Prof. Dr. Uta Störl

63123
Data Engineering für Data Science

Begleittext

Leseprobe

Fakultät für
Mathematik und Informatik

FernUniversität in Hagen

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m^2, weiß), hergestellt aus 100 % Altpapier.

Vorwort

Liebe Fernstudent:innen,

wir begrüßen Sie herzlich zum Modul „Data Engineering für Data Science“.

Sie haben sich entschlossen, Data Science zu studieren oder sich als Informatikstudentin oder -student mit dem Thema Data Engineering zu beschäftigen – herzlichen Glückwunsch zu dieser Entscheidung! Wenn Sie bereits in diesem Gebiet in der Praxis tätig sind oder zukünftig tätig sein werden, werden Sie feststellen, dass Data Scientists einen Großteil (bis zu 80%) ihrer Arbeitszeit damit verbringen, Datensätze zusammenzutragen, zu bereinigen und zu transformieren [For16]. Diese Fähigkeiten und das Wissen um geeignete Technologien für die Verarbeitung großer Datenmengen sind eine essentielle Voraussetzung für eine erfolgreiche Arbeit im Bereich Data Science und deshalb Gegenstand dieses Moduls.

Nach der erfolgreichen Bearbeitung dieses Moduls sollten Sie

- die grundlegenden Programmierkonzepte von Python und Best Practices für guten Python-Code kennen,
- die Fähigkeit besitzen, sich mit neuen Daten vertraut zu machen und die Beschaffenheit und Qualität der Daten analysieren können,
- Methoden zum Daten bereinigen kennen und anwenden können,
- die Fähigkeit besitzen, Daten zu transformieren,
- Methoden zur Datenanalyse kennen und anwenden können,
- die Konzepte verschiedener Big-Data-Analyse-Referenzarchitekturen und deren Vor- und Nachteile kennen,
- Daten geeignet in verschiedenen NoSQL-Datenmodellen modellieren und unterschiedliche Varianten der Modellierung bewerten können,
- wichtige Techniken zur Realisierung einer horizontalen Skalierung der Datenverarbeitung verstanden haben und diese erläutern können,
- die Vor- und Nachteile verschiedener Datenbanktechnologien für die Verarbeitung von Big Data erläutern und einschätzen können, wann Sie welche Datenbanktechnologie einsetzen.

Hinweise zur Durchführung des Moduls

Dieses Modul teilt sich in zwei große Teile:

Teil I (Lektionen 1–4) beschäftigt sich mit den Grundlagen der Vorbereitung der Daten (Analyse der Beschaffenheit der Daten, Datenbereinigung etc.) und Methoden der Datenanalyse und Datenvisualisierung. Begleitend werden dabei Programmierkenntnisse in Python erworben.

Dieser Teil wird anhand des folgenden Buches [McG21] bearbeitet:
Susan E. McGregor:
Practical Python Data Wrangling and Data Quality
O'Reilly Media, Inc., 2021
ISBN: 9781492091509
`https://www.oreilly.com/library/view/practical-python-data/9781492091493/`

Dieses Buch benötigen Sie zwingend zur Bearbeitung dieses Moduls! Das Buch ist in einem gut verständlichen Englisch geschrieben und nicht in einer deutschen Übersetzung verfügbar.

In diesem *Begleittext* geben wir Ihnen zu den jeweiligen Lektionen eine kurze Einführung und Lesehinweise zu den jeweiligen Kapiteln des Buches.

Teil II (Lektionen 5–7) führt in die Grundlagen der Systeme ein, mit denen große Datenmengen gespeichert und verarbeitet werden können (Technologien für Big Data).

Dieser Teil wird anhand von **Videoaufzeichnungen** bearbeitet, welche Ihnen über die **Moodle**-Lernumgebung zur Verfügung gestellt werden.

In diesem *Begleittext* geben wir Ihnen zu den jeweiligen Lektionen eine kurze Einführung zu den jeweiligen Themengebieten sowie weitere Literaturhinweise.

