COMPUTERWOCHE
VOICE OF DIGITAL

Link: https://www.computerwoche.de/a/deduplizierung-wie-geht-das,2358569

EMC sitzt an der Quelle

Deduplizierung - wie geht das?

Datum: 29.11.2010
Autor(en): Thomas Pelkmann

Muss man angesichts der sinkenden Speicherpreise überhaupt noch über Datenmengen sprechen? Oder handelt es sich um ein Randthema, das die ungeteilte Aufmerksamkeit kaum lohnt? Für die Antwort sollte man zumindest wissen, worüber man redet. Eine Übersicht.

Platz satt durch Deduplizierung: Stellen Sie sich vor, Sie könnten 90 Prozent Ihrer Storage-Infrastruktur sparen.

In der IT der Neuzeit steigen nicht nur die **Storage**[1]-Kapazitäten stark an, sondern - und das in noch explosiverem Maße - auch die Mengen der gespeicherten Daten. Laut IDC ergab die Menge der im Jahr 2006 weltweit erzeugten und replizierten digitalen Daten eine Summe von 173 Milliarden GByte. Verglichen mit dem Jahr 2011 ist das eher wenig: Für das kommende Jahr schätzt IDC ein weltweites Speicheraufkommen von stolzen 1.773 Milliarden GByte - das entspricht einer Verzehnfachung der Datenmengen in nur fünf Jahren.

Das Internet mit seinem endlosen Gezwitscher ("**Twitter**[2]") in den sozialen Netzwerken sowie die Unternehmen mit zunehmend digitalisierten Prozessen tragen zu diesem Datenwachstum in nicht unerheblichem Maße bei. Dazu kommt, dass nur wenige Informationen irgendwann dem digitalen Vergessen anheim fallen. Aufbewahrungsfristen und Dokumentationspflichten sorgen dafür, dass kaum etwas von dem gelöscht wird, was einmal seinen Weg auf Festplatten und Bandlaufwerke gefunden hat.

Dass das meiste von diesen exorbitanten Speichermengen dennoch überflüssig ist, zeigt eine einfache Rückwärtsrechnung. Wenn sich, was Anbieter und Analysten übereinstimmend vorrechnen, durch die Reduktion redundanter Daten bis zu 90 Prozent Speicherplatz sparen lässt, heißt das umgekehrt, dass gerade einmal zehn Prozent der Daten wirklich einzigartig sind. Der Rest besteht - zumindest technisch gesehen - aus nichts als überflüssigen Wiederholungen und Doppelungen.

Wer zum Beispiel seine soeben erstellte Firmenpräsentation per E-Mail nicht nur an den Chef, sondern auch ans Marketing und an die drei Kollegen aus der eigenen Abteilung weiterleitet, verursacht damit ein Fünffaches der Menge an Daten, die seine Präsentation eigentlich produziert hat. Und wer ein - zugegebenermaßen hoffnungslos veraltetes - Backup-System einsetzt, dass Tag für Tag den kompletten Firmenspeicherinhalt sichert, erzeugt in kürzester Zeit locker das Vielfache davon.

Alles Doppelte muss raus

Dabei würde es völlig ausreichen, die exemplarisch erwähnte Präsentation nur ein einziges Mal zu sichern, selbst wenn die Kollegen den ein oder anderen Änderungsvorschlag einarbeiten. Denn damit ändert sich nicht die gesamte Vorführung, sondern allenfalls einzelne Teile.

Diesem Grundgedanken ist die so genannte **Deduplizierung**[3] ("Dedup") von Daten verpflichtet: Bei Informationen, die ihren Weg auf dauerhafte Speichermedien finden sollen, filtern Deduplizierungsprogramme alle doppelten Daten heraus. Statt einmal erzeugte Informationsblöcke immer wieder zu speichern, legen Dedup-Anwendungen einen Block einmal ab und verweisen im Wiederholungsfalle einfach auf diese Informationseinheit. Das spart in der Summe enormen Speicherplatz und verringert so die abgelegte Datenmenge um bis zu den eben vorgerechneten 90 Prozent.

Wer gerade wieder eine Rechnung über die Anschaffung von Bandlaufwerken oder Festplatten abgezeichnet hat, wird ermessen können, was eine Einsparung in dieser Größenordnung im Budget ausmacht. Dabei ist der Spareffekt paradoxerweise umso höher, je mehr Daten ein Unternehmen produziert.

Per Definition, etwa von den Marktbeobachtern von IDC, ist Datendeduplizierung "eine Technologie, die doppelt vorhandene Daten in ein einziges gemeinsames Datenobjekt normalisiert, um Speicherkapazitätseffizienz zu erzielen." Im einfachsten Fall vergleichen Dedup-Algorithmen komplette Dateien und sortieren vollständig identische Kopien aus. Sobald ein Kollege in Ihrer Präsentation also auch nur ein Zeichen ändert, weil Sie zum Beispiel ein Komma vergessen haben, würde die Datei erneut gesichert.

