Escuela de Datos Vivos

Curso Data Engineer: Creando el pipeline de datos

MÓDULO A-2

Diseño

Diseño de bases relacionales

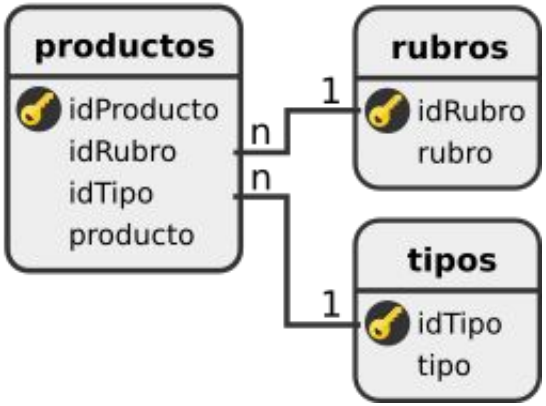
Online Transaction Processing (OLTP)

- 1 Forma Normal
- 2 Forma Normal
- 3 Forma Normal

Online Analytical Processing (OLAP)

Desnormalizado

Normalizado



Diseño de bases relacionales

- 1 Forma Normal
  - Cada columna en la tabla debe tener un valor atómico (simples e indivisibles)
  - Clave primaria no repetida
  - Todos los atributos son dependientes de la claveprimaria
  - No hay grupos repetidos en la tabla. Cada fila/columna

    contiene un solo valor, no un conjunto de ellos

Diseño de bases relacionales

- 1 Forma Normal

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

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

fuente: https://bookdown.org/

Diseño de bases relacionales

- 2 Forma Normal
  - Incluye la primera forma normal y crear una tabla separada para valores que aplican a múltiples filas y crear un link a través de claves foráneas*

*Una clave foránea es una o más columnas ordenadas que corresponde a una clave primaria en otra tabla

Diseño de bases relacionales

- 2 Forma Normal

Cliente

ID Cliente	Nombre	Apellido	Teléfono
123	Rachel	Ingram	555-861-2025
456	James	Wright	555-403-1659 555-776-4100
789	Cesar	Dure	555-808-9633

Cliente

ID Cliente	Nombre	Apellido	Teléfono 1	Teléfono 2	Teléfono 3
123	Rachel	Ingram	555-861-2025		
456	James	Wright	555-403-1659	555-776-4100	
789	Cesar	Dure	555-808-9633		

Cliente

ID Cliente	Nombre	Apellido
123	Rachel	Ingram
456	James	Wright
789	Cesar	Dure

Teléfono del cliente

ID Cliente	Teléfono
123	555-861-2025
456	555-403-1659
456	555-776-4100
789	555-808-9633

fuente: https://bookdown.org/

Diseño de bases relacionales

Online Transaction Processing (OLTP)

- 3 Forma Normal
  - Incluye la segunda forma normal y elimina cualquier columna de una tabla que no dependa de la clave

Diseño de bases relacionales

- 3 Forma Normal

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

CURSO

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

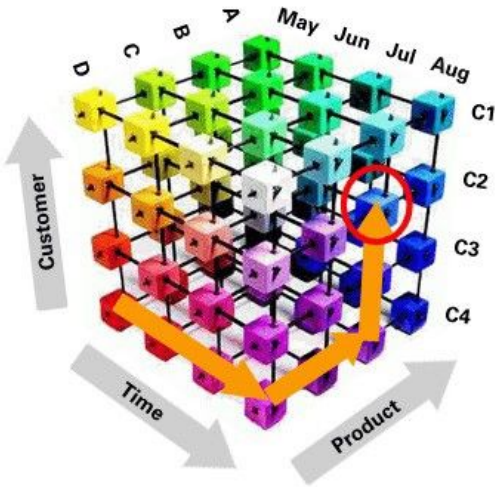
MATRICULA

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

Diseño de bases relacionales

Online Analytical Processing (OLAP)

- Desnormalizada
- Orientada a análisis de negocio
- Permite hacer drill down
- Permite analizar data en diferentes dimensiones
- Grandes volúmenes de data y registros



fuente: medium.datadriveninvestor.com

Diseño de bases no SQL

- Key-value
- Documentales
- Wide column
- Gráficas

Diseño de bases no SQL

- Key-value
  - Data es consultada rápidamente mediante keys
  - Data aleatoria en tiempo real
  - Aps. diseñadas consultas de Keys
  - Manejo de arreglos
  - Valores JSON

```
{
    name: "John",
      age : 35,
      dob : ISODate("01-05-1990"),
      profile_pic : "https://example.com/john.jpg",
      social : {
             twitter : "@mongojohn",
                 linkedin : "https://linkedin.com/abcd_mongojohn"
               }
}
```

Diseño de bases no SQL

- Documentales
  - No relacional, datos semi estructurados
  - Estructuras complejas
  - Información guardada en documentos jerárquicamente
  - Documentos: JSON, XML, texto
  - Documentos diseñados para agrupar información leída junta

mongoDB

Diseño de bases no SQL

- Wide column
  - Grandes volúmenes de data
  - Escritura de baja latencia
  - Más escrituras que lecturas
  - Limitada cantidad de consultas
  - Buscar por clave

Diseño de bases no SQL

- Gráficas
  - Basadas en modelado de entidades y relaciones
  - Nodos y links de manera gráfica
  - Las redes sociales son un claro ejemplo del uso de estas DBs
  - Navegar profundas jerarquías
  - Buscar conexiones ocultas entre ítems distantes
  - Descubrir interrelaciones entre ítems

Data Governance

Data Governance

- ¿Qué es Data Governance?

Es una colección de políticas, procesos, roles, estándares y métricas

- ¿Para qué se utiliza?

Para asegurar el uso eficiente de la información

- ¿Qué pasaría si no hay Data Governance?

Existiría incongruencia en la información, no sería confiable, no contariamos con data de calidad

Data Governance

- ¿Qué beneficios obtengo de Data Governance?
    - Data limpia, confiable y segura para realizar buenos análisis y obtener buenos resultados
    - Única fuente de verdad
- ¿Quienes son responsables de Data Governance?

Debemos estar todos involucrados y en mayor grado Chief data officer (CDO) y comité de DG

- Algunos ejemplos que podría ocurrir sin DG

Clientes con diferentes Ids, diferentes valores para un mismo campo (Argentina - Argentino)

Data Lineage

MARQUEZ

Amundsen

Apache Atlas

- La importancia del orden en los datos
- ¿Qué es Data Lineage?
- ¿Cómo nos puede ayudar en nuestras tareas diarias?
- Data en desuso o histórica donde almacenarla