

Bases de datos distribuidas

Dr. José Luis Zechinelli Martini

jose Luis.zechinelli@udlap.mx

LIS – 3071

*Administración de datos y de conocimiento
LAFMIA – UDLAP*

Plan

- De las bases de datos centralizadas a las distribuidas
- Integración de datos
- Bases de datos distribuidas (BDD):
 - Arquitecturas
 - Funcionamiento general
 - Técnicas de diseño
- Sistemas de acceso a información modernos

¿Por qué BD distribuidas? (1)

- **Bases de datos organizacionales y económicas:**
 - Muchas organizaciones se **descentralizan** y un modelo distribuido se adapta mejor a la estructura de la organización
- **Interconexión de bases de datos existentes:**
 - Solución natural cuando existen diferentes bases de datos en una organización y se quiere ofrecer un **acceso global** a esas bases
- **Crecimiento incremental:**
 - Solución natural cuando la organización crece con **nuevas unidades** que desarrollan sus propias bases de datos
 - En una solución centralizada habría que modificar el esquema, en una solución distribuida el crecimiento es más o menos **transparente**

¿Por qué BD distribuidas? (2)

- **Desempeño y costo de comunicación:**
 - Fomentar el **acceso multi-punto** a los datos sin tener que pasar por un sólo servidor (cuello de botella)
 - **Acercar los datos a los clientes** para evitar comunicaciones remotas
- **Fiabilidad y disponibilidad:**
 - Posibilidad de tener **datos redundantes** para asegurar la disponibilidad a pesar de posibles fallas de comunicación y de los servidores
 - Posibilidad de distribuir la **carga** de diferentes servidores de datos para asegurar mejor calidad de servicio

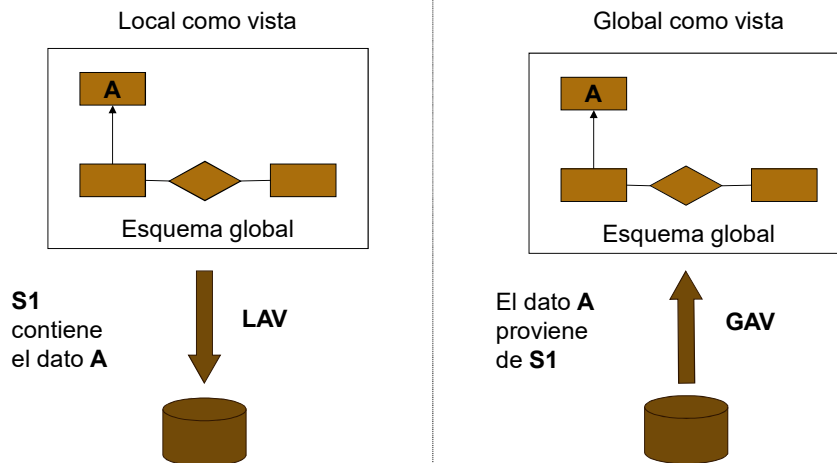
Plan

- ✓ De las bases de datos centralizadas a las distribuidas
- Integración de datos
- Bases de datos distribuidas (BDD):
 - Arquitecturas
 - Funcionamiento general
 - Técnicas de diseño
- Sistemas de acceso a información modernos

Integración de datos y esquemas

- Esquema global con vistas integradas
- Esquema global sin vistas integradas
- Datos a partir de datos estructurados, semi-estructurados y no estructurados
- Datos materializados en el mediador (*Data Warehouse*, Caché)
- Modificaciones (administración de transacciones)

Local as View vs. Global as View



Global como vista

- *Global as View* (e.g., TSIMMIS, SIMS, Garlic):
 - La calidad global del sistema depende de cómo las fuentes se integran para construir el esquema global
 - La integración de nuevas fuentes causa la modificación del esquema global
 - Se reduce la complejidad de la fase de reescritura

Local como vista

- *Local as View (e.g., Information Manifold):*
 - La calidad global del sistema depende de cómo se caracterizan las fuentes
 - Un esquema global, bien especificado a priori, no debe ser modificado cuando las fuentes son actualizadas o nuevamente integradas
 - Se incrementa la complejidad de la fase de reescritura


