Data Engineering

Research Paper

Erster Autor[1], Zweiter Autor[2]

[1] Institution, Bereich, Stadt, Land
{erster.autor}@example.com
[2] Andere Institution, Anderer Bereich, Andere Stadt, Anderes Land
{zweiter.autor}@example.com

Abstract. Data Engineering wird als Teilbereich von Data Science im Zuge der Digitalisierung und Big Data immer wichtiger. Das Data Engineering beschäftigt sich mit der Gewinnung, Speicherung, Transformation und Bereitstellung von Daten für die anschließende Analyse und Informationsgewinnung. Mit Hilfe einer Literaturanalyse werden aktuelle wissenschaftliche Veröffentlichungen qualitativ ausgewertet und in die Phasen des Data Engineering Lifecycle eingeordnet. Anschließend wird eine quantitative Auswertung in einer Konzeptmatrix nach Webster und Watson (2002) durchgeführt. Die Ergebnisse identifizieren die Aufgabengebiete Ingestion und Transformation als Hauptbereiche aktueller Forschung, gefolgt von den Aufgabengebieten der Generierung und Speicherung. Im Aufgabengebiet der Bereitstellung wird auf Basis der durchgeführten Literaturanalyse aktuell weniger aktiv geforscht.

Schlüsselwörter: *Data Engineering, Data Science, Information Systems*

1 Einführung

1.1 Ausgangssituation

Mit der voranschreitenden Digitalisierung werden in Unternehmen riesige Datenmengen generiert (vgl. Zerfaß et al. 2022:291), um auf Basis der daraus gewonnen Informationen bessere Geschäftsentscheidungen zu treffen (vgl. Leimeister 2021:138). Das daraus entstanden Wissenschaftsfeld von Data Science hat sich durch das stetige Wachstum an zu verarbeitenden Daten stark erweitert (vgl. Liebregts et al. 2023:21). Dabei sind innerhalb von Data Science weitere Teilbereiche entstanden, die detaillierter

19th International Conference on Wirtschaftsinformatik,
September 2024, Würzburg, Germany

auf die einzelnen Methoden und Prozesse eingehen, um aus Rohdaten Informationen zu gewinnen (vgl. Tamir et al. 2015:2). Für die Beschreibung des Teilbereichs für die Gewinnung, Speicherung, Transformation und Bereitstellung von Daten für die nachfolgende Analyse ist der Begriff „Data Engineering“ entstanden (vgl. Reis/Housley 2023:30). Der 2023 von Joe Reis und Matt Housley definierte „Data Engineering Lifecyle“ teilt die Aufgabenbereiche des Data Engineerings in fünf Phasen eines Lebenszyklus ein. Von dem Ursprung der Generierung über die Speicherung, Ingestion, Transformation bis zur finalen Bereitstellung (vgl. Reis/Housley 2023:31). Für die weitere Forschung im Bereich des Data Engineerings ist es notwendig, den aktuellen wissenschaftlichen Status zu identifizieren, um zu erkennen, in welchen Bereichen des Data Engineerings aktiv geforscht wird und in welchen Bereichen es Forschungsbedarf gibt.

1.2 Zielsetzung

Das Ziel der vorliegenden Arbeit besteht in einer ausführlichen Analyse der aktuellen Forschungen im Bereich von Data Engineering, um dadurch einen aktuellen wissenschaftlichen Stand und Status abzuleiten. Ein zentrales Ziel dabei ist die Identifizierung der Aufgabenbereiche, in denen aktuell viel Forschung betrieben wird, sowie Aufgabenbereiche, in denen möglicherweise Forschungsbedarf besteht.

Hieraus leitet sich folgende Forschungsfrage ab:

Welche Aufgabenbereiche im Bereich des Data Engineerings stehen im Mittelpunkt aktueller Forschung?

1.3 Aufbau der Arbeit

Die vorliegende Arbeit besteht aus fünf Kapitel. Die Einführung erläutert die Ausgangssituation, die Zielsetzung, den Aufbau der Arbeit sowie die verwendete Methodik. Kapitel 2, Grundlagen und Begriffsdefinitionen, führt zunächst in das Wissenschaftsfeld von Data Science ein und spezifiziert dann weiter in den Teilbereich des Data Engineerings und schlussendlich den Data Engineering Lifecycle. Im dritten Kapitel werden die Ergebnisse der systematischen Literaturanalyse dargestellt und anschließend die Forschungsfrage beantwortet. Kapitel vier fasst die gewonnen Erkenntnisse zusammen und beinhaltet eine darauf basierende Diskussion, sowie kritische Würdigung und möglichen weiteren Forschungsbedarf. Das letzte Kapitel präsentiert die verwendeten Quellen dieser Arbeit.

