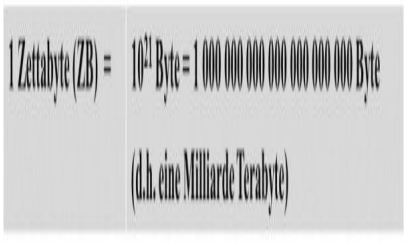
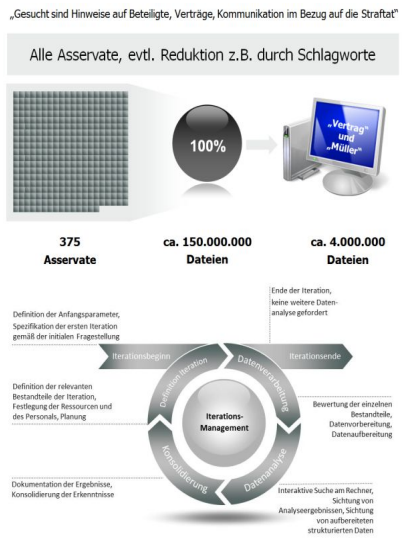
Im Einsatz – im Thema.
POLIZEI PRAXIS

ANALYSE VON GROSSEN UND KOMPLEXEN DATENMENGEN





1 DIE WELT DER DATEN – DIE DATEN DER WELT

Wir leben in einer Welt, in der sich die Technik immens schnell weiterentwickelt. Eben wurde noch über das Festnetz kommuniziert, dann hat das Handy die Telefonie mobil gemacht. Briefe wurden in vielen Bereichen durch E-Mails ersetzt, die Smartphones gaben uns dann die Möglichkeit, E-Mails auch von unterwegs zu lesen und zu schreiben. Das Internet erweiterte die Möglichkeit, sich Informationen zu beschaffen, online einzukaufen, oder zu chatten. Zunächst auf dem PC, dann auch auf mobilen Geräten wie Smartphones oder Tablet Computer.

Unser Kommunikationsverhalten wurde durch diese schnelle Entwicklung nachhaltig verändert, sie wurde immer schneller und mobiler: Wir telefonieren beim Autofahren, schicken E-Mails vom Strand, senden zwischendurch mal eine SMS von der Couch aus. Aber eigentlich ist das auch schon nicht mehr schnell genug, soziale Netzwerke wie Facebook oder Twitter bieten uns permanent die Möglichkeit, mit unseren Kontakten in Verbindung zu bleiben. Und wenn es etwas persönlicher sein soll, dann können wir auch per Skype Videotelefonate über das Internet führen.

Viele Geräte tauschen dabei Informationen miteinander aus. So verbindet sich das Smartphone mit dem Auto nicht nur, um dessen Freisprecheinrichtung zu nutzen. Auch wird das Adressbuch zur Verfügung gestellt, welches die Telefonnummern für Telefonate und die Adressen für die Verwendung mit dem eingebauten Navigationsgerät enthält. Andere Verbindungen erfolgen indirekt, wie beispielsweise das Speichern der Fotos aus dem Smartphone

in der Cloud, sodass z.B. der heimische PC ebenfalls darauf zugreifen und sie anzeigen kann.

Damit die Technik mit ihren vielen verschiedenen Möglichkeiten beherrschbar bleibt, muss sie einfach einzusetzen sein. Viele Geräte kommen uns daher entgegen, indem sie „mitdenken“ und sinnvolle Vorschläge oder Konfigurationen anbieten. Dieses „Mitdenken“ beruht natürlich auf vorab gesammelten Daten, die als Erfahrungswerte verwandt werden, wie beispielsweise die angezeigten Suchvorschläge beim Eingeben eines Suchbegriffes bei Suchmaschinen wie etwa Google.

