Self-made
Big Data Analytics

Die Analyse unstrukturierter Daten mit Hadoop und der Hybrid-Cloud-Architektur IBM Cloud™ Pak for Data

Inhaltsverzeichnis

Abbildungsverzeichnis

Seite

1 Einleitung

„2013 wurden mit knapp 4,5 Milliarden Terrabyte so viele Daten produziert, wie in der gesamten Menschheitsgeschichte zuvor – und jeden Tag kommen etwa 2,5 Millionen Terrabyte hinzu.“[1]

Die Suche auf Google nach dem Begriff ‚Big Data‘ liefert ungefähr 7.620.000.000 Ergebnisse innerhalb von 0,35 Sekunden. Angefangen bei der Definition auf Wikipedia bis hin zum ‚Big Data in drei Minuten‘-Video auf YouTube ist alles dabei. Es ist nicht wirklich überraschend, wie viel über dieses Thema diskutiert wird. Doch wer kennt sich wirklich mit Big Data aus? Was ist Big Data und wofür ist das gut? Wie bekommen Wirtschaft und Wissenschaft heutzutage den immer größer und unübersichtlicher werdenden Datenberg in den Griff?

Der Begriff Big Data beschreibt große Datenmengen und das schnelle Anwachsen von Daten, die sowohl in strukturierter Form von Zahlen und Tabellen als auch in unstrukturierter Form von Texten, Webseiten und Videos vorliegen. Diese Daten verändern sich rasant und besitzen daher eine hohe Halbwertzeit.[2]

Auf Unternehmen hat Big Data insofern Einfluss, dass das klassische Controlling auf dynamische Art und Weise erweitert und durch die Datenmengen die Entscheidungsgeschwindigkeit deutlich erhöht wird. „Diesen tiefgreifenden Wandel gilt es zu meistern und auch zu gestalten, um in Unternehmen nicht an Bedeutung zu verlieren.“[3] Der Ruf nach technologisch anspruchsvolleren Tools und Programmen State of the Art wird größer, um die immense Datenflut zähmen zu können.[4] Allerdings schrecken Fragen wie ‚Sind die Anderen bereits viel weiter?‘ in Kombination mit übereilten Aussagen, wie ‚Wir sind zu klein, um Big Data zu nutzen!‘ und ‚Es ist zu teuer, Big Data einzuführen!‘ viele Unternehmen davon ab, Big Data tatsächlich einzusetzen.

Hadoop ist eine vollkommen kostenlose, enorm skalierbare Lösung, Big-Data-Infrastrukturen im Unternehmen bereitzustellen. Auf Basis des MapReduce-

---

1 Kneser & Dietsche, 2014
2 vgl. Freiknecht & Papp, 2018, S. 10; Dumbill, 2012
3 vgl. Weichel & Herrmann, 2016, S. 9
4 vgl. Radtke & Litzel, 2016

Algorithmus von Google werden in einem Cluster mit vielen Knoten riesige Datenmengen parallel verarbeitet. Heterogene Daten werden in kleine Päckchen aufgeteilt, auf mehreren Clusterknoten parallel verarbeitet und später wieder zusammengeführt. Durch diese Technologie ist es nahezu ausgeschlossen, an die Grenze der Möglichkeiten von Hadoop zu stoßen.[5]

Mit der IBM Cloud™ Pak for Data betreiben und managen Unternehmen ihr eigenes Hadoop-Cluster selbst. IBM Cloud™ Pak for Data ist eine auf Unternehmen abgestimmte Container-Software für Kubernetes. Cloud-basierte Anwendungen können problemlos entwickelt und im eigenen Rechenzentrum wahlweise in einer Private oder in einer Public Cloud eingesetzt werden.[6]

Alle Unternehmen stehen vor derselben Herausforderung: riesige Mengen an Anwendungen, Plattformen, Netzwerke, Verwaltungstools, Sicherheits-anforderungen sowie Daten, Daten und nochmal Daten dominieren das tägliche Geschäft. Hybrid-Cloud-Architekturen, wie die IBM Cloud™ Pak for Data, ergänzen und kombinieren vorhandene Strukturen mit modularen, skalierbaren und flexiblen Services. Der wertvollste Vorteil der IBM Cloud™ Pak for Data besteht allerdings darin, dass Unternehmen die vollständige Kontrolle über ihre Daten zurückgewinnen.[7]

---

5 vgl. Joos, 2014
6 vgl. Ostler, 2018
7 vgl. Hurwitz & Kirsch, 2018, S. 22; 36

2 Big Data

„Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.“[8]

2.1 Big Data – eine leere Worthülse?

Der Begriff „Big Data“ bezeichnet große, komplexe, schnelllebige und schwach strukturierte Datenmengen.[9] Für Unternehmen fällt unter Big Data auch der Einsatz und die Analyse großer Datenmengen, die aus verschiedenen Quellen gewonnen werden, um auf Basis entscheidungsrelevanter Erkenntnissen einen wirtschaftlichen Nutzen und einen Wettbewerbsvorteil zu erlangen. Big Data ist also mehr als nur eine leere Worthülse. Big Data umfasst qualitativ vielfältige sowie in unbekanntem Umfang anfallende strukturierte und unstrukturierte Informationen in Form von Daten.[10]

Abb. 1: Big Data

Quelle: Altmann, 2019

[8] Gartner, Inc., 2019
[9] vgl. Christl, 2014, S. 12
[10] vgl. BITKOM, 2012, S. 7

Das „Big“ in der Definition von Big Data bezieht sich hauptsächlich auf folgende drei Dimensionen:

- **volume:** beschreibt den Umfang und das Datenvolumen. Dieser liegt in der Regel im Tera- bis Zettabyte-Bereich (Megabyte=10^6 Byte, Gigabyte = 10^9 Byte, Terabyte = 10^{12} Byte, Petabyte = 10^{15} Byte, Exabyte = 10^{18} Byte, Zettabyte = 10^{21} Byte),
- **velocity:** beschreibt die Geschwindigkeit, mit der die Datenmengen generiert und transferiert werden. Datenströme können in Echtzeit ausgewertet und analysiert werden,
- **variety:** beschreibt die Vielfalt der Speicherung von strukturierten, semi-strukturierten und unstrukturierten Daten in jeglicher Form, also die Bandbreite der Datentypen und -quellen (Text, Bilder, Audio und Video).[11]

Experten erweitern die drei V’s um zwei weitere Dimensionen:

- **value:** beschreibt unternehmerischen Mehrwert. In Unternehmen wird nur in den Bereichen investiert, wo eine Hebelwirkung besteht bzw. der Unternehmenswert dadurch gesteigert wird,
- **veracity:** beschreibt die Sicherstellung der Datenqualität bei der Auswertung von Daten, denn sämtliche Datenbestände liegen in unterschiedlicher Datenqualität vor.[12]

Die Verarbeitung von Daten erfordert heutzutage erheblichen Aufwand, der mit rein menschlichem Leistungsvermögen nicht zu bewältigen ist. Um valide Aussagen z.B. über Trends im Zeitverlauf treffen zu können, muss die Vielfalt der Daten identifiziert, strukturiert, miteinander verknüpft und anschließend analysiert werden. Erst dann lassen sich aus den daraus generierten Informationen logische Schlüsse ziehen und weitere Maßnahmen planen.