1.4 Methodik

In der vorliegenden Forschungsarbeit wurde das methodische Vorgehen von Fettke (2006) als zentraler Leitfaden für die Literaturanalyse verwendet. Dabei bietet das fünfstufige Phasenmodell eine strukturierte Grundlage für die Literatursuche und -auswahl, wodurch eine systematische Analyse relevanter Forschungsliteratur ermöglicht wird.

Mit Hilfe einer Konzeptmatrix nach Webster und Watson (2002) wird die Forschungsfrage beantwortet sowie quantitativ ausgewertet.

2 Grundlagen und Begriffsdefinitionen

2.1 Data Science

Die branchendurchdringende Digitalisierung und die damit einhergehende stark steigende Sammlung von Daten ordnen wir heute als Zeit von Big Data ein. Gemeint sind hierbei die potenziell unendlich großen Datenmengen, die durch die Digitalisierung, teilweise in Echtzeit, aus zahlreichen Quellen gesammelt, gespeichert und analysiert werden können (vgl. Schallmo 2023:557). Ziel der Unternehmen ist es, mit den Erkenntnissen, die durch die Auswertung der Datenmengen gewonnen werden, Vorteile im Wettbewerb aufzubauen und damit zum Erfolg des Unternehmens beizutragen (vgl. Schallmo 2023:556). Das Sammeln der Daten allein reicht jedoch nicht aus, um Erkenntnisse für den unternehmerischen Vorteil zu generieren. Die Daten aus verschiedensten Quellen in unterschiedlichster Qualität und Art müssen zuerst aufbereitet und auch analysiert werden. Für diese komplexe Aufgabe ist das interdisziplinäre Wissenschaftsfeld des „Data Science“ entstanden. Ziel ist es, aus Daten neues Wissen zu generieren, welches neue Geschäftsmodelle ermöglichen kann (vgl. Papp et al. 2019:XIII).
Im Jahr 2000 veröffentlichte Colin Shearer das Vorgehensmodell CRISP-DM, welches heute als Standard im Bereich Data Science gilt. Sechs Phasen beschreiben den Weg, um aus Daten wertvolle Informationen zu gewinnen, die dann als Entscheidungsgrundlage für Geschäftsentscheidungen dienen können. Darunter das Geschäftsverständnis, das Datenverständnis, die Datenvorbereitung, die Modellierung, die Evaluierung und die schlussendliche Bereitstellung (vgl. Shearer 2000:13).

2.2 Teilbereich des Data Engineering

Durch die stetig wachsenden Datenmengen und die Popularität von Data Science Methoden hat sich das Wissenschaftsfeld „Data Science“ schnell entwickelt und neue Prozesse, Modelle und Tools entstanden. Dabei viel auf, dass Data Scientists den Großteil ihrer Arbeit damit verbringen, Rohdaten zu importieren, aufzubereiten und für die Bearbeitung bereitzustellen (vgl. Rose 2016:6). Dieser Prozess der Bereitstellung von qualitativ hochwertigen und verarbeitbaren Daten wurde im Vorgehensmodell CRISP-DM schlichtweg vorausgesetzt. Um sich diesem Prozess im Detail zu widmen, ihn zu automatisieren und zu optimieren entstand die Disziplin des „Data Engineering“.

Während die Aufgaben eines Data Engineers bereits lange klar waren, wurden diese lange Zeit als zusätzliche Aufgabe zum CRISP-DM Prozess den Data Scientists zugeschrieben. Immer mehr und umfangreichere Datenquellen machten es jedoch notwendig, den Bereich des Data Engineerings als Teilbereich des Data Science auszugliedern.

Joe Reis und Matt Housley definieren Data Engineering als „die Entwicklung, Implementierung und Wartung von Systemen und Prozessen, die Rohdaten aufnehmen und hochwertige, konsistente Informationen erzeugen, die nachgelagerte Anwendungsfälle wie Analysen und Machine Learning unterstützen (Reis/Housley 2023:30). Im Bereich des Data Science legt Data Engineering demnach den Fokus auf die Vorbereitung der Daten, mit denen anschließend mit Data Analytics die Datenauswertungen durchgeführt werden.

