

Presentación del manual

La cualificación profesional es el “conjunto de competencias profesionales con significación en el empleo que pueden ser adquiridas mediante formación modular u otros tipos de formación, así como a través de la experiencia laboral” (Ley 5/2002 de las Cualificaciones y de la Formación Profesional).

Cada cualificación se organiza en unidades de competencia, siendo la unidad de competencia el agregado mínimo de competencias profesionales, susceptible de reconocimiento y acreditación parcial.

Así mismo, cada unidad de competencia lleva asociado un módulo formativo, donde se describe la formación necesaria para adquirir esa unidad de competencia.

Siguiendo esta secuencia, el presente manual “Desarrollo de componente software y consultas dentro del sistema de almacén de datos”, está basado en los contenidos del Módulo formativo: “Creación y mantenimiento de componentes software en sistemas de planificación de recursos empresariales y de gestión de relaciones con clientes” asociado a la Unidad de Competencia: “Realizar y mantener componentes software en un sistema de planificación de recursos empresariales y de gestión de relaciones con clientes” según el Real Decreto correspondiente.

MÓDULO FORMATIVO: Creación y mantenimiento de componentes software en sistemas de planificación de recursos empresariales y de gestión de relaciones con clientes

Nivel: 3

Código: MF1215_3

Asociado a la UC: UC1215_3

Horas: 210

UNIDAD FORMATIVA: Desarrollo de componente software y consultas dentro del sistema de almacén de datos

Nivel: 3

Código: UF1890

Asociado a la UC: UC1215_3

Horas: 30

UF1890
DESARROLLO DE
COMPONENTES
SOFTWARE Y
CONSULTAS DENTRO
DEL SISTEMA DE
ALMACÉN DE DATOS

ÍNDICE

UNIDAD DIDÁCTICA 1. CARGA DE DATOS.....	9
1. Exploración del sistema de almacén de datos Estructuras de información, cubos y multicubos11	
1.1. Identificación de tipos de estructuras de información y sus relaciones para almacenar información.....	14
2. Procesos de carga de datos al sistema de almacén de datos	25
2.1. Identificación de orígenes de datos para la carga de datos.....	31
2.2. Creación de componentes de software para extraer información de un sistema de almacén de datos.....	32
RECUERDA.....	37
Preguntas de Autoevaluación.....	39
UNIDAD DIDÁCTICA 2. EXTRACCIÓN DE DATOS (DATA WAREHOUSE)	41
1. Herramientas para la carga y extracción de datos de sistemas de almacén de datos.....	43
1.1. Mecanismos que se utilizan para la extracción de datos.....	44
1.2. Estructuración de la información para adecuarse a las necesidades de la empresa.....	48
2. Creación de extractores de datos	51
2.1. Obtención de información de fuentes internas o externas	53
2.2. Agrupación, transformación y homogeneización de la información para su posterior estudio	57
RECUERDA.....	63
Preguntas de Autoevaluación.....	65
UNIDAD DIDÁCTICA 3. HERRAMIENTAS DE OBTENCIÓN DE INFORMACIÓN.....	67
1. Herramientas de visualización y difusión	69
1.1. Aplicaciones de visualización genéricos	73
1.2. Wizards, librerías, API	77
1.3. Herramientas de visualización geoespacial.....	82
1.4. Herramientas de visualización de datos temporales	84
RECUERDA.....	87
Preguntas de Autoevaluación.....	89
ACTIVIDADES PRÁCTICAS	91
Actividad Práctica RP1	93
RESPUESTAS A LAS PREGUNTAS DE AUTOEVALUACIÓN	95

UD1 Carga de datos



UF1890 Desarrollo de componentes de software y consultas dentro del sistema de almacén de datos

1. Exploración del sistema de almacén de datos Estructuras de información, cubos y multicubos

Los almacenes de datos son también conocidos como **Data Warehouse**.

Son una colección de datos con las siguientes características:

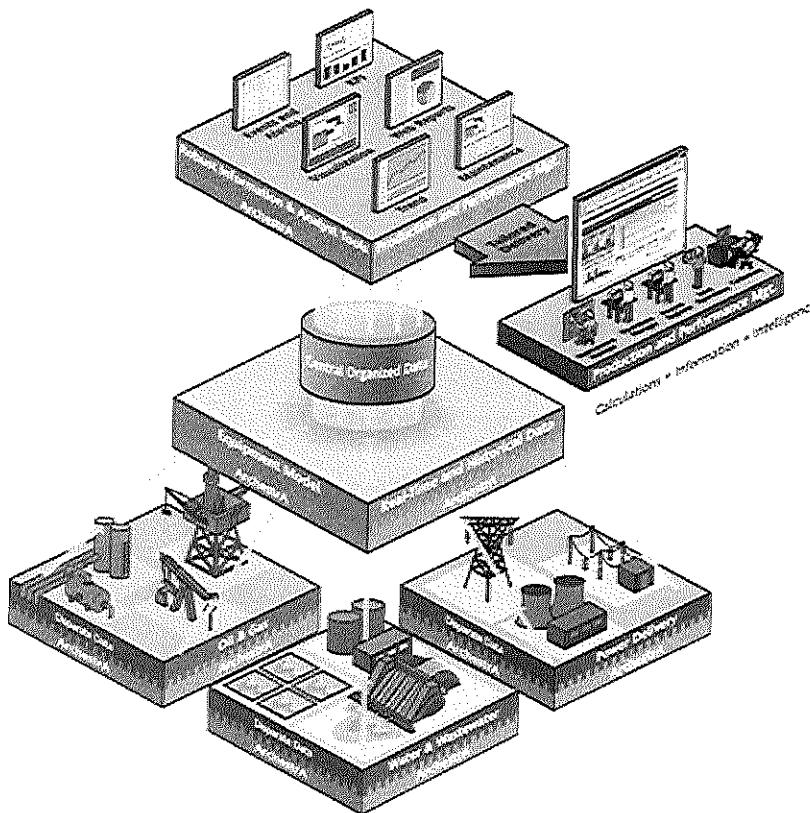
- **Organizado en torno a temas:** la información se clasifica en base a los aspectos que son de interés para la empresa.
- **Integrado:** es el aspecto más importante. La integración de datos consiste en convenciones de nombres, codificaciones consistentes, medida uniforme de variables, etc.
- **Dependiente del tiempo:** esta dependencia aparece de tres formas:
 - La información representa los datos sobre un horizonte largo de tiempo.
 - Cada estructura clave contiene (implícita o explícitamente) un elemento de tiempo (día, semana, mes, etc.).
 - La información, una vez registrada correctamente, no puede ser actualizada.
- **No volátil:** el Almacén de Datos sólo permite cargar nuevos datos y acceder a los ya almacenados, pero no permite ni borrar ni modificar los datos.

Los almacenes de datos suponen una ayuda en la toma de decisión por parte de la empresa o la organización. Este tipo de almacenes son sobre todo un expediente de una empresa que va más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos.

La función principal que se le otorga a un almacén de debe entregar la información correcta a la gente indicada en el momento oportuno en el formato deseado. El almacén de datos ofrece respuesta a las diferentes necesidades del usuario, utilizando para ello sistemas de ayuda en la decisión (DSS), sistemas de información Ejecutiva (EIS) o bien herramientas para realizar consulta o informes.

Los usuarios finales fácilmente pueden hacer consultas sobre sus almacenes de datos sin tocar o afectar la operación del sistema.

Estructura



Los cubos de información también son conocidos como **DataMart**, siguen una lógica de los datos en bruto, de los datos provistos por su sistema de operaciones/finanzas hacia el almacén de datos con la adición de nuevas dimensiones o información calculada.

Se les denominada DataMart porque son la representación de un conjunto de datos relacionados con un tema en particular como por ejemplo:

- Ventas.
- Operaciones.
- Recursos humanos.

Todo ello a disposición de los clientes a quienes les pueda interesar.

Esta información puede ser organizada en tablas mediante el uso de tablas dinámicas de MS-Excel o programas personalizados. Las **tablas dinámicas**

permiten manipular las vistas de la información con relativa facilidad. Los cubos de información se producen con bastante rapidez. A ellos se les aplican las reglas de seguridad de accesos necesarios.

La información se puede clasificar en:

- Estática.
- Dinámica.

El análisis está basado en las dimensiones y por lo tanto se denomina **análisis multidimensional**.

Por tanto relacionando todo esto con el almacén de memoria, se deduce que un data warehouse es una colección de datos que está formado por dimensiones y variables, entendiendo como dimensiones a aquellos elementos que participan en el análisis y variables a los valores que se desean analizar.

Generalmente, las variables son representadas por valores detallados y numéricos para cada instancia del objeto o evento medido.

Por el contrario, las dimensiones son atributos relativos a las variables, y son utilizadas para ordenar, agrupar o abreviar los valores de las mismas. Las dimensiones poseen una granularidad menor y toman como valores un conjunto de elementos menor que el de las variables.

Las **dimensiones** son atributos relativos a las variables, son las perspectivas de análisis de las variables. Forman parte de las tablas de dimensiones.

Las **variables** son conocidas como indicadores de gestión; son datos que están siendo analizados. Forman parte de la tabla de hecho. Más formalmente, las variables representan algún aspecto cuantificable o medible de los objetos o eventos a analizar.

La **estructura lógica** de un almacén de Datos está compuesta por los siguientes niveles:

Metadatos: describen la estructura de los datos contenidos en el almacén. Están en una dimensión distinta al resto de niveles.

Datos detallados actuales: Obtenidos directamente del procesado de los datos. Forman el nivel más bajo de detalle. Ocupan mucho espacio. Se almacenan en disco, para facilitar el acceso.

Datos detallados históricos: igual que los anteriores, pero con datos correspondientes al pasado. Se suelen almacenar en un medio externo, ya que su acceso es poco frecuente.

Datos ligeramente resumidos: primer nivel de agregación de los datos detallados actuales. Corresponden a consultas habituales. Se almacenan en disco.

Datos muy resumidos: son el nivel más alto de agregación. Corresponden a consultas que se realizan muy a menudo y que se deben obtener muy rápidamente. Suelen estar separados del Almacén de datos, formando Supermercados de Datos (Data Marts).

La **estructura física** puede presentar cualquiera de las siguientes configuraciones:

Arquitectura centralizada: todo el almacén de datos se encuentra en un único servidor.

Arquitectura distribuida: los datos del almacén se reparten entre varios servidores. Asignando cada servidor a uno o varios temas lógicos.

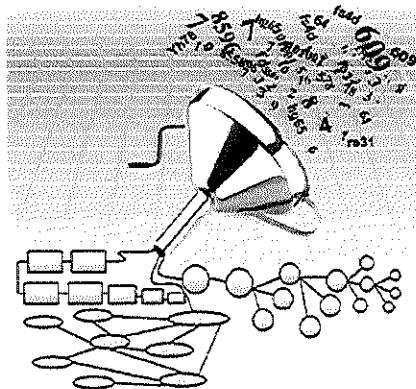
Arquitectura distribuida por niveles: refleja la estructura lógica del Almacén, asignando los servidores en función del nivel de agregación de los datos que contienen. Un servidor está dedicado para los datos de detalle, otro para los resumidos y otro para los muy resumidos.

Cuando los datos muy resumidos se duplican en varios servidores para agilizar el acceso se habla de Supermercados de datos (Data Marts).

1.1. Identificación de tipos de estructuras de información y sus relaciones para almacenar información

La información está constituida por mensajes, es decir, es un conjunto de datos que representan ideas mediante las cuales se incrementa nuestra conciencia, inteligencia o conocimiento. Los mensajes pueden adoptar diferentes

manifestaciones físicas. En líneas generales se definen los mensajes como manifestaciones físicas de la información.



Se pueden definir dos tipos generales de información:

- **Continua:** se caracteriza porque sus datos pueden adoptar un número infinito de valores.
- **Discreta:** se caracteriza porque sus datos pueden adoptar sólo un número finito de valores.

Una necesidad básica de la vida moderna es el intercambio de información. La satisfacción de esta necesidad se obtiene con el transporte o envío de la información.

El problema reside en la transmisión de información, el cual se ha resuelto gracias al empleo de sistemas eléctricos de comunicaciones, que representan grandes ventajas sobre otros posibles.

Los sistemas eléctricos de comunicación son aquellos que utilizan dispositivos eléctricos, electromagnéticos u ópticos, o la combinación de éstos, para transmitir información desde donde se produce hasta donde se utiliza.

Para que la información se pueda transmitir a través de un sistema eléctrico de comunicación se debe convertir de su forma física original a energía eléctrica. En esta forma de energía la información se conoce como señal, y el proceso de conversión recibe el nombre de transducción.

Los sistemas de transmisión de datos constituyen un apoyo para los sistemas de cómputo para el transporte de la información que manejan. Sin estos sistemas no se podría haber desarrollado las redes avanzadas de cómputo de procesamiento distribuido, en las que se comparte y transfiere la información de

datos entre ordenadores que se encuentran separados a más de 20 metros, incluso en zonas geográficas diferentes.

Las redes de transmisión de datos e información pueden ser sencillas, como puede ser el caso de un ordenador y la conexión con sus periféricos, o pueden ser mucho más complejas pasando por la conexión de punto a punto de larga distancia que se satisface con la utilización de módems, o redes ligeramente más complejas que conectan varias terminales de cómputo de edificios lejanos con el ordenador principal de un centro especializado de datos.

La comunicación de datos presupone mayores requisitos en su red básica que el servicio de señal analógica o de voz para conseguir la transferencia correcta de datos.

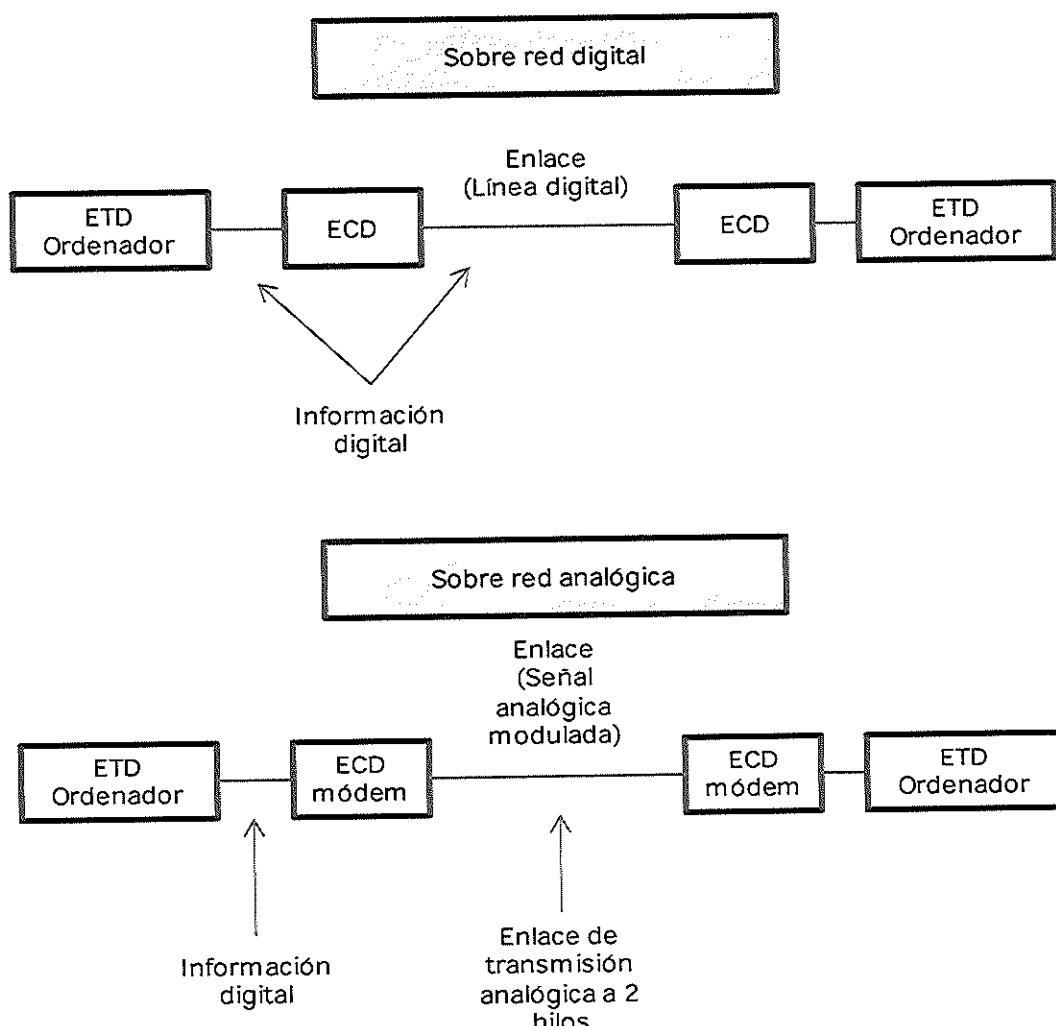
Los ordenadores son considerados inteligentes, pero pese a ello no son seres humanos capaces de realizar juicios, entablar una comunicación organizada con la información apropiada y ordenadamente para representar una sesión coherente y con significado.

La señal de los datos es bastante estricta con el contenido de su información, de tal modo que cualquier error, por pequeño que éste sea puede ser desastroso para la comunicación, es decir, la transmisión de datos debe ser más confiable.

