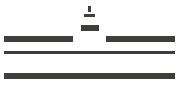


Förderkreis der
Angewandten
Informatik
an der
Westfälischen
Wilhelms-Universität Münster e.V.

Big Data in kleinen und mittleren Unternehmen — eine empirische Bestandsaufnahme

Abschlussbericht an den Förderverein des Instituts für Angewandte Informatik (IAI), Münster

Prof. Dr. Gottfried Vossen, Dr. Jens Lechtenbörger, David Fekete, M.Sc.
Institut für Wirtschaftsinformatik
Westfälische Wilhelms-Universität Münster

April 2015

Inhaltsverzeichnis

Executive Summary

Dieser Bericht untersucht das Potential von *Big Data* in und für kleine und mittlere Unternehmen (KMU). Dafür werden mittels einer Umfrage, Fokus-Interviews und zusätzlichen Recherchen die Potentiale und Herausforderungen bei Big Data untersucht, welche für KMU zu Tragen kommen.

Der Begriff *Big Data* selbst wird genauer betrachtet und es wird erklärt, welche Dimensionen abseits schierer Größe ebenfalls dazu zählen, wie etwa Vielfalt oder Schnelligkeit der Daten. Weiter wird mit eingängigen Praxisbeispielen, etwa von der US-Restaurant-Kette The Cheessecake Factory, die Nutzung von Big Data illustriert und die Anwendbarkeit hiervon auf KMU besprochen.

Die Potentiale und Herausforderungen für KMU werden in technische, organisatorische sowie wirtschaftliche und rechtliche Aspekte unterteilt und untersucht. Ein zusammenführendes Fazit für KMU wird im Anschluss gebildet und diskutiert. Insgesamt soll der Bericht helfen zu klären, ob und wie KMU sich dem Thema Big Data stellen sollten und weshalb das Thema Big Data bei KMU bisher zögerlich angenommen wird.

Es stellt sich heraus, dass es für KMUs grundsätzlich sinnvoll ist, den Blick in die Richtung von Big Data zu lenken, denn die Entwicklungen in Bereichen wie Internet of Things, Industrie 4.0, Smart Car, Smart Home oder Smart City lassen heute bereits klar erkennen, dass die Digitalisierung unserer Welt weiter fortschreiten wird. Damit werden die anfallenden Datenbestände ständig und immer schneller wachsen und neue Formen von Interaktion, Kollaboration, Kooperation, Marketing, Vertrieb und Kundenorientierung ermöglichen.

1. Einleitung

Dem Thema Big Data wird medial, in Forschung, Wissenschaft und in der Praxis spätestens seit 2013 große Aufmerksamkeit zuteil; selbst nicht-technische Medien greifen das Schlagwort laufend auf (z.B. Der Spiegel 20/2013 oder die Berichterstattung über das das PRISM-Projekt der amerikanischen NSA). Als zentrale Charakteristika werden verschiedene, im Englischen mit „V“ beginnende Eigenschaften wie Vielfalt (Variety), Volumen (Volume) und Geschwindigkeit (Velocity) von Big Data gesehen. Von „Big Data“ spricht man typischerweise dann, wenn die Kapazitäten nicht nur großer Organisationen bzgl. Datenhaltung oder der anschließenden Analyse sowohl auf technischer als auch organisationaler Ebene überschritten werden. Neben klassischen Datenbanksystemen, wie etwa MySQL, werden durch Big Data auch etablierte Data-Warehouse-Ansätze in Frage gestellt, da sie nicht die nötige Flexibilität und Geschwindigkeit bieten können, die man sich in diesem neuen Kontext wünscht. Für Big Data werden daher neue Paradigmen propagiert. Das Potential des großen und heterogenen Datenberges soll explorativ von Fachkräften mit vielfältigem Spezialwissen, den sog. „Data Scientists“, mit Hinblick auf die Unternehmensstrategie ergründet werden, um letztlich verwertbare Erkenntnisse zu erhalten. Technisch werden NoSQL-Datenbanken, Sprachen wie R, Programmier-Paradigmen wie Map-Reduce und Tools wie Apache Hadoop als Enabler gesehen.

Insbesondere die Industrie propagiert bereits vielfältige und teils komplexe Big-Data-Lösungsgefüge, welche neue Technologien, Produkte und Paradigmen in die Organisation bringen sollen. Diese interagieren teils auch mit bestehenden Systemen. Big-Data-Lösungen, die bei großen Firmen wie Google und Facebook Wirkung zeigen, sollen schon bald für jedermann geeignet und relevant sein. Jüngste Umfragen suggerieren allerdings, dass einige Unternehmen, insbesondere kleine und mittlere Unternehmen (KMUs), deren Kernkompetenz nicht im IT-Bereich liegt, sich reserviert in Bezug auf Big Data und dessen Nutzen zeigen[1]. Mittels einer Umfrage wurde in dem Projekt, über das hier berichtet wird, die Sichtweise speziell von KMUs auf Big Data erforscht. Dabei wurde speziell auf das Verständnis von Big Data, dessen Nutzen für die KMUs und wahrgenommene Herausforderungen abgezielt. Zielgruppe der Umfrage waren primär kleine und mittlere Unternehmen, wobei hierfür die Größeneinordnung der Europäischen Union[2] (EU) herangezogen wird (Umsatz nicht größer als 50 Mio. € und weniger als 250 Mitarbeiter). Es ist jedoch anzunehmen, dass sich Unternehmen beteiligt haben, die ein Interesse am

[1] http://www.heise.de/ix/meldung/Studie-Unternehmen-zoegern-bei-Big-Data-1884840.html

[2] http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index_de.htm

Thema „Big Data für KMU“ haben, selbst wenn sie nicht in den Definitionsbereich von KMU fallen. Ergänzt wurde die Umfrage mit Interviews mit Umfrageteilnehmern, um die Ergebnisse anzureichern.

Im Folgenden wird in Abschnitt 2 zunächst das Thema „Big Data“ konkretisiert und hinsichtlich einiger seiner relevanten Dimensionen erläutert, sowie exemplarische Anwendungsfälle für Big Data vorgestellt und deren Umsetzung besprochen. Darauf folgend wird in Abschnitt 3 die hier verwendete Untersuchungsmethode näher erläutert, bevor in Abschnitt 4 Potentiale und Herausforderungen anhand der Ergebnisse zunächst herausgestellt und im Anschluss daran in Abschnitt 5 diskutiert werden. Im Abschnitt 6 wird die Studie als Ganzes zusammengefasst und ein Ausblick auf weitere Forschungsgebiete geboten.

2. Hintergrund: Big Data

Neben der fortschreitenden Digitalisierung und der Tatsache, dass digitale Objekte (z. B. E-Mail-Dateianhänge oder Digitalbilder sowie -videos) im Laufe der Zeit immer größer geworden sind, hat die technologische Entwicklung immer schnelleren Datentransport sowie umfassendere Datenspeicherung ermöglicht und die Erzeugung von Daten sowohl automatisch, also durch Maschinen, als auch durch User dramatisch zunehmen lassen. In einer neueren Statistik[3] hat die Firma Intel festgestellt, dass in einer einzigen Internet-Minute rund 650 TB an IP-Daten über das Netz transportiert werden Das Ergebnis dieser Entwicklung ist so überwältigend, dass die Bezeichnung „Big Data“ angemessen erscheint; dieser Abschnitt beschreibt die Herausforderungen, Techniken und Anwendungen von Big Data und wirft dabei einen Blick auf mögliche zukünftige BI-Architekturen.

2.1. Zentrale Charakteristika und Dimensionen von Big Data

Das Volumen (engl. Volume) stellt ein erstes Charakteristikum von Big Data dar. Daten werden laut [CP14] als „big“ betrachtet, wenn sie groß genug sind, um signifikant von paralleler Verarbeitung in einer Armada von Rechnern zu profitieren, wobei die Verwaltung der Berechnung selbst eine beachtliche Herausforderung darstellt. Weitere Charakteristika von Big Data, die zusammen mit dem Volumen als die „fünf Vs“ bezeichnet werden, sind die (hohe) *Geschwindigkeit* (engl. Velocity), mit der Daten produziert werden und verarbeitet werden müssen, die (hohe) *Vielfalt* (engl. Variety), die Daten annehmen können, die (nicht immer gegebene) Präzision, Genauigkeit oder *Vertrauenswürdigkeit* (engl. Veracity), in der Daten vorliegen, und der Wert (engl. Value), den Daten (hoffentlich) darstellen. (Die ersten drei dieser Charakteristika werden dem Analysten Doug Laney[4] von Gartner zugeschrieben.) Geschwindigkeit bezieht sich auf die Tatsache, dass Daten heute oft in Form von Datenströmen auftreten, welche der die Daten analysierenden Einrichtung keine Chance lassen, die Daten zu Analysezwecken dauerhaft zu speichern; stattdessen muss die Verarbeitung unmittelbar erfolgen. Vielfalt bedeutet, dass die Daten in unterschiedlichsten Formen und Formaten auftreten, darunter als unstrukturierter Text, als semi-strukturierte Daten (z.B. XML-Dokumente) oder als strukturierte Daten (z.B. relationale Tabellen). Diese Eigenschaften von Big Data sind in Abbildung 1 zusammengefasst. Wert ist eine Eigenschaft, die Daten erst im Nachhinein, also durch angemessene Analyse, Verdichtung oder sonstige Verarbeitung erhalten.

[3] http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html
[4] http://blogs.gartner.com/doug-laney/

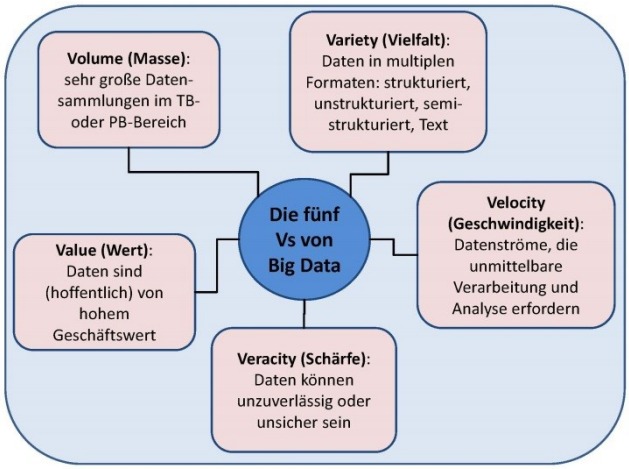


Abbildung 1: Definierende „Fünf Vs“ von Big Data..

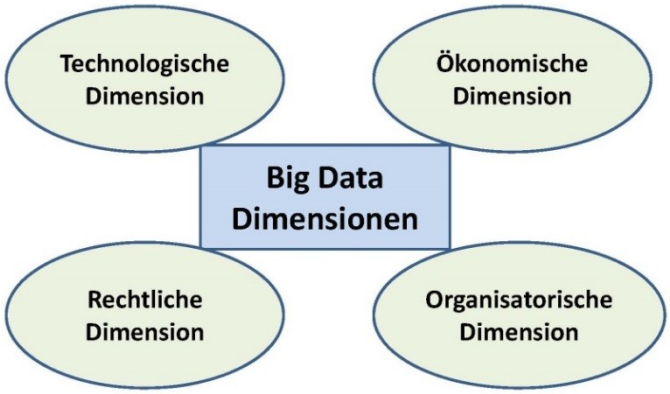


Abbildung 2: Big-Data-Betrachtungsdimensionen.

Der Begriff Big Data lässt sich aus verschiedenen Dimensionen betrachten und diskutieren[5]. Grundsätzlich lassen sich – wie in anderen Bereichen auch – die in Abbildung 2 gezeigten vier Dimensionen zugrunde legen, von denen zwei im Anhang diskutiert werden: Anhang C behandelt die technologische Dimension und mit NoSQL, NewSQL und Map-Reduce diejenigen Technologien, die vornehmlich zur Verarbeitung von Big Data eingesetzt werden. Anhang D widmet sich der organisatorischen Dimension und stellt eine Big-Data-Adoptionsstrategie vor. Die Ausführungen dieser Anhänge sind angelehnt an [Vos13, Vos14, VL15]. Die wirtschaftliche Dimension ist je nach Anwendungsfall individuell zu betrachten und interessiert sich primär für den (Mehr-) Wert, den man als Unternehmen durch eine Beschäftigung mit Big Data erzielen kann. Die rechtliche Dimension schließlich befasst sich mit Fragen des Datenschutzes, der Datensicherheit, der Urheberschaft und verwandten Gebieten ähnlich denen, die man im Cloud Computing betrachtet [VHH12].

[5] `http://datascience.berkeley.edu/what-is-big-data/`

2.2. Anwendungen von Big Data

Als nächstes sollen verschiedene Aspekte von Big Data im Einsatz in einer Organisation beleuchtet werden. Dafür werden zunächst exemplarisch typische Anwendungsfälle von Big-Data-Nutzung in Großunternehmen vorgestellt, um zu illustrieren, wie Big Data in der Praxis eingesetzt wird. Zum Abschluss wird die Anwendbarkeit der typischen Anwendungsfälle von Big Data bei KMU diskutiert. Dass Big Data nicht ein reiner Selbstzweck ist, veranschaulichen die Beispiele von Großunternehmen, die Big-Data-Technologien einsetzen und damit Erfolge erzielen, die sie mit konventionellen Technologien nicht auf diese Art realisieren hätten können. Die Cheesecake Factory aus den Vereinigten Staaten, ein Vertreiber von Käsekuchen und Betreiber der gleichnamigen Restaurant-Kette ebenda, verwendet Big-Data-Technologien, um konventionelle, relationale Daten mit unstrukturierten Informationen zu verbinden. Konkret werden strukturierte Handelsdaten (d.h., z.B. Verkäufe) kombiniert mit Feedback aus sozialen Netzwerken sowie digital erfassten, soften Qualitätskriterien von Lieferungen, wie etwa Temperatur, Geruch und Geschmack. Falls es z.B. bei einer Lieferung eine Geschmacksabweichung beim Senf gibt, kann diese Information mit entsprechenden Lieferantendaten korreliert und so z.B. ein Qualitätsproblem bestimmter Chargen eines Lieferanten ausgemacht werden. Das Beispiel zeigt, wie durch Hinzunahme zusätzlicher Information zu den bekannten kaufmännischen KPIs auch die Produktqualität besser verstanden und gegebenenfalls erhöht werden kann. Auch deutsche Großunternehmen setzen bereits Big-Data-Technologien ein, wie etwa die Drogeriemarktkette „dm“ oder der Versandhändler „Otto“.

Eines der ältesten Beispiele für das, was man mit einer sorgfältigen Analyse von – seinerzeit nicht einmal als „Big“ bezeichneten – Daten erreichen kann, stammt vom amerikanischen Baseball-Team der Oakland Athletics und ihrem Coach Billy Beane, der Statistiken und Spieler-Daten dazu nutzen konnte, durch „Einkauf“ unbekannter Spieler innerhalb kurzer Zeit ein wenig erfolgreiches Team zu einem erfolgreichen zu machen. Dieses Beispiel ist gut dokumentiert [Lew04] und wurde sogar unter dem Titel “Moneyball” mit Brad Pitt in der Hauptrolle verfilmt. Ein weiteres Beispiel aus dem Sport liefert das Indianapolis 500-Meilen-Rennen, das jährlich am Memorial Day Wochenende in den USA stattfindet. Ein moderner Indy 500-Rennwagen ist mit rund 200 Sensoren ausgestattet, die permanent die Motorleistung, Kupplung, Schaltung, Differential, Kraftstoffversorgung, Öl, Lenkung, Reifen, DRS und zahlreiche andere Komponenten und sogar den Gesundheitszustand des Fahrers überwachen. Sie produzieren rund 1 GB an Telemetrie-Daten pro Rennen, die von Renningenieuren während und nach einem Rennen analysiert werden. Ähnlich bei McLaren, wo Rechner rund eintausend Simulationen während eines

Rennens durchführen; nach nur wenigen Runden kann man dadurch die Leistung jedes Systems mit bis zu 90%-iger Genauigkeit vorhersagen. Da die meisten dieser Systeme sogar während eines Rennens verändert werden können, können Boxenmannschaft und Fahrer minutengenaue Anpassungen vornehmen, wenn sich Veränderungen am Auto oder an den Rennbedingungen ergeben. Details findet man z.B. im Blog von Doug Laney[6], und die Situation ist bei Formel-1-Autos[7] oder der NASCAR-Serie ähnlich.

Ein Beispiel aus dem Rettungswesen lieferte im Herbst 2012 der Hurrikan Sandy[8] in der Karibik und an der amerikanischen Ostküste. Die Organisation Direct Relief nutzte Big-Data-Technologie zur Koordination von Rettungsaktivitäten, was auf ihrer Webseite[9] nachzulesen ist. Mit Analyse- sowie Kartensoftware war man in der Lage, Bedarfe zu erfassen und geeignete Maßnahmen zu initiieren; u.a. konnte man Kliniken, die über wichtige medizinische Ressourcen verfügten, miteinander verbinden, um so z.B. Stromausfälle oder Medikamentenmängel zu kompensieren.

Ein weiteres Feld, welches sich dank Big Data endlich bewegen wird, ist die Hausautomatisierung, mit der man sich forschungsseitig bereits seit mehr als 10 Jahren beschäftigt [Vos01]. Hier ist zu erwarten, dass man durch die Fähigkeit, Daten von Klimaanlagen, Heizung, Licht oder Haushaltsgeräten wie Waschmaschine, Trockner oder Kühlschrank zusammen mit Informationen über die in dem betreffenden Haus lebenden Personen zu verarbeiten, Lebensbedingungen schaffen kann, die optimal an die jeweilige Alters- oder Gesundheitssituation angepasst sind.

Auch das Gesundheitswesen verändert sich durch Big Data, etwa dadurch, dass man durch Sammlung und Analyse von Daten über einen Patienten, dessen tägliche Aktivität und dessen Ernährung sowie Informationen von einem Medikamentenhersteller, ggfs. sogar unter Rückgriff auf das Genom, also die Erbinformation des Patienten, ideal konfigurierte Behandlungen schaffen kann. Die zunehmende Verbreitung von persönlichen Trackern wie Fitbit, das Nike+ Fuelband oder das Jawbone Up wird weitere Datenquellen liefern, die für Gesundheitsexperten wie Nutzer gleichermaßen von Interesse sind [Mac15].

Andere Gebiete, die sich bereits intensiv mit der Analyse von Big Data befassen, sind Marktforschung, Verkehrssteuerung (etwa in Ländern wie Singapur) oder autonome sowie vernetzte Fahrzeuge, die selbst fahren und mit anderen Fahrzeugen kommunizieren. Als Beispiel aus der Unterhaltungsindustrie hat Disney Parks & Resorts das MyMagic+ System

[6] http://blogs.gartner.com/doug-laney/the-indy-500-big-race-bigger-data/
[7] http://www.quantumblack.com/formula-1-race-strategy-2/
[8] http://de.wikipedia.org/wiki/Hurrikan_Sandy
[9] http://www.directrelief.org/emergency/hurricane-sandy-relief-and-recovery/

entwickelt, das über die Webseite My Disney Experience und die zugehörige mobile App aktuellste Informationen über Angebote an potentielle Besucher eines Disney-Parks liefern kann. Disneys MagicBand[10] kann von Gästen als Zimmerschlüssel in einem der Park-Hotels, Eintrittskarte in den Park, Zutritt zu FastPass+ Eingängen oder zum Einkaufen in Park-Shops verwendet werden. Besucher, die an diesem Programm teilnehmen, können damit Warteschlangen überspringen, Vorausreservierungen vornehmen und später über ihr Smartphone ändern, und sie werden von Disney-Figuren mit Namen begrüßt. Das MagicBand unterliegende System sammelt Daten über den Besucher, dessen aktuellen Ort, seine Kaufhistorie sowie die Historie der besuchten Attraktionen im Park. Auch für eine Vorhersage der Oscar-Preisträger der amerikanischen Filmindustrie verwendet man heute sogar recht erfolgreich Big Data-Techniken[11].

Soziale Medien sowie Suchmaschinen sind ebenfalls intensiv mit einer Analyse der Daten, die in ihrem Kontext anfallen, befasst. Twitter analysiert beispielweise die Tweets seiner Nutzer, um Nutzergruppen zu identifizieren und zu vergleichen, um Nutzergewohnheiten zu erkennen oder zur Durchführung von Stimmungsanalysen bei Text-Tweets. Facebook interessiert sich für die Anzahl an „Likes“, die eine Seite im Laufe der Zeit bekommt, und unterhält einen Zähler für empfohlene URLs; zwischen einem Click und einer Aktualisierung des Zählers dürfen nicht mehr als 30 Sek. vergehen. Google schließlich führt Clustering von Texten in Google News durch und versucht, ähnliche Nachrichten nahe beieinander zu platzieren; Google klassifiziert ferner E-Mails in Gmail und wendet darauf verschiedene Analysen an, etwa im Zusammenhang mit AdWords.

