1. BIG DATA Y BUSINESS INTELLIGENCE

1.1 BIG DATA

1.1.1 Definición y características

Si buscamos la definición de Big Data podemos encontrar distintas versiones. Aquí presentamos dos definiciones de las más relevantes:

- Big Data se puede definir como información de mucho volumen procesada a gran velocidad y muy variada que requiere sistemas de información innovadores y efectivos para poder facilitar la obtención de conocimiento y la toma de decisiones (Gartner; 2012).
- Big Data son datos que sobrepasan la capacidad de procesamiento de las bases de datos tradicionales. Los datos se mueven demasiado rápido o no cuadran en la arquitectura tradicional. Para obtener valor de estos datos, deben buscarse sistemas tradicionales para procesarlos (O'Reilly).

Si nos fijamos en estas definiciones que hemos dado, vemos que hay tres características comunes en ellas; las 3V's de Big Data:

- **Volumen**: se refiere a la cantidad de datos que tenemos. No hace mucho se utilizaba la medida en Gigabytes y ahora hablamos de Zettabytes o Yottabytes a causa del crecimiento que comentábamos anteriormente.
- **Velocidad**: se requiere que los datos sean procesados en el mínimo tiempo posible, a veces, a tiempo real. Se necesita que los datos se produzcan, procesen, analicen rápidamente para conocer el resultado en muy poco tiempo y así llevar a cabo la mejor acción para nosotros y nuestro negocio. Por ejemplo, si un negocio está en plena campaña, requerirá quizás analizar los comentarios de sus seguidores a tiempo real, para ir modificando y mejorando su actuación promocional.

  Además de ello, debemos considerar el hecho que las empresas reciben a diario una gran cantidad de datos sobre sus clientes a una velocidad

realmente alta, incluso a tiempo real, implicando ello la necesidad inmediata de análisis.

- **Variedad**: aquí nos referimos a las distintas fuentes de datos y a los diferentes tipos de archivo o formato de estos. Hay que diferenciar entre datos estructurados y datos no estructurados. Aunque trataremos más detenidamente este tema en el apartado siguiente, comentar que los datos estructurados son aquellos que pueden clasificarse y que se encuentran en las bases de datos, mientras que los no estructurados son aquellos que no pueden clasificarse y que provienen de comentarios en las redes sociales, vídeos, críticas, etc.

Con el tiempo, algunos autores han ido añadiendo más V's conforme investigaban más sobre esto. Estas otras V's son: veracidad, valor, visualización, verificación, variabilidad y viabilidad. Estas se definen de la siguiente manera:

- **Veracidad**: es muy importante que los datos sean reales y lo más exactos posibles para no dejarnos nada por el camino y poder actuar en consecuencia una vez los hayamos analizado.
- **Valor**: resulta de gran utilidad que la gestión y el análisis de los datos e informaciones nos ayuden a crear valor, que será percibido por los clientes gracias a las acciones que llevemos a cabo.
- **Visualización**: las plataformas que gestionan los datos deben tener en cuenta la forma en que se presentan los datos. Para que podamos extraer de forma sencilla la información de los datos masivos, es importante que visualmente se muestren de un modo práctico y que además vayan acompañados de un contexto para que el análisis no sea tan complejo.

  Se conoce por *visual analytics* (VA) al campo de investigación que estudia y explora soluciones de visualización. Este se encarga de que la complejidad que reside en los datos masivos y el exceso de información que podemos encontrar, se transforme en una oportunidad para nosotros y nuestro negocio.

- **Verificación**: hay que asegurar y confirmar la integridad de los datos, especialmente la de aquellos que proceden de fuentes externas o los que proceden de la nube. La verificación puede darse mediante certificados, firmas digitales, etc.
- **Variabilidad**: esta característica de los datos masivos tiene mucho que ver con otra de las V's; la velocidad. Y es que los datos van cambiando, surgen de nuevos y otros resultan entonces obsoletos.
- **Viabilidad**: cuando queremos llevar a cabo un proyecto de Big Data, debemos tener en cuenta con qué herramientas e infraestructuras contamos, las que necesitamos para llegar a nuestro objetivo y calcular los costes de estos, ya que cada plataforma y software tiene unas características distintas y por ello, su coste también es distinto. Debemos poder hacer frente a los gastos, y que estos estén justificados y sean los necesarios para lograr nuestro objetivo y sacar beneficios para nuestro negocio.

1.1.2 Tipos de datos

Como hemos comentado, hay que diferenciar entre dos grandes grupos de datos: los datos estructurados y los no estructurados. Además consideraremos también los híbridos o semiestructurados.

- **Estructurados**: Son los que tienen un formato ya predefinido, en el que los campos ocupan un sitio fijo y, por lo tanto, conocemos de forma anticipada su organización, tipo, etc. Se almacenan en tablas y la información se representa por datos elementales.

Los datos estructurados pueden tener diferentes fuentes:

- **Creados por la empresa**: registros en tablas, ficheros XML, etc.
- **Provocados**: datos creados de manera indirecta a partir de una acción previa, como pueden ser valoraciones de restaurantes o películas.

- **Dirigidos por transacciones**: datos que tienen lugar al finalizar una acción previa de manera correcta. Son este tipo de datos las facturas de compra o recibos de un cajero.
- **Compilados**: resúmenes de datos de empresa o servicios públicos de nivel grupal, como son el censo electoral, vehículos matriculados, etc.
- **Experimentales**: datos generados como parte de un análisis.

- **No estructurados:** Son aquellos que no tienen una estructura específica. Manipular estos es algo más complejo y no se pueden almacenar en una tabla como sí sucede con los estructurados. Son datos no estructurados los archivos multimedia, archivos PDF o Word, contenido de emails, comentarios en las redes sociales, interacciones con otros usuarios, etc.

  Principalmente existen dos fuentes:

  - **Capturados**: datos creados a partir del comportamiento de un usuario. Estos pueden ser extraídos a partir de aplicaciones de seguimiento de actividades (carrera, ciclismo, natación), posición GPS, etc.
  - **Generados por usuarios**: datos que especifica un usuario, como son las publicaciones en redes sociales, vídeos reproducidos en Youtube, etc.

- **Semiestructurados o híbridos:** Son aquellos datos que sí tienen alguna estructura autodefinida pero, no tienen una estructura fija. Ofrecen información poco regular ya que a veces, debido a su complejidad en el proceso de carga, se pueden perder datos. El tipo de archivo es XML o JSON.

Veamos en una tabla las diferencias entre los datos estructurados y los no estructurados:

DESCRIPCIÓN	ESTRUCTURADOS	NO ESTRUCTURADOS
Descriptivos	Edad, Sexo, Salario medio, Estudios	Actitudes
Social	Definición del usuario	Área de influencia Contenido de sus posts
Localización	Dirección	Localización a tiempo real
Interacción con la empresa	Con el próximo agente disponible	Agente basado para la personalidad del cliente
Customer journey	Transacciones Contacto con los empleados	Interacción con los diferentes puntos de contacto
Próxima acción	Resolver la opción en base a patrones	Entre las diferentes ofertas, hacer la oferta personalizada

1.1.3 Calidad de los datos

Es importante cuidar la calidad de los datos para que estos sean analizados y obtengamos un resultado consistente para poder tomar decisiones al respecto y logremos el objetivo que marcamos en un principio evitando problemas y errores.

La calidad de los datos está fundamentada en una serie de propiedades: precisión, completitud, relevancia, validez de los datos, proveniencia, autenticidad, veracidad, exactitud, reputación y credibilidad.

La Organización Internacional de Normalización (ISO – Internacional organization for Standardization) ha desarrollado una norma para orientar la gestión de la calidad de los datos. Se trata de la ISO 8000 y las relacionadas. Esta norma tiene en cuenta los siguientes aspectos:

- Las personas que deben participar en la gestión de los datos.

- Los procesos que se responsabilizan de la gestión efectiva de los datos.
- La mejora continua de los procesos dedicados a garantizar la calidad de los datos.

Esta norma también tiene en cuenta tres roles que son los que siguen:

- **Gestor de datos**: es el responsable de factores organizativos para gestionar la calidad de los datos.
- **Administrador de datos**: es el responsable de coordinar y supervisar el trabajo de los técnicos de datos, alineados con las directrices del gestor de datos.
- **Técnico de datos**: es el responsable de los cambios de datos que se llevan a cabo, de la corrección de datos y la medición de la calidad.

Una buena gestión de calidad de datos se rige por lo ahora comentado y por una buena gestión de datos, que ahora comentaremos, en la que importa no perder datos y obtener una información de calidad para lograr nuestro objetivo con nuestro negocio y organización.

1.1.4 Gestión de los datos

Aquí veremos diferentes aspectos o pasos a tener en cuenta en la gestión de datos para que podamos garantizar cierta calidad a todo el proceso y así lograr nuestro objetivo de la mejor manera posible.

1.1.4.1 Generación de datos

En la actualidad se generan datos desde diferentes fuentes o entornos. Podemos obtener datos a partir de las búsquedas en Internet, e-mails de una organización, mensajes que se dejan en foros o webs, etc. Estos datos en bruto deben extraerse de las infraestructuras que los producen y después traspasarse a los sistemas de almacenamiento correspondiente para ser analizados en su contexto y no pierdan calidad ni valor.

1.1.4.2 Adquisición de datos

Para la obtención o adquisición de datos, lo primero que hay que tener claro es nuestro objetivo, para así saber qué datos necesitamos y por lo tanto buscarlos en las fuentes de datos que nos interesan. Una vez tengamos esto claro, deberemos evaluar lo siguiente:

- **Heterogeneidad de los datos**: estos deben conservar la estructura, jerarquía y diversidad.
- **Redundancia de los datos:** el espacio que ocupan los datos en los dispositivos de almacenamiento se multiplica y con ello los costes.
- **Calidad:** hay que verificar que los datos sean exactos y reales, y que su procedencia sea válida.

Todo esto es un proceso algo complejo y en el que diferenciamos tres subprocesos:

- **Recogida:** aquí hay que tener en cuenta la procedencia de los datos y lo diversos que son. Existen diferentes métodos de recogida de datos y podemos dividirlos en dos categorías:
  - **Basados en enfoque pull**: se extrae la información de forma proactiva mediante un agente.
  - **Basados en enfoque push**: los datos se obtienen mediante una distribución selectiva en una fuente concreta.

  Tres métodos habituales son: el uso de ***sensores***, que recogen datos tales como la medición de la temperatura o las vibraciones, ***ficheros log***, que son aquellos ficheros generados de manera automática por las aplicaciones y aparatos digitales, y que graban datos de las actividades que se realizan en ellos, como pueden ser el número de clics en un enlace o el número de visitas que recibe una web, y ***web crawler***, que se refiere a aquellas técnicas empleadas para extraer datos disponibles en la red. Dichas técnicas son el rastreo de páginas web, los sistemas de segmentación de palabras y los sistemas de indexación, entre otras.

- **Transmisión**: este es un subproceso en el que aún se están investigando tecnologías y trabajando para mejorar. Los datos los transferimos a un centro de datos para después poder tratar y analizar estos. Este centro está asociado a una arquitectura de red y a un protocolo de transmisión.
- **Preprocesamiento:** en este punto, se tratan los datos en bruto para asegurar su calidad y fiabilidad. En la investigación de este preprocesamiento se llevan a cabo tres técnicas; integración, limpieza y eliminación de redundancia.

    - **Integración**: entendemos por integración la combinación de datos que proceden de distintas fuentes de datos. Esta integración tiene lugar mediante herramientas de procesamiento de flujos y búsqueda:

    - ETL (Extract, Transform, Load) integrado en un datawarehouse incluye tres procesos:

Extracción: se conectan sistemas de fuentes de datos y, según los objetivos que hemos marcado, se analizan, recogen y procesan los datos.

