

WHITE PAPER

Deduplicación de datos para backup: aceleración de la eficiencia y reducción de los costos de TI

Auspiciado por EMC Corporation

Laura DuBois

Julio de 2009

RESUMEN EJECUTIVO

La deduplicación de datos está mejorando considerablemente la economía de TI mediante la optimización de los requisitos de espacio de almacenamiento, las ventanas de backup y el ancho de banda de red en ubicaciones distribuidas y en data centers por igual. En ambientes reales, la deduplicación ha logrado acelerar la eficiencia de los procesos de backup y reducir los costos de TI. Este white paper analiza los distintos enfoques con respecto a la deduplicación para backup de datos y describe las consideraciones que deben tenerse en cuenta al elegir una solución. Asimismo, destaca el portafolio de backup de ofertas de deduplicación y casos de uso específicos de EMC para optimizar la eficiencia de los procesos de backup y reducir los costos.

Adopción de la deduplicación

La demanda de deduplicación de datos en ambientes empresariales y de tamaño mediano es cada vez mayor, ya que las empresas buscan la manera de mantenerse a la par del ritmo de crecimiento del almacenamiento, que casi se duplica año a año. Este crecimiento está impulsado por las nuevas aplicaciones, la proliferación de la virtualización, la creación de almacenes de documentos electrónicos y el uso compartido de documentos, la utilización de tecnologías Web 2.0 y la retención o preservación de registros digitales. Frente a los limitados presupuestos de TI, se intensifica la necesidad de controlar el crecimiento, ya que las empresas buscan reducir los costos operativos y de capital. Desde una perspectiva física, muchos administradores de data centers trabajan con infraestructuras limitadas en lo que respecta al espacio físico, el enfriamiento y la energía. La deduplicación es una tecnología que no solo ayuda a acelerar la eficiencia del almacenamiento mediante la disminución de costos, sino que, además, alivia los data centers físicamente limitados.

La deduplicación también enfrenta los retos relacionados con la falta de eficiencia de la red, los procesos de backup y la administración. A medida que el volumen de datos aumenta, existe una relación cada vez más desproporcionada entre la cantidad de personal de TI y la cantidad de almacenamiento que requiere administración. La deduplicación reduce el espacio para datos, lo que mantiene esta proporción en equilibrio. Del mismo modo, a medida que la brecha entre el disco y la potencia de procesamiento del servidor continúa ampliándose, las empresas buscan maneras de mejorar el performance en todo su ambiente por medio de una WAN, dentro de los subsistemas de almacenamiento en disco, y en ventanas de backup limitadas. La tecnología de deduplicación de datos puede optimizar la infraestructura virtual y física disponible mediante el envío de menor cantidad de datos por medio de enlaces a las redes remotas o locales.

Además, puede mejorar los tiempos de respuesta de nivel de servicio y ayudar a lograr la reducción de las ventanas de backup. La deduplicación también usa medios de acceso aleatorio (disco), lo que mejora los tiempos de recuperación, la seguridad de los datos y la confiabilidad.

Los retos más recientes son el resultado de la virtualización. Mientras las empresas continúan con la implementación de tecnología de máquinas virtuales para colaborar con la consolidación de servidores y la recuperación de desastres, las máquinas virtuales procesan datos redundantes, que necesitan protección. Para justificar los diferentes escenarios de falla o recuperar una imagen, generalmente, se requieren un servidor físico y archivos específicos en una sola solución de backup y proceso de backup. Los enfoques estándar, como la implementación de un agente de backup tradicional en una máquina virtual huésped o el uso de un backup de proxy VCB, no contribuyen a reducir el volumen de datos de las máquinas virtuales al que se le debe realizar backup ni los requisitos de ancho de banda de la red. La deduplicación ofrece importantes ahorros en capacidad de almacenamiento de backup. Además, algunas formas de deduplicación también reducen la cantidad de datos a los que se les debe realizar backup, lo que genera procesos de backup más acelerados y un menor impacto en la red. La deduplicación de manera conjunta con el software de backup enfrenta la necesidad de protección rentable, eficiente y completa de los ambientes de máquinas virtuales.

Los beneficios de la deduplicación

Las empresas están implementando la deduplicación de datos en una serie de ubicaciones en la plataforma de infraestructura a fin de enfrentar estos retos prácticos del mundo real. Entre los beneficios de la deduplicación se incluyen los siguientes:

- ☒ **Reducción de costos.** La deduplicación ofrece eficiencia de recursos y ahorros de costos que incluyen una disminución de las exigencias en cuanto a las baldosas del suelo, el enfriamiento y la energía de los data centers, además de la capacidad de almacenamiento, el ancho de banda de la red y el personal de TI.
- ☒ **Reducción de la emisión de dióxido de carbono.** La deduplicación reduce los requisitos de espacio, enfriamiento y energía para el almacenamiento, lo que disminuye la emisión de dióxido de carbono y da prueba de su responsabilidad ambiental.
- ☒ **Mejora de los niveles de servicio de backup y recuperación.** La deduplicación mejora de manera significativa el performance de los procesos de backup a fin de cumplir con ventanas de backup limitadas. La tecnología de deduplicación también aprovecha el almacenamiento en disco de acceso aleatorio para optimizar el performance de recuperación en comparación con los métodos de acceso secuencial (cinta).
- ☒ **Modificaciones en la rentabilidad del disco en comparación con la cinta.** La deduplicación hace posible los backups basados en disco para un conjunto más amplio de aplicaciones. La cinta tuvo una función destacada en los data centers empresariales debido a sus propiedades de archivo y beneficios económicos. Sin embargo, es posible que la disminución de costos/GBs para el disco cuando se usa con deduplicación haga que los costos de disco sean iguales o menores que los de las cintas.

La tecnología de deduplicación enfrenta muchos de los tradicionales retos de backup que las pequeñas y grandes empresas han tenido desde hace más de una década. Estos retos incluyen mantenerse a la par de la duplicación del crecimiento de datos, cumplir con ventanas de backup más breves, lograr recuperaciones más rápidas ante fallas operacionales y relacionadas con desastres, etc.

La tabla 1 describe la gran cantidad de retos de backup que existen y cómo la deduplicación puede enfrentarlos. Además, especifica el enfoque de la deduplicación más adecuado para enfrentar cada reto.

