Konzeption einer Big-Data-Architektur zur Optimierung und Flexibilisierung der Automobilproduktion

Von der Fakultät für Maschinenbau, Elektro- und Energiesysteme
der Brandenburgischen Technischen Universität Cottbus–Senftenberg
zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

genehmigte Dissertation

vorgelegt von
Master of Science

Anh Duc Tran

geboren am 18. Juni 1986 in Phu Tho, Vietnam

Vorsitzender: apl. Prof. Dr.-Ing. habil. Dr. paed. Annette Hoppe

Gutachter: Prof. Dr.-Ing. Uwe Meinberg

Gutachter: Prof. Dr. rer. pol. habil. Magdalena Mißler-Behr

Tag der mündlichen Prüfung: 16. Juli 2021

Vermerk

Die vorgelegte Dissertation beinhaltet vertrauliche und inhaltliche Daten. Aus diesem Grund wird das Unternehmen, welches gleichzeitig als Projektpartner für die Herausgabe der erarbeiteten Daten zuständig war, pseudonymisiert. Gleiches gilt für Zulieferer und Unternehmenspartner, die an dem Projekt beteiligt waren.

In dieser Arbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mit gemeint, soweit es für die Aussage erforderlich ist.

Inhaltsverzeichnis

I. Abbildungsverzeichnis

II. Tabellenverzeichnis

III. Abkürzungsverzeichnis

BI:	Business Intelligence
bspw.:	beispielsweise
bzw.:	beziehungsweise
DB:	Datenbank
DBMS:	Datenbank-Managementsystem
DBS:	Datenbanksystem
BMW:	Bayerische Motoren Werke
BMWi:	Bundesministerium für Wirtschaft
CoxPH:	Cox Proportional Hazard
CPS:	Cyber-physical Systems
CRM:	Customer-Relationship-Management
DLZ:	Durchlaufzeit
EDI:	Electronic Data Interchange
EHB:	Elektrohängebahn
ERP:	Enterprise Ressource Planning
ETL:	Extract, Transform, Load
FA:	Fertigungsabschnitt
GP:	Geschäftsprozess
GWT:	Großer Werkstückträger
IIoT:	Industrielles Internet der Dinge
IKT:	Informations- und Kommunikationstechnologien
IoT:	Internet of Things
IT:	Informationstechnologie
JIS:	Just in Sequence
JIT:	Just in Time
KPI:	Key Performance Indicators
KVP:	Kontinuierlicher Verbesserungsprozess
MES:	Manufacturing Execution System
NIST:	National Institute of Standards and Technology viii
OEE:	Overall Equipment Efficiency
OEM:	Original Equipment Manufacturer
OLAP:	Online Analytical Processing
PKW:	Personenkraftwagen
PM:	Predictive Maintenance
PMON:	Process Monitor
RFID:	Radio Frequency Identification
SCADA:	Supervisory Control and Data Acquisition
SCM:	Supply-Chain-Management
SPS:	Speicherprogrammierbare Steuerung
TPM:	Total Productive Maintenance
TQM:	Total-Quality-Management
TUL:	Transport, Umschlag, Lagerung
u.a.:	unter anderem
z.B.:	zum Beispiel

1. Einführung

"In a time of drastic change, it is the learners who inherit the future. The learned find themselves equipped to live in a world that no longer exists."

— Warren G. Bennis

Seit dem Tag, an dem die Bundesregierung das Thema Industrie 4.0[1] ins Leben berufen hat, um die Digitalisierung und deren Potenziale zu fördern, wurden innovative Ansätze in der Wirtschaft vorangetrieben. Dies betrifft den Primärsektor, weitestgehend den Sekundärsektor und letztlich auch den Tertiärsektor. Aus der historischen Entwicklung heraus nahm der Primärsektor mit der Agrarwirtschaft, der sogenannten „Urproduktion", stetig ab und der Sekundärsektor durch die Industrialisierung im späten 19. Jahrhundert zu. Jedoch ist die Entwicklung des Tertiärsektors nicht zu vernachlässigen. Genau in diesem steckt auch das Potenzial des Digitalisierungszeitalters (siehe Abbildung 1). Mit der rasanten Entwicklung der Technologien in Bezug auf Robotik und Informatik wird ein Großteil der Fertigungs- und Produktionsprozesse zunehmend automatisiert. Datenwissenschaften (im Engl. „Data Science" genannt) ist unlängst ein Trendphänomen des 21. Jahrhunderts mit unglaublichem Potenzial in Wirtschaft und Gesellschaft. Das gilt zunehmend auch in der Automobilproduktion, wo Prozesse und Produkte digitaler werden.

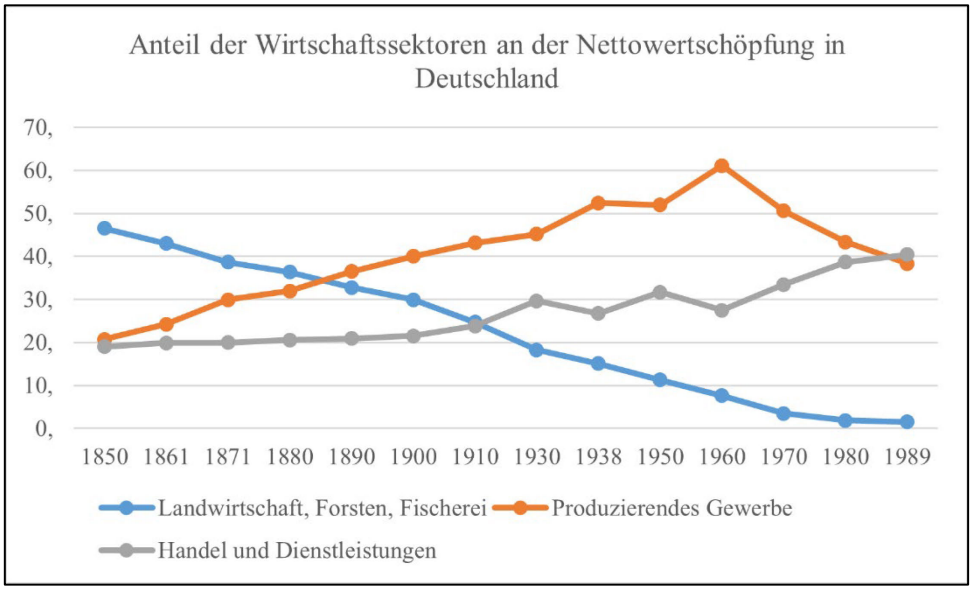


Abbildung 1: Entwicklung Wirtschaftssektoren in Deutschland (in Anlehnung an die Studie der Uni Münster, 2012)

[1] Bekanntgabe in der Öffentlichkeit erstmals 2011 auf der Hannover-Messe.

Die allgegenwärtige Verzahnung von Technologie und Wirtschaft unterstreicht den Stellenwert und die Bedeutung von Daten. Nun ist die Digitalisierung einer Herausforderung ausgesetzt, die sowohl die Technologie als auch den Menschen betrifft. Umso wichtiger erscheint das Verständnis von Maschinen und der damit verbundenen Daten. Vor diesem Hintergrund befasst sich diese Arbeit mit dieser eben genannten Herausforderung. Das Zusammenspiel zwischen dem Sekundärsektor mithilfe eines Vertreters aus der Automobilindustrie und dem Tertiärsektor anhand der Bereitstellung von Dienstleistungen aus dem IT-Sektor wird als Grundlage genommen. Daraus entstand das Big-Data-Projekt mit einem der führenden Automobilhersteller weltweit, welches in den nächstfolgenden Kapiteln beschrieben wird.

1.1 Problemstellung

Die Automobilindustrie steht vor einem bedeutsamen und zugleich herausfordernden Zeitalter – dem Zeitalter der Daten, Technologien und der Globalisierung. Die in dieser Arbeit behandelte Problematik zielt sowohl auf die strategische Ausrichtung der Automobilhersteller als auch in der operativ wirkenden Produktion ab. Die steigenden Anforderungen an Produkte vonseiten der Konsumenten stellen die Produktion vor neue Aufgaben, dies gilt sowohl für die Anforderungen an den Menschen als auch an die Maschine. In der aktuellen Situation stehen die traditionellen Automobilhersteller und ihre Lieferanten unter Zugzwang. Die Digitalisierung fördert junge Unternehmen, die in den Markt drängen, und treibt zugleich die Innovation an. Als Pioniere in ihrem Segment befinden sich große Konzerne, in diesem Fall die Automobilindustrie, in einer Vorreiterrolle. Vor dem Hintergrund der umweltpolitischen Lage soll hier das Ziel einer klimaneutralen Produktion erarbeitet werden. In diesem Kontext stellt sich die Frage, ob die Automobilindustrie dieser Aufgabe überhaupt gerecht werden kann. In der Dissertation wird diese Problematik jedoch nur hinsichtlich der Produktion behandelt. Im Speziellen wird auf die operative Ebene eingegangen und das Beispiel von unvorhersehbaren Maschinenausfällen aufgegriffen. Dabei soll ein Konzept erarbeitet werden, welches die Datenflut beherrschbar macht. In der Automobilindustrie besteht besonders bei den Prozessen der Lackiererei und im Karosseriebau ein sehr hoher Automatisierungsgrad. Durch die Produktion mit Robotern und Maschinen entstehen dabei Produktionsdaten, die in unterschiedlichen Systemen bearbeitet werden. Die Schwierigkeit in der Analyse ist demnach eine umfassende Datenaufbereitung. Das Projekt, welches mit dem Automobilhersteller initiiert

wurde, befasst sich nun mit ebendieser Problematik. Mit einem Maschinenausfall geht immer ein Bandstillstand und somit ein temporärer Produktionsstopp einher. Die Folgen sind ineffiziente Produktionsprozesse, intralogistische Lieferschwierigkeiten und nicht optimal planbares Ressourcen-Management Eine Optimierung der Produktion wird durch datengetriebene Prozesse erreicht. Dazu ist eine technische Betrachtungsweise in Verbindung mit den vorhandenen IT-Infrastrukturen notwendig. Somit entstehen drei Forschungsfragen, die im Laufe der Arbeit beantwortet werden sollen:

1. Ist eine Optimierung hochautomatisierter Produktionsprozesse mit einer sehr geringen Ausfallquote durch neue Ansätze der Informationstechnik möglich?
2. Ist eine Implementierung unter Berücksichtigung der technischen Restriktion möglich?
3. Welche Strategien resultieren aus den Ergebnissen der Arbeit?

1.2 Zielsetzung der Arbeit

Das oberste Ziel ist eine Teiloptimierung in der Produktion. Hierbei ist eine interdisziplinäre Herangehensweise sehr wichtig, da das Zusammenspiel von Produktion, IT und Instandhaltung essentiell für die Optimierung ist. Der Autor wird in der Arbeit ein Konzept ausarbeiten, welches das IT-Management zur Orchestrierung der Datensysteme nutzen kann, um eine automatische Analyse zu erzeugen. Als Evaluierung dient die Fallstudie im operativen Produktionsmanagement. Der Produktionsprozess einer vollautomatisierten Kleberanlage soll demnach nach Plan ohne unvorhersehbare Stopps ausgeführt werden. Mit einem Predictive-Maintenance-Ansatz sollen Maschinenfehler aufgezeigt werden, die möglicherweise in der nahen Zukunft auftreten können. Um dies zu erreichen, wird eine Prozessanalyse des Karosseriebaus bezüglich des Klebeprozesses durchgeführt. Zu dieser klassischen Analyse wird der Autor Werkzeuge aus der Informationstechnik und Informatik heranziehen, um neue Optimierungsmöglichkeiten im Datenmanagement herauszufinden. Dazu wird ein Konzept für ein automatisiertes Datenmanagement erarbeitet. Mit einer Big-Data-Architektur soll die Grundlage dafür geschaffen werden. Dabei soll das Konzept sowohl den Anforderungen in der Praxis gerecht werden als auch Allgemeingültigkeit beanspruchen. Unternehmensinterne und externe Analysen sowie Maschinen- und Prozessanalysen werden dadurch ermöglicht. Mit der Hilfe von Machine Learning soll der Datenprozess zudem automatisiert und optimiert werden. Die Arbeit soll somit den Bogen von der IT bis hin zur Optimierung im Produktions- und

Prozesssinne spannen. Disruptive Ansätze in der Nutzung von Software-Tools aus dem Open-Source-Segment sollen eine Ergänzung zum bisherigen Informationssystem darstellen.

Die Validierung soll durch eine Implementierung der in der Arbeit erörterten Big-Data-Architektur zur Anwendung von ML für Predictive Maintenance am Beispiel des Klebeprozesses im Karosseriebau erfolgen.

1.3 Aufbau der Arbeit

Der Aufbau der Arbeit folgt dem analytischen Gedankengang des Autors. Die Vorgehensweise orientiert sich am Weg vom strategischen zum operativen Management und mündet in der Fertigung. Die ersten beiden Kapitel dienen zur theoretischen Einordnung des Themenkomplexes. Am Anfang steht eine ganzheitliche Betrachtung der Produktion, um ein Grundverständnis für den Leser zu schaffen. Demnach beginnt das Kapitel 2 mit der Produktion und geht auf die Einordnung und den Prozess des Karosseriebaus ein. Unter diesem Aspekt wird die Instandhaltung als Analysethema miteinbezogen. Dort werden unterschiedliche Strategien aufgezeigt, die zur Erarbeitung eines Produktionsprozesses herbeiführen. Aus dieser Sicht wird die informationstechnische Betrachtung und Analyse Aufschluss darüber geben, inwiefern eine computer- oder datenbasierte Unterstützung existiert. Eine weitere Möglichkeit wird sodann in Kapitel 3 behandelt, denn dort wird auf die Datenwissenschaft eingegangen. Hier wird u.a. ein Big-Data-Konzept erarbeitet, welches auf einer vom Autor entwickelten Architektur beruht. Mit dieser wird dann ein Konzept des Predictive Maintenance und Möglichkeiten der Datenanalyse mit Machine Learning und Ereigniszeitanalysen im Deep-Learning-Ansatz vorgestellt. Zum Ende der Arbeit werden die Ergebnisse zusammengefasst und durch statistische Methoden validiert. Des Weiteren wird das Modell im laufenden Betrieb mit Daten angereichert, sodass eine Validierung im Werk geschieht. Zum Abschluss werden noch Auswirkungen auf die Automobilindustrie insgesamt und auf die Produktion im Speziellen dargelegt sowie disruptive Potenziale der Technologie aufgezeigt.

2. Prozessbetrachtung in der Automobilindustrie

Der in der Dissertation analysierte Produktionsprozess wurde anhand der Methode des Design Thinking mit dem Projektpartner erörtert. Im Zuge dessen wurden Workshops veranstaltet, um die Definition für die Anwendung von Big-Data-Technologien einzugrenzen. Mithilfe der Design-Thinking-Vorgehensweise der Stanford School wurden die Anwendung für die Big-Data-Technologie eingegrenzt und dazu Problemstellungen klar definiert. Das Konzept des Design Thinking ist ein iteratives Modell, bei dem jeder einzelne Prozess wiederholt und optimiert wird.

Wir gehen zunächst gezielt auf die Produktion ein, um einen Überblick über das Geschehen und die Prozesse zu erhalten. Dabei ist eine differenzierte Betrachtung der Prozesse in der Produktion und der unterstützenden IT-Systeme notwendig. Von Bedeutung ist das in der Abbildung 2 aufgezeigte Säulenmodell. Das in der Arbeit entwickelte Konzept des Predictive Maintenance im Karosseriebau und der Einfluss auf das Unternehmen werden anhand dessen deutlich gemacht. Diese detaillierte Betrachtung wird von Relevanz sein, um ein ganzheitliches Verständnis zu erhalten.

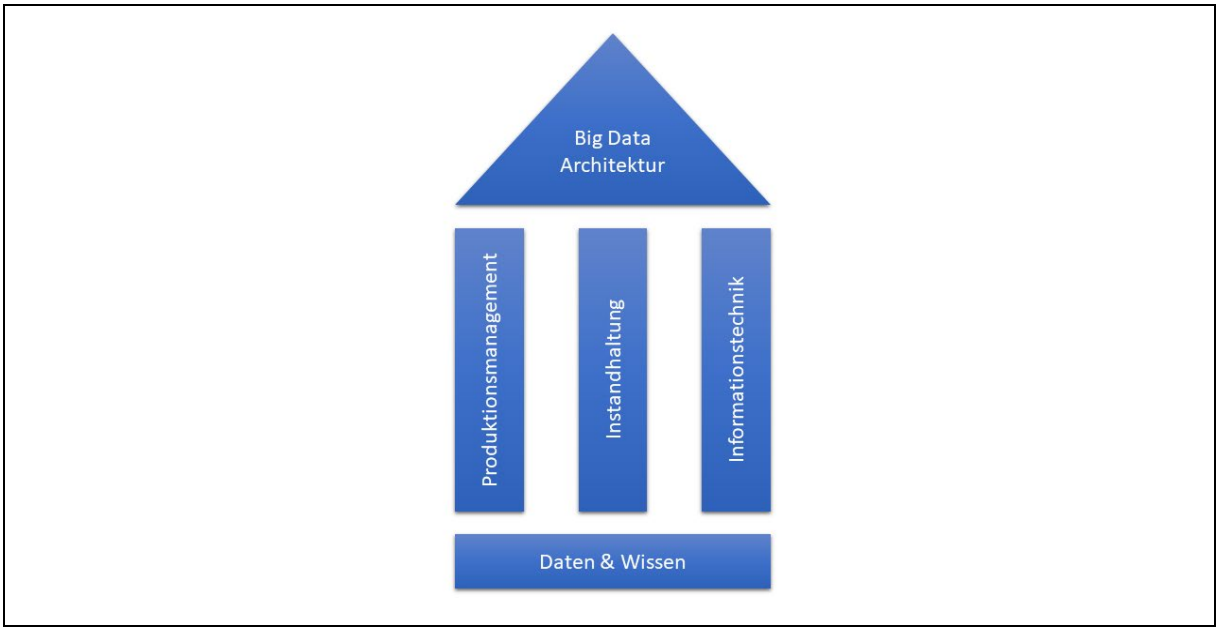


Abbildung 2: Säule der Architektur (eigene Darstellung)

2.1 Instandhaltung & Produktion

2.1.1 Produktion

Die Produktion ist ein Wertschöpfungsprozess, bei dem materielle und immaterielle Elementarfaktoren zu materiellen und immateriellen Gütern transformiert werden. Unter Elementarfaktoren verstehen wir Materialien, Dienstleistungen, Rechte und Informationen (vgl. Gutenberg, 1983, S. 1). Wertschöpfung wird dadurch erreicht, dass auf dem Beschaffungsmarkt erworbene Elementarfaktoren in höherwertige, auf dem Absatzmarkt nachgefragte Produkte und Leistungen umgewandelt werden (vgl. Schuh & Schmidt. 2014, S. 2–3). Das oberste Ziel der Produktion ist die Generierung von Kundennutzen (vgl. Ehrenmann, 2015, S. 51).

Nach GUTENBERG, 1983 werden die Elementarfaktoren nach ihren gemeinsamen Merkmalen kategorisiert. Wie in Abbildung 3 zu sehen ist, können diese in dispositive und operative Faktoren unterteilt werden.

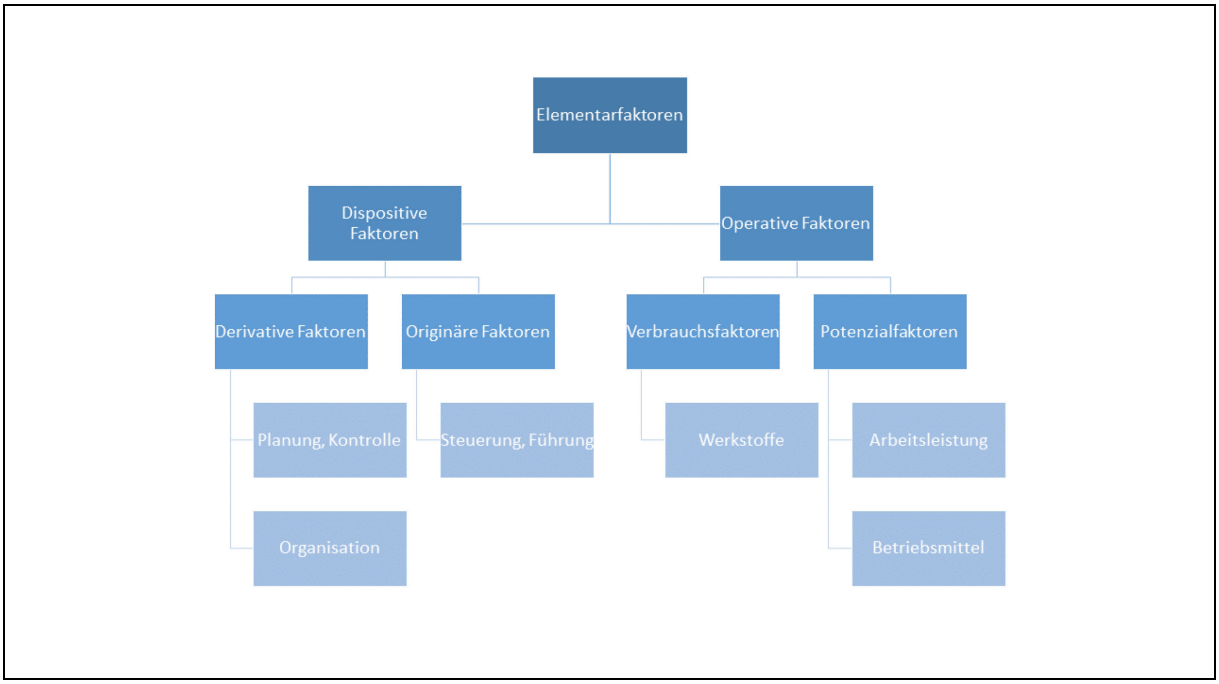


Abbildung 3: Elementarfaktoren (eigene Darstellung in Anlehnung an Kiener, 2017, S. 6)

Die dispositiven Elementarfaktoren umfassen den Teil der menschlichen Arbeitsleistung. Dieser hat zur Aufgabe, die operativen Faktoren mithilfe der Planung, der Kontrolle, der Organisation und der Steuerung optimal zu gestalten. Dispositive Faktoren differenzieren sich grundsätzlich in derivative und originäre Faktoren. Die derivativen Faktoren haben die Entscheidungsunterstützung zum Ziel und umfassen die Planung, Kontrolle und Organisation.

Die Planung dient der Vorbereitung von Entscheidungen. Mittels der Kontrolle wird das realisierte mit dem geplanten Ergebnis verglichen und dabei auftretende Abweichungen werden analysiert. Für die Umsetzung von Entscheidungen ist eine Organisation notwendig. Diese verteilt Aufgaben und überträgt Verantwortungen auf ihre Mitarbeiter. Originäre Faktoren indes führen, steuern und legen die Ziele fest. Dabei werden Entscheidungen über die einzusetzenden Mittel und die Kombination der Produktionsfaktoren getroffen (vgl. Kiener, Maier-Scheubeck, Obermeier & Weiß, 2017, S. 5).

Unter den operativen Elementarfaktoren sind Verbrauchs- und Potenzialfaktoren definiert. Verbrauchsfaktoren werden bei ihrem Einsatz im Wertschöpfungsprozess verbraucht. Sie werden zu Bestsandteilen der Produkte oder Dienstleistungen, zu denen Rohstoffe, Hilfsstoffe oder Halberzeugnisse zählen. Dabei können sie aber auch als Betriebsstoffe Anwendung finden und gelten in diesem Fall nicht als Bestandteil der Produkte. Potenzialfaktoren hingegen bringen den Input für den Erfolg des Wertschöpfungsprozesses mit ihrer Arbeitsleistung und den Betriebsmitteln. Das Leistungsvermögen, z.B. die Produktionskapazität, wird hier bestimmt. Somit haben sowohl Betriebsmittel als auch menschliche Arbeitsleistung Potenzialcharakter. Zu den Betriebsmitteln zählen Maschinen und Einrichtungen, während menschliche Arbeitsleistung ausführende, unmittelbar mit der Leistungserstellung in Zusammenhang stehende Tätigkeiten umfasst (vgl. Kiener et al., 2017, S. 5).

Die Produktion ist somit als Kombination der Elementarfaktoren zum Zweck der Leistungserstellung zu verstehen. Sie impliziert sämtliche wertschöpfende Aktivitäten in einem Unternehmen. Die Produktion besteht aus der Bereitstellung materieller und immaterieller Güter (vgl. Schuh & Schmidt. 2014, S. 3).

Produktionssystem

Im Zusammenhang mit den zuvor genannten Elementarfaktoren wird nun das Produktionssystem näher erläutert. Das Produktionssystem bildet die Grundplattform im wertschöpfenden System. Ein Produktionssystem umfasst mehrere verteilte Wertschöpfungsprozesse. Dabei können die Prozesse auch geographisch verteilt sein (vgl. Ehrenmann, 2015, S. 3–4). Vernetzt sind die Wertschöpfungsprozesse durch die Infrastruktur im Hinblick auf Material- und Informationsfluss (vgl. Dyckhoff & Spengler, 2010, S. 4). In der näheren Betrachtung besteht ein Produktionssystem aus einer operativen und einer dispositiven Instanz (vgl. Schuh & Schmidt, 2014, S. 4). Die operative Instanz beinhaltet die Abwicklung der Transformationsprozesse, die wiederum durch die Kombination menschlicher Arbeitsleistung und den Potenzialfaktoren realisiert werden. Hierbei werden aus materiellen

und immateriellen Gütern, auch Verbrauchsfaktoren, Produkte erzeugt (vgl. Dyckhoff & Spengler, 2010, S. 4). Hingegen sind die Aufgaben der dispositiven Instanz Gestaltung und Steuerung der Transformationsprozesse (vgl. Schuh & Schmidt, 2014, S. 4).

Ein Produktionssystem besteht aus Teilsystemen. Dies kann strukturell hierarchisch und funktional unterteilt werden (vgl. Dyckhoff & Spengler, 2010, S. 4). Der Grund für die Unterteilung ist der höhere Detaillierungs- und Genauigkeitsgrad, welcher modelliert wird (vgl. Ehrenmann, 2015, S. 52). Somit werden u.a. auch Wertschöpfungsnetzwerke bis hin zum elementaren unternehmensinternen Prozess als Produktionssystem aufgefasst (vgl. Nyhuis, 2008, S. 93). Als Beispiel für eine hierarchische Unterteilung ist die schrittweise Differenzierung eines Unternehmensnetzwerks in Fertigung, Montage, Lackiererei und deren Arbeitsplätze zu nennen.

KUNATH betrachtet ein Produktionssystem als funktional. Dieses besteht aus Fertigungssystem, Wertstrom, Materialflusssystem, Informationssystem, Betriebsstoffen und menschlicher Arbeitsleistung. In der Wertschöpfung bildet das Fertigungssystem den Kern. Anlagen, Maschinen und Werkzeuge bilden das Fertigungssystem ab. Die Hauptaufgabe ist, wie von Dyckhoff und Spengler beschrieben, die Umwandlung von materiellen und immateriellen Gütern in fertige Produkte. Wertstoffe und somit Elemente des Wertstroms sind Rohmaterial, Werkstücke, Produkte und Reststoffe. Die Wertstoffe werden vom Materialflusssystem gehandhabt und gelagert. Hauptaufgabe des Materialflusssystems ist die Versorgung des Fertigungsanlagesystems mit den benötigten Materialien unter Berücksichtigung von Terminplanung und Reihenfolge, Entsorgung von Abfällen sowie die Lagerung und der Transport von Produkten. Die Aufgaben des Fertigungs-, des Materialfluss- und des Informationssystems werden sowohl durch menschliche als auch durch maschinelle Arbeitsleistung ausgeführt. Dabei unterstützt das Informationssystem die Ausführung der zuvor genannten Aktivitäten. Es ermöglicht die horizontale und die vertikale Integration der Systeme. Zu seinen Aufgaben gehören Erfassung, Verarbeitung, Speicherung und Bereitstellung von physischen und digitalen Informationen, es wird also – kurz gesagt – der Informationsfluss dargestellt. Eine nähere Erläuterung dessen folgt im Abschnitt 2.2. der Prozessintegration. Damit all die genannten Systeme in Betrieb genommen werden können, sind einerseits Betriebsstoffe notwendig, bspw. Energie, Schmiermittel, Kühlmittel etc. Andererseits muss die Steuerung durch das Management ausgeführt und kontrolliert werden (vgl. Kunath, 2018, S. 227).

Produktions- und Fertigungsmanagement

In der Systemübersicht ist nun die Überführung in das Produktions- und Fertigungsmanagement elementar für die weiteren Schritte zur Einordnung. In der vorliegenden Arbeit werden nach dem Zentralverband Elektrotechnik und Elektronikindustrie e.V. die Begrifflichkeiten „Fertigungsmanagement“ und „Produktionsmanagement“ als Synonyme verwendet (vgl. ZVEI, 2017, S. 8).

Zu den Aufgaben des Produktionsmanagements gehören die zielgerichtete Gestaltung, Entwicklung, Planung, Organisation, Steuerung und Überwachung der Wertschöpfungsprozesse. Das Ziel ist, die Produkte und Dienstleistungen des Unternehmens in der erforderlichen Menge und Qualität zu generieren. Dabei sind der Zeitfaktor und die finanziellen Aspekte als Restriktion zu beachten. Das Verhältnis von Menge, Qualität zu Zeit, zu Kosten/Aufwand muss bestimmt werden (vgl. Schuh et al. 2014, S. 2). Die Aufgabe des Produktionsmanagements ist demnach sehr komplex. Daher wird Komplexität zerlegt und in ihrer Aufgabe unterteilt. Die Aufgaben können stark in ihrem sachlichen und zeitlichen Umfang variieren (vgl. Zäpfel 2001, S. 48).

Daraus folgend wird das Produktionsmanagement hierarchisch in Entscheidungsfelder unterteilt. Die Prozesse auf den unterschiedlichen Ebenen können mithilfe des Produktionscontrollings untereinander rückgekoppelt werden (vgl. Dyckhoff, 1994, S. 3352). Somit stehen sie in einer Wechselbeziehung zueinander (vgl. Zäpfel, 2000, S. 3). In der folgenden Abbildung 4 wird das Produktionsmanagement in Hierarchiestufen unterteilt und als vermaschten Regelkreis dargestellt.

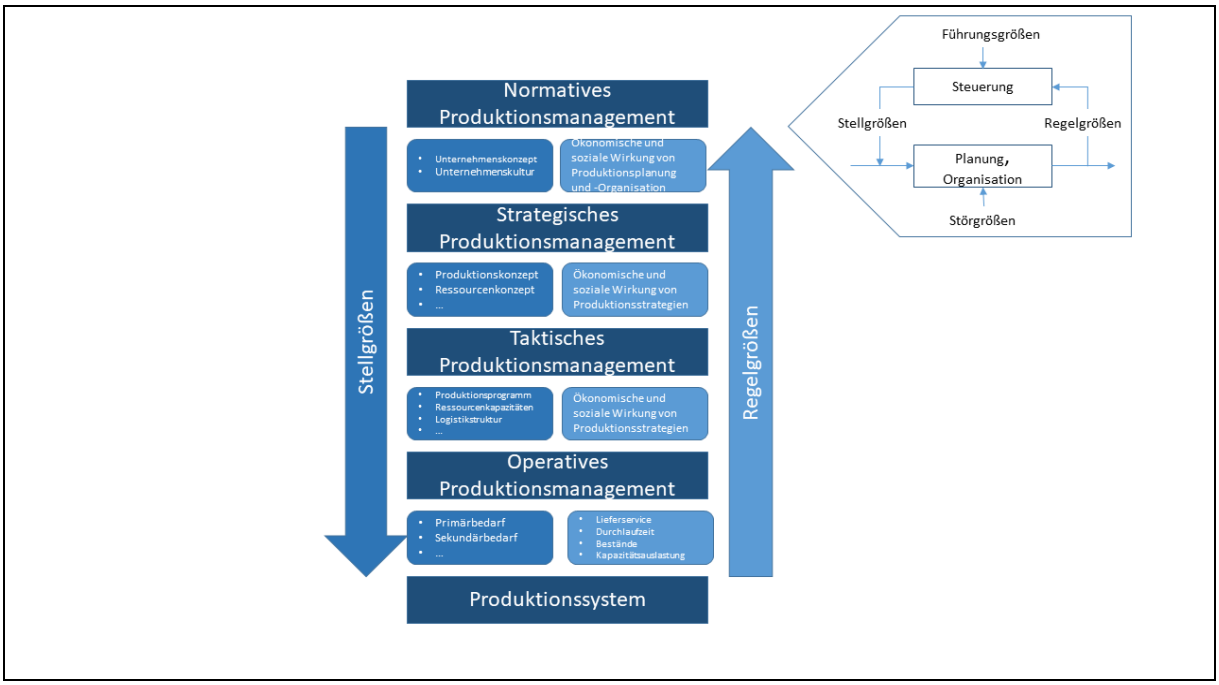


Abbildung 4: Hierarchie des Produktionsmanagements (eigene Darstellung in Anlehnung an Zäpfel, 2001, S. 49)

Die Zuordnung der Aufgabenbereiche zu den einzelnen Hierarchiestufen kann nach Trageweite der Entscheidung, Weisungsbefugnis der Entscheidungsperson, Planungszeitraum, Detaillierungsgrad der Planung oder Vollständigkeit und Sicherheit der Informationen erfolgen (vgl. Dyckhoff, 1994, S. 352).

Die Aufgaben einer höheren Entscheidungsebene sind allgemeiner Natur. Sie haben jedoch eine größere Bedeutung als jene einer untergeordneten Führungsebene. Somit ist der betrachtete Planungshorizont länger. Jedoch sind Detaillierungsgrad und die daraus resultierende Komplexität der Entscheidungsproblematik kleiner als in der Führungsebene (vgl. Zäpfel, 2000, S. 3).

Die Komplexität eines Entscheidungsproblems kann u.a. durch den Grad der Struktureinheit einer Entscheidungssituation beschrieben werden. Die Komplexität nimmt zu, je mehr sie von einer wohlstrukturierten Entscheidungssituation abweicht (vgl. Müller-Wiegand, 2019, S. 15). Merkmale einer wohlstrukturierten Entscheidungssituation sind nach ADAM und WITTE nach ihrer Art und ihrem Umfang definiert. Hierbei sind folgende Faktoren ausschlaggebend und müssen bekannt sein: die zu verfolgenden Ziele und ihre Präferenzrelationen sowie alle entscheidungsrelevanten Alternativen, mögliche Umweltzustände und die Ergebnisse der Alternativen zu jedem Umweltzustand. Darüber hinaus existiert eine eindeutige, anwendbare, systematische, effiziente und bekannte Methode oder auch eine Zielfunktion zur eindeutigen Bewertung der Handlungsalternativen. Die Zielfunktion kann in angemessener Zeit die

Handlungsalternativen eindeutig in eine Rangfolge ordnen. Der Bewertungszusammenhang der Zielfunktion ist vollständig und exakt definiert. Die Methode besteht aus einer endlichen Menge an Entscheidungsregeln, welche bei Anwendung auf die Entscheidungssituation zu einer eindeutigen Abfolge von Lösungsschritten führen. Nach Ablauf einer bestimmten Zeitspanne wird eine optimale Lösung identifiziert und der Lösungsvorgang ist beendet (vgl. Adam, 1993, S. 9 f.; Witte, 1979, S. 73–76).

Die Hierarchieebenen haben aufgrund des unterschiedlichen Detaillierungsgrades und dessen Komplexität verschiedene Informationsbedürfnisse. Anfallende Aufgaben, die den physischen Produktionsprozess unmittelbar steuern, benötigen aktuelle und präzise Informationen über den betrachteten Produktionsbereich. Aufgaben, welche die Wettbewerbsposition des Unternehmens sichern, sind von selektierten und verdichteten Informationen über das gesamte Unternehmen hinweg über die zu erwartenden Entwicklungen der Umwelt abhängig. Dazu zählt bspw. die Nachfrageentwicklung in den unterschiedlichen Märkten (vgl. Zäpfel, 2000, S. 3).

Die Unterteilung der Hierarchiestufe im Produktionsmanagement erfolgt nach der Bedeutung bzw. Tragweite der zu treffenden Entscheidungen (vgl. Kiener et al., 2017, S. 7). In der Literatur existieren unterschiedliche Einteilungen der Hierarchiestufen des Produktionsmanagements. Tabelle 1 gibt einen Überblick über die möglichen Einteilungsvarianten. Potenzielle Hierarchiestufen sind das normative, strategische, taktische und operative Produktionsmanagement.

Autor	Datum	Ebenen	Bezeichnung
Hutzschenreuter	2015	2	strategisch, operativ
Kellner et al.	2020	3	strategisch, taktisch, operativ
Hachtel & Holzbaur	2010	3	strategisch, taktisch, operativ
Kiener et al.	2017	4	strategisch, taktisch, operativ, Produktionssteuerung
Zäpfel	2000	3 + 1	(normativ), strategisch, taktisch, operativ
Schuh et al.	2014	3	normativ, strategisch, operativ
Dyckhoff	1994		

Tabelle 1: Literaturübersicht Hierarchiestufen des Produktionsmanagements (eigene Darstellung)

Die oben genannten Varianten unterscheiden sich in der Zuordnung des normativen Managements zum Produktionsmanagement und dessen Unterteilung in zwei (strategisches und operatives Management) bzw. in drei Ebenen mit strategischem, taktischem und operativem Management.

