ENAE
BUSINESS SCHOOL

TEMA 1: Introducción a las Tecnologías del Dato

MÓDULO /CURSO

Fundamentos de Data Science y Big Data

PROFESOR:

Mario De Felipe

ÍNDICE TEMÁTICO

CONTENIDOS

Objetivos:

- Entender el contexto de la transformación digital
- Conocer las tecnologías que van a intervenir en la ciencia de los datos: Datos, Cloud y Machine Learning.
- Conocer el uso actual de las tecnologías del dato en las organizaciones: los sistemas de analítica de negocio

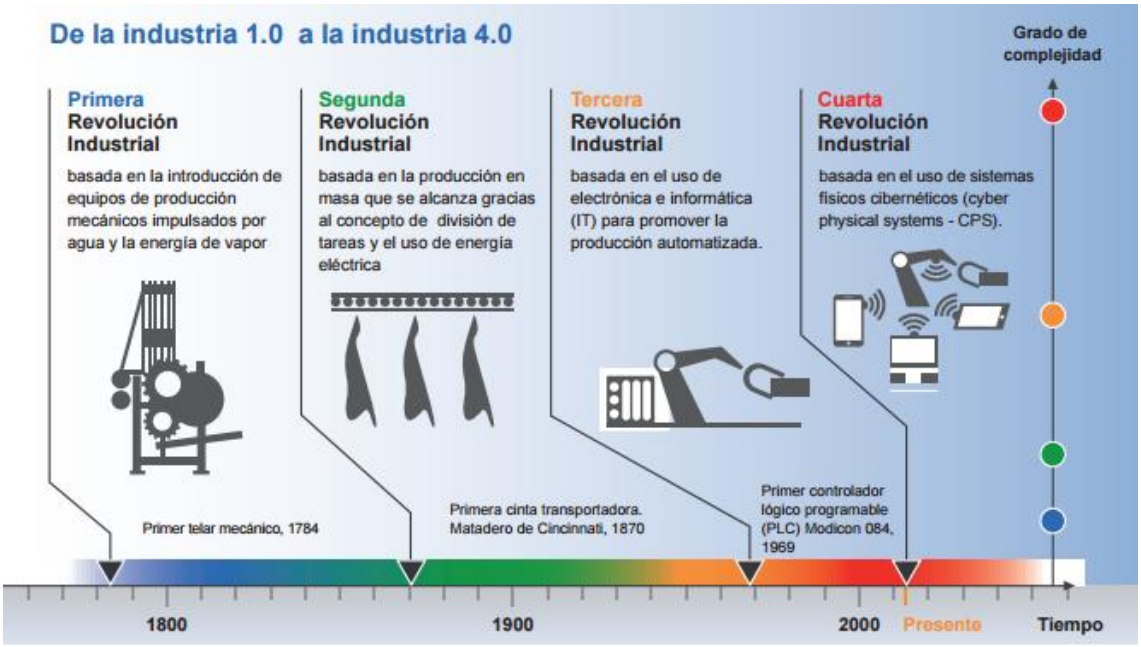
Contenido:

Vamos a poner en contexto la ciencia de los datos en la situación actual de transformación digital que estamos viviendo en todas las organizaciones, veremos qué realmente es importante y cuáles son las tecnologías del dato con las que vamos a trabajar y dónde aplicarlas. El objetivo es que tengáis claro cuál es la situación real de las organizaciones con respecto al uso de tecnologías del dato y conocer las que están a vuestra disposición para implantar.

TEMA 1: INTRODUCCIÓN A LAS TECNOLOGÍAS DEL DATO

LA TRANSFORMACIÓN DIGITAL

A nadie se le escapa que estamos viviendo hoy día una transformación digital que viene a profundizar en la automatización de tareas en los diferentes procesos de negocio de una compañía, utilizando como herramienta las tecnologías de la información. En pocas palabras, se trata de extender el proceso de mecanización industrial al resto de procesos organizativos, buscando no sólo la reducción de costes sino la mejora organizativa en general.



https://grupofranja2.com/index.php/oftalmica/item/1763-de-la-industria-1-0-a-la-4-0

Sin duda, hay muchas otras implicaciones de la transformación digital que tienen que ver con la sociedad o la manera en que nos relacionamos con nuestros clientes, proveedores, empleados, incluso con la administración pública. Os recomiendo un excelente artículo de Manuel Ortigosa que lo explica de manera muy sencilla: http://www.blog.andaluciaesdigital.es/5-claves-para-impulsar-tu-empresa-hacia-la-transformacion-digital/ .

¿Qué ha sucedido para que se produzca este cambio? Hay tres factores tecnológicos que han aparecido en el mundo empresarial y que influyen directamente en la transformación de la toma de decisiones en la compañía:

- **Más datos disponibles (Big Data)**. Tanto por la cantidad de datos que producimos internamente en nuestras organizaciones (sensores, logs, bases de datos, ERP, CRM, etc.) como los que existen en Internet (redes sociales, IoT, proveedores de datos, etc.), tenemos a nuestro alcance la posibilidad de obtener más información y conocimiento que en ningún otro momento de nuestra historia.
- **Más capacidad de procesamiento y de almacenamiento (Cloud e IaaS)**. La irrupción de la nube nos permite disponer de más recursos de infraestructura IT de forma más flexible y reduciendo la barrera que suponía el coste inicial de una inversión en este tipo de tecnología.
- **Más algoritmos sofisticados (Machine Learning e Inteligencia Artificial)**. Como consecuencia de la mayor cantidad de datos disponibles, se empieza a popularizar el uso de algoritmos de aprendizaje automático como herramientas para realizar minería de datos y análisis predictivo aplicadas al mundo empresarial. Es ahora cuando empezamos a hablar de la Ciencia de los Datos (Data Science), cuando se está convirtiendo en una disciplina científica aplicar matemáticas y estadísticas al tratamiento de datos digitales.

Aplicadas estas tecnologías al proceso de toma de decisiones de la compañía, lo que nos van a permitir es lo siguiente:

- **Que las decisiones se realicen de forma objetiva**. La decisión debe ser un proceso lógico que extrae conclusiones a partir del análisis de unos datos, y estos datos deben ser empíricos y no subjetivos; es decir, la decisión debe estar basada en hechos y no en experiencia.
- **Maximizar el éxito de la decisión**. Como consecuencia de lo anterior, tomar la decisión basándose en datos relevantes y válidos minimiza el riesgo de error.
- **Sistematizar o mecanizar el proceso de decisión**. Aplicar un método a la toma de decisiones mediante sistemas que nos permitan la captura de datos, su gestión, procesamiento y pongan a disposición de los decisores de toda la información relevante que precisan en cada momento.
- **Automatizar la decisión en sí**. Este es el santo grial de toda organización: que un sistema tome la decisión por nosotros. Pues bien, aunque hay decisiones a nivel operativo que ya se pueden automatizar, lo que sí estamos en condiciones de hacer es sistemas que nos recomienden la decisión a tomar.

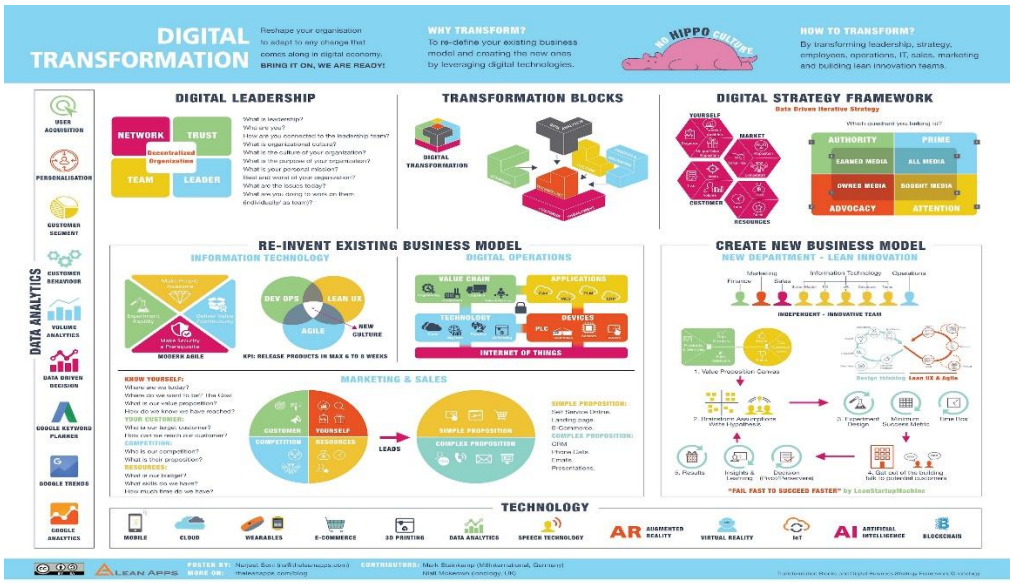
A su vez, estas facultades influyen decisivamente (y nunca mejor dicho) en la mejora y en el crecimiento de la organización, ya que van a tener como consecuencia:

1. **Lograr el cumplimiento de la estrategia empresarial**. Cualquier persona de la compañía a cualquier nivel (desde ejecutivo a operativo) disponga en cada momento de la máxima información fiable, objetiva, verificable y de fácil interpretación para tomar cualquier

decisión, constituye la mejor herramienta para que cada individuo pueda desempeñar su trabajo de la manera más óptima para la compañía.