TABLA 1

Los retos de backup y el impacto de la deduplicación

Retos de backup	Impacto de la deduplicación	Deduplicación más adecuada
Los tiempos de recuperación son cada vez más cortos para minimizar los costos del tiempo fuera.	La deduplicación reduce el costo de almacenar más datos de backup en disco. Mantener los backups en disco en lugar de en la cinta mejora considerablemente los tiempos de recuperación para una amplia serie de aplicaciones.	Deduplicación en el origen o en el destino
La confiabilidad de los procesos de backup pone la recuperación de datos en riesgo.	La dependencia de los medios de cinta para los procesos de backup introduce el riesgo de errores de medios (medios defectuosos, dispositivos contaminados, etc.), la falta de medios disponibles o fallas de hardware. La deduplicación usa el disco en el proceso de protección de datos, lo que elimina o reduce estas situaciones de fallas.	Deduplicación en el origen o en el destino
Las ventanas de backup se reducen, ya que las operaciones se ejecutan 24x7 para cumplir con las exigencias de los clientes de todo el mundo.	Los backups tradicionales implican la transferencia de importantes cantidades de datos redundantes, lo que puede exceder ventanas de backup exigentes o inexistentes. La deduplicación reduce la cantidad de datos a los que se les debe realizar backup, lo que permite que se les haga backup a más datos en una ventana disponible.	Deduplicación en el origen
Una mayor virtualización de servidores implica menos recursos disponibles para los procesos de backup, lo que puede aumentar los tiempos de backup y acentuar las ventanas de backup.	La deduplicación en el origen implica que los datos duplicados no requieran el procesamiento de recursos compartidos, lo que disminuye la contención y acelera los backups de máquinas virtuales.	Deduplicación en el origen
El crecimiento de datos significa que no se les puede realizar backup a todos los datos en las ventanas de backup disponibles.	Las empresas enfrentan un crecimiento anual promedio del 50% en la cantidad de datos que requieren protección. Este crecimiento se contrapone con las limitaciones de las ventanas de backup nocturnas y los métodos tradicionales. La deduplicación enfrenta este reto de crecimiento y permite procesos de backup eficientes de los crecientes conjuntos de datos.	Deduplicación en el origen
La copia segura fuera del sitio mediante métodos de cinta tradicionales pone los datos en riesgo por pérdida o robo.	La dependencia de medios de cinta desmontables para el almacenamiento fuera del sitio en caso de desastres presenta riesgos que comprometen los medios físicos. La deduplicación junto con los procesos de replicación seguros permiten que una copia electrónica permanezca fuera del sitio, lo que elimina la necesidad del manejo manual de los medios de cinta y mejora la seguridad.	Deduplicación en el origen o en el destino

TABLA 1

Los retos de backup y el impacto de la deduplicación

Retos de backup	Impacto de la deduplicación	Deduplicación más adecuada
Los costos de infraestructura de backup aumentan para mantenerse a la par del crecimiento de capacidad y de las ventanas de backup.	La mayoría de las empresas enfrentan los retos relacionados con las ventanas de backup y el crecimiento de datos mediante la implementación de más infraestructura de cintas. La automatización y los discos de cinta pueden enfrentar los actuales cuellos de botella de performance y realizar los procesos de backup más rápidamente, aunque con sobrecarga en la administración y en los costos. La deduplicación reduce el continuo gasto en infraestructura de cintas para mantenerse a la par de estas tendencias.	Deduplicación en el origen o en el destino

Fuente: IDC, 2009

DEDUPLICACIÓN: QUÉ, DÓNDE, CUÁNDO Y CÓMO

Qué es la deduplicación

IDC define la deduplicación de datos como una tecnología que normaliza los datos duplicados en un solo objeto de datos compartido a fin de lograr eficiencia en la capacidad de almacenamiento. Específicamente, la deduplicación de datos hace referencia a todo algoritmo que busca objetos de datos duplicados (por ejemplo, bloques, fragmentos, archivos) y desecha los datos duplicados cuando los ubica. Cuando se detectan datos duplicados, no se conservan; por el contrario, se modifica un "puntero de datos" de manera que el sistema de almacenamiento indique una copia exacta del objeto de datos ya almacenado en el disco. Además, la deduplicación de datos opera solamente en datos únicos y elimina los costos relacionados con el mantenimiento de múltiples copias del mismo objeto de datos.

A diferencia del almacenamiento de una instancia (SIS), que deduplica los datos al nivel de los objetos o de los archivos, la deduplicación de datos se asocia más frecuentemente con procesos de deduplicación de subarchivos. La deduplicación de subarchivos examina un archivo y lo divide en "fragmentos". Luego, estos fragmentos más pequeños se examinan para detectar la presencia de contenido de datos redundantes en múltiples sistemas y ubicaciones. La deduplicación también se diferencia de la compresión, que reduce el espacio de un solo objeto en vez de en archivos o partes de un archivo. No obstante, los datos deduplicados también se pueden comprimir para un mayor ahorro de espacio.

Dónde se lleva a cabo la deduplicación

La deduplicación de datos de backup puede llevarse a cabo en el origen o en el destino. Un ejemplo de deduplicación en el origen es la disminución del tamaño de los datos de backup en el cliente (por ejemplo, Exchange o servidor de archivos) para que solo los datos únicos de subarchivos se envíen por la red durante el proceso de backup. Un ejemplo de deduplicación en el destino sería reducir el tamaño de los datos de backup después de que pasan por la red cuando llegan a un dispositivo de deduplicación. La deduplicación en el origen brinda ahorros en el almacenamiento, las ventanas de backup y el ancho de banda de red. La deduplicación en el destino proporciona ahorros en el almacenamiento, funciona con el software de backup existente y puede reducir el impacto en la red, aunque requiere un dispositivo de hardware en cada ubicación. En las ubicaciones donde se lleva a cabo la deduplicación no solo se obtienen diferentes beneficios, sino que los costos y los tiempos de implementación también se ven afectados. Las empresas deben evaluar sus actuales problemas de backup y vincular estos retos con los diferentes enfoques de la deduplicación (consulte nuevamente la tabla 1).

