technische universität dortmund

itpl

Technische Universität Dortmund

Fachwissenschaftliche Projektarbeit

Data Mining-Werkzeuge und ihre Schnittstellen zu Datenbankmanagementsystemen

abgegeben von

Thomas Rellensmann

Matrikelnr. 175791

Maschinenbau (BA)

betreut von

M. Sc. J. Hunker

Abgabedatum: 1. Februar 2019

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ACID Atomarity Consistency Isolation Durability

ADO ActiveX Data Objects

ADO.NET ActiveX Data Objects .NET

API Application Programming Interface

BI Business Intelligence

COM Component Object Model

JDBC Java Database Connectivity

JSON Java Script Object Notation

KDD Knowledge Discovery in Databases

NoSQL NotOnlySQL

ODBC Open Database Connectivity

OLAP Online Analytical Processing

OLE-DB Object Linking and Embedding - Database

SSAS Microsoft SQL Server Analysis Services

SQL Structured Query Language

XML eXtensible Markup Language

Inhaltsverzeichnis

1 Einleitung

Die weltweit generierte Datenmenge wird sich nach Schätzungen der International Data Corporation von 16,1 Zettabyte im Jahr 2016 auf 163 Zettabyte im Jahr 2025 verzehnfachen [RGR17, S. 3]. Dieser Umstand verdeutlicht das Ausmaß und die Geschwindigkeit der digitalen Transformation, in der sich unsere Gesellschaft befindet. Durch die massenhafte Generierung, Auswertung und Bereitstellung von Daten erschließen sich im industriellen wie privaten Umfeld Entwicklungs- und Optimierungspotentiale. Dies hat zur Folge, dass neben den klassischen Produktionsfaktoren menschliche Arbeit, Betriebsmittel und Werkstoffe, auch die Information den wirtschaftlichen Erfolg eines Unternehmens mitbestimmt und daher zunehmende Berücksichtigung findet. Infolgedessen erweitert sich das Aufgabenspektrum der Unternehmen: Auf der einen Seite muss die Erhebung von Daten geplant, gesteuert und überwacht werden und auf der anderen Seite erfordert die Extraktion von Wissen aus diesen Daten fachliche Expertise und entsprechende Hard- und Software [MK16, S. 3]. Um diesen Aufgaben gerecht zu werden, hat sich eine Vielzahl von Software-Anwendungen etabliert. Es werden Datenbanksysteme genutzt, um Daten verschiedenster Art in persistenten Strukturen zu speichern und zu verwalten. Dazu dienen zwei Komponenten: Die Datenbank selbst enthält und speichert die Daten sowie die Beschreibung ihrer Struktur und eine Verwaltungskomponente - das Datenbankmanagementsystem – stellt die Schnittstelle des Benutzers zur Datenbank her und kann über eine Abfrage- und Manipulationssprache auf die Daten zugreifen und sie verändern [MK16, S.2]. Die etablierten Datenbankmanagementsysteme basieren auf verschiedenen Datenmodellen und weisen unterschiedliche Eigenschaften etwa hinsichtlich der Performanz des Systems und der Konsistenz der Daten auf. Des Weiteren existieren Anwendungen, um aus den in einer Datenbank abgelegten Daten mithilfe von Techniken des Data Minings Wissen zu extrahieren. Der Zugriff auf eine Datenbank vonseiten des Data Mining-Programms ist hierbei zwingend notwendig und wird über unterschiedlich implementierte Schnittstellen realisiert.

Das Ziel dieser Projektarbeit ist, den Stand der Technik in Bezug auf diese Schnittstellen zu untersuchen. Die Arbeit soll eine Hilfestellung bieten, um je nach Art und Umfang der anfallenden Daten eine günstige und kompatible Kombination von Datenbankmanagementsystem und Data Mining-Programm auszuwählen. Diese Kombinationen sollen als Ergebnis der Arbeit in einer kompakten grafischen Übersicht dargestellt werden. Bei der Auswahl eines Datenbanksystems ist zunächst zwischen verschiedenen Datenmodellen zu unterschieden, auf welchen die Datenbankmanagementsysteme basieren. Auch die Schnittstelle zwischen Data Mining-Programm und Datenbank wird in Abhängigkeit des zugrundliegenden Datenmodells, wie in Kapitel

3.1 gezeigt wird, unterschiedlich realisiert. Aus diesem Grund werden in Kapitel 2 zunächst die Eigenschaften des relationalen und verschiedener postrelationaler Modelle dargestellt. Als zweiter Schritt erfolgt in Kapitel 3 die Untersuchung von Data Mining-Anwendungen in Hinblick auf ihre Möglichkeiten, Daten aus Datenbanksystemen zu importieren. Hierfür dienen unterschiedliche Programmierschnittstellen wie etwa Open Database Connectivity (ODBC) oder Object Linking and Embedding - Database (OLE-DB), welche zunächst hinsichtlich ihrer Funktionsweise und ihrer Eigenschaften untersucht werden. Auf dieser Basis erfolgt im Anschluss die detaillierte Betrachtung der Schnittstelle von einer Auswahl etablierter Data Mining-Programme. Hierbei liegt der Fokus darauf, zu welchen Datenbanksystemen eine Schnittstelle existiert beziehungsweise eingerichtet werden kann und wie dies geschieht. Dazu wird untersucht, welche Programmierschnittstelle(n) das Data Mining-Programm unterstützt und welche Datenbanken sich darüber anbinden lassen. Die Ergebnisse der beiden Kapitel werden im letzten Schritt in Kapitel 4 mit der Erstellung einer grafischen Übersicht zusammengeführt und gebündelt.

2 Datenbanksysteme

Die Aufgabe eines Datenbanksystems ist nach Steiner (2017) die Verwaltung von beliebigen Daten, die Bereitstellung von Informationen aus diesen Daten sowie deren Sicherung vor dem Zugriff unbefugter Personen [Ste17, S. 5]. Die Verwaltungsaufgaben umfassen dabei „das Eingeben von neuen Daten, das Löschen veralteter Daten sowie das Nachführen bestehender Daten“ [Ste17, S. 5]. Hierfür kommen zwei Komponenten zum Einsatz: Die gespeicherten Daten in Form einer Datenbank, sowie das Datenbankmanagementsystem. Die existierenden Datenbanksysteme unterscheiden sich hinsichtlich des Aufbaus und der Funktionsweise dieser beiden Komponenten. Insbesondere zwei verschiedene Modellansätze können dabei unterschieden werden: Die am meisten verbreiteten Datenbankmanagementsysteme basieren auf dem relationalen Datenmodell und werden, abgeleitet von der dabei verwendeten Abfragesprache SQL (Structured Query Language), auch SQL-Datenbanken genannt [Mei18, S. 9]. Durch die veränderten Anforderungen, die sich aus der digitalen Transformation an Datenbankmanagementsysteme ergeben, rücken neuerdings auch alternative, postrelationale Ansätze in das Interesse von Industrie und Forschung, welche unter dem Begriff NoSQL-Datenbanken (NotOnlySQL) zusammengefasst werden [Mei18, S. 9]. Beide Konzepte sollen im Folgenden in ihren Grundzügen dargestellt werden.

2.1 Relationale Datenbanksysteme

Das relationale Datenmodell wurde Anfang der Siebzigerjahre durch den englischen Mathematiker Edgar Frank Codd konzipiert [MK16, S. 6]. Es kennt zur Abbildung von Daten ein einziges Konstrukt: die Tabelle, auch Relation genannt [Stu16, S. 9]. In Abbildung 1 ist der grundsätzliche Aufbau und die Nomenklatur einer Relation dargestellt. Ein Datensatz, ein sogenanntes Tupel, entspricht einer Zeile der Tabelle [Ste17, S. 14]. Die Spalten werden als Attribute bezeichnet und die einzelnen Zellen enthalten den jeweiligen Attributwert [Ste17, S. 14]. Die Reihenfolge der Zeilen und Spalten ist dabei regellos und hat, angelehnt an die Unordnung von Mengen im mathematischen Sinn, keine Bedeutung [Mei18, S. 16]. Um jeden Datensatz eindeutig identifizieren zu können, wird ihm ein Identifikationsschlüssel, bestehend aus einem Attributwert (in Abbildung 1 das Attribut PNr.) oder einer minimalen Kombination verschiedener Attributwerte zugeordnet [MK16, S. 4]. Beziehungen zwischen mehreren Tabellen lassen sich abbilden, indem in einer Tabelle die Identifikationsschlüssel einer anderen Tabelle referenziert werden. Auch die Erstellung einer Beziehungstabelle, welche lediglich die Identifikationsschlüssel der miteinander in Beziehung stehenden Tabellen als Fremdschlüssel enthält, ist möglich [MK16, S. 19]. So ließe sich etwa die Tabelle ‚Personen‘ aus Abbildung 1 mit einer zweiten Tabelle ‚Autos‘ verbinden, um nachzuhalten, welche Person welches Auto fährt.

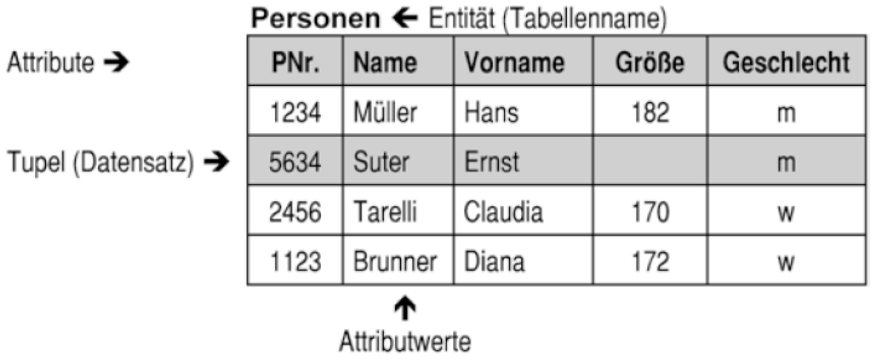

Abbildung 1: Aufbau einer Relation [Ste17, S. 15]

Als Grundlage für die Bearbeitung und Manipulation von Tabellen dient die ebenfalls von Codd vorgeschlagene relationale Algebra. Diese bietet aus der Mengenlehre abgeleitete Operationen an, um aus einer oder mehreren Tabellen eine Ergebnistabelle mit den gewünschten Daten zu berechnen [Stu16, S. 41]. Diese Operationen sind die Vereinigung, die Differenz, das kartesische Produkt, die Projektion und die Selektion von Daten [MK16, S. 104]. Die am häufigsten verwendete und vom American National Standard Institut (ANSI) als Standard für relationale Datenbanken erklärte Manipulations- und Abfragesprache ist die Structured Query Language [Ste17, S. 141]. SQL ist eine relational vollständige Sprache, das heißt sie kann alle Operatoren der Relationenalgebra darstellen [MK16, S. 104]. Eine Abfrage in SQL folgt einer festen Struktur [Mei16, S. 417]:

SELECT	Attribut/e der Ergebnistabelle
FROM	Tabelle/n, die betrachtet werden sollen
WHERE	Selektionsbedingung

Betrachtet man die Beispieltabelle in Abbildung 1 würde die Abfrage

SELECT	Name
FROM	Personen
WHERE	Größe = 182

eine Resultattabelle mit einer Spalte ‚Namen' und dem Datensatz ‚Müller' ausgeben. An der Abfragestruktur zeigt sich, dass es sich bei SQL um eine deskriptive Sprache handelt. Der Benutzer muss lediglich angeben, welche Daten ausgegeben werden sollen, nicht jedoch, durch welchen Aktionen das Datenbankmanagementsystem die entsprechenden Datensätze findet [MK16, S. 8]. Neben einem Abfrageteil, zu dem der vorgestellte Ausdruck gehört, bietet SQL einen Sprachenteil zur Datendefinition

(Data Definition Language), Datenmanipulation (Data Manipulation Language) und Datenschutz (Data Security Language) [Ste17, S. 6f.].

Eine wesentliche Forderung, die relationale Datenbanken von alternativen Ansätzen unterscheidet, ist die Gewährleistung der Datenkonsistenz [MK16, S. 187]. Nach Meier & Kaufmann (2016) sind Daten konsistent, wenn sie korrekt sind und zwischen ihnen keine Widersprüche bestehen [MK16, S. 56]. Diese Forderung ist notwendig, um schwerwiegende Fehler bei Berechnungen mit den Daten auszuschließen, kann jedoch bei einem hohen Datenaufkommen und Mehrbenutzerbetrieb nicht immer gewährleistet werden. Im relationalen Datenmodell wird Konsistenz erreicht, indem die Daten in Normalformen strukturiert werden [Ste17, S. 52]. Da hierzu meist die Aufteilung der Daten auf mehrere Tabellen notwendig ist, geht dies zulasten der Übersichtlichkeit und verlängert die Bearbeitungszeit von Abfragen. Um die Konsistenz auch bei Veränderungen der Daten durch Benutzer und insbesondere dem gleichzeitigen Zugriff mehrere Benutzer auf dieselben Daten sicherzustellen, werden sogenannte Transaktionen verwendet [MK16, S. 136]. Dabei handelt es sich um eine Folge von Datenbankanweisungen, welche nur vollständig ausgeführt werden dürfen [Stu16, S. 141]. Falls während der Ausführung ein Fehler auftritt, werden die schon bearbeiteten Anweisungen rückgängig gemacht und der ursprüngliche Zustand der Datenbank wiederhergestellt [Stu16, S. 141]. Diese Eigenschaft von Transaktionen wird als Atomarität bezeichnet [MK16, S. 136]. Darüber hinaus müssen Transaktionen drei weitere Eigenschaften aufweisen [MK16, S. 136]:

Konsistenz (Consistency)	Die Datenbank muss in einen konsistenten Zustand überführt werden.
Isolation	Parallel ablaufende Transaktionen von mehreren Benutzern müssen dieselben Ergebnisse liefern wie im Einbenutzerbetrieb.
Dauerhaftigkeit (Durability)	Datenbankzustände müssen so lange bestehen bleiben, bis sie von einer Transaktion verändert werden.

Diese vier Eigenschaften werden als ACID-Prinzip bezeichnet [MK16, S. 136].

Eine Datenbank auf Basis des relationalen Datenmodells kann einfach um neue Daten oder Beziehungen ergänzt werden, indem eine neue Tabelle erstellt wird. Daher weist dieses Konzept eine große Flexibilität auf und kann reale Systeme gut abbilden. Gleichzeitig wird es bei großen Datenmengen und komplexen Beziehungen schwer überschau- und berechenbar, da für eine Abfrage unter Umständen viele verschiedene Tabellen betrachtet werden müssen [Ste17, S. 10]. Relationale Datenbanken finden in den meisten kleineren und mittleren Betrieben Verwendung, stoßen

aber insbesondere bei massivem Datenaufkommen und vielen parallel zugreifenden Benutzern, wie etwa bei Web-Anwendungen, an ihre Grenzen [Mei18, S. 10].

2.2 Postrelationale Datenbanksysteme

Unter dem Begriff ‚postrelational' werden Datenbanken zusammengefasst, die nicht (ausschließlich) auf dem relationalen Datenmodell basieren [MK16, S. 188]. Bei postrelationalen Datenbankmanagementsystemen wird die Forderung nach stetiger Konsistenz und Redundanzfreiheit gelockert, um auch bei großen Mengen zu verarbeitender Daten und vielen parallel auf die Daten zugreifenden Nutzern eine hohe Ausfallsicherheit sowie Verfügbarkeit zu garantieren. Grundlage dieser Priorisierung ist das CAP-Theorem von Eric Brewer aus dem Jahr 2000 [Mei18, S. 33]. Darin stellt er fest, dass bei einem massiv verteilten Datenbanksystem nur zwei der drei Forderungen Konsistenz (Consistency), Verfügbarkeit (Availabilty) und Ausfallsicherheit (Partition Tolerance) gleichzeitig gewährleistet werden können [Mei18, S. 33]. Je nach Anwendungsfall werden unterschiedliche Kombinationen angestrebt. Viele Webdienste müssen beispielsweise dauerhaft verfügbar und gegen Ausfälle gesichert sein, dazu wird in Kauf genommen, dass die Datenbank zwischenzeitlich inkonsistente Zustände durchläuft. Bei den meisten postrelationalen Ansätzen wird dazu auf eine allzu strenge Strukturierung der Daten verzichtet. Es existiert eine Vielzahl an postrelationalen Datenmodellen (siehe etwa [MK16]), im Folgenden sollen die vier häufigsten, die sogenannten Core-NoSQL-Modelle, sowie das multidimensionale Datenmodell dargestellt werden [MK16, S. 222].

2.2.1 Schlüssel-Wert-Datenbank

Ein Schlüssel-Wert-Datenbank speichert binäre Relationen [HSS18, S. 667]. Unter einem Schlüssel (key) werden bestimmte Nutzdaten (value) abgelegt [HSS18, S. 667]. Eine typische Anwendung für eine Schlüssel-Wert-Datenbank ist der Einkaufswagen in einem Webshop. Als Schlüssel dient etwa eine personalisierte Kundennummer, unter der die Nutzdaten, in diesem Fall die einzelnen Produkte im Einkaufswagen des Kunden, abgelegt sind. Die Nutzdaten sind schemafrei. Es müssen daher keine Metadaten über die Struktur der Daten, wie dies bei Tabellen im relationalen Datenmodell der Fall ist, oder ihre Art definiert werden [MK16, S. 223]. Auch Referenzen zwischen einzelnen Datensätzen können nicht abgebildet werden [MK16, S. 223]. Durch diesen einfachen Aufbau sind Schlüssel-Wert-Datenbanksysteme in der Lage, große Datenmengen performant zu verarbeiten und die Speicherorte auf verschiedene Server zu verteilen [Wie15, S. 105]. Dieser Prozess wird Sharding genannt [Wie15, S. 105]. Die Unabhängigkeit der Daten untereinander ermöglicht den Einsatz paralleler Auswertungsverfahren, wie zum Beispiel Map/Reduce, bei denen

die partitionierten Daten von unterschiedlichen Rechnern gleichzeitig verarbeitet und die berechneten Ergebnisse anschließend zentral zusammengeführt und ausgegeben werden [HSS18, S. 668]. Durch die Verteilung von Teilaufgaben auf mehrere Rechner ist somit eine effiziente und schnelle Abfrage und Verarbeitung großer Datenmengen möglich.

2.2.2 Dokument-Datenbank

Bei Dokument-Datenbanken handelt es sich um eine Unterform der Schlüssel-Wert-Datenbanken. Unter einem eindeutigen Schlüssel kann ein Datensatz abgelegt werden. Im Gegensatz zu Schlüssel-Wert-Datenbanken ist dieser allerdings nicht von beliebiger Form, sondern ein strukturiertes Dokument [MK16, S. 229]. Die Struktur wird über ein Dateiformat definiert. In der Praxis kommt meist JSON (JavaScript Object Notation) zum Einsatz, seltener auch XML (eXtensible Markup Language) [HSS18, S. 671]. In Abbildung 2 ist der Aufbau einer Dokument-Datenbank im JSON-Format

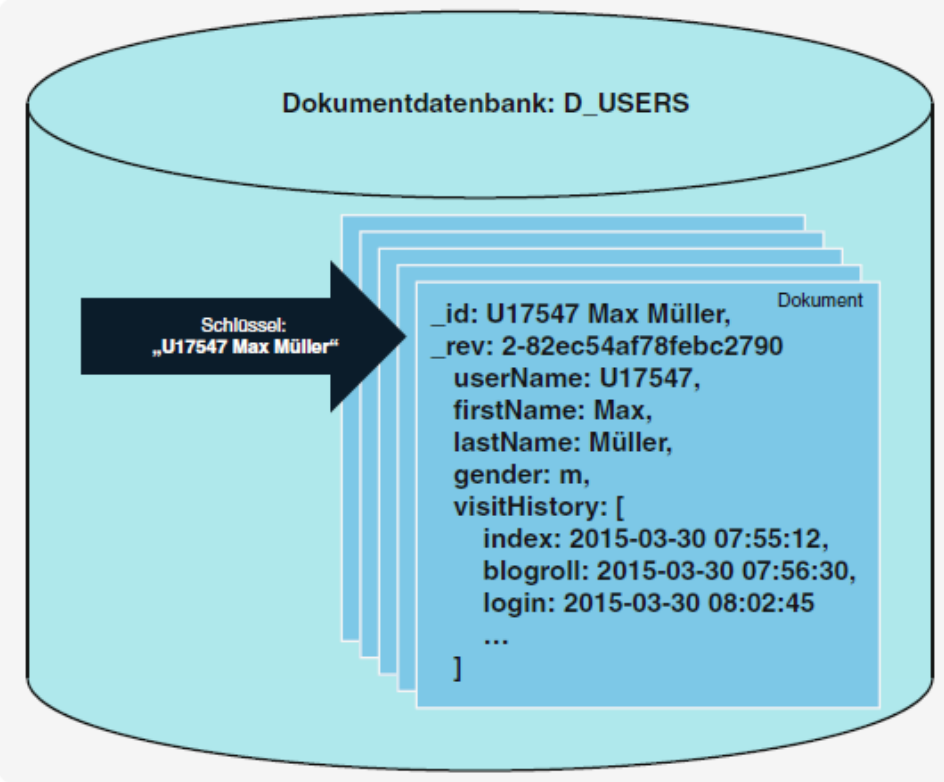

Abbildung 2: Dokument-Datenbank [MK16, S. 230]

abgebildet. Über den Schlüssel '_id' kann jedes Dokument eindeutig identifiziert und Benutzerdaten strukturiert gespeichert werden. Wie an dem Attribut 'visitHistory' zu sehen ist, können als Attributwert auch verschachtelte Schlüssel-Wert-Kombinationen abgelegt werden. Trotz ihrer Struktur sind Dokument-Datenbanken schemafrei, das

heißt es muss nicht definiert werden, welche Attribute in den Dokumenten verwendet werden dürfen [MK16, S. 230].