2. **Control y optimización de los procesos organizativos**. Medir el rendimiento de personas, recursos y procesos dentro la organización nos permite detectar ineficiencias y fortalezas y, en consecuencia, iniciar tareas correctoras para mejorar lo malo y tareas promotoras que potencien lo bueno. Básicamente lo que se persigue es velar por la rentabilidad a nivel operativo, reduciendo costes y/o invirtiendo en aquello que produce más beneficio.
3. **Monetización del dato**. Un nuevo mundo se ha abierto ante nosotros a raíz de disponer de tecnologías que nos permiten analizar datos masivamente y es la posibilidad de utilizar la información que extraemos de ellos no sólo para la toma de decisiones sino para extraer valor de ellos. ¿Y qué valor podemos obtener? Pues idealmente existe la posibilidad de obtener un retorno económico directo (vender el dato, ya hay compañías que se dedican a ellos) pero lo más factible es que nos permita descubrir cómo vender más y mejor, nuevos clientes, diseñar nuevos productos, nuevas estrategias,... en definitiva, construir nuevos procesos empresariales basados en el análisis de datos.

Por tanto, vemos que estos tres factores (Datos, Nube y Algoritmos) son los que están permitiendo hacer que el uso que las organizaciones hacen de los datos sea algo diferencial, una ventaja competitiva. A estas compañías que sacan provecho y aplican las tecnologías del dato son las que se denominan *Data-Driven Business* o Negocios Dirigidos por los Datos:

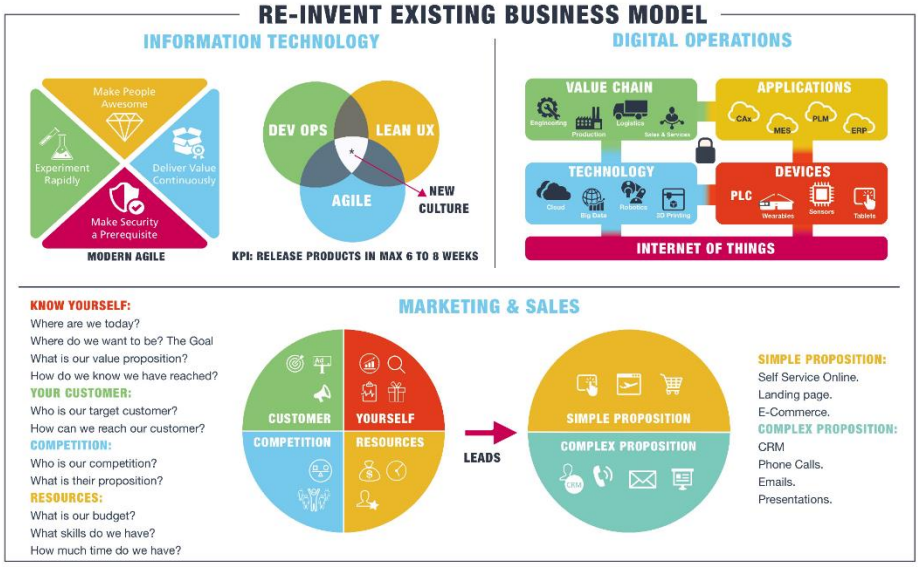


Fuente: A graphic guide to understand Digital Transformation, Narjeet.

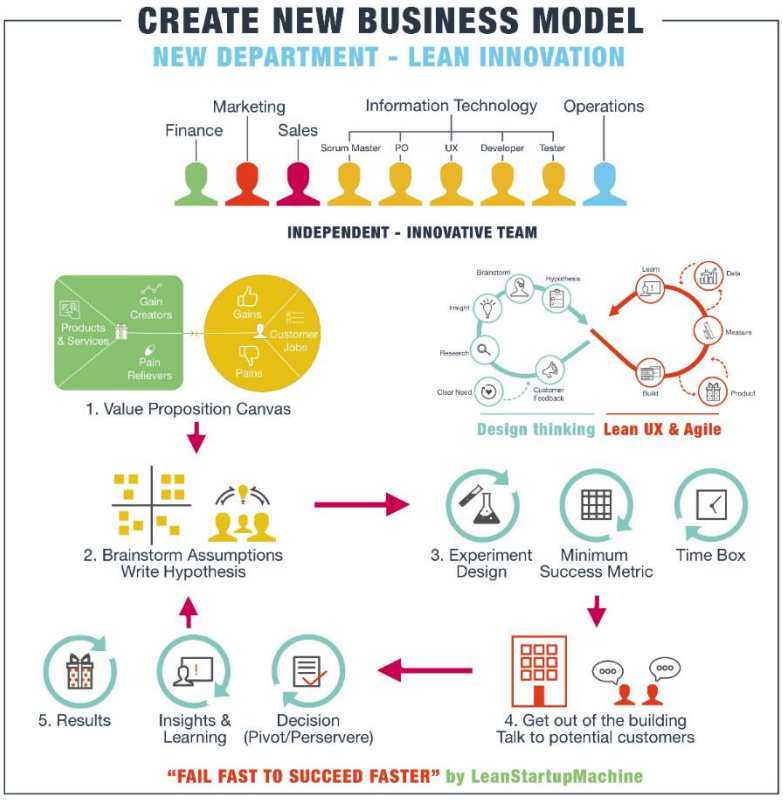
Fijaos que en el centro de todo está el cambio de paradigma en la organización de las compañías:

- **Reinventar el modelo de negocio de existente**. Se trata de digitalizar todos y cada uno de los proceso de negocio para lograr su automatización y poder aprovechar por un lado, la reducción de costes y mejora productividad que ello supone, y por otro,

responder de manera ágil y efectiva a las necesidades cambiantes del mercado y de nuestros clientes. A mi modo de ver, es un estadio intermedio entre el modelo de negocio actual (heredado de la tercera revolución industrial) y el próximo modelo 4.0, íntegramente digital.



- **Crear un nuevo modelo de negocio**. Cada vez más, se impone la tendencia que las empresas deben "copiar" el modelo de emprendimiento o *startup*. Dados los cambios vertiginosos y constantes que se producen en el mercado, es importante contar con un departamento que se ocupe de estar en constante generación de nuevas ideas, de nuevos productos y servicios, de nuevos mercados y clientes. No se trata sólo de un I+D+i, sino llevarlo un paso más allá, explorando la viabilidad comercial y empresarial. Para lograr esto se deben utilizar metodologías ágiles, que permitan prototipar rápido, lanzar el producto al mercado, medir la respuesta del cliente y descartar o validar.



¿Y cómo se logra este cambio de modelo de negocio? En la parte que nos atañe, con dos factores clave:

- **Infraestructura Tecnológica**. Como paso previo al análisis de datos, es necesario aprovisionarse de tecnología que permita la obtención y generación de datos: dispositivos móviles, sensores, IoT, e-Commerce, wearables, etc. y como es obvio, de infraestructura propia para poder realizar el tratamiento, almacenamiento, procesado y comunicación de los datos.
- **Análisis de Datos**. Implantar en la compañía una cultura de toma de decisiones basadas en datos que permita sacar valor de ellos.

Esta es la visión que me gustaría trasmitiros: todo lo que vamos a estudiar en este máster son un conjunto de estrategias, técnicas y herramientas que, mediante el análisis de datos permite a las organizaciones:

- **Reinventar sus modelos de negocio**, a través de la toma de decisiones basada en datos objetivos y no sujeto a la experiencia del decisor, e incluso automatización dichas decisiones, pudiendo optimizar así sus procesos de negocio, reduciendo costes y así lograr ser más competitivos.
- **Crear nuevos modelos de negocio**, desde detectar nuevas oportunidades de negocio, analizando clientes y productos, generando ventas a través del *marketing automation* hasta la propia venta de datos o soluciones analíticas.

Daos cuenta de que **el objetivo último es crear valor económico[1] al negocio**, ya sea por la vía de la reducción de costes como por la de generación directa de ingresos. He aquí la verdadera razón de por qué las organizaciones están corriendo para adoptarla. Debéis tenerlo siempre presente en vuestro día a día: IT ya no es una "actividad soporte" sino una "actividad primaria" y por tanto, todo lo que hagáis debe estar destinado a producir beneficio a la compañía.

Con que tengáis esto presente en vuestro día a día, yo ya me doy por satisfecho. En los siguientes puntos de este tema, vamos a terminar de introducir las tecnologías en las cuales nos apoyaremos para la Ciencia de los Datos.

[1] En el caso de las Administraciones Públicas y de las Organizaciones No Gubernamentales, además de valor económico, podemos hablar de valor social, ya que el fin último debería ser crear o mejorar el bienestar de los ciudadanos.

LAS TECNOLOGÍAS DEL DATO CLÁSICAS: LOS SISTEMAS DE INTELIGENCIA DE NEGOCIO

¿Qué son los Sistemas de Inteligencia de Negocio?

Vamos a comenzar presentando el principal uso de las tecnologías de análisis de datos en las organizaciones, que son los Sistemas de Inteligencia de Negocio o Business Intelligence.

