Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Personenbezogene Daten im Data Lake

Felix Ebinger

Studiengang:	Informatik
Prüfer/in:	PD Dr. rer. nat. habil. Holger Schwarz
Betreuer/in:	Corinna Giebler, M.Sc.
Beginn am:	14. Juni 2018
Beendet am:	14. Dezember 2018

Kurzfassung

Big-Data-Analysen bieten Wettbewerbsvorteile, ermöglichen Innovationen und können zu einer höheren Qualität von Produkten oder Serviceleistungen beitragen. Insbesondere die Analyse von Kundendaten und des Kundenverhaltens eröffnet vielfältige Möglichkeiten, um dem Kunden auf ihn zugeschnittene Angebote zu unterbreiten und um so zu höheren Umsätzen und zu einer höheren Kundenzufriedenheit beizutragen.

Für die dafür benötigten Daten werden geeignete Speichersysteme benötigt. Ein solches System stellt der Data Lake dar. Neben der gut skalierenden und günstigen Speicherung von Daten ist auch die Auswertung der Daten mittels explorativer Analysen bereits im Design angelegt.

Gleichzeitig steht aber auch der Schutz, genauer der fehlende Schutz der Privatsphäre, des Einzelnen bei Big Data Verarbeitungen im Mittelpunkt der öffentlichen Aufmerksamkeit und Kritik. Insbesondere wird vor dem so entstehenden „gläsernen Menschen“ und den daraus resultierenden gesellschaftlichen Folgen gewarnt.

Die sich daraus ergebenden Fragen, in welchem Umfang und auf welche Art personenbezogene Daten verarbeitet werden dürfen, bedürfen, neben einer ethisch-moralischen, vor allem einer rechtlichen Antwort. Die europäische Datenschutzgrundverordnung stellt hierzu den rechtlichen Rahmen dar, in dem personenbezogene Daten verarbeitet werden dürfen.

In dieser Arbeit werden die gesetzlichen Anforderungen mit dem Konzept des Data Lakes abgeglichen und es wird aufgezeigt, wo Herausforderungen beim Design und bei der Implementierung eines Data Lakes entstehen (z.B. Transparenz, Zweckbindung, Recht auf Löschung). Zudem werden Lösungsansätze für diese Herausforderungen entwickelt und vorgestellt.

Aus den einzelnen Lösungsansätzen werden zwei Lösungskonzepte für einige der identifizierten Herausforderungen entwickelt. Eines der Konzepte, ein Metadaten-Modell, wird dabei prototypisch umgesetzt und anhand von Use Cases beispielhaft getestet.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Listings

1 Einleitung

Big Data nimmt, in all seinen Ausprägungen und Einsatzgebieten, im Alltag der Menschen und damit auch in der öffentlichen Wahrnehmung einen immer größeren Stellenwert ein. Big-Data-Analysen durchdringen das Leben vom Privaten (z.B. Sprachassistenten) über das Berufliche (z.B. Industrie 4.0) bis hin zum Handeln der öffentlichen Hand (z.B. Predictive Policing [Kno18]).

Die riesigen Datenmengen und die ganz unterschiedlich strukturierte Daten, die bei Big Data entstehen, bringen klassische Speichersysteme, wie z.B. relationale Datenbankmanagementsysteme (RDBMS), an ihre Grenzen. Daher wurde das Konzept des *Data Lakes* entwickelt, das gut skaliert und somit eine kostengünstige Speicherung großer Datenmengen ermöglicht. Im Data Lake werden alle Daten zunächst in ihrer Rohform, das heißt unverändert, gespeichert. Dieser Ansatz ermöglicht es zum einen mit den heterogenen Datenstrukturen umzugehen und eröffnet zum anderen für die explorative Analyse ganz neue Möglichkeiten, da ohne eine Aufbereitung der Daten nicht die Gefahr besteht, dass Informationen verloren gehen, die später gebraucht werden.

Die zunehmende Verbreitung von Big-Data-Analysen in den unterschiedlichsten Lebensbereichen birgt große Herausforderungen, insbesondere im Hinblick auf den Datenschutz. Spätestens seit den Snowden-Enthüllungen 2013 wird in der öffentlichen Debatte verstärkt diskutiert, wie viel Organisationen, ob staatliche (z.B. NSA) oder private (z.B. Facebook, Google), über den Einzelnen wissen und wissen dürfen. Verstärkt wurde diese Debatte im Frühjahr 2018 durch die Datenschutzgrundverordnung (DSGVO), die im Mai 2018 anwendbar wurde.

Die Diskussion um die DSGVO wurde in der Öffentlichkeit und in den Medien nicht immer faktisch korrekt geführt. So war beispielsweise in der Berichterstattung fälschlicherweise immer wieder die Rede davon, dass für jede Verarbeitung von Daten über eine Person eine Einwilligung erforderlich sei [ZDF18]. Dennoch birgt die DSGVO als datenschutzrechtliche Regulierung große Herausforderungen im Bereich Big Data.

Exemplarisch dafür seien hier nur vier Grundsätze des Datenschutzrechts angesprochen. Es muss dem Einzelnen verständlich erklärt werden, wie und wofür Daten über ihn verarbeitet werden (*Grundsatz der Transparenz*), und er muss mit der Verwendung der Daten für eine bestimmte Verarbeitung rechnen können (*Grundsatz der Fairness*). Eine nachvollziehbare Erklärung der Verarbeitung ist im Bereich der massenhaften und explorativen Analysen, ganz zu schweigen vom maschinellen Lernen, kaum möglich. Zudem stehen der Idee der massen- und dauerhaften Speicherung, wie sie im Data Lake vorgesehen ist, zwei weitere Grundsätze entgegen: zum einen der *Grundsatz der Zweckbindung*, also dass die Daten nur für den Zweck, für den sie erhoben wurden, verwendet werden dürfen, und zum anderen der *Grundsatz der Speicherminimierung*, also dass nur so wenig Daten wie nötig erhoben werden dürfen.

1.1 Aufgabenstellung

Diese Arbeit widmet sich einem Teilaspekt des Themas Big Data und Datenschutz. Sie setzt sich mit der Fragestellung auseinander, welche Anforderungen sich aus der DSGVO beim Design und bei der Implementierung eines Data Lakes ergeben, wenn in diesem personenbezogene Daten - also der datenschutzrechtlichen Regulierung unterliegende Daten - gespeichert werden sollen.

Ziel dieser Arbeit ist daher zunächst die Analyse des Gesetzestextes und die Identifikation der Herausforderungen beim Entwurf eines DSGVO-konformen Data Lakes. Diese Herausforderungen werden anhand von Beispielen aus einem Anwendungsszenario veranschaulicht. Zudem werden auf Basis dieser Analyse Lösungsvorschläge erarbeitet und ein Konzept erstellt, das den Nutzer des Data Lakes bei der Einhaltung der Anforderungen der DSGVO unterstützt. Dieses wird anschließend prototypisch implementiert.

1.2 Gliederung

Diese Arbeit ist wie folgt strukturiert:

Kapitel 2 - Verwandte Arbeiten In diesem Kapitel werden bereits vorhandene Arbeiten vorgestellt. Dies betrifft insbesondere den Bereich Big Data und Datenschutz. Zudem werden Auslegungshilfen für die DSGVO vorgestellt.

Kapitel 3 - Grundlagen In diesem Kapitel werden die dem Thema zugrunde liegenden Konzepte Big Data, Data Lake und Datenschutz vorgestellt. Zudem werden die DSGVO als gesetzliche Implementierung des Datenschutzes, die juristische Zitierweise und datenschutzrechtliche Begrifflichkeiten eingeführt.

Kapitel 4 - Anwendungsszenario Um die folgende, rechtliche Diskussion mit Beispielen zu veranschaulichen, wird ein Anwendungsszenario konstruiert. Zudem werden Use Cases für die prototypische Implementierung vorgestellt.

Kapitel 5 - Pflichten des Verantwortlichen In diesem Kapitel werden die sich aus der DSGVO ergebenden Pflichten bei der Implementierung eines Data Lakes anhand von Anwendungsfällen diskutiert und Kollisionen zwischen den rechtlichen Anforderungen und dem Konzept des Data Lakes herausgearbeitet und veranschaulicht. Zudem werden Lösungsansätze erarbeitet.

Kapitel 6 - Konzept Auf Basis der erarbeiteten Lösungsvorschläge wird in diesem Kapitel zunächst eine Veränderung am Konzept des Data Lakes vorgeschlagen. Zuletzt wird ein Metadaten-Modell entwickelt, das den Nutzer bei der Einhaltung der gesetzlichen Pflichten unterstützt.

Kapitel 7 - Implementierung In diesem Kapitel wird eine prototypische Implementierung des Metadaten-Modells vorgestellt. Außerdem wird für die Use Cases eine prototypische Implementierung der Governance-Prozesse präsentiert.

Kapitel 8 - Diskussion der Implementierung Die prototypische Implementierung wird gegen das erwartete Verhalten überprüft, um zu zeigen, dass das entworfene Metadaten-Modell den Nutzer des Data Lakes bei der Einhaltung der Anforderungen der DSGVO unterstützt.

Kapitel 9 - Zusammenfassung und Ausblick Die Arbeit schließt mit einer Zusammenfassung und einem Ausblick auf weitere Forschungsmöglichkeiten, die sich im Bereich „Personenbezogene Daten im Data Lake“ eröffnen.

2 Verwandte Arbeiten

Zum DSGVO-konformen Data Lake gibt es derzeit noch kaum wissenschaftliche Analysen. Es existieren einzelne Übersichtsartikel oder Blogbeiträge, die Herausforderungen bei der Verarbeitung personenbezogener Daten im Data Lake herausgreifen und aufzeigen (z.B. [Bir18]). Zudem gibt es eine Bachelorarbeit [Kad15], in der auf Basis eines DSGVO-Entwurfs ein kommerzielles Data Lake Konzept gegen einige Anforderungen der DSGVO geprüft wurde. Es wurden Design-Erweiterungen und Änderungen an diesem Produkt vorgeschlagen, um in diesen Bereichen DSGVO-Konformität zu erreichen.

Auf dem Markt gibt es einige kommerzielle Anbieter, die einen DSGVO-konformen Data Lake vermarkten[Arc18; MT17]. Das oft verwendete Schlagwort dafür ist „Governed Data Lake“, auch wenn jeder Data Lake, der sich nicht in kürzester Zeit in einen „Data Swamp“ verwandeln soll, eine starke Governance benötigt [CJL+15]. Welche Aspekte der DSGVO für die Konformität betrachtet wurden und wie diese umgesetzt wurden, ist nicht öffentlich dokumentiert.

Die meisten Arbeiten zum Data Lake beschreiben die Einsatzmöglichkeiten und Architektur (z.B. [CJL+15; LS14; PWD17]) von Data Lakes. Datenschutz stellt dabei, wenn überhaupt, nur eine Randnotiz dar. Es wird lediglich darauf hingewiesen, dass es Regulierungen, wie z.B. den Datenschutz gibt, dass die Anforderungen eingehalten werden müssen und dass dafür entsprechende Governance-Prozesse benötigt werden (z.B. [CJL+15]).

Aus juristischer Perspektive scheint der Data Lake noch nicht wissenschaftlich betrachtet worden zu sein. Es existieren jedoch Arbeiten, die sich mit dem Phänomen Big Data beschäftigen. Da die DSGVO noch verhältnismäßig jung ist, beruhen viele Arbeiten noch auf der alten Rechtslage (z.B. [Wei13]). Ebenso können Arbeiten über das Schweizer Datenschutzrecht hinzugezogen werden, da das dortige Datenschutzrecht sich eng am europäischen orientiert (z.B. [HB14]).

Zunehmend gibt es auch Arbeiten, die sich mit Big Data unter der DSGVO beschäftigen. Zumeist werden die Unterschiede zwischen der alten und der neuen Rechtslage herausgearbeitet. Ein besonderer Schwerpunkt liegt auf Arbeiten, die sich mit einem Grundsatz des europäischen Datenschutzrechts, der Zweckbindung, auseinandersetzen [FHS17]. Dies beruht zum einen darauf, dass der Grundsatz der Zweckbindung für die explorative Analyse, die einen wichtigen Teil von Big Data darstellt, ein großes Hindernis darstellt und zum anderen darauf, dass sich hier interessante Änderungen zur alten Rechtslage bezüglich der Zweckänderung ergeben haben.

Neben dem Gesetzestext selbst, Urteilen und wissenschaftlichen Arbeiten, stellen Gesetzeskommentare eine wichtige Grundlage bei der Arbeit mit Gesetzen dar. Gesetzeskommentare werden auf Basis des Gesetzestextes, Dokumenten aus dem Gesetzgebungsprozess, Urteilen und rechtswissenschaftlichen Arbeiten geschrieben und setzen sich kritisch mit den möglichen Auslegungen des Gesetzes auseinander. Sie werden meist von mehreren Autoren aus verschiedenen Blickrichtungen

(wie z.B. Wissenschaft, Politik oder Praxis) geschrieben. So stellen Kommentare für den Anwender des Gesetzes eine wichtige Auslegungshilfe dar. In dieser Arbeit wird insbesondere auf den 2018 von Gierschmann, Schlender, Stentzel und Veil herausgegebenen „Kommentar Datenschutz-Grundverordnung“ [GSSV18] zurückgegriffen.

Speziell im Datenschutz gibt es darüber hinaus noch weitere Auslegungshilfen. Die Datenschutzaufsichtsbehörden haben nicht nur eine Aufsichtsfunktion, sondern haben gleichzeitig einen Beratungsauftrag inne. Es werden von den Behörden daher regelmäßig Papiere veröffentlicht, in denen die Rechtsauffassung der Datenschutzbehörden dargestellt wird und Handlungsempfehlungen ausgesprochen werden.

Zuerst sind in diesem Bereich die Leitlinien des Europäischen Datenschutzausschusses zu nennen. Dieser besteht aus Vertretern der nationalen Aufsichtsbehörden und erlässt Leitlinien um eine einheitliche Auslegung der DSGVO zu gewährleisten [AE17a; AE17b; AE18]. Diese werden auf der Webseite des Europäischen Datenschutzausschusses veröffentlicht [Eur18].

Die deutschen Aufsichtsbehörden arbeiten in der Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder (DSK) zusammen und veröffentlichen dort weitere Auslegungshilfen, wie Kurzpapiere [Kon17] oder Orientierungshilfen. Diese werden auf der Webseite der DSK veröffentlicht [Kon18].

Zudem werden Leitfäden, Diskussionsbeiträge und Praxisbeispiele wie [Bay17; Bay18] durch die einzelnen Aufsichtsbehörden der Bundesländer veröffentlicht.

3 Grundlagen

In einer Arbeit, die ein Systemkonzept aus der Informatik mit den Anforderungen, die sich aus einem Gesetz ergeben, vergleicht, müssen zunächst die technischen (Abschnitt 3.1) und rechtlichen (Abschnitt 3.2) Grundlagen des Themas „Personenbezogene Daten im Data Lake“ betrachtet werden.

3.1 Technische Grundlagen

In diesem Abschnitt wird das dem Data Lake zugrunde liegende Konzept von Big Data (Abschnitt 3.1.1) und der Data Lake (Abschnitt 3.1.2) selbst vorgestellt.

3.1.1 Big Data

Bei Big Data handelt es sich um keine neuartigen Daten oder Datenarten, vielmehr beschreibt der Begriff ein Phänomen, nach dem immer mehr Daten in immer diverserer Ausprägung immer schneller entstehen - es beschreibt zunächst also Charakteristiken der heute an vielen Stellen entstehenden Daten. Durch die Analyse dieser Daten kann Mehrwert generiert werden, allerdings bringen sowohl die Datenmenge als auch die unterschiedliche Natur der Daten klassische Analysewerkzeuge und Speichersysteme zunehmend an ihre Grenzen [GH15].

Bereits 2001 wird vom Marktforschungsinstitut Garnter in einer Analyse die Entwicklung zu dem, was heute als Big Data bezeichnet wird, vorweggenommen [Lan01]. Die dort beschriebenen Eigenschaften von Daten sind als die „drei V’s“ zur Definition von Big Data bekannt geworden:

Volume Volume beschreibt die Menge der Daten. Es entstehen in kürzeren Zeiträumen immer mehr Daten.

Velocity Velocity bezieht sich auf die Geschwindigkeit, in der Daten anfallen, verarbeitet und analysiert werden. Zudem wird der Zeitraum, in dem auf die Daten reagiert werden muss, immer kürzer.

Variety Variety bezieht sich auf die unterschiedliche Natur der entstehenden Daten. Es fallen strukturierte (z.B. Datenbanktabellen), semi-strukturierte (z.B. Logdateien) und unstrukturierte Daten (z.B. Bilder) an.

In den folgenden Jahren wurde die Idee der Charakterisierung von Big Data mit verschiedenen Vs beibehalten. Die drei Vs wurden zunächst durch ein viertes V ergänzt [SSS+12; ZDP+13]. Später wurde die Charakterisierung auf sechs [GH15], sieben [Fer17; Rij13] und sogar zehn Vs erweitert [Fir17; KAS+18]. Gemeinsam ist allen Erweiterungen der ursprünglichen Charakterisierung die Ergänzung um „Veracity“:

Veracity Bei Daten aus unterschiedlichen Quellen, insbesondere aus Quellen, die außerhalb der Kontrolle des Verarbeiters liegen (z.B. Social Media), stellt sich immer die Frage nach der Korrektheit der Daten und wie weit der Datenquelle vertraut werden kann. Die Korrektheit der Ausgangsdaten ist für den Wert und die Vertrauenswürdigkeit von Analyseergebnissen entscheidend.

Ein anderer Definitionsansatz für Big Data wird als das *HACE-Theorem* bezeichnet [WZWD14]. Demnach handelt es sich um Big Data, wenn große Datenmengen aus verschiedenartigen und unabhängigen Quellen verarbeitet werden, um komplexe und sich stetig verändernde Zusammenhänge zwischen den Daten zu entdecken („Big Data starts with **h**eterogenus and **a**utonomous sources […], and seeks to explore **c**omplex and **e**volving relationships among data“).

Die vorgestellten Definitionen von Big Data haben einige Gemeinsamkeiten: Es handelt sich um schnell enstehende, große und komplexe Datenmengen aus verschiedenen Quellen mit unterschiedlicher Datenstruktur und einem unterschiedlichem Level an Datenqualität. Diese Daten lassen sich aufgrund ihrer Diversität und großen Datenmenge mit den klassischen Werkzeugen zur Datenverarbeitung und -analyse nicht mehr verarbeiten, weshalb neue Ansätze benötigt werden.

Durch die Nutzung von Big Data entstehen viele Chancen. Es wird möglich, bisher unbekannte Zusammenhänge zu erkennen und daraus belastbare Vorhersagen für die Zukunft zu treffen (z.B. predictive Maintenance). Mit Big Data wird es zudem einfacher, Ressourcen zielgerichteter und sparsamer einzusetzen (z.B. Smart Cities). Zudem können Produkte durch eine Auswertung des Nutzungsverhaltens der Kunden verbessert und neue Produkte kreiert werden.

Mit diesen Chancen gehen jedoch auch Risiken einher. Grundsätzlich besteht die Herausforderung aus den gesammelten Daten tatsächlich einen Wert zu generieren. Werden Zusammenhänge in den Daten gefunden, so handelt es sich nicht unbedingt um Kausalitäten, sondern oft nur um Korrelationen. Zudem besteht die Gefahr, dass bei der Nutzung von Daten aus verschiedenen Quellen mit unterschiedlicher Datenqualität bei der Analyse Scheinkorrelationen auftreten. Außerdem müssen die aus den verschiedenen Datenquellen gesammelte Daten zusammengebracht und integriert werden, damit eine Auswertung möglich wird. Anderenfalls entstehen unverbundene Datensilos, die letztlich nur unnütze Datengräber sind.

Ein zentraler Kritikpunkt an Big Data ist die Gefahr, dass der Mensch zum gläsernen Menschen wird. Das bedeutet, dass - ob von staatlicher oder privater Seite - immer mehr Daten über den Einzelnen gesammelt werden und es keine Privatsphäre mehr gibt. Dadurch neigt der Mensch zu opportunem, also von der Gesellschaft oder dem Staat als passend erachteten, Verhalten und wird in seiner Freiheit als Individuum eingeschränkt. Ein Extrembeispiel, wohin dies führen kann, zeigt die derzeitige Diskussion um „Social Scoring“ in China [Dor17].

3.1.2 Data Lake

Ein Data Lake ist ein Speichersystem für Daten, insbesondere im Big Data Kontext. Es speichert Daten, im Gegensatz zu Data Warehouses, in ihrer originalen Rohform. Ein Data Lake ermöglicht es, Daten aus verschiedenen Quellen zu speichern und integrieren und soll so der Problematik voneinander getrennter Datensilos entgegenwirken. Im Data Lake ist es unerheblich, ob die Datenquellen strukturierte, semi-strukturierte oder unstrukturierte Daten erzeugen.

Da die Daten im Data Lake in ihrer ursprünglichen Form und nicht für bestimmte Use Cases aufgearbeitet gespeichert werden, können die Daten im Data Lake auch für andere, zum Zeitpunkt der Speicherung noch unbekannte Zwecke und Analysen genutzt werden. Teil des Konzepts des Data Lakes ist die Nutzung von explorativen Analysen, um zwischen den im Data Lake gespeicherten Daten neue Zusammenhänge zu erkennen und so einen Mehrwert zu generieren.

Damit Data Lakes diesen Mehrwert bieten können, müssen sie jedoch aus mehr als nur einem großen Speicherverbund bestehen. Ein Data Lake stellt eine Daten-Management-Plattform dar. Damit die zunächst einfach gespeicherten Daten tatsächlich zugänglich, auswertbar und nutzbar sind, wird ein Datenkatalog benötigt. In diesem werden die Daten katalogisiert und kategorisiert. Je besser der Datenkatalog organisiert ist, desto größer ist der Mehrwert, den ein Data Lake bieten kann. Dazu ist ein geeignetes Metadaten-Management unerlässlich, die Metadaten stellen den Kern des Datenkatalogs dar. Ebenso wird ein Governance-System benötigt, das den Datenzugriff kontrolliert und steuert, die Einhaltung von Richtlinien und Gesetzen erzwingt und die Datenqualität sicherstellt. Ohne diese Komponenten erhält man einen unnutzbaren Datenfriedhof, im Bereich von Data Lakes wird dann auch von einem Data Swamp („Datensumpf“) gesprochen.

Im Folgenden werden die drei Kernbestandteile eines Data Lakes, das Datenspeichersystem (Abschnitt 3.1.2), die Metadaten (Abschnitt 3.1.2) und das Governance-System (Abschnitt 3.1.2) vorgestellt.

Datenspeichersystem

Da nicht nur wichtige, sondern alle Daten gespeichert werden, sind niedrige Kosten eine Grundanforderung an das Datenspeichersystem eines Data Lakes. Das Speichersystem muss gut skalieren und in der Lage sein, Daten im Bereich vieler Petabytes zu speichern. Diese Anforderungen lassen sich nur mit einem horizontal skalierenden System erreichen, das heißt, es werden viele einfache, aber günstige Systeme miteinander verbunden, anstatt ein teures, aber sehr leistungsfähiges, zu nutzen.

Auch wenn im Detail unterschiedlich ausgestaltet, so liegt vielen Data-Lake-Architekturen die Unterteilung des eigentlichen Speichersystems in unterschiedliche Zonen [LS14; PWD17] zugrunde. Das in dieser Arbeit genutzte Zonenmodell basiert primär auf dem Modell von LaPlante und Sharma [LS14], die genaue Zonenabgrenzung folgt eher dem Modell von Patel, Wood und Diaz [PWD17]. Das Modell wird in Abbildung 3.1 dargestellt.

Landing Zone Die Daten aus den verschiedenen Quellsystemen werden zunächst in einer Landing Zone zwischengespeichert. Diese Zone ist nicht-persistent und dient als einheitlicher Eingang zum Data Lake. Sie ermöglicht es, die Einhaltung bestimmter Governance-Regeln zu erzwingen, sodass alle Daten, die in den persistenten Speicher des Data Lakes eingepflegt werden, bestimmten Mindeststandards unterliegen. Um die Daten in Zukunft möglichst flexibel einsetzen zu können, werden die Transformationen auf ein Minimum beschränkt. Beispiele für solche Transformationen sind:

Verschlüsselung Sofern Daten verschlüsselt gespeichert werden sollen, muss die Verschlüsselung angewandt werden, bevor die Daten im persistenten Speicher des Data Lakes gespeichert werden.

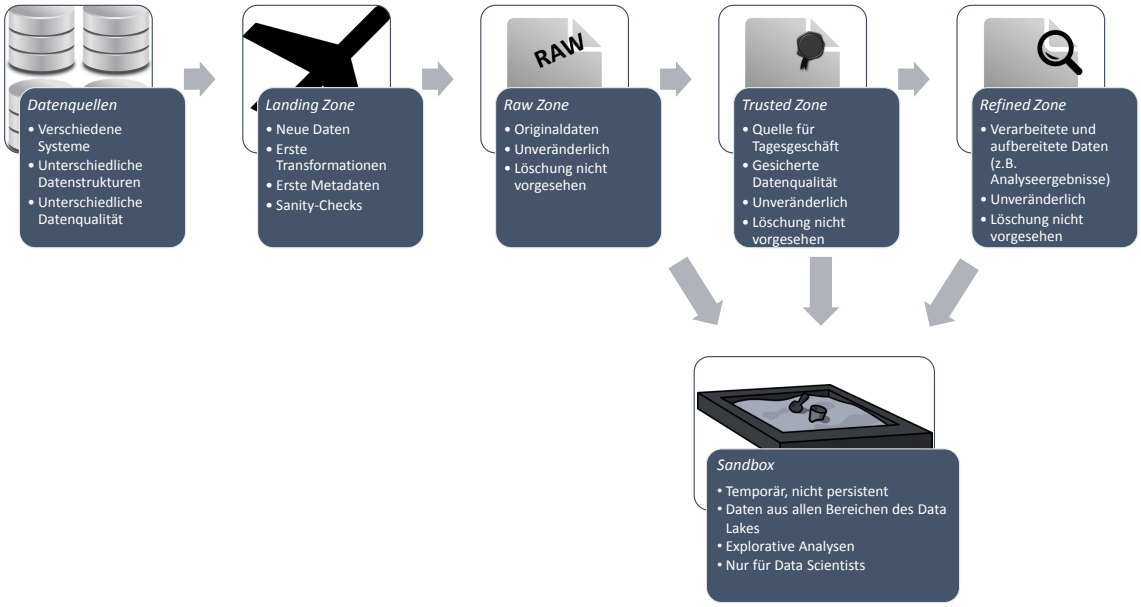

Abbildung 3.1: Zonenmodell für den Data Lake nach [LS14; PWD17]

Entfernung sensitiver Daten Sofern sensitive Daten, wie bspw. Gesundheitsdaten, Kreditkartendaten oder andere personenbezogene oder sensitive Daten nicht im Data Lake gespeichert sein sollen, müssen diese entfernt werden, bevor die Daten im persistenten Speicher gespeichert werden.

Datenbereinigung Sofern bestimmte Qualitätsanforderungen an Daten im Data Lake bestehen, kann hier eine Datenbereinigung erfolgen. Oft erfolgt dieser Schritt jedoch später und es werden zunächst alle Daten möglichst in der Originalform gespeichert.

Datenherkunft Für die weitere Nutzung und Analyse der Daten ist es meist unerlässlich zu wissen, aus welcher Quelle sie stammen, bzw. bei abgeleiteten Datensätzen, auf welchen Originaldaten der Datensatz beruht (sog. Datenprovenienz). Gegebenenfalls kann auch eine Sonderbehandlung oder gar Aussortierung von Daten unbekannter Herkunft angedacht werden.

Metadaten Für die weitere Nutzung der Daten im Data Lake sind Metadaten notwendig. Bei einer gesteuerten Datenaufnahme (engl. managed ingestion) können Metadaten bereits hinzugefügt werden, bevor die Daten im persistenten Speicher des Data Lakes gespeichert werden.

Raw Zone Nachdem die Daten in der Landing Zone angekommen sind und ggf. Transformationen durchgeführt wurden, die vor der persistenten Speicherung durchgeführt werden müssen, werden die Daten in die Raw Zone transferiert. Die Raw Zone enthält die Daten möglichst in ihrer Originalform, um flexibel weitere Analysen und Transformationen durchführen zu können. Die Raw Zone ist persistent und eine Veränderung der einmal dort gespeicherten Daten ist nicht vorgesehen.

Trusted Zone Die Daten in der Trusted Zone gelten als verlässlich und werden als Grundlage für das Tagesgeschäft genutzt. Bevor die Daten von der Raw Zone in die Trusted Zone transferiert werden, werden die Daten validiert und es wird sichergestellt, dass sie den vorgegebenen Qualitätsstandards entsprechen. Patel, Wood und Diaz sprechen von der den Daten in der Trusted Zone als „Quelle der Wahrheit“ für alle folgenden Systeme. Die Trusted Zone ist, wie die Raw Zone, persistent.

Refined Zone In der Refined Zone werden angereicherte und verarbeitete Daten gespeichert. Dabei handelt es sich z.B. um Analyseergebnisse oder Auswertungen von Daten aus der Trusted Zone. Auch die Refined Zone ist persistent.

Sandbox Neben der Landing Zone gibt es noch eine weitere nicht-persistente Zone im Data Lake, die Sandbox. Diese dient Data Scientists zur explorativen Analyse und kann, entsprechend den Wünschen und Zielen des Data Scientists, aus allen persistenten Zonen mit Daten gefüllt werden. Sie ist ein wichtiges Werkzeug um mehr Wertschöpfung aus den vorhanden Daten zu kreieren und bisher unbekannte Zusammenhänge zu entdecken. Sandboxes sind nur temporär. Entstehen hier neue Erkenntnisse, so werden die Analysen künftig im Hauptbereich des Data Lakes ausgeführt und die Analyseergebnisse in der Refined Zone gespeichert.

Implementierung Data Lakes werden typischerweise auf Basis des Hadoop-Frameworks implementiert [LS14]. Als Datenspeichersystem dient das Hadoop Distributed File System (HDFS). HDFS ist ein verteiltes und hochverfügbares Dateisystem zur Speicherung sehr großer Datenmengen. Es skaliert sehr gut horizontal und ermöglicht so den Einsatz günstiger und massenhaft verfügbarer Hardware.

HDFS beruht auf dem Konzept „write once, read many“ (WORM), das heißt, dass Daten, nachdem sie gespeichert wurden, unveränderlich sind [Kad15]. Veränderungen werden dadurch umgesetzt, dass eine veränderte Kopie der Datei erzeugt und die bisherige Datei als inaktiv markiert wird. Ebenso werden gelöschte Daten als inaktiv markiert.

Metadaten

Metadaten sind „Daten über Daten“. Sie enthalten Informationen über Daten, die diese Daten beschreiben und die es ermöglichen, die Daten zu nutzen und mit ihnen zu interagieren [Ril17]. Ein Alltagsbeispiel sind bei Musikdateien die sogenannten Tags wie Titel, Komponist, Künstler, Album. Ohne diese Informationen wäre es nicht möglich, nur die Musik eines bestimmten Künstlers oder eines Komponisten zu hören.

Metadaten werden auf vielfältige Weise gespeichert. Viele Dateiformate (z.B. PDF, Bilder oder Musikformate) ermöglichen es, die Metadaten direkt in der jeweiligen Datei zu speichern. Oft werden Metadaten auch durch die Namensgebung der Ordner und Dateien in der Dateisystemhierarchie abgelegt. Um diese Daten dann jedoch effektiv zu nutzen, werden sie durch bspw. eine Medienverwaltungssoftware einmalig ausgelesen und dann in einer Datenbank zur besseren Durchsuchbarkeit gespeichert und zur Verfügung gestellt. So wird es ermöglicht, auch große Sammlungen effizient zu durchsuchen und die gewünschten Dateien zu finden. Eine andere Möglichkeit zur Speicherung von Metadaten stellt die direkte Speicherung in einer Metadaten-Datenbank dar.

Die Metadaten im Data Lake lassen sich in drei Kategorien unterteilen [LS14]:

Technische Metadaten Technische Metadaten geben Auskunft über die Form und Struktur der Daten (z.B. Informationen über vorhandene Datenfelder und deren Datentypen). Zudem beinhalten sie Informationen, wie Daten zu lesen sind, z.B. wie Videos kodiert sind.

Betriebliche Metadaten Betriebliche Metadaten (engl. operational metadata) sind Metadaten, die im weitesten Sinne zum Betrieb des Data Lakes und des zugehörigen Governance-Systems benötigt werden. Dies reicht von Zugriffsberechtigungen über Informationen zur Datenqualität und Datenprovenienz bis hin zu Informationen, die zur Durchsetzung von Richtlinien oder Gesetzen benötigt werden.

Business Metadaten Bei Business Metadaten handelt es sich um die Metadaten, die der Nutzer benötigt, um Daten für die Analyse oder andere Nutzung zu finden (z.B. Beschreibung der Daten, Tags). Sie vermitteln dem Nutzer die Bedeutung der Daten.

Ohne Metadaten handelt es sich bei einem Data Lake um eine große, unüberschaubare Sammlung von Daten. Eine solche Datensammlung bietet keinen Mehrwert, sondern ist in den meisten Fällen schlicht nicht nutzbar. Um die Daten zugänglich zu machen, müssen sie daher mittels Metadaten erschlossen werden. Eine ähnliche Funktion wie die Datenbank der Medienverwaltungssoftware nimmt im Data Lake der Datenkatalog wahr. Er ermöglicht es anhand der Metadaten, bestimmte Daten zu finden, ihre Qualität und Herkunft einzuordnen und somit letztlich die Daten zu analysieren und nutzbar zu machen.

Governance-System

Das Governance-System dient dazu, die Einhaltung von Regeln im Data Lake sicherzustellen. Diese Regeln können unterschiedlichster Natur sein und reichen von betrieblich definierten Anforderungen (z.B. Datenqualität) über interne Richtlinien (Vertraulichkeit von Daten) bis hin zu gesetzlichen Vorgaben (z.B. Compliance-Regeln oder Datenschutz).

