Estudis de Ciències de la Informació i de la Comunicació
Grau de Comunicació

TRABAJO DE FIN DE GRADO

MENTIRAS A LO GRANDE

Una introducción al Big Data en el marketing y el efecto de los *dirty data*

Autora:

Ara Ayora Díaz

Directora del proyecto:

Nati Tomàs Estrada

Curso: 2015-2016

Resumen

El objetivo de este trabajo es introducir el concepto de Big Data, con las implicaciones que tiene en cuanto al modo en que vemos y entendemos el mundo que nos rodea; presentando un breve apunte sobre la tecnología que lo soporta así como ejemplos de algunos de los sectores en los que su uso es más habitual. Para, a continuación, centrarse en el marketing, profundizando en las distintas oportunidades que los datos masivos pueden ofrecer.

Por último, se trata un problema que ha de ser tenido en cuenta, el de los *dirty data*, o datos sucios, aquellos erróneos, duplicados o, lo que resulta más interesante, los que son falsos debido a la decisión consciente de los ciudadanos a la hora de proporcionarlos y se analizan las motivaciones que llevan a esa actitud.

Palabras clave: Big Data, marketing, dirty data, errores, información, muestra, población, producto, precio, comunicación, mentiras, motivaciones

Abstract

The goal of this paper is to introduce the concept of Big Data, including the implications on the way we see and understand the world around us; including a brief outline of the supporting technology and some examples of areas where it is regularly in use. Next it focuses on marketing, giving a more in-depth view of the opportunities offered by massive amounts of data

To close, the paper covers a problem that must be taken into account: dirty data. Data that are erroneous, duplicated or, most interestingly, given deliberately wrong by citizens, whose motivations for such behavior are subjected to analysis.

Keywords: Big Data, marketing, dirty data, errors, information, sample, population, product, price, communication, lies and motivations.

Tabla de contenido

Tabla de ilustraciones

1. Introducción

Big Data es un concepto de moda pero no es realmente algo tan nuevo. Si lo tomamos desde una traducción literal podemos hablar de datos masivos y con ellos ya se trataba hace años. La verdadera diferencia que ha hecho que se ponga de moda es que ahora se cuenta con tecnología que permite tratar grandes cantidades de datos en un tiempo muy limitado, de manera que el análisis de dichos datos pueda ser utilizado casi en tiempo real para tomar decisiones, tanto comerciales como de cualquier tipo, en cualquier campo en que sean de aplicación.

"De la misma forma que el telescopio nos permitió vislumbrar el universo y el microscopio nos permitió comprender los gérmenes, las nuevas técnicas de recopilación y análisis de enormes volúmenes de datos nos ayudará, a ver el sentido de nuestro mundo de una forma que apenas intuimos."(Mayer-Schönberger & Cukier, 2013)

Haciendo un poco de historia, desde el comienzo los humanos se han esforzado por almacenar la información que les era útil, así apareció la escritura que permitió transmitir las leyendas y los mitos, así como otras informaciones, de generación en generación sin las pérdidas que suponía la transmisión oral.

La invención de la imprenta supuso un gran avance que permitió el acceso masivo a la información, colocó el conocimiento al alcance no sólo de unos pocos privilegiados sino del gran público.

Aunque se puede argumentar que la historia de la informática da comienzo mucho antes, es en la década de los años 30 del siglo pasado cuando se empiezan a construir lo que se podría identificar como los primeros ordenadores. Una figura clave en dicho avance fue la de Alan Turing con su "Máquina Universal de Turing" que si bien era una abstracción ponía los cimientos para que se fabricasen los primeros ordenadores en Gran Bretaña durante la Segunda Guerra Mundial.

En la actualidad, se cuenta con tecnología informática que mejora día a día a velocidad vertiginosa; cualquier persona lleva en su bolsillo un ordenador con una capacidad de procesamiento cientos de miles de veces superior al ordenador que llevaba el Apolo XI que en 1969 llevó al hombre a la Luna.

Casi todos somos, en mayor o menor medida, conocedores de la idea heraclítea del cambio, su ***ποταμοῖς τοῖς αὐτοῖς ἐμβαίνομεν τε καὶ οὐκ ἐμβαίνομεν, εἶμεν τε καὶ οὐκ εἶμεν τε***[1] *(Heraclito & Aguirre, 1956)*, nos ha enseñado a pensar en el cambio como algo permanente, cuando en realidad lo verdaderamente permanente es la aceleración de dicho cambio y esto es especialmente cierto en el caso concreto de la tecnología.

El conocimiento en sentido estricto se alcanza al añadir a la información de la que disponemos el ingrediente esencial que forman nuestras experiencias y saberes previos. Pero la información, a su vez, está formada por datos que carecen de significado y es el tratamiento adecuado de dichos datos lo que les otorga ese significado. De manera que los datos no son los ladrillos de nuestra construcción de saber, sino que son tan sólo la materia prima de la que están hechos esos ladrillos que son la información. Con ellos, colocándolos de la manera precisa que nos dicta nuestra experiencia al analizarlos, conseguimos construir el conocimiento (Wernicke, 2015).

Este trabajo se centra en los datos, masivos eso sí, pero sobre todo en cómo se procesan para darles sentido y que tengan verdadera utilidad. Para ello se hablará de lo que nos ha llevado hasta la situación actual y cuáles son las implicaciones que tiene el Big Data en nuestra manera de estudiar y entender el mundo que nos rodea, después se comentarán algunos aspectos sobre la tecnología que actualmente se usa para el tratamiento de los datos masivos; como es lógico, no entra en las ambiciones de este trabajo hacer una análisis

[1] En los mismos ríos entramos y no entramos, [pues] somos y no somos [los mismos].

exhaustivo de las cuestiones técnicas porque es nuestra intención que lo primordial sea comprender cuáles son los fundamentos para su posterior utilización. A partir de ese punto, se enumerará una serie de campos en los que la aplicación del Big Data reporta grandes beneficios. En concreto se mencionarán los sectores relacionados con la seguridad y la defensa, el campo de la medicina y la salud pública y su uso en los deportes.

Y después de todo ello, el trabajo se centrará en el uso del Big Data en el marketing, ya que es uno de los campos que nos permite ser testigos directos de su utilización y cómo afecta a nuestra vida diaria.

Para terminar se realizará un pequeño apunte sobre uno de los problemas técnicos más importantes en el Big Data, el *dirty data*, o los datos sucios, que son aquellos datos erróneos que aparecen en esas grandes bases de datos. La manera en que pueden llegar a producirse es diversa, generalmente se trata de errores involuntarios de transcripción o duplicaciones, pero también se producen por la cada día mayor tendencia entre los ciudadanos a dar datos falsos cuando se cumplimentan formularios a través de internet.

2. Marco teórico

2.1. Antecedentes

Como se ha comentado en la introducción, la historia de la acumulación y almacenamiento de datos va ligada al propio desarrollo del ser humano como ser social pero centrándonos un poco más en lo que actualmente se conoce como datos masivos o *big data,* es necesario nombrar un concepto hasta ahora no tratado en este trabajo: el de la inteligencia de negocio (también conocida como inteligencia empresarial) o *business intelligence*, BI.

Se entiende por inteligencia de negocio el conjunto de actividades encaminadas a generar conocimiento sobre todo lo que influye en la empresa, para que la toma de decisiones sea lo más adecuada posible y lleve al cumplimiento de los objetivos estratégicos marcados.

Fue Hans-Peter Luhn de IBM quien en 1958 escribió en un artículo lo que se puede considerar la primera aproximación a una definición de BI. El artículo se publicó en el IBM Journal of Research and Development y se titulaba "A Business Intelligence System"[2], en él se decía: "…*business* is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, etcetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an *intelligence system.* The notion of *intelligence* is also defined here, in a more general sense, as the "ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal."[3] (Elliot, 2007)

[2] Un sistema de inteligencia de negocio

[3] "…el negocio es un conjunto de actividades que se lleva a cabo con un propósito, ya sea ciencia, tecnología, derecho, gobierno, defensa, etc. Se puede denominar Sistema de Inteligencia a aquella parte del negocio (entendido de manera amplia) encargada de atender la vertiente comunicativa del mismo. El concepto de inteligencia se puede definir, de forma genérica, como la "capacidad de comprender las interrelaciones de hechos conocidos de forma que permita guiar las acciones necesarias para la consecución de un objetivo"

A partir de los años 60 se comenzaron a desarrollar las bases de datos. Hoy el concepto de base de datos es algo común y se ha instalado en el vocabulario popular pero fue Codd quien lo formuló, dando lugar a la teoría de bases de datos relacionales o SQL, las que todos hemos usado hasta ahora y están basadas en tablas que almacenan datos y que se relacionan entre sí mediante claves.

Con estas bases de datos aparecieron las aplicaciones que usan los datos para que las empresas puedan ajustar su operativa y conseguir mejores resultados. Son los sistemas de planificación de recursos empresariales o ERP, como por ejemplo: SAP, Siebel o PeopleSoft que sirven principalmente para gestionar el día a día de las compañías y manejar cosas como la logística o las tareas de producción, llevar el inventario, la contabilidad o generar las nóminas y las facturas. Así dio comienzo el uso de la BI en la empresa, pero estos sistemas tenían y tienen limitaciones, a saber:

- el tiempo necesario para procesar esos datos era muy elevado, por lo que sólo servían para hacer ajustes en la estrategia a medio y largo plazo;
- además los datos que utilizaban dichas aplicaciones habían de ser cuidadosamente añadidos y tratados previamente para que fueran coherentes y no tuvieran errores o duplicaciones, es decir, era necesario que tuvieran integridad y,
- por último, la cantidad de datos que se podían manejar con estos sistemas era pequeña (sobre todo comparada con la que a día de hoy se genera).

Esta situación se mantuvo hasta ya entrado el presente milenio, pero esas limitaciones unidas a la eclosión de internet y la gran cantidad de datos que se recogen a través de la red, ya sea con objetivos específicos o de manera general, nos lleva a la aparición del Big Data y lo que ello representa.

Aunque otros muchos ya habían nombrado el término Big Data con anterioridad, en 1998, John Mashey publicó una presentación titulada "Big Data and the Next Wave of Infrastress"[4] (Mashey, 1998) con la que se popularizó el concepto.

A pesar del tiempo transcurrido, aún no hay un acuerdo en cuanto a la definición del Big Data, tanto es así que en un estudio realizado por IBM en colaboración con la Escuela de Negocios Saïd en la Universidad de Oxford, en la que se preguntaba a más de 1.100 profesionales, expertos y académicos relacionados con el mundo de la TIC (Tecnologías de la Información y la Comunicación) por qué cuestiones consideraban nucleares en la definición de Big Data los resultados fueron los siguientes (IBM & Oxford, 2012):

Ilustración 1. Definición de Big Data (IBM & Oxford, 2012)

[4] Big Data y la próxima ola de Infrastress

Un poco después de la presentación de Mashey, en 2001, Doug Laney introdujo el concepto de las 3 "v" que caracterizan el Big Data (Laney, 2001):

- Velocidad, que hace referencia a la rapidez con que los datos se generan y, eventualmente, pierden valor. Algunos hablan de que los datos están en movimiento. Ante esta situación necesitamos poder procesarlos a una velocidad que nos permita sacarles partido.
- Variedad, que tiene que ver con la heterogeneidad que caracteriza los datos que son tratados con esta tecnología. Los datos provienen de múltiples fuentes y están en todos los formatos imaginables.
- Volumen, que se refiere, lógicamente a la gran cantidad de datos que se manejan.

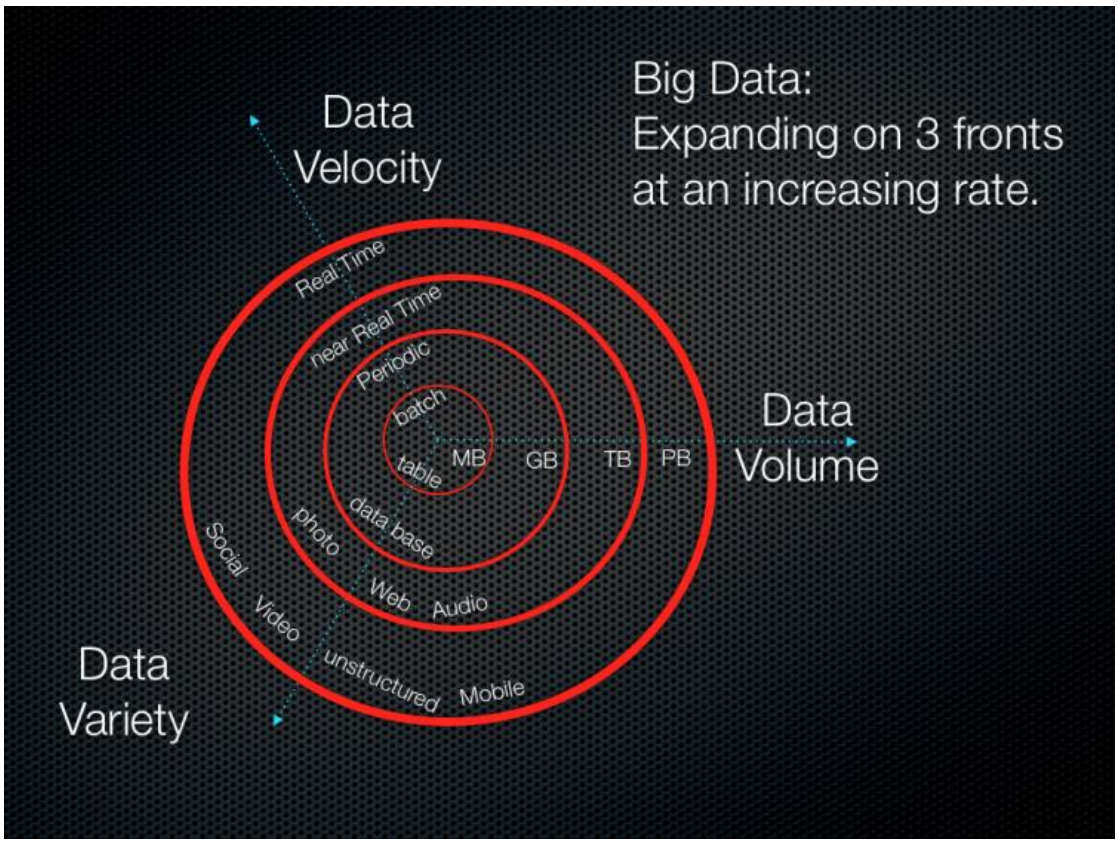


Ilustración 2. Big Data: Expanding on 3 fronts at an increasing rate (Soubra, 2012)

En la actualidad, algunos incluyen una cuarta v, la de la veracidad pero esa es una característica que otros cuestionan y trataremos más adelante.

2.1.1. ¿Qué cambia el Big Data?

En la actualidad, el Big Data es un mundo en crecimiento que constantemente está recibiendo atención por una miríada de empresas, profesionales y académicos que van aportando nuevas perspectivas, una de las interesantes desde mi punto de vista, es la presentada por Viktor Mayer-Schönberger, profesor de Internet Governance and Regulation en Oxford y Kenneth Cukier, data editor en The Economist, en su libro "Big Data - La Revolucion De Los Datos Masivos", en él exponen los tres grandes cambios que trae el Big Data (Mayer-Schönberger & Cukier, 2013):

- El paso de estudiar muestras (subconjuntos escogidos de casos o individuos de una población) a estudiar la población en su conjunto.
- El cambio de trabajar con datos lo más precisos posible a tener unos datos "confusos" o no del todo exactos.
- La transformación de tratar de conocer el "porqué" a saber el "qué".

