LECCIÓN 1:

INTRODUCCIÓN AL BIG DATA

Como hemos visto, cuando hablamos de Big Data lo hacemos de un gran volumen de datos. Estos datos pueden estar estructurados, semi estructurados o no estar estructurados. Vamos a ver cómo son este tipo de datos:

1. **Datos estructurados.** Son los que contienen longitud y formato, y que pueden ser ordenados y procesados de forma sencilla y almacenados en una base de datos. Podemos formarnos una imagen visual e imaginar un archivo en el que los datos se ordenan por fecha o por orden alfabético, por ejemplo. Serían los que proceden de un censo, de operaciones bancarias, de compras online, de encuestas… Entre otros orígenes.

2. **Datos no estructurados.** Son los que no tienen un formato y que no pueden ser almacenados en una tabla. Pueden ser textos, como los comentarios en redes sociales, o un documento procedente de un editor de textos, o un audio, un vídeo o una fotografía.

3. **Datos semi estructurados.** Aunque tienen una cierta organización interna o marcadores para poder procesar sus elementos, no pueden encuadrase dentro de los datos estructurados. Por ejemplo, una página Web, que cuenta con datos estructurados, como el contenido en HTML o los metadatos, y no estructurados como el texto, las imágenes o los vídeos.

Para entender mejor este asunto, hagamos un punto y aparte con los datos no estructurados. Nos servirá para entender mucho mejor este concepto.

1.1 LOS DATOS NO ESTRUCTURADOS

Los datos no estructurados son aquellos que se presentan en bruto y que carecen de organización. ¿Podrían ser convertidos en datos estructurados? En teoría sí, pero no compensaría el tiempo, y, por ende, el dinero que conllevaría. Además, en algunos casos es más sencillo estructurar esos datos sin estructura y en otros es francamente complicado.

Por ejemplo, dentro de un correo electrónico hay cierta información que se puede estructurar sin grandes complicaciones, como el remitente, la hora de envío o el destinatario. Pero hay otra que es más complicado estructurar, como el texto del correo.

Estos son los datos no estructurados, como vemos es un listado concreto:

- Correos electrónicos
- Archivos de procesadores de texto
- Hojas de cálculo
- Imágenes digitales
- Archivos PDF
- Vídeo
- Audio
- Publicaciones en medios sociales

Son archivos que se pueden almacenar sin necesidad de que el sistema entienda el formato del archivo en cuestión. Por lo que, al no contar con un contenido

organizado, no se pueden almacenar de forma estructurada.

Es en el tema de los datos no estructurados donde el Big Data más tiene que avanzar para conseguir sacar más rendimiento de ellos.

1.2 LAS UNIDADES DE INFORMACIÓN, DEL BIT AL GEOPBYTE

Para entender el Big Data tenemos que tener claro un primer concepto, qué es un bit. Se trata de la mínima unidad de información y solo puede tener dos valores, cero o uno. Cada grupo de 8 bits recibe el nombre de byte o de octeto.

Veamos ahora la clasificación de bits, nos servirá para entender mejor este concepto:

Nombre:	**Símbolo:**	**Equivale a:**
Byte	B	
Kilobyte	KB	1024 Byte
Megabyte	MB	1024 Kilobyte
Gigabyte	GB	1024 Megabyte
Terabyte	TB	1024 Gigabyte
Petabyte	PB	1024 Terabyte
Exabyte	EB	1024 Petabyte
Zettabyte	ZB	1024 Exabyte
Yottabyte	YB	1024 Zettabyte
Brontobyte	BB	1024 Yottabyte
Geopbyte	GeB	1024 Bronobyte

Además, existe otro valor llamado Nibble que equivale a 4 bits. Este valor se emplea a la hora de realizar representaciones hexadecimales y también en los códigos binarios decimales, BCD. El BCD se emplea para representar un cierto número de decimales. No olvidemos que cada dígito se encuentra representado por un Nibble.

Es decir, en un Byte hay 8 bits y está compuesto por dos Nibbles, superior e inferior, que a su vez están compuestos por 4 bits cada uno.

