Bases de datos distribuidas

Alex Di Genova

15/11/2023

PMD

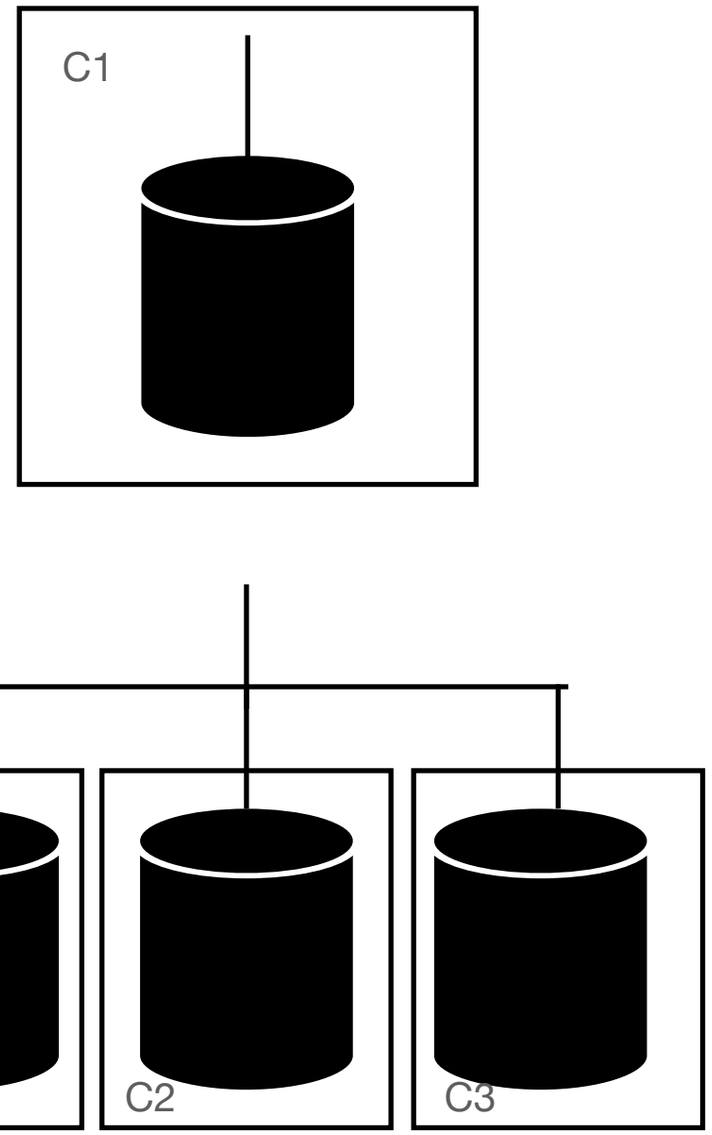
Contenidos finales

- Introducción a las bases de datos distribuidas
- Estrategias de data-placement: sharding (partitioning), replication, duplication
- Procesamiento de consultas distribuidas y su optimización

Bases de datos Distribuidas

Sistemas centralizados

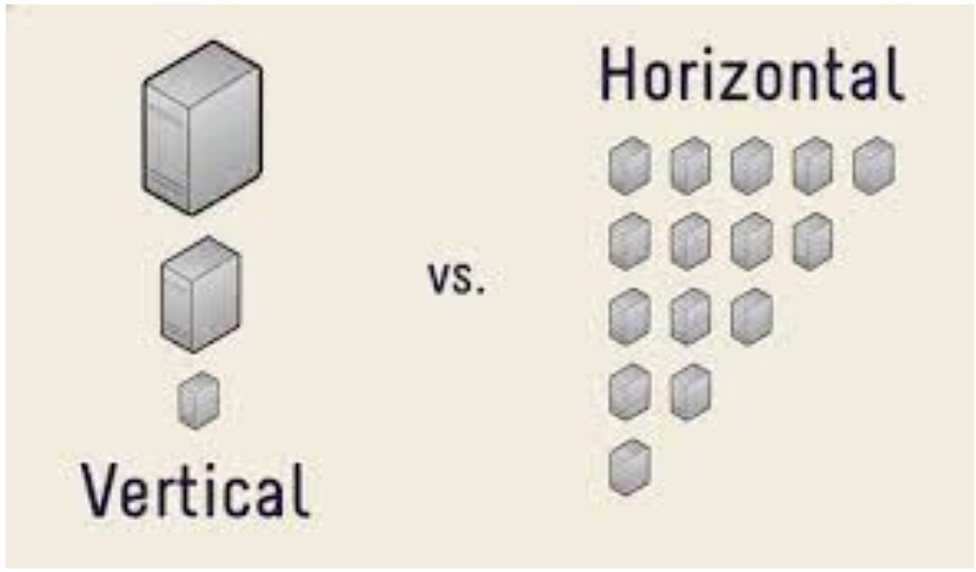
- Durante varias décadas, los sistemas de gestión de bases de datos centralizados, que se ejecutan en un único servidor de bases de datos, han sido predominantes.
    - La complejidad de un sistema de servidor único es menor y la administración más fácil
    - El caso de uso frecuente era realizar consultas cortas (lecturas frecuentemente) en un conjunto coherente de datos, mientras que las modificaciones de datos ocurrian rara vez (escrituras poco frecuentes)
    - La velocidad de transmision de la red era lenta y, por lo tanto, enviar datos entre diferentes servidores era demasiado costoso.
    - La paralelización requería reescribir una consulta en subconsultas y combinar los resultados y esta carga disminuyó los efectos positivos de una ejecución paralela de subconsultas.



