



eDiscovery with Hadoop: Is Open Source an option?

Diplomarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieur

eingereicht von

Marcel Miersebach, BSc
IS101507

im Rahmen des
Studiengangs Information Security an der Fachhochschule St. Pölten

Betreuer: Dr. Alexander Wöhrer

Wien, 24.10.12

(Unterschrift Autor)

(Unterschrift Betreuer)

Ehrenwörtliche Erklärung

Ich versichere, dass

- ich diese Diplomarbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich sonst keiner unerlaubten Hilfe bedient habe.
- ich dieses Diplomarbeitsthema bisher weder im Inland noch im Ausland einem Begutachter/einer Begutachterin zur Beurteilung oder in irgendeiner Form als Prüfungsarbeit vorgelegt habe.
- diese Arbeit mit der vom Begutachter/von der Begutachterin beurteilten Arbeit übereinstimmt.
- Der Studierende/Absolvent räumt der FH St. Pölten das Recht ein, die Diplomarbeit für Lehre- und Forschungstätigkeiten zu verwenden und damit zu werben (z.B. bei der Projektevernissage, in Publikationen, auf der Homepage), wobei der Absolvent als Urheber zu nennen ist. Jegliche kommerzielle Verwertung/Nutzung bedarf einer weiteren Vereinbarung zwischen dem Studierenden/Absolventen und der FH St. Pölten.

Wien, 24.10.12

(Unterschrift Autor)

Zusammenfassung

Diese Diplomarbeit beschäftigt sich mit dem Thema eDiscovery. Jedes Jahr sammeln sich in Unternehmen mehr und mehr elektronische Daten an, die bei einem eDiscovery Verfahren analysiert werden müssen. Für die ständig wachsenden Datenmengen werden früher oder später die traditionellen Verarbeitungsmechanismen nicht mehr ausreichen, daher benötigt es neue innovative Lösungen um diesen Datenmengen gerecht zu werden. Das Open Source Projekt FreeEed versucht dem ständig steigenden Datenvolumen mithilfe von Cloud Technologien entgegenzutreten. FreeEed kann mit den Bestandteilen MapReduce und dem Hadoop Distributed File System (HDFS) mehrere Clusterknoten für die Analyse der Daten verwenden und verspricht so einen enormen Zeitvorteil. Diese Arbeit beschäftigt sich zunächst allgemein mit der Thematik von eDiscovery und zeigt auf, inwiefern dieses Thema für uns in Europa relevant ist. Auch die Frage inwieweit eDiscovery bisher standardisiert wurde, wird geklärt. Nachdem Themenblock eDiscovery werden die verwendeten Technologien und Bestandteile, die bei FreeEed zum Einsatz kommen, erklärt. Der Hauptteil dieser Arbeit beschäftigt sich mit dem Vergleich von FreeEed mit einer traditionellen kommerziellen eDiscovery Software. Dabei wird ein großes Augenmerk auf die Verarbeitungsdauer, unterstützte Dateiformate und der Umgang mit Fremdsprachen gelegt. Zusätzlich werden die verschiedenen Standardfunktionen der beiden Rivalen verglichen und bewertet. Am Schluss dieser Arbeit werden nochmals die Stärken und Schwächen dieser Hadoop basierenden eDiscovery Software herausgearbeitet.

Abstract

This thesis discusses eDiscovery. Each year more and more electronic data is collected by enterprises that have to be analyzed for an eDiscovery purposes when there is a lawsuit. For the continual increasing amount of Data the traditional processing mechanism would not be sufficient. That's why we need new innovative solutions to accomplish the task. The Open Source project FreeEed is trying to solve this issue with cloud technologies. FreeEed uses MapReduce and the Hadoop Distributed File System (HDFS) to split the data on multiple cluster nodes to process and analyses the information. This technic promises an enormous time advantage by parallel processing. The first chapter of this thesis deals with the subject eDiscovery and shows how relevant this topic is for companies in Europe. Additionally the question of how the eDiscovery process can be standardized is also covered in this chapter. In the next section all components and technologies that FreeEed uses are explained. The main part of this paper compares the Open Source eDiscovery software FreeEed with traditional commercial eDiscovery software. The focus of this comparison is the processing time, supported file formats and the handling of foreign languages. Furthermore, the standard eDiscovery features of both competitors are tested and evaluated. In the last section the strengths and weaknesses of a Hadoop based eDiscover software is worked out.

Inhaltsverzeichnis

1	EINLEITUNG.....	8
2	ELECTRONIC DISCOVERY.....	10
2.1	WAS IST ELECTRONIC DISCOVERY?.....	10
2.2	DIE AMERIKANISCHE ZIVILPROZESSORDNUNG	12
2.3	EDISCOVERY IN DEUTSCHLAND UND ÖSTERREICH	14
2.4	PROBLEME UND GEFAHREN BEI EDISCOVERY	15
2.5	DAS ELECTRONIC DISCOVERY REFERENCE MODEL	19
2.6	ETABLIERTE EDISCOVERY SOFTWARE.....	20
3	CLOUD COMPUTING.....	27
3.1	EEDISCOVERY IN DER CLOUD.....	32
3.2	MAPREDUCE	35
3.3	HADOOP.....	38
3.4	CLOUDERA.....	44
3.5	FREEEED/SHMCLOUD	46
4	VERGLEICH	52
4.1	BEWERTUNG.....	60
5	AUSBLICK	63
6	FAZIT.....	65
	ANHANG 1 - INSTALLATION.....	67
	ANHANG 2 - DEPLOYMENT.....	70
	ANHANG 3 - FUNKTIONSTEST.....	71
	ANHANG 4 – FREEEED	77
	ANHANG 5 – SHMCLOUD	80
	ANHANG 6 - ÜBERSICHT FILEFORMATTEST.....	82
	ANHANG 7 - ÜBERSICHT MAPREDUCE JOBS	96
	GLOSSAR.....	98
	LITERATURVERZEICHNIS.....	104

Abbildungsverzeichnis

Abbildung 1: Einflüsse auf das Thema: ‘eDiscovery with Hadoop: Is Open Source an option?’	9
Abbildung 2: Anwendungsbeispiele für eDiscovery [21].....	12
Abbildung 3: Ablauf einer amerikanischen Zivilprozessordnung nach Federal Rules of Civil Procedure [24]	13
Abbildung 4: The Electronic Discovery Reference Model [33]	19
Abbildung 5: Magic Quadrant for E-Discovery Software [40]	21
Abbildung 6: LexisNexis eDiscovery Lösung [47]	22
Abbildung 7: Die verschiedenen Entwicklungen, die Cloud Computing ermöglichen [59].....	28
Abbildung 8: Cloud Service Architektur [67]	29
Abbildung 9: eDiscovery Plan für Cloud Migration [73]	33
Abbildung 10: X1 Rapid Discovery [45]	34
Abbildung 11: MapReduce Überblick [10].....	37
Abbildung 12: Ablauf eines MapReduce Jobs unter Hadoop [86] (S. 168).....	42
Abbildung 13: Cloudera’s Hadoop Distribution [93].....	44
Abbildung 14: Cloudera Manager Architektur [94]	45
Abbildung 15: Funktionsweise FreeEed	47
Abbildung 16: Funktionsweise SHMCloud [101].....	48
Abbildung 17: FreeEed SHMCloud Timeline [103].....	49
Abbildung 18: FreeEed/SHMCloud Prozedur im Vergleich Zukunft und Aktuell [105].....	50
Abbildung 19: Enkive Webfrontend	52
Abbildung 20: Die am meisten angeforderte Medien bei Gerichtsverfahren [108]	53
Abbildung 21: Nuix CPU	54
Abbildung 22: Nuix Memory.....	55
Abbildung 23: Nuix Disk Write Bytes/sec	55
Abbildung 24: Nuix Disk Read Bytes/sec	55
Abbildung 25: SHMCloud CPU	56
Abbildung 26: SHMCloud Memory	56
Abbildung 27: SHMCloud Disk Write Bytes/sec	57
Abbildung 28: SHMCloud Disk Read Bytes/sec.....	57
Abbildung 29: Amazon EC2 Instanzen für SHMCloud	58
Abbildung 30: CPU Auslastung Amazon EC2	59
Abbildung 31: Nuix Encoding und Zeitzone.....	60
Abbildung 32: Hadoop-Cluster Aufbau im Netzwerklabor der FH St. Pölten.....	67
Abbildung 33: Ubuntu MAAS Web GUI.....	68
Abbildung 34: Cloudera Manager Installation des Clusters.....	70
Abbildung 35: FreeEed Stage	77
Abbildung 36: SHMCloud S3 Setup.....	80
Abbildung 37: SHMCloud EC2 Setup	81
Abbildung 38: SHMCloud Test mit 10 Nodes.....	96
Abbildung 39: SHMCloud Test mit 1 Node	97

Tabellenverzeichnis

Tabelle 1: Metadatenfelder [2]	24
Tabelle 2: Vergleich Data Center Größe und Preisspanne der Ressourcen [69]	31
Tabelle 3: Strompreis pro kWh in verschiedenen US Bundesstaaten [69]	31
Tabelle 4: Editionsübersicht FreeEed/SHMCloud [105]	51
Tabelle 5: Datenanalyse NUX vs. SHMCloud.....	54
Tabelle 6: Bearbeitungsdauer SHMCloud in der Amazon Cloud	58
Tabelle 7: Erkannte Fileformate	59
Tabelle 8: Überblick Bearbeitungsdauer und erkannte Fileformate	60
Tabelle 9: Ergebnis des Vergleichs.....	62
Tabelle 10: Hadoop Testprogramme [80].....	71
Tabelle 11: Hadoop Beispielprogramme [80].....	74
Tabelle 12: Übersicht Fileformattest Nuix	88
Tabelle 13: Übersicht Fileformattest SHMCloud/FreeEed	95

1 Einleitung

Jedes Jahr werden von Unternehmen enorm große Mengen an elektronischen Daten aus einer Vielzahl von Quellen gesammelt, verarbeitet und für Entscheidungen herangezogen. Alle Unternehmen, die eine Geschäftsbeziehung mit den USA pflegen, müssen spätestens bei einer juristischen Auseinandersetzung alle elektronisch relevanten Informationen bereitstellen. In der sogenannten Pre-Trial-Discovery-Phase kann ein Kläger vom Beklagten nach US-amerikanischem Zivilrecht (Federal Rules of Civil Procedure, (FRCP)) elektronisch gespeicherte Informationen verlangen. Die Aufklärung dieses Sachverhaltes bezieht sich dabei auf alle relevanten elektronischen Informationen. Dies basiert auf der Erweiterung der FRCP um die Rule 34 (A) im Zuge der vorgerichtlichen Informationsbereitstellung. [1]

Um nun alle relevanten Informationen und Dokumente in Bezug auf Metadaten [2] [3] und Inhalt möglichst automatisiert eingrenzen und durchsuchen zu können, gibt es entsprechende Softwarelösungen. Diese müssen in der Lage sein unterschiedliche Dokumenttypen zu interpretieren, den Text zu extrahieren und diesen dann zu analysieren. Dies ist ein sehr ressourcenintensiver Prozess und kann bei großen Datenmengen sehr viel Zeit in Anspruch nehmen. Heutzutage werden immer mehr Daten produziert und gespeichert und auch die Größe der Datensätze nimmt zu. Laut einer Studie vom EMC ist damit zu rechnen, dass die jährlich erzeugte Datenmenge bis 2020 um den Faktor 44 steigen wird. Um diese Menge an Daten automatisiert und effektiv analysieren zu können, muss man sich künftig mit skalierbaren Verfahren wie Cloud Computing [4] beschäftigen. [5]

Cloud Computing ermöglicht es verschiedene Ressourcen wie Festplattenspeicherplatz, CPU und Arbeitsspeicher lediglich bei Bedarf kurzfristig und hochdynamisch für die Zielerreichung einzusetzen. [6]

Hierfür kann man flexible Dienste nutzen, bei denen die Dauer der verwendeten Ressourcen verrechnet wird, wie beispielsweise Amazon Web Services (AWS) [7], Amazon Elastic Compute Cloud (EC2) [8] oder eine firmeneigene Private Cloud.

Eine bekannte Lösung für Cloud Computing ist das in Java geschriebene Hadoop [9] Framework. Es beruht auf den bekannten MapReduce-Prinzipien [10] von Google und ermöglicht es intensive Rechenoperationen mit großen Datenmengen, sogenannten Big Data im Petabyte-Bereich auf Computerclustern fehlertolerant durchzuführen.

Viele der großen Internet Firmen wie Facebook, Yahoo, IBM, und Google verwenden diese Technik inzwischen. In dieser Arbeit soll nun evaluiert werden inwiefern man sich diese neue Technologie für eDiscovery [11] zu Nutzen machen kann.

Das Projekt FreeEed [12] ist eine Softwarelösung, mit der eDiscovery mit Hilfe von Hadoop betrieben werden kann. In dieser Arbeit soll nun die etablierte eDiscovery Software mit dem Open Source Tool FreeEed verglichen werden. Daraus ergibt sich folgende Fragestellung:

Was sind die Stärken und Schwächen von Hadoop basiertem eDiscovery im Vergleich zu bereits etablierten Lösungen?

Um dies zu evaluieren, sollen die wichtigsten Bewertungskriterien, welche für eine eDiscovery Software relevant sind, definiert werden um daraufhin eine systematische Bewertung durchzuführen. Zusätzlich soll eine Reihe von Use-Cases implementiert werden, in denen man Hadoop-basiertes eDiscovery mit einem konkurrierenden etablierten Produkt vergleicht. Abschließend soll eine Bewertung abgegeben werden aus der hervorgeht, wie gut die jeweilige Software die Eigenschaften unterstützt.

Das Thema 'eDiscovery with Hadoop: Is Open Source an option?' hat Einflüsse aus den unterschiedlichsten Themengebieten. Wenn man sich mit dem Thema eDiscovery beschäftigt, muss man sich zwangsweise auch mit den entsprechenden Gesetzen beschäftigen. Hierzu gehören sowohl die amerikanischen Federal Rules of Civil Procedure wie auch der europäische Datenschutz. Die Open Source Software FreeEed ermöglicht das Cloud basierende eDiscovery auf Basis von Apache Hadoop. Somit ergibt sich der Einfluss von Cloud Computing, Hadoop und Open Source. Um nun feststellen zu können, ob diese Open Source Lösung wirklich eine Alternative darstellt, wird FreeEed in dieser Arbeit mit etablierten Software Produkten verglichen. Somit ergibt sich eine sehr breitgefächerte Mischung aus den unterschiedlichsten Themengebieten.

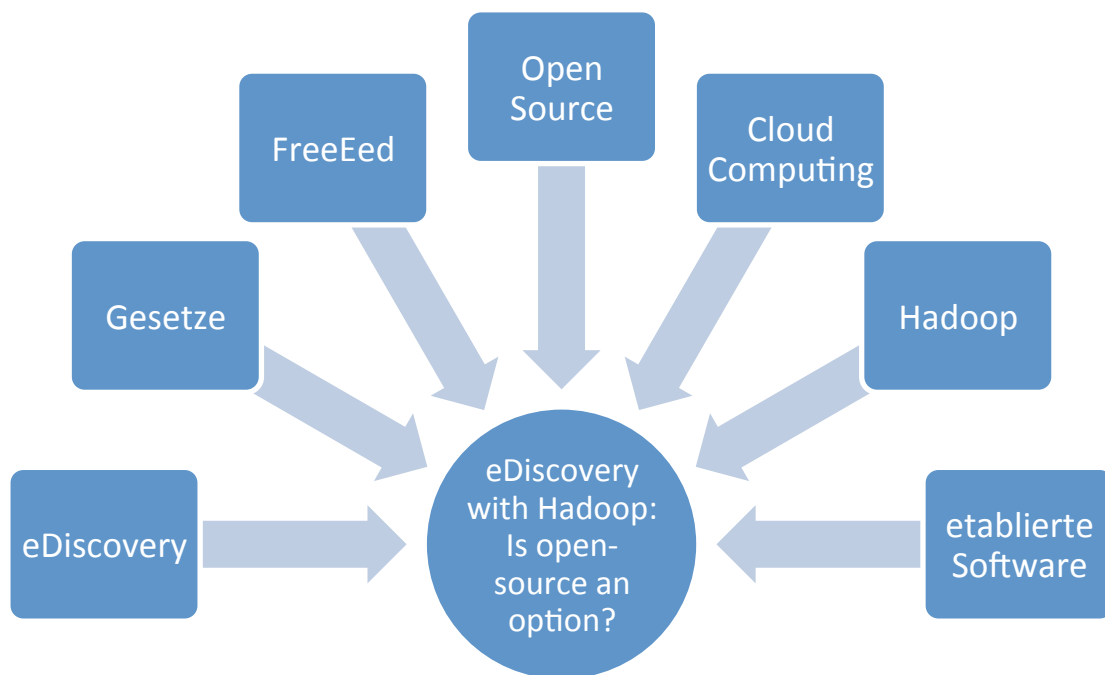


Abbildung 1: Einflüsse auf das Thema: 'eDiscovery with Hadoop: Is Open Source an option?'

Im ersten großen Kapitel wird der Begriff eDiscovery definiert und einige Anwendungsbeispiele aufgezeigt, welche die Relevanz von eDiscovery verdeutlichen sollen. Die darauf folgenden Kapitel beschäftigen sich mit der Gesetzeslage rund um eDiscovery. Sowohl die amerikanische als auch die europäische Rechtslage wird behandelt. Im Kapitel 2.4 werden Gefahren und Probleme, die bei eDiscovery auftreten, aufgezählt. Dazu werden auch Maßnahmen genannt um die Probleme und Gefahren zu minimieren. Das darauf folgende Kapitel erläutert das 'Electronic Discovery Reference Model'. Das Kapitel 2.6 betrachtet die bereits etablierten eDiscovery Software Lösungen. Das umfangreiche Kapitel Cloud Computing beschäftigt sich nicht nur mit den Grundlagen von Cloud Computing, sondern auch mit 'eDiscovery in der Cloud'. Auch die Cloud Technologien MapReduce und Hadoop werden in diesem Kapitel erläutert. Im Kapitel 3.5 wird die eDiscovery Software FreeEed und SHMCloud beschrieben. Im Kapitel 4 wird die Open Source eDiscovery Software mit einer etablierten eDiscovery Software verglichen und bewertet. Das Kapitel 5 gibt einen Ausblick auf die zukünftige Entwicklung des Open Source Projektes. Die Arbeit schließt mit einem Fazit in Kapitel 6 ab, indem nochmals die Stärken und Schwächen einer Hadoop basierten eDiscovery Software herausgearbeitet wurden.

2 Electronic discovery

Die rasante Ausbreitung der Informationstechnik und des Internets haben die Wirtschaft verändert. Vor nicht allzu langer Zeit waren sämtliche Informationen nur in Papierform verfügbar. Diese Informationen wurden in Aktenschränken aufbewahrt. Es war eine sehr mühevoll Aufgabe, bestimmte Informationen in den etlichen Aktenschränken wieder zu finden. [2]

Glücklicherweise haben sich die Zeiten geändert und der technische Fortschritt ermöglicht es uns heutzutage Informationen in Dateien zu speichern und auf Fileservern oder in der Cloud abzulegen. In Unternehmen sammeln sich eine Menge an elektronischen Daten aus einer Vielzahl von internen und externen Quellen an. Diese werden verarbeitet und für Geschäftsentscheidungen herangezogen. Wenn das Unternehmen eine Geschäftsbeziehung mit den USA pflegt, müssen spätestens bei einer gerichtlichen Auseinandersetzung elektronische Informationen gemäß Electronic Discovery-Rules bereitgestellt werden, für den Fall, dass diese als Beweismittel in Betracht kommen würden. [13] „Das amerikanische Rechtssystem verfolgt den Grundgedanken, dass durch die Bereitstellung aller Informationen die Wahrheit im Prozess ans Licht kommt. Dazu bedarf es der bedingungslosen Offenlegung aller relevanten Dokumente.“ [5] Der Anspruch zur Offenlegung aller relevanten Informationen kann schon erhoben werden, bevor es zu einem gerichtlichen Verfahren kommt. [1] Electronic Discovery hat seinen Ursprung in der Neufassung der US Federal Rules of Civil Procedure (FRCP) 34 (A) mit Wirkung vom 01.12.2006. In dieser Rule heißt es: „Rule 34 (A) any designated documents or electronically stored information including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form.“ [14] Es gibt noch eine ganze Reihe anderer Rules aus den Federal Rules of Civil Procedure (FRCP) und Federal Rules of Evidence (FRE), die im Zusammenhang mit eDiscovery relevant sind: FRCP 16, FRCP 26, FRCP 33, FRCP 34, FRCP 37, FRCP 45, FRE 502 und FRE 901. [15] In dieser Arbeit wird auf die genannten Rules nicht weiter eingegangen. Die entsprechenden Passagen können unter der Website des United States Courts¹ eingesehen werden.

2.1 Was ist Electronic discovery?

Electronic discovery (oder e-discovery, eDiscovery) ist der Prozess, um aus einer sehr großen Menge von Daten juristisch relevante Informationen herausfiltern. Hierzu müssen Informationen identifiziert, in Beziehung gesetzt, aufbereitet und zur Verfügung gestellt werden. [16] Zur Unterstützung des eDiscovery Prozesses gibt es bereits zahlreiche Softwarelösungen. Eine eDiscovery Software muss die Spreu möglichst schnell und zuverlässig vom Weizen trennen und die entsprechenden Dokumente in Bezug auf Metadaten sowie auf Inhalt geeignet eingrenzen und durchsuchen. Die Software muss also in der Lage sein, unterschiedlichste Dokumententypen interpretieren zu können, Texte zu extrahieren und diese dann zu analysieren. [5]

Analysiert werden alle elektronisch gespeicherten Informationen - auch Electronically Stored Information (ESI) [17] genannt. ESI sind elektronische Dokumente oder Datensammlungen. Dabei kann es sich auch um gelöschte Dateien oder Daten handeln, welche mithilfe der Computer-Forensik [18] wieder hergestellt werden können. Ebenso sind jene Daten relevant, die ein Computersystem automatisiert generiert, ohne Zutun und teilweise ohne Kenntnisse der Nutzer wie z.B. Zugriffsprotokolle, Arbeitszeitaufzeichnungen usw.. [16] Trotz der explosionsartigen Zunahme elektronischer Inhalte gibt es immer noch jede Menge papierbasierter Informationen in vielen Unternehmen, die sich den heutigen Such- und Ermittlungsarten entziehen. Daher sollte eine gute eDiscovery Software dafür sorgen, dass gescannte und grafische Inhalte ebenso verwertbar sind, wie elektronisch erstellte und gespeicherte Informationen (ESI). Durch das Einscannen von papierbasierten Verträgen, Schaltbildern, Diagrammen, Faxesendungen u.v.m., wird der Inhalt jeder einzelnen Datei per optischer Zeichenerkennung (OCR) durchsuchbar. Darüber hinaus können diese Dokumente um durchsuchbare Metadaten erweitert werden, um die Datei- und Dokumenteigenschaften zu simulieren, die elektronische Dateien

¹ <http://www.uscourts.gov/RulesAndPolicies/FederalRulemaking/RulesAndForms.aspx>

standardmäßig besitzen. Es kann so bei der eDiscovery Suche ein Datum gefunden werden, das in der Kopfzeile eines digitalisierten Faxes steht. [19]

2.1.1 Anwendungsbeispiele für eDiscovery

In den USA entwickelt sich eDiscovery teilweise zu einer Möglichkeit, um unliebsame Mitbewerber unter Ausnutzung der Instrumentarien der US-Gerichtsbarkeit vor einem Prozess in Bedrängnis zu bringen. Inzwischen gibt es eine mächtige Klägerindustrie. Spezialisierte Anwaltsfirmen betreiben eDiscovery Klagen als Geschäftsmodell, werben öffentlich für Sammelklagen und bieten Beteiligungsmodelle für Investoren mit hohen Gewinnerwartungen an. Das Thema eDiscovery spielt nicht nur in amerikanischen Gerichtsverfahren eine Rolle. Die Aufklärung von Sachverhalten durch IT gestützte Beweissammlung erfährt auch im europäischen Kartellrecht und bei der Umsetzung von Compliance-Maßnahmen in Unternehmen immer größere Bedeutung. In Europa sind es die polizeilichen Behörden, die eDiscovery betreiben. Dort gehört die Beweismittelsicherung zum Tagesgeschäft.

Ein häufiger Auslöser für eDiscovery ist die Produkthaftung. [20] „Ein Beispiel hierfür zeigt der folgende Fall einer Zivilklage gegen einen großen deutschen Automobilhersteller: In der Nähe von Houston, Texas, kommt eine 17-jährige Fahrerin auf gerader Straße durch Unachtsamkeit von der Fahrbahn ab. Sie verliert die Gewalt über das Fahrzeug, schleudert in den Gegenverkehr und stirbt noch vor Ort. Eine Klage wird angestrengt, weil der Unfall hätte vermieden werden können, wenn der Hersteller ein Fahrassistenzsystem (ESP) serienmäßig eingebaut gehabt hätte. Schließlich sei der Nutzen von ESP bewiesen und es wäre auch für den betroffenen Wagentyp verfügbar gewesen. Der Automobilhersteller musste nun innerhalb kürzester Zeit alle elektronisch gespeicherten Daten, die für eine Produkthaftung relevant sind oder auch nur zu wichtigen Informationen führen können, bereitstellen. Kommt man solchen Anforderungen nicht im geforderten Umfang nach, steigt das Verurteilungsrisiko und es drohen empfindliche Geldbußen. Auch Patentrechtsstreitigkeiten kommen oft vor. Hier treten häufig sogenannte ‘Patent Trolls’ auf: Patentjäger, die mit teils unüblichen, aber legalen Mitteln Lizenzgebühren einnehmen wollen. Ein solcher ‘Patent Troll’ verklagte zum Beispiel 2007 vor einem amerikanischen Bezirksgericht einen deutschen Sportartikelhersteller und weitere 45 Unternehmen aus der Sportbranche wegen Verletzung eines US-Patents. Es erfolgte eine ‘Discovery Order’, die den deutschen Hersteller zwang, innerhalb von acht Wochen für zirka 450 Schuhmodelle seit 2001 sämtliche Dokumente vorzulegen, die verfahrensrelevant sind. Dafür mussten acht Terabyte Daten von Festplatten, Mailboxen und File Server durchsucht und dem Kläger übergeben werden. E-Mails spielen überhaupt eine große Rolle in solchen Verfahren, wie auch der Fall Arthur Anderson zeigt. Dieses Unternehmen, das immerhin zu den Big Five Prüfungsgesellschaften der Welt zählte, verschwand letztlich aufgrund einer internen E-Mail vom Markt. Die Firmenjuristin Nancy Temple empfahl per E-Mail David Duncan, einem Anderson-Büroleiter, einige Dokumente zum Enron-Fall zu vernichten, obwohl die Untersuchung der US-Börsenaufsicht bereits in vollem Gange war. Diese Mail wurde ‘discovered’ und Arthur Andersen wegen Behinderung der Justiz verurteilt. Kurz darauf löste sich die Gesellschaft auf.“ [20] Dies sind nur einige Beispiele aus der Praxis. Es gibt noch eine ganze Reihe anderer Rechtsstreitigkeiten, die eine ‘Discovery Order’ auslösen können. Die wichtigsten sind in der Abbildung 2: Anwendungsbeispiele für eDiscovery [21] aufgeführt.

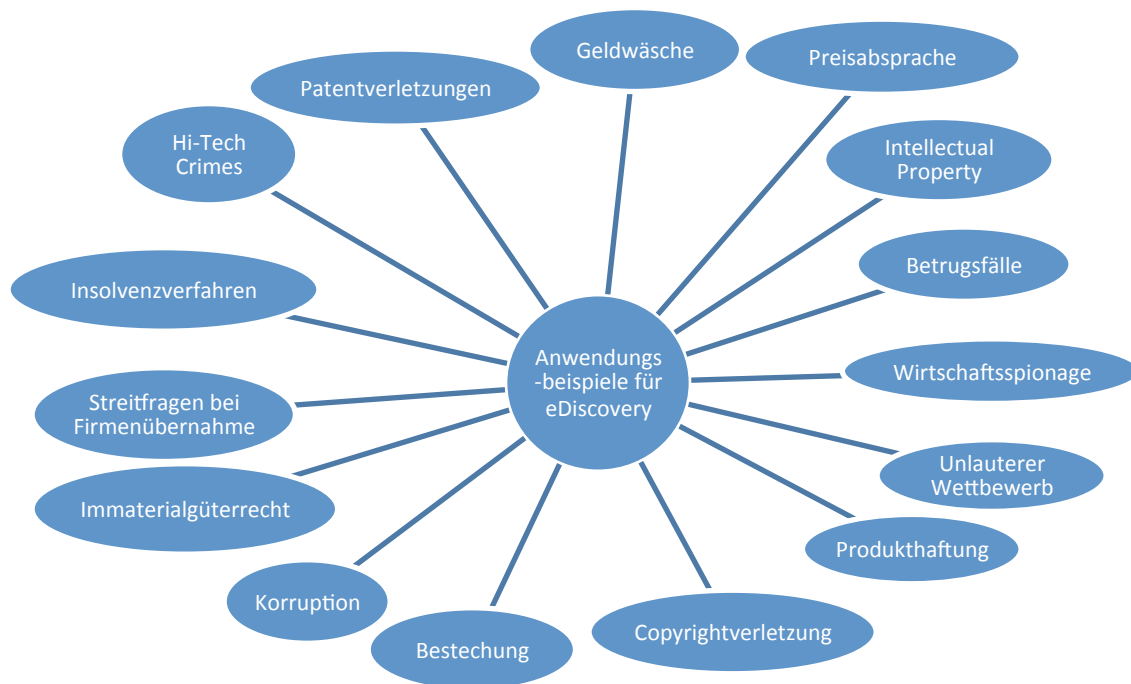


Abbildung 2: Anwendungsbeispiele für eDiscovery [21]

2.2 Die amerikanische Zivilprozessordnung

An dieser Stelle ein kurzer Exkurs in die amerikanische Zivilprozessordnung, damit man die Zusammenhänge von eDiscovery mit der Gesetzgebung versteht. Die amerikanische Zivilprozessordnung ist in drei Teile gegliedert: 'Discovery & Disclosure (Pre-Trial)', 'Trials' und 'Judgment' siehe Abbildung 3. eDiscovery kann als ein Teilprozess der 'Pre-Trial' Phase (Voruntersuchung) betrachtet werden. Der Auslöser für eDiscovery können amerikanische Rechtsanwälte, Insolvenzverwalter, Revisoren, Steuerfahnder oder auch Behörden sein, die durch einen Schriftsatz an das Gericht eine Klage gegen ein Unternehmen erheben und damit jede Information abfragen, die für den Klageanspruch von Bedeutung sein könnte. Dieser Anstoß ermöglicht dem Gericht dem Beklagten einen 'Preservation Letter' oder eine 'Discovery Order' (eine schriftliche Anforderung) zu senden, wodurch er aufgefordert wird, bestimmte Beweismittel ordnungsgemäß der eDiscovery Regelungen zu sichern. Gleichzeitig wird auch ein 'Litigation Hold' ausgegeben, was im Kern eine Pflicht zum Löscherbot prozessrelevanter Daten und zur Datensicherung für Prozesszwecke ist. [13]

Kommt bei den 'Trials' (Gerichtsverhandlungen) heraus, dass die beklagte Prozesspartei die prozessrelevanten Daten nicht aufbewahrt hat, gilt dies als Beweisvereitelung 'Spoliation'. Dies kann zu prozessualen Sanktionen und Geldstrafen führen. So kann der Richter eine 'Adverse Interference Order' erlassen, mit der die Jury belehrt wird, dass das vernichtete Dokument gegen die Partei spricht, die es vernichtet hat. [22]

Die letzte Phase im Verfahren ist das 'Judgment' (Gerichtsurteil). Hier wird das Urteil gefällt und entsprechende Sanktionen (Strafmaßnahmen) verhängt. [23]

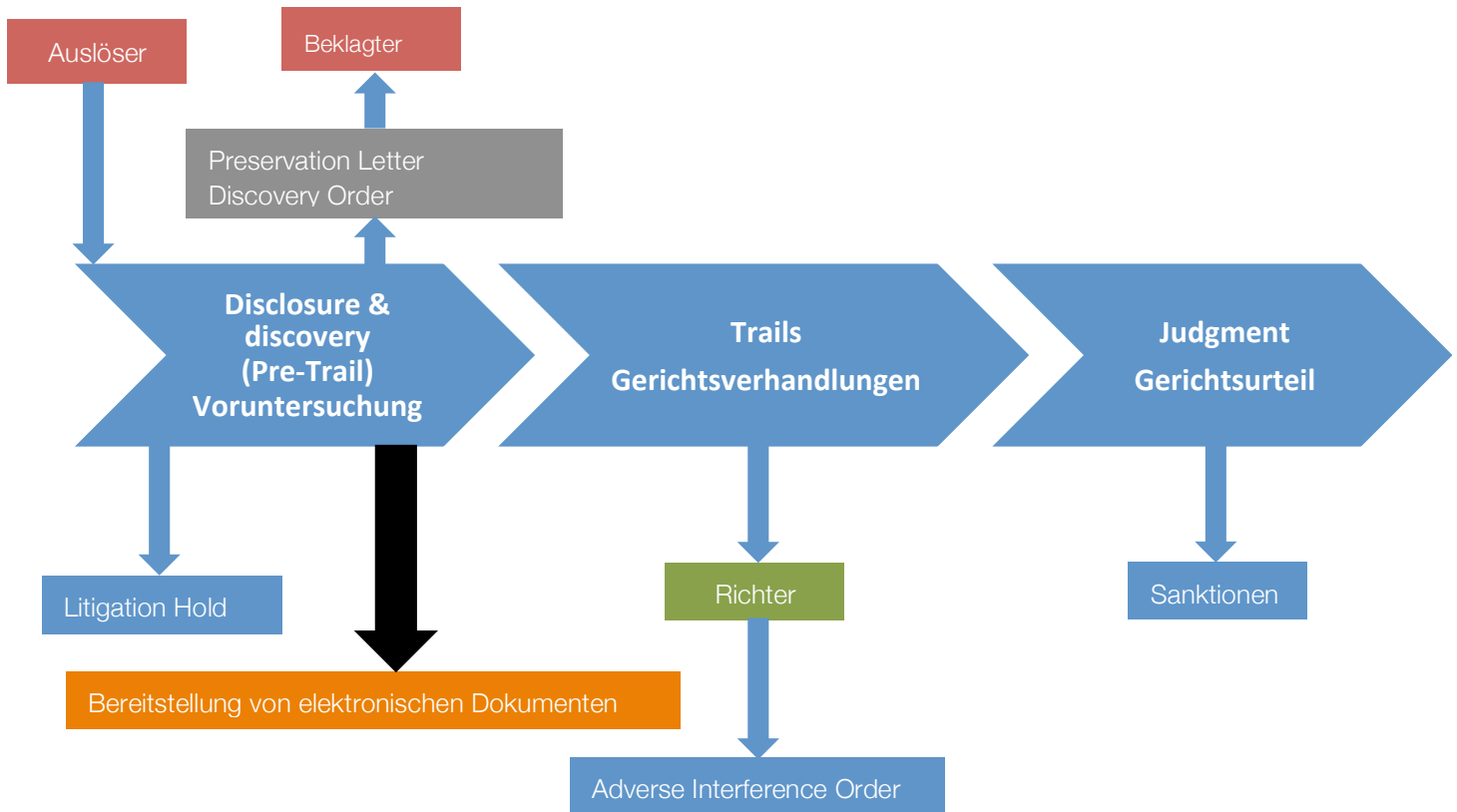


Abbildung 3: Ablauf einer amerikanischen Zivilprozessordnung nach Federal Rules of Civil Procedure [24]

Einen interessanten Einblick wie hoch Sanktionen bei Nichteinhaltung der 'Discovery Order' sein können, zeigt die Publikation 'Sanctions for E-Discovery violations: By the numbers im DUKE Law Journal'. Ein bekannter Fall hieraus war der Fall Qualcomm Inc. vs. Broadcom Corp. [23]

Die Klägerpartei Qualcomm klagte Broadcom wegen mehreren Patentverletzungen im Bereich Digitaler Video Standards an. [25] Die Angeklagte Partei Broadcom forderte während der Verfahrensvorbereitung Informationen von Qualcomm. Während den Verhandlungen kam heraus, dass die Qualcomm Anwälte Kenntnis von 21 Fallrelevanten E-Mails hatten, welche sie aber vorenthielten. Als ein Zeuge die Existenz dieser E-Mails bezeugte, verlor Qualcomm den Fall. Qualcomm wies die Schuld ihren externen Anwälten zu und das Gericht berücksichtigte, dass Qualcomm über ein umfassendes firmeninternes Rechtsberatersteam verfügte und über die entsprechenden Möglichkeiten verfügte, die geforderten Informationen zur Verfügung zu stellen. Da Qualcomm dem Gericht und dem Beklagten Broadcom diese Informationen vorenthalten hatte, bestrafte das Gericht Qualcomm mit einer Sanktionszahlung an Broadcom für diese Missachtung mit \$ 8 568 633,24.

Zusätzlich berichtete das Gericht diese Angelegenheit dem 'State Bar of California' für mögliche Disziplinarverfahren gegenüber den Qualcomm Anwälten. [3] Dieser Fall zeigt sehr deutlich welche hohen Strafen und Sanktionen auf ein Unternehmen zukommen können, wenn diese die 'Discovery Order' missachten. Dies gilt nicht nur für US Unternehmen sondern auch für europäische Unternehmen. Eine Missachtung der 'Discovery Order' kann wie man sieht, schwerwiegende Folgen haben. In wie weit eDiscovery uns in Deutschland oder Österreich betrifft und wie die Gesetzgebung hiermit umgeht wird im Kapitel '2.3 eDiscovery in Deutschland und Österreich' beschrieben.

2.3 eDiscovery in Deutschland und Österreich

Für Daten, die in Deutschland gespeichert sind, wäre das Thema eDiscovery schnell vom Tisch, wenn hier das Haager Beweisübereinkommen (HBÜ) greifen würde. [26] Das US-Gericht müsste dann ein Rechtshilfeersuchen an die Zentrale Landesbehörde richten und die Beweiserhebung durch das zuständige Amtsgericht beantragen.

Deutschland hat allerdings nach Art. 23 des Haager Beweisübereinkommens den Vorbehalt erklärt, bei Rechtshilfeersuchen der 'pretrial discovery of documents' nicht nachzukommen. Ein auf eDiscovery gerichtetes Rechtshilfeersuchen eines US-Gerichts würde somit schlicht abgelehnt werden. Der US Supreme Court hat allerdings entschieden, dass US-Gerichte einer Partei, die ihrer Gerichtsbarkeit unterliegen, die Beschaffung von im Ausland gelegenen Beweisen auferlegen kann, ohne ein Rechtshilfeersuchen nach dem Haager Beweisübereinkommen erlassen zu müssen. [27]

Deutsche oder österreichische Organe können nicht zur Amtshilfe verpflichtet werden. Es existieren auch keine unmittelbaren Pflichten für deutsche oder österreichische Unternehmen einer Discovery Order nachzukommen. Schließlich besitzen Länder in der Europäischen Union ein anderes Rechtssystem und haben ein völlig anderes Verständnis von Datenschutz. [1]

Rein rechtlich betrachtet könnte sich daher eine europäische Gesellschaft, die nicht an einem Prozess in den USA beteiligt ist, weigern, die angeforderten Unterlagen herauszugeben. Bei Nichteinhaltung einer 'Discovery Order' kann es nach der 'Doctrine of Spoliation' zu erheblichen Sanktionen für die am Prozess beteiligte Konzerngesellschaft kommen. Weigert sich die europäische Konzerngesellschaft der 'Discovery Order' nachzukommen, kann dies für die Prozesspartei in den USA zu folgendem führen: Zum Ausschluss eigener Beweismittel und einer Art Beweislastumkehr oder finanzielle Sanktionen wegen 'Contempt of Court'.

Daher können sich die Vorschriften rund um eDiscovery auch auf in Europa ansässige Unternehmen auswirken. [28] So ist es für multinationale Konzerne heute kaum möglich, eDiscovery einfach zu ignorieren. Die besondere Schwierigkeit, die sich diesen Unternehmen stellt, liegt darin, dass eDiscovery und der Europäische Datenschutz einander scheinbar widersprechen. Die Pre-trial Discovery Phase nach US-amerikanischem Vorbild verlangt die schonungslose Offenlegung aller im weitesten Sinn für den Fall relevanten Unterlagen eines Unternehmens noch vor dem eigentlichen Prozess. Der Datenschutz nach europäischem Vorbild schränkt eine solche Handlung massiv ein. [16]

In Deutschland, wo das Datenschutzgesetz sehr streng ist, hat man laut Bundesdatenschutzgesetz (BDSG) das Recht auf informationelle Selbstbestimmung. Das heißt, wenn ein Unternehmen seinem Mitarbeiter einen Laptop oder PC zur Verfügung stellt mit dem er arbeiten soll, so sind sämtliche gespeicherte Daten als persönlich anzusehen. Das Unternehmen braucht die individuelle Erlaubnis des Mitarbeiters, bevor jemand drittes auf die Informationen zugreifen darf. [13]

Doch multinationale Unternehmen müssen diesen Widerspruch trotzdem irgendwie lösen. [16] Ein Ansatz hierfür wird im folgenden Kapitel 2.3.1 Standardprozedur zur Durchführung von eDiscovery in Europa vorgestellt.

2.3.1 Standardprozedur zur Durchführung von eDiscovery in Europa

Die Erkenntnis, dass zwischen europäischem Datenschutz und dem US-amerikanischen Erfordernis der vollumfänglichen Offenlegung ein Spannungsverhältnis besteht, hilft dem multinationalen Konzern wenig. Es muss ein Mittelweg gesucht werden, der zwar nicht perfekt sein wird aber möglichst praktikabel. Die eDiscovery Prozedur kann in vielerlei Hinsicht variieren und angepasst werden. Die eDiscovery Prozedur lässt sich grob in folgende Schritte zusammenfassen:

1. Im ersten Schritt werden in allen europäischen Ländergesellschaften, die über möglich relevante Daten verfügen, forensisch korrekte Kopien der zuvor identifizierten Daten gesammelt (targeted collection). Im Nachgang zur schriftlichen Legal Hold-Mitteilung werden die Mitarbeiter mittels Fragebogens aufgefordert, die Systeme zu identifizieren, auf denen sie für den Fall potenziell relevante Dokumente und Informationen gespeichert haben. So kann vermieden werden, dass es zu einem unverhältnismäßigen Einsammeln

irrelevanter Daten kommt. Mit der Legal Hold-Mitteilung und solchen Interviews kann nicht nur vermieden werden, dass möglicherweise relevante Dokumente gelöscht oder verändert werden, sondern auch dass klar irrelevante Daten gar nicht eingesammelt werden. Die Information und der Befragungsprozess der Mitarbeiter dient auch dem Datenschutz und erlaubt es dem Unternehmen gleichzeitig seinen datenschutzrechtlichen Transparenzpflichten nachzukommen. Mitarbeiter werden daran erinnert, dass sie ihre privaten Inhalte entsprechend der im Betrieb geltenden Richtlinien gar nicht oder nur an bestimmten Stellen (z.B. besonderen Verzeichnissen auf der lokalen Festplatte) speichern dürfen.

2. Die in den europäischen Ländergesellschaften gesammelten Daten werden häufig an einer zentralen Stelle in Europa gesammelt. Dies kann in einer konzerneigenen oder von einem eDiscovery-Service-Provider angebotenen Datenbank zusammengetragen (early case assessment databases) werden. Der Datentransfer innerhalb Europas ist datenschutzrechtlich normalerweise unproblematisch. Der Zugriff auf die sichergestellten Daten sollte aus Datenschutzgründen nur einem kleinen Kreis von entsprechend geschulten und pro Fall autorisierten Personen erlaubt sein.
3. Die gesammelten Informationen können in der early case assessment database einer oder mehreren semi-automatischen Filterungen unterzogen werden, mit dem Ziel, irrelevante Dokumente zu identifizieren und physisch oder zumindest logisch zu entfernen. Die Filterung (culling) erfolgt manuell, indem eDiscovery Experten zusammen mit Personen, die über das entsprechende Fachwissen verfügen, Filter- und Suchparameter definieren, diese testen und soweit nötig präzisieren (z.B. Stichwörter, Datumangaben, Speicherort, Dokumentenbezeichnung, Sender- und Empfängernamen). Das Ziel ist es tatsächlich relevante Dokumente bzw. klar irrelevante Dokumente möglichst gut identifizieren und extrahieren zu können. Dieses Vorgehen dient nicht nur dem Datenschutz, sondern ist letztlich auch eine Maßnahme zur Kostensenkung, denn je geringer die Menge an Daten, desto tiefer werden letztlich die Kosten deren Sichtung sein. Das Ergebnis dieses Schrittes ist eine deutlich reduzierte Datensammlung vermutlich relevanter Daten.
4. In diesem Schritt werden die für den US-Prozess gesammelten und vorgefilterten 'vermutlich relevanten' (likely relevant) Daten normalerweise den eigenen Anwälten in den USA übermittelt oder diese dort via Fernzugriff auf das Review System eines eDiscovery Service Providers in den USA oder Europa zugänglich gemacht. Erst in diesem Schritt werden die nach der Filterung verbliebenen Daten im Hinblick auf den Fall manuell gesichtet. Die Sichtung dient verschiedenen Zwecken: Zum einen werden Dokumente, die aus Gründen des 'legal privilege' nicht offengelegt werden müssen oder klar irrelevant sind, ausgesondert. Zum anderen dient die Sichtung der eigentlichen Ermittlung des Sachverhaltes dem späteren Prozess. Die Sichtung kann jedoch auch benutzt werden, um datenschutzrechtlich problematische Dokumente auszuordnen. So etwa private oder sonst irrelevante Dokumente.
5. Im fünften Schritt erfolgt schließlich die Offenlegung der im vorherigen Schritt gesichteten und ggf. geschwärzten Daten durch die Anwälte in den USA an die Gegenpartei. [16]

2.4 Probleme und Gefahren bei eDiscovery

Laut einer Studie von Symantec verlieren unzählige Anwälte Fälle durch mindere Qualität bei der elektronischen Beweissicherung. Für diese Studie wurden Anwälte im gesamten EMEA (Europe, Middle-East, Africa) Wirtschaftsraum befragt. Wie die Studie aufdeckt, hatte jeder der insgesamt 5 000 befragten Anwälte Schwierigkeiten bei der Beweissicherung elektronischer Daten. Oft konnten die befragten Juristen wichtige elektronisch gespeicherte Daten zur Beweisführung überhaupt nicht oder nicht fristgerecht vorlegen. Die Hälfte der befragten Anwälte gab zu, allein in den vergangenen drei Monaten auf dieses Problem gestoßen zu sein. [29] Damit die oft kurzen Fristen, welche im 'Preservation Letter' mitgeteilt werden auch eingehalten werden können bzw. der eDiscovery Prozess überhaupt funktioniert, muss sich ein Unternehmen im Vorhinein richtig vorbereiten. Wenn beurteilt werden soll, wie reaktionsfähig und kompetent ein Unternehmen auf ein angedrohtes eDiscovery Verfahren reagieren kann, spricht man auch von eDiscovery Readiness. [16] Bei einer mangelnden Vorbereitung und schlechter Information Governance können schnell Kosten in Millionenhöhe pro eDiscovery Fall anfallen. Ein großer eDiscovery Fall und gleichzeitig auch ein Negativbeispiel ist der Fall von Intel und AMD.

Einer der größten eDiscovery Fälle traf Intel total unvorbereitet. AMD klagte im Jahre 2005 gegen seinen Rivalen Intel wegen wettbewerbsfeindlichen Praktiken auf dem Mikrochipmarkt. Beide Parteien bemerkten, dass sie wohl dem größten eDiscovery Fall überhaupt gegenüberstehen würden.

Hätte man wirklich alle Dokumente in gedruckter Form vorgelegt, wäre ein ca. 220 km hoher Papierberg entstanden. Im Jahre 2007 trug der Richter Joseph A. Farnan, Jr. Intel auf dem Gericht 1 000 E-Mails auszuhändigen, welche im Besitz von Intel sein müssten. Intel wies jedoch in einem Brief darauf hin, dass dies problematisch sei, da das von Intel verwendete Mail System Microsoft Outlook so konfiguriert war, dass alle E-Mails von Angestellten nach 35 Tagen gelöscht werden und E-Mails von leitenden Angestellten nach 60 Tagen. Dazu kam noch das Intel Nonindexed Backup Tapes verwendet. Diese Backup Strategie ist nicht für eDiscovery entwickelt worden sondern ausschließlich für das Disaster Recovery. Um nun die relevanten E-Mail Nachrichten mit den entsprechenden Stichworten und Metadaten zu finden, wäre eine gigantische Aufgabe in dieser kurzen Zeit gewesen. Jedes Backup Tape muss gemounted und wiederhergestellt werden, um den Inhalt zu prüfen und zu durchsuchen. Hier ein Ausschnitt aus dem Brief: "The overall scope of the e-mails and documents Intel will be producing is sweeping in breadth and magnitude - and will encompass the equivalent of tens of millions of pages of material from many hundreds of employees with overlapping involvement in communications, both internal and external." Das Gericht verschob daraufhin den Verhandlungstermin auf Februar 2010. [3]

Neue Technologien wie Smartphones, Tablets, Laptops und GPS-Systeme stellen heute eine weitere Schwierigkeit dar. Diese Geräte machen den heutigen Mitarbeiter erst richtig produktiv, doch genau diese Systemvielfalt erschwert den eDiscovery Prozess erheblich. All diese Informationen aus der vielfältigen Gerätelandschaft forensisch korrekt zu sichern, analysieren und zu durchsuchen stellt die eDiscovery Ermittler vor eine enorme Herausforderung. Zusätzlich kommt hinzu, dass diese mobilen Geräte oft verschlüsselt sind und das die Daten vor dem forensischen Speichern noch entschlüsselt werden müssen.[2]

Nicht nur die vielen mobilen Geräte stellen eine Herausforderung dar. Auch die verschiedenen Softwaresysteme wie beispielsweise das Vertriebsinformationssystem, CAD, Product-Lifecycle-Management (PLM), Enterprise-Resource-Planning (ERP), auf denen unter Umständen unterschiedliche Benutzernamen für den selben Benutzer verwendet werden oder auch verschiedene Zugriffssteuerungen zum Einsatz kommen. Oft kann ein Produkt je nach Bearbeiter (z.B. Konstrukteur, Einkäufer, Fertigungsplaner) oder durch unterschiedliche Bezeichnungen und unterschiedliche Landessprachen oder Identifikatoren die Suche erschweren. Die Suchroutinen erfassen in der Regel Metadaten oder Angaben im Dateinamen. Informationen, nach denen gesucht wird, stehen aber meist innerhalb von digitalen Dokumenten (z.B. CAD-Zeichnungen, Grafiken, Videoaufzeichnungen, Prüfberichten, E-Mails). Diese Files werden normalerweise in einem Format gespeichert, welche nur durch das erzeugende Programm gelesen werden können. Werden sie von einem Dokumenten-Management-System oder auf einem Fileserver verwaltet, sind die Dokumente zudem nur mit den entsprechenden Berechtigungen zu öffnen. Vor allem die Automobilindustrie arbeitet im Rahmen eines Collaborative Engineerings (CE) unternehmensübergreifend mit Zulieferern eng zusammen. Während des gesamten Produktentstehungsprozesses werden Daten und Files auf allen möglichen Kommunikationswegen und Plattformen ausgetauscht, redundant gespeichert, protokolliert und wieder zu unterschiedlichen Zeitpunkten verändert weitergereicht. Die Frage, welcher Informationsstand zu einem bestimmten Zeitpunkt welchem Anwender bekannt war, kann meist nicht genau oder nur mit erheblichem Rechercheaufwand beantwortet werden. Informationen verändern entlang der Zeitschiene sowohl ihren Inhalt (z.B. durch fortschreitenden Reifegrad in der Vorentwicklung, Erprobung und Produktion) als auch ihre Bedeutung (z.B. durch neue Testberichte, Marktanalysen oder Risikobewertungen). Policies, welche festlegen, wann nachhaltig gespeichert oder gelöscht werden muss, sind meist nicht definiert und dem Anwender nicht bekannt oder werden individuell oder situationsgerecht ausgelegt. [1]

Auch in technischer Hinsicht ergeben sich bei eDiscovery eine ganze Reihe von Fragen: Wo sind alle potentiell relevanten Daten gespeichert? Was soll ein Unternehmen tun, um Daten von veralteten und abgelösten Systemen wiederherzustellen? Wie extrahiert man die Daten von modernen Systemen wie Enterprise Portalen, sozialen Netzwerken, Blogs, Wikis oder Cloud Speichern? Müssen alte Daten konvertiert werden? Wenn ja, verändert der Konvertierungsprozess die wichtigen Metadaten? [30]

Eine weitere Gefahr bei eDiscovery ist, dass viele europäische Anwälte ihren Klienten oft die Empfehlung aussprechen im Zweifel lieber mehr offenzulegen als zu wenig. Datenschützer und Datenschutzspezialisten heben

jedoch den Finger und weisen darauf hin, dass es gegen das europäische Gesetz verstößt, Personendaten einfach so in die USA zu senden und dort offenzulegen. Dass eDiscovery in großem Umfang nicht nur fallrelevante Personendaten sondern auch viele irrelevante Personendaten zu Tage fördert, ist offenkundig. [16] Man darf nicht vergessen, dass wenn fallirrelevante Daten an die gegnerische Partei ausgehändigt werden (z.B. einem Mitbewerber) dieser aus den irrelevanten Informationen Betriebsgeheimnisse zu neuen Produkten oder andere geheime Informationen aus diesen Daten ableiten kann.

