T. Grust et al. (Hrsg.): Datenbanksysteme für Business, Technologie und Web (BTW 2019),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 435

Ganzheitliches Metadatenmanagement im Data Lake: Anforderungen, IT-Werkzeuge und Herausforderungen in der Praxis

Christoph Gröger[1], Eva Hoos[2]

Abstract: Data Lakes haben sich in der industriellen Praxis als Plattformen für die Speicherung und Analyse aller Arten von (Roh-)daten etabliert. Erweiterte Anforderungen hinsichtlich Governance und Self-Service machen das Metadatenmanagement im Data Lake zum kritischen Erfolgsfaktor. Bisher gibt es dazu jedoch nur wenige wissenschaftliche Arbeiten, es mangelt insbesondere an einer ganzheitlichen Betrachtung zur Konzeption und Realisierung des Metadatenmanagements im Data Lake. Diese Arbeit adressiert das Thema und basiert auf praktischen Erfahrungen aus einem Industriekonzern beim Aufbau eines unternehmensweiten Data Lake. Es werden praktische Anforderungen und Anwendungsbeispiele für das Metadatenmanagement im Data Lake diskutiert und die unterschiedlichen Arten von Metadaten anhand des Praxisbeispiels analysiert. Zur Umsetzung des Metadatenmanagements werden anschießend unterschiedliche IT-Werkzeuge anhand definierter Kriterien analysiert. Das Analyseergebnis zeigt, dass Datenkataloge grundsätzlich die geeignete Werkzeugart darstellen, wobei noch technische Unzulänglichkeiten existieren. Abschließend werden die in der Praxis bestehenden Herausforderungen für ein ganzheitliches Metadatenmanagement im Data Lake zusammengefasst und zukünftige Forschungsbedarfe aufgezeigt.

Keywords: Metadaten, Meta Data, Data Lake, Datenkatalog, Data Catalog, Governance, Self-Service

1 Einleitung

Die digitale Transformation, Big Data und Advanced Analytics verändern die Datenlandschaft in Unternehmen erheblich [Lv17]. Die großen Mengen heterogener Daten sowie die Vielzahl unterschiedlicher analytischer Anwendungsfälle erfordern neue Konzepte und Plattformen für das Datenmanagement [OL13]. In diesem Zuge hat sich der Data Lake in den letzten Jahren als ein neuer Typ von Datenplattform für die Speicherung, Integration und Analyse aller Arten von (Roh-)daten etabliert [Ma17]. Data Lakes erfreuen sich in der industriellen Praxis zunehmender Beliebtheit und werden in unterschiedlichen Branchen eingesetzt. Ein prominentes Anwendungsfeld für Data Lakes stellt die Fertigungsindustrie dar, da im Zuge der Industrie 4.0 [Ba14], also der Digitalisierung der industriellen Wertschöpfungskette, enorme Datenmengen generiert und ausgewertet werden [GCA15]. Es geht beispielsweise um die Mustererkennung in Maschinendaten zur Optimierung von

[1] Robert Bosch GmbH, 70469 Stuttgart, christoph.groeger@de.bosch.com
[2] Robert Bosch GmbH, 70469 Stuttgart, eva.hoos@de.bosch.com

doi:10.18420/btw2019-26

Fertigungsprozessen oder die Ad-hoc-Exploration von Produktentwicklungs- und Felddaten zur Verbesserung des Produktdesigns [Gr18].

Mit dem zunehmenden Einsatz von Data Lakes ergeben sich in der industriellen Praxis erweiterte Anforderungen, insbesondere hinsichtlich der Sicherstellung von Transparenz, Qualität und Compliance der Daten im Data Lake sowie der Unterstützung von Self-Service-Szenarien für Fachanwender. Diese Anforderungen machen Metadatenmanagement[3] im Data Lake zum kritischen Erfolgsfaktor [QHV16, HGQ16]. Es soll damit verhindert werden, dass aus dem Data Lake ein Datensumpf (engl. data swamp) [HGQ16] entsteht, also eine Datenplattform mit nicht mehr sinnvoll nutzbaren Daten.

Bisher gibt es jedoch nur wenige wissenschaftliche Arbeiten zum Metadatenmanagement im Data Lake. Existierende Arbeiten fokussieren auf einzelne Teilaspekte, insbesondere die allgemeine Bedeutung von Metadaten im Data Lake [Te15, MT16, Sh18] sowie die Erfassung und Anreicherung von Metadaten im Data Lake mittels semantischer Technologien [HGQ16, QHV16, Al16]. Es mangelt an einer ganzheitlichen Betrachtung zur Konzeption und Realisierung des Metadatenmanagements im Data Lake, insbesondere hinsichtlich in der Praxis relevanter Anforderungen, zu adressierender Arten von Metadaten sowie passender IT-Werkzeuge zur Umsetzung.

Die vorliegende Arbeit adressiert diese Aspekte und basiert auf praktischen Erfahrungen aus einem weltweit operierenden Industriekonzern beim Aufbau eines unternehmensweiten Data Lake mit ganzheitlichem Metadatenmanagement. Zuerst werden in Kapitel 2 das Unternehmen sowie die Data-Lake-Architektur vorgestellt, die technisch auf einem Cross-Plattform-Ansatz basiert. Zusätzlich wird das ganzheitliche Metadatenmanagement im Data Lake als Ziel dargestellt. Auf dieser Grundlage werden in Kapitel 3 Kernanforderungen und praktische Anwendungsbeispiele für das Metadatenmanagement im Data Lake abgeleitet und analysiert. Davon ausgehend werden in Kapitel 4 die unterschiedlichen Arten relevanter Metadaten im Kontext von Data Lakes anhand des Praxisbeispiels identifiziert und diskutiert. Anschließend geht es in Kapitel 5 um die praktische Umsetzung des Metadatenmanagements mittels unterschiedlicher IT-Werkzeuge. Für die Auswahl geeigneter IT-Werkzeuge werden zuerst produktunabhängige Werkzeugarten definiert und anhand von aus den Anforderungen abgeleiteten Bewertungskriterien analysiert. Das Ergebnis zeigt, dass Datenkataloge grundsätzlich die geeignete Werkzeugart für das ganzheitliche Metadatenmanagement im Data Lake darstellen, wobei noch technische Unzulänglichkeiten existieren. Abschließend werden in Kapitel 6 die in der Praxis bestehenden technischen und organisatorischen Herausforderungen für ein ganzheitliches Metadatenmanagement im Data Lake zusammengefasst und zukünftige Forschungsbedarfe aufgezeigt. Die Arbeit schließt mit einem Fazit in Kapitel 7.

[3] Unter dem Begriff Metadaten werden im Kontext des Datenmanagements allgemein Daten zur Verwaltung und Nutzung von Daten verstanden [VVS00]. Metadatenmanagement bezieht sich auf die systematische Verwaltung und Bereitstellung von Metadaten [He17].

2 Praxisbeispiel: Unternehmensweiter Data Lake und ganzheitliches Metadatenmanagement

Die vorliegende Arbeit basiert auf Erfahrungen aus einem realen Praxisbeispiel aus der Fertigungsindustrie, das die Grundlage für die Ableitung der Anforderungen sowie die Analyse relevanter Arten von Metadaten im weiteren Verlauf der Arbeit darstellt. Im Folgenden wird zuerst das zugrundeliegende Unternehmen kurz vorgestellt, um anschließend auf den Ansatz des unternehmensweiten Data Lake mit ganzheitlichem Metadatenmanagement einzugehen.