Bases de datos Distribuidas

Sistemas centralizados

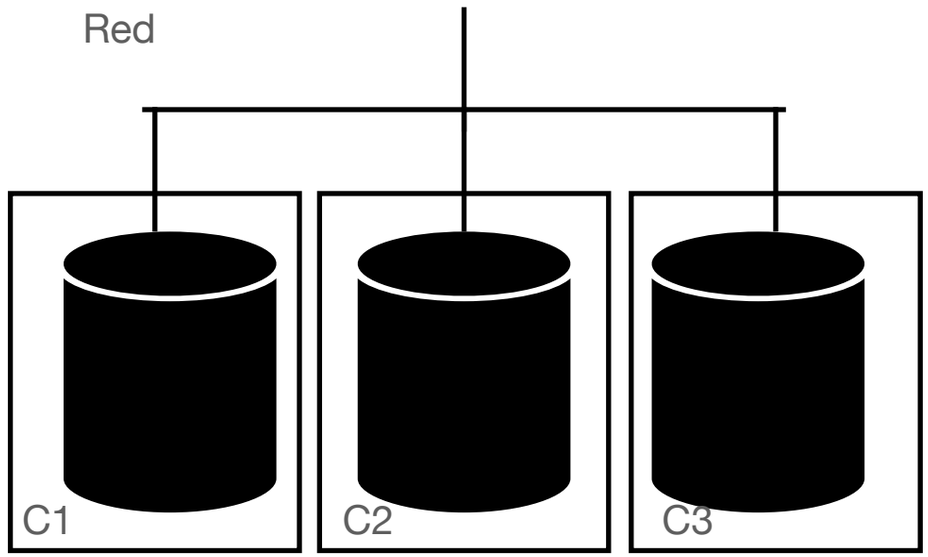
- Escalamiento vertical
    - Cada vez que había requisitos de mayor volumen de datos o escrituras más frecuentes, la reacción obvia era equipar el servidor de base de datos con mayor capacidad en términos de velocidad de procesador, tamaño de memoria o espacio en disco.
- Escalamiento horizontal
    - Conectar varios servidores más en una red permite mejorar el rendimiento y la latencia de un sistema de base de datos a costa de la **coordinación y sincronización** de los servidores.



Una base de datos distribuida es una colección de datos que se distribuyen físicamente en varios servidores interconectados por una red mientras que lógicamente estan relacionados.

Bases de datos Distribuidas

Sistemas distribuidos



- Un motor de datos distribuido (MDD) puede resultar beneficioso no solo cuando se manejan grandes volúmenes de datos, sino también cuando se busca mejorar la disponibilidad y la confiabilidad en sistemas de menor escala.

- **Balanceo de carga**: las consultas de los usuarios y otros procesos deben asignarse a los servidores de la red de modo que todos los servidores tengan aproximadamente la misma carga

- **Escalabilidad flexible**: los servidores pueden abandonar y unirse a la red de manera flexible en cualquier momento para que el MDD pueda reconfigurarse de acuerdo con las demandas actuales de almacenamiento o rendimiento.

- **Nodos heterogéneos**: el MDD puede ejecutarse en una red de servidores donde algunos servidores pueden tener más capacidades que otros.

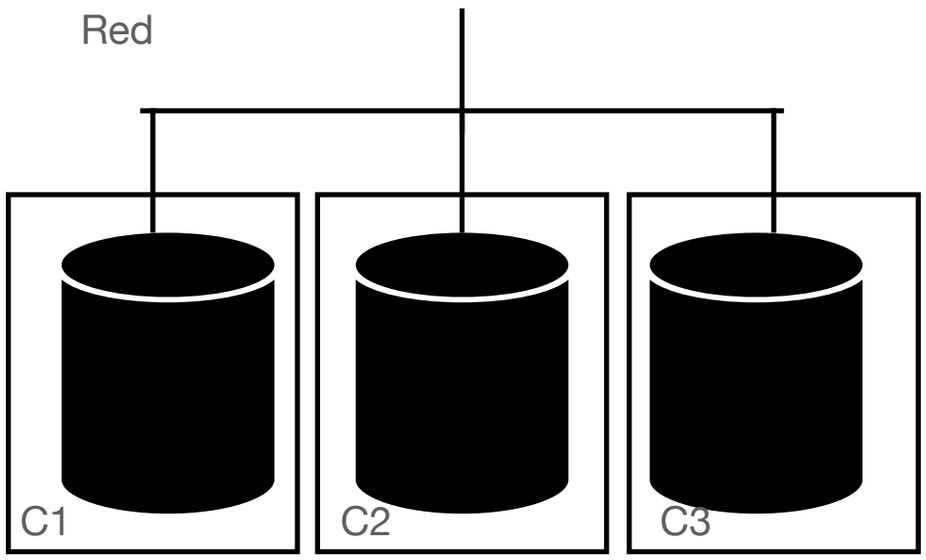
- **Configuración simétrica:** Cada nodo se configura de forma idéntica a los demás; por lo tanto, cada nodo tiene la capacidad de reemplazar un nodo fallido. En particular, las consultas de los usuarios pueden ser manejadas por cualquier servidor del sistema.

