udp Posgrados

FACULTAD DE CIENCIAS SOCIALES E HISTORIA

Modelamiento de Datos para la Reproducibilidad y Comunicación de Resultados

Dr (c) René Lagos Barrios
Programa de Doctorado en Salud Pública UCH

Jueves 5 de diciembre de 2024

Referencias

- Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling. John Wiley & Sons. (Capítulo 1)
- Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Leveraging the data lake: Current state and challenges. Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21, 179–188.
- Kaur, H., & Kaur, G. (2019). Comprehensive Survey of OLAP Models. En N. Yadav, A. Yadav, J. C. Bansal, K. Deep, & J. H. Kim (Eds.), Harmony Search and Nature Inspired Optimization Algorithms (Vol. 741, pp. 415–422). Springer Singapore. https://doi.org/10.1007/978-981-13-0761-4_40
- Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. Ouellette F, editor. PLoS Comput Biol. 22 de junio de 2017;13(6):e1005510.
- Granger, B. E., & Perez, F. (2021). Jupyter: Thinking and Storytelling With Code and Data. Computing in Science & Engineering, 23(2), 7–14. https://doi.org/10.1109/MCSE.2021.3059263

¿Cuál es el proyecto en que has participado en que el manejo de datos ha sido más desafiante o problemático?

¿Por qué?

Gestión de datos en computación científica (Wilson et al, 2017)

1. Guarda los datos originales (raw data)
    a. No los limpies ni alteres
    b. Guarda los procedimiento para descargarlos
2. Respalda los datos en más de una ubicación (pc local y en la nube)
3. Crea datos fáciles de entender por humanos
    a. Nombres de variables, códigos
4. Crea datos amigables para el análisis:
    a. Una columna por variable
    b. Una observación por fila
5. Registra todos los pasos para procesar los datos
6. Usa múltiples tablas con identificadores únicos para cada registro (ej. sexo)
7. Comparte los datos en repositorios que otros puedan usar y citar

Arquitecturas de Datos

Arquitectura de ETL: Extracción, Transformación y Carga (Kimball y Ross, 2013)

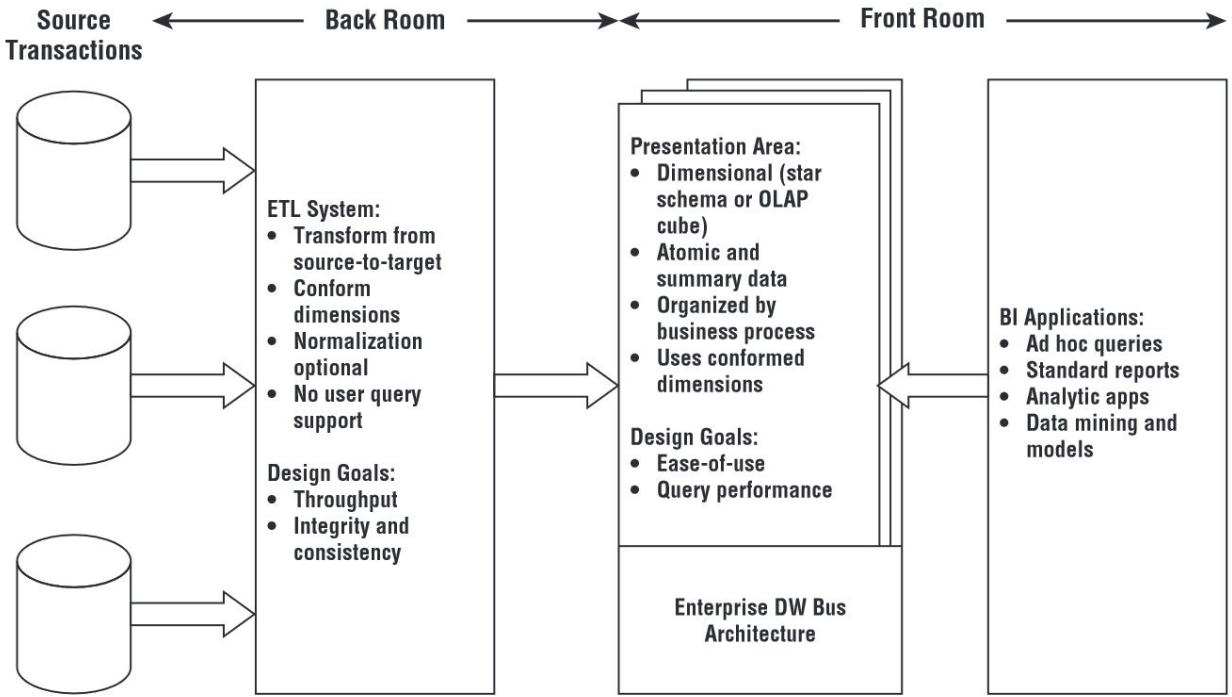


Figure 1-7: Core elements of the Kimball DW/BI architecture.