Das Praxisbeispiel bezieht sich auf einen *global tätigen Industriekonzern* mit mehreren hunderttausend Mitarbeitern und einer weltweit verteilten Fabriklandschaft. Der Industriekonzern gliedert sich in mehrere Geschäftsbereiche, die ein breites Spektrum an Produkten entwickeln und fertigen, speziell in den Domänen Industrie- und Gebäudetechnik, Mobilität und Konsumgüter. Dementsprechend vielfältig gestaltet sich die Prozesslandschaft des Konzerns, von der Massenfertigung hoch standardisierter Produkte bis zur Einzelfertigung von Spezialprodukten nach Kundenanforderung.

Ein wesentliches strategisches Ziel des Konzerns ist die digitale Transformation zum datengetriebenen Unternehmen als Teil der Industrie 4.0. Unterschiedliche Digitalisierungsinitiativen des Industriekonzerns haben in den letzten Jahren zu enormen Mengen heterogener Daten entlang der Wertschöpfungskette geführt. Diese Daten sollen insbesondere genutzt werden, um Produkte und Prozesse ganzheitlich zu optimieren. Es geht beispielsweise um Produktlebenszyklus-Analysen [Ka15], um Produktentwicklungs-, Fertigungs-, Feld- und Kundendaten umfassend zu analysieren. Ein weiteres Anwendungsfeld stellt die Ende-zu-Ende-Analyse von Geschäftsprozessen im Rahmen von Process Mining [Aa16] dar, um Engpässe und Wartezeiten zu eliminieren und die Prozessqualität zu erhöhen. Diese Anwendungsfelder erfordern die unternehmensweite Integration von klassischen transaktionalen Unternehmensdaten, speziell aus Enterprise Ressource Planning (ERP) Systemen, mit Maschinen-, Sensor- und Kundendaten, z.B. aus Manufacturing Execution Systemen (MES) und dem Web (siehe [GSMb14, GSMa14] zur Datenintegration in Industrieunternehmen sowie [GP14] für eine Beschreibung der IT-Systeme in Industrieunternehmen).

Zu diesem Zweck wird ein *unternehmensweiter Data Lake* aufgebaut, der die unternehmensweite Speicherung, Integration und Analyse aller Arten von (Roh-)daten unterstützt. Eine vereinfachte Darstellung der konzeptionellen Architektur des Data Lake zeigt Abb. 1. Die Architektur basiert auf dem Lambda-Architektur-Paradigma [MW15] und umfasst Komponenten zur Batch- und Streaming-Datenverarbeitung zur Umsetzung unterschiedlicher Analyseanwendungen, von Reporting über Exploration bis zu Data Mining (siehe [HKP12, KBM10] für eine Beschreibung der Analyseanwendungen). Die Quelldaten umfassen sämtliche betrieblich relevanten strukturierten und unstrukturierten Daten, die in vier Kategorien unterteilt werden: klassische transaktionale Unternehmensdaten, benutzergenerierte Daten, Maschinen- und Sensordaten sowie Webdaten.

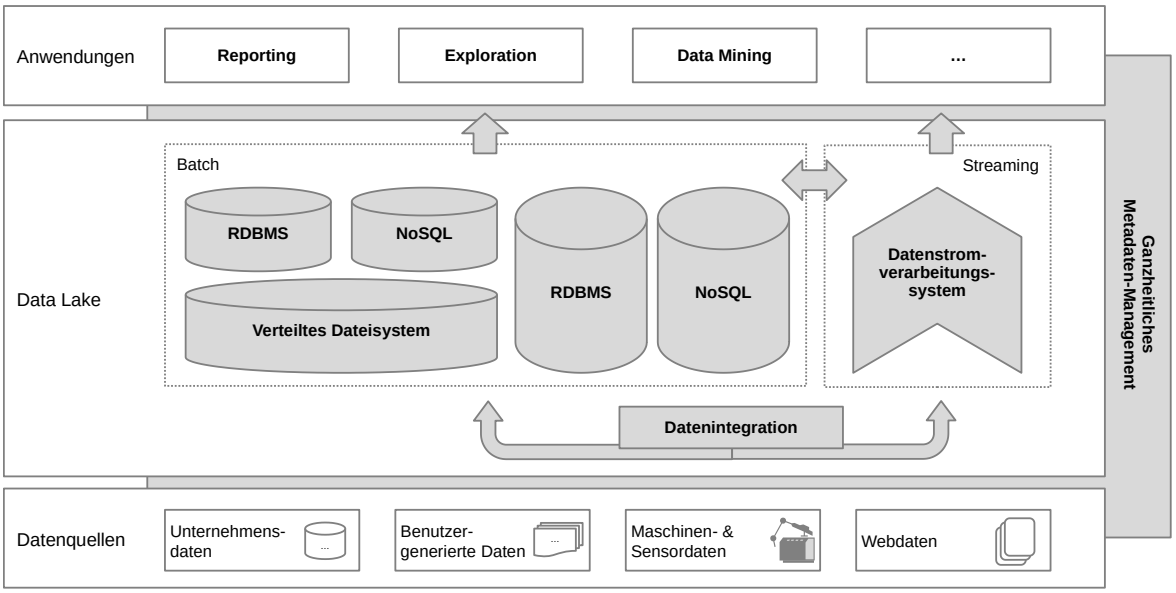

Abb. 1: Konzeptionelle Data-Lake-Architektur des Industriekonzerns (vereinfacht)

Eine wesentliche Besonderheit des Data Lake besteht darin, dass ein umfassender Cross-Plattform-Ansatz verfolgt wird. Dies bezieht sich auf folgende Merkmale:

- *Kombination von Batch- und Streaming-Komponenten*, um sowohl Stapelverarbeitung als auch echtzeitnahe Datenverarbeitung zu unterstützen
- *Kombination unterschiedlicher Datenhaltungstechnologien*, insbesondere von relationalen Datenbanken, spalten- und dokumentenorientierten NoSQL-Datenbanken sowie verteilten Dateisystemen, um eine polyglotte Persistenz [GR15] umzusetzen und die für den jeweiligen Anwendungsfall passende Datenhaltungstechnologie zu nutzen
- *Kombination unterschiedlicher Bereitstellungsvarianten der Komponenten*, um sowohl On-Premise- als auch Cloud-Bereitstellungen zu nutzen

Eine zentrale Komponente des Data Lake, die sich über sämtliche Schichten und Komponenten erstreckt, ist das Metadatenmanagement. Das Ziel ist, ein *ganzheitliches Metadatenmanagement* für den Data Lake zu realisieren, das sämtliche Kernanforderungen (siehe Kapitel 3), sämtliche relevanten Arten von Metadaten (siehe Kapitel 4) über sämtliche Datenhaltungssysteme des Data Lake mit geeigneten IT-Werkzeugen (siehe Kapitel 5) unterstützt. Diese Aspekte stehen in den folgenden Kapiteln der Arbeit im Vordergrund.

3 Kernanforderungen und Anwendungsbeispiele für Metadatenmanagement im Data Lake

Auf der Basis der praktischen Erfahrungen im Industriekonzern lassen sich die *Unterstützung von Self-Service* (siehe Kapitel 3.1) sowie die *Unterstützung von Governance* (siehe Kapitel 3.2) im Data Lake als Kernanforderungen für das Metadatenmanagement im Data Lake identifizieren. Diese werden im Folgenden beschrieben und anhand praktischer Anwendungsbeispiele aus dem Industriekonzern konkretisiert (siehe Abb. 2).

