Taugt Backup als Archiv?

Die Trennung von Backup und Archiv gehört zu den bekannten Grundprinzipien in der IT – in der Praxis wird sie häufig nicht konsequent umgesetzt. Warum es sich lohnt, diesen Ansatz neu zu betrachten, zeigt die RWTH Aachen University. Sie macht deutlich, wie sich Datenbestände gezielter steuern, Wiederherstellungen beschleunigen und Anforderungen aus Verwaltung, Forschung und IT sauber voneinander trennen lassen.

Text: **Thomas Eifert** (RWTH Aachen University)

Illustration: axllll/iStock

Die Frage, ob ein Backup auch als langfristige Datenaufbewahrung dienen kann, taucht immer wieder auf. Eigentlich – zumindest theoretisch – ist klar, dass ein „Langzeit-Backup" ein überholtes Konzept ist. Und dann holt uns die Realität ein …

Häufig wird argumentiert, dass bestimmte Daten wie etwa Geschäftsdaten aus rechtlichen oder organisatorischen Gründen über lange Zeiträume vorgehalten werden müssen. Oder jemand möchte nach langer Zeit eine vermisste Datei wiederfinden: „Die muss doch noch im Backup sein!" Die daraus resultierenden hohen Anforderungen an die Kapazität des Backup-Systems führen dann häufig zu

der Unterscheidung in „heiße“ und „kalte“ Daten und zu der Lösung, Letztere auszulagern, um unterschiedliche Zugriffshäufigkeiten und Aufbewahrungszeiten abzubilden. Auf den ersten Blick wirkt dieser Ansatz plausibel. Bei genauerer Betrachtung zeigt sich jedoch, dass dadurch eine Reihe von Problemen entstehen. Aus diesem Grund hat sich das IT-Center der RWTH Aachen University dafür entschieden, unterschiedliche Lösungen für die Aufbewahrung von Daten (möglichst) klar zu trennen.

Der Zweck eines Backups ist eigentlich klar definiert: Es dient dazu, den letzten bekannten guten Zustand eines Systems zu sichern. Ziel ist es, im Falle eines Fehlers, eines Unfalls oder eines Angriffs – beispielsweise durch Schadsoftware – einen Datenverlust zu vermeiden und das System

> In jedem Fall enthält das Backup-System den oder die (wenigen) Zustände, aus denen im Ernstfall die IT wieder hergestellt werden kann.

möglichst schnell wieder in einen funktionsfähigen Zustand zu versetzen. Es handelt sich also um einen systemgetriebenen Ansatz. Im Mittelpunkt steht das IT-System und dessen Wiederherstellbarkeit.

Was macht Backup?

Backup-Systeme arbeiten in der Regel mit periodischen Sicherungen. Wenn ein Langzeit-Backup erstellt wird, werden zusätzliche, termingebundene Sicherungen zu bestimmten Anlässen durchgeführt. Zwischen diesen Sicherungsarten entstehen komplexe Abhängigkeiten, die für Anwender meist unsichtbar bleiben, weil sie tief im Innern der Backup-Software verborgen sind. Dennoch bestimmen sie maßgeblich, wie Daten gesichert werden und wie eine Wiederherstellung funktioniert.

Ein weiterer wichtiger Punkt ist, dass eine Sicherung immer zum Zustand des Systems im Moment der Sicherung passt. Das bedeutet: Das Backup bildet ein konkretes System inklusive seiner Struktur, seiner Konfiguration und seiner Daten zu einem bestimmten Zeitpunkt ab. Genau darin liegt auch ein Zielkonflikt. Einerseits möchte man für die langfristige Speicherung eine möglichst generische Sicherung, die viele verschiedene Anforderungen erfüllt. Andererseits soll das Backup effizient sein – sowohl bei der Sicherung als auch bei der Wiederherstellung. In manchen Fällen ist es sogar notwendig, nicht nur die Daten, sondern auch die Anwendung selbst mitzusichern, um später wieder auf einen konsistenten Zustand zurückkehren zu können.

