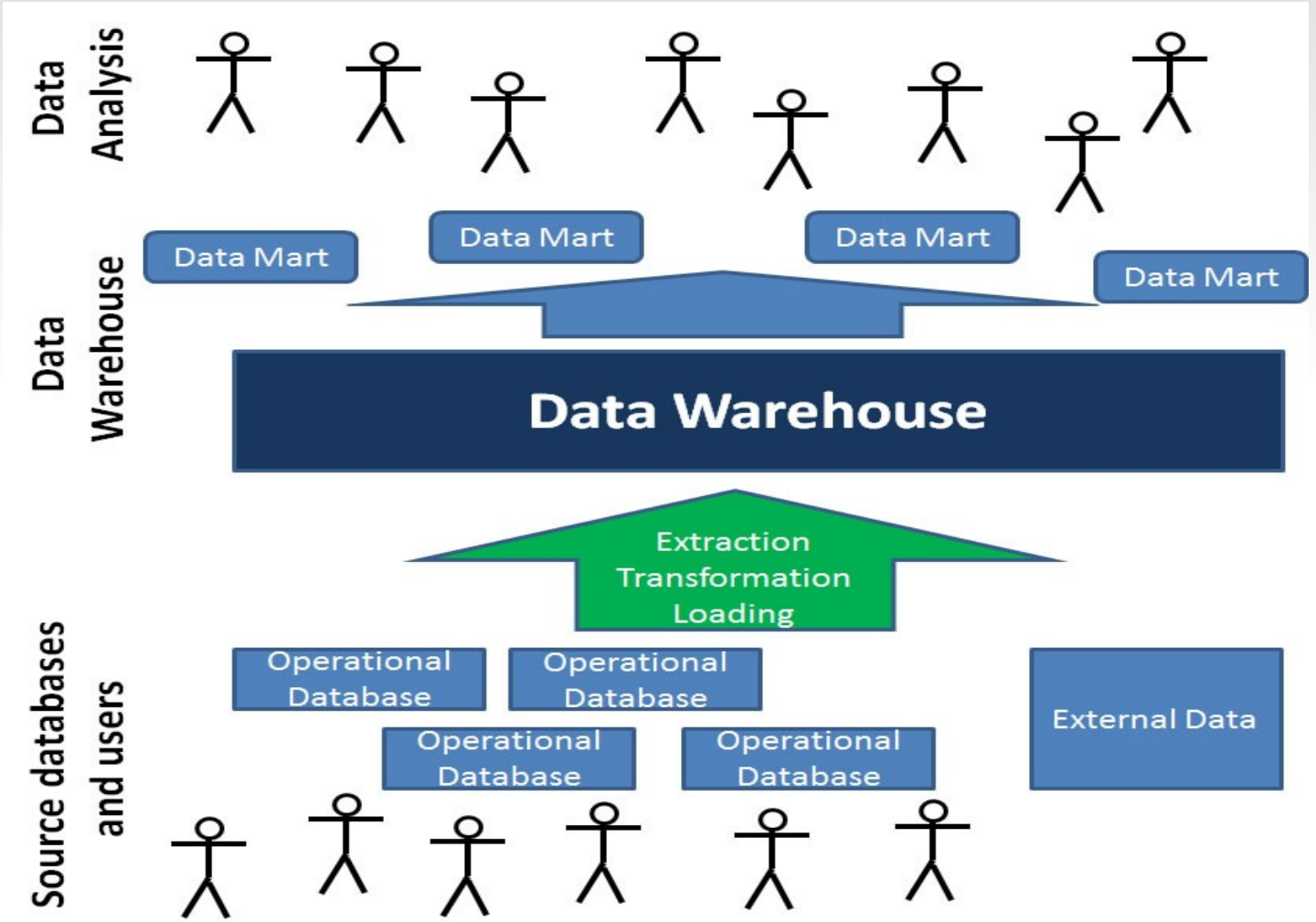
CI2355 – Almacenes de datos y OLAP

UNIVERSIDAD DE
COSTA RICA

Arquitectura



Propiedades esenciales

- Separación
  - El procesamiento transaccional y el analítico deben mantenerse lo más separados posible
- Escalabilidad
  - Las arquitecturas de hardware y software deben ser fáciles de actualizar cuando aumentan el volumen de datos, que tienen que ser administrados y procesados, o aumentan los requerimientos de los usuarios, requerimientos que deben ser atendidos