Eine eindeutige Zuordnung von Aufgabenbereichen des normativen Managements zu den exakt trennbaren Funktionsbereichen der Produktion, des Personals, des Finanzwesens oder des Marketings ist kaum darstellbar. Die Vorgaben des normativen Managements sind nicht an einzelne Funktionen des Unternehmens ausgerichtet. Sie dienen der Organisation als Gesamtes und bestimmen das Handeln aller Mitglieder (vgl. Dyckhoff, 1994, S. 354). Der Vollständigkeit halber wurde hier auch der Aspekt der normativen Betrachtung hinzugezogen. Jedoch wird in dieser Arbeit eher auf die Wechselwirkungen zwischen der operativen, der taktischen und der strategischen Ebene eingegangen. Die Unterteilung unterscheidet sich nämlich nur im Detaillierungsgrad der Aufgabenzuweisung. Damit ergibt sich ein Bild von drei Ebenen. In den folgenden Abschnitten werden diese Ebenen erläutert, woraus sodann die Einordnung der Informationstechnik im Produktionsmanagement resultiert.

Normatives Management

Inhalt des normativen Managements ist die Gestaltung und (Weiter-)Entwicklung von Prinzipien, Normen und Verhaltensregeln. Diese sind darauf ausgerichtet, die langfristige Lebens- und Entwicklungsfähigkeit einer Organisation als Ganzes sicherzustellen. Dabei wird die Lebensfähigkeit einer Organisation durch Gewährleistung ihrer Identität gewahrt. Die Umsetzung erfolgt durch Strukturen, welche die dysfunktionale Verhaltensweise minimieren sollen. Somit steht die Legitimität des Verhaltens im normativen Management im Vordergrund (vgl. Bleicher, 1991, S. 4). Bestandteile des normativen Managements sind die Unternehmensverfassung, -politik und -kultur (vgl. Schuh & Schmidt., 2014, S. 5 f.)

Der Unternehmensverfassung liegt ein Verhaltensrahmen zugrunde, mit dessen Hilfe sie dem Unternehmen Ordnung und Orientierung vermittelt. Die Unternehmensverfassung beinhaltet Verhaltenserwartungen bei Umgang mit der Unternehmensumwelt. Daraus werden im Unternehmen inhaltliche und formelle Handlungsräume und Formen der Zusammenarbeit bereitgestellt (vgl. Bleicher, 1991, S. 15).

Die Unternehmenspolitik beschreibt Ziele und Grundsätze für die gesamte Organisation. Sie beschäftigt sich mit Zwecken, Zielen und Strategien für das Unternehmen. Die Art der Führung eines Unternehmens oder einer Organisation wird dabei von der Unternehmenspolitik jedoch

nicht vorgeschrieben (vgl. Brauchlin & Pichler, 2000, S. 399). Zu ihren Aufgaben zählt vielmehr die Harmonisierung der Interessen von unterschiedlichen Anspruchsgruppen und die Bestimmung der zu verfolgenden Unternehmensziele. Unternehmensvision, -leitbild und -konzept sind Bestandteile einer Unternehmenspolitik (vgl. IGC, 2010, S. 238). BRAUCHLIN definiert sie als Gesamtheit aller Grundsatzentscheidungen, die den Unternehmenszweck festhält, die langfristigen Ziele festlegt und allgemeine Grundsätze aufstellt (vgl. Brauchlin & Pichler, 2000, S. 399).

Zukünftige Ziele des Unternehmens sind in der Unternehmensvision beschrieben und verankert (vgl. Collins & Porras, 1996, S. 66). Die Unternehmensvision legt den Existenzgrund eines Unternehmens fest. Bei dieser Überlegung werden die wirtschaftlichen Aspekte zunächst außen vor gelassen. Sie legt das Geschäftsfeld fest, in dem das Unternehmen tätig ist, und entwickelt es zudem weiter. Im Gegensatz zu klar definierten Unternehmenszielen ist der Vision kein festgelegter Zeitraum zugeordnet (vgl. Collins & Porras, 1996, S. 66). Zweck der Unternehmensvision ist die Inspiration und Motivation der Mitarbeiter. Sie dient als Leitbild und als Orientierung (vgl. Berson et al., 2001, S. 54 f.). Die Unternehmensvision ist demnach auch Grundlage für die Entwicklung und Weiterentwicklung eines Unternehmensleitbildes (vgl. IGC, 2010, S. 238).

Das Unternehmensleitbild wird schriftlich festgehalten und erklärt das Selbstverständnis und die Verhaltensgrundsätze eines Unternehmens aus Sicht der Entscheidungsträger (vgl. Eicher et al., 2018, S. 40). Das Leitbild bildet die Prinzipien für die Verwirklichung der Vision ab (vgl. Dirksen, 2020, S. 31). Ein Unternehmensleitbild gibt den Nutzen gegenüber Mitgliedern und Mitarbeitern, den sogenannten Anspruchsgruppen einer Organisation, eines Unternehmens wieder. Somit setzt sich das Unternehmensleitbild aus einer bewussten Willensäußerung des Managements bzw. der Unternehmensidentität sowie aus unbewussten Werten und Normen zusammen. Diese Faktoren beeinflussen das Unternehmensgeschehen der daraus resultierenden Unternehmenskultur (vgl. Durstberger & Most, 1997, S. 37).

Die direkte Umsetzung einer Unternehmensvision und eines Unternehmensleitbildes bedarf aufgrund der Abstraktion beider genannten eines Unternehmenskonzepts. Durch ein Unternehmenskonzept werden Vision und Leitbild konkretisiert (vgl. IGC, 2010, S. 238). Als Synonym werden die Begriffe „Geschäftsmodell“ oder „Businessplan“ verwendet (vgl. Kieser & Walgenbach, 2003, S. 19; Nagl, 2015, S. 9). Ein Geschäftsmodell ist die ganzheitliche Beschreibung der Geschäftstätigkeit eines Unternehmens (vgl. Umbeck, 2009, S. 48). Ein Unternehmenskonzept wird in leistungswirtschaftliche, finanzwirtschaftliche und soziale

Konzepte differenziert (vgl. Thommen, Achleitner, Gilbert, Hachmeister & Kaiser, 2007, S. 177). Das leistungswirtschaftliche Konzept legt die zu erbringende Leistung, die einzusetzenden Betriebsmittel und die zu verwendenden Verfahren nach Art und Umfang fest. Die zu erbringende Leistung wird anhand von Markt- und Produktzielen spezifiziert. Das leistungswirtschaftliche Konzept mündet in der Strategie für Forschung und Entwicklung, Beschaffung, Produktion und Absatz (vgl. Kieser & Walgenbach, 2003, S. 18). Das finanzwirtschaftliche Konzept definiert die finanziellen Ziele, die einzusetzenden finanziellen Mittel und deren Strategien (vgl. Thommen et al., 2007, S. 177). Finanzielle Ziele legen hier die Erwartungen über die Liquidität, Rentabilität, Gewinn und Shareholder-Value fest. Daraus resultieren Kapitalbedarf und die Kapitalstruktur eines Unternehmens. Das soziale Konzept definiert hingegen die humanitären und moralischen Ziele des Unternehmens. Diese sind als mitarbeiterbezogene und gesellschaftsbezogene Ziele zu betrachten (vgl. Thommen et al., 2007, S. 178). Das soziale Konzept erfasst demnach die menschlichen und gesellschaftlichen Wertvorstellungen der Unternehmensleitung (vgl. Kieser & Walgenbach, 2003, S. 18).

Die Unternehmenskultur ist ein dynamisches System, in dem die Mitglieder gemeinsam handeln. Sie agieren und reagieren in ihrem Umfeld (vgl. Wien & Franzke, 2014, S. 5). Dabei teilen sie Muster des Denkens, des Fühlens, des Handelns sowie die vermittelten Normen, Werte und Symbole innerhalb eines Unternehmens. Normen, Werte und Symbole prägen somit die Entscheidungen, die Handlungen und das Verhalten aller Mitglieder (vgl. Berner, 2012, S. 16 ff.).

Strategisches Produktionsmanagement

Das Ziel des strategischen Produktionsmanagements ist der Aufbau und die Erhaltung von Erfolgspotenzialen (vgl. Bleicher, 1991, S. 5). Dabei bezieht sich der Aufbau auf die innerbetrieblichen Fähigkeiten, die es dem Unternehmen erlauben, langfristig wettbewerbsfähig zu bleiben (vgl. Zäpfel, 2001, S. 53). Als Informationen gehen in die Entscheidungsprozesse des strategischen Produktionsmanagements Erkenntnisse aus Umwelt- und Unternehmensanalysen ein. Die Umweltanalysen sollen Erkenntnisse über Chancen und Risiken im Wettbewerb gewinnen. Die Unternehmensanalysen hingegen sollen Anpassungs- und Handlungsnotwendigkeiten in Abhängigkeit von den unternehmensinternen Ressourcen ausarbeiten (vgl. Zäpfel, 2000, S. 5).

Zum Aufgabenbereich des strategischen Produktionsmanagements gehören die Ziel- und Strategiefindung für das Produktionssystem (vgl. Bleicher, 1991, S. 6). Die Ziele und Strategien sind hier immer in Abhängigkeit von der autorisierten Wertvorstellung des normativen

Managements und der Bestimmung eines Produkt-Markt-Konzepts zu sehen (vgl. Dyckhoff, 1994, S. 354). Das Produkt-Markt-Konzept legt fest, mit welchen Produktfeldern das Unternehmen auf welchen Märkten agiert und die Absatzkanäle definiert (vgl. Zäpfel, 2000, S. 5). Das Produktionsprogramm ist der Leistungskatalog des Unternehmens. Darin werden Produkte und Dienstleistungen festgelegt, welche das Unternehmen den Verbrauchern auf den Absatzmärkten anbietet. Aus Sicht des Produktionsprogramms unterscheiden wir außerdem zwischen einem strategischen, mittelfristigen und kurzfristigen Produktionsprogramm. Das strategische Produktionsprogramm beinhaltet Produktfelder, auf denen das Unternehmen tätig sein wird, während das mittelfristige Produktionsprogramm die Vorgabe der produzierenden Menge der Produkte oder Produktgruppen aufstellt. Im kurzfristigen Produktionsprogramm ist hingegen die genau zu produzierende Menge aller Produktvarianten und ihrer Halbfabrikate enthalten (vgl. Dyckhoff, 1994, S. 354 f.).

Erfolgspotenziale werden in der Produktionsstrategie festgehalten (vgl. Zäpfel, 2001, S. 53). In der Literatur existieren zwar unterschiedliche Definitionen für den Begriff der Produktionsstrategie, dennoch ist deren Kernaussage identisch. HEYES und WHEEL WRIGHT grenzen demnach die Produktionsstrategie als eine Abfolge von Entscheidungen ab, die es einer Geschäftseinheit im Laufe der Zeit ermöglicht, eine gewünschte Fertigungsstruktur und Fertigungsinfrastruktur zu erreichen (vgl. Wheel Wright, 1984, S. 77 ff.). COX und BLACKSTONE beschreiben die Produktionsstrategie als ein kollektives Muster von Entscheidungen, welches sich auf die Formulierung und den Einsatz von Produktionsmitteln auswirkt. Dabei muss die Fertigungsstrategie die strategische Gesamtausrichtung des Unternehmens unterstützen. Dadurch werden Wettbewerbsvorteile erlangt (vgl. Blecker & Kaluza, 2003, S. 3). Nach ZÄPFEL gibt die Produktionsstrategie Aufschluss darüber, „welche Fähigkeiten und Potenziale im Bereich der Produktion zu entwickeln bzw. zu bewahren sind, um so einen Beitrag zur Wettbewerbsfähigkeit eines Unternehmens zu leisten“ (vgl. Zäpfel, 2000, S. 115). Alle Definitionen geben der Produktionsstrategie also die Aufgabe, die verfolgte Wettbewerbsstrategie zu unterstützen (vgl. Blecker & Kaluza, 2003, S. 3).

Zusammengefasst ist die Kernaussage, dass die Produktionsstrategie die Ausrichtung eines Produktionssystems beschreibt. Sie dient der Umsetzung der Unternehmensziele und festigt die Wettbewerbsfähigkeit oder verbessert diese. Für die Umsetzung der Ziele sind Produktionskonzepte und deren Produktionsinstrumente festzulegen (vgl. Akc & Ilas, 2005, S. 6). Zu den zentralen Entscheidungsfeldern der Produktionsstrategie gehören Entscheidungen über:

- die Anzahl und geographische Verteilung von Produktionsstätten bzw. Produktionskapazitäten (vgl. Kellner, Lienland & Lukesch, 2020, S. 165),
- die Fertigungstiefe des Unternehmens (vgl. Zahn, 1988, S. 534),
- die Modifikation und Entwicklung von Fertigungssystemen in Bezug auf Art und Umfang von Kapazität, Fertigungstypen und Fertigungstechnologien (vgl. Zäpfel, 2000, S. 5),
- die Organisation des Fertigungssystems durch die Gestaltung und Entwicklung von Informations- und Kommunikationssystemen, Planungs- und Kontrollsystemen, organisatorischen Abläufen und personellen Fähigkeiten (vgl. Zahn, 1988, S. 533),
- die Lieferantenbeziehungen (vgl. Zahn, 1988, S. 534).

Taktisches Produktionsmanagement

In der Entscheidungshierarchie folgt nun das taktische Produktionsmanagement. Das Ziel des taktischen Produktionsmanagements ist die Realisierung der laufenden Erfolgspotenziale in der Produktion. Die Ziele werden von dem strategischen Management festgelegt und werden in der taktischen Ebene konkretisiert. Damit ist die Präzisierung des strategischen Produktionsprogramms und des Produktionssystems gemeint (vgl. Dyckhoff, 1994, S. 354).

Im mittelfristigen Produktionsprogramm werden Produkte und Produktgruppen nach ihrer Art und Qualität festgelegt. Dabei wird von einem Planungszeitraum von einem Jahr ausgegangen (vgl. Zäpfel, 2000, S. 2).

Nachdem im strategischen Produktionsmanagement die geographische Verteilung und Produktionskapazitäten von Standorten festgelegt wurden, wird im taktischen Produktionsmanagement über die konkrete Ausgestaltung der Fertigungs- und Materialflusssysteme in und zwischen Standorten entschieden (vgl. Kellner et al., 2020, S. 165).

Operatives Produktionsmanagement

Die Produktion kann in groben Zügen als Regelkreis modelliert werden. Er besteht aus zwei miteinander verknüpften Prozessen. In der Abbildung 5 sind die Prozesse als Produktionsplanung und -steuerung sowie als Fertigung dargestellt. Der Fertigungsprozess beinhaltet die Kombination und Umwandlung der Einsatzfaktoren menschliche Arbeitsleistung, Betriebsmittel und Werkstoffe. Die Fertigung wird im Produktionssystem ausgeführt, die Ergebnisse sind Leistungen und die in Beziehung stehenden Rückmeldungen in Form von Informationen. Das übergeordnete operative Produktionsmanagement beinhaltet dabei

informationsbearbeitende und informationsverarbeitende Prozesse. Die eben genannten Prozesse dienen als zielgerichtete Gestaltung und Lenkung des Produktionssystems. Als Führungsgrößen fließen Produktionsziele und Planvorgabe in das operative Produktionsmanagement ein. Die Rückmeldung aus dem Fertigungsprozess geht als Regelgröße nach außen in das Produktionsmanagement. Die Rückmeldungsinformationen werden mit den Sollwerten verglichen. Abweichungen initiieren einen Zyklus im Regelkreis. Die Ergebnisse des Produktionsmanagements sind wiederum Stellgrößen in Form von Produktionszielen und Planvorgaben (Zäpfel, 2000, S. 1 f.).

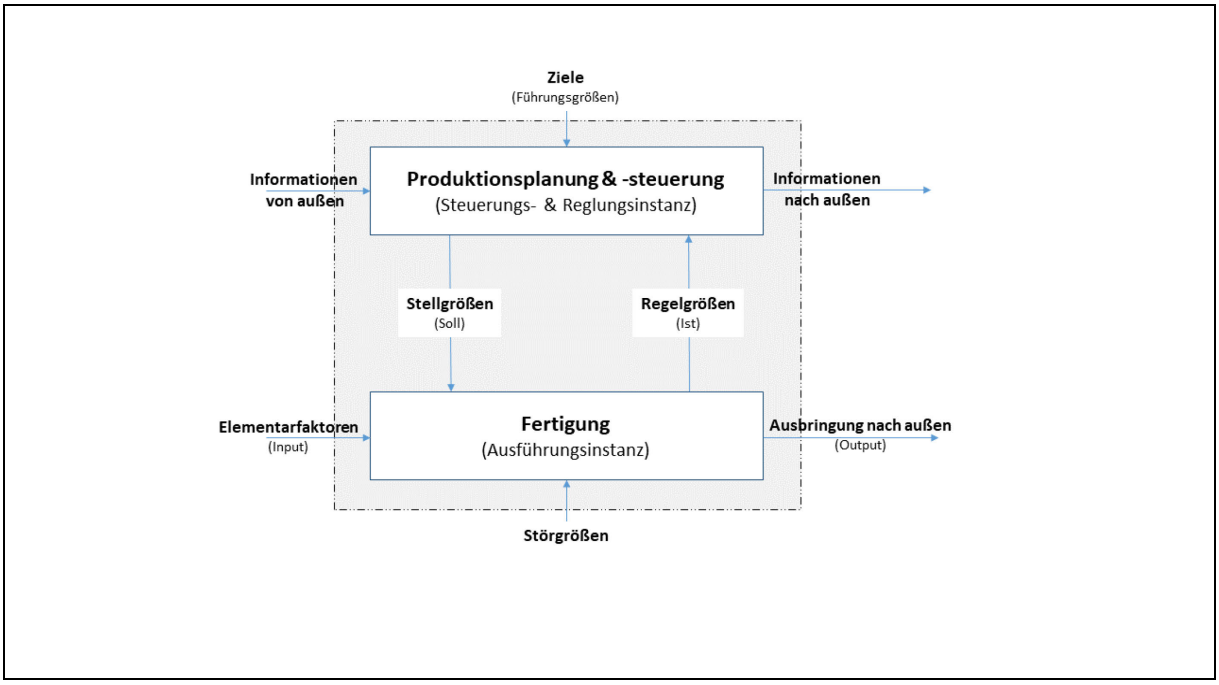


Abbildung 5: Modellierung der Produktion im Regelkreis (eigene Darstellung in Anlehnung an Dyckhoff, 1994, S. 352)

Aufgaben des operativen Produktionsmanagements

Im Wesentlichen betreffen die Aufgaben des operativen Produktionsmanagements die Lenkung des Produktionssystems. Sie beruhen auf den vom taktischen Produktionsmanagement entwickelten Zielen und Strategien. In die Entscheidungsprozesse gehen Informationen wie mittelfristige Produktionsprogramme, Wettbewerbs-, Aktivitäts- und Ressourcenstrategie ein. Der Regelungsprozess beinhaltet die zielgerichtete Planung, Steuerung und Organisation der Produktion. Damit verbunden sind die Aufgaben des operativen Produktionsmanagements in Bezug auf Planung, Steuerung und die damit einhergehende Verbesserung der leistungs-, finanz- und informationswirtschaftlichen Prozesse eines Unternehmens. Hierfür werden betriebswirtschaftliche Produktionspläne aufgestellt und deren Ausführungen überwacht.

Produktionspläne legen die Erzeugung von Gütern in einer bestimmten Menge und Qualität sowie den Produktionszeitpunkt fest (vgl. Kellner et al., 2020, S. 161). Darüber hinaus werden Leistungs- und Kooperationsverhalten der Mitarbeiter auf Basis der Wertvorstellungen des normativen Managements gefördert und beeinflusst, sodass die Ziele des Unternehmens erreicht werden können (vgl. Schuh & Schmidt, 2014, S. 6 f.).

Ziele des operativen Managements

Die Ziele des operativen Produktionsmanagements können unterteilt werden in Markt- und Betriebsziele (vgl. Schuh & Schmidt, 2014, S. 20).

Unternehmen differenzieren sich am Markt nicht ausschließlich über die Merkmale ihrer Produkte. Neben Qualität und Preis spielen auch weitere Faktoren wie Liefertreue, Lieferzeit und Lieferfähigkeit eine erhebliche Rolle. Zu den Betriebszielen sind eine möglichst effiziente Nutzung der Betriebsmittel und die daraus resultierende Kapazitätsauslastung bei gleichzeitig niedrigen Bestands- und Kapitalbindungskosten zu zählen. Diese Ziele stehen zueinander jedoch in Konkurrenz. Unternehmen müssen sich folglich in diesem Zielkonflikt positionieren (vgl. Schuh & Schmidt, 2014, S. 20 f.).

In diesem Zusammenhang ist die Aufgabe des operativen Managements die Wahl eines individuell geeigneten Betriebspunktes für das Produktionssystem. Dabei müssen die Umlaufbestände berücksichtigt werden. Daraufhin folgt die Bestimmung des Betriebspunktes innerhalb der Zieldimensionen und der konkurrierenden Ziele:

- möglichst hohe Auslastung der Produktionskapazitäten,
- möglichst hohe Erfüllung der Termintreue,
- Minimierung der Prozess- bzw. Gesamtkosten und
- Minimierung der Durchlaufzeit.

Für die Ermittlung des Betriebspunktes entwickelte sich aus der Praxis und der Wissenschaft ein Diagramm, welches die einzelnen Betriebskennlinien der Auslastung der Zieldimensionen zusammenführt. In der nächsten Abbildung sind die Zusammenhänge der Wirkung der gegenläufigen Zielgrößen dargestellt (siehe Abbildung 6).

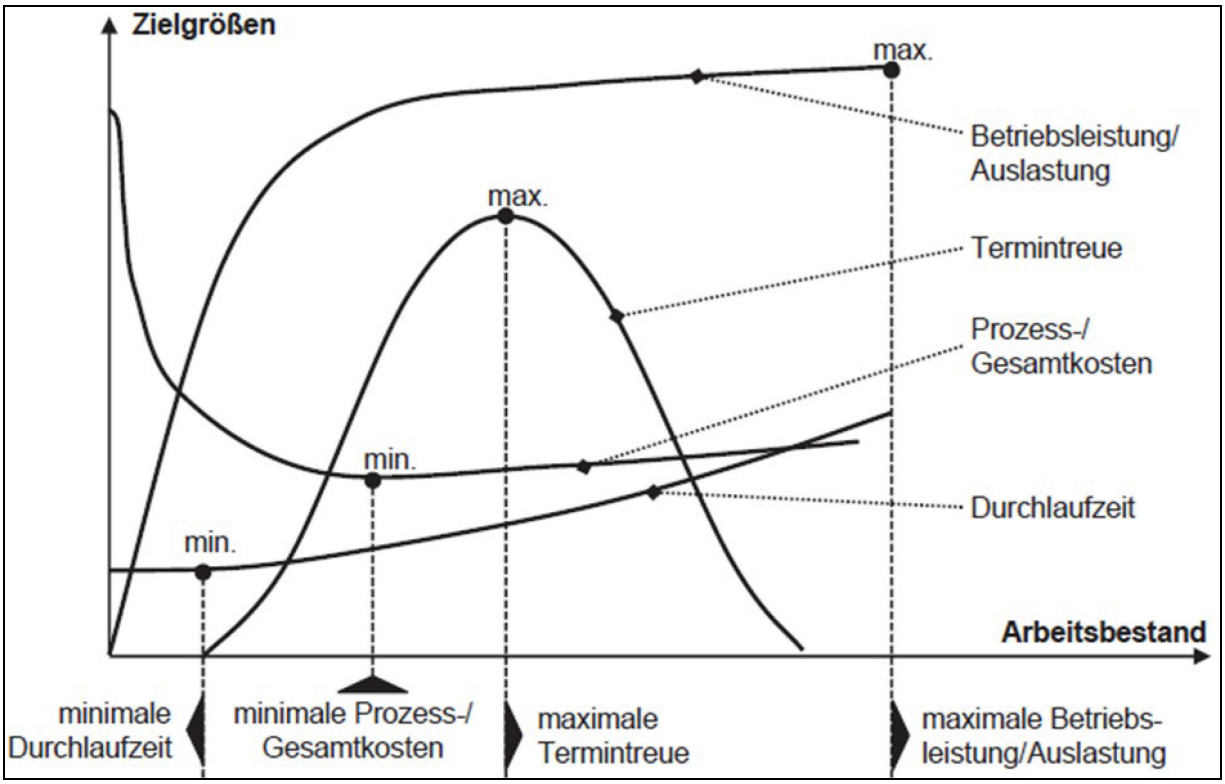


Abbildung 6: Bestimmung des Betriebspunktes anhand Betriebskennlinien (Schuh & Schmidt, 2014, S. 22)

Die Soll-Betriebspunkte sind so zu wählen, dass die Ziele des strategischen und des taktischen Produktionsmanagements erfüllt werden können. Als Ergebnis dieser Positionierung können Ziel- bzw. Führungsgrößen für Termintreue, Durchlaufzeit, Material- und Umlaufbestände sowie die Kapazitätsauslastung der Betriebsmittel festgelegt werden. Das operative Management plant und steuert mithilfe dieser Führungsgrößen das gesamte Produktionssystem in der gegebenen Konfiguration der Planungs- und Produktionsprozesse. An diesem Modell wird deutlich, dass ein Optimum aller Führungsgrößen nicht möglich erscheint, weshalb der Fokus auf die Anpassung einer Führungsgröße beschränkt wird. Damit werden bei Nichterreichung der zuvor genannten Konfiguration Prozesse von dem vorgelagerten strategischen und taktischen Produktionsmanagement angepasst. Dies hat zur Folge, dass die Verläufe der Betriebskennlinien sich ändern. Damit sind wiederum neue Kombinationen von Betriebspunkten möglich (vgl. Nyhuis, 2008, S. 189).

Die Produktionsplanung und -steuerung wird zur Abgrenzung und Erläuterung von Modellen im nächsten Abschnitt näher betrachtet.

Deterministische Produktionsplanung und -steuerung

Die Aufgabe der PPS besteht darin, den mengenmäßigen und zeitmäßigen Produktionsablauf zukünftiger Perioden unter Beachtung verfügbarer Ressourcen zu planen und die gegenwärtige und in naher Zukunft liegende Produktion zu steuern und zu überwachen. Dabei soll das vom vorgelagerten Produktionsmanagement vorgegebene Ziel erreicht werden (vgl. Corsten, Gössinger & Spengler, 2018, S. 1272). Der Prozess der Produktionsplanung und -steuerung wird in der Abbildung 7 veranschaulicht.

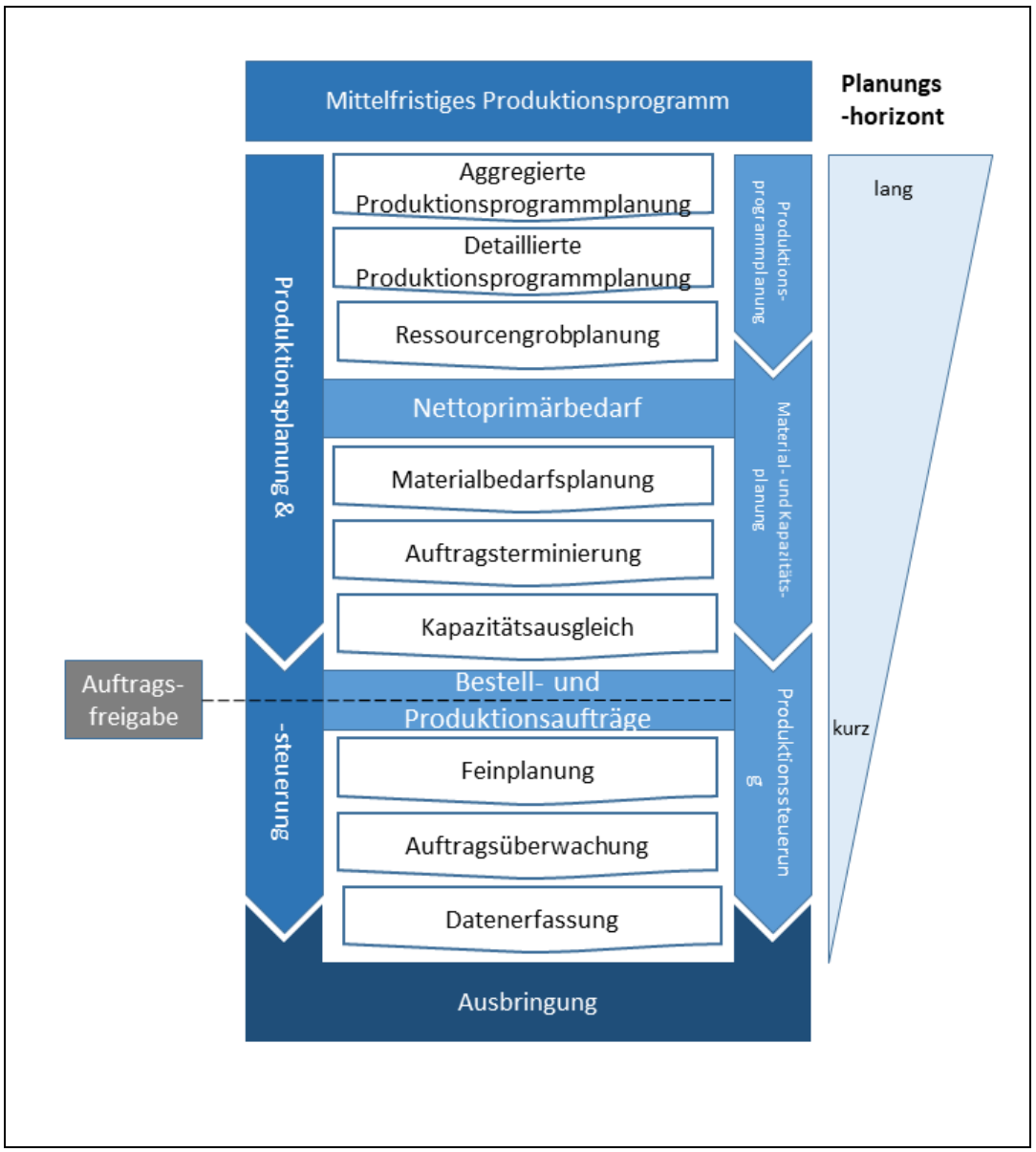


Abbildung 7: Ablauf von PPS (eigene Darstellung in Anlehnung an Kiener, 2017, S. 126)

Das Ziel der PPS ist die effiziente Koordination des Produktionsprozesses (vgl. Jahn, 2016, S. 9). Effizienz ist das Verhältnis zwischen Mitteleinsatz (Input) und Mittelnutzen (Output). Zum Mitteleinsatz gehören Elementarfaktoren. Der Mittelnutzen wird durch Produkte,

Dienstleistungen und Emissionen beschrieben (vgl. Kern, 1996, S. 863). Zur Erreichung von Effizienz werden das Maximum- und Minimumprinzip angewandt. Eine Produktion ist effizient, wenn für den gegebenen Input ein maximaler Output erzeugt oder wenn vice versa mit minimalem Input eine vorgegebene Ausbringungsmenge erreicht wird.

Für ein Unternehmen gibt es unterschiedliche Möglichkeiten, mit den beiden Prinzipien zu agieren. Für die Effizienz der Produktionsplanung und -steuerung werden Kennzahlen wie Durchlaufzeit, Termintreue, Kapazitätsauslastung und Bestände verwendet. Das betriebswirtschaftlich optimale Verhältnis dieser Kennzahlen kann theoretisch in Echtzeit auf Grundlage von Rückmeldedaten mithilfe von MES ermittelt werden (vgl. Jahn, 2016, S. 9–13).

Aufgrund der Komplexität der PPS wird die Gesamtplanungsaufgabe in mehrere Teilaufgaben differenziert. Dabei handelt es sich um unterschiedliche Zieldefinitionen und Aktivitäten, die zeitlich und organisatorisch sukzessiv aufeinanderfolgend ausgeführt werden (vgl. Jahn, 2016, S. 18). Durch die Zerlegung werden hierarchisch miteinander vernetzte Teilaufgabenbereiche gebildet. Jedoch wird bei Zusammenführung der einzelnen Teilergebnisse keine optimale Lösung der Aufgabe der PPS erbracht (vgl. Jahn, 2020, S. 18). Die Komplexität ist aus der Anzahl der zu berücksichtigenden Planungsobjekte und der in Beziehung stehenden Entscheidungen begründet (vgl. Kellner et al., 2020, S. 161).

In der Literatur existieren unterschiedliche Ansätze zur Systematisierung der Aufgaben der PPS. In ihrer Grundstruktur ist aber kein ausschlaggebender Unterschied auszumachen. Die Differenzierung in der Arbeit beruht daher auf einem der im Folgenden genannten Modelle:

- Aachener PPS-Modell nach Luczak und Jahn
- Modellierung nach Kiener, Wiendahl und Kellner.

Die Kernaufgaben der PPS werden somit in zwei Bereiche unterteilt. Bei dem Prozess der Produktionsplanung handelt es sich im Sinne der Reglungstechnik um einen Steuerungsprozess, da keine Rückkopplung existiert. In der Produktionssteuerung hingegen ist eine Rückkopplung elementarer Bestandteil des Prozesses. Aus diesem Grund ist die Bezeichnung „Produktionsreglung“ zutreffender (vgl. Jahn, 2016, S. 20 f.).

Die Produktionsplanung ist ein rationaler Entscheidungsprozess. Entscheidungen werden auf Grundlage von relevanten Alternativen, Präferenzen und Zielen getroffen. Für die Entscheidungsfindung müssen im Vorfeld alle relevanten Handlungsmöglichkeiten systematisch identifiziert und erfasst werden. Zudem gehen ökonomische, personale und soziale Ziele den Entscheidungen voraus (vgl. Zäpfel, 2000, S. 1). Die Produktionsplanung

kann dabei in die Aufgabengebiete Produktionsprogrammplanung sowie Material- und Kapazitätsplanung gegliedert werden (vgl. Kellner et al., 2020, S. 166).

Produktionsprogrammplanung

In der Produktionsplanung werden die zu produzierenden Leistungen nach Art, Menge und Termin für einen definierten Planungszeitraum festgelegt. Für die Festlegung ist das taktische Produktionsmanagement zuständig (vgl. Luczak & Ewersheim, 1999, S. 31). Als Ergebnis ist der Produktionsplan zu nennen (vgl. Jahn, 2016, S. 21).

In produzierenden Unternehmen existieren unterschiedliche Bedarfsarten. Nach der Erzeugnisstruktur wird in Primär-, Sekundär- und Tertiärbedarf unterschieden (vgl. Hartmann, 2002, S. 276 ff.).

Der Primärbedarf ist der Bedarf an verkaufsfähigen Produkten. Dieser umfasst Endprodukte und Halbfertigerzeugnisse. Halbfertigerzeugnisse können Ersatzteile in Form von Baugruppen sein. Der Sekundärbedarf deckt den Bedarf an Rohstoffen, Einzelteilen und Baugruppen ab, die zur Fertigung des Primärbedarfs notwendig sind. Somit ist der Tertiärbedarf hierarchisch ein Bedarf an Hilfs- und Betriebsstoffen, den der Sekundärbedarf benötigt.

Bei der Betrachtung der Bedarfsart wird zwischen Brutto- und Nettobedarf unterschieden (vgl. Hartmann, 2000, S. 276 ff.).

Der Bruttobedarf gibt den absoluten periodenbezogenen Bedarf wieder. Der Nettobedarf berücksichtigt zusätzlich bei der Ermittlung bestehende Bestände wie Lager und Umlaufbestände. Errechnet wird die Differenz aus Bruttobedarf und Beständen in Abhängigkeit von der Lagerstrategie.

Eine Übersicht der Bedarfsarten wird in der Abbildung 8 veranschaulicht.

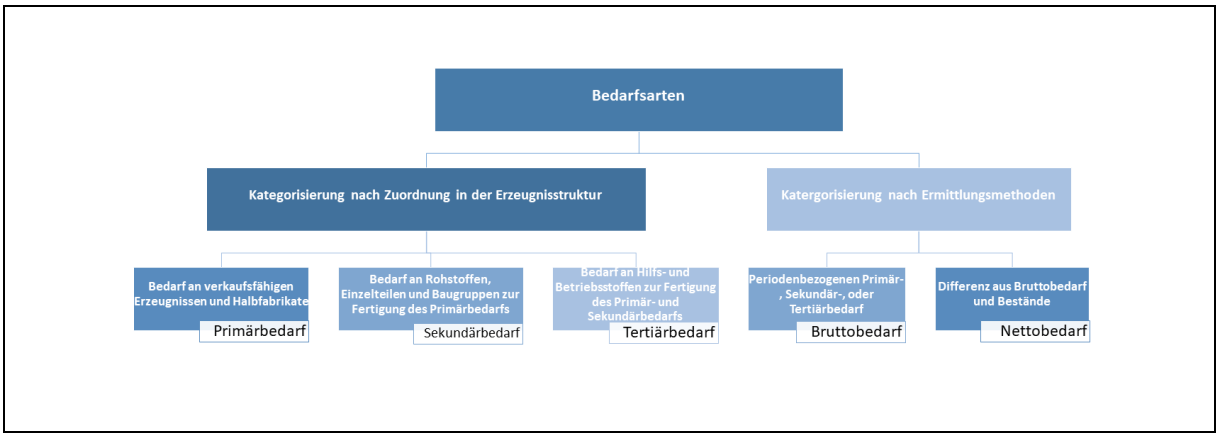


Abbildung 8: Bedarfsarten (eigene Darstellung in Anlehnung an Hartmann, 2002, S. 278)

Die Produktionsplanung erfolgt zweistufig in der aggregierten und detaillierten Produktionsprogrammplanung sowie in der anschließenden Ressourcengrobplanung. Die aggregierte Planung umfasst ähnliche Planungstätigkeiten wie die detaillierte Planung mit dem Unterschied in Bezug auf den Detaillierungsgrad (vgl. Kiener et al., 2017, S. 125). Die Teilung des Planungsprozesses dient dazu, den Planungsprozess sukzessiv mit abnehmendem Zeitabstand zum geplanten Produktionsbeginn zu verfeinern. Gründe dafür sind einerseits die Unsicherheit über Informationen und andererseits der unangemessene Aufwand für die Informationsbeschaffung. Eine Planung der Produktionsaufträge auf Wochen oder Tage schon Monate im Voraus ist nicht sinnvoll, solange Unsicherheiten über Nachfrage oder Produktionskapazitäten existieren (vgl. Kellner, 2020, S. 167).