2.2 Variety: Die Vielfalt der Daten

Big Data entsteht vor allem durch „variety“. Die Vielfalt der Daten und der Datenquellen zeigen, dass es nicht nur um große Datenmengen geht, sondern

---

[11] vgl. Gartner, Inc., 2019; Faser & Meier, 2016, S. 6
[12] vgl. Bachmann, Kemper, & Gerzer, 2014, S. 23ff.; Faser & Meier, 2016, S. 6

auch um verschiedene Arten von Daten. Diese können sowohl in strukturierter als auch in unstrukturierter Form vorliegen.

Strukturierte Daten sind in relationalen Datenbanken organisiert. Sie haben ein vorgegebenes Format, in das alle Informationen abgelegt werden. Die bekanntesten Formen strukturierter Daten sind Structured Query Language (SQL)- oder Access-Datenbanken.[13]

Unstrukturierte Daten sind alle Daten und Informationen, die nicht in Datenbanken oder anderen speziellen Datenstrukturen abgelegt sind. Die Formate unstrukturierter Daten sind vielfältig. Zu den unstrukturierten Daten in Textform zählen z.B. E-Mails, PowerPoint-Präsentationen, Word-Dokumente, Sofortnachrichten und Chatmitteilungen sowie Daten aus einer Kollaborationssoftware. Unstrukturierte Daten in Nicht-Textform sind beispielsweise Bilder im JPEG-Format, MP3-Audio-Dateien oder Videos im Flash-Format.[14]

Daten strömen unsortiert und unstrukturiert durch unsere Netze. Relationale Datenbanken können diese Datenmengen nicht mehr erfassen, verwalten und verarbeiten. Eine wesentliche Herausforderung ist es daher, das Potenzial aus dieser Gemengelage auszuschöpfen und für Unternehmen verwertbare und zum eigenen Vorteil nutzbare Muster abzuleiten.[15]

2.3 Wie Unternehmen von Big Data profitieren können

Gerade für Wissenschaft und Wirtschaft schlummern in unstrukturierten Daten wertvolle Informationen, die sich oftmals schwer herausfiltern lassen.[16]

Der Fortschritt in der Informationstechnik ermöglicht es, große Datenmengen – Big Data – in weniger Zeit strukturieren und analysieren zu können. Diese Analysen können Unternehmen zu Ihrem Vorteil und als Basis für wichtige Unternehmensentscheidungen nutzen. Können also aus Datenmengen Informationen, Muster und Korrelationen abgelesen und aus den daraus gewonnenen Erkenntnissen Maßnahmen abgeleitet werden, sind Unternehmen

---

[13] vgl. TechTarget Germany GmbH, 2014
[14] vgl. TechTarget Germany GmbH, 2015
[15] vgl. Lesniak, 2017; Fachinger, 2018
[16] vgl. TechTarget Germany GmbH, 2015

in der Lage ihre Prozesse zu optimieren, die Wertschöpfung zu steigern und dadurch Wettbewerbsvorteile zu erzielen.[17] In den meisten Fällen tragen die Optimierungsmaßnahmen sogar zu einer Kostensenkung, Zeitersparnis und Qualitätssteigerung von Produkt und Dienstleistung bei.[18] Zusätzlich können mithilfe von Daten aus unterschiedlichen Zusammenhängen neue Informationen aufgewertet und die strategische Planung von Unternehmen neu ausgerichtet werden oder sogar neue Geschäftszweige und -modelle entstehen.[19]

Wer in Wissenschaft und Wirtschaft also dauerhaft in der ersten Liga spielen möchte, muss Big Data in den Griff bekommen. Denn je größer die Datenvielfalt, desto schwieriger wird es, diese Daten zu speichern und zusammen auszuwerten (variety). Hadoop schafft bei der Datenbewältigung Abhilfe.[20]

---

[17] vgl. Gadatsch & Landrock, 2017
[18] vgl. Fachinger, 2018
[19] vgl. Gadatsch & Landrock, 2017
[20] vgl. Schön, 2015

3 Apache Hadoop

„There's a revolution happening in the use of big data, and Apache Hadoop is at the center of it."[21]

Die Open-Source-Software Apache Hadoop ist ein in Java geschriebenes Framework zum verteilten Speichern riesiger Datenmengen. Hadoop basiert sowohl auf dem MapReduce-Algorithmus von Google Inc. als auch auf Vorschlägen des Google-Dateisystems. Intensive Rechenprozesse verarbeiten große Datenmengen im Petabyte-Bereich[22] – Big Data – auf einer sog. Commodity-Hardware[23] in vielen kleinen Prozessschritten parallel. Hadoop wird in einem horizontal skalierbaren Cluster betrieben. Diesem Cluster können problemlos weitere Knoten – Nodes – hinzugefügt werden. Ein Knoten – NameNode – übernimmt im Cluster die Steuerung, die anderen Knoten – DataNodes – die Berechnungen.[24] Im BI-Umfeld verarbeitet Hadoop so große, inhomogene Datenmengen in hoher Geschwindigkeit. Die für Reports und Analysen herangezogenen und in sehr unterschiedlicher Struktur vorliegenden Daten stammen dabei aus vielen verschiedenen Datenquellen.[25] Zu den bekanntesten Unternehmen, bei denen Hadoop im Einsatz ist, gehören u.a. eBay, LinkedIn, Facebook, Twitter, Last.fm, AOL, Lineberger Comprehensive Cancer Center, die New York Times, Yahoo und die IBM.[26]

---

[21] Henschen, 2011

[22] In der nächstgrößeren Einheit entspricht 1 Petabyte 1.000 Terabyte bzw. 1.000.000 Gigabyte bzw. 1.000.000.000.000.000 Bytes. Ein Papierstapel mit 100 Seiten (A4) misst etwa eine Höhe von 1 cm. Würde der gesamte Inhalt von 1 Petabyte einseitig auf Papier in ASCII-Symbolen mit 80 Zeichen pro Zeile und 53 Zeilen pro Seite (≙ 4.240 Zeichen auf einer einseitig bedruckten DIN-A4-Seite) ausgedruckt, wäre dieser Stapel dann 250.000.000.000 cm hoch. In Kilometer umgerechnet würde dieser Papierstapel bei einem Erdumfang von etwa 40.000 km ca. 62-mal um den Erdäquator verlaufen (vgl. (Wiedermann, 2011); (Aschermann, 2017)).