Para un usuario, el MDD distribuido debería aparecer como si estuviera interactuando con un único servidor centralizado. En particular, debe permitir que el usuario envíe su consulta a un unico nodo del sistema y el MDD adapta y redirige la consulta a uno o más nodos de datos.

Bases de datos Distribuidas

Sistemas distribuidos

- Para el usuario, debe ser transparente cómo el MDD maneja internamente el almacenamiento de datos y el procesamiento de consultas de manera distribuida.
- **Acceso:** el sistema de base de datos distribuida proporciona una sistema consulta y una interfaz de administración uniforme para los usuarios, independientemente de la estructura de la red o la organización de almacenamiento.
- **Localización:** la distribución de datos en el sistema de base de datos (y, por lo tanto, la ubicación exacta de cada elemento de datos) está oculta para el usuario. El usuario puede consultar datos sin tener que especificar qué elemento de datos se recuperará de qué servidor de base de datos en la red.
- **Replicación:** si se almacenan varias copias de un elemento de datos en diferentes servidores (por motivos de recuperación y disponibilidad), el usuario no debe saberlo y no debe preocuparse por las copias a las que accede.
- **Fragmentación:** si un gran conjunto de datos debe dividirse en varios elementos de datos (generalmente llamados fragmentos o particiones), el sistema de base de datos distribuida realiza esta división internamente y el usuario puede consultar la base de datos como si contuviera todo el conjunto de datos sin fragmentar. En particular, para responder a la consulta de un usuario, las subconsultas se redireccionan a diferentes servidores y el sistema de base de datos recombina los elementos de datos relevantes para la consulta.
- **Migración:** si algunos elementos de datos deben moverse de un servidor a otro, esto no debería afectar la forma en que un usuario accede a los datos.
- **Concurrencia**: cuando varios usuarios acceden al sistema de base de datos, su operación no debe interferir ni dar lugar a datos incorrectos en el sistema de base de datos. La concurrencia es mucho más difícil de administrar para un sistema distribuido que para uno centralizado. Un problema importante es cómo resolver los conflictos debido a la naturaleza distribuida del sistema.
- **Fallas**: como un sistema de base de datos distribuido es más complejo que uno centralizado, pueden surgir muchos más casos de falla.Por lo tanto, el sistema de base de datos distribuida debe hacer todo lo posible para continuar procesando las solicitudes de los usuarios incluso en presencia de fallas.



Para un usuario, el MDD distribuido debería aparecer como si estuviera interactuando con un único servidor centralizado.

Bases de datos Distribuidas

Errores

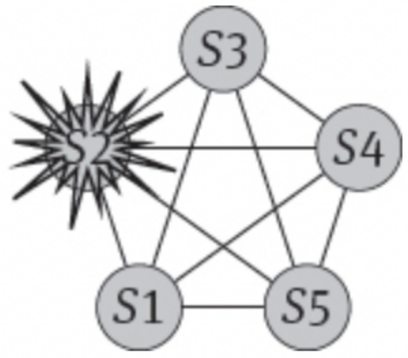
- En un sistema distribuido con varios componentes independientes conectados por una red, partes de la red pueden fallar.

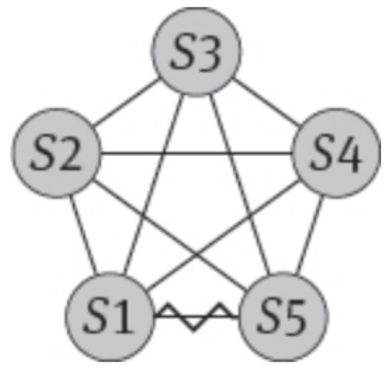
- **Servidor**: un servidor de base de datos puede fallar al procesar los mensajes que recibe, por ejemplo, debido a un componente de red defectuoso; o el servidor falla por completo y debe reiniciarse. Los servidores también pueden retrasarse en el procesamiento de mensajes (debido a sobrecarga) o pueden enviar mensajes incorrectos debido a errores al procesar los datos.

