Desduplicación.

Integrantes:
Cruz Martínez Raúl.
Villanueva Corona Miguel Angel.

IDEA GENERAL

- Mecanismo de ahorro de almacenamiento.
- Se crea a finales de los años ochenta.
- Poca explotación del mecanismo.
- Utilización en servidores.

IDEA GENERAL

La desduplicación consiste en el almacenamiento de una sola copia de archivos, para evitar su repetición.

- Comparación de secuencias de datos.
- En caso de tener una secuencia duplicada, se coloca en su lugar un apuntador a la primer secuencia almacenada.

RATIOS

Existe una forma de medir el ahorro de almacenamiento. Resulta siendo una muy buena ayuda visual para apreciarlo.

- Se expresa a través de una relación.
- Cantidad de datos totales sin desduplicar.
- Cantidad de datos desduplicados

RATIOS

Expresión para la obtención de los ratios de desduplicación:

$$Ratio = \frac{Bytes\ de\ entrada}{Bytes\ de\ salida}$$

RATIOS

En caso de querer expresar esta cantidad como un porcentaje, podremos usar:

$$Ratio\ como\ Porcentaje = \left[1 - \left(\frac{1}{Ratio}\right)\right] \times 100\%$$

RATIOS

Existe otra forma de expresar a los Ratios, y esta es muy similar al cociente convencional:

$$n : m$$

Donde:

n = cantidad de datos duplicados

m = cantidad de datos desduplicados

RATIOS

Ejemplificación del uso de los Ratios.

Si tenemos 500 datos totales sin duplicar.

Pero estos pueden referirse únicamente a un archivo almacenado, se representa como:

$$Ratio = \frac{500}{1}$$

Expresado de otra forma:

$$500 : 1$$

RATIOS

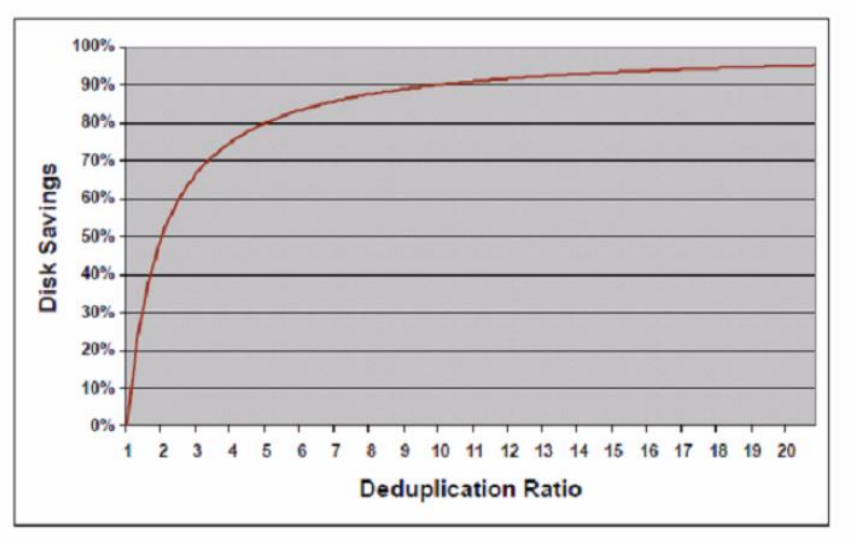
Si queremos expresar el ejemplo anterior como porcentaje:

$$Ratio\ como\ Porcentaje = \left[1 - \left(\frac{1}{\frac{500}{1}}\right)\right] \times 100\%$$

$$Ratio\ como\ Porcentaje = 99.8\%$$

RATIOS

La optimización del almacenamiento, se ve reducida después de cierta cantidad de ratios:



Dutch. M. (2008). Understanding data deduplication ratios

HASHING CRIPTOGRÁFICO.

¿Cómo sabremos cuál secuencia ya se encuentra dentro del disco?

Se utilizan algoritmos Hash, para asociar las secuencias de datos a llaves.

- Cada secuencia de datos genera una llave única que servirá como identificador de la secuencia.
- Si comparamos una llave generada con las existentes en el disco y son iguales, significa que esa secuencia de datos ya existe.