In der aggregierten Produktionsplanung wird die Schnittstelle zum taktischen Produktionsmanagement gebildet (vgl. Kellner, 2020, S. 167). Dabei wird im taktischen Produktionsmanagement das mittelfristige Produktionsprogramm für die zu produzierende Menge der Produktgruppen für den Zeitraum von einem Jahr festgelegt (vgl. Voigt, 2008, S. 532). Die aggregierte Produktionsplanung ermittelt, welche Mengen von welchen Produktgruppen in welchen Planungsperioden produziert werden sollen. Der Planungshorizont beträgt zwischen sechs und achtzehn Monaten. Daraus resultiert ein aggregierter Produktionsplan (vgl. Kellner, 2020, S. 200) Dieser ist verbindlich und schränkt die Freiräume der detaillierten Produktionsprogrammplanung ein (vgl. Kiener, 2017, S. 125).

Die detaillierten Produktionsprogrammplanungen werden auch Primärbedarfsplanungen genannt. Sie konkretisieren den groben Produktionsplan. Ein Primärbedarfsplan ermittelt auf Wochen bzw. teilweise auf Tagesbasis den Nettoprimärbedarf. Dieser stellt den detaillierten Produktionsplan dar. In der detaillierten Produktionsprogrammplanung werden die Mengen des aggregierten Produktionsprogramms auf einzelne Endproduktvarianten heruntergebrochen. Dabei wird die Kalenderwoche oder der Kalendertag für die Fertigstellung festgelegt. Der Planungshorizont der Primärbedarfsplanung liegt zwischen drei und zwölf Monaten (vgl. Kellner, 2020, S. 200). Damit der Nettobedarf ermittelt werden kann, wird zunächst der Bruttoprimärbedarf kalkuliert. Der Bruttoprimärbedarf ist der Bedarf an Endprodukten, der für eine bestimmte Betrachtungsperiode bereitgestellt werden soll. Der Nettoprimärbedarf wird durch die Differenz des Bruttoprimärbedarfs und der Lagerbestände bestimmt. Dieser ermittelte Wert stellt die zu produzierende Bedarfe der Endprodukte dar (vgl. Jahn, 2016, S. 21).

Bei der Ressourcengrobplanung wird vom Auftrag unabhängig überprüft, ob der Produktionsplan mit den vorhandenen Ressourcen realisierbar ist. Die nach Art, Menge und

Termin festgelegten Bedarfe der Enderzeugnisse werden grob eingeplant und mit den verfügbaren Ressourcen abgeglichen. Wird in der Ressourcengrobplanung festgestellt, dass der Primärbedarf nicht gedeckt werden kann, ist eine Ressourcenabstimmung notwendig. Ein Abgleich lässt sich durch eine zeitliche Verschiebung oder durch eine Anpassung der Ressourcen, bspw. durch Sonderschichten, vornehmen (vgl. Luczak & Eversheim, 1999, S. 36 f.). Auf die Anpassung von Ressourcen werden wir im Laufe dieser Arbeit noch näher eingehen.

Materialbedarfs- und Kapazitätsplanung

Die Materialbedarfs- und Kapazitätsplanung hat die Umsetzung des Produktionsplans zur Aufgabe. Sie plant die notwendigen Ressourcen und stellt mit geeigneten Beschaffungsprogrammen den Produktionsprozess sicher. Im Sinne der Materialbedarfs- und Kapazitätsplanung sind Ressourcen alle Mittel, die in den Produktionsprozess einfließen, sie umfassen Betriebsmittel, Material, Personal und Transportmittel (vgl. Luczak & Eversheim, 1999, S. 37 f.). Ziel der Materialbedarfs- und Kapazitätsplanung ist die Bestimmung von Materialien und Ressourcen für die Umsetzung des Produktionsauftrages (vgl. Kellner et al., 2020, S. 168). Die Materialbedarfs- und Kapazitätsplanung wird demnach in folgende Kategorien unterteilt: Materialbedarfsplanung, Auftragsterminierung und Kapazitätsplanung.

Die Materialbedarfsplanung bestimmt für das geplante Produktionsprogramm den zu fertigenden oder zu beschaffenden Sekundärbedarf nach Art, Menge und Termin (vgl. Kiener, 2017, S. 125). Sie lässt sich in Bruttosekundärplanung, Nettosekundärplanung, Durchlaufterminierung sowie Losgrößen- und Bestellmengenplanung aufgliedern (vgl. Jahn, 2016, S. 21). Eine Übersicht über den Prozessablauf der Materialbedarfsplanung ist in der Abbildung 9 zu finden.

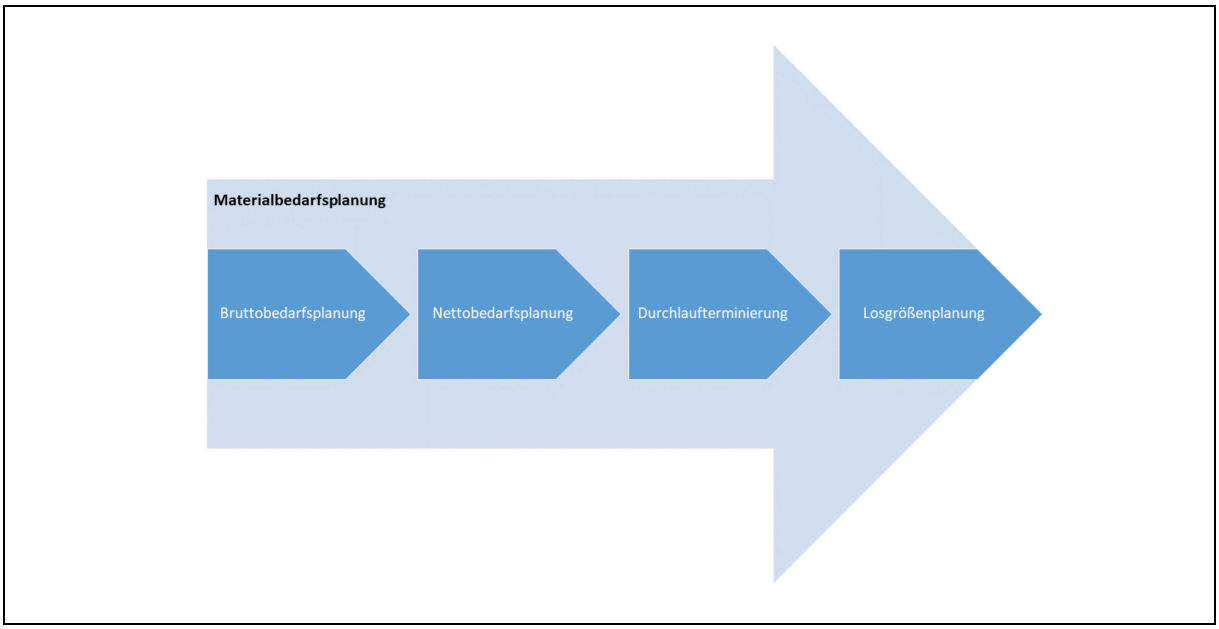


Abbildung 9: Prozessablauf der Materialbedarfsplanung (eigene Darstellung in Anlehnung an Thonemann, 2015, S. 309)

Der fertigungsstufenbezogene Bruttosekundärbedarf wird bestimmt durch die Auflösung von Stücklisten. Das Ergebnis der Bruttobedarfsplanung ist der mengenmäßige Bedarf an Materialien und Zwischenprodukten (vgl. Jahn, 2016, S. 21). Der Bruttosekundärbedarf wird unter Berücksichtigung von Lager-, Umlauf-, Sicherheits- und Meldebeständen sowie Reservierungen und Bestellungen auf den Nettosekundärbedarf reduziert (vgl. Luczak & Eversheim, 1999, S. 39). Nachdem der Nettosekundärbedarf ermittelt wurde, folgt die Bestimmung, wann der Bereitstellungsprozess beginnen muss, um die Bedarfe bereitzustellen. Die Dauer des Bereitstellungsprozesses wird als Vorlaufzeit bezeichnet (vgl. Kellner et al., 2020, S. 255). Die Durchlaufzeitterminierung ordnet den Nettobedarf der einzelnen Werkstoffe und Zwischenprodukte den einzelnen Perioden zu. In diesen Zeitabschnitten wird mit der Fertigung und Beschaffung begonnen. Nur so ist eine Fertigstellung durch das Produktionsprogramm planmäßig realisierbar (vgl. Kiener, 2017, S. 125). In der Losgrößen- und Bestellmengenplanung werden die Nettosekundärbedarfe nach wirtschaftlichen Kriterien zu Fertigungs- und Bestelllosen zusammengefasst. Die Losgröße ist die Menge, die als Auftrag den Fertigungs- oder Bestellprozess durchläuft. Ziel der Losgrößenplanung ist die Zusammenfassung der Nettobedarfsmenge in Lose, sodass die Prozessgesamtkosten minimiert werden (vgl. Kellner et al., 2020, S. 255).

Mithilfe der Durchlaufterminierung werden Beschaffungsaufträge, insbesondere Fertigungsaufträge, durch die Auftragsterminierung geplant. Die Auftragsterminierung legt Start- und Endtermine für jeden Arbeitsvorgang fest. Als Grundlage dienen die technologisch

bedingten Arbeitsabläufe (vgl. Kiener, 2017, S. 127). Die Durchlaufzeit wird ermittelt aus Belegungszeiten, z.B. Rüst- und Bearbeitungszeiten, sowie aus hinterlegten Übergangszeiten wie Wartezeiten vor und nach Bearbeitung, Kontroll- und Transportzeiten. Belegungszeiten sind in Arbeitsplänen und die Übergangszeiten in Übergangsmatrizen hinterlegt. Bei der Auftragsterminierung werden Belastungen und Kapazitäten nicht berücksichtigt. Dabei wird von unbegrenzten Kapazitäten ausgegangen. In der Übergangsmatrix sind Planwerte der Übergangszeiten für den Übergang zwischen Prozessschritten festgehalten (vgl. Luczak & Eversheim, 1999, S. 41).

Die Aufgabe der Kapazitätsplanung besteht darin, den Bedarf und das Angebot aufeinander abzustimmen. Die Kapazitätsplanung ermittelt auf Grundlage der terminierten Arbeitsvorgänge den Kapazitätsbedarf auf der entsprechenden Kapazitätseinheit und stellt diese dem Kapazitätsangebot gegenüber (vgl. Kiener, 2017, S. 127).

Die Auftragsfreigabe wird nach LUCZAK als Bindeglied zwischen der Produktionsplanung und der Produktionssteuerung angesehen. Sie erfolgt nach festgelegten Freigaberegeln oder Verfahren wie die belastungsorientierte Auftragsfreigabe oder die Kanban-Methode (vgl. Luczak & Eversheim, 1999, S. 50). Mit der Auftragsfreigabe ist eine Verfügbarkeitsprüfung verbunden. Sie soll sicherstellen, dass die für die freigegebenen Aufträge erforderlichen Ressourcen wie Personal, Maschinen, Betriebsmittel und Sekundärbedarfe bereitstehen (vgl. Mertens, 2013, S. 201 f.). Auf Grundlage der Auftragsfreigabe wird die Bereitstellung der Ressourcen veranlasst (vgl. Luczak & Eversheim, 1999, S. 50).

Produktionssteuerung

Für die Umsetzung und Kontrolle der Produktionsplanung ist die Produktionssteuerung von Nöten. Sie ist auch für die Optimierung dieser Umsetzung verantwortlich. Bei der Umsetzung von Handlungsmöglichkeiten können Abweichungen und Störungen auftreten. In diesem Zusammenhang sind die weiteren Aufgaben der Produktionssteuerung das Feststellen von bestimmten Zuständen bei der Umsetzung. Diese Zustände sind als Fortschritt des Erfüllungsgrades oder eine Soll-Ist-Abweichung bezüglich Zeit-, Kapazitäts- und Mengenvorgaben hervorzuheben. Basierend auf den Zuständen sollen Maßnahmen abgeleitet werden, welche Störungen und Abweichungen von der Produktionsplanung reduzieren und gegebenenfalls verhindern (vgl. Zäpfel, 2000, S. 21).

Die Produktionssteuerung kann unterteilt werden in Feinplanung, Auftrags- und Ressourcenüberwachung sowie Datenmanagement (vgl. Kiener, 2017, S. 127 f.).

Die Feinplanung besteht aus der Aufteilungs- und Reihenfolgeplanung. Als Beispiel ist hier die Ablauf- bzw. Maschinenbelegungsplanung zu nennen. In der Aufteilungsplanung werden freigegebene Aufträge den entsprechenden Ressourcen zugewiesen. Aufgabe der Reihenfolgeplanung ist die zielspezifische optimale zeitliche Zuordnung der Reihenfolge von Aufträgen zu den Ressourcen des Produktionssystems. Das Ziel der Reihenfolgeplanung ist die Minimierung der Terminabweichung, der Durchlaufzeit (DLZ), der Prozessgesamtkosten sowie die Maximierung der Kapazitätsauslastung. Ergebnis der Reihenfolgenplanung ist der Ablauf bzw. der Maschinenbelegungsplan (vgl. Kellner et al., 2020, S. 300 f.).

Die Auftrags- und Kapazitätsüberwachung beginnt mit der Umsetzung eines Produktionsauftrages. Die Datenerfassung unterstützt dabei mit Informationen die Überwachung. Die Auftragsüberwachung überprüft den Fortschritt von Aufträgen in Abhängigkeit ihrer Planwerte hinsichtlich Qualität, Menge und Zeit. Damit ist ein Soll-Ist-Vergleich der Daten möglich. Für die Zustandserkennung wertet die Kapazitätsüberwachung maschinen- und personalbezogene Daten aus. Für die Identifizierung von Störungen dienen somit diese eben genannten Daten als Grundlage (vgl. Kiener, 2017, S. 127).

Das Datenmanagement auf der PPS-Ebene besteht aus Datenerfassung, Datenaufbereitung und Datenanalyse. Prognose-, Soll- und Ist-Daten unterstützen die einzelnen Module der PPS mit ihrem dazugehörigen Produktionsprozessen und ermöglichen ein optimal angepasstes Datenmanagement (vgl. Kiener, 2017, S. 127). Ist-Daten liefern grundlegende Beschreibungen der Produkt- und Prozessstruktur. Sie können kategorisiert werden in Stamm- und Bewegungsdaten. Die Stammdaten können Teiledaten, Erzeugnisstrukturen, Stücklisten, Arbeitspläne, Betriebsmitteldaten etc. umfassen. Bewegungsdaten können aktuelle Lagerbestände, Kundenaufträge, Bestellungen, auftragsfortschritts-, material-, personal- oder maschinenbezogene Daten umfassen. Prognosedaten beinhalten Informationen zur Vorhersage zukünftiger Zustände. Soll-Daten sind hingegen Steuerungsgrößen. Sie werden auf Basis von Ist- oder Prognosedaten ermittelt (vgl. Kiener, 2017, S. 128).

2.1.2 Einordnung des Karosseriebaus in die Produktion

Die Wertschöpfungskette in der Automobilproduktion umfasst das Presswerk, den Karosseriebau, die Lackiererei und die Endmontage. Im engeren Sinne beginnt die eigentliche Produktion beim Zusammenfügen einzelner Teile eines Automobils mit dem Karosseriebau. In

der Abbildung 10 ist die logische Reihenfolge in der Automobilproduktion in ihrer Grobeinteilung zu sehen.

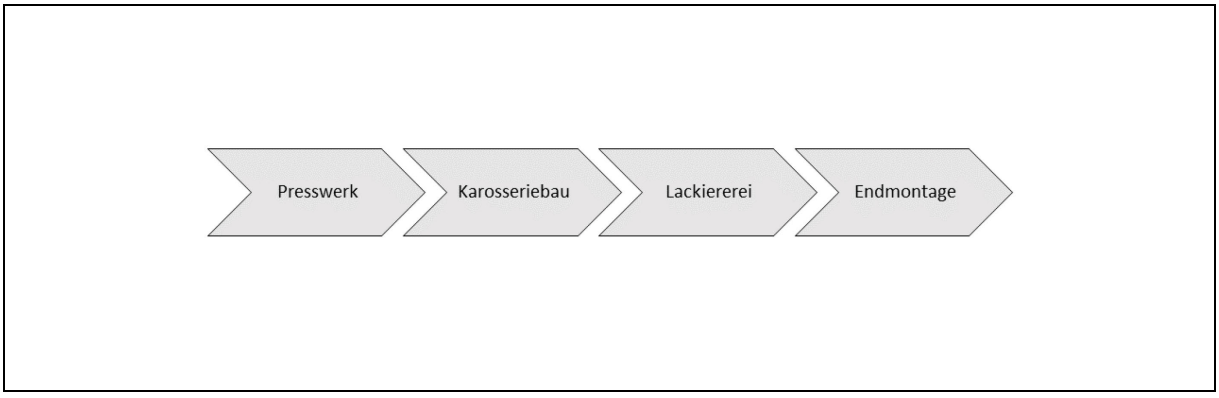


Abbildung 10: Wertschöpfungskette Automobilproduktion (eigene Darstellung)

Der in der Arbeit behandelte Abschnitt des Karosseriebaus umfasst wiederum weitere Prozesse, die im Folgenden erläutert werden. Im Karosseriebau werden Prozesse des Fügens mit den Elementen Kleben, Schweißen, Schrauben, Löten und Nieten als zentrale Prozesse betrachtet (vgl. Fritz & Schulze, 2015, S. 125). Ausgehend von der innerbetrieblichen Supply Chain gelangen die Karosserieteile vom Presswerk zum Karosseriebau. Der Karosseriebau wird in vier Bereiche unterteilt: Unterbau, Aufbau, Anbauteile und Finish. Diese Gliederung wird in der Abbildung 10 anhand eines Beispiels schematisch dargestellt. Farblich sind die Bereiche Unterbau, Aufbau, Anbauteile und Finish voneinander abgegrenzt (siehe Abbildung 11).

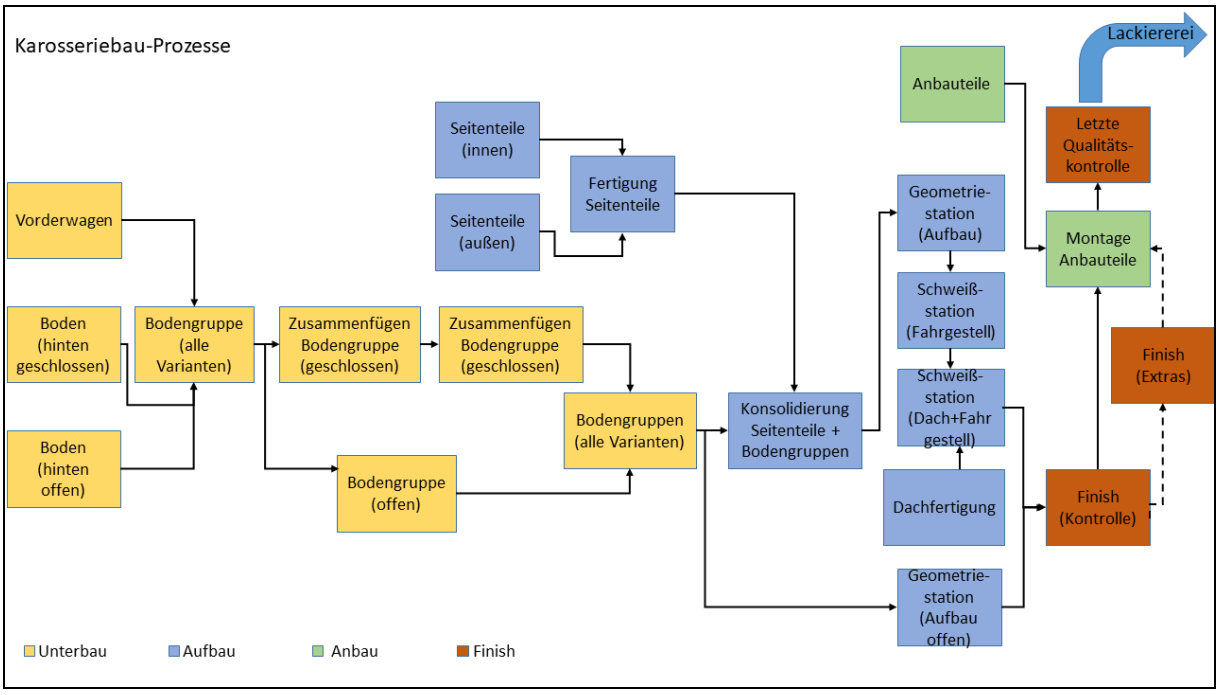


Abbildung 11: Karosseriebauprozess (eigene Darstellung)

Die Linienfertigung beginnt mit dem Unterboden und den Seitenteilen. Parallel starten hierbei die Prozesse, die zu einem späteren Zeitpunkt zusammengeführt werden. Beim Unterboden werden der vordere und der hintere Boden, die Radhäuser sowie der Vorderwagen zusammengebracht. Mit diesem Teilprozess beginnt der Produktionsprozess des Zusammenbauens eines Fahrzeuges. In diesem Schritt erhält jedes Fahrgestell einen Transponder, mit dem alle vollautomatisierten Anlagen und deren Roboter und Maschinen das Fahrzeug bearbeiten. Der Unterboden wird nun in den Bereich Aufbau weitergeleitet, wo das Karosseriegerippe aufgerichtet wird. In diesem Prozess werden das Dach und die Seitenwände montiert. Die Seitenwände werden an den Unterbau der Karosse angebracht. Gleiches gilt für den Dachriegel, der die Seitenwände verbindet. In der Geometriestation wird die Karosse exakt verschweißt. Dieser Schritt beginnt mit den inneren Seitenwänden und geht weiter zu den äußeren. Im nächsten Prozessschritt werden Dach und Karosse zusammengeführt. Nachdem das Fahrzeuggestell von außen komplett ist, werden optional noch spezielle Sonderausstattungen angebaut/verbaut. Die letzte Instanz im Aufbau besteht in der Zusammenführung von beweglichen Anbauteilen. Hier werden Türen, Motorhaube, Heckklappe oder Hecktüren, je nach Variante, montiert. Im Finish-Prozess findet die Abnahme statt. Die Karosse wird auf Qualität wie Spaltmaße, Fügeverbindung oder auch Oberflächenqualität geprüft. Danach folgt die Weitergabe zur Lackiererei.

Zwischenfazit

In diesem Kapitel wurde die Produktion mit dem dazugehörigen Produktionsmanagement näher betrachtet. Der Leser soll damit einen Überblick über die Produktionsprozesse erhalten, um im kommenden Kapitel die Instandhaltung inhaltlich einordnen zu können.

2.1.3 Instandhaltung

Im zuvor behandelten Kapital wurden das Produktionsmanagement und die Produktionsprozesse detailliert betrachtet. Zur Einhaltung der Vorgaben der Produktionsplanung wird der Instandhaltung eine wichtige Rolle zugeschrieben (vgl. Leidinger, 2017, S. 6). In diesem Kapitel wird zunächst auf die allgemeine Instandhaltung eingegangen, um dann die Grundlage für das Predictive-Maintenance-Konzept im Kapitel 5 zu bilden.

Instandhaltung wird überall dort eingesetzt, wo die Betriebsbereitschaft eines Objektes aufrechterhalten werden muss (DIN 31.051:2012-09, 4.1). Sie drückt „die Kombination aller

technischen, administrativen und Managementhandlungen während des Lebenszyklus eines Gegenstandes aus, die dazu bestimmt sind, ihn in einem Zustand zu erhalten oder wiederherzustellen, in dem er die geforderte Funktion erfüllen kann“ (DIN EN 13306:2015-09, 2.1). Als Gegenstand (oder auch Vermögenswert) sollte in diesem Zusammenhang jedes technische Bauteil, Gerät, Teilsystem, jede Funktionseinheit, jedes Gerät oder System definiert werden, das einzeln beschrieben und betrachtet werden kann (DIN 31.051:2012-09, 4.2.1).

Ein Gegenstand kann eine Vielzahl von Zuständen annehmen. Die beiden Haupttypen werden als Aufwärts- oder Abwärtszustand definiert. Ersterer bedeutet, dass der Gegenstand in der Lage ist, seine erforderliche Funktion zu erfüllen, solange die externen Ressourcen zur Verfügung stehen. Der zweite ist entweder durch einen Fehler oder durch eine mögliche Unfähigkeit gekennzeichnet, eine geforderte Funktion während des Wartungsdienstes zu erfüllen. Beide Zustände können von einem deaktivierten Zustand betroffen sein, in dem ein Element aus irgendeinem Grund nicht in der Lage ist, eine geforderte Funktion zu erfüllen, je nachdem, ob es durch ein externes oder internes Ereignis initialisiert wurde. Darüber hinaus kann der Aufwärtszustand in einen Leerlauf-, Bereitschafts- oder Betriebszustand unterteilt werden. Abbildung 2 stellt diese genannten Zustände dar und spezifiziert die Anforderungen an die Zeitintervalle, in denen sich ein Element im Betriebszustand befinden muss (DIN EN 13306:2015-09, 5-6).

Im Lebenszyklus einer Anlage oder einer Maschine kommt es unweigerlich zu Zustandsänderungen durch Degradation, Abnutzung, Alterung und Korrosion. Solche Zustandsänderungen führen letztendlich zu einem Ausfall. Daher verfolgt die Instandhaltung das primäre Ziel, den Materialverschleiß zu verzögern und den Verfall dieser Gegenstände zu vermeiden (vgl. Strunz, 2012, S. 2). Aus einer breiteren Geschäftsperspektive betrachtet, dient sie dem Zweck, eine ausreichende Verfügbarkeit der Anlagen zu gewährleisten und eine hohe Rentabilität der Produktionssysteme sicherzustellen. Die Erzielung einer maximalen Betriebsverfügbarkeit bei minimalen Kosten erfordert die Einhaltung mehrerer Teilziele wie:

- Vermeidung von ungeplanten Stillständen,
- Reduzierung der gesamten Instandhaltungskosten,
- Einsparung natürlicher Ressourcen,
- Erhöhung der Lebensdauer von Maschinen und Anlagen (vgl. Schenk, 2013, S. 16).

Im Rahmen der DIN 31051 (2012-09, 3) wird die Instandhaltung in vier Hauptaufgaben unterteilt, die sich wie folgt zusammensetzen: Inspektion, Wartung, Instandsetzung und Verbesserung (siehe Abbildung 12).

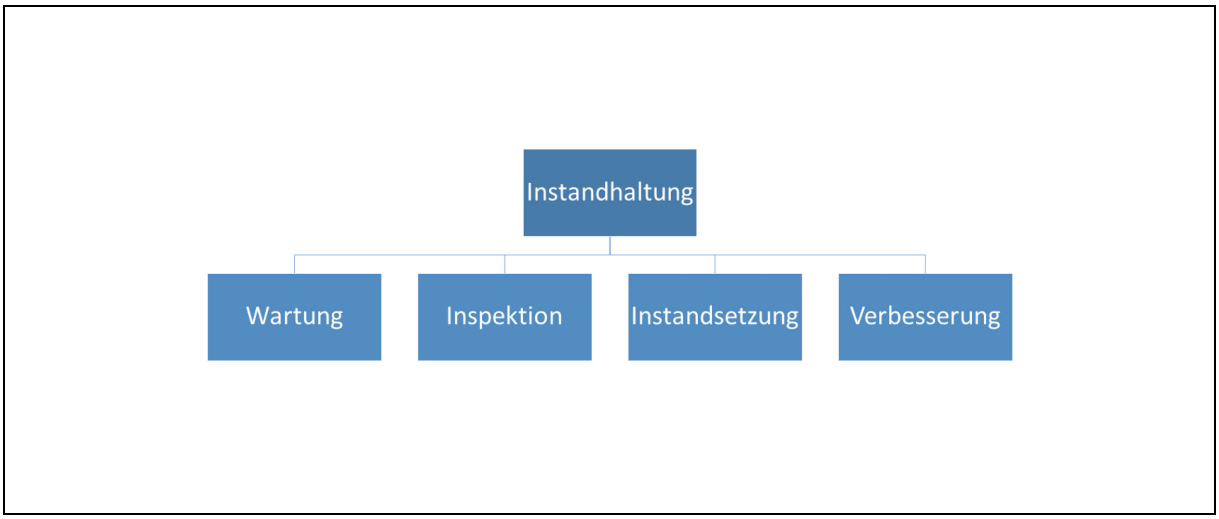


Abbildung 12: Organigramm Instandhaltung (eigene Darstellung nach DIN 31051)

Inspektionen sind Maßnahmen, um den Ist-Zustand eines Gegenstandes zu beurteilen und zu erkennen. Darunter fallen auch Untersuchungen, um die Ursache eines Fehlers oder Verschleißes herauszufinden. Daraus folgt die Ableitung notwendiger Konsequenzen für zukünftige Einsätze.

Die Wartung wird nach DIN31051 zur Bewahrung des Soll-Zustandes durchgeführt. Als Grundlage dienen die entsprechenden Richtlinien, Normen und Gesetze. Als Maßnahmen bzw. dazugehörige Prozesse können das Nachstellen, Reinigen, Auswechseln, Auslösen und Protokollieren genannt werden.

Die Instandsetzung umfasst Prozesse, die einen Gegenstand wieder in eine funktionsfähige und ursprüngliche Form bringen. Damit soll der Soll-Zustand erreicht werden. Dazu zählt bspw. der Austausch einer Maschinenkomponente mit einem gleichwertigen Ersatzteil. Abgegrenzt werden laut dieser Definition davon Maßnahmen, die zu einer Verbesserung führen würden.

Eine Verbesserung umfasst technische und administrative Maßnahmen. Sie erhöht die Funktionssicherheit eines Gegenstandes, ohne die geforderten Eigenschaften zu verändern, z.B. die Beseitigung einer Fehlerstelle durch eine konstruktive Veränderung (vgl. DIN 31051:2012-09, S. 4).

Mit den genannten Maßnahmen können die nachfolgenden Ziele nach LEIDINGER erreicht werden:

- Sicherheit der Anlage,
- Verfügbarkeit der Anlage,
- Zuverlässigkeit der Anlage,

- Werterhaltung der Anlage (vgl. Leidinger, 2017, S. 15).

Eine Instandhaltungsstrategie ist definiert als eine angewandte Managementmethode, um die Instandhaltungsziele zu erreichen (vgl. DIN EN 13306:2015-09, 2.4). Eine solche Strategie legt fest, welche Maßnahmen an welchem Gegenstand wie oft und zu welchem Zeitpunkt durchgeführt werden sollen. Die Vernachlässigung von Instandhaltungsmaßnahmen kann zu einer übermäßigen Anzahl kostspieliger Ausfälle und einer schlechten Systemleistung und damit zu einer verminderten Zuverlässigkeit führen. Wenn sie jedoch zu oft durchgeführt werden, kann sich die Zuverlässigkeit zwar verbessern, aber die Wartungskosten werden stark ansteigen. Die Restnutzungsdauer (RUL) eines Gegenstandes wird dadurch nicht vollständig ausgekostet. Abbildung 13 stellt die Beziehung zwischen der Zuverlässigkeit und der RUL eines Gegenstandes sowie die Wartungskosten dar. Wenn die Zeit bis zum Ausfall eines Systems gegen 0 geht, nimmt die Zuverlässigkeit des Systems ebenso wie die Wartungskosten ab. Sobald die Zeit bis zum Ausfall gleich 0 ist, geht das System in einen Ausfallzustand über. In der Folge steigen die Wartungskosten aufgrund der hohen Folgekosten enorm an (vgl. Peng, Dong, Zuo, 2010, S. 298). Diese beiden konkurrierenden Ziele, die Nutzung der RUL und die Gewährleistung der Zuverlässigkeit, müssen durch ein kosteneffizientes Schema ausgeglichen werden (vgl. Endrenyi et al., 2001, S. 683). Bei der Auswahl der richtigen Strategie müssen neben wirtschaftlichen Aspekten auch rechtliche, sicherheitstechnische und technische Anforderungen berücksichtigt werden. Die Wahl eines geeigneten Instandhaltungsprogramms hat einen entscheidenden Einfluss auf die Häufigkeit von Ausfällen und die vielen unerwünschten Folgen solcher Unterbrechungen.

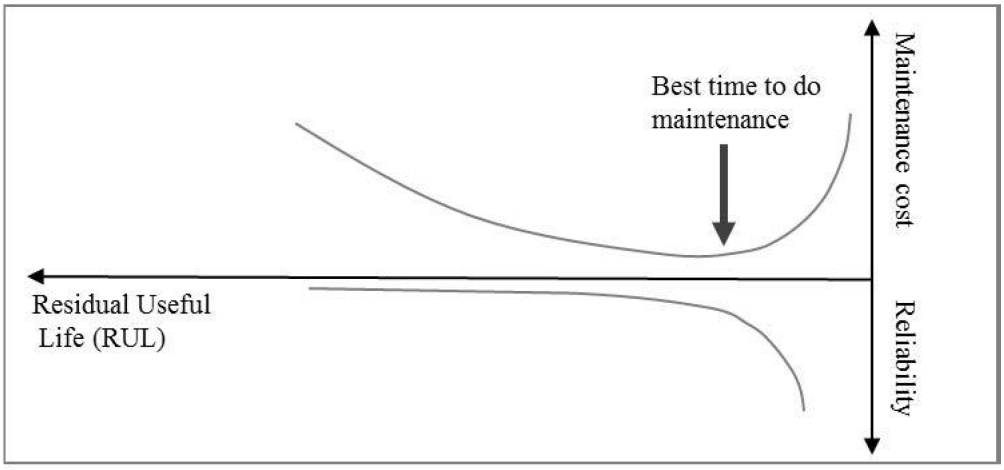


Abbildung 13: Restnutzungsdauer (eigene Darstellung)

Der Zeitpunkt, zu dem die Wartung durchgeführt wird, kann systematisch oder unsystematisch gewählt werden. Bei einer unsystematischen Vorgehensweise wird die Wartung nur dann

durchgeführt, wenn sie notwendig ist bzw. wenn ein Fehler auftritt. Bei einem systematischen Ansatz werden die Maßnahmen periodisch nach Zeit- oder Betriebsintervallen oder in Abhängigkeit vom bewerteten Zustand eines Gegenstandes durchgeführt (vgl. Schenk, 2013, S. 26). Instandhaltungsstrategien können im Allgemeinen in zwei Haupttypen unterschieden werden: korrektive und präventive Instandhaltung (siehe Abbildung 14). Letztere lässt sich zusätzlich in eine vorgegebene, zustandsabhängige und vorausschauende Strategie unterteilen (vgl. DIN EN 13306:2015-09, 7).