Transformación: los datos mediante una serie de reglas, se convierten en formatos normalizados que pueden ser tratados.

Carga: es el proceso con mayor dificultad. En este, se importan los datos ya transformados a la estructura de almacenamiento que hemos seleccionado.

Comentar que el proceso de ETL será tratado en profundidad en otro manual.

- Federación de datos: crea una base de datos virtuales que contiene información o metadatos sobre los datos y su localización. Esta base de datos puede contener datos de diferentes fuentes y se puede ir consultando.

- **Limpieza (cleaning):** si queremos garantizar la calidad de nuestro proyecto este es un proceso clave para ello. En este se identifican aquellos datos imprecisos, que deben actualizarse, que están incompletos, etc. y que

pueden provocar que el resultado de nuestro proyecto no sea el deseado. Este proceso tiene cinco procedimientos: definir y determinar los tipos de errores, buscarlos e identificarlos, corregirlos, documentar los ejemplos de error y los tipos de error, y modificar los procedimientos de introducción de datos para reducir futuros errores. Estos procedimientos están acompañados de pautas de revisión que tienen en cuenta: formato, integridad, racionalidad y límite.

- **Eliminación de redundancia:** cuando hablamos de redundancia nos referimos a la repetición de datos comunes en diferentes conjuntos, hecho que provoca gastos innecesarios, almacenamiento de más y por tanto un desaprovechamiento de este y falta de veracidad.

  Hay diferentes métodos para disminuir esta redundancia, pero es un tema en el que aún se está investigando. Aquí podemos nombrar algunas metodologías como: detección de la redundancia, filtraje de datos y comprensión de datos. Escoger una u otra va a depender del problema que debamos resolver, las características de esos mismos datos, etc.

1.1.4.3 Almacenamiento de los datos

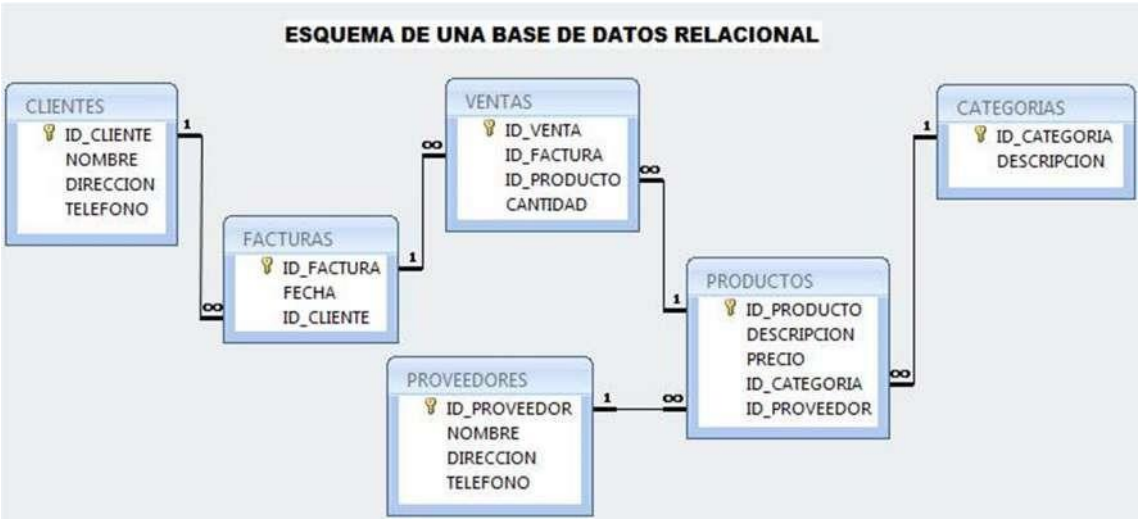
Los datos, una vez obtenidos, deben ser almacenados en una plataforma o centro de datos que contiene diferentes servidores con discos duros de gran capacidad para almacenar estos datos masivos. Más adelante veremos las distintas herramientas que se usan en Big Data, pero un ejemplo de almacenamiento de datos sería Hadoop. En el almacenamiento tiene un papel importante cómo guardamos los datos de cara al uso que podremos hacer de ellos después, su recuperación, etc. Por ello se deben explorar:

- **Lenguajes controlados**, para trabajar la precisión y el registro de todos los conjuntos de datos.
- **Usabilidad** de los almacenes de datos.
- **Visualización** de los datos.

Para que la calidad de los datos sea óptima, tienen lugar varias tareas en esta fase. Algunas son: recopilar, analizar y actualizar metadatos, comprobar la transformación de los datos y calidad, detectar incorrecciones y corregirlas encontrando su causa, analizar fuentes de datos, etc.

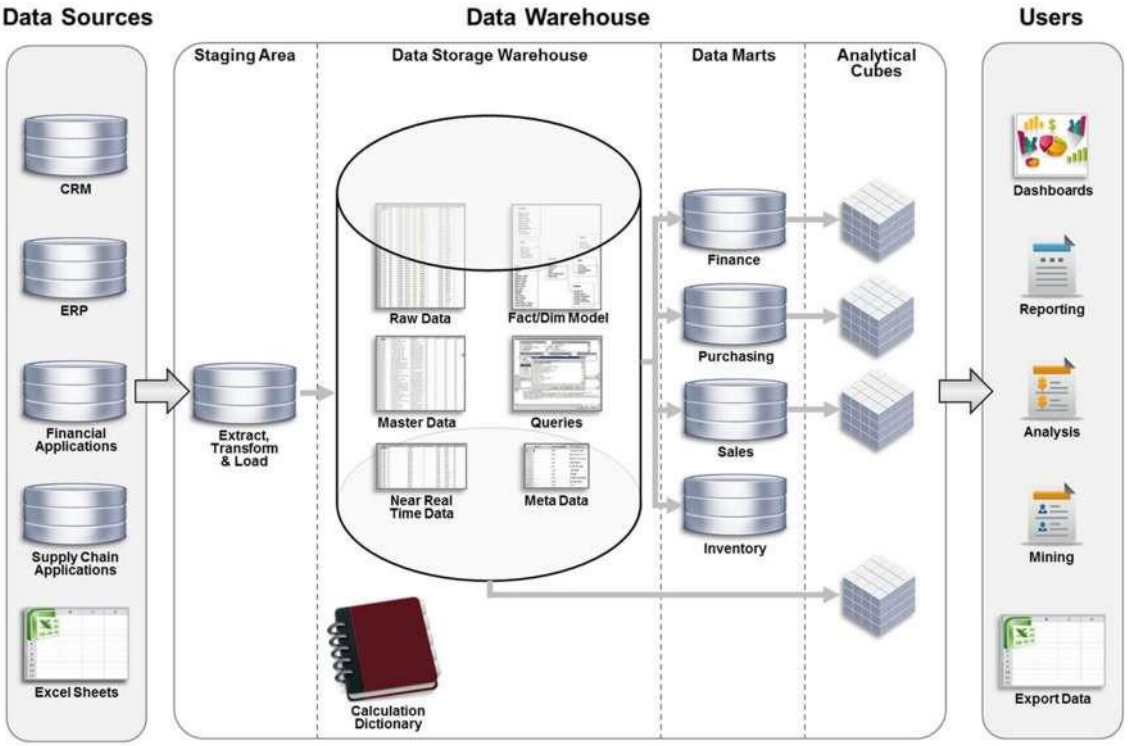
El almacenamiento se compone de infraestructura tecnológica (dispositivos y arquitectura de red) y de marco de gestión de datos para organizar la información de manera adecuada en aras de un procesamiento eficiente. La tecnología asociada a esta fase son los sistemas de almacenamiento masivo, los sistemas de almacenamiento distribuido y los mecanismos de almacenamiento (Chen y otros, 2014).

Los almacenes de datos han ido evolucionando para así hacer frente al gran volumen de datos, el reto de los datos masivos. Tradicionalmente, el software y hardware para almacenar, gestionar y analizar estos datos estaban vinculados a bases de datos relacionales, que tienen grandes limitaciones. Las bases relacionales solo trabajan con datos estructurados y se basan en el lenguaje SQL (structured query language).



**Representación gráfica de una base de datos relacional (lenguaje SQL). Fuente: http://rudeofmagic.blogspot.com.es/2014/11/4-que-es-una-base-de-datos-relacional.html*

Más adelante, se dispuso de los almacenes de datos (data warehouse) que integran datos de distinta procedencia y están orientados a temas de interés y no en funcionalidades.



**Representación gráfica de un sistema data warehouse. Fuente:*

http://www.data-warehouse.com.au/

Finalmente, para trabajar con datos masivos, aparece el lenguaje NoSQL en sustitución del SQL. En otro apartado hablaremos sobre este nuevo lenguaje.

**Representación gráfica de una base de datos NoSQL. Esta es una base de datos documental MongoDB.*

Fuente: https://es.slideshare.net/dipina/nosql-cassandra-couchdb-mongodb-y-neo4j

El almacenamiento de los datos masivos tiene lugar en:

- **Sistemas de archivo**

Los sistemas de almacenamiento tienen que almacenar, organizar, nombrar, compartir y proteger los archivos. Por ello la gestión de datos masivos debe cumplir los siguientes requisitos:

- Rendimiento de la lectura y escritura.
- Accesos a datos simultáneos.
- Creación de sistemas de archivo según demanda.
- Técnicas eficientes para sincronizar archivos.

Cuando hay que diseñar sistemas de archivo hay que tener en cuenta lo siguiente:

- **Acceso distribuido y transparencia en la localización**: los ficheros están distribuidos, pero deben disponer de directorios unificados para que el usuario acceda a ellos sin complicaciones, como lo hace cuando accede a un archivo local. Debe haber consistencia entre los nombres de los ficheros locales y de los remotos.
- **Gestión de fallos**: aunque alguna cosa falle o algún componente no funcione como es debido, los programas de las aplicaciones y los usuarios tienen que seguir operando sin problema. Esto se logra con reproducción y redundancia.
- **Heterogeneidad**: el sistema de archivos debe estar compuesto de variedad de hardware y plataformas de sistemas operativos.
- **Distribución muy definida de los datos**: es aconsejable ubicar objetos individuales cerca de los procesos que los emplearán para lograr así un mejor rendimiento.
- **Tolerancia a la partición de la red**: en algún momento puede ocurrir que para un usuario no sea posible acceder a toda la red o parte de ella. Puede ocurrir, por ejemplo, que un portátil se desconecte en medio de una operación. Por esto es importante que el sistema de archivos logre gestionar

la situación y aportar una solución mediante la aplicación de sistemas de sincronización.

- **Tecnologías de bases de datos**

En las bases de datos es donde tiene lugar el almacenamiento y la recuperación de datos e información. Como hemos comentado más arriba, para gestionar datos masivos, especialmente datos no estructurados, las bases de datos que se emplean son diferentes a las bases de datos relacionales, dado que estas últimas no pueden con el reto que plantea el Big Data en cuanto a volumen y organización. Estas nuevas bases de datos, de las que hablaremos más adelante detalladamente, tienen dos características relevantes:

- No siguen el esquema entidad-relación, esto significa que carecen de una estructura de datos ya prefijada de tablas y relaciones.
- No utilizan lenguaje SQL. Por ello, las nuevas bases de datos llevan el nombre NoSQL *(not only SQL).*

- **Modelos de programación**

Los macrodatos habitualmente se almacenan en centenares o miles de servidores comerciales, que operan con modelos de programación paralelos para procesar los datos. Los modelos paralelos tradicionales, como MPI *(message passing interface)* y OpenMP *(open multi-processing)*, pueden ser inadecuados para operar con los programas paralelos a gran escala. Por eso, han surgido otros modelos de programación paralelos para mejorar el rendimiento de NoSQL y reducirle vacío de rendimientos de las bases de datos relacionales. Estos modelos son clave para la fase posterior de análisis de datos masivos. Algunos de estos modelos de programación son: MapReduce, Dyrad, Ajo-Pairs y Pregel (Chen y otros, 2014).

