GRADO EN COMERCIO

TRABAJO FIN DE GRADO

“Big data: Explicación, principales herramientas y su aplicación en Amazon”

Miguel García Vázquez

FACULTAD DE COMERCIOVALLADOLID,

Fecha: Junio 2023

FACULTAD DE COMERCIO
Universidad de Valladolid

UNIVERSIDAD DE VALLADOLID

GRADO EN COMERCIO

Curso académico 2022/2023

TRABAJO FIN DE GRADO

“Big data. Explicación, principales herramientas y su aplicación en Amazon”

Trabajo presentado por: Miguel García Vázquez

Tutor: María Sol Velasco Sacristán

FACULTAD DE COMERCIO

Valladolid, Junio 2023

Índice de contenidos

Índice de ilustraciones

Lista de acrónimos

1. IA: Inteligencia artificial
2. MIT: Massachusetts Institute of Technology/ Instituto de Tecnología de Massachusetts
3. IBM: International Business Machines
4. GPS: Global Positioning System/Sistema de Posicionamiento Global
5. GDPR: General Data Protection Regulation /Reglamento General de Protección de Datos
6. KPI: Key Performance Indicators/Indicadores Clave de Rendimiento

Resumen

En el trabajo he hablado del significado del Big data, no solamente para las empresas sino qué opinan las organizaciones y diversas áreas económicas del mismo, como lo interpretan desde su posición, porque esté debería de importarnos y que peligros para los consumidores puede acabar acarreando sino se gestiona correctamente.

He hablado sobre la forma y origen de los datos, cómo son clasificado dentro del Big data y dónde encontramos dichos datos almacenados. Con un enfoque centrado en Amazon, he hablado sobre las principales técnicas de análisis de datos, incluyendo el análisis web.

He explicado algunas de las principales formas de análisis de las empresas, por qué son tan importantes estos métodos, qué ventajas les ofrecen a estas, y cómo se pueden aplicar al mundo real.

Finalmente, he explicado sobre Amazon y su proceso de compra, cómo captura y qué datos le interesa obtener.

Palabras clave: Big data, datos, Amazon, Web, análisis, organizaciones, información.

Abstract

In the paper I have talked about the meaning of Big data, not only for companies but also what organizations and different economic areas think about it, how they interpret it from their position, why it should matter to us and what dangers it can end up bringing to consumers if it is not managed correctly.

I have talked about the form and origin of the data, how it is classified within Big data and where we find this data stored. With a focus on Amazon, I have talked about the main data analysis techniques, including web analytics.

I have explained some of the main forms of analytics for companies, why these methods are so important, what advantages they offer to companies, and how they can be applied to the real world.

Finally, I have explained about Amazon and its buying process, how it captures and what data it is interested in obtaining.

Keywords: Big data, data, Amazon, Web, analytics, organizations, information.

1. Introducción

1.1 Justificación

La principal razón por la cual decidí elegir el tema del "Big data", sería debido a que pese a ser un tema, que como veremos en este trabajo, no es tan nuevo como podría parecer, todavía resulta novedoso y refrescante al estar constantemente surgiendo diferentes tecnologías y técnicas, evolucionando e incluso resurgiendo con el pasar del tiempo, como puede ser el resurgimiento en 2022 de las IA.

Esta es una de las razones, por las cuales, "Big data" es un campo tan rico y con tanta variedad de puntos de vista, todos válidos dependiendo de quien lo cuente. Mi objetivo es centrarme en aquellos aspectos que son más usados por las empresas, bases de datos, análisis avanzados de los datos y forma de gestionar el "Big data" haciendo énfasis en las técnicas y herramientas más utilizadas por las organizaciones, que centran sus actividades en el ámbito online. Otra razón por la cual he querido realizar este trabajo, es para aumentar mi propio conocimiento dentro de este tema, dado que muchos de los autores y entidades que veremos en el trabajo, Bernard Marr, Luis Aguilar Joyanes, *MIT*, *IBM*… predicen un aumento sustancial en los empleos relacionados con los datos y en la importancia de estos en el ámbito empresarial.

1.2 Objetivos:

La cantidad de datos que generamos y destruimos en cada momento es gigantesca, tal es este volumen que resulta casi imposible de gestionar sin herramientas especializadas. Esto acaba resultando la razón, por la cual muchas empresas han querido aprovechar y gestionar este flujo de datos, para convertirlo en conocimiento útil que le servirá para la toma de decisiones dentro de la empresa. De las cuales podemos destacar, desde *Uber* hasta *Google*

Mi objetivo principal, se podría resumir de la siguiente manera, comprender el "Big data", pasando por sus principales técnicas y herramientas

Asimismo como objetivos más específicos serán:

- Entender el concepto de “Big data”, sus diferentes definiciones.
- Comprender qué son los datos y cuáles trata el “Big data”.
- Identificar las principales herramientas dentro del “Big data”.
- Comprender cómo se analizan los datos para obtener valor de los mismos.
- Analizar el uso del “Big data” por parte de Amazon y cómo afecta a su empresa.

1.3 Metodología

En primer lugar, se va a realizar una búsqueda y revisión de la bibliografía pertinente, incluyendo libros y trabajos escritos sobre el tema a tratar, los cuales mencionarán y después realizar un análisis cualitativo de las lecturas y conocimientos obtenidos, todo ello realizado con cohesión y coherencia.

Esto se va a realizar con el objetivo de entender y aumentar mi conocimiento sobre “Big data”, comprender en más profundidad el tema, su teoría, las tecnologías y las metodologías más importantes. Seguidamente identificaré aquellas áreas que con el marco teórico de comercio, considero que son más importantes para comprender el tema.

El análisis cualitativo, se va centrar en su mayoría en explicar, el concepto de “Big data” y sus principales herramientas y técnicas utilizadas. Con esto pretendo explicar tanto los conceptos básicos del tema, como su importancia y por qué la gente está usando estas herramientas y sus principales razones.

Finalmente tomaré el ejemplo de una empresa que ha utilizado intensamente el “Big data”: *Amazon*, tanto en su Marketplace, como en su decisión de abrir un proveedor de servicios web centrados principalmente, en “Big data” y bases de datos.

2. Definición de Big data

Antes de hablar del tema, de dónde proviene el término “Big data”, o datos masivos, normalmente el origen del mismo se le atribuye a Francis Dievold, argumentando que surgió a mediados de la década de 1990 en conversaciones informales de la empresa *Silicon Graphics*. Sin embargo, en su documento (Diebold, 2012) el mismo argumenta que si bien fue de los primeros en mencionarlo, el concepto comenzó a surgir a finales de los noventa por parte de diversos autores, los cuales simplemente observaban el entorno añadiendo que fue (Laney, 2001) cuando se terminó de materializar el concepto.

Una vez conocemos el término, me gustaría aclarar qué no es el Big data, según *IBM* (IBM, 2017).

- No es solo muchos datos
- No son solo herramientas, técnicas y una serie de tecnologías de la información,
- No es una moda actual
- No es una solución universal a todos los problemas,

Sin embargo, el Big data si es un concepto muy extenso y variado, como veremos en este documento, abarca desde el marketing de las empresas, hasta la investigación genética (Rojas, 2022). Y si bien es cierto que muchas veces es utilizado por las empresas, todo tipo de organizaciones pueden hacer uso de sus capacidades.

Esta heterogeneidad es la razón por la cual existe un debate entre muchos expertos sobre el tema, cada uno cuenta con una o varias versiones de su propio campo sobre que serían los daos masivos. De hecho, es tanto el problema, que diferentes trabajos y encuestas a profesionales han sido realizados para aclarar el problema. Yo mismo he encontrado dos de estas, una realizada por *Harvard* (McAfee & Brynjolfsson, 2012) y la otra por la universidad *Bielefed* en Alemania (Favaretto, 2020).

La segunda, la de la universidad de *Bielefed*, es la más completa y la que más expone el tema. En la misma encontramos multitud de definiciones, dadas en su mayoría por ejecutivos anónimos de sectores de Big data, pero también expertos cómo el ya mencionado Diebold, Kitchin y Merrifield de diversos campos. Sin embargo, considero que la parte más importante de la encuesta, se encuentra en las ideas que surgen de las conclusiones (Favaretto, 2020):

- El Big data es un campo interdisciplinario y requiere de la influencia de varios campos que pueden no estar relacionados con el tuyo, para poder funcionar.
- Es necesaria una definición clara para la creación de leyes que la dirijan.
- El término evoluciona con el tiempo dependiendo d la cultura, lugar y sector donde se encuentre.

Es esta última conclusión la cual considero más importante, el Big data evoluciona con el pasar del tiempo. Para una persona del año 2000, un Terabyte sería considerado datos masivos o algo cercano a estos por esto, dentro de unos años lo que nosotros consideramos datos masivos, seguramente sea la capacidad estándar de almacenamiento. De todas maneras, trataré de dar algunas de las mejores definiciones que podemos encontrar por autores e instituciones.

“The term Big data refers to data so large, fast and complex that it is almost impossible to process with traditional methods.” (SAS, 2023, p 1).

Otra encuesta realizada para la revista *Harvard Business Review* por el *MIT Center for Digital Business,* y *McKinsey’s business technology office* preguntó a 330 ejecutivos de empresas estadounidenses descubriendo que aquellas empresas que se identificaban como dirigidas por los datos, alcanzaron mejor sus objetivos y eran un 6% más productivas y rentables que sus competidores (McAfee & Brynjolfsson, 2012, p 5).

Para las empresas, ser capaz de gestionar y extraer información es considerado una ventaja competitiva clave e incluso un recurso de valor por sí mismo, muchas empresas basan ahora su actividad en su capacidad para recopilar, analizar la información y sacar conclusiones empresariales, entre ellas contamos a gigantes como *Meta* o *Google* (Mayer-Schnberger & Cukier, 2013).

Traducción, (SAS, 2023, p 1) “El termino Big data se refiere a los datos tan grandes, rápidos y complejos que es casi imposible de procesar con los métodos tradicionales”

Finalmente me gustaría mencionar el concepto de datificar, este concepto, en palabras de sus autores, quiere significar lo siguiente "datificar un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado" (Mayer-Schönberger & Cukier, 2013, p 78) esto es en esencia, tomar un fenómeno o evento cualquiera y clasificarlo de tal forma que pueda ser analizado mediante unos parámetros prestablecidos racionales, estos parámetros suelen ser números u otras formas racionales de clasificación.

Un ejemplo que nos da (Mayer-Schönberger & Cukier, 2013), trata sobre Shigeomi Koshimizu, un profesor en Japón, que consiguió datificar la forma de sentarse de las personas, distribuyo varios puntos de presión alrededor del asiento, y dependiendo de cuanta presión era ejercida en estos, daban un número del 1 al 256 en cada uno de los puntos, que seguidamente era asignado a una de las personas del estudio, consiguiendo una identificación eficaz de más de un 95% (Mayer-Schönberger & Cukier, 2013).

Según las herramientas y la forma de ver el Big data se expandan, cambiaran las ideas preconcebidas sobre la experiencia y la administración de las organizaciones. Pero como todos los grandes cambios, esto será algo gradual y que cambiará con el pasar del tiempo y de los sectores, pese a esto no podemos ignorar el efecto que supone en la administración de empresas. (McAfee & Brynjolfsson, 2012).

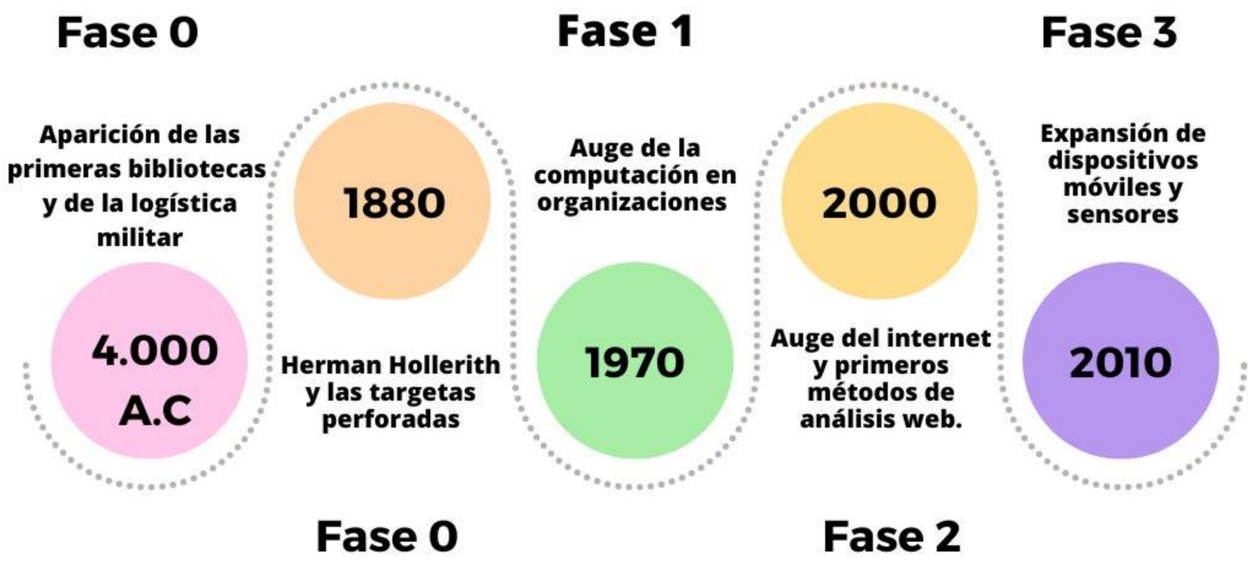
2.1.1 Evolución del Big data

"Big Data basically refers to the fact that we can now collect and analyse data in ways that was simply impossible even a few years ago" (Marr, 2016, p 1).

El Big data se puede dividir en cuatro fases diferenciadas históricamente: **la fase cero**, sería aquellas técnicas antecesoras del Big data**, la fase uno c**on el auge de la computación y las bases de datos sencillas, **la fase dos** en los primeros años de internet con el aumento de usuarios y **la fase tres**, centrada en los dispositivos móviles y sensores (Big data Framework, 2021).

Traducción, (Marr, 2016, pag 1)"Big data es simplemente el hecho que ahora podemos recolectar y analizar los datos de formas que nos resultaban imposibles hace unos años"

Ilustración 1: Línea de tiempo del Big data en fases.



Fuente elaboración propia. Según Big Data Framework, 2021.

Como **la fase cero**, tenemos los primeros usos de bibliotecas aquí se tiene todo el conocimiento de un país en el mismo lugar, la logística militar y civil, con especial auge en Roma (de hecho los romanos fueron de las primeras civilizaciones con estadistas centrados en analizar datos), con la creación de las primeras logísticas avanzadas y métodos de recaudación de impuestos. Por ejemplo, los generales romanos, usaban métodos de análisis sofisticado para predecir cuantos suministros necesitarían en campaña, y cuáles serían los frentes que se encontraban en peligro más inmediato y aquellos otros, que podrían resistir (Big data Framework, 2021).