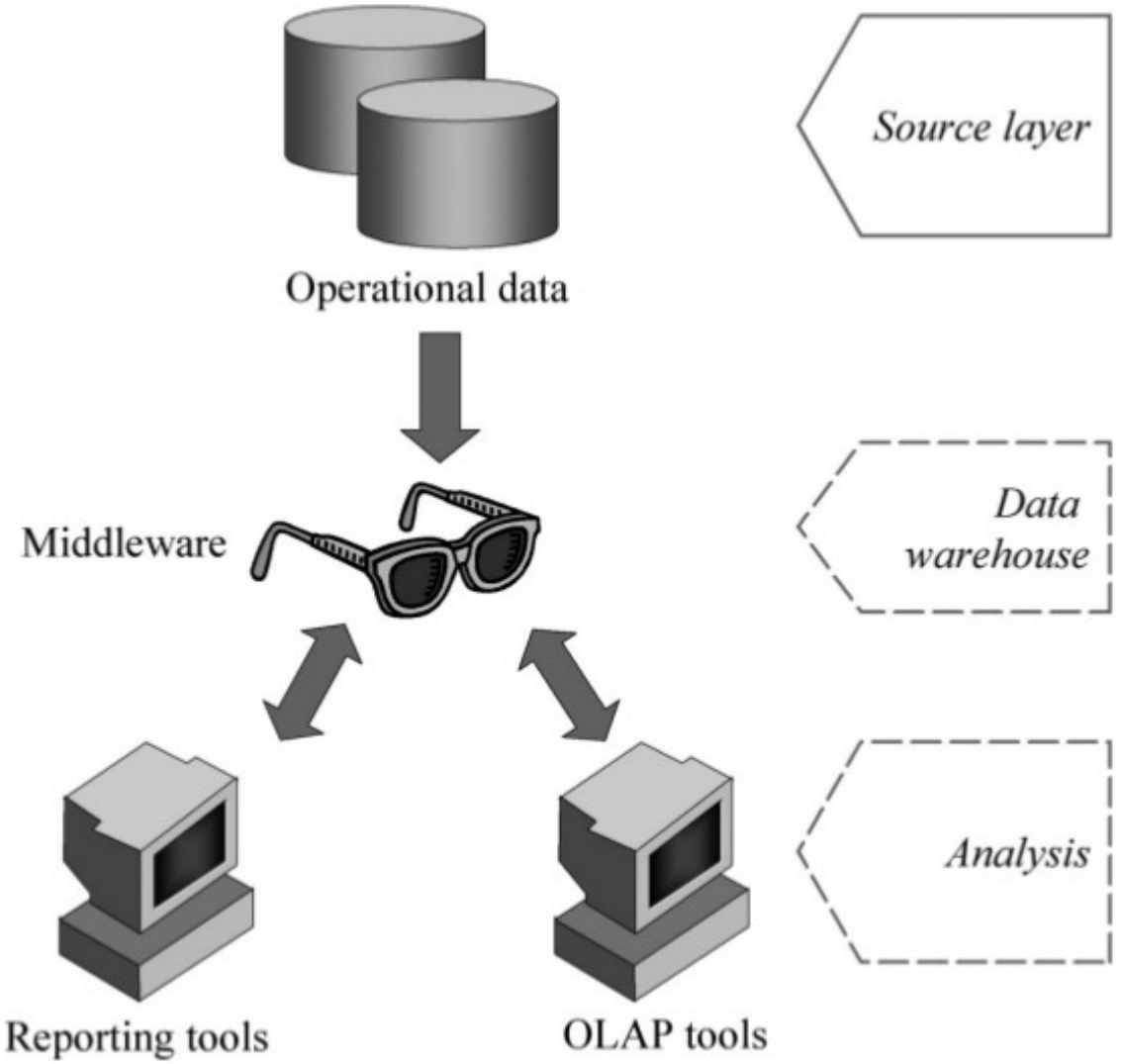
Propiedades esenciales (2)

- Extensibilidad
  - La arquitectura debe ser capaz de alojar nuevas aplicaciones y tecnologías sin tener que rediseñar todo el sistema
- Seguridad
  - Monitorear accesos es esencial porque lo que se guarda en los almacenes de datos es estratégico
- Administrabilidad
  - La gestión del almacén de datos no debe ser particularmente difícil

Clasificación de las arquitecturas

- Orientada a la estructura
  - Una capa
  - Dos capas
  - Tres capas
- Empleo de las capas
  - Orientada a la empresa
  - Orientada a departamentos

