Data Lakes und europäischer Datenschutz

Sicher unterwegs im Data Lake

Ein Beitrag von
Emrah Birsin

Sonderdruck aus
BI-SPEKTRUM 1/2018

Mit dem Inkrafttreten der Europäischen Datenschutzgrundverordnung stehen Unternehmen vor enormen Herausforderungen. Der Schutz personenbezogener Daten bildet ein wichtiges Grundrecht eines jeden Bürgers. Dieser Verantwortung gilt es sich nun zu stellen. Wichtig ist, dass die Europäische Datenschutzverordnung (EU-DSGVO) nur Daten natürlicher Personen schützt. Sie gilt demnach für Daten, die sich einer Person, deren Kennnummern oder Online-Kennungen zuweisen lassen. Nicht betroffen sind Daten juristischer Personen oder Daten, die sich keiner Person zuweisen lassen, sprich anonymisiert sind. Personenbezogene Daten dürfen nicht gespeichert oder verarbeitet werden, es sei denn es liegen besondere Umstände vor, in den meisten Fällen eine Einwilligung der Person oder eine rechtliche Verpflichtung. Wenn es keine Möglichkeit gibt, ein Datum einer natürlichen Person zuzuordnen, gelten die Daten als nicht personenbezogen. Dies gilt auch mit Vorbehalt für anonymisierte Daten von Personen.

Werden personenbezogene Daten gespeichert oder verarbeitet, so gelten die folgenden Grundsätze:

- Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz
- Zweckbindung
- Datenminimierung
- Richtigkeit
- Speicherbegrenzung
- Integrität und Vertraulichkeit

Bild: Shutterstock

Wenn ein klassisches Data Warehouse um einen Data Lake erweitert werden soll, um zum Beispiel Vorteile wie kostengünstigen Speicher, erhöhte Flexibilität bei Analysen, Skalierbarkeit, Kostensenkung durch Pay-as-you-use-Rechnungsmodelle, keine Hardwarewartung etc. zu nutzen, stellen einen diese Grundsätze vor besondere Herausforderungen.

Grundlegende Überlegungen zur Einführung eines Data Lake

Gesetzestext vs. Realität: Die DSGVO [Eur16] versucht allgemeine Regeln für die Verarbeitung personenbezogener Daten aufzustellen. Aber in der Realität tauchen immer wieder Situationen auf, in denen man als Entwickler nicht ganz klar sagen kann, wie die DSGVO auszulegen ist. Daher sollte man immer, wenn personenbezogene Daten verarbeitet werden sollen, einen Datenschutzbeauftragten mit zu Rate ziehen, gegebenenfalls auch einen Juristen, der sich mit der aktuellen Rechtsprechung auskennt.

Anonymisierung vs. Pseudonymisierung: Eine grundlegende Überlegung ist die Frage nach der Möglichkeit der Anonymisierung bzw. der Pseudonymisierung. Der Unterschied ist in Abbildung 1 schematisch veranschaulicht. Wenn Daten komplett anonymisiert werden, also keine Identifizierung einer natürlichen Person mehr möglich ist, greift auch die DSGVO nicht mehr.

Allerdings gehen bei der Anonymisierung Informationen verloren, die eventuell zur Erfüllung diverser Aufgaben gebraucht werden. So sollen zum Beispiel Bestellungen bestimmten Personen zugewiesen werden können, um den aktuellen Stand der Bearbeitung überprüfen zu können.

In Fällen, in denen eine Anonymisierung nicht möglich ist, muss stattdessen eine Pseudonymisierung vorgenommen werden, das heißt, Personen werden Pseudonyme zugeordnet. Bei einer Pseudonymisierung gehen keine Informationen verloren, deswegen bleiben die Daten letztlich einer Person zugeordnet und unterliegen damit der DSGVO.

Warum ein Data Lake?

