Information Lifecycle Management - Eine Methode zur Wertzuweisung von Dateien

Vom Fachbereich
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertationsschrift

von

Dipl.-Math. Lars Arne Turczyk

geboren am 15. Dezember 1969 in Frankenberg/Eder

Darmstadt 2009
Hochschulkennziffer D 17

Vorsitzender: Prof. Dr.-Ing. Gerd Balzer
Erstreferent: Prof. Dr.-Ing. Ralf Steinmetz
Korreferent: Prof. Dr. rer. nat. Paul Müller

Tag der Einreichung: 11. Dezember 2008
Tag der Disputation: 02. April 2009

Danksagung

Die vorliegende Dissertationsschrift entstand während meiner Zeit als externer Doktorand am von Prof. Dr.-Ing. Ralf Steinmetz geleiteten Fachgebiet Multimedia Kommunikation (KOM) an der Technischen Universität Darmstadt.

An erster Stelle möchte ich mich ganz herzlich bei Prof. Dr.-Ing. Ralf Steinmetz für die Betreuung und Förderung meiner Arbeit bedanken. Besonders bedanken möchte ich mich in diesem Zusammenhang dafür, dass Prof. Dr.-Ing. Ralf Steinmetz mir als externem Doktorand die Möglichkeit gegeben hat, Forschung an seinem Fachgebiet zu betreiben. Ebenfalls bedanken möchte ich mich bei Herrn Prof. Dr. rer. nat. Paul Müller (TU Kaiserslautern) für die Übernahme des Korreferats.

Bedanken möchte ich mich bei meinen Kollegen und ehemaligen Kollegen der Siemens AG, die mir die Teilnahme am kooperativen Promotionsprogramm des E-Finance Lab e.V. ermöglichten und mich während der Promotion Ziel führend unterstützten. Zu nennen sind insbesondere meine ehemaligen Vorgesetzten Horst Westerfeld, mittlerweile Staatssekretär der Hessischen Landesregierung, und Prof. Dr. Heinz-Theo Wagner, inzwischen Professor an der Hochschule Heilbronn, sowie Dr. Sertac Son, Mitstreiter in E-Finance Lab, und Dr. Martin Spreitzhofer, Vertreter der Siemens AG im Board des E-Finance Lab e.V.

Ich bedanke mich auch bei allen Professoren und Kollegen im E-Finance Lab e.V., allen voran Herrn Prof. Dr. Wolfgang König sowie seinen Mitarbeitern im Cluster 1, für die Zusammenarbeit im spannenden Umfeld des E-Finance Lab e.V.

Mein besonderer Dank gilt meinen Kollegen der Forschungsgruppe IT-Architekturen am Fachgebiet KOM Nicolas Repp, Julian Eckert, Stefan Schulte und Andrè Miede für die hervorragende, freundschaftliche Zusammenarbeit. Auch möchte ich mich bei Dr.-Ing. Oliver Heckmann, Dr. Andreas U. Mauthe, Dr.-Ing. Rainer Berbner und Dr.-Ing. Nicolas Liebau bedanken, deren Ideen und Anregungen diese Arbeit maßgeblich mitgeprägt haben.

Bedanken möchte ich mich weiterhin bei den von mir betreuten Studenten, die sich auf das Thema ILM eingelassen und in hervorragender Weise bearbeitet haben. Zu nennen sind hier Roswitha Gostner, Marcel Gröpl, Martin Behrens und Christian Frei. Zusammen mit Linda Moore als Lektorin waren sie an vielen Publikationen beteiligt.

Nicht zuletzt möchte ich mich bei meiner Lebensgefährtin Nina und unseren Kindern Arne und Greta bedanken, die mir durch ihre Unterstützung die Promotion an der TU Darmstadt erleichtert haben.

Inhaltsverzeichnis

1 Einführung

1.1 Motivation

Der wirtschaftliche Erfolg eines Unternehmens hängt in hohem Maße von qualifizierten Entscheidungen über die Informationstechnologie (IT) ab [47]. Heutzutage sind diese Entscheidungen insbesondere vor dem Hintergrund eines stetigen organisatorischen Wandels zu treffen [104, 106]. Konventionelle IT-Architekturen betrieblicher Informationssysteme stoßen in diesem beständigen Wandel in der Regel schnell an ihre Grenzen [104, 106]. Um strukturelle Analogien zwischen Unternehmensorganisationen und Informationssystemen zu entwickeln, wurden in den letzten Jahren unterschiedliche Ansätze verfolgt. Diese reichen von Mainframe-Konzepten, Server-Architekturen, intelligenten Clients, Virtualisierung bis zu Service-orientierten Architekturen (SOA) [104]. Eine ganzheitliche Problemsicht ist jedoch erst dann gegeben, wenn auch die Speicherung gemäß den Anforderungen des Geschäftsprozesses Berücksichtigung in einer IT-Architektur findet [73].

Unter den Stichworten „Tiered Storage“ und „Virtualisierung“ haben bereits zwei grundlegende Entwicklungen Einzug in die Speicherung gehalten [56]. Tiered Storage basiert auf verschiedenen Speicherebenen, wobei jede Ebene spezifische Serviceziele abbildet. Dadurch können Daten entsprechend ihrer Anforderungen gespeichert werden [56]. „Virtualisierung“ trennt die logische von der physikalischen Sicht auf die Infrastruktur. So können die Daten im Hintergrund verschoben werden, ohne den Zugriff der Anwendungen zu beeinträchtigen [56].

Information Lifecycle Management (ILM) ist ein auf diese Entwicklungen aufsetzendes Speicher Management-Konzept, welches die Vorteile dieser Trends ausnutzt und Informationen automatisch entsprechend ihres Wertes auf dem jeweils kostengünstigsten Speichermedium speichert und langfristig sicher aufbewahrt [51, 104]. Da berücksichtigt wird, dass sich der Wert von Informationen mit der Zeit ändert, ist ILM dynamisch [1, 7, 31, 51, 60]. Ein extremes Datenwachstum [45, 125] bei gleichzeitig länger werdenden Aufbewahrungszeiten [42, 43, 92] und ein immer größerer Kostendruck lassen bisherige Methoden der Speicherverwaltung an ihre Grenzen stoßen. Längst geht es nicht mehr darum, ausschließlich immer mehr Speicher bereitzustellen [2, 3, 104]. Gefragt sind vielmehr umfassende, dynamische Konzepte, die sich an Lebenszyklus und Nutzung der Informationen orientieren. Laut einer Studie von Glasshouse denken 90% der befragten IT-Entscheider an einen Einsatz von ILM [49]. Gleichzeitig sagen laut enginio 80% der dort befragten IT-Manager, ihnen fehle die Zeit, sich mit neuen Konzepten ausreichend auseinanderzusetzen [58]. Im Moment fehlt es noch an Erfahrungen mit ILM, um ILM beurteilen zu können. Insbesondere die Wertzuweisung ist ein offenes Thema. Diese sollte automatisiert erfolgen und idealerweise eine zuverlässige Prognose über zukünftige Dateizugriffe liefern [54].

Diese Arbeit stellt daher erstmals eine Methode vor, die die Wertzuweisung von Dateien mittels mathematischer Prognosen über das Nutzerverhalten durchführt.

1.2 Herausforderung

Ein wesentlicher Faktor für den Erfolg von ILM ist die automatisierte Wertzuweisung [54, 68, 116]. Hierzu liefert diese Arbeit einen Beitrag. Matthesius und Stelzer untersuchten bekannte Methoden der Wertzuweisung für ILM und verglichen diese hinsichtlich notwendiger Anforderungen. Sie konstatierten in ihren Analysen, dass keine Wertzuweisungsmethode für ILM existiert, die rechtliche Aspekte berücksichtigt oder Dateizugriffe prognostiziert (siehe Tabelle 1) [54]. Ersteres bedarf umfangreicher Kenntnis der aktuellen Gesetzeslage derjenigen Länder, in denen das betreffende Unternehmen aktiv ist [8, 112]. Diese Berücksichtigung liegt ausdrücklich nicht im Fokus dieser Arbeit. Stattdessen soll durch die vorliegende Arbeit die zweite Lücke „Prognose von Zugriffshäufigkeiten“ geschlossen werden.

Konzept / Anforderungen	Chen	Shah et al.	Baghwan et al.	Verma et al.	Mesnier et al.	Zadok et al.	Chandra et al.	Tanaka et al.
Zugriffscharakteristika								
Verwendung der Zugriffshäufigkeit	X	X	X	-	-	-	-	-
Klassifizierung								
Einteilung in Klassen	X	X	X	X	X	X	X	X
Compliance								
Berücksichtung rechtlicher Aspekte	-	-	-	-	-	-	-	-
Automatisierung								
Wertzuweisung	X	X	X	X	X	X	X	X
Verwendung Nutzer- und Administratorenwissen	-	X	-	X	X	X	X	X
Prognose von Zugriffshäufigkeiten	-	-	-	-	-	-	-	-
Kosten								
Berücksichtigung der Systemperformance	-	X	X	X	-	-	-	-
Kostenreduktion	X	X	X	X	X	X	X	X

Tabelle 1: Vergleich bekannter Wertzuweisungsansätze (angelehnt an [54])

1.3 Beitrag dieser Arbeit

Diese Arbeit entwickelt Erkenntnisse über ILM aufgrund realer Unternehmensdaten der Siemens AG. Ein Dateienpool mit Daten über mehr als 70.000 Dateien bildet dabei die Auswertungsbasis. Die Arbeit befasst sich mit der Wertzuweisung von Dateien, dem Hauptaspekt von ILM. Dazu wird eine neuartige Methode, die erstmals Prognosen von Dateizugriffen ermöglicht, entwickelt und ihre Leistungsfähigkeit belegt. Diese Methode wird in einem grundsätzlichen Kontext eines IT-Projekts eingebunden, wofür ein Vorgehensmodell entwickelt wird. In strikter Umsetzung des Vorgehensmodells wird zuerst mit Hilfe einer Fallstudie ein grundsätzliches Potenzial von 90% für ILM diagnostiziert. Die allgemeine Konzeption einer ILM-

Lösung wird vorgestellt, bevor das Thema der Wertzuweisung eingehend erforscht wird. Dieses ist essentiell für ILM und für das wesentliche Ziel von ILM, der Einsparung von Kosten. Alle bekannten Methoden führen eine Dateibewertung durch, jedoch ist keine in der Lage, Prognosen über die Wert bestimmende zukünftige Nachfrage des Dokuments zu erstellen. Die in dieser Arbeit vorgestellte Methode der Wertzuweisung von Dateien unterscheidet sich grundlegend von den bekannten Methoden, weil sie konsequent Zugriffswahrscheinlichkeiten prognostiziert und somit zukunftsgerichtete Aussagen ermöglicht.

Die Leistungsfähigkeit der neuen Methode wird mittels eines Simulators untersucht. Weiterhin wird anhand eines konkreten Beispiels zur Entscheidungsfindung über Migrationsregeln die Methode der Wahrscheinlichkeit zukünftiger Zugriffe mit vier anderen Methoden der Wertzuweisung verglichen. Dabei setzt sich diese Methode gegen die Alternativen durch.

Die Ziele dieser Arbeit liegen in der Beantwortung folgender Forschungsfragen:

1. Existiert Potenzial für ILM?
   In einer Fallstudie wird geklärt, wie groß der Prozentsatz an Dateien in einem Unternehmen ist, der von einer Einführung von ILM Kosten senkend auf preiswerte Speicherebenen migriert werden kann. Dies erlaubt Schlussfolgerungen für die Wertzuweisung und über das Einsparungspotenzial, welches durch Einsatz von ILM angestrebt werden kann.

2. Wie kann man das identifizierte Potenzial nutzbar machen?
   Es wird ein allgemeines Vorgehensmodell entwickelt, welches die Umsetzung von ILM in Form eines IT-Projektes darstellt.

3. Wie können Dateien auf Basis von automatisierten Prognosen bewertet werden?
   Es werden aus einer Unternehmensdatenbank Stichproben entnommen und mit Mitteln der angewandten Statistik auf mögliche Wahrscheinlichkeitsverteilungen untersucht.

4. Ist eine derartige Methode in ILM verwendbar?
   Die gefundene Methode muss ihre Leistungsfähigkeit in der gleichzeitigen Bewertung von tausenden von Dateien belegen. Dazu wird eine Simulationsumgebung implementiert.

1.4 Aufbau der Arbeit

In Kapitel 2 werden die grundlegenden, zum Verständnis der Arbeit notwendigen Begriffe erläutert. So werden ILM zugrunde liegende Definitionen aufgeführt und ein allgemeines Framework zur Umsetzung dieser Speicherstrategie vorgestellt.

In Kapitel 3 wird anhand einer Datenbank der Siemens AG untersucht, ob ausreichend Potenzial für ILM als Speicherkonzept vorliegt. Anschließend wird ein Vorgehensmodell entwickelt und die Konzeption einer ILM-Lösung allgemein beschrieben.

Die bekannten Methoden zur Wertzuweisung werden in Kapitel 4 vorgestellt und einem Vergleich anhand ausgewählter Kriterien unterzogen.

Die im Rahmen dieser Arbeit entwickelte „Methode der Wahrscheinlichkeit zukünftiger Zugriffe“ wird in Kapitel 5 erläutert. Diese Methode auf mathematischer Basis erlaubt die dynamische Wertzuweisung und unterstützt somit die Automatisierung von ILM basierend auf dem entwickelten Vorgehensmodell. Sie ermöglicht die dynamische Speicherung von Dateien auf die jeweilige Speicherhierarchie unter Berücksichtigung der zugesicherten Dienstgüteeigenschaften jeder einzelnen Hierarchie.

In Kapitel 6 wird, als konsequente Weiterentwicklung des zugrunde liegenden Ansatzes des ILM Prozesses, ein Simulationsmodell erarbeitet, das ermöglicht, tausende Dateien gleichzeitig über einen Lebenszyklus von mehreren Jahren zu beobachten. Dieses Kapitel endet mit der Darstellung eines Proof of Concept.

Jedes Unternehmen, das ILM einführen möchte, muss sich für eine oder mehrere Methoden der Wertzuweisung entscheiden. Diese Entscheidung ist aber mit „bloßem Auge“ nicht suffizient zu treffen, da die tatsächlichen Charakteristika von Wertzuweisungsmethoden sich erst im Betrieb oder vorher in der Simulation offenbaren. Daher wird in Kapitel 7 die Methode der Wahrscheinlichkeit zukünftiger Zugriffe mit verschiedenen Methoden der Wertzuweisung verglichen. Es werden die verursachten Speicherkosten, die Qualität der Methode und die erzielte Dienstgüte der Methoden an verschiedenen Szenarien ermittelt. Über eine Nutzenfunktion wird dann die Entscheidungsfindung über die passende Methode gemäß der nutzerspezifischen Restriktionen und Präferenzen des Unternehmens ermittelt.

Kapitel 8 fasst die wesentlichen Ergebnisse der Arbeit zusammen und gibt einen Ausblick auf Erweiterungsmöglichkeiten und weitere Einsatzgebiete.

2 ILM-Grundlagen

In diesem Kapitel wird der Schlüsselbegriff dieser Arbeit, ILM, vorgestellt. Es werden eine Definition, die Herausforderungen an die IT und die daraus folgenden ILM-Ziele behandelt. Anschließend wird das aus der Storage Networking Industry Association (SNIA) entstandene ILM-Framework, das die beteiligten ILM-Teilsegmente darstellt, erläutert. Das Kapitel schließt mit einer Abgrenzung zu Hierarchischem Storage Management, HSM, einem Vorläufer von ILM, und einer Zusammenfassung.

2.1 Definition von ILM

Zu ILM existieren verschiedene Definitionen [8, 46, 73, 117]. Mit dem Begriff und den Inhalten von ILM beschäftigen sich sowohl Beratungshäuser und Speicherhersteller als auch Verbände wie der BITKOM (Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V.) oder die SNIA. Alle sind bestrebt, ILM zu definieren und zu interpretieren.

Im Sommer 2004 stellte die SNIA ihre Definition des Begriffes ILM vor, welche von der ILM-Arbeitsgruppe erarbeitet wurde: Diese beschäftigt sich überwiegend mit den „best practices" rund um ILM.

Definition 1 (Information Lifecycle Management nach SNIA): *ILM beinhaltet Policies, Prozesse, Anwendungen und Werkzeuge, die den Wert einer Information innerhalb eines Geschäftsmodells mit der adäquatesten und kosteneffizientesten IT-Infrastruktur verbinden; und zwar von der Erstellung einer Information bis zur endgültigen Archivierung oder Löschung. Informationen sind abhängig von Geschäftsmodellen, Managementstrategien und Dienstgütevereinbarungen, welche mit Applikationen, Metadaten, Informationen und Daten verknüpft sind [73].*

Diese Definition gibt das Grundverständnis dieser Arbeit von ILM wieder. Informationen stellen ein zentrales Element von ILM dar. Die SNIA hat diesen Begriff ebenfalls eigens definiert:

Definition 2 (Information): *Informationen sind Daten, die durch eine Anwendung oder einen Prozess ausgetauscht, ausgedrückt oder dargestellt werden können [73].*

Mit „Strategien, Prozessen, Anwendungen, Leistungen und Werkzeugen" ist die Vorgehensweise des Datenzentrums gemeint, in dem die Migration von Daten zwischen verschiedenen hierarchischen Speicherebenen durchgeführt wird. Dort werden ebenfalls beschreibende Daten über Daten, so genannte Metadaten, erstellt. Informationen werden beim ILM von der „Erstellung bis zur endgültigen Archivierung oder Löschung" betrachtet, das heißt, ILM beinhaltet die Bereitstellung einer Infrastruktur, die den jeweiligen Wert einer Information vom Zeitpunkt der Erstellung bis zur endgültigen Archivierung oder Löschung berücksichtigt [73]. Da sich der Wert der Informationen mit der Zeit ändert und somit auch deren Verfügbarkeitsnotwendigkeit, sorgt ILM für Automatismen, die die Informationen optimal in einer hierarchischen Speicherlandschaft ablegen. ILM behandelt das hierarchische Speichermanagement nicht nur aus Sicht der IT-Infrastruktur, sondern auch aus der Management-Perspektive, in-

dem „Geschäftsmodelle, Managementstrategien und Dienstgütevereinbarungen" berücksichtigt werden, welche sich am Geschäftsmodell orientieren. Hierbei werden Richtlinien für die Informationsverwaltung aufgestellt, nach welchen dann die Klassifizierung des Wertes einer Information erfolgt. Dadurch können die Informationen in verschiedene Wertigkeitsklassen eingeteilt und in einem kosteneffizienten Speichermedium gespeichert werden, ohne die Dienstgütevereinbarungen zu verletzen. Auch die Personen, die mit den Daten umgehen, werden beim ILM miteinbezogen. Es müssen für jeden zugängliche Richtlinien für den Einzelanwender vorhanden sein. ILM stellt also eine Verbindung zwischen der Speicherinfrastruktur und dem Geschäftsmodell her und berücksichtigt außerdem die Personen innerhalb der Prozesse. Insgesamt ist diese Definition treffend, aber sehr umfangreich, weil sich möglichst alle Aspekte von ILM widerspiegeln sollen.

2.2 Abgrenzung von ILM zu Hierarchischem Speichermanagement (HSM)

Hierarchisches Speichermanagement (HSM) verfolgt seit den 70er-Jahren bereits die Verwendung mehrerer Ebenen zur Speicherung [12]. HSM dient der automatischen Migration von Dateien, auf die eine bestimmte Zeit nicht zugegriffen wurde, von schnellen auf langsamere, billigere Speicher [18].

Durch die Verknüpfung des Wertes einer Information mit der IT-Infrastruktur bringt ILM eine neue Qualität in die Verwaltung von Information, die weit über das bisherige Speichermanagement hinausgeht. In der Vergangenheit wurden Speichermanagementlösungen mehr oder weniger nur als Hardwarekomponenten angesehen, die nach und nach mit Informationen gefüllt wurden. Darüber hinaus benötigte Speicherkapazitäten wurden durch die Anschaffung neuer Hardwarekomponenten beschafft. Sicherheitsanforderungen, Auslagerungsstrategien und die Einbindung in Netzwerke führten zwar zu kombinierten Hardware-Softwarelösungen wie HSM, aber letztlich war der Fokus der Lösungen auf die Verwaltung der Speicherkomponenten ausgerichtet [104]. HSM migriert Dateien bei Erreichen eines Schwellwertes (z.B. 80% Auslastung der Speichersysteme). Entscheidungskriterien sind dabei Dateigröße und Dateialter.

Die heute angewandten Prozesse in der Speicherverwaltung werden den Anforderungen der Zukunft nicht mehr gerecht. ILM nimmt das auf und orientiert sich bei der Speicherung von Informationen an deren Lebenszyklus und Nutzung. ILM berücksichtigt damit sowohl die Aspekte dynamisch veränderlicher Information zu Beginn des Lebenszyklus als auch die Langzeitarchivierung. Es erfolgt keine Einengung auf bestimmte Ausprägungen und Typen von Informationen. Damit wird es auch möglich, Programmversionen, Daten, Datenbanken und beliebige Inhalte in die Verwaltung einzubeziehen [68, 104].

Fred Moore, ein ILM-Pionier, hat den HSM-Ansatz erweitert und ILM geprägt. Gemäß den von Moore formulierten neuen Ansätzen ist es immer wichtiger zu verstehen, wo Daten während ihres Lebenszyklus idealer Weise gespeichert und wie sie verwaltet werden sollten. Er postuliert für die Mehrzahl digitaler Daten die Maxime „90 Tage auf Platte und 90 Jahre auf Magnetband“ [62] und stellt dabei die These auf, Daten sollten 30 Tage nach ihrer Erstellung auf Ebene-1-Speichersystemen, vom 31. bis 90. Tag ihrer Existenz auf Ebene-2-Speicher und danach auf Ebene-3-Speicher aufbewahrt werden [63]. Die von Moore postulierte Fixierung auf 90 Tage unterstellt einen simplifizierten Verlauf der Wiederverwendungswahrscheinlichkeit, die nach 90 Tage gegen Null tendiert (siehe Abbildung 1).

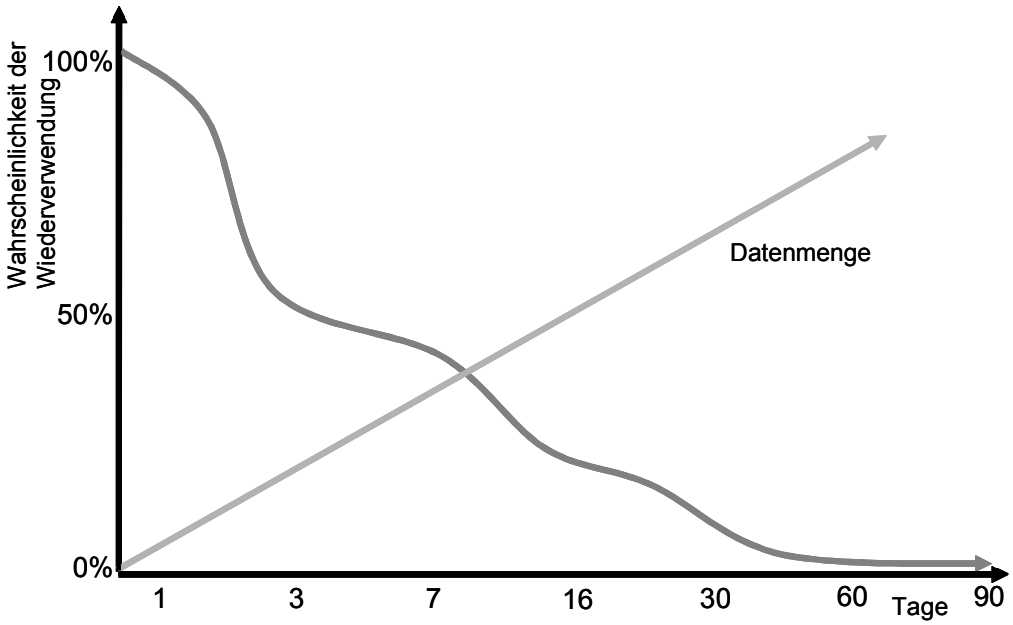


Abbildung 1: Verlauf der Wahrscheinlichkeit der Wiederverwendung nach Moore (2003) [62]

Dieser Ansatz aus 2003 postuliert zwar die Berücksichtigung von Wiederverwendungswahrscheinlichkeiten, diese werden aber zu vereinfacht dargestellt, indem man davon ausgeht, die Wahrscheinlichkeiten konvergieren nach 90 Tage gegen Null.

Auch noch im Jahr 2007 vertreten Sollbach und Thome einen ähnlichen Ansatz der Wahrscheinlichkeiten [104]. Das Problem an diesen Ansätzen ist, dass die Wahrscheinlichkeitsverteilungen nicht spezifiziert werden. Folglich wird in Ermangelung der Wahrscheinlichkeitsverteilungen weiterhin mit den Moore’schen Speicherfristen gearbeitet (90 Tage). Selbst in wissenschaftlichen ILM-Abhandlungen wird dieser stark vereinfachte Ansatz angenommen [102].

2.3 Herausforderungen

Die Definition von ILM ist sehr breit angelegt. Sie spricht damit eine große Zielgruppe an. Dies spiegelt sich darin wieder, dass laut einer Studie von Glasshouse 90% der befragten IT-Entscheider an einen Einsatz von ILM denken [49]. Gleichzeitig sagen aber laut enginio 80%

der dort befragten IT-Manager, ihnen fehle die Zeit, sich mit neuen Konzepten ausreichend auseinanderzusetzen [58]. Dass neue Konzepte benötigt werden, ist unstrittig angesichts der drängenden Herausforderungen, denen sich IT-Manager bezüglich ihrer IT konfrontiert sehen. Die Herausforderungen, mit welchen sich heutige Unternehmen konfrontiert sehen, sind [8, 45, 46, 106, 125]:

- Datenwachstum
- Vorschriften und Regeln
- Wert der Daten im Laufe des Lebenszyklus
- Kostendruck
- Neue Technologien

Diese werden in den folgenden Abschnitten näher erläutert.

2.3.1 Datenwachstum

Der Speicherbedarf nimmt immer weiter zu, denn die Menge von Informationen steigt stetig. Eine Schätzung der Universität von Kalifornien besagt, dass allein im Jahr 2002 ungefähr fünf Exabytes (10^{18} Bytes) neuer zu speichernder Informationen produziert wurden [50]. Mit der Menge von Informationen steigen auch die Systemkosten, die Hardwarekosten und vor allem die Verwaltungskosten. Das betrifft in besonderem Maße Unternehmen und Institutionen, aber auch Privathaushalte. Das bekannteste Beispiel ist der Email-Bereich: Seit 2001 hat sich die Zahl der täglich versandten Emails ungefähr verdreifacht. Man geht davon aus, dass ein Unternehmen mit ca. 3000 Mitarbeitern ungefähr ein Terabyte an Email-Daten pro Jahr produziert [8].

Ein weiterer Bereich mit extrem hohem Datenwachstum ist das Gesundheitswesen [8, 46]. Dort entstehen große Datenmengen durch neue Technologien, wie z.B. Computertomographen. Diese Daten verbrauchen extrem viel Speicherplatz und müssen zudem sehr lange archiviert werden. Auch in anderen Bereichen, wie z.B. im Bankensektor oder bei Versicherungen, gibt es ein ähnlich hohes Datenwachstum [112].

2008 hat IDC neue Erkenntnisse zu Wachstum und Inhalten des weltweiten digitalen Datenvolumens bis 2011 veröffentlicht [28]. 2007 betrug die Datenmenge des "Digitalen Universums" 281 Exabyte und damit zehn Prozent mehr als ursprünglich vorhergesagt. Für diese Datenmassen bräuchte man rund 17 Milliarden iPhones® mit 8-GB-Speicherplatz. Derzeit wächst die digitale Informationsflut, so IDC weiter, jährlich um 60 Prozent und wird bis 2011 rund 1.800 Exabyte (1,8 Zettabytes) erreichen. Dies entspricht einer Verzehnfachung gegenüber 2006 [28].

Weiter sind laut der Studie die IT-Abteilungen von Organisationen jeder Größenordnung und Branche bei rund 85 Prozent der entstehenden Daten in irgendeiner Form in die Speicherung,

Bereitstellung, Übermittlung, Einhaltung von Datenschutzrichtlinien sowie den Schutz der Daten involviert [28].

2.3.2 Vorschriften und Regeln

Ein nationales Unternehmen in Deutschland muss verschiedene Gesetze beachten. Zu beachten sind von Seiten des Gesetzgebers vor allem die Grundsätze der ordnungsgemäßen Buchführung (GoB), die Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GDPdU), das Handelsgesetzbuch (HGB) und die Abgabenordnung (AO) [15].

Für in den USA tätige Unternehmen kommen zusätzliche Vorschriften hinzu. Die Storage Networking Industry Association (SNIA) führt 11 verschiedene amerikanische Regularien auf, wie zum Beispiel solche der Securities and Exchange Commission (SEC) (z.B. SEC 34-47638[1], SEC 17a-4[2] oder Sarbanes-Oxley-Act) oder des US-Senats zum Thema Gesundheitswesen (Health Insurance Portability and Accountability Act (HIPAA)) [72]. Die Erfüllung der Mindestanforderungen solcher Regelwerke auf Basis der IT wird auch als „Compliance" bezeichnet.

Es ist Aufgabe der Verantwortlichen, in Zusammenarbeit mit externen und internen Revisionen, die Aufbewahrungsfristen festzulegen. Aber bereits heute stoßen viele IT-Chefs angesichts der Fülle von Vorschriften und Regeln an ihre Grenzen. So besagt eine aktuelle Studie unter 100 britischen IT-Leitern, dass 88 Prozent einen direkten Einfluss der Gesetzgebung auf ihre IT-Budgets sehen, aber 71 Prozent fühlen sich nicht vollständig darüber informiert [58].

Für ILM spricht, dass generell die meisten Bestimmungen technologie-neutral zu sein scheinen, so dass ein großer Handlungsspielraum in der technischen Umsetzung gegeben ist [46].

2.3.3 Wert der Daten im Laufe des Lebenszyklus

Angesichts des starken Datenwachstums und der regulatorischen Bedingungen gilt es, sehr viele Daten unter Umständen sehr lange aufzubewahren. Weiterhin sollen Daten, die für das Unternehmen keine weitere Bedeutung haben, aussortiert und gelöscht werden. Dies entspricht einer ständigen Wertermittlung von Daten. Als Resultat wird der Lebenszyklus der Daten festgelegt sowie das passende Speichermedium ermittelt.

Der Wert von Daten ändert sich in der Zeit von ihrer Erstellung bis zu ihrer Löschung in der Regel erheblich [8]. Ein alltägliches Beispiel dafür sind Emails, die typischerweise am Tag ihres Eintreffens von Bedeutung sind, aber sobald sie einmal gelesen wurden, praktisch keinen Wert mehr haben. Ein weiteres Beispiel sind Daten elektronischer Terminplanung, die sehr wertvoll sein können. Aber sobald ein Tag vorbei ist, haben die Daten der Terminplanung in der Regel nur noch dokumentatorische Aufgaben. Ein Gegenbeispiel, bei dem die

---

[1] "Interagency Paper on Sound Practices to Strengthen the Resilience of the U.S. Financial System" herausgegeben von der Securities and Exchange Commision (SEC)

[2] Records to Be Preserved by Certain Exchange Members, Brokers and Dealers

Daten kaum an Wert verlieren, sind Stammdaten von Kunden, die evtl. über Jahrzehnte gesammelt wurden und die für ein Unternehmen von existentieller Bedeutung sind und es auch in Zukunft bleiben. In manchen Fällen müssen dagegen Daten nach einer bestimmten Zeit gelöscht werden. Dazu gehören zum Beispiel die Aufzeichnungen einer Videoüberwachung oder Einzelgesprächsnachweise der Telekommunikationsprovider [15].

Die Wertermittlung von Microsoft-Office-Dateien steht bei dieser Arbeit im Mittelpunkt. Es zeigt sich, dass eine Wertfindung von Dateien über das Zugriffsverhalten erfolgen kann [7, 11, 13, 90]. Bei einer Studie zur Wertbestimmung von Informationsobjekten wurde das Zugriffsverhalten von drei NFS-Dateiservern der Harvard University beobachtet. Für den untersuchten Bereich wurden drei typische Zugriffsmuster identifiziert [13]:

- Zugriffshäufungen zu Beginn des Lebenszyklus (bursty),
- periodische Zugriffshäufungen (periodic)
- konstante Zugriffe (constant)

Auch bei beliebigem Zugriffsverhalten jenseits dieser Zugriffsmuster auf Dateien kann eine passende Methode den Wert abbilden [114]. Dies ermöglicht es IT-Administratoren, jede Datei bewertet zu bekommen. Mit Hilfe von IT-Management-Werkzeugen kann daraufhin eine Wert entsprechende Speicherung über einen definierten Zeitraum erfolgen.

2.3.4 Kostendruck

Unternehmen stehen unter stetigem Kostendruck [49, 58, 74]. Der verschärfte Wettbewerb vor allem durch die Globalisierung und die Öffnung der Märkte in Europa wird begleitet von häufig zu spät eingeleiteten oder unzureichenden Maßnahmen der Umstrukturierung in Politik und Wirtschaft. Diese Umstände zwingen Unternehmen dazu, die Produktivität zu steigern und die Kosten zu senken. Der Bereich des Speichermanagements birgt Potential zur Senkung von Personal-, Technologie- und Prozesskosten. Durch automatisierte Prozesse im Speichermanagement werden vorhandene Technologien optimal ausgenutzt, wodurch sich z.B. die Kosten für neue Speicherkapazität reduzieren. Außerdem ergibt sich durch geringeren Verwaltungsaufwand und optimale Bereitstellung von Informationen eine Reduzierung der Personalkosten.

Die IT-Systeme deutscher Unternehmen sind zu zwei Dritteln bereits abgeschrieben. Zudem haben drei von vier CIOs keine Budgethoheit, sondern sind relativ häufig IT-fremden Vorstandsressorts unterstellt [122]. Hierzulande gilt die Informationstechnik in den meisten Unternehmen immer noch als Kostenstelle. Entsprechend schlecht sind Geschäft und IT aufeinander abgestimmt [122].

2008 sorgt der wirtschaftliche Aufschwung zu Jahresbeginn für eine zunehmende Bedeutung von Wachstum und Innovation. Bei den IT-Investitionen stehen Wachstum, die Verbesserung der Leistungsfähigkeit und die Umsetzung innovativer Geschäftsmodelle ganz oben auf der Prioritätenliste. Gleichzeitig mangelt es aber an einer adäquaten Aufstockung der IT-Budgets.

Die widersprüchlichen Anforderungen an die CIOs haben allerdings zu einem wahren Aufschwung bei IT-Beratern geführt [9]. Denn neben dem unvermindert hohen Kostendruck wird ein CIO zunehmend auch immer mehr als Prozessgestalter und Prozessintegrator im Unternehmen aktiv. Die Unterstützung der CIOs bei dieser Aufgabe liegt bei externen Beratern im direkten Fokus [52].

2.3.5 Neue Technologien

In einer zunehmend dynamischen Welt bei einer Intensivierung des Wettbewerbs sehen sich Unternehmen ständig neuen Herausforderungen ausgesetzt. Das Erlangen von Wettbewerbsvorteilen stellt sich zur Absicherung der weiteren Existenz eines Unternehmens als zentrale Aufgabe dar [47]. Vor allem fundamentale technologische Neuerungen aus dem Bereich der Informationstechnologien werden hierbei als strategischer Wettbewerbsfaktor zur Erhöhung der betrieblichen Leistungsfähigkeit gesehen [47]. Die Auswahl und Nutzung geeigneter Innovationen aus diesem Bereich verläuft jedoch nicht immer erfolgreich und wird durch die stets kürzer werdenden Innovationszyklen in Verbindung mit steigender Komplexität erschwert, wodurch eine ständige Auseinandersetzung mit neuen IT-Innovationen zunehmende Bedeutung erlangt [99].

Neuerungen der Speichertechnologien, wie zum Beispiel S-ATA-Platten (Serial-Advanced Technology Attachement) und Fibre Channel/SCSI-Platten, erweitern das klassische Speichermodell [8]. S-ATA wurde aus dem parallelen ATA (Advanced Technology Attachement), oder auch IDE (Integrated Drive Electronics), weiter entwickelt und überträgt Daten nun nicht mehr in 16-Bit Paketen sondern Bit für Bit seriell. Durch die Weiterentwicklung besitzt S-ATA drei Vorteile, die die Festplatten auch für Speicherlösungen interessant machen. Diese sind Geschwindigkeiten bis zu 600 MB/sec, Verkabelung mit 7-adrigen seriellen Kabeln statt 80-adrigen parallelen Kabeln sowie Plug and Play-Funktionalität, womit die Platten mit einem unterstützenden Betriebssystem (BS) im laufenden Betrieb des Busses getrennt werden können.

SCSI (Small Computer System Interface) wurde 1986 vom ANSI (American National Standards Institute) standardisiert und wird ständig weiter entwickelt. Bei professionellen Storagelösungen kommen praktisch nur SCSI-Festplatten zum Einsatz. Sie sind schnell und besonders langlebig und daher erste Wahl in Servern, RAID-Verbünden und SANs. Man unterscheidet bei SCSI zwischen zwei Typen von Geräten, den Initiatoren und den Targets. Eine Datenübertragung geht immer von einem Initiator aus. Deshalb bildet in einem Server der SCSI-Controller den Initiator ab, denn er beginnt eine Transaktion. Festplatten stellen intern die Targets dar. Sie teilen ihren Speicherplatz in Logical Units (LUN) auf [56, 103].

Fibre Channel (FC) ist eine eigenständige Netzwerktechnik und wurde ausschließlich für schnelle Datenübertragungen entwickelt. Dabei bietet es die Möglichkeit, verschiedene Protokolle (z.B. SCSI, ESCON, SNMP) zu übertragen. Mit der Entwicklung von Fibre Channel wurde im Jahr 1988 begonnen. 1994 wurde FC dann durch ANSI standardisiert und verfügt heute über eine hohe Marktreife [81]. Hauptsächlich wird FC zum Transport von SCSI-3-Paketen eingesetzt. Da es auch Festplatten mit FC-Anschlüssen gibt, wird es auch als direkte

Verbindungsmöglichkeit zwischen CPU und Peripherie angesehen. Dabei überträgt es in diesem Fall „nur“ SCSI-Signale. Genau genommen ist FC eine Mischung aus Bus und Netzwerk und vereinigt die Vorzüge beider Technologien miteinander. FC kommt, wie bei Bussen üblich, mit wenig Overhead aus und bietet eine schnelle, serielle Verbindung zwischen den Teilnehmern. Aus der Netzwerktechnik wurde die große Anzahl von verwaltbaren Endgeräten und die hohe Flexibilität übernommen [56, 103].

Wenn man die verschiedenen Technologien im Online- und Nearline-Bereich betrachtet, gibt es heutzutage mindestens drei Haupthierarchien. Als erstes kommt der hoch performante Online-Bereich, der sich durch die unverzügliche Verfügbarkeit von Daten auszeichnet. Im Online-Bereich können in wenigen Millisekunden Daten ein- bzw. ausgelesen werden. Die zweite Hierarchieebene bilden Systeme mit S-ATA-Platten, die auch einen direkten Zugriff gewährleisten, allerdings mit geringeren Durchsatzwerten und weniger Kosten pro Megabyte. Früher waren IDE-Festplatten langsam und nur für den Einsatz in PCs konzipiert. Heute hat sich das Einsatzgebiet gewandelt. Zwar reichen S-ATA-Platten bei der mechanischen Belastung immer noch nicht an SCSI-Platten heran und darum wird weiterhin davon abgeraten, auf ihnen wichtige, produktive Daten dauerhaft zu speichern. Die S-ATA-Platten werden hauptsächlich zur Archivierung und in der Platte-zu-Platte Datensicherung eingesetzt. Die dritte Hierarchieebene ist der Nearline-Bereich mit den Bandtechnologien, der durch geringe Kosten pro Megabyte und hohe Zugriffszeiten gekennzeichnet ist. Wenn die Bänder in automatischen Bibliotheken verwaltet werden, betragen die Zugriffszeiten zwischen 10 Sekunden und einigen Minuten. Als eine vierte Hierarchieebene kann man noch eine Auslagerungsebene betrachten, die die geringsten Kosten pro Megabyte verursacht. Die Zugriffszeiten dieser Ebene sind deutlich größer als im Falle des Nearline-Bereiches, da die Datenträger manuell bereitgestellt werden. Für einen Vergleich der Kosten verschiedener Speicherebenen empfiehlt es sich, die Kosten aller Ebenen auf die Kosten der ersten oder der letzten Ebene zu normieren. So ergeben sich zum Beispiel für ein Speichersystem mit den drei Speicherebenen bestehend aus Enterprise Disk (FC), Low Cost Disk (S-ATA) und Automated Tape folgende Kostenrelationen [62]:

FC : S-ATA : Automated Tape = 45 : 7,5 : 1

Die Herausforderung an neuen Technologien ist, dass es für Unternehmen unumgänglich ist, sich mit der Nutzung bereits vorhandener Innovationen intensiv auseinander zu setzen, um wettbewerbsfähig zu bleiben. Richtet sich der Fokus der Betrachtung speziell auf IT-Innovationen, so zeigt sich, dass aufgrund immer kürzerer Innovationszyklen [99] aus Unternehmenssicht gerade dem Bereich des Managements von Informationstechnologien eine Schlüsselrolle zukommen sollte, da die IT auch zukünftig ein entscheidender Faktor bleiben wird [47].

Zusammenfassend lässt sich feststellen, dass die Herausforderungen vielschichtig genährt sind. ILM ist demzufolge als ein neues Konzept zur Datenspeicherung entstanden, welches konkreten Zielen gerecht werden soll. Aus der Definition und den Herausforderungen ergeben sich nun die Ziele von ILM.

2.4 Ziele von ILM

Als strategischer Lösungsansatz für die Herausforderungen moderner Unternehmen muss ILM folgende Hauptziele erfüllen [46]:

- Aufbewahrung der Information zur richtigen Zeit am richtigen Ort
- Kostenreduktion
- Einhaltung gesetzlicher und regulatorischer Vorgaben
- Erfüllung von SLAs und QoS-Anforderungen

ILM hat dafür zu sorgen, dass die Informationen jederzeit mit möglichst geringer zeitlicher Verzögerung zur Verfügung gestellt werden können. Üblicherweise geschieht dies durch Enterprise-Speicher-Technologie. Andererseits ist es aber auch ein Ziel von ILM, die Informationen kosteneffizient zu speichern. Dies geschieht durch Verwendung mehrerer Speicherhierarchien und Migration von Dateien auf günstigere Speicherebenen. ILM versucht, ein Optimum hinsichtlich beider Aufgabenstellungen zu ermöglichen [46]. Als zusätzliche Nebenbedingungen sind außerdem gesetzliche und organisatorische Vorgaben, sowie SLAs und Dienstgüteanforderungen einzuhalten.

Die individuellen Anforderungen eines Unternehmens wirken sich ebenfalls auf die Ziele von ILM aus. So könnten beispielhafte Nebenziele einer Implementierung folgendermaßen lauten [4]:

- Optimale Ausnutzung vorhandener Speicherressourcen
- Loadbalancing

2.5 ILM-Framework

In diesem Abschnitt wird das von der SNIA erarbeitete ILM-Framework erläutert (siehe Abbildung 2). Angesichts der Komplexität von ILM verfolgt die SNIA den Ansatz der „best practices“, womit relativ schnell Erfahrungen über Teilgebiete von ILM generiert werden sollen. Dass verschiedene Teilergebnisse sich ergänzen können, gewährleistet ein vorgegebenes ILM-Framework [73].

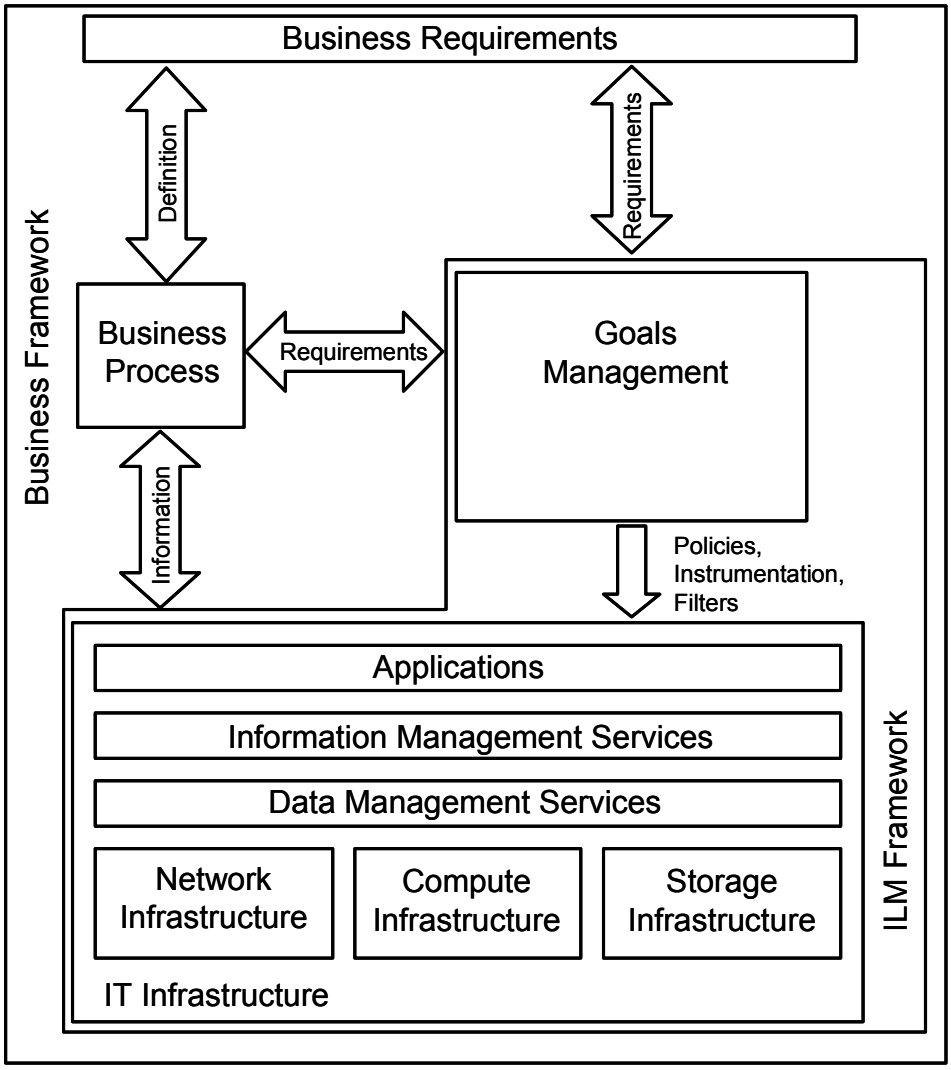


Abbildung 2: ILM-Framework [73]

Das ILM-Framework stellt die Beziehung zwischen den betriebswirtschaftlichen Anforderungen eines Geschäftsprozesses und den dafür erforderlichen Informationen sowie der IT-Infrastruktur her [73].

Das dem ILM-Framework zugrunde liegende Business Framework bildet die Ziele des Geschäftsmodells ab und legt somit die Verwendung der Informationen fest. Über eine Analyse werden individuelle Prozesse extrahiert, um zu prüfen, in welchem Zusammenhang sie mit dem ILM-Modell stehen. Dabei sind die betriebswirtschaftlichen Anforderungen die Basis, auf der das ILM-Framework ansetzt.

2.5.1 Goals Management

Das Goals Management beschreibt die Unternehmensziele in Form von Strategien für die Implementierung in Netzwerk-, Computer- und Speicherinfrastrukturen. Es gibt dem Unternehmen Rückmeldungen bezüglich Kosten und Risiken.

Das Goals Management wird in vier Aufgabenbereiche eingeteilt [73]:

- Business Classification: Auf Applikationsebene werden Geschäftsklassifikationen festgelegt. Diese Klassifikationen sind herstellerspezifisch.
- Transformation: Herstellerspezifische Algorithmen transformieren die Geschäftsklassifikationen in herstellerneutrale Service-Klassifikationen.
- Service-Classification: Service-Klassifikationen beschreiben die kontroll- und prozessbezogenen Anforderungen, ohne dabei konkrete technische Lösungen zu definieren.
- Service Level Objectives (SLO): Ein SLO (Dienstgütekriterium) ist ein Aspekt oder eine Dimension, wie Datenmanagement-Ressourcen spezifiziert und messbar bereitgestellt werden können. Zusammen mit den Richtlinien kann dann eine allgemein gültige Perspektive der Business-Klassifikation erstellt werden.

2.5.2 IT Infrastructure

Applikationen im Bereich der IT-Infrastruktur sind alle Anwendungen, die Daten von der IT-Infrastruktur in Informationen transformieren können, die im Geschäftsprozess benötigt werden [73]. Zukünftige Applikationen sollen eine Schnittstelle zum ILM-Framework bekommen, damit der Benutzer durch Beeinflussung der Dienstgütekriterien die automatisierten Prozesse unterstützen kann.

Der *Information Management Service* führt die einzelnen Informationen und Speicherobjekte durch die verschiedenen Stadien innerhalb der Geschäftsprozesse. Dieser Service nutzt inhaltliche Informationen oder Metadaten für seine Entscheidungsfindung. Für die Ausführung seiner Dienste greift er auf den Data Management Service oder die Netzwerk-, Rechner- und Speicherinfrastruktur zurück.

Data Management Services übernehmen die Kontrolle der Daten vom Zeitpunkt ihrer Entstehung bis zur Löschung. Die Services sind nicht selbst Teil des Datenflusses, sondern lenken ihn. Zu den Services gehören Datenbewegung, Datenabgleich auf Redundanz und Datenlöschung.

Mit *Netzwerkinfrastruktur* wird jene Software und Hardware beschrieben, die eine Verbindung zwischen den einzelnen Computern und Speicherzentren herstellt. Sie beinhaltet LAN- und SAN-verbundene Elemente sowie Schnittstellen zu Servern.

Die *Rechnerinfrastruktur* definiert die verfügbare Rechenleistung eines Unternehmens. Darin sind neben den Servern auch die Rechner der einzelnen Mitarbeiter, deren Betriebssysteme und Hilfswerkzeuge für die Bedienung enthalten.

Die *Speicherinfrastruktur* stellt Speicherkapazitäten in Form von Software und Hardware zur Verfügung und schließt dabei die aktuellen Technologien wie NAS (Network Attached Storage), SAN (Storage Area Network), CAS (Content Addressed Storage) sowie Volume Management ein, wobei im Modell die konkreten Technologien flexibel austauschbar sind.

Die preisliche Darstellung im Storagebereich ist aufgrund eines geschlossenen Marktes, in dem quasi nur die Hersteller selbst agieren, sehr schwer. Man kann eine grobe Faustregel aufstellen, dass sich die Anschaffungskosten eines SAN im Midrange-Bereich in ca. drei gleich große Teile aufteilen lassen [123]:

1. Infrastruktur

Darin sind enthalten: SAN-Switche, Host Bus Adapter (HBAs), Kabel, etc.

2. Speicherhardware

Darin sind enthalten: Redundanter Controller und Festplatten

3. Software

Werksseitig ist die Speicherhardware immer nur mit einer rudimentären Managementsoftware ausgestattet. Benötigt man noch Zusatzfunktionen wie Snapshots oder SRM (Storage Resource Management), so muss man diese dazu kaufen.

Das ILM-Framework repräsentiert das Schaubild einer funktionierenden ILM-Lösung. Wie man zu einer Lösung kommt, beschreibt ein Vorgehensmodell.

2.6 Zusammenfassung

Information Lifecycle Management ist ein 2003 aufgekommenes Storage-Konzept, das die Hauptherausforderungen von IT-Managern adressiert. Durch den Standardisierungsansatz der Storage Networking Industry Association ergeben sich unterschiedliche Forschungsbereiche zum Thema ILM, die langfristig bearbeitet werden müssen. Das ILM-Framework bietet die Basis, die unterschiedlichen Aktivitäten zu strukturieren und erste Erfahrungen mittels „best-practice“-Vorgehen zu generieren.

Die vorliegende Arbeit fokussiert sich auf Abbildung der IT-Unternehmensziele (Goals Management) und die angrenzenden Bereiche, der IT-Infrastruktur sowie der Geschäftsanforderungen (Business Requirements) (siehe Abbildung 3).

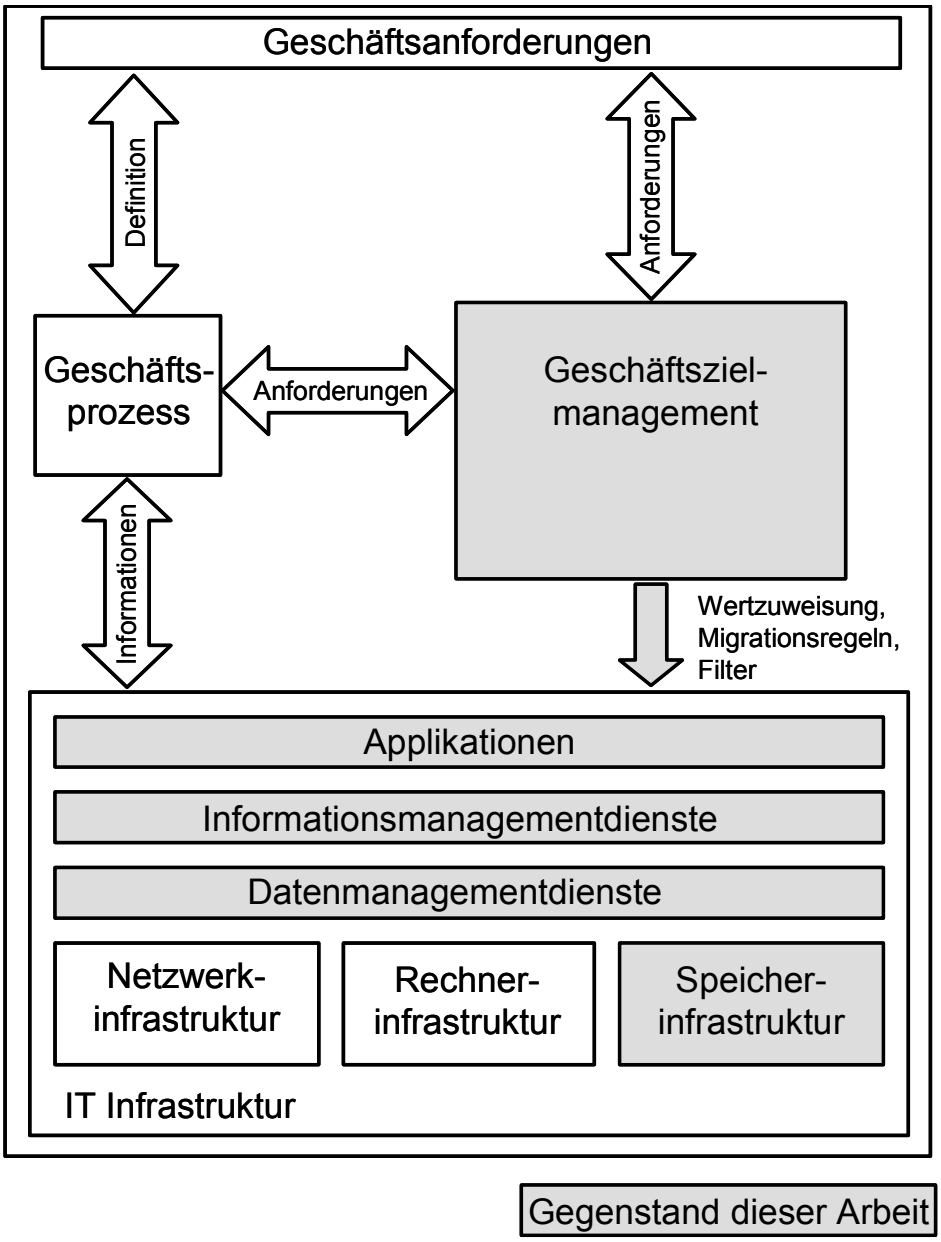


Abbildung 3: ILM-Framework angelehnt an [73]

ILM basiert auf einer Wertbetrachtung von Informationen. Hierin unterscheidet sich ILM vom bereits bekannten HSM, bei dem Dateien nach zeitlichen Kriterien migriert werden. Über das Kriterium Zeit lässt sich ebenfalls der Wert von Dateien definieren. Diese Idee ist heutzutage am weitesten verbreitet, basiert aber auf einem vereinfachten Wahrscheinlichkeitsverlauf von Dateizugriffen. Im weiteren Verlauf wird eine neue Methode der Wertfindung hergeleitet, die diese Wahrscheinlichkeiten konkret mathematisch beschreibt.

Im nächsten Kapitel wird ein Vorgehensmodell für ILM entwickelt, welches einem realen IT-Projekt zugrunde liegt. Damit wird zum einen der „best practice“-Ansatz verfolgt, zum anderen wird ein Vorgehen beschrieben, wie man zu einer dem Framework gerechten ILM-Lösung kommt.