Entsprechend der vielfältigen Aufgaben setzt das Governance-System an unterschiedlichen Punkten an. Um bspw. die Datenqualität im Data Lake oder die Einhaltung der Zonenstruktur sicherzustellen, überwacht das Governance-System das Einfügen und Speichern von Daten. Um z.B. die Einhaltung von Zugriffsregeln sicherzustellen, überwacht und dokumentiert das Governance-System den Datenzugriff im Data Lake. Zudem kann es periodische Überprüfungen geben, die regelmäßig die Einhaltung von Regelungen sicherstellen.

Wie die Metadaten und der Datenkatalog ist auch das Governance-System existenziell für einen funktionierenden Data Lake. Ohne Governance wäre es nicht möglich, Daten im Data Lake zu speichern, auf die nur von einem eingeschränkten Nutzerkreis zugegriffen werden darf. Ebenso wären Analysen, die auf eine bestimmte Datenqualität angewiesen sind, nicht oder nur unter großem Aufwand möglich.

3.2 Rechtliche Grundlagen

In diesem Abschnitt wird zunächst die juristische Zitierweise aus Gesetzen erklärt (Abschnitt 3.2.1). Zudem werden die Idee des Datenschutzes (Abschnitt 3.2.2) und die DSGVO als gesetzliche Ausgestaltung des Datenschutzes (Abschnitt 3.2.3) vorgestellt. Zuletzt werden einige Begriffe aus dem Datenschutzrecht, die für das Verständnis der Arbeit wesentlich sind, erklärt (Abschnitt 3.2.4).

3.2.1 Zitieren aus Gesetzen

In einer Arbeit, die ein Systemkonzept aus der Informatik mit den Anforderungen, die sich aus einem Gesetz ergeben, vergleicht, treffen zwei Welten der Zitierweise aufeinander. Die Arbeit folgt an und für sich der in der Informatik üblichen Zitierweise. Soweit allerdings aus Gesetzen zitiert wird, wird dabei auf die juristische Zitierweise zurückgegriffen, da dies ein einfaches Auffinden der jeweiligen Fundstelle im Gesetz ermöglicht.

Gesetze werden in Klammer direkt an der jeweiligen Stelle im Text zitiert. Dabei wird mitangegeben, an welcher Stelle im Gesetz die Regelung zu finden ist. Die Formulierung *Art. 1 DSGVO* meint dabei Artikel 1 der DSGVO. Mit einer Zahl in Klammer hinter der Artikelnummer kann innerhalb eines Artikels auf einen bestimmten Absatz verwiesen werden, z.B. meint *Art. 1 (1) DSGVO* den Artikel 1, Absatz 1 der DSGVO. Um genauer auf eine Stelle innerhalb eines Absatzes verweisen zu können, gibt es verschiedene Möglichkeiten: es kann auf einen bestimmten Satz (z.B. Satz 2: S. 2) oder auf ein bestimmtes Element einer Aufzählung verwiesen werden (Bei Zahlennummerierung: Nr., bei Buchstabennummerierung: lit.).

Ebenso werden Gesetze zitiert, die nicht aus Artikeln, sondern aus Paragraphen bestehen. Dabei wird statt Art. das Paragraphzeichen § verwendet, so verweist z.B. §7 (3) UWG auf Paragraph 7, Absatz 3 des „Gesetz gegen den unlauteren Wettbewerb“.

Thematisch zusammengehörende Aspekte sind in Gesetzen manchmal an verschiedenen Stellen geregelt. Um zu verdeutlichen, dass sich eine Bestimmung oder ein Zusammenhang aus verschiedenen Stellen im Gesetz ergibt, wird die Formulierung *i.V.m.*, „in Verbindung mit“ genutzt.

Der DSGVO vorangestellt sind sogenannte Erwägungsgründe, die die Idee oder den Willen des Gesetzgebers bei bestimmten Normen verdeutlichen oder erklären sollen. Sie sind kein bindendes Recht, stellen aber eine wichtige Auslegungshilfe dar. Verweise auf die Erwägungsgründe zur DSGVO erfolgen mit der Abkürzung *ErwG*, so verweist z.B. *ErwG 1* auf den Erwägungsgrund 1 zur DSGVO.

3.2.2 Datenschutz

Der Begriff Datenschutz wird oft als „Schutz der Daten“ verstanden und umgangssprachlich mit dem Begriff der Informationssicherheit verwechselt. Der Datenschutz schützt jedoch keine Daten, sondern Menschen vor dem Missbrauch personenbezogener Daten.

Personenbezogene Daten sind alle Informationen, die sich auf eine natürliche Person, also einen Menschen, beziehen. Es ist dabei unerheblich, ob die Person mittels Namen direkt identifiziert ist, über eine Identifikationsnummer wie eine Matrikelnummer oder Personalausweisnummer oder über eine andere eindeutige Angabe wie beispielsweise „aktuelle Kanzlerin der Bundesrepublik Deutschland“ identifizierbar ist.

Es geht beim Datenschutz - je nach Herleitung - um den Schutz der Privatsphäre, den Schutz des Persönlichkeitsrechts oder um den Schutz der informationellen Selbstbestimmung. Dabei wird Datenschutz als individuelles Abwehrrecht gedacht, damit der Einzelne in seinen Rechten geschützt wird.

Zunehmend wird in der Literatur diskutiert, dass Datenschutz aus der Perspektive des Datenmachtproblems gedacht werden muss [Poh18]. Das meint, dass wer Daten und Informationen über andere Personen besitzt, Macht über diese hat, da er mithilfe dieser Daten das Verhalten dieser Personen beeinflussen kann. Als Ziel des Datenschutzes muss daher auch die gesellschaftliche Beherrschbarkeit der personenbezogenen Datenverarbeitung verstanden werden.

Geschichte der Datenschutzregulierung

Historisch wird Datenschutz in Deutschland vor allem als das Recht auf informationelle Selbstbestimmung verstanden. Informationelle Selbstbestimmung soll den Menschen auch im Digitalzeitalter ihre Entscheidungs- und Handlungsfreiheit erhalten [Bun83]. Wer keine Möglichkeit hat zu kontrollieren, wer bei welcher Gelegenheit welche Informationen über die eigene Person besitzt und ob das eigene Verhalten permanent dokumentiert oder festgehalten wird, kann keine selbstbestimmten Entscheidungen mehr treffen, sondern versucht, sich möglichst angepasst und unauffällig zu verhalten (sog. chilling effect oder Abschreckungseffekt). Dieses Grundrecht auf informationelle Selbstbestimmung ist nicht direkt im Grundgesetz (GG) verankert, sondern vom Bundesverfassungsgericht (BVerfG) im Volkszählungsurteil 1983 als Teil des Rechts auf freie Entfaltung der Persönlichkeit und der Menschenwürde hergeleitet (Art. 2 (1) i.V.m. Art. 1 (1) GG).

Die in diesem Urteil vom BVerfG aufgestellten Grundsätze sind bis heute grundlegender Teil des Datenschutzverständnisses. Das Gericht stellt fest, dass der Einzelne vor einer unbeschränkten Verarbeitung der ihn betreffenden Daten geschützt werden muss. Zudem wird klargestellt, dass es keine unwichtigen personenbezogene Daten gibt, sondern dass die Bedeutung vom Zweck der Verarbeitung und der möglichen Weiterverarbeitung oder Verkettung mit anderen Daten abhängt.

Dieses Urteil und das daraus abgeleitete Grundrecht stellten in der Folge die Grundlage für den deutschen Datenschutz, insbesondere die Überarbeitung des BDSG im Jahr 1990 dar [Mic17]. Die europäische Datenschutzrichtlinie von 1995 (vollständiger Titel: *Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr*) sollte EU-weite Mindeststandards für den Datenschutz festschreiben. Rechtliche Grundlage für die Regelungen der Datenschutzrichtlinie war das in der *Europäischen Menschenrechtskonvention* festgelegte Recht auf Achtung des Privatlebens. Die Datenschutzrichtlinie wurde in Deutschland durch die Novelle des BDSG im Jahr 2001 in nationales Recht überführt.

Seit 2009 gilt durch den Vertrag von Lissabon die *Charta der Grundrechte der Europäischen Union*. Diese postuliert, neben dem Recht auf Achtung des Privatlebens in Artikel 7, in Artikel 8 zusätzlich ein eigenes Grundrecht auf Datenschutz. Dieses ist die grundrechtliche Basis der europäischen *Datenschutzgrundverordnung*, die am 24. Mai 2016 in Kraft trat und nach zweijähriger Übergangsfrist seit dem 25. Mai 2018 anwendbar ist. Das BDSG wurde daher 2018 erneut vollständig überarbeitet und ist in seiner neuen Fassung gemeinsam mit der Anwendbarkeit der DSGVO in Kraft getreten.

3.2.3 Datenschutzgrundverordnung

Die DSGVO (vollständiger Titel: *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*) ist eine EU-Verordnung, also ein in den Mitgliedsstaaten unmittelbar geltendes Gesetz der Europäischen Union. Sie regelt die Verarbeitung personenbezogener Daten durch öffentliche (z.B. Behörden) und nicht-öffentliche (z.B. Unternehmen, Vereine) Stellen und vereinheitlicht die Regeln für die Verarbeitung personenbezogener Daten innerhalb der EU und des Europäischen Wirtschaftsraum (EWR). Der EWR umfasst neben den 28 EU-Staaten auch Liechtenstein, Norwegen und Island. Es handelt sich um eine Freihandelszone, in der es keine Zölle mehr gibt und in der die meisten der EU-Binnenmarkt-Vorschriften gelten.

Die Vorgängerregelung der DSGVO, die Datenschutzrichtlinie, war - im Gegensatz zur heute geltenden Verordnung - kein unmittelbar geltendes Recht, sondern musste von den Mitgliedsstaaten in nationalstaatliches Recht umgesetzt werden, in Deutschland durch das Bundesdatenschutzgesetz (alte Fassung). Die Richtlinie gab dabei nur Mindeststandards vor, wobei die nationalen Gesetzgeber jedoch auch strengere Regeln erlassen konnten. Dadurch entstand im gemeinsamen Binnenmarkt ein sehr uneinheitliches und für grenzüberschreitend tätige Unternehmen nur schwer einhaltbares Regelungsdickicht.

Das Ziel der Harmonisierung der Datenschutzvorschriften wurde jedoch nur teilweise erreicht. Die DSGVO enthält zahlreiche sogenannte Öffnungsklauseln, die vom jeweiligen nationalen Gesetzgeber genutzt werden können, um die Datenschutzvorschriften im jeweiligen Mitgliedsstaat zu konkretisieren. Die Existenz der Öffnungsklauseln hat verschiedene Gründe. Zum Teil beruht dies auf der fehlenden Regelungskompetenz der EU, da die Mitgliedstaaten nur bestimmte Kompetenzen an die EU abgetreten haben, und zum Teil auf Uneinigkeit und gewachsenen rechtlichen Traditionen, weshalb bestimmte Rechtsgebiete von den Mitgliedsstaaten selbst geregelt werden dürfen.

Die DSGVO hat insgesamt 99 Artikel und ist in elf Kapitel aufgeteilt:

Kapitel 1 Allgemeine Bestimmungen
: In Kapitel 1 wird definiert, was durch das Gesetz geregelt werden soll und welche Ziele damit verfolgt werden. Zudem wird festgelegt, wo und unter welchen Umständen das Gesetz angewendet werden muss (sog. räumlicher und sachlicher Anwendungsbereich). Das Kapitel schließt mit Begriffsdefinitionen, die im Gesetz genutzte Begrifflichkeiten definieren und erklären.

Kapitel 2 Grundsätze
: In Kapitel 2 werden Grundsätze und Grundregeln festgelegt, die bei der Verarbeitung personenbezogener Daten beachtet werden müssen.

Kapitel 3 Rechte der betroffenen Person
: Werden personenbezogene Daten über eine Person verarbeitet, so hat diese gegenüber dem für die Verarbeitung Verantworlichen bestimmte Rechte (z.B. Recht auf Auskunft über die Verarbeitung). Außerdem hat der Verantwortliche Pflichten gegenüber dieser Person (z.B. Informationspflichten). Diese Rechte und Pflichten sind in Kapitel 3 geregelt.

Kapitel 4 Verantwortlicher und Auftragsverarbeiter
: Der Verantwortliche für die Verarbeitung personenbezogener Daten unterliegt zudem Pflichten, wie personenbezogene Daten verarbeitet werden dürfen und welche Rahmenbedingungen er dafür beachten muss (z.B. Anforderungen an die IT-Sicherheit). Werden Teile der Verarbeitung an einen Dienstleister ausgelagert, so handelt es sich um eine Auftragsverarbeitung, das heißt der Dienstleister handelt im Auftrag und auf Weisung des Verantwortlichen. Das Verhältnis zwischen Verantwortlichem und Auftragsverarbeiter unterliegt gesetzlichen Anforderungen. Diese Pflichten und Anforderungen werden in Kapitel 4 geregelt.

Kapitel 5 Übermittlungen personenbezogener Daten an Drittländer oder an internationale Organisationen
: Werden personenbezogene Daten in Länder außerhalb der EU und des EWR oder an internationale Organisationen (z.B. UN-Organisationen) übermittelt, so besteht die Gefahr, dass diese Daten keinen oder zumindest keinen ausreichenden datenschutzrechtlichen Schutz mehr genießen, da dort europäisches Recht keine Anwendung findet. Daher werden in Kapitel 5 für diese Übermittlungen spezielle Regelungen getroffen, die auch dann ein entsprechendes Datenschutzniveau sicherstellen sollen.

Kapitel 6 Unabhängige Aufsichtsbehörden
: In Kapitel 6 wird die Einrichtung unabhängiger Datenschutzaufsichtsbehörden durch die Mitgliedsstaaten gefordert. Es wird definiert, was Unabhängigkeit bedeutet und festgelegt, welche Regelungen die Mitgliedsstaaten bezüglich der Aufsichtsbehörden zu treffen haben und wie die Besetzung der entsprechenden Stellen erfolgen muss. Zudem werden die Aufgaben, Pflichten und Befugnisse der Aufsichtsbehörden festgelegt.

Kapitel 7 Zusammenarbeit und Kohärenz
: Da in jedem Mitgliedsstaat eigene Aufsichtsbehörden eingerichtet werden, die alle die Einhaltung einheitlichen EU-Rechts überwachen sollen, werden in Kapitel 7 Mechanismen zur Zusammenarbeit der Behörden und der Sicherstellung einheitlicher Rechtsauslegung (sog. Kohärenzverfahren) festgelegt.

Kapitel 8 Rechtsbehelfe, Haftung und Sanktionen
: In Kapitel 8 wird geregelt, wie Betroffene gegen Verstöße gegen das Gesetz vorgehen können und wie sich Betroffene und Verantwortliche gegen das Verhalten einer Aufsichtsbehörde zur Wehr setzen können. Es werden zudem der Schadensersatzanspruch und Bußgelder im Falle eines Verstoßes gegen die DSGVO geregelt. Außerdem werden die Mitgliedsstaaten aufgefordert, den Aufsichtsbehörden weitere Sanktionsmöglichkeiten außer einem Bußgeld zur Verfügung zu stellen.

Kapitel 9 Vorschriften für besondere Verarbeitungssituationen
: In Kapitel 9 werden Spezialregelungen für bestimmte Kontexte der Verarbeitung personenbezogener Daten getroffen, z.B. bei Geheimhaltungspflichten oder im Beschäftigungsverhältnis.

Kapitel 10 Delegierte Rechtsakte und Durchführungsrechtsakte
: Bei delegierten Rechtsakten und Durchführungsrechtsakten handelt es sich um spezielle Regelungen des EU-Rechts, in denen der EU Kommission in engem Rahmen Regelungskompetenzen übertragen werden können, obwohl die Kommission die Exekutive der EU ist. In Kapitel 10 werden die Regeln für die Nutzung dieser Kompetenzen festgelegt.

Kapitel 11 Schlussbestimmungen
: Die Schlussbestimmungen in Kapitel 11 umfassen allgemeine Regelungen wie das Inkrafttreten der DSGVO, die Aufhebung der Vorgängerregelung und ähnliches.

Besonders bemerkenswert ist bei den allgemeinen Regelungen der örtliche Geltungsbereich. Grundsätzlich gelten Gesetze meist für Menschen oder Unternehmen, die sich in einem Land aufhalten oder dort ihren Sitz haben (sog. Territorialprinzip). Die DSGVO dagegen hat einen sehr weiten Geltungsbereich. Sie gilt, sobald das Angebot sich an Betroffene innerhalb des Geltungsbereichs der DSGVO richtet oder das Verhalten von Menschen im Geltungsbereich beobachtet werden soll (Art. 3 (2) DSGVO). Dies trägt der Tatsache Rechnung, dass Datenverarbeitung im Allgemeinen, wie auch die Verarbeitung personenbezogener Daten, zunehmend global und international erfolgt, weshalb für einen effektiven Grundrechtsschutz ein weitreichender Geltungsbereich benötigt wird.

3.2.4 Begriffsdefinitionen

Der Begriff der *Person* wird in dieser Arbeit zu Gunsten des besseren Verständnisses als vereinfachende Formulierung für natürliche Personen verwendet und meint schlicht einen Menschen, da nur diese das Recht auf Datenschutz haben.

Juristische Personen sind Organisationen (z.B. Vereine, Stiftungen, AGs, GmbHs), die rechtlich selbstständig sind. Diese rechtliche Selbstständigkeit wird vom Gesetz zuerkannt und bedeutet, dass diese Organisationen z.B. selbst klagen oder verklagt werden können. Sind juristische Personen gemeint, so werden diese explizit so bezeichnet.

Im Datenschutzrecht werden einige Begriffe genutzt, die auch in dieser Arbeit unerlässlich sind. Diese werden in der DSGVO insbesondere im Artikel 4 und hier im Folgenden kurz definiert:

Personenbezogene Daten Als personenbezogene Daten werden alle Daten bezeichnet, die einer Person direkt (z.B. da der Name bekannt ist) oder indirekt (z.B. über ein Kennzeichen) zugeordnet werden können (Art. 4 Nr. 1 DSGVO).

Die Personenbeziehbarkeit ist immer aus der Sicht des Verantwortlichen zu sehen. Daten sind also nicht nur deshalb personenbezogen, weil ein Dritter diese einer Person zuordnen kann, sondern er benötigt auch die Mittel, diese Zuordnung grundsätzlich durchzuführen [Eur16].

Besondere Kategorien personenbezogener Daten Die besonderen Kategorien personenbezogener Daten sind besonders sensible personenbezogene Daten, deren Verarbeitung den Betroffenen stigmatisieren oder benachteiligen könnten. Die Aufzählung der umfassten Daten im Gesetz ist abschließend. Es handelt sich dabei um Daten, aus denen die „rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen

oder die Gewerkschaftszugehörigkeit“ hervorgeht, sowie biometrische und genetische Daten, Gesundheitsdaten, Daten zum Sexualleben und der sexuellen Orientierung (Art. 9 (1) DSGVO).

Verarbeitung Eine Verarbeitung ist jeder Vorgang, der mit personenbezogenen Daten geschehen kann, wie z.B. Erhebung, Speicherung, Verwendung, Übermittlung oder Löschung, unabhängig davon, ob dies automatisiert oder manuell erfolgt (Art. 4 Nr. 2 DSGVO). Ebenso kann eine Vorgangsreihe dieser Vorgänge eine Verarbeitung darstellen.

So können beispielsweise alle Vorgänge, die zum Newsletterversand benötigt werden, zu einer Verarbeitung zusammengefasst werden. Diese Verarbeitung umfasst dann sowohl das Auslesen der gespeicherten Daten als auch die Verwendung für den Versand.

Verantworlicher Verantwortlicher ist, wer über die Mittel und Zwecke der Verarbeitung personenbezogener Daten entscheidet. Verantwortlich können sowohl natürliche als auch juristische Personen wie z.B. Unternehmen sein (Art. 4 Nr. 7 DSGVO).

Betroffener Der Betroffene, alternativ die betroffene Person, ist derjenige, auf den die personenbezogenen Daten beziehbar sind (Art. 4 Nr. 1 DSGVO).

Technische und organisatorische Maßnahmen Technische und organisatorische Maßnahmen oder kurz Maßnahmen (TOMs) , meint alles, was der Verantwortliche tut, um eine dem Datenschutzrecht entsprechende Verarbeitung personenbezogener Daten sicherzustellen.

Der Begriff „technische Maßnahmen“ bezieht sich dabei auf alle physisch oder in Computersystemen umsetzbare Maßnahmen, wie z.B. die Zugangskontrolle zu Räumen oder die Verschlüsselung von Daten. „Organisatorische Maßnahmen“ umfassen Maßnahmen in der Organisation, wie z.B. Schulung der Mitarbeiter oder¸ entsprechende Arbeitsanweisungen.

Vereinfachend wird in der Arbeit von „den Daten des Betroffenen“ gesprochen. Allerdings gibt es kein Dateneigentum, die Formulierung ist eine der Lesbarkeit dienenden Verkürzung der eigentlichen Formulierung „die den Betroffenen betreffenden personenbezogenen Daten“.

4 Anwendungsszenario

Um die in Kapitel 5 folgende Diskussion der sich aus der DSGVO ergebenden Pflichten mit Beispielen zu veranschaulichen, wird in diesem Kapitel ein beispielhaftes Anwendungsszenario vorgestellt. Zudem werden Use Cases für die Implementierung (Kapitel 7) erstellt.

Neben den in der Einleitung angesprochenen Einsatzbereichen kommen Big-Data-Analysen auch beispielsweise im Onlinehandel zum Einsatz [Som13]. Ziel, vor allem direkt dem Kunden gegenüber, ist eine optimierte Werbeansprache, um z.B. passende Werbemails zu versenden oder geeignete Produkte als „ähnliche Produkte“ im Onlineshop anzubieten und so die Werbekonvertierungsrate zu erhöhen. Aber auch hinter den Kulissen lohnt sich der Einsatz von Big-Data-Analysen. So ermöglicht eine Analyse des Kaufverhalten aller Kunden eine optimierte Lagerhaltung und -sortierung. Ebenso kann der saisonale Bedarf genauer prognostiziert werden. Ein dritter Aspekt ist der Einsatz von Big-Data-Analysen zur Fraud Prevention (Betrugsprävention). Dazu wird bei der Bezahlung im Hintergrund über verschiedene Mechanismen (z.B. Bestelllimits pro Stunde, Bonitätsauskünfte) überprüft, ob eine Bestellung möglicherweise ein Betrugsversuch ist.

Als Anwendungsszenario wird im Folgenden ein fiktiver, global agierender Onlinehändler angenommen. Im Onlinehandel fallen in den unterschiedlichsten Systemen Daten an, diese Daten sollen in einem zentralen Data Lake gespeichert werden. Abbildung 4.1 zeigt einige der Quellsysteme des Data Lakes. Die Quellsysteme decken dabei den ganzen Geschäftsbereich des Onlinehändlers ab, vom Kontakt mit dem Kunden (z.B. Bestellsystem, Kundensupport, Social Media Kanäle, Tracking-Tools auf der Webseite) über Kontakt mit den Lieferanten (Einkaufssystem) bis hin zu internen Systemen (z.B. Lagerverwaltung oder Mitarbeiterverwaltung). Dabei fallen in diesen Systemen ganz unterschiedliche Datenarten an. Im Rezensionssystem und über die Social Media Kanäle entstehen unstrukturierte Daten, wie z.B. Texte, Bilder oder Videos. In Systemen wie dem des Kundensupports entstehen dagegen semistrukturierte Daten, während Systeme wie die Kundenverwaltung oder Mitarbeiterverwaltung auf strukturierten Daten beruhen.

Die Integration verschiedener, zuvor getrennter Datenquellen ermöglicht neue Analysemöglichkeiten. So kann bspw. die Auswertung des durch Tracking-Tools aufgezeichneten Surfverhaltens eines Kunden im Onlineshop in Verbindung mit dem Kaufverhalten Hinweise geben, weshalb der Kunde einen Kauf letztlich nicht getätigt hat. Dadurch kann der Onlineshop den Anforderungen der Kunden entsprechend optimiert werden.

Use Cases

Im Folgenden werden die Use Cases für die Implementierung des Prototypen (Kapitel 7) vorgestellt und das erwartete Verhalten des Data Lakes, konkreter der Governance-Prozesse, beschrieben.

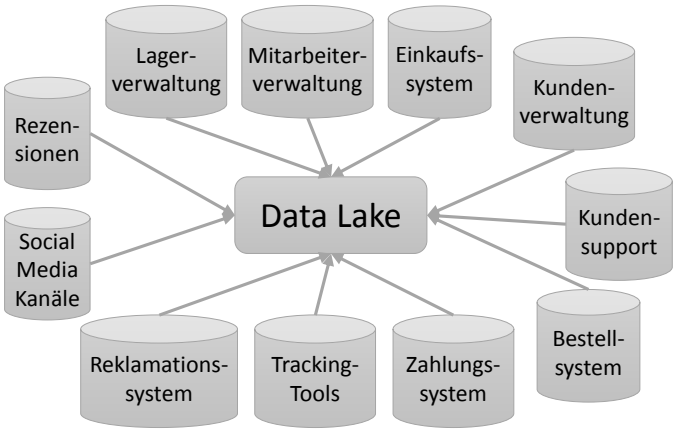

Abbildung 4.1: Quellsysteme für den Data Lake bei einem Onlinehändler

Datenidentifikation von Daten über Betroffenen Für die Erfüllung von Betroffenenrechtsansprüchen (z.B. Recht auf Erhalt einer Datenkopie) ist es erforderlich, dass alle einer Person zugeordneten personenbezogenen Daten identifiziert werden können.

Erwartetes Verhalten: Das System sollte zunächst den Betroffenen eindeutig identifizieren. Wenn dies gelingt, sollte es eine Auflistung aller Daten liefern, die zu dieser Person gespeichert sind. Daten, die mehreren Personen zugeordnet sind, sollten getrennt ausgewiesen werden, da bei diesen auch der Datenschutz anderer Personen betroffen ist.

Newsletter versenden Um einen Newsletter zu versenden, werden alle für diese Verarbeitung und diesen Verarbeitungszweck benötigten Daten, konkret Name und E-Mail-Adresse, angefordert.

Erwartetes Verhalten: Das System sollte, unter Berücksichtigung von z.B. Widersprüchen des Betroffenen oder Widerruf der Einwilligung, eine Liste der benötigten Daten generieren.

Newsletterabmeldung Ein potentieller Kunde meldet sich vom Newsletter ab und seine Daten werden zu keinem anderen Zweck mehr verarbeitet, es gibt keine weitere Geschäftsbeziehung mehr. Zudem verlangt er die Löschung seiner Daten.

Erwartetes Verhalten: Die hinterlegte Einwilligung für den Newsletterempfang sollte gelöscht werden und der Betroffene darf für den weiteren Newsletterversand nicht mehr berücksichtigt werden.

Das System muss prüfen, ob das Datum für weitere Zwecke benötigt wird. Falls dem, wie in diesem Fall, nicht so ist und somit keine Rechtsgrundlage für die Speicherung mehr besteht, muss das Datum gelöscht werden (vgl. Abschnitt 5.3.3). Zudem muss eine Liste der Empfänger der Daten generiert werden, die über das Löschbegehren zu informieren sind (vgl. Abschnitt 5.3.5).

Einschränkung der Verarbeitung Ein Kunde hat gegenüber dem Onlinehändler erwirkt, dass die Verarbeitung eines Datums eingeschränkt werden muss, das heißt für die weitere Verwendung gesperrt werden muss (vgl. Abschnitt 5.3.4). Nun soll auf dieses Datum zugegriffen werden.

Erwartetes Verhalten: Das System sollte den Datenzugriff mit Hinweis auf die Einschränkung der Verarbeitung blockieren.

Streit um Einwilligung Ein potentieller Kunde erhält einen Newsletter auf Basis einer Einwilligung. Er beschwert sich beim Unternehmen, da er meint, keine Einwilligung erteilt zu haben.

Erwartetes Verhalten: Das System muss in der Lage sein zu überprüfen, ob eine Einwilligung gespeichert ist und diese, so sie vorhanden ist, auszugeben.

5 Pflichten des Verantwortlichen

Aus der Datenschutzgrundverordnung ergeben sich für den Verantwortlichen, der personenbezogene Daten verarbeitet, vielfältige Pflichten. Diese gelten damit auch im Data Lake, sofern dort personenbezogene Daten verarbeitet werden. Verstöße gegen die DSGVO sind bußgeldbewehrt und können, je nach Pflicht, mit Bußgeldern von bis zu 10 Millionen Euro, bzw. 2% des weltweiten Jahresumsatzes oder bis zu 20 Millionen Euro, bzw. 4% des weltweiten Jahresumsatzes geahndet werden.

Die Pflichten des Verantwortlichen hat Dr. Winfried Veil, der für die Bundesregierung an den Verhandlungen zur DSGVO beteiligt war, in einer Übersicht [Vei18] zusammengestellt. Diese Darstellung eignet sich besonders, um die Kompatibilität der Regulierungen der DSGVO und des Konzepts Data Lake zu prüfen und bildet die Grundlage für dieses Kapitel.

Nicht alle dort aufgeführten Pflichten sind im Kontext der Erstellung und des Designs eines Data Lakes, der die Anforderungen der DSGVO erfüllt, relevant. Die meisten Benachrichtigungspflichten im Rahmen der Erfüllung von Betroffenenrechten sind beispielsweise nicht beim Design des Data Lakes, sondern erst im darauffolgenden Prozess der Bearbeitung von Betroffenenanfragen relevant. Diese Arbeit beschränkt sich auf die für das Design eines Data Lakes relevanten Pflichten.

Die Pflichten werden zum Teil etwas verkürzt dargestellt, da mit dieser Arbeit ein grundsätzliches Verständnis für die Regelungen der DSGVO und deren Bedeutung im Kontext *Design eines Data Lakes* geschaffen werden soll. Ob und wie eine Pflicht in bestimmten Konstellationen aufgrund z.B. einer Ausnahme in der DSGVO im jeweiligen Einzelfall zu handhaben ist, kann und soll nicht Thema dieser Arbeit sein.

Betrachtet man die Anforderungen der DSGVO im Hinblick auf den Data Lake, so gibt es grundsätzlich zwei Perspektiven, die beachtet werden müssen:

Data Lake als Mittel der Verarbeitung Aus Sicht der bereits etablierten Geschäftsprozesse und Verarbeitungen handelt es sich beim Data Lake zunächst nur um ein Speichersystem, das zur Speicherung der erzeugten Daten, bzw. als Datenquelle dient.

Sofern diese Verarbeitungen personenbezogene Daten im Data Lake speichern oder daraus lesen, unterfällt der Data Lake als Mittel der Verarbeitung den Anforderungen der DSGVO (detailliert in Abschnitt 5.4).

Data-Lake-Verarbeitungen Mit dem Konzept des Data Lakes gehen, neben der Speicherung von Daten, einige Verarbeitungen dieser Daten einher, die zwar streng genommen nicht Teil des Speichersystems Data Lake sind, jedoch Teil des Ökosystems Data Lake sind und einen wichtigen Grund für die Entscheidung für einen Data Lake als Speichersystem darstellen. Diese oft unter dem Schlagwort „Advanced Analytics“ zusammengefassten Analysemethoden,

die insbesondere die zweckunabhängige Untersuchung und Analyse der gespeicherten Daten umfassen, sollen helfen, neue Geschäftsmodelle und bessere Vorhersagemodelle für das Markt- oder Kundenverhalten zu kreieren.

Sofern im Data Lake personenbezogene Daten gespeichert werden und diese Daten untersucht und analysiert werden sollen, so stellen diese Untersuchungen und Analysen selbst eigenständige Verarbeitungen personenbezogener Daten dar und unterfallen damit sämtlichen Regelungen der DSGVO.

Beide Aspekte der Verarbeitung personenbezogener Daten im Data Lake werden im Folgenden betrachtet.

In den folgenden Unterkapiteln werden jeweils zunächst die betrachteten Pflichten für den Verantwortlichen vorgestellt und erklärt. Danach wird die Problematik im Bezug auf Data Lakes herausgearbeitet, soweit möglich an einem Beispiel aus dem Anwendungsszenario aus Kapitel 4 erläutert, und schließlich werden Lösungsvorschläge entwickelt.

5.1 Grundsätze der Verarbeitung personenbezogener Daten

Die erste inhaltliche Norm der DSGVO formuliert die *allgemeinen Grundsätze der Verarbeitung personenbezogener Daten* (Art. 5 (1) DSGVO). Diese Regelung hat eine zentrale Bedeutung in der DSGVO und gibt den Rahmen vor, in dem personenbezogene Daten verarbeitet werden dürfen. Die Formulierung dieser Grundsätze ist sehr abstrakt, nichtsdestotrotz ist die Einhaltung verpflichtend und bei Nicht-Einhaltung drohen hohe Bußgelder. Diese Grundsätze werden in weiteren Artikeln der DSGVO konkretisiert. Die Grundsätze der Rechtmäßigkeit (Abschnitt 5.1.1), der Fairness (Abschnitt 5.1.2) und der Zweckbindung (Abschnitt 5.1.4) sind bereits in der *Charta der Grundrechte der Europäischen Union* festgelegt.

5.1.1 Rechtmäßigkeit

Die Verarbeitung personenbezogener Daten unterliegt einem sogenannten *Verbot mit Erlaubnisvorbehalt*, das heißt, dass die Verarbeitung personenbezogener Daten grundsätzlich verboten ist und die Verarbeitung nur aufgrund einer Rechtsgrundlage, einer im Gesetz geregelten Erlaubnis, erfolgen darf. Dieser Grundsatz wird mit den einzelnen Rechtsgrundlagen und deren Ausgestaltung in den Artikeln 6-9 der DSGVO und im Abschnitt Abschnitt 5.2 in dieser Arbeit konkretisiert.

