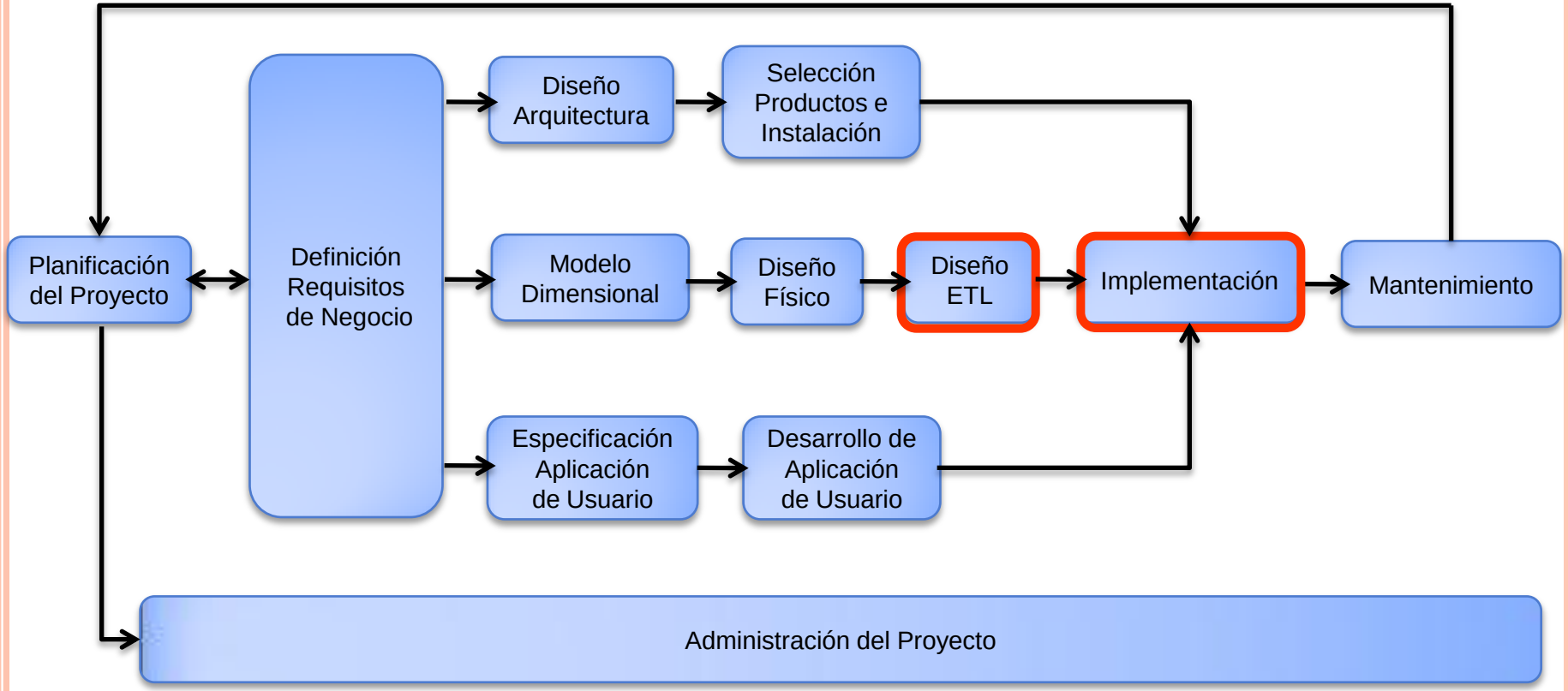
UNIVERSIDAD NACIONAL DE LA MATANZA
INTELIGENCIA DE NEGOCIOS

Proceso ETL - Implementación

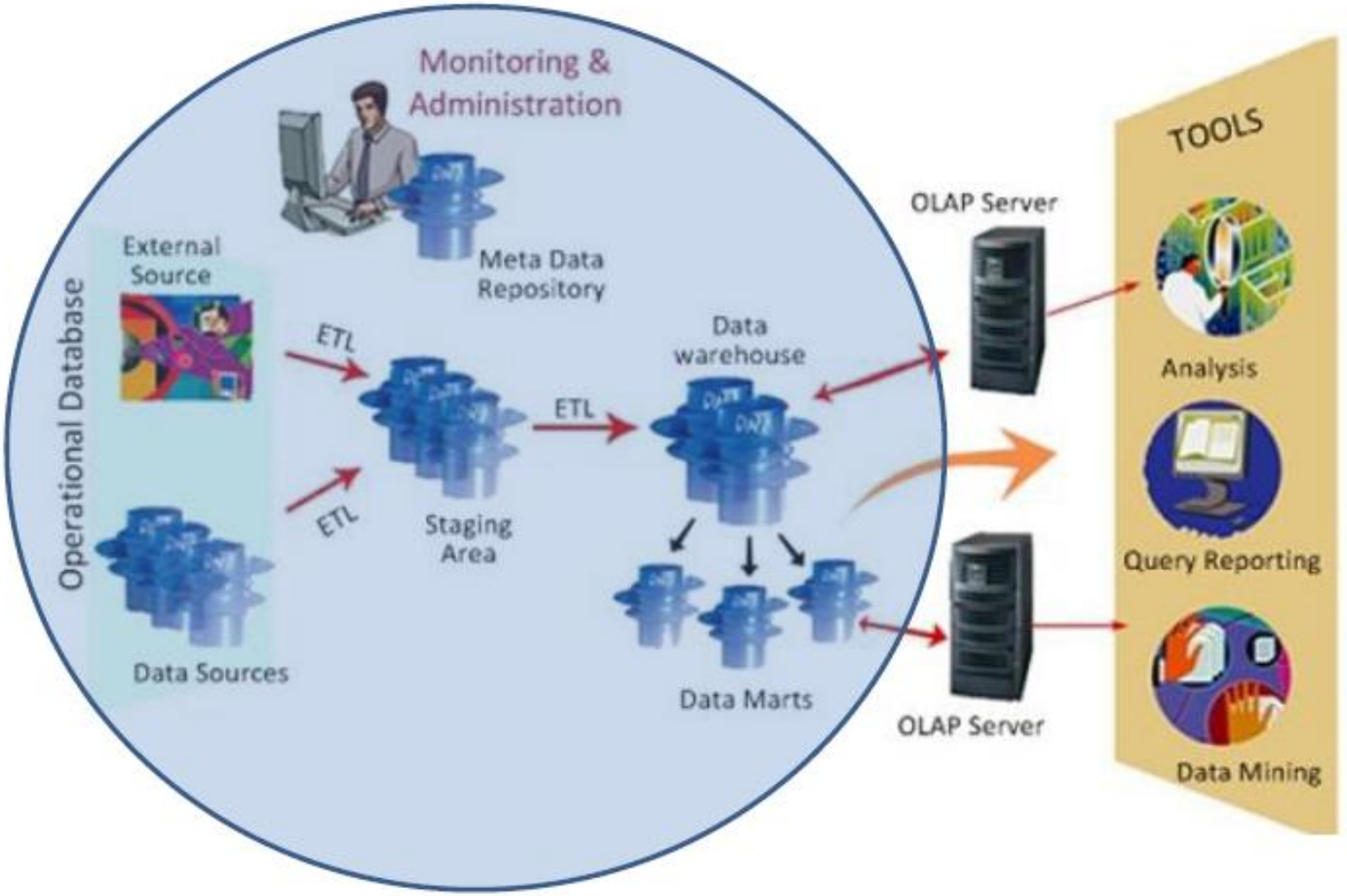
Profesor: Mg. Diego Basso

Curso 2017

Ciclo de Vida de un Proyecto de BI



Proceso ETL - Ubicación



Proceso ETL

- ¿Qué es ETL?
  - **E -** Extract (Extracción)
  - **T -** Transform (Transformación)
  - **L -** Load (Carga)
- Es el proceso que se define para tomar los datos de los sistemas fuente y cargarlos en el DW.
- Consume el 70-80% del tiempo y esfuerzo de la construcción de un DW.
- El proceso ETL es muy complejo y consume muchos recursos.

Proceso ETL - Funciones

- Transfiere información de los sistemas transaccionales y fuentes externas al DW.
- Guarda los datos luego de transferirlos al área de preparación o *Staging Area*.
  - Minimizar al máximo nivel los posibles errores o problemas en la fase de carga de los procesos ETL, normalmente se reserva un área de disco para poder recuperar los datos por etapas.
- Normaliza diversas fuentes de datos.
- Corrige datos incorrectos.
- Completa información faltante.
- Detecta problemas de calidad en los datos.

Proceso ETL - Funciones

- Mantiene las dimensiones del modelo dimensional.
- Agrega filas a las tablas de hechos.
- Genera métricas
  - Tiempos
  - Volúmenes de datos
  - Incidentes
- Permite puntos de control y auditoría.

Proceso ETL

- Alguno de los subsistemas que tiene que incluir:
  - **Data Quality**: son procesos y tecnologías que permiten asegurar la calidad de los datos a las necesidades de negocio.
  - **Data Profiling**: es el proceso de examinar los datos que existen en las fuentes de origen de una organización y recopilar estadísticas e información sobre los mismos.
  - **Data Cleansing**: es el proceso de detectar o descubrir y corregir datos corruptos, incoherentes o erróneos.

Proceso ETL

- Extracción
  - Se extraen los datos desde los distintos sistemas fuentes (internos y externos).
  - Datos con diferentes organización y formatos.
  - Se programa en horarios en que el impacto sea mínimo.
- Transformación
  - Aplica reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados.
  - Los datos se filtran, limpian, completan, homogenizan y se agrupan.