2.1.2. De "n" a "N"

En el primer cambio nos encontramos ante una situación novedosa y que, según ellos, define lo que podemos o no considerar Big Data. Hasta ahora los estudios que se realizaban para conocer cuestiones como, por ejemplo, la prevalencia de una determinada caracteristica física en una población se basaban en tomar una muestra de dicha población ("n"), más o menos al azar, para proceder a estudiar esa condición en esa muestra y los resultados obtenidos se daban por buenos para el conjunto total de la población ("N"). Ahora tenemos la posibilidad de estudiar esa población en su conjunto lo que nos permite no hacer suposiciones estadísticas sino, realmente, conocer esa prevalencia. Con ello elimminamos los sesgos clásicos debidos a la elección de la muestra, lo que se conoce como errores de muestreo.

Además, al tener los datos completos de toda la población a estudio podemos centrar nuestra atención en subconjuntos mucho más pequeños que no tendrían

suficiente representación en una muestra estadística como para sacar conclusiones válidas con respecto a ellos, esto es especialmente útil cuando nos planteamos cuestiones como la segmentación publicitaria.

2.1.3. Datos "confusos"

Los datos con los que se trabaja en Big Data provienen de múltiples fuentes con distintos grados de fiabilidad, eso implica que algunos serán más exactos que otros y que tendremos muchos datos de los que no podemos estar seguros.

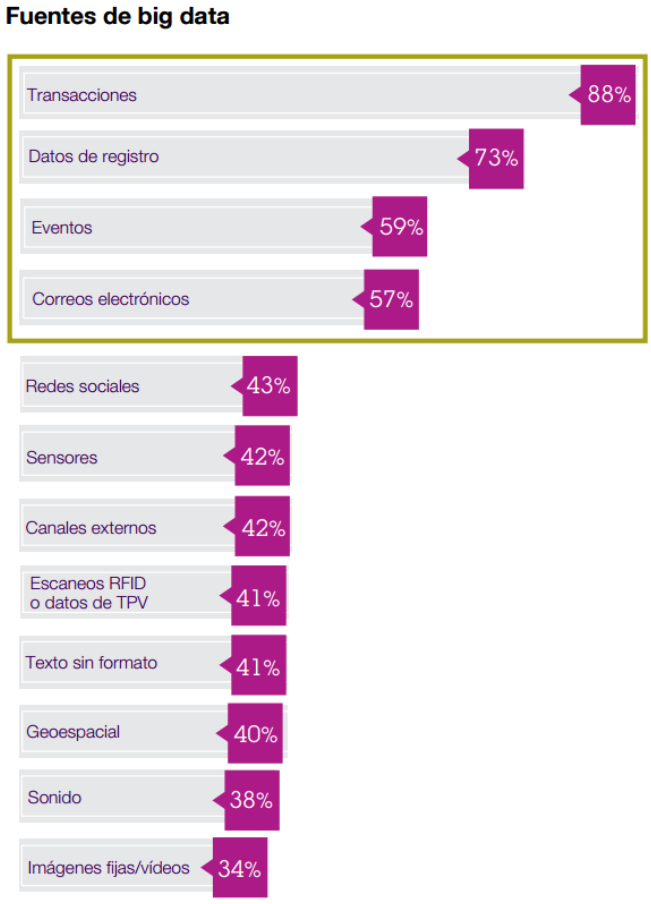


Ilustración 3. Fuentes de Big Data (IBM & Oxford, 2012)

No obstante, según Mayer-Schönberger y Cukier, como hemos eliminado el error de muestro y el volumen de datos es tal podemos permitirnos esa cierta

"confusión" en los datos. Porque el hecho es que esos datos menos precisos quedan cubiertos por la gran cantidad de datos correctos que obtenemos. Contamos con una herramienta de gran valor a la hora de enfrentarnos a este problema, la matemática avanzada que nos permite usar planteamientos de lógica difusa además de técnicas sobradamente probadas de optimización de datos.

Si aceptamos esta premisa debemos plantearnos la utilidad de la cuarta "v" de la que hablabamos anteriormente, la veracidad, y si acaso no debemos acostumbrarnos a la rompedora idea de que con el Big Data la incertidumbre debe ser considerada y utilizada en nuestro favor (IBM & Oxford, 2012).

2.1.4. La correlación como herramienta

Desde el principio, el ser humano se ha sentido impelido a hallar el porqué de las cosas, en ello hemos basado nuestro desarrollo, el conocimiento científico que nos ha permitido evolucionar y adaptarnos a un entorno cambiante sin desaparecer en el intento.

Bajo esa premisa se sigue investigando y resulta esencial y muy útil, de eso no cabe duda alguna, pero el Big Data nos plantea una nueva manera de saber, nos muestra cómo son algunas cosas y no el porqué de ellas. Nos muestra correlaciones entre distintos fenómenos.

Sacarle partido a esa nueva perspectiva requiere ser consciente de que las correlaciones matemáticas no implican ningún conocimiento sobre causalidad.

Una correlación indica la relación que se da entre dos o más cosas, ideas, personas o variables. Se traduce en que cuando una variable se comporta de un determinado modo la otra varía su comportamiento de manera sistemática con la primera. Y es que son las correlaciones matemáticas las que nos permiten explotar la potencialidad de los datos masivos, de ese modo sabemos que

determinadas cosas ocurren cuando ocurren otras pero **no** podemos decir que las primeras ocurren porque ocurren las segundas.

"Journalists are constantly being reminded that "correlation doesn't imply causation;" yet, conflating the two remains one of the most common errors in news reporting on scientific and health-related studies. In theory, these are easy to distinguish—an action or occurrence can cause another (such as smoking causes lung cancer), or it can correlate with another (such as smoking is correlated with high alcohol consumption). If one action causes another, then they are most certainly correlated. But just because two things occur together does not mean that one caused the other, even if it seems to make sense."[5] (Goldin, 2015)

Un ejemplo de correlación que muestra que no se pueden sacar conclusiones sobre causalidad es:

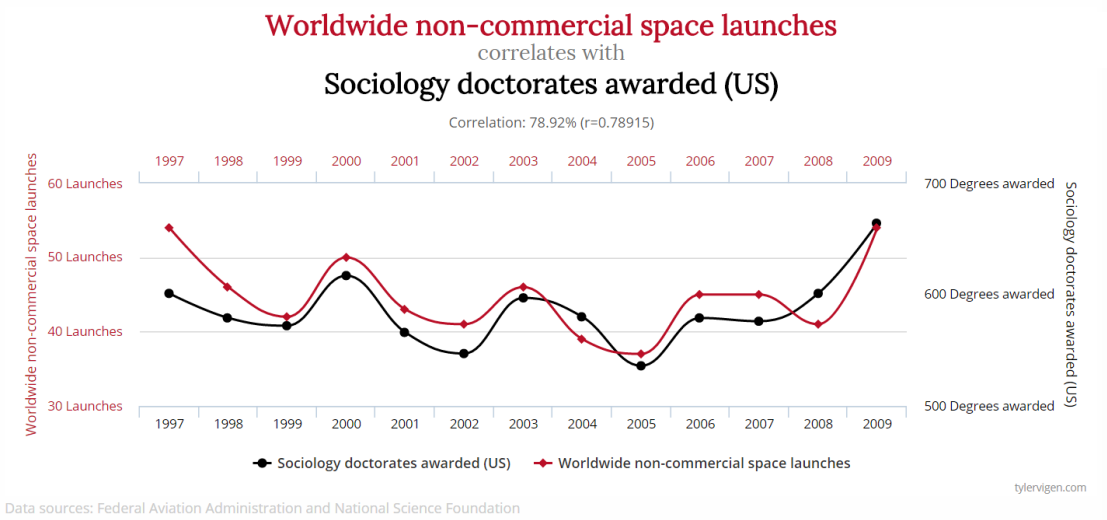


ILUSTRACIÓN 4. LA CORRELACIÓN EXISTENTE ENTRE LOS LANZAMIENTOS DE NAVES ESPACIALES NO COMERCIALES Y LOS DOCTORADOS EN SOCIOLOGÍA EN EEUU (TYLERVIGEN.COM, 2010)

[5] "A los periodistas se les recuerda constantemente que la 'correlación no implica causalidad' pese a lo cual mezclar ambos conceptos sigue siendo uno de los errores más frecuentes en las noticias sobre estudios científicos o relacionados con la salud. En teoría son fáciles de distinguir – un hecho o una acción pueden causar otra (como por ejemplo que fumar provoca cáncer de pulmón) o puede estar correlacionada con otra (como por ejemplo fumar está correlacionado con alto consumo de alcohol). Si una causa la otra, con certeza estarán correlacionadas. Pero que dos cosas ocurran a la vez no significa que una cause la otra, aunque parezca que tenga sentido."

2.2. Introducción a la tecnología

El desarrollo humano ha consistido siempre en buscar las soluciones a los problemas a los que nos enfrentamos y cuando no tenemos disponible una solución, entonces, la inventamos. Si además existe la posibilidad de generar una oportunidad de negocio entonces no se inventa una, se inventan muchas.

En lo relacionado con el Big Data nos encontramos en el segundo caso y las tecnologías asociadas se cuentan por decenas.

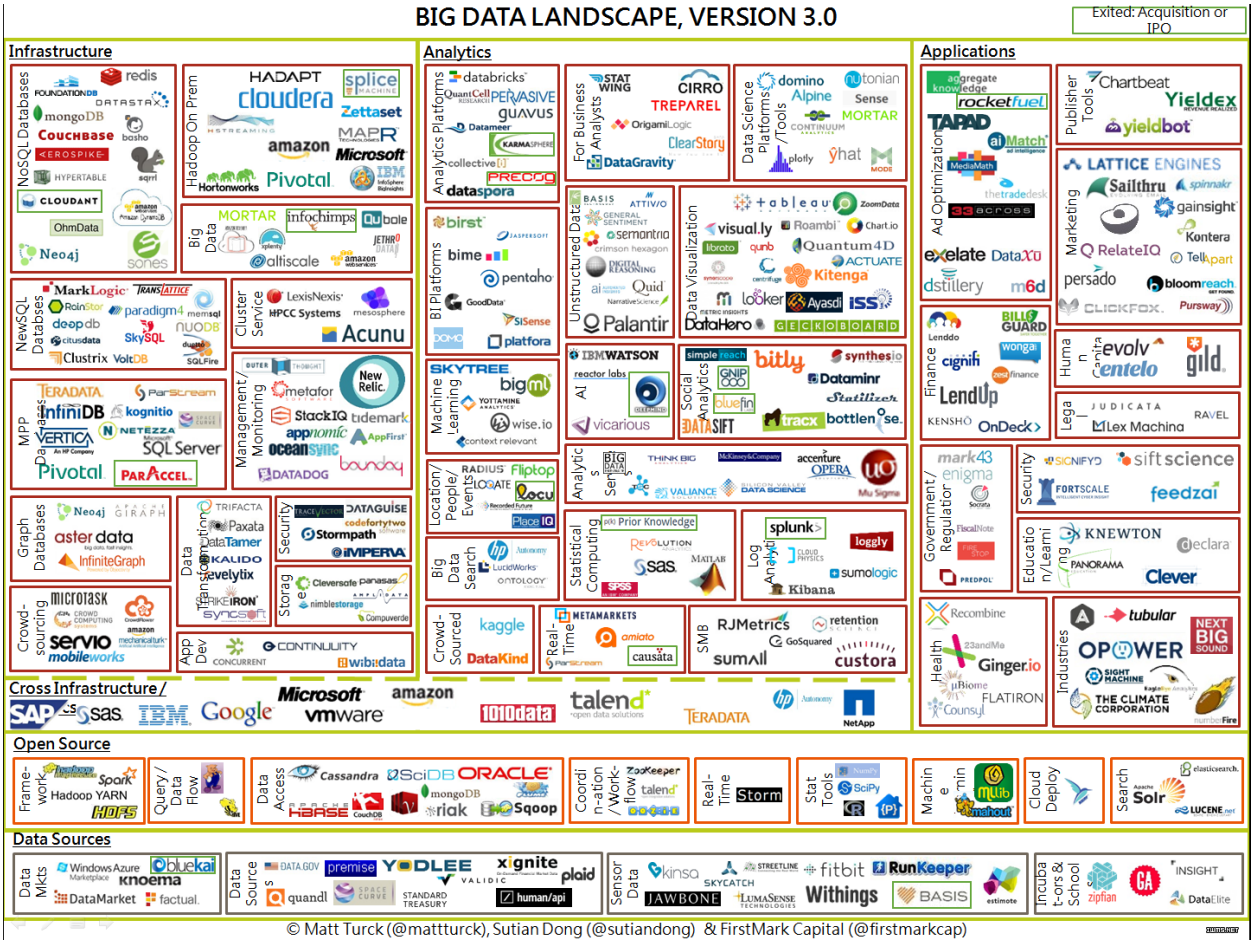


Ilustración 5. Soluciones para Big Data (Turck, 2014)

Debemos tener en cuenta que existen innumerables campos en los que se puede utilizar el Big Data; para cada sector y cada empresa u organización que lo use habrá una arquitectura que puede variar en el tiempo según se produzcan

nuevos avances o descubrimientos o, simplemente, cambien sus objetivos estratégicos y con ello sus necesidades tecnológicas.

De todos modos, a grandes rasgos, el dibujo común de cualquier sistema de tratamiento de datos masivos es el siguiente:

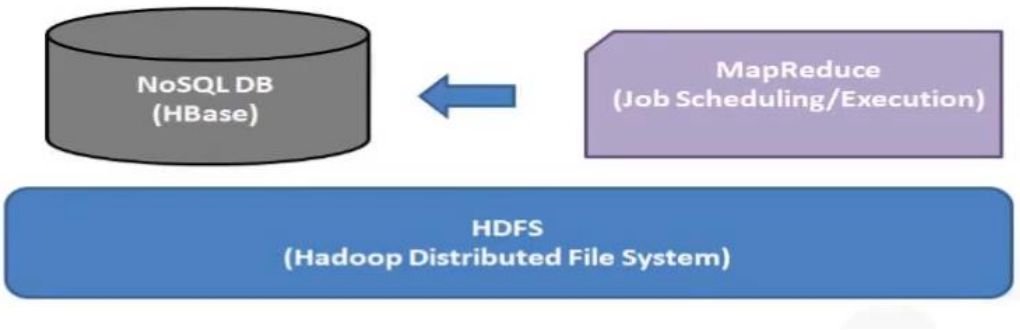


Ilustración 6. Sistema de Big Data (UOC, 2015)

En definitiva, en un sistema de Big Data existen unas aplicaciones de software que le son propias: en primer lugar contará con un sistema de ficheros distribuidos o HDFS, que permita manejar grandes cantidades de datos, en segundo lugar, existirá un motor que ejecutará las tareas y, por último, tendrá una base de datos de las denominadas NoSQL.

Con este tipo de soluciones se consigue dar respuesta a las necesidades implícitas del tratamiento de datos masivos: tolerancia a fallos, gran capacidad de almacenamiento y gran velocidad de procesamiento.

2.2.1. Sistema de archivos distribuido

Aunque no existe un estándar formalmente definido podemos considerar Hadoop como un estándar de *facto*. Es una plataforma bajo licencia Apache, es decir, se trata de software libre. Fue ideado por Doug Cutting.

La característica de ser un sistema distribuido indica que es capaz de almacenar grandes archivos de datos colocándolos en diferentes máquinas que pueden estar en diferentes emplazamientos físicos, estos son llamados nodos.

Esta característica permite que los datos se puedan procesar en paralelo y por ello Hadoop tiene tolerancia a fallos ya que al estar esos datos en diferentes nodos pueden estar duplicados o en distintas fases de procesamiento al mismo tiempo.