2.3 Data Engineering Lebenszyklus

Ausgehend von der Definition von Data Engineering und den Tätigkeiten, welche dem Data Engineer zugeordnet werden konnten, erstellten Reis und Housley das Modell des „Data Engineering Lifecycle“. Dieses Modell aus 5 Phasen beschreibt den Prozess, um aus Rohdaten qualitativ hochwertige und verarbeitbare Daten herzustellen, die dann von Analysten, Data Scientists und Machine Learning Engineers genutzt werden können (vgl. Reis/Housley 2023:63).

Die fünf Phasen des Data Engineering Lifecycle lassen sich folgendermaßen beschreiben (vgl. Reis/Housley 2023:63):

Generierung: Auch wenn die konkrete Generierung der Daten nicht die Aufgabe eines Data Engineers ist, muss dieser doch die Funktionsweise, Strukturen und Schnittstellen der Quellsysteme verstehen. Hauptaufgabe ist es hierbei, den fehlerlosen Import von Daten aus den Quellsystemen mit Hilfe von Schnittstellen und Datapipelines zu ermöglichen. Die Aufgabe wird mit zunehmender Anzahl an Quellsystemen sehr komplex, weil die einströmenden Daten unterschiedliche Schemata besitzen und in unterschiedlichsten zeitlichen Abständen eintreffen.

Speicherung: Bevor die Daten aus den Quellsystemen aufgenommen werden können, müssen sich Data Engineers mit der passenden Speicherung auseinandersetzen. Dabei können verschiedenste Speicherlösungen für unterschiedliche Quelldaten zum Einsatz kommen. Die Wahl einer geeigneten Speicherlösung ist entscheidend, da die Datensicherung sich über den ganzen Data Engineering Lifecycle erstreckt. Schon bei der Auswahl muss beachtet werden, dass die Folgephasen ihre Aufgabe mit der ausgewählten Speicherlösung auf den Daten durchführen können. Während Objektspeicher beispielsweise oft nur eine reine Speicherlösung darstellen, gehen die Funktionalitäten eines Cloud Data Warehouse weit darüber hinaus und ermöglichen komplexe Abfragemuster. Auch die Zugriffshäufigkeit ist bei der Wahl der Speicherlösung zu beachten. Wird auf Daten sehr häufig zugegriffen, werden diese als „heiße Daten“ bezeichnet. Im Gegensatz dazu werden „kalte Daten“ nur selten abgefragt und können in kostengünstigere Archivierungssysteme gespeichert werden. Die Entscheidungen für eine geeignete Speicherlösung sind bereits vor dem eigentlichen Eintreffen der Daten zu treffen.

Ingestion: Die Phase der Ingestion beschäftigt sich dann mit der konkreten Sammlung der Daten in den definierten Speicherlösungen. Da die Quellsysteme meist außerhalb

der Kontrolle der Data Engineers liegen kann es vorkommen, dass die Quellsysteme nicht mehr erreichbar sind oder falsche Daten bzw. Daten mit schlechter Qualität das eigene System erreichen. Auch die Art, wie die Daten versendet werden kann unterschiedlich sein. Während einige Quellsysteme die Daten gebündelt als Batch-Ingestion senden erreichen die eigenen Speicherlösungen bei der Streaming-Ingestion kontinuierlich Daten. Eine große Herausforderung stellt dabei die Echtzeitverarbeitung von einströmenden Streamingdaten dar. Eine der Hauptfragen in der Ingestionphase ist die Frage zur Datenzufuhr selbst. Wird ein Push-Modell gewählt, werden die Daten vom Quellsystem direkt auf das Zielsystem geschrieben. Beim Pull-Modell werden die Daten hingegen aktiv vom Zielsystem bei dem Quellsystem angefragt und fließen anschließend ins Zielsystem. Der bekannte ETL-Prozess (Extrahieren, Transformieren, Laden) verwendet beispielsweise, wie bereits das Wort „extrahieren“ andeutet, die Pull-Methode für die Datenzufuhr. IoT-Geräte hingegen senden die gewonnen Daten häufig gleich los, statt sie lange für einen Abruf zwischenzuspeichern, weswegen dann das Push-Modell zum Einsatz kommt.

Transformation: Sind die Daten aus den Quellsystemen erfolgreich auf dem Zielsystem gespeichert müssen diese transformiert werden. Ziel ist es, die empfangenen und gespeicherten Daten aus ihrer ursprünglichen Form in eine neue Form zu bringen, die für die folgenden Anwendungsfälle nützlicher ist. Ohne diese Datenaufbereitung sind die Daten für Berichte, Analysen und das maschinelle Lernen (ML) unbrauchbar und somit wertlos. Die ersten Schritte hierbei sind beispielsweise die Umwandlung der Zeichenketten in korrekte Datentypen und die Entfernung ungültiger Datensätze. Nachgelagert kann aber auch das Datenschema umgewandelt oder die Daten mit Features angereichert werden, sodass diese in ML-Prozessen verarbeitet werden können.