En un bit solo tenemos dos posibles combinaciones:

- 1
- 0

Mientras que en 4 bits la cifra de combinaciones es 16:

- 0000
- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
- 1000

- 1001
- 1010
- 1011
- 1100
- 1101
- 1110
- 1111

Averiguar cuántas combinaciones tenemos es muy sencillo, se toma como base "2" y se eleva a "n", donde "n" es el número de veces que se repite la cantidad de bits. Así:

1 bit = 2^1 = 2 combinaciones.

2 bits = 2^2 = 4 combinaciones.

3 bits = 2^3 = 8 combinaciones.

4 bits = 2^4 = 16 combinaciones.

6 bits = 2^6 = 64 combinaciones.

8 bits = 2^8 = 256 combinaciones.

Etc.

Esta tabla nos resultará de utilidad:

Nibble en binario:	Valor Hexadecimal:	Valor Decimal:
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0100	6	6
0111	7	7
1000	8	8
1001	9	9
1010	A	10
1011	B	11
1100	C	12
1101	D	13
1110	E	14
1111	F	15

No olvidemos que cualquier letra, y por tanto cualquier palabra, se puede expresar en Bytes.

1.3 BIG DATA: VOLUMEN, VARIEDAD, VELOCIDAD, VERACIDAD Y VALOR DE LOS DATOS

La base del Big Data se conoce como sus 5 Vs:

- **Volumen de los datos.** Ya lo hemos dicho y lo volveremos a repetir a lo largo de este curso, la cantidad de datos con la que contamos cada vez es mayor, y además aumenta de manera exponencial. Para analizar cualquier asunto es necesario contar con un buen número de datos. De hecho, gracias al Big Data la tendencia en los últimos tiempos es analizar la totalidad de los datos con los que se cuenta, en lugar de tomar tan solo una muestra de ellos. Hay que tener en cuenta que cuando hablamos de grandes cantidades de datos los hacemos de Terabytes o Petabytes. Gracias al Big Data conseguimos controlar este volumen tan alto de datos.

- **Variedad de los datos.** También lo hemos apuntado con anterioridad, pero lo repetimos, existen datos estructurados, semi estructurados y no estructurados. Debemos atender a esta diferenciación y tenerla en cuenta a la hora de interpretar los datos.

- **Velocidad de los datos.** Una de las mayores ventajas que presenta el Big Data es sin lugar a dudas que podemos analizar los datos prácticamente en tiempo real. Es lógico que esto sea así. Decimos esto porque el flujo de entrada de datos es constante, por lo que no podríamos esperar a que su recopilación haya concluido. De hecho, no va a suceder nunca.

- **Veracidad de los datos.** Dentro de todos los datos que vamos recopilando, no todos serán verídicos y algunos se habrán recolectado de manera incorrecta. Por lo que uno de los desafíos que presenta el Big Data es detectar precisamente esos fallos y solo extraer los patrones reales.

- **Valor de los datos.** El valor está muy relacionado con la veracidad, solo

que en este caso se analiza la valía de la información antes de recopilarla. Una vez que se han captado dichos datos, se procederá a contrastar su veracidad.

1.4 ¿CUÁNTOS DATOS SON BIG DATA?

¿Cuántos datos hay que manejar para hablar de Big Data? No existe una cifra, pero lo habitual es que se utilice para referirse a petabytes y exabytes de datos como mínimo, ya que son difíciles de integrar.

Según los expertos, el Big Data se basa en tres pilares fundamentales:

- **Tecnología.** Se parte de tratar de maximizar la potencia de cálculo y la exactitud algorítmica, con el fin de reunir, analizar, vincular y comparar grupos de datos masivos.

- **Análisis.** Se busca identificar patrones dentro de grandes datos.

- **Mitología.** Se basa en que los datos masivos dan una mayor inteligencia y conocimiento, lo que puede crear ideas que antes eran imposibles y que son verdaderas, objetivas y precisas.

