. Recuperado 13 abril, 2019, de <https://hablemosdeempresas.com/grandes-empresas/big-data-smart-data/>
- Samaniego, J. (2018b, 18 mayo). Big Data: cómo se pasa de cantidades ingentes de datos a información valiosa que ayude en la toma de decisiones. Recuperado 13 abril, 2019, de <https://hablemosdeempresas.com/grandes-empresas/big-data-smart-data/>
- SAS. (s.f.). Big Data, mejor Marketing. Recuperado 1 junio, 2019, de [https://www.sas.com/es\\_mx/insights/big-data/big-data-marketing.html](https://www.sas.com/es_mx/insights/big-data/big-data-marketing.html)
- Telefonica Digital. (2018). *Big data Learn4Sales*. Recuperado de Material de formación de la compañía
- Tena, M., & BBVA. (2018, 28 noviembre). ¿Qué es la metodología 'agile'? Recuperado 7 mayo, 2019, de <https://www.bbva.com/es/metodologia-agile-la-revolucion-las-formas-trabajo/>