- Áreas de trabajo
- Bases de datos
- Procesos

Arquitectura de ETL: Extracción, Transformación y Carga (Kimball y Ross, 2013)

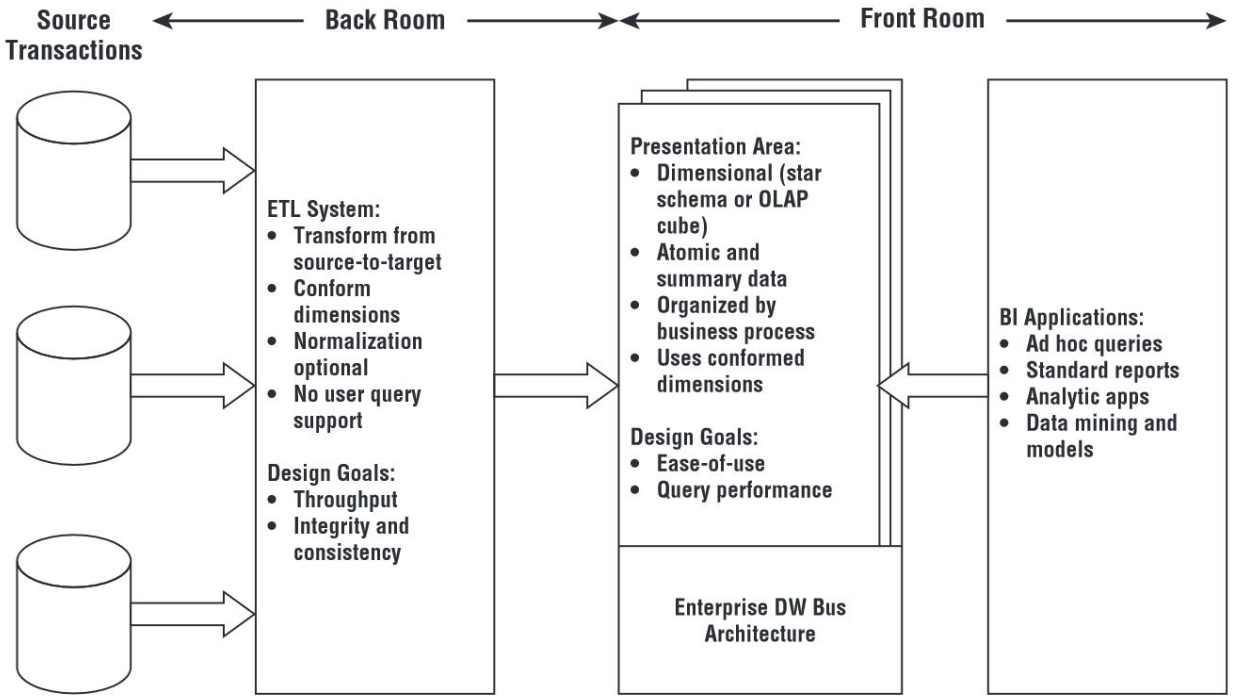


Figure 1-7: Core elements of the Kimball DW/BI architecture.

Extracción

-Obtener y entender los datos para poder procesarlos en el área de trabajo

Transformación

-Limpieza: eliminar duplicados, imputar datos faltantes, geocodificar
-Normalización
-Integración

Carga

-Estructurar datos para análisis
-Cargar en área de presentación (análisis)

Área de trabajo

Ambiente privado, donde se realiza la extracción y transformación de datos.

Datos normalizados (Data warehouse)

Consolidan entidades/eventos con datos limpios, codificados, integrados, sin redundancias ni ambigüedades.

Área de presentación

Ambiente donde los datos son compartidos con otros usuarios.

Datos multidimensionales (Data mart)

Orientados al análisis de relaciones entre variables para responder preguntas.