Podemos definir[2] la Inteligencia de Negocio o *Business Intelligence* (BI) como el conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa que tienen como característica común:

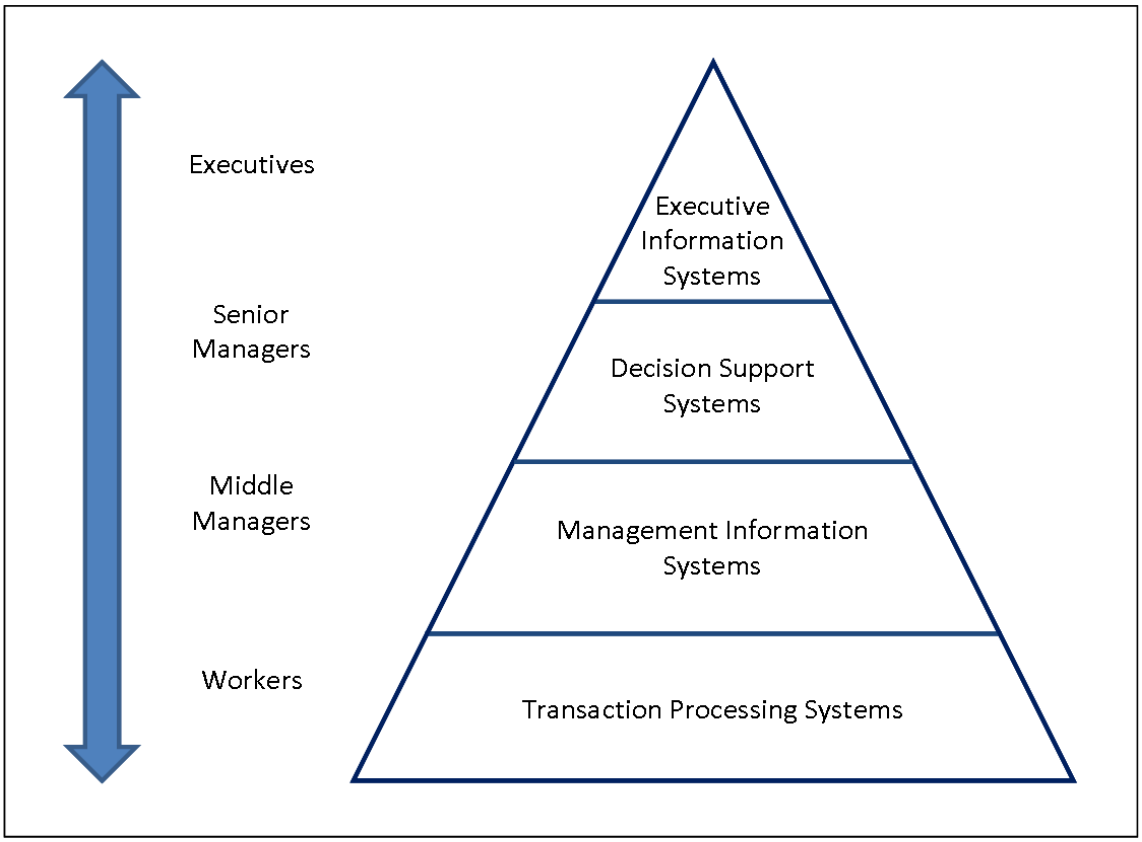
- **Proporcionar acceso** a la **información**
- **Apoyar** a la toma de **decisiones**
- **Orientación** al **usuario** final

Daos cuenta de que de esta definición se deduce que el BI no sólo tiene un impacto tecnológico sino que permite:

- Que los decisores tengan **acceso rápido y sencillo** a la información.
- **Crear**, **manejar** y **monitorizar** las **métricas** fundamentales de la organización para ayudar a tomar la mejor decisión.
- Mejorar la competitividad de la organización.

Dentro de los Sistemas de Información Empresariales, los Sistemas de Inteligencia de Negocio vienen a cubrir los procesos de toma de decisiones y por tanto, su función principal es precisamente esta: el soporte o apoyo a la toma de decisiones:

[2] **Fuente Wikipedia:** https://es.wikipedia.org/wiki/Inteligencia_empresarial



http://www.chris-kimble.com/Courses/World_Med_MBA/Three-Level-Pyramid-model.png

Aunque seguramente muchos de vosotros ya tendréis conocimiento de ellos, permitidme que en este punto hagamos una breve descripción de cada uno de ellos para que todos tengamos claro el contexto:

- **Sistemas Transaccionales[3] o Sistemas de Información Operativa** (*Transactional Process Systems o TPS*). Son los sistemas que se utilizan al nivel más bajo/operativo de la compañía. Normalmente se trata de son las aplicaciones que usan los operarios de planta o empleados sin responsabilidad de mando, para realizar sus tareas del día a día: controlar el stock, informar de las órdenes de producción, reservar una mercancía, hacer un pedido, etc. Esos sistemas son los que generan la mayor parte de los datos de las operaciones de la compañía y por tanto, van a ser la principal fuente sobre la que se va a sustentar cualquier aplicación analítica. Hoy día hablamos del ERP (Planificación de Recursos Empresariales o *Enterprise Resource Planning* de sus siglas en inglés), para referirnos al sistema que soporta, registra, valida, planifica, informa y genera cualquier acción en la compañía.
- **Sistemas de Información de Gestión** (*Management Information System o MIS*). En un nivel superior, por encima del Sistema Transaccional, tenemos las aplicaciones que dan

[3] El nombre de "transaccionales" viene de los antiguos sistemas ("mainframes") utilizados principalmente en banca donde cada operación (transacción) debía completarse o era descartada.

cobertura a los procesos de negocio que están por encima de los procesos más operativos de la compañía y que suelen ser utilizadas por personal administrativo[4] y mandos intermedios. Como ejemplos son aplicaciones como por ejemplo Sistemas de Presupuestación, Control de Tesorería, Gestión de Recursos Humanos o Control de Inventario. Estos sistemas se alimentan de la información de los transaccionales y permiten a los usuarios generar nuevos datos para poder realizar su trabajo. Además, estos sistemas también permiten generar Informes para poder analizar los datos. Actualmente los citados ERP, cubren la mayor parte de esta funcionalidad. La tendencia es que los fabricantes de ERP intentar expandirse desde su nicho transaccional hasta el nivel más alto de los Sistemas de Información.

- **Sistemas de Soporte a la Decisión** (*Decision Support Systems o DSS*). Esos sistemas permiten procesar los datos disponibles en los TPS y MIS y generar la información que necesita cualquier persona de la compañía que tenga que tomar cualquier tipo de decisión en un momento dado. Suele verse como el sistema de informes de la compañía (*Reporting*). Su objetivo fundamental es producir informes para que los usuarios puedan extraer conocimiento de ellos y tomar decisiones. También podemos englobar en esta categoría, aquellas aplicaciones que permiten realizar simulaciones, análisis estadístico o predictivo.
- **Sistemas de Información Ejecutiva** (*Executive Information Systems o EIS*). Son sistemas que facilitan la toma de decisiones al más alto nivel de la compañía. Son la herramienta de trabajo del día Cualquier gerente o propietario, independientemente del tamaño de su empresa. Consiste en el conjunto de indicadores que le permiten conocer el estado (típicamente económico y financiero) de su compañía.

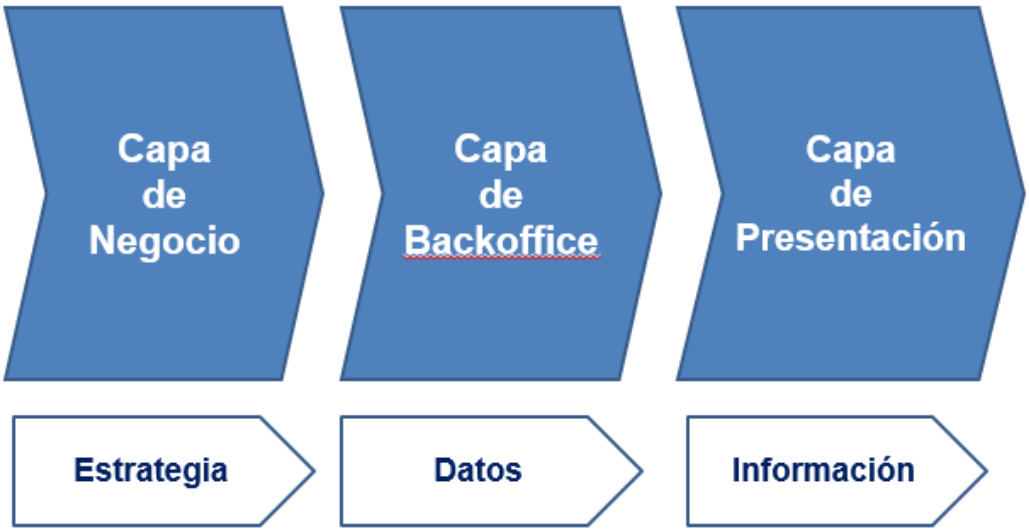
En definitiva, los dos primeros trabajan en el nivel de **GESTIÓN DEL DATO** y los dos últimos en el de **EXPLOTACIÓN O CONSUMO DEL DATO**. Como veremos un poco más adelante, los sistemas transaccionales y de gestión son las fuentes u orígenes del dato y los sistemas de soporte a la decisión son los que permiten analizar los datos residentes en ellos, procesarlos y poner el resultado a disposición de las personas. Dicho con otras palabras, los sistemas de soporte a la decisión transforman datos en información, para que las personas (decisores) pueden extraer conocimiento de dicha información, y por tanto sacar valor de los datos.

[4] Entendiendo administrativo como todo aquel trabajador no dedicado directamente a la producción.

Arquitectura

Como ya podréis intuir, los Sistemas de Inteligencia de Negocio van a constituir el elemento base sobre el cual girará todos los procesos y técnicas de la Ciencia de los Datos. La mayor parte de las organizaciones (muchas pymes incluidas) cuentan con este tipo de sistemas en mayor o menor grado de implantación, y constituyen el principal punto de apoyo (o los cimientos si queréis verlo así) sobre los cuales empezar a desplegar el proceso de la ciencia de datos e las organizaciones. En el siguiente tema, entraremos un poco en la cuestión terminológica y entenderéis todavía mejor por qué.