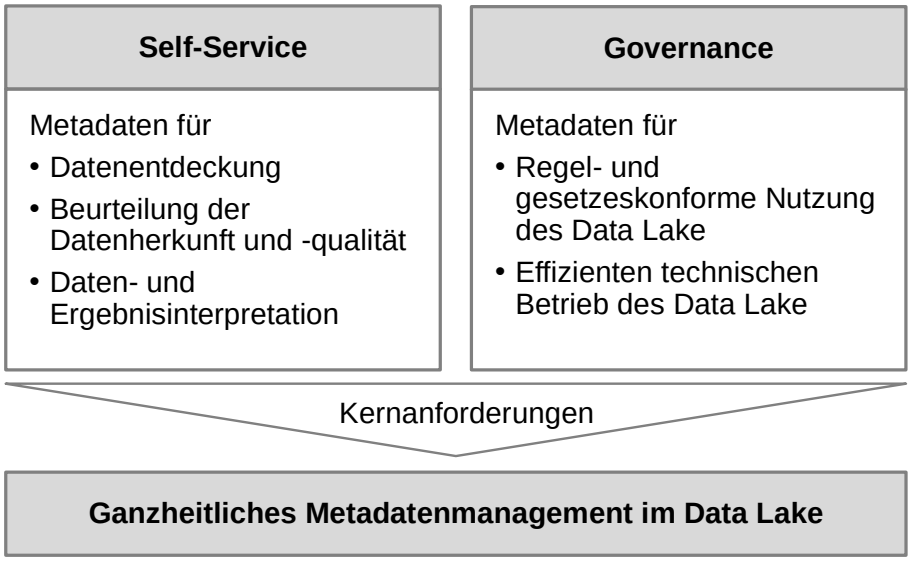

Abb. 2: Kernanforderungen für das Metadatenmanagement im Data Lake

3.1 Unterstützung von Self-Service

Ein wesentliches Ziel des unternehmensweiten Data Lake des Industriekonzerns besteht darin, eine zentrale Datenplattform für sämtliche Arten von Endbenutzern zu realisieren, um eine Demokratisierung, d.h. eine möglichst umfassende Durchdringung sämtlicher Prozesse und Entscheidungsebenen des Konzerns mit datengetriebenen Methoden und Anwendungsfällen, zu erreichen. Es geht folglich darum, nicht nur Data Scientists und IT-Experten als kleinere Gruppen spezialisierter Data-Lake-Nutzer zu adressieren, sondern auch Fachanwendern, z.B. Prozessingenieuren und Marketing-Spezialisten, die Nutzung des Data Lake zu ermöglichen. In diesem Rahmen strebt der Industriekonzern auch an, die Rolle des Citizen Data Scientist [Gr18] im Sinne von Fachanwendern mit vertieften Kenntnissen in Data Science zu etablieren. Ein typisches Anwendungsbeispiel umfasst einen Prozessingenieur, der als Citizen Data Scientist unterschiedliche Analysen der Fertigungsqualität durchführt.

Für diese Demokratisierung der Data-Lake-Nutzung sind insbesondere Konzepte und IT-Werkzeuge für Self-Service-Szenarien relevant, die unter den Schlagwörtern Self-Service-Business-Intelligence [AS16] und Self-Service-Analytics [Di16] diskutiert werden.

Das Ziel besteht allgemein darin, Fachanwendern die Datenaufbereitung und die Datenanalyse mit einfach nutzbaren IT-Werkzeugen zu ermöglichen, ohne dass eine umfassende Unterstützung durch IT-Experten erforderlich ist. Metadaten spielen eine zentrale Rolle bei der Umsetzung von Self-Service-Szenarien [Te15]. Ausgehend von den praktischen Erfahrungen im Industriekonzern geht es hierbei insbesondere um die Unterstützung der Datenentdeckung, der Beurteilung der Datenherkunft und Datenqualität sowie der Daten- und Ergebnisinterpretation durch Fachanwender.

Bei der *Datenentdeckung* (engl. data discovery) steht die Identifikation relevanter Datenbereiche im Data Lake im Vordergrund, um sämtliche für eine spezifische Fragestellung relevanten Daten zu selektieren. Der Prozessingenieur sucht beispielsweise sämtliche Qualitätsdaten zu einem bestimmten Produkt. Es geht einerseits um MES-Daten zu Qualitätsprüfungen, die in einer relationalen Datenbank des Data Lake gespeichert sind. Andererseits geht es um Sensordaten aus Fertigungsmaschinen, die in einem verteilten Dateisystem des Data Lake abgelegt sind. Metadaten ermöglichen nun eine einfache Identifikation und Wiederauffindung dieser Daten im Data Lake, indem z.B. sämtliche Qualitätsdaten im Data Lake einheitlich gekennzeichnet sind. Der Prozessingenieur erhält anhand einer metadatenbasierten Suche beispielsweise eine Liste von Tabellen sowie Dateien im Data Lake, die Qualitätsdaten enthalten, und kann diese nun explorieren.

Zur *Beurteilung der Datenherkunft* (engl. data provenance) und der *Datenqualität* sind diverse Metadaten erforderlich. Es geht beispielsweise um Details zu den Quellsystemen der Daten, zu erfolgten Transformationsprozessen sowie um Datenprofile, z.B. Werteverteilungen und Statistiken. Metadaten zur Herkunft und Qualität der Daten sind ein kritischer Erfolgsfaktor, um eine zuverlässige und vertrauenswürdige Nutzung der Daten im Data Lake zu ermöglichen. Der Prozessingenieur untersucht beispielsweise anhand von Metadaten, ob ein Teil der MES-Daten aus einem produktiven MES oder einem Testsystem stammt. Zudem prüft er, welche Transformationen der Daten im Data Lake vorgenommen werden, da die MES-Daten bereits als vorverarbeitete aggregierte Tabellen zur Verfügung gestellt werden. Darüber hinaus beurteilt der Prozessingenieur die Sensordaten hinsichtlich ihrer Datenqualität anhand des Anteils fehlender Sensorwerte.

Zur *Daten- und Ergebnisinterpretation* sind insbesondere Details zur fachlichen Bedeutung der Daten relevant. Es geht beispielsweise um die Bedeutung betriebswirtschaftlicher Kennzahlen, die Abgrenzung von Fachbegriffen im Rahmen von Glossaren und die Struktur von Aggregationshierarchien. All diese Details stellen zentrale Metadaten zur Dateninterpretation dar und können sich auch auf Analyseergebnisse, wie z.B. Dashboards und Berichte beziehen. Zur Auswertung der MES-Daten benötigt der Prozessingenieur beispielsweise Angaben zur fachlichen Bedeutung einzelner Attribute der Tabellen, da diese keine sprechenden Namen, sondern nur technische Kürzel enthalten. Außerdem fügt der Prozessingenieur neue Tabellen in der relationalen Datenbank des Data Lake hinzu, um die Sensordaten mit den MES-Daten zu integrieren und zu bereinigen. Diese Tabellen enthalten auch berechnete Kennzahlen auf der Basis von MES- und Sensordaten, die in einem Dashboard grafisch veranschaulicht werden. Die fachliche Bedeutung dieser Kennzahlen dokumentiert der

Prozessingenieur anhand von Metadaten und kennzeichnet die neu erstellten Tabellen und das Dashboard mittels Metadaten, um die Ergebnisse wiederauffindbar zu machen und zukünftige Qualitätsanalyen damit zu unterstützen.

Alles in allem geht es beim Metadatenmanagement zur Unterstützung von Self-Service-Szenarien sowohl um eine ganzheitliche Verwaltung und Bereitstellung der Metadaten für Fachanwender entlang des gesamten Datenaufbereitungs- und Analyseprozesses als auch darum, zusätzliche Metadaten durch Fachanwender zu erfassen und wiederverwendbar zu machen.

3.2 Unterstützung von Governance

Mit zunehmender Nutzung des Data Lake gewinnen organisatorische und technische Regelungsanforderungen, d.h. Governance-Anforderungen, im Industriekonzern an Bedeutung. Diese werden allgemein auch unter dem Stichwort Data Governance [KB10] diskutiert.