De este modo, dos de las principales medidas adicionales que se deben tomar para satisfacer los mayores requisitos de la transmisión de datos son el control adecuado del flujo de datos durante la transmisión y la codificación para la detección y corrección de errores.

Estos aspectos están incluidos en lo que se conoce como protocolo de comunicación, cuya función es asegurar la comunicación correcta, completa y entendible para los ordenadores.

Las condiciones existentes en la actualidad de las redes de datos, implican la existencia de redes analógicas trabajando al lado de redes digitales, la forma más sencilla de la comunicación de datos entre dos ordenadores se pueden basar en alguna de las siguientes configuraciones que muestra la imagen.



Los ordenadores de ambos extremos se conocen como equipo terminal de datos (ETD) y los equipos de transmisión como equipo terminal de circuito de datos (ECD).

En el caso de la red analógica la transmisión sobre el enlace es digital en tanto que la información con estructura digital del ordenador se debe convertir a la forma apropiada para su transmisión a través de una red analógica. El ECD analógico en este sistema se encarga de esta conversión, es el famoso módem que transmite los datos binarios digitales imponiéndolos sobre una señal portadora de audiofrecuencia.

En el caso de la red digital, el ECD es digital, y por supuesto, con funciones diferentes a las del ECD analógico:

- **En transmisión:** regenerar y convertir la señal del ETD a un formato, nivel y código de línea apropiados para su transmisión sobre la línea digital.

- **En recepción:** establecer el voltaje de referencia que emplea el ETD y reconvertir la señal de línea a la forma apropiada para su aplicación al ETD.

Los ECD digitales pueden suministrar diferentes velocidades digitales de bits que van de 2.4kbits/s, pasando por el canal estándar de 64kbits/s, sistemas de orden superior como 1.544 Mbit/s, 2.048Mbit/s o 45Mbit/s.

El ECD también es conocido como:

- Unidad terminal de red.
- Unidad de servicio de canal.
- Unidad de servicio de datos.
- Unidad terminal de línea.

Otra diferencia importante entre los ECD digitales y analógicos consiste en que, como en la red analógica la tasa de bits de la señal que se recibe no se establece con precisión, en el caso de transmisión analógica el ECD no puede confiar en la red para obtener información precisa de reloj.

Es por esta razón que se necesita un reloj interno del ECD para poder mantener la tasa de bits de transmisión precisa.

En el caso de una red de señal digital, la señal de reloj se recibe desde la red y se deriva de un reloj maestro altamente preciso que funciona como patrón para toda la red de la empresa pública de telecomunicaciones (EPT).

Una **estructura de datos**, o un tipo de datos estructurado, es un tipo de dato construido a partir de otros. Un dato de tipo estructurado está compuesto por una serie de datos de tipos elementales y alguna relación existente entre ellos. Normalmente, la relación suele ser de orden aunque puede ser de cualquier otro tipo.

Se dice que una estructura de datos que es homogénea cuando todos los datos elementales que la forman son del mismo tipo. En caso contrario, se dice que la estructura es heterogénea. Por ejemplo, el tipo de datos complejo es una estructura homogénea, tanto la parte real como la imaginaria se representan con datos reales.

Siempre que se utilice un dato en un programa debe estar determinado su tipo, para que el traductor sepa como debe tratarlo y almacenarlo. En el caso de datos de tipos elementales, el tipo de dato determina el espacio que se utiliza en memoria. Esto, puede no ocurrir si el dato es de un tipo estructurado.

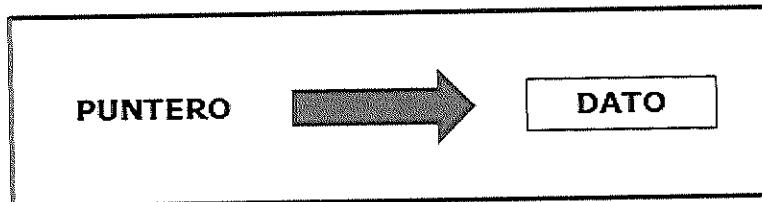
Una estructura de datos que siempre ocupa el mismo espacio en memoria, se dice que es **estática**. Por el contrario, si la memoria asignada a una determinada estructura de datos va variando durante la ejecución del programa, es decir, se realiza una asignación dinámica de memoria, se dice que es una estructura de datos **dinámica**.

1.1.1. Estructura de datos estática

Dentro de los tipos de estructuras de datos que ocupan siempre el mismo espacio en memoria tenemos los punteros, las cadenas y los arrays.

Punteros

Un puntero es un dato que indica la posición de otro dato. Su utilidad se pone de manifiesto en la construcción de estructuras de datos, ya que son ellos los que proporcionan los lazos de unión entre los elementos que constituyen las estructuras. Son importantes los punteros al principio y al final de la estructura.



Si, ocasionalmente, se necesita que un puntero no señale a ningún dato, se dice que el puntero tiene un valor nulo (puntero nulo).

Cadenas

Una cadena es una secuencia de caracteres que se interpretan como un dato único. Las cadenas pueden tener longitud fija o variable. La longitud de la cadena se indica tanto por el número de caracteres que contiene ésta, indicado al principio de la misma, como por un carácter especial denominado fin-de-cadena. Sobre datos de tipo cadena se pueden realizar las siguientes operaciones:

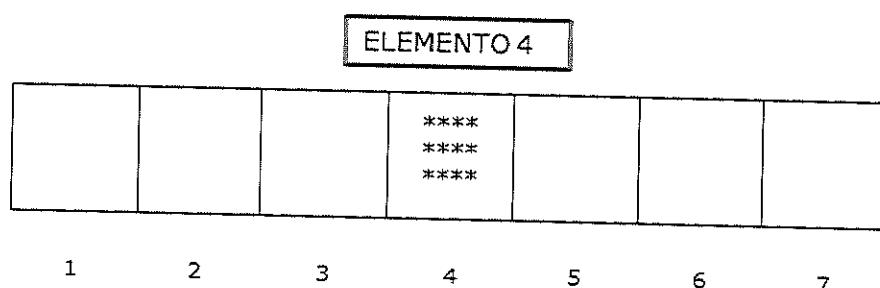
- **Concatenación:** consiste en formar una cadena a partir de dos ya existentes, juntando los caracteres de ambas.
- **Extracción de subcadena:** permite formar una cadena a partir de otra ya existente. La subcadena se forma tomando un tramo consecutivo de la cadena inicial.
- **Comparación de cadenas:** es posible comparar dos cadenas. Se considera menor aquella en que el primer carácter en que difieren ambas es menor.
- **Obtención de la longitud:** la longitud de una cadena es un dato de tipo entero, cuyo valor es el número de caracteres que contiene ésta.

Arrays

El *array* (también llamado formación o matriz), es la estructura de datos más usual. Existe en todos los lenguajes de programación y en algunos es de las pocas estructuras de datos existentes (BASIC y FORTRAN).

Un array es una estructura de datos formada por una cantidad fija de datos del mismo tipo, cada uno de los cuales tiene asociado uno, o más índices, que determinan de forma única la posición del dato en el array.

Podemos imaginar un array como una estructura de celdas donde se pueden almacenar valores. En la figura podemos ver una matriz de un sólo índice que toma valores de 1 a 7.



En el array de la figura siguiente utilizamos dos índices con valores entre 1 y 3, el primero, y entre 1 y 5, el segundo. Cada elemento de esta matriz está representado por un par ordenado de números, el valor de los dos índices.

ELEMENTO 1,4				
1	2	3	4	5

2				
3				

En general, al número de índices del array se le denomina número de dimensiones del array. La dimensión de la formación está dada por los valores máximos de los índices y, el número total de elementos es el producto de estos valores máximos. En los dos ejemplos anteriores, el número de dimensiones de los arrays son 1 y 2, las dimensiones son (7) y (3; 5) y los números totales de elementos son 7 y 15, respectivamente.

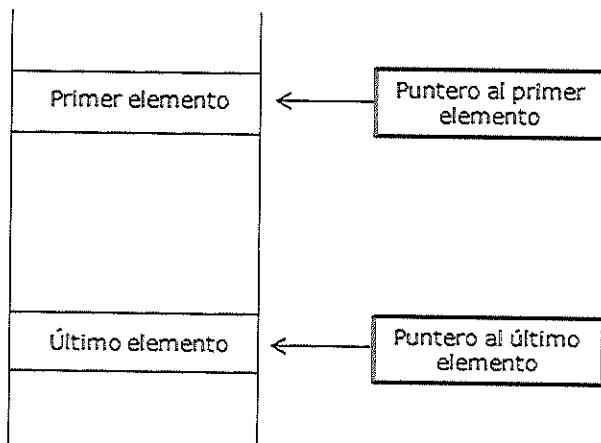
La principal operación que se puede realizar con los arrays es la selección, que consiste en especificar un elemento determinado del array. Esta operación se efectúa dando un valor para todos y cada uno de los índices del array. Con el elemento seleccionado se pueden realizar las operaciones propias de su tipo. Así, con cada elemento de un array real, una vez seleccionado, se pueden realizar las operaciones definidas para datos de tipo real (operaciones aritméticas).

1.1.2. Estructura de datos dinámica

Como ya se ha dicho, este tipo de estructuras ocupan un espacio en memoria que va evolucionando según el tamaño que dicha estructura vaya adquiriendo. Las estructuras de dinámicas que vamos a estudiar son: colas, pilas, listas encadenadas y árboles.

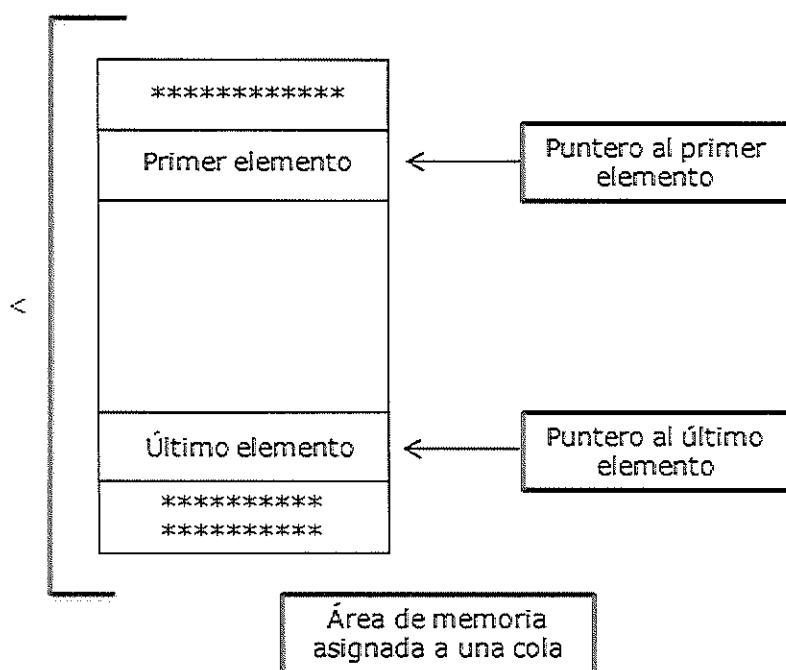
Colas (FIFO)

Una cola es una estructura de datos en la que el primer dato en entrar es el primer dato en salir. Es decir, es una estructura FIFO (*First In First Out*). Todo el mundo conocemos como funciona una cola, los nuevos se ponen al final, los servicios se prestan al principio y no está permitido "colarse". Las mismas reglas se aplican a las colas de datos almacenadas en la memoria de un computador.



Hay varias formas de implementar una cola en la memoria de un computador. Una forma simple consiste en almacenar los datos en posiciones de memoria adyacentes y utilizar punteros para el principio y el fin de la cola. Cuando un elemento se añade a la cola, el puntero de la parte posterior se ajusta para que señale al nuevo elemento. De manera similar, cuando un elemento se elimina de la cola, se ajusta el puntero delantero para que señale al nuevo primer elemento.

El problema de este método para implementar las colas es que las posiciones de memoria que ocupan, varían a medida que se añaden y eliminan elementos de la misma. La solución habitual consiste en asignar un área fija para almacenar la cola y permitir que se mueva en este área de manera circular. Un área de almacenamiento de esta forma se denomina buffer circular, y puede apreciarse en la figura siguiente.



Entre las aplicaciones que tienen las colas se encuentran el almacenamiento de datos en camino, entre un procesador y un periférico, o actuar como punto intermedio en las redes de comunicación de datos.

Pilas (LIFO)

Una pila es una colección ordenada de datos a los que sólo se puede acceder por un extremo, denominado tope o cima de la pila. La pila es una estructura en la que el último elemento en entrar será el primero en salir, es decir, es lo que se denomina estructura LIFO (*Last In First Out*).

Podemos comparar esta estructura con una pila de platos colocada sobre un muelle. Cuando se añade un nuevo plato en lo alto de la pila, los demás bajan, cuando se retira un plato de la pila, los demás suben.

Igual que en el caso de las colas, van a existir dos punteros, uno que indica la posición tope de la pila, denominado puntero de pila, y otro que señala su base, denominado base de pila, y que mantiene el mismo valor mientras existe la pila. Cuando la pila está vacía el puntero de pila tiene el mismo valor que la base de pila.

La pila es una de las estructuras más importantes en computación. Se usa en cálculos, para pasar de un lenguaje de computador a otro y, para transferir el control de una parte del programa a otra.

Las operaciones que se pueden realizar tanto con las colas como con las pilas, son las siguientes:

- **Añadir o eliminar un elemento:** si es una cola podremos añadirla o eliminarla al final de la misma, y si es una pila al principio.
- **Acceder al primer elemento:** normalmente es el único al que se va a poder acceder directamente.
- **Acceder al elemento siguiente del último procesado:** este es el mecanismo normal de acceso tanto a colas como a pilas.
- **Saber si está vacía:** están vacías si no contienen ningún elemento.

Listas encadenadas

Una lista es un conjunto ordenado de datos. Los elementos de la lista pueden insertarse o eliminarse en cualquier punto de la misma, por lo que es menos restrictiva que una pila o una cola.

La forma más sencilla de implementar una lista es hacer uso de un puntero que señale desde un dato al siguiente. También hay un puntero que señala al primer elemento de la lista, mientras que para el último se emplea un puntero nulo.

Una estructura de este tipo se denomina lista encadenada. Cada elemento de la lista consiste en una parte de datos y un puntero.

Una variación sobre la idea de una lista es el caso en el que el puntero del final de la lista señale al primer elemento. Esto crea lo que se denomina una lista circular.

Si los elementos de la lista están en orden alfabético o numérico, dichas listas se conocen como listas ordenadas.

Árboles

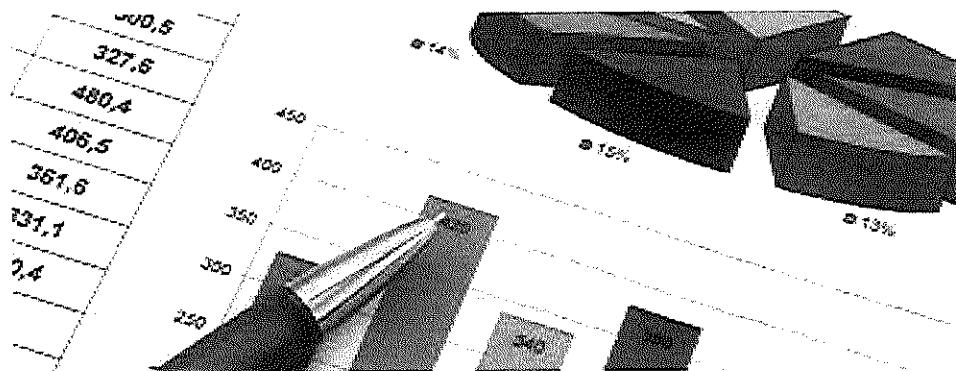
Un árbol es una estructura que implica una jerarquía, en la que cada elemento está unido a otros bajo él. Cada dato en un árbol es un nodo de dicho árbol. El nodo más alto se denomina **raíz**. Cada nodo puede estar conectado a uno o más subárboles, que también responden a la estructura de un árbol. Un nodo, en la parte inferior, del que no cuelga ningún subárbol se denomina nodo terminal u hoja.

Un tipo especial de árboles muy usados en computación son los árboles binarios. En ellos, de cada nodo pueden colgar, a lo más, dos subárboles, denominados subárbol derecho y subárbol izquierdo, que también son árboles binarios.

La forma usual de representar los árboles supone el uso de punteros. En un árbol binario cada nodo está constituido por una parte de datos y dos punteros. Uno, o ambos punteros, pueden tener un valor nulo si del nodo no cuelgan subárboles.

Son muy utilizados en informática. Las partes de muchos programas se enlazan como si se tratara de árboles. Los árboles se utilizan para representar operaciones aritméticas, y en búsquedas y ordenaciones.

2. Procesos de carga de datos al sistema de almacén de datos



Los principales componentes de un sistema de almacén de datos son los siguientes:

- **Sistema ETL (Extraction, Transformation, Load):** realiza las funciones de extracción de las fuentes de datos, transformación y carga del AD, realizando:
 - Extracción de los datos.
 - Filtrado de los datos: limpieza, consolidación, etc.
 - Carga inicial del almacén: ordenación, agregaciones, etc.