2.4.1 Informationsmanagement

Der heutige Durchschnittsuser löscht kaum mehr ein Dokument endgültig und unwiederbringlich von seiner Festplatte. Allein die Vorstellung, auf ein Dokument in naher Zukunft nicht mehr zugreifen zu können, treibt vielen von uns die Schweißperlen auf die Stirn. [31] Speicherplatz ist billig und wird immer billiger. 1 Terrabyte Festplatten bekommt man heute schon für ca. 80 €. [32] Somit besteht also keinerlei Grund sich in seiner Datensammelleidenschaft zu beschränken. Dies sehen scheinbar die meisten Computeruser ähnlich. So wächst die Menge an elektronisch gespeicherten Informationen derzeit laut einer IDG-Studie jedes Jahr um 60%. Die durchschnittliche Datenmenge pro Erdbürger beträgt heute ca. 45 Gigabyte.

Aus Sicht der Compliance sollten möglichst alle Daten gespeichert werden, um den gesetzlichen Auflagen in Bezug auf die Archivierung gerecht zu werden.

Aus der Sicht von eDiscovery stellt sich die Situation genau umgekehrt dar. Im Fall von eDiscovery stehen Firmen vor der Herausforderung, meist große Datenmengen fallspezifisch zu filtern und in kürzester Zeit bereitzustellen. Siemens beispielsweise hat die Kanzlei Debevoise & Plimpton beauftragt, archivierte Datenbestände zu analysieren, um die für die Ermittlungen und die interne Untersuchung relevanten Dokumente zu identifizieren. Siemens musste der Kanzlei geschätzte 150 Millionen € dafür bezahlen. Dabei haben Dutzende von Anwälten jedes einzelne Dokument händisch geöffnet und gelesen, um bei ca. 98% festzustellen, dass sie keinerlei Relevanz für die laufenden Ermittlungen haben. [31] Siemens ist hier sicher kein Einzelfall, die für verschlafenes eDiscovery Lehrgeld zahlen mussten. Siemens hat jedoch aus seinen Fehlern gelernt und inzwischen ein eigenes eDiscovery System namens SIinvestigate entwickelt, da keines der auf dem Markt verfügbaren Produkte den Ansprüchen gerecht geworden ist. [5]

Der von der IDG-Studie beschriebene rasante Datenanstieg macht das Auffinden benötigter Informationen schwieriger denn je. Wie die sprichwörtliche Nadel im Heuhaufen liegen relevante Informationen unter Terabytes von Inhalten auf den unterschiedlichsten Systemen, Dateiservern, Cloudspeichern und mobilen Geräten begraben. Um diesen Herausforderungen entgegen zu wirken, kann man auf neue Technologien setzen und die Cloud Techniken zur Analyse der riesigen Datenbestände nutzen. Auch eine strategisch gute Information Governance Politik und entsprechende Unternehmens Policies können den Datenbestand reduzieren und eDiscovery Kosten reduzieren. [16]

Ein anschauliches Framework für Information Governance ist das von Electronic Discovery Reference Model (EDRM) [33] veröffentlichte 'Information Governance Reference Model (IGRM)' [34]

So gilt auch bei eDiscovery das einfache Prinzip: Daten, die nicht existieren, können auch nicht offengelegt werden. Verboten man beispielsweise die private Nutzung des Internets und die private Kommunikation per E-Mail, so müssen dienstliche Daten oder E-Mails nicht von privaten getrennt werden. Zudem kann sich der Arbeitgeber absichern, nicht gegen das z.B. deutsche Fernmeldegeheimnis § 88 TKG zu verstoßen, wenn er z.B. Mitarbeiter E-Mails bei einem eDiscovery Prozess offenlegt. Auch eine Document Lifecycle Policy reduziert die Datenmenge. Nicht benötigte Daten sollten regelmäßig gelöscht werden, solange keine gesetzlichen Aufbewahrungspflichten dies verbieten. Auch aus amerikanischer Sicht ist das Löschen von Daten grundsätzlich unbedenklich, solange man sich an gemessene interne Richtlinien hält. Allerdings dürfen Daten nicht mehr gelöscht werden, wenn sich ein Rechtsstreit anbahnen könnte.“ [16]

Die Einführung eines Enterprise-Content-Management (ECM) Systems ist eine gute Möglichkeit, entsprechende Policies zur Speicherung und Verwaltung von Dokumenten umzusetzen. So kann verhindert werden, dass Terabytes an elektronisch gespeicherten Informationen doppelt vorhanden sind oder unstrukturiert abgelegt sind.

Ein ECM System ermöglicht auch eine bessere Nachvollziehbarkeit. Es lässt sich leichter feststellen, wer an welchem Dokument was verändert hat und welcher Wissensstand zu dem jeweiligen Zeitpunkt verfügbar war. [35]

2.4.2 Datenexport in die USA

Wenn elektronische Dokumente in die USA übermittelt werden sollen, egal ob an die eigenen Anwälte, an die Beweisgegner oder an das Gericht, entsteht ein Konflikt mit dem europäischen Datenschutzgesetz. Denn nach europäischem Recht dürfen Daten nur in Länder mit angemessenem Datenschutzniveau exportiert werden, was auf die USA nicht zutrifft. Eine Lösung für diesen Konflikt wird in den Binding Corporate Rules (BCR) [36] gesehen, welche die Speicherung der Dokumente entsprechend den eDiscovery Rules unter Berücksichtigung des europäischen Datenschutzrechts definiert. [22]

- Binding Corporate Rules (BCR) gelten EU-rechtlich als Lösung für die Übermittlung personenbezogener Daten außerhalb der Europäischen Union. Diese Rules sind verbindliche unternehmensweit geltende Richtlinien, die es multinationalen Konzernen ermöglichen, interne personenbezogene Daten über Landesgrenzen sowie die EU-Grenze hinweg zu exportieren und dabei das geltende europäische Datenschutzrecht einzuhalten. Diese Rules machen den Umgang und den Transport für den Betroffenen transparent. [26] Jedoch erlauben die BCRs lediglich den Transfer von personenbezogenen Daten im Unternehmensverbund. Für die Weiterleitung an Prozessgegner fordert die Artikel-29-Arbeitsgruppe, welche die BCRs veröffentlicht hat, wiederum die Unterzeichnung der [18] „Standardvertragsklauseln für die Übermittlung personenbezogener Daten.“ [37]
- Das Haager Beweisübereinkommen (HBÜ) ist ein Abkommen, welches den Beweismitteltransfer zwischen Staaten unter bestimmten Auflagen zulässt. Dafür muss jedoch das zuständige Gericht dem schriftlichen Ersuchen stattgeben, damit die Daten verarbeitet und weitergegeben werden dürfen. Der Genehmigungsprozess kann allerdings 6 – 12 Monate in Anspruch nehmen. Daher eignet sich diese Möglichkeit nur sehr bedingt für eDiscovery Anfragen. [16] Erschwerend kommt hinzu, dass beispielsweise Deutschland nach Art. 23 des Haager Beweisübereinkommens den Vorbehalt erklärt hat bei Rechtshilfeersuchen der ‘pretrial discovery of documents’ nicht nachzukommen. [27]
- Die Safe-Harbor-Regelung [36] soll die Unterschiede der Datenschutzrichtlinien zwischen den USA und Europa überbrücken. Diese Regelung wurde vom US-amerikanischen Handelsministerium und der EU-Kommission entwickelt und im Jahre 2000 von der EU gebilligt. Ursprünglich wurde diese Regelung dafür entwickelt, dass Unternehmen mit europäischen Tochtergesellschaften in Europa so operieren können, als würde es keine Grenzen geben. Inzwischen haben unzählige andere Organisationen eine Akkreditierung nach der Safe-Harbour-Regelung beantragt. Darunter zahlreiche Anbieter aus dem eDiscovery Sektor. Um diese Akkreditierung zu bekommen, muss sich das zu akkreditierende Unternehmen zertifizieren und seinen Namen in der Datenbank mit den Safe-Harbor-Unternehmen registrieren lassen. Es werden jedoch Zweifel geäußert ob die Safe-Harbour-Zertifizierung für eDiscovery Anbieter wirklich anwendbar ist. [16] Zusätzlich ist hierbei noch anzumerken, dass sich verantwortliche Stellen in den USA oftmals nur ungern den Safe-Harbor-Regelungen unterwerfen, da sie dadurch der US-Behördenaufsicht für ordnungsgemäße Umsetzung des Datenschutzes unterliegen würden und sich somit zeit- und kostenintensiven jährlichen Audits unterziehen müssten. [16] Für US-Unternehmen ist es sehr einfach an Safe-Harbor teilzunehmen. Es genügt, dass das betreffende Unternehmen versichert, an Safe-Harbor teilzunehmen. Hierfür gibt es keine unabhängigen Prüfungen oder Kontrollen. [16]

Schlussendlich muss von Fall zu Fall individuell geprüft werden, welche entsprechenden Regelungen am besten anwendbar sind. Allerdings sollten im Voraus entsprechende Vorbereitungen getroffen werden, um nicht komplett unvorbereitet vor der Herausforderung zu stehen wie man den europäischen Datenschutz und den Datenexport in die USA am besten mit sich vereint. Es empfiehlt sich schon im Voraus die Binding Corporate Rules umzusetzen und der Safe-Harbor-Regelung beizutreten.

2.5 Das Electronic Discovery Reference Model

Ein geeignetes Framework für das eDiscovery Management bildet das Electronic Discovery Reference Model (EDRM). [38] Im Mai 2005 wurde das EDRM Projekt von George Socha und Thomas Gelbman ins Leben gerufen, um eindeutige Standards und Richtlinien für den eDiscovery Markt vorzugeben. Dieses Modell ist inzwischen zum absoluten non plus ultra für den eDiscovery Prozess geworden. So gut wie jedes Unternehmen, das einen eDiscovery Fall durchführt, folgt diesen Richtlinien. [30] [39]

Das EDRM Projekt umfasst nicht nur das in Abbildung 4 gezeigte 'Electronic Discovery Reference Model' sondern auch eine ganze Reihe weitere Projekte, die von EDRM ins Leben gerufen wurden. An diesen können sich alle eDiscovery Interessensgruppen wie Rechtsberater, IT Manger, Anwaltskanzleien und Softwarehersteller beteiligen. Zu diesen Projekten gehört zum Beispiel das Projekt EDRM XML. Das Ziel dieses Projektes ist es, basierend auf dem XML Standard, Schnittstellen zur Verfügung zu stellen um Electronically Stored Information (ESI) von einem Schritt des eDiscovery Prozesses zum nächsten zu schieben. Somit auch von einer Software zur nächsten oder von der einen Organisation zur anderen.

Ein weiteres Projekt ist das EDRM Metrics Project. Das Ziel ist hier ein effektives Hilfsmittel bereitzustellen, um den eDiscovery Aufwand in Bezug auf finanzielle und administrative Hinsicht zu messen.

Ein weiteres Projekt, welches für diese Arbeit sehr interessant ist, nennt sich EDRM Data Set Project. Hier werden große Datenmengen gesammelt um unterschiedlichste Aspekte von eDiscovery Software und Services zu testen. [39] Um zum eigentlichen Electronic Discovery Reference Model aus Abbildung 4 zurückzukehren. Das Modell unterteilt den eDiscovery Prozess mit unterschiedlichen Farben in sechs Abschnitte. [30] Nachfolgend werden die Phasen des EDRM-Modells erläutert:

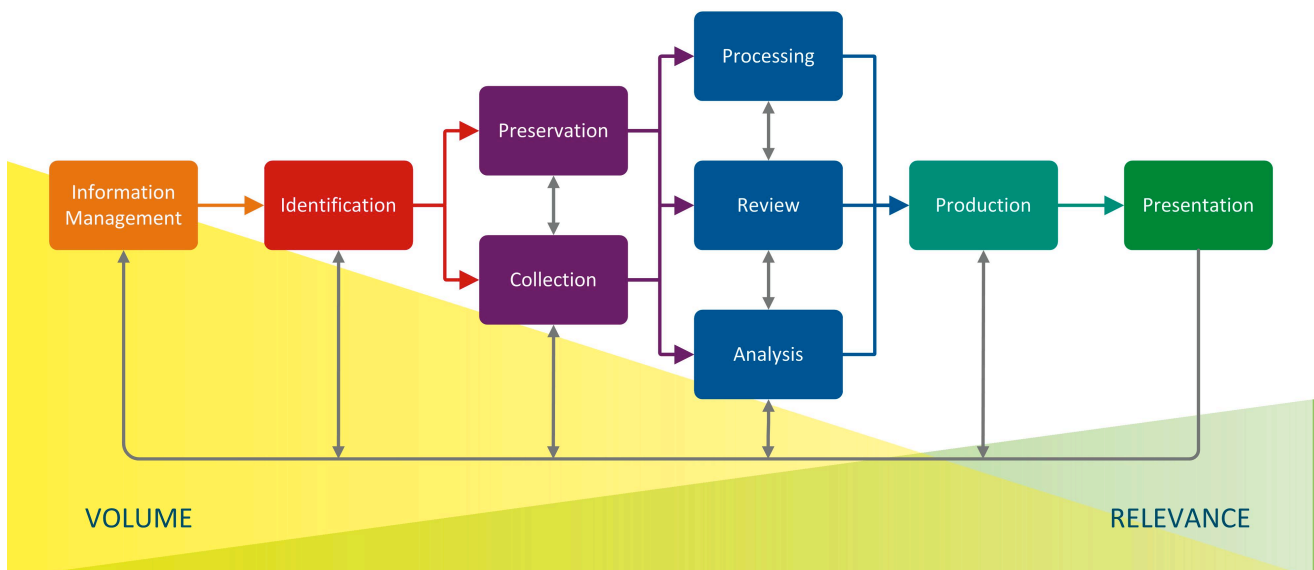


Abbildung 4: The Electronic Discovery Reference Model [33]

1. „Das 'Information Management' ist nicht explizit eDiscovery-spezifisch. Es beinhaltet den ordnungsgemäßen Umgang mit Informationen. Dazu gehört, dass Unternehmen über ein konsistentes Informationsmodell verfügen, die informationstragenden Prozesse beschrieben sind und es eine Retention Policy gibt, in der Aufbewahrung und Vernichtung von Dokumenten geregelt sind.
2. Bei 'Identification' beginnen die spezifischen eDiscovery-Aufgaben. Hier werden Rahmen, Tiefe und Breite der Untersuchung festgelegt. Es gilt, die verfahrensrelevanten Organisationseinheiten, Personen und ihre

Datenträger zu identifizieren. Letztere können Exchange Mailboxen oder Lotus Notes Archive, File Shares, aber auch SMS-Nachrichten im Handy sein.

3. Die 'Preservation-Phase' stellt sicher, dass Informationen nicht verändert oder gar vernichtet werden. Vor allem müssen Löschroutinen, die standardmäßig erfolgen, zwingend außer Kraft gesetzt werden.
4. Zum gleichen Zeitpunkt setzt die 'Collection' ein, das Einsammeln der Informationen – notfalls vor Ort oder im Home Office der betreffenden Mitarbeiter.
5. Im Schritt 'Processing' werden die verfügbaren Informationen aufbereitet, redundante und irrelevante Daten aussortiert. Beim 'Review' geht es bereits mehr um eine inhaltliche Relevanzanalyse, zum Beispiel um den Schutz von privilegierten Informationen wie der Korrespondenz zwischen Anwalt und Mandat.
6. In der 'Analyse' untersuchen und bewerten die Verantwortlichen jedes einzelne Dokumente nach Inhalt und Kontext.
7. In der 'Production-Phase' werden schließlich die relevanten Informationen gedruckt oder auf DVD gebrannt. In Einzelfällen wird ein Onlinezugriff gewährt.
8. Auch eine 'Presentation', also Vorstellung der möglichen Beweismittel vor Gericht kann vorkommen.“ [20]

2.6 Etablierte eDiscovery Software

Früher war die Auswahl von eDiscovery Produkten und Services sehr begrenzt. Der Markt entwickelte sich zögerlich und passte sich nur sehr langsam den Bedürfnissen von Rechtsseite und Großunternehmensseite an. In diesen Zeiten war es üblich alle Informationen und Dokumente in Papierform zu archivieren. Die auf Papier geschriebenen oder gedruckten Informationen mussten dann von etlichen Anwälten geprüft werden. Da diese Methode die Anwaltskosten extrem in die Höhe steigen ließ, begann man damit, hunderte Aktenordner von ortsansässigen Copyshops von Papier in elektronische Dokumente zu konvertieren. Die digitalisierten Informationen ließen sich dann schneller und günstiger durchsuchen. Aus den zu Anfang kleinen Copyshops entwickelte sich ein eigener Markt mit großen spezialisierten Anbietern für Dokumentendigitalisierung. Ein weiterer eDiscovery Geschäftszweig entstand als Anbieter aus der Forensic Branche wie AccessData Group, Microforensics, Guidance Software und ZyLAB Forensic Software Lösungen mit integrierten eDiscovery Mechanismen oder Schnittstellen auf den Markt brachten. Zuvor wurde Forensic Software hauptsächlich von Behörden zur Strafverfolgung verwendet. Durch den immer interessanter werdenden eDiscovery Markt wurden die Forensic tools auch für den zivilen Markt interessant. Als dann so gut wie alle Informationen nur noch elektronisch verarbeitet wurden, betraten auch einige traditionelle IT Anbieter den eDiscovery Markt mit Produkten, die spezielle Features implementiert hatten um z.B eDiscovery auf Mail Systemen oder Backup Storage Systemen durchzuführen. Zu diesen Herstellern gehören Größen wie IBM, EMC und Symantec.

Seit dem Jahre 2006, als die Federal Rules of Civil Procedure die zusätzliche Rule 34 (A) erhalten hat, wuchs der eDiscovery Markt schnell. Inzwischen ist der eDiscovery Markt fast überfüllt mit Herstellern für 'Standard eDiscovery'. Viele der angebotenen Produkte unterscheiden sich von der Funktionalität kaum. [40] Andere Produkte wiederum bieten wichtige Features wie z.B. eDiscovery für Soziale Netzwerke (X1 Social Discovery) oder eDiscovery Funktionen für das Cloud basierende Microsoft Office 365 (ZyLAB Cloud Collector). [41], [42], [43] Firmen, die spezielle Technologien wie hier im Beispiel Office 365 verwenden, haben eine sehr limitierte Auswahl an Produkten, die für sie und ihre verwendeten Technologien in Frage kommen.

Wie bereits im Kapitel 2.4 erwähnt, hat die Siemens AG beispielsweise sehr schnell herausgefunden, dass es für ihre hohen Anforderungen keine allumfassende Lösung gibt. Am Ende hätte man eine Sammlung von 14 unterschiedlichen Tools benötigt, die man manuell konfigurieren hätte müssen und aufgrund fehlender Interoperabilität mit großem Aufwand einzeln bedienen hätte müssen. Daher beschloss man bei der Siemens AG eine eigene eDiscovery Lösung mit allen notwendigen Anforderungen selbst zu entwickeln. [5]

Die Frage, welche eDiscovery Lösung nun die beste ist, die man als Referenzprodukt für den späteren Vergleich mit der Open Source Lösung FreeEed verwenden soll, lässt sich so nicht beantworten. Generell sollte sehr genau evaluiert werden, welche eDiscovery Lösung am besten zu einem Unternehmen passt oder ob man wie die Siemens AG eine eigene Lösung entwickeln sollte, um allen Anforderungen gerecht zu werden. Um die gesamten Anforderungen zu identifizieren, empfiehlt es sich eine Karte der gesamten IT Landschaft anzufertigen und alle Systeme, welche für eDiscovery relevant sind, festzuhalten, egal ob Serversysteme wie Datenbanksysteme, Mail-

und File Server oder Mobile Geräte wie Smartphones Tablets und Notebooks. Nachdem alle für eDiscovery relevanten Systeme identifiziert wurden, sollte nicht vergessen werden, wie man die Daten zusammensammelt, wenn mobile Geräte oder E-Mails verschlüsselt sind oder Server und Informationsmanagement Systeme in die Cloud ausgelagert wurden.

Einen guten Überblick über den eDiscovery Markt gibt die Gartner Studie 'Magic Quadrant for E-Discovery Software'. In dieser Studie sind die großen eDiscovery Anbieter sowohl für Software als auch für eDiscovery als Service aufgeführt. Die Abbildung 5: Magic Quadrant for E-Discovery Software [40] aus der Gartner Studie teilt die in der Studie erwähnten eDiscovery Anbieter in vier Kategorien ein: Marktführer, Herausforderer, Visionäre und Nischenanbieter. [40]

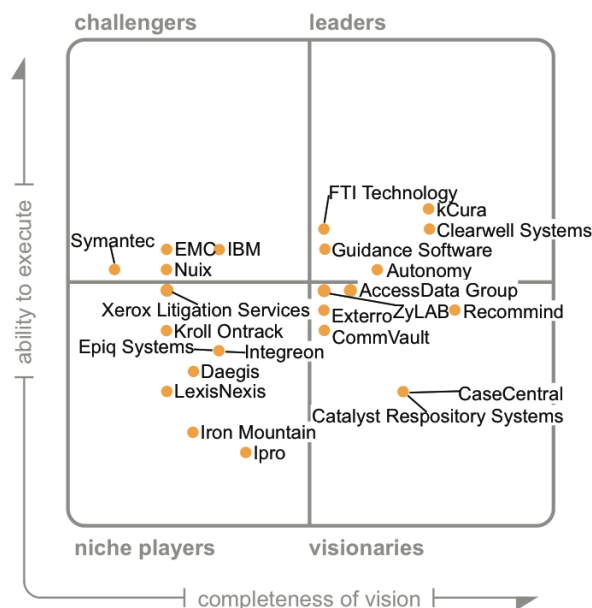


Abbildung 5: Magic Quadrant for E-Discovery Software [40]

'The EDM Buyer's Guide' wurde 2011 von der AccessData Group veröffentlicht und bietet eine sehr gute Hilfestellung bei der Auswahl einer eDiscovery Software. In diesem Guide wird zu jeder Phase, aus dem in Abbildung 4 dargestellten EDM-Modell, eine Checkliste mit Software Anforderungen aufgeführt. Diese Checkliste bietet so eine sehr gute Hilfestellung bei der eDiscovery Produktauswahl. Obwohl dieser Guide von einem eDiscovery Software Anbieter selbst veröffentlicht wurde, bietet die Liste eine sehr gute Basis und zeigt worauf zu achten ist. In Kapitel 4 wird ein Teil dieser Anforderungen aus der 'The EDM Buyer's Guide' Checkliste für den Vergleich der traditionellen eDiscovery Software mit der Open Source Lösung FreeEed herangezogen. [44]

Weitere Fragen in Bezug auf das EDM-Modell sind, welche Phasen die Software überhaupt abdecken soll und welche Schritte manuell ausgeführt werden sollen. Außerdem stellt sich die Frage, ob ein anderes Software Produkt bei bestimmten Schritten sinnvoll wäre.

Nehmen wir als Beispiel die 'Collection' Phase: Können die Collection Aufgaben von der eDiscovery Software übernommen werden oder ist die Infrastruktur so komplex, dass man hierfür mehrere unterschiedliche Software Produkte benötigt?

Ein interessantes Softwarebeispiel für die 'Collection Phase' ist das Open Source Projekt 'GRR Rapid Response Framework (GRR)'. GRR ist bisher ein 'Proof of concept' multi-platform, Open Source Tool für 'enterprise forensic investigations'. Dabei ist es möglich, remote auf den Arbeitsspeicher oder die Festplatte eines Clients zuzugreifen. [45]

Ein mögliches Szenario für den Einsatz von GRR wäre folgendes: Man erhält eine 'Discovery Order' für das Notebook eines Users. In Szenario 1 ist das Notebook des Users nicht erreichbar. In Szenario 2 ist das Notebook erreichbar aber der User soll nicht alarmiert werden (regionale Rechtslage beachten). Mit GRR kann das Problem folgendermaßen gelöst werden: In GRR wird ein sogenannter 'flow' für das Notebook des Users eingestellt. Dieser 'flow' erstellt ein Festplatten Image des Notebooks sobald die Netzwerkverbindung für die Übertragung ausreichend ist. [46] Eine ganz ähnliche kommerzielle Software Lösung bietet auch die Firma F-Response² an. [30]

Der bekannte eDiscovery Anbieter LexisNexis bietet vier einzelne Produkte für den gesamten eDiscovery Prozess an, siehe Abbildung 6.

Für das 'Early Data Assessment' wird die Software 'Early Data Analyzer' verwendet um einen Einblick in Größe und Umfang der Daten zu erhalten. Dabei können Dateien direkt auf dem Zielsystem über das Netzwerk gefiltert und durchsucht werden. Auch Duplikate und irrelevante Dateien können so vor dem zeitintensiven 'Process Data' Schritt entfernt werden.

Danach wird 'LAW PreDiscovery' verwendet, was als eigentliche 'processing engine' fungiert. Hier ist das Ziel alle potentiell für den Fall relevanten Dokumente, egal ob elektronisch oder in Papierform, nochmals detailliert zu filtern. 'LAW PreDiscovery' erlaubt es nochmalig Duplikate und irrelevante Dateien zu entfernen, um später nur noch die wirklich erforderlichen Daten im kostenintensiven Dokumenten Review Prozess zu verarbeiten.

Nachdem wirklich nur noch die relevanten Dokumente übrig sind, werden diese in die Software 'Concordance' exportiert, die dabei hilft, die große Menge an Dokumenten zu managen und zu überprüfen. Die Software erlaubt es mehreren Benutzern die Dokumente zu durchsuchen, zu überprüfen und an andere Benutzer zu verteilen, wie z.B. an eine bestimmte Expertengruppe.

Das letzte Teil in dem LexisNexis eDiscovery Puzzle ist 'CaseMap'. 'CaseMap' ist ein Tool, um effektiv fallrelevante Fakten, Akteure und Streitfragen zu sammeln, organisieren und auszuwerten. Zusätzlich bietet die Software Kollaborationsmechanismen um Teamarbeit zu ermöglichen. Außerdem gibt es die Möglichkeit, die Fakten, Akteure und Streitfragen zusammengefasst in ein PDF Report zu exportieren um dies vor Gericht präsentieren zu können. [47] (S.1–4)



Abbildung 6: LexisNexis eDiscovery Lösung [47]

Ob man nun eine 'All in One' Lösung oder eine Lösung, die aus mehreren Teilen bzw. auch von verschiedenen Herstellern zusammengestellt wurde bevorzugt, ist Geschmackssache und muss den entsprechenden Bedürfnissen individuell angepasst werden. Wichtig hierbei ist nur, dass die entsprechenden Schnittstellen miteinander kompatibel sind.

Es ist daher wichtig, dass möglichst alle Hersteller die Standards des EDRM XML Projekts einhalten. [48]

Das Beispiel mit der eDiscovery Lösung LexisNexis soll demonstrieren, wie unterschiedlich eine eDiscovery Lösung aufgebaut sein kann und dass es für den später folgenden Vergleich in Kapitel 4 wichtig ist, dass man möglichst faire Vergleichsbedingungen schafft.

² <http://www.f-response.com>

2.6.1 Funktionen zur Datenreduktion

In diesem Kapitel wird die Funktionsweise der Mechanismen erklärt, die bei eDiscovery Lösungen verwendet werden um die Datenmenge zu reduzieren. Die wichtigste Methode, um Informationen zu finden, ist die Stichwortsuche.

Die Stichwortsuche ist eine Methode um fallrelevante Informationen in der Masse von Dokumenten zu finden. Dabei wird in allen Dokumenten nach dem entsprechenden Stichwort gesucht, sofern das Dateiformat von der eDiscovery Software unterstützt wird. Die Suche kann mit Wildcards und Boolean Expressions erweitert werden. Manche Software Produkte bieten die Möglichkeit für 'Keyword Field Searching', so können z.B. bei E-Mails speziell die Felder From, To, CC, oder BCC durchsucht werden. [30]

Jedoch kratzt die Stichwortsuche lediglich an der Oberfläche. Es kann nicht garantiert werden, dass man hiermit alle relevanten Informationen findet. Meist sind die Ergebnisse nicht vollständig, da sich die Wörter vom eigentlichen Stichwort unterscheiden. Gründe hierfür sind z.B. eine andere Schreibweise, Abkürzungen, Umgangssprache und Tippfehler. Es kommt auch vor, dass die Autokorrektur ein anderes Wort einsetzt. Einige Suchmechanismen bieten eine sogenannte 'Fuzzy Search' Funktion, wodurch übliche Rechtschreibfehler berücksichtigt werden. Einen ähnlichen Ansatz bietet das sogenannte 'Concept Searching'. Dabei werden anhand des Thesaurus Synonyme des Stichworts bei der Suche berücksichtigt. Bei 'Concept Searching' werden nicht nur Synonyme verwendet sondern auch Zusammenhänge innerhalb einer Kategorie. Wenn man als Keyword z.B. den Begriff 'Kraftfahrzeug' verwendet, werden bei der Suche auch automatisch weitere Wörter in der Kategorie miteinbezogen, das kann z.B. der Begriff 'Fahrzeug', Auto, PKW und unterschiedliche Automarken sein. [3]

Die Kategorisierung von Suchergebnissen stellt ein Problem dar, da man entscheiden muss, welches Resultat am ehesten der Suchanfrage entspricht. Im Vergleich zur klassischen Internetsuche mit Google zum Beispiel wird die Relevanz einer Website anhand des PageRanking-Ansatzes bemessen. Das heißt, wie oft wurde auf die entsprechende Website verwiesen. Dieser Ansatz ist jedoch in der elektronischen Beweissicherung nicht anwendbar. [5]

Das sogenannte Predictive coding, das im Folgenden erläutert wird, könnte hierfür einen Lösungsansatz bieten. Das manuelle Sichten der Dokumente ist der teuerste Schritt im eDiscovery Prozess. Gartner schätzt die Review Kosten für ein Gigabyte Daten auf etwa \$ 18 750. Daher arbeitet man an immer besseren Methoden die Daten im Vorhinein zu reduzieren. Eine der vielversprechendsten Entwicklungen auf diesem Gebiet ist das sogenannte 'Predictive Coding'. Diese Suchfunktion klassifiziert Dokumente nach der Übereinstimmung bestimmter Kriterien. Akademische Forschungen haben gezeigt, dass 'Predictive Coding' das Potential hat, exaktere Resultate zu produzieren und signifikant Zeit und Kosten reduzieren kann. In den letzten Jahrzehnten gab es bedeutende Fortschritte im Bereich 'Machine learning'. Das Ziel der Forschungen ist es, dass die Software selbst vom Input der Umgebung lernt. Diese Technologie wird auf verschiedene Weisen eingesetzt. Das reicht von Anti-Spam Filtern bis zur personalisierten Werbung auf Shoppingsites. Der US Postal Service setzt diese Technik ein um täglich 703 Millionen Briefe und Pakete zu sortieren. Dabei wird eine selbstlernende Schrifterkennungssoftware eingesetzt. Diese Technik ist auch für eDiscovery Applikationen nutzbar und kann den Review Prozess kosteneffektiver machen und die Genauigkeit optimieren. Das ganze funktioniert indem die Software mit dem menschlichen Gutachter interagiert. Der Gutachter kennzeichnet Dokumente in einem Musterset und die Software lernt dabei, die Kriterien der Dokumente zu bemessen und kann genauere Vorschläge erzeugen, welche Kennzeichnung für die restlichen Dokumente zutrifft. Schlussendlich sind die Dokumente so genauer vorkategorisiert und der manuelle Aufwand wird dadurch reduziert. 'Predictive coding' ist jedoch kein Ersatz für einen menschlichen Gutachter. Inzwischen wurde bei gerichtlichen Entscheidungen schon auf 'Predictive coding' hingewiesen (referenced). [49]

Eine andere Technologie die für den eDiscovery Einsatz sehr interessant ist wird bereits bei der Google Suche eingesetzt und nennt sich 'Search by Image'. Diese Suchmethode ist vor allem dann interessant, wenn man Bildmaterial hat und dieses nicht zuordnen kann. Hierfür wird das vorhandene nicht zuordenbare Bild auf den Google Server geladen. Google gleicht dann die Daten ab und schlägt einen 'best guess for this image' mit den dazu passenden Webinhalten oder Bilder mit ähnlich markanten Punkte wie das Referenzbild vor. Wenn man diese Suchmethode in ein eDiscovery System implementieren würde, könnte man eine eingescannte Unterschrift,

Zeichnung oder Bauplan als Suchreferenz in das eDiscovery System laden und bekäme dann alle Dokumente, die eine ähnliche Unterschrift aufweisen oder zum Beispiel E-Mails, Word Dokumente, Bilddateien oder PDFs in denen eine entsprechende Zeichnung oder Bauplan gespeichert ist. [50]

In den meisten Fällen filtert man die Dokumente zusätzlich zur Stichwortsuche noch anhand der Metadaten um die Datenmenge zusätzlich nochmals zu reduzieren. Metadaten bezeichnet man auch als 'data about data', auf Deutsch: Metadaten enthalten Informationsdaten zu einer Datei. [2]

Metadaten können aus den unterschiedlichsten Quellen erzeugt werden: Metadaten können automatisch vom Betriebssystem erstellt werden, von einem User selbst erstellt werden oder durch die Beziehung eines anderen Dokuments gefolgert werden. Einige der Metadaten wie z.B. das Erstelldatum kann einfach von Usern gelesen werden. Andere Metadaten können versteckt oder eingebettet sein und sind daher für unwissende User nicht auffindbar. [51]

Metadaten stellen nützliche Details bereit wie Änderungsdatum, Name des Autors oder das Empfangsdatum einer E-Mail. In der folgenden Tabelle 1: Metadatenfelder [2] findet man die am häufigsten vorkommenden Metadatenfelder. [2]

Feld Name	Beschreibung
Datei Name und Pfad	Der Gesamte Pfad und der Dateiname der Datei
Datei Name	Der Dateiname der Datei
Datei Pfad	Der gesamte Pfad der Datei
MD5 oder SHA1	Der eindeutige Kennung der Datei
Dokument Typ	Dokument Typ definiert das Dateiformat also die Syntax und Semantik innerhalb einer Datei
E-Mail	Stellt Details über die E-Mail bereit
Parent	Dies definiert, ob die Datei einen Anhang hat
Child	Dieses Feld kennzeichnet, dass diese Datei ein Anhang ist
Duplikat	Dieses Feld legt fest, ob die Datei ein Duplikat unter dem gesamten Datenbestand ist oder eine Teilmenge des Datenbestands.
Zugriffsdatum	Das Datum des letzten Zugriffs
Änderungsdatum	Das Datum der letzten Änderung
Erstelldatum	Das Datum wann die Datei erstellt wurde
Text	Der extrahierte Text aus der Datei
Autor	Der Ersteller und Autor der Datei
Zeichen Anzahl	Die Anzahl der Zeichen in der Datei

Tabelle 1: Metadatenfelder [2]

In Kombination mit der Stichwortsuche und den Metadaten ist nun eine viel gezieltere Suche möglich. Ein mögliches Suchszenario könnte wie folgt aussehen: Suche alle PDF Dateien, die von dem Autor 'Max Mustermann' im Zeitraum zwischen '01.03.2006 – 30.05.2006' erstellt wurden und das Stichwort 'Lambdasonde' enthalten. Da man nun den Dokumententyp, den Autor und den Erstellungszeitraum eingegrenzt hat, liefert die Suche nach dem Stichwort 'Lambdasonde' eine sehr viel geringe Anzahl an Dokumenten zurück.

Eine andere wirksame Methode um die Datenmenge zu reduzieren, ist die sogenannte Deduplizierung. Hierbei werden doppelte, also gleiche Dateien identifiziert. Grundsätzlich gibt es zwei wesentliche Deduplizierungsmodelle. Man unterscheidet zwischen der 'per-custodian deduplication' und dem 'global deduplication' Modell. Im 'per-custodian' Modell werden alle Dubletten in den Daten von User X entfernt. Wenn sich eine Kopie der Datei auf der Festplatte von User X befindet und eine andere Kopie auf dem Fileserver im Homeverzeichnis und mehrere Kopien der Datei von unterschiedlichen Backup Tapes wieder hergestellt werden, dann sollte der Anwender oder Gutachter in der Datenmenge von User X die Datei trotzdem nur einmal vorliegen haben. Andere Mitarbeiter haben evtl. auch eine Kopie der Datei von User X gespeichert. Diese wird jedoch nicht dedupliziert. Das 'global deduplication' Modell identifiziert mit einem speziellen Deduplikationsalgorithmus alle

doppelten Kopien einer Datei, egal von welchem Mitarbeiter. Hat sowohl User X als auch User Y eine Kopie der Datei lokal auf der Festplatte gespeichert, wird die erste Kopie, die verarbeitet wird, im Review Environment verwendet. Die zweite Kopie wird dann unterdrückt und nicht angezeigt. Jedoch sollte die Software irgendwo vermerken, wie viele Kopien einer Datei gefunden wurden und wo. Der meist verwandteste Algorithmus zur Deduplizierung von elektronischen Dokumenten ist der MD5 Hash. Bei E-Mails werden meist die Werte von verschiedenen E-Mail-Feldern wie Sender, Empfänger, CC, BCC, Betreff, Body Text und Attachment Titel in einer fest vorgegebenen Reihenfolge gehasht und danach mit anderen E-Mails verglichen. [30]

2.6.2 Zertifizierungen

Um das eDiscovery Personal für die Besonderheit von personenbezogenen Daten zu sensibilisieren ist beispielsweise eine eForensik Zertifizierung eine sinnvolle Basis. Die eForensik Zertifizierung hat einen technischen Teil, indem die Nutzung verschiedener Tools für eine forensische Analyse eingeübt wird. Dabei geht es um richtiges Vorgehen bei einem Verdacht auf einen Systemeinbruch, Datenmissbrauch, Datendiebstahl, Datenlöschung und bei unberechtigter Nutzung von Kommunikationsmöglichkeiten. Im rechtlichen/organisatorischen Teil der Zertifizierung werden die rechtlichen Rahmenbedingungen für Ermittlungen gegen Verdächtige behandelt. Dabei wird genau erläutert, wie man adäquate Reaktionen im Fall eines Systemeinbruchs und die Anforderungen an eine gerichts feste Sammlung, Speicherung und Auswertung digitaler Spuren als Beweismittel ausübt. Eine Zertifizierung für den kompletten eDiscovery Prozess existiert bisher noch nicht. Es werden lediglich Teilprozesse und die entsprechende Software zertifiziert. [16]

Ganz egal welche Zertifizierung letztendlich durchgeführt wird, die Mitarbeiter im eDiscovery Team sollten in der Zertifizierung mindestens folgende Kenntnisse vermittelt bekommen.

- Umgang mit digitalen Beweismitteln und 'chain of custody' (Beweismittelkette)
- Basiswissen über 'digital Forensics and Investigations'
- Ablauf und Kenntnisse in eDiscovery
- Länderspezifische Datenschutzrichtlinien

Vor dem Einsatz einer Forensik Software bzw. eDiscovery Software muss geprüft werden, dass das Software Produkt die Integrität des Beweismaterials nicht beeinflusst.

Organisationen, die eine allgemeine digitale Forensik Zertifizierung durchführen, in der Grundlagen und Techniken Software unabhängig vermittelt werden, sind im folgenden aufgelistet:

- SANS GIAC Certified Forensic Analyst (GCFA) ³
- IACIS Certified Forensic Computer Examiner (CFCE) ⁴
- EC-Council Certified Hacker Forensic Investigator (CFCI) ⁵

Softwarespezifische Zertifizierungen erhält man beispielsweise von folgenden Organisationen:

- AccessData AccessData Certified Examiner (ACE) ⁶
- Guidance Software Encase Certified Examiner (EnCE) ⁷ [2]

Aktuell gibt es keine Organisation, die eine eDiscovery Software zertifiziert und somit garantieren würde, dass die Software die Integrität der Daten nicht beeinflusst, die Software fehlerfrei und zuverlässig arbeitet. Wird eine eDiscovery Software für einen Fall verwendet, kann diese angefochten werden. Somit entsteht ein Präzedenzfall. Prinzipiell kann man sagen, je mehr Präzedenzfälle eine Software gewonnen hat, desto besser. Letztendlich gilt die Software dann vor den Gerichten als akzeptiert. [52]

³ <http://www.sans.org>

⁴ <https://www.iacis.com>

⁵ <http://www.eccouncil.org>

⁶ <http://www.accessdata.com>

⁷ <http://www.guidancesoftware.com>

2.6.3 Forensik

Setzt man sich mit dem Thema eDiscovery auseinander, so gehört auch das Themengebiet der Forensik dazu, denn es kommt des Öfteren vor, dass Daten, die für eine 'Discovery Order' relevant wären, nicht zugänglich sind. Dies kann die unterschiedlichsten Ursachen haben. Erhebliche Schwierigkeiten bereiten dabei vor allem die obligatorische Verschlüsselung von Datenträgern, passwortgeschützten Dokumenten und gelöschten Dokumenten. Sind Datenträger oder Dokumente verschlüsselt, ist eine automatisierte Auswertung nicht möglich. Das Aufbrechen der Verschlüsselung ist abhängig vom verwendeten Verfahren und nur sehr selten und in speziellen Fällen möglich. [5]

Daher sollte man als IT-Verantwortlicher darauf achten, dass die Festplattenverschlüsselung bei vergessenem Passwort durch einen entsprechenden Administrator entsperrt oder das Passwort zurückgesetzt werden kann. Wie es z.B. mit dem Microsoft Bitlocker in einer Active Directory Domain möglich ist. [53]

Der häufigste Einsatz von Forensik Tools sind defekte oder gelöschte Dateien. Hierbei versucht man mit einer speziellen Software oder auch manuell, möglichst viele Informationen der ursprünglichen Datei wiederherzustellen. [5] Die am häufigsten eingesetzten Tools, die Forensiker hierfür nutzen, sind beispielsweise EnCase von Guidance Software, Forensic Tool Kit von Access Data und Helix3. Letzteres ist ein umfangreiches Open Source Tool Kit. [54] Mit diesen mächtigen Tools kann man gelöschte Daten oft wiederherstellen oder mit EnCase Forensics sicherstellen, dass Dateien entdeckt werden deren Datentyp nachträglich verändert wurde um diese zu verstecken. EnCase Forensics findet Dateien anhand der Dateisignatur. Dabei handelt es sich um die sogenannten 'Magic Numbers'. Dies ist eine Byte-Sequenz, die für ein jeweiligen Dateitypen charakteristisch ist z.B. D0CF11E81A1BB11AE für Microsoft Word. [5]

2.6.4 Trends

„Der Trend der Hersteller zu webbasierten Lösungen stößt auch angesichts der aktuellen Datenskandale bei deutschen Unternehmen weiterhin auf Ablehnung. Die Diskussionen aus der Praxis zeigen, dass viele Firmen auf keinen Fall wünschen, dass sensible Unternehmensdaten und personenbezogene Daten ihr Haus verlassen. Dies trifft auch auf das Hosting von Daten zu, welches in Deutschland von vielen Unternehmen völlig abgelehnt wird. Daher geht der Trend zu vor Ort genutzten isolierten Analyseumgebungen. In diesem Fall wird eine vollständige Analyseumgebung vom Serversystem bis hin zu den Review-Arbeitsplätzen an einem zu bestimmenden Standort aufgebaut. Die Analyseumgebung wird soweit möglich physisch vollständig vom Firmennetz und öffentlichen Netzwerken getrennt. Ist die hierzu benötigte Hardware vorhanden, das Know-How verfügbar und die zu analysierenden Daten, kann eine solche isolierte Umgebung bereits in wenigen Tagen funktionsfähig aufgebaut werden.“ [38] Eines der größten Probleme im Bereich eDiscovery ist die enorm steigende Datenmenge. „Laut einer Studie von EMC ist damit zu rechnen, dass die jährlich erzeugte Datenmenge bis 2020 um den Faktor 44 steigen wird. Mehr als 70 % davon entfallen auf einzelne Anwender. Trotzdem tragen Unternehmen die Verantwortung für etwa 80 % der Datenmengen, da sie dafür Sorge tragen müssen, dass Informationen gespeichert, verwaltet und geschützt werden. Um diese Menge an unstrukturierten Daten im Sinne der rechtskonformen Verwaltung administrieren zu können, muss man sich künftig zwangsläufig mit Cloud Computing und anderen effizienten Verfahren für das Informations-Management beschäftigen. [5]

3 Cloud Computing

Cloud Computing ist eines der Modeworte im heutigen IT Zeitalter. Inzwischen setzen viele Hersteller auf das Cloud Computing und bieten Cloud-Plattformen an. Auch traditionelle IT-Unternehmen wie Microsoft, die ihre primären Umsätze durch Software generieren, bieten zunehmend Cloud Dienste und Plattformen an. [55] Dies ist jedoch kein Wunder, wenn man sich die Wachstumsprognosen von Cloud Computing genauer anschaut. Die Marktforscher von Gartner beispielsweise bewerten den weltweiten Markt für das Jahr 2008 mit über 46 Milliarden US-Dollar mit einer Steigerung bis 2013 auf über 150 Milliarden US-Dollar. [56]

Daher verwundert es nicht, dass alle Hersteller etwas vom Kuchen abhaben wollen und auf den Cloud Markt drängen. Bevor wir uns mit den technologischen Grundlagen von Cloud Computing beschäftigen, muss Cloud Computing korrekt definiert werden. Der Begriff Cloud Computing ist nicht einfach zu definieren. Bisher konnte sich keine Definition als allgemeingültig durchsetzen. In den Publikationen werden Definitionen verwendet, die sich zwar meist ähneln, aber doch immer wieder variieren. Eine Definition, welche in Fachkreisen gerne herangezogen wird, ist die Definition der US-amerikanischen Standardisierungsstelle NIST (National Institute of Standards and Technology). Diese Definition wird auch von der ENISA (European Network and Information Security Agency) genutzt:

„Cloud Computing ist ein Modell, dass es erlaubt bei Bedarf, jederzeit und überall bequem über ein Netz auf einen geteilten Pool von konfigurierbaren Rechnerressourcen (z.B. Netze, Server, Speichersysteme, Anwendungen und Dienste) zuzugreifen, die schnell und mit minimalem Managementaufwand oder geringer Serviceprovider-Interaktion zur Verfügung gestellt werden können.“

„Folgende fünf Eigenschaften charakterisieren gemäß der NIST-Definition einen Cloud Service:

1. On-demand Self Service: Die Provisionierung der Ressourcen (z. B. Rechenleistung, Storage) läuft automatisch ohne Interaktion mit dem Service Provider ab.
2. Broad Network Access: Die Services sind mit Standard-Mechanismen über das Netz verfügbar und nicht an einen bestimmten Client gebunden.
3. Ressource Pooling: Die Ressourcen des Anbieters liegen in einem Pool, aus dem sich viele Anwender bedienen können (Multi-Tenant Modell). Dabei wissen die Anwender nicht, wo sich die Ressourcen befinden. Sie können aber vertraglich den Speicherort, also z.B. Region, Land oder Rechenzentrum, festlegen.
4. Rapid Elasticity: Die Services können schnell und elastisch zur Verfügung gestellt werden, in manchen Fällen auch automatisch. Aus Anwendersicht scheinen die Ressourcen daher unendlich zu sein.
5. Measured Services: Die Ressourcennutzung kann gemessen und überwacht werden und entsprechend bemessen auch den Cloud-Anwendern zur Verfügung gestellt werden.