Es stellt sich in allen diesen Anwendungsbereichen heraus, dass unterschiedliche Kenntnisse und Fertigkeiten gefordert sind: Es reicht nicht, allein die notwendige Rechentechnik zu kennen oder die statistischen Methoden zu beherrschen; entscheidend ist häufig der „richtige Mix“ an Kenntnissen aus den Bereichen IT und Informatik, Statistik, Betriebswirtschaft, Anwendungsbereich, kombiniert mit Problemlösungskompetenz und Kommunikationsfähigkeit. IBM beschreibt die neue Spezies des „Data Scientist“ so[12]: „A data scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems, they will pick the right problems that

[10] https://disneyworld.disney.go.com/faq/bands-cards/understanding-magic-band/

[11] http://www.popsci.com/data-still-beats-expertise-oscar-predicting

[12] http://www-01.ibm.com/software/data/infosphere/data-scientist/

have the most value to the organization."

2.3. Anwendbarkeit für kleine und mittlere Unternehmen

Während Großunternehmen typischerweise über eigene IT-Abteilungen verfügen und auf eine umfassende Infrastruktur zurückgreifen können, sind kleine und mittlere Unternehmen oft in einer anderen Situation: Man ist gezwungen, sich auf die Kernkompetenzen und das Kerngeschäft zu konzentrieren und hat kaum Gelegenheit, die Möglichkeiten neuer Technologien umfassend zu evaluieren. Der Bitkom hat in einer im Herbst 2014 veröffentlichten Studie[13] hierzu herausgefunden: „Neun von zehn Unternehmen werten IT-gestützt Daten für Entscheidungsprozesse aus. Dabei ist das Spektrum der analysierten Daten breit. Stammdaten (36 Prozent), Transaktionsdaten (33 Prozent) und Logdaten (31 Prozent) werden am häufigsten genannt. Auch Sensor- und CRM-Daten werden vergleichsweise häufig ausgewertet (25 bzw. 14 Prozent). Texte/Publikationen (9 Prozent), Web Content (9 Prozent) und Social Media (8 Prozent) folgen mit Abstand." Aber: „Fast jedes zweite Unternehmen verzichtet bewusst auf bestimmte Datenanalysen. Neben der Sorge vor Kritik durch Kunden (31 Prozent) fürchten viele hohe Kosten (23 Prozent) sowie Imageschäden in der Öffentlichkeit (23 Prozent). Ethisch-moralische Gründe spielen für jedes siebte Unternehmen eine Rolle (14 Prozent)." Eine weitere wichtige Erkenntnis dieser Studie war, dass Mittelständler typischerweise bereits verstärkt in Speicherplatz investiert haben, Großunternehmen dagegen eher in Analyse-Tools.

Bereits zwei Jahre zuvor hatte das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) in einer ähnlichen Studie Nutzung und Potenzial für Big Data in deutschen Unternehmen untersucht[14]. Es zeigt sich dabei, dass einerseits Neugier gegenüber dem Thema besteht, andererseits aber auch Unsicherheit bzgl. vieler damit zusammenhängender Fragen. In der vorliegenden Ausarbeitung wird die These vertreten, dass eine Beschäftigung mit und Nutzung von Big Data nicht von der Unternehmensgröße abhängig sollte, sondern für alle Unternehmen interessant ist, die ihre Kunden besser kennenlernen wollen und ihre Produkte noch gezielter als bisher entwickeln und vermarkten wollen. Eine Limitierung von In-House-Ressourcen wird dabei zwar häufig als Hemmschuh gesehen, ist aber durch eine Cloud-Orientierung leicht zu kompensieren. Entscheidend wird für die nahe Zukunft sein, durch Aufbau von entsprechendem Know-How unter den Mitarbeitern für fachgebietsübergreifende Aktivitäten auf diesem Gebiet gerüstet zu sein.

[13] http://www.bitkom.org/files/documents/Studienbericht_Big_Data_in_deutschen_Unternehmen.pdf

[14] http://www.iais.fraunhofer.de/bigdata-umfrage.html

3. Untersuchungsmethode der Studie

Im Folgenden wird erörtert, wie das Potential von Big Data in und für KMU empirisch mittels einer Umfrage und Fokus-Interviews untersucht wurde. Hintergrund ist die eingangs erwähnte Unsicherheit gerade bei KMUs, wie sie mit diesem Thema umgehen sollen und ob es für sie überhaupt von Bedeutung ist. Dafür wird zunächst die Datenerhebung erörtert. Anschließend wird dargelegt, wie die Datenanalyse der empirischen Daten zusammen mit den bisherigen Erkenntnissen zu Big Data (vgl. z.B. Abschnitt 2.3) erfolgt.

3.1. Datenerhebung

Die Datenerhebung für die Untersuchung fand mittels einer Umfrage sowie darauf aufbauenden Fokus-Interviews statt, sodass quantitative und qualitative Daten erfasst werden konnten.

3.1.1. Umfrage

Um das Potential von Big Data in und für KMU zu erheben, wurde eine elektronische Umfrage aufgesetzt. Diese wurde in Zusammenarbeit mit der IHK Münster über folgende Quellen an potentielle Interessenten verbreitet und war von Anfang April bis zum 23. Mai 2014 verfügbar.

- Schriftlicher Verteiler der IHK Münster (1444 Unternehmen per Post, ca. 44 per E-Mail)
- Schriftlicher Verteiler des IAI Münster (ca. 40 Mitglieder); ab 7. Mai 2014.
- Ankündigung auf Webseite des Instituts für Wirtschaftsinformatik der Universität Münster

Die Umfrage wurde von insgesamt 33 Interessenten explizit gestartet, wovon 24 mindestens eine Antwort abgegeben haben. Letztere Zahl wird daher als „Anzahl der Umfrageteilnehmer“ bestimmt. Insgesamt 13 Teilnehmer haben die Fragen vollständig beantwortet (ausgenommen Fragen zum Umsatz und zur eigenen Abteilung). 11 Teilnehmer haben den Fragenbogen nicht vollständig ausgefüllt. Nicht vollständig ausgefüllt bedeutet, dass diese nicht bis zur letzten Frage vorgedrungen sind, sondern z.B. die Webseite vorzeitig geschlossen haben. Die Fragen aus der Umfrage sind in Anhang A: Fragebogen der elektronischen Umfrage zusammengestellt.

Die primäre Zielgruppe der Umfrage waren zwar KMU, andere (d.h. große) Unternehmen wurden allerdings nicht daran gehindert abzustimmen. 14 Umfrageteilnehmer haben die entsprechenden Fragen beantwortet, die zur Einordnung des Unternehmens in die EU-Größenkategorien erforderlich sind (Mitarbeiteranzahl und Umsatz). Es sind demnach 9 Umfrageteilnehmer als Großunternehmen nach EU-Definition einzuordnen und 5 als KMU. Zu den weiteren 10 Umfrageteilnehmern kann keine weitere Aussage getroffen werden. Anzumerken ist allerdings, dass unter den Großunternehmen auch (Software-)Dienstleister sein könnten, die ihre Dienste KMUs anbieten. Von den 9 Großunternehmen sind 3 explizit als Dienstleistungsunternehmen einzuordnen. Die Antworten der 9 Großunternehmen werden ebenfalls in die Auswertung mit einbezogen. Kleinstunternehmen mit weniger als 10 Beschäftigten und einem Umsatz von weniger als 2 Mio. €, die nicht zu KMU gezählt werden, waren anhand vorliegender Daten nicht unter den Teilnehmern zu identifizieren.

Die Response-Rates der Umfrage (bzw. die Nicht-Teilnahme der angeschriebenen Parteien), die Vollständigkeit der Angaben sowie die Verteilung der Teilnehmer innerhalb der Unternehmensgrößenklassen werden im Rahmen der abschließenden Analyse gesondert betrachtet.

3.1.2. Fokus-Interviews

Nachdem die elektronische Umfrage geschlossen war, wurden ausgewählte Umfrageteilnehmer persönlich, schriftlich oder telefonisch tiefergehend zum Umfragethema interviewt. Dazu wurde ein semi-strukturierter Fokus-Fragebogen entworfen, der aus 10 Fragen bestand und auch vorläufige Erkenntnisse aus der elektronischen Umfrage berücksichtigte und die Teilnehmer dazu befragte (siehe Anhang B: Fokus-Interview-Fragebogen). Die Antworten auf die Fragen konnten frei formuliert werden. Ziel war es, durch freie Antworten eine qualitativ tiefere Betrachtung von Big Data in KMU zu ermöglichen, als dies allein durch strukturierte Einschätzungen (z.B. mittels einer Likert-Skala wie in der elektronischen Umfrage) möglich ist. Weiterhin waren so Fragen möglich, die genauer auf die konkrete Situation im Unternehmen eingehen. Von allen Umfrageteilnehmern wurden diejenigen 11 gezielt per E-Mail angefragt, die ihre E-Mail-Adresse in der elektronischen Umfrage hinterließen. Ein Teilnehmer beantworte den Fokus-Fragebogen schriftlich per E-Mail und zwei Teilnehmer beantworteten die Fokus-Fragen telefonisch. Im letzteren Fall wurden die Antworten vom Interviewer schriftlich zusammengefasst. Alle drei Interview-Teilnehmer sind als Großunternehmen einzustufen, da alle drei mindestens 250 Mitarbeiter haben, wobei eines mehr als 50 Mio. € Umsatz erwirtschaftete, eines zwischen 10 und 50 Mio. € und zum dritten keine Angaben vorliegen.

3.2. Datenanalyse

Der Hauptfokus der Datenanalyse betrifft die strukturierten Ergebnisse der elektronischen Umfrage. Diese werden ergänzt oder abgewogen mit den Antworten in den drei Fokus-Interviews.

Die Response-Rate, bei angenommenen 1528 einmalig angeschriebenen Parteien, beträgt bei 24 Umfrageteilnehmern 1,6% oder niedriger. Dies ist zwar ein vergleichsweise niedriger Wert [CHT00], aber nicht grundsätzlich überraschend, betrachtet man beispielsweise auch die in letzten Jahrzehnten systematisch fallenden Response-Rates bei anderen Umfragearten (z.B. vgl. [Dey97, S. 216]). Die absolute Anzahl der Teilnehmer sowohl bei der Umfrage (N=24) als auch bei den Fokus-Interviews (N=3) ist ebenfalls relativ gering. Aus diesem Grund sind definitive Schlüsse oder Kausalzusammenhänge, insbesondere in Bezug auf KMU, im Allgemeinen aus dem vorliegenden Material nur bedingt ableitbar. Stattdessen soll die Auswertung Tendenzen zu den jeweiligen Fragen aufzeigen, um so weitere Untersuchungen zu ermöglichen. Weiterhin ist die Auswahl der angeschriebenen Parteien fokussiert auf den Bereich der IHK Nord-Westfalen sowie des IAI aus Münster und ist deren Größenzusammensetzung nicht bekannt und kann daher nicht repräsentativ die Grundgesamtheit der KMU in Nordrhein-Westfalen oder Deutschland abdecken.

Zur Analyse der Umfrage werden die jeweils gegebenen Antworten nach Häufigkeit strukturiert, sodass ersichtlich wird, welche Antworten wie häufig gegeben wurden und inwiefern (eindeutige) Tendenzen zu einer oder mehreren Antwortmöglichkeiten innerhalb der Umfrageteilnehmer auszumachen sind. Sofern möglich werden diese Ergebnisse mit weiteren Angaben, wie etwa Unternehmensgröße korreliert.

Zum Abschluss der Analyse werden die empirischen Ergebnisse mit den bisherigen Erkenntnissen über Big Data in Deutschland, insbesondere zu Big Data bei KMU, in Zusammenhang gebracht und zu einem Gesamtbild verdichtet, um ein abschließendes Fazit für die Studie zu bilden.

4. Identifizierte Potentiale und Herausforderungen

In diesem Abschnitt werden die empirischen Umfrage- und Fokus-Interview-Ergebnisse strukturiert nach den zuvor vorgestellten Dimensionen (technisch, organisatorisch, wirtschaftlich, rechtlich) von Big Data präsentiert und analysiert. Aufbauend darauf werden mögliche Potentiale und Herausforderungen von Big Data auf diesen Ebenen herausgestellt und diskutiert.

4.1. Technische Herausforderungen

Die Umfrageteilnehmer wurden gebeten, auf einer Skala von 0 (nicht relevant) bis 5 (sehr relevant) einzuordnen, welche technischen Rahmenbedingungen für oder gegen die Nutzung von Big Data sprechen. Für Big Data sind insbesondere die Rahmenbedingungen „Integration in eigene Infrastruktur“ und „Integration in bestehende Software“ zu nennen, wo mehr als 50% der abgegebenen Stimmen diese mit „4“ oder „5“ einordneten (vgl. Abbildung 3). Diese Tendenz ist auch bei der Rahmenbedingung „Performance“ beobachtbar, wenn auch weniger stark ausgeprägt. Für die „Integration in andere Angebote“ ist keine ausschlaggebende Tendenz ablesbar. Die beiden Interview-Partner, die am Telefon befragt wurden, betonten die Wichtigkeit der Integration auf technischer Ebene von Big-Data-Software im Speziellen und neuen Lösungen im Allgemeinen in die eigene Infrastruktur.

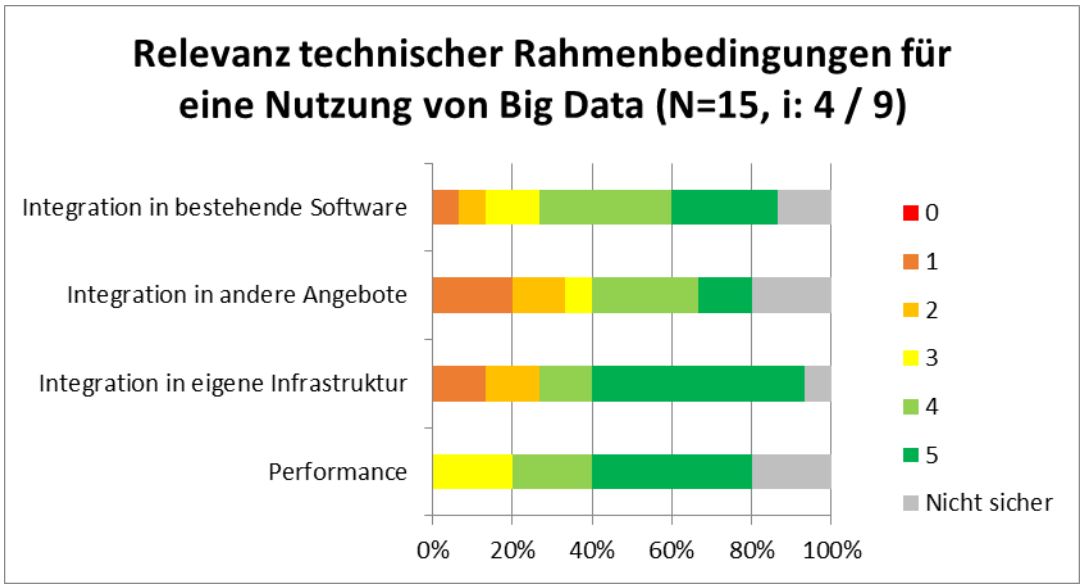


Abbildung 3: Relevanz technischer Rahmenbedingungen für eine Nutzung von Big Data. Technische Rahmenbedingungen sind hier 4 von 9 Rahmenbedingungen für die Nutzung von Big Data insgesamt.

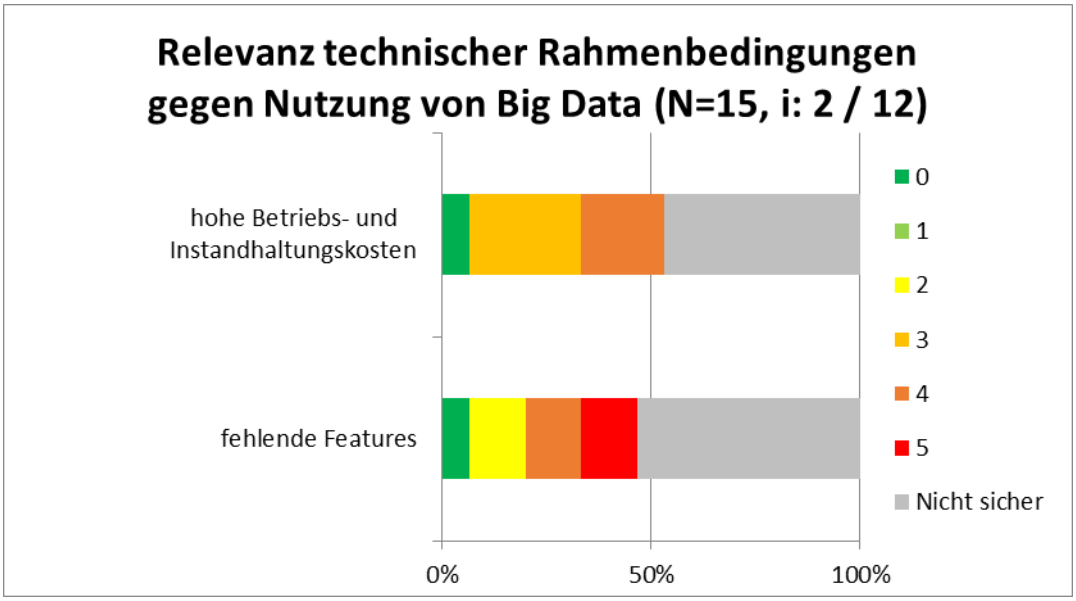


Abbildung 4: Relevanz ausgewählter technischer Rahmenbedingungen gegen eine Nutzung von Big Data. Technische Rahmenbedingungen sind hier 2 von 12 Rahmenbedingungen gegen die Nutzung von Big Data insgesamt.

Für technische Rahmenbedingungen gegen Big Data konnte für beide Kriterien, die Big-Data-Technologie betreffen, keine Tendenz ausgemacht werden (vgl. Abbildung 4). Bei beiden Rahmenbedingungen waren sich ca. 50% der Umfrageteilnehmer auch „nicht sicher“, wie relevant diese Rahmenbedingungen als Argument gegen Big Data sind. Unter den Umfrageteilnehmern lässt sich für technische Rahmenbedingungen festhalten, dass die Umfrageteilnehmer eine höhere Konfidenz bei der Einschätzung positiver Rahmenbedingung für Big Data besitzen als solcher dagegen.

Während nur ein geringer Teil der Teilnehmer (20%) bereits regelmäßig Big-Data-Software einsetzt (vgl. Abbildung 5), ist für den Rest ein Einsatz solcher gerade einmal mittel- bis langfristig (1 bis über 2 Jahre in der Zukunft) oder überhaupt nicht absehbar.

Die Umfrageteilnehmer setzen beinahe durchweg am Markt übliche, traditionelle Technologien für ihre BI- und Analyse-Aktivitäten ein: Tabellenkalkulationen (wie etwa Microsoft Excel; 92%) und klassische SQL-Datenbanken (77%). Messbar vertreten sind auch Data-Warehouse-Lösungen (54%), während andere spezialisierte und neuartige Lösungen, wie etwa Column Stores oder Produkte mit In-Memory-Technologie, nur vereinzelt (einfache bis zweifache Nennungen) oder gar nicht von den Teilnehmern genutzt werden (vgl. Abbildung 6).

Beim hauptsächlich genutzten Software-Typ, welcher für „Datenbanken, Reporting und BI“ im Unternehmen eingesetzt wird, lässt sich keine prägende Tendenz unter den

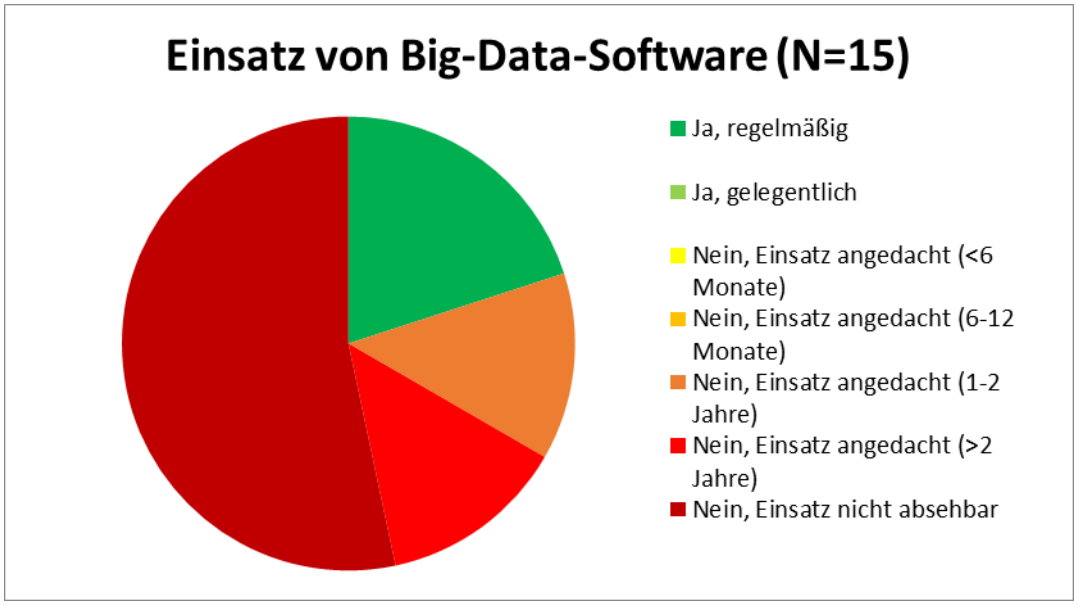


Abbildung 5: Einsatz von Big-Data-Software.

