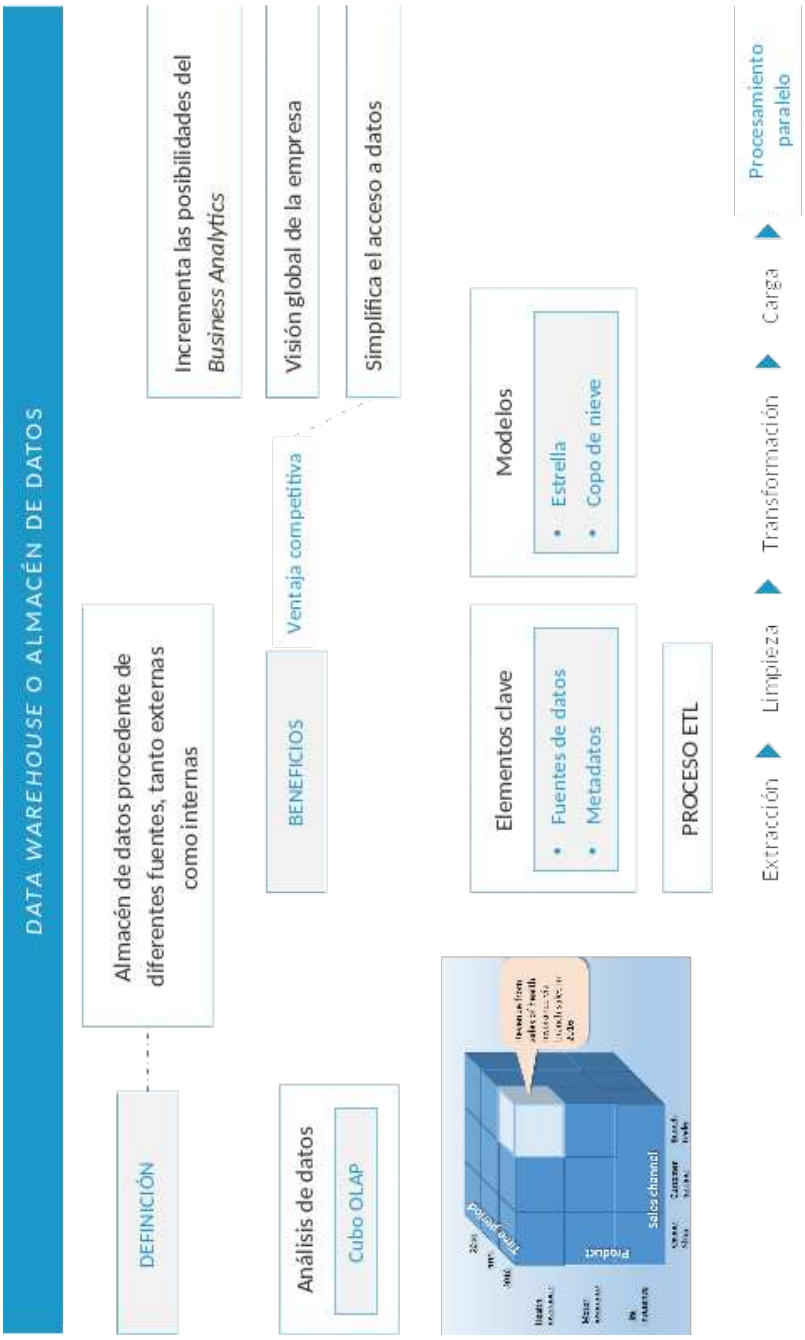
Estrategia y Gestión Empresarial Basada en Análisis de Datos

Tema 5. Data warehouse o almacén de datos

Índice

Esquema



5.1. Introducción y objetivos

La utilización de un *data warehouse* (a poder ser con datos en tiempo real), el sistema de apoyo a la toma de decisiones y los instrumentos de *business analytics* configuran los tres vectores clave sobre los que apoyar una gestión basada en datos o un sistema de *business intelligence*.

Un ***data warehouse*** no es más que el contenedor de los datos procedentes de las diferentes fuentes de la organización una vez que son integrados, depurados y estructurados en una única base de datos centralizada, que permite tener listos para la analítica de negocio (ya sea para su explotación o para el *reporting*). Por tanto, es el fruto de una decisión del sistema de apoyo a la toma de decisiones que considera que el disponer de un repositorio de datos actuales e históricos es crucial para la gestión estratégica de la empresa.

