# Cloud Data Leakage Prevention mit Methoden der automatischen Datenklassifizierung

Anna Hamberger
*Fakultät für Informatik*
*Technische Hochschule Rosenheim*
Rosenheim, Germany
anna.hamberger@stud.th-rosenheim.de

***Zusammenfassung*—Abstract**
***Index Terms*—component, formatting, style, styling, insert**

## I. Einführung

Im Jahr 2022 gaben bereits 84% der befragten 552 Unternehmen in Deutschland an, dass sie Cloud-Dienste in ihrem Unternehmen einsetzen [1]. Cloud Computing hat sich in der Zeit der digitalen Transformation zu einem wichtigen Bestandteil der Informationsverarbeitung entwickelt. Die Nutzung von Cloud-Diensten wird immer beliebter, da sie die Möglichkeit zur effizienten Speicherung großer Datenmengen, schnellen Zugang zu Ressourcen und nahtlosen Datenaustausch bietet. Durch den Verbreitung von digitaler Technologie in der Gesellschaft und in Unternehmen werden immer mehr Daten geteilt und gesammelt. Um diese großen Datenmengen sammeln und verarbeiten zu können, nutzen Unternehmen die Vorteile von Cloud-Diensten. Die Möglichkeit, Daten in Echtzeit zu teilen, verbessert Geschäftsprozesse und erleichtert die Zusammenarbeit im Unternehmen [2].

Da Informationen das wertvollste Gut eines Unternehmens sind, ist ihr Schutz von größter Bedeutung. Beim Sammeln von Daten ist ein Unternehmen zudem verpflichtet, sie vor Diebstahl, Verlust und Missbrauch zu schützen. Es gibt zahlreiche Datenschutzgesetze und -vorschriften, wie die EU-Datenschutz-Grundverordnung (DSGVO), um sensible Daten wie personenbezogene Daten zu schützen. Ziel dieser Vorschriften ist es, strengen Regeln für das Sammeln von Daten vorzugeben und der Einzelperson eine vergleichsweise hohe Kontrolle über ihre personenbezogenen Daten zu geben [3]. Unabhängig des Speicherorts besteht also das Risiko, dass die Datensicherheit verletzt wird.

Eines der Hauptziele der Informationssicherheit ist die Verhinderung der Offenlegung von Daten gegenüber Unbefugten. Datenlecks können jedoch aufgrund der Notwendigkeit, auf Informationen zuzugreifen, diese zu teilen und zu nutzen, nicht immer verhindert werden. Diese Bedrohung kann von böswilligen Außenstehenden ausgehen, die versuchen, sensible Daten zu erhalten. Umgekehrt können auch interne Mitarbeiter eine Gefahr darstellen, wenn sie beabsichtigt oder unbeabsichtigt Informationen preisgeben [4]. Bereits im Jahr 2018 haben Studien gezeigt, dass 53% der befragten Unternehmen Insider-Angriffe in den letzten 12 Monaten bestätigten. Dabei sind Bedrohungen von innen häufig schwerwiegender als von außen, da sie meist schwieriger zu erkennen sind [5]. Die Offenlegung von sensiblen Daten kann erheblichen Schaden verursachen. Unternehmen können ihren Wettbewerbsvorteil verlieren, ihr Image beeinträchtigen, Umsatzeinbußen erleiden oder sogar Geldstrafen und Sanktionen erhalten.

Um das Risiko von Datenschutzverletzungen zu minimieren, werden immer häufiger Data-Leakage-Prevention (DLP) Lösungen eingesetzt. Gartner prognostiziert, dass bis 2027 etwa 70% der größeren Unternehmen eine DLP-Lösung einsetzen werden, um die Datensicherheit vor Insider-Risiken und externen Angreifern zu schützen [6]. DLP-Systeme überwachen den Zugriff und Austausch vertraulicher Daten, um unbefugte Offenlegung oder missbräuchliche Nutzung zu erkennen.