Dokument-Datenbanken eignen sich, wie auch Schlüssel-Wert-Datenbanken, für die Verarbeitung großer, heterogener Datenmengen und unterstützen Sharding sowie Replikationen [Mei16, S. 421]. Dadurch sind auch parallele Berechnungen wie das Map/Reduce-Verfahren möglich und sorgen für eine hohe Performanz und Effizienz dieser Datenbanken.

2.2.3 Spaltenfamilien-Datenbank

In Spaltenfamilien-Datenbanken werden Daten nicht zeilenweise, wie im relationalen Datenmodell, sondern spaltenweise abgespeichert [MK16, S. 226]. Hieraus resultieren einige Vorteile gegenüber relationalen Datenbanken: In der Regel sind bei einer Abfrage nur wenige Spalten von Interesse. Diese können in sogenannten Spaltenfamilien gemeinsam abgelegt und einzeln abgefragt werden, wodurch sich die Bearbeitungszeit einer Abfrage verringert [Wie15, S. 143]. Da alle Einträge einer Spalte aus demselben Wertebereich – der Domäne – stammen, können die Daten bei Wiederholungen verdichtet und somit Speicherplatz gespart werden [Wie15, S. 143 f.]. Außerdem sind spaltenweise Berechnungen einfacher durchführbar, da hierzu nur ein Datensatz abgefragt werden muss [Wie15, S. 144]. Gleichzeitig ist das Schema der Spaltenfamilien-Datenbank weniger streng definiert, als bei relationalen Alternativen. Dies macht Abbildung 3 deutlich, welche das Modell der BigTable-Datenbank von Google zeigt. Dargestellt ist eine Spaltenfamilie 'Contact', welche Kontaktinformationen von Personen enthält. Die Kontaktdaten einer Person werden über den sogenannten Zeilenschlüssel definiert, in Abbildung 3 lautet der Zeilenschlüssel für die Person 'Max Müller' 'U17547'. Eine einzelne Information oder auch Zelle adressiert man zusätzlich über den Spaltenschlüssel, wie etwa 'Contact:Mail' oder 'Contact:Name'. Es fällt auf, dass in einer Spaltenfamilie verschiedene Spaltenschlüssel auftreten können. Das einzige Schema, dass in Spaltenfamilien-Datenbanken definiert wird, sind die Spaltenfamilien selbst. Innerhalb der Spaltenfamilien können im Gegensatz zu relationalen Tabellen jedoch beliebige Spaltenschlüssel verwendet werden [MK16, S. 227]. Außerdem sind die Datensätze im BigTable-Modell mit Zeitstempeln versioniert, sodass sich eine dreidimensionale Struktur ergibt und Änderungen der Daten nachvollzogen werden können. Spaltenfamilien-Datenbanken stellen einen Kompromiss dar zwischen logischer Struktur und Zusammenfassung ähnlicher Daten in Spaltenfamilien bei gleichzeitiger Flexibilität der Daten innerhalb dieser Struktur.

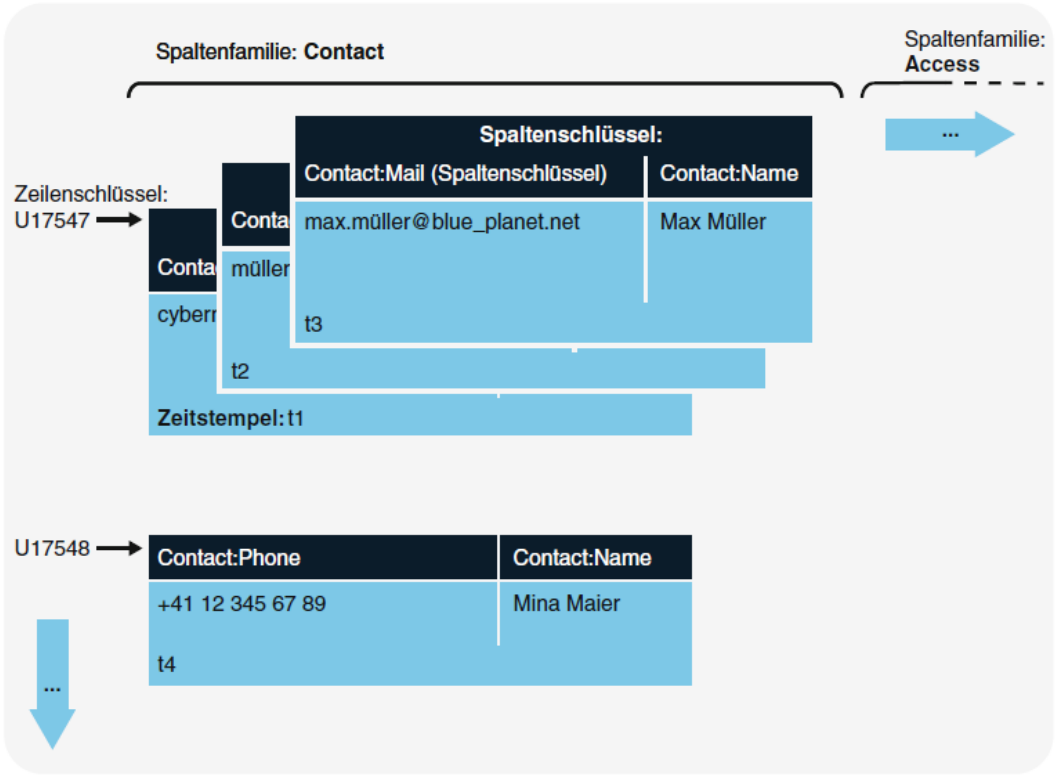

Abbildung 3: Google-Datenbank Big Table [MK16, S. 228]

2.2.4 Graph-Datenbank

Die Stärke von Graph-Datenbanken liegt in der Abbildung von Beziehungen zwischen Daten [HSS18, S. 689]. Sie basieren auf einem intuitiven, visuellen Ansatz, welcher in Abbildung 4 dargestellt ist. Eine Graph-Datenbank besteht aus zwei wesentlichen Komponenten: Knoten, die Entitäten darstellen, sowie Kanten, welche die Beziehung zwischen den Knoten beschreiben. Sowohl die Knoten, als auch die Kanten können Informationen speichern, häufig, wie in Abbildung 4, in Form von Schlüssel-Wert-Kombinationen [Wie15, S. 41]. Durch diesen Aufbau eigenen sich Graph-Datenbanken sehr gut zur Beschreibung von sozialen Medien, Infrastruktur- oder Kommunikationsnetzen. Eine typische Fragestellung ist zum Beispiel die Ermittlung des kürzesten Weges zwischen zwei Knoten oder die Prüfung der Existenz eines sogenannten Eulerkreises, in dem jede Kante genau einmal enthalten ist [Wie15, S. 45].

Graph-Datenbanken unterscheiden sich von relationalen Datenbanken wesentlich durch ihre Eigenschaft der indexfreien Nachbarschaft: Zu jedem Knoten, kann das Datenbankmanagementsystem alle Nachbarn finden, ohne sämtliche existierenden Kanten zu prüfen [MK16, S. 238]. Dabei nutzt es sogenannte Adjazenzlisten aus der Graphentheorie, in der zu jedem Knoten die damit verbundenen Kanten gespeichert

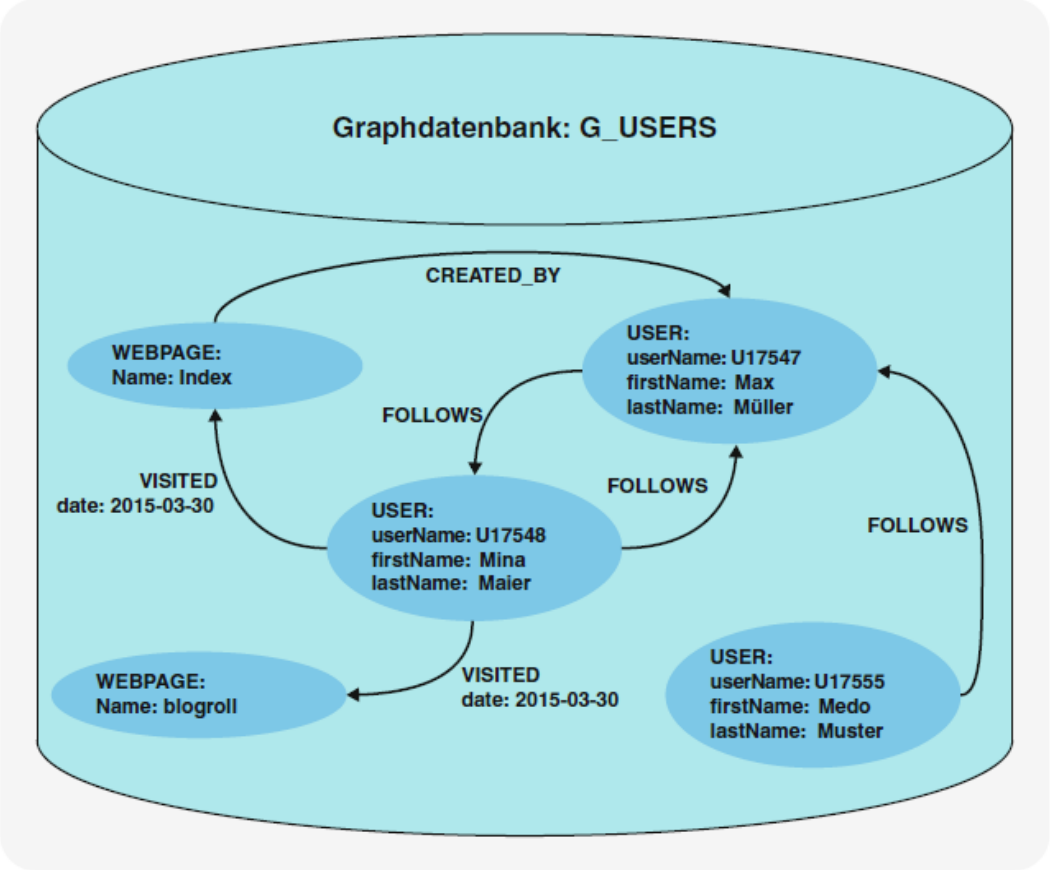

Abbildung 4: Graph-Datenbank [MK16, S. 237]

sind [HSS18, S. 690 f.]. Hierdurch ist der Aufwand für eine Abfrage unabhängig von der Größe und Komplexität der Datenbank immer gleich groß, während bei relationalen Datenbanken der Aufwand mit zunehmender Zahl von Datensätzen steigt [MK16, S. 238]. Dies macht den Einsatz von Graph-Datenbanken bei massivem Datenaufkommen und netzartigen Strukturen sehr effizient. Nachteilig wirkt sich allerdings die Schwierigkeit aus, den Graphen zu partitionieren. Durch die vielen Beziehungen zwischen den Daten gibt es keine effiziente Methode, den Graphen in Teilgraphen zu unterteilen und auf unterschiedlichen Rechnern abzulegen [MK16, S. 239]. Aus diesem Grund unterstützen heutige Graph-Datenbanken kein Sharding [MK16, S. 239].

2.2.5 Multidimensionale Datenbank

Multidimensionale Datenbanken werden gemäß des Online Analytical Processing (OLAP) genutzt, um die Datenanalyse und Entscheidungsfindung zu unterstützen [MK16, S. 196]. Hierzu werden Daten nach beliebigen Entscheidungsdimensionen, wie zum Beispiel Zeit, Produkt und Ort, abgelegt. Das Ergebnis ist ein mehrdimensionaler

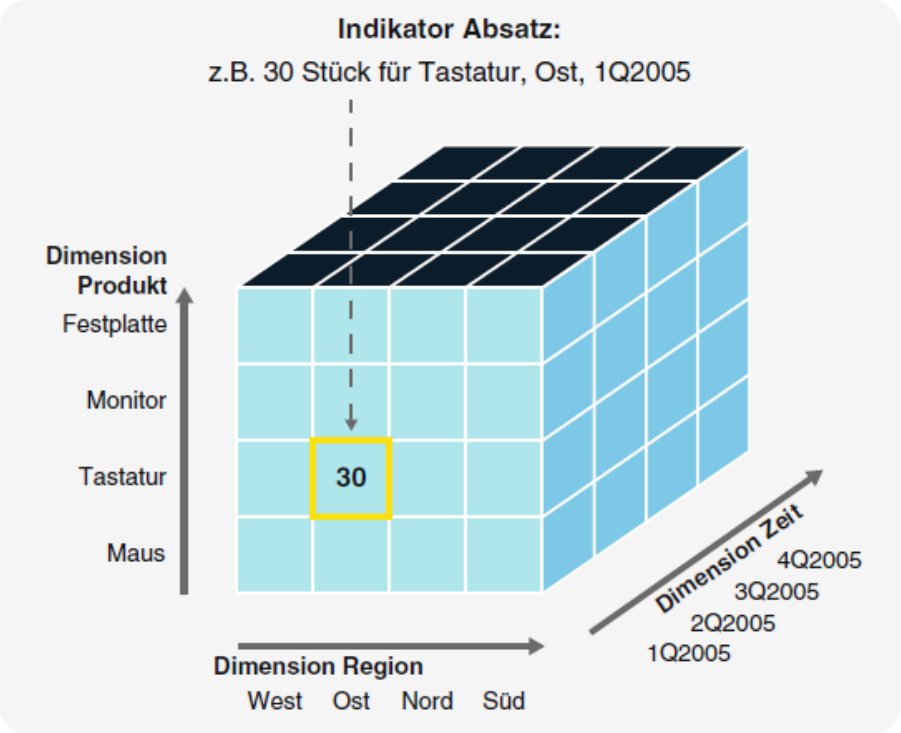

Abbildung 5: Mehrdimensionaler Datenwürfel [MK16, S. 196]

Datenwürfel, wie in Abbildung 5 dargestellt. Auf Grundlage des Datenwürfels lassen sich entscheidungsrelevante Kennwerte, in Abbildung 5 sind dies die Absatzzahlen, hinsichtlich verschiedener Dimensionen auswerten. Diese Kennwerte werden Indikatoren genannt [MK16, S. 197]. Meier & Kaufmann (2016) nennen drei Kernfunktionen von mehrdimensionalen Datenbankmanagementsystemen [MK16, S. 199]. Für die Dimensionsachsen müssen sich beliebige Aggregationsstufen festlegen lassen, das heißt die Einteilung und Gruppierung der Daten entlang der Achsen muss in beliebigen Intervallen möglich sein. Dies umfasst auch eine Strukturierung der Achsen in Ober- und Untergruppen. Innerhalb des Intervalls 'Quartal' der Zeit-Dimension muss es möglich sein, auch Monate, Wochen oder Tage modellieren zu können. Darüber hinaus muss die Auswertungssprache das sogenannte Drill-Down (Erhöhung des Detaillierungsgrades) und Roll-Up (Hinzunahme weiterer Aggregationsstufen) unterstützen. Dies ist bei SQL nicht der Fall [MK16, S. 199]. Die dritte Funktion ist die Auswahl einer einzelnen Datenscheibe (Slicing) sowie der Wechsel der Dimensionsreihenfolge.

3 Schnittstellen von Data Mining-Anwendungen

3.1 Data Mining im Kontext von Big Data

In der Einleitung wurde bereits ausgeführt, dass die digitale Transformation mit der massenhaften und stetig steigenden Generierung und Speicherung von Daten einhergeht. Beispielhaft sei an dieser Stelle die Erhebung von Prozessdaten in der industriellen Fertigung, wie etwa Signale von Sensoren und Aktuatoren und Regelungen und Steuerungen oder die Analyse von Kundendaten im Marketing, genannt [Run10, S. 1]. Diese umfangreichen Datenbestände werden als Big Data bezeichnet [Mei18, S. 5]. Es existiert keine präzise Definition für den Big Data-Begriff, allerdings werden für dessen Charakterisierung von viele Autoren drei V's angeführt [MK16, S. 416]:

Volume	Die Datenmenge ist sehr groß und liegt im Tera- bis Zettabytebereich.
Variety	Die Daten sind sehr vielfältig und nicht von einheitlicher Struktur.
Velocity	Die Daten werden in Echtzeit ausgewertet und analysiert.

Einige Experten führen ausgehend von der Intention, mit der die Daten erfasst werden, sowie ihrer Qualität, noch zwei V's hinzu [Mei18, S. 6]:

Value	Es werden Daten generiert, die den Unternehmenswert steigern sollen.
Veracity	Bei der Auswertung muss die meist unterschiedliche Qualität der Datenbestände berücksichtigt werden.

Aufgabe des Data Minings ist es nach Runkler (2010) „Wissen aus Daten zu extrahieren" [Run10, S. 2]. Ein Datum ist in diesem Kontext definiert als eine „Ansammlung von Zeichen mit der dazugehörigen Syntax" [CL16, S. 37]. Ein Beispiel für ein Datum ist etwa eine Datenzelle in einer relationalen Datenbank. Ist dieses Datum mit einer Bedeutung gekoppelt, spricht man von einer Information [CL16, S. 38]. Informationen sind also interpretierbare Daten. Cleve & Lämmel (2016) führen weiter aus, dass eine Information dann zu Wissen wird, wenn der Anwender die Fähigkeit besitzt, die Information zu benutzen [CL16, S. 38]. Es lässt sich zusammenfassen, dass Data Mining dazu dient, dem Anwender aus einer Datenmenge interpretierbare Aussagen abzuleiten, aus denen er eine Reaktion, resp. Handlung ableiten kann. Peterson (2005) ergänzt die Anforderungen an den Data Mining-Prozess noch insofern, als dass das Wissen neu, statistisch sicher und für den Anwender nützlich sein soll [Pet09, S. 9]. Einige Autoren bezeichnen den gesamten Prozess der Datenverarbeitung und Wissensentdeckung sowie Auswertung als Data Mining [Pet05, S. 10]. Fayyad dagegen definiert das Data Mining als Teilschritt eines übergeordnetem Prozesses, dem Knowledge Discovery in Databases (KDD) [Pet09, S. 9]. Das KDD-Modell nach Fayyad ist in Abbildung 6 dargestellt. Neben dem eigentlichen Data Mining umfasst

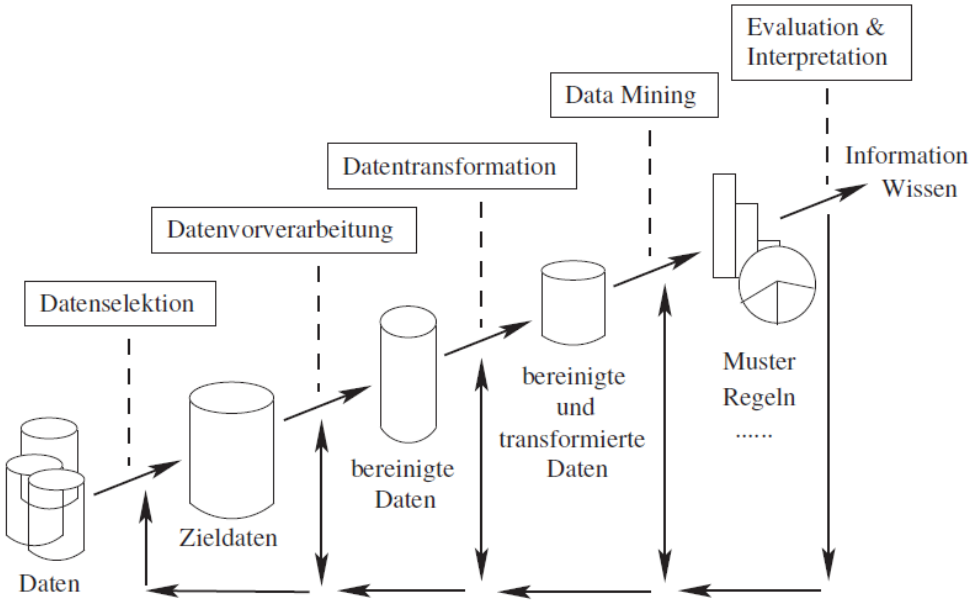

Abbildung 6: Knowledge Discovery in Databases [CL16, S. 6]

es noch die folgenden Schritte [CL16, S. 5f.]:

Selektion	Auswahl und Export der für die Analyse benötigten Daten.
Vorverarbeitung	Bereinigung und Korrektur fehlender und widersprüchlicher Daten.
Transformation	Umwandlung der Daten in für die Analyse geeignete Formate (z.B. Gruppierung in Intervalle).
Data Mining	Suche nach Mustern und Entwicklung eines Modells.
Evaluation	Interpretation und Auswertung der Ergebnisse.

