

# Base de datos distribuidas

Prof. Alejandro Reyes Ortiz

**“PROCESAMIENTO DE CONSULTAS”**

# Introducción: Procesamiento de consultas

- El procesamiento de consultas es necesario aplicarlos para traducir las peticiones del usuario en procesamiento de datos a bajo nivel.
- El procesamiento de consultas significa mejorar el desarrollo de aplicaciones y la productividad del usuario final.
- Se permiten la expresión de consultas complejas de manera concisa y simple.

# Introducción: Procesamiento de consultas

- Los usuarios no especifican el procedimiento a seguir en la construcción de la respuesta a una consulta.
- Este procedimiento es responsabilidad de un módulo, dentro del Sistema de Base de Datos Distribuidas, llamado “procesador de consultas”.
- Este módulo libera al usuario de la optimización de la consulta, una tarea **que consume mucho tiempo** y que es mejor manejada por el “procesador de consultas” ya que puede **explotar grandes cantidades de información** acerca de los datos.

# Introducción: Procesamiento de consultas

- En un ambiente de un SBDD, el procesador de consultas el problema del procesamiento de consultas es complejo, debido a una gran lista de parámetros que afecta el desempeño de las consultas distribuidas:
  - Relaciones fragmentadas.
  - Replicación de relaciones o fragmentos.
  - Costos de comunicación (cantidad de datos a enviar)
  - Tiempo de respuesta de las consultas.

# Problema del procesamiento de consultas

1. Transformar una consulta de alto nivel (solicitud del usuario final) en su consulta equivalente en bajo nivel (SQL distribuida), pasando por la representación en algebra relacional.
2. La transformación debe ser **correcta** y **eficiente**.
3. La transformación será correcta si la consulta es semánticamente correcta y produce un resultado correcto.

# Problema del procesamiento de consultas

Considerar el siguiente esquema

EMP(ENO, ENAME, TITLE)  
ASG(ENO, PNO, RESP, DUR)

La siguiente consulta simple de usuario:

"Encontrar los nombres de los empleados que están administrando un proyecto"

```
SELECT ENAME
FROM EMP, ASG
WHERE EMP.ENO = ASG.ENO
AND RESP = 'Manager'
```

$$\Pi_{ENAME}(\sigma_{RESP='Manager' \wedge EMP.ENO=ASG.ENO} (EMP \times ASG))$$

# Problema del procesamiento de consultas

Considerar el siguiente esquema

EMP(ENO, ENAME, TITLE)  
ASG(ENO, PNO, RESP, DUR)

La siguiente consulta simple de usuario:

"Encontrar los nombres de los empleados que están administrando un proyecto"

```
SELECT ENAME
FROM EMP, ASG
WHERE EMP.ENO = ASG.ENO
AND RESP = 'Manager'
```

$$\Pi_{ENAME}(\sigma_{RESP='Manager' \wedge EMP.ENO=ASG.ENO}(EMP \times ASG))$$
$$\Pi_{ENAME}(EMP \bowtie_{ENO}(\sigma_{RESP='Manager'}(ASG)))$$

El rol del procesador de consultas es elegir, para una consulta dada, la mejor representación en álgebra relacional entre dos resultados equivalentes.

# Problema del procesamiento de consultas

Ahora un escenario de SBDD (base de datos fragmentada)

$\Pi_{ENAME}(EMP \bowtie_{ENO} (\sigma_{RESP="Manager"}(ASG)))$

Asuma que la relación EMP y ASG están horizontalmente fragmentadas:

$EMP_1 = \sigma_{ENO \leq "E3"}(EMP)$

$EMP_2 = \sigma_{ENO > "E3"}(EMP)$

$ASG_1 = \sigma_{ENO \leq "E3"}(ASG)$

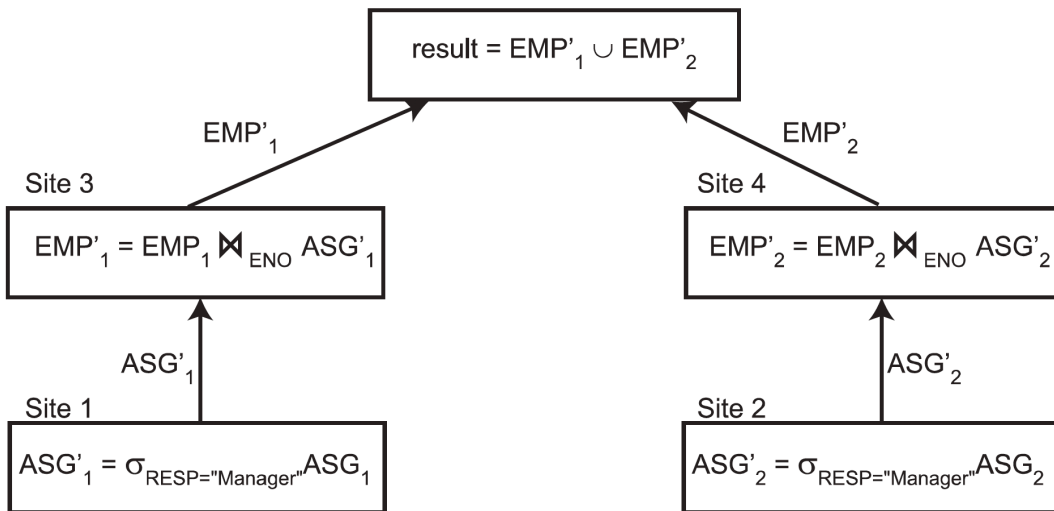
$ASG_2 = \sigma_{ENO > "E3"}(ASG)$

Los fragmentos ASG<sub>1</sub>, ASG<sub>2</sub>, EMP<sub>1</sub>, EMP<sub>2</sub> están almacenados en los sitios 1, 2, 3, 4 respectivamente.

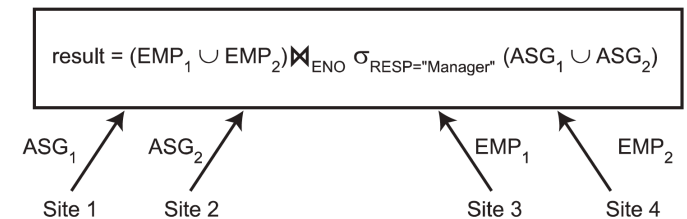


# Problema del procesamiento de consultas

## Estrategias de solución en un procesamiento distribuido



(a) Strategy A



(b) Strategy B

El rol del procesador de consultas es elegir, para una consulta dada, le mejor representación en algebra relacional entre dos resultados equivalentes.

# Objetivos del procesamiento de consultas en un contexto distribuido

- **Transformar** una consulta de alto nivel (solicitud del usuario final) sobre una base de datos distribuida, en su consulta equivalente en bajo nivel (operadores de álgebra relacional).
- Esta transformación **implica capas** (tareas), que serán presentadas más adelante.
- Uno de los aspectos importantes es seleccionar una alternativa, ya que hay muchas alternativas (transformaciones) correctas.
- Se debe **seleccionar aquella alternativa** que minimice (optimización) el consumo de recursos.