Diese Definition gibt die Vision von Cloud Computing wieder, wobei davon abgesehen werden sollte, die einzelnen Punkte zu dogmatisch zu sehen. So wird z. B. eine ubiquitäre Verfügbarkeit bei Private Clouds eventuell gar nicht angestrebt.“ [57]

Cloud Computing konnte erst durch den Fortschritt vieler einzelner Technologien entstehen siehe Abbildung 7:

- Ohne die Virtualisierung [58] wäre ein physikalischer Server nur einmal für einen bestimmten Cloud Service nutzbar. Durch den Einsatz von Virtualisierungssoftware können mehrere virtuelle Server auf einem physikalischen Server betrieben werden. Damit kann nun ein physikalischer Server unterschiedliche Cloud Services an mehrere Kunden preisgünstig bereitstellen. [56] Der Durchbruch der Virtualisierung gelang jedoch erst mit der zunehmenden Vorbereitung der Multicore CPUs. [59]
- Cloud Computing hat wesentliche Technologie Impulse aus dem Grid Computing [60] übernommen. Die Grid Computing Technologie liefert den fundamentalen Bauplan für standardisierte, hochskalierbare Systemarchitekturen, die abstrahiert als virtuelle IT-Services zur Verfügung gestellt werden können. Die

gemeinsame, effektive Nutzung von Ressourcen und der Zugriff über das Internet haben sich im Grid Computing etabliert und fanden ebenso Einzug in das Cloud Computing. [56]

- Bei Web 2.0 [61] [62] geht es hauptsächlich um die kommunikativen Aspekte des Webs mit starker Betonung auf spontaner Interaktion. Die Web 2.0-Technologien, die auf globale Zusammenarbeit von Usern bei der Informationsverarbeitung setzen, hält weiter Einzug in die Unternehmen. Dies gilt auch für service-orientierte Architekturen (SOA) [63]. Bei SOA werden vorhandene Komponenten wie Server, Datenbanken und Websites in Services gekapselt und dann so koordiniert, dass ihre Leistungen zu höheren Services zusammengefasst und anderen Abteilungen oder Kunden zur Verfügung gestellt werden können. Diese Bündelung von IT-Anforderungen zu Services im Sinne von Service orientierten Architekturen (SOA) ist kostensparend und verleiht eine gewisse Agilität. Damit ist es möglich einzelne Services voneinander zu koppeln und in Abhängigkeit von den Erfordernissen der Anwender flexibel neu zu kombinieren. Durch diese lose Koppelung von Services ist es möglich den Ort der Ausführung zu ändern, solange die Regeln für einen Service-Aufruf beibehalten werden. Damit können Software Services ausgelagert werden wie in diesem Fall in die Cloud. [56]
- Die steigende Komplexität von Computer Systemen motivierte die Forschung am Thema Autonomic Computing [64] [65] zu arbeiten. Ziel dieser Forschungen war es die Systeme zu verbessern, indem der menschliche Einfluss minimiert wird. Anderst ausgedrückt, die Systeme sollen sich selbst managen. IBMs Automatic Computing Initiative steuerte zur Definition der vier Eigenschaften von Autonomic Computing bei: Self-configuration, self-optimization, self-healing, und self-protection. Die riesigen Datacenter von Cloud Computing Providern wie Amazon Web Services müssen effizient gemanagt werden. Die Konzepte des Autonomic Computing inspirierten die Virtualisierungs- und Cloud Unternehmen. Die Unternehmen entwarfen Konzepte wie proactive disaster recovery oder automatisiertes VM provisioning. [59]

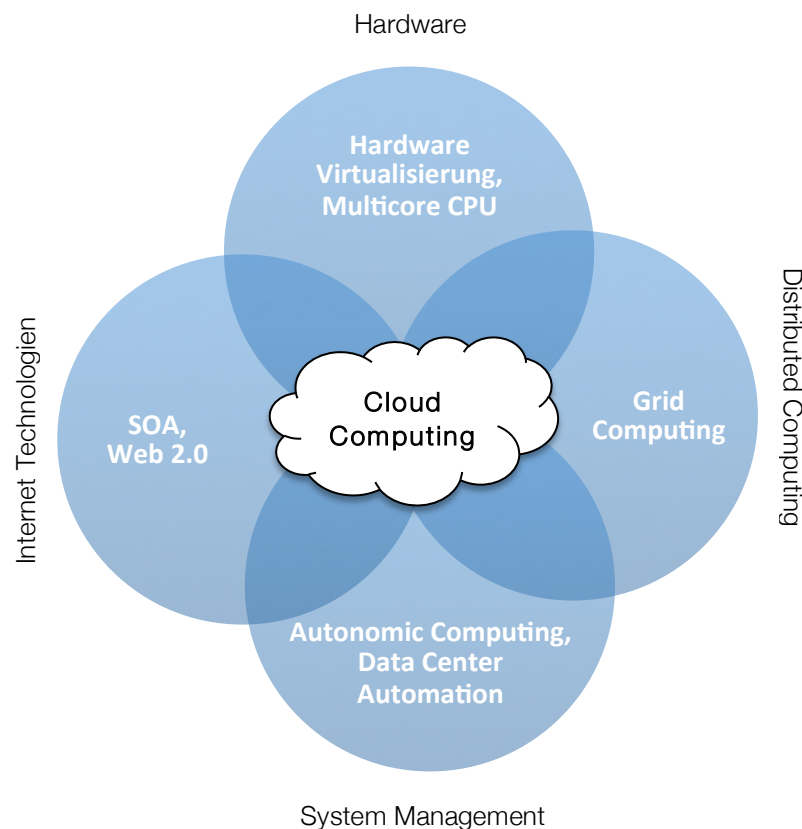


Abbildung 7: Die verschiedenen Entwicklungen, die Cloud Computing ermöglichen [59]

Bei Cloud Computing unterscheidet man drei grundsätzlich verschiedene Servicemodelle:

- **Infrastructure as a Service (IaaS)**
Bei IaaS werden IT-Ressourcen wie Rechenleistung, Arbeitsspeicher oder Netze als Service angeboten. Ein Kunde kauft diese virtualisierten und in hohem Maß standardisierten Services und baut darauf eigene Services zum internen oder externen Gebrauch auf. Der Kunde kann dann z.B. Rechenleistung, Arbeitsspeicher und Datenspeicher anmieten und darauf ein Betriebssystem mit Anwendungen seiner Wahl ausführen. [57] Anbieter für IaaS sind z.B. Amazon EC2, AppNexus, HP Cloud Enabling Computing, Microsoft Windows Azure platform und Sun Cloud. [56]
- **Platform as a Service (PaaS)**
Ein PaaS Anbieter stellt eine komplette Infrastruktur bereit und bietet dem Kunden auf der Plattform standardisierte Schnittstellen an, die von Services des Kunden genutzt werden. Die Plattform kann z.B. Skalierbarkeit, Zugriffskontrolle, Datenbankzugriffe, etc. als Service zur Verfügung stellen. Der Kunde hat keinen Zugriff auf die darunterliegenden Schichten wie Betriebssystem, Hardware usw. Der Kunde kann aber auf der Plattform eigene Anwendungen laufen lassen, für deren Entwicklung der Cloud Provider in der Regel eigene Werkzeuge anbietet. [57] Beispiele für PaaS sind Force.com, Google App Engine und einzelne Microsoft Azure Services wie SQL Azure oder .Net Services. [56]
- **Software as a Service (SaaS)**
Alle Anwendungen, die den Kriterien des Cloud Computing entsprechen, fallen in die SaaS Kategorie. [57] Als Beispiele hierfür können Google Apps for Business, Microsoft Online Services, CRM Online, Salesforce.com und WebEx genannt werden. [66]

Der Begriff "as a Service" wird für eine Vielzahl weiterer Angebote benutzt, wie z.B. für Security as a Service oder Storage as a Service, so dass häufig auch von "XaaS" geredet wird.

Die Servicemodelle unterscheiden sich auch in sofern, wie viel Einfluss der Kunde auf den angebotenen Service hat. Die Abbildung 8 zeigt, welche Schichten selbst verwaltet werden (blau dargestellt) und welche Schichten vom Cloud Provider verwaltet werden (grau dargestellt). Bei der traditionellen IT hat der Anwender die volle Kontrolle über alle Schichten hinweg. Bei IaaS kann der Kunde alles ab dem Betriebssystem aufwärts verwalten. Alle darunterliegenden Schichten wie Virtualisierung oder Networking verwaltet der Cloud Provider. Bei PaaS hat der Kunde nur noch Kontrolle über seine Anwendungen inklusive Daten, die auf der Plattform laufen. Verwendet man SaaS übergibt man praktisch die ganze Kontrolle an den Cloud Provider.

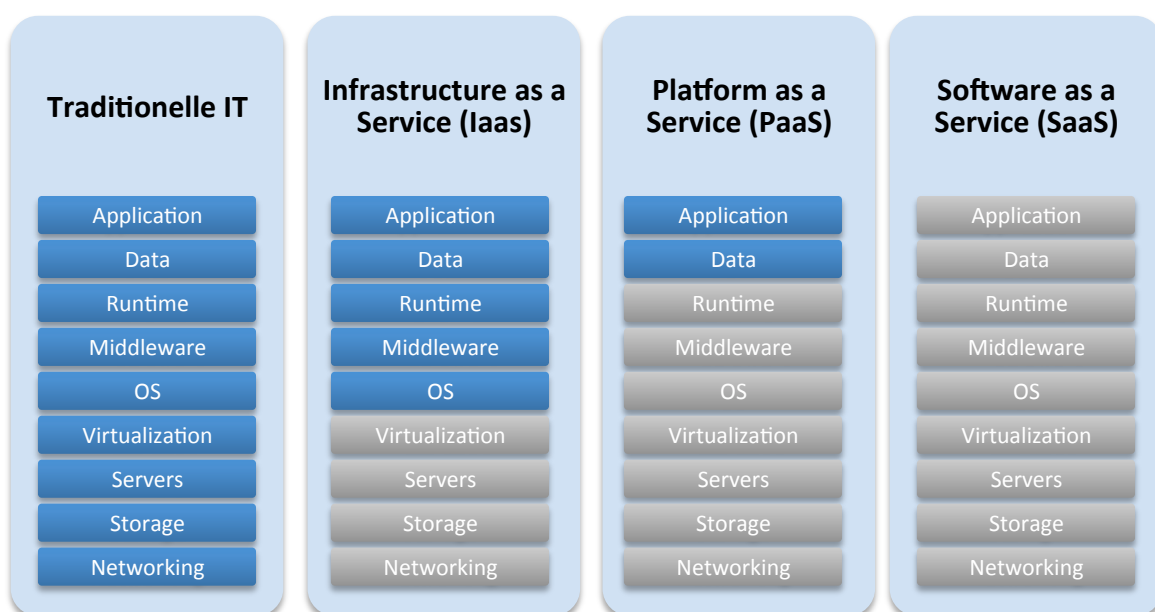


Abbildung 8: Cloud Service Architektur [67]

Unter den Gesichtspunkten, wie eine Cloud betrieben bzw. organisiert ist, unterscheidet man zwischen zwei reinen Cloud Formen: Der Private Cloud (auch Enterprise Cloud genannt) und der Public Cloud, wobei sich beide Formen in der technischen Realisierung grundsätzlich nicht unterscheiden. Virtualisierung, Service-Orientierung und die Nutzung der Services über das Internet sind wesentliche Eigenschaften der technischen Implementierung.

- Die Private Cloud ist eine vom Unternehmen selbst betriebene Cloud-Umgebung. Der Zugang zu dieser Cloud ist beschränkt und nur für das Unternehmen selbst, autorisierte Geschäftspartner, Kunden und Lieferanten zugänglich. Der Zugriff erfolgt in der Regel über eine Virtual Private Network (VPN) Verbindung. Bei der Private Cloud handelt es sich um eine effiziente, standardisierte und sichere Umgebung, die unter der vollen Kontrolle des Unternehmens steht.
- Die Public Cloud wird von einem IT-Dienstleister betrieben. Der Zugriff auf diese Cloud erfolgt in der Regel über das Internet. Viele Kunden oder auch Unternehmen teilen sich eine virtualisierte Infrastruktur. Die Nutzung erfolgt flexibel und schnell. Eine Public Cloud stellt eine Auswahl von hochstandardisierten Services für Geschäftsprozesse, Anwendungen und/oder Infrastruktur bereit.

Eine weitere Cloud Computing Form ist eine Mischform aus der Private und der Public Cloud. Die sogenannte Hybrid Cloud. Diese Cloud Form wird voraussichtlich in absehbarer Zeit breite Akzeptanz finden. Die Herausforderung bei der Hybrid Cloud besteht darin, die traditionelle IT-Umgebung, Private Cloud und/oder Public Cloud auf den Schichten Applikation, Middleware und der Infrastruktur in Bezug auf Services und Sicherheit so zu integrieren, dass eine heterogene Umgebung für den User als homogen wahrgenommen wird. Der User erwartet eine integrierte, den Geschäftsprozess optimal unterstützende Anwendung, unabhängig davon, ob diese Applikation oder Teile davon in einer Private Cloud, Public Cloud oder auf der traditionellen IT-Umgebung zur Verfügung gestellt wird. Letztendlich werden die Punkte Sicherheit und Service über die Akzeptanz und somit auch über den Erfolg des Cloud Computings entscheiden. [56]

Historisch wurde der Begriff „Cloud-Computing“ maßgeblich durch die rasant wachsenden Internetfirmen wie Google, Amazon und Yahoo geprägt. Diese Firmen standen vor dem Problem, die schnell wachsende Anzahl an Usern immer mit genug Ressourcen zu versorgen. Damit die Performance der bereitgestellten Services auch zu Spitzenlastzeiten garantiert ist. [68]

Amazon ist heute der größte Anbieter im Bereich Cloud Computing IaaS. Amazon ist vielen vor allem durch seinen Onlineshop bekannt. Weitaus weniger bekannt ist, dass Amazon sich in den letzten Jahren zu einem bedeutenden Dienstleister für IT entwickelt hat. Der Ursprung dieser Entwicklung liegt im stetigen Wachstum des Onlineshops von Amazon in Bezug auf Sortiment und Kundenstamm. Dieses ständige Wachstum hatte zur Folge, dass Amazon eine immer leistungsfähigere IT-Infrastruktur benötigte. Besonders vor großen Feiertagen, wie in der Vorweihnachtszeit, musste die IT-Infrastruktur enorme Zugriffsspitzen ausgleichen. Die Folge daraus war, dass Amazon das ganze Jahr über enorme zusätzliche Kapazitäten vorhielt um die kurze Zeit der Zugriffsspitzen zu überbrücken. Da die Vorweihnachtszeit zu den lukrativsten Zeiten für Amazon gehört, wurden die enormen IT-Infrastrukturkosten in Kauf genommen. Die hohen Kosten senkten dementsprechend langfristig die Profitabilität des Unternehmens. Amazon nutzte das Problem und entwickelte daraus ein neues Geschäftsmodell. Die ungenutzten Ressourcen der eigenen IT-Infrastruktur wurden einfach weitervermietet. Das Ziel des ab 2002 verfügbaren 'Amazon Web Services' war zunächst die Vermarktung der eigenen Überkapazitäten. Unter dem Namen 'Amazon Web Services' wird heute die gesamte Produktpalette der Amazon Cloud Computing Lösungen angeboten.

Der große Vorteil an Cloud Computing ist, dass man je nach Bedarf quasi sofort Zugriff auf zusätzliche Ressourcen hat. Oft sind die angemieteten Ressourcen beim Cloud Provider sogar billiger wie die selbst gehosteten Ressourcen. In der Publikation 'Above the Clouds: A Berkeley View of Cloud Computing' wird gezeigt, wie stark Netzwerk, Storage und Administrationskosten zwischen mittelgroßen ($\approx 1\,000$ Server) und sehr großen Data Center ($\approx 50\,000$ Server) variieren, siehe Tabelle 2. [69]

Technologie	Kosten in mittelgroßem Data Center	Kosten in sehr großem Data Center
Netzwerk	\$ 95 pro Mbit / sec / Monat	\$ 13 per Mbit / sec / Monat
Storage	\$ 2.20 pro GByte / Monat	\$ 0.40 pro GByte / Monat
Administration	≈ 140 Server / Administrator	> 1000 Server / Administrator

Tabelle 2: Vergleich Data Center Größe und Preisspanne der Ressourcen [69]

Durch die enorm großen Rechenzentren der Cloud Provider wie Amazon, ist eine ganz andere Preispolitik möglich als bei herkömmlichen Hosting Providern oder wenn man die Ressourcen selbst hostet. Amazon betreibt mehrere weltweit verteilte Data Center für ihre AWS Cloud Services. Verwunderlich auf den ersten Blick ist jedoch, dass sich die Preise pro Ressource zwischen den verschiedenen Standorten, selbst innerhalb Nord Amerikas, sehr stark unterscheiden. [70] Auch dieser interessante Aspekt wird in der Berkeley Studie beleuchtet. Laut der Studie ist die Preisdifferenz hauptsächlich auf die unterschiedlichen Strompreise in den Bundesstaaten zurückzuführen. Auch haben einzelne Gesetze der Bundesstaaten Einfluss auf die Ressourcen Preis der Cloud Provider. Wie die Tabelle 3 zeigt, ist der Strompreis in Kalifornien mehr als doppelt so hoch wie in Idaho. Die mögliche Ursache für diesen Preisunterschied könnte damit erklärt werden, dass es in Kalifornien gesetzlich verboten ist Kohlekraftwerke zu betreiben. Dieses Beispiel zeigt deutlich von welchen externen Einflüssen der Preis einzelner Ressourcen abhängig sein kann.

Preis pro kWh	Wo	Mögliche Ursachen warum
3.6 ¢	Idaho	Wasserkraft, keine weiten Distanzen
10.0 ¢	Kalifornien	Weite Distanzen über das Stromnetz, Einschränkungen der Überlandleitungen im Küstengebiet, es sind keine Kohlekraftwerke in Kalifornien erlaubt
18.0 ¢	Hawaii	Keine fossilen Brennstoffe vorhanden alles muss vom Festland geliefert werden

Tabelle 3: Strompreis pro kWh in verschiedenen US Bundesstaaten [69]

Die großen Cloud Provider wie Amazon, Google, Salesforce, Yahoo und Microsoft benutzen innovative und neue Technologien, um den Kunden eine möglichst sichere, performante und skalierbare Umgebung zu bieten. Der Cloud Provider kümmert sich sowohl um Backups und Disaster Recovery als auch um die Netzwerkinfrastruktur wie Load Balancing usw. Durch die skalierbare Infrastruktur und enorme Menge an verfügbaren Ressourcen ist die Cloud auch ein wirksames Mittel gegen DDOS Angriffe. Die Cloud Provider verfügen über großes technisches Know-how und garantieren so eine hohe Verfügbarkeit der Services. Man sollte jedoch nicht vergessen, wenn man seine Daten einem Cloud Provider anvertraut, dass man die eigentliche Kontrolle über geschäftskritische Daten verliert. Bevor man Daten in einer Public Cloud ablegt, sollte man sich immer Fragen, wie man jederzeit an seine Daten kommt, auch in einem entsprechenden Dateiformat, das man verarbeiten kann. Wo genau und in welchem Land werden die Daten gespeichert? Dies kann besonders bei persönlichen Daten kritisch sein, die dann plötzlich außerhalb der EU gespeichert werden. Wer administriert die Services, kann man demjenigen vertrauen? Werden die Daten verschlüsselt gespeichert? Wer hat Zugriff auf die Daten? Haben z.B. Mitarbeiter des Cloud Providers Zugriff auf die Daten? Reicht eine vertragliche Garantie, wie ein SLA [71] aus? Was passiert bei einer Vertragsverletzung? [54] Ein gutes Beispiel, wie schnell man die Kontrolle über seine Daten verlieren kann, war der Skandal um die AGB des Cloud Services Google Drive. Es stellte sich zwar heraus, dass es sich hier um ein großes Missverständnis handelte, allerdings kann solch ein Fall auch schnell zur bitteren Realität werden. In den AGB von Google Drive hieß es in einer Passage, dass alle auf 'Google Drive' gespeicherten Daten automatisch zum geistigen Eigentum des Suchmaschinenkonzerns werden. Offenbar handelte es sich jedoch bei der Passage um eine juristische Standardformulierung, die Google die notwendigen Rechte einräumte, um die Service überhaupt durchführen zu können. Wenn beispielsweise mehrere Leute an einem Dokument arbeiten und Anmerkungen machen, benötigt Google die entsprechenden Rechte, um die Veränderungen durchzuführen. [72]

In diesem Fall ist nochmal alles gut ausgegangen und die ganze Aufregung war umsonst. Dies muss jedoch nicht heißen, dass es immer so glimpflich verläuft.

3.1 eDiscovery in der Cloud

Die rasante Zunahme des Datenvolumens hat die Situation von Rechtsstreitigkeiten dramatisch verändert und einen Paradigmenwechsel in der eDiscovery und Computer Forensic verursacht. Nicht nur die enorme Datenzunahme verändert die bisherige Forensic und eDiscovery Welt, sondern auch das Cloud Computing bietet ganz neue Möglichkeiten und bringt zugleich auch neue Probleme mit sich.

Wenn man von 'eDiscovery in der Cloud' spricht, kann man diesem Begriff drei ganz unterschiedliche Bedeutungen zuordnen. Daher sollen diese drei unterschiedlichen Sichtweisen zunächst genauer definiert werden, bevor man sich mit den Details im Einzelnen beschäftigt.

- SaaS-based eDiscovery: Unter 'eDiscovery in der Cloud' versteht man die Cloud als Werkzeug für verschiedene Aufgaben im eDiscovery Prozess. Hierbei wird SaaS angeboten, das typischerweise einen Schritt im eDiscovery Prozess abdeckt wie z.B. den Collection oder Review Prozessschritt.
- Hosted eDiscovery: 'eDiscovery in der Cloud' bedeutet, einen hosted Service eines Cloud Anbieters zu nutzen, um die in der Cloud gespeicherten Daten zu analysieren. Es gibt zwei unterschiedliche Arten von hosted eDiscovery. Im ersten Szenario speichert der Kunde seine Daten in der Cloud und der Cloud Anbieter führt den eDiscovery Prozess bei Bedarf durch. Im zweiten Szenario werden alle Daten im eigenen Rechenzentrum gehostet. Kommt es nun zu einer 'eDiscovery Order' lädt der Kunde alle relevanten Daten auf den Datenspeicher des Cloud Providers, der dann den eDiscovery Prozess durchführt.
- eDiscovery von in der Cloud gespeicherten Daten: Bei diesem Szenario von 'eDiscovery in der Cloud' speichert der User seine Daten in der Cloud ohne jegliche Vereinbarung mit dem Cloud Provider, wie bei einer 'eDiscovery Order' verfahren wird.

Im Jahre 2009 beschloss die Stadt Los Angeles mit einem Budget von \$ 7.25 Millionen ihr E-Mail System und weitere Systeme in die Cloud umzuziehen. Genauer gesagt sollen alle Office Anwendungen mit der Google Apps Cloud Lösung betrieben werden. Einige Behörden, darunter auch das Los Angeles Police Department (LAPD), sind wegen Sicherheitsbedenken bisher noch nicht komplett auf die Google Cloud Lösung umgestiegen. Das LAPD äußerte einige Sicherheitsbedenken in Bezug auf die Umsetzung des Vorstrafenregisters. Einige Behörden bemängelten auch die fehlende eDiscovery Implementierung. Google hat inzwischen nachgezogen und die Migration soll voraussichtlich im November 2012 abgeschlossen sein. Die Stadt Los Angeles ist kein Einzelfall. Viele andere Organisationen setzen Cloud Lösungen gerade in den Bereichen E-Mail-, Collaboration- und Storage hosting um, ohne sich über die Tragweite ihrer Entscheidung bewusst zu sein. [73]

In den Federal Rules of Civil Procedure heißt es, dass alle fallrelevanten ESI, die unter dem 'Besitz oder Kontrolle' einer Partei stehen, vorzulegen sind. Was auch für alle Daten in der Cloud zutrifft. Aber sind die Daten auch technisch wirklich unter der eigenen Kontrolle? [74] Denn bei Cloud Computing bedeutet: 'data may always be in transit, never anywhere, always somewhere'. „Aus dieser Aussage folgen Fragen wie: Wie können Unternehmen erkennen 'wer, wann und wo' in der Cloud arbeitet? Wie können Organisationen auf aufbewahrte Dokumente zugreifen? Ist sich der 'Storage Manager' bewusst, dass sich manche Informationen im eDiscovery Prozess befinden?“ [13] Sind die Daten überhaupt im richtigen Format zugänglich? Möglich ist auch, dass eine Organisation die Kontrolle über ihre Daten in der Cloud komplett verliert, z.B. wenn der Cloud Provider in Konkurs gehen sollte. [75]

Viele Organisationen vergessen diese Aspekte wenn sie Ihre Services in die Cloud auslagern. Vor der Entscheidung Daten in die Cloud zu migrieren sollte ein Konzept erstellt werden, wie die Daten aus der Cloud bei einer 'Discovery Order' zusammengetragen und analysiert werden können. Oder wie man in der Cloud Lösung z.B. ein Legal Hold durchsetzen kann? Auch entsprechende SLAs können hilfreich sein diese Probleme zu lösen. Die wenigsten Firmen haben jedoch eine entsprechende Strategie vorliegen, wie auch die Umfrage des

eDiscovery Journal in Abbildung 9 ergab. Nur 16% der 172 befragten Unternehmen haben entsprechende eDiscovery Maßnahmen in ihrem Migrationskonzept bedacht. 26% der Befragten verfügen über kein eDiscovery Konzept für ihre Cloud Daten. Was jedoch wirklich überraschte, war das 58% der Befragten nicht wussten, ob ein eDiscovery Konzept existiert. Dies bedeutet viele der Unternehmen werden bei einem drohenden Gerichtsverfahren vor dem Problem stehen große Datenmengen aus der Cloud in kürzester Zeit zusammenzutragen und zu analysieren. Vermutlich wird dies ohne eine entsprechende Vorbereitung zu einer kostspieligen Angelegenheit oder sogar Sanktionen nach sich ziehen.

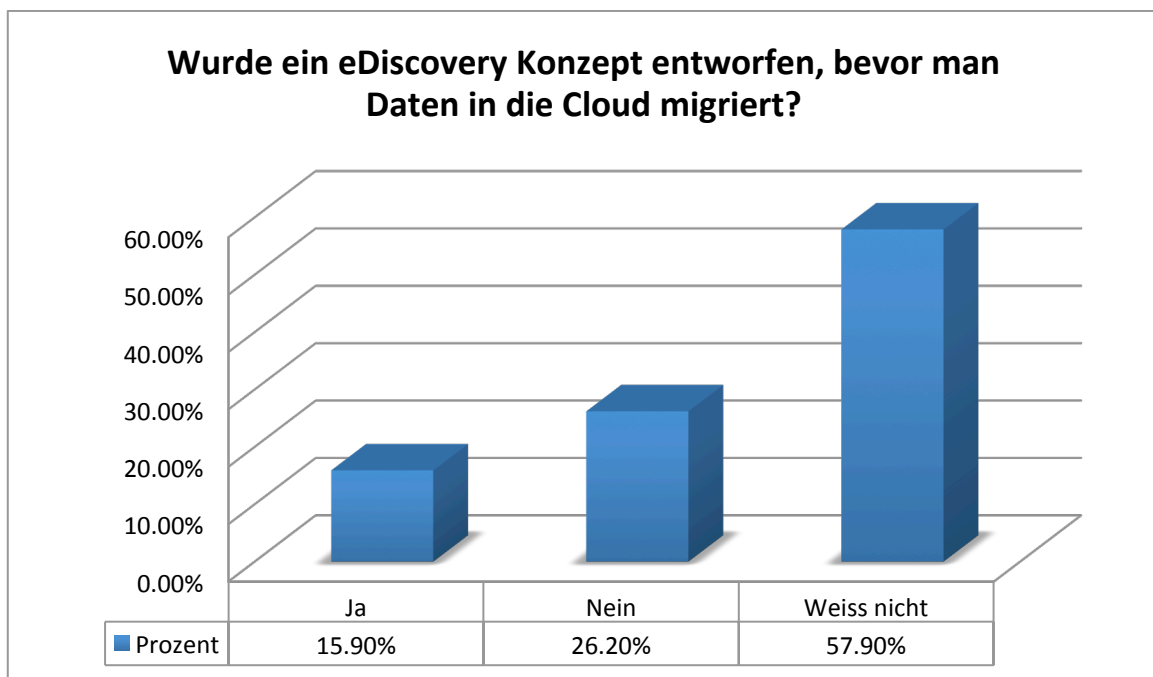


Abbildung 9: eDiscovery Plan für Cloud Migration [73]

Viele Organisationen sind der Auffassung, dass die Daten immer noch durchsucht werden können wenn es einen eDiscovery Fall in ihrer Organisation gibt. Recherchen zeigen jedoch, dass Mitarbeiter ihre Daten trotz Cloudspeicher immer noch lokal oder im internen Netzwerk ablegen. Dies bedeutet die relevanten Informationen liegen unter Umständen an verschiedensten physikalischen Standorten innerhalb der Organisation und nicht unbedingt ausschließlich in der Cloud. [73]

In der Zwischenzeit bieten einige eDiscovery Hersteller und Cloud Provider eDiscovery Lösungen für bestimmte Systeme an. Wie z.B. Google die inzwischen eine Lösung zur Archivierung und für eDiscovery in ihrem Portfolio aufgenommen haben. Google Apps Vault soll nun die Möglichkeit bieten ein 'Legal hold' durchzusetzen und Nachrichten aus Google Mail oder offiziellen Chats zu archivieren. Bei Bedarf lassen sich die Informationen gezielt durchsuchen. Die eDiscovery-Funktionen filtern Daten z.B. nach Absender, Empfänger, Datum, Betreff oder Labels. Im Gegensatz zu unternehmenseigenen eDiscovery Systemen lässt sich Google Apps Vault in Minuten aktivieren. [74]

Eine andere Möglichkeit bietet der eDiscovery Softwarehersteller ZyLAB mit seinem eDiscovery Cloud Collector. Dieser erfasst E-Mails von Microsofts Cloud Exchange Lösung Office 365 und anderen Cloud-basierten E-Mail-Lösungen. Dabei verarbeitet der Cloud Collector nicht nur die Nachrichten, sondern auch Anhänge, Kalendereinträge, Aufgaben und Kontakte, die in einem Mail Account abgespeichert sein können. [42]

Die Firma X1 Discovery bietet mit ihrem Produkt 'Rapid Discovery' eine weitere interessante Lösung an. Zunächst stellen wir uns folgendes Szenario vor: Ein Unternehmen lagert sämtliche Services in die Amazon AWS Cloud aus. Zu diesen Services gehören folgende Server: Microsoft Exchange Server, Microsoft SharePoint Server und Storage Server. Das Unternehmen erhält eine Klage und muss alle fallrelevanten Daten zur Verfügung stellen. Der klassische Ansatz hier wäre wohl wie in der Abbildung 10 links dargestellt, dass alle Daten aus der Cloud heruntergeladen und lokal analysiert werden. X1 bietet eine virtuelle Cloud Instanz an, die man bei Bedarf aktivieren kann. Diese Instanz ermöglicht nun über den Browser gesteuert die eDiscovery Prozedur in der Cloud durchzuführen wie in Abbildung 10 rechts dargestellt. Dabei werden die Daten direkt in der Cloud analysiert und gefiltert. Danach können alle relevanten Daten exportiert und heruntergeladen werden. [43], [76]

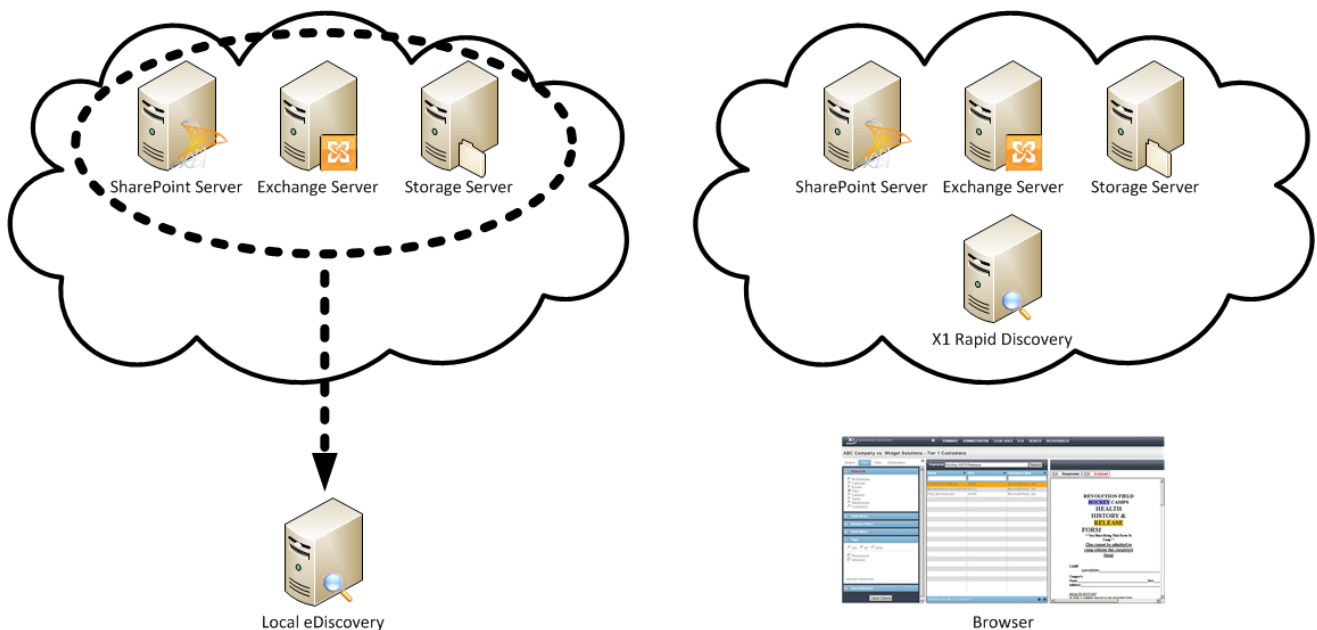


Abbildung 10: X1 Rapid Discovery [45]

Der Unterschied zwischen Public und Privat Cloud ist äußerst wichtig für alle, die sich damit beschäftigen eDiscovery in der Cloud zu betreiben. Eine Public Cloud benutzt shared Hardware, Software und Applikationen, welche für die Öffentlichkeit verfügbar sind, wie z.B. Amazon EC2, AWS und Google Apps. Dieser Ansatz ist sehr effektiv in Bezug auf die optimale Nutzung der Ressourcen. Eine Privat Cloud, egal ob diese intern hinter den eigenen Firewalls betrieben wird oder von einem Provider zur Verfügung gestellt wird, nutzt Hardware, Software und Applikationen ausschließlich für sich selbst.

Private Clouds haben einige spezielle Vorteile gegenüber Public Clouds hinsichtlich eDiscovery. Bei einer Public Cloud hat man meist keinen Einfluss bzw. keine Kenntnis wo genau oder in welchem Land die Daten gespeichert werden. Sind die Daten in der Public Cloud gespeichert, hat man keine Kontrolle wie die Daten aufbewahrt werden oder dann wieder richtig gelöscht werden. Ein weiteres Manko könnte ein zu niedriges disaster recovery oder business continuity [77] Niveau sein. Die Kunden erwarten, dass ihre Daten sicher und hochverfügbar sind. Es bieten jedoch nicht alle Cloud Provider entsprechend hohe Sicherheitsmaßnahmen an. Um zu garantieren, dass ein möglichst hohes Maß an Sicherheit gewährleistet wird, sollte der Cloud Provider folgende Sicherheitsmaßnahmen anbieten: Enterprise-class disaster recovery, welches SAS-70 Type II [78] zertifiziert ist und über ein repliziertes Data Center verfügt. Dies garantiert dem Kunden des Cloud Anbieters eine gewisse Ausfallsicherheit im Falle einer Naturkatastrophe. Natürlich sollte der Provider über ein entsprechendes Business Continuity Konzept verfügen. Jeder Service Provider sollte eine Übersicht über seine Service Level Agreements zur Verfügung stellen, in denen Recovery Point Objectives (RPO) und Recovery Time Objectives (RTO) festgelegt sind. Man sollte sich auch informieren über welche Security und Compliance Zertifizierungen der Cloud Provider

verfügt wie z.B. International Safe Harbor Certification, ISO 27001 oder ISO 27002. Jeder der mit eDiscovery und Forensic zu tun hat, sollte eine gewisse Vorsicht und Misstrauen in Bezug auf Datenschutz, Kontrolle über die Daten und Sicherheit mitbringen, egal ob sie die Daten in der Cloud oder lokal verarbeiten. Einige Cloud Provider bieten eine sichere Cloud Umgebung für eDiscovery an. Dabei hat so manch einer einen höheren IT Sicherheitsstandard als eine firmeninterne IT oder Anwaltskanzlei. Wer eine möglichst sichere eDiscovery Umgebung in der Cloud betreiben möchte, sollte auf eine Private Cloud setzen. Nur so hat man die absolute Kontrolle über alle Sicherheitsaspekte und muss sich keine Sorgen über die Sicherheit bei Shared Resources machen. [79]

3.2 MapReduce

Applikationen wie z.B. eine eDiscovery Software braucht oft mehr Ressourcen als ein Standard Server zur Verfügung stellen kann. Einige Unternehmen stehen vor dem Problem, dass ihre Geschäftsprozesse nicht mehr von einem einzelnen kostengünstigen Server bearbeitet werden können. Eine einfache Methode aber teure Lösung ist, einen speziellen Hochleistungs-Server mit sehr viel Arbeitsspeicher und etlichen CPUs anzuschaffen. Diese Lösung skaliert sehr gut, solange die Anwendung die Hardware voll ausnutzen kann. Die einzige Limitierung sind oft nur die finanziellen Mittel. Eine andere Alternative ist ein Hochverfügbarkeits-Cluster. Diese Art von Cluster benötigt eine spezielle Installation und ist meist schwer zu administrieren. Vieler dieser Hochverfügbarkeits-Cluster sind proprietär und teuer.

Eine sehr viel ökonomischere Lösung, um die benötigten Ressourcen bereitzustellen ist das MapReduce [10] Verfahren. [80] MapReduce ist eine schnelle, robuste, einfach zu nutzende, skalierbare und weit einsetzbare Lösung, um große Datenmengen im Tera- bis Petabyte-Bereich auf Rechnerclustern zu verarbeiten. Bei solchen großen Daten spricht man oft von Big Data. [81]

„Die Grundidee von MapReduce ist schnell skizziert: Man verwende die funktionalen Bausteine Map und Reduce und bilde um diese herum ein Framework mit folgenden Eigenschaften:

- automatische Parallelisierung, Koordinierung und Verteilung von Berechnungen (Jobs)
- Fehlertoleranz gegenüber Hardware- und Softwareausfällen
- automatische Lastverteilung
- Optimierung von Netzwerk und Datentransfer
- Überblick durch Status- und Monitoringnachrichten

So entsteht eine Anwendung, die für viele Einsatzgebiete geeignet ist. Dies zeigen zum einen die zahlreichen Implementierungen von MapReduce, die seit 2003 entstanden sind und zum anderen die vielen Anwendungsbereiche, in denen MapReduce zum Einsatz kommt:

- Indizierung großer Datenmengen für die Suchaufbereitung
- Maschinelles Lernen zur künstlichen Erzeugung von Wissen aus Erfahrung
- Verteiltes Suchen von Pattern, Sortieren von Daten
- Auswertung von großen Logdateien
- Kategorisierung von Daten zur Nachrichtenaufbereitung
- Datenaufbereitung für häufige Anfragen
- Erzeugung von Graphen über die Aufrufhäufigkeit, über die Verlinkung von Webseiten
- Extrahieren von Daten aus Webseiten um neue Einsichten zu erlangen

Gemeinsam haben diese Anwendungen, dass sie mit sehr großen Datenmengen arbeiten müssen, typischerweise im Tera- bis Petabyte-Bereich, und somit von einem einzelnen Rechner nicht in vertretbarer Zeit gelöst werden können. Zu verdanken ist der weite Einsatzbereich von MapReduce der Trennung von Framework- und Anwendungslogik. Die komplexen Details der Berechnung bleiben im Framework verborgen. Der Anwender braucht lediglich die Funktionen Map und Reduce für seine Zwecke zu definieren“. [81] Map und Reduce stammen aus der funktionalen Programmierung und erfüllen dort folgende Aufgaben:

- Map wendet eine benutzerdefinierte, unäre Funktion auf jedes Element der Input Sequenz an. Das Resultat ist wiederum eine Sequenz.
- Reduce verarbeitet eine Input Sequenz mit einer benutzerdefinierten, binären Funktion zu einem einzelnen Ausgabewert. [82]

Der gesamte MapReduce Prozess läuft wie folgt ab: „MapReduce realisiert eine Funktion, welche aus einer Liste von Key/Value Paaren besteht. Dieses Key/Value Paar stellt die Input Liste dar. Eine neue Liste von Key/Value Paaren berechnet die Ausgabeliste. Dabei werden zuerst die zu bearbeitenden Daten in kleine Teile, in sogenannte Blöcke aufgeteilt. Diese werden anschließend zur gleichzeitigen Verarbeitung auf unterschiedliche Knoten im Cluster verteilt. Daraus ergibt sich die parallele Verarbeitung der Daten. Nach der Verarbeitung erhält man mehrere Teilergebnisse. Die Teilergebnisse werden dann wieder zusammengeführt, damit der Nutzer ein komplettes Endergebnis der Berechnung erhält“ [83] Im folgenden Abschnitt wird das soeben genannte Funktionsprinzip von MapReduce ausführlich anhand der Abbildung 11 erklärt. Die Abbildung 11 zeigt einen Gesamtüberblick einer MapReduce Operation, wenn ein User Programm die MapReduce Funktion aufruft. Die folgenden Schritte werden von MapReduce sukzessive abgearbeitet (die nummerierten Beschriftungen in Abbildung 11 entsprechen den folgenden Nummern im Text):

1. Die MapReduce Library im User Programm teilt zuerst die Input Files in M Teile. Jedes dieser Teile ist typischerweise zwischen 16 und 64 MB groß (dies kann mit einem optionalen Parameter gesteuert werden). Dann werden mehrere Kopien des Programms auf den Cluster Maschinen gestartet.
2. Eine der gerade gestarteten Programmkopien wird vom Framework gewählt, um die Koordination von MapReduce durchzuführen. Dieser Knoten wird auch als Master bezeichnet. Die restlichen Programmkopien sind sogenannte workers, die vom Master alle Anweisungen erhalten. Es gibt M map-tasks und R reduce-tasks, die zugewiesen werden. Der Master sucht unproduktive workers und vergibt an jeden entweder einen map-task oder reduce-task.
3. Ein worker, der einen map-task zugewiesen bekommen hat, liest den Inhalt des jeweiligen Input Teiles ein und führt nun auf jedem Key/Value Paar die benutzerdefinierte Map-Methode aus und puffert die Ergebnisse zunächst im eigenen Hauptspeicher. In regelmäßigen Abständen werden die gepufferten Daten partitioniert und auf die lokale Festplatte geschrieben. Die Partitionierung erfolgt mit der Partitionierungsfunktion in R Partitionen, wobei R der Anzahl der folgenden reduce-jobs entspricht. Terminiert ein map-job, so werden die Informationen zur lokalen Speicherung (Speicherort, Dateiname und Größe) an den Master gemeldet. Der Master ist verantwortlich die Speicherortinformation an die reduce-workers weiterzuleiten.
4. Wenn der reduce-worker vom Master über den Speicherort benachrichtigt wurde, benutzt der reduce-worker Remote Procedure Calls (RPCs) um die gepufferten Daten von der lokalen Festplatte des map-workers zu lesen. Wenn der reduce-worker alle intermediate Daten gelesen hat, werden diese anhand der intermediate Keys sortiert, damit alle Teile mit demselben Key zusammen gruppiert werden.
5. Der reduce-worker iteriert über die sortierten Eingabepaare und wendet reduce auf jedes Eingabepaar an. Pro Partition legt der reduce-worker eine Ausgabedatei an und die Paare werden jeweils an diese Datei angehängt.
6. Nachdem sowohl der map- und reduce-job terminiert haben, weckt der Master das User Programm wieder auf. Der Benutzer hat nun Zugriff auf die Ausgabe und kann diese weiter verarbeiten [10] (S. 6)

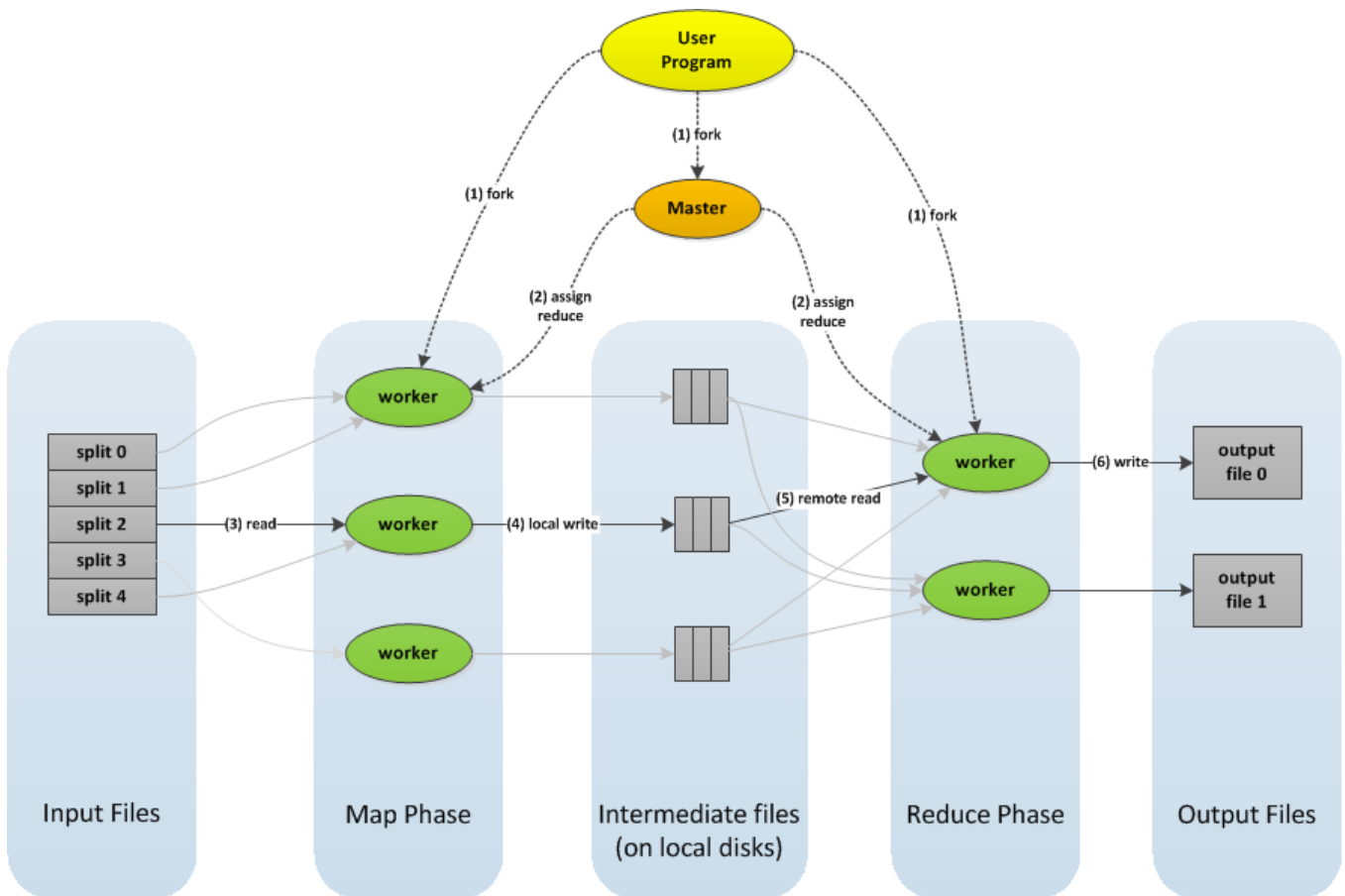


Abbildung 11: MapReduce Überblick [10]

MapReduce wurde im Jahre 2004 von den Google Inc. Mitarbeitern Jeffrey Dean und Sanjay Ghemawat in Mountain View Kalifornien ins Leben gerufen. Mit der Publikation 'MapReduce: Simplified Data Processing on Large Clusters' legten die beiden einen Vorschlag für ein Informationssystem zur Verarbeitung großer Datenmengen auf der Basis des eigenen verteilten Datensystem Google File System (GFS) vor. [84]

Im November 2008 vermeldete Google Inc. über ihren eigenen Weblog, dass mit einem MapReduce-basierten, parallelen Sortiervorgang der 1 TB-Sort-Benchmark nur 68 Sekunden lang gedauert hat. Dafür wurden 1 000 Rechner verwendet. Google verbesserte die bis dato beste Zeit einer ebenfalls parallelen Implementierung auf dem Open Source MapReduce Framework Hadoop um den Faktor 3, bei einer vergleichbaren Größe des Clusters. Der vorherige Rekord lag bei 209 Sekunden auf 910 Computern. Auf dem Weblog wird weiter berichtet, dass mit der MapReduce Implementierung 1 Petabyte innerhalb von 6 Stunden sortiert werden konnte. Dieses Datenvolumen entspricht ungefähr dem 12-fachen aller von der U.S. Library of Congress archivierten Webseiten. [85] [82]

Für MapReduce gibt es bereits unzählige Implementierungen, wobei Hadoop bisher die Bedeutendste darstellt. [83] Das folgende Kapitel 3.3 beschäftigt sich näher mit dem OpenSource Java Framework Hadoop.

3.3 Hadoop

Das Open Source Java Framework Hadoop wurde von dem Schöpfer des Apache Lucene Projektes Doug Cutting ins Leben gerufen. Der Name Hadoop ist kein Akronym sondern eine Phantasiebezeichnung. Doug Cutting erklärt, dass der Name von einem gelben Stoffelefanten kommt, den sein Sohn Hadoop genannt hatte. Da dieser Name einfach zu schreiben und auszusprechen, ohne Bedeutung und sonst nirgends verwendet wurde, fand Doug Cutting diesen Namen passend für sein Projekt. Doug Cutting ist der Meinung, dass Kinder perfekt für die Generierung solcher Projektnamen sind. So beschloss Doug Cutting Teilprojekte von Hadoop nicht nach ihrer Funktion zu benennen, sondern einen Name zu verwenden, der mit Elefanten oder anderen Tieren zu tun hat wie z.B. Pig. Kleineren Komponenten wurden anschaulichere Namen gegeben. So wird zum Beispiel die Komponente, welche die MapReduce Jobs überwacht, Jobtracker genannt. Hadoop hat seinen Ursprung in Apache Nutch, einer freien Open Source Suchmaschine. Eine eigene Suchmaschine von Grund auf zu entwickeln war ein gewagtes Ziel. Es ist eine Herausforderung eine Software zu schreiben, die leistungsfähig genug ist möglichst das gesamte World Wide Web zu indexieren und das ohne ein dediziertes Team. Solch ein Projekt ist nicht nur schwer zu realisieren sondern auch sehr kostspielig. Mike Cafarella und Doug Cutting planten für die Indexierung einer Milliarde Webseiten eine halbe Million Dollar an Hardwarekosten ein. Die monatlichen Betriebskosten schätzten sie auf \$ 30 000. Dennoch glaubten sie, dass es das Geld wert war um dieses Ziel zu erreichen. Das Nutch Projekt wurde im Jahre 2002 gestartet. Ein funktionierender Crawler und ein Such System entstanden in kürzester Zeit. Die Zwei begriffen sehr schnell, dass ihre Architektur mit den Milliarden an Webseiten nicht gut skalieren würde. Man versprach sich dann Besserung mit der im Jahre 2003 erschienenen Publikation von Google's verteiltem Datei System (GFS). Das Google File System sollte das Speicherproblem lösen, das durch die enorm große Datenmenge entsteht. Diese große Datenmenge wird durch die Indexierung und das Crawling erzeugt. Genauer gesagt würde GFS die administrativen Aufgaben wie das Managen der einzelnen Knoten reduzieren. Im Jahre 2004 schrieben die beiden eine Open Source Implementation, basierend auf der Idee von GFS, das Nutch Distributed Filesystem (NDFS). Auch im Jahre 2004 veröffentlichte Google das Paper, das der Welt das MapReduce Verfahren vorstellte. Ein Jahr später im Jahre 2005 hatten die Nutch Entwickler eine funktionierende MapReduce Implementation in Nutch mit dem dazugehörigen Nutch Distributed Filesystem (NDFS). Im Februar 2006 wurde aus dem ursprünglichen Teilprojekt von Lucene ein eigenes Projekt mit dem Namen Hadoop. Zu etwa der gleichen Zeit fing Doug Cutting bei Yahoo! an. Yahoo! Stellte ihm ein Team und Ressourcen zur Verfügung, um Hadoop weiter zu entwickeln. Das Ergebnis wurde dann im Februar 2008 demonstriert. In dieser Demonstration gab Yahoo! bekannt, dass ihr Suchindex von 10 000-core Hadoop Clustern generiert wurde. Hadoop wurde durch den Erfolg zum eigenen Apache Top-Level Projekt gemacht und bekam eine sehr große Community. Mittlerweile wird Hadoop von vielen anderen Unternehmen neben Yahoo! eingesetzt, wie z.B. Last.fm, Facebook und der New York Times. Die New York Times benutze Amazon's EC2 Cloud um 4 TB an eingescanntem Archiv Material in PDFs zu konvertieren und diese ihren Lesern online zur Verfügung zu stellen. Die Verarbeitung dauert weniger als 24 Stunden auf 100 Maschinen. Dies wäre vermutlich nicht möglich gewesen ohne das Amazon pay-by-hour-model und Hadoops leicht zu bedienendem Modell für parallele Verarbeitung. [67]

„Hadoop ist ein Framework für skalierbare, verteilt arbeitende Datenverarbeitungssysteme. Es basiert auf dem Algorithmus von Googles MapReduce und von Googles Filesystem. Diese Technologien werden dem hohen Datenvolumen gerecht indem sich neue Knoten leicht hinzufügen lassen. Dies ermöglicht riesige Rechenprozesse mit Petabytes an Daten auf Computerclustern, da die Verarbeitung der Daten parallel auf mehreren Computern durchgeführt werden kann.“[83] (S. 11)

Hadoop besteht aus drei wesentlichen Komponenten:

- Hadoop Common
- Hadoop Distributed FileSystem
- Hadoop MapReduce

Mit diesen drei Komponenten lässt sich ein vollständiger Hadoop Cluster betreiben, welcher nach dem Master, Slave Prinzip arbeitet. Hadoop Common stellt sämtliche Grundfunktionen bereit, welche andere Komponenten benötigen, wie eine implementierungsneutrale Filesystem Schnittstelle oder eine Schnittstelle für die Remote Procedure Call Kommunikation im Cluster und Bibliotheken. Das Hadoop Distributed Filesystem (HDFS) ist das

primäre Dateisystem von Hadoop und es folgt dem Vorbild des Google Filesystems (GFS). Es ist speziell für die Entwicklung von MapReduce Anwendungen optimiert. Hadoop MapReduce bietet alle Funktionen um nach dem Programmiermodell zu entwickeln und es basiert auf dem in Kapitel 3.2 beschriebenen Google MapReduce Verfahren und dessen Funktionsweise. [83]

3.3.1 Hadoop Ecosystem

„Seit den Anfängen von Hadoop sind viele Anwendungen entstanden, die mit dem oder für das Framework entwickelt wurden. Viele haben als firmeninterne Lösungen für bestimmte Aufgaben gedient, bevor sie der Hadoop Entwickler Community als Open Source Projekt präsentiert wurden. So wundert es auch nicht, dass die meisten Projekte dieses Hadoop Ökosystems von großen Firmen stammen, die selber große Hadoop Installationen betreiben.“ [84] Im folgenden Abschnitt werden die einzelnen Projekte zur besseren Orientierung kurz beschrieben.

