arcserve®

El ABC de la deduplicación de datos: desmistificando los diferentes métodos

20 de enero de 2016

Christophe Bertrand

El crecimiento exponencial de los datos es una realidad y está golpeando de lleno a empresas de todos los tamaños. Las pequeñas y medianas empresas son las más vulnerables a este desafío costoso y omnipresente. Aunque preservar los datos y su integridad sea tan importante como lo es para las grandes empresas, el problema es que sus presupuestos no pueden acompañar la constante necesidad de comprar más espacio de almacenamiento. Además, tampoco basta con aumentar la capacidad. Deben cumplir con los objetivos de punto de recuperación, sin mencionar que el tiempo es dinero cuando se trata de la recuperación de datos. El hecho es que, si se quiere administrar adecuadamente, proteger apropiadamente y recuperar por completo los datos, es necesario **reducir** su volumen.

Se requiere un cierto grado de reducción en el volumen de datos, también conocida como deduplicación, para responder al problema permanente de conjuntos de datos que no paran de crecer. La deduplicación, o eliminación de varias copias de los mismos datos, es posible con la ayuda de algoritmos que identifican y eliminan los datos redundantes. Las empresas tienen varias opciones cuando se trata de optimizar sus procesos con la deduplicación, y todas combaten el problema de una manera diferente. Pero vale la pena recordar que existen dudas sobre cómo funcionan las distintas opciones, los inconvenientes que deben tenerse en cuenta y cuáles otorgan un mayor beneficio (a un menor costo).

Comparación entre los diferentes métodos de deduplicación de datos

Para entender mejor la deduplicación y cómo puede ser utilizada de una mejor manera en su empresa, es importante entender y comparar los diferentes métodos.

El proceso de deduplicación se inicia con la creación de una firma digital única ("hash") para un bloque dado de datos. Este valor hash se guarda en la base de datos para que se pueda comparar con los valores hash creados para los nuevos bloques de datos entrantes. Al comparar los valores hash, se determina si el bloque de datos es único o un duplicado.

Cómo crear un Valor Hash

El proceso de crear un valor hash es fácil de entender y requiere una cantidad nominal de recursos informáticos. Lo que consume una gran cantidad de recursos informáticos es el proceso de comparar nuevos valores hash por cada valor hash almacenado en la base de datos. Cuando los valores hash se cuentan por millones, el proceso de búsqueda en la base de datos puede requerir muchos recursos informáticos.

Diferencias en la deduplicación de datos post-proceso e inline

En términos de lo bien que se realiza la deduplicación, es importante tener en cuenta las diferencias en las deduplicaciones "post-proceso" e "inline". Como su nombre lo dice, la deduplicación post-proceso significa que los datos entrantes se almacenan primero en el disco y se procesan para la deduplicación en un momento posterior. Por otra parte, cuando los datos se procesan para la deduplicación antes de ser escritos en el disco, esto se llama deduplicación **inline.**

La deduplicación **inline** tiene la ventaja de escribir los datos en el disco sólo una vez, pero implica el riesgo de desacelerar el tiempo de escritura en el disco si los recursos informáticos no son suficientes. Con el aumento de potencia de la CPU, la memoria RAM del sistema y las unidades de disco de estado sólido, la deduplicación **inline** es el método preferido de deduplicación en comparación con la deduplicación post-proceso, que requiere espacio de almacenamiento adicional y más disco para la escritura.

Deduplicación "en el destino" y "en el origen"

El segundo aspecto a tener en cuenta es dónde se realiza el proceso de deduplicación. Dada la existencia de la base de datos de hash, es razonable creer que los valores hash, que son pequeños valores de 16 bytes, se pueden compartir en un entorno de red de forma más rápida y fácil que si se comparte un bloque completo de datos. Aquí es donde las definiciones de deduplicación "en el destino" y deduplicación "en el origen" se vuelven importantes. La deduplicación en el destino significa que el conjunto completo de datos se comparte en la red y se deduplica cuando llega al dispositivo de deduplicación de destino. La deduplicación en el destino fue el primer método que logró un gran éxito cuando se combinó con la protección de datos. Los Purpose Build Backup Appliances (PBBA) son los dispositivos de backup de destino que los usuarios finales instalaban con su software de backup para reducir el espacio de almacenamiento de los datos de backup.

Por otro lado, la deduplicación en el origen significa que el proceso se inicia en la fuente de datos. Sólo cuando se determina que los datos son únicos, se transfieren al dispositivo de almacenamiento de backup. En una solución tradicional de backup, este proceso es administrado por el servidor de backup. Este servidor mantiene la base de datos de hash y trabaja con los agentes instalados en los clientes de backup. El agente en el cliente de backup ("la fuente de datos") calcula los valores hash de sus datos locales y envía los valores hash al servidor de backup para compararlos con los valores hash existentes almacenados en la base de datos. Luego, el servidor de backup le dice al agente qué datos son únicos y, por lo tanto, qué datos enviar al servidor de backup para su almacenamiento.

La ventaja de la deduplicación en el origen es la reducción de los datos que se envían a través de la red, y la mejora resultante en el rendimiento. En particular, las aplicaciones en el origen con grandes archivos de datos, como las aplicaciones de bases de datos, se beneficiarían enormemente al no tener que transferir archivos muy grandes a través de la red. El desafío de la deduplicación en el origen es que requiere una importante actualización del servidor de backup. Los servidores tradicionales de backup no procesan datos para deduplicación, sino que dependen de un PBBA para este propósito. Para ahorrar dinero y dolores de cabeza, es importante buscar soluciones de backup de última generación que hayan integrado la deduplicación en el servidor de backup y no utilicen un PBBA.

Deduplicación de datos global

Por último, debería ser considerado el método de deduplicación global, ya que está optimizado para la deduplicación enel origen. Con este método, cada computadora, máquina virtual o servidor en sitios locales, remotos y virtuales se comunica con un servidor de backup, que administra un índice de base de datos global de todos los archivos asociados, mientras determina intuitivamente qué incluir en el backup. El servidor de backup sólo obtiene nuevos datos, mientras elimina las copias duplicadas, y comparte la "inteligencia" deduplicada en todos los sistemas de origen. Ya que los datos del backup se deduplican a nivel global antes de ser transferidos al dispositivo de almacenamiento de backup de destino, sólo se envían los cambios a través de la red, lo que mejora significativamente el rendimiento y reduce el consumo de ancho de banda.

Para obtener mejores resultados

La cuestión de si las organizaciones deben o no aprovechar la deduplicación para administrar y proteger los datos de forma eficaz, es algo así como una obviedad. La deduplicación, cuando se aplica correctamente, es una tecnología fantástica que puede beneficiar en gran medida al rendimiento del backup y la recuperación, al tiempo que reduce los costos de almacenamiento. El proceso de deduplicación en sí utiliza muchos recursos informáticos y requiere una consideración especial cuando se implementa en su red. Al utilizar los últimos recursos de la CPU, RAM y SSD, puede satisfacer los requisitos de rendimiento de la deduplicación. Esto le permite disfrutar de los beneficios de la deduplicación inline, combinados con la deduplicación global en el origen, para obtener los mejores resultados.

Autor: Christophe Bertrand
Fuente:
http://www.dbta.com/Editorial/Think-About-It/The-ABCs-of-Data-Deduplication-Demystifying-the-Different-Methods-108274.aspx