Fundamentos de Análisis de Datos

Unidad 4

Limpieza de datos

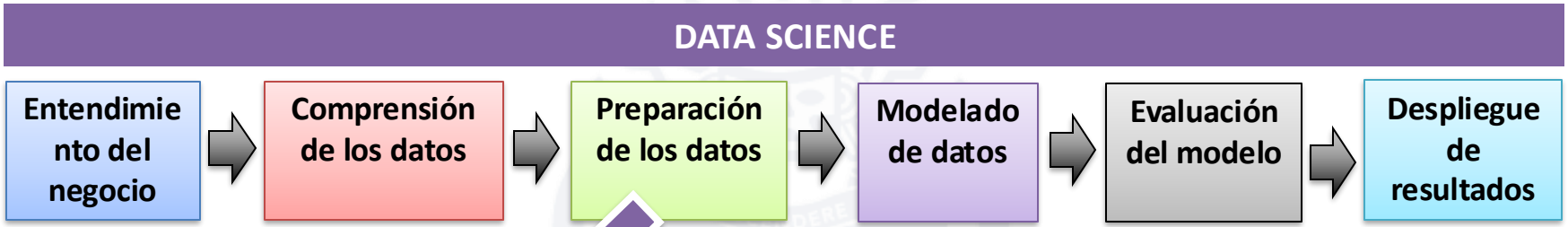
Referencias

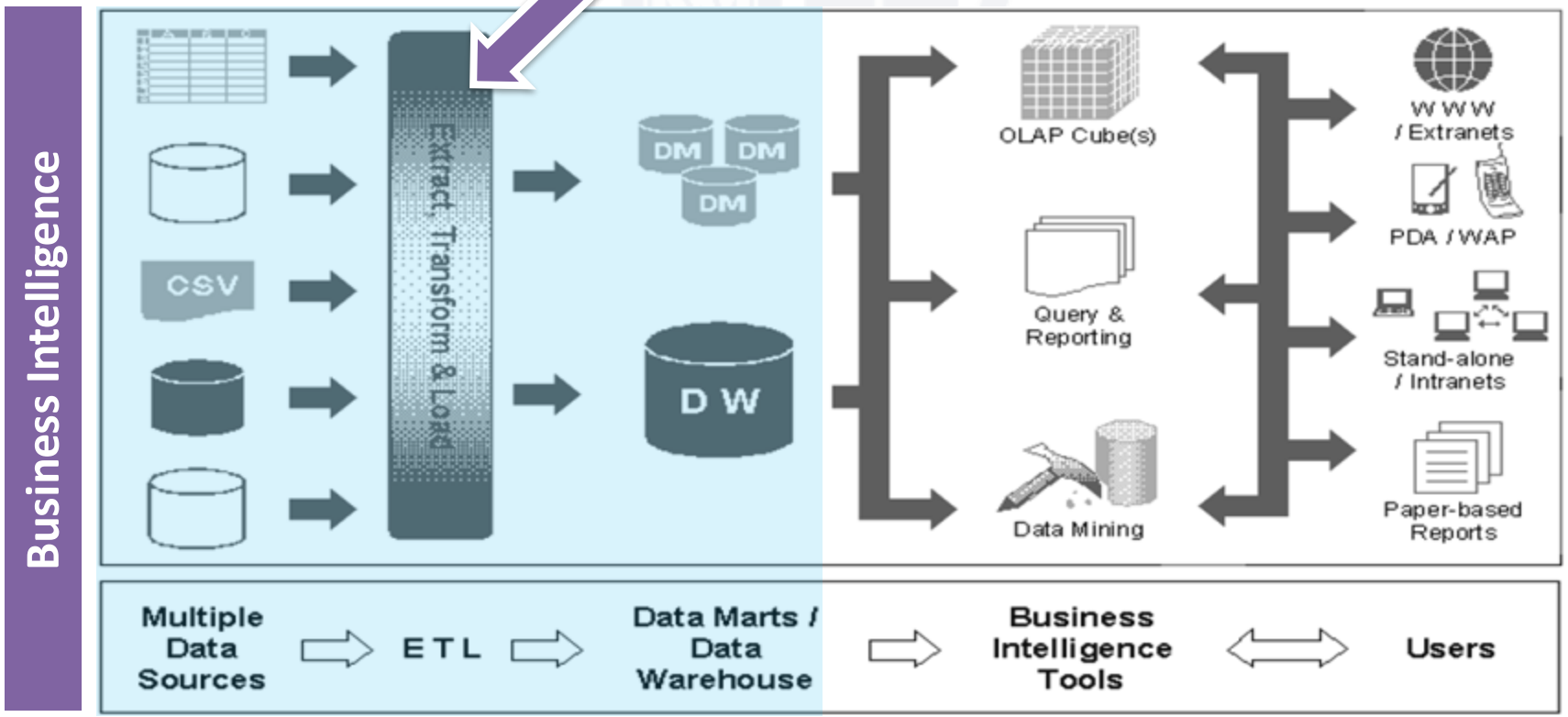
- Menoyo Ros, D. García López, E. & García Cabot, A. (2021). *Fundamentos de la ciencia de datos:* (ed.). Editorial Universidad de Alcalá. https://elibro.net/es/ereader/bibliotecautpl/177631?page=11 (apartado 12.1, pág 359)