1.1.4.4 Análisis de datos

Esta es la fase final y es clave para saber qué valor tienen los datos y generar el conocimiento que nos llevará a actuar del mejor modo para lograr el objetivo que marcamos con nuestro negocio.

El análisis engloba un conjunto de procedimientos y modelos estadísticos para extraer información de un amplio conjunto de datos (Kune y otros, 2016). En este sentido, hay una serie de métodos desarrollados, como son minería de datos, factores, correlaciones, regresiones test A/B, estadística, etc. (Chen y otros, 2014).

Todavía existen limitaciones en el análisis de datos por parte de las plataformas/tecnologías tradicionales, por lo que se sigue investigando sobre ello.

Hay tres ámbitos de investigación:

- Diseño de **tecnologías y software** que según las características de los datos, faciliten el análisis de estos: análisis de datos estructurados, análisis de texto, análisis de datos en web, análisis de datos multimedia, análisis de datos en redes y análisis de datos móviles.
- Diseño de métodos de **análisis** según el formato y la estructura de los datos.
- **Visualización** de la información en forma de gráficos y de la información resultante de los datos masivos para ayudar en el diseño de algoritmos y desarrollo de software. (Hu y otros, 2014).

Vamos a ver algunos métodos de análisis, la arquitectura de análisis y los criterios de selección y visualización:

- **Métodos de análisis**

Dependiendo de la tipología de los datos escogeremos un método de análisis u otro. Estos métodos los podríamos agrupar en dos grupos, los tradicionales y los específicos para datos masivos.

Análisis estadísticos

El análisis estadístico se basa en la teoría de la probabilidad, la cual tiene en consideración la aleatoriedad y la incertidumbre. La estadística descriptiva y la estadística inferencial son las que ofrecen métodos para analizar datos. La estadística descriptiva es la que resume y caracteriza el conjunto de datos que necesitamos y queremos analizar. Por su parte, la estadística inferencial nos permite extraer conclusiones de los datos sujetos a variaciones aleatorias. Las inferencias ayudan a generar modelos, predicciones, responder preguntas basadas en hipótesis, hacer estimaciones o correlaciones, etc.

Podemos ver tres tipos de análisis según el número de variables que se escogen para analizar:

- **Análisis univariante:** se usa para ver cómo se distribuyen los datos y cuál es la dispersión o variabilidad interna de estos. Para calcular la distribución se utiliza el cálculo de frecuencias, la media, la moda y la mediana. Para calcular la dispersión se usa la desviación típica o la varianza.
- **Análisis bivariante:** sirve para estudiar el efecto de una variable sobre otra. Las técnicas que se usan son la comparación de medias, análisis de correlaciones, análisis de varianza y tablas de contingencia.
- **Análisis multivariante:** se utiliza para estudiar y analizar más de dos variantes. Este es más complejo que los dos anteriores y cuenta con diferentes técnicas estadísticas y algoritmos de cálculo más amplio que las otras. Las técnicas que se emplean son el análisis de varianza, para ver el efecto de dos factores sobre una variable técnica; análisis multivariante de la varianza, análisis discriminante, análisis de regresión lineal múltiple, análisis de regresión logística, análisis de covarianza, modelo lineal general, análisis factorial y análisis de conglomerados *(o clusters).*

Minería de datos (Data mining)

Conjunto de métodos, técnicas y procesos que permiten analizar datos que proceden de diferentes fuentes. No es válida para documentos de texto, aunque hay conjuntos de datos estructurados que pueden existir en forma de documento de texto CSV o ARFF. Se usa con conjuntos de datos estructurados por valores o atributos.

Obtenemos la información analizando las estructuras de los datos, de las que emergen patrones de comportamiento y tendencias. Estos patrones están basados en la observación del pasado, y las técnicas de predicción nos aportan la información de cara a tendencias futuras.

La minería de datos empieza con una fase tratada anteriormente, la adquisición u obtención de datos, ya que hay que tener claro cuál es nuestro objetivo para saber qué datos nos interesan y necesitamos analizar. Una vez tengamos claro qué datos nos interesan, hay que evaluarlos para ver qué propiedades tienen, frecuencias, dispersión, valores atípicos y la ausencia de datos. Y si es necesario transformarlos, también se hará. Una vez se hayan evaluado los datos, se aplicará la técnica de minería de datos.

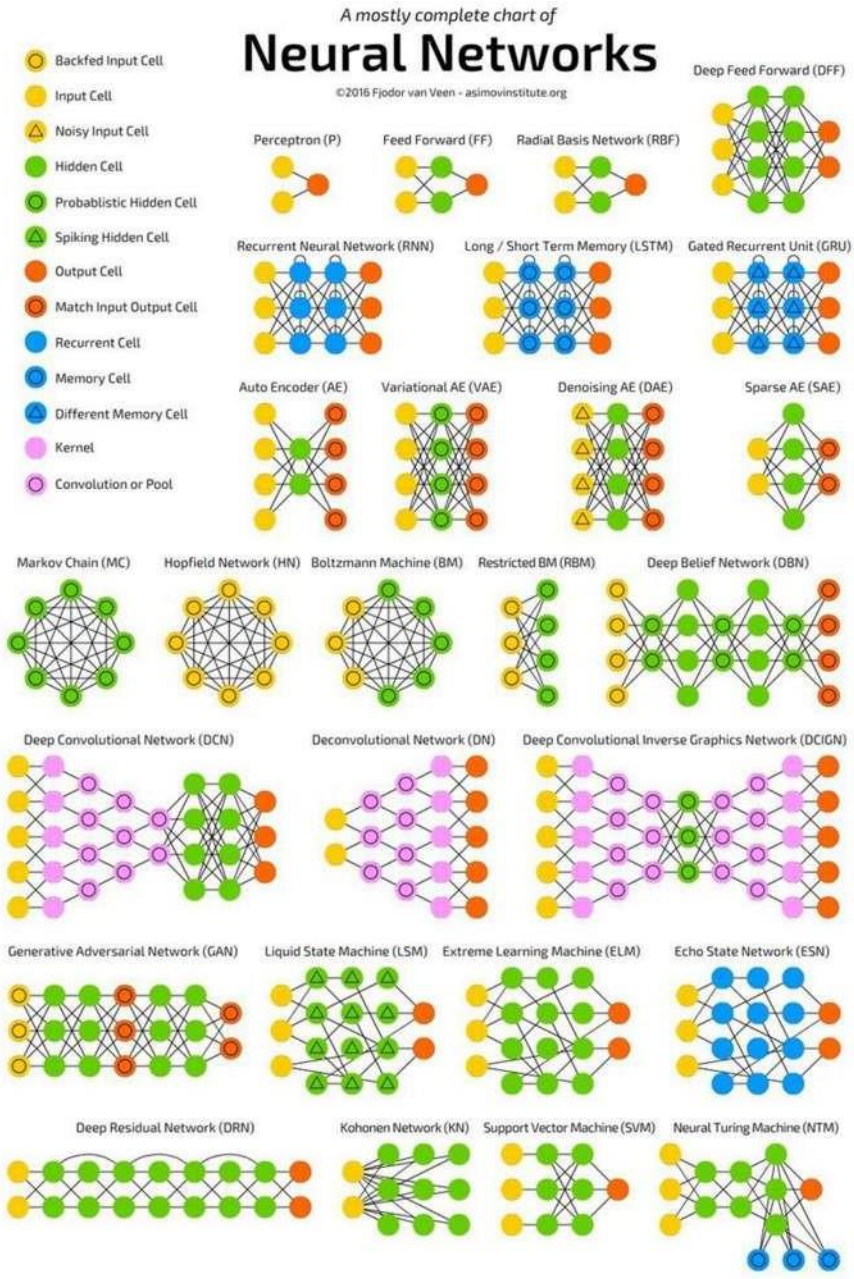
Las técnicas de minería de datos se basan en inteligencia artificial y estadística. Estas dos facilitan la creación de algoritmos que permiten modelizar los datos. Los algoritmos pueden basarse en clasificación supervisada y predictiva o clasificación no supervisada y descriptiva (o de descubrimiento de conocimiento). Sus características son (Gironés, 2013a):

- **Los algoritmos supervisados** tienen como objetivo la obtención de un modelo válido para predecir casos futuros a partir del aprendizaje de casos conocidos. A partir de un conjunto de objetos descritos por un vector de características y del que conocemos la clase a la que pertenece cada objeto, se construye un grupo de datos denominado de entrenamiento o de aprendizaje. Por lo tanto, parten de conocimiento ya existente.

- **Los algoritmos no supervisados** tienen como objetivo obtener un modelo válido para clasificar objetos sobre la base de similitud de sus características, pero sin partir de modelos predictivos. Se basan en un conjunto de objetos descritos por un conjunto de características, y a partir de una métrica que define la similitud entre objetos, se construye un modelo o regla general que clasificará todos los objetos. Por lo tanto, se descubre conocimiento.

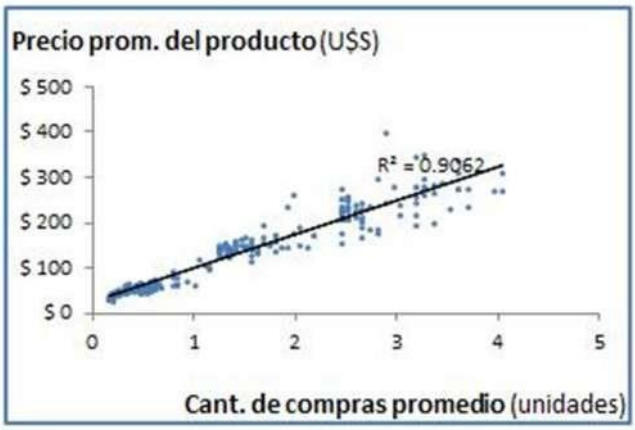
Los algoritmos más representativos son:

- **Redes neuronales**, sirven para ver conexiones en una red, son una buena aproximación a problemas en los que el conocimiento es impreciso o variante en el tiempo. Se basan en la clasificación supervisada, aunque no necesariamente debe ser un clasificador.