Finalmente como un hito en **la fase cero**, encontramos al creador de *IBM*, Herman Hollerith, el cual en 1880 ayudaría a la oficina del censo americana, optimizando los procesos utilizados hasta ese momento. Ésta tenía problemas con la rapidez con la que recopilaban y categorizaban la población, tardaban unos ocho años, en terminarlo y este quedaba desactualizado nada más salir de la oficina, esto suponía muchos problemas para el gobierno, dado que no sabían cuánta y como vivía la gente en el gigantesco país por este retraso. Finalmente contrataron a Hollerith en 1880 para que les diese una solución, este utilizo una de las primeras técnicas de computación, tarjetas perforadas las cuales, disminuyeron el tiempo de creación del censo a un año (Lohr, 2012).

La primera fase, la vemos en la década de 1970, con la aparición de las primeras bases de datos, los datos estructurados y los antecesores de las "Warehouses" donde se empezó a usar la computación, de forma más exhaustiva en las organizaciones y en el análisis (Big data Framework, 2021).

La segunda fase ocurre en la década de los 2000s con la normalización del internet, la llegada de la web 2.0 sobre el 2005, significando básicamente un aumento en el contenido creado por los usuarios, por lo tanto vemos que la generación de datos se multiplicó exponencialmente, principalmente por que, hasta ese entonces la mayoría de datos habían sido estructurados y creados por empresas o gobiernos, sin embargo a partir de este punto, los usuarios serán el mayor generador de datos y las organizaciones se verán obligadas a cambiar radicalmente la forma de gestionar los datos.

Aquí ya encontramos empresas como *Yahoo* o *Amazon* que empiezan a utilizar análisis rudimentarios de comportamiento para descubrir patrones en sus páginas (Russom,2011).

La tercera fase, empieza alrededor del 2010 y continua hasta la fecha, surge por la adopción masiva de dispositivos móviles y la normalización de sensores y medidores en los productos. Gracias a esto la cantidad de datos que se venía formando desde la anterior fase, se ve multiplicada exponencialmente, puesto que ahora las empresas tienen información casi constante de los usuarios, los cuales voluntariamente se encuentran enviando información constantemente a las organizaciones, en contraposición a la fase anterior de ordenadores fijos.

Esto también es llamado el internet de las cosas, llamado así porque encontramos todos los dispositivos conectados entre sí, desde televisiones, relojes hasta lavavajillas (Big data Framework, 2021).

2.1.2 Las Vs del Big data

Las Vs fueron introducidas junto a la aparición oficial del término Big data, por el ya mencionado (Laney, 2001), consultor de datos en Gartner, éste dio una de las primeras y más simples definiciones del Big data, Volumen, Velocidad y Variedad. Con estas tres palabras, el autor buscaba condensar el fenómeno de la digitalización, el cambio de fase y el surgimiento de empresas digitales, más especialmente las de venta online (Laney, 2001).

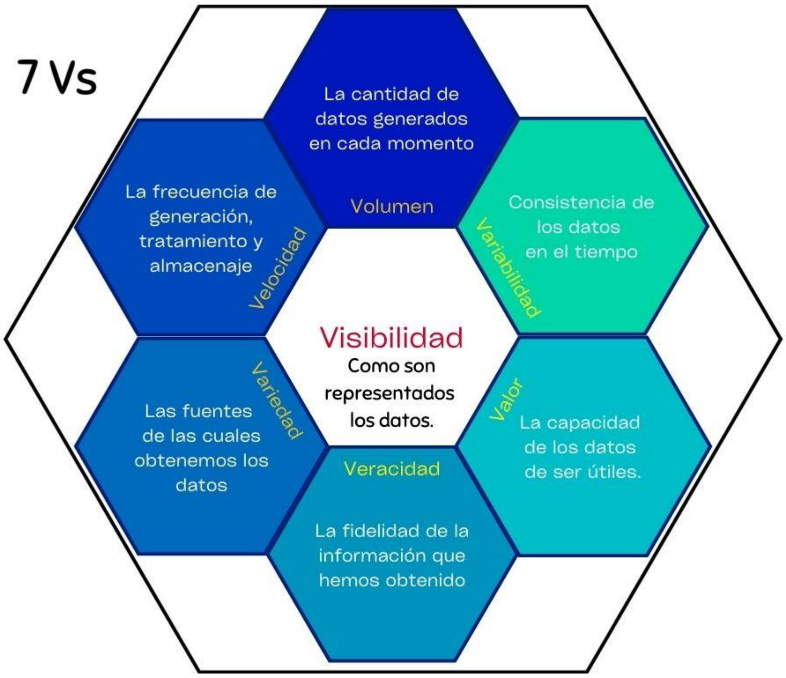
Estas tres palabras han servido de base para muchas de las definiciones del Big data, como por ejemplo la de la consultora tecnológica *IDC* “Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad” (IDC, 2012, p 1).

Si bien es cierto que originalmente solamente eran tres, *IBM* añadió una cuarta, Veracidad mientras que *Oracle* añadiría Valor, quedando la iteración más común de los últimos años para describir al Big data, las 5V´s del Big data, Volumen, Velocidad, Variedad, Veracidad y Valor (Gandomi y Haider, 2015). Con el paso de los años estas se han ido quedando incompletas e inexactas, el Big Data ha ido evolucionado con las tecnologías y la sociedad, dando lugar a que diferentes autores y compañías, añadieran más Vs a las tres originales. Por ejemplo, Kirk Borne, director Científico de *DataPrime.ai*, definió el Big data en 10 Vs (Panimalar.et al, 2017).

Actualmente el enfoque de las Vs se considera desactualizado y autores como Floridi o Marr consideran que es una versión desactualizada del medio, (Favaretto, 2020) sin embargo todavía es utilizado, y se continúan realizando trabajos e investigaciones sobre estas.

Por esta razón, considero que el modelo de las 5Vs que ha sido el más común en los últimos años, dará paso al modelo más actualizado, de las 7Vs Volumen, Velocidad, Variedad, Veracidad, Valor, Variabilidad y visualización. Lleva unos años ganado importancia frente al modelo de las 5Vs. Tiene elementos más complejos en cuenta y si bien es cierto que pueden existir varias interpretaciones, dependiendo de la empresa y el autor, por ejemplo, Maheshwari o Kirk Borne, he escogido, aquellas dos que me parecían más importantes para las empresas.

Ilustración 2: Las 7Vs del Big data: Volumen, Velocidad, Variedad, Variabilidad, Veracidad, Valor, and Visibilidad.



Fuente elaboración propia. Referencia Niculescu V. 2020

Volumen: La cantidad de datos o información creados y almacenados cada segundo que tenemos a nuestra disposición. La única V indispensable para la existencia del Big data, dado que siempre van a ser necesarias grandes cantidades de información para que este surja (Powerdata,2022). Para poner un poco de perspectiva de cuantos datos se generan en internet, en 2012 se creaban aproximadamente unos 2.5 exabytes de información cada día y esa cantidad se duplicaba cada 40 meses. Si tenemos que cada uno de esos exabytes serían aproximadamente 20 millones de archivadores llenos de documentos, nos quedaríamos con unos 50 millones de archivadores diarios (McAfee y Brynjolfsson, 2012, p 4).

Es por esto, que se puede afirmar que uno de los mayores impulsores del Volumen, es la digitalización y la nube, ya que conseguir esos volúmenes físicamente para las empresas, resultaría imposible (Joyanes, 2013).

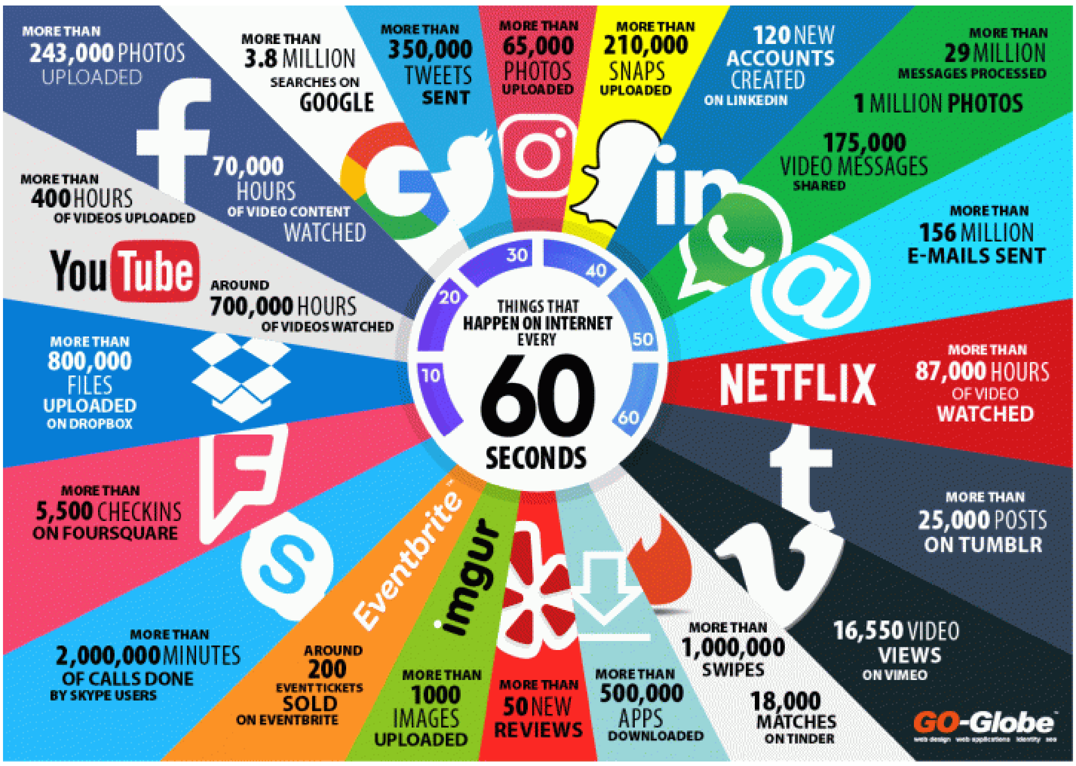
Para que sepamos un poco que implica esa cantidad, en un mundo predigital, en 2010 el entonces CEO de *Google*, Erik Schimdt hizo el siguiente comunicado “Every 2 Days We Create As Much Information As We Did Up To

2003” “Cada dos días creamos tanta información, como habíamos creado desde el principio de la civilización hasta 2003*”* (Siegler, 2010).

Como norma para el volumen, cuanto mayor sea la cantidad de datos mejor, debido a un aumento en la claridad y menos irregularidades dentro de la muestra sin embargo, esto también significa que estaremos limitando aspectos como la velocidad o la veracidad de los datos, dado que los dispositivos electrónicos, cuentan con capacidades limitadas (Keskar et al., 2020). Sin embargo en los últimos años, con el avance de la nube y las mejoras tecnológicas de los servidores, han permitido grandes niveles de almacenamiento a costes muy bajos (Joyanes, 2013).

Con todo esta cantidad de información no sería posible sin la capacidad actual de las empresas, de conseguir más información de sus transacciones en internet que de lo que serían capaces en cualquier operación física realizada (Laney, 2001). Imaginemos que una librería física quisiera conseguir información similar a la que consigue *Amazon*, tendría que apuntar todos los movimientos de los clientes por la tienda, cada libro que han ojeado y una vez vendido, realizar llamadas para que les diese su opinión del libro.

Ilustración 3: Cuanta información se genera en cada minuto.



Referencia Mimmo 2020

Traducción, (Siegler, 2010) “Cada dos días creamos tanta información, como habíamos creado desde el principio de la civilización hasta 2003*”*

Velocidad: La frecuencia de creación, almacenaje, tratamiento y visualización de los datos, casi tan importante como la cantidad es la calidad, tener información actualizada es necesario en cualquier decisión que se desee tomar. Si los bytes eran la forma de medida predilecta del Volumen, el tiempo es la variable más importante en la Velocidad, cuanto más consigamos reducir el tiempo y asemejarlo a la inmediatez, mejores resultados obtendremos en nuestras decisiones. Tomemos el ejemplo de las tarjetas de crédito, cuanto más rápidamente manden las transacciones, menos tiempo tardan en descubrirse fraudes y menos cola de un supermercado haremos (Marr,2014). Esta V es especialmente útil para la detección de estafas, monitorización de los flujos de datos y evitar los errores diarios que pueden ocurrir en el ejercicio de las actividades (Unir, 2020).

Igualmente cuando abordemos esta V, debemos tener en cuenta que Velocidad va a necesitar la organización, si bien es cierto que como norma cuanto menor sea el tiempo mejor, esto no siempre es posible por lo tanto existen diferentes velocidades que usan las organizaciones (Keskar et al., 2020).

- **Flujo de datos:** se están recibiendo los datos en el mismo momento que se crean, sin necesidad de ningún tratamiento, tenemos las cámaras de vigilancia o los servicios de “streaming” como *Twich*, transmiten de forma continua información.
- **Datos en tiempo real:** los datos son creados y consumidos casi al momento por ejemplo, conversaciones telefónicas, videoconferencias o retiradas en cajeros.
- **Datos casi en tiempo real:** existe cierto retraso entre el envío y el recibimiento, recibir la confirmación por email de una compra.
- **Procesamiento en Packs:** los datos llegan periódicamente en grandes cantidades, como los emails diarios sobre novedades que suele mandar Aliexpress o resúmenes diarios en las empresas.

Variedad: La diversidad de fuentes, formatos y tipologías de los cuales podemos obtener datos. Tenemos la capacidad de sacar información de diversos tipos, y de un gran número de fuentes, las más importantes veremos en el apartado 3.1 como pueden ser, archivos, blogs, bases de datos, sensores, webs y similares. Esto provoca una gran diversidad de tipos de datos desde texto, vídeos, imágenes y diferencias en los formatos, desde Word, Access, Excel, JPG, PNG y similares (Unir, 2020). Asimismo los datos se pueden categorizar, en estructurados y no estructurados (Bernard Marr, 2014).

Esta V cobra especial importancia, con la misma expansión de internet, cuantos más sitios sean creados más información saldrá y más variedad existirá, de hecho, cada persona resulta ser una fuente independiente de datos por sí misma (Rulanitee, 2018,).

Un ejemplo de la importancia de la variedad, viene de investigadores de la escuela de medicina John Hopkins, donde encontraron que usando los datos de búsquedas en Google relacionados con la gripe, podían predecir aumentos en el flujo de enfermos de gripe. Asimismo *Wallmart*, usa geolocalización de los móviles, para predecir picos horarios en sus parkings de forma similar a como Google, ahora nos informa del tráfico o de cuánta gente se encuentra presente en una localización, basándose en cuantos móviles se encuentran allí (McAfee y Brynjolfsson, 2012).