Die Gründe für die Nutzung eines Data Lake sind vielfältig [CNK 15; StM 14; Val 17]. Insbesondere die Lagerung verschiedener Daten in unterschiedlichen Strukturen für zukünftige Auswertungen, die noch erarbeitet werden müssen, ist oft ein Grund für die Verwendung eines Data Lake. Hierbei ist jedoch Vorsicht geboten, da eine zweckfreie Speicherung personenbezogener Daten nicht erlaubt ist.

Ein weiteres Szenario besteht darin, dass ein bereits etabliertes DWH mit Inkrafttreten der DSGVO den regulatorischen Anforderungen nicht mehr genügt. Der komplette Neuaufbau eines verordnungskonformen Systems dauert lange und droht entsprechend kostenintensiv zu werden. Hier kann das Offloading in einen Data Lake, der den entsprechenden Regelungen Rechnung trägt, bei gleichzeitig sehr stringenten Zugriffsbeschränkungen auf das bestehende DWH das Problem abfedern, wodurch man sich ein bisschen Zeit für notwendige Anpassungen des DWH erkaufen kann. Ein Data-Lake-Konstrukt ermöglicht es, die Daten relativ unabhängig von im vorhandenen DWH existierenden Strukturen und Mechanismen zu extrahieren, um sie mit individuellen und flexiblen Programmierungsalgorithmen rechtskonform übertragen zu können. Gleichzeitig schafft man die Möglichkeit, einen weiteren Schritt in Richtung einer Zentralisierung der Unternehmensdaten mit der Integration weiterer Datenquellen wie zum Beispiel Streamingdiensten oder Sensordaten zu unternehmen (Abbildung 2).

Schon während der Übertragung sollten die Daten pseudonymisiert oder anonymisiert sein, das heißt, die Daten sollten vor dem Übertragen pseudonymisiert bzw. anonymisiert werden. Wenn mehrere Daten für die Verarbeitung einer bestimmten Person zugeordnet werden müssen, sollte die Person durch ein Pseudonym (ID) repräsentiert werden. Doch eine Pseudonymisierung allein reicht zum Schutz der Daten nicht aus. Wenn ein Dritter die Daten mit Daten aus anderen Quellen verbindet, könnte es sein, dass Personen identifiziert werden können (siehe [NaS08; Swe02]). Daher müssen personenbezogene Daten verschlüsselt gelagert und transferiert werden, damit selbst im Fall eines Datenlecks keine personenbezogenen Informationen aus den Daten gezogen werden können.

Wenn es möglich ist, sollten Daten, die für verschiedene Analysen genutzt werden und nicht miteinander verknüpft werden müssen, auch mit einer eigenen Pseudonymisierung versehen werden, um die Verknüpfung von Datensätzen durch Dritte zu erschweren. Pseudonymisierungsdaten – also Daten, die Informationen enthalten, wie pseudonymisiert wurde – und nicht anonymisierte Daten sollten nicht im Data Lake gespeichert werden und eine eigene Verschlüsselung erhalten, um die Sicherheit der personenbezogenen Daten weiter zu steigern. Gegebenenfalls sollten diese Daten auch physisch separat gespeichert und treuhänderisch verwaltet werden.

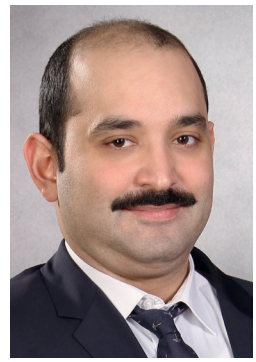
Der Zugriff auf Daten sollte durch ein Rechtevergabesystem auf Grundlage der Verantwortung der einzelnen Mitarbeiter geregelt sein. Dazu kann der Data Lake in verschiedene Bereiche eingeteilt werden, in denen nur bestimmte Daten abgelegt werden (siehe Abbildung 2). Besonders der Zugriff auf Pseudonymisierungsdaten und nicht anonymisierte Daten sollte auf ein Minimum beschränkt sein. Auch hier kann eine treuhänderische Verwaltung in Betracht gezogen werden.

