

Almacenamiento y Procesamiento de la Información

Índice

1	Introducción.....	3
2	Definición del Datawarehouse.....	3
3	Uso del Almacén de Datos.....	4
4	Ventajas e Inconvenientes del Almacén de Datos.....	4
5	Arquitectura del Datawarehouse.....	6
6	ETL. Conceptos y Componentes.....	7
7	Herramientas del mercado.....	12
8	Conclusiones.....	12
9	Bibliografía Esencial.....	13
10	Bibliografía Recomendada.....	13

Objetivos

Los objetivos principales perseguidos en esta clase son:

- Comprender el concepto de datawarehouse y la importancia que cobra en la empresa entendiendo los posibles usos que puede tener.
- Ser capaces de definir las ventajas de disponer de un datawarehouse frente a no disponer de él.
- Conocer la arquitectura completa del DWH entendiendo los procesos ETL, así como cada uno de sus componentes y algunas de las herramientas más importantes del mercado.

1 Introducción

La visión práctica del Business Intelligence es capacitar a las organizaciones para aprovechar y **transformar los datos** con el fin de **lograr resultados** de alto rendimiento.

2 Definición del Datawarehouse

Un **almacén de datos** (del inglés data warehouse) es una **colección de datos** orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. No siendo un producto, lenguaje o copia del sistema transaccional.

Se trata, sobre todo, de un **expediente completo de una organización**, 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 (especialmente OLAP, procesamiento analítico en línea). El almacenamiento de los datos no debe usarse con datos de uso actual.

Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas dependiendo del subsistema de la entidad del que procedan o para el que sea necesario (data marts).

Un DWH es un expediente completo de la empresa

3 Uso del Almacén de Datos

El almacén de datos debe contener los datos de la empresa para poder convertirlos en información útil.

En un almacén de datos lo que se quiere es contener datos que son necesarios o útiles para una organización, es decir, que se utiliza como un **repositorio de datos** para posteriormente **transformarlos en información útil** para el usuario. Un almacén de datos debe entregar la **información correcta** a la **gente indicada** en el **momento óptimo** y en el formato adecuado.

El almacén de datos da respuesta a las necesidades de usuarios expertos, utilizando Sistemas de Soporte a Decisiones (DSS), Sistemas de información ejecutiva (EIS) o herramientas para hacer consultas o informes. Los usuarios finales pueden hacer fácilmente consultas sobre sus almacenes de datos sin tocar o afectar a la operación del sistema.

En el funcionamiento de un almacén de datos son muy importantes las siguientes ideas:

- **Integración de los datos** provenientes de bases de datos distribuidas por las diferentes unidades de la organización y que con frecuencia tendrán diferentes estructuras (fuentes heterogéneas).
- **Separación de los datos** usados en operaciones diarias de los datos usados en el almacén de datos para los propósitos de explotación, de ayuda en la toma de decisiones, para el análisis y para operaciones de control.
- **Períodicamente, se importan datos** al almacén de datos desde las distintas fuentes de información (aplicaciones ERP, ficheros, etc.) realizando las transformaciones necesarias y aplicando las reglas de negocio oportunas.

4 Ventajas e Inconvenientes del Almacén de Datos

El DWH será el repositorio corporativo, aunque este tipo de proyectos tiene un alto costo y dificultad.

Aunque una empresa que no posee un datawarehouse tiene riesgo de seguir siendo competitiva y que sea capaz de gestionar eficientemente su negocio, debemos analizar las ventajas y los inconvenientes que se tienen a la hora de abordar un proyecto de datawarehousing.

Ventajas

Hay muchas ventajas por las que es altamente recomendable usar un almacén de datos son:

- Los almacenes de datos hacen **más fácil el acceso** a una gran variedad de datos a los usuarios finales.