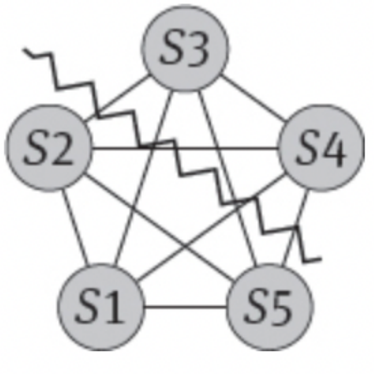
- **Errores de mensajes**: cuando los mensajes se transmiten a través de enlaces de comunicación en la red, los mensajes pueden retrasarse o perderse durante momentos de alta congestión de la red.

- **Errores de enlace**: es posible que un enlace de comunicación entre dos servidores no pueda transmitir mensajes, o puede corromper o duplicar mensajes. Por lo tanto, una falla de enlace puede causar una falla de mensaje.

- **Partición de red**: una red se particiona cuando se divide en dos o más subredes que no pueden comunicarse porque todos los enlaces de comunicación entre ellas están rotos.







Bases de datos Distribuidas

Comunicación

- Debido a las muchas propiedades (como tolerancia a fallas y escalabilidad) que debe tener un sistema de base de datos distribuida, la propagación de información (como listas de miembros o actualizaciones de datos) en la red de servidores de bases de datos es difícil de administrar.

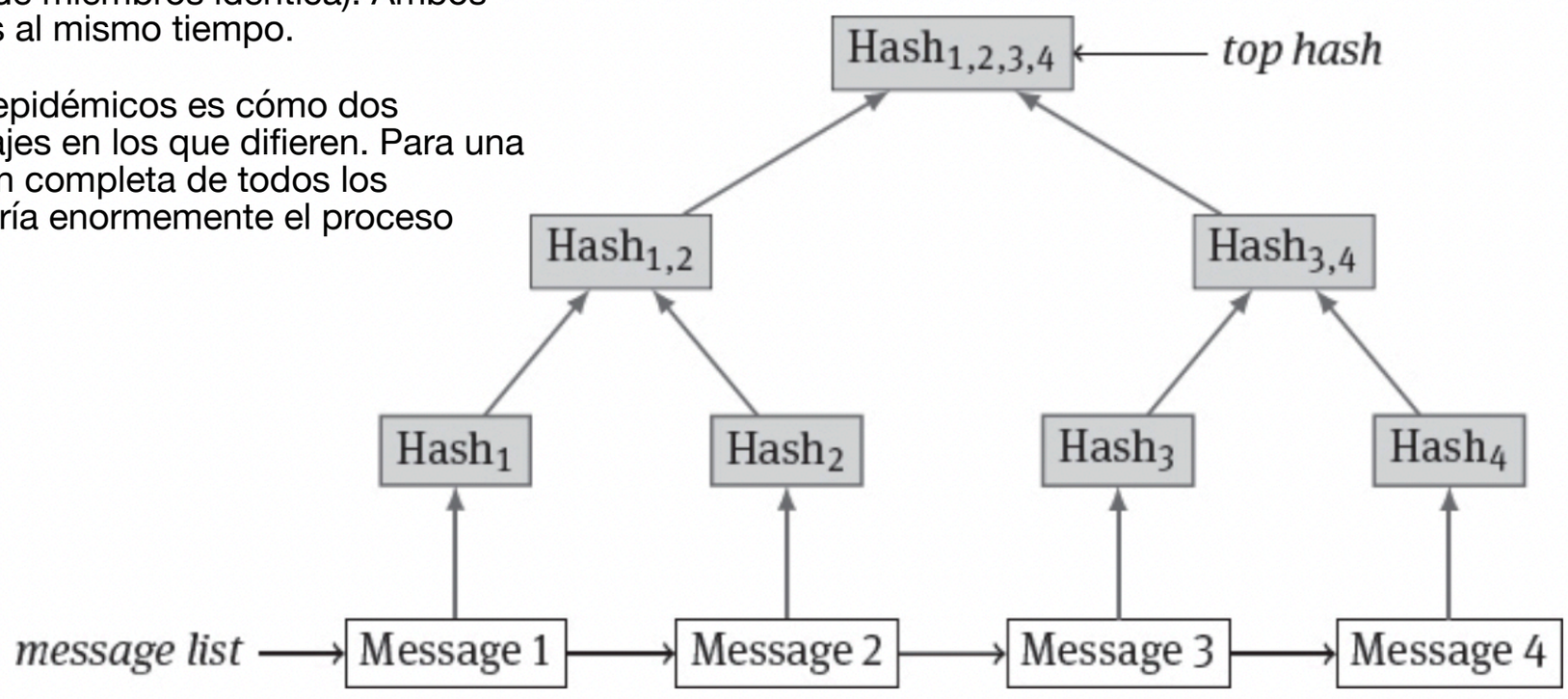
- Los protocolos epidémicos son una categoría de algoritmos peer-to-peer, donde la información (por ejemplo, actualizaciones de datos) se propaga como una infección por toda la red de servidores.

    - push-only: Un servidor infectado contacta a otro servidor y transmite todos los mensajes nuevos que ha recibido. Para propagar la infección, el servidor infectado tiene que encontrar otro servidor que sea susceptible.