HASHING CRIPTOGRÁFICO.

¿Por qué utilizar Hash Criptográfico?

- Genera llaves con poca probabilidad de colisión.
- Permiten manejar grandes cantidades de información.
- Evitamos que existan errores al seleccionar cuáles secuencias se encuentran duplicadas.

HASHING CRIPTOGRÁFICO.

Algoritmos utilizados para el Hash Criptográfico:

- MD5
- SHA-1

La selección de cada uso dependerá de características específicas del sistema. Un ejemplo es la longitud de las secuencias.

HASHING CRIPTOGRÁFICO.

Longitud de las secuencias.

- Cortas
- Largas
- Longitud variables

HASHING CRIPTOGRÁFICO.

Algoritmo MD5.

- El algoritmo permite la entrada de secuencias de longitud variable.
- Entrega una llave de 128 bits (16 bytes)
- Se representa con 32 caracteres hexadecimales

Ejemplo: 5df9f63916ebf8528697b629022993e8

HASHING CRIPTOGRÁFICO.

Algoritmo SHA-1.

- El algoritmo permite la entrada de secuencias de longitud variable.
- Entrada máxima de 2^{64} bits
- Entrega una llave de 160 bits (20 bytes)
- Se representa con 40 caracteres hexadecimales
- Ejemplo: 110518500fa165c1859df82d3e32c8a127f93c1f

HASHING CRIPTOGRÁFICO.

¿Qué pasa cuando tenemos secuencias sumamente largas?

- Debe ser procesada en memoria para conseguir la mayor velocidad.
- Debe almacenarse parte de la secuencia en disco, lo que hará más lento el procesamiento de la llave.

GRANULARIDAD

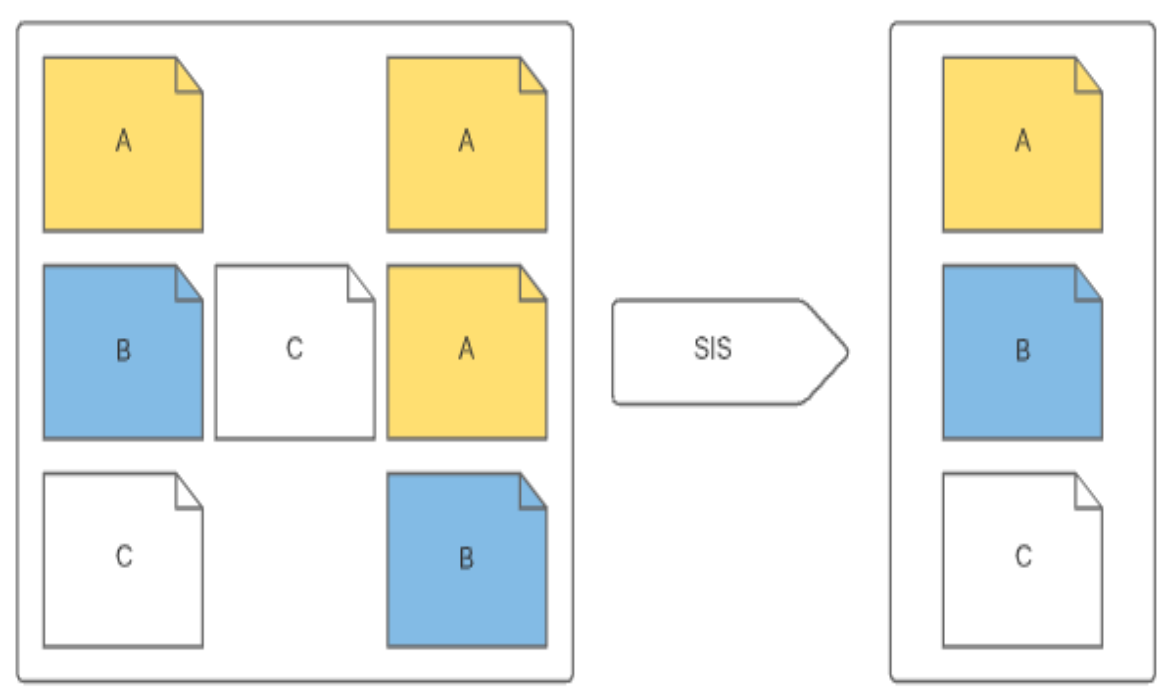
La granularidad es de mucha importancia para la desduplicación y esta tiene tres tipos los cuales son:

-A nivel de archivo

-A nivel de bloque

-A nivel de byte

GRANULARIDAD A NIVEL DE ARCHIVOS



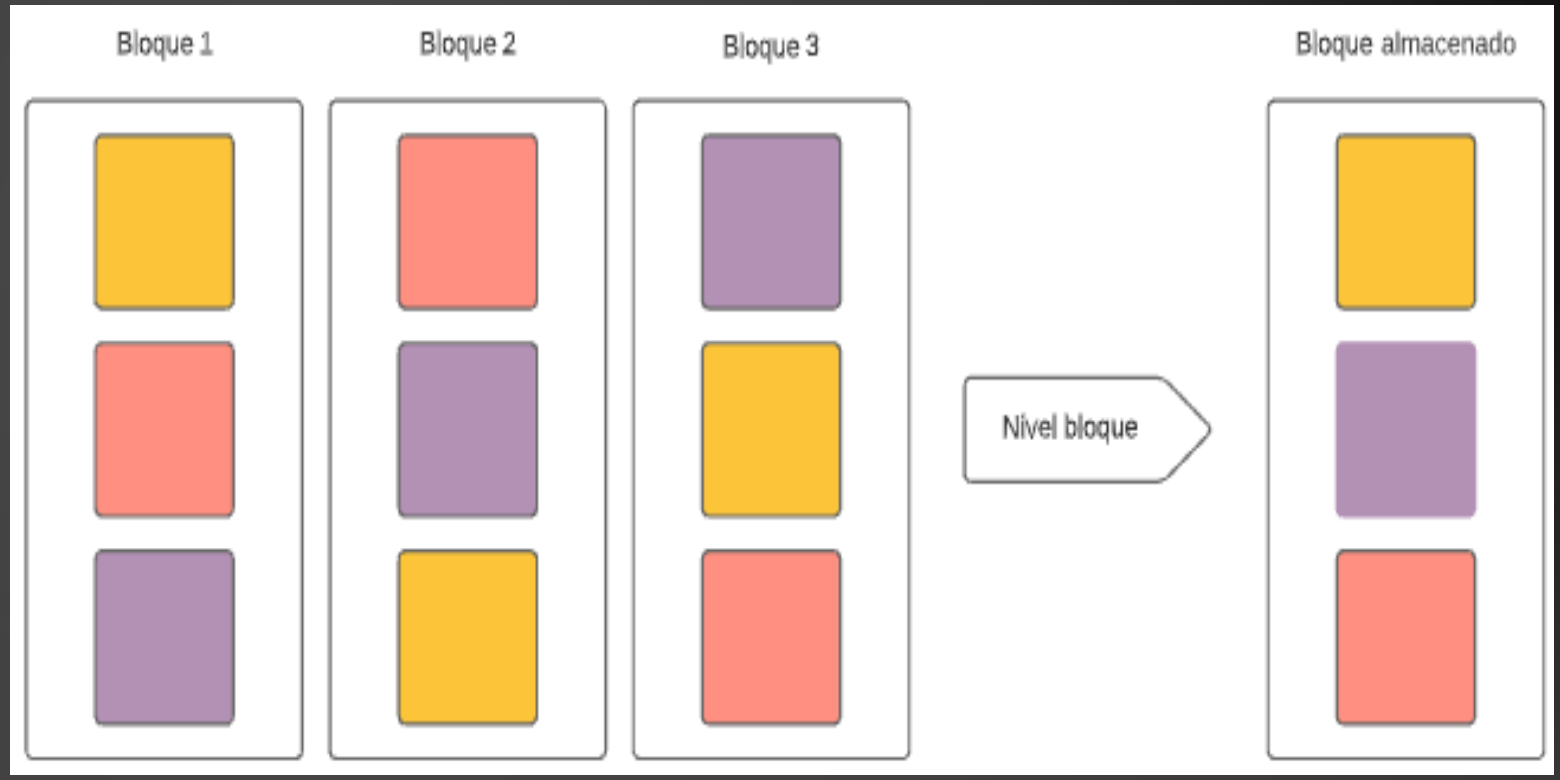
-También llamado SIS
-Hace comparaciones atreves de hashes
-genera identificadores para cada dato

