DIPLOMA
PRIVATE STAATLICH ANERKANNTE HOCHSCHULE
University of Applied Sciences

Hartwig

Big Data

Studienheft Nr. 986

1. Auflage 03/2020

Verfasser

Dr. Michael Hartwig (Dipl.-Informatiker, FH)
Senior Technical Consultant (Intershop Communication AG)

Diploma Hochschule
University of Applied Sciences
Am Hegeberg 2
37242 Bad Sooden-Allendorf
Tel. 05652/587770, Fax 05652/5877729

Hinweise zur Arbeit mit diesem Studienheft

Der **Inhalt** dieses Studienheftes unterscheidet sich von einem Lehrbuch, da er **speziell für das Selbststudium aufgearbeitet** ist.

In der Regel beginnt die Bearbeitung mit einer Information über den Inhalt des Lehrstoffes. Diese Auskunft gibt Ihnen das **Inhaltsverzeichnis**.

Beim Erschließen neuer Inhalte finden Sie meist Begriffe, die Ihnen bisher unbekannt sind. Die **wichtigsten Fachbegriffe** werden Ihnen übersichtlich in einem dem Inhaltsverzeichnis nachgestellten **Glossar** erläutert.

Den einzelnen Kapiteln sind **Lernziele** vorangestellt. Sie dienen als Orientierungshilfe und ermöglichen Ihnen die Überprüfung Ihrer Lernerfolge. Setzen Sie sich **aktiv** mit dem Text auseinander, indem Sie sich Wichtiges mit farbigen Stiften kennzeichnen. Betrachten Sie dieses Studienheft nicht als "schönes Buch", das nicht verändert werden darf. Es ist ein **Arbeitsheft**, **mit** und **in** dem Sie arbeiten sollen.

Zur **besseren Orientierung** haben wir Merksätze bzw. besonders wichtige Aussagen durch Fettdruck und/oder Einzug hervorgehoben.

Lassen Sie sich nicht beunruhigen, wenn Sie Sachverhalte finden, die zunächst noch unverständlich für Sie sind. Diese Probleme sind bei der ersten Begegnung mit neuem Stoff ganz normal.

Nach jedem größeren Lernabschnitt haben wir Übungsaufgaben eingearbeitet, die mit **„SK = Selbstkontrolle"** gekennzeichnet sind. Sie sollen der Vertiefung und Festigung der Lerninhalte dienen. Versuchen Sie, die ersten Aufgaben zu lösen und die Fragen zu beantworten. Dabei werden Sie teilweise feststellen, dass das dazu erforderliche Wissen nach dem ersten Durcharbeiten des Lehrstoffes noch nicht vorhanden ist. Gehen Sie diesen Inhalten noch einmal nach, d. h. durchsuchen Sie die Seiten gezielt nach den erforderlichen Informationen.

Bereits während der Bearbeitung einer Frage sollten Sie die eigene Antwort schriftlich festhalten. Erst nach der vollständigen Beantwortung **vergleichen Sie Ihre Lösung mit dem** am Ende des Studienheftes **angegebenen Lösungsangebot.**

Stellen Sie dabei fest, dass Ihre eigene Antwort unvollständig oder falsch ist, müssen Sie sich nochmals um die Aufgabe bemühen. Versuchen Sie, jedes behandelte Thema vollständig zu verstehen. **Es bringt nichts, Wissenslücken durch Umblättern zu übergehen.** In vielen Studienfächern baut der spätere Stoff auf vorhergehendem auf. Kleine Lücken in den Grundlagen verursachen deshalb große Lücken in den Anwendungen.

Zudem enthält jedes Studienheft **Literaturhinweise.** Sie sollten diese Hinweise als ergänzende und vertiefende Literatur bei Bedarf zur Auseinandersetzung mit der jeweiligen Thematik betrachten. Finden Sie auch nach intensivem Durcharbeiten keine zufriedenstellenden Antworten auf Ihre Fragen, **geben Sie nicht auf. Wenden Sie sich** in diesen Fällen schriftlich oder fernmündlich **an uns**. Wir stehen Ihnen mit Ratschlägen und fachlicher Anleitung gern zur Seite.

Wenn Sie **ohne Zeitdruck** studieren, sind Ihre Erfolge größer. Lassen Sie sich also nicht unter Zeitdruck setzen. **Pausen** sind wichtig für Ihren Lernfortschritt. Kein Mensch ist in der Lage,

stundenlang ohne Pause konzentriert zu arbeiten. Machen Sie also Pausen: Es kann eine kurze Pause mit einer Tasse Kaffee sein, eventuell aber auch ein Spaziergang an der frischen Luft, sodass Sie wieder etwas Abstand zu den Studienthemen gewinnen können.