Zum einen geht es darum, die *regel- und gesetzeskonforme Nutzung des Data Lake* (engl. compliance) sicherzustellen. Ein wesentlicher Faktor sind *Anforderungen aus der EU-Datenschutz-Grundverordnung (DSGVO)* [Eu16], die eine transparente und zweckgebundene Verarbeitung personenbezogener Daten vorschreibt und nachweispflichtig macht. Dies erfordert eine umfassende und ständig aktuell gehaltene Metadatendokumentation über sämtliche Batch- und Streaming-Komponenten im Data Lake. Es geht insbesondere darum, sämtliche Datenbereiche mit personenbezogene Daten im Data Lake zu inventarisieren, also beispielsweise alle relevanten Schemata und Tabellen, Verzeichnisstrukturen und Message Queues. Ein praktisches Anwendungsbeispiel sind Anforderungen von Kunden zur Löschung ihrer personenbezogenen Daten, z.B. Adressdaten oder Produktnutzungsdaten, die mittels entsprechender Metadaten im Data Lake effektiv umgesetzt werden können.

Ein weiterer Faktor sind *Regelungsanforderungen, die sich aus der digitalen Transformation* des Industriekonzerns ergeben. Daten sind ein zentraler Wertgegenstand (engl. asset) im datengetriebenen Unternehmen und erfordern ein systematisches Datenqualitätsmanagement sowie organisatorische Verantwortlichkeiten und Prozesse. Es geht insbesondere um unternehmensinterne Regelungen für Dateneigentum und -nutzung, um beispielsweise zu dokumentieren, welche Geschäftsdivisionen des Industriekonzerns welche Daten im unternehmensweiten Data Lake bereitstellen und verantworten. Hierbei sind auch umfassende Details zur Datenherkunft relevant. Sämtliche Informationen rund um die Herkunft, Qualität und um die organisatorischen Verantwortlichkeiten der Daten sind Metadaten im Data Lake, die systematisch verwaltet werden müssen.

Zum anderen geht es bei Governance-Anforderungen darum, einen *effizienten technischen Betrieb des Data Lake* zu gewährleisten. Der im Industriekonzern verfolgte Cross-Plattform-Ansatz mit seiner Vielzahl an Datenhaltungssystemen bedingt *Datenredundanzen im Data Lake, die systematisch verwaltet und transparent gemacht werden* müssen, um Inkonsistenzen

zu vermeiden und die Speicherkosten im Data Lake zu optimieren. Datenredundanzen entstehen beispielsweise bei der kombinierten Analyse von Stammdaten zu Maschinen und Produkten mit Sensordaten. Die Stammdaten werden im Data Lake des Industriekonzerns meist in relationalen Datenbaken gehalten, wohingegen die Sensordaten in dokumenten- oder spaltenorientierten Datenbanken abgelegt sind. Zur kombinierten Analyse werden die Daten in einer Datenbank zusammengeführt, um darauf z.B. Data-Mining-Verfahren anzuwenden. Die daraus resultierenden Redundanzen sind zu dokumentieren. Darüber hinaus sind für einen effizienten Betrieb des Data Lake *effiziente Änderungsprozesse bei Änderungen der Quellsysteme* wesentlich. Änderungen bei den Quellsystemen des Data Lake ergeben sich beispielsweise durch Systemkonsolidierungen oder Softwareaktualisierungen. Diese Änderungen müssen im Data Lake nachvollzogen werden, um z.B. Transformationsprozesse anzupassen und die Datenbereitstellung im Data Lake aufrecht zu erhalten. Die aufgeführte Transparenz über Datenredundanzen sowie über Auswirkungen von Quellsystemänderungen basiert auf entsprechenden Metadaten zu Datenbereichen sowie Quellsystemen des Data Lake, die es zu verwalten gilt.

Summa summarum stellt ein ganzheitliches Metadatenmanagement im Data Lake ein zentrales Instrument zur Umsetzung verschiedenster Governance-Anforderungen dar und ist deswegen in der industriellen Praxis meist Teil umfangreicher Data-Governance-Projekte.

4 Heterogene Metadaten im Kontext von Data Lakes

In der Literatur werden unterschiedliche allgemeine Klassifikationen für Metadaten in analytischen IT-Systemen beschrieben, insbesondere aktive und passive Metadaten sowie technische und betriebswirtschaftliche Metadaten [KBM10]. Diese Klassifikationen erfassen allerdings nur einen Teil der im Praxisbeispiel relevanten Metadaten. Auf der Basis der im Industriekonzern durchgeführten Workshops zum Metadatenmanagement lassen sich die in Abb. 3 dargestellten Arten von Metadaten im Data-Lake-Kontext ableiten, die nachfolgend beschrieben und diskutiert werden.