Plan

- ✓ De las bases de datos centralizadas a las distribuidas
- ✓ Integración de datos
- Bases de datos distribuidas (BDD):
 - Arquitecturas
 - Funcionamiento general
 - Técnicas de diseño
- Sistemas de acceso a información modernos

BDD: Aspectos a considerar (1)

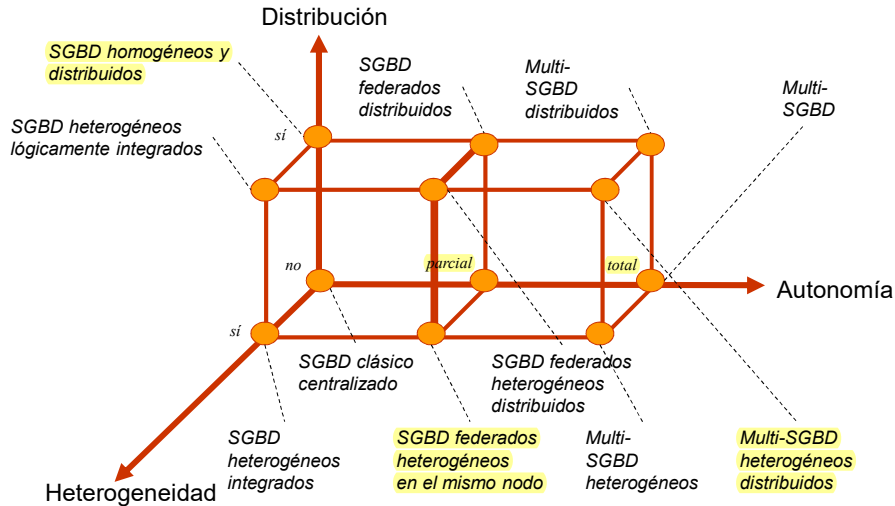
- El tipo de red
- El nivel de heterogeneidad de datos
- La distribución controlada de los datos
- El nivel de heterogeneidad de los sistemas existentes
- Esconder o no la distribución a los usuarios:
 - Transparencia de localización
 - Transparencia de fragmentación
 - Transparencia de duplicación

BDD: Aspectos a considerar (2)

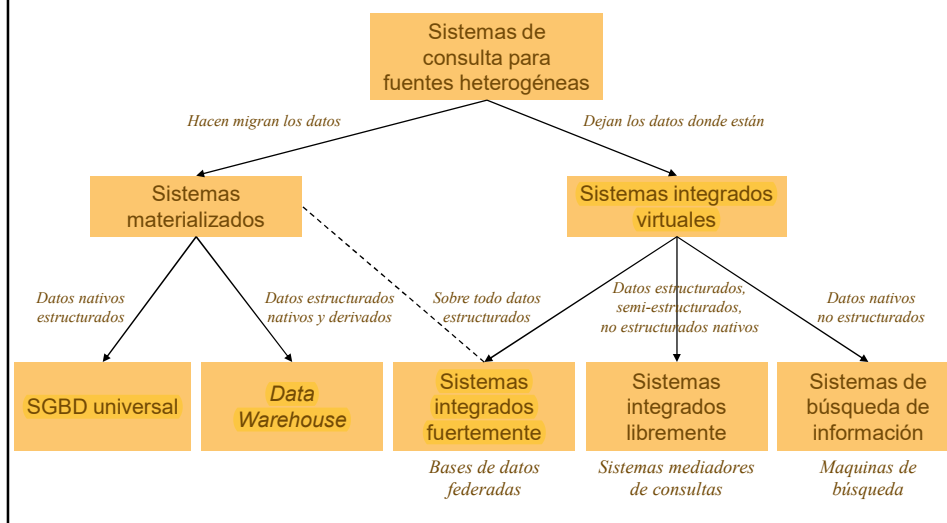
- La transparencia impacta:
 - La consulta distribuida: expresión, optimización, formateo de resultados
 - Las transacciones distribuidas: atomicidad, aislamiento, durabilidad 
 - La autorización