Es gibt zwischen Verarbeitungen und den dafür erhobenen personenbezogenen Daten und der Rechtsgrundlage keine 1:1-Beziehung, sondern eine Verarbeitung darf auf Basis unterschiedlicher Rechtsgrundlagen für unterschiedliche Zwecke erfolgen.

Problem Für jede Verarbeitung personenbezogener Daten wird zumindest eine Rechtsgrundlage benötigt. Dies muss im Data Lake aus zwei Perspektiven betrachtet werden.

Betrachtet man den Data Lake zunächst als Speichersystem, in dem Daten aus verschiedenen Quellen und Verarbeitungen gespeichert werden, so ist nicht die Wahl der Rechtsgrundlage, sondern die technische Unterstützung, dass Daten nur rechtskonform verarbeitet werden, interessant.

Die andere Perspektive wirft einen Blick auf die Analysen, die auf den Daten im Data Lake durchgeführt werden sollen. Jede dieser Analysen ist, wie oben bereits ausgeführt, eine eigenständige Verarbeitung, die folglich auch eine eigene Rechtsgrundlage benötigt.

Anwendungsszenario Die E-Mail-Adresse eines Kunden wird zur Vertragsabwicklung gespeichert, damit dem Kunden Bestellbestätigungen, Rechnungen und Informationen über den Versand zu gesendet werden können. Im Rahmen der Bestellung werden die Kunden nach einer Einwilligung für Werbe-E-Mails gefragt. Sollen nun alle im Data Lake gespeicherten E-Mail-Adressen für eine Werbekampagne für ein neues Produkt genutzt werden, so sollte das System die E-Mail-Adressen, die nicht für die Kampagne genutzt werden dürfen, herausfiltern.

Lösungsansatz Die Verantwortung dafür, dass für jede Verarbeitung personenbezogener Daten eine Rechtsgrundlage vorliegt, bleibt beim Verantwortlichen. Das Governance-System des Data Lakes sollte ihn jedoch darin unterstützen, die gesetzlichen Rahmenbedingungen einzuhalten. Das heißt, es sollte beispielsweise prüfen, ob bei der Verarbeitung aufgrund einer Einwilligung tatsächlich die Einwilligung für diese Verarbeitung vorliegt. In einer einfachen Form kann dies mittels Metadaten geschehen, in komplexeren und größeren Systemen sollte dafür ein Permission Management System genutzt werden, das die vorliegende Nutzungserlaubnis von Daten gegen den Anwendungsfall prüft.

Die Wahl der Rechtsgrundlage für eine Data Lake Verarbeitung kann durch den Data Lake nicht direkt unterstützt werden, da die rechtliche Abwägung, ob und auf welcher Basis Daten verarbeitet werden dürfen, nicht automatisiert werden kann. In Abschnitt 5.2 werden die möglichen Rechtsgrundlagen für die Verarbeitung personenbezogener Daten detaillierter diskutiert. Die Einhaltung der gewählten Rechtsgrundlage sollte durch das Governance-System unterstützt werden.

Da der Anknüpfungspunkt für das Datenschutzrecht und damit auch für das Verbot mit Erlaubnisvorbehalt der Personenbezug von Daten ist, stellt die Anonymisierung einen naheliegenden Weg dar, um diesen Regelungen zu entgehen. Dies ist jedoch mit einigen Herausforderungen verbunden. Zunächst muss dabei eine echte Anonymisierung sichergestellt werden. Dient bspw. eine Telefonnummer als Identifizierungsmerkmal (z.B. bei einem Instant Messenger), so stellt das Hashen der Telefonnummer keine Anonymisierung dar, da es problemlos möglich ist, alle gültigen Telefonnummern aufzuschreiben und zu hashen, wodurch die Telefonnummer wiederhergestellt werden kann.

Zudem wird eine echte Anonymisierung, die dem Verantwortlichen keinen Rückschluss mehr auf die Betroffenen erlaubt, im Zeitalter von Big Data zunehmend erschwert. Durch die Speicherung von immer mehr Datensätzen und die Verknüpfung verschiedener Quellen, ist eine De-Anonymisierung der Daten oft möglich [Mar14]. Ein prominentes Beispiel dafür stellt die anonymisierte Netflix-Datenbank dar, die mithilfe öffentlich verfügbarer Daten de-anonymisiert wurde [NS06].

Dieser Ansatz der „Flucht aus dem Datenschutzrecht“ wird in dieser Arbeit nicht weiterverfolgt, sondern wurde nur der Vollständigkeit halber erwähnt. In dieser Arbeit werden stattdessen die Anforderungen an den Data Lake untersucht, um personenbezogene Daten dort verarbeiten zu können.

5.1.2 Fairness („Treu und Glauben“)

Der Grundsatz der Fairness gegenüber dem Betroffenen wird im deutschen Gesetz als „Treu und Glauben“ bezeichnet. Der Betroffene sollte durch Datenverarbeitungen des Verantwortlichen nicht überrascht werden, sondern im Gegenteil von Anfang an wissen, worauf er sich einlässt. Dieser Gedanke spielt insbesondere bei der Frage nach der Zulässigkeit der Zweckänderung eine Rolle.

Problem Die Fairness stellt die Verarbeitung von personenbezogenen Daten im Rahmen der Data-Lake-Verarbeitungen vor Probleme. Die im Konzept des Data Lakes verankerte Idee der explorativen Datenanalyse widerspricht grundlegend der Idee, dem Betroffenen gegenüber fair zu sein, das heißt seine personenbezogenen Daten nur in einem Rahmen zu verarbeiten, „wie er damit rechnen konnte“.

Anwendungsszenario Bietet der Onlinehändler die Möglichkeit zur Bestellung ohne Kundenkonto, legt aber im Hintergrund trotzdem ein Schattenprofil des Kunden an, als ob er sich registriert hätte, so handelt es sich um eine unfaire Datenverarbeitung. Der Betroffene konnte nicht mit dieser Art der Datenverarbeitung rechnen, im Gegenteil hat er durch die Bestellung ohne Kundenkonto versucht einer derartigen Datenverarbeitung zu entgehen.

Lösungsansatz Die Verknüpfung verschiedenster Daten aus unterschiedlichen Quellen und die Analyse der Daten dürfte in den meisten Fällen nicht mit dem Grundsatz der Fairness kompatibel sein. Insbesondere gilt dies im Kontext der Erstellung von Persönlichkeitsprofilen. Ob eine Analyse im Einzelfall zulässig ist, muss jeweils manuell überprüft werden.

Soweit möglich sollte, auch im Hinblick auf andere Grundsätze wie die Datenminimierung, auf Anonymisierung oder Pseudonymisierung der Daten für die Analyse zurückgegriffen werden. Dies eröffnet im Hinblick auf die Fairness mehr Möglichkeiten zur Analyse von Daten.

5.1.3 Transparenz (Nachvollziehbarkeit)

Das Gesetz beschränkt sich beim Grundsatz der Transparenz auf die Formulierung, dass die Verarbeitung personenbezogener Daten „in einer für die betroffene Person nachvollziehbaren Weise“ erfolgen muss. Im zugehörigen Erwägungsgrund wird dies weiter ausgeführt. Es ist Transparenz darüber herzustellen, wie „betreffende personenbezogene Daten erhoben, verwendet, eingesehen oder anderweitig verarbeitet werden und in welchem Umfang die personenbezogenen Daten verarbeitet werden und künftig noch verarbeitet werden“ (ErwG 39). Diese Erklärung gegenüber dem Betroffenen darf dabei nicht in technischer oder juristischer Fachsprache erfolgen, sondern muss allgemein verständlich und „in einer klaren und einfachen Sprache“ formuliert sein (Art. 12 (1) DSGVO, ErwG 39).

Ziel der Transparenz ist ein Ausgleich des Machtgefälles zwischen dem Verantwortlichen, der über die Datenverarbeitung bestimmt, und dem Betroffenen, dessen Daten verarbeitet werden [Sch17b].

Der Begriff Transparenz wird im Datenschutz gegenteilig zur IT verwendet [Sch17b]. Während in der IT Transparenz zumeist bedeutet, dass ein Nutzer ein System nicht wahrnimmt, meint Transparenz im Datenschutz gerade dem Nutzer gegenüber offenzulegen, welche Verarbeitung wie erfolgt.

Konkretisiert wird dieser Grundsatz insbesondere in den Artikeln 12-14 DSGVO mit den der Verarbeitung vorausgehenden Informationspflichten gegenüber dem Betroffenen, sowie Art. 15 DSGVO mit dem Auskunftsrecht des Betroffenen. Die Informationspflichten gehen erheblich weiter als die alten gesetzlichen Regelungen. Sie stehen insbesondere wegen der Schwierigkeit komplexe Datenverarbeitungen allgemein verständlich zu erklären in der Kritik. Es besteht die Gefahr, dass es zu einem Zuviel an Informationen beim Betroffenen kommt, die dieser nicht mehr nachvollziehen kann und die Informationen daher nur abnickt. Erfahrungen aus dem Verbraucherschutz zeigen, dass eine überbordende Informationspflicht sogar zu Verwirrung beim Betroffenen und damit letztlich zu Intransparenz führen kann [GSSV18].

Problem Eine einfache Erklärung, was alles mit den Daten geschieht und wie sie verarbeitet werden, ist zwar bei exemplarischen Beispielen gut machbar, sobald es jedoch um die Verarbeitung in komplexen Prozessen oder gar um eine Verwendung der Daten im Big Data Kontext geht, stellt diese Anforderung den Verantwortlichen vor große Schwierigkeiten. Auch in diesem Fall muss es für den Betroffenen nachvollziehbar sein, was mit seinen Daten geschieht.

Insbesondere im Hinblick auf explorative Analysen, wie sie im Rahmen der Verarbeitung von Big Data im Data Lake explizit vorgesehen sind, stellt sich die Frage, wie die Anforderungen der Transparenz, insbesondere die vorherige Information des Betroffenen über die Verarbeitung, erfüllt werden kann. Der Verantwortliche muss den Betroffenen bereits bei der Erhebung mitteilen, was er mit den Daten tun wird (Art. 13 DSGVO).

Werden bei der explorativen Analyse zudem noch Methoden des Machine Learnings genutzt, so wird das Problem der Herstellung von Transparenz noch evidenter. Selbst die Entwickler von Machine Learning Systemen können nicht mehr erklären, wie ein System zu seinen Entscheidungen kommt [Cas16]. Daher kann auch dem Betroffenen nicht erklärt werden, was mit seinen Daten geschieht.

Sofern die Daten aus Drittquellen stammen und nicht direkt beim Betroffenen erhoben wurden, so muss die Information des Betroffenen innerhalb einer angemessenen Frist unter Berücksichtigung der Art der Verarbeitung erfolgen (Art. 14 DSGVO). Vereinfacht bedeutet dies: je intensiver der Eingriff in die Rechte des Betroffenen durch die Verarbeitung ist, desto früher muss er informiert werden. In jedem Fall muss die Information spätestens einen Monat nach Erhebung der Daten erfolgen. Zusätzlich gibt es Sonderregelungen zur Offenlegung der Daten an Dritte oder falls sie zur Kommunikation mit dem Betroffenen genutzt werden, vergleiche dazu Art. 14 (3) lit. b, c DSGVO.

Anwendungsszenario Werden die über den Kunden erhobenen Daten, wie z.B. die Bestellhistorie, betrachtete Produkte oder die Suchhistorie, zur Profilbildung genutzt, so muss der Kunde darüber aufgeklärt werden, sodass er nachvollziehen kann, was mit den Daten geschieht und wofür sie verwendet werden. Anderenfalls würde es sich um eine Verletzung des Transparenzgrundsatzes handeln.

Lösungsansatz Im Sinne der Transparenz ist eine beliebige Verwendung der Daten für weitere Analysen nur in soweit möglich, wie die Verarbeitung dem Betroffenen so erklärt werden kann, dass er die Verarbeitung nachvollziehen kann und der Verantwortliche so in der Lage ist, seine sich aus der Transparenz ergebenden Informationspflichten zu erfüllen.

Gleichzeitig stellt der Data Lake für bestimmte Aspekte der Herstellung von Transparenz ein hilfreiches Werkzeug dar. Durch die in der Data Lake Governance dokumentierten Informationen wie z.B. die Datenprovenienz kann die Erfüllung der Informationspflichten unterstützt werden [Sch17b].

5.1.4 Zweckbindung

Die Zweckbindung bei der Verarbeitung personenbezogener Daten ist ein zentrales Prinzip des europäischen Datenschutzes. Dies bedeutet, dass Daten nur für den Zweck, für den sie erhoben wurden, verarbeitet werden dürfen.

Die Zweckbindung ist zudem ein grundlegender Baustein für die Grundsätze der Fairness und der Transparenz und hilft dem Betroffenen, die Tragweite und Auswirkungen der Verarbeitung seiner personenbezogenen Daten zu verstehen.

Es gibt zwischen Verarbeitungen und den dafür erhobenen personenbezogenen Daten und dem Zweck keine 1:1-Beziehung, sondern eine Verarbeitung darf zu verschiedenen Zwecken erfolgen. Allerdings sollte jeder Verarbeitungszweck nur auf eine Rechtsgrundlage gestützt werden [Boc18].

Weiterverarbeitung zu anderen Zwecken

In der DSGVO sind Ausnahmen von diesem Grundsatz normiert. Die Weiterverarbeitung von personenbezogenen Daten ist erlaubt, sofern es eine Rechtsvorschrift gibt, die die Weiterverarbeitung erfordert (z.B. im Onlinehandel: handelsrechtliche Aufbewahrungspflichten) oder sofern der Betroffene der Weiterverarbeitung explizit mittels einer Einwilligung zugestimmt (Art. 6 (4) DSGVO, ErwG 50). Dies gilt allerdings nur solange, wie die Gesetze den in Art. 23 (1) DSGVO genannten Zielen, wie z.B. der nationalen Sicherheit oder der Verhütung und Aufklärung von Straftaten, dienen. Da diese Ziele jedoch sehr weitgehend sind und die meisten Gesetze umfassen, wird diese Einschränkung im Rahmen dieser Arbeit nicht weiter beachtet. Bei einer Weiterverarbeitung nach dieser Ausnahme ist es unerheblich, in welchem Verhältnis der alte und der neue Zweck der Verarbeitung stehen.

Sofern weder eine Rechtsvorschrift noch eine Einwilligung zur Weiterverarbeitung vorliegt, so darf eine Weiterverarbeitung nur zu einem mit dem ursprünglichen Zweck kompatiblen Zweck erfolgen. Zur Prüfung, ob Zwecke miteinander kompatibel sind, nennt das Gesetz in Art. 6 (4) DSGVO zu überprüfende Maßstäbe in einer nicht abschließenden Aufzählung, nach denen die Vereinbarkeit des ursprünglichen und des neuen Zwecks zu überprüfen ist. Zusammengefasst geht es um „den Ausschluss ungerechtfertigt diskriminierender, überraschender und benachteiligender sowie nicht unerheblich in das Recht auf Privatleben eingreifender Datenverarbeitungen“ [GSSV18].

Zusätzlich hat der Gesetzgeber definiert, dass die „Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke“ (Art. 5 (1) lit. b DSGVO) grundsätzlich als kompatibel zum Ursprungszweck gelten.

Wie stark die Zweckbindung tatsächlich ist, hängt sehr davon ab, wie die Begriffe wissenschaftliche und statistische Zwecke ausgelegt werden. Sofern der Begriff „Zweck“ hier im Sinne des Begriffs „Ziel“ gemeint ist, so ist die Ausnahme relativ eng zu verstehen, kollidiert jedoch möglicherweise mit der national (Art. 5 (3) GG) wie europäisch (Art. 13 GRCh) grundrechtlich garantierten Wissenschaftsfreiheit [GSSV18]. Legt man den Begriff „Zweck“ dagegen im Sinne des Begriffs „Methode“ oder „Verfahren“ aus, so ist diese Ausnahme sehr weit gefasst, da damit jegliche Verarbeitung, die statistische Methoden nutzt, aus der Zweckbindung herausfiele. Dies würde für alle Profiling- oder Scoringverfahren gelten, ebenso wie für die meisten Big-Data-Analysen, da diese zumeist auf statistischen Methoden beruhen [GSSV18].

Aus den Erwägungsgründen der DSGVO ergibt sich die Idee des Gesetzgebers hinter der Formulierung „statistische Zwecke“. Der Gesetzgeber legt zunächst eine weite Auslegung des Begriffs zugrunde. Jede Verarbeitung und damit auch Erhebung personenbezogener Daten, die für statistische Untersuchungen und Erstellung statistischer Ergebnisse notwendig sind, sind zunächst durch diesen Begriff gedeckt. Allerdings gilt dies nur für die Fälle, in denen die Ergebnisse einer solchen statistischen Untersuchung keine personenbezogene Daten, sondern aggregierte, von den einzelnen Personen unabhängige Daten sind. Zudem dürfen diese Ergebnisse „nicht für Maßnahmen oder Entscheidungen gegenüber einzelnen natürlichen Personen verwendet werden“ (ErwG 162). Eine personalisierte Analyse oder Profilbildung ist durch die statistischen Zwecke nicht erfasst.

Soll eine Weiterverarbeitung der Daten zu einem anderen Zweck erfolgen, so muss wiederum der Grundsatz der Transparenz beachtet und die sich daraus ergebenden Informationspflichten erfüllt werden (Art. 13 (3) DSGVO, bzw. Art. 14 (4) DSGVO).

Problem Jede Erhebung von personenbezogenen Daten muss zu einem oder mehreren Zwecken erfolgen. Die Verarbeitung über diese Zwecke hinaus ist nur im Rahmen erlaubter Zweckänderungen möglich.

Betrachtet man den Data Lake zunächst als Speichersystem, in dem Daten aus verschiedenen Quellen und Verarbeitungen gespeichert werden, so ist die technische Unterstützung, Daten nur zu den erlaubten Zwecken zu verarbeiten, interessant.

Die andere Perspektive wirft einen Blick auf die Analyse von Daten im Data Lake. Der Grundsatz der Zweckbindung steht in fundamentalem Widerspruch zu der Idee hinter dem Data Lake, dass alle verfügbaren Daten zusammengeführt werden und zweckungebunden verarbeitet sowie analysiert werden können.

Eine weitere Herausforderung stellt die Zweckänderung dar. Wenn der Zweck, beispielsweise aufgrund einer gesetzlichen Aufbewahrungspflicht, geändert wird, so ändern sich auch die „Regeln“, was mit den Daten geschehen darf.

Anwendungsszenario Die Daten, die ein Kunde im Rahmen einer Bestellung angibt, werden zunächst zur Durchführung des Vertrages verarbeitet, die Bestellung wird zusammengestellt, adressiert und zur Zustellung an einen Dienstleister übergeben. Die Daten werden danach jedoch nicht gelöscht, sondern müssen aus handelsrechtlichen Gründen aufbewahrt werden. Diese Daten dürfen dann allerdings nur noch für die gesetzlich vorgesehenen Zwecke verwendet werden. Dies muss durch technische und organisatorische Maßnahmen sichergestellt werden.

Ein Beispiel für die Reichweite der Zweckänderung für statistische Zwecke findet sich bei der Empfehlungsfunktion für weitere Käufe im Onlineshop. Aggregiert der Händler das Kaufverhalten aller Kunden, um zu ermitteln, welche Produkte häufig gemeinsam gekauft werden und nutzt dies, um dem Kunden Vorschläge für weitere Einkäufe zu machen, so ist dies eine legitime Zweckänderung. Es geht nicht um das Kaufverhalten des Einzelnen, sondern es werden lediglich häufig zusammen gekaufte Produkte vorgeschlagen. Wird jedoch das Kauf- und Suchverhalten des einzelnen Kunden analysiert, um ein möglichst für ihn passendes Produkt vorzuschlagen, so ist diese personalisierte Analyse nicht mehr von den statistischen Zwecken gedeckt.

Lösungsansatz Das Governance-System des Data Lakes sollte den Verantwortlichen darin unterstützen, die Zweckbindung einzuhalten. Es sollte beispielsweise prüfen, ob eine Verarbeitung von Daten dem erlaubten Zweck entspricht. In einer einfachen Form kann dies mittels Metadaten geschehen, in komplexeren und größeren Systemen sollte dafür ein Permission Management System genutzt werden, das die vorliegende Nutzungserlaubnis von Daten gegen den Anwendungsfall prüft.

Bereits für andere Zwecke erhobene Daten können durch die Ausnahmen von der Zweckbindung auch in engem Rahmen für andere Zwecke verwendet werden. Durch sie sind zumindest aggregierte statistische Analysen möglich. Wie weit die Ausnahmen tatsächlich greifen und wie weit man damit Zweckänderungen begründen kann, wird erst die Praxis durch Stellungnahmen und Bescheide der Aufsichtsbehörden sowie entsprechende Gerichtsurteile zeigen.

Dennoch lässt sich auch bei weiter Auslegung der unklaren Begriffe keine Rechtfertigung finden, wie personenbezogene Daten komplett ergebnisoffen und zweckungebunden analysiert und genutzt werden können. Dies würde den Grundsatz der Zweckbindung komplett aushebeln und würde ebenso schon beispielsweise an den Grundsätzen der Fairness und Transparenz scheitern.

5.1.5 Datenminimierung

Datenminimierung bezieht sich nicht auf die Löschung von Daten, sondern heißt, dass die Verarbeitung der personenbezogenen Daten für den Zweck angemessen und notwendig sein muss. Zudem muss sie auf das erforderliche Maß beschränkt sein. Das Ziel ist, so wenig personenbezogene Daten wie möglich zu verarbeiten [GSSV18]. Eine treffendere Bezeichnung ist die aus dem alten BDSG bekannte Formulierung der „Datenvermeidung und Datensparsamkeit“.

Der Begriff der Notwendigkeit bestimmt, wie weit dieser Grundsatz greift [GSSV18]. Würde an die objektive Bedeutung der Notwendigkeit für die Verarbeitung angeknüpft werden, so wären moderne Entwicklungen wie Smartphones oder soziale Netzwerke nicht möglich gewesen, da diese in hohem Maße auf Datensammlung und Datenauswertung (z.B. Auswertung von Kontakten und Interessen für den Timeline-Algorithmus) beruhen, die mit einem strikten Verständnis der Notwendigkeit

kaum zu vereinbaren wären. Wird die Notwendigkeit jedoch weiter verstanden, so spielen auch gegenläufige Interessen wie der freie Informationsfluss (Art. 1 (3) DSGVO), die Grundrechte des Verarbeiters (insbes. freie Berufsausübung) und Gemeinwohlinteressen eine Rolle [GSSV18].

Problem Auch bei einer weiten Auslegung des Begriffs der Notwendigkeit steht dieser Grundsatz im Widerspruch zur Idee der Data-Lake-Verarbeitungen, nach welcher Daten im Zweifel zunächst gesammelt werden und im weiteren Verlauf versucht wird, Wertschöpfung aus den Daten zu betreiben. Überspitzt formuliert folgt Big Data, wozu auch die Data Lake Analysen gehören, dem Grundsatz der Datenmaximierung. Je mehr Daten zur Verfügung stehen, desto bessere und desto präzisere Ergebnisse sind möglich [BDW15].

Anwendungsszenario Ein Beispiel für Datenminimierung ist die Möglichkeit der anonymen Bestellung. Dadurch wird kein Kundenkonto mit Bestellhistorie oder ähnliches angelegt, sondern es wird nur die Bestellung bearbeitet.

Lösungsansatz Personenbezogene Daten dürfen nicht anlasslos gesammelt und verarbeitet werden. Sie dürfen nur im notwendigen Umfang und gleichzeitig nur so kurz wie möglich gespeichert werden. In dieser Hinsicht sind die Möglichkeiten zur Analyse personenbezogener Daten schlicht begrenzt, da nicht so viele Daten vorliegen dürfen, wie vielleicht möglich oder für die Analyse wünschenswert wäre. Dies schränkt die Data-Lake-Verarbeitungen ein. Allerdings kann die Einhaltung dieses Grundsatzes weder technisch unterstützt noch sichergestellt werden, dies muss durch den Verantwortlichen geschehen.

5.1.6 Datenrichtigkeit

Nach dem Grundsatz der Datenrichtigkeit hat der Betroffene das Recht, dass die ihn betreffenden personenbezogenen Daten korrekt und „auf dem neuesten Stand“ sind. Der Verantwortliche hat für die Datenrichtigkeit zu sorgen, unrichtige Daten müssen vom Verantwortlichen unverzüglich gelöscht oder korrigiert werden. Konkretisiert wird dieser Grundsatz in dem Recht auf Berichtigung (Abschnitt 5.3.2).

Die Datenrichtigkeit ist nicht objektiv als einfache, binäre Frage zu verstehen, ob die gespeicherten Informationen stimmen oder nicht, sondern muss im Kontext des Verarbeitungszwecks gesehen werden (Art. 5 (1) lit. d DSGVO). So können Daten, die historisch korrekt waren, aber nicht mehr aktuell sind, richtig sein, wenn der Zweck der Verarbeitung erfordert, dass die nicht mehr aktuelle Information erhalten bleibt (sog. Historisierung von Daten z.B. in einer Krankenakte). Gleichzeitig dehnt sich der Anwendungsbereich auch auf Kontextverfälschungen aus, das heißt, wenn - an sich richtige - Informationen zu einem falschen Gesamtbild zusammengesetzt werden [GSSV18].

Problem Der Verantwortliche muss dafür Sorge tragen, dass personenbezogene Daten korrekt und aktuell sind.

Die geforderte Veränderbarkeit und Löschbarkeit von Daten widerspricht der dauerhaften Speicherung der Daten in ihrer originalen Fassung in den persistenten Zonen des Data Lakes (vgl. Abschnitt 3.1.2).

Wie in Abschnitt 3.1.2 diskutiert, werden Dateien in HDFS unveränderlich gespeichert. Entsprechend kollidiert die gängigste Implementierung des Speichersystems des Data Lakes mit der Anforderung, dass Daten löschbar und veränderbar sein müssen.

Die Überprüfung, ob Daten im Kontext des Verarbeitungszweckes korrekt sind, hängt von der Verarbeitung und den betroffenen Daten ab.

Anwendungsszenario Eine mögliche Umsetzung der Pflicht für Datenrichtigkeit zu sorgen, wäre eine jährliche Bitte an den Kunden, die hinterlegten Kundendaten auf Korrektheit zu überprüfen.

Lösungsansatz Die Pflicht für Datenrichtigkeit zu sorgen, begründet keine tiefergehende Nachforschungspflicht durch den Verantwortlichen [GSSV18], eine Lösung wie im Anwendungsszenario dürfte ausreichend sein.

Die Anforderung, dass eine Veränderung und Löschung von Daten möglich sein muss, erfordert eine Modifikation des Zonenmodells für Data Lakes, in denen personenbezogene Daten verarbeitet werden. Zonen, in denen solche Daten gespeichert werden sollen, müssen eine Veränderung und Löschung der Rohdaten zulassen.

Da weder eine Veränderung der Daten ohne Beibehaltung der Originaldaten, noch eine echte Löschung im HDFS möglich ist, scheint die Nutzung als Grundlage des Speichersystems für Data Lakes mit personenbezogenen Daten keine Lösung zu sein. Stattdessen muss auf alternative Technologien zurückgegriffen werden.

Bei der Überprüfung, ob Daten im Kontext des Verarbeitungszweckes richtig sind, muss zwischen Daten unterschieden werden, deren Änderung problemlos möglich ist (z.B. Anpassung der Kundenadresse), die keine besondere Überprüfung erfordern und Daten, deren Richtigkeit vom Kontext abhängig ist. Für den ersten Fall kann ein direktes Interface zur Änderung bestimmter Daten implementiert werden. Im anderen Fall ist eine spezifische Einzelfallprüfung erforderlich, die vom Data Lake nur unterstützt werden kann, wenn z.B. in den Metadaten Informationen über den Zweck der Verarbeitung zur Verfügung stehen.

5.1.7 Speicherbegrenzung

Der Grundsatz der Speicherbegrenzung besagt, dass personenbezogene Daten nur solange gespeichert werden dürfen, wie es für die Verarbeitungen, für die sie gespeichert wurden, erforderlich ist. Konkretisiert wird dieser Grundsatz im Betroffenenrecht auf Löschung (Abschnitt 5.3.3).

Eine Ausnahme hiervon gilt nur, wie bei der Zweckbindung, für „im öffentlichen Interesse liegende Archivzwecke oder für wissenschaftliche und historische Forschungszwecke oder für statistische Zwecke“. Voraussetzung für diese Ausnahme ist allerdings, dass der Verantwortliche technische und organisatorische Maßnahmen trifft, um sicherzustellen, dass diese Daten ausschließlich für diese Zwecke verwendet werden. Dies stellt allerdings keine Erlaubnis zur dauerhaften Speicherung dar.

Es gilt nach wie vor der Grundsatz der Datenminimierung, so dass nur die Daten, die tatsächlich benötigt werden, aufbewahrt werden dürfen und jene auch nur so lange gespeichert werden dürfen, wie dies für die Zwecke der Verarbeitung erforderlich ist.

Problem Der Verantwortliche muss dafür Sorge tragen, dass er personenbezogene Daten nur solange und in dem Umfang speichert, wie es für die Verarbeitung erforderlich ist.

Die bereits in Abschnitt 5.1.6 diskutierten Problematiken (unveränderliche Speicherung, Technologiewahl) entstehen ebenso beim Grundsatz der Speicherbegrenzung.

Anwendungsszenario Bei einem Kunden, der sich nur für den Newsletter angemeldet hat, kann nicht von Anfang an gesagt werden, wie lange die E-Mail-Adresse gespeichert und verarbeitet werden darf. Meldet der Kunde sich nach einem Tag vom Newsletter ab, so entfällt ab diesem Zeitpunkt der Zweck für die Verarbeitung, meldet er sich dagegen erst nach 20 Jahren ab, so darf die Adresse natürlich so lange verarbeitet werden. Da besagte E-Mail-Adresse allerdings nur für den Newsletterversand verarbeitet werden darf, kann ein Löschkriterium festgelegt werden. Dieses regelt, dass die E-Mail-Adresse bei Abbestellung des Newsletters gelöscht werden muss.

Lösungsansatz Um diesen Grundsatz umsetzen zu können, muss zu jedem Datum in Abhängigkeit vom Zweck der Verarbeitung eine Löschfrist festgelegt werden, die bestimmt, wie lange es gespeichert werden darf. Wann das Datum tatsächlich gelöscht werden darf, wird somit durch die längste Löschfrist, die für das Datum gilt, bestimmt. Soweit dies nicht möglich ist, so muss ein Kriterium für die Dauer der Speicherung festgelegt werden (vgl. Anwendungsszenario). Um die Umsetzung der Löschung sicherstellen zu können, bietet es sich an, dass diese Daten als Metadaten zu den personenbezogenen Daten gespeichert werden und durch geeignete Governance-Prozesse die tatsächliche Löschung durchgeführt und überprüft wird.

Um die Löschfristen definieren zu können und der Rechenschaftspflicht (vgl. Abschnitt 5.7) zur Einhaltung dieses Grundsatzes Genüge zu tun, sollte der Verantwortliche ein Löschkonzept zur Löschung personenbezogener Daten entwickeln, in dem festgelegt wird, welche Daten zu welchen Zwecken wie lange aufbewahrt werden dürfen. Ebenso sollten die Governance-Prozesse, die zur Umsetzung des Grundsatzes implementiert sind, dokumentiert werden.

Bezüglich der bereits beim Grundsatz der Datenrichtigkeit (Abschnitt 5.1.6) diskutierten Problematiken sind die gleichen Lösungsansätze auch für den Grundsatz der Speicherbegrenzung anwendbar (Modifikation des Zonenmodells, geeignete Technologiewahl).

5.1.8 Integrität und Vertraulichkeit

Der Grundsatz der Integrität und Vertraulichkeit dient der Umsetzung der Datensicherheit. Exemplarisch wird in der DSGVO der Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust, Zerstörung und Schädigung der Daten genannt.

Weshalb der Grundsatz nur zwei der drei klassischen Ziele der IT-Sicherheit nennt und die Verfügbarkeit außen vor lässt, erschließt sich nicht. Im den Grundsatz konkretisierenden Artikel über die Sicherheit der Verarbeitung werden alle drei Ziele genannt. Die detaillierte Diskussion, einschließlich der Probleme und Lösungsansätze, erfolgt daher mit allen Zielen der IT-Sicherheit in Abschnitt 5.5. Die dort diskutierten Probleme und Lösungsansätze gelten ebenso für diesen Grundsatz.

5.2 Rechtsgrundlagen

Die Rechtsgrundlagen stellen die Ausgestaltung des Grundsatzes der Rechtmäßigkeit dar. Das Gesetz kennt, je nach Zählung, insgesamt sechs bzw. sieben Rechtsgrundlagen, auf deren Basis die Verarbeitung personenbezogener Daten möglich ist (Art. 6 (1) DSGVO). In den meisten Fällen werden im Unternehmen Verarbeitungen auf Basis der *Einwilligung*, zur Anbahnung oder Durchführung eines *Vertrages*, auf Basis *gesetzlicher Grundlage* und aufgrund *berechtigten Interesses* durchgeführt. Die weiteren Rechtsgrundlagen sind der Schutz *lebenswichtiger Interessen* und Wahrnehmung einer Aufgabe im *öffentlichen Interesse* sowie in *Ausübung öffentlicher Gewalt*, die dem Verantwortlichen übertragen wurde. Im Kontext der Data-Lake-Verarbeitungen sind insbesondere die Einwilligung und die Verarbeitung aufgrund berechtigten Interesses relevant.

Der Verantwortliche muss für die Data-Lake-Verarbeitungen eine geeignete Rechtsgrundlage auswählen. Die Wahl der Rechtsgrundlage kann vom Data Lake selbst nicht unterstützt werden, es handelt sich um eine rechtliche Abwägung, die der Verantwortliche durchführen muss. Sie ist Voraussetzung für die rechtskonforme Verarbeitung personenbezogener Daten. Allerdings kann der Data Lake den Verantwortlichen in der Einhaltung der Anforderungen an die jeweilige Rechtsgrundlage unterstützen.