Agenda

- Preprocesamiento de datos
  - Limpieza de datos
  - Transformación de datos
  - Enriquecimiento de datos
- Tratamiento de valores nulos
- Tratamiento de outliers

Proceso ETL


DATA SCIENCE
Entendimiento del negocio
Comprensión de los datos
Preparación de los datos
Modelado de datos
Evaluación del modelo
Despliegue de resultados


Business Intelligence
CSV
Extract, Transform & Load
DM
DM
DM
D W
OLAP Cube(s)
Query & Reporting
Data Mining
W W W / Extranets
PDA / WAP
Stand-alone / Intranets
Paper-based Reports
Multiple Data Sources
E T L
Data Marts / Data Warehouse
Business Intelligence Tools
Users

Preprocesamiento de datos

- El preprocesamiento de los datos es crucial para garantizar **calidad y consistencia** de los datos antes de realizar análisis o construir modelos.
- Se realiza una vez cumplida la recolección y consolidación de los datos desde las fuentes (**extracción**)
- Se parte con una **inspección** de los datos para entender la estructura, tipos de datos, y valores iniciales.
- Se puede complementar con la generación de **estadísticas iniciales** (media, mediana, desviación típica) para comprender la distribución y características de los datos origen.

Preprocesamiento de datos

- Conlleva tareas de limpieza, transformación, y enriquecimiento de los datos.
  - **Limpieza de datos**: Asegurar que los datos sean correctos, completos y consistentes.
  - **Transformación de datos**: Adaptar los datos para facilitar su análisis o modelado.
  - **Enriquecimiento de datos**: Mejorar la calidad y utilidad de los datos originales

Limpieza de datos

- La limpieza de datos asegura que los datos estén en un estado adecuado para el análisis.
- Se trata de corregir inconsistencias en cuanto a formato y estructura de los datos origen.
- Se enfoca en la **calidad y validez del dato original**.

Técnicas de Limpieza de Datos

- **Detección y tratamiento de valores faltantes (nulos, vacíos)**: eliminación o imputación con estadísticos simples.
- **Corrección ortográfica y de formato**: nombres, mayúsculas/minúsculas, símbolos.
- **Eliminación de duplicados** exactos o por claves primarias.
- **Filtrado de registros inválidos o inconsistentes**: edades negativas, fechas imposibles.
- **Conversión de tipos erróneos**: texto mal interpretado como número, fechas como string.
- **Tratamiento de valores atípicos (outliers):** Decidir si eliminarlos, ajustarlos o mantenerlos