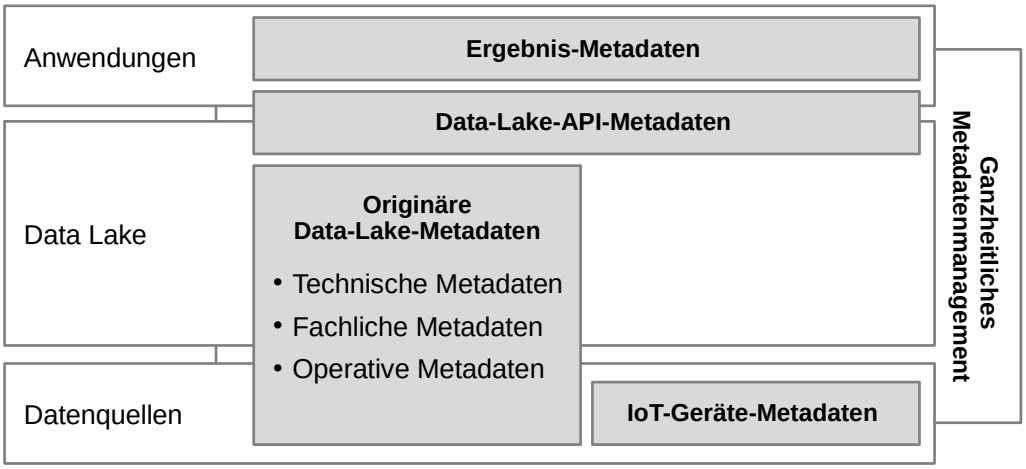

Abb. 3: Metadaten im Kontext von Data Lakes

- *Originäre Data-Lake-Metadaten* sind Metadaten über die im Data Lake gespeicherten Daten, z.B. Daten in relationalen Datenbanken und Message-Queues des Data Lake. Für originäre Data-Lake-Metadaten hat sich in der industriellen Praxis die Unterscheidung in technische, fachliche und operative Metadaten etabliert [He17]:
  - *Technische Metadaten* beziehen sich auf sämtliche technisch-strukturellen Aspekte der Daten sowie der zugrundeliegenden Datenhaltungssysteme im Data Lake, wie z.B. Tabellenstrukturen, Attributnamen, Wertelisten und Zugriffsrechte.
  - *Fachliche Metadaten* beschreiben inhaltliche Bedeutungen und Zusammenhänge der Daten. Dazu gehören beispielsweise Begriffsabgrenzungen, Kennzahlendefinitionen, organisatorische Verantwortlichkeiten und konzeptionelle Datenmodelle.
  - *Operative Metadaten* umfassen technische Details zu Datentransformationen und Datenzugriffen, z.B. Informationen über ETL-Jobs, Quellsysteme und Zugriffsmuster.
- *Ergebnis-Metadaten* stellen Metadaten zu erzeugten Analyseergebnissen, wie z.B. Kennzahlenberichte, Dashboards und Data-Mining-Modelle, dar. Es geht sowohl um Angaben zum Erstellungsprozess der Analyseergebnisse, z.B. verwendete Parameterwerte von Data-Mining-Algorithmen, als auch um Interpretationen der Analyseergebnisse, z.B. die Kennzeichnung bestimmter Knoten in einem Entscheidungsbaum.
- *IoT-Geräte-Metadaten* sind Metadaten über Geräte im Internet der Dinge, wie z.B. Smart-Home-Geräte oder intelligente Fertigungsmaschinen, die Datenquellen für den Data Lake darstellen. IoT-Geräte-Metadaten umfassen Daten zur Einbindung, Verwaltung und Steuerung von IoT-Geräten, beispielsweise ID, Typ, Status und Hersteller eines IoT-Geräts. IoT-Geräte-Metadaten sind folglich von Metadaten über die von IoT-Geräten bereitgestellten Sensordaten, z.B. Datentypen von Sensorwerten, zu unterscheiden. Metadaten über Sensordaten von IoT-Geräten, die im Data Lake verarbeitet werden, werden als originäre Data-Lake-Metadaten betrachtet.
- *Data-Lake-API-Metadaten* sind Metadaten zu den vom Data Lake bereitgestellten Programmierschnittstellen (engl. application programming interface (API)). Diese ermöglichen einen programmatischen Zugriff auf Daten im Data Lake, typischerweise mittels Representational State Transfer (REST). Data-Lake-API-Metadaten umfassen beispielsweise Angaben zu Methoden und Parametern einer API.

Den Kern für das Metadatenmanagement im Data Lake bilden originäre Data-Lake-Metadaten sowie Ergebnis-Metadaten. Ergebnis-Metadaten werden häufig in separaten Analysewerkzeugen, z.B. Data-Mining-Werkzeugen oder Business-Intelligence-Werkzeugen, erzeugt und verwaltet. Wenn Analyseergebnisse wiederum als Daten im Data Lake gespeichert werden, indem beispielsweise Data-Mining-Modelle als Dateien im Predictive-Model-Markup-Language-Format (PMML-Format) im Data Lake abgelegt werden, verschwimmen

die Grenzen zwischen originären Data-Lake-Metadaten und Ergebnis-Metadaten. Eine integrierte Betrachtung dieser Metadaten ist folglich sinnvoll.

Da Data-Lake-APIs den Zugriff auf Daten im Data Lake ermöglichen, ist eine Verknüpfung von Data-Lake-API-Metadaten mit originären Data-Lake-Metadaten sowie Ergebnis-Metadaten vielversprechend. Die Rückgabewerte einer API-Methode können beispielsweise direkt auf fachliche Metadaten der zurückgelieferten Daten verweisen, um die gezielte Verwendung der Daten zu unterstützen.

IoT-Geräte stellen zwar nur eine von mehreren Arten von Quellsystemen des Data Lake dar. Der Unterschied zu traditionellen Quellsystemen wie relationalen Datenbanksystemen besteht aus Data-Lake-Sicht jedoch darin, dass zur effektiven Analyse von Sensordaten von IoT-Geräten häufig IoT-Geräte-Metadaten erforderlich sind. Beispielsweise sind zur Analyse von Sensordaten einer Messstation einer Fertigungslinie Angaben zur Positionierung der Messstation an der Fertigungslinie sowie zum aktuellen Software-Stand der Messstation erforderlich.

Summa summarum sind Metadaten im Data-Lake-Kontext sehr heterogen und stark unterschiedlich hinsichtlich Struktur, erzeugenden Systemen und Verwendungszwecken. Hinzukommt, dass Metadaten auch analytisch gewonnen werden und selbst wiederum Quelldaten für weitere Analysen darstellen. Ein typischer Anwendungsfall im betrachteten Industriekonzern ist beispielsweise die analytische Gewinnung von Metadaten zu Videos aus autonomen Fahrzeugen. Diese Metadaten stellen wiederum die Grundlage für Dashboards und Berichte zur Auswertung des Fahrverhaltens dar.

5 IT-Werkzeuge zur praktischen Umsetzung des Metadatenmanagements im Data Lake

Die Umsetzung eines ganzheitlichen Metadatenmanagements im Data Lake erfordert eine integrierte Betrachtung sämtlicher in Kapitel 4 aufgeführter Arten von Metadaten zur Adressierung der in Kapitel 3 beschriebenen Kernanforderungen um Self-Service und Governance. Das Metadatenmanagement im Data Lake ist damit deutlich umfangreicher und komplexer als das Metadatenmanagement im klassischen Data Warehouse, das hauptsächlich Metadaten zu strukturierten transaktionalen Unternehmensdaten umfasst. Diese Komplexität spiegelt sich auch im Markt für IT-Werkzeuge für das Metadatenmanagement im Data Lake wieder. Der Markt ist sehr heterogen und entwickelt sich sehr dynamisch, da verschiedene Hersteller unterschiedlich ausgerichtete Produkte anbieten und meist ähnliche Funktionalitäten unter verschiedenen, marketinglastigen Begrifflichkeiten referenziert werden.

Als Ausgangspunkt für die praktische Umsetzung im Industriekonzern wurde deswegen eine umfangreiche Softwarestudie durchgeführt. In diesem Rahmen werden produktunabhängige Arten von Werkzeugen für das Metadatenmanagement im Data Lake definiert, die als Strukturierungselemente für die Marktanalyse dienen. Darüber hinaus werden

Bewertungskriterien entwickelt, die sich aus den dargestellten Kernanforderungen für das Metadatenmanagement sowie den relevanten Arten an Metadaten ableiten. Anhand dieser Kriterien wird anschließend die Eignung der Werkzeugarten für das Metadatenmanagement im Data Lake analysiert. Im Folgenden werden die im Rahmen der Studie identifizierten Arten an IT-Werkzeugen für das Metadatenmanagement im Data Lake vorgestellt (siehe Kapitel 5.1), um anschließend deren Eignung zu diskutieren (siehe Kapitel 5.2). Anzumerken ist, dass aus Vertraulichkeitsgründen nicht näher auf einzelne Produkte bestimmter Hersteller eingegangen werden kann.

5.1 Werkzeugarten: Datenverzeichnisse, Datenkataloge und Data-Lake-Management-Plattformen

Im Rahmen der Studie werden ausschließlich produktiv einsetzbare Werkzeuge, sowohl auf Open-Source- als auch auf Closed-Source-Basis, analysiert. Forschungsprototypen werden nicht einbezogen, da sie sich nicht für den Aufbau einer praxistauglichen Lösung eignen. Die Analyse ergibt drei Arten von Werkzeugen für das Metadatenmanagement im Data Lake, die den Markt repräsentieren: systemintegrierte Datenverzeichnisse, Datenkataloge sowie Data-Lake-Management-Plattformen (siehe Abb. 4).