Die speziell Kinder betreffenden Regelungen bezüglich der Rechtsgrundlagen zur Verarbeitung personenbezogener Daten (Art. 6 (1) lit. f und Art. 8 DSGVO) werden in dieser Arbeit nicht behandelt.

5.2.1 Einwilligung

Die Einwilligung ermöglicht als Rechtsgrundlage die größtmögliche Freiheiten für den Verantwortlichen. Sofern der Betroffene zustimmt, können die Daten für die in der Einwilligung genannten Zwecke genutzt werden.

Allerdings ist eine gültige und damit wirksame Einwilligung an verschiedene Bedingungen geknüpft (Art. 7 DSGVO i.V.m. Art. 4 Nr. 11 DSGVO). Sie muss spezifisch für einen oder mehrere Zwecke erteilt werden, das heißt eine generelle Einwilligung zur Nutzung der Daten für beispielsweise Marketing- oder Analysezwecke ist nicht möglich.

Die Einwilligung muss informiert und freiwillig erfolgen. Informiertheit bedeutet, dass der Betroffene weiß, worauf er sich einlässt, wofür welche Daten in welchem Umfang genutzt werden und ist eine Ausgestaltung der Transparenz.

Freiwilligkeit bedeutet, dass kein Zwang, auch kein faktischer, für den Betroffenen bestehen darf, die Einwilligung zu erteilen. Diese Gefahr besteht insbesondere in Abhängigkeitsverhältnissen, wie z.B. in einem Arbeitsverhältnis. Ein weiterer Aspekt der Freiwilligkeit ist mit dem sogenannten Kopplungsverbot verbunden, wonach die Erfüllung eines Vertrages nicht von der Einwilligung zur Verarbeitung personenbezogener Daten, die zur Vertragserfüllung nicht benötigt werden, abhängig gemacht werden soll. Während es im Erwägungsgrund als Verbot formuliert ist (ErwG 43), ist die Formulierung im Gesetzestext kein Verbot, sondern nur, dass dieser Kopplung bei der Bewertung der Freiwilligkeit in „größtmöglichem Umfang Rechnung getragen werden“ muss (Art. 7 (4) DSGVO). Die Reichweite dieser Regelung wird kontrovers diskutiert [Bay17; GSSV18; Här16; Sch17a].

Zudem muss der Verantwortliche nachweisen können, dass der Betroffene in die konkrete Datenverarbeitung eingewilligt hat.

Der Betroffene kann seine Einwilligung jederzeit widerrufen, das heißt er kann zu jedem beliebigen Zeitpunkt entscheiden, dass seine Daten zukünftig nicht mehr für den Zweck, für den eingewilligt wurde, verarbeitet werden dürfen. Eine Kopplung anderer Leistungen des Verantwortlichen an die Einwilligung ist nur bedingt möglich (vgl. Diskussion um sog. Kopplungsverbot oben).

Grundsätzlich kann der Verantwortliche eine Verarbeitung auf mehrere Rechtsgrundlagen stützen. Legt der Verantwortliche sich gegenüber dem Betroffenen allerdings auf die Rechtsgrundlage der Einwilligung fest, so kann er diese Verarbeitung nicht gleichzeitig auf eine andere Rechtsgrundlage stützen [AE18]. Dies würde das Konzept der Einwilligung unterlaufen und wäre dem Betroffenen gegenüber rechtsmissbräuchlich. Der Betroffene muss sich darauf verlassen können, dass eine Verarbeitung beendet wird, wenn er seine Einwilligung dazu widerruft.

Problem Der großen Freiheit bei der Wahl des Verarbeitungszweckes, für den der Verantwortliche die Einwilligung einholt, stehen bei der praktischen Nutzung viele Herausforderungen wie die Informiertheit, die zu erfüllende Nachweispflicht und die Möglichkeit des Widerrufs gegenüber.

Anwendungsszenario E-Mail-Marketing darf bei Bestandskunden - zumindest im engen Umfang - auch ohne Einwilligung auf Basis des berechtigten Interesses (Details im Abschnitt 5.2.2) durchgeführt werden. Bittet man den Kunden jedoch um Einwilligung für das E-Mail-Marketing und der Kunde widerruft diese Einwilligung später, so ist es nicht mehr möglich, sich auf das berechtigte Interesse zu stützen.

Lösungsansatz Es stellt sich, wie bereits beim Grundsatz der Transparenz die Frage, wie weit man dem Betroffenen die Datenverarbeitungen im Rahmen des Data Lake erklären kann, sodass es sich um eine informierte Einwilligung handelt (vgl. dazu auch Abschnitt 5.1.3).

Um der Nachweispflicht nachkommen zu können, ist es ratsam, die Einwilligung des Nutzers zu protokollieren - sowohl den Zeitpunkt und die Art der Einwilligung als auch den spezifischen Einwilligungstext. In einer einfachen Form kann dies mittels Metadaten geschehen, in komplexeren und größeren Systemen sollte dies im Permission-Management-System erfolgen.

Der jederzeit mögliche Widerruf stellt eine Einschränkung dar, inwieweit die Daten verarbeitet werden dürfen. Die Umsetzung muss durch Governance-Prozesse sichergestellt werden.

5.2.2 Berechtigtes Interesse

Die Verarbeitung von personenbezogenen Daten aufgrund berechtigten Interesses ermöglicht dem Verantwortlichen die Verarbeitung personenbezogener Daten ohne Zustimmung des Betroffenen. Kern dieser Rechtsgrundlage ist eine Interessenabwägung, in der der Verantwortliche sein Interesse oder das Interesse eines Dritten an der Datenverarbeitung mit den zu schützenden Interessen des Betroffenen abgewogen werden müssen.

Die Prüfung, ob diese Rechtsgrundlage herangezogen werden kann, kann als vierstufiges Prüfschema dargestellt werden (in Anlehnung an das dreistufige Schema von Assion/Nolte/Veil [GSSV18]):

1. **Definition der Interessen**
   Zunächst müssen die Interessen des Verantwortlichen oder eines Dritten (d.h. nicht der Verantwortliche und nicht der Betroffene) definiert werden. Das kann im Prinzip jeder Grund sein, weshalb personenbezogene Daten verarbeitet werden können. Ziel ist nur zu definieren, weshalb die Daten verarbeitet werden sollen.

2. **Berechtigung der Interessen**
   In der DSGVO wird nicht vorgegeben, was mit berechtigtem Interesse gemeint ist. Stattdessen werden in den Erwägungsgründen Beispiele aufgeführt, die berechtige Interessen darstellen können, z.B. Direktwerbung oder Sicherstellung der IT-Sicherheit (ErwG 47-50). Die Prüfung der Berechtigung ist nicht auf den Datenschutz bezogen, sondern meint eher die Legitimität des Interesses, in der Literatur wird dies als „jedes von der Rechtsordnung anerkannte Interesse“ [Pla16] bezeichnet. Ein Interesse, das gegen geltendes Recht verstößt, kann folglich nicht berechtigt sein [GSSV18].

3. **Erforderlichkeit der Verarbeitung zur Interessenwahrung**
   Die Verarbeitung muss zur Interessenwahrung erforderlich sein, das heißt sie muss nicht nur geeignet sein, den Zweck zu erreichen (zweckmäßig), sondern sie muss das mildeste, das heißt die Interessen des Betroffenen am wenigsten beeinträchtigende, geeignete Mittel sein, um das Interesse zu wahren [GSSV18].

4. **Interessenabwägung**
   Zunächst müssen den oben bestimmten Interessen die entgegenstehenden Interessen des Betroffenen gegenüber gestellt und miteinander abgewogen werden. Ein wesentlicher Teil der Interessenabwägung ist die Abwägung zwischen dem Verwendungsinteresse einerseits und dem Privatheitsinteresse andererseits [GSSV18]. Ein weiterer Maßstab für die Interessenabwägung ist, ob ein Betroffener mit seinem Wissen beim Zeitpunkt der Erhebung „vernünftigerweise“ damit rechnen konnte, dass die Daten für diesen Zweck verarbeitet werden (ErwG 47 S. 3).

Wenn eine Verarbeitung auf berechtigte Interessen gestützt wird, so muss der Betroffene über die berechtigten Interessen informiert werden (Art. 13 (1) lit. d, bzw. Art. 14 (2) lit. b DSGVO). Der Betroffene hat gegen eine Verarbeitung auf Grundlage berechtigten Interesses ein Widerspruchsrecht (detailliert in Abschnitt 5.3.7).

Exkurs E-Mail-Marketing und DSGVO

Auch wenn in den Erwägungsgründen Direktmarketing als berechtigtes Interesse anerkannt wird (ErwG 47), dürfen E-Mail-Adressen aufgrund einer Regelung im Gesetz gegen den unlauteren Wettbewerb (UWG) nur bei Bestandskunden in begrenztem Umfang (Details in §7 (3) UWG) genutzt werden.

Dieses Thema berührt komplexe Zusammenhänge und Geltungshierarchien zwischen Europarecht und nationalem Recht. Vereinfacht dargestellt: Grundsätzlich geht Europarecht nationalem Recht vor und bricht dieses, falls widersprüchliche Regelungen existieren. Allerdings handelt es sich bei der Regelung im UWG um eine Umsetzung einer Regelung aus der ePrivacy-Richtlinie der EU (Art. 13 RL 2002/58/EG). In der DSGVO ist geregelt, dass die DSGVO gegenüber den Umsetzungsregelungen der ePrivacy-Richtlinie zurücktritt (Art. 95 DSGVO). Die Regelung aus dem deutschen UWG gilt daher in diesem Bereich fort.

Problem Die Rechtsgrundlage des berechtigten Interesses erfordert vom Verantwortlichen eine eigenverantwortliche Abwägung der Interessen. Das Ergebnis dieser Abwägung stellt das Fundament der Verarbeitung dar.

Besonders im Bereich der massenhaften Verarbeitung personenbezogener Daten können Persönlichkeitsprofile entstehen, die für den Zweck der Verarbeitung nicht erforderlich sind, aber als Nebenprodukt der eigentlichen Verarbeitung anfallen.

Anwendungsszenario Eine typische Verarbeitung von personenbezogenen Daten auf Grundlage berechtigten Interesses ist E-Mail-Werbung bei Bestandskunden. Bei der Interessenabwägung stehen sich - vereinfacht dargestellt - primär das Interesse des Händlers an Werbung, also letztlich Umsatzsteigerung, und das Interesse des Kunden an Privatheit gegenüber. Da aber bereits eine Geschäftsbeziehung zwischen Kunden und Händler besteht, kann davon ausgegangen werden, dass der Kunde grundsätzlich Interesse an den Angeboten des Händlers hat. Zudem ist es für den Betroffenen auch erwartbar, dass er vom Händler über dessen Angebote informiert wird.

Lösungsansatz Auch wenn die rechtliche Würdigung vom Verantwortlichen manuell durchgeführt werden muss, so bietet die Verarbeitung auf Grundlage des berechtigten Interesses ein mächtiges Werkzeug, da weder das Einverständnis des Betroffenen erforderlich ist, noch der Betroffene durch Widerruf seiner Einwilligung die Verarbeitung beenden kann. Dies ist nur im Fall des Direktmarketings oder in den Härtefallregelungen des Rechts auf Widerspruch möglich.

Oft ist es auch möglich, Verarbeitungen zu entschärfen, indem Daten z.B. pseudonymisiert verarbeitet werden, wodurch die Interessen des Betroffenen nicht so stark tangiert werden, weshalb eine Verarbeitung auf Grundlage des berechtigten Interesses ermöglicht wird. Die technische Umsetzung, dass Daten für bestimmte Verarbeitungen nur pseudonymisiert zur Verfügung gestellt werden, kann vom Data Lake mittels Views, also einer entsprechend modifizierten Ansicht der Daten, bereitgestellt werden. Die Konstruktion der Entschärfung, welche Maßnahmen dafür nötig sind, muss manuell und jeweils im Einzelfall der Verarbeitung erfolgen.

5.2.3 Besondere Kategorien personenbezogener Daten

Eine besondere Herausforderung stellt die Verarbeitung besonderer Kategorien personenbezogener Daten dar. Diese Daten wurden vom Gesetzgeber als besonders schutzwürdig eingestuft und unterliegen daher strengeren Anforderungen an die Rechtsgrundlagen. Diese Daten dürfen weder aufgrund eines Vertrages, noch auf Basis des berechtigten Interesses verarbeitet werden, weshalb so zumeist nur die Einwilligung als Rechtsgrundlage bleibt (Art. 9 (2) DSGVO).

Problem Die Verarbeitung besonderer Kategorien personenbezogener Daten ist für den Data Lake in zweierlei Hinsicht relevant. Zum einen muss dies beim Speichersystem Data Lake beachtet werden. Der Zugriff auf solche Daten darf nicht gewährt werden, sofern die Verarbeitung auf Basis eines Vertrags oder berechtigtem Interesses erfolgt.

Zum anderen muss dies auch bei den Data-Lake-Verarbeitungen beachtet werden. Sensible Daten, insbesondere Gesundheitsdaten, entstehen oft schon durch die Beobachtung von Verhalten über Zeit. Deutliche Verhaltensänderungen ermöglichen oft den Rückschluss auf gesundheitliche Veränderungen. Bei der Analyse von Kundenverhalten können so besondere Kategorien personenbezogener Daten entstehen. Sämtliche Analysen und Verarbeitungen dieser sensible Daten dürfen nicht auf Basis des berechtigten Interesses oder zur Vertragserfüllung erfolgen. Stattdessen wird immer eine Einwilligung benötigt.

Anwendungsszenario Analysen des Kaufverhaltens, insbesondere im Hygienebereich, erlauben schnell Rückschlüsse auf den Gesundheitszustand. Beispielsweise erlaubt das Ausbleiben der regelmäßigen Bestellung von Monatshygiene-Artikeln in Verbindung mit dem Kauf von Elternratgebern den Schluss, dass eine Schwangerschaft vorliegt.

Lösungsansatz Soweit besondere Kategorien personenbezogener Daten im Data Lake gespeichert werden, so müssen diese entsprechend markiert werden. Durch Governance-Prozesse muss sichergestellt werden, dass beim Zugriff auf diese Daten die eingeschränkten Rechtsgrundlagen beachtet werden.

Die Gefahr der Entstehung von besonderer Kategorien personenbezogener Daten muss bei der Verarbeitung personenbezogener Daten beachtet werden. Insbesondere müssen Data Scientists dafür sensibilisiert werden.

5.3 Betroffenenrechte

Die Betroffenenrechte sollen der der personenbezogenen Datenverarbeitung unterworfenen Person die Möglichkeit geben, die Rechtmäßigkeit einer Verarbeitung zu überprüfen und so helfen, die Machtasymmetrie zwischen Verarbeiter und Betroffenem auszugleichen [Sch17b].

Der Betroffene kann diese Betroffenenrechte unentgeltlich in Anspruch nehmen (Art. 12 (5) S.1 DSGVO). Da die Öffentlichkeit zunehmend für das Thema Datenschutz sensibilisiert wird, muss der Verantwortliche mit vielen Anfragen von Betroffenen rechnen und muss diese unverzüglich, im Regelfall spätestens innerhalb eines Monats beantworten (Art. 12 (3) DSGVO). Die Umsetzung

der Betroffenenrechte sollte daher so organisiert werden, dass dies weitgehend automatisiert oder zumindest technisch unterstützt wird. Der Verantwortliche kann die Inanspruchnahme nur in Fällen exzessiver oder offensichtlich unbegründeter Anträge verweigern oder ein Entgelt verlangen (Art. 12 (5) S.2 DSGVO).

Zudem besteht für den Betroffenen immer die Möglichkeit der Beschwerde bei einer Aufsichtsbehörde (Art. 77 (1) DSGVO) oder einer Klage vor Gericht (Art. 79 (1) DSGVO), wenn der Verantwortliche aus seiner Sicht die DSGVO verletzt oder seinen Betroffenenrechten nicht nachkommt.

Eine Grundvoraussetzung für die Erfüllung aller Betroffenenrechte ist die eindeutige Identifikation des Betroffenen. Gelingt es dem Verantwortlichen nicht, den Betroffenen zu identifizieren, so kann er den Antrag nicht umsetzen. Anderenfalls würde er eine Datenschutzverletzung verursachen, wenn bspw. eine Kopie der personenbezogenen Daten an die falsche Person versandt wird.

Neben den im Folgenden diskutierten Betroffenenrechten existiert noch die Informationspflicht des Verarbeiters gegenüber dem Betroffenen (Art. 13, 14 DSGVO). Diese ist eine Ausgestaltung des Grundsatzes der Transparenz (Abschnitt 5.1.3) und enthält grundlegende Informationen über die Verarbeitung der personenbezogenen Daten (z.B. Kontaktdaten, Verarbeitungszwecke, Speicherdauer, Information über Betroffenenrechte). Der Betroffene muss bei Erhebung der Daten oder, falls die Daten nicht direkt beim Betroffenen erhoben wurden, innerhalb einer angemessenen Frist informiert werden. Solange die Grundsätze der Verarbeitung personenbezogener Daten, insbesondere der der Transparenz und der der Fairness (Abschnitt 5.1.2), beachtet werden, entstehen durch die Informationspflichten beim Design des Data Lakes keine weiteren Herausforderungen.

Ein zweites, bereits diskutiertes Betroffenenrecht ist das *Recht auf Widerruf der Einwilligung* (Abschnitt 5.2.1).

5.3.1 Auskunftsrecht

Das Auskunftsrecht ist neben den Informationspflichten die stärkste Ausprägung des Grundsatzes der Transparenz. Es ist in zwei Teile aufgeteilt. Der Betroffene hat das Recht, ein Auskunftsersuchen zu stellen. Der Verantwortliche ist dann verpflichtet, dem Betroffenen mitzuteilen, ob personenbezogene Daten über ihn verarbeitet werden (Art. 15 (1) und (2) DSGVO). Zudem müssen grundlegende Informationen über die Verarbeitungen (z.B. Zwecke, Speicherfristen, Information über weitere Betroffenenrechte und Beschwerderecht bei der Aufsichtsbehörde) beauskunftet werden.

Sofern personenbezogene Daten verarbeitet werden, hat er zudem zusätzlich auf Antrag den Anspruch, eine Kopie dieser Daten zu erhalten (Art. 15 (3) DSGVO).

Die DSGVO kennt keine Ausnahmen vom Auskunftsrecht, weshalb eine Europarechtswidrigkeit der Norm diskutiert wird [GSSV18]. Solange der EuGH diese jedoch nicht festgestellt hat, ist die Norm anzuwenden. Das Recht auf den Erhalt einer Kopie der personenbezogenen Daten „darf die Rechte und Freiheiten anderer Personen nicht beeinträchtigen“ (Art. 15 (4) DSGVO). Im zugehörigen Erwägungsgrund 63 zur DSGVO werden dabei explizit Geschäftsgeheimnisse und Urheberrechte genannt. Allerdings darf dies nicht zu einer vollständigen Verweigerung führen, weshalb anzunehmen ist, dass dies in der Realität bspw. mit Schwärzungen umzusetzen ist.

Durch eine der Öffnungsklauseln (Art. 23 DSGVO) kann der nationale Gesetzgeber Betroffenenrechte einschränken. Eine der wichtigsten Einschränkungen in Deutschland ist, dass Daten, die ausschließlich aufgrund von Aufbewahrungspflichten oder als Backup gespeichert werden, nicht beauskunftet werden müssen, sofern der Aufwand unverhältnismäßig ist und sichergestellt ist, dass die Daten nur für diesen Zweck verarbeitet werden (§34 (1) Nr. 2 BDSG).

Problem Der Verantwortliche muss in der Lage sein, alle personenbezogene Daten, die er zu einer Person gespeichert hat, zu identifizieren und zu beauskunften.

Sofern der Betroffene vom Recht auf eine Kopie der personenbezogenen Daten Gebrauch macht, so muss der Verantwortliche sicherstellen, dass er durch das Bereitstellen der Kopie keine Rechte und Freiheiten Dritter verletzt.

Anwendungsszenario Wenn eine Person von ihrem Auskunftsrecht Gebrauch macht, so muss der Onlinehändler, nachdem er die Person eindeutig identifiziert hat, prüfen, ob Daten über diese Person vorliegen.

Auch wenn die Daten alle im Data Lake gespeichert sind, so stammen sie aus verschiedenen Systemen in unterschiedlichen Formaten, wie z.B. Stammdaten, Supportanfragen, Bestellsystem oder Newsletterverteiler. Der Verantwortliche muss in der Lage sein, diese Daten zu verbinden und zu beauskunften und, falls gefordert, eine Kopie der Daten zur Verfügung stellen.

Lösungsansatz Dies erfordert vom Verantwortlichen genaue Kenntnisse über den Aufbau des Data Lakes und die Datenquellen.

Eine Möglichkeit zur Zuordnung der personenbezogenen Daten zu den jeweiligen Personen ist die Speicherung einer eindeutigen Kennung (z.B. Kundennummer, interne ID) in den Metadaten, sobald die Daten im Data Lake gespeichert werden.

Die Überprüfung, ob die Kopie der Daten Rechte und Freiheiten anderer beeinträchtigt, kann nur teilweise technisch gelöst werden. Durch die Zuordnung von personenbezogenen Daten zu den jeweiligen Betroffenen können direkt die Daten identifiziert werden, die mehreren Betroffenen zugeordnet sind. Allerdings ist die tatsächliche rechtliche Einschätzung, ob dadurch Rechte und Freiheiten anderer beeinträchtigt werden, nicht technisch lösbar. Sie muss durch den Verantwortlichen erfolgen. Er muss entscheiden, welche Daten aus welchen Quellen in dieser Hinsicht möglicherweise Schwierigkeiten bergen. Diese müssen dann im Einzelfall untersucht und ggf. z.B. geschwärzt werden.

Die tatsächliche Auskunftserteilung kann insbesondere für Daten, die zu nahezu allen Betroffenen einer Kategorie, wie z.B. Kunden, vorliegen, durch entsprechende Prozesse des Data Lakes vereinfacht werden. Durch die Metadaten (z.B. Datenkategorien, Empfänger oder Verwendungszwecke und Rechtsgrundlagen) können die Informationen, die beauskunftet werden müssen, automatisiert gesammelt und in ein entsprechendes Antwortschreiben eingefügt werden.

5.3.2 Recht auf Berichtigung oder Vervollständigung

Neben der aus dem Grundsatz der Datenrichtigkeit folgenden Verpflichtung des Verantwortlichen, für Datenrichtigkeit zu sorgen, hat der Betroffene das Recht, unrichtige personenbezogene Daten korrigieren oder unvollständige Daten ergänzen zu lassen (Art. 16 DSGVO).

Es handelt sich hierbei genau genommen um zwei Betroffenenrechte. Zum einen hat der Betroffene das Recht, dass unrichtige personenbezogene Daten korrigiert werden. Eine Ausnahme stellen Verarbeitungen dar, deren Zweck die Dokumentation historischer Daten ist (z.B. Krankenakte). Auch wenn die Daten darin nicht mehr korrekt sind, so fordert der Zweck der Verarbeitung gerade die Verarbeitung der nicht mehr korrekten Daten [GSSV18] (vgl. auch Grundsatz der Datenrichtigkeit, Abschnitt 5.1.6).

Zum anderen hat der Betroffene das Recht, dass unvollständige Daten ergänzt werden. Dieses Recht gilt jedoch nicht schrankenlos, anderenfalls müsste der Verantwortliche beliebige, auch für ihn vollkommen irrelevante Daten über den Betroffenen speichern, was schon mit den Grundsätzen der Datenminimierung und Speicherbegrenzung kollidiert. Die Ergänzung der Daten darf nur soweit verlangt werden, wie die Daten für die Zwecke der Verarbeitung notwendig sind.

Problem Wie bei allen Betroffenenrechten muss der Verantwortliche in der Lage sein, die von der Anfrage betroffenen Daten in seinem System zu identifizieren. Soweit das falsche Datum an verschiedenen Stellen gespeichert ist, so ist dies überall zu korrigieren.

Da es sich beim Recht auf Berichtigung um eine Ausgestaltung des Grundsatzes der Datenrichtigkeit (Abschnitt 5.1.6) handelt, entstehen hier zudem die selben Problematiken (Richtigkeit von Daten im Kontext der Verarbeitung, Kontextverfälschung, Veränderbarkeit von Daten).

Anwendungsszenario Ein Kundenkonto wird vom Händler wegen hoher Rückgabequote gesperrt, da er einen Missbrauch vermutet. Der Kunde hält die Sperrung für ungerechtfertigt, da er zwar viele Lieferungen zurückgesendet hat, die meisten Rücksendungen jedoch Reklamationen waren, da falsche oder defekte Waren geliefert wurden. Die tatsächliche Rückgabequote, also die Rückgaben, da dem Kunde der Artikel nicht gefällt oder er es sich anders überlegt, sei deutlich geringer und liege im üblichen Rahmen. Er verlangt daher die Korrektur dieses Datums und damit verbunden die Entsperrung seines Accounts.

Obwohl die zugrunde liegenden Einzelinformationen (Anzahl Rücksendungen) richtig sind, wird hier aus Sicht des Kunden der Kontext verfälscht, wenn alle Rücksendungen, unabhängig von ihrer Ursache, für die Rückgabequote in einen Topf geworfen werden.

Lösungsansatz Da das System aufgrund des Auskunftsrecht in der Lage sein muss, alle personenbezogene Daten zu einer Person zu identifizieren, ist ein zweistufiger Ansatz zur Identifikation der zu korrigierenden Daten naheliegend. Zunächst werden die gespeicherten personenbezogenen Daten zu dieser Person identifiziert und innerhalb dieses Datensatz werden danach die fehlerhaften oder unvollständigen Daten korrigiert.

Für die bereits beim Grundsatz der Datenrichtigkeit (Abschnitt 5.1.6) diskutierten Probleme können die bereits dort diskutierten Lösungsansätze herangezogen werden (z.B. Modifikation des Zonenmodells).

5.3.3 Recht auf Löschung

Das Betroffenenrecht auf Löschung personenbezogener Daten (Art. 17 DSGVO) verpflichtet den Verantwortlichen nicht nur, wie es der Begriff Betroffenenrecht vermuten lässt, personenbezogene Daten des Betroffenen auf dessen Verlangen zu löschen, sondern beinhaltet gleichzeitig auch die Pflicht des Verantwortlichen diese Löschung proaktiv durchzuführen. In der DSGVO sind Kriterien definiert, wann die Daten gelöscht werden müssen. Dabei können diese Kriterien der Löschung auf Verlangen und der selbstständigen Löschung zugeordnet werden [GSSV18].

Auf Verlangen des Betroffenen müssen personenbezogene Daten gelöscht werden, wenn die Einwilligung für die Verarbeitung widerrufen wird (vgl. Abschnitt 5.2.1) oder wenn der Betroffene der Verarbeitung der Daten erfolgreich widerspricht (vgl. Abschnitt 5.3.7).

Der Verantwortliche muss personenbezogene Daten selbstständig löschen, wenn die Verarbeitung für die Zwecke nicht mehr notwendig ist, die Verarbeitung der Daten rechtswidrig ist oder es für die Löschung eine gesetzliche Pflicht gibt.

Problem Da es sich beim Recht auf Löschung um die Ausgestaltung des Grundsatzes der Speicherbegrenzung handelt, gilt auch hier, dass zu jedem Datum die Information vorliegen muss, wie lange es gespeichert werden darf. Soweit dies nicht möglich ist, so muss ein Kriterium für die Dauer der Speicherung festgelegt werden.

Wie bei allen Betroffenenrechten muss der Verantwortliche in der Lage sein, die betroffenen Daten in seinem System zu identifizieren.

Da es sich beim Recht auf Berichtigung um eine Ausgestaltung des Grundsatzes der Datenrichtigkeit und der Speicherbegrenzung (Abschnitt 5.1.6 und Abschnitt 5.1.7) handelt, entstehen hier zudem die selben Problematiken (Richtigkeit von Daten im Kontext der Verarbeitung, Kontextverfälschung, Löschbarkeit von Daten).

Anwendungsszenario Wie im Anwendungsszenario Abschnitt 5.1.7 erwähnt, ist eine Löschfrist für Daten, die bspw. zum Versand eines Newsletters erhoben werden, nicht sachgerecht. In solchen Fällen kann statt einer Löschfrist auch ein Löschkriterium, in diesem Beispiel die Newsletterabmeldung, festgelegt werden.

Lösungsansatz Wie beim Grundsatz der Speicherbegrenzung diskutiert (vgl. Abschnitt 5.1.7) muss zu jedem Datum eine Löschfrist, bzw. ein Kriterium zur Speicherdauer festgelegt werden, um den Löschanforderungen beim Zweckwegfall gerecht werden zu können. Um die Umsetzung der Löschung sicherstellen zu können, bietet es sich an, dass diese Daten als Metadaten zu den personenbezogenen Daten gespeichert werden. Die Löschung kann dann durch Governance-Prozesse durchgeführt und sichergestellt werden.

Verlangt der Betroffene eine Löschung der Daten nach Widerruf der Einwilligung oder Widerspruch, so muss der Verantwortliche - schon im Rahmen der Bearbeitung dieser Anfragen des Betroffenen - die jeweiligen Daten identifizieren, um sicherzustellen, dass diese für die entsprechenden Verarbeitungen nicht mehr genutzt werden. Als Unterstützung dafür sollte - zusätzlich zur Rechtsgrundlage und zum Zweck - auch jeweils eine Kennung der Verarbeitung in den Metadaten gespeichert werden, sodass alle personenbezogenen Daten, die für eine Verarbeitung verarbeitet werden, einfach identifiziert werden können. Vor der Löschung muss der Verantwortliche überprüfen, ob die Daten aus anderen Gründen noch aufbewahrt werden müssen oder ob die Daten gelöscht werden können. Als Unterstützung dazu dienen die Metadaten. Dort ist hinterlegt, für welche Verarbeitung aus welchem Zweck die Daten noch genutzt werden und welche Löschfristen dabei gelten.

Hinweise auf eine rechtswidrige Verarbeitung von Dritten, z.B. aus einer Auditierung oder Hinweis eines Mitarbeiters, erfordern vom Verantwortlichen die unverzügliche Einstellung der rechtswidrigen Verarbeitung und Löschung der Daten. Die für die Verarbeitung genutzten Daten können, wie oben, beispielsweise über eine Verarbeitungskennung ermittelt werden.

Für die bereits beim Grundsatz der Datenrichtigkeit und Speicherbegrenzung (Abschnitt 5.1.6 und Abschnitt 5.1.7) diskutierten Probleme können die bereits dort diskutierten Lösungsansätze herangezogen werden (z.B. Modifikation des Zonenmodells).

5.3.4 Recht auf Einschränkung der Verarbeitung

Der Betroffene hat das Recht, die Verarbeitung der personenbezogenen Daten einschränken zu lassen (Art. 18 DSGVO). Dieses Recht kann während der Prüfung von Fällen geltend gemacht werden, in denen vom Betroffenen die Richtigkeit von personenbezogenen Daten angezweifelt wird oder in denen Widerspruch gegen eine Verarbeitung aufgrund des öffentlichen Interesses oder einer Interessenabwägung eingelegt wird (vgl. Abschnitt 5.3.7). Außerdem kann der Betroffene die Einschränkung der Verarbeitung anstatt der Löschung der Daten verlangen, um Rechtsansprüche geltend zu machen.

Umgesetzt wird die Einschränkung meist durch eine Sperrung (Beispiele für weitere Möglichkeiten in ErwG 67) - die Daten dürfen nur noch gespeichert und nur in eng begrenzten Ausnahmefällen weiterverarbeitet werden (Details Art. 18 (2) DSGVO). Der Betroffene muss vom Verantwortlichen darüber informiert werden, bevor die Sperrung aufgehoben wird.

Problem Wie bei allen Betroffenenrechten muss der Verantwortliche in der Lage sein, die von der Anfrage betroffenen Daten in seinem System zu identifizieren.

Der Verantwortliche muss sicherstellen, dass die betroffenen Daten gesperrt sind und, abgesehen von den gesetzlich vorgesehenen Ausnahmen, nicht weiterverarbeitet werden.

Anwendungsbeispiel

Der Kunde, der die Sperrung seines Kundenkontos aufgrund der hohen Rückgabequote für ungerechtfertigt hält und daher die Korrektur dieses Datums fordert (vgl. Anwendungsbeispiel in Abschnitt 5.3.2), erwartet, dass die Klärung dieses Vorgangs Zeit benötigt und verlangt daher zudem die Sperrung des strittigen Datums.

Da der Händler die Beschwerde ernst nimmt und den kompletten Prozess überprüft, nutzt er dieses Datum bis zur Klärung des Sachverhalts nicht und entsperrt folglich das Kundenkonto.

Stellt der Händler nach Prüfung des Sachverhalts fest, dass aus seiner Sicht das Datum korrekt ist und er es wieder verwenden möchte, so muss er den Betroffenen davor darüber informieren, um ihm so die Möglichkeit zu geben, sich gegen diese Entscheidung beispielsweise mit einer Beschwerde bei der Aufsichtsbehörde zu wehren.

Lösungsansatz Die Identifikation der betroffenen Daten ist im Fall des Rechts auf Einschränkung im Rahmen der Betroffenenrechte auf Berichtigung oder auf Widerspruch einfach, da sich die Einschränkung nur auf Daten beziehen kann, deren Status in der Bearbeitung des entsprechenden Antrags des Betroffenen noch ungeklärt ist. In diesen Fällen sind die betroffenen Daten bereits in der Handhabung der zugrunde liegenden Betroffenenrechte identifiziert. Diese Daten müssen für den Zeitraum der Einschränkung der Verarbeitung in den Metadaten als gesperrt markiert werden.

Im Fall der Einschränkung statt Löschung betrifft das Recht auf Einschränkung Daten, die der Verantwortliche proaktiv löschen muss (Wegfall des Zwecks, Rechtswidrigkeit der Verarbeitung). Da der Verantwortliche auf diese Fälle üblicherweise auf andere Weise aufmerksam wird als durch einen Antrag des Betroffenen (vgl. auch Abschnitt 5.3.3), erschließt sich nicht, wie der Betroffene in diesem Fall sein Recht auf Einschränkung geltend machen soll. Je nach Einzelfall ist in diesen Fällen eine manuelle Lösung notwendig.

