

# Bases de Datos Heterogéneas

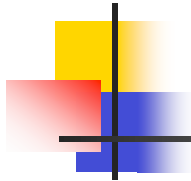
## CI6318



# Agenda

---

- Introducción al curso.
  - Conceptos Básicos.
  - Modelos de Datos
    - Modelo ER, ERE, Relacional.
  - DBMS.
    - Funcionalidades

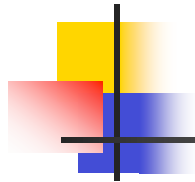


# Objetivos del Curso

---

- Fundamentos Lógicos de Bases de Datos:
  - Lógica de Predicados.
  - Complejidad del Problema de responder una consulta contra una BD.
- Problema de Evaluación y Optimización de Consultas:
  - Ambientes centralizados.
  - Ambientes Heterogéneos
    - Integración de Datos
      - Arquitecturas de Mediadores-Adaptadores
      - Técnicas de Reescrituras de Consultas

# Evaluaciones



- Exámen: 30%.
- Tareas: 30% 5-tareas, c/u 6 puntos.
- Proyecto: 40%
  - Diseño: 10%
  - Implementación: 20%
  - Resultados Experimentales: 10%

# Cronograma



## Semana 1:

- Introducción al curso.
- Introducción a Bases de datos- Conceptos Básicos.
- Introducción al problema de integración de datos.

## ■ Semana 2: MODELOS DE DATOS y LENGUAJES DE CONSULTAS

- Modelo Relacional, Álgebra Relacional y SQL
- Fundamentos Lógicos de las BD.
- Estructuras de Almacenamiento de Datos
- Tarea 1: Problemas sobre consultas en BD.

## ■ Semana 3: EVALUACIÓN DE CONSULTAS

- Evaluación de Consultas en Ambientes Centralizados.
  - Nested-Loop Join, Block Nested Loop Join, Index Nested Loop Join, Hash Join, etc.

## ■ Semana 4-5: TÉCNICAS DE OPTIMIZACIÓN

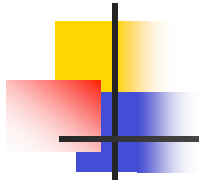
- Definición del proyecto.
- Optimización de Consultas en Ambientes Centralizados.
  - Optimización basada en heurísticas.
  - Modelos de Costos
- Tarea2: Optimización de Consultas basadas en heurísticas y Modelos de costos.

# Cronograma

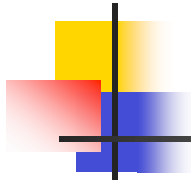


- **Semana 6: TÉCNICAS DE OPTIMIZACIÓN**
  - Optimización de Consultas en Ambientes Centralizados.
    - Optimización basada en costos.
      - Sistema R. Programación Dinámica.
      - Optimizadores Aleatorios.
- **Semana 7:**
  - **Proyecto**
- **Semana 8: TÉCNICAS DE OPTIMIZACIÓN**
  - Optimización de Consultas en Ambientes Centralizados.
    - Optimización basada en costos.
      - Sistema R. Programación Dinámica.
      - Optimizadores Aleatorios.
  - Tarea3: Optimización de Consultas en Ambientes centralizados.

# Cronograma



- **Semana 9: INTEGRACIÓN DE DATOS**
  - Arquitectura de Mediadores y Adaptadores (Mediators and Wrappers).
  - Integración de Datos
  - Tarea 4: Sistemas de Integración.
- **Semana 10: INTEGRACIÓN DE DATOS**
  - Integración de Datos (LAV)
  - Tarea 4: Sistemas de Integración.
- **Semana 11 INTEGRACIÓN DE DATOS Y OPTIMIZACIÓN DE CONSULTAS:**
  - Integración de Datos (LAV)
  - Optimización de Consultas en Ambientes Heterogéneos
    - Tarea 5: optimización de consultas en ambientes heterogéneos.
- **Semana 12:**
  - Examen.
- **Semana 13:**
  - Exposiciones.

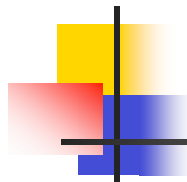


## Referencias

---

- **Database Management Systems,**  
Ramakrishnan and Gehrke.
- Foundations of Databases  
Abiteboul, Hull, Vianu. Addison-Wesley.
- Artículos





# Qué es una Base de Datos?

---

**Universo de Discurso:** porción del mundo real a ser modelado.

**Bases de Datos:** colección de datos relacionados.

**Modelo de Datos:** herramienta de especificación de bases de datos. Los datos son definidos en función de:

- Propiedades estructurales, dinámicas y de comportamiento.

**Esquema Conceptual:** representación de un situación haciendo uso de un modelo de datos particular.



# Modelos de Datos

---

- Fuertemente Tipeados versus Debilmente Tipeados.
  - Relacional, ER versus Datalog
- Semánticos versus poco semánticos
  - Orientados por Objetos (UML), Ontologías, ERE versus Relacional.



## Restricciones de Integridad

---

- **Inherentes**: expresiones que representan las propiedades de las estructuras ofrecidas por el modelo de datos.
- **Implícitas**: expresiones que representan las propiedades especificadas explícitamente a través del modelo.
- **Explícitas**: propiedades que no pueden ser modelados directamente a través del modelo. Se requiere de un lenguaje lógico para representar estas propiedades.

