Risiken des Datenverlustes bei der Langzeitarchivierung

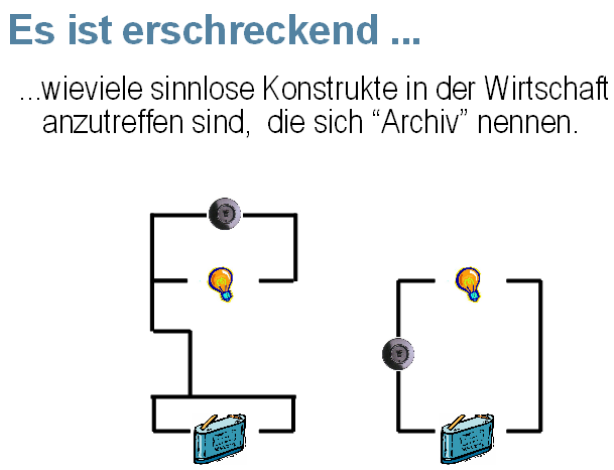
Thorsten Lange
Oracle Deutschland B.V. & Co. KG
Nagelsweg 55, 20097 Hamburg

Schlüsselworte:

Archiv, Archivierung, Langzeitarchivierung, Governance, Risk Management, Compliance, GRC, Aufbewahrung, Datenverlust, Risiko, Daten, Funktionen, Informationen, Inhalte, Migration, Wiederverwendung, Technologien, Formate, Aufbewahrungsdauer, Datenträger, Speicher, Speichermedien, Ordnung, Algorithmus, Linear, Logarithmisch, Exponentiell, SNIA, 100YATF, 100 Year Archive Task Force, Archive Requirements Survey

Einleitung

Dieser Vortrag befaßt sich mit den Problemstellungen, wie sie bei der Langzeitarchivierung von Daten auftreten, sowie den Risiken, die Daten ungewollt auf der langen Reise durch die Zeit zu verlieren. Wir stellen zunächst fest, was unter Langzeitarchivierung zu verstehen ist, behandeln die Arten eines Datenverlustes sowie die Gründe für eine solche Entwicklung. Eine mathematische (keine Angst, leicht verständliche) Betrachtung eines sich auftürmenden Problems veranschaulicht die kritische Entwicklung bei der Aufbewahrung unserer Datenbestände.



Vortragstext

Begriffsbestimmung

Wir archivieren Informationen, um sie zu einem späteren Zeitpunkt wiederverwenden zu können. Gelingt das nicht, habe ich meine Information verloren. Das beinhaltet deutlich mehr als nur den Defekt oder Verlust eines Datenträgers. Man kennt das geflügelte Wort "Niemand braucht ein Backup – frag nach dem Restore!". Die Archivierung kennt das gleiche Problem. Was nützt es, wenn ich Informationen heute in Fülle archiviere, aber für eine durchführbare Wiederverwendung nicht hinreichend Sorge getragen habe?

Ich teile ein Archivierungsvorhaben zunächst ein anhand des wichtigsten Kriteriums: Der vorgesehenen Aufbewahrungsdauer:

- Kurz (bis: 3-4 Jahre)
  Zur Wiederverwendung in der gleichen technischen Umgebung
  Ohne Migration
- Mittel (bis: 10-15 Jahre)
  Zur Wiederverwendung in einer technisch vergleichbaren Umgebung
  Mit "einfacher" Migration
- Lang (bis: 100- "einige hundert" Jahre)
  Zur Wiederverwendung in technologisch veränderter Umgebung
  Komplexe Migration von Technologien und Formaten
- Endlos (1.000 Jahre und darüberhinaus)
  Zur Verwahrung von Informationen für eine unbekannte Nachwelt
  Hinterlassen von Informationen ohne technische Voraussetzungen

In der Langzeitarchivierung befassen wir uns also mit einem Zeithorizont, der über die möglichen Perspektiven jedweder beteiligter Komponenten meines Archivsystems hinausgeht.

Arten des Datenverlustes

Wir sind übereingekommen, daß unter Datenverlust nicht nur zu verstehen ist der Defekt oder Verlust von Datenträgern, sondern generell eine nicht mehr mögliche Wiederverwendung. Die SNIA hat dazu in 2007 eine Erhebung veröffentlicht, (100 Year Archive Task Force, „Archive Requirements Survey") mit einem erschreckenden Ergebnis: So, wie bislang Daten aufbewahrt werden, wird es in den kommenden Dekaden zu flächendeckenden Datenverlusten kommen:

Zitat:

> "It is the contention of the 100 Year Archive Task Force that migration as a discrete long-term preservation methodology is broken in the data center. Today's migration practices do not scale cost-effectively and won't be done until a crisis erupts. This means that today's reliance on migration is taking us down a 'dead-end path'. Hear this clearly. Under these practice guidelines, the world's digital information is at great risk!"

100 Year Archive Requirements Survey
http://www.snia.org/forums/dmf/programs/ltacsi/100_year/

Das Papier nennt auch die Top 4 Arten, seine Daten zu verlieren:

- Can not read it

Dies ist der offensichtlichste Teil. Überlagerte Datenträger, Bit-rotting, aber auch nicht mehr verfügbare Laufwerke für alte Datenträger kommen in Frage.
Abhilfe: Zeitgerechtes Erneuern und Umkopieren.

- Can not interpret it correctly

Ich habe die Daten noch, aber ich kann sie nicht nutzen, weil mir eine Funktionalität fehlt. I.d.R. die ursprüngliche Software. Das kann eine einfache Formatfrage sein (Wie öffne ich eine 20 Jahre alte Harvard's Graphics Datei?), aber auch komplexe Datenstrukturen, mit verteilten Inhalten, mittels spezieller Software verknüpfter Metadaten – wie soll ich die nutzen, wenn es diese Software nicht mehr gibt?
Abhilfe: Schon bei der Archivierung die Wiederverwendung ohne die aktuell beteiligten Komponenten planen. Das bedarf individueller Betrachtungen. Trennen von Daten und Funktionen. Reduktion von Daten auf aufbewahrungsfähige Inhalte.

- Can not validate its authenticity

Ein interessanter Punkt. Ich kann die Daten noch nutzen, aber ich kann die Authentizität nicht prüfen. Damit habe ich die angedachte Aussagekraft verloren.
Abhilfe: Die richtige Strategie. WORM kann als Technik nur bedingt angewendet werden in der Langzeitarchivierung. Qualifizierte Signaturen sind möglicherweise eine Lösung.

- Can not find it

Neben Punkt 2 der am häufigsten unterschätzte. Insbesondere bei unstrukturierten oder schwach strukturierten Daten erfolgt oftmals eine unkontrollierte schlichte „Auslagerung“ statt organisierter Verwahrung. Eine wachsende Menge an Daten ist nur noch über „weiche“ Kriterien auffindbar, z.B. eine Volltextsuche. Zum einen verliert man dadurch die Eindeutigkeit („Welche Dokumente genau gehören zu Projekt xyz?“), zum anderen bewirkt die schlichte Menge eine sich stetig verschlechternde Auffindbarkeit.
Abhilfe: Datenklassifizierung, Definieren des archivierungswürdigen Materials, durchgängige Verwendung von Metadaten und Ablagestrukturen.

Gründe des Datenverlustes

Nun haben wir gesehen, was den Daten widerfahren kann, und auch was man dagegen tun kann. Warum machen wir es dann nicht einfach richtig? Wo ist das Problem?

- Falsches Gewerk

Dokumentenmanagementsysteme zeichnen sich dadurch aus, daß sie Daten und Funktionen miteinander verbinden, und das recht komplex. Das entgegengesetzte Ziel eines Archivs, welches möglichst geringe Voraussetzungen für die Wiederverwendung in ferner Zukunft erreichen möchte. Bei vielen (nicht allen!) als Archiv eingesetzten DMS'en ist das Weiterleben der Daten ohne das DMS nicht gewährleistet, oder bedarf eines „Export After Lifespan“, das weiter unten behandelt wird.

- Falsche Governance

In wessen Verantwortung und Steuerung liegt das Archiv? Oftmals bei der IT, und das ist der Fehler. Archivierung ist eine organisatorische Aufgabe, keine IT Task. Eine aus der Geschäftsleitung delegierte Verantwortung ist erforderlich in die Fachbereiche, die die Daten erzeugen. Nur dort

können die Informationen fachlich durchdrungen, Entscheidungen getroffen werden. Ein durch die IT betriebenes Archiv zeichnet sich typischerweise aus durch schlecht geordnete Haufenbildung.

Quiz: Welches sind die Top 2 Lösungsansätze für nebenstehendes Problem?

1. Putzkolonne draufschicken
2. Nachbarzimmer zumieten

- Falsche Motivation

Mängel in der Aufbewahrung archivierungspflichtiger und sonstiger unternehmenswichtiger Daten sollten im Prüfungsbericht ihre Wirkung zeigen, und z.B. über Basel II die Finanzmittel verteuern. Nach meiner Beobachtung ist der Druck durch die Wirtschaftsprüfung in diesem Thema jedoch nicht hinreichend spürbar. Ich zweifle, ob die Mittel und Methoden dafür vorhanden sind. Damit handelt die Geschäftsleitung tatsächlich im besten Sinne des Unternehmens, wenn sie sich für die kostengünstigste Lösung im Hier und Jetzt entscheidet und das Problem der Wiederverwendung dem Nachfolger in ferner Zukunft überläßt.

Prognosen

Was wird also nun voraussichtlich geschehen? Was meint die SNIA mit „Dead End Path“ bezüglich der Migrationsstrategie?

- Wachsende Berge

Wir werden weiterhin Daten ansammeln und die oben beschriebenen Schwächen beibehalten. Zunächst wird dies durch stetigen neuen Zukauf von Speicher sowie Weiterentwicklung in der Kapazität desselben aufgefangen.

- Migrationsproblem

Datenträger müssen regelmäßig erneuert werden, um mit der technischen Entwicklung Schritt zu halten und um Überlagerung zu vermeiden. Nun wächst mit jedem Zyklus auch die Menge der zu migrierenden Daten. Das betrifft im einfachsten Fall den physikalischen Datenträger, dessen Inhalt 1:1 zu kopieren ist. Es betrifft aber auch Daten, die nun endgültig von mitgeführten Funktionalitäten (Software) getrennt werden müssen.

- Daten in proprietären Appliances
- Daten in proprietären Storage Management Systemen
- Daten in einem DMS (die selbiges für die Wiederverwendung benötigen)

Wo immer es notwendig wird, Daten in einer Großen Iteration aus einem System zu exportieren, laufen wir in ein exponentielles Problem. Fragen Sie sich selbst: Was auch immer 100 Jahre benötigte, um einverleibt zu werden, wie lange wird es brauchen für den Export? („Export-to-Ingest Ratio“).

Eines der wesentlichen Folgeprobleme der o.g. Fehler ist, daß die Migration nicht mehr linear skalieren kann. Wer einmal Programmierung gelernt hat, kennt den Begriff der „Ordnung eines

Algorithmus“. Klassischerweise erklärt man am Beispiel verschiedener Sortier-Algorithmen, wie sich der Zeitbedarf verhält, wenn man die Menge der Eingabe verdoppelt, verzehnfacht, vertausendfacht, etc.. Der „Quicksort“ hat eine Ordnung von O(n * log(n)).

<Es folgt eine kleine Live Demo über die Bedeutung der Ordnung für einen Algorithmus, anhand eines kleinen Programms „permute(s)“, welches Anagramme finden soll.>

In der Archivierung habe ich auch so eine Aufgabenstellung. Ich gedenke, mein Archiv stark wachsen zu lassen, und zwar in Bezug auf

- Anzahl der Bytes
- Anzahl der einzeln adressierbaren Objekte (Dokumente o.a.)

Es wäre sinnvoll zu prüfen, wie sich der Migrationsaufwand voraussichtlich entwickeln wird, wenn der Bestand sich verdoppelt, verzehnfacht, vertausendfacht.

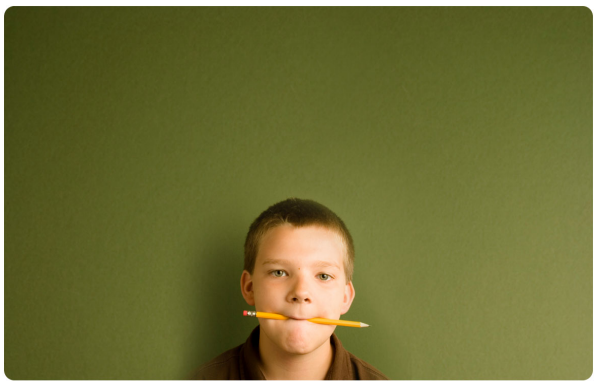
Im Ergebnis zeigt sich, daß Verfahren, bei denen nur der Inhalt von Datenträgern übertragen werden muß (die Bytefolge bleibt erhalten), sich konstant oder logarithmisch verhalten (also nicht einmal linear!), da sie in vollem Umfang von technischer Weiterentwicklung profitieren können. Iterative Ansätze hingegen sind oberhalb eines linearen Verhaltens ($n^{\mu,\ \mu>1}$). Das bedeutet, je größer mein Archiv wird, desto mehr Zeit brauche ich schon für die Migration *je Objekt.* Eine Entwicklung, die in eine undurchführbare Situation führen muß („Dead End Path“).

Stand heute (Fazit)

Ist das nur eine graue Theorie, ein Szenario, daß vielleicht nie eintritt? Nein! Noch sind undurchführbare (oder zumindest problematische) Archivmigrationen selten anzutreffen, doch es gibt sie bereits. Kundensituationen, in denen die Ablösung einer Komponente und die damit verbundene Migration nach erster Prüfung *Jahrzehnte* dauert, sind schon vorgekommen. Und auch wenn diese noch durch entsprechendes Geschick zu lösen sind, so ist doch klar, daß bei einer Verdopplung oder Verzehnfachung des Bestandes keine Chance mehr bestanden hätte.

Es wird mehr dieser Situationen geben, sie werden an Menge und Dramatik zunehmen. Wer wird schon kampflos unternehmenskritische Daten aufgeben? Die Aufmerksamkeit auf dieses Thema wird zunehmen, breiter werden, die Archivierungssysteme kritischer geprüft. Eines Tages erreicht es dann auch Sie.

Sind Sie bereit?



Kontaktadresse:

Thorsten Lange
Oracle Deutschland B.V. & Co. KG
Nagelsweg 55
D-20097 Hamburg

Telefon:	+49 (40) 251523-202
Fax:	-
E-Mail	thorsten.lange@oracle.com
Internet:	www.oracle.de