Deduplicación en el origen

La deduplicación en el origen (cliente) brinda una amplia serie de beneficios que exceden la optimización de la capacidad. Esto implica el envío de una cantidad considerablemente menor de datos, lo que alivia la infraestructura física/virtual congestionada y los enlaces LAN/WAN. Dado que solamente se envían segmentos nuevos o cambiados de datos en los subarchivos, se reduce considerablemente la cantidad de datos transferidos, lo que permite backups

completos diarios extremadamente rápidos. La sobrecarga incremental en el CPU del cliente para llevar a cabo la deduplicación de origen puede ser de hasta un 15%, pero el proceso de backup se completa más rápido que mediante los métodos tradicionales. El impacto general de la deduplicación de origen es realmente mucho menor en comparación con el producido por los agentes tradicionales en un período de siete días. En cambio, es posible que los ambientes con bases de datos muy grandes o bases de datos con altas tasas diarias de modificación opten por una solución orientada al destino. Afortunadamente, los proveedores, por lo general, cuentan con herramientas de evaluación de datos a fin de ayudar a los clientes a elegir la mejor opción. La deduplicación en el origen también ofrece flexibilidad de implementación, ya que las oficinas remotas más pequeñas pueden implementar simplemente el agente de backup de software, sin necesidad de incorporar hardware local adicional.

Deduplicación en el destino

La deduplicación en el destino optimiza la capacidad de almacenamiento en disco de backups, ya que solamente los datos nuevos y únicos en los subarchivos se almacenan en disco. Sin embargo, los datos de backup redundantes aún se envían al destino de deduplicación mediante software de backup tradicional. De modo que no proporciona alivio a una ventana de backup disponible. Un factor fundamental que debe tenerse en cuenta cuando se usa un enfoque orientado al destino es la capacidad de mantenerse a la par del performance de las ventanas de backup, y si la deduplicación posterior al proceso o en línea se requiere frente a una determinada carga de trabajo. (Consulte la siguiente sección para obtener más información sobre las diferencias entre la deduplicación posterior al proceso y la deduplicación en línea).

Además, un enfoque orientado al destino requiere la adquisición de un dispositivo de deduplicación incremental, que debe presupuestarse y administrarse como cualquier otro sistema. Cuando se agota la capacidad del dispositivo, se requiere la implementación de otro dispositivo de deduplicación. Algunas soluciones ofrecen el cluster de varios dispositivos para moderar este problema. La deduplicación orientada al destino puede utilizarse en data centers centrales para grandes volúmenes de datos y en ubicaciones remotas. No obstante, esto se traduce en la implementación de un dispositivo de deduplicación en cada ubicación remota con replicación remota desde varias oficinas a un dispositivo central y de mayor tamaño en el data center. La deduplicación orientada en el origen puede eliminar esta inversión en hardware en las oficinas remotas.

Cuándo se lleva a cabo la deduplicación

En la actualidad, hay dos enfoques diferentes disponibles para determinar *cuándo* se lleva a cabo el proceso de deduplicación: en línea o posterior al proceso. Algunos proveedores también están trabajando en un tercer enfoque llamado deduplicación adaptable o híbrida. La deduplicación en línea elimina los datos redundantes antes de que se escriban en disco, por lo tanto, no se requiere un área de staging del disco. La deduplicación posterior al proceso analiza y reduce los datos una vez almacenados en disco, de manera que requiere un área de staging de capacidad completa donde comenzar el proceso de deduplicación. Al elegir un enfoque, las organizaciones deben tener presente

cuestiones relacionadas con la capacidad del disco y la velocidad de los procesos de backup.

Un proceso en línea es más eficiente en cuanto a capacidad y no hay tiempo de retraso para que se inicie el proceso de deduplicación. En el caso de ambientes de gran capacidad con consideraciones acerca de las ventanas de backup, la deduplicación posterior al proceso prioriza la finalización del proceso de backup, pero requiere mayor capacidad inicial de almacenamiento. Estos enfoques pueden implicar una interrelación de requisitos de capacidad y performance. Un tercer enfoque, que aún se encuentra en las etapas de desarrollo, es la deduplicación adaptable o híbrida. Este método de deduplicación prioriza un enfoque en línea hasta que se alcance un umbral de performance y, luego, cambia automáticamente a un enfoque posterior al proceso, lo que optimiza el método para la carga de trabajo actual en el ambiente. Algunas soluciones líderes ofrecen deduplicación basada en políticas que permite que la configuración de deduplicación por parte del Cliente ocurra de manera inmediata o en función de un programa, o se desactive según las características de un conjunto de datos. Por ejemplo, los datos no estructurados y los conjuntos de datos más pequeños se pueden configurar para tareas de deduplicación inmediata y grandes trabajos de backup configurados para un procesamiento posterior, mientras que es posible desactivar la deduplicación para procesos de backup de medios enriquecidos o datos encriptados. La deduplicación basada en políticas les brinda a los usuarios el mayor nivel de flexibilidad para configurar la deduplicación según las condiciones del ambiente.

Cómo se lleva a cabo la deduplicación

La manera en que se lleva a cabo el proceso de deduplicación depende de la implementación. Un método de deduplicación basado en hash divide un flujo de backup o archivo en fragmentos de longitud variable o fija de datos en los subarchivos. Se calcula un valor hash para cada fragmento. Este proceso calcula un número específico para cada fragmento, el cual después se almacena en un índice. Si se actualiza un archivo, se guardan solo los datos cambiados en el subarchivo; dichas modificaciones no requieren que se guarde un archivo completamente nuevo. Una importante diferenciación que se puede hacer en las implementaciones basadas en hash es si el tamaño del fragmento tiene una extensión fija o variable. Un enfoque de longitud variable puede ajustar, de manera dinámica y según el tipo de contenido, el tamaño de un fragmento a fin de acomodar los fragmentos de datos redundantes, cuya posición se modificó o compensó en un flujo de bytes durante un cambio en los archivos. Un enfoque de longitud fija no detectará los datos redundantes que se reposicionaron o compensaron, por lo tanto, repetirá las operaciones de backup de fragmentos de manera ineficiente, si bien ya se encuentran en el catálogo de backup. Un problema potencial con un enfoque basado en hash es el I/O del disco y el performance. El índice de hash se conserva en la memoria, pero a medida que el índice de hash crece, es posible que desborde de la memoria al disco, por lo que requerirá I/O del disco para la búsqueda y la recuperación de fragmentos. Los proveedores tienen distintas maneras de enfrentar estos retos prácticos tecnológicos.