Unternehmen sammeln häufig große Datenmengen, ohne zu wissen, was erfasst wird oder wie sie nach personenbezogenen Daten suchen oder diese abrufen können. Das erschwert den Schutz der Privatsphäre. DLP-Systeme benötigen die Information, ob bestimmte Daten besonders schützenswert sind oder nicht. Im Zeitalter von Big Data ist es jedoch kaum noch möglich, die enormen Datenmengen manuell zu analysieren. Der Fortschritt im Bereich künstliche Intelligenz (KI) bietet hierbei einen vielversprechenden Ansatz. KI-basierte Methoden zur automatischen Datenklassifizierung können in DLP-Systemen eingesetzt werden, um sensible Informationen zu erkennen.

Aufgrund der neuen Möglichkeiten mit dem Einsatz von KI im Bereich Datenschutz liegt der Fokus in dieser Arbeit auf der Anwendung von Methoden der automatischen Datenklassifizierung zur Erkennung sensibler Informationen, um den Schutz sensibler Daten zu gewährleisten. Diese Arbeit beschäftigt sich mit der Frage, wie sensible Daten in große Datenmengen am besten erkannt werden können. Dabei wird zunächst die Bedrohung durch versehentliche Offenlegung von Daten beschrieben und anschließend die Abwehrmaßnahme 'Data Leakage Prevention' vorgestellt. Dabei liegt der Fokus auf der Erkennung von sensiblen Informationen. Es werden verschiedene KI-basierte Methoden und ihre Funktionsweise im Bezug auf Datenklassifizierung vorgestellt. Anschließend wird deren Einsatz in der Cloud Sicherheit diskutiert.

## II. Accidental Cloud Data Disclosure

Die Cloud Security Alliance (CSA) veröffentlicht jährlich einen Bericht über die größten Bedrohungen der Cloud Security. Dabei werden über 700 Experten zu verschiedenen

Themen in der Cloud Security befragt. Ziel dieses Berichts ist es, auf Bedrohungen, Risiken und Schwachstellen in der Cloud aufmerksam zu machen. Der aktuellste Bericht von 2022 zeigt, dass sich die Verantwortung bezüglich der Sicherheit in der Cloud weg vom Cloud Service Provider und hin zum Cloud-Kunden bewegt [7]. Durch die Verlagerung der Verantwortung und die Komplexität der Cloud steigt das Risiko von Fehlern durch Unwissen. Deshalb ist das achte Sicherheitsproblem des Berichts 'Accidental Cloud Data Disclosure'.

Die versehentliche Offenlegung von Cloud Daten ist eine Sicherheitsbedrohung, die auftritt, wenn sensible Informationen unbeabsichtigt öffentlich zugänglich gemacht werden. Dies kann durch menschliches Versagen, Konfigurationsfehler oder unzureichende Sicherheitsmaßnahmen verursacht werden [7]. So wurde beispielsweise 2023 bekannt, dass bei Toyota Motor die persönlichen Daten von Kunden über mehrere Jahre offengelegt wurden. Der Grund war eine Fehlkonfiguration, wodurch die Datenbank in der Cloud öffentlich zugänglich war [8]. Im August 2023 wurden personenbezogene Daten der derzeitigen Beamten des nordirischen Polizeidienstes versehentlich von einem internen Mitarbeiter auf einer Online-Plattform veröffentlicht, der die Datei verwechselt hatte [9].

Allein die beiden Beispiele zeigen, wie schnell Fehler passieren und so großer Schaden angerichtet werden kann. Durch die Verlagerung der Sicherheits-Verantwortung auf die Cloud-Kunden steigt das Risiko von menschlichem Versagen. Durch Social Engineering und Phishing-Attacken können Mitarbeiter eines Unternehmens unbeabsichtigt sensible Daten wie Zugangsdaten offenlegen. Wie im Beispiel der Polizei in Nordirland kann es Mitarbeitern auch passieren, Daten unwissentlich zu veröffentlichen. Auch der Verlust von unzureichend geschützten Geräten wie Laptop oder Smartphone kann zu einer Daten Offenlegung führen. Der einfache Zugang zu Cloud-Ressourcen kann außerdem dazu verleiten, neue Ressourcen anzulegen oder Services zu nutzen, ohne sich über die nötigen Möglichkeiten der Absicherung zu informieren, wodurch Daten durch Fehlkonfigurationen offengelegt werden können. Doch nicht nur der Mensch ist ein Risiko, sondern auch das Zielsystem. Durch schwache Passwörter, fehlende Authentifizierung bei sicherheitsrelevanten Systemen und weitere Fehlkonfigurationen können Daten in der Cloud unwissentlich offengelegt werden. Aber auch ungeschlossene Sicherheitslücken in verwendeten Cloud-Services sind ein Sicherheitsrisiko [10] [11].