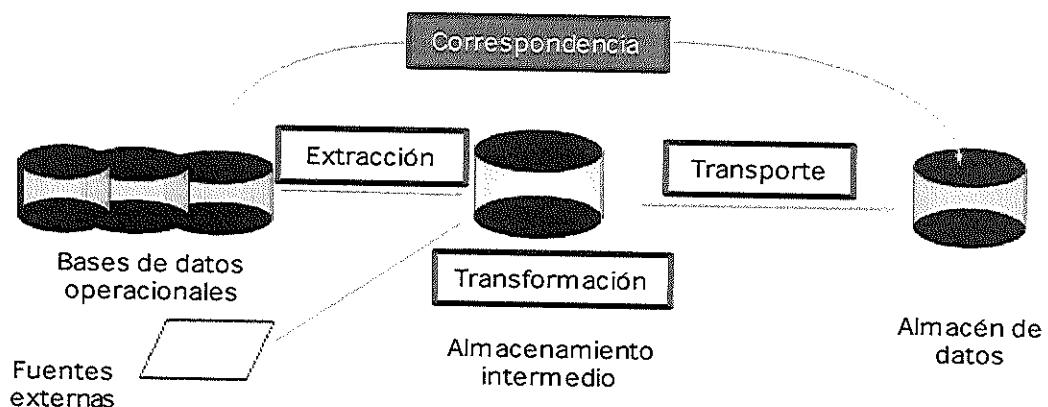
- Refresco del almacén: operación periódica que propaga los cambios de las fuentes externas al almacén de datos.
- **Repositorio propio de datos:** información relevante, metadatos.
- **Interfaces y gestores de consulta:** permiten acceder a los datos y sobre ello se conectan herramientas más sofisticadas, por ejemplo OLAP.
- **Sistemas de integridad y seguridad:** se encargan del mantenimiento global, y copias de seguridad.

El sistema ETL en español se conoce con las siglas ETT (extracción, Transformación y transporte). Este es el sistema encargado del mantenimiento y almacén de datos, por lo que:

- La construcción del sistema E.T.T. es responsabilidad del equipo de desarrollo del almacén de datos.
- El sistema E.T.T. es construido específicamente para cada almacén de datos. Aproximadamente el 50% del esfuerzo.
- En la construcción del E.T.T. se pueden utilizar herramientas del mercado o programas diseñados específicamente.

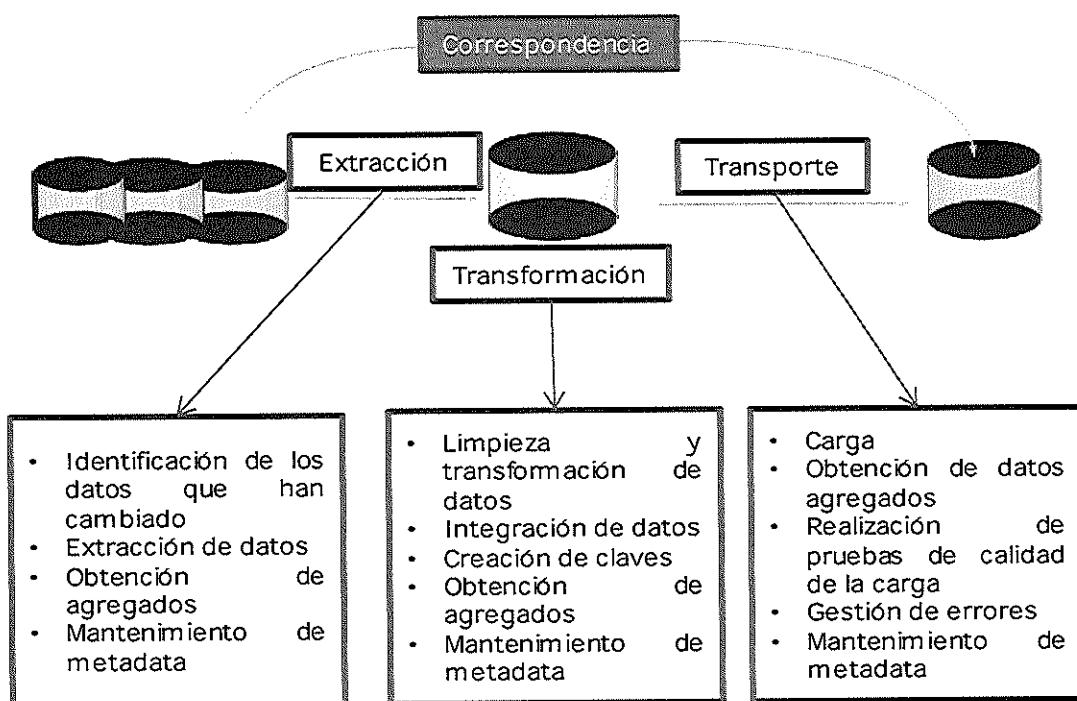
Entre las funciones que cumple este sistema cabe destacar:

- **Carga inicial:** initial load.
- **Mantenimiento o refresco periódico (refreshment):** inmediato, diario, semanal, mensual etc.



El almacenamiento intermedio permite:

- Realizar transformaciones sin paralizar las bases de datos operacionales y el almacén de datos.
- Almacenar metadatos.
- Facilitar la integración de fuentes externas.

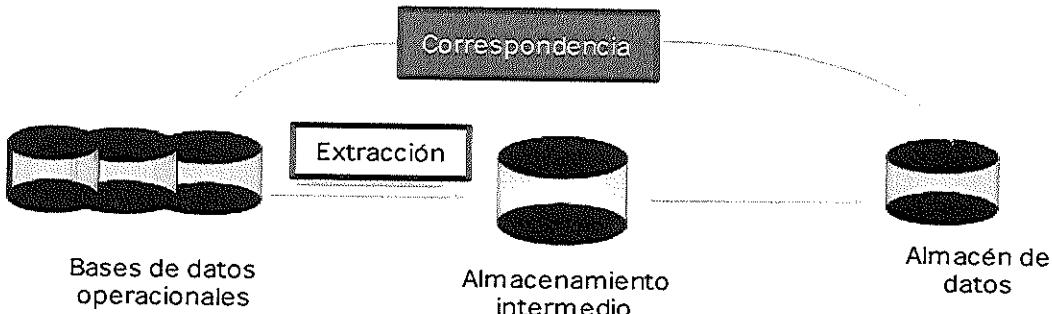


La calidad de los datos es la clave del éxito de un almacén de datos.

Para definir una estrategia de calidad:

- Actuación sobre los sistemas operacionales: modificar las reglas de integridad, los disparadores y las aplicaciones de los sistemas operacionales.
- Documentación de las fuentes de datos.
- Definición de un proceso de transformación.
- Nombramiento de un responsable de calidad del sistema.

Extracción



En la extracción se produce la lectura de datos del sistema operacional, ya sea durante la carga inicial o durante la etapa de mantenimiento del almacén de datos.

La ejecución de la extracción:

- Cuando los datos operaciones están mantenidos en un SGBDR, la extracción de datos se puede reducir a consultas en SQL o rutinas programadas.
- Cuando los datos operacionales están en un sistema propietario o en fuentes externas textuales, la extracción puede ser más complicada y puede tener que realizarse a partir de informes o volcados de datos proporcionados por los propietarios que deberán ser procesados posteriormente.

Durante el mantenimiento del almacén de datos se requiere precisión en la extracción de los datos, y para ello es oportuno realizar la identificación de los cambios.

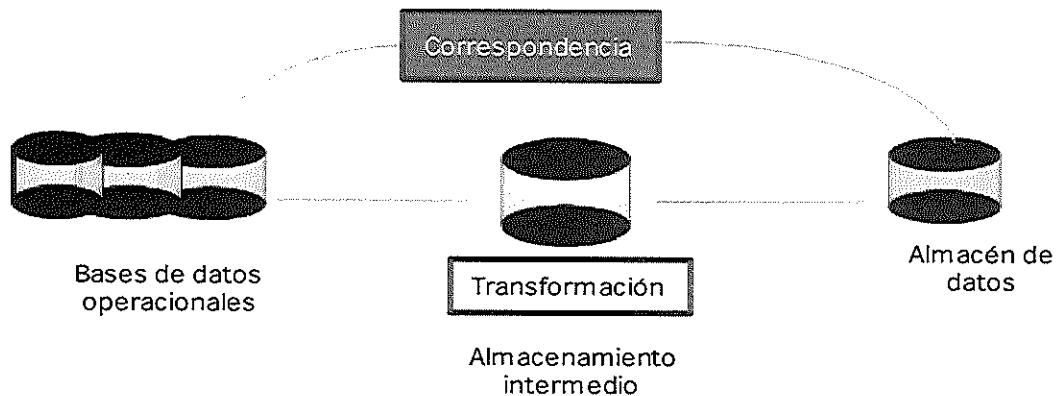
La **identificación de los cambios** se realiza al detectar aquellos datos operacionales, o relevantes, que han podido sufrir una modificación desde la fecha del último mantenimiento.

Para ello existen diferentes métodos:

- Carga total: cada vez que empieza de cero.
- Comparación de instancias de la base de datos operacional.
- Uso de marcas de tiempo en los registros del sistema operacional.
- Uso de disparadores en el sistema operacional.

- Uso del fichero de gestión de transacciones del sistema operacional.
- Uso de técnicas mixtas.

Transformación



La transformación se refiere al cambio de los datos extraídos de las fuentes operacionales (limpieza, estandarización, etc.) y al proceso de cálculo de los datos derivados al aplicar las leyes de derivación.

En los datos operacionales pueden existir anomalías, desde desarrollos independientes a lo largo del tiempo, fuentes heterogéneas, etc.

Es necesario e importante eliminar dichas anomalías, mediante alguno de los siguientes procesos:

- **Limpieza de datos:** eliminar datos, corregir, completar datos, eliminar duplicados.
- **Estandarización:** codificación, formatos, unidades de medida, etc.

Transporte

La fase de transporte o de carga consiste en mover los datos desde las fuentes operacionales o el almacenamiento intermedio hasta el almacén de datos y cargar los datos en la correspondiente estructura de datos.

El proceso de carga puede consumir mucho tiempo, en la carga inicial del almacén de datos se mueven grandes volúmenes de datos. En los mantenimientos que se realizan de forma periódica del almacén de datos se mueven pequeños volúmenes de datos.

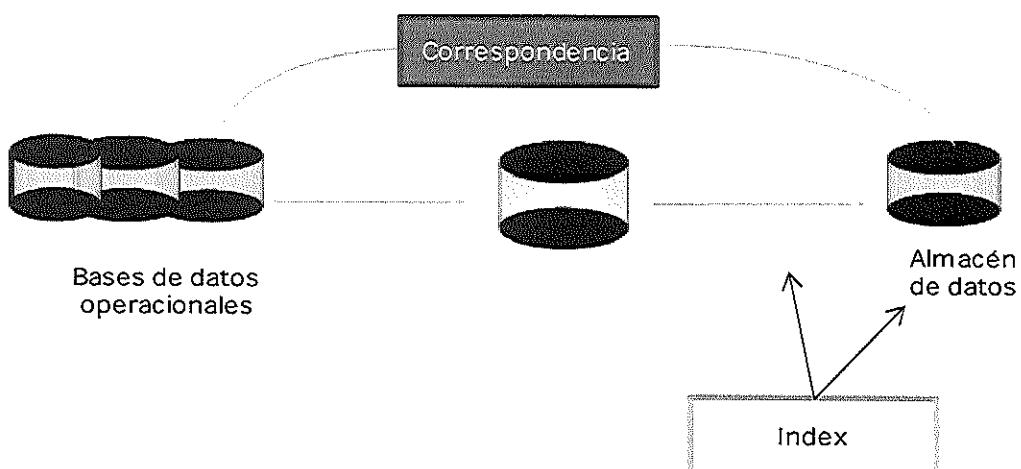
La frecuencia del mantenimiento periódico está determinada por el gránulo del almacén de datos y los diferentes requisitos de los usuarios.

El primer paso es la creación y mantenimiento de la base de datos, posteriormente se definen los intervalos fijos de tiempo a los que añadir cambios al almacén de datos. Se deben determinar las "ventanas de carga" más convenientes para no saturar la base de datos operacional.

Ocasionalmente se pueden archivar o eliminar datos obsoletos que ya no interesa para el análisis.

Posteriormente, tras la carga viene el proceso de indización.

- **Durante la carga:** la carga con el índice habilitado y el proceso de tupla a tupla, es un proceso lento.
- **Después de la carga:** se produce la carga con el índice deshabilitado y se produce la creación del índice de forma total o parcial.

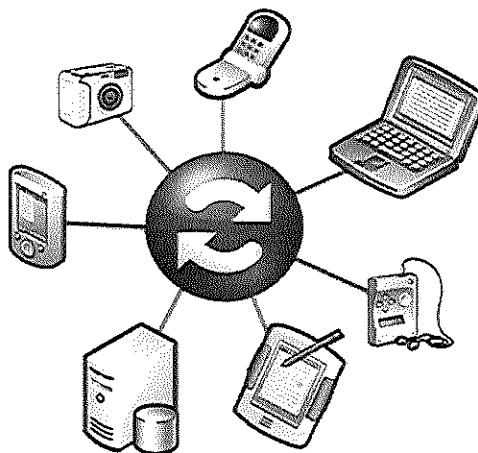


Finalmente se realiza la obtención de agregados.

- Durante la extracción.
- Despues de la carga o transporte.

2.1. Identificación de orígenes de datos para la carga de datos

Un sistema de almacén de datos puede verse como una jerarquía de datos, que tiene su origen en los datos almacenados en los sistemas operacionales y que termina con los datos almacenados en el almacén de datos.



El repositorio de datos operacionales es la fuente donde se encuentran los datos primitivos, actuales e integrados, por lo tanto es el encargado de suministrar datos al sistema, estos datos operacionales pueden ser:

- Procedentes de sistemas mainframe.
- Datos de estaciones de trabajo o servidores privados.
- Sistemas externos como las bases de datos comerciales, de proveedores o clientes, o incluso de Internet.
- Datos departamentales almacenados en sistemas propietarios.

El gestor de carga es el encargado de la extracción y la carga de los diferentes datos que se encuentran en el repositorio de datos, además de realizar algunas transformaciones simples a los datos con el fin de que estén adaptados a las necesidades del almacén de datos.

La **carpeta de origen** es la carpeta del servidor en la que se encuentran los archivos de la bases de datos que se van a transferir. Estos datos como se explicó con anterioridad pueden tener diferentes estructuras, y ello marcará en consecuencia la forma de trabajar con ello.

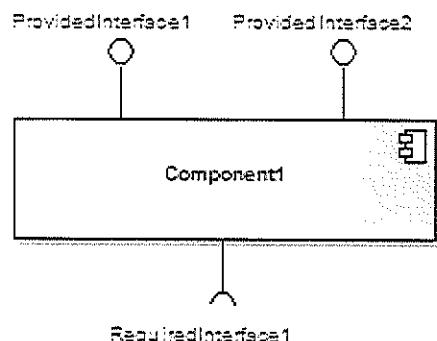
2.2. Creación de componentes de software para extraer información de un sistema de almacén de datos

Un **componente** es un elemento de un sistema software que ofrece un conjunto de servicios o funcionalidades, a través de interfaces definidas.

Estos componentes deben cumplir con una serie de características:

- Ser reutilizables.
- Ser intercambiables.
- Poseer interfaces definidas.
- Ser cohesivos.

Estos componentes son la piedra angular de diferentes paradigmas de programación. Esto ha dado lugar a la aparición en el mercado de gran variedad de especificaciones que plantean la forma de construir, utilizar y distribuir componentes..



La **extracción** de información es un tipo de recuperación de la información cuyo objetivo es extraer automáticamente información estructurada o semiestructurada desde documentos legibles del ordenador.

Los principales objetivos del desarrollo de componentes de software ya sea concretamente en sistemas de almacenamiento de datos, o en cualquier otro tipo de sistemas, es en general, reducir el tiempo de trabajo, el esfuerzo que requiere implementar la aplicación, y los costos del propio proyecto, y de esta forma, incrementar el nivel de productividad de los grupos desarrolladores y minimizar los riesgos globales.

El desarrollo de la descripción de la secuencia de actividades que debe seguirse en un equipo para poder generar un conjunto coherente de productos, en este caso de información es el objetivo del desarrollo de software. Con ello se pretende hacer predecible el trabajo, tanto el costo, como mantener un nivel de calidad.