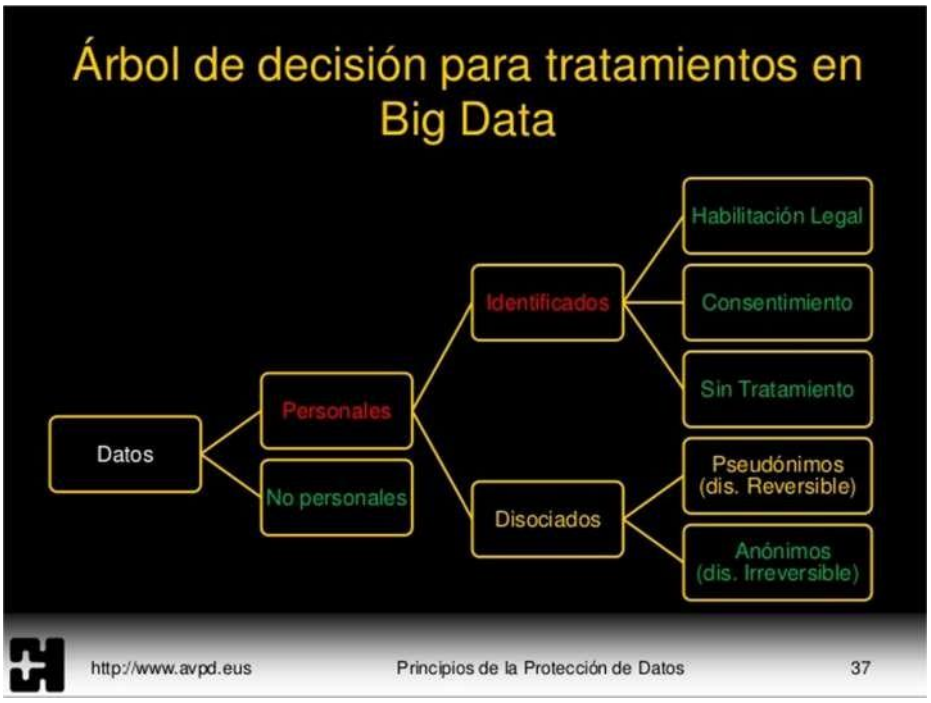
**Representación gráfica de ejemplos de redes neuronales. Fuente: https://www.pinterest.es/pin/318981586099467372/?lp=true*

- **Regresión lineal**, para formar relaciones entre datos. No obstante, es insuficiente en espacios multidimensionales donde intervienen más de dos variables. Se basan en la clasificación supervisada.



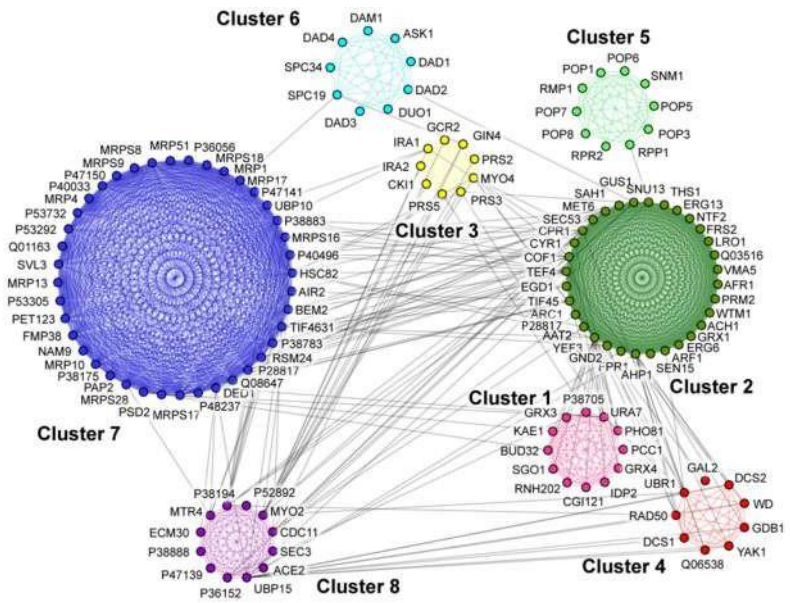
Representación gráfica de una regresión lineal. Fuente: http://www.dataprix.com/blog-it/business-intelligence/mineria-datos/data-mining-basico-correlaciones-regresiones-mercado-valores-excel

- **Árboles de decisión,** para hacer modelos de predicción, representan y categorizan una serie de condiciones, cuya visualización tiene forma de árbol y facilita la comprensión del modelo. Se basan en la clasificación supervisada.



Representación gráfica de un árbol de decisión. Fuente: https://es.slideshare.net/pagonzalez/20161121bigprivacybigdataprivacidadeustatseminariointernacional

- **Agrupamiento**, *clustering*, sirve para ver agrupaciones de datos según criterios de distancia. El agrupamiento se hace sobre la base de jerarquías y partiendo de una fragmentación completa de los datos, estos se van agrupando. Esta técnica analiza los datos que no tienen ninguna etiqueta o información añadida. Por lo tanto, se tienen que descubrir grupos similares en los grupos de datos. Los datos que quedan más cercanos son los que tienen características comunes. Se basan en la clasificación no supervisada.



**Representación gráfica del clustering. Fuente:*
http://www.cs.toronto.edu/~juris/data/rnsc/cluster.png

  - **Segmentación,** sirve para dividir grupos previamente existentes. Se basan en la clasificación no supervisada.
  - **Reglas de asociación** que sirven para encontrar relaciones entre combinaciones de valores en un conjunto de datos. Se basan en la clasificación no supervisada.

Minería web

La minería web extrae conocimiento e información de los enlaces, del contenido que hay en las páginas web y los logs de uso de los recursos de Internet mediante diferentes técnicas.

Creemos interesante comentar qué es la **web semántica**. Y es que la web semántica se podría definir como una nueva forma de web en la que el usuario encontrará respuestas a sus preguntas de forma menos compleja a como lo hace en la actualidad. El futuro de la web 3.0 es que la protagonista sea la semántica, así la búsqueda será más sencilla e intuitiva. En relación con esto, debemos mencionar también el **crawler** o el también conocido como araña de la web, que es un software o webbot que se encarga de recolectar URL's para procesarlas posteriormente. Cuando un crawler visita un sitio web puede:

- Elaborar un índice de las webs que hay en su sitio, explorando el contenido del texto visible, de diversas etiquetas HTML y los hipervínculos en listados en la página.
- Buscar el archivo robots.txt y la meta etiqueta robots para ver las reglas que se han establecido.

La minería web ha desarrollado sus propias técnicas debido a la heterogeneidad de la web, las estructuras de los enlaces y los datos no estructurados. Dichas técnicas son:

- **Minería de la estructura web** *(web structure mining):*

  Esta técnica extrae información de la estructura web. Se pueden identificar los usuarios con los mismos intereses y la relevancia de las páginas web analizando los enlaces.

- **Minería del contenido web** *(web content mining):*

  Analizando el contenido de la página web se extraen patrones. En función del contenido que hay en las páginas web, estas se pueden clasificar, así como se pueden extraer las opiniones que los usuarios dejan en una página.

- **Minería del uso de la web** *(web usage mining):*

  Extrae patrones de uso de los recursos de la red a partir de los logs que registran la actividad del usuario.

Minería de texto

Conjunto de técnicas para identificar y extraer conocimiento de un cuerpo de texto que contiene datos no estructurados. Es una aplicación de la lingüística computacional y del procesamiento de textos. Veamos en distintos puntos de qué se encarga la minería de texto:

- Identifica hechos y datos puntuales partiendo del documento de texto.
- Agrupa textos, a partir de las similitudes que encuentra entre la terminología de los escritos.
- Determina el tema o los temas del documento mediante la categorización automática.
- Identifica conceptos en los documentos y crea redes entre estos conceptos.
- Facilita el acceso a la información del cuerpo de texto.
- Facilita la visualización y navegación entre los diferentes textos.

**Representación gráfica sobre lo que realiza la minería de texto. Fuente: https://analisisdocumental2011.wikispaces.com/Miner%C3%ADa+de+Texto*

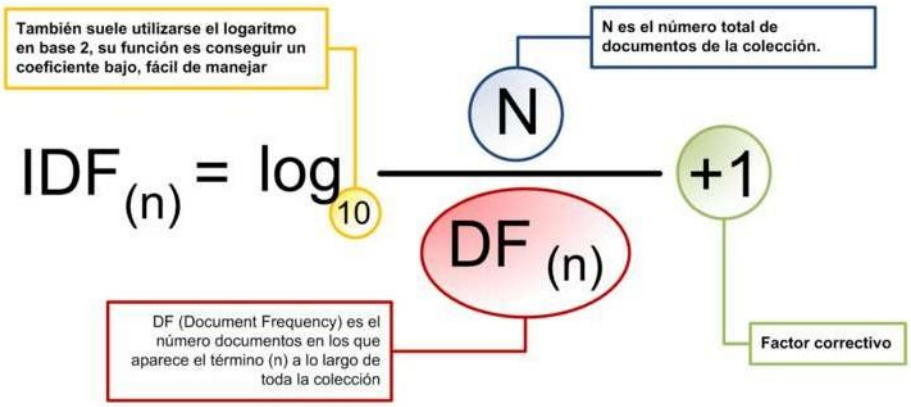
En la minería de texto también hay una fase de recuperación, preparación y valoración de la información para eliminar aquello que no necesitamos. Se deberán eliminar palabras y signos que pertenezcan a las siguientes categorías:

- Conjunciones y preposiciones *(stopwords),* únicamente tienen la función de conectar palabras y frases, y por sí solas no tienen sentido. Para eliminarlas del cuerpo de texto, se carga en el software un diccionario en el idioma correspondiente.
- Palabras derivadas (*stemming* o lematización), nos referimos a plurales, conjunciones verbales, sufijos, prefijos, etc. Se identifica la raíz de la palabra y se tiene en cuenta la que tiene mayor significado.
- Signos de puntuación, mayúsculas y números excepto en casos especiales, como cuando se trate de fechas y horas.

- Objetos específicos de páginas web, en el caso que el texto que se quiere analizar esté almacenado en webs. Algunos ejemplos son el código HTML o tags como <body>, etc.

La minería de texto dispone de modelos de representación de documentos que permiten la aplicación de técnicas numéricas en ellos. El modelo vectorial propuesto por Salton (1971) permite representar los documentos a partir de un vector de pesos asociados a un conjunto de características seleccionadas del documento (Cobo y otros, 2009). La ponderación de las características seleccionadas de cada documento se realiza con diferentes estrategias, siendo la más habitual el llamado *esquema tf.idf*, donde el peso de una característica se obtiene como producto de dos factores:

- ***Factor tf*:** mide la frecuencia de aparición de la característica en el documento.
- ***Factor idf:*** es la frecuencia inversa del documento y permite rebajar significativamente el valor de los pesos correspondientes a características con poco valor discriminante para aparecer en muchos documentos de la colección.



**Representación gráfica de cómo puede trabajar la minería de textos. Fuente: http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com.es/2012/11/frecuencias-y-pesos-de-los-terminos-de.html*

Los glosarios, tesauros, taxonomías y ontologías que hay tras la minería de datos y sus herramientas, facilitan establecer relaciones semánticas entre los términos y técnicas para la extracción de conocimiento. Ejemplos de estos vocabularios controlados son el tesauro muntilingue Eurovoc de la unión Europea, NACE (Nomenclaure statistique des

Activités économiques Dans la Communauté Européene) o la Clasificación Internacional de Patentes (CIP) (Wartena y García, 2015).

Opinion mining y sentiment analysis

Gracias a la minería de texto ahora es más fácil analizar las opiniones o comentarios de los usuarios, ya que reside una gran dificultad en analizarlos debido a que lo que se transmite es una opinión personal, y por tanto la lectura es muy subjetiva y especulativa; no es un lenguaje objetivo. Por ello se estudia el lenguaje orientado a la opinión y a la interpretación de este. Hay que saber si se habla con doble sentido, si se usa ironía, qué se quiere expresar realmente y esto es todo un reto en el que se continúa trabajando mediante técnicas de lenguaje natural (NLP, *natural language processing*).

Figura 1: Etapas del análisis de Minería de Opinión

**Representación gráfica del proceso que lleva a cabo la minería de opinión. Fuente: http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2223-30322016000200003*

Análisis de redes sociales o social network analysis (SNA)

Se estudia la interacción entre personas y organizaciones, a los que llamamos actores. Hay que tener en cuenta tres aspectos:

- **La centralidad**: analiza de qué modo se relaciona un actor con su entorno.

- **La proximidad**: calcula la distancia entre un actor y los actores que hay en su entorno.
- **La intermediación:** calcula la influencia que tiene un actor entre los suyos.

Un ejemplo de esto es analizar las empresas de un mismo sector. Analizando estos aspectos podemos ver el prestigio e influencia que tiene un actor frente a terceros.