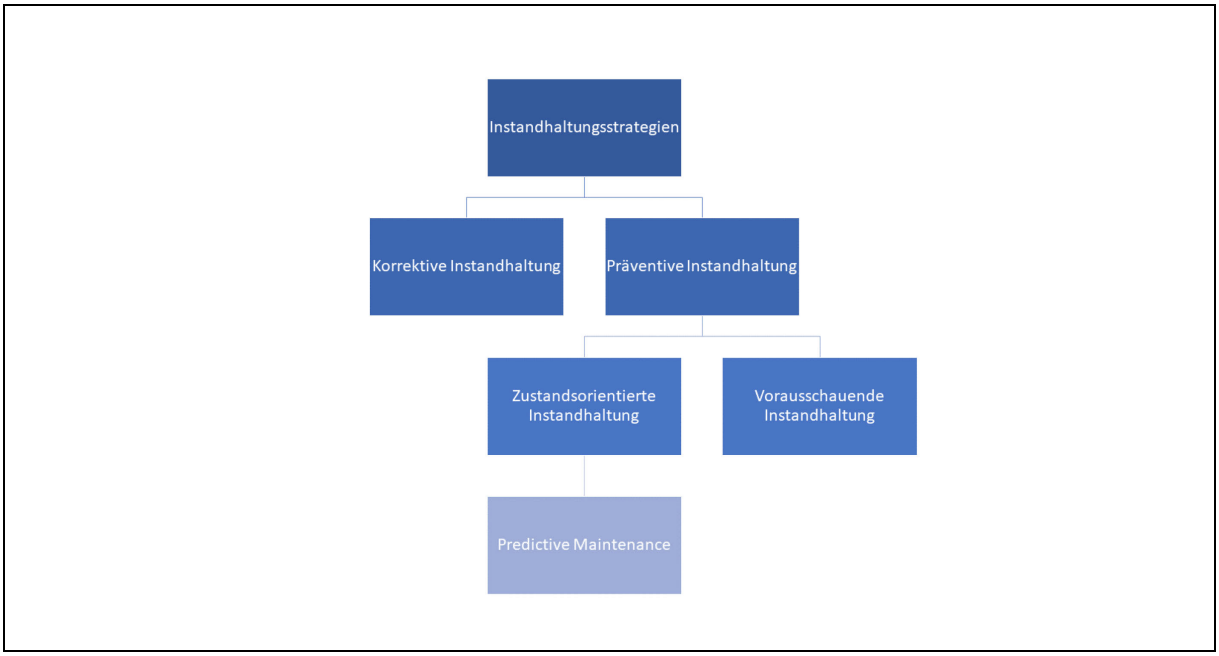


Abbildung 14: Instandhaltungsstrategien (eigene Darstellung nach DIN EN 13306)

Korrektive Instandhaltung

Bei der Anwendung der Korrekturstrategie (auch Ausfallstrategie genannt) werden Instandhaltungsmaßnahmen nur dann ergriffen, wenn ein direkter Ausfall einer Anlage/Maschine eingetreten ist oder wenn ein vordefinierter Schadensumfang erreicht wurde. Auf zwischengeschaltete Dienste oder Inspektionen wird bei diesem Ansatz verzichtet (vgl. Shin & Jun, 2015, S. 119). Im engeren Sinne kann eine solche Methode kaum als „Strategie“" bezeichnet werden, da es keine wirkliche Planung gibt, wenn sie eher auf plötzliche Zwischenfälle reagiert. Die Pannenhilfe ist eine recht primitive Instandhaltungsmethode; sie erfordert jedoch hohe Standards, um effektiv zu arbeiten. Bei einem Ausfall müssen die richtigen Maßnahmen sofort durchgeführt werden und das Instandhaltungspersonal muss über eine hohe Fachkompetenz sowie über ein ausgezeichnetes Bewusstsein verfügen, um die richtige Ausfallursache schnell einschätzen zu können. Darüber hinaus kann diese

Vorgehensweise gefährlich sein und ist daher aufgrund des hohen Risikopotenzials, das von defekten Maschinen ausgeht, in mehreren Bereichen unzulässig (vgl. Schenk, 2013, S. 14).

Auf den ersten Blick erscheint die Instandsetzung als kosteneffizient, da kein Aufwand für die Planung betrieben wird und nur wirklich beschädigte Teile ersetzt werden. Jeder Ausfall ist jedoch immer ein ungeplantes und spontanes Ereignis. Aufgrund dieser Tatsache werden die Wartungsarbeiten in der Regel unter großem Zeitdruck durchgeführt. In solchen unvorbereiteten Szenarien stehen die benötigten Ressourcen wie Personal, Ersatzteile, Werkzeuge und Geräte jedoch selten vollständig zur Verfügung (vgl. Schenk, 2013, S. 14). Das vielleicht gravierendste Risiko besteht in der unzureichenden Berücksichtigung von Interdependenzen zwischen den einzelnen technischen Komponenten. Die Störung eines Elements kann zu enormen Folgeschäden an anderen Vermögenswerten, Ausrüstungen oder Komponenten führen. Ein bekanntes Beispiel ist ein gerissener Zahnriemen eines Viertaktmotors. Der Zahnriemen selbst kann kostenlos ersetzt werden. Ein ausgefallener Zahnriemen kann jedoch zum Bruch der Kolbenstange führen, was einen Totalschaden des Motors zur Folge hat.

Aus den oben genannten Gründen ist die Instandsetzung im Vergleich zu alternativen Strategien durch hohe Ausfallzeiten und Folgekosten gekennzeichnet. Eine weitere Unannehmlichkeit ergibt sich aus Schwierigkeiten bei der Planung der Verfügbarkeit von Betriebssystemen, die durch die Unvorhersehbarkeit von Fehlern verursacht werden. Der Einsatz dieser Strategie ist nur unter folgenden Voraussetzungen für unkritische Bereiche geeignet: Die Folgen eines Ausfalls sind gering, es besteht kein Sicherheitsrisiko und der Ausfall wird schnell erkannt. Auch das Vorhandensein hochredundanter Systeme oder die schnelle Verfügbarkeit aller erforderlichen Instandhaltungsressourcen könnten diese Strategie anwendbar machen. Leider werden diese Voraussetzungen in modernen Industriegebieten selten gleichzeitig erfüllt (vgl. Fu et al., 2004, S. 179).

Vorausschauende Instandhaltung

Die vorgegebene oder planmäßige Wartung ist eine der beliebtesten Wartungsstrategien. Sie umfasst periodische Inspektionen, die für die Instandhaltung der Ausrüstung erforderlich sind. Daher wird jeder Wartungsdienst auf der Grundlage vordefinierter Intervalle unabhängig vom tatsächlichen Zustand oder den Abnutzungsmerkmalen des Geräts durchgeführt. Solche Intervalle basieren ausschließlich auf dem Alter oder der Betriebsdauer (vgl. Ben-Daya & Raouf, 2000, S. 5). Es ist ratsam, diese Strategie zu praktizieren, wenn die erwartete Lebensdauer eines Artikels bekannt ist. Auf diese Weise wird das Fehlerrisiko eines

Gegenstandes erheblich reduziert. Außerdem können Service- und Inspektionsmaßnahmen sowie die erforderlichen Instandhaltungsressourcen besser geplant werden. Solche Maßnahmen werden üblicherweise dann geplant, wenn sich das System im Ruhezustand befindet, und da bestimmte Wartungsschritte häufig praktiziert werden, kann die Dauer des daraus resultierenden Stillstands erheblich reduziert werden (vgl. Matyas, 2002, S. 14).

Bei der Anwendung einer vorher festgelegten Strategie könnte ein Nachteil darin bestehen, dass der Austausch von Gegenständen in der Regel zu früh erfolgt. Dadurch werden Verschleißreserven (RUL) verschwendet und der Verbrauch von Ersatzteilen steigt. Das bedeutet, dass dieser Ansatz zwar die Ausfallkosten reduzieren, gleichzeitig aber die Kosten durch vorbeugende Maßnahmen steigern kann. Um den optimalen Kompromiss zu finden, muss die Haltbarkeit jedes Artikels in einer Anlage angemessen eingeschätzt werden. Dies stellt eine bedeutende Herausforderung dar. Bei falsch berechneten Intervallen oder nicht berücksichtigten Änderungen der Arbeitsbelastung kann es zu einem vorzeitigen Ausfall kommen, sodass die Strategie unter den gleichen Nachteilen leiden kann wie die korrektive Instandhaltung (vgl. Schenk, 2013, S. 29). Darüber hinaus basiert die planmäßige Wartung auf der folgenden Annahme: Das Ausfallverhalten aller Geräte besitzt die Badewannenkurve. Damit eignet sich die Strategie nur für Ausfälle, die ein deutliches Verschleißverhalten aufweisen, nicht aber für zufällige Ausfälle (vgl. Fu et al., 2004, S. 179). Trotz dieser Vorbehalte lässt die vorgegebene Strategie letztlich jedoch weniger Ausfälle zu als der korrigierende Ansatz. Daher wird sie in den meisten Situationen als die überlegenere Methode angesehen (vgl. Schenk, 2013, S. 29).

Eine genaue und wiederholte Analyse der Anlagengeschichte stellt eine wichtige Grundlage für eine wirksame vorbestimmte Instandhaltung dar. Um die Auslastung der gesamten betrieblichen Lebensdauer zu optimieren, müssen periodische Instandhaltungsmaßnahmen mit individuellen Intervallen durchgeführt werden. Ein obligatorischer Schritt ist daher die detaillierte Dokumentation früherer Störungen. Zweitens müssen statistische Methoden angewendet werden, um eine aussagekräftige Prädikation abzuleiten und ein geeignetes Ersatzintervall zu bestimmen (vgl. Schenk, 2013, S. 30).

Eine planmäßige Instandhaltung ist insbesondere dann erforderlich, wenn ein potenziell suboptimaler Zustand eines Gegenstandes eine Gefahr für Mensch oder Umwelt darstellt. Andere Anwendungsfälle stehen in Zusammenhang mit gesetzlichen Regelungen, die eine regelmäßige Überprüfung oder einen regelmäßigen Austausch vorschreiben, z.B. die alle zwei Jahre stattfindende Hauptuntersuchung für Kraftfahrzeuge in Deutschland. Darüber hinaus wird

eine solche Vorgehensweise für Gegenstände empfohlen, die im Vergleich zu den Folgekosten nach einem Ausfall niedrige Wartungskosten haben, z.B. Luft- und Ölfilter. Schließlich stellt die vorgegebene Strategie eine gangbare Option dar, wenn bei Ausfällen die bisherigen Erfahrungen fehlen (vgl. Schenk, 2013, S. 30).

Zustandsorientierte Instandhaltung

Die zustandsabhängige oder auch zustandsorientierte Wartung (auch Inspektionsstrategie genannt) kombiniert die Vorteile kurzer Ausfallzeiten und einer gut ausgenutzten RUL. Der Hauptunterschied zwischen dieser Strategie und der vorgegebenen ist der Parameter, der zur Initialisierung von Instandhaltungsmaßnahmen führt. Während die spätere Strategie in der Regel die Zeit als bestimmenden Faktor verwendet, verwendet die zustandsabhängige Instandhaltung die tatsächliche Form eines Artikels. Sie baut auf der Idee auf, regelmäßig den Zustand eines Artikels und seiner Komponenten zu ersetzen, was nicht bis zur nächsten Inspektion dauern würde (vgl. Leidinger, 2017, S. 18).

Die Wirksamkeit dieser Politik beruht auf der Aufrechterhaltung von verschlechterungsrelevanten Maßnahmen wie Temperatur, Vibrationen, akustischen Emissionen eines Betriebssystems oder Geräts, um dessen Gesundheitszustand zu bestimmen. Sobald diese Maßnahmen eine vordefinierte Ausfallschwelle überschreiten, wird das System zur Reparatur abgeschaltet (vgl. Kaiser & Gebraeel, 2009, S. 840). Eine regelmäßige Inspektion und Überprüfung solcher Variablen durch Mitarbeiter stellt die einfachste Form der Überwachung dar. Solche Maßnahmen können von einfachen Sichtkontrollen bis hin zur vollständigen Demontage reichen. Eine durchdachte Option zur Erfassung verschleißrelevanter Informationen kann durch spezifische Zustandsüberwachungssysteme (CMS) wie ein SCADA-System (Supervisory Control and Data Acquisition) erfolgen. Solche Tools sind in der Lage, regelmäßig automatische Systemtests durchzuführen und den aktuellen Zustand einer Anlage kontinuierlich darzustellen.

Ein wichtiges Ziel beim Einsatz solcher Systeme ist es, mit einer minimalen Anzahl von Sensoren möglichst viele Komponenten zu überwachen, die Investitionskosten niedrig zu halten und die Installation neuer potenzieller Fehlerquellen zu verhindern (vgl. Schenk, 2013, S. 31). Neben der Beobachtung degradationsrelevanter Maßnahmen ermöglichen einige Condition-Monitoring-Systeme auch eine intelligente Diagnose, die eine Fehlererkennung (Feststellung, dass etwas nicht in Ordnung ist), eine Fehleridentifikation (Feststellung der Fehlerursache) und eine Fehlerisolierung (Feststellung des Fehlerortes) umfasst (vgl. Schwabacher & Goebel, 2007, S. 107).

Die zustandsabhängige Wartung liefert wertvolle Hinweise auf bevorstehende Fehler und erhöht die Präzision der Fehlerdiagnose. Dadurch kann sie das Auftreten von Teilefehlern im Vergleich zu den früheren Ansätzen wirksam reduzieren (vgl. Shin & Jun, 2015, S. 120). Der größte Vorteil dieser Strategie ist jedoch die intensive Nutzung der RUL einer Anlage/Maschine. Dadurch können Kosten für Ersatzkomponenten eingespart werden.

Trotz dieser Vorteile erfordert diese Strategie hohe Anfangsinvestitionen für technische Geräte und die Implementierung von IT-Systemen. Zudem ist sie auf gut ausgebildetes und qualifiziertes Personal angewiesen, das in der Lage ist, aus den bereitgestellten Informationen die richtigen Schlüsse zu ziehen. Zustandsabhängige Wartung sollte außerdem nur dann angewendet werden, wenn das Verhalten einer Anlage/Maschine messbar ist. Sie muss also technisch durchführbar und die Erfassung zustandsrelevanter Größen wirtschaftlich sein, da sonst der Aufwand an Zeit und Kosten die möglichen Fehlerkosten übersteigen könnte (vgl. Schenk, 2013, S. 31).

Predictive Maintenance

Das Konzept des Predictive Maintenance ist eine Weiterentwicklung der zustandsorientierten Strategie. Mit Predictive Maintenance ist eine automatisierte Zustandsüberwachung mit Vorhersagen (Prognosen) von Fehlern oder auch Ausfällen möglich. Sowohl die Zeitprognose als auch die Auslöser bzw. Fehlerursachen können hiermit festgestellt werden. Predictive Maintenance ist ein Ergebnis einer rechnergestützten Auswertung der Eingangsdaten (vgl. Schwabacher & Goebel, 2007, S. 1). Während die zustandsbasierte Instandhaltung nur den aktuellen Zustand einer Anlage/Maschine beschreibt, kann die Predictive-Maintenance-Strategie auch voraussichtliche Zustandsänderungen abschätzen. SHIN und JUN definieren diese Strategie als „eine Instandhaltungsstrategie, die Wartungsmaßnahmen durchführt, bevor Produktausfälle auftreten, indem sie den Maschinenzustand einschließlich der Betriebsumgebung bewertet und das Risiko von Ausfällen in Echtzeit auf der Grundlage gesammelter Produktdaten vorhersagt“ (vgl. Shin & Jun, 2015, S. 120). Die Vorhersage einer Verschlechterung einer Anlage/Maschine basiert auf der Annahme, dass viele Anomalien nicht sofort auftreten, sodass in der Regel eine stetige Entwicklung vom Normalzustand hin zu Anomalien stattfindet (vgl. Shin & Jun, 2015, S. 119). Selbst wenn keine direkten Beweise für eine Verschlechterung der Anlage/Maschine vorliegen, kann Predictive Maintenance die Prozess- und Logistikdaten, die während des Betriebes anfallen, sammeln um daraus Muster erkennen, die zu einer Verschlechterung führen (vgl. Susto et al., 2015, S. 812). Der Prozess wird überwacht, indem der Zustand der Anlage/Maschine stetig beobachtet wird, damit

Abweichungen von der durchschnittlichen Leistung erkennbar werden. Voraussetzungen sind hierfür sowohl aktuelle als auch historische Daten über den gesamten Prozess der Anlage/Maschine.

Die Predictive-Maintenance-Strategie bringt somit viele Vorteile mit sich. Jedoch ist zu beachten, dass eine reibungslose Instandhaltung nur mit dem Know-how des für die Anlage/Maschine vorgesehenen Ingenieurs realisierbar ist. Die richtige Interpretation und mögliche Ursache eines Fehlers anhand eines Bauteils oder einer Sensorik sollte durch das Werkzeug des Predictive Maintenance bestimmt werden. Damit können Informationen über den Schweregrad und daraus die richtigen Schlussfolgerungen für den Ingenieur gezogen werden, z.B. wird anhand eines Predictive-Maintenance-Modells eine mögliche Ausfallwahrscheinlichkeit errechnet, die vorhersagt, dass innerhalb von zwei oder drei Tagen ein elementarer Fehler beim Drucksensor auftritt. Mit diesem Indiz können die Bauteile, die mit dem Drucksensor zusammenhängen, im Vorfeld möglichst genau inspiziert und ausgetauscht werden. All dies stellt einen Mehrwert für das Unternehmen dar. Die dadurch veranlasste Instandhaltung ermöglicht zeitlich genaue Service- und Reparaturmaßnahmen vor Eintreffen eines Ausfalls. Das Ziel ist die Aufrechthaltung der aktiven Betriebszeit und das Maximieren der Restnutzungsdauer. Darüber hinaus werden die Planung von Instandhaltungsmaßnahmen und die damit verbundenen Ressourcen optimiert. Des Weiteren besteht die Möglichkeit, durch geeignete Analyse Software-Probleme nach Schweregrad zu priorisieren und daraus Maßnahmen zu entwickeln und zu empfehlen (vgl. Thomson, Edwards, Britton & Rabenau 2014, S. 4). Diese Unterstützung reduziert menschliche Fehler und Fehleinschätzungen sowie die Abhängigkeit von erfahrenem Wartungspersonal. In der Studie von Daugherty et al. wurden die Instandhaltungskosten bei der Durchführung von Predictive Maintenance ermittelt. Die Ergebnisse sind einerseits eine Kostensenkung um 30 % und andererseits Reduzierung der Ausfälle um 70 % (vgl. Daugherty, Banerjee, Negm & Alter, 2015, S. 4). Eine weitere Studie des US-Energieministeriums hat die Umsetzung einer Predictive-Maintenance-Strategie in der Öl- und Gasindustrie erörtert und kam zu dem Ergebnis, dass die Instandhaltungskosten um 25 % gesenkt werden konnten. Die Ausfälle wurden um 70 % verringert und damit eine Produktivitätssteigerung um bis zu 45 % erreicht (vgl. Sullivan, Pugh, Melendez & Hunt, 2010, S. 52). Neben den zuvor genannten Vorteilen existieren jedoch auch Einschränkungen. Zu nennen sind zunächst die hohen Investitionskosten (vgl. Sullivan et al., 2010, S. 52). Für die Implementierung einer Predictive-Maintenance-Strategie müssen die technischen Voraussetzungen gegeben sein, denn sie ist nur mit

ausreichend Sensoriken und Informationstechnik umsetzbar. Neben der Hardware-Beschaffung sind Experten oder qualifizierte Mitarbeiter aus dem Informatikbereich, speziell Datenwissenschaftler, erforderlich. Auf Basis der gesammelten Daten werden Algorithmen und Methoden für die Entscheidungsstrategien entwickelt. Weiterhin gibt ein Predictive-Maintenance-Modell keine 100%ige Garantie für die Genauigkeit. Somit existieren Einschränkungen in der Diagnostik und der Prognose (vgl. Shin, Jun, 2015, S. 121). In der Tabelle 1 sind nochmals die oben genannten Eigenschaften des Predictive Maintenance und der klassischen Instandhaltung zusammengefasst. Die Tabelle 2 gibt einen kurzen Überblick über die Merkmale, Anforderungen und deren Vor- und Nachteile.

	Korrektive Instandhaltung	Zustandsorientierte Instandhaltung	Vorausschauende Instandhaltung	Predictive Maintenance
Merkmale	Ausführung zeitlich nach einem Ausfall oder Fehlermeldung	Ausführung nach Plan oder Intervall	Ausführung durch ständige Beobachtung der Anlage/Maschine	Ausführung vor Eintritt eines Ereignisses durch Prognosen
Anforderungen	• Qualifizierte Mitarbeiter, • Teileverfügbarkeit • kurze Reaktionszeit	• Grundlegendes Wissen über die Bauteile der Anlage/Maschine. • Genaue Planung von Ressourcen	• Überwachungssystem • IT-Infrastruktur • Qualifizierte Mitarbeiter	• Überwachungssystem • IT-Infrastruktur • Daten • Modellierung und Algorithmen
Vorteile	• Maximierung der Lebensdauer • Geringe Planungskosten	• Minimierung der Ausfälle • Reduzierung von Fehlerketten • Hohe Planbarkeit	• Maximierung der Betriebsdauer • Maximierung der Lebensdauer	• Maximierung der Betriebsdauer • Maximierung der Lebensdauer • Hohe Planbarkeit
Nachteile	• Hohe Folgekosten durch Fehler/Ausfälle • Ungeplante Kosten	• Vergeudung der RUL • Unvorhersehbare Fehler existieren • arbeitsaufwändig	• Hohe Investitionskosten für Technik	• Hohe Investitionskosten für Technik • Eventuell neues Personal

Tabelle 2: Gegenüberstellung der Instandhaltungsstrategien (Tran Anh et al., 2018)

2.2 Informationstechnik & Produktion (Prozessintegration)

Zur Einordnung des in der Dissertation entwickelten Big-Data-Konzepts für Predictive Maintenance wird in diesem Kapitel die Informationsverarbeitung innerhalb eines Unternehmens durch elektronische Datenübertragung erläutert. Dazu werden Instrumente aus technischer Sicht sowie aus Prozesssicht betrachtet. Darüber hinaus erfolgt ein Einblick in das Datenmanagement. Zunächst wird das Informationssystem als Ganzes betrachtet und erläutert.

Danach wird anhand der Automatisierungspyramide ein Überblick über die Unternehmensstruktur in Bezug auf Daten und Informationen gegeben. Anhand der Automatisierungsperiode werden die Integrationen und Schnittstellen analysiert und es wird auf das Zusammenwirken von ERP, MES, SCADA und SPS-System eingegangen. Anschließend wird die Maschinenebene genauer betrachtet. Abschließend werden die Instrumente und Konzepte aus dem Bereich des Datenmanagements zusammengefasst.

2.2.1 Integration der Informationstechnik

Im folgenden Abschnitt werden die Zusammenhänge von Informationssystemen zur Produktion veranschaulicht. Dabei ist zunächst die technologische Betrachtung von Bedeutung. Zunächst wird das Informationssystem als Ganzes untersucht, um danach Schritt für Schritt die einzelnen Komponenten zu durchleuchten.

Ein Informationssystem ist ein zweckbezogenes, offenes und dynamisches System (vgl. Teubner, 1999, S. 26). Die Funktion eines Informationssystems ist die bedarfsgerechte und zeitnahe Informationsversorgung (vgl. Krüger, 1994, S. 143). Aus betriebswirtschaftlicher Sicht dienen Informationen generell dem Zweck, Unternehmensaufgaben zu erfüllen. In diesem Zusammenhang werden Informationen als explizites Wissen bereitgestellt, welches der Mensch zur Erfüllung betrieblicher Aufgaben nutzt. Das Wissen im Sinne von Daten wird demnach als eine gewisse Anforderung an das Informationssystem betrachtet. Daten sind maschinell erfasste, verarbeitete und gespeicherte Informationen (vgl. Abts & Mülder, 2017, S. 11). Auf der Grundlage dieser Definition wird eine Kommunikation zwischen Mensch und Maschine generiert. Dabei wird der Austausch von Informationen ermöglicht (vgl. Teubner, 1999, S. 17–18). Aus Unternehmenssicht ist ein Informationssystem ein offenes System. Es bietet die Möglichkeit der Kommunikation der Maschinen und IT-Systemen untereinander, aber darüber hinaus auch der Mensch-Maschine-Interaktion. IS führen zum einen Aufgaben automatisiert selbständig aus und zum anderen unterstützen sie die manuellen Ausführungen von Aufgaben durch die Erfassung, Be- und Verarbeitung sowie der Bereitstellung von Daten (vgl. Abts & Mülder, 2017, S. 15). Die Entwicklung eines solchen Informationssystems wird unternehmensspezifisch durchgeführt (vgl. Seibt, 1991, S. 253). Dies geschieht durch die Anpassung der organisatorischen und personellen Rahmenbedingungen (vgl. Laudon, Laudon & Schoder, 2016, S. 14–15). Im gesamten Unternehmen ist daher ein IS ein Teilsystem, welches in die Aufbau- und Ablauforganisation integriert ist. Somit wird das Unternehmen in und mit

seiner Umwelt vernetzt und verhält sich dynamisch in Abhängigkeit von der Interaktion. In einer weiteren Betrachtungsweise ist ein Informationssystem ein sozio-technisches System, das sich in ein soziales und technisches System unterteilen lässt (vgl. Teubner, 1999, S. 26). In einem solchen sozio-technischen System wirken Mensch und Maschine als Aufgabenträger zusammen (vgl. Ferstl, 2013, S. 71).

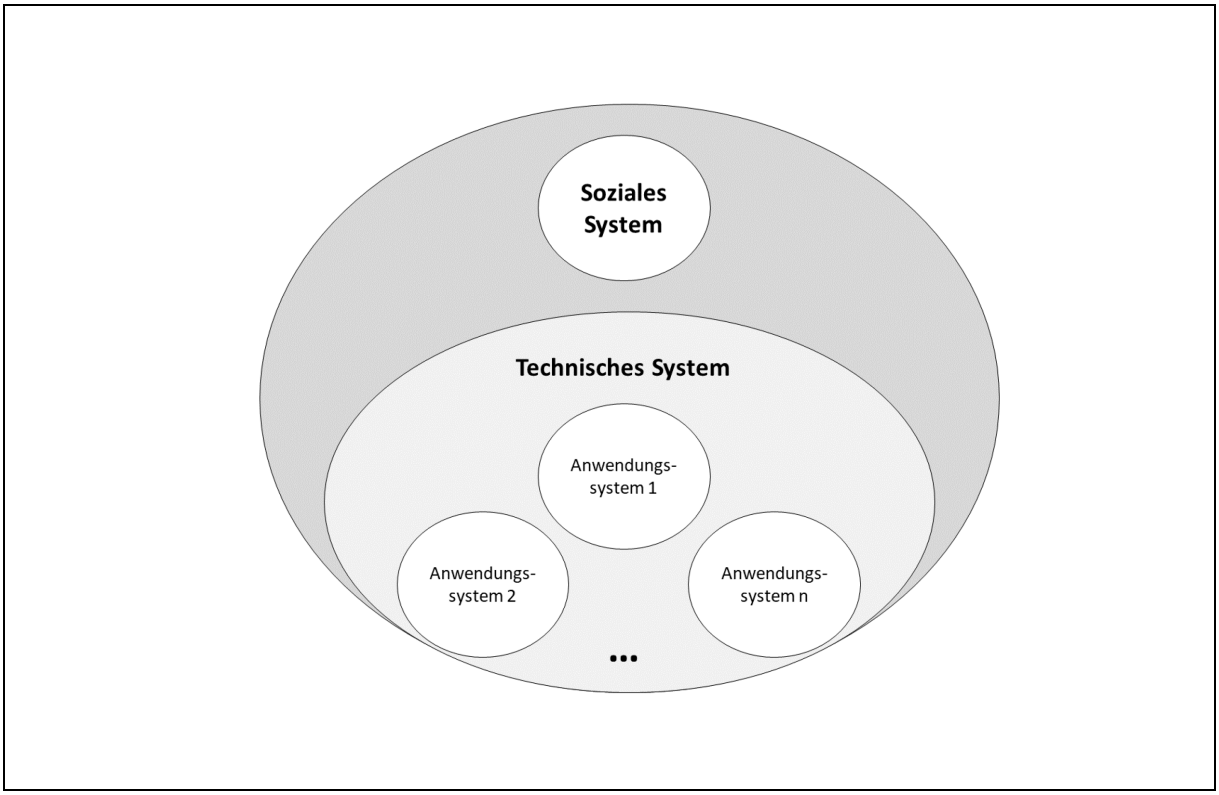


Abbildung 15: Struktur von Informationssystemen (eigene Darstellung in Anlehnung an Teubner, 1999, S. 26)

Das soziale System umfasst den Faktor Mensch, der Aufgaben und Prozesse durch Informationen ausführt (vgl. Abts & Mülder, 2017, S. 15). Des Weiteren zählt auch die Organisationsstruktur dazu, die in ein IS eingebettet ist (vgl. Laudon et. al., 2016, S. 14–15). In der Abbildung 15 ist eine vereinfachte Darstellung zur Veranschaulichung zu sehen. Zur weiteren Betrachtung besteht ein soziales System aus den Faktoren Aufgaben, Mensch und Organisationsstruktur (siehe Abbildung 16).

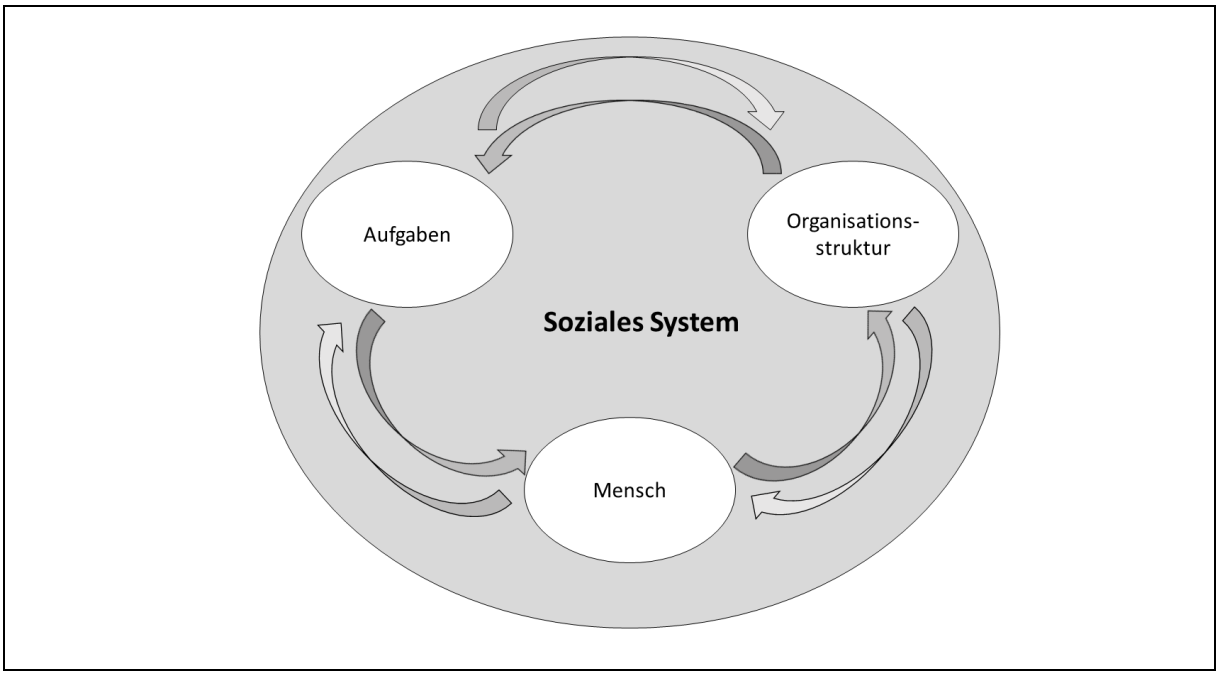


Abbildung 16: Übersicht über das soziale System (eigene Darstellung)

Das technische System hingegen umfasst alle Anwendungssysteme, die zum Betrieb eines Unternehmens oder eines bestimmten Bereiches im Unternehmen eingesetzt werden (vgl. Mertens et al., 1997, S. 38–19). Ein Anwendungssystem erfüllt eine Aufgabe oder einen Aufgabenbereich eines IS (vgl. Ferstl, 2013, S. 6). Daraus folgend ist das technische System die partielle technische Umsetzung eines IS, die zur Lösung einer betrieblichen Aufgabe oder eines Aufgabenbereiches automatisch erfasst, bearbeitet und bereitstellt (vgl. Mertens et al., 1997, S. 38–39). Unter Anwendungssysteme fallen alle Anwendungssoftwares, die zur automatisierten Erfüllung bestimmter betrieblicher Aufgaben und Prozesse eingesetzt werden (vgl. Laudon et al. 2016, S. 14). Eingeschlossen werden hier auch die Daten, die von der Software erzeugt, bearbeitet und bereitgestellt werden (vgl. Abts & Mülder, 2017, S. 15). Anwendungssysteme nutzen die IT-Infrastruktur, auf der die Software ausgeführt wird, und sind zumeist standardisiert (vgl. Laudon et al., 2016, S. 14). Sie bestehen wiederum aus den Elementarfaktoren Aufgaben, Daten, Anwendungsprogramm und IT-Infrastruktur (siehe Abbildung 17).

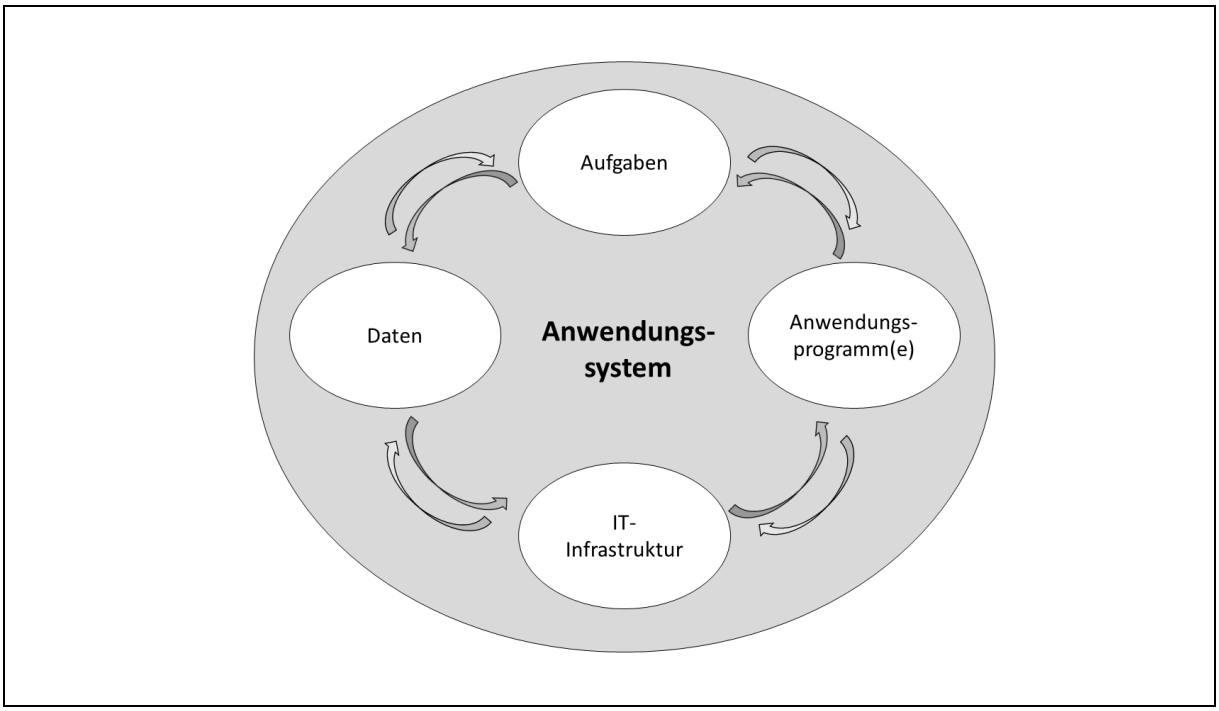


Abbildung 17: Elemente von Anwendungssystemen (eigene Darstellung)

Die IT-Infrastruktur ist in Elemente zu unterteilen. In der folgenden Abbildung 18 wird die IT-Infrastruktur, unterteilt in Informations- und Kommunikationstechnik, dargestellt. Die Bestandteile beider Bereiche werden in Hardware und Software aufgelistet. In der Informationstechnik ist die Kategorie der Hardware durch Endgeräte und Peripherie zusammengesetzt. Unterschieden wird hier wiederum die Nutzung und technische Zusammenstellung beider zuvor genannten Hardwarekomponenten. Ein Endgerät verfügt über Bauteile, aus denen es zusammengesetzt ist, während Peripheriegeräte an ein Endgerät angeschlossen und zur Speicherung von Daten sowie zur Informationseingabe und Informationsausgabe fähig sind. Diese sind bspw. Maus, Tastatur, Monitor, Drucker etc. Die Software wird indessen unterschieden in Systemsoftware und Anwendungssoftware (vgl. Mertens, 1995, S. 37–38).

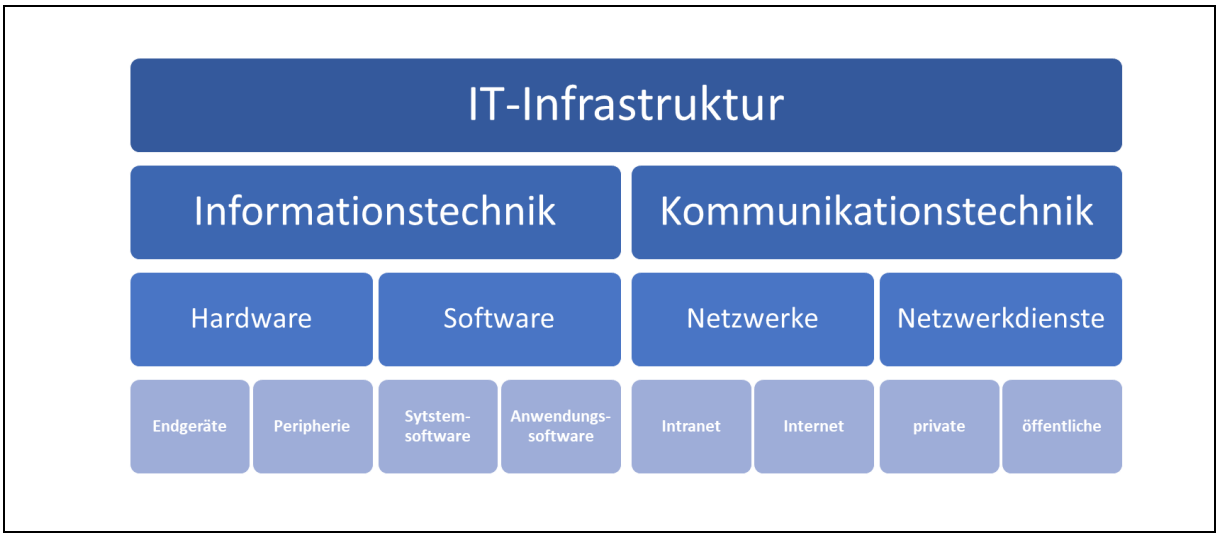


Abbildung 18: Übersicht der IT-Infrastruktur (eigene Darstellung in Anlehnung an Mertens, 1995, S. 37 f.)

Systemsoftware umfasst Betriebssysteme, Compiler und Interpreter, Dienstprogramme und systemnahe Software (vgl. Teuber, 1999, S. 23). Ein Betriebssystem umfasst alle Programme, die für die Nutzung eines Endgerätes erforderlich sind. Sie steuern den Datenfluss zwischen Hardware und Software sowie zwischen den einzelnen Hardwarekomponenten. Ein Betriebssystem bildet eine Schnittstelle zwischen Mensch und der Hardware. Durch Standardisierung stellt ein Betriebssystem eine Plattform da, von der aus Anwendungsprogramme installiert und genutzt werden. Zudem erfüllt das Betriebssystem auch die Funktion der Steuerung und Verwaltung im Sinne der Hardware (vgl. Abts & Mülder, 2017, S. 44).

Aufgaben:

- Steuerung und Überwachung der Ausführung von Programme (Software),
- Verwaltung der Hardware-Komponente bspw. Prozessor, Arbeitsspeicher, Peripherie,
- Datenverwaltung,
- Bereitstellung einer Benutzerschnittstelle.

Die Übersetzung des Programmcodes in Maschinensprache wird durch Compiler und Interpreter vollzogen. Basisdienste sind Programme für die allgemeine Benutzung und Wartung. Systemnahe Software sind Programme zur Datenverwaltung, z.B. Datenbankverwaltungssysteme und Software-Entwicklung (vgl. Teubner, 1999, S. 23).

Die Anwendungssoftware unterstützt konkrete betriebliche Aufgaben. Hiernach wird Anwendungssoftware auch in Standard- und Individualsoftware-Programme gegliedert (vgl. Hansen & Neumann, 2005, S. 29). Sie umfasst Programme, die für den Massenmarkt entwickelt

wurden, sogenannte Standardsoftware. Aus betriebswirtschaftlicher Sicht unterstützen sie eine Funktion und können von mehreren Anwendungsbereichen genutzt werden. Dies gilt für die in der Abbildung des Integrationsmodells dargestellten Funktionsbereiche wie Produktion, Vertrieb, Logistik etc. Somit können sie modular aufgebaut sein und auf eine lokale oder gemeinsame zentrale Datenbasis zurückgreifen. Standardsoftware wird an individuelle Einsatzanforderungen und unternehmensabhängig angepasst. Auch kann eine Erweiterung der Funktionen unter gegebenen Umständen durch die Schnittstellen erzielt werden. Individualsoftware hingegen wird speziell für eine besondere betriebliche Anforderung mit der zugehörigen Hard- und Software-Umgebung entwickelt (vgl. Mertens et al., 2017, S. 16–18).

Die Hardware der Kommunikationstechnik umfasst alle physisch-technischen Geräte, die die Verbindung zwischen Endgeräten wie Computer oder Smartphone herstellen. Diese umfassen kabelgebundene Übertragungswege oder kabellose Übertragungseinrichtungen. Als Beispiel sind Repeater oder Hubs zu nennen. Bei den Vermittlungseinrichtungen ist der Router zu erwähnen. Dabei wird auch zwischen Netzwerken in private und öffentliche Dienste unterschieden (vgl. Teubner, 1999, S. 23). In einem Netzwerk angebotene Leistungen werden als Netzwerkdienste bezeichnet. Diese ermöglichen einerseits Nachrichtendienste in Form von E-Mails und andererseits Datentransfers. Auch diese Dienste können öffentlich und privat einem eingeschränkten Personenkreis angeboten werden (vgl. Hansen & Neumann, 2005, S. 29).

Die IT-Infrastruktur stellt somit die Grundlage dar, auf der die Organisationen, Anwendungs- und Informationssysteme funktionieren (vgl. Laudon et al., 2016, S. 24).

In der Automatisierungspyramide sind die relevanten Systeme abgebildet. Sie ist durch eine Hierarchie über sechs verschiedene Ebenen gekennzeichnet, die verschiedene Aufgaben haben und unterschiedliche Unternehmensdaten zusammenführen. Zudem ist die Automatisierungspyramide nach der Norm DIN ISO 62264 definiert, welche das Schichtenmodell und die Einsatzgebiete der angewandten IT-Systeme für das Produktionsumfeld charakterisiert (vgl. Schöning & Dorchain, 2014, S. 543). Abbildung 19 stellt eine klassische Automatisierungspyramide dar.

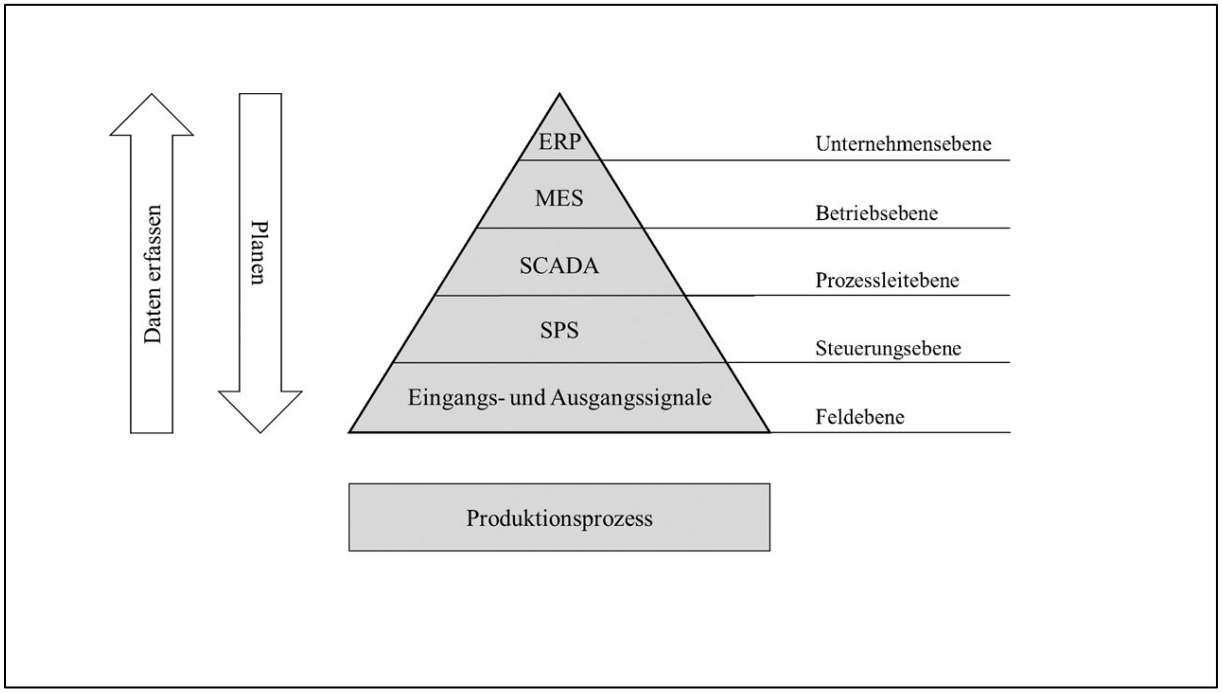


Abbildung 19: Automatisierungspyramide (eigene Darstellung in Anlehnung an Siepmann, 2016, S. 49)

Die oberste Ebene der Automatisierungspyramide beinhaltet das ERP-System. Diese Unternehmensebene unterstützt die Produktionsplanung und -steuerung. Hierunter fallen bspw. Aufgaben der Bestellabwicklung (vgl. Fallenbeck & Eckert, 2014, S. 405). Ein ERP-System ist ein betriebswirtschaftliches System aus einer standardisierten Software und der dazu benötigten IT-Infrastruktur. Das ERP-System wird zum Planen, Steuern und Verwalten eines Unternehmens verwendet. Die Aufgaben eines ERP-Systems sind in zwei Kategorien zu unterteilen. Betriebswirtschaftlich unterstützt das ERP die bedarfsgerechte Planung und Verteilung von Ressourcen. Darunter sind Kapital, Personal, Betriebsmittel und Materialien zu nennen. Informationstechnisch dient ein ERP-System als Informations- und Kommunikationswerkzeug zwischen den einzelnen Funktionsbereichen eines Unternehmens und darüber hinaus (horizontale Integration). Auf dieser Unternehmensebene werden Ziele und Vorgaben sowie Ressourcen und Fortschritte geplant, verwaltet und kontrolliert (vertikale Integration). Ein ERP-System enthält Module der Finanz- und Buchhaltung, des Personalwesens, der Produktionsplanung und -steuerung, des Einkaufs sowie der Logistik (vgl. Wannenwetsch, 2014, S. 458 f.).

Das MES stellt nun die Betriebsebene dar. Auf dieser Ebene werden die einzelnen Produktionsprozesse genauer definiert. Anhand von KPIs werden diese Prozesse überwacht. Darüber hinaus findet auch das Material- und Qualitätsmanagement auf dieser Ebene statt (vgl.

Fallenbeck & Eckert 2014, S. 405). Das Manufacturing Executive System ist die Schnittstelle zwischen der betriebswirtschaftlichen Planung und der Fertigungsebene. Das MES stellt die Feinplanung der Produktion auf Basis und Zielvorgaben des ERP-Systems her und leitet diese an die Steuerungsebene weiter (vgl. Kletti, 2007, S. 119 f.). Auf der Prozessebene werden SCADA-Systeme eingesetzt. Diese Ebene dient dem Monitoring des gesamten Produktionssystems, da hier alle Messwerte der unteren Schichten zusammenkommen. Eine weitere Funktion dieser Ebene ist die Archivierung von Messwerten (vgl. Fallenbeck & Eckert, 2014, S. 405). Auf Steuerungsebene werden die einzelnen Maschinen durch SPS-Systeme bedient (vgl. Fallenbeck, Eckert 2014, S. 405), während auf Feldebene Ein- und Ausgabesignale als technische Schnittstelle zu den Prozessen der Produktion fungieren. Die unterste Ebene stellt der Produktionsprozess selbst dar. Hier werden mittels binärer Signale an Sensoren und Aktoren einfach und schnell Daten in Maschinen gesammelt (vgl. Fallenbeck & Eckert, 2014, S. 405).

Ein Kritikpunkt der Automatisierungspyramide ist, dass es wenige Schnittstellen zwischen den Ebenen gibt. Insofern ist keine direkte Rückkopplung an die einzelnen Prozesse möglich. Da durch die Statik des Modells die Optimierung der Prozesse ex post erfolgt, kann keine Lösung in Echtzeit gefunden werden (vgl. Schöning & Dorchain, 2014, S. 550).

In der folgenden Tabelle 3 werden die einzelnen Systeme als Übersicht dargestellt.

	Prozessebene	**Aufgaben**
ERP-System	Unternehmensebene	• Administrative Aufgaben • Grobe Produktionsplanung
MES-System	Betriebsebene	• Detaillierte Produktionsplanung • Schnittstellenbildung von Unternehmensebene zu Steuerungsebene
SCADA-System	Prozessleitebene	• Überwachung der Steuerungsebene
SPS-System	Steuerungsebene	• Steuern und Regeln der Anlagen/Maschinen • Ausführung der MES-Vorgaben
Eingangs- & Ausgangssignale	Feldebene	• Schnittstellenfunktion

Tabelle 3: Übersicht der Produktivsysteme (eigene Darstellung)

2.2.2 Systemschnittstellen

Das SCADA als Herzstück der Fabrik

Der folgende Abschnitt behandelt die eingesetzten Softwares und IT-Systeme, die beim Karosseriebau bei einem der führenden Automobilhersteller eingesetzt werden. Zudem werden

die wechselseitigen Beziehungen zueinander hergestellt und die Problemantik der Insellösungen der aktuellen Situation an einem Standort herausgearbeitet.

SCADA ist ein intelligentes IT-Software-System, das im Werk eingesetzt wird. Das Lösungspaket umfasst die visuelle Darstellung und bietet über den webbasierten Zugang zu jeder Zeit Zugriff auf die Informationen. Basierend auf der ERP-Schnittstelle zu der Produktionsanlage wird eine gemeinsame Datenbank erzeugt, um eine Vielzahl an übersichtlichen und aktuellen Auswertungen zu generieren sowie zur Verfügung zu stellen. Die Leittechnik ist das einzige IT-Element der Fertigung zur Archivierung und Auswertung von Anlagenzuständen.

Abbildung 20 gibt eine grobe Übersicht über die implementierte Schnittstelle von SCADA in der Fertigung.

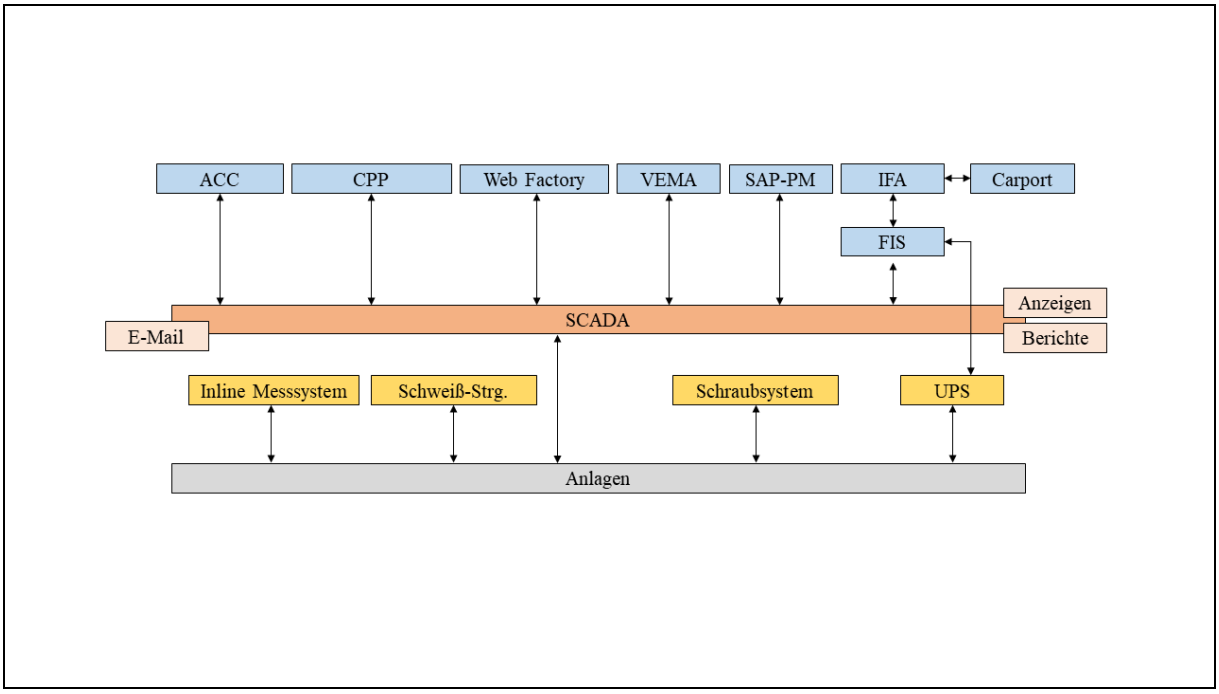


Abbildung 20: MES/SCADA-Schnittstelle (eigene Darstellung)

Das SCADA-System dient als Schnittstelle zwischen der Anlage und diversen Anwenderprogrammen. So bezieht das Programm unterschiedliche Informationen aus der gesamten Anlage und sammelt sie für eine Aufbereitung. Zu den Anlagenschnittstellen zählen Fördertechnikanlage, Frequenzumrichter und Schraubersteuerung. So werden hier bspw. Daten von Messsystemen zusammengetragen. Hierzu zählen Temperaturen, Maschinenauslastungen und -daten. Aktuell werden über SCADA ausschließlich Daten aus der Fertigung archiviert und somit für eine Auswertung zur Verfügung gestellt. Somit bildet die IT-Software das Zentrum für die Kommunikation zwischen der Anlage und Benutzern. Jedoch fällt in Abbildung 20 auf,

dass es mehrere Schnittstellen gibt, die keine direkte Verbindung haben. So müssen bspw. die Informationen aus dem UPS erst an das FIS geleitet werden. Hier umgeht das System EcoEmos bzw. leitet die Informationen über das FIS in das SCADA-System. Anschließend werden die Informationen über die einzige Schnittstelle IFA an den Carport geleitet. Hier lässt sich bereits erahnen, dass die Kompatibilität der einzelnen Programme und Systeme eine Herausforderung sowohl für die direkte Kommunikation als auch für die gesamten Datenmassen darstellt. Um den Zusammenhang des SCADA in der gesamten Fertigung des Werks zu verdeutlichen, folgt eine weitere detaillierte Übersicht. Folgende Abbildung zeigt die Schnittstellen für die Vernetzung der einzelnen Abteilungen mit dem SCADA-System für die einzelnen Fertigungsbereiche.

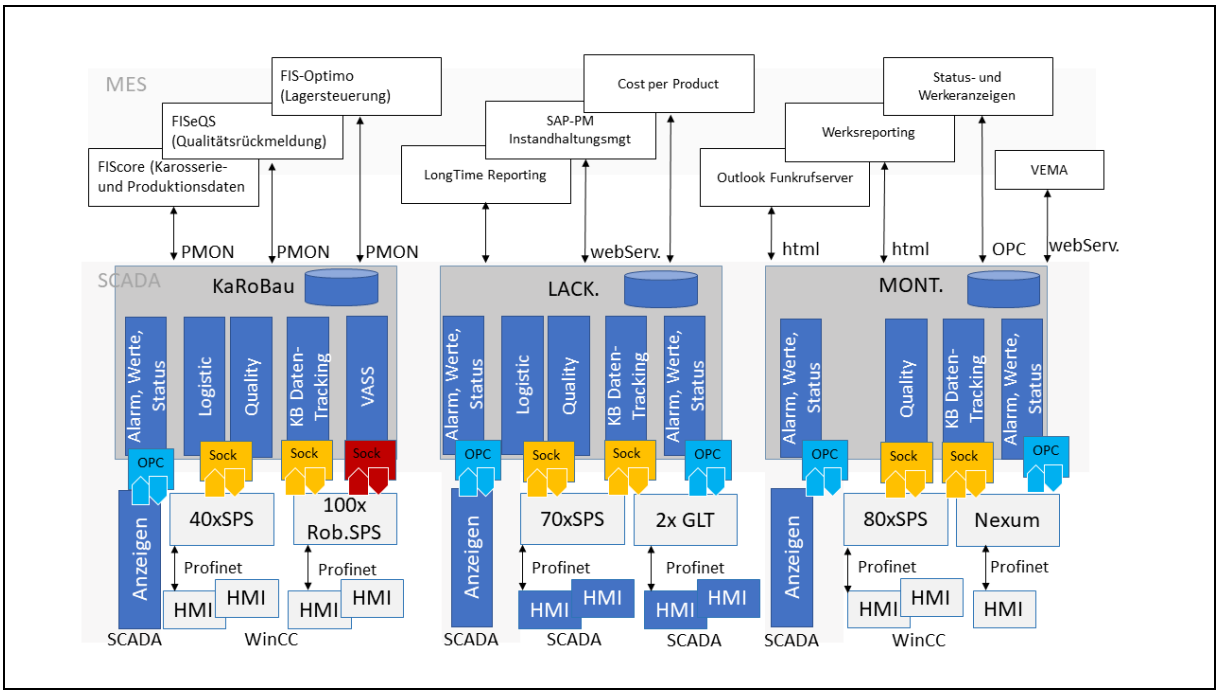


Abbildung 21: SCADA-Öko-System (eigene Darstellung)

Abbildung 21 zeigt die zentralen Schnittstellen der Werksabteilungen zum SCADA-System. Auf dieser Prozessleitebene werden alle Messwerte aus der Steuerungs-, Feld- und Prozessebene archiviert und für das Monitoring aufbereitet. Das Besondere an der Anzeigenschnittstelle zum SCADA-System ist, dass die drei Produktionsschritte unabhängig voneinander fungieren. Der Fokus liegt hierbei beim Karosseriebau auf SCADA-KaRoBau und der Verbindung zum FIS, dem sogenannten Fertigungs-, Informations- und Steuerungssystem, auch gleichzusetzen mit dem in der Literatur und der Arbeit beschriebenen Manufactoring Executive System (MES). FIS wird zentral gesteuert und ist die direkte Verbindung zwischen dem ERP-System und dem SCADA-System im Werk. Das SCADA-System beinhaltet die

Funktion, zahlenwertcodierte Informationen über Störungen im Betriebsablauf an die Leittechnik zu melden und dort anschließend auszuwerten.

Fehlermeldungen, die im Karosseriebau verursacht werden, werden in der SPS aufbereitet und anschließend an ein Visualisierungssystem übergeben. Hierbei werden die Arten der Stillstände klassifiziert und priorisiert. Tabelle 4 gibt einen Überblick über die Gliederung der unterschiedlichen Stillstandarten.

Stillstandart	Farbe	Priorität	Erklärung	Beispiel
Technik	rot	1	Technischer Defekt, Fehlfunktion führt zum Stillstand der Anlage	Phasenausfall, Melderkontrolle
Fremdsystem	magenta	2	Technischer Defekt, Fehlfunktion von Fremdsystemen/Prozessgeräten	Schraubtechnik Master-PC
System	gelb	3	Sicherheitseinrichtung greift durch Not-Halt-Schalter, Scanner etc.	Fördergutmangel
Qualität	orange	4	Qualitätsalarm (Reisleine) ausgelöst	Fixposition Stopp
Organisatorisch	blau	5	Bandstopp durch fehlende Fertigmeldung	Anlage im Handbetrieb
Warnmeldung	violett	6	Störungen ohne Stillstand der Anlage	SPS-Systemfehler

Tabelle 4: Hierarchie der Fehlermeldungen aus SCADA (eigene Darstellung)

Hier wird ein Auszug an detaillierteren Fehlerklassen, der jeweiligen Meldung und den Meldungstypen aufgelistet, um einen Überblick der unterschiedlichen Fehler in der Fertigung zu erhalten. Neben den Stillstandarten, welche hierarchisch in Technik, Fremdsystem, System, Qualität, organisatorisch und Warnmeldung angesiedelt sind, werden diesen auch entsprechende Farben und numerische Prioritäten zugeordnet. Das Farbenspektrum erstreckt sich von Rot, Magenta, Gelb, Orange über Blau bis Violett in absteigender Relevanz. Die Prioritäten sind ebenfalls in absteigender Bedeutung numerisch gekennzeichnet. Darüber beinhaltet Tabelle 4 auch jeweils eine Erklärung der Stillstände und wird durch ein Beispiel ergänzt.

Nachdem die Fehlerhierarchie tabellarisch dargestellt wurde, erfolgt nun ein Einblick in einzelne Fehlermeldungen aus dem System. Tabelle 5 zeigt einen Auszug aus den unterschiedlichen Fehlerklassen.

Fehlerklasse	Meldung	Meldungstyp
01	Gerätestörung	Stillstandmeldung
02	Gerätewarnung	Warnmeldung
03	Not-Aus	Stillstandmeldung
04	Bandstopp	Stillstandmeldung
05	Qualitätsstopp	Stillstandmeldung
06	Organisatorische Störung	Stillstandmeldung
07	Organisatorische Warnung	Warnmeldung
08	Systemstörung	Stillstandmeldung
09	Systemwarnung	Warnmeldung
10	Betriebsart	Warnmeldung

Tabelle 5: Meldungstypen in SCADA (eigene Darstellung)

Es ist bei der Betrachtung von Tabelle 5 nicht notwendig, alle Betriebsstörungen zu nennen, da dies nicht zielführend ist. Die Tabelle dient lediglich der Auflistung von Meldungen und daher als Übersicht. Numerisch werden Fehlerklassen bestimmt, welchen eine spezifische Meldung zugeordnet wird. Darüber hinaus zeigt die Tabelle 5 zwei unterschiedliche Meldungstypen. Hierbei handelt es sich um Stillstandmeldungen und Warnmeldungen. Eine Stillstandmeldung hat eine Aktion zur Folge, da der Betriebsablauf gestört wird, wohingegen eine Warnmeldung eine Überprüfung der betreffenden Stelle impliziert.

Des Weiteren werden im SCADA-Monitor nicht nur die Fehlertypen anhand der Kategorien Warnmeldungen und Stillstandmeldungen bestimmt, sondern auch die Verantwortlichkeiten dafür. Diese sind in Tabelle 6 aufgelistet.

Zuordnung	Verantwortlichkeit
A	Produktion
B	Wartung
C	Produktionssteuerung
D	Motor
E	Lieferant
F	Teileentwicklung
G	Produkt
H	IT
I	Verwaltung
J	Leittechnik
K	Qualitätskontrolle
L	Sonstige

Tabelle 6: Verantwortlichkeiten für Störmeldung (eigene Darstellung)

Anhand der Zuordnung der Verantwortlichkeiten aus Tabelle 6 kann eine Störung direkt einem Bereich zugeordnet werden und die Problemlösung anschließend veranlasst werden. Die Zuordnung erfolgt alphabetisch und beinhaltet die Verantwortlichkeiten Produktion, Wartung,

Produktionssteuerung, Motor, Lieferant, Teileentwicklung, Produkt, IT, Verwaltung, Leittechnik, Qualitätskontrolle und Sonstige.

Die Informationen der vorherigen Tabellen 4–6 zu Betriebsmeldungen über Stillstände können wie folgt in Tabelle 7 zusammen abgebildet werden:

Stillstandmeldungen	Kennung	EcoEmos Fehlerklasse		Definition in EcoEmos	EcoEmos Farbschema
Stillstand Technik	ST	1	Geräte-störung	Alle technischen Meldungen, die zu einem Stopp der Produktion führen	
Stillstand Systembedingt	SS	3	Not-Aus	Alle Meldungen, die einer Sicherheits-einrichtung zugeordnet werden	
Stillstand Systembedingt	SS	8	Systemstörung		
Stillstand Organisatorisch	SO	6	Organisatorische Störung	Produktionsstopp Werkerbedingt z.B. Einlaufsperre, Fahrschalter	
Stillstand Fremdsystem (Stillstand Technik)	SF	4	Externe Geräte-störung	Produktionsstopp Fremdanlage im entsprechenden Takt, z.B. Befüllanlage hat eine Störung	
Stillstand Qualität (Stillstand Organisatorisch)	SQ	5	Qualitätsstörung	Qualitätsstopp, z.B. Reißleine	

Tabelle 7: Zusammenfassung der Stillstände in SCADA (eigene Darstellung)

Diese Übersicht gibt einen tabellarischen Überblick über die verschiedenen Stillstände in der Fertigung. Hierbei werden die Stillstandmeldungen in fünf Kategorien differenziert. Dazu zählen Technik, systembedingt, organisatorisch, Fremdsystem und Qualität. Hinsichtlich der Störungen hat jeder Stillstand eine eigene Kennung. Darüber hinaus werden die Stillstand-meldungen in verschiedene Fehlerklassen, sowohl numerisch als auch mit entsprechender Klassifizierung, eingeordnet. Die Tabelle 7: Zusammenfassung der Stillstände in SCADA beinhaltet weiterhin Informationen über jeweilige Störungsdefinitionen und die Zuweisung der Farbkategorie in EcoEmos. Die Meldungen über Störungen im Betriebsablauf in der Produktion erscheinen anschließend in EcoEmos in einer definierten Struktur.

Neben den Stillstandmeldungen existieren ebenfalls Warnmeldungen in EcoEmos. Diese werden in Tabelle 8 näher betrachtet.

Warnmeldung	Kennung	SCADA-Fehlerklasse		EcoEmos Farbschema
Meldung Technik	MT	2	Gerätewarnung	
Meldung Systembedingt	MS	9	Systemwarnung	
Meldung Systembedingt (Hand)	MS	10	Betriebsart	
Meldung Systembedingt (Automatik)	MS	10	Betriebsart	
Meldung Organisatorisch	MO	7	Organisatorische Warnung	
Meldung Fremdsystem	MF	4	Externe Störung	
Meldung Qualität	MQ	5	Qualitätsstörung	

Tabelle 8: Zusammenfassung der Warnmeldungen in SCADA (eigene Darstellung)

Dabei werden die unterschiedlichen Warnmeldungen vom Monitoring veranschaulicht. Hierzu erfolgen die Meldungen analog zu den Störungen aus Tabelle 7. Zu den fünf Arten von Warnmeldungen zählen Technik, systembedingt, organisatorisch, Fremdsystem und Qualität. Auch diese haben eine spezifische Kennung und werden in verschiedene Fehlerklassen untergliedert. Zudem wird eine Warnmeldung anhand des definierten Farbschemas in der Anzeige dargestellt.

Nachfolgend stellt Abbildung 22 die Visualisierung des gesamten Karosseriebaus durch das SCADA-System für den Benutzer dar. Zusammenfassend visualisiert SCADA somit das Monitoring in der gesamten Endmontage.

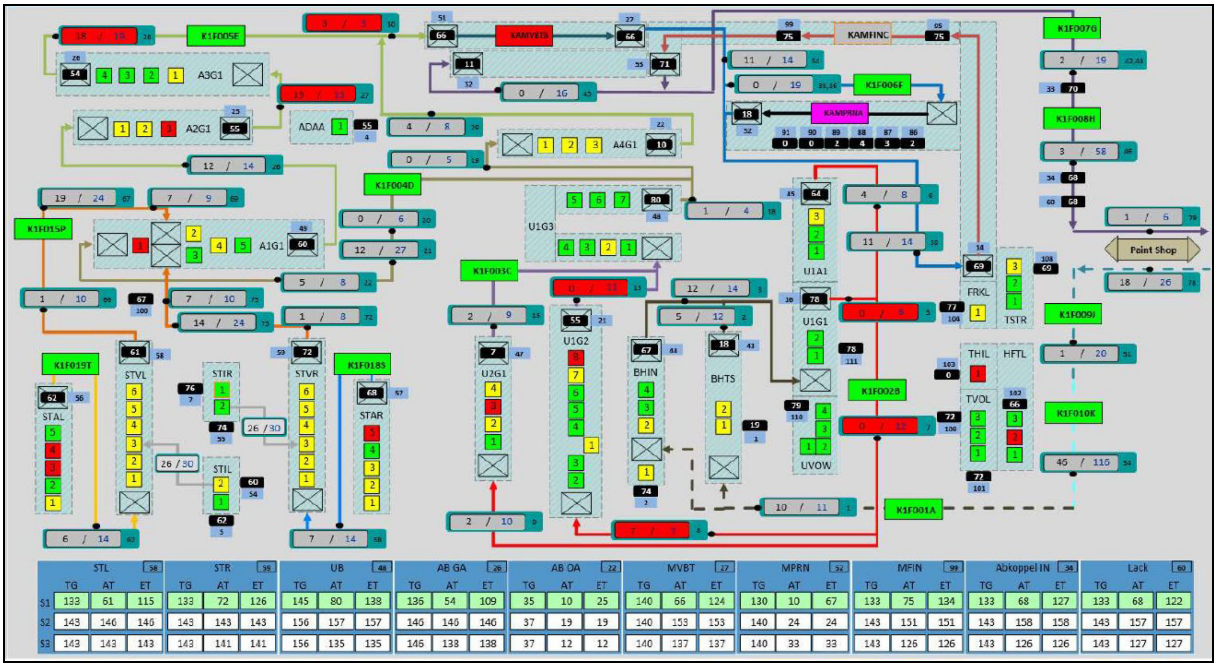


Abbildung 22: Darstellung des Karosseriebauprozesses durch SCADA (Auszug aus SCADA)

2.2.3 Lösungsansatz zur Aufhebung der Restriktion des Schichtenmodells

Eine Lösung für die bestehenden Restriktionen des Schichtenmodells ist die Abbildung der Strukturen mithilfe von cyber-physischen Systemen. Auf technischer Ebene werden dadurch System-, Produktions-, Logistik-, Engineering-, Koordinations- und Managementprozesse sowie das Internet der Dienste ineinander eingebunden (vgl. Kagermann et al., 2013). Physisch betrachtet besteht ein CPS aus Sensoren, Aktoren, Benutzerschnittstellen und einer Kommunikationstechnologie für die Interaktion. Aus logischer Sicht können CPS Informationen erfassen, transportieren, verarbeiten und bereitstellen (vgl. Lucke, Defranceski & Adolf, 2017, S. 75). Durch die Änderung der Automatisierungsstrukturen können Daten aus verschiedenen Ebenen direkt mit der Feldebene gekoppelt werden. Dies hat zur Folge, dass die fehlenden Rückkanäle implementiert werden und Loops für Feedback genutzt werden können (vgl. Schöning & Dorchain, 2014, S. 550). Abbildung 23 stellt die Struktur unter Einsatz von CPS beispielhaft dar.

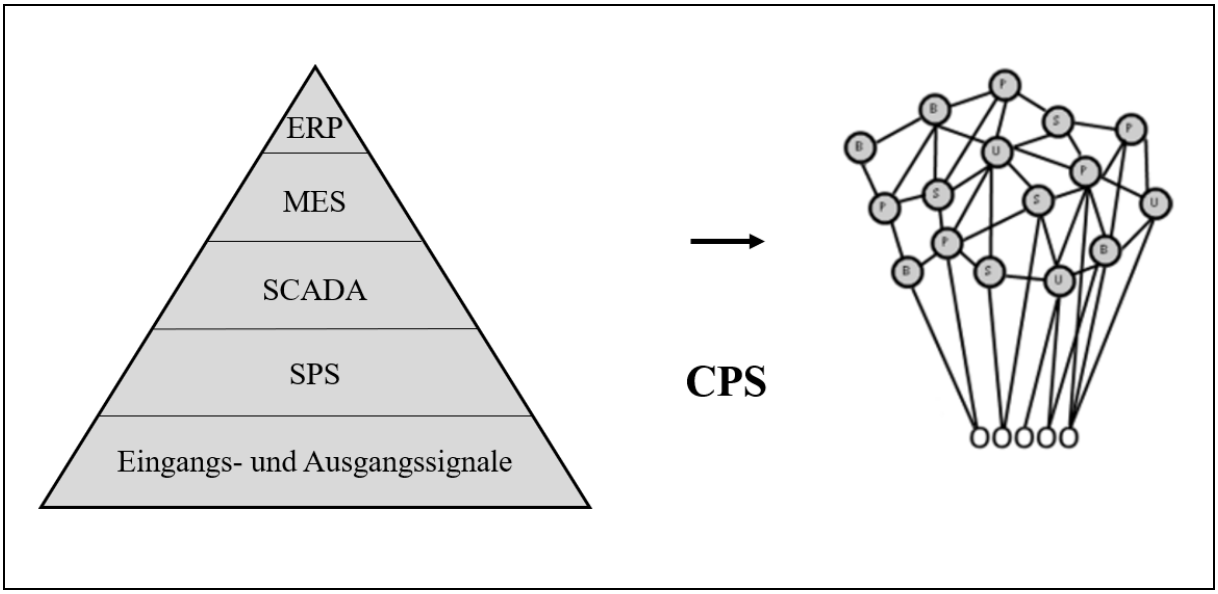


Abbildung 23: Schichtenmodell mit CPS-Automation (eigene Darstellung in Anlehnung an Schöning & Dorchain, 2014, S. 550)

Der Einsatz eines effizientes CPS erfordert bestimmte technische Besonderheiten. So müssen zum einen die vertikale und zum anderen die horizontale Integration vorhanden sein. Die vertikale Integration in Bezug auf die Nutzung von CPS bedeutet, dass die im Schichtenmodell enthaltenen Systemebenen miteinander verknüpft sind. Somit ist eine technische Struktur in Form einer Systemlandschaft Voraussetzung (vgl. Siepmann & Graef, 2016, S. 37). Diese vertikale Integration dient dem Austausch von Daten zwischen den Hierarchieebenen. Dafür müssen die Systeme standardisierte Schnittstellen aufweisen, damit die Maschine-zu-Maschine-Kommunikation möglich ist. Hierzu müssen Sensoren, Aktoren, eingebettete Systeme bis hin zu ganzen Produktionsanlagen sowie Planungs- und Steuerungssysteme miteinander verknüpft sein (vgl. Siepmann & Graef, 2016, S. 37). Nach BROSSARDT ist dabei die Erhebung und Speicherung der Produktionsdaten von Bedeutung. Dabei werden die Daten automatisiert bearbeitet und ausgewertet. Auf Grundlage dessen werden die Produktion und die Ressourcenverwaltung verbessert, denn so können notwendige Informationen an die Produktionsprozesse der Fertigung weitergeleitet werden (vgl. Brossardt, 2014, S. 8–9).

In der horizontalen Integration ist die Vernetzung der innerbetrieblichen Systeme mit außerbetrieblichen Systemen möglich. Systeme außerhalb des Unternehmens, z.B. jene von Kunden, Lieferanten oder auch an anderen Standorten, sowie externe Dienstleister sind auf diese Weise mit dem betrieblichen System verbunden (vgl. Siepmann & Graef, 2016, S. 37). Damit ist eine Grundlage für ein in Echtzeit dynamisches Wertschöpfungsnetzwerk über die Unternehmensgrenzen hinweg möglich (vgl. Brossardt, 2014, S. 9–10).