Um die Risiken für eine versehentliche Offenlegung von Daten zu minimieren, gibt es verschiedene Schutzmaßnahmen. Mit einem kontrollierten Identity Access Management (IAM) kann der interne und externe Zugriff auf die Daten geregelt und kontrolliert werden. Mit einer genauen Kontrolle der Zugriffsrechte auf die Cloud-Ressourcen können unbefugte Benutzer daran gehindert werden, auf sensible Daten zuzugreifen. Durch die Einführung von strengen Passwortrichtlinien und dem Einsatz von Passwort-Manager-Software kann das Risiko des unbefugten Zugriffs auf Geräte, Benutzerkonten oder Cloud-Ressourcen minimiert werden. Durch den Einsatz des Prinzips des geringsten Privilegs erhalten Benutzer zudem nur die Berechtigungen, die sie unmittelbar für ihre Aufgaben benötigen. Das minimiert das Risiko von Fehlkonfigurationen oder missbräuchlichem Zugriff. Neben der Kontrolle der Zugriffe sollten auch die möglichen Schwachstellen überwacht werden. Regelmäßige Schwachstellen-Scans helfen dabei, Sicherheitslücken in der Cloud-Infrastruktur zu identifizieren und zu beheben, bevor diese ausgenutzt werden können. Die Überprüfung und Optimierung von Cloud-Konfigurationen gewährleistet, dass Sicherheitseinstellungen korrekt konfiguriert sind. Zudem ermöglicht eine zentrale Aufstellung und Management aller in der Cloud vorhandenen Assets eine bessere Kontrolle und Überwachung der Daten, Dienste und Einstellungen. Um bekannte Sicherheitslücken zu schließen, sollte eingesetzte Software regelmäßig aktualisiert werden. Um menschliche Fehler zu minimieren, sollten außerdem die Mitarbeiter mit Schulungen für sicherheitsrelevante Themen sensibilisiert werden [11].

Im Bezug auf diese Bedrohung werden die Begriffe Data Loss (Datenverlust) und Data Leakage (Datenleck) häufig als Synonym verwendet, aber sie haben einige Unterschiede. Datenverlust ist der Verlust von Daten, der nicht wiederherstellbar ist, wie z.B. durch Schäden an Speichermedien, unbeabsichtigtes Löschen oder Hardwarefehler. Datenlecks hingegen beziehen sich auf die unbeabsichtigte oder absichtliche Übertragung von Daten aus einem gesicherten Bereich. Daher können Datenlecks auftreten, wenn unbefugte Personen sensible oder vertrauliche Informationen erhalten [12]. Aus diesem Grund wird in dieser Arbeit der Ausdruck Datenleck oder Data Leakage verwendet, um die unbeabsichtigte Offenlegung von Daten zu beschreiben.

## III. Cloud Data Leakage Prevention System

Im vorherigen Kapitel II über die Bedrohung durch versehentliche Datenoffenlegung wurden deutlich, wie schnell ein Datenleck in Unternehmen auftreten kann. Die Bedeutung effektiver Maßnahmen zur Vermeidung von Datenlecks hat sich aufgrund der wachsenden Datenmengen und des damit verbundenen Risikos einer Datenschutzverletzung erhöht. Dieser Bedarf wurde 2022 erkannt, als in der neuesten Version der Norm ISO 27001:2022 die Data Leakage Prevention eingeführt wurde. Die internationale Norm ISO 27001 definiert die Bedingungen für die Einrichtung, Umsetzung und kontinuierliche Verbesserung eines dokumentierten Informationssicherheits-Managementsystems. Die Norm gibt außerdem Vorschriften für die Beurteilung und Behandlung von Informationssicherheitsrisiken, die an die spezifischen Bedürfnisse jedes Unternehmens angepasst werden müssen [13]. Ein Datenleck kann auf verschiedene Weise auftreten. Trotz der Tatsache, dass es nicht immer möglich ist, das Auftreten vollständig zu verhindern, können Maßnahmen ergriffen werden, um die Wahrscheinlichkeit eines Auftretens zu verringern. Diese Maßnahmen werden als Data Leakage Prevention (DLP) bezeichnet [13]. Dabei handelt es sich um eine Reihe von Technologien, Produkten und Methoden, die dazu dienen, zu verhindern, dass vertrauliche Informationen ein Unternehmen verlassen. In den letzten Jahrzehnten wur-