Antworten der Umfrageteilnehmer ausmachen (vgl. Abbildung 7). Sowohl Individual- als auch angepasste sowie Standard-Software und eine Mischung dieser Typen wird mit jeweils ähnlicher Häufigkeit eingesetzt. Generell findet sich unter den Umfrageteilnehmern Client/Server-basierte Software, die in Rechenzentren gehostet wird (vgl. Abbildung 8) Softwarezugriff erfolgt hauptsächlich sowohl über Desktop- als auch mobile Rechner (vgl. Abbildung 9), wobei der Zugriff auf zumindest einige Software auch über Web-Browser erfolgt (vgl. Abbildung 10). Die Umfrageteilnehmer sind also schon mit verteilter Softwarebereitstellung über Web an ihre üblichen festen oder mobilen Arbeitsrechner vertraut.

Die Ergebnisse aus der Umfrage und den Fokus-Interviews stützen die vermutete zögerliche Haltung der Unternehmen zu Big Data. Klassische Tools herrschen unter den Umfrageteilnehmern bisher vor, Big Data befindet sich noch im Experimentierstadium. Auf technischer Ebene ist für die Teilnehmer insbesondere die Integration in die eigene Infrastruktur und damit Software-Landschaft ein Kernthema. Dass Performanz auf technischer Ebene erwartet wird, deckt sich mit den üblichen Performanz-Versprechen der kommerziellen Big-Data-Produkte. Auf technischer Seite wurden auch vereinzelt Unterhaltungskosten und fehlende Features als Hinderungsgrund genannt. Auch wenn bei den Teilnehmern keine eindeutige Tendenz zu Kosten und fehlenden Features erkennbar ist, würden diese Aspekte dennoch zu einem recht jungen Softwareumfeld passen, wie Big Data es ist, wo neuartige Software noch nicht die Reife und Integrationspotenziale der länger am Markt befindlichen Produkte aufweist. Auch wären Unsicherheiten bei der Kostensituation bei

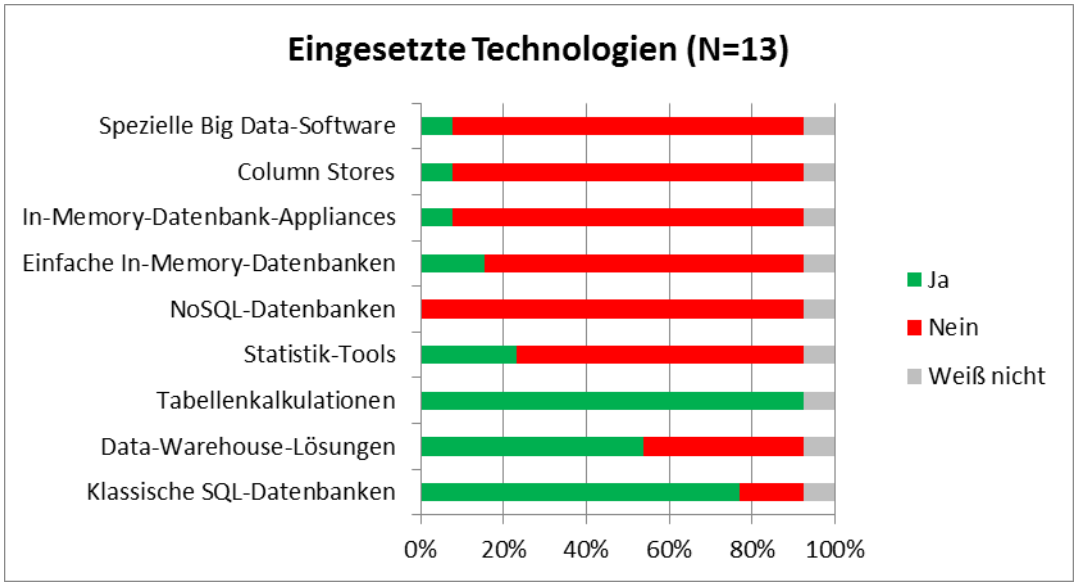


Abbildung 6: Eingesetzte Technologien bei den Umfrageteilnehmern.

neuartigen und unbekannten Produkten nicht grundlegend überraschend, da unter anderem auch Erfahrungswerte nicht vollständig vorliegen.

Technisch bietet sich jedoch bereits die Erfahrung der Umfrageteilnehmer mit verteilter Software als Grundlage für den Einsatz von Big Data an, da verteilte Architekturen die Grundprämisse vieler solcher Produkte wie etwa Apache Hadoop bilden. Die Bedenken hinsichtlich einer Integration in bestehende Infrastrukturen und Produktlandschaften sind hingegen noch näher zu untersuchen. In zwei Telefoninterviews mit Großunternehmen betonten diese jedoch die Herausforderungen, alle bisherigen, strukturierten Datenquellen im Unternehmen überhaupt zusammenzuführen – dies sei eine zentrale, noch ungelöste Aufgabe, bevor man weiterreichende Big-Data-Infrastrukturen aufsetzen könne.

Hier kann zusätzlich die stetig fortschreitende Produktentwicklung in diesem dynamischen Produktumfeld behilflich sein, die bestehende Nachteile durch rasche Produktzyklen angeht. Einen weiteren Beitrag können wissenschaftlich fundierte Use Cases und Kataloge bilden, die die zusammenhängende Nutzung vieler heterogener Analyseprodukte im Rahmen einer BI-Initiative untersuchen und näher betrachten, sodass diese auch als Leitfaden dienen können.

4.2. Organisatorische Rahmenbedingungen

Die Umfrageteilnehmer schätzten ihre Kenntnisse in Bezug auf Big Data (auf einer Skala von 0 (keine Kenntnisse) bis 5 (sehr gute Kenntnisse)) jeweils unterschiedlich ein, sodass sich ein heterogenes Erkenntnisfeld ergibt (vgl. Abbildung 11). Allerdings schätzte die-

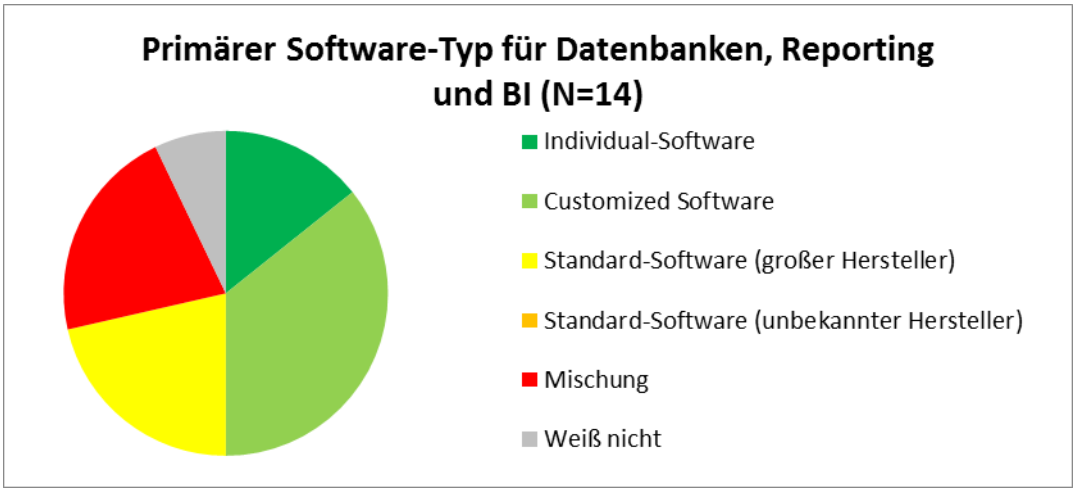


Abbildung 7: Primär eingesetzter Software-Typ für Datenbanken, Reporting und BI bei den Umfrageteilnehmern.

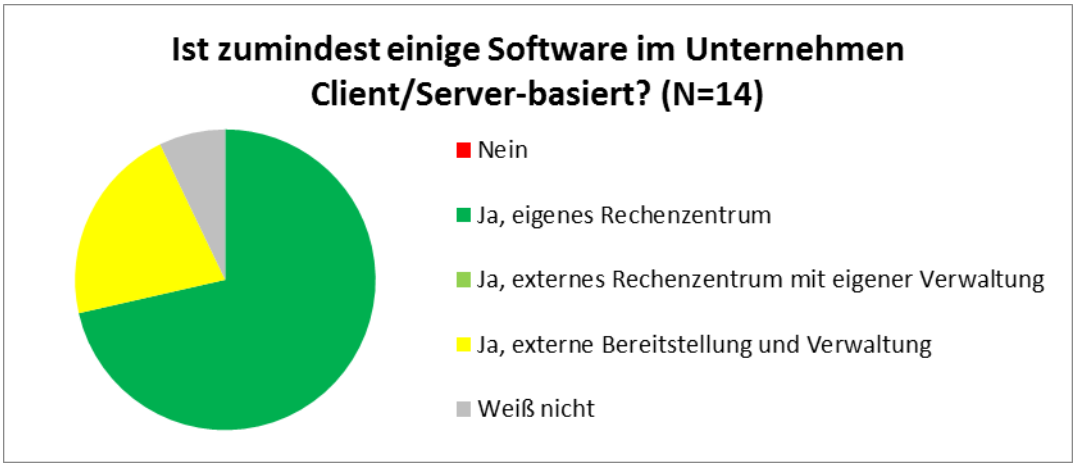


Abbildung 8: Client-/Server-basierte Software bei den Umfrageteilnehmern.

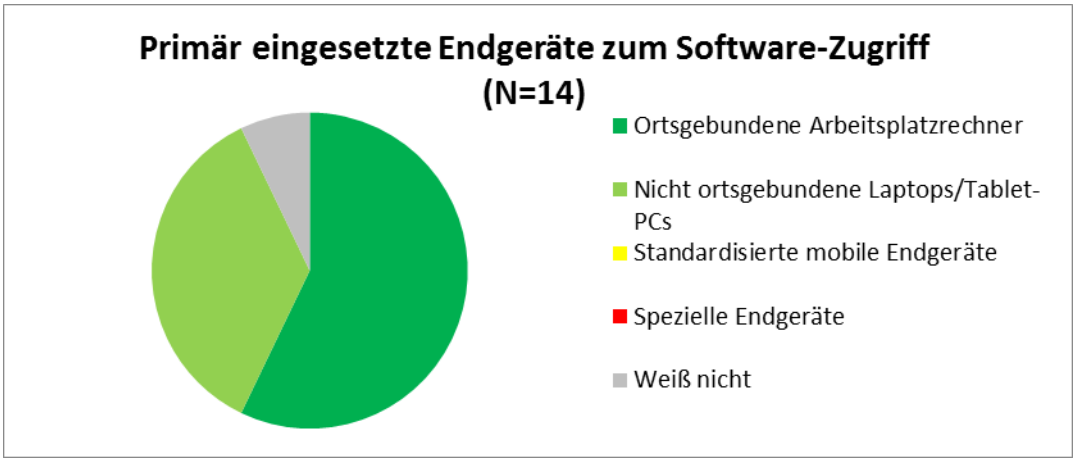


Abbildung 9: Endgeräte im Einsatz bei den Umfrageteilnehmern.

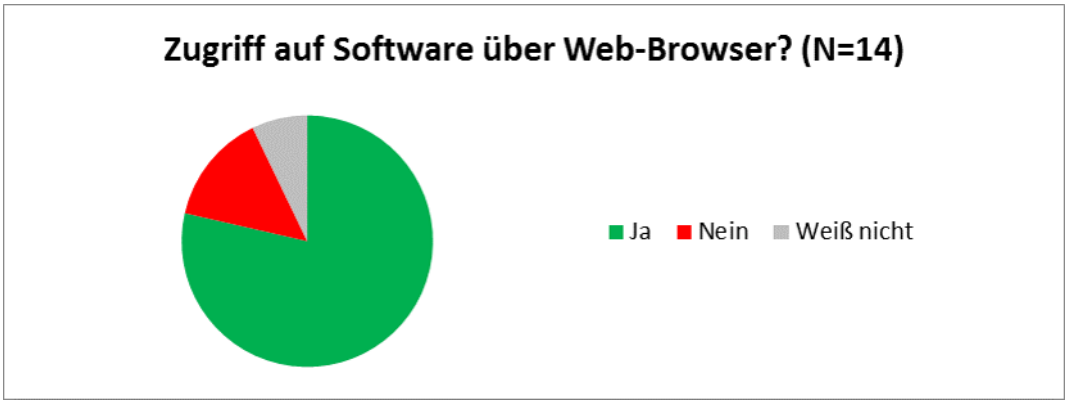


Abbildung 10: Zugriff auf Software über Web-Browser bei den Umfrageteilnehmern.

ses Feld tendenziell Big Data als wenig einfach, aber dennoch als tendenziell attraktiv ein. Bei den anderen Eigenschaften gab es kein einstimmiges Meinungsbild (vgl. Abbildung 12). Allerdings passt die Einschätzung „nicht einfach, aber attraktiv“ zur vermuteten zögerlichen Haltung bei gleichzeitigem Interesse am Thema Big Data in KMU bzw. den Umfrageteilnehmern.

Auch wenn die Vertraulichkeit der bisherigen Big-Data-Produkte nicht eindeutig geschätzt wurde (vgl. Abbildung 12), wurde dies beinahe einstimmig als sehr wichtige Rahmenbedingung für den Einsatz von Big Data genannt (vgl. Abbildung 13). An dieser Stelle sind die Hersteller von Big Data gefragt, das nötige Maß an Vertraulichkeit zu schaffen, damit diese Rahmenbedingung erfüllt werden kann.

Bei der Einschätzung organisatorischer Rahmenbedingungen, die gegen eine Nutzung von Big Data sprechen, herrschte, wie bei den technischen Rahmenbedingungen, Unsicherheit bei der Fragenbeantwortung. Teilweise bis zu 50% der Beantworter waren sich bei ihrer Einschätzung „nicht sicher“ (vgl. Abbildung 14). Contra-Tendenzen bei den Faktoren lassen sich weiterhin bei „keine Zeit“ und „fehlende Kompetenzen im Unternehmen ausmachen“, die überwiegend mit 3 oder höher eingeschätzt wurden. Dass Kompetenzen in Unternehmen fehlen würden, wäre angesichts der vielen neuen Produkte und neuartiger

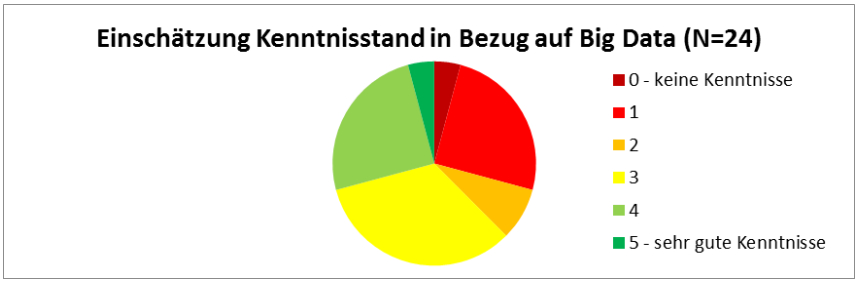


Abbildung 11: Selbsteinschätzung der Umfrageteilnehmer in Bezug auf Big Data.

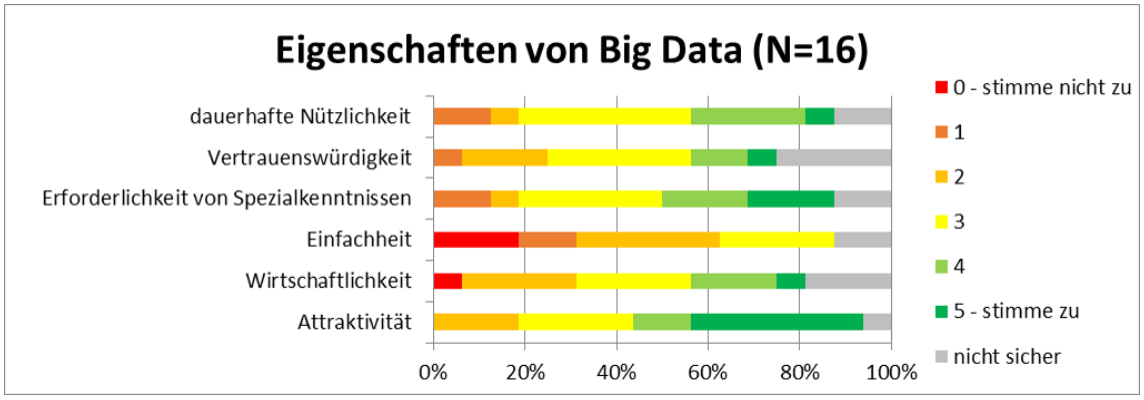


Abbildung 12: Einschätzung verschiedener Eigenschaften von Big Data durch die Umfrageteilnehmer.

Paradigmen mit Big Data verständlich. Andererseits muss dann allerdings, bei Entscheidung für Big Data, die notwendige Kompetenz geschaffen oder beschafft werden. Obwohl der Faktor Zeit bei den Umfrageteilnehmern relevant ist, ist keine entscheidende Aussage zu fehlenden Kompetenzen am Markt getroffen worden. Hier wurden aber auch negative Stimmen abgegeben. Zu betrachten wäre noch die Begründung, warum die Teilnehmer und Unternehmen keine Zeit für das Thema aufwenden wollen. Eine Vermutung ist, dass der Aufwand als nicht lohnenswert betrachtet wird. Dieser Aspekt wird ihm Rahmen der wirtschaftlichen Betrachtung erneut aufgegriffen. Was laut den Teilnehmern tendenziell nicht gegen Big Data spricht, ist dessen Fokus auf Statistik.

Innerhalb der Unternehmungen der Teilnehmer sind überwiegend einzelne Fachabteilungen für die Analyse von Daten zuständig (vgl. Abbildung 15). Wenige Nennungen (jeweils 2) erhielten die IT- bzw. Datenbank- sowie eigene Analytik-Abteilungen. Die in einem der Telefon-Interviews getätigte Aussage zum allgemeinen Problem der Datenintegration der verschiedenen Quellen im Unternehmen unterstreicht das Bild, dass unter den Umfrageteilnehmern dezentral analysiert wird. Sollte dies überwiegend und auch bei KMU der Fall sein, wäre die zögerliche Herangehensweise an Big Data besser erklärbar, denn insbesondere die explorative Analyse von Big Data beruht auf dem Zusammenbringen von vielen, abteilungsübergreifenden Daten. Es wäre zu prüfen, in wie fern abteilungsübergreifende, unternehmensweite Analyse-Use-Cases für Big Data für solche Unternehmen geeignet sind oder ob stärker auf einzelne Unternehmensteile fokussierte Use Cases (beispielsweise für die Vertriebs- oder Produktionsabteilung) die Realität in deutschen Unternehmen besser abbilden würden.

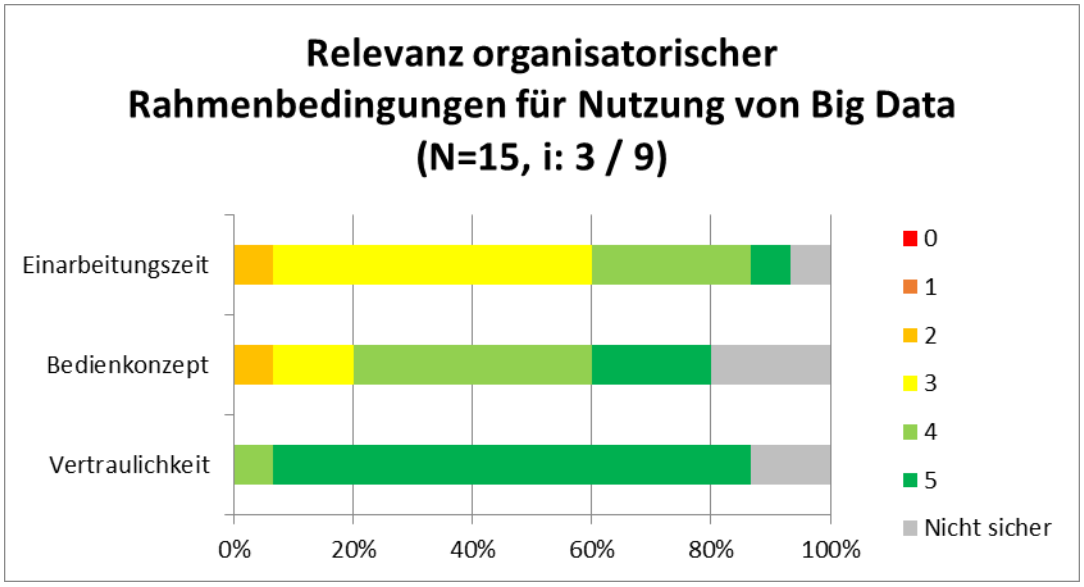


Abbildung 13: Relevanz organisatorischer Rahmenbedingungen für eine Nutzung von Big Data. Organisatorische Rahmenbedingungen sind hier 3 von 9 Rahmenbedingungen für die Nutzung von Big Data insgesamt.

