ComplianceBusiness
Ausgabe 2 // Juni 2017

Elasticsearch meets Compliance

Anwendungsmöglichkeiten von Enterprise Search im Complianceumfeld

Von Thomas Langthaler und Olly Salzmann

2013 wurde das „Digital Universe“, also die Gesamtheit der durch die Menschheit verursachten und gesammelten Daten, noch auf 4,4 Zettabyte (oder 1 Milliarde Terabyte) geschätzt. Bis 2020 soll diese Zahl auf das Zehnfache, 44 Zettabyte, ansteigen und bis 2025 gar auf 180 Zettabyte. Jedes Unternehmen trägt seinen Teil zu dieser Entwicklung bei, indem immer mehr Daten gesammelt werden. Um mit diesen wachsenden Datenmengen umgehen zu können und die gewünschten Informationen zu finden, sind mächtige Suchtechnologien nötig. Eine solche Technologie bietet Elasticsearch.

„You know, for Search ...“ lautet der Slogan von Elasticsearch und weist somit zunächst auf eine hochskalierbare und blitzschnelle Enterprise-Suchmaschine hin. Entwickelt wird die Elasticsearch-Technologie von dem Unternehmen Elastic BV. Die Plattform „Information Retrieval Engine“ von Apache Lucene stellt dabei den technischen Kern der Suchfunktionalität zur Verfügung. Unter dieser Apache-Lucene-Lizenz wird auch Elasticsearch veröffentlicht und steht damit auch für die kommerzielle Nutzung kostenfrei zur Verfügung. Zu den Nutzern von Elasticsearch zählen namhafte Organisationen aus verschiedensten Themenbereichen: darunter Facebook, Adobe Systems, The Guardian und Netflix.

Um mit den wachsenden Datenmengen umgehen zu können, sind mächtige Suchtechnologien nötig.

ComplianceBusiness
Ausgabe 2 // Juni 2017

Der nachfolgende Artikel bietet einen Erfahrungsbericht über die Nutzung einer solchen Technologie und deren Chancen und Grenzen auch im Forensic- und Complianceumfeld. Neben den Einsatzmöglichkeiten in diesem Bereich – dem Schwerpunkt der Darstellungen – bietet sich die Nutzung von Elasticsearch auch als Applikation zur Effizienzsteigerung der interdisziplinären Zusammenarbeit der Bereiche Technology, Business und Compliance in Unternehmen an.

Überblick über die Funktionalität

Elasticsearch erlaubt die Volltextsuche in Echtzeit in riesigen Datenmengen. Daten in Elasticsearch sind im Gegensatz zu herkömmlichen SQL-Datenbanken dokumentenorientiert, das bedeutet, dass Daten beinahe beliebiger Art indiziert und anschließend durchsucht werden können. Dadurch können zum einen strukturierte Daten aus eben solchen SQL-Datenbanken verarbeitet werden, aber auch (nach geringem Aufbereitungsaufwand) unstrukturierte Daten, wie beispielsweise PDF- oder Microsoft-Office-Dokumente.

Elasticsearch ist oftmals schneller als andere Technologien. Teile seines Geschwindigkeitsvorteils gegenüber anderen Such- und Datenbanktechnologien gewinnt Elasticsearch aus der Art der Speicherung der Daten. Neben dem Originaltext (beispielsweise einem Dokumenteninhalt) bildet Elasticsearch einen sogenannten Index, vergleichbar mit der Suche nach einem Wort im Index eines Buchs. Suchanfragen können entweder in einer Elasticsearch-eigenen Suchbeschreibungssprache oder über zahlreiche Programmierschnittstellen in unterschiedlichen Programmiersprachen (Ruby, PHP und Python, um nur einige zu nennen) abgesetzt werden. Darüber hinaus entstehen zunehmend Konnektoren, um Daten aus Elasticsearch auch in etablierten Big-Data-Tools verarbeiten zu können, beispielsweise Tableau oder Hadoop.

Elasticsearch selbst wird in Java entwickelt und fügt sich daher gut in bestehende Enterprise-Systemlandschaften und Administrationen ein. Der vergleichsweise einfache Datenzugriff im Gegensatz zu anderen Suchlösungen ermöglicht es kleinen Entwicklerteams, in kurzer Zeit beeindruckende Ergebnisse zu produzieren.

Bei Suchen in Elasticsearch sind sogenannte „Fuzzy-Searches“ möglich, bei denen mit einer gewissen Fehlertoleranz gesucht werden kann, um Rechtschreib- oder Tippfehler in Suchbegriffen oder indizierten Daten Rechnung zu tragen. Neben der herkömmlichen Suche bietet Elasticsearch ebenfalls Features zur Autovervollständigung sowie die Erzeugung von Suchvorschlägen an.

Um Geschwindigkeit und Ausfallssicherheit zu erreichen, ist Elasticsearch ein verteiltes System, das auf mehreren Servern läuft. Wird mehr Speicherplatz benötigt, um zusätzliche Daten vorzuhalten, können dem Netzwerk kurzerhand zusätzliche sogenannte Knoten zu anderen Servern hinzugefügt werden.