den verschiedene Sicherheitssysteme wie Firewalls, Intrusion-Detection-Systeme (Einbrucherkennung) und virtuelle private Netzwerke (VPN) eingeführt, um das Risiko von Datenlecks zu reduzieren. Wenn die zu schützenden Daten klar definiert, strukturiert und konstant sind, erfüllen diese Systeme ihren Zweck. Jedoch sind sie unzuverlässig für Daten, die sich ändern oder unstrukturiert sind. Durch einfache Regeln kann beispielsweise eine Firewall den Zugriff auf ein sensibles Datenobjekt verhindern. Die Firewall erkennt jedoch nicht, wenn das Datenobjekt über einen E-Mail-Anhang gesendet wird. DLP-Systeme hingegen sind darauf spezialisiert, vertrauliche Daten zu identifizieren, zu überwachen und zu schützen und unerwünschte Datenbewegungen zu verhindern. [4].

### *A. Cloud Data Leakage Prevention System*

Ein DLP-System umfasst eine Reihe von Regeln und Richtlinien, die Daten nach ihrem Typ klassifizieren, um sicherzustellen, dass sie nicht böswillig oder versehentlich weitergegeben werden. Das System überwacht Endbenutzeraktivitäten, den Datenfluss sowie die über das Netzwerk gesendeten Daten. Wenn verdächtige Aktivitäten erkannt werden, wird eine Systemwarnung ausgelöst. DLP-Lösungen identifizieren sensible Inhalte mithilfe von Datenklassifizierungs-Label, Techniken zur Inspektion von Inhalten und Kontextanalysen. Sie überwachen die Datenaktivität und kontrollieren sie anhand vordefinierter DLP-Richtlinien. Die Richtlinien definieren, ob die Verwendung bestimmter Inhalte oder Daten in bestimmten Situationen erlaubt sind [6].

Gartner klassifiziert DLP-Lösungen in drei Kategorien. Eine Enterprise-DLP-Lösung ist ein zentrales System, das darauf ausgelegt ist, komplexe Anforderungen und Strukturen großer Unternehmen zu bewältigen. Sie verfügt über fortschrittliche Technologien zur Identifikation, Klassifizierung und Markierung sensibler Daten und ist in der Lage, verschiedene Datenquellen zu integrieren. So kann diese Lösung den gesamten Lebenszyklus von Daten in einem Unternehmen abdecken. DLP-Richtlinien werden dabei an zentraler Stelle verwaltet und durchgesetzt. Dagegen werden integrierte DLP-Lösungen direkt in einen Dienst, wie bspw. ein E-Mail-Gateway, integriert und verfügen deshalb nur über begrenzte Richtlinienfunktionen. Das Management von mehreren integrierten DLP-Systemen ist ein manueller Aufwand, jedoch werden diese Systeme im jeweiligen Dienst speziell an die Anforderungen angepasst und können Inhaltsüberprüfungen besser durchführen. Cloud-native DLP-Lösungen sind die dritte Kategorie, zu der sowohl SaaS-Lösungen als auch Cloud-Anbieter mit integrierten DLP-Funktionen gehören. Sie sind speziell für den Einsatz in Cloud-Umgebungen entwickelt und darauf ausgerichtet, sensible Daten in Cloud-Diensten zu schützen. Diese Lösungen verfügen über Mechanismen zur automatischen Erkennung von sensiblen Daten, die in Cloud-Anwendungen und -Speicherplätzen gespeichert sind. Dies umfasst die Identifikation von Daten in Form von Dokumenten, E-Mails, Datenbanken und anderen Formaten [6]. Im weiteren Verlauf der Arbeit wird der Begriff DLP-System für alle drei Kategorien verwendet.