En ocasiones, debido a la complejidad de la organización o por razones operativas, es necesario crear subconjuntos de datos más pequeños para departamentos o áreas específicas de la empresa. Son los llamados ***data mart***, que pueden ser independientes (si los datos son tomados directamente por el área del *data mart*) o dependientes (si utilizan los datos del *data warehouse*):

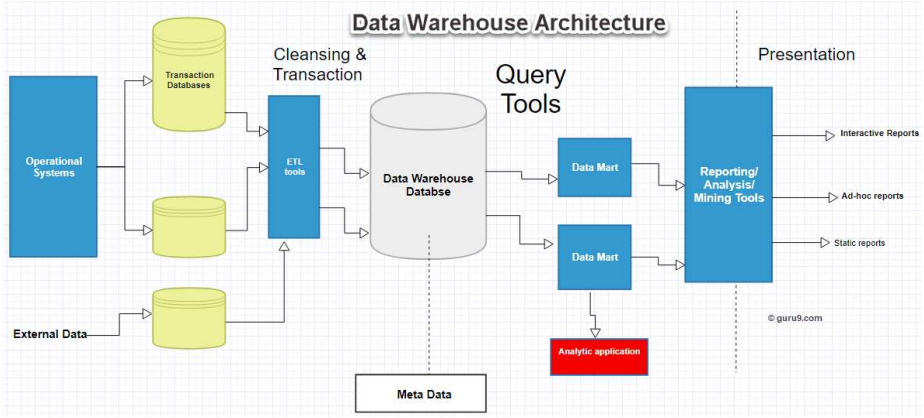


Figura 1. Arquitectura de un *data warehouse*. Fuente: https://towardsdatascience.com/discerning-data-

Ideas clave

warehouses-an-introduction-to-databases-and-warehousing-cc9fc10f3283

Desde un punto de vista teórico, estos almacenes de datos pueden ser **orientados a objetos** (es decir, datos organizados por *subjects* como ventas, productos o clientes, entre otros) o **integrados**, en los que los datos proceden de diferentes fuentes y deben ser integrados de manera consistente.

Entre los **beneficios** de la disposición de un *data warehouse* se deben incluir, al menos, los siguientes:

- Expande las posibilidades del *business analytics*.
- Permite obtener una mejor visión global de la corporación, así como una mejor y más actualizada información.
- Simplifica el acceso a los datos.
- Se transforma en una ventaja competitiva.
- Facilita la toma de decisiones.

En este tema se analizarán los principales elementos y herramientas de extracción, transformación y carga (**ETL** de sus siglas en inglés *extract*, *transform and load*) de datos.

Los **objetivos** que se pretenden conseguir en este tema son los siguientes:

- Comprender en qué consiste el sistema de almacenamiento de datos.
- Identificar los diferentes modelos de *data warehouse*.
- Entender los procesos ETL.
- Conocer las herramientas de procesamiento analítico en línea (OLAP, por sus siglas en inglés).

Ideas clave

5.2. Almacén de datos o data warehouse (DW)

Como vimos en el tema dedicado a los sistemas de información, el *data warehouse* o almacén de datos es un elemento clave del sistema de apoyo a la toma de decisiones. De manera ideal, este almacén ha de ser **accesible, de fácil gestión y flexible** para responder a las necesidades de los usuarios del sistema de apoyo quienes requieren de una información de calidad y disponible para, tras su análisis, ser usada en la toma de decisiones. Es decir, esta información ha de estar depurada, ser consistente y, a poder ser, generar series lo bastante largas y con la suficiente periodicidad y frecuencia como para permitir análisis estadísticos rigurosos.

El sistema de almacenamiento habrá de extraer información de diferentes sistemas, incluyendo fuentes externas, para almacenarla en un entorno integrado, que ha de permitir contextualizar el análisis de la información y relacionarla dentro de la organización.

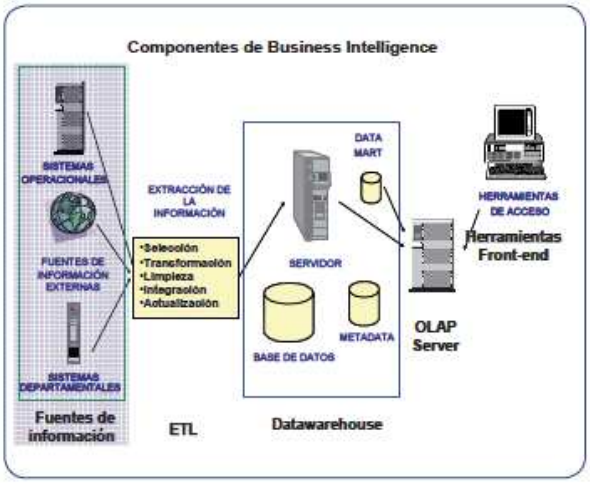


Figura 2. Componentes del *data warehouse*.Fuente: Cano (2007, p. 95).

Pasamos a continuación a analizar dos de los elementos claves del *data warehouse*: las fuentes de datos y los metadatos.

Ideas clave

Fuentes de datos

Los datos pueden proceder de la actividad propia de la empresa o de fuentes externas. Estas fuentes externas que nutren el sistema de información del *data warehouse* pueden ser adquiridas a otras empresas (es el caso habitual de las empresas que capturan y almacenan información de consumidores y hogares para la investigación comercial) o pueden proceder de la labor de agencias oficiales de estadística o, incluso, de información contenida en Internet o en redes sociales.

Los metadatos o metadata

Los metadatos son un componente esencial de un sistema de almacenamiento de datos. El metadato es algo así como el repositorio de toda la información incorporada en el *data warehouse* o *data mart*: descripciones y significado, variables, atributos, formatos, cantidad, características, máximos y mínimos. En definitiva, un metadato contiene la sintaxis, estructura y semántica de los datos almacenados.