- **Facilitan el funcionamiento** de las aplicaciones de los sistemas de apoyo a la decisión tales como informes de tendencia, por ejemplo: obtener los ítems con la mayoría de las ventas en un área en particular dentro de los últimos dos años; informes de excepción, informes que muestran los resultados reales frente a los objetivos planteados a priori, etc.
- Los almacenes de datos pueden **trabajar en conjunto** y, por lo tanto, aumentar el valor operacional de las aplicaciones empresariales, en especial la gestión de relaciones con clientes.

Inconvenientes

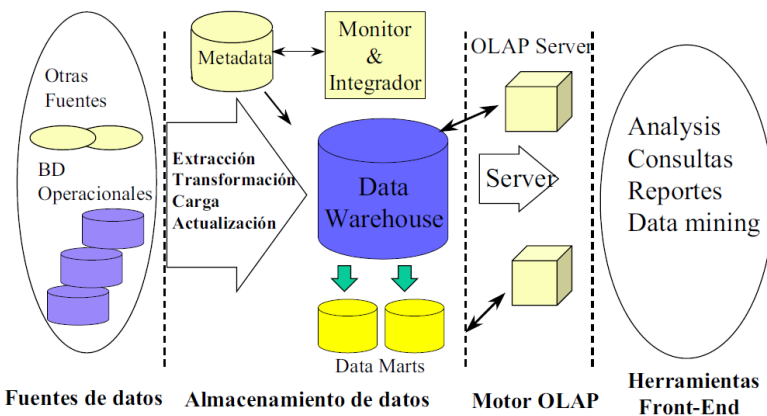
Utilizar almacenes de datos también plantea algunos inconvenientes, como:

- A lo largo de su vida los almacenes de datos pueden suponer **altos costos**. El almacén de datos no suele ser estático. Los costos de mantenimiento son elevados.
- Los almacenes de datos se pueden quedar **obsoletos relativamente pronto**.
- A veces, ante una petición de información estos devuelven una **información poco óptima**, que también supone una pérdida para la organización.
- A menudo existe una delgada línea entre los almacenes de datos y los sistemas operacionales. Hay que determinar qué funcionalidades de estos se pueden aprovechar y cuáles se deben implementar en el data warehouse. Resultaría **costoso implementar operaciones no necesarias o dejar de implementar alguna que sí vaya a necesitarse**.

5 Arquitectura del Datawarehouse

Los datos de diversas fuentes se integran en el DWH mediante procesos ETL.

En la figura siguiente se ilustran los **componentes** que conforman un **sistema completo de BI**, éstos son:



Marco de Arquitectura de sistema informacional

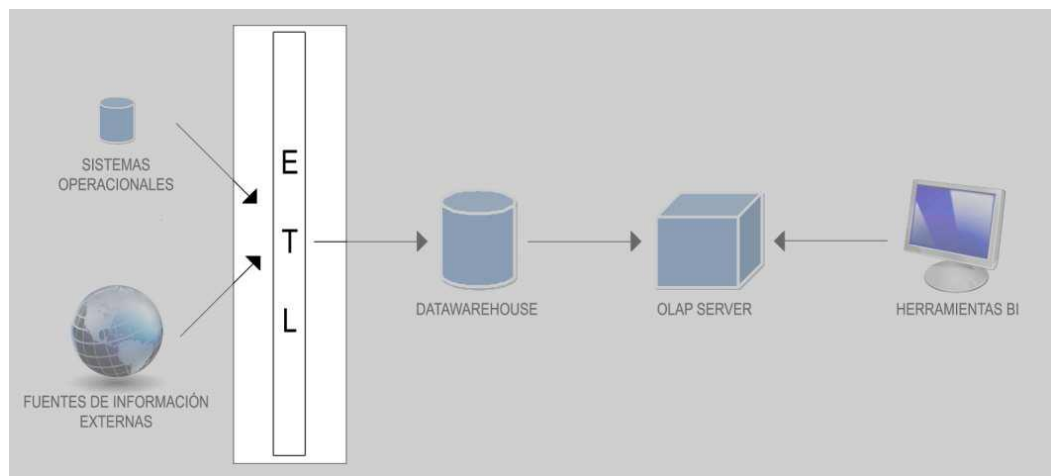
Los componentes principales:

- **Fuentes de información**, que nutrirán de datos al sistema.
- **Proceso ETL**, que extrae, transforma y carga los datos de las fuentes de información al datawarehouse.
- **Datawarehouse**, que reúne los datos de la organización. Es una colección de datos integrada, esto es, contiene los datos de todos los sistemas operacionales de la organización, y es no volátil, la información no se modifica ni se elimina.
- **Motor OLAP** que permite capacidad de cálculo, consultas, análisis de escenarios en grandes volúmenes de datos.
- **Herramientas Front-end**, que son las herramientas de visualización que permiten la navegación por los datos para su análisis por parte del usuario final

6 ETL. Conceptos y Componentes

ETL: Extraer, transformar y cargar

La migración de datos de las fuentes de datos al datawarehouse requiere un proceso para extraer, transformar y cargar dichos datos. Este proceso se conoce como ETL, del inglés Extract, Transform and Load.



Arquitectura-Proceso ETL

Esta parte del proceso es la más costosa, pero es necesaria por la necesidad de conciliar y limpiar los datos de origen.

Importancia de los procesos ETL

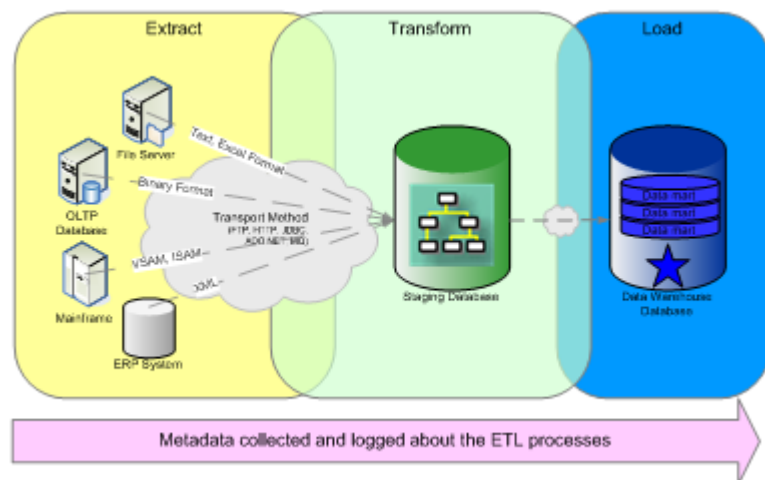
El **diseño, construcción y prueba** de los componentes de ETL es una de las **mayores tareas y más complejas** en proyectos de almacenamiento de datos conllevando la mayoría de **esfuerzo, tiempo y costo**.

El proceso de ETL se utiliza para **rellenar continuamente** y actualizar el almacén de datos, con las consideraciones de **mantenimiento** asociados a la vulnerabilidad y el uso del sistema / **cambios de diseño**.

Por qué utilizar una herramienta ETL

En la actualidad existen muchas **herramientas ETL** que pueden **simplificar** y acelerar significativamente la implementación de un proceso de ETL. Tienen la ventaja de **mantenimiento más fácil** y que la **formación** a los usuarios es **más rápida** puesto que son ayudados en gran medida por tener una representación gráfica del proceso global, e individuales asignaciones de las ETL.

Estas herramientas también ayudan con la **recolección de metadatos** que describen las acciones que se han realizado sobre los datos. Hay que destacar que las herramientas ETL se pueden utilizar en **combinación con otros productos** (por ejemplo, herramientas dedicadas a la calidad de datos), y ocasionalmente con código personalizado en PL / SQL o servicios web.