3 Vorgehensmodell für die Einführung von ILM

ILM umfasst Aufgaben und Prozesse. Aus diesem Grund ist für die Einführung von ILM von Bedeutung, Unternehmen Handlungsempfehlungen in Form von Vorgehensmodellen zur Verfügung zu stellen, aus denen eine unternehmensspezifische Einführungs- und Umsetzungsstrategie erarbeitet oder abgeleitet werden kann [8, 75].

Um ILM konkret als ein Projekt angehen zu können, wird in diesem Kapitel ein Vorgehensmodell entwickelt. Auf Basis dieses Vorgehensmodells wurde eine Fallstudie initiiert und anschließend die ILM-Konzeption, das heißt die Entscheidungsvorlage für das IT-Management, aufgezeigt.

3.1 Verwandte Arbeiten

In verschiedenen Veröffentlichungen werden Vorgehensmodelle zur Einführung von ILM vorgestellt und beschrieben. So hat beispielsweise die SNIA einen Plan entwickelt, der mittels fünf Phasen ein strategisches Einführungskonzept für ILM in Unternehmen beschreibt [75]. Die Phasen beschreiben das Vorgehen von der Konsolidierung der Speicherung bis zu einem unternehmensweiten ILM. BITKOM hat ein so genanntes „ILM-Modell“ entwickelt, welches durch einzelne Prozessschritte beschreibt, wie ILM eingeführt und durch operative Maßnahmen umgesetzt werden kann [8]. Beide Vorgehensmodelle haben noch nicht den Stand erreicht, sich in der Praxis durchzusetzen [104].

Matthesius und Stelzer haben ein vierphasiges Vorgehensmodell entwickelt [54]. Darin werden in der ersten Phase *Analyse der Systemlandschaft* zunächst die relevanten Systeme, wie z.B. Content Management-Systeme oder Data Warehouse-Systeme, in dem betreffenden Unternehmen erfasst. Anschließend werden die Datenflüsse zwischen diesen Systemen analysiert. Bei der *Untersuchung der relevanten Systeme* ist von besonderem Interesse, wie groß die Systeme sind, welches Systemwachstum zu erwarten ist und wer die Anwender der Systeme sind. Ziel dieser Phase ist, die Bedeutsamkeit und Tragweite der Systeme im Unternehmen zu ermitteln. In der Phase der *Datenbewertung* werden zuerst Bewertungskriterien festgelegt, wie z.B. die Zugriffshäufigkeit auf bestimmte Daten oder das Wissen der Anwender. Anschließend wird den Daten der jeweilige Wert zugewiesen. Darauf aufbauend erfolgt die Einteilung der Daten in Klassen. In der Phase der *Datenverlagerung* werden die Daten auf die dafür vorgesehenen Speichermedien verlagert oder gelöscht. Grundlage hierfür ist die Klassifikation der Daten, die in der Phase Datenbewertung erfolgte. Dieses Vorgehensmodell mit seinen dedizierten Betrachtungen der Datenströme ist für ILM im Data-Warehouse-Bereich geeignet. So hat z.B. Oßmann dieses Modell für ILM im Rahmen des SAP® Business Information Warehouse angewandt [68].

Kaiser et al. haben einen Prozess beschrieben, der die Implementierung von ILM in einer Organisation erleichtern soll [44]. Diese Arbeit ist rein theoretisch gehalten und verfolgt das grundsätzliche Ziel, ILM mit Hilfe grafischer Artefakte logisch zu strukturieren, einen Vorschlag für den Umfang und die Inhalte dieses Konzepts zu liefern und Prozesse zu konzipieren, die dieses Verständnis von ILM unterstützen. Der vorgestellte Prozess umfasst die orga-

nisatorischen Vorbedingungen einer ILM-Implementierung mit der Bestimmung der Anforderung in den relevanten Gebieten Kosten, Usability und Compliance. Der Prozess fährt fort mit der Definition von Service Level Objectives und endet mit der Festlegung der System-Policies. Zusammen mit einer Methodik zur Definition und Zuweisung von Dokumentenklassen soll auf diese Weise ein Lebenszyklus umfassend und flexibel abgebildet werden können [44].

Thome und Sollbach entwickelten einen Projektansatz für ILM, der ihre eigenen Erfahrungen widerspiegelt [104]. Sie beschreiben den „Weg zur Zielerreichung“ von ILM in vier Stufen. Die erste ist die *Klassifizierung* von Daten und Applikationen, die basierend auf „Business Rules“ erfolgt. Es folgt die *Implementierung* auf Basis von Policies mit Informationsmanagement-Tools. Das *Management* der gesamten Speicher- und IT-Umgebung bildet die nächste Stufe, die in die Stufe *Tiered* mündet. In dieser Stufe werden die Speicherressourcen in Klassen migriert. Diese vier Stufen sind in Abhängigkeit vom Gesamtziel mehr oder weniger häufig zu durchschreiten [104]. Außerdem definieren Thome und Sollbach zusätzlich fünf „Schritte zum Ziel ILM“. Diese sind die *Ermittlung der Requirements*, die *Definition der Service Levels,* die *Etablierung von Policies und Verfahren,* die *Zuordnung von Applikationen und Service Level Agreements* und das *Messen der Resultate* [104].

Der Speicherhersteller EMC versteht unter ILM ein Konzept, das die unternehmensweite, intelligente Speicherung und Verwaltung von Daten zu den jeweils geringsten Kosten umsetzt [14]. Die Realisierung erfolgt in drei Phasen. Phase 1 sind *Produkte für ILM*. Diese Phase steht für ein automatisiertes Speichernetz, welches Grundvoraussetzung für den Aufbau einer ILM-Umgebung ist. Phase 2 besteht aus *Beratung für ILM*. In dieser Phase erfolgt eine detaillierte Bestandsaufnahme der Daten- und Anwendungsarten in der Organisation. Anschließend wird definiert, wie diese Daten - von der Entstehung über deren Nutzung bis zur Vernichtung - gespeichert und verwaltet werden sollen, damit die geforderte Verfügbarkeit jederzeit gesichert bleibt. Den Abschluss bildet die Phase *Services für ILM*. Dabei bilden die in Phase 2 definierten Regeln das Gerüst für die technologische Implementierung einer integrierten, automatisierten ILM-Umgebung [14].

3.2 Vorgehensmodell

Im Jahr 2004 wurde im Rahmen des E-Financelab e.V. ein IT-Projekt der Siemens AG auf die Einsatzmöglichkeiten von ILM untersucht. Die oben beschriebenen Herausforderungen des Wachstums und der steigenden Kosten waren gegeben und erzeugten Handlungsbedarf.

Im ersten Schritt wurde folgendes Vorgehensmodell entwickelt [106]. Das Vorgehensmodell besteht aus verschiedenen Phasen. Die Phasen lauten wie folgt [106, 107]:

- Erfassung: Die Speicherlandschaft des Unternehmens wird mittels Verwaltungssystemen untersucht, um die Daten auf den Speichermedien zu erfassen (Ist-Analyse). Ziel dieser Phase ist, das grundsätzliche Potenzial von ILM zu identifizieren.

- Sozialisierung: Die Ergebnisse der Ist-Analyse werden der Unternehmensleitung präsentiert und gemeinsam die Verwahrungsorte der Dateien mit den Geschäftsprozessen abgeglichen (Soll-Zustand).
- Klassifizierung: Den Informationen werden Werte zugewiesen. Dazu gibt es verschiedene Möglichkeiten. Das konkrete Vorgehen basiert auf dem Ergebnis der Phase Sozialisierung.
- Automatisierung: Eine Optimierungsfunktion verteilt die Informationen entsprechend ihrer Klassifikation auf ein adäquates Speichermedium unter Berücksichtigung der Dienstgütevereinbarungen.
- Überprüfung: Nach Einführung von Automatisierungsprozessen wird in regelmäßigen Abständen überprüft, ob die Abbildung des Geschäftsmodells noch gültig ist, da dieses sich über den Zeitraum verändern kann, was dann in ILM eingearbeitet werden muss.

Diese Phasen stehen untereinander in Abhängigkeit, wie folgende Abbildung zeigt:

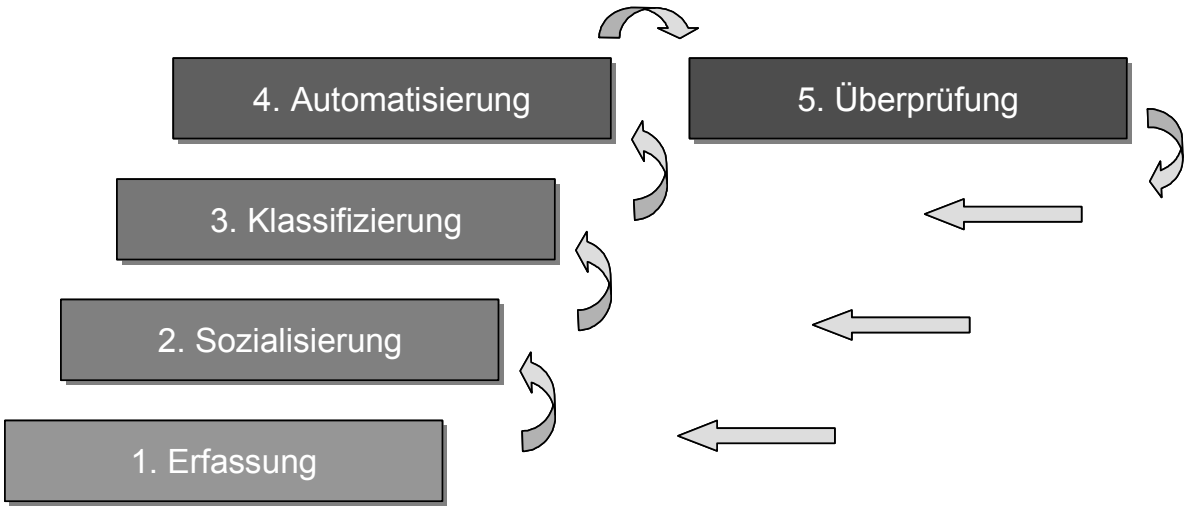


Abbildung 4: ILM Vorgehensmodell [108, 109, 110]:

In der ersten Phase gilt es zu erfassen, welche Daten sich wo auf der Speicherlandschaft befinden. Speicher-Administratoren können dazu Storage Resource Management (SRM)-Tools verwenden, die diesen Prozess unterstützen. SRM Lösungen helfen dem Administrator zu begreifen, welche Daten sich auf den Speicher-Komponenten befinden. Die meisten SRM-Tools können Berichte über Nutzungsmuster generieren.

Wenn die IT-Abteilung ermittelt hat, welche Daten sie hat und wo diese Daten leben, können sie die nächsten Schritte des ILM Prozesses beginnen.

Mit den generierten Berichten aus Phase 1 verschafft sich der Speicher-Administrator einen umfangreichen Gesamtüberblick. Dieser Gesamtüberblick ist an die Geschäftsleitung zu kommunizieren. Ziel dieser Phase ist es, ein gemeinsames Verständnis über die Ist-Situation

zu erreichen. Sobald die IT-Abteilung und die Geschäftsleitung ein gemeinsames Verständnis über die Speicher-Situation erlangt haben (Sozialisierung), müssen Abteilungsleiter bestimmen, wie relevant die Daten für das Geschäft zu jedem gegebenen Zeitpunkt sind.

Aufgrund der Unternehmensrelevanz der Daten ist man in der Lage, Daten basierend auf Geschäftserfordernissen zu klassifizieren (z.B. auftragskritisch, geschäftskritisch oder abteilungsrelevant). Diese Klassifizierung ermöglicht es der IT-Leitung zu bestimmen, wo die Daten während ihres Lifecycles aufbewahrt werden sollen. Die Klassifizierung dient dazu, Regeln (Policies) zu schaffen, um Daten auf die passenden Speicher- "Klassen" im Laufe der Zeit zu migrieren. Als Ergebnis dieser Phase liegt ein Klassifizierungsschema für die Daten des betreffenden Unternehmens vor.

In der Phase *Automatisierung* werden alle bisher gesammelten Daten verwendet, um die Policies einzuführen, die die Migration der Daten automatisieren sollen. Die Automatisierung führt zur Minimierung der manuellen Datenverwaltung. Sobald die Daten klassifiziert worden sind und die IT-Leitung mit der Geschäftsleitung sich auf einen Plan geeinigt hat, wo die Daten während ihres Lebenszyklus aufbewahrt werden sollen, müssen die Policies in Form von Migrationsregeln erstellt werden.

Automated Data Migration (ADM) -Tools können die Migration von Daten von einer Speicher-Klasse zur anderen basierend auf Migrationsregeln automatisieren. ADM Lösungen sind im Grunde genommen eine Kombination von intelligentem SRM und Hierarchischem Storage Management (HSM). HSM wurde dafür entworfen, Datenmigration für Archivierungszwecke zu automatisieren. Es ist rein zugriffsgesteuert. Im Gegensatz dazu ermöglichen ADM-Lösungen Datenmigration über verschiedene Speicher-Ressourcen basierend auf einer Kombination von anwenderdefinierten Kriterien. Es ist somit wertgesteuert. ADM Tools helfen der IT die Wertzuweisung durchzuführen. Administratoren erstellen Kriterien, um die Dateien zu bewerten, basierend auf Art, Eigentum, letztem Zugang oder Alter der Dateien.

Dateien von einer bestimmten Applikation können prinzipiell auf dem primären Speicher abgelegt, und können später gegebenenfalls auf den sekundären Speicher migriert werden, wenn auf sie nicht in einem gewissen Zeitraum zugegriffen worden ist. Eine andere Option kann sein, Dateien von einer bestimmten Applikation zu spiegeln oder auf den sekundären Speicher zu kopieren. ADM Lösungen haben auch die Fähigkeit, Dateien in einer höheren Speicherebene wiederherzustellen, wenn diese Daten migriert worden sind und danach wiederholt auf sie zugegriffen wurde.

Die letzte Phase ist die *Überprüfung*. In regelmäßigen Abständen oder bei Einsatz neuer Unternehmensapplikationen gilt es, die Nutzungsmuster zu überprüfen. Die konstante und korrekte Umsetzung der Policies hält den ILM Prozess am Leben.

Gemäß diesem Vorgehensmodell wird in dieser Arbeit sukzessiv vorgegangen. Anhand einer Fallstudie der Siemens AG wird Phase 1 des vorgestellten Modells, die Erfassung, eingeleitet.

3.3 Fallstudie zur Erfassung

Im Rahmen einer Fallstudie, genannt Fallstudie 1, wurde ein Datenbestand der Siemens AG analysiert, um zu prüfen, ob genügend Potenzial vorhanden ist, so dass ILM nutzbringend angewendet werden kann [115]. Nachfolgend wird die Fallstudie 1 kurz vorgestellt und die Ergebnisse werden präsentiert. Nähere Details zur Fallstudie 1 finden sich in Anhang A.

3.3.1 Ausgangssituation

Im Rahmen der Fallstudie 1 wurde eine Analyse über ein Dokumentenverwaltungssystem der Siemens AG vorgenommen. Das Dokumentenverwaltungssystem wird von einem deutschlandweit tätigen Consulting-Bereich benutzt. Dieser Bereich unterteilt sich in neun Regionen, die jeweils ein eigenes Consulting Team haben, welches in der Region Rhein-Main etwa 90 Mitarbeiter zählt. Deutschlandweit sind circa 700 Personen in dem Bereich tätig. Das Unternehmen unterstützt die überregionale Zusammenarbeit der einzelnen Consulting-Teams. Hierfür wurde ein einheitliches Dokumentenverwaltungssystem, nachfolgend „Datenbank“ genannt, eingeführt, in der sämtliche Berater ihre Projekte ablegen müssen. Damit wird ein überregionaler Zugang zu allen Dokumenten gewährleistet.

Die gesamte Datenbank umfasst über 150.000 Dokumente. Die Dokumente können in der Datenbank editiert, angesehen, versioniert und gelöscht werden. Beim Versionieren werden mehrere physikalische Dateien für ein Dokument im System vorgehalten. Ein Löschvorgang macht diese nicht mehr über die üblichen Clients zugänglich, obwohl sie drei Tage in der Datenbank vorgehalten werden und nur mit einem Administratorzugang wiederhergestellt werden können.

Eine Untergruppe des Consulting-Bereiches, das Projektmanagement (PM), hat klare Richtlinien für die Dokumentation und Ablage, welche für eine einheitliche Archivierung sorgen sollen. Aus diesem Grund wurden Stichproben vom PM-Bereich entnommen mit dem Ziel zu identifizieren, wie häufig Dateien aufgerufen werden. Abgeleitet daraus sollen nicht nachgefragte Dateien als Potential für ILM identifiziert werden.

3.3.2 Datenbasis

Am 01.11.2004 wurden in der Datenbank 134 Projekte gezählt. Die Datenbank selbst unterteilt sich in drei Bereiche, benannt Mittelwest (MW), Nordost (NO) sowie Süd-Südwest (SS), in denen die neun Regionen ihre Projektablage vornehmen. Um keinen Bereich zu bevorteilen und weil keine Mitarbeiterverteilung über Regionen bekannt war, steuerte jeder Bereich den jeweils gleichen Anteil an der Stichprobe bei.

Eine alphabetische Liste der Projekte wurde durchnummeriert. Die vergebenen Ordnungsnummern identifizierten die Projekte. Mittels der Random-Funktion von Java wurde eine Pseudozufallszahl im Bereich [0, n-1] (n = Anzahl Projekte) bestimmt. Nach der Auswahl wurde dieses aus der Liste gestrichen und selbige neu durchnummeriert. Dieser Schritt wurde für jeden Bereich dreimal ausgeführt mit dem Ergebnis einer neunelementigen Stichprobe.

Es wurden 1762 Einzeldokumente, die insgesamt 942 MB Speicher benötigen, protokolliert. Aus Gründen der Anonymisierung wurden die untersuchten Projekte mit Großbuchstaben beschriftet, mit denen sie nachfolgend referenziert werden (siehe Tabelle 2).

Projekt	A	B	C	D	E	F	G	H	I
Region	MW	MW	MW	NO	NO	NO	SS	SS	SS
Dateien	196	170	121	430	592	6	116	66	65
Menge (kB)	54687	267840	26725	338295	191848	640	36483	12952	12969

Tabelle 2: Untersuchte Projekte

Der protokollierte Zeitraum eines Dokumentes erstreckt sich von der Erstellung bis zum 31. Januar 2005. Über die Funktion „Notification“, einem Logbuch der Zustandsveränderungen eines Dokumentes in der Datenbank, sowie zur Konsistenzprüfung über die Funktion „History“, eine Zugriffsaufzeichnung, wurde der Zugriffsverlauf jedes Einzeldokuments bestimmt. Aus diesen Rohdaten wurde eine relative Zugriffshäufigkeit für jedes Intervall berechnet.

Die Einteilung der zu betrachtenden Zeitintervalle erfolgt in Anlehnung an die Moore’sche Studie und lautet [63]:

(0, 1], (1, 3), [3, 7), [7, 15), [15, 30), [30, 60), [60, 90), [90, ∞[3]) Tage.

Im Anhang A wird der Zeitraum jenseits der 90 Tage weiter unterteilt und untersucht, weil sich zeigen wird, dass eine Aussage jenseits von 90 Tagen einer weiteren Differenzierung bedarf.

3.3.3 Ergebnisse der Fallstudie 1

Die Erfassung der Daten brachte folgendes Ergebnis. Es finden signifikant viele Zugriffe nach 90 Tagen nach Dateierstellung statt (siehe Abbildung 5)

[3] Exakt endet das letzte Intervall mit dem Tag der Datenerhebung

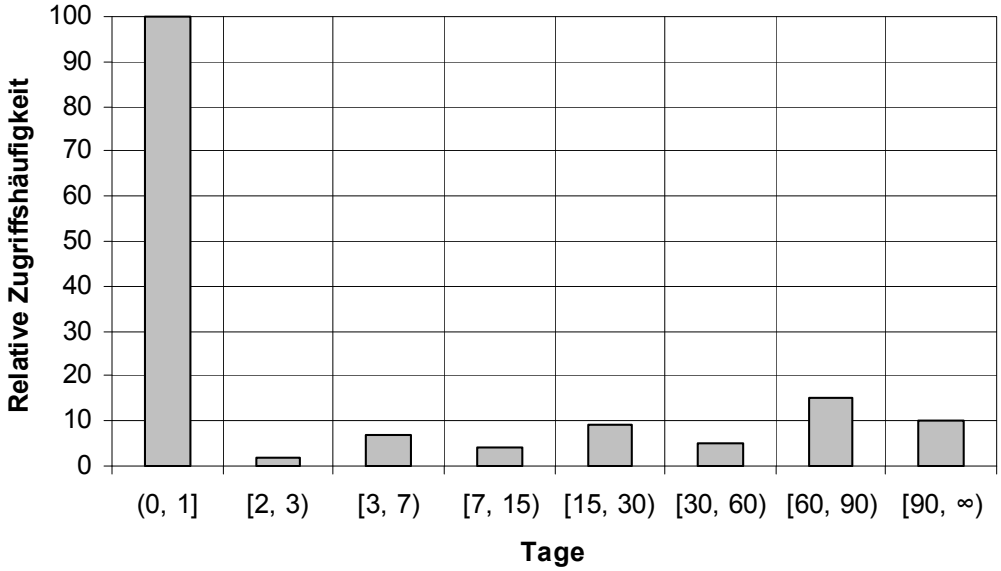


Abbildung 5: Relative Zugriffshäufigkeiten der Dokumente

Die Entwicklung der Zugriffe weist ein deutliches Gefälle auf. Die Zugriffshäufigkeiten im Intervall drei bis sieben Tage liegt bei 6,59 Prozent. Die Zugriffshäufigkeit im Intervall [30, 60) liegt unter 5 Prozent, nämlich bei 4,61 Prozent. Werden die beiden nächsten Intervalle untersucht, ist ein Zuwachs ersichtlich, nämlich 15,37 Prozent im Intervall [60, 90) und 9,61 Prozent im Intervall [90, ∞).
Somit kann die Aussage Moores, dass nach 90 Tagen die Zugriffshäufigkeiten nahe Null liegen, hier nicht bestätigt werden. Dies bedeutet, dass Moore allein nicht als Grundlage für ILM bei der Datenbank herangezogen werden sollte.

Ausgehend von den Dokumenten, die nach 90 Tagen keinen Zugriff zu verzeichnen haben, ist der von ihnen belegte Anteil am Gesamtspeicherbedarf der Stichprobe 88%. Dieses Ergebnis bestätigt die Untersuchungen von Gibson et al. aus dem Jahre 1998 [30].

Um die Charakteristika der Zugriffe noch genauer zu ermitteln, wurde das Intervall [90,∞) weiter unterteilt und eine Auswertung aller Projekte auf 400-Tage-Basis angestellt. Dabei zeigte sich, dass ca. 9 Prozent aller Zugriffe zwischen dem 250. und 300. Tag nach Erstellung erfolgte (siehe Abbildung 6). Eine Erklärung dafür ist nicht eindeutig zu geben. Es kann an unternehmenseigenen Prozeduren der Revision oder des Controllings liegen.

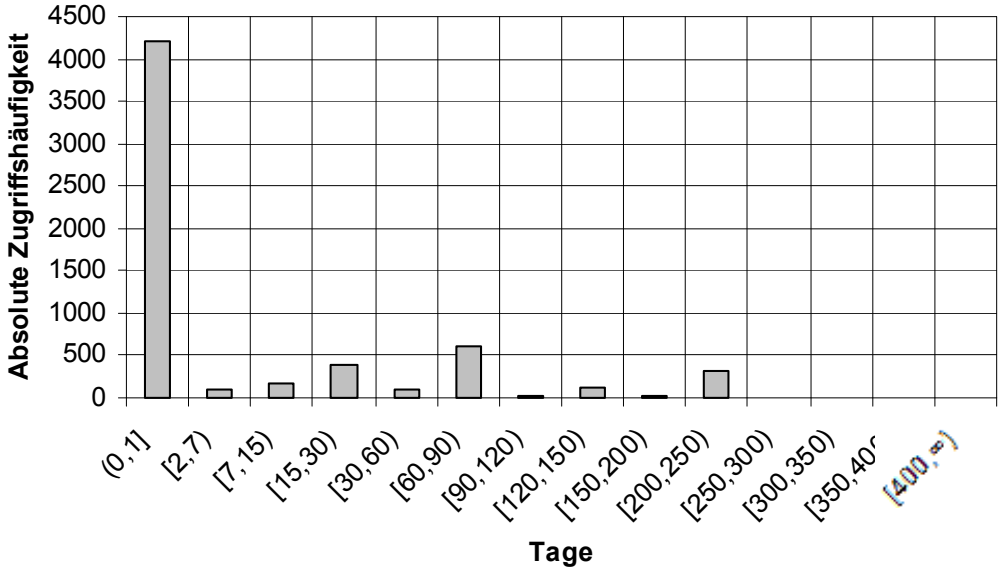


Abbildung 6: Histogramm der absoluten Zugriffshäufigkeiten

Die Betrachtung auf 400-Tage-Basis zeigt, dass, obwohl 88% der Dateien keine Zugriffe 90 Tage nach Erstellung erfahren, man dennoch die Zugriffe über einen längeren Zeitraum als 90 Tage betrachten muss. Damit kann eine kanonische Migrationsregeln wie z.B. *Migriere alle Dokumente 90 Tage nach Erstellung* nur bedingt leistungsfähig sein, was sich im späteren Verlauf dieser Arbeit bestätigen wird.

Mit diesen Ergebnissen aus der Phase Erfassung lässt sich in die nächste Phase, die Sozialisierung, starten. Im Laufe dieser Phase erstellt die IT-Abteilung zusammen mit der Geschäftsleitung eine gemeinsame Konzeption über eine ILM-Lösung. Wie ein derartiges Konzept, das sog. ILM-Konzept, welches Ist-Situation und Soll-Situation für den IT-Entscheider konsolidiert, strukturiert ist, zeigt der nachfolgende Abschnitt.

3.4 ILM-Konzeption zur Sozialisierung

Information Lifecycle Management (ILM) hat das Potenzial, Kostenpositionen (Total Cost of Ownership, Produktivität, usw.) und den Umgang mit Service Level Agreements wesentlich zu verbessern. Bei der Entwicklung eines ILM-Konzeptes werden neben den Kostenuntersuchungen prinzipiell zwei Aspekte beleuchtet, ein organisatorischer Teil (OrgTeil) und ein technischer Teil (TechTeil) [112] (siehe Abbildung 7).

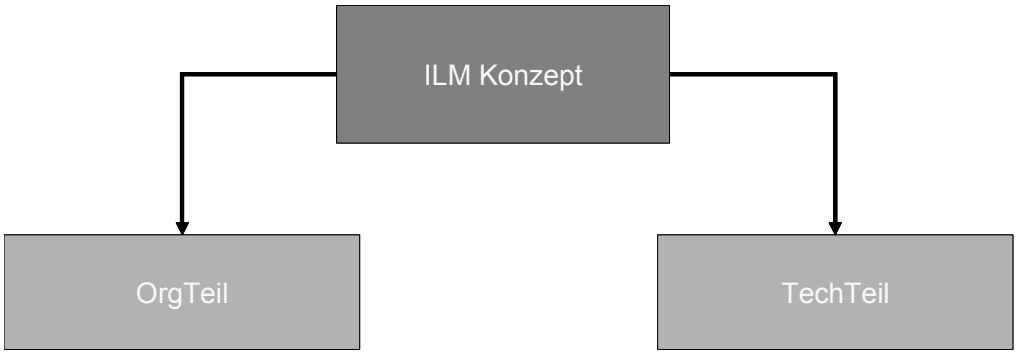


Abbildung 7: Die Hauptaspekte eines ILM-Konzeptes [112]

Der organisatorische Teil (OrgTeil) dient der Identifizierung der relevanten Gesetze und Regelungen sowie der Geschäftsprozesse und der benötigten Informationen. Ziel ist eine Zuordnung der Informationen in Kategorien je nach Relevanz der Informationen für die Geschäftsprozesse des Unternehmens.

Der technische Teil (TechTeil) stellt die Abbildung des OrgTeils auf eine Infrastruktur zwecks optimaler Speicherung der Informationen in den ihnen zugewiesenen Kategorien dar.

Da kein Unternehmen in neue Lösungen und Produkte investieren wird, wenn die wirtschaftlichen Ziele dadurch nicht erreicht werden, sind TCO-Kalkulationen der Gradmesser für Wohl und Weh der konzipierten Lösung.

3.4.1 Der organisatorische Teil (OrgTeil)

Der OrgTeil befasst sich mit allem, was die in einem Unternehmen benötigten „Informationen“ betrifft. Der Begriff der Information ist sehr abstrakt und muss demnach im Rahmen des Konzepts erst noch ILM-gerecht definiert werden. In der Regel sind Informationen gespeicherte Daten. Diese Daten werden zur Abwicklung der Geschäftsprozesse gebraucht. In der vorliegenden Arbeit wird als Information eine Datei angesehen.

Im OrgTeil gilt es, alle für die Existenz des Unternehmens relevanten Informationen zu identifizieren. Dazu betrachtet man primär die Geschäftsprozesse des Unternehmens. Diese zerfallen in einzelne Prozessschritte in verschiedenen Abteilungen des Unternehmens und beinhalten Schnittstellen zu externen Organisationen von z.B. Kunden oder Lieferanten. Die Prozesse werden außerdem von der geltenden Gesetzgebung beeinflusst, auf deren Einhaltung nicht verzichtet werden kann. Die Prozesse bedienen sich Applikationen, die wiederum Daten und somit Informationen erzeugen und verarbeiten. Folglich sind Informationen die Basis der Existenz des Unternehmens.

Der OrgTeil identifiziert diejenigen Informationen, die für das Unternehmen wichtig sind und entsprechend ihres Wertes gesichert werden müssen. Aspekte des OrgTeils sind insbesondere [112]:

- Organisation
- Regelungen
- Vorschriften
- Geschäftsprozesse
- Applikationen
- Informationen

Unter dem Aspekt „Regelungen“ versteht man interne Regelungen und Vorgaben. Unter „Vorschriften“ versteht man Gesetze und Bestimmungen, die extern beeinflusst werden (siehe Abschnitt 2.3). Die Erfüllung aller dieser Regeln bezeichnet man auch als „Compliance“. Dass Compliance nicht trivial ist, ergibt sich daraus, dass man z.B. bei einer international aufgestellten Bank von mehr als 700 zu erfüllenden Regeln ausgeht [112]. Einige dieser Regeln finden sich im nächsten Abschnitt. Das Risiko der Compliance besteht darin, dass das betroffene Unternehmen gegen die Regeln verstößt, weil es sie intern technisch oder organisatorisch nicht umsetzen kann. Hinzu kommt, dass IT-Leiter sich überfordert sehen mit dem Thema Compliance. So besagt eine Studie von enginio vom Februar 2005 unter 100 britischen IT-Leitern, dass 88% einen direkten Einfluss der Gesetzgebung auf ihre IT-Budgets sehen. Gleichzeitig fühlen sich aber 71% von ihnen nicht vollständig informiert über die Einflüsse der Gesetzgebung auf ihr Geschäft [58]. Dieser Umstand liefert einigen Speicher-Herstellern die Gelegenheit, die Absätze zu steigern mit dem Problem, dass die adressierten Entscheider auf Kundenseite nicht wissen, was sie tatsächlich benötigen. Hierin offenbart sich Bedarf an Beratung, der vor der Beschaffung von Komponenten in einem Konzept gedeckt werden sollte [111].

Compliance zielt auf die Einhaltung der rechtsverbindlichen Mindestanforderungen in Bezug auf die Sicherheit und Verfügbarkeit von Informationen. Diese Mindestanforderungen werden unter anderem festgelegt in Regelwerken wie den SEC-Regeln mit dem Sarbanes-Oxley-Act (SOX) oder FSA-Regeln (Financial Services Authority) in den USA, der Baseler Eigenkapitalübereinkunft (Basel II) oder dem deutschen KonTraG (Gesetz zur Kontrolle und Transparenz im Geschäftsverkehr). Durch die Vielzahl an Anforderungen bedarf es zur Sicherstellung der Compliance eines automatisierten Prozesses. Ein solcher automatisierter Prozess liegt jeder ILM-Lösung zugrunde. Die Sicherstellung von Compliance in Unternehmen kann ebenfalls durch organisatorische Maßnahmen gestützt werden. Dazu werden bei Banken und Finanzdienstleistern Compliance-Abteilungen eingerichtet. Weiterführende Informationen zur Compliance und ebenso zu den Auswirkungen unzureichender Compliance finden sich in einem von mir verfassten Fachartikel [112].

3.4.2 Der technische Teil (TechTeil)

Der TechTeil im ILM-Konzept liegt zeitlich nach dem OrgTeil. Aus dem OrgTeil entstehen die Anforderungen an den TechTeil. Die Aufgabe des TechTeils ist die Abbildung dieser An-

forderungen an die IT-Infrastruktur. Als Resultat des TechTeils sind insbesondere folgende Punkte untersucht und geklärt worden [112]:

- Architektur
- Systeme
- Hardware
- Software
- Protokolle
- Performance

Aus dem TechTeil resultieren konkrete Antworten auf die technischen Fragen. In der Speicherwelt gibt es ähnlich wie beim OrgTeil mannigfaltige Aspekte. Die aus dem OrgTeil resultierenden Anforderungen an den TechTeil sind ausnahmslos zu erfüllen. Wechselwirkungen beider Teile sind selbstverständlich nicht ausgeschlossen, die klare Priorität liegt jedoch auf dem OrgTeil. Grund dafür ist die längere Relevanz der Aspekte Regelungen (z.B. Gesetze), Prozesse (z.B. Produkthaltung, Kundenbeziehung) und Organisation (z.B. Unternehmensstruktur).

Dem gegenüber sind die Lebenserwartungen elektronischer Datenträger kurzlebiger. Zwar wird die Lebensdauer von z.B. CD-RW und DVD-RW mit bis zu 70 Jahren angegeben, allerdings variieren derartige Angaben und es kann sich nur um theoretische Werte handeln. Weil kein Speichermedium-Hersteller eine Lebensdauer garantiert, muss ein Unternehmen damit rechnen, die Medien zeitgerecht ersetzen zu müssen. Andernfalls droht eine Verletzung der Compliance, die strikt zu vermeiden ist.
Aus diesem Grund ist die Technik-Betrachtung bewusst in einem eigenen Teil ausgeklammert. Ein Vorteil der Trennung besteht darin, dass zu erwartende Innovationen bei der Technik bei gleich bleibendem OrgTeil eingearbeitet werden können. Ein Beispiel dafür ist die bereits genannte WORM-Technologie. In Verbindung mit einem Speichermedium sorgt diese Technologie dafür, dass Daten einmal an einer Stelle ablegt und die Informationen dann an diesem Platz eingefroren werden. Damit sind die Daten vielfach wieder lesbar, aber nicht zu verändern oder zu löschen.
Mit der Nutzung von Tapes als Speichermedium, die im Vergleich zu Festplatten immer noch preiswerter sind, steht kleinen und mittelständischen Betrieben verhältnismäßig günstiger Speicherplatz zur Verfügung. Da dieser sich auch zur Archivierung von Daten eignet, zeigt sich, dass Compliance nicht zwangsläufig teuer sein muss.

3.4.3 Die ILM-Lösung

Basis jeder ILM-Lösung ist ein ILM-Konzept. Das Konzept, bestehend aus OrgTeil und TechTeil, gibt dem Unternehmen Klarheit über folgende Aspekte der Datenhaltung:

1. Was sind die relevanten Informationen für das Geschäft? (Information)

Beispielsweise SAP®-Daten, CRM-Daten, Voice/Mail-Messaging oder Dateien aus Office-Applikationen

2. Wie ist die zeitliche Entwicklung des Wertes der relevanten Informationen? (Lifecycle)

Zum Beispiel: Wann ist eine E-Mail unwichtig geworden und sollte gelöscht werden?

3. Wie bildet man die relevanten Informationen auf die technische Umgebung gemäß ihrem Wert ab? (Management)

Beispielsweise sind Telefongespräche im Aktienhandel per Gesetz so wichtig, dass diese auf ein eigenes System gespeichert werden müssen – hingegen in anderen Bereichen ist die Aufzeichnung von Gesprächen in der Regel rechtswidrig und somit irrelevant.

Die wichtigste Frage, die im Konzept abschließend zu klären ist, ist die Frage der wirtschaftlichen Zielerreichung. Kosten (Preis und TCO) sind das Hauptkriterium bei Investitionen im Speicherbereich. Im Ranking des SNIA End User Councils (EUC) über Einflussfaktoren bei Investitionen liegen Kosten noch vor dem Reagieren auf steigenden Kapazitätsbedarf [74]. Es sind also TCO-Analysen der ILM-Lösung zu erstellen. Dazu ist in Anbetracht der wirtschaftlichen Rahmenbedingungen zu prüfen, ob der gewünschte Zweck der Konzeption erreicht worden ist. Ohne Wirtschaftlichkeitsbetrachtungen wird IT-seitig nichts realisiert.

Kriterien der TCO-Analyse sind unter anderen [112, 113]:

- Hardware und Software Beschaffung
- Hardware Wartung
- Software Wartung
- Raumbedarf
- Stromverbrauch
- Administration

Verläuft der TCO Check erfolgreich, ist das ILM-Konzept fertig gestellt.

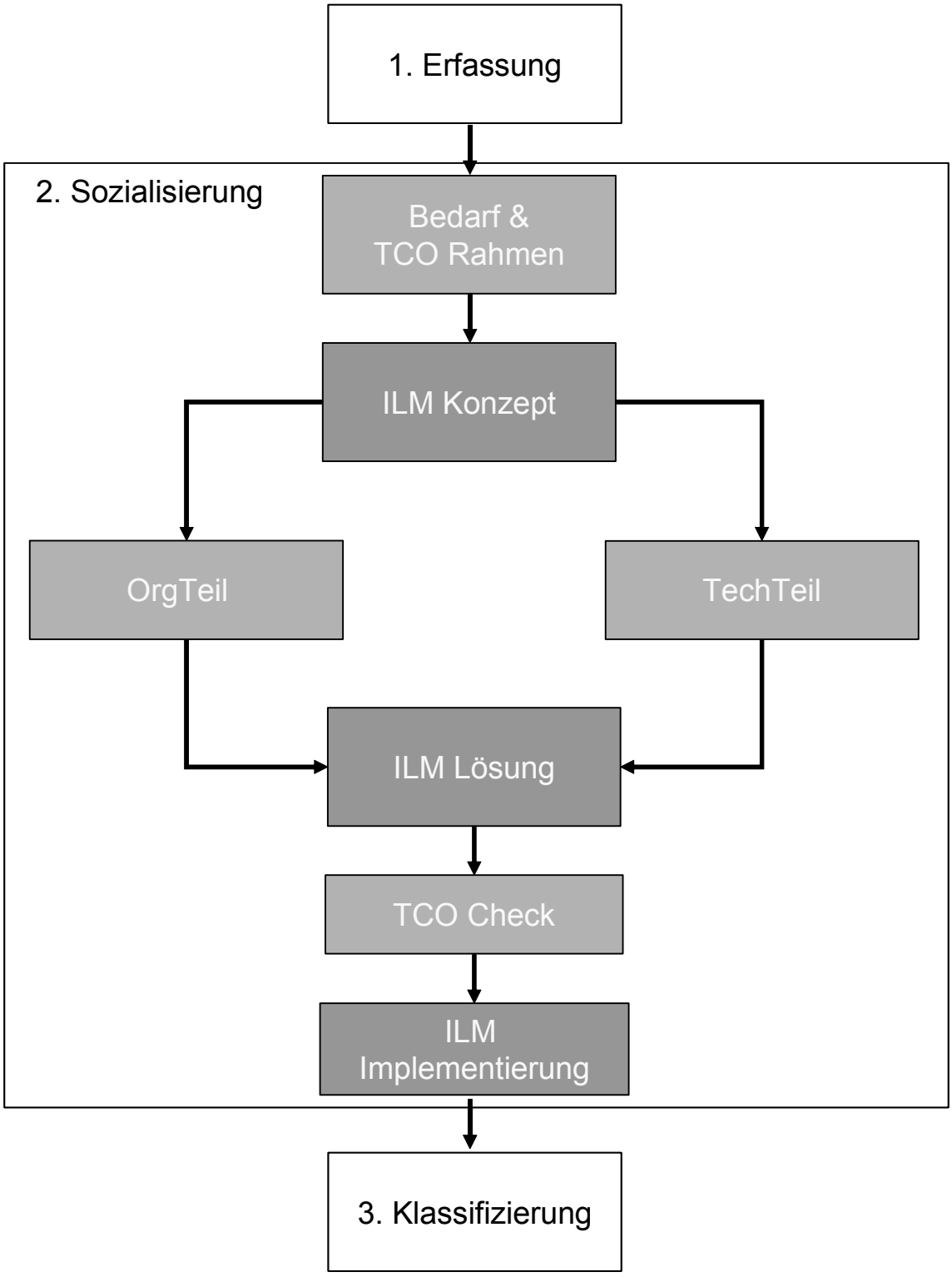


Abbildung 8: Ablauf der ILM-Konzeption

Mit der positiven Entscheidung für eine im ILM-Konzept erarbeitete ILM-Lösung geht das Vorgehensmodell in seine nächste Phase.

3.5 Zusammenfassung

In diesem Kapitel wurde ein allgemeines Vorgehensmodell für ILM Projekte entwickelt, welches in einem IT-Projekt der Siemens AG Anwendung fand. Das Vorgehen besteht aus den fünf Phasen „Erfassung“, „Sozialisierung“, „Klassifizierung“, „Automatisierung“ und „Überprüfung“. Die Phasen „Erfassung“ und „Sozialisierung“ wurden exemplarisch durchlaufen. Dazu wurde in der Phase „Erfassung“ eine Fallstudie an einer Datenbank der Siemens AG durchgeführt. Die zwei Hauptergebnisse waren, dass erstens fast 90% der Dateien 90 Tage nach ihrer Entstehung keine Zugriffe mehr erfahren und zweitens die bekannten Moore’schen Zugriffsmuster hier nicht zutreffen.

Für die Phase „Sozialisierung“ wurde die allgemeine Struktur eines ILM-Konzepts vorgestellt. Das Konzept bildet als Dokument die Entscheidungsgrundlage für ILM-Projekte und ist in seiner Struktur auch außerhalb des Siemens-Projektes verwendbar. Hauptcharakteristikum des Konzepts ist die Trennung der organisatorischen Anforderungen von der technischen Realisierung. Diese Trennung ermöglicht und verlangt eine Konzentration auf die ILM-Projektziele, die aus den Geschäftsanforderungen resultieren und nicht technisch formuliert sind.

Nach Abschluss der beiden ersten Phasen wird im folgenden Kapitel dargestellt, wie die „Klassifizierung“ durchgeführt werden kann. Kernpunkt dieser Phase ist die Bewertung von Dateien.

4 Bekannte Methoden der Klassifizierung

ILM speichert Dateien in Abhängigkeit ihres Wertes auf verschiedene Speicherhierarchien. Praktisch erweist sich diese Wertzuweisung als nicht trivial. In diesem Kapitel werden bekannte Ansätze zur Wertzuweisung vorgestellt und anhand von objektiven Kriterien gegenübergestellt. Klassifizierung findet auch in anderen System-Bereichen statt. Wie dieses Thema dort behandelt wird, zeigt der nächsten Abschnitt.

4.1 Kriterien für nicht-wertbasierte Klassifizierung

Die Klassifizierung auf Basis von Nutzungsinformation findet in verschiedenen Bereichen statt. Google® benutzt den PageRank-Algorithmus, um die Wichtigkeit einer Internetseite aufzulisten [69, 79]. Dabei wird eine Seite hauptsächlich auf Basis der Anzahl, wie viele andere Seite mit ihr verlinkt sind, klassifiziert. Hierbei stellen die Links eine Art der Nutzung der Seite durch auf sie verweisende Seiten dar. Caching-Algorithmen benutzen häufig Nutzungsinformationen, um festzulegen, welche Daten wichtig sind. Die wichtigen werden daraufhin in den Cache-Speichern von File-Systemen, Datenbanken oder Speicher-Controlern vorgehalten [16, 17, 22]. Im Bereich von Web-Caching wurden der bekannte Least-Recently-Used-Algorithmus und seine Ableger zur Klassifizierung von Web-Inhalten entwickelt [41, 82]. LRU verwirft die am wenigsten genutzten Einträge zuerst, um die Speichergrenzen des Caches einzuhalten.

Derartige Algorithmen von Suchmaschinen oder Cache-Speichern können für ILM nicht angewendet werden, weil zum einen keine Verweise von Dateien auf Dateien existieren und zum anderen ILM keine Begrenzung der Speicherkapazitäten wie beim Caching berücksichtigen muss.

4.2 Kriterien für wertbasierte Klassifizierung

In der Literatur werden verschiedene Kriterien für die Dateibewertung verwendet. Die Bewertung der Dateien anhand der monetären Erträge, die aus der Nutzung der darin enthaltenen Informationen entstehen, wurde vielfach propagiert [8, 61, 72, 117]. Auch eine Bewertung von Dateien anhand der entstandenen Kosten, die bei der Erstellung anfallen, ist denkbar [61]. Je höher die entstandenen Kosten sind, desto höher ist der Wert, der den Informationen zugewiesen wird. Derartige Bewertungen erweisen sich als sehr aufwändig, da sich aufgrund der Informationskomplexität ein direkter Bezug zu bestimmten finanziellen Größen nur schwer herstellen lässt. Zudem gibt es innerhalb jedes Unternehmens Informationen, welche zwar häufig genutzt werden, aber keine direkten monetären Erträge bewirken bzw. sich monetär kaum oder nur schwer bewerten lassen [13, 90]. Auch das Nutzungsverhalten kann in die Wertzuweisung einbezogen werden, beispielsweise die Anzahl der Lese- und Schreibzugriffe auf die Dateien [13, 23].

Zusätzlich sind die gesetzlichen Anforderungen zur Aufbewahrung der Informationen [15, 90] sowie die Gewohnheiten der Anwender [59, 90] hinsichtlich der Informationsnutzung von Relevanz. Die Erhebung und permanente Aktualisierung dieser Informationen ohne automati-

sierte Unterstützung ist praktisch kaum möglich und wirtschaftlich nicht sinnvoll [54]. Weiterhin ist ihre Eignung und Aussagekraft für die Wertzuweisung im Rahmen von ILM nicht evident [54].

Ein aussagekräftiges Kriterium zur Bewertung von Dateien ist die Verwendung von Zugriffshäufigkeiten. Diese lassen sich zusätzlich noch einfach ermitteln [7, 11, 13, 90]. Die Aussagekraft steckt in der Ermittlung von tatsächlichen Zeitpunkten, an denen auf die Dateien zugegriffen wurde. Eine Migration noch benötigter Dateien auf langsame Speicherhierarchien ist prinzipiell zu vermeiden. 2008 formulierten Matthesius und Stelzer Anforderungen an Bewertungsmechanismen für Dateien in ILM [54]. Diese sind folgende:

- Bewertung von Dateien
- Klassifikation der Dateien
- Kostenreduktion
- Berücksichtigung der Systemperformance
- Verwendung der Zugriffshäufigkeit als Bewertungskriterium
- Verwendung von Anwender- und Administratorenwissen
- Berücksichtigung rechtlicher Anforderungen
- Prognose der Zugriffshäufigkeit

In der Literatur finden sich verschiedene Konzepte zur Gestaltung der Dateibewertung im ILM [7, 11, 13, 59, 90, 102, 120, 125]. Diese sind teilweise grundsätzlich verschieden, so dass sie nachfolgend gemäß den oben genannten Kriterien von Matthesius und Stelzer gegenübergestellt werden.

4.3 Gegenüberstellung bekannter Methoden

In diesem Abschnitt werden verschiedene Konzepte zur Dateibewertung vorgestellt und untersucht. Abschließend wird überprüft, inwieweit die Konzepte bei der Bewertung von Dateien die Kriterien erfüllen.

Chen [13] verwendet zur Bewertung von Dateien die Zugriffshäufigkeit als Bewertungskriterium. Basierend auf der Zugriffshistorie einer Datei wird ihr Wert für einen Zeitpunkt t berechnet. Die Zugriffshäufigkeiten der Vergangenheit werden dabei in verschiedene Zyklen aufgeteilt. Anhand dieser Zyklen wird mit Hilfe der Bewertungsfunktion der Wert einer Datei berechnet, wobei aktuelle Zyklen stärker gewichtet werden als ältere. Je häufiger in der Vergangenheit auf eine Datei zugegriffen wurde und je öfter diese Zugriffe kurz vor dem Bewertungszeitraum t erfolgten, desto höher der Wert. Die Dateien werden anschließend entsprechend ihrer Zugriffshäufigkeiten klassifiziert.

Shah et al. bewerten und klassifizieren Dateien ebenfalls anhand der Zugriffshäufigkeit [90]. Wichtige Anforderungen sind dabei die Berücksichtigung des Anwender- und Administratorenwissens, der Systemperformance und die Reduktion der Kosten für Datenspeicher. Anwender und Administratoren wirken aktiv an der Gestaltung der Bewertungsfunktionen und der anschließenden Verlagerung der Dateien mit.

Bhagwan et al. verfolgen ein Konzept zur Bewertung von Informationen [7] mittels Zugriffshäufigkeit. Durch Verwendung des Kriteriums Zugriffshäufigkeit werden Dateien bewertet und klassifiziert. Im Vordergrund stehen die Performance der Systeme sowie die Kostenreduktion, die durch eine optimale Verlagerung der Dateien erreicht werden kann. Hierbei spielt das Anwender- und Administratorenwissen eine untergeordnete Rolle.

Verma et al. [120] und Mesnier et al. [59] verwenden innerhalb ihrer Konzepte zur automatisierten Bewertung anstelle des Bewertungskriteriums Zugriffshäufigkeit den Dateityp. Dateien gleichen Typs, wie beispielsweise .txt, besitzen nach Meinung der Autoren dieselben Eigenschaften und können einer gemeinsamen Klasse zugeordnet werden. Mesnier et al. unterscheiden zusätzlich, ob auf eine Datei ein schreibender oder ein lesender Zugriff erfolgt. Auf log-Dateien erfolgt in der Regel ein schreibender Zugriff, weshalb diese Information entsprechend bewertet und danach auf Speichermedien verlagert wird, die für den Schreibzugriff optimiert sind. Allerdings vernachlässigen beide Konzepte die Zugriffshäufigkeit auf die Dateien. Somit könnten Dateien auf kostenintensive Speichermedien gespeichert sein, obwohl seit einem längeren Zeitraum keine Zugriffe darauf erfolgten und eine Speicherung auf kostengünstigen Speichermedien sinnvoller wäre.

Eine starke Konzentration auf das Anwender- und Administratorenwissen bezüglich der Systeme und Dateien erfolgt innerhalb der Konzepte von Zadok et al. [125] und Chandra, Gehani und Yu [11]. Jeder Anwender und Administrator erzeugt eigene Bewertungsfunktionen bzw. bewertet seine Dateien selbst. Dabei wird festgelegt, für welchen Zeitraum eine Information einen hohen Wert besitzt und ab wann eine Information auf kostengünstigere Speichermedien verlagert werden kann oder gelöscht werden soll. Dies führt zu Kosteneinsparungen, da Dateien entsprechend ihres Wertes verlagert bzw. behandelt werden. Allerdings setzen beide Konzepte eine angemessene Vorgehensweise der Anwender und Administratoren voraus. Denkbar ist aber, dass aus subjektiver Sicht der Anwender und Administratoren alle Informationen wichtig sind, obwohl nur selten darauf zugegriffen wird [13]. Die Verwendung technischer und objektiver Bewertungskriterien wird bei diesen Konzepten vernachlässigt. Dies führt auch dazu, dass für die Informationsbewertung ein hoher Aufwand entsteht, da mögliche Kosteneinsparungen durch eine automatisierte Unterstützung bei der Bewertung nicht genutzt werden.

Tanaka et al. [102] ähneln in ihrem Bewertungsansatz dem von Moore postuliertem zeitbasierten Ansatz der 90-Tage-Frist. Ebenso wie Moore wird das Entstehungsdatum als Kriterium verwendet. Im Unterschied betrachten Tanaka zusätzlich noch Anwenderwissen, um z.B. Projektdateien eine längere Frist auf der obersten Speicherebene zu gewähren.

In Tabelle 3 werden die Konzepte gemäß der erläuterten Kriterien gegenübergestellt.

Konzept / Anforderungen	Chen	Shah et al.	Baghwan et al.	Verma et al.	Mesnier et al.	Zadok et al.	Chandra et al.	Tanaka et al.
Zugriffscharakteristika								
Verwendung der Zugriffshäufigkeit	X	X	X	-	-	-	-	-
Klassifizierung								
Einteilung in Klassen	X	X	X	X	X	X	X	X
Compliance								
Berücksichtigung rechtlicher Aspekte	-	-	-	-	-	-	-	-
Automatisierung								
Wertzuweisung	X	X	X	X	X	X	X	X
Verwendung Nutzer- und Administratorenwissen	-	X	-	X	X	X	X	X
Prognose von Zugriffshäufigkeiten	-	-	-	-	-	-	-	-
Kosten								
Berücksichtigung der Systemperformance	-	X	X	X	-	-	-	-
Kostenreduktion	X	X	X	X	X	X	X	X

Tabelle 3: Gegenüberstellung der Methoden nach den Kriterien von Matthesius und Stelzer [54]

Alle vorgestellten Konzepte führen eine Dateibewertung durch, die anschließend in eine Klassifikation mündet. Gemäß Klassifikation werden die Dateien auf das jeweils optimale Speichermedium verlagert. Eine Kostenreduktion für die Administration der Datenspeicher und die Aufbewahrung der Dateien ist ebenfalls Ziel aller Konzepte.

Die Mehrzahl der Konzepte lässt das Anwender- und Administratorenwissen in die Bewertung einfließen. Das Bewertungskriterium Zugriffshäufigkeit sowie die Systemperformance wird lediglich von drei der insgesamt acht betrachteten Konzepte berücksichtigt.

Keines der vorgestellten Konzepte erstellt Prognosen der Zugriffshäufigkeit oder berücksichtigt die rechtlichen Anforderungen.

4.4 Bewertung der vorgestellten Methoden

Die Verwendung der Zugriffshistorie bei der Wertzuweisung von Dateien hat einen entscheidenden Vorteil: Diese Daten sind kostengünstig für jede Datei vorhanden. Die Verwendung von Adminstratorenwissen ist sicherlich Ziel führend bei der Einzel-Bewertung, jedoch bei der mehrfachen Bewertung nicht praktikabel. Methoden, die mit von Administratoren festgelegten Zeitfristen arbeiten wie jene von Tanaka, sind sicherlich für ILM praktikabel. Ihre Leistungsfähigkeit bleibt zu testen. In Kapitel 7 wird der zeitbasierte Bewertungsansatz von Moore, dem Tanakas Ansatz gleicht, genauer untersucht.

4.5 Zusammenfassung

Die Wichtigkeit der Wertfindung für ILM ist evident. Es gibt mehrere Methoden zur Dateibewertung in ILM. Keine hat sich bislang durchgesetzt. Aspekte wie Eignung, Leistungsfähigkeit und Praktikabilität der Methoden bleiben unbetrachtet.

Der Blick auf die Zugriffshäufigkeit scheint viel versprechend, wird in den bekannten Methoden aber nicht konsequent weiterverfolgt. Die Auswertung der Zugriffshäufigkeit obliegt den Administratoren, die bei der Betrachtung der Zugriffe wieder ein Zeitschema anwenden müssen und dann doch wieder gemäß Moore agieren. Da stellt sich der Ansatz von Tanaka als ebenbürtig heraus.

Es mangelt an einer Prognose über die Zugriffe. Dies gilt es im nächsten Abschnitt zu beheben. Mit einer prognostizierenden Methode kann man aus der Wahrscheinlichkeit zukünftiger Zugriffe treffendere Schlüsse ziehen. Statt einer Zeit, die auf die Datei nicht mehr zugegriffen wurde, erhält man nun eine Wahrscheinlichkeit weiterer Zugriffe auf diese Datei. Die Auswertung obliegt immer noch dem Administrator, der nun jedoch ein besseres Verständnis über den zugriffsbasierten ermittelten Wert besitzt.

5 Methode der Wahrscheinlichkeit zukünftiger Zugriffe

In den vorangegangenen Kapiteln wurden die Notwendigkeit neuer Konzepte und die daraus entstandene Idee des Information Lifecycle Management erläutert. Weiterhin wurde dargestellt, dass die Wertzuweisung von Dateien zur Klassifikation benötigt wird und automatisiert erfolgen sollte. Zusätzlich ist keine Wertzuweisungsmethode bekannt, die Dateizugriffe prognostiziert.

