

UD 1 Gestión de Soluciones - Diseño y Construcción de Soluciones

1. Introducción de los datos al conocimiento

El **dato** es una representación sintáctica, generalmente numérica, que puede manejar un dispositivo electrónico - normalmente un ordenador - sin significado por sí solo. Sin embargo, el dato es a su vez el ingrediente fundamental y el elemento de entrada necesario en cualquier sistema y/o proceso que pretenda extraer información o conocimiento sobre un dominio determinado. En este sentido, 7 es un dato, como también lo es π o como son los términos aprobado o suspenso.

Por su parte, la **información** es el dato interpretado, es decir, el dato con significado. Para obtener información, ha sido necesario un proceso en el que, a partir de un dato como elemento de entrada, se realice una **interpretación** de ese dato que permita obtener su significado, es decir, información a partir de él. La información es también el elemento de entrada y de salida en cualquier proceso de toma de decisiones. Partiendo de los datos del ejemplo anterior, información obtenida a partir de los mismos puede ser: El 7 es un número primo, π es una constante cuyo valor es 3, 141592653..., María ha aprobado el examen de conducir, Pablo está suspenso en matemáticas.

A partir de información, es posible construir conocimiento. El **conocimiento** es información aprendida, que se traduce a su vez en reglas, asociaciones, algoritmos, etc. que permiten resolver el proceso de toma de decisiones. Así pues, la información obtenida a partir de los datos permite generar conocimiento, es decir, aprender. El conocimiento no es estático, como tampoco lo es siempre el aprendizaje. Aprender, construir conocimiento, implica necesariamente contrastar y validar el conocimiento construido con nueva información que permita, a su vez, guiar el aprendizaje y construir conocimiento nuevo. Siguiendo con los ejemplos anteriores, el conocimiento que permite obtener que el 7 es un número primo puede ser el *algoritmo de Eratóstenes*. Por otra parte, el conocimiento que permite obtener el valor del número π puede extraerse de los resultados de los trabajos de *Jones, Euler o Arquímedes*, mientras que el aprobado de María en el examen de conducir y el suspenso de Pablo en matemáticas, se pueden obtener de la regla que en una escala de diez asigna el aprobado a notas mayores o iguales que 5 y el suspenso a notas menores.

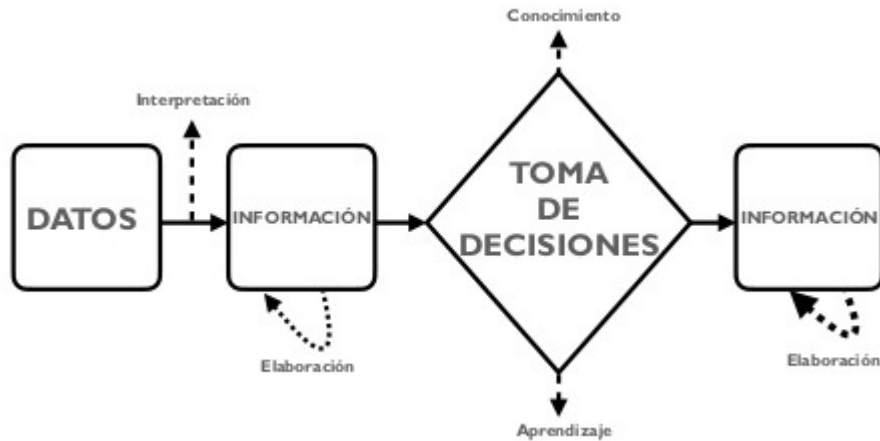


Figura 1.1: Relación entre datos, información y conocimiento en el proceso de toma de decisiones. (Fuente: UCLM)

Por tanto, datos, información y conocimiento están estrechamente relacionados entre sí y dirigen cualquier proceso de toma de decisiones. La [figura 1.1](#) muestra la relación entre datos, información y conocimiento, en un proceso genérico de **toma de decisiones**. Más concretamente, en el ejemplo del suspenso de Pablo en matemáticas, el proceso de toma de decisión acerca de la calificación de Pablo se estructuraría de la siguiente forma:

1. El profesor corrige el examen de Pablo, que ha sacado un 3. Esta calificación, por sí sola, es simplemente un dato.
2. A continuación, el profesor calcula la calificación final de Pablo, en base a la nota del examen, sus trabajos y prácticas de laboratorio. *La nota final de Pablo es un 4*. Esto último es información.
3. ¿Ha aprobado Pablo? La información de entrada al proceso de decisión es su calificación final de 4 puntos, obtenida en el paso anterior. El conocimiento del profesor sobre el sistema de calificación le indica que una nota menor a 5 puntos se corresponde con un suspenso y, en caso contrario, con un aprobado.
4. La información de salida tras este proceso de decisión es que *Pablo está suspenso en matemáticas*.

Question

Siguiendo el ejemplo anterior ¿Cómo se produce el proceso de toma de decisiones para determinar si un número es primo?

Aunque se trate de un ejemplo trivial, la importancia del proceso de toma de decisiones no lo es. En **marketing**, por ejemplo, se analizan bases de datos de clientes para identificar distintos grupos e intentar predecir el comportamiento de estos. En el mundo de las **finanzas**, las

inversiones realizadas por grandes empresas responden a un proceso complejo de toma de decisiones donde los datos son el eje fundamental de este proceso. En **medicina**, existe una gran cantidad de sistemas de ayuda a la decisión que permiten a los doctores contrastar y validar sus diagnósticos de forma precoz. En definitiva, no hay área de conocimiento ni ámbito de aplicación que escape al proceso de toma de decisiones.

✓ Success

Siguiendo Hablando de datos en información en Big Data. Observemos un fragmento de esta entrevista a [@JaimeObregon](#) hablando sobre una parte del gran trabajo que realiza



La información y los datos (Fuente: [x.com](#))

2. La carrera entre los datos y la tecnología

Que los datos son el elemento fundamental en cualquier proceso y/o sistema de **toma de decisiones** no es algo nuevo. Sin embargo, los datos no siempre han estado al alcance de los expertos y no siempre ha sido posible ni sencillo procesarlos según las necesidades concretas de cada caso de aplicación.



La información, por tanto, siempre ha sido poder y el gran reto ha sido y sigue siendo extraer información a partir de datos para generar conocimiento. Para ello, es necesario contar con dos factores que deben estar alineados: **datos y tecnología**.

Obtener **datos** no ha sido siempre una tarea fácil. Esto es debido principalmente a que la gran cantidad de sensores disponibles en la actualidad, que permiten registrar magnitudes de cualquier proceso, no existía como a día de hoy. Además, los sensores existentes en esta

época (finales del siglo XX y comienzos del siglo XXI) no estaban ampliamente extendidos, ya que sus prestaciones estaban lejos de las que ofrecen hoy y sus precios no estaban al alcance de cualquier usuario. Por tanto, los procesos que se monitorizaban y de los cuales se recogían datos eran, sobretodo, procesos industriales realizados en grandes empresas. Por todos estos motivos, tradicionalmente se recurría a **modelos de simulación** que, a través de la implementación de un modelo matemático, permitían generar datos realistas de un proceso.

Los datos generados mediante simulación son conocidos como **datos sintéticos** mientras que los datos provenientes de las lecturas de un sensor se conocen como **datos reales**.

