druva

White Paper

Deduplizierung - Funktion und Umgang mit Unternehmensdaten

Dieses WhitePaper erläutert Deduplizierungstechniken und beschreibt, wie DRUVAs Technologie Speicherbedarf und Leitungsbelastung bis zu 80% reduziert.

Inhaltsverzeichnis

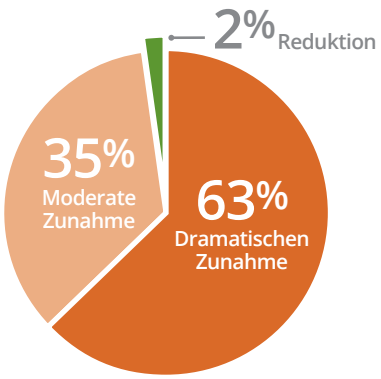
Einführung

Unternehmen suchen heute neue Wege und Methoden, um mit dem immer schnelleren und vor allem unkontrollierten Datenwachstum der verteilten Unternehmensdaten in Bezug auf Management, Schutz und Sicherheit Schritt halten zu können - insbesondere gilt dies für Daten auf Laptops, mobilen Geräten (zusammen als Mobile-Devices definiert) und für Daten in Aussenstellen. Die Datenmenge in Unternehmen verdoppelt sich derzeit alle 14 Monate, die Speicherplätze der Daten sind zunehmend verteilt und die Verknüpfung zwischen Dateien ist komplexer geworden. Diese Faktoren haben erheblichen Auswirkungen auf den Speicherbedarf - und die Bandbreitenkosten. Daten-Deduplizierung bietet Unternehmen die Möglichkeit, die Menge an Speicher und Bandbreite für Backups und andere Daten - Speicherung und Zugangsarbeitsbelastungen drastisch zu reduzieren. Druva verfügt über einen einzigartigen Ansatz in Zusammenhang mit der Daten-Deduplizierungstechnologie, um den Kunden bei Herausforderungen der Datenverwaltung zu helfen.

Zunahme der Unternehmensdaten verursacht einen starken Anstieg der Speicher - und Bandbreitenkosten

In den letzten 5 Jahren haben die Unternehmensdaten enorm zugenommen, deshalb gibt es einen signifikanten Anstieg der Speicher - und Bandbreitenkosten. Eine Umfrage von AFCOM (Rechenzentrum Handelsorganisation) hat festgestellt, dass über 63% der befragten IT - Manager einen dramatischen Anstieg in ihrem Speicherkosten gesehen haben.

Veränderung der Speicheran-forderungen in letzten 5 Jahren,



Zwei Hauptgründe für diesen Anstieg sind die Ausbreitung von mobilen Geräten im Unternehmen und die im letzten Jahrzehnt geografisch immer mehr verteilte Funktionsweise des Unternehmens. Glücklicher- oder unglücklicherweise für IT - Manager, ist ein Großteil des Datenzuwachses das direkte Ergebnis der replizierten Dateien auf mehreren Datenspeichern und Endpunktgeräten.

Deduplizierung von Daten ermöglicht reduzierten Speicherbedarf und geringere Bandbreiten

Daten-Deduplizierung bezieht sich auf die Beseitigung der überflüssigen Daten. Deduplizierungsalgorithmen identifizieren und löschen das Duplikat, so dass nur eine Kopie (oder "einzelne Instanz") der Daten gespeichert werden. Jedoch wird die Indexierung aller Daten noch behalten, sollten diese Daten je erforderlich sein.