Modelos de arquitecturas



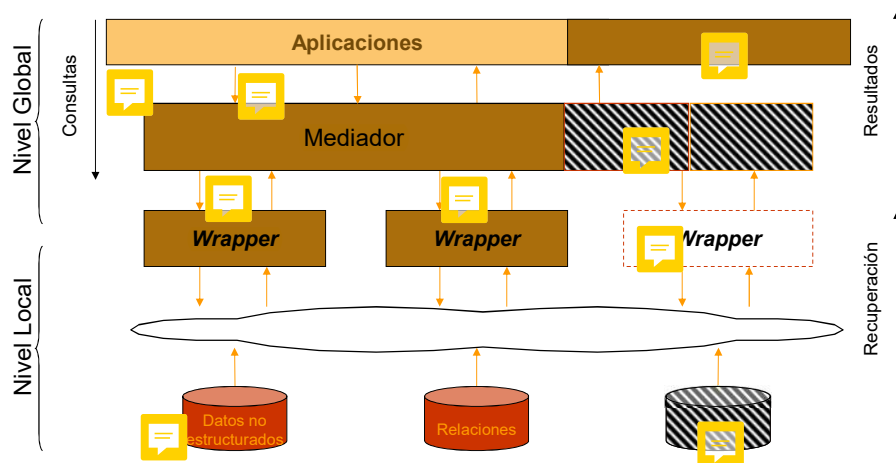
Espectro de sistemas distribuidos



Plan

- ✓ De las bases de datos centralizadas a las distribuidas
- ✓ Integración de datos
- Bases de datos distribuidas (BDD):
 - ✓ Arquitecturas
 - Funcionamiento general
 - Técnicas de diseño
- Sistemas de acceso a información modernos

Sistemas Multi-SGBD



Ejemplo

- R1: artículos
 - A(NA, DESCR, PV, PR, P)
 - El artículo de número NA, descrito por la descripción DESCR cuyo precio de venta PV, precio de recuperación PR y es vendido por el proveedor P
- R2: tiendas
 - T(NT, CIUDAD, DIRM)
 - La tienda de número NT se localiza en la ciudad CIUDAD y tiene de director al empleado de número DIRM
- R3: ventas mensuales
 - V(NT, NA, MES, CANT)
 - La tienda de número NT vendió una cantidad CANT de artículos de tipo NA durante el mes MES
- R4: empleados
 - E(NE, NOM, EDAD, FECHAEMPLO, SAL, NT)
 - El empleado con número NE, de nombre NOM, de edad EDAD fue contratado en la fecha FECHAEMPLO, gana un salario SAL en la tienda NT
- R5: histórico de empleados
 - HE(NE, FECHAEMPLO, FIN, SAL)
 - El empleado identificado por el número NE, contratado en la fecha FECHAEMPLO ganaba un salario SAL en la fecha FIN
- R6: proveedores
 - F(NF, CIUDAD)
 - El proveedor identificado por el número NF se encuentra en la ciudad CIUDAD

Integración débil: Consulta

- Las ventas mensuales de cada tienda organizadas por ciudad:

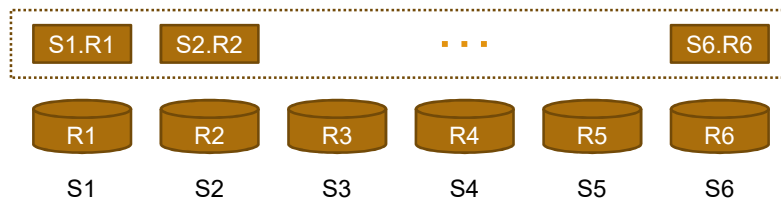
```

SELECT      T.NT, V.Mes, T.Ciudad,
            SUM( V.CANT ) Ventas

FROM        T, V

WHERE       T.NT = V.NT

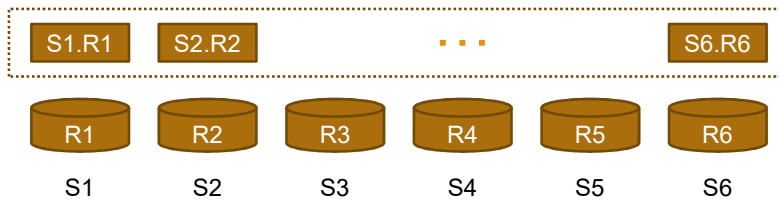
GROUP BY    T.NT, V.Mes, T.Ciudad
  
```



Integración débil: Integración (1)

- Las ventas mensuales de cada tienda organizadas por ciudad:

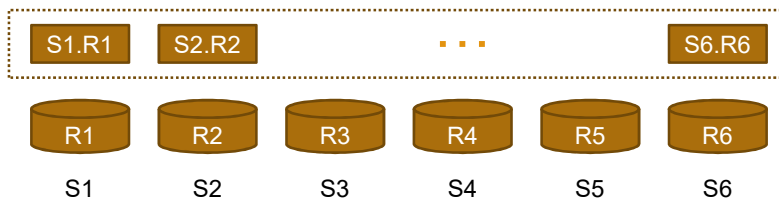
```
SELECT      S2.R2.NT, S3.R3.Mes, S2.R2.Ciudad,
            SUM( S3.R3.CANT ) Ventas
FROM        S2.R2, S3.R3
WHERE       S2.R2.NT = S3.R3.NT
GROUP BY    S2.R2.NT, S3.R3.Mes, S2.R2.Ciudad
```



Integración débil: Integración (2)

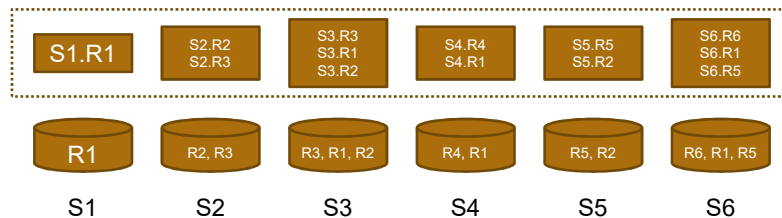
- Las ventas mensuales de cada tienda organizadas por ciudad:

```
SELECT      struct( N: NT, M: Mes, C: Ciudad,
                    V: (select p.v from partition p) )
FROM        S2.R2 t, S3.R3 v
WHERE       t.NT = v.NT
GROUP BY    t.NT, t.Mes, t.Ciudad
```



Integración débil: Duplicación

- Las ventas mensuales de cada tienda organizadas por ciudad
- Los artículos que se pueden vender sin duplicar datos en la respuesta
- Los artículos que vende cada tienda organizados por ciudad
- La historia de los salarios de los directores de las tiendas organizados por ciudad



Plan

- ✓ De las bases de datos centralizadas a las distribuidas
- ✓ Integración de datos
- Bases de datos distribuidas (BDD):
 - ✓ Arquitecturas
 - ✓ Funcionamiento general
 - Técnicas de diseño
- Sistemas de acceso a información modernos

Distribución de datos

- Sea R una relación de una BDR (global) y R1, R2 fragmentos locales:
 - **Centralización:** $R = R1$ en el sitio S1
 - **Duplicación total:**
 - $R = R1$ en S1 igual $R2$ en S2
 - $R = \text{escoger}(R1, R2)$
 - **Fotografía:**
 - $R = R1$ en S1 (actualizada)
 - $R2$ en S2 (no actualizada)
 - **Fragmentación:** Horizontal / Vertical

Fragmentación de los datos

- **Fragmentación horizontal:** un subconjunto de las tuplas de la relación (condición sobre uno o mas atributos)
- **Fragmentación vertical:** un subconjunto de los atributos de la relación (no se requieren todos los atributos)
- **Fragmentación híbrida:** fragmentación horizontal y vertical
- Ejemplo:
 - DEPARTMENT(DNO, DNAME, MGRSSN)
 - PROJECT(PNO, PNAME, PLOC, DNO)
 - EMPLOYEE(SSN, NAME, SEX, BDATE, ADDRESS,
SALARY, MGRSSN, DNO)

Fragmentación horizontal

- **Fragmentación horizontal completa:** conjunto de fragmentos horizontales cuyas condiciones C_1, C_2, \dots, C_n incluyen todas las tuplas en R (cada agrupación tiene cierto significado lógico)
- **Fragmentación horizontal completa y disjunta:** si no existen tuplas en R que satisfagan C_i y C_j para cualquier $i \neq j$
- Para reconstruir la relación R, aplicar la operación UNION a los fragmentos