Um Elasticsearch herum wurden weitere Tools entwickelt, die zusätzliche Auswertungen, Analysen und die Visualisierung von Ergebnissen möglich machen. Diese Applikationen ermöglichen es, nicht nur Dokumente, sondern auch zum Beispiel Zeitreihendaten aus unterschiedlichsten Quellen zu indizieren und außerhalb der erzeugenden Systeme revisionssicher zu speichern. Dies ermöglicht auch eine Auswertung von Log-Informationen aus Betriebssystemen oder Netzwerkkomponenten oder von Transaktionsdaten aus Rechnungslegungssystemen.

Use-Cases für Elasticsearch im Complianceumfeld

Mit seiner Fähigkeit, beliebige Daten nahezu in Echtzeit zu indizieren, zu durchsuchen und Warnhinweise („Alerts“) auf Veränderungen solcher Daten zu erzeugen, bietet sich Elasticsearch für unterschiedliche Anwendungsfälle im Complianceumfeld an. So können derart archivierte Daten zu Zwecken des Continuous Monitoring oder Continuous Auditing von Transaktionsdaten in Echtzeit ausgewertet werden. Auf diese Art könnten beispielsweise auch Indikatoren für Embargoverstöße in Buchungstexten erkannt werden.

Im Fall von Legal Holds, im Kontext von Businessapplikationen, ist es möglich, die Anwendungsdaten in Elasticsearch auszulagern und so komplexe und teure Maßnahmen zur Verhinderung der Löschung dieser Daten in den Anwendungen selbst einzusparen. Darüber hinaus können die archivierten Daten genutzt werden, um eine schnelle Einschätzung treffen zu können, ob ein System für das Verfahren/die Untersuchung überhaupt relevant ist oder ignoriert werden kann.

ComplianceBusiness
Ausgabe 2 // Juni 2017

Im Zusammenhang mit E-Discovery ist die Nutzung für sogenannte Early-Case-Assessments („ECAs“) möglich, um schnell und einfach Einsicht in die Daten von Anwendungssystemen und unstrukturierten Daten (etwa PDF, Microsoft Office, Chatprotokolle etc.) zu erlangen und eine initiale Einschätzung der Relevanz dieser Daten für einen Sachverhalt treffen zu können. Dabei kann zeitnah und proaktiv eine Risiko-Nutzen-Einschätzung für die weitere Verfolgung eines Sachverhalts getroffen werden. Elasticsearch kann dabei unterstützen, derartige Einschätzungen zu beschleunigen oder gar erst zu ermöglichen.

Bei Daten aus betrieblichen Anwendungssystemen ist es notwendig, diese erst auf eine Art und Weise aufzubereiten, die es ermöglicht, sie in Elasticsearch zu indizieren. Nach dieser Aufbereitung ist es grundsätzlich möglich, Datenexporte aus einer großen Anzahl unterschiedlicher unternehmensbezogener IT-Anwendungen durchzuführen.

Außer für ECAs kann Elasticsearch auch genutzt werden, um Informationen zu sammeln und auszuwerten, die für die Informationssicherheit relevant sind. Häufig werden Protokollinformationen von Serversystemen, Netzwerkkomponenten (beispielsweise Firewalls und Proxies) und Angriffserkennungssystemen (Intrusion-Detection-Systems, IDS) archiviert, um proaktiv Monitoringaktivitäten durchzuführen, sicherheitsrelevante Parameter in Echtzeit zu visualisieren sowie bei Sicherheitsvorfällen Nachforschungen zur Ursache anstellen zu können. Dies unterstützt eine zügige Wiederherstellung von Systemen und die Rückführung betroffener Anwendungen in den Normalbetrieb sowie die Ableitung von weiteren Maßnahmen zur zukünftigen Verhinderung ähnlicher Vorfälle und Angriffe. Auf diese Art können beispielsweise auch Anforderungen aus ISO/IEC 27001:2013 oder dem PCI DSS erfüllt werden.

Ebenso ist es möglich, jegliche compliancerelevanten Dokumente und Informationen in Elasticsearch in einer schnell durchsuchbaren Form vorzuhalten, um so kurzfristig compliancebezogene Abfragen durchzuführen. Goldman Sachs beispielsweise nutzt Elasticsearch als Kerntechnologie für verschiedene suchverwandte Fragestellungen. So wird es eingesetzt, um Verträge zu indizieren und der Rechtsabteilung zu ermöglichen, nach bestimmten Vertragsklauseln oder Formulierungen zu suchen; eine Aufgabe, die ohne Einsatz von Technologien sehr personal- und zeitintensiv sein kann. Zusätzlich wird Elasticsearch genutzt, um Trades zu verfolgen und zu jedem Zeitpunkt deren Status zu tracken und verfügbar zu machen.

Ein Fokusbereich ist auch eine sinnvolle Verknüpfung von transaktionellen Daten mit Kommunikationsdaten (Chats oder E-Mails). Damit könnte unter anderem nun auch die Herleitung zu eventuellen Absprachen bei Preisänderungen, Zahlfrist und Verträgen abgebildet werden.

Thomas Langthaler,

Senior Manager,
Deloitte GmbH,
Düsseldorf

tlangthaler@deloitte.de

Olly Salzmann,

Partner, Deloitte GmbH,
Düsseldorf

osalzmann@deloitte.de
www.deloitte.de