Ventaja: requiere menor potencia de proceso

Desventaja: si se cambia algún dato se debe de almacenar de nuevo los archivos

GRANULARIDAD A NIVEL DE BLOQUES

Característica
-como su nombre lo dice separa los datos en bloques
-hace varias comparaciones
-al modificar no los afecta tanto
-entre mas bloques mayor va a ser la cantidad de comparaciones



GRANULARIDAD A NIVEL DE BYTES

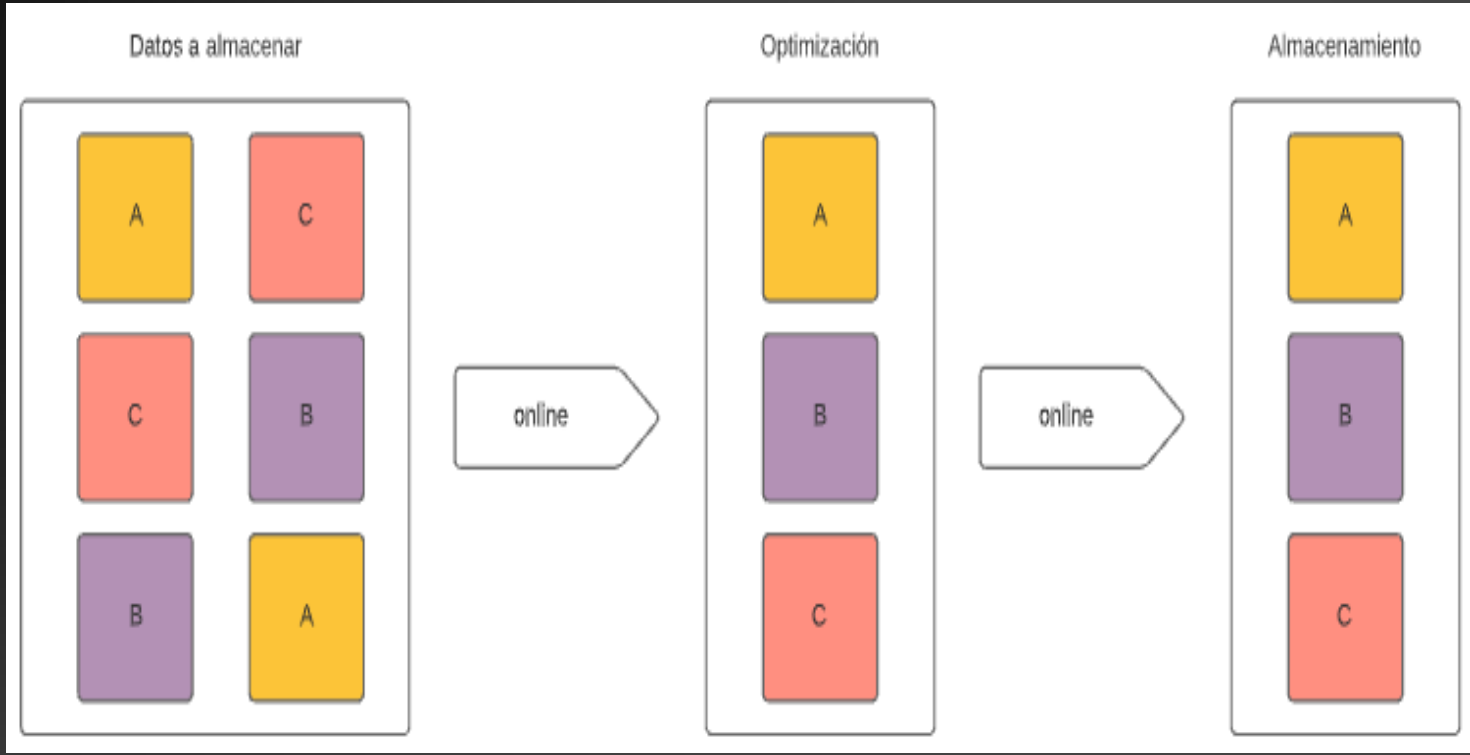
Característica
-tiene mayor granularidad
-separa byte por byte
-mayor precisión
-peor rendimiento