Proceso ETL

- Ejemplos de Transformación
  - Seleccionar sólo ciertas columnas para su carga
    - No cargar columnas con valores nulos
  - Traducir códigos
    - si la fuente almacena una "H" para Hombre y "M" para Mujer pero el DW tiene que guardar "1" para Hombre y "2" para Mujer.
  - Codificar valores libres
    - Convertir "Hombre" en "H“
  - Obtener nuevas medidas calculadas
    - Importe_Venta = cantidad * precio
  - Calcular totales de múltiples filas de datos
    - Ventas totales por cada producto

Proceso ETL

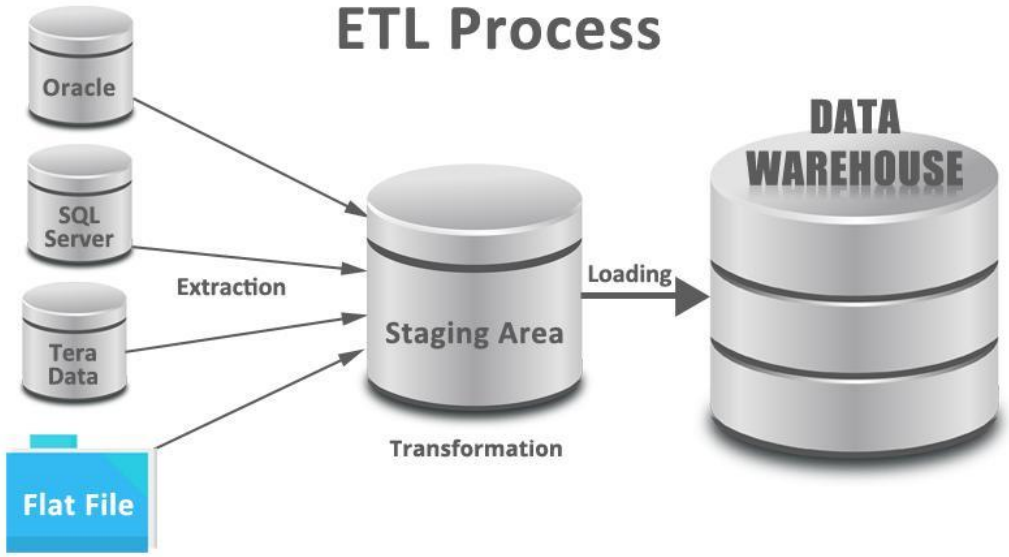
- Carga
  - Los datos transformados se ordenan, se consolidan, se verifica la integridad y se incorporan al DW (conjunto de tablas que van a consultar los usuarios).
  - En las dimensiones en general se hacen Inserts o Updates.
  - Lógica para manejar claves SK y claves operacionales.
  - Tabla de Hechos:
    - Separar inserts de updates: ideal es que sólo hayan inserts.
    - Carga incremental: distintos períodos de tiempo para la carga de datos.
    - Recuperación de fallas: Estar preparado ante fallas.
- Es un proceso repetitivo.

Staging Area

- Es un área temporal de almacenamiento de datos utilizada para el procesamiento de los mismos durante los procesos de ETL.
  - Se toman los datos necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos.
  - Una vez que los datos han sido traspasados, el DW se independiza de los sistemas fuentes hasta la siguiente carga.
  - Se suele añadir algún campo de fecha para que en el caso de falla evita empezar todo desde el principio.

Staging Area

- Se construyen y se implementan los procesos de extracción, limpieza, transporte, transformación y carga de los datos.
- Se utiliza una herramienta especializada en el tratamiento de grandes volúmenes de datos.



Errores Típicos en los Datos

- Datos Incompletos
  - Registros o campos faltantes
- Datos Incorrectos
  - Códigos erróneos
  - Cálculos o agregaciones incorrectas
  - Registros duplicados
- Datos Incomprensibles
  - Varios campos dentro de un campo
  - Formato extraño en los datos
  - Códigos desconocidos
  - Formatos de archivo extraños

Errores Típicos en los Datos

- Datos Inconsistentes
  - Uso inconsistente de la codificación
  - Inconsistencia en el significado de los códigos
  - Códigos superpuestos
  - Códigos distintos con igual significado
  - Inconsistencia en nombres y direcciones
  - Reglas de negocio inconsistentes
  - Agregaciones inconsistentes
  - Inconsistencia en la granularidad del nivel atómico de datos

Herramientas ETL

- Adoptar una herramienta o hacerlo nosotros mismos.

Herramientas ETL	Codificación Manual
Específicas	Flexibilidad
Metodologías ETL	Técnicas de programación estándar
Personal capacitado en herramienta	Recursos propios
Mayor costo inicial	Menor costo inicial
Conectores múltiples	Se deben desarrollar conectores
Simplifica el mantenimiento	El mantenimiento se complejiza con el tiempo
Auto documentación (meta data)	Se debe desarrollar
Trazabilidad de los datos	Muy difícil de conseguir trazabilidad