    - pull-only: un servidor susceptible se pone en contacto con otro servidor y solicita nuevos mensajes. Para propagar la infección, el servidor susceptible tiene que encontrar un servidor infectado.

    - push-pull: un servidor contacta a otro servidor y ambos intercambian sus nuevos mensajes. Después de este intercambio, ambos servidores tienen el mismo estado (por ejemplo, una lista de miembros idéntica). Ambos servidores son susceptibles e infectados al mismo tiempo.

- Un problema importante con los protocolos epidémicos es cómo dos servidores pueden identificar aquellos mensajes en los que difieren. Para una gran cantidad de mensajes, una comparación completa de todos los mensajes no es factible ya que esto ralentizaría enormemente el proceso epidémico.



Bases de datos Distribuidas

fragmentación de datos

- En un sistema de base de datos distribuida, dos preguntas importantes son (1) cómo se puede dividir el conjunto completo de datos en la base de datos en subconjuntos (**fragmentación de datos**) y (2) cómo se pueden distribuir los subconjuntos entre los servidores de base de datos en la red (**asignación de datos**).

- Qué buenas fragmentaciones y buenas asignaciones son para una base de datos determinada?
  - depende en gran medida de las características de ejecución del sistema.
  - El tipo de accesos (mayormente lecturas o mayormente escrituras),
  - Los patrones de acceso (a qué registros de datos se accede regularmente y a cuáles rara vez),
  - La afinidad de los registros (a qué registros de datos se accede y junto con qué otros registros de datos),
  - a frecuencia de los accesos y
  - La duración de los accesos.

Bases de datos Distribuidas

fragmentación de datos

- Una vez que se ha establecido una buena fragmentación y asignación, una base de datos distribuida puede aprovechar

- Localidad de datos (idealmente, los registros de datos en el mismo fragmento a menudo se acceden juntos),

- Minimización de los costos de comunicación (idealmente, no es necesario mover registros de datos a otro servidor para responder una consulta),

- Mejor eficiencia de la gestión de datos (idealmente, las consultas en fragmentos más pequeños se pueden ejecutar más rápido que en el gran conjunto de datos y las estructuras de índice pueden ser más pequeñas y, por lo tanto, los registros de datos se pueden encontrar más rápido),

- Balanceo de carga (idealmente, a todos los servidores se les asigna la cantidad óptima de datos y solo tienen que procesar las consultas de los usuarios de acuerdo con las capacidades de cada servidor y sin el peligro de los puntos de acceso, es decir, sin sobrecargar algunos servidores con la mayoría de los usuarios). consultas mientras que los otros servidores están en su mayoría inactivos)

Bases de datos Distribuidas

fragmentación de modelos relacionales

- Los dos enfoques básicos son la fragmentación vertical y la fragmentación horizontal. Además, la fragmentación derivada se basa en una fragmentación horizontal determinada y la fragmentación híbrida combina tanto la fragmentación vertical como la horizontal.

Original data	A	B	C	D
	a_1	b_1	c_1	d_1
	a_2	b_2	c_2	d_2
	a_3	b_3	c_3	d_3

Fragment 1	ID	A	B
	1	a_1	b_1
	2	a_2	b_2
	3	a_3	b_3

Fragment 2	ID	C	D
	1	c_1	d_1
	2	c_2	d_2
	3	c_3	d_3

Original data	A	B	C	D
	a_1	b_1	c_1	d_1
	a_2	b_2	c_2	d_2
	a_3	b_3	c_3	d_3

Fragment 1	A	B	C	D
	a_1	b_1	c_1	d_1
	a_2	b_2	c_2	d_2

Fragment 2	A	B	C	D
	a_3	b_3	c_3	d_3

Bases de datos Distribuidas

Typos de fragmentación

- **Manual**: una fragmentación manual requiere que el administrador de base de datos identifique fragmentos en un conjunto de datos y configure el sistema utilizando estos fragmentos.
- **Random**: cada registro de datos tiene la misma probabilidad de ser asignado a un fragmento.
- **Basada en la estructura**: Analiza la definición del esquema de datos (o el modelo de datos en general) e identifica las subestructuras que constituyen los fragmentos.
- **Basada en valores**: Analiza los valores contenidos en los elementos de datos para definir los fragmentos.
- **Basada en rangos:** como una forma especial de fragmentación basada en valores, la fragmentación basada en rangos divide la clave principal (u otro atributo que se puede ordenar) en intervalos inconexos pero consecutivos. Cada intervalo define un fragmento.
- **Basada en hashing**: si cada registro de datos tiene una clave el hash de esta clave puede determinar el servidor de base de datos al que se asigna el registro.
- **Basada en costos**: Se basa en una función de costos para encontrar una buena fragmentación. El objetivo es encontrar una fragmentación con un costo total mínimo.
- **Basada en afinidad:** Se basa en una especificación de qué tan afines son ciertos registros de datos; con qué frecuencia se accede a ellos juntos en una solicitud de lectura.
- **Clustering**: se pueden usar algoritmos especializados para encontrar subconjuntos coherentes en los datos (los llamados clústeres).