Das Cybersecurity Framework des National Institute of Standards and Technology (NIST CSF) bietet freiwillige Standards und Best Practices, die Unternehmen dabei helfen, Cybersecurity-Risiken zu managen und zu reduzieren. Es gibt Unternehmen eine Struktur, um ihre aktuelle Cybersicherheitssituation zu bewerten, verbesserungsbedürftige Bereiche zu identifizieren, Maßnahmen zu priorisieren, Fortschritte zu bewerten und mit den Stakeholdern zu kommunizieren. Die CSF besteht aus fünf Kernfunktionen: Identifizieren, Schützen, Erkennen, Reagieren und Wiederherstellen. DLP-Systeme konzentrieren sich hauptsächlich auf die Identifizierung, die Erkennung und den Schutz und ergänzen diese Funktionen durch den Bereich der Überwachung. Die spezifischen Funktionen eines DLP-Systems können je nach Hersteller variieren [14].

Die Literatur-Recherche ergab die folgende Auswahl an Best-Practises, die in DLP-Systemen eingesetzt werden sollten. Um sensible Daten schützen zu können, müssen diese zuerst identifiziert werden. Die Aufgabe besteht darin, ein Dateninventar zu erstellen, die Daten nach ihrer Sensibilität zu klassifizieren und sie entsprechend zu kennzeichnen. Zum Schutz der sensiblen Daten sollten Maßnahmen ergriffen werden, die den Zugriff auf die Daten einschränken. Das bedeutet, dass Richtlinien wie minimale Zugriffsrechte, starke Authentifizierungsmethoden und strenge Zugriffskontrolllisten eingeführt werden sollten. Außerdem sollten Daten sowohl im Ruhezustand als auch während der Übertragung verschlüsselt werden. So wird sichergestellt, dass die Daten selbst dann, wenn sie abgefangen werden, für unbefugte Benutzer unlesbar bleiben. Zusätzlich sollte ein DLP-System die Datenströme innerhalb und nach außen überwachen, um potenzielle Datenschutzverletzungen oder Richtlinienverstöße in Echtzeit erkennen zu können. Dies ermöglicht eine schnelle Reaktion auf potenzielle Probleme und begrenzt den daraus resultierenden Schaden [15] [16] [17].

Die Funktionen eines DLP-Systems basieren alle darauf, dass sensible Daten erkannt und in irgendeiner Art markiert sind. Derr erste Schritt bei DLP-Systemen ist daher die Identifizierung sensibler Daten. Es gibt verschiedene Strategien und Methoden zur Klassifizierung dieser Daten, die durch den Einsatz von KI weiter verbessert wurden.

### *B. Erkennung von sensiblen Daten*

Unternehmen setzen immer mehr auf SaaS-Produkte, anstatt sie als Produkt zu kaufen [18]. In ihrem Tagesgeschäft verlassen sich Unternehmen oft auf mehrere Softwareprodukte, um verschiedene Anforderungen zu erfüllen. Das hat zur Folge, dass die Daten des Unternehmens über verschiedene Apps und Cloud-Plattformen verstreut sind. Die Herausforderung besteht darin, den Überblick zu behalten und zu wissen, wo sich die sensiblen Daten befinden. Das Sammeln und Identifizieren von Daten in DLP-Systemen stellt aufgrund von Verschlüsselung, verborgenen Kanälen, nicht unterstützten Datenformaten und großer Mengen an Daten eine große Herausforderung dar [19].

Die Methoden zur Erkennung und Klassifizierung von sensiblen Daten unterscheiden sich je nach Art und Format der

Daten, sowie deren Zustand. Außerdem gibt es die Möglichkeit, Daten manuell oder automatisiert zu klassifizieren.

*1) Daten Art und Format:*

*2) Daten Zustand:*

*3) Manuelle Datenklassifizierung:*

*4) Automatische Datenklassifizierung:*

*a) content-based:*

*b) context-based:*

### *C. Anwendung in der Cloud Security*

*1) Labeling für andere Maßnahmen:*

*2) Anwendungsbeispiele:*

## IV. Ausblick

## Literatur


[1] KPMG, "Nutzung von cloud computing in unternehmen in deutschland in den jahren 2011 bis 2022," 2022. [Online]. Available: https://de.statista.com/statistik/daten/studie/177484/umfrage/einsatz-von-cloud-computing-in-deutschen-unternehmen-2011/

[2] C. Surianarayanan and P. R. Chelliah, "Introduction to cloud computing," in *Essentials of Cloud Computing*, ser. Texts in Computer Science, C. Surianarayanan and P. R. Chelliah, Eds. Cham: Springer International Publishing and Imprint: Springer, 2023, pp. 1–38.