Pero con los datos no es suficiente. Es necesario también contar con la tecnología necesaria para su procesamiento. Generar, almacenar y procesar todos estos datos no es una tarea trivial, y plantea una serie de problemas tecnológicos a resolver.

- Primer problema tecnológico a resolver, el **almacenamiento**. Algunas soluciones propuestas pasan por los **sistemas de información distribuida**, entendidos como un conjunto de ordenadores separados físicamente y conectados en red destinados al almacenamiento de datos o por los **sistemas de información en la nube**, que permiten adquirir espacio de almacenamiento en servidores privados, dejando la gestión de estos servidores en manos del proveedor.
- El segundo problema tecnológico es el **procesamiento** de los datos almacenados. Este aspecto cobra especial relevancia en función del caso de aplicación, pudiendo distinguirse entre procesamiento on-line (en línea/**stream**) y procesamiento off-line (fuera de línea/**batch**).
-  En el caso del **procesamiento on-line/stream processing**, los datos son procesados a medida que son generados, ya que se requiere una respuesta en tiempo real. Por ejemplo, en un sistema de control del tráfico que permite regular los semáforos en función del tráfico actual, el sistema debe regular el semáforo a medida que se van generando e interpretando los datos del tráfico en un instante de tiempo dado.
-  Por otra parte, en el caso del **procesamiento off-line/batch processing**, no es necesario que los datos se procesen a medida que se generan. Por ejemplo, en un sistema de detección del fraude bancario, comprobar si un cliente ha realizado algún movimiento fraudulento es una tarea que puede llevarse a cabo off-line, por ejemplo, haciendo un análisis de los movimientos del cliente en un momento dado, sin tener por qué diagnosticar cada movimiento que este va realizando.


En este sentido, la **computación distribuida**, en donde múltiples máquinas realizan el procesamiento optimizando el rendimiento o la **computación en la nube**, que permite adquirir recursos de procesamiento al igual que se puede adquirir espacio de almacenamiento, son dos soluciones al problema del procesamiento.

Otras alternativas son la **programación paralela** y la **programación multi-procesador**, que permiten, respectivamente, aprovechar el paralelismo de múltiples hilos de ejecución dentro de

un procesador y realizar el procesamiento dividiéndolo en múltiples hilos en diferentes procesadores


Question

Piensa en procesos cotidianos que requieran un procesamiento on-line y en otros que requieran un procesamiento off-line.




 En la actualidad, la proliferación de una gran cantidad de sensores con altas prestaciones y precios asequibles que permiten monitorizar y generar datos sobre cualquier proceso ha supuesto un **incremento exponencial en la cantidad de datos generados**. Es posible monitorizar casi cualquier proceso, incluyendo los domésticos como el consumo eléctrico de un hogar, la presencia dentro del mismo o procesos cotidianos como la actividad física, entre otros muchos. Hoy, los datos llevan la delantera en la carrera entre datos y tecnología. Si bien es cierto que la tecnología ha experimentado grandes avances en los últimos años, la cantidad de datos generada no deja de crecer. Esto supone un **reto permanente para la tecnología**, que sigue evolucionando a nivel hardware con la aparición de arquitecturas con mayores posibilidades de procesamiento, almacenamiento y a nivel software, con la aparición de modelos de programación que optimizan el procesamiento de los datos.

3. Los datos: los de ayer y los de hoy

Al igual que la tecnología ha ido evolucionando para dar respuesta a la ingente cantidad de datos que ha comenzado a generarse, estos últimos también han experimentado una gran evolución. Esta evolución, o revolución, no está únicamente relacionada con la **cantidad** de datos (como se expuso en el anterior apartado) sino también con el **tipo** y el **formato** de los mismos.

 Tradicionalmente, el tipo y formato de datos con el que se ha trabajado para extraer información y conocimiento a partir de ellos era ciertamente **limitado**. En muchas ocasiones se trataba de ficheros de datos estructurados de forma tabular, donde cada fila del conjunto de datos representaba una instancia del mismo y cada columna una variable o atributo de la instancia. El formato de archivo que se manejaba solían ser **formatos de hojas de cálculo** (.xlsx, .ods, .numbers etc) o **ficheros separados por comas** (.csv). Muy pocos eran los procesos en los que se trabajaba con otros tipos de datos como texto, imágenes, audio e incluso vídeos, ya que los formatos de estos tipos de datos eran limitados hace unos años, su procesamiento más complejo y la tecnología para ello aún en desarrollo.

Aunque a día de hoy también se sigue trabajando con archivos de datos en forma de hojas de cálculo y/o archivos tradicionales para generar conocimiento a partir de ellos, las posibilidades actuales son prácticamente **ilimitadas**.

-  En cuanto al **texto**, las técnicas de inteligencia artificial y procesamiento del lenguaje natural hacen posible la extracción de conocimiento a partir de **grandes volúmenes de textos**, que pueden provenir de páginas web, archivos .pdf, redes sociales, etc.
-  El desarrollo de hardware con mejores prestaciones y los nuevos modelos de programación permiten procesar en la actualidad **grandes cantidades de imágenes, audios y vídeos** con una gran variedad de técnicas de inteligencia artificial en tiempos razonables.
-  Finalmente, han aparecido nuevos tipos y formatos de datos, como por ejemplo, aquellos datos generados a partir de **grafos**, los cuales se tratarán en próximas secciones y capítulos con más detenimiento. Estos datos se corresponden, por ejemplo, con datos geográficos obtenidos a partir de mapas como los generados en aplicaciones como Google Maps u Open Street Maps o datos de seguimiento y actividad en redes sociales de gran valor en campañas publicitarias entre otros muchos.

Question

Haz una búsqueda y elabora un listado con distintos tipos de datos y los formatos de almacenamiento más utilizados con los que se trabaja en ciencia de datos y big data.

Los diferentes tipos y formatos de datos, los de ayer y los de hoy, son la materia básica fundamental en cualquier proceso de extracción de información y de conocimiento. Después, las metodologías empleadas para ello y arquitecturas hardware sobre las que se realice el procesamiento de los mismos, permitirán definir **procesos y metodologías de big data**, aplicadas a un ámbito concreto.

4. Soluciones Big Data

En esta nueva era tecnológica en la que nos hayamos inmersos, a diario se generan enormes cantidades de datos, del orden de **petabytes** (más de un millón de gigabytes) **en muy cortos períodos de tiempo**. Hoy en día, cualquier dispositivo como puede ser un reloj, un coche, un smartphone, etc está conectado a Internet generando, enviando y recibiendo una gran cantidad de datos. Tanto es así, que se estima que el 90 % de los datos disponibles en el mundo ha sido generado en los últimos años. Sin lugar a dudas, esta y las próximas generaciones serán las generaciones del big data.

Esta realidad descrita anteriormente demanda la capacidad de enviar y recibir datos e información a gran velocidad, así como la capacidad de almacenar tal cantidad de datos y procesarlos en tiempo real. Así pues, la gran cantidad de datos disponibles junto con las herramientas, tanto hardware como software, que existen a disposición para analizarlos se conoce como **big data**.

👉 No existe una definición precisa del término **big data**, ni tampoco un término en castellano que permita denominar este concepto. A veces se usan en castellano los términos *datos masivos* o *grandes volúmenes de datos* para hacer referencia al big data. Por este motivo, a menudo el concepto de big data es definido en función de las características que poseen los datos y los procesos que forman parte de este nuevo paradigma de computación. Esto es lo que se conoce como 👉 **las Vs del big data** 👉.