[23] Hadoop ist darauf ausgelegt, Commodity-Hardware einzusetzen. Sie ist verhältnismäßig günstig, leicht zu beziehen, mit anderen Geräten kompatibel und einfach auszutauschen. Ein bestehendes Cluster kann so problemlos um weitere Knoten erweitert werden, um mehr Rechenleistung erbringen zu können (Scale-Out; vgl. Freiknecht & Papp, 2018, S. 22).

[24] vgl. Freiknecht & Papp, 2018, S. 21; Luber & Litzel, 2016

[25] Luber & Litzel, 2016

[26] vgl. Hadoop Wiki, 2018

3.1 Die Bestandteile von Hadoop

Die Grundlage für das Software-Framework ist Hadoop Common. Dieser liefert die Grundfunktionen für die anderen Module des Frameworks (z.B. die Java-Archiv-Files und -Scripts für den Start der Software). Die Kommunikation innerhalb der Cluster sowie die Zugriffe auf darunterliegende Dateisysteme erfolgt über Schnittstellen.[27] Die drei zentralen Bestandteile bzw. Module von Apache Hadoop sind:

- **Hadoop Distributed File System (HDFS):** Das HDFS ist ein über dem gesamten Cluster verteiltes Dateisystem. Auf ihm werden die Daten abgelegt. HDFS erlaubt den schnellen Zugriff auf große Datenmengen sowie die zuverlässige und dauerhafte Speicherung der Daten.
- **MapReduce-Algorithmus:** Mit dem MapReduce-Algorithmus können Daten gemäß einer zweiphasigen Verteilung durch Mapper und Reducer verteilt verarbeitet werden.
- **Yet Another Resource Negotiator (YARN):** Der Ressourcen-Manager YARN ist eine Cluster-Verwaltungstechnik. YARN bestimmt, welche Prozesse welchen Ressourcen eines Clusters zugeteilt werden. Zusätzlich legt YARN fest, wie die Jobs abgearbeitet werden.[28]

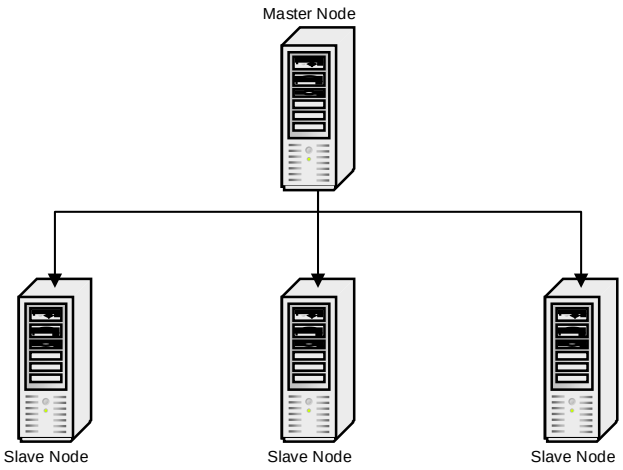
3.2 Das Hadoop-Prinzip

Das Prinzip von Hadoop ist recht einfach erklärt: Ein Hadoop-Cluster funktioniert nach dem Master-Slave-Prinzip. Der Master übernimmt die Rolle des NameNodes und verwaltet das Cluster. Er verteilt Jobs und Daten an die (optimalsten) Slaves – den DataNodes.[29]

---

[27] vgl. Joos, 2015; Freiknecht & Papp, 2018, S. 22
[28] vgl. Freiknecht & Papp, 2018, S. 22; Luber & Litzel, 2016; Printz, 2013
[29] vgl. Bayer, 2013

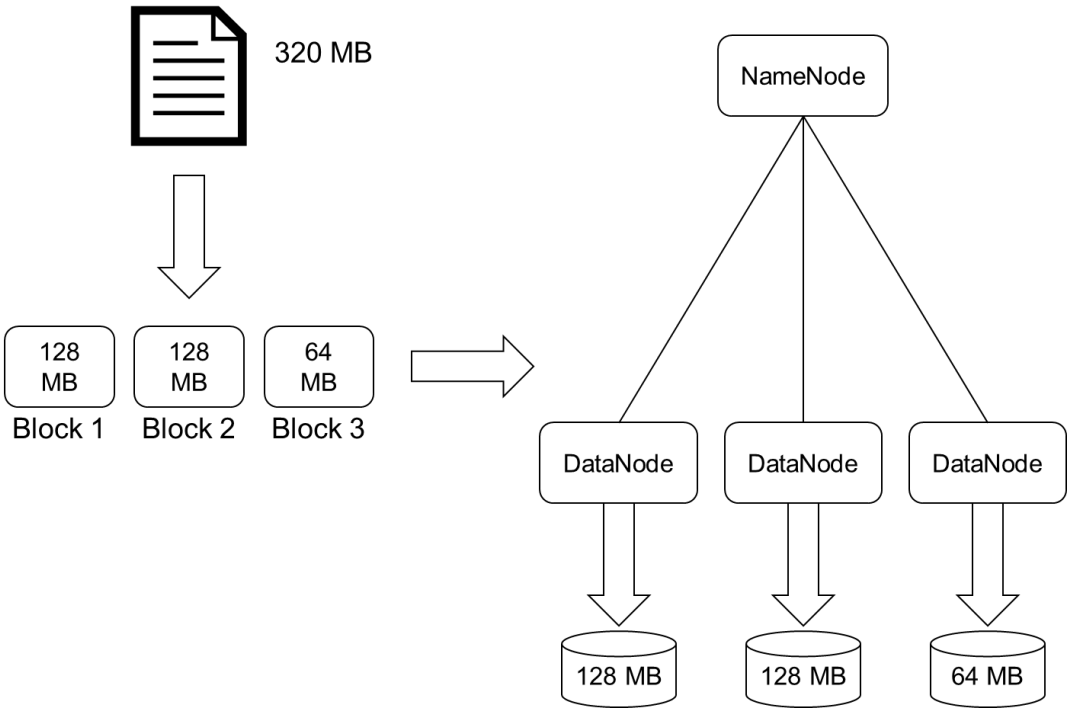
Abb. 2: Hadoop’s Master-Slave-Prinzip



Quelle: in Anlehnung an Dabhoiwala, 2018

Um Datenverluste zu vermeiden, zerlegt der Master große Datenmengen (Big Data) in einzelne Datenblöcke und setzt von diesen Kopien auf. HDFS verteilt die Daten auf mehrere Clusterknoten.