Die nachfolgende Tabelle zeigt die Zuordnung der sieben Lektionen[1] zu den jeweiligen Themengebieten::

Teil	Lektion	Inhalt
I	1	Einführung
I	2	Beschaffenheit von Daten und Datenqualität
I	3	Datenbereinigung (Data Cleaning)
I	4	Datenanalyse und -visualisierung
II	5	Analyse von Big Data
II	6	NoSQL: Datenmodelle und Techniken
II	7	Column Stores und In-Memory-Datenbanksysteme

Voraussetzungen

Dieses Modul setzt grundlegende Kenntnisse im Bereich Datenbanken, wie sie beispielsweise im Modul „Datenbanken“ vermittelt werden, voraus. Ebenfalls vorausgesetzt werden Programmierkenntnisse in einer „beliebigen“ Programmiersprache.

[1] Bitte beachten Sie: Bis zum Wintersemester 2023/24 wurden *Lektionen* an der FernUniversität als *Kurseinheiten* bezeichnet. Wir haben versucht, die Unterlagen größtenteils anzupassen – beachten Sie bitte dennoch die Äquivalenz dieser beiden Begriffe.

Praktische Übungen

Zur Vertiefung des Verständnisses der in diesem Modul vorgestellten Inhalte bieten wir in diesem Modul **praktische Übungen** an. Sie finden diese Aufgaben im **Online-Übungssystem** bzw. im **Moodle** zu diesem Modul. Im Moodle finden Sie (im Abschnitt Infrastruktur) auch Hinweise zur Installation der benötigen Software. **Nutzen Sie unbedingt unsere Softwareumgebung für die Bearbeitung der Aufgaben!** Verwenden Sie **nicht** die im obigen Buch vorgeschlagene Softwareumgebung (d.h. **ignorieren** Sie die Seiten 14-23 aus dem Buch „Practical Python Data Wrangling and Data Quality“).

Die Bereitstellung und Bearbeitung der praktischen Übungen erfolgt zu den üblichen Bearbeitungszeiträumen für die einzelnen Lektionen.

Für die Prüfung wird die gleiche Softwareumgebung verwendet wie für die praktischen Übungen. Aus diesem Grund müssen von den Einsendeaufgaben zu den Lektionen 2–6 mindestens die Einsendeaufgaben zu einer Lektion bestanden sein.

Beachten Sie bitte, dass einige Fragen in den Aufgaben bewusst eher „offen“ gestellt sind und es nicht immer DIE eine eindeutige Lösung gibt. Das liegt im Charakter der praktischen Aufgabenstellungen im Bereich Datenvorverarbeitung und Datenqualität.

Rechnerausstattung für praktische Übungen und Klausur

Die Anforderungen an Hardware, die Sie im Masterstudiengang Data Science verwenden, sind u.U. etwas höher als die Anforderungen in anderen Studiengängen. Für die Bearbeitung dieses Moduls und die Teilnahme an der Online-Klausur ist eigene Hardware notwendig. Sie benötigen Hardware mit den folgenden Parametern:

- 16 GB RAM
- CPU mit 4 Cores
- 4 GB freier Festplattenspeicher

Prüfung

Dieses Modul wird in Form einer schriftlichen Online-Klausur geprüft. Die Inhalte aller Lektionen sind prüfungsrelevant. **Für die Prüfung müssen Sie die gleiche Softwareumgebung verwenden wie für die praktischen Übungen**.

Literatur

Auf hilfreiche und weiterführende Literatur wird in den Literaturhinweisen am Ende jedes Kapitels hingewiesen.

Die Autorin

Prof. Dr. Uta Störl ist seit April 2021 an der FernUniversität in Hagen tätig und leitet das Lehrgebiet „Datenbanken und Informationssysteme“. Sie hat an der Friedrich-Schiller-Universität in Jena studiert und promoviert. Danach war sie mehrere Jahre bei der Dresdner Bank in Frankfurt am Main tätig. Von 2005 bis 2021 war sie Professorin für Datenbanken an der Hochschule Darmstadt und hat dort das Big Data Competence Center aufgebaut und geleitet. Ihr Forschungsschwerpunkt sind Big-Data-Technologien, insbesondere NoSQL-Datenbanksysteme und Data Engineering für Data Science.

Danksagung

Die Erstellung eines neuen Moduls ist nicht möglich, ohne die Unterstützung von engagierten und kompetenten Mitarbeiterinnen und Mitarbeitern. Ich möchte mich ganz herzlich bei **Valerie Restat** und **André Conrad** für die vielen Diskussionen, den wertvollen inhaltlichen Input und vor allem die Erstellung der praktischen Übungen und der dafür benötigten Infrastruktur bedanken.