Arquitectura de una capa

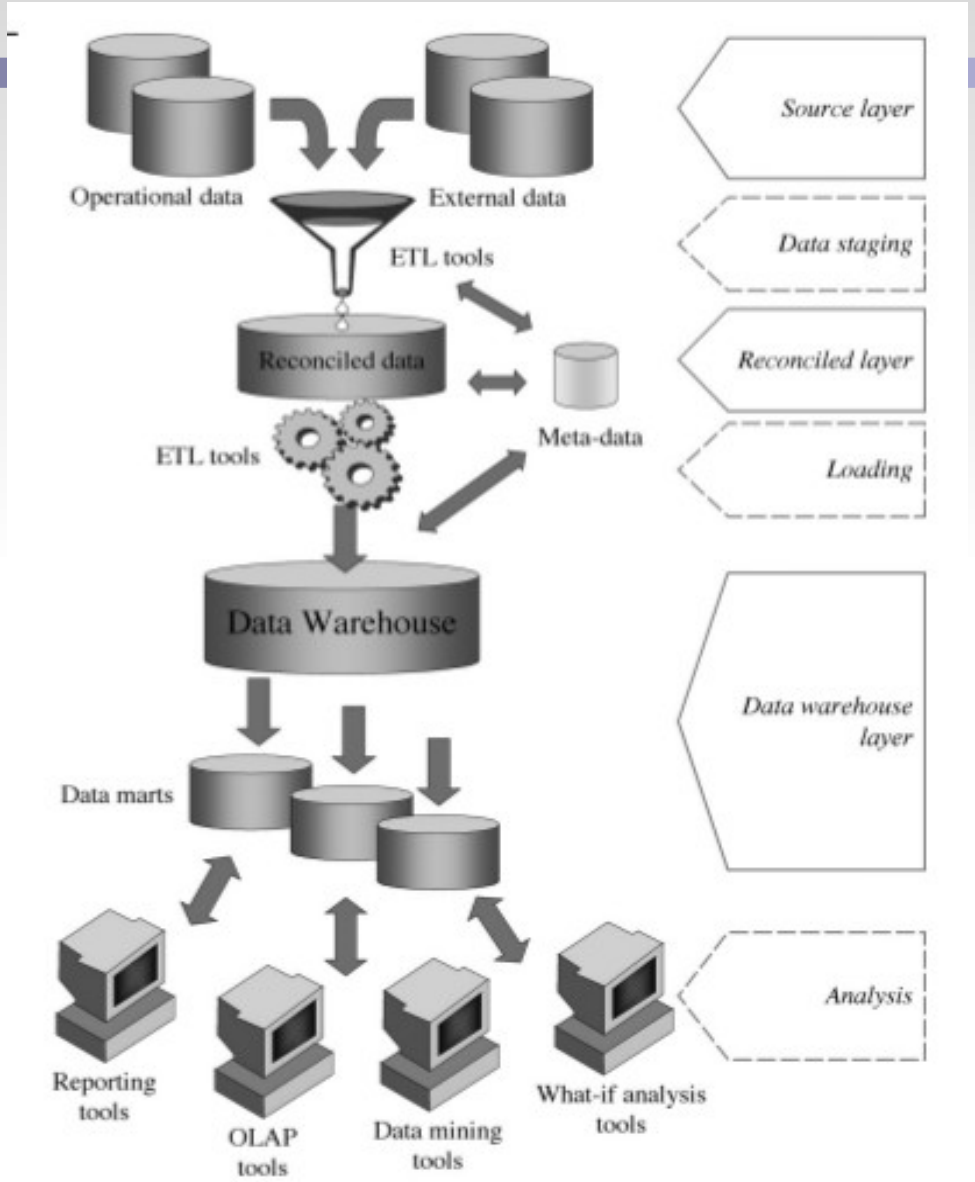


Arquitectura de dos capas

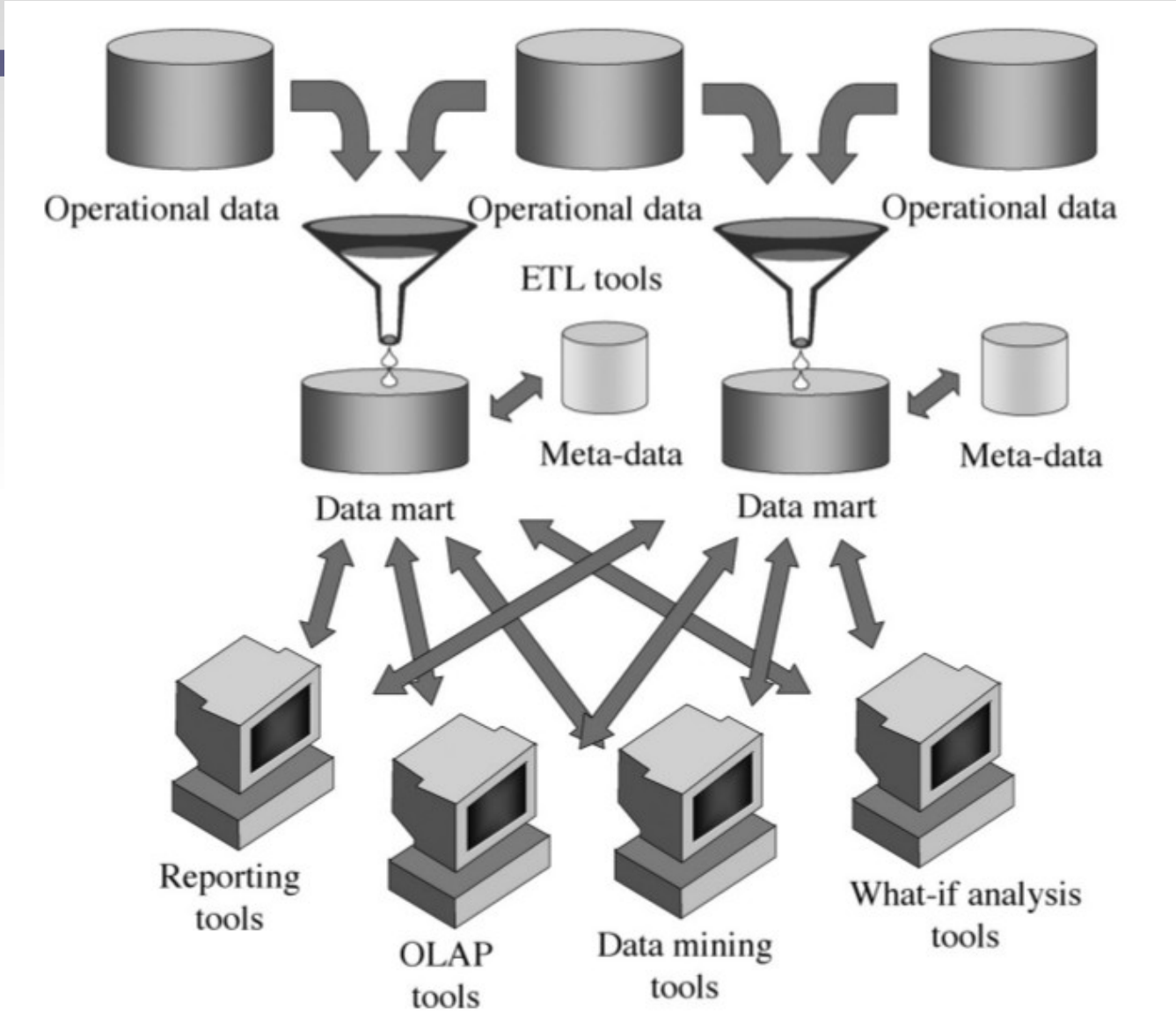
Ventajas

- Información de calidad siempre disponible en el AD, aunque tengamos el acceso a las fuentes temporalmente denegado
- Las consultas al AD no afectan el rendimiento de la BD
- AD están lógicamente estructurados de acuerdo con el modelo multidimensional
- Se evitan las diferencias de tiempo y granularidad entre sistemas OLTP y sistemas OLAP
- AD pueden diseñarse para optimizar para aplicaciones de análisis y reportes