Überall fallen dabei Benutzer-abhängige Daten an - Kommunikationsdaten, Dokumente, Bilder, Daten aus dem Surf-Verhalten, Geo-Daten usw. Die Speicherorte dieser Daten werden dabei immer vielfältiger. Auf der einen Seite sammeln viele an der Benutzung der Daten beteiligte Stellen (z.B. Suchmaschinen, Shopping-Seiten) Daten, die allerdings häufig nicht im direkten Zugriff der Benutzer liegen. Auf der anderen Seite verteilen sich aber auch die Daten im persönlichen Zugriff über immer mehr Speicherorte: Computer (in Form von Desktop-PCs, Laptops oder auch Servern), Mobilgeräte (Handys, Smartphones, Tablets, Navigationsgeräte), mobile Speichermedien (wie USB-Sticks, Speicherkarten, externe Festplatten), Cloud-Speicher (z.B. Dropbox, Microsoft OneDrive, Google Drive, Apple iCloud) und so weiter.

Die Datenmengen werden dabei immer umfangreicher. Das lässt sich gut beobachten, wenn man die aktuell üblichen Speichergrößen betrachtet: Auch in Privat-PCs sind Festplatten in Terabyte-Größen mittlerweile üblich, selbst der Speicherplatz von Mobilgeräten wie Smartphones umfasst viele Gigabyte.

Untersuchungen haben ergeben, dass sich das Datenvolumen weltweit alle zwei Jahre verdoppelt und bis zum Jahr 2020 auf 40 Zettabyte erhöhen wird. Die Einheit Zettabyte ist folgendermaßen definiert:

Als Versuch der Einordnung: Bei 40 Zettabyte entspricht die Anzahl der Bytes geschätzt 57?mal der Anzahl sämtlicher Sandkörner aller Strände der Erde – damit ist die Anzahl zugegebenermaßen immer noch unvorstellbar groß...

Wir erzeugen also eine immer größer werdende Flut von Daten, die auf diverse Speicherorte verteilt werden, welche sich technisch zunehmend unterscheiden – das ist als Ausgangsbasis für Ermittlungen natürlich eine Herausforderung.

2 DIE GRUNDSÄTZLICHE HERAUSFORDERUNG BEI DER DATENANALYSE

Die Akquise der auszuwertenden Daten ist stets die Grundlage für deren Analyse. Da sich die Daten in den verschiedensten Speicherorten befinden, die teilweise technisch sehr unterschiedlich sind, ist ein umfangreiches technisches Knowhow und das entsprechende Equipment erforderlich.

Die eigentliche Herausforderung beginnt jedoch bei der Datenauswertung. Meist hat man es bei den hierfür zu verarbeitenden Datenmengen zwar aktuell noch nicht mit der Größenordnung Zettabyte zu tun, jedoch ist ein Datenvolumen im Bereich von Terabyte heutzutage durchaus üblich. Speziell bei größeren Ermittlungsverfahren steht man sehr schnell vor der Herausforderung einer erheblichen Datenmenge, die es zu sichten gilt: Es ist schlichtweg nicht möglich, in einer vertretbaren Zeit alle vorliegenden Daten durchzuschauen. Gesucht ist also eine effiziente Methode, um die relevanten Daten ausfindig zu machen.

3 DER HERKÖMMLICHE LINEARE ANSATZ

Die klassische lineare Vorgehensweise illustriert folgendes Beispiel: Bei einem Ermittlungsverfahren im Bereich Wirtschaftskriminalität ging es um die Zahlung von illegalen Provisionen zwischen mehreren Personen. In diesem Verfahren wurden 375 Asservate gesichert, die insgesamt 150 Millionen Dateien enthielten.
Zunächst wurden die gesicherten Daten aller Asservate zur Sichtung aufbereitet, dies nahm aufgrund der Datenmenge schon erhebliche Zeit in Anspruch.