- Common: Eine Reihe von Komponenten und Interfaces für verteilte Datensysteme und allgemeine I/O (Serialisierung, Java RPC, Persistent Data Structures).
- Avro: Ist ein Daten Serialisierungsframework für effiziente, Cross Language und Persistent Data Storage.
- MapReduce: Ein Modell zur verteilten Daten Verarbeitung und Execution Environment für große Cluster.
- HDFS: Ein verteiltes Filesystem, welches auf großen Clustern läuft.
- Pig: Kann zur Analyse von großen Datenmengen eingesetzt werden. Hierfür stellt Pig sowohl eine Programmiersprache, die High Level-Abfragen auf Datenbeständen ermöglicht, als auch einen Compiler zur Ausführung.
- Hive: Ein verteiltes Data Warehouse. Hive managet die gespeicherten Daten auf dem HDFS und ermöglicht einfache SQL-ähnliche Abfragen, die in MapReduce Jobs umgewandelt werden.
- HBase: Ist eine skalierbare, einfache Datenbank zur Verwaltung von großen Datenmengen innerhalb eines Hadoop Clusters. HBase verwendet das HDFS als zugrunde liegenden Speicher. Die HBase Datenbank basiert auf Google BigTable. Die HBase Datenstruktur ist für Daten geeignet, die selten geändert werden, jedoch sehr häufig ergänzt werden.
- ZooKeeper: Ein verteilter hochverfügbarkeits Koordinierungsservice, der Konfigurationen für andere Systeme bereitstellt.
- Sqoop: Ein Tool um effizient Daten zwischen relationalen Datenbanken und HDFS zu verschieben. [86]
- Hama und Mahout: Diese beiden Projekte sind für wissenschaftliche Datenverarbeitung unter Hadoop gedacht. Hama ist ein Matrix Berechnungspaket um z.B. Inverse, Eigenwert und Eigenvektoren und andere Matrix Operationen zu berechnen. Der Begriff Mahout bedeutet im indischen Sprachraum, Elefantenführer. Mahout stellt eine Sammlung von Algorithmen aus dem Bereich Maschinelles Lernen zur Verfügung. [87]
- CloudBase: Ist für den Datawarehouse Layer konzipiert um Logdateien zu analysieren.
- Hue: Mit Hue lassen sich Prozessabläufe in einer webbasierten Benutzeroberfläche automatisieren. Über diese Oberfläche kann man mit anderen Hadoop Tools wie Hive und Pig interagieren.
- Flume: Ein verteilter Service um effektiv große Datenmengen zu verschieben. Flume wird hauptsächlich für die Logdateianalyse eingesetzt, dabei werden die Logdateien auf verschiedene Knoten aufgeteilt. Am Ende werden die analysierten Daten via HDFS bereitgestellt.
- Wirr: Ziel von Wirr ist die Bereitstellung von einheitlichen Tools für den Betrieb Cloudbasierender Services. Dabei wird der einheitliche Zugriff auf die Kommandozeiletools und Programmierschnittstellen der verschiedenen Cloud Provider ermöglicht. [84]
- Oozie: Ist ein Server basierendes Workflow Management System für Hadoop. Die Cloudera Version bringt eine Unterstützung für Hive- und Sqoop-Jobs mit.
- Snappy: Ist für Kompressions- und Dekompressionsaufgaben zuständig [88]

Es sind noch eine ganze Reihe weiterer Projekte für Hadoop verfügbar und es entstehen ständig neue Projekte. Diese Liste soll lediglich einen kurzen Überblick über die wichtigsten Projekte und deren Aufgabe geben.

3.3.2 Building Blocks von Hadoop

- Job Tracker: Der Job Tracker koordiniert die MapReduce Jobs und ist auch zuständig für das Splitten der Aufgaben.
- Task Tracker: Der Task Tracker führt die MapReduce Aufgaben aus. Als erstes wird der Job Tracker informiert, ob der Task Tracker bereit für eine neue Aufgabe ist. Bevor der Task Tracker eine Aufgabe zugeteilt bekommt, muss der Job Tracker einen Job wählen. Task Tracker haben eine limitierte Anzahl an Slots für MapReduce Aufgaben. Wieviele MapReduce Aufgaben ausgeführt werden können, hängt von der Hardware ab wie z.B. die Anzahl der Prozessoren, Cores und vom Speicher des Knoten.
- NameNode: Der NameNode wird als Namespace Manager verwendet. Er verwaltet die Metadaten des Dateisystems wie Dateien und Verzeichnisstrukturen und führt Namensraum Operationen wie das Öffnen, Umbenennen und Schließen von Ordnern und Dateien aus. Er führt auch das Mapping von Blöcken zu DataNodes aus. In einem HDFS existiert nur ein NameNode. Falls der NameNode nicht mehr verfügbar sein sollte, geht HDFS offline. In der Master/Slave Architektur von HDFS übernimmt der NameNode die Rolle des Masters.
- DataNode: Data Nodes haben die Slave Rolle und sind für die Verwaltung der Nutzdaten im HDFS auf den einzelnen Clusterknoten verantwortlich. Diese verwalten nur die einzelnen Blöcke, die auf ihnen abgelegt wurden. Die Information welcher Block von wem bearbeitet wird und welche Dateien enthalten sind, hat nur der NameNode. Die DataNodes sind zuständig für alle Schreib- und Lesezugriffe vom Dateisystem. Sie können Blöcke erstellen, löschen, und nachbilden. [89]

3.3.3 Hadoop Distributed File System

Hadoop Distributed File System (HDFS) ist ein Dateisystem, das für den Einsatz von MapReduce Jobs konzipiert wurde. HDFS kommt mit Direktzugriff nicht besonders gut zu recht. Für die Ausfallsicherheit werden die Daten einfach auf mehrere Storage Nodes gespiegelt. Das ist in der Hadoop Community auch als Replication bekannt. Solange mindestens eine fehlerfreie Replikation verfügbar ist, ist es nicht notwendig den User darüber in Kenntnis zu setzen. [80] Die Verarbeitung einer großen Datenmenge läuft idealerweise auf mehreren Kernen einer Multi-Core-CPU, dabei werden die Daten von einer oder mehreren Festplatten gelesen oder geschrieben. Bei der parallelen Verarbeitung sehr großer Datenmengen, ist der I/O beim Festplattenzugriff der limitierende Faktor und somit Flaschenhals der gesamten Architektur. Daher besitzen moderne Server Systeme einen RAID-Controller, um die Datenlast auf mehrere physikalische Festplatten aufzuteilen. Allerdings hat auch dieser Ansatz irgendwo seine Grenzkapazität. HDFS macht sich bekannte Eigenschaften verteilter Datensysteme zunutze.

- Parallelisierung: Bei einem verteilten Filesystem wie es bei Hadoop mittels HDFS der Fall ist, wird der Geschwindigkeitsvorteil durch das gleichzeitige Lesen und Schreiben auf vielen Rechnern im Netzwerk erreicht.
- Cluster-lokale Verarbeitung: Die MapReduce Verarbeitung läuft auf den einzelnen Cluster Knoten, welche die Daten enthalten. Dadurch wird die zu übertragende Datenmenge auf ein Minimum reduziert.
- Failover: Sollte ein Cluster Knoten ausfallen, wird dies von Hadoop registriert und startet den MapReduce Prozess auf einem anderen Knoten, der bereits über die replizierten Daten verfügt.

Durch diese Funktionen ist es Hadoop möglich auf kostengünstiger Standardhardware sehr leistungsfähig und sicher zu arbeiten. [84]

„HDFS basiert, wie MapReduce, auf der Master-Slave Architektur. Der Master in einem HDFS-Cluster, Namenode genannt, ist die zentrale Verwaltungsstelle des HDFS Clusters. Er speichert keine Nutzdaten, sondern ausschließlich Metadaten (z.B. die Zuordnung, welcher Slave welche Daten speichert). Die Slaves in einem HDFS Cluster, die als Datanodes bezeichnet werden, speichern nur die Nutzdaten. Durch die Verwendung von Standard Hardware als Basis, ist eine häufige Ausfallquote der Hardware gegeben. Daher überwacht der Namenode, wie der Jobtracker auch, die Verfügbarkeit der einzelnen Datanodes und verteilt die Daten neu, sobald ein Datanode ausfällt. Die Speicherung der Daten erfolgt, genau wie bei einem nicht verteilten Dateisystem, in Form von Dateien und Verzeichnissen. Die Daten eines HDFS Clusters sind in Blöcken gleicher Größe auf den Datanodes verteilt. Ein Daten-Block hat in der Standardeinstellung die Größe von 64MB und ist drei Mal im Cluster repliziert. Die Anzahl der Replikationen lässt sich mittels eines Konfigurationsparameters einstellen. Namenodes lassen sich mit einem sogenannten Master-File-Table eines nichtverteilten Dateisystems vergleichen. Beide speichern nur Metadaten des Dateisystems. Ohne diese Metadaten ist das Dateisystem nicht funktionsfähig. Daher ergibt sich bei HDFS auch ein Single-Point-of-Failure: Fällt der Namenode eines HDFS Clusters aus oder verliert die Metadaten, ist das Dateisystem des Clusters zerstört. Bei der Kommunikation zwischen einem Client, der Daten anfragt oder sendet, agiert der Namenode nur als Verzeichnisdienst.“ [90] Das HDFS Filesystem ist für die Verwaltung großer Dateien ausgelegt, zwar kann HDFS auch kleine Dateien handeln, jedoch ist das System hierfür nicht optimiert und kann daher auch keine guten Performancewerte erreichen. [90]

3.3.4 Funktionsweise

Dieses Kapitel beschäftigt sich mit dem Ablauf eines MapReduce-Jobs auf einem Hadoop Cluster und den dabei ablaufenden Prozeduren. Betrachtet man die Abbildung 12 sieht man vier zentrale Komponenten: Client, Jobtracker, Tasktracker und Shared Filesystem. Eine genaue Beschreibung dieser Komponenten findet man in den Kapiteln 3.3.2 und 3.3.3.

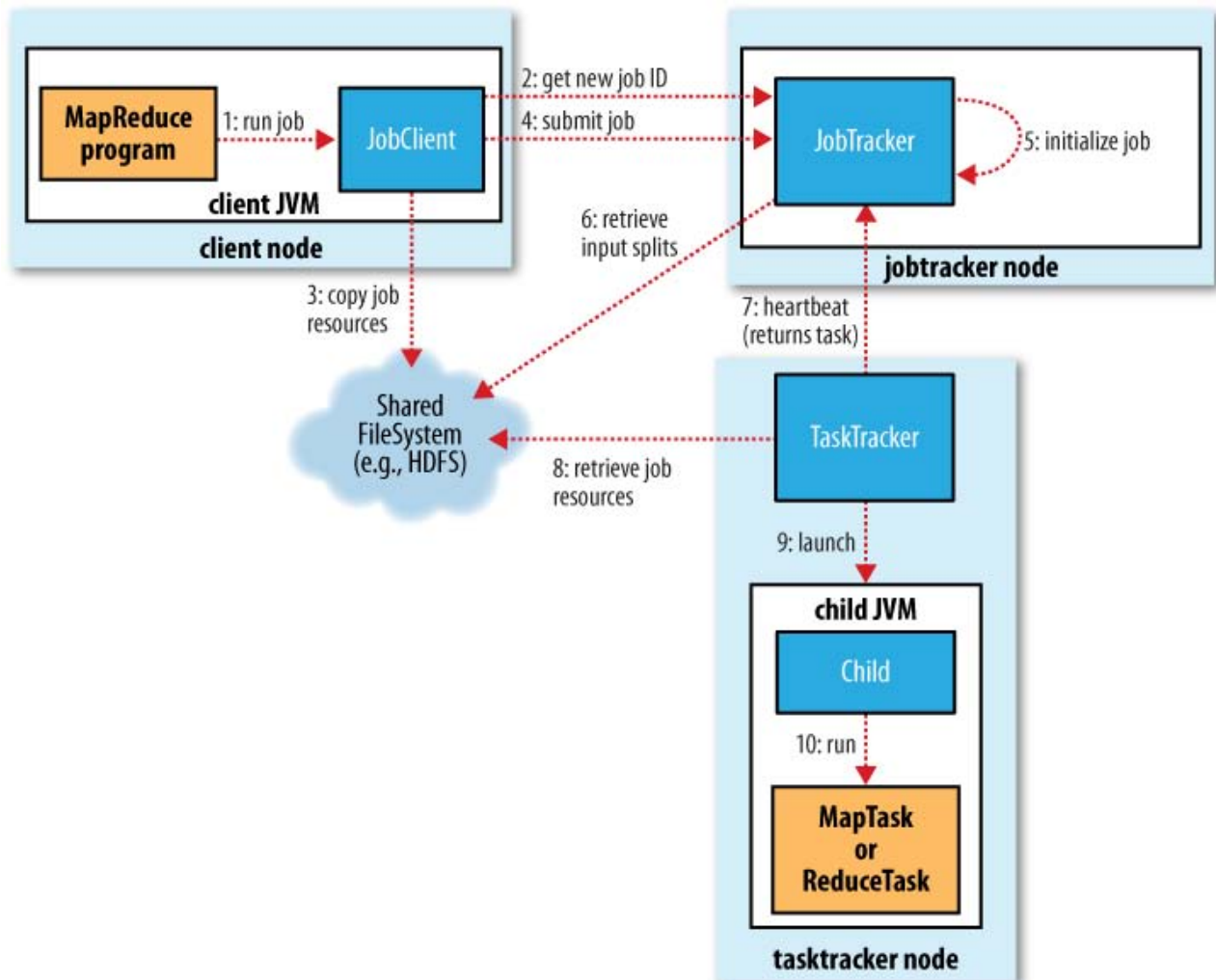


Abbildung 12: Ablauf eines MapReduce Jobs unter Hadoop [86] (S. 168)

„Beim Aufruf der Methode runJob wird eine neue JobClient-Instanz auf dem Client erzeugt (Schritt 1 in Abbildung 12). Anschließend fragt der JobClient mit der Methode getNewJobID beim Jobtracker eine neue Job-ID an (Schritt 2). Daraufhin kopiert der JobClient alle benötigten Job-Daten in das gemeinsame Dateisystem. Dies umfasst das auszuführende Programm und die dafür nötigen Bibliotheken, sowie individuelle Konfigurationsdaten (Schritt 3). Nachdem dieser Prozess abgeschlossen ist, wird der Job durch den Aufruf von submitJob an den Jobtracker übergeben. Dabei werden die festgelegten Spezifikationen für Input und Output überprüft. Falls diese nicht den Vorgaben entsprechen, wird der Job abgebrochen und ein entsprechender Fehler zurückgegeben (Schritt 4). Nach dem Aufruf von submitJob muss der Job auf dem Jobtracker initialisiert werden. Der Jobtracker verwaltet dabei alle übergebenen Jobs mit einer Job-Warteschlange. Der Jobtracker ist ebenfalls für die Einteilung eines Jobs in mehrere Tasks zuständig, welche jeweils eine eigene Task-ID erhalten. Die eigentliche Initialisierung

beinhaltet die Instanziierung eines Job-Objekts, dem nachfolgend die einzelnen Tasks eines Jobs sowie deren Status- und Verlaufsmeldungen zugeordnet sind (Schritt 5). Anschließend greift der Jobtracker auf die vom JobClient in den Konfigurationsdaten festgelegten Input Splits zu und übergibt sie dem jeweiligen Job. Input Splits sind logische Referenzen auf einem Block im gemeinsamen Dateisystem und repräsentieren die vom Job zu verarbeitenden Daten. Jedem (Map-)Task wird dabei genau ein Input Split zugeordnet (Schritt 6). Die einzelnen Tasks werden den Tasktrackern gemäß dem Job Scheduling zugeteilt. Die einzelnen Tasktracker kommunizieren über periodisch gesendete heartbeats mit dem Job Tracker. Der sogenannte heartbeat ist dafür zuständig Statusmeldungen an den Jobtracker zu melden. Dies umfasst beispielsweise den aktuellen Fortschritt der Task-Bearbeitung, ob der Tasktracker überhaupt zur Verfügung steht aber auch jede andere Art von Nachricht. Wenn ein Tasktracker dem Jobtracker meldet, dass er zur Verfügung steht, kann ihm ein neuer Task zugeteilt werden. Dabei gilt zu beachten, dass Tasktracker nur eine begrenzte Anzahl an Tasks gleichzeitig bearbeiten können. Standardmäßig gilt, dass nur ein Task pro Prozessorkern erlaubt ist, jedoch kann es unter Umständen auch sinnvoll sein eine höhere Anzahl an Threads zu definieren. Die Anzahl der einzelnen Map- und Reducetasks lässt sich in Hadoop unabhängig von der Anzahl der Prozessorkerne konfigurieren (Schritt 7). Wenn einem Tasktracker ein Task zugeordnet wurde, werden zuerst die für den Task benötigten Daten und Bibliotheken, sowie das eigentliche Programm aus dem gemeinsamen Dateisystem abgeholt. Hierbei gilt, dass Tasks nur lokal vorhandene Input Splits bearbeiten und jeder Map-Task nur einen Input Split verarbeitet. Die aus dem gemeinsamen Dateisystem überführten Dateien werden in einem lokalen Arbeitsverzeichnis gespeichert (Schritt 8). Daraufhin startet die TaskRunner Instanz eine neue Java Virtual Machine (JVM) zur Ausführung eines Tasks. Für jeden Task wird dabei eine eigene JVM gestartet (Schritt 9). Im Anschluss wird der eigentliche Task ausgeführt. Die JVM des Child-Prozesses kommuniziert dabei periodisch mit dem Parent-Tasktracker und tauscht Statusmeldungen aus. Ein Task wird anschließend als 'finished' gemeldet, falls der Task erfolgreich beendet wurde, oder als 'failed' gemeldet, falls der Task abgebrochen werden musste. Sobald alle Tasks eines Jobs erfolgreich beendet wurden, wird der Job vom Jobtracker als 'successful' gemeldet (Schritt 10).“ [91]

3.4 Cloudera

Um einen Hadoop Cluster für die eDiscovery Software FreeEed installieren zu können, wurde die Cloudera Distribution gewählt (siehe Anhang 2 - Deployment), welche man auf Basis von Ubuntu installiert hat (siehe Anhang 1 - Installation).

Cloudera war eine der ersten Firmen, die sowohl ihre eigene Hadoop-Distribution entwickelte als auch das Thema Support und Services kommerzialisierte. So lassen sich kostenpflichtige Beratungsservice ebenso wie Schulungen rund um Hadoop und darauf aufsetzende Anwendungen bei Cloudera buchen. Namhafte Mitarbeiter wie der Hadoop-Erfinder Doug Cutting selbst stehen für die Professionalität, zu der auch die beachtliche Kundenkartei gehört. Vor allem Kunden aus dem amerikanischen Online-Marketing-Umfeld gehörten dazu, wie etwa AOL, Adconion oder ComScore. [84]

Die Cloudera Hadoop Distribution bringt folgende Softwarekomponenten aus dem Hadoop Ökosystem mit: Flume, Sqoop, Hive, Pig, HBase, Oozie, Hue, Zookeeper und Wrr. [92] Die einzelnen Komponenten wurden bereits im Kapitel 3.3.1 Hadoop Ecosystem beschrieben. Einige dieser Komponenten wurden von Cloudera selbst entwickelt, diese sind trotzdem über die Apache-Lizenz als Open Source zugänglich. In der Abbildung 13 findet man nochmals die wichtigsten Komponenten der Cloudera Hadoop Distribution kategorisiert und in einem Schichtenmodell angeordnet.

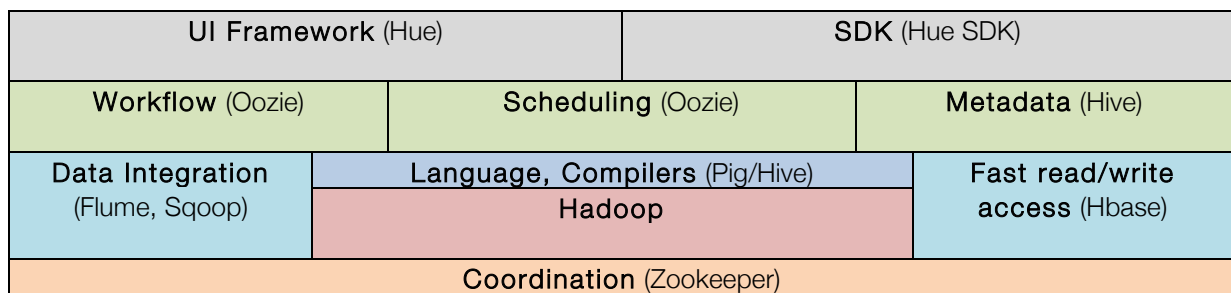


Abbildung 13: Cloudera's Hadoop Distribution [93]

Cloudera bzw. Hadoop kann in folgenden Modes betrieben werden:

- Pseudo-Distributed Mode: Beim diesem Mode laufen sämtliche Hadoop-Services auf einem Host. Dabei kommunizieren die Services aber untereinander wie im Cluster-Mode.
- Cluster Mode: Dieser Mode ist der eigentliche Mode wenn man einen Cluster mit mehreren Hosts betreibt. [84]

Dabei kann man Cloudera als Linux Paket installieren, wobei jedoch die komplette Konfiguration der Knoten manuell durchgeführt werden muss. Verwendet man den Cloudera Manager, wird einem diese Arbeit abgenommen. Der Cloudera Manager liegt in einer kommerziellen Version namens Cloudera Manager und einer freien Version namens Cloudera Manager Free Edition vor. Die Free Edition ist auf 50 Hosts limitiert und einige Features wie LDAP, Kerberos Authentifizierung, Monitoring, Log Search, Statistiken und Oracle Database Support sind nicht verfügbar. Der Cloudera Manager unterstützt nun seit der neusten Version 4.0 weitaus mehr Linux Distribution als sein Vorgänger. Zu den unterstützten Distributionen gehören die folgenden Distributionen: Red Hat, CentOS, SUSE Linux Enterprise Server, Oracle Enterprise Linux, Debian und Ubuntu. Alle Cloudera Softwarekomponenten sind 64-Bit fähig.

Der Cloudera Manager Server verrichtet die folgenden Aufgaben:

- Stellt das Cloudera Data Manager Modell bereit, welches in der Cloudera Manager Server Datenbank gespeichert ist. Das Data Modell ist ein Katalog der verfügbaren Knoten im Cluster und der Services, Rolle und die Konfiguration die jedem Knoten zugewiesen ist.
- Kommuniziert mit den Agents und sendet Konfigurationsanweisungen und überwacht den Heartbeat der Agents.
- Führt Befehle für Tasks aus.

- Stellt eine Admin Konsole bereit für Management- und Konfigurations- Aufgaben.
- Erstellt, liest, validiert, aktualisiert und löscht Konfigurationseinstellungen.
- Ermittelt und zeigt den Health Status des Clusters.
- Zeigt den belegten Speicherplatz, CPU und RAM Auslastung.
- Stellt ein flächendeckendes Set von APIs für die verschiedensten Cloudera Manager Features bereit.

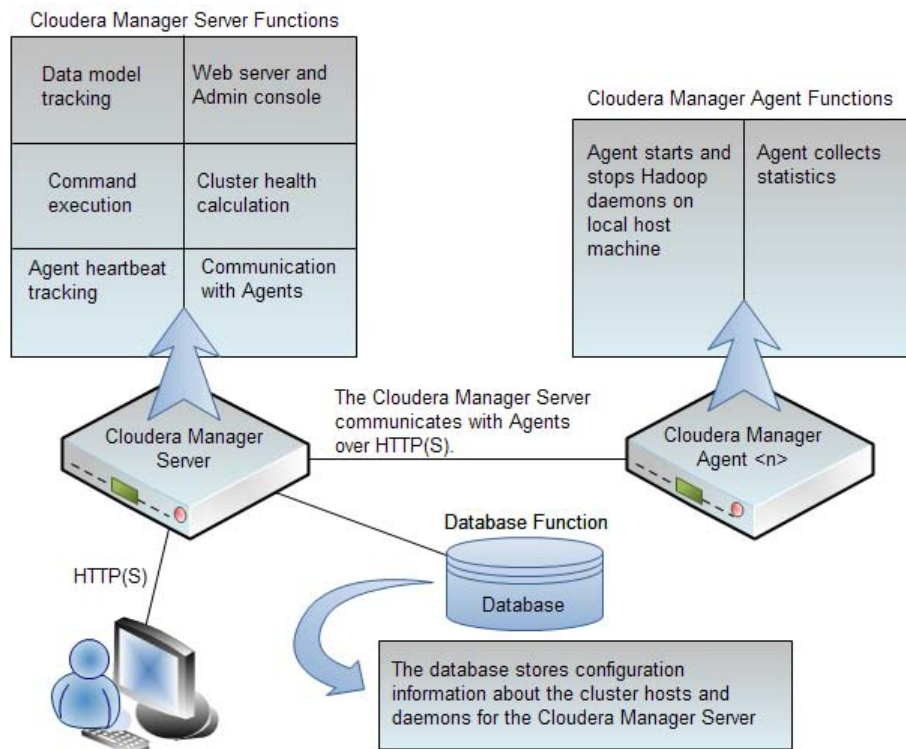


Abbildung 14: Cloudera Manager Architektur [94]

3.5 FreeEed/SHMCloud

Oftmals kostet nicht nur der gesamte eDiscovery Prozess viel Geld, sondern auch das Erwerben einer entsprechenden eDiscovery Software kann teuer sein. Dass dies nicht immer der Fall ist, will das Open Source Projekt FreeEed beweisen. Das kleine unabhängige Team aus Houston hat sich zum Ziel gesetzt, Firmen eine eDiscovery Software zur Verfügung zu stellen, die für jedes Budget und für jede Verfahrensgröße ausgelegt ist. Laut dem Chefentwickler Mark Kerzner besteht die Philosophie in dem Projekt darin, dass jeder die Open Source Software ausprobieren kann, ohne teure Lizenzen zu kaufen. Die Software kann dann je nach Datengröße auf einem einzelnen Server oder aber auch auf einem Cloud-based Cluster ausgeführt werden. [95] Um solch eine flexible Software zu entwickeln, wurde auf das Apache Hadoop Framework gesetzt. FreeEed ist eine Multiplattform Software und kann sowohl auf Windows, Mac OS X und Linux eingesetzt werden. Die grundlegenden Building Blocks von FreeEed sind HDFS, Hadoop, HBase, Hive, Mahout und Pig Tika, Lucene und Solr. [12]

- FreeEed setzt auf das Hadoop Distributed File System (HDFS), welches für die zuverlässige Speicherung von großen Datenmengen auf verschiedenen Knoten in einem großen Cluster konzipiert wurde. HDFS wurde dem Google Filesystem nachempfunden, weitere Details und die Funktionsweise von HDFS findet man im Kapitel 3.3.3 Hadoop Distributed File System.
- Hadoop ist ein in Java geschriebenes Open Source Framework für skalierbare, verteilt arbeitende Software. Mehr über Hadoop findet man im Kapitel 3.3 Hadoop.
- HBase ist eine einfache, skalierbare Datenbank um sehr große Datenmengen innerhalb eines Hadoop-Clusters zu verwalten.
- Hive managet die gespeicherten Daten auf dem HDFS und ermöglicht einfache SQL-ähnliche Abfragen, die in MapReduce Jobs umgewandelt werden.
- Mahout stellt eine Sammlung von Algorithmen aus dem Bereich Maschinelles Lernen zur Verfügung.
- Pig kann zur Analyse von großen Datenmengen eingesetzt werden. Hierfür stellt Pig sowohl eine Programmiersprache, die High Level-Abfragen auf Datenbestände ermöglicht, als auch einen Compiler zur Ausführung bereit.
- Tika ist eine Inhalt Analyse, MIME Detection und erweiterbares Parsing Framework, das bei der Metadatenerkennung und Extrahierung hilft. Zusätzlich stellt Tika eine allgemeine Parser API für strukturierte Textinhalte für viele unterschiedliche Dokumentformate bereit. Tika ist auch mit bereits existierenden Parser Libraries wie Apache POI für Microsoft Formate, PDFBox für Adobe PDF ausgestattet. Zu den Features von Tika gehört das Erkennen von Content Types, Character Encoding, Sprache, Charaktereigenschaften und das Extrahieren von strukturierten Textinhalten aus dem Dokument. Wie auch XML benutzt auch Tika SAX Events, um die Daten zu extrahieren. [96] (S. 2) Die Idee für Tika stammt eigentlich aus dem Apache Nutch Projek und wurde wie auch Hadoop von Dough Cutting ins Leben gerufen. Apache Nutch ist ein Open Source Framework für eine Large-scale Suchmaschine. [97]
- Lucene ist eine leistungsstarke, skalierbare Information Retrieval (IR) Library. Information Retrieval ist der Prozess nach Dokumenten zu suchen, Informationen in Dokumenten oder Metadaten zu Dokumenten. Mit Lucene kann man verschiedene Suchfunktionen in seine Applikationen einbinden. Bei Lucene handelt es sich um ein in Java geschriebenes, freies Open Source Projekt, das unter der Apache Software License steht. Lucene ist inzwischen schon seit mehreren Jahren die am häufigsten verwendete Freie IR Library. So wird Lucene zum Beispiel bei der Wikipedia Suchfunktion verwendet und wird ebenso bei Twitter eingesetzt. Lucene kann alle Daten indizieren und durchsuchbar machen, aus denen sich Text extrahieren lässt. Dabei ist es egal aus welcher Quelle die Daten stammen, welches Format oder Sprache verwendet wird, so lange Text daraus abgeleitet werden kann. Das bedeutet, man kann Daten indizieren oder durchsuchen egal ob diese auf Webseiten, Remote Webservern, oder in lokal gespeicherten Dokumenten, einfachen Textdateien, Word Dokumenten, XML, E-Mail Archiven, Instant Messenger Chats, HTML, PDF oder in jeglichen anderen Dateiformaten, aus denen Text extrahiert werden kann,

gespeichert sind. Des Weiteren ist es möglich Daten aus Datenbanken zu indizieren und eine Volltextsuche durchzuführen. Diese Möglichkeit bieten Datenbankhersteller oft nur in einem begrenzten Umfang an. [98]

- Solr ist ein Open Source Enterprise Search Server, der auf Lucene basiert. Solr wurde von CNET entwickelt und wird heute noch auf deren Website eingesetzt, des Weiteren wird Solr bei Zappos, Netflix und zahlreichen anderen Staatlichen und Unternehmenswebseiten eingesetzt. Der Name stammt aus dem Akronym von Search on Lucene and Resin. Solr ist in Java geschrieben und kommuniziert über Standardprotokolle wie HTTP, XML und JSON. [99]

Nachdem die einzelnen Building Blocks erläutert wurden, soll nun die Interaktion der Hauptkomponenten anhand der Abbildung 15 erklärt werden. FreeEed nutzt das Hadoop Framework um die Datenverarbeitung abzuwickeln. Hadoop ist für die Verarbeitung, Speicherung und Analyse der großen Menge an verteilten und unstrukturierten Daten zuständig. Jede Datei wird aus dem Archiv eingelesen und erhält eine eindeutige ID. Danach werden alle Daten in Tika verarbeitet, wobei Tika Text und Metadaten der Daten extrahiert. Dieser Prozess wird auch als Staging bezeichnet. Nachdem sowohl Text und Metadaten extrahiert wurden, werden die extrahierten Daten mithilfe von Lucene, Solr durchsucht und indiziert. Die Suchresultate werden in einem Output Ordner auf dem HDFS abgelegt.

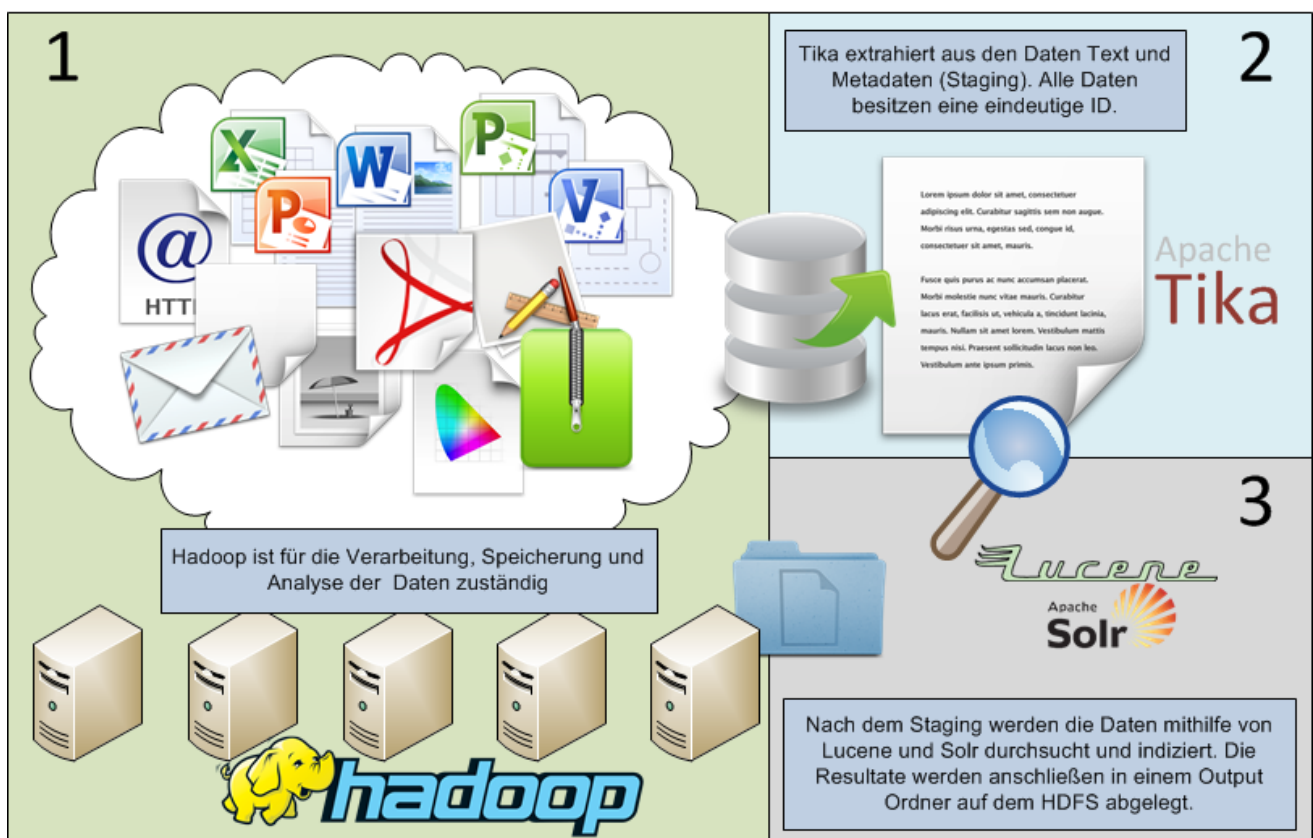


Abbildung 15: Funktionsweise FreeEed

Da FreeEed zur Extrahierung von Text und Metadaten Apache Tika verwendet, werden sämtliche Dateiformate unterstützt, die auch von Tika unterstützt werden. Dazu gehören HTML, XML, Microsoft Office, OpenOffice, Apple iWork, PDF, EPUB, RTF, Text Formate, Komprimierte Dateien, Archive, Bild Formate, Video Formate, Java Class

Files, Mail Formate, AutoCAD und einige weitere Formate. [100] Nähere Informationen zu den unterstützten Dateiformaten findet man auf der Apache Tika Website⁸ und unter der Github Projektseite von FreeEed⁹.

Zusätzlich zum Open Source Projekt FreeEed entwickelte der Entwickler Mark Kerzner das kommerzielle Produkt SHMCloud. SHMCloud ist eine Software as a Service Applikation, mit der es möglich ist dieselben Funktionen wie FreeEed zu nutzen, jedoch in der Amazon Cloud anstatt auf einem lokalen Hadoop Cluster. So verfügt SHMCloud über eine GUI mit deren Hilfe die Verbindung zur Amazon Cloud hergestellt werden kann, um dann die Daten in ein Amazon S3 Bucket hochzuladen und anschließend mit Hadoop EC2 Instanzen zu verarbeiten. Am 31.05.2012 wurde die erste geschlossene Beta von SHMCloud veröffentlicht. Die Abbildung 16 zeigt die Funktionsweise von SHMCloud. Die zu analysierenden Daten werden mithilfe von SHMCloud in einen Amazon S3 Bucket hochgeladen. SHMCloud streamt die Daten aus dem S3 Bucket und startet je nach Einstellung eine gewisse Anzahl von Amazon EC2 Instanzen um die Daten parallel zu verarbeiten. Die Amazon EC2 Instanz mit der Amazon Machine Images ID ami-2f812e46 ist eine speziell für SHMCloud angepasste Hadoop Installation auf Ubuntu Linux. [101] SHMCloud ist inzwischen auch auf dem Amazon Marketplace zu finden¹⁰.

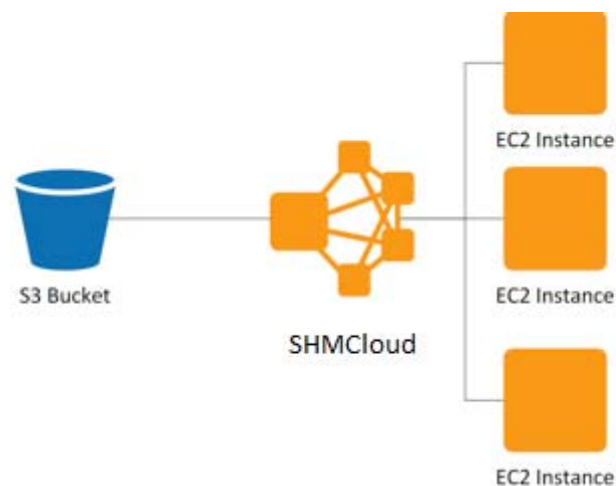


Abbildung 16: Funktionsweise SHMCloud [101]

Mithilfe von SHMCloud dauert die Analyse von 50 GB gepackten Enron Daten (entpackt 100 GB) mit 49 Knoten lediglich 30 Minuten. Für SHMCloud müssen jetzt noch keine Lizenzkosten bezahlt werden, es wird pro Maschinenstunde abgerechnet. Zusätzlich zu den Amazon AWS Kosten für Cloud Speicher im S3 Bucket und für die EC2 Instanzen müssen \$ 10 pro Instanz an die Firma SHMsoft, welche hinter SHMCloud steht, gezahlt werden. Das bedeutet bei 1h x \$ 10 x 50 Knoten = \$ 500. Zuzüglich müssen noch die Ressourcen für den S3 Bucket und die EC2 Instanzen von Amazon bezahlt werden. [102]

Sowohl FreeEed als auch SHMCloud sind bisher noch keine fertigen Produkte und stecken noch in den Kinderschuhen. Beide Projekte befinden sich in der Entwicklung und erhalten noch das eine oder andere Feature.

⁸ <http://tika.apache.org/1.2/formats.html>

⁹ <https://github.com/markkerzner/FreeEed/wiki/k-Supported-file-formats>

¹⁰ https://aws.amazon.com/marketplace/pp/B0097CE8GO/ref=srh_res_product_title

Die Abbildung 17 zeigt eine grobe Timeline für das Projekt. Im Jahre 2012 soll eine Analytics und Review Plattform implementiert werden, welche äußerst wichtig für ein eDiscovery Tool ist und leider bisher noch fehlt.

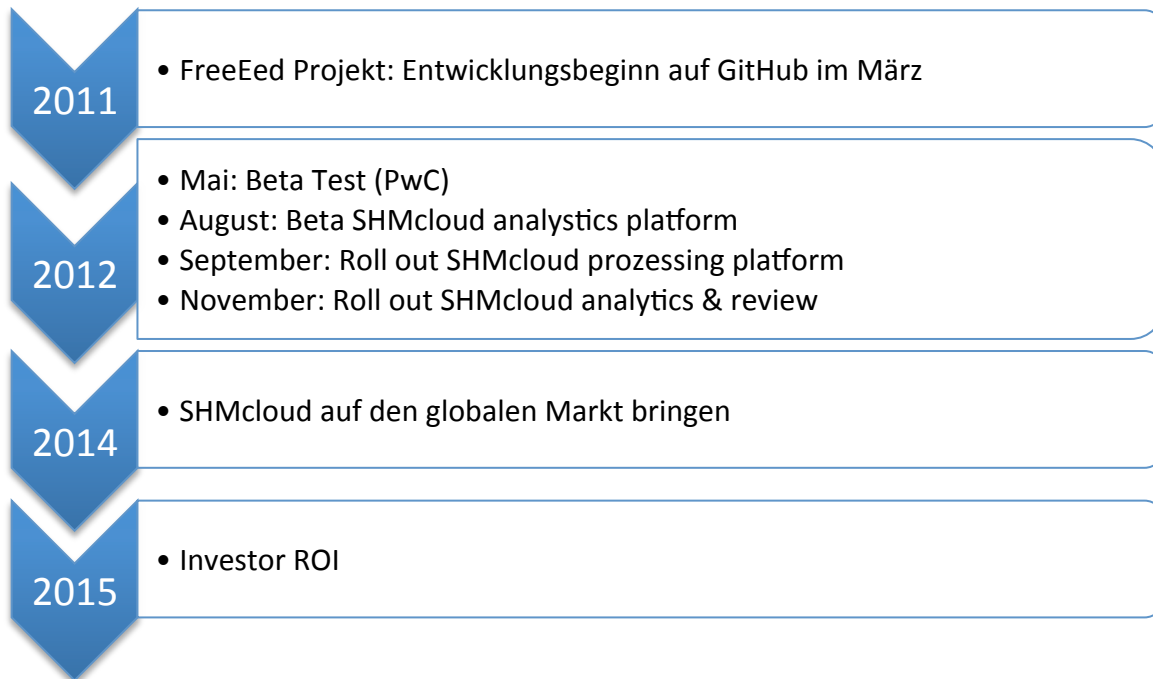


Abbildung 17: FreeEed SHMCloud Timeline [103]

Zukünftig soll auch die DataStax [104] Technologie für eine Schnellsuche und Review Funktion integriert werden. [105] Bei DataStax handelt es sich um eine gesamte Big Data platform, die auf der Basis von Apache Cassandra [106] Architektur gebaut wurde. [106] [104] „DataStax schafft eine smart integrierte Big Data platform, indem es die Leistung von Cassandra mit Analytics von Apache Hadoop und Enterprise Search mit Apache Solr verbindet. Mit DataStax erhalten Echtzeit-, Analyse und Such-Workloads eine maximale Leistung ohne Konflikte. Zusätzlich bietet es den Vorteil, nur eine Datenbank verwalten zu müssen.“ [107]

Die FreeEed/SHMCloud Software wird durch die Desktop Anwendung ‘Player’ gesteuert. Der Player ermöglicht es dem User Projekte zu verwalten, Daten hinzuzufügen und das Setzen von bestimmten Verarbeitungsparametern. Die Daten zu ‘Stagen’ (kopieren der Daten auf den Hadoop Cluster oder die Amazon AWS Cloud) und dann die Verarbeitung zu starten und zu steuern. Die Abbildung 18 zeigt genau dies auf der rechten Seite, was der aktuellen Version von SHMCloud entspricht. Zukünftig soll die Verarbeitung jedoch so aussehen wie in Abbildung 18, auf der linken Seite zu sehen. Dabei sollen folgende Vorteile zum Tragen kommen:

1. Browser Interface, damit kann die Software via Workstation bis hin zum Tablet gesteuert werden
2. Optimierte Datensammlung
3. Proprietär data Quellen und Datenbanken
4. Suche und erst Review direkt auf dem Cluster
5. Erlaubt zusätzliches culling, basierend auf vorherigen Ergebnissen

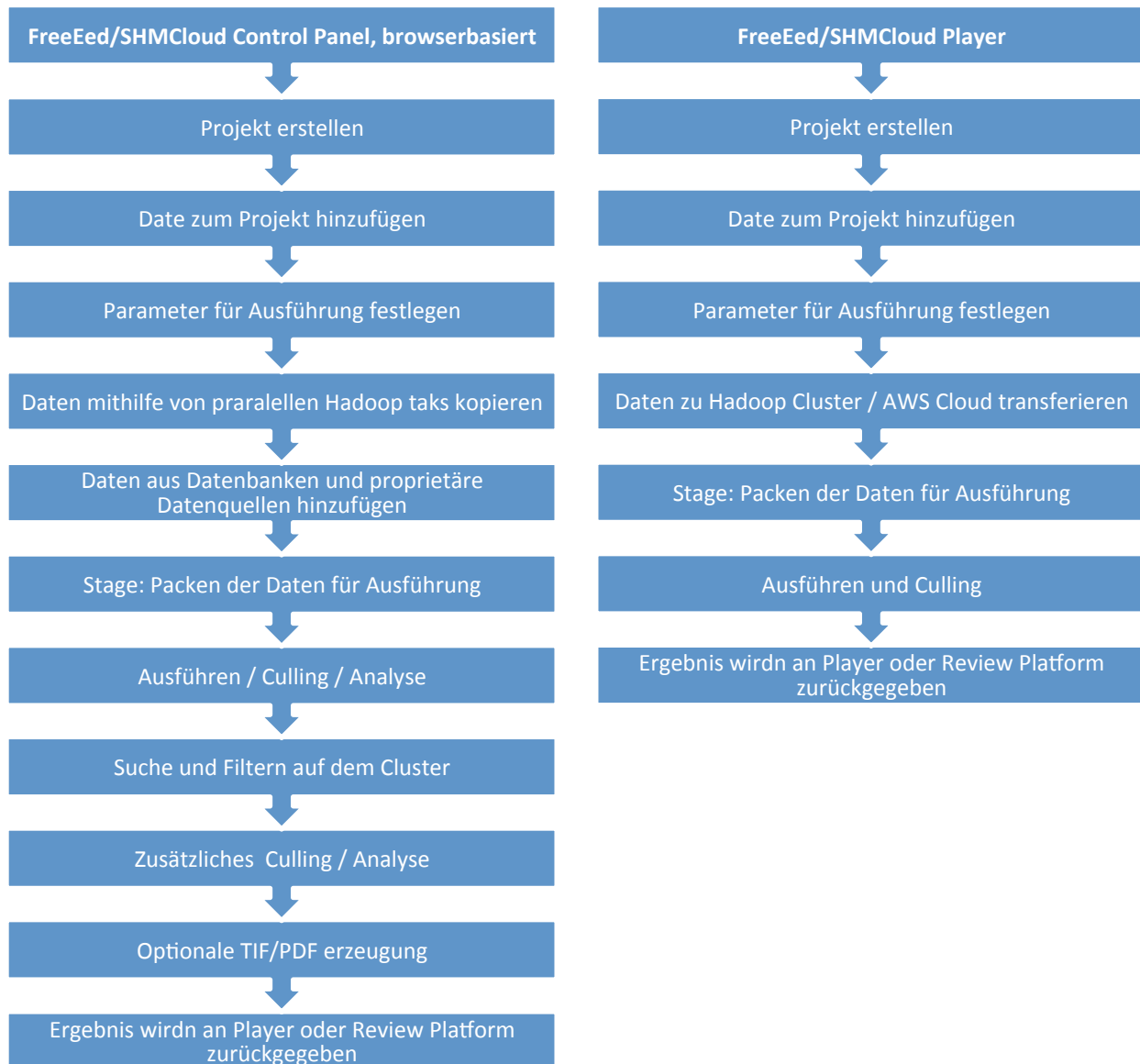


Abbildung 18: FreeEed/SHMCloud Prozedur im Vergleich Zukunft und Aktuell [105]

Am 19.12.2012 gab SHMsoft bekannt, dass es zukünftig drei Editionen geben wird. Die Open Source Lösung FreeEed unter der Apache Lizenz als Single Node Lösung und zwei Closed Source Produkte. Die Closed Source Produkte unterscheiden sich zum einen im Preis und der gewünschten Ausführungsart. Bei der Ausführungsart kann man zwischen der Amazon Cloud und dem eigenen Hadoop Cluster wählen. Neben diesen drei Versionen soll weiterhin eine Open Source Lösung entstehen, die ebenfalls ohne Lizenzkosten auf einem Hadoop Cluster lauffähig sein soll. Dafür wird die Open Source Cluster Lösung einige Limitierungen im Funktionsumfang haben. Konkrete Pläne zur Open Source Cluster Lösung gibt es derzeit noch nicht. Vorerst wird sich die Entwicklung auf die folgenden drei Editionen in Tabelle 4 konzentrieren.

Edition/Feature	FreeEed	SHMCloud for Amazon AWS	SHMCloud for Hadoop cluster
License	Apache	Closed Source	Closed Source
Player (desktop application) for local one-workstation processing	Free	Free	Free
Player app to control cluster processing	It works, but you do it yourself	Enterprise support	Enterprise support
Levels of support	Email, community	Training, implementation, and support, 8 through 5, or 24x7	Training, implementation, and support, 8 through 5, or 24x7
Pricing	Free	\$400/month+\$5 / server instance hour	Yearly: \$2,500 per node on Hadoop cluster
Text and metadata extraction, culling, load file	Yes	Yes	Yes
OCR	No	Yes	Yes
Imaging	No	Yes	Yes
Deduplication	No on Windows, yes on Linux	Yes	Yes
Speed	2 GB / hour	2 GB / hour * number of machines in the cluster (which is limited only by your AWS account)	2 GB / hour * number of machines in the cluster
Formats: MS Office, PST, PDF, images	Yes	Yes	Yes
Custom formats	No	Optional	Optional
Databases	No	Optional	Optional
Integration support	No	Available	Available
Training	No	Available	Available
Scalability	Limited by one workstation, or by your own support on the cluster	Basically, unlimited - you only need to increase your maximum number of instances assigned by Amazon	Like any Hadoop cluster, depends on your hardware

Tabelle 4: Editionsübersicht FreeEed/SHMCloud [105]

Um FreeEed/SHMCloud eine Reviewplattform zur Verfügung zu stellen gibt es unterschiedliche Überlegungen. Zum einen könnte man sich eine Zusammenarbeit mit der australischen Firma Sintelix¹¹ vorstellen, die an einem sehr leistungsfähigem Such- und Analysesystem arbeitet. Eine andere Möglichkeit sieht man in der Open Source Email Archivierung und eDiscovery Software namens Enkive¹². Mithilfe von Enkive ist es möglich über ein Webfrontend E-Mail Archiv nach verschiedenen Parametern zu durchsuchen, wie man dies in Abbildung 19 sehen kann.