Algunos autores coinciden en que big data son datos cuyo **volumen** es demasiado grande como para procesarlos con las tecnologías y técnicas tradicionales, requiriendo nuevas arquitecturas hardware, modelos de programación y algoritmos para su procesamiento. Además, se trata de datos que se presentan en una gran **variedad** de estructuras y formatos: datos sintéticos, provenientes de sensores, numéricos, textuales, imágenes, audio, vídeo... Finalmente, se trata de datos que requieren ser procesados a gran **velocidad** para poder extraer valor y conocimiento de ellos. Esta concepción se conoce como las tres Vs del big data (ver [figura1.2](#)).

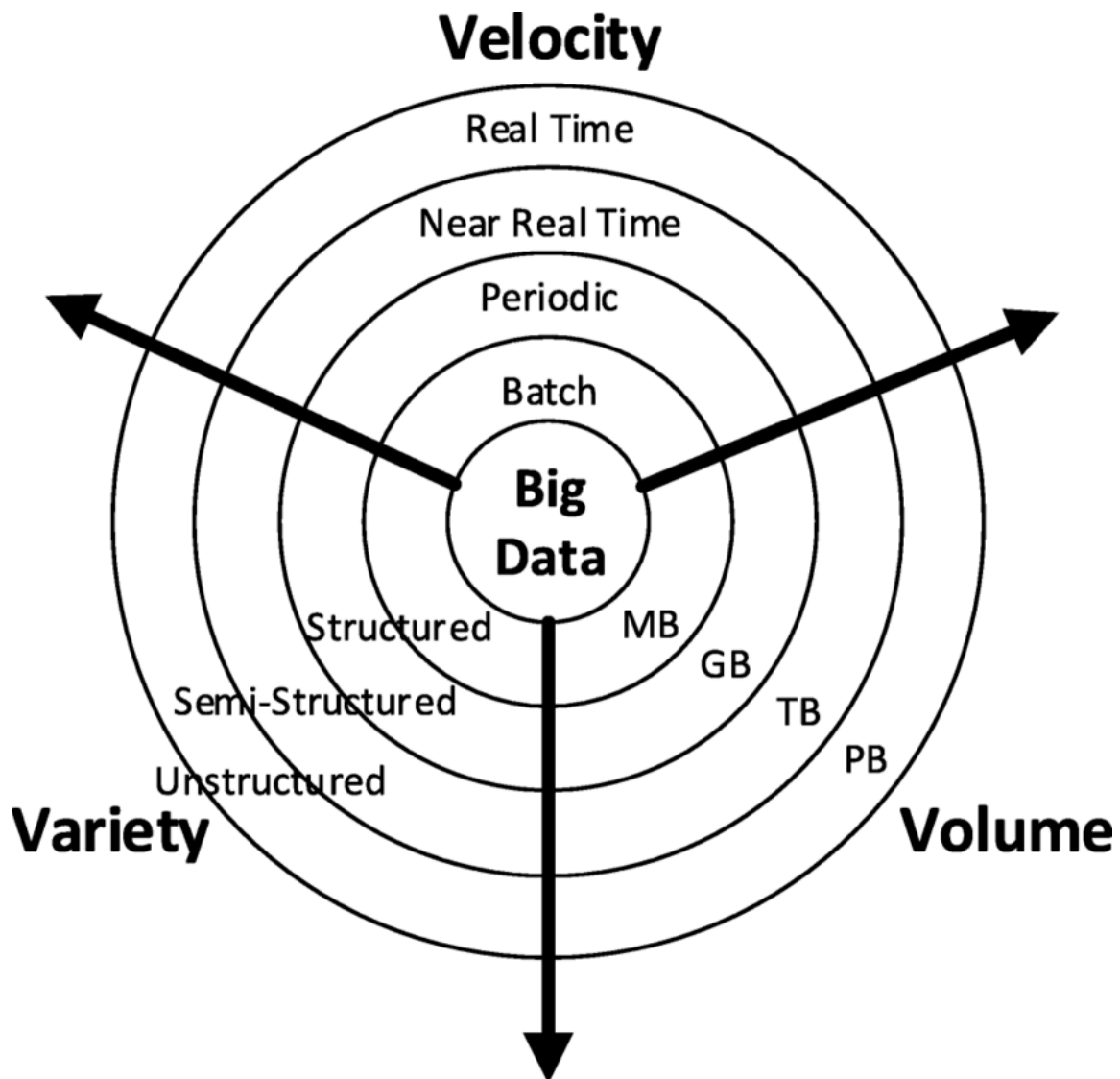


Figura 1.2: Definición de big data en base a “Las tres Vs del big data”. (Fuente: Researchgate)

Otros autores amplían las características que han de tener los datos que forman parte del big data, incluyendo “otras Vs” como lo son:

- 🙌 **Volatilidad**, referida al tiempo durante el cual los datos recogidos son válidos y a durante cuánto tiempo deberán ser almacenados.
- 🙌 **Valor**, referido a la utilidad de los datos obtenidos para extraer conocimiento y tomar decisiones a partir de ellos.
- 🙌 **Validez**, referida a lo precisos que son los datos para el uso que se pretende darles. El uso de datos validados permitirá ahorrar tiempo en etapas como la limpieza y el preprocesamiento de los datos.
- 🙌 **Veracidad**, relacionada con la confiabilidad del origen del cual provienen los datos con los que se trabajará así como la incertidumbre o el ruido que pudiera existir en ellos.
- 🙌 **Variabilidad**, frente a la variedad de estructuras y formatos, hace referencia a la complejidad del conjunto de datos, es decir, al número de variables que contiene. Estas características, unidas a las tres Vs descritas anteriormente, se conocen como las ocho Vs del big data (ver [figura1.3](#)).

The 8V's of Big Data





Figura 1.3: Definición de big data en base a "Las ocho Vs del big data". (Fuente: LinkedIn)


Dado que no existe una definición uniforme para el término big data, muchos autores definen el término en función de aquellas características que consideran más relevantes, por lo que es común encontrar "las cinco Vs del big data", "las siete Vs del big data" o "las diez Vs del big data" según cada autor, apareciendo distintos términos para describir el big data, como también pueden ser **visualización** o **vulnerabilidad**, entre otros.


Son muchas las **soluciones a nivel hardware y software**, que se han propuesto a los problemas derivados del almacenamiento y el procesamiento de big data. A continuación, se describen los fundamentos de tres de ellas, las cuales serán desarrolladas a nivel teórico, tecnológico y práctico en los siguientes capítulos.


5. Almacenes de datos


 Tradicionalmente hablando, cuando nos referimos a almacenes de datos, podemos hablar de las **bases de datos relacionales** son colecciones de datos integrados, almacenados en un soporte secundario no volátil y con redundancia controlada. La definición de los datos y la estructura de la base de datos debe estar basada en un modelo de datos que permita captar las interrelaciones y restricciones existentes en el dominio que se pretende modelizar. A su vez, un **Sistema Gestor de Bases de Datos (SGBD)** se compone de una colección de datos estructurados e interrelacionados (una base de datos) así como de un conjunto de programas para acceder a dichos datos.

 Las bases de datos tradicionales, siguiendo la definición anterior, están basadas generalmente en sistemas relacionales u objeto-relacionales. Para el acceso, procesamiento y recuperación de los datos, se sigue el modelo **Online Transaction Processing (OLTP)**. Una transacción es una interacción completa con un sistema de base de datos, que representa una unidad de trabajo. Así pues, una transacción representa cualquier cambio que se produzca en una base de datos.