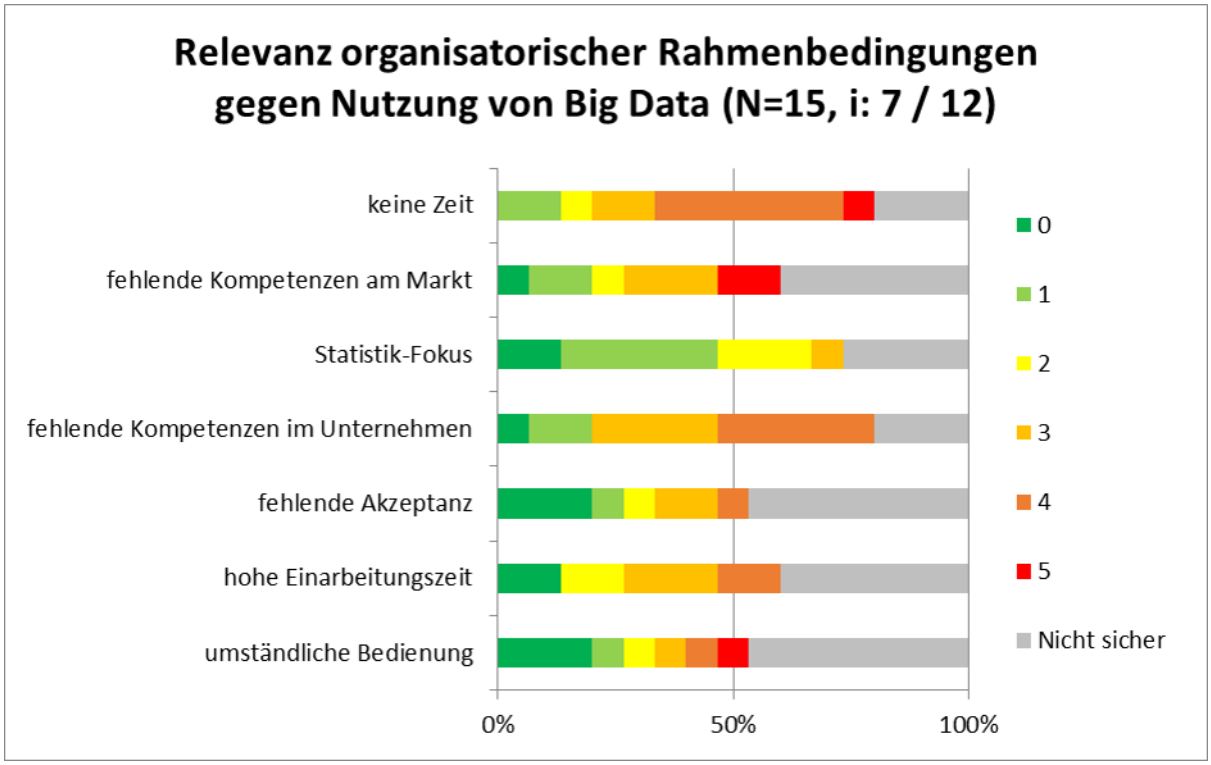


Abbildung 14: Relevanz organisatorischer Rahmenbedingungen gegen eine Nutzung von Big Data. Organisatorische Rahmenbedingungen sind hier 7 von 12 Rahmenbedingungen gegen die Nutzung von Big Data insgesamt.

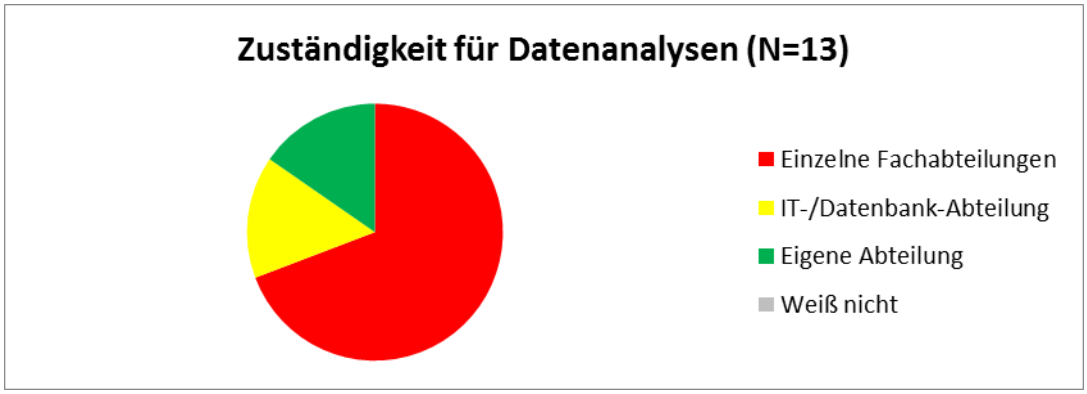


Abbildung 15: Relevanz organisatorischer Rahmenbedingungen gegen eine Nutzung von Big Data. Organisatorische Rahmenbedingungen sind hier 7 von 12 Rahmenbedingungen gegen die Nutzung von Big Data insgesamt.

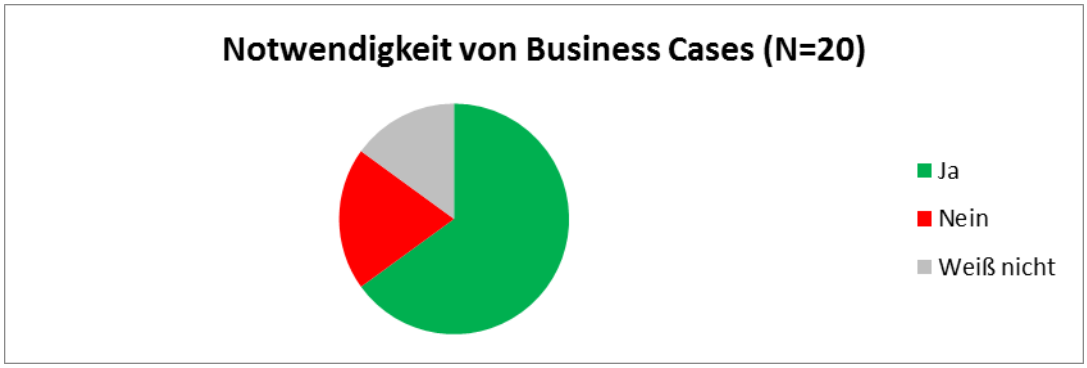


Abbildung 16: Einschätzung zur Notwendigkeit von Business Cases beim Einsatz von Big Data.

4.3. Wirtschaftlichkeit

Selbst wenn die technischen und organisatorischen Voraussetzungen vorliegen, ist für den Einsatz neuer Produkte eine Analyse der Wirtschaftlichkeit bzw. des Nutzens entscheidend. Die meisten Umfrageteilnehmer setzen hierbei die im Wirtschaftsbereich üblichen Anwendungsfälle („Business Cases“) voraus, die das Einsatzszenario im Unternehmen beschreiben und den Nutzen, sofern vorhanden, herausarbeiten (vgl. Abbildung 16). Auf dieser Basis können Projekte oder Programme angestoßen werden.

Auch wenn sich ein Drittel der Teilnehmer in Bezug auf die Relevanz von Business Cases nicht sicher war, war unter denen, die eine Einschätzung abgegeben haben, die Tendenz dahingehend, dass es für Big Data für diese Unternehmen keine relevanten Business Cases gibt. Dieser Eindruck erhärtet sich in Bezug auf die Aussage „kein erkennbarer Nutzen“, wo fast die Hälfte der Teilnehmer diese als „relevant“ oder „sehr relevant“ bewertete. Ebenso relevant gegen Big Data ist laut den Teilnehmern die Amortisationszeit von Investitionen in

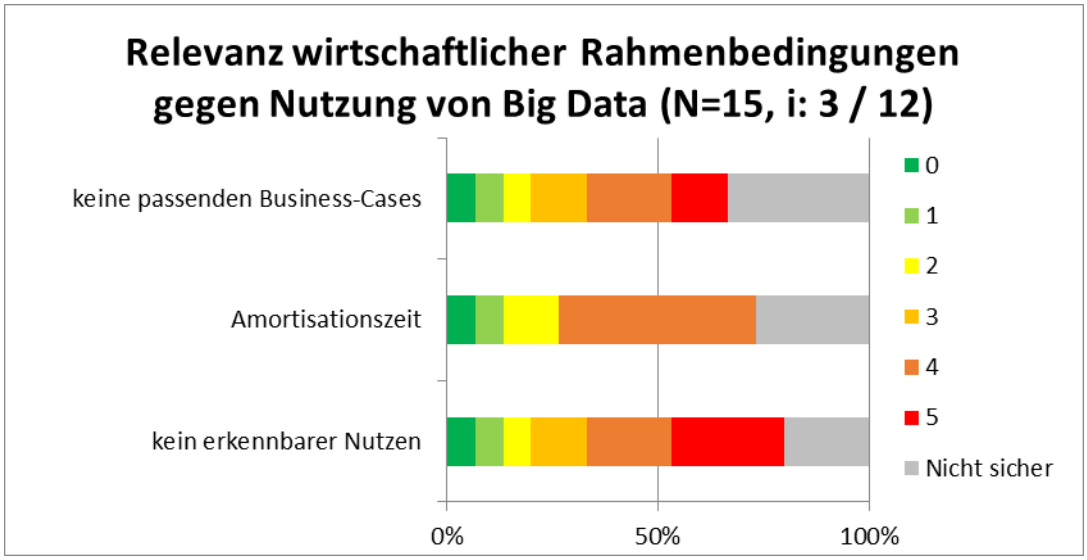


Abbildung 17: Relevanz wirtschaftlicher Rahmenbedingungen gegen eine Nutzung von Big Data. Wirtschaftliche Rahmenbedingungen sind hier 3 von 12 Rahmenbedingungen gegen die Nutzung von Big Data insgesamt.

Big Data, wobei ein geringer Preis an sich wiederum nicht eindeutig als Rahmenbedingung für einen Einsatz bewertet wurde (vgl. Abbildung 17, Abbildung 18).

Das wirtschaftliche Urteil im Teilnehmerfeld fällt recht eindeutig zu Ungunsten von Big Data aus: Der Nutzen wird von den Unternehmen trotz positiv eingeschätzter Attraktivität nicht gesehen, auch wenn allgemein nicht unbedingt ein niedriger Preis ausschlaggebend wäre (vgl. Abbildung 18). Dieses Bild bestätigte sich auch in den Fokus-Interviews, wo der geringe Nutzen ebenfalls herausgestellt wurde bzw. die dringende Notwendigkeit von Business Cases und das Fehlen dieser hervorgehoben wurde.

Im Telefon-Interview betonte ein Interviewpartner, dass die internen Kosten aufgrund

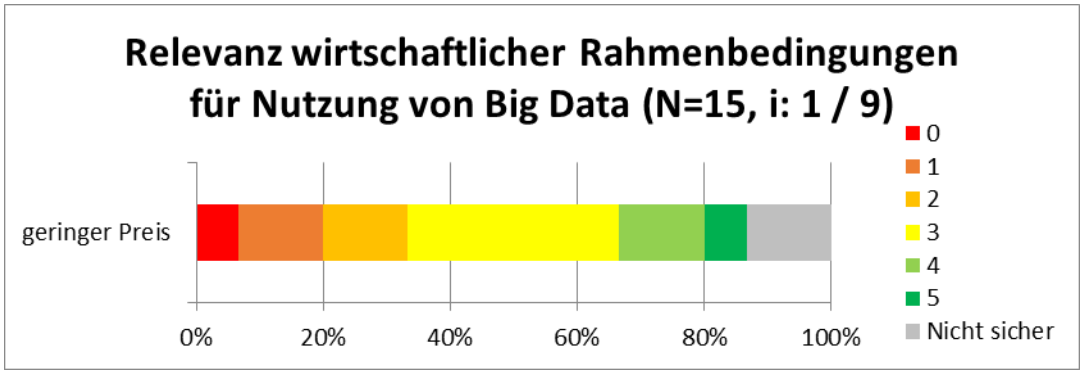


Abbildung 18: Relevanz wirtschaftlicher Rahmenbedingungen für die Nutzung von Big Data. Wirtschaftliche Rahmenbedingungen ist hier 1 von 9 Rahmenbedingungen für die Nutzung von Big Data insgesamt.

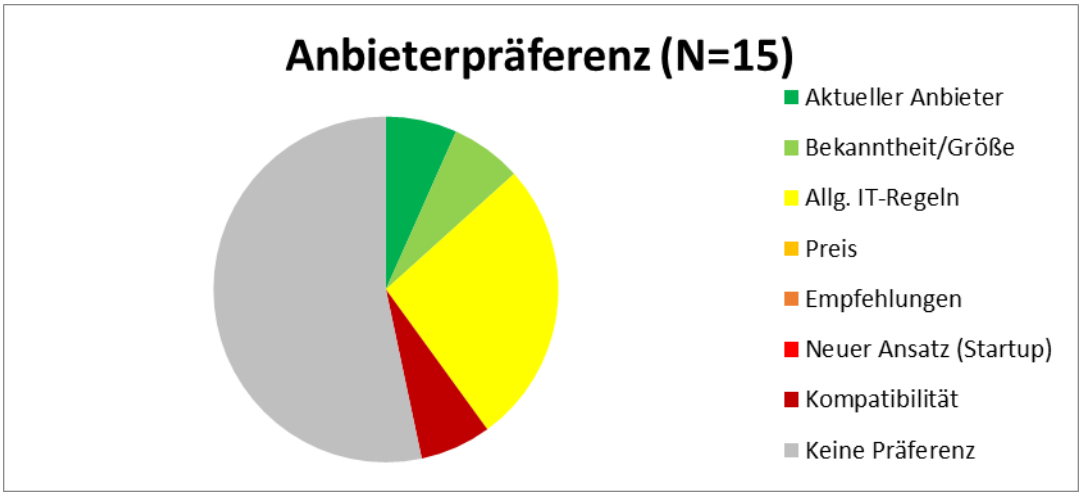


Abbildung 19: Präferenz von Software-Anbietern bei den Umfrageteilnehmern.

des unklaren Nutzens ebenfalls unklar seien und man sich daher auf andere Aspekte konzentriere.

Dieser Eindruck geht einher mit der Eingangsvermutung, dass KMU und deutsche Unternehmen noch sehr zögerlich beim Thema Big Data sind. Die Gründe, warum dies trotz mehrerer vorhandener Use-Case-Beispiele von Herstellerseite so ist, sind noch weiter zu untersuchen. Doch es kristallisiert sich der Eindruck heraus, dass zwischen den existierenden Use Cases und den Erwartungen der Unternehmen an die Use Cases noch eine Lücke klafft. Big Data könnte einerseits ungeeignet für ein Unternehmen sein, aber es ist genauso möglich, dass sich ein Unternehmen nicht in den Use Cases wiederfindet. Letzteres wird durch einen Telefon-Interviewpartner gestützt, der herausstellte, dass bisherige Big-Data-Leitfäden Praxisfragen oft unbeantwortet ließen. Wäre erst einmal eine Entscheidung für Big Data getroffen, hätten zumindest die Umfrageteilnehmer entweder keine Anbieterpräferenz oder würden nach üblichen IT-Regeln entscheiden (vgl. Abbildung 19).

4.4. Rechtliche Aspekte

Auch wenn die rechtliche Dimension nicht im Fokus dieser Umfrage stand, lässt sich festhalten, dass für die Teilnehmer ein Anbieter nicht unbedingt seinen Sitz in Deutschland haben muss (vgl. Abbildung 20). Unabhängig davon muss ein solcher aber die Vertraulichkeit sicherstellen können.

Ein Telefon-Interviewpartner betonte weiterhin, dass geltende Datenschutzregeln eine wirklich durchreichende Big-Data-Analyse verhindern würden. Personenbezogene Daten seien besonders geschützt und es läge am Gesetzgeber, eine Situation zu schaffen, wo es nicht nur komplette Nutzungsverbote gäbe. Hier ist noch ergänzen, dass es auch durchaus

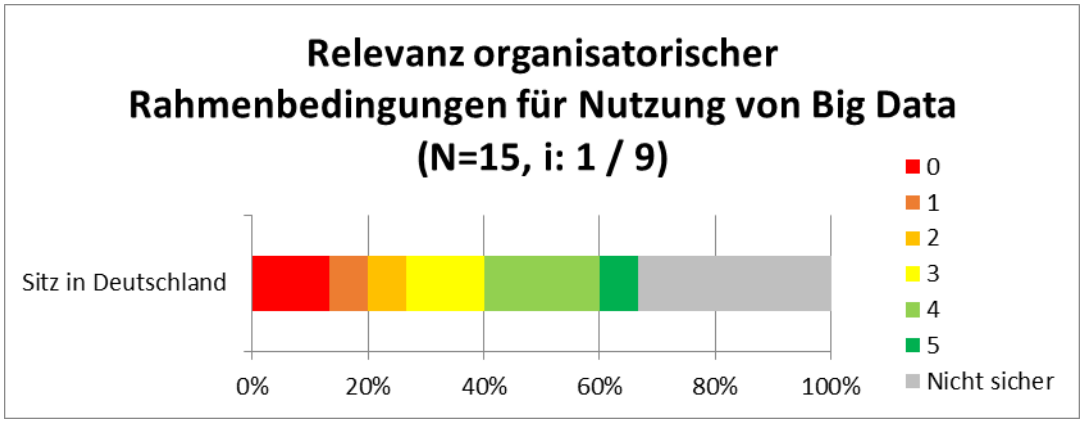


Abbildung 20: Relevanz organisatorischer Rahmenbedingungen für die Nutzung von Big Data im Zusammenhang mit der rechtlichen Einschätzung von Big Data. Organisatorische Rahmenbedingung mit Bezug auf die rechtliche Einschätzung ist hier 1 von 9 Rahmenbedingungen für die Nutzung von Big Data insgesamt.

legale Möglichkeiten gibt, mit personenbezogenen Daten zu arbeiten (wie z.B. bei der Schufa oder anderen Auskunfteien).

5. Diskussion

Die Umfrageergebnisse bestätigen das Bild, dass das Thema „Big Data“ gerade von KMU bisher nur zögerlich angenommen wird. Der durchgehende Eindruck ist, dass der Nutzen von Big Data für die eigene Unternehmung nicht oder nicht hinreichend gesehen wird und die passenden Business Cases fehlen. Die Vermutung liegt nahe, dass dies auch das vergleichsweise geringe Interesse, sich an der Umfrage zu beteiligen, erklärt. Denn wenn der Nutzen für eine Unternehmung von vorne herein unklar ist, hat dies möglicherweise einen Einfluss darauf, an einer Umfrage wie der vorliegenden mitzuwirken. Es darf unterstellt werden, dass, wenn dies bei Großunternehmen mit der entsprechenden Ressourcensituation der Fall ist, dies erst recht bei KMU der Fall ist.

Wenn sich die Unternehmen nicht in den passenden Leitfäden oder Use Cases zu Big Data wiederfinden oder überhaupt nicht die geeigneten Informationen auffinden können, kann keine positive Entscheidung für bzw. eine informierte Entscheidung gegen Big Data erfolgen. Insbesondere die beiden Telefoninterview-Partner äußerten den Wunsch nach einem wissenschaftlich fundierten Leitfaden, der Big-Data-Produktkombinationen mit passenden Anwendungsfällen verbindet. Solange ein Gap zwischen den Big-Data-Use-Cases, die den Unternehmen bekannt sind, und deren Anforderungen an Use-Cases existiert, können diese nur schwer eigene Business-Cases für Big Data aufbauen, und die Einstiegshürde bleibt hoch. Beachtenswert ist hier insbesondere die Verortung von Analysetätigkeiten im Unternehmen. Wenn diese, wie bei den Umfrageteilnehmern, bei den jeweiligen Fachabteilungen durchgeführt werden, sollten sich auch anwendungsfallgetriebene Leitfäden zunächst an dieser Situation orientieren und eine entsprechende Verortung im jeweiligen Unternehmen beachten und nicht unbedingt ein Szenario mit voll-integriertem Datenbestand anstreben oder gar voraussetzen. Beispielsweise könnten Leitfäden bestimmte Probleme isoliert im Vertrieb oder Marketing betrachten und auf deren lokaler Datenbasis aufbauen und damit spezifische Potentiale aufzeigen. Allerdings ist hierzu anzumerken, dass der erfolgreiche und nutzbringende Umgang mit Big Data in zunehmendem Maße Querschnittswissen erfordert, welchen neben IT-Kenntnissen auch Statistik, Problemlösungskompetenz, Kommunikationsfähigkeit, Marketing-Verständnis und ggfs. weiteres umfasst. Weitere Parameter könnten hierbei rechtliche Erwägungen, insbesondere aus dem Datenschutzrecht, sein, die bei Use Cases aus dem angloamerikanischen Raum nicht im Fokus liegen. Insgesamt wird das positive Potential von Big Data schon teilweise erkannt, was dies eben so attraktiv wirken lässt.

Abgesehen von den wirtschaftlichen Voraussetzungen sind auch organisatorische Herausforderungen erkennbar, denn ein erfolgsversprechender Umgang mit Big Data erfordert

ähnlich wie ein Data Warehouse eine ganzheitliche Betrachtung der Aufgabenstellung und des für sie relevanten Datenbestandes. Aus diesem Grund wird in vielen Veröffentlichungen der Ruf nach dem „Data Scientist“[15] immer lauter, von dem man sich Kernkompetenzen in unterschiedlichen Gebieten erhofft, deren Zusammenwirken erst einen fruchtbaren Zugang zu Big Data ermöglicht. Welche Kompetenzen das im Einzelnen genau sind, wird derzeit insbesondere von Aus- und Weiterbildern, aber auch von Unternehmen selbst und den Umfrageteilnehmern intensiv diskutiert. Wenn derartige Kompetenzen nicht vorliegen, erhöht dies die Einstiegshürde in Big-Data-Initiativen und steigert die Kosten, weil Training oder Recruiting nötig sind. Potentiale bestehen hier insbesondere auch für die Hersteller von entsprechenden Produkten, die den Wünschen der Unternehmen nachkommen und Produkte entwerfen, die ein gutes Bedienkonzept haben. Dies wirkt sich positiv auf die nötige Einarbeitungszeit aus, denn insbesondere die Umfrageteilnehmer haben keine generellen Akzeptanzprobleme mit Big-Data-Lösungen.