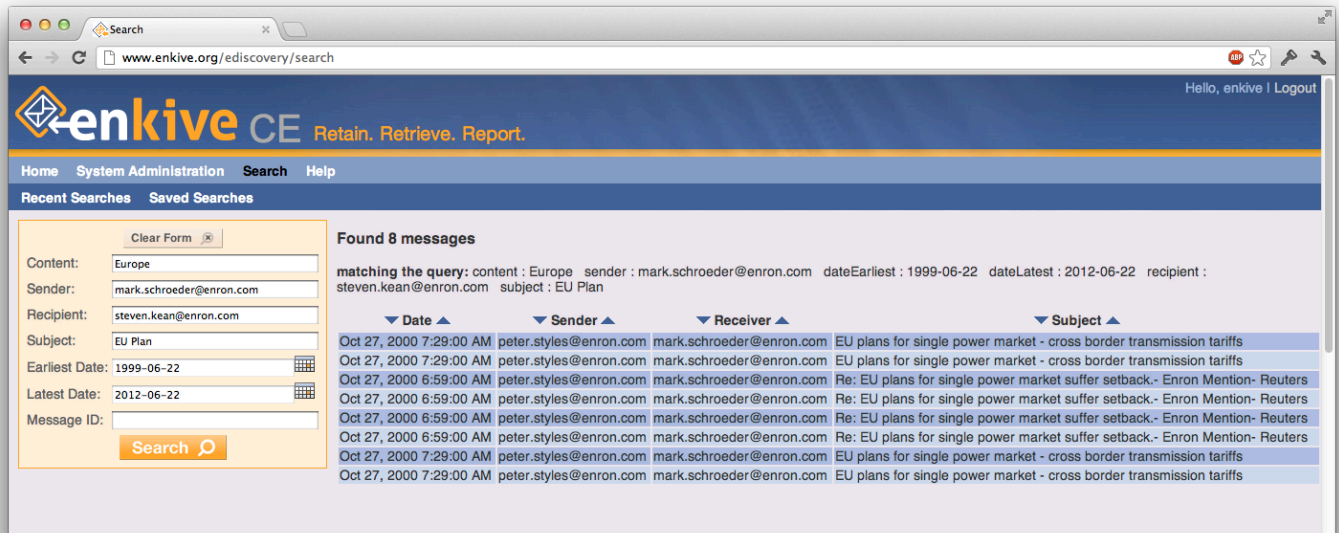


Abbildung 19: Enkive Webfrontend

4 Vergleich

In diesem Kapitel wird die Cloud basierte eDiscovery Software FreeEed/SHMCloud mit einer etablierten eDiscovery Lösung verglichen. Die Suche nach einem etablierten eDiscovery Produkt war komplexer als zuerst gedacht, da viele eDiscovery Anbieter keine Trail- oder Demoversionen anbieten oder die Lösung nur in Verbindung mit einer Appliance erhältlich sind, wie beispielsweise bei Symantec/Clearwell das der Fall ist. Bei vielen Anbietern ist die einzige Möglichkeit für eine Produkt Demo eine Cisco WebEx Vorführung. Für die Suche nach einer Testversion wurde das Gartner Magic Quadrant for E-Discovery als Übersicht der Anbieter herangezogen. Letztendlich wurden zwei Anbieter gefunden: LexisNexis und Nuix. Nuix Enterprise Workstation 4 ist für den Vergleich am sinnvollsten, da LexisNexis für jeden Verarbeitungsschritt eine eigene Softwarekomponente benötigt, wie bereits in Abbildung 6 erläutert.

Leider war der Vergleich zwischen Nuix und FreeEed auf einem lokalen Hadoop Cluster nicht möglich, da FreeEed zum aktuellen Zeitpunkt diese Funktion nicht unterstützt hat (siehe Anhang 4 – FreeEed). Somit wurden einige der Tests mittels SHMCloud auf der Amazon AWS Cloud durchgeführt.

¹¹ <http://www.sintelix.com>

¹² <http://www.enkive.org>

Dieses Kapitel enthält lediglich eine Beschreibung des Vergleichs. Die genaue Bewertung und weitere Bewertungskriterien folgen im Unterkapitel 4.1 Bewertung.

Folgende Punkte wurden bei diesem Vergleich berücksichtigt:

- Dauer der Datenanalyse auf Workstation
- Hardwareauslastung
- Analyse in der Cloud (SHMCloud)
- Unterstützte Dateiformate

Laut EMC sind E-Mails und deren Attachments die am meisten angeforderten Medien bei Gerichtsverfahren und behördlichen Untersuchungen.

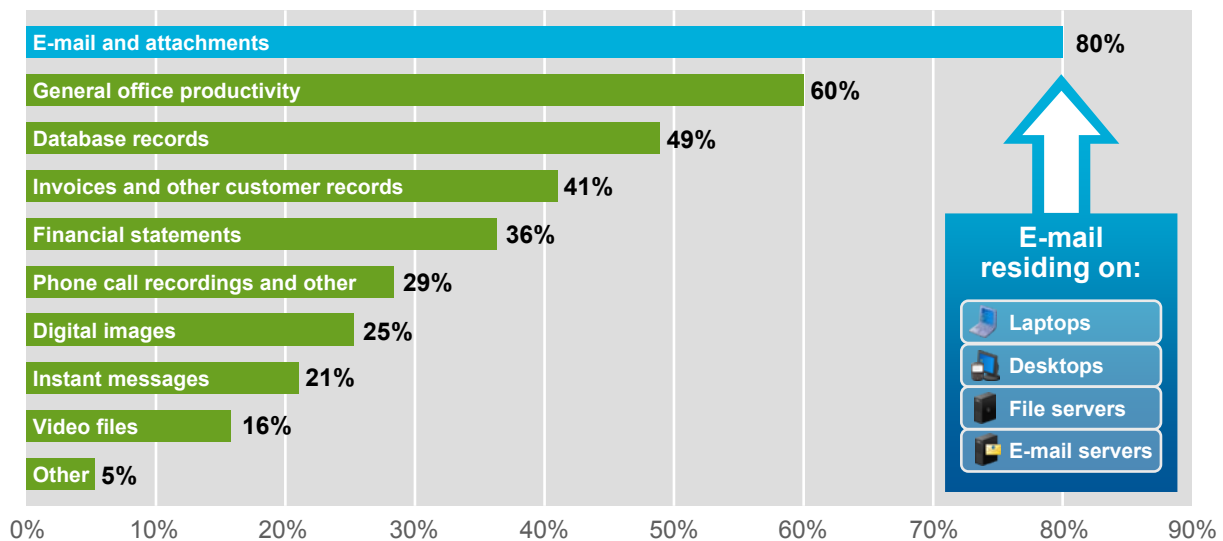


Abbildung 20: Die am meisten angeforderte Medien bei Gerichtsverfahren [108]

Daher wurde als Testdatensatz ein 16,36 GB großer Enron E-Mail Datensatz samt Attachments verwendet. Der Enron E-Mail Datensatz stammt von dem CALO Projekt der Carnegie Mellon Universität und enthält die E-Mail Daten von mehr als 150 Enron Angestellten, meist aus dem höheren Management. Die Daten wurden im Rahmen des Enron Prozesses von der Federal Energy Regulatory Commission offengelegt und stellen ein einzigartiges großvolumiges realistisches E-Mail Archiv dar. [109]

Enron war ein Energiekonzern mit Firmensitz in Houston, Texas und der siebtgrößte Konzern in den USA. Der Konzern beschäftigte etwa 22 000 Mitarbeiter. Im Jahr 2001 ereignete sich durch Bilanzfälschungen der größte Unternehmensskandal und die größte Unternehmenspleite in der US-Geschichte. Aus diesem Skandal entwickelte sich auch der Sarbanes-Oxley Act. [110]

In der folgenden Tabelle befinden sich alle Details zum Datenanalysetest:

	Nuix	SHMCloud
Hardware	Intel Core 2 CPU 1.86GHz, 4 GB, 2TB SATA 5400 rpm, 64MB Cache	Intel Core 2 CPU 1.86GHz, 4 GB, 2TB SATA 5400 rpm, 64MB Cache
OS	Windows 7	Windows 7
Anzahl der Knoten	1	1
Datengröße im ZIP Format	16.36 GB	16.36 GB
Start und Ende	17.09.2012 13:12 18.09.2012 01:29	Stage: 18.09.2012 11:03 18.09.2012 11:25 Ausführung: 18.09.2012 11:32 18.09.2012 18:04
Bearbeitungsdauer gesamt	12:16	06:54

Tabelle 5: Datenanalyse NUX vs. SHMCloud

Während des Ausführens der eDiscovery Analyse wurden die Maschinen mithilfe der Netzwerkmonitoring Software PRTG Network Monitor über Windows Management Instrumentation (WMI) und Simple Network Management (SNMP) überwacht. Damit ist es möglich zu sehen wie die Hardware während des Prozesses beansprucht wird.

Hardwaremonitoring Nuix:

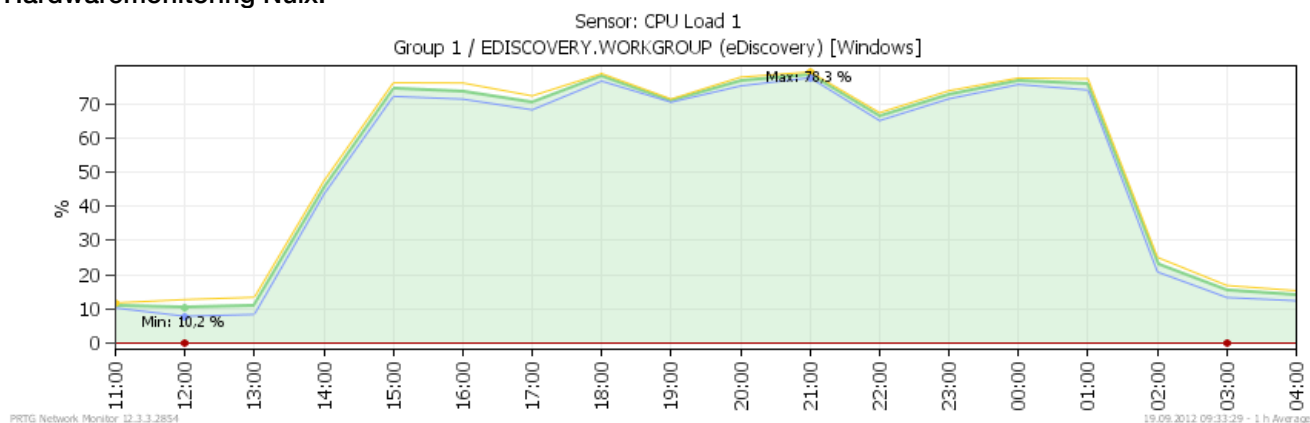


Abbildung 21: Nuix CPU

Downtime (%)
Total (%)
Processor 1 (%)
Processor 2 (%)



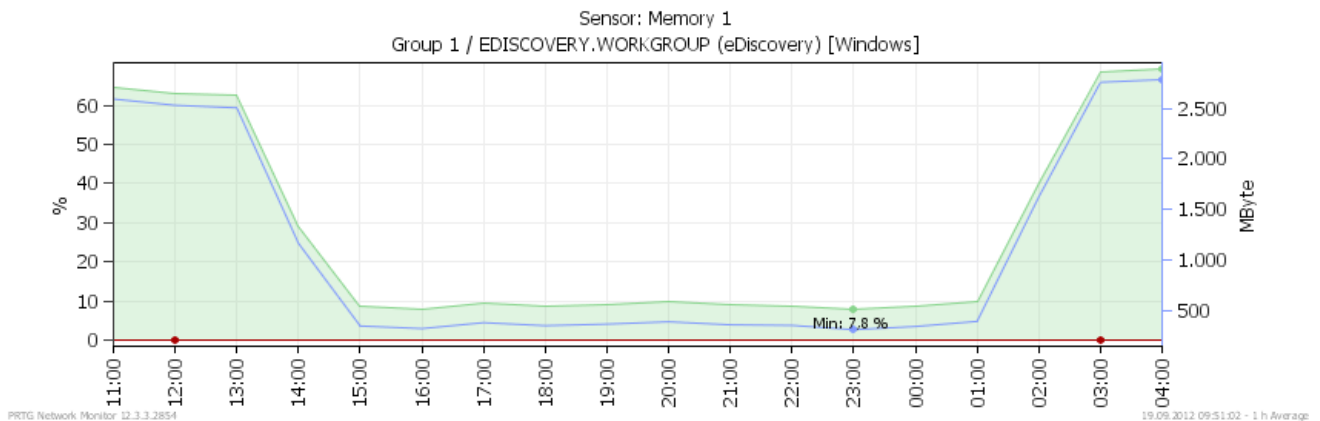


Abbildung 22: Nuix Memory

Downtime (%)
Percent Available Memory (%)
Available Memory (MByte)

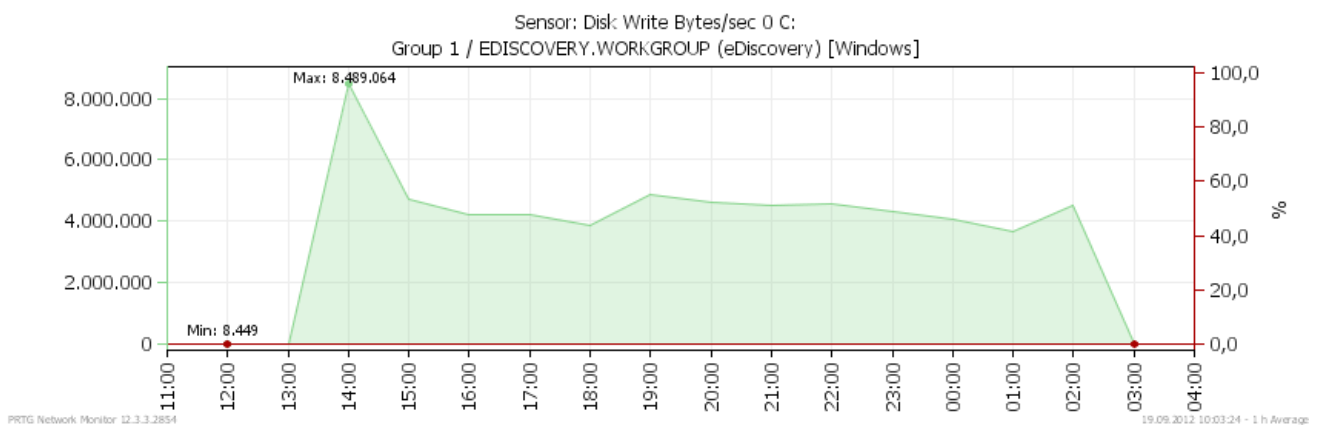


Abbildung 23: Nuix Disk Write Bytes/sec

Downtime (%)
Disk Write Bytes/sec

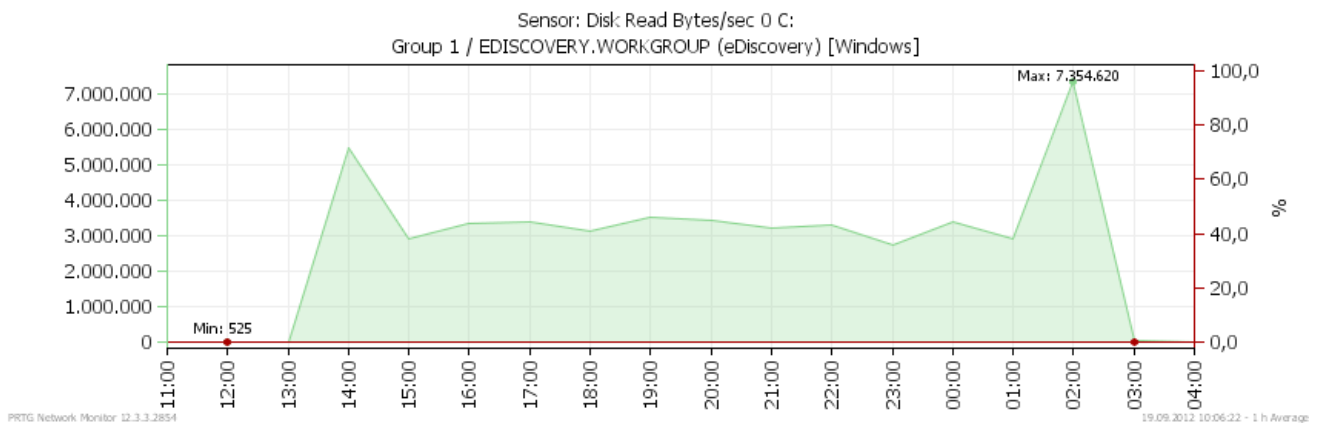


Abbildung 24: Nuix Disk Read Bytes/sec

Downtime (%)
Disk Read Bytes/sec

Hardwaremonitoring SHMCloud:

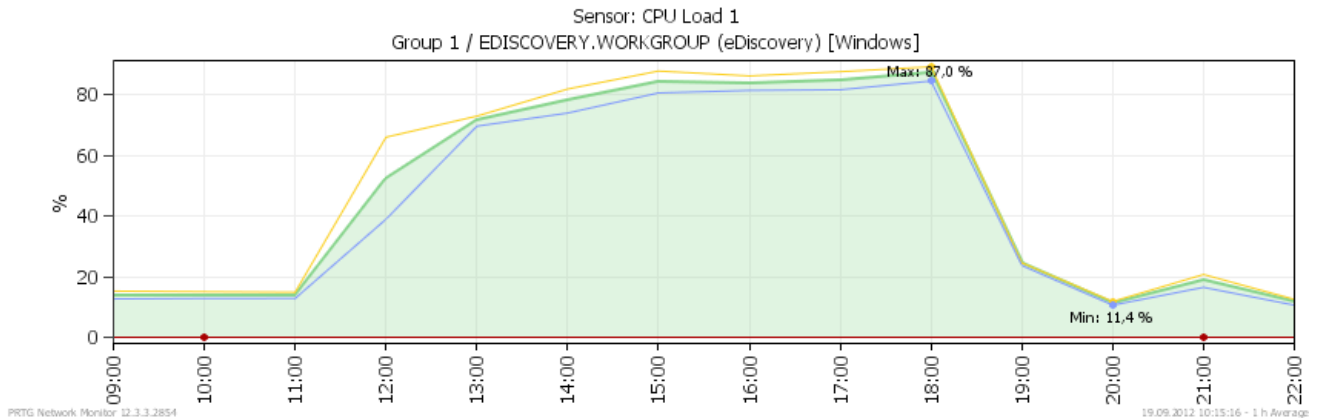


Abbildung 25: SHMCloud CPU

Downtime (%)
Total (%)
Processor 1 (%)
Processor 2 (%)

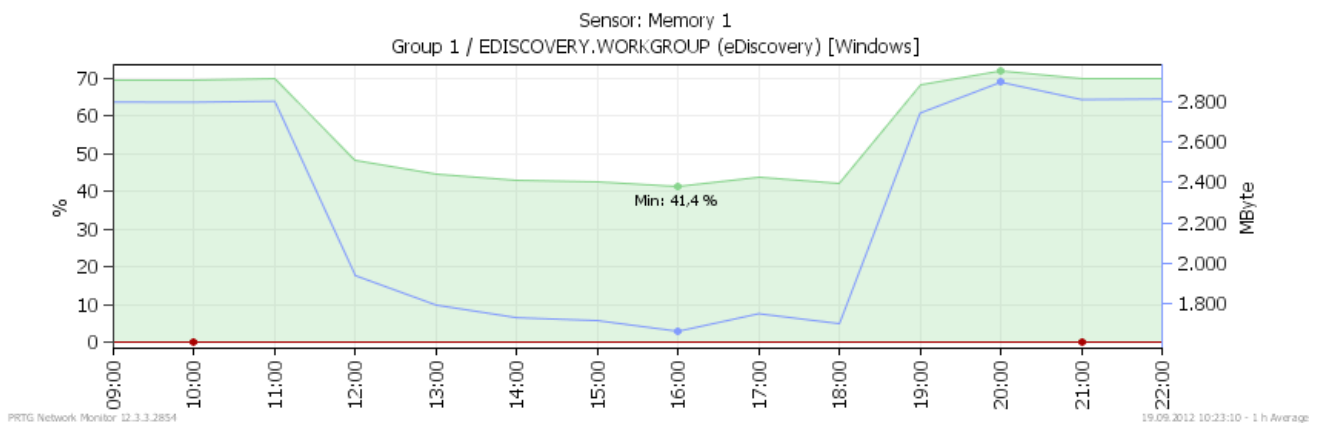


Abbildung 26: SHMCloud Memory

Downtime (%)
Percent Available Memory (%)
Available Memory (MByte)

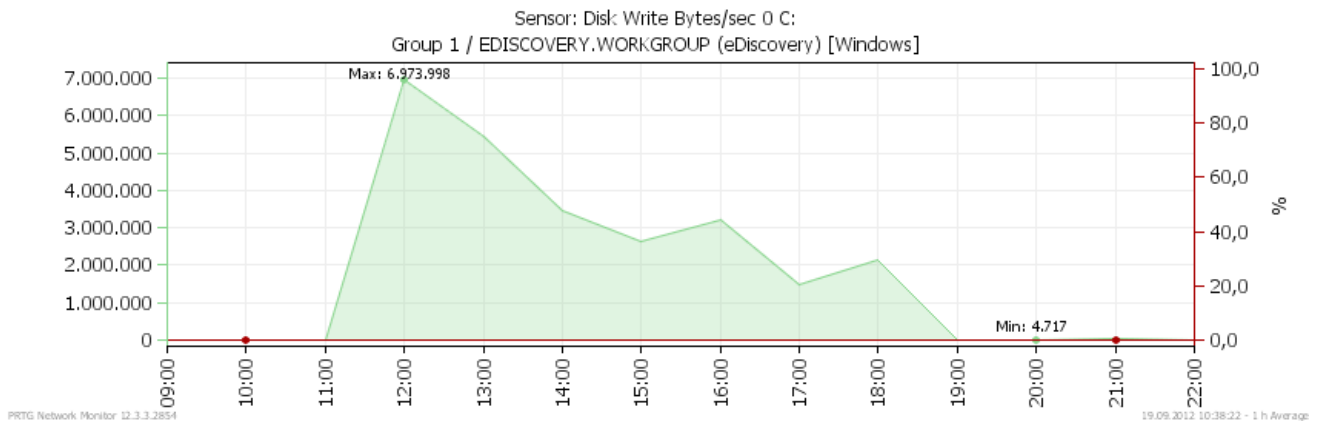


Abbildung 27: SHMCloud Disk Write Bytes/sec

Downtime (%)
Disk Write Bytes/sec

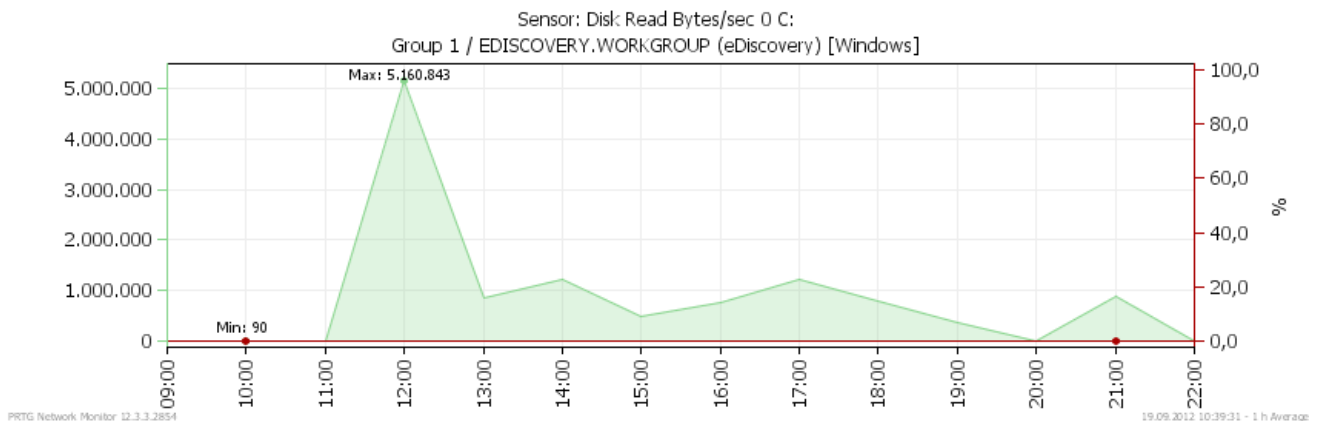


Abbildung 28: SHMCloud Disk Read Bytes/sec

Downtime (%)
Disk Read Bytes/sec

Analyse in der Cloud (SHMCloud)

Auch für den Test in der Amazon Cloud (siehe Anhang 5 – SHMCloud) wurde das 16,36 GB große Enron Testdatenset verwendet. Dieser Test wurde zunächst auf einem Cluster bestehend aus 10 Knoten durchgeführt und danach auf einem einzelnen Knoten um auch hier wieder einen Zeitvergleich anstellen zu können (siehe Tabelle 6 und Anhang 7 - Übersicht MapReduce Jobs). Als Instanztyp wurde der High-CPU Extra Large Instance Type verwendet kurz auch als c1.xlarge bekannt. Dieser Instanztyp hat folgende Daten:

„7 GB of memory

20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each)

1690 GB of instance storage

64-bit platform

I/O Performance: High“ [111]

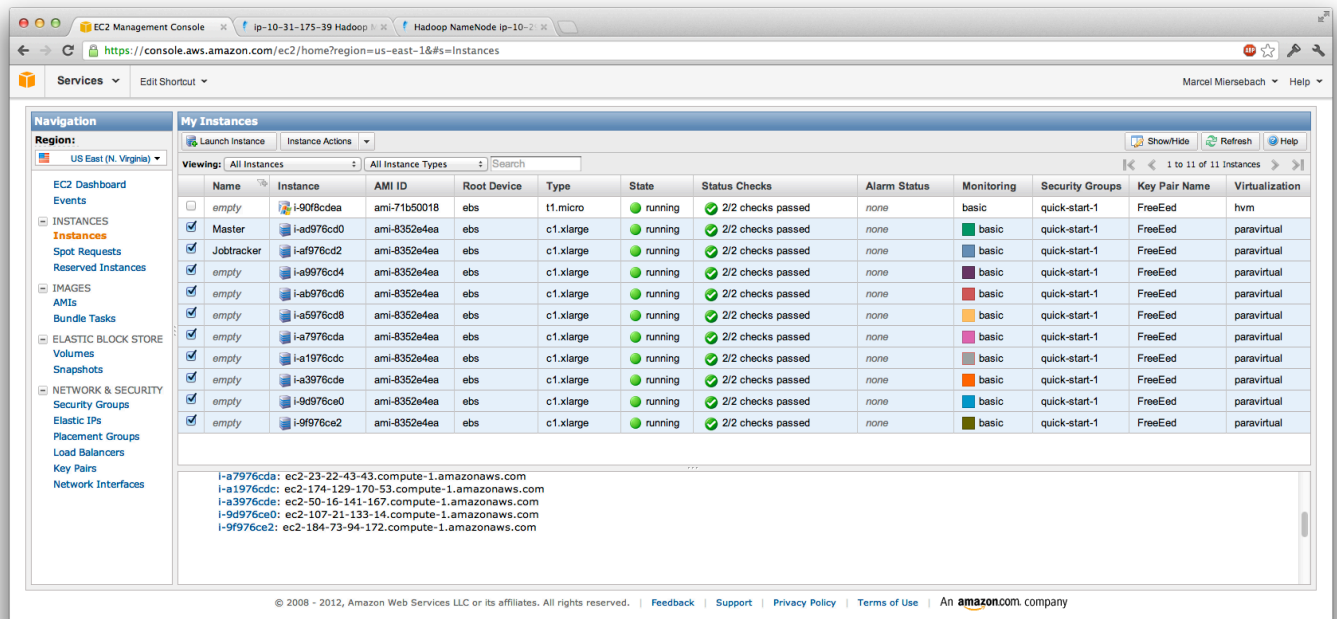


Abbildung 29: Amazon EC2 Instanzen für SHMCloud

Die Bearbeitungsdauer mit einem Knoten von 6:41 Stunden war ähnlich wie auf der lokalen Windows 7 Maschine mit 06:54. Die deutliche Erhöhung der Knotenanzahl hat die Bearbeitungsdauer stark reduzieren können. Ursprünglich wurde eine noch kürzere Bearbeitungsdauer erwartet.

	10 Knoten	1 Knoten
Datengröße im ZIP Format	16.36 GB	16.36 GB
OS	Ubuntu 12.04 + Hadoop	Ubuntu 12.04 + Hadoop
Hardware	c1.xlarge	c1.xlarge
Bearbeitungsdauer	2:27	6:41

Tabelle 6: Bearbeitungsdauer SHMCloud in der Amazon Cloud

Die Abbildung 30 zeigt die CPU Auslastung aller 10 Knoten über 3 Stunden hinweg. Bei Knoten i-ad076cd0 (grün) handelt es sich um den Master Knoten, i-af976cd2 (blau) fungiert als Jobtracker. Dieses Diagramm zeigt, dass die CPU Auslastung auf dem Masterknoten im Gegensatz zu den anderen Knoten sehr gering ist.

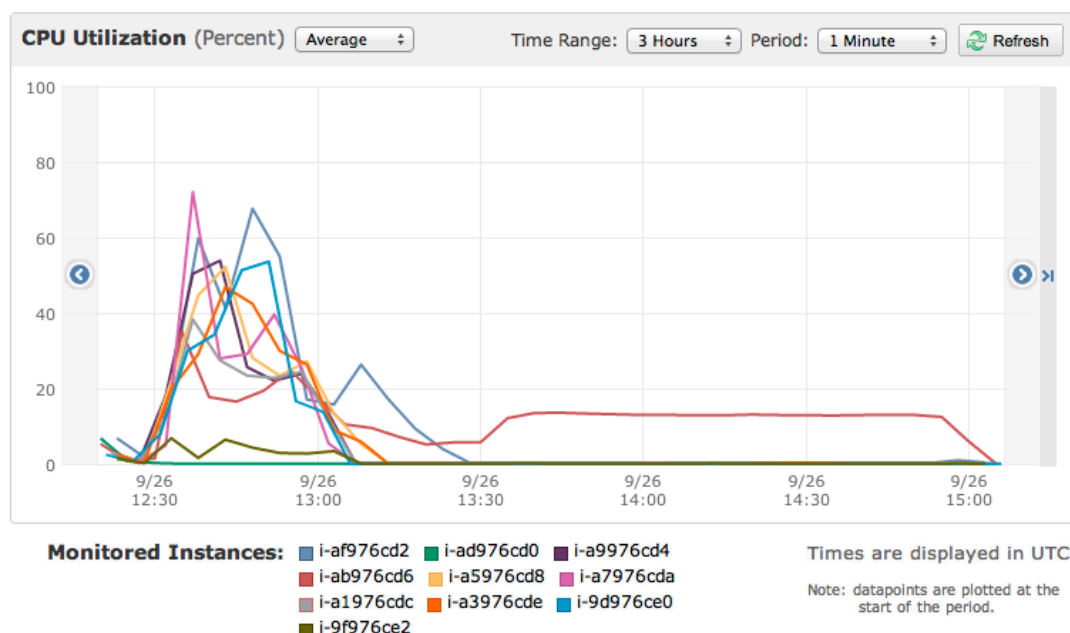


Abbildung 30: CPU Auslastung Amazon EC2

Unterstützte Dateiformate

Für die Evaluierung der unterstützten Dateiformate wurde das EDM File Formats Data Set 1.0.1¹³ herangezogen. Dieses Data Set beinhaltet 381 Dateien mit 200 Dateiformaten. Zusätzlich wurden einige weitere relevante Testdaten hinzugefügt. Bei diesem Test wurde geprüft, aus wievielen Dateien Text extrahiert werden kann. Zu beachten ist jedoch, dass in einigen Dateien kein Text vorhanden ist, da es sich z.B. um reine Bild- oder Grafikdateien handelt. Eine Aufstellung der Resultate findet man im Anhang 6 - Übersicht Fileformattest.

	Nuix	SHMCloud
Dateianzahl	418	418
Text extrahiert	344	89
Kein Text extrahiert	74	329

Tabelle 7: Erkannte Fileformate

Unterstützung von Fremdsprachen

Da bei multinationalen Konzernen der Schriftverkehr in verschiedenen Sprachen abläuft, kommt der Behandlung von Fremdsprachen eine wichtige Rolle bei eDiscovery Software zu. Besonders interessant sind Sprachen, die kein Lateinisches Alphabet verwenden. Für diesen Test wurde das EDM Internationalization Data Set verwendet.¹⁴ Bei diesem 724 MB großen Data Set handelt es sich um einen Snapshot von ausgewählten Ubuntu localization mailing list archives in 23 Sprachen.

¹³ <http://www.edrm.net/resources/data-sets/edrm-file-format-data-set>

¹⁴ <http://www.edrm.net/resources/data-sets/edrm-internationalization-data-set>

4.1 Bewertung

In diesem Kapitel werden nun die im vorherigen Kapitel gesammelten Vergleichsdaten analysiert und bewertet. Was zu Beginn am deutlichsten hervorsticht, dass Nuix für die Bearbeitung des 16.36 GB großen Daten Sets fast doppelt so lange gebraucht hat wie SHMCloud. Wenn man jedoch die Anzahl der Dokumente betrachtet aus denen Text extrahiert werden konnte, wird schnell klar weshalb SHMCloud wesentlich weniger Zeit benötigte. SHMCloud konnte von den 418 Dateien lediglich aus 89 Dateien Text extrahieren.

	Nuix	SHMCloud
Bearbeitungsdauer gesamt	12:16	06:54
Text Extrahiert	344	89

Tabelle 8: Überblick Bearbeitungsdauer und erkannte Fileformate

Vergleicht man die Diagramme aus dem Hardware monitoring, fällt auf, dass SHMCloud die CPU mehr als Nuix beansprucht. Nuix wiederum verwendet bei der Analyse sehr viel Arbeitsspeicher. Die Lese- und Schreibaktivitäten sind bei Nuix ebenfalls höher als bei SHMCloud. Dies lässt sich ebenfalls dadurch erklären, dass Nuix viel mehr Text extrahieren konnte, und somit mehr Content zum Lesen und Schreiben hatte. Leider konnte SHMCloud nur Dokumente mit Lateinischen Buchstaben verarbeiten. Bei Fremdsprachen mit einem anderen Alphabet wie z.B. Griechisch, Russisch usw. wurden keine Ergebnisse erzielt. Bei Nuix hingegen konnte man für das Datenset das entsprechende Source Encoding angeben. Positiv ist auch aufgefallen, dass man die Zeitzone der Source Daten einstellen kann, wie man dies in Abbildung 31 sehen kann. Sowohl bei Nuix als auch bei SHMCloud wäre eine automatische Erkennung für das Encoding sinnvoll, da bei einem Mail Archiv, das verschiedenst codierte Nachrichten beinhaltet, gleichzeitig nach dem englischen und z.B. griechischen oder russischen Stichwort gesucht werden kann. Bei der aktuellen Lösung muss ein Evidence für jede entsprechende Sprache geöffnet werden und die komplette Verarbeitung einzeln durchgeführt werden.

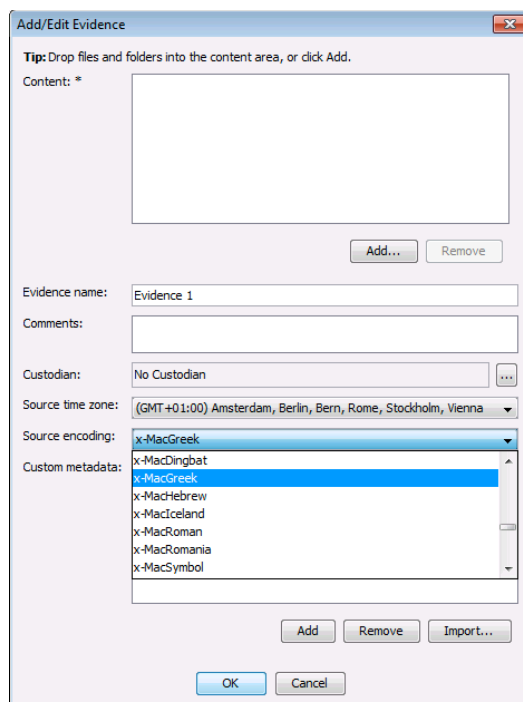


Abbildung 31: Nuix Encoding und Zeitzone

Sowohl SHMCloud als auch Nuix verfügen über eine OCR Funktion. Nuix hatte Probleme mit einigen schief oder unscharf gescannten Dokumenten. SHMCloud hingegen kam mit diesen Dokumenten wesentlich besser zurecht. SHMCloud verwendet die Open Source OCR engine von Tesseract. Tesseract war 1995 auf Platz drei beim UNLV accuracy test. Seitdem hat Google die engine noch weiter verbessert. Die Tesseract OCR engine unterstützt viele unterschiedliche Bildformate und mehr als 40 Sprachen. [112]

Die Duplikaterkennung funktionierte bei beiden Produkten problemlos, wobei SHMCloud die Duplikaterkennung nur auf Linux und Mac OS X unterstützt. Die Windows Version kann dies aufgrund von fehlenden Bibliotheken nicht. Beide Produkte verwenden für die Deduplikation MD5, SHMCloud kann wahlweise auch SHA-1 verwenden. Man muss sich natürlich im Klaren sein, dass SHMCloud kein fertiges Produkt ist und noch in der Entwicklungsphase steckt. Daher ist es auch nicht unbedingt verwunderlich, dass noch Teile der Software fehlen wie z.B. die wichtige Reviewplattform. Trotzdem soll dies nicht außer Acht gelassen werden. Es soll in die Bewertung nicht nur der aktuelle IST Zustand der Software bewertet werden sondern auch Augenmerk auf Eigenschaften wie die Erweiterbarkeit gelegt werden.

Zunächst ist das sehr gelungene Graphical user interface (GUI) von Nuix zu erwähnen. Es bietet sehr viele Funktionen, auch in Bezug auf Collaboration mit mehreren Benutzern und ist einfach und intuitiv zu bedienen. Darüber hinaus bietet Nuix eine sehr gute Reviewplattform an, die sehr viele Möglichkeiten zur Suche und Filterung hat. Zusätzlich gibt es Funktionen wie Word Lists, Metadaten View, Native View, PDF View, Diff View und Tree View für die einzelnen Dateien. Nuix verfügt auch über einen gut realisierten Result View, so können die Ergebnisse statistisch, auf einer Event Map oder einer Network Map sehr schön dargestellt werden. Diese Statistiken und Maps sind später für den Export und die Präsentation vor Gericht äußerst sinnvoll. Nuix unterstützt auch den Export der Ergebnisse für viele unterschiedliche Systeme, darunter das standardisierte EDRM XML Format. SHMCloud besitzt bisher keine Reviewplattform. Die Ergebnisse werden lediglich in einer CSV Datei mit allen Metadaten Informationen gesammelt. Zusätzlich werden die nativen Dateien und die extrahierten Texte in ZIP Dateien abgelegt. Hierbei handelt es sich um eine sehr simple Darstellung der Resultate und kann bei weitem nicht mit dem enormen Funktionsumfang von Nuix Enterprise Workstation 4 mithalten. Äußerst positiv an SHMCloud ist zunächst einmal die einfach gehaltene Benutzeroberfläche. Man findet sich sehr schnell und intuitiv zurecht, ansonsten gibt es auch wie bei Nuix eine entsprechende Dokumentation. SHMCloud bietet durch die Solr, Lucene Suchfunktion ein sehr mächtiges Suchwerkzeug. Hier können entsprechend gute Filter- und Suchanfragen zusammengestellt werden, jedoch kann eine solche Such- bzw. Filterfunktion auch schnell in einer sehr komplexen und schwer verständlichen Zeichenkette enden. An dieser Stelle wäre z.B. eine Standardsuchoption mit entsprechender GUI Unterstützung für nicht bewandte Benutzer sinnvoll. Parallel sollte trotzdem eine manuelle Sucheingabe in einem Expertenmodus möglich sein, um komplexe Suchfunktionen selbst zu schreiben. Ein großer Vorteil von SHMCloud ist die Plattformunabhängigkeit, da das Projekt in Java programmiert wurde, läuft es auf den unterschiedlichsten Plattformen. Noch flexibler wird man zukünftig sein, wenn der SHMCloud Player über ein Webfrontend zu bedienen ist. Dann kann man den eDiscovery Prozess in der Cloud bzw. auf dem Cluster ganz einfach von einem Tablet oder Smartphone steuern. SHMCloud setzt auf sehr interessante Open Source Projekte, die äußerst mächtig sind und unheimlich viele Funktionen bieten. Projekte wie Hadoop, Solr, Lucene und Tika sind für große Datenmengen optimiert und bereits erfolgreich im Big Data Sektor. Mit diesen Technologien im Rücken sollte sehr viel möglich sein und das Projekt auch sehr gut erweiterbar sein. Momentan werden nicht sehr viele verschiedene Dateiformate und fremdcodierte Sprachen unterstützt. Jedoch ist dies durch die offene Architektur und die sehr schnell und gut weiterentwickelten Open Source Projekte, auf denen SHMCloud aufbaut, relativ einfach möglich weitere Dateiformate und Sprachen zu implementieren. Laut SHMSoft können neue Dateiformate relativ einfach über zusätzliche Bibliotheken implementiert werden. Sollte Tika das entsprechende Format unterstützen, sind nur sehr kleine oder sogar gar keine Änderungen notwendig. Im Test war es nicht möglich fremdcodierte Texte mit SHMcloud zu extrahieren. Dies sollte jedoch möglich sein, da alle verwendeten Softwarekomponenten die Unterstützung für fremdcodierte Sprachen mitbringen. Vermutlich ist hier noch ein wenig Entwicklungsarbeit und Bug fixing nötig. Grundsätzlich ist der Culling Prozess, welcher mithilfe von Lucene durchgeführt wird, Sprachsensitiv. Weder Nuix noch SHMCloud verwendet Benutzerrollen oder ein Rechtesystem mit dem man den Zugriff für bestimmte Benutzer einschränken kann. Auch die Anbindung an ein LDAP oder Active Directory System ist nicht möglich. Beide Applikationen sind 64 Bit fähig. Wünschenswert für beide Produkte wäre die Funktion für ein Litigation Hold System und ein Collection System, das die Daten auf

Clients, Mobilien Geräten und Servern zusammensammelt. Die Nuix Reviewplattform bietet im Gegensatz zur unfertigen SHMCloud Lösung, Vorteile wie Funktionen für In-document hit highlighting oder eine Anzeige für die Trefferzahl eines Suchbegriffs. Vermutlich werden solche Funktionen später mit der SHMCloud Reviewplattform mit implementiert. SHMCloud steht in drei verschiedenen Editionen zur Verfügung und beginnt mit einer Gratis Open Source Version und geht bis zur eigenen Cluster Edition. Somit sollte jedes Budget abgedeckt sein. SHMCloud kann theoretisch mit ca. 300 bis 400 Knoten umgehen, um mehrere Terabyte große Daten zu analysieren. Genaue Limitierungen sind jedoch nicht bekannt und konnten auch nicht evaluiert werden. Nuix ist für Cases mit einem Datenvolumen von ca. 750 GB geeignet. Das Ergebnis dieser Bewertung ist in der Tabelle 9 nochmals übersichtlich zusammengefasst.

	Nuix	SHMCloud
Analyse Dauer	○	●
Unterstützte Formate	●	○
Plattformabhängigkeit	-	✓
Cloud/Cluster fähig	-	●
Fremdsprachen Encoding	◐	-
OCR	○	●
Duplikaterkennung	●	●
Reviewfunktion	●	-
Dokumentation	●	●
Bedienkomfort	●	◐
Interface GUI	●	●
Collaboration	◐	k. A.
Suchoptionen	●	●
Präsentation der Resultate	●	○
Exportfunktion	●	◐
Erweiterbarkeit	k. A.	●
Benutzerrollen und Rechte	-	-
Active Directory, LDAP	-	-
64-bit fähig	✓	✓
Litigation Hold System	-	-
Collection System	-	-
E-Mail Felder Auswertung	✓	✓
In-document hit highlighting	✓	-
keyword counts and summaries	✓	-
Text Extraction	✓	✓
Max. Case Größe in GB	750 GB	k. A.
Preis	k. A.	●

✓ vorhanden - nicht vorhanden k. A. Keine Angabe
 ○ schlecht ◐ zufriedenstellend ● sehr gut

Tabelle 9: Ergebnis des Vergleichs

5 Ausblick

Die Entwicklung der Informationstechnologie tendiert zu weltweit verteilten Cloud Anwendungen. Daraus resultiert eine gewollte oder ungewollte Verbreitung von Unternehmensdaten in der Cloud, die das Daten- und Informationsproblem dramatisch verschärft. Die Cloud bzw. Cloud Technologien wie Hadoop bieten ganz neue Chancen, sowohl auf dem normalen IT-Sektor als auch für eDiscovery. Man kann sich der Cloud nur sehr schwer entziehen, da die Cloud zukünftig ein fester Bestandteil unserer IT sein wird. Firmen verlagern immer mehr Services in die Cloud und viele haben keine konkreten Pläne, wie man bei einem anstehenden eDiscovery Verfahren mit den Daten in der Cloud verfahren soll. Auch hier sind zukünftig die Hersteller der Cloudlösungen wie auch eDiscovery Hersteller gefragt, entsprechende Produkte anzubieten. Entweder man schafft die Möglichkeit den eDiscovery Prozess direkt in der Cloud durchzuführen, oder man implementiert entsprechende Exportfunktionen, mit denen man die Daten in sein lokales eDiscovery System importieren kann. Ein gutes Beispiel dafür ist Google. Sie stellen ihren Business Kunden inzwischen ein eDiscovery System für ihre Apps suite als kostenpflichtige Zusatzoption zur Verfügung. Anbieter sollten solch eine eDiscovery Funktion in Zukunft nicht erst nach einigen Monaten nachreichen, sondern von Beginn an implementieren. Gewinnen kann man mit der Cloudtechnologie nicht nur in den main business Segmenten sondern auch in Nischensegmenten wie eDiscovery. Die Cloud ist ein zusätzlicher Faktor für den rasanten Anstieg des Datenvolumens. Um dieses große Datenaufkommen zu analysieren, muss man seine äußerst sensiblen Unternehmensdaten nicht zwingend in fremde Hände geben, sondern man kann auch lediglich neue Technologien wie z.B. Hadoop für sich nutzen. Genau das versucht SHMsoft mit seiner eDiscovery Lösung zu realisieren. Mit der SHMsoft Lösung ist es möglich die Daten sowohl in der Cloud als auch auf einem eigenen Cluster zu analysieren. Bisher hat man nur den wichtigsten Teil der Software fertiggestellt und befindet sich noch mitten in der Entwicklungsphase.

Es fehlen noch wichtige Bestandteile für eine eDiscovery Software, da bis jetzt nur eine geringe Anzahl an Dateiformaten unterstützt wird und noch keine Reviewplattform zur Verfügung steht. Die aktuellen Pläne und die Roadmap sind sehr vielversprechend. Es gibt bereits interessante Ansätze und Ideen, um die Reviewplattform und ein komplettes Webinterface für die Softwaresteuerung zu ermöglichen. Die Zukunft der Open Source Cluster Lösung von FreeEed/SHMCloud wird besonders im Hinblick auf ihre Limitierung im Vergleich zur kommerziellen Version interessant werden. Ob sich das einzige Open Source Produkt auf dem eDiscovery Markt letztendlich behaupten kann, wird sich noch herausstellen. Bei SHMsoft setzt man jedenfalls auf eine sehr gute Mischung von Open Source Projekten, die sich im Big Data Bereich bereits bewährt haben und als zukunftsweisend einzustufen sind. Da es sich um ein Open Source Projekt handelt, bietet es die Möglichkeit sich selbst in diesem Projekt mit interessanten neuen Features einzubringen, wie beispielsweise mit der folgenden Funktion: Zum aktuellen Zeitpunkt gibt es keine eDiscovery Software, welche eine Search by Image Funktion bereitstellt. Diese Funktion ermöglicht es Bilder oder Grafiken anhand ihrer Ähnlichkeit zu identifizieren. Denkbar wäre beispielsweise eine eingescannte Unterschrift als Suchparameter einzufügen, um nach allen Dokumenten zu suchen, die mit dieser Unterschrift signiert wurden oder auch nach allen Dokumenten mit einer entsprechenden Grafik. Google stellt solch einen Algorithmus bereits für seine Suchmaschine zur Verfügung.

Es gibt auch einfacher zu realisierende Features, welche die eDiscovery Software noch leistungsfähiger machen würden wie z.B. die Möglichkeit eine Passwortliste für verschlüsselte Dateien oder Archive zu hinterlegen. Kommen für eine passwortgeschützte ZIP Datei 20 verschiedene Passwörter in Frage, so kann man diese hinterlegen und die eDiscovery Software versucht automatisch diese ZIP Datei mit den hinterlegten Passwörtern zu entpacken. So ist kein manueller Eingriff mehr nötig. Eine weitere Verbesserung sowohl für Nuix als auch für SHMCloud wäre die Funktion für eine Fuzzy Search (Unschärfe Suche). Damit werden bei einer Stichwortsuche auch ähnlich geschriebene Zeichenketten erfasst, in denen sich ein Schreibfehler eingeschlichen hat, ein Synonym oder eine Abkürzung verwendet wurde.

Man kann das Problem der steigenden Datenmengen nicht nur durch die Rechenpower von unzähligen Maschinen in der Cloud bekämpfen sondern auch mit Technologien wie Predictive coding und Machine learning, die bereits bei AntiSpam Software zum Einsatz kommen. Predictive coding und Machine learning kann als Bestandteil für eine intelligente eDiscovery Software genutzt werden. Um die enormen Datenmengen in der Zukunft zu handeln, reicht eine simple Such- und Analysesoftware nicht mehr aus. Es bleibt daher spannend, wie

sich die Zukunft des jungen eDiscovery Marktes weiter entwickelt, da die eDiscovery Anbieter ständig neue Lösungen entwickeln. Mit großer Spannung kann auch die Zukunft von SHMSoft verfolgt werden. Offen bleibt, ob sich dieses Projekt in der eDiscovery Branche, vor Gericht und in multinationalen Unternehmen behaupten kann oder ob das Projekt von einem der großen eDiscovery Anbieter aufgekauft wird.

Nicht nur auf dem eDiscovery Markt wird sich zukünftig viel verändern, auch die Cloud Technologien wie Hadoop werden sich weiterentwickeln müssen. Besonders im Bereich der Security hat gerade Hadoop sehr viel Nachholbedarf. Anfänglich wurden Hadoop Cluster von einer kleinen Gruppe von Menschen betrieben, die sich gegenseitig vertrauten, daher wurden keine wirklichen Sicherheitsmechanismen implementiert. Inzwischen wurde ein Permission Control System auf Basis von Active Directory mittels Kerberos nachgerüstet. Zahlreiche weitere Security Verbesserungen wurden vorgeschlagen aber nicht implementiert. Bei anderen Big Data Tools sieht es nicht besser aus. Cassandra und MongoDB verfügen über wenige Sicherheitsmechanismen und sind somit z.B. anfällig für SQL Injections. In dieser Hinsicht sollte zukünftig noch einiges passieren, damit unsere Daten sicher sind, egal ob für eDiscovery oder andere Big Data Anwendungen.[113] [114] [115]

Die Zukunft ist nicht nur in technischer Hinsicht interessant, sondern auch im Bereich der Legislative werden die neuen Technologien wie Cloud Computing Veränderungen mit sich bringen. Für das Thema eDiscovery sind eindeutige und einfache EU Gesetze wünschenswert, damit der Datenexport für ein eDiscovery Verfahren einfacher zu realisieren wird.