Gestión de la reputación

Combinando minería de texto, minería de opinión y análisis de redes sociales junto con el procesamiento del lenguaje natural, se pueden recuperar fuentes de información (blogs, artículos…) para conocer la visión que hay sobre determinadas compañías u organizaciones.

Social media analytics

Cada vez es más habitual que las empresas analicen los datos que proceden de sus redes sociales para emprender acciones futuras. Este análisis se efectúa en tiempo real. Hay varios aspectos a tener en cuenta:

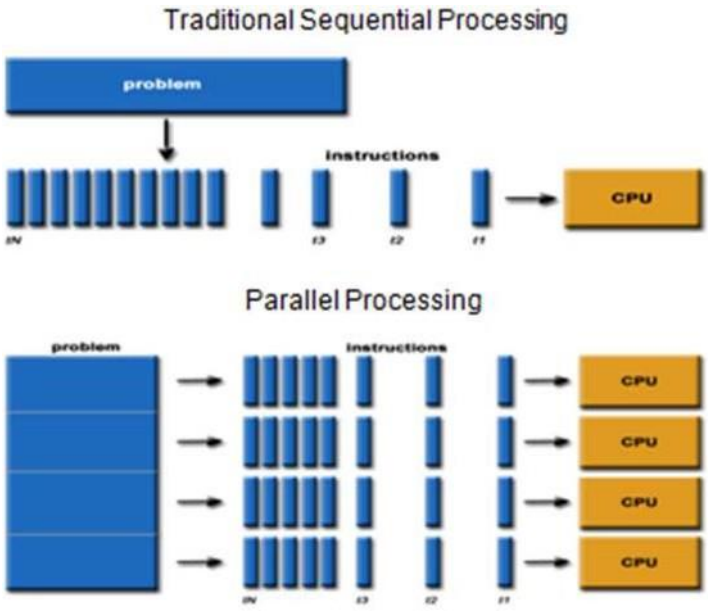
- **Visibilidad y exposición**: cantidad de tráfico hacia la web, cantidad de visitas y páginas visitadas, cantidad de seguidores y suscriptores.
- **Sentimiento y notoriedad:** número de conversaciones sobre una marca y las respectivas comparaciones con sus principales competidoras.
- **Influencia**: capacidad de un usuario de modificar el comportamiento de sus seguidores o suscriptores.
- **Engagement o vinculación emocional**: cómo interactúa la gente con una empresa u organización y con el contenido que difunde.
- **Popularidad**: número de suscriptores o seguidores que tiene una empresa. En las RRSS nos podemos fijar en el número de “me gusta”, seguidores, suscripciones y en el feedback que tenemos.

**Representación gráfica de un ejemplo de Análisis de redes sociales. Fuente: https://blog.bufferapp.com/social-media-analytics-tools*

Las mismas redes sociales ya tienen sus propias herramientas para que podamos analizar cómo funcionamos, qué feedback tenemos con nuestros seguidores, cómo son estos, etc.

Para gestionar los datos masivos y, especialmente los no estructurados y semiestructurados o híbridos, han surgido nuevos métodos de análisis para extraer la información con rapidez. Los métodos son los siguientes:

- ***Bloom filter***: almacena datos de encriptación *(hash).*
- ***Hashing***: transforma los datos en valores numéricos de longitud fija, o en valor de indexación.
- ***Index*:** mejora la velocidad de las acciones de insertar, borrar, modificar e interrogar.
- ***Triel**:* se utiliza para recuperar con rapidez información y hacer estadísticas de frecuencia de palabras.
- ***Parallel computing**:* utilización en paralelo de diversos recursos informáticos para completar las tareas computacionales. Descompone un problema y asigna las partes a varios procesos separados, para ser completados de manera individual.



**Representación gráfica de la diferencia entre el proceso tradicional y el Parallel computing. Fuente: http://www.techdarting.com/2013/07/what-is-parallel-programming-why-do-you.html*

- **Arquitectura de análisis**

La arquitectura tiene en cuenta las características de los datos masivos de velocidad, variabilidad, volumen y valor. Las arquitecturas se construyen considerando los siguientes aspectos (Chen y otros, 2014):

- La **presión de tiempo**: cuando los datos van cambiando de manera constante y tenemos la necesidad de analizarlos con rapidez, el análisis lo haremos a tiempo real. En cambio, cuando no necesitamos lo anterior, el análisis se hará *offline*. Dependiendo de qué requiramos, usaremos una cosa u otra, la que sea mejor en ese momento y, por tanto, el coste se verá reducido, seremos más eficaces y tendremos un mejor resultado.
- El nivel de **volumen de memoria** requerido para hacer el análisis: desde este punto hay que ver tres tipos de arquitectura:

  - **Análisis del nivel de memoria**: el volumen total de datos es pequeño y no sobrepasa el máximo de memoria del clúster.
  - **Análisis de inteligencia de negocio**: se usa cuando la escala de datos sobrepasa el nivel de memoria, pero puede ser importante en el entorno de análisis de la inteligencia de negocio.
  - **Análisis masivo**: el volumen de datos sobrepasa la capacidad de productos y bases de datos relacionales tradicionales.

- La **complejidad** de los diferentes datos que deben analizarse influye en el tipo de algoritmo que seleccionemos para hacer el análisis.

- **Criterios de selección**

  Como ya hemos repetido en numerosas ocasiones, hay que tener claras cuáles son nuestras necesidades, así escogeremos la opción que mejor nos vaya para lograr un buen resultado en cuanto a lograr nuestro objetivo se refiere. Lo mismo sucede a la hora de seleccionar la herramienta de análisis y visualización, ya que hay varias. Hay que saber qué tipo de análisis haremos, qué información queremos obtener, cómo lo queremos visualizar, etc., ya que no es lo mismo hacer un análisis predictivo que un análisis a tiempo real.

  Hay varios tipos de aplicaciones para mostrar el resultado de análisis, entre las que destacan: *scorecards* o cuadros de mando, informes predefinidos, informes a medida, consultas *(queries)* o cubos OLAP *(online analytic processing),* alertas, análisis estadístico, pronóstico, modelado predictivo o minería de datos, optimización y minería de procesos.

- **Visualización**

Hay que destacar la importancia del análisis visual cuando hablamos de un estudio numérico, ya que nos facilitará la comprensión del contenido informativo resultante de este.

Hay aplicaciones que ofrecen acceso a los datos de forma interactiva, simulaciones de tipos <<What if>> con la posibilidad de guardar los diferentes estudios hechos en forma de versiones, facilidad de diseño de informes con posibilidad de combinar

gráficos con resúmenes numéricos de frecuencias, disponibilidad de columnas con operadores estadísticos, gráficos con movimiento de variables, funciones de optimización de objetivos, etc. (Gironés, 2013b).

1.1.5 Tecnologías de los Big Data

Existen varias herramientas o tecnologías para la gestión de macrodatos. Dependiendo de las necesidades de nuestra organización o negocio contaremos con unas u otras. Aquí vamos a ver las más relevantes.

1.1.5.1 Hadoop

Es un software de código abierto que almacena, procesa y analiza todo tipo de datos. Diseñado para pasar de servidores individuales a miles de máquinas. Este puede crecer añadiendo módulos, que ahora se van a tratar, como son: Hadoop Distributed File System y Hadoop MapReduce. Hadoop está disponible bajo licencia Apache 2.0, que es compatible con otras licencias de código abierto. Los conjuntos de datos están distribuidos en grupos de ordenadores que utilizan modelos sencillos de programación.

1.1.5.2 Hadoop Distributed File System (HDFS)

Es un sistema de archivos basado en una arquitectura maestra-esclavo. Es escalable, tolera los fallos, y cuenta con una arquitectura distribuida. Los archivos están distribuidos en varias máquinas para su procesamiento.

1.1.5.3 MapReduce

Es un modelo de programación que permite procesar grandes conjuntos de datos en un cluster mediante un algoritmo distribuido y en paralelo. Puede ejecutarse en varios lenguajes de programación: Java, Ruby, Python y CC++.

Contienen un procedimiento *Map* para filtrar y ordenar y un procedimiento *Reduce* para crear resúmenes. Este sistema controla los servidores distribuidos, ejecutando las distintas tareas en paralelo, gestionando todas las comunicaciones y transferencias de

datos entre las distintas partes del sistema, ocupándose de redundancias y fallos, y administrando procesos en general.

1.1.5.4 Hadoop Eco System

Se llama Hadoop Eco System al conjunto de tecnologías desarrolladas para incrementar la eficiencia y funcionalidad de Hadoop. Está formado por: Apache PIG, Apache HBase, Apache Hive, Apache Sqoop, Apache Flume y Apache Zookeeper. Expliquemos algunos de estos:

- **Apache HBase**: es un modelo de base de datos de tipo no relacional, distribuido y de código abierto escrito en Java y que se ejecuta sobre HDFS. Permite almacenar grandes cantidades de datos dispersos y es tolerante a los fallos. Destaca por su comprensión, las operaciones en memoria y los filtros Bloom a nivel de columna. Las tablas de HBase pueden funcionar como entradas y salidas para las tareas MapReduce ejecutadas en Hadoop, y se puede acceder a ellas a través de la API de Java.
- **Apache Hive**: se trata de una plataforma de almacenamiento de datos que permite hacer resúmenes, consultas y análisis de los datos. Soporta análisis de grandes conjuntos de datos almacenados en sistemas de ficheros compatibles con Hadoop. Proporciona el lenguaje HiveQL, que es del tipo SQL. Mantiene un soporte completo para MapReduce y proporciona índices en los que se incluyen índices de mapas de bits. Una de las compañías que la usa en la actualidad es Netflix.
- **Apache PIG**: abstrae el lenguaje de programación de MapReduce en construcción de nivel más alto. Trabaja con tareas en paralelo pensado para crear programas para MapReduce. Pig puede ampliarse utilizando funciones definidas por el usuario, que el desarrollador puede escribir en Java, Python, JazaScript o Ruby y luego llamar desde el lenguaje.

1.2 BUSINESS INTELLIGENCE

1.2.1 Definición y características

Se conoce por Business Intelligence al conjunto de métodos que son utilizados para obtener conocimiento sobre una organización. El proceso consiste en recopilar **datos** y modelarlos para convertirlos en **información** que será analizada para convertirla en **conocimiento** útil para el negocio. Esta herramienta permite tomar decisiones en base a datos y así mejorar distintas partes del negocio: marketing, comunicación, oferta comercial, etc.

- **Datos:** mínima unidad semántica. Elementos primarios de información que por sí solos no sirven para poder tomar decisiones.
- **Información**: conjunto de datos ya procesados y que por tanto, al tener un significado nos ayudan a la toma de decisiones.
- **Conocimiento:** mezcla de experiencia, información y saber hacer que sirve para tomar decisiones.

Un ejemplo sería el siguiente:

- Datos: los tickets y facturas de compra de nuestros clientes durante el período de un año o el historial de productos vendidos.
- Información: la suma de los importes agrupados por meses o trimestres, o los artículos comprados.
- Conocimiento: descubrir aquellos meses en los que las ventas han sido más altas en cuanto a importe o número, y aquellos meses en los que han sido más bajas, así como determinar aquellos productos más vendidos y menos vendidos, o los productos que se han vendido a la vez.

Además de esta definición existen versiones de diferentes personalidades. Veamos algunas de estas:

- Business intelligence (BI) se puede definir como el uso de los datos recopilados con el fin de generar mejores decisiones de negocio, esto implica accesibilidad,

análisis y revelar nuevas oportunidades (Almeida, Ishikawa, Reinschmidt & Roeber, 1999).