Beim Anlegen der Daten im Data Lake sollte auch immer beachtet werden, dass betroffene Personen ein Widerspruchsrecht und Einschränkungsrecht besitzen, Daten also vielleicht zeitweise eingeschränkt oder gar nicht verarbeitet werden dürfen. Daher sollte die Information, ob bestimmte Daten überhaupt verarbeitet werden dürfen, ebenso hinterlegt sein, am besten auf der Datenebene. Analyse-Tools können dann Daten, die nicht zur Verfügung stehen, automatisch filtern und von der Verarbeitung ausschließen.

Löschung von Daten (Recht auf Vergessenwerden)

Wenn Daten gelöscht werden sollen, sei es, weil der Zweck der Datenerhebung erfüllt wurde oder aufgrund eines Löschungsantrags, muss darauf geachtet werden, dass sämtliche Daten, die mit der

DR. EMRAH BIRSIN ist Business Intelligence & Analytics Developer bei OPITZ CONSULTING Deutschland GmbH. Im Rahmen seiner mehrjährigen Tätigkeit in klinischen Studien erwarb er sich umfassende Erfahrungen im Umgang mit personenbezogenen Daten.
E-Mail:
Emrah.Birsin@opitz-consulting.com



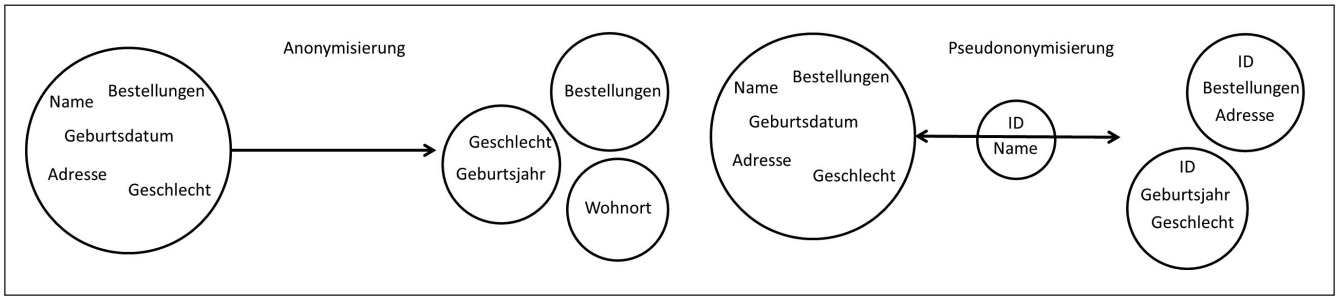


Abb. 1: Vergleich von Anonymisierung (links) und Pseudonymisierung (rechts). Bei der Anonymisierung gehen Informationen verloren. Einzelne Daten lassen sich nicht mehr einer Person zuordnen. Auch verschiedene Datensätze lassen sich nicht mehr verbinden. Bei der Pseudonymisierung lassen sich über zusätzliche Daten die Daten wieder mit den Originaldaten verbinden. Auch reduzierte Daten können wiederhergestellt werden und getrennte Datensätze lassen sich verbinden.