In Data Mining-Anwendungen ist meist der gesamte KDD-Prozess implementiert. Die Datenanalyse findet dabei in der Regel automatisiert statt, bei der Datenselektion und - vorbereitung ist Unterstützung durch den Anwender notwendig [CL16, S. 39].

3.2 Programmierschnittstellen

Der Zugriff auf eine Datenbank von einem Anwendungsprogramm heraus geschieht normalerweise nicht direkt, sondern mithilfe einer Zwischenschicht. Diese von Geisler (2014) Middleware genannte Schicht koppelt beide Systeme miteinander [Gei14, S. 77]. Durch die Verwendung von Middleware muss sich der Programmierer nicht mit den Implementierungsdetails spezifischer Datenbanken auseinandersetzen, sondern kann mit der Wahl einer datenbankunabhängigen Programmierschnittstelle den Zugriff auf eine Vielzahl von Datenbanken realisieren [Gei14, S. 78]. Abbildung 7 zeigt das

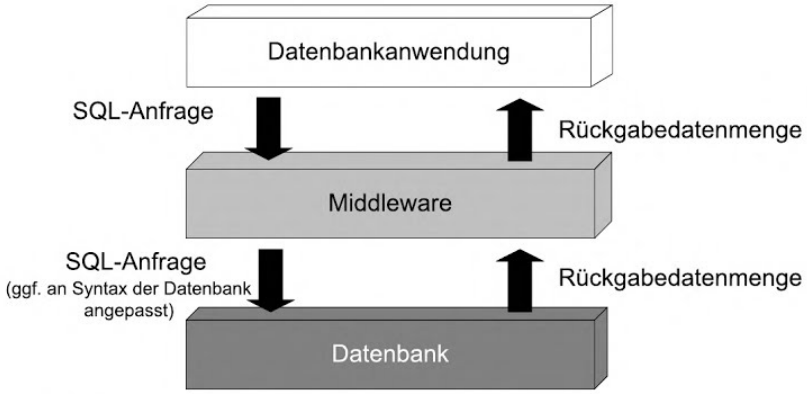

Abbildung 7: Funktionsprinzip von Middleware [Gei14, S. 77]

Funktionsprinzip von einer auf SQL basierenden Middleware. Für den Zugriff der Middleware auf die Datenbank benötigt diese datenbankspezifische Treiber, welche von den Datenbankherstellern zur Verfügung gestellt werden [Gei14, S. 78]. Im Folgenden sollen die meist verwendeten Programmierschnittstellen dargestellt werden.

3.2.1 Open Database Connectivity (ODBC)

Die Open Database Connectivity wurde 1992 von der SQL Access Group und Microsoft entwickelt [Her02, S. 216]. Sie nutzt für den Zugriff auf Datenbanken eine standardisierte Version von SQL und kann daher in Verbindung mit relationalen Datenbanken genutzt werden [Gei14, S. 78]. Die ODBC-Architektur, dargestellt in Abbildung 8, besteht aus vier Komponenten:

1. Die Datenbankanwendung ist das Data Mining-Programm, das der Benutzer bedient. Es ruft nach einer Benutzereingabe eine ODBC-Funktion auf, um eine SQL-Anweisung an die Datenbank abzusetzen [Her02, S. 218].

2. Der Treiber-Manager lädt und entlädt den notwendigen Treiber und leitet die ODBC-Funktion an diesen weiter [Gei14, S. 78].

3. Der Treiber führt die ODBC-Funktion aus, übermittelt die SQL-Anfrage an die Datenbank und liefert das Ergebnis zurück [Gei14, S. 79].

4. Die Datenbank enthält die gewünschten Daten.

Laut Herbolsheimer (2002) entstehen bei optimal eingerichteten ODBC-Treibern kaum Geschwindigkeitsverluste im Vergleich zu direktem Zugriff auf die Datenbank [Her02, S. 217]. Im Gegensatz dazu nennt Schwichtenberg (2010) neben eingeschränkter Flexibilität durch die Beschränkung auf relationale Datenbanken auch

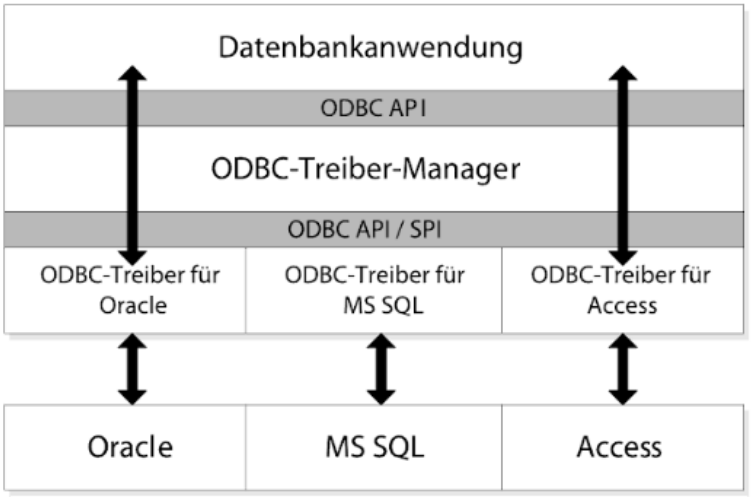

Abbildung 8: ODBC-Architektur [Gei14, S. 79]

die Geschwindigkeit der ODBC-Schnittstelle als Grund für dessen Ablösung durch OLE-DB [Sch10, S. 646].

3.2.2 Object Linking and Embedding - Database (OLE-DB)

Das von Microsoft entwickelte OLE-DB basiert auf dem Component Object Model (COM) [Her02, S. 218]. Herbolsheimer (2002) beschreibt COM als eine „Ansammlung von Spezifikationen, Datenstrukturen und Schnittstellen“ [Her02, S. 218]. Über verschiedene Schnittstellen und Methoden kann auf Objekte unterschiedlicher Herkunft zugegriffen werden. Da die Komponenten untereinander auf binärer Ebene kommunizieren, ist COM unabhängig von einer spezifischen Programmiersprache [Her02, S. 218]. OLE-DB definiert eine Reihe von COM-Schnittstellen, über die Softwarekomponenten auf eine Datenbank zugreifen können [Her02, S. 218]. Die OLE-DB-Architektur unterscheidet zwischen Datenanbietern (Data Provider), welche auf Datenquellen zugreifen können, Dienstanbietern (Service Provider), die Daten verarbeiten und weiterleiten und Konsumenten (Consumer), das heißt Anwendungen, welche OLE-DB nutzen [Sch10, S. 648]. Die Data Provider ersetzen somit die Treiber der ODBC-Schittstelle. Als Weiterentwicklung von ODBC ist OLE-DB abwärtskompatibel und unterstützt über eine sogenannte 'OLE-DB-ODBC-Bridge' alle existierenden ODBC-Treiber [Sch10, S. 648]. Im Allgemeinen ist der Zugriff über die OLE-DB-ODBC-Bridge jedoch langsamer, als der direkte Zugriff über einen OLE-DB-Provider, sodass dieser - falls vorhanden - zu bevorzugen ist [Gei14, S. 80].

3.2.3 Java Database Connectivity (JDBC)

Die Java Database Connectivity (JDBC) ist Teil der Standard-API (Application Programming Interface) von Java [SSH18, S. 413]. Wie auch ODBC basiert sie auf SQL und ermöglicht somit den Zugriff auf relationale Datenbanken [Sal16, S. 433]. Durch die objektorientierte Struktur von Java zeichnet sich JDBC nach Saake et al. (2018) gegenüber ODBC durch eine bessere Übersichtlichkeit und leichtere Bedienbarkeit aus, da einzelne Mechanismen, wie die Verbindungsherstellung, SQL-Anweisungen oder das Anfrageergebnis in eigenen Klassen unterteilt und typisiert sind [SSH18, S. 314]. Die wichtigsten Klassen sind hierbei [SSH18, S. 314]:

java.sql.DriverManager	Laden des Datenbanktreibers und Aufbau einer Verbindung.
java.sql.Connection	Repräsentation der Datenbankverbindung.
java.sql.Statement	Ausführung von SQL-Anweisungen.
java.sql.ResultSet	Verwaltung der Anfrageergebnisse und Zugriff auf einzelne Spalten.

Bei den zur Herstellung der Verbindung benötigten Treibern unterscheidet JDBC zwischen vier Typen [Sal16, S. 434f.]:

Treibertyp I	Dieser Treibertyp wird auch 'JDBC-ODBC-Bridge' genannt und verwendet für den Datenbankzugriff die schon vorgestellte ODBC-Schnittstelle. Der Umweg über eine zusätzliche Schnittstelle wirkt sich negativ auf die Effizienz aus und beschränkt JDBC auf die Möglichkeiten von ODBC.
Treibertyp II	Wie Treibertyp I ist auch Typ II nicht direkt an die Datenbank angekoppelt. Stattdessen erfolgt die Kommunikation mit der Datenbank über native Binärdaten.
Treibertyp III	Dieser Treibertyp stellt die Verbindung zur Datenbank über eine Middleware und Netzwerk-Sockets her. Auch hierbei erfolgt kein direkter Austausch mit der Datenbank.
Treibertyp IV	Der Treibertyp IV wird auch Pure Java Driver genannt und besteht aus reinem Java-Code. Im Gegensatz zu den Treibertypen I-III stellt er über die Netzwerkschnittstelle des Datenbankmanagementsystems einen direkten Kontakt zur Datenbank her.

3.2.4 ActiveX Data Objects (ADO)

Aufgrund der Komplexität und Systemnähe der OLE-DB-Schnittstelle hat Microsoft mit den 'ActiveX Data Objects' (ADO) eine vereinfachte, auf OLE-DB basierende

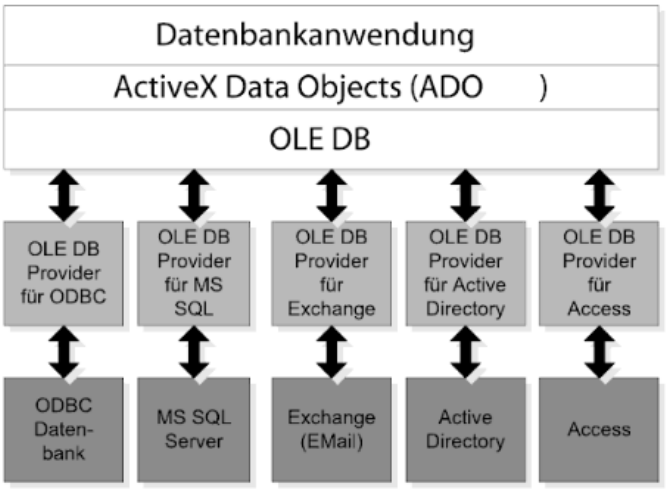

Abbildung 9: ADO-Architektur [Gei14, S. 80]

und objektorientierte Datenschnittstelle geschaffen [Sch10, S. 648f.]. ADO ist somit nicht als eigenständige Programmierschnittstelle zu sehen, sondern vereinfacht lediglich die Benutzung der OLE-DB-Schnittstelle. In Abbildung 9 ist die ADO-Architektur dargestellt. Wie in JDBC werden Interaktionen mit der Datenbank in verschiedenen Klassen definiert [Sch10, S. 656]. Bei der Nomenklatur der Klassen zeigt sich die Nähe zu JDBC noch deutlicher: In einem gleichnamigen 'Connection'-Objekt werden Verbindungsinformationen gespeichert, das Objekt 'Command' beinhaltet die Beschreibung der Abfrage (in JDBC das 'Statement'-Objekt) und 'RecordSet' enthält die Ergebnismenge der Abfrage ('ResultSet' in JDBC) [Her02, S. 228].

3.2.5 ADO.NET

Bei ADO.NET handelt es sich um Microsofts Weiterentwicklung der 'ActiveX Data Objects' [Gei14, S. 81]. Hierbei wurde die Grundannahme von ADO angepasst, dass durchgehend eine Verbindung zwischen Client und Server besteht, ein Anwendungsprogramm also stets auf die Datenbank zugreifen kann [Gei14, S. 81]. ADO.NET hingegen ist für Intra- und Internetanwendungen konzipiert, bei denen keine dauerhafte Verbindung gegeben sein muss, sondern Daten stattdessen lokal zwischengespeichert, bearbeitet und zu einem späteren Zeitpunkt wieder mit der Datenbank synchronisiert werden [Gei14, S. 81]. Dies spiegelt sich in der ADO.NET-Architektur wieder, siehe Abbildung 10. Wie auch bei ADO, resp. OLE-DB, wird bei ADO.NET zwischen Datenprovidern und Datenkonsumenten unterschieden [Dob+18, S. 702]. Die Datenprovider stehen stets mit der Datenbank in Verbindung, ihre Objekte werden daher als 'verbundene Objekte' bezeichnet. Die Datenkonsumenten hingegen bestehen unabhängig von der Datenbank und bilden somit 'unverbunde-

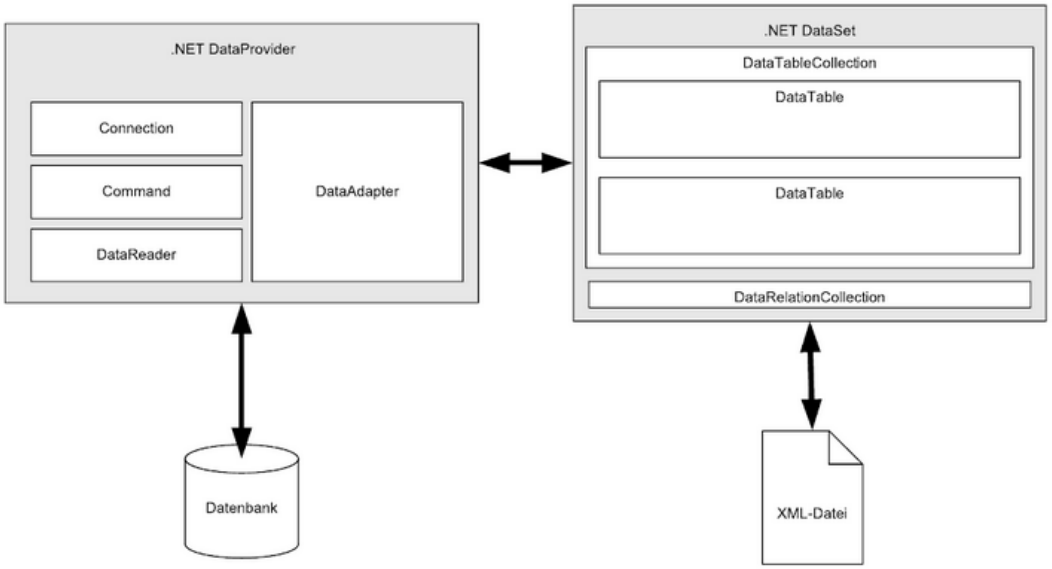

Abbildung 10: ADO.NET-Architektur [Gei14, S. 82]

ne Objekte' [MNK14, S.974f.]. Das Kernobjekt von ADO.NET ist das 'DataSet'. Es ist ein Datenkonsument und vergleichbar mit dem 'RecordSet'-Objekt in ADO [Dob+18, S. 702]. Es kann jedoch lokal bearbeitet und sogar erstellt werden und enthält sämtliche Klassen, die für die Arbeit mit dem 'DataSet' auf dem Client notwendig sind [Gei14, S. 82]. Die Schnittstelle zwischen dem 'DataSet' und dem Datenprovider ist die Klasse 'DataAdapter', welche somit als Verbindungsglied zwischen den unverbundenen 'DataSet'-Objekten und den Datenprovidern fungiert [Dob+18, S. 702].

3.3 Untersuchung von Data Mining-Anwendungen

Im Folgenden werden einige Data Mining-Anwendungen hinsichtlich ihrer Schnittstellen zu Datenbanksystemen untersucht. Aufgrund des großen Marktes von Data Mining-Anwendungen und des begrenzten Umfanges dieser Projektarbeit musste vorab eine Auswahl von Programmen vorgenommen werden. Diese gestaltete sich wie folgt:

Im Februar 2018 veröffentlichte das amerikanische Marktforschungsunternehmen Gartner Inc. den 'Magic Quadrant for Data Science and Machine-Learning Platforms' [Gar]. Hierbei bewertet Gartner 16 IT-Anbieter von Data Science-Anwendungen. Dies sind Alteryx, Anaconda, Angoss, Databricks, Dataiku, Domino, H20.ai, IBM, KNIME, Mathworks, Microsoft, Rapidminer, SAP, SAS, Teradata und TIBCO Software [Gar]. Es stellt sich an dieser Stelle die Frage, ob die Betrachtung von Data Science- und Machine Learning-Programmen angesichts des Fokus dieser Arbeit auf

das Data Mining zulässig ist. Hierbei hilft eine genaue Betrachtung der jeweiligen Begriffe. Said & Torra (2019) definieren Data Science als die Ableitung von Handlungen und Vorhersagen auf der Grundlage von aus Daten extrahiertem Wissen [ST19, S. 1]. Dies ist weitgehend synonym mit der in Kapitel 3.1 dargestellten Data Mining-Definition von Runkler. Maschinelles Lernen (Machine Learning) fasst nach Frochte (2018) Techniken zusammen, mit derer Hilfe Computer Verhalten aus Daten erlernen [Fro18, S. 13]. Diese Techniken werden unter anderem für das Data Mining genutzt [Fro18, S. 16; Pet09, S. 19]. Auf dieser Grundlage scheint es schlüssig, die von Gartner untersuchten Programme in die Schnittstellenbetrachtung mitaufzunehmen. Ergänzt wird die Auswahl noch durch eine Studie des Frauenhofer Instituts für Produktionstechnik und Automatisierung IPA aus dem Jahr 2014 über den Einsatz und die Nutzenpotentiale von Data Mining in Produktionsunternehmen [Wes+14]. Neben den von Gartner schon aufgeführten Anbietern der Programme RapidMiner, SPSS von IBM, Statistica von Statsoft und SAP BI wurde dabei auch Oracle Data Mining in den Unternehmen genutzt und soll daher in dieser Projektarbeit betrachtet werden [Wes+14, S.19]. Abschließend wurde zudem eine Studie der mayato GmbH betrachtet, welche im Jahr 2009 den Data Mining-Markt auf einen aus zwölf Programmen bestehenden Querschnitt reduzierte und diesen einem Funktionsvergleich unterzog [Dil09, S. 3]. Dieser Querschnitt enthielt neben einigen schon genannten Programmen noch die folgenden Anwendungen [Dil09, S. 4]:

- Universität Waikato: Weka
- KXEN Analytic Framework
- Viscovery SOMine
- prudsys Discoverer / Basket Analyzer
- Bissantz Delta Master

Insgesamt ergibt sich somit eine Auswahl von 22 Anwendungen. Die Schnittstellen dieser Programme zu Datenbanksystemen werden im Folgenden betrachtet. Viele Data Mining-Programme bieten neben den schon vorgestellten Schnittstellen noch einige Importfunktionen für besondere oder seltene Dateiformate und Schnittstellen zu bestimmten Anwendungen oder Diensten. Die ausführliche Behandlung dieser Schnittstellen ist im Rahmen dieser Projektarbeit nicht möglich, die jeweiligen Funktionen sollen im Folgenden aber unter dem Reiter 'Sonstige' mit aufgeführt werden.