Un enfoque alternativo es la deduplicación de datos basada en delta, que almacena o transfiere los datos por medio de diferencias a partir de una copia de base. La base es una copia completa de los datos usados para recrear otras versiones de los datos. La deduplicación de datos basada en delta puede

realizarse a nivel de bloques o de bytes. El enfoque utilizado, ya sea basado en delta o hash, se convierte en una interrelación entre una tasa de deduplicación obtenida frente al performance. Los fragmentos de mayor tamaño tienden a reducir el índice de deduplicación de los datos, pero los fragmentos de menor tamaño generan mayor sobrecarga de indexación.

Otro factor que afecta la tasa de deduplicación es la posible capacidad del engine de reconocer un formato de datos particular (aplicación de backup particular, datos de Microsoft Exchange, etc.). La capacidad de detectar el formato de los datos requiere comprender dónde los metadatos específicos de la aplicación se inyectan en el flujo. El engine de deduplicación puede adaptar el tamaño del fragmento de manera que sea ideal para el formato de datos según la aplicación natural, lo que genera resultados de deduplicación potencialmente mayores.

CONSIDERACIONES AL EVALUAR LA TECNOLOGÍA DE DEDUPLICACIÓN

En la actualidad, hay disponibles en el mercado distintos tipos de productos con capacidades de deduplicación. Las aplicaciones de backup, los dispositivos, las librerías de cintas virtuales, las soluciones de optimización de WAN, los subsistemas de almacenamiento en disco primario, todos pueden presentar algún tipo de funcionalidad de deduplicación. Es importante que los miembros de una empresa estén de acuerdo en qué problemas se intentan resolver por tipo de datos o aplicación antes de elegir un tipo de deduplicación. Los diferentes enfoques de la deduplicación brindan diferentes beneficios en cuanto a eficiencia de la red, performance y capacidad.

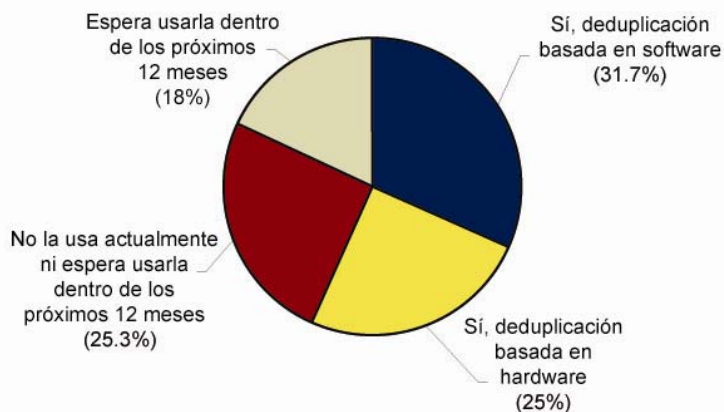
1. **Tasas de deduplicación.** La tasa de deduplicación obtenida variará de acuerdo con una gran cantidad de factores, incluidos el tipo de datos, la tasa de cambio de datos, los períodos de retención, los segmentos de longitud fija frente a los de longitud variable, las políticas de backup, el conocimiento del formato de archivos, etc. De acuerdo con la investigación de IDC, las tasas totales de deduplicación de almacenamiento en disco back-end del mundo real equivalen a 8:1 y 22:1 según los factores antes mencionados. Las soluciones de deduplicación en el origen pueden reducir el ancho de banda de red diario requerido en un porcentaje mucho mayor en comparación con los métodos de backup completos diarios tradicionales; varios proveedores líderes afirman que la relación es de 500:1. No obstante, al igual que con todas las métricas de performance, los beneficios variarán según el ambiente. Las empresas deben tener cuidado con las garantías de performance, escalación y rendimiento ofrecidas, y probar la deduplicación en las instalaciones con sus propios conjuntos de datos.
2. **Función de la compresión, la encriptación y la multiplexión.** La compresión (la codificación de datos para reducir su almacenamiento) puede ser una tecnología complementaria a la deduplicación. La compresión está optimizada para un único objeto y reduce su espacio, mientras que la deduplicación funciona en todos los objetos. Sin embargo, la compresión puede aplicarse a los datos que ya se deduplicaron a fin de proporcionar más ahorros de espacio. No obstante, si la deduplicación se aplica a un archivo ya comprimido (o encriptado), el beneficio de la deduplicación será insignificante o inexistente. Las empresas que usan la encriptación o la compresión a nivel del software y la deduplicación basada en el destino

posiblemente necesiten desactivar estas funciones para obtener beneficios de deduplicación o utilizar compresión basada en hardware. También se debe considerar el actual uso de la multiplexión para backups que intercala datos de múltiples clientes en un único flujo enviado a un disco de cinta. Sin embargo, este proceso dificulta la detección de segmentos de datos que ya existen. Es necesario desactivar la multiplexión si las empresas desean beneficiarse de la deduplicación.

3. **Deduplicación para máquinas virtuales.** El uso de máquinas virtuales en ambientes de producción ha intensificado la necesidad de proteger y recuperar la máquina virtual, el host físico y los archivos. Las opciones de backup de máquinas virtuales incluyen un agente en cada huésped, un backup de servidor proxy VCB o un agente en la consola de servicios. Las soluciones de backup tradicionales son ineficientes para los backups de máquinas virtuales porque transfieren grandes cantidades de datos redundantes y requieren gran cantidad de ciclos de CPU para la ejecución de un backup, lo que genera una menor consolidación de servidores y performance de backup deficiente. La deduplicación tiene la capacidad de enfrentar estas limitaciones. La deduplicación en el origen implica que los datos duplicados nunca se trasladan por la infraestructura física subyacente compartida, por lo que los backups completos diarios son rápidos y eficientes. La deduplicación también puede ocurrir a escala global, en VMDKs, para eliminar los backups de datos redundantes en los sistemas. Otro factor que se debe considerar es que un archivo VMDK, cuando se actualiza, se compensa. Solamente la deduplicación de longitud variable puede explicar esta compensación, al encontrar y transferir solamente los cambios únicos dentro del VMDK.
4. **Deduplicación para oficinas remotas.** Al igual que las operaciones del data center, las oficinas remotas requieren recuperación de desastres (remota) y local. No obstante, las características de las oficinas remotas presentan retos. Las ubicaciones de las oficinas remotas, generalmente, presentan un ancho de banda WAN limitado, personal de TI no especializado y una cantidad desproporcionada de oficinas y sucursales remotas en relación con la cantidad de data centers regionales o principales. La deduplicación puede minimizar la transferencia de datos mediante WAN, y la deduplicación global elimina los datos redundantes del data center y las oficinas. Dado que muchas empresas cuentan con personal de TI limitado en oficinas remotas, buscan la manera de reducir el espacio de hardware de almacenamiento en ubicaciones distribuidas. La deduplicación en el origen puede implementarse por medio de software, lo cual modera este problema. Un estudio reciente de IDC analizó el uso de la deduplicación en oficinas remotas y el tipo de deduplicación implementado (consulte la figura 1).