In diesem Kapitel wird nun anhand real erhobener Daten eine neue Wertzuweisungsmethode entwickelt, die als Resultat für jede Datei eine Wahrscheinlichkeit zukünftiger Zugriffe ausgibt. Die Bewertung der Dateien erfolgt also in Form von Wahrscheinlichkeiten. Je höher die Wahrscheinlichkeit zukünftiger Zugriffe, desto höher der Wert der Datei. Anhand des ermittelten Wertes lassen sich die Dateien klassifizieren. Legt man nun Schwellwerte für die Wahrscheinlichkeiten fest, so erhält man Migrationsregeln, wie sie für ILM benötigt werden.

Es wird die praktische Herangehensweise mit Hilfe statistischer Methoden aufgezeigt. Das Ziel ist, mathematisch das Dateizugriffsverhalten zu beschreiben. Dieses Kapitel stellt zu Beginn die Herausforderungen und das Vorgehen vor. Anschließend werden die erhobenen Daten mit den bekannten Methoden der deskriptiven und angewandten Statistik beschrieben.

5.1 Herausforderungen

Bislang existiert keine Wertzuweisungsmethode von Dateien, die Zugriffe prognostiziert. Wenn man nun eine Methode der Prognose entwickeln möchte, braucht man dazu den Begriff der Wahrscheinlichkeit. Dieser Begriff zieht den Begriff der Wahrscheinlichkeitsverteilung nach sich [32, 35]. Es gibt theoretische Wahrscheinlichkeitsverteilungen und empirische Wahrscheinlichkeitsverteilungen. Die empirische Verteilung hat den Nachteil, dass sie nur für eine bestimmte Stichprobe Gültigkeit hat. Der Vorteil der empirischen Verteilung ist, dass sie stets existiert. Anders verhält sich die theoretische Verteilung. Der Vorteil ist, dass sie auf andere Stichproben der Grundgesamtheit übertragbar ist. Der Nachteil ist, dass der Zusammenhang zwischen Stichprobe und theoretischer Verteilung erst nachgewiesen werden muss [32, 35]. Dies resultiert in folgenden Herausforderungen:

- Definition einer Zufallsvariablen
- Herleitung theoretischer Wahrscheinlichkeitsverteilungen der Zufallsvariablen

5.2 Vorgehen

Angelehnt an die Herausforderungen wird Schritt für Schritt vorgegangen. Als erstes wird eine Stichprobe aus der Datenbank der Siemens AG entnommen. Diese dient dem Untersuchungszweck. Dazu werden die Stichprobenmerkmale beschrieben und anschließend wird eine passende Zufallsvariable definiert. Im nächsten Schritt werden mittels Korrelationsanalysen die Abhängigkeiten zwischen verschiedenen Stichprobenmerkmalen untersucht.

Danach soll ein theoretisches Verteilungsmodell für die definierte Zufallsvariable gefunden werden. Aus dem theoretischen Modell werden Hypothesen über tatsächliche Verteilungen

formuliert und getestet. In einem iterativen Prozess wird die Verteilung weiter präzisiert und getestet, bis signifikante Wahrscheinlichkeitsverteilungen der Zufallsvariablen gefunden werden.

5.3 Verwandte Arbeiten

Bereits zu Beginn der 80er Jahre wurden Untersuchungen zum Zugriffsverhalten auf Dateien unternommen [48, 84, 94]. So hat schon 1981 Satyanarayanan Daten von acht 200 MB-Festplatten eines Systems an der Carnegie-Mellon Universität gesammelt. Es handelte sich um ein PDP-10 (Programmed Data Processor Model 10) System mit dem Betriebssystem TOPS-10 (Total Operating System). Er zeichnete jeweils Dateigröße, Zeitstempel und Dateityp gemäß Datei-Endung auf. Er entwickelte den Begriff der funktionalen Lebenszeit und kam zu dem Ergebnis, dass mit steigender Dateigröße oder steigender funktionaler Lebenszeit, die Anzahl der Zugriffe abnimmt [84].

1984 zeichneten Mullender and Tanenbaum Daten über Dateigrößen in einem UNIX-System der belgischen Vrije Universität auf [64]. Die Ergebnisse waren mit denen Satyanarayanans vergleichbar, obwohl das Betriebssystem ein anderes war.

1991 haben Bennett et al. drei Fileserver an der Universität von Western Ontario untersucht. Sie zeichneten jeweils Dateigröße, Zeitstempel und Dateityp auf [5]. Die Ergebnisse waren unter anderem, dass die mittlere Größe von txt-Dateien um 5% größer war als bei den Aufzeichnungen von Satyanarayanan. Im Gegensatz dazu sank die Größe von exe-Dateien um 38%. Die Anzahl der Dateien pro Nutzer nahm um eine Größenordnung zu.

Smith und Seltzer haben 1994 über zehn Monate täglich Snapshots von vier Fileservern der Harvard Universität aufgezeichnet [95]. Auch sie zeichneten Dateigrößen und Altersinformationen auf, jedoch lag ihr Fokus in der Untersuchung von Datei-Fragmentierung.

Ebenfalls 1994 haben Sienknecht et al. eine umfangreiche Studie unternommen. Daten von 46 HP-UX Systemen der Firma Hewlett-Packard mit 54 GB belegtem Speicher und 2,3 Millionen Dateien wurden untersucht [93]. Primär wurde die Dateigröße untersucht mit dem Resultat, dass die mittlere Dateigröße der einzelnen Dateisysteme zwischen 10 kB und 40 kB lag. Dieses Ergebnis war wiederum eine Steigerung zu den von Bennett et al. generierten Ergebnissen.

Den Trend der steigenden Dateigrößen haben Douceur und Bolosky 1999 weiter bestätigen können [20]. Sie untersuchten 10.568 Dateisysteme von 4.801 Windows PCs. Sie analysierten wiederum die Dateigröße, das Dateialter sowie die Verzeichnisgröße und die Pfadtiefe. Bezüglich der Dateigröße war das Ergebnis eine vierfache mittlere Dateigröße bezogen auf die von Bennet et al. ermittelten Werte für das Unix-System. Weiterhin erkannten Douceur und Bolosky, dass der Dateityp mit der Dateigröße korreliert. In ihrer umfangreichen Arbeit versuchten sie, Verteilungsfunktionen für die beobachteten Größen herzuleiten. Es gelang ihnen, grafische Approximationen zu finden. So fanden sie, dass die Dateigröße einer Log-Normal-Verteilung folgt, das Dateialter einer Hyperexponential-Verteilung folgt, die Verzeichnisgrö-

ßen inversen Polynomial-Verteilungen folgen und die Pfadtiefe der Poisson Verteilung folgt. Sie messen die Qualität ihrer grafischen Approximationen mittels MDCC-Werts (maximum displacement of the cumulative curves), der die größte absolute vertikale Abweichung zwischen den verglichenen Kurven angibt. Ausdrücklich schlugen bei einem Signifikanzniveau von 0,01 bei den approximierten Verteilungen alle Kolmogoroff-Smirnov Tests und alle χ^2-Tests fehl [20].

Weiterhin werden nun einige bereits existierende neuere Arbeiten vorgestellt, die, ebenso wie diese Arbeit, das langfristige Zugriffsverhalten ganzer Datenbanken oder Teilen davon anhand von Stichproben mit statistischen Methoden analysieren. Zweck aller Arbeiten ist es, Erkenntnisse für ein effizientes hierarchisches Speichermanagement (HSM) mit automatischer Datenmigration zu ermöglichen.

Das langfristige Zugriffsverhalten wurde von Strange über einen Zeitraum von 84 Tagen [100], von Gibson, Miller und Long über einen Zeitraum von 120 bis 280 Tagen [29, 30] und von Schmitz über einen Zeitraum von 12 Wochen [88] hinweg analysiert. Die untersuchten Daten stammen bei den genannten Autoren aus Dateisystemen deutscher [88] sowie amerikanischer [29, 30, 100] Forschungseinrichtungen.

Strange untersuchte das Zugriffsverhalten von insgesamt 203 meist kleinen Dateien aus sieben UNIX-Dateisystemen [100]. Er stellte einige Migrationsalgorithmen vor und führte Simulationen mit zwei Speicherebenen durch, um die Algorithmen zu vergleichen. Der so genannte „only-file-size-algorithm" verschiebt Dateien ab einer bestimmten Größe auf eine zweite Speicherebene. Er orientierte sich an der Größe, weil es nach seiner Meinung besser war, wenige große Dateien zu verschieben als viele kleine. Ebenso war es sein Interesse, die Wahrscheinlichkeit, dass eine bereits migrierte Datei doch noch gebraucht wird, zu reduzieren. Der „least-recently used algorithm" verschiebt die Dateien zuerst, die am längsten nicht benutzt wurden. Die dafür maßgebliche Anzahl der Tage zwischen zwei aufeinander folgenden Zugriffen bezeichnet Strange als „interreference interval". Er unterscheidet nur, ob an einem Tag auf eine Datei zugegriffen wurde oder nicht (binäre Auswertung). Es entstand also ein Informationsverlust, da mehrere Zugriffe am gleichen Tag nicht einzeln betrachtet wurden. Außerdem wurden keine weitergehenden statistischen Analysen durchgeführt.

Strange betrachtet auch zwei Kombinationen der beiden genannten Migrationsalgorithmen der Formen $Zeit \cdot x + Dateigröße$ und $Zeit \cdot Dateigröße^x$ (Zeit = Zeit zw. Zugriffen, x = Gewichtungsfaktor). Der zuletzt genannte Algorithmus erzielte bei den Simulationen die besten Ergebnisse.

Gibson und Miller untersuchten den so genannten „file-aging algorithm", der zusätzlich zur Dateigröße und Zeit seit der letzten Benutzung einen zuvor berechneten Migrationswert berücksichtigt [29, 30, 100]. Dieser Migrationswert soll die Intensität der Benutzung im Zeitverlauf erfassen: Er erhöht sich an einem Tag, an dem die betreffende Datei benutzt wurde, und verringert sich mit jedem Tag der Nichtbenutzung.

Schmitz nahm ebenfalls nur eine binäre Auswertung der erfolgten Zugriffe vor [88]. Sie ordnete die Dateien anhand der Anzahl der Tage seit dem letzten Zugriff in Alterskategorien und anhand der Dateigrößen in Größenkategorien ein. Es folgte eine Analyse der Zusammenhänge zwischen Zugriffswahrscheinlichkeiten und Zeit seit letztem Zugriff bzw. Größe bzw. beider Merkmale. Schmitz definiert als Zugriffswahrscheinlichkeit den Quotient aus der Anzahl zugegriffener Dateien der entsprechenden Kategorie am Tag i und der Anzahl aller zugegriffener Dateien am Tag i.

Diese Arbeit ist bezüglich des beobachteten Zeitraumes langfristiger als die erwähnten angelegt. Im Folgenden wird der komplette Lebenszyklus von Dateien im Alter von bis zu 1771 Tagen (4,85 Jahre) analysiert. Dies wird dadurch ermöglicht, dass das betrachtete System Zugriffe grundsätzlich aufzeichnet und diese Historie bereits für jede Datei vorhält. Die Dateien für die Analysen in dieser Arbeit wurden einer deutschen Unternehmensdatenbank entnommen (siehe nächster Abschnitt) und sind zudem um ein bis 13 Jahre aktueller.

Im Unterschied zu den genannten Arbeiten wird hier zusätzlich zum statistischen Ansatz der deskriptiven Statistik der wahrscheinlichkeitstheoretische Ansatz verfolgt. Das führt insbesondere dazu, dass den Migrationsregeln signifikante statistische Resultate zugrunde liegen. Dies bietet keine der erwähnten Arbeiten.

5.4 Erzeugung der Stichprobe

Die Siemens AG verfügt über eine so genannte Erfahrungsdatenbank (EDB), in der sowohl die Dateien gespeichert als auch alle Zugriffe auf diese Dateien protokolliert werden. Insgesamt enthält diese Datenbank schätzungsweise 150.000 Dateien und ihre Zugriffsprotokolle. Für die Durchführung der folgenden statistischen Analysen wurde aus der Datenbank eine Zufallsstichprobe von 1000 Dateien entnommen. Die Auswahl erfolgte zufällig mit Hilfe des Zufallszahlengenerators in Microsoft-Excel®.

Über jede Datei sind folgende Informationen bekannt: Dateityp, Dateigröße, Datum und Uhrzeit (minutengenau) der Dateierstellung sowie Datum, Uhrzeit und Art der einzelnen Zugriffe. Diese Informationen wurden nach der Stichprobenentnahme mit Microsoft-Excel® aufbereitet, so dass die für die jeweiligen Analysen benötigten Daten zur Verfügung standen. Für die Datenanalyse selbst wurden die Programme R und MATLAB® verwendet.

R ist sowohl eine Programmiersprache, als auch der Name eines Software-Systems, das im Internet frei verfügbar ist. Die Programmiersprache R wurde speziell für die Statistik und stochastische Simulationen entwickelt. Konkret heißt die Sprache S, ihre Implementierung und das System heißen R [85].

5.5 Beschreibung der Stichprobenmerkmale

In diesem Teilkapitel erfolgt zunächst die nähere Betrachtung einzelner Merkmale der Stichprobe mit den Hilfsmitteln der deskriptiven Statistik. Häufigkeitstabellen charakterisieren die Stichprobe und stellen dadurch die Grundlage für weitergehende Analysen dar.

Tabelle 4 bis Tabelle 9 charakterisieren die Stichprobe, indem sie die Häufigkeitsverteilungen der Anzahl der Zugriffe pro Datei, der Größe der Dateien, der Größe der Zugriffe, des Alters der Dateien sowie die in der Stichprobe enthaltenen Dateitypen und Zugriffsarten darstellen.

Anzahl Zugriffe	[1;2)	[2;3)	[3;4)	[4;5)	[5;10)
Anzahl Dateien	307	152	99	79	209
Anzahl Zugriffe	[10;20)	[20;50)	[50;100)	[100;200)	[200;292)
Anzahl Dateien	77	53	14	6	4

Tabelle 4: Anzahl der Zugriffe je Datei

Auf die 1000 Dateien der Stichprobe wurde seit ihrem jeweiligen Bestehen bis zur Stichprobenentnahme insgesamt 7911-mal zugegriffen (siehe Tabelle 4). Bei der Anzahl der Zugriffe ist jedoch zu beachten, dass in der untersuchten Datenbank stets der erste Zugriff auf eine Datei zum Zeitpunkt der Dateientstehung protokolliert ist. Auf 307 der 1000 Dateien wurde demzufolge nach dem Entstehungszeitpunkt kein einziges mal mehr zugegriffen. Diese „nicht verwendeten" Dateien einmal ausgenommen, wurde auf die meisten Dateien, nämlich 152, nur einmal nach dem Entstehungsdatum zugegriffen.

Dateigröße	[1kB;10kB)	[10kB;50kB)	[50kB;100kB)	[100kB;500kB)	[500kB;1MB)
Anzahl Dateien	22	265	158	267	108
Dateigröße	[1MB;2MB)	[2MB;5MB)	[5MB;10MB)	[10MB;50MB)	[50MB;115MB)
Anzahl Dateien	81	48	36	12	3

Tabelle 5: Größe der Dateien

Folgende Werte charakterisieren die Größe der untersuchten Dateien zusätzlich zu Tabelle 5: Das Maximum beträgt 114,85 MB, die durchschnittliche Größe 1139,96 kB, die Standardabweichung 5174,05 kB und der Median 117,50 kB. Abbildung 5.2 zeigt, dass nur 22 Dateien (2,2 %) unter 10 kB groß sind und nur 15 Dateien (1,5 %) größer als 10 MB sind.

Weiterhin sind die durch Zugriffe auf Dateien entstandenen Größenänderungen im vorliegenden Fall so gering, dass sie vernachlässigt werden können.

Zugriffsgröße	[1kB;10kB)	[10kB;50kB)	[50kB;100kB)	[100kB;500kB)	[500kB;1MB)
Anzahl Zugriffe	169	2357	1228	1408	1458

Zugriffs-größe	[1MB;2MB)	[2MB;5MB)	[5MB;10MB)	[10MB;50MB)	[50MB;115MB)
Anzahl Zugriffe	440	426	322	65	28

Tabelle 6: Größe der Zugriffe (Größe der jeweils zugegriffenen Datei)

Als die Größe der Zugriffe wird die Größe der jeweils zugegriffenen Datei bezeichnet. Durch die Gewichtung jedes Zugriffes mit der Dateigröße enthält diese Darstellung mehr Informationen als die vorigen Betrachtungen der Anzahl der Zugriffe auf eine Datei oder der Dateigrößen. Im Durchschnitt wurde auf 1167,96 kB zugegriffen, während der Median nur 123 kB beträgt und die Standardabweichung 4992,42 kB.

In Tabelle 6 ist zu sehen, dass bei knapp 30 % der Zugriffe, nämlich 2357-mal, auf lediglich 10 bis 50 kB zugegriffen wird.

Dateialter	[0;1 W)	[1 W;1 M)	[1 M;¼ J)	[¼ J;½ J)	[½ J;1 J)
Anzahl Dateien	7	37	87	109	231
Dateialter	[1 J;1½ J)	[1½ J;2 J)	[2 J;3 J)	[3 J;4 J)	[4 J;5 J)
Anzahl Dateien	247	80	138	36	28

Tabelle 7: Alter der Dateien (W = Woche, M= Monat, J= Jahr)

Es handelt sich hier um eine langfristige Analyse des Dateizugriffsverhaltens mit entsprechend hohem Alter der untersuchten Dateien. Die meisten Dateien sind zwischen einem halben und eineinhalb Jahren alt.

Über die Hälfte der Dateien (52,9 %) sind älter als ein Jahr. Das Durchschnittsalter beträgt 463,3 Tage, das maximale Alter beträgt 1771,09 Tage (4,85 Jahre), der Median 388,3 Tage und die Standardabweichung 357,2 Tage.

Dateityp	doc	xls	ppt	pdf	zip	msg	Sonstige
Anzahl Dateien	335	185	164	140	41	24	111

Tabelle 8: Dateitypen

Die Dateitypen doc, xls, ppt, pdf und zip sind am häufigsten in der Stichprobe enthalten (siehe Tabelle 8). Unter „Sonstige” fallen die Dateitypen avi, cfg, csv, cti, dot, exe, gif, htm, jpg, log, mdb, mmap, mmp, mp3, mpg, mpp, pps, pst, rtf, sql, tif, trc, txt, vsd, vss, wav, wbk, wf2 und xml.

Zugriffsart	Version Fetched	View	Version Added	Move
Anzahl Zugriffe	169	2357	1228	1408
Zugriffsgröße	Reserve	Unreserve	Permission Changed	Sonstige
Anzahl Zugriffe	440	426	322	65

Tabelle 9: Zugriffsarten

Die meisten Zugriffe auf Dateien der Stichprobe, nämlich 46,22 % von 7911, sind vom Typ „Version Fetched" (siehe Tabelle 9). Die Zugriffsarten „View" und „Version Added" sind mit 19,20 % bzw. 17,60 % am zweit- bzw. dritthäufigsten vertreten. Weitere häufig auftretende Zugriffsarten sind „Move", „Reserve", „Unreserve" und „Permission Changed". Die deutlich seltener vorkommenden unter „Sonstige" fallenden Zugriffsarten sind „Attributes Changed", „Rename", „Copy", „Version Deleted", „Alias Created" und „Generation Created".

5.6 Untersuchung des Kriteriums „Vergangene Zeit seit Dateierstellung"

In diesem Abschnitt wird für alle Zugriffe die Zeit betrachtet, die seit Erstellung der Datei vergangen ist. Mit anderen Worten ist dies das Alter einer Datei zum Zeitpunkt eines Zugriffes. Im Gegensatz zur statischen Betrachtungsweise im vorherigen Abschnitt, die das Alter der Dateien zum Zeitpunkt der Stichprobenentnahme darstellte, handelt es sich hier um eine dynamische Betrachtungsweise, die die Entwicklung der Anzahl der Zugriffe im Zeitverlauf erkennen lässt.

Ziel dieser Betrachtung ist es festzustellen, ob das Alter einer Datei als Kriterium für eine Migrationsregel in der Datenbank in Frage kommt. Dieses Kriterium ist für eine Regelerstellung geeignet, wenn ab einer überschaubaren Zeit seit der Erstellung einer Datei nur noch sehr wenige Zugriffe auf eine Datei zu erwarten sind.

Die durchschnittlich vergangene Zeit seit Dateierstellung beträgt 201,43 Tage, das Maximum beträgt 1748,8 Tage, der Median 71,9 Tage und die Standardabweichung 305,3 Tage.

Da die Stichprobe Dateien verschiedenen Alters enthält (siehe vorheriger Abschnitt), wurden Teilstichproben mit einem jeweiligen Mindestalter von einem halben, einem und zwei Jahren separat betrachtet. Das entspricht einer Anzahl von 760, 529 bzw. 202 Dateien. Das Zugriffsverhalten wurde dementsprechend nur bis zum Mindestalter der Dateien analysiert, nämlich bis zu einem halben bzw. einem bzw. zwei Jahren nach Dateierstellung. So wird der systematische Fehler vermieden, Zugriffe auf Dateien, die z.B. nur sechs Monate alt sind, über einen Zeitraum von z.B. einem Jahr zu betrachten. In Abbildung 9 ist die Verteilung der Zugriffe der mindestens ein Jahr alten Dateien über einen Zeitraum von einem Jahr dargestellt.

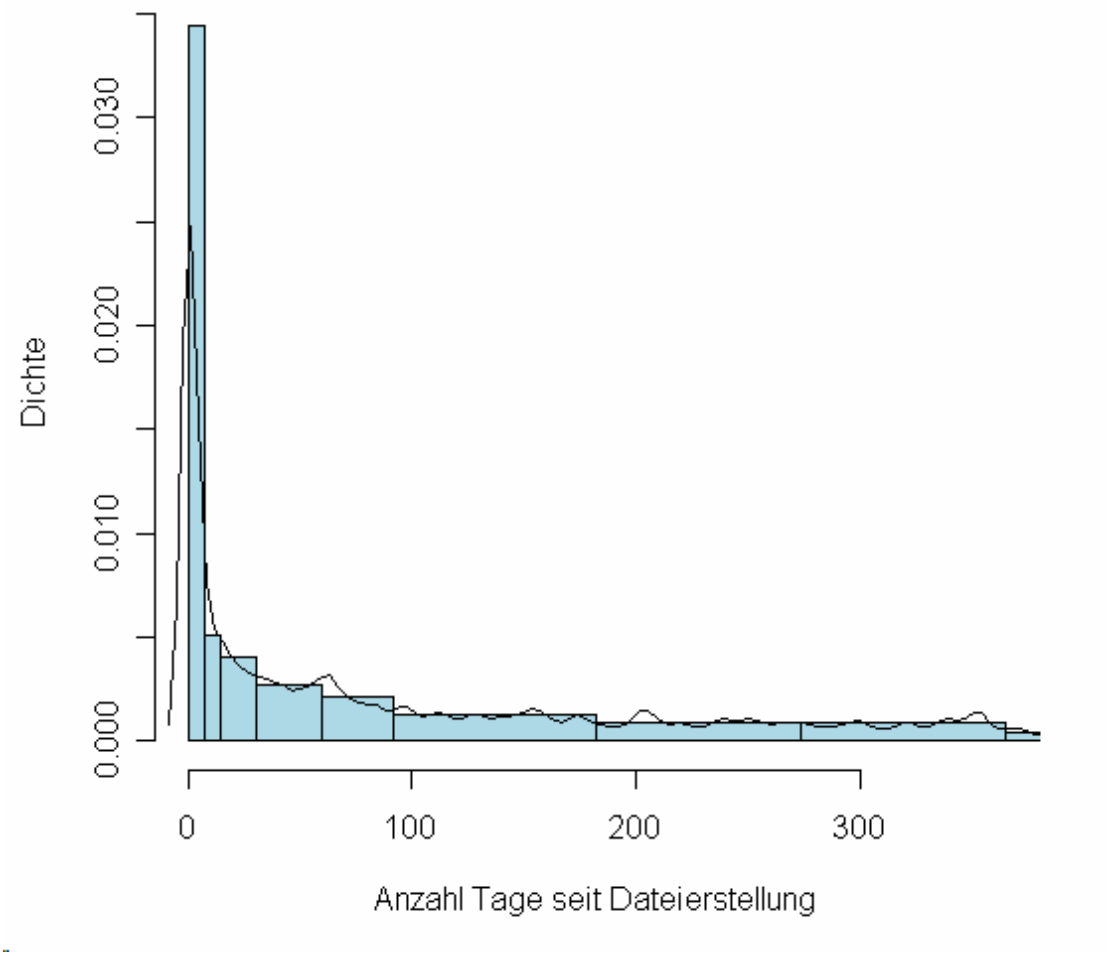


Abbildung 9: Vergangene Zeit seit Erstellung von Dateien mit Mindestalter 1 Jahr

Zeit seit Erstellung	[0;1 W]	(1W;2 W]	(2W;1M]	(1M;2M]
Anzahl Zugriffe	1155	170	306	389
Zeit seit Erstellung	(2M;1/4J]	(1/4J;1/2J]	(1/2J;3/4J]	(3/4J;1J]
Anzahl Zugriffe	323	540	402	373

Tabelle 10: Vergangene Zeit seit Erstellung von Dateien mit Mindestalter 1 Jahr

An den Abszissen der Histogramme wird die Dichte abgetragen, die wie folgt definiert ist:

$$Dichte = \frac{relative\ Häufigkeit}{Klassenbreite} \tag{1}$$

Durch Verwendung der Dichte anstelle der Häufigkeit wird erreicht, dass die Flächen der Rechtecke eines Histogramms auch bei unterschiedlichen Klassenbreiten proportional den absoluten bzw. relativen Häufigkeiten sind. Histogramme können die Verteilung eines Merkmals gut veranschaulichen; allerdings hängt die Form eines Histogramms auch von der Wahl der Klassen ab. Aus diesem Grund ist in den Histogrammen zusätzlich die Dichtekurve dargestellt. Diese durch einen Kerndichteschätzer geschätzten Dichtekurven glätten zwar je nach gewählter Bandbreite mehr (bei großer Bandbreite) oder weniger (bei kleiner Bandbreite) Informationen heraus, aber dadurch wird der Kurvenverlauf nur wenig beeinflusst.

Am linkssteilen[4] Verlauf von Histogramm und Dichtekurve in Abbildung 9 erkennt man, dass die große Mehrzahl der Zugriffe kurz nach Erstellung einer Datei erfolgt. 865 Zugriffe bzw. 18 % aller Zugriffe erfolgen noch am ersten Tag. Die Anzahl der Zugriffe nach Ablauf des ersten Tages ist zwar deutlich geringer, nimmt aber nur sehr langsam mit der Zeit ab. An einigen Stellen steigt sie sogar wieder etwas an. Die Histogramme und Dichtekurven der Dateien der anderen beiden Altersgruppen haben einen ähnlichen Verlauf, weshalb sie hier nicht abgebildet sind.

Aufgrund der beschriebenen Entwicklung der Anzahl der Zugriffe im Zeitablauf ist die vergangene Zeit seit Dateierstellung im Falle der untersuchten Datenbank kein geeignetes Merkmal, um es in einer Migrationsregel zu verwenden.

5.7 Untersuchung des Kriteriums „Vergangene Zeit seit dem letzten Zugriff“

Nachdem sich das in Abschnitt 5.4 behandelte Merkmal der vergangenen Zeit seit Dateierstellung als ungeeignet für die Verwendung in einer Migrationsregel herausgestellt hat, soll nun in diesem Abschnitt die vergangene Zeit seit dem letzten Zugriff für alle Zugriffe untersucht werden.

Die Abbildung 10 veranschaulicht die unterschiedliche Betrachtungsweise von Abschnitt 5.6 und diesem Abschnitt.

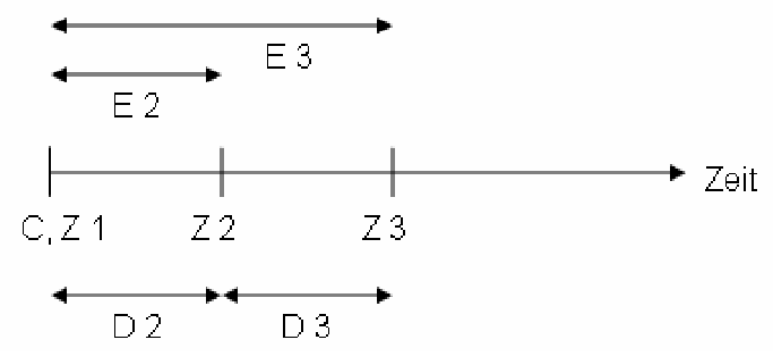


Abbildung 10: Veranschaulichung: C = Create, Z = Zugriff, E = Zeit seit Erstellung, D = Zeit seit letztem Zugriff

Am Beginn des Dateilebenszyklus steht das „Create” und der erste Zugriff; im weiteren Verlauf sind exemplarisch zwei weitere Zugriffe markiert. Die Zeitspannen E2 und E3 sind die Zeiten seit Erstellung für Zugriff 2 und 3. Die Zeitspannen D2 und D3 sind die Zeiten seit dem letzten Zugriff für Zugriff 2 und 3. Gemäß Definition ist E1 für alle Dateien Null. D1 ist nicht definiert, denn Zugriff 1 hat keinen Vorgänger und deshalb kann keine Zeit seit dem letzten Zugriff berechnet werden. Aufgrund dieser Tatsache fallen alle ersten Zugriffe der

[4] Eine eingipflige Häufigkeitsverteilung heißt linkssteil, falls $\bar{x} > \tilde{x}_{0,5} > x_{\text{mod}}$. Dabei bezeichnet $\bar{x}$ das arithmetische Mittel, $\tilde{x}_{0,5}$ den Median und x_{mod} den Modalwert (häufigster Wert) [24].

1000 Dateien aus der Betrachtung, wodurch sich die Anzahl der Zugriffe auf 6911 reduziert. 307 Dateien, die nur diesen einen Zugriff (Z1) zu verzeichnen hatten (siehe Abschnitt 5.5), fallen dadurch ganz weg, so dass von nun an 693 Dateien analysiert werden.

Im Gegensatz zum vorigen Abschnitt spielt hier das unterschiedliche Alter der in den Stichproben enthaltenen Dateien keine Rolle, da in diesem Fall die zeitliche Differenz zwischen den Zugriffen zugrunde liegt. Folglich wird die gesamte Stichprobe mit ihrer bestehenden Altersstruktur der Dateien analysiert.

Histogramm und Dichtekurve in Abbildung 11 haben einen sehr linkssteilen[5] Verlauf: Fast die Hälfte aller Zugriffe erfolgten innerhalb eines Tages nach dem letzten Zugriff. Die durchschnittlich vergangene Zeit seit dem letzten Zugriff beträgt 33,71 Tage.

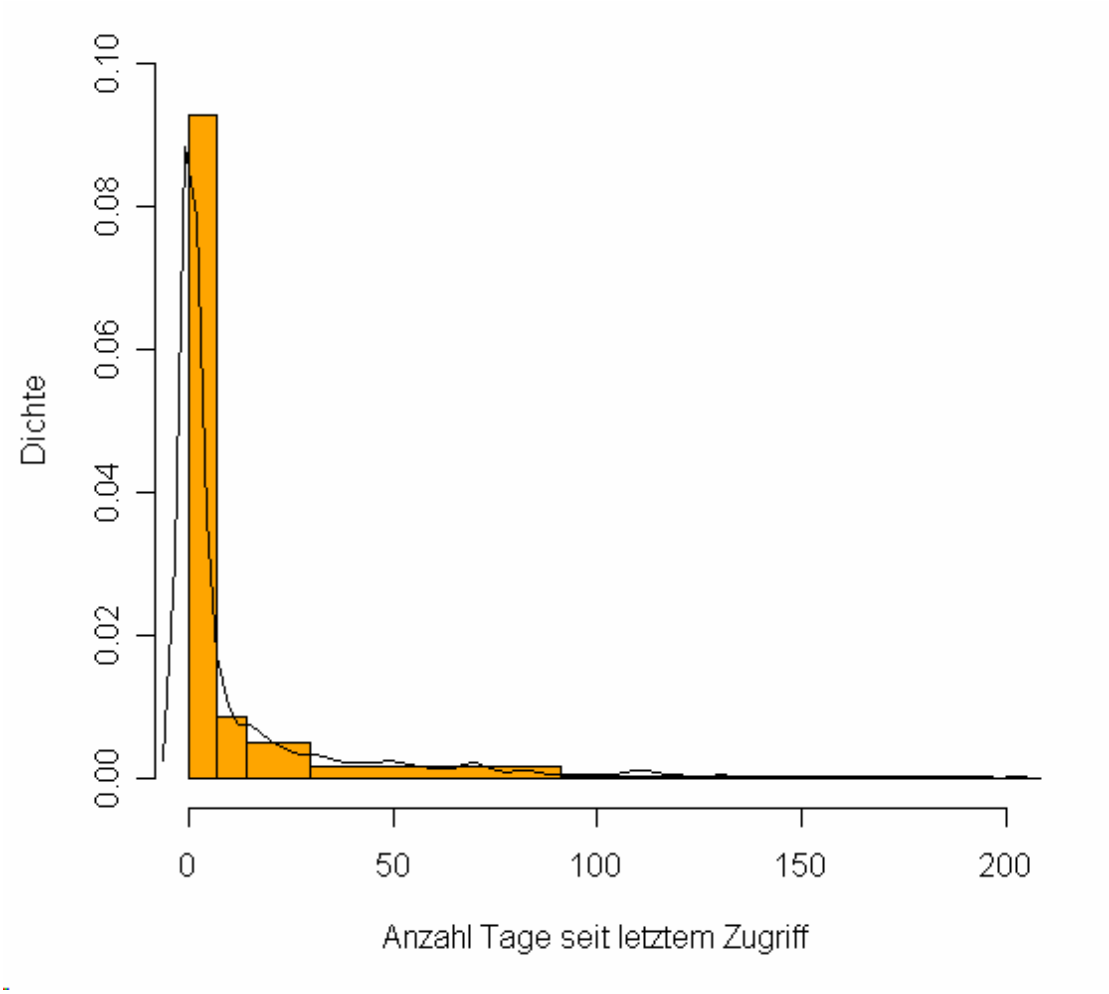


Abbildung 11: Vergangene Zeit seit dem letzten Zugriff

[5] Eine eingipflige Häufigkeitsverteilung heißt linkssteil, falls $\bar{x} > \tilde{x}_{0,5} > x_{\text{mod}}$. Im vorliegenden Fall ist 33,71 > 1, 02 > 0.

Zeit seit letztem Zugriff	[0;1 W]	(1W;2W]	(2W;1M]	(1M;1/4J]	(1/4J;1/2J]
Anzahl Zugriffe	4485	419	568	785	301
Zeit seit letztem Zugriff	(1/2J;3/4J]	(3/4J;1J]	(1J;2J]	(2J;3J]	(3J;3,7J]
Anzahl Zugriffe	131	125	66	24	7

Tabelle 11: Vergangene Zeit seit dem letzten Zugriff

Die Dichtekurve nähert sich bereits nach einem halben Jahr einer sehr kleinen Dichte: Nach Ablauf eines halben Jahres nach dem letzten Zugriff erfolgten nur noch 5,11 % der Zugriffe und nach einem dreiviertel Jahr nur noch 3,21 %. Es können also Zeiten seit dem letzten Zugriff angegeben werden, wie z.B. ein dreiviertel Jahr, nach deren Ablauf die Wahrscheinlichkeit weiterer Zugriffe sehr gering ist. Die vergangene Zeit seit dem letzten Zugriff kann daher als Charakteristikum für eine Migrationsregel verwendet werden und wird deshalb im nächsten Abschnitt auf Korrelation mit anderen Merkmalen überprüft.

5.8 Korrelationsanalyse

In diesem Abschnitt wird der Zusammenhang zwischen dem Merkmal „Anzahl der Tage seit dem letztem Zugriff" und den Merkmalen „Dateigröße" bzw. „Dateialter" untersucht. Als grafisches Hilfsmittel dient das *Streudiagramm*, aus dem ein möglicher Zusammenhang am besten zu ersehen ist. Dabei werden die Werte des einen Merkmals gegen die Werte des anderen in einem Koordinatensystem eingezeichnet. Die so entstehende Punktwolke gibt je nach ihrer Form Hinweise auf z.B. einen linearen, polynomialen oder exponentiellen Zusammenhang. Der Grad des linearen Zusammenhangs wird als *Korrelation zwischen zwei Merkmalen* bezeichnet und mit *Korrelationskoeffizienten* berechnet. Für die folgende Korrelationsanalyse wird der Pearsonsche Korrelationskoeffizient ausgewählt.

Der funktionale Zusammenhang zweier Merkmale wird mit einer *Regressionsanalyse* spezifiziert. Je nach dem, ob man einen linearen, polynomialen oder sonstigen Zusammenhang vermutet, wird eine lineare, polynomiale oder sonstige nichtlineare Regression durchgeführt. Das Ergebnis einer erfolgreichen Regressionsanalyse ist eine mathematische Funktion, die es ermöglicht, den Wert eines Merkmals aus dem Wert des anderen Merkmals zu schätzen.

Im Falle von ILM können diese Erkenntnisse für die Erstellung von solchen Migrationsregeln genutzt werden, die zwei Merkmale berücksichtigen.

5.8.1 Korrelationskoeffzienten

Als Schätzer für die Korrelation zweier Merkmale kann der *Pearsonsche Korrelationskoeffizient* oder ein *Rangkorrelationskoeffizient* (z.B. von Spearman) verwendet werden. Bei Rangkorrelationskoeffizienten schätzt man die Korrelation nur aufgrund von Ranginformationen, wobei die Abstände zwischen den Merkmalswerten nicht berücksichtigt werden. Das hat zur

Folge, dass ein Rangkorrelationskoeffizient unempfindlich gegenüber Ausreißern ist. Er ist jedoch nicht geeignet, wenn, wie in der vorliegenden Stichprobe, sehr viele Bindungen (gleiche Merkmalswerte) auftreten. Außerdem entsteht durch die Nichtbeachtung der Abstände zwischen den Merkmalswerten ein großer Informationsverlust. Aus diesem Grund scheiden Rangkorrelationskoeffizienten im vorliegenden Fall aus.

Für die folgende Korrelationsanalyse wird der Pearsonsche Korrelationskoeffizient *r* ausgewählt, der in Gleichung 2 definiert und in Tabelle 12 interpretiert ist [86].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

Korrelationskoeffizient	Interpretation
|r| = 0	Keine Korrelation
|r| < 0,5	Schwache Korrelation
0,5 ≤ |r| < 0,8	Mittlere Korrelation
0,8 ≤ |r| < 1	Starke Korrelation
|r| = 1	Perfekte Korrelation

Tabelle 12: Interpretation des Pearsonschen Korrelationskoeffizienten

Der Pearsonsche Korrelationskoeffizient hat den Vorteil, dass er ohne Informationsverlust berechnet wird, aber er ist empfindlich gegenüber Ausreißern. Aus diesem Grund wird in den folgenden Abschnitten jede Punktwolke auf Ausreißer untersucht. Die Ausreißer werden dann nicht bei der Berechnung des Pearsonschen Korrelationskoeffizienten berücksichtigt.

5.8.2 Korrelation zwischen den Merkmalen „Anzahl Tage seit letztem Zugriff“ und „Dateigröße“

In diesem Abschnitt wird der Einfluss der Dateigröße auf das Zugriffsverhalten untersucht bzw. ob die Anzahl der Tage seit dem letzten Zugriff und die Dateigröße in einem bestimmten Zusammenhang stehen. Dazu wird jedem Zugriff die Größe der zugegriffenen Datei zugeordnet, was auch als Zugriffsgröße bezeichnet werden kann.

Zur Berechnung des Pearsonschen Korrelationskoeffizienten wurden als Ausreißer[6] die vier mit Abstand größten Dateien[7] und der Zugriff mit der größten Anzahl der Tage seit dem letz-

[6] Die Ausreißerregel geht von dem Quartilsabstand $s_{Q=X_{0,75}-X_{0,25}}$ der empirischen Daten aus. Es werden alle Werte zu Ausreißern erklärt, die mehr als $1{,}5 \cdot s_Q$ vom unteren bzw. oberen Quartil entfernt sind [76] .

[7] Diese Dateien sind ca. 115 MB, 59 MB, 57 MB und 48 MB groß.

ten Zugriff[8] „aussortiert". Im Streudiagramm (siehe Abbildung 12) ist kein linearer Zusammenhang zu erkennen.

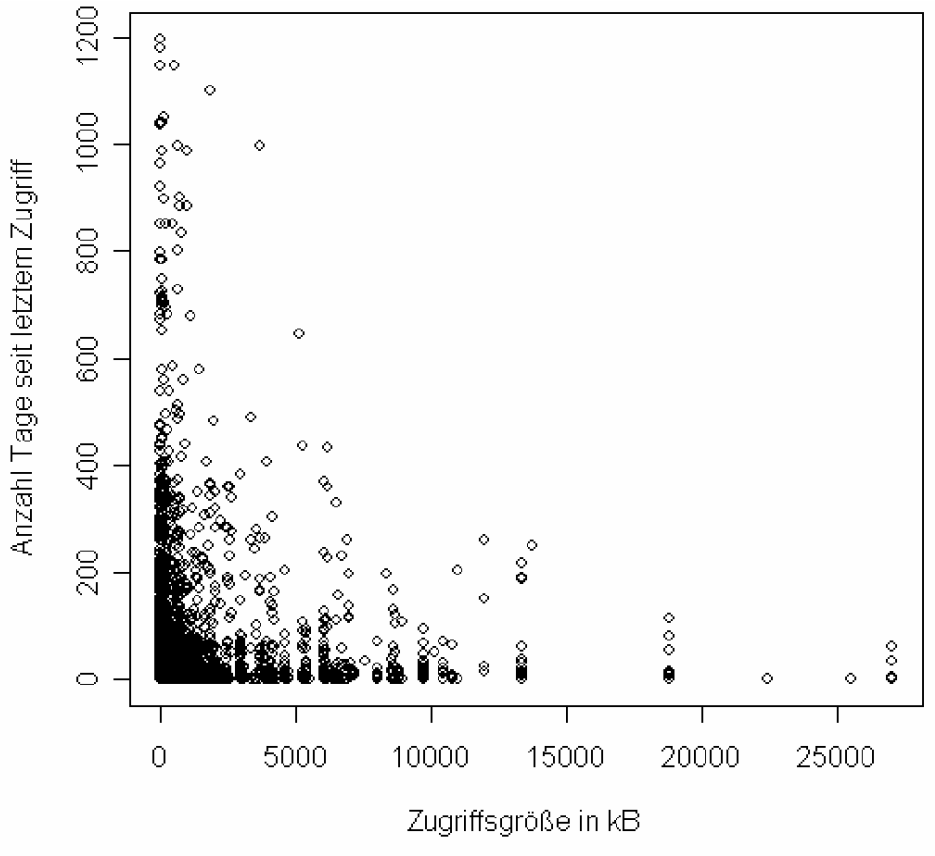


Abbildung 12: Streudiagramm: Anzahl Tage seit letztem Zugriff und Zugriffsgröße

Der Pearsonsche Korrelationskoeffizient bestätigt dies mit einem sehr kleinen Wert: Er beträgt 0,0107 mit einem 95%-Konfidenzintervall von [-0,0129; 0,0343], das bedeutet, es liegt keine Korrelation vor. Aufgrund der Form der Punktwolke kommt auch kein nichtlinearer Zusammenhang (polynomial, exponentiell, etc.) in Frage. Aus der Dateigröße können also keine Schlussfolgerungen für eine Migrationsregel gezogen werden.

Die Tabelle 13 bestätigt diese Erkenntnisse, da in allen drei Klassen der Dateigrößen[9] die Zugriffe gleichmäßig über die Klassen der Tage seit dem letzten Zugriff verteilt sind.

[8] Sie beträgt ca. 1327 Tage bzw. 3,64 Jahre.

[9] Die drei Klassen der Dateigrößen wurden so ausgewählt, dass auf jede Klasse ein Drittel aller Dateien entfallen. Auf diese Weise wird die Repräsentativität der einzelnen Größenklassen gewahrt.

Tage s. l. Zugriff / Dateigröße	[0T;1M)	[1M;1/2J)	[1/2J;1J)	[1J;2J)	[2J;3,7J)
[1kB;60kB)	2189	316	79	18	12
[60kB;600kB)	1465	392	90	26	8
[600kB;115MB)	1818	378	87	22	11

Tabelle 13: Verteilung der Anzahl der Zugriffe auf Tage seit letztem (s.l.) Zugriff und Dateigröße (T = Tage, M = Monat, J = Jahr)

Das Ergebnis dieses Abschnittes stimmt mit dem von Strange [100] überein, der bei seiner Analyse ebenfalls keine Korrelation zwischen der Zeit seit dem letztem Zugriff und der Dateigröße beobachten konnte. Im Gegensatz dazu stellte Schmitz [88] bei den von ihr untersuchten Daten fest, dass ein Zusammenhang zwischen der Zeit seit dem letzten Zugriff und der Dateigröße existiert. Sie stellte jedoch keine Funktion zur mathematischen Beschreibung dieses Zusammenhangs auf.

5.8.3 Korrelation zwischen den Merkmalen „Anzahl Tage seit letztem Zugriff“ und „Dateialter“

Für die Untersuchung des Zusammenhangs zwischen der Anzahl der Tage seit dem letztem Zugriff und dem Dateialter wurde jedem Zugriff das Alter der zugegriffenen Datei zugeordnet. Anschließend wurde bei beiden Merkmalen der mit Abstand größte Wert[10] als Ausreißer von der weiteren Betrachtung ausgeschlossen. Der Pearsonsche Korrelationskoeffizient beträgt in diesem Fall 0,1992, was lediglich eine schwache Korrelation bedeutet, mit einem 95%-Konfidenzintervall von [0,1765; 0,2218]. An der Form der Punktwolke (siehe Abbildung 13) ist kein nichtlinearer Zusammenhang (polynomial, exponentiell, etc.) zu erkennen, weshalb keine Regressionsanalyse in Betracht kommt.

[10] Ein Zugriff, der ca. 1327 Tage nach dem letzten Zugriff erfolgte, und eine ca. 1771 Tage (4,85 Jahre) alte Datei.

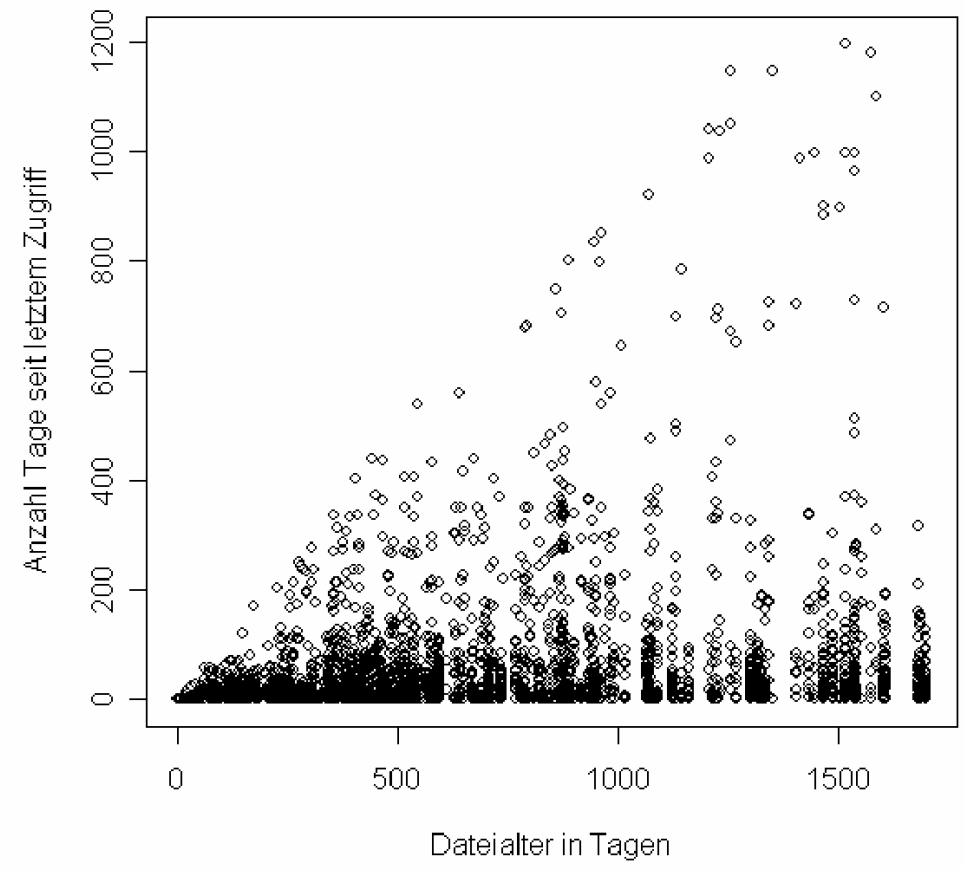


Abbildung 13: Streudiagramm: Anzahl Tage seit letztem Zugriff und Dateialter

Die Tabelle 14 weist keine Besonderheiten bezüglich der verschiedenen Altersklassen[11] auf, die für die Erstellung einer Migrationsregel relevant sind: Die große Anzahl von 2425 Zugriffen im ersten Zahlenfeld der Tabelle kommt durch die vielen Dateien mit nur geringem Alter zustande.

Tage s. l. Zugriff / Dateialter	[0T;1M)	[1M;1/2J)	[1/2J;1J)	[1J;2J)	[2J;3,7J)
[0T;1J)	2425	197	27	-	-
[1J;2J)	1511	385	65	19	-
[2J;3,7J)	1536	504	164	47	31

Tabelle 14: Verteilung der Anzahl der Zugriffe auf Tage seit letztem (s.l.) Zugriff und Dateialter

Tendenziell haben bei Dateien höheren Alters die Zugriffe größere zeitliche Abstände als bei weniger alten Dateien, ohne dass dies jedoch mit einer (mathematischen) Regel beschreibbar wäre.

[11] Die drei Klassen des Dateialters wurden so ausgewählt, dass auf jede Klasse ein Drittel aller Dateien entfallen. Auf diese Weise wird die Repräsentativität der einzelnen Altersklassen gewahrt.

Das Ergebnis dieses Kapitels ist, dass die betrachteten Zusammenhänge für die Verwendung in Regeln für die Wertzuweisung nicht geeignet sind, da jeweils nur eine schwache Korrelation vorliegt und keine Spezifikation mittels Regressionsanalyse möglich ist.

5.9 Theoretische Verteilungsmodelle

In Abschnitt 5.7 wurde bereits festgestellt, dass die vergangene Zeit seit dem letzten Zugriff als Charakteristikum für eine Migrationsregel geeignet ist. Die Definition der entsprechenden Zufallsvariablen lautet wie folgt:
Sei (Ω, F, P) ein beliebiger Wahrscheinlichkeitsraum. Die Abbildung $X : \Omega \rightarrow \mathbb{R}$

$$A : \{\omega \in \Omega \mid X(\omega) \leq x\} \in F \quad \forall x \in \mathbb{R} \tag{3}$$

heißt Zufallsvariable, falls
mit

- F: σ-Algebra
- P: Wahrscheinlichkeitsmaß
- Ω: Grundgesamtheit, hier: Menge aller Zugriffe
- ω: Beliebiges Element aus Ω, hier: Zugriff
- X(ω): Hier: Anzahl der Tage seit dem letztem Zugriff

5.9.1 Eingrenzung geeigneter Verteilungsmodelle

Im Folgenden wird für die *Zufallsvariable X: „Anzahl der Tage seit dem letzten Zugriff"* ein geeignetes theoretisches Verteilungsmodell gesucht.

Häufig können Zufallsvariablen nur diskret gemessen werden, lassen sich aber aufgrund der feinen Abstufung wie stetige Zufallsvariablen behandeln. In solch einem Fall spricht man von *quasi-stetigen* Zufallsvariablen [24]. Demnach ist die Zufallsvariable X *quasi-stetig*, da die Zeitspanne zwischen den Zugriffen zwar stetig ist, aber nur diskret, nämlich minutengenau [12] gemessen wurde. Eine quasi-stetige Zufallsvariable ist wie eine stetige Zufallsvariable zu behandeln [24]. Es wird also ein stetiges Verteilungsmodell gesucht.

Zugriffe auf Informationsobjekte werden häufig mit der so genannten *Zipf-Verteilung* modelliert. Studien haben gezeigt, dass Internetabfragen und Filesharing-Anwendungen wie Napster und Gnutella dieser Verteilung folgen [40]. Es gibt allerdings auch gegenteilige Erkenntnisse [80]. Die Zipf-Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die natürlichen Zahlen Wahrscheinlichkeiten zuordnet. In diesem Fall werden den Informationsobjekten gemäß der Häufigkeit ihrer Aufrufe Rangzahlen zugeordnet, die ihre Popularität ausdrücken. Dabei verliert man jedoch alle Informationen über die zeitliche Entwicklung sowie die zeitlichen

[12] Aufgrund der langen betrachteten Zeiträume wird für die minutengenau gemessenen Daten die Einheit „Tage" verwendet (siehe Abschnitte 5.6 und 5.7).

Zusammenhänge der Zugriffe. Gerade Letzteres ist jedoch in dieser Arbeit von großer Bedeutung, denn aus der zeitlichen Entwicklung der Zugriffe soll mit Hilfe von Verteilungsmodellen das Zugriffverhalten modelliert werden. Des Weiteren ist die Zipf-Verteilung eine diskrete Verteilung, aber in dieser Arbeit kommen nur stetige Verteilungen in Frage. Die Modellierung von Zugriffen mit der Zipf-Verteilung ist daher im vorliegenden Fall nicht geeignet.

Zur weiteren Eingrenzung der in Frage kommenden stetigen Verteilungen wird der grundsätzliche Verlauf der empirischen Dichte- und Verteilungsfunktion mit den Kurvenverläufen stetiger Verteilungsmodelle verglichen. Dichtekurve und Histogramm in Abbildung 11 (Abschnitt 5.7) zeigen, dass die vorliegende Häufigkeitsverteilung extrem linkssteil ist. Demzufolge sind symmetrische Verteilungen wie z.B. die Normalverteilung nicht zur Modellbildung geeignet. Stetig und linkssteil sind dagegen die so genannten *Lebensdauerverteilungen*. Ferner deutet auch der Sachverhalt - es wird eine Zeitspanne zwischen zwei Ereignissen modelliert - auf eine Lebensdauerverteilung hin.

Nach Hartung [32] und Schlittgen [86] gehören zu den Lebensdauerverteilungen

• die Exponentialverteilung,

• die Weibullverteilung,

• die Rayleighverteilung,

• die Gammaverteilung,

• die Erlangverteilung und

• die IDB-Verteilung (Hjorth-Verteilung).

Die IDB-Verteilung kann bereits ausgeschlossen werden, da sie (unabhängig von der Wahl der Parameter) immer eine badewannenförmige und keine linkssteile Dichtefunktion besitzt [32].

In den nächsten Abschnitten werden die Lebensdauerverteilungen im Einzelnen vorgestellt und überprüft. Dabei wird folgendermaßen vorgegangen:

1. Zuerst wird die theoretische Verteilung an die empirische Verteilung angepasst. Dafür müssen die Parameter der theoretischen Verteilungsfunktion mit entsprechenden Methoden geschätzt werden, so dass der Verlauf der empirischen Verteilungsfunktion möglichst exakt nachgebildet wird.

2. Danach folgt die Überprüfung der jeweiligen Verteilungsannahme mit Hilfe geeigneter statistischer Tests.

Das Ziel statistischer Tests ist es festzustellen, ob eine Hypothese über z.B. die Verteilung einer Zufallsvariablen bei einem bestimmten Signifikanzniveau α angenommen oder abgelehnt wird.

Im folgenden Abschnitt werden einige ausgewählte statistische Testverfahren vorgestellt.

5.9.2 Statistische Tests

Nachfolgend werden die für diese Arbeit relevanten Testverfahren erläutert: Der Q-Q-Plot, der Kolmogoroff-Smirnov-Anpassungstest und der χ^2-Anpassungstest[13].

5.9.2.1 Der Q-Q-Plot

Wie bereits im vorherigen Abschnitt beschrieben, können durch den Vergleich der Kurvenverläufe von empirischer und theoretischer Verteilungsfunktion die in Frage kommenden Verteilungen eingegrenzt werden. Eine genauere optische Überprüfung, ob die angenommene Verteilung auch tatsächlich vorliegt, wird jedoch durch die Krümmung im Verlauf der Verteilungsfunktionen wesentlich erschwert. Für das Auge ist lediglich die Abweichung von einer Geraden gut zu erkennen. Der Quantil-Quantil-Plot, kurz Q-Q-Plot, ist so konstruiert, dass bei Übereinstimmung von empirischer und theoretischer Verteilung eine Gerade resultiert und Unterschiede in Form von Abweichungen von der Geraden zu erkennen sind [32]. Auf diese Weise können Verteilungsannahmen optisch und mit geringem Aufwand überprüft werden.

