UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Sistemas Operativos
Deduplicación de sistemas operativos modernos

EXPOSITORES

Hernández
Ramírez Miguel
Ángel
Adrian Alejandro
Pacheco Pacheco

PROFESOR

Ing. Gunnar Eyal Wolf
Iszaevich

INTRODUCCION

Para comenzar esta investigación debemos de plantear como concepto fundamental que los sistemas operativos son conjuntos de programas que gestionan los recursos de hardware de un sistema informático y proveen servicios a los programas de aplicación. Su función principal es administrar los recursos del ordenador, coordinar el hardware y organizar archivos y directorios en los dispositivos de almacenamiento de nuestro ordenador.

Estos servicios del sistema operativo consisten en la ejecución de de programas o aplicaciones que funcionan en segundo plano y ofrecen características específicas. Estos servicios son cargados por el propio sistema operativo y pueden cambiar según las necesidades del usuario. Algunas de las funciones de los servicios del sistema operativo incluyen:

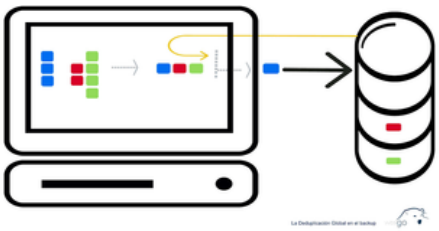
- Gestión de procesos: Controlan la ejecución de programas y procesos en el sistema.
- Gestión de memoria: Administran la asignación y liberación de memoria para los programas y procesos en ejecución.
- Gestión de entrada/salida (E/S): Controlan la comunicación entre el hardware y el software para facilitar la interacción con dispositivos periféricos.
- Gestión del sistema de archivos: Administran la organización, acceso y manipulación de archivos en el sistema de almacenamiento.

Ahora bien, conociendo que son los sistemas operativos podemos centrarnos en qué es la deduplicación

En esta investigación entenderemos la deduplicación como un proceso utilizado en la gestión de datos para identificar y eliminar duplicados o epeticiones innecesarias dentro de un conjunto de datos. Este procedimiento es fundamental para mejorar la eficiencia en el almacenamiento de información, reducir costos y evitar confusiones causadas por datos duplicados. La deduplicación puede realizarse a nivel de archivos, bases de datos o incluso en sistemas completos, ayudando a optimizar el rendimiento y la organización de la información.

La deduplicación de datos es una técnica especializada que se puede entender que reemplaza las copias adicionales con metadatos que apuntan al original, reduciendo así las necesidades de almacenamiento y mejorando la eficiencia del ancho de banda al guardar y transmitir solo los datos útiles y necesarios para la empresa

Esto se puede lograr mediante el uso de herramientas especializadas que identifican y eliminan las copias idénticas o similares de los sistemas operativos.

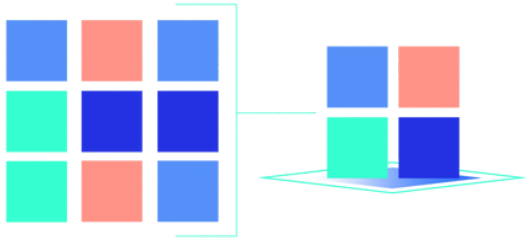


¿De dónde surge?

Cómo todo lo que conocemosl a necesidad de deduplicación en sistemas operativos modernos surge por la existencia de múltiples copias de datos similares en diferentes fuentes, como sistemas operativos, máquinas virtuales y aplicaciones. Estas duplicaciones pueden generar una sobrecarga en el almacenamiento y la red, lo que resulta en un uso ineficiente de recursos. La deduplicación ayuda a reducir los costos de almacenamiento y el uso de ancho de banda al eliminar bloques de datos duplicados durante las copias de seguridad y la transferencia de datos.

Además, optimiza la red WAN y evita la necesidad de invertir en software de deduplicación de datos. Al eliminar las redundancias, la deduplicación no solo reduce los costos operativos, sino que también mejora la eficiencia de la gestión de datos al simplificar la administración de sistemas y la realización de copias de seguridad, lo que beneficia tanto a las organizaciones como a los usuarios finales.