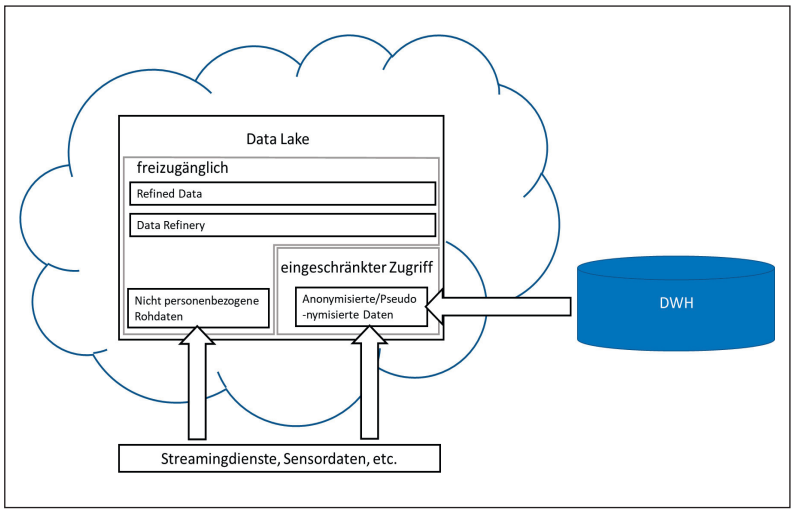


Abb. 2: Exemplarische Struktur eines Data Lake: Personenbezogene Daten werden in einem extra Bereich des Data Lake gespeichert, der besondere Zugriffsrechte verlangt

Person in Bezug stehen, gelöscht werden – wobei gelöscht in den meisten Fällen bedeutet, dass die Daten komplett anonymisiert werden, da ein Interesse an den Daten zwecks historischer Datenanalyse besteht. Daten, die ihren Verwendungszweck erfüllt haben, sollten automatisch einer vollständigen Anonymisierung der Daten unterzogen werden, wobei darauf zu achten ist, dass keine indirekte Identifizierung aufgrund verknüpfter Daten möglich sein darf. Bei einem Löschungsantrag müssen alle Daten der Person identifiziert und entsprechend anonymisiert werden. Pseudonymisierungsinformationen müssen gelöscht werden. Selbst zum Zweck der Dokumentation darf keine Information darüber erhalten bleiben, ob jemand Teil der Datenbank war, da selbst die Information, dass jemand Teil einer Datenbank war, eine personenbezogene Information ist.

Darüber hinaus ist auch darauf zu achten, dass die Löschung in allen Kopien und Replikationen unverzüglich durchzuführen ist, was Backups miteinschließt. Daher sollten nach der Löschung der Daten im Produktivsystem alte Backups aktualisiert werden. Da die Sicherheit der Daten anderer Personen hier eine Rolle spielt, kann das Backup so lange verschoben werden, bis die Stabilität des Systems und die Sicherheit der Daten anderer Personen sichergestellt wurde. Kriterien zur Identifizierung dieses Zeitpunkts sollten in entsprechenden Dokumenten festgelegt sein.

Ein gewisser Konflikt entsteht hier bei der Benachrichtigung von Empfängern der Daten. Bei der Löschung personenbezogener Daten muss allen Empfängern der Daten die Löschung mitgeteilt werden. Das Speichern der Benachrichtigung würde aber bedeuten, dass dokumentiert wurde, dass eine Person Teil eines Datensatzes war. Jede Nachricht, die den Löschungswunsch weitergibt, müsste sofort wieder aus dem System gelöscht werden. Die sofortige Löschung könnte aber zu Problemen führen: Sollte es zum Beispiel zu Fehlern beim Zustellen der Nachricht kommen, könnte man nicht mehr nachvollziehen, welcher Löschungsantrag an den Empfänger gesendet wurden.

Eine mögliche Lösung könnte darin bestehen, beim ersten Informationsaustausch mit dem Empfänger eine ID für jede Person anzulegen, die für die Kommunikation mit dem Empfänger genutzt wird. Diese ID sollte jedoch nicht identisch mit der eigenen internen ID sein. Die ID der betroffenen Person und der Löschungswunsch können übertragen und gespeichert werden, da nur pseudonymisierte Daten in der Benachrichtigung verwendet werden, die, nachdem die Pseudonymisierungsdaten aus dem System gelöscht wurden, keiner Person mehr zugeordnet sind und damit nicht mehr dem Schutz personenbezogener Daten unterliegen.