Sofern der Betroffene den Verantwortlichen auf den Wegfall des Zwecks oder die Rechtswidrigkeit hinweist und gleichzeitig die Einschränkung statt Löschung fordert, so müssen die Daten in den Metadaten ebenfalls als gesperrt markiert werden. Die Identifikation der Daten wird im Einzelfall meist manuell erfolgen müssen.

Durch Governance-Prozesse im Data Lake muss sichergestellt werden, dass diese Daten nicht weiterverarbeitet oder gar gelöscht werden.

5.3.5 Mitteilungspflicht im Fall einer Berichtigung, Löschung oder Einschränkung

Auch die Mitteilungspflicht im Fall einer Berichtigung, Löschung oder Einschränkung (Art. 19 DSGVO) besteht aus einer Pflicht des Verantwortlichen, der dieser proaktiv nachkommen muss und einem Betroffenenrecht, das der Betroffene einfordern kann.

Sofern personenbezogene Daten im Rahmen der Verarbeitung durch den Verantwortlichen an Dritte weitergegeben wurden, muss der Verantwortliche diese Empfänger der Daten darüber informieren, dass die Daten ergänzt, berichtigt oder gelöscht werden müssen oder dass für die Daten eine Verarbeitungseinschränkung gilt (Art. 19 S. 1 DSGVO).

Ziel dieser Regelung ist, dass diese Betroffenenrechte effektiv durchgesetzt werden können und somit nicht nur im Bezug auf den Verantwortlichen gelten, sondern auch gegenüber den Empfängern, an die die Daten weitergegeben wurden [GSSV18]. Die Empfänger müssen die Umsetzung des entsprechenden Betroffenenrechts jedoch nicht blind durchführen, sondern prüfen selbstständig, ob die Voraussetzung für die Umsetzung der Betroffenenrechte auch bei ihnen gegeben sind [GSSV18].

Neben dieser proaktiven Pflicht des Verantwortlichen kann der Betroffene außerdem verlangen, dass ihm alle Empfänger der betroffenen Daten genannt werden (Art. 19 S. 2 DSGVO).

Ziel der Informationspflicht gegenüber dem Betroffenen ist es, den Betroffenen in die Lage zu versetzen, die Durchsetzung seiner Rechte bei den Empfängern durch bspw. ein Auskunftsbegehren zu kontrollieren [GSSV18].

Problem Um der Mitteilungspflicht gegenüber den Empfängern und dem Betroffenenrecht auf Mitteilung der Empfänger bei Berichtigung, Löschung oder Sperrung gerecht zu werden, muss der Verantwortliche wissen, welche Daten er an wen weitergegeben hat.

Anwendungsszenario Für die Auslieferung einer Bestellung oder die Zahlungsabwicklung greift der Onlinehändler auf externe Dienstleister zurück. Damit diese ihre Aufgabe erfüllen können, erhalten sie vom Händler die benötigten personenbezogenen Daten über den Betroffenen. Macht dieser nun eines der drei Betroffenenrechte (Berichtigung, Löschung oder Einschränkung) gegenüber dem Händler geltend, so ist der Händler verpflichtet, dies auch seinen Dienstleistern mitzuteilen.

Lösungsansatz Um zu wissen, welche Daten an wen weitergegeben wurden, sollte diese Information bei jeder Datenweitergabe an Dritte in den Metadaten des jeweiligen Datums gespeichert werden. Dies ermöglicht eine einfache Identifikation der zu informierenden Empfänger und ermöglicht außerdem dem Betroffenen mitzuteilen, an wen seine Daten weitergegeben wurden. Durch entsprechende Governance-Prozesse sollte sichergestellt werden, dass die Metadaten bei Weitergabe der Daten immer gespeichert werden.

Ungelöst bleibt damit zunächst die praktische Frage, wie die Mitteilungen an die Empfänger weitergegeben werden sollen. Das Gesetz macht dazu keine Vorgaben. Es ist daher ratsam mit dem Empfänger - im Rahmen der Vereinbarung zur Datenübermittlung - auch eine Vereinbarung zu treffen, wie die Mitteilungspflicht erfüllt werden soll. Aus Sicht des Verantwortlichen wäre, insbesondere bei vielen Empfängern, eine Lösung, die die Erfüllung der Mitteilungspflicht automatisiert, zu bevorzugen. Denkbare, einfache Möglichkeiten stellen verschlüsselte, automatisierte E-Mails oder HTTP-Requests dar.

5.3.6 Recht auf Datenportabilität

Der Betroffene hat das Recht Daten, die er bereitgestellt hat, in einem „strukturierten, gängigen und maschinenlesbaren Format“ zu erhalten, sofern die Daten aufgrund einer Einwilligung oder eines Vertrages verarbeitet werden und die Verarbeitung mittels automatisierter Verfahren erfolgt (Art. 20 DSGVO). Auf Wunsch des Betroffenen können die Daten auch direkt an einen anderen Verantwortlichen übermittelt werden.

Wie weit dieses Recht reicht und welche Daten umfasst sind, ist unklar und Teil der fachlichen Diskussion [AE17a; Dur17; Pla16]. Auch da es sich um ein neues Recht ohne Vorläufer in den alten Regelungen handelt, wird eine endgültige Klärung bis zu einem (höchst-)richterlichen Urteil offen bleiben.

Auch wenn die Formulierung des Gesetzes offen ist und grundsätzlich jede Form der Datenverarbeitung unter die Datenportabilität fallen kann, so wird aus den Gesetzgebungsunterlagen und der Idee hinter dem Recht auf Datenportabilität klar, dass es vor allem im Hinblick auf Internetdienste, im Speziellen soziale Netzwerke, konzipiert und gedacht ist.

Einige der strittigen Formulierungen werden im Folgenden kurz dargestellt.

Zunächst sind für das Recht auf Datenportabilität nur personenbezogene Daten des Betroffenen relevant, das bedeutet, anonymisierte oder andere nicht personenbezogene Daten sind nicht Teil des Portabilitätsanspruchs. Außerdem können nur Daten, die nur den Betroffenen selbst betreffen, portiert werden. Daten, die auch Dritte betreffen, wie E-Mails (Absender, CC-Empfänger), Chatprotokolle (Chatpartner) und Bilder (andere abgebildete Personen), dürfen nicht portiert werden - dies würde das Recht auf Datenschutz dieser Personen verletzen[1] (Art. 20 (4) DSGVO).

Da das Gesetz den Portabilitätsanspruch nur für Daten vorsieht, die der Betroffene zur Verfügung stellt, bezieht er sich nicht auf Daten über den Betroffenen, die der Verantwortliche aus anderen Quellen erhebt, oder Daten, die aus der Weiterverarbeitung von Daten des Betroffenen entstehen (z.B. Arztdiagnosen oder Bewegungsprofile) [GSSV18]. Ebenso entfällt der Portabilitätsanspruch, sofern die Verarbeitung auf einer anderen Rechtsgrundlage als Vertrag oder Einwilligung basiert. Strittig ist, wie weit der Begriff bezüglich Nutzungsdaten, wie bspw. aufgezeichnete Bewegungsdaten bei einem Navigationsdienst auszulegen ist. Eine Ansicht, insbesondere gestützt durch die Leitlinie der europäischen Aufsichtsbehörden, sieht in diesen vom Verantwortlichen beim Betroffenen „beobachteten Daten“ Daten, die vom Betroffenen dem Verantwortlichen bereitgestellt werden [AE17a]. Die andere Ansicht stützt sich dagegen vor allem auf die Formulierung der „Bereitstellung durch den Betroffenen“, die ein aktives, freiwilliges und wissentliches Handeln des Betroffenen nahelegt, was bei aufgezeichneten Nutzungsdaten eben nicht der Fall sei [GSSV18].

Ebenso unklar ist, wie genau die Daten formatiert sein sollen. In einem Erwägungsgrund wird angegeben, dass es nahegelegt wird, ein Austauschdatenformat zu entwickeln. Das ist jedoch nicht verpflichtend vorgegeben und es ist auch nicht klar ob und wie weit bei einer direkten Weitergabe die Verantwortlichen kooperieren müssen [GSSV18].

Um die Portabilität der Daten zu einem anderen Verantwortlichen zu erleichtern, sieht die DSGVO zudem vor, dass die Daten auf Wunsch des Betroffenen direkt an den anderen Verantwortlichen übermittelt werden müssen (Art. 20 (2) DSGVO). Wie weit die Ausnahme reicht, dass die direkte Übermittlung „technisch machbar“ sein muss, ist unklar, allerdings legt die Formulierung nahe, dass es sich um eine echte Ausnahme handelt, da auf die technische Machbarkeit und nicht z.B. auf einen unverhältnismäßigen Aufwand abgestellt wird.

Problem Die beschriebenen Unklarheiten, wie das Recht auf Datenportabilität zu verstehen ist, insbesondere im Hinblick auf die umfassten Daten und das Austauschformat, werden im Data Lake dadurch verschärft, dass Daten aus unterschiedlichen Anwendungskontexten und System gesammelt werden.

Anwendungsbeispiel Ein langjähriger Kunde möchte einen anderen Onlinehändler nutzen. Da er jedoch im Laufe der Jahre viele Daten in seinem Kundenkonto gespeichert hat, um ein möglichst bequemes und einfaches Einkaufen zu ermöglichen, möchte er diese Daten zum neuen Händler

[1] Andere Ansicht: Arbeitsgruppe der europäischen Aufsichtsbehörden in der Leitlinie zur Datenportabilität [AE17a]. Dies ist jedoch eine sehr freie Auslegung des Wortlauts des Gesetzes.

mitnehmen, um dort ein ähnliches Einkaufserlebnis zu haben ohne die Daten manuell eingeben zu müssen. Diese Daten können z.B. Zahlungsinformationen, Lieferadressen, Wunschlisten und ähnliche Daten umfassen.

Lösungsansatz Auch wenn es aus den genannten Gründen schwierig ist, einzuschätzen wie weit das Recht auf Datenportabilität reicht und in wie weit dies tatsächlich relevant wird, erscheint es doch ratsam, in dieser Hinsicht Vorkehrungen zu treffen und einen Data Lake entsprechend auszulegen. Es bedarf daher einer geeigneten Definition, welche Daten vom Portabilitätsanspruch umfasst sind. Naheliegend ist zudem eine entsprechende Markierung in den Metadaten, da die Frage nach der Bereitstellung bei der Datenerfassung am besten beantwortet werden kann.

Problematisch bleibt dabei die Frage nach personenbezogenen Daten Dritter, insbesondere wenn diese nicht strukturiert erfasst sind, beispielsweise in einem Freitextfeld oder auf einem Bild. Diese müssten im Falle einer Weitergabe der Daten herausgefiltert werden.

Als Formate für die Datenübermittlung bieten sich für strukturierte und semi-strukturierte Daten entsprechende allgemein genutzte und, da auf einfachem Text basierend, für jeden lesbare Formate wie z.B. CSV, XML oder JSON an.

5.3.7 Recht auf Widerspruch

Der Betroffene hat das Recht, gegen Verarbeitungen, die auf Grundlage berechtigten Interesses erfolgt, Widerspruch einzulegen (Art. 21 (1) DSGVO). Ebenso gilt das Widerspruchsrecht, wenn die Verarbeitung in Ausübung einer Aufgabe im öffentlichen Interesse oder in Ausübung öffentlicher Gewalt erfolgt. Der Lesbarkeit halber wird im Folgenden nur die Verarbeitung auf berechtigtem Interesse erwähnt, die Regelung gilt jedoch ebenso für die beiden anderen Rechtsgrundlagen. Die Idee des Widerspruchrechts ist, dem Betroffenen die Möglichkeit zu geben, sich gegen Verarbeitungen zu wehren, an deren Entstehung er nicht beteiligt ist (z.B. durch Vertragsschluss oder Einwilligung) und deren Durchführung nicht verpflichtend ist (z.B. Verarbeitung aufgrund gesetzlicher Verpflichtung).

Sofern es sich um einen Widerspruch gegen Direktwerbung handelt, so muss dieser nicht begründet werden, sondern muss auf jeden Fall umgesetzt werden.

Bis auf diese Ausnahme handelt es sich beim Recht auf Widerspruch um eine Härtefalllösung, im Gesetz wird das als „aus Gründen, die sich aus ihrer besonderen Situation ergeben“ beschrieben. Das bedeutet, der Verantwortliche muss prüfen, ob seine Interessen an der Datenverarbeitung die genannten, spezifischen Gründe des Betroffenen gegen die Verarbeitung überwiegen.

Das Widerspruchsrecht schafft so einen Ausgleich zwischen der generellen Interessenabwägung des Verantwortlichen, welche die allgemeinen Interessen an der Verarbeitung und die entgegenstehenden Interessen der Betroffenen miteinander abwägt, und dem Einzelfall, in dem bestimmte Sondersituationen (z.B. schwere Krankheit oder Geheimhaltungsinteresse) eine andere Abwägung erfordern.

Wichtig ist die Trennung vom Recht auf Widerspruch und dem Recht auf jederzeitigen Widerruf der Einwilligung. Während ersteres, wie eben diskutiert im Wesentlichen eine Härtefallregelung ist, kann das Recht auf Widerruf der Einwilligung jederzeit und unbegründet wahrgenommen werden (Abschnitt 5.2.1).

Problem Je mehr ein Verantwortlicher sich, insbesondere im Werbebereich, auf das berechtigte Interesse stützt, desto mehr Widersprüche wird er zu verarbeiten haben.

Zudem muss sichergestellt werden, dass die Widersprüche umgesetzt und berücksichtigt werden.

Anwendungsszenario Bestandskunden wird regelmäßig auf Basis berechtigten Interesses ein Newsletter zugesendet. Durch einen Link am Ende des Newsletters kann der Kunde der Nutzung seiner E-Mail-Adresse für diesen Zweck widersprechen. Da es sich um den Widerspruch gegen eine Direktmarketingmaßnahme handelt, muss der Widerspruch nicht geprüft werden, sondern kann direkt umgesetzt werden.

Lösungsansatz Um den Arbeitsaufwand bei der Umsetzung von Widersprüchen möglichst gering zu halten, sollte der Verantwortliche die Bearbeitung von Widersprüchen gegen Direktmarketing von den Widersprüchen im Härtefall organisatorisch möglichst trennen. Denkbar wäre z.B. eine eigene Kategorie für „Werbewiderspruch“ in einem Onlineformular zur Geltendmachung der Betroffenenrechte. Da es in diesem Fall keine Abwägung gibt und der Widerspruch unbedingt umgesetzt werden muss, kann die Bearbeitung dieser Widersprüche so automatisiert werden.

Der Widerspruch gegen Direktmarketingmaßnahmen muss im System hinterlegt werden. Eine Möglichkeit stellt dabei die Speicherung in den Metadaten von Kontaktinformationen dar. Soweit Kontaktdaten nicht für Direktmarketing genutzt werden dürfen, so werden die Daten in den Metadaten entsprechend markiert.

In den anderen Fällen muss die Prüfung, ob ein Widerspruch gegen eine Verarbeitung begründet ist, manuell erfolgen, da Härtefallentscheidungen immer Einzelfallentscheidungen sind.

Auch in diesem Fall sollte der Widerspruch gegen eine bestimmte Verarbeitung der Daten im System verankert werden. Dies kann ebenfalls in den Metadaten geschehen, in dem man zusätzlich zu den hinterlegten Datenfeldern Verarbeitungskennung, Zweck und Rechtsgrundlage die Sperrung für die jeweilige Verarbeitung hinterlegt, sofern es sich um eine Verarbeitung im öffentlichen Interesse oder aufgrund berechtigen Interesses handelt.

Auch beim Recht auf Widerspruch sollte die tatsächliche Umsetzung mittels Governance-Prozessen sichergestellt werden.

5.4 Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen (Data Protection by Design and by Default)

Datenschutz durch Technikgestaltung (Data Protection by Design, Art. 25 (1) DSGVO) verpflichtet den Verantwortlichen von Anfang an, das heißt vom Design über Entwurf und Implementierung bis hin zum laufenden Betrieb, Datenschutz mitzudenken. Dies gilt insbesondere für die Grundsätze der Verarbeitung personenbezogener Daten. Datenschutz muss bei der Wahl der Mittel zur Verarbeitung, das heißt unter anderem bei der Wahl der zur Verarbeitung genutzten Systeme und Technologien, berücksichtigt werden. Ziel ist es, dass Datenschutz technischen Systemen inhärent sein soll und so der Betroffene geschützt wird [GSSV18].

Datenschutz durch Technikgestaltung bedeutet jedoch nicht, dass jede mögliche TOM, unabhängig vom Preis oder der Menge und Sensibilität der Daten, getroffen werden muss. Stattdessen müssen bei der Wahl geeigneter TOMs die Implementierungskosten, die Art, Umfang, Umstände und Zwecke der Verarbeitung sowie Eintrittswahrscheinlichkeit und Schwere der Risiken für den Betroffenen berücksichtigt werden. Zugleich wird auch ein technischer Bezugsrahmen für die Maßnahmen festgelegt, sie müssen den „Stand der Technik“ berücksichtigen. Diese offene Formulierung, ohne Vorgabe konkreter Maßnahmen, gewährleistet zum einen Technikneutralität und vermeidet, dass das Gesetz in einigen Jahren vollkommen veraltet ist, lässt allerdings den Verantwortlichen über die Maßstäbe, die genau an den Begriff angelegt werden müssen und welche Maßnahmen daher notwendig sind, im Unklaren [GSSV18].

Legt man die vom Bundesverfassungsgericht im Streit um die Genehmigung eines Atomreaktors (sog. Kalkar 1-Entscheidung) konstruierte Drei-Stufen-Theorie zugrunde, so ist der „Stand der Technik“ zwischen den „allgemein anerkannten Regeln der Technik“ und „Stand von Wissenschaft und Technik“ einzuordnen [Bun78]. Die „allgemein anerkannten Regeln der Technik“ meinen dabei, dass die Wirksamkeit der Maßnahme zur Erreichung eines bestimmten Zieles allgemein anerkannt ist (allgemein bezieht sich dabei nicht die allgemeine Öffentlichkeit, sondern „technische Praktiker“). Ein Beispiel dafür ist die Nutzung von HTTPS zur Eingabe von personenbezogenen Daten im Internet. Es ist allgemein anerkannt, dass dadurch Datendiebstahl (z.B. durch Sniffing) verhindert wird. Der „Stand von Wissenschaft und Technik“ meint dagegen, dass neueste wissenschaftliche Erkenntnisse berücksichtigt werden müssen, auch wenn die daraus abgeleiteten Maßnahmen noch nicht marktreif sind [Tel18]. Offensichtlich kann dies nur in Bereichen mit sehr großem Risiko eine Rolle spielen. Ein Beispiel ist dafür der Fall, indem diese Theorie entwickelt wurde: Sicherungsmaßnahmen für Atomkraftwerke müssen diesem Stand entsprechen. Der dazwischenliegende „Stand der Technik“ muss sich nicht an dem wissenschaftlichen Stand orientieren, ist jedoch auch noch nicht allgemein anerkannt. Vereinfacht ist damit die beste am Markt verfügbare Maßnahme gemeint [Tel18].

Bedingt durch diese Definition verschiebt sich der Stand der Technik stetig, weshalb die geeigneten Maßnahmen nicht nur zu Beginn der Verarbeitung definiert werden können und danach nicht mehr verändert werden müssen. Stattdessen ist eine regelmäßige Überprüfung erforderlich, ob die Maßnahmen noch den Anforderungen entsprechen oder es mittlerweile geeignetere Maßnahmen gibt [GSSV18; Tel18].

Die DSGVO fordert vom Verantwortlichen datenschutzfreundliche Voreinstellungen (Data Protection by Default, Art. 25 (2) DSGVO). Das heißt, er soll Verarbeitungen und Systeme so auslegen, dass, wenn der Betroffene keine Voreinstellungen ändert, personenbezogene Daten nur soweit

verarbeitet werden, wie es für den Zweck erforderlich ist, sowie nur so wenig Daten wie möglich erhoben und nur solange wie nötig verarbeitet werden. Im Grunde bedeutet dies, dass die Grundsätze der Zweckbindung, Datenminimierung und Speicherbegrenzung technisch umgesetzt werden müssen [GSSV18].

Problem Datenschutz durch Technikgestaltung bedeutet, dass der Data Lake zur Verarbeitung personenbezogener Daten so gestaltet werden muss, dass er die Einhaltung der gesetzlichen Vorgaben zum Datenschutz erleichtert und unterstützt.

Diese Norm ist der rechtliche Anknüpfungspunkt, weshalb die datenschutzrechtlichen Anforderungen der DSGVO, obwohl zunächst an die einzelnen Verarbeitungen gerichtet, für das Speichersystem Data Lake relevant sind. Da der Data Lake für verschiedene Verarbeitungen als Speichersystem dient, muss er als Mittel einer anderen Verarbeitung ebenso den Grundsätzen der Datenverarbeitung und damit natürlich auch deren Präzisierung und Konkretisierung in den anderen Pflichten des Verantwortlichen genügen.

Durch den sich ändernden Stand der Technik müssen getroffenen Maßnahmen regelmäßig auf ihre Tauglichkeit überprüft werden.

Anwendungsbeispiel Neben dem beim Grundsatz der Datenminimierung vorgestellten Beispiel der Bestellung ohne Kundenkonto ist auch eine standardmäßige Verschlüsselung des Webseitenzugriffs eine Maßnahme zum Datenschutz durch Technikgestaltung.

Um die datenschutzfreundlichen Voreinstellungen umzusetzen, kann der Händler die beim Aufruf des Onlineshops gesetzten Cookies auf die zur Funktion der Seite benötigten Cookies beschränken und den Kunden um die Zustimmung zur Setzung weiterer Cookies, wie z.B. Tracking-Cookies, bitten.

Lösungsansatz Da ein Data Lake ein Mittel zur Verarbeitung von personenbezogenen Daten ist, müssen die in diesem Kapitel diskutierten Pflichten des Verantwortlichen bereits bei der Konstruktion des Data Lakes beachtet werden. Der Data Lake sollte durch geeignete Wahl der genutzten Technologien und durch Governance-Prozesse den Verantwortlichen bei der Einhaltung seiner Pflichten unterstützen.

Die getroffenen Maßnahmen müssen regelmäßig z.B. im Rahmen eines kontinuierlichen Verbesserungsprozesses überprüft und gegebenenfalls angepasst werden.

5.5 Sicherheit der Verarbeitung

Die Datenschutzgrundverordnung ist kein IT-Sicherheitsgesetz, stellt aber dennoch in Artikel 32 bestimmte Anforderungen an die IT-Sicherheit von Systemen, die bei der Verarbeitung personenbezogener Daten zum Einsatz kommen. IT-Sicherheit wird dabei als ein notwendiges Werkzeug verstanden, um Datenschutz gewährleisten zu können.

Die Maßstäbe, die an die IT-Sicherheitsmaßnahmen angelegt werden müssen, sind grundsätzlich dieselben wie beim Datenschutz durch Technikgestaltung. Das heißt, bei der Wahl der Maßnahmen müssen Implementierungskosten, Stand der Technik sowie Art, Umfang, Umstände und Zwecke der Verarbeitung wie auch Eintrittswahrscheinlichkeit und Schwere der Risiken für den Betroffenen berücksichtigt werden.

Die DSGVO stellt keine expliziten Anforderungen, welche IT-Sicherheitsmaßnahmen ergriffen werden müssen, da geeignete und empfohlene Sicherheitsmaßnahmen sich im Lauf der Zeit verändern und um Technikneutralität zu gewährleisten. Stattdessen stellt sie auf den „Stand der Technik“ ab, der Verantwortliche muss die Entscheidung treffen, welche Maßnahmen im jeweiligen Verarbeitungskontext notwendig sind, um ein angemessenes Schutzniveau zu gewährleisten. Ebenso ist es regelmäßig erforderlich, die getroffenen Maßnahmen zu überprüfen und gegebenenfalls anzupassen.

Auch wenn die Anforderungen an die Maßnahmen gleich formuliert sind, so muss bei der Umsetzung klar zwischen Datenschutz durch Technikgestaltung und der Sicherheit der Verarbeitung unterschieden werden, da beide je ein eigenes Schutzziel verfolgen und somit andere Maßnahmen notwendig sein können. Es ist problemlos möglich, eine Verarbeitung aus Sicht der IT-Sicherheit angemessen sicher zu gestalten, während grundlegende datenschutzrechtliche Anforderungen, wie z.B. Speicherminimierung oder Zweckbindung nicht beachtet werden.

Während Datenschutz durch Technikgestaltung die Einhaltung der Grundsätze der Verarbeitung personenbezogener Daten sicherstellen soll, geht es bei der Sicherheit der Verarbeitung vor allem um die Datensicherheit, die insbesondere durch die klassischen Ziele der IT-Sicherheit, das heißt Vertraulichkeit, Integrität und Verfügbarkeit, erreicht werden soll. Ergänzt werden diese Ziele durch die explizite Nennung von Teilaspekten dieser Ziele wie Verschlüsselung und Pseudonymisierung (als Maßnahme für die Vertraulichkeit), Belastbarkeit (als Teil der Verfügbarkeit) und Wiederherstellung der Daten nach einem Zwischenfall (ebenfalls Teil der Verfügbarkeit) [GSSV18].

Problem Auch wenn diese Regelung zunächst auf die Sicherheit der einzelnen Verarbeitung abstellt, so ist sie dennoch bei der Konstruktion des Data Lakes sehr relevant, da der Data Lake qua Konstruktion als Speichersystem für viele verschiedene Verarbeitungen genutzt wird. Er ist daher besonderen Anforderungen an die IT-Sicherheit unterworfen, er muss schließlich allen Sicherheitsanforderungen gerecht werden, die an die einzelnen Verarbeitungen gestellt werden.

Außerdem muss beachtet werden, dass durch die Speicherung von Daten aus verschiedenen Kontexten und Verarbeitungen am selben Ort möglicherweise weitere Risiken für den Betroffenen entstehen, falls beispielsweise Daten unbeabsichtigt in die Hände Dritter gelangen. Dies führt ggf. zu weiteren und höheren Anforderungen an die IT-Sicherheit des Data Lakes.

Auch wenn die DSGVO die IT-Sicherheitsmaßnahmen grundsätzlich auch die Implementierungskosten als Maßstab für die Wahl der Maßnahmen nennt, so scheint es doch naheliegend zu sein, dass dieser Maßstab, wenn überhaupt, nur sehr bedingt für Sicherheitsmaßnahmen gelten kann, die nur aufgrund der Wahl eines komplexeren Speichersystems durch den Verantwortlichen notwendig sind. Anderenfalls wäre hier eine große Regelungslücke gegeben, dass man durch besonders komplex oder schlecht designte Systeme die Anforderungen an die IT-Sicherheit unterlaufen könnte.

Zudem wird bei den Maßnahmen zur IT-Sicherheit im Gesetz explizit ein Verfahren zur regelmäßigen Überprüfung, Bewertung und Evaluierung der IT-Sicherheitsmaßnahmen gefordert (Art. 32 (1) lit. d DSGVO).

Lösungsansatz Um den Anforderungen der einzelnen Verarbeitungen im Data Lake gerecht zu werden, muss bereits beim Design des Data Lakes eine Vorstellung darüber herrschen, welche Verarbeitungen den Data Lake zukünftig nutzen und welche personenbezogenen Daten dort gespeichert werden sollen. Entsprechend der daraus resultierenden Anforderungen an die IT-Sicherheit müssen geeignete Maßnahmen getroffen werden.

Zudem muss in übergeordneten Prozessen die Überprüfung der IT-Sicherheitsmaßnahmen verankert werden. Eine Möglichkeit dafür ist die Orientierung an oder Zertifizierung nach Standards wie dem BSI-Grundschutzstandard oder der ISO 27001 [GSSV18].

5.6 Datenschutz-Folgenabschätzung

Die Datenschutzgrundverordnung sieht für Verarbeitungen, die „voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen“ darstellen können, eine vorherige Datenschutz-Folgenabschätzung (DSFA oder DPIA/PIA von engl. Data Protection Impact Assesment) vor (Art. 35 DSGVO). Dadurch soll das Risiko auf ein akzeptables Niveau gesenkt werden.

Der Begriff des Risikos darf dabei nicht mit dem sonst im Unternehmenskontext verwendeten Risikobegriff, dem Risiko aus Sicht des Unternehmen, verwechselt werden, sondern stellt ausschließlich auf die Auswirkungen auf den Betroffenen ab und nennt als zu berücksichtigende Kriterien Art, Umfang, Umstände und Zwecke der Verarbeitung.

Zunächst nennt die DSGVO drei Kategorien von Verarbeitungen, in denen eine Datenschutz-Folgenabschätzung (DSFA) erforderlich ist (Art. 35 (3) DSGVO).

Bewertung persönlicher Aspekte Eine Bewertung persönlicher Aspekte erfolgt insbesondere bei der Erzeugung von Datengrundlagen für automatisierte Entscheidungen im Einzelfall (vgl. Art. 22 DSGVO) und bei der automatisierten Erstellung von Persönlichkeitsprofilen (z.B. Bewerberauswahl) [GSSV18]. Diese Verarbeitungen greifen besonders tief in die Privatsphäre und das Persönlichkeitsrecht ein und begründen so ein hohes Risiko.

Umfangreiche Verarbeitung sensibler Daten Sensible Daten umfasst zum einen besondere Kategorien personenbezogener Daten (Art. 9 (1) DSGVO) und zum anderen personenbezogene Daten über strafrechtliche Verurteilungen (Art. 10 DSGVO). Aufgrund der besonderen Sensitivität der Daten und des Schadens, der für die Betroffenen entstehen kann, wenn diese Daten z.B. öffentlich werden, muss für solche Verarbeitungen ein hohes Risiko angenommen werden.

Überwachung öffentlich zugänglicher Bereiche Dies gilt insbesondere für die Videoüberwachung von öffentlichem Raum (z.B. Straßen oder Plätze) und öffentlich zugänglichem Raum (z.B. Einkaufszentren, Flughäfen). Insbesondere die Möglichkeit, aus den Überwachungsdaten massenhaft Bewegungsprofile zu erstellen, begründet hier ein hohes Risiko für den Betroffenen.

Des Weiteren verpflichtet die DSGVO die Aufsichtsbehörden eine Blacklist von Verarbeitungen bzw. Kategorien von Verarbeitungen zu erstellen, bei denen in jedem Fall eine DSFA durchzuführen ist (Art. 35 (4) DSGVO). Außerdem können die Aufsichtsbehörden eine Whitelist von Verarbeitungen erstellen, bei denen keine DSFA erforderlich ist (Art. 35 (5) DSGVO). Beide Listen müssen zwischen den Aufsichtsbehörden abgestimmt werden, soweit die dort genannten Verarbeitungen Angebote im gemeinsamen Binnenmarkt betreffen können (Art. 35 (6) DSGVO). Bisher wurden nur Entwürfe der Blacklist veröffentlicht, die Grundlage für den europäischen Abstimmungsprozess sein sollen.

Sofern die oben genannten Kriterien nicht anwendbar sind, so muss der Verantwortliche die Risikoeinschätzung eigenverantwortlich durchführen. Dies bedeutet für den Verantwortlichen, dass er unter Berücksichtigung der geplanten TOMs für jede Verarbeitung personenbezogener Daten zunächst eine vorläufige Risikoeinschätzung durchführen muss, um zu bestimmen, ob eine DSFA notwendig ist. Auch wenn keine DSFA notwendig ist, so ist die Risikoabschätzung zu dokumentieren, um der Rechenschaftspflicht genüge zu tun. Eine hilfreiche Unterstützung dazu ist die Leitlinie zur Datenschutzfolgenabschätzung des Europäischen Datenschutzausschusses, in der Kriterien definiert werden, die die Erfordernis einer DSFA begründen [AE17b].

Die Prüfung, wann eine DSFA durchzuführen ist, lässt sich wie in Abbildung 5.1 als ein vierstufiges Prüfschema darstellen.

Eine DSFA muss vollständig dokumentiert werden und muss mindestens Folgendes umfassen (Art. 35 (7) DSGVO):

Beschreibung der Verarbeitung Die systematische Beschreibung der Verarbeitung muss zumindest die geplanten Verarbeitungsvorgänge, die Zwecke der Verarbeitung, die Rechtsgrundlage und, sofern der Verantwortliche sich auf berechtigtes Interesse stützt, die Interessenabwägung umfassen.

Notwendigkeit und Verhältnismäßigkeit Der Verantwortliche muss die Notwendigkeit und Verhältnismäßigkeit der Verarbeitung in Bezug auf den Zweck darlegen. Das heißt er muss nachweisen, dass die Verarbeitung geeignet und erforderlich ist, um den Zweck zu erreichen, also das mildeste Mittel dafür darstellt [GSSV18]. Des Weiteren muss die Angemessenheit nachgewiesen werden, das heißt, dass die Vorteile, die sich aus der Verarbeitung ergeben, und die Nachteile für den Betroffenen in angemessenem Verhältnis stehen [GSSV18].

Risikobewertung Es muss eine Risikobewertung der für den Betroffenen entstehenden Risiken durchgeführt werden, das heißt es müssen die Risiken für den Betroffenen anhand der Schwere des Eingriffs in die Rechte und Freiheiten des Betroffenen und der Eintrittswahrscheinlichkeit bewertet werden. Als Grundlage kann die vorläufige Risikobewertung dienen, mit der die Notwenigkeit einer DSFA bestimmt wurde.

Abhilfemaßnahmen Den in der Risikobewertung ermittelten Risiken müssen geeignete Abhilfemaßnahmen gegenübergestellt werden, die das Risiko senken, das heißt entweder die Eingriffsschwere (z.B. Verschlüsselung der Daten, sodass beim Verlust der Daten die Daten nicht missbraucht werden können) oder die Eintrittswahrscheinlichkeit (z.B. strikte IT-Sicherheitsmaßnahmen, um Verlustgefahr zu minimieren) senken.