 El modelo OLTP, traducido al castellano como **procesamiento de transacciones en línea**, permite gestionar los cambios de la base de datos mediante la inserción, actualización y eliminación de información de la misma a través de transacciones básicas que son procesadas en tiempos muy pequeños.

 Con respecto a la recuperación de información de la base de datos, se utilizan operadores clásicos (concatenación, proyección, selección, agrupamiento...) para realizar consultas básicas y sencillas (realizadas, mayoritariamente, en lenguaje SQL y extensiones del mismo).

 Finalmente, las opciones de visualización de los datos recuperados son limitadas, mostrándose fundamentalmente los resultados de forma tabular y requiriendo un procesamiento adicional y más complejo en caso de querer presentar datos complejos.

 La revolución en la generación, almacenamiento y procesamiento de los datos, así como la irrupción del **big data**, han puesto a prueba el modelo de funcionamiento, rendimiento y escalabilidad de las bases de datos relacionales tradicionales. En la actualidad, se requiere de soluciones integradas que aúnen datos y tecnología para almacenar y procesar grandes cantidades de datos con diferentes estructuras y formatos con el objetivo de facilitar la consulta, el análisis y la toma de decisiones sobre los mismos. En este sentido, la **inteligencia de negocio**, más conocida por el término inglés **business intelligence**, investiga en el diseño y desarrollo de este tipo de soluciones. La inteligencia de negocio puede definirse como *la capacidad de una empresa de estudiar sus acciones y comportamientos pasados para entender dónde ha estado la empresa, determinar la situación actual y predecir o cambiar lo que sucederá en el futuro, utilizando las soluciones tecnológicas más apropiadas para optimizar el proceso de toma de decisiones*.

Estas nuevas soluciones requerirán un modelo de procesamiento diferente a **OLTP**. Esto es así, ya que el objetivo perseguido por la inteligencia de negocio está menos orientado al ámbito transaccional y más enfocado al ámbito analítico. Las nuevas soluciones utilizan el modelo **Online analytical processing (OLAP)**.

La principal diferencia entre OLTP y OLAP estriba en que mientras que el primero es un sistema de procesamiento de transacciones en línea, el segundo es un sistema de **recuperación y análisis** de datos en línea. Por tanto, OLAP complementa a SQL aportando la capacidad de analizar datos desde distintas variables y dimensiones, mejorando el proceso de toma de decisiones. Para ello, OLAP permite realizar cálculos y consolidaciones entre datos de distintas dimensiones, creando modelos que no presentan limitaciones conceptuales ni físicas, presentando y visualizando la información de forma flexible, esto es, en diferentes formatos.

Los sistemas OLAP están basados, generalmente, en sistemas o interfaces multidimensionales que proporcionan facilidades para la transformación de los datos, permitiendo obtener nuevos datos más combinados y agregados que los obtenidos mediante las consultas simples realizadas por OLTP. Al contrario que en OLTP, las unidades de trabajo de OLAP son más complejas que en OLTP y consumen más tiempo.

Finalmente, en cuanto a la visualización de los mismos, los sistemas OLAP permiten la visualización y el análisis multidimensional a partir de diferentes vistas de los datos, presentando los resultados en forma matricial y con mayores posibilidades estéticas y visuales. La tabla 1.1 muestra un resumen con las principales diferencias entre los sistemas **OLTP y OLAP**.

	Bases de datos relacionales(OLTP)	Soluciones Business Intelligence(OLAP)
Concepto	Sistema de procesamiento de transacciones en línea	Sistema de recuperación y análisis de datos en línea
Funciones	Gestión de transacciones: inserción, actualización, eliminación...	Análisis de datos para dar soporte a la toma de decisiones
Procesamiento	Transacciones cortas	Procesamientos de análisis complejos
Tiempo	Las transacciones requieren poco tiempo de ejecución	Los análisis requieren mayor tiempo de ejecución

	Bases de datos relacionales(OLTP)	Soluciones Business Intelligence(OLAP)
Consultas	Simples, utilizando operadores básicos tradicionales	Complejas, permitiendo analizar los datos desde múltiples dimensiones
Visualización	Básica. Muestra los datos en forma tabular	Muestra los datos en forma matricial. Mayores posibilidades gráficas

Tabla 1.1: Tabla resumen y comparativa entre OLTP y OLAP

5.1. Sistemas de ayuda a la decisión

En una empresa u organización, los datos generados a diario son, principalmente, aquellos derivados de las operaciones rutinarias de la empresa. Estos datos, tradicionalmente, se almacenaban en **bases de datos relacionales** y su manipulación se correspondía con **transacciones** realizadas sobre la base de datos. Sin embargo, el objetivo de cualquier organización es seleccionar esos datos para realizar estudios y análisis que permitan generar informes que, a su vez, permitan a la empresa **extraer información para tomar decisiones estratégicas** que conduzcan a la organización al éxito.

El crecimiento exponencial de los datos manejados por una organización ha hecho que los computadores sean las únicas herramientas capaces de procesar estos datos para obtener información y ofrecer ayuda en la toma de decisiones. En este contexto, aparecen los **sistemas de ayuda a la decisión** o *Decision Support Systems (DSS)* que ayudan a quienes ocupan puestos de gestión a tomar decisiones o elegir entre diferentes alternativas.



Sistema de ayuda a la decisión



Sistema de ayuda a la decisión: Conjunto de técnicas y herramientas tecnológicas desarrolladas para procesar y analizar datos para ofrecer soporte en la toma de decisiones a quienes ocupan puestos de gestión o dirección en una organización. Para ello, el sistema combina los recursos de los gestores junto con los recursos computacionales para optimizar el proceso de toma de decisiones.

Mientras que las bases de datos relacionales han sido tradicionalmente el componente del *back-end* en el diseño de sistemas de ayuda a la decisión, los almacenes de datos se han convertido en una opción mucho más competitiva como elemento *back-end* al mejorar el rendimiento de éstas.

Los **campos de aplicación** de los almacenes de datos no se reducen únicamente al ámbito empresarial, sino que cubren multitud de dominios como las **ciencias naturales, demografía, epidemiología o educación, entre otros muchos**. La propiedad común a todos estos campos y que hace de los almacenes de datos una adecuada solución en estos ámbitos **es la necesidad de almacenamiento y herramientas de análisis que permitan obtener en tiempos razonables información y conocimiento útiles para mejorar el proceso de toma de decisiones**.

5.2. Almacenes de datos: Concepto

La aparición de los **almacenes de datos** está ligada, principalmente, a una serie de retos que es necesario abordar para **convertir los datos transaccionales** con los que trabaja una base de datos relacional en **información para generar conocimiento y dar soporte al proceso de toma de decisiones**

- **Accesibilidad:** Desde cualquier dispositivo, a cualquier tipo de usuario y a gran cantidad de información que no puede ser almacenada de forma centralizada. La accesibilidad, en este sentido, debe hacer frente al problema de la **escalabilidad** del sistema y de los datos que este maneja.
- **Integración:** Referente a la gestión de datos heterogéneos, con distintos formatos, y provenientes de distintos ámbitos de la organización. Una correcta integración debe garantizar a su vez la corrección y completitud de los datos integrados.
- **Consultas mejoradas:** Permitiendo incluir operadores avanzados y dar soporte a herramientas y procedimientos que posibiliten obtener el máximo partido de los datos existentes. De este modo, será posible obtener **información precisa para realizar un análisis eficiente**.
- **Representación multidimensional:** Proporciona herramientas para analizar de forma multi-dimensional los datos del sistema, incluyendo datos de **diferentes unidades** de la organización con el objetivo de proporcionar herramientas de **análisis y visualización multi-dimensional** para mejorar el proceso de toma de decisiones.

