

WHITE PAPER

Backup und Recovery: Höhere Effizienz und geringere IT-Kosten durch Datendeduplizierung

Gefördert von: EMC Corporation

Laura DuBois

Robert Amatruda

Februar 2010

EXECUTIVE SUMMARY

Datendeduplizierung verbessert die IT-Wirtschaftlichkeit enorm, indem sie die Anforderungen an die Netzwerkbandbreite, die Backup-Zeitfenster und den Speicherplatz in verteilten Unternehmen und Rechenzentrumstandorten gleichermaßen minimiert. In echten Umgebungen erhöht die Deduplizierung die Backup- und Recovery-Effizienz und senkt die IT-Kosten. Dieses White Paper beschäftigt sich mit den verschiedenen Ansätzen zur Deduplizierung von Backup-Daten und gibt Hilfestellung bei der Auswahl einer Lösung. Ebenso werden das EMC Portfolio mit Deduplizierungsprodukten für Backup und Recovery sowie spezifische Einsatzbeispiele für eine optimale Backup-Effizienz und Kostensenkungen dargestellt.

Die Akzeptanz der Deduplizierung

Die Nachfrage nach der Datendeduplizierung wächst sowohl in mittelständischen als auch in Großunternehmen, denn dort wird nach Möglichkeiten gesucht, um mit dem Speicherbedarf Schritt zu halten, der sich in jedem Jahr fast verdoppelt. Dieses Wachstum wird verursacht von neuen Anwendungen, der Verbreitung der Virtualisierung, der Erstellung elektronischer Dokumentenspeicher, der gemeinsamen Nutzung von Dokumenten und Web 2.0-Technologien sowie der Aufbewahrung bzw. Vorhaltung digitaler Aufzeichnungen. Wenn IT-Budgets knapp sind, steigt die Notwendigkeit, das Datenwachstum unter Kontrolle zu halten, denn Unternehmen müssen ihre Kapital- und Betriebskosten senken. Auf der physischen Seite haben viele Rechenzentrumsmanager auch mit einer eingeschränkten Infrastruktur im Hinblick auf die Performance, Kühlung und die Stellfläche zu tun. Deduplizierung ist eine Technologie, die nicht nur zur Verbesserung der Speichereffizienz durch Kostensenkungen beiträgt, sondern auch physisch eingeschränkte Rechenzentren entlastet.

Die Deduplizierung löst auch Probleme in Verbindung mit Ineffizienzen in den Bereichen Management, Backup und Netzwerk. Wenn die Menge der Daten wächst, steigt das Missverhältnis zwischen der Anzahl der IT-Mitarbeiter und der Speicherkapazität, der gemanagt werden muss, überproportional an. Durch Deduplizierung wird die Datenmenge verringert, sodass dieses Verhältnis ausgewogen bleibt. Wenn sich in ähnlicher Weise die Lücke zwischen der Serververarbeitungsleistung und den Festplatten vergrößert, suchen Unternehmen nach Möglichkeiten, die Performance ihrer gesamten Umgebung über ein WAN, innerhalb von Festplattenspeicher-Subsystemen und über begrenzte Backup-Zeitfenster hinweg zu verbessern. Datendeduplizierungstechnologien können die verfügbare physische und virtuelle Infrastruktur optimieren, indem weniger Daten über lokale oder Remote-Netzwerkverbindungen gesendet werden. Ebenso können dadurch die Service-

Level-Reaktionszeiten verbessert und immer kleiner werdende Backup-Zeitfenster eingehalten werden. Deduplizierung verwendet auch Direktzugriffsmedien (Festplatten). Dies verbessert die Recovery-Zeiten, die Datensicherheit und die Zuverlässigkeit.

Die neuesten Herausforderungen resultieren aus der Virtualisierung. Immer mehr Unternehmen stellen die Virtual-Machine-Technologie bereit, um die Serverkonsolidierung und die Disaster Recovery (DR) zu unterstützen – und virtuelle Maschinen verarbeiten Daten, die hochredundant sein können, aber dennoch abgesichert werden müssen. Um verschiedene Ausfallszenarien zu berücksichtigen oder ein Image wiederherzustellen, sind in einer einzelnen Backup-Lösung und in einem einzelnen Backup-Prozess normalerweise ein physischer Server und separate Dateien erforderlich. Deduplizierung bietet deutliche Einsparungen bei der Backup-Speicherkapazität, da sie die normalerweise in VMDK-Dateien vorhandene Redundanz eliminieren kann. Standardansätze, wie die Bereitstellung eines herkömmlichen Backup-Agenten in einer virtuellen Gast-Maschine oder die Verwendung eines VCB-Proxy zum Erstellen eines Backups auf Image-Ebene, tragen in keiner Weise dazu bei, die Menge der Daten von virtuellen Maschinen, die gesichert werden müssen, oder die Anforderungen an die Netzwerkbandbreite zu reduzieren. Die Deduplizierung erfüllt in Verbindung mit Backup-Software die Anforderungen an eine vollständige, effektive und kosteneffiziente Absicherung von Umgebungen mit virtuellen Maschinen.

Die Vorteile der Deduplizierung

Unternehmen stellen Dateneduplizierung auf mehreren Ebenen der Infrastruktur bereit, um diese praktischen, realistischen Herausforderungen zu meistern. Zu den Vorteilen der Deduplizierung zählen:

- ☒ **Kostensenkungen.** Die Deduplizierung ermöglicht Ressourceneffizienz und Kosteneinsparungen durch die Verringerung des Energiebedarfs, der Kühlungs- und Stellflächenanforderungen im Rechenzentrum sowie der Speicherkapazität, Netzwerkbandbreite und der Personaleffizienz.
- ☒ **Verbesserung von Backup- und Recovery-Service-Levels.** Die Deduplizierung kann die Backup-Performance erheblich verbessern und ermöglicht die Einhaltung von begrenzten Backup-Zeitfenstern. Die Deduplizierungstechnologie nutzt außerdem Festplattenspeicher mit Direktzugriff, wodurch sich im Vergleich mit sequenziellen Zugriffsmethoden (Band) eine bessere Recovery-Performance ergibt.
- ☒ **Die neue Wirtschaftlichkeit: Festplatte im Vergleich zu Band.** Die Deduplizierung ermöglicht festplattenbasierte Backups für eine größere Anzahl von Anwendungen. Wegen ihrer Wirtschaftlichkeit und der Archivierungseigenschaften spielen Bänder in Rechenzentren von Unternehmen immer noch eine wichtige Rolle. Die Kosten pro Gigabyte für Festplatten, die in Verbindung mit Deduplizierung verwendet werden, sinken jedoch und führen zu Festplattenkosten, die unter den Kosten für Bänder liegen.

- ☒ **Reduzierung der Treibhausgasemissionen.** Die Deduplizierung verringert den Energieverbrauch sowie die Kühlungs- und Stellflächenanforderungen für Speicher und trägt so zur Senkung der Treibhausgasemissionen und zum verantwortlichen Umgang mit der Umwelt bei.

Durch die Deduplizierungstechnologie lassen sich viele der Herausforderungen im Bereich Backup bewältigen, mit denen große und kleine Unternehmen schon mehr als ein Jahrzehnt konfrontiert sind. Zu diesen Herausforderungen zählen beispielsweise die Verdoppelung des Datenwachstums, die Einhaltung kürzerer Backup-Zeitfenster sowie die Ermöglichung einer schnelleren Recovery bei betrieblichen und anderen Ausfällen.

Tabelle 1 führt die vielfältigen Herausforderungen im Bereich Backup auf und zeigt, wie diese durch die Deduplizierung gelöst werden können.