Los Sistemas de Inteligencia de negocio están formados por 3 bloques o capas[5]:



Capa de Negocio

Si en lugar de un sistema de soporte a la decisión, estuviéramos hablando de arquitectura de un edificio, os diría que la Capa de Negocio es a la Inteligencia de Negocio, lo que los planos a la arquitectura.

La capa de negocio **consiste en definir exactamente la información que queremos que el sistema nos genere**; es decir, se trata de determinar que la información que vamos a obtener del sistema **es realmente la que necesitamos** para posteriormente analizarla y que nos ayude a tomar las decisiones que requiere nuestro trabajo.

Fijaos que esto tiene varias implicaciones:

1. La información no es la misma para cada persona, dependerá de su función en la empresa y de los procesos o tareas que lleve a cabo. Distintas funciones o tareas implican distintas decisiones; distintas decisiones implican distintas informaciones a analizar.
2. La información a generar está determinada por el modelo de negocio de la compañía. No le vale la misma información para la Directora de Recursos Humanos de una empresa de limpieza que para el propietario de una tienda online de ropa. Distintos negocios implican distintas informaciones y por tanto, distintas decisiones.

En el TEMA 3, cuando hablemos de una de las tareas de la ciencia de los datos, volveremos a profundizar en este asunto. Por ahora, para que podáis entender rápidamente en qué consiste la capa de negocio, se trata de definir:

1. **Objetivos**. Qué decisiones queremos analizar con la información que esperamos obtener del sistema; es decir, nos planteamos primero las preguntas que queremos que responda el sistema. Un ejemplo muy claro: ¿Cuánto dinero gano con cada uno de mis clientes?. Además de esto, nos podemos plantear otras cuestiones como el beneficio o retorno de la inversión que esperamos obtener del sistema.
2. **Información**. Qué información queremos que el sistema nos genere para responder a nuestras preguntas. O visto de otro modo, se trata de definir **Métricas** que permitan cuantificar nuestros objetivos. Siguiendo el ejemplo anterior: para saber cuánto dinero ganamos con cada cliente, deberíamos poder medir esa cantidad monetaria. Para ello debemos definir una métrica (la rentabilidad) y cómo calcularla (importe neto de ventas menos el coste de las ventas). Además de esta definición, debemos establecer un valor que nos sirva de referencia o **Meta**, que nos dé una orientación de forma intuitiva y rápida de la bondad o del cumplimiento del objetivo. Como veis, se trata de establecer una o varias métricas que cuantifiquen nuestro objetivo, una forma de cálculo y una meta para cada una de ellas. Así mismo, también debemos definir las variables o **Dimensiones** por las que queremos analizar la métrica. Matemáticamente, si la métrica es una función, la dimensión es la variable. Siguiendo con el ejemplo, la **dimensión** en este caso sería el Cliente.
3. **Acciones.** Son las iniciativas que se van a tomar para optimizar el resultado de la decisión. Por ejemplo; si resulta que para un determinado cliente, el sistema nos informa de que tiene una rentabilidad muy baja, podemos determinar las acciones que deberemos tomar en tal caso: informar al comercial de la cuenta, reducir costes logísticos, minimizar los descuentos, aumentar el margen vendiendo más volumen, etc. Daos cuenta de que puede tratarse de acciones "informatizadas" que desencadene el propio sistema de IN o bien, acciones que deben ser acometidas por las propias personas. Lo importante en este caso es que se considere como un todo, como que los procedimientos, el modelo de negocio y la estrategia son parte integrante del Sistema de Inteligencia de Negocio de la compañía.

Capa de Backoffice

Es el núcleo del Sistema de Inteligencia de Negocio, se encarga básicamente de cumplir las siguientes funciones:

- **Identificación de los Datos**. A partir de las métricas y su forma de cálculo hemos de identificar qué datos necesitamos y donde encontrarlos. Nótese que los datos pueden estar dispersos en una o varias fuentes de datos, en distintas localizaciones y en diferentes formatos. Siguiendo con el ejemplo de la rentabilidad de los clientes, para poder calcularla, necesitamos obtener primero el importe neto de las ventas y luego el coste de la ventas. Típicamente el importe neto de ventas lo extraeremos de las facturas de venta que estarán en nuestro ERP y para calcular el coste, normalmente no será tan sencillo, ya que algunos parámetros los obtendremos directamente del ERP (coste de los materiales, gasto de personal, etc.) y otros tendremos que basarnos en otras fuentes de datos (reparto de costes logísticos, comisiones comerciales, amortizaciones, etc.

Métrica	Rentabilidad de Cliente	
Forma de Cálculo	Importe Ventas Cliente – Coste Ventas Cliente	
Datos	Definición	Fuente
	Importe Neto de Línea	Facturas Venta (ERP)
	Coste Artículo	Facturas Venta (ERP)
	Rappel	Contabilidad (ERP)
	Portes	Hoja Excel

- **Extracción, Transformación y Carga de los datos** (*Extract, Transformation and Load* o ETL). Una vez identificados los datos que necesitamos para construir nuestras métricas, tenemos que programar el sistema para que vaya a las fuentes, extraiga la información, realice las validaciones que consideremos oportunas, transforme aquellos datos que lo requieran y los cargue en un repositorio adecuado (normalmente, una base de datos). Además este proceso nos permite homogeneizar y consolidar los datos. En nuestro ejemplo, haríamos un proceso para cargar y transformar del ERP el maestro de clientes, las facturas de venta, los datos contables y la hoja Excel con los portes.

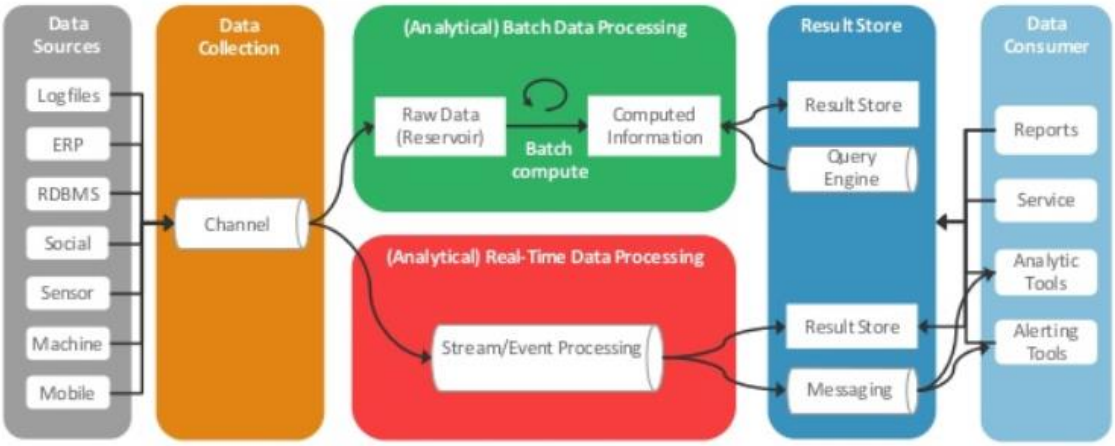
- **Almacenamiento de los datos**. Se construye un repositorio independiente donde se alojan los datos en un formato útil para el análisis. En nuestro ejemplo, sería una base de datos donde almacenaríamos las ventas, los datos de clientes, la contabilidad y los datos de portes.

- **Procesamiento de los datos**. Más allá de las transformaciones que se realizan en el proceso de ETL meramente con el objetivo de asegurar un conjunto de datos consistente y coherente, el motor de los sistemas de inteligencia de negocio debe realizar los cálculos necesarios que den lugar a las métricas y dimensiones definidas en la capa de negocio. Esto incluye pasa desde operaciones matemáticas simples hasta el despliegue de minería de datos, aprendizaje automático y deep learning.

- **Validación de los datos**. Este proceso nos permite garantizar la calidad, la autenticidad y la integridad de los datos.

Para ilustrar los procesos que se llevan a cabo en la capa de *backoffice* quiero poneros como ejemplo la llamada **arquitectura lambda** para procesamiento de datos en tiempo real[6]:

[6] Aunque no quiero entrar en polémicas, el tema del tiempo real es un tema bastante conflictivo para mí. La realidad es que debemos distinguir entre sistemas de "near-real time" (que son los que habitualmente implementamos), streaming de datos y de conexión directa o en vivo.