- BI es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio (Sinnexus, 2007).
- BI se puede ver como el proceso en el que intervienen personas y sistemas con la meta de obtener, recopilar, analizar y presentar la información que soporte de mejor forma la toma de decisiones de negocio. El proceso se puede dividir en cuatro etapas: *extracción, consolidación, explotación y visualización* (Dávila, 2006).

1.2.2 DataWareHouse (DWH)

Un sistema de Business Intelligence cuenta con varios elementos de importancia (ETL, OLAP, etc.) que vamos a ir viendo más adelante, pero de entre todos hay uno que sobresale. Esa pieza clave es el DataWareHouse (DWH) o almacén de datos.

Veamos cómo definía *Bill Inmon* el concepto **DataWareHouse**:

"DataWareHouse es una colección de datos orientados al tema, integrados no volátiles e históricos cuyo objetivo es servir de apoyo en el proceso de toma de decisiones gerenciales" (Inmon, 1996).

Nosotros definimos DataWareHouse como un gran repertorio de datos, que a menudo está constituido por una base de datos relacional, aunque no es la única opción ya que puede estar constituido por bases de datos orientadas a columnas, por ejemplo. Esta gran cantidad de datos que puede almacenar, será después debidamente analizada y procesada para obtener así el conocimiento necesario para la toma de decisiones respecto a nuestro negocio u organización.

El DataWareHouse presenta las siguientes características:

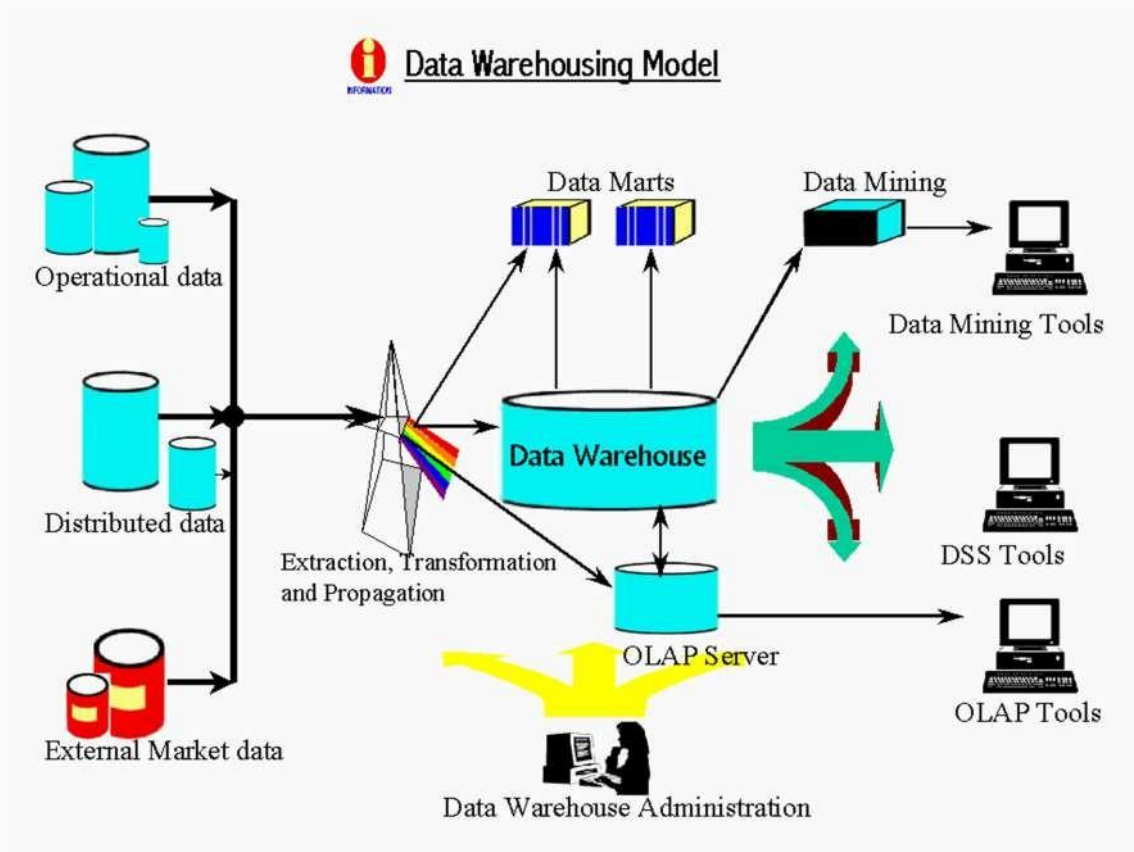
- **Orientado a un tema**: los conjuntos de datos se organizan por temas, para así facilitar el acceso a la información y el entendimiento por parte de los usuarios.
- **Integrado**: como hemos comentado, los datos proceden de diferentes fuentes de datos o entornos y presentan una estructura consistente de datos.
- **Variable en el tiempo**: se realizan fotografías de los datos basadas en fechas o hechos. El tiempo es algo muy importante, ya que teniendo esta información se pueden hacer tendencias.
- **No volátil**: es únicamente de lectura para los usuarios finales, no para ser modificado.

Hay que tener en cuenta que existen otros elementos en el contexto de un DataWareHouse:

- **Data Warehousing:** es el proceso de extraer y filtrar datos de las operaciones comunes de la organización, procedentes de los distintos sistemas de información operacionales y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un almacén de datos con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones de una organización.
- **Data Mart**: subconjunto de los datos del DataWareHouse. Tiene como objetivo responder a un determinado análisis, función o necesidad, con una población de usuarios específica. Los datos están estructurados en modelos de estrella o copo de nieve, como sucede con el DWH. Puede ser independiente o no de un DataWareHouse. El Data Mart está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de una organización.
- **Operational Data Store:** almacén de datos que únicamente proporciona los últimos valores de los datos y generalmente admite un pequeño desfase sobre los datos operacionales.
- **Staging Area**: sistema que permanece entre las fuentes de datos y el DataWarehouse con el objetivo de:
    - Facilitar la extracción de datos desde fuentes de origen con una heterogeneidad y complejidad grande.

- Mejorar la calidad de los datos.
- Ser utilizado como caché de datos operacionales con el que posteriormente se realiza el proceso de data warehousing.
- Utilización de la misma para acceder en detalle a información no contenida en el data Warehouse.

- **Procesos ETL**: tecnología de integración de datos basada en la consolidación de datos que se une tradicionalmente para alimentar DataWareHouse, Data Mart, Staging Area y ODS. A menudo se combina con otras técnicas de consolidación de datos.
- **Metadatos:** datos estructurados y codificados que describen características de instancias; aportan informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias.



**Representación gráfica de una arquitectura DataWareHouse. Fuente:*

http://www.intellibusiness.com/DWWP_anju.htm

1.2.3 Tipos de datos

Es importante saber con qué tipo de datos trabajamos. Que el proceso sea más sencillo y fiable dependerá de la tipología de datos que analicemos o con los que trabajemos.

Como hemos visto en Big Data, podemos distinguir entre datos estructurados, no estructurados y semiestructurados:

- **Estructurados:** son aquellos datos que provienen de sistemas informáticos. Normalmente tienen una calidad muy alta y se pueden usar fácilmente, como son tickets de caja, datos de sensores, etc.
- **No estructurados**: son datos que no provienen de una estructura clara y que para poder ser transformados y analizados se precisan algoritmos. Son datos no estructurados los documentos de texto, comentarios en páginas web, interacciones en redes sociales, etc.
- **Semiestructurados:** son aquellos datos que provienen de entornos estructurados, pero que necesitan una transformación más compleja. Con un tweet puedes tener fácilmente información sobre el nombre del usuario, ubicación, el día y hora que publicó el tweet. Estos son datos estructurados, pero el comentario que hay en ese tweet sí es algo que hay que descifrar, por lo que es un dato no estructurado.

En Business Inteligence los datos también se pueden diferenciar según su origen:

- **Datos internos**: son los que provienen de la empresa.
- **Datos externos:** son aquellos que provienen de sistemas analíticos. Los más usados son los que provienen de Google Analytics.

FUENTES DE DATOS	**Externas**	GPS Censo Teléfonos móviles Historial a crédito Registro inmobiliario Meteorología	Blogs Twitter Pinterest Facebook Instagram Externo a sensor
	Internas	Registros RRHH Registro ventas Financiero Perfiles web CRM Inventario	Foros online Documentos de texto Feeds de las webs Sensor Data
		Estructurados	**Desestructurados**
		TIPOS DE DATOS	

1.2.4 Calidad de los datos

Los procesos de calidad son los que se encargan de comprobar y garantizar la totalidad y validez de los datos de diferentes entornos o fuentes de datos. Estos procesos han sido cada vez más necesarios al haber aumentado los datos de entornos externos a la empresa y red 2.0 en la que las empresas y usuarios tienen canales de comunicación bidireccional.

Tales procesos pueden ser, por ejemplo, comprobar que los campos de moneda tengan posiciones decimales y así evitar redondeos y que se pierda información, o pueden tratarse de procesos más complejos como terminar de rellenar información de un cliente.

Veamos dos ejemplos para aclarar este punto:

- Si tenemos un negocio en el que usuarios se pueden dar de alta completando un formulario, puede ser que en ocasiones no se rellenen todos los campos. Pongamos de ejemplo, que un chico llamado Héctor se ha dado de alta sin rellenar el campo de género, por el nombre sabemos que es hombre. Pero si una persona llamada Andrea se ha dado de alta y su nacionalidad la ha marcado como italiana, no podemos saber si es mujer o hombre. Por este motivo es necesario crear diccionarios que en función de los valores de distintas dimensiones, nos ayuden a terminar de garantizar la integridad de la información.
- Analizar los comentarios de los usuarios en redes o páginas web es uno de los procesos más complejos puesto que dependiendo del contexto, puede significar una cosa u otra. Si por ejemplo, alguien utiliza la palabra “helado” al hablar de un restaurante cualquiera, quizás se refiere a algo negativo, en cambio, si habla de una horchatería, puede que la palabra sea positiva. El uso de la ironía y el doble sentido complican la interpretación de los comentarios. Estos datos se manejan a partir de diccionarios y se crean algoritmos de aprendizaje automático.