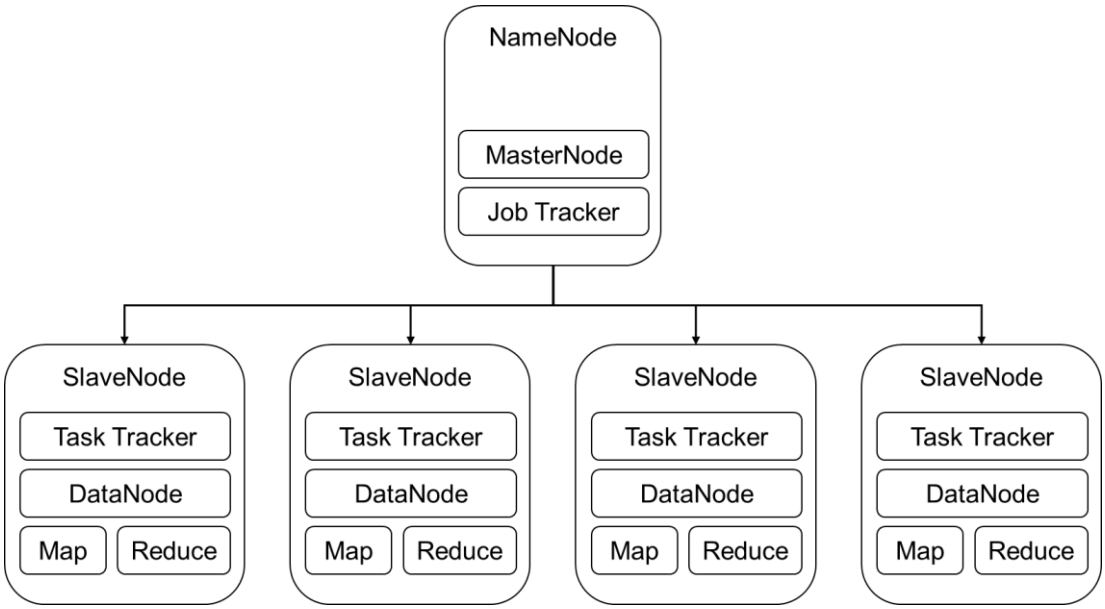
Abb. 3: Datenverteilung und Datenspeicherung in Hadoop



Quelle: in Anlehnung an Dabhoiwala, 2018

MapReduce verarbeitet diese dann an ihrem Ablageort parallel und führt sie schließlich wieder zusammen.[30] Sollte ein Knoten ausfallen, geht durch diese Vorgehensweise keine Information verloren, da der NameNode noch weiß, in welche Datenblöcke die Dateien zerlegt und wo diese noch abgelegt sind.[31]

Abb. 4: Hadoop's Master-Slave-Architektur

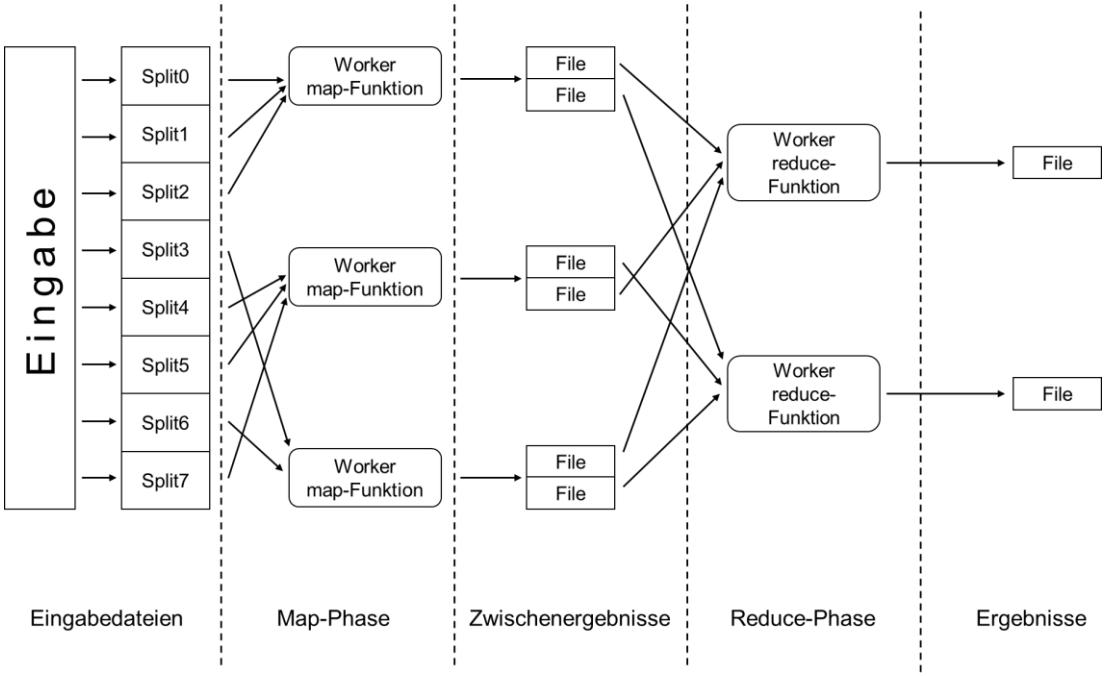


Quelle: in Anlehnung an Jumani, 2017

In einem speziell für hoch parallelisierte Datenverarbeitung in einem Cluster ausgelegtem Algorithmus wird die zu verarbeitende Datei auf mehrere Map-Prozesse verteilt. Diese berechnen in der Map-Phase parallel die Zwischenergebnisse. In der darauffolgenden Reduce-Phase werden diese Zwischenergebnisse eingesammelt und daraus eine Ergebnisdatei ermittelt.[32]

[30] vgl. Bayer, 2013
[31] vgl. Freiknecht & Papp, 2018, S. 74f.; Bayer, 2013
[32] vgl. Bayer, 2013

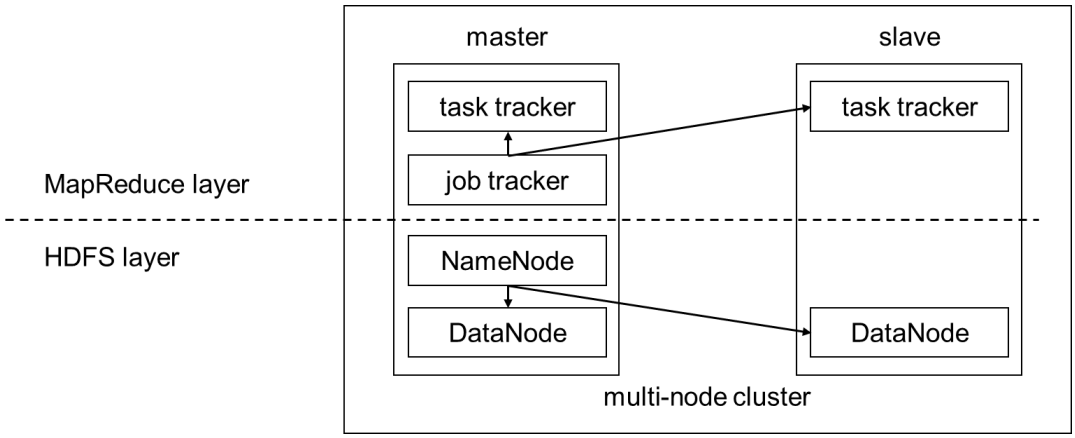
Abb. 5: Verarbeitungsschritte einer Hadoop-MapReduce-Anwendung