Auf der technischen Seite besteht die Herausforderung, eine Gesamtlösung zur Datenanalyse zu erhalten, sodass die neuen Big-Data-Produkte in bestehende Analysesysteme integriert werden können – sowohl seitens der Hard- als auch der Software. Letztlich bestätigt das Bild aus der Umfrage die Verhältnisse am Markt, wo viele neuen Produkte und Produktkategorien zu finden sind. Neuartige Produkte können prinzipiell nicht die Reife der traditionellen Produkte erreichen, und auch entsprechende Erfahrungen im Zusammenspiel der verschiedenen Produkte können nicht vollumfänglich vorliegen. Auch in dieser Hinsicht sind sowohl Praxis als auch Wissenschaft gefragt, das Zusammenspiel mehrerer Produkte in einer gemeinsamen Architektur (vgl. z.B. Abbildung 24 im Anhang) näher zu eruieren und praxisnah aufzubereiten. Nur dann können die Performanz-Potentiale und die weiteren Versprechen von Big-Data-Produkten eingelöst werden. Bis dahin liegt die Herausforderung auf technischer Ebene darin, die höhere Einstiegshürde zu überwinden, um das Potential von Big Data zu erschließen und Vorteile zu realisieren, die von anderen Unternehmen möglicherweise noch nicht auf breiter Front umgesetzt werden.

[15] http://news.dice.com/2015/02/12/what-does-it-mean-to-be-a-data-scientist/

6. Zusammenfassung und Ausblick

Wenn man dem Gartner Hype Cycle (Ausgabe 2014[16]) glauben darf, befindet sich Big Data mittlerweile auf dem Weg von „Peak of Inflated Expectations“ in den „Trough of Disillusionment“, also von einem Höhepunkt der überhöhten Erwartungen in die Senke der Desillusionierung, und erleidet damit das typische Schicksal, das bereits viele IT-Entwicklungen ereilt hat (u. a. Service-Orientierung oder Cloud Computing, um nur einige zu nennen). Nach anfänglicher Euphorie stellt sich allmählich heraus, was sinnvoll machbar ist, was lediglich eine frühere Entwicklung fortsetzt und was einen völlig neuen Ansatz erfordert. Vor diesem Hintergrund sind die Ergebnisse unserer Umfrage wenig überraschend. Gerade KMUs und unter diesen wieder vor allem jene, deren Kernkompetenz nicht im IT-Bereich liegt, stehen dem Thema reserviert gegenüber. Man nimmt zwar die diversen Erfolgsgeschichten zur Kenntnis, sieht aber gleichzeitig Herausforderungen, denen man sich nicht unmittelbar, sondern am ehesten mittelfristig stellen muss:

- Neue technologische Konzepte wie Map-Reduce und Hadoop
- Weiterentwicklung von bestehenden Data Warehouse-Lösungen
- Mitarbeiter-Schulung gleichzeitig in unterschiedlichen Gebieten
- Projektziele, die abteilungsübergreifende Kooperation erfordern
- Ungeklärte rechtliche, möglicherweise sogar ethische Fragestellungen

Sobald die Herausforderungen identifiziert sind, kann man sich ihnen stellen, und dies ist auch für KMUs machbar. Es ist für KMUs sogar sinnvoll, den Blick allmählich in die Richtung von Big Data zu lenken, denn die Entwicklungen in Bereichen wie Internet of Things, Industrie 4.0, Smart Car, Smart Home oder Smart City lassen heute bereits klar erkennen, dass die Digitalisierung unserer Welt weiter fortschreiten wird. Damit werden die anfallenden Datenbestände ständig und immer schneller wachsen und neue Formen von Interaktion, Kollaboration, Kooperation, Marketing, Vertrieb und Kundenorientierung ermöglichen.

Allerdings wird nicht nur „der Kunde“ immer gläserner, weil man mehr und mehr über ihn in digitaler Form weiß, auch die Person an sich wird immer durchsichtiger; es gibt bereits warnende Stimmen, die voraussagen, dass in Zukunft ein Fehlen von Daten über eine Person diese verdächtig macht, nicht das Vorhandensein der Daten. Insofern wird

[16] `http://www.gartner.com/newsroom/id/2819918`

sich nicht nur der Umgang mit Objektdaten, sondern auch der mit Personendaten ändern (müssen), und es ist nicht falsch, sich rechtzeitig auf diese Entwicklungen einzustellen.

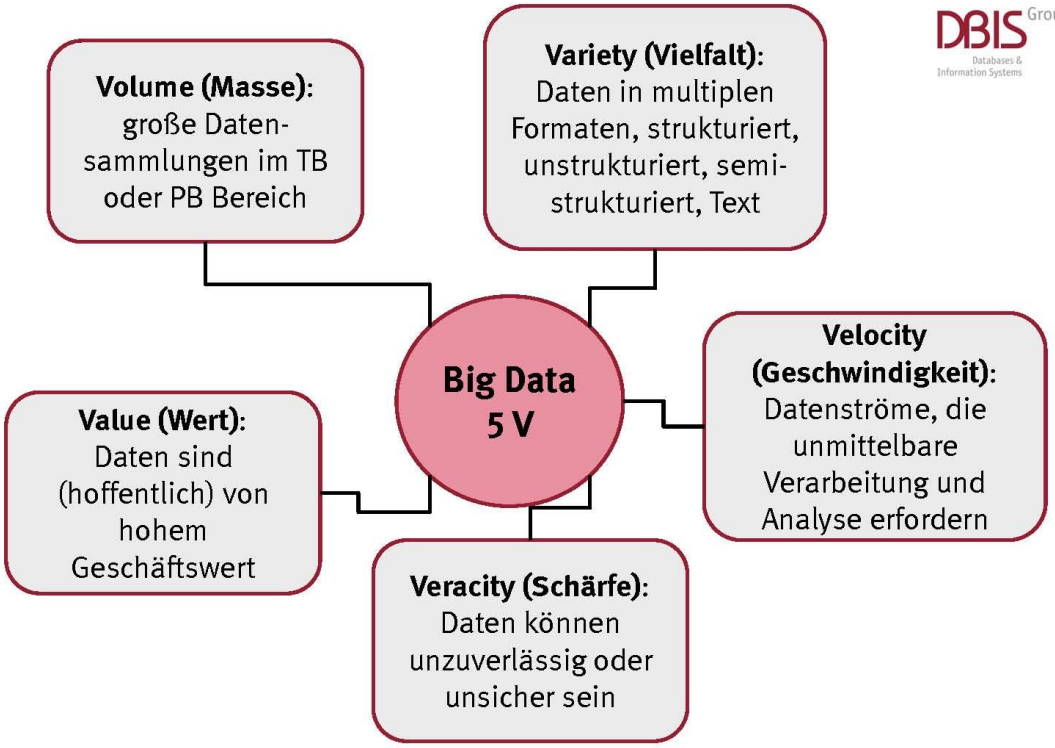
A. Fragebogen der elektronischen Umfrage

Im Folgenden werden die Fragen aus dem elektronischen Fragebogen wiedergegeben. Dieser wurde mit dem Tool „Limesurvey“, gehostet bei der WWU Münster, online erfasst. Die visuelle Darstellung der einzelnen Fragen bzw. der Frageoptionen unterscheidet sich daher von der nun folgenden textuellen Darstellung. Die der Umfrage vorgelagerte „Willkommensnachricht“ wurde an dieser Stelle weggelassen. Die Reihenfolge der Fragen entspricht der der elektronischen Fassung.

I. Kenntnisstand im Hinblick auf Big Data

Um ein einheitliches, gemeinsames Verständnis des Begriffs "Big Data" sicherzustellen, geben wir hier in einer Übersicht an, was im Rahmen dieser Umfrage darunter zu verstehen ist: Big Data ist gekennzeichnet durch die nachfolgend gezeigten Eigenschaften, von denen vor allem Vielfalt und Schärfe in vielfacher Hinsicht neue Herausforderungen bedeuten (mehr als Masse und Geschwindigkeit). Man erhofft sich neue Formen von Wertschöpfung und Erkenntnisgewinn von einer Verknüpfung und Analyse großer Datenbestände. Zahlreiche Fallstudien deuten darauf hin, dass diese Hoffnung berechtigt ist.

Verweis: http://de.wikipedia.org/wiki/Big_Data



1) Wie beurteilen Sie Ihre Kenntnisse im Hinblick auf Big Data-Anwendungen?
5 bedeutet Sie haben sehr umfängliches Wissen über Big Data-Anwendungen und 0 bedeutet Sie haben kein Wissen bzgl. Big Data-Anwendungen.

	0	1	2	3	4	5
Kenntnisse über Big Data-Anwendungen						

2) Betreiben Sie bereits eine Big Data-Anwendung bzw. haben Sie vor, dies zu tun?

- Ja, regelmäßig
- Ja, ausprobiert
- Nein

3) Glauben Sie, dass in Ihrem Unternehmen zunächst ein angemessener Business Case erarbeitet werden muss, bevor eine Big Data-Anwendung aufgesetzt wird?

- Ja
- Nein
- Weiß nicht

4) Wie stehen Sie Big Data bzw. Big Data-Anwendungen gegenüber?

	0 – Stimme nicht zu bis 5 – stimme zu "N.S." – Nicht sicher						
	N. S.	0	1	2	3	4	5
Die Nutzung von Big Data finde ich attraktiv.							
Die Nutzung von Big Data-Anwendungen ist wirtschaftlich.							
Die Nutzung von Big Data-Anwendungen ist einfach.							
Sicherheit von Daten-Backups ist ein Schlüsselfaktor bei der Nutzung von Big Data-Anwendungen.							
Applikationskompatibilität ist ein Schlüsselfaktor bei der Nutzung von Big Data-Anwendungen.							
Big Data-Anwendungen sind vertrauenswürdig.							
Big Data-Software ermöglicht es mir, Dinge schneller zu erledigen.							

Big Data-Anwendungen verbessern unsere Unternehmensleistung.							
Big Data-Anwendungen verbessern unsere Wettbewerbsfähigkeit.							
Big Data ist mehr als ein Hype, ist nützlich.							

Bitte begründen Sie Ihre negative Erwartungshaltung kurz (nur wenn sie negativ sind):

_____

5) Wird bereits spezifische Big-Data-Software in Ihrer Firma genutzt?

Big-Data-Software ist z.B. eine Hadoop-Distribution von MapR, Apache, Hortonworks oder Cloudera, IBM InfoSphere BigInsights, Amazon Elastic MapReduce oder auch eine Anwendung von Wibidata, Splunk oder Palantir etc.

- o Ja, regelmäßig
- o Ja, gelegentlich
- o Nein, aber über den Einsatz wird nachgedacht
- o Nein, der Einsatz ist nicht absehbar

Falls nein, zu welchem ungefähren Zeitpunkt werden die Big Data-Anwendungen Ihrer Erwartung nach eingeführt werden?

- o in den nächsten 6 Monaten
- o in 6 bis 12 Monaten
- o in 1 bis 2 Jahren
- o später als in 2 Jahren

6) Gibt es grundsätzliche Vorbehalte gegenüber Big Data in Ihrem Unternehmen, die den Einsatz bzw. einen intensiveren Einsatz von Big Data-Anwendungen verhindern?

- o nein
- o ja

Wenn ja, welche Vorbehalte sind dies Ihrer Meinung nach?

_____

7) Gibt (bzw. gäbe) es bei Einsatz von Big Data-Software grundsätzlich eine Präferenz hinsichtlich des Softwareanbieters?

- Der aktuelle Softwareanbieter
- Bekannte und große Softwareanbieter
- Die Auswahl eines Anbieters wird durch allg. Regeln der IT-Abteilung bestimmt
- Der günstigste Anbieter
- Ein Anbieter, der von Nutzern empfohlen wird (bspw. in Diskussionsforen)
- Keine besondere Präferenz
- Andere Präferenz: _____

II. Gründe für und wider Big Data-Anwendungen

Als nächstes fragen wir Sie zuerst nach den Rahmenbedingungen, also nach den Voraussetzungen, die eine Big Data-Applikation erfüllen muss, bevor ein Einsatz in Ihrem Unternehmen überhaupt in Frage kommt. Im Anschluss fragen wir Sie dann nach Gründen, die aus Ihrer Sicht für oder gegen einen Einsatz von Big Data-Produkten in Ihrem Unternehmen sprechen.

8) Bitte beurteilen Sie die Relevanz der nachfolgenden Rahmenbedingungen für die Nutzung von Big Data.

	0 – nicht relevant bis 5 – sehr relevant "N.S." – Nicht sicher						
	N.S.	0	1	2	3	4	5
Vertraulichkeit der Daten gegenüber Dritten							
Bedienkonzept der Big-Data-Software angelehnt an klassische Desktop-Anwendungen (Drag & Drop, Fenster u. ä.)							
Performance der Anwendung (Darstellung, Interaktion und Verarbeitung ohne Zeitverzögerung)							
Einarbeitungszeit für die IT-Mitarbeiter							
Kostenlose bzw. preisgünstige Softwarenutzung							
Möglichkeit einer Installation der Big Data-Software auf der eigenen Infrastruktur							
Möglichkeit der Integration in andere Big Data-Angebote (auch anderer Anbieter)							
Möglichkeit der Integration in bestehende Software							
Anbieter hat Sitz in Deutschland							

9) Falls es weitere Gründe gibt, die aus Ihrer Sicht für eine Nutzung von Big Data-Produkten im Unternehmen sprechen, so können Sie diese hier eintragen und bewerten:

_____

10) Was sind aus Ihrer Sicht die Gründe, die gegen eine Nutzung von Big Data-Anwendungen sprechen, und welche Relevanz haben diese Gründe?

	0 – nicht relevant bis 5 – sehr relevant "N.S." – Nicht sicher						
	N.S.	0	1	2	3	4	5
Bedienung umständlich							
Hohe Einarbeitungszeit							
Fehlende Mitarbeiterakzeptanz (etablierte Arbeitsabläufe, Einarbeitung in neue Software)							
Nutzen nicht erkennbar							
Einschlägige Kompetenzen im Unternehmen nicht vorhanden							
Big Data-Anwendungen sind primär etwas für Statistiker							
Anschaffungskosten für neuartige Software und leistungsfähige Hardware müssen erst amortisiert werden							
Big Data-Anwendungen enthalten bisher nicht alle Features, die das Unternehmen sich wünscht							
Einschlägige Kompetenzen am Markt nicht verfügbar							
Keine Zeit sich neben Tagesgeschäft damit zu beschäftigen							
Prominente Business-Cases für Big-Data passen nicht zu uns							
Hohe Betriebs- und Instandhaltungskosten über die Zeit (Personal, Hard-/Software)							

11) Falls es weitere Gründe gibt, die aus Ihrer Sicht gegen eine Nutzung von Big Data-Produkten im Unternehmen sprechen, so können Sie diese hier eintragen und bewerten:

_____

Welches ist die größte Herausforderung für die Datenanalyse- und Reporting-Fähigkeiten im Unternehmen in den nächsten 12 bis 24 Monaten?

_____

III. Softwareunterstützung im Unternehmen

12) Welchen Typ von Software verwenden Sie in Ihrem Unternehmen hauptsächlich?

- Individual-Software (speziell für das Unternehmen programmiert)
- Customized Software (auf das Unternehmen angepasste Standard-Software)
- Standard-Software eines großen Herstellers (ohne Anpassungen)
- Standard-Software eines unbekannteren Herstellers
- Mischung aus obigen Kategorien ohne klaren Schwerpunkt
- Weiß nicht

13) Wie viele Mitarbeiter nutzen die Software in Ihrem Unternehmen?

- bis 10 Nutzer
- 10 -39 Nutzer
- 40-149 Nutzer
- 150 oder mehr Nutzer
- Weiß nicht

14) Über welche Endgeräte haben die Nutzer hauptsächlich regelmäßig Zugriff auf die Software?

- Ortsgebundene Arbeitsplatzrechner

- o Nicht ortsgebundene Laptops/Tablet-PCs
- o Standardisierte mobile Endgeräte (Smartphones, PDAs)
- o Spezielle Endgeräte (z. B. mobile RFID-Lesegeräte, spezialisierte Bedienterminals, besonders widerstandsfähige PDAs)

  Weiß nicht

15) Basieren zumindest einige der Anwendungen auf einer Client-Server-Architektur?

- o Nein, es wird kein zentraler Server verwendet
- o Ja, der oder die Server werden im eigenen Rechenzentrum bzw. Serverraum durch Ihre Firma bereitgestellt und verwaltet
- o Ja, der oder die Server werden in einem externen Rechenzentrum bereitgestellt und durch Ihre Firma verwaltet
- o Ja, der oder die Server werden komplett durch einen externen Anbieter bereitgestellt und verwaltet
- o Weiß nicht

16) Erfolgt der Zugriff auf die Software ganz oder zu einem guten Teil über einen Web-Browser?

- o ja
- o nein

17) Haben Sie spezielle Wünsche an die Hersteller von Big Data-Software bzw. –Applikationen, die bisher nicht berücksichtigt sind?

_____

18) Welche der folgenden Technologien setzen setzt Ihr Unternehmen bereits ein?
(Mehrfachnennungen sind möglich)

- ☐ Klassische SQL-Datenbanken (z.B. MySQL, IBM DB2, Oracle Database)
- ☐ Data-Warehouse-Lösungen (DWH) (z.B. SAP BW)
- ☐ Tabellenkalkulationen (z.B. Excel)
- ☐ Statistik-Tools (z.B. SPSS)
- ☐ NoSQL-Datenbanken (z.B. MongoDB, CouchDB, Apache Cassandra)
- ☐ Einfache In-Memory-Datenbanken (z.B. memcached)
- ☐ In-Memory-Datenbank-Appliances (z.B. SAP HANA, Oracle Exalytics)

☐ Column Stores (z.B. Sybase IQ)

☐ Spezielle Big-Data-Software (z.B. Wibidata, Splunk, Palantir)

o Keine davon

o Weiß nicht

19) Welche Abteilung im Unternehmen ist hauptsächlich für die Durchführung von Datenanalysen (inklusive Reporting) zuständig?

o Einzelne Fachabteilungen jeweils für ihren Bereich

o IT- oder Datenbank-Abteilung

o Eigene Abteilung für Datenanalyse ("Business Analytics")

o Andere: _____

IV. Zum Unternehmen und zu Ihrer Person

Die Beantwortung der folgenden Fragen ist optional.

20) Aus welcher Branche stammt Ihr Unternehmen?

o Einzel- und Großhandel

- Dienstleistungen:
  - o Freiberufliche, wissenschaftliche und technische Dienstleistungen
  - o Grundstücks- und Wohnungswesen
  - o Information und Kommunikation
  - o Andere Dienstleistungen

o Baugewerbe

o Industrie, Verarbeitendes Gewerbe

o Gastgewerbe, Tourismus

o Gesundheits- und Sozialwesen

o Transport & Verkehr

o Energie

o Land- & Forstwirtschaft, Fischerei

o Andere Branche: _____

21) Wie hoch war der Jahresumsatz Ihres Unternehmens im vergangenen Jahr?

o bis 500.000 €

o 500.000-2 Mio. €

o 2 Mio - 10 Mio. €

o 10 Mio - 50 Mio. €

o mehr als 50 Mio. €

22) Wie viele Mitarbeiter werden in Ihrem Unternehmen beschäftigt?

- 1 bis 9 Mitarbeiter
- 10 bis 24 Mitarbeiter
- 25 bis 49 Mitarbeiter
- 50 bis 149 Mitarbeiter
- 150 bis 249 Mitarbeiter
- 250 oder mehr Mitarbeiter

23) Wie viele Mitarbeiter beschäftigen Sie, die sich um Aufbau, Betrieb und Wartung der IT-Infrastruktur kümmern?

- 1 bis 5 Mitarbeiter
- 6 bis 10 Mitarbeiter
- 11 bis 20 Mitarbeiter
- mehr als 20 Mitarbeiter
- Ausschließlich externes Personal

24) Betreiben Sie ein eigenes Rechenzentrum für Ihr Unternehmen?

- Nein
- Ja, eins
- Ja, mehrere. Anzahl:
- Weiß nicht

25) Zu Ihrer Person: In welcher Abteilung des Unternehmens sind Sie tätig?

- IT-Abteilung
- Fachabteilung
- Geschäftsführung
- Andere: _____

Zustellung der Auswertung

Wir lassen Ihnen die Ergebnisse unserer Studie gerne zukommen, sobald sie vorliegen, falls Sie Ihre E-Mail-Adresse angeben. Wir versichern, diese lediglich für das Versenden der Ergebnisse zu verwenden.

An welche E-Mail-Adresse sollen wir Ihnen später die Ergebnisse zukommen lassen?