TABELLE 1	
Herausforderungen im Bereich Backup und Auswirkungen der Deduplizierung	
Backup-Herausforderungen	Auswirkungen der Deduplizierung
Backup-Zeitfenster werden kürzer, wenn der Betrieb rund um die Uhr läuft, um den Bedarf weltweiter Kunden zu erfüllen.	Bei herkömmlichen Backups werden riesige Mengen redundanter Daten übertragen, wodurch die Einhaltung von knappen Backup-Fenstern erschwert wird bzw. wofür nicht existente Backup-Fenster erforderlich wären. Deduplizierung in Backup-Software kann die zu sichernde Datenmenge reduzieren, und schnelle Inline-Deduplizierungsspeichersysteme können die Performance des Backup-Ziels steigern, da sie dafür sorgen, dass mehr Daten in einem verfügbaren Fenster gesichert werden können.
Die Recovery-Time-Anforderungen werden reduziert, um die Kosten von Ausfallzeiten zu minimieren.	Die Deduplizierung senkt die Kosten zur Speicherung von mehr Backup-Daten auf Festplatte. Wenn Backups auf Festplatte statt auf Band vorgehalten werden, verbessert dies die Recovery-Zeiten für eine große Anzahl von Anwendungen erheblich.
Die mangelnde Zuverlässigkeit von Backups gefährdet die Daten-Recovery.	Wenn man von Bandmedien für Backups abhängig ist, bringt dies Risiken wie Medienfehler (fehlerhafte Medien, verschmutzte Köpfe usw.), nicht mehr erhältliche Medien oder Hardwareausfälle mit sich. Die Deduplizierung verwendet Festplatten im Data-Protection-Prozess, sodass diese Fehlerquellen beseitigt oder verringert werden. Die Nutzung von Festplatten ermöglicht außerdem Statusüberprüfungen und andere Maßnahmen zur automatischen Reparatur und Fehlervermeidung.
Eine höhere Servervirtualisierung bedeutet, dass weniger Ressourcen für Backups verfügbar sind. Dies kann Backup-Zeiten verlängern und Backup-Fenster belasten.	Die Deduplizierung kann eingesetzt werden, um die Verarbeitung redundanter Daten durch gemeinsame Ressourcen zu eliminieren, physische Ressourcen zu entlasten und Backups von virtuellen Maschinen zu beschleunigen. Die Deduplizierung ermöglicht außerdem eine längere Aufbewahrung der Backup-Daten von virtuellen Maschinen mit geringerem Speicherplatzbedarf, sodass eine Recovery schnell von einer Festplatte statt von Band stattfinden kann.
Datenwachstum kann bedeuten, dass nicht alle Daten innerhalb verfügbarer Backup-Fenster gesichert werden können.	Unternehmen sehen sich einem jährlichen durchschnittlichen Datenwachstum von 50 % gegenüber. Alle diese Daten müssen abgesichert und geschützt werden. Dieses Wachstum ist mit begrenzten nächtlichen Backup-Zeitfenstern und herkömmlichen Methoden nicht zu bewältigen. Die Deduplizierung löst die Herausforderungen dieses Wachstums und ermöglicht ein effizientes Backup der wachsenden Datenmengen.

TABELLE 1

Herausforderungen im Bereich Backup und Auswirkungen der Deduplizierung

Backup-Herausforderungen	Auswirkungen der Deduplizierung
Sichere externe Kopien unter Nutzung herkömmlicher Bandmethoden setzen Daten dem Risiko von Verlust oder Diebstahl aus.	Wenn man sich bei einem Notfall auf extern aufbewahrte Wechselmedien (Band) verlassen muss, entsteht dadurch ein Risiko für die physischen Medien. Die Deduplizierung ermöglicht in Verbindung mit sicheren Replikationsprozessen die externe Aufbewahrung einer elektronischen Kopie, sodass die Notwendigkeit der manuellen Handhabung von Bandmedien entfällt und damit die Sicherheit erhöht wird.
Verteilte Daten an entfernt gelegenen Standorten benötigen zentralen Schutz und zentrale Recovery.	Entfernt gelegene Standorte ersetzen eigenständige Bandsicherungsvorgänge durch einen zentralisierten „Edge-to-Core“-Ansatz, um Backup, Recovery und Management zu verbessern. Durch die Deduplizierung wird das Senden dann geringerer Backup-Datenmengen über hoch belastete WAN-Verbindungen an ein zentrales Rechenzentrum sinnvoll.
Backup-Infrastrukturkosten steigen , um mit dem Kapazitätswachstum und den Backup-Zeitfenstern Schritt zu halten.	Die meisten Unternehmen lösen Herausforderungen in Verbindung mit dem Datenwachstum und den Backup-Zeitfenstern, indem sie die Bandinfrastruktur vergrößern. Zusätzliche Bandlaufwerke und Automatisierung können momentane Performance-Engpässe beseitigen und Backups sind zunächst schneller durchführbar, doch geht dies zu Lasten der Kosten und des Management-Overheads. Die Deduplizierung löst das Problem von Grund auf, um die laufenden Kosten für die Bandinfrastruktur zu reduzieren und gleichzeitig mit dem Kapazitätswachstum und der Trendentwicklung der Backup-Fenster Schritt zu halten.

Quelle: IDC, 2010

DEDUPLIZIERUNG: WAS, WO, WANN UND WIE

Was ist Deduplizierung?

IDC definiert die Dateneduplizierung als eine Technologie, die doppelt vorhandene Daten in ein einziges gemeinsames Datenobjekt normalisiert, um Speicherkapazitätseffizienz zu erzielen. Genauer gesagt, bezieht sich Dateneduplizierung auf jeden Algorithmus, der nach doppelten Daten (z. B. Blöcke, Dateien, Segmente) sucht und Duplikate verwirft, sobald diese gefunden werden. Wenn Duplikate erkannt werden, werden sie nicht aufbewahrt. Stattdessen wird ein „Data Pointer“ modifiziert, mit dessen Hilfe das Speichersystem auf eine exakte Kopie des bereits auf der Festplatte gespeicherten Datenobjekts verweist. Außerdem vermindert die Dateneduplizierung das Problem der Kosten für mehrere Kopien desselben Datenobjekts.

Dateneduplizierung wird am häufigsten mit Subdatei-Vergleichsprozessen in Zusammenhang gebracht. Dies unterscheidet sich vom Single-Instance-Speicher (SIS), der Daten auf Datei- oder Objektebene vergleicht. Bei der Deduplizierung von Subdateien wird eine Datei untersucht und in „Segmente“ aufgeteilt. Diese kleineren Segmente werden dann auf das Auftreten redundanter Dateninhalte über mehrere Systeme und Standorte hinweg untersucht. Die Deduplizierung unterscheidet sich auch von der Komprimierung, bei der die Größe eines einzelnen Objekts und nicht die Anzahl der Dateien oder von Teilen einer Datei reduziert wird. Deduplizierte Daten können zusätzlich komprimiert werden, um noch mehr Speicherplatz einzusparen.

Wo erfolgt die Deduplizierung?

Die Deduplizierung von Backup-Daten kann an der Quelle oder am Ziel erfolgen. Ein Beispiel für die Deduplizierung an der Quelle wäre die Verringerung der Größe von Backup-Daten auf dem Client (z. B. Exchange- oder Dateiserver), sodass während des Backup-Prozesses nur einzigartige Subdateidaten über das Netzwerk übertragen werden. Mit der Deduplizierung an der Quelle verfügt die Backup-Anwendung über eine Deduplizierungstechnologie, die in ihrer Architektur eingebettet ist. Ein Beispiel für Deduplizierung am Ziel wäre die Verringerung der Größe von Backup-Daten, nachdem diese über das lokale Netzwerk übertragen wurden und ein Deduplizierungsspeichersystem erreicht haben. Die Deduplizierung an der Quelle bietet Einsparungspotentiale in den Bereichen Netzwerkbandbreite, Backup-Zeitfenster und Speicherplatz. Bei der Deduplizierung am Ziel verfügt das Speichersystem über eine Deduplizierungstechnologie, die im Speicher-Controller eingebettet ist. Die Deduplizierung am Ziel ermöglicht Speichereinsparungen, funktioniert zusammen mit vorhandener Backup-Software und kann die WAN-Auswirkungen der Replikation reduzieren. Die Deduplizierung hat nicht nur die oben erwähnten Vorteile, sondern wirkt sich auch positiv auf die Implementierungszeiten und -kosten aus. Unternehmen sollten ihre aktuellen Backup-Probleme analysieren und diese den unterschiedlichen Deduplizierungsmethoden zuordnen.