Bereitstellung: In der letzten Phase des Data Engineering Lifecycles gilt es die transformierten Daten so bereitzustellen, dass die folgenden Anwender ohne große Aufwände mit den Daten arbeiten können, um daraus einen Mehrwert zu ziehen. Dabei erfordern unterschiedliche Anwendungsbereiche oft eine speziell angepasste Bereitstellung der Daten. Beispielsweise müssen die Daten für Operational Analytics möglichst aktuell sein, jedoch ist die Masse an Daten oft kleiner. Im Gegensatz dazu fordern Business Intelligence Anwendungen die Bereitstellung von einer großen Menge an Daten aus der Vergangenheit, um diese zusammen mit der Geschäftslogik auszuwerten. Dabei ist es jedoch weniger relevant, dass die Daten auf die Sekunde genau aktuell sind. Für ML-Systeme ist es hingegen wichtiger, dass die Daten beispielsweise mit Features angereichert und im passenden verarbeitbaren Format vorliegen. Auch hier wird die Umsetzung komplex, wenn z.B. mehrere Anwender gleichzeitig auf Daten zugreifen, oder sich Daten durchgängig ändern bzw. aktualisieren.

Neben den fünf Phasen im Modell müssen Data Engineers parallel auch Aufgabenbereiche wie Sicherheit, Datenmanagement, Data-Ops, Datenarchitektur, Orchestrierung und Softwareentwicklung abdecken, die den Data Engineering Lifecycle in jeder Phase begleiten.

3 Resultate

Im Rahmen der Forschungsmethodik für die vorliegende Arbeit wurde eine sorgfältige Auswahl und Analyse relevanter Literatur im Bereich der Informationssysteme und Informationstechnologie durchgeführt. Ein zentrales Element war die Verwendung eines stark eingeschränkten zeitlichen Filters, wobei nur Arbeiten berücksichtigt wurden, die im Zeitraum von 2022 bis 2024 veröffentlicht wurden. Um eine möglichst breite Übersicht über aktuelle Forschungsarbeiten im Bereich des Data Engineerings zu erzielen, wurde der sehr allgemeine Suchterm „Data Engineering“ verwendet und nach Konferenzbeiträgen gefiltert. Die nachfolgende Tabelle quantifiziert die Ergebnisse der detaillierten Literaturrecherche:

Tabelle 1. *Quantifizierte Darstellung der identifizierten Publikationen*

	Suchterm
Quelle	
Springer Link	32
AMCIS	23
Abfrageergebnisse ∑	55
Nach Sichtung des Titels & Zugänglichkeit	39
Nach Entfernung der Duplikate ∑	39
Nach wissenschaftlicher Eignung	33
Nach inhaltlicher Relevanz	19
Relevante Publikationen ∑	**19**

Die Auswahl der Datenbanken wurde gezielt auf Veröffentlichungen in der internationalen Konferenz Americas Conference on Information Systems (AMCIS) und dem renommierten wissenschaftlichen Verlag SpringerLink beschränkt.

3.1 Auswertung der Ergebnisse

Nachfolgend werden die Ergebnisse der Literaturanalyse präsentiert und auf Basis des Data Engineering Lifecycles (vgl. Reis/Housley 2023:63) den einzelnen Aufgabenbereichen bzw. Phasen innerhalb des Bereichs Data Engineering zugeordnet.

Generierung
Die gefundenen Forschungsarbeiten, die sich der Phase der Generierung zuordnen lassen, beschäftigen mit den Technologien und den Schemata der einzelnen Datenquellen, sowie deren sichere Anbindung an das Zielsystem. Darunter die Analyse von Web-Services als Datenquelle (vgl. Sarkar/Sawale 2022:301) aber auch die Anbindung von

drahtlosen Datenquellen und die verteilte Datenverarbeitung der einströmenden Daten ins Zielsystem (vgl. Madhuri et al. 2022:125). Zusätzlich liegt der Fokus auch auf einer möglichst sicheren Anbindung von fremden Datenquellen an eigene Systeme. Mit automatisierter Erkennung von Cyberattacken (vgl. Purnaye/Kulkarni 2022:11), sowie Clusteranalysen bei einströmenden Daten von kabellosen Sensornetzen (vgl. Jaiswal/Anand 2022:293) sollen die Daten bereits vor der Speicherung einer Sicherheitskontrolle unterzogen werden, um mögliche Attacken oder Falschdaten abzufangen.