1.2.5 Etapas de un proceso de Business Intelligence

Como hemos comentado anteriormente, se pueden distinguir cuatro etapas dentro de un proyecto de Business Intelligence:

- **Etapa de extracción**

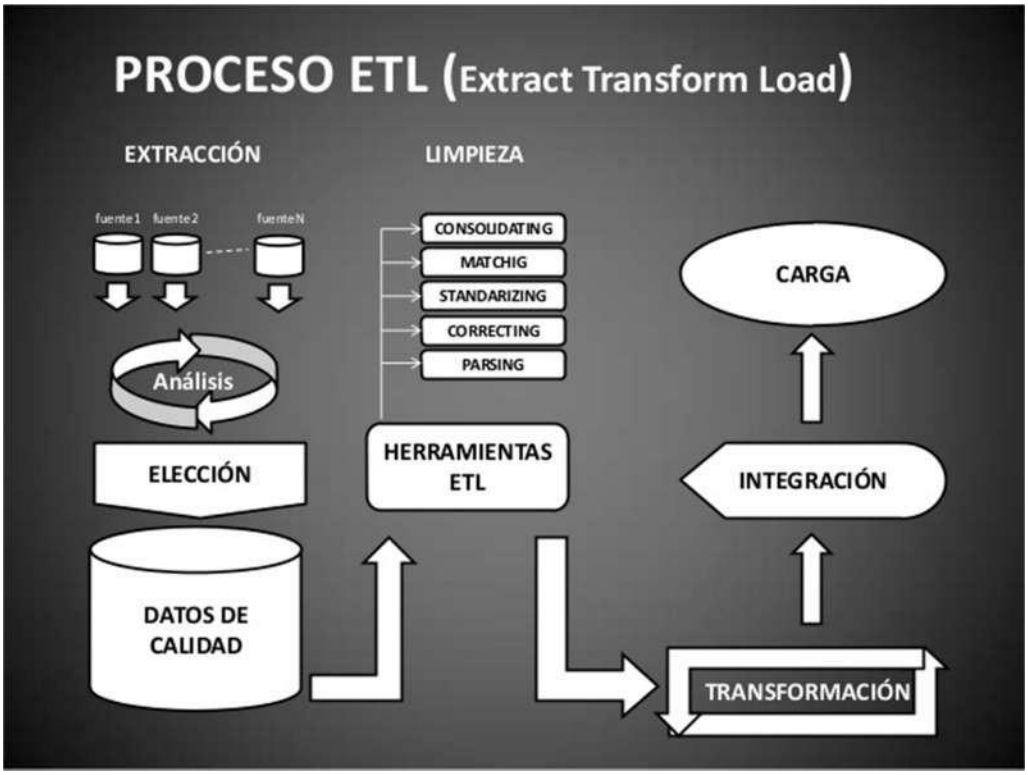
  Implementar un proceso de Business Intelligence en una organización o empresa, empieza con la selección de información relevante para la toma de decisiones y como ya hemos ido diciendo, para seleccionar esa información lo que necesitamos son datos. Como ya hemos comentado, para saber qué datos necesitamos analizar, hay que tener claro qué queremos lograr con nuestro negocio, qué nos interesa y cuál es nuestra necesidad.

Una vez ya identificada y seleccionada esa información, pasamos a la siguiente etapa, la consolidación.

- **Etapa de consolidación**

  Es en esta etapa donde tiene lugar **el proceso ETL** *(Extract, Transformation y Load),* que es un conjunto de procesos por medio de los cuales los datos de la fuente operacional son preparados para colocarse en el DWH (Kimball, 2002).

  El proceso ETL es una tecnología que tiene la función de integración de los datos, lo que significa que debe ofrecer una única visión de los datos. Este proceso se encarga de la extracción, transformación y carga de esos datos como su propio nombre indica. También tiene la función de gestionar estos datos. Debe asegurar su integridad, coherencia y disponibilidad en el destino.



**Representación gráfica de un Proceso ETL. Fuente: https://es.slideshare.net/dfernang/proceso-etl*

Teniendo en cuenta que en la etapa anterior hemos establecido cuáles son nuestras necesidades de información y hemos visto qué datos necesitamos,

esta segunda etapa consiste en recopilar los datos de distinta procedencia con el fin de normalizarlos, depurarlos y estructurarlos, almacenándolos en el DWH. En esta etapa de consolidación se precisa utilizar diferentes metodologías, técnicas, hardware y los componentes de software que proporcionan en conjunto la infraestructura para soportar el proceso de información.

Data Mart es una de las metodologías más utilizadas, de la que ya hemos explicado anteriormente cuál es su función en un apartado anterior. No obstante, veremos aquí cómo lo define Kimball.

"Data Mart es el subconjunto lógico y físico del área de presentación de datos en un DWH. "

Finalizada la etapa de consolidación, pasamos a la de explotación.

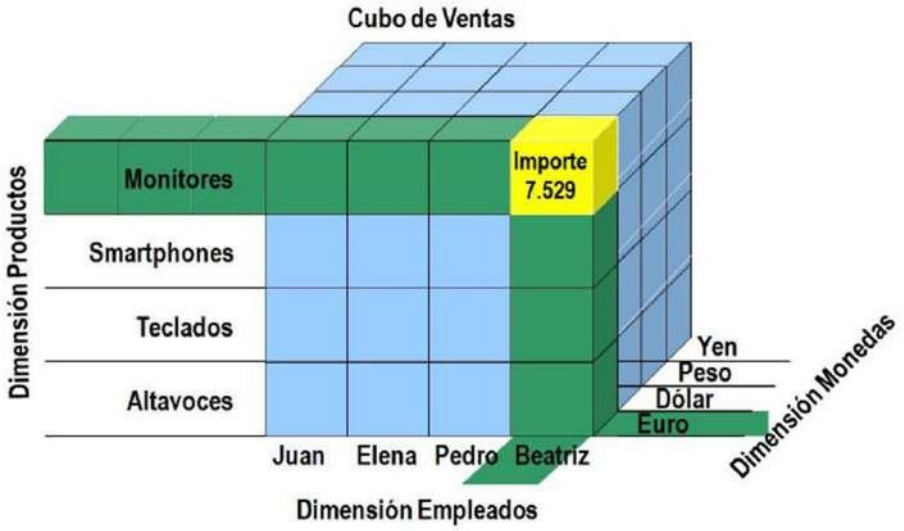
- **Etapa de explotación**

  En esta tercera etapa es en la que se aplican una serie de herramientas para dejar listos los datos del DWH en manos de los usuarios. Estos deberán tener capacidad suficiente para aprovechar y explotar la información ya filtrada que tenemos en el DWH. Existen dos tecnologías que nos permiten llevar a cabo esta explotación de los datos:

  - **Cubos OLAP**: es un sistema que permite el análisis multidimensional. Dicho análisis consiste en modelar la información en medidas, dimensiones y hechos.

    Las medidas son los valores de un dato, las dimensiones son las descripciones de las características que definen ese dato y los hechos corresponden a la existencia de valores específicos de una o más medidas para un combinación particular de dimensiones.

    Es un método ágil y flexible para organizar datos, en especial los metadatos. Mediante consultas o informes, tiene el objetivo de recuperar y manipular datos y combinaciones de estos.



** Representación gráfica de un cubo OLAP. Fuente: http://codebotic.blogspot.com.es/2015/12/cubo-olap-cinema-i.html*

El almacenamiento físico de los datos tiene lugar en un vector multidimensional. En un sistema OLAP puede haber más de tres dimensiones, por ello, también reciben el nombre de hipercubos.

Veamos los tipos de sistemas OLAP:

- **ROLAP**: almacena los datos en un motor relacional. Los datos son detallados, se evitan agregaciones y las tablas se encuentran normalizadas.
- **MOLAP**: los datos se almacenan en una base de datos multidimensional. Se calcula de antemano el resumen de información para optimizar los tiempos de respuesta.
- **HOLAP**: los datos se almacenan en una base de datos multidimensional y en un motor relacional.

- **Minería de datos**: conjunto de métodos y técnicas para procesar y analizar los datos con el fin de obtener el conocimiento, a priori, no visible.

  Las técnicas de minería de datos provienen de la inteligencia artificial y de la estadística. Estas técnicas o métodos son algoritmos más o menos sofisticados, que como hemos comentado ya, se aplican sobre un

conjunto de datos para obtener un resultado que se convierte en conocimiento para nosotros y nuestro negocio. Según su finalidad, estos algoritmos pueden clasificarse en:

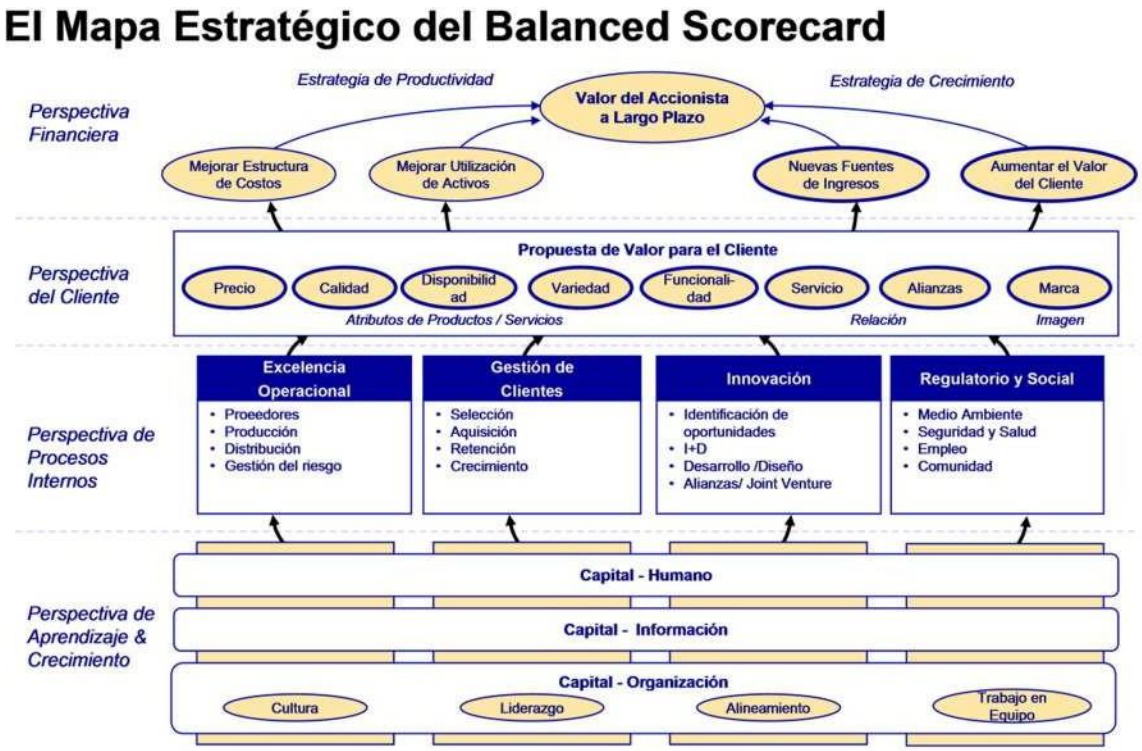
- Regresión.
- Clasificación.
- Agrupamiento.

Se usan algoritmos de redes neuronales, algoritmos genéticos, árboles de decisión, máquinas de soporte vectorial, etc. Algunos de estos algoritmos precisan de supervisión humana para obtener unos resultados fiables y de calidad.

- **Etapa de visualización**

  En la etapa de visualización, los usuarios pueden saber mediante ciertas herramientas gráficas, qué sucede en su empresa u organización y así tomar las decisiones pertinentes. En esta etapa intervienen las siguientes metodologías y/o herramientas:

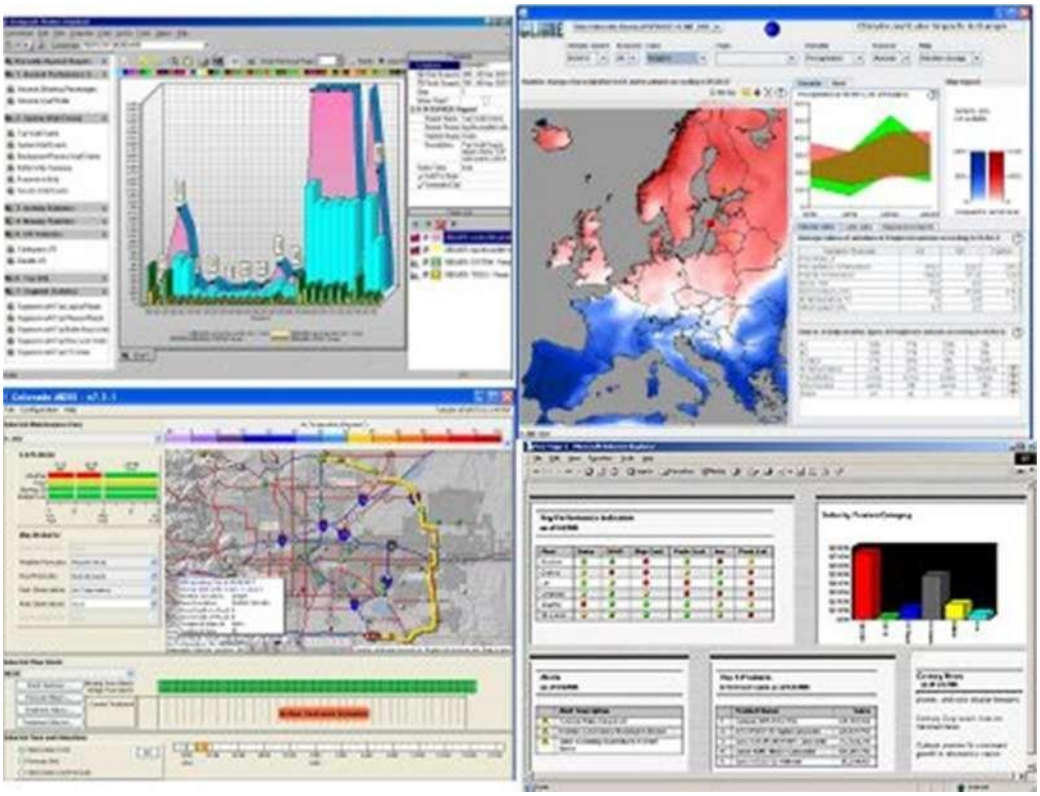
  - **Balance Scored Card o Cuadro de Mando Integral (CMI):** es de gran ayuda en la planificación estratégica de las empresas. Quiere evitar que los objetivos que tiene la empresa se centren únicamente en el corto plazo y en los rendimientos actuales, e intenta que se tenga una perspectiva más amplia.