Email-Adresse: _____

Falls Sie keine Zustellung der Ergebnisse wünschen, lassen Sie dieses Feld leer und klicken Sie auf "Weiter".

B. Fokus-Interview-Fragebogen

Westfälische Wilhelms-Universität Münster

DBIS Group Databases & Information Systems

Big Data in KMU: Fokus-Fragen

Ergänzende Fokus-Fragen zur „Big Data in KMU“-Umfrage von der WWU Münster.

Allgemeine Fragen: Nachfragen zu bisherigen Ergebnissen

1. Viele Umfrageteilnehmer hielten Big Data für grundsätzlich attraktiv, aber es gab Vorbehalte wie etwa fehlende Business Cases oder geringe Einfachheit. Welches Versprechen von Big Data macht dies Ihrer Meinung nach für Unternehmen so attraktiv? Trifft dies für Sie ebenfalls zu?

2. Die Vielfalt der Big-Data-Software ist groß. In wie fern und unter welchen Voraussetzungen würden Use-Case-getriebene Leitfäden zur Big-Data-Softwareauswahl helfen, den Zugang zum Thema Big Data zu erleichtern?

3. Denken Sie, dass sich das Kosten-Nutzen-Verhältnis von Big-Data-Lösungen schwieriger bewerten lässt als das von anderer Software?

4. Die Integration in die eigene Infrastruktur wurde häufiger als nötige Rahmenbedingung für Big Data genannt. Welche Herausforderungen mit aktueller Big-Data-Software sehen Sie an dieser Stelle? An welcher Stelle hapert es hier noch? Wo wäre Hilfe wünschenswert, z.B. mit Hilfe von Workshops?

5. In wie fern würde weiter verbreitetes Wissen bei den (Fach)mitarbeitern um Big-Data-Software dessen Einsatz fördern?

Spezielle Fragen: Weitergehende und firmenspezifische Fragen

6. Unter welchen Voraussetzungen könnten Sie sich eine Data-Warehouse-Fortentwicklung durch einen (ausgiebigeren) Big-Data-Einsatz (z.B. Apache Hadoop, MapReduce) in Ihrem Unternehmen vorstellen?

7. Welche zentralen Herausforderungen sehen Sie für die Datenanalyse / das Reporting in Ihrem Unternehmen in Zukunft – sowohl technisch als auch organisatorisch?

8. Gibt es nicht-erschlossene Datenquellen, die Sie ohne neuartige Technik oder organisatorische Maßnahmen nicht erschließen können?

9. Welche Informationen würden für Sie einen Wettbewerbsvorteil gerne nutzen, welche Sie derzeit nicht auswerten können?

10. Bei Big Data geht es auch um sog. explorative Analyse, wo viele Daten erst gesammelt werden und dann erst nach passenden Erkenntnissen (auch im Nachhinein) gesucht wird. Was halten Sie von diesem Ansatz? Wäre dieser (ggf. als Ergänzung) geeignet für Ihr Unternehmen?

Databases & Information Systems Group - Prof. Dr. Gottfried Vossen - Leonardo-Campus 3 - 48149 Münster, Germany
Umfrage-Feedback: david.fekete@wi.uni-muenster.de, Tel.: +49 251 83-38161

C. Zur technologischen Dimension von Big Data

Zum Umgang mit Big Data sind in den letzten Jahren zahlreiche Techniken, Methoden und Technologien entwickelt worden, über die wir im Folgenden einen kurzen Überblick geben. Wenn Daten in so hohen Quantitäten auftreten, dass lokal vorhandene Speicher- und Verarbeitungskapazitäten nicht mehr ausreichen, kann es nicht verwundern, dass „traditionelle“ Technologie, die auf einer zentralen Datenbank fußt, nicht mehr ausreicht. Um zu verstehen, was stattdessen benötigt wird bzw. angemessen ist, stellen wir zunächst Anforderungen an die Verarbeitung von Big Data zusammen und betrachten dann Technologien, die diesen genügen. Diese Anforderungen lassen sich wie folgt zusammenfassen:

- Angemessene, hohe Verarbeitungsleistung für komplexe Rechnungen;
- skalierbare, verteilte und fehlertolerante Datenverarbeitung mit temporärem oder dauerhaftem Speicher;
- parallele Programmier- und Verarbeitungsparadigmen, die mit großen Datenmengen angemessen umgehen können;
- angemessene Implementierungen und Ausführungsumgebungen für diese Programmiermodelle und Paradigmen.

Bzgl. Hardware-Lösungen zur Verarbeitung von Big Data sei der Leser auf [SM13] verwiesen. Ebenfalls relevant in diesem Zusammenhang ist ein Wiederaufkommen von Hauptspeicher- bzw. In-Memory-Datenbanktechnologie, eine Entwicklung, die ursprünglich aus den 1980er Jahren stammt [Eic89] und die mittlerweile – dank enormer technologischer Fortschritte in den letzten 30 Jahren – in kommerziellen Produkten verfügbar ist [LLV+11, PZ11]. Bei Datenbanken sind ferner NoSQL- und NewSQL-Systeme entstanden, die wir im folgenden Abschnitt skizzieren, bevor wir uns im Anschluss Map-Reduce zuwenden.

C.1. Neue Entwicklungen im Datenbank-Bereich

Seit einigen Jahren gibt es vielfältige Datenbank-Aktivitäten und -Produkte unter dem Schlagwort „NoSQL“. Dennoch (oder gerade deswegen) gestaltet sich eine genaue Abgrenzung der zugehörigen Bestrebungen schwierig. Zunächst ist „NoSQL“ der Name einer seit Ende der 1990er Jahre weiterentwickelten relationalen Open-Source-Datenbank[17], die keinen Zugriff über die standardisierte Datenbanksprache SQL anbietet. Nach heutigem

[17] `http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/nosql/Home%20Page`

Verständnis gehört diese Datenbank allerdings nicht der NoSQL-Bewegung an, denn das „No“ wird mehrheitlich nicht mit „nein“, sondern mit „not only“ interpretiert. Die Bewegung wurde insbesondere durch zwei Konferenzen im Jahre 2009 geprägt, das NOSQL Meeting[18] im Juni 2009 und die no:sql(east)[19] im Oktober 2009.

Die erste dieser Konferenzen hatte zum Ziel, Beschränkungen gängiger relationaler Datenbanken und deren Überwindung mit Hilfe verteilter, nicht-relationaler Open-Source-Datenbanken zu diskutieren. Ausgangspunkt dieser Fokussierung war die Tatsache, dass im Web-Umfeld große Datenbestände aus Performanz-, Flexibilitäts- und Skalierbarkeitsgründen häufig nicht mit relationalen Datenbanken verwaltet und ausgewertet wurden, sondern durch alternative Systeme, die auf SQL und starre Datenbankschemata oder ACID-Transaktionen verzichten. Einflussreich waren hier insbesondere die Kernideen von Bigtable [CDG+08] und Map-Reduce [DG04] (Datenhaltung und -auswertung bei Google) sowie von Dynamo [DHJ+07] (Datenhaltung bei Amazon), die in einer steigenden Anzahl von Open-Source-Projekten aufgegriffen und in populären Web-Auftritten eingesetzt wurden. Im Folgenden werden wir zunächst Grundlagen der verteilten Datenhaltung und insbesondere den für viele NoSQL-Systeme charakteristischen Begriff der Eventual Consistency einführen, bevor wir genauer auf diese Systeme eingehen.

Parallel zu den NoSQL-Entwicklungen gab es allerdings bereits frühzeitig kritische Stimmen, die vor dem Verzicht auf SQL-Standardisierung und auf ACID-Transaktionen warnten. Entsprechend wurden und werden SQL-Datenbanken, die weiterhin auf ACID-Transaktionen setzen, aber die Skalierbarkeit von NoSQL-Systemen erreichen sollen, unter dem Schlagwort NewSQL beworben; derartige Systeme werden abschließend thematisiert.

C.2. Partitionierung, Replikation, CAP-Theorem, Eventual Consistency

Partitionierung und Replikation sind gängige Techniken zur Gewährleistung von Skalierbarkeit und Fehlertoleranz in verteilten Systemen (etwa im Cloud Computing und im Datenbankumfeld).

Verteilte Anwendungen verlassen sich hinsichtlich der notwendigen Rechenleistung typischerweise auf große Ansammlungen gebräuchlicher Hardware, bestehend aus konventionellen Prozessoren („Rechenknoten“), die zu Clustern zusammengefasst und über geeignete Netzwerktechnologie miteinander verbunden sind; derartige Cluster werden häufig innerhalb eines Datacenters oder über mehrere Datacenter hinweg repliziert. Re-

[18] `http://nosql.eventbrite.com/`
[19] `https://nosqleast.com/2009`

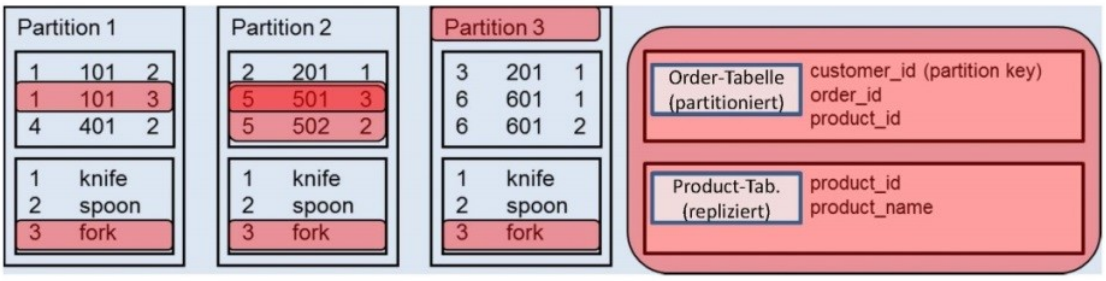


Abbildung 21: Partitionierung vs. Replikation.

plikation, also das gezielte Kopieren identischer Daten auf mehrere Server, als Form von Redundanz ist in diesem Zusammenhang der Schlüssel zur Zuverlässigkeit des betreffenden Systems und des fehlertoleranten Rechnens; analog werden Daten durch Replikation gegen Verlust geschützt. Das Ergebnis ist dann entweder ein verteiltes File-System (wie das Hadoop Distributed File System, kurz HDFS) oder eine global verteilte Datenbank (wie Google Spanner).

Neben Fehlertoleranz und (hoher) Verfügbarkeit ermöglicht Verteilung eine parallele Verarbeitung von Daten, speziell dann, wenn Rechenvorgänge unabhängig voneinander auf unterschiedlichen Teilen der Daten ausgeführt werden können. In einem solchen Fall werden Daten häufig über mehrere Cluster oder sogar Datacenter hinweg partitioniert; Abbildung 21 illustriert den Unterschied zwischen Partitionierung und Replikation. In dem in dieser Abbildung gezeigten Beispiel sind diejenigen Daten aus einer relationalen Datenbank, die Kunden-Order (Bestellungen) betreffen, über drei Speicherorte partitioniert so, dass jedem Ort eindeutige Kundennummern zugeordnet sind. Die Product-Tabelle ist dagegen an denselben Orten repliziert (d.h. identisch kopiert). Anfragen und Updates können damit an eine oder gleichzeitig an mehrere Partitionen gerichtet werden, wobei Replikation und Partitionierung für Benutzer transparent sind.

Es sollte unmittelbar einleuchten, dass schreibende Zugriffe auf replizierten Daten besonderer Vorkehrungen bedürfen, um die Konsistenz der Daten sicherzustellen. Während klassische relationale Datenbanken mit ACID-Transaktionen die Serialisierbarkeit als Konsistenzbegriff verfolgen, gilt im Kontext verteilter Systeme die Ein-Kopien-Konsistenz (engl. *single copy consistency*), also die Illusion der Abwesenheit von Replikation, als starke Konsistenz, und es gibt eine ganze Reihe schwächerer Formen von Konsistenz wie *Eventual Consistency*, *Read-Your-Writes*, *Monotonic Reads*, *Monotonic Writes* oder *kausale Konsistenz* [Vog08].

Zur Durchsetzung von starker Konsistenz auf replizierten Daten kommen oftmals Konsens-Protokolle wie Paxos [Lam98] zum Einsatz, etwa bei Google in Bigtable [CDG+08]

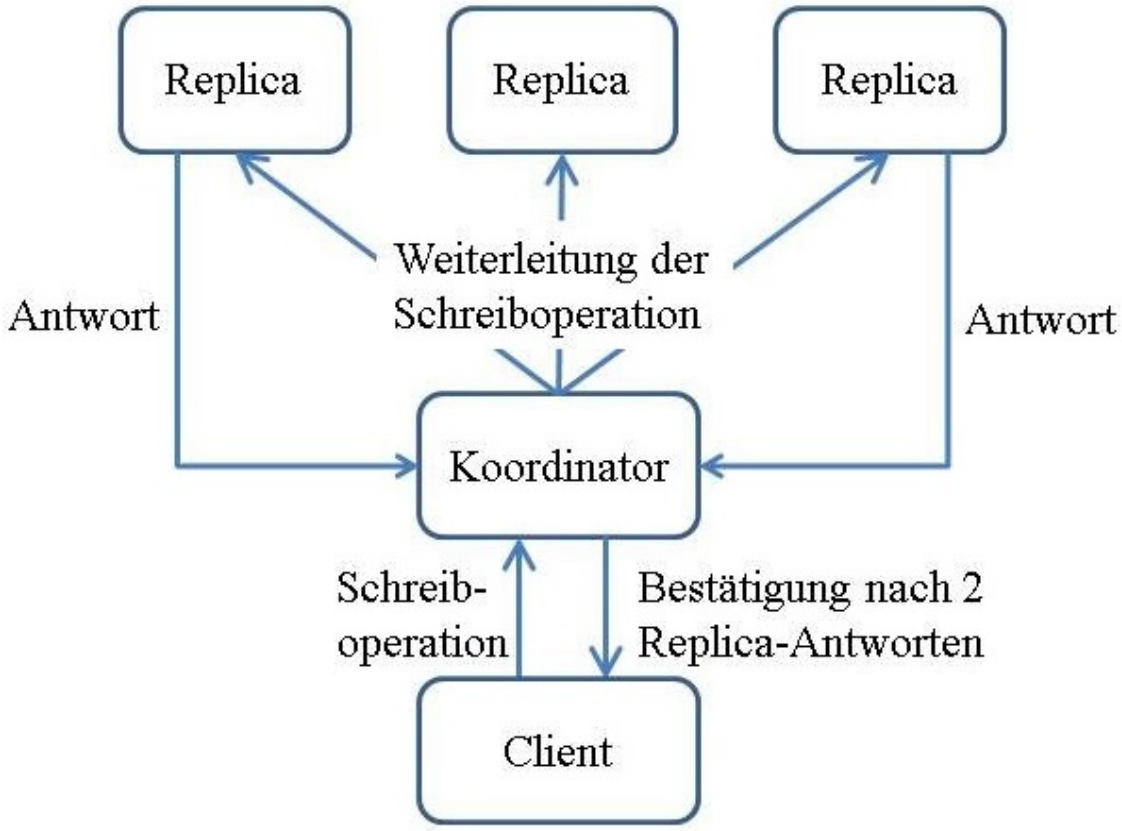


Abbildung 22: Schreiboperation in Quorum-System mit N=3 und W=2.

und Spanner [CDE+12]. *Eventual Consistency* beruht wie im Falle von Amazon Dynamo [DHJ+07] typischerweise auf Quorum-Protokollen, unter denen Lese- und Schreiboperationen auf mehreren (möglicherweise allen) Kopien ausgeführt werden müssen, bevor sie als erfolgreich durchgeführt gelten: Bei Dynamo werden Daten auf N Knoten repliziert, und es gibt zwei konfigurierbare Parameter, R und W. R spezifiziert die minimale Anzahl von Knoten, die für eine erfolgreiche Lese-Operation erforderlich sind, während W diese Anzahl für Schreib-Operationen vorgibt. Im Falle von $R+W > N$ liegt ein klassisches Quorum-System vor, in dem sich Lese- und Schreiboperationen immer überlappen und wo daher starke Konsistenz garantiert werden kann. Da jedoch viele Knoten beteiligt sind, ist eine hohe Latenz zu erwarten. Eventual Consistency liegt vor, falls $R+W \leq N$ gilt; hier ist die Latenz reduziert, allerdings können Leseoperationen auf Knoten ausgeführt werden, die noch nicht alle Schreiboperationen nachvollzogen haben, was dann in veralteten Werten resultiert. In Abbildung 22 wird [BVF+14] folgend eine Schreiboperation für den Fall N=3 und W=2 illustriert: Die Schreiboperation wird erfolgreich abgeschlossen, wenn zwei der drei Kopien die Schreiboperation erfolgreich durchgeführt haben. Im Falle R=1 könnten nachfolgende Leseoperationen dann den Wert der noch nicht aktualisierten Kopie liefern.

Schwächere Formen von Konsistenz (wie Eventual Consistency) sind vor allem aus zwei Gründen interessant. Erstens können sie Latenzen vermeiden, die aus Synchronisationsmechanismen resultieren. Zweitens besagt das *CAP-Prinzip* [FB99], das später in [GL02] als *CAP-Theorem* bewiesen wurde, dass man von den drei Eigenschaften Konsistenz (engl. *Consistency*), Verfügbarkeit (engl. *Availability*) und Partitionstoleranz (engl. *Partition tolerance*) höchstens zwei gleichzeitig sicherstellen kann. Partitionstoleranz liegt vor, wenn Anwendungen mit der Situation umgehen können, dass die zugrunde liegenden Server aufgrund von Netzwerkproblemen in verschiedene Partitionen zerfallen, zwischen denen zeitweise keine Kommunikation möglich ist. Für hochgradig (weltweit) verteilte Anwendungen lassen sich Netzwerkpartitionierungen in der Regel nicht ausschließen. Wenn dann noch hohe Verfügbarkeit garantiert werden soll, müssen laut CAP-Theorem also Kompromisse bezüglich der Konsistenz eingegangen werden.

Entsprechend verfolgen viele NoSQL-Datenbanken nicht mehr das Ziel von starker Konsistenz, sondern versprechen „irgendwann konsistent“ (engl. *eventually consistent*) zu werden: Wenn Partitionierungen behoben sind und hinreichend lange keine neuen Aktualisierungen vorgekommen sind, ist sichergestellt, dass irgendwann sämtliche Updates so durch das System propagiert wurden, dass alle Knoten wieder konsistent sind. NoSQL-Datenbanken greifen damit das *BASE*-Prinzip auf (anstelle von ACID), das im Kontext verteilter Internet-Anwendungen bereits lange bekannt ist [FGC+97]: *Basically Available, Soft State, Eventually Consistent*.

Bemerkenswert ist allerdings, dass das CAP-Theorem keine Rechtfertigung liefert, auf starke Konsistenz generell zu verzichten: Wie Brewer in [Bre12] erläutert, ist der Verzicht auf Konsistenz nur während der Dauer von Partitionierungen notwendig. Solange keine Partitionierung auftritt, was dem Regelbetrieb entspricht, können auch hochverfügbare Systeme starke Konsistenz garantieren.

C.3. NoSQL

Wie bereits erwähnt subsumiert „NoSQL“ verschiedene Formen der Datenhaltung, die auf die eine oder andere Art von „klassischen“ SQL-DBMS abweichen, wobei in der Regel Open Source, horizontale Skalierbarkeit und hohe Verfügbarkeit bei Vernachlässigung von Konsistenz erwartet werden. Die Bandbreite[20] reicht von Key-Value Stores (z.B. Amazon SimpleDB und Dynamo, LinkedIn Voldemort, Riak, Redis) und Column Stores (z.B. Google Bigtable [CDG+08], Apache Hbase oder Cassandra, Yahoo! PNUTS) über Dokument-Datenbanken (z.B. MongoDB oder Couchbase) bis zu Graphen-Datenbanken

[20] `http://nosql-database.org/`

(z.B. Neo4J der Allegro) [RW12]. Wir stellen nachfolgend einige Besonderheiten von Column Stores am Beispiel Bigtable sowie von Key-Value Stores am Beispiel Dynamo vor.

Das Datenmodell von Bigtable [CDG+08] beruht auf Tabellen, denen Spalten zugewiesen werden, die ihrerseits aus verwaltungstechnischen Gründen in sogenannte *Spaltenfamilien* gruppiert werden. Während eine Tabelle typischerweise wenige Spaltenfamilien umfasst, kann die Anzahl der Spalten unbeschränkt sein. Eine Zeile einer Tabelle (also ein Datensatz) darf Werte für eine beliebige Auswahl von Spalten annehmen. Werte gehören einem einheitlichen Datentyp „String“ an und können im Zeitverlauf versioniert werden. Konzeptionell kann eine Tabelle somit als Abbildung von Tripeln der Form (Zeilen-ID, Spalten-ID, Zeitpunkt) auf beliebige Zeichenketten verstanden werden. Lese- und Schreibzugriffe auf einzelne Zeilen sind atomar, und es werden Transaktionen auf einzelnen Zeilen unterstützt (jedoch nicht zeilenübergreifend).

Technisch setzt Bigtable auf dem Google File System (GFS) auf und läuft in Server-Clustern, wobei einzelne Tabellen dynamisch in Zeilenbereiche, sogenannte *Tablets*, partitioniert und auf unterschiedliche Tablet-Server verteilt werden. Die Daten werden mit Hilfe des Paxos-Algorithmus konsistent repliziert, und neue Tablet-Server können zur Laufzeit hinzugefügt werden, wodurch Skalierbarkeit und Hochverfügbarkeit sichergestellt werden.