Fragmentación vertical

- **Fragmentación vertical completa:** conjunto de fragmentos verticales cuyas listas de atributos L_1, L_2, \dots, L_n incluyen todos los atributos de la relación R, pero que comparten solamente la llave primaria de R
- Condiciones:
 - $L_1 \cup L_2 \cup \dots \cup L_n = \text{ATTRS}(R)$
 - $L_i \cap L_j = \text{PK}(R)$ para cualquier $i \neq j$ y $\text{PK}(R)$ es la llave primaria
- Para reconstruir la relación R, aplicar la operación JOIN a los fragmentos (sin fragmentación horizontal)

Ejemplos de fragmentaciones

- Fragmentación horizontal:
 - **Primaria:** Crear tres fragmentos de EMPLOYEE y de PROJECT usando los predicados (DNO = 5), (DNO = 4), (DNO = 1)
 - **Derivada:** Aplicar algún predicado utilizado en una fragmentación horizontal primaria, por ejemplo usar la información almacenada en DEPARTMENT para fragmentar EMPLOYEE y PROJECT
- Fragmentación vertical de EMPLOYEE:
 - $L_1 = \{ \text{SSN, NAME, SEX, BDATE, ADDRESS} \}$
 - $L_2 = \{ \text{SSN, SALARY, MGR, DNO} \}$

Fragmentación híbrida

- Combinar los fragmentos horizontales (DNO = 5), (DNO = 4), (DNO = 1) y verticales de la relación EMPLOYEE: seis fragmentos
- Los fragmentos son especificados por una combinación de selecciones y proyecciones de la forma $\pi_L(\sigma_C(R))$:
 - $C = \text{true}$ y $L \neq \text{ATTRS}(R)$ selecciona un fragmento vertical
 - $C \neq \text{true}$ y $L = \text{ATTRS}(R)$ selecciona un fragmento horizontal
 - $C \neq \text{true}$ y $L \neq \text{ATTRS}(R)$ selecciona un fragmento mezclado
 - $C = \text{true}$ y $L = \text{ATTRS}(R)$ selecciona toda la relación R

Ejemplo

- R1: artículos
 - A(NA, DESCR, PV, PR, P)
 - El artículo de número NA, descrito por la descripción DESCR cuyo precio de venta PV, precio de recuperación PR y es vendido por el proveedor P
- R2: tiendas
 - T(NT, CIUDAD, DIRM)
 - La tienda de número NT se localiza en la ciudad CIUDAD y tiene de director al empleado de número DIRM
- R3: ventas mensuales
 - V(NT, NA, MES, CANT)
 - La tienda de número NT vendió una cantidad CANT de artículos de tipo NA durante el mes MES
- R4: empleados
 - E(NE, NOM, EDAD, FECHAEMPLO, SAL, NT)
 - El empleado con número NE, de nombre NOM, de edad EDAD fue contratado en la fecha FECHAEMPLO, gana un salario SAL en la tienda NT
- R5: histórico de empleados
 - HE(NE, FECHAEMPLO, FIN, SAL)
 - El empleado identificado por el número NE, contratado en la fecha FECHAEMPLO ganaba un salario SAL en la fecha FIN
- R6: proveedores
 - F(NF, CIUDAD)
 - El proveedor identificado por el número NF se encuentra en la ciudad CIUDAD

Plan

- ✓ De las bases de datos centralizadas a las distribuidas
- ✓ Integración de datos
- ✓ Bases de datos distribuidas (BDD):
 - ✓ Arquitecturas
 - ✓ Funcionamiento general
 - ✓ Técnicas de diseño
- Sistemas de acceso a información modernos

Sistemas Peer to Peer

- **Sistemas para compartir archivos:** Estos sistemas asumen que hay suficientes *peers* en el sistema que ofrecen el recurso requerido
 - *Lime Wire*
 - *Ares (Gnutella)*
 - *Freenet*
 - *Torrent File*
- **Sistemas de mensajería instantánea:** Cada *peer* guarda información única, las consultas son específicas a un *peer*
 - *Skype*
 - *ICQ*