**Representación gráfica del mapa estratégico del Cuadro de Mando Integral. Fuente: https://margaritaberdugo.wordpress.com/2015/10/29/el-balanced-scorecard/*

También se puede considerar como una aplicación que ayuda a una compañía a expresar los objetivos e iniciativas necesarias para cumplir con su estrategia, mostrando de forma continuada cuándo la empresa y los empleados alcanzan los resultados definidos en su plan estratégico. (Sinnexus, 2007).

- **Sistemas de Soporte a la Decisión (DSS):** es una herramienta enfocada al análisis de los datos de una empresa u organización. Analizar los datos es algo complejo. Las aplicaciones como esta normalmente disponen de informes predefinidos en los que presentan la información de forma estática, pero no permiten profundizar en los datos, navegar entre ellos, manejarlos desde distintas perspectivas, etc. (Sinnexus, 2007).



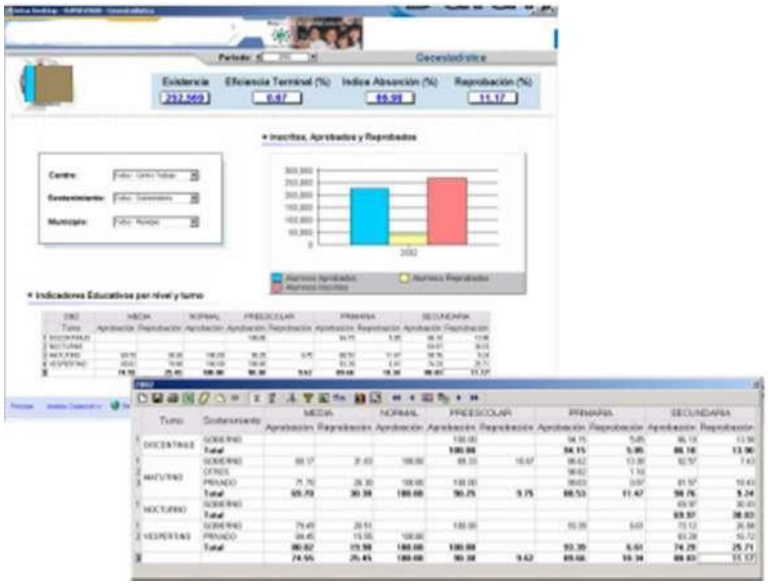
**Representación gráfica de diferentes Sistemas de Soporte de la Decisión. Fuente: http://www.iiia.csic.es/udt/es/artificialintelligence/sistemas-soporte-decisiones*

Gracias a los Sistemas de Soporte de la Decisión no es necesario ser un experto técnico ni tener necesidad de recurrir al departamento de sistemas, puesto que permiten que los usuarios generen informes dinámicos y flexibles y así, tener a mano la información histórica que requieren.

- **Sistemas de Información Ejecutiva (EIS):** es una aplicación que muestra informes y listados *(query & reporting)* de las distintas áreas de negocio, y lo hacen de forma consolidada facilitando así la monitorización de la empresa o parte de ella si estamos trabajando en una unidad de esta. El EIS permite que el usuario final visualice cuál es el panorama de su negocio y puede hacerlo de forma general o en aspectos concretos.

  El EIS se caracteriza por ofrecer al ejecutivo un acceso rápido y efectivo a la información compartida, utilizando interfaces gráficas visuales e

intuitivas. Suele incluir alertas e informes basados en excepción, así como históricos y análisis de tendencias. También es frecuente que permita la domiciliación por correo de los informes más relevantes. (Sinnexus, 2007).



**Representación gráfica de un Sistema de Información Ejecutiva. Fuente: http://www.sinnexus.com/business_intelligence/sistemas_informacion_ejecutiva.aspx*

Esta última etapa tiene como objetivo ofrecer al usuario una visualización de los informes y consultas, pudiendo personalizar estas y acceder a ello fácilmente.

1.2.6 Herramientas de Business Intelligence

En el momento de elegir una herramienta de Business Intelligence, hay que tener en cuenta diferentes aspectos, ya sea a nivel funcional como a nivel tecnológico. Algunas de las cosas a tener en cuenta son las siguientes:

- **El uso que se le dará a la herramienta**: hay que tener en cuenta para qué sirve, Si sirve para analizar, ejecutar un proceso periódicamente, para publicar información en una intranet, etc.
- **Es de gran importancia el mapa de sistemas de la compañía**: lo lógico es que si todas las herramientas de una empresa o compañía se desarrollan en un

lenguaje específico, las herramientas que se busquen se integren bien con este lenguaje.

- **Base implantada de la solución**: es todo un reto encontrar profesionales que entiendan de ciertas tecnologías, esto puede crear cierta dependencia de terceros o tener que pagar tarifas más altas.
- **Presupuesto**: es un punto clave. No nos referimos solo al presupuesto del que disponemos para llevar a cabo el proyecto, sino a no salirnos de ese presupuesto establecido. Dependiendo de qué software vayamos a usar, por ejemplo, ese presupuesto puede aumentar. Cualquier decisión puede afectar.

En un proyecto de Business Intelligence son varias las herramientas que pueden utilizarse. Veamos cuáles son esas herramientas:

- **Bases de datos**: Se usan para almacenar el DWH *(DataWareHouse),* en algunos proyectos es algo imprescindible tener una base de datos para crear el DWH y guardar todos los pasos intermedios del proceso ETL.
  - Oracle.
  - Microsoft SQL Server.
  - Amazon Redshift.
  - SQL Database.
  - Teradata, HANA, Google BigQuery, HP Vertica, Netezza, etc.

  Hay una gran cantidad de bases de datos, pero tienen un precio más elevado o no son tan conocidas.

- **Herramienta ETL**: es necesario usar esta herramienta cuando precisamos integrar datos de distinta procedencia o cuando la explotación de los datos no será de forma directa. Algunas de estas herramientas son:
  - Microsoft SqlServer Integration Services.
  - SAP DataServices.
  - IBM Infosphere DataStage.
  - Pentaho Data Integration.
  - PowerCenter.

  - Talend.
- **Plataforma de análisis**: son herramientas que se nutren de información y que nos permiten navegar por los datos de manera gráfica. También se llaman datadiscovery.
- **Reporting**: se utiliza para la creación de listas o fichas corporativas.
- **Cuadros de mando**: se utilizan para hacer el seguimiento, de manera visual, de los KPIs estratégicos de la compañía.
- **Herramientas de front**: estas herramientas pueden ser de reporting y análisis o de análisis y cuadros de mando.
  - Microsoft PowerBI.
  - Sap Business objects.
  - SAS.
  - Qlik.
  - Tableau.
  - R Studio.
- **Presupuestación**: esta herramienta sirve para introducir presupuestos, estimaciones, objetivos, etc.
  - SAP BPC.
  - Hyperion.

1.3 BIG DATA VS. BUSINESS INTELLIGENCE

En los apartados anteriores hemos visto qué es Big Data y Business Intelligence y sus características, tipología de datos, etc. Ahora hablaremos de cómo se pueden complementar, sus similitudes y sus diferencias.

Cuando se empezó a hablar de Big Data, se pensaba que este terminaría sustituyendo a Business Intelligence ya que resolvía problemas que este no había sido capaz de resolver. Sin embargo, el tiempo ha demostrado que para llevar a cabo una buena estrategia se requiere el uso de ambas tecnologías. Su combinación es la clave, ya que el Business Intelligence nos aporta el seguimiento y control de la estrategia de la compañía

y por su parte, el Big Data nos aporta la capacidad de encontrar patrones disruptivos sobre los que innovar.

A partir de un Terabyte (TB) de información es cuando consideramos que hay un gran volumen de datos. Recordemos que *Volumen* es una de las 3V's de Big Data y que cuando encontramos una de estas recurriremos por tanto a esa tecnología. Si no tenemos un gran volumen de datos, utilizaremos Business Intelligence ya que mantener la estructura Big Data con menos información resultará caro e innecesario cuando Business Intelligence puede soportar perfectamente ese volumen.

Se requieren bases de datos nuevas (lo veremos en el siguiente apartado) cuando combinamos proyectos de Big Data y Business Intelligence. Las bases de datos tradicionales tienen la capacidad de insertar mucha información y para garantizar esta información tienen muchas reglas. A causa de esto, los procesos de carga de información son largos e ineficientes cuando hablamos de grandes volúmenes. Dependiendo de nuestras necesidades vamos a combinar bases de datos. Esta combinación de datos de diferente naturaleza se llama *"polyglot persistence".*

Los principales fabricantes de software de Business Intelligence han empezado a integrar las soluciones de Big Data para poder completar los Data WareHouse (DWH).

La información interna y estructurada se sigue cargando en el DWH y la externa o desestructurada se carga en la de Big Data, que está mejor dimensionada para ese tipo de tareas. Mediante técnicas como MapReduce se puede añadir esta información y cargarla en el DWH. Así la información que no tenga gran volumen y sea menos cambiante se seguirá cargando en el DWH, y la otra se cargará en la parte de Big Data.

A través de las herramientas ETL, presentadas anteriormente, cargamos los datos de la arquitectura de Big Data agregados en el DWH.

En la parte de front, las visualizaciones o reporting corporativo se nutren del DWH y, las aplicaciones de la compañía pueden acceder al DWH o a la arquitectura de Big Data. Si son simples estadísticas con una gran velocidad podemos utilizar el BI. Si por el

contrario, necesitamos más velocidad o realizar consultas o bases de datos más extensas, se utilizarán bases de datos más acordes a nuestras necesidades.

A través de la siguiente tabla vamos a mostrar las diferencias que hay entre estos dos conceptos

	BUSINESS INTELLIGENCE	**BIG DATA**
Tipología de datos	Estructurados	No Estructurados o Semiestructurados
Volumen de datos	Grande	Enorme
Generación de datos	Transaccional	Bajo demanda
Para qué sirve	Para encontrar respuestas concretas a problemas concretos	Permite descubrir tendencias y patrones que no siempre se conocen las respuestas
Complementariedad	Proporciona herramientas de análisis para tablas de Big Data	Permite filtrar y organizar datos no estructurados o semiestructurados para Business Intelligence