1.5 LOS 4 PASOS DEL BIG DATA

Veamos ahora los 4 pasos del Big Data. Lo haremos de una forma casi esquemática. Un pequeño esbozo que nos servirá para ir entendiendo mejor el curso.

1.5.1 Conseguir los datos

Como es lógico, el primer paso dentro del proceso del Big Data es recopilar los datos. Eso sí, los datos no se recopilan sin ton ni son, sino que es necesario contar con información que resulte confiable, de calidad y que tenga sentido.

Vamos a poner un ejemplo para que nos sea más sencillo comprender lo importante que resulta este primer paso. Imaginemos que estamos asistiendo a un concierto, muchos de los asistentes hacen comentarios utilizando un hashtag, lo hacen desde el mismo lugar, y casi al mismo tiempo. Para poder recabar información interesante de lo que ha ocurrido es necesario que solo tomemos en cuenta los datos que sean precisos. Es decir, que estén relacionados con lo que deseamos investigar.

1.5.2 Procesar los datos

Cuando ya hemos recopilado todos los datos, llega el momento de procesar la información. Se hace necesario debido a que todos esos datos estarán en distintos formatos, no seguirán ningún orden y además utilizarán varios registros. De no modificarlos no podríamos acceder a ellos.

Por ejemplo, si vamos a recopilar datos que tienen que ver con el idioma, en algunos datos figurará como español, en otros como castellano y algunos otros incluso como España. Es el momento de cambiarlos todos a español.

1.5.3. Almacenar los datos

Una vez que hemos recopilado y procesado los datos, llega el momento de almacenarlos en una gran base de datos.

Las bases de datos empleadas pueden tener diferentes estructuras y en su interior los datos se almacenarán de diversas formas. Por claves, nodos o variables, son algunas de ellas.

La manera en la que organizamos la información va a depender de cómo posteriormente vayamos a analizarla.

1.5.4. Analizar los datos

El paso fundamental dentro del Big Data es el análisis, puesto que si no sabemos realizar un análisis adecuado de todos los datos que recabamos, todo el proceso habrá sido baldío.

La forma en la que analizaremos los datos va a depender de la base de datos que hayamos elegido para almacenarlos.

1.6 EL BIG DATA Y EL INTERNET DE LAS COSAS

Si al principio decíamos de que seguro que habíamos escuchado hablar del término Big Data, podemos decir lo mismo de otro, nos referimos al Internet de las cosas. ¿A qué se refiere? El Internet de la Cosas nos habla de conectar todo tipo de objetos que tenemos en nuestro hogar, desde las ventanas a la cafetera, pasando por cualquier otro aparato inteligente con el que contemos.

El fin no es solo ejercer control sobre ellos, sino sacar el máximo rendimiento a cada uno de esos aparatos. En el mundo de los negocios se le podrá sacar un gran rendimiento a la unión de ambos. Será una herramienta fundamental para conocer, por ejemplo, los hábitos del consumidor.

Pongamos un ejemplo, vamos a imaginar que tenemos una empresa que se dedica a la venta de cafeteras con cápsulas. Con el uso del Big Data unido al Internet de las Cosas podemos saber qué tipo de café toman nuestros clientes, a qué hora, lo que nos puede llevar a discernir a qué hora se levantan o a qué hora comen.

Pero, además, si está conectado con otros aparatos, como es lógico, también vamos a conocer otros datos de los demás aparatos. Debemos dejar claro que toda la información recabada será confidencial y anónima, no sabremos a qué cliente pertenece.

Pero en otras ocasiones, cuando nuestro cliente lo autorice, dejará de ser una información anónima y servirá para que le demos a cada cliente justo aquello que necesita.

Si sabemos analizar de la manera adecuada todos esos datos obtendremos una información tan valiosa que puede hacer que nuestra empresa crezca de manera exponencial.

Sin duda, una puerta que esconde innumerables posibilidades y que, si somos sinceros, pocos han traspasado por ahora.

1.7 ¿QUIÉN SE OCUPA DEL BIG DATA?