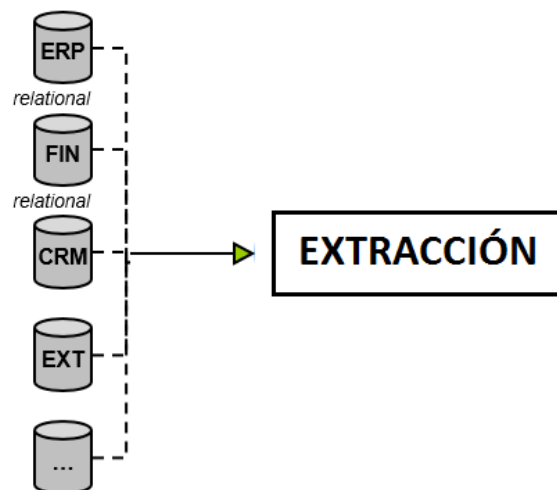
Procesos principales ETL

Extracción

Es el **subproceso ETL** encargado de **recuperar los datos de las fuentes de información**. Se extraen los datos con el uso de una herramienta ETL, pero se podrían extraer también de manera manual, con lenguajes de programación como COBOL.

Lo más sencillo sería duplicar los datos de las fuentes de información y pasar este conjunto a la herramienta ETL, pero de esta forma se hace trabajar a la herramienta con enormes cantidades de datos, cuando en realidad no se necesitarán todos.

La manera óptima sería limpiar, filtrar y ordenar los datos requeridos al datawarehouse, y si es posible en un sólo paso. Pero esto cuenta con el inconveniente de que es posible que tuvieran que desmontarse las bases de datos de la organización durante unas horas, teniendo que parar las actividades que se estén realizando en ese momento contra las fuentes de información. Esta es la razón por la que **se divide el proceso en extracción, transformación y carga**.



Extracción desde distintos orígenes

En este subproceso, **extraeremos los datos necesarios y sólo los necesarios** de las fuentes de información. Para ello tenemos que convenir cuáles son las mejores fuentes, de entre las que se cuentan. Terminado de ejecutar el subproceso contamos con los datos originales.

Ha de tenerse en cuenta que, si no todos los datos origen son actuales, es decir, **si queremos incluir** en el proceso ETL **datos origen históricos**, **será necesario aplicar algunos cambios a los programas encargados de la carga inicial**, porque al tratarse de datos antiguos es posible que no se solapen muy bien con los actuales, dando lugar a muchos posibles rechazos de datos que quieren cargarse en el proceso de limpieza.

Transformación

Después de limpiar los datos se procede al **paso de transformación**, que está **destinado a la integración, derivación, agregación y totalización de los datos**.

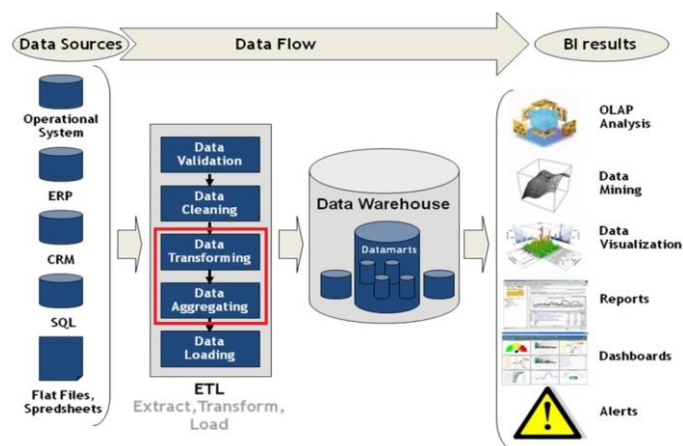
- **Integración.** Cada elemento debe conocerse con un nombre estándar.
 - Renombraremos, de ser necesario, conforme al estándar que se siga en el datawarehouse. Así se renombrará por ejemplo DNI a DNI_Usuario.
- **Derivación.** Partiendo de datos individuales se calcularán nuevos datos, mediante la realización de cálculos o búsquedas. Un ejemplo sería a partir de un código sacar una descripción. Opuestamente, también puede crearse un nuevo dato a partir de un solo dato de origen, como por ejemplo podrían crearse los campos, día, mes, año, hora, minuto, y segundo, a partir de un campo timestamp.