Deduplizierung an der Quelle

Wenn die Deduplizierung an der Quelle (oder mit Client-Backup-Software) erfolgt, hat dies außer der Speicherkapazitätsoptimierung eine Reihe von weiteren Vorteilen. Es werden wesentlich weniger Daten vom Quellgerät zum Speicher-

Repository übertragen, was überlastete virtuelle/physische Infrastrukturen und LAN-/WAN-Verbindungen entlastet. Da nur neue oder geänderte Subdateidatensegmente vom Quellgerät zum Speicher-Repository übertragen werden, ist die übertragene Datenmenge erheblich geringer. Dadurch sind äußerst schnelle tägliche komplette Backups möglich. Der inkrementelle Overhead auf der Client-CPU zur Durchführung der Deduplizierung an der Quelle kann um bis zu 15 % höher sein, doch das Backup wird dafür viel schneller abgeschlossen als bei herkömmlichen Methoden – und einige Architekturen stellen Drosselmechanismen bereit, um kurzfristige Overhead-Zunahmen zu bewältigen. Die Auswirkungen der Deduplizierung an der Quelle sind insgesamt tatsächlich wesentlich geringer als bei Verwendung herkömmlicher Agenten über einen Zeitraum von sieben Tagen. Die Deduplizierung an der Quelle ermöglicht auch eine flexible Bereitstellung, weil kleinere, entfernt gelegene Standorte nur einen Software-Backup-Agenten bereitstellen müssen. In Umgebungen mit sehr großen Datenbanken oder Datenbanken mit vielen täglichen Änderungen sollte stattdessen eine Deduplizierung am Ziel in Betracht gezogen werden. Anbieter verfügen normalerweise über Datenbewertungs-Tools, um Kunden bei der Entscheidungsfindung zu unterstützen.

Deduplizierung am Ziel

Bei der Deduplizierung am Ziel wird die Backup-Festplattenspeicherkapazität optimiert, da auf der Festplatte nur neue, einzigartige Subdateidaten gespeichert werden. Alle Backup-Daten werden weiterhin an das Deduplizierungsziel unter Nutzung herkömmlicher Backup-Software gesendet. Dies ermöglicht eine nahtlose Integration in die vorhandene IT-Infrastruktur. Verfügbare Backup-Fenster werden nur dann entlastet, wenn das vorherige Backup-Ziel, normalerweise Bänder, die Engstelle der Performance der Backup-Lösung war. Bei der Deduplizierung am Ziel führt das Speichersystem (auch als Deduplizierungsspeichersystem bezeichnet) die Deduplizierung durch, um die Data Protection und die Performance der Disaster Recovery zu optimieren, während die Anwendungsserver vom Deduplizierungsprozess entlastet werden. Die Deduplizierung am Ziel ist einfach zu implementieren, indem ein schnelles, anwendungsunabhängiges Speichersystem erstellt wird (als NAS [Network-Attached Storage] über Ethernet oder als VTL [Virtual Tape Library] über Fibre Channel anzubinden). Client-Software oder andere Konfigurationen sind nicht erforderlich. Deduplizierungsspeichersysteme werden häufig bei größeren Datensätzen und Datenbanken verwendet. Die Deduplizierung am Ziel kann auch in zentralen Rechenzentren für große Datenmengen und an entfernt gelegenen Standorten für lokale Backups, gefolgt von einer Replikation auf ein zentrales Rechenzentrum, verwendet werden.

Wann erfolgt die Deduplizierung?

Es gibt heute zwei unterschiedliche Methoden, um zu bestimmen, *wann* der Deduplizierungsprozess erfolgen soll: Inline oder Post-Process. Die Inline-Deduplizierung beseitigt redundante Daten, bevor sie auf die Festplatte geschrieben werden, sodass kein Festplatten-Staging-Bereich erforderlich ist. Bei der Post-Process-Deduplizierung werden die Daten analysiert und reduziert, nachdem sie auf der Festplatte gespeichert wurden. Daher ist ein Staging-Bereich mit voller Kapazität erforderlich, auf dem ein Deduplizierungsprozess gestartet wird. Bei der Auswahl einer dieser Methoden sollten Unternehmen die Backup-Geschwindigkeit und die Festplattenkapazität berücksichtigen.

Die Inline-Deduplizierung stellt eine Deduplizierungsmethode dar, die eine schnellere Effizienz und Wirtschaftlichkeit ermöglicht. Sie reduziert die im System erforderliche „Rohkapazität“ der Festplatte erheblich, da der komplette, noch nicht deduplizierte Datensatz nie auf Festplatte geschrieben wird. Wenn Replikation als Teil des Inline-Deduplizierungsprozesses unterstützt wird, optimiert die Inline-Deduplizierung im Vergleich zu anderen Methoden auch die Recovery deutlich, da das System nicht darauf warten muss, den gesamten Datensatz zu absorbieren und dann zu deduplizieren, bevor mit der Replikation am entfernt gelegenen Standort begonnen werden kann.

Die Post-Process-Deduplizierung erfordert eine Wartezeit, bis sich die Daten auf der Festplatte befinden, bevor der Deduplizierungsprozess gestartet wird. Dieser Ansatz erfordert eine größere anfängliche Kapazität als Inline-Lösungen. Außerdem verursacht ein Post-Process-Ansatz eine Verzögerung bis zum Abschluss der Deduplizierung und der Replikation. Außerdem besteht die Gefahr der Inkonsistenz zwischen einem lokalen und einem entfernt gelegenen System, da es zwei Speicherzonen gibt, jeweils mit Policies und einem Status, die gemanagt werden müssen.

Wie erfolgt die Deduplizierung?

Der Ablauf des Deduplizierungsprozesses hängt von der Implementierung ab. Bei einer Hash-basierten Deduplizierungsmethode wird eine Datei oder ein Backup-Stream in Blöcke fester oder variabler Länge mit Subdateidaten aufgeteilt. Für jedes Segment wird ein Hash-Wert berechnet. Dieser Prozess berechnet eine eindeutige Zahl für jedes Segment, die dann in einem Index gespeichert wird. Wenn eine Datei aktualisiert wird, werden nur die geänderten Subdateidaten gespeichert; Änderungen erfordern nicht, dass eine völlig neue Datei gespeichert wird. Ein bedeutender Unterschied bei Hash-basierten Implementierungen besteht darin, ob die Segmentgröße fest oder variabel ist. Ein Ansatz mit variabler Länge kann die Segmentgröße je nach Content-Typ dynamisch anpassen, um redundante Datensegmente zu berücksichtigen, deren Position sich bei Änderung einer Datei in einem Byte-Stream verschoben hat. Bei einem Ansatz mit fester Länge werden redundante Daten, deren Position sich geändert hat, nicht erkannt. Segmente werden daher erneut gesichert, da sie eindeutig zu sein scheinen, obwohl sich diese bereits im Backup-Repository befinden. Dies ist ineffizient. Der Hash-Index wird im Speicher verwaltet. Wenn der Hash-Index jedoch größer wird, läuft er möglicherweise vom Speicher auf die Festplatte über, sodass Festplatten-I/O für die Suche von Hash-Werten erforderlich ist. Die Anbieter lösen diese praktischen Technologieherausforderungen auf unterschiedliche Weise. Die Ergebnisse reichen vom Eliminieren des Problems bis zu einer deutlich schlechteren Performance.

Ein alternativer Ansatz ist die Delta-basierte Dateneduplizierung (auch als Delta-Differenzierung oder Delta-Codierung bezeichnet), bei der die Unterschiede zu einer Basiskopie gespeichert oder übertragen werden. Die Baseline ist eine komplette Kopie der Daten, die zur Neuerstellung weiterer Versionen der Daten verwendet wird. Die Delta-basierte Dateneduplizierung kann auf Block- oder Byte-Ebene erfolgen. Statt einen Hash zum Ermitteln neuer Daten im Netz zu verwenden, scannt und indiziert eine Delta-Differenzierungsmethode den eingehenden Datenstrom und sucht nach Daten, die den bereits gespeicherten Daten ähnlich sind. Zu den Vorteilen des Delta-basierten Ansatzes gehört eine geringe CPU-Beanspruchung, da kein starker Hash berechnet werden muss. Ein Delta-Differenzierungsprozess erfordert allerdings viele Festplatten-I/O-

Vorgänge zum Vergleich der alten Daten mit den eingehenden neuen Daten. Darum hängt der langfristige Vorteil der einzelnen Ansätze möglicherweise von den relativen Performance-Verbesserungen der CPU im Vergleich zur Festplattentechnologie ab.