Beim Q-Q-Plot werden die empirischen Quantile[14] gegen die entsprechenden theoretischen Quantile[15] der interessierenden Verteilung in einem Koordinatensystem abgetragen. Bei übereinstimmenden Verteilungen liegen alle Punkte auf der Ursprungsgeraden mit der Steigung 1.

In dieser Arbeit dient der Q-Q-Plot vor allem der Veranschaulichung. Für eine genauere Überprüfung der Verteilungsannahmen werden Anpassungstests durchgeführt.

5.9.2.2 Der Kolmogoroff-Smirnov-Anpassungstest

Soll bei stetigen Zufallsvariablen überprüft werden, ob eine empirische Verteilungsfunktion mit der (angepassten) Verteilungsfunktion einer ausgewählten Verteilung übereinstimmt, so kann man z.B. den Kolmogoroff-Smirnov-Anpassungstest, kurz KS-Test, verwenden. Er ist ein exakter Test, bei dem die Daten nicht klassiert werden müssen und deshalb kein Informationsverlust entsteht. Der KS-Test ist allerdings nur zur Überprüfung vollständig spezifizierter Verteilungen geeignet. Wenn also die Parameter der Verteilung wie im vorliegenden Fall unbekannt sind und geschätzt werden müssen, kann der Test in seiner ursprünglichen Form nicht mehr durchgeführt werden. Wegen der großen praktischen Bedeutung einer Anwendung des Tests auch bei zu schätzenden Parametern gibt es die so genannte Lilliefors-Modifikation des

---

[13] Testet man nicht Hypothesen über den Parameter einer Verteilung, sondern über den Typ einer Verteilung, so spricht man von einem Anpassungstest.

[14] Der i-te Wert x(i) der geordneten Beobachtungsreihe x(1),...,x(n) entspricht gerade dem empirischen i/(n+1)-Quantil der Beobachtungsreihe.

[15] Im Falle einer stetigen Verteilungsfunktion F(x) bezeichnet man $\xi_{i/(n+1)} = F^{-1}(i/(n+1))$ als das i/(n+1)-Quantil einer Verteilung. Es gilt: $F(\xi_{i/(n+1)}) = i/(n+1)$.

KS-Tests [83, 86]. Diese Modifikation ermöglicht die Anwendung des KS-Tests im Fall zu schätzender Parameter einer Normal- oder einer Exponentialverteilung, aber nicht bei weiteren Verteilungsmodellen. Aus diesem Grund kommt in dieser Arbeit nicht der KS-Test, sondern der im nächsten Abschnitt beschriebene χ^2-Anpassungstest zur Anwendung.

5.9.2.3 Der χ^2-Anpassungstest

Liegen n unabhängige Beobachtungen $x_{(1)},...,x_{(n)}$ einer Zufallsvariablen X vor, kann man mit Hilfe des χ^2-Anpassungstests die Hypothese

> H_0: Die Beobachtungen $x_{(1)},...,x_{(n)}$ entstammen einer Verteilung mit der Verteilungsfunktion F(x)

gegen die Alternative

> H_1: Die Beobachtungen $x_{(1)},...,x_{(n)}$ entstammen nicht einer Verteilung mit der Verteilungsfunktion F(x)

zum Signifikanzniveau α testen.

Der Nachteil des χ^2-Anpassungstests ist, dass die Daten klassiert werden müssen, wodurch ein Informationsverlust entsteht. Die Klassierung erlaubt - streng genommen - auch nicht, die Hypothese H_0: $F(x) = F_0(x)$ zu überprüfen[16]. Nur die Gleichheit der Wahrscheinlichkeiten für die Klassen wird getestet.

Der Vorteil des χ^2-Anpassungstests gegenüber dem KS-Test ist seine problemlose Anwendbarkeit bei allen Verteilungsmodellen, wenn ein oder mehrere Parameter der theoretischen Verteilung geschätzt werden müssen. In diesem Fall wird einfach die Anzahl der Freiheitsgrade um die Zahl der geschätzten Parameter reduziert (siehe unten).

Bei der Durchführung des χ^2-Anpassungstests geht man folgendermaßen vor:

1. Schritt: Unterteilung des Wertebereiches der Beobachtungen $x_{(1)},...,x_{(n)}$ in k disjunkte Klassen.

In diesem Fall erstreckt sich der Wertebereich von 0 bis 1326,81 Tage seit dem letzten Zugriff. Es wurden bei den durchgeführten Tests jeweils 30 bis 50 Klassen gebildet.

2. Schritt: Bestimmung der Anzahl O_i (i=1,...,k) der Beobachtungswerte, die in jeder Klasse liegen.

3. Schritt: Berechnung der Wahrscheinlichkeiten p_i (i=1,...,k), mit denen eine Beobachtung unter der Hypothese H_0 in der i-ten Klasse liegt. Die Zahl $E_i = np_i$ (i=1,...,k) ist dann die Zahl der unter H_0 erwarteten Beobachtungen in der i-ten Klasse.

---

[16] F(x): Theoretische Verteilungsfunktion, F0(x): Empirische Verteilungsfunktion

4. Schritt: Berechnung der folgenden Testfunktion:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

T ist unter H_0 approximativ χ^2-verteilt mit $k - 1 - m$ Freiheitsgraden, wobei m für die Anzahl der geschätzten Parameter steht. Die Faustregel, ab wann die Approximation gilt, lautet $E_i \geq 5$ für alle i=1,...,k.

5. Schritt: Die Nullhypothese wird zum Signifikanzniveau α verworfen, falls gilt:

$$T > \chi^2_{k-1;-m;\alpha} \tag{5}$$

Aus der Liste der in Frage kommenden Verteilungen (siehe oben) wird nachfolgend die Exponentialverteilung mit Hilfe des Q-Q-Plots und des χ^2-Anpassungstests auf ihre Eignung hin überprüft.

5.9.3 Annahme einer Exponentialverteilung

In Abschnitt 5.9.1 wurden die in Frage kommenden Verteilungen der Zufallsvariablen X: „Anzahl der Tage seit dem letzten Zugriff" auf die Lebensdauerverteilungen eingegrenzt. Die einfachste und bekannteste Lebensdauerverteilung ist die Exponentialverteilung, die im Folgenden vorgestellt wird. Anschließend wird überprüft, ob die Zufallsvariable X aus einer exponentialverteilten Grundgesamtheit stammt.

5.9.3.1 Allgemeines zur Exponentialverteilung

Die Exponentialverteilung eignet sich zur Modellierung der Lebensdauer von Objekten, die nicht altern, und von Wartezeiten zwischen zwei Ereignissen. Eine stetige Zufallsvariable X ist *exponentialverteilt* mit dem Parameter $\lambda > 0$, wenn ihre Dichtefunktion durch

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6}$$

gegeben ist. Die entsprechende Verteilungsfunktion lautet

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{7}$$

Der Parameter λ kann mit Hilfe des arithmetischen Mittels $\bar{x}$ der Beobachtungen $x_{(1)},...,x_{(n)}$ geschätzt werden:

$$\hat{\lambda} = \frac{1}{\bar{x}} \tag{8}$$

5.9.3.2 Überprüfung der Annahme einer Exponentialverteilung

Zur Überprüfung einer Verteilungsannahme werden zunächst die Parameter der entsprechenden Verteilung geschätzt. Für den Parameter λ der Exponentialverteilung gilt im vorliegenden Fall $\overline{\lambda} = 0{,}024$.

In Abbildung 14 sieht man, dass sich die Verteilungsfunktion der Exponentialverteilung nicht an die empirische Verteilungsfunktion anpassen lässt: Sie kreuzen sich an einer Stelle, weichen aber sonst deutlich voneinander ab.

Der Q-Q-Plot verschafft Gewissheit darüber, dass die Zufallsvariable X nicht exponentialverteilt ist (siehe Abbildung 15), da die empirischen und theoretischen Quantile nicht auf einer Geraden mit der Steigung 1 liegen.

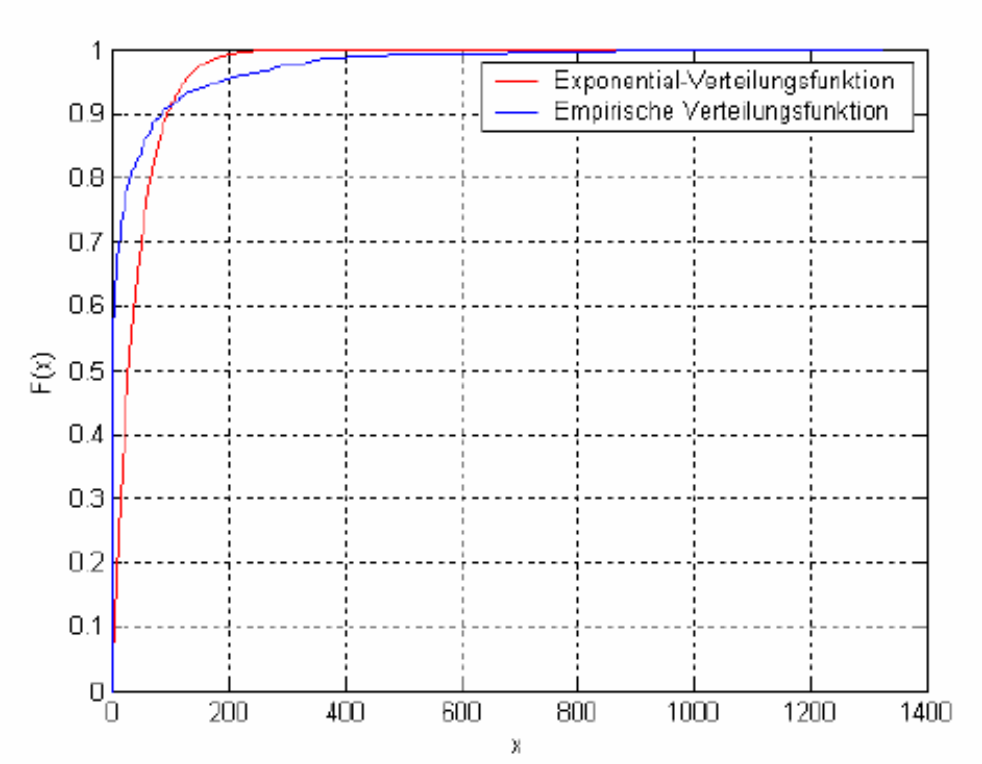


Abbildung 14: Exponential-Verteilungsfunktion mit $\overline{\lambda}$ = 0,024 und empirische Verteilungsfunktion

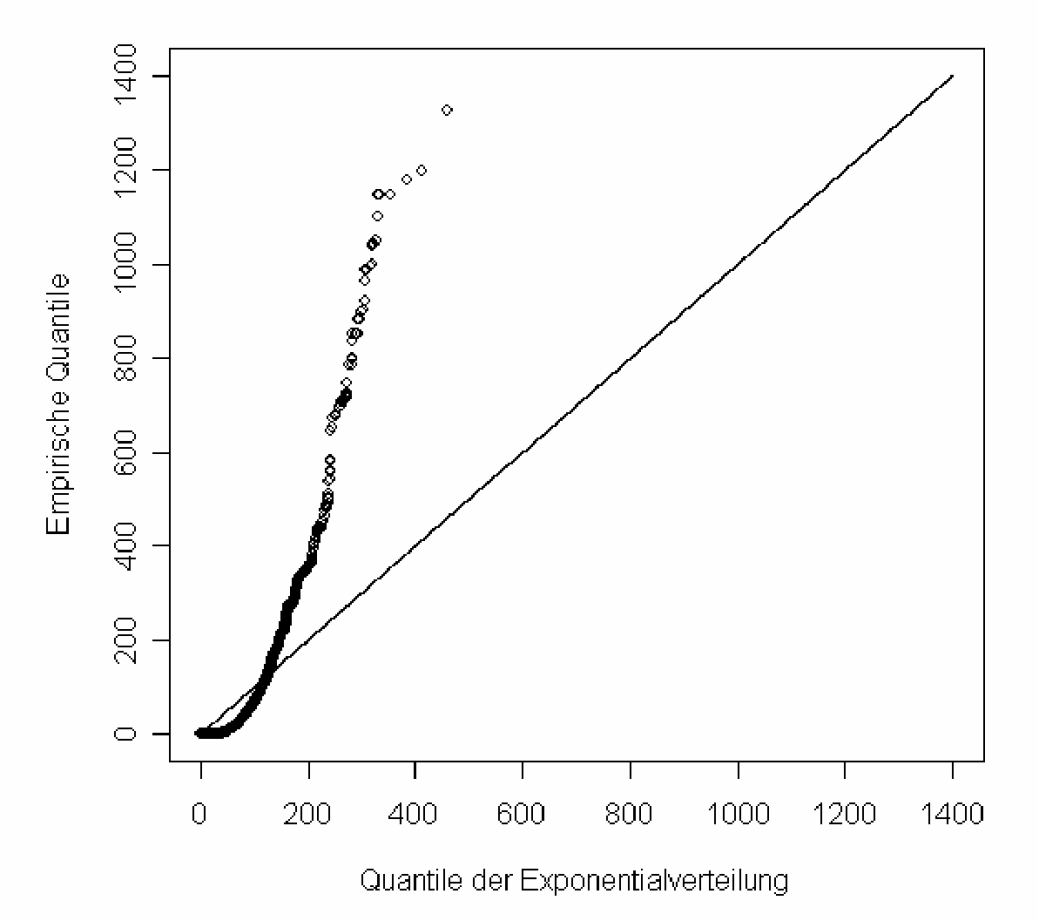


Abbildung 15: Q-Q-Plot der Exponentialverteilung mit $\bar{\lambda}$ = 0,024 und der empirischen Verteilung

Um das Ergebnis des Q-Q-Plots in Zahlen zu fassen und zur besseren Vergleichbarkeit im Hinblick auf die Testergebnisse der folgenden Abschnitte, wird außerdem der χ^2-Anpassungstest durchgeführt: Es wird die Hypothese

H_0: Die Zufallsvariable X entstammt einer Ex(0,024)-Verteilung

gegen die Alternative

H1: Die Zufallsvariable X entstammt nicht einer Ex(0,024)-Verteilung

zum Signifikanzniveau α = 0,001 getestet[17]. Die Prüfgröße T beträgt 32014,47 und es gilt:

$T > \chi^2_{31;0,001} \Leftrightarrow 32014,47 > 61,10$

Das heißt, *H_0* muss verworfen werden, was das Ergebnis des Q-Q-Plots bestätigt.

[17] Das Signifikanzniveau α ist die Wahrscheinlichkeit, mit der eine richtige Nullhypothese abgelehnt wird. Je näher α an der Null liegt, umso eher behält man die Nullhypothese bei.[83] [15]

5.9.4 Annahme einer Weibullverteilung

Im vorherigen Abschnitt zeigten die statistischen Tests, dass die Zufallsvariable X nicht aus einer exponentialverteilten Grundgesamtheit stammt. Daher wird nun die Weibullverteilung betrachtet, die über einen Parameter mehr als die Exponentialverteilung verfügt, wodurch sich ihre Verteilungsfunktion besser an eine empirische Verteilungsfunktion anpassen lässt. Ob diese größere Flexibilität für positive Testergebnisse ausreicht, soll nachfolgend geklärt werden.

5.9.4.1 Allgemeines zur Weibullverteilung

Die Weibullverteilung wird wie die Exponentialverteilung zur Beschreibung der Lebensdauer von Objekten verwendet, allerdings berücksichtigt sie auch Abnutzungserscheinungen bzw. Materialermüdungserscheinungen [4]. Eine Zufallsvariable X ist *weibullverteilt* mit den Parametern $\alpha > 0$ und $\beta > 0$, wenn ihre Dichtefunktion durch

$$f(x) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{9}$$

gegeben ist. Die Verteilungsfunktion der Weibullverteilung ist

$$F(x) = \begin{cases} 1 - e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{10}$$

Mit $\alpha = 1$ erhält man die Verteilungsfunktion der *Exponentialverteilung*, die daher einen Sonderfall der Weibullverteilung darstellt. Ein weiterer Sonderfall ist die *Rayleighverteilung*, die man mit $\alpha = 2$ erhält. Da im Folgenden die allgemeinere Weibullverteilung untersucht wird, braucht die Rayleighverteilung in dieser Arbeit nicht extra behandelt zu werden.

Der Vorteil der Weibullverteilung ist, dass sie im Gegensatz zur Exponential- und Rayleighverteilung über zwei einstellbare Parameter verfügt. Dadurch lässt sich ihre Verteilungsfunktion deutlich besser an eine empirische Häufigkeitsverteilung anpassen.

Die beiden Parameter der Weibullverteilung lassen sich jedoch nur mit aufwändigen Verfahren schätzen, deshalb wurde dafür das Programm MATLAB® eingesetzt. Die mit MATLAB® ermittelten Parameterwerte wurden dann wie im Falle der Exponentialverteilung bezüglich des χ^2-Anpassungstests und hier auch bezüglich einer gestutzten Verteilungsfunktion (siehe Abschnitt 5.9.4.3) mittels Rekursion weiter optimiert.

5.9.4.2 Überprüfung der Annahme einer Weibullverteilung

Die geschätzten Parameter der Weibullverteilung betragen $\hat{\alpha} = 0{,}31$ und $\hat{\beta} = 6{,}45$. Die angepasste Verteilungsfunktion der Weibullverteilung und die empirische Verteilungsfunktion haben einen sehr ähnlichen Verlauf (siehe Abbildung 16).

Der Q-Q-Plot zeigt eine bessere Anpassung als im Fall der Exponentialverteilung, jedoch liegen die meisten Punkte auch hier nicht auf der Geraden (siehe Abbildung 17), was eine nicht ausreichende Anpassung bedeutet.

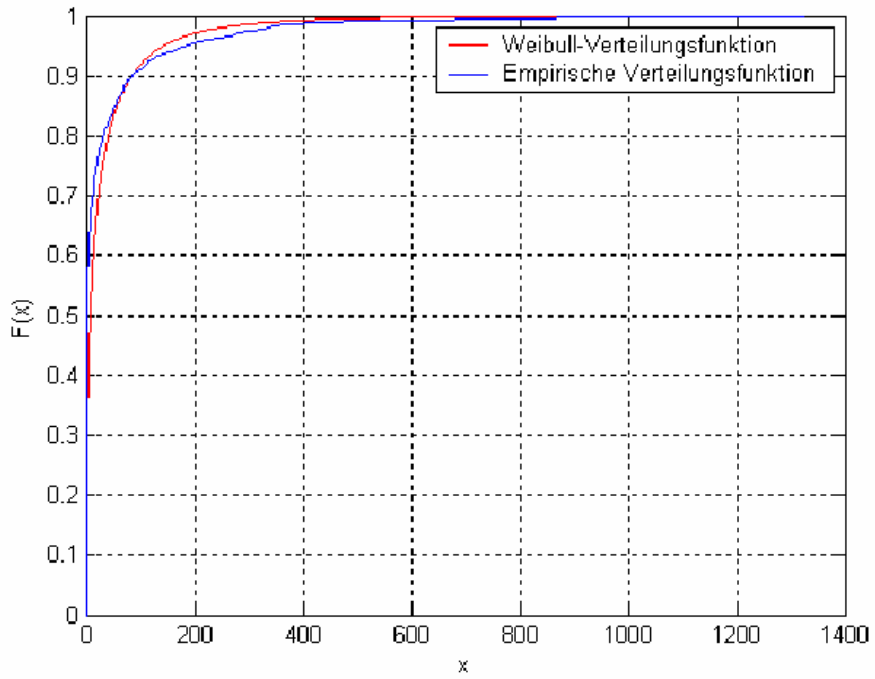


Abbildung 16: Weibull-Verteilungsfunktion mit $\hat{\alpha}$ = 0,31, $\hat{\beta}$ = 6, 45 und empirische Verteilungsfunktion

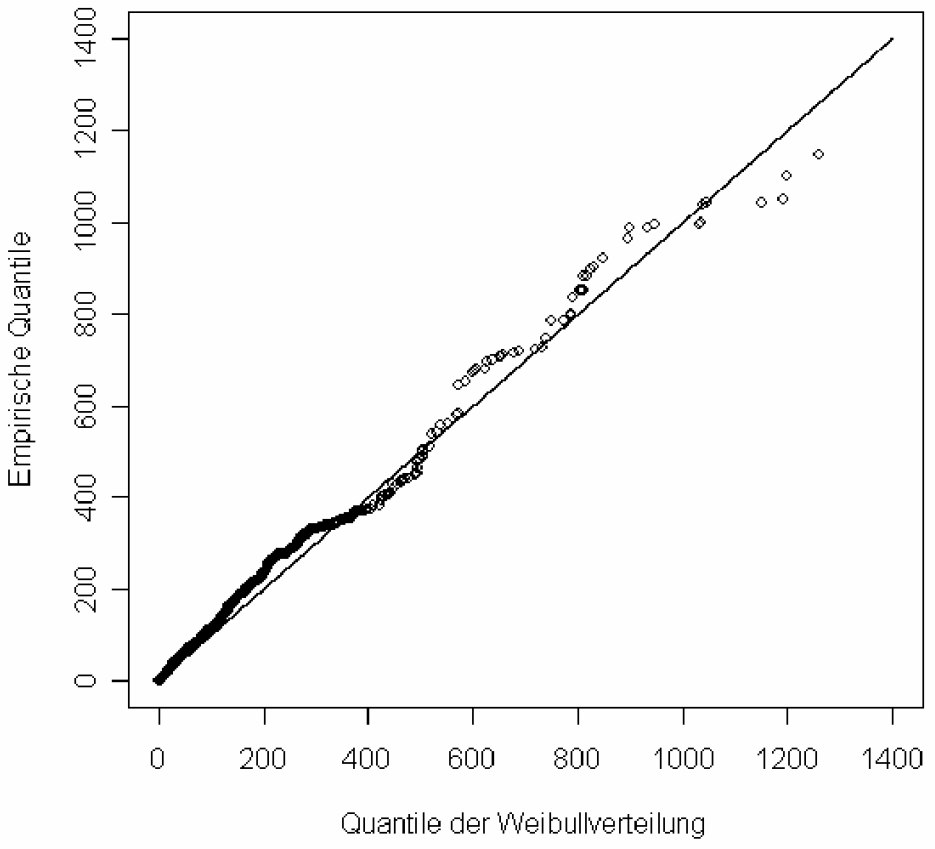


Abbildung 17: Q-Q-Plot der Weibullverteilung mit $\hat{\alpha}$ = 0,31, $\hat{\beta}$ = 6,45 und der empirischen Verteilung

Der χ^2-Anpassungstest gibt Aufschluss darüber. Die Hypothesen lauten:

H_0: Die Zufallsvariable X entstammt einer We(0,31;6,45)-Verteilung

gegen die Alternative

H_1: Die Zufallsvariable X entstammt nicht einer We(0,31;6,45)-Verteilung.

Es wird zum Signifikanzniveau 0,001 getestet. Die Prüfgröße T beträgt 307,89 und es gilt:

$T > \chi^2_{47;0,001} \Leftrightarrow 307,89 > 82,72$

Das heißt, H_0 muss verworfen werden.

Die bereits durch den Q-Q-Plot gewonnene Vermutung, dass die Anpassung an die empirischen Daten besser gelingt als im Fall der Exponentialverteilung, lässt sich beim χ^2-Anpassungstest an dem deutlich geringeren Wert der Testgröße T ablesen: Dem Wert von 32014,47 im Fall der Exponentialverteilung steht hier ein Wert von 307,89 gegenüber.

5.9.4.3 Gestutzte Weibullverteilung

Aufgrund des Verlaufs der Verteilungsfunktion der Weibullverteilung mit optimierten Parametern existieren - anders als bei der Exponentialverteilung – erwartete Beobachtungen bei über 1327 Tagen seit dem letzten Zugriff, also über dem Maximum der beobachteten Anzahl Tage seit dem letzten Zugriff. Das bedeutet, dass die Güte der Anpassung der Weibullverteilungsfunktion an die empirische Verteilungsfunktion dadurch beeinträchtigt wird, dass sie jeweils über einen unterschiedlichen Wertebereich hinweg positive Häufigkeiten enthalten.

Um die Weibullverteilungsfunktion auf den Wertebereich der Beobachtungen zu begrenzen, wird sie im Folgenden *gestutzt*, das heißt, Werte außerhalb dieses Wertebereichs werden vernachlässigt. Allgemein lässt sich dies wie folgt formulieren [35]:

Gegeben sei eine Zufallsvariable X mit der Verteilungsfunktion F(x) und $a \leq x \leq b$. Betrachtet man nur die Werte einer Zufallsvariablen, die im Wertebereich $\gamma \leq x \leq \delta$ liegen, erhält man eine *gestutzte Verteilung*. Die ganze Masse der Wahrscheinlichkeit muss jetzt im Wertebereich $[\gamma; \delta]$ konzentriert werden. Deswegen ist eine Normierung auf diesen Bereich notwendig, die durch Formel 11 realisiert wird.

Die Verteilungsfunktion der gestutzten Verteilung F*(x) mit $\gamma \leq x \leq \delta$ und $\gamma > a$ sowie $\delta < b$ ist durch

$$F^*(x) = \frac{F(x) - F(\gamma)}{F(\delta) - F(\gamma)} \quad \gamma \leq x \leq \delta \tag{11}$$

definiert, wobei F die Verteilungsfunktion von X ist. Im vorliegenden Fall wurde die Verteilungsfunktion der Weibullverteilung auf den Wertebereich [0; 1327] gestutzt bzw. beim Wert

von 1327 „abgeschnitten" (siehe Gleichung (12)), weil die maximale Anzahl Tage seit dem letzten Zugriff 1327 beträgt.

$$F^*(x) = \frac{F(x) - F(\gamma)}{F(\delta) - F(\gamma)} = \frac{1 - e^{-(x/\beta)^\alpha}}{1 - e^{-(1327/\beta)^\alpha}} \quad \gamma \le x \le 1327 \tag{12}$$

Die optimierten Parameter lauten $\hat{\alpha} = 0{,}30$ und $\hat{\beta} = 6{,}67$.

Nun wird der χ^2-Anpassungstest durchgeführt. Das Signifikanzniveau beträgt wieder 0,001 und die Hypothesen entsprechen bis auf die Werte der Parameter denen des vorherigen Abschnittes. Für die Testgröße ergibt sich ein Wert von T = 264,07.

Es gilt der Ablehnungsbereich

$T > \chi^2_{47;0,001} \Leftrightarrow 264{,}07 > 82{,}72.$

Das heißt, H_0 muss auch im Falle der gestutzten Verteilungsfunktion verworfen werden. Es wurde jedoch eine weitere Verbesserung der Testgröße erzielt: Sie konnte von 307,89 auf 264,07 gesenkt werden.

Trotz den gegenüber der Exponentialverteilung besseren Testergebnissen muss man feststellen, dass die Zufallsvariable X nicht einer Weibullverteilung entstammt. Es ist also die Untersuchung weiterer Verteilungsannahmen notwendig, wobei in allen nachfolgenden Abschnitten immer die gestutzte Variante der jeweiligen Verteilung betrachtet wird.

5.9.5 Annahme einer Gammaverteilung

In diesem Abschnitt wird die Gammaverteilung, die eine weitere Lebensdauer- bzw. Wartezeitverteilung mit zwei Parametern ist, vorgestellt und bezüglich der Zufallsvariablen X: „Anzahl der Tage seit dem letztem Zugriff" getestet.

5.9.5.1 Allgemeines zur Gammaverteilung

Eine stetige Zufallsvariable X ist *gammaverteilt* mit den Parametern $\alpha > 0$ und $\beta > 0$, wenn ihre Dichtefunktion durch

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & x \ge 0 \\ 0 & x < 0 \end{cases} \tag{13}$$

gegeben ist. Das Integral $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ heißt Gammafunktion mit dem Parameter α. Die Verteilungsfunktion der Gammaverteilung lautet

$$F(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^x u^{\alpha-1} e^{-u/\beta} du & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{14}$$

Die Verteilungsfunktion der gestutzten Gammaverteilung erhält man analog zu Abschnitt 5.10.3.3:

$$F^*(x) = \frac{F(x) - F(0)}{F(1327) - F(0)} = \frac{\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^x u^{\alpha-1} e^{-u/\beta} du}{\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{1327} u^{\alpha-1} e^{-u/\beta} du} \qquad \alpha \leq x \leq \beta \tag{15}$$

Schätzungen für die Parametern $\alpha > 0$ und $\beta > 0$ der (nicht gestutzten) Gammaverteilung erhält man mit Hilfe des arithmetischen Mittels $\bar{x}$ und der Varianz[18] s^2 der Beobachtungen $x_{(1)},\ldots,x_{(n)}$ [86].

$$\hat{\beta} = \frac{s^2}{\bar{x}} \tag{16}$$

$$\hat{\alpha} = \frac{\bar{x}}{\hat{\beta}} \tag{17}$$

Mit der Schätzung der Parameter beginnt in Abschnitt 0 die Überprüfung der Verteilungsannahme, doch zunächst wird im folgenden Abschnitt auf die mit der Gammaverteilung verwandte *Erlangverteilung* hingewiesen.

5.9.5.2 Zusammenhang von Gammaverteilung und Erlangverteilung

Eine Zufallsvariable X ist *erlangverteilt* mit den Parametern $\alpha \in \mathbb{N}$ und $\beta > 0$, wenn ihre Dichtefunktion durch

$$f(x) = \begin{cases} \frac{1}{(\alpha-1)!\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{18}$$

gegeben ist. Die Verteilungsfunktion der Erlangverteilung ist

[18] 7Die Varianz ist definiert als $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$F(x) = \begin{cases} 1 - e^{-x/\beta} \sum_{i=0}^{\alpha} \frac{(x/\beta)^i}{i!} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{19}$$

Der Zusammenhang zwischen Gamma- und Erlangverteilung lässt sich durch Gleichsetzen der Dichtefunktionen herleiten:

$$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \overset{!}{=} \frac{1}{(\alpha-1)!\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

$$\Gamma(\alpha) \overset{!}{=} (\alpha-1)! \tag{20}$$

Die Gleichung (20) ist erfüllt, wenn $\alpha \in \mathbb{N}$ [35]. Demnach entsteht die Erlangverteilung aus einer Gammaverteilung mit den Parametern $\alpha \in \mathbb{N}$ und $\beta > 0$. Mit anderen Worten: Beschränkt man bei der Gammaverteilung den Parameter α auf natürliche Zahlen, so entspricht diese der Erlangverteilung. Aufgrund dieses Zusammenhangs wird die Erlangverteilung in dieser Arbeit nicht gesondert untersucht, sondern nur die allgemeinere Gammaverteilung.

5.9.5.3 Überprüfung der Annahme eine Gammaverteilung

Die Schätzwerte für die Parameter der Gammaverteilung betragen $\hat{\alpha} = 0{,}14$ und $\hat{\beta} = 268$. In Abbildung 18 sieht man, dass Gammaverteilungsfunktion und empirische Verteilungsfunktion nur wenig voneinander abweichen. Jedoch weist der Q-Q-Plot ähnlich wie im Falle der Weibullverteilung viele Punkte auf, die nicht auf der Geraden liegen (siehe Abbildung 19).

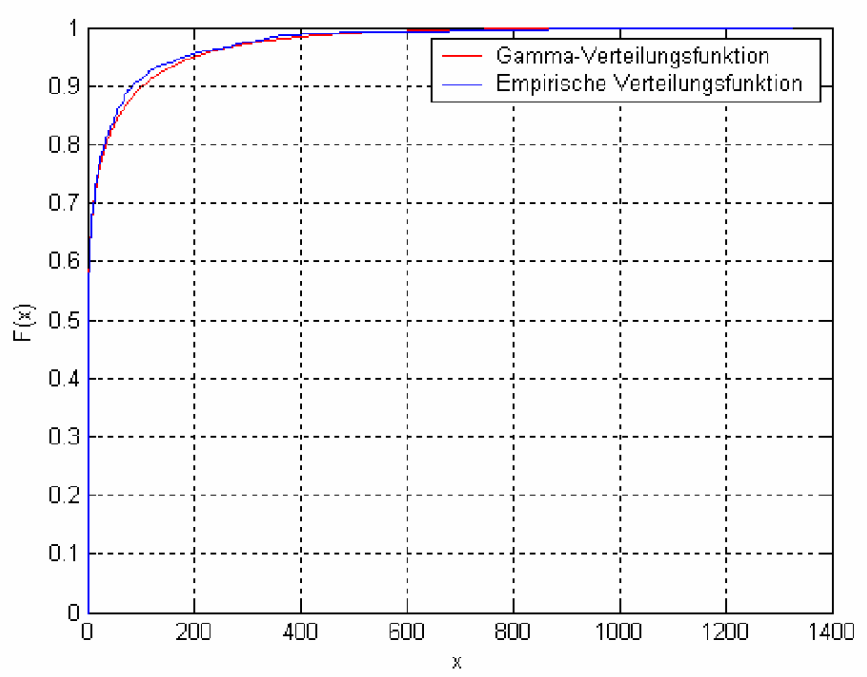


Abbildung 18: Gamma-Verteilungsfunktion mit $\hat{\alpha} = 0,14$, $\hat{\beta} = 268$ und empirische Verteilungsfunktion

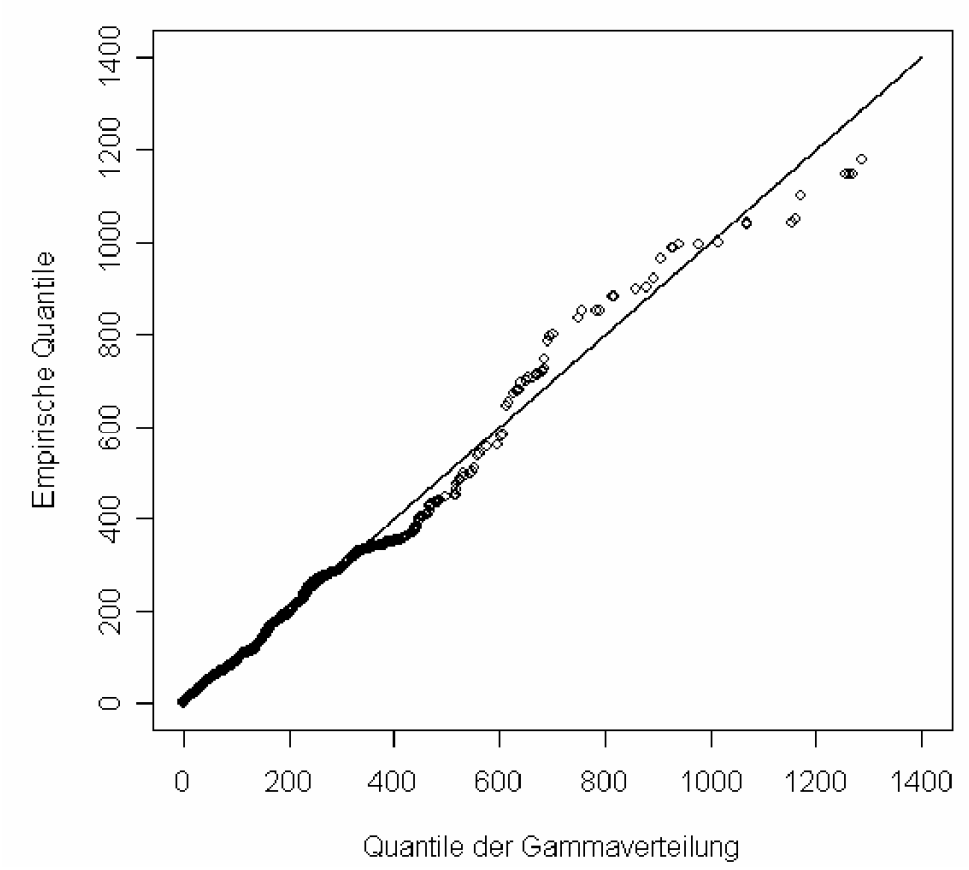


Abbildung 19: Q-Q-Plot der Gammaverteilung mit $\hat{\alpha}$ = 0, 14, $\hat{\beta}$ = 268 und der empirischen Verteilung

Mit dem χ^2-Anpassungstest soll nun die Annahme einer Gammaverteilung rechnerisch überprüft werden: Es wird die Hypothese

H_0: Die Zufallsvariable X entstammt einer Ga(0,14; 268)-Verteilung

gegen die Alternativhypothese

H_1: Die Zufallsvariable X entstammt nicht einer Ga(0,14; 268)-Verteilung

zum Signifikanzniveau 0,001 getestet. Für die Testgröße T erhält man einen Wert von 203,73. Der Ablehnungsbereich stellt sich wie folgt dar:

$T > \chi^2_{47;0,001} \Leftrightarrow 203, 73 > 82, 72$

Das heißt, H_0 muss abgelehnt werden.

Anhand der Testgrößen im Fall der Weibullverteilung und im Fall der Gammaverteilung kann man feststellen, dass mit diesen beiden Verteilungen die Anpassung an die empirische Verteilungsfunktion zwar wesentlich besser gelingt als mit der Exponentialverteilung, aber trotzdem die jeweilige Nullhypothese verworfen werden muss. Es ist demnach die Anwendung weitergehender statistischer Methoden erforderlich.

5.10 Gemischte Verteilungsfunktionen

Alle im vorherigen Kapitel aufgestellten Verteilungsannahmen der Zufallsvariablen X: „Anzahl Tage seit dem letzten Zugriff" und der entsprechenden transformierten Zufallsvariable mussten verworfen werden. Anhand der Größe der Testfunktion des χ^2-Anpassungstests konnte jedoch festgestellt werden, dass sich die gestutzte Weibullverteilung und die gestutzte Gammaverteilung am besten zur Modellierung der vorliegenden Daten eignen, auch wenn kein positives Testergebnis erzielt werden konnte.

In diesem Kapitel werden daher die Verteilungsfunktionen der genannten Verteilungen weiter optimiert, indem gemischte Verteilungsfunktionen eingeführt werden. Unter einer gemischten Verteilung versteht man die Summe von mehreren mit Gewichten versehenen Verteilungsfunktionen [76].

Im Folgenden werden zunächst gemischte Verteilungsfunktionen für die gesamte Stichprobe konstruiert und anschließend für einzelne Untergruppen der Stichprobe.

5.10.1 Allgemeines

Von beliebigen Verteilungsfunktionen $F_k(x)$ $(k = 1, 2, ...)$ kann eine mit den Gewichten p_k versehene Mischung angegeben werden [76]:

$$F(x) = \sum_{k=1}^{\infty} p_k F_k(x) \tag{21}$$

$F(x)$ ist dann ebenfalls eine Verteilungsfunktion.

Die Gewichte p_k sind nichtnegative Zahlen und es gilt:

$$\sum_{k=1}^{\infty} p_k = 1 \tag{22}$$

Die einzelnen Gewichte entsprechen dem jeweiligen Anteil der Beobachtungen.

Zur Veranschaulichung wird nun der Fall der Mischung einer diskreten und einer stetigen Verteilungsfunktion anhand eines Beispiels betrachtet [65]:

Es soll die Verteilungsfunktion der Wartezeiten W an einer Verkehrsampel bestimmt werden. Es sei p die Wahrscheinlichkeit, dass ein Fahrer freie Durchfahrt hat. $F_{W>0}$ sei die Verteilungsfunktion der Wartezeit, wenn man warten muss. Es gilt dann:

$$\begin{aligned} P(W = 0) &= p \\ P(W \le x \mid W > 0) &= F_{W>0}(x) \end{aligned} \tag{23}$$

$P(W \le x \mid W > 0)$ ist die Bezeichnung für die bedingte Wahrscheinlichkeit, dass $W \le x$ unter der Voraussetzung, dass $W > 0$.

Insgesamt ergibt sich aus der Verknüpfung der beiden Formeln:

$$F(x) = p \cdot 1_{[0,\infty)}(x) + (1-p)\, F_{W>0}(x) \tag{24}$$

Dabei ist $1_{[0,\infty)}(x)$ die Sprungfunktion mit der Sprungstelle bei x = 0. Sie ist wie folgt definiert:

$$1_{[0,\infty)}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{25}$$

Im oben genannten Beispiel hat die gemischte Verteilungsfunktion im Falle keiner Wartezeit *(W = 0)* den Wert *p*. Nur wenn gewartet werden muss *(W > 0)*, ist die Wartezeit gemäß der stetigen Verteilungsfunktion $F_{W>0}$ verteilt. Man sieht, dass eine gemischte Verteilungsfunktion ein geeignetes Mittel sein kann, um „Sonderereignisse", wie z.B. eine Wartezeit der Dauer Null, in einem ansonsten stetigen Verteilungsmodell zu berücksichtigen.

Die Methode des Mischens von Verteilungsfunktionen wird im Folgenden auf die zu untersuchenden Daten angewandt.

5.10.2 Konstruktion einer gemischten Verteilungsfunktion

Die empirische Dichtefunktion besitzt eine Eigenschaft, die die Anpassung eines klassischen stetigen Verteilungsmodells erschwert und dadurch schlechte Testergebnisse verursacht: Von den 6911 untersuchten Zugriffen beträgt in 1079 Fällen die Anzahl der Tage seit dem letzten Zugriff Null. Dies führt dazu, dass die Dichtefunktion beim Übergang von $x = 0$ (Null Tage) zum nächst höheren Wert $x = 6,94 \cdot 10^{-4}$ ($6,94 \cdot 10^{-4}$ Tage entsprechen einer Minute) einen großen Sprung aufweist, der mit einem stetigen Verteilungsmodell nur unzureichend modelliert werden kann. Eine bessere Anpassung soll nachfolgend mit Hilfe der Mischung einer diskreten und einer stetigen Verteilungsfunktion realisiert werden.

Zur Lösung des beschriebenen Problems wird nun analog zum Beispiel der Wartezeiten an einer Verkehrsampel in Abschnitt 5.10.1 eine Mischung einer diskreten und einer stetigen Verteilungsfunktion aufgestellt und anschließend getestet.

Die Wahrscheinlichkeit *p*, dass die Wartezeit *W* bzw. die Anzahl Tage seit dem letzten Zugriff Null beträgt, berechnet sich zu $p = {}^{1079}\!/_{6911}$. $F_{W>0}(x)$ ist die Verteilungsfunktion der Zufallsvariablen X: „Anzahl Tage seit dem letzten Zugriff", allerdings unter der Voraussetzung, dass $W > 0$. Mit anderen Worten: $F_{W>0}(x)$ ist die Verteilungsfunktion der 6911−1079 = 5832 Beobachtungen von mehr als Null Tagen seit dem letzten Zugriff. Damit ergibt sich folgende gemischte Verteilungsfunktion:

$$F(x) = \frac{1079}{6911} \cdot 1_{[0,1327)}(x) + \frac{5832}{6911} \cdot F_{W>0}(x) \tag{26}$$

Im Folgenden wird unter Verwendung der Ergebnisse aus Kapitel 5.9.1 eine geeignete Verteilungsfunktion $F_{W>0}(x)$ ermittelt. In Kapitel 5.9.1 wurden mit der gestutzten Weibullverteilung und der gestutzten Gammaverteilung die besten Resultate erzielt. Aus diesem Grund bilden diese beiden Verteilungen die Basis für die nachfolgenden statistischen Analysen.

5.10.3 Überprüfung auf Weibullverteilung

In diesem Abschnitt wird die Mischung einer diskreten und einer stetigen Verteilungsfunktion getestet, wobei für die stetige Verteilungsfunktion eine gestutzte Weibullverteilung angenommen wird. In Anlehnung an Abschnitt 5.10.1 gelten folgende Bezeichnungen:

- $F_{W>0}(x)$: Verteilungsfunktion der Wartezeiten größer Null (hier: Weibullverteilungsfunktion)
- $F^*_{W>0}(x)$: Gestutzte Variante von $F_{W>0}(x)$
- $F(x)$: Verteilungsfunktion der Gesamtwartezeit

Die Weibullverteilung wird auf den Wertebereich [6, $94 \cdot 10^{-4}$; 1327] gestutzt (siehe Gleichung 27), weil die minimale Anzahl Tage seit dem letzten Zugriff $6,94 \cdot 10^{-4}$ beträgt und die maximale Anzahl 1327. Die Gleichung 24 präsentiert sich nun wie folgt:

$$F(x) = \frac{1079}{6911} \cdot 1_{[0,1327)}(x) + \frac{5832}{6911} \cdot F^*_{W>0}(x)$$

mit

$$F^*_{W>0}(x) = \frac{F_{W>0}(x) - F_{W>0}(6{,}94 \cdot 10^{-4})}{F_{W>0}(1327) - F_{W>0}(6{,}94 \cdot 10^{-4})} \qquad 6{,}94 \cdot 10^{-4} \le x \le 1327 \tag{27}$$

Die optimierten Parameter der gestutzten Weibullverteilung betragen $\hat{\alpha} = 0{,}33$ und $\hat{\beta} = 9{,}90$.

Nun wird der χ^2-Anpassungstest durchgeführt. Es wird die Hypothese

> H_0: Die Zufallsvariable X entstammt einer Grundgesamtheit mit der gemischten Verteilung gemäß Gleichung 27 mit der gestutzten Weibullverteilungsfunktion $F^*_{W>0}(x)$

zum Signifikanzniveau 0,001 getestet. Für die Testgröße T erhält man einen Wert von 236,29. Der Ablehnungsbereich lautet:

$$T > \chi^2_{48;0,001} \Leftrightarrow 236,29 > 84,04$$

Das heißt, H_0 muss abgelehnt werden. Die Testgröße konnte im Vergleich zur gestutzten Weibullverteilung (siehe Abschnitt 5.9.4.3) ein wenig gesenkt werden, nämlich von 264,07 auf 236,29, aber trotzdem wurde auch im Falle der gemischten Verteilungsfunktion, bestehend aus Sprungfunktion und gestutzter Weibullverteilung, kein positives Testergebnis erzielt.

5.10.4 Überprüfung auf Gammaverteilung

Analog zum vorherigen Abschnitt wird hier eine gemischte Verteilungsfunktion getestet, allerdings wird in diesem Fall für die stetige Verteilungsfunktion eine gestutzte Gammaverteilung angenommen.

Die Gammaverteilung wird auf den Wertebereich [6, $94 \cdot 10^{-4}$; 1327] gestutzt (siehe Gleichung 27). Die zu testende gemischte Verteilungsfunktion entspricht der Gleichung 27 des

vorherigen Abschnittes mit der Bezeichnung $F^*_{W>0}(x)$ für die Verteilungsfunktion der gestutzten Gammaverteilung. Die optimierten Parameter der gestutzten Gammaverteilung lauten $\hat{\alpha} = 0{,}14$ und $\hat{\beta} = 260{,}0$.

Die Nullhypothese des χ^2-Anpassungstests lautet wieder:

> H_0: Die Zufallsvariable X entstammt einer Grundgesamtheit mit der gemischten Verteilung gemäß Gleichung 27 mit der gestutzten Gammaverteilungsfunktion $F^*_{W>0}(x)$

Es wird zum Signifikanzniveau 0,001 getestet. Für die Testgröße T erhält man einen Wert von 200,58. Der Ablehnungsbereich lautet:

$$T > \chi^2_{45;0,001} \Leftrightarrow 200{,}58 > 80{,}08$$

H_0 muss demnach abgelehnt werden.

Das Ergebnis dieses Abschnittes ist, dass mit einer gemischten Verteilungsfunktion, bestehend aus Sprungfunktion und gestutzter Weibull- bzw. Gammaverteilung, keine positiven Testergebnisse erreicht wurden. Für die gesamte Stichprobe kann demnach keine Verteilungsfunktion angegeben werden, deshalb werden nachfolgend Verteilungsannahmen für Teilstichproben überprüft.

5.10.5 Betrachtung von Zufallsstichproben aus der Gesamtstichprobe

In diesem Abschnitt werden sechs Zufallsstichproben zu je 1000 Beobachtungen aus der Gesamtstichprobe von 5832 Beobachtungen mit Hilfe der entsprechenden Funktion in Microsoft Excel® entnommen. Da mit der Mischung einer diskreten und einer stetigen Verteilungsfunktion die besten Testergebnisse erzielt wurden, wird auch hier ein stetiges Verteilungsmodell nur für die Beobachtungen mit mehr als Null Tagen seit dem letzten Zugriff gesucht.

Ziel dieser Stichprobenentnahme ist es festzustellen, ob das Zugriffsverhalten einiger Dateien so von den übrigen Dateien abweicht, dass dadurch positive Testergebnisse verhindert werden. Ist dies der Fall, werden die Testergebnisse zufällig zusammengestellter Teilmengen der gesamten Stichprobe im Durchschnitt besser ausfallen als die bisherigen Testergebnisse.

5.10.5.1 Überprüfung auf Weibullverteilung

An die sechs Zufallsstichproben zu je 1000 Beobachtungen wird im Folgenden jeweils die Mischung einer diskreten und einer stetigen Verteilungsfunktion angepasst, wobei als stetige Verteilung eine gestutzte Weibullverteilung angenommen wird. Dann wird ein χ^2-Anpassungstest auf die Nullhypothese

> H_0: Die Zufallsvariable X entstammt einer Grundgesamtheit mit der gemischten Verteilung gemäß Gleichung 27 mit der gestutzten Weibullverteilungsfunktion $F^*_{W>0}(x)$

zum Signifikanzniveau 0,001 durchgeführt. Die Schätzungen der Parameter, die Ablehnungsbereiche und die Testergebnisse für die sechs Teilstichproben sind Tabelle 15 zu entnehmen.

Stichprobe Nr.	$\hat{\alpha}$	$\hat{\beta}$	Ablehnungsbereich	Ergebnis: H_0…
1	0,34	10,8	$T > \chi^2_{28;0,001} \Leftrightarrow 49,49 < 56,89$	nicht verwerfen
2	0,31	10,6	$T > \chi^2_{30;0,001} \Leftrightarrow 57,85 < 59,70$	nicht verwerfen
3	0,35	10,3	$T > \chi^2_{28;0,001} \Leftrightarrow 48,03 < 56,89$	nicht verwerfen
4	0,31	10,1	$T > \chi^2_{30;0,001} \Leftrightarrow 65,71 > 59,70$	verwerfen
5	0,33	9,8	$T > \chi^2_{27;0,001} \Leftrightarrow 46,02 < 55,48$	nicht verwerfen
6	0,34	11,8	$T > \chi^2_{29;0,001} \Leftrightarrow 57,25 < 58,30$	nicht verwerfen

Tabelle 15: X^2-Anpassungstests der Zufallsstichproben auf eine gemischte Verteilungsfunktion mit gestutzter Weibullverteilung

Die Testergebnisse in Tabelle 15 überraschen, denn während der χ^2-Anpassungstest der gesamten Stichprobe negativ ausfällt, muss hier in nur einem von sechs Fällen die Nullhypothese verworfen werden. An diesem Ergebnis erkennt man, dass einzelne Beobachtungen in der Stichprobe so stark von den übrigen abweichen, dass dadurch der χ^2-Anpassungstest negativ ausfällt, obwohl für die Mehrheit der Beobachtungen die getroffene Verteilungsannahme zutrifft.

5.10.5.2 Überprüfung auf Gammaverteilung

Analog zum vorherigen Abschnitt werden nun die sechs Zufallsstichproben zu je 1000 Beobachtungen mit dem χ^2-Anpassungstest getestet, wobei als stetige Verteilung eine gestutzte Gammaverteilung angenommen wird. Die Nullhypothese lautet dementsprechend:

> H_0: Die Zufallsvariable X entstammt einer Grundgesamtheit mit der gemischten Verteilung gemäß Gleichung 27 mit der gestutzten Gammaverteilungsfunktion $F^*_{W>0}(x)$

Das Signifikanzniveau ist 0,001. Die Schätzungen der Parameter, die Ablehnungsbereiche und die Testergebnisse für die sechs Teilstichproben sind Tabelle 16 zu entnehmen.

Stichprobe Nr.	$\hat{\alpha}$	$\hat{\beta}$	Ablehnungsbereich	Ergebnis: H_0…
1	0,15	225	$T > \chi^2_{27;0,001} \Leftrightarrow 34,83 < 55,48$	nicht verwerfen
2	0,13	295	$T > \chi^2_{29;0,001} \Leftrightarrow 48,11 < 58,30$	nicht verwerfen
3	0,15	213	$T > \chi^2_{27;0,001} \Leftrightarrow 33,66 < 55,48$	nicht verwerfen
4	0,13	273	$T > \chi^2_{29;0,001} \Leftrightarrow 42,97 < 58,30$	nicht verwerfen
5	0,14	214	$T > \chi^2_{27;0,001} \Leftrightarrow 31,61 < 55,48$	nicht verwerfen
6	0,15	248	$T > \chi^2_{29;0,001} \Leftrightarrow 42,52 < 58,30$	nicht verwerfen

Tabelle 16: χ^2-Anpassungstests der Zufallsstichproben auf eine gemischte Verteilungsfunktion mit gestutzter Gammaverteilung

Im Unterschied zum vorherigen Abschnitt muss hier sogar in keinem der sechs Fälle die Nullhypothese verworfen werden. Die Testergebnisse der Zufallsstichproben zeigen, dass die Annahme einer Weibull bzw. einer Gammaverteilung grundsätzlich sinnvoll ist, auch wenn der χ^2-Anpassungstest der vorliegenden Gesamtstichprobe negativ ausfällt. Es wäre demnach fahrlässig, bei der Untersuchung der Zeiten seit dem letzten Zugriff einer Datenbank nicht Weibull- bzw. Gammaverteilungsannahmen zu überprüfen. Motiviert durch diese Erkenntnis

sollen im folgenden Abschnitt bestimmte Untergruppen an Stelle einer zufälligen Teilstichprobe untersucht werden.

5.11 Aufteilung der Stichprobe nach Dateitypen

Im Folgenden wird die nach Dateitypen[19] aufgeteilte Stichprobe entsprechend der Vorgehensweise der vorangegangenen Abschnitte untersucht. Die Teilmengen sind im einzelnen 1358 Beobachtungen der doc-Dateien, 2645 Beobachtungen der xls- Dateien, 1323 Beobachtungen der ppt-Dateien, 857 Beobachtungen der pdf-Dateien und 728 Beobachtungen „sonstiger" Dateitypen[20].

5.11.1 Überprüfung auf Weibullverteilung

Zunächst wird die Annahme einer gemischten Verteilungsfunktion mit einer gestutzten Weibullverteilung als stetige Verteilung überprüft. Mit dem χ^2-Anpassungstest wird folgende Nullhypothese zum Signifikanzniveau 0,001 getestet:

> H_0: Die Zufallsvariable X entstammt einer Grundgesamtheit mit der gemischten Verteilung gemäß Gleichung 27 mit der gestutzten Weibullverteilungsfunktion $F^*_{W>0}(x)$

Im Falle der Beobachtungen der xls-Dateien, der ppt-Dateien und der „sonstigen" Dateien wurde die Nullhypothese nicht verworfen (siehe Tabelle 17). Für doc- und pdf-Dateien musste sie dagegen verworfen werden.

Dateityp	$\hat{\alpha}$	$\hat{\beta}$	Ablehnungsbereich	Ergebnis: H_0…
doc	0,38	23,6	$T > \chi^2_{33;0,001} \Leftrightarrow 91,85 > 63,87$	verwerfen
xls	0,25	1,1	$T > \chi^2_{29;0,001} \Leftrightarrow 49,59 < 58,30$	nicht verwerfen
ppt	0,38	14,3	$T > \chi^2_{31;0,001} \Leftrightarrow 39,69 < 61,10$	nicht verwerfen
pdf	0,48	21,9	$T > \chi^2_{28;0,001} \Leftrightarrow 79,59 > 56,89$	verwerfen
sonstige	0,46	27,7	$T > \chi^2_{30;0,001} \Leftrightarrow 38,83 < 59,70$	nicht verwerfen

Tabelle 17: χ^2-Anpassungstests der nach Dateitypen aufgeteilten Stichprobe auf eine gemischte Verteilungsfunktion mit gestutzter Weibullverteilung

Die xls-Dateien, ppt-Dateien und „sonstige" Dateien machen zusammen 53,89 % aller Dateien der Stichprobe aus. Für diese Dateitypen ist die gemischte Verteilungsfunktion, bestehend aus Sprungfunktion und gestutzter Weibullverteilungsfunktion, ein geeignetes Verteilungsmodell. Diese Erkenntnisse können unmittelbar für die Erstellung einer Migrationsregel im Rahmen von ILM verwendet werden.

---

[19] Die verschiedenen Dateitypen wurden in Abschnitt 5.5 vorgestellt.

[20] Da die Stichprobe von den nicht genannten Dateitypen nur wenige enthält, wurden sie unter „sonstige" zusammengefasst.

5.11.2 Überprüfung auf Gammaverteilung

In diesem Abschnitt wird die Annahme einer gemischten Verteilungsfunktion mit der gestutzten Gammaverteilung als stetige Verteilung überprüft. Mit dem χ^2-Anpassungstest wird die Nullhypothese

> H_0: Die Zufallsvariable X entstammt einer Grundgesamtheit mit der gemischten Verteilung gemäß Gleichung 27 mit der gestutzten Gammaverteilungsfunktion $F^*_{W>0}(x)$

zum Signifikanzniveau 0,001 getestet.