Arquitectura de Datamarts independientes (Kimball y Ross, 2013)

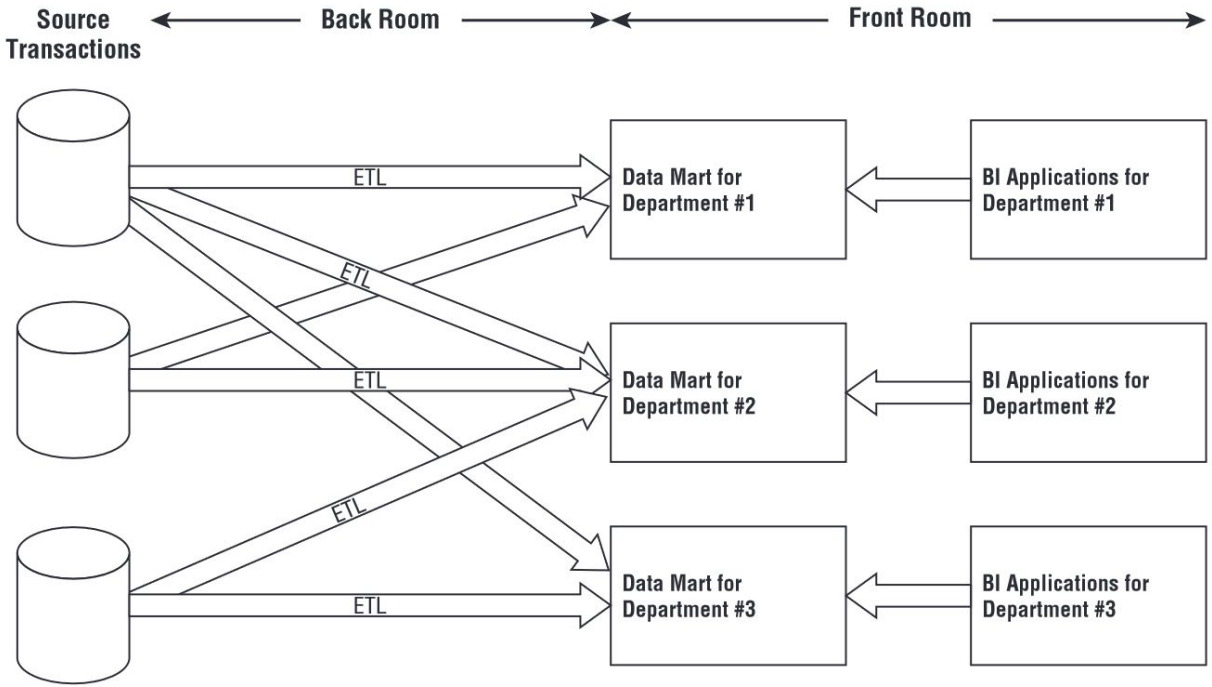


Figure 1-8: Simplified illustration of the independent data mart "architecture."

Arquitectura de Data lake (Nargesian et al, 2019)