FIGURA 1

Uso de la tecnología de deduplicación para la protección de datos en oficinas remotas



n = 300

Fuente: Estudio especializado en oficinas remotas de IDC, 2009

- 5. Deduplicación para data centers de producción/recuperación de desastres.** En el caso de los ambientes de data centers, los volúmenes de datos escalan considerablemente, aunque con el beneficio de conectividad LAN y conexiones más rápidas con los sitios de recuperación de desastres. Los grandes data centers aún se esfuerzan por cumplir con las ventanas de backup para, por lo menos, algunas de sus aplicaciones y no pueden permitirse poner en riesgo el performance de los procesos de backup. Esta realidad puede requerir un enfoque de la deduplicación basada en políticas que incluye deduplicación en el origen y el destino según la aplicación y el ambiente. Es posible que la optimización del ancho de banda de la red en un data center sea menos prioritario que la replicación remota a un sitio de recuperación de desastres. Sin embargo, si las ventanas de backup continúan reduciéndose, el ancho de banda de la red se transformará en un problema con el paso del tiempo.
- 6. Deduplicación global.** La definición de lo que constituye la deduplicación global puede variar. Algunos usuarios observan la deduplicación global entre sitios, mientras que otros lo hacen dentro de un único frame de almacenamiento. Sin embargo, la deduplicación global, en el mejor de los casos, debería ser entre sitios y frames a fin de obtener el máximo beneficio. La deduplicación en el destino puede ofrecer deduplicación global de varias oficinas remotas a un frame único y sus pares de réplicas. Sin embargo, cuando se alcanza un máximo de performance o capacidad, se debe configurar un nuevo dispositivo, que introduce otro dispositivo de deduplicación independiente. La deduplicación en el origen también puede ofrecer deduplicación global de varios servidores de data centers de oficinas remotas y sus pares. Las empresas deben recordar que la deduplicación global es un término que tiene un significado diferente según el proveedor y su enfoque.

7. **Deduplicación y replicación.** La replicación es realmente el siguiente campo de batalla para la tecnología de deduplicación. Tanto los proveedores innovadores como los ya establecidos han comprobado que funciona, y las reacciones positivas de la evaluación de la tecnología por parte de los usuarios generan entusiasmo y demanda. La deduplicación se está implementando en ambientes empresariales, en ubicaciones edge y core, con un aumento de la eficiencia y una reducción de los costos de infraestructura. Dado que más empresas desean adoptar enfoques más estratégicos (es decir, favorecer los casos de uso como el cumplimiento de normas) en relación con el aprovechamiento de las cintas en los data centers y la minimización de su uso en las ubicaciones remotas, la función de la replicación remota pasa a ser de suma importancia. Los requisitos de replicación de los usuarios son cada vez más sofisticados e incluyen:
- ❑ **Replicación basada en deduplicación, que replica un conjunto de datos deduplicados y no un volumen completo.** Algunos proveedores ofrecen servicios de replicación con un producto activado para deduplicación. No obstante, las empresas deben asegurarse de que la función de replicación se base en la deduplicación.
 - ❑ **Replicación "todo o nada" a nivel de cintas/directorios.** Algunos casos de uso requieren una replicación completa del sistema, mientras que otros pueden requerir flexibilidad para determinar qué espacios compartidos o cintas replicar.
 - ❑ **Monitoreo de la replicación, optimización del performance y resolución de problemas.** A pesar de la deduplicación, la mayoría de las grandes empresas aún tiene muchos datos que replicar. Esto se administra con un proceso de replicación asíncrono o programado, y con monitoreo del proceso de replicación y del ancho de banda usado. Las herramientas de resolución de problemas y optimización ayudan a garantizar que el proceso de replicación continúe en una ventana de replicación disponible.
 - ❑ **Replicación en tiempo real y programada para enlaces de menor y mayor latencia.** Algunos enlaces/sitios requieren replicación en tiempo real, mientras que otros pueden funcionar adecuadamente con un proceso de replicación programado. Las características de las oficinas remotas varían significativamente y pueden tener enlaces de menor latencia, mientras que es posible que los enlaces entre dos data centers no enfrenten el mismo problema.
8. **Seeding y migración.** Si bien la deduplicación es muy buena para reducir el almacenamiento y/o la transmisión de datos redundantes, requiere el establecimiento de un primer backup o de una base inicial. En el caso de la deduplicación del edge al core, los usuarios deben considerar cómo crear esta base en enlaces de ancho de banda limitado. La mayoría de los proveedores ofrece alguna clase de servicio de seeding para crear rápidamente esta base, ya sea mediante un proceso de replicación masiva basada en deduplicación con sistemas paralelos o mediante el restore local en un sistema de deduplicación de una serie de cintas de un último backup completo. Con los ciclos de actualización de almacenamiento en un ciclo de rotación de entre tres y cinco años, otras consideraciones que deben tenerse en cuenta son la manera en que se realiza una migración y la interrupción que causará en el ambiente existente.

9. **Selección de proveedores.** Es posible que los proveedores realicen varias reclamaciones y declaraciones en relación con su enfoque de deduplicación. La investigación de IDC demuestra que por lo general no todos los productos de deduplicación disponibles funcionan realmente según lo indica su publicidad. Las empresas deben considerar el tiempo durante el cual se comercializó un determinado producto activado para deduplicación, la manera en que diversos clientes usan el producto para la producción y la madurez del producto en ambientes reales. Las empresas deben investigar en profundidad la escalabilidad de un producto y preguntar por la matriz de soporte de un sistema y/o una aplicación. Las empresas que deciden no llevar a cabo una prueba de concepto (POC) corren el riesgo de encontrarse con resultados inesperados de performance y confiabilidad.