Quelle: in Anlehnung an Fischer, 2010

In Hadoop werden Rechenaufgaben als Job bezeichnet, wobei der Master als „JobTracker“ der Kopf des Systems ist. Er verteilt und verwaltet die Jobs im Cluster. Der NameNode weiß, wo die Daten in den DataNodes liegen. Der JobTracker verteilt dann die Aufgaben an die TaskTracker auf den Slave-Systemen, die schließlich die eigentliche Jobabwicklung übernehmen.[33]

Abb. 6: Der Master als Kopf eines Hadoop-Systems



Quelle: in Anlehnung an Bayer, 2013

[33] vgl. Bayer, 2013

Im Big Data-Umfeld ist Hadoop mit seinen drei Komponenten HDFS, MapReduce und YARN speziell für komplexe Analysen die ideale Lösung, sog. Data Lakes[34] problemlos realisieren bzw. große Datenmengen speichern und verarbeiten zu können.[35]

3.3 Die Vorteile von Hadoop

Dem CAP-Theorem[36] nach kann ein verteiltes Dateisystem nur zwei von folgenden drei Eigenschaften gleichzeitig garantieren und erfüllen:

- **Consistency (Konsistenz):** Daten sind immer aktuell und synchronisiert. Jeder Benutzer erhält jederzeit, unabhängig davon, welcher Knoten ihn zurückgibt, die gleiche Antwort auf seine Leseanfrage.
- **Availability (Verfügbarkeit):** Ein Benutzer erhält innerhalb einer angemessenen Zeit immer eine Antwort vom System.
- **Partition Tolerance (Ausfalltoleranz):** Das System funktioniert auch dann weiter, wenn einige seiner Komponenten ausfallen.[37]

Hadoop erfüllt zwei dieser Eigenschaften in vollem Umfang: Availability und Partition Tolerance. Konsistenz wird nicht unterstützt, da Informationen nur auf dem NamenNode liegen, auf dem die Replikate platziert sind und diese Informationen nicht für jeden einzelnen Knoten des Clusters verfügbar sind.[38]

Unternehmen profitieren durch den Einsatz von Hadoop:

- Verfügbarkeit: Daten können aufgrund der ständigen Verfügbarkeit auf vielfältige und flexible Weise analysiert werden. Analysen stehen so für wichtige Unternehmensentscheidungen zeitnah zur Verfügung.
- Skalierbarkeit: HDFS ist hoch skalierbar. Es können sehr große Datensätze gespeichert und auf viele Knoten verteilt werden. Diese großen Datensätze können schließlich parallel verarbeitet werden.

---

[34] Ein Data Lake ist ein sehr großer Datenspeicher, der daten aus den unterschiedlichsten Quellen in ihrem ursprünglichen Rohformat aufnimmt (vgl. Luber & Nitzel, 2018).
[35] vgl. Luber & Nitzel, 2018
[36] Die Abkürzung CAP steht für Consistency, Availability und Partition Tolerance (Anm. d. Verf.).
[37] vgl. Bekker, 2018
[38] vgl. Nazrul, 2018

- Ausfalltoleranz: Jeder Datenblock wird im Cluster standardmäßig dreimal repliziert und auf drei Maschinen bzw. Datenknoten gespeichert. Dadurch werden die Daten vor dem Ausfall eines DataNodes geschützt.
- Speicherplatz: Die Lösung ist problemlos skalierbar und leicht erweiterbar. Wenn mehr Festplattenspeicher benötigt wird, können einem Hadoop-Cluster einfach weitere DataNodes hinzugefügt werden.
- Flexibilität: Sämtliche heterogene Daten können unabhängig davon, ob sie strukturiert, teilstrukturiert oder unstrukturiert sind, gespeichert werden.
- Kosteneffizient: Die Open Source-Software Hadoop bietet durch die Nutzung von Commodity-Hardware und ohne Bindung an einen bestimmten Hersteller eine kostengünstige Speicherung großer und vielfältiger Datenmengen. Zusätzlich lohnt sich dadurch die Vorratshaltung von Rohdaten, deren Wert noch unklar ist.[39]

Um Big Data in vollem Umfang beherrschen zu können, benötigen Unternehmen eine technisch ausgereifte Plattform. Mit der IBM Cloud™ Pak for Data können die Datenmassen vereinheitlicht und vereinfacht erfasst, organisiert und analysiert werden.

[39] vgl. Freiknecht & Papp, 2018, S. 74f.; Bayer, 2013; Müller, 2016; Dabhoiwala, 2018

4 IBM Cloud™ Pak for Data

IBM Cloud™ Pak for Data ist die ideale Daten- und Analyseplattform mit integrierter Governance. Daten und Dienste können mit einer Public-Cloud verknüpft und sensible, unternehmensinterne Daten aber weiterhin in einer „privaten“ Cloud verwaltet werden. Diese Hybrid-Cloud-Lösung ist für Unternehmen ein riesengroßer Fortschritt: Daten erscheinen wieder vollwertig und sie übernehmen so wieder die komplette Kontrolle über ihre wertvollen Daten.[40]

4.1 Das Zusammenspiel von Hadoop und IBM Cloud™ Pak for Data

Mit der IBM Cloud™ Pak for Data betreiben und managen Unternehmen ihr eigenes Hadoop-Cluster selbst. IBM Cloud™ Pak for Data ist eine auf Unternehmen abgestimmte Container-Software für Kubernetes. Cloud-basierte Anwendungen können problemlos entwickelt und im eigenen Rechenzentrum wahlweise in einer Private oder in einer Public Cloud eingesetzt werden.[41] Im Zusammenspiel mit Hadoop wird die IBM Cloud™ Pak for Data auf einem sog. EdgeNode[42] des Hadoop-Clusters installiert.[43] Administratoren und Entwickler arbeiten bei dieser Lösung zusammen, damit der Cluster optimal funktioniert. Dienste können installiert oder über die Cloud betrieben werden.[44] Nutzer greifen schließlich über die Cloud sicher auf sämtliche Daten im Hadoop-Cluster zu, übermitteln interaktive Spark-Aufträge und planen Jobs, die schließlich als YARN-Anwendung im Hadoop-Cluster ausgeführt werden.[45]

Mit IBM Cloud™ Pak for Data und Hadoop können umfangreiche Analysen komplexer Daten heterogener Datenquellen in einer Cloud-Umgebung

---

[40] vgl. IBM Deutschland GmbH, 2019; SIGS DATACOM GmbH, 2018, S. 15

[41] vgl. Ostler, 2018

[42] Ein leerer EdgeNode ist ein virtueller Linux-Computer, auf dem die gleichen Clienttools installiert und konfiguriert sind wie auf den Hauptknoten. Auf einem EdgeNode werden allerdings keine Apache Hadoop-Dienste ausgeführt. Ein EdgeNode kann zum Zugreifen auf ein Cluster und zum Testen und Hosten von Clientanwendungen verwendet werden (vgl. Microsoft Corporation, 2018).

