Unidad 2
Tecnologías

Bases de Datos en Memoria

Es una base de datos cuyos datos están almacenados en la memoria principal (RAM) para facilitar tiempos más rápidos de respuesta. Los datos de origen se cargan a la memoria del sistema en un formato comprimido no relacional.

- Es un tipo de base de datos de análisis
- De solo lectura que almacena datos históricos sobre indicadores para aplicaciones de inteligencia empresarial/análisis de negocios
- Se actualiza regularmente para incorporar datos de transacción recientes
- Puede reducir o eliminar la necesidad de indexar datos

On-Disk Vs In-Memory

- Todos los datos se almacenan en el disco, se necesita I/O de disco para mover los datos a la memoria principal cuando sea necesario.
- Los datos siempre se conservan en el disco.
- Estructuras de datos tradicionales como B-Trees diseñadas para almacenar tablas e índices de manera eficiente en el disco.
- Admite un conjunto muy amplio de cargas de trabajo, por ejemplo, OLTP, almacenamiento de datos, cargas de trabajo mixtas, etc.

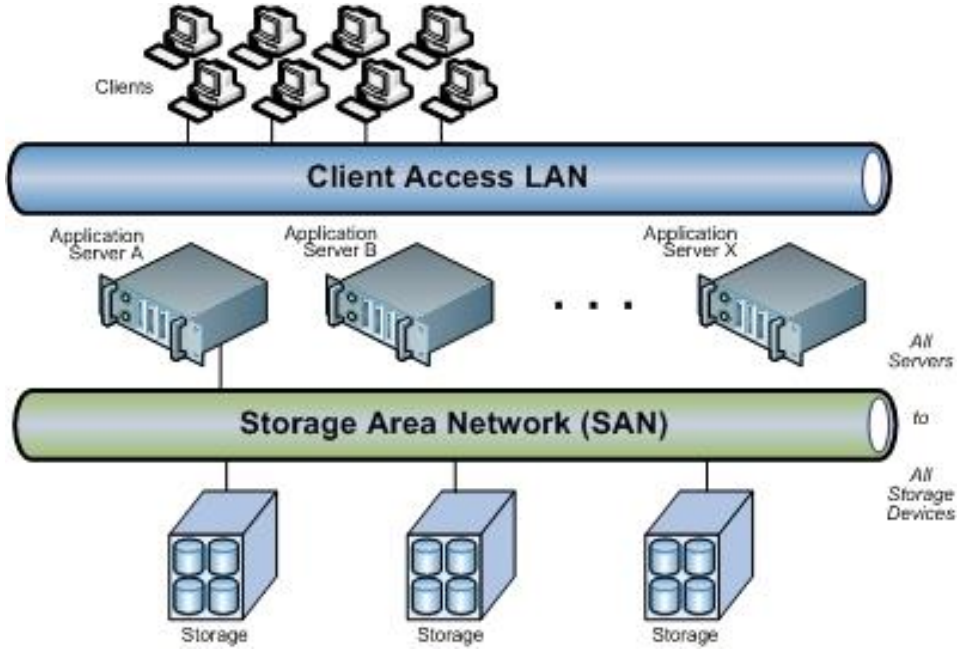
- Todos los datos almacenados en la memoria principal, no es necesario realizar I/O de disco para consultar o actualizar datos.
- Los datos son persistentes o volátiles según el producto de la base de datos en memoria.
- Las estructuras de datos especializadas y las estructuras de índice asumen que los datos siempre están en la memoria principal.
- Optimizado para alto rendimiento.