3.3.1 Alteryx

Alteryx bietet eine aus mehreren Softwarepaketen zusammengesetzte Plattform für die Datenanalyse an [Altb]. Im Alteryx Designer sind mehr als 40 Data Mining-

Funktionen implementiert [Altd].

Alteryx listet alle unterstützten Datenquellen in der Alteryx Dokumentation [Alte]. Dabei werden folgende Quellen genannt [Alte]: Amazon Aurora (Verbindung über ODBC), Amazon Redshift (ODBC),Amazon S3 (Verbindung über ein von Alteryx implementiertes Tool, vergleiche [Alta]), Apache Cassandra (ODBC), DataStax (ODBC), dBase, ESRI GeoDatabase, Exasol (ODBC), HP Vertica (ODBC), IBM DB2 (ODBC oder OLE-DB), Microsoft Access (.mdb, .accdb), Microsoft Azure Data Lake Store (Verbindung über Alteryx Tool), Microsoft Azure SQL Database (ODBC, OLE-DB), MongoDB (Alteryx Tool), MySQL (ODBC), Oracle (ODBC, OLE-DB, OCI), Pivotal Greenplum (ODBC), PostgreSQL (ODBC), SAP HANA (ODBC) und Snowflake (ODBC). Über XML- und JSON-Dokumente kann auch auf Dokument-Datenbanken zugegriffen werden [Alte].

Sonstige: Alteryx besitzt außerdem Schnittstellen zu den folgenden Anwendungen und Dateiformaten [Alte]: Adobe Analytics, Amazon Athena, ASCII (.flat, .asc), Apache Hadoop Avro (.avro), Hadoop Distributed File System (HDFS), Apache Spark, Apache Hive, Autodesk, Textdateien (.csv, txt), Databricks, GIS, Google Analytics, Google BigQuery, Google Earth/Maps, Google Sheets, GZip-Dateien (.tar, .gz), HTML, MapInfo, MapR, Marketo, Microsoft Analytics Platform System, Microsoft Azure ML, Microsoft SQL Data Warehouse, Microsoft Cognitive Services, Microsoft Dynamics CRM, Microsoft Excel (.xls, .xlsx, .xlsb, .xlsm), Microsoft OneDrive, Microsoft Power BI, Microsoft SharePoint, Microsoft SQL Server, Netsuite Analytics, OpenGIS (.gml), Qlik (.qvx), Salesforce.com, SAS (.sas7bdat), SQLite (.sqlite), SRC Geography (.geo), Tableau (.tde, .hyper), ZIP-Dateien (.zip).

3.3.2 Anaconda

Bei Anaconda handelt es sich um eine Open-Source Data Science Distribution, welche mehr als 1400 Pakete basierend auf den Programmiersprachen Python und R unterstützt [Anah]. Durch die offene Architektur und die direkte Programmierung der gewünschten Analyseaufgabe samt Schnittstelle in Python oder R ist Anaconda sehr vielseitig und flexibel. Es sind keine vordefinierten Schnittstellen vorhanden, wie dies bei Data Mining-Programmen der Fall ist, in denen der Anwender lediglich über die Bedienungsoberfläche mit dem Programm kommunizieren kann. Über die Anaconda Cloud können Pakete und Bibliotheken gesucht und mit anderen Nutzern geteilt werden [Anaac]. Um im Rahmen dieser Arbeit einen Überblick über die Möglichkeiten der Anaconda-Distribution zu geben, wird im Folgenden für alle in diesem Kapitel genannten Datenbanksysteme (einschließlich der in den noch folgenden Unterkapitel genannten) die Existenz von Paketen in der Anaconda Cloud

untersucht. Diese sind vorhanden für die Datenquellen Amazon Redshift [Anag], Amazon S3 [Anaj], Cassandra [Anaf], Ceph [Anam], Couchbase [Anaaa], Elastic Search [Anao], Esri GeoDatabase [Anao], Exasol [Anax], Google BigQuery [Anap], Google Cloud Storage [Anaq], HBase [Anal], HP Vertica [Anay], MariaDB [Anau], Microsoft Azure Blob Storage [Anan], Microsoft SQL Server [Anaz], Minio [Anai], MongoDB [Anab], MySQL [Anat], Neo4j [Anar], Oracle [Anac], Pivotal Greenplum [Anak], PostgreSQL [Anas], Redis [Anav], Snowflake [Anaw], SQLite [Anaab], Sybase [Anad] und Teradata [Anae]. Es ist zu betonen, dass diese Liste nicht vollständig ist. Eine Befragung aus dem '2018 Anaconda State of Data Science Report' über die verwendeten Datenquellen der Anaconda-Nutzer zeigt, wie vielseitig Anaconda genutzt wird [Anaa, S. 4]. Unter den Anwendungsfällen finden sich sowohl SQL-, als auch NoSQL-Datenbanken, Cloud-Dienste oder Apache Hadoop und Spark.

3.3.3 KnowledgeSEEKER (Datawatch)

Untersucht wurde die Datenmanagement-Plattform Datawatch Angoss KnowledgeSEEKER der Firma Datawatch Corporation. Diese wirbt insbesondere mit der benutzerfreundlichen Erstellung von Entscheidungsbäumen [Datj].

Die Software KnowledgeSEEKER verwendet laut der Produktbroschüre des Programms die ODBC-Schnittstelle [Datj]. Darüber hinaus können Excel-, und Textdateien (.csv) und XML-Dokumente eingelesen werden [Datj]. Auch eine Verbindung zu dem Programm SPSS von IBM und SAS ist möglich [Datj].

3.3.4 Databricks Unified Analytics Platform

Untersucht wurde die Software Databricks Unified Analytics Platform. Diese beinhaltet Funktionen aus den Bereichen Maschinelles Lernen, Neuronale Netze und der Analyse von Graphen [Date; Datc; Datd].

Databricks beschreibt die unterstützen Datenquellen in der Online-Dokumentation: Amazon Redshift, Amazon S3, Azure Blob Storage, Azura Data Lake Storage, Azure Cosmos DB, Azure SQL Data Warehouse, Cassandra, Couchbase, ElasticSearch, MongoDB, Neo4j, Oracle, Redis und Snowflake [Datb].

JDBC: Darüber hinaus kann unter Verwendung von JDBC eine Verbindung zu einer relationalen Datenbank eingerichtet werden [Data]. Die JDBC-Treiber für MySQL, Microsoft SQL-Server und der Azure SQL Database sind in der Databricks Runtime ab Version 3.4 enthalten [Data].

Sonstige: Zusätzlich unterstützt Databricks noch die folgenden Datenquellen und Dateiformate: Bilder, Avro-Dateien, Textdateien (.csv), JSON-Dokumente, Parquet-Dateien, LZO komprimierte Dateien, Zeitreihen und Zip-Dateien [Datb].

3.3.5 Dataiku

Dataiku DSS ist eine Analyseplattform mit Schwerpunkt auf Maschinellem Lernen, in der frei zugängliche Programmbibliotheken wie Scikit-Learn, MLlib und XGBoost oder selbstständig in Python oder R implementierte Modelle mit einer grafischen Bedienoberfläche von Dataiku kombiniert werden [Datg]. Ein Anwendungsgebiet von Dataiku DSS ist zum Beispiel die Cluster-Analyse von Textdateien [Gre14]. Die Datenbankschnittstellen spezifiert Dataiku in dem Produktdatenblatt [Datf, S. 4]:

SQL-Datenbanken: MySQL, PostgreSQL, Vertica, Amazon Redshift, Pivotal Greenplum, Teradata, IBM Netezza, SAP HANA, Oracle, Microsoft SQL Server, Google BigQuery, IBM DB2, Exasol, MemSQL und Snowflake. Darüber hinaus kann eine neue Verbindung über JDBC eingerichtet werden.

NoSQL-Datenbanken: MongoDB, Cassandra und ElasticSearch.

Zudem kann auf die Cloud-basierten Datenspeicher Amazon S3, Google Cloud Storage, Azure Blob Storage und Azure Data Lake Store zugegriffen werden.

Sonstige: Cloudera, Hortonworks, MapR, AmazonEMR, Textdateien (.csv), Parquet-Dateien, ORC-Dateien, SequenceFiles, RCFiles, FTP, SCP, SFTP, HTTP.

3.3.6 Domino Data Sience Platform

Untersucht wurde die Software Domino Data Science Platform. Wie die Anaconda Distribution setzt diese auf eine offene Infrastruktur, in der frei verfügbare Data Sience-Werkzeuge und Datenkonnektoren in der Domino-Software eingebunden und zusammengeführt werden [Domb]. In Abbildung 11 ist die Grundstruktur der Domino Data Sience Platform dargestellt.

Domino unterstützt wie Anaconda die Programmiersprachen Python und R [Domc]. Aus diesem Grund lassen sich die in Kapitel 3.3.2 recherchierten Pakete und Bibliotheken auch für die Verbindung mit der Domino Plattform verwenden. Darüber hinaus benennt Domino auf der Supportwebsite die empfohlenen Pakete für einige Datenbanksysteme und beschreibt, wie die Verbindung hergestellt werden kann [Doma]. Folgende Datenquellen werden aufgeführt: Amazon S3, PostegreSQL, IBM DB2,

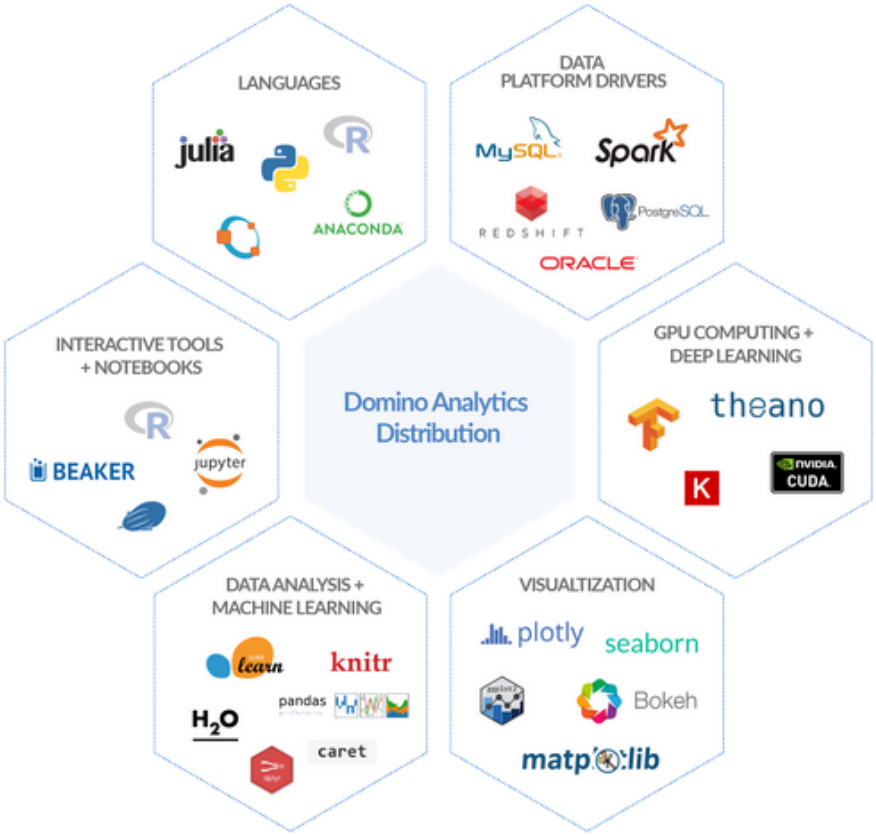

Abbildung 11: Aufbau der Domino Data Science Platform [Domc]

Oracle, Snowflake, MSSQL, Google BigQuery und Amazon Redshift.

Sonstige: Auch eine Verbindung zu Apache Spark und Apache Hadoop ist möglich [Doma].

3.3.7 H2O (H2O.ai)

Die Software H2O der Firma H2O.ai schließt sich in die Reihe der Open-Source Plattformen an. Entgegen Anaconda und Domino legt H2O den Schwerpunkt hierbei allerdings auf Algorithmen aus dem Bereich Maschinelles Lernen und spezifiziert die unterstützten Pakete und Funktionen genau [H2Oc; H2Oa]. Diese umfassen auch klassische Data Mining-Funktionen wie etwa die Cluster-Analyse oder Klassifikation [H2Oa]. Obwohl H2O die Programmiersprachen Python und R unterstützt und somit auf eine große Auswahl an Paketen zugreifen kann, grenzt es in der Online-Dokumentation auch die verwendbaren Schnittstellen und Datenbanksysteme ein:

JDBC: Relationale Datenbanken können über JDBC eingebunden werden [H2Ob].

Unterstützt werden MySQL, PostgreSQL, MariaDB, Netezza, Amazon Redshift und Hive [H2Ob].

Darüber hinaus ist Amazon S3 als 'Default Data Source' eingerichtet [H2Ob]. Weitere Datenquellen können über ein API der Firma Alluxio eingelesen werden [H2Ob]. Dies unterstützt laut der Produktwebsite die Cloud-Dienste Amazon S3, Google Cloud Storage, Microsoft Azure und Alibaba Object Storage Service, die Objektdatenbanken EMC Elastic Cloud Service, IBM Cloud Object Storage, Ceph, FusionStor und Minio sowie die Datenbank HBase [All17].

Sonstige: H2O kann die folgenden Dateiformate einlesen: CSV, ORC, SVMLight, ARFF, XLS, XLSX, Avro und Parquet [H2Ob].

3.3.8 SPSS Modeler (IBM)

Untersucht wurde die Version 17.1 des SPSS Modeler der Firma IBM. Dieser bietet eine breite Palette an Data Sience-Funktionen, wie etwa Klassifierungs-, Segmentierungs und Assoziationsalgorithmen, die Analyse von Texten oder Geodaten bis hin zur Unterstützung von neuronalen Netzen und Regressionsmodellen [IBMf].

IBM stellt dem Anwender im SPSS Modeler sogenannte Quellenknoten zur Verfügung, über die unterschiedliche Daten importiert werden können [TSC15, S. 7]. Dies umfasst folgende Datenbanken:

ODBC: Der SPSS Modeler bietet einen Datenbankknoten an, mit dem SQL-Datenbanken auf Basis von ODBC verwendet werden können [TSC15, S. 7]. Für kompatible Datenbanken stellt IBM die jeweiligen Treiber im sogenannten SPSS Data Access Pack zur Verfügung [TSC15, S. 18]. Während der Installation kann ausgewählt werden, welche Treiber installiert werden sollen. Unterstützt werden DB2, Informix, Oracle, Microsoft SQL Server, Sybase, GreenPlum, Teradata, MySQL und Redshift.

XML: Über einen XML-Quellenknoten können XML-Dokumente importiert werden [TSC15, S. 8].

IBM: Über zwei Quellenknoten können auf Daten der IBM-Datenbanken Cognos BI und Cognos TM1 zugegriffen werden [TSC15, S. 7f.].

Sonstige: Unterstützung von HDFS, Einlesen von Textdateien mit freien und festen Feldern, Einlesen von Statistikdateien (.sav, .zsav), Import von Formaten aus der

Marktforschungssoftware, Import von SAS- und Excel-Dateien, Einlesen von Karten- oder Geodaten [TSC15, S. 7f.], Unterstützung von SalesForce und Hadoop.

3.3.9 KNIME Analytics Platform

Untersucht wurde die Version 3.7.0 der KNIME Analytics Platform. KNIME ist eine kostenlose Open-Source Datenanalysesoftware, in der Analyseaufgaben durch die Kombination von mehr als 2000 sogenannten Modulen erstellt werden können [KNI]. Die folgenden Informationen ergaben sich aus der Untersuchung des Programmes.

In KNIME kann der Anwender im sogenannten 'Node Repository' über verschiedene Knoten eine Verbindung zu einer Datenbank einrichten:

JDBC: Über sogenannte 'Database-Nodes' ist der Zugriff auf Datenbanken mit JDBC-Schnittstelle möglich. Für sechs Datenbanken existiert in der KNIME-Grundversion eine vordefinierte 'Database-Node': H2, Microsoft SQL Server, MySQL, PostgreSQL, SQLite und Vertica. Über die Installation der Erweiterungen 'KNIME & Extensions' und 'KNIME Big Data Extensions' werden die Schnittstellen von Amazon Athena, Amazon Redshift, Hive und Impala implementiert. Darüber hinaus stellt KNIME einen allgemeinen JDBC-Connector-Knoten bereit, mit dem unter Angabe des entsprechenden Treibers und der JDBC-URL der spezifischen Datenbank eine neue JDBC-Verbindung definiert werden kann.

NoSQL: KNIME bietet 'Structured Data-Nodes' an, mit denen auf Dateien in den Formaten XML und JSON zugegriffen werden kann.

Sonstige: Einlesen von Zeitreihen, Zugriff auf Google Analytics und Twitter.

3.3.10 MATLAB for Data Analytics (MathWorks)

Untersucht wurde die Software MATLAB for Data Analytics von der Firma MathWorks. Bei MATLAB handelt es sich um eine Produktfamilie, in der eine Desktop-Umgebung mit einer eigenen, für Matrix-basierte Mathematik ausgelegte Programmiersprache verbunden wird [Matd]. Über kombinierbare 'Toolboxes', kann MATLAB an unterschiedliche Anwendungsfälle angepasst werden [Mate]. MATLAB for Data Analytics enthält die Toolboxes Global Optimization, Parallel Computing, Curve Fitting, Deep Learning, Statistics and Machine Learning, Optimization, Database, Text Analytics und Symbolic Math [Matc]. Über die 'Database Toolbox' kann auf relationale und postrelationale Datenbanksysteme zugegriffen werden [Matc; Matb].

SQL-Datenbanken: Für die Verbindung mit relationalen Datenbanken unterstützt MATLAB die Schnittstellen ODBC und JDBC [Matf].

NoSQL-Datenbanken: MATLAB unterstützt die Datenbanken Cassandra, MongoDB und Neo4j [Matb].

Sonstige: MATLAB kann außerdem die folgenden Dateiformate importieren: Textdateien, Microsoft Excel, Bilder (unter anderem JPEG, TIFF, PNG), NetCDF, HDF, FITS, CDF, Audio- und Videodateien und XML [Matg] sowie JSON-Dokumente und Binärdateien [Mata]. Auch auf Internetinhalte (TCP/IP, RESTful Webservices, E-Mail, FTP) kann zugegriffen werden [Mata].

3.3.11 Microsoft SQL Server Analysis Services

Untersucht wurde die Software Microsoft SQL Server 2017. Die Microsoft SQL Server Analysis Services (SSAS) unterstützen drei Arten der Data-Mining-Modellbildung: relationale Datenmodelle in Form von Tabellen, mehrdimensionale Datenkonstrukte gemäß dem Online Analytical Processing (OLAP) und das visuelle Modell 'Power Pivot', das allerdings auf einer relationalen Infrastruktur basiert [Mic18b]. Je nach Art des Modells, das erstellt werden soll, werden unterschiedliche Datenquellen unterstützt:

Relational: SSAS 2017 unterstützt das sogenannte Kompatibilitätslevel 1400 [Mic18b]. Aus der Familie der Cloud-Dienste Microsoft Azure unterstützt SSAS die Datenquellen Azure SQL Database, SQL Data Warehouse, Blob Storage, Table Storage, Cosmos DB, Data Lake Store, HDInsight HDFS und HDInsight Spark [Mic18e]. Für die Datenquellen Microsoft SQL Server, Microsoft SQL Server Data Warehouse, Oracle und Teradata kann im 'In-memory'-Modus, bei dem die zu untersuchenden Daten lokal in den Speicher des SSAS-Servers kopiert werden [Mic18f], sowohl über einen OLE-DB-Provider als auch über die ADO.NET-Schnittstelle zugriffen werden [Mic18e]. Für den Zugriff im Modus 'DirectQuery', in dem Anfragen direkt auf dem Datenbank-Server verarbeitet werden [Mic18f], werden dagegen lediglich die ADO.NET-Provider unterstützt [Mic18e]. Im 'In-memory'-Modus unterstützt SSAS außerdem die Datenquellen Access, IBM Informix, JSON-Dokumentdatenbanken, MySQL, PostgreSQL, SAP HANA, SAP Business Warehouse und Sybase [Mic18e]. Dateien können in den folgenden Formaten importiert werden: Excel, JSON, Text/CSV und XML [Mic18e]. Unter Vewendung von OLE-DB oder ODBC kann auch eine neue Datenbankverbindung eingerichtet werden [Mic18e].