✓ Almacén de datos (Data Warehouse)

Por tanto, un **almacén de datos**, más conocido por el término **data warehouse** (en inglés), es una solución de **business intelligence** que combina tecnologías y componentes con el objetivo de ayudar al uso estratégico de los datos por parte de una organización. Esta solución debe proveer a la empresa, de forma integrada, de capacidad de almacenamiento de una gran cantidad de datos así como de herramientas de análisis de los mismos que, frente al procesamiento de transacciones, permita transformar los datos en información para ponerla a disposición de la organización y optimizar el proceso de toma de decisiones.

O bien, más resumidamente, según W. Inmon (conocido por ser el “padre” del concepto de **Almacén de datos (Data Warehouse)**): Colección de datos orientados a temas, integrados, variante en el tiempo y no volátil que da soporte al proceso de toma de decisiones de la dirección.

Para entender correctamente esta definición, es necesario ahondar en las características que incluye la misma.

- **Orientados a temas:** Es decir, no orientado a procesos (transacciones), sino a entidades de mayor nivel de abstracción como “artículo” o “pedido”.
- **Integrados:** Almacenados en un formato uniforme y consistente, lo que implica depurar o limpiar los datos para poder integrarlos.
- **Variante en el tiempo:** Asociados a un instante de tiempo (mes, trimestre, año...)
- **No volátiles:** Se trata de datos persistentes que no cambian una vez se incluyen en el almacén de datos.

El diseño y funcionamiento de los almacenes de datos se basa en el sistema de procesamiento analítico en-línea, **OLAP**. Este sistema se encarga del análisis, interpretación y toma de decisiones acerca del negocio, en contraposición a los sistemas de procesamiento de transacciones en línea, **OLTP**.

Así pues, los sistemas **OLTP están dirigidos por la tecnología y orientados a automatizar las operaciones del día a día** de la organización, mientras que los sistemas **OLAP están dirigidos por el negocio y proporcionan herramientas para tomar decisiones a largo plazo**, mejorando la estrategia y la competitividad de la organización. La tabla 1.2 muestra una comparativa entre las principales características de las bases de datos operacionales (OLTP) y los almacenes de datos (OLAP).

Característica	BBDD Operacionales(OLTP)	Almacén Datos(OLAP)
Objetivo	Depende de la aplicación	Toma de decisiones

Característica	BBDD Operacionales(OLTP)	Almacén Datos(OLAP)
Usuarios	Miles	Cientos
Trabajo con...	Transacciones predefinidas	Consultas y análisis específicos
Acceso	Lectura y escritura a cientos de registros	Principalmente lectura. Miles de registros
Datos	Detallados, numéricos y alfanuméricos	Agregados, principalmente numéricos
Integración	En función de la aplicación	Basados en temas, con mayor nivel de abstracción
Calidad	Medida en términos de integridad	Medida en términos de consistencia
Temporalidad Datos	Solo datos actuales	Datos actuales e históricos
Actualizaciones	Continuas	Periódicas
Modelo	Normalizado	Desnormalizado, multidimensional
Optimización	Para acceso OLTP a parte de la BBDD	Para acceso OLAP a gran parte de la BBDD

Tabla 1.2. Diferencias entre BBDD Operacionales y Almacenes de Datos

5.3 Almacenes de datos: Arquitectura

Las arquitecturas disponibles para el **diseño de almacenes de datos** se basan, principalmente, en garantizar que el sistema cumpla una serie de propiedades esenciales para su óptimo funcionamiento

- **Separación:** De los datos transaccionales y los datos estratégicos que sirven como punto de partida a la toma de decisiones.
- **Escalabilidad:** A nivel hardware y software, para actualizarse y garantizar el correcto funcionamiento del sistema a medida que el número de datos y usuarios aumenta.

- **Extensiones:** Permitiendo integrar e incluir nuevas aplicaciones sin necesidad de rediseñar el sistema completo.
- **Seguridad:** Monitorizando el acceso a los datos estratégicos guardados en el almacén de datos.



Almacén de datos



Las arquitecturas de **almacenes de datos** se clasifican, fundamentalmente, en dos tipos: arquitecturas orientadas a la estructura y arquitecturas orientadas a la empresa.

5.3.1 Arquitecturas orientadas a la estructura

Las arquitecturas orientadas a la estructura reciben su nombre debido a que están diseñadas poniendo especial énfasis en el **número de capas y elementos que componen la arquitectura del sistema de almacén de datos**. De acuerdo con este criterio, es posible distinguir las siguientes arquitecturas.

Arquitectura de una capa

El objetivo principal de esta arquitectura, poco utilizada en la práctica, es **minimizar la cantidad de datos almacenados eliminando para ello los datos redundantes**. La [figura 1.4](#) muestra un esquema de este tipo de arquitectura. En ella, el almacén de datos creado es virtual, existiendo un middleware que interpreta los datos operacionales y ofrece una vista multidimensional de ellos.

El principal inconveniente de esta arquitectura es que su simplicidad hace que el sistema **no cumpla la propiedad de separación**, ya que los procesos de análisis se realizan sobre los datos operacionales.