Die Tabelle 18 zeigt, dass für ppt-Dateien und „sonstige" Dateiarten das angenommene Verteilungsmodell mit der gestutzten Gammaverteilung zur Modellierung der Zufallsvariablen X: „Anzahl Tage seit dem letzten Zugriff" geeignet ist. Dagegen muss für die anderen drei Dateitypen die Nullhypothese verworfen werden.

Dateityp	$\hat{\alpha}$	$\hat{\beta}$	Ablehnungsbereich	Ergebnis: H_0…
doc	0,21	279	$T > \chi^2_{32;0,001} \Leftrightarrow 71,31 > 62,49$	verwerfen
xls	0,10	141	$T > \chi^2_{25;0,001} \Leftrightarrow 172,59 > 52,62$	verwerfen
ppt	0,19	221	$T > \chi^2_{30;0,001} \Leftrightarrow 57,36 < 59,70$	nicht verwerfen
pdf	0,29	146	$T > \chi^2_{28;0,001} \Leftrightarrow 67,46 > 56,89$	verwerfen
sonstige	0,29	181	$T > \chi^2_{29;0,001} \Leftrightarrow 27,61 < 49,59$	nicht verwerfen

Tabelle 18: χ^2-Anpassungstests der nach Dateitypen aufgeteilten Stichprobe auf eine gemischte Verteilungsfunktion mit gestutzter Gammaverteilung

Das Ergebnis des Abschnittes 5.11 ist, dass für die in der Stichprobe enthaltenen xls-Dateien, ppt-Dateien und „sonstige" Dateien ein geeignetes Verteilungsmodell konstruiert werden kann: Bei der Erstellung einer Migrationsregel für ILM sollte im Falle von xls-Dateien eine gemischte Verteilungsfunktion mit einer gestutzten Weibullverteilung benutzt werden, im Falle von ppt-Dateien und „sonstigen" Dateitypen stehen Weibull- und Gammaverteilung (jeweils gestutzt) zur Wahl.

Wenn eine Regel für doc- und pdf-Dateien aufgestellt werden muss, sollte die gemischte Verteilungsfunktion mit einer gestutzten Gammaverteilung ausgewählt werden. Trotz negativer Testergebnisse eignet sich für diese Fälle die Gammaverteilung besser als die Weibullverteilung, da erstere niedrigere Testgrößen beim χ^2-Anpassungstest erzielte, was eine bessere Anpassung bedeutet.

5.12 Zusammenfassung der Testergebnisse

Nachdem eine geeignete Zufallsvariable definiert wurde, sollte eine zugehörige allgemeine Verteilungsfunktion hergeleitet werden. In Kapitel 5.9.1 wurde festgestellt, dass es kein passendes Verteilungsmodell für die gesamte Stichprobe gibt. Aus diesem Grund wurden in Kapitel 5.10 gemischte Verteilungsfunktionen eingeführt und die nach Dateitypen aufgeteilte Stichprobe auf ihre Verteilung hin untersucht. Analog wurde außerdem die nach Dateialter, nach Anzahl der Zugriffe, nach Zugriffsart und nach Dateigröße aufgeteilte Stichprobe unter-

sucht. Dabei ist es gelungen, für einige Untergruppen geeignete Verteilungsmodelle zu konstruieren. Die besten Testergebnisse ließen sich dabei mit der Aufteilung nach Dateitypen erzielen. So konnte für 81,4% der Dateien ein Verteilungsmodell mit positiven Testergebnissen konstruiert werden. Lediglich für doc-Dateien und pdf-Dateien mit jeweils weniger als 7 Zugriffen und einem Alter größer als 364 Tage wurde kein passendes Modell gefunden.

Die Tabelle 19 gibt einen Überblick über die Testergebnisse. Nicht aufgeführt sind in der Tabelle die Kriterien Zugriffsart und Dateigröße, da in diesen Fällen die χ^2-Anpassungstests negativ ausgefallen sind.

Kriterium	Klasse	Verteilungsmodell
Dateialter	[0 Tage; 365 Tage)	W
	[365 Tage; 730Tage)	-
	[730 Tage; 1772 Tage)	-
Anzahl Zugriffe	[1 Zugriff; 7 Zugriffe)	-
	[7 Zugriffe; 15 Zugriffe)	G
	[15 Zugriffe; 292 Zugriffe)	W
Dateityp	doc	-
	xls	W
	ppt	W,G
	pdf	-
	sonstige	W,G

Tabelle 19: Zusammenfassung der Testergebnisse. W = Weibullverteilung, G = Gammaverteilung, „-" = Kein geeignetes Verteilungsmodell

Es liegt nun eine Methode vor, die die Bewertung von Dateien in Form von Prognosen über zukünftige Zugriffe ausgibt. Damit ist die dritte Forschungsfrage „Wie können Dateien auf Basis von automatisierten Prognosen bewertet werden?" beantwortet. Somit lassen sich Dateien bewerten und anschließend klassifizieren. Im Vorgehensmodell ist die Phase der Klassifizierung damit abgeschlossen. Ob diese Methode für die ILM-Phase 4 „Automatisierung" verwendet werden kann, wird im nächsten Kapitel untersucht.

6 Simulationsmodell für die ILM-Automatisierung

Dieses Kapitel beschreibt den vierten ILM-Prozessschritt „Automatisierung“. Nach Herleitung der Methode der Wahrscheinlichkeit zukünftiger Zugriffe, gilt es nun zu prüfen, ob diese für ILM verwendbar ist. ILM benötigt Bewertungsmethoden, die permanent Bewertungen von mehreren tausend Dateien durchführen können. Um dies zu überprüfen, wird eine Simulationsumgebung implementiert. In diesem Kapitel werden zunächst die Grundlagen des Simulationswerkzeuges erläutert. Zu Beginn werden simulationsvorbereitende Begriffsdefinitionen von ILM geliefert und anschließend die Simulationsziele ausführlich dargelegt sowie notwendige Annahmen und Vereinfachungen der Simulation erläutert. Zusätzlich werden anschließend das Simulationsmodell und die Struktur des Simulators beschrieben. Das Kapitel endet mit einem Proof of Concept und der Beantwortung der vierten Forschungsfrage: „Ist eine derartige (prognostizierende) Methode in ILM verwendbar?“.

6.1 Begriffsdefinitionen

Die SNIA beschreibt mit ihrer Definition den Begriff ILM zutreffend und umfassend (siehe Abschnitt 2.1). Für Simulationen ist sie jedoch nicht konkret genug, so dass weitere Festlegungen notwendig sind, die zu einer für Simulationszwecke nutzbaren Begriffsbildung führen.

Als Informationen betrachtet diese Arbeit konkret Microsoft-Office-Dateien, weil nur für diese die Methode spezifiziert wurde. Dennoch sprechen die nachfolgenden Definitionen allgemeiner von „Informationen“ bzw. „Informationsobjekten“. Das heißt, Definition 2 (Information) (siehe Abschnitt 2.1) findet in der Simulation unmittelbare Anwendung. Hierbei werden Informationen durch Dateien dargestellt. In einem anderen realen Umfeld können diese beispielsweise auch in Form von Datenbankeinträgen oder Datenobjekten, die zwischen Anwendungen oder Prozessen ausgetauscht werden, aufgefasst werden [68].

Die Vision von ILM basiert auf der Annahme, dass Informationen gemäß ihrem Wert aufbewahrt werden können. Hierzu muss der Wert verschiedener Informationen vergleichbar und messbar gemacht werden. Das ist nur möglich, wenn er eindeutig beschrieben werden kann.

Definition 3 (Wert einer Information): *Der Wert einer Information V(I) beschreibt die Relevanz einer Information für die Geschäftstätigkeit eines Unternehmens. V(I) lässt sich als wesentliche Eigenschaft einer Information I auffassen und ist für Zeitpunkte nach der Erstellung einer Information definiert ($t \geq t_0$). Dabei ist der Zeitpunkt der Erstellung bestimmt als $t_0 \geq 0$.*

Eine weitere Voraussetzung für die wertgerechte Speicherung von Informationen besteht darin, dass ihr Wert zweifelsfrei ermittelt wird. Hat man eine passende Methode der Wertzuweisung von Informationen zur Hand, kann der Wert einer Information eindeutig quantifiziert werden. Somit ist es auch möglich, Informationsobjekte mit ähnlicher Bewertung in Gruppen zusammenzufassen. Informationen werden klassifiziert und bilden so genannte Informationsklassen.

Definition 4 (Informationsklasse): *Eine Informationsklasse C ist eine Gruppierung aller Informationen $I_1,\ldots,I_n$ deren Wert $V(I_i(t))$ zum Zeitpunkt t in einem vordefinierten Wertebereich liegt.*

$$C := C_{i,j} := \left\{ I_i(t_j) \mid a \le V(I_i(t_j)) < b \,;\, a,b \in IR \right\} \tag{28}$$

Da der Informationswert dynamisch ist und sich im Laufe der Zeit ändert, sind die Informationsklassen ebenfalls dynamisch. Die Gruppierung der Informationsobjekte wird mit jeder Bewertung der Informationen neu zusammengestellt.

Ein zentrales Element einer ILM-Lösung sind Speichermedien. Sie dienen als Träger der Informationen und lassen sich ebenso zu Gruppen zusammenfassen, den Speicherebenen bzw. Speicherhierarchien.

Definition 5 (Speicherhierarchie): *Eine Speicherhierarchie S ist eine Gruppierung von Speichermedien mit ähnlichen Eigenschaften hinsichtlich Kosten und Dienstgütekriterien, insbesondere der Sicherheit, Sicherungshäufigkeit und Zugriffsgeschwindigkeit.*

Informationsklassen und Speicherhierarchien stehen in einer festen Zuordnung, d.h. Informationen, deren Wert V(I) in einem bestimmten Intervall liegen, werden einer definierten Speicherhierarchie zugeordnet. Ändert sich nun der Wert eines Informationsobjektes im Laufe der Zeit, so wird das Objekt gegebenenfalls migriert. Die zugehörige Definition von Migration lautet:

Definition 6 (Migration): *Unter Migration versteht man einen Prozess, bei dem eine Information I aufgrund einer Veränderung ihres Wertes V(I) auf eine andere Speicherhierarchie S verschoben wird. Gleichzeitig wird die Information Bestandteil einer anderen Informationsklasse C [4].*

In der Simulation wird der Lebenszyklus als dynamische Veränderung des Informationswertes betrachtet. Er ist folgendermaßen definiert:

Definition 7 (Lebenszyklus): *Der Lebenszyklus L einer Information bezeichnet die Abbildung des Informationswertes V(I(t)) während der Zeit t im Intervall $[t_1; t_2]$ mit $t_0 \le t_1 \le t_2$.*

$$L(I) := \left\{ V(I_i(t)) \mid t_1 \le t \le t_2 \right\} \tag{29}$$

Zusammenfassend ergibt sich für die Simulation die nachfolgende Definition von ILM:

Definition 8 (Information Lifecycle Management): *ILM bezeichnet die Abbildung von Informationen $I_1,\ldots,I_n$ auf Informationsklassen $C_1,\ldots,C_m$ in Abhängigkeit von ihrem Wert $V(I_1),\ldots,V(I_n)$ im Zeitintervall $[t_1; t_2]$.*

Wenn ILM entsprechend Definition 8 aufgefasst wird, kann ILM in mehrere Schritte zerlegt werden (siehe Abbildung 20):

1. Bestimmung des Informationswertes V(I).

2. Zuordnung einer Information zu einer Informationsklasse aufgrund des Wertes V(I).

3. Feste Abbildung der Informationsklasse auf eine Speicherhierarchie.

4. Speicherung der Information.

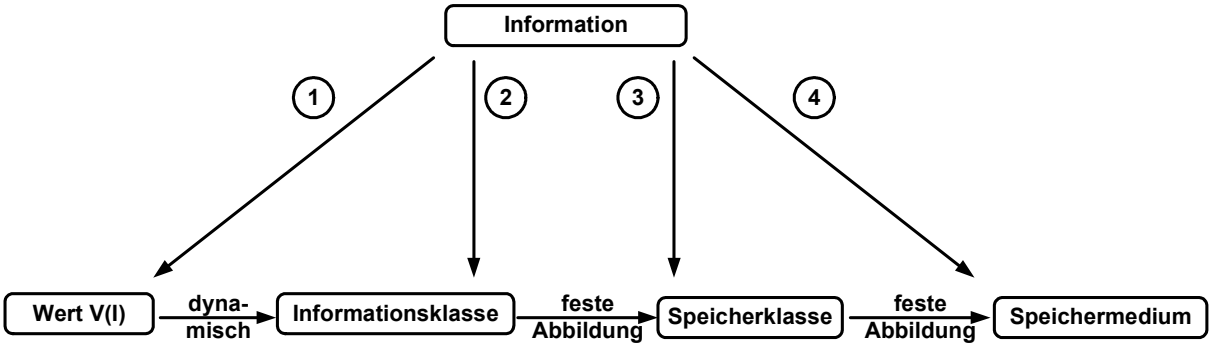


Abbildung 20: Simulationsframework

Das Simulationsframework in Abbildung 20 wird nachfolgend vertieft. Dazu werden zuvor die Ziele und Nebenziele der Simulation erörtert.

6.2 Ziele der Simulation

ILM verspricht erhebliche Kosteneinsparungen durch Hierarchisierung, weshalb rund 90% der in einer Studie befragten IT-Entscheider den Einsatz von ILM erwägen [49]. Nichtsdestotrotz existieren bislang zu wenige Erfahrungsberichte und die Forschung an realen ILM-Systemumgebungen ist sehr kostenaufwendig. Zusätzlich sagen laut einer weiteren Studie 66% der IT-Manager, sie hätten keine Zeit, ein rudimentäres Kosten- oder Datenbewertungsmodell für ILM-Projekte zu erstellen [25]. Diese Arbeit adressiert diese Umstände und trägt dazu bei, IT-Manager in ihrem Entscheidungsprozess über ILM zu unterstützen. Dazu soll ein Simulationstool entwickelt werden, welches die Untersuchung der nachstehend aufgeführten Hauptziele ermöglicht [4]:

- Betrachtung von ILM über den gesamten Lebenszyklus (ganzheitliche Betrachtung)
- Beobachtung von Kapazitätsanforderungen je Hierarchie
- Untersuchung verschiedener Migrationsregeln
- Untersuchung verschiedener Szenarien bei gleicher Migrationsregel
- Analyse des dynamischen Verhaltens von ILM-Szenarien
- Verwendung realer und künstlicher Dateienpools

Die Betrachtung von ILM über den gesamten Lebenszyklus (ganzheitliche Betrachtung) bedeutet, dass jeder Tag des gesamten definierten Lebenszyklus individuell dargestellt wird.

Die zwei Hauptbeweggründe von IT-Managern, im Storagebereich neue Konzepte einzusetzen, sind zum einen das Wachstum der Kapazitätsanforderung und zum anderen die Notwendigkeit, mögliche Kosteneinsparungspotenziale zu identifizieren [74]. Erkenntnisse über Kapazitätsanforderungen sind dazu unerlässlich. Aus diesem Grund ist die Betrachtung von Kapazitätsanforderungen für IT-Manager wichtig, um Entscheidungen zu treffen.

Neben der vorgestellten Methode der Wahrscheinlichkeit zukünftiger Zugriffe sollen auch andere Migrationsregeln untersucht werden können. Die Untersuchung verschiedener Migrationsregeln erlaubt IT-Managern, diese Methoden zu vergleichen und in ihren Entscheidungsfindungsprozess einzubeziehen.

Die Untersuchung verschiedener Szenarien bei gleicher Migrationsregel erlaubt die Feinkonzeption eines potentiellen ILM-Szenarios. Dazu gehören die Identifikation der passenden Anzahl von Speicherhierarchien sowie die Bestimmung der Parameterwerte der ausgewählten Migrationsstrategie.

Der Simulator soll grundsätzlich mit realen und auch künstlichen Daten bestückt werden können. Das erlaubt zum einen eine konkrete Betrachtung existierender Daten und deren Resultate in der ILM-Simulation. Zum anderen können mit künstlichen Daten transferierbare Aussagen erarbeitet werden, die grundsätzliche Erkenntnisse über ILM liefern, wie z.B. die langfristige Dynamik von ILM-Szenarien. Die zugehörigen Fragen lauten: „Ist ein ILM-Szenario stabil?“ und „Wann schwingt es ein, wenn es überhaupt einschwingt?“. Diese Fragestellungen sollen allgemeingültig und unabhängig von bestimmten technischen oder produktspezifischen Implementierungen betrachtet werden können, so dass die gewonnenen Erkenntnisse auf verschiedene Implementierungen abbildbar sind.

Nach den Hauptzielen werden nun die Nebenziele der Simulation erläutert.

6.3 Nebenziele der Simulationen

Als Nebenziele werden Aspekte bezeichnet, die nicht im direkten Fokus des Simulators liegen, jedoch inhaltlich nicht zu vernachlässigen sind. Sie bedürfen einer separaten Betrachtung in Abhängigkeit konkreter Situationen und Unternehmensumfelder. Nebenziele sind [4]:

- konkrete Kosten
- Verkehrsverhalten
- Dienstgütekriterien (QoS)
- Nutzerzufriedenheit

Das Simulationsmodell lässt konkrete Kosten außen vor. Die Ermittlung dieser Kosten hängt sehr stark von den individuellen Kostenfaktoren eines Unternehmens ab, wie zum Beispiel Miete, Strom oder die gewählte Technik [113]. Patel und Shah kommen zu der Feststellung, dass, wenn man mit dem Design und der Implementierung einer ILM-Lösung erfolgreich sein möchte, man bedenken sollte, dass nur rund 25% der Aufgabe sich mit der Auswahl von Pro-

dukten befasst [71]. Die in Kapitel 3 dargestellte Konzeption sieht die Trennung zwischen Konzept und technischer Realisierung vor, so dass der Versuch, konkrete Kosten zu ermitteln, konsequenterweise hier nicht unternommen wird. Dass dennoch nutzbringende Aussagen bezüglich Kosten, die ja jeder IT-Entscheidung zugrunde liegen, mit diesem Simulator möglich sind, wird Kapitel 7.3 verdeutlichen. Dort wird gezeigt, wie aufgrund von relativen Kostenbeziehungen zwischen den Speicherhierarchien eine Entscheidung hinsichtlich der passenden Migrationsstrategie getroffen werden kann.

Das Verkehrsverhalten beschreibt die Auswirkungen von ILM auf die IT-Infrastruktur einer Organisation, z.B. auf die verfügbare Bandbreite von Netzwerkkomponenten. Offensichtlich belegen ILM-bedingte Datentransfers Bandbreite, deren Betrachtung im Simulator jedoch außer Acht gelassen wird, weil dies kein Kriterium für ILM-Entscheidungen ist [74].

Für den Begriff *Dienstgüte* (engl. *Quality of Service*, Abk. *QoS*) existieren verschiedene Definitionen. Die Dienstgüte setzt sich aus der Betrachtung verschiedener Teilaspekte zusammen. In Abhängigkeit davon, welcher Dienst mit welchem Fokus betrachtet wird, werden die Teilaspekte ausgewählt. Das CCITT (Commité Consultatif Internationale Télégrafique et Téléfonique) veröffentlichte 1989 im Zusammenhang mit der ISDN-Technologie eine allgemeine Definition von Dienstgüte. Als Dienstgüte definiert das CCITT den “Gesamteffekt von Systemeigenschaften, die den Grad der Nutzerzufriedenheit eines Dienstes bestimmen” [10]. Hierbei wird unter „Nutzer“ nicht zwangsweise eine Person verstanden.

In dieser Arbeit wird die Grundintention der CCITT, dass Dienstgüte direkt die Nutzerzufriedenheit bestimmt, aufgenommen. Die genauen Aspekte der Dienstgüte im Kontext von ILM ergeben sich als technische Parameter der Diensterbringung. Diese sind z.B. Verfügbarkeit, Zugriffsgeschwindigkeit oder Wiederherstellungszeit. Diese technische Sichtweise entspricht der Dienstgüte in Netzwerken, wo z.B. die Übertragungsgeschwindigkeit und Verfügbarkeit der Netzkomponenten betrachtet werden [96]. Schmitt definiert Dienstgüte wie folgt: „Dienstgüte kennzeichnet das definierte, kontrollierbare Verhalten eines Systems bezüglich quantitativ messbarer Parameter [87]“.

Diese Art der Dienstgüte kann als technische Dienstgüte bezeichnet werden, wohingegen unter Dienstgüte im Sinne der CCITT auch nicht-technische Aspekte von Dienstgüte fallen können, wie z.B. die Reputation eines Dienstanbieters [33, 34]. Unter Dienstgütekriterien werden in dieser Arbeit technische Faktoren wie Verfügbarkeit, Zugriffsverzögerung oder Wiederherstellungsziele (Recovery-Time-Objectives (RTO)) von Speicherhierarchien verstanden. Durch das Dienstgüteverhalten eines ILM-Szenarios werden die Nutzerzufriedenheit und damit die Akzeptanz eines solchen Systems direkt beeinflusst.

Diese genannten Nebenziele sind von der genauen Implementierung eines ILM-Systems abhängig. Sie können deshalb nicht in einem allgemeingültigen Kontext analysiert werden, sondern müssen von Fall zu Fall einzeln untersucht werden. Ein Beispiel dazu gibt, wie bereits erwähnt, Kapitel 7.3.

6.4 Annahmen und Vereinfachungen

ILM ist aufgrund der beteiligten Unternehmensbereiche sehr komplex. Deshalb werden für die Simulation von ILM-Szenarien Annahmen und Vereinfachungen getroffen. Diese sind im Einzelnen [4]:

- Das Konzept von ILM ist unabhängig von einer speziellen technischen Implementierung.
- ILM ist automatisiert durchführbar.
- Betrachtung von Microsoft® Office-Files (doc, xls, usw.).
- Verwendung einer rundenbasierten Simulation.
- Ein Simulationszyklus entspricht einem Tag.
- Dateizugriffe wirken sich nicht auf die Dateigröße aus.
- Dienstgütekriterien wirken sich auf die Nutzerzufriedenheit und die Akzeptanz eines ILM-Systems aus, jedoch nicht auf dessen Funktion.

ILM ist ein strategisches Konzept für die wertgemäße Speicherung von Daten und somit unabhängig von produktspezifischen Lösungen wie in Kapitel 3.4 beschrieben. Wichtiger als die technischen Komponenten ist für den Erfolg von ILM der organisatorische Teil (OrgTeil) [112]. Die Anforderungen des OrgTeils müssen unabhängig von der technischen Realisierung betrachtet werden, um ILM Ziel führend einzusetzen. Technische Details wirken sich primär auf die Nebenziele der Simulation, insbesondere auf die konkreten Kosten, aus. Sie haben keinen direkten Effekt auf das grundsätzliche ILM-Konzept. Aus diesem Grund werden sie im Rahmen der Simulation nicht berücksichtigt.

Ferner wird angenommen, dass ILM automatisiert durchführbar ist. Hierzu muss die Informationsklassifikation ohne Benutzereingriffe erfolgen können. Diese Vereinfachung setzt voraus, dass die Methode der Wertfindung den Wert einer Information automatisch bestimmt. Weiterhin wird in der Simulation angenommen, dass Wertveränderungen der Dateien ausschließlich durch das Zugriffsverhalten bestimmt werden. Diese Annahme macht eine automatisierte Informationsbewertung und somit die Simulation einer Migration ohne Nutzereingriff erst ermöglicht. Diese Voraussetzung bedingt jedoch die Anwendung der Simulationsergebnisse lediglich auf solche Dateien, die der Annahme entsprechen.

Weiterhin wird angenommen, dass nur Dokumente von Microsoft-Office-Anwendungen betrachtet werden. Unternehmen, die den Einsatz von ILM erwägen, benötigen primär neue Aufbewahrungsstrategien für Geschäftsdokumente. Diese Einschränkung ist demnach sinnvoll und steht nicht in Widerspruch zur Transferierbarkeit der Ergebnisse [68].

Die Simulationen von ILM erfolgt rundenbasiert. Es wird keine herkömmliche diskrete ereignisorientierte Simulation verwendet, sondern eine Variation, bei der ein Simulationszyklus

einem Tag entspricht. Die Verwendung einer rundenbasierten Simulation beruht auf der Annahme, dass automatisierte Migrationen auf niedrigere Speicherebenen jeweils nachts durchgeführt werden und entsprechende Rückmigrationen auf höhere Ebenen innerhalb des zugehörigen Zyklus durchgeführt werden können [4].

Es wird weiterhin, dass Dateizugriffe keine Auswirkung auf die Dateigröße haben. Dies ergibt sich sowohl aus den Untersuchungen in Kapitel 5.5 als auch auf den Ergebnissen von Gibson und Miller, die besagen, dass sich die Dateigröße durch einen Zugriff typischerweise um nicht mehr als 32 KB verändert [30].

Zuletzt wird angenommen, dass die Dienstgüte sich zwar auf die Nutzerzufriedenheit und die ILM-Akzeptanz auswirkt, nicht aber auf die ILM-Funktion. Das heißt zum Beispiel, dass ein ILM-System funktionieren kann, unabhängig davon ob die Zugriffszeit 10 Sekunden oder 10 Minuten dauert.

6.5 Der Simulationsplan

Nach Festlegung der Ziele und Annahmen gibt nun der Simulationsplan einen Überblick über die Vorgehensweise zur Simulation von ILM (vgl. Abbildung 21). Zunächst werden die ILM-Szenarien festgelegt. Dazu werden Parameter wie z.B. die Simulationsdauer oder die Speicherebenenanzahl definiert. Danach wird die Szenario-Konfiguration an den Simulator übergeben. Gemäß dem Simulationsmodell wird das ILM-Szenario durchgespielt und das Ergebnis in Form von Logfiles protokolliert. Diese Dateien dienen der Auswertung hinsichtlich der Untersuchungsziele. Diese betreffen einerseits die Dynamik des Szenarios und die Anzahl der Speicherhierarchien. Andererseits stehen Vergleiche zwischen verschiedenen Szenarien und Migrationsstrategien im Fokus der Auswertungen. Kosten auf Volumenbasis betrachtet, d.h. mit Hilfe der Kapazitätsbelegung α_i und den relativen Preisverhältnissen β_i der Speicherebenen (siehe Kapitel 2.3) kann das jeweilige Kostenreduktionspotenzial untersucht werden.

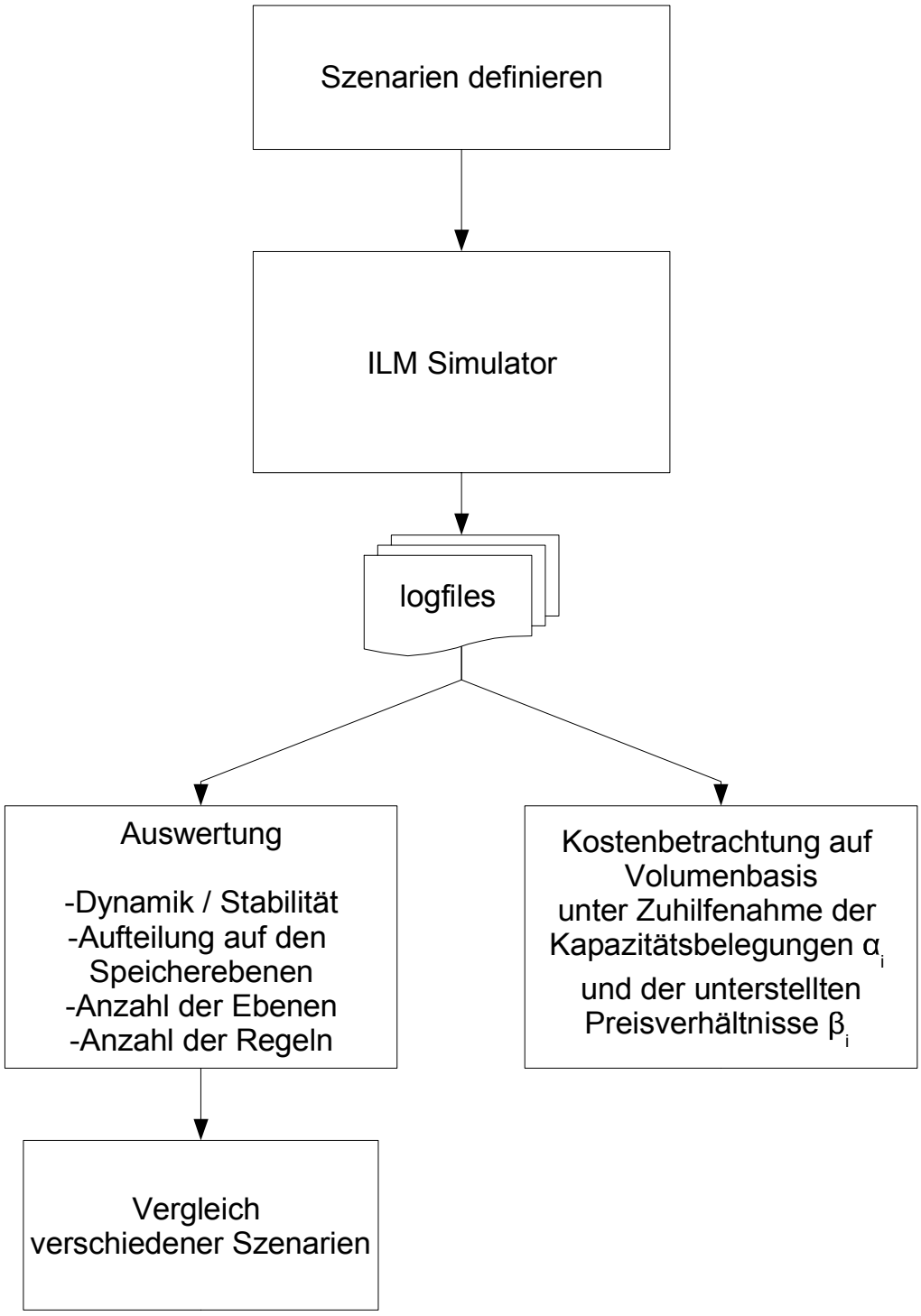


Abbildung 21: Simulationsplan [4]

6.6 Das Simulationsmodell

Der Simulator simuliert das grundlegende Verhalten eines ILM-Systems unter Anwendung definierter Migrationsregeln und Szenarien für Microsoft Office-Dokumente (siehe Abbildung 22). Er dient zur Generierung von Analysedaten von ILM.

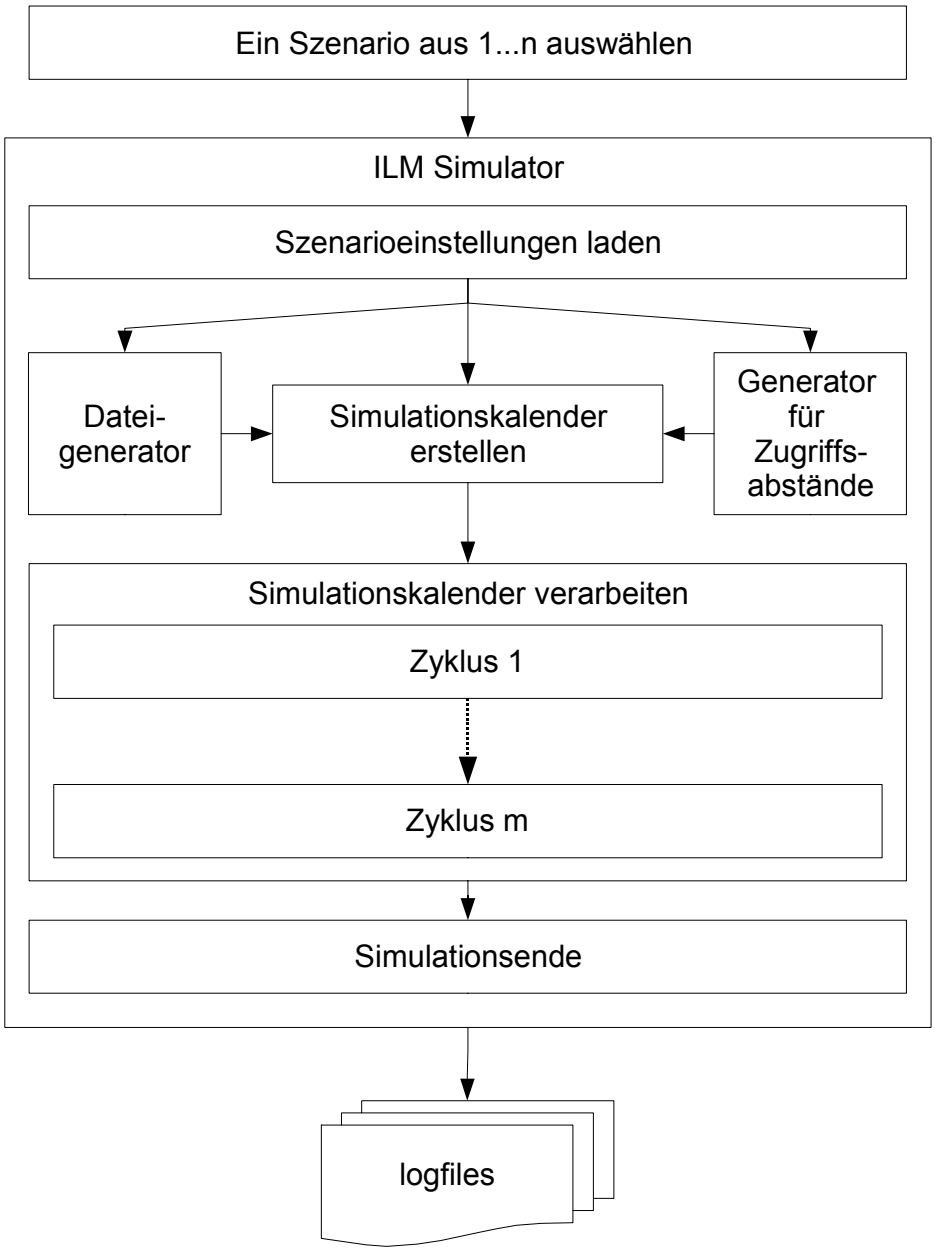


Abbildung 22: Simulationsmodell [4]

Nachdem ein Szenario definiert wurde, lädt das Simulationstool im ersten Schritt die Szenarioeinstellungen. Diese Konfiguration liefert die notwendigen Parameter für die Dateigeneration und die Erzeugung von Dateizugriffen. Danach wird mittels Dateigenerators und des Generators für Zugriffsabstände der Simulationskalender erstellt. Es beginnt daraufhin die zentrale Phase der Simulation, die das charakteristische Verhalten von ILM darstellt. Bei der Verarbeitung des Simulationskalenders werden alle im Simulationskalender geplanten Dateioperationen zyklisch vollzogen. Ist das Simulationsende erreicht, werden die Ergebnisprotokolle (logfiles) abgeschlossen.

6.6.1 Simulationsszenario

Das Simulationsszenario, kurz Szenario, dient der Beschreibung aller wichtigen Elemente einer ILM-Untersuchung. Es liefert dabei eine Festlegung der Umstände, unter denen ILM zum Einsatz gebracht werden soll. Weiterhin legt ein Szenario die Parameter für die Simulation von ILM fest. Aus implementierungstechnischer Sicht ist ein Szenario durch folgende Elemente gekennzeichnet [4]:

- Startsituation
- Anzahl der Simulationszyklen (Simulationsdauer)
- Anzahl der Speicherhierarchien
- Auswahl von Migrationsregeln
- Anzahl der simulierten Dateien

Durch die Startsituation wird festgelegt, welches Dateienwachstum für die Simulation angenommen wird. Die Anzahl der simulierten Dateien und die Simulationsdauer definieren den Umfang der Simulation. Die Anzahl der Speicherhierarchien bildet die Basis der technischen Realisierung des simulierten ILM-Systems. Sie ist Eingangsparameter und gleichzeitig Untersuchungsgegenstand der Simulationen. Die ausgewählten Migrationsregeln legen fest, welche Migrationsregeln genutzt werden.

6.6.2 Startsituation

Die Startsituation beschreibt den Ausgangszustand, in dem sich die ILM-Speichersysteme befinden. Sie legt weiterhin fest, welches Dateienwachstum angenommen wird. Tabelle 20 stellt die in realen Systemen möglichen Kombinationen dar. Diese Typ-Bezeichnungen werden im weiteren Verlauf verwendet. Die Spaltenbezeichner mit den Zuständen „konzentriert auf Speicherhierarchie 1“ und „vorsortiert auf verschiedene Hierarchien“ beschreiben die Hierarchieebenen, auf denen die Dokumente zu Beginn der Simulation abgelegt werden. Die Zeilenbeschriftung definiert, ob die Hierarchien anfangs leer oder gefüllt sind. Ferner gibt sie an, ob Datenwachstum vorliegt.

	konzentriert auf Speicherhierarchie 1	vorsortiert auf verschiedene Hierarchien
gefüllte Hierarchien	Typ 1	Typ 4
ungefüllte Hierarchien mit Datenwachstum	Typ 2	Typ 5
gefüllte Hierarchien mit Datenwachstum	Typ 3	Typ 6

Tabelle 20: Mögliche Typen als Startsituation

Typ 1 bezeichnet eine Startsituation, bei der bereits alle Dateien existieren und auf der ersten Hierarchieebene gespeichert sind. Praktisch hat diese Situation eine geringe Bedeutung, weil Unternehmen, die ILM einsetzen wollen, einem Datenwachstum unterliegen, sodass bedingt durch die Geschäftsaktivitäten immer neue Dateien erzeugt werden.

Typ 2 beschreibt die Ausgangssituation eines leeren Systems. Dieser Typ bedarf einer Wachstumsrate. Neu zu speichernde Dateien werden auf der ersten Speicherebene abgelegt. Diese Situation ist realistisch und gibt den Fall einer Neuanschaffung wieder, bei der die alte Speicherinfrastruktur weiterhin betrieben wird und neu erstellte Dokumente auf das neue System gespeichert werden.

Typ 3 ist eine Kombination aus Typ 1 und Typ 2. Dabei sind zu Beginn bereits Dateien im System vorhanden. Diese liegen auf der obersten Ebene. Durch das angenommene Datenwachstum nimmt der Datenbestand im Laufe der Zeit zu. Dieser Typ ist ebenfalls realistisch und entspricht dem Vorgehen, bei dem die alte Speicherinfrastruktur nicht weiter betrieben wird.

Die Startsituation nach Typ 4 ist eine Variation von Typ 1. Der Unterschied zwischen beiden Ausgangssituationen liegt darin, dass die Dateien in Typ 4 bereits bei Einführung des Systems nach ihrem Wert den Speicherhierarchien zugeordnet werden. Praktisch hat diese Variante eine geringe Bedeutung.

Typ 5 ist eine Variation von Typ 2 und Typ 6 eine Variation von Typ 3, jeweils mit dem Unterschied, dass die Dateien vorsortiert auf den Speicherebenen abgelegt werden. Die Aufteilung der Dateien auf die unterschiedlichen Speicherebenen zu Beginn eines ILM-Systems bedarf einer Wertzuweisungsmethode, die nicht auf Zugriffscharakteristika beruht. Eine manuelle Wertzuweisung könnte für diese Vorsortierung zum Einsatz kommen. Aufgrund des zu erwartenden Aufwands manueller Methoden wird den Startsituationen Typ 4 – Typ 6 in der Realität wenig Bedeutung zugeschrieben.

Daten zur manuellen Bewertung können nicht künstlich erzeugt werden, sodass Simulationen der Typen 5 und 6 nicht möglich sind. Typ 4 kann, wenn ein ILM-System stabil ist und gegen einen endlichen Wert konvergiert, simuliert werden. Dazu wird zunächst eine Phase mit Datenwachstum simuliert bis die definierte Anzahl Dateien vorhanden ist. Danach wird die Simulation ohne Datenwachstum fortgeführt. Ist das System eingeschwungen, kann die Beobachtung zu Typ 4 erfolgen. Auf diese Weise haben die Dateien bereits zu Beginn der Beobachtungsphase einen definierten Informationswert [4, 118].

6.6.3 Simulationskalender

Der Simulationskalender ist der zentrale Bestandteil des Simulationstools. Nachdem das Szenario in den Simulator geladen wurde, wird als im nächsten Simulationsschritt der Simulationskalender erstellt. Der Simulationskalender enthält zwei Listen für jeden Zyklus. Die erste Liste enthält Einträge über im jeweiligen Zyklus zu erstellende Dateien. Jeder Eintrag dieser Dateianlageliste besteht aus Dateityp, Dateigröße und den Bezeichner der Datei (Dateina-

men). Die Einträge der zweiten Liste legen fest, dass in dem zugehörigen Zyklus auf die zugehörigen Dateiobjekte zugegriffen wird (siehe Abbildung 23). Die Einträge der Dateizugriffsliste bestehen aus den Dateinamen der jeweiligen Datei.

Der Simulationskalender an das Konzept einer Diskreten Ereignisorientierten Simulation (DES) angelehnt. Um das Simulationsmodell einer ereignisorientierten Simulation eindeutig zu definieren, werden ein Zustandsmodell, ein Ereigniskalender und Ereignisroutinen zur Verarbeitung der Ereignisse benötigt [4, 53].

Der Kalender zur Simulation von ILM hat eine ähnliche Funktion. Er legt die Dateianlage- und Zugriffsereignisse zur späteren Weiterverarbeitung im Simulator fest. Das Modell des Simulationstools unterscheidet sich von der klassischen DES. Die Ereignis verarbeitenden Routinen werden rundenbasiert aufgerufen und nicht durch das Eintreten eines Ereignisses ausgelöst.

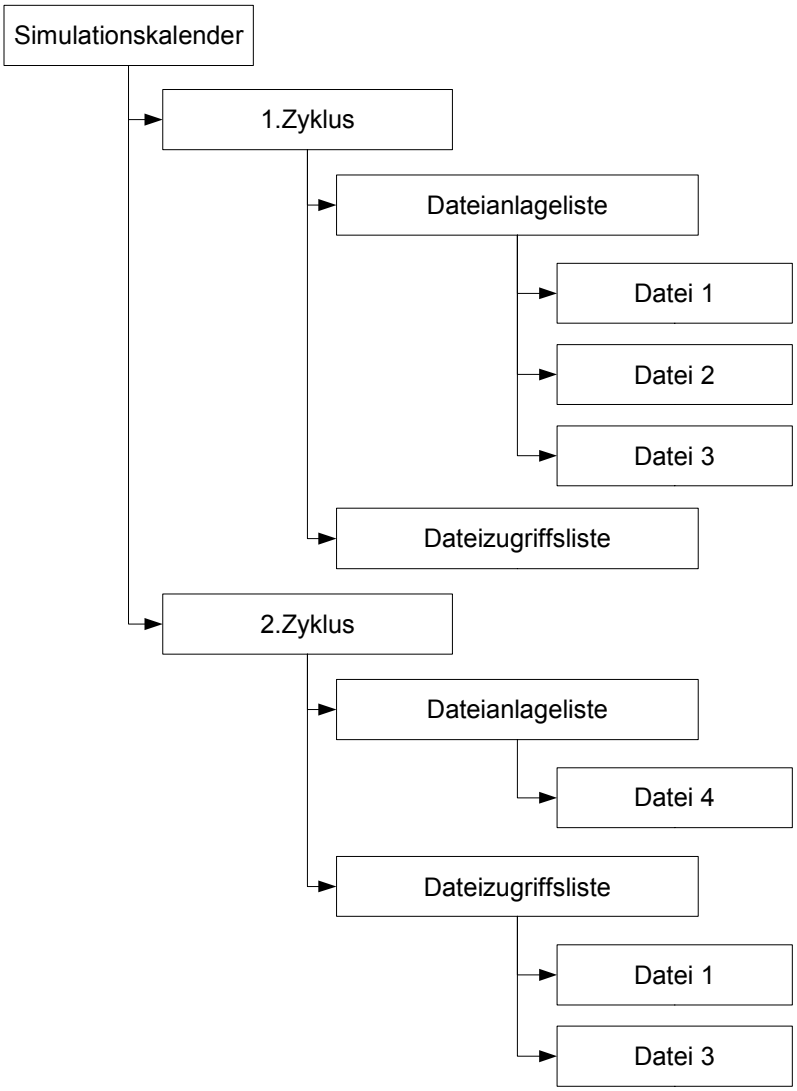


Abbildung 23: Simulationskalender [4]

Die Erstellung des Simulationskalenders erfolgt in zwei Durchläufen. Zuerst werden mittels Dateigeneratormoduls für jede Runde die Dateianlagelisten erstellt. Danach folgt die Erzeugung der Dateizugriffslisten, wobei der Generator für Dateizugriffe zum Einsatz kommt. Beide Generatoren werden nachfolgend detailliert betrachtet.

6.6.4 Dateigenerator

Der Dateigenerator wird zum Aufbau der Dateianlagelisten benutzt. Er kann Kalendereinträge auf deterministischer als auch auf stochastischer Basis generieren. Im deterministischen Fall wird der Generator mit Protokolldaten von Dokumenten aus einem realen System gespeist. Damit kann ILM für einen realen Datenbestand simuliert werden und dessen Verhalten über einen längeren Zeitraum simuliert werden.

Im stochastischen Falls werden die Attribute der Dateien durch Erzeugung von Zufallszahlen bestimmt. Folgende Merkmale der Dateien werden generiert, um ein Dateiobjekt vollständig zu simulieren [4]:

- Dateityp
- Dateigröße
- Abstand zur nächsten Dateierstellung

Jeder Dateityp tritt mit einer fest definierten Wahrscheinlichkeit auf. Als Schätzer für die Wahrscheinlichkeit p_i, wird die relative Häufigkeit h_i der Dateitypen i in der Stichprobe aus Kapitel 5.5 genutzt. Dementsprechend gilt

$$\hat{p}_i = h_i = \frac{n_i}{\sum_{i=1}^{k} n_i} \text{ mit k} \in \mathbb{N} \tag{30}$$

Die in der Simulation relativen Häufigkeiten entsprechen den Werten aus Kapitel 5 (siehe Tabelle 21):

Dateityp	doc	xls	ppt	pdf	Sonstige
Anzahl Dateien n_i	335	185	164	140	176
Relative Häufigkeiten h_i	0,335	0,185	0,164	0,14	0,176

Tabelle 21: Häufigkeiten der Dateitypen

In der Simulation wird angenommen, dass die Dateigröße (μ,σ^2)-normalverteilt ist. Die durchschnittliche Dateigröße beträgt μ=1139,96 kB bei einer Standardabweichung von σ=5174,05 kB. Die Aufgabe des Generators ist ferner, den Abstand zwischen der Erstellung von zwei Dateien festzulegen. Dazu wird eine exponentialverteilte Zufallszahl generiert. Die Exponentialverteilung wird in der Naturwissenschaft häufig verwendet, um Zeitabstände zwischen bestimmten Ereignissen nachzustellen [35]. Verteilungsparameter ist die Ankunftsrate der Ereignisse λ. Der Erwartungswert E(X) der exponentialverteilten Zeitabstände konvergiert bei einer hinreichend großen Datenerhebung gegen Kehrwert der Ankunftsrate, d.h. $E(X) = 1/\lambda$.

Dieses Phänomen wird zur Steuerung des simulierten Datenwachstums ausgenutzt. Unter der Annahme, dass in der Simulation keine Dateien gelöscht werden, kann ein lineares Wachstum angenommen werden. Hierzu wird als Verteilungsparameter λ die gewünschte Wachstumsrate eingesetzt.

6.6.5 Generator für Dateizugriffe

Der Generator für Dateizugriffe dient zur Erzeugung und Terminierung von Zugriffen auf Dateien. Wie der Dateigenerator kann dieser Generator Kalendereinträge deterministisch oder stochastisch generieren [4].

Aufgabe des Generators ist es, Dateizugriffslisten des Kalenders mit zufällig generierten Ereignissen zu füllen. Dazu wird der Abstand zwischen zwei Dateizugriffen stochastisch erzeugt. Bekannt sind der Dateiname sowie der letzte Zugriffzeitpunkt. Durch Addition des zufällig ermittelten zeitlichen Abstands wird der Zeitpunkt des nächsten Zugriffs errechnet. Die Abstände werden abhängig von Dateityp, Dateigröße und Zugriffshistorie mit gemischt-gestutzten weibullverteilten bzw. gemischt-gestutzten gammaverteilten Zufallsvariablen modelliert. Die Verteilungsparameter α und β der jeweiligen Verteilung hängen von diesen Kriterien ab (siehe Kapitel 5). Eine zusätzliche Rolle bei der Simulation spielt die Parametrisierung der Verteilungen. Die Ergebnisse der χ^2-Tests aus Kapitel 5 dienen hier zur Orientierung. In den Tabellen 22 und 23 sind die zutreffenden Verteilungen und Parameter für Dateien jünger als 1 Jahr bzw. älter aufgeführt. W steht für eine Weibullverteilung und G für eine Gammaverteilung mit den jeweiligen Parametern α und β.

Zugriffe / Dateityp	[1-6]	[7-14]	[15-∞)
doc	W(0,35;3,5)	G(0,32;183)	W(0,35;3,5)
xls	W(0,25;1,1)	W(0,25;1,1)	W(0,25;1,1)
ppt	W(0,38;14,3)	W(0,38;14,3)	W(0,38;14,3)
pdf	W(0,35;3,5)	G(0,32;183)	W(0,35;3,5)
sonstige	W(0,46;27,7)	G(0,29;181)	W(0,46;27,7)

Tabelle 22: Verteilungsmatrix für Dateien jünger als 1 Jahr [4]

Zugriffe / Dateityp	[1-6]	[7-14]	[15-∞)
doc	G(0,23;236)	G(0,32;183)	W(0,36;4,0)
xls	W(0,25;1,1)	W(0,25;1,1)	W(0,25;1,1)
ppt	W(0,38;14,3)	W(0,38;14,3)	W(0,38;14,3)
pdf	G(0,27;132)	G(0,32;183)	W(0,36;4,0)
sonstige	W(0,46;27,7)	G(0,29;181)	W(0,46;27,7)

Tabelle 23: Verteilungsmatrix für Dateien 1 Jahr oder älter [4]

Die Zufallsvariable X: „Anzahl der Tage seit dem letzten Zugriff“ ist quasi-stetig. Damit die erzeugten Zufallszahlen in einer rundenbasierten Simulation verwendet werden können, müssen sie diskretisiert werden. Hierzu rundet der Generator sie immer auf die nächst kleinere ganze Zahl ab.

6.6.6 Migrationsregeln

Migrationsregeln sorgen bei Wertänderungen von Dateien für die Migration zwischen den Speicherhierarchien. Nachfolgend die zugehörige Definition.

Definition 9 (Migrationsregel): *Eine Migrationsregel ist eine formale Vorschrift, die eine Information I aufgrund des Informationswertes V(I) einer Informationsklasse C zuordnet. Bei einer Änderung des Informationswertes V(I) bewirkt die Migrationsregel, dass eine Information ggf. einer anderen Informationsklasse C* und dadurch einer anderen Speicherhierarchie S* zugeordnet wird [4].*

Es wird angenommen, dass der Informationswert nur vom Zugriffsverhalten abhängig ist. Die hergeleitete Bewertungsmethode verwendet Verteilungsfunktionen für die Zugriffe. Es wird in jedem Zyklus die prognostizierte Zugriffswahrscheinlichkeit je Datei ermittelt. Fällt diese unter den Schwellwert 1-p, wird die Datei auf eine niedrigere Speicherhierarchie migriert.

Für die Simulationen können Wahrscheinlichkeitsverteilungen für die Zugriffsabstände verwendet werden [4]:

- Sprungfunktion
- lineare Verteilung
- gemischt-gestutzte Weibullverteilung
- gemischt-gestutzte Gammaverteilung

Die Sprungfunktion stellt eine zeitbasierte Wertzuweisung dar. Nach einer festen Zeitspanne ohne Zugriff wird die Datei migriert (siehe Kapitel 2). Diese Migrationsregel ist weit verbreitet und intuitiv.

Die lineare Verteilung nimmt an, dass die Zugriffswahrscheinlichkeit linear abnimmt. Diese Migrationsregel basiert auf einen einfachen Algorithmus zur Wertzuweisung. Seine Komplexität ist im Vergleich zu den in Kapitel 5 dargestellten Algorithmen deutlich geringer. Dadurch ist auch diese Migrationsregel intuitiv.

Die Migrationsregeln auf Basis der gemischt-gestutzten Weibull- bzw. Gammaverteilung sind Hauptuntersuchungsgegenstand des Simulators.

6.6.7 Verarbeitung des Simulationskalenders

Nach Darstellung der Struktur und Konstruktion des Simulationskalenders wird in diesem Abschnitt beschrieben, wie das Simulationstool den Simulationskalender verarbeitet.

Der Simulator durchläuft eine Schleife, in der der Simulationskalender Zyklus für Zyklus abgearbeitet wird. Je Zyklus werden verschiedene Aufgaben durchgeführt (siehe Abbildung 24).

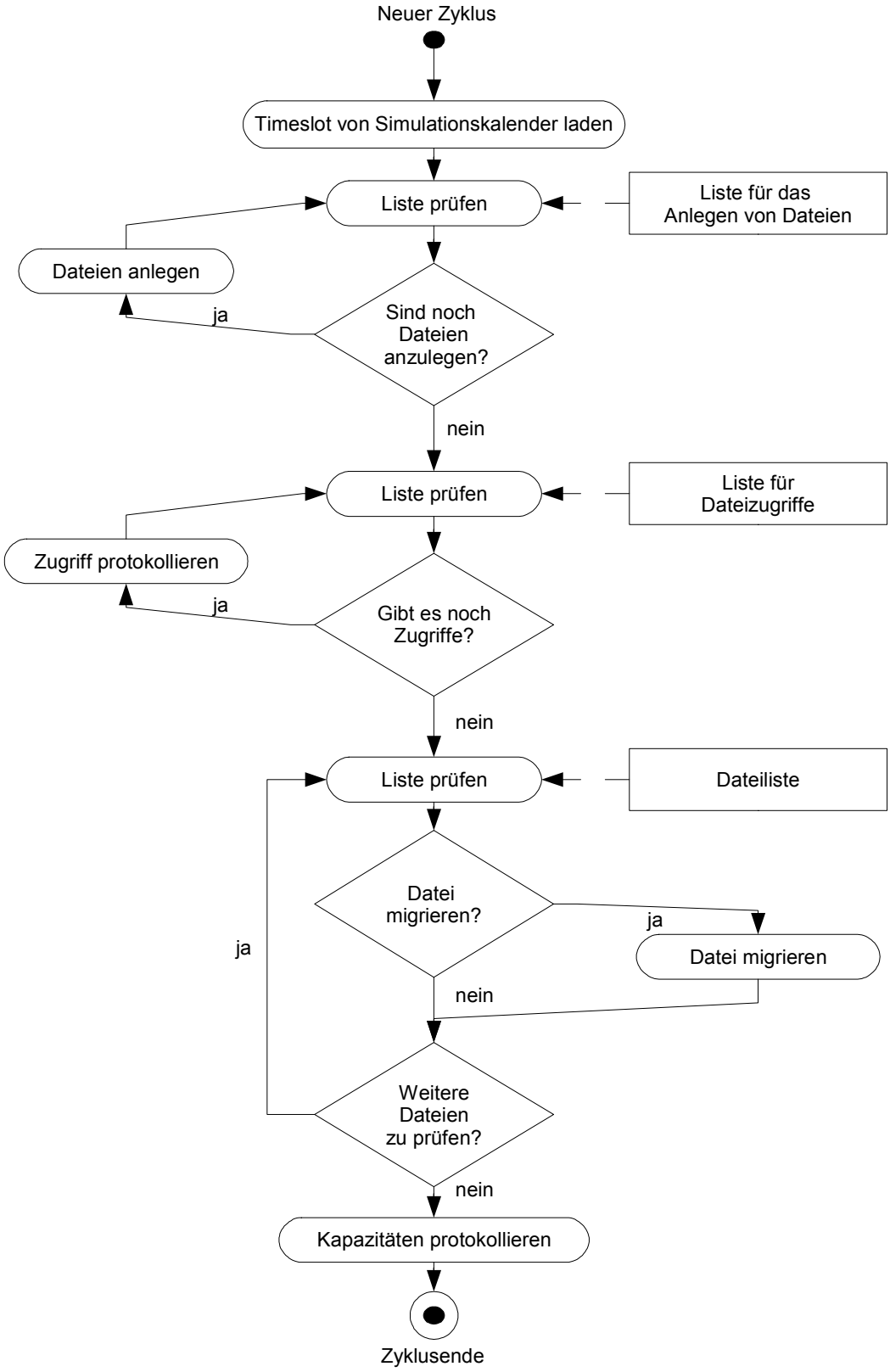


Abbildung 24: Ablaufplan eines Simulationszyklus [4]

Am Anfang wird die Dateianlageliste des jeweiligen Zyklus betrachtet. Der der vorhandenen Einträge entsprechenden Dateiobjekte angelegt. Danach wird die Zugriffsliste überprüft. Alle Zugriffe werden in den Logfiles und der Historie einer Datei protokolliert. Sind alle Zugriffe eines Zyklus verarbeitet, werden für alle im Simulator angelegten Dateiobjekte der Informationswert bestimmt. Die Migrationsregeln dienen dabei der Festlegung, ob eine Datei zu migrieren ist. Im Falle einer Migration wird die betroffene Datei im Simulator einer anderen Speicherhierarchie zugeordnet.