Abschließend unterstützt SSAS im relationalen Datenmodell die hier nicht wei-

ter behandelten Online-Dienst-Formate Dynamics 365, Exhange Online, Saleforce Objects, Salesfoce Reports und SharePoint Online Lists, sowie Active Directory, Exhange, OData Feed und SharePoint Lists [Mic18e].

Mehrdimensional: Für mehrdimensionale Modelle unterstützt SSAS die Datenbanken Access (OLE-DB), SQL Server (OLE-DB, ADO.NET), Oracle (OLE-DB, ADO.NET), Teradata (OLE-DB, ADO.NET), Informix (OLE-DB), IBM DB2 (OLE-DB), Sybase (OLE-DB) [Mic18h]. Zusätzlich kann auf jede Datenbank zugegriffen werden, für die ein OLE-DB-Provider existiert [Mic18h]. ODBC-Datenquellen werden für mehrdimensionale Datenmodelle nicht unterstützt [Mic18h].

3.3.12 RapidMiner Studio

Untersucht wurde die Version RapidMiner Studio 9.0. RapidMiner Studio ist eine Data Science-Anwendung, welche mehr als 1500 Algorithmen und Funktionen aus den Bereichen Data Mining und Maschinelles Lernen anbietet [Rapb]. Es kann auf Grundlage der Programmiersprachen Python und R erweitert werden und unterstützt externe Pakete und Bibliotheken [Rapb; Rapc].

RapidMiner Studio verwendet für die Einbindung von Datenbanken laut der Unternehmenswebsite die JDBC-Schnittstelle [Rapc]. RapidMiner Studio liefert für einige SQL-Datenbanken die benötigten Treiber mit, sodass diese im Programm ohne weitere Konfiguration auswählbar sind. Folgende Datenbanken lassen sich verwenden:

JDBC: Mitgeliefert und im Programm auswählbar sind MySQL, PostgreSQL, Sybase, HSQLDB, Ingres, Microsoft Access, Microsoft SQL Server und Oracle. Im Allgemeinen sind aber alle Datenbanken mit JDBC-Unterstützung kompatibel [Rapc].

JDBC-ODBC-BRIDGE: RapidMiner Studio bietet im Programm außerdem die Verwendung einer JDBC-ODBC-Bridge an, sodass sich auch ODBC-kompatible Datenbanken einlesen lassen.

NoSQL: Darüber hinaus bietet RapidMiner Studio die Möglichkeit über Erweiterungen die NoSQL-Datenbankenen Cassandra und MongoDB zu verwenden [Rapa]. Auch auf andere XML-fähige Dokument-Datenbanken kann zugegriffen werden [Rapc].

Sonstige: Einlesen von SAS-, ARFF-, Excel-, Stata-Dateien; Zugriff auf Dropbox und Amazon S3; Einlesen von Textdokumenten, Internetseiten, PDF und HTML; Zugriff auf Twitter und Salesforce.com; Zugriff auf Audiodaten, Bilder, Zeitreihen [Rapc].

3.3.13 SAP BW/4HANA

Laut der eingangs genannten Studie des Frauenhofer Instituts für Produktionstechnik und Automatisierung verwendeten 13 % der befragten Produktionsunternehmen für Data Mining-Aufgaben im Jahr 2014 ein Business Intelligence (BI)-Tool von SAP [Wes+14, S.19]. Unter dem Begriff Business Intelligence werden Prozesse zusammengefasst, die der Entscheidungsfindung von Unternehmen auf Grundlage der Analyse vorhandener Daten dienen [MK16, S. 199 f.]. Auch im Querschnitt der mayato GmbH aus dem Jahr 2009 wird mit der SAP Netweaver Data Mining Workbench, die eine Komponente der Plattform SAP Netweaver BI ist, eine BI-Lösung genannt [Dil09, S. 3f.]. In der Studie der mayato GmbH wurde schon 2009 eine Fusion von Datenverwaltungs- und Datenanalysesystemen beobachtet [Dil09, S. 3f.]. Es zeigt sich, dass sich dieser Trend weiter fortgesetzt hat. Zwar bietet SAP weiter BI-Lösungen wie etwa die Software SAP BusinessObjects Business Intelligence oder SAP Lumira an, diese sind jedoch auf einzelne Anforderungen spezialisiert [SAPa]. Die Plattform SAP Netweaver BI hingegen ist auf der SAP-Website nicht mehr zu finden, stattdessen wird die Data Warehouse-Komplettlösung SAP BW/4HANA beworben, die die relationale Datenbank SAP HANA beinhaltet [SAPc]. Ein Data Warehouse dient nach Meier & Kaufmann (2016) dazu, Daten aus verschiedenen Datenquellen zusammenzufassen, die Daten entlang einer Zeitachse abzubilden und eine Datenanalyse gemäß dem OLAP-Prinzip zu ermöglichen [MK16, S. 201]. SAP HANA enthält Data Mining-Algorithmen, mit denen sich unter anderem Textdateien, räumliche Daten, Prozess-, Serien- und Streaming-Daten sowie vernetzte Daten in Form von Diagrammen untersuchen lassen [SAPf].

Die Integration von Daten in SAP HANA ist in dem Paket SAP HANA Smart Data Integration geregelt [SAPd]. Auf der Supportwebsite des Pakets verweist SAP auf eine 'Product Availability Matrix' in der die unterstützte Hard- und Software spezifiziert ist [SAPe]. Bedauerlicherweise gehört die Matrix zu einem Bereich, der nur für Kunden von SAP einsehbar ist. Leider hat sich SAP nicht bereit erklärt, die Information für diese Projektarbeit herauszugeben, sodass die unterstützten Datenbanksysteme nicht benannt werden können.

3.3.14 SAS Enterprise Miner

Untersucht wurde die Version 15.1 des Data Mining-Programms SAS Enterprise Miner. Dessen Daten-Zugriffs-, Manipulations- und Management-Funktionalitäten beruhen auf der 'SAS 9.4 Intelligence Platform' [SASa]. Für die Verbindung mit Datenbanksystemen verwendet diese die ODBC-Schnittstelle [SASc]. Auch XML-Dateien können über die 'SAS 9.4 Intelligence Platform' eingelesen werden [SASd].

Darüber hinaus stellt SAS Enterprise Miner für den Import von Dateien einen 'File Import Node' zur Verfügung [SASb]. Hierüber lassen sich folgende Dateiformate einlesen: dBase (.dbf), Stata (.dta), Microsoft Excel (.xls, .xlsx), SAS JMP (.jmp), Paradox. DB (.db), SPSS (.sav), Lotus (.wk1, .wk3, .wk4), Textdateien (.txt, .csv), .dlm-Dateien [SASb].

Optional kann mithilfe der Software SAS/ACCESS auf weitere Datenbanken zugriffen werden. SAS/ACCESS stellt vordefinierte Schnittstellen zu einer großen Anzahl an Datenbanksystemen, sowie offene Schnittstellen auf Basis von JDBC, ODBC und OLE-DB, bereit [SASd].

3.3.15 Teradata

Teradata bietet mit dem Softwareprodukt Ventage eine Komplettlösung für die Speicherung, Verwaltung und Analyse von Daten an [Terb, S. 1]. Als Teil der 'Teradata SQL engine' enthält Ventage auch die Datenbank Teradata Database [Tera]. Der Import von Daten aus externen Datenbanken und entsprechende Schnittstellen sind nicht vorgesehen.

3.3.16 Statistica (StatSoft/TIBCO)

Untersucht wurde die Version 13.5 der Software Statistica von der Firma StatSoft. Statistica wird auch von der Partnerfirma TIBCO Software Inc. vertrieben [TIBb]. Es handelt es sich dabei um eine modulare Softwarelösung, in der der Leistungsumfang durch mehrere Produktvarianten an individuelle Anforderungen angepasst werden kann [Staa]. In dem Paket 'Statistica Modeler' sind verschiedene Data Mining-Algorithmen enthalten. Eine noch größere Auswahl enthält das Paket 'Statistica Data Scientist', das unter anderem um Text Mining und Funktionen der Prozessoptimierung ergänzt wurde [Staa].

Statistica kann über einen sogenannten 'Streaming Database Connector' auf eine externe, über das Netzwerk verbundene Datenbank zuzugreifen [TIBb]. Hierbei hat der Anwender im Programm die Möglichkeit eine Datenbankverbindungen über OLE-DB oder ADO.NET einzurichten.

OLE-DB: Der Benutzer kann über verschiedene OLE-DB-Provider auf Datenbanken zugreifen.

1. OLE-DB-Provider für ODBC-Treiber: Über die OLE-DB-ODBC-Bridge können alle Datenbanken mit ODBC-Schnittstelle verwendet werden.
2. OLE-DB-Provider für Microsoft SQL Server.

3. Microsoft OLE-DB Simple Provider: Über den OLE-DB Simple Provider kann eine Verbindung zu Datenquellen hergestellt werden, die lediglich einen grundlegenden OLE-DB-Support benötigen [Mic18g]. Dies sind zum Beispiel XML-Dokumente [Mic18g].

ADO.NET: Auch über die ADO.NET-Schnittstelle lässt sich eine Verbindung zu allen ODBC-Datenbanken und Microsoft SQL Server herstellen. Zusätzlich wird ein Data-Provider für die Datenbanken Oracle und Microsoft Access angeboten. Außerdem hat der Benutzer die Möglichkeit eine Datenbankverbindung über einen OLE-DB-Provider selbst zu definieren.

Darüber hinaus können in Statistica die 'Spotfire Data Connections' der Firma TIBCO verwendet werden [Stab]. Diese unterstützen in der Version 10.0 die Datenbanken Amazon Redshift, IBM DB2, IBM Netezza, Microsoft SQL Server, Oracle, Oracle Essbase, Pivotal Greenplum, PostgreSQL, SAP HANA, Teradata und Vertica [TIBa].

Sonstige: Lokale Dateien der folgenden Formate können eingelesen werden: Statistica (.sta, .smx, .scr, .sta, .css), Excel (.xls, .xlsx, .xlsm, .xlsb), dBASE (.dbf), Lotus/Quattro (.wk1, .wk3, .wq1), Textdateien (.txt, .csv), HTML (.htm), SPSS (.sav, .por), SAS (.sd, .ssd, .sas7, .tpt, .xpt), JMP (.jmp), Minitab (.mtw). Folgende weitere Anwendungen/Systeme werden unterstützt: Apache Drill, Apache Spark SQL, Attivio, Cloudera Hive, Cloudera Impala, Dremio, Google Analytics, Hortonworks, OData, Oracle Essbase, Pivotal HAWQ und Salesforce [TIBa].

3.3.17 Oracle Data Mining

Oracle Data Mining ist eine Komponente der Oracle Advanced Analytics-Option, die in der Enterprise Edition der relationalen Datenbank Oracle Database 12c zum Einsatz kommt [Orae; Oraa]. Es handelt sich somit nicht um ein eigenständiges Data Mining-Programm, sondern beinhaltet Algorithmen, welche auf die in der Datenbank gespeicherten Daten angewandt werden können [Orab]. Die Oracle Data Mining-Option kommt daher ohne Datenimport aus einer fremden Datenbank aus [Orac]. Hieraus ergibt sich, dass sich Oracle Data Mining lediglich bei Verwendung der Oracle Database einsetzen lässt.

3.3.18 Weka

Untersucht wurde die Version 3.8.3 der Data Mining-Software Weka der University of Waikato. Weka benutzt für die interne Datenspeicherung das Dateiformat ARFF [FHW16, S. 17]. Um Dateien in anderen Formaten zu importieren oder eine Ver-

bindung zu einer Datenbank herzustellen, bietet Weka 'Converter' an [FHW16, S. 23]:

JDBC: Für den Zugriff auf Datenbanken bietet Weka einen 'Converter' an, der auf die Daten von JDBC-Datenbanken zugreifen kann [Bou+, S. 185].

Sonstige: Einlesen von C4.5-Dateien (.names, .data), .bsi-Dateien, Textdateien (.csv), LIBSVM-Dateien (.libsvm), in XML oder JSON geschriebene ARFF-Dateien (.xrff, .json), SVM-Dateien (.dat), Matlab-Dateien (.m) [Bou+, S. 185; FHW16, S. 23].

3.3.19 KXEN Analytic Framework

Die Firma KXEN Inc. wurde im Jahr 2013 von SAP übernommen [Kal13]. Die KXEN-Technologie sollte laut Kalenda (2013) in verschiedene SAP-Lösungen integriert werden [Kal13]. Die Software KXEN Analytic Framework wird nicht mehr vertrieben.

3.3.20 Viscovery SOMine

Untersucht wurde die Version 7.2 der Data Minig-Software Viscovery SOMine.

In der Grundversion von Viscovery SOMine 7 können als Datenquellen Dateien verschiedener Formate importiert werden: Textdokumente (.txt, .csv), Excel-Dateien (.xlsx, .xls), SPSS-Dateien (.sav), Viscovery-Dateien (.dms) und von Viscovery definierte XML-Dokumente (.xml) [Visb, S. 33].

Mit der Erweiterung 'Enterprise Data' ist es möglich ein Datenbanksystem als Datenquelle einzurichten [Visb, S. 34]. Als Schnittstelle verwendet Viscovery SOMine OLE-DB und ODBC [Visa].

3.3.21 prudsys Discoverer / Basket Analyzer

Die Software prudsys Discoverer und prudsys Basket Analyzer sind im aktuellen Portfolio der Firma prudsys nicht mehr enthalten [pru]. Im Jahr 2008 waren diese laut eines Artikels des Internetmagazins ixtenso Teil der prudsys Expert Mining Suite [bet08], allerdings hat eine Recherche keine Informationen zu dem Verleib der Software oder einer Neuausrichtung der Firma prudsys ergeben.

3.3.22 Bissantz Delta Master

Untersucht wurde die BI-Software Delta Master der Firma Bissantz. Trotz ihrem Schwerpunkt auf BI-Elemente beinhaltet sie auch Data Mining-Funktionen, wie

zum Beispiel multidimensionale Rangfolgen, Komponentenvergleiche, Warenkorbanalysen und Bayes-Verfahren [Bisa]. Die Firma Bissantz stellt ein 'Factsheet' über die unterstützten Schnittstellen auf der Unternehmenswebsite zur Verfügung [Bisb]. Hierbei sind im Bereich der Datenbanksysteme insbesondere OLE-DB und ODBC zu nennen [Bisb]. Mit der Schnittstelle ODP.NET stellt Bissantz darüber hinaus für das Datenbanksystem Oracle eine ADO.NET-Schnittstelle bereit [Oraf; Bisb]. Außerdem können Microsoft Access-Datenbanken (.mdb) und Excel-Dateien (.xls) importiert werden [Bisb].

Sonstige: Schnittstelle zu Microsoft SQL Server Analysis Services über XMLA und ADOMD.NET [Bisb; Mic17; Mic18a], Schnittstelle zu SAP-Produkten über SAP BAPI [Bisb], Unterstützung der Schnittstellen ODBO und OCI [Bisb].

3.4 Unterstützte Schnittstellen der Datenbanksysteme

Aus der Schnittstellenbetrachtung des vorherigen Kapitels ergeben sich 47 Datenbanksysteme und Speicherdienste, auf die eines oder mehrere der genannten Data Mining-Anwendungen zugreifen können. Aus diesen Ergebnissen eine Kompatibilitätstabelle zu erstellen ist aber nicht sinnvoll, da einige Data Mining-Programme lediglich die Unterstützung einer Programmierschnittstelle spezifizieren. Für die meisten Datenbanksysteme aus Kapitel 3.3 ist jedoch nicht bekannt, für welche Schnittstellen Treiber oder Data Provider existieren. Dies soll daher zunächst untersucht werden. Die Ergebnisse dieses Kapitels sind als Ergänzung der Ergebnisse des Kapitels 3.3 zu sehen. Es werden lediglich die Schnittstellen untersucht und im folgenden genannt, deren Unterstützung sich aus Kapitel 3.3 nicht ergibt.

ODBC: Amazon Redshift [Amab], Amazon S3 [Sima], Apache Cassandra [Simb], Couchbase [Cou], DataStax [Prob], dBase [Proc], ElasticSearch [Ela18b], Exasol [EXA], HBase [Simc], H2 [H2], HP Vertica [Mica], PostgreSQL [The13], IBM DB2 [IBMa], IBM Informix [IBMc], Ingres [Actc], MariaDB [Marc], Microsoft Access [Mich], Microsoft Azure Cosmos DB [Mic18c], Microsoft Azure SQL Database [Micb], Microsoft Azure SQL Data Warehouse [Mic18d], Microsoft Azure Table Storage [CDaa], Microsoft SQL Server [Micb], MongoDB [Pay18], MySQL [Orad], Pivotal GreenPlum [Pivb], Redis [Mici], Teradata [Ter15], SAP HANA [SAPb], Snowflake [Sno], SQLite [Dev] und Sybase [SAP13b].

JDBC: Amazon Redshift [Amaa], Apache Cassandra [Simb], Couchbase [Cou], DataStax [Proa], ElasticSearch [Ela18a], Exasol [EXA], Google BigQuery [Goo], HBase [CDac], H2 [H2], IBM DB2 [IBMa], IBM Informix [IBMb], Ingres [Actb], MariaDB [Marb], Microsoft Azure SQL Database [Micb], Microsoft Azure SQL Data Ware-

house [Mic18d], Microsoft SQL Server [Micb], MySQL [Orad], Pivotal Greenplum [Piva], PostgreSQL [Theb], PostgreSQL [Theb], Redis [CDaf], SAP HANA [SAP13a], Snowflake [Sno], SQLite [SQL], Sybase [SAP13a], Teradata [Ter15] und Neo4j [Hun16].

OLE-DB: Vertica [Mica], IBM DB2 [IBMa], IBM Informix [IBMd], Microsoft Access [Mich], Microsoft Azure SQL Database [Altc], Microsoft SQL Server [Micb], PostgreSQL [Thec], Sybase [SAP13c] und Teradata [Ter15].

ADO.NET: Apache Cassandra [CDab], Couchbase [Micc], DataStax [Dati], Elastic-Search [Mice], Exasol [EXA], Google BigQuery [Micf], HBase [Micg], Vertica [Mica], IBM Informix [IBMe], Ingres [Acta], MariaDB [Mara], Microsoft Azure SQL Database [Micb], Microsoft Azure SQL Data Warehouse [Mic18d], Microsoft SQL Server [Micb], Minio [Min], MongoDB [CDad], MySQL [Orad], Neo4j [Neo], PostgreSQL [Thea], Redis [CDae], SAP HANA [SAPg], Snowflake [Sno], SQLite [Micd], Sybase [Prod] und Teradata [Ter15].

Es fällt auf, dass auch für viele NoSQL-Datenbanken, wie etwa MongoDB und Cassandra, ODBC- und JDBC-Treiber existieren. Dies wird realisiert, indem die Treiber SQL-Funktionalitäten auf die jeweilige Programmierschnittstellen der NoSQL-Datenbanken abbilden [Sim12]. Im Fall von MongoDB nutzt der MongoDB BI Konnektor den ODBC-Treiber um SQL-Anfragen in die programmeigene Sprache MongoDB Query Language zu übersetzen [Pay18].