10. **Casos de uso para la deduplicación.** La deduplicación es una tecnología que promete incrementar la plataforma de infraestructura de almacenamiento. Hasta la fecha, esta tecnología se implementó de manera generalizada en el ámbito de los procesos de backup, donde ya existe una gran cantidad de datos redundantes. Cada semana se realizan backups de estos mismos datos, para lo cual se usan innecesariamente recursos de almacenamiento, redes y servidores. Algunas empresas están comenzando a considerar y probar la deduplicación existente en ambientes de almacenamiento primario desde un enfoque de almacenamiento en red (NAS). No obstante, esta implementación requiere un performance mejorado para evitar las consecuencias de tiempo de respuesta y latencia. En la actualidad, la tecnología de deduplicación está bien posicionada para las operaciones de backup de ambientes de data centers, oficinas remotas y máquinas virtuales.

PORTAFOLIO DE EMC DE SOLUCIONES ACTIVADAS CON DEDUPLICACIÓN

EMC ofrece una amplia gama de productos activados con deduplicación para ayudar a los clientes a reducir los costos de TI y acelerar la eficiencia de los procesos de backup. Las soluciones de deduplicación de backup incluyen EMC Avamar, que brinda un enfoque orientado al origen respecto de la deduplicación; EMC Disk Library, que ofrece un enfoque orientado al destino respecto de la deduplicación; y EMC NetWorker, que se puede implementar con el enfoque orientado al origen o al destino, o ambos. Además, aunque no está incluida en el alcance de este white paper, EMC ofrece una solución de deduplicación para almacenamiento primario y datos de backup con su sistema EMC Celerra de almacenamiento conectado en red, y una solución de deduplicación de archivo en disco con su línea de productos Centera.

EMC Avamar

Las soluciones de backup y recuperación de EMC Avamar incluyen tecnología de deduplicación integrada para identificar datos redundantes en el origen, lo que minimiza los datos de backup antes de enviarlos por medio de la LAN o WAN. Con Avamar, una empresa obtiene reducción de datos y backups diarios completos y rápidos para ambientes VMware, oficinas remotas y servidores LAN y NAS de data centers. Avamar también deduplica los datos de backup a nivel global en todos los sitios y servidores, y con el transcurso del tiempo. A diferencia

de los productos que usan métodos de recuperación tradicionales, Avamar puede hacer restore de los datos rápidamente en un solo paso, ya que elimina la complicación de recuperar el último backup completo satisfactorio y los backups incrementales posteriores a fin de alcanzar el punto de recuperación deseado. Las capacidades de Avamar representan una variación fundamental de las aplicaciones de backup tradicionales.

El agente de Avamar realiza un seguimiento de los archivos que son nuevos o que se han modificado. El agente no necesita recorrer la totalidad del árbol de sistemas de archivos para identificar datos nuevos o cambiados y controlará, en primer lugar, la memoria caché local en busca de dichos archivos. Al identificarlos, el agente dividirá los archivos nuevos o modificados en segmentos de datos de longitud variable en subarchivos y asignará un valor hash (ID único) a cada segmento. El agente se comunicará con el servidor de Avamar para determinar si el hash es único o ya existe. Si el segmento de datos es nuevo, se enviará a toda la LAN/WAN durante el backup diario completo.

Estos procesos aumentarán la utilización de CPU en el host en comparación con un agente de backup tradicional. Sin embargo, debido a que el backup protege de manera eficiente solo los segmentos de datos nuevos, los backups de Avamar se completan considerablemente más rápido que los backups tradicionales completos e incrementales. Por ejemplo, un backup incremental que generalmente se lleva a cabo en 10 horas puede llevar cerca de 1 hora con Avamar, lo que reduce el impacto semanal del backup de 50 horas a 5 horas para backups incrementales de lunes a viernes. Además, los backups diarios completos de Avamar son considerablemente más rápidos que los backups completos tradicionales.

Las soluciones de backup y recuperación de Avamar brindan deduplicación global y en el origen, lo que las transforma en la opción ideal para empresas con los siguientes ambientes:

- Empresas que implementan **máquinas virtuales** y evalúan una nueva estrategia de protección para recuperar los servidores físicos, virtuales y objetos específicos.
- Empresas que intentan mejorar el backup de **oficinas remotas** para obtener backups diarios rápidos y completos, administración centralizada, confiabilidad mejorada, replicación segura y tráfico de backup reducido en enlaces WAN congestionados
- Empresas que buscan detener el crecimiento de datos, reducir las ventanas de backup y el tráfico de red para el backup de ambientes **NAS y de servidor de archivos** locales.

EMC Avamar puede implementarse en cuatro tipos de configuraciones:

- Software Avamar.** En el caso de oficinas remotas más pequeñas, el agente de software Avamar puede implementarse en los sistemas para que estén protegidos (clientes) sin hardware adicional local.
- Servidor de Avamar de otros fabricantes.** El software Avamar puede adquirirse e implementarse en una variedad de servidores estándar de la industria certificados con almacenamiento de disco interno.

- ☒ **Avamar Data Store.** Esta solución escalable e integral incluye el software Avamar preinstalado y preconfigurado en hardware de EMC para solicitud, implementación y servicio simplificados.
- ☒ **Avamar Virtual Edition for VMware.** Esta configuración, de las mejores en la industria, hace posible que se implemente un servidor de Avamar como dispositivo virtual en un servidor ESX existente, lo que permite aprovechar los recursos adjuntos y el almacenamiento en disco.

Avamar es diferente a otros enfoques de deduplicación en el origen disponibles en el mercado. Por ejemplo, la deduplicación de Avamar usa segmentos de datos de longitud variable de subarchivos, que brindan eficiencia y performance de nivel superior, en comparación con soluciones que usan segmentos de longitud fija. Avamar usa arquitectura grid para escalar la capacidad y el performance; cada nodo incremental aumenta el CPU, la memoria, el I/O y el almacenamiento para todo el sistema.