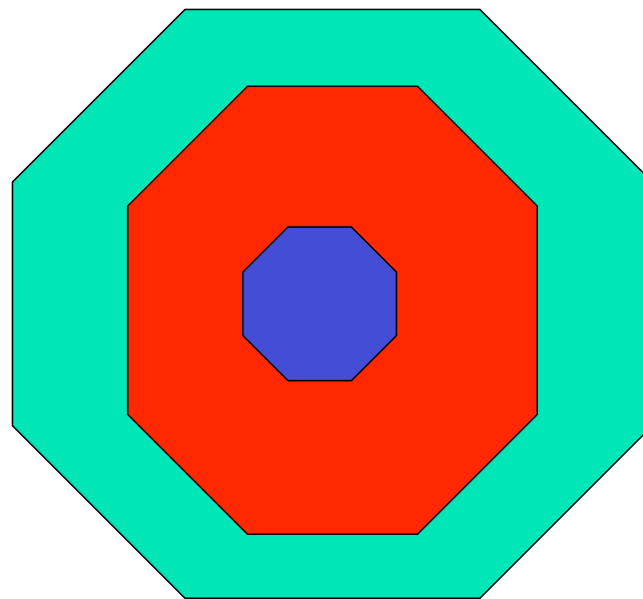
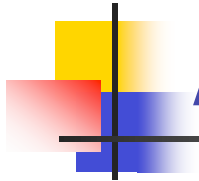


# Lógica de Primer Orden- Esquema de Datos

---

- Un **esquema de datos** puede ser formalizado como un **conjunto de restricciones** que deben ser respetados.
- Desde el punto de vista **formal**, un **esquema de datos** corresponde a un conjunto de **fórmulas de la Lógica de Primer Orden**. Ese conjunto de fórmulas se denominan **teoría**.
- La **interpretación** de un **esquema de datos** es definida como una **colección** de todas las **estructuras** que **respetan** el conjunto de **restricciones impuestas**.
- Las **interpretaciones legales** de un **esquema de datos** son todos los **modelos de la teoría** de la Lógica de Primer Orden.

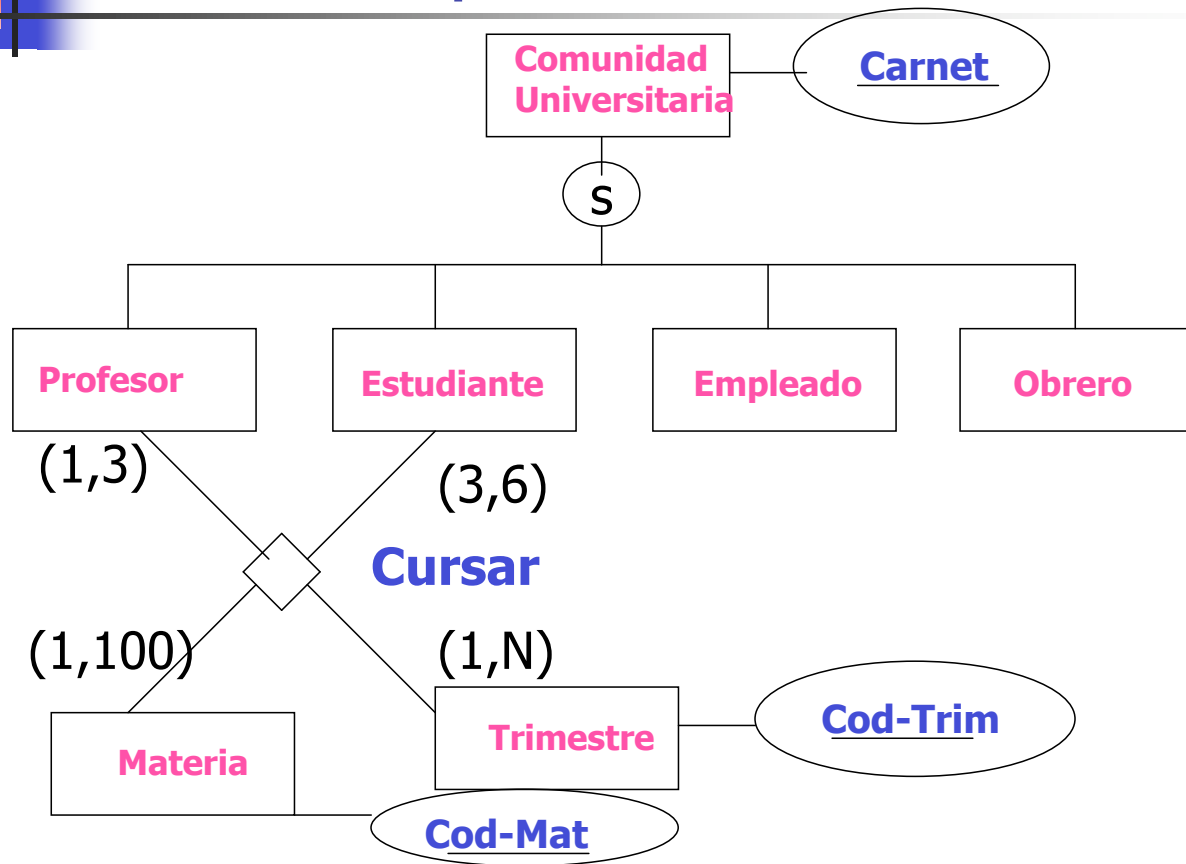
# Esquema de Datos-Sistema Axiomático



-  Teoría
-  Teoremas
-  Axiomas

Un **esquema de datos** o **teoría** se compone de un conjunto de **axiomas** y **reglas de inferencia** que caracterizan a un conjunto de **teoremas**.

# Ejemplo de una axiomatización de un Esquema ERE

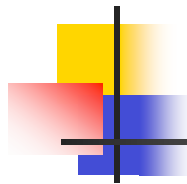




# Conocimiento Representado en el esquema

---

- **Comunidad Universitaria** conjunto de elementos que representan a la comunidad universitaria.
- **Profesor** conjunto de elementos de la comunidad universitaria que representan a los profesores.
- **Estudiante** conjunto de elementos de la comunidad universitaria que representan a los estudiantes.
- **Empleado** conjunto de elementos de la comunidad universitaria que representan a los empleados administrativos.
- **Obrero** conjunto de elementos de la comunidad universitaria que representan a los obreros.
- **Materia** conjunto de elementos que representan a las materias.
- **Trimestre** conjunto de elementos que representan a los trimestres.