Darüber hinaus ist eine weitere Lösung aus der Software-Betrachtung zur Optimierung der Verknüpfung der zuvor genannten Systeme denkbar. Daraus entstand die Überlegung einer Big-Data-Architektur. Die Dissertation betrachtet in diesem Zusammenhang die Systemlandschaft eines der führenden Automobilhersteller und optimiert die vorliegende Diskrepanz in Bezug auf die Datenintegrität, die Schnittstellenproblematik und den daraus resultierenden Informationsflussprozess mit der Konzeptionierung und Implementierung einer neuen Softwarearchitektur. Durch die Big-Data-Architektur und deren Tools für Datenbeschaffung, Datenverarbeitung und Datenanalyse werden die Systemebenen miteinander verknüpft, um somit eine bessere Aussagekraft und Entscheidungsfindung zu erreichen. Im Besonderen wird dabei die Optimierung der vertikalen Integration angestrebt. Der Vorteil bei der Verwendung bestehender Systeme und der Optimierung durch das „Overlaying" eines oder mehreren Software-Tools aus Unternehmenssicht besteht einerseits in der Umsetzungsgeschwindigkeit, Kosten sowie Nutzbarkeit und andererseits kann die Software-Architektur beliebig angepasst und weiterentwickelt werden. Der Autor richtet sich hierbei an das Schichtenmodell nach SCHÖNING und DORCHAIN, um die Implementierung kurz darzustellen, siehe Abbildung 24. Die nächsten Kapitel geben Ausschluss über die Durchführung des Lösungsansatzes. Daher wird nun Kapitel 3 die Thematik der Datenwissenschaft einleiten, um daraufhin das Konzept der Big-Data-Architektur vorzustellen.

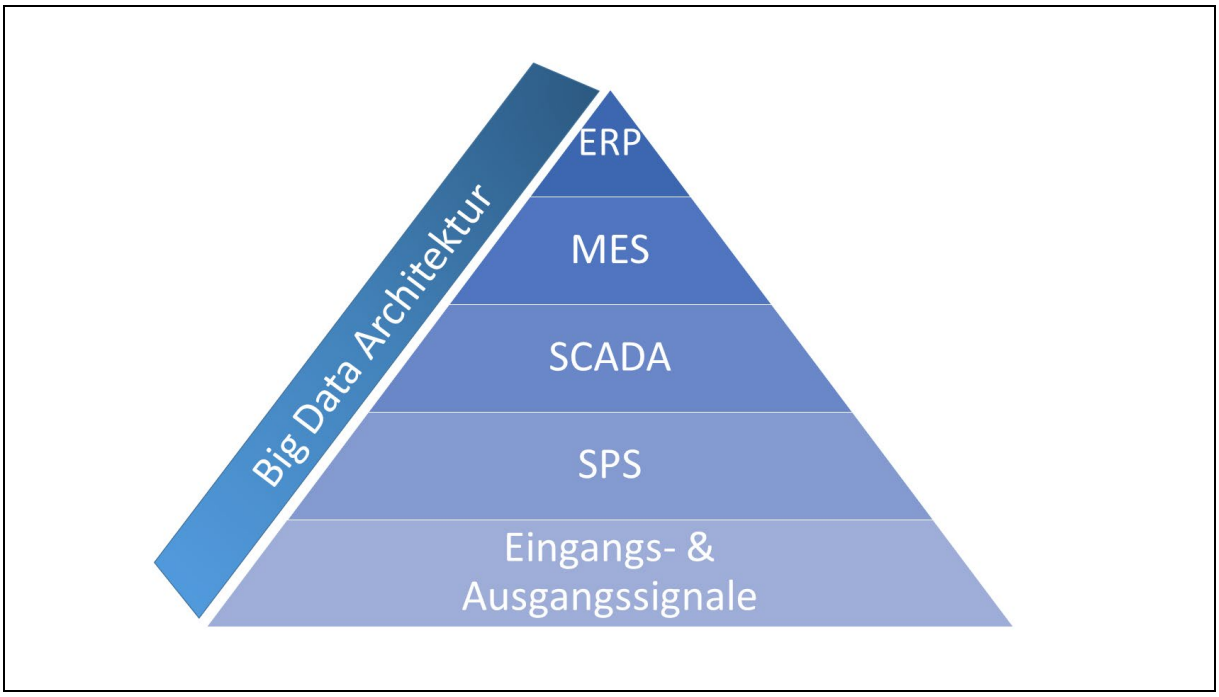


Abbildung 24: Implementierungsansatz an das Schichtenmodell (eigene Darstellung)

3. Data Science

Der Begriff „Data Science“ wurde in den 60er Jahren aufgegriffen und erstmalig vom dänischen Informatik-Pionier Peter Naur erwähnt. Für NAUR war Data Science ein Synonym für Informatik als Wissenschaft (vgl. Goldschmidt, 2019, S. 36). Vom Englischen ins Deutsche übersetzt bedeutet Data Science auch wörtlich „Datenwissenschaft“. Nun ist die Datenwissenschaft breit gefächert und muss zunächst im branchen- und unternehmensspezifischen Kontext klar definiert werden.

3.1 Begriffserklärung Data Science

Data Science kann als eine Art Weiterentwicklung und somit Vertiefung der Informatik gesehen werden. Die Schnittstellenfunktion und die auf Datenwissen basierten Inhalte und Analysen sind die wichtigsten Eigenschaften von Data Science. Aus der Historie heraus sind immer wieder unterschiedliche Definitionen und damit einhergehend unterschiedliche Betrachtungsweisen erkennbar (vgl. Haneke, Trahasch & Zimmer, 2019, S. 6). Eine verständliche Definition nach DONOHO lautet: *„This coupling of scientific discovery and practice involves the collection of management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.“* (Donoho, 2015, S. 745) Hierbei rückt neben der Interdisziplinarität auch der Ansatz der Entdeckung neuer wissenschaftlicher Erkenntnisse aus Daten in den Vordergrund. Für DONOHO ist Data Science eine Wissenschaft, die Daten erzeugt, validiert und erklärlich aufbereitet. Dabei kommen Werkzeuge wie Maschine Learning und Algorithmen zum Einsatz, um große Datenmengen zu analysieren (vgl. Donoho, 2015, S. 745).

In der Abbildung 25 ist die Interdisziplinarität dieses Themenbereichs dargestellt. Im Venn-Diagramm nach CONWAY werden unterschiedlichste Disziplinen und deren Know-how miteinander in Verbindung gebracht. Daraus entstehen neue Schnittstellen und neues Wissen, welches sich im Laufe der Zusammenarbeit entwickelt. Data Science ist der Mittelpunkt, an dem Wissen als Informationen zusammenkommt. Damit stellt Data Science folgende Anforderungen in Bezug an einen Datenwissenschaftler (Data Scientist) auf:

- Mathematisches Wissen,

- Informatikwissen,
- Betriebswirtschaftliche Kenntnisse.

Die Abgrenzung zum Berufsbild des klassischen Informatikers, Statistikers oder auch Mathematikers ist durch die entstandene Schnittstellenfunktion und das dafür benötigte Know-how abbildbar (siehe Abbildung 25).

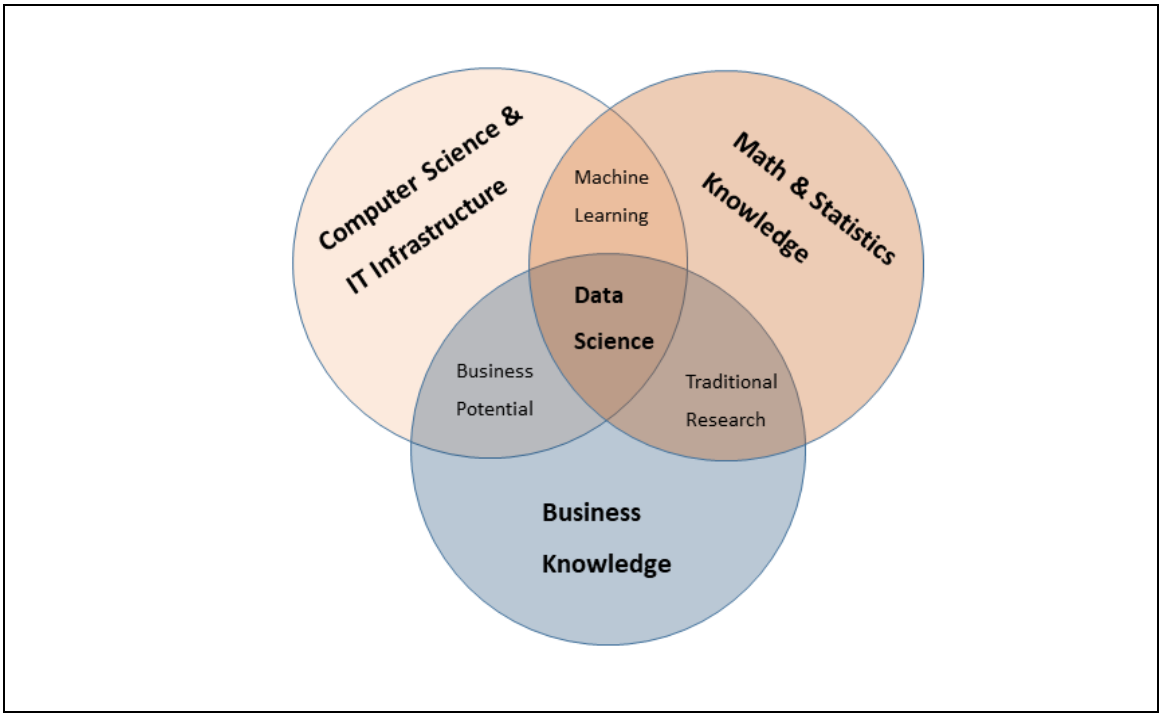


Abbildung 25: Interdisziplinarität von Data Science als Venn-Diagramm (eigene Darstellung in Anlehnung an Conway, 2010)

3.2 Data-Science-Projekte

Nun wird anhand zweier Modelle – des CRISP-DM-Modells und des ASUM-DM-Modells – erläutert, wie der Prozess eines Data-Science-Projekts aussehen kann. Das in der Dissertation beschriebene Projekt wird auf der Basis dieser Modelle erarbeitet.

Das CRISP-DM-Modell (engl. Cross Industry Standard Process for Data Mining) beinhaltet folgende sechs Phasen: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment (siehe Abbildung 26). Im Business Understanding werden die Ziele und der Analyseprozess in Abhängigkeit vom Geschäftsumfeld und dessen Zielen festgelegt. Data Understanding und Data Preparation sind iterative Teilprozesse, in denen der Datenbestand vorbereitet wird. Dabei wird die Konstruktion des finalen Datensatzes für die Modellierung bestimmt. Danach folgen das Modellieren und die Analyse der Daten im

Modeling-Prozess. Hier werden geeignete Data-Mining-Verfahren, optimierte Parameter und mehrere Modelle erzeugt. Im Prozessschritt der Evaluation wird im Anschluss daran das geeignete Modell auf seine Validität geprüft, um dann in der letzten Prozessphase (Deployment) die Ergebnisse bereitzustellen und das Modell schlussendlich in einen Entscheidungsprozess zu integrieren (vgl. Shearer, 2000, S. 13–22).

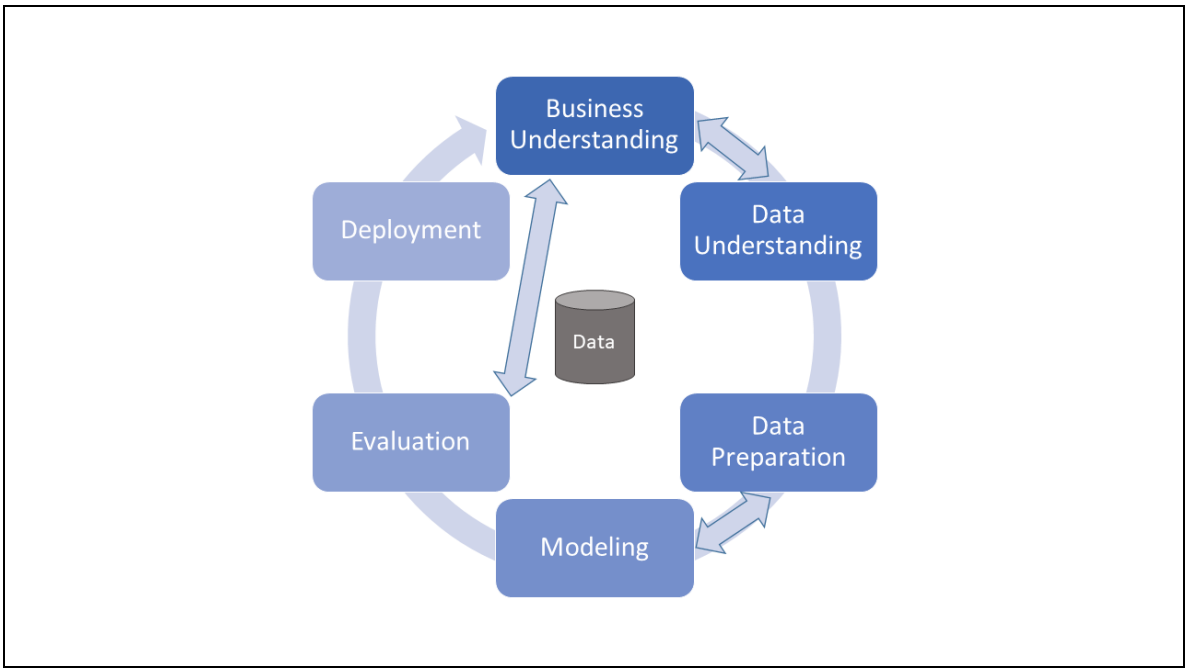


Abbildung 26: CRISP-DM-Modell (eigene Darstellung in Anlehnung an Shearer, 2000)

Das ASUM-DM-Modell ist die Weiterentwicklung des zuvor erläuterten CRISP-DM-Modells. Das Ziel des von IBM entwickelten Modells ist, die statischen Schwachstellen des CRISP-DM-Modells zu eliminieren. Zu erkennen ist nun der Projektmanagementprozess, der als begleitend und kontrollierend parallel zum genannten Modell eingesetzt wird. Das ASUM-DM-Modell ist agil. Auch in den Prozessschritten ist die Kommunikation aller Beteiligten ein Mittel, um Schwachstellen im Projekt von Anfang an zu erkennen und dadurch die Prozesse zu verbessern. Ein weiterer Unterschied zum erstgenannten Modell liegt darin, dass im ASUM-Prozess die Chronologie nicht entscheidend ist und jeder Prozess mehrmals durchlaufen wird. Das ASUM-DM-Modell (engl. Analytics Solutions Unified Method for Data Mining/Predictive Maintenance) beinhaltet hierbei fünf Phasen (siehe Abbildung 27):

1. Analyze
2. Design
3. Configure & Build
4. Deploy
5. Operate & Optimize

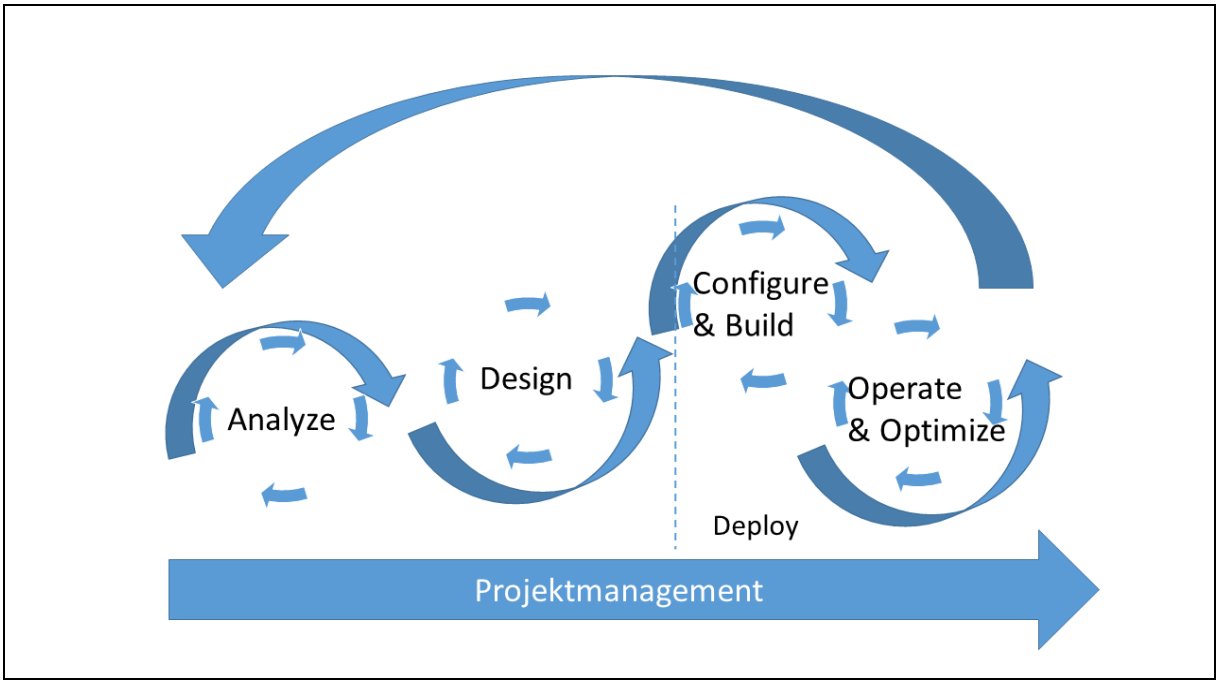


Abbildung 27: Analytics-Solutions-Unified-Methode (eigene Darstellung in Anlehnung an IBM Corporation)

Wie schon erwähnt, werden im ersten Analyseprozess zur Bestimmung von Anforderungen und Zielen alle am Projekt beteiligten Instanzen herangezogen. Danach folgt der zweite Prozessschritt Design, in dem Ressourcen und die Entwicklungsumgebung definiert werden. Der Prozess Configure & Build beinhaltet dann das schrittweise Umsetzen der Komponenten. Dabei wird auch eine Testphase implementiert, bevor die Integration (Deploy) in die Produktivumgebung genehmigt wird. Der letzte und damit fünfte Prozess des ASUM-DM-Modells ist Operate & Optimize. Hier geht es um die stetige Optimierung der Prozesse.

Die nähere Betrachtung des ersten Prozessschrittes, der Analyse, im ASUM-DM-Modell in einem Data-Science-Projekts gibt Auskunft über die Lebensdauer der Datenanalyse. Mit dem nun folgenden Modell „Lebenszyklus der Datenanalyse“ wird das zuvor beschriebene ASUM-DM-Modell vertieft.

Die Projekte, die sich mit der Datenanalyse befassen, unterscheiden sich von den traditionellen Projekten der Bildverarbeitung und der Datenbank-Managementsysteme. Bei den datenanalytischen Projekten ist mehr Forschung erforderlich. Ein Prozess oder ein Lebenszyklus, der die verschiedenen Aktivitäten des Projekts regelt, ist im Wesentlichen notwendig, soll aber den Prozess der Forschung nicht behindern. Anfänglich werden die Probleme, die groß erscheinen, weiter in kleinere Teile zerlegt, damit sie leichter angegangen

werden können. Wenn ein Prozess zeit- und phasengesteuert ist, dann hilft es, ihn für kleinere Teile des Problems leicht zu wiederholen. Daher muss der Lebenszyklus eines Datenanalyseprojekts sorgfältig geplant werden (vgl. Demchenko, Ngo & Membrey, 2013, S. 1–6). Der Umfang der Arbeit sollte klar definiert werden, wobei die Anforderungen zu verstehen und rechtzeitig zu berücksichtigen sind. In einigen der Szenarien kann es während der Datenanalysephase vorkommen, dass das zu einem früheren Zeitpunkt im Lebenszyklus definierte Ziel nicht geeignet ist und das Projekt neu begonnen oder abgebrochen werden muss. Daher ist ein gut definierter Prozess oder Lebenszyklus erforderlich, der die notwendige Flexibilität bietet, um Analysemethoden für verschiedene Teile des Projekts hinzuzufügen. In diesem Kapitel wird der Lebenszyklus der Datenanalyse mit fünf Schlüsselphasen vorgestellt, nämlich objektive Definition, Verständnis der Daten und Anforderungen, Datenbereinigung sowie Durchführung der Analyse und Visualisierung der Ergebnisse (vgl. Srinivasa, Siddesh & Srinidhi, 2018, S. 11). Die Phasen des Lebenszyklus eines Datenanalyseprojekts sind in der Abbildung 28 dargestellt. Die Phasen sind für die Probleme im Zusammenhang mit Big Data und IoT definiert. Die Phasen des Projekts werden iterativ definiert, da frühere Phasen des Projekts bei Bedarf noch einmal wiederholt werden können. Jede der Phasen des Lebenszyklus sowie seine Teilschritte werden in den folgenden Abschnitten vorgestellt.

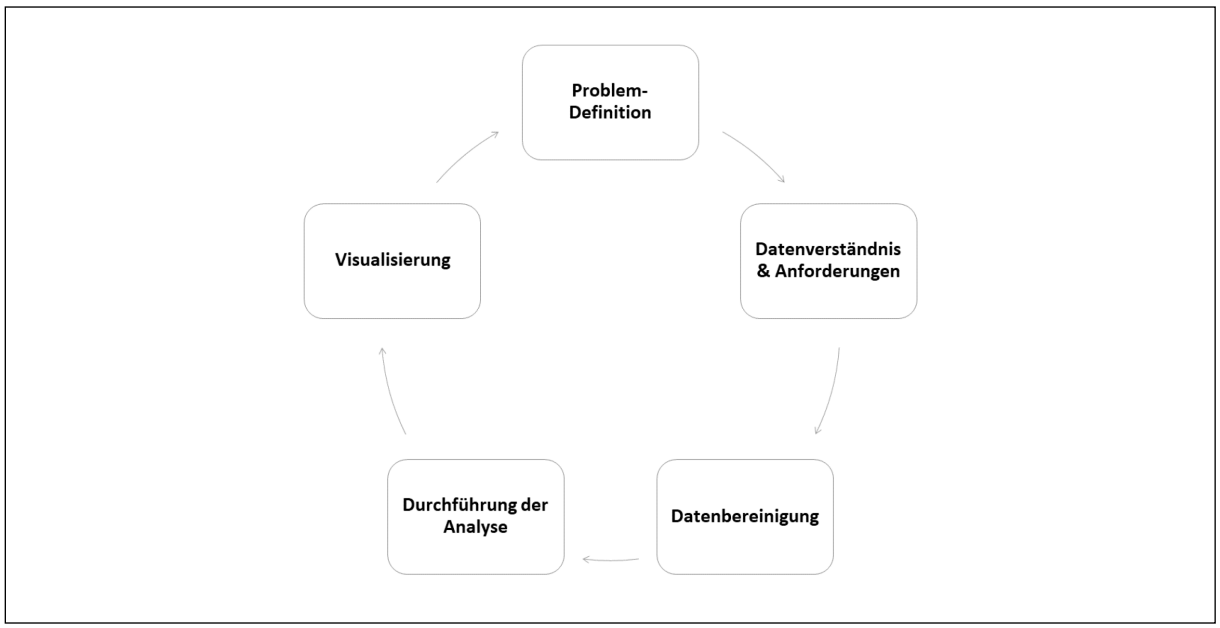


Abbildung 28: Lebenszyklus der Datenanalyse zum ASUM-DM-Modell (eigene Darstellung in Anlehnung an Srinivasa et al., 2018, S. 11–25)

In der ersten Phase ist die Definition eines Problems im jeweiligen Geschäftsbereiches für das Projektmanagement erforderlich. Die Probleme und Ziele müssen klar formuliert sein. Um ein Gleichgewicht im Team und daraus innovative Lösungen zu schaffen, ist die Zusammenarbeit

unterschiedlicher Experten notwendig (siehe Kapitel 3, Abschnitt 3.1). Neben der Problemdefinition sind in der ersten Phase u.a. auch die gegebenen Ressourcen zu betrachten. Hier wird klar angesprochen, welche Tools, Technologien, Mitarbeiter, Systeme und Daten zur Anwendung kommen. Die Vorüberlegung wird für die spätere Phase der Durchführung der Analyse noch von Bedeutung sein. Je eher eine Thematik angesprochen wird, desto langfristiger und strategischer kann in Bezug auf Know-how und Technologie geplant werden (vgl. Srinivasa et al., 2018, S. 13). In der ersten Phase wird zudem auch die Datenquelle identifiziert. Dabei gilt es, die geeigneten Systeme oder Quellen der zu analysierenden Daten, u.a. auch Rohdaten, zu definieren. Der abschließende Schritt in dieser Phase ist dann, das Ziel in der Theorie zu erarbeiten. Mit dieser Hypothese kann der Prozessschritt wiederum in Bezug auf Ressourcen verfeinert und optimiert werden.

Die zweite Phase beinhaltet das Datenverständnis und die Anforderungen bzw. auch Restriktionen zur Erreichung des in der ersten Phase definierten Ziels. Das Datenverständnis hängt zugleich mit dem Prozessverständnis zusammen. Hierbei wird die zweite Phase somit in Identifikation der Datenmerkmale und Identifikation der Datenstruktur unterteilt. Die Datenmerkmale müssen analysiert werden und ein Verständnis erzeugen, damit die dazugehörigen Prozesse optimiert werden. Je nach Anwendungsgebiet ist die Datenverfügbarkeit somit festgelegt. Die Bestimmung der Daten nach ihren Eigenschaften ist Basis für die nächstfolgende Anwendung wie bspw. Für das Machine Learning. Daraus folgt zugleich auch die Identifikation der Datenstruktur. Sie gibt Auskunft darüber, welches geeignete Tool zur Anwendung kommt. Daten können in ihrer Struktur als CSV-Dateien, Textdateien, Matrix, JSON-Dateien etc. vorliegen (vgl. Srinivasa et al., 2018, S. 16).

In der nächsten Phase werden die Daten bereinigt. Da sie in unterschiedlichen Formen und Ausprägungen vorhanden sind, gilt es in diesem Schritt, die Daten für die Analyse zu homogenisieren oder auch umzuwandeln. Die Datenaufbereitung ist ein essentieller Prozess in der Datenanalyse, um die zuvor ausgewählten Tools nutzen zu können. In dieser Phase ist hervorzuheben, dass ein ständiges Feedback bezüglich der Datenqualität erfolgen muss. Auf diese Weise werden Datenverluste minimiert oder auch ausgeschlossen.

Nachdem die Daten aufbereitet sind, folgt die Phase der Durchführung der Analyse. In dieser Phase werden Variable (Werte) identifiziert, die für das Modell notwendig sind. Ein Datensatz beinhaltet sehr viele Variable, jedoch sind nur einige von Nutzen. In Bezug auf Analysen zur Korrelation können dadurch Werte entfallen, die keinen Mehrwert oder neue Erkenntnisse mit sich bringen. Durch die Visualisierung der Daten kann dies herausgefunden werden. Sobald

dieser Schritt erfolgt, ist ein geeignetes analytisches Modell zur Datenanalyse erforderlich (vgl. Srinivasa et al., 2018, S. 22). Die Auswahl des Modells hängt mit dem vordefinierten Ziel aus der ersten Phase zusammen. In der Erstellung eines Modells wird der Datensatz in Trainingsdatensatz und Testdatensatz aufgeteilt. Damit wird erreicht, dass sowohl Training als auch Test in einem Modell ausgeführt werden.

Die Ergebnisse aus den analytischen Modellen können anhand von Darstellungen (Visualisierung) abgebildet werden. Damit ist die Visualisierung die letzte Phase in einem Daten-Analyse-Lebenszyklus. Die Darstellung der Ergebnisse erzeugt Klarheit und dient dem allgemeinen Verständnis über den gesamten Zyklus hinweg.

3.3 Business Intelligence

Schon lange bevor der Begriff „Big Data“ als Buzzword in der Wirtschaft aufkam, gab es technische Lösungen, mit denen Datenmengen analysiert werden konnten, jedoch mit Unterschieden zu den Möglichkeiten der heutigen Big-Data-Technologien. Diese technischen Lösungen sind Business Intelligence (BI) und Data Mining. Der große Unterschied zwischen BI und heutigen Big-Data-Technologien ist, dass BI nur bereits vorliegende Daten nutzen kann, wohingegen Big-Data-Technologien auch die Datengenerierung übernehmen können. Big-Data-Technologien können dabei auf Datenmengen aus verschiedensten Quellen zugreifen, was bei BI nicht möglich ist. Außerdem können Big-Data-Technologien mit deutlich größeren Datenmengen umgehen als BI. Ein weiterer Unterschied zu Big-Data-Technologien liegt darin, dass das Data Mining seine Daten schon vor der eigentlichen Nutzung aufbereitet und im zugehörigen Data Warehouse speichert. Die Big-Data-Technologien sind hingegen viel flexibler, sie können auf unterschiedliche Datenspeicher zugreifen und benötigen keine im Voraus aufbereiteten Daten, um mit ihnen umgehen zu können. Sie können vielmehr mit strukturierten und unstrukturierten Daten umgehen, wohingegen das Data Mining nur strukturierte Daten verarbeitet. Weitere Unterschiede und auch Gemeinsamkeiten von BI, Big-Data-Technologien und Data Mining sind in Abbildung 29 dargestellt (vgl. Freiknecht, 2014, S. 15–18).

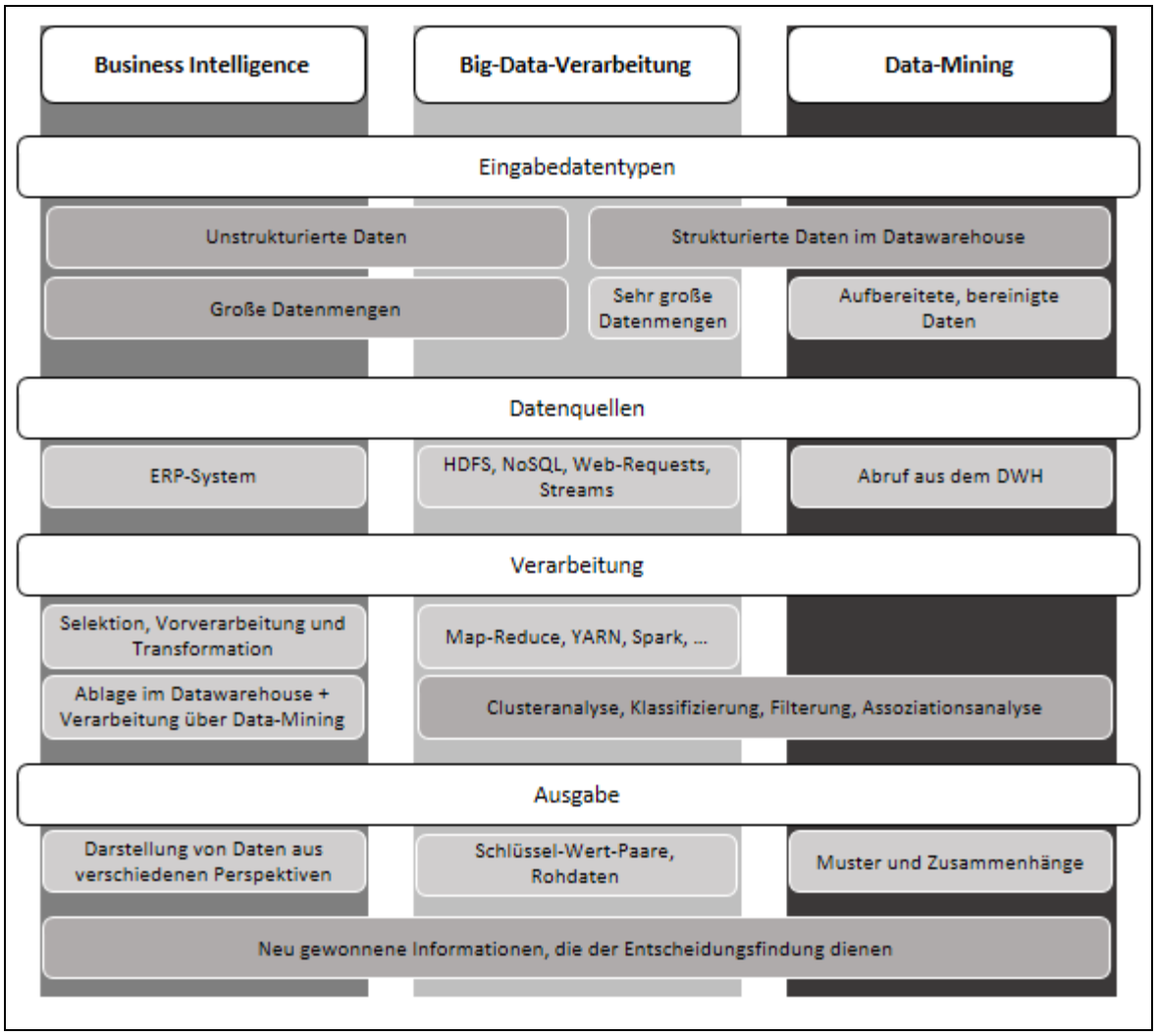


Abbildung 29: Definitionsvergleich von Business Intelligence, Big-Data-Verarbeitung und Data Mining (Freiknecht, 2014, S. 18)

Data Mining

Ein seit den 90er Jahren aufkommendes Forschungsfeld ist das Data Mining. Die Entwicklungen in dieser Richtung sind stark anwendungsorientiert und umfassten zunächst Mining-Herangehensweisen im akademischen Umfeld der Wirtschaftswissenschaften und Medizin. Mit der zunehmend kostengünstigen Verfügbarkeit von Rechenleistung einerseits und dem steigenden Angebot an anwenderorientierten Software-Lösungen andererseits finden Data-Mining-Prozesse heute praktisch überall dort Einsatzmöglichkeiten, wo Daten verarbeitet werden. Dabei bilden vor allem folgende verwandte Forschungsfelder die Grundlage für das Data Mining (vgl. Gorunescu, 2011, S. 1–2).

- Statistik:

Speziell die explorative Datenanalyse spielt eine wesentliche Rolle, um unbekannte Daten zu ordnen und zielgerichtet zu untersuchen. Expertenwissen über den Datenursprung ist dabei sekundär und es ist möglich, Zusammenhänge in Daten anhand mathematischer Merkmale zu identifizieren. Die Herausforderungen in der Applikation bestehen darin, die richtigen Methoden auszuwählen, sowie in der Ergebnisinterpretation. Zahlreiche Ansätze zur Visualisierung von Daten sind dabei verfügbar und können dabei helfen, diesen Herausforderungen zu begegnen.

- Künstliche Intelligenz:

Künstliche Intelligenz als Teilgebiet der Informatik greift auf heuristische Verfahren zurück, um Informationen zu verarbeiten. Durch den Aspekt der Heuristik grenzt sich dieses Feld von der Statistik ab. Gleichzeitig gibt es deutliche Überschneidungen mit dem Maschinellen Lernen. Dies ist essentieller Bestandteil von Data-Mining-Lösungen, deren Mehrwert von dem Automatisierungsgrad einer Methode bestimmt ist.

- Datenbanksysteme:

Als drittes Einflussfeld wirken Datenbanksysteme direkt auf die Leistungsfähigkeit des Data-Mining-Prozesses ein. Eine oder mehrere Datenbanken müssen hierbei so angelegt sein, dass die Daten mit den genannten Verfahren weiterverarbeitet werden können (vgl. Gorunescu, 2011, S. 3).

Unter der Zielsetzung, Konzepte für Monitoring-Methoden aufzuzeigen, werden dafür die Bereiche der Statistik und des Maschinellen Lernens (als Teil des Feldes Künstliche Intelligenz) im Zentrum der Analyse stehen. Data Mining ist ein Konzept, das sowohl für Big Data als auch Small Data angewandt werden kann.

Das allgemeine Lernmodell als Grundlage des Maschinellen Lernens

Maschinelles Lernen ist eine Unterdisziplin der Künstlichen Intelligenz. Der Begriff „Lernen“ kann hier nicht einheitlich definiert werden, da in unterschiedlichen Forschungsfeldern angepasste Definitionen üblich sind. Gemein ist ihnen aber, dass Lernen immer etwas mit Veränderung zu tun hat.

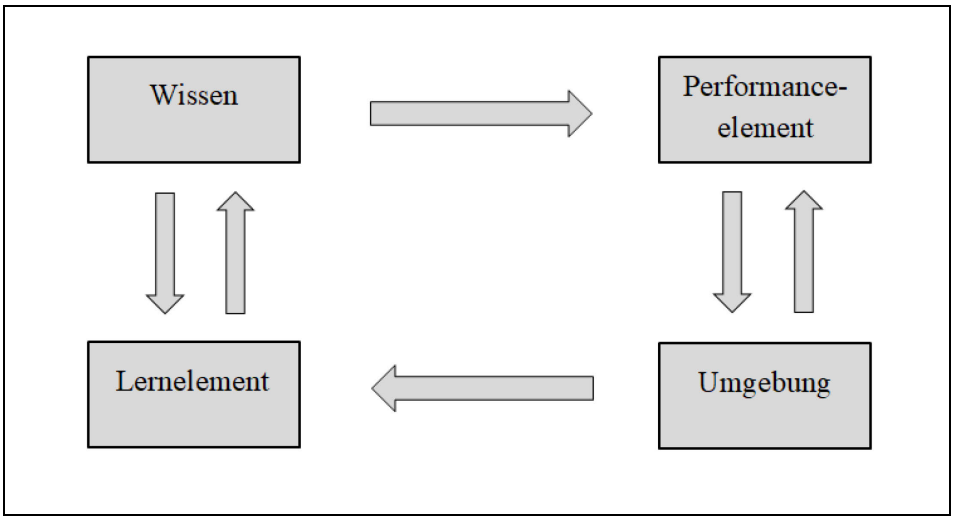


Abbildung 30: Allgemeiner Lernprozess (eigene Darstellung in Anlehnung an Beierle & Kern-Isberner, 2014, S. 100)

Lernen kann als ein adaptives Element betrachtet werden, das dafür sorgt, dass die gleiche Aufgabe in Zukunft besser gelöst wird als zuvor. Andere Auslegungen sehen das Lernen bereits in der bloßen Sammlung von Wissen. Weiterhin kann zwischen dem Lernen beim Menschen und dem Lernen von Maschinen unterschieden werden. Entsprechende computerbasierte Systeme sind bisher regelmäßig nicht in der Lage, aus eigenen Erfahrungen eine Leistungssteigerung für das eigene Handeln zu erreichen. Maschinelles Lernen soll nun genau dies ermöglichen (vgl. Beierle & Kern-Isberner, 2014, S. 98).