2.2.2. Motor de trabajos

En este caso nos encontramos con MapReduce, es un motor de procesamiento que permite la ejecución en paralelo de datos masivos. Fue creado por Yahoo y actualmente se utiliza en conjunción con Hadoop en el proyecto de software libre Apache Hadoop.

Como hemos visto en el HDFS también MapReduce presenta tolerancia a fallos. Como su nombre indica el motor tiene dos partes: Map, que realiza la extracción de los datos y les asigna un par clave/valor y Reduce, que combina los valores que tienen idéntica clave para generar un resultado único.

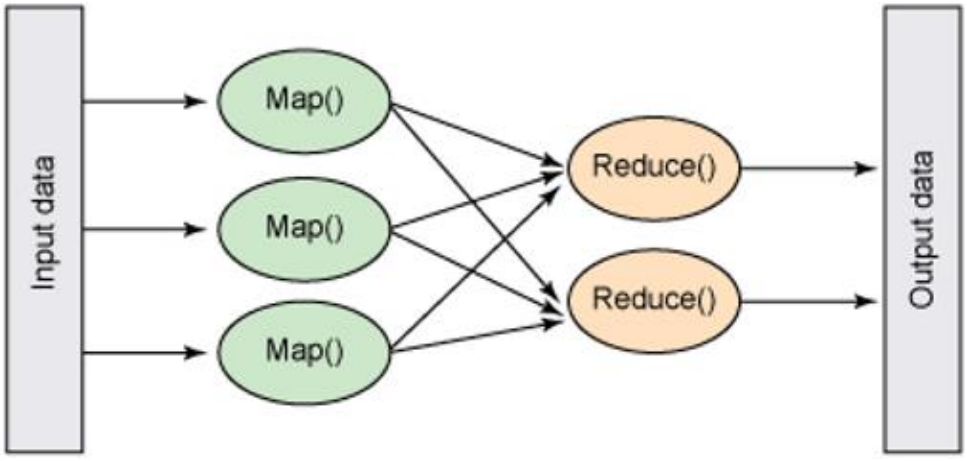


Ilustración 7. MapReduce (Niño, 2015)

2.2.3. Base de datos NoSQL

En este caso tenemos más dispersión en cuanto a las BBDD NoSQL que más se utilizan, por ejemplo HBase de Apache o BigTable de Google con su versión en software abierto, LevelDB. Además dependiendo del uso prioritario y del tipo de

dato que se vaya a utilizar existen diferentes bases de datos NoSQL en el mercado, por ejemplo para documentos o datos en grafos.

Las principales características de una base de datos NoSQL son que puede almacenar cualquier tipo de datos, que les asigna una clave única y que, al contrario que en las bases de datos tradicionales SQL, no se asegura la integridad de los mismos.

2.3. Usos preferentes

Los campos de uso del Big Data son innumerables, cualquier cuestión en la que los datos sean de utilidad podrá ser objetivo de aplicación del Big Data pero al ser esta una aproximación al conocimiento un tanto novedosa existen sectores que han tomado la delantera a la hora de implementar técnicas de Big Data.

Algunos de estos sectores resultan evidentes para todo el mundo aunque quizás no se conozca realmente el alcance al que han llegado, como por ejemplo, las redes sociales o el marketing al que dedicaremos el siguiente apartado. Sin embargo hay otros campos en los que el uso del Big Data nos pasa más desapercibido, y de ellos se hará un breve repaso a continuación.

2.3.1. Seguridad y Defensa

El uso en sistemas militares es casi siempre un punto inicial del desarrollo de los avances tecnológicos, en el caso del Big Data no fue así, de hecho, el inicio del Big Data, tal y como lo conocemos hoy, fue la investigación, por un lado la de los astrónomos y astrofísicos cuando al enviar satélites que podían recoger datos estelares se vieron inundados por los mismos y tuvieron que idear la manera de tratarlos; o los investigadores del genoma que se encontraron en un problema parecido. Sin embargo, la industria militar y de seguridad es una de las que más partido saca al Big Data.

Las aplicaciones más obvias son aquellas que sirven para el tratamiento de imágenes y al hablar de ello es inevitable pensar en el guiado de misiles o en el espionaje mediante satélites pero no es menos cierto que esas aplicaciones saltan al mundo civil proporcionando grandes beneficios como en el caso de la detección temprana de incendios forestales (Wheatley, 2013) o las actuaciones preventivas ante la formación y llegada de grandes tormentas como ciclones o huracanes (Dotson, 2012). En estos casos se usan datos provenientes no sólo de la toma de imágenes vía satélite sino también de sensores a nivel de superficie o

incluso de análisis de comunicaciones vía twitter de ciudadanos cercanos a la zona en cuestión.

Concretamente en España, el uso de los datos masivos recogidos por múltiples vías: sensores y datos de incendios, producción agrícola, densidad de población, sensores meteorológicos, etc. nos permite evaluar las zonas con mayor riesgo de desertificación.

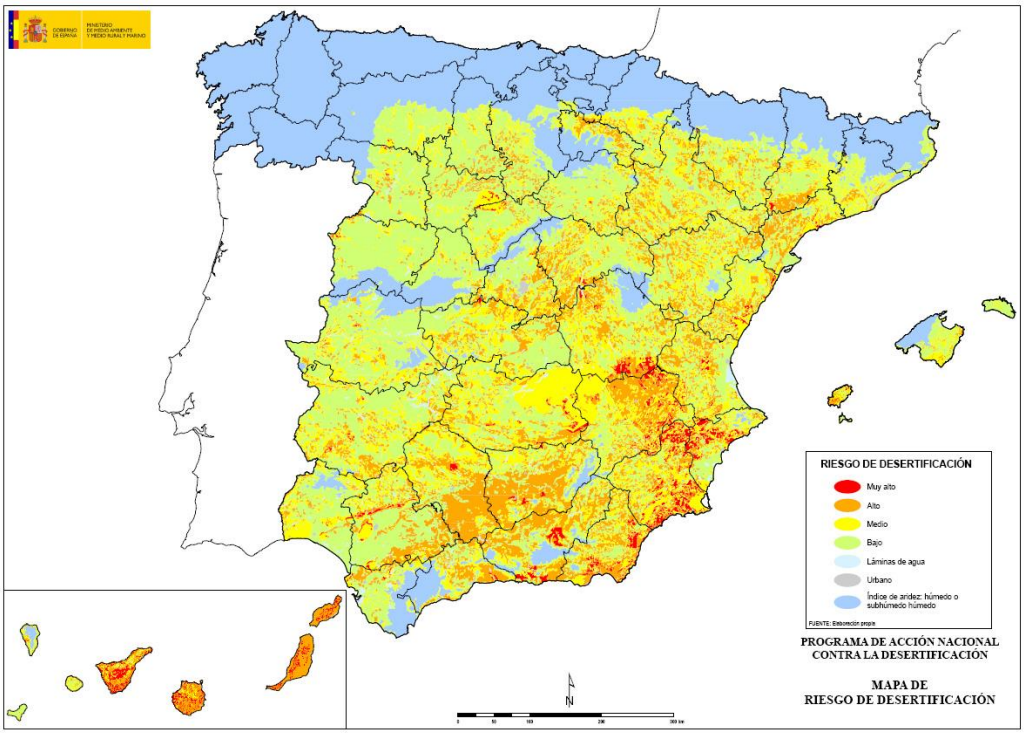


Ilustración 8. Riesgo de desertificación (Magrama, 2015)

Por supuesto, el análisis de comunicaciones a través de redes sociales se usa sin descanso en la lucha contra el terrorismo yihadista. Algunos expertos han realizado estudios sobre cómo el Daesh utiliza twitter para diseminar su mensaje y captar nuevos miembros, se cifra en más de 90.000 las cuentas de miembros o simpatizantes de este grupo terrorista por lo que su análisis requiere el uso de las técnicas de Big Data, por ejemplo para poner de manifiesto las relaciones entre unos y otros (Berger & Morgan, 2015).

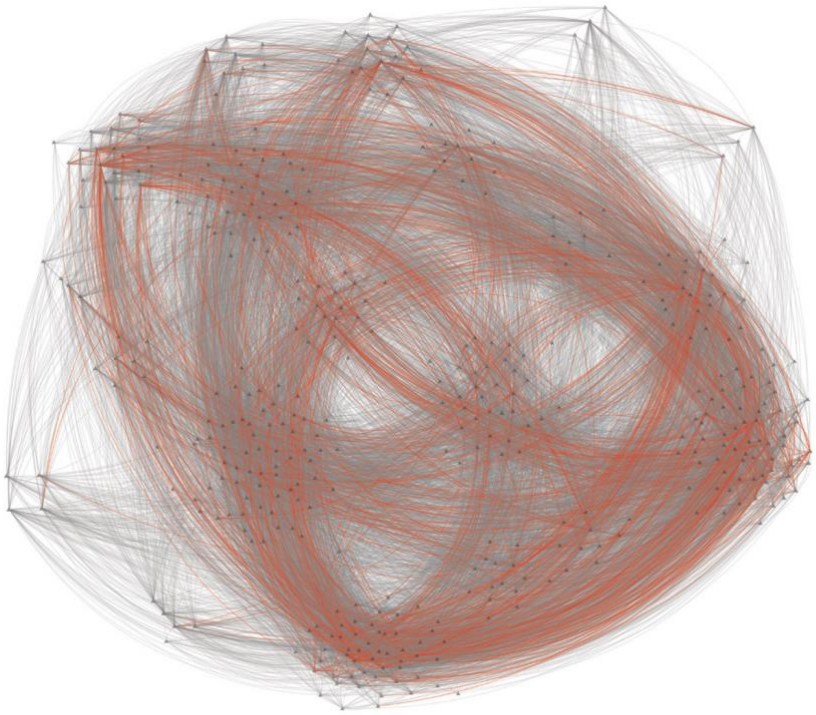


Ilustración 9. Relaciones entre las 500 cuentas principales de simpatizantes del Daesh en Twitter (Berger & Morgan, 2015)

2.3.2. Medicina y Salud Pública

Como hemos comentado anteriormente uno de los motivos del primer impulso del Big Data fue el estudio del genoma humano y es que se trataba de almacenar y tratar 3200 millones de pares de bases de ADN. Se tardó unos 10 años en hacerlo y hoy una persona puede secuenciar su genoma en un solo día; eso implica que la cantidad de datos que se irán almacenando según más y más personas se vayan realizando su estudio genético será enorme. Actualmente, se considera que se han secuenciado 250.000 perfiles genéticos lo que implica 25 petabytes de espacio de almacenamiento, o lo que es lo mismo, un 1 seguido de 15 ceros de bytes.

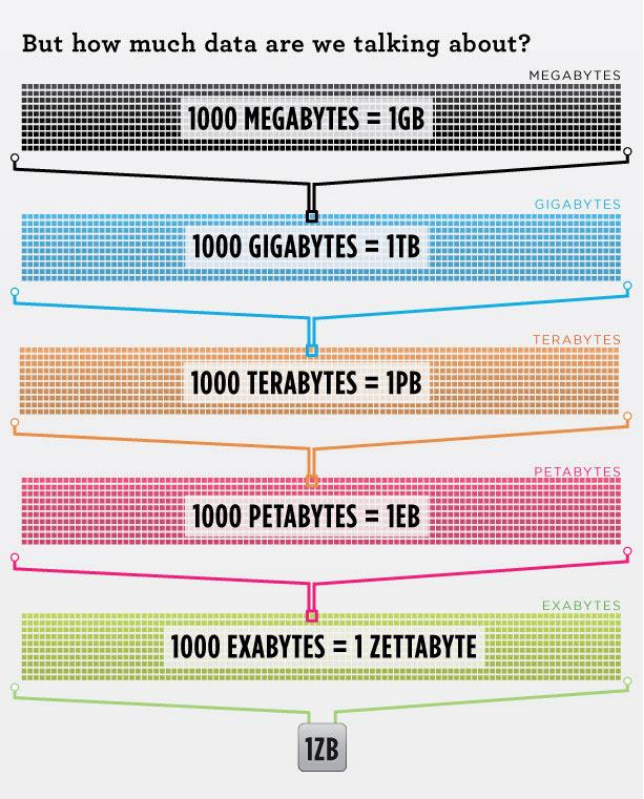


Ilustración 10. Unidades de medida de almacenamiento de datos (Take ad way, 2011)

Pero el uso del Big Data en medicina tiene como ámbito natural la capacidad predictiva de sus análisis, de tal manera que tomando los datos se puedan poner en marcha acciones preventivas para mejorar la salud de la población, tanto de manera general, mediante campañas globales; como de manera individual, activando protocolos de detección precoz en personas con características que presenten correlaciones claras con determinados problemas de salud.

Sin olvidar la capacidad para tratar datos de distintas fuentes y formatos tales como los que se almacenan hoy en los sistemas sanitarios, por ejemplo: imágenes analógicas de radiografías, informes, imágenes digitales provenientes de pruebas como RMI o TAC, listados propios de los análisis clínicos, así como la ventaja que ofrece el conocimiento del perfil genético de cada uno de los pacientes a la hora de la aplicación de la medicina personalizada; uno de los

pilares de la medicina del futuro, las 4P: personalizada, preventiva, predictiva y participativa (García Barbosa, 2014).

Entre 2014 y 2015, África ha sufrido uno de los peores brotes de ébola de la historia, ha matado a más de 11.000 personas. Para poder parar una epidemia como ésta resultaba crucial conocer cómo se propagaba y así poder colocar equipos de despliegue rápido que podían contener la enfermedad evitando su extensión pero esto era difícil de saber, el simple viaje de alguien contagiado antes de que presentara síntomas resultaba suficiente para llevar el virus de una zona a otra y el tiempo de respuesta de los equipos sanitarios en el lugar, principalmente los pertenecientes a la organización Médicos Sin Fronteras, era esencial. Lamentablemente, no se utilizaron las posibilidades que el Big Data ofrece y acabar con el brote se convirtió en una pesadilla. A posteriori, se han realizado análisis de la información en redes sociales que han mostrado que se podría haber utilizado para ser más eficiente en esa labor, se hubiera podido conocer prácticamente en tiempo real cómo se expandía la enfermedad. Pero como siempre en la historia de la humanidad se aprende fallando y ahora se tiene la experiencia, el conocimiento y las herramientas tecnológicas que harán que ante una situación similar se ofrezca una respuesta más rápida y eficaz.