Speicherung
Im Bereich der Speicherung gibt es einen klaren Fokus zur Erforschung der Frage, wie verschiedene Arten von Datenquellen effizient und sicher gespeichert werden können. Die große Nachfrage an der Speicherung von großen Datenmengen zeigt sich in Forschungsarbeiten zum generellen Aufbau einer Big Data Datenbank (vgl. Alnafoosi/Steinbach 2023:31), sowie Vergleichen von unterschiedlichen NoSQL-Datenbanken (vgl. Asaad et al. 2024:214). Mit einem Clustering von JSON-Dokumenten nach deren Aufbau, soll die Bearbeitung der großen Datenmengen verbessert werden (vgl. Uma Priya/Santhi Thilagam 2023:51). Auch bei der Speicherung ist die Sicherheit der Daten essenziell. Mit sicheren Methoden verteilter Speicherung von Bildern (vgl. Tandon/Sharma 2022:553) soll die Privatsphäre bei möglichst gleichbleibender Bildqualität erreicht werden.

Ingestion
Aktuelle Forschungsarbeiten in der Phase der Ingestion beschäftigen sich mit der Optimierung von bestehenden Datenübertragungstechniken, um hohe Datenmengen in kurzer Zeit verarbeiten zu können, und neuen Techniken, die eine sichere Datenübertragung trotz schwieriger Bedingungen ermöglichen. Mit der Erforschung von verzögerungstoleranten Netzwerken können Daten auch erfolgreich übertragen werden, wenn die Verbindung zwischen den Konten kurzzeitig abbricht und Datenpakete verloren gehen (vgl. Gantayat et al. 2022:453). Auch Unterwassersensoren und deren gesammelte Daten stellen eine Herausforderung für die Datenübertragung dar. Mit zeitbasierten Übertragungsprotokollen und Minimierung des Rauschens können auch dort möglichst fehlerfreie Datenübertragungen gewährleistet werden (vgl. Bhujange et al. 2023:1). Neue Techniken für eine bessere Lastenverteilung werden aufgrund der wachsenden Anzahl an Daten von intelligenten Geräten (vgl. Kaur/Aron 2022:213) und sozialen Medien (vgl. Jain et al. 2022:469) wichtiger, um Datenstau und hohe Latenzen bei Datencentern zu verhindern. Moderne mehrstufige Verbindungsnetzwerke sind dabei eine Lösung für schnelle Datenübertragung, haben jedoch noch mit Unschärfe (Fuzziness) zu kämpfen (vgl. Amutha/Pritha 2022:165).

Transformation
Im Bereich der Transformation beschäftigen sich aktuelle Forschungsarbeiten hauptsächlich mit der Qualität der gespeicherten Daten. Eine Verbesserung der Datenqualität wird zum einen in der Speicherarchitektur selbst erforscht (vgl. Altendeitering et al. 2022:3), zum anderen aber auch in der Analyse der gespeicherten Daten selbst gesucht (vgl. Pohl et al. 2023:1). Die Reinigung der fehlerhaften oder unvollständigen Daten

spielt eine immer größere Rolle, um Data Scientists saubere Daten für die Erstellung von Auswertungen zu liefern (vgl. Sahay et al. 2022 & Jouseau et al. 2022:25). Für die Erstellung von Machine Learning Modellen ist die Anreicherung der Daten mit Features ein essenzieller Schritt für ein gutes Modell, weswegen bereits in der Phase der Transformation die Daten mit Informationen angereichter werden (vgl. Khatri/Bansal 2022:177).

Bereitstellung
Die gefundene Veröffentlichung, die sich der Phase der Bereitstellung zuordnen lässt, versucht die Verarbeitung und Analyse von Echtzeitdaten zu verbessern und die Ergebnisse möglichst schnell bereitzustellen (vgl. Dashora/Babu 2023:547). Die Herausforderung besteht dabei darin, dass die Bereitstellung der analysierten Daten innerhalb kürzester Zeit erfolgen muss, da angeschlossene Informationssysteme möglichst in Echtzeit auf Veränderungen reagieren müssen.

3.2 Konzeptmatrix

Die nachfolgende Tabelle veranschaulicht die Zuordnung der identifizierten Forschungsarbeiten im Bereich Data Engineering zu den Aufgabebereichen des Data Engineerings nach dem Vorbild der Konzeptmatrix von Webster & Watson (2002).