6 Fazit

Wie sich in dieser Arbeit gezeigt hat, ist für multinationale Unternehmen das Thema eDiscovery schwer zu umgehen. Ohne entsprechende Prozesse und Policies ist ein eDiscovery Verfahren kaum handzuhaben. Auch die richtige Softwareauswahl ist ein entscheidender Teil für ein erfolgreiches eDiscovery Konzept. Die Firma SHMSoft wird zukünftig sowohl eine frei verfügbare Open Source Edition als auch eine Closed Source Variante anbieten. So sollte aus reiner Kostensicht für jede Unternehmensgröße und jedes Budget etwas dabei sein. Etwas enttäuschend ist, dass für diese Arbeit leider keine funktionsfähige Version von FreeEed/SHMCloud zur Verfügung stand, die auf einem lokalen Hadoop Cluster lauffähig war, wie dies zunächst von FreeEed angepriesen wurde. Genau dieser Fall kann besonders bei Open Source Software eintreten. Dies bestätigt auch eine Studie der Fraunhofer-Gesellschaft, dass durch die ständige Implementierung neuer Features der Reifegrad der Software manchmal auf der Strecke bleibt. [116] Im konkreten Fall hat man sich stark auf die Lauffähigkeit in Amazon EC2 Instanzen konzentriert und die Funktionalität für den lokalen Hadoop Cluster vernachlässigt. Das Kapitel 4 hat gezeigt, welche wichtigen Eigenschaften sowohl von Nuix als auch von SHMCloud unterstützt werden und wie gut sich die jeweilige Software bei der Analyse verhält. Anhand dieser Ergebnisse lässt sich nun folgende Forschungsfrage beantworten:

Was sind die Stärken und Schwächen von Hadoop basiertem eDiscovery im Vergleich zu bereits etablierten Lösungen?

Die größte Stärke der Hadoop basierten eDiscovery Applikation FreeEed/SHMCloud ist die flexible Erweiterbarkeit. Die Anzahl an Clusterknoten kann je nach Datenvolumen flexibel angepasst werden. Somit kann die Verarbeitungszeit für die Analyse direkt beeinflusst werden, egal ob es sich um die Amazon AWS Cloud handelt oder um einen lokalen Hadoop Cluster im Unternehmen. Letzteres ist zum aktuellen Zeitpunkt jedoch noch nicht möglich und soll erst in einer späteren Version realisiert werden. Da die FreeEed/SHMCloud Architektur auf sehr erfolgreichen Open Source Produkten aus dem Big Data Umfeld wie Hadoop, Tika, Lucene und Solr aufbaut, hat sie ein enormes Potential und lässt sich einfach mit neuen Projekten aus diesem Umfeld erweitern. Zusätzlich ist die günstige und flexible Preisgestaltung ein enormer Pluspunkt für die Software. Ein weiterer Vorteil den FreeEed/SHMCloud bietet, ist die Plattformunabhängigkeit. Zukünftig soll die gesamte Software über eine Weboberfläche steuerbar sein. So kann man den SHMCloud Player von jedem System überwachen und steuern. Die wichtigsten eDiscovery Funktionen wie Duplikaterkennung, E-Mail Auswertung und Text Extraction werden von der Open Source Software unterstützt. Die Such- und Filterfunktion sowie die OCR Schrifterkennung konnten sogar bessere Ergebnisse erzielen als die kommerzielle Lösung. Die einfache und intuitiv gestaltete GUI und eine gute Dokumentation sorgen für den entsprechenden Bedienkomfort.

Der Reifegrad wird als die größte Schwäche dieser eDiscovery Software eingestuft. Zum einen wird bisher die Ausführung auf einem lokalen Hadoop Cluster noch nicht unterstützt, auch die Anzahl der unterstützten Dateiformate ist bisweilen noch eher gering und Fremdsprachen bereiten der Software ebenfalls Probleme. Eine Reviewplattform muss unbedingt nachgereicht werden, da man bisher die Suchresultate nur in Form einer Auflistung von Dateinamen ausgegeben bekommen hat und dann die Stichworte erst in den extrahierten Textdokumenten suchen musste, um an die betreffende Stelle im Text zu gelangen. Eine andere Möglichkeit oder Zwischenlösung für dieses Problem wäre die Resultate in einem speziellen Format auszugeben wie z.B. in dem EDRM XML Interchange Format, indem die gesuchten Passagen mittels 'In-document hit highlighting' markiert werden, um dann das Ergebnis in einem anderen Reviewprogramm auszuwerten. So könnte die automatische Analyse von großen Datenmengen in kurzer Zeit in der Cloud oder auf einem lokalen Hadoop Cluster durchgeführt werden und das Ergebnis auf einer bereits fertigen Reviewplattform ausgewertet werden. Sowohl bei dem kommerziellen Produkt Nuix als auch bei FreeEed/SHMCloud werden Funktionen für die Benutzerrollen, das Zugriffsrechtssystem oder eine Active Directory, LDAP Anbindung vermisst. Wenn mehrere User auf einer Plattform arbeiten und nur einen gewissen Teilbereich sehen sollen, kann dies bei beiden Produkten nicht eingeschränkt werden. Wünschenswert wäre auch, dass beide Produkte eine Funktion mitbringen, um ein

Litigation Hold durchzusetzen. Zusätzlich wären auch Werkzeuge für das Sammeln der Daten auf den verschiedenen Systemen nützlich.

FreeEed/SHMCloud ist ein leistungsfähiges, zukunftsweisendes eDiscovery Projekt. Dieses ist noch nicht komplett ausgereift und weist daher einige Mängel auf, da bestimmte Funktionen noch nicht oder schlecht unterstützt werden. Dennoch ist es wahrscheinlich, dass das Projekt eine interessante Zukunft vor sich hat. Mit einiger Entwicklungsarbeit wird es zu einem guten Produkt für eDiscovery und kann der traditionellen eDiscovery Software Konkurrenz machen. Der eDiscovery Markt bleibt somit weiterhin spannend und man wird sehen ob sich ein Open Source Projekt behaupten kann.

Anhang 1 - Installation

Um die Leistungsfähigkeit des Hadoop basierenden eDiscovery Programms FreeEed zu testen, wurden 12 handelsübliche Standardrechner verwendet. Diese waren jeweils mit einem Intel Core 2 CPU 1.86GHz, 4 GB Arbeitsspeicher und je einer 80 GB Festplatte bestückt. Als Betriebssystem wurde Ubuntu Server 12.04 LTS¹⁵ verwendet.



Abbildung 32: Hadoop-Cluster Aufbau im Netzwerklabor der FH St. Pölten

Um die 12 Rechner mit Ubuntu Server 12.04 LTS zu installieren, wurde die neue Funktion Metal as a Service (MAAS), die von dem neuen 12.04 Release bereitgestellt wird, verwendet. MAAS ist ein Werkzeug für Hyperscale-Implementierungen. Damit sollen Rechenzentren in die Lage versetzt werden, die Hardware und die darauf laufenden Dienste einfach einzurichten und dynamisch bereitzustellen. MAAS verwendet einige Technologien, die

¹⁵ <http://www.ubuntu.com/download/server>

bereits in Canonicals voriger Bereitstellungsplattform JuJu eingesetzt wurden. MAAS ist für Umgebungen gedacht, in denen viele Serverknoten zu verwalten sind. Mit Hilfe der in Python implementierten Software lassen sich innerhalb kurzer Zeit Knoten via Netzwerk und PXE installieren und später aktualisieren. Nachdem der Server eingerichtet wurde, kann mithilfe von JuJu die nötige Software installiert werden.

MAAS stellt eine einfach zu benutzende Weboberfläche bereit, über die ein Administrator bis zu einigen tausend Serverknoten einrichten und verwalten kann. Die Knoten werden mit Hilfe des Tools Cobbler aufgelistet. [117]

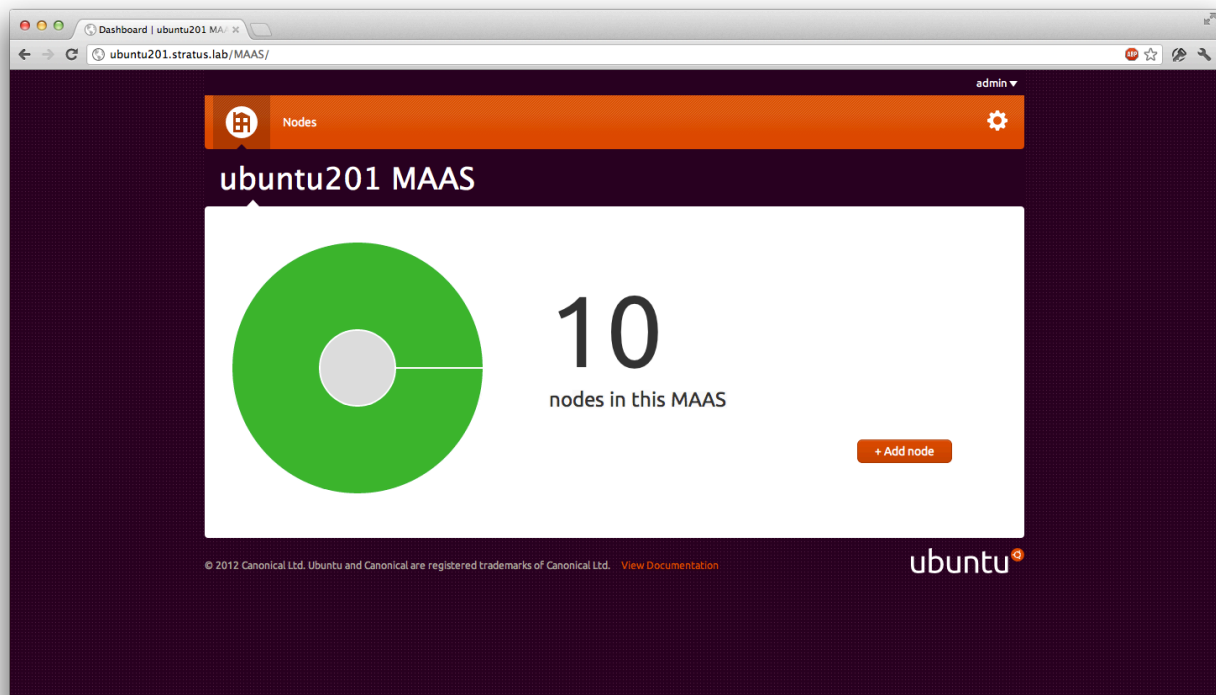


Abbildung 33: Ubuntu MAAS Web GUI

Um MAAS zu installieren, wählt man beim Booten mit dem Ubuntu 12.04 Server Image folgende Bootoption aus 'Ubuntu MAAS Installation or Enlistment'. Bei der Installation wählt man noch die Option 'Create a new MAAS on this server '. Nachdem die Installation abgeschlossen ist, müssen folgende Befehle ausgeführt werden¹⁶:

Anlegen eines Administrator Accounts für MAAS:

```
1. $ sudo maas createsuperuser
[sudo] password for marcel:
Username (Leave blank to use 'root'):
E-mail address: is101507@fhstp.ac.at
Password:
Password (again):
Superuser created successfully.
```

Hier wird der DHCP Server 'dnsmasq' installiert um den PXE boot zu ermöglichen:

```
1. $ sudo apt-get install maas-dhcp
```

¹⁶ <https://wiki.ubuntu.com/ServerTeam/MAAS/>

Da 'dnsmasq' die /etc/hosts als Quelle für die DNS Namensauflösung verwendet, muss hier IP und Hostname eingetragen werden:

```
1. $ cat /etc/hosts
2. 127.0.0.1        localhost
3. 10.110.11.71     node1.sky.lab   node1
4. 10.110.11.72     node2.sky.lab   node2
5. 10.110.11.73     node3.sky.lab   node3
6. ...
```

Da beim Installationsprozess von 'maas-dhcp' nicht alle Optionen konfiguriert werden konnten, wurde manuell noch ein NTP Zeitserver eingetragen und die IP Adressen anhand der MAC-Adresse reserviert.

```
1. $ cat /etc/cobbler/dnsmasq.template
2. read-ethers
3. addn-hosts = /var/lib/cobbler/cobbler_hosts
4. domain=sky.lab
5.
6. dhcp-range=10.110.11.70,10.110.11.90
7. dhcp-option=3,10.110.11.254
8. dhcp-lease-max=1000
9. dhcp-authoritative
10. dhcp-boot=pxelinux.0
11. dhcp-boot=net:normalarch,pxelinux.0
12. dhcp-boot=net:ia64,$elilo
13. dhcp-option=42,130.133.1.10 # Timeserver time.fu-berlin.de
14. $insert_cobbler_system_definitions
15.
16. #Reserved Addresses
17. dhcp-host=00:19:D1:B1:D5:5A,node1,10.110.11.71,infinite
18. dhcp-host=00:19:D1:89:38:EA,node2,10.110.11.72,infinite
19. dhcp-host=00:19:D1:89:E8:56,node3,10.110.11.73,infinite
20. ...
```

Hiermit lädt der MAAS Server die aktuellen Ubuntu Images herunter:

```
1. $ sudo maas-import-isos
```

Da Ubuntu MAAS bei der automatischen Installation der Knoten einen Benutzer namens Ubuntu mit einem zufällig generierten Passwort anlegt, ist es nicht möglich sich auf den einzelnen Knoten später anzumelden. Dies ist normalerweise auch nicht notwendig, da die gesamte Verwaltung der Knoten über MAAS und JuJu geregelt wird. Für die später notwendige Fehlersuche wurde eine Möglichkeit offengehalten, um sich auf den einzelnen Knoten anzumelden. 'd-i passwd/root-login boolean' wurde daher von 'false' auf 'true' umgestellt.

```
1. $ cat /var/lib/cobbler/kickstarts/maas.preseed
2. ...
3. # User Setup
4. d-i passwd/root-login boolean true
5. d-i passwd/make-user boolean true
6. d-i passwd/user-fullname string ubuntu
7. d-i passwd/username string ubuntu
8. d-i passwd/user-password-crypted password !
9. d-i passwd/user-uid string
10. d-i user-setup/allow-password-weak boolean false
11. d-i user-setup/encrypt-home boolean false
12. d-i passwd/user-default-groups string adm cdrom dialout lpadmin plugdev sambashare
13. d-i debian-installer/allow_unauthenticated string false
```

Anhang 2 - Deployment

Nun hat man verschiedene Möglichkeiten Hadoop zu 'deployen'. Man könnte Apache Hadoop auch mit der Ubuntu Paketverwaltung installieren, mithilfe von MAAS und JuJu ein voll automatisches 'deployment' initialisieren oder aber man verwendet die Hadoop Distribution Cloudera. Die Wahl fiel diesem Fall auf Cloudera. Cloudera kann wahlweise mit Paketen installiert werden oder auch mithilfe des Cloudera Manager automatisch 'deployed' werden.

Zunächst wurde auf dem ersten Knoten der Cloudera Manager heruntergeladen und installiert:

```
1. $ wget http://archive.cloudera.com/cm4/installer/latest/cloudera-manager-installer.bin
2. $ chmod u+x cloudera-manager-installer.bin
3. $ sudo ./cloudera-manager-installer.bin
```

Nach der Installation des Cloudera Managers kann dessen Weboberfläche unter seiner IP-Adresse oder dem Hostnamen und dem Port 7180 aufgerufen werden. In unserem Fall lautet die Adresse `http://10.110.11.78:7180`. Der Standardbenutzer und das Passwort ist jeweils bei einer Neuinstallation 'admin'. In der Weboberfläche werden die IP Adressen, der zu installierenden Cloudera Knoten festgelegt. Nun muss man noch eine Möglichkeit angeben, wie man sich auf den einzelnen Knoten via SSH anmelden kann. Dann wählt man noch die zu installierenden Komponenten aus. Bevor die eigentliche Installation beginnt, wird noch ein Inspector Programm ausgeführt, das alle Installationsparameter überprüft. Hier hat das Inspector Programm die Namensauflösung bemängelt. Daraufhin wurde bei allen Knoten in der `/etc/hosts` Datei die Zeile `127.0.1.1` auskommentiert.

```
1. $ cat /etc/hosts
2. 127.0.0.1    localhost
3. #127.0.1.1  nodeX
```

Nach dieser Änderung konnte das Inspector Programm keinerlei Fehler feststellen, und die Installation der Knoten verlief fehlerfrei.

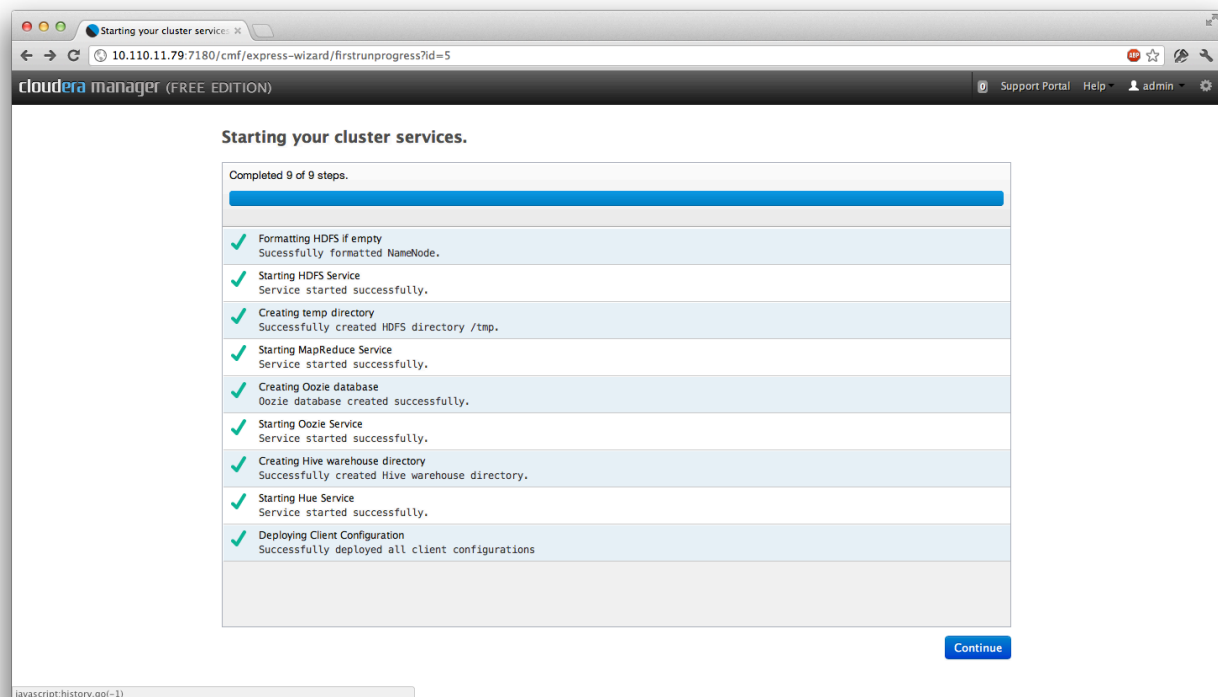


Abbildung 34: Cloudera Manager Installation des Clusters

Anhang 3 - Funktionstest

Nach der erfolgreichen Installation der ersten 3 Knoten wurden zunächst einige allgemeine Funktionen und Performance Tests durchgeführt. Hadoop stellt für diesen Zweck einige Beispiel- und Testprogramme zur Verfügung. Alle verfügbaren Beispiel- und Testprogramme sind in Tabelle 10 und Tabelle 11 aufgelistet. Einige dieser Programme wurden für Funktions- und Performancetests ausgeführt und werden im weiteren Verlauf genauer erläutert.

Programm	Beschreibung
DFSCIOTest	Verteilter I/O Benchmark für libhdfs
DistributedFSCheck	Verteilte Überprüfung des Filesystem Konsistenz
JHLogAnalyzer	Job History Log Analyzer
MRReliabilityTest	Ein Programm, das die Zuverlässigkeit des MR Frameworks testet, indem Ausfälle und Fehler produziert werden.
SliveTest	Ein HDFS Stess Test und Live Daten Verifikation
TestDFSIO	Verteilter I/O Benchmark
fail	Ein Job der immer Fehlschlägt
filebench	Benchmark SequenceFile (Input Output) Format (block ,record compressed und uncompressed), Text (Input Output) Format (compressed und uncompressed)
loadgen	Exemplarischer Map/Reduce Loadgenerator
mapredtest	Ein Programm um Map/Reduce zu prüf und testen.
mrbench	Ein Map/Reduce Benchmark, der viele kleine Jobs erstellen kann
nnbench	Ein Benchmark Test der den NameNode belastet
sleep	Ein Job der bei jedem Map/Reduce Task schläft
testbigmapoutput	Ein Map/Reduce Programm, das mit großen und unteilbaren Dateien arbeitet
testfilesystem	Ein Test für das Filesystem read/write
testmapredsort	Ein Map/Reduce Programm, welches das sortierten überprüft
testsequencefile	Ein Test für flat Files mit binären Key Valu Paaren
testsequencefileinputformat	Ein Test für das sequence file input format
testtextinputformat	Ein Map/Reduce Benchmark der die Performance von Maps mit mehreren Schwellen mit Maps mit 1. Schwelle vergleicht

Tabelle 10: Hadoop Testprogramme [80]

Die Hadoop Beispiele und Testprogramme befinden sich bei CDH4 im folgenden Verzeichnis:

```
1. # pwd
2. /usr/lib/hadoop-mapreduce
```

Mit dem Programm TestDFSIO kann man die Geschwindigkeit des Hadoop-Filesystem HDFS prüfen. Dieses Programm testet sowohl die Lese- und Schreibgeschwindigkeit mithilfe von zufällig erzeugten Daten, ganz ähnlich wie es auch das Programm 'sysbench' auf Linux Filesystemen macht.

Mit dem Aufruf dieses Befehls werden fünf Dateien mit je 200 MB Daten im HDFS und für jede zu schreibende Datei ein Mapper-Task erzeugt. Insgesamt werden also 1GB Daten generiert.

```
1. # hadoop jar hadoop-mapreduce-client-jobclient-2.0.0-cdh4.0.1-tests.jar TestDFSIO -write -
   nrFiles 5 -fileSize 200
2. ...
3. 12/07/23 21:35:10 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write
```

```

4. 12/07/23 21:35:10 INFO fs.TestDFSIO:      Date & time: Mon Jul 23 21:35:10 CEST 2012
5. 12/07/23 21:35:10 INFO fs.TestDFSIO:      Number of files: 5
6. 12/07/23 21:35:10 INFO fs.TestDFSIO: Total MBytes processed: 1000.0
7. 12/07/23 21:35:10 INFO fs.TestDFSIO:      Throughput mb/sec: 7.680786512538884
8. 12/07/23 21:35:10 INFO fs.TestDFSIO: Average IO rate mb/sec: 9.988778114318848
9. 12/07/23 21:35:10 INFO fs.TestDFSIO:   IO rate std deviation: 6.657976523127291
10. 12/07/23 21:35:10 INFO fs.TestDFSIO:      Test exec time sec: 57.645

```

Um den Datendurchsatz bei Lesezugriff zu ermitteln führt man diesen Befehl aus. Bei diesem Aufruf werden fünf Dateien mit je 200 MB Daten eingelesen.

```

1. # hadoop jar hadoop-mapreduce-client-jobclient-2.0.0-cdh4.0.1-tests.jar TestDFSIO -read -
   nrFiles 5 -fileSize 200
2. ...
3. /07/23 21:38:23 INFO fs.TestDFSIO: ----- TestDFSIO ----- : read
4. 12/07/23 21:38:23 INFO fs.TestDFSIO:      Date & time: Mon Jul 23 21:38:23 CEST 2012
5. 12/07/23 21:38:23 INFO fs.TestDFSIO:      Number of files: 5
6. 12/07/23 21:38:23 INFO fs.TestDFSIO: Total MBytes processed: 1000.0
7. 12/07/23 21:38:23 INFO fs.TestDFSIO:      Throughput mb/sec: 227.79043280182233
8. 12/07/23 21:38:23 INFO fs.TestDFSIO: Average IO rate mb/sec: 236.41342163085938
9. 12/07/23 21:38:23 INFO fs.TestDFSIO:   IO rate std deviation: 48.66240859523409
10. 12/07/23 21:38:23 INFO fs.TestDFSIO:      Test exec time sec: 24.713

```

Um alle Testdaten zu löschen, ruft man TestDFSIO wie folgt auf:

```

1. # hadoop jar hadoop-mapreduce-client-jobclient-2.0.0-cdh4.0.1-tests.jar TestDFSIO -clean
2. 12/07/23 21:39:07 INFO fs.TestDFSIO: TestDFSIO.0.0.6
3. 12/07/23 21:39:07 INFO fs.TestDFSIO: nrFiles = 1
4. 12/07/23 21:39:07 INFO fs.TestDFSIO: fileSize (MB) = 1.0
5. 12/07/23 21:39:07 INFO fs.TestDFSIO: bufferSize = 1000000
6. 12/07/23 21:39:07 INFO fs.TestDFSIO: baseDir = /benchmarks/TestDFSIO
7. 12/07/23 21:39:07 INFO fs.TestDFSIO: Cleaning up test files

```

Der MapReduce-Benchmark führt beliebige Mapper- und Reduce-Tasks aus um die Verarbeitungsgeschwindigkeit zu messen. Dieses Beispiel führt 2 Test-Mapper und Reducer aus. Am Ende erhält man dann die durchschnittliche Ausführungszeit in Millisekunden.

```

1. # hadoop jar hadoop-mapreduce-client-jobclient-2.0.0-cdh4.0.1-tests.jar mrbench -numRuns 2
2. ...
3. ...
4. DataLines      Maps      Reduces      AvgTime (milliseconds)
5. 1              2          1             22586

```

Der NameNode Benchmark führt einen Loadtest auf dem NameNode aus. Dabei wird sowohl die Konfiguration als auch die Hardware getestet. Der folgende Aufruf von NNbench erzeugt 500 Dateien mit Hilfe von 10 Mapper- und 5 Reducer-Tasks.

```

1. # hadoop jar hadoop-mapreduce-client-jobclient-2.0.0-cdh4.0.1-tests.jar nnbench -operation
   create_write -maps 10 -reduces 5 -blockSize 1 -bytesToWrite 0 -numberOfFiles 500 -
   replicationFactorPerFile 3 -readFileAfterOpen true
2. NameNode Benchmark 0.4
3. 12/07/23 21:43:48 INFO hdfs.NNbench: Test Inputs:
4. 12/07/23 21:43:48 INFO hdfs.NNbench:      Test Operation: create_write
5. 12/07/23 21:43:48 INFO hdfs.NNbench:      Start time: 2012-07-23 21:45:48,200
6. 12/07/23 21:43:48 INFO hdfs.NNbench:      Number of maps: 10
7. 12/07/23 21:43:48 INFO hdfs.NNbench:      Number of reduces: 5
8. 12/07/23 21:43:48 INFO hdfs.NNbench:      Block Size: 1
9. 12/07/23 21:43:48 INFO hdfs.NNbench:      Bytes to write: 0
10. 12/07/23 21:43:48 INFO hdfs.NNbench:      Bytes per checksum: 1
11. 12/07/23 21:43:48 INFO hdfs.NNbench:      Number of files: 500
12. 12/07/23 21:43:48 INFO hdfs.NNbench:      Replication factor: 3
13. 12/07/23 21:43:48 INFO hdfs.NNbench:      Base dir: /benchmarks/NNbench
14. 12/07/23 21:43:48 INFO hdfs.NNbench:      Read file after open: true
15. 12/07/23 21:43:49 INFO hdfs.NNbench: Deleting data directory
16. 12/07/23 21:43:49 INFO hdfs.NNbench: Creating 10 control files

```

```

17. ...
18. 12/07/23 21:46:47 INFO hdfs.NNBench: ----- NNBench ----- :
19. 12/07/23 21:46:47 INFO hdfs.NNBench: Version: NameNode
    Benchmark 0.4
20. 12/07/23 21:46:47 INFO hdfs.NNBench: Date & time: 2012-07-23
    21:46:47,985
21. 12/07/23 21:46:47 INFO hdfs.NNBench:
22. 12/07/23 21:46:47 INFO hdfs.NNBench: Test Operation: create_write
23. 12/07/23 21:46:47 INFO hdfs.NNBench: Start time: 2012-07-23
    21:45:48,200
24. 12/07/23 21:46:47 INFO hdfs.NNBench: Maps to run: 10
25. 12/07/23 21:46:47 INFO hdfs.NNBench: Reduces to run: 5
26. 12/07/23 21:46:47 INFO hdfs.NNBench: Block Size (bytes): 1
27. 12/07/23 21:46:47 INFO hdfs.NNBench: Bytes to write: 0
28. 12/07/23 21:46:47 INFO hdfs.NNBench: Bytes per checksum: 1
29. 12/07/23 21:46:47 INFO hdfs.NNBench: Number of files: 500
30. 12/07/23 21:46:47 INFO hdfs.NNBench: Replication factor: 3
31. 12/07/23 21:46:47 INFO hdfs.NNBench: Successful file operations: 0
32. 12/07/23 21:46:47 INFO hdfs.NNBench:
33. 12/07/23 21:46:47 INFO hdfs.NNBench: # maps that missed the barrier: 0
34. 12/07/23 21:46:47 INFO hdfs.NNBench: # exceptions: 99
35. 12/07/23 21:46:47 INFO hdfs.NNBench:
36. 12/07/23 21:46:47 INFO hdfs.NNBench: TPS: Create/Write/Close: 0
37. 12/07/23 21:46:47 INFO hdfs.NNBench: Avg exec time (ms): Create/Write/Close: 0.0
38. 12/07/23 21:46:47 INFO hdfs.NNBench: Avg Lat (ms): Create/Write: Infinity
39. 12/07/23 21:46:47 INFO hdfs.NNBench: Avg Lat (ms): Close: NaN
40. 12/07/23 21:46:47 INFO hdfs.NNBench:
41. 12/07/23 21:46:47 INFO hdfs.NNBench: RAW DATA: AL Total #1: 108962
42. 12/07/23 21:46:47 INFO hdfs.NNBench: RAW DATA: AL Total #2: 0
43. 12/07/23 21:46:47 INFO hdfs.NNBench: RAW DATA: TPS Total (ms): 0
44. 12/07/23 21:46:47 INFO hdfs.NNBench: RAW DATA: Longest Map Time (ms): 0.0
45. 12/07/23 21:46:47 INFO hdfs.NNBench: RAW DATA: Late maps: 0
46. 12/07/23 21:46:47 INFO hdfs.NNBench: RAW DATA: # of exceptions: 99

```

Unter den Example Programmen finden sich Tools, mit deren Hilfe sich die Geschwindigkeit des Hadoop-Clusters für ausgesuchte Anwendungsfälle bestimmen lässt.

Programm	Beschreibung
aggregatewordcount	Ein Aggregatbasierendes MapReduce Programm, welches die Wörter des Input Files zählt
aggregatewordlist	Ein Aggregatbasierendes MapReduce Programm, welches das Histogramm der Wörter des Input Files berechnet
dbcount	Ein Beispiel Job der die pageviews von einer Datenbank zählt.
grep	Ein MapReduce Programm das die passenden Regular Expressions im Input File zählt
join	Ein Job der einen Join der sortierten und gleich groß partitionierten Datensets durchführt
multifilewc	Der Job zählt die Anzahl der Wörter von mehreren Files
Pentomino	Ein MapReduce Programm das Lösungen für Pentomino Probleme sucht
pi	Ein MapReduce Programm, das den pi Wert mithilfe der Monte Carlo Methode bestimmt
randomtextwriter	Ein MapReduce Programm, welches 10GB Zufallstextdaten pro Knoten produziert
randomwriter	Ein MapReduce Programm, welches 10GB Zufallsdaten pro Knoten produziert
secondarysort	Ein Beispielprogramm, welches einen secondary sort für den Reduce definiert.
sleep	Ein Job der bei jedem map und reduce taks schläft
sort	Ein MapReduce Programm, welches die Daten sortiert, die von dem randomwriter erzeugt wurden
sudoku	Ein Sodoku Löser
teragen	Generiert Daten für den Terasort
terasort	Führt den Terasort aus
teravalidate	Überprüft das Resultat des Terasort
wordcount	Ein MapReduce Programm, dass die Wörter in einem Input File zählt

Tabelle 11: Hadoop Beispielprogramme [80]

Das Beispielprogramm 'Sort', sortiert den Inhalt von Dateien mit Hilfe von Map- und Reduce Tasks. Um den Sort Test durchzuführen, benötigt es folgende Schritte:

- zufällige Beispieldaten erzeugen (randomwriter)
- erzeugte Daten sortieren (sort)

Dieser Aufruf erzeugt Testdaten im Verzeichnis /random-test-data:

```
1. # hadoop jar hadoop-mapreduce-examples-2.0.0-cdh4.0.1.jar randomwriter -
   Dtest.randomwriter.maps_per_host=2 random-test-data
2. Running 6 maps.
3. ...
4. Job ended: Mon Jul 23 21:53:01 CEST 2012
5. The job took 220 seconds.
```

Der Aufruf um die Testdaten zu sortieren:

```
1. # hadoop jar hadoop-mapreduce-examples-2.0.0-cdh4.0.1.jar sort random-test-data sorted-test-
   data
2. Running on 3 nodes to sort from hdfs://node9.sky.lab:8020/user/root/random-test-data into
   hdfs://node9.sky.lab:8020/user/root/sorted-test-data with 2 reduces.
3. ...
4. Job ended: Mon Jul 23 22:09:03 CEST 2012
5. The job took 775 seconds.
```

[84]

Der PI Estimator benutzt die Monte Carlo Methode um PI zu bestimmen. Das Programm verwendet zwei Integer Argumente, zum einen die Anzahl von Maps und die Anzahl der Samples pro Map. [80]

```
1. # hadoop jar hadoop-mapreduce-examples-2.0.0-cdh4.0.1.jar pi 2 10
```

```

2. Number of Maps = 2
3. Samples per Map = 10
4. Wrote input for Map #0
5. Wrote input for Map #1
6. Starting Job
7. ...
8. Job Finished in 22.799 seconds
9. 12/07/23 22:11:11 INFO util.NativeCodeLoader: Loaded the native-hadoop library
10. Estimated value of Pi is 3.80000000000000000000

```

Welche Hadoop-Konfigurationsverzeichnisse innerhalb von CDH4 genutzt werden, zeigt der folgende Befehl:

```

1. $ sudo update-alternatives --display hadoop-conf
2. hadoop-conf - auto mode
3.   link currently points to /etc/hadoop/conf.cloudera.mapreduce1
4.   /etc/hadoop/conf.cloudera.hdfs1 - priority 90
5.   /etc/hadoop/conf.cloudera.mapreduce1 - priority 91
6.   /etc/hadoop/conf.empty - priority 10
7.   Current 'best' version is '/etc/hadoop/conf.cloudera.mapreduce1'.

```

Der HDFS Status lässt sich über das folgende Kommando abrufen. Dabei kann man alle HDFS Details wie z.B. den gesamten verfügbaren Speicherplatz einsehen.

```

1. $ sudo -u hdfs hadoop dfsadmin -report
2. DEPRECATED: Use of this script to execute hdfs command is deprecated.
3.   Instead use the hdfs command for it.
4.
5. Configured Capacity: 191644925952 (178.48 GB)
6. Present Capacity: 191016632320 (177.9 GB)
7. DFS Remaining: 112483745792 (104.76 GB)
8. DFS Used: 78532886528 (73.14 GB)
9. DFS Used%: 41.11%
10. Under replicated blocks: 0
11. Blocks with corrupt replicas: 0
12. Missing blocks: 0
13.
14. -----
15. Datanodes available: 3 (3 total, 0 dead)
16.
17. Live datanodes:
18. Name: 10.110.11.77:50010 (node7.sky.lab)
19. Hostname: node7.sky.lab
20. Rack: /default
21. Decommission Status : Normal
22. Configured Capacity: 63881641984 (59.49 GB)
23. DFS Used: 26161532928 (24.36 GB)
24. Non DFS Used: 0 (0 KB)
25. DFS Remaining: 37720109056 (35.13 GB)
26. DFS Used%: 40.95%
27. DFS Remaining%: 59.05%
28. Last contact: Mon Jul 23 22:13:34 CEST 2012
29.
30. Name: 10.110.11.78:50010 (node8.sky.lab)
31. Hostname: node8.sky.lab
32. Rack: /default
33. Decommission Status : Normal
34. Configured Capacity: 63881641984 (59.49 GB)
35. DFS Used: 26209665024 (24.41 GB)
36. Non DFS Used: 0 (0 KB)
37. DFS Remaining: 37671976960 (35.08 GB)
38. DFS Used%: 41.03%
39. DFS Remaining%: 58.97%
40. Last contact: Mon Jul 23 22:13:33 CEST 2012
41.
42. Name: 10.110.11.79:50010 (node9.sky.lab)
43. Hostname: node9.sky.lab

```

```
1. Rack: /default
2. Decommission Status : Normal
3. Configured Capacity: 63881641984 (59.49 GB)
4. DFS Used: 26161688576 (24.36 GB)
5. Non DFS Used: 628293632 (599.19 MB)
6. DFS Remaining: 37091659776 (34.54 GB)
7. DFS Used%: 40.95%
8. DFS Remaining%: 58.06%
9. Last contact: Mon Jul 23 22:13:34 CEST 2012
```

Diese Details lassen sich auch unter der Weboberfläche des Namenode einsehen, dazu gibt man im Browser einfach die IP-Adresse des NameNodes und den Port 50070 an, der Port für den Secondary NameNode lautet 50090. Des Weiteren stellt Cloudera 4 noch die folgende weiteren webbasierten Elemente bereit: Unter der IP-Adresse des Jobtracker und Port 50030 findet man eine Übersicht aller Jobs und ihren Status. Unter der IP-Adresse des Cloudera Managers und Port 7180 erreicht man den Cloudera Manager, indem man Service hinzufügen, entfernen kann und auch neue Knoten hinzufügen oder entfernen kann. Die Weboberfläche des HUE Servers erreicht man unter der entsprechenden IP-Adresse mit der Portnummer 8888.

Anhang 4 – FreeEed

In den folgenden Schritten wird die eDiscovery Software FreeEed/SHMCloud eingerichtet. Zunächst wird die Software von der Projektseite heruntergeladen und extrahiert.

```
1. # wget http://freeeed.org/images/download/FreeEed.zip
2. # unzip FreeEed.zip
```

Nachdem man FreeEed heruntergeladen und entpackt hat, öffnet man ein Projekt und führt das Staging durch. Dabei kompiliert FreeEed alle input Dateien in mehrere ZIP Pakete. Dies hat den Vorteil, dass die Originaldateien nicht verändert werden bei der weiteren Verarbeitung. Die Dateien wurden bereits komprimiert für den Upload in den Cluster.

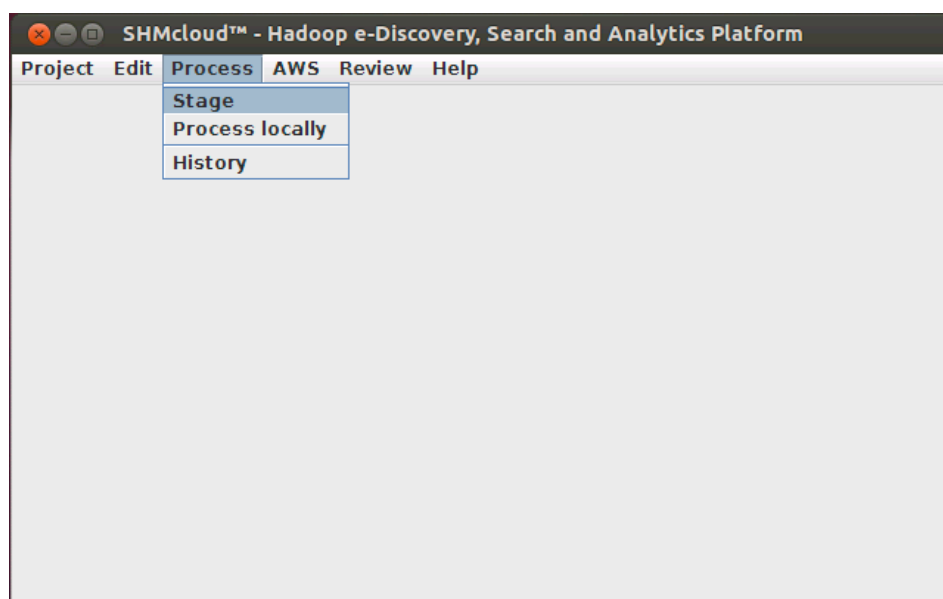


Abbildung 35: FreeEed Stage

Nachdem das Staging abgeschlossen wurde, soll ein Testlauf mit der run_hadoop.sh ausgeführt werden. Jedoch erhält man folgende Fehlermeldung.

```
1. #./run_hadoop.sh
2. rmr: DEPRECATED: Please use 'rm -r' instead.
3. rmr: `/freeeed/output': No such file or directory
4. SHMCloud V4.0.4
5. Build time: Wed Jul 11 07:40:45 CEST 2012
6. Running Hadoop job
7. Input project file = small_hadoop.project
8. Output path = /freeeed/output
9. Reading project file small_hadoop.project
10. project.isEnvHadoop() = true
11. 12-07-11 18:42:10 Running command: hadoop fs -rmr /freeeed_work_area/SHM003
12. 12-07-11 18:42:12 rmr: DEPRECATED: Please use 'rm -r' instead.
13. 12-07-11 18:42:12 rmr: `/freeeed_work_area/SHM003': No such file or directory
14. 12-07-11 18:42:12 Running command: hadoop fs -mkdir /freeeed_work_area/SHM003
15. Exception in thread "main" java.lang.NoSuchMethodError:
    com.google.common.io.Files.deleteDirectoryContents(Ljava/io/File
16.
17. ...
```

Dieser Fehler konnte behoben werden indem man Guava Jar Dateien in '*.Backup' umbenennt, damit diese nicht mehr aufgerufen werden. Bei Guava handelt es sich um Google core Bibliotheken, welche Java basierten Projekten, Funktionen wie collections, caching, string processing, I/O, zur Verfügung stellt. [118]

Nachdem dieser Fehler behoben war, wurde die run_hadoop.sh erneut ausgeführt. Jedoch erhält man einen weiteren Fehler, bei dem man darauf hingewiesen wird, dass eine Datei oder Ordner fehlt.

```

1. # ./run_hadoop.sh
2. rmr: cannot remove /freeeed/output: No such file or directory.
3. SHMCloud V3.8.8
4. Build time: Thu Jun 21 01:53:35 CEST 2012
5. Running Hadoop job
6. Input project file = sample_hadoop.project
7. Output path = /freeeed/output
8. Reading project file sample_hadoop.project
9. project.isEnvHadoop() = true
10. 12-06-27 17:58:19 Running command: hadoop fs -rmr /freeeed_work_area/0001
11. 12-06-27 17:58:20 Running command: hadoop fs -mkdir /freeeed_work_area/0001
12. 12-06-27 17:58:21 Running command: hadoop fs -copyFromLocal /mnt/tmp/tmpinput
    /freeeed_work_area/0001/input1
13. 12-06-27 17:58:22 Running command: hadoop fs -copyFromLocal /mnt/tmp/tmpinput
    /freeeed_work_area/0001/input2
14. 12-06-27 17:58:23 Running command: hadoop fs -copyFromLocal /mnt/tmp/tmpinput
    /freeeed_work_area/0001/input3
15. 12-06-27 17:58:24 Running command: hadoop fs -copyFromLocal /mnt/tmp/tmpinput
    /freeeed_work_area/0001/input4
16. 12-06-27 17:58:26 Running command: hadoop fs -copyFromLocal /mnt/tmp/tmpinput
    /freeeed_work_area/0001/input5
17. Ready to run, inputPath =
    /freeeed_work_area/0001/input1,/freeeed_work_area/0001/input2,/freeeed_work_area/0001/input3,/
    freeeed_work_area/0001/input4,/freeeed_work_area/0001/input5, outputPath = /freeeed/output
18. 12/06/27 17:58:28 INFO input.FileInputFormat: Total input paths to process : 5
19. 12/06/27 17:58:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
    platform... using builtin-java classes where applicable
20. 12/06/27 17:58:28 WARN snappy.LoadSnappy: Snappy native library not loaded
21. 12/06/27 17:58:28 INFO mapred.JobClient: Running job: job_201204241022_0022
22. 12/06/27 17:58:29 INFO mapred.JobClient: map 0% reduce 0%
23. 12/06/27 17:58:39 INFO mapred.JobClient: Task Id : attempt_201204241022_0022_m_000001_0,
    Status : FAILED
24. java.io.IOException: No such file or directory
25. ...
    
```

Nach einiger Zeit der Evaluierung stellte sich heraus, dass die 'gestageten' Files nicht im HDFS System vorhanden sind sondern nur auf dem lokalen File System. Daher hat man versucht die kompilierten Daten manuell auf das HDFS Filesystem zu kopieren, jedoch hat dies leider auch nicht funktioniert, da die Daten bei dem automatischen Kopierprozess eine andere Bezeichnung erhalten.

```

1. # hadoop fs -copyFromLocal /home/ubuntu/SHMCloud/test-data/staged/* /freeeed/input-data/
    
```

Um das Staging manuell über die Kommandozeile durchzuführen, wurde mit dem Build-Management-Tool Maven das FreeEed Projekt neu erstellt und kompiliert. Leider war auch dieser Versuch erfolglos. [119]

```

1. # mvn install:install-file -DgroupId=com.independentsoft \
2. -DartifactId=JPST \
3. -Dversion=1.0 \
4. -Dfile=JPST/lib/jpst.jar \
5. -Dpackaging=jar \
6. -DgeneratePom=true
7. # mvn -Pcloudera clean install
    
```

Nach Rücksprache mit dem FreeEed Entwickler Mark Kerzner ist dies momentan noch ein generelles Problem bei FreeEed. FreeEed funktioniert bisher ausschließlich in der Amazon Cloud mit speziell angepassten Hadoop Maschinen. Da jedoch das Ziel dieser Arbeit darin bestand einen eigenen Cluster bzw. eine Private Cloud zu

betreiben, wurde überlegt, ob es möglich ist das Image der Amazon Instanz zu exportieren und auf einem VMware ESXi¹⁷ Open Stack¹⁸ oder Eucalyptus¹⁹ Umgebung zu betreiben. Um die Amazon AMI zu exportieren wurden die Amazon EC2 API Tools installiert. [120] [121]

```
1. $ ec2-describe-images -a | grep ami-2f812e46
2. IMAGE    ami-2f812e46    shmsoft/SHMCloud-ami-image_nokey/image.manifest.xml 959297014381
   available public        x86_64 machine aki-825ea7eb          instance-
   store    paravirtual  xen
3. TAG      image    ami-2f812e46    Name    SHMSOFT
```

```
1. $ ec2-create-instance-export-task ami-2f812e46 -e vmware -f vmdk -c ova -b mierrebach
```

Nach genauerer Rücksprache mit Mark Kerzner wurde der Plan jedoch wieder verworfen, da der Anpassungsaufwand der AMI auf eine andere Umgebung wie VMware ESXi, Open Stack oder Eucalyptus definitiv zu groß sei.

Daraufhin wurde nach mehreren erfolglosen Versuchen schlussendlich entschieden, FreeEed nicht auf dem Hadoop Cluster zu betreiben sondern die Amazon Cloud zu nutzen.

¹⁷ <http://www.vmware.com/products/vsphere-hypervisor/overview.html>

¹⁸ <http://www.openstack.org>

¹⁹ <http://www.eucalyptus.com>

Anhang 5 – SHMCloud

Hierfür wurde SHMCloud heruntergeladen und entpackt.

```
1. # wget https://s3.amazonaws.com/shmsoft/SHMCloud.zip
2. # unzip SHMCloud.zip
```

Als erstes muss man unter dem Menüpunkt AWS das S3 Setup und EC2 Setup durchführen. Dazu trägt man die Access Key ID und den Secret Access Key ein, den man in der AWS Management Console findet. Danach muss noch ein Bucket erstellt werden, in dem die Daten gespeichert werden.

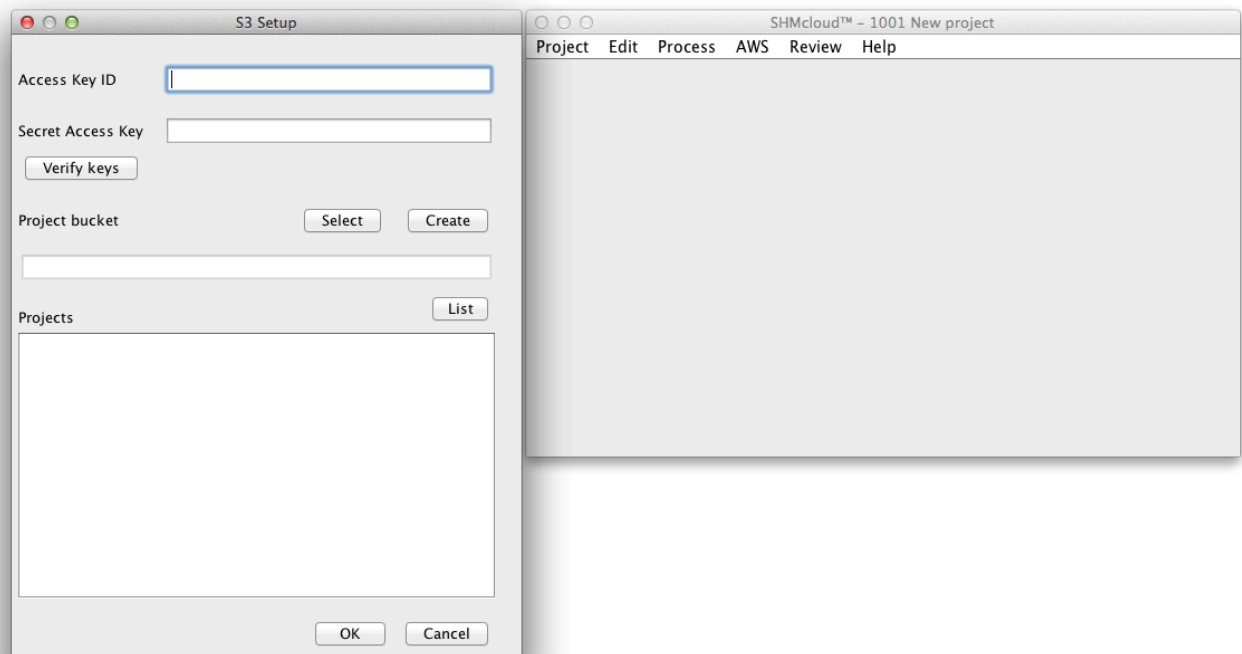


Abbildung 36: SHMCloud S3 Setup

In der AWS Management Console erstellt man zunächst eine Security Group in der alle benötigten Hadoop Ports und SSH eingetragen werden.

```
1. TCP Port: 0 - 6000 Source: Security Group Name
2. TCP Port: 22, 3389, 50030, 50060, 50070 Source: WAN IP
```

Ebenso muss man sich ein Key pair erstellen. Den Namen der Security group und des Key pair trägt man unter EC2 setup ein. Dann wählt man noch die gewünschte Zone und den Instance type aus, hier stehen medium und large zur Auswahl. Der Wert für Cluster size legt fest, wie viele Nodes der Cluster haben soll.

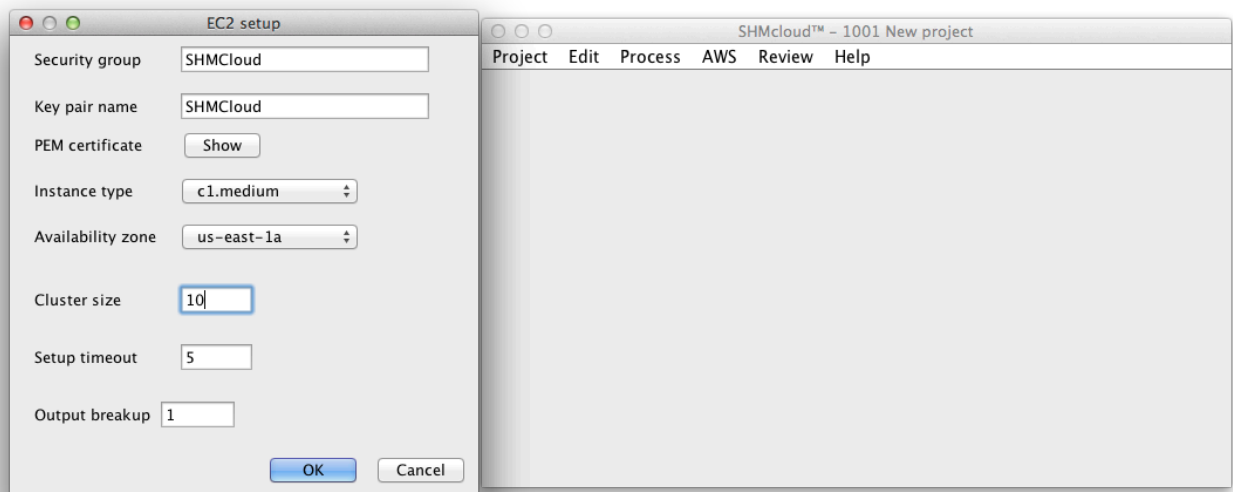


Abbildung 37: SHMCloud EC2 Setup

Nach diesen Konfigurationsschritten kann man ein Projekt erstellen oder öffnen und das Staging durchführen. Danach kann man im AWS Menü den Cluster starten. Nachdem dann alle Nodes generiert und online sind, kann die Verarbeitung auf dem Cluster durchgeführt werden.