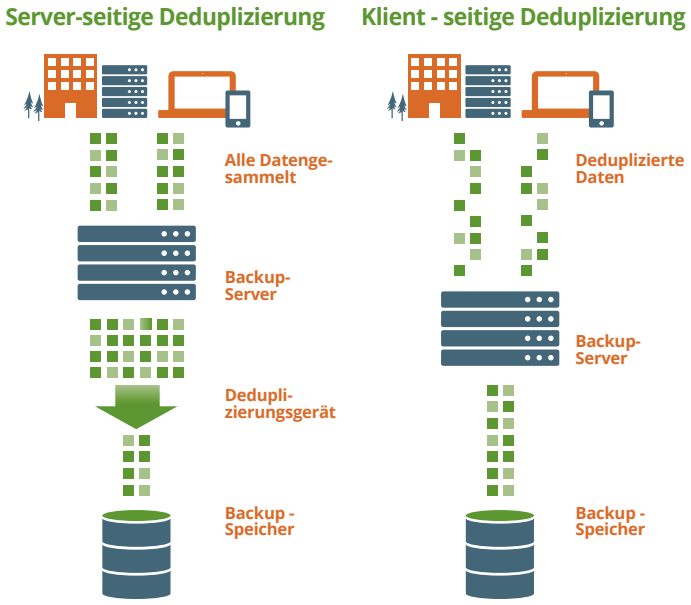
Die Deduplizierung kann die erforderliche Bandbreite und Speicherkapazität verringern, da nur die einzigartigen Daten gespeichert werden. Beispielsweise könnte ein typisches E-Mailsystem 100 Fälle der gleichen 1 MB Dateianlage enthalten. Wenn die E-Mail Plattform gesichert oder archiviert wird, werden alle 100 Fälle gespeichert, so dass ein 100 MB Speicherplatz erforderlich wird. Mit der Daten-Deduplizierung wird nur eine Instanz der Anlage tatsächlich gespeichert; jeder darauffolgende Fall wird sich auf die gespeicherten Kopie beziehen. In diesem Beispiel könnte der erforderliche Speicherplatz und Bandbreite von 100 MB auf nur 1 MB verringert werden.

Die praktischen Vorzüge dieser Technologie hängen von verschiedenen Faktoren ab, wie zB Angriffspunkt, Algorithmus, Datentyp und Daten aufbewahrung / Schutzrichtlinien. Einige der Möglichkeiten, wo die Deduplizierungstechnologien abweichen. Diese Technologien unterscheiden sich in der folgender Weise: Wo die Deduplizierung erfolgt (Server- oder Klientseite), durch Granularität der Deduplizierung (Datei oder Unterdatei basierend) und schließlich durch die Logik der Entdeckung von Duplikaten-Daten (Block basiert oder Anwendung- bewusst).

Verschiedene Methoden für die Daten-Deduplizierung

Es gibt verschiedene Methoden für die Daten-Deduplizierung.

Serverbasiert vs. Clientbasiert



Die Methode der **SERVERSEITIG BASIERTE DEDUPLIZIERUNG** wirkt sich auf die Daten auf dem Server. In diesem Fall ist der Klient nicht betroffen und profitiert nicht von irgendeinem Deduplizierung.

Die Deduplizierungsmaschine kann in der Hardware - Array eingebettet werden, die als NAS / SAN - Gerät mit Deduplizierungsfähigkeiten verwendet werden kann.

Alternativ kann er auch als eigenständige Software oder Hardware Appliance, die als Vermittler zwischen Backup - Server und Speicher - Arrays wirkt angeboten werden.

In beiden Fällen verringert diese Methode die Menge der übermittelten Daten nicht und bietet keine Verbesserungen in Bandbreitennutzung es verbessert nur die Speicherauslastung.

Die **KLIENTSEITIGE DEDUPLIZIERUNGSMETHODE** wirkt auf die Daten auf dem Klient, dh, bevor es auf den Server verschoben wird. Ein Deduplizierung bewusstes backup - Agent wird auf dem Klient installiert, der Agent speichert nur einzigartige Daten. Dieser Ansatz ergibt eine verbesserte Bandbreite und Speichernutzung. Jedoch bedeutet diese Methode eine zusätzliche Computerlast auf dem Backup-Klient.

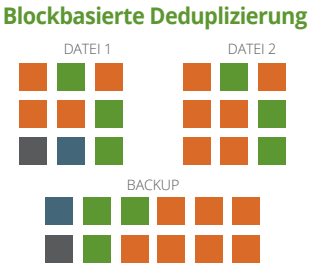
Daten-Deduplizierung auf Datei- vs. Unterdateiebene

Das Algorithmus der Duplikat-Entfernung kann auf vollständige Dateien oder Unterdatei-Ebenen angewendet werden. Dateiebene-Duplikate können leicht beseitigt werden, indem eine einzelne Prüfsumme der kompletten Dateidaten berechnet und diese mit vorhandenen Prüfsummen der gespeicherten Dateien verglichen wird. Dieser Ansatz ist einfach und schnell, aber der Umfang der Deduplizierung ist sehr klein, da er das Problem der duplikaten Inhalte in verschiedenen Dateien oder der Datensätzen nicht löst (z.B., E-Mail). Die Deduplizierungstechnik auf Unterdatei-Ebene bricht die Datei in kleinere festere oder variablere Blöcke und verwendet dann standarde - Hasch - basierte Algorithmen, um ähnliche Blöcke zu finden.