Sofern auch nach der DSFA ein hohes Risiko für den Betroffenen besteht, das heißt, dass keine geeignete Abhilfemaßnahmen getroffen werden können, so kann der Verantwortliche die Aufsichtsbehörde konsultieren (Art. 36 DSGVO). Die Verarbeitung darf dann nur realisiert werden, sofern die Aufsichtsbehörde dem zustimmt und die ggf. gestellten Auflagen erfüllt werden.

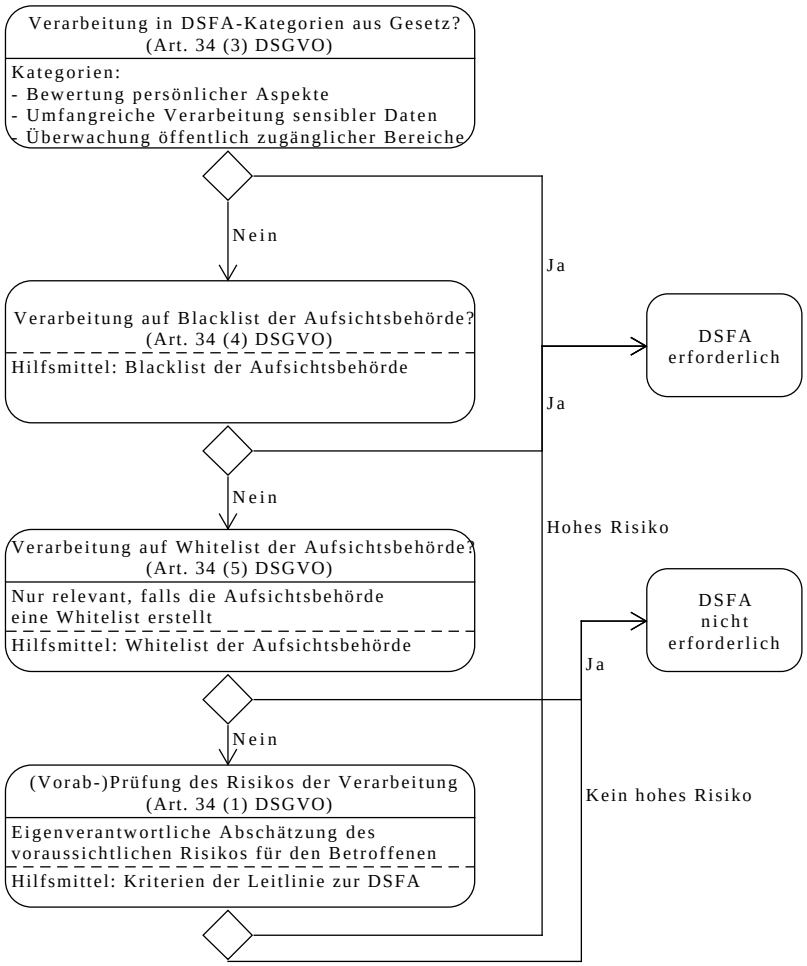

Abbildung 5.1: Prüfschema Datenschutz-Folgenabschätzung nach Art. 35 DSGVO

Bei einer Änderung des Risikos muss die DSFA überprüft und gegebenenfalls überarbeitet werden (Art. 35 (11) DSGVO). Die Risikoanalyse sollte daher regelmäßig überprüft werden, ob die Risikoeinschätzungen nach wie vor aktuell ist und damit auch ob die TOMs noch dem Stand der Technik entsprechen und das Risiko nach wie vor adäquat eindämmen. Außerdem empfiehlt sich eine regelmäßige Überprüfung des DSFA, insbesondere im Hinblick auf die tatsächliche Wirksamkeit der implementierten TOMs und ob die Verarbeitung entsprechend der DSFA durchgeführt wird.

Problem Im Data Lake gibt es die oben als Data-Lake-Verarbeitungen bezeichnete Verarbeitungen, bei denen die Daten aus den verschiedenen, internen wie externen, Quellen gemeinsam analysiert werden sollen, um Wertschöpfung für den Betreiber des Data Lakes zu generieren. Diese Verarbeitungen werden vom Entwurf der Blacklist des Landesbeauftragten für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (Datenschutzaufsichtsbehörde Baden-Württemberg) erfasst [Lan18]. Dort werden im Kontext von Big Data folgende Kriterien für eine Verarbeitung genannt, bei der in jedem Fall eine DSFA durchzuführen ist:

- Zusammenführung und Weiterverarbeitung von personenbezogenen Daten aus verschiedenen Quellen in großem Umfang
- Nutzung von Daten für Zwecke, für die die Datenerhebung nicht beim Betroffenen erfolgte (z.B. Bonitätsdaten von Auskunfteien, Analyse des Verhaltens bei Werbeansprache)
- Nutzung von für den Betroffenen nicht nachvollziehbaren Algorithmen
- Ziel: Entdeckung von Zusammenhängen zwischen Daten für nicht vorher bestimmbare Zwecke

  oder

  Ziel: Erzeugung von Datengrundlagen für automatisierte Entscheidungen im Einzelfall (vgl. Art. 22 DSGVO)

Eine durch die strengen Kriterien der Blacklist nicht erfasste Data-Lake-Verarbeitung muss im Einzelfall auf das bestehende Risiko überprüft werden. Die Leitlinie zur Datenschutzfolgenabschätzung nennt neun Kriterien, von denen mindestens zwei erfüllt sein müssen, um meist ein hohes Risiko annehmen zu können. Zwei dieser Kriterien sind schon durch die Idee und Konstruktion des Data Lakes zumeist erfüllt: „Datenverarbeitung in großem Umfang“ und „Abgleichen oder Zusammenführen von Datensätzen“. Daher sind diese Verarbeitungen wohl meist als DSFA-pflichtig zu betrachten.

Anwendungsszenario Im Onlinehandel ist es, insbesondere als großer Anbieter, unerlässlich, Fraud-Prevention-Systeme einzusetzen, das heißt Systeme, die beispielsweise Bonitäts- und Plausibilitätschecks durchführen, um in Echtzeit Betrug zu verhindern. Dazu sollen alle verfügbare Daten genutzt werden, um möglichst sicherzustellen, dass kein Kunde fälschlich abgewiesen wird, aber dennoch gleichzeitig möglichst jeder Betrugsversuch erkannt wird. Dies kann vom Vergleich mit bisherigem Kaufverhalten über Standortanalysen und Checks gegen Blacklists und bis hin zu Bonitätsauskunft von Auskunfteien (z.B. Schufa) reichen.

Der Vorschlag für die Blacklist der Aufsichtsbehörde nennt als Verarbeitungen, die eine DSFA erfordern, sogar explizit „Big-Data-Analysen von Kundendaten, die mit Angaben aus Drittquellen angereichert wurden“ und Fraud-Prevention-Systeme.

Lösungsansatz Eine DSFA ist nicht technisch zu lösen. Sie ist eine Analyse und Bewertung, die vom Verantwortlichen manuell durchgeführt werden muss. Es muss sichergestellt werden, dass die Prüfung, ob eine DSFA notwendig ist, und die regelmäßige Überprüfung der DSFA in den Prozessen des Unternehmens verankert ist.

Zur Durchführung und Dokumentation der Datenschutzfolgenabschätzung hat die Commission Nationale de l’Informatique et des Libertés (CNIL, Datenschutzaufsichtsbehörde Frankreichs) eine Software entwickelt, die sich durch eine strukturierte Prozessführung mit ausführlichen Erläuterungen und Beispielen auszeichnet [Com18].

5.7 Nachweis- und Dokumentationspflichten

Einer der größten Unterschiede zwischen der alten Rechtslage nach dem Bundesdatenschutzgesetz (alte Fassung) und der neuen nach der DSGVO sind deutlich ausgeweitete und verschärfte Nachweis- und Dokumentationspflichten.

Die DSGVO kennt zwölf einzelne solcher Pflichten [Vei18]. Dies reicht von der grundsätzlichen Rechenschaftspflicht, dass die Grundsätze der Verarbeitung personenbezogener Daten eingehalten werden, über die Nachweispflicht auf Verarbeitungsebene, dass z.B. geeignete IT-Sicherheitsmaßnahmen gewählt wurden, bis hin zu Nachweispflichten im Einzelfall, dass z.B. dieser Betroffene für diese Verarbeitung dieser Daten eine Einwilligung erteilt hat.

Ein wichtiger Baustein dieser Pflichten ist das Verzeichnis der Verarbeitungstätigkeiten. Es handelt sich dabei um ein vom Verantwortlichen zu führendes Register aller Verarbeitungen personenbezogener Daten, die er durchführt (Art. 30 DSGVO). Dort müssen unter anderem Ansprechpartner, Zwecke der Verarbeitung, Kategorien der Betroffenen (z.B. Kunden oder Mitarbeiter) und Kategorien der Daten (z.B. Kontaktdaten), aber auch eine Beschreibung der getroffenen IT-Sicherheitsmaßnahmen dokumentiert werden.

Problem Die umfangreichen Dokumentations- und Nachweispflichten unter der DSGVO stellen Verantwortliche vor große Herausforderungen.

Bei der isolierten Betrachtung des Systems Data Lake in dieser Arbeit werden Informationen wie der Zweck und die Rechtsgrundlage einer Verarbeitung, aber auch die Löschfrist einzelner Daten jeweils direkt in den Metadaten des Data Lakes verankert. Dadurch entsteht in diesen Bereichen eine doppelte Dokumentation, da diese Informationen auch im Verzeichnis der Verarbeitungstätigkeiten dokumentiert werden müssen.

Lösungsansatz Auch wenn ein Data Lake als Datenspeichersystem grundsätzlich nicht dafür geeignet ist, Nachweis- und Dokumentationspflichten für Verarbeitungsprozesse zu erfüllen, so kann die umfangreiche Governance in einem Data Lake die Einhaltung dieser Pflichten unterstützen.

Wird ein Data Lake in einem Unternehmen implementiert, so sollte die doppelte Dokumentation in den Metadaten und im Verzeichnis der Verarbeitungstätigkeiten vermieden werden. Wird im Rahmen des Datenschutzmanagements bereits eine Software für das Verzeichnis verwendet, so kann über eine Verknüpfung der Systeme nachgedacht werden und es ist sogar denkbar, das Verzeichnis der Verarbeitungstätigkeiten direkt im Rahmen des Data Lakes zu führen. Dabei muss jedoch beachtet werden, dass auch der DSGVO unterfallende, dokumentationspflichtige Datenverarbeitungen außerhalb des Data Lakes bestehen.

5.8 Drittlandübertragungen

Das Kapitel 5 der DSGVO (Art. 44-50 DSGVO) widmet sich der Übermittlung von Daten an Verantwortliche in Drittländern, das heißt Staaten außerhalb der EU und des EWR.

Die Anforderungen an eine Übermittlung in ein Drittland sind zweistufig. Für die Übermittlung als Verarbeitung gelten zunächst alle allgemeine Regelungen der DSGVO, das heißt insbesondere, dass eine Rechtsgrundlage für die Übermittlung benötigt wird (vgl. Abschnitt 5.2). Zusätzlich müssen die Bestimmungen dieses Kapitels eingehalten werden. Diese sollen das Datenschutzniveau für den Betroffenen garantieren, auch wenn die Daten außerhalb des Geltungsbereichs der DSGVO verarbeitet werden. Für eine Übermittlung von personenbezogenen Daten in Drittländer muss mindestens eine der folgenden Bedingungen erfüllt sein:

Angemessenheitsbeschluss Die EU-Kommission kann einen Angemessenheitsbeschluss erlassen, der einem Drittstaat, einem Teilgebiet des Drittstaates oder bestimmten (Wirtschafts-) Sektoren in diesem Drittstaat ein angemessenes Schutzniveau bescheinigt (Art. 45 (1) DSGVO). Diese Beschlüsse werden von der Kommission veröffentlicht [EU 18]. Für Kanada gilt dieser Beschluss beispielsweise nur für den kommerziellen Sektor, da die Datenverarbeitung dort einem Datenschutzgesetz unterfallen, das einen ausreichenden Schutz bietet.

Geeignete Garantien Sofern kein Angemessenheitsbeschluss vorliegt, ist es auch möglich sich mittels Garantien des Gegenübers abzusichern, dass dieser ein angemessenes Datenschutzniveau für den Betroffenen sicherstellt (Art. 46 (1) DSGVO). Gemeinsam ist allen Garantien, dass sie einen durchsetzbaren und rechtlich einklagbaren Datenschutzstandard sicherstellen sollen. Die wichtigsten Garantien sind:

Standarddatenschutzklauseln Die Standarddatenschutzklauseln sind von der EU-Kommission beschlossene Vertragsklauseln, in denen sich das Gegenüber zur Einhaltung von Datenschutzmindeststandards verpflichtet. Werden diese dem Vertrag hinzugefügt, so stellen sie eine geeignete Garantie dar und ermöglichen eine Datenübermittlung in Drittländer (Art. 46 (2) lit. c, d DSGVO).

Verbindliche interne Datenschutzvorschriften Innerhalb von Unternehmensgruppen können verbindliche, von der Aufsichtsbehörde genehmigte Datenschutzvorschriften geeignete Garantien darstellen (Art. 46 (2) lit. b i.V.m. Art. 47 DSGVO).

Verbindliche Verhaltensregeln (Code of Conduct) Unterwerfen sich Drittlandempfänger von der Aufsichtsbehörde genehmigten, verbindlichen Verhaltensregeln, die die Einhaltung eines Datenschutzmindeststandards sicherstellen, so können diese eine geeignete Garantie darstellen (Art. 46 (2) lit. e DSGVO)

Genehmigte Zertifizierung Ist der Drittlandempfänger nach einer von der Aufsichtsbehörde genehmigte Zertifizierung zertifiziert und stellt so die Einhaltung eines Datenschutzmindeststandards sicher, so kann dies eine geeignete Garantie darstellen (Art. 46 (2) lit. f DSGVO)

Ausnahmen für bestimmte Fälle Im Einzelfall dürfen Daten auch ohne Angemessenheitsbeschluss und ohne geeignete Garantien in Drittländer übermittelt werden. Es handelt sich dabei um Ausnahmeregelungen, die nicht für Übermittlungen im Regelfall genutzt werden dürfen [Kon17]. Die wichtigsten sind die explizite Einwilligung zur Übermittlung, Erforderlichkeit zur Vertragserfüllung und - unter strengen Voraussetzungen - ein zwingendes, berechtigtes Interesse (Art. 49 DSGVO)

Problem Die Anforderungen bezüglich der Drittlandübermittlung betrifft im Kontext des Data Lakes vor allem die Wahl der Dienstleister, insbesondere im Hinblick auf Hosting- und Cloudprovider, da diese ihren Sitz meist in den USA haben. Ebenso gilt dies für die meisten Werbenetzwerke und Trackinganbieter.

Anwendungsszenario Als global agierender Onlinehändler soll Kunden rund um den Globus ein möglichst schneller und latenzarmer Zugang zum Onlineshop ermöglicht werden. Die Daten, Kunden- wie Angebotsdaten, werden daher von einem global agierenden Cloudanbieter gehostet, sodass auch Lastspitzen an Weihnachten oder am Black Friday abgefangen werden können.

Lösungsansatz Bei der Thematik Drittlandübertragung von personenbezogenen Daten handelt es um ein technisch nicht lösbares Problem. Es muss durch geeignete Wahl des Dienstleisters oder des Standorts und durch entsprechende organisatorische Maßnahmen wie dem Nachweis geeigneter Garantien gelöst werden. Nichtsdestotrotz muss es bei der Implementierung des Data Lakes beachtet werden.

5.9 Fazit

Beim Design eines Data Lakes, in dem personenbezogene Daten verarbeitet werden, müssen datenschutzrechtliche Regelungen beachtet werden. Diese stellen den Bereich der Big Data Datenverarbeitung im Allgemeinen und den Data Lake im Besonderen vor große Herausforderungen. Tabelle 5.1 zeigt die in dieser Arbeit herausgearbeiteten Probleme, die Pflichten, auf denen diese beruhen und die herausgearbeiteten Lösungsansätze.

Die herausgearbeiteten Lösungsansätze lassen sich grundsätzlich in folgende Kategorien unterteilen:

Manuell: Die Umsetzung vieler Pflichten, insbesondere derjenigen, die eine rechtliche Würdigung oder Abwägung benötigen, kann nicht automatisiert werden. Der Data Lake kann die Umsetzung solcher Pflichten nur durch Bereitstellung zusätzlicher Informationen z.B. aus den Metadaten unterstützen. Die Umsetzung muss außerhalb des Data Lakes erfolgen.

Metadaten und Governance-Prozesse: Die Umsetzung der Pflichten, deren Erfüllung primär von Wissen über Daten oder Datenstrukturen abhängig ist, kann durch den Data Lake unterstützt werden. Insbesondere kann dies durch die Speicherung des benötigten Wissens in z.B. Metadaten erfolgen. Governance-Prozesse können die Umsetzung solcher Pflichten unterstützen oder sogar übernehmen. Zudem können die Governance-Prozesse die Einhaltung von Nutzungseinschränkungen wie der Zweckbindung unterstützen.

Änderung am Data Lake Modell: Einige Pflichten erfordern Anpassungen des Designs und der Idee eines Data Lakes. Dazu gehören zum einen die Anpassungen an den Data Lake Zonen und zum anderen betrifft dies die explorativen Analysen. Diese sind mit personenbezogenen Daten nur in sehr begrenztem Umfang im Rahmen der gesetzlichen Vorgaben möglich.

Im folgenden Kapitel werden die Lösungsansätze *Änderung am Data Lake Modell* und *Metadaten und Governance-Prozesse* näher betrachtet und jeweils ein Konzept dafür erarbeitet.

Tabelle 5.1: Zusammenfassung Probleme & Lösungsansätze

Problem	**Pflichten**	**Lösungsansatz**
Planung rechtskonformer Verarbeitung	Grundsatz der Rechtmäßigkeit Wirksame Einwilligung einholen Interessenabwägung Besondere Kategorien personenbezogener Daten	Manuell
Sicherstellung rechtskonformer Verarbeitung	Grundsatz der Rechtmäßigkeit Wirksame Einwilligung nachweisen Besondere Kategorien personenbezogener Daten	Metadaten und Governance-Prozesse Permission Management System und Governance-Prozesse
Explorative Datenanalyse	Grundsatz der Fairness Grundsatz der Transparenz Grundsatz der Zweckbindung Grundsatz der Datenminimierung	Nur sehr eingeschränkt möglich, ggf. im Einzelfall statistische Zwecke, Zweckänderung
Sicherstellung der Zweckbindung	Grundsatz der Zweckbindung	Metadaten und Governance-Prozesse
Änderung von Daten	Grundsatz der Datenrichtigkeit Grundsatz der Speicherbegrenzung Recht auf Berichtigung Recht auf Löschung	Änderung am Data Lake Modell Geeignete Technologiewahl
Löschung von Daten	Grundsatz der Datenrichtigkeit Grundsatz der Speicherbegrenzung Recht auf Löschung	Änderung am Data Lake Modell Geeignete Technologiewahl
Proaktive Löschung von Daten	Grundsatz der Datenrichtigkeit Grundsatz der Speicherbegrenzung Recht auf Löschung	Löschkonzept, Umsetzung: Metadaten und Governance-Prozesse Änderung am Data Lake Modell Geeignete Technologiewahl
Kontextabhängige Richtigkeit	Grundsatz der Datenrichtigkeit Recht auf Berichtigung	Manuell
Sicherstellung der Korrektheit von Daten	Grundsatz der Datenrichtigkeit	Keine tiefergehende Nachforschungspflicht
Identifikation der Daten über einen Betroffenen	Betroffenenrechte	Metadaten und Governance-Prozesse

Tabelle 5.1: Zusammenfassung Probleme & Lösungsansätze

Problem	**Pflichten**	**Lösungsansatz**
Kopie ohne Daten Dritter	Recht auf eine Kopie der Daten	Teilweise manuell: Identifikation der problematischen Daten durch Metadaten und Governance-Prozesse, dann manuelle Einzelfallprüfung
Sperrung von Daten	Recht auf Einschränkung	Metadaten und Governance-Prozesse
Identifikation von Datenempfängern	Mitteilungspflicht an Empfänger Informationsrecht über Empfänger	Metadaten und Governance-Prozesse
Mitteilung an Datenempfänger	Mitteilungspflicht an Empfänger	Einzelfalldefinition des Übermittlungswegs, danach: automatisierte Übermittlung durch Governance-Prozesse
Reichweite des Portabilitätsanspruchs	Recht auf Datenportabilität	Unklar, manuelle Definition erforderlich Umsetzung durch Metadaten und Governance-Prozesse
Austauschformat	Recht auf Datenportabilität	Allgemeingebräuchliche Formate, ggf. textbasiert
Portabilität ohne Daten Dritter	Recht auf Datenportabilität	Teilweise manuell: Identifikation der problematischen Daten durch Metadaten und Governance-Prozesse, dann manuelle Einzelfallprüfung
Weitergabe an anderen Verantwortlichen	Recht auf Datenportabilität	Manuelle Einzelfalllösung
Prüfung Widerspruch	Recht auf Widerspruch	Marketingwidersprüche automatisiert, sonst manuell
Umsetzung Widerspruch	Recht auf Widerspruch	Metadaten und Governance-Prozesse
Auswahl TOMs (Regelmäßige) Überprüfung TOMs	Grundsatz der Integrität & Vertraulichkeit Datenschutz durch Technikgestaltung Datenschutzfreundliche Voreinstellungen Sicherheit der Verarbeitung	Detaillierte Kenntnisse über (geplante) Verarbeitungen Detaillierte Kenntnisse über gespeicherte Daten Risikoabwägung Kontinuierlicher Verbesserungsprozess
Datenschutz in Mitteln der Verarbeitung	Datenschutz durch Technikgestaltung	Einhaltung der Grundsätze und deren Ausgestaltung (Pflichten) auch durch das Speichersystem Data Lake
Datensicherheit eines zentralen Speichersystems	Grundsatz der Integrität & Vertraulichkeit Sicherheit der Verarbeitung	Wahl geeigneter TOMs
DSFA für Data-Lake-Verarbeitungen	Datenschutzfolgenabschätzung	Manuell

Tabelle 5.1: Zusammenfassung Probleme & Lösungsansätze

Problem	**Pflichten**	**Lösungsansatz**
Erfüllung Nachweis- und Dokumentationspflichten	Rechenschaftspflicht Grundsätze Nachweis geeigneter IT-Sicherheit Dokumentation Einwilligung Verarbeitungsverzeichnis …	Erfüllung teilweise durch Governance-Strukturen und Dokumentation/Metadaten aus dem Data Lake (z.B. Provenienz) erleichtert
Doppelte Dokumentation in Metadaten und Verarbeitungsverzeichnis	Verarbeitungsverzeichnis	Schnittstelle zwischen den Systemen

6 Konzept

Auch wenn sich viele der in Kapitel 5 herausgearbeiteten Anforderungen der DSGVO nicht technisch in einem Data Lake umsetzen lassen, so können zumindest einige durch Anpassungen des Data Lakes-Designs, konkreter der Speicherzonen, oder durch Governance in Verbindung mit z.B. Metadaten oder einem Permissionmanagementsystems umgesetzt werden. Im Folgenden wird zunächst das Zonenmodell aus Abschnitt 3.1.2 entsprechend der im Kapitel zuvor erarbeiteten Lösungsansätze angepasst (Abschnitt 6.1). Zudem wird aus den in den Lösungsansätzen vorgeschlagenen Metadaten ein Metadaten-Modell erstellt (Abschnitt 6.2).

6.1 Zonenmodell

Das Zonenmodell aus Abschnitt 3.1.2 stellt Anforderungen an die Zonen Raw-Zone, Trusted-Zone und Refined-Zone, die im Data Lake mit personenbezogenen Daten nicht erfüllbar sind.

Im Zonenmodell sind die Daten in der Raw-Zone unveränderlich und auch in den beiden anderen Zonen ist grundsätzlich keine Datenveränderung vorgesehen. Dies widerspricht, wie in Kapitel 5 diskutiert, dem Grundsatz der Datenrichtigkeit (Abschnitt 5.1.6) und dem Betroffenenrecht auf Berichtigung (Abschnitt 5.3.2). Personenbezogene Daten im Data Lake müssen veränderbar sein.

Zudem ist in diesen Zonen auch keine Datenlöschung vorgesehen, Ziel ist primär die Sammlung möglichst vieler Daten, um daraus später einen Mehrwert zu generieren. Die Speicherung personenbezogener Daten erfordert jedoch neben der Veränderbarkeit auch die Löschbarkeit dieser Daten. Dies ergibt sich, wie in Kapitel 5 diskutiert, aus den Grundsätzen der Datenrichtigkeit (Abschnitt 5.1.6) und der Speicherbegrenzung (Abschnitt 5.1.7) sowie aus dem Betroffenenrecht auf Löschung (Abschnitt 5.3.3). Personenbezogene Daten im Data Lake müssen löschbar sein. Zudem werden Governance-Prozesse benötigt, die eine Löschung sicherstellen, sobald diese Daten nicht mehr benötigt werden. Das Ziel, möglichst viele Daten für eine mögliche zukünftige Wertschöpfung zu kreieren muss hinter die gesetzlichen Anforderungen zur Verarbeitung personenbezogener Daten zurücktreten.

Im ursprünglichen Zonenmodell sind dem Data Scientist keine Grenzen gesetzt, welche Daten aus den persistenten Zonen herangezogen werden sollen, um diese in der Sandbox zu analysieren. Im Data Lake mit personenbezogenen Daten ist dies nicht möglich. Stattdessen wird für diese Verarbeitung eine Rechtsgrundlage benötigt und der Zweck der Verarbeitung muss feststehen. Nur Daten, die dafür genutzt werden dürfen, dürfen in die Sandbox des Data Scientists gelangen. Es werden Governance-Prozesse benötigt, die dies sicherstellen. Abbildung 6.1 zeigt das Zonenmodell für einen Data Lake mit personenbezogenen Daten.

Wenn die strengeren Governance-Anforderungen und die geänderten Anforderungen an die Datenintegrität nur für personenbezogene Daten im Data Lake gelten sollen und nicht für die anderen Daten, so müssen personenbezogene Daten in einem getrennten Bereich im Data Lake abgelegt werden.

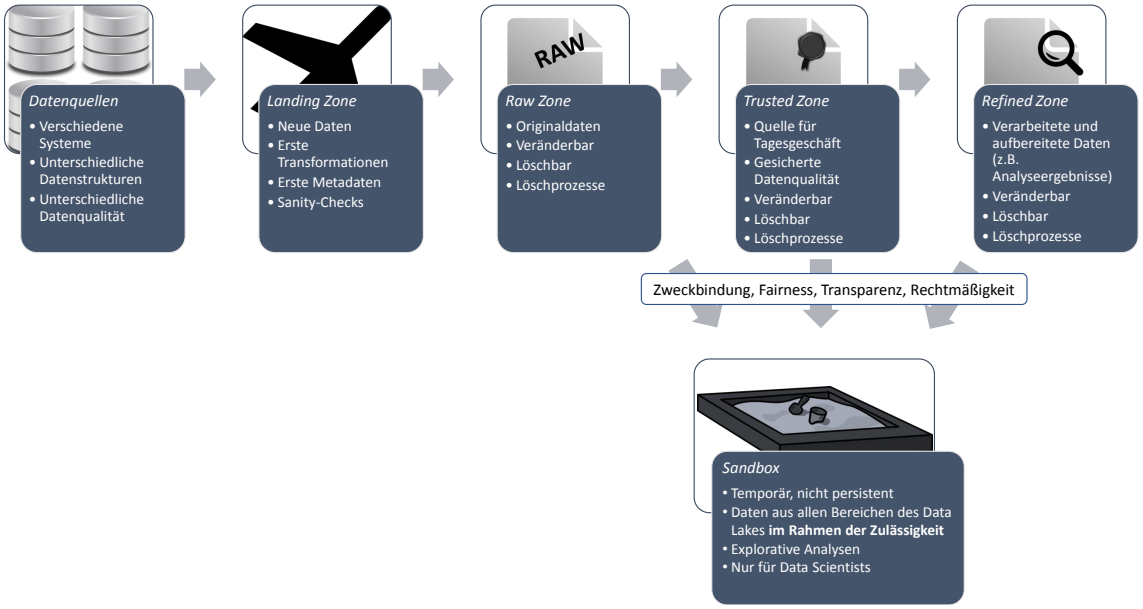

Abbildung 6.1: Zonenmodell für den Data Lake mit personenbezogenen Daten

Nachdem die Daten in der Landing Zone zwischengespeichert wurden, müssen die personenbezogenen Daten in die Raw Zone des personenbezogenen Speicherbereichs transferiert werden, während die nicht personenbezogenen Daten in die Raw Zone des nicht personenbezogenen Speicherbereichs transferiert werden können.

Die beiden Bereiche des Data Lakes sind jedoch nicht hermetisch voneinander abgetrennt. Auch bei Analysen im nicht personenbezogenen Bereich des Data Lakes muss beachtet werden, dass nicht personenbezogene Daten durch Kombination verschiedener Datenquellen möglicherweise personenbezogen werden. Diese Daten müssen im personenbezogenen Bereich des Data Lakes abgelegt werden. Ebenso ist es möglich, dass anonymisierte personenbezogene Daten oder statistische Analyseergebnisse nicht mehr personenbezogen sind. Diese Daten müssen dann auch nicht mehr im personenbezogenen Bereich abgelegt werden.

6.2 Metadaten-Modell

Die in Kapitel 5 erarbeiteten Lösungsansätze, die auf Metadaten basieren, unterstützen den Verantwortlichen darin, folgende Pflichten aus der DSGVO zu erfüllen:

- Grundsatz der Rechtmäßigkeit (Rechtsgrundlage, Einwilligung, besondere Kategorien personenbezogener Daten)
- Grundsatz der Zweckbindung
- Grundsatz der Datenrichtigkeit
- Grundsatz der Speicherbegrenzung
- Recht auf Auskunft

- Recht auf Berichtigung (Datenidentifikation)
- Recht auf Löschung
- Recht auf Einschränkung der Verarbeitung
- Mitteilungspflicht im Fall einer Berichtigung, Löschung oder Einschränkung
- Recht auf Datenportabilität
- Recht auf Widerspruch

Indirekt dienen die Metadaten, insbesondere im Verbund mit den Governance-Strukturen sowie Eigenschaften eines Data Lakes wie der Dokumentation der Datenprovenienz, auch der Einhaltung der Nachweis- und Dokumentationspflichten.

Die benötigten Metadaten lassen sich in zwei Kategorien unterteilen: die Metadaten, die zur Verarbeitung gehören und die Metadaten, die tatsächlich dem jeweiligen personenbezogenen Datum zugeordnet werden.

Die Unterteilung ergibt sich aus den zwei Sichtweisen, die hier aufeinander stoßen. Die erste Sichtweise ergibt sich aus der DSGVO, die mit Verarbeitungen arbeitet und daher nahezu alle Pflichten an die Verarbeitung koppelt. Die Frage, für welchen Zweck welche Kategorie von Daten verarbeitet werden darf, wird auf dieser Ebene geklärt. Die zweite Sichtweise ergibt sich aus der praktischen Umsetzung. Hier erfolgt der Zugriff auf die Daten selbst, weshalb immer im Einzelfall überprüft werden muss, welcher Kategorie ein spezifisches Datum angehört und ob es daher auch tatsächlich für diesen Zweck verwendet werden darf.

Den Metadaten zur Verarbeitung sind zuerst die Informationen über die Verarbeitung selbst, wie Name und Beschreibung, zuzuordnen. Zudem sind die zugehörigen Verarbeitungszwecke und Rechtsgrundlagen, sowie die jeweils verarbeiteten Datenkategorien mit Löschfristen, bzw. -kriterien der Verarbeitung zuzuordnen. Soweit Einwilligungen erforderlich sind oder gegen bestimmte Verarbeitungstätigkeiten Widerspruch eingelegt wurde, so sind auch diese Teil der Metadaten zur Verarbeitung, da Einwilligungen und Widersprüche nicht einzelne Daten betreffen, sondern die Verarbeitung von Daten zu spezifischen Zwecken.

Den Metadaten zu den personenbezogenen Daten sind Informationen über das Datum selbst, wie Speicherort, die Empfänger des Datums und die jeweilige Datenkategorie, zugeordnet. Zudem wird auf die Verarbeitungen referenziert, die dieses Datum nutzen können. Ebenso gehört die Markierung der eingeschränkten Verarbeitung (Sperrung) und die Markierung, ob das Datum dem Portabilitätsanspruch unterfällt, zu diesen Metadaten.

Zunächst müssen die einzelnen *Verarbeitungen* verwaltet werden. Jede Verarbeitung hat eine Verarbeitungs-ID (VID). Zur besseren Verständlichkeit für den Nutzer wird der Name der Verarbeitung (VName) und eine Beschreibung (VBeschreibung) hinterlegt.

Die gesetzlich vorgesehenen *Rechtsgrundlagen* werden mit einer ID (RID) und der textuellen Bezeichnung (RName) gespeichert. Optional kann zum besseren Verständnis der Nutzungszwecke der Rechtsgrundlage eine Beschreibung (RBeschreibung) als weiteres Attribut modelliert werden.