Abschließend noch ein formaler Hinweis: Sofern in diesem Studienheft bei Professionsbezeichnungen und/oder Adressierungen aus Gründen der besseren Lesbarkeit ausschließlich die männliche Form Verwendung findet (z. B. „Rezipienten“), sind dennoch alle sozialen Geschlechter, wenn kontextuell nicht anders gekennzeichnet, gemeint.

Wir wünschen Ihnen viel Erfolg bei der Bearbeitung dieses Studienheftes.

Ihre

DIPLOMA
Private Hochschulgesellschaft mbH

Inhaltsverzeichnis Seite

Glossar

Fachtermini	Erläuterungen
RAM	Random Access Memory, flüchtiger Arbeitsspeicher im Prozessor
RDBMS	Relationales Datenbank Management System
Data Warehouse, DWH	Datenanalysesystem mit einer Betonung auf einer mehrdimensionalen Datensicht
ETL	Extraction-Transfer-Load, Datenaufbereitungs- und Einspeicherungsprozesse für ein DWH
Data Lake	Datenspeicherungssystem für Dateien und Daten verschiedenster Formate, nutzt normale Datei-und Ordnerstrukturen
3V, 4V, 7V	Big-Data-Definitionen
Datafizierung	der Drang intensiv Daten zu speichern und analysieren zu wollen
Persistenz	eine nichtflüchtige, dauerhafte Speicherung von Daten
Schema-on-Read	das Verändern von Daten nur während und zum Zwecke der Analyse
Schema-on-Write	das Verändern von Daten vor Analyseverfahren
(Near) Real-time Analytics	Analyseverfahren mit äußerst schnellem Antwortzeitverhalten
Predictive Analytics	Vorhersageanalyseverfahren
Explorative Analytics	Analyseverfahren zur Erklärung und Verbesserung von Prozessen
Korrelationsanalyse	Analyseverfahren zur Bestimmung des Zusammenhangs gemeinsamer Erscheinungen
Hybride Infrastruktur	Datenspeicher und -analyseinfrastruktur, die auf eine vielfältige Kombination von operativen und analytischen Datenbanken setzt
Skalierbarkeit	die Eigenschaft, ein System leicht auch mit höheren Anforderungen arbeiten zu lassen
Latency, Latenz	Zeitverlust bei Zugriffen auf Daten
Cloud Computing	virtuelle Computing-Ressourcen im Internet
Virtualisierung	Abbilden von Ressourcen (v. a. ganze Betriebssysteme) über Software
Batch/Stream-Verarbeitung	Verarbeitungsmodalitäten für Daten mit konstanter oder ohne Zwischenspeicherung
Immutable Datasets	Datenstrukturen, die nach einer Initialisierung nicht mehr verändert werden
Lazy Evaluation	Auswertungsprinzip mit spätestmöglicher Berechnung

Big Data

Data Preparation	Datenaufbereitungstätigkeiten vor einer Verwendung in Analyseschritten
Data Cleaning	Datenbereinigung um fehlerhafte Daten, findet während des Data Preparation statt
Feature Selection	Reduktion der Daten auf zur Analyse notwendige direkte und indirekte Attribute
Data Munging	Prozess des Findens von Datenquellen und der Feature Selection
Datenreduktion	Wegnahme nicht zur Analyse beitragender Datensätze
Data Sink	Zwischenspeicher (analog zu einem Data Lake) während des Analyseprozesses
Processing-to-Data	Datenauswertung, welches das Auswertungsprogramm physisch zum Ort der Daten bringt
Data-to-Processing	Datenauswertung, welches die Daten physisch zum Ort des Auswertungsprogrammes bringt
Hadoop	Analyseplattform für Daten aus einem Data Lake mit hoher Skalierbarkeit
HDFS	Hadoop Distributed File System
Yarn	(Default-)Aufgabenmanager im Hadoop
Pig, Hive, Spark, Spark SQL	Analysesysteme zur Programmierung von Analyseprozessen im Hadoop
Zeppelin	Web-Notebook für die Präsentation von Daten
MapReduce	Abstraktes Datenauswertungsprinzip im Hadoop
Mittelwertanalysen	Analyseverfahren zur Berechnung verschiedenster Mittelwerte
Assoziationsanalyse	Analyseverfahren zur Ermittlung von passenden Eigenschaftskombinationen
Entscheidungsbaum	Analyseverfahren zur Klassifizierung
Random Forest Algorithm	Analyseverfahren zur Klassifizierung
Vier-Felder-Tafel, Confusion Matrix	Hilfsmittel zur Auswertung der Korrektheit von Klassifizierungsverfahren
Maschinelles Lernen	Bereich der Informatik, der sich mit der Vorhersage, Klassifikation, Wissensgenerierung aus Daten befasst