El grid de Avamar usa una configuración de arreglos redundantes de nodos independientes (RAIN) para obtener tolerancia a fallos integrada y alta disponibilidad en todo el grid, y elimina los puntos de falla únicos. Avamar distribuye su índice interno en los nodos de Avamar para obtener confiabilidad, equilibrio de carga y escalabilidad. Además, diariamente y de manera automática, Avamar verifica que los datos de backup sean totalmente recuperables, y el servidor de Avamar se autoverifica dos veces al día para garantizar la integridad del servidor. Por último, Avamar ofrece una amplia gama de servicio de soporte para aplicaciones y clientes, incluso soporte para Exchange, SQL, Oracle, DB2, SharePoint, Lotus Notes y NDMP.

Avamar brinda varias maneras de proteger las máquinas virtuales y físicas. Las opciones para el backup de Avamar de ambientes de máquinas virtuales de VMware incluyen lo siguiente:

- ☒ **Agente Avamar en SO huésped.** Un agente Avamar dentro de cada SO huésped brinda un enfoque de backup que es considerablemente más eficiente que los enfoques de backup de agentes tradicionales. Los agentes Avamar livianos reducen los datos de backup en el huésped, lo que disminuye los requisitos de red y la contención para CPU, NIC, disco y recursos de memoria compartidos. Debido a que solo se hace backup de los datos en subarchivos nuevos o únicos, Avamar permite backups completos diarios y rápidos.
- ☒ **Backup de Avamar para VCB.** Un agente Avamar que se ejecuta en el servidor proxy VCB hace backup solo de los datos únicos y descarga el procesamiento para las máquinas huésped. La deduplicación se lleva a cabo todos los archivos VMDK y en ellos, y soporta backup a nivel de archivos e imágenes VCB. La eficiente replicación de Avamar permite que los archivos VMDK se transfieran rápidamente a toda la WAN para respaldar objetivos de recuperación de desastres.
- ☒ **Agente Avamar en la consola ESX.** Un agente Avamar en la consola ESX puede llevar a cabo la deduplicación entre todos los archivos VMDK y en ellos. Este método brinda backup a nivel de imagen y una opción de restore, sin depender de VMware VCB ni del almacenamiento de uso compartido. Sin embargo, puede utilizarse para restore a nivel de archivos.

EMC Disk Library

La familia EMC Disk Library (DL) ofrece deduplicación basada en políticas con sus sistemas 1500, 3000 y 4000. EMC Disk Library 1500 y 3000 brindan backup a disco basado en LAN con deduplicación incluida. El diseño de DL1500 está orientado a clientes de tamaño mediano que desean performance mejorado, retención más prolongada en sitio y menores costos de replicación. DL1500 comienza a una capacidad utilizable de 4 TB y se expande a 36 TB, con una tasa de recopilación de backup sostenida de 0.72 TB/hora con deduplicación de datos inmediata, o hasta 0.84 TB/hora cuando se retrasa el proceso de deduplicación.

DL3000 comienza a 8 TB de capacidad utilizable y se expande a 148 TB, con una tasa de recopilación de backup sostenida de 1.44 TB/hora con deduplicación de datos inmediata. El sistema de DL1500 y DL3000 incluye deduplicación basada en políticas. A diferencia de los modelos DL1500 y DL3000, la deduplicación de DL4000 se realiza por medio de una opción adicional de hardware para sistemas nuevos e instalados de librerías de cintas virtuales de DL4000. Las empresas pueden implementarlo para reducir los requisitos de capacidad para el backup a disco y disminuir el tráfico de red para la replicación entre data centers.

La deduplicación de EMC Disk Library es ideal para el data center, los grandes volúmenes de almacenamiento y los ambientes de bases de datos con un alto nivel de cambio que consideran la incorporación de disco para los procesos de backup. Las empresas que utilizan la deduplicación de Disk Library:

- Buscan detener el crecimiento de **datos de gran volumen** para el backup de los ambientes existentes de las librerías de cintas virtuales de EMC.
- Implementan una nueva **estrategia de backup a disco** para obtener recuperación y confiabilidad mejorados y, a la vez, minimizar los costos de almacenamiento.
- Incorporan tecnología de discos y deduplicación en un ambiente **existente de EMC Disk Library**.
- Utilizan la **bóveda electrónica de backups** para un data center de recuperación de desastres y minimizan el uso de la cinta física.
- Buscan **reemplazar la cinta con el disco para backup** con una mínima interrupción en las operaciones de backup actuales.

Los sistemas de Disk Library usan un método de deduplicación en el destino. La misma capacidad de deduplicación funciona en toda la familia Disk Library y brinda deduplicación a nivel de bloque, de longitud variable, basada en hash en el destino. La deduplicación de Disk Library usa "filtros de sensibilidad de aplicaciones" que pueden detectar el formato del flujo de datos y comprende dónde se incorporan metadatos específicos de una aplicación en el flujo. El filtro ubicará marcadores en torno a estos metadatos y los filtrará para un mayor impacto de deduplicación.

Los sistemas usan un proceso de deduplicación basado en políticas, configurable por el Cliente. La deduplicación puede configurarse para que se lleve a cabo de manera inmediata o en función de lo programado, o se puede desactivar totalmente. La política está configurada a nivel de recursos

compartidos de archivos o de librería virtual. La deduplicación en "modo inmediato" o en línea es ideal para grupos más pequeños de datos y datos no estructurados. La deduplicación en "modo programado" permite que el proceso de deduplicación dé prioridad al backup, que se completa antes de que comience la deduplicación. Este procedimiento, ideal para grupos más grandes de datos, permite que los backups se completen entre un 150% y un 200% más rápidamente (según EMC) que si se realizan en modo inmediato.

Para tipos de datos que no se deduplican de manera adecuada, la capacidad puede desactivarse. Para replicación remota activada con deduplicación para fines de recuperación de desastres, se puede configurar la replicación a nivel de sistemas, aplicaciones, directorios o cartucho de cinta virtual. El índice de deduplicación de Disk Library se agrupa en clusters en oficinas, y los objetos similares se agrupan en categorías para lograr búsquedas de índices eficientes y minimizar el I/O de disco. La compresión de hardware brinda otro nivel de optimización de almacenamiento.