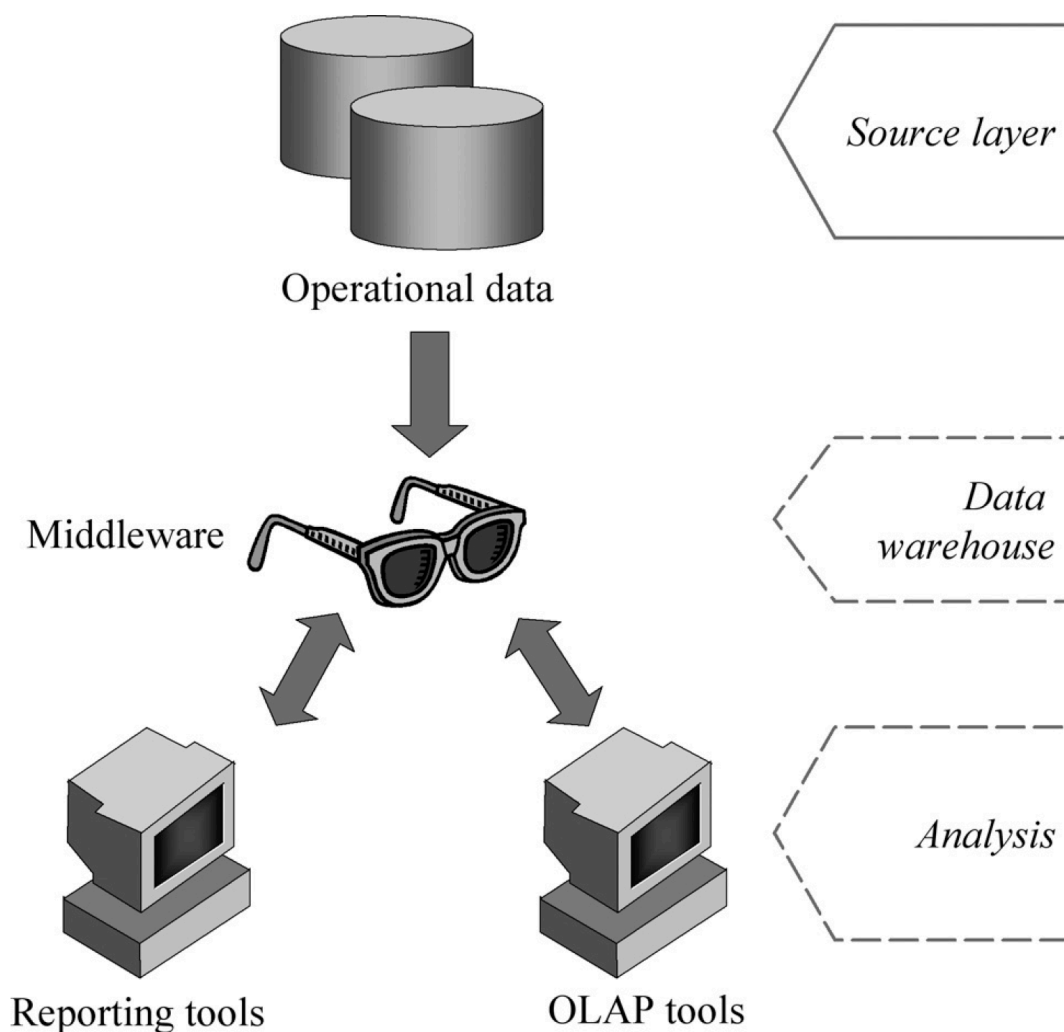


Figura 1.4. Almacén de datos. Arquitectura de una capa. (Fuente: UCLM)

Arquitectura de dos capas

Fue diseñada con el objetivo de solucionar el problema de la separación que presentaba la arquitectura de una capa. Este esquema consigue **subrayar la separación entre los datos disponibles y el almacén de datos** a través de los siguientes componentes (ver [figura1.5](#):

- **Capa de origen (fuente):** Se corresponde con los orígenes y fuentes de los datos heterogéneos que se pretenden incorporar al almacén de datos.
- **Puesta a punto:** Proceso por el cual se utilizan herramientas de Extracción, Transformación y Carga (ETL) para extraer, limpiar, filtrar, validar y cargar datos en el almacén de datos.
- **Capa de almacén de datos:** Almacenamiento centralizado de la información en el almacén de datos, el cual puede ser utilizado para crear *data marts* o repositorios de metadatos.

- **Análisis:** Conjunto de procesos a partir de los cuales los datos son analizados de forma eficiente y flexible, generando informes y simulando escenarios hipotéticos para dar soporte a la toma de decisiones.

Data mart

Data mart es un subconjunto o agregación de los datos almacenados en un almacén de datos primario que incluye información relevante sobre un área específica del negocio.

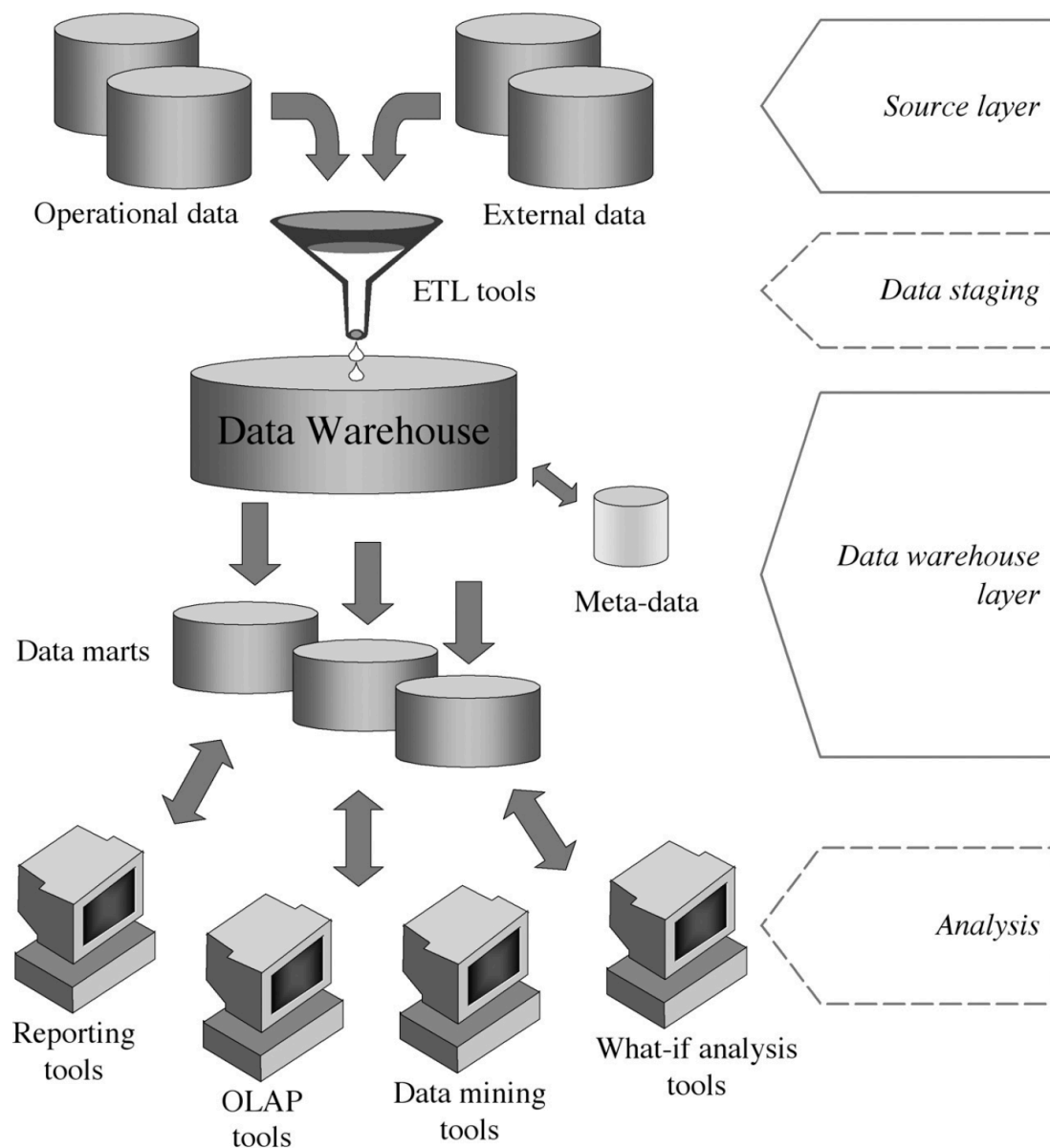


Figura 1.5. Almacén de datos. Arquitectura de dos capas. (Fuente: UCLM)

Arquitectura de tres capas

Este tercer tipo de arquitectura incluye una capa llamada de **datos reconciliados** o almacén de datos operativos. Con esta capa, los datos operativos obtenidos tras la limpieza y depuración son integrados y validados, proporcionando un modelo de datos de referencia para toda la organización.

De este modo, **el almacén de datos no se nutre de los datos de origen directamente, sino de los datos reconciliados generados**, los cuales también son utilizados para realizar de forma más eficiente tareas operativas, como la realización de informes o la alimentación de datos a procesos operativos.