Jede Migration wird in den Logfiles vermerkt. Die Protokollierung der Migrationen wird am Ende jedes Zyklus aktualisiert und die resultierenden Kapazitätssalden jeder Speicherhierarchie notiert.

6.7 Erfolgsgrößen

Nachdem das Simulationsmodell erläutert wurde, dient dieses Kapitel der Erfolgsgrößen der Simulation. Diese sind im Hinblick auf die Simulationsziele wichtig (siehe Kapitel 6.2). Die zentralen Größen sind [4]:

- Durchschnittlicher Jitter
- Anzahl der Rückmigrationen pro Runde
- Absoluter Kapazitätsbedarf pro Ebene
- Relativer Kapazitätsbedarf pro Ebene
- Wachstumsrate des Kapazitätsbedarfs pro Ebene

Der durchschnittliche Jitter $\overline{\theta(t)}$ eine wichtige Messgröße. Er dient als Indikator für die Zuverlässigkeit von Migrationsregeln. Der Begriff Jitter[21] wird aus der Hin- und Herbewegung der Dateien zwischen den Speicherebenen abgeleitet. Der Jitter ist folgendermaßen definiert:

Definition 10 (Jitter): *Jitter $\theta_k(t)$ beschreibt die Anzahl der Migrationen einer Datei k von einer niedrigen Speicherklasse auf eine höhere Speicherklasse während des Beobachtungszeitraums t. Eine solche Migration wird auch als Rückmigration bezeichnet [4].*

Je größer der Jitter, desto häufiger hat das System eine Datei auf eine niedrigere Speicherhierarchie verschoben, obwohl danach noch auf sie zugegriffen wurde. Für ILM ist das Verhalten des gesamten Dateikorpus interessant, weshalb in den späteren Untersuchungen vorrangig der durchschnittliche Jitter betrachtet wird. Dabei wird während des Simulationszyklus der Jitter über alle Dateien im System gemittelt.

[21] Englisch: „Zittern“

Neben dem Jitter wird auch die Anzahl der Rückmigrationen pro Runde $\rho(t)$ betrachtet. Dies ist ein alternatives Zuverlässigkeitsmaß und kann besonders bei der aggregierten Betrachtung der Dateiengesamtheit eingesetzt werden. Im Vergleich zum durchschnittlichen Jitter besteht der Vorteil, dass nur über die Rückmigrationen während eines Zyklus gemittelt wird, nicht aber über alle Simulationszyklen wie beim Jitter. Dadurch werden Veränderungen im Zeitbereich leichter erkennbar.

Eine wichtige Erfolgsgröße ist der absolute Kapazitätsbedarf pro Ebene $a_i(t)$. Er gibt an, wie viel Speicherplatz in einer Hierarchieebene durch Dateien belegt ist. Als Basiskennziffer stellt er die Grundlage für weitere Erfolgsgrößen und ist wie folgt definiert:

Definition 11 (absoluter Kapazitätsbedarf): *Der absolute Kapazitätsbedarf einer Speicherebene i ergibt sich aus der Summe der Dateigrößen der auf i gespeicherten Dateien [4].*

$$a_i(t) = \sum_{j=1}^{n(t)} s_{ij} \tag{31}$$

n(t) : Anzahl der Dateien auf Speicherebene i

s_{ij} *: Größe der Datei j auf Speicherebene i*

Zum einen lässt sich aus dem Kapazitätsbedarf das Verhältnis zwischen Speicherbedarf einer Ebene und der Gesamtkapazität α_i errechnen. Diese Relation wird als relativer Kapazitätsbedarf bezeichnet und ist insbesondere für Kostenabschätzungen interessant.

$$\alpha_i = \frac{a_i}{\sum_{i=1}^{k} a_i} \tag{32}$$

Zum anderen kann aus dem zeitlichen Verlauf der belegten Speicherkapazität jeder Ebene das Wachstum des absoluten Kapazitätsbedarfs $g_i(t)$ ermittelt werden.

$$g_i(t) = \frac{a_i(t)}{a_i(t-1)} - 1 \tag{33}$$

Diese Größe ist insbesondere bei der Analyse des Verhaltens eines ILM-Systems im zeitlichen Verlauf nützlich. Im Falle von ILM-Szenarien mit Datenwachstum ermöglicht erst die Wachstumsrate die Identifikation einer stationären Phase.

6.8 Die Nomenklatur der Simulationsszenarien

Dieses Kapitels gibt nun einen Überblick über die Bezeichnungen der für die Simulation verfügbaren Szenarienausprägungen.

In Abschnitt 6.6.1 wurde bereits erläutert, wodurch ein Szenario eindeutig charakterisiert und identifiziert werden kann. Um eine bessere Übersichtlichkeit zu ermöglichen, sollen die in

Simulationen der nachfolgenden Kapitel untersuchten Szenarien jeweils durch eine Kurzbeschreibung identifiziert werden. Eine solche Kurzbeschreibung könnte beispielsweise wie folgt aussehen: Szenario T1-D3000-E3-R2-I2000. Die Abkürzung definiert dabei sämtliche Einzelfaktoren, die zur Charakterisierung eines Szenarios relevant sind. Der Szenarientyp wird in der Simulationsbeschreibung durch den ersten Kürzelabschnitt beginnend mit „T“ beschrieben. Simuliert werden die Typen 1-3 (vgl. Abschnitt 6.6.2). Der zweite Abschnitt der Kurzbeschreibung gibt die Simulationsdauer des Szenarios an. In dem Beispiel beträgt sie 3000 Simulationszyklen. Der nächste Abschnitt gibt die Anzahl der Speicherebenen an. Im vierten Abschnitt ist die Art der verwendeten Migrationsregel hinterlegt. Abschließend gibt der letzte Teil der Abkürzung die Anzahl der Informationsobjekte an, die anfänglich durch den Simulator verwaltet werden.

Abkürzung	Beschreibung
T	Typ des Szenarios
D	Simulationsdauer
E	Anzahl der Speicherebenen
R	Art der Migrationsregel
I	Startanzahl der Dateien

Tabelle 24: Abkürzungen der Szenarienausprägungen

6.9 Proof of Concept

Die ersten zwei Ziele der Simulation von ILM sind „eine ganzheitliche Untersuchung des Verhaltens von ILM-Szenarien“ sowie „die Beobachtung der Kapazitätsanforderungen je Hierarchie“ (siehe Abschnitt 6.2). Dazu gehören sowohl die Betrachtung jeder einzelnen Datei (Mikroebene) als auch die Betrachtung von gesamten Informationsklassen (Makroebene) entlang ihres Lebenszyklus. Nachfolgend werden diese beiden Kriterien untersucht.

6.9.1 Mikroebene

Der Lebenszyklus einer Information ist zeitlich durch beliebige Punkte t_1 und t_2 repräsentiert (siehe Abschnitt 6.1). Der Simulator ist in der Lage, den Wert einer Datei $V(I(t))$ über das komplette Zeitintervall nachzubilden. Damit kann die Simulation als ganzheitlich angesehen werden. Bei der ganzheitlichen Analyse auf Mikroebene spielen folgende Aspekte eine Rolle:

- Zugriffsaktivität
- Migrationsaktivität

Um diese Aspekte darstellen und erläutern zu können, wurde das Szenario T2-D4000-E5-R5-I100 simuliert und anschließend eine beliebige Datei als Untersuchungsobjekt entnommen.

6.9.1.1 Zugriffsaktivität

Zugriffsaktivität bezeichnet die Häufigkeit von Zugriffen an einem bestimmten Tag. Abbildung 25 stellt die Zugriffsaktivität einer Datei im Laufe des Lebenszyklus graphisch dar. An diesem Beispiel sieht man, wie die Zugriffe auf eine Datei verteilt sind. Im Zyklus 0 beginnt die Beobachtung und die Datei angelegt. Die weiteren Zugriffe sind durch Markierer gekennzeichnet. Erfolgen in einer Runde mehrere Zugriffe auf eine Datei, ist dies anhand der größeren Ausschläge erkennbar. So wurde zum Beispiel auf die Datei am 3117. Tag dreimal zugegriffen. Die Zugriffsaktivität bestimmt die Migrationsaktivität, die nachfolgend beschrieben wird.

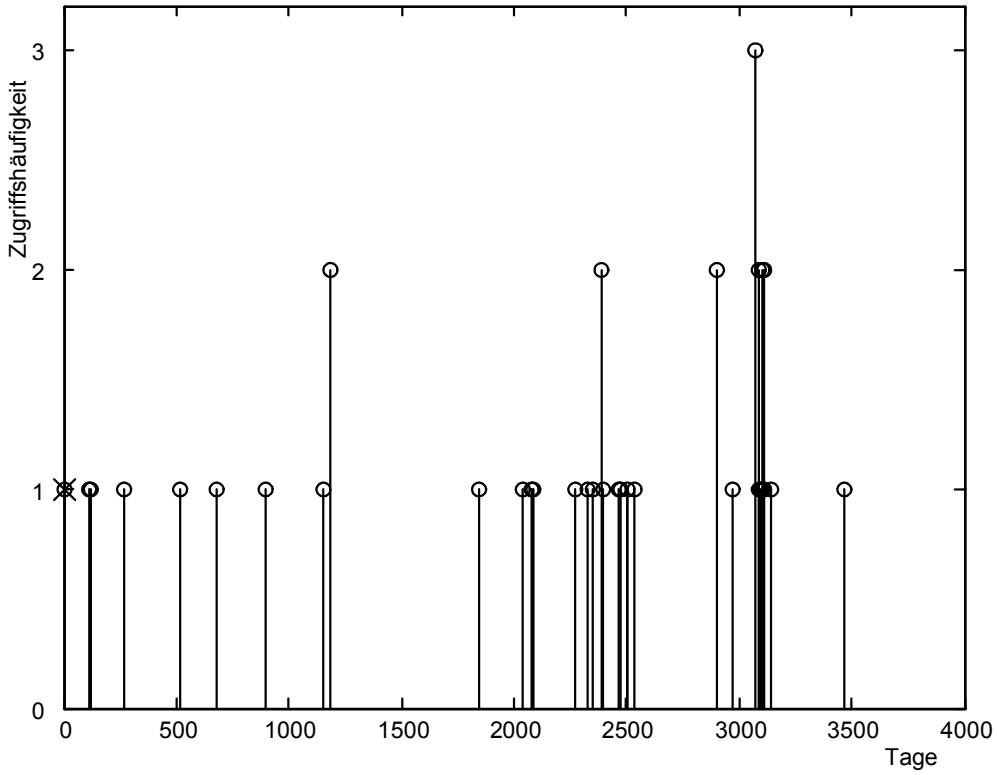


Abbildung 25: Zugriffsaktivität einer exemplarisch gewählten Datei

6.9.1.2 Migrationsaktivität

Die Migrationsaktivität beschreibt das Auftreten von Migrationen einer Datei. Sie ist hauptsächlich bestimmt durch das Zugriffsverhalten auf eine Datei. Weitere Einflussfaktoren sind die Migrationsregeln. In Abbildung 26 ist die Migrationsaktivität abgebildet, die zu dem Beispiel aus Abbildung 25 gehört. Die negativen Ausschläge markieren Migrationen auf niedrigere Speicherebenen, die in Folge eines sinkenden Informationswerts ausgelöst werden. Die positiven Ausschläge stehen für Rückmigrationen auf eine höhere Speicherebene. Für die Analyse ist es unerheblich, ob die Datei zurückmigriert wird, um einen Zugriff technisch erst zu ermöglichen oder wegen der Informationswertänderung durch einen Zugriff (siehe Abschnitt 6.1).

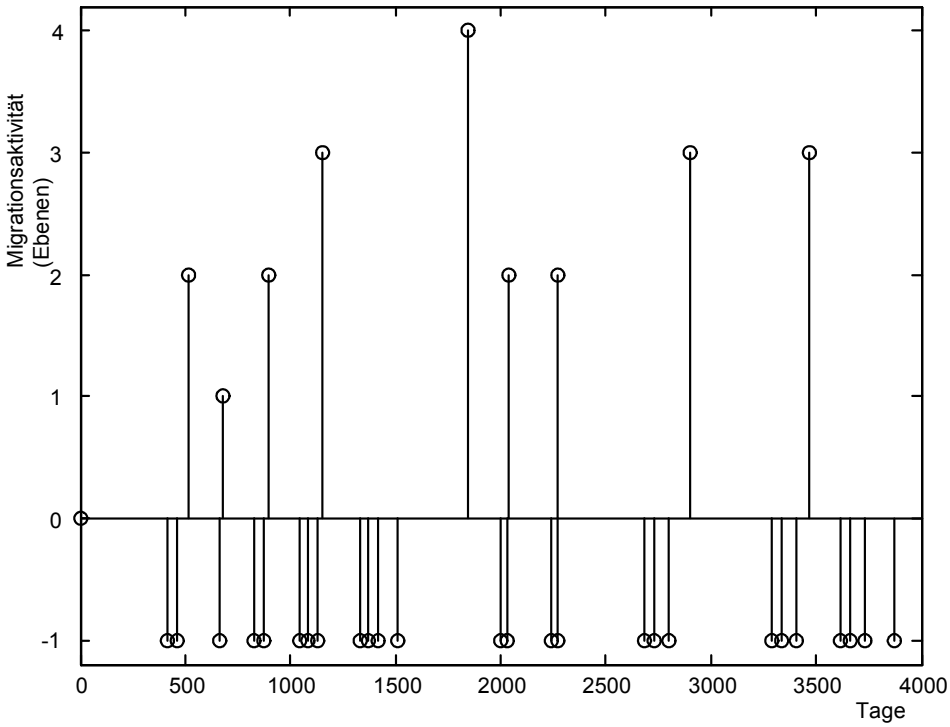


Abbildung 26: Migrationsaktivität einer exemplarisch gewählten Datei

Mit einer anderen Darstellung kann gezeigt werden, wie sich der Aufbewahrungszustand einer Datei im Laufe ihres Lebenszyklus verändert. In Abbildung 27 ist zu sehen, dass zwischen 1500 und 1800 Tagen nach Erstellung der absinkt. Darum wird die Datei in mehreren Schritten auf die niedrigste Speicherhierarchie verschoben. Nach ca. 1800 Tagen wird jedoch auf die Datei zugegriffen, so dass sie auf die höchste Speicherebene migriert wird. Nach diesem Zeitpunkt dauert es bis etwa 3800 Tage bis die Information wieder auf der niedrigsten Speicherebene abgelegt wird.

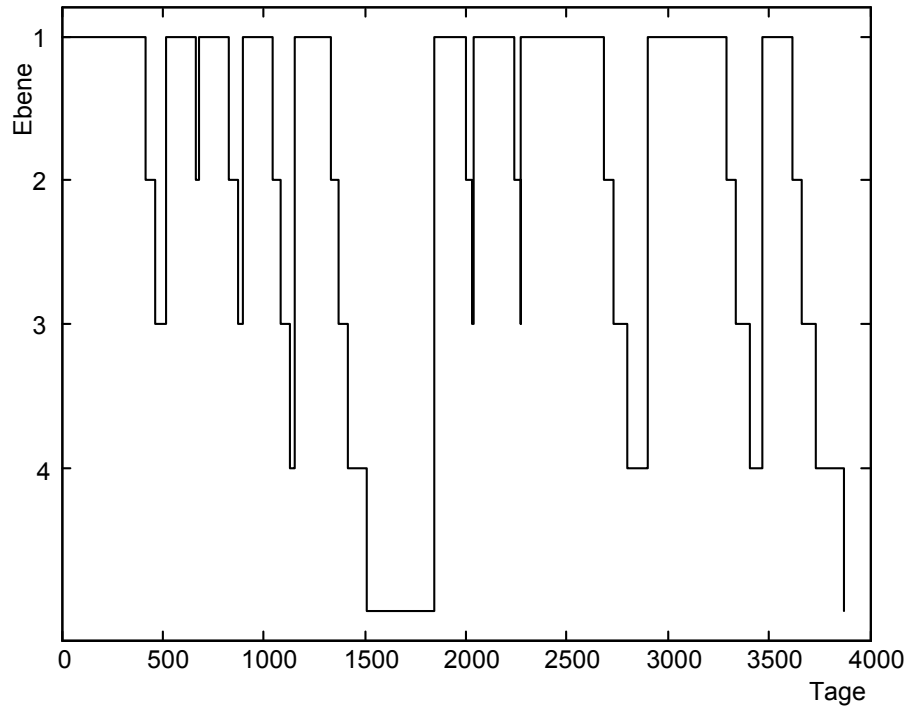


Abbildung 27: Aufbewahrungsebenen einer Datei im Laufe des Lebenszyklus

Da das Informationsobjekt zwischen den Speicherklassen hin und her migriert wird, dient dieses Beispiel auch zur Demonstration von Jitter (vgl. Abschnitt 6.1). Bis die Datei endgültig auf Speicherebene 5 abgelegt wurde, wurde sie vorher neunmal auf eine höhere Hierarchie zurückmigriert. Diese Datei hat demnach einen Jitter von $\theta(4000) = 9$.

Ein weiteres Beispiel zeigt den charakteristischen Verlauf einer Datei mit einer Häufung der Zugriffe zu Beginn des Lebenszyklus (siehe Abbildung 28). Die Datei hat lediglich zu Beginn des Lebenszyklus einen hohen Wert. Ihr Wert sinkt nach der Anfangsphase deutlich ab. Nach ungefähr 1600 Simulationszyklen erfolgen keine weiteren Zugriffe mehr, so dass die Datei bis zum Ende des Lebenszyklus auf der niedrigsten Hierarchieebene verweilt.

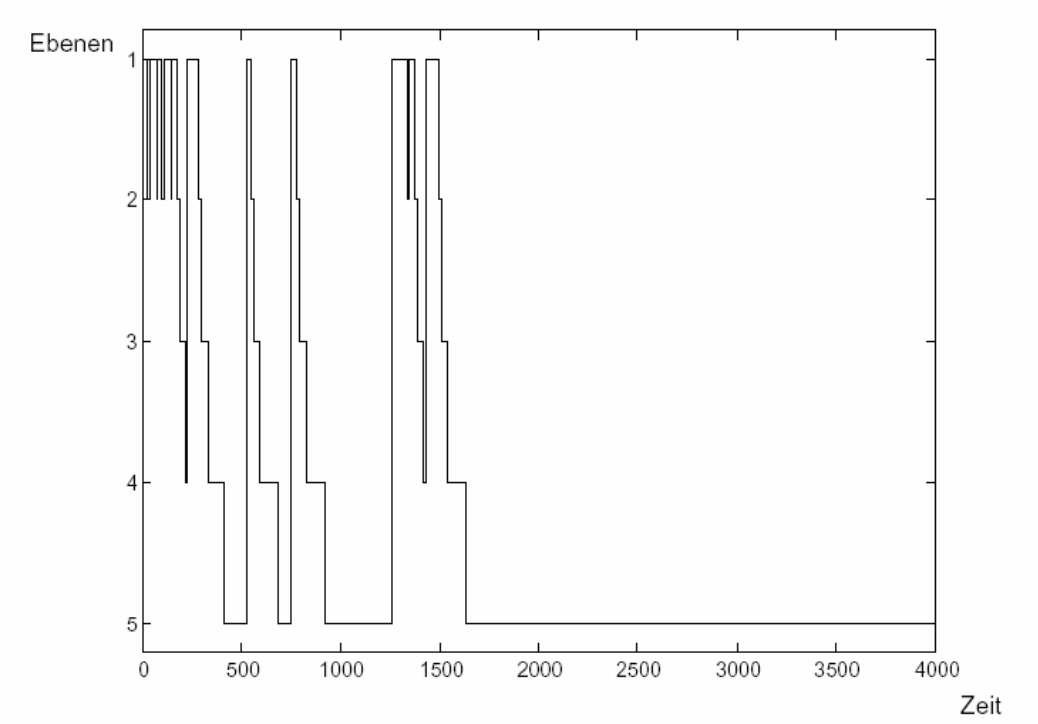


Abbildung 28: Aufbewahrungsebenen einer Datei im Laufe des Lebenszyklus

Die Darstellungen in diesem Kapitel sind nur sinnvoll auf Mikroebene, also bei der Betrachtung des Lebenszyklus einer einzelnen Datei. Aus aggregierter Sichtweise ist das Gesamtverhalten von ILM interessant, welches in den nachfolgenden Kapiteln untersucht wird.

6.9.2 Makroebene

Die Makroebene gibt das Gesamtszenario aller Dateien wieder. Die Gesamtkapazität aller Hierarchien über den gesamten Simulationszyklus wird ausgegeben. Zusätzlich dient der beobachtete Jitter als Maß für die Zuverlässigkeit der Migrationsregeln und somit des Gesamtsystems.

Abbildung 29 zeigt die Resultate einer Simulation mit 3 Hierarchien. Als Migrationregeln wurden verwandt:

Regel 1: Migriere die Datei, wenn ihre zukünftige Zugriffswahrscheinlichkeit unter 10% fällt (p_1=10%).

Regel 2: Migriere die Datei, wenn ihre zukünftige Zugriffswahrscheinlichkeit unter 5% fällt (p_2=5%).

Regel 3: Migriere die Datei auf Hierarchie 1, wenn auf die Datei zugegriffen wird.

Das jährliche Datenwachstum betrug 20%.

Nach 3200 Simulationszyklen war der relative Kapazitätsbedarf der einzelnen Ebenen:

Hierarchie 1: 46%, Hierarchie 2: 17% und Hierarchie 3: 37%.

Der beobachtete Jitter betrug: 1,89.

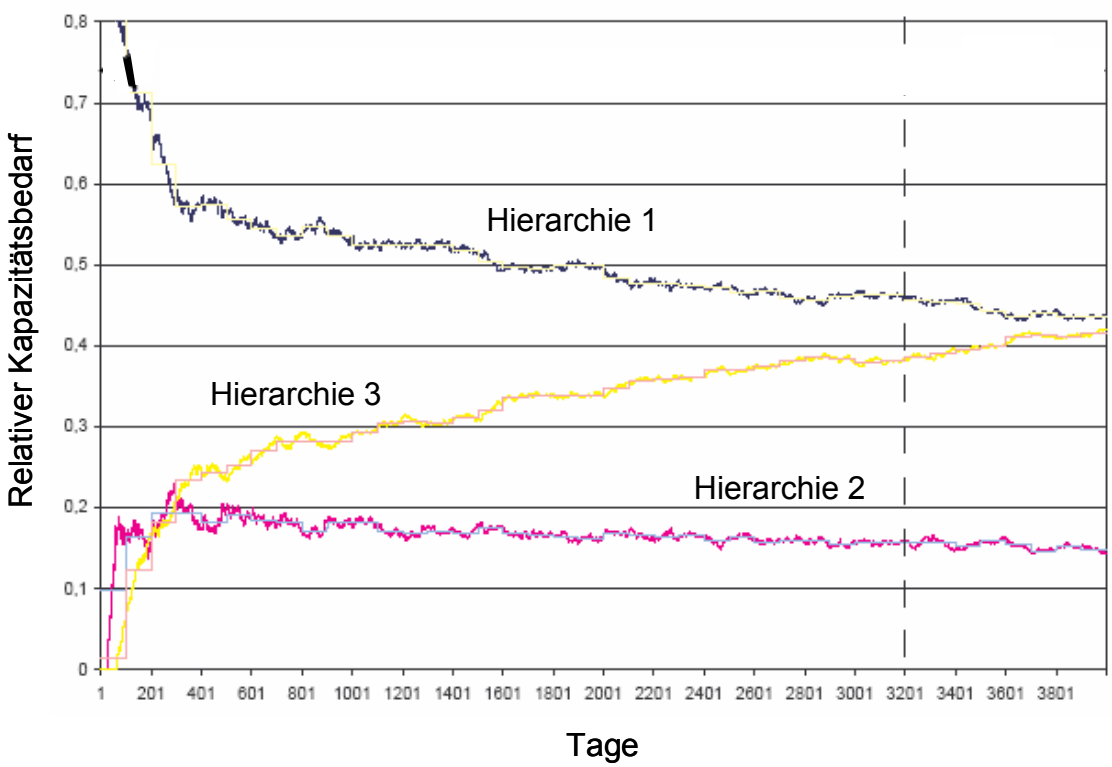


Abbildung 29: Typ 1 –Szenario

6.10 Zusammenfassung

Nach Herleitung der Methode der Wahrscheinlichkeit zukünftiger Zugriffe war es möglich, Dateien zu klassifizieren. In diesem Kapitel wurde geprüft, ob diese Methode automatisiert täglich tausende Dateien bewerten kann und somit für ILM verwendbar ist. Dazu wurde eine Simulationsumgebung implementiert. Es wurde zunächst das Simulationswerkzeug erläutert. Die Funktion des Simulators wurde mit einem Proof of Concept belegt. Darin wurde die Erreichung der ersten zwei der angegebenen Simulationsziele überprüft. Ebenso wurde damit auch die vierte Forschungsfrage „Ist eine derartige (prognostizierende) Methode in ILM verwendbar?“ positiv beantwortet.

Nachdem die Funktion des Simulators sichergestellt ist, werden in den nächsten Kapiteln die weiteren Ziele der Simulation verfolgt. Diese sind (siehe Abschnitt 6.2):

- Untersuchung verschiedener Migrationsregeln
- Untersuchung verschiedener Szenarien bei gleicher Migrationsregel
- Analyse des dynamischen Verhaltens von ILM-Szenarien
- Verwendung realer und künstlicher Dateienpools

7 Anwendung der Methode der Wahrscheinlichkeit zukünftiger Zugriffe

Mit Hilfe der vorgestellten „Methode der Wahrscheinlichkeit zukünftiger Zugriffe“ und des implementierten Simulators lassen sich nun ILM-Szenarien simulieren und daraus Erkenntnisse ziehen. Im Moment fehlt es noch an Erfahrungen mit ILM, um ILM beurteilen zu können. Der Bedarf nach Erfahrungen spiegelt sich in einer Studie von Glasshouse, wonach 90% der befragten IT-Entscheider an einen Einsatz von ILM denken [49]. Die hier gezeigten Simulationsergebnisse unterstützen IT-Manager in ihrer Entscheidungsfindung bezüglich des Designs von ILM-Umgebungen.

In diesem Kapitel wird zum einen die Analyse der notwendigen Anzahl von Hierarchien durchgeführt und zum anderen wird ein konkretes Beispiel zur Entscheidungsfindung über Migrationsregeln vorgestellt. In diesem Beispiel wird die Methode der Wahrscheinlichkeit zukünftiger Zugriffe mit vier anderen Methoden der Wertzuweisung verglichen und bewertet.

Die Analyse des dynamischen Systemverhaltens wurde ebenfalls durchgeführt. Die Ergebnisse werden nachfolgend kurz zusammenfassend dargestellt [4].

7.1 Analyse des dynamischen Systemverhaltens

Eines der Ziele von ILM-Simulation ist, das Verhalten eines ILM-Szenarios bezüglich seiner Dynamik zu analysieren. In Abschnitt 6.7 bereits betrachtet, welche Größen das dynamische Verhalten bestimmen. Weiter wurde untersucht, ob das System in Bezug auf diese Messgrößen Stabilitätskriterien erfüllt. Daraus kann das typische Verhalten des Systems bestimmt werden.

Die Dynamik eines ILM-Systems wurde mit Hilfe mehrerer Simulationsreihen explorativ untersucht. Es wurden Ausgangssituationen vom Typ 1 bis 3 betrachtet (T1-D4000-E3-R5-I500, T2-D4000-E3-R5-I500, T3-D4000-E3-R5-I500). Der relative Kapazitätsbedarf stellte sich als zentrale Untersuchungsgröße heraus. Zusätzlich fasst Tabelle 25 die Ergebnisse der übrigen Kennzahlen kurz zusammen.

	stabil			stationär		
Szenario / Erfolgsgröße	1	2	3	1	2	3
$a_i(t)$	ja	nein	nein	nein	nein	nein
$\alpha_i(t)$	ja	ja	ja	nein	nein	nein
$g_i(t)$	ja	ja	ja	ja	ja	ja
$\rho(t)$	ja	nein	nein	ja	nein	nein
$\overline{\theta(t)}$	nein	nein	nein	nein	nein	nein

Tabelle 25: Zusammenfassung der Simulationsergebnisse

7.2 Analyse der notwendigen Anzahl der Speicherhierarchien

Dieses Kapitel untersucht nun den Einfluss der Anzahl verwendeter Speicherhierarchien auf das typische Verhalten von ILM. Außerdem wird die Fragestellung betrachtet, welche Anzahl der Speicherhierarchien in einem ILM-System sinnvoll ist.

Zur Untersuchung werden zuerst im Rahmen einer Sensitivitätsanalyse vergleichbare ILM-Szenarien mit einer unterschiedlichen Anzahl der Speicherhierarchien simuliert Danach werden die Auswirkungen der Speicherhierarchien analysiert. Dazu wird mittels analytischer Methoden das Systemverhalten untersucht, wenn die Anzahl der Speicherebenen im laufenden Betrieb verändert wird.

7.2.1 Sensitivitätsanalyse

Sensitivitätsanalysen dienen der Überprüfung des Einflusses von Inputfaktoren auf die zu untersuchenden Ergebnisgrößen. Mit Hilfe von Sensitivitätsanalysen kann systematisch die Empfindlichkeit gegenüber Variationen der Inputfaktoren ermittelt werden [105]. Die Sensitivitätsanalyse kann auf zwei Arten durchgeführt werden. Zum einen durch mathematische Analyse von Modellgleichungen und andererseits durch Anwendung von variierten Inputfaktoren und Simulation der Ergebnisgrößen mit anschließendem Vergleich der Ergebnisse gegenüber den Referenzergebnissen [105].

Nachfolgend wird mittels Sensitivitätsanalyse der Einfluss der Anzahl der Speicherebenen auf ILM-Systeme erforscht.

7.2.2 Simulation der Speicherebenen

Zur Untersuchung des Effektes der Speicherebenenanzahl werden vier Simulationsdurchläufe durchgeführt. Als Grundlage wird dabei ein T3-D2000-E2-R5-I500-Szenario simuliert und die Komponente *E* variiert. Das Datenwachstum beträgt dabei 20% pro Jahr. Der Simulator beginnt die Simulation mit einem Datenbestand von 500 Dateien. Die Simulationsdauer beträgt 2000 Tage. Um Schwankungen der Messwerte zu reduzieren, werden pro Simulation Beobachtungen aus zehn simulierten Datenreihen gemittelt [4]. Als Migrationsregeln kommen jene nach der entwickelten Methode der Wahrscheinlichkeit zukünftiger Zugriffe zum Einsatz.

Variabler Eingangsfaktor ist die Anzahl der Speicherebenen. In jeder Simulation lautet die Schwellwertwahrscheinlichkeit der ersten Ebene p_1= 10%. Die Abstände der Schwellwertwahrscheinlichkeiten d_{p1} der übrigen Ebenen *i* sind äquidistant aufgeteilt (siehe Abbildung 30 mit p_1= 11% bzw. p_2= 5,5%.).

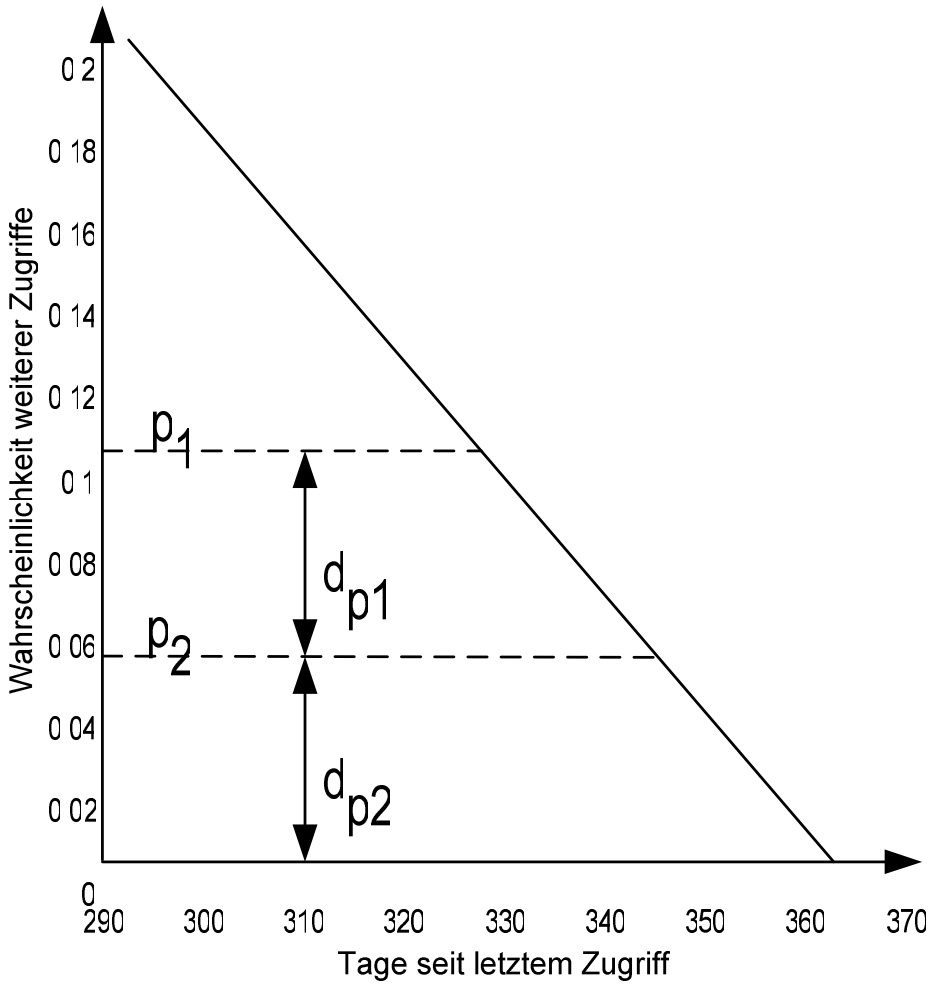


Abbildung 30: Exemplarische Darstellung der Schwellwertwahrscheinlichkeit

Die Schwellwertwahrscheinlichkeit definiert die Grenze zwischen zwei Speicherklassen. Passiert die Zugriffswahrscheinlichkeit einer Datei die Schwellwertwahrscheinlichkeit, so wird die Datei auf eine andere Speicherhierarchie migriert. Beim Einfügen von zusätzlichen Speicherhierarchien werden die Schwellwertwahrscheinlichkeiten derart angepasst, so dass sie wieder den gleichen Abstand zueinander haben.

Die Auswirkungen der Eingangsvariablen lassen sich am sinnvollsten anhand des relativen Kapazitätsbedarfs und des durchschnittlichen Jitters beobachten. Da beide Messgrößen nicht stationär sind, ist es schwierig, die Beobachtungswerte zu quantifizieren. Als alternativer Punktschätzer wird der Durchschnitt aller beobachteten Messwerte verwendet. Dabei sind lediglich Ausreißer bezüglich des relativen Kapazitätsbedarfs problematisch, die im ersten Simulationszyklus durch plötzliche Migration von Dateien der ersten Ebene auf günstigere Hierarchien entstehen. Sortiert man diese Beobachtungswerte aus, so stellt der Durchschnittswert ein sinnvolles Maß für beide Erfolgsgrößen dar. Beim durchschnittlichen Jitter kann auf das Filtern Messwerte komplett verzichtet werden. Als Referenzpunkt wird der durchschnittliche Jitter am 1000. Tag verwendet.

Am Anfang wird zuerst eine Simulation mit zwei Speicherebenen durchgeführt. Sukzessiv wird die Anzahl der Speicherhierarchien um jeweils eine Stufe pro Durchlauf erhöht bis zur

maximalen Anzahl von fünf Ebenen. Bei der ersten Simulation gibt es nur eine Schwellwertwahrscheinlichkeit. Diese beträgt p_1= 10% und liegt zwischen Ebene eins und Ebene zwei.

Tabelle 26 zeigt das Simulationsergebnis. Hierzu wurden jeweils die Ausreißerwerte zu Beginn der Datenreihen entfernt und der Durchschnitt der verbleibenden Beobachtungswerte gebildet. Es ergibt sich etwa ein Verhältnis von 1:1 für den Speicherbedarf der beiden Ebenen zueinander. Der durchschnittliche Jitter beträgt $\overline{\theta(1000)}$ =2,136.

Relativer Speicherbedarf bei 2 Ebenen	α_1	α_2
	0,508	0,492

Tabelle 26: Mittlerer relativer Kapazitätsbedarf bei 2 Ebenen (T3-D2000-E2-R5-I500)

Im nächsten Simulationsdurchlauf stehen drei Hierarchien zur Verfügung. Die zugehörigen Schwellwertwahrscheinlichkeiten liegen bei p_1= 10% und p_2= 5%. Tabelle 27 stellt das Simulationsergebnis des relativen Kapazitätsbedarfs α_i dar.

Relativer Speicherbedarf bei 3 Ebenen	α_1	α_2	α_3
	0,510	0,166	0,323

Tabelle 27: Mittlerer relativer Kapazitätsbedarf bei 3 Ebenen (T3-D2000-E3-R5-I500)

51% der Dateien liegen auf der ersten Speicherhierarchie. Auf der zweiten Ebene werden 16,6% und auf der dritten knapp ein Drittel des gesamten Dateienbestandes aufbewahrt. Es wurde ein mittlerer Jitter von $\overline{\theta(1000)}$ =2,093 gemessen.

Bei der dritten Simulation können vier Speicherebenen genutzt werden. Der Schwellwerte betragen p_1= 10%, p_2= 6,66% und p_3=3,33%. Die sich aus der Simulation ergebenden Kapazitätsanforderungen sind in Tabelle 28 notiert. 51,1% des gesamten Kapazitätsbedarfs werden auf die erste Speicherebene verwendet. Gut 10,3% der Dateien sind auf Hierarchie zwei gespeichert. 13,9% werden auf der dritten Ebene gespeichert. Die übrigen 24,7% des Dateienbestandes werden auf der vierten Hierarchie aufbewahrt. Es wurde ein Jitter in Höhe von $\overline{\theta(1000)}$ =2,16 gemessen.

Relativer Speicherbedarf bei 4 Ebenen	α_1	α_2	α_3	α_4
	0,511	0,103	0,139	0,247

Tabelle 28: Mittlerer relativer Kapazitätsbedarf bei 4 Ebenen (T3-D2000-E4-R5-I500)

Zu Abschluss wird eine Simulation mit fünf Speicherebene durchgeführt. Die Schwellwertwahrscheinlichkeit betragen p_1= 10%, p_2= 7,5%, p_3= 5% und p_4= 2,5%. Im Ergebnis liegen wieder etwa die Hälfte der Daten auf Ebene eins. Ebene zwei enthält 7,4%, Ebene drei 9,2% und Ebene vier 12,3%. Die verbleibenden 20,9 %, liegen auf der fünften Speicherebene (siehe Tabelle 29).

Relativer Speicherbedarf bei 5 Ebenen	α_1	α_2	α_3	α_4	α_5
	0,502	0,074	0,092	0,123	0,209

Tabelle 29: Mittlerer relativer Kapazitätsbedarf bei 5 Ebenen (T3-D2000-E5-R5-I500)

Ferner beträgt der gemessene Jitterwert $\overline{\theta(1000)} = 2{,}17$.

Insgesamt weichen die beobachteten Jitterwerte über alle vier Simulationen um weniger als 5% voneinander ab. Es kann somit davon ausgegangen werden, dass die Zuverlässigkeit unabhängig von der Anzahl der Speicherebenen ist.

Als Übersicht sind die Simulationsergebnisse der vier Durchläufe zusammen in Abbildung 31 dargestellt. Unter Anwendung der Sensitivitätsanalyse werden nun die Auswirkungen der Variation der Speicherebenen untersucht. Beginnend mit zwei Hierarchien wird die Ebenenanzahl schrittweise um eins erhöht. Man erkennt, dass der Speicherbedarf der ersten Ebene trotz Änderung der Anzahl der Speicherhierarchien annähernd konstant bleibt. Der Kapazitätsbedarf der zweiten Ebene verringert sich stetig mit zunehmender Anzahl der Speicherebenen.

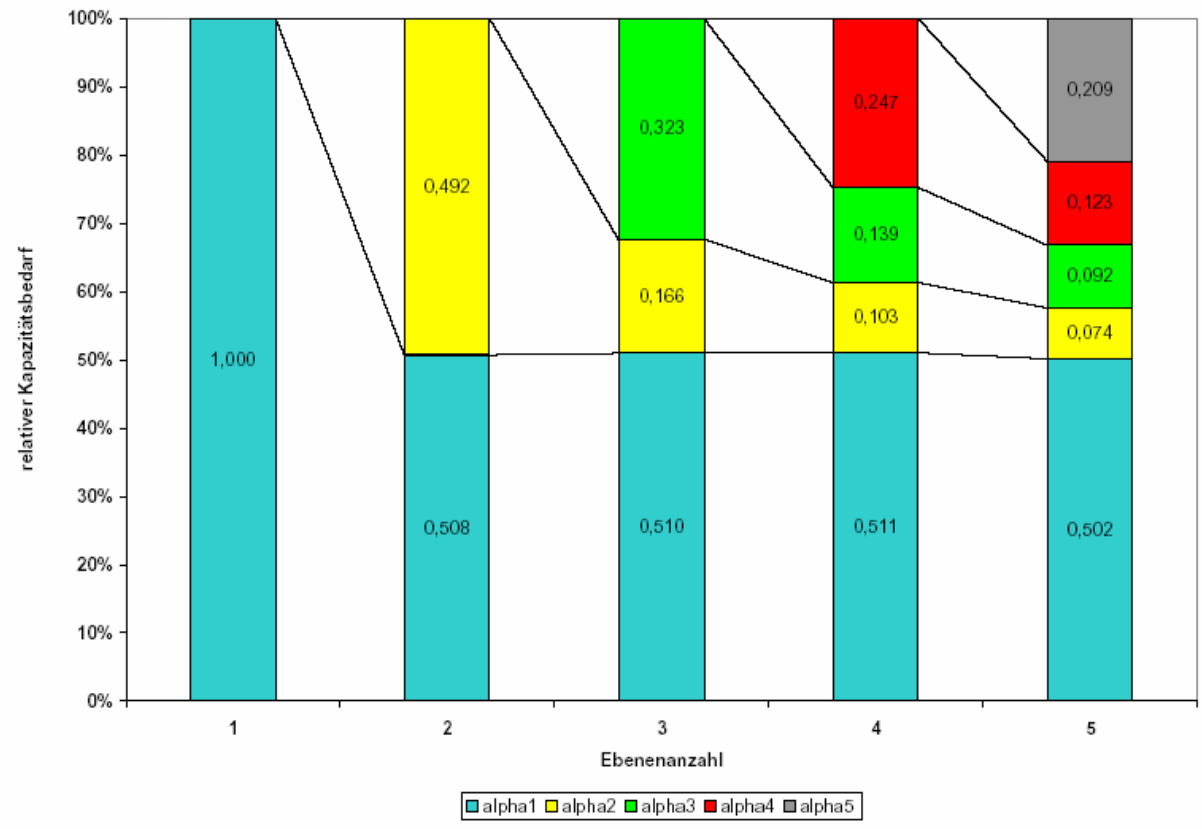


Abbildung 31: Mittlerer relativer Kapazitätsbedarf in Abhängigkeit von der Anzahl der Speicherhierarchien

Es fällt auf, dass bei äquidistanten Schwellwerten die niedrigste Ebene stets den größten Teil des Restdatenbestandes aufbewahrt. Insgesamt werden also auf der obersten und der untersten Hierarchie die meisten Dateien gespeichert. Dieses Ergebnis deckt sich mit den Beobachtungen in realen IT-Systemen und ist ein wesentlicher Treiber für ILM.

Grund für dieses Verhalten ist die Verteilung der Zugriffe auf die Dateien. Mit abnehmender Wahrscheinlichkeit für weitere Zugriffe sinkt die Verteilungsfunktion langsamer ab. Nach diesem Prinzip bildet sich auch die Populationsgröße auf den Einzelebenen ab, wenn die Schwellwertwahrscheinlichkeiten äquidistant sind.

Die optimale Anzahl der Speicherebenen ist natürlich von den Anforderungen der Geschäftsprozesse an die Speicherung abhängig. Insgesamt jedoch erscheint eine Anzahl von zwei bis maximal drei Speicherhierarchien für normale Anwendungsfälle als empfehlenswert [113, 118].

7.2.3 Rekombination von Hierarchien

Zusätzlich zur passenden Anzahl von Hierarchien wurden nun die Auswirkungen diskutiert, wenn die Anzahl der Speicherebenen im laufenden Betrieb variiert würde. Dieses wird analytisch betrachtet basierend auf den simulativen Ergebnissen aus Abschnitt 7.2.2.

Es kommen drei Fälle von Rekombinationen der Speicherklasseninfrastruktur in Frage. Zum einen können bestehende Hierarchien zusammengefasst werden. Des Weiteren können neue Ebenen hinzugefügt werden. Schließlich kann es auch zu einer Verschiebung der Schwellwertwahrscheinlichkeiten kommen bei gleich bleibender Anzahl der Speicherhierarchien. Unabhängig davon, welcher Fall im realen Betrieb auftritt, sind für das Verhalten von ILM die Schwellwertwahrscheinlichkeiten ausschlaggebend. Dies soll an zwei Beispielen verdeutlicht werden:

Beispiel 1: Der Betreiber eines ILM-Systems mit zwei Speicherebenen entscheidet sich dazu eine der beiden Speicherebenen stillzulegen. In einem solchen Fall kann einfach die Schwellwertwahrscheinlichkeit auf einen Wert kleiner 0% bzw. größer 100% geändert werden. Infolge der Änderung kommt es zur Konzentration sämtlicher Dateien auf eine der beiden Ebenen, so dass die andere Ebene leer stillgelegt werden kann.

Beispiel 2: Der Betreiber eines ILM-Systems mit zwei Speicherebenen mit Schwellwert 10% entschließt sich dazu, eine zusätzliche Ebene einzufügen. Die Schwellwerte sollen zukünftig bei 5% und 2,5% liegen. Nun stellt sich die Frage, wie die Struktur des relativen Speicherbedarfs nach der Umstellung aussieht.

Mit Hilfe einer logischen Analyse und den bisherigen Untersuchungsergebnissen lässt sich diese Frage beantworteten. Dateien, die zum Zeitpunkt der Umstellung eine Zugriffswahrscheinlichkeit zwischen 5% und 10% besitzen, werden auf die erste Speicherebene migriert. Ebenso werden alle Dateien mit einer Zugriffswahrscheinlichkeit zwischen 0 und 2,5% der dritten Hierarchieebene zugeordnet. Lediglich die Dateien, die eine Wahrscheinlichkeit größer 10% oder zwischen 2,5% und 5% haben, bleiben ihren ursprünglichen Speicherklassen zugeordnet und werden nicht migriert. Mit Hilfe der Simulationsergebnisse aus Abbildung 31 lässt sich dieses Beispiel quantifizieren. Vor der Rekombination sind auf beiden Ebenen jeweils ca. 50% der Daten gespeichert. Nach der Rekombination sind auf der ersten Hierarchie rund 67%

des Gesamtdatenbestands gespeichert. Die zweite Ebene würde etwa 13% und die dritte Ebene ungefähr 20% der Dateien aufbewahren.

Ferner lässt sich aus den Untersuchungen zur Dynamik von ILM-Szenarien in Kapitel 7.1 ableiten, dass es beim Einsetzen neuer Speicherebenen zu keinen unerwünschten Einschwingeffekten kommt [4].

7.2.4 Zusammenfassung

In diesem Kapitel wurden die Auswirkungen der Anzahl der verwendeten Speicherebenen in einem ILM-Szenario auf das Verhalten des Systems analysiert. Hierzu wurden vier Simulationen mit unterschiedlicher Speicherklasseninfrastruktur durchgeführt und die Effekte anhand einer Sensitivitätsanalyse überprüft. Dabei zeigte sich, dass für die Zuweisung der Informationsobjekte zu den Speicherklassen und somit auch für das Verhalten von ILM die Schwellwertwahrscheinlichkeit zwischen den Ebenen maßgeblich ist. Nur wenn sich die Dimensionierung der Speicherinfrastruktur bzw. der Anzahl der Speicherebenen auf die Schwellwerte auswirkt, ist die Anzahl der Speicherebenen relevant für das Verhalten von ILM.

Solche Auswirkungen können insbesondere bei der Rekombination von Speicherhierarchien vorkommen. Theoretische Analysen haben gezeigt, dass eine Veränderung der Speicherstruktur im laufenden Betrieb durchgeführt werden kann, ohne dass ein Hin- und Hermigrieren der Dokumente zwischen den Speicherebenen ausgelöst wird. Demzufolge kann eine problemlose Überführung vom Infrastrukturausgangszustand in den gewünschten Endzustand gewährleistet werden.

Als weitere Erkenntnis aus diesem Kapitel wird festgehalten, dass die Zuverlässigkeit unabhängig von der Anzahl der Ebenen ist. Zusätzlich wurde anhand von logischen Annahmen die notwendige bzw. sinnvolle Anzahl der Speicherhierarchien für die meisten regulären ILM-Anwendungsfälle hergeleitet. Üblicherweise sind zwei bzw. drei Ebenen angemessen. Diese Erkenntnis dient ausdrücklich lediglich als Richtwert und entbehrt nicht der notwendigen Überprüfung der Anforderungen im individuellen Implementierungsfall.

7.3 Anwendungsszenario

Der vorgestellte ILM-Prozess mit seinen fünf Phasen stellt die allgemeine Vorgehensweise dar. Für die konkrete Umsetzung eines ILM-Szenarios bedarf es, wie ebenfalls beschrieben, einer Konzeption. Diese klärt neben verschiedenen Punkten insbesondere die Anzahl der ILM-Hierarchien, den avisierten Zeithorizont des Projektes und auch die zu verwendenden Migrationsregeln. In den Abschnitten zuvor wurde der zeitliche Verlauf von ILM-Szenarien betrachtet und ebenso die Thematik der Anzahl an Hierarchien beleuchtet. In diesem Abschnitt werden nun Aspekte zur Entscheidung über Migrationregeln betrachtet. An einem realen Beispiel wird die Entscheidungsfindung unter Verwendung von Simulationsergebnissen durchgeführt.

7.3.1 Rationales Entscheiden

Ob eine Entscheidung richtig oder falsch ist, zeigt stets die Zukunft. Wichtig ist, rational zu entscheiden. Dadurch lassen sich zumindest Fehler beim Entscheidungsvorgang vermeiden.

Allgemein lässt sich unabhängig von der Entscheidungssituation die Entscheidungsfindung in sechs Schritte aufteilen [19]:

1. Schritt: Definition des Entscheidungsproblems

2. Schritt: Auflistung der Alternativen

3. Schritt: Identifikation der möglichen Auswirkungen jeder Alternative bezüglich festgelegter Entscheidungskriterien

4. Schritt: Auflistung des Nutzens sämtlicher Alternativen

5. Schritt: Auswahl eines mathematischen Modells der Entscheidungstheorie

6. Schritt: Anwendung des Entscheidungsmodells und Treffen der Entscheidung

Nachfolgend werden diese Schritte vollzogen.

7.3.2 Ausgangssituation und Entscheidungsproblem

Ein IT-Manager hat in seinem Unternehmen Potential für eine ILM-Lösung identifiziert. Er hat sich beraten lassen und sich für eine ILM-Lösung mit 3 Hierarchien entscheiden. Weiterhin ist sein zeitlicher Horizont, über den die Lösung ihre Leistungsfähigkeit beweisen soll, mit 1000 Tagen, also ca. 3 Jahren, festgelegt.

Die Basis des Vergleichs bildet ein Auszug des Datenbestands des Unternehmens. Es handelt sich hierbei um 10000 Dateien mit einem Gesamtvolumen von 5,529 GB.

Der IT-Manager befindet sich in der Phase Sozialisierung. Sein Bedarf und sein TCO-Rahmen bringen ihn dazu, sich für die Festlegung der Migrationsregel im ILM-Konzept für eine Alternative entscheiden zu müssen.

Das Entscheidungsproblem lautet: Welche Migrationsregel ist die beste?

7.3.3 Migrationsregeln

Wie in Kapitel 4 dargestellt existieren verschiedene Methoden der Wertzuweisung von Dateien für ILM. Die wenigsten eignen sich für die automatisierte Wertzuweisung, die für ILM als Hauptpunkt identifiziert wurde.

Die in dieser Arbeit vorgestellte Methode eignet sich für ILM. Aus diesem Grund wird diese Methode mit vier anderen verglichen. Die ersten zwei sind sehr einfache Migrationregeln, die nicht wirklich mit ILM vereinbar sind. Ihre Betrachtung als mögliche Alternativen ist dennoch wichtig, weil sie in der Realität sehr wohl Optionen sind und ernsthaft diskutiert werden müssen. Migrationsregel 1 (M1) lautet:

Definition Migrationsregel 1 (M1): *Alle Dateien werden auf der obersten Hierarchie gespeichert.*

Die Anwendung dieser Regel bedeutet, dass kein tatsächliches ILM betrieben wird, weil nur die Hierarchie genutzt wird. Diese Option steht dafür, dass ein existierendes monolithisches System weiter genutzt wird.

Migrationsregel 2 (M2) ist sinngemäß das Gegenteil von M1. M2 lautet:

Definition Migrationsregel 2 (M2): *Alle Dateien werden auf der untersten Hierarchie gespeichert. Erfolgt ein Zugriff, so wird die betroffene Datei für 1 Tag auf der obersten Hierarchie gespeichert. Danach wird sie wieder auf der untersten Hierarchie gespeichert.*

Wie M1 ist M2 sehr trivial. Als Entscheidungsoption legt M2 fest, wie sich eine fast ausschließliche Nutzung der kostengünstigsten Hierarchie auf den Nutzen auswirkt.

Migrationsregel 3 (M3) ist keine reale Migrationsregel und funktioniert nur im Labor. Sie basiert auf perfekter Antizipation von Dateizugriffen. Sie lautet:

Definition Migrationsregel 3 (M3): *Alle Dateien werden auf der untersten Hierarchie gespeichert. Wenn auf eine Datei am nächsten Tag zugegriffen wird, wird sie tags zuvor auf die oberste Hierarchie migriert (perfekte Antizipation).*

Ähnlichkeiten zu M2 sind klar vorhanden. M3 wartet mit dem Vorteil auf, dass die jeweils auf die Datei zugreifende Person nicht langen Zugriffszeiten ausgesetzt ist.

Die meist angewandte Migrationsregel ist die zeit-basierte Migration. Hierbei werden feste Zeitintervalle definiert, die die Migrationsaktivitäten festlegen. Die bekannteste dieser Regel ist die bereits erwähnte Moore'sche Migrationsregel „90 Tage auf Enterprise Speicher - 90 Jahre auf Band" [63]. Wegen des festen Zeitintervalls bezeichnet man die zeit-basierten Migrationsregeln auch als statische Migrationsregeln,

Hier nun die Definition der zur Entscheidung stehenden zeit-basierten Migrationsregel.

Definition Migrationsregel 4 (M4): *Eine Datei, auf die 90 Tage nicht zugegriffen wurde, wird auf die nächst niedrigere Hierarchie migriert.*

Im Gegensatz zu den statischen Migrationsregeln legt die vorgestellte Methode der Wahrscheinlichkeit zukünftiger Zugriffe Datei-individuelle Zeitpunkte fest, an denen eine Datei migriert wird. Die individuelle Wahrscheinlichkeit ändert sich täglich. Aus diesem Grund spricht man von einer dynamischen Migrationsregel.

Die zur Entscheidung stehende dynamische Migrationsregel lautet:

Definition Migrationsregel 5 (M5): *Eine Datei wird von Hierarchie 1 auf Hierarchie 2 migriert, wenn ihre Wahrscheinlichkeit zukünftiger Zugriffe unter 10% sinkt. Eine Datei wird weiterhin von Hierarchie 2 auf Hierarchie 3 migriert, wenn ihre Wahrscheinlichkeit zukünftiger Zugriffe unter 2,5% sinkt.*

Die Berechnung der Wahrscheinlichkeiten erfolgt nach der in Kapitel 5 beschriebenen Methode.

Die Migrationsregeln M1-M5 sind die dem IT-Manager zur Verfügung stehenden Alternativen, zwischen denen er sich zu entscheiden hat.

7.3.4 Die Entscheidungskriterien

Standardgemäß werden in der Industrie IT-Entscheidungen auf Basis von TCO Betrachtungen und OPEX Minimierung getroffen [113]. Diesem Vorgehen folgend gibt es verschiedene Kostenmodelle [57, 71, 113]. Derartige Modelle unterstützen Entscheidungen über IT-Architekturen, d.h. „ILM ja oder nein?“ oder „Welche Speicher-Technologien sollen eingesetzt werden?“. Unter Annahme, dass eine Entscheidung für ILM bereits gefallen ist, entscheiden wir hier nun über die Migrationsregeln.