TIPOS DE DESDUPLICACIÓN

La desduplicacion se divide en dos ramas dependiendo de como llegan a su cometido:

- Deduplicación online
- Deduplicación offline

DESDUPLICACIÓN ONLINE



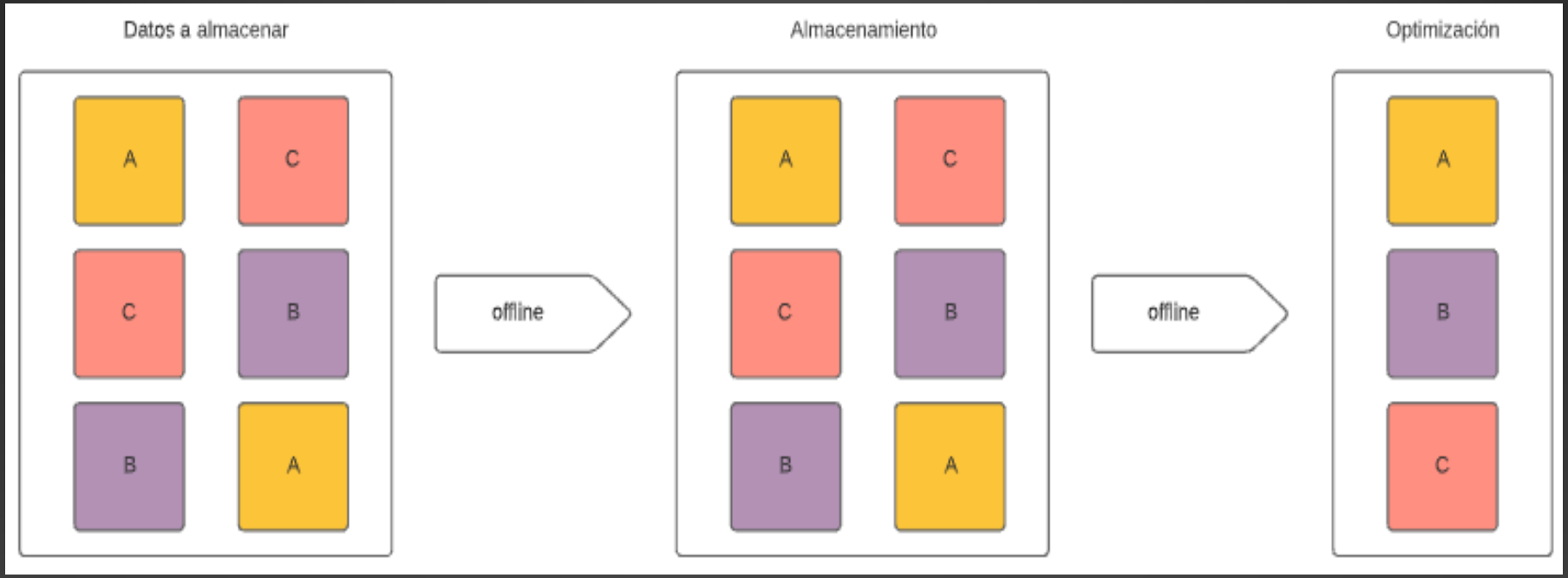
Características

-optimiza y después almacena
-solicita poco almacenamiento
-si llegan demasiados datos disminuye su rendimiento

DESDUPLICACIÓN OFFLINE

Características

-almacena y después optimiza
-si llegan demasiados datos duplicados solo deja una copia
-solicita demasiado almacenamiento