[43] vgl. IBM Corporation, 2019; IBM Deutschland GmbH, 2019

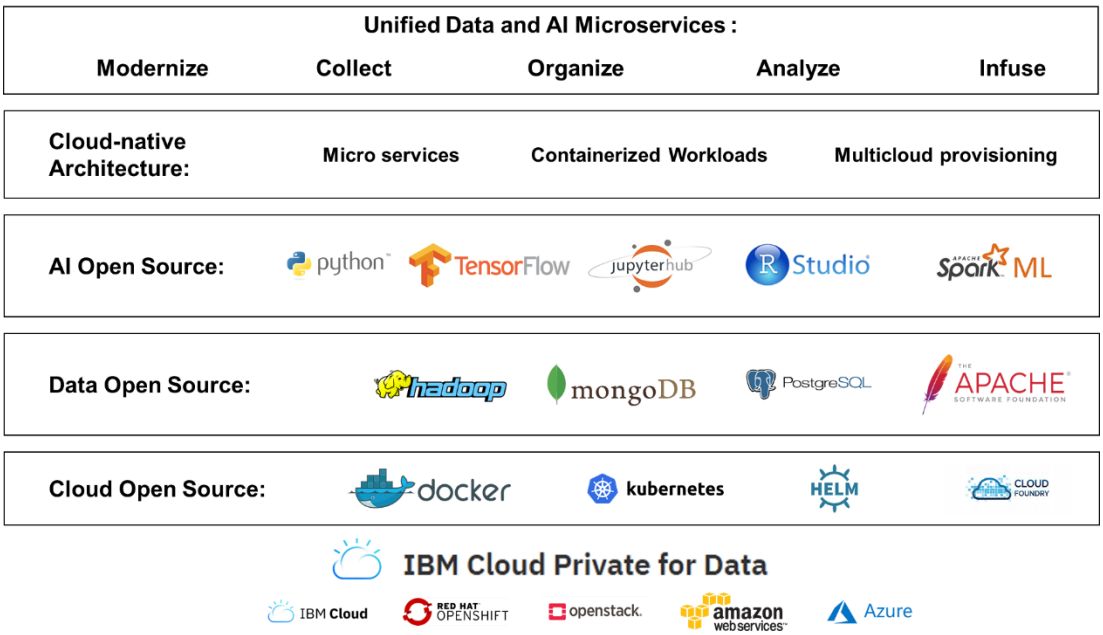
[44] vgl. Joos, 2015

[45] vgl. IBM Corporation, 2019; IBM Deutschland GmbH, 2019

ausgeführt, in Self-Service-Oberflächen übersetzt und Endnutzern zur weiteren Bearbeitung und Nutzung bereitgestellt werden.[46]

Zusätzlich können in der Anwendungsplattform IBM Cloud™ Pak for Data neben Hadoop zahlreiche weitere Services kombiniert und somit eine performante Umgebung für die Entwicklung, Implementierung und Verwaltung von Advanced-Analytics-Anwendungen bereitgestellt werden.[47]

Abb. 7: Mögliche Open Source-Architektur



Quelle: in Anlehnung an IBM Deutschland GmbH, 2019

4.2 Die Vorteile von IBM Cloud™ Pak for Data für Unternehmen

Alle Unternehmen stehen vor derselben Herausforderung: riesige Mengen an Anwendungen, Plattformen, Netzwerke, Verwaltungstools, Sicherheits-anforderungen sowie Daten, Daten und nochmal Daten dominieren das tägliche Geschäft. Hybrid-Cloud-Architekturen ergänzen und kombinieren vorhandene Strukturen mit modularen, skalierbaren und flexiblen Services. Unternehmen steigern so ihr Innovationspotenzial und stärken vor allem langfristig ihre Wettbewerbsfähigkeit. Der wertvollste Vorteil der IBM Cloud™ Pak for Data

[46] vgl. Hurwitz & Kirsch, 2018, S. 30
[47] vgl. IBM Deutschland GmbH, 2019

besteht darin, dass Unternehmen die Kontrolle über ihre Daten zurückgewinnen. Sämtliche Elemente lassen sich wieder in der eigenen Umgebung kontrollieren.[48]

Governance und Compliance

Unternehmen erhalten wieder die volle und direkte Kontrolle über die Sicherheit in ihrem eigenen Rechenzentrum. Der Zugriff auf sämtliche Unternehmensdaten erfolgt einfach und kontrolliert. Die Definition und Umsetzung von unternehmenseigenen Richtlinien und Datenkataloge ermöglichen einen kontrollierten Zugriff auf Datenbestände und analytische Modelle. Fehlerhafte oder unvollständige Daten werden direkt angezeigt. Im Rahmen einer Überprüfung lassen sich diese Daten schließlich manuell vervollständigen oder korrigieren. Das führt zu einer deutlichen Verbesserung der Datenqualität.[49]

Zugriffberechtigung und Teamwork

Mit der IBM Cloud™ Pak for Data arbeiten User, Datenwissenschaftler, Dateningenieure und Entwickler mit rollenspezifischen Bedienoberflächen. Eine personalisierte oder wahlweise auch kollaborativ nutzbare Bedienerschnittstelle gefährdet weder Datenschutz noch Datensicherheit und spart sogar noch wertvolle Arbeitszeit ein, da die Mitarbeiter keine Zeit mehr mit dem Auffinden und Validieren von Daten vergeuden.[50]

Sicherheit

Mit der IBM Cloud™ Pak for Data werden die Daten hinter das eigene Firewall-geschützte Rechenzentrum gelegt. Unternehmen behalten die volle Kontrolle und entscheiden eigenständig, welche Daten sie über Schnittstellen nach Außen bereitstellen wollen. Zusätzlich sorgen Virtual Private Cloud, Verschlüsselung und API-Schlüssel für die Sicherheit von Daten.[51]

---

[48] vgl. Hurwitz & Kirsch, 2018, S. 22; 36
[49] vgl. SIGS DATACOM GmbH, 2018, S. 24
[50] vgl. SIGS DATACOM GmbH, 2018, S. 25
[51] vgl. Baumann, 2018; IBM Deutschland GmbH, 2019

Flexibilität[52]

- Skalierbarkeit: Die Infrastruktur kann bei Bedarf vergrößert oder verkleinert werden.
- Speicheroptionen: Nutzer können abhängig von den Sicherheits-anforderungen zwischen Public-, Private- oder Hybrid-Speicher auswählen.
- Kontrollmöglichkeiten: Unternehmen können zwischen den Optionen Software as a Service (SaaS), Platform as a Service (PaaS) und Infrastructure as a Service (IaaS) bestimmen, wie viel Kontrolle sie ausüben wollen.