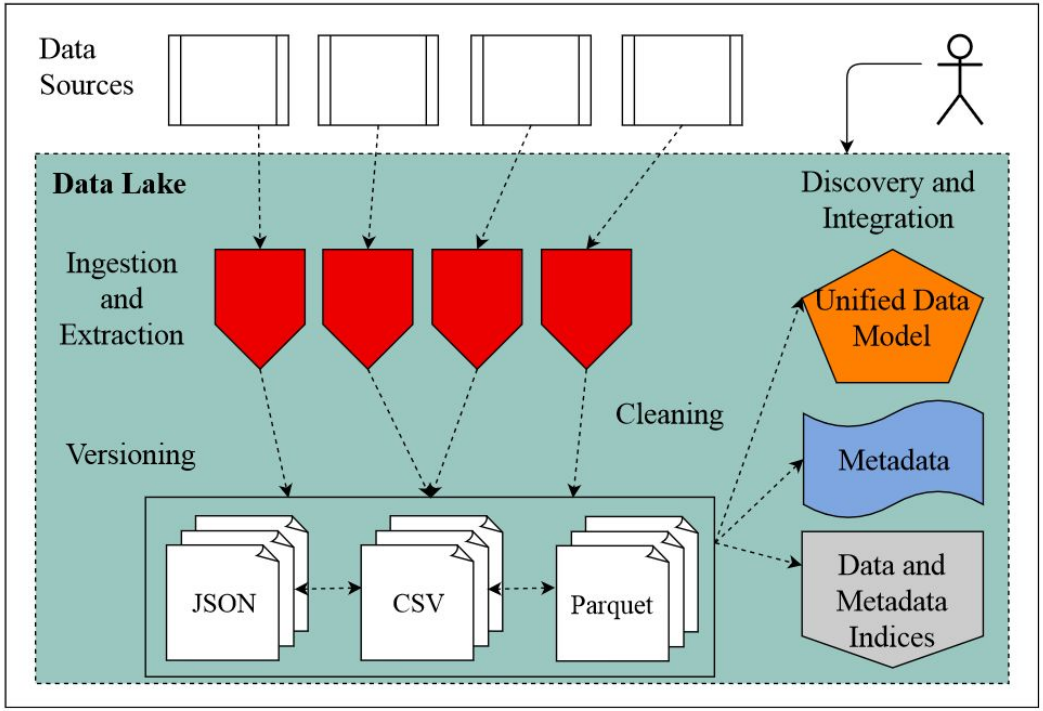


Figure 1: Example Data Lake Management System.

Ejercicio J1

1. Revise el script Ejercicio J1 (Colab) ¿qué hace? ¿cuán complejo le parece?
2. Separe el script en las fases de Extracción, Transformación, Carga y Análisis ¿cuán complejo le parece el código?
3. ¿Cree que vale esfuerzo de separar los bloques en este caso? ¿Por qué?

Normalización de datos

Normalización de Bases de Datos

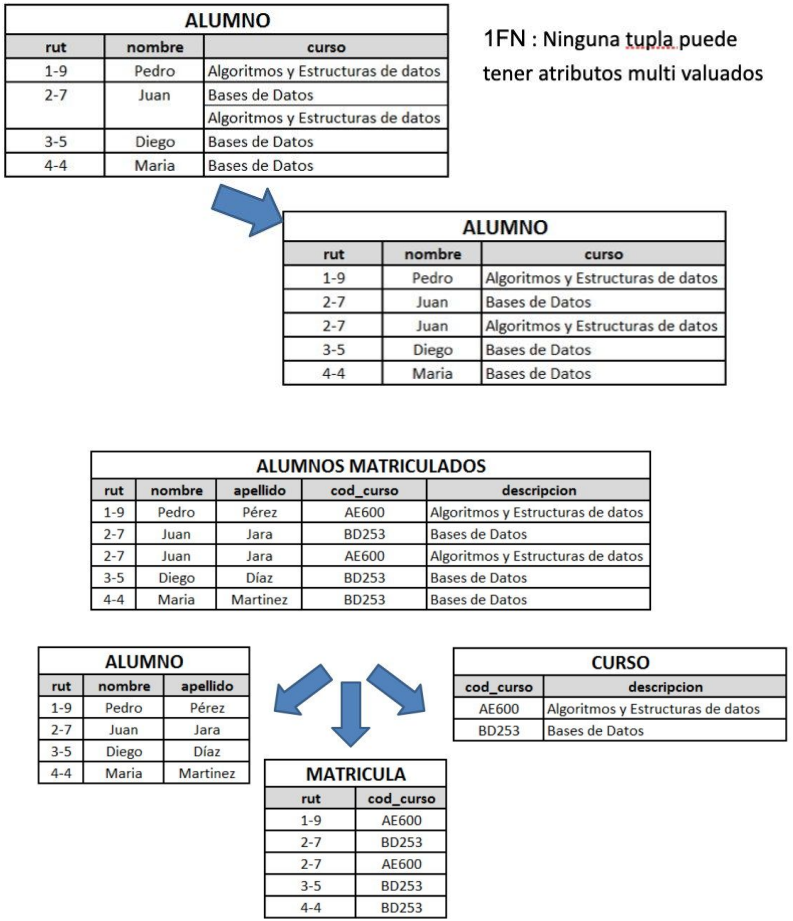
Proceso que busca minimizar la redundancia y ambigüedad de los datos, mejorar la integridad y facilitar el mantenimiento de la información.

Formas Normales

1ra Forma Normal (1FN): Cada celda debe contener un único valor atómico, es decir, no listas ni registros.

2da Forma Normal (2FN): Cada atributo no clave debe depender funcionalmente de la clave primaria completa.