APLICACIONES DE LA DESDUPLICACIÓN

VENTAJAS Y DESVENTAJAS

Como nada es perfecto, la desduplicacion también tiene sus ventajas y desventajas

VENTAJAS

-Optimización del espacio físico
-Rápido procesamiento dependiendo del tipo de desduplicación
-Mantenimiento sencillo
-Rápida recuperación ante un desperfecto

DESVENTAJAS

-Problemas si hay poco espacio físico
-Poco crecimiento de eficiencia por los ratios
-Es poco efectiva para datos volátiles
-Si no hay bastantes datos duplicados, perdemos tiempo al procesar el hash

CONCLUSIONES

En conclusión podemos decir que la desduplicación es un buen método para eliminar los datos que se encuentren duplicados dentro del almacenamiento, aunque, como se vio, tiene unos cuantos defectos, y por tanto es mejor manejarlo con otros tipos de métodos de optimización de almacenamiento.

REFERENCIAS

[1] Wolf. G., Ruiz. E., Bergero. F., Meza E. (2015). *Fundamentos de sistemas operativos*. [Online]. Available: http://sistop.org/pdf/sistemas_operativos.pdf

[2] NetApp. (2019). *¿Qué es la deduplicación de datos?* [Online]. Available: https://www.netapp.com/es/data-management/what-is-data-deduplication/

[3] A. F. Mancheno, "*EVALUACIÓN DE LOS SISTEMAS DE DEDUPLICACIÓN OPENDEDUP (SDFS) Y ZFS PARA OPTIMIZAR EL SISTEMA DE ALMACENAMIENTO EN UN SERVIDOR DE BACKUPS*". Facultad de Informática y Electrónica. ESPC. Riobamba. Ecuador. 2015.

[4] Whitehouse. L. (2009). *Cuando y donde utilizar la tecnología de deduplicación de datos en el safeguard de disco*. [Online]. Available: https://searchdatacenter.techtarget.com/es/consejo/Cuando-y-como-utilizar-la-tecnologia-de-deduplicacion-de-datos-en-el-safeguard-de-disco

[5] Jiménez. F. (2009). *TÉCNICAS DE DEDUPLICACIÓN DE DATOS Y APLICACIÓN EN LIBRERÍAS VIRTUALES DE CINTAS*. [Online]. Available: https://oa.upm.es/1803/1/PFC_FRANCISCO_JAVIER_JIMENEZ_PATRICIO.pdf

REFERENCIAS

[6] S. T. Ahmed, L. E. George. "*Lightweight hash-based de-duplication system using the self detection of most repeated patterns as chunks divisors*". *Journal of King Saud University - Computer and Information Sciences*. pp. 6-7. April 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157821000914

[7] Menezes. A., Van Oorschot. P., Vanstone S. (1996). *Hash Functions and Data Integrity*. [Online]. Available: https://cacr.uwaterloo.ca/hac/about/chap9.pdf

[8] IBM. (2016). Active Memory deduplication. [Online]. Available: https://www.ibm.com/docs/en/linux-on-systems?topic=linuxonibm/liabd/virtcon-active-memory-de-duplication.htm

[9] Dutch. M. (2008). Understanding data deduplication ratios- [Online]. Available: https://www.snia.org/sites/default/files/Understanding_Data_Deduplication_Ratios-20080718.pdf

UN POCO MAS DE DESDUPLICACION

- Oracle. (2004). 4.10. Memory Overcommitment. [Online]. Available: https://docs.oracle.com/en/virtualization/virtualbox/6.0/user/guestadd-memory-usage.html
- Facebook. (2021). información sobre la desduplicación de eventos offline. [Online]. Available: https://www.facebook.com/business/help/1772588746090250?id=565900110447546
- IBM. (1993). Deduplication. [Online]. Available: https://www.ibm.com/docs/en/spectrum-protect/8.1.12?topic=reference-deduplication
- Wbsgo. (2014). Deduplicacion en el backup de datos. [Online]. Available: https://www.whitebearsolutions.com/la-deduplicacion-en-el-backup/