Ein weiterer Faktor, der sich auf die Deduplizierungsrate auswirken kann, ist die Frage, ob ein Deduplizierungs-Engine Markierungen, die von der Backup-Anwendung in den Datenstrom eingefügt wurden, oder bestimmte Datenformate (z. B. eine Backup-Anwendung, Microsoft Exchange-Daten usw.) erkennen kann. Die Fähigkeit, Markierungen und das Datenformat zu erkennen, erfordert ein Verständnis dafür, wo anwendungsspezifische Metadaten in einen Stream einfließen. Wenn der Deduplizierungs-Engine die Markierungs-Offsets und das Format der Daten erkennt, kann der Engine die Segmentgröße so anpassen, dass sie ideal für das Datenformat der ursprünglichen Anwendung ist, was zu potenziell besseren Deduplizierungsergebnissen führt. Die Nutzung dieses Ansatzes erfordert allerdings, ein Verständnis für die sich ändernden Formate jeder einzelnen Backup-Anwendung (NetWorker, NBU, TSM usw.) und jeder Benutzer-Anwendung (Oracle, Exchange usw.) zu entwickeln und aufrecht zu erhalten.

EVALUIERUNG VON DEDUPLIZIERUNGSTECHNOLOGIEN

Heute sind unterschiedliche Produkttypen mit Deduplizierungsfunktionen erhältlich. Backup-Anwendungen, Appliances, virtuelle Bandbibliotheken, WAN-Optimierungslösungen und primäre Festplattenspeicher-Subsysteme können alle über Deduplizierungsfunktionen verfügen. Bevor sich ein Unternehmen für eine bestimmte Art der Deduplizierung entscheidet, muss geklärt werden, welche Probleme je nach Anwendung oder Datentyp zu lösen sind. Unterschiedliche Deduplizierungsansätze haben unterschiedliche Vorteile in Bezug auf die Kapazität, die Performance und die Netzwerkeffizienz.

1. **Deduplizierungsraten.** Die erzielte Deduplizierungsrate ist von einer Vielzahl von Faktoren abhängig, einschließlich Datentypen, Datenänderungsraten, Aufbewahrungszeiträumen, Segmenten mit variabler oder fester Länge, Backup-Policys, Dateiformaterkennung u. Ä. IDC-Untersuchungen weisen bei Back-End-Festplattenspeicherung auf realistische Deduplizierungsraten zwischen 8:1 und 22:1 hin, die auf den oben erwähnten Faktoren basieren. Die Lösungen für die Deduplizierung an der Quelle können die erforderliche tägliche Netzwerkbandbreite im Vergleich zu herkömmlichen Methoden für tägliche komplette Backups erheblich reduzieren. Wie bei allen Performance-Kennzahlen unterscheidet sich das tatsächliche Ergebnis je nach Umgebung. Unternehmen sollten sich nicht auf Durchsatz-, Skalierungs- oder Performance-Versprechen verlassen und die Deduplizierung vor Ort mit ihren eigenen Datenmengen testen.
2. **Die Bedeutung von Komprimierung, Verschlüsselung und Multiplexing.** Die Komprimierung – die Codierung von Daten, um die erforderliche Speicherkapazität zu reduzieren – kann die Deduplizierung ergänzen. Die Komprimierung ist für ein einzelnes Objekt optimiert und reduziert dessen Größe, während die Deduplizierung objektübergreifend arbeitet. Die Komprimierung kann jedoch auf Daten angewendet werden, die bereits dedupliziert wurden, um noch mehr Speicherplatz zu sparen. Wenn jedoch

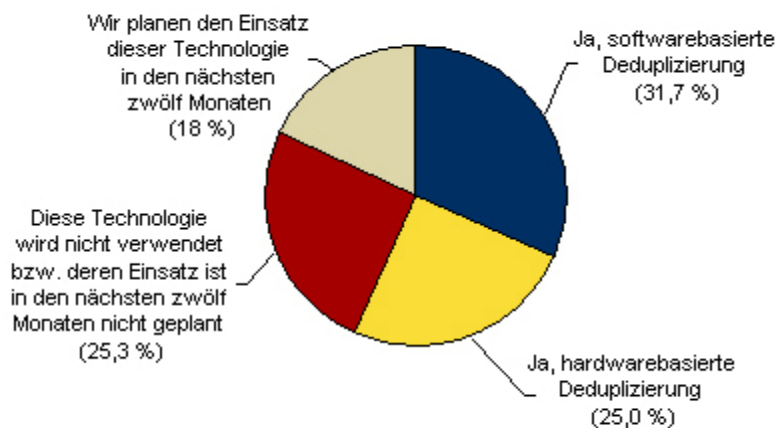
die Deduplizierung auf eine bereits komprimierte (oder verschlüsselte) Datei angewendet wird, sind die Vorteile der Deduplizierung vernachlässigbar bzw. nicht vorhanden, es sei denn, die komprimierte Datei wird erneut gesichert. Unternehmen, die Deduplizierung zusammen mit der Komprimierung verwenden, erzielen möglicherweise zusätzliche Vorteile, wenn zuerst die Deduplizierung der Daten stattfindet. Außerdem muss die aktuelle Praxis berücksichtigt werden, für Backups Multiplexing zu verwenden, bei dem Daten von mehreren Clients kombiniert und in einem einzigen Stream an ein Bandlaufwerk gesendet werden. Bei diesem Prozess ist es jedoch schwierig, Datensegmente zu erkennen, die bereits vorhanden sind. Multiplexing ist eine Funktion, die in den meisten Backup-Anwendungen verwendet wird, um den Shoeshining-Effekt zu vermeiden und die Performance zu verbessern, wenn Daten auf ein Bandmedium geschrieben werden. Multiplexing muss deaktiviert werden, wenn Unternehmen von der Deduplizierung profitieren möchten. In diesem Szenario werden durch das Deaktivieren des Multiplexing die Performance-Gewinne, die durch aktiviertes Multiplexing entstanden sind, allerdings nicht aufgehoben.

3. **Deduplizierung bei virtuellen Maschinen.** Die Verwendung von virtuellen Maschinen in der Produktion hat die Notwendigkeit verstärkt, die virtuelle Maschine, den physischen Server und Dateien abzusichern und wiederherzustellen. Zu den Optionen für Backups virtueller Maschinen gehört das Backup auf Image- und/oder Dateiebene unter Nutzung eines Agenten, der auf einem Guest oder einer Servicekonsole ausgeführt wird oder der die Backup-API eines Virtualisierungsanbieters nutzt. Herkömmliche Backup-Lösungen sind beim Backup virtueller Maschinen ineffizient, weil sie große Mengen redundanter Daten übertragen und zur Ausführung eines Backups sehr viele CPU-Zyklen benötigen. Daraus ergibt sich eine schlechte Backup-Performance und eine begrenzte Serverkonsolidierung. Die Deduplizierung kann diese Einschränkungen beseitigen. Die Deduplizierung an der Quelle bedeutet, dass Duplikate niemals über die zu Grunde liegende, gemeinsam genutzte physische Infrastruktur übertragen werden. Tägliche Komplett-Backups erfolgen so schnell und effizient. Die Deduplizierung kann auch global über VMDKs hinweg erfolgen, um Backups redundanter Daten über mehrere virtuelle Systeme hinweg zu vermeiden. Auf diese Weise wird die Fähigkeit von Anwendern, virtuelle Maschinen ohne Band wiederherzustellen, erheblich verbessert und eine effektive Disaster Recovery unter Verwendung der Replikationsfähigkeiten von Deduplizierungssystemen wird bereitgestellt.
4. **Deduplizierung für entfernt gelegene Standorte.** Genau wie Rechenzentren erfordern entfernt gelegene Standorte sowohl lokale als auch (Remote) Disaster Recovery. Die Charakteristiken von entfernt gelegenen Standorten bringen jedoch ganz neue Herausforderungen mit sich. Entfernt gelegene Standorte haben in der Regel eine beschränkte WAN-Bandbreite, kein dediziertes IT-Personal und eine überproportionale Anzahl von Zweigstellen im Vergleich zu regionalen oder Hauptrechenzentren. Die Deduplizierung kann die Datenübertragung über das WAN minimieren und beseitigt sogar redundante Daten über Zweigstellen und Rechenzentren hinweg. Weil IT-Personal an entfernt gelegenen Standorten rar ist, möchten Unternehmen möglichst wenig Speicherhardware an verteilten Standorten einsetzen. Die Deduplizierung an der Quelle kann durch Software bereitgestellt werden und trägt so zur Lösung dieser Herausforderung bei.