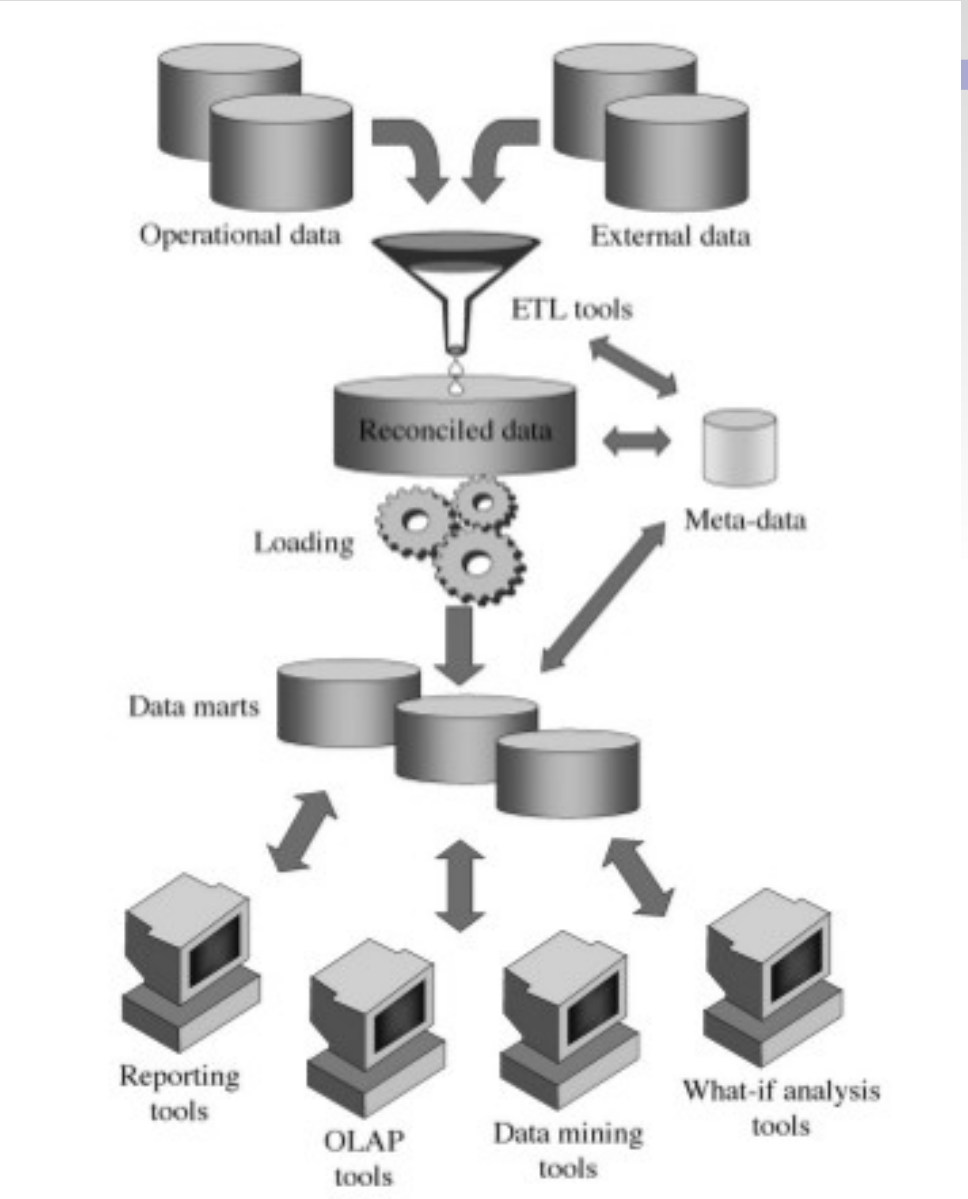
Arquitectura de tres capas



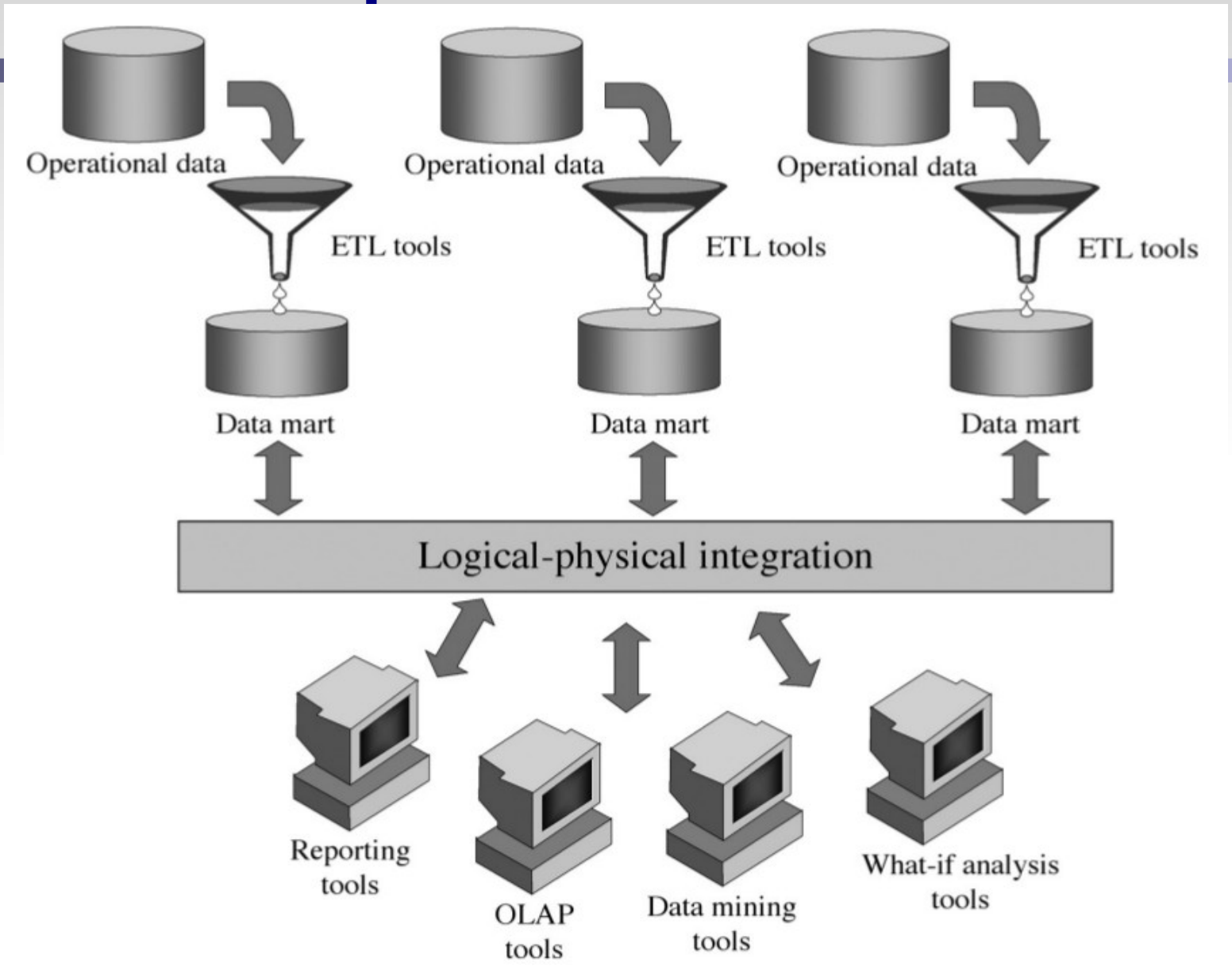
Data marts independientes



Hub-and-spoke



Arquitectura federada

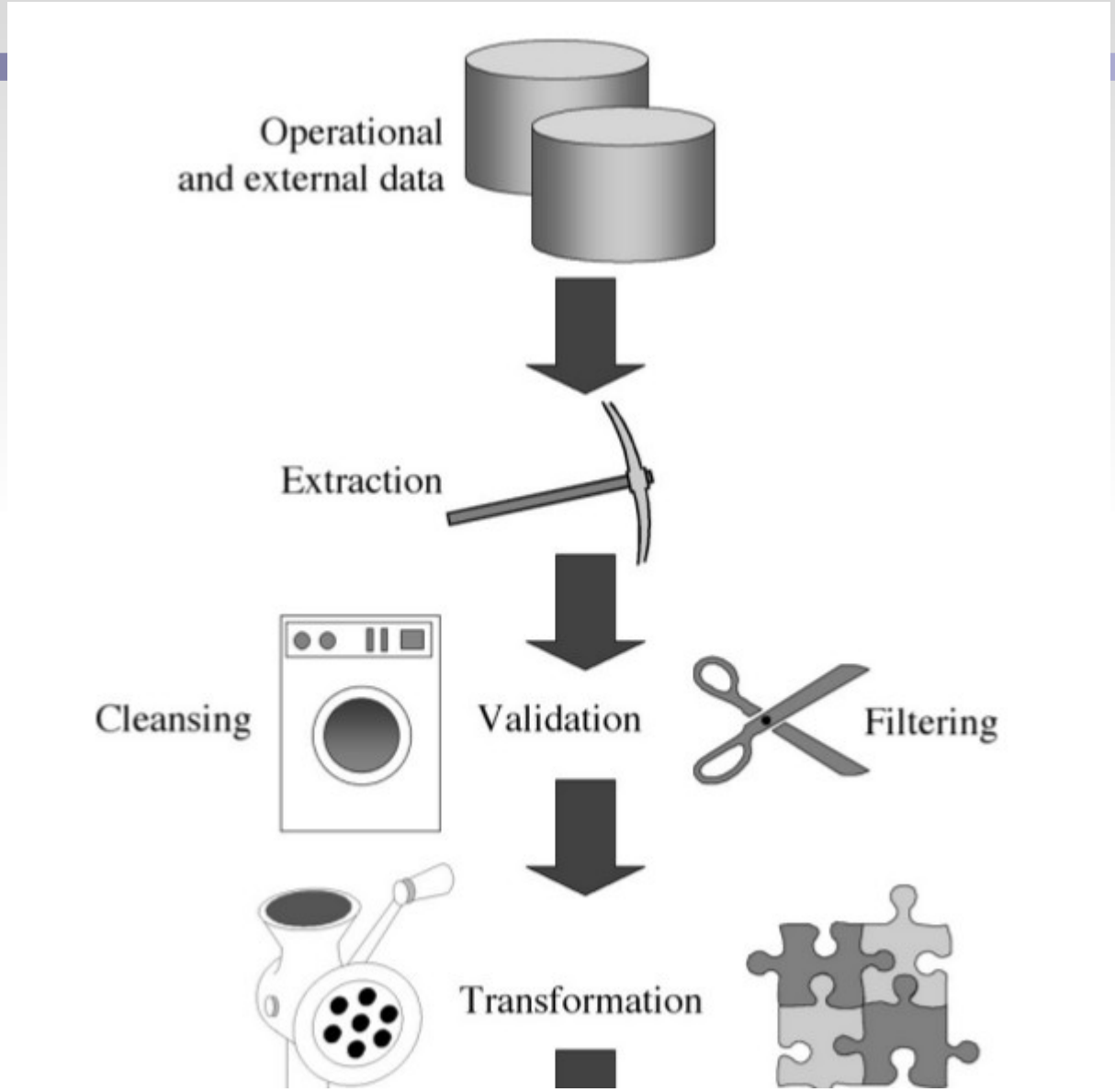


Capa de preparación de datos

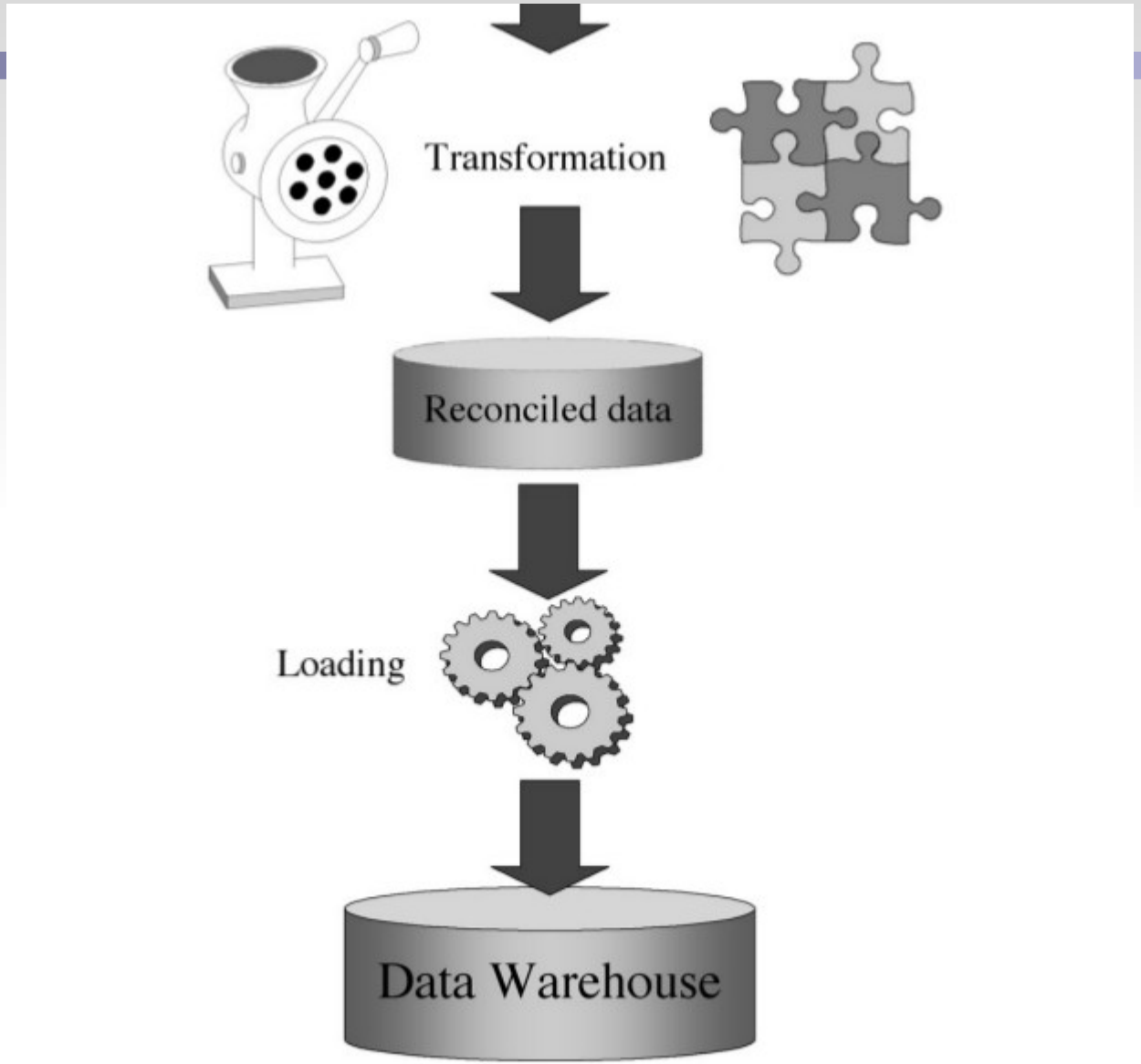
- Preparación de datos (data staging)
  - Extracción o captura
  - Limpieza (cleansing, cleaning o scrubbing)
    - Rectificación y homogenización de datos
  - Transformación
    - Normalización de formatos
  - Carga (loading)

→ ETL

ETL


Operational
and external data
Extraction
Cleansing
Validation
Filtering
Transformation

ETL (cont.)



Extracción

- Obtener datos desde las fuentes
  - Estática
  - Incremental
- Calidad de los datos a ser extraídos
  - Exactitud de las restricciones
  - Formatos adecuados
  - Claridad del esquema

Cleansing

- Errores e inconsistencias
  - Datos duplicados
    - Misma persona, diferente ID
  - Valores inconsistentes asociados lógicamente
    - Dirección y código postal
  - Datos faltantes
    - Dirección de trabajo
  - Uso inesperado de los campos
    - Número telefónico en lugar de número de ID

Cleansing (2)

- Errores e inconsistencias (cont.)
  - Valores erróneos o imposibles
    - 30 de febrero, 2012
  - Valores inconsistentes con una única entidad porque se usan diferentes prácticas
    - Estados Unidos – EE.UU. – U.S.A.
  - Valores inconsistentes para una entidad individual por errores de transcripción
    - Costarricense – costarricence