Por otra parte, debemos conocer que la construcción de un *data warehouse* se acomoda a diferentes modelos o estructuras. Las más conocidas, por la frecuencia de uso, son las correspondientes al modelo estrella y al llamado copo de nieve. En empresas de gran dimensión y en presencia de múltiples *data marts* (subconjuntos de datos de áreas específicas), es importante analizar las dimensiones y las necesidades de los diferentes usuarios para garantizar un único diseño de los *data marts* para facilitar la integración y uniformidad de la información.

El modelo estrella

El modelo estrella contiene una tabla central, la llamada tabla de hechos y varias tablas asociadas a esta que capturan diferentes dimensiones, de forma que estas tablas de dimensiones no tienen información relacionada entre sí, sino tan solo con la llamada tabla de hechos.

Lo más representativo de la arquitectura de estrella es que solo existe una tabla de

dimensiones para cada dimensión. Esto quiere decir que la única tabla que tiene relación con otra es la de hechos, esto es, que toda la información relacionada con una dimensión debe estar en una sola tabla. En la figura 3 se observa un ejemplo de este modelo.

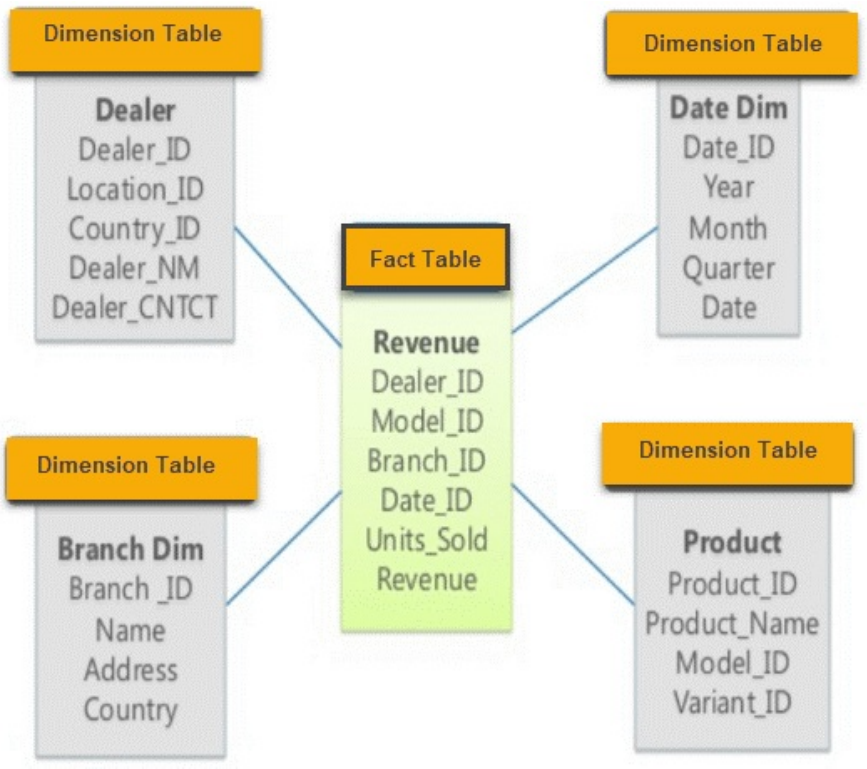


Figura 3. Ejemplo de modelo estrella.Fuente: https://www.guru99.com/star-snowflake-data-warehousing.html

Supongamos que se trata del diseño de un *data warehouse* de ventas, en el que los hechos son las ventas. Cada venta de la tabla de hechos queda configurada por un identificador de producto, una sucursal, un vendedor y una fecha.

Se trata pues de un modelo multidimensional y no normalizado que permite un análisis simple y rápido de un análisis multidisciplinar y con buen rendimiento, ya que permite indexar las dimensiones de manera individual.

El modelo copo de nieve

En esta estructura (figura 4), la tabla de hechos ya no es la única que se relaciona con las tablas de dimensiones, dado que hay tablas de dimensiones que se relacionan entre sí y no tienen relación directa con la de hechos. Si la información necesita disponer de varios niveles de granularidad, se crean jerarquías entre las diferentes dimensiones.

