FACULTAD DE INGENIERÍA

Ingeniería en Computación

Bases de Datos Avanzadas

Datawarehouse

Elaborado por: MARÍA DE LOURDES RIVAS ARZALUZ

Septiembre 2015

Propósito

Actualmente las empresas necesitan contar con alguna herramienta que les apoye en la toma de decisiones tanto tácticas como estratégicas haciendo uso de grandes cantidades de sus datos para conseguir este propósito.

La información contenida en este trabajo permitirá al alumno conocer los conceptos fundamentales, la arquitectura y la explotación de un almacén de datos (datawarehouse), que es un componente de la inteligencia de negocios.

Contenido

- Introducción
- Conceptos
- Modelo Multidimensional
- OLAP
- Cube/Rollup
- ETL
- Conclusiones
- Bibliografía

Introducción

- Una vez que cubiertas las necesidades operacionales, surgen nuevos requerimientos sobre los sistemas, como obtener información que sirva de base para la toma de decisiones.
- Estás necesidades se basan en el análisis de un número ingente de datos
- Su finalidad es realizar análisis detallados sobre los datos
- Un datawarehose no es un producto, es un proceso dentro de la inteligencia de negocios.

Conceptos

Inteligencia de Negocios (BI, Business Intelligence)

Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

Fig. 1. Inteligencia de Negocios

Conceptos

Características de BI

- ✓ Accesibilidad a la información: Los datos son la fuente principal de este concepto.
- ✓ Apoyo en la toma de decisiones: Acceso a herramientas de análisis que permitan seleccionar y manipular aquellos datos de interés.
- ✓ Orientación al usuario final. Independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

Conceptos

Datawarehouse

Una colección de datos clasificada por temas, integrada, variable en el tiempo y no volátil que se utiliza como ayuda al proceso de toma de decisiones por parte de quienes dirigen una organización.

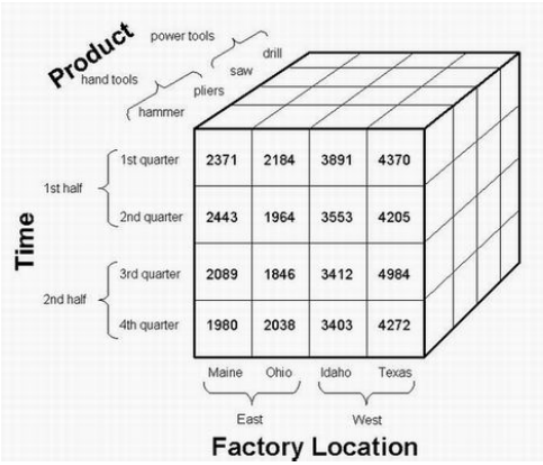

Fig. 2. Representación de un datawarehouse.

Conceptos

Datawarehouse (DWH)

- ✓ Extrae datos de las bases de datos operacionales o fuentes externas, transforma, consolida, integra, verifica la integridad y centraliza los datos.
- ✓ Permite el acceso y manipulación de la información, a través de análisis multivariables, con el objetivo final de dar soporte al proceso de toma de decisiones.
- ✓ Se utilizan nuevas técnicas y estrategias de diseño.

Conceptos

Características de los DWH

✓ Temática: Se organizan los datos temas o hechos (por ejemplo, ventas, clientes) de en lugar de áreas de aplicación.

✓ Integrada: Es debido a la mezcla de datos procedentes de diferentes sistemas de aplicación.

✓ Variable en el tiempo: Intervalo de tiempo durante el que se almacenan los datos y el hecho de que los datos representan una serie de instantáneas.

✓ No volátil: Los datos no se actualizan en tiempo real sino que se refrescan en forma periódica partir de los sistemas operacionales.

Conceptos

Ventajas de los DWH

- ✓ Proporciona información clave para la toma de decisiones empresariales.
- ✓ Mejora la calidad de las decisiones tomadas.
- ✓ Especialmente útil para el medio y largo plazo.
- ✓ Útiles para el almacenamiento de análisis y consultas de históricos.
- ✓ Las empresas obtienen un aumento de la productividad.
- ✓ Transforma los datos en información y la información en conocimiento

Conceptos

Desventajas de los DWH

- ✓ Subestimación de los recursos necesarios para la carga de datos.
- ✓ Surgen problemas ocultos de los sistemas de origen.
- ✓ No se capturan los datos requeridos.
- ✓ Pueden generarse problemas en la homogeneización de los datos.
- ✓ Existe una alta demanda de recursos.
- ✓ Genera altos costos de mantenimiento.
- ✓ Requiere de continua limpieza, transformación e integración de datos.
- ✓ Tiene un diseño complejo y multidisciplinar.