1 Aus Daten kann man lernen, aus vielen Daten kann man viel lernen (Einleitung)

> *Data is power. It differentiates and becomes the basis for new products, sales and customer relationships. A company's 'optimal exploitation of data' is key, but more importantly, that exploitation drives revenue (Zimmermann, 2011, in KING2014).*

Bevor wir uns mit dem Begriff Big Data auseinandersetzen, sollen einige Einführungsbeispiele auf die Möglichkeiten und Probleme der Nutzung gewaltiger Datenmengen hinweisen. In (SCHÖNB2017) findet sich ein schon als klassisch einzustufendes und bekannt gewordenes Beispiel. Hierbei geht es um die Frage nach der Ausbreitung von Grippewellen. Durch intensive Nutzung von Flugreisen und Mutationen der Krankheitserreger, die eine schnelle Ansteckung erwirken, gelangen Erreger immer schneller in ferne Ländern und eine Ausbreitung von Krankheiten bedarf äußerst raschen Handelns.

Bisher haben die WHO und auch daie US-amerikanischen Centers for Disease Control and Prevention (CDC) dazu Aufzeichnungen von Ärzten aus vielen Ländern ausgewertet. Wer soll es besser wissen als Ärzte? Diese behandeln die Patienten, verschreiben Medikamente und sehen ihre Patienten. Tatsächlich lassen sich auf dieser Basis gute Rückschlüsse auf Krankheiten und ihre Ausbreitungen schließen. Nur, wie schnell? Zunächst muss dem Arzt ein außergewöhnlicher Anstieg von Krankheiten erst einmal auffallen. Dann muss er seine Akten sauber führen und die entsprechenden Informationen auch weiterleiten. Auch wenn die WHO sich über viele Jahre ein gutes Netz an Amtsärzten und Informanten in allen Teilen der Welt aufgebaut hat, eine gewisse Verzögerung lässt sich nicht vermeiden.

Im Jahre 2009 stellten sich einige Ingenieure von Google dann die Frage, ob sich solche Krankheiten nicht auch in den Abfragen an Google zeigen könnten. In allen Teilen der Welt nutzen Menschen ihre Webseite, um im Internet nach Informationen zu suchen. Es liegt nahe, anzunehmen, dass auch kranke Menschen oder ihre Angehörigen sich über Krankheiten und deren Anzeichen informieren. Die Softwareentwickler suchten daher gezielt nach solchen Anfragen und deren Häufigkeitsverteilungen. Dabei probierten sie viele Modelle aus, bis sie schließlich ein Modell fanden, welches die Grippewellen von 2003 bis 2008 tatsächlich exakt spiegelten. Die Genauigkeit des Ergebnisses überraschte. Mittlerweile nutzen daher auch die eben genannten CDC in ihren Vorhersagen und Analysen sowohl lokale Ärzte als auch Big-Data-Verfahren.

Ebenso hat das folgende Beispiel zur Popularisierung der Nutzung von Big Data beigetragen. Das Nationale Institut für Sturmwarnung der USA wertet seit vielen Jahren Daten aus Wetterstationen aus, um Warnungen über herannahende Stürme auszusprechen. Walmart, eine große Supermarktkette in den USA, nutzt diese Hinweise. Zum einen müssen bestimmte Lebensmittel für bald betroffene Gebiete bereitgestellt werden, zum anderen die Geschäfte der Kette in den betroffenen Gebieten rechtzeitig geschlossen und gegebenenfalls sturmsicher gemacht werden. Walmart speichert seit vielen Jahren die Einkäufe seiner Kunden und versucht damit, ihnen ein immer besseres (und für Walmart profitableres) Angebot an Waren und Preisen zu offerieren. Auch hier haben sich einige Mitarbeiter frühzeitig dafür interessiert, ob sich nicht die herannahenden Stürme auch im Einkaufsverhalten der Bürger zeigen könnten. Tatsächlich ist dem so. Während es nicht in jedem Ort der USA eine entsprechende Wetterstation gibt und Wetter nach wie vor ein komplexes Phänomen mit kurzfristigen, abrupten Veränderungen ist, gibt es aber so gut wie überall in den USA einen Walmart Supermarkt. Und ja, bei ersten Anzeichen von Sturm gerade in entfernten Gebieten greifen die Menschen zu anderen Lebensmitteln und gehen notwendige Reparaturen am Haus und ihren Schutzräumen an.

Nun kann man sich fragen: Was unterscheidet die Analyse der eingeschickten Daten der Ärzte von einer Analyse aller Google-Anfragen? Oder, bezogen auf das zweite Beispiel: Was unterscheidet eine Analyse der Daten der Wetterstationen von einer Analyse der Lebensmitteleinkäufe?