3ra Forma Normal (3FN): Ningún atributo no clave debe depender funcionalmente de otro atributo no clave.



ALUMNO

rut	nombre	curso
1-9	Pedro	Algoritmos y Estructuras de datos
2-7	Juan	Bases de Datos
		Algoritmos y Estructuras de datos
3-5	Diego	Bases de Datos
4-4	Maria	Bases de Datos

1FN : Ninguna tupla puede tener atributos multi valuados

ALUMNO

rut	nombre	curso
1-9	Pedro	Algoritmos y Estructuras de datos
2-7	Juan	Bases de Datos
2-7	Juan	Algoritmos y Estructuras de datos
3-5	Diego	Bases de Datos
4-4	Maria	Bases de Datos

ALUMNOS MATRICULADOS

rut	nombre	apellido	cod_curso	descripcion
1-9	Pedro	Pérez	AE600	Algoritmos y Estructuras de datos
2-7	Juan	Jara	BD253	Bases de Datos
2-7	Juan	Jara	AE600	Algoritmos y Estructuras de datos
3-5	Diego	Díaz	BD253	Bases de Datos
4-4	Maria	Martinez	BD253	Bases de Datos

ALUMNO

rut	nombre	apellido
1-9	Pedro	Pérez
2-7	Juan	Jara
3-5	Diego	Díaz
4-4	Maria	Martinez

MATRICULA

rut	cod_curso
1-9	AE600
2-7	BD253
2-7	AE600
3-5	BD253
4-4	BD253

CURSO

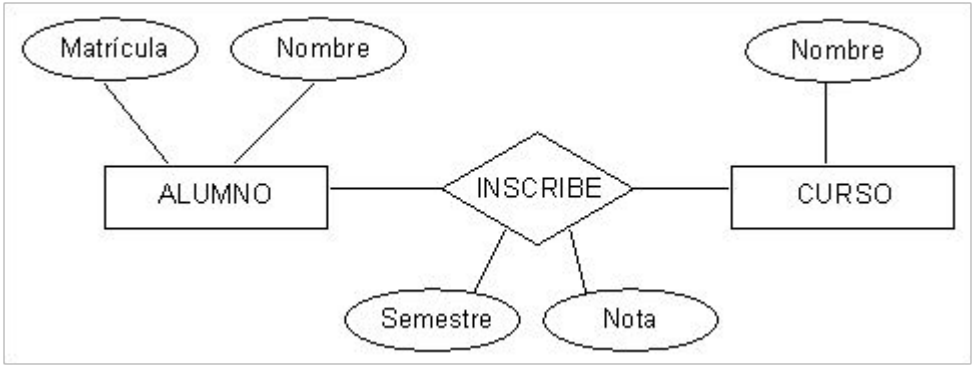
cod_curso	descripcion
AE600	Algoritmos y Estructuras de datos
BD253	Bases de Datos

Fuente de la imagen: https://bookdown.org/paranedagarcia/database/normalizacion.html

Modelo Entidad-Relación

Herramienta gráfica que representa la **estructura conceptual** de una base de datos. Permite capturar el **significado de los datos**, mediante la visualización de las entidades, los atributos de esas entidades y las relaciones entre ellas que describen los datos.

Permite aplicar las reglas de normalización y obtener un diseño lógico de la base de datos.



- **Entidad**: "cosa" material o abstracta con existencia propia
  - **Llave primaria**: atributo o combinación de atributos, que identifica a cada entidad en forma única
- **Atributos**: propiedades que describen a las entidades
- **Relación**: asociación entre entidades, puede tener atributos propios
  - **Llave foránea**: identifican entidades asociadas