Blockbasierte Deduplizierung

Die Block-basierte Deduplizierungsalgorithmen funktionieren wie folgt: die Deduplizierungsmaschine betrachtet eine Sequenz von Daten, segmentiert sie in Blöcke mit variabler Länge und sucht Blöcke, die wiederholt werden. Die Maschine speichert einen Zeiger zum ursprünglichen Block, anstatt, den duplikaten Block wieder zu speichern. Es gibt zwei Hauptarten vom Block - basierten Ansatz.

Ansatz mit fester Blocklänge, hier werden die Dateien wie der Name schon sagt, in Blöcke mit fester Blocklänge unterteilt und eine einfache Prüfsumme (MD5 / SHA etc.) basierte Ansatz wird verwendet um Duplikate zu finden.



Obwohl es möglich ist, nach wiederholten Blöcken zu suchen, ist der Ansatz sehr begrenzt wirksam, da die Primärgelegenheit für die Datenreduktion das Finden von duplikaten Blöcke in zwei übertragenen Datensätzen ist, welche größtenteils — aber nicht vollständig — von den gleichen Datensegmenten gebildet werden. Zum Beispiel können ähnliche Datenblöcke an verschiedenen Offsets in zwei verschiedenen Dateien vorhanden sein. Das heißt, dass die Blockgrenze von ähnlichen Daten unterschiedlich sein kann.

Dies ist sehr häufig der Fall, wenn einige Bytes in eine Datei eingefügt werden, und die geänderte Datei wieder verarbeitet und in fester Länge - Blöcke unterteilt wird. Alle Blöcke scheinen sich geändert haben. Deshalb mögen zwei Datensätze mit wenig Unterschied sehr wenige identische Blöcke mit fester Länge haben.

Die Datensegmenttechnologie mit variabler Länge unterteilt den Datenstrom in Datensegmente mit variabler Länge mit der Verwendung einer Methodologie, die die gleichen Blockgrenzen in den verschiedenen Orten und in den Zusammenhängen finden können. Dies ermöglicht die Grenzen innerhalb des Datenstreams zu "schweben", so dass Änderungen in einem Teil des Datensatzes wenig oder keine Auswirkung auf die Grenzen in anderen Orten des Datensatzes haben.

Durch diese Methode können duplikate Datensegmente an den verschiedenen Orten innerhalb einer Datei oder zwischen verschiedenen Dateien gefunden werden, welche durch die gleiche / unterschiedliche Anwendung hergestellt werden.

Einschränkungen der blockbasierte Deduplizierung

1 Die Blockgröße (fest oder schwimmend) wird verwendet, um die Datengrenze zu bestimmen, normalerweise ist sie eine "beste" Schätzung und kann daher mit der tatsächlich angewendete Blockgröße nicht vollständig übereinstimmen.

2 Verschiedene Anwendungen haben verschiedene Möglichkeiten, um die Daten auf der Festplatte zu schreiben, der blockbasierte Algorithmus kann oft nicht identische Blöcke in unterschiedlichen Anwendungsdateitypen nicht erkennen (z.B. den gleichen TextBlock gespeichert in MS Word-Datei und eine .PST e-Mail - Datei).

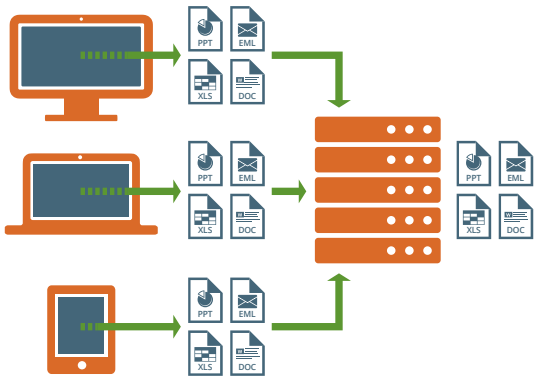
3 Anwendungen wie Microsoft Outlook und Office verwenden eine komplexe Datenbank, basierend auf der Festplattendatenstruktur, die jeden Block mit einem einzigartigen Kopf - und Fußzeile "stempelt", der die Suche nach duplikaten Datenblöcke weiter erschwert.