Ventajas y Desventajas

- No es necesario lidiar con el almacenamiento secundario, que puede ser un orden de magnitud más lento que acceder a los datos almacenados en la memoria principal.
- Utilización de un árbol AVL en lugar de un árbol b, lo que reduce (o elimina) la necesidad de duplicar datos, pero aumenta el número de filas a las que se accede durante el recorrido.
- La memoria no es persistente. Cuando una aplicación se cierra, ya sea limpia o inesperadamente, los datos almacenados en una base de datos en memoria se perderán. Ciertas aplicaciones no requieren persistencia de datos entre ejecuciones, pero otras sí. Esas aplicaciones que requieren un alto grado de persistencia pueden no ser adecuadas para usar una base de datos en memoria.
- Almacena todos los datos en la memoria principal, lo que puede limitar severamente la cantidad de datos que se pueden almacenar.

Almacenamiento

- ***Storage Area Network (SAN),*** es una red de almacenamiento integral conectada a las redes de comunicación de una compañía. Además de contar con interfaces de red tradicionales, los equipos con acceso a la SAN tienen una interfaz de red específica que se conecta a la SAN que agrupa los siguientes elementos:



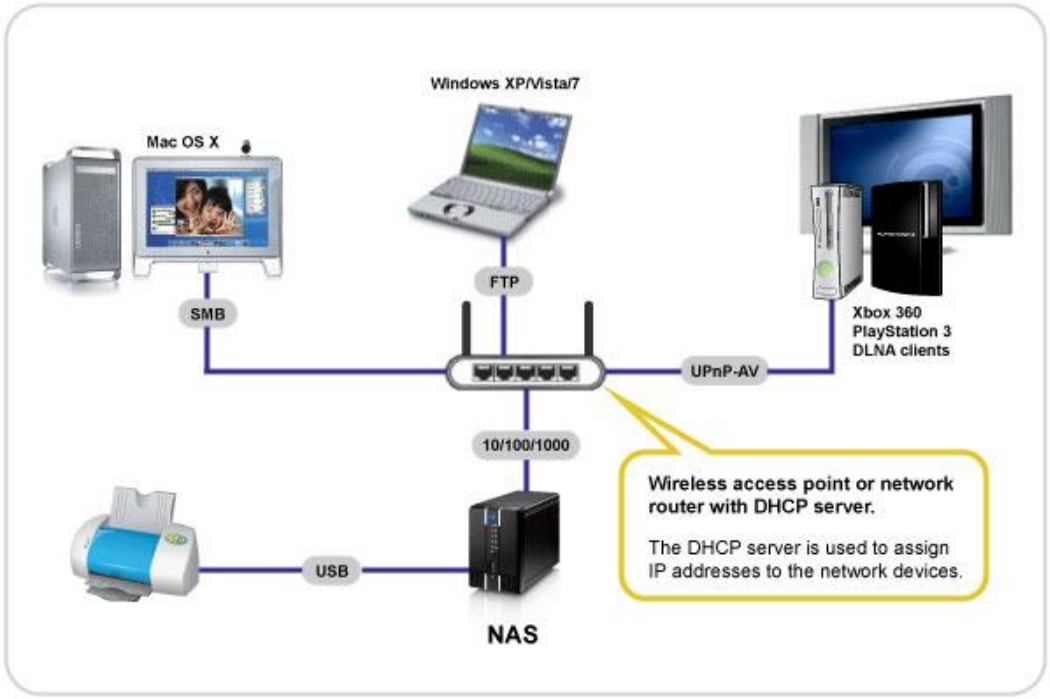
- Canal de fibra o iSCSI.
- Interconexión dedicado (Switches, routers, etc).
- Discos duros mecánicos, SSD e híbridos.

Beneficios Importantes

- Mejora la disponibilidad; múltiples rutas de datos
- Mejora el rendimiento de la aplicación
- Aumente la utilización y la eficacia del almacenamiento
- Mejora la protección y la seguridad de los datos.
- Diseñadas para el almacenamiento en bloque dentro de bases de datos, también conocidas como datos estructurados.

Almacenamiento

- ***Network Attached Storage (NAS),*** son dispositivos de almacenamiento a los que se accede desde los equipos a través de protocolos de red (normalmente TCP/IP). También se podría considerar un sistema NAS a un servidor (Microsoft Windows, Linux, etcétera) que comparte sus unidades por red, pero la definición suele aplicarse a sistemas específicos.



- Canal de fibra o iSCSI.
- Interconexión dedicado (Switches, routers, etc).
- Discos duros mecánicos, SSD e híbridos.

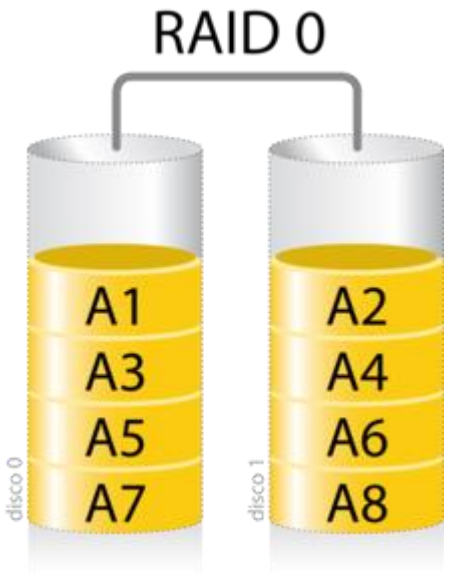
Beneficios Importantes

- Maneja datos no estructurados, como audio, video, sitios web, archivos de texto y documentos de Microsoft Office.
- Permite a los usuarios colaborar y compartir datos de manera más efectiva.
- Se conecta por wireless.

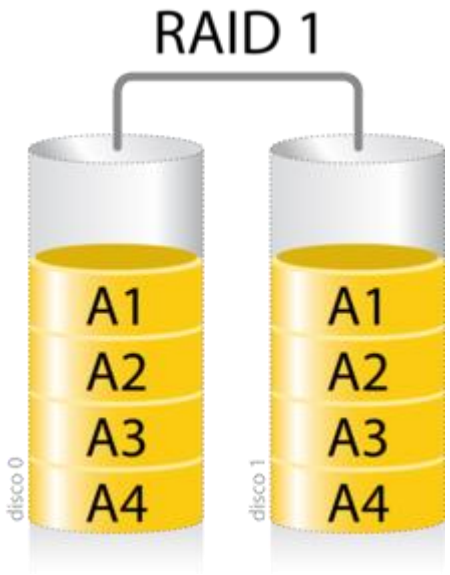
Almacenamiento

- Redundant Array of Independent Disks (RAID), hace referencia a un sistema de almacenamiento de datos que utiliza múltiples unidades (discos duros o SSD), entre las cuales se distribuyen o replican los datos. Dependiendo de su configuración (nivel), los beneficios de un RAID:
  - Mayor integridad
  - Tolerancia frente a fallos
  - Tasa de transferencia y capacidad
- En sus implementaciones originales, su ventaja clave era la habilidad de combinar varios dispositivos de bajo costo y tecnología más vieja en un conjunto que ofrecía mayor capacidad, fiabilidad, velocidad o una combinación de éstas que un solo dispositivo de última generación y costo más alto.

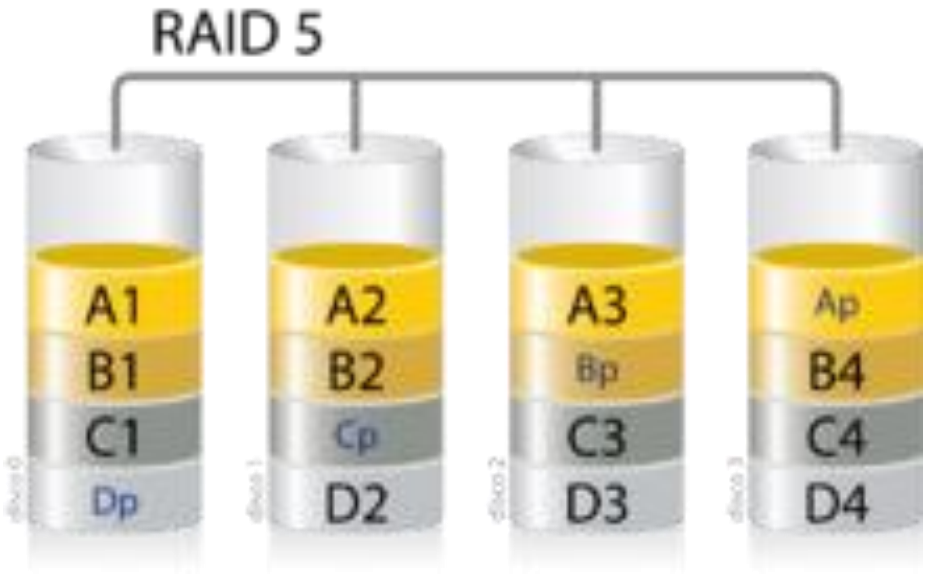
RAID



Distribuye los datos equitativamente entre dos o más discos



Crea una copia exacta (o espejo) de un conjunto de datos en dos o más discos



Un RAID 5 (también llamado distribuido con paridad) es una división de datos a nivel de bloques que distribuye la información de paridad entre todos los discos miembros del conjunto.

NoSQL

- Un NoSQL que originalmente se refería a no SQL o no relacional es una base de datos que proporciona un mecanismo para el almacenamiento y la recuperación de datos. Estos datos se modelan en medios distintos de las relaciones tabulares utilizadas en bases de datos relacionales, se utilizan en aplicaciones web en tiempo real y big data y su uso aumenta con el tiempo. Incluye simplicidad de diseño, escalamiento horizontal más simple a clusters y un control más fino sobre la disponibilidad. La capacidad de adaptación de una base de datos NoSQL depende del problema que debe resolver. Las estructuras de datos utilizadas por las bases de datos NoSQL a veces también se consideran más flexibles que las tablas de bases de datos relacionales.

NoSQL

- Comprometen la consistencia por la disponibilidad, velocidad y tolerancia de partición.
- Carecen de transacciones ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad).
- Ofrecen un concepto de consistencia eventual en el que los cambios en la base de datos se propagan a todos los nodos, por lo que las consultas de datos podrían no devolver datos actualizados inmediatamente o podrían dar como resultado la lectura de datos que no son precisos, lo cual es un problema conocido como lecturas obsoletas.
- Pueden exhibir escrituras perdidas y otras formas de pérdida de datos.
- Registro de escritura anticipada para evitar la pérdida de datos.

Ventajas y Desventajas

Ventajas

- *Escalabilidad*. Usa fragmentación para el escalado horizontal. Particionar los datos y colocarlos en varios servidores de tal manera que se mantenga el orden de los datos.
- *Disponibilidad*. La función de replicación automática lo hace altamente disponible porque en caso de falla, los datos se replican al estado consistente anterior.

Desventajas

- *Enfoque limitado*. Diseñado principalmente para el almacenamiento, proporciona muy poca funcionalidad.
- *Open Source*. No hay un estándar confiable
- *Administración*. Tiene una reputación de ser difícil de instalar y aún más para administrar a diario.
- *No GUI.* No están disponibles.
- Respaldos. No tiene un enfoque para la copia de seguridad de datos de manera consistente.
- *Documentos grandes*. Almacenan datos en formato JSON

- Tipos:
  - *Key Value Store*: Memcached, Redis, Coherence
  - *Tabular*: Hbase, Big Table, Accumulo
  - *Document Based*: MongoDB, CouchDB, Cloudant
  - *Graph*

- ¿Cuándo sar NoSQL?
  - Cuando se necesita almacenar y trabajar con una gran cantidad de datos.
  - La relación entre los datos que almacena no es tan importante
  - Los datos cambian con el tiempo y no están estructurados.
  - No se requiere soporte de restricciones y uniones a nivel de base de datos
  - Los datos crecen continuamente y necesita escalar la base de datos regularmente para manejar los datos.

SQL Vs NO SQL

- En una base de datos ***relacional***, un registro de libros a menudo se enmascara (o "normaliza") y se almacena en tablas separadas, y las relaciones se definen mediante restricciones de claves primarias y externas. En este ejemplo, ***la tabla Libros*** tiene las columnas ISBN, Título del libro y Número de edición, la ***tabla Autores*** tiene las columnas IDAutor y Nombre de autor y, finalmente, ***la tabla Autor-ISBN*** tiene las columnas IDAutor e ISBN. El modelo relacional está diseñado para permitir que la base de datos aplique la integridad referencial entre tablas en la base de datos, normalizada para reducir la redundancia y, generalmente, está optimizada para el almacenamiento.
- En una base de datos ***NoSQL***, el registro de un libro generalmente se almacena como un ***documento JSON***. Para cada libro, el elemento, ISBN, Título del libro, Número de edición, Nombre autor y IDAutor se almacenan como atributos en un solo documento. En este modelo, los datos están optimizados para un desarrollo intuitivo y escalabilidad horizontal.

Appliances All-In-One

- Es una combinación de software y hardware utilizados para almacenar y dar acceso a las bases de datos, brinda servicios de almacenamiento preparados para bases de datos, tales como la posibilidad de descargar el procesamiento del servidor de la base de datos al almacenamiento de forma transparente.
- Es una solución completa denominada end-to-end para bases de datos
- Una solución fácil de implementar y lista para usar que con los más altos niveles de rendimiento del mercado
- Escalabilidad lineal de I/O

Appliances All-In-One

Ventajas

- Se configuran asegurando costos mínimos, incorporando el mejor material y produciendo la mejor infraestructura de acuerdo con la demanda del negocio.
- Reduce en gran medida los problemas
- Optimización en costos
- Minimiza los recursos y el tiempo.

Desventajas

- “Todos los huevos en una sola canasta”
- Escalabilidad

Lenguaje R; Características

- También conocido como “GNU S”, es un entorno y un lenguaje de programación para el cálculo estadístico y la generación de gráficos.
- Dispone de almacenamiento y operadores para cálculo sobre arreglos y en particular matrices
- La capacidad de combinar, sin fisuras, análisis “preempaquetados” (por ejemplo, una regresión logística) con análisis ad-hoc, específicos para una situación.
- La capacidad de manipular y modificar datos y funciones.
- Los gráficos de alta calidad: visualización de datos y producción de gráficos.
- La comunidad de R es muy dinámica, con gran crecimiento del número de paquetes, e integrada por estadísticos de gran renombre.
- Hay extensiones específicas a nuevas áreas como bioinformática, geoestadística y modelos gráficos.
- Es un lenguaje orientado a objetos.
- Se parece a Matlab y Octave y su sintaxis recuerda a C/C++.
- Es gratuito y su descarga e instalación son sencillas

Lenguaje R; Variables

Todas las cosas que manipula R se llaman objetos. En general, éstos se construyen a partir de objetos más simples. De esta manera, se llega a los objetos más simples que son de cinco clases a las que se denomina atómicas:

- character (cadenas de caracteres)
- numeric (números reales)
- integer (números enteros)
- complex (números complejos)
- logical (lógicos o booleanos, que sólo toman los valores True o False)

En el lenguaje, sin embargo, cada uno de estas clases de datos no se encuentran ni se manejan de manera aislada, sino ***encapsulados*** dentro de la clase de objeto más básica del lenguaje: ***el vector***. Un vector puede contener cero o más objetos, pero todos de la misma clase. En contraste, la clase denominada list, permite componer objetos también como una secuencia de otros objetos, pero, a diferencia del vector, cada uno de sus componentes puede ser de una clase distinta.