Conceptos

La arquitectura de un DHW viene determinada por su situación central como fuente de información para las herramientas de análisis.

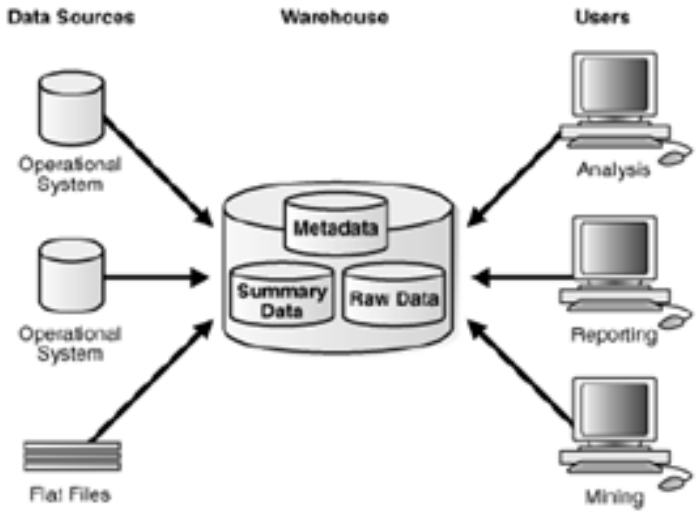

Fig. 3. Arquitectura básica de datawarehouse.

Conceptos

Datamart

- ✓ Base de datos departamental, enfocada en el almacenamiento de los datos de un área de negocio específica.
- ✓ Estructura óptima de datos para analizar la información al detalle que afectan a los procesos de un área o departamento.
- ✓ Puede ser alimentado desde los datos de un DataWareHouse

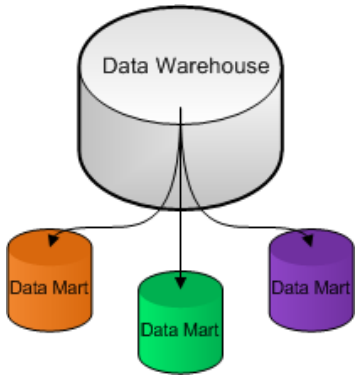

Fig. 4. Los Datamarts se alimentan de un Datawarehouse.

Modelo multidimensional

- ✓ En un esquema multidimensional se representa una actividad que es objeto de análisis (hecho) y las dimensiones que caracterizan la actividad (dimensiones) .
- ✓ La información relevante sobre el hecho (actividad) se representa por un conjunto de indicadores (medidas o atributos del hecho).
- ✓ La información descriptiva de cada dimensión se representa por un conjunto de atributos (atributos de dimensión).
- ✓ Entre los atributos de una dimensión existen jerarquías.

Modelo multidimensional

- ✓ Se pueden utilizar distintos modelos de datos (conceptuales o lógicos).
- ✓ La representación gráfica del esquema multidimensional dependerá del modelo de datos utilizado (relacional, ER, UML, OO)

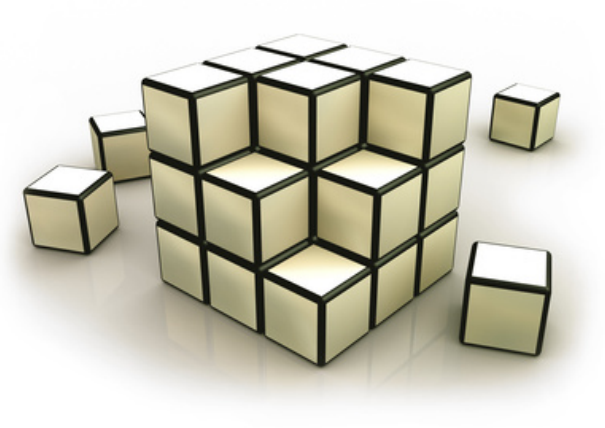

Fig. 5. Modelo multidimensional

Modelo multdimensional

Cadena de supermercados con 300 almacenes en la que se ofrecen unos 30.000 productos.
Actividad: Ventas