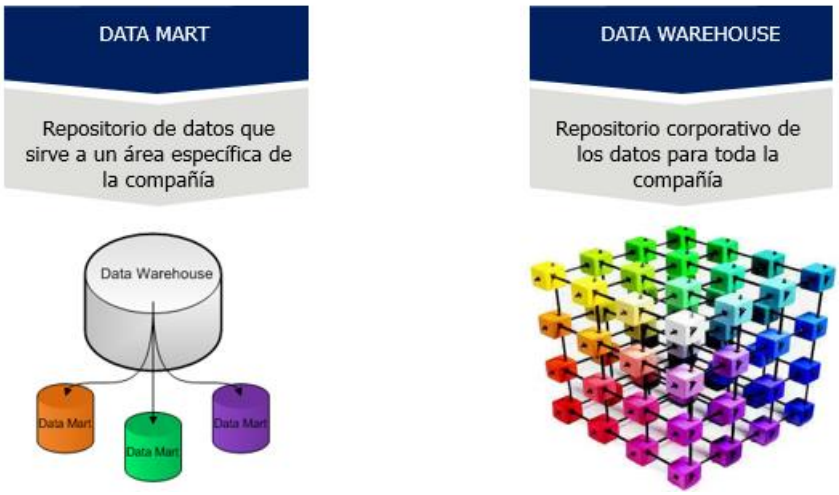


Fijaos en los diferentes elementos que los componen:

- **Fuentes de Datos (Data Sources)**. Son los orígenes o repositorios que tenemos disponibles para buscar los datos que necesitamos para poder calcular nuestras métricas. Estos repositorios o fuentes pueden ser internos o externos a la organización, y también podemos clasificarlas como estructuradas (los datos contenidos en ella están organizados según un esquema lógica y por tanto, es factible definir un proceso o algoritmo para automatizar su procesado, como ocurre en las bases de datos) o no estructuradas (los datos no siguen ningún orden lógico y requieren un tratamiento complejo para extraer información, por ejemplo un simple fichero Word). Son ejemplos de fuentes de datos: los TPS (nuestro ERP, sistemas transaccionales, etc.), los MIS (CRM, Sistema Presupuestario, Tesorería, etc.)
- **Ingestión / Captura de Datos (Data Collection).** Son los sistemas encargados de captura el dato allá de donde se encuentren, darles cierto tratamiento y almacenarlos. Tened en cuenta que ya no hablamos sólo de sistemas que puedan leer de bases de datos o ficheros, sino que podemos hablar de sistemas que lean de sensores o se conecten a maquinaria industrial. Personalmente me gusta mucho el concepto de Refinería de Datos de Kiran Donepudi, donde hablamos de sistemas que se encargan de captura el dato, limpiarlo y enriquecerlo. Podéis leer su artículo aquí: https://www.linkedin.com/pulse/data-refinery-kiran-donepudi**.**
- **Almacenamiento (Data Store).** Son los repositorios intermedios de la información entre el origen y el destinatario de la misma. Sirven para albergar y consolidar todos los datos que hemos extraído de las fuentes de datos, procesarlos, calcular nuestras métricas evaluadas por las diferentes dimensiones y almacenar dicho resultado[7]. Aunque en módulos posteriores profundizaréis (y mucho) sobre los almacenes de datos, permitidme ahora que os presente algunos conceptos relativos a los diferentes sistemas de almacenamiento:

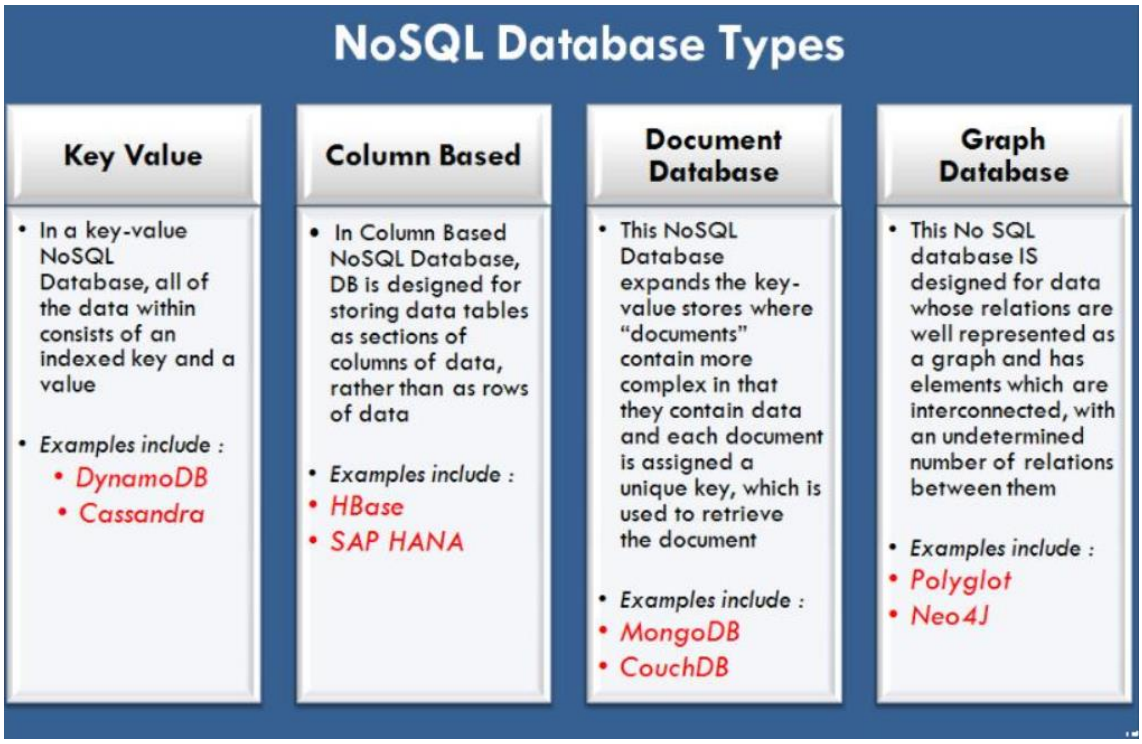
[7] Tradicionalmente, en todo sistema de BI el formato de almacenamiento de las métricas calculadas son unas estructuras multidimensionales llamadas cubos y que están basadas en la tecnología de bases de datos OLAP (On-Line Analytical Processing http://es.wikipedia.org/wiki/OLAP), que permite agilizar las consultas a grandes conjuntos de datos.

- **Datawarehouse / Datamart**. Son los sistemas de almacenamiento de datos basados en tecnologías de bases de datos relacionales. Usamos el término *datamart* cuando el propósito del repositorio es almacenar la información extraída sirve a un objetivo departamental y el *datawarehouse* cuando los datos contenidos en él cubren a diversas áreas de la compañía. Como curiosidad comentaros que hay un debate permanente a este respecto sobre que aproximación es mejor[8]. Imaginaos como es de complejo y extenso este tema, que dedicaremos cuatro módulos enteros a que aprendáis todo al respecto de estos elementos fundamentales.

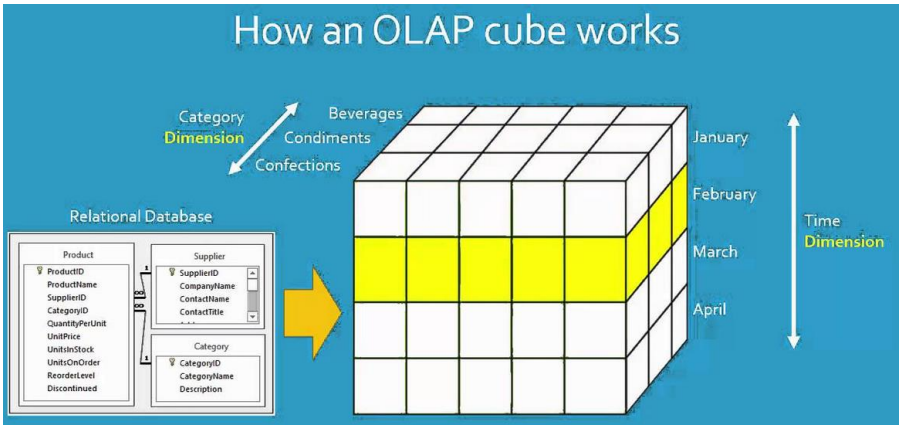


- **Data Lake**. Frente a los datawarehouse/datamart que están pensados para optimizar el almacenamiento de datos estructurados, hablamos de data lake como aquellos sistemas que almacenan los datos en su estado natural, lo cual los convierte en el repositorio adecuado para conservar datos no estructurados. Como tecnologías no encontramos los sistemas basado en fichero (tipo Hadoop) y las **Bases de Datos No SQL**. Como veréis en módulos siguientes, las tecnologías NoSQL consisten en formas alternativas de estructurar la información que las relacionales: bases de datos de grafos, documentales, clave-valor, tabulares, orientada a objetos, etc.)

[8] Datawarehouse vs Datamart, o lo que ya casi es lo mismo, la aproximación de Bill Inmon o la de Ralph Kinball: ¿construir un datawarehouse completo o construir datamarts que solucionen problemas?

