Introducción a cs de datos · Conceptos

I. tipo de dato

A. Estructurado : - organizado en un formato predefinido y homogeneo
(Excel, BDD sql,) - está en tablas con filas y columnas

B. semiestructurados: - no están ordenados con filas-columnas
(JSON, XML, mails) - Están jerarquizados; utilizan sistema de llaves / etiquetas
similar a un diccionario en python. tiene "key-values"

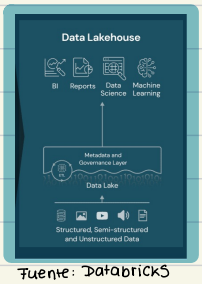
C. no estructurada: - no tiene estructura ni jerarquía
(logs, post rrss, ★ Se pueden almacenar en DATA LAKES!!
PDF)

OBS: un dataset es un formato tabular en donde los datos (registros) se organizan por atributos.

II. ¿Cómo Guardamos y organizamos los datos?

a. Data warehouse → enfoque con fin analítico
- Integra distintas bases de datos
- Cumple modelo ACID: Atomicidad Consistencia Aislamiento Durabilidad
↳ en resumen Garantiza la integridad y fiabilidad en procesos dentro de la base de datos

b. Data lakehouse
- Permite Guardar y organizar datos estructurados, semiestructurados y no estructurados en un solo lugar / sistema
- Flexible y escalable



Fuente: Databricks

III. ¿De Donde sacamos datos?

III.a Web

Ⓐ→ URL (clásica descarga archivos)

Ⓑ→ API ("interfaz de programación de aplicaciones"): Permite la comunicación entre 2 componentes de software a través de reglas y protocolos (funcionan en sentido cliente-servidor)

OBS: Para el curso se utilizan las librerías requests y beautifulSoup
→ métodos en requests: .get(), .post() / peticiones protocolo http

NOTA: apis pueden ser públicas o privadas, las respuestas que se obtienen pueden estar en formato JSON.

Ⓒ→ Webscrapping!!: extraer automatizadamente los datos de una página web
★ Aquí entra beautifulSoup → trae datos html o XML !!

Por: Francisca Meepsen

★ Concepto: DATA WRANGLING ⟶ Proceso de limpiar, transformar y organizar los datos

★ concepto2: ETL (extraer transformar cargar)

Ambos tienen el obj de preparar los datos pero varían en la ejecución y formato

- ETL es más repetitivo, para grandes volúmenes, formal y planificado
- Data wrangling es más flexible y enfocado en análisis específicos que se requieran. semi automatizado/manual

Por: Francisca Yepsen