Daten aus sozialen Netzwerken und anderen öffentlichen Quellen

Ein sehr heikles Thema stellen personenbezogene Daten aus öffentlichen Quellen dar [Brü17]. Zwar beinhalten die Nutzungsbedingungen sozialer Netzwerke entsprechende Passagen, die darauf hinweisen, dass eigene Beiträge und je nach den eigenen Einstellungen auch Nutzerdaten für andere öffentlich zugänglich sind. Somit sind Beiträge und Nutzerdaten, die öffentlich zugänglich sind, von der betroffenen Person offensichtlich öffentlich gemacht worden, was bedeutet, dass persönliche Informationen verarbeitet werden dürfen. Aber die EU-DSGVO entbindet einen in diesen Fällen nicht von der Informationspflicht. Den betroffenen Personen muss weiterhin mitgeteilt werden, dass entsprechende Daten erhoben werden und zu welchem Zweck; weitere Informationen, die es der betroffenen Person erlauben, ihre Rechte in Anspruch zu nehmen, müssen bereitgestellt werden. Nur wenn die Erteilung der Information unmöglich oder der Aufwand unverhältnismäßig hoch ist, kann davon abgesehen werden, die betroffene Person zu benachrichtigen. Aber da keine genauen Richtlinien existieren, wann die Mitteilung der Informationen als unmöglich anzusehen ist oder was ein unverhältnismäßiger Aufwand wäre, ist eher davon abzuraten, personenbezogene Daten aus sozialen Netzwerken zu verarbeiten.

Anders sieht die Sache natürlich aus, wenn ein Unternehmen einen Dienst anbietet, bei dem darauf hingewiesen wird, dass Daten aus sozialen Netzwerken zwecks Personalisierung des Dienstes herangezogen werden, da es sich hier um gezielte Suchen nach bestimmten Personen handelt, die ihr Einverständnis gegeben haben und informiert wurden. Aber auch hier ist darauf zu achten, dass den Personen die Möglichkeit gegeben werden muss, diese zusätzliche Datensammlung zu unterbinden, wenn sie für den Dienst nicht zwingend notwendig ist.

Als Letztes gibt es noch die Möglichkeit, nur nicht personenbezogene Informationen aus öffentlichen Quellen abzufragen und zu verarbeiten, wie die Anzahl von Erwähnungen bestimmter Begriffe oder komplett anonymisierte Daten. Aber auch hier gilt es aufzupassen. Die Daten sollten nicht durch Kombination anderer öffentlich zugänglicher Daten zur Identifizierung einzelner Personen

nutzbar sein. Zum Beispiel wenn die Zeiträume, über die die Daten erfasst und aggregiert werden, zu klein gewählt werden, können diese Zeiträume nur einzelne Beiträge enthalten, die den Suchkriterien genügen, und sind daher leicht einer einzelnen Person zuzuordnen.

Datenreduzierung im Data Lake

Zwecks Einhaltung der Zweckbindung, Datenminimierung (Datensparsamkeit) und Speicherbegrenzung sollten die Daten im Data Lake kontinuierlich darauf überprüft werden, ob sie noch gebraucht werden. Ist schon beim Anlegen der Daten ein Verfallsdatum für den Zweck vorherzusehen, sollten die Daten mit einem Verfallsdatum versehen werden. Wenn möglich ist eine automatische Löschung vorzuziehen, aber auch ein Verfahren, in dem vor der Löschung die Bestätigung einer verantwortlichen Person eingeholt wird, ist möglich. Da je nach erhobenen Daten verschiedene Zeiträume in Frage kommen könnten, sollte das Verfallsdatum auf Datenebene gesetzt werden.

Für Daten, deren Verfallsdatum an Ereignisse gekoppelt ist, bei denen der Eintrittszeitpunkt nicht klar ist, sollten, wenn möglich, Trigger hinterlegt werden, die beim Eintreten des Ereignisses entsprechendes Personal darauf hinweisen und gegebenenfalls ein Verfallsdatum vergeben.