Da eine Verarbeitung mehreren Zwecken dienen kann, werden die *Verarbeitungszwecke* getrennt von der Verarbeitung verwaltet. Jeder Zweck hat eine Zweck-ID (ZID), sowie den Zweck selbst (ZName) und eine zugehörige Beschreibung (ZBeschreibung). Über eine 1:n-Beziehung wird die

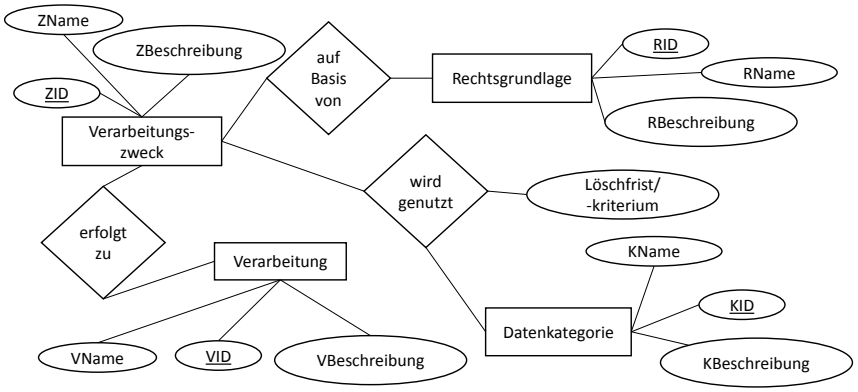

Abbildung 6.2: Metadaten Verarbeitung als Entity-Relationship-Diagramm

Zuordnung der jeweiligen Zwecke zur zugehörigen Verarbeitung abgebildet. Die Zuordnung der für den Zweck der Verarbeitung benötigten Datenkategorien wird mit einer n:m-Beziehung modelliert, wobei jede Zweck-Kategorie-Zuordnung noch ein Attribut, die Löschfrist, bzw. -kriterium erhält. Abbildung 6.2 zeigt den zur Verarbeitung gehörenden Teil des Metadaten-Modells als Entity-Relationship-Diagramm.

Die möglichen *Datenkategorien*, die von Verarbeitungen genutzt werden können, bzw. denen die personenbezogenen Daten zugeordnet werden müssen, werden mit einer ID (KID), dem Namen (KName) und einer Beschreibung (KBeschreibung) der Kategorie verwaltet. Da eine Datenkategorie zu verschiedenen Verarbeitungszwecken genutzt werden kann, besteht zwischen der Datenkategorie und dem Verarbeitungszweck eine n:m-Beziehung. Diese Beziehung bekommt zwei Attribute zugeordnet, eine Löschfrist und ein Löschkriterium, wobei nur einem von beiden ein Wert zugeordnet sein darf.

Der *Betroffene* wird in aller Regel bereits in vorhandenen Systemen oder Datenbanken hinterlegt sein, z.B. Kundendatenbanken oder Personaldatenbanken. Die in diesem Metadaten-Modell benötigte Betroffenen-ID (BID) zur Identifikation der betroffenen Person wird meist eine aus diesen Systemen stammende, eindeutige ID sein. Daher muss eine entsprechende Verknüpfung zwischen der BID, den betreffenden Systemen und den jeweiligen internen IDs, wie z.B. Kundennummer oder Personalnummer, sichergestellt werden.

Sofern *Einwilligungen* nicht in einem Permission-Management-System verwaltet werden, kann dies auch in den Metadaten erfolgen. Eine Einwilligung wird von einem Betroffenen für einen bestimmten Zweck erteilt, die Einwilligung wird daher als Beziehung „willigt ein“ zwischen dem Betroffenen und dem Verarbeitungszweck abgebildet. Um sicherstellen und dokumentieren zu können, dass eine wirksame Einwilligung vorliegt, muss zudem der Einwilligungstext, der Zeitpunkt sowie die Art und der Ort der Einwilligung (z.B. Double-Opt-In mit zugehöriger URL) dokumentiert werden. Da eine Einwilligung jederzeit widerrufen werden kann, muss das Vorliegen einer Einwilligung vor jeder Verarbeitung auf Grundlage einer Verarbeitung überprüft werden.

Analog zu Einwilligungen sollten auch *Widersprüche* in einem Permission-Management-System verwaltet werden. Auch der Widerspruch einer Person richtet sich gegen die Verarbeitung von Daten zu einem bestimmten Zweck. Daher wird der Widerspruch als Beziehung „widerspricht“ zwischen Betroffenem und Verarbeitungszweck modelliert. Weitere Attribute sind nicht erforderlich, sind aber denkbar, wie z.B. der Verweis auf die Vorgangsnummer des Widerspruchs. Ob ein Widerspruch

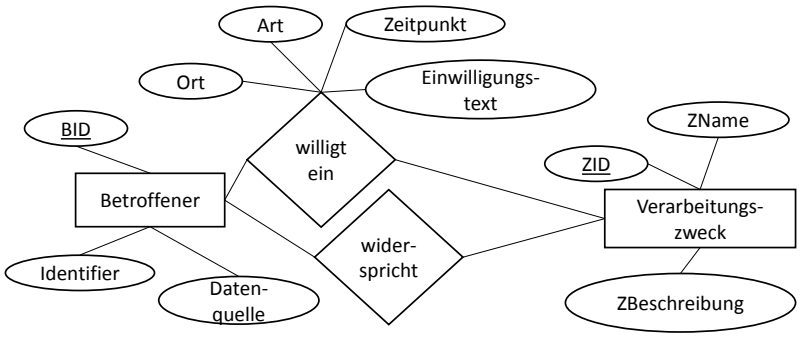

Abbildung 6.3: Modellierung Einwilligungen und Widersprüche als Entity-Relationship-Diagramm

gegen eine Verarbeitung eingelegt wurde, muss überprüft werden, bevor eine Verarbeitung auf der Grundlage öffentlichen Interesses oder berechtigten Interesses erfolgt. Abbildung 6.3 zeigt die Modellierung der Einwilligungen und der Widersprüche als Entity-Relationship-Diagramm.

In den Metadaten zum *personenbezogenen Datum* selbst wird zuallererst der Speicherort des tatsächlichen Datums gespeichert. Dabei kann es sich beispielsweise um einen Dateipfad oder um einen Verweis auf ein Datenfeld in der Kundendatenbank handeln. Zudem wird dem Datum eine Data-ID (DID) zugeordnet.

Um den Löschanforderungen genüge zu tun, muss ein Löschdatum festgelegt werden. Dieses ergibt sich aus den beim Verarbeitungszweck gespeicherten Löschfristen. Da ein Datum zu unterschiedlichen Verarbeitungszwecken und somit auch auf Basis unterschiedlicher Löschfristen verarbeitet werden darf, müssen Governance-Prozesse sicherstellen, dass dieses Löschdatum nur genau dann verändert wird, die durchgeführte Verarbeitung die Löschfrist dieses Datums tatsächlich verlängert. Ist keinerlei Löschfrist für das Datum festgelegt, sondern nur Löschkriterien, so wird das aktuelle Datum eingetragen. Wenn die Löschfrist abgelaufen ist, müssen die Governance-Prozesse vor der Löschung zunächst alle anwendbaren Löschkriterien überprüfen. Nur wenn kein Kriterium mehr die Speicherung verlangt, wird das personenbezogene Datum gelöscht.

Wie in Abschnitt 5.3.6 diskutiert, sollte bei der Erhebung eines Datums entschieden werden, ob es dem Portabilitätsanspruch unterfällt. Ein entsprechendes Flag sollte in den Metadaten gespeichert werden. Ebenso wird ein Flag benötigt, ob der Betroffene eine Einschränkung der Verarbeitung, also im Wesentlichen eine Sperrung des Datums zur weiteren Verarbeitung, erwirkt hat.

Da ein personenbezogenes Datum nicht zwangsläufig nur einer Person zuzuordnen ist und über eine Person zumeist mehr als ein Datum gespeichert ist, wird die Zuordnung zwischen Datum und Betroffenem als n:m-Beziehung abgebildet.

Um Datenempfänger über Berichtigungen, Löschungen oder Einschränkungen der Verarbeitung informieren zu können (Mitteilungspflicht, Abschnitt 5.3.5), müssen zuletzt noch die *Empfänger* verwaltet werden, an die das jeweilige personenbezogene Datum weitergegeben wurde. Diese erhalten jeweils eine Empfänger-ID (EID) und einen Namen. Zudem wird der Übertragungsweg benötigt, um die Empfänger der Daten tatsächlich benachrichtigen zu können. Welche Daten an welche Empfänger weitergegeben wurden, wird durch eine n:m-Beziehung modelliert. Abbildung 6.4 zeigt die Modellierung der dem einzelnen personenbezogenen Datum zugeordneten Metadaten.

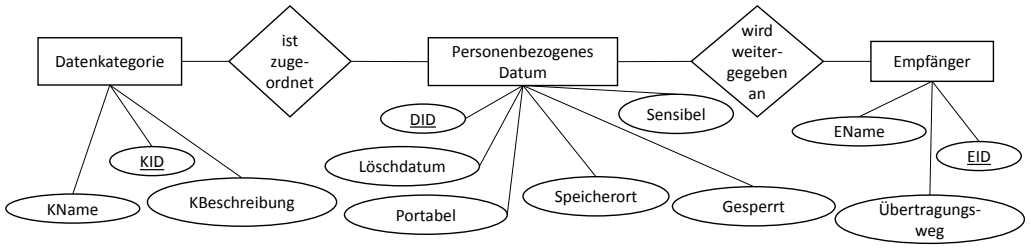

Abbildung 6.4: Metadaten personenbezogener Daten als Entity-Relationship-Diagramm

Zusammengefasst lässt sich das gesamte Metadaten-Modell wie in Abbildung 6.5 als Entity-Relationship-Diagramm darstellen.

Für die Implementierung in der Praxis sollte beachtet werden, dass die Modellierung der Verarbeitung mit Verarbeitungszweck und Datenkategorie, sowie die Empfängerliste auch Teil des zu führenden Verarbeitungsverzeichnisses sind (vgl. Abschnitt 5.7). Um doppelte und damit potenziell inkonsistente Dokumentation zu vermeiden, sollten diese Daten über Schnittstelle aus dem dafür genutzten Datenschutzmanagementsystem stammen.

Abhängig von der konkreten Implementierung des Data Lakes, den gespeicherten Daten und den genutzten Systemen (z.B. Permission-Management-Systeme) kann eine Anpassung des Metadaten-Modells erforderlich sein.

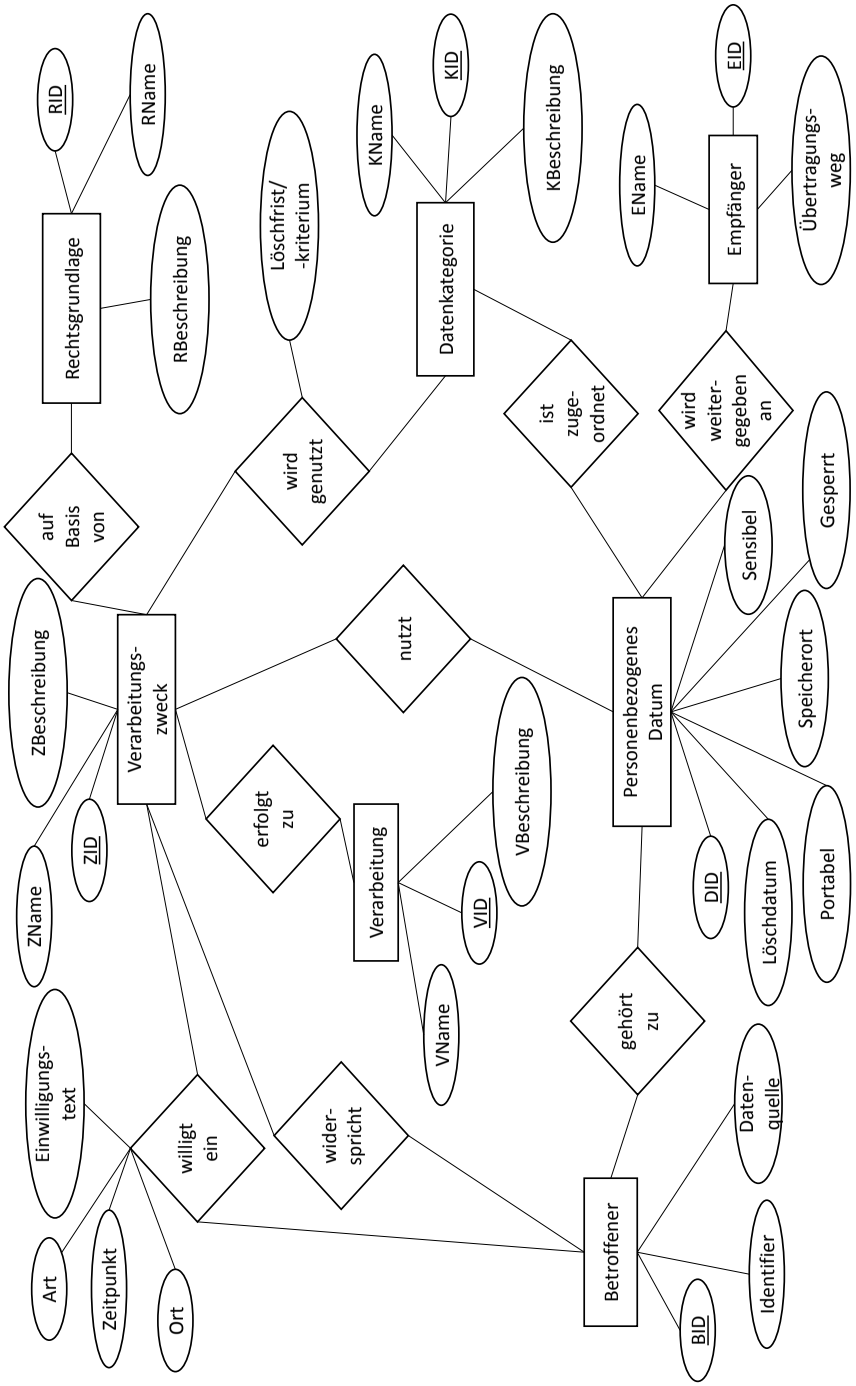

Abbildung 6.5: Metadaten-Modell als Entity-Relationship-Diagramm

7 Implementierung

Die Implementierung des veränderten Zonenmodells würde die vollständige Implementierung des Speicherbereichs eines Data Lakes erfordern. Da dies im Rahmen dieser Arbeit nicht umsetzbar gewesen wäre, wird in diesem Kapitel nur die prototypische Implementierung des in Abschnitt 6.2 erarbeiteten Metadaten-Modells und von Governance-Prozessen anhand der Use Cases aus Kapitel 4 beschrieben. Zunächst werden die zur Implementierung genutzten Systeme (Abschnitt 7.1) und die logische Modellierung des Metadaten-Modells vorgestellt (Abschnitt 7.2). Danach wird die Generierung der Testdaten (Abschnitt 7.3) erklärt und schließlich die Implementierung der Governance-Prozesse (Abschnitt 7.4) vorgestellt.

7.1 Systeme

Das Metadaten-Modell wird für den Prototypen in einem relationalen Datenbankmanagementsystem (RDBMS) implementiert. Dazu wird MariaDB[Mar18b], ein Fork von MySQL, genutzt. MariaDB ist ein Open Source RDBMS wird von den ursprünglichen MySQL Entwicklern entwickelt[Mar18a]. Es ist eines der verbreitetsten RDBMS und hat auf vielen Linux-Distributionen und bei vielen Webprojekten MySQL als Standarddatenbanksystem abgelöst. Einer der prominentesten Nutzer ist Wikipedia.

Der Einfachheit halber wird für den Prototypen auf semi-strukturierte und unstrukturierte Daten verzichtet. Dadurch können die Daten in einem RDBMS gespeichert werden. Dies stellt keine Einschränkung dar und dient nur der Vereinfachung des Prototyps, da die Metadaten des Metadaten-Modells so modelliert sind, dass sie unabhängig vom tatsächlichen Datenspeicher gespeichert werden. Die Metadaten enthalten stattdessen nur einen Link zu den tatsächlichen Daten (vgl. Abschnitt 6.2). Es kann daher jederzeit ein anderes für den jeweiligen Anwendungszweck und die anfallenden Daten geeignetes Speichersystem genutzt werden. Es muss nur der tatsächliche Datenzugriff entsprechend angepasst werden.

Es wird zudem die Laufzeitumgebung Python 3.6 vorausgesetzt, da der Zugriff auf die Daten und Metadaten sowie die Governance-Prozesse in Python implementiert wurde.

7.2 Realisierung des Metadaten-Modells

Zunächst wird das konzeptionelle Datenmodell aus Abschnitt 6.2 in ein logisches Datenmodell transformiert. Die n:m-Beziehungen werden jeweils in einer eigenen Tabelle abgebildet. Die 1:n-Beziehungen werden durch Fremdschlüssel realisiert. Abbildung 7.1 zeigt das vollständige logische Datenmodell des Metadaten-Modells.

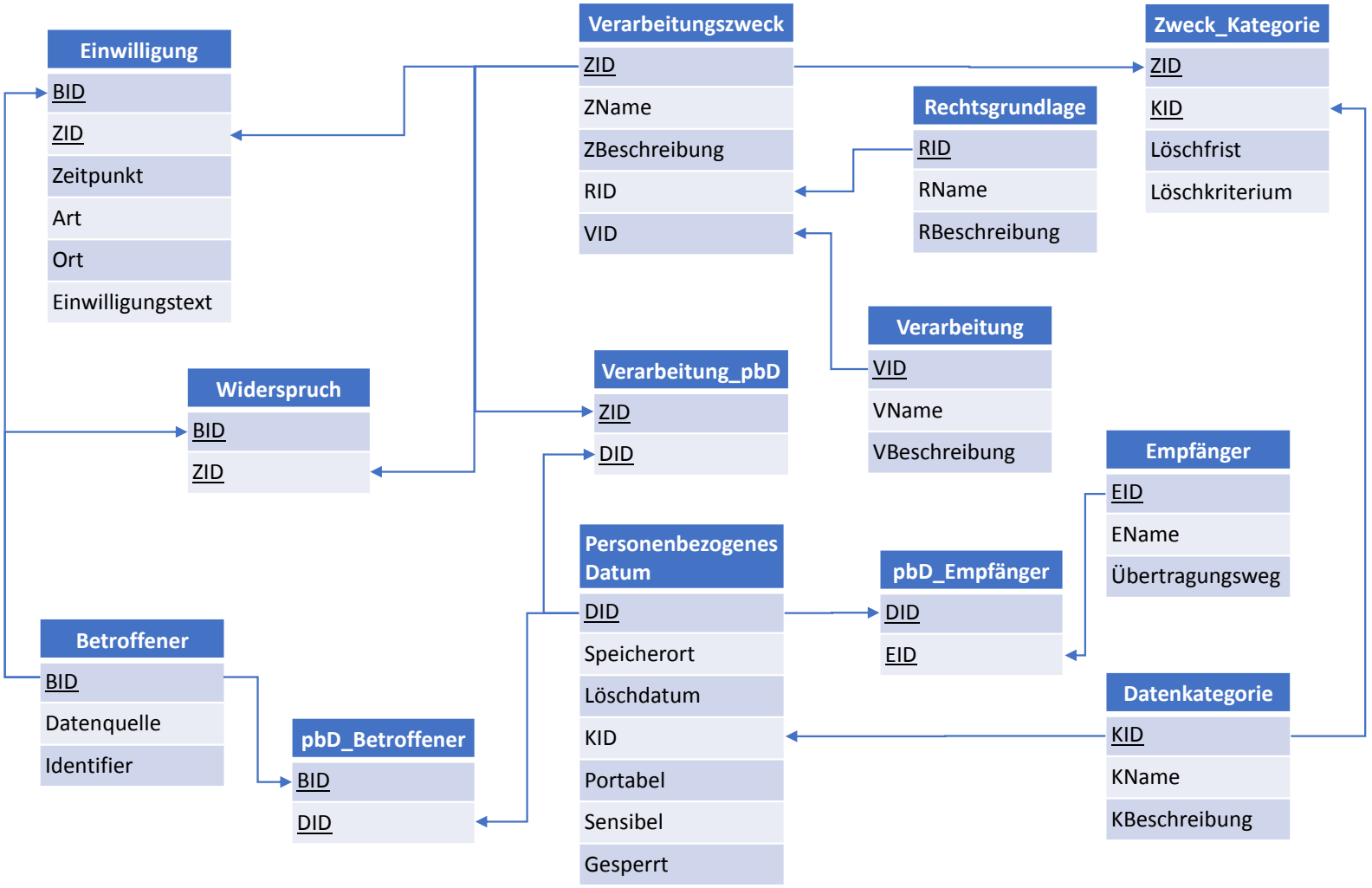

Abbildung 7.1: Logische Modellierung des Metadaten-Modells

Listing 7.1 SQL-Befehl zur Erstellung der Verarbeitungszweck-Tabelle

```
CREATE TABLE IF NOT EXISTS PURPOSE (
  PURPID INTEGER NOT NULL AUTO_INCREMENT PRIMARY KEY,
  PURPNAME VARCHAR(50),
  PURPDESCRIPTION VARCHAR(1000),
  LID INTEGER NOT NULL references LEGAL_GROUND(LID),
  PROCID INTEGER NOT NULL references PROCESSING(PROCID)
);
```

Listing 7.2 Datenbankauszug der Datenkategorien

```
+-----+--------------+-------------------------------------------------------------+
| CID | CNAME        | CDESCRIPTION                                                |
+-----+--------------+-------------------------------------------------------------+
|   1 | Name         | Name des Betroffenen (z.B. Kunde oder Mitarbeiter)          |
|   2 | Kontaktdaten | Kontaktdaten des Betroffenen (z.B. Kunde oder Mitarbeiter)  |
|   3 | Geburtsdatum | Geburtsdatum des Betroffenen (z.B. Kunde oder Mitarbeiter)  |
[...]
```

Listing 7.1 zeigt exemplarisch den SQL-Befehl zur Erzeugung der Verarbeitungszweck-Tabelle (engl. purpose of processing, Kurzform in der Implementierung: purpose) des Metadaten-Modells. Die Erzeugung aller anderen Tabellen erfolgt analog. Alle IDs werden nicht manuell gesetzt, sondern über `AUTO_INCREMENT` von der Datenbank selbst aufsteigend vergeben.

7.3 Testdaten

Verarbeitungen Die Test-Metadaten für die Rechtsgrundlagen, Datenkategorien, Verarbeitungen und Verarbeitungszwecke werden manuell generiert. Als Rechtsgrundlagen wurden die sieben im Gesetz vorgesehenen Rechtsgrundlagen (vgl. Abschnitt 5.2) hinterlegt. Listing 7.2 zeigt einen Auszug aus der Datenkategorie-Tabelle. Für diesen Prototypen wurden zwei Verarbeitungen, Kundenverwaltung und Newsletter, mit je zwei Verarbeitungszwecken implementiert.

Die Rechtsgrundlagen und Datenkategorien können in einer CSV Datei abgelegt und dann vom System importiert werden. Die Metadaten zu den Verarbeitungen und Verarbeitungszwecke können aus einer JSON-Datei importiert werden.

Betroffenendaten Die Betroffenen-Tabelle wird nicht in der im Modell vorgeschlagenen Art modelliert, da es in der prototypischen Implementierung keine vorhandenen Kundendatenbanken oder Personaldatenbanken gibt. Die vorgeschlagene Tabelle, die die vorhandenen Systeme und das Metadaten-System miteinander verknüpfen sollte, wird daher nicht benötigt.

Stattdessen wird eine Betroffenen-Datenbank mit Daten benötigt. Für den hier isoliert betrachteten Data Lake wird daher eine denkbar einfache Personenverwaltung genutzt. Jeder Betroffene wird mit einer ID, dem Vor- und Nachnamen, sowie einer E-Mailadresse modelliert. Abbildung 7.2 zeigt das logische Modell der implementierten Betroffenen-Tabelle.

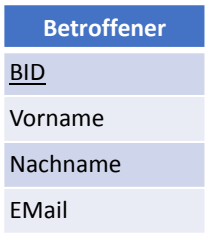

Abbildung 7.2: Logisches Modell der implementierten Betroffenen-Tabelle

Listing 7.3 Auszug der generierten Betroffenen-Datenbank

```
+-----+-------------+--------------+---------------------------------+
| SID | FIRST_NAME  | LAST_NAME    | EMAIL                           |
+-----+-------------+--------------+---------------------------------+
|   1 | Marni       | Nadia        | Marni.Nadia@example.com         |
|   2 | Vannie      | Nay          | Vannie.Nay@example.com          |
|   3 | Corene      | Vivien       | Corene.Vivien@example.com       |
|   4 | Feodora     | Zane         | Feodora.Zane@example.com        |
|   5 | Cristin     | Themis       | Cristin.Themis@example.com      |
|   6 | Bevvy       | Jaf          | Bevvy.Jaf@example.com           |
|   7 | Salli       | Rosemary     | Salli.Rosemary@example.com      |
|   8 | Suzanna     | Bremble      | Suzanna.Bremble@example.com     |
|   9 | Dianna      | Pantheas     | Dianna.Pantheas@example.com     |
|  10 | Mara        | Durand       | Mara.Durand@example.com         |
|  11 | Imogene     | Esmond       | Imogene.Esmond@example.com      |
[...]
```

Die benötigten Betroffenendaten werden zufällig aus Namenslisten[Dom16] generiert. Zudem können Betroffene manuell eingepflegt werden. Listing 7.3 zeigt einen Auszug aus der generierten Betroffenen-Datenbank.

Die Zuordnung, welche Daten für welche Verarbeitung genutzt werden dürfen, erfolgt für den Prototypen zufällig durch das System. Der zufälligen Zuordnung durch das System liegt nur die Annahme zugrunde, dass in dieser Betroffenen-Tabelle keine Mitarbeiter, sondern nur Kunden abgelegt sind. Etwa 75% davon sollen Bestandskunden sein, während die verbleibenden 25% potentielle Neukunden sind, die sich z.B. für einen Newsletter angemeldet haben.

Zudem werden für Verarbeitungen auf Basis berechtigten Interesses etwa 20% Widersprüche und bei Verarbeitungen auf Basis einer Einwilligung 20% fehlende Einwilligungen ins System eingepflegt. So kann die Funktionalität der Zugriffsüberprüfung demonstriert werden.

7.4 Governance-Prozesse

Governance-Prozesse unterstützen den Nutzer bei der Einhaltung der Anforderungen der DSGVO, indem sie bspw. den Datenzugriff überwachen und dabei die Einhaltung der Grundsätze der Zweckbindung und Rechtmäßigkeit überprüfen. Durch die Überwachung der Löschfristen bzw. -kriterien wird zudem die Einhaltung des Grundsatzes der Speicherbegrenzung und des Rechts auf Löschung

unterstützt. Zudem ermöglichen die Prozesse die Identifikation des Betroffenen und aller ihn betreffenden Daten, wodurch die Beantwortung von Betroffenenanfragen unterstützt wird. Die implementierten Governance-Prozesse demonstrieren die Zweckmäßigkeit des Metadaten-Modells anhand der Use Cases.

Exemplarisch wird hier der Prozess zum Datenzugriff auf ein spezifisches Datum detailliert vorgestellt. Um Zugriff auf ein personenbezogenes Datum zu erhalten, werden der Speicherort, die geplante Verarbeitung und der Verarbeitungszweck als Parameter benötigt. Um Zugriff auf das Datum zu erhalten, müssen mehrere Prüfungen durchgeführt werden:

Verarbeitung und Verarbeitungszweck Der angegebene Verarbeitungszweck muss zu der angegebenen Verarbeitung passen.

Überprüfung Datenkategorie Die Datenkategorie des Datums muss eine der Datenkategorien sein, die für diesen Verarbeitungszweck genutzt werden dürfen.

Identifikation betroffener Personen Da personenbezogene Daten nicht zwangsläufig nur einer Person zugeordnet sind, müssen alle Personen identifiziert werden, denen das Datum zugeordnet ist.

Besondere Kategorien personenbezogener Daten Es muss überprüft werden, ob es sich um ein Datum der besonderen Kategorien personenbezogener Daten handelt. Falls ja, so muss sichergestellt werden, dass das Datum nur auf der Basis der dafür zulässigen Rechtsgrundlagen verarbeitet wird (vgl. Abschnitt 5.2.3).

Überprüfung Rechtsgrundlage Bei bestimmten Rechtsgrundlagen müssen spezifische Überprüfungen durchgeführt werden, ob eine Verarbeitung zulässig ist. Ist das Datum mehreren Personen zugeordnet, so muss diese Überprüfung alle Betroffenen umfassen:

Einwilligung: Im Falle einer Einwilligung muss überprüft werden, ob die Einwilligung der Betroffenen (noch) vorliegt (vgl. Abschnitt 5.2.1).

Öffentliches Interesse, öffentliche Aufgabe oder berechtigtes Interesse: Im Falle einer Verarbeitung aufgrund öffentlichen Interesses, in Wahrnehmung einer öffentlichen Aufgabe oder auf Basis berechtigten Interesses, muss überprüft werden, ob einer der Betroffenen Widerspruch gegen diese Verarbeitung eingelegt hat (vgl. Abschnitt 5.3.7).

Überprüfung Einschränkung der Verarbeitung (Sperrung) Es muss überprüft werden, ob das Datum zur Verarbeitung gesperrt ist (vgl. Abschnitt 5.3.4).

Schlägt eine der Prüfungen fehl, wird kein Datenzugriff gewährt. Stattdessen wird eine Exception (dt. Ausnahme) ausgelöst. Listing 7.4 zeigt die Überprüfung der Rechtsgrundlage im Falle einer Einwilligung oder berechtigten Interesses. Der erfolgreiche Datenzugriff wird durch einfache Ausgabe der Daten simuliert. Abbildung 7.3 zeigt den Datenzugriff als Ablaufdiagramm.

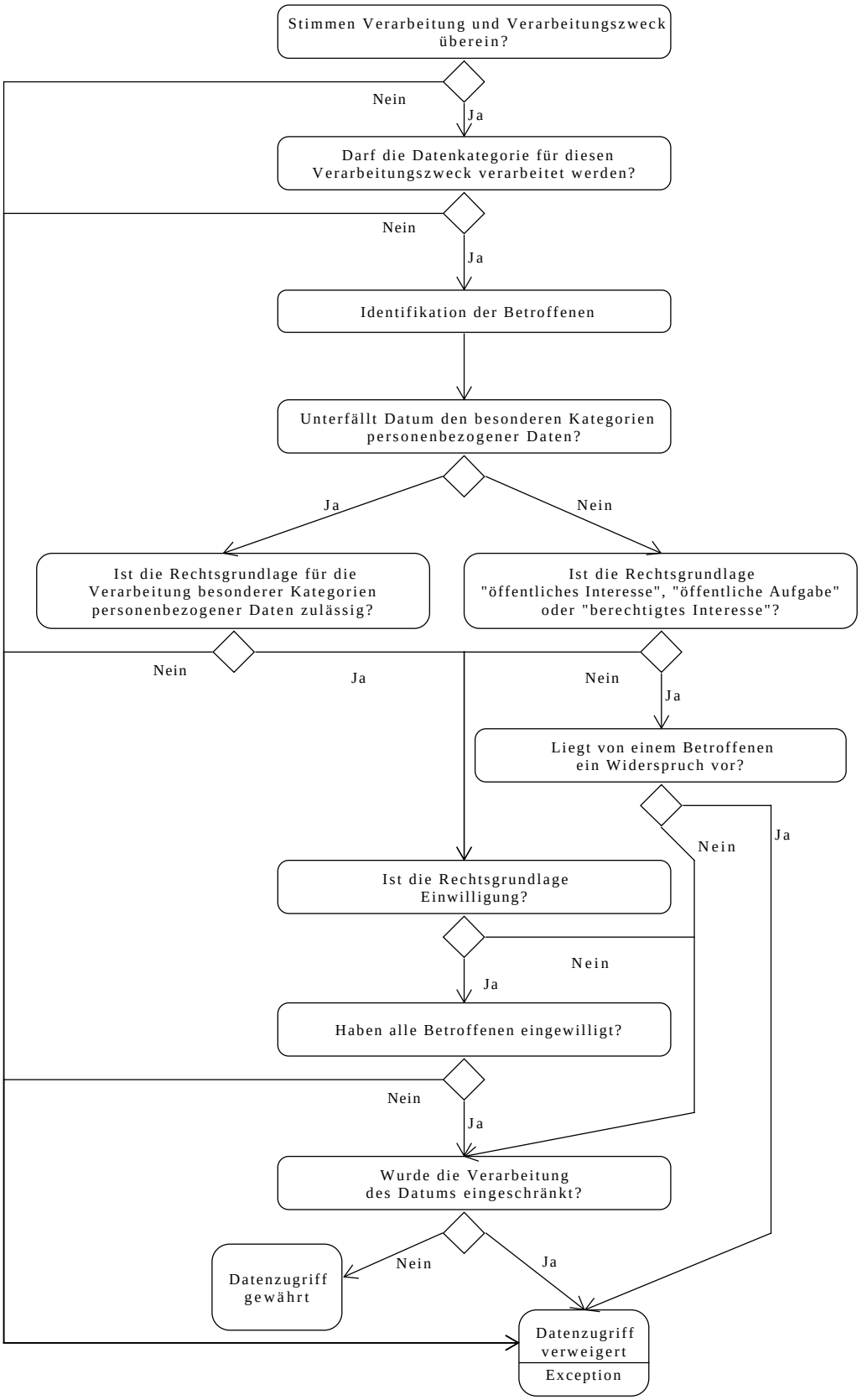

Abbildung 7.3: Ablauf des Datenzugriffs im Data Lake mit personenbezogenen Daten

Listing 7.4 Ausschnitt aus der Zugriffsprüfung für Datenzugriff

```
# Check if consent or objection apllies:
if legalGroundID == getLegalGroundID("Einwilligung"):
    for subjectID in subjectIDs:
        if not consented(subjectID, purpID):
            raise ValueError("Access denied. Data subject didn't consent to processing for the
 purpose " + purpose)
[...]
elif legalGroundID == getLegalGroundID("Berechtigtes Interesse"):
    for subjectID in subjectIDs:
        if objected(subjectID, purpID):
            raise ValueError("Access denied. Data subject objected to processing for the
purpose " + purpose)
```

8 Diskussion der Implementierung

Das Metadaten-Modell wurde entworfen, um den Nutzer des Data Lakes bei der Einhaltung der gesetzlichen Anforderung bei der Verarbeitung personenbezogener Daten zu unterstützen. Es kann den Nutzer bei den in Abschnitt 6.2 genannten Pflichten unterstützen.