Uno de esos estudios (Odlum & Yoon, 2015) muestra cómo la actividad en twitter sobre el ébola en Nigeria se adelantó al anuncio oficial del primer caso: "Tweets started to rise in Nigeria prior to the official announcement of the first probable EVD case. On July 24, Twitter users discussed the first case of EVD in related tweets (eg, "#EbolaVirus 1st case discovered Lagos, pls spread the word" and "Guys,#EbolaVirus is in Lagos. Be informed. Be careful."). The first probable EVD case was announced by the Nigerian Ministry of Health on July 27 and by the CDC on July 31."[6]

[6] Los tuits se incrementaron en Nigeria antes del anuncio oficial del primer caso probable de Ébola. El 24 de julio los usuarios de twitter comentaban el primer caso de Ébola en tuits (por ejemplo, #EbolaVirus primer caso descubierto en Lagos. Xfavor difundidlo" y "Gente, el #EbolaVirus está en

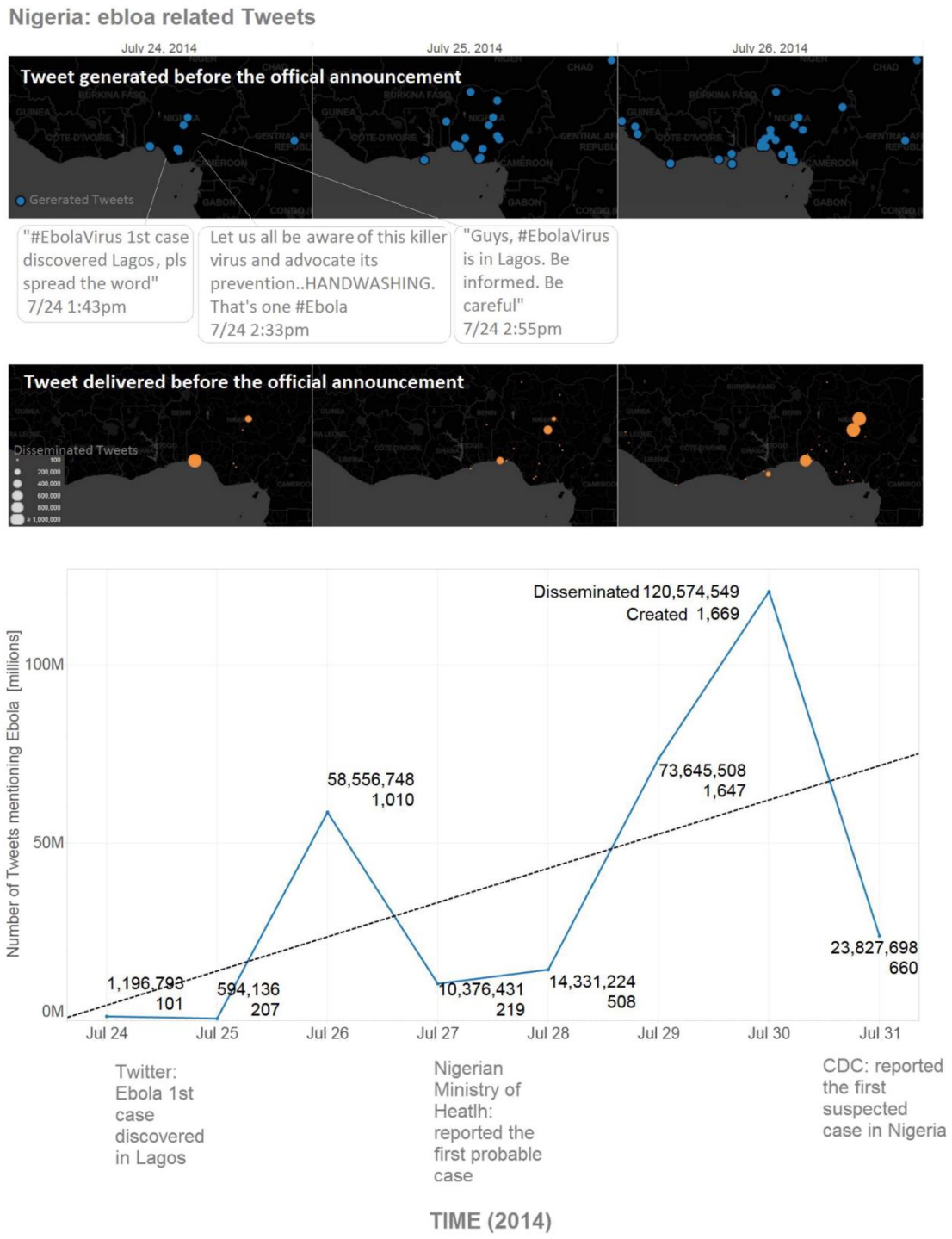


Ilustración 11. Tuits relacionados con el Ébola en Nigeria (Odlum & Yoon, 2015)

Lagos. Informaos. Tened cuidado"). El ministerio de sanidad nigeriano anunció el primer caso probable de Ébola el 27 de junio y el CDC lo hizo el 31 de julio.

2.3.3. El mundo del deporte

Hace unos años una película, Moneyball, mostró el, para algunos, sorprendente caso del éxito de un equipo de baseball, el Oakland Athletics, que confeccionó una plantilla competitiva partiendo de un presupuesto muy limitado gracias al uso de los datos. El baseball es un deporte en el que la estadística forma parte del juego, no sólo los entrenadores y jugadores sino que también los aficionados revisan los números para evaluar las posibilidades de ganar de su equipo pero además las estadísticas de bateo o de lanzamiento de los *pitchers* (lanzadores) se muestran en pantalla durante los partidos televisados.

RK	Player	Pos	G	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	AVG ▾	OBP	SLG	OPS
1	Cobb, T	CF	3035	11429	2246	4191	723	297	117	1938	1249	357	892	178	.367	.433	.513	.946
2	Hornsby, R	2B	2259	8173	1579	2930	541	169	301	1584	1038	679	135	64	.358	.434	.577	1.010
3	Delahanty, E	LF	1835	7505	1599	2596	522	185	101	1464	741	244	455	0	.346	.411	.505	.917
4	Speaker, T	CF	2789	10195	1881	3515	792	222	117	1529	1381	220	432	129	.345	.428	.500	.928
5	Williams, T	LF	2292	7706	1798	2654	525	71	521	1839	2019	709	24	17	.344	.482	.634	1.115
6	Hamilton, B	OF	1591	6269	1691	2159	242	95	40	739	1187	218	912	0	.344	.455	.432	.888
7	Brouthers, D	1B	1673	6711	1523	2296	460	205	106	1296	840	238	256	0	.342	.423	.519	.942
8	Ruth, B	OF	2503	8399	2174	2873	506	136	714	2213	2062	1330	123	117	.342	.474	.690	1.164
9	Heilmann, H	OF	2147	7787	1291	2660	542	151	183	1539	856	550	113	64	.342	.410	.520	.930
10	Keeler, W	RF	2123	8591	1719	2932	241	145	33	810	524	36	495	0	.341	.388	.415	.802

RK	Player	W	L	ERA ▲	G	GS	SV	SVO	IP	H	R	ER	HR	BB	SO	AVG	WHIP
1	Walsh, E	195	126	1.82	430	315	35	-	2964.1	2346	873	598	23	617	1736	.218	1.00
2	Joss, A	160	97	1.89	286	260	5	-	2327.0	1888	730	488	19	364	920	.223	0.97
3	Brown, M	239	130	2.06	481	332	49	-	3172.1	2708	1044	725	43	673	1375	.233	1.07
4	Ward, J	164	102	2.10	292	261	3	-	2461.2	2317	1183	575	26	253	920	-	1.04
5	Mathewson, C	373	188	2.13	635	551	29	-	4780.2	4218	1617	1133	89	844	2502	.236	1.06
6	Waddell, R	193	143	2.16	407	340	5	-	2961.1	2460	1063	711	37	803	2316	.228	1.10
7	Johnson, W	417	279	2.17	802	666	34	-	5914.1	4913	1902	1424	97	1363	3508	.227	1.06
8	Bond, T	193	115	2.25	322	314	0	-	2779.2	2857	1339	695	32	178	860	######	1.09
9	White, W	229	166	2.28	403	401	0	-	3542.2	3440	1844	896	65	496	1041	1.097	1.11
10	Reulbach, E	182	106	2.28	399	300	13	-	2632.1	2117	887	668	33	892	1137	.224	1.14

Ilustración 12. Estadísticas de bateo y lanzamiento por año de los mejores de la historia de la MLB (MLB, 2015)

Hoy en día nos hemos acostumbrado a ver cómo los deportistas llevan sensores GPS en sus entrenamientos, y no sólo los deportistas profesionales sino los aficionados utilizan detectores que les permiten conocer su rendimiento y que

recopilan información de todo tipo: pulsaciones, pasos, distancia recorrida, trayectos, modo de natación, desarrollo de la bicicleta, etc. Toda esa información se trata mediante aplicaciones más o menos complejas y sirve para decidir estrategias ante partidos importantes o modificaciones en la rutina de entrenamientos. Además es frecuente que esa información se vuelque en la red y se ponga en relación con datos similares de miles de usuarios de esas aplicaciones, vemos a continuación un ejemplo de ello:

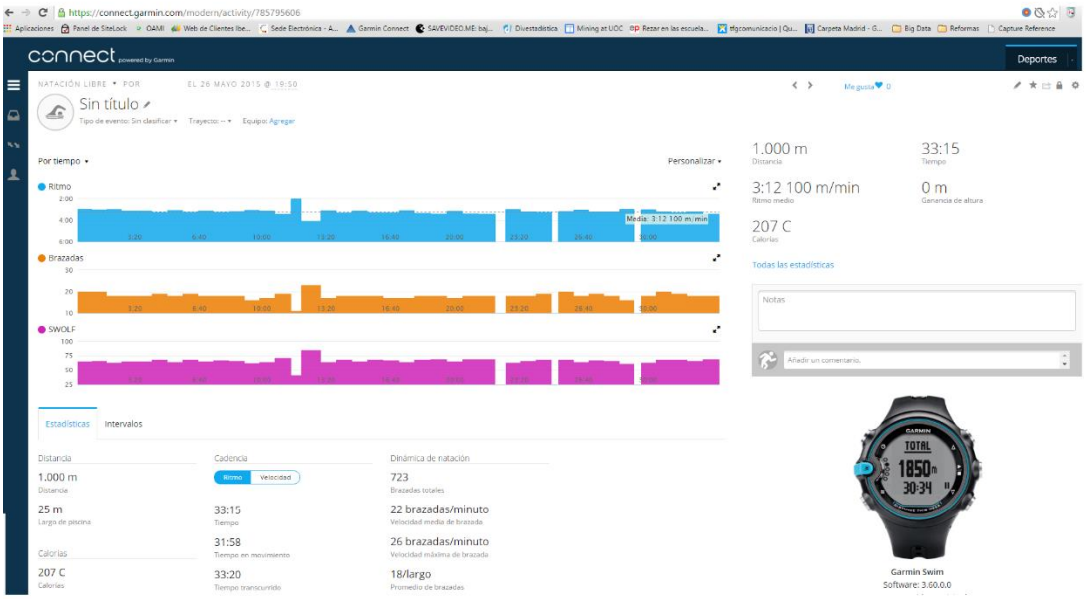


Ilustración 13. Aplicación GarminConnect con datos de una sesión de natación

Toda esa información se puede ya usar en tiempo real, así vimos al ayudante del entrenador del club Atlético de Madrid durante un partido, usando unas Google Glass para ver sobre la marcha estadísticas de los datos recogidos en ese mismo encuentro.

Ilustración 14. Uso de Google Glass en el fútbol (Cazón, 2014)

3. El uso del Big Data en el marketing

Según Philip Kotler, considerado el padre del marketing moderno, la definición de esta disciplina es la siguiente:

"El marketing es un proceso social y administrativo mediante el cual grupos e individuos obtienen lo que necesitan y desean a través de generar, ofrecer e intercambiar productos de valor con sus iguales". (Kotler & Armstrong, 2003)

Así pues, las empresas, organizaciones y particulares se relacionan con los consumidores para llevar a cabo intercambios de productos y servicios con un determinado valor económico y es el marketing el proceso que se encarga de facilitar esos intercambios que constituyen el núcleo de nuestra economía de mercado.

Para alcanzar sus objetivos, el marketing se apoya en una serie de herramientas básicas e interrelacionadas conocidas como el Marketing Mix. Éste concepto fue introducido en los años 50 por Neil Borden (Borden, 1984) y aunque en aquel momento se trataba de un conjunto de 12 variables, en 1964, Jerome McCarthy las redujo a cuatro, hoy comúnmente conocidas como las 4P, a saber:

- producto,
- precio,
- distribución (en inglés *place*) y
- promoción (hoy definida como comunicación) (McCarthy, 1964).

Si bien es cierto que existen autores que realizan una ampliación de las variables introduciendo tres más: personas, procesos y presentación (*physical evidence*), vamos a centrarnos en las cuatro anteriormente citadas para explorar las posibilidades que el Big Data ofrece al marketing y en qué manera puede ser utilizado para mejorar el proceso.

3.1. Producto

Volviendo a Kotler, producto es "todo aquello que se puede ofrecer en un mercado para su atención, adquisición o consumo, y que satisface un deseo o una necesidad". (Kotler et al., 2008)

Es decir, el producto está en el núcleo del proceso de negocio y el estudio de sus componentes es esencial para conocer si satisface o no a los consumidores. De este modo se han identificado componentes intrínsecos, que son los que han formado parte de su elaboración y componentes extrínsecos, que han sido añadidos con posterioridad, como son: (Ammetller Montes, Rodríguez Ardura, & Universitat Oberta de Catalunya, 2009)

- La calidad
- El diseño (envase y etiqueta)
- La marca
- El servicio al cliente

Se entienden como producto no sólo los objetos físicos, como son los bienes de consumo tangibles, sino que también se incluyen otros que, a pesar de ser intangibles cumplen con la idea de producto, como es el caso de los servicios o incluso las ideas.

El Big Data presenta la capacidad de poner de manifiesto necesidades o deseos no cubiertos, lo que puede dar lugar a la aparición de nuevos productos. Así mismo, facilita el análisis de los componentes extrínsecos lo que permite modificarlos para hacer más atractivo el producto.

A continuación se exponen un par de ejemplos de lo anterior.

Netflix, compañía de VoD[7] y actualmente también una productora audiovisual, utiliza el análisis de datos masivos de manera exhaustiva y muy exitosa.

[7] Vídeo bajo demanda

Recopilan todo tipo de datos de sus suscriptores, sin limitarse a lo más obvios como pueda ser el historial de compras o visionados, sino que hacen uso máximo de la tecnología de *streaming* que les permite conocer si los espectadores se saltan determinadas escenas, retroceden para verlas de nuevo, usan o no la pausa y qué hacen tras una pausa (continuar en ese u otro momento o bien volver a la página de inicio y seleccionar otro contenido).

Esta cantidad masiva de datos, a los que se unen los obtenidos mediante recopilación y análisis de los comentarios en redes sociales, han servido para que Netflix produzca dos de las series de más éxito de los últimos tiempos, "House of Cards" y "Orange is the New Black", asegurándose de que el desarrollo posterior de las tramas, a partir de la segunda temporada ya que las dos primeras estaban grabadas por completo, está completamente alineado con el gusto de los consumidores. Saben también que a sus usuarios no les gusta esperar semana a semana para ver el siguiente capítulo por lo que cuando estrenan la temporada liberan todos los episodios a la vez, dando lugar a maratones de visionado. Netflix está, además, analizando datos obtenidos de imágenes concretas para obtener conclusiones sobre gustos de intensidad de imagen, volumen o paletas de color preferidas por los usuarios. (Hegde, 2014) (Wernicke, 2015)

ILUSTRACIÓN 15. IMAGEN DE CABECERA DE "HOUSE OF CARDS"

En resumen, está creado un producto desde sus componentes intrínsecos, basándose en los datos tratados mediantes técnicas de Big Data para asegurar la satisfacción de sus clientes. (Jarreño, 2014)

A continuación se presentan dos ejemplos de modificación de componentes extrínsecos, como es la calidad y el servicio al cliente.

Para asegurar la calidad de su servicio, el grupo Air France-KLM utiliza el tratamiento de datos masivos que recopilan durante el vuelo, en sus aviones, para anticiparse a las posibles averías y efectuar mantenimientos que las eviten, consiguiendo así reducir el tiempo de inactividad de su flota con el consiguiente ahorro en costes. Los sensores a bordo de sus 10 modelos Airbus A380 son un total de 300 000, de los que sólo se utilizan los datos de 24 000 para hacer esos mantenimientos predictivos. En un solo vuelo se recoge un total de 1,6 Gb de datos, lo que hace que anualmente esa cifra llegue a los 9 Tb (Terabites). Sin duda, una enorme cantidad de datos que permiten aumentar la seguridad de sus vuelos y reducir los costes, algo que hace sólo unos años sonaba absolutamente contradictorio.