„Wie lange habe ich denn nun Backup?“

Die Aufbewahrungszeit von Backups richtet sich nicht primär nach organisatorischen oder rechtlichen Anforderungen, sondern eher nach technischen Überlegungen. Ein Beispiel dafür ist die Frage, wie lange eine mögliche Malware unentdeckt in einem System verbleiben kann. Daraus ergibt sich eine heuristische Entscheidung darüber, wie lange Sicherungen verfügbar bleiben sollten, um im Ernstfall auf einen Zustand zurückgreifen zu können, der noch nicht kompromittiert war. Gleichzeitig ist die Maximalzeit dadurch beschränkt, wie lange sich mit einer solchen systemseitigen Sicherung sinnvoll noch etwas anfangen lässt. In jedem Fall enthält das Backup-System den oder die (wenigen) Zustände, aus denen im Ernstfall die IT wieder hergestellt werden kann.

Ganz anders sieht es bei einer Wiederherstellung sehr alter Daten aus, etwa alter Buchungsdaten. Aus betrieblicher Sicht ist es meist nicht sinnvoll, solche Daten aus einem Backup in das Primärsystem wiederherzustellen, sondern beispielsweise für eine Revision bereitzustellen. Damit gehört diese Aufgabe nicht zum eigentlichen Zweck eines Backups, sondern in den Bereich einer geordneten Datenaufbewahrung im Rahmen von Geschäftsprozessen. Wenn Daten über lange Zeiträume vorgehalten werden müssen, geschieht dies aus Gründen der Dokumentation, der Nachvollziehbarkeit oder aufgrund gesetzlicher Anforderungen – nicht weil das unterliegende IT-System sie in einer Sicherung gespeichert hat. Überdies kann – je nach Aufbewahrungsfrist – das IT-System in der Zwischenzeit erneuert oder ersetzt worden sein, was die Verwendbarkeit einer solchen systembezogenen Sicherung infrage stellt.

Damit wird deutlich: Diese Form der Aufbewahrung von Daten passt nicht zu einem Backup. Die Notwendigkeit ergibt sich aus dem jeweiligen Geschäftsprozess und nicht aus dem IT-System selbst. Daraus folgt die logische Konsequenz, systemgetriebene und prozessgetriebene Aufbewahrung voneinander zu trennen. Diese Trennung führt zu den komplementären Ansätzen Backup und Datenarchiv.

Für eine Einrichtung wie die RWTH Aachen ergibt sich daraus ein klarer Lösungsansatz. Statt zu versuchen, ein einziges System für alle Zwecke zu nutzen, werden separate Lösungen für Backup und Archivierung eingesetzt. Das Backup übernimmt die Sicherung von Systemzuständen und dient der schnellen Wiederherstellung im Störfall. Das Archiv hingegen kümmert sich um die langfristige und strukturierte Aufbewahrung von Daten, deren Haltezeiten durch organisatorische, wissenschaftliche oder gesetzliche Anforderungen bestimmt werden.

Innerhalb des Archivs selbst ist eine weitere Differenzierung sinnvoll. Ein Bereich betrifft Dokumentenmanagementsysteme und Verwaltungsarchive. Hier werden Haltezeiten in erster Linie durch Geschäftsprozesse festgelegt sowie durch Compliance-Anforderungen oder gesetzliche Regelungen. Ein anderer Bereich betrifft Forschungsdaten. Für diese gelten in der Regel andere Kriterien. Oft orientiert sich die Aufbewahrungsdauer an den Regeln guter wissenschaftlicher Praxis, die beispielsweise eine Mindestaufbewahrung von zehn Jahren vorsehen können. Darüber hinaus kann es Forschungsdaten geben, die einen dauerhaften wissenschaftlichen Wert besitzen und deshalb in eine langfristige Datenhaltung oder Langzeitverfügbarkeit überführt werden sollten. Eine entsprechende Infrastruktur haben wir in die Landesinitiative für Forschungsdatenmanagement – fdm.nrw integriert. Hier existiert ein von mehreren NRW-Hochschulen kollaborativ erbrachter Service für alle Forscherinnen und Forscher an NRW-Hochschulen.

Schließlich gibt es noch einen dritten Bereich: Die „digitalen Reste“ sind Daten, die keinem klar definierten Prozess zugeordnet sind oder deren zukünftige Bedeutung unklar ist – die dennoch zumindest vorläufig

> Ein wichtiger Punkt ist die Verlagerung von Single-Item-Recovery-Funktionen auf die Ebene des primären datenhaltenden Systems.

aufgehoben werden sollen. Auch für diese Daten muss eine Lösung gefunden werden, allerdings nicht unbedingt innerhalb eines klassischen Backup-Systems.