# Restricciones Inherentes

- Si A es un conjunto:
  - No existen repeticiones
  - El orden es irrelevante.
- En ERE los elementos de un conjunto se diferencian por su clave:
  - $\forall x( x \in A \text{ and clave}(B,A))$   
 $\Rightarrow \text{not } (\exists x1( x1 \in A \text{ and } x1 \neq x \text{ and } x.B = x1.B))$

## Cursar:

- es una relación que representa a la asociación entre **Profesor** **Estudiante** **Materia** **Trimestre**
- es un conjunto.
- **Cursar**  $\subseteq$  **Profesor** **X** **Estudiante** **X** **Materia** **X** **Trimestre**





# Restricciones Inherentes- Axiomas

---

- $\forall x( x \in \text{ComunidadUniversitaria} \Rightarrow \text{not } (\exists x1( x1 \in \text{ComunidadUniversitaria} \text{ and } x1 \neq x \text{ and } x.\text{Carnet}=x1.\text{Carnet})))$
- $\forall x( x \in \text{Profesor} \Rightarrow \text{not } (\exists x1( x1 \in \text{Profesor} \text{ and } x1 \neq x \text{ and } x.\text{Carnet}=x1.\text{Carnet})))$
- $\forall x( x \in \text{Estudiante} \Rightarrow \text{not } (\exists x1( x1 \in \text{Estudiante} \text{ and } x1 \neq x \text{ and } x.\text{Carnet}=x1.\text{Carnet})))$
- $\forall x( x \in \text{Empleado} \Rightarrow \text{not } (\exists x1( x1 \in \text{Empleado} \text{ and } x1 \neq x \text{ and } x.\text{Carnet}=x1.\text{Carnet})))$
- $\forall x( x \in \text{Obrero} \Rightarrow \text{not } (\exists x1( x1 \in \text{Obrero} \text{ and } x1 \neq x \text{ and } x.\text{Carnet}=x1.\text{Carnet})))$



# Restricciones Inherentes

---

- **Comunidad Universitaria** se especializa de forma solapada en:
  - **Profesor**
  - **Estudiante**
  - **Obrero**
  - **Empleado**
- **Profesor  $\cup$  Estudiante  $\cup$  Obrero  $\cup$  Empleado**  
 **$\subseteq$  Comunidad Universitaria**



# Restricciones Implícitas

Cardinalidades en **Cursar**:

- **Card(Profesor, Cursar)=(1,3)**
  - $\forall x (x \in \text{Profesor} \Rightarrow 1 \leq |\{(x,y,z,j) / (x,y,z,j) \in \text{Cursar}\}| \leq 3)$
- **Card(Estudiante, Cursar)=(3,6)**
  - $\forall x (x \in \text{Estudiante} \Rightarrow 3 \leq |\{(y,x,z,j) / (y,x,z,j) \in \text{Cursar}\}| \leq 6)$
- **Card(Materia, Cursar)=(1,100)**
  - $\forall x (x \in \text{Materia} \Rightarrow 1 \leq |\{(y,z,x,j) / (y,z,x,j) \in \text{Cursar}\}| \leq 100)$
- **Card(Trimestre, Cursar)=(1,N)**
  - $\forall x (x \in \text{Trimestre} \Rightarrow 1 \leq |\{(y,z,j,x) / (y,z,j,x) \in \text{Cursar}\}|)$



# Restricciones Implícitas- Axiomas

- **Cursar**  $\subseteq$  **Profesor**  $\times$  **Estudiante**  $\times$  **Materia**  $\times$  **Trimestre**  
 $\forall x (x \in \text{Profesor} \Rightarrow 1 \leq |\{(x,y,z,j) / (x,y,z,j) \in \text{Cursar}\}| \leq 3)$
- $\forall x (x \in \text{Estudiante} \Rightarrow 3 \leq |\{(y,x,z,j) / (y,x,z,j) \in \text{Cursar}\}| \leq 6)$
- $\forall x (x \in \text{Materia} \Rightarrow 1 \leq |\{(y,z,x,j) / (y,z,x,j) \in \text{Cursar}\}| \leq 100)$
- $\forall x (x \in \text{Trimestre} \Rightarrow 1 \leq |\{(y,z,j,x) / (y,z,j,x) \in \text{Cursar}\}|)$
- **Profesor**  $\cup$  **Estudiante**  $\cup$  **Obrero**  $\cup$  **Empleado**  
     $\subseteq$  **Comunidad Universitaria**  
  
| **Comunidad Universitaria** | =  
| **Profesor** | + | **Estudiante** | + | **Materia** | + | **Trimestre** | -  
( | **Profesor**  $\cup$  **Estudiante** | + | **Profesor**  $\cup$  **Empleado** | +  
| **Profesor**  $\cup$  **Obrero** | + | **Estudiante**  $\cup$  **Empleado** | +  
| **Estudiante**  $\cup$  **Obrero** | + | **Empleado**  $\cup$  **Obrero** | )



# Instancias de un Esquema de Datos-Informalmente

---

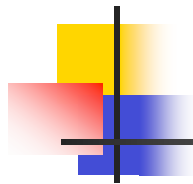
- Bases de Datos:
  - Consistentes.
  - No consistentes.
  - Transacciones.
- Qué significa contestar una consulta?
  - Complejidad.



# Modelo Relacional

---

- Estructura básica:
  - Relación:
    - Se define sobre un conjunto de dominios  $D_1, \dots, D_n$ .
    - Corresponde a un subconjunto sobre el producto cartesiano de  $D_1, \dots, D_n$
  - Propiedades:
    - No existen tuplas repetidas.
    - El orden de los elementos es irrelevante.
    - Si las columnas se nombran, en orden de las columnas es irrelevante.
  - Restricciones:
    - Integridad Referencial
    - Clave Primaria



## Modelo Relacional-Terminología

---

- Sea  $R$  una relación con los atributos  $A_1, \dots, A_n$  sobre los dominios  $D_1, \dots, D_n$ , respectivamente, entonces
$$R \subseteq D_1 \times \dots \times D_n$$
  - No hay tuplas repetidas.
  - El orden de las tuplas es irrelevante.
  - El orden de las columnas es irrelevante.