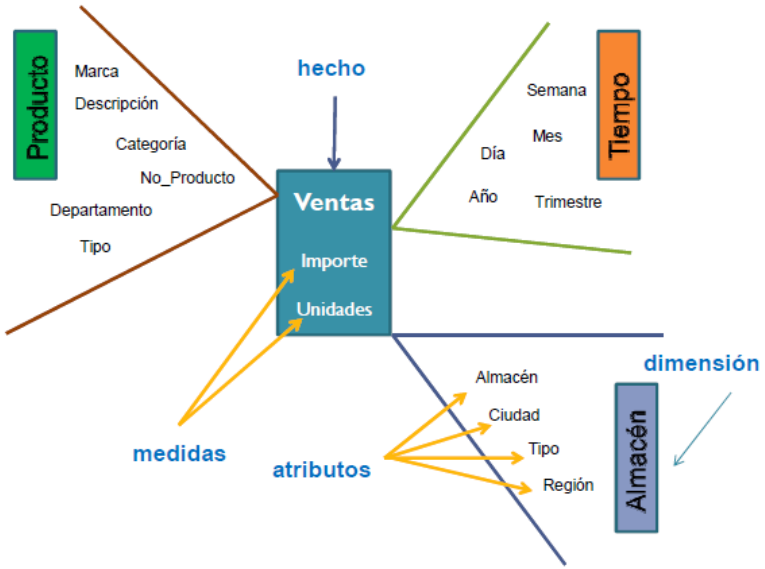

Fig. 6. Ejemplo de un modelo multidimensional

OLAP

OLAP (On-Line Analytical Processing, por sus siglas en inglés).

- ✓ Es una solución utilizada en el campo de la Inteligencia de negocios (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

- ✓ Utiliza estructuras multidimensionales (Cubos OLAP o hipercubos) que contienen datos resumidos de grandes Bases de datos o Sistemas Transaccionales (OLTP).

- ✓ Los cubos, las dimensiones y las jerarquías son la esencia de la navegación multidimensional del OLAP.

OLAP

Tipos de implementaciones OLAP

- ✓ MOLAP (Multidimensional OLAP)
- ✓ ROLAP (Relational OLAP)
  - Esquema en estrella (star)
  - Esquema en copo de nieve (snowflake)
  - Constelaciones de hechos

OLAP

MOLAP (Multidimensional OLAP)

- ✓ Con estructuras multidimensionales (matrices multdimensionales) la información puede ser visualizada en varias dimensiones de análisis.

- ✓ Utiliza una arquitectura de dos niveles:
  - La bases de datos multidimensionales
  - El motor analítico.

- ✓ La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato.

- ✓ Requiere unos cálculos intensivos de compilación

OLAP

ROLAP (Relational OLAP)

- ✓ Usa bases de datos relacionales
- ✓ La base de datos relacional maneja los requerimientos de almacenamiento de datos
- ✓ El motor ROLAP proporciona la funcionalidad analítica.
- ✓ Los usuarios ejecutan sus análisis multidimensionales, y el motor ROLAP transforma dinámicamente sus consultas a sentencias SQL.

OLAP

ROLAP

- Esquema en estrella (star)