- Die Lebensmitteleinkäufe und die Google-Suchen waren als Daten bereits vorhanden.
- Diese Daten wurden allerdings nicht mit dem Ziel gespeichert, Krankheiten oder Stürme vorherzusagen.

- Die Lebensmitteleinkäufe und Google-Anfragen mussten zunächst geschickt bearbeitet und gefiltert werden.
- Insofern sich eine kleine Anzahl von Amtsärzten für eine Epidemie ausspricht, ist dies bereits ein ernst zu nehmender Indikator für eine neue Krankheitswelle. Nur weil vier Schüler im gleichen Land während eines Schulprojektes vermehrt im Internet nach der gleichen Krankheit suchen, spricht dies noch lange nicht für eine Epidemie. Die Aussagekraft der Analyse der Google-Anfragen und Lebensmitteleinkäufe hängt also stark davon ab, ob enorm viele (und korrekte, „saubere" und aussagekräftige) Daten vorliegen.

In einem weiteren Beispiel demonstrieren die Autoren (SCHÖNB2017), was die Konsequenzen einer Zunahme von Daten bewirken kann. Es entsteht eine neue Dimensionserkenntnis.

> Peter Norvig, Experte für künstliche Intelligenz bei Google, zieht zum Vergleich gerne Bilder heran. „Stellen Sie sich als Erstes die bekannte Abbildung eines Pferdes aus den altsteinzeitlichen Höhlenmalereien von Lascaux in Frankreich vor, die etwa 17.000 Jahre alt sind. Dann denken Sie an die Fotografie eines Pferdes – oder besser noch, an Pablo Picassos Farbtupfer, die den Höhlenmalereien gar nicht so unähnlich sind. Als Picasso die Zeichnungen von Lascaux sah, soll er bemerkt haben: ‚Wir haben seitdem nichts Neues erfunden.'
> Einerseits hat Picasso recht, andererseits irrt er aber. Denn es dauert sehr lange, ein Pferd zu zeichnen; fotografiert ist es sehr viel schneller. Dies stellt zwar eine Veränderung dar, aber noch keine grundlegende, da das Ergebnis immer noch dasselbe ist: die Abbildung eines Pferdes. Stellen Sie sich dagegen 24 Bilder eines Pferdes pro Sekunde vor. Hier führt die quantitative zu einer qualitativen Veränderung, denn ein Film ist etwas grundlegend anderes als ein fotografisches Standbild. So ist es auch mit Big Data: Indem wir die Menge verändern, verändern wir das Wesen der Aufzeichnung."

Merke: **Big Data erlauben Schlussfolgerungen aus Daten, die nur aufgrund des gro-ßen Datenvolumens zu erklären sind. Es sind insbesondere Schlussfolgerun-gen, die oft gar nichts mit dem Grund der Datenerhebung zu tun haben. Be-reits kleinste Einflussparameter erwirken dennoch eine statistische Signifi-kanz, die wir aufspüren wollen. Zusammen mit sich weiter verbilligendem Speicherplatz berechtigt dies eine zentrale Forderung aller wirtschaftlichen Akteure, Daten auf Vorrat zu speichern. Im Zweifelsfalle ob noch unklarer Auswertungen sollten infrage kommende Daten zunächst immer erst einmal gespeichert werden.**

Die derzeitigen technologischen Möglichkeiten sprechen daher klar für eine gewisse „Sammel-Taktik": Selbst, wenn noch keine klaren Auswertungs- und Analyseziele definiert sind, sollten bestehende Daten zunächst aufbewahrt werden. Sowohl ein Anbinden an Data Warehouses, ein Umformatieren und Umwandeln, als auch ein Verwenden unter neuen Gesichtspunkten ist relativ leicht und einfach möglich.

All dies entbindet aber nicht von der notwendigen Beachtung gesetzlicher Vorschriften. Wie auch bereits im Heft Data Warehousing klar herausgearbeitet: Gesetzliche Vorgaben zur Einhaltung des Datenschutzes und dem gewissenhaften Umgang mit personenbezogenen Daten sind einzuhalten! Dies trifft nicht nur auf Datenanalysen zu, sondern beginnt bereits mit der Speicherung von Daten, auch wenn Analysen noch gar nicht vorgenommen wurden.

1.1 Big Data sind überall

Einige weitere Beispiele sollen die enorme Bandbreite der Anwendungsmöglichkeiten der Datenanalyse (SCHÖNB2017) aufzeigen:

- Biologie: Mittlerweile kann sich fast jeder eine persönliche Analyse seiner DNA erlauben. Diese sind zumeist auf Abschnitte und Marker beschränkt. Steve Jobs war einer der Ersten, der eine komplette Analyse seiner ganzen DNA hat vornehmen lassen.