Anwendungs-„bewusste" Daten-Deduplizierung

Die Anwendung-bewusste (App - aware) Deduplizierung ist ein revolutionäres Konzept, das die Einschränkungen überwindet, die durch die Block-Ebene Deduplication auferlegt werden. Es profitiert von der Kenntnis des Formats der Daten, die gesichert wurden. Anstatt die optimale Blockgröße einzuschätzen, interpretiert es die Datei wie die tatsächliche Anwendung sie tun würde, und identifiziert die logischen Blöcke oder Nachrichten in den Dateien, die sich geändert haben.

Der Deduplizierungsalgorithmus entfernt Duplikate aus der logischen Block- bzw. Nachrichtenebene und ist absolut genau, da es sich auf das Verständnis der Struktur der Daten auf der Festplatte stützt.

Anwendung-bewusste Daten-Deduplizierung



Dieser Ansatz erweist sich als höchst effizient, wenn es um komplexe Anwendungen wie Microsoft Outlook und Office geht, die zu mehr als 95 % der Daten auf den Unternehmens - PCs ausmachen. Die Anwendung-bewusste Deduplizierung identifiziert und entfernt nicht nur alle Duplikate in E-Mails und Anlagen innerhalb einer einzelnen PST - Datei, aber identifiziert und entfernt auch effektiv die Duplikate in verschiedenen Anwendungen. Mit diesem neuen Ansatz kann ein in einem Microsoft Word - Dokument eingebettetes Bild als das Duplikat eines Bildes vorhanden als Anlage in einer PST - Datei identifiziert werden.

Die Anwendung-bewusste Deduplizierung bringt auch Leistungsverbesserungen in der Geschwindigkeit der Deduplizierung, da es wenig oder gar kein Scannen von Daten gibt , um die "schwebende" Blockgrenzen zu finden.

Vorteile der Anwendung-bewussten Block-basierten Deduplizierung

HOCHEFFI-ZIENT
in der anwendungsüber-greifenden Entfernung von Duplikaten.

BIS 300% MEHR EFFIZIENT
als ein einfacher blockbasierter Ansatz bei der Entfernung von Duplikaten in komplexen Anwendungen wie Microsoft Outlook.

BIS 200% SCHNELLERE
Datenverarbeitung im Vergleich zum variablen Block-basierten Ansatz.

BEISPIEL:

Nehmen Sie an, dass eine kurze E-Mail in der folgenden Weise auf der Platte gespeichert wird.

Datum 07.01.2014; Von: Bill Gates; Zu: Warren Buffett, Betreff: Verkaufen; Text: alles verkaufen!

Gehen Sie jetzt davon aus, dass, wenn Bill sein Outlook in einer Woche öffnet und Outlook das Datum auf 07/17 ändert, wird die Nachricht wie folgt gespeichert:

Datum 07.01.2014; Von: Bill Gates; Zu: Warren Buffett, Betreff: Verkaufen; Text: alles verkaufen!

Die Methode der Block-basierten Deduplizierung wird die Nachrichten in den folgenden Blöcken vor und nach der Änderung speichern

VOR:

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5
Datum 07.01.2014;	Von: Bill Gates;	Zu: Warren Buffett ;	Betreff: Verkaufen;	Text: alles verkaufen!