Fuente de la imagen: https://users.dcc.uchile.cl/~nbaloian/cc20a/post/e-r.html

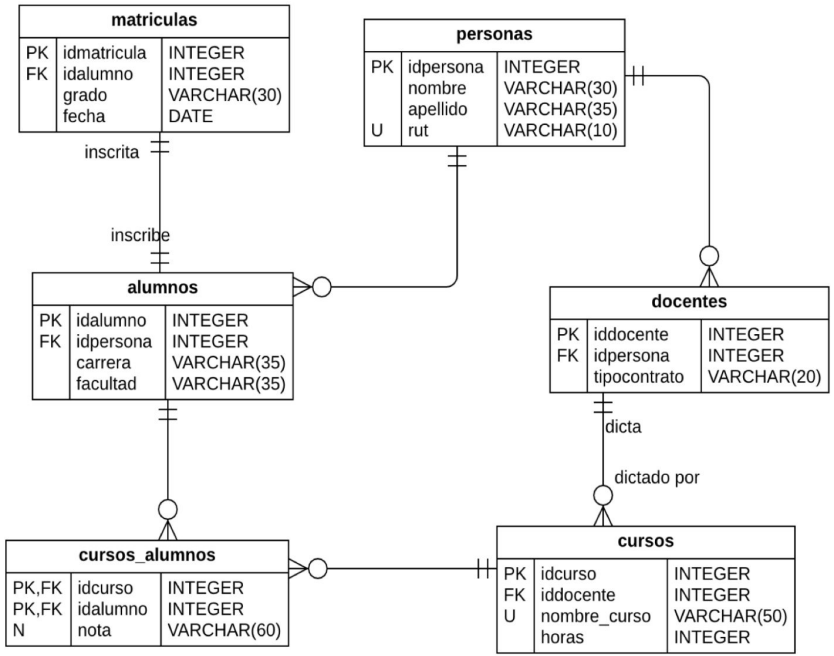
Modelo Físico de una Base de Datos

Representación detallada de cómo se almacenan los datos en el disco duro. Permite optimizar el rendimiento de la base de datos, considerando factores como el tamaño de los datos, la frecuencia de acceso y las operaciones más comunes.

Tablas: Estructuras para almacenar los datos, con filas y columnas.

Tipos de datos: Los tipos de datos que se asignan a cada columna (entero, texto, fecha, etc.).

Relaciones: relaciones entre tablas y llaves



Fuente de la imagen: https://bookdown.org/paranedagarcia/database/modelamiento-de-datos.html

Ejercicio J2: Modelo de datos de población beneficiaria de APS Universal

Suponga que usted participa en la reforma de Atención Primaria Universal y necesita dimensionar la población beneficiaria de esta política

1. Formar grupos de 3-4 personas
2. Descargue y revise la base de datos de:
   a. Población Beneficiaria de Fonasa y población inscrita en Atención Primaria de Salud en 2022: https://www.fonasa.cl/sites/fonasa/datos-abiertos/estadisticas-anuales
   b. Población total por comuna (INE) https://www.ine.gob.cl/docs/default-source/proyecciones-de-poblacion/cuadros-estadisticos/base-2017/estimaciones-y-proyecciones-2002-2035-comunas.xlsx?sfvrsn=8c87fc3f_3
3. Elabore un diagrama Entidad-Relación que identifique las entidades y atributos que describen las bases de datos
4. Construya un script para normalizar la base de datos
5. Construya una tabla para cada entidad

udp Posgrados

FACULTAD DE CIENCIAS SOCIALES E HISTORIA

Modelamiento de Datos para la Reproducibilidad y Comunicación de Resultados

Dr (c) René Lagos Barrios
Programa de Doctorado en Salud Pública UCH

Sábado 7 de diciembre de 2024

Agenda

Jueves:

- Recapitulación
- Modelamiento multidimensional de datos
- **Break 10:15**
- Ejercicio: crear un datamart

-Raw data
-Data normalizada
-Análisis de datos
-Data warehouse
-Data mart

Normalización y limpieza de Comunas

Script en Colab: Normalizar Comunas.ipynb

- ✓ Combina código, texto y visualizaciones de resultados en un mismo documento
- ✓ Código estructurado con lógica ETL: extracción > transformación > carga
- ✓ Llegar y correr. Sin necesidad de instalar software o carpetas.
- ✓ Códigos generados con ayuda de inteligencia artificial

Quiz

¿Para qué sirve separar el área de trabajo y el área de presentación?

¿Para qué sirve la normalización de los datos?

¿Para qué sirve un diagrama de Entidad-Relación de una base de datos?

¿Para qué sirve un data warehouse?