Abbildung 30 zeigt eine schematische Darstellung eines Lernmodells. In Lernsystemen wird zwischen den Elementen Lernen und Leistung bzw. Performance unterschieden. Während das Leistungselement (Performanceelement) auf der Grundlage des vorhandenen Wissens mit der Umgebung interagiert, berücksichtigt das Lernelement Veränderungen in der Umgebung. Infolge dieser Veränderungen wird neues Wissen generiert und vorhandenes Wissen angepasst. Darüber hinaus kann ein Kritikelement eingesetzt werden, um den Lernprozess zu bewerten und Auskunft über seine Wirksamkeit geben, während ein Problemgenerator neue Aufgaben erschafft, die zu neuen Erfahrungen führen (vgl. Beierle & Kern-Isberner, 2014, S. 98–99).

3.4 Big-Data-Ökosystem

Das Big-Data-Ökosystem ist durch die Open-Source-Logik weit verbreitet und bietet Anwendern unterschiedlichster Branchen Zugriff. Die Nutzungsvielfalt und die unterschiedlichen Anwendungsmöglichkeiten bezüglich der Besonderheiten und deren

Eigenschaften sind schwer in einer kurzen Beschreibung abbildbar. Aus dem Englischen übersetzt wird Big Data mit großen Datenmengen gleichgesetzt. Bei Big Data handelt es sich um ein sehr vielschichtiges Thema mit entsprechend vielen unterschiedlichen Technologien. Aus diesem Grund ist es schwer, eine eindeutige Definition auszumachen (vgl. Fasel & Meier, 2016, S. 3). In der Fachwelt lassen sich aus diesem Grund zahlreiche Begriffsklärungen finden.

So kann unter Big Data eine durch die Nutzung elektronischer Technologien anfallende Datenmenge verstanden werden, die sich durch ihre Größe, hohe Komplexität und Schnelllebigkeit auszeichnet (vgl. König, Schröder & Wiegand, 2018, S. 7). Ebenso knapp umschreibt MEIER Big Data. Er versteht darunter umfangreiche Datenbestände, welche mit herkömmlichen Softwarewerkzeugen kaum zu bewältigen sind. Die Datenbestände stammen aus diversen Quellen und sind meist unstrukturiert (vgl. Meier, 2018, S. 5).

FREYTAG hingegen beschreibt Big Data etwas umfassender. Ihm zufolge steigt seit Jahren die Tendenz, große Datenmengen zu generieren, zu speichern und zu analysieren. Die Treiber dieses Phänomens sind vor allem technologische Fortschritte wie die Reduzierung von Speicherkosten, die wachsende Vernetzung, hohe und gleichzeitig kostengünstige Übertragungsgeschwindigkeiten oder Ansätze zur Parallelbearbeitung von Datenmengen. Das Themengebiet Big Data ist somit sehr umfangreich. Es umfasst:

- das Nachverfolgen und Auswerten,
- das Suchen und Identifizieren,
- das Analysieren,
- das Vorhersagen und Planen,
- sowie Datenmanagement und -integration (vgl. Freytag, 2014, S.97–99).

Laut DORSCHEL muss Big Data aus verschiedenen Perspektiven beschrieben werden. In Bezug auf Business-Intelligence-Verfahren wird von Big-Data-Technologien eine neue Form der Analysen von Datenmengen hervorgebracht. Unterstützt wird diese Entwicklung dabei von der technologischen Weiterentwicklung von Speicher- und Prozesstechnologien. Datenmengen können dadurch im Terabyte-Bereich und höher ausgewertet werden. Völlig neu ist die Nutzung von großen Datenmengen zur Unterstützung des eigenen Geschäftsmodells nicht, denn Unternehmen wie Amazon oder Google verwenden schon seit längerer Zeit die Nutzungsdaten ihrer Kunden, um z.B. Voraussagen zu deren Verhalten zu treffen. Zum Schlagwort Big Data gehören auch hochleistungsfähige Datenbank-Managementsysteme. Diese stellen operative Arbeitsplattformen her. Sie arbeiten nach anderen Effizienz- und

Performancekriterien als konventionelle Datenbanken und können so massive Verarbeitungsgeschwindigkeiten hervorbringen. Dies birgt vor allem für Unternehmen mit einem Bedarf an On-Demand-Zugriffen auf große Datenmengen starke Vorteile. Weiterhin sind unter Big Data auch Analyseprozesse der Vergangenheit, Gegenwart und Zukunft zu verstehen. Dafür werden diverse Methoden, Verfahren und Werkzeuge genutzt. Herangezogen werden können diese Analyseprozesse z.B. für die Optimierung industrieller Prozesse, in der Spieleentwicklung, bei sozialen Netzwerken oder auch für nachrichtendienstliche Zwecke (vgl. Dorschel, 2015, S. 1–4).

Die vorgenannten Ansichten verdeutlichen, dass Big Data viele Facetten hat und die gängigen Schlagworte oft synonym zueinander verwendet werden. Wenn in dieser Arbeit von Big-Data-Technologien gesprochen wird, dann sind damit alle Verfahren, Werkzeuge, Hardware- und Software-Komponenten gemeint, die notwendig sind, um die Nutzung von großen Datenmengen zu realisieren.

Die 5 Vs von Big Data

Wenn in der Literatur über Big-Data-Technologien gesprochen wird, werden meist die fünf Vs miteinbezogen. Diese sollen die Eigenschaften von Big Data widerspiegeln. Nachfolgend werden diese fünf Eigenschaften kurz vorgestellt.

Volume

Diese Eigenschaft entspricht ins Deutsche übersetzt der Menge an Daten, also dem Datenbestand. Dieser ist in der Regel sehr umfangreich und liegt im Tera- bis Zettabyte-Bereich (vgl. Meier, 2018, S. 6). Anlass für diese hohen Mengen sind die sich ständig weiterentwickelnden IT-Systeme. Computerchips werden immer kleiner und gleichzeitig auch leistungsfähiger. Dies ermöglicht ihre Integration in immer mehr Lebensbereiche, in denen sie Steuerungsaufgaben ausführen und dabei Daten erzeugen. Gleichzeitig steigt auch das Niveau der Vernetzung in den verschiedenen Lebensbereichen (vgl. Dorschel, 2015b, S. 7). Das Datenvolumen auf der Welt steigt somit relativ mit dem exponentiellen Wachstum der Rechenkapazitäten. Studien kamen zu dem Ergebnis, dass sich das weltweite Datenvolumen alle zwei Jahre verdoppelt (vgl. Rudolph & Linzmajer, 2014, S. 13).

Die Volume-Eigenschaft stellt folgende Anforderungen an Big-Data-Lösungen (vgl. Lanquillon & Mallow, 2015b, S. 264):

- Speicherung und Verarbeitung von sehr großen Datenmengen muss realisierbar sein,

- gute Skalierfähigkeit und Fehlertoleranz beim Ausfall von Komponenten (Robustheit),
- es muss mit kleinen Datenpaketen wie auch riesigen Datenmengen umgegangen werden können.

Variety

Dieser Begriff steht für die Vielfalt an Datenquellen und Datenformaten. Die Daten können strukturiert, semistrukturiert und unstrukturiert vorliegen (vgl. Meier, 2018, S. 6). IT-Systeme speichern Daten in verschiedenen Lebensbereichen zu verschiedensten Zwecken. Dadurch entsteht ein breites Spektrum an Dateninhalten und Datenformaten (vgl. Dorschel, 2015, S. 8).

Die Quellen, aus denen Daten bezogen werden, sind vielfältig. Daten können zentral in Datenbanksystemen abgelegt sein. Sie sind aber auch dezentral an anderen Speicherorten und in verschiedenen Dateiformaten vorzufinden. Zu potenziellen Datenquellen gehören u.a. (vgl. Freytag, 2014, S. 99):

- Produktions- bzw. Prozessdaten,
- Webdaten,
- Wissenschaftsdaten,
- Geschäftsdaten,
- Sensordaten,
- Daten von mobilen Endgeräten,
- Daten aus sozialen Netzwerken/Blogs/Foren.

Innerhalb dieser Datenquellen befinden sich Daten in verschiedenen Formen. So können sie bspw. als Texte, Graphiken, Bilder, Audios oder Videos vorliegen. Diese liegen wiederum in verschiedenen Formen vor. Beispiele sind in der Abbildung 31 dargestellt.

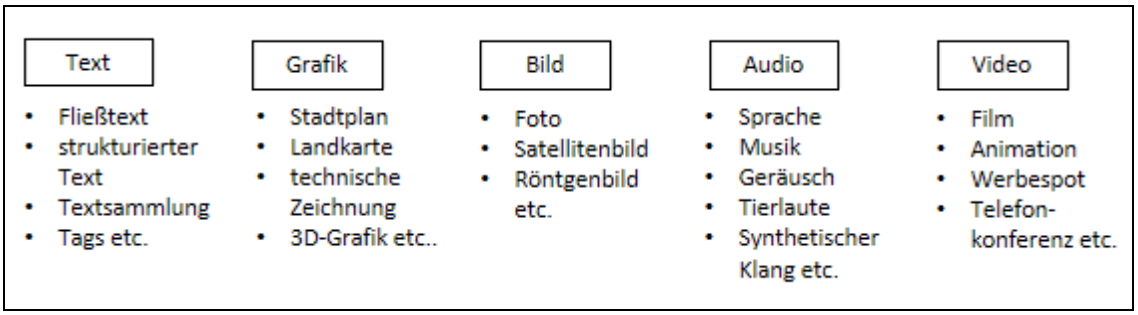


Abbildung 31: Formen von Daten (eigene Darstellung in Anlehnung an Meier & Kaufmann, 2016)

In Bezug auf Variety müssen Big-Data-Lösungen folgenden Anforderungen gerecht werden (vgl. Lanquillon & Mallow, 2015b, S. 264):

- es müssen beliebige Datenquellen und Formate beherrscht werden,
- mit Schemalosigkeit muss umgegangen werden können, also müssen Daten, deren Struktur nicht bekannt ist, geladen werden können,
- wenn Daten in Datenquellen häufig geändert werden, soll dies beherrschbar sein (Agilität).

Velocity

Das Wort „Velocity“ steht für die Geschwindigkeit und spielt darauf an, dass Datenströme in Echtzeit analysiert werden können (vgl. Meier, 2018, S. 6). Manche Autoren verstehen darunter auch die Geschwindigkeit, in der neue Daten produziert werden. Zwischen den Eigenschaften Volume und Velocity besteht eine Wechselbeziehung. Mit steigender Geschwindigkeit der Systeme wachsen die Produktion von Daten und damit das Volumen an Daten in immer kürzerer Dauer. Velocity steht auch für eine kurze Halbwertszeit bezüglich der Erkenntnisse, die aus Daten gezogen werden. Begründet wird dies damit, dass sich Themen, die durch digitale Daten beschrieben werden, zunehmend schneller verändern. Aus diesem Grund müssen in den existierenden Datenbeständen häufig neue bzw. veränderte Daten ergänzt werden (vgl. Dorschel, 2015, S. 7–8).

Aufgrund der Velocity werden folgende Anforderungen an Big-Data-Lösungen gestellt (vgl. Lanquillon & Mallow, 2015b, S. 264):

- Datenströme in großen Dimensionen müssen verarbeitet werden,
- eine Echtzeitreaktion auf Ereignisse bzw. Ereigniskonstellationen innerhalb von Datenströmen soll gewährleistet sein.

Veracity

Diese Eigenschaft steht für die Aussagekraft bzw. Richtigkeit der Daten. Viele Daten sind ungenau und müssen daher bewertet werden, denn höhere Datenbestände allein garantieren keine bessere Aussagekraft. Für die Bewertung werden spezifische Algorithmen verwendet (vgl. Meier, 2018, S. 7). Big-Data-Technologien beziehen häufig auch Daten ein, deren objektiver Erkenntniswert nicht eindeutig messbar ist. Vor allem von Menschen verfasste Daten enthalten subjektive Wahrnehmungen sowie verschiedene inhaltliche und zeitliche Kontexte. Daher gehört es auch zu Big Data, diese Störfaktoren herauszufiltern, um verlässliche Daten zu erhalten (vgl. Dorschel, 2015, S. 8).

Die Veracity führt damit zu folgenden Anforderungen an Big-Data-Lösungen (vgl. Lanquillon & Mallow, 2015b, S. 264):

- es wird eine Systemkomponente benötigt, welche die Datenqualität misst und eine Datenbereinigung durchführen kann,
- Daten müssen in ihrem ursprünglichen Format verfügbar und inhaltlich nicht verändert worden sein,
- eine Integration von Metadaten ist erforderlich.

Value

Der Begriff „Value“ ist mit dem Informationskapital bzw. dem Vermögenswert von Daten gleichzusetzen. Die Anwendung von Big-Data-Technologien soll den Unternehmenswert steigern. Daher werden insbesondere dort Investitionen in Personal und technische Infrastruktur getätigt, wo eine Hebelwirkung für den Unternehmenserfolg erwartet wird (vgl. Meier, 2018, S.7).

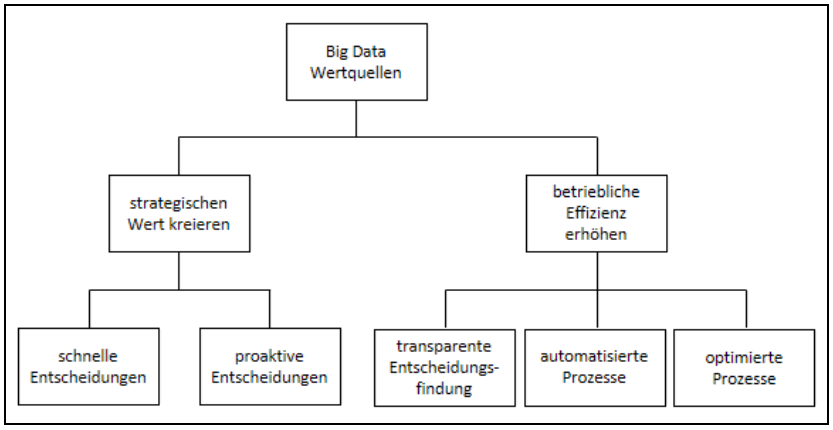


Abbildung 32: Betriebliche Wertquellen von Big Data (in Anlehnung an Omri, 2015, S. 104)

Wie in der Abbildung 32 dargestellt, kann der Mehrwert der Datennutzung für ein Unternehmen in verschiedenen Formen auftreten. Der Wert von Daten kann bspw. strategischer Natur sein. Die Datennutzung kann dazu beitragen, Entscheidungen schneller zu treffen, und beeinflusst auch proaktiv die Entscheidungsfindung. Außerdem kann der Mehrwert von Daten darin bestehen, die betriebliche Effizienz zu erhöhen. So können Daten Entscheidungsfindungen transparenter machen, Prozesse automatisieren und optimieren (vgl. Omri, 2015, S. 104).

3.5 Interdisziplinäre Verschmelzung

In diesem Abschnitt werden die Voraussetzungen für ein Big-Data-Projekt erläutert. Im Mittelpunkt steht das Interdisziplinäre Team und dessen Einflussfaktoren. Daher wird zunächst kurz auf die Technologie eingegangen, die in den vorherigen Abschnitten erwähnt wurde, um den Überblick zu festigen. Danach folgt eine Prozessanalyse eines Big-Data-Projekts, um dann auf die menschliche Komponente einzugehen.

Die oben genannten Einschränkungen von Business Intelligence und Data Mining sind hinsichtlich der Daten und der Prozesse bekannt. Eine Erweiterung des Analysesystems durch Big-Data-Werkzeuge bringt jedoch nicht nur ausschließlich Analysevorteile. Technisch betrachtet können dadurch auch BI und DM entscheidend genutzt werden. Bezüglich der Tools und der Vorgehensweise mit Machine-Learning-Algorithmen zur Datenbeschaffung oder auch der erweiterten Anwendung durch Künstliche Intelligenz sind vorhandene Analysen durch BI oder DM und das dazugehörige Data Warehouse eine potente Datenquelle.

Auf strategischer Unternehmensebene sind Big-Data-Projekte interdisziplinär zu verorten (vgl. Gao, Koronios & Selle, 2015, S. 1). Die Big-Data-Projekte sind Innovationsprojekte, in denen die Datenanalyse die Grundlage für die Durchführung neuer Ansätze und Optimierung vorhandener Prozesse bildet. Ein Big-Data-Team muss flexibel auf Veränderungen im Unternehmen und durch die Analyse aufgedeckte Probleme eingehen können (vgl. Sicular, 2012, S. 10). Eine Herausforderung dabei ist der Umgang mit ständig wachsenden großen Datenmengen, denn hier müssen der relative Nutzen dieser Daten und die daraus möglichen Schlussfolgerungen identifiziert. Entscheidend ist die Wichtigkeit dieser Daten für die Aufbewahrung. Dabei sollen diejenigen Daten aufbewahrt werden, die einen potenziellen Geschäftswert widerspiegeln (vgl. SAS, 2013). Im Zuge dessen spielt die Datenqualität eine zentrale Rolle, so werden für die Unternehmen entscheidende Prozesse angestoßen. Der mit der Big-Data-Analyse ausgeführte Prozess wirkt sich auf das gesamte Unternehmen aus. Daraus

folgt eine weitere Herausforderung für das Big-Data-Team, nämlich, die Qualität der unternehmensinternen, aber auch unternehmensexternen Daten für die Analyse zu gewährleisten (vgl. Sathi, 2012, S. 4–5). Aufgrund der innovativen Prozesse während eines Big-Data-Projekts müssen die dazugehörigen Protagonisten eine innovative Denkweise mitbringen, damit kreative Lösungsansätze verfolgt werden können (vgl. Gao et al., 2015, S. 3). Dennoch ist eine Eingrenzung wichtig, um das Projekt- und das damit verbundene Unternehmensziel nicht aus dem Fokus geraten zu lassen. Innerhalb der Zielvorgabe sollen jedoch der Freiheit und Kreativität keine Grenzen gesetzt werden (vgl. Sicular, 2012, S. 29). Dabei unterstützt das Big-Data-Projekt ein klar definiertes Unternehmensziel. Für den Erfolg oder auch Misserfolg müssen außerdem Analysen bezüglich der Auswirkungen betrieben werden. Diese Analyse kann auf Grundlage einer Kosten-Nutzen-Analyse oder anhand statistischer Analysen erfolgen (vgl. Gopalkrishan, Steier, Lewis & Guszcza, 2012). Eine weitere Herausforderung ist auch die Erwartungshaltung des Unternehmens. Die Sicherstellung der Funktionsfähigkeit und Ausbaufähigkeit der Lösungen sollen das Unternehmen voranbringen (vgl. Sicular, 2012, S. 26).

Ein zentraler Bestandteil eines Big-Data-Teams ist die oder der Datenwissenschaftler/in (vgl. Davenport & Patil, 2012, S. 7). Sie sind die Schnittstelle im Unternehmen (siehe Abschnitt 3.1.), die dazu fähig ist, Einzeldisziplinen in einem Unternehmen miteinander zu kombinieren, um dadurch innovative Lösungen zu erzeugen. Datenwissenschaftler (engl. Data Scientists) erwerben Fach- und Branchenkenntnisse, die dann in die Analysen hineinfließen. Sie entwickeln ein detailliertes Verständnis der spezifischen Geschäftsprozesse innerhalb des Funktionsbereiches (vgl. Giannikas, 2011, S. 9). Neben den „Hauptakteuren“ sind Experten der jeweils beteiligten Fachbereiche notwendig. Sie sollen mit in den Prozess des Big-Data-Projekts eingebunden werden. Dadurch werden Unternehmensziele vorgegeben und die Restriktionen definiert. Durch die Aufstellung eines interdisziplinären Teams wird zu Beginn eines Projekts der Fokus auf Unternehmenswert und Unternehmensvorstellung in die Ideenfindung eingeschlossen. Daraus kann der maximale Erfolg für die Organisation und das gesamte Unternehmen generiert werden (vgl. Sicular, 2012, S. 11). Ein Big-Data-Team erfordert das Zusammenwirken von Akteuren unterschiedlicher Unternehmensbereiche. Bezogen auf Unternehmensziele bringt diese Diversität großes Innovation- und Erfolgspotenzial mit sich. CLUTTERBUCK definiert das Resultat der Diversität als „[…] means of overcoming injustice – righting wrongs – and at the other as a means of enhancing individual and group contribution to the organization’s goals“ (vgl. Clutterbuck & Ragins,

2002, S. 55). Aus diesem Grund stellt sich ein interdisziplinäres Team auch der Herausforderung von Sicherheits- und Ethikfragen, die mit dem Big-Data-Projekt einhergehen. Dafür ist das Einbeziehen des Risikomanagements und der Rechtsexperten in frühen Projektstadien zu empfehlen.

4. Big-Data-Lösungskonzeption – systematischer Aufbau

4.1 Verständnis von Big Data im industriellen Kontext

Der Begriff „Big Data“ ist in der heutigen Wirtschaftswelt allgegenwärtig. Trotzdem kann aufgrund seiner Vielschichtigkeit, wie bereits erläutert, keine eindeutige Definition dafür festgehalten werden (Fasel & Meier, 2016, S. 3). Stattdessen existieren viele verschiedene Ansätze. Big-Data-Lösungskonzeption – systematischer Aufbau

So kann wie im Kapitel zuvor Big Data als eine durch die Nutzung elektronischer Technologien anfallende Datenmenge verstanden werden, die sich durch ihre Größe, Komplexität und Schnelllebigkeit auszeichnet (vgl. König, Schröder & Wiegand 2018, S. 7) Des Weiteren versteht MEIER darunter umfangreiche Datenbestände, welche mit herkömmlichen Software-Werkzeugen kaum zu bewältigen sind. Die Datenbestände stammen aus diversen Quellen und sind meist unstrukturiert (vgl. Meier, 2018, S. 5).

Eine umfassendere Beschreibung geht von FREYTAG aus. Die Generierung großer Datenmengen sowie das Speichern und Analysieren stehen im Mittelpunkt. Technologische Fortschritte wie die Reduzierung von Speicherkosten, die wachsende Vernetzung, hohe und gleichzeitig kostengünstige Übertragungsgeschwindigkeiten oder Ansätze zur Parallelbearbeitung von Datenmengen sind Treiber von Big Data. Zusammenfassend beschreibt Freytag das Themengebiet Big Data mit dem Nachverfolgen und Auswerten von Daten. Darunter fallen auch die Begrifflichkeiten des Suchens und Identifizierens, des Analysierens, der Vorhersagen und der Planung sowie das Datenmanagement und dessen Integration (vgl. Freytag, 2014, S. 97–99).

In der Industrie muss Big Data aus verschiedenen Perspektiven beschrieben werden. Neben Business-Intelligence-Verfahren bringen Big-Data-Technologien neue Formen der Analysen von großen Datenmengen mit. Die Entwicklung der Automatisierung in der Produktion wird dabei von der technologischen Weiterentwicklung von Speicher- und Prozesstechnologien unterstützt. Mit Big-Data-Analysen werden die erzeugten Datenmengen im Tera-Bereich und höher ausgewertet. Wie schon erwähnt, ist die Nutzung von großen Datenmengen zur Unterstützung des eigenen Geschäftsmodells nicht neu. Tech-Unternehmen wie Google oder der Online-Handel mit Amazon treffen ihre Voraussagen mithilfe von Kundendaten. Dabei spielen hochleistungsfähige Datenbank-Managementsysteme im Big-Data-Ökosystem eine wesentliche Rolle. Sie stellen operative Arbeitsplattformen her. Die Effizienz- und

Performancekriterien sind leistungsfähiger als konventionelle Datenbanken und können so massive Verarbeitungsgeschwindigkeiten hervorbringen. Die Vorteile sind einerseits On-Demand-Zugriffe, andererseits lösen sie das Problem des Schichtenmodells. Im Big-Data-Analyseprozess werden historische und aktuelle Betriebsdaten wie bspw. Maschinen-, Prozess- und extern erhobene Daten konsolidiert, um einen Zusammenhang zwischen ihnen darzustellen und diese zu verstehen und zu nutzen. Diverse Methoden, Verfahren und Werkzeuge werden herangezogen und dafür eingesetzt. Anwendungsszenarien der Analyseprozesse finden in der Optimierung industrieller Prozesse, in der Spieleentwicklung, bei sozialen Netzwerken statt. Sie stellen auch Nutzen für nachrichtendienstliche Zwecke dar (vgl. Dorschel, 2015a, S. 1–4).

Die vorgenannten Meinungen verdeutlichen, dass Big Data viele Facetten hat und die gängigen Schlagworte oft synonym zueinander verwendet werden. Wenn in dieser Arbeit von Big-Data-Technologien gesprochen wird, dann sind damit alle Verfahren, Werkzeuge, Hardware- und Softwarekomponenten gemeint, die notwendig sind, um die Nutzung von großen Datenmengen zu realisieren.

Die in der Arbeit verwendeten Big-Data-Technologien werden anhand einer Nutzwertanalyse abgeleitet. Dabei ging der Autor in der Bewertung auf die in der Literatur vorhandenen Merkmale ein. Ebenfalls wurden Aspekte zur Anwendung in einem Werk berücksichtigt. Somit wurden sowohl die Merkmale als auch die Bewertungsskala durch das Zusammenführen von theoretischen und praxisnahen Anforderungen erstellt. Die Bewertungsskalierung erfolgt mithilfe eines Punktesystems von 1 bis 5. Die höchste Bewertung wird bei 5 Punkten je Merkmal erreicht. Daraufhin werden die Tools mit der höchsten Gesamtpunktzahl in die Architektur (Abschnitt 4.2) integriert. Ausschnitte aus der Nutzwertanalyse werden in der Abbildung 33 und Abbildung 34 dargestellt. Die Auswahl der geeigneten Werkzeuge für die Big-Data-Architektur beläuft sich zunächst auf die theoretischen Merkmale, die jedes Tool einzeln vorweist. Demnach ist die erste Bewertung für die Datenbeschaffung vorzunehmen. In der Analyse wurden die Tools Apache Kafka, Apache Nifi, Apache Flume, Apache Storm und Apache Hadoop im direkten Vergleich analysiert. Als Eigenschaften sind bspw. Verarbeitungsgeschwindigkeit, Schnittstellen und Skalierbarkeit aufgelistet. Weitere Eigenschaften wurden auch herangezogen, um eine Entscheidung bezüglich der Auswahl treffen zu können. Unter den Punkten Datenintegrität und Support versteht der Autor die Nutzbarkeit der Tools unter dem Gesichtspunkt der Praxisanwendbarkeit im Werk.

Nutzwertanalyse Big Data Tools											
Skala 1-5	1=gering, 5=hoch										
		Data Ingestion									
Attribute	**Gewichtung [%]**	**Kafka**	**afka gewic**	**Nifi**	**Nifi gewich**	**Flume**	**ume gewic**	**Storm**	**orm gewic**	**Hadoop**	**adoop gewich**
Geschwindigkeit	**15**	**5**	**75**	**5**	**75**	**4**	**60**	**5**	**75**	**3**	**45**
Datenintegrität	15	5	75	5	75	2	30	3	45	5	75
Schnittstelle (API)	**10**	**5**	**50**	**4**	**40**	**1**	**10**	**3**	**30**	**5**	**50**
Streaming	10	4	40	1	10	4	40	5	50	4	40
Batchverarbeitung	**10**	**2**	**20**	**1**	**10**	**1**	**10**	**1**	**10**	**5**	**50**
Replikation	10	5	50	1	10	1	10	1	10	5	50
Skalierbarkeit	**15**	**5**	**75**	**5**	**75**	**5**	**75**	**3**	**45**	**5**	**75**
Unabhängigkeit	5	5	25	5	25	1	5	3	15	5	25
Support (lib/prjct)	**10**	**5**	**50**	**5**	**50**	**4**	**40**	**5**	**50**	**5**	**50**
	100	41	**460**	32	**370**	23	280	29	330	42	**460**

Abbildung 33: Nutzwertanalyse Data-Ingestion-Tools (eigene Darstellung)

Nachdem die Auswahl in der Datenbeschaffung eingegrenzt wurde, befasst sich der Autor mit der Analyse der Tools im Processing Layer und im Serving Layer (siehe Abbildung 34). In diesem Kontext ist hervorzuheben, dass der eigentliche Analyseprozess möglichst auf ein Analysewerkzeug begrenzt wird, um die Übersichtlichkeit und eine effiziente Bearbeitung der Daten zu gewährleisten. Deshalb wurden aus drei Tools (Apache Spark, Apache Flink, Apache Storm) ein geeignetes Tool für die Anwendung ausgewählt. Die Auswahl fiel auf Apache Spark aufgrund seiner enormen Funktionalität und seines Potenzials, möglichst alle Datenformen analysieren zu können. Ein weiterer Faktor ist simultane Verarbeitung im Speed- und im Batch-Layer, ausgehend aus der Lambda-Architektur. Als letzte Instanz in der vom Autor vorgestellten Big-Data-Architektur (Abschnitt 4.2) wurden NoSQL-Datenbanken miteinander verglichen, um geeignete Speichermedien im Serving Layer anzubieten. Dabei wurden die NoSQL-Datenbank-Managementsysteme wie Cassandra, HBase, Hive, MongoDB und Neo4j ausgewählt und selektiert. Das Ergebnis der Nutzwertanalyse bildet hierbei die Basis für die im Folgenden vorgestellte Big-Data-Architektur.

	Processing Layer			Serving Layer (noSQL)				
Attribute	**Spark**	**Flink**	**Storm**	**Cassandra**	**Hbase**	**Hive***	**MongoDB**	**Neo4j**
Geschwindigkeit	**75**	**75**	**75**	**60**	**75**	**60**	**30**	**60**
Datenintegrität	75	30	30	75	75	60	30	30
Schnittstelle (API)	**50**	**40**	**40**	**40**	**40**	**20**	**20**	**20**
Streaming	50	50	50	40	40	20	20	20
Batchverarbeitung	**50**	**10**	**20**	**50**	**50**	**50**	**50**	**50**
Replikation	10	20	10	50	50	40	50	40
Skalierbarkeit	**75**	**60**	**60**	**75**	**75**	**60**	**60**	**60**
Unabhängigkeit	25	25	15	25	25	5	10	5
Support (lib/prjct)	**50**	**50**	**50**	**40**	**50**	**40**	**50**	**20**
	460	360	350	**455**	**480**	**355**	320	305

Abbildung 34: Nutzwertanalyse Processing Layer & Serving Layer (eigene Darstellung)

4.2 Architektur

Die in diesem Kapitel untersuchte Architektur wird anhand von vier Kategorien betrachtet:

- Die Datenquelle als Herkunft der Daten und somit der dazugehörigen Datenbeschaffung,
- die Datenverarbeitung innerhalb des Clusters, welches in dieser Arbeit das Hadoop-Cluster ist,
- die Datenanalyse,
- und die Visualisierung, woraus die Interpretation und die Entscheidungsfindung resultieren.

Dieser Abschnitt untersucht auch in zwei Ansätzen das Datenmanagement. Als klassische IT-Architektur ist die IT-Infrastruktur mit dem Data Warehouse zu nennen. Darauf basierend betrachtet die Wissenschaft die Analysemöglichkeiten in den zuvor genannten Kategorien der Business Intelligence und des Data Mining. Für eine einheitliche und darüber hinaus echtzeitliche Betrachtung der Daten ist der Ansatz der Big-Data-Analyse von Bedeutung. Dahingehend wird im zweiten Ansatz die Big-Data-Architektur mit der Big-Data-Station durchleuchtet.

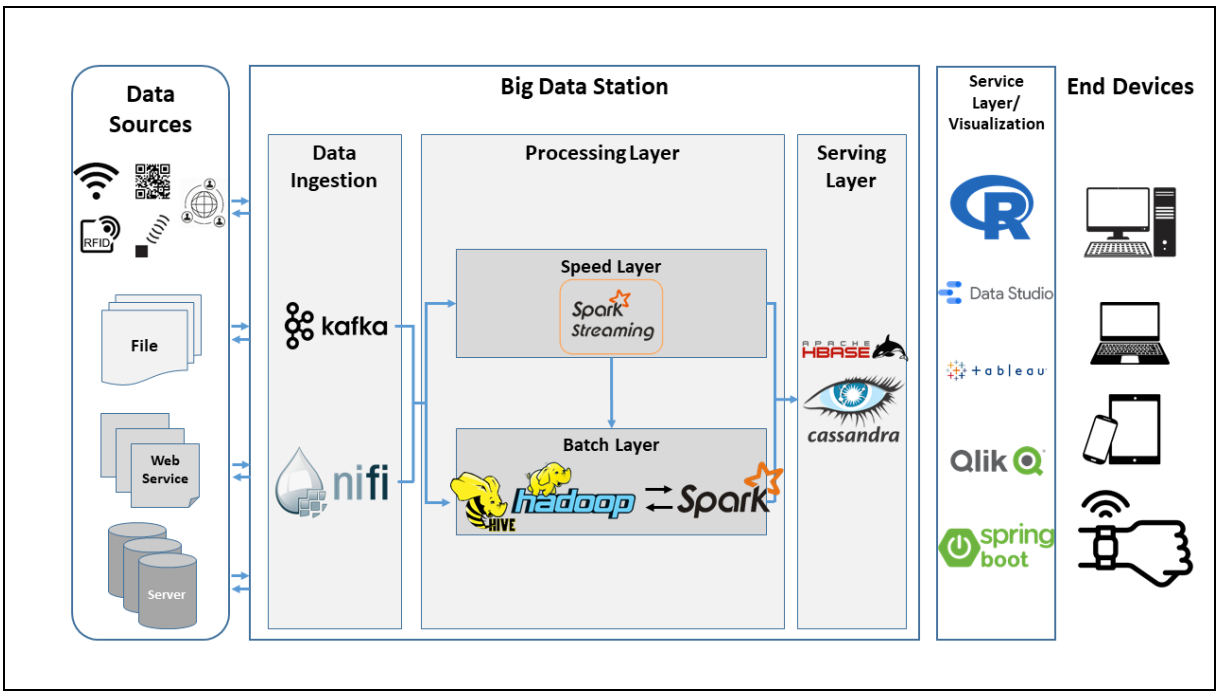


Abbildung 35: Big-Data-Referenzarchitektur (eigene Darstellung)

In der Referenzarchitektur (siehe Abbildung 35) sind die vier Schichten abgebildet. Die erste Schicht ist die Datenquelle. Sie legt die Quelle fest, an der das Big-Data-Cluster angebunden ist, um somit die nächsten Prozessschritte einleiten zu können. Im Cluster selbst befinden sich Data Ingestion, Processing Layer und Serving Layer. Der Serving Layer ist für die bildliche Darstellung, also für die Visualisierung für den Anwender, zuständig. Im Folgenden betrachten wir zunächst das Konzept des klassischen Data Warehouse, um dann auf das Konzept des Data Lakes zu erläutern. In diesem Kapitel werden dann auch die Prozesse des Big-Data-Clusters näher erklärt. Diese Architektur erlaubt dem Unternehmen auch direkte Analyseprozesse von Echtzeitdaten im Sinne der Streaming-Analyse (Abschnitt 4.4). Eine weitere Analysemethode wird außerdem durch Machine Learning erreicht. Dies ist ein zentraler Bestandteil der Arbeit, in dem die Ergebnisse vorgestellt werden. Im Abschnitt 4.3 befassen wir uns mit dem Datenmanagement. Darauf aufbauend wird die Analyse und Darstellung im Kapitel 5 behandelt. Als Abschluss des Kapitels 4 wird auf die Integration eingegangen, welche die Voraussetzung für die Durchführung der in der Arbeit vorgestellten Fallstudie (Kapitel 5) abbildet.

4.2.1 Data-Warehouse-Architektur

Eine der ersten und in der Literatur wohl am häufigsten verwendete und zitierte Definition des Begriffs „Data Warehouse“, kurz DWH, stammt von William H. INMON. Er definiert Data Warehouse wie folgt:

„A data warehouse is a subject oriented, integrated, non-volatile, and time varying collection of data in support of management's decisions.” (Inmon, 1996, S. 33)

Ausgehend von dieser Definition lassen sich folgende vier fundamentalen Merkmale beschreiben:

- eine Themenorientierung (englisch: subject orientation)
- eine integrierte Datengrundlage (englisch: integration)
- eine unveränderliche Datengrundlage (englisch: non-volatile)
- eine Zeitbezogenheit (englisch: time-related)

Data Warehouse beinhaltet also ausschließlich Daten, die für die Unterstützung von Entscheidungsprozessen relevant sind. Diese genannten Daten, welche aus heterogenen Systemen stammen, werden zuerst in eine einheitliche und bereinigte Struktur gebracht. Anschließend werden die aufbereiteten Daten in einer integrierten Datengrundlage in Form

eines Data Warehouse gespeichert. Diese Daten sind nach dem Import in das Data Warehouse unveränderlich. In der Regel werden sie also nicht mehr grundlegend verändert oder aus dem Datenbestand gelöscht. Somit werden eine konsistente Datengrundlage und die Reproduzierbarkeit der Analyseergebnisse gewährleistet. Zudem ermöglicht eine Zeitbezogenheit wie Zeitstempel oder Historiendaten eine Analyse von Veränderungen und Entwicklungen im Laufe der Zeit. Es besteht daher die Notwendigkeit, Daten über einen längeren Zeitraum zu speichern (vgl. Inmon, 2005, S. 32 ff.).