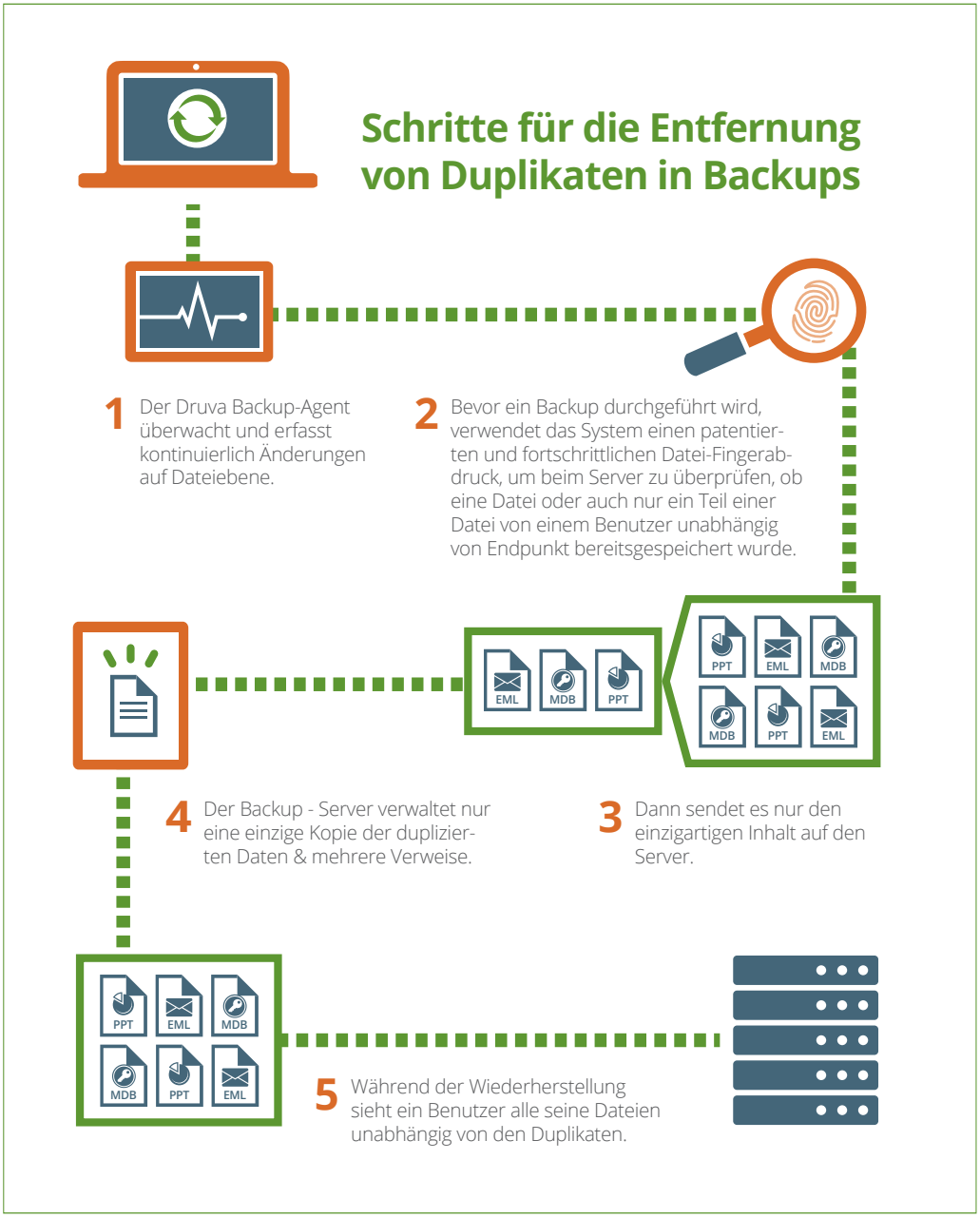
NACH:

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5
Datum 07.01.2014;	Von: Bill Gates;	Zu: Warren Buffett ;	Betreff: Verkaufen;	Text: alles verkaufen!

Beachten Sie, dass die Änderung des Tages auf zwei Zeichen verursacht, dass alle Daten verschieben werden und alle Blöcke unterschiedlich sein werden. Die blockbasierte Deduplizierungsmaschine wird eine ganz neue Kopie der Nachricht erstellen, obwohl nur ein Zeichen geändert wurde. In diesem Fall ist die variable blockbasierte Deduplizierung ein wenig besser, jedoch noch nicht optimal.

Erläuterung der DRUVA Deduplizierungs-Technologie

Die Druva patentierte globale und Anwendung-bewusste Deduplizierungstechnologie bietet unübertroffene Bandbreite und Speichereinsparungen, wenn es auf Backup-und File-Sharing kommt.



