SCHÜRMANN
ROSENTHAL
DREYER
RECHTSANWÄLTE

DIGITALES BUSINESS · TECHNOLOGIE · MEDIEN

DSGVO-konforme Datennutzung mit Data Lakes

Stiftung Datenschutz – Datenschutz am Mittag

Philipp Müller-Peltzer
Rechtsanwalt, Partner

Ilan Leonard Selz
Rechtsanwalt, Senior Associate

Empfehlungen:

JUVE HANDBUCH 2020|2021

NOMINIERT
JUVE Awards 2022
Kanzlei des Jahres für Technologie und Medien

JUVE AWARDS 2020
Kanzlei des Jahres für IT und Datenschutz

JUVE AWARDS 2020
Kanzlei des Jahres für Technologie und Medien

The LEGAL 500 DEUTSCHLAND
FÜHRENDE KANZLEI 2022

Agenda

1 Überblick

2 Personenbezug

3 RGL und Zweckbindung

4 Betroffenen-rechte

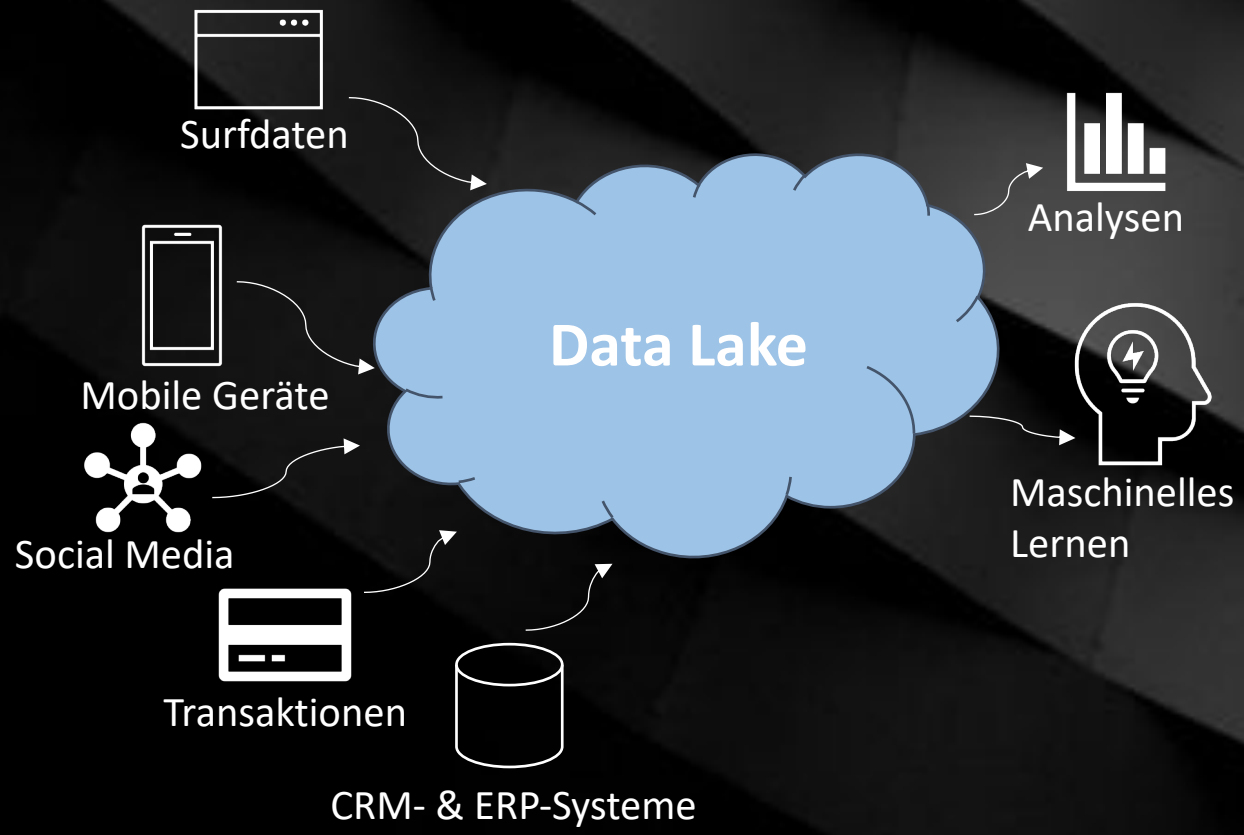
5 TOM

6 Fazit und Best Practices

I. Was sind Data Lakes?

Was sind Data Lakes?