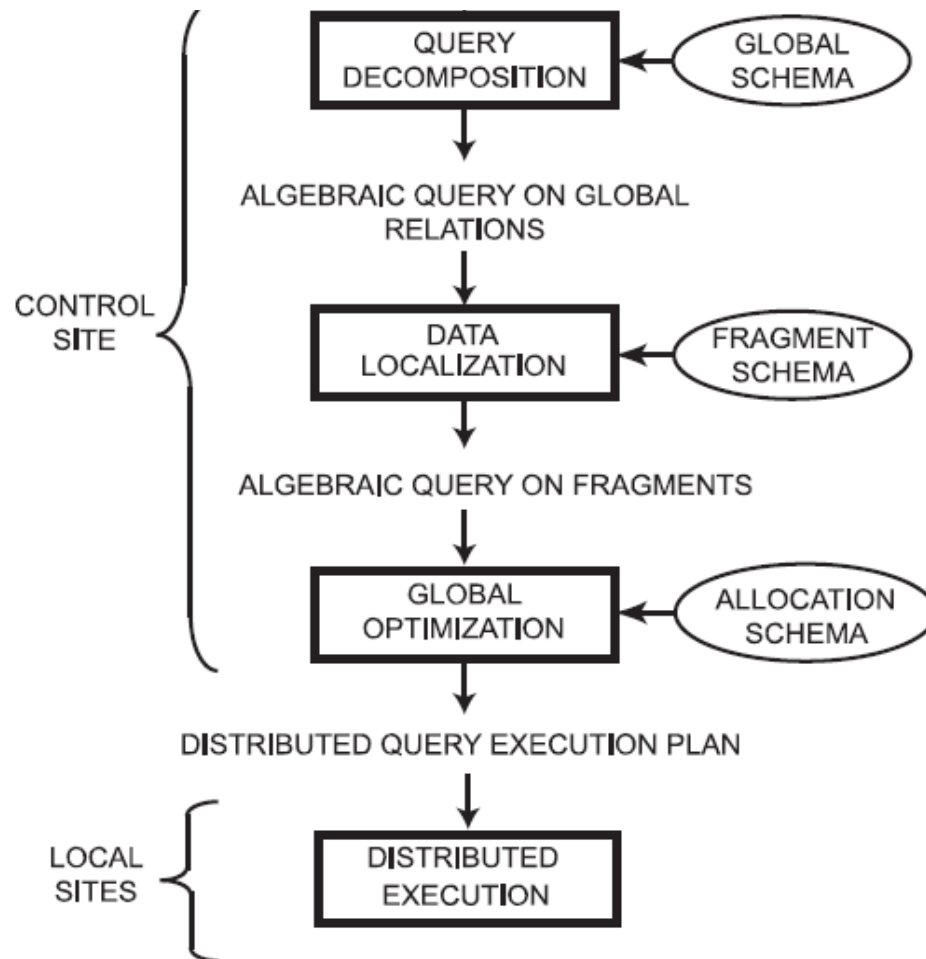
# Objetivos del procesamiento de consultas en un contexto distribuido

- Una buena medida del consumo de recursos es el **costo total** que se incurrirá en el procesamiento de la consulta.
  1. **CPU:** costo por el desempeño de las operaciones en memoria principal
  2. **I/O:** costo por acceder a disco
  3. **Comunicación:** costo de intercambiar datos entre servidores

# Características del procesador de consultas

- **Lenguaje:** Calculo relacional es una propuesta para el procesamiento de consultas; lenguajes para representar el modelo de fragmentación. Cualquier lenguaje donde se puedan implementar reglas de decisión.
- **Tipos de optimización :** elegir la mejor estrategia para la ejecución de la consulta, el espacio de búsqueda puede ser muy exhaustivo. Para evitar altos costos en las búsquedas: estrategias aleatorias.
- **Explotación de la topología:** ancho de banda , considerar el costo por la comunicación.
- **Explotación de fragmentos replicados:** localización de los datos involucrados en la consulta para minimizar los tiempos de comunicación.

# Capas del procesamiento de consultas distribuidas



# Descomposición de la consulta

- Crea la consulta algebraica, hace uso del esquema global de las relaciones.
- La información necesaria para dicha transformación es encontrada en el esquema global.
- La consulta transformada debe ser re-escrita en una forma **normalizada** (adecuada para su manipulación posterior), **analizada semánticamente** (detectar consultas incorrectas), **simplificar** (eliminar redundancia en los predicados) y **re-estructuración** (transformación final)

# Localización de datos

- La entrada de esta capa es la consulta algebraica sobre las relaciones globales.
- Su función es localizar los datos de la consulta utilizando la información de la distribución de los datos en el esquema de fragmentos.
- Determina qué fragmentos están implicados en la consulta y transforma la consulta distribuida en una consulta sobre fragmentos.

# Optimización global de la consulta

- Encontrar la estrategia para que la consulta sea óptima.
- Encuentra los mejores operadores para realizar la consulta.
- Toma en cuenta los costos.
- Depende de la localización de los fragmentos.
- Suele representarse y guardarse (para futuras ejecuciones) como un plan de ejecución de consultas distribuidas



# Ejecución de la consulta distribuida

- Esta la última capa del procesamiento de consultas, es realizada por todos los sitios que tienen fragmentos implicados en la consulta.
- Cada sub-consulta que se ejecuta en un sitio, llamada consulta local, se ejecuta utilizando el esquema local del sitio.

# Conclusiones

- En resumen, ...
- ¿Cómo lo aplicarían a su sistema de BDD?