Der bekannteste Bigtable-Klon im Open-Source-Umfeld ist das in Java programmierte HBase[21], das aufbauend auf dem Hadoop Distributed File System (HDFS) Bigtable-ähnliche Funktionalität bereitstellt und beispielsweise bei Adobe und Twitter eingesetzt wird. Im Gegensatz zu dieser tabellenorientierten Speicherung verwaltet Amazon Dynamo [DHJ+07] als Key-Value Store Schlüssel-Wert-Paare. Zu einem Schlüsselwert kann ein beliebiger Wert als Blob (Binary Large Object) per PUT-Operation gespeichert und per GET-Operation erfragt werden. Es gibt weder Operationen, die mehrere Schlüssel betreffen (also auch keine Transaktionen) noch ein relationales Schema. Die Besonderheit von Dynamo ist, dass PUT-Operationen unabhängig von Netzwerk- und Serverausfällen jederzeit möglich sein sollen (always writable; z. B. sollen Kunden jederzeit neue Artikel in ihre Warenkörbe legen können, auch wenn der Server abstürzt, mit dem ein Browser bisher kommuniziert hat). Dies bedeutet insbesondere, dass im Falle von Netzwerkpartitionierungen inkonsistente Wert-Versionen zu einem Schlüssel entstehen können. Entsprechend bietet Dynamo lediglich Eventual Consistency als Konsistenzgarantie. Technisch nutzt Dynamo Partitionierung und Replikation der Daten basierend auf verteilten Hash-Tabellen (Distributed Hash Tables, DHT), wobei Vektoruhren eingesetzt werden, um unterschiedliche Wert-Versionen zu einem Schlüssel erkennen und verwalten zu können.

[21] `http://wiki.apache.org/hadoop/Hbase`

Viel zitierte Open-Source-Projekte, die ebenfalls auf verteilten Hash-Tabellen und Eventual Consistency aufbauen, sind Project Voldemort[22], das bei LinkedIn eingesetzt wird, das ursprünglich bei Facebook entwickelte Apache Cassandra[23] und Basho Riak[24].

C.4. NewSQL

Die NoSQL-Entwicklungen, insbesondere der Fokus auf Eventual Consistency, wurde bereits frühzeitig von kritischen Kommentaren begleitet. Beispielsweise hat Michael Stonebraker schon im Jahre 2010 in einem Blogeintrag der Communications of the ACM[25] Gründe angegeben, warum Unternehmen NoSQL-Systeme entweder nicht kennen würden oder nicht interessiert seien; seine Kern-Statements lauteten damals bereits (1) „No ACID Equals No Interest“, da für viele Unternehmen OLTP das Kerngeschäft ausmacht und dieses nicht ohne harte Konsistenzgarantien auskommt, (2) „A Low-Level Query Language is Death“, da Anwendungsprogrammierer nicht auf Block- oder Record-Ebene arbeiten wollen, und (3) „NoSQL Means No Standards“, was für viele Unternehmen schon deshalb von Bedeutung ist, weil sie viele Datenbanken betreiben und einheitliche Schnittstellen zwischen diesen benötigen.

In einem weiteren Blogeintrag[26] benutzte Michael Stonebraker 2011 den Begriff „New SQL“ für die wieder aufkommenden SQL-Datenbanken mit ACID-Konsistenz, die versprachen, den NoSQL-Datenbanken gleichwertige Skalierbarkeitseigenschaften aufzuweisen (etwa Clustrix, NimbusDB/NuoDB, VoltDB). Derartige Datenbanken zielen also auf den klassischen Datenbankmarkt, allerdings mit besonderem Fokus auf extreme Skalierbarkeit, etwa im Big-Data-Umfeld.

Eine etwas andere Art von Datenbank, die ebenfalls der Kategorie NewSQL zugerechnet wird, ist die F1-Datenbank [SEC+13] von Google. Selbst in Unternehmen wie Google, die ihr Geld auf völlig andere Weise als „Brick-and-Mortar“-Firmen verdienen, gibt es viele Anwendungen, die z.B. strenge Konsistenz in Gegenwart von weit verteilter Replikation erfordern, die sich ein dem Relationenmodell zumindest ähnliches Datenmodell wünschen und die ferner auf Versionierung setzen. Bei Google trifft dies z.B. auf das Anzeigengeschäft zu; das Unternehmen hat darauf mit der Entwicklung von Spanner [CDE+12] reagiert, einer hoch skalierbaren, global verteilten Mehrversionen-Datenbank

[22] http://project-voldemort.com/
[23] http://cassandra.apache.org/
[24] http://www.basho.com/riak/
[25] http://cacm.acm.org/blogs/blog-cacm/99512-why-enterprises-are-uninterested-in-nosql
[26] http://cacm.acm.org/blogs/blog-cacm/109710-new-sql-an-alternative-to-nosql-and-old-sql-for-new-oltp-apps/fulltext

mit synchroner, anwendungsbezogen konfigurierbarer Replikation. Spanner garantiert Hochverfügbarkeit durch Replikation über Data Center oder sogar über Kontinente hinweg. Erster Kunde für Spanner ist Googles F1-Datenbank [SEC+13], das Backend von Google AdWords mit SQL-Unterstützung, verteilten SQL-Anfragemöglichkeiten, optimistischer Transaktionskontrolle sowie Möglichkeiten zur Datenbank-Reorganisation und dynamischen Schema-Veränderungen.

Es stellt sich heraus, dass selbst Web-Applikationen wie das AdWords-System (welches wirtschaftliche Daten verwaltet und daher harten Anforderungen an Daten-Integrität und Daten-Konsistenz genügen muss) nicht ohne klassische ACID-Transaktionsgarantien auskommen, während Eventual Consistency von NoSQL-Systemen an dieser Stelle nicht ausreichend ist. F1-Transaktionen bestehen aus vielen Lese- und optional einer Schreib-Operation, welche die Transaktion beendet. Es werden Snapshot-, pessimistische und optimistische Transaktionen unterschieden und vom Scheduler unterschiedlich behandelt, wobei letzterer Typ die Voreinstellung ist. Optimistische Transaktionen können lange laufen, bestimmte Client-Fehler ausgleichen sowie Server-Failovers mitmachen. Da F1 die synchrone Replikation von Spanner mit dem Ziel höchster Verfügbarkeit über Rechenzentrumsgrenzen hinweg einsetzt (laut [SEC+13] mit Kopien in fünf Rechenzentren, je zwei an der US-Ost- und -Westküste, eines im Landesinneren), sind die Commit-Latenzen von schreibenden Transaktionen mit 50-150ms deutlich höher als bei den von Stonebraker genannten NewSQL-Datenbanken, was bei Google durch spezielle Schema-Cluster-Techniken und Programmierstrategien für Anwendungen ausgeglichen wird.

Die New SQL-Entwicklung und insbesondere Google zeigen, dass durch eine angemessene Systemarchitektur sowie eine geschickte Kombination von Datenbank-Technologien auch global verfügbare Informationssysteme gebaut werden können, die höchsten Anforderungen an Ausfallsicherheit, Skalierbarkeit, Konsistenz und Usability genügen können. Es ist davon auszugehen, dass andere Hersteller dieser Entwicklungsrichtung folgen werden, so dass nicht unbedingt von einem „Back to the Roots“ bei relationalen Datenbanken gesprochen werden kann, aber von einer klaren und zeitgemäßen Evolution, die mit ACID-Transaktionen, Skalierbarkeit und Hochverfügbarkeit bei globaler Verteilung die Stärken zweier paralleler Entwicklungslinien zusammenführt.

C.5. Map-Reduce

Während Replikation als Maßnahme zum Erhalt von Datenverfügbarkeit gilt, erweist sich Partitionierung als Schlüssel zur algorithmischen Behandlung zahlreicher Big-Data-Probleme. Partitionierung folgt im Wesentlichen dem „alten“ Prinzip des *Teile und Herr-*

sche (engl. *Divide and Conquer*), das bereits von Julius Caesar verwendet wurde und in Algorithmen der Informatik eine lange Tradition hat (z.B. bei Such- und Sortierverfahren). Wenn Daten in mehrere unabhängige Teile zerlegt werden, kann eine Verarbeitung dieser Teile parallel erfolgen, z.B. in verschiedenen Kernen einer Multicore-CPU, in unterschiedlichen CPUs eines Clusters oder sogar in verschiedenen Data Centern; die dort erzielten Einzelresultate müssen dann zu einem Gesamtresultat zusammengesetzt werden. Dies ist bereits die grundlegende Idee von Googles Map-Reduce-Ansatz [DG04] (US Patent 7,650,331 von Januar 2010), der aus der funktionalen Programmierung bekannte Funktionen höherer Ordnung zur Spezifikation von Berechnungen auf sehr großen Datenmengen heranzieht.

Map-Reduce kombiniert zwei Funktionen, *map* und *reduce*, die auf Schlüssel-Werte-Paare angewendet werden und die im Grundsatz dem Group-By von SQL gefolgt von einer Aggregation ähneln. Eine Map-Reduce-Berechnung arbeitet grundsätzlich wie in Abbildung 23 gezeigt: Eingabedaten werden in sog. „Chunks“ (Partitionen) bereitgestellt und stammen typischerweise aus einem verteilten Dateisystem. Die Chunks werden zunächst von Map-Tasks auf *Mapper* genannten Komponenten verarbeitet. Mapper konvertieren Chunks in eine Folge von Schlüssel-Wert-Paaren; wie genau dies passiert, hängt von der spezifischen Aufgabe ab und wird durch den Code bestimmt, welchen der Benutzer für die Map-Funktion bereitstellen muss. Als nächstes werden die von den Mappern erzeugten Zwischenergebnisse durch einen Master-Controller gesammelt und anhand ihrer Schlüssel gruppiert. Die Schlüssel und die ihnen zugeordneten Gruppen werden dann so an Reduce-Tasks weitergeleitet, dass alle Schlüssel-Wert-Paare mit demselben Schlüssel derselben Reducer-Komponente zugewiesen werden. Die Reducer bearbeiten jeweils einen Schlüssel und kombinieren alle mit diesem assoziierten Werte in der Weise, wie die Anwendung es erfordert und der zugehörige Code spezifiziert.

Als Beispiel aus [Whi12] betrachten wir die Analyse von Wetterdaten, die als langer String von Wetterstationen geliefert werden; wir interessieren uns für die Höchsttemperatur pro Jahr. Input-Daten können in diesem Fall wie in Abbildung 24 gezeigt aussehen. Die betreffende Wetterstation sendet (als Datenstrom!) regelmäßig lange Zeichenreihen, die entsprechend interpretiert werden müssen; jeder solche String enthält u.a. die Stations-ID, das Datum der Messung, Längen- sowie Breitengrad des Orts der Station und die aktuelle Temperatur.

Wir nehmen nun an, der folgende Input liege vor; die für die Bestimmung der maximalen Jahrestemperatur relevanten Teile sind kenntlich gemacht (und Temperaturwerte auf ganze Zahlen gerundet):

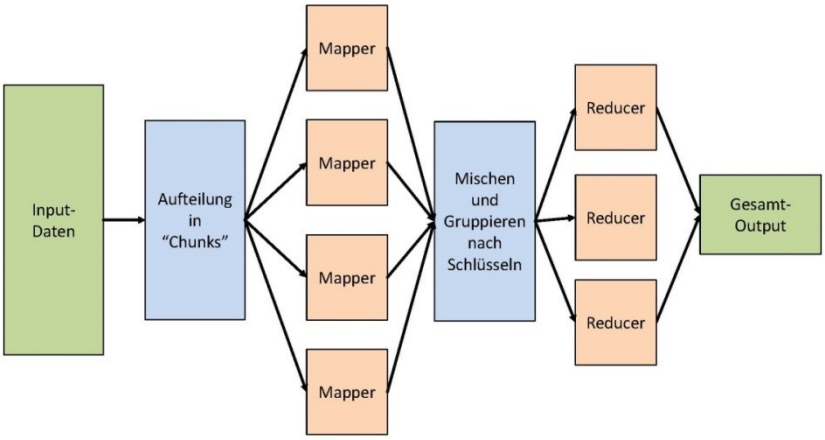


Abbildung 23: Prinzip einer Map-Reduce-Berechnung.

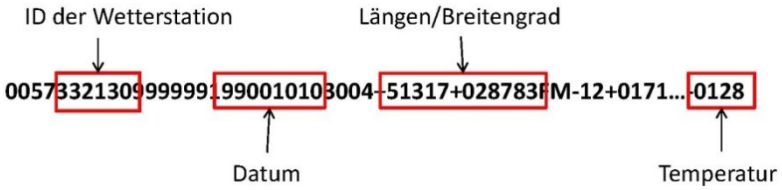


Abbildung 24: Exemplarische Wetterdaten.

```
0067011990999991990051507004+51317+028783FM-12+0171...+0000
0043011990999991990051512004+51317+028783FM-12+0171...+0022
0043011990999991990051518004+51317+028783FM-12+0171...-0011
0043011990999991989032412004+51317+028783FM-12+0171...+0111
0043011990999991989032418004+51317+028783FM-12+0171...+0078
```

Falls dies als ein Chunk an einen Mapper geleitet wird, extrahiert dieser das Jahr (als Schlüssel) und die Temperatur (als Wert) wie gewünscht und gibt aus: (1990, 0), (1990, 22), (1990, -11), (1989, 111), (1989, 78). Gruppierung anhand der Schlüsselwerte liefert (1989, [111, 78]), (1990, [0, 22, -11]), woraus ein Reducer nun die Maxima bestimmen kann als (1989, 111) und (1990, 22).

Es sollte einleuchten, dass eine Aufgabe wie diese, die in der Realität sehr große Datenmengen von Wetterstationen umfassen wird, die aber alle unabhängig voneinander verarbeitet werden können, ein perfekter Kandidat für eine Map-Reduce-Verarbeitung ist. Andere Aufgaben dieser Art sind z.B. das Zählen des Vorkommens bestimmter Wörter in einer Textsammlung (was z.B. für Indexerzeugung und -wartung bei einer Suchmaschine relevant ist), Matrix-Vektor-Multiplikation (relevant für eine Berechnung des PageRank zur Anordnung von Suchergebnissen) oder Operationen der Relationenalgebra in Datenbanken

(wie Joins oder Aggregat-Operationen, die z.B. im Kontext von Anfrage-Optimierung relevant sind) [LRU14].

C.6. Hadoop

Damit eine solche Map-Reduce-Berechnung bzw. -Ausführung funktioniert, sind verschiedene Fragen zu klären: Wie wird der Code für eine bestimmte Map-Reduce-Task erstellt? Wie lässt sich ein gegebenes Problem datenseitig so in Chunks zerlegen, dass diese parallel verarbeitet werden können? Wie lassen sich Tasks angemessen an Rechenknoten (die einen Mapper oder einen Reducer ausführen) zuweisen? Wie werden die Teile einer Berechnung, die auf verschiedenen Knoten stattfindet, synchronisiert? Wie lässt sich ein solches Szenario robust gegen Fehler bzw. Ausfälle machen?

Die erste dieser Fragen kann nur durch einen Benutzer beantwortet werden, der die benötigten Map- bzw. Reduce-Funktionen – typischerweise in einer Hochsprache wie Java – programmiert. Für das oben gezeigte Wetter-Beispiel findet sich der entsprechende Code beispielsweise in [Whi12]. Die weiteren Fragen wurden in den letzten Jahren auf unterschiedliche Weisen beantwortet; die bekannteste Antwort ist die Software-Bibliothek Hadoop[27], die bereits in Version 2 verfügbar ist. Hadoop [Whi12] unterstützt skalierbare Berechnungen, die auf verteilten Rechner-Clustern laufen. Die Kernkomponenten der ersten Hadoop-Version sind die Map-Reduce Engine sowie das Hadoop Distributed File System (HDFS). Erstere ist für Ausführung und Kontrolle von Map-Reduce-Jobs zuständig; HDFS ist ein verteiltes Dateisystem, in welchem große Datenmengen abgelegt, gelesen und ausgegeben werden können. Nutzerdaten werden in Blöcke unterteilt, die über die lokalen Speicher von Cluster-Knoten repliziert werden. HDFS basiert auf einer Master-Slave-Architektur, bei welcher ein Namenode als Master den Datei-Namensraum einschließlich der Datei-Block-Zuordnung und der Lokalisierung von Blöcken verwaltet und Datanodes als Slaves die eigentlichen Blöcke verwalten; Einzelheiten hierzu entnehme man z.B. dem HDFS Architecture Guide[28]. Neben diesen Hauptkomponenten gibt es zahlreiche Erweiterungen von Hadoop um spezifische Funktionalität, die zusammen das sog. Hadoop Ecosystem ergeben. Inzwischen wurden auch verschiedene Alternativen zu Hadoop vorgeschlagen (z.B. Disco, Skynet, Twister oder FileMap) sowie dessen Weiterentwicklung zu Hadoop NextGen Map-Reduce (YARN) bzw. Hadoop 2.0. Wer sich für Anwendungen bzw. Anwender von Hadoop interessiert, sei auf die Apache Website[29] verwiesen.

[27] `http://hadoop.apache.org/`
[28] `http://hadoop.apache.org/docs/stable/hdfs_design.html`
[29] `http://wiki.apache.org/hadoop/PoweredBy`

Das Map-Reduce-Paradigma und seine Implementierung Hadoop haben in den letzten Jahren nicht nur zahlreiche Entwicklungen angestoßen, die zu neuen kommerziell verfügbaren Produkten geführt haben, sondern auch eine Reihe von Forschungsaktivitäten [ASSU12, SASU13, ABE+14, Fed13, LRU14, SH13, Shi13], von denen hier Apache Flink, Apache Spark und AsterixDB kurz skizziert seien.

Eine typische Erweiterung das grundlegenden Map-Reduce-Paradigmas ist PACT, ein in [BEH+10] beschriebenes Programmiermodell, welches das Map-Reduce-Modell verallgemeinert durch Hinzunahme weiterer Funktionen sowie durch die Möglichkeit, Verhaltensgarantien für Funktionen zu spezifizieren. PACT wurde im Rahmen des Projekts Stratosphere [ABE+14] entwickelt, das inzwischen unter dem Namen Flink als Apache-Top-Level-Projekt entwickelt wird.[30] Während PACT „nur“ auf eine erweiterte Map-Reduce-Ausführungsumgebung abzielt, besteht die Kernidee von Flink darin, Datenverarbeitungs-Pipelines nicht länger prozedural auszuprogrammieren, sondern deklarativ zu spezifizieren (z. B. in der Skriptsprache Meteor). Wie in [ABE+14] im Detail dargestellt wird, werden diese deklarativen Programme automatisch optimiert und in PACT-Programme übersetzt, die ihrerseits weiter optimiert und dann in der Nephele genannten Ausführungsumgebung parallelisiert verarbeitet werden.

Das Apache Top-Level-Projekt Spark[31] geht aus von der Beobachtung, dass Map-Reduce für die Datenstromverarbeitung, für iterative Berechnungen und für interaktive Analysen weniger geeignet ist und stellt eine alternative verteilte Analyseumgebung bereit [ZCF+10]. Kernkomponente von Spark sind sog. Resilient Distributed Datasets (RDDs) [ZCD+12], die read-only Datenstrukturen im Hauptspeicher bereitstellen und sich in verschiedenen Programmiersprachen wie übliche Datenstrukturen nutzen lassen. Die im Cluster verteilt im Hauptspeicher vorliegenden Daten eignen sich dann insbesondere für Anwendungen des maschinellen Lernens, wobei Spark die Fehlertoleranz garantiert. Das Apache Top-Level-Projekt Mahout für maschinelles Lernen ist von den Vorteilen von Spark gegenüber Map-Reduce so überzeugt, dass die verwendete Ausführungsumgebung im April 2014 von Map-Reduce auf Spark umgestellt wurde und keine neuen Map-Reduce-Implementierungen mehr akzeptiert werden.[32]

Das Open-Source-System AsterixDB[33] wird von seinen Entwicklern als Big Data Management System bezeichnet, das folgende Anforderungen erfüllen soll [AB14]:

- Ein flexibles, semistrukturiertes Datenmodell

[30] `https://flink.apache.org/`
[31] `https://spark.apache.org/`
[32] `https://mahout.apache.org/`
[33] `http://asterixdb.ics.uci.edu/`

- Eine mindestens ebenso mächtige Anfragesprache wie SQL
- Eine effiziente, parallele Ausführungsumgebung
- Unterstützung für Datenmanagement und automatische Indexierung
- Unterstützung für verschiedene Anfragegrößen
- Unterstützung für kontinuierlich eingehende Daten
- Skalierbarkeit angesichts großer Datenvolumina in großen Clustern
- Unterstützung für Variabilität in Big-Data-Szenarien, z. B. textuelle, temporale und räumliche Daten

Um diese Anforderungen zu unterstützen, die weder von Map-Reduce- noch NoSQL-Systemen noch parallelen Datenbanken gleichzeitig erfüllt werden, setzt AsterixDB auf eine Shared-Nothing-Cluster-Architektur mit eigenem Datenmodell, eigener deklarativer Anfragesprache (AQL) und eigener Ausführungsumgebung (Hyracks). AQL-Anfragen werden in eine algebraische Darstellung übersetzt und optimiert und dann in Hyracks zur Ausführung gebracht. Die in [AB14] berichteten experimentellen Ergebnisse zeigen ein Leistungsverhalten, das mit dem spezialisierter Systeme (verglichen wurde gegenüber einer nicht benannten relationalen Datenbank, MongoDB und Hive) mithalten kann, was vor allem deshalb vielversprechend erscheint, weil AsterixDB das kombinierte Anwendungsspektrum dieser Systeme abdecken soll. Die Publikation weitergehender Leistungsmessungen ist angekündigt; es bleibt also spannend.