4 Schnittstellen gängiger Data Mining-Werkzeuge

Im Folgenden werden die Ergebnisse der Kapitel 3.3 und 3.4 in einer grafischen Übersicht zusammengefasst. Aufgrund der Größe der Tabelle wurde diese in der Druckversion auf drei Teile aufgeteilt. Für eine bessere Übersicht empfiehlt sich die Betrachtung des Excel-Sheets, das dem Fachgebiet vorliegt. In den Tabellen 1 bis 3 sind die betrachteten Datenbanksysteme und Speicherdienste den Data Mining-Anwendungen gegenübergestellt und bei Kompatibilität mit einem 'X' versehen. Neben SQL- und NoSQL-Datenbanksystemen enthält die Liste auch einige Data Warehouses sowie Cloud-Objektspeicherdienste. Objektspeicher sind nicht mit objektorientierten oder objektrelationalen Datenbanken zu verwechseln, sondern basieren auf einem grundsätzlich verschiedenem Speicherprinzip. Der Speicher wird, statt in einzeln addressierbare Blöcke von konstanter Größe, in Objekte von beliebiger Größe aufgeteilt [MGR03, S. 84]. Neben den Daten selbst werden in den Objekten auch Metadaten abgelegt [MGR03, S. 86]. Da die Struktur der Daten beliebig ist, lassen sich Dateien, Bilder, Multimedia-Inhalte und sogar Datenbanken in einem Objekt speichern [MGR03, S. 86]. Da viele Data Mining-Anwendungen Schnittstellen zu Objektspeichern aufweisen, wurden diese in die Kompatibilitätsliste mit aufgenommen.

Zum besseren Verständnis sind einige Begriffe und Punkte der Tabelle mit Anmerkungen gekennzeichnet, die nachfolgend erläutert werden:

- $*^1$ Diese Datenbanksysteme basieren zwar auf dem relationalen Datenmodell, speichern die Daten physisch aber nicht in Zeilen, sondern in Spalten ab. Dieser spaltenorientierte Ansatz verbindet damit die Vorteile der spaltenweisen Speicherung mit der Struktur relationaler Datenbanksysteme.
- $*^2$ Die Rubrik 'Multi-Modell' enthält Datenbanksysteme, die sowohl relationale als auch postrelationale Datenmodelle unterstützen. Das Datenbanksystem DataStax Enterprise unterstützt beispielsweise die Modellbildung auf Grundlage von Tabellen, Schlüssel-Wert-Kombinationen, JSON-Dokumenten und Graphen [Dath].
- $*^3$ Statistica sieht im Programm die Verwendung einer OLE-DB-ODBC-Bridge vor. Da die Performanz bei einer solchen Verbindung geringer ist, als bei direkter Verwendung einer Schnittstelle, sind Datenbanksysteme und Speicherdienste, für die kein OLE-DB- oder ADO.NET-Provider existiert, blau eingefärbt.
- $*^4$ Wenn die Anwendung die Einrichtung einer neuen Datenbankverbindung über ODBC, JDBC, OLE-DB oder ADO.NET erlaubt, ist das an dieser Stelle gekennzeichnet.

			Alteryx	Anaconda	Knowledge SEEKER	Databricks	Dataiku	Domino	H2O
SQL-Datenbank		Amazon Aurora	X		X				
		Esri GeoDatabase	X	X				X	
		H2				X	X		
		HSQLDB				X	X		
		IBM DB2	X		X	X	X	X	
		Ingres			X	X	X		
		MariaDB		X	X	X	X	X	X
		MemSQL					X		
		Microsoft Access	X		X	X	X		
		Microsoft Azure SQL Database	X		X	X	X		X
		Microsoft SQL Server	X	X	X	X	X	X	
		MySQL	X	X	X	X	X	X	X
		Oracle	X	X	X	X	X	X	
		Pivotal Greenplum	X	X	X	X	X	X	
		PostgreSQL	X	X	X	X	X	X	X
		SQLite		X	X	X	X	X	
		SAP HANA	X		X	X	X		
		SAP Sybase ASE		X	X	X	X	X	
	spalten-orientiert *1	Exasol	X	X	X	X	X	X	
		HP Vertica	X	X	X	X	X	X	
		Teradata		X	X	X	X	X	
Multi-Modell *2		DataStax	X		X	X	X		
		IBM Informix			X	X	X		
		Microsoft Azure Cosmos DB			X	X			X
NoSQL-Datenbank	Schluessel-Wert	HBase		X	X	X	X	X	X
		Microsoft Azure Table Storage			X				
		Redis		X	X	X	X	X	
	Doku-ment	Couchbase Server		X	X	X	X	X	
		ElasticSearch		X	X	X	X	X	
		MongoDB	X	X	X	X	X	X	
	Spal.-F.	Apache Cassandra	X	X	X	X	X	X	
	Graph	Neo4j		X		X	X	X	
	OLAP	IBM Cognos TM1							
Data-Warehouse		Amazon Redshift	X	X	X	X	X	X	X
		Google BigQuery	X	X	X	X	X	X	
		Microsoft Azure SQL Data Warehouse			X	X	X		
		Snowflake	X	X	X	X	X	X	
Cloud-Objektspeicher		Alibaba Object Storage Service							X
		Amazon S3	X	X	X	X	X	X	X
		Ceph		X				X	X
		EMC Elastic Cloud Service							X
		Google Cloud Storage		X			X	X	X
		IBM Cloud Object Storage							X
		Microsoft Azure Blob Storage		X		X	X	X	X
		Microsoft Azure Data Lake Store	X			X	X		X
		Minio		X				X	X
Unterstuetzte Schnittstelle *4		ODBC			X				
		JDBC				X	X		
		OLE-DB							
		ADO.NET							

Tabelle 1: Kompatibilitätsliste - Teil 1

			SPSS Modeler	KNIME	MATLAB	Microsoft SQL Server		RapidMiner
						Relationales Modell	Mehrdim. Modell	
SQL-Datenbank		Amazon Aurora			X	X		
		Esri GeoDatabase						
		H2		X	X	X		X
		HSQLDB		X	X			X
		IBM DB2	X	X	X	X	X	X
		Ingres		X	X	X		X
		MariaDB		X	X	X		X
		MemSQL						
		Microsoft Access		X	X	X	X	X
		Microsoft Azure SQL Database		X	X	X	X	X
		Microsoft SQL Server	X	X	X	X	X	X
		MySQL	X	X	X	X		X
		Oracle	X	X	X	X	X	X
		Pivotal Greenplum	X	X	X	X		X
		PostgreSQL		X	X	X	X	X
		SQLite		X	X	X		X
		SAP HANA		X	X	X		X
		SAP Sybase ASE	X	X	X	X	X	X
	spalten-orientiert *1	Exasol		X	X	X		X
		HP Vertica		X	X	X	X	X
		Teradata	X	X	X	X	X	X
Multi-Modell *2		DataStax		X	X	X		X
		IBM Informix	X	X	X	X	X	X
		Microsoft Azure Cosmos DB			X	X		
NoSQL-Datenbank	Schluessel-Wert	HBase		X	X	X		X
		Microsoft Azure Table Storage			X	X		
		Redis		X	X	X		X
	Doku-ment	Couchbase Server		X	X	X		X
		ElasticSearch		X	X	X		X
		MongoDB			X	X		X
	Spal.-F.	Apache Cassandra		X	X	X		X
	Graph	Neo4j		X	X			
	OLAP	IBM Cognos TM1	X					
Data-Warehouse		Amazon Redshift	X	X	X	X		X
		Google BigQuery		X	X	X		X
		Microsoft Azure SQL Data Warehouse		X	X	X	X	X
		Snowflake		X	X	X		X
Cloud-Objektspeicher		Alibaba Object Storage Service						
		Amazon S3			X	X		
		Ceph						
		EMC Elastic Cloud Service						
		Google Cloud Storage						
		IBM Cloud Object Storage						
		Microsoft Azure Blob Storage				X		
		Microsoft Azure Data Lake Store				X		
		Minio						
Unterstuetzte Schnittstelle *4		ODBC			X	X		
		JDBC		X	X			
		OLE-DB				X	X	
		ADO.NET						

Tabelle 2: Kompatibilitätsliste - Teil 2

			SAS Enterprise Miner	Teradata	Statistica	Oracle	Weka	SOMine	Delta Master
SQL-Datenbank		Amazon Aurora			X *3			X	X
		Esri GeoDatabase							
		H2	X		X *3		X	X	X
		HSQLDB					X		
		IBM DB2	X		X		X	X	X
		Ingres	X		X *3		X	X	X
		MariaDB	X		X *3		X	X	X
		MemSQL							
		Microsoft Access	X		X		X	X	X
		Microsoft Azure SQL Database	X		X		X	X	X
		Microsoft SQL Server	X		X		X	X	X
		MySQL	X		X *3		X	X	X
		Oracle	X		X	X	X	X	X
		Pivotal Greenplum	X		X *3		X	X	X
		PostgreSQL	X		X		X	X	X
		SQLite	X		X *3		X	X	X
		SAP HANA	X		X		X	X	X
		SAP Sybase ASE	X		X		X	X	X
	spalten-orientiert *1	Exasol	X		X *3		X	X	X
		HP Vertica	X		X		X	X	X
		Teradata	X	X	X		X	X	X
Multi-Modell *2		DataStax	X		X *3		X	X	X
		IBM Informix	X		X		X	X	X
		Microsoft Azure Cosmos DB			X *3				
NoSQL-Datenbank	Schluessel-Wert	HBase	X		X *3		X	X	X
		Microsoft Azure Table Storage	X		X *3				
		Redis	X		X *3		X	X	X
	Doku-ment	Couchbase Server	X		X *3		X	X	X
		ElasticSearch	X		X *3		X	X	X
		MongoDB	X		X *3			X	X
	Spal.-F.	Apache Cassandra	X		X *3		X	X	X
	Graph	Neo4j					X		
	OLAP	IBM Cognos TM1							
Data-Warehouse		Amazon Redshift	X		X		X	X	X
		Google BigQuery	X		X *3		X	X	X
		Microsoft Azure SQL Data Warehouse	X		X		X	X	X
		Snowflake	X		X *3		X	X	X
Cloud-Objektspeicher		Alibaba Object Storage Service							
		Amazon S3	X		X *3			X	X
		Ceph							
		EMC Elastic Cloud Service							
		Google Cloud Storage							
		IBM Cloud Object Storage							
		Microsoft Azure Blob Storage							
		Microsoft Azure Data Lake Store							
		Minio							
Unterstuetzte Schnittstelle *4		ODBC	X		X *3			X	X
		JDBC					X		
		OLE-DB			X			X	X
		ADO.NET			X				

Tabelle 3: Kompatibilitätsliste - Teil 3

5 Fazit

Es zeigt sich, dass die Programmierschnittstellen ODBC, JDBC, OLE-DB und ADO.NET ihre Kernaufgaben, die Gewährleistung des Zugriffs auf eine Bandbreite von Datenbanksystemen, erfüllen. Insbesondere im Bereich der SQL- und NoSQL-Datenbanksysteme ergibt sich in der grafischen Übersicht aus Kapitel 4 ein insgesamt homogenes Bild. Bei den betrachteten Data Mining-Anwendungen lassen sich zwei Ausrichtungen unterscheiden. Die Mehrheit der Data Mining-Anwendungen unterstützt eine oder mehrere der vorgestellten Programmierschnittstellen. Einige Anwendungen setzen dagegen auf Open Source-Programmbibliotheken in Python oder R. Dieser Bereich ist aufgrund der großen Flexibilität und der Vielzahl an unterstützten Datenbanksystemen und Speicherdiensten nicht zu unterschätzen. Die Einbettung und Verwaltung der Programmbibliotheken in eine grafische Oberfläche erhöht zudem die Bedienerfreundlichkeit und macht diese Anwendungen auch für Nutzer ohne fortgeschrittene Programmierkenntnisse interessant. Des Weiteren hat sich gezeigt, dass es keine strikte Trennung von Schnittstellen für relationale und postrelationale Datenmodelle gibt. Statt für den Zugriff auf NoSQL-Datenbanksysteme auf die Schnittstellen OLE-DB oder ADO.NET zurückzugreifen, bieten die Datenbankhersteller häufig adaptierte ODBC- und JDBC-Treiber an. Diese Beobachtung bestätigt sich auch in der Treiberrecherche in Kapitel 3.4. Für die älteste Schnittstelle ODBC existieren mehr Treiber, auch im postrelationalen Bereich, als für dessen modernere Alternativen. Dementsprechend ist die Unterstützung für SQL- und NoSQL-Datenbanksysteme ähnlich groß. Lediglich im Bereich der multidimensionalen und Graph-Datenbanken ist eine geringere Kompatibilität erkennbar. Hinsichtlich der digitalen Transformation macht diese Arbeit deutlich, dass SQL-Datenbanksysteme bei den von Data Mining-Anwendungen unterstützten Datenbanken noch eine dominierende Stellung einnehmen. Der Abstand zu NoSQL- und Multi-Modell-Datenbanksystemen ist jedoch nicht groß. Zieht man zudem noch neuere Ansätze wie Objektspeicher und Data Warehouses hinzu, bilden die SQL-Datenbanksysteme sogar die Minderheit. In diesem Zuge ist zudem die breite Auswahl an Objektspeicherdiensten zu betonen, die sich aus der Schnittstellenrecherche ergeben hat. Auch wenn diese insgesamt in geringerem Maße unterstützt werden, als klassische Datenbanksysteme, zeigen die Ergebnisse doch, dass ihnen für Big Data-Anwendungen eine nicht zu vernachlässigende Rolle zukommt. Die digitale Transformation ist außerdem noch in der enormen Schnelllebigkeit des Data Science-Marktes sichtbar geworden. Die Tatsache, dass es viele der etablierten Unternehmen aus dem Jahr 2009 heute nicht mehr gibt und die teils starke Veränderung des Produktportfolios der noch existenten Firmen zeigt das Ausmaß und die Geschwindigkeit des gerade stattfindenden Wandels auf.

6 Zusammenfassung und Ausblick

Das Ziel der vorliegenden Projektarbeit war es, unterschiedliche Data Mining-Werkzeuge im Kontext der unterstützten Datenbankmanagementsysteme darzustellen. Zu diesem Zweck wurde auf der Grundlage von drei Studien eine Auswahl an zu untersuchenden Data Mining-Anwendungen getroffen. Bei der Analyse ergab sich, dass nicht nur die Schnittstellen von Programm zu Programm unterschiedlich implementiert werden, sondern deren Aufbau und Funktionsweise auch teils mehr und teils weniger transparent dokumentiert und kommuniziert werden. Während von einigen Herstellern lediglich die unterstützten Datenbanksysteme benannt werden, stellen Andere den gesamten zugrunde liegenden Code zur Verfügung und Dritte dokumentieren die verwendete Schnittstelle gar nicht. Dennoch konnten zu fast allen Data Mining-Programmen die unterstützten Datenbanksysteme ermittelt werden. Anhand der grafischen Übersicht aus Kapitel 4 wurde gezeigt, dass insgesamt eine breite Unterstützung von SQL- und NoSQL-Datenbanksystemen vorhanden ist. Allerdings werfen die Ergebnisse der Projektarbeit auch Fragen auf. Trotz der Tatsache, dass viele Hersteller postrelationaler Datenbanksysteme auf SQL basierende ODBC- und JDBC-Treiber unterstützen, bleibt unklar, ob diese hinsichtlich ihrer Funktionalität und Performanz an Data Provider für OLE-DB oder ADO.NET heranreichen können. Auch ein Vergleich der Verbindungsqualität bei Verwendung der herkömmlichen Programmierschnittstellen gegenüber von Programmbibliotheken in Python oder R wäre interessant und würde einen detaillierten Vergleich der Data Mining-Anwendungen ermöglichen. Hier ist weitere Forschung notwendig, um die Entscheidungsfindung auf einer noch fundierteren Wissensgrundlage unterstützen und erleichtern zu können. Insgesamt lässt sich abschließen, dass diese Arbeit als erste Entscheidungshilfe dienen kann, um auf Basis einer bestimmten Schnittstelle oder für eines der betrachteten Data Mining-Anwendungen oder Datenbanksysteme günstige Kombinationsmöglichkeiten einzugrenzen und auszuwählen.

Literatur

[Acta] Actian Corporation, Hrsg. *Electronic Software Distribution: Actian X, Ingres & Vector Drivers .Net Data Provider*. URL: https://esd.actian.com/product/drivers/.Net_Data_Provider/Windows_32-Bit/.Net_Data_Provider_GA (besucht am 30. 12. 2018).

[Actb] Actian Corporation, Hrsg. *Electronic Software Distribution: Actian X, Ingres & Vector Drivers JDBC*. URL: https://esd.actian.com/product/drivers/JDBC/java/JDBC (besucht am 30. 12. 2018).

[Actc] Actian Corporation, Hrsg. *Electronic Software Distribution: Actian X, Ingres & Vector Drivers ODBC*. URL: https://esd.actian.com/product/drivers/ODBC/Windows_32-Bit/ODBC_Driver_3.50 (besucht am 30. 12. 2018).

[All17] Alluxio Inc., Hrsg. *Alluxio Subscriptions*. 2017. URL: https://www.alluxio.com/products (besucht am 28. 12. 2018).

[Alta] Alteryx Inc., Hrsg. *Amazon S3*. URL: https://help.alteryx.com/current/DataSources/AmazonS3.htm (besucht am 27. 12. 2018).

[Altb] Alteryx Inc., Hrsg. *Die moderne Analytics-Plattform*. URL: https://www.alteryx.com/de/plattform (besucht am 12. 01. 2019).

[Altc] Alteryx Inc., Hrsg. *Microsoft Azure SQL Database*. URL: https://help.alteryx.com/2018.3/DataSources/SQLDB.htm (besucht am 30. 12. 2018).

[Altd] Alteryx Inc., Hrsg. *Solutions: Data Mining*. URL: https://www.alteryx.com/de/node/22256 (besucht am 12. 01. 2018).

[Alte] Alteryx Inc., Hrsg. *Supported Data Sources*. URL: https://help.alteryx.com/current/DataSources/SupportedDataSources.htm?tocpath=Data%20Sources%7CSupported%20Data%20Sources%7C_____0 (besucht am 27. 12. 2018).

[Amaa] Amazon Web Services, Hrsg. *Configure a JDBC Connection*. URL: https://docs.aws.amazon.com/redshift/latest/mgmt/configure-jdbc-connection.html (besucht am 30. 12. 2018).

[Amab] Amazon Web Services, Hrsg. *Install and Configure the Amazon Redshift ODBC Driver on Microsoft Windows Operating Systems*. URL: https://docs.aws.amazon.com/redshift/latest/mgmt/install-odbc-driver-windows.html (besucht am 30. 12. 2018).

[Anaa] Anaconda Inc., Hrsg. *2018 Anaconda State of Data Science Report*. URL: https://know.anaconda.com/State-of-Data-Science-2018_Report-Registration.html (besucht am 28. 12. 2018).

[Anab] Anaconda Inc., Hrsg. *anaconda / packages / mongodb 4.0.3*. URL: https://anaconda.org/anaconda/mongodb (besucht am 16. 01. 2019).

[Anac] Anaconda Inc., Hrsg. *anaconda / packages / oracle-instantclient 11.2.0.4.0*. URL: https://anaconda.org/anaconda/oracle-instantclient (besucht am 16. 01. 2019).

[Anad] Anaconda Inc., Hrsg. *anaconda / packages / python-sybase 0.40*. URL: https://anaconda.org/anaconda/python-sybase (besucht am 16. 01. 2019).

[Anae] Anaconda Inc., Hrsg. *anaconda / packages / teradata 15.10.0.21*. URL: https://anaconda.org/anaconda/teradata (besucht am 16. 01. 2019).

[Anaf] Anaconda Inc., Hrsg. *Anaconda Cloud: Search: cassandra*. URL: https://anaconda.org/search?q=cassandra (besucht am 15. 01. 2019).

[Anag] Anaconda Inc., Hrsg. *Anaconda Cloud: Search: Redshift*. URL: https://anaconda.org/search?q=redshift (besucht am 15. 01. 2019).

[Anah] Anaconda Inc., Hrsg. *Anaconda Distribution: The Most Popular Python/R Data Science Distribution*. URL: https://www.anaconda.com/distribution/ (besucht am 28. 12. 2018).

[Anai] Anaconda Inc., Hrsg. *anaconda-platform / packages / minio-server 2017.06.13*. URL: https://anaconda.org/anaconda-platform/minio-server (besucht am 16. 01. 2019).

[Anaj] Anaconda Inc., Hrsg. *auto / packages / amazons3 0.1: Django Storage Backend for Amazon S3*. URL: https://anaconda.org/auto/amazons3 (besucht am 15. 01. 2019).