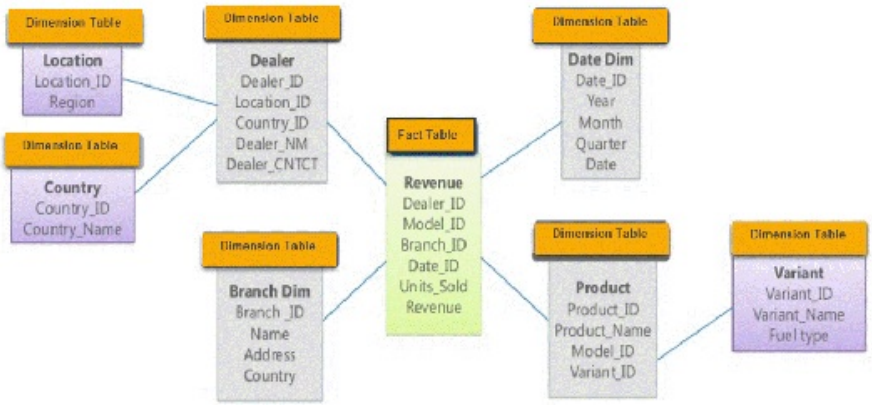
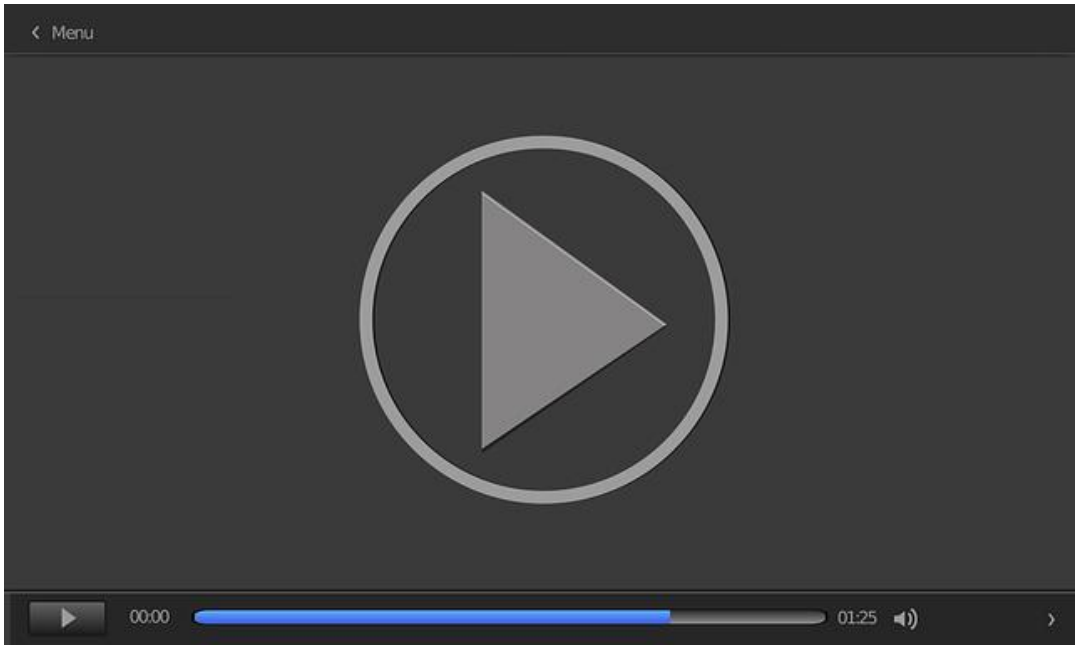


Figura 4. Ejemplo de modelo copo de nieve.Fuente: https://www.guru99.com/star-snowflake-data-warehousing.html

El modelo en estrella es multidimensional, pero no tiene jerarquías. Si las tuviera, se denominaría modelo copo de nieve, es decir, relaciones entre las tablas de dimensiones.

Este modelo ha de estar normalizado, evitando así la redundancia de datos. El principal inconveniente del modelo es el elevado tiempo de respuesta.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=c510e110-d539-4e06-b3ac-b15d0080233f

5.3. Los procesos ETL

Como sus siglas indican, consiste en la **extracción, transformación y carga de los datos** en el *data warehouse*, de modo que se puede afirmar que es una parte fundamental de este. Antes de guardar los datos, deben ser transformados, limpiados, filtrados y redefinidos. Este proceso es, en opinión de los expertos, el que consume más recursos en un proyecto de *business intelligence*, razón por la cual su diseño es importante.

Lo fundamental del proceso ETL es capturar datos de las diferentes fuentes de información disponibles y almacenarla sin errores en el *data warehouse*.

Con la aparición de los *data warehouse* (bases de datos que integran todos los datos de una compañía), las empresas necesitan combinar los datos desde sus diferentes repositorios origen, cargándolos en un único repositorio destino, por lo que el uso de los lenguajes de programación clásicos (COBOL, RPG, PL-SQL o SAS/BASE), que requería del mantenimiento de programas tremendamente largos y de costoso mantenimiento, provocó que las grandes empresas tecnológicas comenzasen a desarrollar sus propias herramientas.

En este momento, empresas como IBM, Informatica, Oracle o SAS comienzan a lanzar potentes herramientas orientadas al diseño y desarrollo de procesos ETL sin la necesidad de programarlas en código. Así nacen Informatica Power Center, IBM Datastage, ODI (*Oracle Data Integrator*) o SAS Data Integrator.

Como bien es sabido, el coste de las licencias de este tipo de *software* es muy elevado, por lo que su implementación se limitó a las grandes corporaciones. Recientemente, y gracias a la progresiva implantación de sistemas de BI por parte de

empresas de menor tamaño, han sido empresas dedicadas a ofrecer *software* de código abierto o libre las que han democratizado el uso de estas herramientas ETL. Algunos ejemplos lo constituyen las herramientas como Talend, KETL, Scriptella, Jaspersoft ETL o, la más conocida de todas ellas, Kettle (Pentaho Data Integrator).