Inhaltsverzeichnis

Teil I.

Datenvorverarbeitung und Datenqualität

1. Einführung

Bitte lesen Sie zur Einführung in Teil I des Moduls im Buch das Vorwort (*Preface*), Kapitel 1 *Introduction to Data Wrangling and Data Quality*, Kapitel 2 *Introduction to Python* und Kapitel 3 *Understanding Data Quality*. Überspringen Sie dabei bitte die Seiten 14–23, da wir eine andere Softwareumgebung nutzen.

Für die Einsendeaufgaben machen Sie sich bitte mit der Umgebung vertraut, in der diese durchgeführt werden. Detaillierte Informationen dazu finden Sie im Abschnitt Infrastruktur im Moodle zum Modul. **Nutzen Sie unbedingt unsere Softwareumgebung für die Bearbeitung der Aufgaben!** Verwenden Sie **nicht** die im Buch vorgeschlagene Umgebung (d.h. ignorieren Sie die Seiten 14-23 aus dem Buch).

Das vorliegende Modul ist ein Pflichtmodul im Rahmen des Studiengangs Data Science. Uns ist bewusst, dass Sie mit sehr unterschiedlichen Vorkenntnissen im Bereich Programmieren diesen Studiengang beginnen. Für einige von Ihnen wird diese erste Lektion daher eher den Charakter einer Wiederholung haben. Für andere wird sie hingegen herausfordernd sein. Schließen Sie daraus, wie leicht oder schwer Ihnen diese erste Einheit fällt, nicht auf den weiteren Verlauf des Moduls. Diese Anmerkung gilt ebenso für die Studierenden der Informatik. Ihnen werden die ersten Lektionen eher leicht fallen – schließen Sie daraus nicht auf die nachfolgenden Lektionen.

1.1. Einführung in Data Wrangling und Datenqualität

Für die Qualität der Ergebnisse einer Datenanalyse ist neben den verwendeten Analysemethoden die Qualität der verwendeten Daten von absolut entscheidender Bedeutung. In dieser einführenden Lektion werden Sie die verschiedenen Prozessschritte in der Datenvorverarbeitung kennenlernen. Außerdem wird der Begriff der Datenqualität eingeführt und verschiedene Aspekte der Datenqualität erläutert.

1.2. Einführung in Python

Es gibt verschiedene Methoden, um Daten vorzuverarbeiten und zu analysieren. Aufgrund ihrer Einfachheit, guten Lesbarkeit und der vielen verfügbaren Bibliotheken ist derzeit die Programmiersprache *Python* für die Lösung dieser Aufgaben sehr weit verbreitet. Dennoch möchten wir betonen, dass es auch Alternativen wie *R* und *SQL* für diese Aufgaben gibt. Ziel dieses Moduls ist also nicht, Sie zu Python-Expert:innen auszubilden, sondern die Konzepte der Datenvorverarbeitung und Datenanalyse anhand einer ausgewählten Programmiersprache – in diesem Fall Python – zu vermitteln. In dieser Lektion wird deshalb in die grundlegenden Konzepte von Python eingeführt.

1.3. Lernziele

Nach der Bearbeitung der ersten Lektion sollten Sie

- die Grundkonzepte von Data Wrangling (Datenvorverarbeitung) verstanden haben,

- ein erstes Verständnis für den Begriff der Datenqualität erhalten haben,
- die grundlegenden Programmierkonzepte von Python kennen,
- die praktische Fähigkeit erworben haben, mit Jupyter Notebooks arbeiten zu können.

2. Beschaffenheit von Daten und Datenqualität

Lesen Sie für diese Lektion aus Kapitel 4 *Working with File-based and Feed-based Data in Python* die Seiten 91–97 (bis zum Abschnitt *Working with Structured Data*) und in Kapitel 5 *Accessing Web-based Data* die Seiten 141–148 (bis zum Abschnitt *Specialized APIs*). Kapitel 6 *Assessing Data Quality* lesen Sie bitte vollständig.