Decíamos que de momento no se está sacando todo el rendimiento posible al Big Data unido al Internet de las Cosas, pero podríamos decir lo mismo del Big Data en sí. Uno de los problemas es que no existen por el momento el suficiente número de profesionales que sean capaces de ocuparse de este tipo de trabajo.

Los expertos auguran que será una de las profesiones más demandadas a corto plazo, de hecho, ya lo es.

No podemos obviar que analizar una gran cantidad de datos conlleva tiempo y dinero, por supuesto también el trabajo de un profesional especializado en la materia.

Hoy por hoy un gran número de empresas no cuentan con un profesional especializado en este asunto, por lo que no están sacando el rendimiento adecuado a los datos con los que cuentan. Dentro de poco esto cambiará y controlar los datos no será una opción, será imprescindible.

Pongamos un ejemplo para entender mejor lo importante que resulta para las empresas manejar de la manera adecuada los datos que se encuentran a su disposición. Imaginemos que contamos con una cadena de tiendas. Una de ellas ha agotado cierto producto, del que quedan pocas unidades en el almacén. Puede que lo lógico sea reponer el producto en dicha tienda.

Claro que, si analizamos bien los datos, comprobamos que hay otras dos tiendas más que han vendido el modelo, puesto que están ubicadas en sendas localidades en las que el tiempo es mejor, y es una prenda orientada al verano.

Dentro de esas dos tiendas, una de ellas está situada en una ciudad en la que se espera que en los próximos días llegue un buen número de turistas. Por lo que, analizando todos estos datos, concluimos que la mejor opción es reponer

la prenda en esta última tienda. Es un ejemplo muy simplificado, pero que sirve para ilustrar lo útil que resulta analizar bien los datos.

De ahí, que sea tan importante prestar atención a todo lo que vamos a aprender a lo largo de este curso.

1.8 ¿CÓMO ALMACENAMOS TANTOS DATOS?

Uno de los grandes desafíos a los que se enfrenta el Big Data es sin duda conseguir almacenamiento de cantidades de datos ingente. No sirven los sistemas tradicionales, por lo que hay que buscar otras alternativas.

Estas son algunas de las opciones que están a nuestra disposición para almacenar datos:

- **Nube híbrida**. Es un sistema más avanzado con respecto a la nube tradicional. Estas nubes híbridas cuentan con un hardware o software propio y se puede acceder a la nube desde donde prefiramos. Debemos aclarar que existen tres tipos de nube. La primera es la nube privada, que es aquella que está gestionada para un único cliente que controla todo lo relacionada con ella. La segunda es la nube pública, en la que diferentes clientes comparten el almacenamiento. En tercer lugar, encontramos la nube híbrida, que es una mezcla de las dos anteriores. De esta forma, quien usa la nube tiene una parte que le pertenece y comparte otras, pero siempre de forma controlada. Los datos se almacenarán teniendo en cuenta su importancia.

- **Memoria Flash.** Gracias a la memoria Flash y a los SSD que se basan en ella, es posible la existencia de centros de datos que han sido creados para almacenar grandes cantidades de datos. La memoria Flash almacena la información desde un semiconductor.

- **I-SDS.** Nos ayuda a organizar mejor y en menos tiempo los datos gracias a infraestructuras que se gestionan mediante un software inteligente. Es capaz de controlar una gran cantidad de datos, conseguir los resultados más frecuentes y aproximados a aquellos que se buscan.

- **Almacenar archivos en frío.** Consiste en almacenar los datos que revisten una menor importancia en discos que sean más lentos. El fin de esta

táctica es liberar los discos que trabajan más rápido, donde nos interesa almacenar los datos que más se utilizan. Se suele utilizar dentro de los negocios para almacenar los datos que ya tienen un tiempo y que no es preciso que estén constantemente asequibles. El reto de este tipo de archivos es discernir cuáles son los datos más importantes.

Por supuesto, con independencia de la fórmula que hayamos escogido para almacenar los datos, debemos asegurarnos de que en ningún momento se vea comprometida la seguridad, la integridad y el rendimiento de los datos que manejamos.