- Skalierbare Systeme im Big-Data-Kontext zur Speicherung großer Mengen an strukturierten und unstrukturierten Daten aus unterschiedlichen Quellen
- Rohdaten + Metadaten (Data Lineage, Data Quality etc.)
- Zielsetzung: prädiktive Analytik, BI Analytics, Modellierung, Data Discovery unter Ermöglichung eines Selfservice und Unterstützung der Governance
- Verwendung durch unterschiedliche Nutzer, um
  - Daten zusammenführen und analysieren zu können
  - Modelle des Maschinellen Lernens zu trainieren
  - Speicherung historischer Daten
- Endgültige Verwendung der Daten steht zum Zeitpunkt der Speicherung noch nicht immer fest
- Herausforderung: Verhinderung der Entstehung eines Data Swamp



Data Lakes vs. Data Warehouses

	Data Lakes	Data Warehouse
Datenstruktur	Roh, unverarbeitet, heterogen	Verarbeitet und strukturiert
Fokus	Schema-on-read	Schema-on-write
Fokus	Skalierbarkeit, Flexibilität	Geschwindigkeit, Qualität
Benutzer	Datenanalysten: Muster und Zusammenhänge erkennen	Business-Anwender: Bestimmte Relationen verlässlich beauskunften
Zweck der Daten	Nicht spezifisch vorab definiert	Festgelegte Analysezwecke

→ Speicherstrukturen für unterschiedliche Anforderungen – Kombinationen denkbar – kein Ausschluss

II. Personenbezug

Personenbezug

- Art. 4 Abs. 1 DSGVO: alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen
- Pseudonymisierung: Bezug zu spezifischer Person nur unter Heranziehung zusätzlicher Informationen möglich → Personenbezug bleibt bestehen
- Anonymisierung: wenn keine Identifizierbarkeit mehr möglich → Kein Personenbezug und keine Anwendbarkeit der DSGVO (solange anonymisiert)
- Relevant für den gesamten Lebenszyklus im Data Lake:
  1. Einspeisen von Daten aus diversen Quellen in den Data Lake
  2. Datenhaltung im Data Lake
  3. Datenentnahme für Analysen und andere Nutzungen
  4. Ggf. Weiterverwendung von Ergebnissen und Zurückspeisen von Daten in den Data Lake

Anonymisierung I

- Maßstab nach Erwägungsgrund 26 DSGVO: alle Mittel, die *„nach allgemeinerem Ermessen wahrscheinlich genutzt werden“*
  → objektive Faktoren, wie Kosten der Identifizierung und Zeitaufwand nach Stand der Technik
- Absolute und relative Anonymisierung z.B. durch Trust Center
- Verfahren zur Anonymisierung, z.B.:
  - Entfernung identifizierender Merkmale
  - Aggregieren von Daten
  - Masking, Shuffling
  - Kryptografische Hash-Verfahren
- Besondere Herausforderungen bei unstrukturierten Daten (z.B. Audio, Video, Bild und Freitext)
- Motivated-Intruder-Test

Anonymisierung II

- Relevant für den gesamten Lebenszyklus im Data Lake:
  - Einspeisen von Daten aus diversen Quellen in den Data Lake→ soweit Ursprungs-Daten personenbezogen, in der Regel nur Pseudonymisierung denkbar
  - Datenhaltung im Data Lake → in der Regel höchstens Pseudonymisierung
  - Datenentnahme für Analysen und andere Nutzungen → Pseudonymisierung; ggf. Anonymisierung
  - Ggf. Weiterverwendung von Ergebnissen und Zurückspeisen von Daten in den Data Lake → In Regel Anonymisierung

III. Rechtsgrundlagen und Zweckbindung

Rechtsgrundlagen

- Rechtsgrundlagen für die Datenverarbeitung sind gem. Art. 6 Abs. 1 S. 1 DSGVO:
  - Einwilligung der betroffenen Person
  - Datenverarbeitung dient der Erfüllung eines Vertrags
  - Vorliegen von berechtigten Interessen
- Erforderlich für die Verarbeitung personenbezogener Daten ist jedoch nur das Vorliegen einer einzigen Rechtsgrundlage
- Bei besonderen Kategorien personenbezogener Daten ist zusätzlich noch Art. 9 Abs. 2 DSGVO zu beachten
- RGL für jede „Verarbeitung“ erforderlich → Unterscheidung insbesondere zwischen:
  - RGL für Datenhaltung im Data Lake
  - RGL für Nutzungen und Analysen

Zweckbindung bei Nutzung I

- Grundsatz der Zweckbindung, Art. 5 Abs. 1 lit. b DSGVO: Verarbeitung personenbezogener Daten für festgelegte, eindeutige und legitime Zwecke
- Anforderungen der DSGVO:

Festlegung **vor** Datenerhebung	Transparenz gegenüber der betroffenen Person	Hinreichend bestimmter Zweck

- Speicherung in Data Lakes erfolgt häufig mit der Intention, über die Verarbeitungszwecke erst *nach* Datenerhebung zu entscheiden
- Wurden Daten jedoch ursprünglich für andere Zwecke erhoben, dürfen diese nur dann für andere Zwecke verwendet werden, wenn zulässige Zweckänderung nach Art. 6 Abs. 4 DSGVO vorliegt.
- Einwilligung, Spezialvorschrift oder neuer Zweck muss kompatibel mit ursprünglichem Verarbeitungszweck sein → Einzelfallprüfung

Zweckbindung bei Nutzung II

- Die Nutzung der Daten muss i.d.R. den Anforderungen des Art. 6 Abs. 4 DSGVO genügen
- Durchführung einer Kompatibilitätsprüfung: Vergleich des ursprünglichen Erhebungszwecks mit dem neuen Analysezweck
- Kriterien (nicht abschließend):
  - Inhaltlicher Zusammenhang zwischen dem initialen Zweck und dem neuen Zweck
  - Zusammenhang mit der Datenerhebung → Was konnte ein Betroffener vernünftigerweise erwarten?
  - Art der Daten: bei Daten nach Art. 9 DSGVO Vermutung der Inkompatibilität
  - Folgen: Eintrittswahrscheinlichkeit und Ausmaß von Gefahren bei Zweckänderung
  - Vornahme von angemessenen Schutzvorkehrungen für Weiterverarbeitung

→ Rechtsfolgen:
  - Wenn Kompatibilität (+), ist keine neue Rechtsgrundlage erforderlich (h.M.)
  - Dennoch Information über Zweckänderung

IV. Betroffenenrechte

Informationspflichten und Transparenz

- Grundsatz der Transparenz (Art. 5 Abs. 1 lit. a DSGVO) und Informations- und Aufklärungspflichten (Art. 12 ff. DSGVO) über Art, Zweck, Ausmaß der Datenverarbeitung
  - Unterscheidung zwischen Direkterhebung (Art. 13 DSGVO) und Dritterhebung (Art. 14 DSGVO)
  - Besondere Informationspflichten bei Zweckänderung nach Art. 13 Abs. 3 DSGVO bzw. Art. 14 Abs. 4 DSGVO (auch bei Kompatibilität der Zwecke)
  - Ggf. erweiterte Informationspflichten bei „einer automatisieren Entscheidungsfindung einschließlich Profiling“, Drittlandsübermittlungen und Darstellung berechtigter Interessen
- Herausforderungen:
  - Data Lakes für Laien verständlich zu erläutern und nachvollziehbar zu machen, vor allem, da die Zwecke der späteren Datenanalysen noch nicht feststehen
  - Prozess für nachhaltige Pflege der Informationen bei Evolution der Zwecke
  - Darüber hinaus: Geheimhaltungsinteressen des Unternehmens
- Nutzen:
  - Inventarisierung / Auffinden Datenpunkte in unstrukturierten Quellen

Löschpflichten

- Löschung auf Antrag und Löschung nach Zweckerfüllung/Zeitablauf
- Data Governance: Kohärenter Löschprozess
  - Auffindbarkeit im Data Lake
  - Erstreckung auf Data Lake und Analyse-Umgebung
- Löschen durch Anonymisierung
- Pseudonymisierungskonzepte und Prozesse, die für Löschung Beziehungen flächendeckend aufheben
- Tool-basierte Ansätze:
  - Identity Services/ Data Catalogs: Schema, welche Datenquellen und -Felder Identifier enthalten; Unterstützung beim initialen Taggen und systemübergreifenden Metadaten-Management
  - Data Lake Management Plattformen