Pero la parte primordial de la transformación se basa en el pre-cálculo de datos, para optimizar las futuras consultas al datawarehouse.

- **Totalización.** Se basa en el procesamiento de datos de origen de carácter numérico para obtener otro valor numérico que corresponderá a promedios, totales, máximos, mínimos, etc.

- **Agregación.** Consiste en crear datos a partir de tuplas de dos tablas diferentes, por ejemplo, a partir de la tabla 'usuario' y de la tabla 'cargos', crear un historial con los puestos que ha desempeñado el usuario en la empresa.

Como es evidente, los datos se totalizan o agregan en base a la información que se quiera reportar.



Transformación y agregación de la información

Por otro lado, los datos deben **conciliarse** antes de cargarse en el datawarehouse, es decir, comprobar que **la información que entra** en el proceso ETL coincide con la que sale del mismo. Cada programa que lea datos y los escriba luego debe crear lo que se llaman **totales de conciliación**.

Existen tres tipos de totales:

- **Cantidad de registros.** Para que las cifras coincidan, **el número total de registros que entran debe ser igual al número de registros que salen** más el número de registros que se rechazan, suponiendo que algún registro en el proceso de limpieza sea rechazado y no sea añadido. Este tipo de conciliación se debe llevar a cabo cuando se han extraído y mezclado filas de datos.
- **Totales.** Versa lo mismo que el primer tipo, pero en este caso lo que **se suma y totaliza es una determinada columna de entrada**. Es decir, si en el tipo anterior lo que se contabilizaba era el número de registros de una tabla, aquí se contabiliza el número de atributos de la misma.

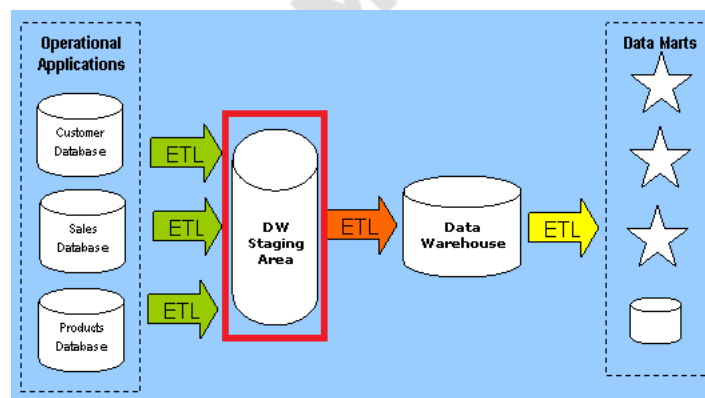
- **Cantidad de dominios.** La cantidad de registros para un determinado dominio (o valor) del campo de entrada, debe ser la misma que el número de registros para el mismo dominio en la salida.

Área staging

Un área de ensayo, o área provisional, es un **área de almacenamiento intermedio** utilizado para el procesamiento de datos durante la extracción, transformación y el proceso de carga (ETL). El área de preparación de datos se encuentra entre la fuente de datos(s) y el objetivo(s) de datos, que a menudo son almacenes de datos, u otros repositorios de datos.

Áreas staging son **a menudo de naturaleza transitoria (volátiles)**, donde sus contenidos se borran antes de ejecutar un proceso ETL o inmediatamente después de la finalización con éxito de un proceso de ETL ejecutado.

Hay arquitecturas de área staging puestas en escena **de manera no transitorias**, sin embargo, **son muy pocas** las que están diseñadas para mantener los datos durante largos periodos de tiempo para emplearlas con fines de archivo o de solución de problemas.