No existe un único proceso de desarrollo universal. Se ha de configurar en función de la naturaleza del producto y de la experiencia de la empresa. Existen diferentes tipos de aplicaciones:

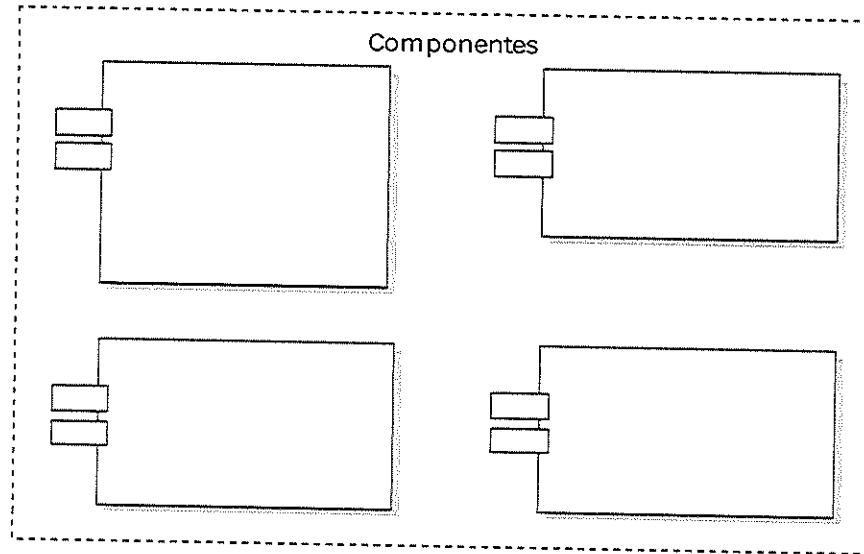
- **Aplicación monoprocesadora:** se realiza en un solo ordenador, no se produce comunicación con otras aplicaciones.
- **Aplicaciones embebidas:** se ejecuta en un entorno automático especial.
- **Aplicaciones de tiempo real:** tiene entre sus especificaciones requerimientos temporales.
- **Aplicaciones distribuidas:** se ejecutan en varios procesadores, por lo que requiere de intercomunicación a través de la red.

Modelo de componentes

Este modelo ilustra los componentes de software que se usarán para construir el sistema. Se pueden construir a partir del modelo de clases y escribir desde cero para el nuevo sistema o se puede importar de otros proyectos y de productos de terceros.

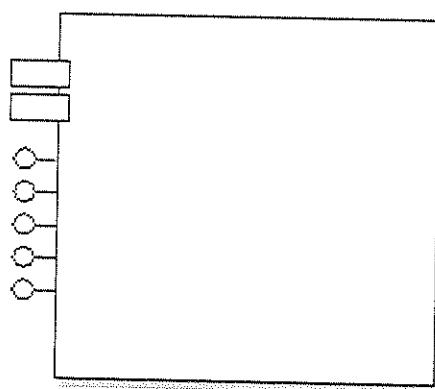
Los componentes son agregaciones de alto nivel de las piezas de software más pequeñas y proveen un enfoque de construcción de bloques de "caja negra" para la elaboración del software.

Los componentes se suelen representar gráficamente como a continuación se detalla, pudiendo ser estos desde controles hasta interfaz de usuario o como servidores de reglas de negocio.

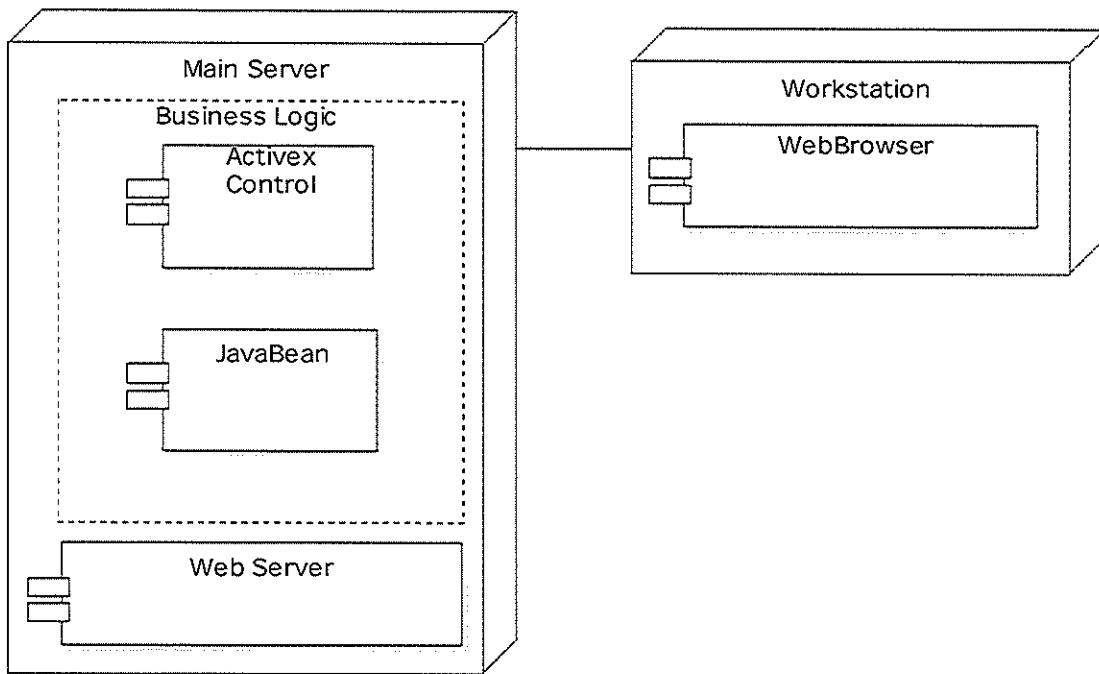


La relación entre los diferentes componentes y el software se puede observar en el diagrama de componentes, sus dependencias, su comunicación así como sus condiciones y la ubicación-

No obstante los componentes pueden exponer las interfaces. Estas son todos los puntos visibles de entrada o los servicios que un componente ofrece y deja disponible a otros componentes de software y clases. Típicamente un componente está compuesto por numerosas clases y paquetes de clases internos. También se puede crear a partir de la colección de componentes más pequeños.



Es en el diagrama de despliegue en el que se puede observar el despliegue físico del sistema en un ambiente de producción o de prueba. Se observa dónde se ubican los diferentes componentes, en qué servidores, máquinas o hardware. Puede representar los enlaces de redes, el ancho de banda de LAN, etc.



Los componentes pueden ir acompañados de ciertos requisitos para indicar sus obligaciones contractuales, es decir, qué servicios son los que proveen en el modelo. Los requisitos suponen una ayuda para poder documentar el comportamiento funcional de los elementos del software.

Además de los requisitos, también pueden verse incluidas ciertas restricciones que indican el entorno o lugar en el que pueden operar. Las pre-condiciones especifican lo que debe ser verdadero antes de que un componente pueda realizar alguna función: las post-condiciones indican lo que debe ser verdadero durante la vida útil del componente.

Las descripciones textuales y procedimentales de las acciones de un objeto a lo largo del tiempo, son los escenarios, y describen la forma en la que un componente trabaja. Se pueden crear múltiples escenarios para describir tanto el camino básico como las excepciones, errores y otras condiciones.

Un componente puede implementar otro elemento del modelo o un componente puede ser implementado por otro elemento. Al emplear las relaciones de realización desde y hacia los componentes, se pueden seguir las dependencias entre los elementos del modelo y la trazabilidad desde los requisitos iniciales hasta la implementación final.

RECUERDA

- Los almacenes de datos suponen una ayuda en la toma de decisión por parte de la empresa o la organización. Este tipo de almacenes son sobre todo un expediente de una empresa que va más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos.
- Los cubos de información también son conocidos como DataMart , siguen una lógica de los datos en bruto, de los datos provistos por su sistema de operaciones/finanzas hacia el almacén de datos con la adición de nuevas dimensiones o información calculada.
- La información está constituida por mensajes, es decir, es un conjunto de datos que representan ideas mediante las cuales se incrementa nuestra conciencia, inteligencia o conocimiento. Los mensajes pueden adoptar diferentes manifestaciones físicas. En líneas generales se definen los mensajes como manifestaciones físicas de la información.
- Una estructura de datos, o un tipo de datos estructurado, es un tipo de dato construido a partir de otros. Un dato de tipo estructurado está compuesto por una serie de datos de tipos elementales y alguna relación existente entre ellos. Normalmente, la relación suele ser de orden aunque puede ser de cualquier otro tipo.
- Durante el mantenimiento del almacén de datos se requiere precisión en la extracción de los datos, y para ello es oportuno realizar la identificación de los cambios.
- La identificación de los cambios se realiza al detectar aquellos datos operacionales, o relevantes, que han podido sufrir una modificación desde la fecha del último mantenimiento.
- Un sistema de almacén de datos puede verse como una jerarquía de datos, que tiene su origen en los datos almacenados en los sistemas operacionales y que termina con los datos almacenados en el almacén de datos.
- La carpeta de origen es la carpeta del servidor en la que se encuentran los archivos de la bases de datos que se van a transferir. Estos datos como se explicó con anterioridad pueden tener diferentes estructuras,

y ello marcará en consecuencia la forma de trabajar con ello.

- Un componente es un elemento de un sistema software que ofrece un conjunto de servicios o funcionalidades, a través de interfaces definidas.
- La extracción de información es un tipo de recuperación de la información cuyo objetivo es extraer automáticamente información estructurada o semiestructurada desde documentos legibles del ordenador.

Preguntas de Autoevaluación

1. Indica si es verdadero o falso el siguiente enunciado:

"Los almacenes de datos son también conocidos como Data Warehouse".

- a) Verdadero.
- b) Falso.

2. Segundo la clasificación general de la información, ¿cómo puede ser esta? Selecciona las respuestas correctas.

- a) Estática.
- b) Dinámica.
- c) Negativa.

3. ¿Qué tipo de estructura lógica supone el primer nivel de agregación de los datos detallados actuales?

- a) Datos detallados.
- b) Datos detallados actuales.
- c) Dato ligeramente resumidos.

4. ¿Qué tipos generales de información se puede observar en relación con su almacenamiento? Selecciona las respuestas correctas.

- a) Continua.
- b) Discreta.
- c) Discontinua.

5. ¿Cómo se denomina a la secuencia de caracteres que se interpretan como un dato único?

- a) Cadena.
- b) Puntero.
- c) Array.

UD2 Extracción de datos (data Warehouse)



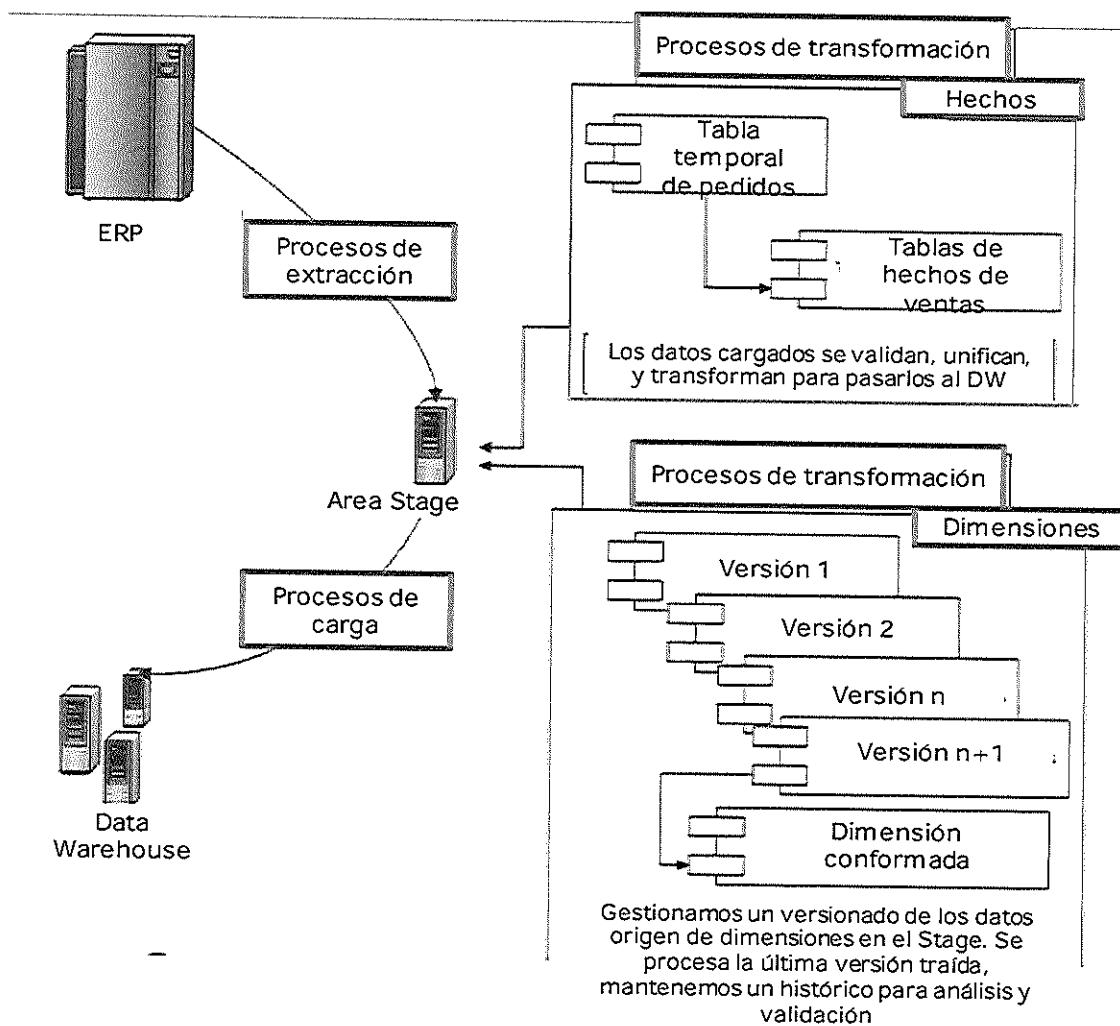
UF1890 Desarrollo de componentes de software y consultas dentro del sistema de almacén de datos

1. Herramientas para la carga y extracción de datos de sistemas de almacén de datos

Como se observó en el tema anterior, el proceso de extracción, transformación y carga, conocido como ETL, es importante ya que es la forma en que los datos se guardan en un almacén de datos.

Implican las siguientes operaciones:

- **Extracción:** acción de obtener la información deseada a partir de los datos almacenados en fuentes externas.
- **Transformación:** cualquier operación realizada sobre los datos para que puedan ser cargados en el data warehouse o se puedan migrar de éste a otra base de datos.
- **Carga:** consiste en almacenar los datos en la base de datos final.

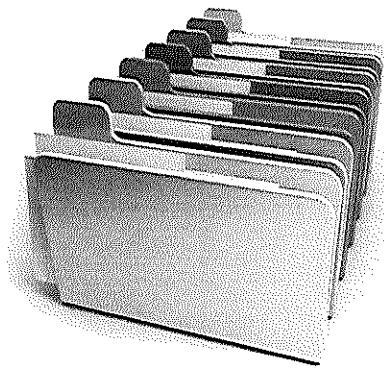


1.1. Mecanismos que se utilizan para la extracción de datos

La primera parte del proceso ETL consiste en **extraer** los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos.

Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba. De no ser así los datos son rechazados.



Un requerimiento importante que se debe exigir a la tarea de extracción es que ésta cause un impacto mínimo en el sistema origen. Si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que éste no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde este impacto sea nulo o mínimo.

La extracción de datos se define como una estrategia o proceso central que trata de ensamblar datos desde sistemas y fuentes dispares, enriqueciendo estos datos de manera que se produzca información valiosa y reutilizable.

Hay ciertos aspectos a tener en cuenta en este proceso:

- Extraer los datos eficiente y eficazmente desde los sistemas fuentes.
- Identificar y documentar el nivel de servicio de acuerdo con los sistemas fuentes.
- Brindar una guía sobre el mecanismo de transferencia de datos, escalando los procedimientos en caso de fallos durante la transmisión de los datos.
- Suplementar/enriquecer los datos desde los sistemas fuentes para una fácil actualización y cambio que sea aplicable en el ambiente de data warehouse.
- Proveer un marco para permitir una estrategia de reconciliación favorable de acuerdo a la fuente y el destino final.

En esta fase lo más adecuado es utilizar una herramienta ETT (en inglés ETL), que permita manipular una amplia variedad de datos del sistema fuente pudiendo conseguir que todos los datos tengan un formato común.

Hay que tener en cuenta que la herramienta de ETT que se seleccione ha de permitir lograr los resultados deseados en un tiempo relativamente menor que la forma tradicional de codificar y mantener los objetos del data warehouse.

En la selección del ETT se debe tener en cuenta:

- Fácil de usar y comprensible desde el punto de vista del mantenimiento y el desarrollo de la perspectiva.
- El proceso ETL debe integrarse con el proceso de negocio.
- Debe soportar el procesamiento de grandes volúmenes de datos.
- Poder extraer datos desde distintas fuentes heterogéneas.
- Puede ser necesario que la herramienta ETL soporte procesamiento paralelo.
- Debe poseer un amplio espectro de conectividad y la habilidad de estandarizar los datos tomados desde diversas fuentes, que pueden estar incluso almacenadas en bases de datos soportadas sobre una plataforma diferente.

Una posible metodología para la extracción de conocimiento a partir de los datos sería la que aparece a continuación descrita en la siguiente tabla.

Fases	Objetivo	Técnica/Herramienta
Identificar	Seleccionar las bases de datos que puedan aportar la información necesaria para obtener el conocimiento.	Experiencia de los expertos.
Extraer	Ensamblar datos desde fuentes dispares, enriqueciéndolos de manera que cree información valiosa.	Herramientas ETL/ Data Warehouse.
Procesar	Construir por medio de algoritmos de Minería de Datos, modelos de comportamiento.	Minería de Datos
Almacenar	Validar, seleccionar y mantener los modelos de Minería de Datos/ Experiencia del comportamiento.	Experiencia del Ingeniero del Conocimiento.
Compartir	Poner a disposición de la Organización el conocimiento descubierto.	Portal del Conocimiento.