Esta capa de datos reconciliados también puede implementarse de forma virtual en una arquitectura de dos capas, ya que se define como una vista integrada y coherente de los datos de origen. La [figura1.6](#) muestra de forma gráfica este tipo de arquitectura

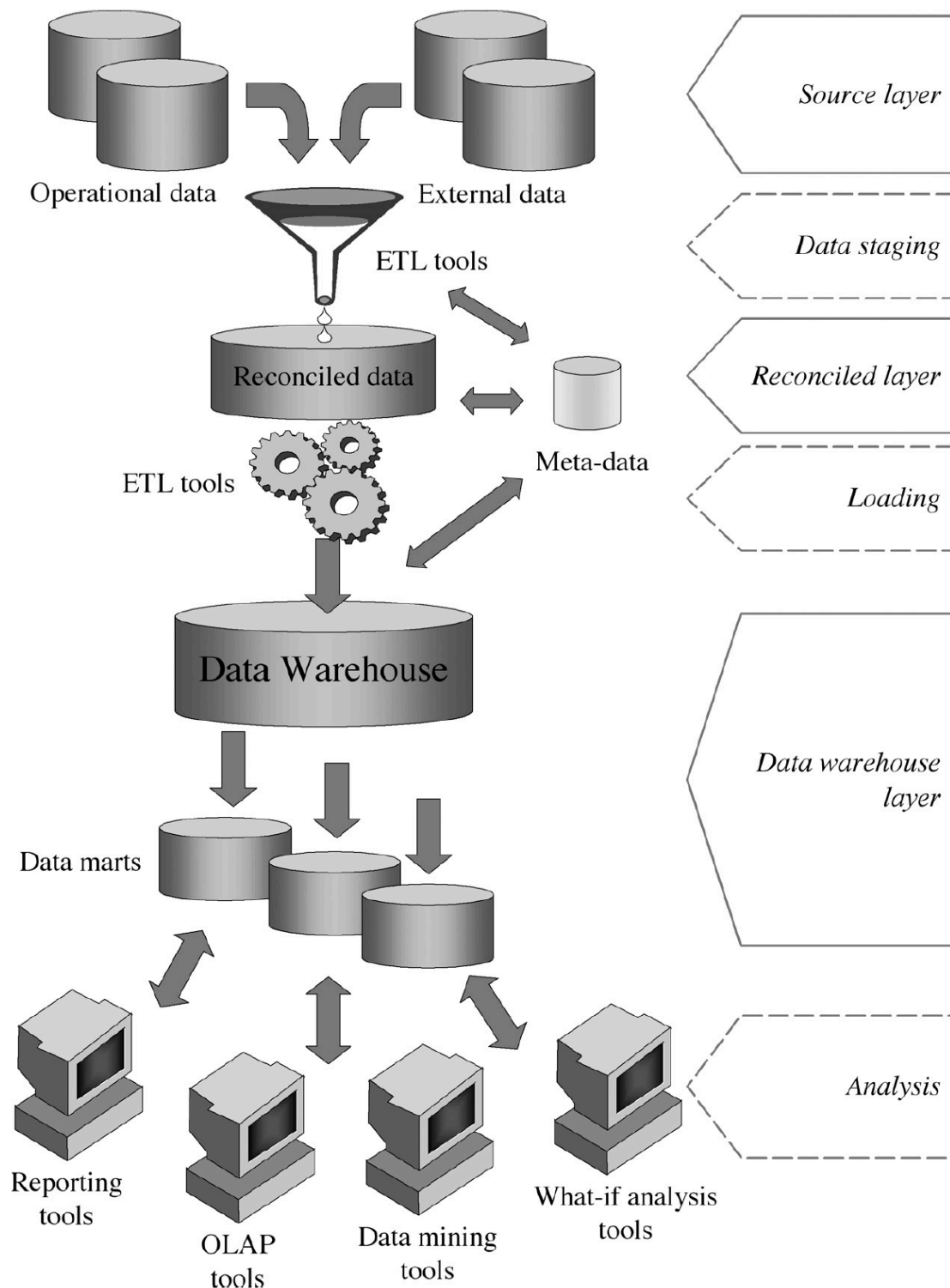


Figura 1.6. Almacén de datos. Arquitectura de tres capas. (Fuente: UCLM)

5.3.2. Arquitecturas orientadas a la empresa

Esta clasificación distingue **cinco tipos de arquitecturas** que combinan las capas mencionadas en la primera clasificación para diseñar almacenes de datos.

1. *Arquitectura de data marts independientes*

Arquitectura preliminar en la que **los distintos data marts son diseñados de forma independiente y contruidos de forma no integrada**. Suele utilizarse en los inicios de implementación de proyectos de almacenes de datos y reemplazada a medida que el proyecto va creciendo.

2. *Arquitectura en bus*

Similar a la anterior, **asegura la integración lógica de los data marts creados**, ofreciendo una visión amplia de los datos de la empresa y permitiendo realizar análisis rigurosos de los procesos que en ella se llevan a cabo.

3. *Arquitectura hub-and-spoke (centro y radio)*

Esta arquitectura es **muy utilizada en almacenes de datos de tamaños medio y grande**. Su diseño pone especial énfasis en garantizar la escalabilidad del sistema y permitir añadir extensiones al mismo.

Para ello, **los datos se almacenan de forma atómica y normalizada en una capa de datos reconciliados que alimenta a los data marts** contruidos que contienen, a su vez, los datos agregados de forma multidimensional. Los usuarios acceden a los data marts, si bien es cierto que también pueden hacer consultas directamente sobre los datos reconciliados.

4. *Arquitectura centralizada*

Se trata de un caso particular de la arquitectura hub-and-spoke. En ella, **la capa de datos reconciliados y los data marts se almacenan en un único repositorio físico**.

5. *Arquitectura federada*

Se trata de un tipo de arquitectura **muy utilizada en entornos dinámicos, cuando se pretende integrar almacenes de datos o data marts existentes con otros para ofrecer un entorno único e integrado de soporte a la toma de decisiones**. De esta forma, cada almacén de datos y cada data mart es integrado virtual o físicamente con lo demás. Para ello, se utilizan una serie de técnicas y herramientas avanzadas como son las ontologías, consultas distribuidas e interoperatividad de metadatos, entre otras.

6. Data Lake

6.1 Concepto

Un data lake o lago de datos es un repositorio centralizado que permite almacenar, compartir, gobernar y descubrir todos los datos estructurados y no estructurados de una organización a cualquier escala. Es el lugar en el que se vuelcan los datos en bruto.

Los data lakes no requieren un esquema predefinido, se pueden almacenar y procesar datos sin esquema y en cualquier formato sin la necesidad de conocer cómo se van a explotar en el futuro. Esta característica evita que sean necesarios complejos procesos **ETL (Extracción, Transformación y Carga)** (que explicaremos en el próximo punto) de limpieza y preparación.

Entre las características más importantes de los data lakes se encuentra su flexibilidad en almacenar diferentes tipos de datos, que proporciona la agilidad necesaria para los procesos de ingesta. También es muy importante que proporcione suficiente trazabilidad, y de esta manera poder determinar los cambios que han sufrido los datos en los procesos de transformación o ingesta.