Bases de datos Distribuidas

fragmentación de modelos relacionales

- **Vertical**: Subconjuntos de columnas forman los fragmentos (**basada en estructura**) e identificables por una tupla clave (recombinacion mediante **joins**). En el ejemplo suponemos que A,B y C,D son más afines que los otros subset de columnas.

- **Horizontal**: Subconjuto de filas forman los fragmentos (**basada en valores**). La recombinación se logra tomando la **unión** de las tuplas en los diferentes fragmentos.

- **Derivada**: Una fragmentación horizontal en una tabla primaria induce una fragmentación horizontal en una tabla secundaria relacionada. Los fragmentos primarios y derivados con valores coincidentes para los atributos de unión se pueden almacenar en el mismo servidor

- **Hibrida**: Combinación arbitraria de fragmentacion horizontal y vertical.

- Fragmentaciones para bases de datos basadas en archivos, clave-valor y grafos siguen las mismas ideas.

Original data	A	B	C	D
	a_1	b_1	c_1	d_1
	a_2	b_2	c_2	d_2
	a_3	b_3	c_3	d_3

Fragment 1	ID	A	B
	1	a_1	b_1
	2	a_2	b_2
	3	a_3	b_3

Fragment 2	ID	C	D
	1	c_1	d_1
	2	c_2	d_2
	3	c_3	d_3

Original data	A	B	C	D
	a_1	b_1	c_1	d_1
	a_2	b_2	c_2	d_2
	a_3	b_3	c_3	d_3

Fragment 1	A	B	C	D
	a_1	b_1	c_1	d_1
	a_2	b_2	c_2	d_2

Fragment 2	A	B	C	D
	a_3	b_3	c_3	d_3

Bases de Datos Distribuidas

Asignación de datos

- Los fragmentos idealmente deben distribuirse uniformemente en todos los servidores (balanceo de carga).
- Tipos de asignación:
  - **Basada en rangos**
  - **Basada en costos**: Problema de optimización que intentamos programación lineal.
    - Similar al problema de ensamble de contenedores (bin pa
      - K servidores corresponden a K contenedores
      - Los contenedores tienen una capacidad maxima W
      - n framentos corresponden a n objetos.
      - Cada objeto tiene u peso (capacidad) wi < W
      - Los objetos deben ser ubicados en un numero minim sin exeder la capacidad maxima.
    - En una formulación simple podemos considerar la capacidad de almacenamiento (W) de los servidores y el consumo de cada fragmento (wi).

$$\text{minimize} \sum_{k=1}^{K} y_k \qquad \text{(minimize amount of servers)}$$

$$\text{subject to} \sum_{k=1}^{K} x_{ik} = 1 \qquad \text{(each fragment } i \text{ assigned to one server)}$$

$$\sum_{i=1}^{n} w_i \cdot x_{ik} \le W \cdot y_k \qquad \text{(capacity of each server } k \text{ not exceeded)}$$

$$y_k \in \{0, 1\} \qquad k = 1, \ldots, K$$

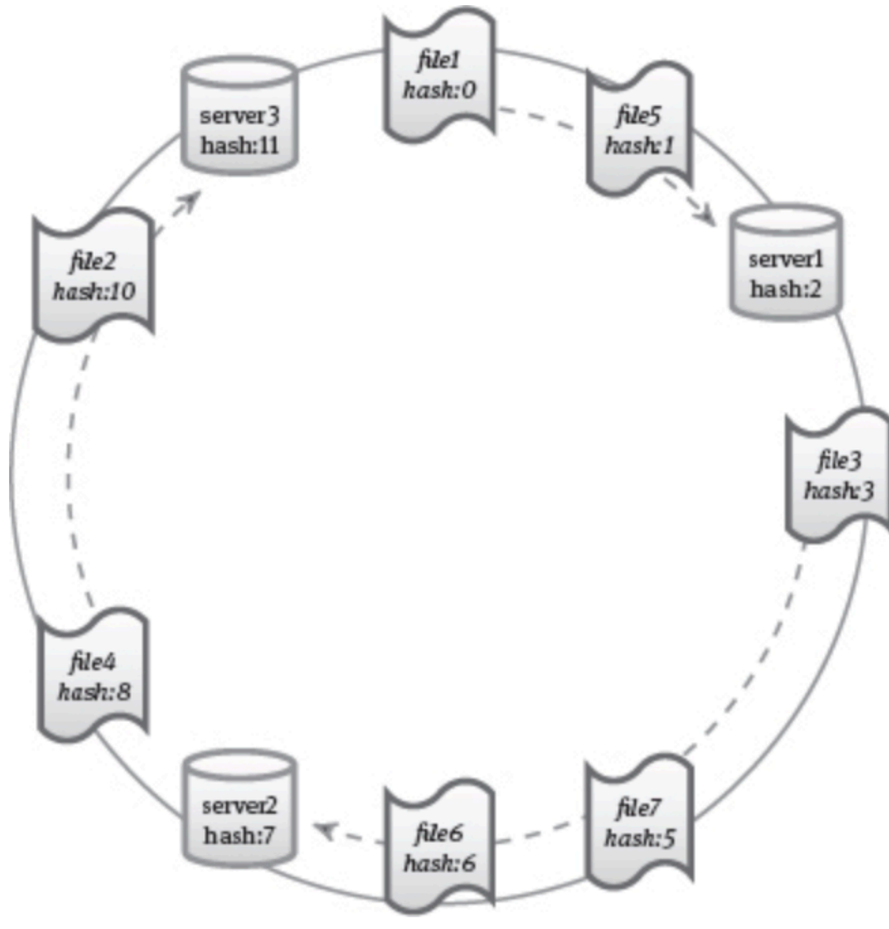
$$x_{ik} \in \{0, 1\} \qquad k = 1, \ldots, K, \;\; i = 1, \ldots, n$$