[3] V. Kužina, A.-M. Petric, M. Barišić, and A. Jović, "Cassed: Context-based approach for structured sensitive data detection," *Expert Systems with Applications*, vol. 223, p. 119924, 2023.

[4] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, "A survey on data leakage prevention systems," *Journal of Network and Computer Applications*, vol. 62, pp. 137–152, 2016.

[5] CA Technologies, "2018 insider threat report," 2018. [Online]. Available: https://crowdresearchpartners.com/wp-content/uploads/2017/07/Insider-Threat-Report-2018.pdf

[6] R. Chugh and A. Bales, "Marktführer für data loss prevention," 2023. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-2EYVL2C2&ct=230912&st=sb

[7] Cloud Security Alliance, "Top threats to cloud computing pandemic eleven," 2022. [Online]. Available: https://cloudsecurityalliance.org/artifacts/top-threats-to-cloud-computing-pandemic-eleven

[8] Z. Whittaker, "Toyota confirms another years-long data leak, this time exposing at least 260,000 car owners," *TechCrunch*, 31.05.2023. [Online]. Available: https://techcrunch.com/2023/05/31/toyota-customer-data-leak-years/

[9] PSNI, "Police service of northern ireland statement on data breach | psni," 12.11.2023. [Online]. Available: https://www.psni.police.uk/latest-news/police-service-northern-ireland-statement-data-breach

[10] S. Trabelsi, "Monitoring leaked confidential data," in *2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, IEEE, Ed., 2019, pp. 1–5.

[11] T. Brindha and R. S. Shaji, "An analysis of data leakage and prevention techniques in cloud environment," in *2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, IEEE, Ed. IEEE, 2015, pp. 350–355.

[12] Proofpoint, "Dlp - data loss prevention: Schutz vor datenverlust | proofpoint de," 2021. [Online]. Available: https://www.proofpoint.com/de/threat-reference/dlp

[13] V. Monev, "Data leakage prevention in iso 27001: Compliance and implementation," in *2023 International Conference on Information Technologies (InfoTech)*, IEEE, Ed. IEEE, 2023, pp. 1–5.

[14] NIST, "Framework for improving critical infrastructure cybersecurity," 2014. [Online]. Available: https://www.nist.gov/system/files/documents/cyberframework/cybersecurity-framework-021214.pdf

[15] M. E. Hussain and R. Hussain, "Cloud security as a service using data loss prevention: Challenges and solution," in *Internet of Things and Connected Technologies*, ser. Lecture Notes in Networks and Systems, R. Misra, N. Kesswani, M. Rajarajan, B. Veeravalli, and A. Patel, Eds. Cham: Springer International Publishing and Imprint: Springer, 2022, vol. 340, pp. 98–106.

[16] I. Herrera Montano, J. J. García Aranda, J. Ramos Diaz, S. Molina Cardín, I. de La Torre Díez, and J. J. P. C. Rodrigues, "Survey of techniques on data leakage protection and methods to address the insider threat," *Cluster Computing*, vol. 25, no. 6, pp. 4289–4302, 2022.

[17] B. S. Shishodia and M. J. Nene, "Data leakage prevention system for internal security," in *2022 International Conference on Futuristic Technologies (INCOFT)*. IEEE, 2022, pp. 1–6.

[18] Gartner, "Umsatz mit software-as-a-service (saas) weltweit von 2010 bis 2022 und prognose bis 2024 (in milliarden us-dollar)," 2023. [Online]. Available: https://de.statista.com/statistik/daten/studie/194117/umfrage/umsatz-mit-software-as-a-service-weltweit-seit-2010/

[19] B. Hauer, "Data and information leakage prevention within the scope of information security," *IEEE Access*, vol. 3, pp. 2554–2565, 2015.