EMC NetWorker

EMC NetWorker es una aplicación de backup para empresas que centraliza las operaciones de backup y recuperación. NetWorker brinda una plataforma común que soporta una amplia variedad de opciones de protección de datos, entre ellas, backup a disco, replicación, protección continua de datos y deduplicación en ambientes físicos y virtuales. La versatilidad de NetWorker lo convierte en el software de backup ideal para los clientes que buscan simplificar la administración en los ambientes, desde grandes data centers hasta oficinas remotas. La aplicación principal de NetWorker brinda deduplicación en el origen mediante la integración con la tecnología de deduplicación de EMC Avamar y también puede aprovechar las soluciones de deduplicación de destino, como EMC Disk Library, dentro del alcance de sus operaciones.

Las empresas que usan NetWorker:

- Buscan detener el crecimiento de datos de gran volumen para los ambientes **existentes de Networker.**
- Implementan una nueva estrategia de backup a disco para lograr una recuperación mejorada que requiere el uso de cintas físicas para archivo a largo plazo.
- Incorporan tecnología de discos y deduplicación en un ambiente **existente de EMC Disk Library.**
- Cumplen con diversos requisitos: algunos de ellos son ideales para deduplicación en el origen y otros son más adecuados para deduplicación en el destino.
- Reducen los costos y la complejidad al consolidar múltiples estrategias de protección de datos en una sola aplicación.

El enfoque de deduplicación de la aplicación NetWorker ha avanzado en el mercado desde el punto de vista de su integración de la deduplicación con una aplicación tradicional de backup. El software cliente NetWorker para ambientes de deduplicación y no deduplicación es un agente único. Se han integrado completamente las capacidades de deduplicación en el origen, lo que minimiza los requisitos de implementación y mantenimiento. La consola de NetWorker puede administrar y monitorear ambos tipos de backup: tradicional y de deduplicación. Para los usuarios de NetWorker que desean obtener los beneficios de la deduplicación, no existe un costo adicional por el uso del cliente.

A diferencia de otros productos, NetWorker no tiene SKUs incrementales de software ni costos adicionales por la integración de deduplicación. Los clientes de NetWorker pueden incorporar el engine de deduplicación apropiado en el ambiente de backup: la solución de back-end EMC Disk Library o Avamar. Uno de los beneficios de usar la deduplicación activada para NetWorker es el soporte para cintas físicas, que garantiza que los usuarios que necesitan usar cintas puedan hacerlo con la misma aplicación. Otro beneficio de la deduplicación de la aplicación de backup es el uso correcto del aprovisionamiento y de las secuencias de encriptación y compresión. NetWorker brinda a las empresas sólidas funciones de deduplicación sin causar interrupciones en el ambiente actual de backup.

RETOS: ¿QUÉ ENFOQUE ADOPTAR?

Como se muestra en este documento, los diferentes enfoques y tecnologías de deduplicación brindan distintas ventajas para cada caso de uso. Por lo tanto, es importante asignar cada producto de EMC al ambiente en el que brinde el mayor nivel de eficiencia. La tabla 2 describe los criterios que pueden usar las empresas para decidir qué productos de EMC activados para deduplicación se adecuan a sus requisitos.

EMC cuenta con una gran variedad de productos con funcionalidad de deduplicación. Si bien la deduplicación es una función o tecnología, y no un producto independiente, EMC debe incrementar la capacitación para los clientes en relación con cuál es el lugar más indicado para aprovechar la capacidad, según los retos específicos del ambiente del Cliente. La capacitación, junto con los casos de estudio documentados y los parámetros de referencia de pruebas de escala y performance, aumentará la confianza del Cliente para aplicar la tecnología en un producto determinado.

TABLA 2

Selección de un producto de EMC activado para deduplicación

	EMC NetWorker	EMC Disk Library	EMC Avamar
Deduplicación para backup	<ul style="list-style-type: none"> • En el origen • Deduplicación en línea 	<ul style="list-style-type: none"> • En el destino • Basada en políticas y configurable para deduplicación inmediata, programada o desactivada 	<ul style="list-style-type: none"> • En el origen • Deduplicación en línea
Ideal para:	<ul style="list-style-type: none"> • Ambientes NetWorker • Necesidad de soporte físico de cintas • Ambientes grandes y heterogéneos 	<ul style="list-style-type: none"> • Requisitos de alta velocidad para backup y recuperación • Replicación para backup fuera del sitio • Soporte para el ambiente actual de backup: sin cambios operacionales • Soporte para data centers y sitios remotos 	<ul style="list-style-type: none"> • Ambientes virtuales • Oficinas remotas • Servidores LAN/NAS
Opciones de implementación	<ul style="list-style-type: none"> • Un solo agente NetWorker • Agente para oficinas remotas pequeñas • Para el nodo de deduplicación, cualquiera de los siguientes: <ul style="list-style-type: none"> • Avamar Data Store: solución inmediata integral (hardware y software) • Servidor de otros fabricantes: creación de un servidor de Avamar propio • Avamar Virtual Edition: dispositivo virtual que aprovecha el servidor ESX y el disco existentes 	<ul style="list-style-type: none"> • Hardware de dispositivo 	<ul style="list-style-type: none"> • Solo agente: para oficinas remotas pequeñas • Avamar Data Store: solución inmediata integral (hardware y software) • Servidor de otros fabricantes: creación de un servidor de Avamar propio • Avamar Virtual Edition: dispositivo virtual que aprovecha el servidor ESX y el disco existentes

Fuente: IDC, 2009

CONCLUSIÓN

La tecnología de deduplicación puede acelerar la eficiencia de los procesos de backup y reducir los costos de TI. Las empresas están implementando diferentes tipos de soluciones activadas con deduplicación para enfrentar infinitos retos de costos y operaciones frente al creciente volumen de datos de backup. IDC considera que la deduplicación es una función principal y obligatoria para una variedad de soluciones de almacenamiento para enfrentar estos retos. EMC como proveedor está bien posicionado para enfrentar estos problemas de larga data, y ofrece una gama de soluciones para una variedad de ambientes y casos de uso a fin de cumplir con las exigencias de tecnología de los clientes en los próximos cinco años.

Aviso de copyright

Publicación externa de información y datos de IDC: cualquier información de IDC que se utilice en material publicitario o promocional, y comunicados de prensa requiere el permiso previo por escrito del Vicepresidente o Gerente Regional de IDC correspondiente. La solicitud de permiso debe enviarse junto con un borrador del documento propuesto. IDC se reserva el derecho de denegar la aprobación de uso externo por cualquier motivo.

Copyright 2009 IDC. Queda totalmente prohibida la reproducción sin el permiso por escrito.