Tabelle 2. *Konzeptmatrix*

Referenz	Generierung	Speicherung	Ingestion	Transformation	Bereitstellung
Alnafoosi/Steinbach 2023		x			
Amutha/ Pritha 2022			x		
Jain et al. 2022			x		
Purnaye/ Kulkarni 2022	x				
Sahay et al. 2022				x	
Uma Priya/ Santhi Thilagam 2023		x			
Pohl et al. 2023				x	
Dashora/Babu 2023					x
Kaur/Aron 2022			x		
Sarkar/Sawale 2022	x				
Tandon/Sharma 2022		x			
Bhujange et al. 2023			x		
Jaiswal/Anand 2022	x				
Altendeitering et al. 2022				x	
Jouseau et al. 2022				x	
Asaad et al. 2024		x			
Khatri/Bansal 2022				x	
Gantayat et al. 2022			x		
Madhuri et al. 2023	x				

3.3 Auswertung der Konzeptmatrix

In der nachfolgenden Tabelle wird die erarbeitete Konzeptmatrix analysiert. Die Spalte „Anzahl“ gibt dabei die absolute Häufigkeit der identifizierten Forschungsarbeiten innerhalb des Aufgabenbereichs wieder. Die relative Häufigkeit in Prozent zeigt die prozentuale Verteilung der Ausprägungen im Verhältnis zu allen in der Konzeptmatrix ausgewerteten Publikationen.

Tabelle 3**.** *Auswertung der Konzeptmatrix*

Dimension	*Generierung*	*Speicherung*	*Ingestion*	*Transformation*	*Bereitstellung*
Anzahl	4	4	5	5	1
Relative Häufigkeit in %	21%	21%	26,5%	26,5%	5%

Die Analyse der aktuellen Forschungsarbeiten im Bereich des Data Engineerings zeigt, dass im Aufgabenbereich der Ingestion und Transformation in den Jahren 2022 bis 2024 mit jeweils 26,5% am meisten publiziert wurde. Dies lässt darauf schließen, dass die Wissenschaft dort mit weiterer Forschung das höchste Potenzial sieht, die Daten für nachfolgenden Auswertungen zu verbessern. Ebenfalls relevant sind die Aufgabenbereiche Generierung und Speicherung. Mit jeweils 21% bilden die beiden Bereiche den Durschnitt bei insgesamt 5 Phasen ab, was darauf hinweist, dass die Aufgabenbereiche weiterhin relevant für die aktuelle Forschung sind. Der Bereitstellungsphase konnte in der durchgeführten Literaturanalyse nur eine Quelle (5%) zugeordnet werden. Damit ist der Aufgabenbereich der Bereitstellung weit abgeschieden auf dem letzten Platz. Dies deutet weniger auf eine Forschungslücke, sondern vielmehr auf einen bereits viel erforschten Aufgabenbereich hin, der aktuell weniger Potenzial für Optimierungen bietet.

Die Ergebnisse verdeutlichen die aktive Forschung in vier von fünf Aufgabenbereichen des Data Engineerings. Dies unterstreicht nochmals die Anforderungen und auch die Nachfrage an Lösungen, um den großen Datenmengen im Big Data Umfeld gerecht zu werden.

3.4 Beantwortung der Forschungsfrage

Die ursprüngliche Forschungsfrage „Welche Aufgabenbereiche im Bereich des Data Engineerings stehen im Mittelpunkt aktueller Forschung?“ kann folgendermaßen beantwortet werden:

Die Aufgabenbereiche Ingestion und Transformation stehen im Fokus aktueller Forschung im Bereich des Data Engineerings. Dabei beschäftigt sich die Forschung hauptsächlich mit Datenübertragungen unter schwierigen Bedingungen, Lastenverteilung im

Kontext von Big Data und der Transformation von Daten durch Datenreinigung und Anreicherung der Daten. Weniger im Fokus, aber dennoch aktiv erforscht, werden die Aufgabenbereiche Generierung und Speicherung, in denen zu neuen Datenquellen und deren Anbindung, sowie einer optimierten Speicherung der empfangenen Daten geforscht wird. Weniger erforscht wird der Aufgabenbereich Bereitstellung, in dem vermutlich bereits etablierte Methoden und Prozesse vorhanden sind und der Forschungsbedarf weitreichend gedeckt ist.