Kleinere Deduplizierungssysteme können auch an entfernt gelegenen Standorten bereitgestellt werden, wenn eine lokale Recovery benötigt wird, und einige Anbieter unterstützen das Replizieren von solchen Systemen auf ein zentrales Rechenzentrum. Eine aktuelle IDC-Studie hat sich mit dem Einsatz der Deduplizierung an entfernt gelegenen Standorten und dem verwendeten Typ (siehe Abbildung 1) befasst.

ABBILDUNG 1

Verwendung von Deduplizierungstechnologien für die Absicherung von Daten von entfernt gelegenen Standorten



n = 300

Quelle: IDC Remote Branch Special Study, 2009

5. Deduplizierung für Produktions-/Disaster-Recovery-Rechenzentren.

Große Rechenzentren haben immer noch Probleme, Backup-Zeitfenster zumindest für einige ihrer Anwendungen einzuhalten und können es sich nicht leisten, die Backup-Performance zu beeinträchtigen. Dies könnte einen Deduplizierungsansatz rechtfertigen, der je nach Anwendung und Umgebung die Deduplizierung an der Quelle und am Ziel beinhaltet. Die Optimierung der Netzwerkbandbreite innerhalb eines Rechenzentrums ist eventuell weniger wichtig als die Remote-Replikation auf einen Disaster-Recovery-Standort. Wenn jedoch die Backup-Zeitfenster weiter schrumpfen, wird die Netzwerkbandbreite mit der Zeit zu einem Problem.

6. Deduplizierung und Replikation.

Replikation kann realistisch als nächste Stufe der Deduplizierungstechnologie angesehen werden. Etablierte Anbieter haben bewiesen, dass sie funktioniert, und Anwender, die die Technologie bewertet haben, sind begeistert und fragen danach. Die Deduplizierung wird in Enterprise-Umgebungen sowohl in peripheren als auch Core-Standorten eingesetzt und steigert die Effizienz bei gleichzeitiger Senkung der Infrastrukturkosten. Die Verwendung der Remote-Replikation wird besonders wichtig, da immer mehr Unternehmen den Einsatz von Bändern an entfernt gelegenen Standorten minimieren, sie aber immer noch

an einem zentralisierten Standort für Archivierungs- und Compliance-Zwecke unterstützen müssen. Die Anforderungen von Anwendern an die Replikation werden immer komplexer und beinhalten Folgendes:

- ❑ **Deduplizierungsorientierte Replikation, die einen deduplizierten Datensatz und nicht ein komplettes Volume repliziert.** Einige Hersteller bieten Replikationsservices zusammen mit einem deduplizierungsfähigen Produkt an. Unternehmen müssen sich jedoch vergewissern, dass die Replikationsfunktion deduplizierungsorientiert ist.
- ❑ **„Alles oder Nichts“ und Replikation auf Verzeichnis-/Bandebene.** Einige Einsatzbereiche benötigen eine komplette Systemreplikation, während für andere flexibel festgelegt werden soll, welche freigegebenen Verzeichnisse oder virtuellen Bänder repliziert werden sollen.
- ❑ **Replikationsüberwachung, Performance-Tuning und Fehlerbehebung.** Trotz der Deduplizierung müssen in den meisten großen Unternehmen weiterhin sehr viele Daten repliziert werden. Dies wird mithilfe eines geplanten oder asynchronen Replikationsprozesses gemanagt, der den Replikationsprozess und die verwendete Bandbreite überwacht. Tuning- und Fehlerbehebungs-Tools tragen dazu bei, dass der Replikationsprozess innerhalb des verfügbaren Replikationsfensters abgewickelt wird.
- ❑ **Geplante und Echtzeitreplikation für Verbindungen mit hoher und geringerer Latenz.** Einige Verbindungen/Standorte erfordern Echtzeitreplikation, während für andere ein geplanter Replikationsprozess angemessen sein kann. Die Charakteristiken von entfernt gelegenen Zweigstellen variieren erheblich und können über Verbindungen mit geringer Latenz verfügen, während Verbindungen zwischen zwei Rechenzentren dieses Problem nicht unbedingt aufweisen müssen.

7. **Seeding und Migration.** Die Deduplizierung ist zwar sehr gut für die Reduzierung von Speicher und/oder der übertragenen redundanten Daten geeignet, doch muss dazu erst eine Baseline oder ein erstes Backup erstellt werden. Für die Edge-to-Core-Deduplizierung und -Replikation müssen Anwender überlegen, wie diese Baseline über Verbindungen mit beschränkter Bandbreite erstellt werden kann. Die meisten Anbieter stellen eine Form von Seeding-Service bereit, um diese Baseline schnell zu erstellen. Dies erfolgt entweder über einen deduplizierungsorientierten Bulk-Replikationsprozess mit parallelen Systemen oder unter Nutzung einer Reihe von Bändern aus dem letzten kompletten Backup, die lokal auf einem Deduplizierungssystem wiederhergestellt werden. Andere Überlegungen berücksichtigen angesichts von Speichererneuerungsraten in drei- oder fünfjährigem Rhythmus, wie eine Migration erfolgt und wie sehr sie eine vorhandene Umgebung beeinträchtigt.
8. **Anbieterauswahl.** Anbieter machen viele Versprechungen und Aussagen bezüglich ihres Deduplizierungsansatzes. IDC-Untersuchungen zeigen, dass nicht alle verfügbaren Deduplizierungsprodukte wie beworben funktionieren. Unternehmen müssen berücksichtigen, wie lange ein bestimmtes deduplizierungsfähiges Produkt bereits am Markt ist, wie viele Kunden das Produkt in der Produktion verwenden und wie ausgereift das Produkt in realistischen Umgebungen ist. Unternehmen sollten die Skalierbarkeit eines

Produkts umfassend untersuchen. Fragen Sie nach Referenzen von Nutzern und einer Anwendungs- und/oder System-Support-Matrix. Unternehmen, die kein Proof-of-Concept (POC) durchführen, gehen das Risiko ein, später Überraschungen in den Bereichen Performance und Zuverlässigkeit zu erleben.

9. **Einsatzbereiche für die Deduplizierung.** Die Deduplizierung ist eine Technologie, die verspricht, das Speicherinfrastrukturniveau weiter zu heben. Bisher wurde diese Technologie in großem Umfang in Backup-Szenarien eingesetzt, wo bereits eine große Menge redundanter Daten vorhanden ist. Die gleichen Daten werden jede Woche gesichert, was unnötige Server-, Netzwerk- und Speicherressourcen beansprucht. Einige Unternehmen untersuchen oder testen jetzt die Deduplizierung in primären Speicherumgebungen innerhalb eines NAS-Ansatzes. Diese Implementierung erfordert jedoch eine verbesserte Performance, um Auswirkungen auf die Latenz und die Reaktionszeiten zu vermeiden. Heute ist die Deduplizierungstechnologie bereits gut geeignet für das Backup von virtuellen Maschinen, entfernt gelegenen Standorten und Rechenzentrumsumgebungen.