- Finanzwirtschaft/Versicherung: Xoom untersucht riesige Mengen von Finanztransaktionen und entdeckte so 2011 Anstrengungen einer Bande. Jede einzelne Transaktion schien völlig legal, erst die Summe der Transaktionen ergab ein ungewöhnliches Muster.

- Sport: Steven Levitt untersuchte fast alle Sumo-Wettkämpfe der letzten Jahre und fand heraus: Es gab ausgehandelte, korrumpierte Wettkämpfe. Glücklicherweise erfolgte das nicht bei den hoch angesehenen Meisterschaften, sondern eher unbedeutenderen Punktekämpfen am Ende der Turniere. (Was Levitt natürlich verwunderte, bevor er den Grund dafür herausfand.)

- Transportoptimierung: UPS reduzierte 2011 nach Analyse aller durchgeführten Fahrten seine Routen um 20 Mio. km und mehr als 12 Mio. Liter Kraftstoff.

- Sozialwissenschaften: Albert-László Barabási kam zu vielen neuen Erkenntnissen über das Verhalten menschlicher Gruppen nach Analyse von Abertausenden von Freundschaftsbeziehungen. So erkannte er, dass Menschen „an den Rändern“ einer festen Gruppe mit einer Beziehung zu Außenseitern tatsächlich eine größere Rolle für die Stabilität der Gruppe spielen als zentrale Mitglieder. Das war eine wichtige Erkenntnis für Werbung und Wahlen!

1.2 Big Data sind wirklich groß

Dass es sich bei solchen Anwendungen tatsächlich um die Verarbeitung von fast unvorstellbar groß gewordenen Datenmengen handeln kann, soll hier anhand einiger Beispiele aufgezeigt werden. Diese sind u. a. erneut (SCHÖNB2017) entnommen.

- Allein Google sammelt wohl pro Tag 24 Petabyte an Daten, ungefähr tausendmal so viel wie alle gedruckten Werke in der US-Kongressbibliothek zusammen. Pro Tag!

- Facebook, ein Unternehmen, das es vor dem jahr 2004 noch gar nicht gab, erhält pro Stunde über zehn Millionen neue Fotos. Facebook-Nutzer geben pro Tag etwa drei Milliarden Kommentare oder „Gefällt-mir“-Klicks ab; die digitale Spur, die sie so hinterlassen, kann der Konzern auswerten, um die Vorlieben der einzelnen Kunden zu erfassen.

- Die 1,9 Milliarden monatlichen Nutzer (Stand 2020) des Google-Videodienstes YouTube laden pro Sekunde eine Stunde Videos hoch.

- Die Anzahl der Twitter-Kurznachrichten wächst (noch) jährlich und liegt 2020 bei über 500 Millionen Tweets pro Tag.

Gerade im Bereich der Unternehmensdaten geht man von Mengensteigerung um bis zu 650 % oder gar 1000 % in den nächsten Jahren aus (ZEUS2012 und Gantz, Reisel, in KING2014). So verwundert es wirklich nicht, dass die bereits seit vielen Jahren geläufigen Größenbezeichnungen für Daten (Bit, Byte, Kilobyte, Megabyte, Gigabyte, Terabyte, Petabyte, Exabyte, Zettabyte, Yottabyte) in naher Zukunft um neue Bezeichner erweitert werden. Einige der eingereichten Vorschläge sind Brontobyte, Xonabyte,

Ronnabyte (alle für 1024 Yottabyte) und Quecca (dann für 1024 Ronnabyte). Die zuletzt genannten scheinen die aussichtsreichsten Kandidaten für eine feste Normierung zu sein (SCHMITT2019).

(JODLB2018) stellt der Zunahme der Bevölkerung (durchschnittlich etwa 1,5 % pro Jahr) die jährliche Zuwachsrate von Büchern (bisher etwa 10 %) und die Zuwachsrate der neu generierten Daten entgegen (geschätzt etwa 50 %). Dieser enormen Explosion der Menge vorhandener Daten stehen laut Experten gerade einmal drei Prozent (geschätzt) gegenüber, die den Anteil der tatsächlich intensiv genutzten und analysierten Daten darstellen. Aber dies wird sich, auch aufgrund Ihrer Mithilfe nach dem Ende Ihres Studiums, sicherlich ändern.