Exzellente Eigenschaften der Druva Deduplizierungstechnologie

1 **Allgemein:** Die Datenredundanz wird für alle Benutzer und Endpunkte eliminiert.

2 **Klient-Seite:** Duplikatsüberprüfungen und die Zwischenspeicherung dieser Überprüfungen werden an der Klient-Seite durchgeführt, was die Bandbreite wesentlich verringert und die Backups beschleunigt.

3 **Anwendung-bewusst:** Verständnis von On-Disk-Formate von Anwendungen, beispielsweise führt dies in Outlook zur 100% korrekten, schnelleren Deduplizierung und reduziert den Speicherbedarf.

4 **Hohe Leistung:** Die globale und Anwend-ung-bewusste Deduplizierung stellt sicher, dass nur einzigartige Daten auf den Server übertragen werden, was zu bis zu 6 X schnelleren ersten Sicherungen und einer optimalen WAN - Bandbreitennutzung führt.

5 **Einheitliches anwendungsüber-greifendes Backup & Datei - Sharing:** Die Deduplizierung funktioniert über Backup - und File - Sharing - Funktionen, damit massive Speicher - und Bandbreitene-insparungen erzielt werden können.

Eine ausführliche Erläuterung der Deduplizierungstechniken und des Innovationsansatzes des Druva zur Deduplizierung finden Sie in einem White Paper über Datendeduplizierung für Endpunkte hier: ***www.druva.com/resources****.*

Globale Deduplizierung

Der Klient - Agent führt zweifache Überprüfungen auf dem Klient - Gerät durch einen Vergleich der Daten von allen Benutzern im Unternehmen und von allen ihren Geräten aus. Nur eine einzelne Instanz eines Blocks aus über alle Geräte/ Benutzer wird auf dem Server gespeichert. Diese Kombination des globalen und Klient-seitigen Deduplizierung bietet 80 % Einsparungen bei Bandbreite und Speicher.

Deduplizierung auf Client-Seite

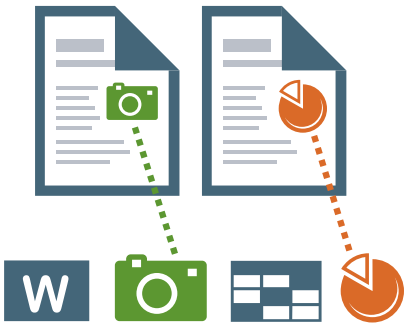
Druva verwendet eine Klient-ausgelöste Architektur mit Deduplizierung auf dem Klient, die hohe Ebenen von Skalierbarkeit und von Sicherheit ermöglicht. Durch die Ausführung von Deduplizierungstests auf dem Klient, ist Druva in der Lage erhebliche Bandbreite zu speichern. Zusätzlich die Klient-seitige Zwischenspeicherung dieser Datei und die Prüfungen der Unterdatei-Deduplizierung machen die Backups deutlich schneller.

Der Klient hat auch eine leistungsfähige WAN Optimierungsmaschine , welche die Netzwerkverfügbarkeit automatisch priorisiert und die Backup - Bandbreite als Prozentsatz der insgesamt verfügbaren Bandbreite festlegt. Dadurch stellt Druva sicher, dass ein Backup weder einen großen Prozentsatz der Bandbreite verbraucht, noch die Endbenutzereraktivität stört.

Objektbasierte, Anwendungs-spezifische Deduplizierung

Druva kennt das On-Dist-Format der Anwendungen und nutzt diese Kenntnisse, um den Deduplizierungsprozess deutlich zu verbessern und eine 100% Genauigkeit zu garantieren. Die Anwendungbewusset Deduplizierungstechnologie erkennt gängige Anwendungen wie Outlook (PST). Die Anwendung-bewusste Deduplizierung beseitigt die Abhängigkeit von mehreren Prüfsummen, um schnellere Deduplizierungsresultate zu erzielen. Für andere Anwendungen wird eine Methode der Deduplizierung mit variabler Blocklängen verwendet.

