Procesos de extracción, transformación y carga de datos

ETL

- Procedimientos (herramientas) destinados a obtener los datos de las fuentes operacionales, limpiarlos, convertirlos a los formatos de utilización y cargarlos en el repositorio final.

Integridad de datos

- Los datos cumplen condiciones de integridad cuando se ajustan a todos los stándares de valor y completitud.
- Todos los datos del DW son correctos
- El DW está completo (no existen más datos fuera de él).

Integridad de datos

- La credibilidad del DW depende de la integridad de sus datos
- El uso del DW depende de la percepción de los usuarios y de la confianza que tengan en su contenido.
- De la integridad de datos depende el éxito del proyecto.

Controles de Integridad

- Controles de Prevención : controlan la integridad antes de cargar los datos en el DW.
- Controles de Detección : aseguran la exactitud y completitud de la información una vez cargada en el DW.

Etapas del proceso ETL

- Migración de datos
- Limpieza
    - Transformación

(cálculos,agregados,sumarizaciones,desnormalización).

- Carga
- Conciliación - Validación

Migración

- Staging area : área de trabajo fuera del DW.
- El propósito de la migración es mover los datos de los sistemas operacionales a las áreas de trabajo (staging areas).
- NO se debe mover datos innecesarios (control preventivo).

Limpieza (Data cleaning)

- Corregir, estandarizar y completar los datos
- Identificar datos redundantes
- Identificar valores atípicos (outliers)
- Identificar valores perdidos (missings)

Limpieza (actividades)

- Se debe uniformar las tablas de códigos de los sistemas operacionales y simplificar esquemas de codificación
- Datos complejos, que representan varios atributos a la vez, deben ser particionados.

Transformación

- Son procesos destinados a adaptar los datos al modelo lógico del DW
- Se generan “reglas de transformación”.
- Las reglas deben validarse con los usuarios del DW

Transformación

- Generalmente el DW no contiene información de las entidades que - en los sistemas operacionales - son muy dinámicas y sufren frecuentes cambios.
- Si es necesario se utilizan Snapshots (fotos instantáneas)

Transformación

- La des-normalización de los datos tiene como propósito mejorar el rendimiento.
- Otro propósito es el de reflejar relaciones estáticas, es decir, que no cambian en una perspectiva histórica. Por ejemplo: producto - precio vigente al momento de facturación.

Transformación (sumarizaciones)

- Los datos sumarizados aceleran los tiempos de análisis.
- Las sumarizaciones también ocultan complejidad de los datos.
- Las sumarizaciones pueden incluir joins de múltiples tablas
- Las sumarizaciones proveen múltiples vistas del mismo conjunto de datos detallados (dimensiones).

Sumarizaciones (mantenimiento)

- El mantenimiento de las sumarizaciones es una tarea crítica.
- El DW debe actualizarlas a medida que se cargan nuevos datos.
- Debe existir alguna forma de navegar los datos hasta el nivel de detalle (drill down).
- La definición de la granularidad es un problema serio de diseño.

El nivel de granularidad: problema de diseño del DW

- Cúal es la unidad de tratamiento (fila)
- ¿Qué es un cliente? Una cuenta, un individuo, una familia
- ¿Cómo se sumariza la dimensión tiempo? Días, semanas, meses …?

Carga (Loading)

- Dos aproximaciones:
  - Full Refresh
  - Incremental
- Aunque el Full Refresh parece más sólido desde el punto de vista de la integridad de los datos, a medida que crece el DW se vuelve cada vez más difícil de realizar.

Controles de detección

- La validación de la carga del DW identifica problemas en los datos no detectados en las etapas anteriores.
- Existen dos maneras de hacer la validación:
  - completa (al final del proceso)
  - por etapas a medida que se cargan los datos

Controles de detección

- Los controles incluyen reportes que comparan los datos del DW con las fuentes operacionales a través de:
  - totales de control
  - número de registros cargados
  - valores originales vs valores limpios (transformados), etc.

Herramientas ETL

- Pueden ser procesos manuales diseñados a medida (querys SQL, programas en Visual Basic, etc).
- Existen herramientas que proporcionan interfaces visuales para definir joins, transformaciones, agregados, etc. sobre las plataformas mas comunes.