Ein letzter Hinweis sei vor den folgenden Kapiteln gestattet. Die Analyse großer Datenmengen zeichnet sich durch eine Vielzahl an verwendeten Technologien aus. Sie werden also in diesem Heft auf sehr viele Begriffe, Verfahren und Tools stoßen. Es ist bekannt, dass dies den Einstieg in das Lernen des Themas erschwert. Wir haben daher versucht, das daher recht umfangreiche Lernmaterial um kleine humorvolle Bemerkungen und Geschichten zu erweitern. Wir würden uns freuen, wenn Ihnen diese das Lernen erleichtern. Sie sollten diese Vorgehensweise aber nicht als Vorlage für Ihre spätere Bachelor- oder Masterarbeit verwenden.

Übungsaufgaben zur Selbstkontrolle SK

1. Suchen Sie in der Literatur oder im Internet nach weiteren Beispielen für Big-Data-Anwendungen. Suchen Sie gezielt nach Beispielen, wo die Datenerhebung eigentlich aus einem anderen Grund als der darauffolgenden Analyse erfolgte.

2. Angenommen sei ein Besucher eines Webshops.

a. Nennen Sie interessante, anfallende und zu speichernde Daten über den Verlauf seines Besuches.
b. Geben Sie eine Größenordnung dieser Daten an und reflektieren Sie darüber.

2 Big Data: Chancen und Risiken (Definitionen)

Im zweiten Kapitel dieses Lehrheftes werden wir zunächst den Begriff Big Data genau charakterisieren. Daran anschließend wollen wir uns Analysemöglichkeiten und -varianten anschauen, die auf diesen Daten basieren. Das Kapitel schließt mit Bemerkungen zu Chancen, aber auch Risiken der Nutzung solcher Daten und Analyseverfahren ab.

Lernziele dieses Abschnitts:

Sie sollten daher am Ende des Kapitels in der Lage sein,

- Big Data als Begriff klar zu definieren,
- Big Data von anderen Technologien zur Speicherung und Analyse von Daten wie relationalen Datenbanken, Data-Warehousing-Systemen abzugrenzen,
- Möglichkeiten der Datenanalyse zu klassifizieren und erste Beispiele zu nennen und
- Risiken zu nennen, die mit der Speicherung und Analyse solcher Daten einhergehen.

2.1 Zum Begriff Big Data

Ursprünglich wurde unter dem Begriff Big Data einmal jede Datenmenge verstanden, die zu groß für den sie verarbeitenden Computer war (SCHÖNB2017, Dumbill in KING2014). Würde man darunter den RAM des Computers verstehen, dann müsste man mittlerweile die meisten Datenbanken ebenso hier einordnen. Dieser erste Ansatz gab zumindest den Start für die Entwicklung von Technologien wie Hadoop and MapReduce, die für die Verarbeitung von Daten aus vielen Dateiquellen entwickelt wurden und damit umgehen konnten, dass die Daten von fernen Datenquellen konstant geholt und wieder dort abgelegt werden mussten. Inzwischen betont man neben der Größe der Daten vor allem die Möglichkeit der raschen Verarbeitung. Diese Eigenschaft ist für die Datenanalyse oft wichtiger als die reine Größe.

> Big data enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time to gain the right insights. The key to understanding big data is that data has to be managed so that it can meet the business requirement a given solution is designed to support (HURWITZ2013).

Gerade die Möglichkeit der zeitnahen Verarbeitung der Daten gibt Big Data seine Berechtigung. Firmen sind so in der Lage, schnell in den Daten Muster zu finden, die vielleicht auf

- sich abzeichnende Veränderungen im Kunden- oder Kaufverhalten,
- Veränderungen im Verhalten von Maschinen oder deren Bauteilen,
- bereits kleinste Anzeichen von Auswirkungen auf menschliche Organe nach der Einnahme von Medikamenten oder
- den Einsatz einer höchstwahrscheinlich gerade gestohlenen Kreditkarte

hindeuten. Hier spielt das schnelle Reagieren auf solche Ereignisse eine große Rolle, da sich daraus große wirtschaftliche oder gesundheitliche Konsequenzen ergeben können. Das letzte Beispiel verdeutlich dies: Ein Erkennen des Kreditkartenmissbrauchs am Folgetag ist nicht ausreichend, denn dann wurde die Karte sicherlich bereits mehrfach eingesetzt. Der Diebstahl und damit fälschliche Einsatz müssen sofort, beim ersten Missbrauch, erkannt werden.

Dabei greifen Big Data auf wichtige technologische Errungenschaften zurück. Hierzu zählen wir vor allem

- große Fortschritte in der Speichertechnik,
- enorme Kostensenkungen in den Bereichen Speicher und Hardware allgemein,

- enorme Verbesserungen in der Datenübertragung,
- die Entwicklung einfacher Virtualisierungstechnologien,
- das Entstehen von kostengünstigen Cloudangeboten.