El almacenamiento debe permitir que extraigamos los datos de manera sencilla y utilizando diversos métodos. Además, los datos deben ser útiles.

1.9 MÁS SOBRE EL BIG DATA

Debemos dejar claro que cuando hablamos de Big Data lo hacemos de un número de datos tan elevado que nada tiene ver con lo que estamos acostumbrados y que para manejarlos necesitamos contar con nuevas herramientas.

Es necesario contar con las infraestructuras que sean las adecuadas y con una tecnología preparada para poder procesar esas grandes cantidades de datos. Con independencia de que estén estructurados o no.

Es necesario que ese compendio de datos se convierta en información que seamos capaces de leer y nos lleve a tomar las decisiones correctas, aún en los casos que tengamos que hacerlo en tiempo real.

Parece mentira, pero a día de hoy no todas las empresas tienen implantado un sistema de Big Data, pero las que sí lo han hecho marcan la diferencia. Ya que pueden entender cómo es su cliente, qué hábitos e intereses tiene y qué necesita. Cuando averigüemos esto último, estaremos preparados para dárselo y esto marcará la diferencia con la competencia.

No es un aspecto menor. Pongámonos en la piel del consumidor, que no es muy difícil, ya que todos los somos. Ahora pensemos en una empresa que nos ofrece justo lo que necesitamos, es decir, sabe lo que nos hace falta y además hace que llegue a nosotros su oferta. Sin duda, estaremos encantados de tratar con una compañía así.

Las empresas no han descuidado el asunto del manejo de datos, ya que llevan tiempo trabajando con el Data Warehouse, veamos en qué consiste.

1.10 DATA WAREHOUSE

La necesidad de almacenar información no es nueva, aunque nunca antes habían existido cantidades de datos tan enormes como en la actualidad. Hace tiempo se creó el concepto de Data Warehouse, fue término creado por Bill Inmon. Este Almacén de datos aportaba soluciones extras a los tradicionales Centros de Información que se utilizaban únicamente para almacenar datos.

Según Inmon, el Data Warehouse se basa en almacenar la información de manera homogénea y confiable. Estas son sus características principales:

- **Integración.** Los datos que almacenamos en este sistema es necesario que estén integrados en una estructura consistente. A la vez que la información se estructura en diferentes niveles.

- **Temático.** Es aconsejable ordenar los datos por tema, de esta forma es más sencillo acceder a ello y entenderlos.

- **Histórico.** Los datos que almacenamos en el Data Warehouse sirven para analizar las tendencias. Por lo que se estudian dentro de una variable de tiempo.

- **Información permanente.** La información que se guarda en Data Warehouse tiene el fin de ser consultada, no de ser modificada posteriormente. Es decir, cuando se incorporan nuevos datos, no se alteran los anteriores que ya estaban almacenados.

1.10.1 Los metadatos y el Data Warehouse

Dentro del almacenamiento Data Warehouse tiene su espacio los metadatos. ¿Qué son los metadatos? Son los datos que tienen información de otros datos.

Con ayuda de ellos podemos controlar cuándo llegó la información, de dónde procede o su fiabilidad, por poner solo algunos ejemplos.

Los metadatos tienen que cumplir una serie de requisitos, que dependen en buena medida de la persona a la que van dirigidos:

- **Dar soporte al usuario final.** Deben dar soporte al usuario final, permitiéndole acceder al Data Warehouse utilizando su propio lenguaje. Dándole a conocer el tipo de información que existe y cuál es su significado. Sirve para elaborar informes, consultas o análisis.

- **Dar soporte a los técnicos.** Dar soporte a los técnicos que son responsables del Data Warehouse en la gestión de la información, en la administración de esta herramienta o en la extracción de la información, por ejemplo.

Aunque no es un sistema nuevo, se le puede seguir sacando todavía partido, ya que tiene diferentes ventajas:

- Nos da información para tomar buenas decisiones dentro de nuestro negocio.

- Nos ayuda a encontrar relaciones ocultas dentro de los datos, lo que será de gran utilidad para nosotros.