Transformación

- Formato operacional de las fuentes

  →

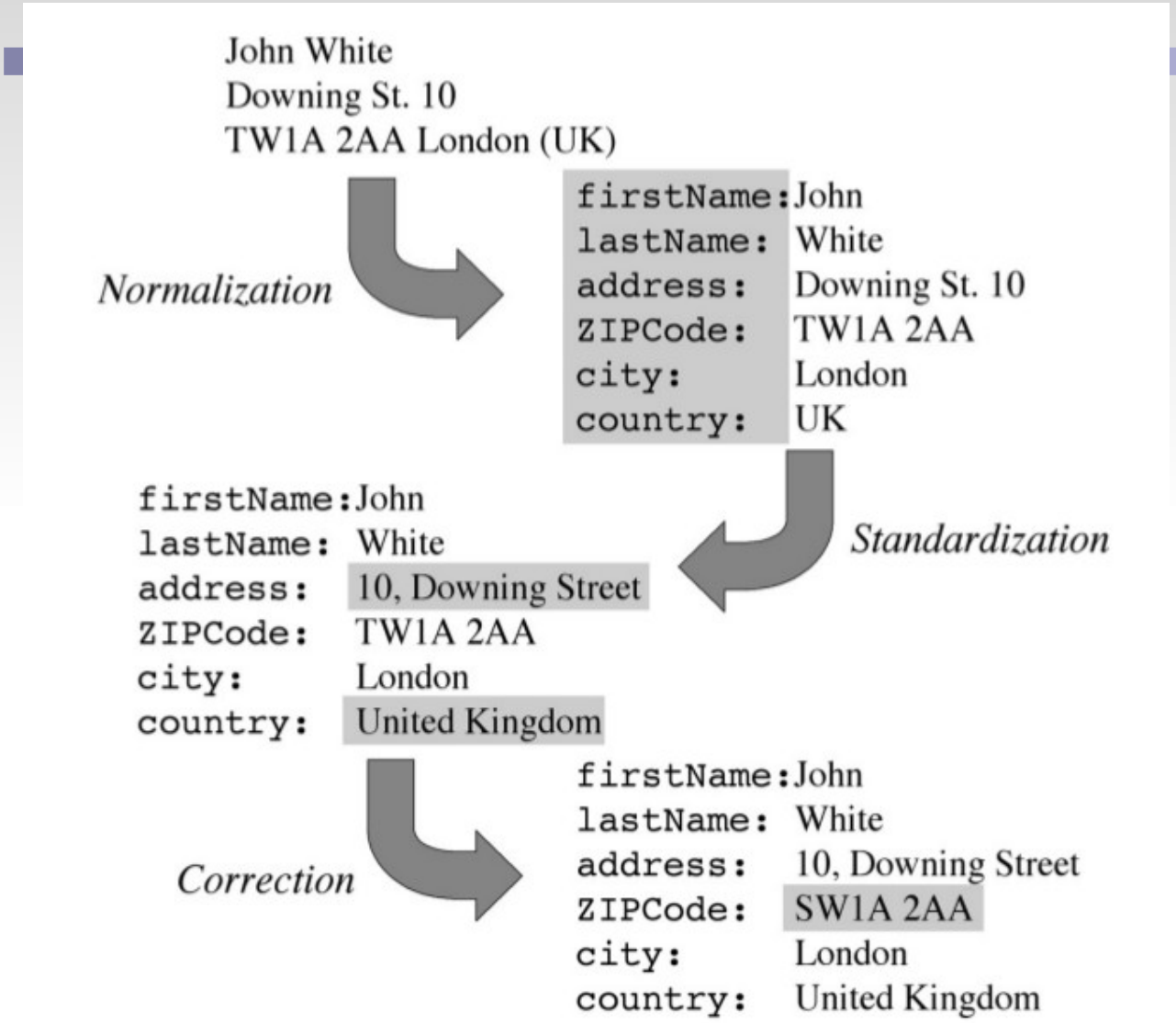
- Formato del almacén de datos

- Texto libre podría esconder datos valiosos
  - Pérez e Hijos, S.R.L
- Diferentes formatos para datos individuales
  - 2012-03-12
  - Año: 2012, mes: 3, día: 12

Tareas de transformación

- Conversión y normalización
  - Formatos de almacenaje y unidades de medida para uniformar los datos
- Pareo (matching)
  - Campos equivalentes en diferentes fuentes
- Selección
  - Reducir el número de campos y registros en la fuente

Ejemplo


John White
Downing St. 10
TW1A 2AA London (UK)
Normalization
firstName: John
lastName: White
address: Downing St. 10
ZIPCode: TW1A 2AA
city: London
country: UK
Standardization
firstName: John
lastName: White
address: 10, Downing Street
ZIPCode: TW1A 2AA
city: London
country: United Kingdom
Correction
firstName: John
lastName: White
address: 10, Downing Street
ZIPCode: SW1A 2AA
city: London
country: United Kingdom

Carga

- Refrescar
  - Datos son reescritos completamente
  - Datos anteriores son reemplazados
  - Normalmente utilizada en combinación con la extracción estática para poblar el AD
- Actualizar
  - Sólo se agregan al AD los datos que cambiaron en las fuentes
  - Normalmente, no se borran ni modifican datos
  - Utilizada en combinación con la extracción incremental

Referencias

- Matteo Golfarelli, Stefano Rizzi. *Data Warehouse Design: Modern Principles and Methodologies.* McGraw-Hill, 2009