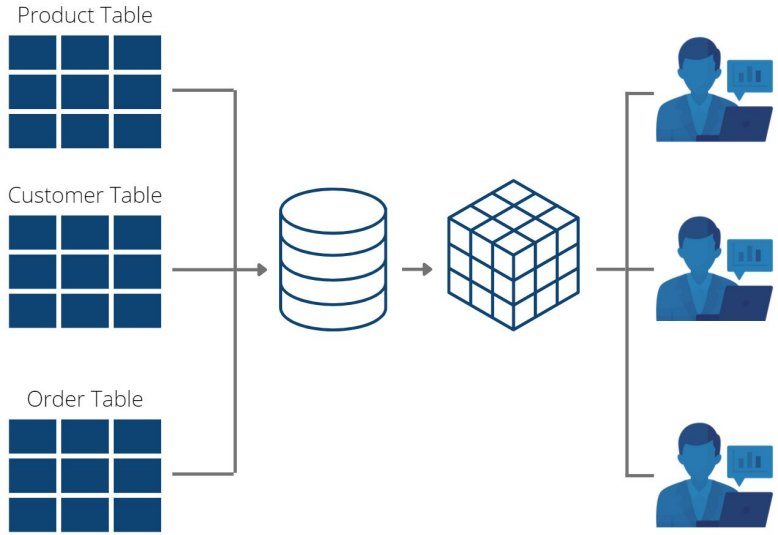
Modelamiento Multidimensional de datos

Análisis Multidimensional de Datos (Online Analytical Processing, OLAP)

Permite analizar los datos desde múltiples perspectivas.

Facilita la agrupación de datos y la creación de resúmenes a diferentes niveles de detalle.

Optimiza el rendimiento en consultas sobre grandes volúmenes de datos.



Fuente imagen:
https://towardsdatascience.com/online-analytical-processing-olap-and-its-influence-on-data-science-c386bc96a736

Cubos de datos

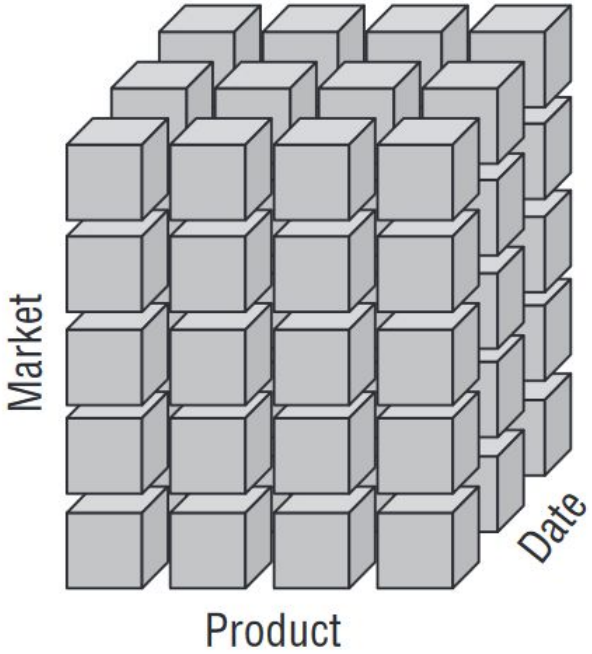
Representación **conceptual** de la estructura de datos para análisis

Métricas: los valores numéricos que se analizan (eventos, individuos, costos).

Dimensión: categorías para estratificar los datos (ejemplo, tiempo, localización, características demográficas).

Operaciones para analizar datos:

- *Slice and dice*: cortar y picar = estratificar y pivotear
- *Drill down*: desglosar
- *Roll-up*: agrupar



Ejemplo: Tabla Dinámica

Ejercicio V1: modelos de datos multidimensionales

1. Formar grupos de 3-4 personas
2. Seleccione uno de los siguientes dashboards:
   a. Población Beneficiaria Fonasa
   b. Población Inscrita en Atención Primaria de Salud
   c. Otro dashboard con indicadores sociales
3. Identifique las métricas y dimensiones del dashboard
4. ¿Qué operaciones permite realizar sobre los datos? (filtrar, estratificar, desglosar, agregar)

Modelo Estrella

Fact table: tabla con los “hechos” de interés, las métricas que lo cuantifican y los atributos básicos que lo caracterizan. Cada fila representa un hecho único (grano).

- *Granularidad*: nivel de detalle máximo de los hechos.

Dimensión: agrupación de atributos relacionados que caracterizan los hechos (tiempo, localización, etc)

- *Jerarquías*: relaciones entre atributos según nivel de agregación (por ejemplo, año, trimestre, mes).