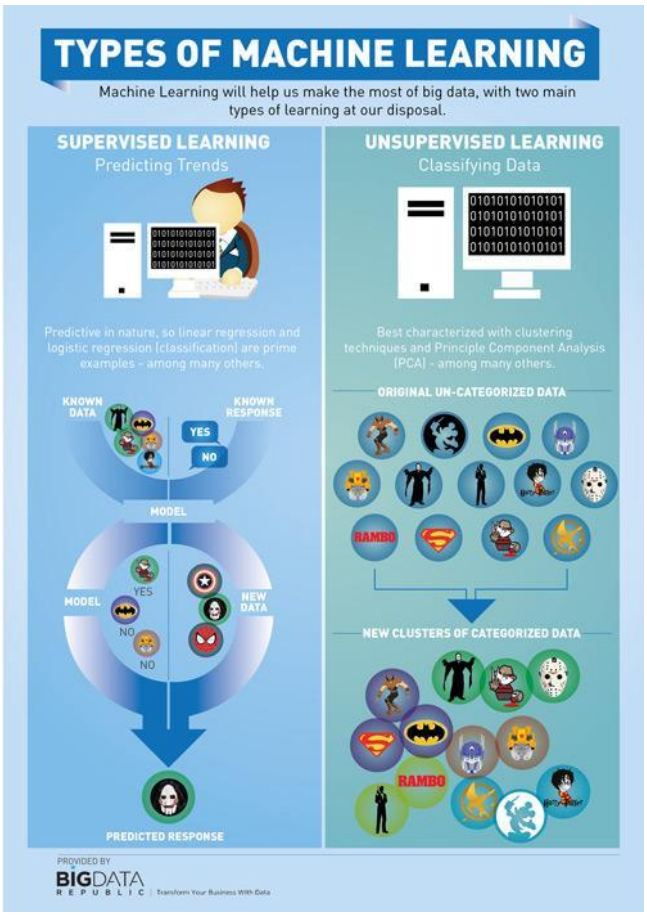


- **Procesamiento y Análisis (Data Processing)**. Es el motor de todo sistema de inteligencia de negocio: se encarga de procesar los datos almacenados (o no) y realizar los cálculos para dar lugar a las métricas y dimensiones definidas en la capa de negocio. Bajo este proceso encontramos algunos conceptos a destacar:
    - **Tecnología OLAP.** La tecnología OLAP (*OnLine Analytical Proccessing*) y sus variantes han reinado durante las últimas dos décadas como el estándar para agregar los datos en métricas y dimensiones. Seguramente la mayoría de vosotros está familiarizado con el concepto de "cubo", como el elemento de información que contiene información agregada de diferentes indicadores y variables de negocio. Aunque la mayor parte de fabricantes de software de BI siguen manejando el concepto de cubo, en la actualidad se ha popularizado la tecnología columnar o tabular.



    - **Minería de datos**. Es un conjunto de técnicas que consisten en la creación de programas que a partir de una información o conjunto de datos mediante la aplicación de un algoritmo obtenga una conclusión. En la actualidad esta

disciplina se engloba dentro de lo que se denomina Data Science y que incluye machine learning para realizar análisis descriptivo y predictivo de forma automática.



- **Deep Learning.** Aunque en realidad es un caso particular de machine learning, se está popularizando este término para referirse a la utilización de redes neuronales muy sofisticadas, sobretodo en aplicaciones de procesamiento de datos no estructurados como procesamiento natural de lenguaje y reconocimiento de imágenes. Hoy día el uso de Deep Learning ha propiciado la aparición de lo que se conoce como Inteligencia Artificial Generativa, que se ha popularizado con la publicación de ChatGPT al gran público.

Capa de Presentación

La capa de presentación o explotación del dato consiste en el proceso que permite consumir, comunicar y presentar los datos resultantes de un proceso previo de análisis, el cual, basándose en un conjunto de datos base y por medio de técnicas de análisis y procesamiento digitales, permiten la generación de información a partir de estos datos.

Como hemos visto, si la capa de backoffice es la encargada de convertir los datos en información mediante su tratamiento, la capa de presentación debe facilitar a las personas que consumen la información la creación de conocimiento. Por tanto, el factor humano es decisivo para el éxito

de un Sistema de Inteligencia de Negocio, y por tanto, la forma en que representamos la información es decisiva.

Existen tres factores claves a tener en cuenta en la presentación de la información:

- El **Formato**. Básicamente la información la podemos representar de dos formas: Numérica (Tablas) o Gráfica. Como sabéis, existen multitud de representaciones gráficas: barras, líneas, radar, dispersión, histogramas, etc. Debemos determinar qué forma de representación es la más adecuada para cada métrica, de tal forma que su interpretación resulte lo más intuitiva posible y al mismo tiempo permita extraer la mayor cantidad de conocimiento posible.

- El **Tipo de Aplicación** o **Herramienta de Visualización**. Tenemos muchas posibilidades a la hora de distribuir eso gráficos o tablas:

  - Informes: Crystal Reports, PDF, Microsoft Word, etc.
  - Hojas de cálculo: Microsoft Excel, LibreOffice Calc, etc.
  - Presentaciones: Microsoft Powepoint, Prezi, etc.
  - Páginas Web: HTML5, Flash, Siverlight, etc.
  - Cuadros de Mando
  - Correo Electrónico
  - Etc.

  Como estamos destacando, lo importante es elegir la forma de comunicación que mejor se adapte al usuario, la que le sea más cómoda, más sencilla de interpretar o algo tan sencillo como que sea la que más le guste. Es el sistema el que debe adaptarse al usuario y no al revés.

- El **Dispositivo**. Hace unos años ni siquiera se tenía en cuenta este factor. Ahora es vital saber si se va a acceder a la información desde un móvil, un ordenador o una tablet, para de esta forma representar la información en la pantalla del dispositivo de la forma más ordenada e intuitiva posible. A este respecto comentaros que hoy día, la inmensa mayoría de soluciones comerciales tienen mecanismos para adaptar automáticamente la información según el dispositivo que se conecte al sistema.

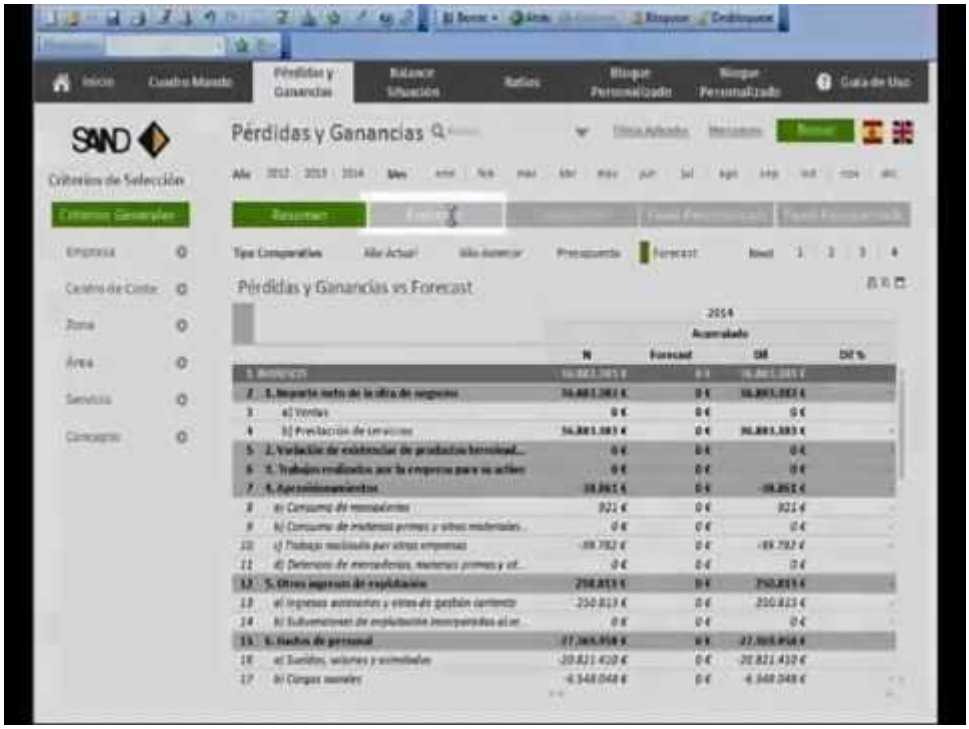
Algunos ejemplos y casos de uso

Para los que estéis menos familiarizados con los Sistemas de Inteligencia de Negocio, os voy a dejar algunos vídeos para que los veáis en acción:

CUADRO DE MANDO COMERCIAL

https://www.youtube.com/watch?v=Pl8zmlkicsU

CUADRO DE MANDO FINANCIERO



https://www.youtube.com/watch?v=BENpQZegbkU

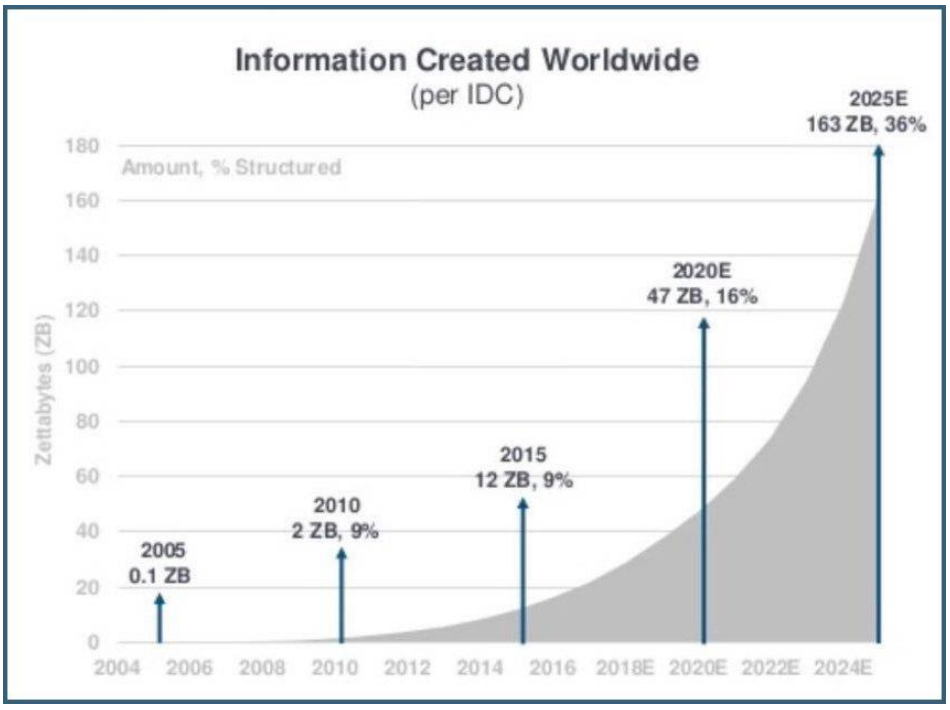
CUADRO DE MANDO INTEGRAL

https://www.youtube.com/watch?v=xSaV5u1fyzk

LAS NUEVAS TECNOLOGÍAS DEL DATO: BIG DATA

El problema Big Data

El problema de analizar más datos de los que podemos manejar con las tecnologías de las que disponemos siempre ha existido. Lo que ha cambiado es que desde hace unos años el ritmo con el que producimos los datos es exponencial, fundamentalmente debido a la acción humana (redes sociales y producción de contenidos audiovisuales) y a la de sistemas y dispositivos (móviles, servicios web, sistemas, sensores, dispositivos conectados a internet, etc.)