ILUSTRACIÓN 16. BIG DATA PARA MEJORAR LA SEGURIDAD EN LOS VUELOS.

"Lorsqu'une panne se produit sur une plateforme lointaine, la compagnie doit héberger et nourrir les passagers, envoyer un autre avion et dépêcher sur place l'équipe de réparation avec les pièces détachées nécessaires, ce qui coûte cher, très cher."[8] (Loukil, 2015)

Además de las ventajas del mantenimiento preventivo se han encontrado, sin esperarlo con una reducción en los tiempos de reparación de las averias, ya que ahora son capaces de identificar rápidamente cuál es la pieza que ha fallado, pasando de 6 horas de búsqueda a tan sólo 5 minutos.

[8] "Cuando se produce una avería en una plataforma distante, la empresa debe acomodar y alimentar a los pasajeros, enviar otro avión y darse prisa en hacer llegar al lugar el equipo de reparación con los repuestos necesarios, lo que resulta caro, muy caro."

Para cuidar de manera excelente el servicio al cliente, British Airways[9] puso en marcha, ya hace algunos años, el programa "Know Me".

Es un programa que pone al cliente y sus necesidades en el centro de atención, para ello utiliza datos de 200 fuentes distintas. (Del Rey, 2012). El programa busca crear valor para el cliente ofreciendo servicios personalizados, centrándose, más que en otras cuestiones que también se ofrecen (reservas personalizadas o similares), en la experiencia en vuelo. Así, se presta atención especial a un cliente que vuela en *business* por primera vez en la compañía, dando explicaciones sobre las funcionalidades del asiento o de los servicios que tiene a su disposición, o se tiene un detalle con un cliente habitual de *business* pero que en ese vuelo viaja en turista acompañado de su familia. Estas atenciones personalizadas se extienden incluso a las malas noticias: si las maletas no han embarcado en el avión el pasajero recibirá la información, ahorrándose al menos la infructuosa espera ante el carrusel de maletas. También realizan acciones que generan una gran satisfacción en sus clientes desde experiencias previas negativas, por ejemplo si un pasajero ha sufrido retrasos o cualquier otro problema en un vuelo anterior, la tripulación tiene esa información y procede a disculparse y ofrecerle una atención especialmente cuidadosa.

3.1.1. Los datos como producto

Resulta interesante observar cómo los propios datos son un producto. Esto es así desde hace tiempo, pero últimamente ha cobrado una importancia muy superior debido a las enormes posibilidades económicas y de todo tipo que genera el poseer un conjunto de datos que puedan ser tratados para extraer información relevante para las necesidades de las empresas y organizaciones.

[9] http://www.britishairways.com/

En general, se trata de un producto de los denominados organizacionales, es decir, aquellos que son adquiridos por empresas, administraciones u organizaciones para ser usados en su propio proceso de producción.

La mayor parte de las empresas recopilan diversos tipos de datos ya que resultan imprescindibles para llevar a cabo su actividad, pero los datos son reutilizables y poseen la característica de que pueden ser usados para propósitos muy distintos a los que en su día dieron lugar a su recogida, por ello, actualmente, empresas con amplia historia se están dado cuenta del potencial económico de comercializar sus bases de datos construidas a lo largo de los años.

Este es el caso de Telefónica, que incluso creó una empresa en su grupo dedicada a este fin, Telefónica Dynamic Insights[10] (Lunden, 2012).

Utilizan datos anónimos y agregados para crear modelos que permitan a sus clientes evaluar el comportamiento de usuarios de móviles en cuanto a geolocalización como a hábitos comportamentales. Específicamente, en su oferta dicen:

> Los datos de tráfico se enriquecen con **información demográfica y de comportamiento** e incluyen la ubicación sociodemográfica, residencial y laboral, la información sobre los objetivos de los visitantes, el grupo de edad, género y otros atributos que permiten una evaluación por perfil y una segmentación sofisticadas. Esto permite a las empresas segmentar los datos para evaluar únicamente el tráfico de su población meta y añadir así valiosos atributos a sus insights.

ILUSTRACIÓN 17. MODELO DE TDI. (TDI, 2013)

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

De igual forma, Lufthansa[11] se plantea vender los datos sobre sus usuarios, eso sí, la idea es que sus clientes den el consentimiento y para ello les ofrecen determinadas ventajas en los servicios, como descuentos, aparcamiento gratuito o acceso a la sala VIP (Magazin, 2015).

"Nach Umsatz sind wir der größte Luftfahrt-Konzern der Welt. Aber die Märkte bewerten Google, Whatsapp nach ganz anderen Maßstäben - nur dank der Daten, die sie generieren. Unsere Kundendaten dagegen werden an der Börse überhaupt nicht bewertet, folglich müssen wir mehr daraus machen. Wir müssen arbeiten mit diesen Daten."[12] Simone Menne, CFO de Lufthansa.

ILUSTRACIÓN 18. VENTAJAS ASOCIADAS A PERMITIR EL EMPLEO DE LOS DATOS PERSONALES. ACCESO A LAS SALAS VIP DE LUFTHANSA

[11] http://www.lufthansa.com/

[12] "Tras los procesos de ventas [que hemos sufrido] somos la mayor compañía aeronáutica del mundo, pero en los mercados se evalúa a Google, Whatsapp con unos estándares completamente diferentes - sólo por los datos que ellos generan. Los datos de nuestros clientes, por otra parte, no se han valorado en el mercado de valores, por lo tanto, necesitamos hacer lo mismo. Tenemos que trabajar con esos datos".

3.2. Precio

La Real Academia de la Lengua Española define precio como:

precio.

(Del lat. *pretĭum*).

1. m. Valor pecuniario en que se estima algo.

2. m. Esfuerzo, pérdida o sufrimiento que sirve de medio para conseguir algo, o que se presta y padece con ocasión de ello.

Pero poniendo el foco en el marketing:

"El precio es el valor (en forma de dinero o no) que el comprador de un bien entrega a cambio de la utilidad que recibe por la adquisición del mismo." (Rodríguez Ardura & Universitat Oberta de Catalunya, 2013)

Aunque en la actualidad no se considera el precio como el único factor que influye en las decisiones de compra, lo que resulta cierto es que las decisiones sobre los precios de una compañía pueden marcar el éxito o fracaso de la misma. Esto es así porque el precio es el factor sobre el que la empresa puede actuar para cambiar sus resultados de manera más rápida y por ello es la herramienta que más flexibilidad proporciona para alcanzar los objetivos fijados en la estrategia general de la compañía.

Los precios tienen un impacto sobre la percepción de los consumidores que es mayor al de cualquier otra cosa. Permiten construir una imagen tanto si son bajos con respecto al mercado como si son más altos que la media. Además, la facilidad con que pueden ser cambiados les hace ser utilizados de manera habitual como el elemento competitivo básico en cualquier sector y, por último, son el modo en que las empresas consiguen sus ingresos, y conseguir la mayor cantidad de beneficio es el fin último de cualquier negocio.

La política de precios se puede utilizar con distintos objetivos: generar imagen, mejorar la cuota de mercado o, el más habitual, aumentar las ventas. (Ammetller Montes et al., 2009)

Pero la fijación de esa política debe estar basada en datos que permitan acertar con la misma, para ello no sólo se han de tener en cuenta los costes de producción sino que se ha de estimar la demanda del producto en cuestión y también evaluar el entorno, en lo que se incluye el conocimiento lo más preciso posible de la actividad que lleva a cabo la competencia.

Y es en lo anterior en lo que las técnicas de Big Data resultan de gran utilidad.

3.2.1. Evaluación de la demanda

La evaluación de la demanda permite reducir los riesgos de sobreestimación que conllevan un exceso de inventario o de subestimación que limitan la capacidad de la compañía para ofrecer el mejor servicio al cliente debido a las roturas de stock. Para realizar una predicción correcta es necesario contar con la mayor cantidad de información posible, no sólo el histórico de los datos que tenga la compañía.

"Predecir, fundamentados en datos históricos es como conducir a oscuras por un camino desconocido, pero tomando como guía únicamente el camino recorrido" (Boada & Mayorca, 2011)

El Big Data nos permite tratar esa información para ajustar los procesos productivos, pero además nos permite hacerlo con una velocidad hasta ahora desconocida, limitando así posibles desviaciones debidas a las cambiantes condiciones del mercado.

Un ejemplo de ello es el uso de la analítica de datos en la gestión de flotas y rutas aeronáuticas. Como consumidores vemos sus efectos más evidentes en los distintos precios que alcanzan los billetes de avión dependiendo de los días o de las horas de los vuelos, pero la realidad es un poco más compleja que eso.

Las compañías aéreas no sólo utilizan algoritmos basados en los datos masivos para fijar sus precios, y lo de fijar es una mera fantasía ya que los precios de billetes de avión son extremadamente variables (Parra, 2014), sino que también utilizan los grandes datos para la planificación de sus rutas. Dependiendo de la demanda prevista envían al aeropuerto de origen el modelo de avión más adecuado, con un mayor o menor número de plazas y eso implica una cascada de decisiones sobre los mantenimientos programados de la flota; gestión de recursos humanos, todo ello con la dificultad añadida de las distintas habilitaciones con las que cuentan los pilotos que les hace válidos para pilotar un tipo de avión pero no otro, así como los cupos de horas que se les permite trabajar; acuerdos con las compañías de *handling*; o el pago de tarifas de gestión aeroportuaria y del servicio de control aéreo.

Otro ejemplo interesante es el de UBER, que realiza una fijación de sus precios basándose en la demanda en tiempo real (O'Reilly, 2015), hasta ahí todo es de lo más habitual; a lo largo de la historia los precios vienen definidos en su mayor parte por la demanda del bien o servicio que se ofrece, pero UBER va más allá y utiliza no sólo los datos provenientes de los usuarios que entran en su aplicación para pedir un coche sino que analiza el entorno y las circunstancias que rodean esas peticiones, por ejemplo, cuando el tiempo meteorológico es malo aumenta sus precios o cuando se producen eventos programados como conciertos o espectáculos deportivos.

Lo más interesante es que sus algoritmos también son capaces de reaccionar ante eventos inesperados como fue el caso de la crisis de rehenes que tuvo lugar en Sydney en diciembre de 2014. En ese momento y ante la cantidad de personas que deseaban abandonar la zona en la que se estaban produciendo los hechos, UBER cuadriplicó sus tarifas. Posteriormente y ante la avalancha de críticas recibidas se vio obligada a pedir disculpas y devolvió el importe cobrado a sus clientes, no obstante, explicaron que el aumento de la tarifa consiguió que

muchos más de sus conductores se desplazaran a la zona haciendo así posible el desalojo de una mayor cantidad de personas. (Jiménez, 2014)

ILUSTRACIÓN 19. VARIACIÓN HORARIA DE LA DEMANDA DE UBER EN NOCHEVIEJA DE 2014

3.2.1.1. Fijación automática de precios

Una de las estrategias más controvertidas es la que consiste en fijar los precios en función de la máxima cantidad que un consumidor está dispuesto a pagar, el uso del análisis de grandes datos puede ser extraordinariamente eficiente para este tipo de fijación de precios.

Es difícil encontrar alguna empresa que abiertamente reconozca su utilización ya que es algo muy mal recibido por sus potenciales clientes. No obstante, existen algunas voces que han denunciado tales prácticas. Veamos algunos ejemplos:

- en el año 2000, se anunció que Amazon estaba haciendo llegar diferentes descuentos a sus clientes ante la compra del mismo DVD. Amazon negó que estuviera segregando a sus clientes por su

sensibilidad al precio analizada con los datos sacados del histórico de sus perfiles pero ahí quedaron las denuncias (Porter, 2011)

- La compañía Orbitz que se dedica a la venta online de billetes y estancias de hotel realizó un estudio con el que supo que los clientes que acceden a las páginas de venta por internet desde sus ordenadores MAC están dispuestos a gastar un 30% más que los que lo hacen desde PC, eso lo incluyó en sus algoritmos para ofrecer precios superiores a los usuarios de MAC, aumentando su beneficio. (Mattioli, 2012)

- Por último, podemos nombrar a la compañía Progressive, dedicada a los seguros que ofrece la posibilidad de colocar un sensor en el coche que evalúa el modo de conducción para así ajustar la tarifa del seguro, ofertando mejores precios a aquellos conductores que muestren un comportamiento más adecuado y precios más elevados a aquellos que conduzcan más rápido o de manera más agresiva (Meek, 2014). Esta novedosa aproximación cambia de manera drástica el sector de los seguros, ya que modifica por completo las reglas de juego, que anteriormente también se basaban en datos con la diferencia de que no estaban personalizados (Gittleson, 2013).

3.2.2. Conocimiento sobre la competencia

En cuanto al conocimiento sobre la competencia, éste resulta crucial en un entorno de economía competitivo, así tener datos sobre los precios de las empresas que ofrecen productos o servicios similares y poder procesar esos datos en tiempo real proporciona una oportunidad que algunas empresas han sabido aprovechar, es de esperar que eso que ahora es una oportunidad se torne en un futuro inmediato en una necesidad ya que cuando el mercado en sus distintos sectores sea un actor maduro frente al Big Data todas las organizaciones contarán con las capacidades necesarias para sacarle partido y el éxito dependa entonces de la excelencia en el análisis y del diseño de sus soluciones tecnológicas.

Un ejemplo de lo anterior se pudo ver cuando los grandes comercios tradicionales (Walmart, Toys R Us, etc.) decidieron plantar cara a Amazon durante el *Black Friday* de 2013, anunciando que igualarían el precio de cualquiera de sus competidores se desató una guerra de ofertas y contraofertas en las que Amazon demostró su dominio de estas técnicas, realizando más cambios de precio que todos sus competidores juntos.

Y es que Amazon es uno de los pioneros en el uso de Big Data y roza la perfección cuando se trata de usar los datos para sacar ventajas comerciales. Su agresiva política de precios les ha llevado a cambiar en un solo día el precio de hasta 80 millones de productos.

No obstante, no todos estos cambios suponen una bajada de precio. Dado que esta hipercompetitividad genera una disminución de los márgenes de beneficio, ser capaz de tomar la decisión de subir los precios en el momento adecuado sin perder cuota significativa de mercado puede ser lo que marque la diferencia entre la supervivencia de la marca y su desaparición (Loeb, 2014).