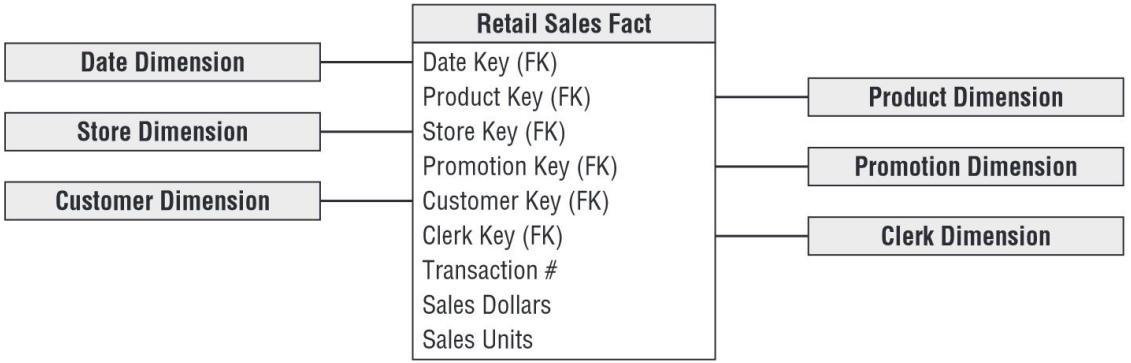
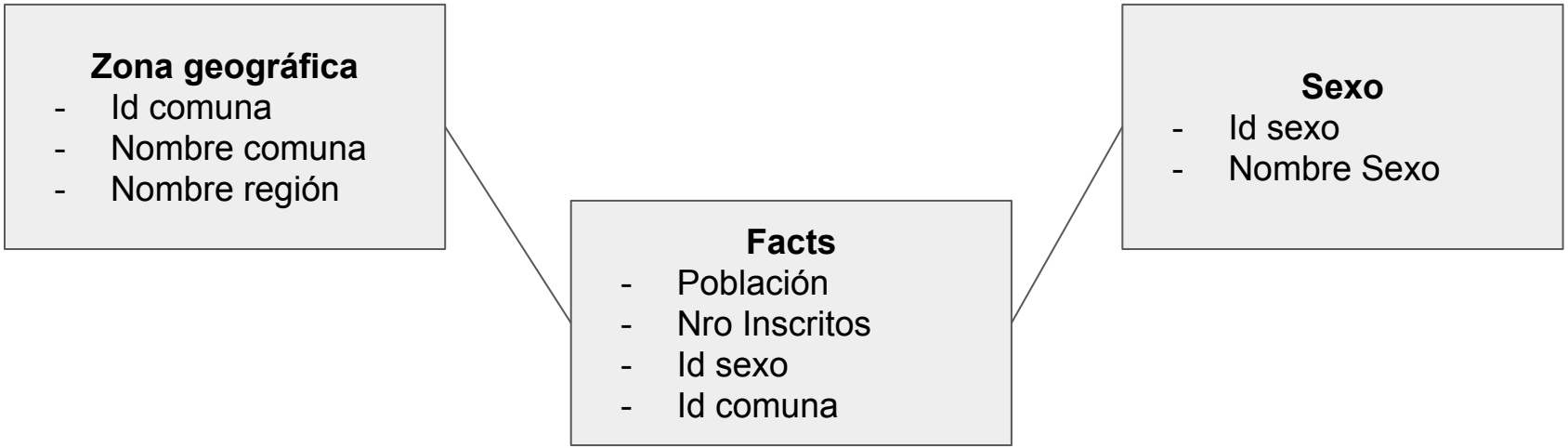


Figure 1-5: Fact and dimension tables in a dimensional model.

Ejercicio V2: Datamart de población beneficiaria de APS Universal

1. Revise los datos en el archivo DW_inscritos.xlsx.
   a. Están normalizados?
2. Construya un modelo estrella de una base de datos multidimensional para caracterizar por sexo y zona geográfica la población que se beneficiaría con la política de Atención Primaria Universal.
3. Construya los scripts para generar:
   a. Tabla fact que considere la población comunal, los beneficiarios Fonasa y los inscritos en APS
   b. Una dimensión con variables geográficas
   c. Una dimensión para sexo
   d. una dimensión para edad
4. Construya una base de datos multidimensional que contenga las métricas y dimensiones
5. ¿Qué tipos de análisis puede realizar con una tabla de este tipo?

Modelo estrella de Inscritos APS



Ver implementación de datamart en notebook de colab: Datamart Inscritos Fonasa