Sin embargo, con el aumento de las necesidades de los datos a tiempo real y la aparición de los sistemas *bigdata*, se está empezando a poner en entredicho el futuro del ETL tradicional. Las herramientas ETL tradicionales funcionaban muy bien en *batch*, pero no en tiempo real. Hay opiniones que apuntan a que la era del *big data* traerá el fin de las herramientas ETL, aunque, por el momento, está provocando que nazcan soluciones híbridas: SAP HANA, Hadoop ETL, Power Center Big Data… o que se creen arquitecturas BI mixtas donde conviven los procesos de ETL tradicionales con tecnologías *big data*.

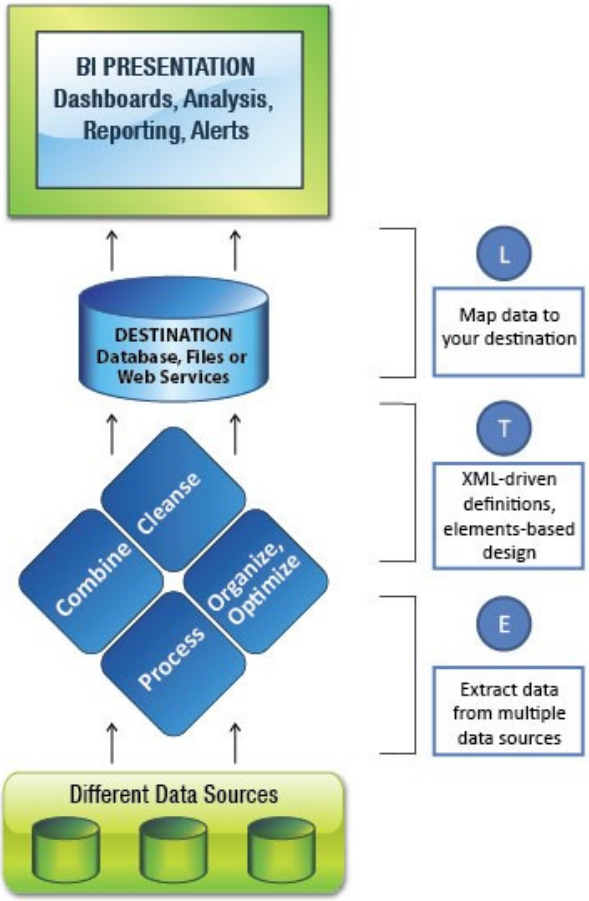


Figura 5. Esquema tradicional de una herramienta ETL.Fuente: https://www.dataprix.com/es/es/herramientas-etl

En la figura 5 reproduce un esquema clásico ETL, en el que, una vez extraídos los datos, se procede al **proceso de limpieza**. En esta fase se comprueba la calidad, se corrigen y se eliminan duplicados, valores erróneos e incluso datos incompletos e inconsistencias. En este campo se está produciendo un desarrollo muy especial en los últimos tiempos.

La tercera de las fases se ocupa de la **transformación de los datos**. Se filtran, recodifican o se generan agregados mediante transformaciones de los datos originales.

Finalmente, la última etapa del proceso es la de **carga**, en la que hay que comprobar la consistencia con definiciones y formatos previamente establecidos. La carga o actualización debe llevarse a cabo mediante sobreescritura, es decir, sin modificar los datos preexistentes.

La mayoría de las herramientas ETL disponibles en el mercado suelen incorporar las siguientes funcionalidades:

- Control de la **extracción de los datos y su automatización** , disminuyendo el tiempo empleado en el descubrimiento de procesos no documentados, minimizando el margen de error y permitiendo mayor flexibilidad.
- Permiten su **compatibilidad** con diferentes tipos de *hardware*, *software*, datos y recursos humanos existentes.
- Permiten una **gestión integrada** del *data warehouse* y los *data marts* existentes, integrando la extracción, transformación y carga para la construcción del *data warehouse* corporativo y de los *data marts*.
- Usan la **arquitectura de metadatos**, facilitando la definición de los objetos de negocio y las reglas de consolidación.
- Permiten acceder a **fuentes** de datos diferentes.

Nos detenemos ahora nuevamente en las diferentes fases del proceso ETL.

Proceso de extracción

Como ya hemos avanzado, la primera parte del proceso ETL consiste en extraer los datos desde las fuentes disponibles que, en la mayor parte de los casos, consisten en la fusión de datos provenientes de diferentes fuentes. Algunos de ellos proceden de bases de datos relacionales o de ficheros planos, pero también es posible que procedan de bases de datos no relacionales o de estructuras distintas.

Una parte fundamental de la extracción es el chequeo de los datos, a partir de los cuales solo los que tienen la estructura esperada pasan a tener el formato necesario para iniciar la siguiente etapa: la de limpieza.

Proceso de limpieza

En esta fase se identifican los datos incompletos, incorrectos o inconsistentes, así como los datos duplicados. De esta forma se elimina la información que no responde a la realidad y se ahorra espacio eliminando duplicaciones.