- Nos ayuda mediante los datos que nos proporciona a aprender de lo que ha sucedido en el pasado, para predecir lo que ocurrirá en el futuro.

- Nos permite implantar sistemas de gestión de las relaciones con el cliente.

- Nos ayuda a optimizar mejor nuestros recursos, gracias a las estadísticas o a los informes que podemos generar.

1.11 DECÁLOGO DEL BIG DATA

Para concluir esta introducción que hemos hecho al Big Data, conoceremos un decálogo creado por la multinacional FICO, que es líder en soluciones de analítica y que ayuda a las organizaciones de más de 80 países. Es toda una experta en el Big Data. Prestemos atención a estos 10 puntos:

1. La necesidad de aprender un nuevo lenguaje de programación como Groovy, Pig o Apache, que puedan desarrollar aplicaciones y sistemas optimizados para gestionar y analizar el Big Data.

2. Reducir las infraestructuras tecnológicas, mejorando la capacidad de estas para digerir grandes cantidades de datos, para lo que se pueden usar herramientas como Hadoop MapReduce de acuerdo con FICO.

3. Además de acumular datos históricos, es importante aplicar sistemas que analicen los datos en el momento en el que se recogen. Como dice FICO y recoge SiliconWeek, la analítica en tiempo real es el futuro del Big Data.

4. Crear una mayor red de relaciones profesionales entre el equipo de analítica y el resto de departamentos de negocio.

5. Explorar nuevas fuentes de datos y nuevos tipos de información, que puede ayudar a mejorar la toma de decisiones.

6. Tener en cuenta que la cantidad de datos sin estructurar (vídeos, blogs, datos procedentes de redes sociales) es cientos de veces mayor que la de los datos estructurados. Y estos datos sin estructurar pueden ser muy importantes para el negocio.

7. Mejorar la forma de mostrar los resultados de un análisis de los datos, lo que facilita la toma de decisiones en la compañía.

8. Es importante compartir información entre los diferentes departamentos de la compañía.

9. Colaborar con expertos fuera de la empresa.

10. Implantar estrategias de aprendizaje en la infraestructura del Big Data acelera los avances en la analítica.

Terminamos aquí esta introducción al Big Data, hemos conocido las bases para poder aprender más acerca de los datos masivos, con el fin de conseguir almacenarlos y gestionarlos de la manera adecuada.

Antes de seguir adelante, vamos a adquirir unas nociones básicas sobre el Marketing Online, que serán imprescindibles para poder entender plenamente el Big Data.

RESUMEN LECCIÓN 1:

- Big Data es el proceso para aglutinar una gran cantidad de datos, para poder analizarlos casi en tiempo real. Se busca encontrar ciertos patrones ocultos, recurrentes o correlaciones diferentes.

- Se necesitan herramientas y técnicas de almacenamiento que estén encaminadas a este gran volumen de datos.

- En el Big Data los datos pueden estar estructurados, semi estructurados o no estar estructurados.

- Los estructurados tienen longitud y formato, pueden ser ordenados y procesados de forma sencilla y almacenados en una base de datos.

- Los datos no estructurados no tienen un formato y no pueden ser almacenados en una tabla.

- Los datos semi estructurados tienen una cierta organización interna o marcadores para poder procesar sus elementos, pero no pueden encuadrase dentro de los datos estructurados.

- Un bit es la mínima unidad de información y solo puede tener dos valores, cero o uno.

- Las 5 Vs del Big Data son: Volumen, Variedad, Velocidad, Veracidad y Valor.

- Los 4 pasos del Big Data son: Conseguir los datos, Procesar los datos, Almacenar los datos y Analizar los datos.

- El Internet de la Cosas hace referencia a que todos los dispositivos inteligentes de un hogar estén conectados entre sí. A partir de ahí se puede

obtener muchos beneficios del conocimiento que nos proporciona.

- Opciones para almacenar datos: Nube híbrida, Memoria Flash, I-SDS y Almacenar archivos en frío.

- El Data Warehouse se basa en almacenar la información de manera homogénea y confiable.

- Los metadatos son los datos que tienen información de otros datos.