Danach wurde schnell klar, dass bei einem solchen Datenvolumen die Sichtung aller Dateien kein gangbarer Weg ist – das Datenvolumen musste eingegrenzt werden. Eine bewährte Methode hierfür ist der automatische Aufbau eines Suchwortindexes unter Verwendung einer forensischen Auswertungssoftware, mit dem sich dann eine

Schlagwortsuche nach geeigneten Suchbegriffen durchführen lässt. Dazu wird eine Suchwortliste erstellt und diese auf die aufbereiteten Daten angewandt, und ggf. das Ergebnis extrahiert, um das Datenvolumen zu reduzieren. Auch dies nahm im vorliegenden Fall aufgrund der Datenmenge viel Zeit in Anspruch. Durch diese Vorgehensweise konnte die zu sichtende Datenmenge auf 4 Millionen Dateien reduziert werden.

Die Hochrechnung der Sichtungszeit ist bei dieser Größenordnung allerdings erschreckend: Selbst wenn man nur 5 Sekunden Betrachtungszeit pro Datei kalkuliert, ergibt sich ein zeitlicher Aufwand von ca. 3,5 Jahren, der allein schon für die Sichtung anfällt. Eine weitere Auswertung oder Verarbeitung der Daten ist darin noch nicht enthalten…

Bei der Sichtung der Daten entsteht außerdem das Phänomen, dass man gleichzeitig zu viele und zu wenige Daten zu sichten hat: Auf der einen Seite wird man bei der Festlegung der Suchwortliste zu Anfang möglichst viele Begriffe aufnehmen, um nicht zu früh schon möglicherweise relevante Daten bei der Sichtung auszuschließen. Das resultiert in einer großen Menge von „falschen“ Treffern, die für die Ermittlungen nicht zielführend sind.

Auf der anderen Seite wird man erst im Laufe der Sichtung sowie der weiteren Ermittlung neue Erkenntnisse gewinnen und damit weitere Suchbegriffe erkennen, die jedoch in der anfänglichen Suchwortliste noch nicht enthalten sind. Dadurch müssen bereits gesichtete Daten aufgrund neuer Erkenntnisse eventuell noch einmal gesichtet werden.

Um die Sichtungszeit auf ein vertretbares Maß zu reduzieren ist es daher sehr wichtig, schnellstmöglich die wichtigsten Daten zu sichten und neue Erkenntnisse zu gewinnen, um die Suche möglichst effizient durchführen zu können.

Aber was ist dazu die geeignetste Vorgehensweise?

4 DER ITERATIVE PROZESS

Zur Sichtung von großen Datenmengen bietet sich ein iteratives Vorgehen an, in dem zunächst die vielversprechendsten Daten mit Hilfe der für sie am sinnvollsten Untersuchungsmethoden gesichtet werden. Unter Berücksichtigung der dabei gewonnenen Erkenntnisse werden dann die nächsten Daten untersucht.

Um sich mit den zu analysierenden EDV-Systemen vertraut zu machen, ist es sinnvoll, den Prozess mit einer explorativen Datenanalyse zu starten, bei der die gesicherten Daten grob gesichtet, eingeschätzt und nach Relevanz bewertet werden. Dazu ist z.B. eine Sichtung der Verzeichnisstruktur hilfreich, da jeder Benutzer seine Dateien auf eine eigene Art sortiert und ablegt, wobei Dateien mit ähnlichen Inhalten aber häufig zusammen in einem Verzeichnis gespeichert werden. Weiterhin besteht beispielsweise die Möglichkeit, das System als virtuelle Maschine aufzusetzen, um die Umgebung nachzustellen, in der sich der Systembenutzer bewegt hat. Dies kann vor allem dann interessant sein, wenn auf dem System Software installiert ist, in deren Datenbasis relevante Informationen vermutet werden (z.B. Abrechnungsdaten in Warenwirtschaftssystemen oder Zahlungsströme in Buchhaltungssystemen).

Beim iterativen Prozess gliedert sich jede Iteration in mehrere Phasen, die vom Iterationsmanagement koordiniert werden:

1) Iterationsbeginn

Um den Einstieg in die Iteration zu ermöglichen, werden Anfangsparameter festgelegt. Hierbei handelt es sich zum Beispiel um die Festlegung der zu durchsuchenden EDV-Systeme und die Definition einer Liste von Schlagworten, nach denen diese Systeme durchsucht werden sollen. Diese Suchwortliste kann Namen, Email-Adressen oder andere Worte beinhalten und leitet sich direkt aus der aktuellen Fragestellung ab. Die initiale Fragestellung kann sich aus dem Anfangsverdacht oder Erkenntnissen einer vorangegangenen Durchsuchungsmaßnahme ergeben.