Una vez extraídos los datos, se integran y almacenan en un data warehouse. La meta de un data warehouse es integrar aplicaciones a nivel de datos. El dato extraído de los sistemas operacionales se procesa, se transforma y ubica de acuerdo a un esquema similar a un modelo entidad/relación. La noción del data warehouse debe ser extendida para incluir no sólo datos orientados a transacciones, sino también aquellos datos creados por los ingenieros del conocimiento.

OLAP On-Line Analytical Processing

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis implica la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil. Este sistema es típico de los datamart.

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, por lo general implica entre dos y cinco años.

- Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL).

En la base de cualquier sistema OLAP se encuentra el concepto de cubo OLAP (también llamado cubo multidimensional o hipercubo). Se compone de hechos numéricos llamados medidas que se clasifican por dimensiones. El cubo de metadatos es típicamente creado a partir de un esquema en estrella o copo de nieve, esquema de las tablas en una base de datos relacional. Las medidas se obtienen de los registros de una tabla de hechos y las dimensiones se derivan de la dimensión de los cuadros.

Los sistemas OLAP se clasifican en:

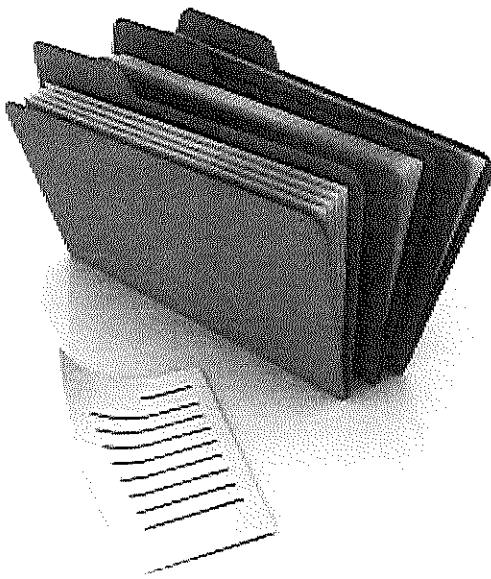
- **ROLAP:** es una implementación OLAP que almacena los datos en un motor relacional. Los datos son detallados para evitar las agregaciones y las tablas se encuentran desnormalizadas. Suelen trabajar sobre esquemas estrella o copo de nieve, aunque es posible trabajar sobre cualquier base de datos relacional.
- **MOLAP:** almacena datos en una base de datos multidimensional. Para de este optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado.
- **HOLAP:** almacena datos en un motor relacional y en otros en una base de datos multidimensional.

Estas herramientas permiten presentar al usuario una visión multidimensional de los datos para cada actividad que es objeto de análisis. El usuario formula consultas a la herramienta OLAP seleccionado los atributos de este esquema multidimensional sin conocer la estructura interna del almacén de datos, es decir sin conocer el esquema físico del almacén de datos.

La herramienta OLAP genera la correspondiente consulta y la envía al gestor de consultas del sistema.

Una consulta a un almacén de datos consiste en la obtención de medidas sobre los hechos parametrizados por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones.

1.2. Estructuración de la información para adecuarse a las necesidades de la empresa



La **información** es un conjunto de datos procesados, que constituyen un mensaje que cambia el estado de conocimiento del sujeto o sistema que recibe dicho mensaje.

Cualquier tipo de material que posea las características de un documento, cuenta tanto con un contexto externo como interno.

En el ambiente de la recuperación de información todo tipo de documentación tiene unos parámetros que permiten ubicarlos de un modo fácil y rápido y esto es lo que se conoce como **palabras claves**, las cuales son identificadas en la búsqueda por los buscadores y toman como referencia para llegar a la información.

Los **descriptores** son los encargados de sintetizar la información para así optimizar la localización. Los **listados de encabezamiento de materia** son una frase que contiene el tema que se quiere referenciar; los **términos** o las palabras que expresan una idea de una disciplina determinada, son los que pueden ser términos compuestos conformados por una unión de letras y además cuentan con un significado.

Los **descriptores libres o no descriptores**, están formados por palabras no relevantes sobre el tema y que no sintetizan la información para su posterior

ubicación. Finalmente las **etiquetas**, son un conjunto de caracteres útiles en la identificación de los datos o instrucciones.

En el proceso de extracción de datos es importante tener en cuenta la tipología genérica de los documentos. La extracción de datos se suele utilizar para el rellenado automático de metadatos o como entrada para otro tipo de software empresarial.

A continuación se describen los tipos de documentos generales.

Documentos estructurados

Son documentos en los que se sabe que información se va a encontrar y la posición que esta información ocupa dentro de las dimensiones físicas del documento.

En este tipo de documentos es relativamente sencillo encontrar y extraer datos, pues podemos saber dónde buscarlos.

Documentos semiestructurados

La dificultad de gestión va creciendo conforme a la información se vuelve estructurada. Los documentos semiestructurados, son aquellos en los que se sabe que información se va a encontrar, pero no se sabe exactamente en dónde se encuentra.

La forma en que se extraen datos de documentos semiestructurados no puede ser la misma que en el caso anterior, aquí es necesario enseñarle al software cómo es lo que se está buscando.

Documento desestructurados

En este tipo de documentos ni se sabe qué vamos a encontrar, ni donde se encuentra.

La dificultad de extracción de datos de este tipo de documentos es máxima. En este grupo se incluye los informes y las cartas. Según algunos autores este grupo de información no existe.

Según algunos autores estos documentos cumplen generalmente tres características:

- La estructura del documento no ha sido diseñada por la empresa que ahora quiere gestionarlos.
- La estructura de este documento puede variar dependiendo de quien la envía.
- No puede ser procesados siguiéndose a un template o plantilla.

Estructura lógica y estructura física

En el intento de estructuración de la información y los datos es imprescindible tener claras las ideas básicas sobre los mecanismos que se encuentran involucrados.

- **Estructura lógica:** se relaciona con la idea inicial de los programados acerca de cómo están organizados los diferentes datos, y coincide aproximadamente con la forma en que son manipulados los datos por programador de alto nivel.
- **Estructura física:** se relaciona con la forma en que están contenidos los datos en la máquina, de la que existen dos versiones:
 - Una corresponde a la que adoptan los datos en memoria.
 - Otra correspondiente a su almacenamiento externo.

Es evidente que ambas estructuras no se corresponde, la **estructura física** de datos en los almacenamientos externos no se corresponden de forma exacta con la estructura lógica. En primer lugar, el documento de clientes antes mencionado, puede estar representado físicamente por varios ficheros que pueden ser multi-volumen. Es decir, que ocupan más de un volumen lógico en la máquina que los contiene. Si son aplicaciones de red, pueden estar incluso en máquinas remotas, distintas de la que ejecuta la aplicación. Además, aunque nos figuremos la estructura lógica es un todo continuo, sabemos que la estructura física correspondiente, incluso si se tratase de un solo fichero, está compuesta por trozos que pueden estar dispersos en el disco.

La **estructura lógica** está ordenada por números o bien por nombres. En cambio la estructura física puede estar construida simplemente por el orden natural, es decir, por el orden de creación de los propios registros. Normalmente, la

apariencia de ordenación es el resultado de un proceso complejo que utiliza índices, tablas y punteros, para proporcionar un acceso ordenado a una estructura mucho más caótica.

El programador con ayuda del programa se encarga de manejar los datos en términos de estructura lógica. Con respecto a ello, las herramientas que proporcionen el lenguaje o entorno de programación, deberán ser de mayor nivel cuanto mayor sea la distancia con que pueda ser manejada la estructura lógica de datos respecto de su verdadera estructura física. Precisamente el manejo de estas estructura a dado lugar a la distancia con que pueda ser manejada la estructura lógica de datos respecto de su verdadera estructura física.

De este modo el manejo de dichas estructuras ha dado lugar a toda una rama de la industria del software que ha logrado alcanzar un elevado nivel de sofisticación y especialización.

2. Creación de extractores de datos

Un **extractor de datos** es un conjunto de motores de extracción que permite obtener mediante la extracción los datos de las bases de datos y utilizar programas de visualización.

Es decir permite la extracción automática de entidades o conceptos del contenido de texto visible de un elemento y asignarlos a una propiedad administrada. Al mismo tiempo, estas propiedades se pueden usar para restringir las consultas mediante filtros de propiedades o como opciones de refinamiento de consultas.



Existen dos tipos de extractores:

- **Extractores de propiedades de coincidencia exactas:** son adecuados para establecer coincidencias de cadenas en todos los idiomas. Las entradas en el diccionario personalizado pueden ser palabras individuales o una cadena de palabras. La búsqueda de cadenas coincidentes se realiza después de una tokenización básicas, que reemplaza los caracteres separados presentes en el textos por espacios en blanco. Tras la tokenización básica, los extractores deben buscar una coincidencia exacta para la cadena.
- **Extractores de propiedades de coincidencia parcial:** son adecuados para establecer coincidencias de cadenas en todos los documentos del este asiático ya que las palabras de estos idiomas no están separados por espacios. También pueden ser usado en aquellos casos específicos en los que se necesitan coincidencias de subcadenas.

En ambos casos es importante tener en cuenta que el establecimiento de coincidencias se distingue mayúsculas y minúsculas.

A continuación se explicará los aspectos claves en el proceso de creación de un extractor de datos:

- Obtención de información de fuentes internas o externas.
- Agrupación, transformación y homogeneización de la información para su posterior estudio.

2.1. Obtención de información de fuentes internas o externas

El proceso de obtención de la información comienza por la selección de las fuentes utilizables, es decir, los lugares en los que se obtiene la información útil para poder realizar las diferentes tareas con ella.

Se puede hablar de dos tipos de fuentes:

- **Internas:** propias de la empresa.
- **Externas:** ajenas a la empresa.

Al mismo tiempo, dentro de ambos tipos de información se distingue:

- **Fuentes primarias:** son aquellas que se adquieren mediante los procesos de investigación de forma directa, ya sea de la propia empresa o por encargo a empresas dedicadas a la elaboración de este tipo de información.
- **Fuentes secundarias:** son aquellos datos que proceden de estadísticas y documentos ya publicados, proporcionan información de tipo general ya elaborado. Es recomendable empezar la investigación acudiendo a ellas. Son fuentes secundarias externas las publicaciones de organismos oficiales, bancos, etc.

Dentro de ambos tipos de información existen amplios grupos de datos, que dan lugar a la clasificación del tipo de información dentro de las fuentes de información primaria y secundaria.

Información cuantitativa

Es la que se extrae mediante métodos diversos a partir de una muestra representativa de la población para proyectar los resultados y conclusiones a toda la población.

- **La encuesta:** ofrece información abundante si se cuenta con un buen cuestionario. Un cuestionario es un documento que recoge una serie de preguntas formuladas con claridad, de fácil comprensión, sin implicación de respuestas y colocadas con un orden lógico. Además se debe determinar la población que se quiere estudiar y la muestra

correcta. Hay diferentes tipos de encuestas: personal, postal, telefónica, en Internet.

- **La encuesta Ómnibus:** consiste en una entrevista personal con varios apartados dentro del cuestionario sobre diferentes temas o productos.
- **Método Delphi:** es una variante específica de la encuesta Ómnibus, que está basada en entrevistas a expertos y consultores que elaboran informes independientes entre sí acerca de las cuestiones investigadas.
- **El panel:** es una encuesta periódica que se realiza a las mismas personas, es decir, la muestra es permanente. Permite identificar cambios y la evolución en el tiempo de las variables investigadas. Se utiliza para medir audiencias de televisión, en otros usos.
- **La observación:** consiste en observar la conducta de los diferentes consumidores y extraer conclusiones a partir de ellas. Las personas son totalmente libres para comportarse y reaccionar de forma espontánea. La observación puede ser directa o indirecta, a través de cámaras.
- **La experimentación o prueba de mercado simulada:** la observación se hace en un escenario preparado, donde se busca la participación interactiva de las personas que componen la muestra. Se pretende reproducir a escala reducida situaciones reales para poder prever los resultados, problemas, ventajas, inconvenientes, etc.

Información cualitativa

Emplea métodos adecuados para investigar o buscar necesidades, hábitos de consumo, etc.

- **Encuesta en profundidad:** es una entrevista abierta realizada por un profesional. Supone establecer un diálogo con un individuo para conocer sus motivaciones, gustos, personalidades, actitudes, etc.
- **Reunión de grupo:** se trata de una reunión activa de un grupo de entre seis a diez consumidores con un moderador para hablar sobre un producto.

- **La pseudocompra:** el investigador se pone en el lugar de posibles compradores o clientes.
- **Técnicas proyectivas:** intenta conocer los impulsos psicológicos que subyacen detrás del comportamiento del consumidor estudiando su reacción ante determinados estímulos externos, por ejemplo ante frases, imágenes, marcas, colores, etc. es necesaria la observación directa del cliente, o consumidor, en el lugar de compra o el uso de medios como las tarjetas de fidelización con las que se puede analizar cómo reaccionan los consumidores o sus políticas comerciales.



La obtención de la información, de los datos, requiere del cumplimiento de diferentes aspectos que configuran el proceso completo.

- **Planteamiento de una necesidad:** la falta de información puede ser un problema, y por tanto es necesario dar solución mediante la obtención de los datos requeridos.
- **Determinación de los objetivos a alcanzar:** determinar qué se quiere conseguir, que resultados son los esperados y cómo se van a emplear.
- **Fijar el contenido del estudio:** dependiendo de los objetivos que se quieran conseguir, se definen en detalle el contenido de la información que se necesita.
- **Búsqueda de información:** se obtiene la información a partir de las fuentes primarias y secundarias.

- **Planificación de la intervención:**

- Determinación de las fuentes secundarias de obtención de la información.
- Determinar las fuentes primarias, técnicas a emplear, selección de la muestra y elaboración del cuestionario.
- Estimación y planificación del tiempo necesario.
- Cuantificación del coste de la investigación.
- Señalar las personas encargadas de la investigación y asignación de responsabilidad.

- **Obtención de la información:** la información es adquirida en base a todos los aspectos anteriormente señalados.
- **Interpretación de resultados:** se le da sentido a la información obtenida, para poder dar respuesta a la necesidad inicial que puso en marcha todo este proceso.

La obtención de información útil, óptima y adecuada, en un momento determinado, es uno de los recursos más importantes para una empresa. El proceso comienza con la captación de datos de la realidad en la que desenvuelven, continúa con su procesamiento y culmina con la emisión de informes que les permiten tomar decisiones con una menor dosis de incertidumbre.

La obtención de los datos comprende varios procesos:

- **Procesamiento de los datos:** consiste en la actividad de clasificación, registro, cálculo y almacenamiento de datos, estableciendo relaciones entre ellos de forma tal que brinden información.
- **Obtención de la información:** es decir los informes de salida del sistema de información. Los informes presentan la información de manera ordenada para que los datos procesados puedan ser interpretados para poder sacar conclusiones válidas.
- **Sistemas de información:** se entiende por sistema al conjunto de elementos que se relacionan entre sí para poder aportar y contribuir a la consecución de un objetivo concreto. Uno de los sistemas más importantes, es el de organización de la información. Se compone de los siguientes elementos:

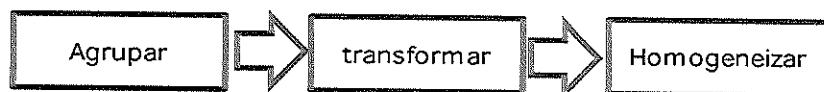
- Datos que ingresan.
- Un proceso de elaboración.
- Combinación.
- Modificación.
- Información emitida de diferentes formatos.

2.2. Agrupación, transformación y homogeneización de la información para su posterior estudio

Según la definición de la Real Academia Española, los siguientes términos deben entenderse del siguiente modo:

- **Agrupar:** es reunir en grupo o apiñar, constituyendo una agrupación.
- **Transformar:** hacer mutar algo en otra cosa.
- **Homogeneizar:** hacer homogéneo, por medios físicos o químicos, un compuesto o mezcla de elementos diversos.

En base a dichas definiciones nos adentraremos en la utilización de dichos términos para la comprensión del proceso de tratamiento de información para su posterior estudio.



Los contenidos se pueden **agrupar** de diferentes modos:

- Arriba a bajo.
- Abajo arriba.
- Orden cronológico.
- Geográfico.
- Jerárquico.
- Por tareas.

Pero siempre con el objetivo de responder a las necesidades que ha dado lugar la necesidad de agrupar la información.

La información frecuentemente se compone de diferentes contenidos que pueden ser desarrollados en diversos formatos:

- Texto.
- Vídeo.
- Audio.
- Fotografías.

Cuando estos elementos forman parte de un listado que no tiene ningún orden, el usuario que va a trabajar con ellos se verá obligado a recorrer toda la lista, en la que todos los elementos pueden aparecer mezclados.

Si sólo tenemos cuatro contenidos quizás pueda funcionar dicha forma de buscar los datos, pero sin duda puede que el usuario los encuentre por casualidad, y por tanto esta opción no parece ser la más operativa, sobre todo si se trata de decenas, cientos o millones de contenidos.