Una tabla de hechos y una tabla adicional por cada dimensión

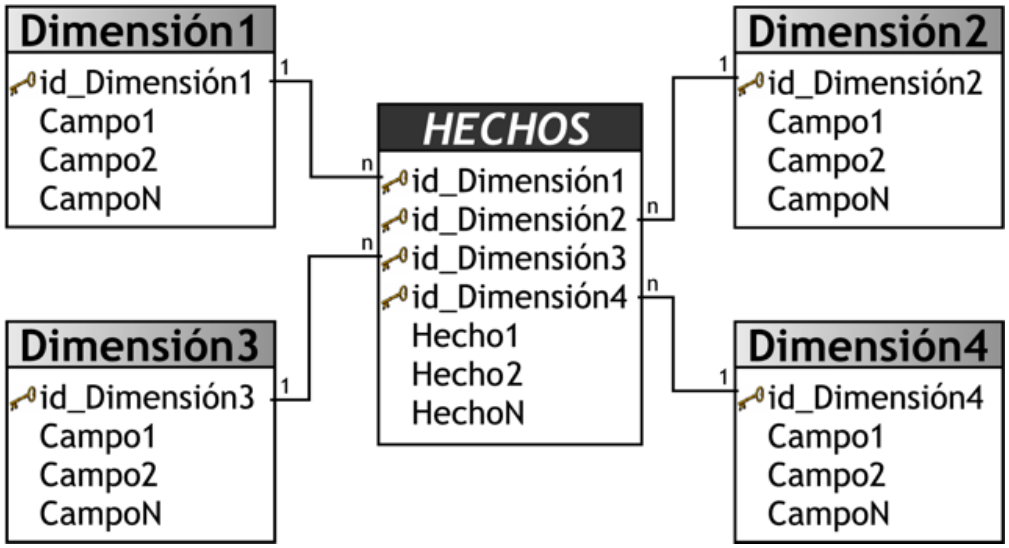

Fig. 7. Estructura de un esquema de estrella

OLAP

ROLAP

- Esquema de copo de nieve

Refleja la organización jerárquica de las dimensiones

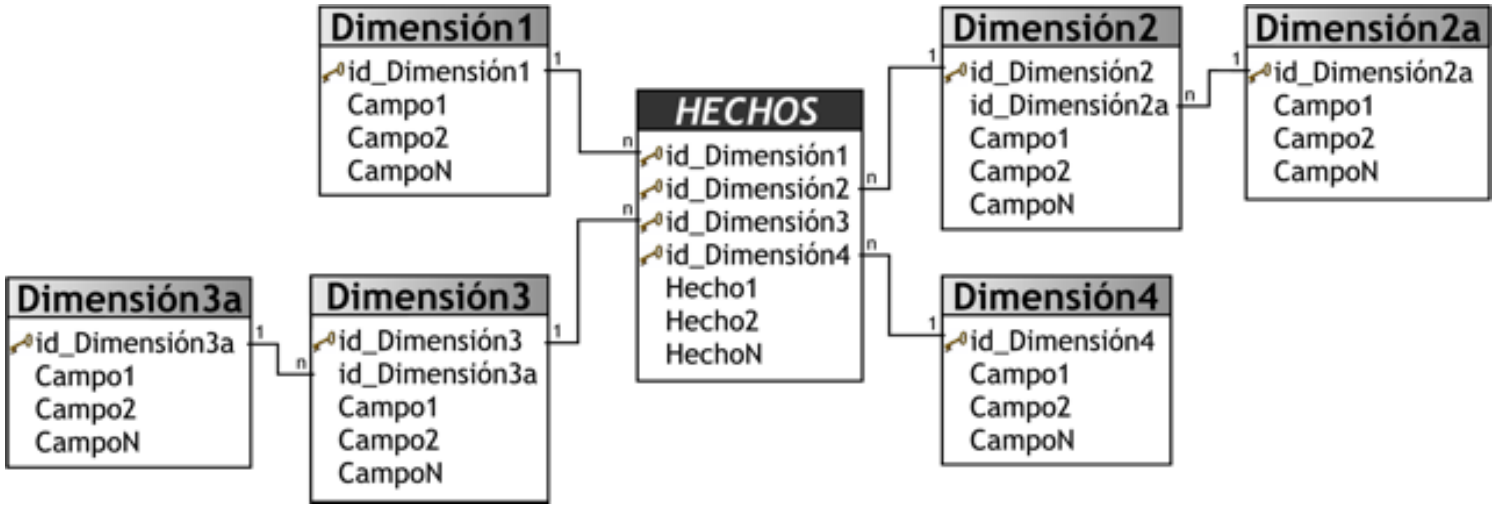

Fig. 8. Estructura de un esquema de copo de nieve

OLAP

ROLAP

- Esquema de constelaciones de hechos

Los esquemas en estrella y bola de nieve pueden generalizarse con la inclusión de distintas tablas de hechos que comparten todas o algunas de las dimensiones.

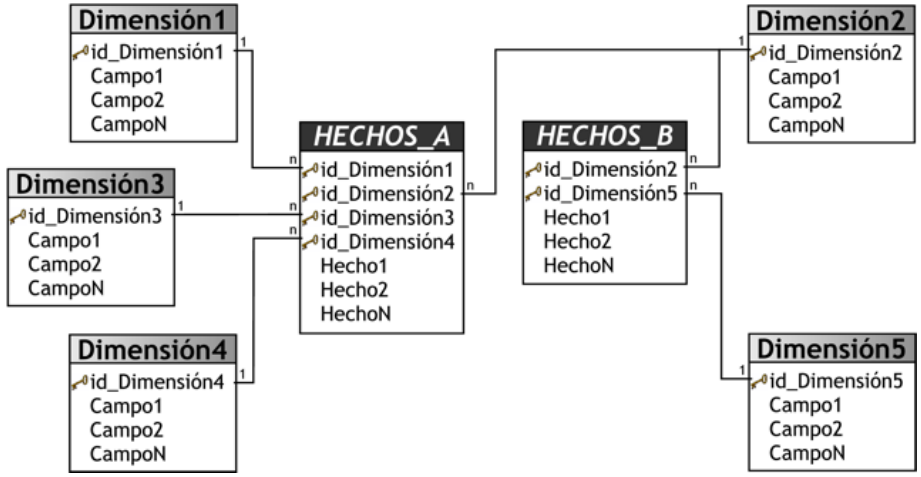

Fig. 9. Estructura de un esquema de constelaciones

OLAP

Operaciones OLAP

Se puede utilizar SQL ampliado (ROLLUP y CUBE)