„Single-Purpose“-Systeme mit klar definierten Aufgaben

Ein weiterer wichtiger Punkt in diesem Konzept ist die Verlagerung von Single-Item-Recovery-Funktionen auf die Ebene des primären datenhaltenden Systems. Ein typisches Beispiel hierfür ist ein E-Mail-System wie Exchange. Wenn einzelne Objekte – etwa eine E-Mail, ein Kalendereintrag oder ein Dokument – wiederhergestellt werden sollen, ist es wesentlich effizienter, dies direkt innerhalb des Systems zu tun, das diese Daten verwaltet. Das Backup dient hier „nur“ der Wiederherstellung nach einem großen Datenverlust.

Der Vorteil dieses Ansatzes liegt darin, dass wir zwar mehrere unterschiedliche Systeme betreiben müssen, jedes dieser Systeme jedoch klar auf einen bestimmten Zweck ausgerichtet ist, sodass nur „Single-Purpose“-Systeme mit klar definierten Aufgaben vorhanden sind. Diese Spezialisierung führt zu einer deutlichen Reduktion von Abhängigkeiten und Komplexität. Backup-Systeme müssen sich nicht mehr um langfristige Archivierungsanforderungen kümmern, und Archivsysteme müssen nicht die Anforderungen an schnelle Systemwiederherstellung erfüllen. Beide Bereiche können unabhängig voneinander optimiert werden.

Darüber hinaus ermöglicht die Trennung Optimierungen auf der Prozess- und der Kostenseite: Durch den geschäftsprozessgetriebenen Ansatz ist klar, welche Daten für welchen Zweck aufbewahrt werden. Darauf folgt unmittelbar, für welche Daten dies nicht der Fall ist, sodass diese aus der langfristigen Speicherung herausgehalten werden können – eine Unterscheidung, die auf Systemebene nicht möglich ist. Überdies ist damit auch ein wichtiger Aspekt des Business Continuity Managements (BCM) adressiert, dass nämlich die geschäftskritischen Datenbestände hier bereits klar identifiziert sind.

Auf der technischen Ebene sorgt diese Trennung für den großen Gewinn: Zum Beispiel ist im Fall einer Erneuerung klar definiert, welche Datenbestände nach einer begrenzten Haltezeit obsolet sind, und für welche Teile eine oft aufwendige Migration gestellt werden muss.

Fazit

Selbstverständlich war die Umsetzung dieses Trennungsansatzes auch mit einem gewissen Aufwand verbunden, insbesondere bei der Einführung auf Prozessebene. In vielen Einrichtungen existieren bereits große Mengen an Daten, die über Jahre hinweg in unterschiedlichsten Strukturen abgelegt wurden. Um die Trennung zwischen Backup und Archiv sauber umzusetzen, ist es unumgänglich, grundlegend aufzuräumen. Alternativ kann ein Stichtag definiert werden, ab dem neue Regeln gelten.

Eine pragmatische Lösung besteht darin, einen „Daten-Dachboden“ einzurichten. Dabei handelt es sich um einen Bereich, in dem unsortierte oder schwer einzuordnende Daten abgelegt werden können. Diese Daten sind zwar weiterhin verfügbar, werden jedoch nicht mehr aktiv in operative Prozesse integriert. Auf diese Weise lässt sich vermeiden, dass wertvolle Informationen verloren gehen, während gleichzeitig eine klare Struktur für neue Daten geschaffen wird.

Langfristig ermöglicht die Trennung von Backup und Archiv eine deutlich verbesserte Steuerung der Datenhaltung. Systeme bleiben übersichtlich, Wiederherstellungen werden effizienter, und die Anforderungen aus Verwaltung, Forschung und IT lassen sich klar voneinander abgrenzen. Genau darin liegt der eigentliche Vorteil eines solchen Ansatzes: Er schafft Klarheit darüber, wofür welche Systeme zuständig sind – und verhindert, dass ein Backup-System zu einer universellen, aber letztlich ungeeigneten Lösung für alle Formen der Datenaufbewahrung wird. ◆