Weitere Betroffenenrechte

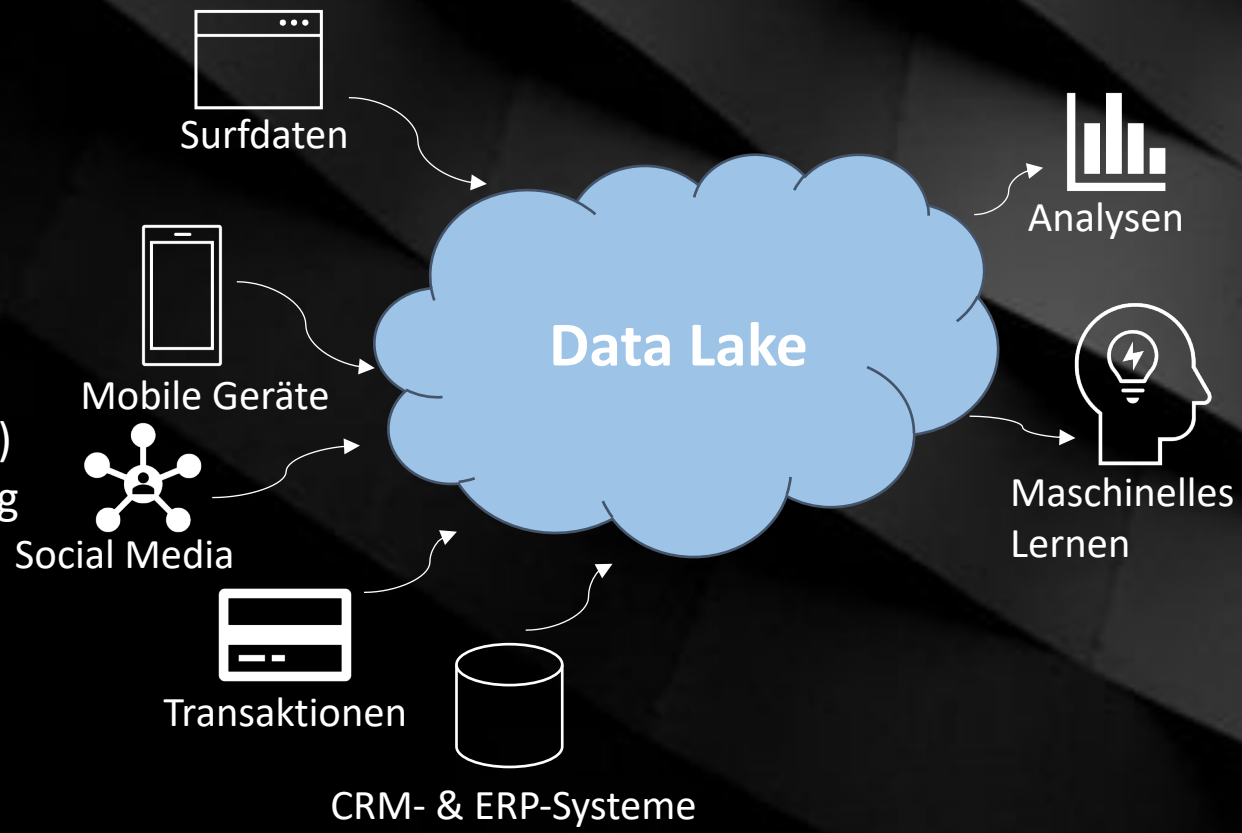
- Recht auf Auskunft: Betroffene Person kann von dem Verantwortlichen Auskunft verlangen, ob und welche personenbezogenen Daten verarbeitet werden
  - Identifizierung der entsprechenden Daten ist erforderlich
  - Verantwortlicher muss Kenntnisse über die Datenquellen und den Aufbau des Data Lakes haben, um die Auskunft geben zu können (bzgl. Zwecke, Empfänger, Drittstaaten, Profiling)
  - Mögliche Lösung: Data Catalogs / ganzheitliches Metadatenmanagement
- Recht auf Berichtigung: Recht auf Korrektur unrichtiger personenbezogener Daten oder Ergänzung unvollständiger Daten
  - Identifizierung der entsprechenden Daten und Lokalisierung
  - Mögliche Lösung: Metadaten und Data Lineage Tools

V. Technische und organisatorische Maßnahmen

TOM

- Zielkonflikt: Datenminimierung nach Art. 5 DSGVO vs. Datenmaximierung?
- Art. 25 DSGVO Privacy by Design und Privacy by Default: Datenschutz bereits für Lake-Architektur maßgeblich
- Diversere Komponenten:
  - Produktivsysteme
  - Data Lake
  - Angebundene Tools (z.B. BI-Tools, Data Bricks, Tableau)
  - Data Catalog zur Inventarisierung und Vorqualifizierung
  - Zwischenschritte (Veredelung, Strukturierung, Pseudonymisierung, etc.)
- Dienstleister und Cloud-Dienste involviert

→ Ziel: Einheitliches (hohes) Sicherheitsniveau nach dem Stand der Technik (Art. 32 DSGVO)