[Anak] Anaconda Inc., Hrsg. *auto / packages / busyflow.pivotal 0.3.4*. URL: https://anaconda.org/auto/busyflow.pivotal (besucht am 16. 01. 2019).

[Anal] Anaconda Inc., Hrsg. *auto / packages / hbase-thrift 0.20.4*. URL: https://anaconda.org/auto/hbase-thrift (besucht am 16. 01. 2019).

[Anam] Anaconda Inc., Hrsg. *auto / packages / python-cephclient 0.1.0.4*. URL: https://anaconda.org/auto/python-cephclient (besucht am 15. 01. 2019).

[Anan] Anaconda Inc., Hrsg. *conda-forge / packages / azure-storage-blob 1.4.0*. URL: https://anaconda.org/conda-forge/azure-storage-blob (besucht am 16. 01. 2019).

[Anao] Anaconda Inc., Hrsg. *conda-forge / packages / elasticsearch-dsl 6.3.0*. URL: https://anaconda.org/conda-forge/elasticsearch-dsl (besucht am 15. 01. 2019).

[Anap] Anaconda Inc., Hrsg. *conda-forge / packages / google-cloud-bigquery 1.8.1*. URL: https://anaconda.org/conda-forge/google-cloud-bigquery (besucht am 16. 01. 2019).

[Anaq] Anaconda Inc., Hrsg. *conda-forge / packages / google-cloud-storage 1.13.0*. URL: https://anaconda.org/conda-forge/google-cloud-storage (besucht am 16. 01. 2019).

[Anar] Anaconda Inc., Hrsg. *conda-forge / packages / neo4j-python-driver 1.6.2*. URL: https://anaconda.org/conda-forge/neo4j-python-driver (besucht am 16. 01. 2019).

[Anas] Anaconda Inc., Hrsg. *conda-forge / packages / postgresql 10.6*. URL: https://anaconda.org/conda-forge/postgresql (besucht am 16. 01. 2019).

[Anat] Anaconda Inc., Hrsg. *conda-forge / packages / pymysql 0.8.1*. URL: https://anaconda.org/conda-forge/pymysql (besucht am 16. 01. 2019).

[Anau] Anaconda Inc., Hrsg. *conda-forge / packages / r-rmariadb 1.0.6*. URL: https://anaconda.org/conda-forge/r-rmariadb (besucht am 16. 01. 2019).

[Anav] Anaconda Inc., Hrsg. *conda-forge / packages / redis-py 3.0.1*. URL: https://anaconda.org/conda-forge/redis-py (besucht am 16. 01. 2019).

[Anaw] Anaconda Inc., Hrsg. *conda-forge / packages / snowflake-connector-python 1.7.3*. URL: https://anaconda.org/conda-forge/snowflake-connector-python (besucht am 16. 01. 2019).

[Anax] Anaconda Inc., Hrsg. *conda-forge / packages / sqlalchemy_exasol 2.0.4*. URL: https://anaconda.org/search?q=Exasol (besucht am 16. 01. 2019).

[Anay] Anaconda Inc., Hrsg. *conda-forge / packages / vertica-python 0.7.4*. URL: https://anaconda.org/conda-forge/vertica-python (besucht am 16. 01. 2019).

[Anaz] Anaconda Inc., Hrsg. *ijstokes / notebooks / sql-server-pyodbc-anac*. URL: https://anaconda.org/ijstokes/sql-server-pyodbc-anaconda-demonstration/notebook (besucht am 16. 01. 2019).

[Anaaa] Anaconda Inc., Hrsg. *ilan / packages / couchbase 2.3.3*. URL: https://anaconda.org/ilan/couchbase (besucht am 15. 01. 2019).

[Anaab] Anaconda Inc., Hrsg. *r / packages / r-rsqlite 2.1.1*. URL: https://anaconda.org/r/r-rsqlite (besucht am 16. 01. 2019).

[Anaac] Anaconda Inc., Hrsg. *User guide*. URL: http://docs.anaconda.com/anaconda-cloud/user-guide/ (besucht am 14. 01. 2019).

[bet08] beta-web GmbH, Hrsg. *prudsys EXPERT MINING SUITE*. 2008. URL: https://ixtenso.de/technologie/prudsys-expert-mining-suite-2008.html (besucht am 26. 12. 2018).

[Bisa] Bissantz & Company GmbH, Hrsg. *Business Intelligence mit DeltaMaster: Sehen, verstehen, handeln*. URL: https://www.bissantz.de/files/products/Business_Intelligence_mit_DeltaMaster_Brochuere.pdf (besucht am 20. 01. 2019).

[Bisb] Bissantz & Company GmbH, Hrsg. *Factsheet*. URL: https://www.bissantz.de/#deltamaster (besucht am 26. 12. 2018).

[Bou+] Remco R. Bouckaert u. a. *WEKA Manual for Version 3-8-3*. Hrsg. von University of Waikato. URL: https://sourceforge.net/projects/weka/files/documentation/3.8.x/WekaManual-3-8-3.pdf/download?use_mirror=netcologne&download= (besucht am 23. 12. 2018).

[CDaa] CData Software Inc., Hrsg. *Azure ODBC Driver: Read, Write, and Update Azure Tables through ODBC*. URL: https://www.cdata.com/drivers/azure/odbc/ (besucht am 30. 12. 2018).

[CDab] CData Software Inc., Hrsg. *Cassandra ADO.NET Provider*. URL: https://www.cdata.com/drivers/cassandra/ado/ (besucht am 30. 12. 2018).

[CDac] CData Software Inc., Hrsg. *HBase JDBC Driver*. URL: https://www.cdata.com/drivers/hbase/jdbc/ (besucht am 30. 12. 2018).

[CDad] CData Software Inc., Hrsg. *MongoDB ADO.NET Provider*. URL: https://www.cdata.com/drivers/mongodb/ado/ (besucht am 30. 12. 2018).

[CDae] CData Software Inc., Hrsg. *Redis ADO.NET Provider*. URL: https://www.cdata.com/drivers/redis/ado/ (besucht am 30. 12. 2018).

[CDaf] CData Software Inc., Hrsg. *Redis JDBC Driver*. URL: https://www.cdata.com/drivers/redis/jdbc/ (besucht am 30. 12. 2018).

[CL16] Jürgen Cleve und Uwe Lämmel. *Data Mining*. 2nd ed. De Gruyter Studium. Berlin: De Gruyter, 2016. ISBN: 978-3-11-045675-2. URL: https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=4793920.

[Cou] Couchbase, Hrsg. *Couchbase ODBC and JDBC Drivers*. URL: https://docs.couchbase.com/server/6.0/connectors/odbc-jdbc-drivers.html (besucht am 30. 12. 2018).

[Data] Databricks, Hrsg. *Connecting to SQL Databases using JDBC*. URL: https://docs.databricks.com/spark/latest/data-sources/sql-databases.html (besucht am 27. 12. 2018).

[Datb] Databricks, Hrsg. *Data Sources*. URL: https://docs.databricks.com/spark/latest/data-sources/index.html (besucht am 27. 12. 2018).

[Datc] Databricks, Hrsg. *Deep Learning Guide*. URL: https://docs.databricks.com/applications/deep-learning/index.html (besucht am 13. 01. 2019).

[Datd] Databricks, Hrsg. *Graph Analysis Guide*. URL: https://docs.databricks.com/spark/latest/graph-analysis/index.html (besucht am 13. 01. 2019).

[Date] Databricks, Hrsg. *Machine Learning*. URL: https://docs.databricks.com/spark/latest/mllib/index.html (besucht am 13. 01. 2019).

[Datf] Dataiku, Hrsg. *Dataiku Datasheet*. URL: http://pages.dataiku.com/hubfs/Dataiku_DataSheet.pdf (besucht am 27. 12. 2018).

[Datg] Dataiku, Hrsg. *Visual Machine Learning and Modeling in Dataiku*. URL: https://www.dataiku.com/dss/features/machine-learning/ (besucht am 13. 01. 2019).

[Dath] DataStax, Hrsg. *The Always-On, Active Everywhere, Distributed Hybrid Cloud Database: Built on Apache Cassandra*. URL: https://www.datastax.com/products/datastax-enterprise (besucht am 26. 01. 2019).

[Dati] DataStax Inc., Hrsg. *ADO.NET*. URL: https://docs.datastax.com/en/developer/csharp-driver/3.6/features/components/adonet/ (besucht am 30. 12. 2018).

[Datj] Datawatch Corporation, Hrsg. *KnowledgeSEEKER Brochure*. URL: https://www.datawatch.com/resource-center/literature/knowledgeseeker-brochure/ (besucht am 27. 12. 2018).

[Dev] Devart, Hrsg. *ODBC Driver for SQLite*. URL: https://www.devart.com/odbc/sqlite/ (besucht am 30. 12. 2018).

[Dil09] Marcus Dill, Hrsg. *Data Mining Software 2009: Funktionsvergleich und Benchmarkstudie*. 2009. URL: http://www.mayato.com/wp-content/uploads/2015/03/mayato_DMSZ_8S.pdf.

[Dob+18] Walter Doberenz u. a. *Visual C# 2017 – Grundlagen, Profiwissen und Rezepte*. München: Hanser, 2018. ISBN: 9783446453593.

[Doma] Domino Data Lab, Hrsg. *CONNECTING TO DATA SOURCES*. URL: https://support.dominodatalab.com/hc/en-us/sections/360000203383-CONNECTING-TO-DATA-SOURCES (besucht am 27. 12. 2018).

[Domb] Domino Data Lab, Hrsg. *Domino Data Science Platform: Built to let data science teams rapidly develop and deliver models.* URL: https://www.dominodatalab.com/platform/#foundation (besucht am 19. 01. 2019).

[Domc] Domino Data Lab, Hrsg. *Drive breakthrough research and deliver high-impact models: Develop and deliver models with open access to the tools you love, on scalable infrastructure that automatically tracks your work.* URL: https://www.dominodatalab.com/data-scientists/ (besucht am 19. 01. 2019).

[Ela18a] Elasticsearch B.V., Hrsg. *Download JDBC Client (Beta)*. 2018. URL: https://www.elastic.co/downloads/jdbc-client (besucht am 30. 12. 2018).

[Ela18b] Elasticsearch B.V., Hrsg. *Elasticsearch SQL ODBC Driver*. 2018. URL: https://www.elastic.co/guide/en/elasticsearch/sql-odbc/master/index.html (besucht am 30. 12. 2018).

[EXA] EXASOL AG, Hrsg. *Clients, Interfaces & Drivers: Drivers*. URL: https://www.exasol.com/portal/pages/viewpage.action?pageId=4030482 (besucht am 30. 12. 2018).

[FHW16] Eibe Frank, Mark Hall und Ian H. Witten. "The WEKA Workbench". In: *Data Mining: Practical Machine Learning Tools and Techniques*. Hrsg. von Ian H. Witten u. a. Morgan Kaufmann, 2016. URL: http://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

[Fro18] Jörg Frochte. *Maschinelles Lernen: Grundlagen und Algorithmen in Python*. München: Hanser, Carl, 2018. ISBN: 978-3-446-45291-6.

[Gar] Gartner Inc., Hrsg. *Magic Quadrant for Data Science and Machine-Learning Platforms*. URL: https://www.gartner.com/doc/3860063/magic-quadrant-data-science-machinelearning (besucht am 27. 12. 2018).

[Gei14] Frank Geisler. *Datenbanken: Grundlagen und Design*. Verlagsgruppe Hüthig Jehle Rehm, 2014. ISBN: 9783826697074.

[Goo] Google Cloud, Hrsg. *Simba-Treiber für Google BigQuery*. URL: https://cloud.google.com/bigquery/partners/simba-drivers/ (besucht am 30. 12. 2018).

[Gre14] Jeremy Greze. *Easy Text Clustering*. 2014. URL: https://blog.dataiku.com/easy-text-clustering (besucht am 13. 01. 2019).

[H2] H2, Hrsg. *H2 Database Engine*. URL: http://www.h2database.com/html/main.html (besucht am 30. 12. 2018).

[H2Oa] H2O.ai, Hrsg. *Algorithms*. URL: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html (besucht am 19. 01. 2019).

[H2Ob] H2O.ai, Hrsg. *Getting Data into Your H2O Cluster*. URL: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/getting-data-into-h2o.html (besucht am 28. 12. 2018).

[H2Oc] H2O.ai, Hrsg. *H2O: The #1 open-source machine learning platform for the enterprise*. URL: https://www.h2o.ai/products/h2o/ (besucht am 19. 01. 2019).

[Her02] Andreas Herbolsheimer. *Datenbank-Programmierung: Beispiellösungen mit Access, SQL Server und PostgreSQL*. 1. Aufl. Programmer's Choice. München und [Erscheinungsort nicht ermittelbar]: Pearson Deutschland und Addison-Wesley, 2002. ISBN: 3827319455.

[HSS18] Andreas Heuer, Kai-Uwe Sattler und Gunter Saake. *Datenbanken: Konzepte und Sprachen*. Sechste Auflage. mitp Professional. Frechen: MITP, 2018. ISBN: 978-3-95845-777-5. URL: https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=5392219.

[Hun16] Michael Hunger. *The All-New, Officially Supported Neo4j-JDBC Driver 3.0*. 2016. URL: https://neo4j.com/blog/official-neo4j-jdbc-driver-3-0/ (besucht am 06. 01. 2018).

[IBMa] IBM, Hrsg. *Db2 driver package*. URL: https://www.ibm.com/support/knowledgecenter/en/SS6NHC/com.ibm.swg.im.dashdb.doc/connecting/connect_driver_package.html (besucht am 30. 12. 2018).

[IBMb] IBM, Hrsg. *IBM Informix JDBC Driver*. URL: https://www.ibm.com/support/knowledgecenter/en/SSGU8G_12.1.0/com.ibm.jdbc_pg.doc/ids_jdbc_013.htm (besucht am 30. 12. 2018).

[IBMc] IBM, Hrsg. *IBM Informix ODBC Driver*. URL: https://www.ibm.com/support/knowledgecenter/de/SSGU8G_11.50.0/com.ibm.gsg.doc/ids_gsg_271.htm (besucht am 30. 12. 2018).

[IBMd] IBM, Hrsg. *Install and configure Informix OLE DB Provider*. URL: https://www.ibm.com/support/knowledgecenter/en/SSGU8G_12.1.0/com.ibm.oledb.doc/ids_oledb_008.htm (besucht am 30. 12. 2018).

[IBMe] IBM, Hrsg. *Installing the IBM Informix .NET Provider*. URL: https://www.ibm.com/support/knowledgecenter/en/SSGU8G_12.1.0/com.ibm.netpr.doc/ids_net_010.htm (besucht am 30. 12. 2018).

[IBMf] IBM Deutschland GmbH, Hrsg. *SPSS Modeler: Details*. URL: https://www.ibm.com/de-de/products/spss-modeler/details (besucht am 19. 01. 2019).

[IBM13] IBM, Hrsg. *IBM Netezza ODBC, JDBC, OLE DB, and .NET installation and configuration*. 2013. URL: https://www.ibm.com/support/knowledgecenter/en/SSULQD_7.2.1/com.ibm.nz.datacon.doc/c_datacon_plg_overview.html (besucht am 30. 12. 2018).

[Kal13] Florian Kalenda. *SAP verstärkt sich mit KXEN im Bereich Predictive Analytics*. 2013. URL: https://www.zdnet.de/88169291/sap-verstaerkt-sich-mit-kxen-im-bereich-predictive-analytics/ (besucht am 23. 12. 2018).

[KNI] KNIME AG, Hrsg. *KNIME Analytics Platform: Open, intuitive, integrative data science*. URL: https://www.knime.com/knime-software/knime-analytics-platform (besucht am 19. 01. 2019).

[Mara] MariaDB Foundation, Hrsg. *ADO.NET*. URL: https://mariadb.com/kb/en/library/adonet/ (besucht am 30. 12. 2018).

[Marb] MariaDB Foundation, Hrsg. *Downloads: MariaDB Connector/J 2.3 Series*. URL: https://downloads.mariadb.org/connector-java/ (besucht am 30. 12. 2018).

[Marc] MariaDB Foundation, Hrsg. *Downloads: MariaDB Connector/ODBC 3.0 Series*. URL: https://downloads.mariadb.org/connector-odbc/ (besucht am 30. 12. 2018).

[Mata] MathWorks, Hrsg. *Data Import and Export*. URL: https://de.mathworks.com/help/matlab/data-import-and-export.html (besucht am 28. 12. 2018).

[Matb] MathWorks, Hrsg. *Database Toolbox*. URL: https://de.mathworks.com/help/database/index.html (besucht am 28. 12. 2018).

[Matc] MathWorks, Hrsg. *Free MATLAB Trial for Data Analytics*. URL: https://de.mathworks.com/campaigns/products/trials/targeted/dan.html (besucht am 28. 12. 2018).

[Matd] MathWorks, Hrsg. *MATLAB*. URL: https://de.mathworks.com/products/matlab.html (besucht am 19. 01. 2019).

[Mate] MathWorks, Hrsg. *Produkte & Dienstleistungen*. URL: https://de.mathworks.com/products.html?s_tid=gn_ps (besucht am 19. 01. 2019).

[Matf] MathWorks, Hrsg. *Relational Databases*. URL: https://de.mathworks.com/help/database/relational-databases.html (besucht am 28. 12. 2018).

[Matg] MathWorks, Hrsg. *Standard File Formats*. URL: https://de.mathworks.com/help/matlab/standard-file-formats.html (besucht am 28. 12. 2018).

[Mei16] Andreas Meier. "Zur Nutzung von SQL- und NoSQL-Technologien". In: *HMD Praxis der Wirtschaftsinformatik* 53.4 (2016), S. 415–427. ISSN: 1436-3011. DOI: 10.1365/s40702-016-0225-x.

[Mei18] Andreas Meier. *Werkzeuge der digitalen Wirtschaft: Big Data, NoSQL & Co: Eine Einführung in relationale und nicht-relationale Datenbanken*. essentials. Wiesbaden: Springer Vieweg, 2018. ISBN: 978-3-658-20336-8. DOI: 10.1007/978-3-658-20337-5. URL: http://dx.doi.org/10.1007/978-3-658-20337-5.

[MGR03] M. Mesnier, G. R. Ganger und E. Riedel. "Storage area networking - Object-based storage". In: *IEEE Communications Magazine* 41.8 (2003), S. 84–90. ISSN: 0163-6804. DOI: 10.1109/MCOM.2003.1222722.

[Mica] Micro Focus, Hrsg. *Client Drivers*. URL: https://www.vertica.com/download/vertica/client-drivers/ (besucht am 30. 12. 2018).

[Micb] Microsoft Corporation, Hrsg. *Connection modules for Microsoft SQL databases*. URL: https://docs.microsoft.com/de-de/sql/connect/sql-connection-libraries?view=sql-server-2017 (besucht am 30. 12. 2018).

[Micc] Microsoft Corporation, Hrsg. *Couchbase ADO.NET Provider*. URL: https://marketplace.visualstudio.com/items?itemName=CDATASOFTWARE.CouchbaseADONETProvider (besucht am 30. 12. 2018).

[Micd] Microsoft Corporation, Hrsg. *dotConnect ADO.NET Data Provider for SQLite Standard Edition*. URL: https://marketplace.visualstudio.com/items?itemName=DevartSoftware.dotConnectADONETDataProviderforSQLiteStandardEditi (besucht am 30. 12. 2018).

[Mice] Microsoft Corporation, Hrsg. *Elasticsearch ADO.NET Provider*. URL: https://marketplace.visualstudio.com/items?itemName=CDATASOFTWARE.ElasticsearchADONETProvider (besucht am 30. 12. 2018).

[Micf] Microsoft Corporation, Hrsg. *Google BigQuery ADO.NET Provider*. URL: https://marketplace.visualstudio.com/items?itemName=CDATASOFTWARE.GoogleBigQueryADONETProvider (besucht am 30. 12. 2018).

[Micg] Microsoft Corporation, Hrsg. *HBase ADO.NET Provider*. URL: https://marketplace.visualstudio.com/items?itemName=CDATASOFTWARE.HBaseADONETProvider (besucht am 30. 12. 2018).