Auch in der Modellierung, Haltung und Verarbeitung von Daten haben sich technologische Veränderungen ergeben. Dabei zeichnen sich klar einige Wellen (oder Umbrüche) ab.

- Zu Beginn der Computerisierung erfolgte die Speicherung von Daten in **einfachen zeilenorientierten Textdateien**. Dies bedingte die typisch lineare Volltextsuche (engl. brute force) für kleinste Analysen und Änderungen.

- Mit Beginn der 1970er-Jahre setzte sich dann das relationale Datenbankmodell durch. Für die **operationale Datenhaltung** ist es, aufgrund solcher Technologien wie der referentiellen Integrität und der Normalisierung nach wie vor der Standard. Hiermit ist das Speichern und Verändern von Daten (bspw. für Lagerhaltungsinformationssysteme, Buchungsverwaltungen, Einkaufsorganisation) sowie deren Darstellung in Berichten gut möglich und entsprach exakt den wirtschaftlichen Anforderungen der damaligen Zeit.

- Mit Beginn der 1990er-Jahre begann sich das wirtschaftliche Umfeld zu verändern. Datenanalyse und Ursachenforschung ergänzten nun die operationale Datenhaltung. Dazu mussten vor allem **analytische Operationen** schnell durchführbar sein. **Data-Warehousing**-Systeme mit ihren ETL-Prozessen zur Anbindung weiterer Datenquellen und -formate sowie Denormalisierungsverfahren zur schnellen Bereitstellung von verschiedenen Sichten auf die Daten füllten diese Lücke.

In einem weiteren Zyklus erkannten erste Firmen das wirtschaftliche Potenzial in **nicht-strukturierten** Daten. **Strukturierte** Daten folgen einem klaren Schema, welches sich mit jeder neuen Einheit (oder auch Entität) wiederholt. Nicht-strukturierte Daten lassen sich nicht in immer wieder gleichartige Einheiten zerlegen. Beispielsweise sind die folgenden Daten nicht-strukturierter Natur:

- Videos,
- Musikdateien,
- Facebook-Kommentare,
- Bilder.

Während es bei medialen Daten schnell einsichtig ist, wollen wir uns kurz die eben erwähnten Facebook-Kommentare anschauen. Auch sie sind grundsätzlich unstrukturierter Natur. Ein einzelner (unbeantworteter) Kommentar kann als String angesehen werden. Ein Kommentar mit medialem Anhang ist eher ein Feld aus Einzelelementen. Ein Kommentar, der Reaktionen anderer Nutzer hervorgerufen hat, ist dagegen ein komplexes Netz aus Antworten und erneuten Reaktionen.

Wir bezeichnen daher im Folgenden:

Definition:

1. **Strukturierte Daten** stellen speicherbare Informationen dar, die sich in kleinere Einheiten zerlegen lassen. Diese kleineren Einheiten folgen einem klaren Aufbau.
2. **Unstrukturierte Daten** bezeichnen speicherbare Informationen, die sich ebenfalls in kleinere Einheiten zerlegbar sind. Diese kleineren Einheiten folgen jedoch **keinem** klaren Aufbau.
3. Strukturierte und nicht-strukturierte Daten können beide ein wirtschaftlich interessantes Potenzial haben. Eine Grundvoraussetzung hierbei ist, dass sie sich **Zielgruppen zuordnen** lassen.

Die gerade dargestellte Definition erlaubt auch erste Erkenntnisse darüber, warum Facebook-Daten einen solch großen wirtschaftlichen Nutzen darstellen (Salopp gesprochen, warum also Mark Zuckerberg

so reich geworden ist.): Die Daten sind sofort und extrem einfach Personen zuordenbar. Hier mussten gar keine Annahmen gemacht und keine kostspieligen Analysen vorgenommen werden. Jeder Kommentar ist eindeutig einem Nutzer zuordenbar.

> *"The issue with data, particularly personal data, is this: context is everything" (LUDL2012, KING2014).*

Mit dem Aufkommen von Cloud-Speichern und der einfachen, billigen Speicherung solcher strukturierten und unstrukturierten Daten (beispielsweise in sogenannten BLOB-Speichern) kam es zu einer nochmaligen Potenzierung von Datenmengen. Experten schätzen, dass nur etwa 20 % der vorhandenen Daten strukturierter Natur sind (HURWITZ2017, ZEUS2012). Die kostenintensive (insbesondere zeitintensive) Umwandlung aller Daten in eine feste, strukturierte Form ist dann nicht immer sinnvoll, da sie ad hoc und schnell (Schlagwort agil) für Analysen verwendet werden sollen. Dies wurde dann durch Verfahren wie MapReduce und Hadoop ermöglicht.