# Terminología

Nombres de Atributos

Producto(Nombre Relación)

	Nombre	Precio	Categoría	Fabricante
	Carton	1000	chocolate	Savoy
	Toronto	6000	chocolate	Savoy
	CafeMocha	3000	alimento	Fama de America

tuplas

(Aridad=4)

Producto(Nombre: string, Precio: real, Categoría: enum, Fabricante: string)






# Interrogando a una Base de Datos

```
SELECT P.Fabricante  
FROM Producto P  
WHERE P.Precio>1000
```

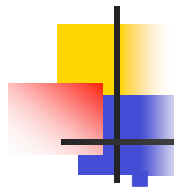
- SQL (Structured Query Language)
  - Basado en Álgebra y Cálculo Relacional.
- Datalog.



# Modelo Relacional-Lenguajes de Interrogación

---

- Álgebra Relacional:
  - Basado en teoría de conjuntos.
  - Cerrada.
- Cálculo Relacional
  - Basado en Lógica de Primer Orden.



# Álgebra Relacional

---

## ■ Álgebra cerrada:

- El resultado de la aplicación de cualquier operador del álgebra relacional a una o más relaciones es también una relación.

## ■ Operadores Básicos:

- Selección
- Proyección
- Producto Cartesiano
- Unión
- Intersección

## ■ Operadores No Básicos:

- Join: Theta, Natural.



# Por qué usar un DBMS?

---

Si cualquier programa puede manejar datos?

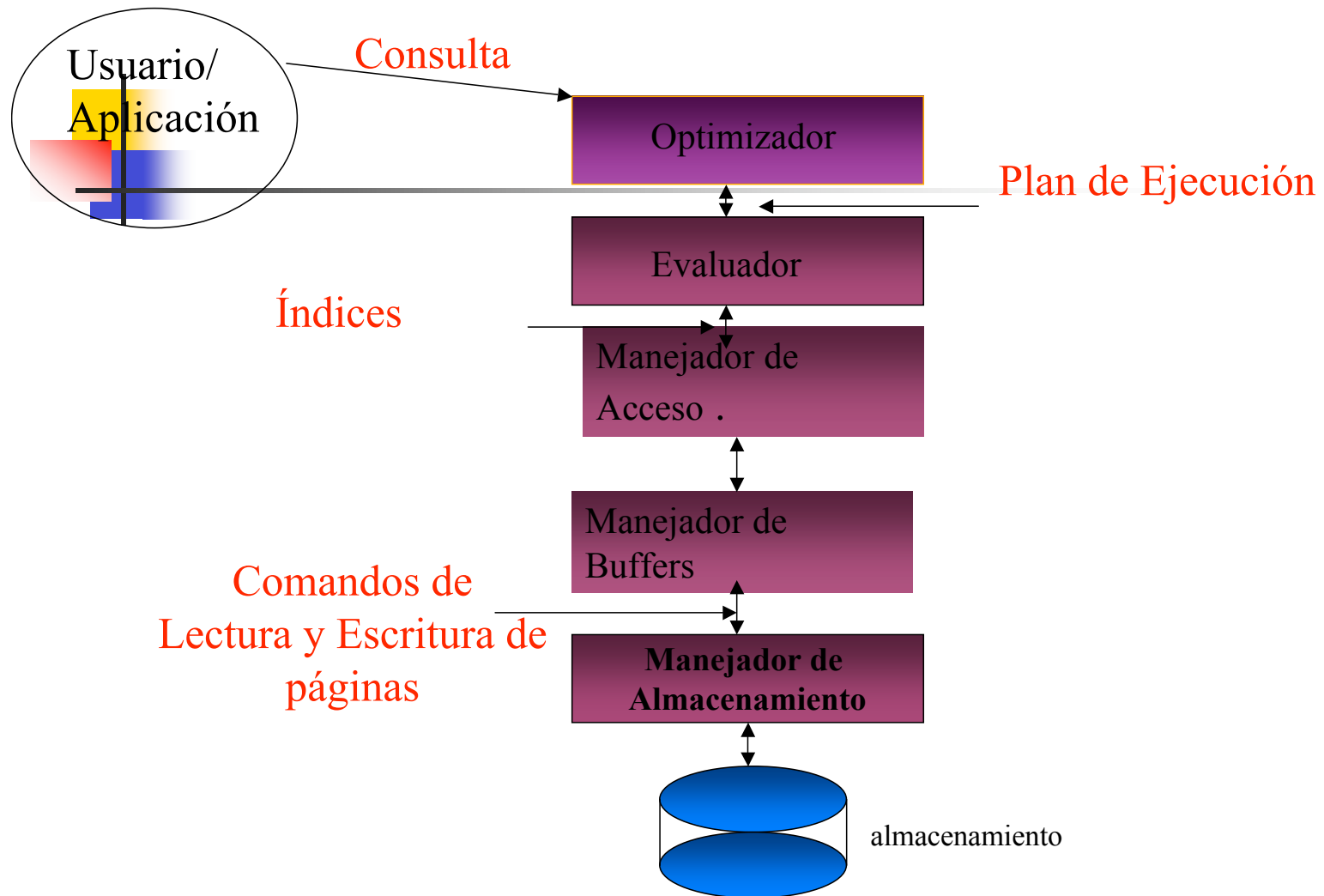
- Grandes volúmenes de datos(Giga's, Tera's)
- Los datos son estructurados
- Los datos deben persistir
- Los datos son valiosos
- Los datos deben ser accedidos eficientemente
- Los datos deben ser accedidos concurrentemente
- Los datos deben ser accedidos sólo por personas autorizadas
- Los datos deben ser manipulados por transacciones que llevan la BD de un estado consistente a otro estado consistente.