Surge de este modo la necesidad de organizar la información de algún modo. Lo más adecuado sería diseñar una forma de agrupación de contenidos con la que el mayor porcentaje de usuarios pueda encontrar lo que necesita lo antes posible y con el menor esfuerzo posible, dependiendo sus necesidades concretas en cuanto a la búsqueda de información.

Hay bastantes formas de llevar a cabo la agrupación de la información:

- **Agrupaciones exactas:** un contenido sólo podrá pertenecer a una categoría y no a otra.
 - Por orden cronológico.
 - Por ubicación geográfica.
- **Agrupaciones ambiguas:** un contenido podría encontrarse en más de una categoría. Qué esté en una u otra va a depender de diversos factores.
 - Por jerarquías.
 - Por facetas.
 - Por tareas.
 - Por audiencias.
 - Por prominencia. Contenidos que son destacados.

- Híbridos. Ocurren cuando se mezclan varios tipos que no se deberían usar en un mismo espacio ya que pueden desorientar al usuario.

La creación de la agrupación o clasificación de los diferentes contenidos, se puede llevar a cabo mediante dos estrategias fundamentalmente:

- **De arriba abajo:** también conocida como tip-down. Primero se deciden las categorías principales, y después, cada una se va desglosando en subcategorías.
- **De abajo arriba:** también se conoce como bottom-up. Primero se listan las subcategorías o temáticas que se van a tratar o para las cuales se tiene información y después se van agrupando en función de los objetivos establecidos, y de cómo se considere que será el modelo mental más probable que utilice el usuario, es decir como creamos que el usuario realizará la búsqueda.

La **transformación**, es también conocida como conversión, y es el proceso de transformación de datos informáticos de una representación concreta a otra, cambiando los bits de un formato a otro, normalmente para lograr la interoperabilidad de aplicaciones o sistemas diferentes. A nivel más simple, la conversión de datos puede ejemplificarse por la conversión de un fichero de texto desde una codificación de caracteres a otra. Son conversiones más complejas las de formatos de ficheros ofimáticas y multimedia, a veces fuera de las capacidades de ordenadores domésticos.

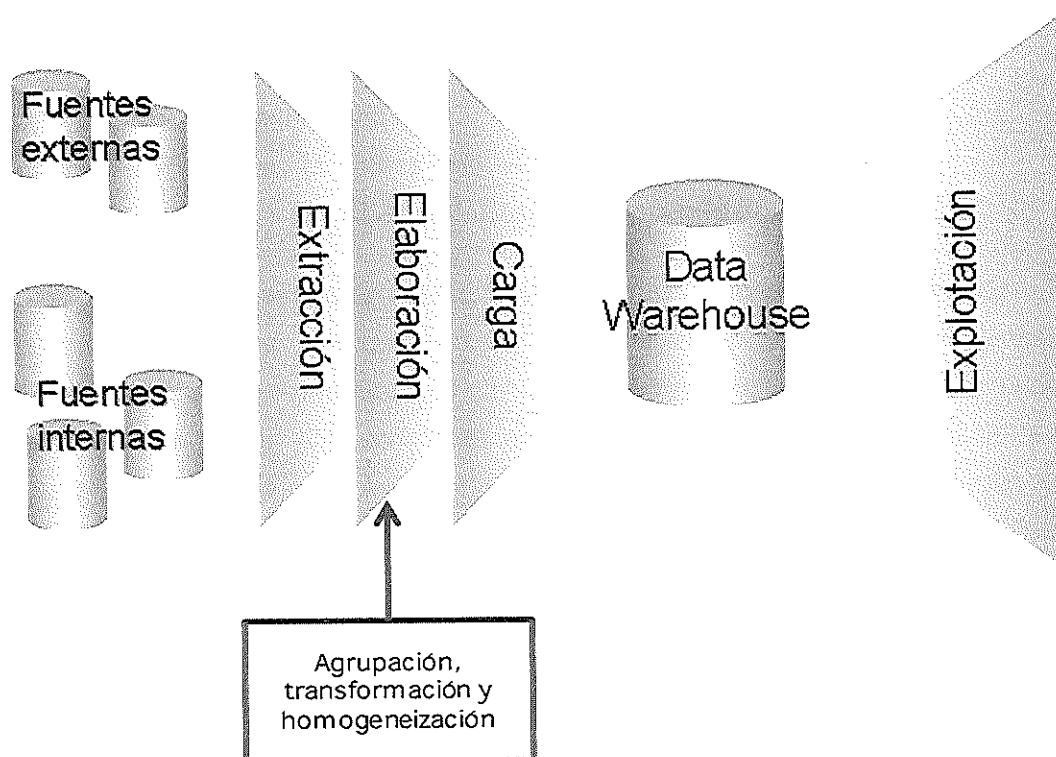
Antes de que pueda efectuarse cualquier tipo de transformación de datos el programador o usuario que sea el encargado de ello debe tener en cuenta unos cuantos conceptos básicos:

- Es fácil descartar información usando un ordenador, pero añadirla requiere esfuerzo.
- El ordenador puede usarse para añadir información sólo con base en reglas; la mayoría de adiciones que interesa a los usuarios sólo puede lograrse con la ayuda de humanos.
- Sobremuestrear los datos o convertirlos en un formato con más posibilidades no añade información: sólo hace hueco para dicha adición, que suele tener que hacer un humano.

El proceso de **homogeneización** es aplicable a diferentes campos, en este contexto hace referencia al proceso mediante el que se produce la mezcla de información, mediante una regla general, para obtener una mejor calidad en los resultados obtenidos del trabajo con la información.

En el proceso de homogeneización se realizan los diferentes ajustes necesarios para poder realizar la comparación y agregación de los diferentes documentos, información, o datos concretos, de un modo adecuado.

Para trabajar con la información es muy importante garantizar que los datos de que se disponen poseen un control adecuado y homogéneo.



Estos tres tipos de procesos tienen lugar en la etapa de elaboración de la información.

En la etapa de extracción se produce la obtención de la información de las fuentes internas y externas como se ha explicado en temas anteriores.

En la fase de elaboración es donde se produce los procesos de agrupación, transformación y homogeneización de la información.

En la fase de carga se organiza y actualizan los datos.

Y en la última fase se produce la extracción y el análisis de la información en los diferentes niveles de agrupación.

Es importante conocer, y diferenciar los siguientes términos:

- **Cubos OLAP (On-Line Analytical Processing):** Estructura cúbica de ordenamiento de los datos con fin de hacer un análisis n dimensional sobre la base de datos.
- **ETL (Extraction, Transformation Load):** Proceso de extracción transformación y carga de datos requerido para alimentar un Data Warehouse.
- **Metadata (datos sobre los datos):** Los metadatos permiten mantener información de la procedencia de la información, la periodicidad de refresco, su fiabilidad, forma de cálculo, etc., relativa a los datos de nuestro almacén.
- **Data Mart:** Base de datos construida para un fin específico. Por lo general, podemos decir que un Data Mart es un subconjunto del Data Warehouse.
- **ODS (operacional data store):** bases de datos operacionales.

UF1890 DESARROLLO DE COMPONENTE SOFTWARE Y CONSULTAS DENTRO DEL SISTEMA DE ALMACÉN DE DATOS

RECUERDA

- El proceso de extracción, transformación y carga, conocido como ETL, es importante ya que es la forma en que los datos se guardan en un almacén de datos.
- La primera parte del proceso ETL consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos.
- Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis implica la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil. Este sistema es típico de los datamart.
- La información es un conjunto de datos procesados, que constituyen un mensaje que cambia el estado de conocimiento del sujeto o sistema que recibe dicho mensaje.
- Los descriptores son los encargados de sintetizar la información para así optimizar la localización. Los listados de encabezamiento de materia son una frase que contiene el tema que se quiere referenciar; los términos o las palabras que expresan una idea de una disciplina determinada, son los que pueden ser términos compuestos conformados por una unión de letras y además cuentan con un significado.
- Un extractor de datos es un conjunto de motores de extracción que permite obtener mediante la extracción los datos de las bases de datos y utilizar programas de visualización.
- Es decir permite la extracción automática de entidades o conceptos del contenido de texto visible de un elemento y asignarlos a una propiedad administrada. Al mismo tiempo, estas propiedades se pueden usar para restringir las consultas mediante filtros de propiedades o como opciones de refinamiento de consultas.
- El proceso de obtención de la información comienza por la selección de las fuentes utilizables, es decir, los lugares en los que se obtiene la información útil para poder realizar las diferentes tareas con ella.

- Según la definición de la Real Academia Española, los siguientes términos deben entenderse del siguiente modo:
 - Agrupar: es reunir en grupo o apiñar, constituyendo una agrupación.
 - Transformar: hacer mutar algo en otra cosa.
 - Homogeneizar: hacer homogéneo, por medios físicos o químicos, un compuesto o mezcla de elementos diversos.

Preguntas de Autoevaluación

1. ¿Cómo se denomina el proceso que consiste en almacenar los datos en la base de datos final?

- a) Extracción.
- b) Transformación.
- c) Carga.

2. Indica si es verdadero o falso el siguiente enunciado:

"La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación".

- a) Verdadero.
- b) Falso.

3. ¿Qué técnica se utiliza en la fase de identificación?

- a) Herramientas ETL.
- b) Experiencia del experto.
- c) Data Warehouse.

4. ¿Qué tipo de estructura se relaciona con la forma en que están contenidos los datos en la maquina?

- a) Lógica.
- b) Física.
- c) Mixta.

5. ¿Qué tipo de información se caracteriza por ser la propia de la empresa?

- a) Externa.
- b) Confidencial.
- c) Interna.

UD3 Herramientas de obtención de información



UF1890 Desarrollo de componentes de software y consultas dentro del sistema de almacén de datos

1. Herramientas de visualización y difusión

Normalmente se suele contar con una gran cantidad de información almacenada. Si está información se observase de forma lineal supondría un largo y tedioso proceso, por lo que mediante las herramientas de visualización lo que se pretende es realizar dicha tarea de una forma más fácil, permitiendo la comprensión e interpretación de la información.

Pero no basta con tener procesos de visualización, para garantizar unos resultados adecuados es necesario tener buenas herramientas de visualización que permitan trabajar los datos con calidad y de un modo adecuado. Es por tanto que no se duda acerca de la importancia que tiene el procesamiento adecuado de la información antes de llevar a cabo el tratamiento de visualización.

El tratamiento de los datos y su visualización forma un tandem atractivo para el usuario que lo solicita, cada vez con mayor frecuencia, para poder interpretar datos de forma ágil y rápida.

Un exceso de información conlleva a una pérdida de interés en las audiencias que se traduce en la mayoría de los casos, en una falta de comprensión de la información expuesta.

Para evitarlo se producen los procesos de visualización de los datos que permite realizar un diseño comprensivo.

Cuando la visualización es de muchos datos el proceso se complica. Además de analizar los datos, saber interpretarlos, contrastarlos con otros datos y estudiarlos, se hace necesario saber comunicarlos. Y algo aun más importante, es saber ubicar la información, dotándola de un contexto y compararlos con algo.

Para poder realizar esta tarea de visualización de un modo más sencillo, y asequible, se cuenta con herramientas de visualización, algunas de ellas gratuitas. Algunos ejemplos de estas herramientas son las siguientes:

- **Tableau:** es de fácil uso y tiene grandes funcionalidades. Cuenta con una versión de pago y una versión libre. Pero presenta una limitación, y es que no permite trabajar con grupos amplios de datos.

- **Weave:** es una potente herramienta de visualización. Permite graficar cualquier tipo de dato ajustándose casi por completo a las necesidades de la persona.
- **Gephi:** es considerada una de las mejores herramientas y permite realizar grafos dinámicos, jerárquicos de forma sencilla.
- **Many Eyes:** tiene muchísimas opciones de personalización y permite compartir las creaciones.
- **NodeXL:** permite trabajar con Excel, y es bastante útil para el trabajo de análisis de redes y estructuras.
- **Data-Driven Documents:** ayuda a manipular documentos completos basados en datos. Usa HTML, SVG y CSS para crear las diferentes visualizaciones. Combina potentes componentes que permiten crear interacciones de un modo sencillo y la integración con CSS hace que sea bastante adaptable a cualquier proyecto web.
- **Axiis:** es uno de los más conocidos y al mismo tiempo uno de los más antiguos. Además de todas las opciones que presenta se caracteriza, porque permite utilizar aquellas que se comparten ampliamente en Internet.
- **Google Fusión Tables:** es la herramienta de visualización de datos propia de google. Es una de las más recientes, solo hace falta una cuenta de google para poder trabajar con esta herramienta. Permite compartir datos de una forma ampliamente abierta y construir gráficos personalizados de difusión.

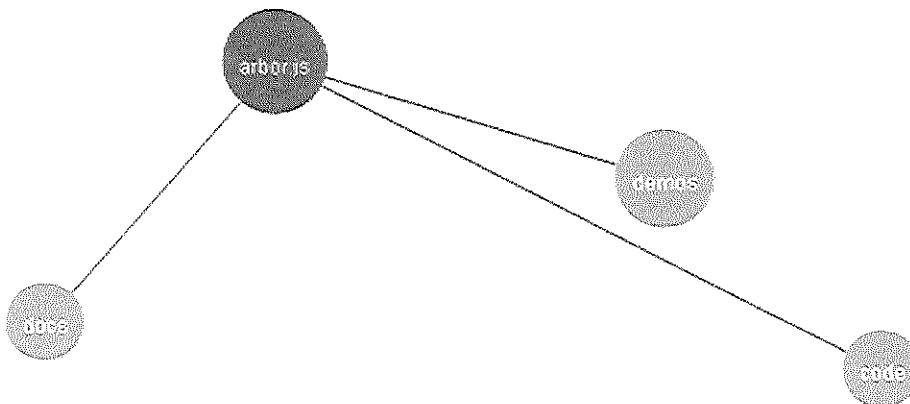
Una de las mayores fortalezas de la visualización de datos tiene que ver con nuestra capacidad de procesar la información visual y grandes volúmenes de datos mucho más rápido que la información textual o verbal. Utilizar buenas herramientas para la visualización de datos es la clave para el éxito del uso de técnicas y tecnologías apropiadas. Se amplía el horizonte de nuestro pensamiento en nuevos campos de interesantes análisis.

Para poder crear una visualización de datos adecuada, es importante conocer y entender las herramientas disponibles y su correcta aplicación en los campos relacionados.

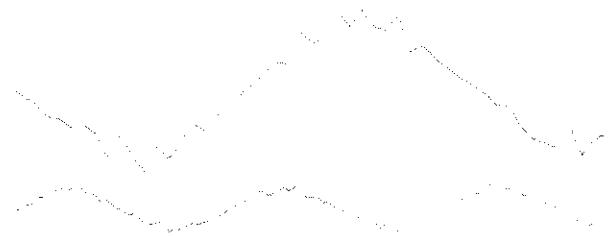
En la actualidad, existen varias bibliotecas de software libre que emplean javascript, que tiene reconocidos beneficios en cuanto a su utilización.

- **Arbor.js:** es una biblioteca adecuada que gestiona de un modo eficaz abstracciones gráficas de organización y manejo de actualizaciones. Cubism.js y Rickshaw son otras que están ayudando a crear gráficos de línea y en tiempo real. Crean interacciones flexibles y eficientemente complicadas, apoyadas en el formato vector SVG para su visualización.

arbor.js a graph visualization library using web workers and jQuery



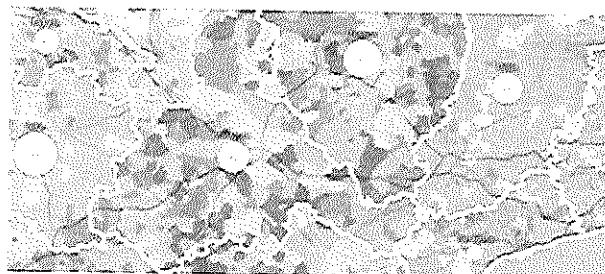
- **Envisión.js:** es muy popular, ya que crea visualización rápida y dinámica de datos en tiempo real. Es muy utilizada cuando se tratan los segmentos de información financiera. En la biblioteca de javascript, se proporcionan herramientas que permiten crear visualizaciones de datos interactivos para la web. En la fabricación de mapas interactivos basados en piezas de escritorio y páginas móviles amigables, se emplea la biblioteca Leaflet JavaScript o Modest Maps, ambos ofrecen una gran comunidad de alcance.



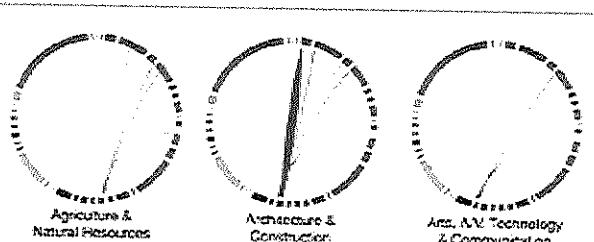
- **Polymaps:** es empleado para realizar mapas dinámicos e interactivos. Esta biblioteca tiene el propósito de dirigirse de un modo específico al público interesado en la visualización de datos, por ejemplo en los datos concernientes a mapas calientes con información estadística por zonas geográficas, siendo esta su mejor referencia.



- **CartoDB:** es un servicio web que permite la conversión de datos CSV en visualizaciones en diferentes ejes, o bien agrupar los datos modificando su tamaño o su color, y plasma toda esta información en un mapa. Su uso no es complejo.