Proceso de transformación

La fase de transformación aplica una serie de reglas a los datos para convertirlos en datos compatibles con los sistemas de carga. Por ejemplo, estas transformaciones pueden consistir en seleccionar solo ciertas columnas, traducir o recodificar, obtener nuevos valores mediante cálculo, dividir columnas en varias o la aplicación de cualquier regla de validación entre otras.

Proceso de carga

En esta fase, los datos transformados son cargados en el sistema de destino. En algunas bases de datos, se sobrescribe la información antigua con nuevos datos, pero los *data warehouse* mantienen un historial para tener un rastro.

Los dos tipos de carga fundamentales son el de **acumulación simple**, que consiste en resumir todos los cambios de un determinado período y transferirlos al *data warehouse* o bien el ***rolling***, en el que se guarda información resumida a distintos niveles.

Procesamiento paralelo

Los nuevos *softwares* ETL incluyen la aplicación de procesamiento paralelo, lo que mejora el rendimiento cuando se trata de grandes volúmenes de datos. Se pueden identificar tres tipos principales de paralelismos:

De datos: consistente en dividir un archivo secuencial en pequeños archivos de datos para proporcionar acceso paralelo.

De segmentación (*pipeline*): permite el funcionamiento simultáneo de varios componentes en el mismo flujo de datos.

De componente: consiste en el funcionamiento simultáneo de múltiples procesos en diferentes flujos de datos en el mismo puesto de trabajo.

Estos tres tipos de paralelismo no son excluyentes, sino que pueden ser usados en una misma operación ETL.

5.4. El análisis de datos en data warehouse: OLAP

Las llamadas herramientas OLAP son instrumentos de procesado analítico que permiten realizar en línea procesado, cálculo, previsión y análisis de la información residente en un *data warehouse* con funcionalidades que permiten analizar y explorar datos sobre la base de un modelo multidimensional.

La estructura *operative* de OLAP se basa en el concepto de cubo, que no es más que una estructura de datos multidimensional (real o virtual) que permite el rápido análisis de los datos.

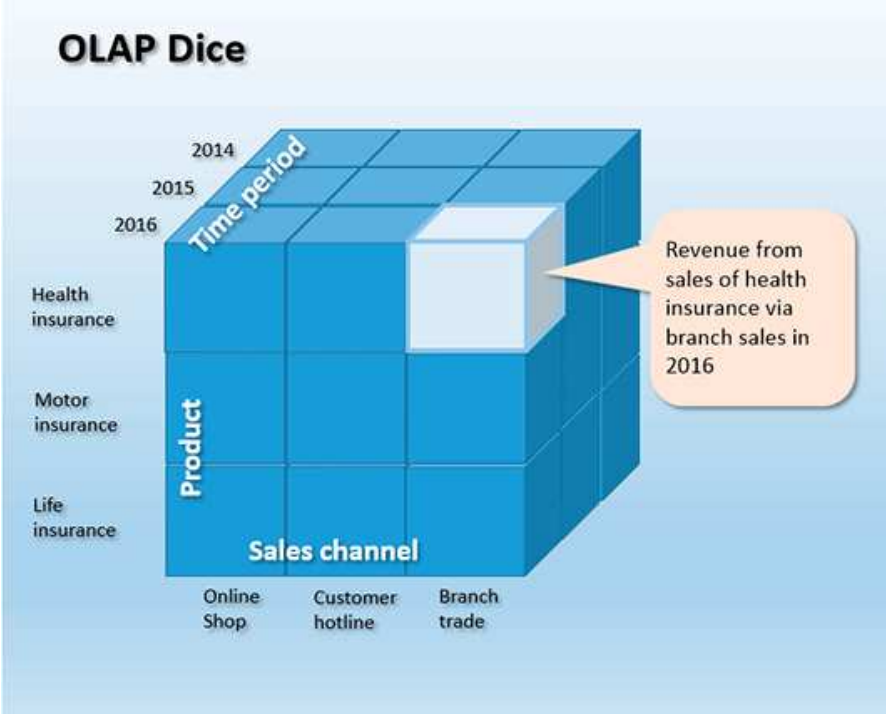


Figura 6. Concepto de cubo. Fuente: https://www.ionos.es/digitalguide/online-marketing/analisis-web/los-data-warehouses-en-la-business-intelligence/

Podemos identificar una serie de **operaciones básicas** de los cubos OLAP:

Consolidación: que supone la acumulación de datos en estructuras más o menos

complejas para producir interrelaciones, por ejemplo, ventas por distritos o por equipo de vendedores, etc.

Drill-down: que permite movernos en direcciones contrarias de agregación, de forma que puede presentar automáticamente datos detallados que abarcan los datos consolidados. Por ejemplo, podría acceder a las ventas de un producto individualizado o por equipos de ventas dentro de un distrito concreto.

***Slicing-dicing*:** se refiere a la capacidad de visualizar la base de datos desde diferentes puntos de vista o dimensiones. Por ejemplo, se podría obtener la gráfica temporal de ventas o las ventas realizadas por diferentes equipos de vendedores, o por coste de las ventas realizadas, etc.