# Funcionalidad de un DBMS



---

- Manejo de memoria persistente
- Manejo de transacciones
- Manejo de recuperación
- Manejo de integridad
- Separación entre la visión lógica y física de los datos.
  - Lenguajes de interrogación de alto nivel.
  - Procesamiento de consultas eficiente
- Interfaces con lenguajes de programación



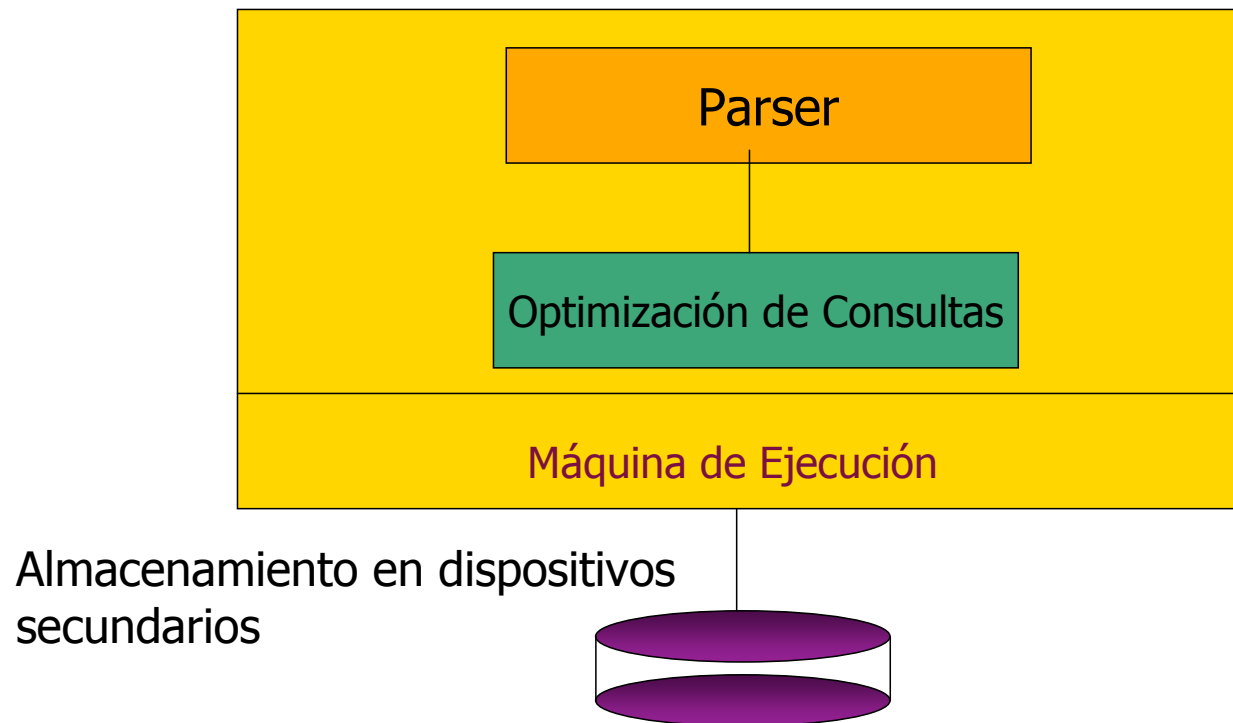


# Optimización y Evaluación

---

- Optimización: proceso mediante el cual se identifica una estrategia eficiente de ejecutar una consulta.
- Evaluación: mecanismo mediante el cual una consulta es ejecutada.

# Arquitectura





# Optimización de Consultas

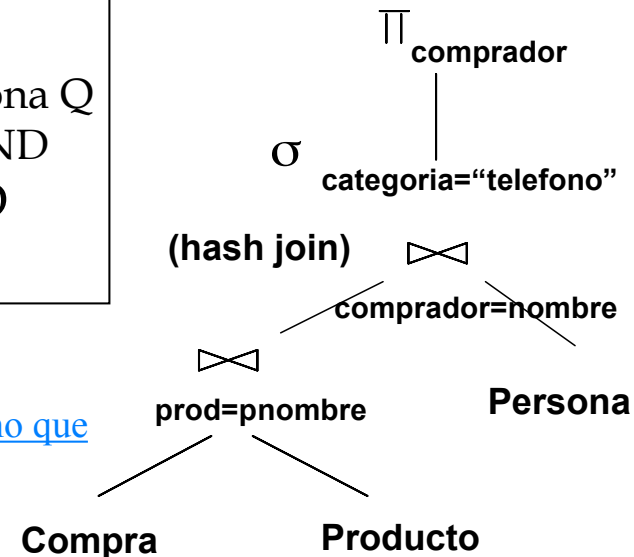
**Objetivo:**

*Consulta en SQL*

```
SELECT C.comprador
FROM Producto P, Compra C, Persona Q
WHERE P.Categoria="Telefono" AND
C.comprador=Q.nombre AND
C.prod=P.pnombre
```

Plan: Árbol de operadores del álgebra relacional  
donde cada operador está anotado con el algoritmo que  
lo implementa