2.1. Datenbeschaffenheit

Neben den technischen Aspekten, wie man sich, also mit Hilfe welcher APIs, Daten beschaffen kann, ist es wichtig, sich mit der unterschiedlichen Beschaffenheit der Daten auseinanderzusetzen. Daten können beispielsweise strukturiert, semi-strukturiert oder unstrukturiert sein. Diese Eigenschaften haben Konsequenzen für die weitere Verarbeitung. Außerdem können die Daten in unterschiedlichen Dateiformaten und/oder Zeichenkodierungen vorliegen. Am Anfang des Data-Engineering-Prozesses ist es folglich wichtig, sich zunächst mit den Daten und ihrer Beschaffenheit vertraut zu machen.

2.2. Datenqualität

Nach der grundsätzlichen Analyse der Beschaffenheit der Daten folgt die Analyse der Qualität der Daten. In dieser Lektion lernen Sie kennen, welche verschiedenen Aspekte der Begriff der Datenqualität beinhaltet. Außerdem lernen Sie verschiedene Methoden kennen, mit deren Hilfe Sie die Qualität der Daten beurteilen können und werden diese praktisch anwenden.

2.3. Lernziele

Nach der Bearbeitung dieser Lektion sollten Sie in der Lage sein,

- die Beschaffenheit von Daten zu analysieren,
- Daten in ihre Arbeitsumgebung laden können,
- die Fähigkeit besitzen, sich mit neuen Daten vertraut zu machen,
- die Qualität der Daten analysieren können.

3. Datenbereinigung (Data Cleaning)

Bitte lesen Sie Kapitel 7 *Cleaning, Transforming, and Augmenting Data* und Kapitel 8 *Structuring and Refactoring Your Code* im Buch. Lesen Sie außerdem zum Thema Fairness und Bias das im Moodle bereitgestellte Paper [SS20].

3.1. Data Cleaning

Nachdem Sie in Lektion 2 kennengelernt haben, wie Sie die Qualität der Daten analysieren können, werden wir in dieser Lektion Methoden kennenlernen, wie Qualitätsprobleme in den Daten ggf. behoben werden können. Dieser Prozessschritt der Datenbereinigung wird in der Literatur i.A. als *Data Cleaning* bezeichnet. In den praktischen Übungen werden Sie unterschiedliche Methoden des Data Cleaning anwenden.

3.2. Data Transformation

Nachdem die Daten bereinigt wurden, müssen sie ggf. noch transformiert werden, um sie in die für die Analyse (siehe nächste Lektion) geeignete Form zu bringen. Verschiedene Methoden zur Transformation der Daten werden eingeführt.

3.3. Lernziele

Nach der Bearbeitung dieser Lektion sollten Sie

- Methoden zum Bereinigen von Daten kennen und anwenden können,
- die Fähigkeit besitzen, Daten zu transformieren,
- Best Practices für guten Python-Code kennen.

3.4. Literaturhinweise

Vertiefende Hinweise zu den Themen *Anomaly Detection* und *Missing Value Imputation* sind in folgenden Papern zu finden:

- Anomaly detection: A survey, Chandola et al. [CBK09]
- Missing value imputation: a review and analysis of the literature (2006–2017), Lin und Tsai [LT20]

4. Datenanalyse und -visualisierung

Bitte lesen Sie Kapitel 9 *Introduction to Data Analysis* und Kapitel 10 *Presenting Your Data* im Buch.

4.1. Datenanalyse

Nachdem wir in den vorigen Lektionen kennengelernt haben, wie wir die Qualität von Daten analysieren, die Daten bereinigen und für die Analyse vorbereiten können, steht nun die Datenanalyse im Mittelpunkt dieser Lektion. Wir werden verschiedene Methoden zur Datenanalyse kennenlernen und praktisch anwenden.

4.2. Datenvisualisierung

Zum Verständnis der Analyseergebnisse von Daten sind Visualisierungen ein wichtiges Mittel. Wir werden deshalb einige Visualisierungstechniken kennenlernen und erproben.

4.3. Lernziele

Nach der Bearbeitung dieser Lektion sollten Sie

- Methoden zur Datenanalyse kennen und anwenden können,
- unterschiedliche Visualisierungskonzepte kennen und anwenden können.

Teil II.

Technologien für Big Data

5. Analyse von Big Data

Für Teil II werden Ihnen Vorlesungseinheiten per Video bereitgestellt sowie zu den jeweiligen Lektionen vertiefende Literaturhinweise gegeben.

Sehen Sie sich die Videos zu dieser Lektion an. Die Links dazu finden Sie im Moodle zu diesem Modul.

5.1. Large Scale Parallel Data Processing

Zunächst wird eine Einführung und Übersicht über Teil II gegeben und in die grundlegenden Prinzipien und Ansätze zur Verarbeitung großer Datenmengen (Big Data) eingeführt. Es werden Programmiermodelle zur parallelen und verteilten Verarbeitung großer Datenmengen (u.a. MapReduce) eingeführt.

5.2. Big-Data-Referenzarchitekturen

Anschließend werden wichtige Big-Data-Referenzarchitekturen und ausgewählte Frameworks zur Big-Data-Analyse (*Apache Hadoop* und *Apache Spark*) vorgestellt. In der Einsendeaufgabe zu dieser Lektion werden wir exemplarisch *Apache Spark* zur Datenanalyse verwenden.

5.3. Lernziele

Nach der Bearbeitung dieser Lektion sollten Sie

- die Konzepte verschiedener Big-Data-Analyse-Referenzarchitekturen und deren Vor- und Nachteile kennen,
- die Konzepte der parallelen Verarbeitung von Daten verstanden haben und diese praktisch anwenden können.

5.4. Literaturhinweise

Vertiefende Informationen zu verschiedenen Konzepten der Big-Data-Analyse finden sich u.a. im Buch von Kleppmann [Kle17]. Dieses Buch ist darüber hinaus für den gesamten Teil II dieses Module empfehlenswert.

6. NoSQL: Datenmodelle und Techniken

Sehen Sie sich die Videos zu dieser Lektion an.

Für das Thema NoSQL-Datenmodellierung (Lektion 6.1) stellen wir Ihnen einen zusätzlichen Lehrtext im Moodle bereit.

6.1. NoSQL-Datenmodelle und -Datenmodellierung

Für die Verarbeitung großer Datenmengen in einer horizontalen Verteilungsarchitektur haben sich in den letzten Jahren NoSQL-Datenbanksysteme etabliert. In dieser Lektion werden wir uns zunächst mit den verschiedenen Arten von NoSQL-Datenbanksystemen, den unterschiedlichen NoSQL-Datenbankmodellen und den Konsequenzen für die Datenmodellierung beschäftigen. Anschließend wird auf einige Aspekte der Anwendungsentwicklung mit NoSQL-Datenbanken eingegangen.

6.2. NoSQL-Techniken

Um die Verarbeitung von großen Datenmengen horizontal verteilt über eine Vielzahl von Servern zu ermöglichen, müssen einige grundlegende Aspekte der Datenbanktechnologie im Vergleich zur klassischen vertikalen Skalierung verändert werden. Wir werden die Wichtigsten davon kennenlernen.

6.3. Lernziele

Nach der Bearbeitung dieser Lektion sollten Sie

- die unterschiedlichen NoSQL-Datenmodelle kennen und ihre Vor- und Nachteile erläutern können,
- Daten geeignet in den verschiedenen NoSQL-Datenmodellen modellieren und verschiedene Varianten der Modellierung bewerten können,
- wichtige Techniken zur Realisierung einer horizontalen Skalierung der Datenverarbeitung verstanden haben und diese erläutern können.

6.4. Literaturhinweise

Eine generelle Einführung in die Konzepte und Ideen von NoSQL findet sich im Buch von Sadalage und Fowler [SF12]. Eine deutschsprachige Einführung sowie weitere Ausführungen zur NoSQL-Datenmodellierung findet sich in Kapitel 12 der 2. Auflage des Taschenbuchs Datenbanken [Kud15]. Für das Thema NoSQL-Datenmodellierung stellen wir Ihnen darüber hinaus einen zusätzlichen Lehrtext im Moodle bereit. Techniken zur Realisierung einer horizontalen Skalierung werden in Kapitel 13 in [Kud15] beschrieben. Detaillierte Beschreibungen dieser Techniken finden sich in den Büchern von Wiese [Wie15] bzw. Gessert et al. [GWR20].