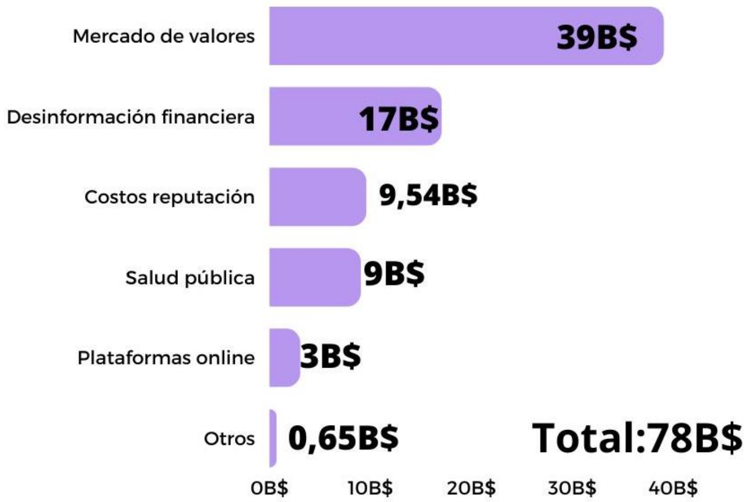
El principal problema de esta V para las empresas y los Marketplace, radica en como igualar, comparar y procesar todos estos datos de una sola vez. Dado que como ya hemos visto, datos que quizás no apreciásemos a simple vista, pueden acabar siendo determinantes para averiguar algo sobre la organización. *Amazon* por ejemplo, no solamente cuenta con la información recogida en sus páginas web, sino que suele utilizar censos, corredores de datos y GPS para mejorar sus recomendaciones (Marr, 2016).

Veracidad: La fidelidad de la información que hemos obtenido, ya que si la información obtenida, no es fidedigna o representativa del mundo real, las aplicaciones de estos dato es casi nula. Sin embargo, la norma en el Big data, no suele ser una inexactitud tan completa de la información, normalmente suele ser el llamado ruido, lo que da más problemas. Esto es básicamente información inútil o incompleta, resultado en su mayoría de los procesos normales del procesamiento y recopilación de datos (Sohaib et al., 2019).

Sin embargo, es una corriente ascendente el número de información falsa al que nos vemos expuestos, y esto sí que puede tener grandes impactos en la empresa.

A continuación, podemos ver los costes aproximados de la desinformación durante el año 2019, esto fue un estudio realizado por *CHEQ*, una empresa de ciberseguridad (CHEQ, 2019).

Ilustración 4: Costes estimados, de las noticias falsas en billones de dólares.



Fuente elaboración propia. Referencia CHEQ 2019

Otro ejemplo, de cómo la falta de veracidad afecta a las empresas, lo encontramos con un tweet, falso, realizado en nombre de la farmacéutica *Eli Lilly*, donde decía que su insulina sería en ese momento gratis, dando lugar a una bajada del 4,4% de sus acciones en bolsa (Lee, 2022).

Valor: Da igual, la cantidad, variedad, veraces o con que velocidad los procesemos si los datos, no tienen valor para nosotros y para nuestra toma de decisiones. Por esto mismo, debemos de preguntarnos siempre antes de empezar con la recopilación de datos, si estos nos van a aportar lo necesario para justificar su coste (Maheshwari, 2015).

Esta V gana mucha importancia, y autores como Maheshwari la colocan en la primera posición de las Vs (Maheshwari, 2015). En esta línea *IBM* plantea que mientras los datos disponible(volumen) aumentan en las organizaciones, el porcentaje de datos que se analiza está en disminución, y cada vez se extrae información importante de un porcentaje menor de los datos capturados (Joyanes, 2013).

Es por esto que debemos de tener siempre en mente nuestro objetivo, la cultura y las necesidades de la organización, antes de realizar cualquier esfuerzo de datos, razonar si estos se encuentran relacionados con lo anterior, y si pueden resultarnos beneficiosos, en especial para la toma de decisiones y para ganar información sobre nuestro entorno y nuestra organización. Si no tenemos esto en cuenta antes de realizar nuestros esfuerzos, estos pueden suponer una carga económica y temporal innecesaria.

Visualización: "La visualización de datos es la representación gráfica de información y datos. Al utilizar elementos visuales como cuadros, gráficos y mapas" (Tableau, 2023, p 1) así define *Tableau*, una de las mayores empresas de visualización de datos, sobre su sector y sus actividades. Es gracias a esta visualización que conseguimos que miles de celdas de información, se conviertan en un formato legible y accesible, que proporcione valor para la organización encontrando patrones, o valores que no habrían sido posibles sin estas herramientas.

La visualización es tremendamente variada, y depende en gran medida de los datos y de aquello que queremos averiguar, pueden ser desde gráficos de barras, hasta arboles de conceptos o histogramas. Por esto es importante que cuando se realice la visualización se elija correctamente el método para su mejor comprensión (Tableau,2023).

Variabilidad: La capacidad de cambio de los datos, los datos en internet se encuentran en todo momento actualizándose o cambiando, esto normalmente significa que mucha de la información que podíamos haber recolectado, ha quedado obsoleta o que ya no es representativa, para cuando llega a nosotros (Sohaib et al., 2019). Por esto las empresas deben de buscar aquellos datos que sean más estables en el tiempo, y que sufran de menos actualizaciones, sin quedar obsoletos, a la hora de la recopilación de datos.

2.2 Importancia del Big data

"El aumento del volumen y la velocidad de datos significa que las organizaciones tendrán que desarrollar herramientas para recopilar, analizar e interpretar los datos" (Davenport et al., 2012 p 23).

Esta frase resume perfectamente la importancia del Big Data y porque las empresas, están haciendo tantos esfuerzos por adoptarla. Los métodos tradicionales de tratamiento de datos, como la estadística básica o los reportes ya no resultan suficientes para gestionar la información con la que se ven bombardeada cada día las organizaciones (Manyika, 2011). Estos métodos tradicionales incluyen aspectos como estadística básica, como varianzas o regresiones, servidores y bases de datos físicos como papel, CD o disquetes. Para ver hasta qué punto, se pueden atascar los procesos por falta de necesidad, tenemos el curioso caso del gobierno Japones, donde unos 1.900 procesos gubernamentales todavía requerían de disquetes en el 2022, puesto que al no tener la necesidad de evolucionar junto al resto de la población, esta organización había acabo quedándose con procesos tan desactualizados, que usaban métodos de almacenaje que ya no se fabrican, esto sería similar a aceptar solamente reportes escritos a máquinas de escribir (BBC, 2022).

Las empresas, cada vez obtienen una mayor cantidad de datos, tanto de sus transacciones, como de los consumidores a través del internet, cada uno de nuestros ordenadores es un pequeño punto que contribuye al tráfico global de datos (Manyika, 2011). Gracias a este aumento de los datos, el Big data tiene un impacto en casi todos los sectores de la sociedad, desde las empresas, gobiernos, salud o construcción, esta marea de datos que nosotros mismo creamos, cambia a quienes toca y les obliga a adaptarse a nuevas cotas de información que no dejan de aumentar (Aguilar,2013).

Pese a la gran cantidad de datos que desde la antigüedad se han visto obligados a gestionar, organizaciones como gobiernos suelen apoyarse en extensas burocracias, métodos y sistemas desactualizados que nublan la visón que tienen del exterior y puede dar lugar a errores o estancamiento como paso en Japón.

Con todo algunos de ellos han podido ver la importancia de estas nuevas técnicas y están invirtiendo en formas de poder adaptarlas a sus propias necesidades. En su mayoría realizan estos esfuerzos para compartirlos y crear una mayor transparencia con el exterior, esto permitiría tanto a empresas como a particulares, mejorar su opinión de los gobiernos y conseguiría un mercado más informado y por lo tanto más competitivo. Por ejemplo tenemos EEUU con su iniciativa Data.gov, en Europa encontramos hasta 183 iniciativas independientes, y muchos países de Sudamérica, Asia África están empezando a realizar esfuerzos similares (Joyanes, 2013).

En cuanto a las empresas podemos encontrar técnicas de Big data en muchísimos campos y áreas laborales, puede ser usado, desde secuenciación de ADN, empresas como *Myheritage* que buscan reconstruir el pasado genético de un individuo, usan estas técnicas para abaratar costes y disminuir el tiempo en algo que antes podía tardar años. (Mayer-Schnberger & Cukier, 2013).

Tenemos el libro escrito por (Marr, 2016) el cual nos da un vistazo en 45 empresas de diferentes tamaños ya áreas que han podido de manera exitosa, utilizar aspectos del Big data en sus operaciones y como estas se han visto afectadas.

Otro aspecto de vital importancia para las empresas, las ventas y el marketing se pueden ver muy beneficiadas por el uso de Big data, por ejemplo en las empresas de tarjetas de crédito, las cuales antes de la adopción de técnicas de Big data y análisis, tardaban semanas en extraer y analizar correctamente los datos que tenían sobre sus clientes para ofrecer un plan personalizado, ahora gracias a herramientas de análisis avanzadas, como el análisis de sentimientos o el predictivo, que veremos en el apartado 4, son capaces de en una llamada ofrecer ofertas personalizadas a las necesidades del cliente (Davenport et al., 2012).

Finalmente, una razón que quizás no sea tan hablada, la encontramos en la evolución, como ya hemos visto el Big data es algo muy relacionado a la sociedad y a la cultura del momento, por lo tanto según estas avanzan han acabado forzando a muchas tecnologías a evolucionar y adaptarse. Por ejemplo, con el aumento del volumen de datos, a partir del 2010 con la llegada de los móviles inteligentes, forzó la aparición de nubes de almacenamiento masivas, y de métodos para poder contener y gestionar estos nuevos volúmenes incrementales de datos y en general trajeron una nueva forma de pensar sobre la información (Mayer-Schnberger & Cukier, 2013).

En general los datos masivos son una revolución que está cambiando la forma de ver el mundo, el Big data ya nos permite realizar tareas que antes llevaban años en unos pocos días (Mayer-Schnberger & Cukier, 2013). Por ejemplo, ahora las empresas son capaces de recibir información en tiempo real del cliente según este se desplaza por su página web, los supermercados son capaces de predecir cuantas ventas tendrán y empresas como *Uber* son capaces de realizar estimaciones casi exactas de tiempo y precio en viajes en ciudades, todo esto resultaba algo imposible para las primeras empresas online (Marr,2016).

2.2.1 Ventajas del uso del Big data

Como ya hemos visto, el Big data es muy importante para las empresas y puede ser un factor decisivo para su éxito en el mercado, por esto quiero destacar alguna de las ventajas más generales que ofrecen los datos masivos (McKinsey, 2011):

Ilustración 5: Principales ventajas del Big data.

Fuente elaboración propia. Referencia McKinsey, 2011

- **Crea transparencia:** hemos visto que muchos gobiernos están adoptando programas de Big data, para crear visibilidad sobre los datos que estos recogen y mejorar la opinión del publico así como ayudar a las empresas a conocer mejor los mercados (Joyanes, 2013).

- **Permite experimentar:** antes de los datos masivos, las empresas necesitaban esperar a la generación de informes y reportes, para saber si sus estrategias habían resultado exitosas y en qué grado estas lo habían sido. Con el uso de los datos masivos, estos tiempos se ven acortados en gran medida, de forma similar a como *IBM* solución el problema del censo.

- **Mejora la segmentación:** segmentación es la principal razón por la cual existen los corredores de datos, que veremos en el siguiente apartado, el Big data permite separar de una forma mucho más exacta y eficaz, las poblaciones en sesgos.

- **Toma de decisiones:** la ventaja más importante, desde la recogida de datos, hasta la decisión final, encontramos técnicas de Big data en todo el proceso de decisión, sea automático o simplemente sirva como apoyo para la persona.

La aparición de los datos masivos, ha creado una ventaja competitiva para aquellas organizaciones que la han adoptado frente al resto de estructuras que no lo han hecho, esta se puede apreciar especialmente con la competitividad de otras empresas que no usan el Big data en sus organizaciones, o que evitan el uso de datos masivos en general (Marr, 2016). Si bien es cierto que depende del sector al cual nos dirigimos, la inmensa mayoría de áreas se verán beneficiadas por la adopción de Big data en sus organizaciones. Por ejemplo las empresas de ventas al público y al por mayor, fueron estimadas que conseguirían una mejora en beneficios de un 60% al adoptar estas técnicas o una capacidad de decisión hasta dos veces más rápida y eficaz (McKinsey, 2011).

Finalmente destacar que ventajas nos aporta a nosotros como usuarios, el Big data, no sirve solamente a las empresas, sino que su beneficio repercute en los usuarios también, primeramente obtenemos más información de nuestras compras, tenemos más opciones de compra, por ejemplo en 2003 Oren Etzioni un científico de computación, descubrió que había pagado mucho más por su asiento de avión que el resto de pasajeros, creo la empresa *Farecast* la cual podía predecir aumentos en los boletos de aviones en tiempo real. Los vendedores se pueden ajustar mejor al mercado y por lo tanto, pueden ofrecer precios y productos más personalizados a las personas (McKinsey, 2011).

Por último tenemos algo clave, la comodidad al tener las empresas información de nuestros gustos y preferencias, pueden aportar una experiencia más completa a nuestra compra y hacerla de la forma que nos resulte más cómoda, desde la selección del producto hasta el pago (Mayer-Schnberger & Cukier, 2013).

2.3 Venta de datos y usuario como producto

Uno de los principales lugares, de donde las empresas, pueden extraer información valiosa, son los mismos consumidores, todos esos cuadritos donde aceptamos los términos y condiciones, o cuando en una llamada a la compañía de internet nos avisan que la llamada puede ser grabada, están recopilando nuestros datos para ser posteriormente recopilados y analizados. Todo esto ha dado lugar a una economía nueva, exclusiva de datos donde las empresas compran y venden la información de los consumidores.

Esta información es tremendamente valiosa y útil para las empresas que lo compran, dado que normalmente les resultaría mucho más complicado y costoso conseguir información similar. Esta información puede ser usada en varios departamentos simultáneos, como ventas y marketing, pueden ser usados varias veces y dependiendo de qué datos tengamos, pueden tardar mucho en quedar obsoletos. El mayor beneficio lo encontramos en los corredores, puesto que estos datos pueden ser vendidos todas las veces que le sea necesario, pueden resultar activos económicos muy valiosos y alcanzan precios altos en los mercados (Latto, 2023).

Finalmente existe otro peligro con esta recogida de datos, el de la sobreinformación y las noticias falsas, ya vimos en el apartado 2.3 los costes que tienen las noticias falsas globalmente, sin embargo estos costes se relacionan estrechamente con la sobreinformación este fenómeno es tan grande, que se podría decir que nos encontramos en la era de la sobreinformación, llegan tantos estímulos y datos a nuestro cerebro, que no somos capaces de filtrar adecuadamente, esto no es un fenómeno nuevo, desde que existen las noticias la gente ha buscado formas de ocultar la verdad que no sirva a sus intereses, sin embargo ahora con internet y la redes sociales, los usuarios se ven bombardeados para hacerles crear aquello que beneficie a las corporaciones (Garay, 2019).