Rollen- und Berechtigungskonzept

- Maßnahmen mit denen sichergestellt und dokumentiert wird, welche Personen Zugriff auf das System haben und die Reichweite ihres Zugriffs
- Berechtigungskonzept sollte dem Need-to-Know-Prinzip folgen
- Speziell für Data Lakes:
  - Differenzierung zwischen Fachbereichen und Data Scientists
  - Abgestufte Berechtigungen
  - Beschränkung: Berechtigung der Data Scientists auf konkrete Analysezwecke
  - Keine/wenig Super-User
  - Keine/wenig doppelte Rollen
- Sicherstellung der Berechtigung durch:
  - Eindeutige Zuordnung von Benutzerrechten
  - Zugangsregelung durch z.B. Passwörter
  - Logging

Verschlüsselung der Daten

- Sicherheitsmaßnahme nach Art. 32 DSGVO: Informationen müssen in eine nicht mehr interpretierbare Zeichenfolge umgewandelt werden
- Wichtig: trotz Verschlüsselung handelt es sich in der Regel um personenbezogene Daten
- Schlüssel-Management
- Um eine hohe Sicherheit gewährleisten zu können, sollten die Daten bereits gleich bei der Ankunft in den Data Lake verschlüsselt werden, bevor sie in den permanenten Speicher des Data Lakes gelangen
- Differenzierung zwischen:
  - data in transit → Verschlüsslung zwingend
  - data at rest → Verschlüsselung empfehlenswert
  - data in use → Verschlüsselung Ausnahme

Quelle: TheDigitalArtist

Cloudanwendungen I – AVV

- Data Lakes befinden sich wegen der besseren Skalierbarkeit in Clouds
- Räumliche Anwendbarkeit der DSGVO
- Cloud-Dienstleister sind typischerweise Auftragsverarbeiter im Sinne des Art. 4 Nr. 8 DSGVO
- Erforderlich ist somit ein Auftragsvereinbarungsvertrag (AVV)
  - Cloud-Dienstleister muss selbst auch Sicherheitsmaßnahmen wie z.B. Verschlüsselung oder Pseudonymisierung der Daten vornehmen
  - Festlegung der Bedingungen, unter denen der Auftragsverarbeiter selbst einen Auftragsverarbeiter in Anspruch nimmt
- Einheitliches Sicherheitsniveau für alle Dienstleister

Quelle: akitada31

Cloudanwendungen II – Drittstaaten

- Hosting- oder Cloudprovider liegen oft in den USA oder in anderen Drittländern
- Auslagerung der Hosting-Services an Subunternehmer in Drittländern ist eine Übermittlung
- Zugriffsmöglichkeit grundsätzlich ausreichend; Speicherort allein löst das Problem nicht.
- Art. 45, 46 DSGVO: Daten dürfen nur in Drittländer übermittelt werden, wenn ein Angemessenheitsbeschluss vorliegt oder geeignete Garantien vorliegen
- EuGH Schrems II Urteil
  - Personenbezogene Daten dürfen nur in Drittländer übermittelt werden, wenn der dortige Schutz dem der EU entspricht
  - In den USA besteht kein angemessenes Schutzniveau
  - Standarddatenschutzklauseln reichen als Garantie alleine nicht aus
  - Zusätzliche Maßnahmen sind erforderlich: nach Empfehlung der EDSA können das vertragliche, technische oder organisatorische Maßnahmen sein
- Trans-Atlantic Data Privacy Framework ab 2023

Quelle: Pixaline

VI. Fazit und Best Practices

Fazit und Best Practices

- Sensibilität der Nutzergruppen sicherstellen
- Rechtsgrundlagen sicherstellen
- Prozesse für Ingest & Datenbeschaffung etablieren und Verantwortlichkeiten festlegen
- Konzepte für Betroffenenrechte aufstellen & Tools prüfen / Metadatenmanagement
- Data Governance: Dokumentation sicherstellen
- Geeigneten Cloud Anbieter wählen
- Ggf. DSFA durchführen
- KI-VO frühzeitig in den Blick nehmen

Vielen Dank für Ihre Aufmerksamkeit!

Referenten

Philipp Müller-Peltzer

mueller-peltzer@srd-rechtsanwaelte.de

Rechtsanwalt, Partner

Ilan Leonard Selz

selz@srd-rechtsanwaelte.de

Rechtsanwalt, Senior Associate

Kontakt

SCHÜRMANN
ROSENTHAL
DREYER
RECHTSANWÄLTE

DIGITALES BUSINESS . TECHNOLOGIE . MEDIEN

Schürmann Rosenthal Dreyer Rechtsanwälte

Am Hamburger Bahnhof 4
10557 Berlin
Deutschland

Tel: +49 (0)30 213 002 80
Fax: +49 (0)30 213 002 849

info@srd-rechtsanwaelte.de
www.srd-rechtsanwaelte.de