Técnicas de implementación



Las técnicas de deduplicación en sistemas operativos modernos incluyen:

Deduplicación a nivel de archivo: Esta técnica elimina duplicados de archivos completos comparando las firmas digitales o hashes de cada archivo y eliminando aquellos que sean idénticos. Es comúnmente utilizada en sistemas de almacenamiento para ahorrar espacio de manera efectiva

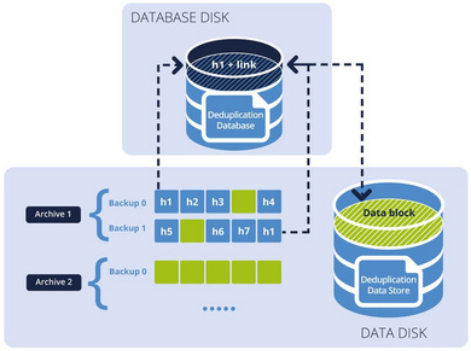
Deduplicación a nivel de bloques: Esta técnica divide los archivos en bloques más pequeños y analiza cada uno de ellos para identificar y eliminar duplicados. Esto permite un ahorro de espacio y una optimización del ancho de banda en la transferencia de datos

Deduplicación posterior al proceso: Este método permite a las empresas utilizar su servicio de reducción de datos sin preocuparse por la sobrecarga de capacidad de almacenamiento, ya que todos los datos se almacenan en su forma completa hasta que se realiza el proceso de deduplicación programado

Deduplicación del lado del origen: Se refiere al proceso de deduplicación que se desarrolla en el origen de los datos, eliminando redundancias antes de que los datos sean transferidos o almacenados

Deduplicación de lado de destino: El proceso de deduplicación se ejecuta en el espacio de almacenamiento de destino, lo que ayuda a optimizar el almacenamiento y la transferencia de datos
Deduplicación basada en hardware frente a deduplicación basada en software: Los dispositivos de deduplicación construidos funcionalmente reducen la carga de procesamiento asociada con los productos de software, mientras que los métodos basados en hardware dan prioridad a la reducción de datos en el nivel de almacenamiento

Deduplicación en línea: Este enfoque implica la eliminación de datos duplicados en tiempo real, a medida que los datos se escriben en el sistema de almacenamiento. Cuando se escribe un nuevo conjunto de datos, el sistema verifica si existe una copia idéntica o similar en el almacenamiento y, si es así, solo se almacena una referencia al dato existente en lugar de duplicarlo. La deduplicación en línea puede requerir un mayor poder de procesamiento y recursos de almacenamiento, ya que implica una verificación constante de duplicados durante las operaciones de escritura.
Deduplicación fuera de línea: En contraste, la deduplicación fuera de línea implica la eliminación de datos duplicados después de que se hayan almacenado en el sistema de almacenamiento. Es decir, los datos se almacenan inicialmente sin ninguna consideración de duplicación, pero luego se ejecuta un proceso periódico o programado para identificar y eliminar los duplicados. Este proceso generalmente implica escanear el almacenamiento en busca de duplicados y luego eliminar las copias redundantes. La deduplicación fuera de línea puede ser menos intensiva en recursos que la deduplicación en línea, ya que no requiere verificaciones en tiempo real durante las operaciones de escritura.