Big-Data-Technologien adressieren oft spezifische Probleme mit bestimmten Charakteristika (wie etwa Hadoop Map-Reduce, das auf Stapelverarbeitung ausgelegt ist) anstatt alle Klassen von Problemen mit einer Lösung zu handhaben („one size fits all“), was insgesamt zwar zu einem breiteren Spektrum an Analysemöglichkeiten führen kann, aber auch die Komplexität erhöht, da z.B. mehrere Produkte und Datenquellen zusammengeschaltet werden müssen, um die Daten aus diesen zu Informationen vereinen zu können. Dies ist ebenso der Fall, wenn z.B. eine bestehende, klassische Data-Warehouse-Lösung mit Big-Data-Produkten für unstrukturierte Daten ergänzt wird.

Wenn für jede Aufgabe während der Datenanalyse, z.B. unstrukturierte, nicht-bereinigte Daten, ein dafür zugeschnittenes System eingesetzt wird, muss man diese auf geeignete Weise verbinden können, um die Informationen daraus adäquat verknüpfen zu können. Dies ist eine technische Herausforderung, da die Systeme nicht nur unterschiedliche Schnittstellen haben, sondern auch der Zugriff teils grundlegend anders erfolgt. Beispielsweise

muss zwischen einer deklarativen SQL-Anfrage, die im traditionellen Data-Warehousing-Kontext üblich ist und das „wie“ der Anfrage nicht vorgibt, und einem imperativ programmierten Map-Reduce-Job in Java vermittelt werden. Entweder können die Produkte den Zugriff auf andersartige Tools vermitteln, wie etwa einige SQL-Datenbanken, die Daten aus HDFS in eine virtuelle Tabelle einbinden können. Die andere Möglichkeit ist eine Vermittlungs- oder Integrationsschicht im Stile einer Middleware, die zwischen den Tools vermittelt und Aufgaben adäquat an diese verteilt.

Nach gängiger Auffassung dient das Data Warehouse als zentraler Ort für die wahrhaftige Darstellung von Information („single point of truth“). Während dies in der Vergangenheit meist bedeutete, dass hierfür ein physisches System bzw. ein einzelnes Produkt eingesetzt wurde, muss dies heute bei Einsatz mehrerer Produkte eher virtuell verstanden werden. Das Data Warehouse der Zukunft kann also eher einem „Logical Data Warehouse“ entsprechen, das aus mehreren Produkten aufgespannt wird. Darunter können klassische SQL-basierte Data Warehouses aber auch neuere Big-Data-Tools, wie NoSQL-Datenbanken oder Tools aus Apache Hadoop sein. Je mehr Tools zusammen zu einem komplexeren gesamten vereint werden, desto komplexer wird auch der Entwicklungs- und Aufbauprozess sowie der Betrieb der Gesamtarchitektur.

C.7. Anwendungsbeispiel Facebook

Die Anforderungen des bekannten sozialen Netzwerks „Facebook“ erforderten an vielen Stellen neue Denkweisen in Bezug auf die Nutzung von Technologie und stellten mit die Weichen für die Weiterentwicklung von z.B. NoSQL-Datenbanken. Des Weiteren ist auch die reine Datenmenge bei Facebook heutzutage in der Big-Data-Kategorie zu verorten, da dort täglich mehr als 500 TB neue Daten erzeugt werden. Quelle sind unter anderem mehr als 300 Millionen hochgeladene Bilder sowie mehr als 2 Milliarden „Likes“ und geteilte Beiträge[34]. Facebook verzichtet hierbei aber auch nicht vollständig auf die Eigenschaften relationaler DBMS, sondern setzt auf neue Lösungen, die besonders skalierbar sind und auf HDFS/Hadoop aufsetzen, um die anfallenden Datenmengen handhaben zu können[35]. Während z.B. „Likes“ in einer modifizierten Version einer relationalen Datenbank gesichert werden, aber Nachrichten vom Messenger in Apache HBASE, verwendet Facebook Apache Hive und Presto als Data-Warehousing-Lösung. Das Apache-Projekt Hive hat das Ziel, die Anfragesprache SQL in einer Hadoop-Umgebung einzusetzen, um die Vorteile relationaler Systeme, wie etwa deklarative Anfrageverarbeitung und schnelle Antwortzeiten, mit einer

[34] `http://internet.org/efficiencypaper`
[35] `http://hortonworks.com/big-data-insights/how-facebook-uses-hadoop-and-hive/`

Abbildung 25: Mögliche Big Data-Adoptionsstrategie.

hohen Skalierbarkeit zu vereinen. Dabei werden im Hintergrund MapReduce-Aufgaben aus dem SQL generiert. Mit Presto, einer Neuentwicklung von Facebook und mittlerweile Open Source, sollen verteilte SQL-Anfragen „ad-hoc“ und interaktiv mit hoher Geschwindigkeit durchgeführt werden[36]. Das System bei Facebook muss dabei sowohl klassische Reports unterstützen, aber auch Ad-hoc-Anfragen, die viele unterschiedliche Datenquellen mit einbeziehen, wie etwa den Messenger oder den Facebook News Feed.

D. Zur organisatorischen Dimension von Big Data

Wir werfen als Nächstes einen Blick auf die Frage, wie ein Unternehmen Nutzen aus Big Data ziehen kann. Was ist zu tun bzw. zu verändern, wenn das Unternehmen bereits ein Data Warehouse für Datenanalysezwecke aufgesetzt hat? Wir skizzieren hier eine Modifikation der „klassischen“ Data Warehouse-Architektur so, dass Big Data-Anforderungen Rechnung getragen wird.

Wie bei vielen anderen Einführungsentscheidungen bzgl. IT auch, die man im Laufe der letzten 30 Jahre zu fällen hatte, macht es Sinn, eine Entscheidung, ob ein Big Data-Projekt aufgesetzt oder ob in Big Data-Technologie investiert werden soll, wohlfundiert zu treffen. Dabei kann zunächst ein Blick auf Anwendungsbereiche helfen, die vergleichbar sind. Ferner erscheine allgemeine Techniken wie z.B. eine SWOT-Analyse sinnvoll, mit welcher sich Stärken, Schwächen, Chancen und Risiken einer bestimmten Technologie oder eines Projektvorhabens analysieren lassen. Alternativ kann zur Entscheidungsfindung eine Kontextanalyse herangezogen werden, die nach Zielen und der Eigenschaft Value fragt, welcher von dem Projekt zu erwarten ist. Beide Werkzeuge sind populär und haben sich etwa im Zusammenhang mit Geschäftsprozessmodellierung bewährt [SVOK12]. Umfassender als spezifische Analysen ist die Entwicklung einer Big Data-Strategie, die z.B. wie in Abbildung 25 gezeigt aussehen kann.

Die Strategie startet mit einer Phase der Informationsbeschaffung und Planung, was z.B. eine SWOT- oder eine Kontextanalyse (oder beides) umfassen kann und die Definition des Business Case enthalten sollte. Falls eine Entscheidung zugunsten eines Big Data-Projekts

[36] https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920

oder der Adoption von Big Data-Technologie gefällt wird, müssen zunächst relevante Datenquellen identifiziert und selektiert werden; ein Unternehmen wird typischerweise eine Reihe interner Quellen wie vorhandene Datenbanken, aber und vor allem externe Quellen insbesondere aus dem Web nutzen wollen, da gerade letztere relevante Daten (etwa aus öffentlichen Blogs) liefern können. Die dritte Phase besteht aus einer Detailplanung und umfasst die Wahl der zu verwendenden Technologie, etwa die Entscheidung für eine bestimmte Hadoop-Implementierung. Sodann kann eine Realisierung stattfinden; schließlich ist das System bzw. das Projekt operational und benötigt ggfs. regelmäßige oder gelegentliche Wartung.

Wir gehen hier nicht auf weitere Einzelheiten einer Strategieentwicklung ein, bemerken jedoch, dass es auch für Unternehmen, deren Kernkompetenz nicht in der IT liegt, sinnvoll sein kann, von den modernen Möglichkeiten zur Datenanalyse Gebrauch zu machen. Man spricht in diesem Zusammenhang gerne von „Business Intelligence“ (BI), und über viele Jahre war das Werkzeug der Wahl für jegliche BI-Anwendung das Data Warehouse [Inm05], gemeinhin verstanden als eine Datenbank, die separat von operationalen Systemen im Rahmen eines ETL-Prozesses aufgebaut wird, der Daten aus den relevanten Quellen extrahiert, in das Warehouse-Schema transformiert und schließlich Daten aus den Quellen ins Warehouse lädt. Das Ergebnis bildet dann die Grundlage für Anwendungen wie Online Analytical Processing (OLAP), Planung, Reporting, Ad-hoc-Anfragen, Spreadsheets, Dashboards oder Data- Mining-Applikationen [HKP11].

Eine BI-Architektur aus abstrakter Sicht zeigt Abbildung 24. Wesentlich ist hier, dass zunächst nicht spezifiziert wird, auf welche Weise die zentrale BI-Funktionalität realisiert wird; dies erfolgt traditionell über ein Data Warehouse, wird jedoch inzwischen häufig zumindest um ein Hadoop-Setup ergänzt, wenn nicht gar abgelöst. Wesentlich ist ferner, dass Abbildung 26 noch keine Festlegung trifft, ob Daten lokal oder etwa in der Cloud gehalten werden.

Abbildung 27 zeigt demgegenüber eine konkrete BI-Architektur. Neben einem Data Warehouse mit seinen Datenquellen, seiner Staging Area und seinen Data Marts ist hier erkennbar, wie sich eine traditionelle Data Warehouse-Architektur für Big Data-Anwendungen erweitern lässt. Neu sind gegenüber einer klassischen Data Warehouse-Architektur z.B. die weiter gefasste Selektion externer Datenquellen sowie die Erweiterung um eine Map-Reduce-Engine wie z.B. Hadoop (linke Hälfte der Abbildung), was auch bedeutet, dass hier Datei- und Datenbanksysteme neben- und miteinander existieren. Es werden natürlich zusätzliche Kommunikationswege zwischen den alten und den neuen Komponenten benötigt; das Ergebnis kann dann wie in der Abbildung gezeigt aussehen.

Es sei erwähnt, dass Business Intelligence- sowie Analyse-Anwendungen nicht notwen-

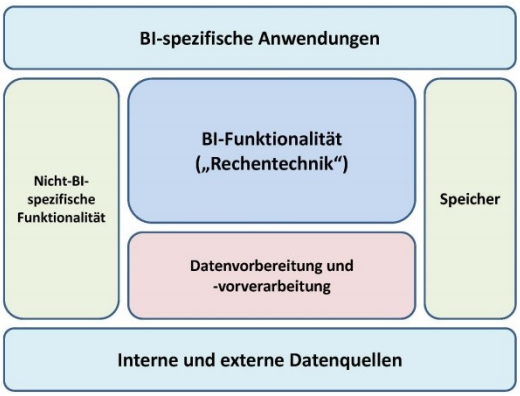


Abbildung 26: Abstrakte BI-Architektur.

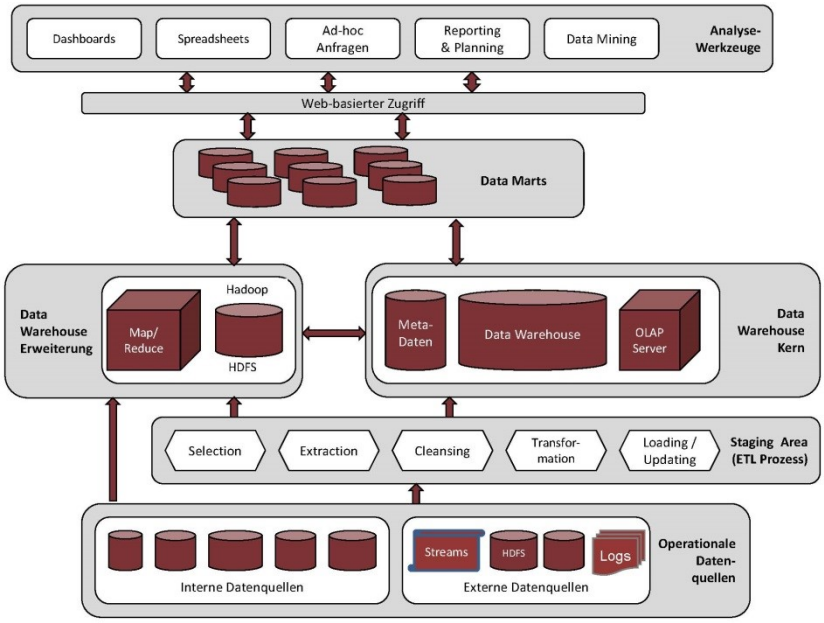


Abbildung 27: Für Big Data-Nutzung erweiterte Data Warehouse-Architektur.

digerweise das Vorhandensein eines Data Warehouse erfordern. Viele der heute verfügbaren Werkzeuge können als Aufsatz oder Erweiterung eines operationalen Systems oder einer Datenbank betrieben werden, das bzw. die im Unternehmen bereits im Einsatz ist. In einem solchen Fall ist ein expliziter Architekturentwurf nicht erforderlich, allerdings zeigt die Erfahrung, dass insbesondere gut dokumentierte Strategie- sowie Architekturüberlegungen ein Unternehmen vor einem Scheitern entsprechender Projekte schützen können.

E. Blogs und News-Seiten zum Thema

- Data Science Central Portal zu Big Data:
  `http://www.datasciencecentral.com/`
- Dataversity News Portal: `http://www.dataversity.net/`
- Datanami News Portal: `http://www.datanami.com/`
- The Big Data Landscape: `http://www.bigdatalandscape.com/`
- DataFloq: `https://datafloq.com`
- AnalyticsWeek: `http://analyticsweek.com`
- "'The Home of Data Science": `https://www.kaggle.com/`
- Radar von O'Reilly Media zum Thema Internet of Things:
  `http://radar.oreilly.com/iot`

Literatur

[AB14] Sattam Alsubaiee and Alexander Behm. Storage Management in AsterixDB. *Proceedings of the . . .* , pages 841–852, 2014.

[ABE+14] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, Felix Naumann, Mathias Peters, Astrid Rheinländer, Matthias J. Sax, Sebastian Schelter, Mareike Höger, Kostas Tzoumas, and Daniel Warneke. The Stratosphere platform for big data analytics. *The VLDB Journal*, 2014.

[ASSU12] Foto N Afrati, Anish Das Sarma, Semih Salihoglu, and Jeffrey D Ullman. Vision Paper: Towards an Understanding of the Limits of Map-Reduce Computation. *Arxiv preprint arXiv*, pages 1–5, 2012.

[BEH+10] D Battré, S Ewen, F Hueske, O Kao, V Markl, and D Warneke. Nephele/PACTs: a programming model and execution framework. In *SOCC*, pages 119–130, 2010.

[Bre12] E. Brewer. CAP twelve years later: How the "rules"have changed, 2012.

[CDE+12] James C Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J J Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. Spanner : Google's Globally-Distributed Database. In *Proceedings of OSDI'12: Tenth Symposium on Operating System Design and Implementation*, pages 251–264, 2012.

[CDG+08] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah a. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A Distributed Storage System for Structured Data, 2008.

[CHT00] Colleen Cook, Fred Heath, and Russel L Thompson. A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys. *Educational and Psychological Measurement*, 60(6):821–836, 2000.

[CP14] Mark Cavage and David Pacheco. Bringing arbitrary compute to authoritative data. *Communications of the ACM*, 57(8):40–48, 2014.

[Dey97] Eric L. Dey. Working with low survey response rates: The efficacy of weighting adjustments. *Research in Higher Education*, 38(2):215–227, 1997.

[DG04] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, page 10, Berkeley, CA, USA, 2004. USENIX Association.

[DHJ+07] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon's highly available key-value store. In *Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles - SOSP '07*, page 205, 2007.

[Eic89] M H Eich. Main memory database research directions. In H Boral and P Faudemay, editors, *Database Machines Sixth International Workshop, IWDM '89*, chapter Main Memor, pages 251–268. Springer-Verlag New York, Inc., New York, NY, USA, 1989.

[FB99] A. Fox and E.A. Brewer. Harvest, yield, and scalable tolerant systems. *Proceedings of the Seventh Workshop on Hot Topics in Operating Systems*, 1999.

[Fed13] Gilles Fedak. Special Issue: MapReduce and its Applications. *Concurrency Computation Practice and Experience*, 25(1):1, 2013.

[FGC+97] A Fox, S Gribble, Y Chawathe, E Brewer, and P Gauthier. Cluster-based scalable network services. *Proceedings of the sixteenth ACM symposium on Operating systems principles*, page 91, 1997.

[GL02] Seth Gilbert and Nancy Lynch. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services, 2002.

[HKP11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[Inm05] William Inmon. *Building the Data Warehouse*. Wiley, Burlington, MA, 4th editio edition, 2005.

[Lam98] Leslie Lamport. The part-time parliament, 1998.

[Lew04] M Lewis. *Moneyball: The Art of Winning an Unfair Game*. Norton & Company, 2004.

[LLV+11] Peter Loos, Jens Lechtenbörger, Gottfried Vossen, Alexander Zeier, Jens Krüger, Jürgen Müller, Wolfgang Lehner, Donald Kossmann, Benjamin Fabian, Oliver Günther, and Robert Winter. In-memory databases in business information systems, 2011.

[LRU14] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of Massive Datasets*. Palo Alto, CA, 2014.

[Mac15] Richard MacManus. *Trackers: How Technology is Helping Us Monitor & Improve Our Health*. David Bateman Ltd, New Zealand, 2015.

[PZ11] Hasso Plattner and Alexander Zeier. *In-Memory Data Management*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[RW12] Eric Redmond and Jim R Wilson. *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*. Pragmatic Bookshelf, 2012.

[SASU13] Anish Das Sarma, Foto N. Afrati, Semih Salihoglu, and Jeffrey D. Ullman. Upper and lower bounds on the cost of a map-reduce computation. *Journal Proceedings of the VLDB Endowment*, pages 277–288, 2013.

[SEC+13] Jeff Shute, Stephan Ellner, John Cieslewicz, Ian Rae, Traian Stancescu, Himani Apte, Radek Vingralek, Bart Samwel, Ben Handy, Chad Whipkey, Eric Rollins, Mircea Oancea, Kyle Littlefield, and David Menestrina. F1: a distributed SQL database that scales. *Proceedings of the VLDB Endowment*, 6(11):1068–1079, 2013.

[SH13] Caetano Sauer and Theo Härder. Compilation of Query Languages into MapReduce. *Datenbank-Spektrum*, pages 1–11, 2013.

[Shi13] Kyuseok Shim. MapReduce Algorithms for Big Data Analysis. In Aastha Madaan, Shinji Kikuchi, and Subhash Bhalla, editors, *Databases in Networked Information Systems SE - 3*, volume 7813 of *Lecture Notes in Computer Science*, pages 44–48. Springer Berlin Heidelberg, 2013.

[SM13] Michael Saecker and Volker Markl. Big data analytics on modern hardware architectures: A technology survey. In *Lecture Notes in Business Information Processing*, volume 138 LNBIP, pages 125–149, 2013.

[SVOK12] Frank Schönthaler, Gottfried Vossen, Andreas Oberweis, and Thomas Karle. *Business Processes for Business Communities*. Springer Berlin Heidelberg, 2012.

[VHH12] Gottfried Vossen, Till Haselmann, and Thomas Hoeren. *Cloud Computing für Unternehmen - Technische, wirtschaftliche, rechtliche und organisatorische Aspekte*. d.punkt Verlag, Heidelberg, 2012.

[VL15] Gottfried Vossen and Jens Lechtenbörger. NoSQL, NewSQL, MapReduce und Hadoop. In P. Chamoni and P. Gluchowski, editors, *Analytische Informationssysteme: Business Intelligence-Technologien und -Anwendungen*. Springer-Verlag, Berlin, 5. auflage edition, 2015.

[Vog08] Werner Vogels. Eventually Consistent, 2008.

[Vos01] Gottfried Vossen. Vernetzte Hausinformationssysteme — Stand und Perspektiven. 2001.

[Vos13] Gottfried Vossen. Big Data as the New Enabler in Business and Other Intelligence. *Vietnam Journal of Computer Science*, 1(1):3–14, 2013.

[Vos14] Gottfried Vossen. Big Data: Der neue Katalysator für Business und andere Intelligenz. In T. Schwarz, editor, *Leitfaden Marketing Automation*, pages 51–72. marketing-BÖRSE GmbH, Wahäusel, 2014.

[Whi12] Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, Sebastopol, CA, 3rd editio edition, 2012.

[ZCD+12] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, page 2, Berkeley, CA, USA, 2012. USENIX Association.

[ZCF+10] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, page 10, Berkeley, CA, USA, 2010. USENIX Association.