DEDUPLIZIERUNG INNERHALB VON ANWENDUNGEN



Viele Anwendungen neigen dazu, die Datenstruktur einer Datei zu ändern, auch wenn ein kleines Element von der Datenstruktur geändert wird. Als Ergebnis sieht die gesamte Datei unterschiedlich aus, wenn es dauerhaft auf der Festplatte gespeichert wird. Betrachten Sie z. B. eine Outlook PST - Datei. Die PST - Datei ändert bis zu 5% der Blöcke auch wenn der Benutzer Outlook ohne Aktualisierung schließt. Die Druva Anwendung-bewusste Technologie erkennt bis zu 87 verschiedene Nachrichtentypen in PST - Dateien, um die tatsächlichen Änderungen abzufangen und nur die einzigartige Inhalte (z. B. e-Mails, Anhänge, Kalenderaktualisierungen, etc..) zu speichern

Dieser Ansatz garantiert 100% Deduplizierungsgenauigkeit in unterstützten Anwendungen und eine optimale Nutzung des Speicher und Bandbreite.

ANWENDUNGSÜBERGREIFENDE DEDUPLIZIERUNG

Jede Anwendung speichert unterschiedlich Daten auf der Festplatte und oft ändert sich die Darstellung der Daten vollständig wenn sie auf der Festplatte indexiert und gespeichert werden. Ein gutes Beispiel ist eine Bilddatei, die in einem Word - dokument sowie in einer PST - Datei als Anlage anwesend ist. Die blockbasierte Deduplizierung ist häufig nicht in der Lage, solche Duplikate anwendungsübergreifend zu identifizieren (wie die Daten selbst geändert hat).

Da Druva die logische Ansicht der Daten versteht, ist es viel effizienter Duplikate anwendungsübergreifend zu entdecken und diese als andere Deduplizierung-basierte Backuplösungen zu beseitigen.

BEISPIEL:

Unter Verwendung der Anwendung- bewussten Deduplizierung kann ein Word - dokument auf dem Desktop eines Benutzers als Duplikat einer E-Mail - Anlage leicht identifiziert werden, die bereits gespeichert wurde, und kann vom Speicher entfernt werden.

Vereinheitlichte Deduplizierung für Backup und File-Sharing

Druva inSync bietet sowohl File Sharing und Zusammenarbeit und Backup - Funktionen und der Deduplizierungsalgorithmus funktioniert in beiden Funktionalitäten und verringert weiter die Bandbreiten - und Speicherkosten.

Organisationen verwenden normalerweise zwei separate Lösungen für Filesharing und Backup. Die gleiche Datei mag in mehrfachen Kopien auf dem Server vorhanden sein —, zuerst von einem Backup-Programm und dann wieder durch ein Filesharing - Tool gespeichert. Während das verwendete Filesharing-Tool möglicherweise zur Deduplizierung fähig wäre, kann es keine anwendungsübergreifende Deduplizierung von Dateien, die durch ein anderes Tool wie ein Backup Utility gespeichert wurden, durchführen. Druva InSync ist die einzige Lösung, die einer Organisation ermöglicht, ein einziges Tool für Filesharing und Backup zu verwenden. InSync weiß, welche Dateien für das Backup und welche für Filesharing gespeichert werden um Überschneidungen zwischen den beiden Funktionen zu beseitigen, was zu erheblichen Bandbreite und Speicherkostenvorteile führt.

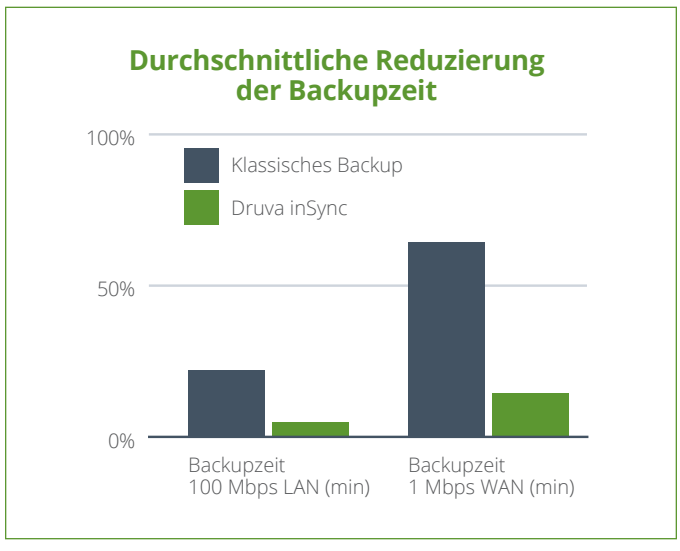
Einsparpotential durch DRUVA Deduplizierung - Bandbreite, Speicherplatz