Anhang 6 - Übersicht Fileformattest

Übersicht der getesteten Fileformate für Nuix. Die mit einem 'x' markierten Dateien konnten nicht erkannt werden.

FOLDER	FILENAME	FILE FORMAT (REPORTED)	UNSUPPORTED FILES
acs	YESNO.MDB	MS Access 2.0	x
ami	POINTS.SAM		
bdr	PRINT1.OBD		x
bmp	4seafoodc.cdr	corel draw 3	
bmp	911.CDR	corel draw 4	
bmp	AQUA-OS2.BMP	OS/2 bmp	(x)
bmp	BLUES.BMP	icon	(x)
bmp	C5MENU.CDR	corel draw 5	
bmp	EYE.CDR	corel draw 2	
bmp	JAZZ.BMP	win bmp	
bmp	mdiframe.ico	OS/2 wrap bmp	x
bmp	PO.cdr	corel draw 6.0	
bmp	SELECT.CUR	cursor	x
cdr	5radialg.cdr		
cdr	award.cdr		
cdr	BOXES.CDR		
cdr	FROMWPG.CDR		
cdr	GOODPEAR.CDR		
cdr	itmenu.cdr		
cdr	UNIFILL.CDR		
cgm	APPLE.CGM	computer graphics metafile	(x)
cmx	NEWSFLSH.CMX	Corel Draw Clipart	
dbf	BUSINESS.DBF	dbase 3	
dbf	PICTURES.DBF	dbase 5	
dbf	SYSTIMES.DBF	dbase 4	
dif	NAVYDIF1	Navy DIF	
drw	TUTOR07.DRW	Micrografx products	
dsf	COASTER.DSF	Micrografx designer	
dwg	3DPolygon.dwg	AC drawing 2000/2002	
dwg	AIRPLANE.DWG	AC drawing 10	
dwg	ASHADE.DWG	AC drawing 9	
dwg	BLKTEST.DWG	AC drawing 12	
dwg	CHROMA.DWG	AC drawing 13	
dwg	colorwh.dwg	AC drawing 14	
dwg	NOZZLE.DWG	AC drawing 2.5	
dx	DEMIDX41.DX	DX3.1	
dx	TEST1.DX2	DX3.0 and below	
dx	CHAIRBIN.DXF		
dx	CHROMA.DXF		
dx	HOME1.DXF		
dx	lcaddwg1.dxf		
dx	NOZZLE.DXF		
emf	allshapes.emf	emf	(x)
emf	BALCONY.EMF	emf	
emf	ObjectFillStyle.emf	emf	
en4	CHARFORM.WPF	enable wp 5	
en4	ENABLE45.WPF		
ens	FINANCE.SSF	enable ss	
eshr	dasharrow.rtf	rtf with esher	
exe2	demet.dll	dll	x
exe2	MAKEVID.EXE	exe	x
fax	GROUP3B.FAX		
fcd	FLDTYPE.FOL	first choice DB	
fcs	FORMATS.SS	first choice SS	x
fft	BOLD1.FFT	IBM DCA/FFT	
flw	FONT.SPR	freelance	
flw	PATTERNS.PRE	freelance 96/97/Millennium Edition	
fmv	123PIE.FMV		
fmv	KITCHEN.FMV		
fpx	catbird.fpx		
fwk	REPORT.FW3		

gdf	GDF	IBM Graphic Data Format	
gem	TRUMPET.GEM	Gem file	
gif	A_BITMAP.GIF	gif	x
gif	bannerimage[1].gif	gif	x
GP4	SCALE.cal		
gzip	uedit32.ini.gz	gzip	
hgs	ORG2.CH3	Harvard 3.0 Chart	
hgw	OBJS2.PRS	Harvard Graphics	
html	GGGR.HTML	html	
hwp	Korean_tab.hwp	hangul 97	x
hwp	LESSON2.HWP	hangul 97	x
hwp2	TextBoxes1.hwp	hangul 2002	
ich2004	flag.JTT	ichitaro higher 7.0 and up	
ich2004	header.jtd	ichitaro higher 7.0 and up	
ich2004	miyazawaKenji.jtt	ichitaro higher 7.0 and up	
ich2004	overlay.jtd	ichitaro higher 7.0 and up	
ich2004	welcome.JTD	ichitaro higher 7.0 and up	
ich6	NUMBER.JBW	ichitaro 5-6	
igs	F408X.IGS	IGES Drawing	
igs	IGESD.IGS	IGES Drawing	
img	RAZORBK.IMG	GEM image	x
iwp	ASCII.DOC	Wang IWP	
iwp	CHARFORM.DOC	Wang IWP	
iwp	INDENTS.DOC	Wang IWP	
iwp	MANUSC.DOC	Wang IWP	
iwp	OUTLINE.DOC	Wang IWP	
jbg2	003288.jb2.pdf	jbig2 embedded pdf	x
jpg	leaf_BG.JPG		(x)
jw	COLS3.JW	just write 1.0	
jw	DEMOQA3.QW	Q A write 3.0	
jw	JWRITE2.JW	just write 2.0	
l123r9	Place Inside Plot Area.123	lotus 123	
leg	GGGR.ws5	word star 5.0	
leg	GUIDE.WSD	word star for windows	
leg	LEGACY.CHP	legacy	
lwp	complex.lwp	word pro 97/millennium	
lzh	H.EXE	Self-Extracting LZH	x
lzh	S.COM	Self-Extracting LZH	
lzh	WINSOCK.LZH	LZH Compress	
m11	TEST.AF0	Mass 11	x
manu	SAMPLE7.DOC	Lotus Manuscript 2.0	
mcw	BOLD	MacWrite II	
met	AIRPLANE.MET	OS/2 Metafile	
met	PAPIE.MET	OS/2 Metafile	
met	USA2.MET	OS/2 Metafile	
mif	46226.MIF	MIFF 4.0	
mif	MSBUG2.MIF	MIFF 5.5	
mif	process.mif	MIFF 6.0	
mif	TAG2.MIF	MIFF 5.0	
mm	COMTAB.DOC	MultiMate Advantage 2	
mm	UNDER1.DOC	MultiMate Advantage 2	
mm4	MATH.DOX	MultiMate 4.0	
MMFN	BLD2.FNT	MultiMate Note	x
mp	MPLAN4.MOD	Multiplan 4.0	
mpp	1task3resources.mpp	Microsoft Project 2000/2002/2003	
mpp	DrawingWithoutTasks.mpp	Microsoft Project 98	
mpp	ENGINEER.mpp	Microsoft Project 2000/2002/2003	
mpp	ProjectInfoWithBaseline.mpp	Microsoft Project 2000/2002/2003	
msg	ManyFields2.msg	MS Outlook	
msg	ManyFields3.msg	MS Outlook	
msg	ManyFields4.msg	MS Outlook	
msg	ManyFields5.msg	MS Outlook	
msg	Pink Note.msg	MS Outlook	
msg	Regular.msg	MS Outlook	
msw	CENTER.DOC	Word for DOS 4.x	
msw	NARROW.WRI	Windows Write	

msw	TABLE1.DOC	Word for DOS 6.x	
msw	WPRD55.DOC	Word for DOS 5.x	
mwkd	MACWRK2.DB	Mac Works 2.0 DB	
mwp2	TABLE2	Mac WordPerfect 3.0	
mwp2	WPM31_C	Mac WordPerfect 3.0	
mwpf	WPFFORM4	Mac WordPerfect 1.x	
mwrk	FLRIGHT.MWK	Mac Works 2.0 WP	
ow	OWTUTOR.WP	OfficeWriter	
pbm	543x543.ppm	PPM (Portable Pixmap)	x
pbm	A6.ppm	PPM (Portable Pixmap)	(x)
pbm	d1.pgm	PGM (Portable Graymap)	x
pbm	disk.pbm	PBM (Portable Bitmap)	x
pbm	FL.pgm	PGM (Portable Graymap)	(x)
pbm	gnu.pbm	PBM (Portable Bitmap)	x
pbm	le.ppm	PPM (Portable Pixmap)	
pbm	mab.pbm	PBM (Portable Bitmap)	
pbm	owl.pgm	PGM (Portable Graymap)	
pbm	tmp.pbm	PGM (Portable Graymap)	
pcd	ICESTORM.PCD	Kodak Photo CD	
pcl	PCFILE50.LTR	PC File 5.0 Doc	
pcx	APPLE.PCX	Paintbrush	(x)
pcx	SKELET4.DCX	Multipage PCX	x
pcx	WATERFAL.PCX	Paintbrush	(x)
pdf	artistic2.ai	AI 9.0	
pdf	LEAR.PDF	PDF	
pdx	ATTFILES.DB	Paradox 2 or 3	
pdx	BOOKORD.DB	Paradox	
pdx	NEWRECS.DB	Paradox 2 or 3	
pdx	SYMBOLS.DB	Paradox	
pdx	TASKLIST.DB	Paradox 3.5	
pdx	TESTPDX4.DB	Paradox	
pdx	TESTUM7.DB3	Paradox 3.5	
pfs	DEMCHOIC.DOC	First Choice WP	
pfs	ITALIC.DOC	First Choice 3 WP	
pfs	MANUSCRT.DOC	First Choice WP	
pfs	SAMPLE.IWA	IBM Writing Assistant	
pfs	TABS.DOC	First Choice 3 WP	
PFSWrite1	GGGR.pfs	Professional Write 1	
PFSWrite2	GGGR.pfs	Professional Write 2	
pgl	AN2.PGL	HP Graphics Language	
pgl	CRTOON01.PGL	HP Graphics Language	
pic	3D.PIC	Lotus PIC	x
pic	BAR10.PIC	Lotus PIC	
pict	BALL-16M.PCT	Mac PICT2	x
pict	DOODLE.PCT	Mac PICT	
pif	BALLS2.PIF	IBM Picture Interchange Format	x
pif	PAPCOLOR.PIF	IBM Picture Interchange Format	(x)
png	pinata.png	Portable Network Graphics Format	
pntg	PIE.MAC	MacPaint	
pp2	0191-385.PPT	PowerPoint 4.0	
pp2	BLUEBOXC.PPT	Mac PowerPoint 4.0	
pp2	OBJS3.PPT	PowerPoint 3.0	
pp7	POINT12.PPT	PowerPoint 7.0	
pp97	AutoShapes.ppt	Microsoft PowerPoint 97-2004	
pp97	Background.ppt	Microsoft PowerPoint 97-2004	
pp97	Creativity.ppt	Microsoft PowerPoint 97-2004	
pp97	Presentation1.ppt		
pp97	Properties.ppt	Microsoft PowerPoint 97-2004	
pp97	Sample_presentation.ppt	Microsoft PowerPoint 97-2004	
pp97	titleandchart.ppt	Microsoft PowerPoint 97-2004	
ppl	COLFMT.PL	PFS: Plan	
ps	3DIMENSN.AI	Post Script	x
ps	LAYERS.AI	Post Script	x
ps	MASK.AI	Post Script	x
ps	SPIRO2.PS	Post Script	
psd	85x11.PDD	Adobe Photoshop	

psp	Beach Scene.psp	Paint Shop Pro	
psp	Mountain Lake.psp	Paint Shop Pro	
psp	Spider Spin.psp	Paint Shop Pro	
pst	sample.pst	MS outlook personal folder	
qa	TEST8	Q&A Write	
qad	CENTERED.DTF	Q&A Database	
qp6	cashbudget.wb3	Quattro Pro Win 7.0/8.0	
qp6	CHART4.WB2	Quattro Pro Win 6.0	
qp6	LESSON5.WQ2	Quattro Pro 5.0	
qp9	fontsizeand.qpw	Quattro Pro Win 10.0	
qp9	insertcellname.qpw	Quattro Pro Win 11.0	
qp9	insertcolrowsheet.qpw	Quattro Pro Win 10.0	
qp9	qp9SPEEDF.qpw	Quattro Pro Win 9.0	
qp9	wines.qpw	Quattro Pro Win 9.0	
ras	AutumnBear.ras	RAS - Sun Raster	
ras	BIRDS.RAS	RAS - Sun Raster	
ras	PumpkinBear.ras	RAS - Sun Raster	
rbs	MAXCOL2.RBF		x
rbs	MINCOL2.RBF		x
rft	INDENT.RFT	IBM DCA/RFT	
rft	SOFT1.RFT	IBM DCA/RFT	
rfx	BANK2.R2D	Reflex	
rfx	CHARTTEST.R2D	Reflex	
rnd	INTER.RND	AutoShade Rendering	
rnd	KITCHEN.RND	AutoShade Rendering	
rnd	Pushpins.rnd	AutoShade Rendering	
rnd	Robot.rnd	AutoShade Rendering	
rtf	MSRTF.RTF	Rich Text Format	
sam	DPCT.SAM	Samna	
sc5	SC5.CAL	SuperCalc 5	
sdw	goldstar.sdw	Ami Draw	x
sdw	moon.sdw	Ami Draw	x
shw3	A day at the Zoo.SHW	Corel Presentations 7.0 - 12.0	
shw3	arrow_shapes.shw	Corel Presentations 7.0 - 12.0	
shw3	line_spacing.shw	Corel Presentations 7.0 - 12.0	
shw3	lotsofsides.SHW	Corel Presentations 7.0 - 12.0	
shw3	simple.shw	Corel Presentations 7.0 - 12.0	
smd	SMARTDB.DB	Smart DataBase	
sms	SMARTSS.WS	Smart Spreadsheet	
smt	FNOTE1.DOC	SmartWare II	
snap	LOTUSNAP.001	Lotus Screen Snapshot	
soc	DatabaseFunctions.sdc	StarOffice 5.2 Calc	
soi	DisplayQualityB&W.sdd	StarOffice 5.2 Impress	
sow	footendnotes.sdw	StarOffice 5.2 Writer	
spt	SPRINT.SPT	Sprint	
taz	1000FILE.TAR	UNIX Tar	
taz	1000FILE.TAZ	UNIX Compress	
text	text.txt	text	
tga	CHEETA.TGA	Targa	(x)
tga	EARTH.TGA	Targa	(x)
tga	MTRCYCLE.TGA	Targa	(x)
tif6	AQUA.TIF	Tagged Image File Format	(x)
tif6	BALL.TIF	Tagged Image File Format	(x)
tif6	BIKE.EPS	EPS (TIFF Header)	
tif6	BROOK.EPS	EPS (TIFF Header)	
tif6	CITYSCPE.EPS	EPS (TIFF Header)	
tif6	EAGLE256.TIF	Tagged Image File Format	(x)
tw	ED.TW	Total Word	
txt	ADJLNEND.TXT	IBM DisplayWrite 2 or 3	
txt	COLUMN.TXT	IBM DisplayWrite 2 or 3	
txt	DEMOW4.DOC	IBM DisplayWrite 4	
txt	DW5.DOC	IBM DisplayWrite 5	
txt	GGR.txt		
vcrd	CCD development meeting.cal	vCard	
vcrd	Newt_Gingrich.vcf	vCalendar	
viso	2shape.vsd	Visio 6.x	
viso	50 US Existing Access Replication Structure.vsd	Visio 5.x	
viso	aircond.vsd	Visio 4.x	

viso	archscale.vsd	Visio 6.x	
viso	arrows3.vsd	Visio 6.x	
viso	axis.vsd	Visio 6.x	
viso	BlkDia.vsd	Visio 2003	
viso	Dataflow Diagram.VSD	Visio 5.x	
viso	draw.vsd	Visio 6.x	
viso	Fishbone Diagram.VSD	Visio 5.x	
viso	fontchnng.vsd	Visio 6.x	
viso	Jacobson Model.VSD	Visio 5.x	
viso	Shadow-Backcolour.vsd	Visio 6.x	x
viso	starts2.vsd	Visio 6.x	
viso2000	city.vsd	Visio 6.x	
viso2000	gradient.vsd	Visio 6.x	
viso2002	3Dshape.vsd	Visio 6.x	
viso2002	linecust.vsd	Visio 6.x	
viso2003	04 Cross Diagonal.vsd	Microsoft Visio 2003	x
viso2003	OrgChart.vsd	Microsoft Visio 2003	
vw3	01560.VW4	Volkswriter	
w2004Mac	Flyer	Microsoft Word 2003/2004	
w2004Mac	Letterhead	Microsoft Word 2003/2004	
w2004Mac	Newsletters.doc	Microsoft Word 2002	
w6	ALIGN.DOC	Word 6.0 or 7.0	
w6	GRPHBORD.DOC	Word 6.0 or 7.0	
w97	CalendarWizard.doc	Microsoft Word 2000	
w97	Columns.doc	Microsoft Word 2003/2004	
w97	HANGING.DOC	Microsoft Word 97/98	
w97	headfoot.doc	Microsoft Word 2000	
w97	InsertDiagram2.doc	Microsoft Word 2003/2004	
w97	J_Word97_6.DOC	Microsoft Word 97/98	
w97	PNUMBER.DOC	Microsoft Word 97/98	
w97	ProfessionalReport.doc	Microsoft Word 2000	
wbmp	Koala52.wbmp	Wireless Bitmap	X
wg2	123OS22.WG2	Lotus 123 Release 2 for OS/2	
wk3	TIME.WK3	Lotus 1-2-3 3.x	
wk6	Seasons.123	Lotus 1-2-3 97 Edition	
wks	BALLET.WKS	Microsoft Works SS	
wks	CHARS.WKS	Windows Works 4.0 SS	
wks	COMMON.wks	Lotus 1-2-3 1.0	
wks	DISTRICT.WKQ	Quattro	
wks	DOLLAR.SYM	Lotus 1-2-3 2.0	
wks	INVDB.wk1	Lotus 1-2-3 2.0	
wks	INVDB.wks	Lotus 1-2-3 1.0	
wks	INVENT.WDB	Microsoft Works DB	
wks	LABEL.WDB	Windows Works 4.0 DB	
wks	ORDERS.WKS	VP-Planner	
wks	PRODUCT.WKS	VP-Planner	
wks	ROTATE.WR1	Symphony	
wks	SAMPLE.WR1	Symphony	
wks	SMALL.WKS	Microsoft Works SS	
wks	SYMPHONY.WR1	Symphony	
wks	WRK95SS.WKS	Windows Works 4.0 SS	
wm	DEMOWMC.ORG	WordMarc	
wmf	AMHAPPY.WMF	Windows Metafile	x
wmf	ASPDEBUG.WMF	Windows Metafile [5000]	x
wmf	BIRD.WMF	Windows Metafile	(x)
wmf	COFFEE.WMF	Windows Metafile	(x)
wmf	DEMOPWP.G01	Windows Metafile [5000]	x
wmf	preview.wmf	Windows Metafile [5000]	(x)
wml	gmeyertestalltags.wml	Wireless Markup Language (WML)	x
wml	Logo.gif		(x)
wml	Stellent_logo.gif		(x)
word	AFTER.DOC	WinWord 1.x	
word	ANTATION.DOC	Win Word 2.0	
word	COLORS	Mac Word 4.0	
word	COMP11.CLX	Mac Word 4.0	
word	Macword5	Mac Word 5.x	
word	PAPERTIT.CLX	Mac Word 4.0	
word	STYLE	Mac Word 5.x	

word	WRDCHARS	Mac Word 4.0	
work	ALDEMO.WPS	Windows Works 4.0 WP	
work	BALOU_C.WPS	Windows Works 3.0 WP	
work	FORMLETR.WPS	Microsoft Works 2.0	
work	INDENTS.WPS	Microsoft Works 1.0	
work	LETTER.WPS	Microsoft Works 1.0	
wp10	datetime.wpd	WordPerfect 6.1 - 12.0	
wp10	Naming.wpd	WordPerfect 6.1 - 12.0	
wp11	BasicTabs.wpd	WordPerfect 6.1 - 12.0	
wp11	insertGraphic.wpd	WordPerfect 6.1 - 12.0	
wp12	All Font Sizes.wpd	WordPerfect 6.1 - 12.0	
wp12	Colors.wpd	WordPerfect 6.1 - 12.0	
wp5	EGE00B3.WP5	WordPerfect 5.1/5.2	
wp7	CWP7FNT1.WPD	WordPerfect 7.0/8.0/10.0	
wp7	Text.wpd	WordPerfect 7.0/8.0/10.0	
wpf	INDENTS.WPF	WordPerfect 4.2	
wpg	ARROW-22.WPG	WordPerfect Graphic	x
wpg	BICYCLE.WPG	WordPerfect Graphic	
wpg	NEWS.WPG	WordPerfect Graphic	
wpg	THANKS.WPG	WordPerfect Graphic	x
wpg2	COT1.WPG	WordPerfect Graphic 2	
wpg2	FACTORY.WPG	WordPerfect Graphic 2	x
wpg2	Gradient.wpg	WordPerfect Graphic 2	x
wpg2	simple.shw	Corel Presentations 7.0 - 12.0	
wpg2	TREE.WPG	WordPerfect Graphic 2	
wpl	ONEPAGE.WPL	Digital WPS Plus	
wpw	BORDERS3.WPW	PerfectWorks for Windows	
wpw	FILL2.WPW	PerfectWorks for Windows	
wpw	TEST.WPW	PerfectWorks for Windows	
ws	11235.BUG	word star 7.0	
ws	ASCII.WS6	word star 6.0	
ws	BOTMAR.WS	wordstar 5.0	
ws	COLS.WS	wordstar 4.0	
ws2	WS2TEST1.WS2	Wordstar 2000	
xbm	TEST1.XBM	X-Windows Bitmap	
xl2000	Bar.xls	Microsoft Excel 2000	
xl2002	linecharts.xls	Microsoft Excel 2002	
xl2003	column.xls	Microsoft Excel 2003	
xl2004Mac	Investment Calculator1.xls	Microsoft Excel 97/98/2004	
xl2004Mac	Lists	Microsoft Excel 97/98/2004	
xl5	CHART64.XLS	Microsoft Excel 5.0/7.0	
xl5	CHART67.XLS	Microsoft Excel 5.0/7.0	
xl5	MACXL4.XLS	Microsoft Excel 4.0	
xl5	MANY.XLS	Microsoft Excel 3.0	
xl5	mergedataerase2.xls	Microsoft Excel 97/98/2004	
xl5	XL3.XLS	Microsoft Excel 3.0	
xl5	XLS4.XLS	Microsoft Excel 4.0	
xpm	35FLOPPY.XPM		
xwd	SMALL.XWD		
xy	XYMISC	XyWrite / Nota Bene	
zip	PCT.ZIP	.ZIP File	
zip	testzip.zip	.ZIP File	
SELF-CREATED TEST DATA SET			
FOLDER	FILENAME	FILE FORMAT (REPORTED)	
own	Archive.tgz	tar.gz	
own	Book1.xlsx	Microsoft Excel 2010	
own	bookmark.nsf	Lotus Notes 8.5	x
own	busytime.nsf	Lotus Notes 8.5	x
own	imap3521.nsf	Lotus Notes 8.5	x
own	names.nsf	Lotus Notes 8.5	x
own	bookmarks_7_27_12.html	Chrome Bookmarks HTML Export	
own	Database11.accdb	Microsoft Access Database 2010	
own	Desktop.rar	Win RAR Archiv	x
own	draw.odg	OpenDocument Drawing	
own	File.cpp	C++ File	
own	FreeEed.zip	Zip File	

own	FreeEed-1.0-SNAPSHOT.jar	Java jar	
own	FreeEedLogging.class	Java Class	
own	FreeEedLogging.java	Java File	
own	keynote.key	Apple Keynote	
own	mail.box	Mail Archiv Box File	
own	math-test.odf	OpenOffice Math	
own	myImage.dd	DD Forensic Disk Image	
own	New Database.odb	OpenOffice Database	x
own	numbers.numbers	Apple Numbers	
own	OCR.pdf	Manual Scanned PDF Document	x
own	pages.pages	Apple Pages	
own	pom.xml	XML File	
own	Präsentation.odp	OpenDocument Presentation	
own	Privat.ics	iCalendar File	
own	Project1.mpp	Microsoft Project 2010	x
own	Publication1.pub	Microsoft Publishe 2010	
own	run_hadoop_s3.sh	Shell Script	
own	Scan 1.tiff	Manual Scanned tiff Document	x
own	Scan 2.jpeg	Manual Scanned jpg Document	x
own	SPS_2010_Design Sample.vsd	Microsoft Visio 2010	
own	tab.ods	OpenDocument Spreadsheet	
own	Te-eDiscovery.pptx	Microsoft Powerpoint 2010	
own	test.bak	MS SQL 2012 DB Backup	x
own	word.docx	Microsoft Word 2010	
own	writer.odt	OpenDocument Text	
No text extracted			74
Text extracted			344
Total			418

Tabelle 12: Übersicht Fileformattest Nuix

Übersicht der getesteten Fileformate für FreeEed/SHMCloud. Die mit einem 'x' markierten Dateien konnten nicht erkannt werden.

FOLDER	FILENAME	FILE FORMAT (REPORTED)	UNSUPPORTED FILES
acs	YESNO.MDB	MS Access 2.0	x
ami	POINTS.Z.SAM		
bdr	PRINT1.OBD		x
bmp	4seafoodc.cdr	corel draw 3	x
bmp	911.CDR	corel draw 4	x
bmp	AQUA-OS2.BMP	OS/2 bmp	x
bmp	BLUES.BMP	icon	x
bmp	C5MENU.CDR	corel draw 5	x
bmp	EYE.CDR	corel draw 2	x
bmp	JAZZ.BMP	win bmp	x
bmp	mdiframe.ico	OS/2 wrap bmp	x
bmp	PO.cdr	corel draw 6.0	x
bmp	SELECT.CUR	cursor	x
cdr	5radialg.cdr		x
cdr	award.cdr		x
cdr	BOXES.CDR		x
cdr	FROMWPG.CDR		x
cdr	GOODPEAR.CDR		x
cdr	itmenu.cdr		x
cdr	UNIFILL.CDR		x
cgm	APPLE.CGM	computer graphics metafile	x
cmx	NEWSFLSH.CMX	Corel Draw Clipart	x
dbf	BUSINESS.DBF	dbase 3	x
dbf	PICTURES.DBF	dbase 5	x
dbf	SYSTIMES.DBF	dbase 4	x
dif	NAVYDIF1	Navy DIF	
drw	TUTOR07.DRW	Micrografax products	x
dsf	COASTER.DSF	Micrografax designer	x
dwg	3DPolygon.dwg	AC drawing 2000/2002	x
dwg	AIRPLANE.DWG	AC drawing 10	x
dwg	ASHADE.DWG	AC drawing 9	x
dwg	BLKTEST.DWG	AC drawing 12	x
dwg	CHROMA.DWG	AC drawing 13	x
dwg	colorwh.dwg	AC drawing 14	x
dwg	NOZZLE.DWG	AC drawing 2.5	x
dx	DEMOMX41.DX	DX3.1	
dx	TEST1.DX2	DX3.0 and below	x
dx	CHAIRBIN.DXF		x
dx	CHROMA.DXF		x
dx	DOVE1.DXF		
dx	lcaddwg1.dxf		
dx	NOZZLE.DXF		
emf	allshapes.emf	emf	x
emf	BALCONY.EMF	emf	x
emf	ObjectFillStyle.emf	emf	x
en4	CHARFORM.WPF	enable wp 5	x
en4	ENABLE45.WPF		x
ens	FINANCE.SSF	enable ss	x
eshr	dasharrow.rtf	rtf with esher	
exe2	demet.dll	dll	x
exe2	MAKEVID.EXE	exe	x
fax	GROUP3B.FAX		x
fcd	FLDTYPE.FOL	first choice DB	x
fcs	FORMATS.SS	first choice SS	x
fft	BOLD1.FFT	IBM DCA/FFT	x
flw	FONT.SPRZ	freelance	x
flw	PATTERNS.PRE	freelance 96/97/Millennium Edition	x
fmv	123PIE.FMV		x
fmv	KITCHEN.FMV		x
fpx	catbird.fpx		x
fwk	REPORT.FW3		x
gdf	GDF	IBM Graphic Data Format	x
gem	TRUMPET.GEM	Gem file	x

gif	A_BITMAP.GIF	gif	x
gif	bannerimage[1].gif	gif	x
GP4	SCALE.cal		x
gzip	uedit32.ini.gz	gzip	
hgs	ORG2.CH3	Harvard 3.0 Chart	x
hgw	OBJS2.PRS	Harvard Graphics	x
html	GGGR.HTML	html	
hwp	Korean_tab.hwp	hangul 97	x
hwp	LESSON2.HWP	hangul 97	x
hwp2	TextBoxes1.hwp	hangul 2002	x
ich2004	flag.JTT	ichitaro higher 7.0 and up	x
ich2004	header.jtd	ichitaro higher 7.0 and up	x
ich2004	miyazawaKenji.jtt	ichitaro higher 7.0 and up	x
ich2004	overlay.jtd	ichitaro higher 7.0 and up	x
ich2004	welcome.JTD	ichitaro higher 7.0 and up	x
ich6	NUMBER.JBW	ichitaro 5-6	x
igs	F408X.IGS	IGES Drawing	
igs	IGESD.IGS	IGES Drawing	
img	RAZORBK.IMG	GEM image	x
iwp	ASCII.DOC	Wang IWP	x
iwp	CHARFORM.DOC	Wang IWP	x
iwp	INDENTS.DOC	Wang IWP	x
iwp	MANUSC.DOC	Wang IWP	x
iwp	OUTLINE.DOC	Wang IWP	x
jbg2	003288.jb2.pdf	jbig2 embeded pdf	x
jpg	leaf_BG.JPG		x
jw	COLS3.JW	just write 1.0	x
jw	DEMOQA3.QW	Q A write 3.0	x
jw	JWRITE2.JW	just write 2.0	x
l123r9	Place Inside Plot Area.123	lotus 123	x
leg	GGGR.ws5	word star 5.0	x
leg	GUIDE.WSD	word star for windows	x
leg	LEGACY.CHP	legacy	x
lwp	complex.lwp	word pro 97/millennium	x
lzh	H.EXE	Self-Extracting LZH	x
lzh	S.COM	Self-Extracting LZH	x
lzh	WINSOCK.LZH	LZH Compress	x
m11	TEST.AF0	Mass 11	x
manu	SAMPLE7.DOC	Lotus Manuscript 2.0	x
mcw	BOLD	MacWrite II	x
met	AIRPLANE.MET	OS/2 Metafile	x
met	PAPIE.MET	OS/2 Metafile	x
met	USA2.MET	OS/2 Metafile	x
mif	46226.MIF	MIFF 4.0	
mif	MSBUG2.MIF	MIFF 5.5	
mif	process.mif	MIFF 6.0	
mif	TAG2.MIF	MIFF 5.0	
mm	COMTAB.DOC	MultiMate Advantage 2	x
mm	UNDER1.DOC	MultiMate Advantage 2	x
mm4	MATH.DOX	MultiMate 4.0	x
MMFN	BLD2.FNT	MultiMate Note	x
mp	MPLAN4.MOD	Multiplan 4.0	x
mpp	1task3resources.mpp	Microsoft Project 2000/2002/2003	x
mpp	DrawingWithoutTasks.mpp	Microsoft Project 98	x
mpp	ENGINEER.mpp	Microsoft Project 2000/2002/2003	x
mpp	ProjectInfoWithBaseline.mpp	Microsoft Project 2000/2002/2003	x
msg	ManyFields2.msg	MS Outlook	
msg	ManyFields3.msg	MS Outlook	
msg	ManyFields4.msg	MS Outlook	
msg	ManyFields5.msg	MS Outlook	
msg	Pink Note.msg	MS Outlook	
msg	Regular.msg	MS Outlook	
msw	CENTER.DOC	Word for DOS 4.x	x
msw	NARROW.WRI	Windows Write	x
msw	TABLE1.DOC	Word for DOS 6.x	x
msw	WPRD55.DOC	Word for DOS 5.x	x

mwkd	MACWRK2.DB	Mac Works 2.0 DB	x
mwp2	TABLE2	Mac WordPerfect 3.0	x
mwp2	WPM31_C	Mac WordPerfect 3.0	x
mwpf	WPFFORM4	Mac WordPerfect 1.x	x
mwrk	FLRIGHT.MWK	Mac Works 2.0 WP	x
ow	OWTUTOR.WP	OfficeWriter	x
pbm	543x543.ppm	PPM (Portable Pixmap)	x
pbm	A6.ppm	PPM (Portable Pixmap)	x
pbm	d1.pgm	PGM (Portable Graymap)	x
pbm	disk.pbm	PBM (Portable Bitmap)	x
pbm	FL.pgm	PGM (Portable Graymap)	x
pbm	gnu.pbm	PBM (Portable Bitmap)	x
pbm	le.ppm	PPM (Portable Pixmap)	x
pbm	mab.pbm	PBM (Portable Bitmap)	x
pbm	owl.pgm	PGM (Portable Graymap)	x
pbm	tmp.pbm	PGM (Portable Graymap)	x
pcd	ICESTORM.PCD	Kodak Photo CD	x
pcl	PCFILE50.LTR	PC File 5.0 Doc	
pcx	APPLE.PCX	Paintbrush	x
pcx	SKELET4.DCX	Multipage PCX	x
pcx	WATERFAL.PCX	Paintbrush	x
pdf	artistic2.ai	AI 9.0	x
pdf	LEAR.PDF	PDF	x
pdx	ATTFILES.DB	Paradox 2 or 3	x
pdx	BOOKORD.DB	Paradox	x
pdx	NEWRECS.DB	Paradox 2 or 3	x
pdx	SYMBOLS.DB	Paradox	x
pdx	TASKLIST.DB	Paradox 3.5	x
pdx	TESTPDX4.DB	Paradox	x
pdx	TESTUM7.DB3	Paradox 3.5	x
pfs	DEMCHOIC.DOC	First Choice WP	x
pfs	ITALIC.DOC	First Choice 3 WP	x
pfs	MANUSCRT.DOC	First Choice WP	x
pfs	SAMPLE.IWA	IBM Writing Assistant	x
pfs	TABS.DOC	First Choice 3 WP	x
PFSWrite1	GGGR.pfs	Professional Write 1	x
PFSWrite2	GGGR.pfs	Professional Write 2	x
pgl	AN2.PGL	HP Graphics Language	
pgl	CRTOON01.PGL	HP Graphics Language	
pic	3D.PIC	Lotus PIC	
pic	BAR10.PIC	Lotus PIC	x
pict	BALL-16M.PCT	Mac PICT2	x
pict	DOODLE.PCT	Mac PICT	x
pif	BALLS2.PIF	IBM Picture Interchange Format	x
pif	PAPCOLOR.PIF	IBM Picture Interchange Format	x
png	pinata.png	Portable Network Graphics Format	x
pntg	PIE.MAC	MacPaint	x
pp2	0191-385.PPT	PowerPoint 4.0	
pp2	BLUEBOXC.PPT	Mac PowerPoint 4.0	x
pp2	OBJS3.PPT	PowerPoint 3.0	
pp7	POINT12.PPT	PowerPoint 7.0	x
pp97	AutoShapes.ppt	Microsoft PowerPoint 97-2004	
pp97	Background.ppt	Microsoft PowerPoint 97-2004	
pp97	Creativity.ppt	Microsoft PowerPoint 97-2004	
pp97	Presentation1.ppt		
pp97	Properties.ppt	Microsoft PowerPoint 97-2004	
pp97	Sample_presentation.ppt	Microsoft PowerPoint 97-2004	
pp97	titleandchart.ppt	Microsoft PowerPoint 97-2004	
ppl	COLFMT.PL	PFS: Plan	x
ps	3DIMENSN.AI	Post Script	x
ps	LAYERS.AI	Post Script	x
ps	MASK.AI	Post Script	x
ps	SPIRO2.PS	Post Script	x
psd	85x11.PDD	Adobe Photoshop	x
psp	Beach Scene.psp	Paint Shop Pro	x
psp	Mountain Lake.psp	Paint Shop Pro	x

psp	Spider Spin.psp	Paint Shop Pro	x
pst	sample.pst	MS outlook personal folder	
qa	TEST8	Q&A Write	x
qad	CENTERED.DTF	Q&A Database	x
qp6	cashbudget.wb3	Quattro Pro Win 7.0/8.0	x
qp6	CHART4.WB2	Quattro Pro Win 6.0	x
qp6	LESSON5.WQ2	Quattro Pro 5.0	x
qp9	fontsizeand.qpw	Quattro Pro Win 10.0	x
qp9	insertcellname.qpw	Quattro Pro Win 11.0	x
qp9	insertcolrowsheet.qpw	Quattro Pro Win 10.0	x
qp9	qp9SPEEDF.qpw	Quattro Pro Win 9.0	x
qp9	wines.qpw	Quattro Pro Win 9.0	x
ras	AutumnBear.ras	RAS - Sun Raster	x
ras	BIRDS.RAS	RAS - Sun Raster	x
ras	PumpkinBear.ras	RAS - Sun Raster	x
rbs	MAXCOL2.RBF		x
rbs	MINCOL2.RBF		x
rft	INDENT.RFT	IBM DCA/RFT	x
rft	SOFT1.RFT	IBM DCA/RFT	x
rfx	BANK2.R2D	Reflex	x
rfx	CHARTEST.R2D	Reflex	x
rnd	INTER.RND	AutoShade Rendering	x
rnd	KITCHEN.RND	AutoShade Rendering	x
rnd	Pushpins.rnd	AutoShade Rendering	x
rnd	Robot.rnd	AutoShade Rendering	x
rtf	MSRTF.RTF	Rich Text Format	
sam	DPCT.SAM	Samna	x
sc5	SC5.CAL	SuperCalc 5	x
sdw	goldstar.sdw	Ami Draw	x
sdw	moon.sdw	Ami Draw	x
shw3	A day at the Zoo.SHW	Corel Presentations 7.0 - 12.0	x
shw3	arrow_shapes.shw	Corel Presentations 7.0 - 12.0	x
shw3	line_spacing.shw	Corel Presentations 7.0 - 12.0	x
shw3	lotsofsides.SHW	Corel Presentations 7.0 - 12.0	x
shw3	simple.shw	Corel Presentations 7.0 - 12.0	x
smd	SMARTDB.DB	Smart DataBase	x
sms	SMARTSS.WS	Smart Spreadsheet	x
smt	FNOTE1.DOC	SmartWare II	x
snap	LOTUSNAP.001	Lotus Screen Snapshot	x
soc	DatabaseFunctions.sdc	StarOffice 5.2 Calc	x
soi	DisplayQualityB&W.sdd	StarOffice 5.2 Impress	x
sow	footendnotes.sdw	StarOffice 5.2 Writer	x
spt	SPRINT.SPT	Sprint	x
taz	1000FILE.TAR	UNIX Tar	x
taz	1000FILE.TAZ	UNIX Compress	x
text	text.txt	text	
tga	CHEETA.TGA	Targa	x
tga	EARTH.TGA	Targa	x
tga	MTRCYCLE.TGA	Targa	x
tif6	AQUA.TIF	Tagged Image File Format	x
tif6	BALL.TIF	Tagged Image File Format	x
tif6	BIKE.EPS	EPS (TIFF Header)	x
tif6	BROOK.EPS	EPS (TIFF Header)	x
tif6	CITYSCPE.EPS	EPS (TIFF Header)	x
tif6	EAGLE256.TIF	Tagged Image File Format	x
tw	ED.TW	Total Word	x
txt	ADJLNEND.TXT	IBM DisplayWrite 2 or 3	x
txt	COLUMN.TXT	IBM DisplayWrite 2 or 3	x
txt	DEMOW4.DOC	IBM DisplayWrite 4	x
txt	DW5.DOC	IBM DisplayWrite 5	x
txt	GGGR.txt		x
vcrd	CCD development meeting.cal	vCard	
vcrd	Newt_Gingrich.vcf	vCalendar	
viso	2shape.vsd	Visio 6.x	x
viso	50 US Existing Access Replication Structure.vsd	Visio 5.x	x
viso	aircond.vsd	Visio 4.x	x
viso	archscales.vsd	Visio 6.x	x
viso	arrows3.vsd	Visio 6.x	x

viso	axis.vsd	Visio 6.x	
viso	BlkDia.vsd	Visio 2003	
viso	Dataflow Diagram.VSD	Visio 5.x	x
viso	draw.vsd	Visio 6.x	
viso	Fishbone Diagram.VSD	Visio 5.x	x
viso	fontchng.vsd	Visio 6.x	
viso	Jacobson Model.VSD	Visio 5.x	x
viso	Shadow-Backcolour.vsd	Visio 6.x	x
viso	starts2.vsd	Visio 6.x	x
viso2000	city.vsd	Visio 6.x	x
viso2000	gradient.vsd	Visio 6.x	x
viso2002	3Dshape.vsd	Visio 6.x	x
viso2002	linecust.vsd	Visio 6.x	
viso2003	04 Cross Diagonal.vsd	Microsoft Visio 2003	x
viso2003	OrgChart.vsd	Microsoft Visio 2003	x
vw3	01560.VW4	Volkswriter	
w2004Mac	Flyer	Microsoft Word 2003/2004	
w2004Mac	Letterhead	Microsoft Word 2003/2004	
w2004Mac	Newsletters.doc	Microsoft Word 2002	
w6	ALIGN.DOC	Word 6.0 or 7.0	
w6	GRPHBORD.DOC	Word 6.0 or 7.0	
w97	CalendarWizard.doc	Microsoft Word 2000	
w97	Columns.doc	Microsoft Word 2003/2004	
w97	HANGING.DOC	Microsoft Word 97/98	
w97	headfoot.doc	Microsoft Word 2000	
w97	InsertDiagram2.doc	Microsoft Word 2003/2004	
w97	J_Word97_6.DOC	Microsoft Word 97/98	
w97	PNUMBER.DOC	Microsoft Word 97/98	
w97	ProfessionalReport.doc	Microsoft Word 2000	
wbmp	Koala52.wbmp	Wireless Bitmap	x
wg2	123OS22.WG2	Lotus 123 Release 2 for OS/2	x
wk3	TIME.WK3	Lotus 1-2-3 3.x	x
wk6	Seasons.123	Lotus 1-2-3 97 Edition	x
wks	BALLET.WKS	Microsoft Works SS	x
wks	CHARS.WKS	Windows Works 4.0 SS	x
wks	COMMON.wks	Lotus 1-2-3 1.0	x
wks	DISTRICT.WKQ	Quattro	x
wks	DOLLAR.SYM	Lotus 1-2-3 2.0	x
wks	INVDB.wk1	Lotus 1-2-3 2.0	x
wks	INVDB.wks	Lotus 1-2-3 1.0	x
wks	INVENT.WDB	Microsoft Works DB	x
wks	LABEL.WDB	Windows Works 4.0 DB	x
wks	ORDERS.WKS	VP-Planner	x
wks	PRODUCT.WKS	VP-Planner	x
wks	ROTATE.WR1	Symphony	x
wks	SAMPLE.WR1	Symphony	x
wks	SMALL.WKS	Microsoft Works SS	x
wks	SYMPHONY.WR1	Symphony	x
wks	WRK95SS.WKS	Windows Works 4.0 SS	x
wm	DEMOWMC.ORG	WordMarc	x
wmf	AMHAPPY.WMF	Windows Metafile	x
wmf	ASPDEBUG.WMF	Windows Metafile [5000]	x
wmf	BIRD.WMF	Windows Metafile	x
wmf	COFFEE.WMF	Windows Metafile	x
wmf	DEMOPWP.G01	Windows Metafile [5000]	x
wmf	preview.wmf	Windows Metafile [5000]	x
wml	gmeyertestalltags.wml	Wireless Markup Language (WML)	
wml	Logo.gif		x
wml	Stellent_logo.gif		x
word	AFTER.DOC	WinWord 1.x	x
word	ANTATION.DOC	Win Word 2.0	x
word	COLORS	Mac Word 4.0	x
word	COMP11.CLX	Mac Word 4.0	x
word	Macword5	Mac Word 5.x	x
word	PAPERTIT.CLX	Mac Word 4.0	x
word	STYLE	Mac Word 5.x	x
word	WRDCHARS	Mac Word 4.0	x
work	ALDEMO.WPS	Windows Works 4.0 WP	x

work	BALOU_C.WPS	Windows Works 3.0 WP	x
work	FORMLETR.WPS	Microsoft Works 2.0	x
work	INDENTS.WPS	Microsoft Works 1.0	x
work	LETTER.WPS	Microsoft Works 1.0	x
wp10	datetime.wpd	WordPerfect 6.1 - 12.0	x
wp10	Naming.wpd	WordPerfect 6.1 - 12.0	x
wp11	BasicTabs.wpd	WordPerfect 6.1 - 12.0	x
wp11	insertGraphic.wpd	WordPerfect 6.1 - 12.0	x
wp12	All Font Sizes.wpd	WordPerfect 6.1 - 12.0	x
wp12	Colors.wpd	WordPerfect 6.1 - 12.0	x
wp5	EGE00B3.WP5	WordPerfect 5.1/5.2	x
wp7	CWP7FNT1.WPD	WordPerfect 7.0/8.0/10.0	x
wp7	Text.wpd	WordPerfect 7.0/8.0/10.0	x
wpf	INDENTS.WPF	WordPerfect 4.2	x
wpg	ARROW-22.WPG	WordPerfect Graphic	x
wpg	BICYCLE.WPG	WordPerfect Graphic	x
wpg	NEWS.WPG	WordPerfect Graphic	x
wpg	THANKS.WPG	WordPerfect Graphic	x
wpg2	COT1.WPG	WordPerfect Graphic 2	x
wpg2	FACTORY.WPG	WordPerfect Graphic 2	x
wpg2	Gradient.wpg	WordPerfect Graphic 2	x
wpg2	simple.shw	Corel Presentations 7.0 - 12.0	x
wpg2	TREE.WPG	WordPerfect Graphic 2	x
wpl	ONEPAGE.WPL	Digital WPS Plus	x
wpw	BORDERS3.WPW	PerfectWorks for Windows	x
wpw	FILL2.WPW	PerfectWorks for Windows	x
wpw	TEST.WPW	PerfectWorks for Windows	x
ws	11235.BUG	word star 7.0	x
ws	ASCII.WS6	word star 6.0	x
ws	BOTMAR.WS	wordstar 5.0	x
ws	COLS.WS	wordstar 4.0	x
ws2	WS2TEST1.WS2	Wordstar 2000	x
xbm	TEST1.XBM	X-Windows Bitmap	
xl2000	Bar.xls	Microsoft Excel 2000	
xl2002	linecharts.xls	Microsoft Excel 2002	
xl2003	column.xls	Microsoft Excel 2003	
xl2004Mac	Investment Calculator1.xls	Microsoft Excel 97/98/2004	
xl2004Mac	Lists	Microsoft Excel 97/98/2004	
xl5	CHART64.XLS	Microsoft Excel 5.0/7.0	x
xl5	CHART67.XLS	Microsoft Excel 5.0/7.0	x
xl5	MACXL4.XLS	Microsoft Excel 4.0	x
xl5	MANY.XLS	Microsoft Excel 3.0	x
xl5	mergedataerase2.xls	Microsoft Excel 97/98/2004	
xl5	XL3.XLS	Microsoft Excel 3.0	x
xl5	XLS4.XLS	Microsoft Excel 4.0	x
xpm	35FLOPPY.XPM		
xwd	SMALL.XWD		x
xy	XYMISC	XyWrite / Nota Bene	
zip	PCT.ZIP	.ZIP File	
zip	testzip.zip	.ZIP File	
SELF-CREATED TEST DATA SET			
FOLDER	FILENAME	FILE FORMAT (REPORTED)	UNSUPPORTED FILES
own	Archive.tgz	tar.gz	x
own	Book1.xlsx	Microsoft Excel 2010	
own	bookmark.nsf	Lotus Notes 8.5	x
own	busytime.nsf	Lotus Notes 8.5	x
own	imap3521.nsf	Lotus Notes 8.5	x
own	names.nsf	Lotus Notes 8.5	x
own	bookmarks_7_27_12.html	Chrome Bookmarks HTML Export	
own	Database11.accdb	Microsoft Access Database 2010	x
own	Desktop.rar	Win RAR Archiv	x
own	draw.odg	OpenDocument Drawing	
own	File.cpp	C++ File	
own	FreeEed.zip	Zip File	
own	FreeEed-1.0-SNAPSHOT.jar	Java jar	

own	FreeEedLogging.class	Java Class	
own	FreeEedLogging.java	Java File	
own	keynote.key	Apple Keynote	
own	mail.box	Mail Archiv Box File	x
own	math-test.odf	OpenOffice Math	x
own	myImage.dd	DD Forensic Disk Image	x
own	New Database.odb	OpenOffice Database	x
own	numbers.numbers	Apple Numbers	
own	OCR.pdf	Manual Scanned PDF Document	x
own	pages.pages	Apple Pages	
own	pom.xml	XML File	
own	Präsentation.odp	OpenDocument Presentation	
own	Privat.ics	iCalendar File	
own	Project1.mpp	Microsoft Project 2010	x
own	Publication1.pub	Microsoft Publishe 2010	
own	run_hadoop_s3.sh	Shell Script	
own	Scan 1.tiff	Manual Scanned tiff Document	x
own	Scan 2.jpeg	Manual Scanned jpg Document	x
own	SPS_2010_Design Sample.vsd	Microsoft Visio 2010	x
own	tab.ods	OpenDocument Spreadsheet	
own	Te-eDiscovery.pptx	Microsoft Powerpoint 2010	
own	test.bak	MS SQL 2012 DB Backup	x
own	word.docx	Microsoft Word 2010	
own	writer.odt	OpenDocument Text	
		No text extracted	329
		Text extracted	89
		Total	418

Tabelle 13: Übersicht Fileformattest SHMCloud/FreeEed

Anhang 7 - Übersicht MapReduce Jobs

SHMCloud Test mit 10 Nodes:

Job Name: MRFreeEedProcess

Job File: hdfs://ip-10-29-1-44.ec2.internal:8020/mnt/tmp/hadoop/mapred/staging/ubuntu/.staging/job_201209261226_0004/job.xml

Submit Host: ip-10-31-175-39.ec2.internal

Submit Host Address: 10.31.175.39

Job-ACLs: All users are allowed

Job Setup: Successful

Status: Succeeded

Started at: Wed Sep 26 12:34:42 UTC 2012

Finished at: Wed Sep 26 15:02:26 UTC 2012

Finished in: 2hrs, 27mins, 44sec

Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	93	0	0	93	0	0 / 1
reduce	100.00%	1	0	0	1	0	0 / 0