Aus diesem Grund werden drei Kriterien betrachtet, die die Leistungsfähigkeit von Migrationsregeln charakterisieren. Das erste Kriterium sind die direkten Kosten, die durch das belegte Speichervolumen unter Anwendung der jeweiligen Migrationsregel entstehen.

Das zweite Kriterium ist der Jitter als Qualitätsindex der Migrationen. Das dritte Kriterium ist die durchschnittliche QoS (Quality of Service, Dienstgüte) pro Datei, die man erhält, wenn die jeweilige Migrationsregel angewandt wird.

Diese drei Kriterien werden von dem Simulator ausgegeben. Die Migrationsregeln wurden implementiert. Nun werden die Kriterien im Detail betrachtet und anschließend die Simulationsergebnisse zusammengefasst.

7.3.4.1 Direkte Kosten

Speicher-Volumina steigen stetig, obwohl IT-Abteilungen nicht immer steigende Budgets rechtfertigen können [70]. Es wird von diesen Abteilungen eine Gesamtaufstellung für die gesamten Speicherdienste erwartet, d.h. Hardware, Software, Infrastruktur und Personal zusammen. Diese Aufstellung soll möglichst konstant über Jahre bleiben – ein Prozentanteil des Gesamt IT-Volumens [70].

Unternehmen halten Ausschau nach besseren Methoden, die Speicherkapazitäten zu verwalten. Automatisierte Dateimigration ist ein Hauptpunkt von ILM und 90% von IT-Entscheidern erwägen ILM [49]. IT ist ständiges Ziel von Kostenreduktionen und soll dennoch unternehmenskritische Services weiterhin bereitstellen. Das Argument hierzu ist der Preisverfall. Und tatsächlich unterliegen im Speicherbereich alle Technologien einem stetigen Preisverfall.

2006 hat Sun Microsystems festgestellt, dass die Preise jährlich um ca. 35% fallen. Die Preise für Enterprise-Speicherplatten rangieren zwischen 25 USD und 40 USD pro GB. Mittelklasse-Platten kosteten demnach zwischen 12 USD und 20 USD pro GB. Standardplatten sind für 3 USD bis 10 USD pro GB zu haben. Die Preise für Kapazitätsband liegen bei 1 USD bis 2 USD pro GB. Wissend, dass diese Preise mittlerweile hinfällig sind, ist ihre Relation beinahe konstant über die Jahre geblieben [101]. Das Verhältnis zwischen Enterprise-Speicherplatten

(FC, SCSI, FICON, ESCON) und Mittelklasse-Platten (SCSI, FC) ist 2:1. Die Relation zwischen Mittelklasse-Platten und Standardplatten (S-ATA) ist 3:1 und zwischen Standardplatten und Kapazitätsband ist 5:1 [101].

Horizon Information Strategies hat schon 2003 ähnliche Relationen konstatiert [62].

Das Verhältnis zwischen Enterprise-Speicherplatten (FC, SCSI, FICON, ESCON) und Mittelklasse-Platten (SCSI, FC) ist 2,5:1. Die Relation zwischen Mittelklasse-Platten und Standardplatten (S-ATA) ist 3:1 und zwischen Standardplatten und Kapazitätsband 6:1 [62].

Die Relationen scheinen sich also konstant zu entwickeln. Aus diesem Grund werden hier die Relationen zur Entscheidungsfindung herangezogen. Nichtsdestotrotz ist der Einzelpreis pro GB eine schwache Metrik für Speicherentscheidungen, wie nachfolgend auch dargestellt wird.

Angenommen die Relationen zwischen Hierarchien H1, H2 und H3 sind:

H1 = 1 Kosteneinheit/MB, H2 = 1/6*H1, H3 = 1/5*H2 [101]. Folglich können die direkten Kosten aus den Simulationsergebnissen ermittelt werden (siehe Tabelle 30).

Migrationsregel	M1	M2	M3	M4	M5
Direkte Kosten	5529	237	237	1122	1746

Tabelle 30: Direkte Kosten je Migrationsregel

Wohl wissend, dass Kosten der wichtigste Punkte bei IT-Entscheidungen sind [91], erkennt man, dass die triviale Migrationregel M2 eine tatsächliche Option ist. Das Problem mit M2 ist, dass die Datei erst in dem Moment hochgeladen wird, wenn sie aufgerufen wird. Daraus entstehen für den Nutzer Ladezeiten.

7.3.4.2 Jitter

Der durchschnittliche Jitter $\overline{\theta(t)}$ ist speziell vor dem Hintergrund einer wertgerechten Speicherung der Informationen eine wichtige Messgröße (siehe Abschnitt 6.7). Sie dient als Indikator für die Zuverlässigkeit der bereits ausgeführten Migrationen. Der Begriff Jitter wird aus der wiederholten Hin- und Herbewegung der Dateien zwischen den Speicherebenen abgeleitet. Jeweils im Fall einer Rückmigration erhöht sich der Jitter. Wissend, dass in diesem Fall ein Datei aufgerufen wurde, die nicht auf der obersten Ebene lag und demnach Ladezeiten erzeugt, ist ein hoher Jitterwert nicht im Interesse von IT-Managern, die auch den Service verantworten. Tabelle 31 zeigt die Simulationsergebnisse hinsichtlich des durchschnittlichen Jitters.

Migrationsregel	M1	M2	M3	M4	M5
Jitter	0	6,97	0	0,5	0,3

Tabelle 31: Simulierter durchschnittlicher Jitter pro Migrationsregel

Man erkennt, dass M2 den größten Jitter erzeugt. M3 hat per Definition keinen Jitter. M4 und M5 haben relativ geringe Jitter-Werte.

7.3.4.3 Dienstgüte (QoS)

In einem Unternehmen haben unterschiedliche Speicherebenen unterschiedliche Dienstgüten (Quality of Service). Dies ist begründet in unterschiedlichen Dienstgütevereinbarungen (Service Level Agreements, SLAs) und führt zu unterschiedlichen Verfügbarkeiten und Wiederherstellungszeiten der einzelnen Speicherebenen. Auch dieser Punkt beschäftigt IT-Manager sehr und wird in Form der Verfügbarkeit in der Entscheidungsfindung repräsentiert.

Unter der Annahme, dass unterschiedliche SLA (Service Level Agreements) für die drei Hierarchien existieren, seien die Verfügbarkeiten wie folgt:

Verfügbarkeit H1 = 0,99, Verfügbarkeit H2 = 0,97 und Verfügbarkeit H3 = 0,93.

Die Dienstgüte (QoS) wird gemäß nachfolgender Formel ermittelt:

$$QoS = \frac{1}{10000} \sum_{i=1}^{3} (Verfügbarkeit\ H_i \cdot durchschnittliche\ Anzahl\ von\ Dateien\ auf\ H_i) \qquad (34)$$

Die Simulationen ergeben folgende Resultate hinsichtlich der durchschnittlichen Anzahl Dateien auf den einzelnen Hierarchien (siehe Tabelle 32):

Migrationsregel	M1	M2	M3	M4	M5
Durchschnittliche Anzahl von Dateien auf Hierarchie 1	10000	141	141	1921	2696
Durchschnittliche Anzahl von Dateien auf Hierarchie 2	0	0	0	1195	2607
Durchschnittliche Anzahl von Dateien auf Hierarchie 3	0	9859	9859	6884	4697

Tabelle 32: Simulierte durchschnittliche Anzahl von Dateien pro Hierarchie

Daraus ergeben sich folgende durchschnittliche Verfügbarkeiten pro Migrationsregel:

Migrationsregel	M1	M2	M3	M4	M5
QoS	0,99	0,931	0,931	0,942	0,957

Tabelle 33: Simulierte durchschnittliche Verfügbarkeiten

7.3.4.4 Zusammenfassung der Kriterien

Zum Abschluss des 3. Schritts zur Entscheidungsfindung werden die einzelnen Entscheidungskriterien zusammengefasst.

Migrationsregel	M1	M2	M3	M4	M5
Direkte Kosten	5529	237	237	1122	1747
Jitter	0	6,97	0	0,5	0,3
Dienstgüte (QoS)	0,99	0,931	0,931	0,946	0,957

Tabelle 34: Zusammenfassung der Simulationsergebnisse je Migrationsregel

Nun liegen die Simulationsergebnisse vor und somit ist Schritt 3 der Entscheidungsfindung abgeschlossen.

7.3.5 Entscheidungsfindung

Im 4. Schritt wird eine Nutzenmatrix für alle Kriterien und Alternativen erstellt. Dazu wird eine Nutzenfunktion $\Phi(M_i)$ definiert.

Im 5. Schritt wird das Entscheidungsmodell gewählt. Hier kommt das Model DMUC (Decision-Making Under Certainty) mit multiplen Zielen zum Einsatz [19]. In diesem Umfeld wissen Entscheider mit Sicherheit die Konsequenzen jeder einzelnen Alternative. Die Wahl fällt dann auf die Alternative, welche den größten Nutzen generiert, womit schon Schritt 6 beschrieben ist.

Nutzenfunktion $\Phi(M_i)$ sei wie folgt definiert:

$$\Phi(M_i) = \sum_{p=1}^{r} g_p \cdot u_{ip} \tag{35}$$

Mit $\Phi(M_i)$ = Nutzen der Alternative M_i

r = Anzahl der Kriterien

g_p = Gewicht des Kriteriums p

u_{ip} = Nutzen von M_i bezüglich Kriterium p

Die Wichtung hat großen Einfluss auf das Entscheidungsergebnis. Hier werden die Gewichte beinahe gleichmäßig verteilt mit einer leichten Übergewichtung der direkten Kosten. Die Gewichte lauten:

$g_1 = 0{,}4$, $g_2 = 0{,}3$ und $g_3 = 0{,}3$ [22]

Bei der Zusammenfassung der drei Kriterien muss man diese noch normieren, um die Signifikanz der einzelnen Kriterien zu gewährleisten [19]. Der beste Wert je Kriterium wird auf 1 normiert, der schlechteste jeweils auf 0.

Dieses Vorgehen führt zu folgender normierten Entscheidungstabelle (siehe Tabelle 35)

Ziel	Direkte Kosten		Jitter		QoS		Nutzen	Rang
Gewicht	0,4	$g_1 \cdot u_{i1}$	0,3	$g_2 \cdot u_{i2}$	0,3	$g_3 \cdot u_{i3}$	$\Phi(M_i)$	
Migrationsregel								
M1								
	0	0	1	0.3	1	0.3	**0.6**	**4**

[22] IT-Manager müssen die Gewichte hinsichtlich der spezifischen Anforderung der jeweiligen Speicherumgebung anpassen. Die logischen Entscheidungsschritte bleiben davon unberührt.

M2	1	0.4	0	0	0	0	**0.4**	**5**
M3	1	0.4	1	0.3	0	0	**0.7**	**2**
M4	0.816	0.326	0.93	0.279	0.254	0.076	**0.682**	**3**
M5	0.715	0.286	0.96	0.288	0.44	0.132	**0.706**	**1**

Tabelle 35: Normierte Entscheidungstabelle

Durch Anwendung des Entscheidungsmodells DMUC (Decision-Making Under Certainty) mit multiplen Zielen fällt die Entscheidung über die beste Alternative der Migrationsregeln. Dieses ist Migrationsregel M5 und die zugrunde liegende Methode der Wahrscheinlichkeiten zukünftiger Zugriffe.

7.3.6 Zusammenfassung des Anwendungsszenarios

ILM ist eine viel versprechende Strategie. Um identifizierte Potenziale zu heben, ist ein breiteres Wissen über die ILM-Prozeduren notwendig. Es gibt wenig Erfahrungsberichte und das Sammeln eigener Erfahrung an realen Systemen ist sehr aufwändig. Insbesondere, wenn man bedenkt, dass laut Techtarget 66% der IT-Manager keine Zeit haben, rudimentäre Kosten- oder Datenbewertungsmodelle zusammenzustellen [25], zeigt sich der Nutzen der Simulation. Integriert in einen Prozess des rationalen Entscheidens erhalten Entscheider eine immense Unterstützung durch die Simulationen.

ILM braucht leistungsfähige Migrationsregeln. IT-Manager müssen über diese Regeln entscheiden. In dem gezeigten realen Beispiel lautete die Frage: „Wie kann die beste Alternative identifiziert werden?“ und „Welches ist die beste Alternative?“

Es wurden fünf Migrationsregeln in den Simulator implementiert und auf ihre Leistungsfähigkeit untersucht. In einem 1000 Tage ILM-Szenario mit 3 Hierarchien und 10000 Dateien wurden die Kriterien Kosten, Jitter und Dienstgüte ausgewertet. Mittels Nutzenfunktion wurde die Entscheidung getroffen.

Die Ergebnisse sind:

Methode M1 ist keine Option für ILM. Die generierten Kosten offenbaren die obere Grenze und stehen für ein Beibehalten existierender monolithischer Strukturen.

Die naive Methode M2 generiert die geringsten direkten Kosten und repräsentiert demnach die untere Kostengrenze. Dadurch, dass der Jitter mehr als zwanzigmal höher ist als bei M5 und dadurch, dass nur zwei Hierarchien genutzt werden, ist M2 nicht für ILM empfehlenswert.

Die Methode M3 ist per Definition von theoretischer Natur. Im konkreten Entscheidungsszenario ist M3 überraschend nicht die beste Alternative.

Migrationsregel M4 mit der zeit-basierten Migration nach Moore ist eine echte Alternative. Mit langer Tradition aus HSM lässt sich dieser einfache Ansatz auch für ILM nutzen. M4

funktioniert mit mehreren Hierarchien im Gegensatz zu M1 bis M3. Die gezeigte Leistungsfähigkeit führte im Beispiel zu einem achtbaren dritten Rang.

Migrationsregel M5 mit ihrer ausdrücklichen Ausrichtung auf ILM-Bewertung und dynamischer Wertzuweisung auf Basis mehrerer Dateieigenschaften inklusive Zugriffshistorie zeigte die beste Leistungsfähigkeit im Beispiel.

Allgemein lässt sich sagen, dass rationales Entscheiden über IT-Architekturen wichtig ist. Unabhängig von der spezifischen Situation ist dadurch der größte Erfolg zu erzielen. Der Aufwand des Entscheidens ist unter Zuhilfenahme des Simulators gering und die Aussagen rational begründbar. Dies macht eine Argumentation innerhalb eines Unternehmens über die Entscheidung einfach.

8 Zusammenfassung und Ausblick

Dieses Kapitel fasst die wesentlichen Ergebnisse der Arbeit zusammen und diskutiert im Ausblick Anknüpfungspunkte für zukünftige, forschungsrelevante Fragestellungen im Umfeld von ILM.

8.1 Zusammenfassung

Information Lifecycle Management (ILM) als Speicher Management-Konzept, welches Informationen automatisch entsprechend ihres Wertes auf dem jeweils kostengünstigsten Speichermedium speichert, erfährt großes Interessen unter IT-Managern [49, 58]. Aktuell fehlt es noch an Erfahrungen mit ILM, um ILM beurteilen zu können. Insbesondere die Wertzuweisung und Klassifizierung von Dateien ist ein offenes Thema, weil existierende Methoden aufwendig anzuwenden sind. Entscheidend für den weiteren Erfolg ist eine wirkungsvolle Wertzuweisungsmethode.

In dieser Arbeit wurden aufgrund realer Unternehmensdaten der Siemens AG Erkenntnisse über ILM entwickelt. Ein Dateienpool mit Daten über mehr als 70.000 Dateien stand als Auswertungsbasis zur Verfügung. Diese Arbeit befasste sich mit der Wertzuweisung von Dateien und behebt das Defizit existierender Methoden, dass diese keine Prognosen über die zukünftige Verwendung von Dateien liefern. Dazu wurde eine neuartige, statistische Methode entwickelt und ihre Leistungsfähigkeit belegt. Diese Methode wurde in den grundsätzlichen Kontext eines IT-Projekts eingebunden, wofür ein Vorgehensmodell entwickelt wurde. Dieses besteht aus fünf Phasen: „Erfassung“, „Sozialisierung“, Klassifizierung“, „Automatisierung“ und „Überprüfung“.

Die erste Forschungsfrage lautete: „Existiert Potenzial für ILM?“. Diese Frage repräsentiert die Phase der „Erfassung“, die die konkrete Ist-Situation eines Unternehmens erfasst und das mögliche Einsparungspotenzial identifiziert. Hierzu wurde in einer Fallstudie an einem konkreten Beispiel der Siemens AG das Potenzial resultierend aus der Identifikation ungenutzter Dateien quantifiziert. Fast 90% der Dateien erfuhren 90 Tage nach ihrer Erstellung keine Zugriffe mehr. Dieses Ergebnis begründete das weitere Vorgehen hinsichtlich des Einsatzes von ILM.

Die zweite Forschungsfrage lautete: „Wie kann man das identifizierte Potenzial nutzbar machen?“. In der Phase „Sozialisierung“ wurde die allgemeine Erstellung eines ILM-Konzepts als Entscheidungsgrundlage erläutert. Innerhalb der ILM-Konzeption werden die unternehmensspezifischen Aspekte hinsichtlich der ILM-Lösung zusammengefasst. Anhand eines spezifischen Szenarios wurde der Aspekt der Entscheidung über Migrationsregeln in seiner generellen Struktur vorgestellt und unternehmensspezifisch durchgeführt.

Zur Beantwortung der dritten Forschungsfrage „Wie können Dateien auf Basis von automatisierten Prognosen bewertet werden?“ wurden existierende Verfahren untersucht. Dabei offenbarte sich insbesondere das Defizit existierender Methoden, dass keine Prognosen über zukünftige Zugriffe erstellt werden. Ohne Prognosen werden Dateien anhand anderer Faktoren, wie z.B. Administratorenwissen, bewertet. Dadurch ist die Bewertung erstens aufwendig und

zweitens schwer automatisierbar. Deshalb wurde eine Methode entwickelt, die Prognosen erstellt und damit das identifizierte Defizit existierender Methoden abstellt. Damit kann die Phase „Klassifizierung“ durchgeführt werden.

Als letzte Forschungsfrage wurde eine Frage beleuchtet, die die Idee der „best practices“ verfolgt. Sie lautete: Ist eine derartige Methode in ILM verwendbar? Die Frage prüft die Eignung der Methode für die Phase „Automatisierung“. Dazu wurde ein Simulator implementiert und ein Dateipool der Siemens AG mit über 70.000 Dateien genutzt. Die Auswertungen zeigten, dass die hier vorgestellte Methode der Wahrscheinlichkeit zukünftiger Zugriffe sowohl auf Dateibasis als auch auf aggregierter Hierarchiebasis die ILM-Prozeduren umsetzt. In einem Anwendungsbeispiel setzte sich diese Methode hinsichtlich der Leistungsfähigkeit insbesondere gegen die weit verbreitete zeitbasierte Regel durch.

Die Phase der „Überprüfung“, welche die Anpassung der Migrationsregeln auf sich ändernde Situationsumgebungen wie z.B. Nutzungsmuster oder Applikationen sicherstellen soll, schließt sich logisch an, wurde in dieser Arbeit aber nicht weiter betrachtet. Dazu bedarf es einer implementierten Lösung, die im Rahmen dieser Arbeit logischerweise nicht vorlag.

8.2 Ausblick

Das vorgestellte ILM-Framework (Abbildung 2) beinhaltet weitere, im ILM-Umfeld relevante Themengebiete, die Gegenstand zukünftiger Forschungsvorhaben sein können. Ziel sollte es sein, zukünftigen Applikationen eine Schnittstelle zum ILM-Framework zu gewähren. Einige dieser Themengebiete werden im Folgenden kurz skizziert.

Das Gebiet der Business Processes, das grundlegend für ILM ist, umfasst mehrere Bereiche, die über die Speicherung von Dateien hinausgehen. Hier lassen sich hinsichtlich der Gesamt-IT Ansätze aus der Forschung von Service-orientierten Architekturen (SOA) einsetzen. Das SOA-Paradigma gewinnt im Zusammenhang mit der Überwindung von Komplexität in heterogenen Anwendungslandschaften und der architekturellen Unterstützung unternehmensübergreifender Workflows zunehmend an Bedeutung. Dies gilt im besonderen Maße für Web Services als Technologie zur Umsetzung des SOA-Paradigmas [6, 21, 77, 78, 89]. Eine Forschungsfrage hierzu könnte lauten: „Wie lässt sich ILM in die Prozess-Ketten von Web Services abbilden?“.

Bezüglich der Netzinfrastruktur lassen sich zwei Ansätze für ILM-Vorhaben weiter untersuchen. Zum einen ist der Grid-Ansatz in Speicherstrategien eine viel versprechende Überlegung [26]. Möglichkeiten zur Realisierung von Grid Services beschreibt der *OGSA (Open Grid Service Architecture)*-Ansatz [66].Weiterhin stellt das Peer-to-Peer-Paradigma [55, 97] eine Alternative dar. Unter einem Peer-to-Peer-System versteht man ein sich selbst organisierendes, dezentrales System gleichberechtigter, autonomer Entitäten (Peers). Peer-to-Peer-Systeme operieren vorzugsweise ohne Nutzung zentraler Dienste auf der Basis eines Rechnernetzes mit dem Ziel der gegenseitigen Nutzung von Ressourcen [67, 98]. Das Peer-to-Peer-Paradigma wird auch bereits im Zusammenhang mit den oben genannten Service-orientierten Architekturen untersucht [27, 39, 121, 124]. Im Rahmen der Venice-Service-

Grid-Architektur [36, 37, 38] wird ein Verzeichnisdienst für Services umgesetzt. Diese Architektur bietet verschiedene Voice-Dienste sowie Mehrwertdienste als Web Services im Rahmen eines Service Grids an. Die zugehörige Forschungsfrage könnte lauten: „Kann ein Speicherservice auf ILM-Basis in einen Voice Service Grid zur Sprachaufzeichnung genutzt werden?

9 Literaturverzeichnis

[1] Abd-El-Malek, M., Courtright II, W. V., Cranor, C., Ganger, G. R., Hendricks, J., Klosterman, A. J., Mesnier, M., Prasad, M., Salmon, B., Sambasivan, R. R., Sinnamohideen, S., Strunk, J. D., Thereska, E., Wachs, M., Wylie, J. J.: *Ursa Minor: versatile cluster-based storage*. In: 4th USENIX Conference on File and Storage Technologies, San Francisco, 2005, S. 59-72.

[2] Abd-El-Malek, M., Courtright II, W. V., Cranor, C., Ganger, G. R., Hendricks, J., Klosterman, A. J., Mesnier, M., Prasad, M., Salmon, B., Sambasivan, R. R., Sinnamohideen, S., Strunk, J. D., Thereska, E., Wachs, M., Wylie, J. J.: *Early experiences on the journey towards self-* storage*. In: Data Engineering Bulletin, 29 (2006) 4, S. 55-62.

[3] Allen, N.: *Don't waste your storage dollars: what you need to know*. Research Note, Gartner Group Inc., Stamford, 2001.

[4] Behrens, M.: *Simulation von Information-Lifecycle-Management-Szenarien*. Diplomarbeit, TU Darmstadt, 2006.

[5] Bennett, J. M., Bauer, M. A., Kinchlea, D.: *Characteristics of Files in NFS Environments.* In: 1991 ACM Symposium on Small Systems, 1991, S. 33-40.

[6] Berbner, R., Grollius, T., Repp, N., Eckert, J., Heckmann, O., Ortner, E., Steinmetz, R.: *Management of Service-oriented Architecture (SoA) based Application Systems*. In: Enterprise Modelling and Information Systems Architectures - An International Journal, 2 (2007) 1, S. 14-25.

[7] Bhagwan, R., Douglis, F., Hildrum, K., Kephart, J. O., Walsh, W. E.: *Time-varying Management of Data Storage*. In: First Workshop on Hot Topics in Systems Dependability, Yokohama, 2005, S. 222-232.

[8] Born, S., Ehmann, S., Hintemann, R., Kastenmüller, S., Schaupp, D., Stahl, H.-W.: *Leitfaden zum Thema „Information Lifecycle Management"*. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM), 2004.

[9] Brien, J.: *Kostendruck auf CIOs: IT-Berater im Aufwind*. http://www.pressetext.at/pte.mc?pte=070524007, 2008, Abruf am 17.11.2008.

[10] CCITT: *CCITT Blue Book, VIII.2, Data Communication Networks: Services and Facilities, Interfaces. Band 2*. Genf, 1989.

[11] Chandra, S., Gehani, A., Yu, X.: *Automated Storage Reclamation Using Temporal Importance Annotations*. In: 27th International Conference on Distributed Computing Systems, Toronto, 2007, S 333-343.

[12] Chen, P.: *Optimal file allocation in multi-level storage hierarchies*. In: National Computer Conference and Exposition, 1973, S. 277-282.

[13] Chen, Y.: *Information valuation for Information Lifecycle Management*. In: The Second International Conference on Autonomic Computing (ICAC'05), 2005, S. 135-146.

[14] contentmanager.de: *EMC zeigt alle Bausteine des Information Lifecycle Managements*. http://www.contentmanager.de/magazin/news_h6591-print_emc_zeigt_alle_bausteine_des_information.html, 2005, Abruf am 17.11.2008.

[15] Dauen, S.: *Aufbewahrungspflichten.* Haufe-Verlag, München, 2004.

[16] Denning, P. J.: *The working set model for program behavior*. In: Communications of the ACM, 11(5) (1968), S.323-333.

[17] Denning, P. J.: *Working sets past and present*. In: IEEE Transactions on Software Engineering, SE-6(1):64-84 (1980) S. 64-84.

[18] documanager.de: *IT-Glossar - Hierarchisches Speichermanagement*. http://www.documanager.de/ressourcen/glossar_545_hierarchisches_speichermanagement.html, Abruf am 03.10. 2008.

[19] Domschke, W., Drexl, A.: *Einführung in Operations Research.* 4. Aufl., Springer, Berlin, Heidelberg, 1998.

[20] Douceur, J. R., Bolosky, William J.: *A Large-Scale Study of File-System Contents.* In: SIGMETRICS '99, Atlanta, 1999, S. 59-70.

[21] Eckert, J., Miede, A., Repp, N., Steinmetz, R.: *Resource Planning Heuristics for Service-oriented Workflows*. In: IEEE/WIC/ACM International Conference on Web Intelligence 2008, Sydney, 2008, S. 591-597.

[22] Effelsberg, W., Haerder, T.: *Principles of database buffer management*. In: ACM Transactions on Database Systems, 9(4) (1984), S. 560-595.

[23] Ellard, D., Mesnier, M., Thereska, E., Ganger, G. R., Seltze, M.: *Attribute-Based File Prediction of File Properties.* Harvard Computer Science Technical Report TR-14-03, 2003.

[24] Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G.: *Statistik - Der Weg zur Datenanalyse.* Band 5. Aufl., Springer, Berlin u.a., 2004.

[25] Foskett, S.: *Best Practices – Belt tightening continues to shift budgets away from security and information management*. Glasshouse Technologies Inc., 2006.

[26] Foster, I., Kesselman, C.: *The Grid: Blueprint for a New Computing Infrastructure.* 2. Auflage, Elsevier, o. O., 2004.

[27] Friese, P., Müller, J. P., Freisleben, B.: *Integrating Peer-to-Peer Technology into a Web Service Environment*. In: Multikonferenz Wirtschaftsinformatik (MKWI 2006), Workshop on P2P and Grid Computing, Passau, 2006, S. 25-40.

[28] Gantz, J. F., Chute, Ch., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., Toncheva, A.: *The Diverse and Exploding Digital Universe - An Updated Forecast of Worldwide Information Growth Through 2011*. IDC White Paper, 2008.

[29] Gibson, T., Miller, E.: *An Improved Long-Term File-Usage Prediction Algorithm*. In: 25th Annual International Conference on Computer Measurement and Performance (CMG 99), Reno, 1999, S. 639-648.

[30] Gibson, T., Miller, E. L., Long, D. D. E.: *Long-term File Activity and Inter-Reference Patterns*. In: 24th International Conference on Technology Management and Performance Evaluation of Enterprise-Wide Information Systems, Computer Measurement Group, Anaheim, 1998, S. 92-103.

[31] Gray, J., Reuter, A.: *Transaction Processing: Concepts and Techniques.* Morgan Kaufmann, San Francisco, 1993.

[32] Hartung, J.: *Statistik - Lehr- und Handbuch der angewandten Statistik.* Band 10. Auflage, Oldenbourg, München u.a., 1995.

[33] Heckmann, O.: *A System-oriented Approach to Efficiency and Quality of Service for Internet Service Providers*. TU Darmstadt, Submitted Ph.D Thesis, 2004.

[34] Heckmann, O.: *The Competitive Internet Service Provider.* Wiley & Sons, 2006.

[35] Heike, H.-D., Târcolea, C.: *Grundlagen der Statistik und Wahrscheinlichkeitsrechnung - Statistik I.* Oldenbourg München u.a., 2000.

[36] Hillenbrand, M., Götze, J., Müller, P.: *Venice – A Lightweight Service Grid*. In: 32nd EUROMICRO Conference, Cavtat, 2006, S. 364-371.

[37] Hillenbrand, M., Götze, J., Müller, P.: *Web Services Directory based on Peer-to-Peer Technology*. In: 32nd EUROMICRO Conference, Cavtat, 2006, S.364-371.

[38] Hillenbrand, M., Götze, J., Zhang, G., Müller, P.: *A Lightweight Service Grid based on Web Services and Peer-to-Peer*. In: 15. ITG/GI-Fachtagung Kommunikation in Verteilten Systemen (KIVS 2007), Bern, 2007, S. 89-100.

[39] Hillenbrand, M., Müller, P.: *Web Services and Peer-to-Peer*. In: Steinmetz, R., Wehrle, K. (Hrsg.): Peer-to-Peer Systems and Applications, Springer, 2005.

[40] Iamnitchi, A., Ripeanu, M.: *Myth and Reality: Usage Behavior in a Large Data-Intensive Physics Project*. University of Chicago, Chicago, 2002.

[41] Jelenkovic, P. R., Radovanovic, A.: *Least-Recently-Used Caching with Dependent Requests*. In: Theoretical Computer Science, Vol. 326 (October 2004) Issues 1-3, S. 293-327.

[42] Johnson, C., Agrawal, R.: *Intersections of Law and Technology in Balancing Privacy Rights with Free Information Flow*. In: 4th International Conference on Law and Technology, Cambridge, 2006, S. 222-232.

[43] Kaarst-Brown, M. L., Kelly, S.: *IT-Governance and Sarbanes Oxley: The latest sales pitch or real challenges for the IT Function?* In: 38th Hawaii International Conference on System Sciences, Big Islands, 2005, S. 236-246.

[44] Kaiser, M. G., Smolnik, S., Riempp, G.: *Konzeption eines Information-Lifecycle-Management-Frameworks im Dokumenten-Management-Kontext*. In: Multikonferenz Wirtschaftsinformatik 2008 (MKWI 2008), München, 2008, S. 483-494.

[45] Kanakamedala, K., Kaplan, J., Srinivasaraghavan, R.: *A smarter approach to data storage*. In: The McKinsey Quarterly (2007), S. 1-4.

[46] Kastenmüller, S.: *Information Lifecycle Management*. Fujitsu Siemens Computers, Bad Homburg, 2006.

[47] Kircher, H.: *IT-Innovationen für Wachstum und Erfolg.* Kircher, H. (Hrsg.), Berlin/Heidelberg, 2007.

[48] Lawrie, D. H., Randal, J. M., Barton, R. R.: *Experiments with Automatic File Migration*. In: IEEE Computer, 15(7) (1982), S. 45-55.

[49] Linden, L.: *Storage Budget Survey 2006*. http://www.glasshouse.com/storage-budget-survey/registration.shtml, Glasshouse Technologies Inc., 2006, Abruf am 17.11.2008.

[50] Lyman, A., Varian, B.: *How Much Information?* University of California, Berkeley, 2003.

[51] Maier, R., Hädrich, T., Peinl, R.: *Enterprise Knowledge Infrastructures.* Springer-Verlag, Berlin, Heidelberg, New York, 2005.

[52] Masuhr, A.: *A.T. Kearney baut Beratungsbereich Strategisches IT-Management weiter aus*. http://www.atkearney.de/content/veroeffentlichungen/pressemitteilungen_detail.php/id/50237, 2008, Abruf am 17.11.2008.

[53] Mattern, F.: *Modellbildung und Simulation*. In: R. Wilhelm. (Hrsg.): Informatik - Grundlagen, Anwendungen, Perspektiven, C. H. Beck, München, 1996.

[54] Matthesius, M., Stelzer, D.: *Analyse und Vergleich von Konzepten zur automatisierten Informationsbewertung im Information Lifecycle Management.* In: Multikonferenz Wirtschaftsinformatik, München, 2008, S. 471-482.

[55] Mauthe, A., Hutchison, D.: *Peer-to-Peer Computing: Systems, Concepts and Characteristics*. In: Praxis in der Informationsverarbeitung & Kommunikation (PIK), 26 (2003) 2, S. 60-64.

[56] Mauthe, A., Thomas, P.: *Professional Content Management Systems – Handling Digital Media Assets.* John Wiley & Sons Ltd, 2004.

[57] Merrill, D.: *Storage Economics Identifying and Reducing Operating Expenses in the Storage Infrastructure*. Hitachi Data Systems, WHP-153-00, 2003.

[58] Merrin, R., Harnetty, D.: *Compliance Confusion Set to Drive Market in 2005*. Engenio Information Technologies, 2005.

[59] Mesnier, M., Thereska, E., Ganger, G. R., Ellard, D., Seltzer, M.: *File classification in self-* storage systems*. In: The First International Conference on Autonomic Computing (ICAC-04), New York, 2004, S. 44-51.

[60] Mont, M.: *On Privacy-aware Information Lifecycle Management in Enterprises: Setting the Context*. In: ISSE 2006 - Securing Electronic Business Processes, Wiesbaden, 2006, S. 405-415.

[61] Moody, D., Walsh, P.: *Measuring the value of information: An asset valuation approach*. In: The 7th European Conference on Information Systems, Copenhagen, 1999, S. 361-373.

[62] Moore, F.: *Storage - New Game New Rules.* Horison Information Strategies, Melbourne, 2003.

[63] Moore, F.: *Information Lifecycle Management.* Horison Information Strategies, Melbourne, 2004.

[64] Mullender, S. J., Tanenbaum, A. S.: *Immediate Files*. In: Software – Practice and Experience, Band 14 (4), 1984, S. 365-368.

[65] Näther, W., Stoyan, D.: *Elementare Stochastik - Skript zur Vorlesung.* Heidelberg, 2005.

[66] Open Grid Services Architecture Work Group: *Defining the Grid: A Roadmap for OGSA Standards.* http://www.gridforum.org/documents/GFD.53.pdf, 2005, Abruf am 17.11.2008.

[67] Oram, A.: *Peer-to-Peer: Harnessing the Power of Disruptive Technologies.* O'Reilly Media, 2001.

[68] Oßmann, K.: *Automatisierte Bewertung von Daten im SAP® Business Information Warehouse im Rahmen des Information Lifecycle Managements*. Technische Universität Ilmenau, 2008.

[69] Page, L., Brin, S., Motwani, R., Winograd, T.: *The pagerank citation ranking: Bringing order to the web.* Stanford University, 1999.

[70] Paquet, R.: *Poll Confirms Companies Aren't Ready for ILM*. Gartner Inc. #G00125956, 2005.

[71] Patel, C. D., Shah, A. J.: *Cost Model for Planning, Development and Operation of a Data Center*. Internet Systems and Storage Laboratory HP Laboratories Palo Alto, HPL-2005-107(R.1), 2005.

[72] Peterson, M.: *Information Lifecycle Management. A Vision for the Future*. http://www.snia.org/forums/dmf/programs/ilmi/ilm_docs/, Storage Networking Industry Association, 2004, Abruf am 17.11.2008.

[73] Peterson, M.: *ILM Definition and Scope - An ILM Framework.* http://www.snia.org/forums/dmf/programs/ilmi/ilm_docs/, Storage Networking Industry Association, 2004, Abruf am 17.11.2008.

[74] Peterson, M.: *Top Ten Pain Points – Survey Results*. http://www.snia.org/forums/dmf/programs/ilmi/ilm_docs/, SNIA End User Council, 2004, Abruf am 17.11.2008.

[75] Peterson, M., St. Pierre, E.: *Information Lifecycle Management Roadmap*. http://www.snia.org/forums/dmf/programs/ilmi/ilm_docs/, 2004, Abruf am 17.11.2008.

[76] Rényi, A.: *Wahrscheinlichkeitsrechnung.* Band 4. Auflage, VEB Deutscher Verlag der Wissenschaften, Berlin, 1973.

[77] Repp, N., Schulte, S., Eckert, J., Berbner, R., Steinmetz, R.: *Service-Inventur: Aufnahme und Bewertung eines Services-Bestands*. In: Workshop MDD, SOA und IT-Management (MSI 2007), Oldenburg, 2007, S. 13-22.

[78] Richter, J.-P., Haller, H., Schrey, P.: *Serviceorientierte Architektur*. In: Informatik Spektrum, 28 (2005) 5, S. 413-416.

[79] Ridings, R., Shishigin, M.: *PageRank Uncovered.* Technical Report, 2002.

[80] Roadknight, C., Marshall, I., Vearer, D.: *File Popularity Characterisation*. BT Research Laboratories, Suffolk, 1999.

[81] Robbe, B.: *SAN - Storage Area Network.* Hanser Fachbuchverlag, 2004.

[82] Robinson, J. T., Devarakonda, N. V.: *Data Cache Management Using Frequency based Replacement*. In: The 1990 ACM SIGMETRICS Conference, 1990, S. 134-142.

[83] Sachs, L.: *Angewandte Statistik - Anwendung statistischer Methoden.* Band 11. Auflage, Springer, Berlin u.a., 2004.

[84] Satyanarayanan, M.: *A Study of File Sizes and Functional Lifetimes*. In: Proceedings of the 8th ACM Symposium on Operating Systems Principles, 1981, S. 96-108.

[85] Sawitzki, G.: *Statistical Computing - Einführung in R.* Heidelberg, 2005.

[86] Schlittgen, R.: *Einführung in die Statistik - Analyse und Modellierung von Daten.* Band 10. Auflage, Oldenbourg, München u.a., 2003.

[87] Schmitt, J. B.: *Heterogeneous Network Quality of Service Systems*. Kluwer Academic Publishers, Boston, 2001.

[88] Schmitz, C.: *Entwicklung einer optimalen Migrationsstrategie für ein hierarchisches Datenmanagement System*. Forschungszentrum Jülich GmbH, Jülich, 2004.

[89] Schulte, S., Berbner, R., Steinmetz, R., Uslar, M.: *Implementing and evaluating the Common Information Model in a relational and RDF-based Database*. In: 3rd International ICSC Symposium on Information Technologies in Environmental Engineering (ITEE 2007), Oldenburg, 2007, S. 109-118.

[90] Shah, G., Voruganti, K., Shivam, P., Alvarez Rohena, M.: *ACE: Classification for Information Lifecycle Management.* IBM Almaden Research Center, 2006.

[91] Short, J. E.: *ILM Survey: What Storage, IT and Records Managers Say*. ISIC Research Report Vol. 06, No. RB06-02, Information Storage Research Center, University of California, Berkeley, 2006.

[92] Short, J. E.: *Information Lifecycle Management: An Analysis of End User Perspectives.* San Diego, 2006.

[93] Sienknecht, T. F., Friedrich, R. J., Martinka, J. J., Friedenbach, P. M.: *The Implications of Distributed Data in a Commercial Environment on the Design of Hierarchical Storage Management*. In: Performance Evaluation, 20 (1994) (1-3), S. 3-25.

[94] Smith, A. J.: *Long Term File Migration: Development and Evaluation of Algorithms*. In: Communications of ACM, 24(8), (1982), S. 521-532.

[95] Smith, K. A., Seltzer, M. I.: *File Layout and File System Performance*. Harvard University Technical Report TR-35-94, 1994.

[96] Steinmetz, R.: *Multimedia-Technologie: Grundlagen, Komponenten und Systeme.* 3. Aufl., Springer, Berlin, Heidelberg, 2000.

[97] Steinmetz, R., Wehrle, K.: *Peer-to-Peer-Networking & -Computing*. In: Informatik-Spektrum, 27 (2004) 1, S. 51-54.

[98] Steinmetz, R., Wehrle, K.: *Peer-to-Peer Systems and Applications.* Springer, 2005.

[99] Steinmüller, K.: *Methoden der Zukunftsforschung – Langfristorientierung als Ausgangspunkt für das Technologie-Roadmapping*. In: Möhrle, M. G., Isenmann, R. (Hrsg.): Technologie-Roadmapping - Zukunftsstrategien für Technologieunternehmen, Berlin/Heidelberg, 2005.

[100] Strange, S.: *Analysis of Long-term Unix File Access Patterns for Application to Automatic File Migration Strategies*. University of California, Berkeley, Technical Report UCB/CSD-92-700, EECS Department, 1992.

[101] SUN Microsystems Inc.: *Storage Optimization - ILM. March 2006*. Sun Microsystems, 2006.

[102] Tanaka, T., Ueda, R., Aizono, T., Ushijima, K., Naitih, I., Komoda, N.: *Proposal and Evaluation of Policy Description for Information Lifecycle Management.* In: International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05), Wien, 2005, S. 261-267.

[103] TecChannel.de: *SANs - Standards und Lösungen*. http://www.tecchannel.de/index.cfm?pid=207&pk=401636, 2008, Abruf am 03.10.2008.

[104] Thome, G., Sollbach W.: *Grundlagen und Modelle des Information Lifecycle Management.* Springer-Verlag, Berlin, Heidelberg, New York, 2007.

[105] Thommen, J.-P., Achleitner, A.-K.: *Allgemeine Betriebswirtschaftslehre.* Gabler Verlag, 2. Auflage, 1998.

[106] Turczyk, L. A.: *Information Lifecycle Management als Weg aus dem Speicherdilemma*. In: Information, Wissenschaft und Praxis Ausgabe 7-2004 (2004), S. 407-410.

[107] Turczyk, L. A.: *Ordnungssystem – Information Lifecycle Management*. In: Content Management Magazin, Ausgabe 4/ 2004 (2004), S. 35-36.

[108] Turczyk, L. A.: *Rumble in the jungle*. In: DOQ Magazin, Ausgabe 5/2004 (2004), S. 90-93.

[109] Turczyk, L. A.: *Wege aus dem zunehmenden Datendschungel*. In: ntz Nachrichtentechnische Zeitung, Ausgabe 11/ 2004 (2004), S. 42-43.

[110] Turczyk, L. A.: *Wenn der Sack voll ist – Information Lifecycle Management*. In: IT-Sicherheit – Management und Praxis, Ausgabe 4/ 2004 (2004), S. 57-58.

[111] Turczyk, L. A.: *ILM - Produkte sind zweitrangig – das Konzept macht's.* In: Speicherguide.de - Das Storage Magazin (2005).

[112] Turczyk, L. A.: *Information Lifecycle Management: Organisation ist wichtiger als Technologie*. In: Information, Wissenschaft und Praxis, Ausgabe 7-2005 (2005), S. 371-372.

[113] Turczyk, L. A., Behrens, M., Liebau, N., Steinmetz, R.: *Cost Impacts on Information Lifecycle Management Design*. In: 13th Americas Conference on Information Systems (AMCIS 2007), Keystone, 2007, S. 1110-1121.

[114] Turczyk, L. A., Frei, Ch., Liebau, N., Steinmetz, R.: *Eine Methode zur Wertzuweisung von Dateien in ILM*. In: Multikonferenz Wirtschaftsinformatik 2008 (MKWI 2008), München, 2008, S. 459-470.

[115] Turczyk, L. A., Gostner, R., Berbner, R., Heckmann, O., Steinmetz, R.: *Analyse von Datei-Zugriffen zur Potentialermittlung für Information Lifecycle Management.* Technische Universität Darmstadt, Fachgebiet KOM. Technical Report KOM 01-2005, 2005.

[116] Turczyk, L. A., Gröpl, M., Liebau, N., Steinmetz, R.: *A Method for File Valuation in Information Lifecycle Management.* In: 13th Americas Conference on Information Systems (AMCIS 2007), Keystone, 2007, S. 1122-1133.

[117] Turczyk, L. A., Heckmann, O., Berbner, R., Steinmetz, R.: *A Formal Approach to Information Lifecycle Management*. In: 17th Annual Information Resources Management Association Conference (IRMA 2006), Washington, 2006, S.531-533.

[118] Turczyk, L. A., Heckmann, O., Steinmetz, R.: *Simulation of Information Lifecycle Management*. In: 18th Annual Information Resources Management Association Conference (IRMA 2007), Vancouver, 2007, S. 1063-1066.

[119] Turczyk, L. A., Liebau, N., Steinmetz, R.: *Modeling Information Lifecycle Management.* In: 13th Americas Conference on Information Systems (AMCIS 2007), Keystone, 2007, S.1134-1146.

[120] Verma, A., Sharma, U., Rubas, J., Pease, D., Kaplan, M., Jain, R., Devarakonda, M., Beigi, M.: *Policy-Based Information Lifecycle Management in a Large-Scale File System*. In: The Sixth IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'05), Stockholm, 2005, S. 139-148.

[121] Verma, K., Sivashanmugam, K., Sheth, A., Patil, A., Oundhakar, S., Miller, J.: *METEOR-S WSDI: A Scalable P2P Infrastructure of Registries for Semantic Publication and Discovery of Web Services*. In: Journal of Information Technology and Management, 6 (2005) 1, S. 17-39.

[122] von Donop, T.: *Accenture-Studie: Deutsche Unternehmen schöpfen das Geschäftspotenzial ihrer Informationstechnologie nicht aus*. Accenture Deutschland, 2008.

[123] Wiedemann, M.: *Vergleichende Analyse und Konzeption von Storage Area Networks (SAN) für mittelständische Unternehmen*. University of Applied Sciences Cologne, 2005.

[124] Wombacher, A.: *Decentralized establishment of consistent, multi-lateral collaborations*. Dissertationsschrift, Technische Universität Darmstadt, Fachbereich Informatik, Darmstadt, 2005.

[125] Zadok, E.: *Reducing Storage Management Costs via Informed User-Based Policies*. In: 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies (MSST 2004), Maryland, 2004, S. 101-105.

10 Abkürzungsverzeichnis

ADM	**Automated Data Migration**
AO	**Abgabenordnung**
ATA	**Advanced Technology Attachments**
BaFin	**Bundesanstalt für Finanzdienstleistungsaufsicht**
CAS	**Content Addressable Storage**
CASSI	**Content Addressable Storage Solutions Initiative**
CD-RW	**Compact Disc-read write**
CIO	**Chief Information Officer**
CRM	**Customer Relationship Management**
DAX 30	**Deutscher Aktien Index**
DCGK	**Deutscher Corporate-Governance-Kodex**
DES	**Diskrete Ereignisorientierte Simulation**
DMF	**Daten-Management-Forum**
DPI	**Data Protection Initiative**
DVD-RW	**Digital Versatile Disc-read write**
EDB	**Erfahrungsdatenbank**
EUC	**End User Councils**
FC	**Fibre Channel**
FSA	**Financial Services Authority**
GDPdU	**Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen**
GoB	**Grundsätze der ordnungsgemäßen Buchführung**
HGB	**Handelsgesetzbuch**
HIPAA	**Health Insurance Portability and Accountability Act**
HSM	**Hierarchisches Speicher Management**

ILM	**Information Lifecycle Management**
kB	**kiloByte**
KonTraG	**Gesetz zur Kontrolle und Transparenz im Geschäftsverkehr**
KS-Test	**Kolmogoroff-Smirnov-Anpassungstest**
LAN	**Local Area Network**
MB	**Megabyte**
Mktg	**Marketing-Komitee der SNIA**
MW	**Mittelwest**
NAS	**Network Attached Storage**
NO	**Nordost**
OrgTeil	**organisatorischer Teil**
PACS	**Picture Archive and Communications Systems**
PM	**Projektmanagement**
Q-Q-Plot	**Quantil-Quantil-Plot**
RTO	**Recovery-Time-Objectives**
SAN	**Storage Area Networks**
S-ATA	**Serial-Advanced Technology Attachement**
SCSI	**Small Computer System Interface**
SdK	**Schutzgemeinschaft der Kapitalanleger**
SEC	**Securities and Exchange Commission**
SLA	**Service Level Agreement**
SLO	**Service Level Objectives**
SNIA	**Storage Networking Industry Association**
SOA	**Service-orientierte Architektur**
SOX	**Sarbanes-Oxley-Act**
SRM	**Storage Resource Management**

SS	**Süd-Südwest**
TC	**Technical Council**
TCO	**Total Cost of Ownership**
TechTeil	**technischer Teil**
TLG	**Technical Liaison Group**
TWG	**Technical Work Group**
URL	**Uniform Resource Locator**
WORM	**Write Once Read Many**

11 Verwendete Bezeichner

α	Signifikanzniveau
α, β	Verteilungsparameter der Gamma- und der Weibullverteilung
$\hat{\alpha}, \hat{\beta}$	geschätzte Verteilungsparameter
λ	Verteilungsparameter der Exponentialverteilung
$\hat{\lambda}$	geschätzter Verteilungsparameter
r	Korrelationskoeffizient
H_0, H_1	Hypothesen der Anpassungstests
O_i	Anzahl der Beobachtungswerte in der i-ten Klasse
E_i	Erwartete Beobachtungen in der i-ten Klasse
$F(x)$	Verteilungsfunktion
$F^*(x)$	Verteilungsfunktion der gestutzten Verteilung
$1_{[0,\infty)}(x)$	Sprungfunktion
Ω	Grundgesamtheit
ω	Element aus Ω
F	Sigma-Algebra
$X(\omega)$	Zufallsvariable
P	Wahrscheinlichkeitsmaß
V(I)	Informationswert
S	Speicherhierarchie
C	Informationsklasse
$\rho(t)$	Rückmigrationen pro Runde
$\theta_k(t)$	Jitter
$\overline{\theta(t)}$	durchschnittlicher Jitter
$a_i(t)$	absoluter Kapazitätsbedarf der Hierarchie i
$n(t)$	Anzahl der Dateien auf Speicherebene i
s_{ij}	Größe der Datei j auf Speicherebene i
$g_i(t)$	Wachstum des absoluten Kapazitätsbedarfs
T	Typ des Szenarios
D	Simulationsdauer
E	Anzahl der Speicherebenen

R	Art der Migrationsregel
I	Startanzahl der Dateien

Anhang

A Fallstudie 1

Im Rahmen der Fallstudie 1 wurde eine Analyse über ein Dokumentenverwaltungssystem eines DAX-30-Unternehmens vorgenommen, um zu evaluieren, ob genügend Potenzial für die Umsetzung von ILM vorhanden ist.

Das Dokumentenverwaltungssystem wird von einem deutschlandweit tätigen Consulting-Bereich benutzt. Dieser Bereich unterteilt sich in neun Regionen, die jeweils ein eigenes Consulting Team haben, welches in der Region Rhein-Main etwa 90 Mitarbeiter zählt; deutschlandweit sind es circa 700 Personen.

Das Unternehmen unterstützt die überregionale Zusammenarbeit der einzelnen Consulting-Teams. Hierfür wurde ein einheitliches Dokumentenverwaltungssystem, nachfolgend „Datenbank“ genannt, eingeführt, in der sämtliche Berater ihre Projekte ablegen müssen. Damit wird ein überregionaler Zugang zu allen Dokumenten gewährleistet. Die gesamte Datenbank umfasst über 150.000 Dokumente. Die Dokumente können in der Datenbank editiert, angesehen, versioniert und gelöscht werden.

Beim Versionieren werden mehrere physikalische Dateien für ein Dokument im System vorgehalten. Ein Löschvorgang macht diese nicht mehr über die üblichen Clients zugänglich, obwohl sie drei Tage in der Datenbank vorgehalten werden und nur mit einem Administratorzugang wieder hergestellt werden können.

Eine Untergruppe, das Projektmanagement (PM), hat klare Richtlinien für die Dokumentation und Ablage, welche für eine einheitliche Archivierung sorgen sollen. Aus diesem Grund wurden Stichproben vom PM-Bereich entnommen, da definierte Vorgaben und Prozesse bestehen und abgeglichen werden können.

Ziel der Fallstudie 1 ist es zu identifizieren, wie häufig Dateien aufgerufen werden. Abgeleitet daraus sollen nicht nachgefragte Dateien als Potential für ILM identifiziert werden.

A.1 Datenbasis

Am 1.11.2004 wurden in der Datenbank 134 Projekte gezählt. Die Datenbank selbst unterteilt sich in drei Bereiche, benannt Mittelwest (MW), Nordost (NO) sowie Süd-Südwest (SS), in denen die neun Regionen ihre Projektablage vornehmen. Um keinen Bereich zu bevorteilen, steuerte jeder Bereich den jeweils gleichen Anteil an der Stichprobe bei, da keine Mitarbeiterverteilung bekannt war.

Die alphabetische Liste der Projekte wurde durchnummeriert. Die vergebenen Ordnungsnummern identifizierten die Projekte. Mittels der Random-Funktion von Java wurde eine Pseudozufallszahl im Bereich [0, n-1] (n = Anzahl Projekte) bestimmt. Nach der Auswahl wurde dieses aus der Liste gestrichen und selbige neu durchnummeriert. Dieser Schritt wurde für jeden Bereich dreimal ausgeführt mit dem Ergebnis einer neunelementigen Stichprobe. Die Auswahlwahrscheinlichkeit für jedes Element der Grundgesamtheit war somit gleich, worauf es hier zunächst ankommt.

Auf diese Weise wurden 1762 Einzeldokumente, die insgesamt 942 MB Speicher benötigen, protokolliert. Aus Gründen der Anonymisierung wurden die untersuchten Projekte mit Großbuchstaben beschriftet, mit denen sie nachfolgend referenziert werden (siehe Tabelle 36).

Projekt	A	B	C	D	E	F	G	H	I
Region	MW	MW	MW	NO	NO	NO	SS	SS	SS
Dateien	196	170	121	430	592	6	116	66	65
Menge (kB)	54687	267840	26725	338295	191848	640	36483	12952	12969

Tabelle 36: Untersuchte Projekte

Der protokollierte Zeitraum eines Dokumentes erstreckt sich von der Erstellung bis zum 31. Januar 2005 und wurde über die Notification sowie zur Konsistenzprüfung über die History der Einzeldokumente bestimmt. Aus diesen Rohdaten wurde eine relative Zugriffshäufigkeit für jedes Intervall berechnet. Die Einteilung der zu betrachtenden Zeitintervalle erfolgt in Anlehnung an Moores Studie [62] und lautet:

(0, 1], (1, 3), [3, 7), [7, 15), [15, 30), [30, 60), [60, 90), [90, ∞) Tage.

In Kapitel A.4 wird der Zeitraum jenseits der 90 Tage weiter unterteilt und untersucht, weil sich zeigen wird, dass eine Aussage jenseits von 90 Tagen einer weiteren Differenzierung bedarf.

A.2 Auswertung der Zugriffsdaten auf 90-Tage-Basis

Um die relativen Zugriffe auf die vorgegebenen Intervalle zu erhalten, wird wie folgt vorgegangen: Es seien D1, D2, D3 drei Dokumente, die folgende Zugriffe aufweisen (Tabelle 37, links). Findet für ein Intervall und ein Dokument ein Zugriff statt, so wird eine Eins notiert, ansonsten eine Null (Tabelle 37, rechts). In jedem Intervall werden die Zugriffe summiert und durch die Anzahl der Dokumente dividiert.

Zugriffe	Intervall 1	Intervall 2	Intervall 3	Zugriffe	Intervall 1	Intervall 2	Intervall 3
D1	12	4	2	D1	1	1	1
D2	10	3	0	D2	1	1	0
D3	9	0	0	D3	1	0	0
				Relative Zugriffe	1	0,66	0,33

Tabelle 37: Erhebung Nutzdaten

Diese Schritte werden für alle untersuchten Dokumente und Intervalle vorgenommen. Die aggregierte Auswertung führt zu folgender relativer Zugriffshäufigkeit (siehe Tabelle 38 und Abbildung 32).