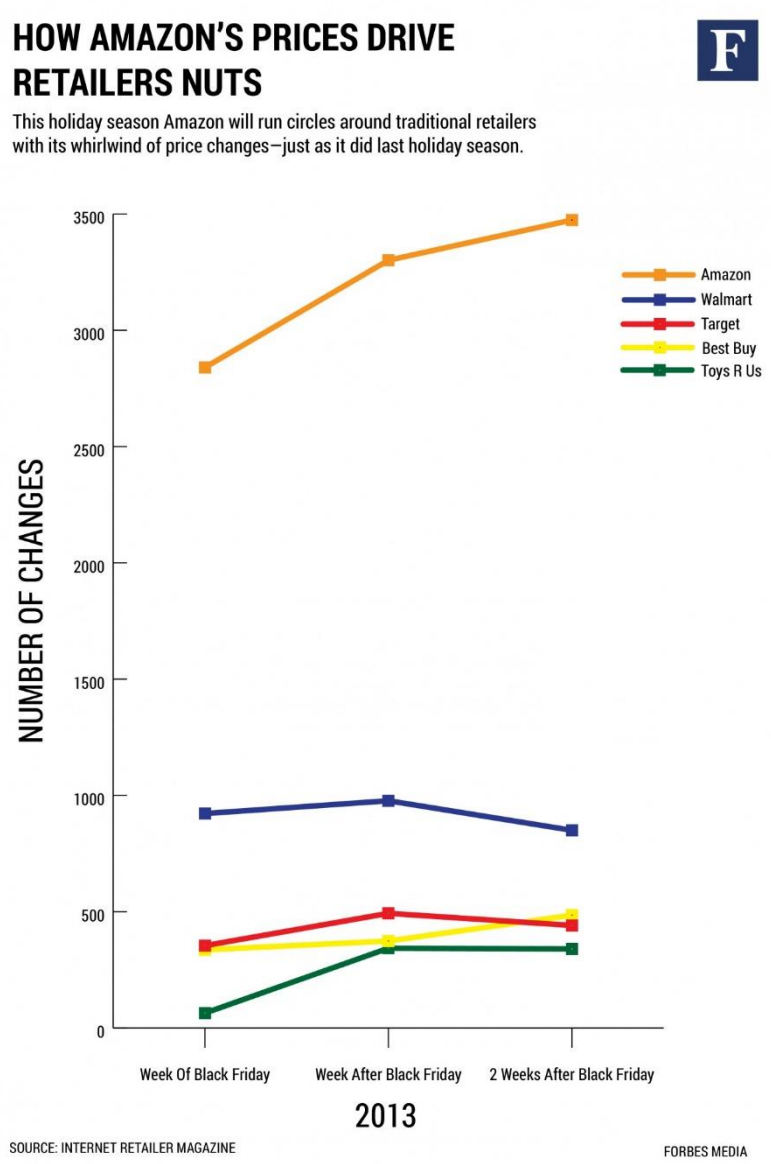


ILUSTRACIÓN 20. NÚMERO DE CAMBIOS DE PRECIO EN LAS SEMANAS DEL BLACK FRIDAY Y LAS DOS SIGUIENTES, EN AMAZON Y SUS COMPETIDORES. (LOEB, 2014)

3.3. Distribución (*Place*)

El proceso de distribución es un elemento estratégico en la actividad comercial, su función es poner a disposición de los clientes el producto, es decir, cambiar su estado, pasando de ser un bien producido a ser un bien de consumo. Dicho así puede parecer una función simple pero nada más lejos. La distribución es, generalmente, poco apreciada por los consumidores. Es una de esas actividades que se dan por sentadas y que sólo se tienen en cuenta cuando fallan.

Involucrados en ese proceso se pueden encontrar numerosos actores: agentes comerciales, representantes, almacenistas, transportistas, mayoristas, minoristas así como, por ejemplo, intermediarios que proveen productos financieros.

Todos ellos aumentan el valor del producto ya que cumplen funciones imprescindibles para su comercialización.

La distribución, además de ser imprescindible, también conlleva decisiones que influyen en el resto de herramientas del marketing, y debido a su naturaleza es una variable que no puede ser modificada de manera rápida y habitual sino que requiere de plazos mucho más amplios que lo que hemos visto con los precios. (Rodríguez Ardura & Universitat Oberta de Catalunya, 2013)

Es por ello que el análisis de los datos para la toma de las decisiones más oportunas es especialmente importante, ya que equivocarse en la política de distribución puede suponer el fracaso de toda la estrategia comercial.

El Big Data vuelve a mostrarse como una herramienta especialmente útil a la hora de apoyar la toma de decisiones. Y no sólo para la empresa responsable de la creación del producto, aquella que le da la marca, sino todas aquellas que son parte del canal de distribución. Así, el Big Data es usado por las entidades de crédito para tomar decisiones rápidas a la hora de otorgar créditos a los consumidores para que adquieran el producto, un ejemplo es la empresa

Kreditech[13] que utiliza algoritmos basados en Big Data para sus servicios de crédito y consigue dar respuesta en menos de un minuto a las solicitudes que recibe.

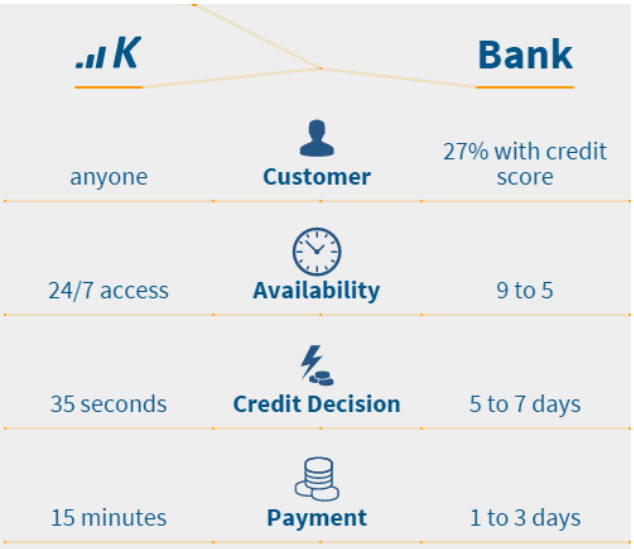


ILUSTRACIÓN 22. DIFERENCIA ENTRE LA BANCA TRADICIONAL Y EL SERVICIO DE KREDITECH BASADO EN BIG DATA

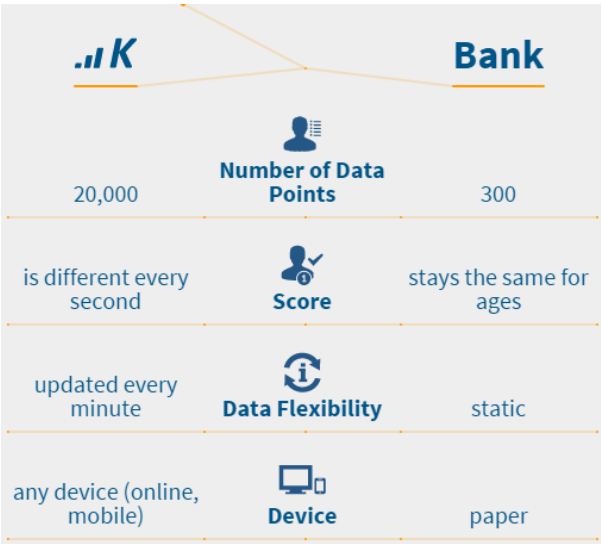


ILUSTRACIÓN 21. BIG DATA SCORING (VALORACIÓN CREDITICIA BASADA EN BIG DATA)

O las empresas de transporte pueden usar técnicas basadas en el análisis de datos masivos para ser más eficientes en el reparto de los productos, un caso paradigmático es el de UPS[14], una de las mayores empresas de paquetería a nivel mundial, que ha conseguido reducir el tiempo de entrega y disminuir sus costes de manera radical evaluando los trayectos de sus camiones de reparto, en esa evaluación toman en cuenta millones de datos, no sólo los proporcionados por los GPS de sus vehículos sino también datos meteorológicos, de tráfico, de posibilidad de retenciones, incluso han llegado al extremo de eliminar, en la medida de lo posible, los giros que sus camiones realizan a la izquierda, ya que en esos giros hay que ceder el paso a los vehículos que vienen de frente lo que les hace parar y perder tiempo. Esa decisión no se toma a la

[13] https://www.kreditech.com/
[14] https://www.ups.com/

ligera, se adopta después de un análisis exhaustivo de los datos anteriormente citados. De hecho, su algoritmo es capaz de analizar 200 000 posibilidades para cada ruta en tiempo real.

Así UPS ahorró entre 2004 y 2012 la cantidad de 37,8 millones de litros de gasolina, casi 100 millones de minutos de espera y 25 millones de horas de trabajo. Esa cantidad de combustible ahorrada supone una ventaja evidente para el medio ambiente. Además redujeron el número de accidentes que sufrían sus conductores ya que los giros a la izquierda son causa frecuente de ellos (BusinessIntelligence.com, 2015).

3.4. Promoción (ahora Comunicación)

Para los profesionales de la mercadotecnia, las herramientas anteriormente citadas: el producto, el precio y la distribución, son tan importantes como la promoción pero cuando se nombra el marketing, el público generalmente lo identifica con la publicidad. Lo cierto es que también es la comunicación de marketing el campo en el que el uso del Big Data resulta más evidente para el gran público, tal vez porque es en donde más rápido se empezó a utilizar o quizás, es que, por su propia naturaleza, es lo que más llega a los consumidores, a fin y al cabo, llegar a los posibles clientes es su máximo objetivo.

La comunicación de marketing abarca una serie de técnicas o herramientas como son:

- La publicidad,
- La venta directa,
- La promoción de ventas,
- El patrocinio,
- Las relaciones públicas o
- El marketing directo

Todas ellas, en mayor o menor medida, son susceptibles de mejorar mediante el uso del Big Data, en parte porque en el proceso de planificación de la comunicación de marketing hay actividades que se aprovecharían del análisis de datos masivos, como por ejemplo en la determinación del público objetivo.

Un ejemplo de ello lo tenemos con el estudio llevado a cabo en Girona con la participación de Telefónica Digital Insights, en dicho estudio se evaluaba el perfil de los turistas que visitaban la ciudad durante el evento conocido como "Temps

de Flors" [15], obteniendo datos relativos a su procedencia, tanto a nivel local, como nacional o internacional, género o edad así como las franjas horarias que se ligaban a su actividad turística en la localidad (Lorente, 2015). Además se pueden comparar los perfiles obtenidos con estudios similares en otras ciudades de manera que se identifiquen las variables que hacen que los turistas se decanten por una u otra.

Con esa información, que se enclava en lo que se denomina geomarketing, se puede mejorar la oferta así como orientarla de manera más eficiente.

Otra de las cuestiones que se benefician del Big Data es la posible reducción de costes lo que influye en el presupuesto que se puede dedicar al conjunto de acciones de comunicación. De igual modo, el análisis de resultados, con las funciones de control que conlleva puede, en ciertas circunstancias realizarse mediante Big Data.

Pero, sin duda alguna, los ámbitos en los que encontramos más ejemplos exitosos de comunicación de marketing basada en Big Data son el marketing directo y las promociones de venta.

En el primero tenemos un claro campeón, Amazon [16] con su sistema de recomendaciones de productos, con el que consigue aproximadamente un tercio de toda su facturación (Mayer-Schönberger & Cukier, 2013). Un sistema

[15] https://youtu.be/XExDvHUOHPA

[16] http://www.amazon.com/

que debido a su eficacia está siendo copiado por todos aquellos que cuentan con la capacidad tecnológica suficiente para hacerlo (Harris, 2013).

El sistema funciona recomendando a los clientes artículos que, a su vez otros clientes han comprado al adquirir el producto que el cliente está mirando. No se para en saber el porqué de dichas relaciones, simplemente sabe que esas compras conjuntas han ocurrido con antelación y las muestra a sus potenciales clientes.

En ocasiones la relación parece obvia, como en este caso, y es probable que alguien que busque una tostadora, un producto clásico para elaborar un desayuno, se anime a comprar también un hervidor de agua.

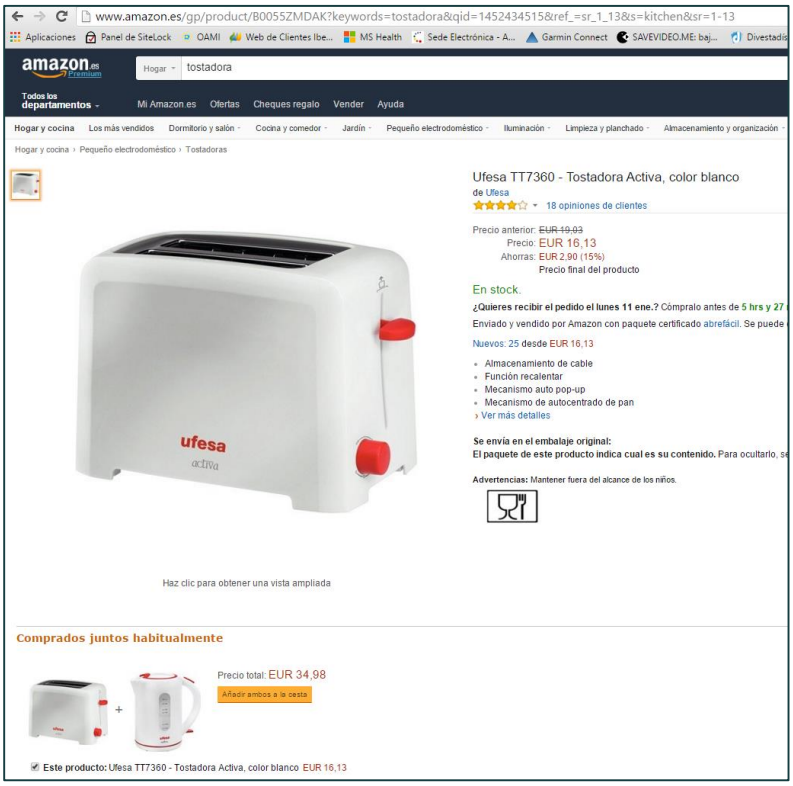


ILUSTRACIÓN 23. EJEMPLO DE RECOMENDACIÓN DE AMAZON

Pero en otros casos el vínculo no resulta tan claro, por ejemplo, es más extraña la posible conexión entre la compra de una flauta dulce con la compra de pinzas de la ropa.

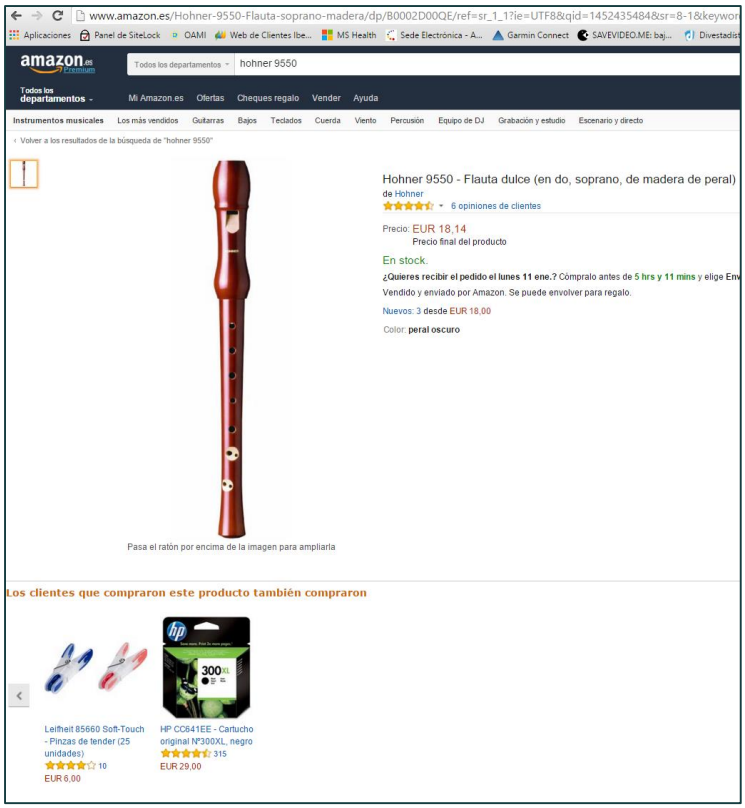


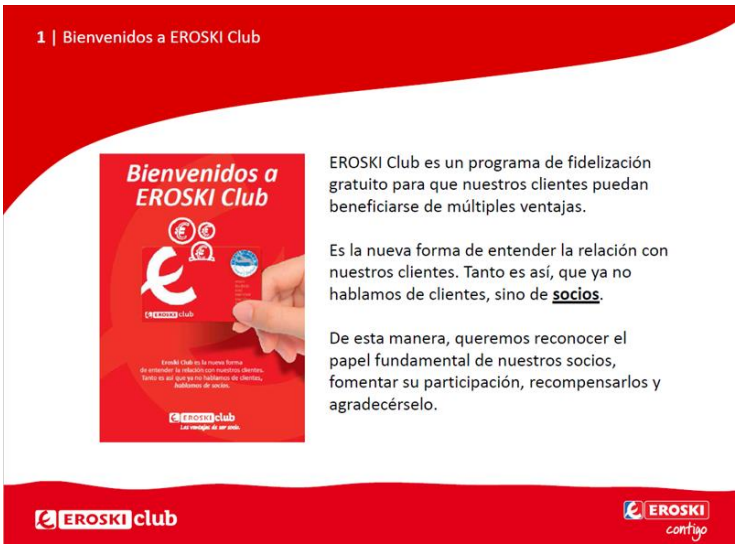
ILUSTRACIÓN 24. EJEMPLO DE RECOMENDACIÓN DE AMAZON

No obstante, y en contra de lo que pueda dictar nuestra lógica, funciona, y funciona maravillosamente bien.