- Agregación(roll): permite un criterio de agrupación en el análisis, agregando grupos.
- Disgregación(drill): Carácter de agrupación de las consultas disgregando los grupos actuales.

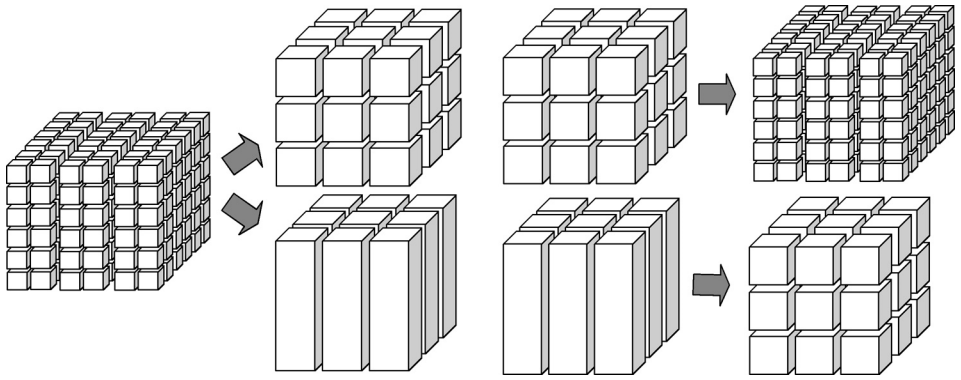

Fig. 10. Operaciones de agregación y disgregación

OLAP

Sistemas OLTP vs Sistemas OLAP

Sistemas OLTP	Almacenes de datos
Almacena datos actuales	Almacena datos históricos
Almacena datos detallados	Almacena datos resumidos en poca o gran medida
Los datos son dinámicos	Los datos son principalmente estáticos
Procesamiento repetitivo	Procedimiento *ad hoc*, no estructurado y heurístico
Alta tasa de transacciones	Tasa media o baja de transacciones
Patrón de uso predecible	Patrón de uso impredecible
Dirigido por transacciones	Dirigido por análisis
Orientado a la aplicación	Orientado a los temas
Soporta las decisiones cotidianas	Soporta las decisiones estratégicas
Sirve a un gran número de usuarios administrativos/operacionales	Sirve a un número relativamente bajo de usuarios de tipo gerencial

Rollup/Cube

Equivalencia en consultas SQL

Conjunto A – Consultas de agregación
Conjunto B – Consultas con gruping sets

```
A1. SELECT a, b, SUM(c) FROM tab1 GROUP BY GROUPING SETS ((a,b))
B1. SELECT a, b, SUM(c) FROM tab1 GROUP BY a, b

A2. SELECT a, b, SUM(c) FROM tab1 GROUP BY GROUPING SETS ((a,b),a)
B2. SELECT a, b, SUM(c) FROM tab1 GROUP BY a, b UNION
    SELECT a, null, SUM(c) FROM tab1 GROUP BY a

A3. SELECT a,b, SUM(c) FROM tab1 GROUP BY GROUPING SETS (a,b)
B3. SELECT a, null, SUM(c) FROM tab1 GROUP BY a UNION
    SELECT null, b, SUM(c) FROM tab1 GROUP BY b

A4. SELECT a, b, SUM(c) FROM tab1 GROUP BY GROUPING SETS ((a,b),a,b,())
B4. SELECT a, b, SUM(c) FROM tab1 GROUP BY a, b UNION
    SELECT a, null, SUM(c) FROM tab1 GROUP BY a, null  UNION
    SELECT null, b, SUM(c) FROM tab1 GROUP BY null, b UNION
    SELECT null, null, SUM(c) FROM tab1
```

Rollup/Cube

Equivalencia en consultas SQL

CUBE: crea un subtotal de todas las posibles combinaciones de los conjuntos de columnas incluidos en su argumentos

```
GROUP BY CUBE(a,b,c) es equivalente a
GROUP BY GROUPING SETS ((a,b,c),(a,b),(b,c),(a,c),(a),(b),(c),())
```

ROLLUP: calcula la agregación en niveles jerarquicos de la dimensión

```
ROLLUP (a, b, c) es equivalente a
GROUPING SETS ((a,b,c),(a,b),(a),())
```

Proceso ETL

ETL (Extract-Transform-Load)

- ✓ Proceso que organiza el flujo de los datos entre diferentes sistemas en una organización.
- ✓ Aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos, data mart ó un almacén de datos.
- ✓ ETL forma parte de la Inteligencia Empresarial (Business Intelligence), también llamado “Gestión de los Datos” (Data Management).