4 Diskussion und Forschungsbedarf

Die durchgeführte Literaturanalyse identifizierte aktuelle wissenschaftliche Veröffentlichungen im Bereich des Data Engineerings, welche mit Hilfe der Phasen des Data Engineering Lifecycle diesen Aufgabengebieten zugeordnet werden konnten. Die Anwendung der Konzeptmatrix nach Webster und Watson (2002) ermöglichte eine transparente Zuordnung, sowie eine anschließende quantitative Auswertung. Das Ergebnis der Auswertung zeigt, dass im Bereich des Data Engineerings aktuell aktive Forschung betrieben wird. Der Fokus der Forschung liegt dabei in den Aufgabenbereichen Ingestion und Transformation, gefolgt von Generierung und Speicherung. Im Aufgabenbereich der Bereitstellung wird aktuell am wenigsten geforscht.

Wichtig zu erwähnen ist, dass die Ergebnisse inklusive der Konzeptmatrix auf der Grundlage von 19 Literaturquellen erstellt wurden. Eine umfangreichere Literaturanalyse könnte differenziertere und aussagekräftigere Ergebnisse liefern und mögliche Extreme glätten oder bestätigen.

Zukünftige Forschungen könnten deshalb die Anzahl an Datenbankquellen erhöhen sowie die Zeitspanne von inkludierten Veröffentlichungen von zwei auf drei Jahre erhöhen.

Literatur

Alnafoosi, Ahmad Basim/Theresa A. Steinbach (2023): Empirical Assessment of Big Data Technology adoption Factors for Organizations with Data Storage Systems, in: *AMCIS 2023 Proceedings*, [online] https://aisel.aisnet.org/amcis2023/sig_adit/sig_adit/31.

Altendeitering, Marcel/Stephan Dübler/Tobias Moritz Guggenberger (2022): Data Quality in Data Ecosystems: Towards a Design Theory, in: *AMCIS 2022 Proceedings*, [online] https://aisel.aisnet.org/amcis2022/DataEcoSys/DataEcoSys/3.

Amutha, A./R. Mathu Pritha (2022): Fuzziness on interconnection networks under ratio labelling, in: *Smart innovation, systems and technologies*, S. 165–173, [online] doi:10.1007/978-981-16-6624-7_17.

Asaad, Chaimae/Karim Bäına/Mounir Ghogho (2023): Investigating the perceived usability of Entity-Relationship quality frameworks for NoSQL databases, in: *Lecture Notes in Computer Science*, S. 214–227, [online] doi:10.1007/978-3-031-49333-1_16.

Bhujange, Ketan/B. R. Chandavarkar/Pradeep Nazareth (2023): Implementing holding time based data forwarding in underwater opportunistic routing protocol using UnetStack3, in: *Smart innovation, systems and technologies*, S. 1–11, [online] doi:10.1007/978-981-19-7524-0_1.

D, Uma Priya/P. Santhi Thilagam (2023): JSON document clustering based on structural similarity and semantic fusion, in: *Lecture notes on data engineering and communications technologies*, S. 51–62, [online] doi:10.1007/978-981-99-0609-3_4.

Dashora, Rajnish/M. Rajasekhara Babu (2022): A Survey on Advancements of Real-Time Analytics Architecture Components, in: *Computational Methods and Data Engineering. Lecture Notes on Data Engineering and Communications Technologies*, S. 547–559, [online] doi:10.1007/978-981-19-3015-7_41.

Fettke, Peter (2006): State-of-the-Art des State-of-the-Art, in: *Wirtschaftsinformatik und Angewandte Informatik*, Bd. 48, Nr. 4, [online] doi:10.1007/s11576-006-0057-3.

Gantayat, Pradosh Kumar/Sadhna Mohapatra/Sandeep Kumar Panda (2022): Secure trust level routing in Delay-Tolerant network with node categorization technique, in: *Smart innovation, systems and technologies*, S. 453–458, [online] doi:10.1007/978-981-16-6624-7_45.

Jain, Tavishi/Bhavya Singh/Rupesh Kumar Dewang (2022): Event Detection in Live Twitter Streams Using Tf-Idf and Clustering Algorithms, in: *Data, Engineering and Applications. Lecture Notes in Electrical Engineering*, S. 469–480, [online] doi:10.1007/978-981-19-4687-5_36.

Jaiswal, Kavita/Veena Anand (2022): Efficient Fault-Tolerant Cluster-Based Approach for Wireless Sensor Networks, in: *Smart innovation, systems and technologies*, S. 293–300, [online] doi:10.1007/978-981-16-6624-7_29.

Jouseau, Roxane/Sébastien Salva/Chafik Samir (2022): On studying the effect of data quality on classification performances, in: *Lecture Notes in Computer Science*, S. 82–93, [online] doi:10.1007/978-3-031-21753-1_9.