Transformación de los datos

- Los datos limpios se someten a una serie de operaciones que los convierte en datos ajustados e idóneos para aplicar distintas técnicas de análisis.
- Adapta datos para **hacerlos analizables o modelables**.
- Transformar datos no es sólo cambiar números, es darle a nuestros modelos y visualizaciones los **insumos más adecuados y significativos** para trabajar

Técnicas de transformación de datos

- **Codificación de variables categóricas**: one-hot encoding, label encoding.
- **Normalización y estandarización** de escalas numéricas.
- **Creación de variables discretas** (binning): agrupar edades en rangos, por ejemplo.
- **Aplicación de funciones matemáticas**: logaritmos, potencias, escalas, para mejorar la distribución, reducir asimetrías
- **Cambio de estructura**: pivotear datos, convertir entre formato ancho/largo.
- **Extracción de componentes de variables**: extraer año/mes de una fecha, dominio de un correo, etc.
- **Reducción de la dimensionalidad**: PCA, selección de variables, agrupación

Enriquecimiento de los datos

- Se refiere al proceso de mejorar, refinar, y potenciar los datos, previo a su análisis.
- Se trata de complementar los datos origen con otros que puedan ayudar a mejorar la calidad y precisión del análisis a realizar.
- Agrega **nueva información útil que no estaba presente**
- El enriquecimiento no modifica los datos originales, sino que **los potencia** con nueva información que permite **analizar mejor, predecir más y decidir con contexto**

Técnicas de enriquecimiento de los datos

- **Creación de variables derivadas**: relaciones, tasas, diferencias, acumulados.
- **Cálculo de agregados por grupo**: promedio por cliente, conteo por categoría.
- **Cruce (join) con otras fuentes** de datos externas o auxiliares.
- **Incorporación de metadatos**: regiones geográficas, clasificaciones externas.
- **Categorización avanzada** con reglas de negocio o scores.
- **Uso de APIs o scraping** para agregar datos faltantes o complementarios.

Herramientas

- R
  - **dplyr**: Paquete para la manipulación de datos
  - **tidyr**: Paquete para la limpieza y reestructuración de datos
  - **stringr**: Paquete para la manipulación de cadenas de texto
  - **janitor**: Proporciona funciones para la limpieza inicial de datos
  - **lubridate**: Paquete para trabajar con fecha y datos temporales.
- Python
  - **Pandas**: Biblioteca esencial para la manipulación y análisis de datos estructurados.
  - **NumPy**: Utilizado para operaciones matemáticas y lógicas en arrays.
  - **SciPy**: Complementa a NumPy con funciones estadísticas y matemáticas avanzadas.
  - **Scikit-learn**: Proporciona herramientas para pre-procesamiento, imputación, escalado y más.

Herramientas

- SQL
  - Funciones de agregación y filtrado: GROUP BY, HAVING, WHERE para limpiar y agregar datos
- Herramientas BI
  - **Tableau**: Ofrece capacidades de preparación de datos y limpieza visual.
  - **Power BI**: Incluye Power Query para la manipulación y limpieza de datos.
- Otras herramientas
  - **OpenRefine**: Herramienta poderosa para la exploración y limpieza de datos.
  - **Trifacta**: Plataforma que facilita la preparación de datos con una interfaz visual y algoritmos avanzados.

Tratamiento de valores faltantes

- Los valores faltantes (nulos, cadenas vacías) pueden afectar negativamente el análisis y los modelos de aprendizaje automático.
- Técnicas:
  - **Eliminación de filas**: Si los valores nulos son pocos y su eliminación no afecta significativamente el análisis
  - **Eliminación de columnas**: Si es hay una alta proporción de valores nulos y no es crucial para el análisis
  - **Imputación de valores nulos**: usar media, mediana, moda, u otros valores constantes. O técnicas de machine learning para predecir los valores faltantes
  - **Relleno adelante o atrás**: Utilizado en series temporales donde los valores nulos se rellenan con el valor anterior o siguiente disponible