# BD Avanzada

## Tema 5 - Bases de Datos Paralelas

# Bases de Datos Paralelas

**Objetivo General**:

- Al completar este tema el estudiante estará en la capacidad de manejar los conceptos sobre Bases de Datos Paralelas.

**Objetivos Específicos**: Al completar este tema el estudiante deberá estar en la capacidad de discutir sobre los siguientes conceptos:

- Bases de datos paralelas
- Algoritmos que se utilizan para manejar los datos.

# Bases de Datos Paralelas

- **Paralelismo**
  - Se utiliza para proporcionar aceleración, y las consultas se ejecutan más rápido debido a que se proporcionan más recursos, como procesadores y discos.
  - Se utiliza para proporcionar ampliabilidad, y las cargas de trabajo crecientes se tratan sin aumentar el tiempo de respuesta mediante un aumento en el grado de paralelismo.
- **Paralelismo de E/S**
  - Se refiere a la reducción del tiempo necesario para recuperar relaciones del disco dividiéndolas en varios discos.

# Bases de Datos Paralelas

- **Técnicas de división**
  - **Turno rotatorio.** Asegura una distribución homogénea de las tuplas entre los discos. Cada disco tiene aproximadamente el mismo número de tuplas que los demás.
  - **División por asociación.** Uno o más atributos del esquema de la relación dada se designan como atributos de la división.
  - **División por rangos.** Distribuye rangos contiguos de valores de los atributos a cada disco.

# Bases de Datos Paralelas

- **Clasificación del acceso a los datos**

1. Explorar la relación completa.

2. **Consultas concretas**, buscan tuplas que tengan un valor concreto para un atributo concreto.

3. **Consultas de rangos**, localizan todas las tuplas cuyo valor de un atributo dado se halle en un rango especificado.

# Bases de Datos Paralelas

- **Tipos de acceso**
  - **Turno rotatorio.** El esquema se adapta perfectamente a las aplicaciones que desean leer secuencialmente la relación completa para cada consulta.
  - **División por asociación.** Este esquema se adapta mejor a las consultas concretas basadas en el atributo de división.
  - **División por rangos.** Este esquema se adapta bien a las consultas concretas y de rangos basadas en el atributo de división.

# Bases de Datos Paralelas

- **Clasificación del sesgo**
  - **Sesgo de los valores de los atributos**, se refiere al hecho de que algunos valores pueden aparecer en los atributos de división de muchas tuplas.
  - **Sesgo de la división**, se refiere al hecho de que puede haber un desequilibrio en la carga de la división, aunque no haya sesgo en los atributos.

# Bases de Datos Paralelas

- **Paralelismo entre consultas**
  - Se ejecutan en paralelo entre sí diferentes consultas o transacciones.
  - Se usa principalmente para ampliar los sistemas de procesamiento de transacciones para permitir un número mayor de transacciones por segundo.
  - Cuando un procesador tiene acceso a los datos o los actualiza, el sistema de bases de datos debe asegurar que el procesador tenga la *última versión de éstos en su memoria intermedia* - **coherencia caché**.

# Bases de Datos Paralelas

- **Paralelismo en consultas**
  - Ejecución en paralelo de una única consulta en varios procesadores y discos.
  - Se usa para acelerar las consultas de ejecución larga.
- **Tipos de consultas en paralelo**:
  - **Paralelismo en operaciones.** Se puede acelerar el procesamiento de consultas haciendo paralela la ejecución de cada una de las operaciones, como puede ser la ordenación, la selección, la proyección y la reunión.
  - **Paralelismo entre operaciones.** Se puede acelerar el procesamiento de consultas ejecutando en paralelo las diferentes operaciones de las expresiones de las consultas.

# Bases de Datos Paralelas

- **Paralelismo en operaciones**
  - Dado que las operaciones relacionales trabajan con relaciones que contienen grandes conjuntos de tuplas, se pueden paralelizar las operaciones ejecutándolas en paralelo en subconjuntos diferentes de las relaciones.
  - Dado que el número de tuplas de una relación puede ser grande, el grado de paralelismo es potencialmente enorme.
  - Se puede ordenar por separado cada partición y concatenar los resultados para obtener la relación completa ordenada.

# Bases de Datos Paralelas

- **Ordenación en paralelo**
  - ***Ordenación con división por rangos***, primero se divide por rangos la relación y después se ordena cada partición.
  - ***Ordenación y mezcla externas paralelas***, cada procesador *P o*rdena localmente los datos en el disco. Las partes ordenadas por cada procesador se mezclan luego para obtener el resultado ordenado final.
- **Reunión paralela**
  - Exige que se comparen pares de tuplas para ver si satisfacen la condición de reunión. Si lo hacen, el par se añade al resultado reunido. Los algoritmos de reunión paralela intentan repartir entre varios procesadores los pares que hay que comparar.
- ***Reunión por división***
  - Para ciertos tipos de reuniones, como las equirreuniones y las naturales, es posible *dividir* las dos relaciones de entrada entre los procesadores y procesar localmente la reunión en cada uno de ellos.
- ***Reunión con fragmentos y réplicas***
  - **1.** Se divide una de las relaciones (digamos *r*). Se puede utilizar en *r* cualquier técnica de división, incluyendo la división por turno rotatorio.
  - **2.** La otra relación, *s*, se replica entonces en todos los procesadores.
  - **3.** El procesador *P* procesa entonces localmente la reunión de $r_i$ con toda *s*, utilizando cualquier técnica de reunión.

# Bases de Datos Paralelas

- **Costo de la evaluación en paralelo de las operaciones**
  - **Costos de iniciar** la operación en varios procesadores.
  - **Sesgo** en la distribución de trabajo entre los procesadores, con algunos procesadores con mayor número de tuplas que otros.
  - **Contención de recursos** —como la memoria, los discos y la red de comunicaciones— que dan lugar a retrasos.
  - **Costo de construir** el resultado final transmitiendo los resultados parciales desde cada procesador.

# Bases de Datos Paralelas

- **Paralelismo entre operaciones**
  - **Paralelismo de encauzamiento**, ejecutar simultáneamente operaciones en procesadores diferentes de modo que la operación *B* consuma las tuplas en paralelo con su producción por la operación *A*.
  - **Paralelismo independiente**, las operaciones en las expresiones de las consultas que son independientes entre sí pueden ejecutarse en paralelo.
  - **Optimización de consultas**, toma una consulta y encuentra el plan de ejecución más económico de entre los muchos planes de ejecución posibles que proporcionan la misma respuesta.
  - **Intercambio de operadores**, modelo que usa implementaciones existentes de operaciones, actuando sobre copias locales de los datos, junto con una operación de intercambio que traslada los datos entre diferentes procesadores.