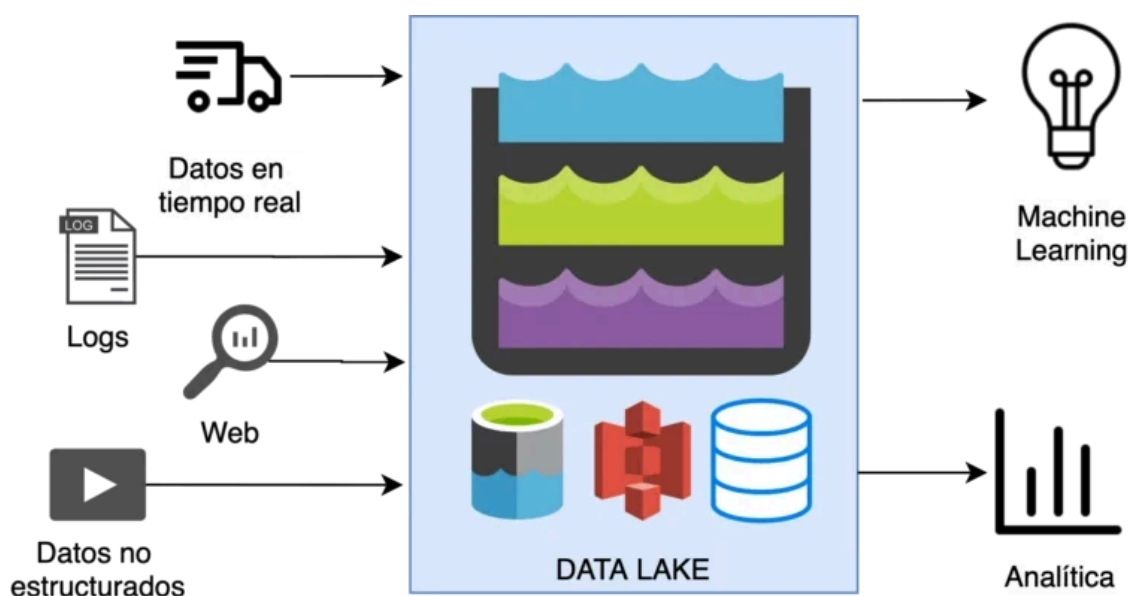


Figura 1.7. Data Lake. (Fuente: AprenderBigData.com)

6.2 Data Lake vs Data Warehouse

Los *data lake* y los *data warehouse* se utilizan de forma generalizada para el **almacenamiento de big data**, pero, aunque ambos son almacenes de datos, estos **no son términos intercambiables**. Un data lake o "lago de datos" es un gran conjunto de datos **en bruto**, que todavía **no tiene una finalidad definida**. En cambio, un data warehouse o "almacén de datos" es un depósito de datos que **ya están estructurados y filtrados** y han sido procesados para un **propósito concreto**.

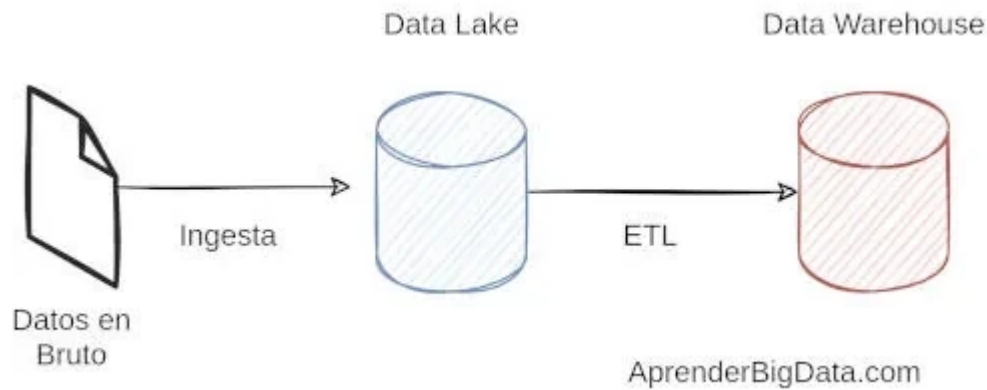


Figura 1.8. Data Warehouse vs Data Lake 1

A menudo se confunden estos dos tipos de almacenamiento de datos, pero son mucho más diferentes de lo que puede parecer a simple vista. De hecho, lo único que tienen en común es que contienen grandes cantidades de datos. Es importante realizar la distinción, ya que **los data lake y los data warehouse atienden a diferentes propósitos**, por lo que requieren un enfoque diferente para ser optimizados adecuadamente.

Así, un **data lake almacena datos sin procesar** y que todavía **no tienen una finalidad determinada**. Sus usuarios finales son los **científicos de datos** y su **accesibilidad es elevada**. Además, en un data lake, justamente por esta fácil accesibilidad, se pueden actualizar los datos rápidamente.

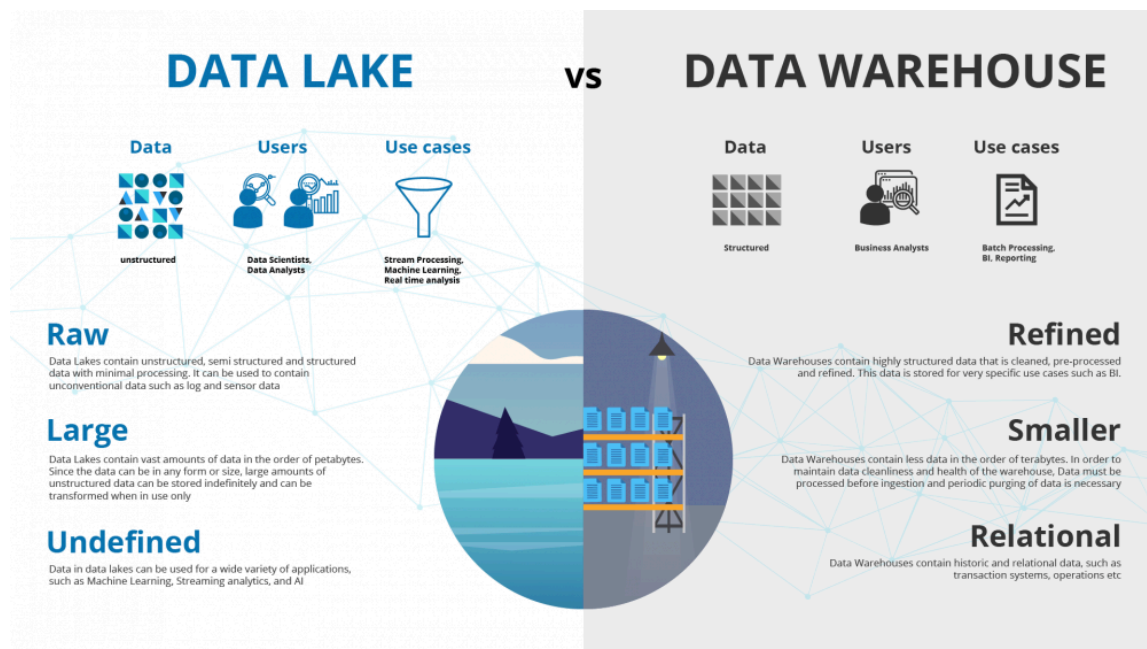


Figura 1.9. Data Warehouse vs Data Lake 2. (Fuente: Huawei)

Por su lado, un **data warehouse cuenta con datos procesados** y que ya se están usando, por lo que tienen una **finalidad concreta**.