En lo que se refiere a las promociones de venta podemos ver un ejemplo, en Eroski [17], la cadena de supermercados realiza promociones de venta personalizadas a los clientes que poseen su tarjeta de fidelización y para ello utilizan técnicas de Big Data. Han llegado a imprimir folletos personalizados con las fotos de sus clientes y enviárselos a sus domicilios o correos electrónicos. Cada quincena lanzan un total de 30 000 promociones (Larrakoetxea, 2016). Y es en esa velocidad en donde radica la diferencia, hace 10 años, El Corte Inglés ya hacía llegar folletos de manera diferenciada a sus clientes, si comprabas alimento para gatos con su tarjeta de compra, al cabo de uno o varios meses te llegaba un folleto denominado mascotas, pero se necesitaba como mínimo un

[17] http://www.eroski.es/

mes para procesar los datos obtenidos de las tarjetas y emitir las promociones. Ahora sobre la marcha, normalmente al efectuar la compra, se imprime el ticket de promoción, en ocasiones es algo genérico pero en otros casos se hace teniendo en cuenta la compra efectuada.



4. El problema del *Dirty Data*

En el mundo de los datos el *Dirty Data* (o "datos sucios") se define como los datos inadecuados, erróneos o duplicados que se encuentran almacenados en las bases de datos.

Como se ha señalado cuando se hablaba de la tecnología, existen dos tipos de bases de datos, las SQL o relacionales y las NoSQL, que son las que se utilizan para el Big Data.

Las primeras, las SQL, necesitan asegurar la integridad de los datos que se introducen, es decir, que los datos sean correctos, para ello se valen de una serie de reglas denominadas restricciones de integridad que, en un principio, se establecen para mantener la base de datos sin errores. Eso implica que, si está bien diseñada, los únicos datos sucios que pueden contener son los que se introdujeron con algún tipo de error, por ejemplo un número de teléfono equivocado; los que dejan de ser válidos debido a la obsolescencia, por ejemplo un número de hijos o nivel de ingresos referido a años anteriores (es necesario hacer la salvedad de que según el propósito del proyecto ese tipo de datos pueden continuar siendo válidos); o bien datos no adecuados debido a que, de manera voluntaria, se mintió al darlos (Rivero Cornelio, Martínez Fuentes, & Alonso Martínez, 2005).

En el segundo tipo de base de datos, por el contrario, no se cuenta con ningún tipo de restricción y eso es necesario ya que en los proyectos de Big Data se toman datos de diferentes fuentes, con diferentes formatos y se mezclan para después ser tratados (Vaish, 2013).

De lo anterior se deduce con facilidad que, cuando se trabaja con datos masivos, es decir con bases de datos NoSQL, resulta algo muy común que existan duplicidades o datos contradictorios que se agregan desde distintas bases de

datos, además de todos los tipos de datos sucios que también se encuentran en las SQL.

Existe mucha controversia sobre la necesidad o no de limpiar las bases de datos de estos *dirty data*, hay quienes dicen que para ajustar los resultados es necesario procesar previamente los datos para asegurar su idoneidad y que eso resulta esencial para el negocio. Ya en 2007 Gartner[18], la empresa líder en consultoría tecnológica, alertaba de que el 25% de los datos críticos para el negocio de las primeras 1000 empresas en la lista Forbes eran fallidos (Moore, 2007).

FUENTE: WWW.FUNNELHOLIC.COM

Por otra parte, como ya comentamos en el apartado denominado Datos "confusos", hay quienes mantienen que el tratar con tal cantidad de datos permite que haya datos incorrectos ya que se han eliminado los errores de muestreo (Mayer-Schoïnberger & Cukier, 2013).

En todo caso, existen dos aproximaciones al problema de limpiar los datos: la primera consiste en poner un especial cuidado en el diseño del proyecto para que los datos que se agreguen a la base de datos desde distintas fuentes sean lo más consistentes que sea posible, en muchos casos esas fuentes son bases de datos SQL y también necesitan ser limpiadas previamente a su inclusión (Rahm & Hong-Hai, 2014); la segunda, se basa en limpiar los datos *a posteriori*, dado que se trata de volúmenes ingentes para hacer ese trabajo se utilizan algoritmos y es ese punto en el que hay que prestar atención ya que es fácil que se introduzcan sesgos no deseados al hacerlo.

[18] http://www.gartner.com/

Lo que queda claro es que la decisión de limpiar o no los datos, depende del propósito para el que vayan a ser usados, no es lo mismo utilizar Big Data para realizar pronósticos personalizados de cáncer de mama que utilizarlo para hacer llegar publicidad a millones de personas de manera más o menos personalizada; en el primer caso, errores como la edad o la condición de fumadora en los datos recogidos puede conllevar resultados nefastos.

"Some errors deserve priority, but which ones are most important is highly study-specific. In most clinical epidemiological studies, errors that need to be cleaned, at all costs, include missing sex, sex misspecification, birth date or examination date errors, duplications or merging of records, and biologically impossible results. For example, in nutrition studies, date errors lead to age errors, which in turn lead to errors in weight-for-age scoring and, further, to misclassification of subjects as under- or overweight"[19] (Van den Broeck, Argeseanu Cunningham, Eeckels, & Herbst, 2005)

Mientras que en el caso de la publicidad, una rápida evaluación coste-beneficio muestra que lo más probable es que no sea necesario invertir demasiado en la limpieza de los datos, por ejemplo, la cadena de supermercados Target[20], envió publicidad sobre artículos de bebé a una adolescente lo que hizo que su airado padre elevara una protesta ya que, según él, su hija no estaba embarazada, a pesar de que después se comprobó que, de hecho, lo estaba, se puede considerar un error no haber tenido en cuenta la edad de la receptora de la publicidad, quizás ni siquiera estaba en su base de datos o era incorrecta, pero

[19] "Algunos errores merecen ser tratados de forma prioritaria, pero cuáles son los más importantes depende mucho del estudio del que se trate. En la mayoría de los estudios clínicos epidemiológicos, los errores que a toda costa se han de limpiar incluyen el sexo no especificado o erróneo, errores en la fecha de nacimiento o fecha del examen, duplicaciones o mezcla de registros y resultados biológicamente imposibles. Por ejemplo, en estudios relacionados con la nutrición, errores en las fechas conducen a errores en las edades que a su vez llevan a errores en ratios de edad-peso y en última instancia a clasificar de forma errónea a los sujetos como con peso demasiado bajo o alto"

[20] http://intl.target.com/

un caso aislado no justifica el coste de revisar los datos que Target utiliza para enviar ese tipo de publicidad o promociones (Hill, 2016).

4.1. El porqué de las mentiras

Un caso que merece una especial atención es el de cuando los ciudadanos mienten conscientemente al proporcionar los datos que les son solicitados.

Hay diferentes estudios al respecto que identifican los motivos más habituales, a saber:

- Para proteger su privacidad.
- Para evitar la publicidad online.
- Por miedo a cómo serán utilizados esos datos.
- Por considerar no adecuados los datos solicitados con respecto a lo que desea obtener el usuario.

La privacidad es una de las preocupaciones en alza en un mundo digital y la globalidad de la red hace que la protección legal sea insuficiente en muchos casos.

Por otro lado, el spam es una de las acciones que los usuarios refieren normalmente entre las más molestas y es normal que se trate de evitar.

Como hemos visto anteriormente, la capacidad de las empresas y de la administración para usar los datos hacen que las personas se sientan vulnerables ante determinadas prácticas, por ejemplo con casos como el de aumento o denegación de seguros de vida por la información personal que las compañías pueden encontrar en la red.

Por último, cada vez es más frecuente que para acceder a determinados contenidos o realizar compras, el proceso de registro incluya la obligación de aportar datos personales que no parecen estar acordes con el objetivo del usuario, si considera que lo que va a obtener no justifica esa petición, miente.

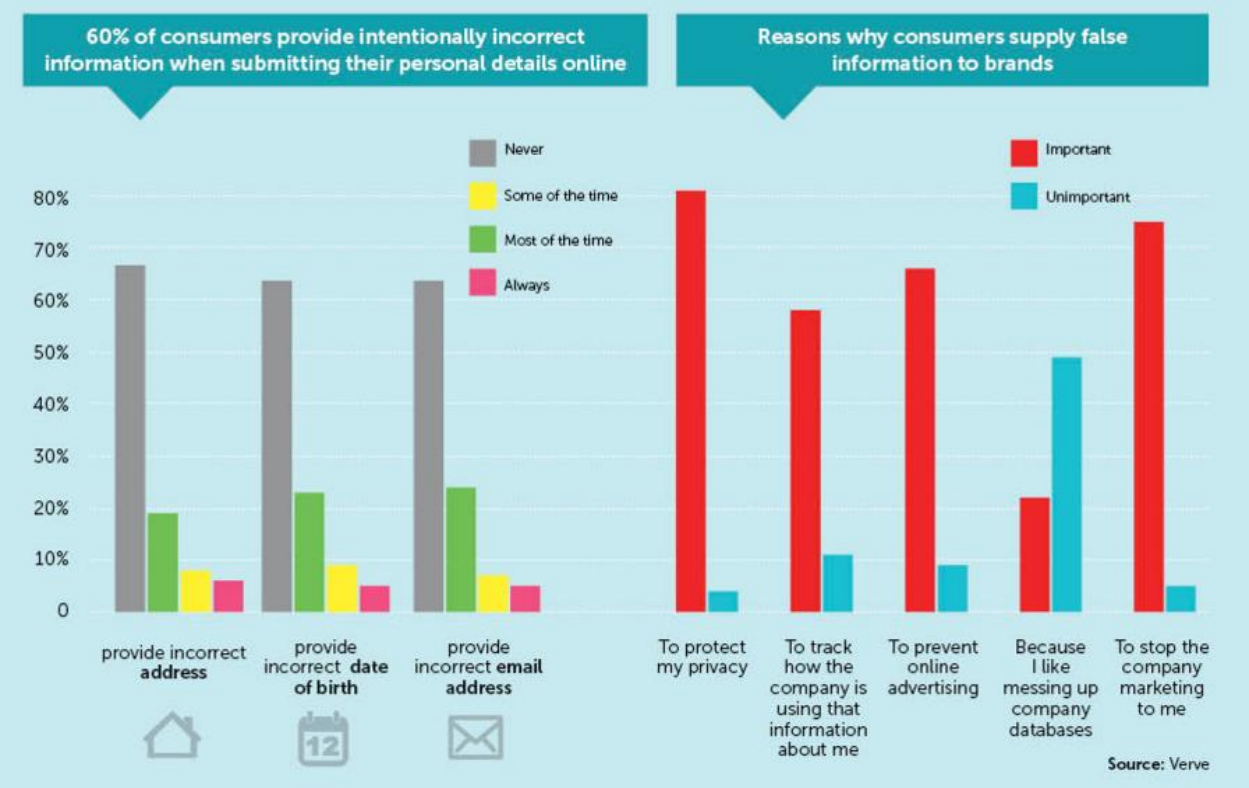


ILUSTRACIÓN 25. DATOS DEL ESTUDIO DE VERNE SOBRE LA POBLACIÓN DE REINO UNIDO

Los datos que con mayor frecuencia se falsean son (Chahal, 2015):

- La dirección
- La edad
- El correo electrónico

Lo que resulta consistente con los motivos para mentir anteriormente citados.

Un motivo no reseñado que genera *dirty data* es el sentirse observado, un ejemplo llamativo de ello fue el fracaso de Flu Trends, de Google[21]. Fu Trends, permitía seguir la evolución de la gripe en más de 25 países, se basaba en el análisis de los términos de búsqueda lo que proporcionaba información de manera considerablemente más rápida que lo que obtenía el Centro de Control de Enfermedades, CDC[22]. Los dos primeros años los datos que obtenían eran muy

[21] https://www.google.com
[22] http://www.cdc.gov/

precisos pero a partir del tercero, Google sobreestimo sus evaluaciones en un factor 2. Dado el secretismo con el que guardan sus algoritmos es difícil asegurar los motivos por los que se produjo ese resultado pero uno de los más factibles es que se produjo un cambio en las formas de búsqueda de los usuarios que Google no supo estimar correctamente (Lazer & Kennedy, 2015).

4.2. El valor de los *dirty data*

No se puede dejar de hacer mención al hecho de que en ocasiones esos datos que se consideran erróneos o incoherentes tienen valor en sí mismos.

A veces, se consideran erróneos valores estadísticamente extremos pero, realmente no lo son, se pueden usar para identificar casos únicos pero relevantes o para evaluar características que si bien se salen de la norma pueden ser útiles (Van den Broeck et al., 2005).

Ese es también el caso de los datos en los que los usuarios mienten o mienten aparentemente, por ejemplo:

"…suppose you have a patient who regularly tweets about sitting around the house, but also retweets exercise tips almost every week. That could mean that the patient wants to exercise, but has a hard time making a lasting commitment. Clean data would say that she leads a sedentary life, and disregard the exercise tips as a minor contradiction to the data."[23] (Spearman, 2015)

Uno de los ejemplos más exitosos del uso de datos sucios es cómo Google recopiló los errores de los usuarios al introducir búsquedas para crear y posteriormente refinar su corrector ortográfico. Ahora cuando se realizan

[23] "...supongamos que hay una paciente que tuitea habitualmente que está sentada en su casa, pero que también retuitea consejos sobre ejercicio casi todas las semanas. Esto podría significar que a la paciente le gustaría hacer ejercicio pero que le cuesta comprometerse a ello. Los datos limpios dirían que lleva una vida sedentaria y desecharían los consejos sobre el ejercicio como una contradicción menor en los datos."

búsquedas que incluyen palabras mal escritas Google es capaz de sugerir la búsqueda correcta.