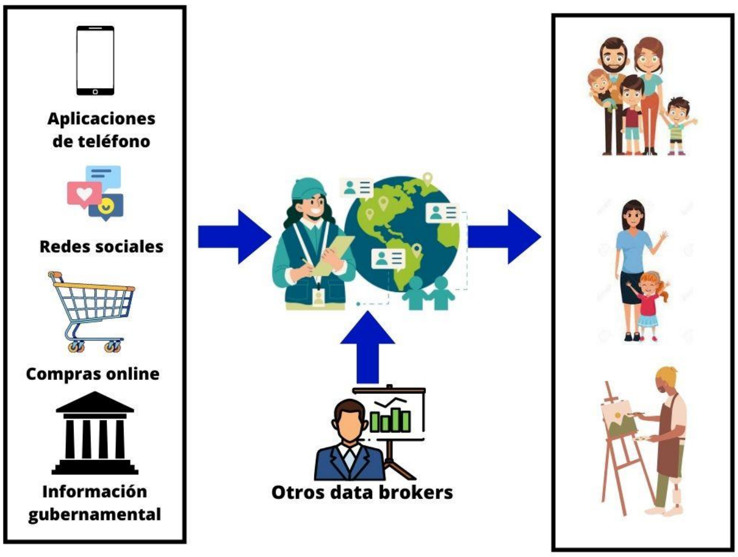
Ya en 2012, *Intel* realizó una encuesta sobre la información en las redes sociales, en las mayores potencias económicas del momento, y concluyo que casi la mitad de las personas preguntadas, se sentían sobrecargados de información (Joyanes, 2013).

2.3.1 Corredores de datos

Aquellos encargados de la compra y venta de estos datos, son los llamados “data Brokers” o corredores de datos, **“**recopilan datos personales, los agrupan en paquetes y los venden a terceros**”** (Latto, 2023, p 1) los corredores, reúnen información tanto dentro como fuera de internet, desde búsquedas de gatitos, hasta impuestos y visitas al hospital. Estos paquetes también incluyen aparte de esta información, datos detallados de los gustos, preferencias, ideales y experiencias importantes de las personas, que puedan resultar relevantes para las empresas (Anthes, 2014).

Primeramente estos datos son recopilados, segundo estos serán etiquetados, clasificados y gestionados, con todo si fuera necesario, por falta de volumen de datos o falta de información sobre algún aspecto, complementados con los datos de otros corredores, para finalmente ser clasificados por paquetes, estos son conjuntos de personas, que comparten alguna característica similar, ya sea demográfica, étnica, cultural o social, por ejemplo madres solteras jóvenes, urbanitas extranjeros de bajos ingresos y multitud de variantes, dependiendo de a quien se los quieran vender. Asimismo, esta clasificación se suele subdividir aún más, ya sea por razones médicas familiares o similares, por ejemplo madres solteras jóvenes con diabetes y con dos padres.

Ilustración 6: Estructura de paquetes de corredores de datos.



Fuente elaboración propia. Referencia Privacy Bee 2020

Pero como reúnen estos información y como evitarlos, según el antivirus Avast, los métodos más usados por los corredores serían los siguientes (Latto, 2023):

- **Cookies de seguimiento:** archivos de datos, que las páginas web usan para registrar información sobre sus visitantes.
- **Huella digital del navegador:** cuando navegamos por internet, dejamos unas huellas por aquellos sitios que visitamos o que buscamos, que pueden ser usada para rastrear nuestra actividad.

- **Balizas de correo y páginas web**: pequeñas imágenes de unos píxeles, que se encargan de seguir nuestra actividad en páginas web o en correos, para mandar anuncios relacionados.
- **Seguimiento de IP**: tu carnet de identidad de internet, cada vez que accedes a internet, lo haces desde esta dirección la cual puede ser rastreada para saber a qué lugares vas.
- **Sitios de compra online:** *Amazon* registra tus preferencias y gustos, para mejorar la experiencia, y estos datos muchas veces pueden ser aprovechados por terceros.

La principal forma de ingreso de los corredores, es la venta de estos paquetes de datos de consumidores sin embargo, de vez en cuando realizan una venta más especializada, con perfiles específicos. Normalmente los compradores, suelen ser otros corredores, anunciantes y encargados de campañas políticas, sin embargo cualquier empresa puede usar sus servicios de vez en cuando (Latto, 2023).

Existen muchas razones por las que estas prácticas son consideradas inmorales o directamente dañinas por ejemplo, en personas con facilidad de adicciones, anuncios de juego o de alcohol pueden ser predominantes, o los hackers pueden robar identidades y datos bancarios.

Sin embargo, los corredores son un mal necesario, dado que mueven una economía billonaria, dan multitud de trabajos y muchas empresas tecnológicas, entrarían en crisis sin sus servicios. Asimismo, estos ayudan a detectar fraudes, hacen el proceso de compra más cómodo para los consumidores y mueven la economía mundial, creando un mercado nuevo basado en los datos (Anthes, 2014).

2.3.2 Entidades que protegen al usuario

La existencia de estas empresas y de otras similares, ha llevado a la creación de organismos y leyes por parte de los gobiernos, para controlar lo que las organizaciones pueden hacer con estos datos.

En el caso de España, tenemos *Incibe* la cual es una organización centrada principalmente en ciberseguridad, sin embargo trata de informar a la ciudadanía de los peligros de la venta de datos, y aboga por un uso responsable de la información y evitar estos casos(Incibe, 2023). Asimismo tenemos la “Ley

Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales." Siendo la iteración más actualizada de leyes sobre la protección de los datos, con aún actualizaciones y cambios(Boe, 2023).

Asimismo existen varios programas de protección de datos en otros países, por ejemplo en la Unión Europea existe el Reglamento general de protección de datos (GDPR), que busca la protección de los datos de los consumidores en la unión (Aguilar, 2013).

2.4 Pirámide DKIW

Ilustración 7: Estructura de la pirámide DKIW.

Fuente elaboración propia. Referencia Soto 2022

La pirámide DKIW "Data Information Knowledge Wisdom" es un modelo que intenta explicar el proceso de obtención de información, desde la obtención, hasta la emisión de una decisión, para ello se basa en la relación entre los datos, información, conocimiento y sabiduría, sirviendo cada parte inferior, como base para la superior. El objetivo final de la pirámide, es que lleguemos al mejor juicio de valor posible tomando como base, tanto los datos actuales como aquellos que ya teníamos (Figuerola, 2013).

Datos: cuando hablamos de datos, lo que queremos decir son hechos o cifras concretas, sin que nosotros tengamos contexto de dónde vienen estos y que muchas veces carecen de un significado concreto. Sin que tengan un contexto que los defina, estos resultan inútiles para las organizaciones. Por ejemplo 100 años o casa roja.

Información: ya un nivel por encima de los datos, estos añaden un contexto a los datos más específico añadiendo significado y cierto valor estos, agrupándolos en ciertos apartados para que estos puedan ser comprendidos o usados en un contexto. Por ejemplo el agua está a 100 grados o las casas rojas de la calle Benito.

Conocimiento: una vez tenemos el contexto, debemos de entender lo que nos quieren decir, esto sería alguna experiencia previa sobre el tema a tratar y que por lo tanto nos a permitir entender los datos en su contexto esto ya nos permite la toma de algunas decisiones. Por ejemplo, si ponemos el agua a menos de 100 grados, esta no se evaporará.

Sabiduría: Conocimiento acumulado que te permite aplicar los conceptos de un campo o de varios a nuevas situaciones o problemas. La sabiduría te permite dar el mejor juicio ante una situación, esto implica una experiencia previa muy amplia y que te da la capacidad de actuar ante una situación en consecuencia. Por ejemplo, el agua hierve a 100 grados, pero el alcohol hierve a 78 grados, por lo tanto hirviendo a 90 grados, separamos alcohol de agua.

3. Manejo de los datos en el Big data

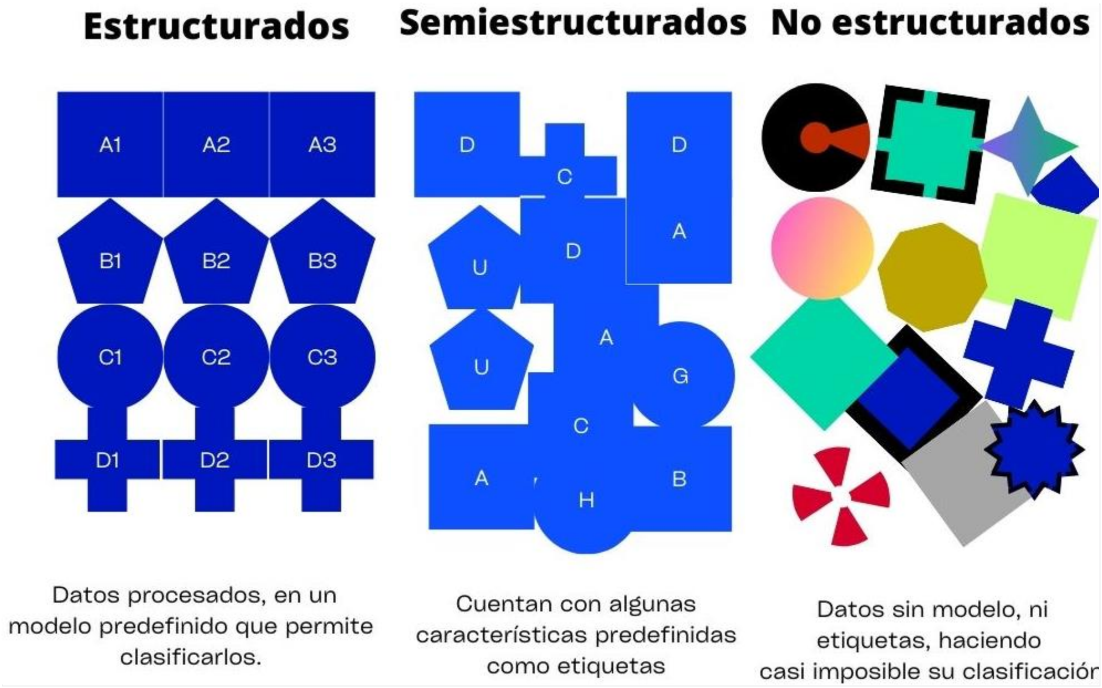
La gestión y control de los datos es vital para la correcta aplicación de técnicas y herramientas, por esto aquellas personas que trabajan diariamente en el sector se ven obligadas a encontrar formas de clasificar racionalmente los datos, para saber cuáles son las mejores formas de abordarlos (Soares, 2013). A continuación explico, cómo se pueden dividir estos conjuntos de datos y que diferencias existen entre ellos. Así como algunas de las herramientas más utilizadas en la gestión, control y como ayudan a las organizaciones a sobrevivir en el mar de datos que se encuentran.

3.1 Tipos de datos y fuentes de datos

Como ya mencionamos en la Variedad los datos pueden tomar multitud de formas y de orígenes, siendo las más comunes las siguientes (Bastidas, 2021):

3.1.1 Forma de los datos

Ilustración 8: Forma de los datos.



Fuente elaboración propia

Estructurados: son datos que han convertidos en algún tipo de formato y forma de clasificación de datos, dándoles una apariencia similar entre los datos con el mismo formato, informes, reportes, hojas de cálculo, son datos con forma y estructura fija, son fáciles de clasificar incluso si no se está familiarizado con su formato (Bastidas, 2021). Estos datos son en su mayoría producidos por organizaciones y no por los usuarios. Por ejemplo un número de teléfono móvil tiene la siguiente forma numérica y esta no se puede cambiar(+XX XXX XXX XXX) (Soares, 2013).

Semiestructurados: tienen una estructura algo definida, sin ser tan clara como los estructurados. Esta estructura, suelen ser etiquetas o identificadores que permite que sean clasificados mediante un criterio racional (Joyanes, 2013). Sin embargo estas no son fáciles de clasificar y normalmente requieren reglas complejas por lo tanto, necesitas estar familiarizado con estas para su clasificación. Algunos ejemplos, serían las imágenes y videos que cuentan con localización, fecha, hora o incluso Hastags, dado que es una forma de etiquetación (Bastidas, 2021).

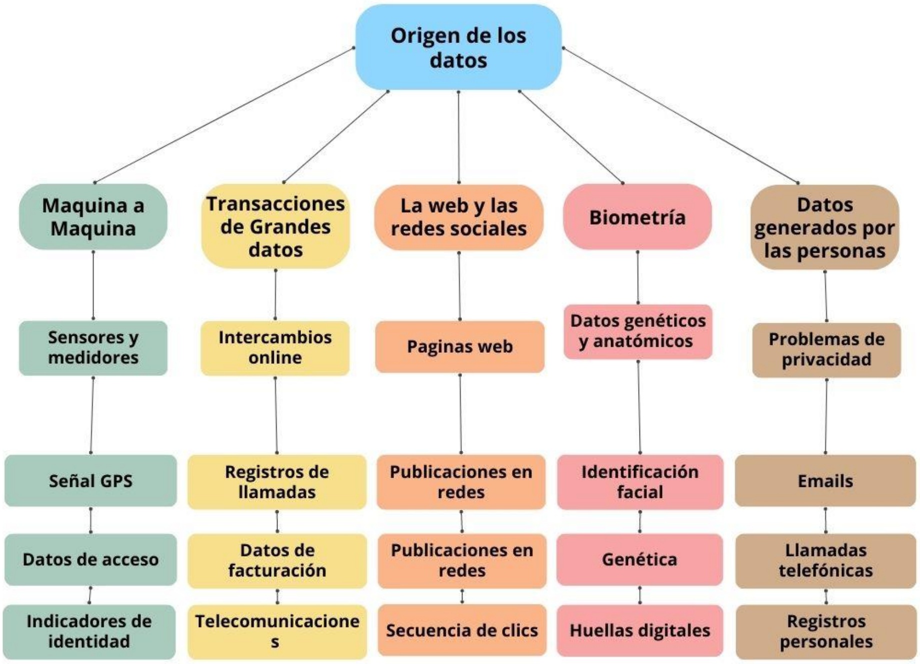
No estructurados: son aproximadamente el 90% de los datos en internet (Marr, 2021) no tienen una forma o estructura predefinida, son casi imposibles de clasificar de forma racional, siendo mucho más complicados de analizar, usando formatos muy diversos, donde se incluyen videos, fotografías, audios, conversaciones, comentarios en redes (Keskar et al., 2020).

Sin embargo sí que existen iniciativas para su etiquetado, como es el sitio web "Image.net" la cual según las palabra de su página web "es una base de datos de imágenes organizada con nombres, donde cada nombre cuenta con cientos o miles de imágenes relacionadas con el tema" (Image, 2023) esta iniciativa, busca convertir las imágenes sin estructurar, en datos semiestructurados, lo cual ha resultado especialmente útil para el aprendizaje de IA y enseñarlas a diferenciar conceptos (Soares, 2013).

3.1.2 Origen de los datos

Una vez que hemos visto que tipos de datos nos podemos encontrar, vamos a ver dónde sería más probable que obtengamos dichos datos, y qué diferencias podemos encontrar cuando los datos tienen uno u otro origen (Joyanes, 2013).

Ilustración 9: Origen de la mayoría de datos de Big data.



Fuente elaboración propia. Referencia Soares 2013

M2M o Máquina a máquina: estas son las tecnologías que permiten conectar dos o más dispositivos entre sí, normalmente estos dispositivos conectados siendo sensores o medidores, podemos encontrar estos medidores, en aparatos como las televisiones inteligentes, relojes inteligentes, automóviles y muchos otros dispositivos que usamos diariamente. Estos dispositivos, registran actividades o eventos, estos eventos pudiendo ser desde aumentos ambientales, señales GPS, indicadores de identidad, datos de acceso y similares.

Una vez estos eventos son registrados, los dispositivos transmiten estos datos a otras máquinas, encargadas de su almacenaje y conversión en datos que puedan ser usados por las organizaciones, o en su defecto las mismas maquinas tomarán decisiones automáticas, este tipo de conexiones se encentran en auge, dado que sensores y medidores son ahora la norma para multitud de productos (Soares, 2013). Por ejemplo, cuando el termómetro enciende la caldera, dado que ha registrado ciertas temperaturas.

Número de conexiones M2M en el mundo desde el 2014 al 2020

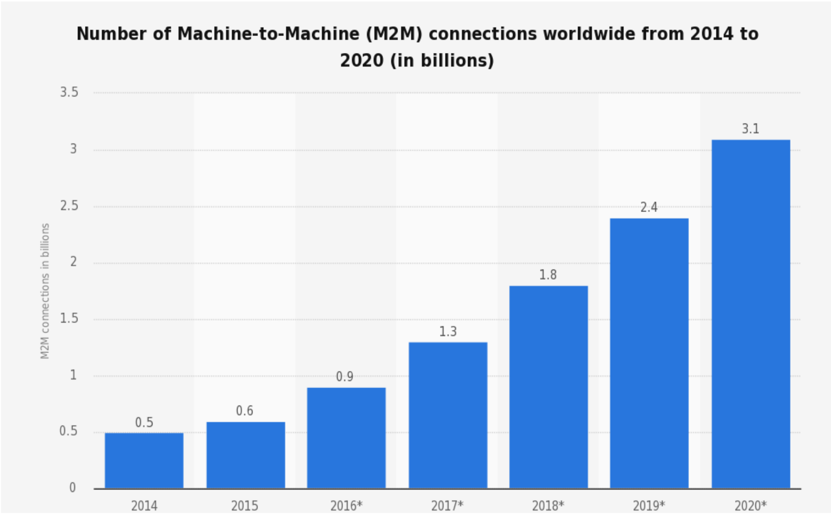


Ilustración 10 Referencia Statista 2016

Transacciones de grandes datos: esto incluye, datos generados durante transacciones, desde registros de llamadas, telecomunicaciones, registros de compras o datos de facturación, siendo normalmente no o semi estructurados. Este tipo de origen, suele tener problemas con la privacidad de los usuarios, dado que muchos de estos datos son personas para los consumidores y a veces son de difícil acceso.

La web y las redes sociales: abarca la mayoría de contenido disponible en internet, desde publicaciones en redes, comentarios, noticias, secuencia de clics[3], blogs o paginas especializadas, esta es normalmente la más extensa fuente de datos. Este origen captura multitud de datos las publicaciones de los usuarios, no solamente su contenido, sino que lugar, hora y fecha de la publicación y si existen antecedentes a esa publicación, haciendo este uno de los orígenes más completos y complejos, por esto encontramos mucha más información de la que podría parecer a primera vista, siendo la inmensa mayoría poco estructurados dado que usaremos comentarios y opiniones de los usuarios (Soares, 2013).

Normalmente va a permitir, una comprensión más profunda de los consumidores sirviendo como base de muchas, de las grandes empresas centradas en internet, o de aquellas que usan el mismo para mejorar sus procesos (Marr, 2016). *Amazon* obtiene una inmensa cantidad de datos útiles, en especial de los comentarios sobre sus productos, tanto en su página web como en las redes y luego, esta información puede ser usada para mejorar la experiencia de los consumidores y elegir las mejores recomendaciones. Estos datos suelen ser analizados, por herramientas de analítica web(Joyanes, 2013).

Biometría: los datos biológicos, estos permiten identificar a cada ser humano como un individuo, dividiéndose en datos anatómicos y genéticos. Los datos genéticos, ADN, sangre y aquello que necesitaría un análisis más exhaustivo siendo casi imposibles de cambiar, por otro lado, tenemos los anatómicos, los cuales pueden ser fácilmente reconocibles, pero más fáciles de alterar, reconocimiento facial, huellas dactilares, ojos, voz (Soares, 2013). Un ejemplo de cómo las empresas están utilizando estas tecnologías, lo tenemos en los teléfonos, las últimas formas de desbloqueo, son la huella dactilar y el reconocimiento facial, lo cual permite personalizar aún más el producto para el consumidor.

Estos datos son en su mayoría muy personales y extremadamente útiles para identificar personas, incluso de formas inesperadas. Para ver en qué grado pueden resultar útiles, tenemos el ejemplo de Shigeomi Koshimizu, profesor del Instituto Avanzado de Tecnología Industrial de Japón, colocó 360 sensores en el asiento de un coche, para ver si podía detectar entre un grupo de personas, quien se había sentado convirtiendo la presión en cada punto en datos, y el resultado acabó siendo de un acierto de un 98% (Mayer-Schnberger & Cukier, 2013, p 51).

Por último, estos datos pueden resultar tremendamente útiles por una serie de razones, con una rápida búsqueda de *Google* podemos encontrar varios ejemplos de la utilización del Big data en la medicina. Desde hace unos años, se llevan usando técnicas de Big data para analizar individualmente las células de tumores y de otros tejidos cancerígenos, a un nivel genético, creando reportes mucho más específicos que permiten personalizar el tratamiento a cada paciente (Rojas, 2022).

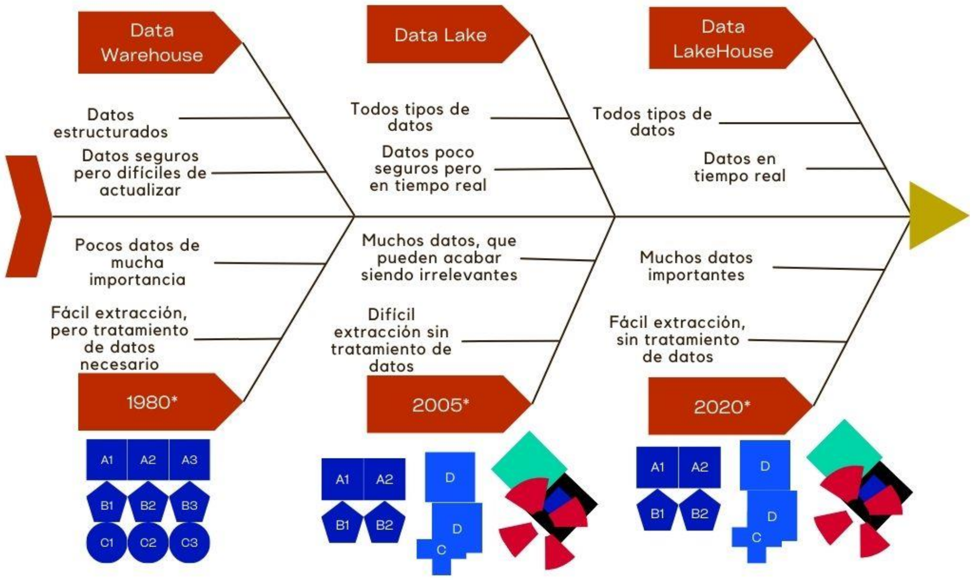
Datos generados por las personas: la información generada directamente por las personas. Siendo como ya hemos comentado, el principal problema de esta información la privacidad, puesto que esta fuente puede contener conversaciones privadas o registros que el consumidor no quería compartir en un principio del dentro de esta encontramos los emails, notas de voz, encuestas o registros médicos (Soares, 2013).

3.2 Bases de datos

La capacidad de almacenar la mayor cantidad de información, en el menor espacio posible, es algo que lleva siglos siendo un obstáculo para la humanidad, encontramos como claros ejemplos las primeras bibliotecas, estos primero intentos de nuestros antepasados de crear un repositorio de datos[4] unificado, fueron los primeros intentos del gobierno por unificar la mayor cantidad de conocimiento posible en el mismo lugar.

Sin embargo estos primero intentos no tuvieron mucho éxito y las grandes bibliotecas, eran proclives a incendios y saqueos constantes (Mayer-Schnberger & Cukier, 2013). Las bases de datos no son muy diferentes des estos primeros intentos, buscan condensar la mayor cantidad de información útil en un formato datificable y permitir a los encargados de las organizaciones, usar esa información contenida para comprender mejor la situación de la empresa y en consecuencia mejorar la toma de decisiones. Sin embargo, las bases de datos sufren de problemas que las bibliotecas apenas si notan, las bases, deben de adaptarse al entorno en el cual se encuentran, deben estar siempre disponibles y deben de poder actualizarse cada poco (Nambiar & Mundra, 2022).

Ilustración 11: Estructura de bases de datos.



Fuente elaboración propia

3.2.1 Data Warehouse

"Is a large data repository wherein data can be stored and integrated from various sources in a well-structured manner" (Nambiar & Mundra, 2022, p 1). Este repositorio, fue una de las primeras bases de datos en surgir y su función principal sería la de realizar consultas, acceder y analizar los datos en tiempo real o ver la historia de los datos que contienen, quien ha sacado los datos y para que los ha sacado.

Esto tiene como objetivo, ayudar a las organizaciones en la toma de decisiones, manteniendo un registro de la información más importante en las empresas (Talend, 2023).

Los datos dentro de este se encuentran estructurados, dado que cuando estos surgen, la mayoría de información se encontraba en este formato, al ser las organizaciones las principales emisoras de información, mientras que internet y los contenidos creados por usuarios no eran todavía tan importantes. El hecho que estén estructurados, da la ventaja de fácil revisión, gestión y almacenaje, igualmente cada departamento cuenta con sus subapartados dentro estos repositorios, llamados "data Marts", estos actúan como puertas con contraseña dentro del repositorio, para asegurar que solamente miembros del departamento asignado podrán acceder a la puerta de su departamento(Joyanes, 2013).

Esto permite mejorar la organización del repositorio, al evitar repeticiones y almacenamiento innecesario en el mismo "data Mart", igualmente al estar separados por claves, mejora la seguridad puesto que los ladrones solamente pueden acceder a una parte de la base de datos. Sin embargo este repositorio tiene algunos problemas, es más complicado compartir información entre departamentos, es poco flexible, difícil de actualizar, solo admite datos estructurados y su proceso de admisión de los datos es complejo (Nambiar & Mundra, 2022).

Traducción, (Nambiar & Mundra, 2022, p 1) "Es un gran repositorio de datos en el que se pueden almacenar e integrar datos de diversas fuentes de forma bien estructurada"

3.2.2 Data Lake

"Data lakes are centralized storage repositories that enable users to store raw, unprocessed data in their original format" (Nambiar & Mundra, 2022, p 2). Con el aumento de variedad y volumen de información en internet, se hacía necesaria una nueva forma de almacenamiento de datos, diferente de las "Warehouses" tradicionales esto serían los "data Lakes" o lago de datos, estos cuentan con mucha mayor flexibilidad y almacenamiento, ideales para un entorno donde no se paran de generar datos en ningún momento y que necesita actualización constante.

Los lagos de datos son la alternativa moderna a los "Warehouses" y a diferencia de estos, pueden tener todo tipo de datos dentro de su repositorio (muy útiles cuando el 90% de los datos son no estructurados) (Marr, 2021) su proceso de integración de datos es más sencillo, solo etiquetándolos, para una posterior búsqueda. Esto permite que estos datos sean más maleables y flexibles, al ser su gestión mucho más sencilla que métodos más tradicionales, esto no vienen sin sus propios problemas sin embargo, da lugar a que exista un mayor número de datos inútiles al estar admitiendo en el sistema, estos datos que no se encuentran estructurados.

Compensan este exceso de datos inútiles, con una capacidad de almacenamiento mucho mayor que las "Warehouses", igualmente al encontrarse centralizados y sin barreras dentro del repositorio, permiten a los departamentos compartir datos y mejorar sustancialmente la comunicación interdepartamental. Sin embargo, esto significa, que la extracción de los datos es más complicada y suele dar lugar a retrasos, repeticiones de datos, inexactitudes en con la información adquirida y una menor seguridad, dado que si un hacker tiene acceso a la base, no existen barreras dentro de esta como en las "Warehouses"(Joyanes, 2013).

Traducción, (Nambiar & Mundra, 2022, p 2) "Los lagos de datos son depósitos de almacenamiento centralizados que permiten a los usuarios almacenar datos sin procesar en su formato original"

Finalmente estos cuentan con un gran problema, cuando su mantenimiento no es el adecuado o es utilizado en exceso sin purgar datos innecesarios, el lago se convertirá en un llamado pantano de datos, siendo esto un lago de datos, que en lugar de contar con datos útiles para la empresa, se encuentra lleno de datos repetitivos, mal integrados y en su gran mayoría sin ningún valor real para la organización, cuando un lago se convierte en un pantano, significa que nos quedamos con una base de datos de capacidades reducidas, que entorpecerá el correcto funcionamiento de la organización(IBM, 2018).

La última gran diferencia entre estos dos métodos de almacenamiento, sería que los lagos, permiten usar mucho mejor herramientas de ”data science” y “machine learning” frente a las “warehouses” (Nambiar & Mundra, 2022).

3.2.3 Data Lakehouse

Finalmente existen los “data Lakehouse”, siendo estos la más moderna iteración de las bases de datos, su forma es la de un lago de datos, pero añade capacidades que le convierte en una versión mejorada de este buscando combinar las ventajas de las dos estructuras anteriores “Toma el almacenamiento flexible de datos no estructurados de un lago de datos y las funciones y herramientas de gestión de almacenes de datos" (Oracle, 2023, p 1).

Las principales características del mismo, radican en combinar las ventajas de los “Warehouse”, buscando seguridad, estructura y fiabilidad con la de los lagos, gran capacidad de almacenamiento, flexibilidad y variedad de datos. Consigue esto a grandes rasgos, tomando un lago de datos al cual le añadimos elementos similares a las “Warehouse”, en especial su estructura y la forma de admisión de los datos, lo cual le permite ser flexible como los lagos pero mejora la extracción de datos, su seguridad y elimina repeticiones que podrían ocurrir al tener tantos datos, como ocurriría con un “Warehouse” (IBM, 2018).

4. Análisis de datos

En el apartado 2.1.2 hablamos sobre el Valor del Big data, como una de las Vs más importantes, no te sirve de nada tener 20.000 celdas de información sobre los mejores tejidos para invierno, cuando eres una empresa tecnológica. Ocurre algo similar con el Big data, de que nos sirven estas masivas cantidades de datos si no podemos usarlas en nuestra organización.

Aquí entra el análisis de datos, esta faceta del Big data se centra en interpretar la información que hemos conseguido y transformarla de tal forma que podamos usarla en las organizaciones. Como el propio Big data, este campo es muy amplio y con los avances en IA, están potenciando las capacidades de estos análisis, a niveles nunca antes vistos (Marr, 2016).

Como mencionamos al principio del trabajo, la toma de decisiones es la razón por la cual existe el Big Data, y nunca se llegaría hasta conocimiento práctico solamente observando estos datos, es necesario un proceso de análisis para conseguir el conocimiento que nos permita realmente mejorar la toma de decisiones. Antes de la digitalización, este proceso de análisis era extremadamente exhaustivo y normalmente requería de muchos recursos, para ser completado usando técnicas tradicionales, el mejor ejemplo lo tenemos con el nacimiento de *IBM* y el censo de los estados Unidos (Russom, 2011).

Sin embargo, ahora tenemos nuevas técnicas de análisis que nos ayudan a reducir tanto el tiempo, cómo los recursos empleados. Es el conjunto de técnicas y estrategias centradas en optimizar el posicionamiento orgánico en buscadores de internet) cual es el mejor y te da su resultado. Otro ejemplo también quizás más claro, sería *Shazam* (aplicación utilizada para la identificación de música, basado en fragmentos de la misma)con su identificación de música (Marrs, 2016).

Una encuesta realizada a 3.000 ejecutivos, muchos de ellos trabajando en las empresas más importantes de sus sectores, coincidían en como el uso de la información y el análisis les diferenciaba del resto de su sector. Estos a diferencia del resto, usaban analíticas en todos los niveles de su organización no solamente en las decisiones importantes, siendo esta la norma en muchas empresas. Asimismo se descubrió que las empresas exitosas, usaban dos veces más análisis que el resto de empresas de su sector.

En este estudio, se demostró que las empresas, cada vez cuentan con menos problemas para obtener datos, solamente dos de cada diez encuestados, tenía problemas en la recopilación en cambio cuatro de cada diez tenían problemas de cómo usar análisis con los datos. La mayoría expresaron su deseo de que las decisiones realizadas fueran dirigidas por os datos. Sin embargo para que estas decisiones relacionadas con los datos sean eficaces, deben de estar relacionadas con la estrategia de la empresa y coincidir con sus objetivos (LaValle, 2010).

4.1 Análisis web

Es la rama de análisis de datos, con un enfoque en la información de páginas web, tanto su contenido como el tráfico y recorrido que pasa por ellas (Joyanes,2011).

La analítica web, según la “digital analitics association” “It is the measurement, collection, analysis and reporting of Internet data in order to understand and optimize web usage” “Es la medición, recopilación, análisis y elaboración de informes de datos de Internet con el fin de comprender y optimizar el uso de la web.” (Digital analytics association, 2023) esta definición condensa los objetivos y la esencia de la analítica web en su conjunto, utiliza herramientas de análisis, para poder optimizar la experiencia web del usuario e ir mejorando el diseño y utilidades de la página.

Una vez sabemos a qué se refiere, vamos a aclarar que existen dos analíticas web, la primera con un énfasis especial en la secuencia de clics o “clickstream”, esto el flujo de los clics realizados en la página, esto te da una gran parte de la información que necesitas sobre los usuarios, donde hacen más clics, que llama más la atención y los lugares más importantes de la página (Joyanes, 2013).

Este enfoque tiene su base en el surgimiento de internet y los primeros portales de venta en Internet pese a esto, sobre el año 2010 y el avance da las páginas web, este análisis se vuelve más complejo y completo, cambiando hacia la analítica web 2.0 la cual cuenta con muchos más aspectos que su antecesor, como podemos ver en la imagen (Kaushik, 2011):

Ilustración 12: La analítica web2.0

Elaboración propia. Referencia Kaushik, 2011

Este nuevo enfoque pese a que la secuencia de clics sigue siendo el aspecto central, toma más indicadores (Kaushik, 2011):

- Uso de la perspectiva, que sería realizar un juicio de valor de cómo le ha ido a nuestra página web.
- La inteligencia competitiva, cómo se comporta la competencia, que decisiones toma y como le ha ido económicamente.
- Los comentarios de los clientes, tiene en cuenta los comentarios de los clientes.
- Experimentación, permite ver los resultados de los cambios rápidamente, dado más flexibilidad al cambio en las páginas web.
- Análisis más exhaustivos y completos de los datos, dado que usamos muchas más métricas que antes, lo cual nos permite tener una visión más completa de la página (Kaushik, 2011).

Existe por ejemplo, la “digital analitics association”, esta es una entidad, encargada de gestionar aquello relacionado con analítica web y técnicas similares, para ello publica periódicamente documentos, sobre avances en el sector o cuestiones teóricas sobre el mismo. Uno de estos documentos, indica las definiciones de los indicadores más representativos del Big data, que si bien está un poco desactualizado, la mayoría de su contenido todavía cuenta con validez (Burby, 2007). Existen tres tipos de métricas en análisis web:

- **Cuentas**: un número entero que registra cierta cantidad de información sobre el sitio
- **Ratio**: la división de una cuenta entre otro indicador, muchas veces no siendo un número entero
- **KPI:** una cuenta o ratio, de especial importancia para un sector, siendo la forma más común de representar el análisis web por ejemplo el margen de beneficio neto sería un KPI de finanzas

4.1.1 KPI

Las KPI son de vital importancia para las empresas, usen o no los datos masivos. Existen multitud de métricas que pueden ser usadas en la analítica web, sin embargo, una de las herramientas más utilizadas, en el análisis web, sería *Google analytics* el cual incluye muchos de los KPI más utilizados por otras herramientas y por esto lo considero el más representativo. Las KPI más comunes serían las siguientes:

Ilustración 13: KPI, más importantes en Google.

Referencia Unaoficinavirtual 2016

- **Sesiones**: número de veces donde alguien a interactuado con tu página web, al menos una vez.
- **Usuarios**: Cuantos visitantes únicos han accedido a nuestra página web, cuenta solamente la primera vez que un usuario ha accedido a nuestra página.

- **Número de páginas vistas:** en cuantos lugares de nuestra página, han entrado los usuarios. Por ejemplo, en una tienda de ropa, si se han visto, vaqueros, chaquetas y camisetas, serían tres páginas.
- **Porcentaje de rebotes:** cuantos visitantes han entrado en la página, pero no han interactuado con su contenido.
- **Nuevo visitante/ visitante recurrente**: miden si los usuarios que han visitado nuestra página web han accedido solamente una vez o en cambio, han accedido por lo menos una segunda vez.

Finalmente queda decir que el análisis web es un proceso continuo, en el cual las organizaciones analizan las KPI y aquellos datos más importantes, para recibir las opiniones de los clientes y en turno adaptar esta para mejorar la experiencia de los consumidores (Waisberg, 2009).

4.2 Análisis predictivo

“El análisis predictivo consiste en estudiar los datos históricos y actuales para hacer predicciones sobre el futuro. Usar una mezcla de técnicas matemáticas, estadísticas y de “machine learning” avanzadas para analizar los datos y así determinar y extrapolar las tendencias ocultas.”(Amazon AWS-P, 2023). Esta sería la definición dada por *Amazon* del análisis predictivo, esta resume perfectamente el análisis y como es aplicado en las empresas. Para saber el principal objetivo de este análisis, nos lo da *Google* el cual lo condensa en “que va a pasar a continuación” (Google,2023).

La analítica predictiva no es nada nuevo y lleva siendo usada años, por economistas, científicos e ingenieros para realizar análisis, antes se utilizaban algoritmos estadísticos complicados e historiales para realizar predicciones, sin embargo con la llegada del “machine learning” sus capacidades se vieron expandidas, a niveles antes desconocidos e imposibles por métodos tradicionales(Amazon AWS-P, 2023).

Pero como se encuentra tan ligado este análisis al ”machine learning”, primeramente sería porque estos análisis suelen ser modelos muy complejos, que requieren un cantidad masiva de información, y los ingenieros encontraron que la forma más eficaz de condensar esta información, sería con el aprendizaje automático, esto implican usar técnicas matemáticas y de computación, para enseñar a las IA a realizar o interpretar los datos, esto se encuentran dentro del campo del “machine learning” y permiten a estos modelos realizar análisis mucho más complejos y completos. Estos modelos son normalmente muy complicados y necesitaran ser revisados periódicamente para no quedar obsoletos (Google,2023).

Este análisis, tiene muchos usos y aplicaciones, pero se utiliza principalmente en estas áreas (Amazon AWS-P, 2023):

- **Finanzas:** predecir precios en bolsa, averiguar los riesgos de una inversión o saber si se aproximan bajadas o subidas en los mercados.
- **Venta minorista:** predecirla demanda estacional, que productos pueden venderse mejor, que lugares pueden responder mejor a la publicidad o promociones estacionales o que productos deben de colocarse al frente.
- **Sanidad:** averiguar cómo se va a comportar una enfermedad, identificación prematura de dolencias y enfermedades o evitar dar tratamientos incorrectos.
- **Logística:** saber los horarios de llegada de paquetes y pedidos, es muy importante para un correcto funcionamiento en este campo, proveedores de logística usaran este tipo de análisis para determinar las horas aproximadas de llegadas, basado en tráfico y otras medidas.

Otra de las principales razones por las que este análisis es tan interesante, viene con las aplicaciones de la inteligencia artificial, esta última irrumpió durante el año 2023, como la última tendencia tecnológica, doblando inversiones en su desarrollo, con un 50% de organizaciones, admitiendo haber incorporado inteligencia artificial en al menos un área de su negocio.

Un ejemplo de la eficacia de este análisis junto a las IA, viene de *PepsiCo*, la cual redujo su tiempo de producción y su tiempo de análisis de la demanda del consumidor sustancialmente usando estas técnicas (Eynde,2023).

Para conseguir este vistazo al futuro, este tipo de análisis debe de seguir una serie de pasos determinados (Amazon AWS-P, 2023):

Ilustración 14 Proceso de creeación del análisis predictivo

Fuente elaboración propia. Referencia Eynde, 2023

1. **Determinar el problema:** antes de desarrollar complicados modelos de computación y algoritmos, necesitaremos saber el objetivo de nuestro análisis y para que necesitamos esa información (Eynde,2023).
2. **Obtener y organizar los datos:** recopilación, normalización y gestión de los datos necesarios.
3. **Preprocesar los datos:** Una vez tenemos los datos en bruto, debemos de prepararlos para su análisis, eliminado irregularidades y aquellos valores, que puedan dar errores en el análisis.
4. **Desarrollar los modelos:** una vez sabemos que datos vamos a analizar, debemos de desarrollar modelos que se ajusten específicamente a la naturaleza de esos datos. Por ejemplo, para predecir las ventas de una empresa de ropa de invierno, los datos meteorológicos, serán mucho más importantes que en una tienda de ropa de lujo.
5. **Comprobar resultados y repetir:** Una vez tengamos el modelo, experimentar para ver la eficacia de sus predicciones, y si estas no resultaran satisfactorias, deberíamos de volver al paso anterior y ajustar el modelo hasta que lo sean.

Pese a ser un tipo de análisis, complicado de implementar en las organizaciones, las empresas se beneficiarán enormemente del uso de análisis predictivo. (Amazon AWS-P, 2023)

- **Comodidad:** para los clientes, tener primero aquellos productos que es más probable que compren, les ahorra tiempo, es conveniente y es más probable que compren compulsivamente (Google,2023).
- **Optimización de la línea de suministros:** prediciendo cuando será necesario reponer stock, gracias a patrones en la demanda, unas mejores capacidades de entrega, con optimización de ruta y de tráfico y en general la capacidad de poder predecir la demanda de los productos y cuantos eran necesarios (Beasley, 2021).

Sin embargo nos encontramos con un problema con este tipo de analíticas, y esto ocurre cuando los datos históricos dejan de ser representativos del momento actual. Por ejemplo el COVID 19, fue completamente inesperado y resultó en reacciones nunca antes documentadas, por ejemplo un producto tan común y poco importante como el papel higiénico, acabase convirtiéndose en uno de los productos más demandadas, las mascarillas en cambio sí que era algo más obvio que su demanda aumentaría, pero ningún modelo de predicción hubiera podido ver ese pico en la demanda. Este es realmente el mayor peligro de estos tipos de análisis (Beasley, 2021).

Finalmente decir que en su base la analítica predictiva es muy sencilla, si tenemos 30.000 registros que cuando el cielo se llena de nubes oscuras, en la siguientes horas a empezado a llover, la próxima vez que ocurra este fenómeno podremos suponer que le seguirá la lluvia.

4.3 Análisis de sentimientos

El “Sentiment analysis”, análisis de sentimientos o análisis de opinión, es un tipo de análisis de datos, ampliamente usado por muchas empresas y organizaciones, que utiliza ampliamente técnicas de IA, es especialmente utilizado por *Amazon* y pese a ser similar en la mayoría de organizaciones, este se aproxima más al enfoque de análisis dado por *Amazon*.

“El análisis de opiniones es el proceso de analizar texto digital para determinar si el tono emocional del mensaje es positivo, negativo o neutral” (Amazon-O AWS,2023). Actualmente no es solo texto digital, sino que incluye audios y llamadas, prácticamente cualquier lugar donde un cliente exprese su opinión. El análisis de sentimientos solía ser exclusivamente de textos y comentarios, pero nuevas tecnologías están permitiendo usar audios y tonos para este mismo objetivo. Las fuentes de donde suelen obtener estos datos serían, redes sociales, blogs, foros, cuadros de comentarios y paginas especializadas.

Prácticamente toda la información y datos conocidos, se podrían dividir en dos categorías, datos objetivos y datos subjetivos, los primeros son datos lógicos y sin subjetividad, por ejemplo el agua hierve a 100 grados, la tierra orbita el sol o tuvimos 200 ventas más que el año pasado, estos datos son fáciles de entender y de aplicar, el problema viene con los datos subjetivos, los cuales tienen en cuenta los sentimientos y estado de ánimo de las personas, estos resultan mucho menos predecibles y difíciles de comprender(Joyanes,2013). El análisis de sentimiento intenta conseguir la mayor cantidad de información lógica de este tipo de datos (Mampel, 2023).

Ya mencionamos el concepto de sobreinformación, en el apartado 2.3 el cual, aunque no lo parezca tiene relación con este análisis, ya mencionamos que estamos en la era de la sobreinformación, pues bien es esta sobreinformación, la que hace que tengamos una opinión de las cosas mucho antes, de comprarlas o de verlas. Un estudio de la revista *Harvard Business Review*, estimó que más del 60% de los consumidores deciden si van a comprar algo, antes de realizar ninguna interacción con la tienda o página, basados principalmente en publicidad y primeras impresiones (Joyanes, 2013).

Por esto es tan importante para las empresas este análisis, si averiguan que es gusta ver a los consumidores antes de la compra, pueden aumentar exponencialmente sus ventas (Mampel, 2023).

Con el análisis de sentimientos, podemos monitorizar las opiniones de los usuarios sobre nuestra organización, nuestros productos, que opinión les ha suscitado nuestra campaña de marketing y como es la relación organización-cliente, en las redes sociales dado que muchas organizaciones tienen cuentas oficiales, en diversas redes con el objetivo de mejorar la comunicación con los usuarios, (Joyanes, 2013). por ejemplo, gobiernos como el español o el italiano usan cuentas en *Twitter* y otras redes para realizar anuncios o avisos.

Esta estructura puede cambiar de empresa a empresa y que necesidades va a cubrir, sin embargo el proceso general que siguen la mayoría de empresas, sería el siguiente (Mampel, 2023):

Ilustración 15 Proceso de creación del análisis de opinión

Fuente elaboración propia. Referencia Mampel, 2023

1. **Procesar el lenguaje natural:** el programa recibe los datos que debe de analizar.
2. **Elimina palabras innecesarias:** el programa elimina aquellas palabras, como los conectores que solamente sirven para darle unidad a la frase.
3. **Analiza la opinión recibida:** El programa, "identifica, extrae y estudia los datos obtenidos" (Mampel, 2023, p 1). Usando los diversos tipos de clasificación, separa las palabras importantes y las agrupa según el tipo de clasificación que estemos usando.
4. **Uso:** una vez son clasificados, los datos servirán para realizar gráficas y otros procesos, para mejorar los procesos de la organización y finalmente serán archivados en la base de datos(Shulex, 2023).

El proceso mediante el cual las IA consiguen extraer esta información, varía de modelo a modelo y de las necesidades de la empresa, para Microsoft, quizás un análisis basado en aspectos resulte lo más recomendable, pero Amazon prefiera uno emocional. A continuación explico a que se refieres esto (Mampel, 2023):

- **Polarizado:** el proceso más común, la base de muchos otros sistemas y el más usado por Amazon, divide las palabras o conjunto de palabras en, normalmente tres categorías, positivo, negativo y neutro (así como sus diversos superlativos y diminutivos, como muy o poco).

Asigna un puntaje emocional dentro a cada una de esas palabras, por ejemplo bueno podría tener un calificación de 2 en positivo mientras que maravilloso o muy bueno podría tener una de 4 y muy malo de -4, una vez se han calificados las palabras dependiendo de cuál es la suma general, la opinión se clasificara en una escala del 0 al 100, donde cada segmento será asignado una de las clasificaciones antes mencionadas, estas y la longitud del segmento dependerá de la empresa, por ejemplo, Amazon puede calificar muy positivo de 89 a 100 y Microsoft muy positivo de 95 a 100(Amazon AWS-O, 2023).

- **Basado en aspectos:** el modelo es similar al anterior, pero este busca opiniones sobre aspectos específicos del producto, por ejemplo "este bolígrafo es muy cómodo al escribir, pero se seca rápido".
- **Intención:** divide las palabras en dos categorías en lugar de tres, interesados y no interesados en el producto.
- **Emoción:** se centra en identificar la emoción que el consumidor quería transmitir, en el momento de la realización del comentario, como pueden ser la alegría, tristeza o ira.
- **Mixtos:** Usamos más de una técnica, por ejemplo, usamos en el mismo análisis una mezcla entre polarizado y emoción, todo para obtener mejores conclusiones.

Cuanto más sabes sobre el consumidor mejor puedes vender tu producto, esto es una máxima del comercio, debes de entender las necesidades y limitaciones de tus consumidores, para generar el ambiente de compra óptimo (Marr, 2016). Por esta razón es tan valioso este análisis, permite convertir todas esas opiniones y comentarios, en información para comprender a tus consumidores y cuáles son sus necesidades.

Con todo este entendimiento no es la única ventaja que ofrece el análisis de opinión a las organizaciones, estas son algunas de las más comunes (Amazon AWS-O, 2023):

- **Proporcionar información objetiva:** cuando las personas leen un comentario, es más común que lo hagan de forma subjetiva, por ejemplo, si te encuentras de mal humor es más probable que no seas imparcial en el análisis o un incluso encontrándote en tus plenas facultades, cuando un comentario empieza positivamente, es mucho más probable que la persona analizándolo lo clasifique como positivo, restándole importancia a la segunda parte lo era o no.

- **Mejores productos y servicios:** gracias a esta información objetiva, las organizaciones son capaces de centrarse en los problemas destacados por los clientes, mejor que con un analista humano.
- **Análisis a escala:** Resultaría físicamente imposible revisar la cantidad necesaria de comentarios con exactitud, para cubrir las necesidades de las empresas sin el uso de modelos de análisis de sentimientos.
- **Resultados en tiempo real:** con la aplicación de estos modelos, el tiempo necesario para el análisis se ve reducido, hasta la casi inmediatez.

Si bien es cierto que cuenta con muchas ventajas, también cuenta con problemas y dificultades, en especial por parte de las IA, las cuales todavía necesitan intervención humana para ciertos aspectos (Shulex, 2023).

Como pueden ser términos abstractos, palabras muy sentimentales o con muchas interpretaciones, sufren muchos problemas con sarcasmos, al no saber interpretarlos y tomarlos literalmente, igualmente sufren un problema similar con negaciones, por ejemplo “no esperaba que me gustase tanto”, sería percibido como algo negativo(Mampel, 2023)

Finalmente me gustaría mencionar dos páginas web, las cuales me gustaría poner como ejemplo de utilización de este tipo de análisis, este primero usa análisis polarizado (Soper, 2023) y este otro, realiza uno más completo y complejo, tomando los datos de *Twitter* Csc2, 2023).

Es importante que las empresas tengan en cuenta no solo los comentarios de sus páginas web, sino que los mensajes e interacciones de las redes sociales son muy importantes para ver cómo la gente percibe su marca, hacía que público o nicho deben de centrar sus esfuerzos y recibir “feedback” sobre sus acciones de marketing.

5. El caso de Amazon y su integración del Big data

Amazon es una empresa masiva en el sentido más literal de la palabra, el solo supuso la mitad del crecimiento online estadounidense en 2016 y un 21% del crecimiento del mercado minorista en este país. Su servicio de “Streaming” es el segundo del mundo por detrás de *Netflix*, con 200 millones de abonados y se encuentra disponible en la mayoría de países del mundo, todo esto sin contar la multitud de otros servicios que ofrece, como *Kindle*, *Amazon Music, Twich* y otros muchos.

Por esto solo es lógico que maneje los datos en escalas gigantescas y que tenga que analizar estos datos en escalas gigantescas para poder mantener su posición global, frente a competidores, como *Alibaba* o diferentes páginas nacionales que han ido surgiendo con el avance del comercio digital (Galloway, 2018).

5.1 De donde obtiene los datos

Ya hemos hablado de cómo *Amazon* utiliza una cantidad ingente de datos, en su modelo de negocio esto lo consigue recogiendo el máximo número de métricas y variables posibles de sus usuarios todo, con el objetivo de mejorar al máximo su servicio y proceso. Alguna de las métricas más importantes es:

Interacciones con la tienda de *Amazon*, interacciones con otras páginas de *Amazon* (incluidas *Kindle*, *Twich* y paginas donde tengas que usar tu cuenta *Amazon*), los historiales de compra, secuencia de clics, carrito, lista de deseos, secciones visitadas, actividad en la página, fecha, hora, ubicación, corredores de datos y que han comprado personas similares a tu perfil. Estos serían los principales lugares de donde *Amazon* extrae su información de los usuarios (Ayudaley, 2020).

Con todo estas métricas, se expanden aún más por ejemplo, la fecha y hora se utiliza para poder comparar con otros usuarios que hayan comprado en esa misma franja horaria y estacional, ofreciendo productos similares a estos. Asimismo, basándose en tu ubicación pueden intuir tu grupo demográfico general, e incluso tu nivel de ingresos aproximado.

Finalmente usará datos de sus filiales, como *Prime Video, Twich o Kindle* para acotar aún más sus resultados, comparando tu tiempo y ocupación en estas plataformas con aquellos productos que has comprado y con los productos adquiridos por otras personas con gustos similares.

Por ejemplo, si ocupas tu tiempo en estas plataformas, viendo películas o libros en inglés, la próxima vez que entres a *Amazon* te saldrán recomendados más producto en ingles que antes de realizar estas acciones. *Amazon* llama a esto, visión 360º dado que le permite tener una visión completa sobre sus usuarios(Marr, 2016). Destacar que este concepto se retroalimenta entre páginas, si recientemente has comprado la saga de libros del señor de los anillos, en *KINDLE*, tus próximas recomendaciones de la página principal de *Amazon* y de *Prime Video* tendrán que ver con este último.

5.1.1 Para que usa los datos

Ya hemos introducido las métricas y el objetivo de *Amazon* como entidad, el cual es la personalización, ahora hablaré de sus usos prácticos, primeramente tenemos la **filtración colaborativa**, esta trata de identificar qué productos o servicios va a comprar basándose en las métricas y comparaciones antes mencionados, con especial énfasis en que han comprado personas similares a ti o en un franja y lugar similar al tuyo han comprado (Marr, 2016).

Este sistema ha sido tan exitoso, que *Amazon* a decidió implementarlo directamente en su página, con “Mi Amazon” una página especializada en mostrarle a la persona recomendaciones de productos, llevando la personalización de la compra online al siguiente nivel, dado que es más probable que realices compras compulsivas (Ayudaley, 2020).

Seguidamente tenemos el “recommendation engine” la cual actualmente utiliza elementos de “machine learning” y otros sistemas algorítmicos complejos, que le permiten mejorar, junto con la información obtenida del consumidor, esta básicamente te muestra las recomendaciones que cree que más te van a interesar, en la barra de recomendación, usando el máximo número de métricas. Permite un mejor uso de sus recursos, al darle la capacidad de optimizar la cadena de distribución y la logística general, dado que al poder predecir aunque sea de manera limitada lo que los consumidores van a demandar, supone ahorros en transporte y almacenaje.

Permite prevenir y detectar fraudes e irregularidades que puedan surgir, por ejemplo vemos una gran cantidad de pedidos realizados por un sola persona a un proveedor nuevo y sin otras ventas, esto puede ser un caso de lavado de dinero o algo similar. Permite optimizar los precios, como ya mencionamos, *Amazon* no coge en exclusiva datos de su página web, sino que complementa esta con corredores de datos, esto le permite tener una visión clara de la competencia y saber qué precios le permiten seguir siendo competitivo. Básicamente recoge datos de la competencia para ajustar sus precios a está (Ampofo, 2023).

5.1.2 El enfoque de Amazon

Amazon ha conseguido tanto éxito empresarial, no solo por su modelo de negocio, el cual no deja de ser envidiable, sino que ha sabido capturar el instinto del consumidor, cuando usted compra en *Amazon* sin importar el producto o el precio, *Amazon* va a tratar de que tengas a mejor y más cómoda experiencia posible. Comprar no es siempre agradable, en especial cuando tienes que adquirir productos caros o delicados, con todo *Amazon* trata en todo momento de hacer esto fácil e intuitivo, dándote toda la información y opciones en una misma página, ya antes de meterlo en el carrito, y una vez se encuentra vas a realizar el pago, si está correctamente configurado bastará con dar un clic (Galloway, 2018).

Esta es una forma de mejorar la experiencia web, la sencillez, sin embargo, existe otra más complicada pero mucho más eficaz por su dificultad, la personalización, está última solo es posible usando datos masivos en los sitios web. Las métricas de las que hablamos anteriormente tienen este uso, personalizar la experiencia del consumidor, conseguir gracias a los datos, la mejor experiencia de compra posible. Cuando un usuario entra en *Amazon*, la página se encuentra configurada casi exclusivamente para él, preferencias, gustos o situación estacional, similar a la huella dactilar no existen dos páginas de usuarios de *Amazon* exactamente iguales.

Un ejemplo de esta personalización de la compra, la tenemos con un programa de IA en Alexa, el cual te elige la ropa que deberías de ponerte para cada ocasión, te tienes que sacar una foto con cada pieza de ropa que tengas, entonces Amazon se creará un modelo virtual tuyo y basado en la información, que tenga recopilada sobre ti esta te recomendará que estilo de verías de llevar, así como sugerencias en las futuras compras (PuroMarketing, 2017).

Finalmente el uso de los datos masivos se encuentra profundamente insertado en el modelo de negocio de *Amazon*, confía en este para gran parte de su proceso empresarial, y sin estas técnicas, no sería tan grande como lo es hoy.

6. Conclusiones

Los datos masivos son una parte ya de nuestra vida diaria y de los procesos de las organizaciones, si bien es cierto que cada sector y organización pueden tener su visón personal de los datos masivos, o de como estos deben de ser gestionados, entre ellos siguen compartiendo muchos de los usos y razones de sus utilización.

Considero que este auge que han tenido en los últimos años, es el resultado de un cambio en la sociedad y en como esta funciona, antes eran las organizaciones las encargadas de informar y de crear información, la inmensa mayoría de ejemplo que tenemos de los primeros usos de Big data, como las bibliotecas, logística militar o las tarjetas de censo, eran creadas por y para los gobiernos y empresas, sin embargo con la llegada de internet esta corriente cambio totalmente y se dio paso a una nueva concepción de la información y de los datos.

Las organizaciones empezaron a utilizar estas técnicas, dado que vieron que podían aprovechar un nuevo nicho casi inexplotado de valor, muchas de las grandes empresas digitales del momento usaron técnicas de Big data desde los inicios, tanto Google como Amazon capturaban los datos de sus usuarios y los de todo el internet, para poder generar valor a partir de estos datos (Marr, 2016). Los datos masivos supusieron y suponen una revolución gigantesca para todos los niveles de la sociedad, para las grandes organizaciones, como los gobiernos o *Amazon* les ha dado la oportunidad de

Considero que el objetivo y conclusión del Big data en sus inicios, era ayudar a la toma de decisiones, en especial aquellas organizaciones grandes y con muchas transacciones, sin embargo, actualmente su uso se ha extendido a la mayoría de áreas de las empresas, especialmente en marketing y el uso de la publicidad, pero otras en áreas siguen dando ventajas frente a aquellas organizaciones que no la utilizan.

El uso de los datos masivos debería de ser considerado obligatorio por las empresas, cada vez se obtiene más información de las transacciones y el correcto aprovechamiento de estos pueden suponer la diferencia entre una empresa líder del sector, y una bancarrota.

El uso de los datos masivos por parte de *Amazon* es uno de los mejores ejemplos, de cómo se deben de gestionar los flujos de información, desde sus inicios en 1994 han estado utilizando la información aportada por sus clientes, para mejorar todos los aspectos posibles de su proceso y método empresarial. Si *Amazon* hubiera simplemente tratado de vender la mayor cantidad de producto, sin centrarse en el porqué, como y quien está comprando, jamás sería el gigante que es hoy.