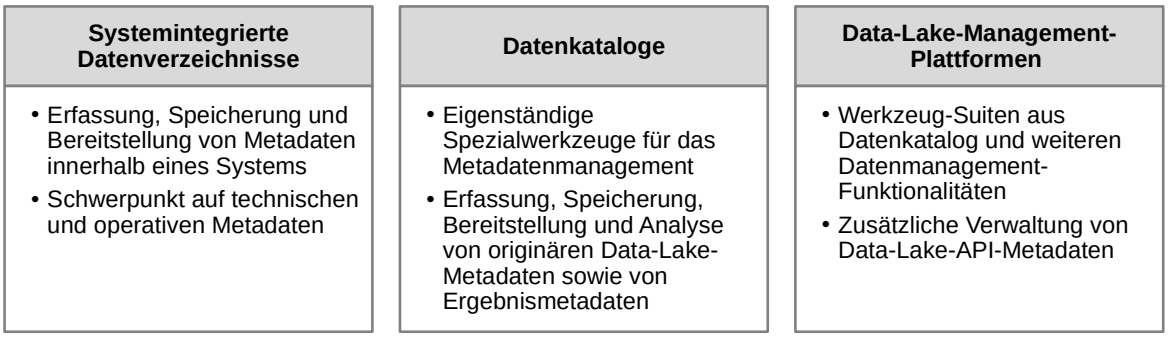

Abb. 4: Arten von IT-Werkzeugen für das Metadatenmanagement im Data Lake

Systemintegrierte Datenverzeichnisse (engl. data dictionary) [KE06] sind in einem Datenhaltungssystem integriert und dienen der Erfassung, Speicherung und Bereitstellung von Metadaten innerhalb dieses Systems. Der Schwerpunkt liegt auf technischen und operativen Metadaten. Typische Beispiele sind Datenverzeichnisse in relationalen Datenbanksystemen, die z.B. Details zu Schemata, Tabellen und Attributen enthalten. Im Hadoop- und NoSQL-Kontext haben sich nach diesem Vorbild ähnliche Ansätze etabliert, wie z.B. der Hive Metastore als Datenverzeichnis für Apache Hive [Ap18a].

Datenkataloge (engl. data catalog) [Sh18] stellen eigenständige Spezialwerkzeuge für das Metadatenmanagement unabhängig von einem spezifischen Datenhaltungssystem dar. Sie ermöglichen die Erfassung, Speicherung, Bereitstellung sowie die Analyse von technischen, fachlichen und operativen Metadaten sowie von Ergebnismetadaten. Dabei unterstützen sie meist eine Vielzahl an Datenhaltungssystemen zur Metadaten-Erfassung, von relationalen

Datenbanken über NoSQL-Datenbanken bis zu Datenhaltungssystemen aus dem Hadoop-Kontext. Datenkataloge basieren auf einem Metadaten-Repository und bieten typischerweise Funktionen zur systemübergreifenden Analyse des Datenverlaufs (engl. data lineage) und der Datenqualität sowie zur fachlichen Klassifikation von Metadaten, beispielsweise mittels Tagging und Glossaren. Es werden darüber hinaus auch Kollaborationsfunktionen zur Einbindung der Endbenutzer bei der Anreichung und Nutzung von Metadaten angeboten, z.B. durch Bewertungs- und Kommentarfunktionen. Beispielshafte Werkzeuge dieser Art sind Waterline Smart Data Catalog [Wa18], Informatica Enterprise Data Catalog [In18] und Collibra Catalog [Co18]. Auch Datenkataloge wie Apache Atlas [Ap18b], die auf unterschiedliche Datenhaltungssysteme im Hadoop-Kontext fokussieren, gehören zu dieser Art.

Data-Lake-Management-Plattformen stellen Werkzeugen-Suiten dar, die auf der Basis eines Datenkatalogs weitere Funktionalitäten für das Datenmanagement im Data Lake integrieren. Typischerweise geht es um ergänzende Funktionalitäten für ETL, Self-Service-Data-Preparation und Datenföderation (engl. data federation), die eng mit dem Datenkatalog integriert sind. Beispielsweise werden Metadaten zu Datentransformationen in ETL-Jobs direkt im Datenkatalog abgelegt. Data-Lake-Management-Plattformen unterstützen häufig auch die Definition von APIs zum Zugriff auf Daten im Data Lake und ermöglichen damit zusätzlich die Verwaltung von Data-Lake-API-Metadaten. Beispielhafte Data-Lake-Management-Plattformen sind Kylo [Te18] und Iguazio [Ig18].

5.2 Bewertung

Um die Eignung der Werkzeugarten für ein ganzheitliches Metadatenmanagement im Data Lake zu bewerten, werden die folgenden grundlegenden Kriterien definiert, die aus den Kernanforderungen (siehe Kapitel 3) sowie den unterschiedlichen Arten von Metadaten (siehe Kapitel 4) abgeleitet werden:

- Die *systemübergreifende Metadaten-Verwaltung* bezieht sich darauf, dass Metadaten über unterschiedliche Datenhaltungssysteme im Data Lake hinweg gemäß des Cross-Plattform-Ansatzes (siehe Kapitel 2) zu verwalten sind.
- Der *offene Metadaten-Austausch* bezieht sich auf standardisierte technische Schnittstellen zum Zugriff auf sowie zum Import und Export von Metadaten, um eine Metadaten-Nutzung in anderen Systemen zu ermöglichen.
- Die *unterstützten Metadaten-Arten* umfassen sämtliche Metadaten für ein ganzheitliches Metadatenmanagement, d.h. *originäre Data-Lake-Metadaten*, *Ergebnis-Metadaten*, *IoT-Geräte-Metadaten* und *Data-Lake-API-Metadaten* (siehe Kapitel 4).
- Der *Funktionsumfang zur Metadaten-Verwaltung* bezieht sich auf grundlegende Funktionalitäten zur Verarbeitung von Metadaten. Hierbei werden die *Speicherung*

und Bereitstellung von Metadaten, *Datenverlaufsanalysen* sowie metadatenbasierte *Mehrwertfunktionen*, z.B. zur automatisierten Erkennung personenbezogener Daten für DSGVO-Szenarien, unterschieden.

Wie in Tab. 1 dargestellt, fokussieren *systemintegrierte Datenverzeichnisse* auf die datenhaltungsinterne Speicherung und Bereitstellung von originären Data-Lake-Metadaten und bieten typischerweise keine standardisierten Schnittstellen für den offenen Metadaten-Austausch. Dementsprechend dienen systemintegrierte Datenverzeichnisse der im Data Lake verwendeten Datenhaltungssysteme zwar als Quellen für Metadaten, eigenen sich aber nicht als eigentliche Werkzeuge zur Realisierung eines ganzheitlichen Metadatenmanagements im Data Lake.

Kriterien / Werkzeugarten	**Systemintegrierte Daten-verzeichnisse**	**Datenkataloge**	**Data-Lake-Management-Plattformen**
Systemübergreifende Metadaten-Verwaltung	-	+	+
Offener Metadaten-Austausch	-	+	+
Unterstützte Metadaten-Arten			
Originäre Data-Lake-Metadaten	+	+	+
Ergebnis-Metadaten	-	+	+
IoT-Geräte-Metadaten	-	-	-
Data-Lake-API-Metadaten	-	-	+
Funktionsumfang zur Metadaten-Verwaltung			
Metadaten-Speicherung/-Bereitstellung	+	+	+
Datenverlaufsanalyse	-	+	+
Mehrwertfunktionen (DSGVO, ...)	-	+	-

Tab. 1: Bewertung der Werkzeugarten für ein ganzheitliches Metadatenmanagement

Datenkataloge ermöglichen als eigenständige Werkzeuge eine systemübergreifende Metadaten-Verwaltung und unterstützen mit standardisierten Schnittstellen, meist APIs, einen offenen Metadaten-Austausch. Der Funktionsumfang zur Metadaten-Verwaltung ist im Vergleich zu den anderen Werkzeugarten sehr ausgeprägt. Sie bieten Kollaborationsfunktionen, um den Endbenutzer im Rahmen von Self-Service-Szenarien einzubinden, und adressieren Governance-Anforderungen, um z.B. mittels Workflowfunktionen eine hohe, kontrollierte Qualität der Metadaten sicherzustellen. Darüber hinaus werden häufig metadatenbasierte Mehrwertfunktionen zur Erkennung personenbezogener Daten angeboten.