	Counter	Map	Reduce	Total
Job Counters	SLOTS_MILLIS_MAPS	0	0	35,076,519
	Launched reduce tasks	0	0	1
	Total time spent by all reduces waiting after reserving slots (ms)	0	0	0
	Rack-local map tasks	0	0	87
	Total time spent by all maps waiting after reserving slots (ms)	0	0	0
	Launched map tasks	0	0	94
	Data-local map tasks	0	0	7
	SLOTS_MILLIS_REDUCE	0	0	8,729,535
FileSystemCounters	FILE_BYTES_READ	71,000,160,499	93,717,314,573	164,717,475,072
	HDFS_BYTES_READ	17,035	0	17,035
	FILE_BYTES_WRITTEN	124,022,226,923	93,717,384,730	217,739,611,653
	HDFS_BYTES_WRITTEN	0	320,599,169	320,599,169
Map-Reduce Framework	Map input records	93	0	93
	Reduce shuffle bytes	0	53,015,532,330	53,015,532,330
	Spilled Records	1,625,328	1,265,312	2,890,640
	Map output bytes	53,012,593,596	0	53,012,593,596
	CPU time spent (ms)	25,272,190	5,895,740	31,167,930
	Total committed heap usage (bytes)	56,274,518,016	2,107,572,224	58,382,090,240
	Combine input records	0	0	0
	SPLIT_RAW_BYTES	12,453	0	12,453
	Reduce input records	0	715,920	715,920
	Reduce input groups	0	715,920	715,920
	Combine output records	0	0	0
	Physical memory (bytes) snapshot	78,901,170,176	2,236,579,840	81,137,750,016
	Reduce output records	0	715,922	715,922
	Virtual memory (bytes) snapshot	293,843,795,968	3,095,748,608	296,939,544,576
	Map output records	715,920	0	715,920

Abbildung 38: SHMCloud Test mit 10 Nodes

SHMCloud mit Test 1 Nodes:

Job Name: MRFreeEedProcess

Job File: hdfs://ip-10-31-188-212.ec2.internal:8020/mnt/tmp/hadoop/mapred/staging/ubuntu/.staging/job_201209270510_0003/job.xml

Submit Host: ip-10-31-188-212.ec2.internal

Submit Host Address: 10.31.188.212

Job-ACLs: All users are allowed

Job Setup: Successful

Status: Succeeded

Started at: Thu Sep 27 05:12:26 UTC 2012

Finished at: Thu Sep 27 11:53:27 UTC 2012

Finished in: 6hrs, 41mins, 1sec

Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	93	0	0	93	0	0 / 0
reduce	100.00%	1	0	0	1	0	0 / 0

	Counter	Map	Reduce	Total
Job Counters	SLOTS_MILLIS_MAPS	0	0	34,783,509
	Launched reduce tasks	0	0	1
	Total time spent by all reduces waiting after reserving slots (ms)	0	0	0
	Total time spent by all maps waiting after reserving slots (ms)	0	0	0
	Launched map tasks	0	0	93
	Data-local map tasks	0	0	93
	SLOTS_MILLIS_REDUCE	0	0	23,325,886
FileSystemCounters	FILE_BYTES_READ	70,854,195,402	93,494,731,779	164,348,927,181
	HDFS_BYTES_READ	17,314	0	17,314
	FILE_BYTES_WRITTEN	123,876,254,454	93,494,802,296	217,371,056,750
	HDFS_BYTES_WRITTEN	0	320,599,199	320,599,199
Map-Reduce Framework	Map input records	93	0	93
	Reduce shuffle bytes	0	53,015,491,867	53,015,491,867
	Spilled Records	1,650,739	1,266,714	2,917,453
	Map output bytes	53,012,553,133	0	53,012,553,133
	CPU time spent (ms)	21,709,370	7,821,300	29,530,670
	Total committed heap usage (bytes)	55,046,176,768	2,124,349,440	57,170,526,208
	Combine input records	0	0	0
	SPLIT_RAW_BYTES	12,732	0	12,732
	Reduce input records	0	715,920	715,920
	Reduce input groups	0	715,920	715,920
	Combine output records	0	0	0
	Physical memory (bytes) snapshot	73,709,379,584	2,271,203,328	75,980,582,912
	Reduce output records	0	715,922	715,922
	Virtual memory (bytes) snapshot	291,395,440,640	3,092,393,984	294,487,834,624
	Map output records	715,920	0	715,920

Abbildung 39: SHMCloud Test mit 1 Node

Glossar

A

A · *Amazon Web Services ist eine Sammlung verschiedener Webservices, die auf dem Webportal von Amazon.com im Internet als Cloud-Lösung angeboten werden. Amazon Web Services wurde im Juli 2002 als Dienst für andere Webseiten oder Clientseitige Anwendungen gestartet. Die Dienstleistungen werden größtenteils über HTTP transportiert, wobei unter anderem REST und SOAP genutzt werden können.*

Active Directory · *Active Directory (AD) ist der Verzeichnisdienst von Microsoft für Windows-Server und bildet eine zentrale Komponente der Netzwerkverwaltung. Im Active Directory sind die Geräte und deren Attribute tabellarisch abgelegt.*

Amazon S3 · *Amazon Simple Storage Service (S3) ist ein key-value-basierter File Hosting Service. Daten werden dabei in sogenannten Buckets organisiert.*

API · *Eine Programmierschnittstelle (API) ist eine Schnittstelle für den Programmierer, auf der bestimmte interne Funktionsabläufe abstrahiert werden. Eine solche Programmierschnittstelle besteht aus Funktionen, Konstanten und Variablen und stellt Befehle, Routinen und Makros, die von dem Betriebssystem oder einer Betriebssystemerweiterung kommen, als Programmierhilfen bereit. Der Programmierer kann auf Programmbibliotheken zugreifen, ohne den Code für die Anwendungsprogramme ändern zu müssen, vorausgesetzt, die Syntax und Semantik der Programmierschnittstelle stimmt mit denen der Programmbibliothek überein.*

B

Backup · *Backup bezeichnet das Kopieren von Daten in der Absicht, diese im Fall eines Datenverlustes zurückkopieren zu können.*

Big Data · *Als Big Data werden besonders große Datenmengen bezeichnet, die mit Hilfe von Standard-Datenbanken und Daten-Management-Tools nicht oder nur unzureichend verarbeitet werden können. Problematisch sind hierbei vor allem die Erfassung, die Speicherung, die Suche, Verteilung, Analyse und Visualisierung von großen Datenmengen. Das Volumen dieser Datenmengen geht in die Terabytes, Petabytes, Exabytes und Zettabytes.*

Binding Corporate Rules · *Binding Corporate Rules sind verbindliche unternehmensweit geltende Richtlinien, die*

von der Artikel 29-Gruppe entwickelt wurden, um es multinationalen Konzernen, internationalen Organisationen und Firmengruppen zu ermöglichen, intern personenbezogene Daten über Landesgrenzen sowie die EU-Grenze hinweg zu transferieren, und dabei das geltende europäische Datenschutzrecht einzuhalten

Blog · *Blog ist aus der Bezeichnung Weblog abgeleitet, einer Art Web-Tagebuch, das als Webseite geführt wird. Ein solches Logbuch entspricht einem Journal und ist mit Einträgen, Kommentaren und Notizen versehen, die chronologisch geordnet sind: der aktuellste Eintrag befindet sich an oberster Stelle, gefolgt von älteren Einträgen.*

C

CAD · *Unter CAD versteht man das Konstruieren eines Produkts mittels EDV. Ursprünglich wurde mit CAD die Verwendung eines Computers als Hilfsmittel beim technischen Zeichnen bezeichnet.*

Cassandra · *Cassandra ist ein einfaches, verteiltes Datenbankverwaltungssystem für sehr große strukturierte Datenbanken (ein sogenanntes*

Cloud Computing · *Cloud Computing ist ein Modell, das es erlaubt bei Bedarf, jederzeit und überall bequem über ein Netz auf einen geteilten Pool von konfigurierbaren Rechnerressourcen (z.B. Netze, Server, Speichersysteme, Anwendungen und Dienste) zuzugreifen, die schnell und mit minimalem Managementaufwand oder geringer Serviceprovider-Interaktion zur Verfügung gestellt werden können.*

Cluster · *In der Rechnerwelt versteht man unter einem Cluster eine Gruppe miteinander vernetzter eigenständiger Computer, die sich wie ein virtueller Prozessor verhalten. Ein Cluster besteht danach aus mindestens zwei Knoten, die über das Scalable Coherent Interface (SCI), über Gigabit-Ethernet oder Infiniband miteinander verbunden sind.*

Compliance · *Compliance beschreibt in der Unternehmensführung die Einhaltung der gesetzlichen, unternehmensinternen und vertraglichen Regelungen.*

Computer-Forensik · *Computer-Forensik verwendet Software zur Ermittlung allgemeiner krimineller Handlungen und speziell zur Aufdeckung von Computerkriminalität.*

CSV · *Comma Separated Values (CSV) ist ein Textformat, das für den Import von Texten in Spreadsheets und SQL-Datenbanken genutzt wird, ebenso wie für den Export aus*

SQL-Datenbanken. Beim CSV-Format werden die einzelnen Texte einer Zeile durch Kommas oder Semikolon getrennt.

D

DataNode · Data Nodes haben die Slave Rolle und sind für die Verwaltung der Nutzdaten im HDFS auf den einzelnen Clusterknoten verantwortlich.

DDoS · Ein *Distributed Denial of Service (DDoS)* ist eine DoS-Attacke, die im Verbund von vielen Computern aus erfolgt. Die DDoS-Attacke wird zum gleichen Zeitpunkt von verschiedenen Computern ausgelöst und ist dadurch nur schwer zu orten und noch schwieriger zu unterbinden. Wie bei der DoS-Attacke wird der attackierte Computer bei der DDoS-Attacke mit fehlerhaften und vorsätzlich falsch adressierten IP-Paketen bombardiert, und zwar so intensiv, bis der angegriffene Server seinen Dienst einstellen muss oder abstürzt.

Disaster Recovery · Der Begriff *Disaster Recovery* (im Deutschen auch *Katastrophen-Recovery* oder *Notfallwiederherstellung* genannt) bezeichnet Maßnahmen, die nach einem Unglücksfall in der Informationstechnik eingeleitet werden. Dazu zählt sowohl die Datenwiederherstellung als auch das Ersetzen nicht mehr benutzbarer Infrastruktur, Hardware und Organisation. Umfassender als Disaster Recovery ist der Begriff *Business Continuity*, der nicht die Wiederherstellung der IT-Dienste, sondern unterbrechungsfreie Geschäftsabläufe in den Vordergrund stellt.

E

EC2 · Die seit 2006 verfügbare EC2 ist eine Art virtueller Host (genauer: eine Anwendung von Cloud Computing und Infrastructure as a service). EC2-Instanzen können über die Amazon Management Console, aber auch programmatisch über Web Services erstellt werden und sind während ihrer Laufzeit über das Internet verfügbar.

eDiscovery · Electronic discovery (oder e-discovery, eDiscovery) ist der Prozess, um aus einer sehr großen Menge von Daten juristisch relevante Informationen herausfiltern. Hierzu müssen Informationen identifiziert, in Beziehung gesetzt, aufbereitet und zur Verfügung gestellt werden.

Electronic Discovery Reference Model · Das *Electronic Discovery Reference Model* ist eine gemeinsame Initiative von E-Discovery-Experten, Juristen, Anbietern und Endanwendern. Das Modell beschreibt die verschiedenen Phasen eines eDiscovery-Prozesses und ist seit seiner Schaffung im Jahr 2005 der Standard für E-Discovery.

Enterprise-Content-Management · *Enterprise-Content-Management* umfasst die Methoden, Technologien und Werkzeuge zur Erfassung, Verwaltung, Speicherung, Bewahrung und Bereitstellung von Content und Dokumenten zur Unterstützung organisatorischer Prozesse im Unternehmen.

Enterprise-Resource-Planning · Ein *Enterprise-Resource-Planning System* ist eine komplexe Anwendungssoftware zur Unterstützung der Ressourcenplanung eines gesamten Unternehmens.

F

Failover · Ein *Failover* ist der ungeplante Wechsel zwischen zwei oder mehreren Server- oder Netzwerkdiensten bei einem einseitigen Ausfall.

Federal Rules of Civil Procedure · 1938 wurden die Rechtssysteme des Common Law und des Equity durch die Federal Rules of Civil Procedure (deutsch: Bundesregeln zu Zivilprozessen) zusammengelegt.

Flume · Ein verteilter Service um effektiv große Datenmengen zu verschieben. Flume wird hauptsächlich für die Logdateianalyse eingesetzt, dabei werden die Logdateien auf verschiedene Knoten aufgeteilt. Am Ende werden die analysierten Daten via HDFS bereitgestellt.

Fuzzy Search · Die *Fuzzy Search* umfasst in der Informatik eine Klasse von *String-Matching-Algorithmen*, die eine bestimmte Zeichenkette in einer längeren Zeichenkette oder einem Text suchen bzw. finden sollen.

G

Github · *GitHub* ist ein webbasierter Hosting-Dienst für Software-Entwicklungsprojekte.

Google File System · Das *Google File System (GFS)* ist ein proprietäres, auf Linux basierendes verteiltes Dateisystem, das Google für seine Anwendungen benutzt. Es ist für Googles Websuche optimiert. Die Daten sind in teilweise mehrere Gigabyte großen Dateien gespeichert, die selten gelöscht, überschrieben oder verkleinert werden. Auch ist es für hohe Datendurchsätze optimiert.

Graphical user interface · Als *Graphical User Interface (GUI)* bezeichnet man die grafische Benutzeroberfläche, die das Bedienen der Computer erleichtert.

Grid Computing · *Grid Computing* ist eine Form des verteilten Rechnens, bei der ein virtueller Supercomputer aus einem Cluster lose gekoppelter Computer erzeugt wird. Es wurde entwickelt, um rechenintensive Probleme zu lösen.

H

- Haager Beweisübereinkommen** · *Das Haager Beweisübereinkommen, ist ein multilateraler völkerrechtlicher Vertrag, der in Den Haag am 18. März 1970 von den Mitgliedern der Haager Konferenz für Internationales Privatrecht geschlossen wurde. Die Konvention erlaubt zu einem gewissen Umfang Gerichten oder Behörden der Konventionsstaaten bei anderen Konventionsstaaten um Rechtshilfe zu ersuchen.*
- Hadoop** · *Apache Hadoop ist ein freies, in Java geschriebenes Framework für skalierbare, verteilt arbeitende Software. Es basiert auf dem bekannten MapReduce-Algorithmus von Google Inc. sowie auf Vorschlägen des Google-Dateisystems und ermöglicht es, intensive Rechenprozesse mit großen Datenmengen (Big Data, Petabyte-Bereich) auf Computerclustern durchzuführen.*
- Hadoop Distributed Filesystem** · *HDFS ist ein hochverfügbares, leistungsfähiges Dateisystem zur Speicherung sehr großer Datenmengen auf den Dateisystemen mehrerer Rechner (Knoten). Dateien werden in Datenblöcke mit fixer Länge zerlegt und diese redundant auf die teilnehmenden Knoten verteilt. HDFS verfolgt hierbei einen Master-Slave-Ansatz.*
- HBase** · *Ist eine skalierbare, einfache Datenbank zur Verwaltung von großen Datenmengen innerhalb eines Hadoop Clusters. HBase verwendet das HDFS als zugrunde liegenden Speicher. Die HBase Datenbank basiert auf Google BigTable. Die HBase Datenstruktur ist für Daten geeignet, die selten geändert werden, jedoch sehr häufig ergänzt werden.*
- Hive** · *Ein verteiltes Data Warehouse. Hive managet die gespeicherten Daten auf dem HDFS und ermöglicht einfache SQL-ähnliche Abfragen, die in MapReduce Jobs umgewandelt werden.*
- Hosting** · *Ganz allgemein bezeichnet man mit Hosting eine Dienstleistung, bei der Programm-, Informations- und Rechner-Ressourcen den Anwendern zur Verfügung gestellt werden. In früheren Rechnerarchitekturen war der Host der zentrale Rechner, auf dem die Ressourcen abgelegt, verarbeitet und abgerufen werden konnten. Heute wird Hosting in aller Regel mit dem Internet und dem Webhosting assoziiert.*
- HTML** · *Hypertext Markup Language (HTML), ist eine textbasierte Auszeichnungssprache zur Strukturierung von Inhalten wie Texten, Bildern und Hyperlinks in Dokumenten.*
- HTTP** · *Das Hypertext Transfer Protocol (HTTP) ist ein allgemeines, statusloses, objektorientiertes Protokoll zur Datenübertragung im Rahmen des World Wide Web (WWW). Das HTTP-Protokoll ist ein einfaches Protokoll, das*

vom Prinzip her dem Gopher-Protokoll ähnelt. Es ist in RFC 2616 aus dem Jahr 1999 beschrieben und definiert einen Satz von Nachrichten und Antworten, Request/Response, mit denen ein Web-Client und ein Webserver während einer HTML-Sitzung miteinander kommunizieren.

I

- I/O** · *Jedes Rechnerkonzept verfügt über Peripheriegeräte für die Ein- und Ausgabe(I/O) von Daten, Befehlen und Programmen. Da sich beide Vorgänge, die Eingabe und die Ausgabe, wesentlich voneinander unterscheiden, arbeiten Prozessoren mit speziellen Steuerungsverfahren mit denen sie die unterschiedlichen Vorgänge abarbeiten. Diese Steuerungsverfahren können synchron arbeiten oder über Befehle, Unterbrechungen, Absprachen, Rückfragen usw. gesteuert werden.*
- Information Retrieval** · *Information Retrieval, ist ein Fachgebiet, das sich mit computergestütztem Suchen nach komplexen Inhalten (also z. B. keine Einzelwörter) beschäftigt und in die Bereiche Informationswissenschaft, Informatik und Computerlinguistik fällt.*
- IT-Governance** · *IT-Governance besteht aus Führung, Organisationsstrukturen und Prozessen, die sicherstellen, dass die Informationstechnik (IT) die Unternehmensstrategie und -ziele unterstützt.*

J

- Java** · *Java ist eine objektorientierte Programmiersprache und eine eingetragene Marke des Unternehmens Sun Microsystems (seit 2010 Oracle).*
- Job Tracker** · *Der Job Tracker koordiniert die MapReduce Jobs und ist auch zuständig für das Splitten der Aufgaben.*
- JSON** · *Die JavaScript Object Notation (JSON) ist ein kompaktes Format zum Austausch von Daten, welches von Douglas Crockford definiert wurde.*

L

- LDAP** · *Lightweight Directory Access Protocol (LDAP) ist ein TCP/IP-basiertes Directory-Zugangsprotokoll, das sich im Internet und in Intranets als Standardlösung für den Zugriff auf Netzwerk-Verzeichnisdienste für Datenbanken, E-Mails, Speicherbereiche und andere Ressourcen etabliert hat.*
- Load Balancing** · *Load-Balancing ist die Lastverteilung von Übertragungsstrecken und Hardware-Einrichtungen. Dieses optimierte Layer-4-7-Switching wird vorwiegend in Verbindung mit Webswitches in Server-Farmen eingesetzt.*

Dabei wird der Datenverkehr über das Server Load Balancing (SLB) an eine Server-Gruppe verteilt. Entsprechend den Vorgaben und Priorisierungen wird festgelegt, welche Server mit der Bearbeitung der Aufgaben betraut werden.

Lotus Notes · Lotus Notes ist ein dokumentenorientiertes, verteiltes Datenbanksystem mit sehr enger E-Mail-Anbindung.

Lucene · Lucene ist eine Programmbibliothek zur Volltextsuche und ein Projekt der Apache Software Foundation.

M

Machine learning · Machine learning ist ein Oberbegriff für die MapReduce · MapReduce ist ein von Google Inc. eingeführtes Framework für nebenläufige Berechnungen über große (mehrere Petabyte) Datenmengen auf Computerclustern.

MD5 · MD5 ist eine weit verbreitete kryptographische Hashfunktion, die aus einer beliebigen Nachricht einen 128-Bit-Hashwert (Prüfsumme) erzeugt.

Metadaten · Metadaten sind Daten über Daten, oder anders gesagt Informationen zu Daten. Es handelt sich um Daten, die andere Daten beschreiben. In einem Papierarchiv sind die Metadaten Indizes, Detaillisten, Register oder andere Suchhilfen. Bezogen auf Bilddateien sind Metadaten Informationen über die Bilddatei. Es kann sein, dass die Metadaten alle Informationen über die Fotos, aber ebenso Informationen für die Bildverwaltung umfassen.

Mit Hue lassen sich Prozessabläufe in einer webbasierten Benutzeroberfläche automatisieren. Über diese Oberfläche kann man mit anderen Hadoop Tools wie Hive und Pig interagieren. · Mit Hue lassen sich Prozessabläufe in einer webbasierten Benutzeroberfläche automatisieren. Über diese Oberfläche kann man mit anderen Hadoop Tools wie Hive und Pig interagieren.

N

NameNode · Der NameNode wird als Namespace Manager verwendet. Er verwaltet die Metadaten des Dateisystems wie Dateien und Verzeichnisstrukturen und führt Namensraum Operationen wie das Öffnen, Umbenennen und Schließen von Ordnern und Dateien aus.

O

Oozie · Ist ein Server basierendes Workflow Management System für Hadoop. Die Cloudera Version bringt eine Unterstützung für Hive- und Sqoop-Jobs mit.

Open Source · Open Source quelloffen nennt man Software, die unter einer Lizenz steht, die besagt, dass der Quelltext öffentlich zugänglich ist und frei kopiert, modifiziert und verändert wie unverändert weiterverbreitet werden darf.

optische Zeichenerkennung · Texterkennung oder auch Optische Zeichenerkennung Optical Character Recognition, OCR ist ein Begriff aus der Informationstechnik und bezeichnet die automatisierte Texterkennung innerhalb von Bildern.

P

PDF · Das Portable Document Format (PDF) ist ein plattformunabhängiges Dateiformat für Dokumente, das vom Unternehmen Adobe Systems entwickelt und 1993 veröffentlicht wurde.

POI · POI ist Open-Source-Software, die Java-APIs zum Lesen und Schreiben von Dateien im Dateiformat von Microsoft Office wie z. B. Word und Excel bereitstellt. Es startete als Teil des Jakarta-Projekts und ist nun ein Apache-Top-Level-Projekt.

Policies · Policies sind Statuten, Richtlinien oder Verhaltensregeln für die Benutzung von Systemen, Netzwerken, Diensten oder für die Zuordnung von Anwendungen. Policies können technische Spezifikationen, Dienste und deren Merkmale oder die garantierten Übertragungskriterien zwischen Providern und Kunden beinhalten, ebenso die Benutzerrechte regeln.

Product-Lifecycle-Management · Product-Lifecycle-Management bezeichnet ein strategisches Konzept zum Management eines Produktes über seinen gesamten Produktlebenszyklus. Dieses Konzept umfasst sowohl unterstützende IT-Systeme als auch Methoden, Prozesse und Organisationsstrukturen.

PXE · Preboot Execution Environment (PXE) ist ein Industrie-Standard für das Booten von netzwerkasierten Clients. Der PXE-Code befindet sich im BIOS auf der Hauptplatine des Personal Computers und erlaubt es dem PC-Client mit dem Netzwerk-Server zu kommunizieren. Über die PXE-Schnittstelle kann das Betriebssystem des Clients remote gebootet und dieser remote gestartet und konfiguriert werden.

Python · Die Programmiersprache Python wurde Anfang der 90er Jahren von Guido van Rossum in Amsterdam veröffentlicht. Die Sprache wurde im Hinblick auf die englische Komikertruppe Monty Python benannt und ist eine dynamisch typisierte objektorientierte Programmiersprache. Python hat den Ruf, eine einfache Programmiersprache zu sein, d.h. sie ist leichter zu erlernen und zu lesen. Mit Python kann man äußerst komplexe Programme schreiben und entsprechende Probleme lösen.

Die Grenzen von Python liegen dort, wo es auf systemnahe Programmierung bspw. Gerätetreiber oder optimale Geschwindigkeit ankommt. Python ist hier als Interpretersprache - d.h. ihre Programme werden stets zur Laufzeit in Maschinensprache übersetzt - naturgemäß Sprachen wie der Programmiersprache C unterlegen.

R

Remote Procedure Calls · Remote Procedure Call (RPC), ist eine Technik zur Realisierung von Interprozesskommunikation. Sie ermöglicht den Aufruf von Funktionen in anderen Adressräumen.

S

Safe-Harbor-Regelung · Safe Harbor ist eine besondere Datenschutz-Vereinbarung zwischen der Europäischen Union und den Vereinigten Staaten, die es europäischen Unternehmen ermöglicht, personenbezogene Daten legal in die USA zu übermitteln.

Simple Network Management · Das Simple Network Management Protocol (SNMP) erlaubt ein zentrales Netzwerkmanagement für viele Netzwerkkomponenten. Die primären Ziele des SNMP-Protokolls sind die Verringerung der Komplexität der Management-Funktionen, die Erweiterbarkeit des Protokolls und die Unabhängigkeit von irgendwelchen Netzwerkkomponenten.

SLA · Der Begriff Service-Level-Agreement (SLA), bezeichnet eine Vereinbarung bzw. die Schnittstelle zwischen Auftraggeber und Dienstleister für wiederkehrende Dienstleistungen. Ziel ist es, die Kontrollmöglichkeiten für den Auftraggeber transparent zu machen, indem zugesicherte Leistungseigenschaften wie etwa Leistungsumfang, Reaktionszeit und Schnelligkeit der Bearbeitung genau beschrieben werden. Wichtiger Bestandteil ist hierbei die Dienstgüte (Servicelevel), welche die vereinbarte Leistungsqualität beschreibt.

SOA · Serviceorientierte Architektur (SOA), ist ein Architekturmuster der Informationstechnik aus dem Bereich der verteilten Systeme, um Dienste von IT-Systemen zu strukturieren und zu nutzen. Eine besondere Rolle spielt dabei die Orientierung an Geschäftsprozessen, deren Abstraktionsebenen die Grundlage für konkrete Serviceimplementierungen sind.

Solr · Solr ist ein Open Source Enterprise Search Server, der auf Lucene basiert. Solr wurde von CNET entwickelt und wird heute noch auf deren Website eingesetzt, des Weiteren wird Solr bei Zappos, Netflix und zahlreichen anderen

Staatlichen und Unternehmenswebseiten eingesetzt. Der Name stammt aus dem Akronym von Search on Lucene and Resin. Solr ist in Java geschrieben und kommuniziert über Standardprotokolle wie HTTP, XML und JSON.

SQL · Structured Query Language (SQL) ist eine spezialisierte Programmiersprache für den Zugriff auf Datenbanksysteme.

SQL Injections · SQL-Injection ist eine Technik mit der Sicherheitslücken von SQL-Datenbanken ausgenutzt werden. Mit dieser Technik können Angreifer unberechtigt auf Datenbanken und ihre Daten zugreifen, diese ausspähen, manipulieren und mit eingeschleusten Datenbankbefehlen unter ihre Kontrolle bringen. Ansatzpunkte für das Einschleusen von eigenen Befehlen sind Lücken in den Benutzereingaben und deren mangelnde Überprüfung.

Sqoop · Ein Tool um effizient Daten zwischen relationalen Datenbanken und HDFS zu verschieben.

T

Task Tracker · Der Task Tracker führt die MapReduce Aufgaben aus.

V

Virtual Private Network · Generell handelt es sich bei einem Virtual Private Network (VPN) um ein geschlossenes, logisches Netz, das auf unterschiedlichen Schichten des OSI-Referenzmodells aufsetzt und für eine bestimmte Benutzergruppe etabliert wird. VPN setzen in der Regel auf den Schichten 2 oder 3 des OSI-Referenzmodells auf und verwenden Tunneling-Mechanismen für den IP-Verkehr.

W

Web 2.0 · Web 2.0 ist ein Schlagwort, das für eine Reihe interaktiver und kollaborativer Elemente des Internets, speziell des World Wide Webs, verwendet wird. Hierbei konsumiert der Nutzer nicht nur den Inhalt, er stellt als Prosument selbst Inhalt zur Verfügung.

Wiki · Wiki ist eine Web-Technologie zur Erstellung und Pflege von Inhalten und zur Kollaboration. Dabei wird das Wissen von vielen Menschen gesammelt und strukturiert. Jeder, der sich berufen fühlt zu einem Thema mit seinem Wissen beizutragen, kann dies unmittelbar tun. Die Autoren und deren textliche Einträge, Ergänzungen und Änderungen werden zeitlich registriert und können von jedem Besucher nachvollzogen werden.

Wildcard · Wildcard ist ein Fachausdruck aus dem Computer-Bereich und bezeichnet einen Platzhalter für andere Zeichen.

Windows Management Instrumentation · Windows Management Instrumentation (WMI) ist eine Management-Infrastruktur für die Verwaltung von Windows-basierten Systemen. Die WMI-Infrastruktur bietet eine einheitliche Schnittstelle für das Monitoring von Windows-Systemen über die auf Systeminformationen von lokalen Windows-PCs aber auch auf Remote-PCs zugegriffen und Konfigurationen geändert werden können.

Wirr · Ziel von Wirr ist die Bereitstellung von einheitlichen Tools für den Betrieb Cloudbasierender Services. Dabei wird der einheitliche Zugriff auf die Kommandozeiletools und Programmierschnittstellen der verschiedenen Cloud Provider ermöglicht.

X

XML · XML ist eine Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten in Form von Textdateien. XML wird u. a. für den plattform- und implementationsunabhängigen Austausch von Daten zwischen Computersystemen eingesetzt, insbesondere über das Internet.

Z

ZIP · Bei dem ZIP-Dateiformat handelt es sich um ein Kompressionsverfahren mit dem Texte und Grafiken komprimiert und archiviert werden können.

ZooKeeper · Ein verteilter hochverfügbarkeits Koordinierungsservice, der Konfigurationen für andere Systeme bereitstellt.

Literaturverzeichnis

- [1] T-Systems, "White Paper Electronic Discovery .," pp. 4, 6, 10, 16.
- [2] J. Wiles, *Techno security's guide to e-discovery and digital forensics*. Burlington, MA: Syngress, 2007, pp. 95–96, 115–119, 215–216, 232.
- [3] L. Volonino and I. Redpath, *e-Discovery for Dummies*. Hoboken, NJ: For Dummies, 2009, pp. 14, 92, 157–158.
- [4] P. Mell and T. Grance, "The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology," 2011.
- [5] D. Mausberger, O. Salzmann, C. Trauner, and M. Wimmer, "iX | Elektronische Informationsbeschaffung - Sammelsurium," pp. 98 – 102, 2012.
- [6] C. Arthur, "Government to set up its own cloud computing system | Technology | guardian.co.uk," 2010. [Online]. Available: <http://www.guardian.co.uk/technology/2010/jan/27/cloud-computing-government-uk>. [Accessed: 19-Apr-2012].
- [7] Amazon, "What is AWS?," 2012. [Online]. Available: <http://aws.amazon.com/what-is-aws/>. [Accessed: 19-Jun-2012].
- [8] Amazon, "Amazon Elastic Compute Cloud (Amazon EC2)," 2012. [Online]. Available: <http://aws.amazon.com/ec2/>. [Accessed: 19-Oct-2012].
- [9] Apache Hadoop, "ProjectDescription - Hadoop Wiki," 2009. [Online]. Available: <http://wiki.apache.org/hadoop/ProjectDescription>. [Accessed: 19-Oct-2012].
- [10] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, pp. 1–13, 2008.
- [11] National Center for State Courts, "Electronic Discovery: Questions and Answers," *Civil Action*, vol. 3, no. 2, p. 9, 2004.
- [12] M. Kerzner, J. Wade, P. Kerr, S. Wiesel, A. Rosenblatt, and Y. Kerzner, "FreeEed.org," 2012. [Online]. Available: <http://freeeed.org/>. [Accessed: 12-Sep-2012].
- [13] E. Muka, "Electronic-Discovery," pp. 1–3, 5–6.
- [14] United States Courts, "Federal Rules of Civil Procedure," 2012. [Online]. Available: <http://www.uscourts.gov/RulesAndPolicies/FederalRulemaking/RulesAndForms.aspx>. [Accessed: 11-May-2012].
- [15] L. Volonino and I. Redpath, "E-Discovery For Dummies Cheat Sheet," 2012. [Online]. Available: <http://www.dummies.com/how-to/content/ediscovery-for-dummies-cheat-sheet.html>. [Accessed: 11-May-2012].
- [16] M. H. Hartmann, *Internationale E-Discovery und Information Governance*. Schmidt Erich Verlag, 2011, pp. 2–3, 10, 16, 24–26, 29, 59–66, 77, 119, 146, 158, .

- [17] P. Groot, "Electronically Stored Information: Balancing Free Discovery with Limits on Abuse," *Duke L. & Tech. Rev.*, vol. 4, 2009.
- [18] Bundesamt für Sicherheit in der Informationstechnik, "Leitfaden „IT-Forensik“,“ vol. 1, no. März, 2011.
- [19] ZyLab, "Profitieren Sie vom Mehrwert Ihrer papierbasierten Informationen," 2012. [Online]. Available: <http://www.zylab.de/Lösungen/DigitalisierenErfassen.aspx>. [Accessed: 25-May-2012].
- [20] H. Damskis and A. Haidinger, "Siemens: eDiscovery Wenn globale Unternehmen unter Beweisdruck geraten," pp. 2–4, 2011.
- [21] B. Weissmann, "E-Discovery zur Aufdeckung von Wirtschaftskriminalitätsfällen," in *IT-Compliance und E-Discovery in der österreichischen Praxis*, 2011, p. 4.
- [22] I. Geis and C. Klas, "Der Beweiswert elektronischer Dokumente im Austausch mit den USA Ein besonderer Aspekt des deutschen Datenschutzrechts," pp. 7–8, 2009.
- [23] R. H. Jones, J. Mencken, E. Logan, A. Walcoff, S. Johnson, and A. Brown, "SANCTIONS FOR E-DISCOVERY VIOLATIONS: BY THE NUMBERS," p. 861, 2010.
- [24] E. Muka, "Electronic-Discovery Präsentation." p. 6, 2011.
- [25] D. Campbell, "E-DISCOVERY: ETHICAL RULES REMAIN UNCHANGED," 2008. [Online]. Available: <http://www.theiplawblog.com/archives/-cyberspace-law-ediscovery-ethical-rules-remain-unchanged.html>. [Accessed: 29-May-2012].
- [26] Haager Übereinkommen, "Haager Übereinkommen," 2012. [Online]. Available: http://www.hcch.net/index_en.php?act=conventions.text&cid=82. [Accessed: 25-May-2012].
- [27] S. Hanloser, "Datenschutzrechtliche Probleme und Lösungen," vol. 29, pp. 785–789, 2008.
- [28] compliancemagazin, "eDiscovery und Sanktionen der US-Gerichte," 2008. [Online]. Available: <http://www.compliancemagazin.de/gesetzstandards/usa/luther280708.html>. [Accessed: 30-Mar-2012].
- [29] M. Piontek and K. Scheurer, "Symantec-Studie: Anwälte verlieren Fälle durch mindere Qualität elektronischer Beweissicherung," 2012. [Online]. Available: http://www.symantec.com/de/de/about/news/release/article.jsp?prid=20100907_01. [Accessed: 29-May-2012].
- [30] E. Casey, "Handbook of digital forensics and investigation," pp. 63–64, 78, 86, 92, 106, 126–127, 2009.
- [31] Funk Wolfram, W. Alge, F. Decker, and A. Tsoikas, "Compliance versus E-Discovery," 2010. [Online]. Available: <http://sectank.net/?p=702>. [Accessed: 30-Mar-2012].
- [32] Amazon.de, "Samsung Spinpoint F3 HD103SJ 1TB 8,9 cm," 2012. [Online]. Available: http://www.amazon.de/Samsung-Spinpoint-HD103SJ-interne-Festplatte/dp/B002MQC0P8/ref=sr_1_4?ie=UTF8&qid=1338631589&sr=8-4. [Accessed: 02-Jun-2012].
- [33] EDRM, "EDRM Stages « The Electronic Discovery Reference Model," 2012. [Online]. Available: <http://www.edrm.net/resources/edrm-stages-explained>. [Accessed: 27-Feb-2012].
- [34] EDRM, "Information Governance Reference Model (IGRM) Guide « The Electronic Discovery Reference Model," 2012. [Online]. Available: <http://www.edrm.net/resources/guides/igrm>. [Accessed: 14-Jun-2012].

- [35] IBM, "IBM InfoSphere eDiscovery Analyzer 2.1.1," pp. 1–4, 2009.
- [36] C. Reyes, "US Discovery-EU Privacy Directive Conflict: Constructing a Three-Tiered Compliance Strategy, The," vol. 361, pp. 357–387, 2008.
- [37] Österreichische Datenschutzkommission, "Musterverträge : Standardvertragsklauseln für die Übermittlung personenbezogener Daten in Drittländer," 2012. [Online]. Available: <https://www.dsk.gv.at/site/6208/default.aspx>. [Accessed: 04-Jun-2012].
- [38] H. Özbek, "Datenschutz konformer Einsatz von E-Discovery Systemen," pp. 576–580, 2010.
- [39] Clearwell, "EDRM Model — Clearwell," 2011. [Online]. Available: <http://www.clearwellsystems.com/e-discovery-101/edrm-project.php>. [Accessed: 06-Jun-2012].
- [40] D. Logan and J. Bace, "Magic Quadrant for E-Discovery Software," May 2011.
- [41] Microsoft, "Microsoft Office 365 - Komplettlösung für kleine und mittelständische Unternehmen," 2012. [Online]. Available: <http://www.microsoft.com/de-at/office365/online-software.aspx>. [Accessed: 10-Jun-2012].
- [42] ZyLab, "ZyLAB stellt neuen eDiscovery Cloud Collector vor," 2012. [Online]. Available: <http://www.zylab.de/NewsEvents/PressInformationen/tabid/2846/articleType/ArticleView/articleId/142/ZyLAB-stellt-neuen-eDiscovery-Cloud-Collector-vor.aspx>. [Accessed: 10-Jun-2012].
- [43] X1 Discovery, "X1 Discovery delivers next generation eDiscovery for social media and effectively supports investigations of cloud-based data," 2012. [Online]. Available: <http://www.x1discovery.com/>. [Accessed: 22-Jun-2012].
- [44] AccessData Group, "The EDRM Buyer's Guide," pp. 1–14, 2011.
- [45] GRR, "GRR Rapid Response Framework (GRR)," 2012. [Online]. Available: <https://sites.google.com/site/grrresponserig/>. [Accessed: 14-Jun-2012].
- [46] M. I. Cohen, D. Bilby, and G. Caronni, "Distributed forensics and incident response in the enterprise," *ScienceDirect*, vol. 8, p. 109, Aug. 2011.
- [47] LexisNexis, "LexisNexis E-Discovery Solutions," 2012. [Online]. Available: <http://www.lexisnexis.com/docs/BREDiscovery-150430.pdf>.
- [48] EDRM, "XML « The Electronic Discovery Reference Model," 2012. [Online]. Available: <http://www.edrm.net/projects/xml>. [Accessed: 14-Jun-2012].
- [49] Symantec, "Symantec Transparent Predictive Coding Cost-Effective and Defensible Technology Assisted Review," pp. 1–12.
- [50] Google, "Google Search by Image," 2012. [Online]. Available: <http://www.google.com/insidesearch/features/images/searchbyimage.html>. [Accessed: 15-Jun-2012].
- [51] John Resnick, *The Applied Discovery BLACK LETTER BOOK*. 2012, pp. 2–6.
- [52] M. Kerzner and M. Miersebach, "Interview with Mark Kerzner," 2012.
- [53] Microsoft technet, "Windows BitLocker-Laufwerkverschlüsselung," 2010. [Online]. Available: <http://technet.microsoft.com/de-de/library/cc731549>. [Accessed: 18-Jun-2012].

- [54] S. Biggs and S. Vidalis, "Cloud computing: The impact on digital forensic investigations," in *Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for*, 2009, pp. 1–6.
- [55] M. Meir-Huber, "Cloud Computing: Praxisratgeber und Einstiegsstrategien," vol. 1, 2010.
- [56] BITKOM, "Cloud Computing-Evolution in der Technik, Revolution im Business," *Berlin: BITKOM*, pp. 8, 21–32, 69, 71, 2009.
- [57] Bundesamt für Sicherheit in der Informationstechnik, "BSI: Cloud Computing Grundlagen," 2012. [Online]. Available: https://www.bsi.bund.de/DE/Themen/CloudComputing/Grundlagen/Grundlagen_node.html. [Accessed: 19-Jun-2012].
- [58] A. Singh, "An Introduction to Virtualization," 2004. [Online]. Available: <http://www.kernelthread.com/publications/virtualization/>. [Accessed: 19-Oct-2012].
- [59] R. Buyya, J. Broberg, and A. Goscinski, "CLOUD COMPUTING Principles and Paradigms," pp. 10–13, 2011.
- [60] R. Buyya and S. Venugopal, "A gentle introduction to grid computing and technologies," *database*, no. July, 2005.
- [61] T. O'Reilly, "Web 2.0 Compact Definition: Trying Again - O'Reilly Radar," 2006. [Online]. Available: <http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html>. [Accessed: 19-Oct-2012].
- [62] T. O'Reilly, "What Is Web 2.0 - O'Reilly Media," 2005. [Online]. Available: <http://oreilly.com/web2/archive/what-is-web-20.html>. [Accessed: 19-Oct-2012].
- [63] P. Bianco, R. Kotermanski, and P. Merson, "Evaluating a service-oriented architecture," no. September, 2007.
- [64] P. Horn, "Autonomic computing: IBM's Perspective on the State of Information Technology," 2001.
- [65] J. Kephart and D. Chess, "The vision of autonomic computing," *Computer*, no. January, pp. 41–50, 2003.
- [66] M. Meir-Huber, *Cloud Computing Praxisratgeber und Einstiegsstrategien*, 2nd ed. Frankfurt am Main: entwickler.Press, 2010, pp. 28–30.
- [67] Microsoft, "The Business of IT - Cloud Services," 2012. [Online]. Available: <http://www.microsoft.com/austria/enterprise/article.aspx?Id=IaaS+PaaS+und+SaaS>. [Accessed: 19-Jun-2012].
- [68] Bloomberg Businessweek, "Jeff Bezos' Risky Bet," 2012. [Online]. Available: http://www.businessweek.com/magazine/content/06_46/b4009001.htm. [Accessed: 21-Jun-2012].
- [69] M. Armbrust, A. Fox, R. Griffith, A. Joseph, and R. Katz, "Above the clouds: A berkeley view of cloud computing," p. 6, Feb. 2009.
- [70] Amazon, "Global Infrastructure Amazon AWS," 2012. [Online]. Available: <http://aws.amazon.com/about-aws/globalinfrastructure/>. [Accessed: 21-Jun-2012].

- [71] E. Keller-Stoltenhoff, "GULP - Knowledge Base: Das Servicelevelagreement," 2010. [Online]. Available: <http://www.gulp.de/kb/lwo/vertrag/servicelevelagreement.html>. [Accessed: 19-Oct-2012].
- [72] M. Liedtke, "Aufregung um AGB von Cloud-Dienst Google Drive," 2012. [Online]. Available: <http://www.welt.de/wirtschaft/webwelt/article106230913/Aufregung-um-AGB-von-Cloud-Dienst-Google-Drive.html>. [Accessed: 23-Jun-2012].
- [73] B. Murphy, "e-Discovery in The Cloud Not As Simple As You Think - Forbes," *Forbes*, 29-Nov-2011. [Online]. Available: <http://www.forbes.com/sites/jasonvelasco/2011/11/29/e-discovery-in-the-cloud-not-as-simple-as-you-think/>. [Accessed: 28-Feb-2012].
- [74] D. Srocke and F. Karlstetter, "Archivierung und eDiscovery per Cloud," 2012. [Online]. Available: <http://www.searchcloudcomputing.de/sicherheit/risk-management/articles/359681>. [Accessed: 22-Jun-2012].
- [75] T. Harbert, "E-discovery in the Cloud - Computerworld," 2012. [Online]. Available: http://www.computerworld.com/s/article/9226375/E_discovery_in_the_Cloud. [Accessed: 22-Jun-2012].
- [76] B. Murphy, "eDiscovery Coming To The Cloud - eDiscovery Journal," 2012. [Online]. Available: <http://ediscoveryjournal.com/2012/04/ediscovery-coming-to-the-cloud/>. [Accessed: 22-Jun-2012].
- [77] ISO, "ISO/IEC 27031:2011 Information technology -- Security techniques -- Guidelines for information and communication technology readiness for business continuity," 2011.
- [78] SAS 70, "SAS 70 OVERVIEW," 2012. [Online]. Available: http://sas70.com/sas70_overview.html. [Accessed: 19-Oct-2012].
- [79] CaseCentral, "Blueprint for Cloud-Based eDiscovery – On Demand Technology," 2012. [Online]. Available: <http://www.casecentral.com/cloudblueprint.php>. [Accessed: 28-Feb-2012].
- [80] J. Venner, *Pro Hadoop*. New York: Springer, 2009, pp. 1, 6, 18–20, 23.
- [81] R. Grimm, "Funktionale Programmierung (3): Das MapReduce-Framework 'Hallo Worker! Ein Petabyte Daten bitte.'," 2009. [Online]. Available: <http://www.linux-magazin.de/Online-Artikel/Funktionale-Programmierung-3-Das-MapReduce-Framework>. [Accessed: 08-Aug-2012].
- [82] S. Bethge and A. Rheinländer, "MapReduce," Humboldt Universität zu Berlin, 2009.
- [83] N. Kuhnt, "Das MapReduce-Framework," Friedrich-Schiller-Universität Jena.
- [84] R. Wartala, *Hadoop: Zuverlässige, verteilte und skalierbare Big-Data-Anwendungen [Broschiert]*. Open Source Press; Auflage: 1, 2012, pp. 17, 22–23, 58–59, 95.
- [85] G. Czajkowski, "Sorting 1PB with MapReduce | Official Google Blog," 2008. [Online]. Available: <http://googleblog.blogspot.co.at/2008/11/sorting-1pb-with-mapreduce.html>. [Accessed: 30-Jun-2012].
- [86] T. White, *Hadoop: The definitive guide*. Sebastopol: Yahoo Press, 2010, pp. 9–12, 168.
- [87] C. Lam, *Hadoop in Action*. Stamford: Manning Publications, 2011, p. 264.
- [88] Cloudera, "CDH3 Installation - Cloudera Support," 2012. [Online]. Available: <https://ccp.cloudera.com/display/CDHDOC/CDH3+Installation>. [Accessed: 02-Jul-2012].
- [89] A. Pavel, "HIVE," Universität Stuttgart, 2011.

- [90] M. Hoecker, "Hadoop as a Service (HaaaS) auf Basis von Eucalyptus und Cloudera," Hochschule Mannheim, 2010.
- [91] T. Findling, "Hadoop," Universität Leipzig, 2010.
- [92] Cloudera, "Apache Hadoop Ecosystem | Apache Hadoop for the Enterprise | Cloudera," 2012. [Online]. Available: <http://www.cloudera.com/resource/apache-hadoop-ecosystem/>. [Accessed: 04-Jul-2012].
- [93] S. Netke and R. Kalakota, "What is a 'Hadoop' ? - Explaining Big Data to the C-Suite," 2011. [Online]. Available: <http://www.saama.com/blog/bid/71286/What-is-a-Hadoop-Explaining-Big-Data-to-the-C-Suite>. [Accessed: 03-Jul-2012].
- [94] Cloudera, *Cloudera Manager Free Edition Installation Guide*. 2012, pp. 2–4.
- [95] Legaltechtoday, "Open Source Could Change the Future of E-Discovery," 2011. [Online]. Available: <http://www.legaltechtoday.com/2011/06/14/open-source-could-change-the-future-of-e-discovery/>. [Accessed: 30-Mar-2012].
- [96] R. Gupta and S. Ahmed, "Project Proposal Apache Tika," pp. 1–4.
- [97] C. A. Mattmann and J. L. Zitting, *Tika in Action*. New York: Manning Publications Co., 2012, p. 15.
- [98] E. Hatcher and O. Gospodnetic, *Lucene in action*. New York: Manning Publications Co., 2004, pp. 20–22.
- [99] D. Smiley and E. Pugh, *Apache Solr 3 Enterprise Search Server*. Birmingham: Packt Publishing, 2009, pp. 7–10.
- [100] Apache Tika, "Apache Tika - Supported Document Formats," 2012. [Online]. Available: http://tika.apache.org/0.9/formats.html#HyperText_Markup_Language. [Accessed: 04-Jul-2012].
- [101] SHMsoft, "SHMcloud 2012 Private Beta AWS S3," pp. 2–4, 2012.
- [102] SHMsoft, "SHMcloud™ Pricing and Support," 2012. [Online]. Available: <http://shmsoft.com/index.php/shmcloud/shmcloud-pricing-support>. [Accessed: 21-Jul-2012].
- [103] Mark Kerzner, "SHMsoft, Inc. Revolutionary Law Tools." Houston, pp. 1–16, 2012.
- [104] DataStax, "DataStax Enterprise | DataStax," 2012. [Online]. Available: <http://www.datastax.com/products/enterprise>. [Accessed: 08-Aug-2012].
- [105] M. Kerzner, "FreeEed-Open Source Big Data Solution for eDiscovery," pp. 1–6, 2012.
- [106] Apache, "The Apache Cassandra Project," 2012. [Online]. Available: <http://cassandra.apache.org/>. [Accessed: 18-Sep-2012].
- [107] Actuate, "Actuate und DataStax: Neue Big-Data-Partnerschaft," 2012. [Online]. Available: <http://www.actuate.com/ger/unternehmen/neuigkeiten/pressemitteilungen-ressourcen/?articleid=24078>. [Accessed: 08-Jul-2012].
- [108] EMC, "Driving ROI with In-House eDiscovery Solutions Information Volume and Complexity are Exploding," p. 11, 2009.

- [109] W. Cohen, "Enron Email Dataset," 2009. [Online]. Available: <http://www.cs.cmu.edu/~enron/>. [Accessed: 06-May-2012].
- [110] B. Seidl, "Enron - Vor 10 Jahren ging das Licht aus - Meldungen - boerse.ARD.de," 2011. [Online]. Available: <http://boerse.ard.de/meldungen/enron-vor-10-jahren-ging-das-licht-aus-100.html>. [Accessed: 08-May-2012].
- [111] Amazon, "Amazon EC2 Instance Types," 2012. [Online]. Available: <http://aws.amazon.com/ec2/instance-types/>. [Accessed: 28-Sep-2012].
- [112] Tesseract OCR, "Tesseract-ocr - How to run UNLV tests on Tesseract. - An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google. - Google Project Hosting," 2012. [Online]. Available: <http://code.google.com/p/tesseract-ocr/wiki/TestingTesseract>. [Accessed: 08-Aug-2012].
- [113] ISACA, "Big Data Hadoopand Security Outline of presentation for ISACA on 09/20/2012." p. 1, 2012.
- [114] Hortonworks, "Motivations for Apache Hadoop Security | Hortonworks," 2012. [Online]. Available: <http://hortonworks.com/blog/motivations-for-apache-hadoop-security/>. [Accessed: 08-Aug-2012].
- [115] J. Vijayan, "IT must prepare for Hadoop security issues - Computerworld," 2012. [Online]. Available: http://www.computerworld.com/s/article/9221652/IT_must_prepare_for_Hadoop_security_issues. [Accessed: 08-Aug-2012].
- [116] T. Renner, M. Vetter, S. Rex, and H. Kett, *Open Source Software Einsatzpotenziale und Wirtschaftlichkeit*. Stuttgart: , 2005, pp. 59–63.
- [117] F. Benthin, "Canonical bringt Metal as a Service (MAAS) - Pro-Linux," 2012. [Online]. Available: <http://www.pro-linux.de/news/1/18238/canonical-bringt-metal-as-a-service-maas.html>. [Accessed: 13-Jul-2012].
- [118] Google, "guava-libraries - Guava: Google Core Libraries for Java 1.6+ - Google Project Hosting," 2012. [Online]. Available: <http://code.google.com/p/guava-libraries/>. [Accessed: 08-Aug-2012].
- [119] Apache, "Maven - Welcome to Apache Maven," 2012. [Online]. Available: <http://maven.apache.org/>. [Accessed: 08-Jul-2012].
- [120] Amazon, "Exporting EC2 Instances - Amazon Elastic Compute Cloud," 2012. [Online]. Available: <http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/ExportingEC2Instances.html>. [Accessed: 08-Sep-2012].
- [121] R. Sosinski, "Starting Amazon EC2 with Mac OS X -," 2008. [Online]. Available: <http://www.robertsosinski.com/2008/01/26/starting-amazon-ec2-with-mac-os-x/>. [Accessed: 19-Sep-2012].