- **Basada en hashing**: utiliza una función hash (como MD5 o MurmurHash) sobre los fragmentos para determinar el servidor al que se asigna cada fragmento.

Bases de Datos Distribuidas

Asignación de datos hashing

- Es la estrategia de asignacion más usada.
- Los valores hash se ven como un anillo: cuando alcanzamos el valor hash más alto, comenzamos nuevamente desde 0.
- El valor hash se calcula no solo para cada fragmento sino también para cada servidor de base de datos.
- Al calcular un valor hash para un servidor de base de datos, cada servidor tiene una posición fija en el anillo; la ventaja de estos valores hash es que presumiblemente distribuyen los servidores uniformemente en el anillo. De manera similar, para cada elemento de datos se calcula un valor hash. De esta manera, los datos también se distribuyen bien en el anillo.



Bases de Datos Distribuidas

Replicas y sincronización

- La replicación se refiere al concepto de almacenar varias copias de un registro de datos en diferentes servidores de bases de datos (numero de copias predeterminado -> factor).

  - Mejora la confiabilidad del sistema de base de datos distribuida al ofrecer una mayor disponibilidad de datos: si una de las réplicas no puede manejar una consulta de usuario, otra réplica puede tomar el relevo.

  - Ofrece una latencia más baja que un sistema no replicado al permitir el balanceo de carga, la localidad de datos y la paralelización: cualquier réplica puede responder a las consultas de un usuario.

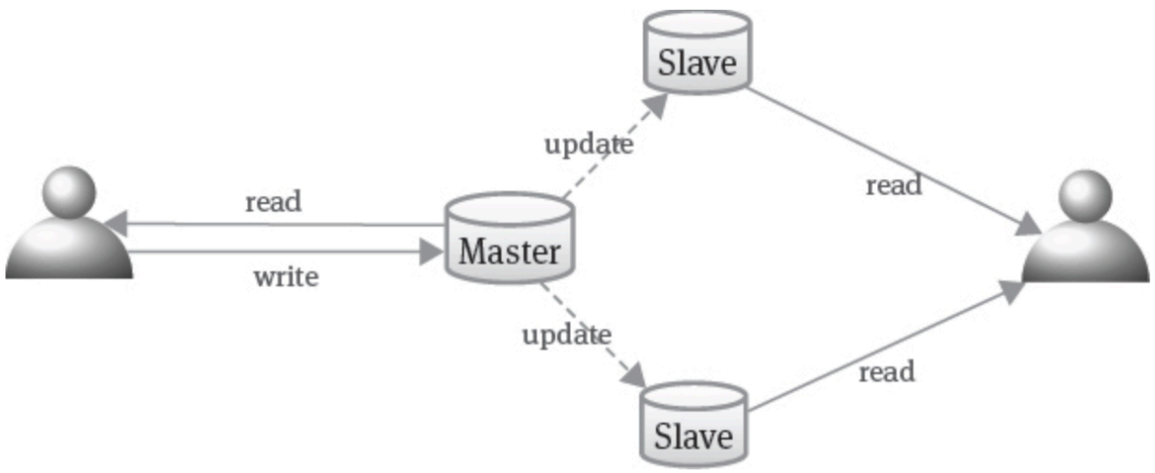
  - Problema de concurrencia: dos o más usuarios pueden actualizar simultáneamente el mismo registro de datos en diferentes réplicas y el sistema de base de datos debe ofrecer un mecanismo para resolver este conflicto.