Tatsächlich arbeiten moderne Deduplikationstechniken wesentlich flexibler: Sie unterteilen Daten in kleine Blöcke ("Chunks") und vergleichen so schon viel kleinteiliger und damit Speicher schonender. Für jedes erfasste Segment erstellen Dedup-Anwendungen eine Prüfsumme ("Hash"), die dann in einem Index gespeichert wird. Bei späteren Speicherungen derselben Datei mit Abwandlungen werden dann nur die tatsächlich geänderten Blöcke gespeichert. Auf die identischen Teile verweisen dann so genannte Data Pointer.

Dedup an der Quelle oder am Ziel

Sozusagen der letzte Schrei bei den Dedup-Algorithmen ist die Unterteilung in Blöcke nicht starrer, sondern variabler Größe. "Ein Ansatz mit variabler Länge", heißt es bei IDC, "kann die Segmentgröße je nach Content-Typ dynamisch anpassen." Damit sei es möglich, redundante Datensegmente zu berücksichtigen, deren Position sich bei Änderung einer Datei in einem Byte-Stream verschoben habe. "Bei einem Ansatz mit fester Länge werden redundante Daten, deren Position sich geändert hat, nicht erkannt." Das aber sei "ineffizient", weil eigentlich redundante Segmente dann erneut gespeichert werden müssen.

Für die Suche nach überflüssigen Redundanzen haben sich zwei Verfahren etabliert: das Deduplizieren an der Quelle und am Ziel. Beim Dedup an der Quelle, ein Verfahren, das zum Beispiel **EMC**[4] bevorzugt, werden Redundanzen ausgefiltert, bevor Daten für das Backup auf die Speichermedien übertragen werden. Die zielbasierte Deduplizierung erfolgt dagegen nach der Übertragung der Daten direkt am Backup-Speicher-Device.

Beide Verfahren sind Ziel führend, haben aber für sich genommen Vor- und Nachteile. So reduziert die Quellvariante die Menge der über das Netz ans Ziel übertragenen Daten um das Zehn- bis Zwanzigfache. In Unternehmen, deren Netzwerke schon im Normalbetrieb an die Leistungsgrenze kommen oder die mit einer Vielzahl von Außenstellen arbeiten, verhindert dieses Verfahren eventuelle Übertragungsengpässe. Der Gewinn sind eine höhere Verfügbarkeit und eine bessere Performance des Firmennetzes.

Zudem ist Dedup an der Quelle flexibler als am Ziel: Hier können Daten aller Art gespeichert werden, egal, ob sie kompatibel zur Deduplikations-Anwendung sind. Zudem benötigt diese Variante keine zusätzliche Hard- oder Software am Ziel. Schließlich verringert die Quell-Variante die Zeit für Backups, weil redundante Daten schon vor dem Transport durchs Netz und dem eigentlichen Backup-Prozess ausgefiltert und damit stark reduziert werden.

Dieser Gewinn muss vor allem mit Prozessorlast bezahlt werden, weil der Desktop-PC oder der Server für die Deduplizierung der Daten zuständig sind, und nicht das Storage-System. Im Extremfall kann das spürbare Leistungseinbußen auf produktiven Maschinen zur Folge haben.

Deduplizierung kann IT-Kosten senken

Umgekehrt liegt genau hier die Stärke der Deduplizierung am Ziel: Die Quellmaschinen werden von aufwändiger Rechenarbeit entlastet. Die Ziel-Deduplizierung wird häufig dann angewendet, wenn Clients mit der Technik inkompatibel sind oder wenn die Prozessorauslastung beim Client die Dauer der Datensicherung überschreiten würde.

Der Nachteil dieser Methode liegt auf der Hand: Zunächst müssen alle Daten - einschließlich der zum größten Teil redundanten - ans Speichermedium übertragen werden. Das führt auf jeden Fall zu einer eigentlich überflüssigen Be-, wenn nicht zu einer Überlastung der Bandbreiten im Netz.

Deduplizierungstechnologien, da sind sich Anbieter und Analysten schon wieder einig, können die Effizienz von Backups verbessern und die IT-Kosten senken. Die Marktforscher von IDC legen dabei Wert auf die Feststellung, "dass die Deduplizierung für eine Vielzahl von Speicherlösungen eine unverzichtbare Core-Funktion ist", weil sich sonst Kosteneffizienz und andere Herausforderungen der IT nicht bewältigen ließen. Entsprechend gehören Dedup-Lösungen nach übereinstimmender Meinung aller Beteiligten schon bald zu den **Commodities**[5].

Links im Artikel:

[1] **https://www.computerwoche.de/hardware/storage/1912768/**
[2] **http://twitter.com/COMPUTERWOCHE**
[3] **https://www.computerwoche.de/hardware/storage/2351494/**
[4] **http://germany.emc.com/collateral/brochure/datadeduplication.pdf**
[5] **https://www.computerwoche.de/schwerpunkt/i/IT-Commodity.html**