5.5. Tendencia y elección de las herramientas

En la última versión del cuadrante mágico de Gartner del 2018 (figura 7), se evalúan y comparan las herramientas ETL más importantes del mercado, clasificadas atendiendo a diferentes criterios tales como conectividad, capacidad de entrega y transformación de datos, capacidades relacionadas con los metadatos y de modelado, de diseño, de gestión de datos y de arquitectura, entre otros.

Si hacemos uso de esta versión, se obtiene una figura que sigue siendo muy parecida a la de la última década, en la que el liderazgo lo siguen manteniendo las aplicaciones desarrolladas por las grandes tecnológicas.

Figura 7. Herramientas para la integración de datos.Fuente: https://www.informatica.com/es/data-

integration-magic-quadrant.html

Las **características** más importantes que ha de incluir un *software* ETL según Gartner son las siguientes:

- Conectividad o capacidad de adaptación, que se refiere a la capacidad para conectar y relacionar datos de diferentes fuentes (bases de datos relacionales y no relacionales, datos procedentes del ERP y del CRM, etc.).
- Capacidades para proporcionar datos en varios formatos.
- Capacidades de transformación de datos, a través de procesos más o menos complejos.
- Capacidades de recuperar datos desde el origen y de actualizar datos y metadatos sincronizando los cambios.
- Capacidades de diseño y entorno de desarrollo, facilitando el trabajo en equipo y la gestión de *workflows*.
- Capacidades de gestión de datos mediante la aplicación de minería de datos, establecimiento de perfiles, etc.
- Capacidades adaptación a las diferentes plataformas hardware y sistemas operativos existentes.
- Capacidad de aplicación de operaciones como la monitorización y control de los procesos de integración de datos, estadísticas, controles de seguridad, etc.
- Capacidad de la arquitectura y para la integración, permitiendo la interoperabilidad de diferentes componentes del sistema y otras herramientas externas.

Creemos que, en cualquier caso, avanzar hacia *software* libre, en la nube con posibilidad de *data warehouse* en tiempo real, que permita actuar de forma óptima

con *big data* y con capacidad de realizar supercomputación serán las líneas maestras que orienten la evolución de estos productos.

5.6. Referencias bibliográficas

Cano, J. L. (2007). *Business intelligence: competir con información*. Madrid: ESADE.

Comparativa de herramientas ETL

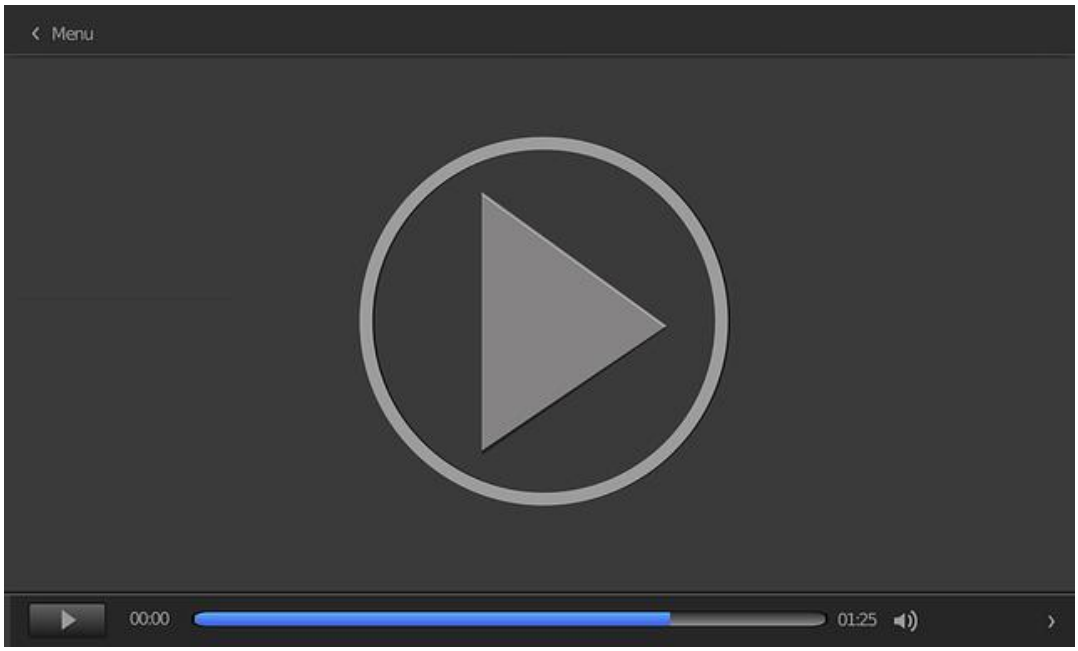
Informatica. Informatica se alza como líder en dos de los cuadrantes mágicos de Gartner para el 2019, en los ámbitos de herramientas de integración de datos e iPaaS empresarial (en línea). https://www.informatica.com/es/data-integration-magic-quadrant.html

En este enlace podrás descargar los informes del cuadrante mágico de Gartner para conocer los principales proveedores de herramientas ETL.

Procesos ETL con Pentaho Data Integration

Auribox Trainin. (16 de junio de 2017). *Procesos ETL con Pentaho Data Integration. Paso a paso* [Archivo de vídeo]. Recuperado de https://youtu.be/DlnmRyACbd4

En este vídeo tutorial se presenta la creación de una ETL paso a paso.