Proceso ETL

Extracción

- ✓ Consiste en extraer los datos desde los sistemas de origen
- ✓ Los formatos de las fuentes pueden estar en bases de datos, archivos planos u otras estructuras
- ✓ Convierte los datos a un formato preparado para iniciar el proceso de transformación

Fig. 11. Extracción en el proceso ETL

Proceso ETL

Transformación

- ✓ Aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados.
- ✓ Algunas fuentes de datos requerirán cierta manipulación de los datos

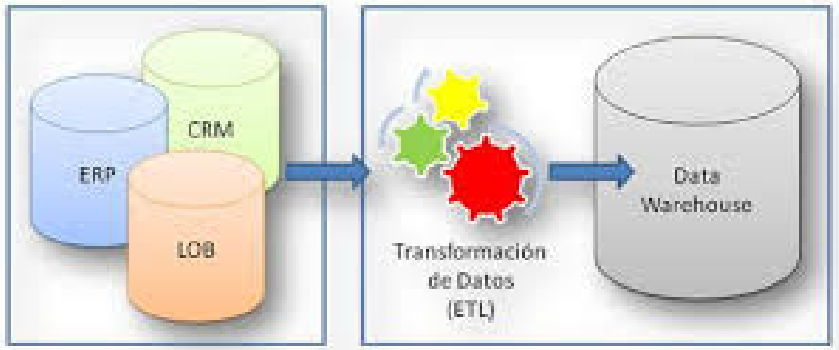

Fig. 12. Transformación en el proceso ETL

Proceso ETL

Carga

- ✓ Formas básicas para el proceso de carga:
  - Acumulación simple: realiza un resumen de todas las transacciones comprendidas en el período de tiempo
  - Rolling: Se aplica para mantener varios niveles de granularidad
- ✓ Interactúa directamente con la base de datos de destino.
- ✓ Se aplican todas las restricciones y disparadoress que se hayan definido en el destino para garantizar la calidad de los datos

Fig. 13. Carga en el proceso ETL

Proceso ETL

Otras aplicaciones de ETL

- ✓ Tareas de Bases de datos: para consolidar, migrar y sincronizar bases de datos operativas.
- ✓ Migración de datos entre diferentes aplicaciones por cambios de versión o cambio de aplicativos.
- ✓ Sincronización entre diferentes sistemas operacionales (por ejemplo, nuestro entorno ERP y la Web de ventas).
- ✓ Consolidación de datos: sistemas con grandes volúmenes de datos que son consolidados en sistemas paralelos para mantener históricos o para procesos de borrado en los sistemas originales.

Proceso ETL

Herramientas y aplicaciones ETL

- ✓ IBM Websphere DataStage
- ✓ Pentaho Data Integration (Kettle ETL) –
- ✓ SAS ETL Studio
- ✓ Oracle Warehouse Builder
- ✓ Cognos Decisionstream
- ✓ BusinessObjects Data Integrator (BODI)
- ✓ Microsoft SQL Server Integration Services (SSIS)

Conclusiones

- Es necesario ampliar las competencias de los alumnos en el uso de tecnologías emergentes como BI.

- Al alumno le permitirá tener un enfoque distinto sobre el modelado de datos y podrá mejorar su perspectiva en la forma en cómo se almacenan.

- Se espera que este trabajo sea un apoyo para entender formas distintas sobre sistemas de bases de datos.

Referencias Bibliografícas

- Connolly, T. (2009). Database Systems (5a. ed.). Mc Graw Hill.
- Date, C. J. (2001). Introducción a los Sistemas de Bases de Datos (7a. ed.). Perason Education.
- Silberschatz, A. (2006). Fundamentos de Bases de Datos (5a. ed.). Mc Graw Hill.
- Kimball Group. http://www.kimballgroup.com/data-warehouse-business-intelligence-courses/on-site-education-training-classes/dimensional-modeling-fundamentals/ Recuperado Ene-2015