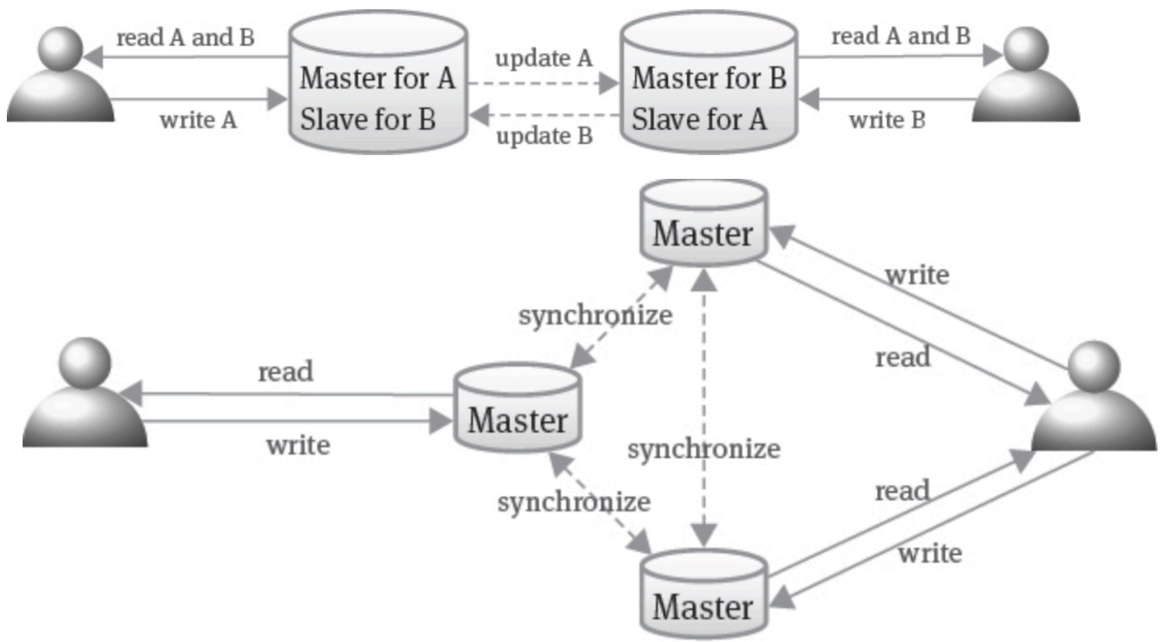
Bases de Datos Distribuidas

Modelos de replicas

- **Modelo maestro-esclavo**: El nodo maestro controla las operaciones de escritura.
    - Tener un único servidor maestro para todas las solicitudes de escritura en el sistema de la base de datos es un cuello de botella. Una solución pragmática es dividir el conjunto de todos los registros en particiones separados y asignar a cada partición un servidor maestro.

- **Modelo de multiples maestros**: Todos los servidores aceptan solicitudes de escritura y lectura para un elemento pero los servidores tienen que sincronizar regularmente su estado entre ellos.





Bases de Datos Distribuidas

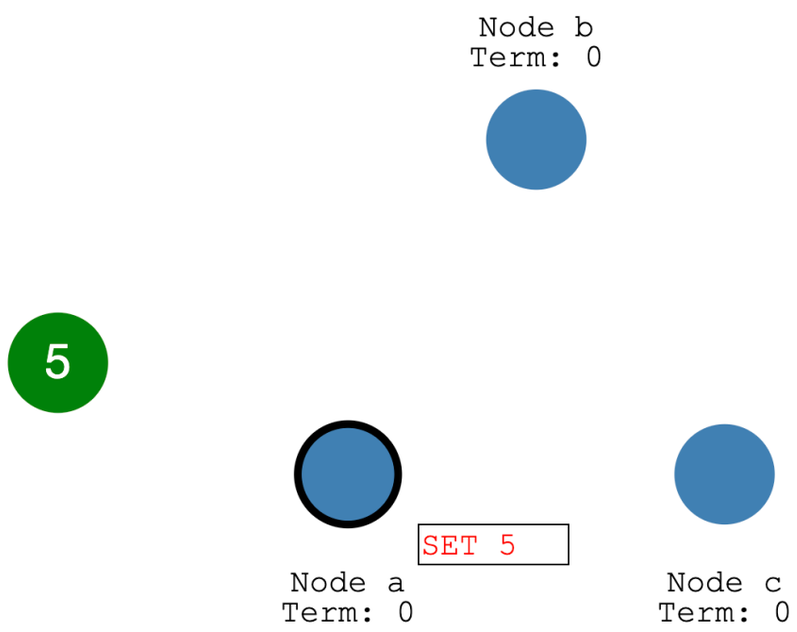
Control de concurrencia distribuido

- El control de concurrencia distribuido asegura la correcta ejecución de las operaciones (más generalmente, las transacciones) que afectan a los datos que se almacenan de forma distribuida en diferentes servidores de bases de datos.
    - En el caso de la replicación de datos, una aplicación típica de un protocolo de control de concurrencia es sincronizar todas las réplicas de un registro cuando se emite una escritura a uno de los servidores de bases de datos.

- Existen distintos protocolos, pero los mas usados son los protocolos de **consenso de quórum** que requieren una cierta mayoría de agentes para acordar un valor propuesto. La definición exacta de mayoría depende de el comportamiento de lectura y escritura y los tipos de fallas contra los que el protocolo debe ser resistente.

Bases datos distribuidas

El algoritmo RAFT



- http://thesecretlivesofdata.com/raft/

Consultas?

Consultas o comentarios?

Muchas gracias