2) Datenverarbeitung

Sind die Arbeitsschritte festgelegt, so werden diese entsprechend der Definition durchgeführt. Ziel ist es, Daten zur Beantwortung einer definierten Fragestellung zu finden. Hierzu werden die Datenquellen automatisiert nach verschiedenen Gesichtspunkten durchsucht. Die gefundenen Daten werden auf mobilen Datenträgern oder forensischen Arbeitsstationen für die Analyse zur Verfügung gestellt. In diesem Schritt werden die Daten, sofern sie nicht direkt interpretiert werden können (z.B. Datenexporte aus Programmen, Datenbankinhalte etc.), so aufbereitet, dass der Ermittler sie verwenden und in seine Analysen einbeziehen kann.

3) Datenanalyse

Sind die Daten durch die IT-Forensiker bereitgestellt worden, kann eine Sichtung der Ergebnisse durch Sachverständige, Polizei oder andere durch den Kunden (z.B. Ermittlungsbehörde) benannte Personen erfolgen, die die Ergebnisse der Datenverarbeitung bewerten. Hierzu stehen dem jeweiligen Benutzer Werkzeuge zur Verfügung, die es ermöglichen, die Datenmenge interaktiv zu bearbeiten, zu durchsuchen und zu sichten. Hierbei gewonnene Erkenntnisse werden festgehalten, relevante Dateien werden markiert und extrahiert.

4) Konsolidierung

Die Ergebnisse der vorangegangenen Begutachtung werden nun unter Berücksichtigung der Fragestellung zusammengetragen, dokumentiert und diskutiert. Nach einer Konsolidierung der Erkenntnisse von möglicherweise mehreren Bearbeitungs-Teams folgt die Dokumentation der gewonnenen Erkenntnisse.

Am Ende einer Iteration muss eindeutig geklärt werden, ob die zugrundeliegende Fragestellung hinreichend beantwortet wurde, oder ob eine weitere Iteration durchlaufen werden muss, bei der die Ergebnisse der bisher erfolgten Analysen mit einbezogen werden.

5) Definition der Iteration

Ist bereits mindestens eine Iteration erfolgt, so werden stets die Ergebnisse und Erkenntnisse in die Definition der folgenden Iterationen berücksichtigt.

Die einzelnen Iterationen (Durchläufe durch den Iterationszyklus) können einerseits sehr stark voneinander abweichen und sich mit unterschiedlichen Fragestellungen beschäftigen, die sich im Analyseschritt vorangehender Iterationen ergeben. Andererseits können sie sich auch mit der gleichen Fragestellung, aber unterschiedlichen Analysetiefen oder in verschiedenen Analyse-Bereichen befassen.

Beispiele hierzu sind:

- Suche nach definierten Schlagworten
  - im Dateisystem
  - in verschlüsselten Containern
  - in gelöschten Dateien
  - in auf dem Datenträger vorhandenen Dateifragmenten (File Carving)

- Datei-Interaktionsanalyse
  - Suche nach Hinweisen auf Erstellung, Zugriff oder Kopie von Dateien
  - Suche nach verwendeten Dateien auf angeschlossenen USB-Geräten

- Surfprofil-Analyse
  - Suche nach betrachteten Webseiten
  - Suche nach gespeicherten Seiteninhalten
  - Analyse der Verbindungsdaten und –wege (z.B. Nutzung von Proxy, VPN)

- Geo-Lokalisation
  - Suche nach Geo-Informationen in Fotos
  - Suche nach GPS-Daten in Mobilgeräten

Wie zuvor beschrieben, können bei komplexen Problemstellungen auch thematisch gruppierte Arbeitspakete gebildet werden, so dass voneinander getrennt eine zielgerichtete Bearbeitung möglich ist.