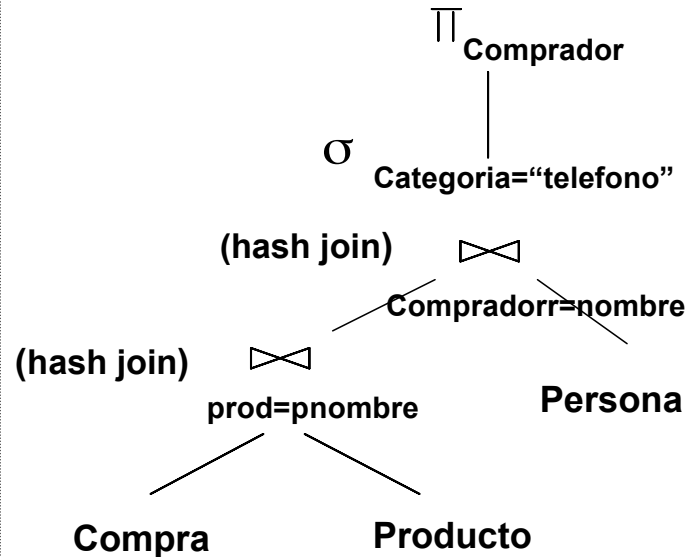
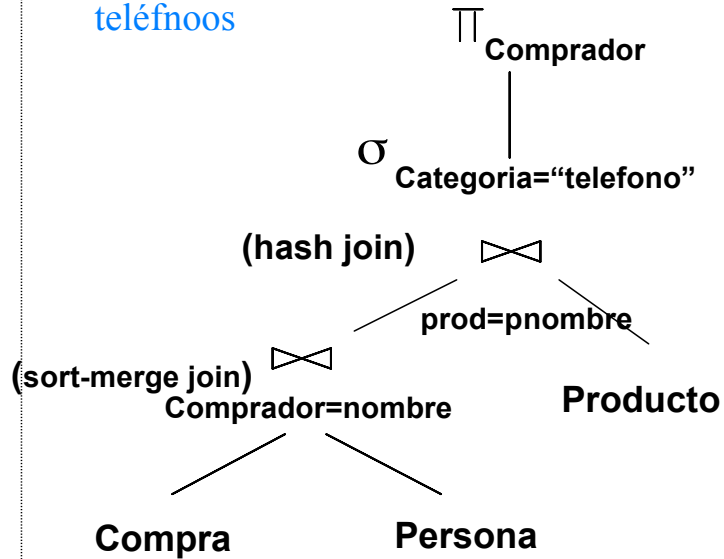
*Plan de Ejecución:*



**Idealmente:** se desea conseguir el mejor. **En la práctica:** se evitan los peores

# Planes de Ejecución

Dar los nombres de las personas que han comprado un producto en la categoría de teléfonos



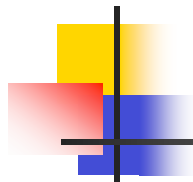
Existen muchas maneras de evaluar una consulta de SQL.



## Bases de Datos en la Industria

---

- Las Bases de Datos Relacionales han sido usadas exitosamente.
- Oracle tiene un mercado de aprox \$200B
- Otras compañías: IBM, MS, Sybase, Informix
- Tendencias:
  - warehousing y soporte de decisiones
  - Integración de datos
  - XML y sus dialectos

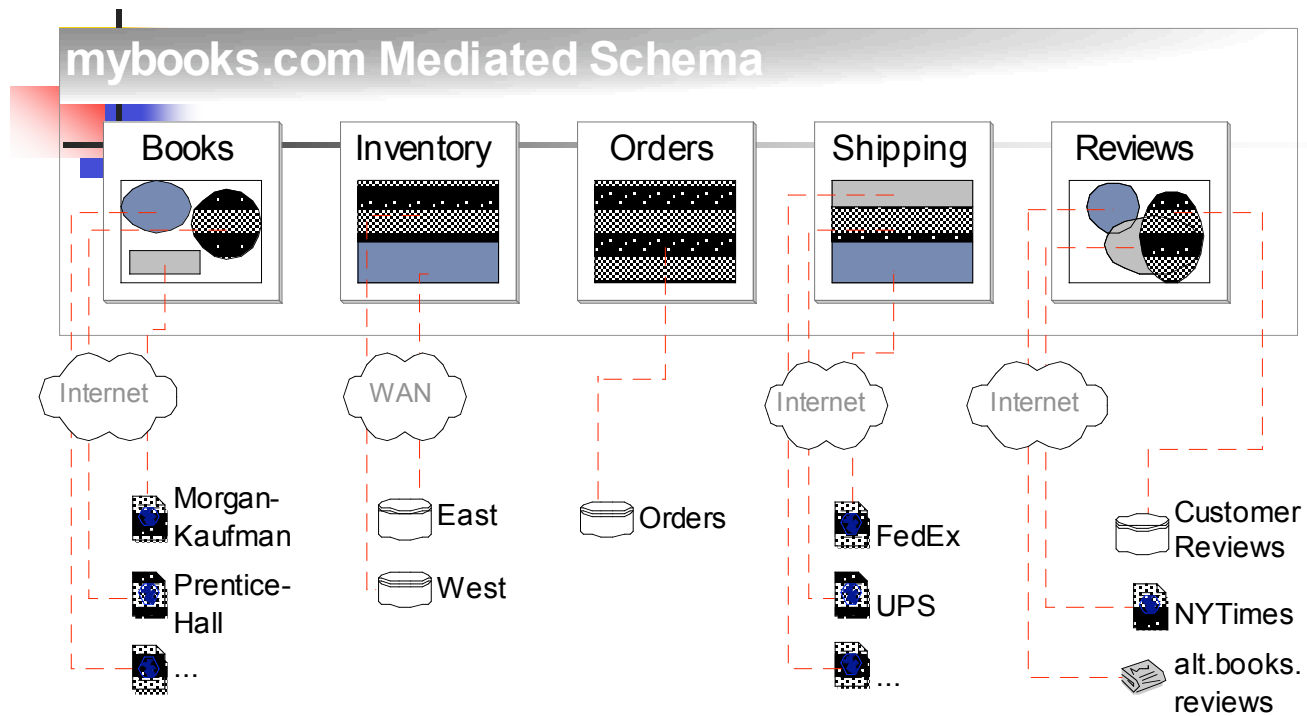


## Integración de Datos

---

- Es el problema de proveer:
  - Acceso (consultar y eventualmente actualizar)
  - uniforme (transparente a los usuarios)
  - a multiples (aún 2 es un problema!)
  - fuentes de datos (no únicamente bases de datos)
  - autónomas (no afecta la conducta de las fuentes de datos)
  - heterogéneos (diferentes modelos de datos y esquemas)
  - estructurados (al menos semi-estructurados)

# Integración de Datos



Acceso transparente a múltiples fuentes de datos heterogéneas

## Motivación

- Integración de los datos de una organización.