Intervall	(0,1]	(1,3)	[3,7)	[7,15)	[15,30)	[30,60)	[60,90)	[90,∞)
Zugriff %	100	0,82	6,59	2,35	8,74	4,1	15,37	9,61
Zugriff abs.	3909	52	151	134	408	142	685	397

Tabelle 38: Relative Zugriffshäufigkeiten der Dokumente

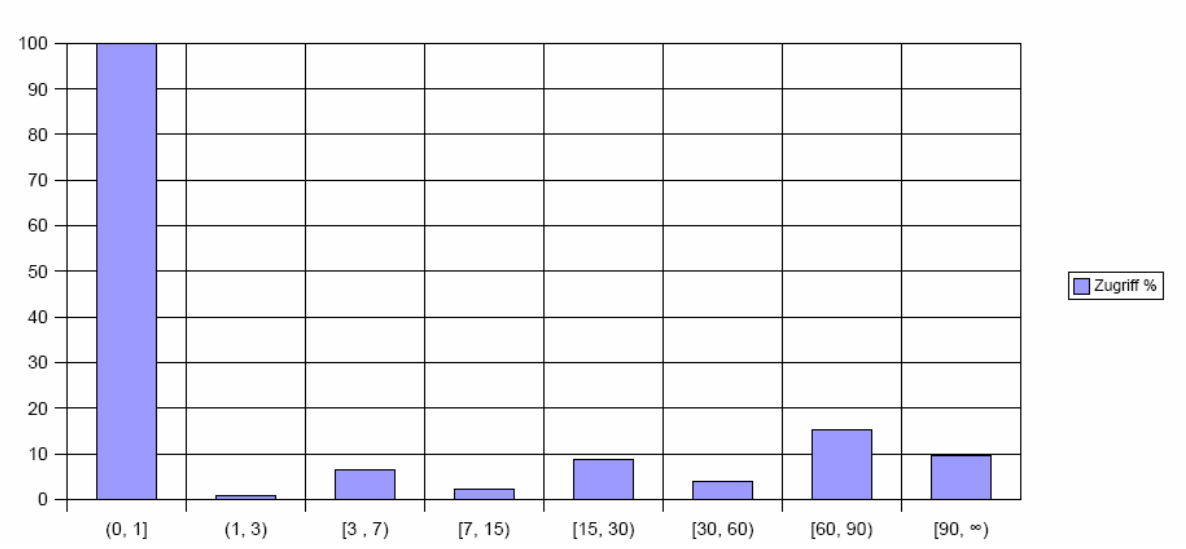


Abbildung 32: Relative Zugriffshäufigkeiten der Dokumente

Die Entwicklung der Zugriffe weist ein deutliches Gefälle auf. Die Zugriffshäufigkeiten im Intervall drei bis sieben Tage liegt bei 6,59 Prozent. Dies entspricht in etwa Moores Hauptaussage [62].

Die Zugriffshäufigkeit im Intervall [30, 60) liegt unter 5 Prozent, nämlich bei 4,61 Prozent. Werden die beiden nächsten Intervalle untersucht, ist ein Zuwachs ersichtlich, nämlich 15,37 Prozent im Intervall [60, 90) und 9,61 Prozent im Intervall [90, ∞).

Somit kann die zweite Moore'sche Hauptaussage, dass nach 90 Tagen die Zugriffshäufigkeiten nahe Null liegen, hier nicht bestätigt werden. Dies bedeutet, dass Moore allein nicht als Grundlage für ILM bei der Datenbank herangezogen werden sollte. Um exaktere Aussagen über die Zugriffe machen zu können, werden alle neun Projekte der Stichprobe einzeln untersucht.

A.3 Ergebnis der Auswertungen auf 90-Tage-Basis

Die Einzelbetrachtung der Projekte zeigt, dass entgegen den Aussagen Moores 30 Tage nach Erstellung eines Dokumentes und später sehr wohl noch Zugriffe in signifikanter Häufigkeit stattfinden.

Auch wenn man über alle Projekte aggregiert, sinkt die Zugriffshäufigkeit erst 90 Tage nach Erstellung unter die 10-Prozent-Schwelle.

Wie man an der großen Streuung später Zugriffshäufigkeiten zwischen den einzelnen Projekten festgestellt hat, stellt sich grundsätzlich die Frage, ob eine solche Aggregierung sinnvoll ist.

Diese Frage ließe sich mit statistischen Mitteln beantworten, indem man z.B. auf Anpassung an die „Moore-Kurve" (siehe Abschnitt 2.2) testet. Da die empirische Verteilung der Moore-Kurve nicht vorliegt, ist eine Anpassung ausgeschlossen. Angesichts dessen, dass die Moore-

Studie schon mittels deskriptiver Statistik als nicht anwendbar erkannt wurde, empfiehlt sich eine Anpassung an andere bekannte Verteilungen.

Für jedes Projekt werden die Aufteilung des Speicherbedarfs und die Anzahl der Dokumente jeweils in Abhängigkeit vom Dokumententyp näher betrachtet. A.5 enthält eine kurze Erklärung, von welcher Applikation dieser stammt.

Die diversen Dokumentkategorien (wie Datenhaltung [MDB, LDB, XLS, MPP], Datenaufbereitung [DOC, PPT, VSS, VSD], Archivierung [PDF, GZ, ZIP], Illustration [TIF, JPG, GIF] sowie Sonstige [XLA, DOT, MSG, TXT, HTM, RTF, TRC]) unterscheiden sich hinsichtlich Anzahl zugeordneter Dokumente und deren Größe teilweise beträchtlich.

Nach Betrachtung der einzelnen Projekte wird nun ein Vergleich mit den aggregierten Daten vorgenommen. Stellt sich heraus, dass die Ergebnisse im Wesentlichen projektunabhängig ausfallen, würde dies ein einfacheres Regelkonzept für ILM erlauben, ohne dass die Effektivität des Konzepts merklich leiden würde.

Die nachfolgenden Statistiken weisen die Abhängigkeit von Dokumentenzahl beziehungsweise -größe nach Dokumententyp aus, jedoch ohne Berücksichtigung der Projektzugehörigkeit. Um die Repräsentativität dieser Ergebnisse einzuschätzen, wird zusätzlich die empirische Streuung bezüglich Projektzugehörigkeit und Dokumentenkategorien ermittelt.

Die Kategorien Illustration und Sonstige tragen zur Gesamtzahl der Dokumente mit 2,64 Prozent einen verschwindend geringen Teil bei, weswegen sie nicht weiter diskutiert werden.

Anhand von Tabelle 39 und Tabelle 40 sowie Abbildung 33 und Abbildung 34 ergeben sich für die Dokumentenkategorien folgende Anteile hinsichtlich Gesamtdokumentzahl und Gesamtspeicherbedarf (siehe Tabelle 41):

Projekte	Aufbereitung	Haltung	Archivierung
Anzahl Dokumente	850	447	401
Anteil Dokumentenzahl	48,24	25,37	22,76
Speicherbedarf (kB)	556864	106604	256801
Anteil Speicherbedarf	59,12	11,32	27,27

Tabelle 39: Dokumentenanzahl und Speicherbedarf nach Kategorien

Die Aufschlüsselung nach Dokumententypen ähnelt bezüglich der Dokumentenzahl stark dem Projekt E. Hinsichtlich der Dokumentengröße gibt es kein Einzelprojekt mit hoher Ähnlichkeit. Damit stellt sich für ein ILM-Regelwerk die Frage, ob die Projektzugehörigkeit berücksichtigt werden soll.

Typ/Projekt	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
Anzahl	748	7	404	68	1	25	9	15	1	27	346
Anzahl %	42,45	0,4	22,93	3,86	0,06	1,42	0,51	0,85	0,06	1,53	19,64
Typ/Projekt	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc	
Anzahl	23	3	19	53	2	2	5	1	1	2	
Anzahl %	1,31	0,17	1,08	3,01	0,11	0,11	0,28	0,06	0,06	0,11	

Tabelle 40: Alle Projekte – Dokumentenanzahl nach Dateityp

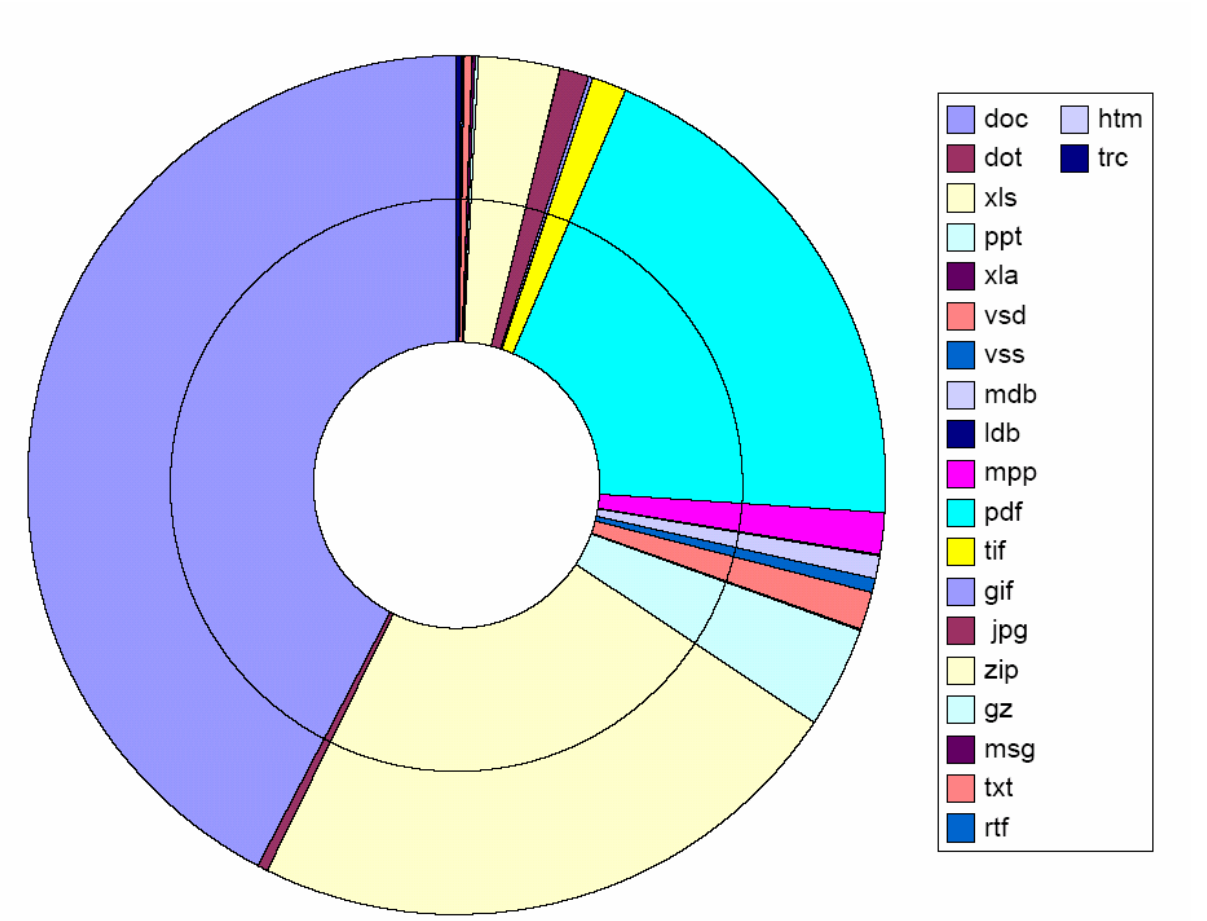


Abbildung 33: Alle Projekte - Dokumentenanzahl nach Dateityp

Typ/Projekt	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
Anzahl	456956	350	95983	91353	35	8294	261	4032	2	6587	82857
Anzahl %	48,51	0,04	10,19	9,7	0	0,88	0,03	0,43	0	0,7	8,8
Typ/Projekt	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc	
Anzahl	5293	11	13506	173026	918	71	15	34	8	2480	
Anzahl %	0,56	0	1,43	18,37	0,1	0,01	0	0	0	0,26	

Tabelle 41: Alle Projekte – Dokumentengröße nach Dateityp

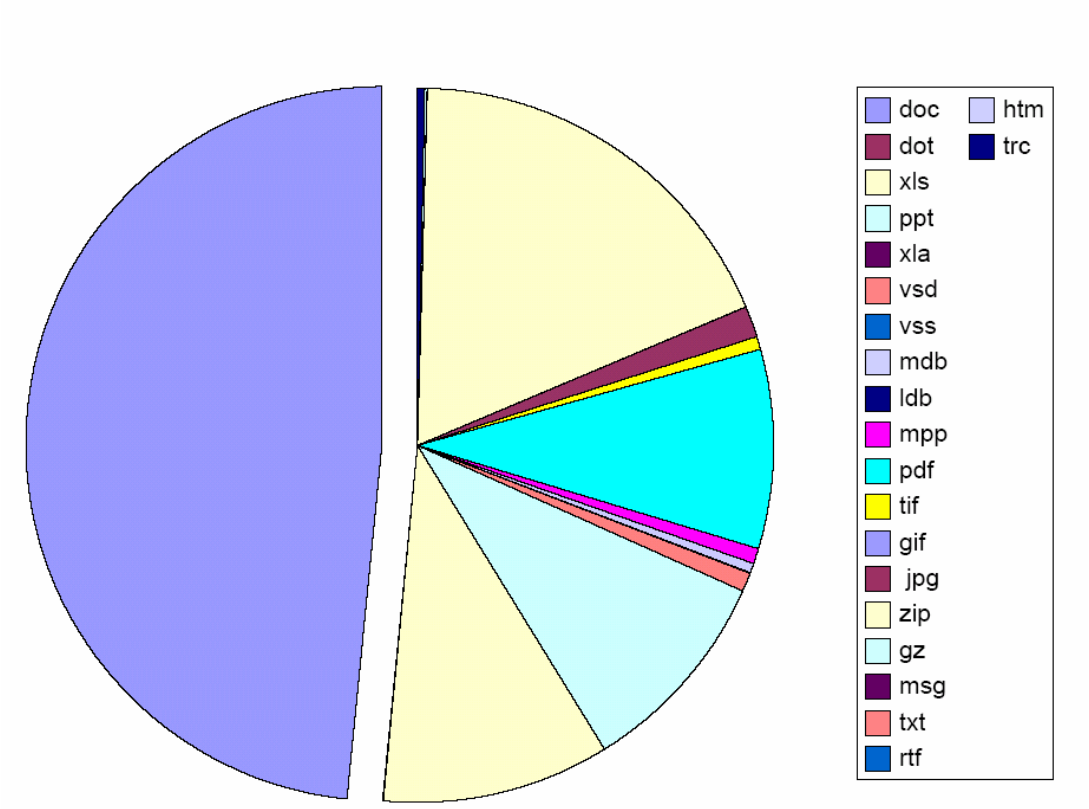


Abbildung 34: Alle Projekte – Dokumentengröße nach Dateityp

Die Inspektion einfacher empirischer Maßzahlen der Daten (Tabelle 42 bis Tabelle 45)[23] deutet darauf hin, dass die Aufteilung nach Dokumentkategorien stark vom Projekt abhängt. Dokumentenzahl und Speicherbedarf streuen ebenfalls stark zwischen den Projekten. Das könnte darauf hinweisen, dass ein einheitliches Regelwerk ein deutlich geringeres Potenzial hebt[24].

Kategorie	Mittelwert (MW)	Standardabweichg.	Standardabweichg./MW
Dokumentenaufbereitung	94,44	83,93	88,87
Dokumentenhaltung	49,67	56,25	113,25
Archivierung	44,56	52,78	118,45

Tabelle 42: Streuung der Dokumentenanzahl

Kategorie	Mittelwert (MW)	Standardabweichg.	Standardabweichg./MW
Dokumentenaufbereitung	61873,78	96632,72	156,18
Dokumentenhaltung	11844,89	13346,55	112,68
Archivierung	28533,44	37851,5	132,66

Tabelle 43: Streuung der Dokumentengröße

[23] Für die auf Dokumentkategorien entfallenden Anteile wurden gerundete Werte verwendet, was bei verschiedenen Einträgen zu weiteren Rundungsfehlern führte.

[24] Valide Aussagen lassen sich nur mit Verteilungstests ermitteln, was den Rahmen dieser Fallstudie übersteigt.

Kategorie	Mittelwert (MW)	Standardabweichg.	Standardabweichg./MW
Dokumentenaufbereitung	0,59	0,2	33,48
Dokumentenhaltung	0,22	0,11	49,01
Archivierung	0,19	0,12	66,29

Tabelle 44: Streuung der Kategorienanteile bzgl. Dokumentenanzahl

Kategorie	Mittelwert (MW)	Standardabweichg.	Standardabweichg./MW
Dokumentenaufbereitung	0,55	0,27	50,22
Dokumentenhaltung	0,16	0,12	72,89
Archivierung	0,29	0,23	78,29

Tabelle 45: Streuung der Kategorienanteile bzgl. Dokumentengröße

Die aggregierte Zugriffsstatistik (Tabelle 46) zeigt, dass mit einem einfachen Regelwerk praktisch kein Potenzial gehoben werden kann. Eine Ausnahme bilden ausgewählte Dateitypen wie LDB, HTM, GIF, TIF und in Grenzen auch die restlichen Dokumentenarten der Kategorie Sonstiges. Zwar verzeichnen diese teilweise Zugriffe in vielen Intervallen, aber die absolute Zahl der Dokumente ist gering. Außerdem handelt es sich bei den Typen häufig um Hilfsdateien mit vornehmlich technischer Bedeutung für die assoziierten Applikationen (zum Beispiel TRC, LDB).

Typ/Projekt	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
(1,3)	100	100	100	100	100	100	100	100	100	100	100
[3,7)	0,88	0	0,84	0	0	2,5	0	0	0	3,33	1,11
[7,15)	7,86	14,29	9,46	8,33	0	4,17	7,35	1,19	0	28,33	11,68
[15,30)	8,31	0	9,8	7,09	50	18,75	12,5	0	0	43,33	7,35
[30,60)	2,03	0	10,98	12,09	0	1,04	0	0,13	0	13,33	2,57
[60,90)	15,71	28,57	24,6	25,12	50	19,79	25	25,25	0	37,08	15,87
[90,∞)	10,58	14,29	10,34	16,64	0	15,42	18,75	12,5	0	25,83	7,39
Typ/Projekt	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc	
(1,3)	100	100	100	100	100	100	100	100	100	100	
[3,7)	0	0	0	0,55	0	0	0	0	0	0	
[7,15)	0	0	0	0,55	0	0	0	0	0	0	
[15,30)	0	0	40	17,03	100	100	0	100	0	100	
[30,60)	50	0	1,54	2,89	0	0	0	0	0	100	
[60,90)	0	0	40	21,43	0	100	37,5	100	0	100	
[90,∞)	0	0	20	13,19	100	100	37,5	0	0	100	

Tabelle 46: Alle Projekte – Relative Zugriffshäufigkeiten nach Dokumenttyp und Zeitintervall

Typ	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
kB %	48,51	0,04	10,19	9,7	0	0,88	0,03	0,43	0	0,7	8,8
[90,∞)	10,58	14,29	10,34	16,64	0	15,42	18,75	12,5	0	25,83	7,39
¬[90,∞)	89,42	85,71	89,66	83,36	100	84,58	81,25	87,5	100	74,17	92,61
Potenzial	43,38	0,03	9,13	8,08	0	0,74	0,02	0,37	0	0,52	8,14
Typ	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	tif	Summe
kB %	0,56	0	1,43	18,37	0,1	0,01	0	0	0	0,26	100
[90,∞)	0	0	20	13,19	100	100	37,5	0	0	100	-
¬[90,∞)	100	100	80	86,81	0	0	62,5	100	100	0	-
Potenzial	0,56	0	1,15	15,94	0	0	0	0	0	0	88,09

Tabelle 47: Potenzial nach Dokumenttyp und gesamt

Ausgehend von den Dokumenten, die nach 90 Tagen keinen Zugriff zu verzeichnen haben, wird nun bestimmt, wie hoch der von ihnen belegte Anteil am Gesamtspeicherbedarf der Stichprobe ist. Tabelle 47 zeigt, dass es ein Potenzial von 88 Prozent des belegten Speichers zu heben gibt. Dieses Ergebnis bestätigt die Untersuchungen von Gibson et al. aus dem Jahre 1998.

Um die Charakteristika der Zugriffe noch genauer zu ermitteln, wird nachfolgend das Intervall [90,∞) weiter unterteilt.

A.4 Auswertung aller Projekte auf 400-Tage-Basis

In diesem Kapitel werden zuerst die Zugriffe bis zu 400 Tage nach Erstellung untersucht. Tabelle 48 zeigt, dass ca. 9 Prozent aller Zugriffe zwischen dem 250. und 300. Tag nach Erstellung erfolgt. Eine Erklärung dafür ist nicht eindeutig zu geben. Es kann an unternehmenseigenen Prozeduren der Revision oder des Controllings liegen.

Intervall	(1,3)	[3,7)	[7,15)	[15,30)	[30,60)	[60,90)	[90,120)
Häufigkeit	3574	89	229	255	296	677	63
Intervall	[120,150)	[150,200)	[200,250)	[250,300)	[300,350)	[350,400)	[400,∞)
Häufigkeit	162	46	146	552	3	11	14

Tabelle 48: **Absolute Zugriffshäufigkeiten auf 400-Tage-Basis**

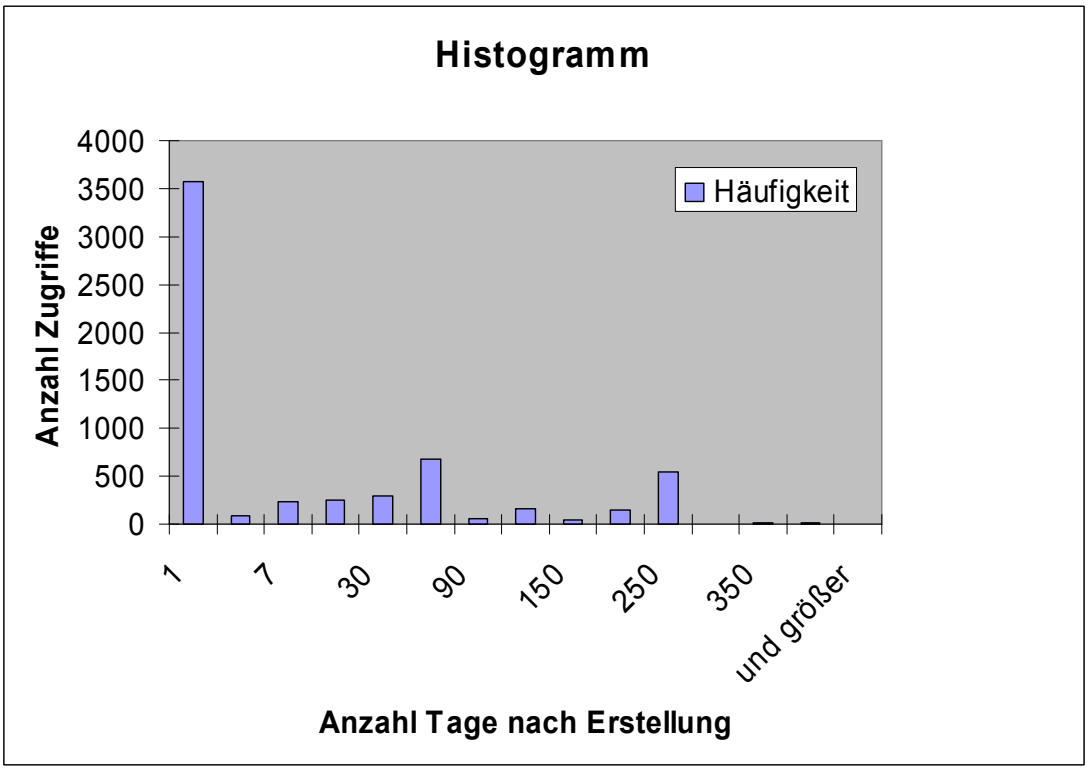


Abbildung 35: Histogramm der absoluten Zugriffshäufigkeiten

Betrachtet man die vergangene Zeit zwischen den Zugriffen, so erkennt man hier eine Häufung im Bereich 200. bis 250. Tag.

Tage s.l.Z.	(1,3)	[3,7)	[7,15)	[15,30)	[30,60)	[60,90)	[90,120)
Häufigkeit	4216	109	179	396	94	613	31
Tage s.l.Z.	[120,150)	[150,200)	[200,250)	[250,300)	[300,350)	[350,400)	[400,∞)
Häufigkeit	118	26	328	1	6	0	0

Tabelle 49: Absolute Zugriffshäufigkeiten seit letztem Zugriff

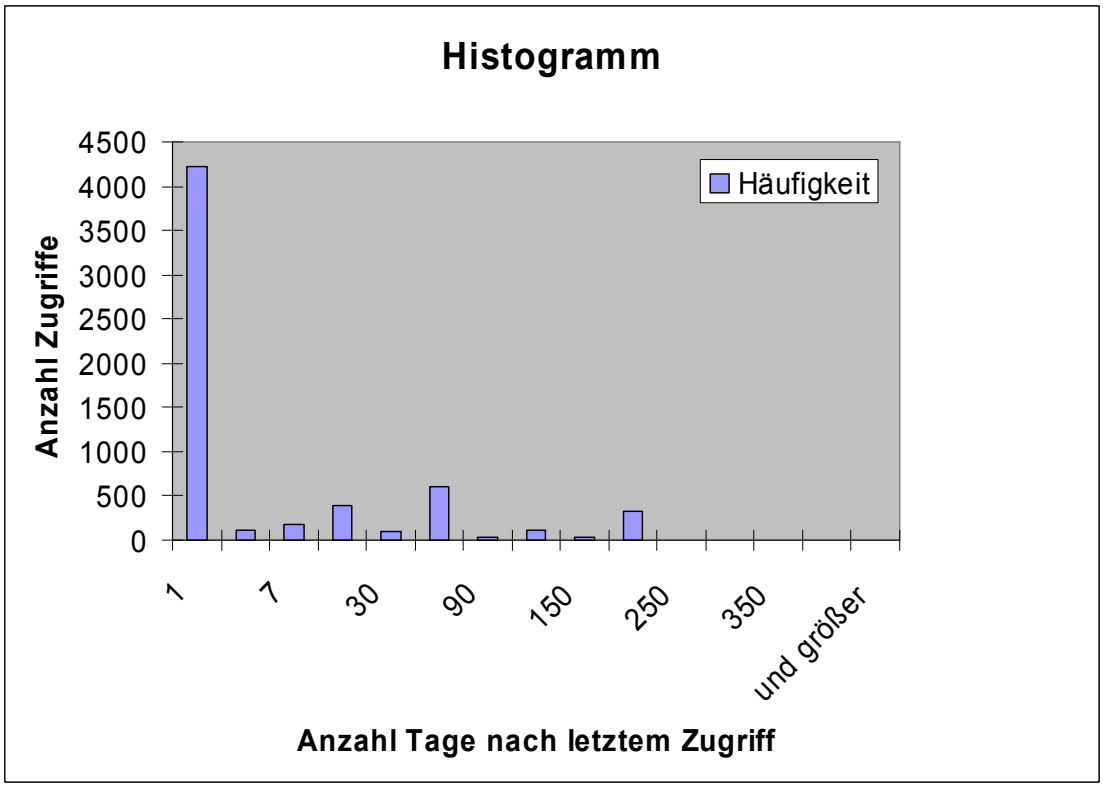


Abbildung 36: Histogramm der absoluten Zugriffe seit letztem Zugriff

Die Betrachtung auf 400-Tage-Basis zeigt, dass, obwohl 89% der Dateien keine Zugriffe 90 Tage nach Erstellung erfahren, man dennoch die Zugriffe über einen längeren Zeitraum als 90 Tage betrachten muss.

A.5 Zusammenfassung der Fallstudie 1

Im Rahmen des Technical Reports wurde ein Datenbestand eines Unternehmens analysiert, um zu prüfen, ob genügend Potenzial vorhanden ist, so dass ILM nutzbringend angewendet werden kann.

Es wurde ein Potenzial von 90 Prozent identifiziert, welches 90 Tage nach Erzeugung brach liegt. Eine kanonische Regel für die Datenbank lautet wie folgt:

Lösche oder verdränge alle Dokumente 90 Tage nach Erstellung.

Eine derartige Regel würde allerdings eine Fehlerrate von circa 10 Prozent verursachen, was unvertretbar ist. Die Fehler ziehen sich durch fast alle Applikationen hindurch.

Es zeigt sich, dass die Moore-Studie [62] im Spezialfall der untersuchten Datenbank des DAX-30-Unternehmens nicht aussagekräftig ist.

Insbesondere die betrachteten Intervalle bei Horison sind zu grob. Man muss das Intervall [90, ∞) in weitere Teilintervalle unterteilen und untersuchen, wie die Betrachtung auf 400-Tage-Basis zeigt.

Obwohl 89% der Dateien keine Zugriffe 90 Tage nach Erstellung erfahren, finden über 16% der Zugriffe (997 von 6117) nach 90 Tagen nach Erstellung statt. Diese Erkenntnis sollte bei einer ILM-Regelerstellung berücksichtigt werden.

Nur mittels der Unterteilung auf 400-Tage-Basis lässt sich das identifizierte Potential zumindest teilweise erschließen.

B Dateienpool

Es wurde ein Dateienpool mit Zugriffsinformationen und Metadaten von über 75.000 Dateien der Datenbank EDB (siehe Fallstudie 1) angelegt. Dieser steht zur Erzeugung von Simulationsstichproben zur Verfügung.

B.1 Allgemeines über den Dateienpool

Der Dateienpool enthält Zugriffsinformation von 75.601 Dateien. Diese Informationen sind in einer Microsoft Access-Datenbank konsolidiert. Zu jeder Datei stehen folgende Informationen zur Verfügung:

Dateiname
Dateityp
Dateigröße
Datei-ID
Besitzer-ID
Anzahl Zugriffe
Zugriff 1 Nutzer-ID, Zugriffstyp
…
Zugriff n Nutzer-ID, Zugriffstyp

Beispiel:

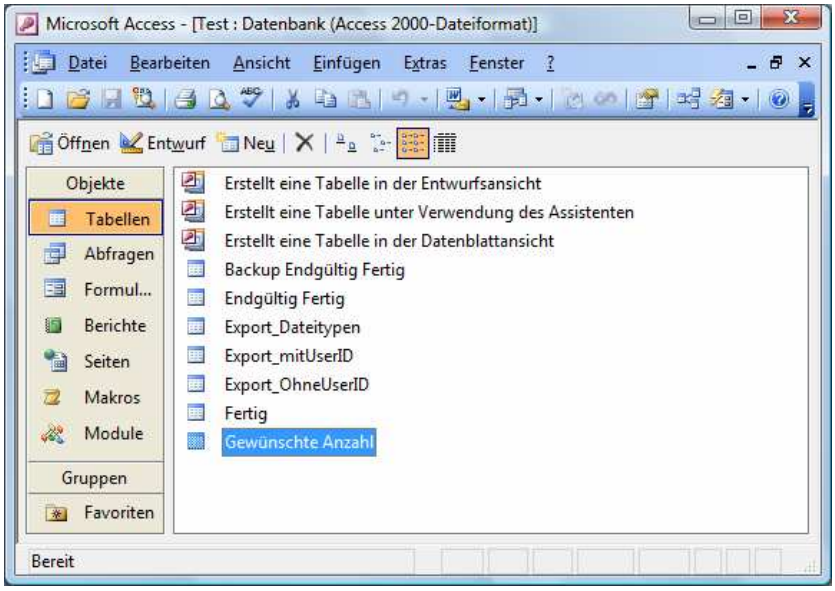
```
Angebot Schulung.doc
DOC
380928
272188
21617
6
21617     05.12.2001 15:27    ADDVERSION
21617     05.12.2001 15:27    CREATE
14235     09.01.2003 12:01    FETCH
14235     19.08.2003 16:00    FETCH
6188      09.10.2003 12:05    FETCH
14235     08.01.2004 15:59    VIEW
```

Von den unterschiedlichen Dateitypen stehen insbesondere folgende zur Verfügung:

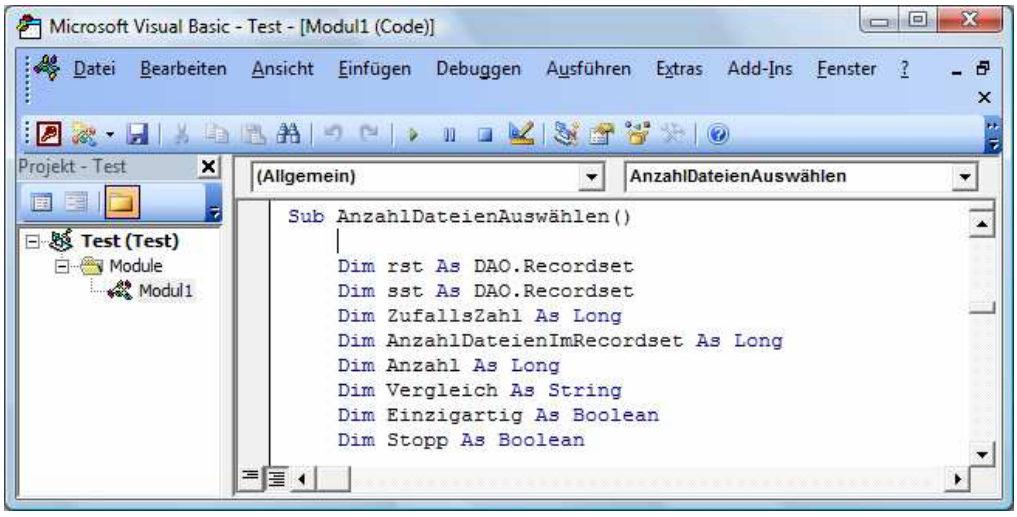
18.217 Microsoft Word Dokumente (.doc)
10.483 Microsoft Excel Tabellen (.xls)
6.898 Microsoft Powerpoint (.ppt)
6.385 Adobe Acrobat Dokumente(.pdf)
2576 zip-Archive (.zip)
2519 e-mail-Nachrichten (.msg)

B.2 Erzeugung einer Stichprobe aus dem Dateienpool

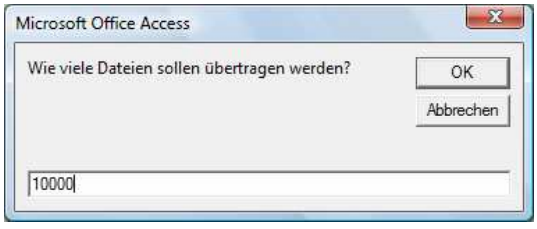
Zur Erzeugung einer Stichprobe wird die Access-Datenbank geöffnet. Es präsentiert sich folgendes Fenster:



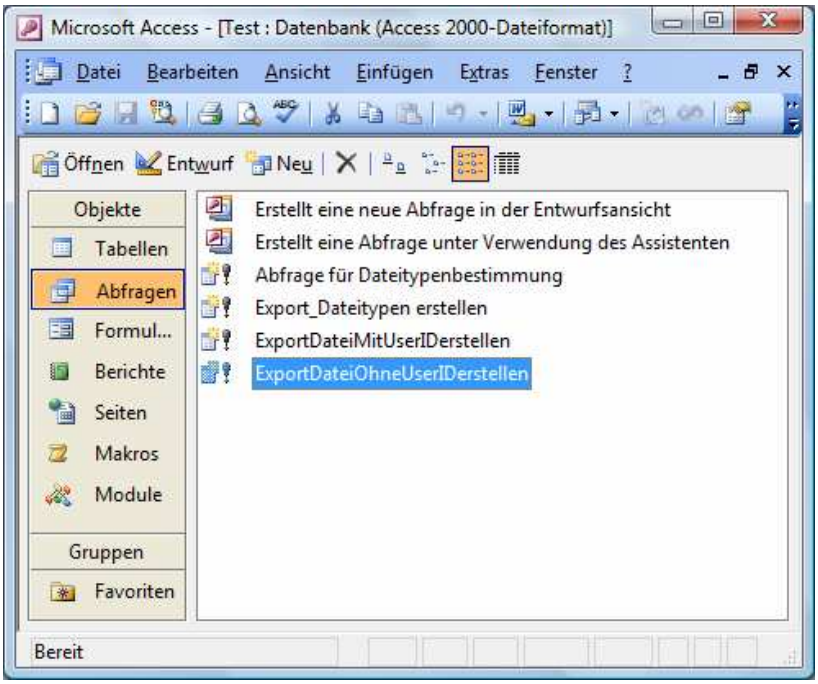
Man wechselt zu „Module“. Hier wird mittels Visual Basic die Erzeugung durchgeführt:



Anschließend gibt man die gewünschte Anzahl von Dateien an.



Nach Erzeugung gelangt man zu „Abfragen“.

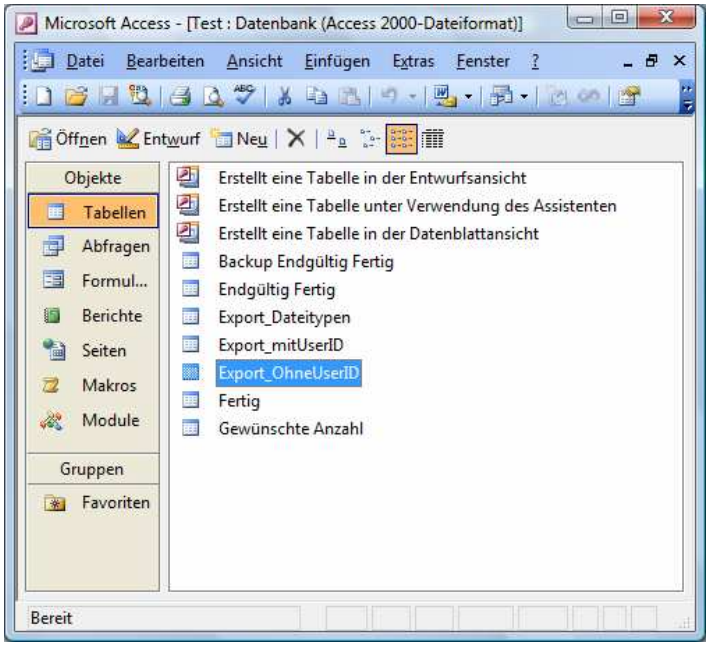


Hier besteht die Möglichkeit, eine Stichprobe mit oder ohne „User ID“ zu erzeugen. Die „User ID“ ist ein eindeutiger Bezeichner, der den Zugreifenden auf eine Datei identifiziert und jedem Zugriff zugewiesen werden kann.

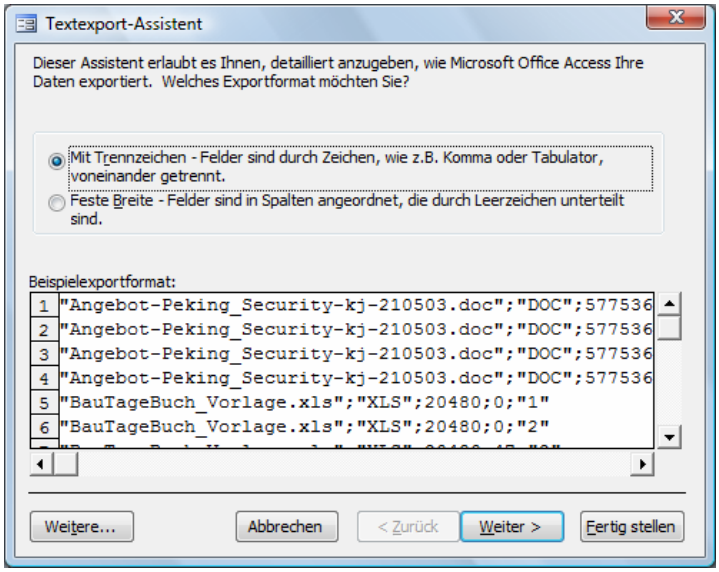
Um eine Stichprobe mit „User ID“ zu erstellen, ist die die Abfrage „ExportDateiMitUserI-Derstellen“ auszuführen.

Um eine Stichprobe ohne User ID zu erstellen, ist die die Abfrage „ExportDateiOhneUserI-Derstellen“ auszuführen.

Nach der Dateierstellung kann man nun die Datei exportieren. Dazu wechselt man in der linken Spalte auf „Tabellen“.

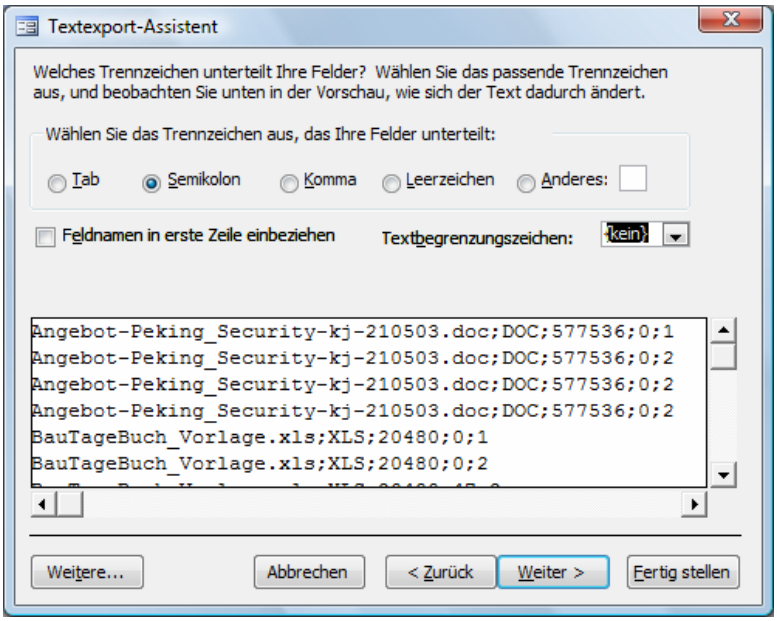


Für eine Stichprobe ohne User ID die Tabelle „Export_OhneUserID“ öffnen



Mit „Weiter“ bestätigen

Als Trennzeichen Semikolon auswählen, als Textbegrenzungszeichen „kein“ auswählen

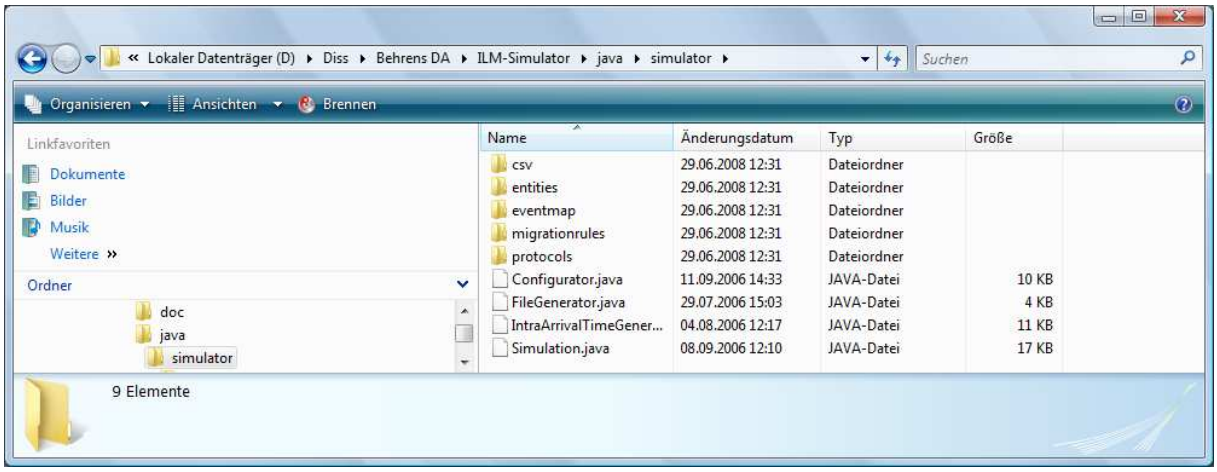


Mit „Weiter“ und „Fertig stellen“ bestätigen.

Damit ist die Erzeugung einer Stichprobe abgeschlossen.

Diese kann nun in den Simulator eingelesen werden.

C Nutzung des Simulators



1. Die erstellte Exportdatei nach G:\ILM\Simulator kopieren. (ggf. anstatt „G“ den Laufwerksbuchstaben eines anderen Datenträgers verwenden)
2. Im Verzeichnis „G:\ILM\Simulator\java\simulator“ die Datei „Configurator.java“ bearbeiten
3. Die Zeile „ this.sourcedatafile = …. “ abändern in „this.sourcedatafile = "x"; “ (x= Kompletter Dateiname der erstellten Exportdatei)
4. In den Zeilen

```
layer1.add("layer1");

layer1.add(0.15);

layer2.add("layer2");

layer2.add(0.075);

layer3.add("layer3");

layer3.add(-0.05);
```

kann man die Prozentwerte der Migrationsregeln eingeben/abändern.

5. In der Zeile „ this.scenario = … " wird dem Programm gesagt, welche Parameter es verwenden soll:

Beispiel: this.scenario = "T2-D1000-E3-R5-I0-W0.2";

Erklärung hierfür findet man Kapitel 6.

Zahl hinter D = Dauer der Simulation

Zahl hinter E = Anzahl Ebenen

Mit dem Abschnitt R5 wird die Migrationsregel bestimmt.

6. Die Änderungen in der Datei speichern und die Datei schließen.

7. Unter „Windows“->„Start“->“Utilities“->“Accessories“ den „Command Prompt“

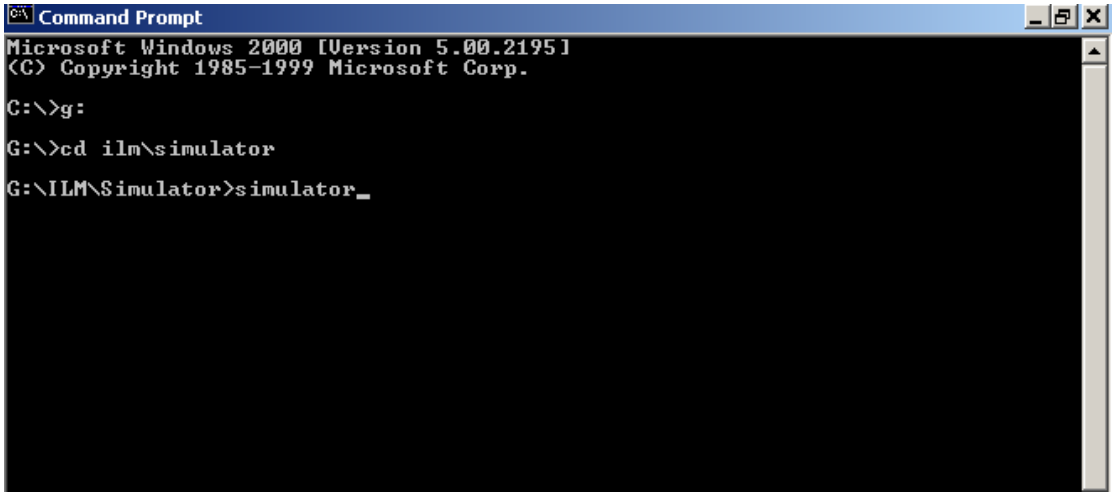
starten

8. Folgende Befehle eingeben:

„g:“ mit „Enter“ bestätigen

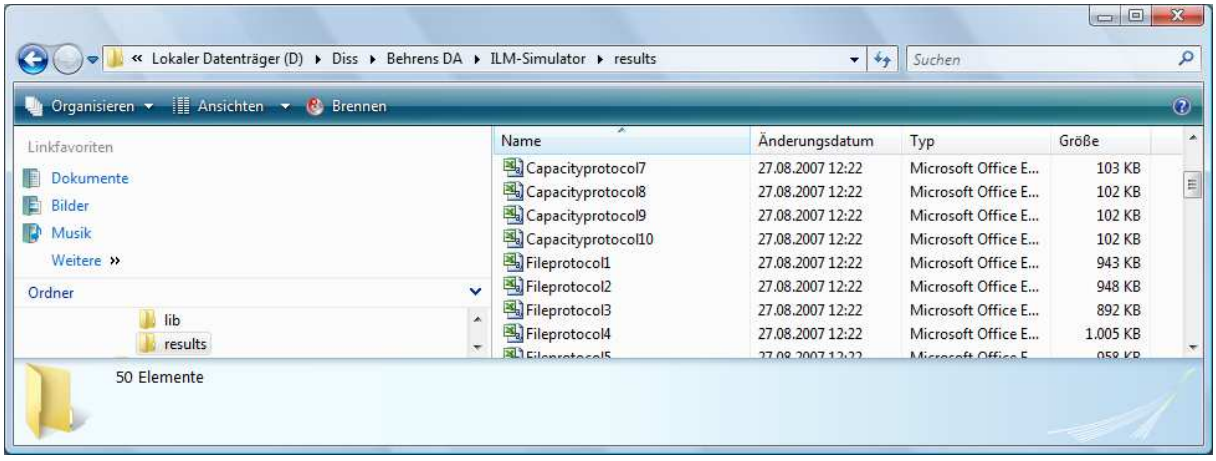
„ cd ilm\simulator“ mit „Enter“ bestätigen

„simulator“ mit „Enter“ bestätigen (Dieser Befehl startet den Simulator)



Die Resultate der Simulation finden sich unter „results“.

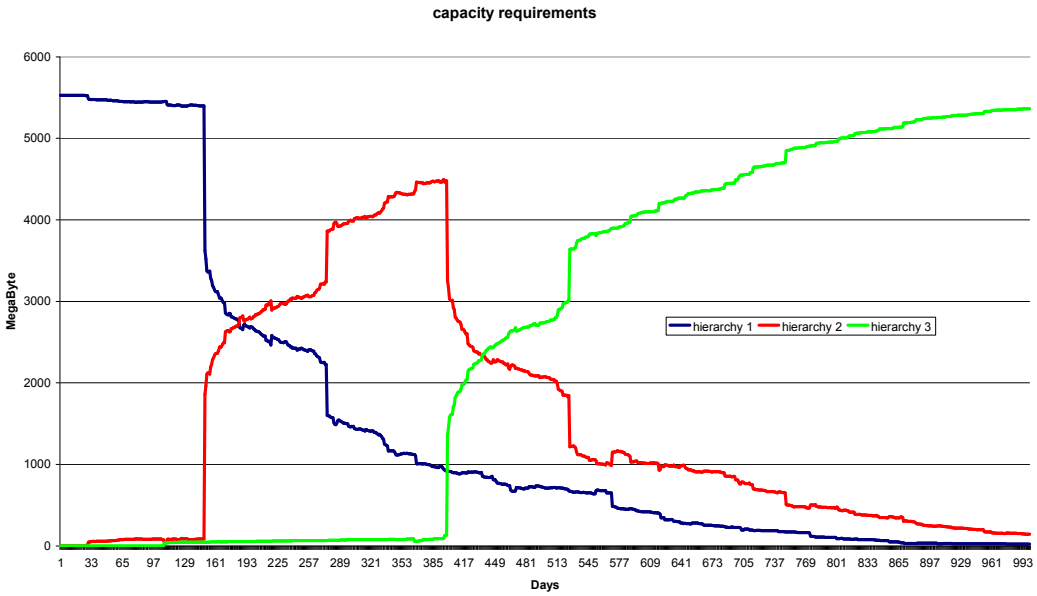
Hier stehen die erzeugten Protokolldateien im csv.Format:



Zur Visualisierung werden Grafiken aus den einzelnen Protokollen erzeugt:

1. Unter „G:\ILM\Simulator\results“ die Datei „Capacityprotocol1.csv“ öffnen
2. Das gesamte Tabellenblatt markieren und kopieren, anschließend die Datei schließen
3. Unter „G:\ILM\Simulator“ die Datei „Grafikmakro-Vorlage.xls“ öffnen
4. Die kopierten Werte in dem Tabellenblatt „Capacityprotocol1“ einfügen

5. „Alt+F8“ drücken und das Makro „GrafikErstellen“ starten
6. Mit „Datei“->“Speichern unter…“ die Datei unter gewünschtem Namen abspeichern
7. Auf dem Tabellenblatt „Diagramm1“ befindet sich nun die gewünschte Grafik:



D Publikationen des Verfassers

D.1 Wissenschaftliche Veröffentlichungen

Turczyk, L. A., Frei, C., Behrens, M., Liebau, N., Steinmetz, R.: *Simulation of Automated File Migration in Information Lifecycle Management.* In: 14th Americas Conference on Information Systems (AMCIS 2008), Toronto, 2008, S. 11-29.

Turczyk, L. A., Frei, C., Liebau, N., Steinmetz, R.: *Eine Methode zur Wertzuweisung von Dateien in ILM.* In: Multikonferenz Wirtschaftsinformatik 2008 (MKWI 2008), München, 2008, S. 459-470.

Turczyk, L. A., Liebau, N., Steinmetz, R.: *Modeling Information Lifecycle Management.* In: 13th Americas Conference on Information Systems (AMCIS 2007), Keystone, 2007, S. 1134-1146.

Turczyk, L. A., Gröpl, M., Liebau, N., Steinmetz, R.: *A Method for File Valuation in Information Lifecycle Management.* In: 13th Americas Conference on Information Systems (AMCIS 2007), Keystone, 2007, S. 1122-1133.

Turczyk, L. A., Behrens, M., Liebau, N., Steinmetz, R.: *Cost Impacts on Information Lifecycle Management Design*. In: 13th Americas Conference on Information Systems (AMCIS 2007), Keystone, 2007, S. 1110-1121.

Turczyk, L. A., Heckmann, O., Steinmetz, R.: *Simulation of Information Lifecycle Management*. In: 18th Annual Information Resources Management Association Conference (IRMA 2007), Vancouver, 2007, S. 1063-1066.

Turczyk, L. A., Heckmann, O., Steinmetz, R.: *File Valuation in Information Lifecycle Management*. In: 18th Annual Information Resources Management Association Conference (IRMA 2007), Vancouver, 2007, S. 347-350.

Turczyk, L. A., Heckmann, O., Berbner, R., Steinmetz, R.: *An Analytical Model of Information Lifecycle Management*. In: 17th Annual Information Resources Management Association Conference (IRMA 2006), Washington, 2006, S. 422-426.

Turczyk, L., Heckmann, O., Berbner, R., Steinmetz, R.: *A Formal Approach to Information Lifecycle Management*. In: 17th Annual Information Resources Management Association Conference (IRMA 2006), Washington, 2006, S. 531-533.

D.2 Weitere Veröffentlichungen

Turczyk, L. A.: *Information Lifecycle Management: Organisation ist wichtiger als Technologie.* In: Information - Wissenschaft und Praxis, Ausgabe 7-2005, S. 371-372.

Turczyk, L. A.: *Produkte sind zweitrangig – das Konzept macht's.* In: Speicherguide.de – Das Storage Magazin, September 2005, http://www.speicherguide.de/magazin/ bacground.asp?todo=de&theID= 1285 Abruf am 20.11.2008.

Turczyk, L. A.: *Information Lifecycle Management als Weg aus dem Speicherdilemma.* In: Information Wissenschaft und Praxis, Ausgabe 7-2004, S. 407-410.

Turczyk, L. A.: *Ordnungssystem – Information Lifecycle Management.* In: Content Management Magazin, Ausgabe 4/ 2004, S. 35-36.

Turczyk, L. A.: *Wenn der Sack voll ist – Information Lifecycle Management.* In: IT-Sicherheit – Management und Praxis, Ausgabe 4/ 2004, S. 57-58.

Turczyk, L. A.: *Wege aus dem zunehmenden Datendschungel.* In: ntz Nachrichtentechnische Zeitung, Ausgabe 11/ 2004, S. 42-43.

Turczyk, L. A.: *Rumble in the jungle.* In: DOQ Magazin, Ausgabe 5/2004, S. 90-93.

D.3 Technical Reports

Turczyk, L. A., Gröpl, M., Heckmann, O., Steinmetz, R.: *Statistische Datenanalyse von langfristigem Dateizugriffsverhalten.* Technische Universität Darmstadt, Fachgebiet KOM. Technical Report KOM 09-2006, 2006.

Turczyk, L. A., Gostner, R., Berbner, R., Heckmann, O., Steinmetz, R.: *Analyse von Datei-Zugriffen zur Potentialermittlung für Information Lifecycle Management.* Technische Universität Darmstadt, Fachgebiet KOM. Technical Report KOM 01-2005, 2005.

E Lebenslauf des Verfassers

Persönliche Daten

Name:	Lars Arne Turczyk
Geburtsdatum:	15. Dezember 1969
Geburtsort:	Frankenberg/Eder
Nationalität:	Deutsch

Schulausbildung

1976 – 1980	Wigand Gerstenberg Grundschule, Frankenberg/Eder
1980 – 1989	Gymnasium Edertalschule, Frankenberg/Eder, Abschluss: Abitur

Wehrdienst

1989 – 1990	Grundwehrdienst beim 3. Panzergrenadierbataillon 51 in Fritzlar

Studium

1990 – 1996	Philipps-Universität Marburg Studiengang: Mathematik, Abschluss: Dipl.-Math.
Seit Oktober 2003	Externer Doktorand am Fachgebiet Multimedia Kommunikation (KOM) an der Technischen Universität Darmstadt

Berufliche Tätigkeiten

1996 – 1999	Produkt Manager bei der Siemens AG in München
1999 – 2000	System Engineer bei der Siemens AG in Frankfurt
Seit 2000	Senior Consultant bei der Siemens Enterprise Communications GmbH & Co. KG in Frankfurt

F Eidesstattliche Erklärung laut §9 PromO

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation alleine und nur unter Verwendung der angegebenen Literatur und Hilfsmittel erstellt habe.

Die Arbeit hat noch nicht zu Prüfungszwecken gedient.

Darmstadt,