Dazu wurde das Metadaten-Modell, mit der in Abschnitt 7.3 diskutierten Abweichung bei der Betroffenen-Tabelle, vollständig in der Datenbank abgebildet. Zudem wurden Governance-Prozesse für die Use Cases (Kapitel 4) implementiert, um die Funktionalität und den Nutzen des Metadaten-Modells zu demonstrieren.

Im Folgenden werden daher die Use Cases nochmals aufgegriffen und für jeden Use Case das Verhalten des Prototyps diskutiert.

Datenidentifikation von Daten über Betroffenen Für die Erfüllung von Betroffenenrechtsansprüchen (z.B. Recht auf Erhalt einer Datenkopie) ist es erforderlich, dass alle einer Person zugeordnete personenbezogene Daten identifiziert werden können.

Erwartetes Verhalten: Das System sollte zunächst den Betroffenen eindeutig identifizieren, da anderenfalls eine solche Anfrage nicht beantwortet werden kann. Wenn dies gelingt, sollte es eine Auflistung aller Daten liefern, die zu dieser Person gespeichert sind. Daten, die mehreren Personen zugeordnet sind, sollten getrennt ausgewiesen werden, da bei diesen auch der Datenschutz anderer Personen betroffen ist.

Implementierung: Zunächst versucht das System den Betroffenen eindeutig zu identifizieren. Im Rahmen des Prototyps gilt ein Betroffener als eindeutig identifiziert, wenn die zur Verfügung gestellten Daten (z.B. Vorname, Nachname und/oder E-Mail-Adresse) nur einen Treffer in der Betroffenen-Datenbank liefern. Bei erfolgreicher Identifizierung des Betroffenen bestimmt das System anhand der Metadaten alle dieser Person zugeordneten personenbezogenen Daten. Danach überprüft das System, welche dieser Daten weiteren Personen zugeordnet sind und gibt danach die Daten, die nur dem Betroffenen zugeordnet sind und Daten, die mehreren Personen zugeordnet sind, getrennt voneinander aus.

Newsletter versenden Um einen Newsletter zu versenden, werden alle für diese Verarbeitung und diesen Verarbeitungszweck benötigten Daten, konkret Name und E-Mail-Adresse, angefordert.

Erwartetes Verhalten: Das System sollte, unter Berücksichtigung von z.B. Widersprüchen des Betroffenen oder Widerruf der Einwilligung, eine Liste der benötigten Daten generieren.

Implementierung: Das System identifiziert zunächst anhand der Verarbeitung und des Verarbeitungszweckes, welche Daten für die Verarbeitung genutzt werden dürfen. Diese Daten werden dann ausgegeben, sofern der Datenzugriff erlaubt ist (detailliert dazu Abschnitt 7.4). Falls der Datenzugriff auf einzelne Daten nicht erlaubt ist, wird dem Nutzer eine diesbezügliche Warnung angezeigt.

Newsletterabmeldung Ein potentieller Kunde meldet sich vom Newsletter ab und seine Daten werden zu keinem anderen Zweck mehr verarbeitet, es gibt keine weitere Geschäftsbeziehung mehr. Zudem verlangt er die Löschung seiner Daten.

Erwartetes Verhalten: Die hinterlegte Einwilligung für den Newsletterempfang sollte gelöscht werden und der Betroffene darf für den weiteren Newsletterversand nicht mehr berücksichtigt werden.

Das System muss prüfen, ob das Datum für weitere Zwecke benötigt wird. Falls dem, wie in diesem Fall, nicht so ist und somit keine Rechtsgrundlage für die Speicherung mehr besteht, muss das Datum gelöscht werden (vgl. Abschnitt 5.3.3). Zudem muss eine Liste der Empfänger der Daten generiert werden, die über das Löschbegehren zu informieren sind (vgl. Abschnitt 5.3.5).

Implementierung: Bei der Löschung der Einwilligung wird die Zuordnung, dass diese Daten für die Newsletter-Verarbeitung genutzt werden dürfen, entfernt. Um dem Löschanspruch gerecht zu werden, wird zudem überprüft, ob die Daten für einen weiteren Zweck genutzt werden dürfen oder müssen. Dazu wird in der Relationstabelle `Verarbeitung_pbD` überprüft, ob das Datum einem weiteren Verarbeitungszweck zugeordnet ist. Wenn dies nicht der Fall ist, werden zunächst die Empfänger der Daten ermittelt. Diese werden, zusammen mit dem jeweiligen Übertragungsweg, wie dieser Empfänger über die Löschung zu benachrichtigen ist, ausgegeben. Zuletzt werden die entsprechenden personenbezogenen Daten gelöscht.

Einschränkung der Verarbeitung Ein Kunde hat gegenüber dem Onlinehändler erwirkt, dass die Verarbeitung eines Datums eingeschränkt werden muss, das heißt für die weitere Verwendung gesperrt werden muss (vgl. Abschnitt 5.3.4). Nun soll auf dieses Datum zugegriffen werden.

Erwartetes Verhalten: Das System sollte den Datenzugriff mit Hinweis auf die Einschränkung der Verarbeitung blockieren.

Implementierung: Das System prüft vor der Gewährung des Datenzugriffs in den Metadaten, ob ein Datum zur Verarbeitung gesperrt ist. Da das Datum gesperrt ist, wird der Datenzugriff mittels einer Exception verweigert.

Streit um Einwilligung Ein potentieller Kunde erhält einen Newsletter auf Basis einer Einwilligung. Er beschwert sich beim Unternehmen, da er meint, keine Einwilligung erteilt zu haben.

Erwartetes Verhalten: Das System muss in der Lage sein zu überprüfen, ob eine Einwilligung gespeichert ist und diese, so sie vorhanden ist, auszugeben.

Implementierung: Das System prüft für die gegebene Verarbeitung und den gegebenen Verarbeitungszweck, ob eine Einwilligung des Betroffenen hinterlegt ist. Diese wird dann, zusammen mit den Informationen über Zeitpunkt der Erteilung und Art der Einwilligung ausgegeben. Listing 8.1 zeigt als Beispiel die Ausgabe einer Einwilligung zum Empfang eines Newsletters.

Tabelle 8.1 nennt auf Basis der Tabelle 5.1 nochmal alle in Kapitel 5 herausgearbeiteten Herausforderungen und zeigt, welche durch den implementierten Prototypen abgedeckt werden. Die mit ”✓”markierten Herausforderungen werden durch den Prototypen adressiert, während die mit ”mar-

Listing 8.1 Ausgabe einer Einwilligung zum Newsletterempfang

```
Einwilligung von Vannie Nay (Vannie.Nay@example.com) zu Verarbeitungszweck 'Newsletter
Neukunden':
Einwilligung erteilt am: 02.12.2017 17:23:15 Uhr
Einwilligung wurde als Double Opt-In erteilt.
Ort der Einwilligung: https://example.com/newsletter
Der Wortlaut der Einwilligung lautet:
Wir verarbeiten Ihre personenbezogene Daten (Name, Vorname, E-Mail), um Ihnen unseren
Newsletter zuzusenden. Mit Ihrer Anmeldung willigen Sie in die Verarbeitung dieser Daten ein
und stimmen zu, künftig unseren Newsletter zu erhalten. Sie können Ihre Einwilligung jederzeit
 widerrufen. Zudem haben Sie das Recht auf Auskunft, Berichtigung, Löschung, Einschränkung der
 Verarbeitung, sowie ein Recht auf Datenportabilität. Sollten Sie der Meinung sein, dass wir
Ihre Daten nicht ordnungsgemäß verarbeiten, so steht Ihnen ein Recht auf Beschwerde bei einer
Aufsichtsbehörde zu.
```

kierten nicht durch die Metadaten und zugehörige Governance-Prozesse abgedeckt werden können. Diese erfordern andere Lösungsansätze, z.B. eine manuelle Abwägung (vgl. dazu auch Kapitel 5 und speziell Abschnitt 5.9).

Herausforderungen, die mit "$\checkmark^*$"markiert sind, wie die Datenkopie oder Datenportabilität ohne Daten Dritter, können nur teilweise technisch gelöst werden (z.B. Datenidentifikation), während der andere Teil manuell erfolgen muss (z.B. rechtliche Abwägung). Der technische Teil ist im Prototypen implementiert, der manuelle seiner Natur nach nicht.

Die mit "($\checkmark^+$)"markierten Herausforderungen der Datenlöschung und -änderung sind im Prototypen insoweit implementiert, als dass der Prototyp keine Zonen des Speichersystems kennt und MariaDB die Änderung und Löschung von Daten ermöglicht. Von den Lösungsvorschlägen wurde daher die Technologiewahl umgesetzt, während die Änderung des Zonenmodells aufgrund der vereinfachten Struktur des Prototypen nicht erfolgt ist.

Die mit "($\checkmark$)"markierten Herausforderungen wurden partiell umgesetzt. Für die „Mitteilung an Empfänger“ (Abschnitt 5.3.5) werden die Empfänger und der jeweilige Übertragungsweg in den Metadaten dokumentiert. Zudem werden bei der Erfüllung des Rechts auf Löschung diese Metadaten auch ausgegeben. Allerdings erfolgt, in Ermangelung entsprechender Gegenstellen, keine Benachrichtigung der entsprechenden Empfänger. Diese kann jedoch auf Basis dieser Metadaten problemlos erfolgen.

Da der Data Lake ein Mittel der Verarbeitung für andere Verarbeitungen darstellt (vgl. Abschnitt 5.4) und in diesem Data Lake Prototypen einige der Pflichten, die sich aus der Verarbeitung personenbezogener Daten ergeben, erfüllt werden, wird die Herausforderung „Datenschutz in Mitteln der Verarbeitung“ teilweise erfüllt. Zudem werden einige Nachweis- und Dokumentationspflichten, wie z.B. Dokumentation der Einwilligung im Prototypen adressiert, während andere wie z.B. das Verarbeitungsverzeichnis nicht durch das Metadaten-Modell abgedeckt werden.

Da der Fokus dieser Arbeit auf den Herausforderungen durch die DSGVO beim Design eines Data Lakes lag, wurden die in der Diskussion der Pflichten (Kapitel 5) nicht berücksichtigten Benachrichtigungspflichten gegenüber dem Betroffenen im Prototypen nicht implementiert. Insbesondere

Problem	Abdeckung durch Prototyp
Planung rechtskonformer Verarbeitung	-
Sicherstellung rechtskonformer Verarbeitung	✓
Explorative Datenanalyse	-
Sicherstellung der Zweckbindung	✓
Änderung von Daten	(✓$^{+}$)
Löschung von Daten	(✓$^{+}$)
Proaktive Löschung von Daten	✓
Kontextabhängige Richtigkeit	-
Sicherstellung der Korrektheit von Daten	-
Identifikation der Daten über einen Betroffenen	✓
Kopie ohne Daten Dritter	✓*
Sperrung von Daten	✓
Identifikation von Datenempfängern	✓
Mitteilung an Datenempfänger	(✓)
Reichweite des Portabilitätsanspruchs	-
Austauschformat	-
Portabilität ohne Daten Dritter	✓*
Weitergabe an anderen Verantwortlichen	-
Prüfung Widerspruch	-
Umsetzung Widerspruch	✓
Auswahl TOMs	-
(Regelmäßige) Überprüfung TOMs	-
Datenschutz in Mitteln der Verarbeitung	(✓)
Datensicherheit eines zentralen Speichersystems	-
DSFA für Data-Lake-Verarbeitungen	-
Erfüllung Nachweis- und Dokumentationspflichten	(✓)
Doppelte Dokumentation in Metadaten und Verarbeitungsverzeichnis	-

Tabelle 8.1: Zusammenfassung Pflichten & Lösungen

im Bereich der automatisierbaren Bearbeitung von Betroffenenanfragen (z.B. Auskunftsrecht oder Widerspruch gegen Marketing) ist eine Automatisierung dieser Benachrichtungspflichten durch Prozesse des Data Lakes naheliegend.

9 Zusammenfassung und Ausblick

Die Verarbeitung personenbezogener Daten im Data Lake birgt viele Vorteile, erfordert jedoch im Licht der Datenschutzgrundverordnung eine besondere Aufmerksamkeit. Insbesondere ermöglicht die zentrale Speicherung von Daten die Integration von Daten aus verschiedenen Systemen und wirkt der Bildung von Datensilos entgegen. Data Lakes mit starker Governance ermöglichen es, personenbezogene Daten im Rahmen der gesetzlichen Anforderungen zu verarbeiten und so von den Vorteilen des Data Lakes zu profitieren.

In dieser Arbeit wurde daher die Datenschutzgrundverordnung im Hinblick auf Herausforderungen beim Design eines Data Lakes analysiert. Dazu wurden zunächst die aus der DSGVO entstehenden Pflichten diskutiert, dann die Herausforderungen im Hinblick auf den Data Lake herausgearbeitet und schließlich Lösungsvorschläge erarbeitet.

Betrachtet man den Data Lake als Speichersystem, so entstehen zunächst die gleichen Herausforderungen wie bei anderen Speichersystemen, in denen personenbezogene Daten gespeichert werden. Es müssen insbesondere geeignete technische und organisatorische Maßnahmen ergriffen werden, um sowohl den Schutz der Betroffenen als auch die Datensicherheit zu gewährleisten.

Allerdings entstehen bei der Nutzung eines Data Lakes durch die zentrale Speicherung von Daten aus verschiedenen Quellen und Systemen zusätzliche Anforderungen an das Speichersystem. Um bei der Verwendung eines Data Lakes eine DSGVO-konforme Verarbeitung personenbezogener Daten zu gewährleisten, sollte der Data Lake den Verantwortlichen bei der Einhaltung weiterer Pflichten unterstützen. Dies gilt insbesondere im Bereich des Datenzugriffs und betrifft die Grundsätze der Rechtmäßigkeit und der Zweckbindung, sowie deren Ausgestaltungen. Durch Metadaten und zugehörige Governance-Prozesse kann der Data Lake den Datenzugriff nur gewähren, soweit diese Daten für diesen Zweck verarbeitet werden dürfen und die entsprechende Rechtsgrundlage vorliegt. Dies betrifft ebenfalls die Einhaltung der Betroffenenrechte. So können Governance-Prozesse und Metadaten beispielsweise bei der Identifikation der zu einem Betroffenen gespeicherten Daten oder der Sicherstellung der Löschung von Daten den Verantwortlichen im Bereich der Betroffenenrechte unterstützen. Daher wurde in dieser Arbeit ein Metadaten-Modell entwickelt, um den Verantwortlichen bei der Einhaltung dieser Pflichten zu unterstützen.

Die Grundsätze der Speicherbegrenzung und der Datenrichtigkeit erfordern zudem eine Änderung am Konzept des Data Lakes. Überall, wo personenbezogene Daten gespeichert werden, muss sichergestellt werden, dass diese Daten verändert und gelöscht werden können. Außerdem ist es nicht möglich, beliebige Daten in die Sandbox zu kopieren und zu analysieren, da dem der Grundsatz der Zweckbindung sowie die Grundsätze der Fairness und Transparenz entgegenstehen. In dieser Arbeit wurde daher das Zonenmodell als verbreitetes Architekturkonzept für Data Lakes entsprechend angepasst.

Explorative Analysen personenbezogener Daten unterliegen durch die Grundsätze der Fairness, Transparenz, Zweckbindung, Rechtmäßigkeit, Datenminimierung und Speicherbegrenzung strengen Anforderungen. Da explorative Analysen eigenständige Verarbeitungen personenbezogener Daten sind, benötigen sie eine eigene Rechtsgrundlage und unterliegen, solange keine extra Einwilligung eingeholt wurde, den strengen Anforderungen zur Weiterverarbeitung personenbezogener Daten. Um den Grundsätzen der Fairness und Transparenz gerecht zu werden, dürfen Daten nur im Rahmen dessen verarbeitet werden, wie der Betroffene dies vernünftigerweise erwarten kann. Zudem muss die Verarbeitung für den Betroffenen verständlich und nachvollziehbar sein. Die Grundsätze der Datenminimierung und der Speicherbegrenzung stehen der Grundidee von Big Data entgegen, dass zunächst so viele Daten wie möglich als Grundlage für spätere Analysen gesammelt werden. Sie erfordern, dass nur so wenig personenbezogene Daten wie nötig erhoben und verarbeitet werden und dass diese gelöscht werden, sobald sie nicht mehr benötigt werden. Dies schränkt die Datenbasis für explorative Analysen ein.

Das im Rahmen dieser Arbeit entwickelte Metadaten-Modell wurde mit zugehörigen Governance-Prozessen prototypisch implementiert. Zudem wurde ein realistisches Anwendungsszenario und Use Cases entwickelt. Anhand dieser Use Cases wurde die Funktionalität des Metadaten-Modells exemplarisch getestet.

Die Diskussion der Implementierung zeigte, dass das Metadaten-Modell die gestellten Anforderungen erfüllen kann und für viele Herausforderungen bei der Verarbeitung personenbezogener Daten im Data Lake ein hilfreiches Werkzeug darstellt. Allerdings können viele Herausforderungen, insbesondere im Bereich der explorativen Datenanalyse, nicht technisch gelöst werden, da z.B. Rechtsfragen oder -abwägungen nicht durch das System erfolgen können.

Die Frage, ob ein konkreter Data Lake DSGVO-konform ist, ist von insbesondere von den dort zu speichernden Daten und den Verarbeitungen, für die der Data Lake genutzt werden soll abhängig. Durch den konkreten Einzelfall kann bei der Prüfung, ob sich ein spezifischer Data Lake mit der DSGVO vereinen lässt, eine andere Bewertung entstehen als bei dieser Arbeit, die sich am Grundkonzept des Data Lakes orientiert und keine spezifische Implementierung zur Grundlage hat. Dennoch kann sie als erster Anhaltspunkt dienen, welche Aspekte eines Data Lakes im Hinblick auf die Speicherung personenbezogener Daten problematisch sind und gelöst werden müssen.

Ausblick

In dieser Arbeit wurden die Herausforderungen durch die DSGVO beim Design eines Data Lakes untersucht. Um in der Praxis personenbezogene Daten in einem Data Lake zu verarbeiten, sollte auch dieser praktische Aspekt wissenschaftlich untersucht werden. Beispielsweise können Konzepte entwickelt werden, um den Verantwortlichen durch Data Lake Prozesse bei der (automatisierten) Bearbeitung der Benachrichtigungspflichten bei Betroffenenrechten zu unterstützen (z.B. Auskunftserteilung).

Zudem eröffnet die Thematik Datenschutzverletzungen (Art. 33, 34 DSGVO) weitere Forschungsfragen wie bspw. „Wie kann der Data Lake den Verantwortlichen bei der Erkennung von Datenschutzverletzungen unterstützen?“oder „Wie kann der Data Lake die Risikoeinschätzung bei einer Datenschutzverletzung und ggf. die Meldung an die Aufsichtsbehörde oder den Betroffenen unterstützen?“.

Diese Arbeit hat das abstrakte Konzept Data Lake anhand eines realistischen, aber fiktiven Anwendungsszenario untersucht. Eine derartige Untersuchung sollte an einer konkreten Implementierung des Data Lakes wiederholt werden. Diese Untersuchung kann insbesondere auch konkreter auf die durch die jeweilige Wahl der Implementierungstechnologie entstehenden Herausforderungen eingehen.

Vom direkten Anwendungsfall Data Lake losgelöst wäre eine interdisziplinäre Untersuchung anwendungsrelevanter Themen, wie z.B. Wirksamkeit von Anonymisierung oder Pseudonymisierung im Kontext von Big Data sehr interessant.

Literaturverzeichnis

[AE17a] Artikel 29 Datenschutzgruppe, Europäischer Datenschutzausschuss. *Leitlinien zum Recht auf Datenübertragbarkeit*. Techn. Ber. WP242 rev.01. engl. Titel: Guidelines on the right to data portability. 5. Apr. 2017. URL: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611233 (zitiert auf S. 18, 57, 58).

[AE17b] Artikel 29 Datenschutzgruppe, Europäischer Datenschutzausschuss. *Leitlinien zur Datenschutz-Folgenabschätzung (DSFA) und Beantwortung der Frage, ob eine Verarbeitung im Sinne der Verordnung 2016/679 „wahrscheinlich ein hohes Risiko mit sich bringt"*. engl. Titel: Data protection impact assessment Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk"for the purposes of Regulation 2016/679. 4. Okt. 2017. URL: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236 (zitiert auf S. 18, 65).

[AE18] Artikel 29 Datenschutzgruppe, Europäischer Datenschutzausschuss. *Leitlinien in Bezug auf die Einwilligung gemäß Verordnung 2016/679*. engl. Titel: Guidelines on Consent under Regulation 2016/679. 10. Apr. 2018. URL: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611233 (zitiert auf S. 18, 47).

[Arc18] Arcadia Data. *Arcadia Data Drives GDPR Compliance With Modern Data Lake BI Architecture*. 23. Mai 2018. URL: https://www.arcadiadata.com/press-release/arcadia-data-drives-gdpr-compliance-with-modern-data-lake-bi-architecture/ (zitiert auf S. 17).

[Bay17] Bayrisches Landesamt für Datenschutzaufsicht. „Verarbeitung personenbezogener Daten für Werbung". In: *EU-DS-GVO - Das BayLDA auf dem Weg zur Umsetzung der Verordnung*. Nr. 12. 4. Mai 2017. URL: https://www.lda.bayern.de/media/baylda_ds-gvo_12_advertising.pdf (zitiert auf S. 18, 47).

[Bay18] Bayrisches Landesamt für Datenschutzaufsicht. *Muster 9: Online-Shop – Verzeichnis von Verarbeitungstätigkeiten*. 2018. URL: https://www.lda.bayern.de/media/muster_9_online-shop_verzeichnis.pdf (zitiert auf S. 18).

[BDW15] U. Bub, V. Deleski, K. Wolfenstetter. *Sicherheit im Wandel von Technologien und Märkten: Tagungsband zur vierten EIT ICT Labs-Konferenz zur IT-Sicherheit*. Springer Fachmedien Wiesbaden, 2015. ISBN: 9783658112745 (zitiert auf S. 43).

[Bir18] E. Birsin. „Data Lakes und europäischer Datenschutz - Sicher unterwegs im Data Lake". In: *BI-SPEKTRUM* 4 (2018) (zitiert auf S. 17).

[Boc18] K. Bock. *Kirsten Bock auf Twitter*. K. Bock ist Senior Legal Counsel beim Unabhängigen Landeszentrum für Datenschutz Schleswig-Holstein. 14. Nov. 2018. URL: https://twitter.com/privacyDE/status/1062731923607429121 (zitiert auf S. 40).

[Bun78] Bundesverfassungsgericht. „Kalkar I“. In: *BVerfGE 49*. 8. Aug. 1978. URL: http://www.servat.unibe.ch/dfr/bv049089.html (zitiert auf S. 61).

[Bun83] Bundesverfassungsgericht. „Volkszählungsurteil“. In: *BVerfGE 65, 1*. 15. Dez. 1983. URL: http://www.servat.unibe.ch/dfr/bv065001.html (zitiert auf S. 26).

[Cas16] D. Castelvecchi. „The Black Box of AI“. In: *Nature* 538 (6. Okt. 2016), S. 20–23 (zitiert auf S. 39).

[CJL+15] M. Chessell, N. Jones, J. Limburn, D. Radley, K. Shank, I. Redbooks. *Designing and Operating a Data Reservoir*. IBM Redbooks, 2015. ISBN: 9780837440668. URL: https://books.google.de/books?id=-BWrCQAAQBAJ (zitiert auf S. 17).

[Com18] Commission Nationale de l'Informatique et des Libertés. *The open source PIA software helps to carry out data protection impact assesment*. 31. Mai 2018. URL: https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment (zitiert auf S. 67).

[Dom16] Dominic Tarr. *GitHub - dominictarr/random-name*. 2016. URL: https://github.com/dominictarr/random-name (zitiert auf S. 86).

[Dor17] A. Dorloff. *China auf dem Weg in die IT-Diktatur*. 9. Sep. 2017. URL: https://www.deutschlandfunk.de/sozialkredit-system-china-auf-dem-weg-in-die-it-diktatur.724.de.html?dram:article_id=395440 (zitiert auf S. 20).

[Dur17] E. Durmus. „Das Recht auf Datenübertragbarkeit nach der DSGVO - Handlungsoptionen und -notwendigkeiten für Unternehmen“. Bachelorarbeit. Hochschule Darmstadt, 2017 (zitiert auf S. 57).

[EU 18] EU Kommission. *Adequacy of the protection of personal data in non-EU countries*. 2018. URL: https://ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/adequacy-protection-personal-data-non-eu-countries_de (zitiert auf S. 69).

[Eur16] Europäischer Gerichtshof. „Breyer“. In: *C-582/14*. ECLI:EU:C:2016:779. 19. Okt. 2016. URL: http://curia.europa.eu/juris/document/document.jsf?text=&docid=184668&pageIndex=0&doclang=DE&mode=lst&dir=&occ=first&part=1&cid=1428447 (zitiert auf S. 29).

[Eur18] Europäischer Datenschutzausschuss. *DSGVO: Leitlinien, Empfehlungen, bewährte Verfahren*. 2018. URL: https://edpb.europa.eu/our-work-tools/general-guidance/gdpr-guidelines-recommendations-best-practices_de (zitiert auf S. 18).

[Fer17] L. Fernando. *7 V's of Big Data*. 17. Jan. 2017. URL: http://blogsofdatawarehousing.blogspot.com/2017/01/7-vs-of-big-data.html (zitiert auf S. 19).

[FHS17] N. Forgo, S. Haenold, B. Schuetze. *The Principle of Purpose Limitation and Big Data*. Sep. 2017. DOI: 10.1007/978-981-10-5038-1_2 (zitiert auf S. 17).

[Fir17] G. Firican. *The 10 Vs of Big Data*. 8. Feb. 2017. URL: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx (zitiert auf S. 19).

[GH15] A. Gandomi, M. Haider. „Beyond the hype: Big data concepts, methods, and analytics“. In: *International Journal of Information Management* 35.2 (2015), S. 137–144. URL: http://www.sciencedirect.com/science/article/pii/S0268401214001066 (zitiert auf S. 19).

[GSSV18] S. Gierschmann, K. Schlender, R. Stentzel, W. Veil, Hrsg. *Kommentar Datenschutz-Grundverordnung*. Bundesanzeiger Verlag, 2018 (zitiert auf S. 18, 39–44, 47, 48, 51, 53, 54, 56–58, 61–65).

[Här16] N. Härting. „Kopplungsverbot - der Einwilligungskiller nach der DSGVO“. In: *CRonline - Portal zum IT-Recht* (11. Okt. 2016). URL: https://www.cr-online.de/blog/2016/10/11/kopplungsverbot-der-einwilligungskiller-nach-der-dsgvo/ (zitiert auf S. 47).

[HB14] O. Heuberger-Götsch, T. Burkhalter. „Datenschutz in Zeiten von Big Data“. In: *HMD Praxis der Wirtschaftsinformatik* 51.4 (2014), S. 480–493 (zitiert auf S. 17).

[Kad15] V. Kadenic. „Compliance of Data Lake Enterprise Architecture Model with the General Data Protection Regulation (GDPR)“. Bachelorarbeit. Luleå University of Technology, 2015 (zitiert auf S. 17, 23).

[KAS+18] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, S. Salehian. „The 10 Vs, Issues and Challenges of Big Data“. In: *Proceedings of the 2018 International Conference on Big Data and Education*. ACM. 2018, S. 52–56 (zitiert auf S. 19).

[Kno18] T. Knobloch. „Vor die Lage kommen: Predictive Policing in Deutschland“. In: (2018). URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/predictive.policing.pdf (zitiert auf S. 13).

[Kon17] Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder. *Datenübermittlung in Drittländer*. Techn. Ber. Kurzpapier Nr. 4. 11. Juli 2017. URL: https://www.datenschutzkonferenz-online.de/media/kp/dsk_kpnr_4.pdf (zitiert auf S. 18, 69).

[Kon18] Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder. *Datenschutzkonferenz*. 2018. URL: https://www.datenschutzkonferenz-online.de/ (zitiert auf S. 18).

[Lan01] D. Laney. „3D data management: Controlling data volume, velocity and variety“. In: *META group research note* 6.70 (2001), S. 1 (zitiert auf S. 19).

[Lan18] Landesbeauftragter für Datenschutz und Informationsfreiheit Baden-Württemberg. *Liste von Verarbeitungsvorgängen nach Art. 35 Abs. 4 DS-GVO*. Techn. Ber. 2018. URL: https://www.baden-wuerttemberg.datenschutz.de/wp-content/uploads/2018/05/Liste-von-Verarbeitungsvorg%C3%A4ngen-nach-Art.-35-Abs.-4-DS-GVO-LfDI-BW.pdf (zitiert auf S. 66).

[LS14] A. LaPlante, B. Sharma. *Architecting Data Lakes*. O'Reilly Media, 2014 (zitiert auf S. 17, 21–24).

[Mar14] M. Martini. „Big Data als Herausforderung für den Persönlichkeitsschutz und das Datenschutzrecht“. In: *Deutsches Verwaltungsblatt* 129.23 (2014), S. 1481–1489 (zitiert auf S. 37).

[Mar18a] MariaDB Foundation. *About MariaDB*. 2018. URL: https://mariadb.org/about/ (zitiert auf S. 83).

[Mar18b] MariaDB Foundation. *MariaDB*. 2018. URL: https://mariadb.org/ (zitiert auf S. 83).

[Mic17] W. Michl. „Das Verhältnis zwischen Art. 7 und Art. 8 GRCh — zur Bestimmung der Grundlage des Datenschutzgrundrechts im EU-Recht“. In: *Datenschutz und Datensicherheit - DuD* 41.6 (Juni 2017), S. 349–353. URL: https://doi.org/10.1007/s11623-017-0789-9 (zitiert auf S. 26).

[MT17] MapR, Talend. *Get Ahead Of General Data Protection Regulation (GDPR) With MapR And Talend*. 2017. URL: https://mapr.com/resources/mapr-talend-gdpr-solution-brief/ (zitiert auf S. 17).

[NS06] A. Narayanan, V. Shmatikov. „How to break anonymity of the netflix prize dataset“. In: *arXiv preprint cs/0610105* (2006) (zitiert auf S. 37).

[Pla16] K.-U. Plath, Hrsg. *BDSG/DSGVO - Kommentar zum BDSG und zur DSGVO sowie den Datenschutzbestimmungen des TMG und TKG*. Dr. Otto Schmidt Verlag, 2016 (zitiert auf S. 48, 57).

[Poh18] J. Pohle. „Datenschutz und Technikgestaltung“. Dissertation. 2018 (zitiert auf S. 26).

[PWD17] P. Patel, G. Wood, A. Diaz. „Data Lake Governance Best Practices“. In: *DZone's Guide to Big Data: Data Science & Advanced Analytics* 4 (2017), S. 6–7. URL: https://dzone.com/articles/data-lake-governance-best-practices (zitiert auf S. 17, 21, 22).

[Rij13] M. van Rijmenam. *Why The 3V's Are Not Sufficient To Describe Big Data*. 7. Aug. 2013. URL: https://datafloq.com/read/3vs-sufficient-describe-big-data/166 (zitiert auf S. 19).

[Ril17] J. Riley. „Understanding metadata“. In: *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)* (2017) (zitiert auf S. 23).

[Sch17a] D. Schätzle. „Zum Kopplungsverbot der Datenschutz-Grundverordnung - Warum auch die DSGVO kein absolutes Kopplungsverbot kennt“. In: *PinG Privacy in Germany* 5 (2017). URL: https://www.pingdigital.de/PinG.05.2017.203 (zitiert auf S. 47).

[Sch17b] E. Schlehahn. *Transparenz als zentrales Element von Datenschutzrecht, Ethik und Technik*. Präsentation. 3. Okt. 2017. URL: https://www.forum-privatheit.de/forum-privatheit-de/publikationen-und-downloads/veroeffentlichungen-des-forums/2017-11-02-Jahrestagung-2017/2.3a_Schlehahn_FINAL_Forum-Privatheit_Transparenz.pdf (zitiert auf S. 38–40, 50).

[Som13] S. Sommer. „Warum Amazon weiß, was Ihre Frau mag“. In: *Manager Magazin* online (29. Nov. 2013). URL: http://www.manager-magazin.de/unternehmen/handel/big-data-analyse-im-online-handel-a-935555.html (zitiert auf S. 31).

[SSS+12] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, P. Tufano. „Analytics: The real-world use of big data“. In: *IBM Global Business Services* 12 (2012), S. 1–20 (zitiert auf S. 19).

[Tel18] TeleTrusT - Bundesverband IT-Sicherheit. *Handreichung zum „Stand der Technik“ technischer und organisatorischer Maßnahmen*. 2018. URL: https://www.teletrust.de/fileadmin/docs/fachgruppen/ag-stand-der-technik/TeleTrusT-Handreichung_Stand_der_Technik_-_Ausgabe_2018.pdf (zitiert auf S. 61).

[Vei18] W. Veil. *GDPR: 68 Obligations of the Controller*. W. Veil ist Referent im Referat IT I 1 (Digitale Agenda; Grundsatz- und Rechtsangelegenheiten der IT und Digitalisierung) im Bundesministerium des Innern, für Bau und Heimat. 17. Feb. 2018. URL: https://www.flickr.com/photos/winfried-veil/25437610017 (zitiert auf S. 35, 68).

[Wei13] T. Weichert. „Big Data und Datenschutz“. In: *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein* (2013) (zitiert auf S. 17).

[WZWD14] X. Wu, X. Zhu, G.-Q. Wu, W. Ding. „Data mining with big data“. In: *IEEE transactions on knowledge and data engineering* 26.1 (2014), S. 97–107 (zitiert auf S. 20).

[ZDF18] ZDF. *ZDF heute Sendung vom 24.05.2018*. 24. Mai 2018. URL: https://www.zdf.de/nachrichten/heute-19-uhr/180524-h19-gesamtsendung-100.html (zitiert auf S. 13).

[ZDP+13] P. Zikopoulos, D. Deroos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan. *Harness the power of big data: The IBM big data platform*. McGraw-Hill New York, NY, 2013 (zitiert auf S. 19).

Alle URLs wurden zuletzt am 01. 12. 2018 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift