DESDUPLICACION

¿Qué es la desduplicación?

Se trata de una técnica de optimización del almacenamiento de datos que libera recursos al eliminar segmentos redundantes dentro de un conjunto de datos.

Mediante un proceso de escaneo, se identifican y almacenan solo una vez las partes duplicadas, con la opción de aplicar compresión para un mayor ahorro. Esta estrategia reduce la redundancia sin comprometer la integridad ni la fidelidad de la información.

¿Cuál es su importancia?

Las grandes empresas manejan enormes volúmenes de datos de distintos tipos a través de diversas operaciones diarias, lo que inevitablemente genera acumulación de datos duplicados en los sistemas de almacenamiento. Esto impacta la necesidad de mantener redundancia intencional para la recuperación ante desastres, la alta disponibilidad y la protección de datos.

Al eliminar los datos duplicados, se libera espacio sin necesidad de adquirir más almacenamiento para el crecimiento continuo de la información. Además, esto agiliza las copias de seguridad, reduciendo el tiempo y los recursos informáticos necesarios. Como resultado, el acceso a los datos es más rápido y fiable, minimizando errores derivados de duplicados o conflictos.

Funcionamiento

- *A NIVEL ARCHIVO*

La función analiza una lista de archivos para identificar duplicados, especialmente aquellos que comparten un origen común. Garantiza que cada archivo único se almacene solo una vez y, opcionalmente, genera referencias a la copia original en lugar de conservar múltiples versiones idénticas.

- *A NIVEL BLOQUE*

La función normalmente escanea el sistema de archivos y divide la información en segmentos de datos fijos llamados bloques. Estos bloques se organizan en tablas

hash, que a su vez se almacenan en una base de datos para ser buscados posteriormente a través de sus índices generados. De este modo, cuando un archivo se actualiza, en lugar de crear una nueva versión completa, solo se guardan los bloques modificados, lo que la convierte en una técnica más eficiente que la deduplicación a nivel de archivo.

Al igual que en el enfoque por archivo, también se pueden generar referencias a la copia original. Sin embargo, esta técnica presenta como desventaja un mayor requerimiento de potencia de procesamiento.

Implementación

- *In line.*

Todos los datos nuevos son analizados en tiempo real, es decir, durante el proceso de escritura. Antes de que cualquier dato se guarde en el almacenamiento, el sistema verifica si ya existe una copia idéntica. Si encuentra duplicados, en lugar de escribir el nuevo dato, simplemente crea una referencia al original. Esto permite ahorrar espacio desde el primer momento, evitando que los datos redundantes lleguen siquiera a ocupar espacio físico.

- Out line.

Los datos son almacenados primero en el sistema sin ser analizados. Luego, posteriormente, el sistema revisa la información en busca de duplicados y elimina las copias redundantes, reemplazándolas por referencias a la versión original.

Permitiendo que la escritura de datos sea más rápida, ya que no se realiza un análisis inmediato. Lo que requiere una fase adicional de procesamiento para optimizar el almacenamiento.

¿Qué es la compresión?

A menudo, la desduplicación se combina con la compresión para optimizar aún más el almacenamiento. Mientras que la desduplicación elimina datos redundantes almacenando solo una copia de cada segmento único, la compresión reduce el tamaño de esos segmentos mediante algoritmos que minimizan la cantidad de bits necesarios para representar la información.

Al trabajar en conjunto, ambas técnicas maximizan el ahorro de espacio y mejoran la eficiencia del almacenamiento sin comprometer la integridad de los datos.

EJEMPLOS:

- *BTRFS*

El sistema de archivos BTRFS utiliza el método de gestión de recursos conocido como Copy-on-Write (CoW), el cual permite mantener múltiples versiones de un archivo sin reemplazar la versión original. En lugar de sobrescribir un archivo al modificarlo, el sistema solo captura las partes modificadas, creando una nueva versión del archivo. Esto se logra mediante una técnica de instantáneas, donde se guarda una copia parcial de los cambios en lugar de duplicar todo el archivo, lo que optimiza el uso del espacio y evita desorden.

Una de las principales ventajas de este sistema es la deduplicación, que busca y elimina bloques de datos repetidos entre archivos. Al hacerlo, el sistema mantiene solo una copia de los bloques redundantes y combina las partes restantes como clones de otros archivos. Esto no solo ahorra espacio, sino que también mejora la eficiencia del sistema.

BTRFS ofrece dos herramientas específicas para la deduplicación, dependiendo de las necesidades del usuario:

- duperemove: Se utiliza para la deduplicación a nivel de archivo, permitiendo eliminar duplicados entre archivos completos.
- BEES: Se emplea para la deduplicación a nivel de bloques, lo que permite gestionar los datos a un nivel más bajo y directo sobre el almacenamiento.

Además de la deduplicación, BTRFS también soporta técnicas de compresión. Sin embargo, su efectividad depende del tipo de datos. Los archivos que ya están comprimidos, como los videos o imágenes, no se benefician mucho de una compresión adicional. Por otro lado, los archivos de texto sin formato pueden reducirse significativamente en tamaño. Por lo tanto, es importante evaluar la naturaleza de los datos antes de aplicar técnicas de compresión para asegurar que se logren los beneficios esperados en términos de ahorro de espacio.

En resumen, BTRFS, con su implementación de CoW, deduplicación y compresión, ofrece una gestión de archivos avanzada que optimiza el uso del espacio en disco y facilita la administración de versiones y cambios en los archivos, especialmente en entornos con grandes volúmenes de datos.

- *NTFS*

New Technology File System (NTFS) es un sistema de archivos de diario patentado y desarrollado por Microsoft, diseñado para optimizar el almacenamiento de datos. Su enfoque se basa en el modelo de escritura offline, en el que los datos se

escriben inicialmente sin optimizar en el disco. Posteriormente, NTFS realiza un proceso de deduplicación y optimización, asegurándose de que, cuando el usuario acceda a los archivos, estos ya estén desduplicados y optimizados para un mejor rendimiento.

El sistema escanea los archivos en el sistema, los divide en un número aleatorio de *clusters*, identifica los bloques de datos únicos y los almacena de manera eficiente. Además, puede aplicar técnicas de compresión a los datos, si es necesario, y finalmente reemplaza el flujo original del archivo con los datos optimizados almacenados.

NTFS emplea la Master File Table (MFT), que mantiene un registro de todos los archivos y sus ubicaciones en el disco. Un archivo puede ocupar múltiples clusters dispersos en el disco, y NTFS utiliza punteros para rastrear estos bloques de datos de manera eficiente.

Sin embargo, su principal desventaja es que solo se puede utilizar en sistemas operativos Windows, lo que limita su compatibilidad en entornos multisoporte. Además, no se puede escribir directamente sobre archivos NTFS en sistemas que no sean Windows, lo que restringe su flexibilidad para usuarios de otros sistemas operativos, como Linux o macOS.

- *ZFS*

ZFS es un sistema de archivos avanzado y transaccional, diseñado para ofrecer alto rendimiento y fiabilidad. Su arquitectura se basa en características innovadoras como la semántica Copy-on-Write (CoW)**, la** deduplicación de datos**, la** integridad de los datos **y una** gestión avanzada del almacenamiento. Estas características permiten que ZFS se destaque en la gestión eficiente y segura de grandes volúmenes de datos.

Uno de los aspectos clave de ZFS es su capacidad para realizar deduplicación de datos, lo que elimina las copias redundantes en el almacenamiento. Utiliza hashes criptográficos para identificar bloques de datos repetidos. Cada bloque de datos tiene un hash único, y cuando ZFS detecta que un nuevo bloque tiene el mismo hash que otro ya almacenado, en lugar de almacenar una copia duplicada, crea una referencia al bloque original. Esto optimiza el uso del espacio de almacenamiento y mejora la eficiencia en sistemas con grandes cantidades de datos redundantes.

En resumen, ZFS es un sistema de archivos robusto y flexible que ofrece una amplia gama de características avanzadas para la gestión de datos. Su capacidad para garantizar la integridad de los datos, su deduplicación eficiente, y su enfoque de escritura transaccional lo convierten en una opción ideal para entornos que

requieren alto rendimiento **y** alta disponibilidad. Sin embargo, sus requisitos de hardware específicos y su compatibilidad limitada con algunos sistemas operativos pueden ser factores que considerar antes de adoptarlo en ciertos entornos.

FUENTES

- Javier, F. (2009). Técnicas de deduplicación de datos y aplicación en librerías virtuales de cintas | Archivo Digital UPM. *Oa.upm.es*. https://oa.upm.es/1803/
- *Desduplication - ES*. (2024, December 19). ISID. https://isid.com/es/tecnologia/desduplication-es/
- wmgries. (2024, November 2). *Introducción a la desduplicación de datos*. Microsoft.com. https://learn.microsoft.com/es-es/windows-server/storage/data-deduplication/overview#what-is-dedup
- Content Studio. (2023, July 10). *¿Qué es la desduplicación de datos?* Purestorage.com; Pure Storage. https://www.purestorage.com/la/knowledge/what-is-data-deduplication.html
- *Deduplication — BTRFS documentation*. (2025). Readthedocs.io. https://btrfs.readthedocs.io/en/latest/Deduplication.html
- *ZFS Deduplication*. (2025, February 12). Truenas.com. https://www.truenas.com/docs/references/zfsdeduplication/
- (2025). Oracle.com. https://docs.oracle.com/cd/E19253-01/820-2314/zfsover-2/index.html
- Freire, A. (2022, June 8). *Qué es NTFTS - El blog de dinahosting*. El Blog de Dinahosting. https://dinahosting.com/blog/que-es-ntfs/#Ventajas_e_inconvenientes_de_NTFS
- wmgries. (2022, February 18). *Understanding Data Deduplication*. Microsoft.com. https://learn.microsoft.com/en-us/windows-server/storage/data-deduplication/understand
- Recoverit. (2023, June 29). *Btrfs Sistema de archivos Btrfs: Definición, Características, Ventajas*. Wondershare.es; Wondershare Recoverit. https://recoverit.wondershare.es/file-system/what-is-btrfs-file-system.html
- TheDatahoarder. (2021, August 7). *Putting BTRFS compression and deduplication to the test - DataHoards*. DataHoards. https://www.datahoards.com/putting-btrfs-compression-and-deduplication-to-the-test/