Data-Lake-Management-Plattformen stellen eine relativ neue Werkzeugart am Markt dar. Der Funktionsumfang zur Metadaten-Verwaltung der Data-Lake-Management-Plattformen orientiert sich an dem von Datenkatalogen, weist aber aktuell nicht denselben Reifegrad und

dieselbe Breite auf, insbesondere was metadatenbasierte Mehrwertfunktionen angeht. Ein Vorteil im Vergleich zu Datenkatalogen besteht jedoch in der integrierten Verwaltung von Data-Lake-API-Metadaten. Diese müssen bei der Verwendung von Datenkatalogen über entsprechende Schnittstellen für die Metadaten-Bereitstellung manuell gepflegt werden.

Summa summarum stellen Datenkataloge grundsätzlich die geeignete Werkzeugart zur praktischen Umsetzung eines ganzheitlichen Metadatenmanagements im Data Lake dar. Die Studienergebnisse weisen jedoch auf funktionale Unzulänglichkeiten und Erweiterungspotentiale von Datenkatalogen hin. Diese werden als Herausforderungen im folgenden Kapitel zusammengefasst.

6 Praktische Herausforderungen und Forschungsbedarfe

Ausgehend von den bisherigen Erfahrungen zum Aufbau eines ganzheitlichen Metadatenmanagements für den unternehmensweiten Data Lake des Industriekonzerns lassen sich IT-technische und fachlich-organisatorische Herausforderungen ableiten und zugehörige Forschungsbedarfe identifizieren (siehe Abb. 5).

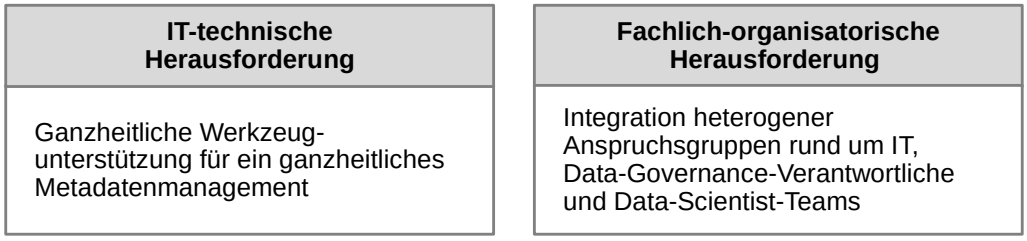

Abb. 5: Herausforderungen für ein ganzheitliches Metadatenmanagement im Data Lake

Die wesentliche *IT-technische Herausforderung* besteht darin, eine *ganzheitliche Werkzeugunterstützung für ein ganzheitliches Metadatenmanagement* zu realisieren. Hierbei geht es insbesondere darum, Unzulänglichkeiten existierender Datenkataloge zu adressieren, wobei folgende Aspekte eine zentrale Rolle spielen:

- *Datengetriebene Analyse des Datenverlaufs*: Die Analyse des Datenverlaufs, d.h. von Zusammenhängen zwischen Metadaten, erfolgt in existierenden Datenkatalogen größtenteils modellgetrieben, d.h. durch die Nutzung gegebener Modelle zu Metadaten und deren Zusammenhängen. Dies wird typischerweise durch das Auslesen von Metadaten aus ETL-Jobs realisiert. Im Rahmen von Self-Service-Szenarien stehen jedoch meist keine umfassenden Modelle zum Datenverlauf aus ETL-Werkzeugen zur Verfügung, da sich der Datenverlauf aus manuellen, iterativen Datentransformationen durch den Endbenutzer ergibt. Es geht folglich darum, Metadaten-Zusammenhänge datengetrieben zu erfassen und zu analysieren. Denkbar wäre beispielsweise, Data-Mining-Verfahren auf relationale Tabellenstrukturen anzuwenden, um mögliche Beziehungen und Ähnlichkeiten zwischen den Tabellen zu erkennen. In existierenden Datenkatalogen ist die datengetriebene Analyse von Metadaten-Zusammenhängen

noch sehr rudimentär ausgeprägt und basiert häufig auf Zeitstempeln und existierenden relationalen Beziehungen.

- *Metadaten-Erfassung von Streaming-Komponenten*: Der Fokus existierender Datenkataloge liegt auf der Erfassung von Metadaten aus Batch-Datenhaltungssystemen, z.B. durch die Anbindung von systemintegrierten Datenverzeichnissen relationaler Datenbanken. Die Unterstützung der Metadaten-Erfassung von Streaming-Komponenten, wie z.B. Apache Kafka, ist bei aktuell verfügbaren Datenkatalogen wenig ausgeprägt. Für einen spezifischen Datenkatalog muss typischerweise ein eigener Adapter für jede Streaming-Komponente entwickelt werden. Streaming-Komponenten sind in Data Lakes für die echtzeitnahe Datenverarbeitung jedoch von großer Bedeutung [Ma17]. Für ein ganzheitliches Metadatenmanagement ist eine integrierte Verwaltung von Metadaten über Batch- und Streaming-Komponenten erforderlich, d.h. der (Batch-)Datenkatalog ist um einen Ereigniskatalog mit Metadaten zu Ereignisströmen (engl. event stream) zu ergänzen. Damit können beispielsweise ganzheitliche Analysen des Datenverlaufs über Batch- und Streaming-Komponenten, z.B. für DSGVO-Nachweispflichten, durchgeführt werden.
- *Integration von IoT-Geräte-Metadaten und Data-Lake-API-Metadaten*: IoT-Geräte-Metadaten sowie Data-Lake-API-Metadaten werden von existierenden Datenkatalogen typischerweise nicht betrachtet. Meist werden IoT-Geräte-Metadaten in dedizierten IoT-Gerätemanagement-Anwendungen verwaltet. Ein ganzheitliches Metadaten-Management erfordert dementsprechend eine integrierte Verwaltung von IoT-Geräte-Metadaten und Data-Lake-API-Metadaten gemeinsam mit originären Data-Lake-Metadaten und Ergebnis-Metadaten.

Die wesentliche *fachlich-organisatorische Herausforderung* für ein ganzheitliches Metadatenmanagement besteht darin, die *heterogenen Anspruchsgruppen* (engl. stakeholder) *rund um IT*, *Data-Governance-Verantwortliche und Data-Scientist-Teams zu integrieren*, um eine erfolgreiche Umsetzung des Metadatenmanagements als Gesamtprojekt sicherzustellen. Data-Governance-Verantwortliche sind in der industriellen Praxis ein wesentlicher Initiator für das Metadatenmanagement, sowohl in analytischen als auch in operativen IT-Systemen, und definieren zentrale Anforderungen für das Metadatenmanagement. Es werden beispielsweise Rollen, Prozesse und Verantwortlichkeiten für Dateneigentum und Datennutzung im Rahmen von Data-Governance-Projekten festgelegt, die im Datenkatalog abzubilden sind. Data Scientists sind in der Praxis meist als unternehmensübergreifende Expertenteams organisiert, die stark self-service-orientiert arbeiten, um Anwendungsfälle schnell umzusetzen [Gr18]. Sie stellen damit eine wesentliche Anspruchs- und Endbenutzergruppe für den Datenkatalog dar. In der Praxis liegt meist eine organisatorische Trennung der IT, die für die Konfiguration und den Betrieb des Datenkatalogs verantwortlich ist, Data-Governance-Verantwortlichen in Fachabteilungen sowie Data Scientist-Teams vor, was zu unterschiedlichen funktionalen Zielsetzungen führt. Die IT fokussiert auf den effizienten und sicheren Betrieb des Datenkatalogs, Data Scientists sind an möglichst weitreichenden kollaborationsgetriebenen Self-Service-Funktionalitäten interessiert, wohingegen Data-

Governance-Verantwortliche meist Transparenz und regelgetriebene Strukturen bevorzugen. Zur zielgerichteten Umsetzung des Metadatenmanagements ist folglich eine kontinuierliche Einbeziehung und Harmonisierung dieser Anspruchsgruppen, d.h. ein systematisches Stakeholder-Management, erforderlich.