ADMINISTRACION

La deduplicación de datos en sistemas operativos modernos se administra a través de tecnologías que eliminan copias redundantes de datos, reduciendo significativamente los requisitos de capacidad de almacenamiento. Esta técnica puede ejecutarse como un proceso en línea mientras los datos se escriben en el sistema de almacenamiento o como un proceso en segundo plano para eliminar duplicados después de que los datos se han escrito en el disco

EJEMPLO:

. En sistemas como NetApp, la deduplicación se implementa como un proceso de cero pérdida de datos que se ejecuta tanto en línea como en segundo plano para maximizar el ahorro. Se ejecuta oportunamente en línea para no interferir con las operaciones del cliente y de manera integral en segundo plano para optimizar los beneficios

Además, la deduplicación suele estar activada por defecto y se ejecuta automáticamente en todos los volúmenes y agregados sin necesidad de intervención manual, con una sobrecarga mínima en el rendimiento debido a su ejecución en un dominio de eficiencia dedicado

Aunque existen varios metodos de implementacion de la deduplicacion, todos tienen algunos pasos en comun, los cuales son

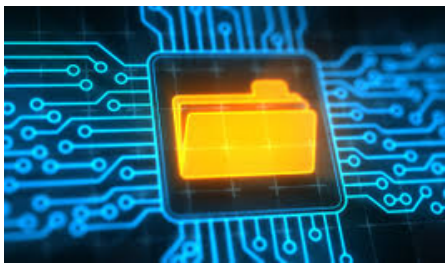
1. **Identificación de datos duplicados:** El sistema escanea y analiza los datos almacenados en el disco para identificar patrones de datos duplicados o similares.
2. **Indexación y almacenamiento de metadatos:** Se crea un índice o una tabla de metadatos que mapea los bloques de datos duplicados y sus ubicaciones en el disco.
3. **Eliminación de datos redundantes:** Los datos duplicados se eliminan físicamente del disco, y en su lugar se establecen referencias o enlaces a los bloques de datos únicos.
4. **Gestión de la integridad de los datos:** Se implementan mecanismos para garantizar la integridad de los datos, como checksums o firmas digitales, para verificar la integridad de los bloques de datos antes y después de la deduplicación.
5. **Optimización del rendimiento:** Se realizan ajustes para optimizar el rendimiento del sistema de archivos y minimizar el impacto en las operaciones de lectura y escritura, como el almacenamiento en caché de datos deduplicados o la paralelización de operaciones de deduplicación.

IMPLEMENTACION

Para la implementación de la deduplicación de datos, es un proceso realizado en bases de datos, sistema de gestión de datos y en almacenamiento de archivos. Existen varias técnicas para hacerlo y algoritmos para implementar la deduplicación de datos y esta dependerá de las necesidades especificas requeridas para el sistema.
Cuando se trata de una deduplicación a nivel de bloque de datos, el primer paso que se realiza en este proceso es dividir dichos datos en porciones mas pequeñas, comúnmente se especifica un tamaño fijo pudiendo ser 4KB o 8Kb según sea el caso. Y de esta manera podríamos observar de forma mas rápida y detallada. Posteriormente, se calcula un Hash criptográfico para cada bloque, su utilidad consiste en funcionar como identificador único para dicho bloque. Estos Hash anteriores y la información sobre donde se ubican los bloques, se van a almacenar en un indice presente en una tabla de búsqueda. Algunos ejemplo de estos podrían ser (MD5, SHA-1, etc...)
En el caso en el que se ingresa un bloque de datos nuevo, se realiza el proceso de calcular su Hash y como se menciono anteriormente se buscara a traves de su indice en una tabla de hashes. En caso de que ya haya un bloque en existencia, significara que se trata de un duplicado. Lo que quedara por realizar es que se cambiara dicho bloque duplicado con un enlace al bloque inicial o original, y así se liberara un poco de espacio en el disco. Algo muy importante a mencionar, es que se debe de realizar un manejo de las colisiones Hash que se trata del caso cuando dos bloques diferentes generan el mismo Hash.

Cuando se trata de una implementación a nivel archivo, este proceso puede variar un poco. Lo que se realizara es calcular el Hash para todo un mismo archivo y posteriormente se comparar con un conjunto de datos de Hashes de un archivo que han sido ingresados con anterioridad. Si un Hash coincide, por obvias razones se identificara como un archivo duplicado.
Para este caso como el anterior se debe de manejar las colisiones de manera correcta, y también ser consientes de que la deduplicación puede llegar a consumir recursos de la CPU, esto debido a que implica calcular y comparar Hashes.