- **Circos:** es un software que permite la visualización de datos circulares, que nos permite interrelacionar datos y crear hermosas visualizaciones. Este formato es bastante conocido y se emplea por profesionales en la visualización de datos.



La selección de un tipo u otro de herramienta de visualización será en función de las necesidades que se presentan en el proceso de presentación de la información.

Para crear un buen análisis y visualización de datos, lo más importante es conocer y entender las herramientas disponibles y su correcta aplicación en los campos relacionados.

Existen muchas herramientas para ayudar a transformar los datos en gráficos, pero éstas pueden conllevar un alto coste.

A continuación se describen algunas de las herramientas de visualización gratuitas clasificadas de acuerdo a sus características tecnológicas:

1.1. Aplicaciones de visualización genéricos

Herramientas que ofrecen diversas opciones de visualización. Aunque algunas siguen apostando por las tablas y gráficos convencionales, muchas otras abogan por ofrecer nuevas opciones tales como diagramas de árbol y nubes de palabras.

Google Fusión Tables

Aplicación web que permite organizar, gestionar, colaborar, visualizar y publicar datos en la web de una manera sencilla.

Gestiona grandes colecciones de datos que deben estar normalizados y guardados en un archivo Excel, .ods, .csv o .kml.

Permite visualizar los datos mediante gráficos circulares, gráficos de barras, diagramas de dispersión y líneas de tiempo; así como mapas geográficos basados en Google Maps.

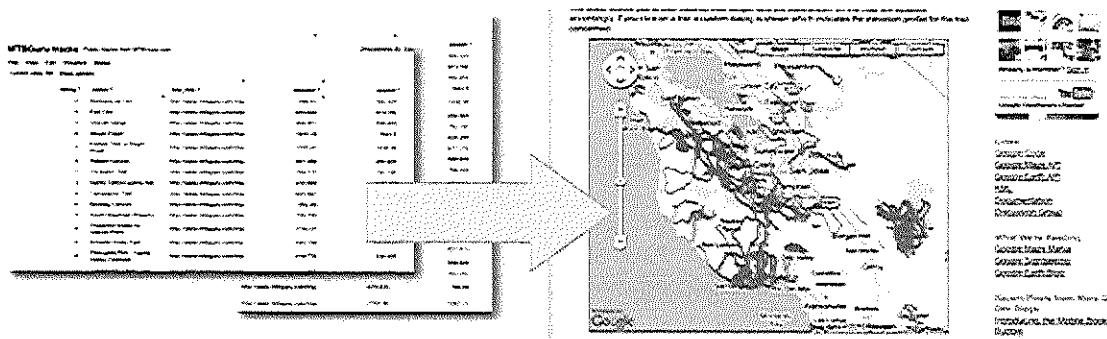
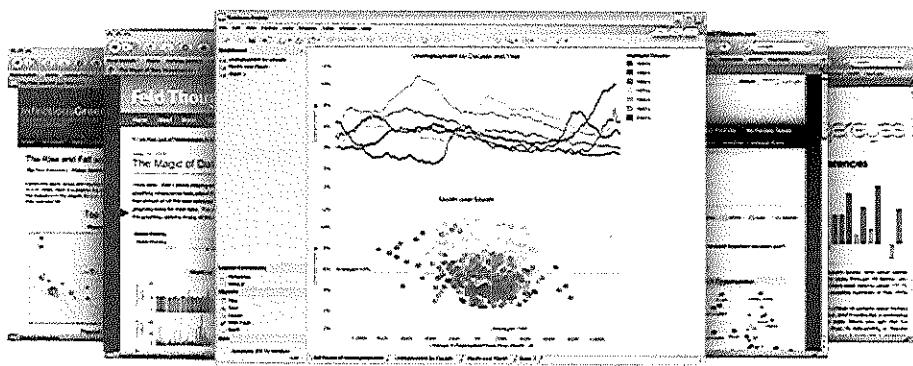


Tableau Public

Herramienta gratuita de visualización de datos mediante gráficos que combina algunos elementos tradicionales de las herramientas de business intelligence como puedan ser el modelo de organización de variables mediante el uso de dimensiones y medidas o la conexión con otros sistemas de gestión de información como las bases de datos o las hojas de cálculo con el uso de un interfaz gráfico atractivo, eficiente y rápido.

Algunas de las características más relevantes de esta herramienta son:

- Detección de datos rápida y fácil. Permite trabajar con bases de datos y hojas de cálculo de cualquier tamaño. Acepta formatos como Excel, Access, y formatos de texto.
- Trabaja con una gran variedad de gráficos: fiebres, barras, barras apiladas, tartas, ta-blas, mapas con polígonos, líneas o puntos, etc.
- Publicación de gráficos interactivos.
- Combinación de diversas fuentes de datos en una sola vista.
- Los datos son públicos.
- Descarga de datos en crudo desde las mismas visualizaciones.



Many eyes

Aplicación web para crear, compartir y discutir la representación gráfica de los datos de usuario cargados.

Esta herramienta de visualización de datos gratuita ha sido puesta a disposición de los usuarios por parte de la empresa IBM.

Many Eyes permite compartir las visualizaciones creadas fomentando las conversaciones alrededor de una visualización y proponiendo otros enfoques a partir de los mismos datos. Se trata de una herramienta de uso público, es decir, que todos los datos y visualizaciones que se realicen estarán a disposición del resto de usuarios; no se puede usar de forma privada.

Permite realizar gran cantidad de tipos de visualizaciones:

- Relaciones entre puntos (Scatterplot, matriz de gráficos y diagramas de red).
- Comparar valores (gráficos de barras, histogramas y gráficos de burbuja).
- Traceo de cambios de tendencia en el tiempo (gráficos de líneas, barras y barras por categorías).
- Ver partes de totales (gráficos de queso, mapas de secciones simples y con comparaciones).
- Analizadores de texto (árbol de conceptos, nube de etiquetas, relaciones de frase y generador de nubes de palabras).
- Gráficos geográficos (gráficos sobre mapas).

CartoDB

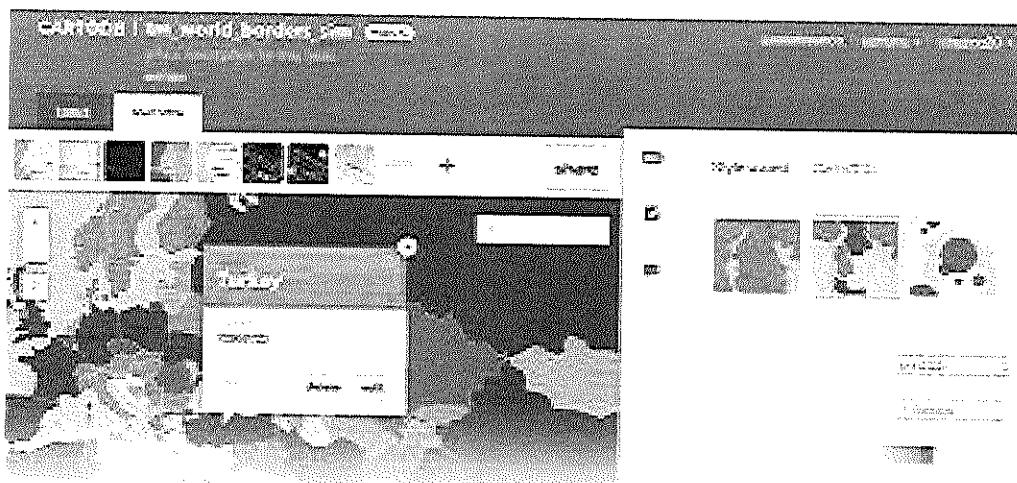
Base de datos geoespacial en la nube, que funciona con los servicios web de Amazon por detrás, permitiendo la escalabilidad, la flexibilidad y la elasticidad de sus servicios. Es un proyecto Open Source que también se ofrece como un servicio bajo demanda.

La finalidad de CartoDB es facilitar la creación de aplicaciones geolocalizadas y la creación de mapas. Permite diseñar y desarrollar mapas en tiempo real que funcionan en todas las plataformas web y móviles.

Entre sus características podemos destacar:

- Diseño de mapas: Para sus capas de datos, es posible utilizar CartoCSS para editar fácilmente el formato y la apariencia con la que se generan.
- Integración con otros servicios cartográficos (Google Maps, MapBox): CartoDB produce las capas de datos, y para la capa de mapa utiliza GoogleMaps y, desde la versión 2.0 MapBox. Estos mapas incluyen las funciones básicas de zoom, desplazamiento, etc.
- Integración con otras librerías: CartoDB cuenta ya con varias librerías a su alrededor que permiten extender su uso o integrar otros servicios.
- Geocodificación: Es posible obtener información geográfica a partir de otros elementos distintos de las coordenadas.

- Capacidad de importar datos fácilmente: CartoDB permite introducir datos directamente en las tablas a partir de su panel, añadir datos vía SQL o lectura desde URLs, pero también se pueden importar colecciones de datos directamente en múltiples formatos.
- Realizar peticiones SQL con componentes espaciales: Gracias al uso de PostGIS, CartoDB permite consultar y combinar conjuntos de datos utilizando los datos geoespaciales para realizar la combinación.
- Tablas públicas y privadas: Como corresponde a un servicio cloud, CartoDB permite diferenciar entre un uso de las tablas público, o un uso privado.



Está orientado a desarrolladores sin experiencia en sistemas de información geoespacial, con una interfaz muy amigable.

Entre sus clientes se cuentan diversas instituciones tales como ONU, Google, la NASA, la Universidad de Oxford, de Yale, entre otros.

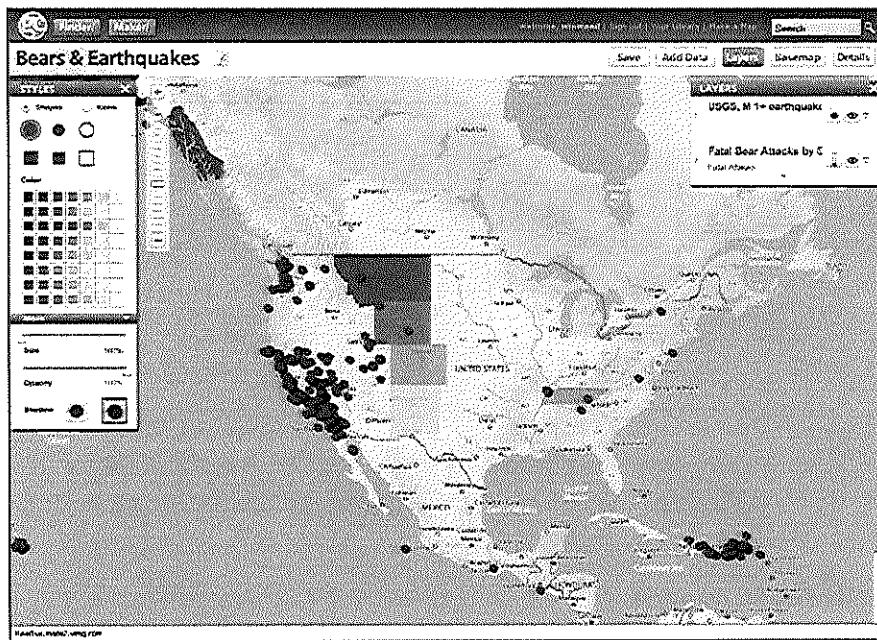
GeoCommons

Plataforma geoespacial de gestión de datos, visualización, creación de mapas y análisis espacial.

Admite la carga de datos desde distintos tipos de fuentes de datos: hojas de cálculo, archivos KML, shape, servidores de bases de datos con soporte espacial, servicios OGC como WMS y TMS, o del repositorio público de la plataforma.

Las técnicas de representación cartográficas que permite son mapas de coropletas, de símbolos proporcionales clasificados en intervalos y color aplicado a

simbología puntual. También destacan las funcionalidades de animación temporal. Los mapas se pueden exportar a formato KML y los datos a formato KML, hoja de cálculo o shape, entre otros, e incrustar en una página web.



1.2. Wizards, librerías, API

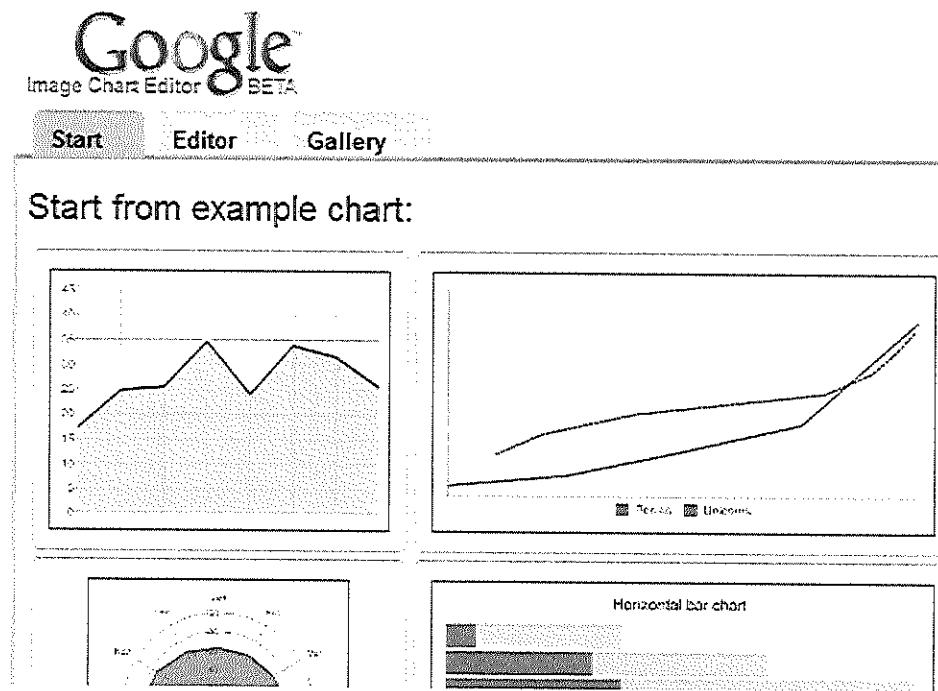
Amplia gama de librerías y APIs disponibles para ayudar al desarrollador a crear sus propias visualizaciones.

Google Chart Tools

Herramienta de Google Developers que permite la creación de gráficas en forma de imágenes PNG. Su funcionamiento se basa en peticiones http a una determinada url.

Es de uso gratuito pero con ciertas limitaciones. Inicialmente, su uso estaba limitado a 50.000 peticiones por url y día, pero, actualmente este límite se sitúa en 250.000. Para evitar esta limitación, almacenar las imágenes generadas en un servidor propio a modo de cache de imágenes.

Dispone de una gran variedad de tipos de gráficas, los cuales vienen dados como clases de JavaScript. Una de las ventajas que tienen este sistema de generación de gráficas es que no se necesita instalar ningún componente en nuestro entorno o servidor, por lo que se puede generar cada gráfica "al vuelo".



JavaScript Infovis Toolkit

Biblioteca de JavaScript que proporciona herramientas para crear visualizaciones de datos interactivas en aplicaciones web (mapas estratégicos, árboles jerárquicos, mapas relacionales, etc.). Debido a su gran diversidad de representaciones, se adapta a cualquier necesidad del desarrollador.

Algunas de las características más relevantes de esta librería son:

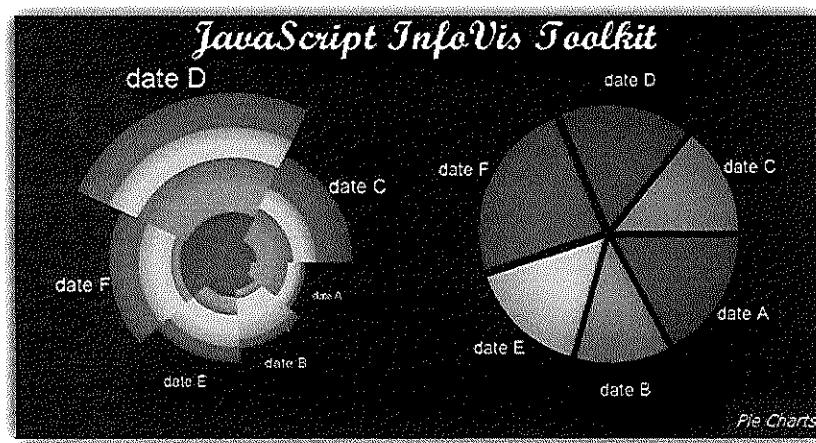
- Posee diferentes tipos de representaciones de datos.
- Permite interactuar con los datos en tiempo real.
- Compatible con la mayoría de navegadores.
- Recurso Open Source de fácil integración en desarrollos web.
- Extensible.
- Permite combinar las visualizaciones para crear nuevas formas de visualización.
- Gran velocidad de proceso para estructuras complejas.

Desde el punto de vista técnico, la representación de los datos a mostrar viene marcada por una estructura JSON (JavaScript Object Notation) un formato ligero de intercambio de datos, el cual se basa en dos estructuras: una colección de pares de nombre/valor (objeto, registro, estructura, diccionario, tabla hash, etc.) y

una lista ordenada de valores (vectores, listas o secuencias), estas son estructuras universales y permite a todos los lenguajes de programación adaptarse con facilidad.