7 Fazit

Die in dieser Arbeit diskutierten Praxiserfahrungen machen deutlich, dass ein ganzheitliches Metadatenmanagement im Data Lake einen zentralen Erfolgsfaktor für die Umsetzung von Governance-Anforderungen und Self-Service-Szenarien im Data Lake darstellt. Gleichzeitig wird deutlich, dass die Anforderungen an ein ganzheitliches Metadatenmanagement sowohl IT-technisch als auch fachlich-organisatorisch sehr komplex sind. Heterogene Metadaten sind über heterogene Datenhaltungssysteme im Data Lake systematisch zu verwalten und organisatorisch getrennte Anspruchsgruppen mit unterschiedlichen Zielsetzungen – IT, Data Scientists und Data-Governance-Verantwortliche – sind für eine erfolgreiche Umsetzung konstruktiv einzubeziehen. Zukünftige Forschungsbedarfe ergeben sich insbesondere aus einer ganzheitlichen Werkzeugunterstützung für das Metadatenmanagement im Data Lake, bei der die integrierte Verwaltung von Metadaten aus Batch- und Streaming-Komponenten sowie die datengetriebene Gewinnung von Metadaten über den Datenverlauf im Vordergrund stehen.

Danksagungen

Die Autoren danken Dieter Neumann, Thomas Müller und Arnold Lutsch für die kontinuierliche Unterstützung und die konstruktiven Diskussionen zu dieser Arbeit.

Literaturverzeichnis

[Aa16] Aalst, W. v.d.: Process Mining. Springer, Heidelberg, 2016.

[Al16] Alserafi, A. et al.: Towards Information Profiling: Data Lake Content Metadata Management: Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW) 2016. IEEE, S. 178–185, 2016.

[Ap18a] Apache Software Foundation: Apache Hive. `http://hive.apache.org`, Stand: 12.09.18.

[Ap18b] Apache Software Foundation: Apache Atlas. `http://atlas.apache.org`, Stand: 12.09.18.

[AS16] Alpar, P.; Schulz, M.: Self-Service Business Intelligence. In Business & Information Systems Engineering, 58(2), S. 151–155, 2016.

[Ba14] Bauernhansl, T.: Die Vierte Industrielle Revolution – Der Weg in ein wertschaffendes Produktionsparadigma. In (Bauernhansl, T.; Hompel, M. t.; Vogel-Heuser, B. Hrsg.): Industrie 4.0 in Produktion, Automatisierung und Logistik. Anwendung, Technologien, Migration. Springer Vieweg, Wiesbaden, S. 5–35, 2014.

[Co18] Collibra: Collibra Catalog. http://www.collibra.com/data-governance-solutions/data-catalog, Stand: 12.09.18.

[Di16] Dinsmore, T. W.: Disruptive Analytics. Apress, Berkeley, 2016.

[Eu16] Europäische Union: EU Verordnung 2016/679 (Datenschutz-Grundverordnung), 2016.

[GCA15] Gölzer, P.; Cato, P.; Amberg, M.: Data processing requirements of industry 4.0 – use cases for big data applications: Proceedings of the European Conference on Information Systems (ECIS) 2015, paper 61, 2015.

[GP14] Gausemeier, J.; Plass, C.: Zukunftsorientierte Unternehmensgestaltung. Strategien, Geschäftsprozesse und IT-Systeme für die Produktion von morgen. Hanser, München, 2014.

[GR15] Gessert, F.; Ritter, N.: Polyglot Persistence. In Datenbank-Spektrum, 15(3), S. 229–233, 2015.

[Gr18] Gröger, C.: Building an Industry 4.0 Analytics Platform. In Datenbank-Spektrum, 18(1), S. 5–14, 2018.

[GSMa14] Gröger, C.; Schwarz, H.; Mitschang, B.: The Manufacturing Knowledge Repository. Consolidating Knowledge to Enable Holistic Process Knowledge Management in Manufacturing: Proceedings of the International Conference on Enterprise Information Systems (ICEIS) 2014. SciTePress, S. 39–51, 2014.

[GSMb14] Gröger, C.; Schwarz, H.; Mitschang, B.: The Deep Data Warehouse. Link-based Integration and Enrichment of Warehouse Data and Unstructured Content: Proceedings of the IEEE International Enterprise Distributed Object Computing Conference (EDOC) 2014. IEEE, Los Alamitos, S. 210–217, 2014.

[He17] Henderson, D. et al.: DAMA-DMBOK. Data management body of knowledge. Technics Publications, New Jersey, 2017.

[HGQ16] Hai, R.; Geisler, S.; Quix, C.: Constance: An Intelligent Data Lake System: Proceedings of the International Conference on Management of Data (SIGMOD) 2016. ACM Press, New York, S. 2097–2100, 2016.

[HKP12] Han, J.; Kamber, M.; Pei, J.: Data mining. Concepts and techniques. Elsevier/Morgan Kaufmann, Amsterdam, 2012.

[Ig18] Iguazio: Iguazio Data Platform. http://www.iguazio.com, Stand: 12.09.18.

[In18] Informatica: Informatica Enterprise Data Catalog. http://www.informatica.com/de/products/big-data/enterprise-data-catalog.html, Stand: 12.09.18.

[Ka15] Kassner, L. et al.: Product Life Cycle Analytics - Next Generation Data Analytics on Structured and Unstructured Data. In Procedia CIRP, 33, S. 35–40, 2015.

[KB10] Khatri, V.; Brown, C. V.: Designing data governance. In Communications of the ACM, 53(1), S. 148–152, 2010.

[KBM10] Kemper, H.-G.; Baars, H.; Mehanna, W.: Business Intelligence – Grundlagen und praktische Anwendungen. Vieweg+Teubner, Wiesbaden, 2010.

[KE06] Kemper, A.; Eickler, A.: Datenbanksysteme. Oldenbourg, München, 2006.

[Lv17] Lv, Z. et al.: Next-Generation Big Data Analytics. State of the Art, Challenges, and Future Research Topics. In IEEE Transactions on Industrial Informatics, 13(4), S. 1891–1899, 2017.

[Ma17] Mathis, C.: Data Lakes. In Datenbank-Spektrum, 17(3), S. 289–293, 2017.

[MT16] Miloslavskaya, N.; Tolstoy, A.: Big Data, Fast Data and Data Lake Concepts. In Procedia Computer Science, 88, S. 300–305, 2016.

[MW15] Marz, N.; Warren, J.: Big data. Principles and best practices of scalable real-time data systems. Manning, Shelter Island, 2015.

[OL13] OLeary, D.: Artificial intelligence and big data. In IEEE Intelligent Systems, 28(2), S. 96–99, 2013.

[QHV16] Quix, C.; Hai, R.; Vatov, I.: Metadata Extraction and Management in Data Lakes With GEMMS. In Complex Systems Informatics and Modeling Quarterly, (9), S. 67–83, 2016.

[Sh18] Sharma, B.: Architecting data lakes. O'Reilly, Sebastopol, 2018.

[Te15] Terrizzano et al.: Data Wrangling: The Challenging Yourney from the Wild to the Lake: Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR) 2015, 2015.

[Te18] Terradata: Kylo. www.kylo.io, Stand: 12.09.18.

[VVS00] Vetterli, T.; Vaduva, A.; Staudt, M.: Metadata standards for data warehousing. In ACM SIGMOD Record, 29(3), S. 68–75, 2000.

[Wa18] Waterline Data: Waterline Smart Data Catalog. https://www.waterlinedata.com/product-overview, Stand: 12.09.18.