DAS EMC PORTFOLIO MIT DEDUPLIZIERUNGSLÖSUNGEN

EMC bietet ein breites Spektrum an Backup- und Recovery-Produkten und -Services, um Kunden bei der Senkung der IT-Kosten und der Steigerung der Backup-Effizienz zu unterstützen. Backup-Deduplizierungslösungen umfassen die Deduplizierungs-Backup-Software EMC Avamar, die Deduplizierungsspeichersysteme EMC Data Domain und EMC NetWorker, die zusammen mit Avamar, Data Domain und anderen Zielsystemen von Drittanbietern bereitgestellt werden können. Außerdem bietet EMC mit dem NAS-System EMC Celerra eine Deduplizierungslösung für Primärspeicher- und Backup-Daten und mit der Produktreihe Centera eine Deduplizierungslösung für Festplattenarchive an. Diese werden jedoch im Rahmen dieses Dokuments nicht behandelt.

EMC Avamar

Die Deduplizierungs-Backup-Software EMC Avamar umfasst integrierte Deduplizierungstechnologie, um redundante Daten an der Quelle zu identifizieren und so die Menge der Backup-Daten zu verringern, bevor sie über das LAN/WAN gesendet werden. Mit Avamar profitieren Unternehmen von Datenreduzierungen und schnellen, täglichen Komplett-Backups für VMware-Umgebungen, entfernt gelegene Standorte, Desktops/Laptops und LAN- und NAS-Server in Rechenzentren. Avamar dedupliziert außerdem Backup-Daten über Standorte und Server hinweg und im Laufe der Zeit. Im Gegensatz zu Produkten, die herkömmliche Recovery-Methoden nutzen, kann Avamar Daten schnell in einem einzigen Schritt wiederherstellen und vermeidet so den Aufwand, das letzte intakte komplette und nachfolgende inkrementelle Backups wiederherstellen zu müssen, um den gewünschten Recovery-Punkt zu erreichen. Die Fähigkeiten von Avamar unterscheiden sich fundamental von herkömmlichen Backup-Anwendungen.

Der Avamar-Agent merkt sich neue und geänderte Dateien. Er muss nicht die gesamte Dateisystemhierarchie durchlaufen, um neue oder geänderte Daten zu erkennen, sondern überprüft zunächst die lokale Cache-Datei auf entsprechende Daten. Nach der Identifizierung unterteilt der Agent die neuen oder geänderten Dateien in Subdateisegmente variabler Länge und weist jedem Segment einen Hash-Wert (eindeutige ID) zu. Der Agent überprüft dann den lokalen Hash-Cache, um festzustellen, ob der Hash bereits vorhanden ist und zuvor gesichert wurde. Falls er gesichert wurde, wird er nicht erneut gesichert. Schließlich kommuniziert der Agent mit dem Avamar-Server, um zu bestimmen, ob der Hash-Wert eindeutig oder bereits vorhanden ist. Wenn das Datensegment neu ist, wird es im Verlauf des täglichen kompletten Backups über das LAN/WAN übertragen.

Diese Prozesse steigern die CPU-Auslastung auf dem Server im Vergleich zu einem herkömmlichen Backup-Agenten. Weil das Backup jedoch effektiv nur neue Datensegmente auf dem Netzwerk sichert, sind Avamar-Backups erheblich schneller abgeschlossen als herkömmliche komplette und inkrementelle Backups. So könnte beispielsweise ein inkrementelles Backup, das in der Regel zehn Stunden benötigt, mit Avamar nur noch bis zu eine Stunde benötigen und dadurch die wöchentliche Beeinträchtigung durch inkrementelle Backups von Montag bis Freitag von 50 auf fünf Stunden senken. Außerdem erfolgen die täglichen kompletten Backups mit Avamar um ein Vielfaches schneller als herkömmliche komplette Backups.

Darüber hinaus bietet Avamar jetzt Desktop-/Laptop-Absicherung, die im Hintergrund aktiv ist, vorhandene Netzwerkverbindungen verwendet und wichtige CPU-Zyklen nicht belastet. Benutzerdaten werden bei der Anmeldung während der normalen Backup-Fenster automatisch gesichert. Backups können aber auch vom Anwender gestartet werden, sodass Anwender ihre Daten bei Bedarf wiederherstellen können.

Avamar-Backup- und Recovery-Lösungen ermöglichen eine Deduplizierung sowohl an der Quelle als auch global, sodass sie für Unternehmen mit folgenden Problemstellungen besonders geeignet sind:

- Bereitstellung virtueller Maschinen und Evaluierung einer neuen Strategie für die Absicherung, um physische Server, virtuelle Server und eigenständige Objekte wiederherzustellen
- Verbesserung von Backups ihrer entfernt gelegenen Standorte, um schnelle, tägliche Komplet-Backups, ein zentrales Management, eine höhere Zuverlässigkeit, eine sicherere Replikation und einen reduzierten Backup-Traffic über belastete WAN-Verbindungen zu erreichen
- Verringerung des Datenwachstums, von Backup-Zeitfenstern und des Netzwerk-Traffics für Backups lokaler NAS- und Dateiserver-Umgebungen
- Absicherung wichtiger Desktop-/Laptop-Daten, auch in Filialen und für mobile Mitarbeiter

Die EMC Avamar-Bereitstellungskonfigurationen sind wie folgt:

- Client-Optionen

- ❑ Avamar-Softwareagenten werden auf den zu abzusichernden Systemen (Clients) bereitgestellt, ohne dass zusätzliche lokale Hardware (z. B. Medienserver) erforderlich ist. Agenten gibt es für die meisten Betriebssysteme und für führende Anwendungen und Datenbanken. Im Gegensatz zu vielen Backup-Anwendungen entstehen bei Avamar keine Kosten für die Client-Software. Avamar verwendet stattdessen eine deduplizierte, kapazitätsbasierte Lizenzierung. Dies kann zu sehr kostengünstigen Implementierungen und nachfolgenden Client-Erweiterungen führen.

☒ Serveroptionen (Backup-Repository)

- ❑ Avamar-Server von Drittanbietern. Avamar-Software kann für eine Reihe zertifizierter Server gemäß Branchenstandard mit internem Festplattenspeicher erworben und bereitgestellt werden.
- ❑ Avamar Data Store. Diese skalierbare All-in-One-Lösung beinhaltet auf EMC Hardware vorinstallierte und vorkonfigurierte Avamar-Software und vereinfacht so die Bestellung, Bereitstellung und den Service.
- ❑ Avamar Virtual Edition for VMware. Diese Branchenneuheit ermöglicht die Bereitstellung eines Avamar-Servers als virtuelle Appliance auf einem vorhandenen ESX-Server und kann so die angeschlossenen Ressourcen und Festplattenspeicher nutzen.

Avamar unterscheidet sich von anderen, auf dem Markt verfügbaren Ansätzen für die Deduplizierung an der Quelle. So verwendet die Deduplizierung mit Avamar Subdateisegmente variabler Länge, die eine höhere Effizienz und Performance ermöglichen. Avamar nutzt eine Grid-Architektur zur Skalierung der Performance und der Kapazität, wobei jeder zusätzliche Node CPU, RAM, I/O und Speicher des gesamten Systems vergrößert.

Das Avamar-Grid verwendet eine Konfiguration mit einem redundanten Array unabhängiger Nodes (Redundant Array of Independent Nodes – RAIN). Dies sorgt für Fehlertoleranz und hohe Verfügbarkeit im gesamten Grid und beseitigt Single-Points-of-Failure. Avamar verteilt seinen internen Index auf alle Avamar-Nodes und verbessert so die Zuverlässigkeit, den Lastausgleich und die Skalierbarkeit. Außerdem prüft Avamar jeden Tag automatisch, ob Backup-Daten vollständig wiederhergestellt werden können, und der Avamar-Server prüft sich selbst zweimal täglich, um für die Serverintegrität zu sorgen. Und schließlich unterstützt Avamar ein breites Angebot an Anwendungen und Clients, einschließlich Exchange, SQL Server, Oracle, DB2, SharePoint, Lotus Notes und NDMP.