- World-wide-web:

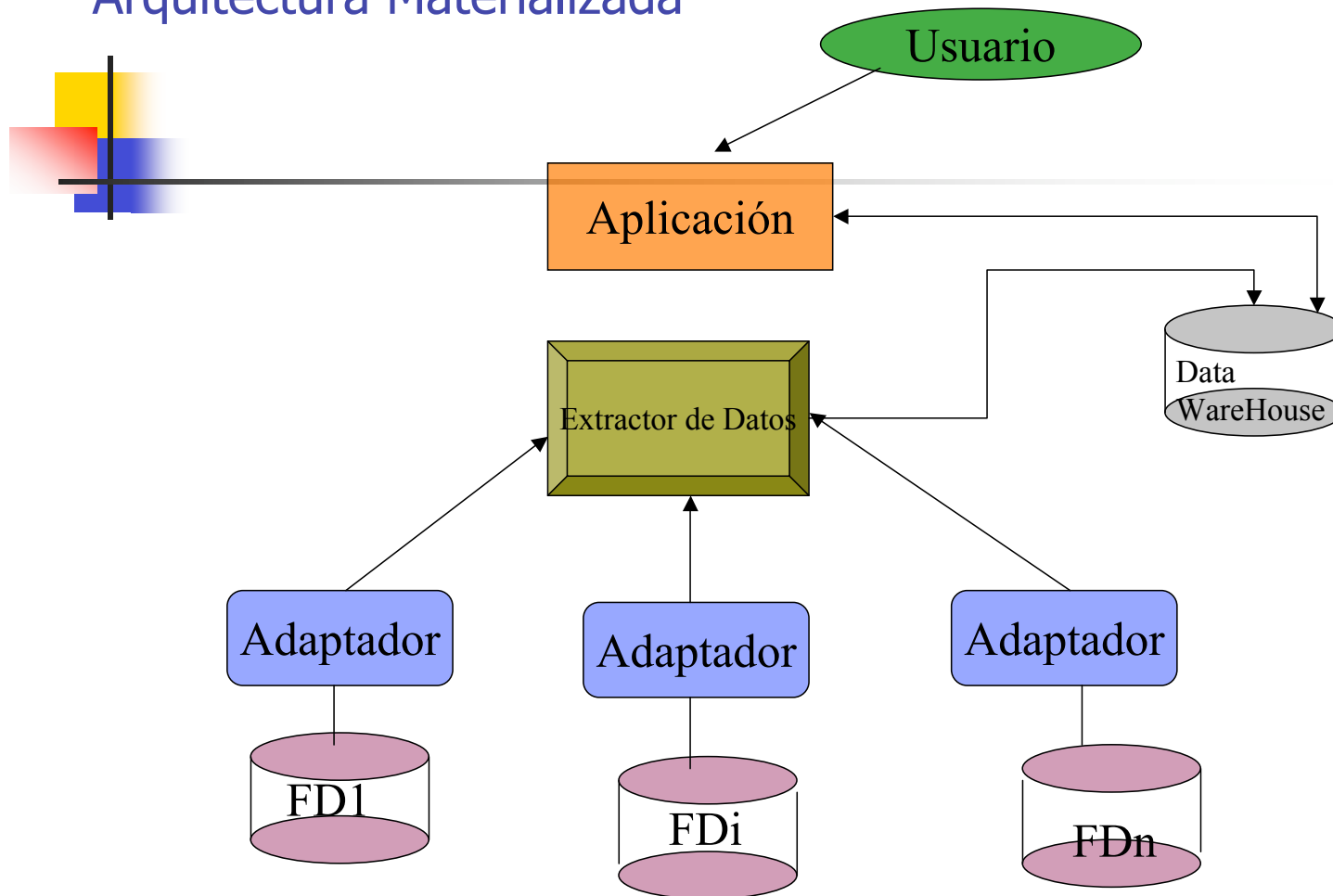
- - Diferentes fuentes de ventas
  - Portales que integran datos desde múltiples fuentes de datos
  - Integración de XML
- Ciencia y cultura:
  - Fuentes de datos con datos Biomoleculares.
  - Fuentes de datos astrofísicas: monitoreando los eventos del espacio.
  - Fuentes de datos Ambientales.
  - Culturales: acceso uniforme a las bases de datos culturales.