Ademas de estos existen ya algoritmos mas actuales y avanzados que pueden adaptarse a distintas aplicaciones y requisitos de rendimiento. Un caso de funcionamiento de dichos algoritmos podría ser que pueden optimizar la dirección de duplicados, haciendo mas eficiente y rápido el proceso. Usar estos algoritmos dependerá de distintos factores como el volumen de datos a manejar, los recursos de hardware y software que se tengan en el sistema y entre otros.



Riesgos

En la deduplicación de datos pueden haber complicaciones y riesgos tales como los siguientes:

- Colisiones de Hash: Existe la posibilidad de que dos bloques diferentes generen el mismo Hash. En caso de colisiones, Debería de existir un mecanismo fuerte para manejar estas situaciones y evitar corrupción de datos.
- Recursos: En este proceso se podría requerir en veces una cantidad grande de almacenamiento y hardware. Estos sistema tienen que tener la capacidad para manejar eficientemente la deduplicación sin afectar al sistema.
- Complejidad: Cuando se trata de entornos distribuidos, que significara que se almacenan datos en muchas ubicaciones, provocaría que la deduplicación pueda volverse mas compleja.
- Seguridad: Cuando se trabaja en entornos donde la seguridad es primordial, la duplicación podría generar complicaciones. La accesibilidad de estos a traves de Hashes podría ser un riesgo a tomar en cuenta.
- Transacción de datos: Los datos son propensos al cambio, lo que podría provocar que la deduplicación se torne difícil en ambientes donde exista muchos cambios. Y mantener la integridad de los datos puede ser de mayor complejidad.
- Degradación del rendimiento: La deduplicación de datos puede aumentar la carga de trabajo en los sistemas de almacenamiento y procesamiento, lo que puede resultar en una degradación del rendimiento del sistema si no se dimensiona adecuadamente.
- Pérdida de datos: Existe el riesgo de pérdida de datos durante el proceso de deduplicación, especialmente si no se implementan adecuadamente mecanismos de respaldo y recuperación. La pérdida accidental de datos puede tener consecuencias graves para las organizaciones.

Beneficios

En la deduplicación de datos pueden haber también beneficios y ventajas tales como los siguientes:

- Gestión de datos optima: Al quitar los datos duplicados, la deduplicación puede ayudar a facilitar la gestión y organización de grandes conjuntos de datos.
- Copia de seguridad y recuperación: La deduplicación de datos nos podría ayudar a acelerar los procesos de copia de seguridad y recuperación esto debido al reducir el tiempo y los recursos necesarios para transferir y almacenar datos.
- Mejora del rendimiento: La deduplicación de datos en algunas ocasiones puede ayudar a mejorar el rendimiento de las operaciones de E/S debido a que se reduce la cantidad de datos que se transfieren y procesan.
- Reducción de costos: Como estamos disminuyendo la cantidad de almacenamiento requerido, la deduplicación de datos puede ayudarnos a bajar los costos asociados con la adquisición y gestión de almacenamiento.
- Ahorro de espacio de almacenamiento: La deduplicación de datos elimina copias redundantes, lo que nos ayuda a dar un uso más eficiente del espacio de almacenamiento.

Sistemas Operativos modernos

Hoy en día existen sistemas operativos con los cuales trabajamos o realizamos tareas con extensiones de estos mismos o variantes, cada sistema operativo teniendo un enfoque diferente a las necesidades para implementar esta técnica basándose en necesidades o funcionalidades particulares. siendo los mas conocidos los siguientes:

- Linux: La deduplicación de datos en Linux es asequible y requiere hardware menor. Las soluciones están disponibles en algunos casos a nivel de bloque, y son capaces de funcionar sólo con flujos de datos redundantes de bloques de datos en lugar de archivos individuales, porque la lógica es incapaz de reconocer archivos separados sobre muchos protocolos como SCSI, SAS Fibre canal e incluso SATA. El sistema de archivos que discutimos aquí es Lessfs-a block level deduplication y sistema de archivos Linux habilitado para FUSE. FUSE es un módulo de kernel visto en sistemas operativos similares a UNIX, que proporciona la capacidad de los usuarios de crear sus propios sistemas de archivos sin tocar el código del kernel. Para utilizar estos sistemas de archivos, FUSE debe estar instalado en el sistema. La mayoría de los sistemas operativos como Ubuntu y Fedora tienen el módulo pre-instalado para soportar el sistema de archivos ntfs-3g.