Effizienz[53]

- Einfacher Zugriff: Der Zugriff auf Cloud-basierte Anwendungen und Daten ist von praktisch jedem Gerät mit Internetverbindung aus möglich.
- Datensicherheit: Dank vernetzter Sicherungen führen Hardwarefehler nicht zu Datenverlusten.
- Einsparpotenziale: Da Cloud-Computing ausschließlich Remote-Ressourcen nutzt, sparen Unternehmen Kosten für Hardware ein und bezahlen durch ein nutzungsabhängiges Zahlungsmodell nur für die tatsächlich verwendeten Ressourcen.

Der größte Nutzen ist schließlich der strategische Nutzen, von dem Unternehmen profitieren. Cloud-Service-Provider (CSPs) betreiben die zugrunde liegende Infrastruktur. Unternehmen können sich so auf die Anwendungsentwicklung und weitere Prioritäten konzentrieren. Regelmäßige Updates der Service-Provider stellen den Nutzern die jeweils neueste Technologie bereit. Der weltweit mögliche Zugriff auf Daten und Services erlaubt Teams an weit entfernten Standorten zusammenzuarbeiten. Unternehmen können also insgesamt schneller als ihre Mitbewerber agieren, die erst wertvolle Ressourcen für eine funktionierende IT-Infrastruktur abstellen müssen.[54]

---

[52] vgl. IBM Deutschland GmbH, 2019
[53] vgl. IBM Deutschland GmbH, 2019
[54] vgl. IBM Deutschland GmbH, 2019

Autoreninfo:

Carmelo Barba arbeitet seit über vier Jahren als Consultant bei der gmc² GmbH. Dieses White Paper entstand begleitend zum Messeauftritt der gmc² GmbH auf der 15. TDWI-Konferenz vom 24.-26. Juni 2019 in München. Auf der TDWI-Konferenz präsentiert die gmc² einen zweiteiligen Hackathon zu „Self-made Big Data Analytics mit Hadoop und RaspberryPi“.

5 Literaturverzeichnis

Altmann, G. (2019). *Big Data*. Abgerufen am 18. Juni 2019 von Pixabay GmbH: https://pixabay.com/de/illustrations/daten-big-data-internet-online-www-4132580/

Aschermann, T. (9. April 2017). *Erdumfang in km genau berechnen - so einfach geht's*. Abgerufen am 29. Mai 2019 von www.chip.de: https://praxistipps.chip.de/erdumfang-in-km-genau-berechnen-so-einfach-gehts_92022

Bachmann, R., Kemper, G., & Gerzer, T. (2014). *Big Data - Fluch oder Segen?: Unternehmen im Spiegel gesellschaftlichen Wandels.* Frechen: Mitp Verlag.

Baumann, T. (01. Mai 2018). *Cloud Transformation mit Kubernetes.* Abgerufen am 19. Juni 2019 von tiri GmbH: https://blog.thomasbaumann.com/blog/ibm-cloud-private-icp-docker-kubernetes-multicloud

Bayer, M. (25. März 2013). *Hadoop - der kleine Elefant für die großen Daten*. Abgerufen am 2019. Juni 2019 von Computerwoche.de: https://www.computerwoche.de/a/hadoop-der-kleine-elefant-fuer-die-grossen-daten,2507037,3

Bekker, A. (10. Juli 2018). *Apache Cassandra vs. Hadoop: Distributed File System: Wann jedes davon besser passt*. Abgerufen am 06. Juni 2019 von ScienceSoft USA Corporation.: https://www.scnsoft.de/blog/cassandra-vs-hadoop

BITKOM. (2012). *Big Data im Praxiseinsatz - Szenarien, Beispiele, Effekte.* (T. u. Bundesverband Informationswirtschaft, Hrsg.) Abgerufen am 08. Mai 2019 von https://www.bitkom.org/Bitkom/Publikationen/Leitfaden-Big-Data-im-Praxiseinsatz-Szenarien-Beispiele-Effekte.html

Christl, W. (2014). *Kommerzielle digitale Überwachung im Alltag.* Wien: Bundesarbeitskammer. Abgerufen am 08. Mai 2019 von http://crackedlabs.org/dl/Studie_Digitale_Ueberwachung.pdf

Dabhoiwala, A. (12. September 2018). *Hadoop for Beginners*. Abgerufen am 06. Juni 2019 von KDnuggets: https://www.kdnuggets.com/2018/09/hadoop-beginners.html

Dumbill, E. (2012). *Planning for Big Data.* Sebastopol (CA, USA): O'Reilly Media.

Fachinger, V. (30. Oktober 2018). *Big Data Analytics – Warum Sie diesen Trend nicht verpassen sollten und wie Sie selbst profitieren*. Abgerufen am 16. Mai 2019 von https://piwikpro.de/blog/was-ist-big-data-und-wie-profitieren-unternehmen-davon/

Faser, D., & Meier, A. (2016). Was versteht man unter Big Data und NoSQL? In D. Fasel (Hrsg.), *Big Data - Grundlagen, Systeme und Nutzungspotenziale* (S. 3-16). Wiesbaden: Springer Vieweg.

Fischer, O. (01. April 2010). *Verarbeiten großer verteilter Datenmengen mit Hadoop*. Abgerufen am 04. Juni 2019 von heise Developer: https://www.heise.de/developer/artikel/Verarbeiten-grosser-verteilter-Datenmengen-mit-Hadoop-964755.html?seite=all

Freiknecht, J., & Papp, S. (2018). *Big Data in der Praxis: Lösungen mit Hadoop, Spark, HBase und Hive - Daten speichern, aufbereiten, visualisieren.* München: Carl Hanser Verlag.

Gadatsch, A., & Landrock, H. (2017). *Big Data für Entscheider: Entwicklung und Umsetzung datengetriebener Geschäftsmodelle.* Wiesbaden: Springer Vieweg.

Gartner, Inc. (2019). *Gartner IT Glossary*. Abgerufen am 08. Mai 2019 von https://www.gartner.com/it-glossary/big-data

Hadoop Wiki. (5. April 2018). *Powered by Apache Hadoop.* Abgerufen am 29. Mai 2019 von Hadoop Wiki: https://wiki.apache.org/hadoop/PoweredBy

Henschen, D. (7. November 2011). *Hadoop Spurs Big Data Revolution*. Abgerufen am 29. Mai 2019 von Information Week: https://www.informationweek.com/database/hadoop-spurs-big-data-revolution/d/d-id/1101160

Hintenhaus, M. (25. Mai 2015). *Volume, Velocity und Variety: Zahl und Vielfalt von Datenquellen erst machen den Big-Data-Bestand aus*. Abgerufen am

16. Mai 2019 von https://www.bigdata-insider.de/zahl-und-vielfalt-von-datenquellen-erst-machen-den-big-data-bestand-aus-a-486927/

Hurwitz, J., & Kirsch, D. (2018). *IBM Cloud Private For Dummies®.* New York, USA: John Wiley & Sons, Inc.