In manchen Fällen mag es sein, dass zwecks statistischer Analysen gewisse Daten über einen Zeitraum gespeichert werden sollen, der weit über den eigentlichen Verwendungszeitraum hinausgeht, zum Beispiel um die Entwicklung der Kundenzahlen über die Jahre hinweg zu analysieren. In dem Fall muss sichergestellt werden, dass die Daten nicht einer Person zugeordnet werden können, also keine Pseudonymisierungsdaten mehr vorliegen und die Daten so weit wie möglich anonymisiert werden. Solche Daten sollten auf ein Minimum reduziert werden, sowohl was die Menge an Daten angeht als auch den Informationsgehalt. Soll zum Beispiel nur die Anzahl von Kunden über die Jahre analysiert werden, so reichen Jahresangaben für den Kundenkontakt und eine interne Kunden-ID, die keine Zuweisung zu einer Person mehr besitzt.

Überwachung der Datensicherheit

Alle Verbindungen zum Data Lake und jede Verarbeitungsanfrage sollten überwacht und gespeichert werden. Zum einem erlauben diese Daten die Identifizierung von Attacken auf den Data Lake sowie die Einleitung von Gegenmaßnahmen, zum anderen hilft die Überwachung dabei, Datenschutzverletzungen frühzeitig aufzudecken, die Auswirkungen einer solchen Verletzung zu minimieren und der Meldepflicht bei den Aufsichtsbehörden nachzukommen.

Die Speicherung und Verarbeitung der dabei anfallenden Daten, die auch personenbezogen sein können, ist durch die rechtliche Verpflichtung zum Datenschutz sowie der Wahrung der Grundrechte und Grundfreiheiten der Personen, deren Daten gespeichert sind, legitimiert. Diese Log-Dateien sollten bezüglich der Sicherung wie personenbezogene Daten behandelt werden.

Fazit

Der Schutz personenbezogener Daten ist ein komplexes Thema. Systeme sollten immer darauf überprüft werden, ob sie auf dem neuesten technischen Stand sind. Prozesse wie Pseudonymisierung und Anonymisierung sowie die Festlegung von Zugriffsrechten und die Verschlüsselung der Daten bilden eine technische Grundlage zum Schutz der Daten. Aber unabdingbar sind auch eine Sensibilisierung der Mitarbeiter und eine genaue Betrachtung der Umstände, unter denen die Daten erhoben, gespeichert und verarbeitet werden. Daher sollten vor dem Aufbau eines Data Lake die Architektur und die einzelnen Lake-Bereiche gemäß dem Vorhaben nach den vorhandenen Datenschutzrichtlinien beleuchtet und entsprechende Vorgehensweisen festgelegt werden.

Literatur

[Brü17] Brünen, B.: Datenschutzgrundverordnung: Endet die Ära der Social Media Plugins? http://www.it-rechtkanzlei.de/social-plugins-datenschutzgrundverordnung-dsgvo.html, abgerufen am 30.1.2018

[CNK15] Chroust, T. / Nemecek, J. / Kürschner, S.: Data Lakes – Möglichkeiten und Herausforderungen für eine effiziente Erkenntnisgewinnung. In: I.VW Management-Information – St. Galler Trendmonitor für Risiko- und Finanzmärkte, 4/2015, S. 21–25

[Eur16] Europäisches Parlament und der Rat: Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Text von Bedeutung für den EWR). 2016

[NaS08] Narayanan, A. / Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: IEEE Computer Society, 2008, S. 111–125

[StM14] Stein, B. / Morrison, A.: The enterprise data lake: Better integration and deeper analytics. Technology Forecast: Rethinking integration, Issue 1, 2014

[Swe02] Sweeney, L.: k-anonymity: a model for protecting privecy. In: International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Oktober 2002, S. 557–570

[Val17] Vallaey, M.: Why do I need a Data Lake. http://info.bigindustries.be/why-do-i-need-a-data-lake, abgerufen am 29.1.2018