- Windows: Para este sistema operativo ofrece un servicio integrado Windows Server para esta implementación, el proceso de deduplicación se realiza a nivel bloque de volumen de almacenamiento. Y para llevarla acabo habilitando la función en el servidor. Este proceso se puede programar en un horario concurrente, definir lo volúmenes y monitorizar los resultados. Básicamente lo que hace Windows es identificar el bloque de datos duplicado dentro del volumen y almacena una sola copia.

- MacOs: En macOS, la gestión de la deduplicación se realiza principalmente mediante utilidades de respaldo, como Time Machine. Estas herramientas tienen la finalidad de detectar cambios en los datos y almacenarlos de manera eficiente, evitando redundancias innecesarias. Time Machine en macOS emplea técnicas de deduplicación a nivel de bloques, lo que implica que durante la realización de copias de seguridad, no replica los bloques de datos que permanecen inalterados desde la última copia. En lugar de ello, guarda referencias a estos bloques, optimizando así el espacio de almacenamiento. La configuración de Time Machine se puede ajustar a través de las Preferencias del Sistema en macOS.

Para sistemas operativos menos comunes o de menor uso, la implementación de la deduplicación puede variar considerablemente y dependerá en gran medida de la disponibilidad de herramientas y tecnologías específicas. En sistemas embebidos, la deduplicación podría depender de la implementación del sistema de archivos en cuestión, podrían no tener soporte por deficiencia de recursos. Por otro lado en entornos mas especializados, la deduplicación puede ser administrada por herramientas del proveedor.

CONCLUSIONES

Como conclusión la investigación que se realizó podemos determinar que la de deduplicación de sistemas operativos modernos se volvió una necesidad a partir del exceso de información almacenado en archivos por lo que la comunidad decidió implementar este método para eliminar redundancias y optimizar el rendimiento de los sistemas con un esfuerzo mínimo en el llenado de sus archivos, la duplicación de datos significará una herramienta fundamental en un futuro muy cercano ya que representará una necesidad fundamental en la creación de cualquier base o sistema que administre cualquier tipo de concepto que se quiera ejecutar

REFERENCIAS

Technologies, V. (s. f.). The Complete Data Deduplication Guide and Why It is Important. https://www.veritas.com/es/mx/information-center/data-deduplication

Deduplicación de copias de seguridad – Acronis. (2019, 14 abril). Acronis. https://www.acronis.com/es-es/blog/posts/deduplication

Ciberseg. (2022, 21 abril). Deduplicación de datos: qué es y cómo funciona. Ciberseguridad. https://ciberseguridad.com/guias/recursos/deduplicacion-datos/

¿Qué es la deduplicación de datos? | Glosario. (s. f.). HPE España. https://www.hpe.com/es/es/what-is/data-deduplication.html

"Data Deduplication for Data Optimization for Storage and Network Systems" por P. M. Chen y E. K. Park.

"A Survey of Data Deduplication Techniques" por R. K. Ghosh y S. J. Patel.

Wmgries. (2023, November 4). Información acerca de Desduplicación de datos. Microsoft Learn. https://learn.microsoft.com/es-es/windows-server/storage/data-deduplication/understand

Anantha Krishnan P T. (2019, febrero 15). Data deduplication with a Linux based file system. Open Source For You. https://www.opensourceforu.com/2019/02/data-deduplication-with-a-linux-based-file-system

Explainer: deduplication. (2021, mayo 29). The Eclectic Light Company. https://eclecticlight.co/2021/05/29/explainer-deduplication/