IBM Corporation. (2019). *Integrating with a Hadoop cluster*. Abgerufen am 17. Juni 2019 von https://www.ibm.com/support/knowledgecenter/en/SSQNUZ_1.2.0/com.ibm.icpdata.doc/zen/admin/hadoopintegration.html

IBM Deutschland GmbH. (2019). *Analytics Engine*. Abgerufen am 14. Juni 2019 von ibm.de: https://www.ibm.com/de-de/cloud/analytics-engine

IBM Deutschland GmbH. (2019). *IBM Cloud Private: Warum Container?* Abgerufen am 17. Juni 2019 von https://www.ibm.com/de-de/cloud/private/why-containers

IBM Deutschland GmbH. (2019). *Vorteile von Cloud-Computing*. Abgerufen am 19. Juni 2019 von ibm.de: https://www.ibm.com/de-de/cloud/learn/benefits-of-cloud-computing

Joos, T. (23. September 2014). *So funktioniert Apache Hadoop*. Abgerufen am 18. Juni 2019 von Big Data Insider: https://www.bigdata-insider.de/so-funktioniert-apache-hadoop-a-457897/

Joos, T. (10. Juli 2015). *10 Dinge, die Sie über Hadoop wissen sollten*. Abgerufen am 29. Mai 2019 von Computerwoche - Voice of digital: https://www.computerwoche.de/a/10-dinge-die-sie-ueber-hadoop-wissen-sollten,3096660

Jumani, A. (01. Oktober 2017). *Hadoop Master/Slave Architecture*. Abgerufen am 06. Juni 2019 von Big Data Analytics and Its Applications - Scientific Figure on ResearchGate: https://www.researchgate.net/figure/Hadoop-Master-Slave-Architecture_fig4_320345031

Kneser, J., & Dietsche, P. (Regisseure). (2014). *Das Ende des Zufalls. Die Macht der Algorithmen* [Kinofilm]. Abgerufen am 02. Mai 2019 von http://www.3sat.de/mediathek/?mode=play&obj=71022

Lesniak, D. (14. Dezember 2017). *Big Data: 5 x V. Die großen fünf Merkmale von Big Data*. Abgerufen am 14. Mai 2019 von https://www.micromata.de/blog/big-data/big-data-v5/

Luber, S., & Litzel, N. (01. September 2016). *Was ist Hadoop?* Abgerufen am 28. Mai 2019 von Big Data Insider: https://www.bigdata-insider.de/was-ist-hadoop-a-587448/

Luber, S., & Nitzel, N. (15. Februar 2018). *Was ist ein Data Lake?* Abgerufen am 06. Juni 2019 von Big Data Insider: https://www.bigdata-insider.de/was-ist-ein-data-lake-a-686778/

Microsoft Corporation. (06. November 2018). *Verwenden leerer Edgeknoten in Apache Hadoop-Clustern in HDInsight*. Abgerufen am 17. Juni 2019 von Microsoft Azure: https://docs.microsoft.com/de-de/azure/hdinsight/hdinsight-apps-use-edge-node

Müller, S. (09. Februar 2016). *Apache Hadoop – ein bewährtes Open Source-Framework*. Abgerufen am 05. Juni 2019 von it-novum GmbH Deutschland: https://it-novum.com/blog/apache-hadoop-ein-bewaehrtes-konzept/

Nazrul, S. (24. April 2018). *CAP Theorem and Distributed Database Management Systems*. Abgerufen am 18. Juni 2019 von Towards Data Science: https://towardsdatascience.com/cap-theorem-and-distributed-database-management-systems-5c2be977950e

Ostler, U. (15. Oktober 2018). *IBM erweitert IBM Cloud Private mit Watson KI-Funktionen*. Abgerufen am 17. Juni 2019 von DataCenter Insider: https://www.datacenter-insider.de/ibm-erweitert-ibm-cloud-private-mit-watson-ki-funktionen-a-765718/

Piazzi, C. (16. Mai 2018). *Installation eines Hadoop Single Node Cluster*. Abgerufen am 04. Juni 2019 von Modius-Techblog: https://www.modius-techblog.de/big-data/installation-eines-hadoop-single-node-cluster/?cookie-state-change=1559639873487

Printz, U. (14. August 2013). *Einführung in Hadoop – Die wichtigsten Komponenten von Hadoop*. Abgerufen am 04. Juni 2019 von codecentric

AG: https://blog.codecentric.de/2013/08/einfuhrung-in-hadoop-die-wichtigsten-komponenten-von-hadoop-teil-3-von-5/

Radtke, M., & Litzel, N. (01. September 2016). *Big Data Insider*. (V. I.-M. GmbH, Herausgeber) Abgerufen am 08. Mai 2019 von Was ist Big Data?: https://www.bigdata-insider.de/was-ist-big-data-a-562440/

Schön, C. (29. April 2015). *Big Data und Hadoop: Apache macht das Unmögliche möglich*. Abgerufen am 28. Mai 2019 von Big Data Blog: https://bigdatablog.de/2015/04/29/big-data-und-hadoop-apache-macht-das-unmoegliche-moeglich/

SIGS DATACOM GmbH. (2018). *Big Data effizient nutzen: Hybrides Datenmanagement mit der IBM Cloud Private for Data.* Troisdorf: SIGS DATACOM GmbH.

TechTarget Germany GmbH. (29. September 2014). *Was sind unstrukturierte und strukturierte Daten und wie unterscheiden sie sich?* Abgerufen am 16. Mai 2019 von https://www.computerweekly.com/de/antwort/Was-sind-unstrukturierte-und-strukturierte-Daten-und-wie-unterscheiden-sie-sich

TechTarget Germany GmbH. (13. Februar 2015). *Unstrukturierte Daten*. Abgerufen am 16. Mai 2019 von https://www.computerweekly.com/de/definition/Unstrukturierte-Daten

Weichel, P., & Herrmann, J. (6. April 2016). Wie Controller von Big Data profitieren können. (P. Schäffer, Hrsg.) *Controlling & Management Review*(Sonderheft 1), S. 8-15.

Wiedermann, N. (17. Juli 2011). *Was ist eigentlich ein Petabyte ?* Abgerufen am 29. Mai 2019 von My tiny TechBlog: https://norwied.wordpress.com/2011/07/17/was-ist-eigentlich-ein-petabyte/

Zikopoulos, P., deRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., & Giles, J. (2013). *Harness the Power of Big Data.* New York: Melnyk, Roman B.