Definition:
Wir sprechen von einem **Data Lake**, wenn Daten vieler Datenquellen, die zu einer Verarbeitung herangezogen werden, möglichst in ihrer originalen Form, das heißt ohne Transformationsprozesse, zusammen gespeichert werden.

Merke: **Die Menge an unstrukturierten Daten übertrifft die Menge an strukturierten Daten um ein Vielfaches. Seine Verwendung in analytischen Verfahren er-fordert neue, agile Verarbeitungsprinzipien. Ein Umwandeln in strukturierte, feste Formate ist nicht immer zielführend.**

Fortschritte in der

Speichertechnik,
Netzwerktechnik und
Virtualisierung

machten die gewünschte agile Nutzung solcher großen Datenmengen möglich.

Bei der nun folgenden Definition des Begriffes Big Data wollen wir die englischen Bezeichnungen voranstellen, da sie recht einprägsam sind.

Definition:
Wir sprechen allgemein von der Nutzung von **Big-Data**-Technologien, wenn die folgenden Eigenschaften zutreffen:

- Extremely large **volumes** of data: riesige Datenmengen,
- extremely high **velocity** of data: schnelle Datenverarbeitung (zumeist mit Zugriff auf parallele Verarbeitungsmechanismen),
- extremely wide **variety** of data: Verarbeitung von verschiedensten Datenarten (in vielen Fällen unstrukturierter Natur),
- high **veracity**: hohe Zuverlässigkeit und Korrektheit der erzielten Aussagen.

Die ersten drei Eigenschaften wurden in früheren Versionen der Definitionen von Big Data herangezogen, jedoch schnell um „Veracity" ergänzt. Durch den gemeinsamen Anfangsbuchstaben der englischen Begriffe wird diese Definition auch gern als 3V oder 4V bezeichnet.

Definition:
Tatsächlich wurden diese Eigenschaften später erweitert (siehe u. a. impact):[1]

- **Value**: Big-Data-Anwendungen sollen hohe Mehrwerte generieren. Die Investitionen in Personal und Infrastruktur müssen sich also lohnen bzw. sogar äußerst attraktiv sein. (Fasel, Grundlagen)
- **Variability**: Die Daten sind nicht nur unterschiedlich, sondern erlauben auch immer wieder neue, andersartige Interpretationen. Sie werden also vielfach genutzt und durchlaufen auch gleiche Verarbeitungsprozesse immer wieder.
- **Visualization**: Nur durch korrekte, saubere, aber auch gestalterisch ansprechende Aufbereitung von Daten in visueller Form kann es gelingen, Entscheidungsträger schnell und umfassend zu informieren. Big-Data-Ergebnisse liegen zumeist nicht in Sichten direkt vor, sondern sind reine Datenkolonnen. Eine Visualisierung ist unumgänglich.

Diese Eigenschaften lassen sich weiter charakterisieren:

Dimension	**Technisch**	**Qualitativ**	**Zielorientiert**
Eigenschaft	volume velocity variety	veracity variability	visualization value

Dennoch gilt ebenso:

Merke: **Auch wenn Big-Data-Analysen gern mit 4V oder 7V charakterisiert werden, heißt dies nicht, dass alle Analysen alle Eigenschaften vollständig erfüllen müssen! Es gibt viele Fälle, bei denen beispielsweise**

- das „Volume" gering, aber die „Velocity" hoch ist,
- „Velocity" gar keine Rolle spielt, dagegen die „Variability" äußerst hoch ist.

Einige Beispiele seien (EADL2019-1) entnommen:

- (Volume) Typische Transaktionsanalysen von Banken sind ein Fall großer Datenmengen, auch das CERN sei hier erwähnt.
- (Velocity) US Express analysiert Routen basierend auf Sensor-Daten und passt diese dann an.
- (Variability) NextBio analysiert menschliche Gene, was mit relationalen Datenbanken aufgrund der immer wiederkehrenden einfachen Struktur (A-G-T…) in riesigen Datenkolonnen nicht möglich war.
- (Value) Geisinger kombiniert automatisch Patientendaten mit Arztnotizen und spart enorme Lizenzkosten.

Damit wird bereits jetzt deutlich, was im folgenden Kapitel genauer zu erläutern sein wird: **Eine Big-Data-Infrastruktur erfordert nicht nur Konzepte für technische Umsetzungen!**

[1] In (JODLB2018) finden sich die folgenden Begriffe in deutscher Sprache, die sich allerdings im IT-Umfeld wenig behaupten konnten: Volumen, Geschwindigkeit, Mannigfaltigkeit, Unsicherheit und Visualisierung, Bedeutungswandel, Wert. Veracity wird zunächst mit Unsicherheit, später in der angegebenen Quelle dann aber auch mit Richtigkeit übersetzt.