Los casos de uso o posibilidades de esta librería son innumerables:

- Desarrollo en entornos BI (Business Intelligence).
- Representación de organigramas.
- Mapas estratégicos en cuadros de mando (Balanced Scorecard).
- Mapas estadísticos de datos.
- Mapas relacionales.



D3.js

Librería de JavaScript que permite crear visualizaciones complejas y gráficos interactivos. Básicamente, la librería permite manipular documentos basados en datos usando estándares abiertos de la web; y los navegadores pueden crear visualizaciones complejas sin depender de un software propietario. Sus desarrollos son abiertos y pueden ser reimplementados por otros desarrolladores. Sus posibilidades son tan amplias como la geometría misma (burbujas, diagramas Chord, links de nodos,...)

D3 permite enlazar datos al DOM (Modelo en Objetos para la Representación de Documentos) y aplicar transformaciones. Por ejemplo, generar una tabla HTML a partir de una serie de números. O bien utilizar los mismos datos para crear un gráfico interactivo SVG con transiciones e interacción.

Protovis

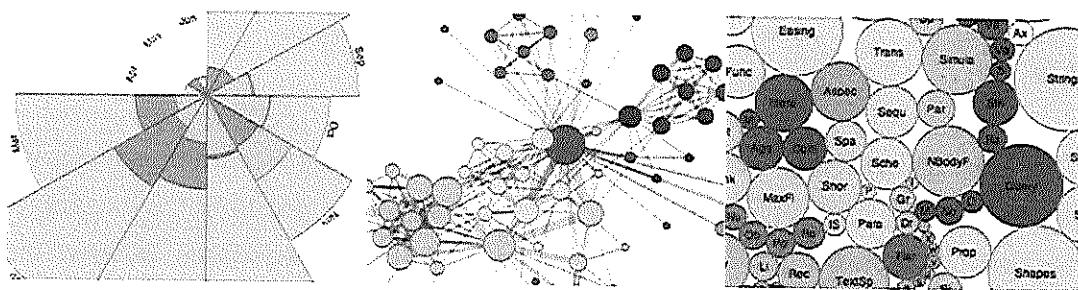
Librería gráfica JavaScript orientada a la realización de visualizaciones.

Proporciona al desarrollador un gran conjunto de componentes y herramientas, y otorga la posibilidad de personalizar las visualizaciones con un control detallado.

Algunas de las características más relevantes de esta librería son:

- Versatilidad prácticamente ilimitada. Se basa en el framework de la gramática de los gráficos.
- Configuración de gráficos sencillo, basado en el método de encadenamiento.
- Enfocada a los gráficos estadísticos, su método de desarrollo permite, además, su uso para visualizaciones más bien estructuradas y basadas en los datos.
- Incorpora algunas funciones estadísticas como preparación de los datos.

El principal inconveniente que presenta es que Protovis es una biblioteca pesada (pesa más de 700 Kb), pensada para Intranets o conexiones rápidas.

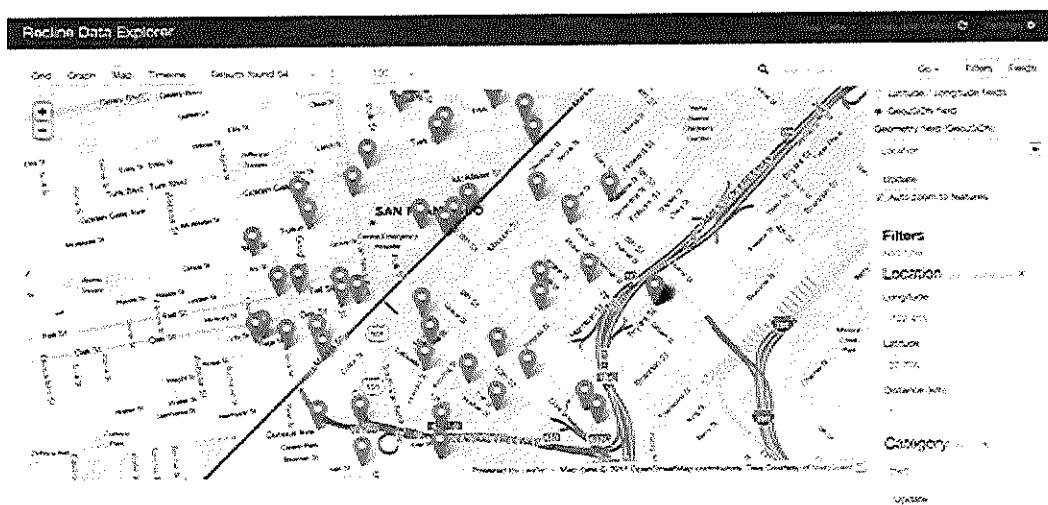


Recline.js

Biblioteca para el desarrollo de aplicaciones basadas en HTML y JavaScript. Diseñada para la integración, por lo que es fácil de integrar en otros sitios web y aplicaciones. Orientada a desarrolladores sin grandes conocimientos de programación, que utiliza una interfaz sencilla para la vista (y edición) de datos. Las visualizaciones se ofrecen en modo gráfico, mapa y líneas de tiempo.

Recline funciona sobre Backbone, esta estructura provee un excelente soporte para la construcción de aplicaciones que manejan importantes cargas de datos, utilizando modelos para la gestión de la información y vistas para mostrarlas. Además, resulta fácilmente extensible a través de nuevos Backends que permiten conectar una base de datos o capa de almacenamiento.

Esta biblioteca cuenta con muchas funciones para la manipulación de bases de datos, incluida su carga, consulta y manipulación. Consta de soporte para cargar datos de archivos CSV, Excel, Google Docs, ElasticSearch, CouchDB y DataHub entre otros.



Provista de mecanismos de limpieza y actualización de datos mediante un sencillo script.

La biblioteca Recline consiste en tres módulos:

- **Modelo:** definición de la estructura de los datos, (por ejemplo: definición del dataset a utilizar según origen y tipo de datos).
- **Backend:** conexión de los datos mediante el API de Recline.js directamente con el origen de datos, que puede ser una base de datos, un archivo separado con comas, etc.
- **Vistas:** muestra de la información obtenida y gestión en las dos instancias anteriores.

1.3. Herramientas de visualización geoespacial

Herramientas para la representación de datos geográficos.

OpenHeatMap

Aplicación web capaz de convertir datos estadísticos de hojas de cálculo en mapas térmicos.

Su funcionamiento es sencillo y soporta diferentes formatos de archivos como fuente: Excel, CVS o documentos vinculados desde Google Docs.

Los ficheros deben tener un formato concreto con una columna que indique la dirección o posición geográfica de cada uno de los datos para poder posicionarlos.

OpenHeatMap permite compartir los mapas a través del correo o redes sociales, o también embeberlo en una página web.

OPENHEATMAP

- 1 - Upload your spreadsheet
- 2 - Get an interactive online map in seconds

Title:

Author: <http://>

Time: -

Key: - tc

Style: Red

Transparency:

Size:

Save & view

OpenLayers

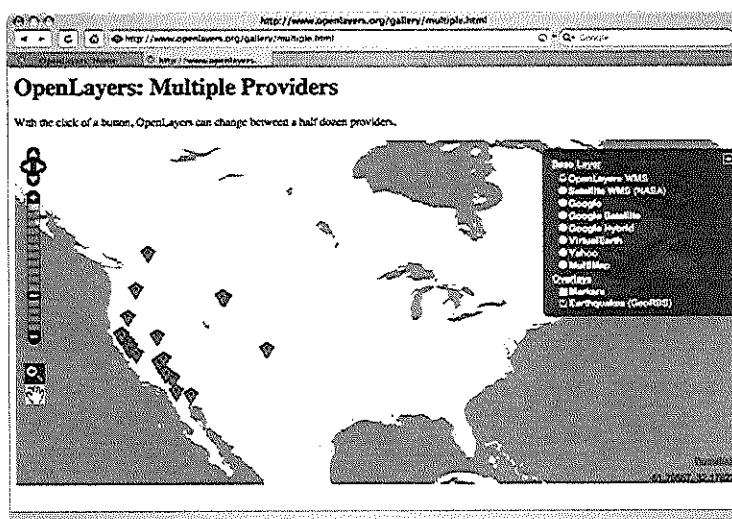
Biblioteca de JavaScript de código abierto que permite la inclusión de un componente tipo mapa en cualquier página web, con georeferencias.

Es una librería del lado del cliente, un visor de mapas en javascript, por lo que la descarga de estos se realiza directamente desde el navegador a través de

Ajax. No genera tráfico en el servidor, los mapas se descargan directamente del servidor de mapas.

OpenLayers permite sobreponer distintas capas sobre una básica, añadir indicadores o puntos en el mapa con leyendas, así como polígonos y proporciona su propio API para dibujarlos de una manera sencilla.

Incorpora un set de controles básicos y una toolbar de controles avanzados y permite incluir los controles necesarios haciendo uso del API.



OpenStreetMap

Proyecto colaborativo de creación de mapas libres y editables.

Los mapas se crean utilizando información geográfica capturada con dispositivos GPS móviles, ortofotografías y otras fuentes libres. Esta cartografía, tanto las imágenes creadas como los datos vectoriales almacenados en su base de datos, se distribuye bajo licencia abierta Open Database License (ODbL).

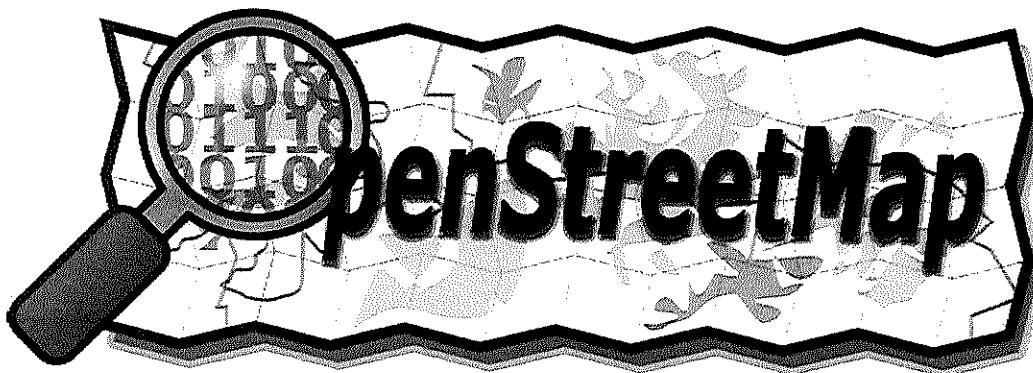
Los usuarios registrados pueden subir sus trazas desde el GPS y crear y corregir datos vectoriales mediante herramientas de edición creadas por la comunidad OpenStreetMap.

OpenStreetMap utiliza una estructura de datos topológica. Los datos se almacenan en el datum WGS84 lat/lon (EPSG:4326) de proyección de Mercator. Los elementos básicos de la cartografía OSM son:

- **Los nodos (nodes):** puntos que recogen una posición geográfica dada.

- **Las vías (ways):** lista ordenada de nodos que representa una polilínea o polígono (cuando una polilínea empieza y finaliza en el mismo punto).
- **Las relaciones (relations):** grupos de nodos, caminos y otras relaciones a las que se pueden asignar determinadas propiedades comunes. Por ejemplo, todas aquellas vías que forman parte del Camino de Santiago.
- **Las etiquetas (tags):** se pueden asignar a nodos, caminos o relaciones y constan de una clave (key) y de un valor (value). Por ejemplo: highway=trunk

Los atributos de los datos siguen un modelo más elaborado que las fórmulas de indexación social. La ontología de las características del mapa (principalmente el significado de las etiquetas) se mantiene mediante una wiki.

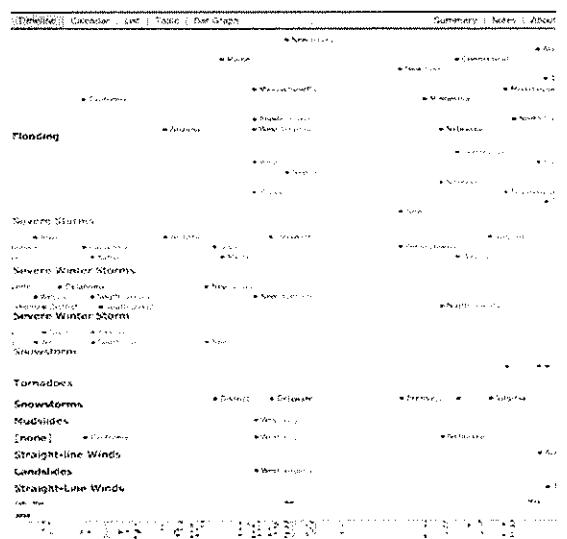


1.4. Herramientas de visualización de datos temporales

Herramientas para el análisis de datos en los que el tiempo es un componente importante.

TimeFlow

Herramienta de visualización para datos temporales. La versión actual es "alfa", por lo que puede contener errores.



Esta herramienta ayuda a analizar los datos temporales a través de cinco vistas diferentes:

- Vista Línea de Tiempo.
- Vista Calendario.
- Vista Diagrama de barras.
- Vista Tabla.
- Vista Lista.

RECUERDA

- Normalmente se suele contar con una gran cantidad de información almacenada. Si está información se observase de forma lineal supondría un largo y tedioso proceso, por lo que mediante las herramientas de visualización lo que se pretende es realizar dicha tarea de una forma más fácil, permitiendo la comprensión e interpretación de la información.
- Pero no basta con tener procesos de visualización, para garantizar unos resultados adecuados es necesario tener buenas herramientas de visualización que permitan trabajar los datos con calidad y de un modo adecuado. Es por tanto que no se duda acerca de la importancia que tiene el procesamiento adecuado de la información antes de llevar a cabo el tratamiento de visualización.
- Las aplicaciones de visualización genéricas son herramientas que ofrecen diversas opciones de visualización. Aunque algunas siguen apostando por las tablas y gráficos convencionales, muchas otras abogan por ofrecer nuevas opciones tales como diagramas de árbol y nubes de palabras.
- Amplia gama de librerías y APIs disponibles para ayudar al desarrollador a crear sus propias visualizaciones.
- Las herramientas de visualización geoespacial son aquellas que se utilizan para representar datos geográficos.
- Las herramientas de visualización de datos temporales son aquellas que permiten analizar datos en los que el tiempo es un componente importante.

Preguntas de Autoevaluación

1. ¿Qué herramienta se caracteriza por ser de fácil uso y tener grandes funcionalidades?

- a) Tableau.
- b) Axiis.
- c) Data tableau.

2. Indica si es verdadero o falso el siguiente enunciado:

"La selección de un tipo u otro de herramienta de visualización será en función de las necesidades que se presentan en el proceso de presentación de la información".

- a) Verdadero.
- b) Falso.

3. ¿Qué herramientas de visualización de datos mediante gráficos que combina algunos elementos tradicionales de las herramientas de business?

- a) Google Fusión Tables.
- b) Tableau Públic.
- c) Many eyes.

4. ¿Qué herramienta de visualización se caracteriza por ser una base de datos geoespacial en la nube, que funciona con los servicios web de amazon?

- a) CartoDB.
- b) Geocommons.
- c) Tableau public.

5. ¿Qué posibles usos presenta JavaScript Infovis Toolkit? Selecciona las respuestas correctas.

- a) Mapas estratégicos en situaciones de mando (Balanced Scorecard).
- b) Mapas estadísticos de datos.
- c) Mapas relacionales.

Actividades Prácticas



**UF1890 Desarrollo de
componente software
y consultas dentro del
sistema de almacén de
datos**

Actividad Práctica RP3

Para la elaboración de esta actividad práctica se atenderá a lo indicado en el Real Decreto 1531/2011, de 31 de octubre, en lo correspondiente a la UC1215_3 (Realizar y mantener componentes software en un sistema de planificación de recursos empresariales y de gestión de relaciones con clientes).

El ejercicio que se expone a continuación, se corresponde con la realización profesional RP3 (Desarrollar componentes y consultas dentro del sistema de almacén de datos (data warehouse) para almacenar y recopilar información (data mining) de acuerdo a especificaciones de diseño establecidas). Y concretamente con la CR3.3 (Los extractores de información sobre el sistema de almacén de datos se generan e integran para extraer la información necesaria de forma eficiente, siguiendo especificaciones técnicas, según normas de la organización y cumpliendo la legislación vigente sobre protección de datos).

Ejercicio

Se propone al alumnado que realice la siguiente actividad seleccionando o creando previamente una base de datos, y que realice la representación esquemática para posteriormente poder realizar la exemplificación de una posible extracción de información.

Para ello se aconseja al alumno que comprenda y represente las diferentes etapas en el proceso de tratamiento de la información, ubicando lo más precisamente posible el momento de la extracción de la información, así como los pasos ha seguir, y aquellos aspectos o consideraciones indispensables para que todo el proceso se realice de forma adecuada.

Se tendrá en cuenta la comprensión, síntesis del tema y la originalidad en la realización de esta actividad.

Respuestas a las Preguntas de Autoevaluación

UNIDAD DIDÁCTICA 1		UNIDAD DIDÁCTICA 2		UNIDAD DIDÁCTICA 3	
1	A	1	C	1	A
2	A, B	2	A	2	A
2	C	3	B	3	B
4	A, B	4	B	4	A
5	A	5	C	5	B, C