Avamar bietet eine Reihe von Möglichkeiten, um virtuelle und physische Maschinen abzusichern. Optionen für das Backup mit Avamar in Umgebungen mit virtuellen Maschinen von VMware sind:

- ☒ **Avamar-Agent im Gast-Betriebssystem.** Ein Avamar-Agent innerhalb jedes Gast-Betriebssystems ermöglicht ein Backup, das um ein Vielfaches effizienter als herkömmliche Backup-Methoden mit Agenten ist. Lightweight-Avamar-Agenten reduzieren die Backup-Daten auf dem Gast, verringern die Netzwerkanforderungen und die Konflikte bei gemeinsam genutzten CPU-, NIC-, Festplatten- und Speicherressourcen. Weil nur neue oder einzigartige

Subdateidaten gesichert werden, ermöglicht Avamar schnelle tägliche komplette Backups.

- ☒ **Avamar für VCB- oder vStorage-API-Backups.** Ein Avamar-Agent, der auf einem Proxy-Server ausgeführt wird, sichert nur einzigartige Daten und entlastet die Gastmaschinen von der Verarbeitung. Die Deduplizierung erfolgt innerhalb von und über alle VMDK-Dateien hinweg und unterstützt Backups auf Datei- und Image-Ebene. Die effiziente Replikation von Avamar ermöglicht eine schnelle Übertragung von VMDK-Dateien über das WAN, um die Disaster-Recovery-Ziele zu unterstützen.
- ☒ **Avamar-Agent auf ESX-Konsole.** Ein Avamar-Agent auf der ESX-Konsole kann innerhalb von und über alle VMDK-Dateien hinweg deduplizieren. Diese Methode ermöglicht eine Backup- und Wiederherstellungsoption auf Image-Ebene, ohne von VMware VCB oder gemeinsamem Speicher abhängig zu sein. Eine Wiederherstellung auf der Dateiebene ist jedoch nicht möglich.

EMC Data Domain

Data Domain-Deduplizierungsspeichersysteme reduzieren den erforderlichen Festplattenspeicher zur Aufbewahrung und Absicherung von Enterprise-Daten. Durch die Identifizierung redundanter Dateien und Daten beim Speichern ermöglichen Data Domain-Systeme einen Speicher, der durchschnittlich zehn- bis 30-mal kleiner als der des ursprünglichen Datensatzes ist. Backup-Daten können dann effizient über vorhandene Netzwerke für eine optimierte Disaster Recovery und einen konsolidierten Bandbetrieb repliziert und abgerufen werden.

Data Domain-Systeme stehen in verschiedenen Konfigurationen, die sich in Performance und Kapazität unterscheiden, zur Verfügung. Verschiedene Softwareoptionen erweitern die Funktionalität und den Nutzen der Systeme. Zum Beispiel:

- ☒ Data Domain Appliance Series stellt kosteneffiziente und skalierbare Deduplizierungsspeichersysteme mit hohem Durchsatz und integriertem Speicher bereit.
- ☒ Data Domain DDX Array Series bietet ein skalierbares Festplatten-Array-Speichersystem mit hoher Performance, einer logischen Kapazität von bis zu 56,7 Petabyte und einem Durchsatz von bis zu 86,4 Terabyte/Stunde. Es beinhaltet Management und Unterstützung für bis zu 2.880 entfernt gelegene Standorte bei Verwendung von DD880-Controllern.
- ☒ Data Domain Gateway Series stellt Enterprise-Gateways bereit, die Deduplizierungs- und Komprimierungsvorteile für Rechenzentren bieten, die bestimmte externe Speichersysteme von Drittanbietern verwenden möchten.
- ☒ Data Domain Replicator-Software ist eine netzwerkeffiziente, automatische Replikationssoftwarelösung, die für die Disaster Recovery, Absicherung von Daten an entfernt gelegenen Standorten und Bandkonsolidierung mehrerer Standorte zur Verfügung steht.

- ☒ Data Domain Virtual Tape Library-Software emuliert mehrere Bandbibliotheken über eine Fibre-Channel-Schnittstelle und stellt so Deduplizierungsspeicher für SAN-Umgebungen bereit.
- ☒ Data Domain OpenStorage-Software stellt eine nahtlose Integration zwischen Data Domain-Deduplizierungs-Speichersystemen und Symantec Veritas NetBackup bereit.
- ☒ Data Domain Retention Lock-Software ermöglicht dem Anwender die einfache Implementierung der Deduplizierung mit Dateisperren, um internen IT- und Compliance-Policies gerecht zu werden. IT-Administratoren erhalten die erforderliche Flexibilität, um den Unternehmensbetrieb Tag für Tag und kostengünstig aufrecht zu erhalten. Die Software bietet sichere Daten-Shredding-Funktionen.

Data Domain-Deduplizierungslösungen stellen eine konsolidierte Speicherebene für Backup-, Nearline- und Archivdaten bereit. Außerdem können Data Domain-Lösungen in vorhandene Infrastrukturen integriert werden. Sie bieten ein gutes Maß an Investitionsschutz, sodass sie ideal für Firmen mit folgenden Umgebungen geeignet sind:

- ☒ Rechenzentren, die ein explosives Datenwachstum für Backup- und Fixed-Content-Daten aufweisen, und immer schwierigeren operativen und Disaster-Recovery-Anforderungen gegenüber stehen
- ☒ Größere Unternehmen, die eine Konsolidierung von entfernt gelegenen Standorten zu Rechenzentren für die Data Protection und die Disaster Recovery unterstützen müssen
- ☒ Backup- und Archivinfrastrukturen, die ein konsolidiertes Deduplizierungsziel erfordern
- ☒ Infrastrukturen, die auf Bandsysteme und Medien aufbauen und hinsichtlich Backup und Recovery neu entwickelt werden müssen, um belastende Kosten- und Managementprobleme zu eliminieren.

Die Data Domain Data Invulnerability-Architektur stellt die Integrität aller Backup-Daten sicher, indem Data Protection, Datenverifizierung und automatische Fehlerkorrektur auf hohem Niveau bereitgestellt werden. Zu den zentralen Bereichen des Datenintegritätsschutzes gehören folgende:

- ☒ Ununterbrochene Recovery-Verifizierung, damit die Korrektheit von Backup-Daten geprüft wird, und diese von jeder Ebene des Systems im gesamten Lebenszyklus der Daten wiederhergestellt werden können
- ☒ Einzigartigkeitsverifizierung bietet Schutz gegen zufällige und bösartige Hash-Kollisionen, um eine erfolgreiche Daten-Recovery zu ermöglichen
- ☒ Dual Disk Parity RAID (RAID 6) bietet Schutz gegen bis zu zwei gleichzeitige Festplattenausfälle

Die Data Invulnerability-Architektur verfügt über zusätzliche Ebenen des Datenintegritätsschutzes, die integriert wurden, um Fehler zu erkennen und zu beheben und so eine risikoarme Recovery der Backup-Daten zu ermöglichen.

Die Data Domain Replicator-Software überträgt nur die deduplizierten und komprimierten eindeutigen Änderungen über ein IP-Netzwerk, sodass nur ein Bruchteil der Bandbreite, der Zeit und der Kosten von herkömmlichen Replikationsmethoden erforderlich ist. Wenn mehrere Data Domain-Systeme im selben Zielsystem repliziert werden, wird der Deduplizierungseffekt effizienter, da das Zielsystem jedes einzelne eindeutige Segment aus allen eingehenden Replikationsströmen nur einmal speichert, was die erforderliche Bandbreite weiter minimiert.