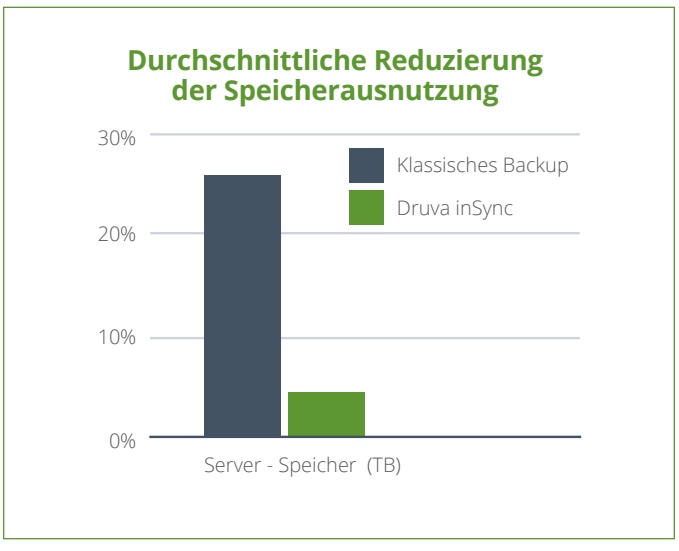
Die folgende Tabelle evaluiert die Leistung der Deduplizierungstechnologie des Druva gegen obliegende Installationen bei vier verschiedenen Kunden in unterschiedlichen Industriebranchen. Diese Benchmarks zeigen deutlich die Vorteile, die die Druva's globale und Anwendungs-bewusste Deduplizierungstechnologie in Bezug auf Backupzeit und Speicherauslastung bereitgestellt werden.

BEISPIEL

SWEISE wenn eine E-Mail mit einer 1 MB Anlage an 1000 - Benutzer gesendet wird, würde die traditionelle inkrementelle Backup-Software diese 1 MB - Anlage aus jeder der 1000 verschiedenen Mail-Boxen speichern. Druva InSync würde dagegen die 1 MB Anlage vom ersten Benutzer speichern, und alle anderen 999 Kopien als Duplikate überspringen, was mehr als 99,9 % Backup - Zeit, Bandbreite und Speicherplatz spart.

KUNDEN	Anzahl der PCs	Durchschnitt LAN BACKUPZEIT (MIN)		Durchschnitt VPN/WAN BACKUPZEIT (MIN)		TOTAL SPEICHER VERWENDET (TB)	
		Alte Anwendung	inSync	Alte Anwendung	inSync	Alte Anwendung	inSync
GROßE FINANZGESELLSCHAFT	2000	24	8	90	20	60	12
ÖL- und GASUNTERNEHMEN	500	10	4	N/A	9	10	1.2
BERATUNGSGRUPPE	300	15	6	40	8	27	2
GRAFIKUNTERNEHMEN	100	45	14	N/A	6	6.8	1.6





Über Druva

Druva ist der Führer im Datenschutz und -verwaltung am Rand und bringt Übersicht und Kontrolle in die Geschäftsinformationen bei den zunehmend mobilen und verteilten Unternehmen. Errichtet für die allgemeinen und privaten Wolken, verhindern die ausgezeichneten inSync und Phoenix - Lösungen des Druva den Datenverlust und befriedigen die Anforderungen der Datenverwaltung, Compliance und eDiscovery auf Laptops, Smartgeräten und Fernservern. Als am schnellsten wachsender Datenschutzanbieter der Industrie, werden Druva von mehr als 3.000 globalen Unternehmen über 3 Millionen Geräte anvertraut. Erfahren Sie mehr unter **www.druva.com** und beteiligen Sie sich im Gespräch auf **twitter.com/druvainc**.

Q415-CON-10473

Druva, Inc.
Americas: +1 888-248-4976
Europe: +44(0)20.3750.9440
APJ: +919886120215
sales@druva.com
www.druva.com