## Aspectos a Considerar



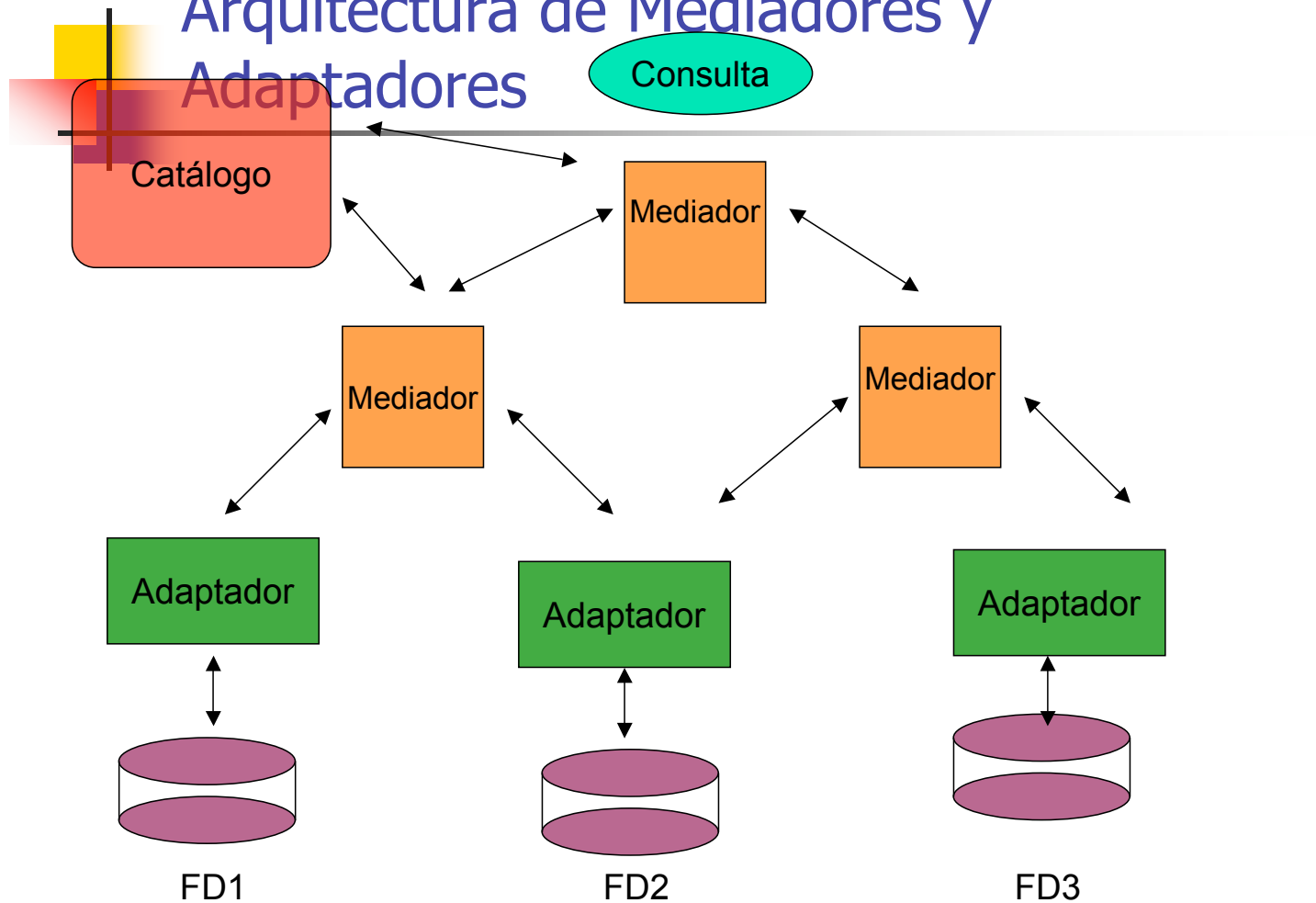
- Arquitectura virtual vs. Arquitectura materializada.
- Acceso: a consultas o a consultas y actualizaciones.
  - Problema similar a la actualización a través de vistas.
  - Necesidad del manejo de transacciones distribuidas.
- Esquemas mediadores:
  - Esquemas mediadores requieren integración del esquema y entonces reformalización de la consulta.
  - Sin esquemas mediadores se pierde algunas de las ventajas de la integración de datos.

## Arquitectura Materializada






# Arquitectura de Mediadores y Adaptadores



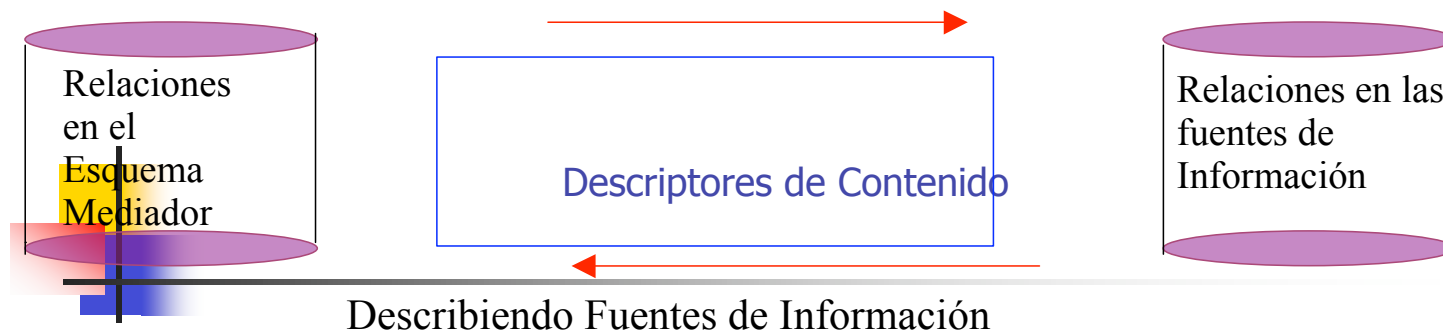
## Tópicos a Considerar:

- 
- Funcionalidad de los Mediadores y Wrappers.
  - Conflictos estructurales y conflictos semánticos.
    - Correspondencia de esquemas y reformulación.
  - Descripción de las fuentes de datos:
    - Modelamiento de la completitud de los datos.
    - Modelamiento de las capacidades de procesamiento de las fuentes de datos.
  - Selección de Fuentes de Datos.
  - Optimización de consultas.
  - Ejecución de consultas.



# Catálogos de un Sistema Mediator

- Descripción del contenido de las fuentes de datos.
- Capacidad de las fuentes de datos.
- Completitud del contenido de las fuentes de datos.
- Propiedades físicas de las fuentes de datos y de la red.
- Estadísticas sobre los datos.
- Equivalencias entre fuentes de datos.



- Consultas de los usuarios referencian a relaciones en el esquema del mediador.
- Fuentes almacenan datos en sus propios esquemas locales.
- Los descriptores de contenido definen una correspondencia entre los esquemas mediadores y locales.

## Algunos Prototipos

- 
- DISCO(INRIA)
  - Garlic (IBM)
  - HERMES/WebSrcMed (U. Maryland)
  - InfoMaster(Stanford)
  - Information Manifold (AT & T)
  - IRO-DB (Versailles)
  - SIMS, ARIADNE (USC/ISI)
  - The Internet Softbot/ Occam/ Razor/ Tukwila (UW)
  - TSIMMIS (Stanford), XMAS (UCSD)
  - WHIRL (AT &T)



## Aspectos a considerar

---

- Definición de las fuentes en función del esquema integrado.
- Algoritmos para reescribir la consulta en el esquema integrado en función de las fuentes de datos.