Die Data Domain-Replikationstechnologie bietet folgende wichtige Vorteile:

- WAN-Vaulting für die Disaster Recovery stellt netzwerkbasierende Data Protection bereit, indem Backup-Daten sicher und automatisch an einen sicheren externen Ort repliziert werden
- Data Protection für entfernt gelegene Standorte ermöglicht das Vaulting von Backup-Daten aus vielen Filialen in einem zentralen Hub oder Rechenzentrum
- Kaskadierte Replikation ermöglicht die Deduplizierung von Daten, die von einem entfernt gelegenen Standort zu einem zentralen Rechenzentrum und dann weiter an zusätzlichen Standorten, wie zum Beispiel einem Disaster-Recovery-Standort, zum Zwecke der verbesserten Recovery und Disaster Protection repliziert werden sollen
- Bandkonsolidierung, wodurch die Duplizierung von Backup-Daten an jedem entfernt gelegenen Standort entfällt, und nur eine deutlich reduzierte Bandinfrastruktur an einem zentralen Hub mit IT-Mitarbeitern erforderlich ist

EMC NetWorker

EMC NetWorker ist eine Enterprise-Backup-Anwendung, die Backup- und Recovery-Vorgänge zentralisiert. NetWorker stellt eine gemeinsame Plattform zur Verfügung, die Unterstützung für eine breite Palette von Data-Protection-Optionen bietet, darunter Backup-to-Disk, Continuous Data Protection und Deduplizierung innerhalb physischer und virtueller Umgebungen. Die Vielseitigkeit von NetWorker macht das Produkt zu einer hervorragenden Backup-Lösung für Kunden, die ihr Management umgebungsübergreifend, von großen Rechenzentren bis zu entfernt gelegenen Standorten, vereinfachen möchten. Die NetWorker-Core-Anwendung ermöglicht durch Integration in die Deduplizierungstechnologie von EMC Avamar die Deduplizierung an der Quelle und kann im Rahmen ihres Betriebs auch Lösungen für die Deduplizierung am Ziel, wie Data Domain-Systeme, nutzen. Unternehmen, die die Deduplizierung mit NetWorker verwenden,

- möchten das Datenwachstum in vorhandenen NetWorker-Umgebungen bremsen.
- stellen eine neue Backup-to-Disk-Strategie für eine verbesserte Recovery bereit, die weiterhin die Verwendung physischer Bänder zur Archivierung oder den langfristigen Bedarf erfordert.
- erfüllen die verschiedensten Anforderungen – einige sind besonders für die Deduplizierung an der Quelle, andere für die Deduplizierung am Ziel geeignet.

- ☒ senken die Kosten und die Komplexität, indem sie mehrere Data-Protection-Strategien mit einer Anwendung konsolidieren.

Der Deduplizierungsansatz der NetWorker-Anwendung hat den Markt im Hinblick auf die Integration der Deduplizierung in eine herkömmliche Backup-Anwendung vorangebracht. Die NetWorker-Client-Software für nicht deduplizierte und deduplizierungsorientierte Backups ist ein einziger Agent. Funktionen zur Deduplizierung an der Quelle sind vollständig integriert; der Aufwand für die Bereitstellung und die Wartung ist daher minimal. Die NetWorker-Konsole kann beide Backup-Arten managen und überwachen – sowohl herkömmliche als auch deduplizierte Backups. Für NetWorker-Kunden, die die Vorteile der Deduplizierung nutzen möchten, entstehen auf Client-Seite keine zusätzlichen Kosten.

Im Gegensatz zu anderen Angeboten gibt es bei NetWorker keine zusätzlichen Softwarelizenzen oder Aufpreise zur Deduplizierungsintegration. NetWorker-Kunden können den entsprechenden Deduplizierungs-Engine zur Backup-Umgebung (Avamar-Software oder die Data Domain-Back-End-Lösung) hinzufügen. Einer der Vorteile bei Verwendung der NetWorker-Deduplizierung ist die Unterstützung für physische Bänder. Wenn Anwender weiterhin Bänder benötigen, können sie diese Anforderung innerhalb ein- und derselben Anwendung erfüllen. Ein anderer Vorteil ist die Möglichkeit, die umfangreiche Anwendungsunterstützung von NetWorker zu nutzen und so die Disaster Recovery, granulare Recovery sowie Snapshot-Management für Host-unabhängige Backups zu ermöglichen. NetWorker stellt Unternehmen die leistungsstarken Funktionen der Deduplizierung zur Verfügung, ohne die aktuelle Backup-Umgebung zu beeinträchtigen.

HERAUSFORDERUNGEN: WELCHER ANSATZ?

Wie in diesem Dokument gezeigt, haben unterschiedliche Deduplizierungstechnologien und -ansätze je nach Einsatzbereich unterschiedliche Vorteile. Es ist daher wichtig, über eine einfache Möglichkeit zu verfügen, jedes EMC Produkt der Umgebung zuzuordnen, in der es eine maximale Effizienz bietet. Tabelle 2 soll Unternehmen die Entscheidung erleichtern, welches deduplizierungsorientierte EMC Backup- und Recovery-Produkt für sie geeignet ist.

EMC bietet viele unterschiedliche Produkte mit Deduplizierungsfunktionen. Da die Deduplizierung kein eigenständiges Produkt ist, besteht der Bedarf, dass EMC Kundens Schulungen entsprechend anpasst, um die Nutzung dieser Funktion auf Basis der spezifischen Herausforderungen in der Kundenumgebung zu vermitteln. Entsprechende Schulungen, zusammen mit dokumentierten Fallstudien sowie Skalierungs- und Performance-Test-Benchmarks, werden das Vertrauen der Kunden in die Anwendung der Technologie im Rahmen der entsprechenden Produkte erhöhen.

TABELLE 2

Auswahl eines EMC Deduplizierungsprodukts

	EMC NetWorker	EMC Data Domain	EMC Avamar
Deduplizierung für Backups	<ul style="list-style-type: none"> • Quelle • Inline-Deduplizierung 	<ul style="list-style-type: none"> • Ziel • Inline-Deduplizierung 	<ul style="list-style-type: none"> • Quelle • Inline-Deduplizierung
Besonders geeignet für Umgebungen mit:	<ul style="list-style-type: none"> • NetWorker-Umgebungen • Unterstützung für physische Bänder erforderlich • Große, heterogene Umgebungen 	<ul style="list-style-type: none"> • Backup/-Recovery mit hoher Geschwindigkeit erforderlich • Replikation für Offsite-Backup • Unterstützung für derzeitige Backup-Umgebung – keine Änderungen im Betriebsablauf • Unterstützung für Rechenzentren und entfernt gelegene Standorte 	<ul style="list-style-type: none"> • Virtuelle Umgebungen • Entfernt gelegene Standorte • LAN-/NAS-Server • Desktop/Laptops
Bereitstellungsoptionen	<ul style="list-style-type: none"> • Einzelner NetWorker-Agent • Siehe Avamar-Bereitstellungsoptionen 	<ul style="list-style-type: none"> • Appliance-Hardware oder Gateway 	<ul style="list-style-type: none"> • Nur Agent – für kleinere, entfernt gelegene Standorte • Avamar Data Store: sofort einsatzbereite All-in-One-Lösung (Hardware und Software) • Server eines Drittanbieters: Erstellen eines eigenen Avamar-Servers • Avamar Virtual Edition: virtuelle Appliance, nutzt vorhandene ESX-Server und Festplatten

Quelle: IDC, 2010

FAZIT

Deduplizierungstechnologie kann die Backup-Effizienz verbessern und die IT-Kosten senken. Unternehmen setzen unterschiedliche Arten von deduplizierungsfähigen Lösungen ein, um eine Vielzahl der durch die stetig wachsende Menge an Backup-Daten entstehenden Kosten zu senken und betrieblichen Herausforderungen zu bewältigen. IDC stellt fest, dass die Deduplizierung für eine Vielzahl von Speicherlösungen eine unverzichtbare Core-Funktion ist, weil sonst diese Herausforderungen nicht bewältigt werden können. EMC als Anbieter ist gut positioniert, um diese schon seit längerem bestehenden Probleme zu adressieren. EMC bietet eine Reihe von Lösungen für eine Vielzahl von Umgebungen und Einsatzbereichen, um die Kundennachfrage nach dieser Technologie in den nächsten fünf Jahren zu erfüllen.

Copyright-Hinweis

Externe Veröffentlichung von IDC-Informationen und -Daten – IDC-Informationen, die in Werbung, Pressemitteilungen oder Promotion-Materialien verwendet werden sollen, bedürfen vorab der schriftlichen Genehmigung durch den zuständigen IDC Vice President oder Country Manager. Ein Entwurf des zur Veröffentlichung bestimmten Dokuments muss jeder Anfrage beiliegen. IDC behält sich das Recht vor, die Genehmigung der externen Nutzung ohne Begründung zu verweigern.

Copyright 2010 IDC. Eine Reproduktion ohne schriftliche Genehmigung ist untersagt.