Fue gracias a que usaron los datos masivos en todos los apartados de su empresa, desde lo más pequeño hasta la toma de decisiones de la junta directiva, que han conseguido adaptarse y crecer con el tiempo, sin importar en qué circunstancias se encontrasen.

Finalmente, me tengo que destacar la capacidad de cambio del Big data, cada año surgen nuevas técnicas, tecnologías y herramientas que ayudan a las empresas a gestionar este flujo, con la llegada de las IA y de modelos de aprendizaje, estamos entrando en una nueva forma de gestionar y analizar los datos, que podría revolucionar la forma de ver la información, no diferente de la aparición de los números árabes.

7. Bibliografía

Aguilar, L. J. (2013). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* Alfaomega Grupo Editor.

Amazon AWS-O (2023). ¿Qué es el análisis de opiniones? *Amazon.com*. Recuperado el 13 de junio de 2023, de https://aws.amazon.com/es/what-is/sentiment-analysis/

Amazon AWS-P (2023). ¿Qué es el análisis predictivo? *Amazon.com.* Recuperado el 18 de junio de 2023, de https://aws.amazon.com/es/what-is/predictive-analytics/

Ampofo., M. (2023). How Amazon uses data science and analytics to drive E-commerce success. *Linkedin.com*. Recuperado de https://www.linkedin.com/pulse/how-amazon-uses-data-science-analytics-drive-success-michael-ampofo/

Anthes, G. (2014). Data brokers are watching you. Communications of *The ACM*, 58, 28-30. https://doi.org/10.1145/2686740

Ayudaley. (2020). Amazon y el Big Data. Una historia de éxito. *AyudaLeyProteccionDatos.* Recuperado el 18 de junio de 2023 https://ayudaleyprotecciondatos.es/big-data/amazon/

Bastidas, J. C. (2021). Datos Estructurados, Semiestructurados y No Estructurados. *Linkedin.com*. Recuperado el 20 de mayo, de 2023 https://acortar.link/BzRg1a

BBC News Mundo. (2022). Japón declara la guerra a los disquetes y otros dispositivos obsoletos. *BBC.* Recuperado el 29 de mayo, de 2023 https://www.bbc.com/mundo/noticias-62758191

Beasley, K. (2022). Unlocking the power of predictive analytics with AI. *Forbes*. Recuperado de https://www.forbes.com/sites/forbestechcouncil/2021/08/11/unlocking-the-power-of-predictive-analytics-with-ai/

Big Data Framework. A short history of Big Data. (2021). *Enterprise Big Data Framework*. Recuperado de khttps://www.bigdataframework.org/knowledge/a-short-history-of-big-data/

BOE-A-2018-16673 (2023) Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. *Boe.es*. Recuperado el 12 de junio de 2023, de https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673

Burby, J. Brown, A & WAA Standards Committee. (2007) Web Analytics Definitions.. *Stuker.com*. Recuperado de https://stuker.com/wp-content/uploads/import/i-2731fa96a2de23f5b57b1471f450b9d0-WAA-Standards-Analytics-Definitions-Volume-I-20070816.pdf

Cristina Crespo Garay (2019). La distopía de Orson Welles: ¿Detonante del inicio de la posverdad? . *National Geographic.* Recuperado de https://www.nationalgeographic.es/historia/2019/10/la-distopia-de-orson-welles-detonante-del-inicio-de-la-posverdad

Csc2. (2023) Tweet sentiment visualization app. . *Ncsu.edu*. Recuperado el 1 de junio de 2023, de https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Data lake vs data warehouse: Key differences. (2023).- A Leader in Data Integration & Data Integrity; *Talend*. Recuperado el 1 de junio de 2023, de https://www.talend.com/resources/data-lake-vs-data-warehouse/

Davenport, T., Bart, P., & Bean, R. (2012). How “Big Data” is Different. *Hbs.edu*, 54, 22-24 https://www.hbs.edu/ris/Publication%20Files/SMR-How-Big-Data-Is-Different_782ad61f-8e5f-4b1e-b79f-83f33c903455.pdf

DeAngelis, S. (2018). The seven “vs” of big data . *Enterra Solutions.* Recuperado el 20 de mayo de 2023, de https://enterrasolutions.com/the-seven-vs-of-big-data/

Diebold, F. X. (2012). A personal perspective on the origin(s) and development of “big data”: The phenomenon, the term, and the discipline, second version. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2202843

Digital analytics association. (2023). *Digitalanalyticsassociation.org*. Recuperado el 15 de junio de 2023, de https://www.digitalanalyticsassociation.org/

El Haj Tirari, H. B. A. T. B. M. (2014). Predictive Analysis of Big Data in Retail Industry. *Rabbat university*. Recuperado el 18 de junio de 2023 https://acortar.link/QLnLFj

Eynde V, K. (2023). AI and predictive analytics: Moving from insights to foresight. *Forbes*. Recuperado de https://www.forbes.com/sites/microsoft_/2023/05/08/ai-and-predictive-analytics-moving-from-insights-to-foresight/

Favaretto, M., De Clercq, E. M., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*. Recuperado de https://doi.org/10.1371/journal.pone.0228987

Figuerola, N. (2013). Gestión del Conocimiento. *Articulospm*. Recuperado 3 de junio de 2023, de https://articulospm.files.wordpress.com/2013/08/gestic3b3n-de-conocimiento-dikw.pdf

Galloway, S. (2018). *Four: The hidden DNA of Amazon, apple, Facebook, and Google*. Conecta.

Google (2023) ¿Qué son las analíticas predictivas y cómo funcionan? . *Google Cloud*. Recuperado el 11 de junio de 2023, de https://cloud.google.com/learn/what-is-predictive-analytics?hl=es

IBM (2017). What is big data? More than volume, velocity and variety *Ibm.com.* Recuperado el 24 de marzo de 2023, de https://developer.ibm.com/blogs/what-is-big-data-more-than-volume-velocity-and-variety/

IBM . (2018) Data lakes and data swamps. *Ibm.com*. Recuperado el 1 de junio de 2023, de https://developer.ibm.com/articles/ba-data-becomes-knowledge-2/

ImageNet. (2023). *Image-net.org*. Recuperado el 10 de junio de 2023, de https://www.image-net.org/

Incibe. (2023). *Incibe.* Recuperado 14 de junio de 2023, de https://www.incibe.es/

Kaushik, A. (2011). *Analítica Web 2.0: El arte de analizar resultados y la ciencia de centrarse en el cliente*. Grupo Planeta

Keskar, V., Yadav, J., & Kumar, A. (2020). 5V's of Big Data Attributes and their Relevance and Importance across Domains. *International Journal of*

Innovative Technology and Exploring Engineering.9. 154-163. Recuperado de https://www.ijitee.org/wp-content/uploads/papers/v9i11/K77090991120.pdf

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Gartner*. Recuperado de https://www.gartner.com/en/chat/blog

Latto, N. (2023). Corredores de datos: todo lo que necesita saber. Corredores de datos: todo lo que necesita saber. *Avast.* Recuperado 14 de junio de 2023, de https://www.avast.com/es-es/c-data-brokers (Latto, 2023)

LaValle, S., Hopkins, M. S., Lesser, E., Shockley, R., & Kruschwitz, N. (2010). Analytics: The new path to value*. MIT Sloan Management Review*. Recuperado de https://sloanreview.mit.edu/projects/analytics-the-new-path-to-value/

LaValle, S., Hopkins, M. S., Lesser, E., Shockley, R., & Kruschwitz, N. (2010). Analytics: The new path to value*. MIT Sloan Management Review.* Recuperado de https://sloanreview.mit.edu/projects/analytics-the-new-path-to-value/ (LaValle, 2010)

Lee, B. Y. (2022). Fake Eli Lilly Twitter Account Claims Insulin Is Free, Stock Falls 4.37%. *Forbes*. Recuperado de https://www.forbes.com/sites/brucelee/2022/11/12/fake-eli-lilly-twitter-account-claims-insulin-is-free-stock-falls-43/

Lohr, S. (2012). How Big Data Became So Big - Unboxed. *The New York Times*. Recuperado de https://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html

Maheshwari, R. (2015). 3 V's or 7 V's - what's the value of Big Data? *Linkedin.com.* Recuperado de https://www.linkedin.com/pulse/3-vs-7-whats-value-big-data-rajiv-maheshwari/

Mampel, P. (2023). Sentiment Analysis: ¿Qué es? La Guía Definitiva Para Monitorizar el Análisis del Sentimiento. *Ringover*. Recuperado de https://www.ringover.es/blog/sentiment-analysis

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey*. Recuperado de http://abesit.in/wp-content/uploads/2014/07/big-data-frontier.pdf

Marr, B. (2014). Big Data: The 5 Vs Everyone Must Know.*Linkedin.com.* Recuperado el 24 de marzo de 2023, de https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/

Marr, B. (2016b). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. John Wiley & Sons.

Marr, B. (2021). What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone. *BernardMarr.com* Recuperado el 24 de marzo de 2023, de https://bernardmarr.com/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*. Recuperado de https://www.tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf

Mimmo (2020) Big Data, l'indagine conoscitiva di tre Autorithy.. *Tech_Parliamone*. Recuperado el 11 de junio de 2023 https://techparliamone.altervista.org/big-data-lindagine-conoscitiva-di-tre-autorita/?doing_wp_cron=1687167789.4904220104217529296875

Nambiar, A. M., & Mundra, D. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big data and cognitive compute*r, 6, 132. https://doi.org/10.3390/bdcc6040132

Niculescu. V. (2020). *Researchgate.net*. Recuperado el 3 de marzo de 2023, de https://www.researchgate.net/figure/The-7Vs-of-Big-Data-Volume-Velocity-Variety-Variability-Veracity-Value-and_fig1_341622174

Oracle. ¿Qué es un Data Lakehouse? (2023). *Oracle.com*. Recuperado el 6 de junio de 2023, de https://www.oracle.com/es/big-data/what-is-data-lakehouse/

Panimalar, A., Shree, V., & Kathrine, V. (2017). The 17 V's Of Big Data. *International Research Journal of Engineering and Technology*, 4. 329-333 https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf

Powerdata (2023). Data lake: definición, conceptos clave y mejores prácticas. *Powerdata.es.* Recuperado el 1 de junio de 2023, de https://www.powerdata.es/data-lake

Powerdata (2023). Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos. *Powerdata.es*. Recuperado el 1 de junio de 2023, de https://www.powerdata.es/data-warehouse

Privacy Bee What are data brokers? (2020,). *Privacy Bee.com.* https://privacybee.com/blog/what-are-data-brokers/

PuroMarketing. (2017). Cómo Amazon usa la Inteligencia Artificial y el Big data para vender más. *PuroMarketing*. https://www.puromarketing.com/12/28900/como-amazon-usa-inteligencia-artificial-big-data-para-vender-mas

Rai, D. (2020,). The 7 vs of Big Data in HR analytics. *CHRMP*. Recuperado el 18 de marzo de 2023, de https://www.chrmp.com/the-7-vs-of-big-data-in-hr-analytics/

Ramon Soto (2022) Datos, patrones y conocimiento. . *BI5ON*. Recuperado el 20 de junio de 2023, de https://www.bi5on.com/blogs/entry/43-datos-patrones-y-conocimiento/

Rayaprolu, A. (2023). How much data is created every day in 2023? *Techjury*. Recuperado el 3 de junio de 2023, de https://techjury.net/blog/how-much-data-is-created-every-day/

Rojas, V. (2022). Nuevas tecnologías en la detección temprana del cáncer: Big Data e Inteligencia Artificial para salvar vidas. *Telefónica*. https://www.telefonica.com/es/sala-comunicacion/blog/nuevas-tecnologias-en-la-deteccion-temprana-del-cancer-big-data-e-inteligencia-artificial-para-salvar-vidas/

Rulanitee. (2018). Big Data, what does it mean? – *Tendai*. Recuperado el 18 de marzo de 2023 http://tendaibepete.com/2018/02/09/big-data-what-does-it-mean/

Russom, P. (2011). BIG DATA ANALYTICS. *TDWI RESEARCH.* Recuperado de https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

SAS (2023) Big Data: What it is and why it matters. *SAS* Recuperado el 1 de junio, de 2023https://www.sas.com/en_us/insights/big-data/what-is-big-data.html#:~:text=Big%20data%20refers%20to%20data,around%20for%20a%20long%20time.

Shulex. (2023) How does Amazon use sentiment analysis? . *Voc.Ai.* Recuperado el 19 de junio de 2023, de https://www.voc.ai/blog/how-does-amazon-use-sentiment-analysis

Siegler, M. G. (2010). Eric Schmidt: Every 2 days we create as much information as we did up to 2003. *TechCrunch*. https://techcrunch.com/2010/08/04/schmidt-data/

Soares, S. (2013). Not Your Type? Big Data Matchmaker On Five Data Types You Need To Explore Today - *DATAVERSITY*. Recuperado el 18 de marzo de 2023, https://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/ (Soares, 2013)

Sohaib, A., Muhammad Jameel Arshad, & Irshad Ahmed Sumra. (2019). 7, Vs of Big Data: A Survey. *Engineering science and technology international research journal*, 3, 74. http://www.estirj.com/Volume.3/No.4/12Sohaib34.pdf

Soper, D. (2023). Free sentiment analyzer. *Danielsoper.com.* Recuperado el 1 de junio de 2023, de https://www.danielsoper.com/sentimentanalysis/default.aspx

Statista. (2016). Photo of the day: Infographic, number of worldwide M2M connections (2014 to 2020). *Electrical Engineering News and Products.* https://www.eeworldonline.com/photo-of-the-day-infographic-number-of-worldwide-m2m-connections-2014-to-2020/

Tableau (2023) Guía de visualización de datos para principiantes: definición, ejemplos y recursos de aprendizaje. *Tableu*. Recuperado el 3 de junio de 2023, de https://www.tableau.com/es-mx/learn/articles/data-visualization

Unaoficinavirtual (2016) Metricas google analytics.. *Unaoficinavirtual.com*. Recuperado el 20 de junio de 2023, de https://unaoficinavirtual.com/metricas-google-analytics/

University of Baltimore (2019). the economic cost of bad actors on the internet *Cheq.ai*. Recuperado de https://s3.amazonaws.com/media.mediapost.com/uploads/EconomicCostOfFakeNews.pdf

Vive. (2020). Las tres V del Big Data: todo un reto por su volumen, variedad y velocidad. UNIR. https://www.unir.net/ingenieria/revista/3-v-big-data/

Waisberg, D., & Kaushik, A. (2009). Web Analytics 2.0: Empowering Customer Centricity. Psu.edu. Recuperado el 11 de junio de 2023, de https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f804b8c0f66b28220d64060e27892dbe0a7a3baa

Walter. (2018). Amazon and big data. *Digital Innovation and Transformation*. Recuperado de https://d3.harvard.edu/platform-digit/submission/amazon-and-big-data/