[Mich] Microsoft Corporation, Hrsg. *Microsoft Access Database Engine 2016 Redistributable*. URL: https://www.microsoft.com/en-us/download/details.aspx?id=54920 (besucht am 30. 12. 2018).

[Mici] Microsoft Corporation, Hrsg. *Redis ODBC Driver*. URL: https://marketplace.visualstudio.com/items?itemName=CDATASOFTWARE.Red (besucht am 30. 12. 2018).

[Mic17] Microsoft Corporation, Hrsg. *XMLA Concepts*. 2017. URL: https://docs.microsoft.com/de-de/sql/analysis-services/multidimensional-models/scripting-language-assl/xmla-concepts?view=sql-server-2014 (besucht am 26. 12. 2018).

[Mic18a] Microsoft Corporation, Hrsg. *ADOMD.NET*. 2018. URL: https://docs.microsoft.com/en-us/bi-reference/adomd/developing-with-adomd-net (besucht am 26. 12. 2018).

[Mic18b] Microsoft Corporation, Hrsg. *Comparing tabular and multidimensional solutions*. 2018. URL: https://docs.microsoft.com/de-de/sql/analysis-services/comparing-tabular-and-multidimensional-solutions-ssas?view=sql-server-2017 (besucht am 16. 12. 2018).

[Mic18c] Microsoft Corporation, Hrsg. *Connect to Azure Cosmos DB using BI analytics tools with the ODBC driver*. 2018. URL: https://docs.microsoft.com/de-de/azure/cosmos-db/odbc-driver (besucht am 30. 12. 2018).

[Mic18d] Microsoft Corporation, Hrsg. *Connect to Azure SQL Data Warehouse: Supported drivers and connection strings*. 2018. URL: https://docs.microsoft.com/fi-fi/azure/sql-data-warehouse/sql-data-warehouse-connect-overview (besucht am 30. 12. 2018).

[Mic18e] Microsoft Corporation, Hrsg. *Data sources supported in SQL Server Analysis Services tabular 1400 models*. 2018. URL: https://docs.microsoft.com/de-de/sql/analysis-services/tabular-models/data-sources-supported-ssas-tabular-1400?view=sql-server-2017 (besucht am 26. 12. 2018).

[Mic18f] Microsoft Corporation, Hrsg. *DirectQuery mode*. 2018. URL: https://docs.microsoft.com/de-de/sql/analysis-services/tabular-models/directquery-mode-ssas-tabular?view=sql-server-2017 (besucht am 26. 12. 2018).

[Mic18g] Microsoft Corporation, Hrsg. *Microsoft OLE DB Simple Provider Overview*. 2018. URL: https://docs.microsoft.com/de-de/sql/ado/guide/appendixes/microsoft-ole-db-simple-provider?view=sql-server-2017 (besucht am 19. 12. 2018).

[Mic18h] Microsoft Corporation, Hrsg. *Supported Data Sources (SSAS - Multidimensional)*. 2018. URL: https://docs.microsoft.com/de-de/sql/analysis-services/multidimensional-models/supported-data-sources-ssas-multidimensional?view=sql-server-2017 (besucht am 27. 12. 2018).

[Min] Minio Inc., Hrsg. *Download*. URL: https://minio.io/downloads.html#download-sdk-dotnet-framework (besucht am 30. 12. 2018).

[MK16] Andreas Meier und Michael Kaufmann. *SQL- & NoSQL-Datenbanken*. 8., überarbeitete und erweiterte Auflage 2016. eXamen.press. Berlin und Heidelberg: Springer Vieweg, 2016. ISBN: 978-3-662-47663-5. DOI: 10.1007/978-3-662-47664-2. URL: http://dx.doi.org/10.1007/978-3-662-47664-2.

[MNK14] Dirk Mertins, Jörg Neumann und Andreas Kühnel. *SQL Server 2014: Das Programmierhandbuch. Inkl. ADO.NET Entity Framework*. 6. Aufl., rev. Ausg. Galileo Computing. Bonn: Galileo Press, 2014. ISBN: 9783836230445.

[Neo] Neo4j, Hrsg. *Chapter 1. Get started: 1.1. About the official drivers*. URL: https://neo4j.com/docs/driver-manual/1.7/get-started/ (besucht am 30. 12. 2018).

[Oraa] Oracle Corporation, Hrsg. *Downloads*. URL: https://www.oracle.com/technetwork/database/options/advanced-analytics/downloads/index.html (besucht am 22. 12. 2018).

[Orab] Oracle Corporation, Hrsg. *Introduction to Oracle Data Mining: 2.1 About Oracle Data Mining*. URL: https://docs.oracle.com/en/database/oracle/oracle-database/18/dmcon/intro-data-mining.html#GUID-7BE45C68-6C87-4E02-A6B3-A52D501B16AD (besucht am 22. 12. 2018).

[Orac] Oracle Corporation, Hrsg. *Introduction to Oracle Data Mining: 2.2 Data Mining in the Database Kernel*. URL: https://docs.oracle.com/en/database/oracle/oracle-database/18/dmcon/intro-data-mining.html#GUID-7BE45C68-6C87-4E02-A6B3-A52D501B16AD (besucht am 22. 12. 2018).

[Orad] Oracle Corporation, Hrsg. *MySQL Connectors*. URL: https://www.mysql.com/de/products/connector/ (besucht am 30. 12. 2018).

[Orae] Oracle Corporation, Hrsg. *Oracle Data Mining: Scalable in-database predictive analytics*. URL: https://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html (besucht am 27. 12. 2018).

[Oraf] Oracle Corporation, Hrsg. *Oracle Data Provider for .NET*. URL: https://www.oracle.com/technetwork/topics/dotnet/index-085163.html (besucht am 26. 12. 2018).

[Pay18] Seth Payne. *Just Released: MongoDB ODBC Driver*. 2018. URL: https://www.mongodb.com/blog/post/odbc-driver-for-the-mongodb-connector-for-business-intelligence (besucht am 30. 12. 2018).

[Pet09] Helge Petersohn. *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur: Zugl.: Leipzig, Univ., Habil, 2004*. München und Wien: Oldenbourg, 2009. ISBN: 978-3-486-57715-0. DOI: 10.1524/9783486593334. URL: http://dx.doi.org/10.1524/9783486593334.

[Piva] Pivotal Software Inc., Hrsg. *DataDirect JDBC Driver for Pivotal Greenplum*. URL: https://gpdb.docs.pivotal.io/550/datadirect/datadirect_jdbc.html (besucht am 30. 12. 2018).

[Pivb] Pivotal Software Inc., Hrsg. *DataDirect ODBC Drivers for Pivotal Greenplum*. URL: https://gpdb.docs.pivotal.io/550/datadirect/datadirect_ODBC_71.html#topic1 (besucht am 30. 12. 2018).

[Proa] Progress Software Corporation, Hrsg. *Powerful DataStax JDBC Driver*. URL: https://www.progress.com/jdbc/datastax-enterprise (besucht am 30. 12. 2018).

[Prob] Progress Software Corporation, Hrsg. *Powerful DataStax ODBC driver*. URL: https://www.progress.com/odbc/datastax-enterprise (besucht am 30. 12. 2018).

[Proc] Progress Software Corporation, Hrsg. *Powerful dBase ODBC driver*. URL: https://www.progress.com/odbc/dbase (besucht am 30. 12. 2018).

[Prod] Progress Software Corporation, Hrsg. *Powerful SAP Sybase ADO.NET driver*. URL: https://www.progress.com/net/sybase (besucht am 30. 12. 2018).

[pru] prudsys AG, Hrsg. *Stay ahead. Make the best decisions with artificial retail intelligence*. URL: https://prudsys.de/wp-content/uploads/prudsys-portfolio-ai-for-retail-intelligent-solutions-pricing-personalization.pdf (besucht am 26. 12. 2018).

[Rapa] RapidMiner Studio, Hrsg. URL: https://docs.rapidminer.com/latest/studio/how-to/nosql/ (besucht am 17. 12. 2018).

[Rapb] RapidMiner Studio, Hrsg. *RapidMiner Studio*. URL: https://rapidminer.com/products/studio/ (besucht am 19. 01. 2019).

[Rapc] RapidMiner Studio, Hrsg. *RapidMiner Studio: Feature List*. URL: https://rapidminer.com/products/studio/feature-list/ (besucht am 17. 12. 2018).

[RGR17] David Reinsel, John Gantz und John Rydning. *Data Age 2025: The Evolution of Data to Live-Critical: Don't Focus on Big Data; Focus in the Data Thats Big*. Hrsg. von IDC. Framingham, 2017. URL: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwi_6KHx_IDfAhVKKlAKHf8PDJYQFjAAegQIABAC&url=https%3A%2F%2Fwww.seagate.com%2Fwww-content%2Four-story%2Ftrends%2Ffiles%2FSeagate-WP-DataAge2025-March-2017.pdf&usg=AOvVaw1xdmo3y6C_WHr_2M9cUxs0 (besucht am 02. 12. 2018).

[Run10] Thomas A. Runkler. *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*. Computational intelligence. Wiesbaden: Vieweg + Teubner, 2010. ISBN: 978-3-8348-0858-5. DOI: 10.1007/978-3-8348-9353-6. URL: http://dx.doi.org/10.1007/978-3-8348-9353-6.

[Sal16] Alexander Salvanos. *Professionell entwickeln mit Java EE 7: Das umfassende Handbuch ; [alle wichtigen APIs, Konzepte und Technologien ; Best Practices für reale Anforderungen geschäftskritischer Software ; inkl. JDBC, Persistence API, Dependency Injection, Java Message Service, Enterprise JavaBeans, Webservices u.v.m.* 2., korrigierter Nachdruck. Rheinwerk Computing. Bonn: Rheinwerk Verlag, 2016. ISBN: 9783836220040.

[SAPa] SAP SE, Hrsg. *Business-Intelligence-Lösungen (BI)*. URL: https://www.sap.com/germany/products/analytics/business-intelligence-bi.html (besucht am 05. 01. 2018).

[SAPb] SAP SE, Hrsg. *Connect to SAP HANA via ODBC*. URL: https://help.sap.com/viewer/0eec0d68141541d1b07893a39944924e/2.0.00/en-US/66a4169b84b2466892e1af9781049836.html (besucht am 30. 12. 2018).

[SAPc] SAP SE, Hrsg. *SAP BW/4HANA – das Echtzeit-Data-Warehouse*. URL: https://www.sap.com/germany/products/bw4hana-data-warehousing.html#why-sap (besucht am 05. 01. 2018).

[SAPd] SAP SE, Hrsg. *SAP HANA Administration Guide: Data Access*. URL: https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.03/en-US/7791e61775f949d9989eafc443158cdb.html (besucht am 23. 01. 2019).

[SAPe] SAP SE, Hrsg. *SAP HANA Smart Data Integration and SAP HANA Smart Data Quality*. URL: https://help.sap.com/viewer/p/HANA_SMART_DATA_INTEGRATION (besucht am 29. 01. 2019).

[SAPf] SAP SE, Hrsg. *SAP HANA: Analyseprozesse*. URL: https://www.sap.com/germany/products/hana/features/advanced-analytics.html (besucht am 20. 01. 2019).

[SAPg] SAP SE, Hrsg. *The SAP HANA Data Provider for Microsoft ADO.NET*. URL: https://help.sap.com/viewer/0eec0d68141541d1b07893a39944924e/2.0.00/en-US/469dee9e6d611014af70d4e9a9cd6b0a.html (besucht am 30. 12. 2018).

[SAP13a] SAP SE, Hrsg. *Connecting Using JDBC*. 2013. URL: http://infocenter.sybase.com/help/index.jsp?topic=/com.sybase.infocenter.dc10083.1601/doc/html/san1282692593610.html (besucht am 30. 12. 2018).

[SAP13b] SAP SE, Hrsg. *Connecting Using ODBC*. 2013. URL: http://infocenter.sybase.com/help/index.jsp?topic=/com.sybase.infocenter.dc10083.1601/doc/html/san1282692597782.html (besucht am 30. 12. 2018).

[SAP13c] SAP SE, Hrsg. *Connecting Using OLE DB*. 2013. URL: http://infocenter.sybase.com/help/index.jsp?topic=/com.sybase.infocenter.dc10083.1601/doc/html/san1282692605063.html (besucht am 30. 12. 2018).

[SASa] SAS Institute Inc., Hrsg. *Data Access Requirements*. URL: https://documentation.sas.com/?docsetId=emag&docsetTarget=p1iz898qpaslysn1pxyfcvpxoc2k.htm&docsetVersion=15.1&locale=de (besucht am 20. 01. 2019).

[SASb] SAS Institute Inc., Hrsg. *File Import Node*. URL: http://documentation.sas.com/?docsetId=emref&docsetTarget=p1rk96oj5sk2tyn1esay58oha0o3.htm&docsetVersion=15.1&locale=de (besucht am 22. 12. 2018).

[SASc] SAS Institute Inc., Hrsg. *Relational Database Sources*. URL: https://documentation.sas.com/?cdcId=bicdc&cdcVersion=9.4&docsetId=bidsag&docsetTarget=p0r68n8gtyzjqen1ddl4r2q3eh4v.htm&locale=de (besucht am 20. 01. 2019).

[SASd] SAS Institute Inc., Hrsg. *XML Data*. URL: https://documentation.sas.com/?cdcId=bicdc&cdcVersion=9.4&docsetId=bidsag&docsetTarget=p1swgjdmcvrbw9n1ozgynxhkz39g.htm&locale=de (besucht am 20. 01. 2019).

[Sch10] Holger Schwichtenberg. *Windows Scripting: Automatisierte Systemadministration mit dem Windows Script Host [5.8] und der Windows PowerShell [2.0] ; [für alle Windows-Versionen (inkl. XP, 2003 R2, Vista, Windows 7 und 2008 R2) ; Visual Basic 6.0, Visual Basic Script 5.8 und Power Shell 2.0 ; über 1000 Praxisbeispiele]*. 6., aktualisierte Aufl. Net.com. München: Addison-Wesley, 2010. ISBN: 9783827329097.

[Sima] Simba Technologies Inc., Hrsg. *Amazon S3 ODBC Driver with SQL Connector*. URL: https://www.simba.com/drivers/amazon-s3-odbc-jdbc/ (besucht am 30. 12. 2018).

[Simb] Simba Technologies Inc., Hrsg. *Cassandra ODBC and JDBC Drivers with SQL Connector*. URL: https://www.simba.com/drivers/cassandra-odbc-jdbc/ (besucht am 30. 12. 2018).

[Simc] Simba Technologies Inc., Hrsg. *HBase ODBC Driver with SQL Connector*. URL: https://www.simba.com/drivers/hbase-odbc-jdbc/ (besucht am 30. 12. 2018).

[Sim12] Simba Technologies Inc., Hrsg. *Simba ODBC Drivers Enable SQL Access to NoSQL Big Data Sources*. 2012. URL: https://www.simba.com/news/simba-odbc-drivers-enable-sql-access-to-nosql-big-data-sources/ (besucht am 24. 01. 2019).

[Sno] Snowflake Computing Inc., Hrsg. *Connecting to Snowflake*. URL: https://docs.snowflake.net/manuals/user-guide-connecting.html (besucht am 30. 12. 2018).

[SQL] SQLite, Hrsg. *SQLite Java: Connect To The SQLite Database Using SQLite JDBC Driver*. URL: http://www.sqlitetutorial.net/sqlite-java/sqlite-jdbc-driver/ (besucht am 30. 12. 2018).

[SSH18] Gunter Saake, Kai-Uwe Sattler und Andreas Heuer. *Datenbanken – Konzepte und Sprachen*. 6., überarbeitete Auflage. mitp Professional. Frechen: MITP, 2018. ISBN: 9783958457768.

[ST19] Alan Said und Vicenç Torra, Hrsg. *Data science in practice*. Bd. volume 46. Studies in big data. Cham, Switzerland: Springer, 2019. ISBN: 978-3-319-97555-9.

[Staa] StatSoft Europe, Hrsg. *Statistica Produktvarianten*. URL: https://www.statsoft.de/de/statistica/statistica-software/ (besucht am 20. 01. 2019).

[Stab] StatSoft Europe, Hrsg. *TIBCO Spotfire*. URL: https://www.statsoft.de/de/statistica/tibco-spotfire/ (besucht am 20. 01. 2019).

[Ste17] René Steiner. *Grundkurs Relationale Datenbanken: Einführung in die Praxis der Datenbankentwicklung für Ausbildung, Studium und IT-Beruf*. 9., erweiterte und aktualisierte Auflage. Lehrbuch. Wiesbaden: Springer Vieweg, 2017. ISBN: 978-3-658-17978-6. DOI: 10.1007/978-3-658-17979-3. URL: http://dx.doi.org/10.1007/978-3-658-17979-3.

[Stu16] Thomas Studer. *Relationale Datenbanken: Von den theoretischen Grundlagen zu Anwendungen mit PostgreSQL*. 1. Aufl. 2016. eXamen.press. Berlin und Heidelberg: Springer Vieweg, 2016. ISBN: 978-3-662-46570-7. DOI: 10.1007/978-3-662-46571-4. URL: http://dx.doi.org/10.1007/978-3-662-46571-4.

[Tera] Teradata Corporation, Hrsg. *Teradata Vantage, the platform for Pervasive Data Intelligence: Software Components*. URL: https://www.teradata.com/Products/Software/Vantage/Components (besucht am 29. 12. 2018).

[Terb] Teradata Corporation, Hrsg. *Teradata Vantage: The Platform for Pervasive Data Intelligence*. URL: http://assets.teradata.com/resourceCenter/downloads/Datasheets/EB9959v2.pdf (besucht am 29. 12. 2018).

[Ter15] Teradata Corporation, Hrsg. *Find downloads for connecting to the Teradata ecosystem.* 2015. URL: http://downloads.teradata.com/download/connectivity (besucht am 30. 12. 2018).

[Thea] The Npgsql Development Team, Hrsg. *Npgsql - .NET Access to PostgreSQL*. URL: https://www.npgsql.org/ (besucht am 30. 12. 2018).

[Theb] The PostgreSQL Global Development Group, Hrsg. *PostgreSQL JDBC Driver 42.2.5 Released*. URL: https://jdbc.postgresql.org/ (besucht am 30. 12. 2018).

[Thec] The PostgreSQL Global Development Group, Hrsg. *PostgreSQL Native OLEDB Provider (PGNP) 1.3.0 32/64-bit released!* URL: https://www.postgresql.org/about/news/1153/ (besucht am 30. 12. 2018).

[The13] The PostgreSQL Global Development Group, Hrsg. *psqlODBC - PostgreSQL ODBC driver*. 2013. URL: https://odbc.postgresql.org/ (besucht am 30. 12. 2018).

[TIBa] TIBCO Software Inc., Hrsg. *Spotfire Connectors 10.0*. URL: https://docs.tibco.com/pub/spotfire/general/sr/GUID-6B7619DB-FD61-4E02-A020-ADC17A7B670A.html (besucht am 20. 01. 2019).

[TIBb] TIBCO Software Inc., Hrsg. *TIBCO Statistica is Now Part of TIBCO Data Science*. URL: https://www.tibco.com/products/data-science/statistica-now-part-tibco-data-science (besucht am 19. 12. 2018).

[TSC15] TSC Germany, Hrsg. *IBM SPSS Modeler 17.1: Quellen-, Prozess und Ausgabeknoten*. 2015. URL: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.1/de/ModelerSPOnodes.pdf.

[Visa] Viscovery Software GmbH, Hrsg. *Viscovery SOMine 7 - Data Sheet: Extension Module - Enterprise Data*. URL: https://www.viscovery.net/download/public/Data_Sheet_Viscovery_SOMine_-_Enterprise_Data.pdf (besucht am 25. 12. 2018).

[Visb] Viscovery Software GmbH, Hrsg. *Viscovery SOMine: User's Manual*. (Besucht am 25. 12. 2018).

[Wes+14] Markus Weskamp u. a. *Studie: Einsatz und Nutzenpotentiale von Data Mining in Produktionsunternehmen: Ergebnisse*. Hrsg. von Frauenhofer-Institut für Produktionstechnik und Automatisierung. Stuttgart, 2014. URL: `http://publica.fraunhofer.de/starweb/pub09/servlet.starweb` (besucht am 17. 12. 2018).

[Wie15] Lena Wiese. *Advanced Data Management*. s.l.: De Gruyter, 2015. ISBN: 978-3-11-044140-6. URL: `http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1107018`.