Sind diese Bedingungen bekannt, kann durch das Iterationsmanagement die Einplanung von Personal und Ressourcen und deren Zuweisung zu den anstehenden Aufgaben erfolgen.

6) Iterationsende

Wenn alle Asservate durchsucht sind bzw. alle erforderlichen Untersuchungen darauf angewandt wurden und keine weiteren Erkenntnisse zu erwarten sind, dann endet der Prozess an dieser Stelle.

Übergreifend: Iterationsmanagement

Wie es in einem Projekt einen Projektmanager gibt, der das Projekt und die Beteiligten steuert, Aufgaben zuweist, deren Erfüllung kontrolliert und Planungsaufgaben übernimmt, so gibt es einen solchen Manager auch hier. Der Iterationsmanager behält den Blick für das „große Ganze“ und stellt sicher, dass das Team in sich funktioniert, während die Team-Mitglieder sich auf ihre Aufgaben und deren Umsetzung konzentrieren und diese effektiv und effizient umsetzen können. Arbeiten verschiedene Teams gleichzeitig an der Bearbeitung von unterschiedlichen Iterationen, sorgt der Iterationsmanager für einen reibungslosen Ablauf, teilt die Arbeiten sinnvoll auf und sichert einen konsistenten Informationsfluss zwischen allen Beteiligten. Der Iterationsmanager sorgt ebenso dafür, dass die bereits bekannten Ergebnisse in nachfolgende Iterationen einfließen und bereits erledigte Suchen und Verarbeitungsschritte nicht mehrfach durchgeführt werden.

Durch die Verwendung dieses iterativen Vorgehensmodells ist es möglich, flexibel auf neue Anforderungen und Gegebenheiten zu reagieren. Die Verarbeitung von Erkenntnissen, die sich während der oben beschrieben Vorgehensweise ergeben, schafft ein wendiges und effektives System für sehr komplexe Herausforderungen. Der Einsatz mehrerer parallel arbeitender Teams ermöglicht zudem die gleichzeitige Durchführung verschiedener Iterationszyklen und damit eine hohe Zeitersparnis.

5 FAZIT

Die Datenmengen bei den Ermittlungen werden immer größer und komplexer. Daher wird es immer schwieriger, die zur Verfügung stehenden Daten in akzeptablem Zeitaufwand auszuwerten.

Hier hilft die vorgestellte iterative Vorgehensweise: Das Datenvolumen wird möglichst frühzeitig der jeweiligen Fragestellung entsprechend sinnvoll reduziert und logisch zugeordnet. Durch ein intelligentes Iterations-Management werden die bereits erlangten Ergebnisse in den nächsten Zyklus einbezogen. Andere Ermittlungsergebnisse sowie nachgelieferte Daten können jederzeit flexibel berücksichtigt werden. Somit kann die Ermittlungsarbeit erheblich erleichtert und enorm beschleunigt werden.

Über den Autor:

Wulf Kollmann, Diplom-Informatiker, studierte Informatik mit Nebenfach BWL an der TH Darmstadt und kombiniert seit 1989 beide Fachrichtungen bei Projekten im In- und Ausland. Er entwickelte seit 1994 maßgeschneiderte kaufmännische Datenbank-Anwendungen für Kunden aus verschiedensten Branchen. Seit 2013 durchforstet er bei der CONTURN Analytical Intelligence Group GmbH in Frankfurt am Main den Datendschungel und leitet dort die Abteilung Forensic Services & Project Management.

Die CONTURN AIG GmbH ist im Auftrag von Justiz, Behörden und Unternehmen bundesweit als Dienstleister für IT-Forensik und Datenanalytik aktiv. Mit digital-forensischen Analysen und Beweissicherungen unterstützt CONTURN bei der Aufklärung von Wirtschaftsdelikten.

E-Mail: infoservice(at)conturn.com

www.conturn.com

Alle Artikel dieser Kategorie