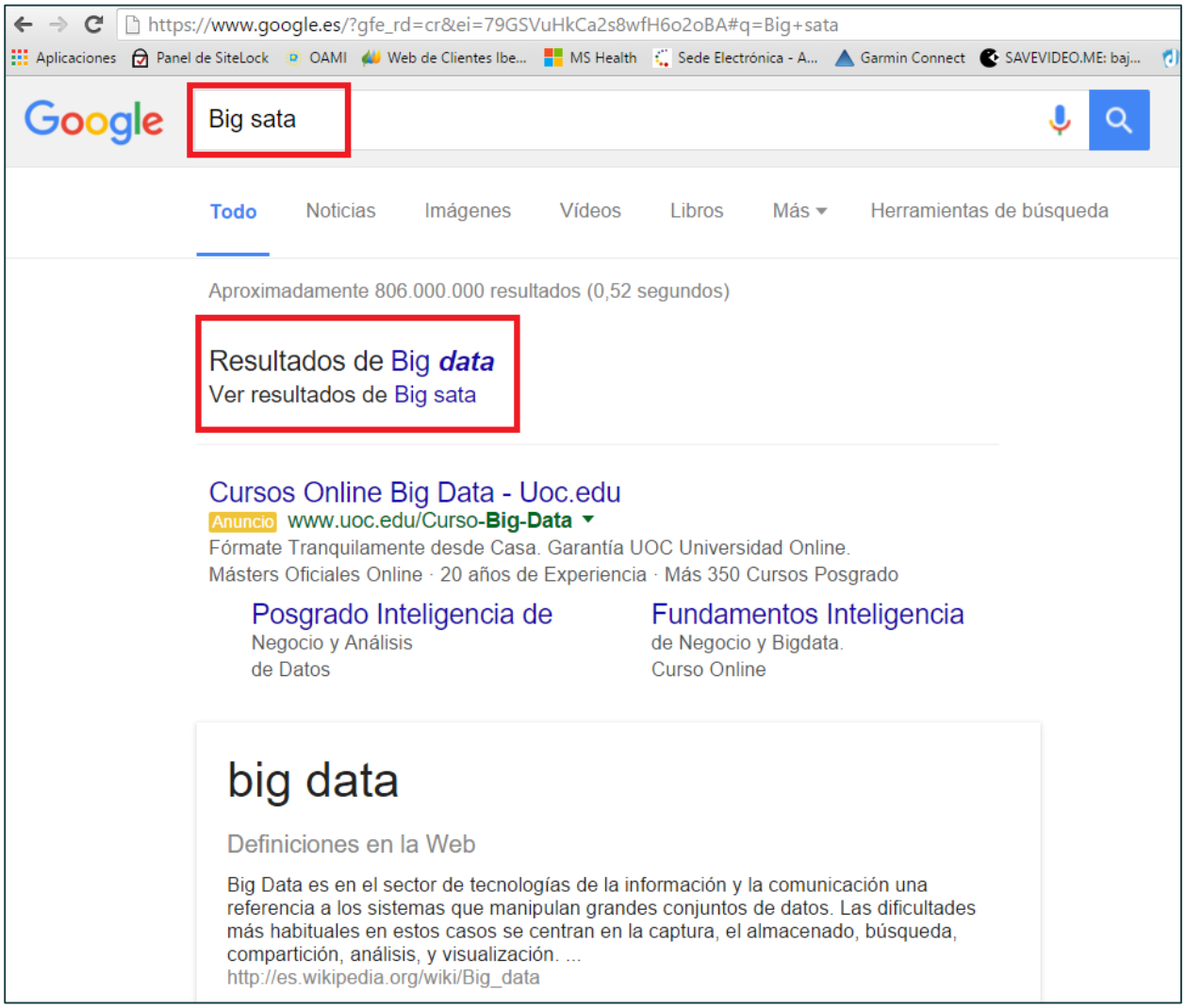


ILUSTRACIÓN 26. BÚSQUEDA ERRÓNEA EN GOOGLE Y SUGERENCIA DE LA CORRECTA

5. Conclusiones

Lo primero que es necesario decir es que el uso del Big Data, tanto en marketing como en cualquier otro ámbito, no es una cuestión de futuro sino de presente. El Big Data se encuentra ya involucrado en aspectos cotidianos de toda la sociedad. No cabe duda de que en los próximos años su utilización sufrirá un incremento considerable y se hará más presente para el público en general. Se abrirán nuevas oportunidades que aún no llegamos a ver pero que, en parte, modificarán nuestra visión sobre el mundo y su funcionamiento.

Su utilidad en lo referido al marketing es enorme, se ha puesto de manifiesto a lo largo de este trabajo que los datos masivos pueden aportar valor en cada punto del proceso de marketing, en cada actividad que los departamentos responsables desarrollan, no sólo en lo más evidente como es la publicidad sino en la producción, la estrategia de precios o la cadena de distribución.

No se pueden obviar los problemas que pueden aparecer, sobre todo en materia de privacidad, por ello es necesario, no sólo que los gobiernos ejerzan una labor de protección de sus ciudadanos sino que éstos sean conscientes de sus acciones y de las consecuencias de las mismas. Es imprescindible que se eduque a la sociedad en su conjunto en el buen uso de la red y en el cuidado que se ha de tener cuando se proporciona a terceros datos personales de todo tipo. Además se debe mantener una actitud vigilante para evitar la discriminación que puede aparecer con un uso sesgado de los datos, sobre todo en cuestiones de género, de procedencia, etnia, religión, estrato socioeconómico, etc.

Los ciudadanos que ya son conscientes de dichos problemas son los artífices de que las bases de datos masivos se estén poblando de datos falsos, en parte porque no se ha implantado un modo sencillo de que has empresas que los recaban se comuniquen con sus usuarios para evaluar cuáles, de esos datos, hacen que los consumidores tengan reticencias a la hora de darlos y trabajar

conjuntamente en limitar o modificar esas peticiones. Por ello, aunque no es el único motivo, el análisis de los datos sucios no es algo que se deba soslayar. Ese estudio puede aportar información muy relevante para que las empresas adapten sus procedimientos en pos de lo mejor para el conjunto de la sociedad y no sólo mirando por un beneficio estrictamente económico.

Por último, considero necesario recalcar que los datos, por muchos que sean, no dan la solución a los problemas, la solución viene de, basándose en esos datos, realizar el mejor análisis y para ello es imprescindible el conocimiento, la experiencia y la inteligencia de quienes han de tomar las decisiones.

6. Bibliografía

Ammetller Montes, G., Rodríguez Ardura, I., & Universitat Oberta de Catalunya. (2009). *Direcció de màrqueting* (5a ed ed.). Barcelona: UOC, Universitat Oberta de Catalunya.

Berger, J. M., & Morgan, J. (2015). The ISIS Twitter Consensus. from http://www.brookings.edu/~/media/Research/Files/Papers/2015/03/isis-twitter-census-berger-morgan/isis_twitter_census_berger_morgan.pdf?la=en

Boada, A. J., & Mayorca, R. (2011, 07/2011). Planificación de demanda, en empresas con estilo de venta por catálogo. *Rev. Lasallista Investig.*, 8, from http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S1794-44492011000200014&lng=en&nrm=iso&tlng=es

Borden, N. H. (1984). The Concept of the Marketing Mix. from http://www.guillaumenicaise.com/wp-content/uploads/2013/10/Borden-1984_The-concept-of-marketing-mix.pdf

BusinessIntelligence.com. (2015). How UPS Uses Big Data With Every Delivery - Business Intelligence. from http://businessintelligence.com/big-data-case-studies/ups-uses-big-data-every-delivery/

Cazón, P. (2014). Mono Burgos: "Las Google Glass son muy útiles para el técnico" | Liga BBVA | AS.com. *as.com*. from http://futbol.as.com/futbol/2014/04/13/primera/1397409854_496947.html

Chahal, M. (2015). Consumers are 'dirtying' databases with false details. from http://www.marketingweek.com/2015/07/08/consumers-are-dirtying-databases-with-false-details/

Del Rey, J. (2012). British Airways: Future of Airline Travel Is Data-Centric Personalization. from http://adage.com/article/special-report-digital-conference-san-francisco-2012/british-airways-future-data-centric-personalization/237357/

Dotson, K. (2012). Hurricane Sandy and the Big Data of Disaster Prediction. from http://siliconangle.com/blog/2012/10/26/hurricane-sandy-and-the-big-data-of-disaster-prediction/

Elliot, T. (2007). The Real Pioneer of Business Intelligence (and BI 2.0)? , from http://timoelliott.com/blog/2007/11/the_real_pioneer_of_business_i.html

García Barbosa, J. (2014). La medicina del futuro pasa por big data. from http://www.aunclicdelastic.com/la-medicina-del-futuro-pasa-por-big-data/

Gittleson, K. (2013). How big data is changing the cost of insurance - BBC News. from http://www.bbc.com/news/business-24941415

Goldin, R. (2015, 2015-08-19). Causation vs Correlation. from http://www.stats.org/causation-vs-correlation/

Harris, D. (2013, 2013-01-29). You might also like … to know how online recommendations work | Gigaom. from https://gigaom.com/2013/01/29/you-might-also-like-to-know-how-online-recommendations-work/

Hegde, H. (2014, 2014-09-22). Netflix and its Revolutionary Use of Big Data. from http://dataconomy.com/netflix-and-its-revolutionary-use-of-big-data/

Heraclito, & Aguirre, R. G. (1956). *Heráclito de Efeso: fragmentos*. Buenos Aires ,.

Hill, K. (2016). How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did - Forbes. from http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

IBM, I. f. B. V., & Oxford, E. d. N. S. e. l. U. d. (2012). Analytics: el uso de big data en el mundo real. from http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf

Jarreño, P. (2014, 2014-02-24). HouseHouse of Cards, Netflix, el huevo y la gallina. from https://www.territoriocreativo.es/etc/2014/02/house-of-cards-netflix-el-huevo-y-la-gallina.html

Jiménez, R. (2014). Uber se disculpa por subir precios durante la toma de rehenes en Sidney. from http://www.qore.com/noticias/32109/Uber-se-disculpa-por-subir-precios-durante-la-toma-de-rehenes-en-Sidney

Kotler, P., & Armstrong, G. (2003). *Fundamentos de marketing* (6a. ed ed.). México etc: Pearson Education.

Kotler, P., Armstrong, G., Moreno López, Y., García de Madariaga Miranda, J., Flores Zamora, J. d. J., & Galli, C. A. (2008). *Principios de marketing : duodécima edición*. Madrid: Pearson Educación.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Larrakoetxea, C. (2016). El 'big data' permite a Eroski lanzar 30.000 promociones por quincena, casi todas diferentes - elEconomista.es. from http://www.eleconomista.es/pais_vasco/noticias/7217840/12/15/El-big-data-permite-a-Eroski-lanzar-30000-promociones-por-quincena-casi-todas-diferentes.html

Lazer, D., & Kennedy, R. (2015). What We Can Learn From the Epic Failure of Google Flu Trends. from http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Loeb, W. (2014). Amazon's Pricing Strategy Makes Life Miserable For The Competition - Forbes. from http://www.forbes.com/sites/walterloeb/2014/11/20/amazons-pricing-strategy-makes-life-miserable-for-the-competition/

Lorente, A. (2015, 2015-12-21). El Big Data ayuda a entender mejor el sector turístico. from http://blogthinkbig.com/el-big-data-ayuda-a-entender-mejor-el-sector-turistico/

Loukil, R. (2015). Air France KLM anticipe les pannes de ses avions A380 au big data. from http://www.usine-digitale.fr/article/air-france-klm-anticipe-les-pannes-de-ses-avions-a380-au-big-data.N365483

Lunden, I. (2012, 2012-10-08 16:04:27). Telefonica Wants To Turn Its Mobile Data Into A Big Data Business, Launches Dynamic Insights Unit. from http://social.techcrunch.com/2012/10/08/telefonica-wants-to-turn-its-mobile-data-into-a-big-data-business-launches-dynamic-insights-unit/

Magazin, M. (2015). Big Data Lufthansa will Kundendaten zu Geld machen - manager magazin. from http://www.manager-magazin.de/unternehmen/artikel/big-data-lufthansa-will-kundendaten-zu-geld-machen-a-1038737.html

Magrama. (2015). La desertificación en España - Desertificación y restauración forestal - Política forestal - Desarrollo Rural - magrama.es. from http://www.magrama.gob.es/es/desarrollo-rural/temas/politica-forestal/desertificacion-restauracion-forestal/lucha-contra-la-desertificacion/lch_espana.aspx

Mashey, J. (1998). Big Data and the Next Wave of Infrastress. from http://usenix.org/legacy/publications/library/proceedings/usenix99/invited_talks/mashey.pdf

Mattioli, D. (2012). On Orbitz, Mac Users Steered to Pricier Hotels. from http://www.wsj.com/articles/SB10001424052702304458604577488822667325882

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data : a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

McCarthy, E. J. (1964). *Basic marketing*. Homewood, Ill.,: R. D. Irwin.

Meek, T. (2014). PTCVoice: In-Car Sensors Put Insurers In The Driver's Seat - Forbes. from http://www.forbes.com/sites/ptc/2014/06/27/in-car-sensors-put-insurers-in-the-drivers-seat/

MLB. (2015). Estadísticas de la Grandes Ligas MLB. from http://mlb.mlb.com/stats/sortable_es.jsp

Moore, S. (2007). 'Dirty Data' is a Business Problem, Not an IT Problem, Says Gartner. from http://www.gartner.com/newsroom/id/501733

Niño, M. (2015). MapReduce: el origen de la era Big Data | Blog de Mikel NiÃ±o: Emprendimiento digital, startups, Big Data Analytics y nuevos modelos de negocio. from http://www.mikelnino.com/2015/02/map-reduce-origen-era-big-data.html

O'Reilly, T. (2015, 2015-09-15). Improving Uber's Surge Pricing — What's The Future of Work? — Medium. from http://medium.stfi.re/the-wtf-economy/improving-uber-s-surge-pricing-3fd2fe108bd6?sf=jxoyko

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? - American Journal of Infection Control. from http://www.ajicjournal.org/article/S0196-6553(15)00137-6/abstract

Parra, S. (2014). Los precios de un vuelo pueden cambiar siete mil millones de veces en una hora. from http://www.yorokobu.es/precios-de-vuelos/

Porter, E. (2011). *Todo tiene un precio : descubre que el valor de las cosas afecta al modo en que nos enamoramos, trabajamos, vivimos y morimos : los precios escriben la historia* (1*® ed.). Madrid: Aguilar.

Rahm, E., & Hong-Hai, D. (2014). Data Cleaning: Problems and Current Approaches. from http://betterevaluation.org/sites/default/files/data_cleaning.pdf

Rivero Cornelio, E., Martínez Fuentes, L., & Alonso Martínez, I. (2005). *Bases de datos relacionales : fundamentos y diseño lógico*. Madrid: Universidad Pontificia de Comillas.

Rodríguez Ardura, I., & Universitat Oberta de Catalunya. (2013). *Principios y estrategias de marketing* (pp. 1 recurs electrònic (455 p.)).

Soubra, D. (2012). The 3Vs that define Big Data. from http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

Spearman, S. (2015, 2015-07-27). In Defense of Dirty Data in Healthcare - Health Security Solutions. from http://www.healthsecuritysolutions.com/2015/07/27/in-defense-of-dirty-data-in-healthcare/

Take ad way. (2011). ¿Cómo será el crecimiento del Vídeo Online desde ahora y hasta 2015? , from http://www.video-on-life.com/index.php/crecimiento-video-online-2015/

TDI. (2013, 2013-08-16). Smart Steps - ES | Telefonica Dynamic Insights. from http://dynamicinsights.telefonica.com/es/488/smart-steps-2

Turck, M. (2014). The state of big data in 2014 (chart). from http://venturebeat.com/2014/05/11/the-state-of-big-data-in-2014-chart/

tylervigen.com. (2010). 15 Insane Things That Correlate With Each Other. from http://tylervigen.com/spurious-correlations

UOC. (2015). Miriada X: Introducción al Business Intelligence. Módulo 5. from https://miriadax.net/es_ES/web/introduccion-al-business-intelligence/inicio

Vaish, G. (2013). *Getting started with NoSQL : your guide to the world and technology of NoSQL* (pp. iii, 123 p.).

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005, Oct). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. *PLoS Med.* 2, from http://dx.doi.org/10.1371/journal.pmed.0020267

Wernicke, S. (2015). How to use data to make a hit TV show.

Wheatley, M. (2013). Big Data Goes Green: How Data Analytics Is Saving the World's Forests. from http://siliconangle.com/blog/2013/07/02/big-data-goes-green-how-data-analytics-is-saving-the-worlds-forests/