Kaur, Mandeep/Rajni Aron (2022): FOG clustering-based architecture for load balancing in scientific workflows, in: *Lecture notes on data engineering and communications technologies*, S. 213–221, [online] doi:10.1007/978-981-16-7182-1_18.

Khatri, Sushma/Pooja Bansal (2022): Feature Selection Using Information Gain for Software Effort Prediction Using Neural Network Model, in: *Data, Engineering and Applications. Lecture Notes in Electrical Engineering*, S. 177–198, [online] doi:10.1007/978-981-19-4687-5_14.

Leimeister, Jan Marco (2021): *Einführung in die Wirtschaftsinformatik*, Springer e-Books, [online] doi:10.1007/978-3-662-63560-5.

Liebregts, Werner/Willem-Jan van den Heuvel/Arjan van den Born (2023): *Data science for entrepreneurship*, Springer Cham, [online] doi:10.1007/978-3-031-19554-9.

Madhuri, A./S. Sindhura/D. Swapna/S. Phani Praveen/T. Sri Lakshmi (2022): Distributed computing meets movable wireless communications in next generation mobile communication networks (NGMCN), in: *Lecture notes on data engineering and communications technologies*, S. 125–136, [online] doi:10.1007/978-981-19-3015-7_10.

Papp, Stefan/Wolfgang Weidinger/Mario Meir-Huber/Bernhard Ortner/Georg Langs/Rania Wazir (2019): *Handbuch Data Science: Mit Datenanalyse und Machine Learning Wert aus Daten generieren*.

Pohl, Matthias/Christian Haertel/Daniel Staegemann/Klaus Turowski (2023): Data Valuation Methods - A Literature review, in: *AMCIS 2023 Proceedings*, [online] https://aisel.aisnet.org/amcis2023/asys/sig_asys/1.

Purnaye, Prasad/Vrushali Kulkarni (2022): Information retrieval for cloud forensics, in: *Smart Innovation, Systems and Technologies*, S. 11–18, [online] doi:10.1007/978-981-16-6624-7_2.

Reis, Joe/Matt Housley (2023): *Handbuch Data Engineering: Robuste Datensysteme planen und erstellen*, 1. Aufl., O'Reilly, [online] https://dpunkt.de/produkt/handbuch-data-engineering/.

Rose, Doug (2016): *Data science*, Apress eBooks, [online] doi:10.1007/978-1-4842-2253-9.

Sahay, Akshat/Sinkon Nayak/Siddharth Swarup Rautaray/Manjusha Pandey (2022): Application-Oriented content quality analysis of data using Python, in: *Lecture notes in networks and systems*, S. 25–32, [online] doi:10.1007/978-981-19-1559-8_4.

Sarkar, Moumita Majumder/Manish Dhananjay Sawale (2022): Pragmatic Analysis of Web Service Discovery Models and Architectures from a Qualitative Perspective, in: *Data, Engineering and Applications. Lecture Notes in Electrical Engineering*, S. 301–316, [online] doi:10.1007/978-981-19-4687-5_23.

Schallmo, Daniel R. A. (2023): *Digitalisierung*, Schwerpunkt Business Model Innovation, [online] doi:10.1007/978-3-658-36634-6.

Shearer, Colin (2000): The CRISP-DM Model: The New Blueprint for Data Mining, in: *Journal of Data Warehousing*, Bd. 4, Nr. 5, S. 13–22.

Stieglitz, Stefan/Christian Wiencierz (2022): Digitalisierung, Big Data und soziale Medien als Rahmenbedingungen der Unternehmenskommunikation, in: *Springer eBooks*, S. 289–309, [online] doi:10.1007/978-3-658-22933-7_10.

Tamir, Mike/Steven Miller/Alessandro Gagliardi (2015): The data engineer, in: *Social Science Research Network*, [online] doi:10.2139/ssrn.2762013.

Tandon, Nandni/A. Sharma (2022): Composite Reversible data hiding scheme for secure image reconstruction, in: *Lecture notes in electrical engineering*, S. 553–564, [online] doi:10.1007/978-981-19-4687-5_43.

Webster, Jane/Richard T. Watson (2002): Analyzing the past to prepare for the future: writing a literature review, in: *MIS Quarterly*, Bd. 26, Nr. 2, S. 3.

Zerfaß, Ansgar/Ulrike Röttger/Manfred Piwinger (2022): *Handbuch Unternehmenskommunikation*, Springer eBooks, [online] doi:10.1007/978-3-658-22933-7.