Volumen de datos generados en el mundo, en Zettabytes:

2005: .1 ZB

2010: 2 ZB

2015: 12 ZB

2020: 47 ZB

2024: 163 ZB

Referencias:

Megabyte = 1,024 Kilobytes

Terabyte = 1,024 Gigabytes

Petabyte = 1,024 Terabytes

Exabyte = 1,024 Petabytes

Zettabyte = 1,024 Exabytes

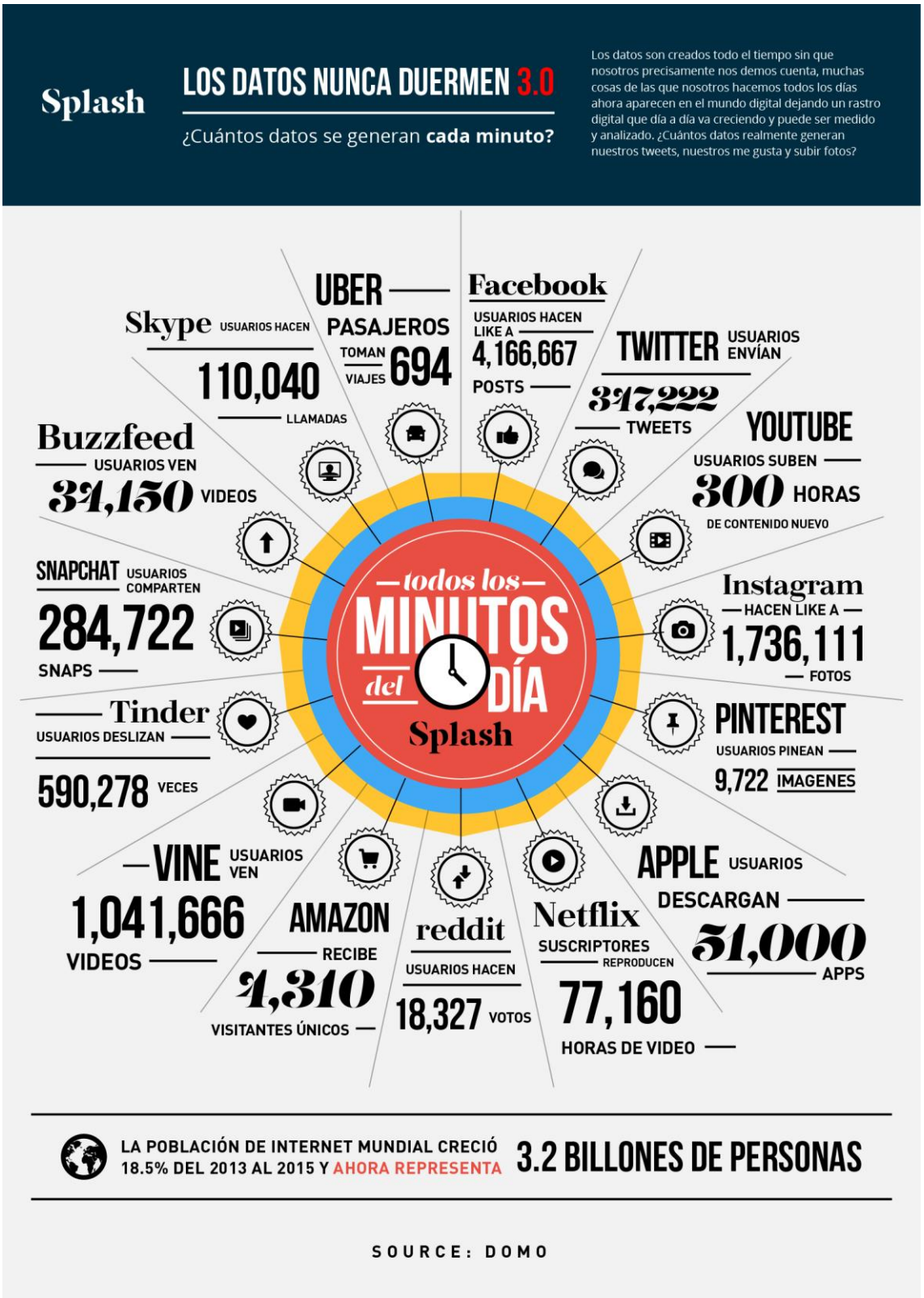
El término Big Data se suele atribuir al ensayo de Viktor Mayer-Schönberger "La revolución de datos masivos"[9] publicado en 2013. Es entonces cuando se empieza a popularizar el término entre las empresas tecnológicas para comercializar tecnologías que permitían sacar valor de esta nueva materia prima y empieza la confusión con el término debido al gran espectro de herramientas, técnicas y disciplinas que caían bajo este paraguas del Big Data.

Todos los fabricantes, consultores, empresas hablaban de Big Data haciendo referencia a cosas muy diferentes, hasta que empezó a imponerse el consenso del llamado "Paradigma de las Vs". ¿Esto qué es? Pues bien se trata de definir el concepto por las características que ha de cumplir y que, curiosamente, dichas características empiezan siempre por la letra V. Más curioso es todavía que se empezó con 3 términos (Volumen, Velocidad y Variedad) y ahora se habla de 8 tal y como podéis ver en la imagen:

[9] "Big Data. A Revolution That Will Transform How We Live, Work, andThink", 2013 https://puntocritico.com/ausajpuntocritico/documentos/Big_Data.pdf

Me vais a permitir que me quede con:

- **Volumen**: Lo más obvio, pero no tanto, procesar datos a una escala que ni la microinformática (a nivel particular) ni la informática (a nivel empresarial) se podía permitir y que ha requerido de un nuevo mercado, las empresas de servicios en la nube para poder procesarlos.
- **Velocidad**: la rapidez con la que se generan nuevos datos, constituye también un problema, ya que en un entorno empresarial la velocidad con la que se generan los datos es la misma a la que transcurren los procesos de negocio (a velocidad humana), ahora ya no es así:



- **Variedad**. Todos los sistemas de información conocidos se han basado en el procesamiento de información estructurada (tablas de datos, estructuradas en filas y columnas como un Excel) y sin embargo, más del 80% de la información existente en las

organizaciones[10] no tiene un formato fijo, es decir, es no estructurada. En Internet todavía este porcentaje es todavía mucho mayor con lo que el problema se agrava.

- **Valor**. La gran promesa del Big Data, es la obtención de valor directo de los datos (económico en el ámbito privado, bien para la sociedad o colectivo en el caso del sector público y las organizaciones sin ánimo de lucro).

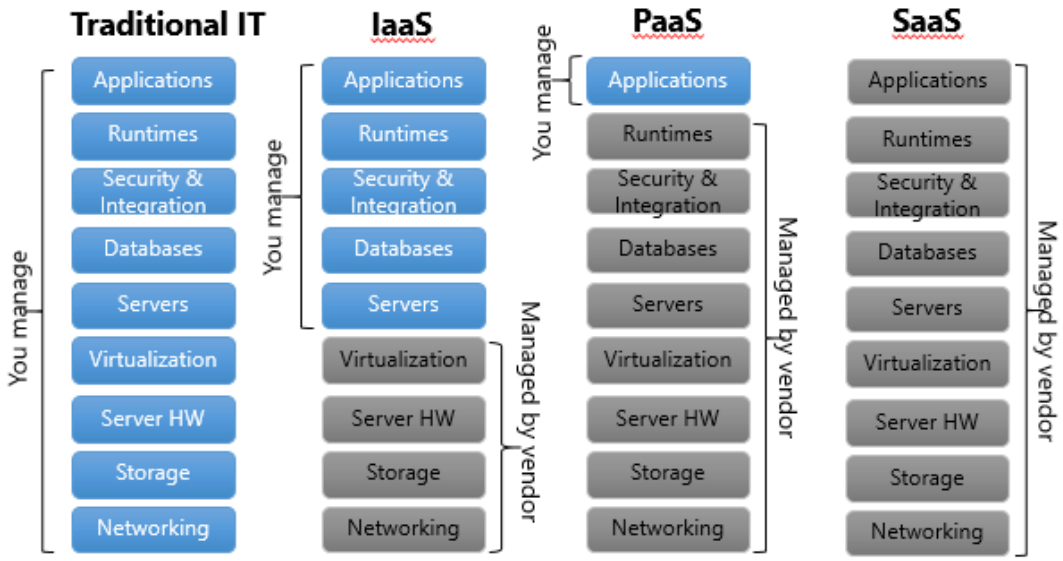
Los 3 pilares: Datos, Cloud y Machine Learning

Ahora que hemos aclarado a qué nos referimos con Big Data, sabemos que nos referimos a la suma de dos elementos:

BIG DATA = DATOS + TECNOLOGÍAS DE DATOS

Acabamos de ver la magnitud del problema de los datos, es el momento ahora de que conozcamos las tecnologías de datos que han supuesto la disrupción Big Data:

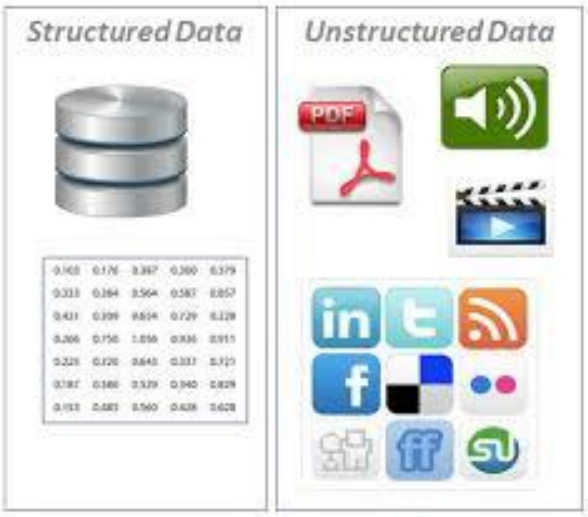
- **Los Servicios en la Nube o Cloud Computing.** Igual que en la revolución industrial, Henry Ford posibilitó convertir el automóvil en un bien de consumo, las tecnologías cloud facilitan lo que antes requería una inversión en infraestructura tecnología convertirlo en un servicio, asequible para muchas más compañías, logrando un abaratamiento de la capacidad de computación y el almacenamiento. Esto se consigue mediante la construcción de grandes Centros de Procesamiento de Datos (CPDs) o Centros de Datos (Data Centers)[11] con decenas de miles de servidores (entre 50.000 y 80.000) que se "alquilan" para su uso por parte de terceros.



[10] https://www.datamation.com/big-data/structured-vs-unstructured-data.html
[11] Para entender la magnitud de estas colosales infraestructuras os referencio a la web de Amazon: https://aws.amazon.com/es/compliance/data-center/data-centers/

- **Aplicaciones de almacenamiento y computación distribuida**. El hecho de disponer de plataformas cloud no serviría de nada, si no dispusiéramos de una tecnología que nos permitiera distribuir el procesamiento de los datos en sistemas diferentes. Aunque siempre han existido sistemas de procesamiento en paralelo, Hadoop se convirtió en el framework estándar para procesar y almacenar datos al menor coste posible. Sobre él, han ido surgiendo versiones comerciales y plataformas completamente diferentes hasta encontrarnos en la actualidad con todo un ecosistemas de soluciones de computación y almacenamiento distribuido.

- **Datos no estructurados**. En la actualidad la mayoría de analítica que se lleva a cabo en las empresas está basada en datos estructurados (ficheros con formato o bases de datos). Hoy día estamos en condiciones de poder sacar información de datos no estructurados (ficheros de texto sin formato como un documento Word o PDF, imágenes, audios y vídeos). Básicamente dos factores han sido los catalizadores: nuevos sistemas de almacenamiento (frente a las bases de datos relacionales ahora contamos con los sistemas de ficheros distribuidos como Hadoop o bases de datos NoSQL) y nuevas técnicas de procesamiento (procesamiento natural del lenguaje y reconocimiento de voz e imágenes)



- **Aprendizaje automático o Machine Learning**. Las técnicas y algoritmos de machine learning existen desde los años 70, pero la mayoría no podían salir del ámbito académico, científico o gubernamental, debido a los elevados requerimientos de computación. Sin embargo, con el abaratamiento de la infraestructura de procesamiento y la mayor cantidad de datos disponibles, ha provocado un auge inusitado (e inesperado) de esta disciplina que forma parte de la Inteligencia Artificial.

En resumen, habéis podido comprobar que bajo el paraguas del **término Big Data** se agrupan un montón de tecnologías diferentes que nos **ayudan a mejorar el procesamiento, análisis y explotación de los datos en las organizaciones**. El fenómeno Big Data está cambiando la forma, las técnicas y las herramientas que empleamos en las compañías para sacar valor de nuestros datos y, por tanto, están haciendo evolucionar los Sistemas de Inteligencia de Negocio, a lo que como veremos en el siguiente tema, podemos llamar de forma generalizada Sistemas de Analítica de Negocio. Estos sistemas van a ser la herramienta, la plataforma, sobre la cual tenemos que desplegar los procesos, la metodología que nos permite sacar valor de ellos. A esta disciplina que se apoya en las tecnologías del dato, es lo que denominamos Ciencia del Dato.

Algunos ejemplos y casos de uso

Para los que estáis menos familiarizados con el mundo de las tecnologías del dato, os cito algunas aplicaciones prácticas que las compañías están llevando a cabo utilizando Big Data:

- Mejora en la toma de decisiones
- Reducción de costes
- Ganar conocimiento del cliente
- Aumentar la disponibilidad y accesibilidad de los datos de una organización

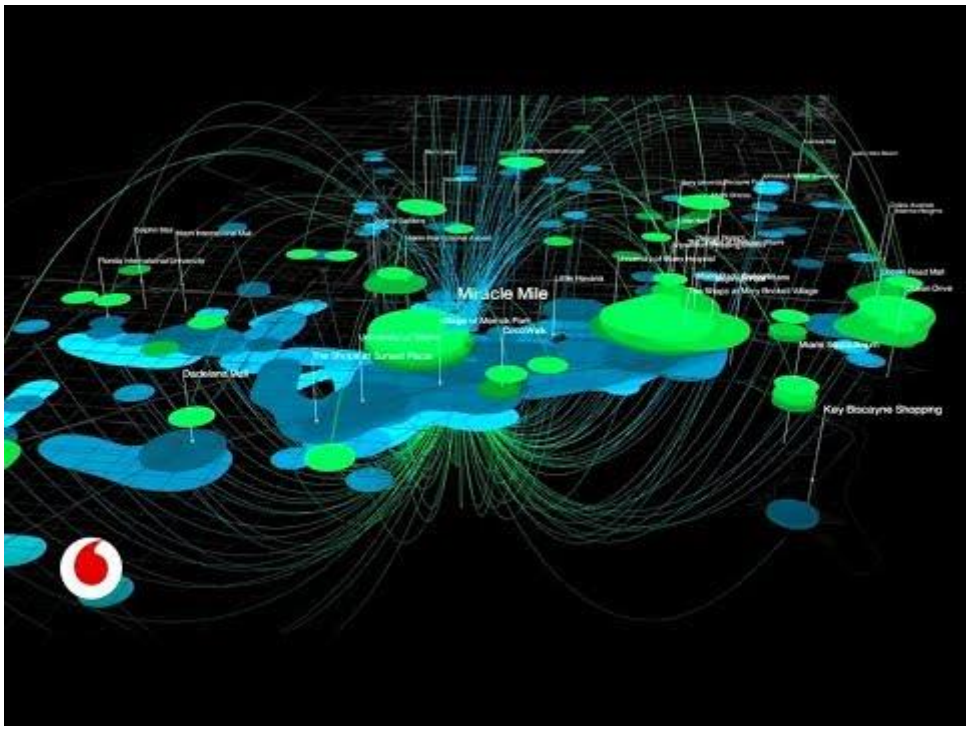
- Generar nuevas fuentes de ingresos
- Aumentar la competitividad
- Garantizar la seguridad y protección de los datos

Os pongo 2 ejemplos de casos de uso muy diferentes:

Cómo ha ayudado el Big Data al Real Madrid CF

https://www.youtube.com/watch?v=DXq30dvE0Xg

Predecir el comportamiento en redes sociales



https://www.youtube.com/watch?v=yoSqojO2-CQ

RESUMEN

Hemos conocido la importancia de las tecnologías del dato en las organizaciones, cómo uno de los elementos clave para desarrollar la transformación digital en las compañías. El uso de tecnologías del dato es vital para poder hacer más competitivas nuestras compañías: bien ayudándonos a ser más eficientes en los modelos de negocio existentes, bien facilitando la creación de nuevos modelos de negocio. El objetivo de las tecnologías del dato es claro: sacar valor del dato.

Hemos presentado cuáles son las tecnologías del dato en las organizaciones:

- Los **Sistemas de Inteligencia de Negocio** (Business Intelligence) dentro de lo que son los Sistemas de Información Empresariales, son los que se ocupan en la mayor parte de organizaciones de sacar valor del dato, mediante su análisis y explotación, a fin de mejorar la toma de decisiones, convirtiendo los datos en información y poniéndola ésta a disposición de las personas, para que puedan extraer conocimiento de ella.
- Las tecnologías asociadas al paradigma **Big Data**: cloud, tecnologías de almacenamiento y procesamiento del dato y machine learning. Ahora sabemos que el Big Data no es un producto, no es una única tecnología, sino que es la suma de muchas diferentes y que lo que nos permite es ahondar en los sistemas de análisis de datos para sacar un mayor valor de ellos. Como veremos en el tema siguiente, estas tecnologías ya no se limitan únicamente a crear información sino que nos van a permitir extraer retorno mucho más directo.

Con todo este contexto que ha hemos aprendido, vamos a pasar en el siguiente tema a presentar la Ciencia de Dato, que ya intuimos que va a ser el conjunto de procesos y técnicas que, mediante el uso de las tecnologías, nos permite obtener un beneficio a partir de los datos.

ENAE BUSINESS SCHOOL