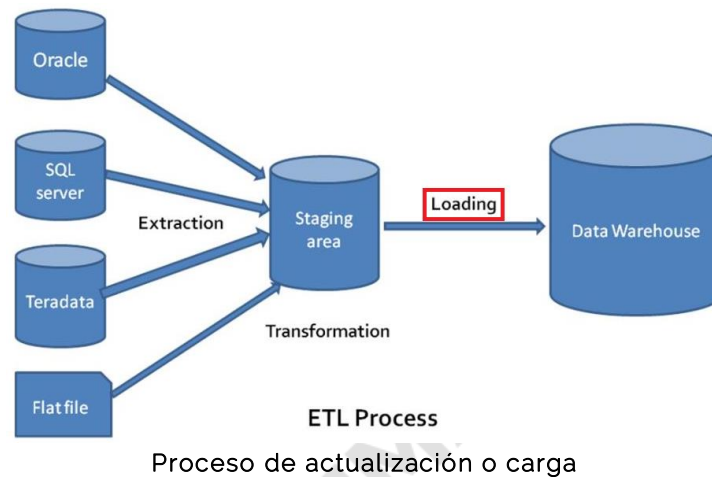
Arquitectura donde se integra el área staging

Carga

Este subproceso se encarga de la **carga de los datos resultantes en el datawarehouse**. Es un proceso sencillo, pero se deben tener en cuenta dos aspectos:

- Una vez cargados los datos en el datawarehouse se debe **activar la integridad referencial** para evitar que los datos almacenados se corrompan.

- Es necesario **definir índices** debido a la gran cantidad de información que se ha cargado facilitando un óptimo acceso.



Actualmente hay numerosos fabricantes de herramientas ETL siendo algunas de las más utilizadas Powercenter, SSIS y Datastage.

7 Herramientas del mercado

Actualmente en el mercado hay numerosos fabricantes de software haciendo que la lista de herramientas ETL elegibles sea amplia. Algunas de las más conocidas son:

- Informática **Powercenter**, **SAP Data Services**, **Oracle Data Integrator**, **IBM Datastage**, **Pentaho Data Integration (open source)** son de uso general, herramientas ETL dedicadas.
- **SQL Server Integration Services (SSIS)** es una herramienta ETL libre disponible cuando se utiliza Microsoft SQL Server 2005.

8 Conclusiones

A modo de resumen algunos de los puntos que debemos tener en mente son:

- **Modularidad** - principios de desarrollo de software. Dividir los procesos por lógica, reutilizables, procesos mantenibles

Los **metadatos** deben ser recogidos durante la ejecución. Estado de éxito/ fallo, tiempo de inicio / final, número de registros procesados, etc.

- La **arquitectura** del DWH influirá en el tipo y la **escalabilidad** de los procesos ETL requeridos.
- **Construcción Off-shore** – requiere encontrar el candidato ideal, ya que ETL es la parte más importante y grande del proyecto de DWH. Asimismo, es importante tener en cuenta el **coste** de las licencias de herramientas, la conectividad, la legalidad de acceso a datos, etc.

9 Bibliografía Esencial

La siguiente bibliografía debería ser leída con atención con el fin de que se afiancen los conocimientos teóricos expuestos el tema.

- **Wikipedia**, enciclopedia digital.
Enlace: https://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos
- **TDWI**, comunidad especializada en BI y DWH.
Enlace: <https://tdwi.org/Home.aspx>
- **TechTarget**, comunidad especializada en BI y DWH
Enlace: <https://searchdatamanagement.techtarget.com>
- **Business Intelligence Knowledge**, blog especializado en BI y DWH.
Enlace: <http://www.business-intelligence-knowledge.com/>

10 Bibliografía Recomendada

En las siguientes comunidades se puede ampliar profundamente sobre los temas vistos en el tema.

- **Kimbal Group**, grupo de consultores expertos y especializados en el diseño de DWH para conseguir soluciones de BI optimizadas.
Enlace: <http://www.kimballgroup.com/>
- **BI LATAM**, comunidad latinoamericana especializada en BI y DWH.

Enlaces:

Web:

<http://bilatam.pass.org/Inicio.aspx>

Youtube: <https://www.youtube.com/user/BILATAM/videos>

Facebook: <https://es-la.facebook.com/BusinessIntelligenceLatam/>

Red SUMMA ©