7. Column Stores und In-Memory-Datenbanksysteme

Sehen Sie sich die Videos zu dieser Lektion an.

7.1. Column Stores

Neben den in Lektion 6.1 vorgestellten Datenbanktechnologien zur vertikalen Skalierung gibt es weitere wichtige Technologien, mit denen die Verarbeitung von Big Data ermöglicht wird. Zunächst werden wir uns mit Column Stores beschäftigen und die diesen Datenbanksystemen zugrundeliegenden Techniken kennenlernen.

7.2. In-Memory-Datenbanksysteme

Anschließend beschäftigen wir uns mit In-Memory-Datenbanksystemen und den in diesen Systemen eingesetzten Technologien.

7.3. Lernziele

Nach der Bearbeitung dieser Lektion sollten Sie

- die Konzepte von Column Stores verstanden haben,
- In-Memory-Datenbanksysteme und ihre Technologien erläutern können,
- die Vor- und Nachteile der verschiedenen Datenbanktechnologien für die Verarbeitung von Big Data erläutern können,
- einschätzen können, wann Sie welche Datenbanktechnologie einsetzen.

7.4. Literaturhinweise

Vertiefende Informationen zu In-Memory-Datenbanksystemen finden sich im Buch von Plattner und Zeier [PZ12] bzw. dem Lehrbuch von Plattner [Pla14] (deutschsprachige Ausgabe: [Pla13]). In diesen Büchern wird auch auf Grundlagen von Column Stores eingegangen. Weitere Beschreibungen von in Column Stores eingesetzten Techniken finden sich in einem umfangreichen Aufsatz von Abadi et al. [ABH+13] bzw. den korrespondierenden Tutorial-Folien[1] ebenfalls von Abadi et al. [ABH09]. Techniken zur hybriden zeilen- und spaltenorientierten Speicherung werden in [RDHF12] beschrieben.

[1] `https://de.slideshare.net/abadid/vldb-2009-tutorial-on-columnstores`

Literaturverzeichnis

[ABH09] Daniel J. Abadi, Peter A. Boncz, and Stavros Harizopoulos. Column oriented database systems. *Proc. VLDB Endow.*, 2(2):1664–1665, 2009.

[ABH+13] Daniel J. Abadi, Peter A. Boncz, Stavros Harizopoulos, Stratos Idreos, and Samuel Madden. *The Design and Implementation of Modern Column-Oriented Database Systems.* Transatlantic Publisher, 2013.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009.

[For16] Forbes. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, 2016.

[GWR20] Felix Gessert, Wolfram Wingerath, and Norbert Ritter. *Fast and Scalable Cloud Data Management.* Springer, 2020.

[Kle17] Martin Kleppmann. *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems.* O'Reilly, 2017.

[Kud15] Thomas Kudraß, editor. *Taschenbuch Datenbanken.* Hanser, 2nd edition, 2015.

[LT20] Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006-2017). *Artif. Intell. Rev.*, 53(2):1487–1509, 2020.

[McG21] Susan E. McGregor. *Practical Python Data Wrangling and Data Quality.* O'Reilly Media, Inc, 2021.

[Pla13] Hasso Plattner. *Lehrbuch In-Memory Data Management: Grundlagen der In-Memory-Technologie.* Springer, 2013.

[Pla14] Hasso Plattner. *A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases.* Springer, 2 edition, 2014.

[PZ12] Hasso Plattner and Alexander Zeier. *In-Memory Data Management: Technology and Application.* Springer, 2 edition, 2012.

[RDHF12] Philipp Rösch, Lars Dannecker, Gregor Hackenbroich, and Franz Faerber. A Storage Advisor for Hybrid-Store Databases. *Proc. VLDB Endow.*, 5(12):1748–1758, 2012.

[SF12] Pramod J. Sadalage and Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence.* Addison-Wesley, 2012.

[SS20] Sebastian Schelter and Julia Stoyanovich. Taming Technical Bias in Machine Learning Pipelines. *IEEE Data Eng. Bull.*, 43(4):39–50, 2020.

[Wie15] Lena Wiese. *Advanced Data Management for SQL, NoSQL, Cloud and Distributed Databases.* DeGruyter, 2015.

Diese Seite bleibt aus technischen Gründen frei!

2002413
(04/25)

63123-01-S#1