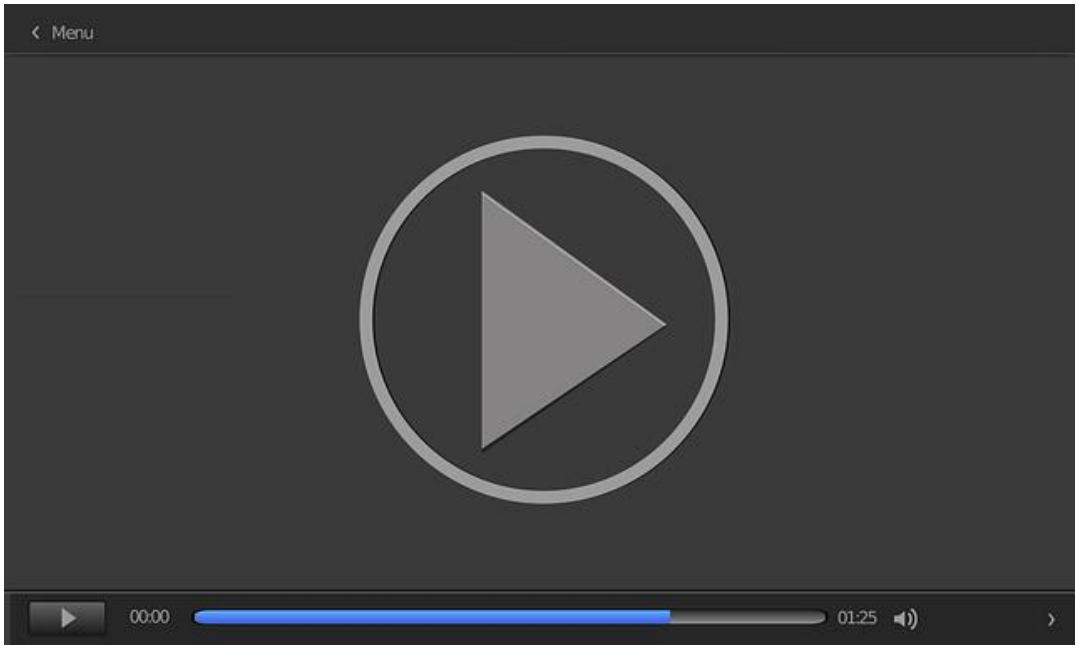
Accede al vídeo:

https://www.youtube.com/embed/DlnmRyACbd4

Desarrollo de un cubo OLAP con Schema Workbench

Sitio21web Desarrollo de sistemas. (15 de enero de 2017). *Cómo hacer un cubo - Pentaho - Mondrian - Schema Workbench - Tutorial* [Archivo de vídeo]. Recuperado de https://youtu.be/_H6jmU5TKQw

En este vídeo podrás observar paso a paso la creación de un cubo con la herramienta Pentaho, de tipo *open source*, que integra todas las etapas de una estrategia BI.



Accede al vídeo:

https://www.youtube.com/embed/_H6jmU5TKQw

Test

1. ¿Es obligatorio realizar la extracción de datos para construir el *data warehouse*?

   A. Sí.

   B. No.

2. ¿Cuál de estas acciones representa una etapa del proceso ETL?

   A. Extracción.

   B. Limpieza.

   C. Carga

   D. Todas las anteriores son correctas.

3. Los *data mart* forman parte del proceso *data warehouse*:

   A. Verdadero.

   B. Falso.

4. Son beneficios asociados a la implementación de un *data warehouse*:

   A. Facilita la toma de decisiones.

   B. Simplifica el acceso a los datos.

   C. Proporciona una visión integral del negocio.

   D. Todas las anteriores son correctas.

5. ¿Cuáles pueden ser dos posibles fuentes de datos?

   A. Bases de datos relacionales y archivos de texto plano.

   B. Archivos XML y codificación de archivos HTML.

   C. Archivos PDF y documentos en papel.

   D. Ninguna de las anteriores.

6. La carga de datos ha de preceder a la extracción y transformación:

A. Verdadero.

B. Falso.

7. El modelo OLAP permite:

A. La acumulación de datos en estructuras más o menos complejas.

B. Movernos en direcciones contrarias de agregación.

C. Visualizar datos desde diferentes puntos de vista o dimensiones.

D. Todas son correctas

8. El *metadata*:

A. Hace referencia a la sintaxis.

B. Hace referencia a la semántica de los datos.

C. Hace referencia a la estructura de los datos.

D. Todas las anteriores son correctas.

9. ¿Cuál es la función del *data warehouse*?

A. Aumentar el trabajo de los usuarios.

B. Ayudar en la toma de decisiones.

C. Centralizar los datos para facilitar el manejo.

D. Ninguna de las anteriores.

10. Para diseñar un *data warehouse* hay que incluir al menos dos *data marts*:

A. Verdadero.

B. Falso.

C. Solo en el caso de los data marts dependientes.

D. Todas las anteriores son correctas.