Demnach ist ein Data Warehouse William H. INMON zufolge eine Sammlung von Daten mit spezifischen Eigenschaften zur Unterstützung von Managemententscheidungen (vgl. Inmon, 2005, S. 1). Jedoch ist diese Definition zum einen nicht aussagekräftig genug, um sie in der Praxis zu nutzen, und zum anderen ist sie so restriktiv, dass zahlreiche Anwendungsbereiche und Vorgehensweisen ausgeblendet werden. Im Laufe der Zeit wurden stetig Anpassungen an der Definition von Data Warehouse vorgenommen. So haben BAUER und GÜNZEL die Definition folgendermaßen erweitert:

„Ein Data Warehouse ist eine physische Datenbank, die eine integrierte Sicht auf beliebige Daten zu Analysezwecken ermöglicht.“ (Bauer & Günzel, 2013, S. 4)

Dementsprechend ist ein Data Warehouse ein Speicher für Daten, die von den verschiedenen Betriebssystemen eines Unternehmens generiert und gesammelt werden. Data Warehousing ist oft Teil einer umfassenderen Datenmanagementstrategie und legt den Schwerpunkt auf die Erfassung von Daten aus verschiedenen Quellen, um den Zugriff und die Analyse durch Geschäftsanalysten, Datenwissenschaftler und andere Endnutzer zu ermöglichen. Typischerweise ist ein Data Warehouse eine relationale Datenbank, die auf einem Großrechner, einer anderen Art von Unternehmensserver oder zunehmend in der Cloud untergebracht ist (vgl. Freiknecht & Papp, 2018, S. 18). Die nachfolgende Abbildung 36 veranschaulicht den Workflow in einem Data Warehouse.

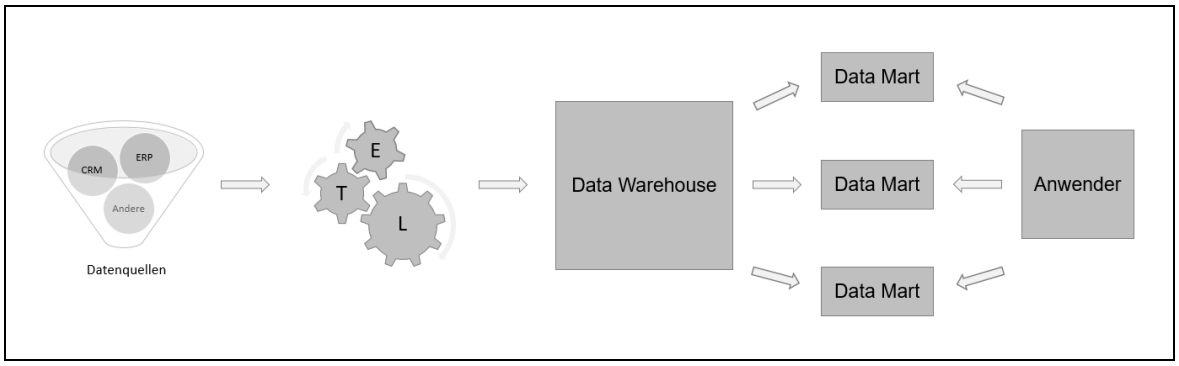


Abbildung 36: Das Data Warehouse (eigene Darstellung in Anlehnung an Freiknecht & Papp, 2018, S. 18)

Die Daten werden zunächst aus den verschiedenen internen und externen Datenquellen gesammelt. Anschließend werden sie mittels ETL[2]-Tools extrahiert, transformiert und als relationale, strukturierte Daten in das Data Warehouse geladen. Aus dem Data Warehouse heraus werden Extrakte in Data Marts abgelegt, auf die Anwender zugreifen können. Data Marts sind kleinere dezentralisierte Systeme, in denen Teilmengen von Daten aus einem Data Warehouse organisiert und bestimmten Gruppen von Geschäftsanwendern zur Verfügung gestellt werden (vgl. Rouse, 2016b). Die Notwendigkeit der Aufarbeitung der gesammelten Daten impliziert den fehlenden direkten Zugriff auf die ursprünglichen operativen Daten (vgl. Gluchowski & Roland, Karsten, 2008, S. 117).

Auf der Grundlage dieser Definitionen werden Datenintegration, Datenanalyse und die Entscheidungsunterstützung als grundlegende Bestandteile eines Data Warehouse angesehen (vgl. Farkisch, 2011, S. 7). Dabei gilt die Entscheidungsunterstützung als übergeordnetes Ziel eines Data Warehouse. Die Bestandteile Datenerfassung, Datenintegration und Datenanalyse führen letztendlich zu den entscheidungsrelevanten und -unterstützenden Informationen.

Komponenten eines Data Warehouse

Es ist nun bekannt, dass ein Data Warehouse Daten speichert, die aus internen oder externen Systemen extrahiert werden. Die Datensätze innerhalb des Data Warehouse müssen eindeutige Spezifika enthalten, um sie recherchierbar und für Geschäftsanwender nützlich zu machen (vgl. Gadatsch, 2013, S. 282). Zusammengefasst gibt es drei Hauptkomponenten des Data Warehouse:

- eine Datenintegrationsschicht, die Daten aus operativen Systemen wie Excel oder SAP-ERP extrahiert,

[2] Extrahieren, Transformieren, Laden.

- einen Datenbereitstellungsbereich, in dem die Daten bereinigt und organisiert werden,
- einen Präsentationsbereich, in dem Daten gelagert und zur Nutzung bereitgestellt werden.

Eine Data-Warehouse-Architektur kann als eine Reihe von Schichten verstanden werden. Die untere Schicht entspricht zumeist dem Datenbankserver, die mittleren Schichten der Analysemaschine und die obere Schicht der Data-Warehouse-Software, die Informationen für die Berichtserstattung und Analyse präsentiert (siehe Abbildung 37).

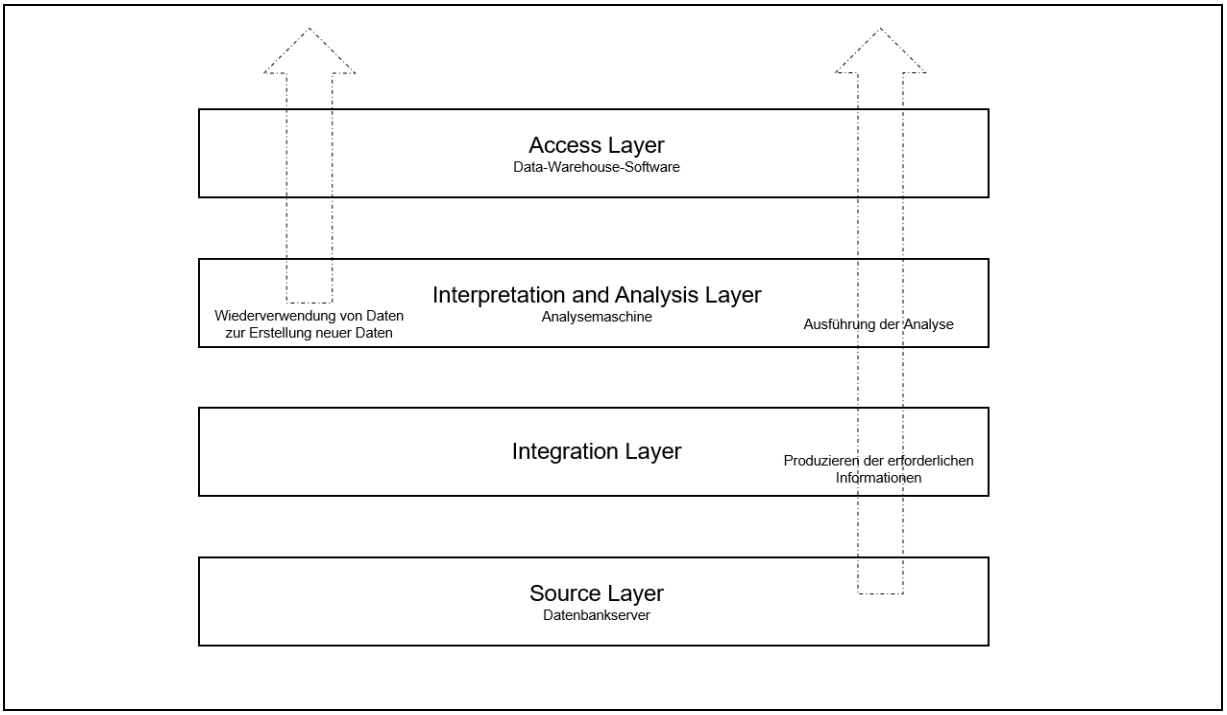


Abbildung 37: Beispielhafte Schichtenarchitektur von Data Warehouses (eigene Darstellung in Anlehnung an Gadatsch, 2013, S. 282–285)

Datenanalysewerkzeuge wie z.B. Tableau ermöglichen den Benutzern den Zugriff auf die Daten im Data Warehouse. Ein Enterprise Data Warehouse speichert analytische Daten für alle Geschäftsvorgänge eines Unternehmens. Alternativ können einzelne Geschäftseinheiten, insbesondere in großen Unternehmen, eigene Data Warehouses haben.

In diesem Kontext ist besonders Hadoop für viele Unternehmen zu einer wichtigen Erweiterung von Data Warehouses geworden, da die verteilte Datenverarbeitungsplattform die Komponenten einer Data-Warehouse-Architektur verbessern kann – von der Datenaufnahme über die Analyseverarbeitung bis hin zur Datenarchivierung. Hadoop ist ein kostenloses, in Java entwickeltes System für skalierbare und verteilte Software. Es basiert im Wesentlichen auf dem MapReduce-Algorithmus von Google Inc. und auf Empfehlungen des Google-Dateisystems.

Hadoop bietet die Möglichkeit, intensive Rechenprozesse mit großen Datenmengen auf mehreren Computerclustern durchzuführen (vgl. Luber, 2018). In einigen Fällen dienen Hadoop-Cluster als Bereitstellungsbereich für traditionelle Data Warehouses. In anderen Fällen werden Systeme, die Hadoop und andere große Datentechnologien enthalten, selbst als vollwertige Data Warehouses eingesetzt (vgl. Rouse, 2016b).

Implementierung von Data Warehouse

Es existieren drei Hauptansätze für die Implementierung eines Data Warehouse, die im Folgenden näher erläutert werden.

Der Top-down-Ansatz wurde von William H. INMON entwickelt (siehe Abbildung 38). Dieser Ansatz erfordert zunächst den Aufbau des Enterprise Data Warehouse. Hierbei werden die Daten zunächst aus operativen und möglicherweise externen Drittsystemen extrahiert und in einem sogenannten Staging-Bereich transformiert und aufbereitet. Anschließend werden sie in ein normalisiertes und standardisiertes Datenmodell (dritte Normalform; 3NF) geladen, um aus den gespeicherten Daten sogenannte Data Marts zu erstellen (vgl. Rouse 2016b).

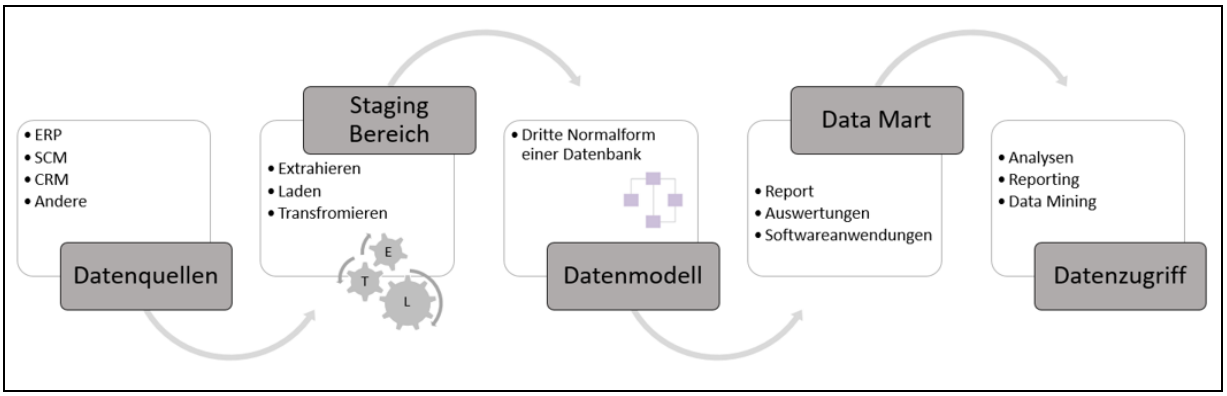


Abbildung 38: Der Top-down-Ansatz nach Inmon (eigene Darstellung in Anlehnung an Rouse, 2016b)

Ein weiterer Ansatz ist die Bottom-up-Methode (siehe Abbildung 39). Ralph KIMBALL hat diese alternative Data-Warehousing-Architektur entwickelt. Analog zum Top-down-Ansatz werden auch hier im Ausgangspunkt Daten aus den operativen oder externen Drittsystemen extrahiert und in einem Staging-Bereich transformiert. Jedoch werden sie im Anschluss nicht in einer 3NF-Datenbank abgelegt, sondern in ein Sternschema überführt. Diesbezüglich werden eine oder mehrere Faktentabellen mit einer oder mehreren Dimensionstabellen verbunden. Die Daten werden dann verarbeitet und in Data Marts geladen, von denen sich jedes auf einen bestimmten Geschäftsprozess konzentriert. Die Data Marts werden mithilfe einer Data-

Warehouse-Busarchitektur integriert, um ein unternehmensweites Enterprise Data Warehouse zu bilden (vgl. Rouse, 2016b).

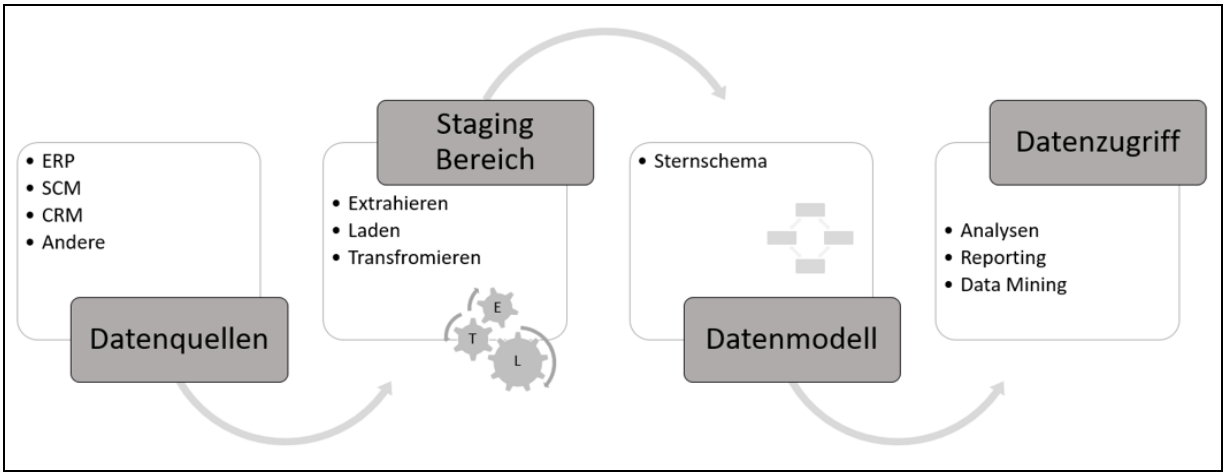


Abbildung 39: Die Bottom-up-Methode nach Kimball (eigene Darstellung in Anlehnung an Rouse, 2016b)

Oftmals versuchen Unternehmen, die Geschwindigkeit des Bottom-up-Ansatzes mit den Datenintegrationsmöglichkeiten des Top-down-Ansatzes zu kombinieren. Dies wird als Hybridmethode bezeichnet (vgl. Rouse, 2016b).

Vorteile und Optionen

Die Implementierung von Data Warehouse bietet für die Unternehmen sowohl aus IT- als auch aus betrieblicher Sicht einen Vorteil.

So kann die Trennung von analytischen und operativen Prozessen die Leistung von operativen Systemen verbessern und es den Datenanalysten und Geschäftsanwendern ermöglichen, schneller auf relevante Daten aus verschiedenen Quellen zuzugreifen und diese abzufragen. Überdies können Data Warehouses eine verbesserte Datenqualität und -konsistenz für Analysezwecke bieten und dadurch die Genauigkeit von Business-Intelligence-Anwendungen verbessern (vgl. Rouse, 2016b).

Unternehmen können grundsätzlich zwischen lokalen Systemen, herkömmlichen Cloudbereitstellungen oder Data-Warehouse-as-a-Service-Angeboten (kurz DWaaS) wählen. Hierbei bieten ein lokales Data Warehouse Flexibilität und Sicherheit, sodass die verantwortlichen IT-Fachbereiche die Kontrolle über die Data-Warehouse-Verwaltung und -Konfiguration behalten können. Im Gegensatz dazu ermöglicht ein cloudbasiertes Data Warehouse dem Unternehmen eine schnelle Skalierung seiner Systeme. Zusätzlich entfallen die anfänglichen Investitionen in die Infrastruktur und die Kosten für die laufende

Systemwartung. Der Data-Warehouse-as-a-Service bietet zudem einen verwalteten Cloudservice, der Unternehmen von der Notwendigkeit befreit, ihre Data Warehouse zu implementieren, zu konfigurieren und zu verwalten.

Ein Schwachpunkt des Data Warehouse besteht darin, dass es sich primär auf die Beschaffung und Bereitstellung herkömmlicher strukturierter Daten, vor allem aus SQL-Datenbanken, konzentriert. Im Big-Data-Umfeld ist es jedoch auch erforderlich, auf eine große Menge an Daten zuzugreifen, welche häufig nur in unstrukturierter Form vorliegt. Darüber hinaus müssen wesentlich größere Datenmengen beschafft und zur Verfügung gestellt werden (vgl. Luber, 2018).

4.2.2 Data-Lake-Architektur

Ein Data Lake ist ein Datenspeicher, der eine riesige Menge an Rohdaten bis zu deren Anwendung in ihrem ursprünglichen Format speichert (vgl. Laskowski, 2016). Während ein hierarchisches Data Warehouse Daten in Dateien oder Ordnern speichert, verwendet ein Data Lake eine flache Architektur zur Speicherung von Daten. In einem Data Lake wird jedem Datenelement eine eindeutige Kennzeichnung zugewiesen und mit einem Satz erweiterter Metadaten-Tags versehen. Sobald eine fachliche Frage aufkommt, kann der Data Lake nach relevanten Daten abgefragt werden. Dieser kleinere Datensatz kann anschließend zur Beantwortung der Frage analysiert werden (vgl. Rouse, 2016a). Mit der Einführung von Hadoop, einer Datenplattform, welche mittels verteilter Dateisysteme unterschiedlichste Möglichkeiten der Datenaufbereitung bietet, hat sich Einiges geändert. Die Daten werden nun sofort auf das Hadoop-Cluster geladen. So ist von ETL (Extrahieren, Transformieren, Laden) die Rede, wenn die Daten bereits vor dem Laden vom Data Warehouse transformiert wurden, bevor sie gespeichert werden konnten. Bei Data Lakes wird jedoch der Begriff ELT (Extrahieren, Laden, Transformieren) verwendet. Dabei werden, wie in Abbildung 40 dargestellt, die Daten aus den Ursprungssystemen entnommen, in den Data Lake geladen und anschließend in ein Zielsystem transformiert.

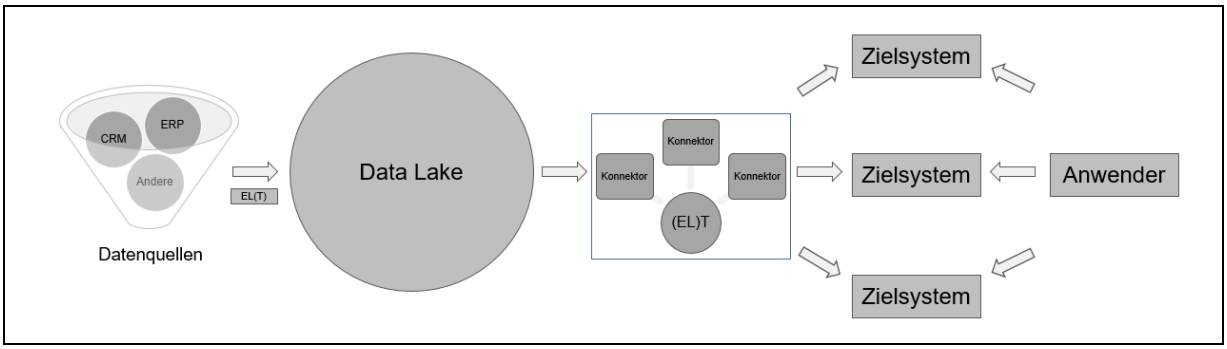


Abbildung 40: Der Data Lake (eigene Darstellung in Anlehnung an Freiknecht & Papp, 2018, S. 18)

Mithilfe des Konzepts des Data Lake kann das Data Warehouse zu einer Plattform für die Analyse großer Datenmengen erweitert werden. Der Data Lake besitzt eine hohe Speicherkapazität und bietet dementsprechend die Möglichkeit, große Datenmengen aufzunehmen. Zugleich ermöglicht er die Verarbeitung unterschiedlichster Datenformate, einschließlich unstrukturierter Daten. Die vielfältigen Daten des Data Lake müssen jedoch in einem ersten Zwischenschritt so aufbereitet werden, dass die Benutzer mit den entsprechenden Tools darauf zugreifen können (vgl. Rouse 2016a). Die unstrukturierten Rohdaten des Data Lake werden also durch geeignete Transformationen in strukturierte Datensätze transformiert, die mit den Tools für den Datenzugriff des Data Warehouse angezeigt und untersucht werden können (vgl. Luber 2018). Das Schema und die Datenanforderungen werden erst bei der Abfrage der Daten definiert.

Der Begriff „Data Lake“ wird oft mit Hadoop-orientierter Objektspeicherung in Verbindung gebracht. In einem solchen Szenario werden die Daten eines Unternehmens zunächst in die Hadoop-Plattform geladen. Anschließend können Geschäftsanalyse- und Data-Mining-Tools auf die Daten angewendet werden, die sich auf den Clusterknoten der Hadoop-Computer befinden (vgl. Rouse 2016a).

Vorteile und mögliche Gefahren

Der Data Lake bringt eine Vielzahl von Vorteilen mit sich. Seine Flexibilität ermöglicht es Entwicklern und Datenwissenschaftlern, ein bestimmtes Datenmodell, eine Anwendung oder eine Abfrage im laufenden Betrieb einfach zu konfigurieren. Gleichzeitig stellt der Data Lake Daten und Informationen für zahlreiche Benutzertypen unabhängig vom zeitlichen Investitionsaufwand zur Verfügung. Auf der einen Seite unterstützt der Data Lake jene Benutzer, die mit der Datenquelle bestehende als auch völlig neue Sachverhalte generieren. Auf der anderen Seite erhalten Anwender notwendige Informationen aus dem Data Lake, um tägliche Reports zu erstellen. Außerdem sind Data Lakes verhältnismäßig kostengünstig zu

implementieren. Die meisten Technologien, die zu ihrer Verwaltung verwendet werden, basieren auf Open-Source-Anwendungen. Ein weiterer Vorteil ist die Tatsache, dass kein konkreter oder vordefinierter Plan für den Anwendungsbereich der gespeicherten Daten existieren muss. Dadurch wird die arbeitsintensive Datenmodellentwicklung und Datenbereinigung so lange aufgeschoben, bis ein Fachbereich einen klaren Bedarf für die Daten identifiziert. Diesbezüglich ermöglicht der Data Lake eine Vielzahl verschiedener Analysemethoden zur Dateninterpretation, einschließlich der Analyse großer Datenmengen, Echtzeitanalyse oder SQL[3]-Abfragen (vgl. Rouse, 2016a).

Trotz der Vorteile dieses kostengünstigen, unstrukturierten Datenspeichers, der einem Unternehmen zur Verfügung steht, bestehen mögliche Gefahren beim Einsatz von Data Lakes. Eine der größten potenziellen Gefahren des Data Lakes ist, dass er sich in einen Datensumpf oder Datenfriedhof verwandeln könnte. Falls nämlich ein Unternehmen eine schlechte Datenverwaltung praktiziert, besteht die Gefahr, dass der Überblick über die im Data Lake vorhandenen Daten verloren geht, selbst wenn stetig Daten hineinfließen. Das Ergebnis ist der Verlust potenziell wertvoller Daten, die sozusagen unbemerkt auf dem Grund des Data Lakes verloren gehen. Dadurch werden die Daten verschlechtert, nicht verwaltet oder gar unzugänglich gemacht. Zwar stehen Data Lakes theoretisch allen Mitarbeitern eines Unternehmens zur Verfügung, in der Praxis sind sie jedoch unter Umständen nicht so leicht zugänglich. Die verantwortlichen Analysten haben möglicherweise Schwierigkeiten, unstrukturierte Daten aus einer Vielzahl von Quellen zu analysieren. Diese praktische Herausforderung des Zugangs zu etwaigen Daten kann zu einem Mangel an angemessener Datenpflege beitragen und zur Entwicklung eines Datenfriedhofs führen. Es ist wichtig, die Investitionen in einem Data Lake zu maximieren und das Risiko einer fehlgeschlagenen Bereitstellung zu verringern (vgl. Rouse, 2016a).

4.2.3 Sensorik

Die IT-Komponenten und auch Sensoren werden immer preiswerter. Gleichzeitig steigt auch die Leistungsfähigkeit der Systeme. In der Produktion sind die Anlagen heutzutage größtenteils ab der Inbetriebnahme mit Sensoren ausgestattet. Auch andere Objekte sind immer häufiger mit Sensoren ausgestattet. Unter dem Internet of Things versteht man die Integration der

[3] Structured Query Language.

Sensordaten aus verschiedensten Objekten, realisiert durch Anwendungen, welche das Internet nutzen (vgl. Winkelhake, 2017, S. 61). Das Internet of Things beruht auf folgenden Grundsätzen (vgl. Bischoff et al., 2015, S. 10):

- der Speicherung individueller Informationen direkt am Objekt,
- der Vernetzung der Objekte,
- der individuellen Entscheidungsfindung auf der Basis lokal ausgewerteter Informationen,
- den individuellen Services auf Abruf zur echtzeitnahen und ereignisorientierten Steuerung von Prozessen.

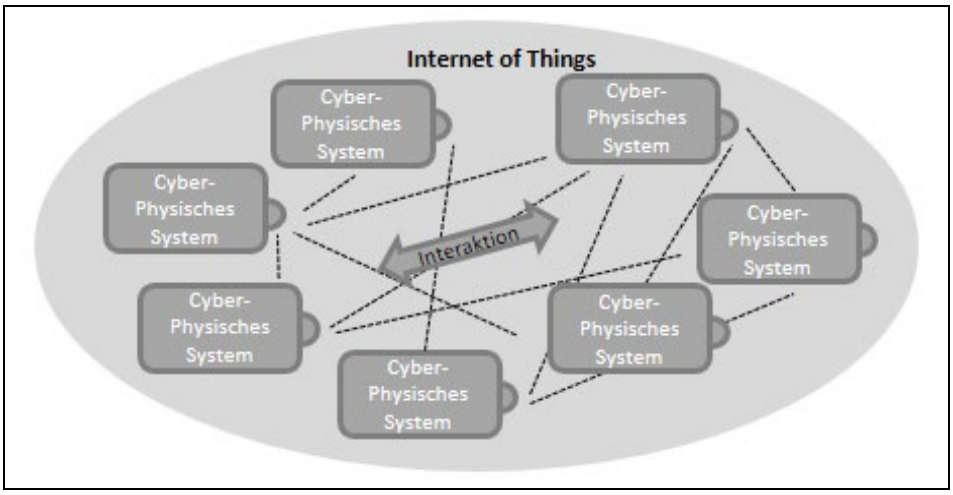


Abbildung 41: Vernetzung der cyber-physischen Systeme im Internet of Things (eigene Darstellung in Anlehnung an Bischoff et al., 2015, S. 10)

Das Internet of Things bildet die Infrastruktur zur Integration cyber-physischer Objekte zu einem globalen Netzwerk, so wie in Abbildung 41dargestellt. Es bietet den cyber-physischen Systemen und den Menschen einen kontrollierenden, koordinierenden und ortsunabhängigen Zugriff auf alle vernetzten cyber-physischen Systeme (vgl. Bischoff et al., 2015, S. 10).

4.2.4 Integration einer Big-Data-Station in ein Informationsökosystem

Um den im vorangegangen Abschnitt des Data Warehouse erwähnten Herausforderungen gerecht zu werden, wurde das Konzept des Big-Data-Shops in Anlehnung an Data Lakes entwickelt.

Das Big-Data-Station-Modell soll zunächst an einem Produktionsstandort in der Praxis zum Einsatz kommen. Dieses Konzept entstand in einem iterativen Prozess im Big-Data-Projekt

durch den Autor und wird in Kooperation im Werk erprobt. Das Modell dient der effizienten Auswertung von Daten in Echtzeit mithilfe einer besonderen Datenselektion und -auswertung durch das Machine Learning. Insbesondere sollen Vorhersagen über mögliche zukünftige Fehlerquellen getroffen und optimale Wartungszeitpunkte durch die Implementierung von Machine Learning und Predictive Maintenance festgelegt werden. Hierzu werden u.a. die Technologien der In-Memory- und Streaming-Systeme genutzt.

Abbildung 42 gibt einen Überblick über die Funktionsweise des Big-Data-Station-Modells.

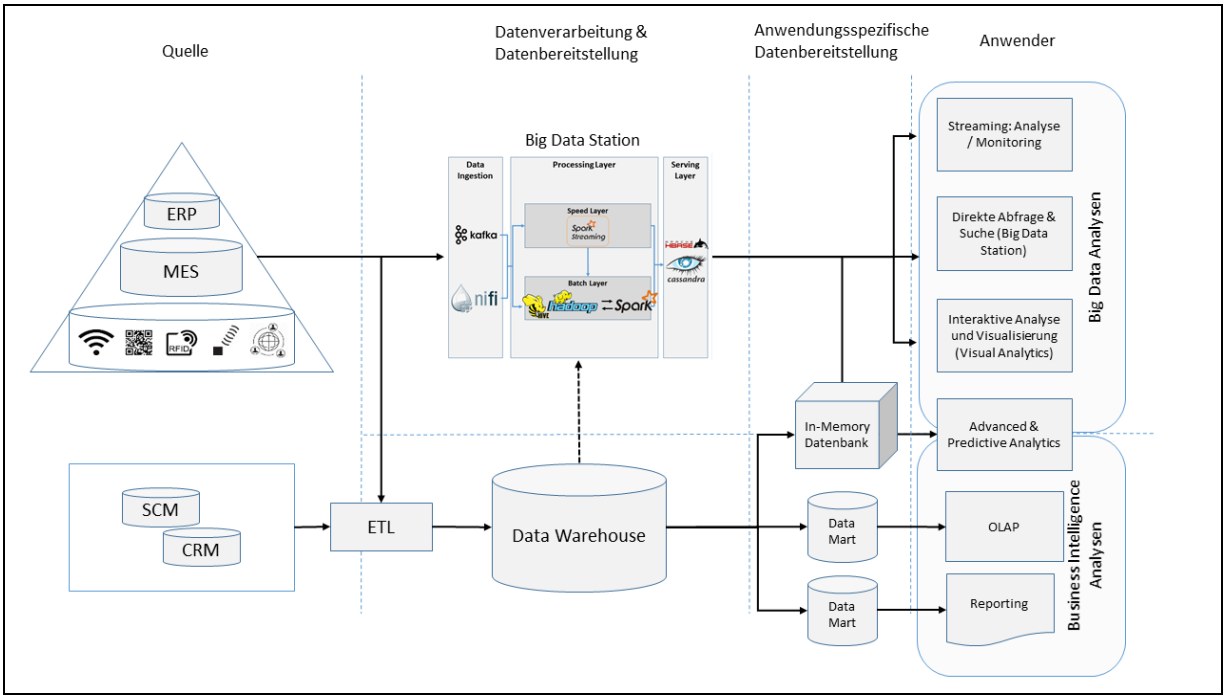


Abbildung 42: Big-Data-Station-Konzept (eigene Darstellung)

Das Modell der Big-Data-Station ist grundsätzlich in drei Ebenen unterteilt. Diese Ebenen stellen Quellsysteme, die Datenbereitstellung und die Informationsgenerierung dar. Darüber hinaus kann eine Einteilung der Prozesse und Tools in Big Data und etablierte BI-Komponenten vorgenommen werden. Diese Gegenüberstellung dient dem Vergleich zwischen der herkömmlichen Vorgehensweise in der Praxis und der Variationsmöglichkeit von Big Data im Kontext der Industrie 4.0. Im Folgendem wird das Modell erläutert. Dabei wird jeweils zuerst auf die herkömmlichen BI-Komponenten und anschließend auf die Big-Data-Komponenten eingegangen.

Die Quellsysteme stellen geeignete Funktionen zur Datenerfassung sowohl interner als auch externer Daten dar. Hierzu existiert keine Abgrenzung zwischen BI- und Big-Data-Komponenten. Zu den Datenquellen zählen W-Lan, QR-Codes, RFID-Chips, Lokalisierung und Daten bzw. Informationen aus dem Internet. Diese Informationen werden über das MES

zu KPIs verarbeitet und anschließend in das ERP-System weitergeleitet. Von dort aus gelangen die Daten direkt in den Big Data Store. Darüber hinaus werden auch unternehmensbezogene Daten aus dem CRM und SCM erfasst und für die Datenanalyse herangezogen, wobei diese den klassischen Weg des Data Warehouse gehen, da in der Arbeit nicht auf SCM- und CRM-Analysen eingegangen wird.

Zur Funktionsweise:

Die erhaltenen bzw. generierten Rohdaten aus den Quellen werden anschließend in der Datenbereitstellung verarbeitet. Das Modell wird in der Datenbereitstellung in zwei verschiedene Teile untergliedert, einerseits die anwendungsübergreifenden und andererseits die anwendungsspezifischen Komponenten. Auf anwendungsübergreifender Seite geschieht dies zum einen über die sogenannten Big-Data-Station und zum anderen im Bereich der herkömmlichen BI-Komponenten auf klassische Weise durch ETL-Prozesse. ETL ist eine Form der Datenintegration. Über den herkömmlichen Weg werden somit verschiedene Rohdaten durch Extraktion, Transformation und Überleitung für die Speicherung im Data Warehouse bereitgestellt. Darüber hinaus kommt der Big Data Store zum Einsatz, welcher durch Machine Learning und Data Mining mit diversen Algorithmen anwendungsübergreifend Daten in Echtzeit verarbeitet. Der Vorteil der Big-Data-Station ist, dass parallel zur Datenbereitstellung die Daten auch darin selektiert und gespeichert werden. Des Weiteren existiert auch eine Verbindung zwischen der Big-Data-Station und den ETL-Prozessen. Durch diese Verbindung können auch Daten aus der Big-Data-Station mithilfe von ETL in das Data Warehouse importiert werden. Auf anwendungsspezifischer Seite werden für die Datenbereitstellung auf Seiten der BI-Komponenten Data Marts genutzt. Diese fungieren als Kopien einzelner Teilbestände an Daten aus dem Data Warehouse. Diese Teildatenbestände können sowohl für einzelne Organisationsbereiche als auch zweckorientiert für bestimmte Analysen wie bspw. OLAP verwendet werden. Verschiedene Data Marts erlauben also unterschiedlichen Benutzern den Zugriff auf spezielle Teildatenbestände. Auf Seiten der Big-Data-Komponenten werden Graphendatenbanken und weitere NoSQL-Datenbanken im Serving Layer genutzt. Eine besondere Form der anwendungsspezifischen Datenbereitstellung ist die In-Memory-Datenbank. Diese kann sowohl als Big-Data-Komponente als auch von Seiten der herkömmlichen BI-Komponente genutzt werden und ist keinem der beiden Bereiche eindeutig zurechenbar, da sie Komponenten aus beiden Aspekten beinhaltet.

Auf der Ebene der Informationsgenerierung gilt es, die aus den Daten gewonnenen Informationen auszuwerten und zu visualisieren. Bei der Informationsgewinnung kommen

verschiedene Tools zum Einsatz. Auf der Seite der etablierten BI-Komponenten wird sowohl OLAP als auch das Reporting herangezogen. Neben den herkömmlichen Analysemethoden der BI gibt es Advanced und Predictive Analytics. Bei Advanced und Predictive Analytics handelt es sich um über den Standard hinausreichende Auswertungen und Vorhersagen zu Entwicklungen. Daraus resultierend können Entscheidungsträger mögliche Maßnahmen zur Optimierung einleiten. Dieses wichtige Analysewerkzeug kann sowohl im Bereich der etablierten BI-Komponenten als auch für die Big-Data-Komponenten eingesetzt werden.

Bei den Big-Data-Komponenten gibt es insgesamt drei verschiede Möglichkeiten der Informationsgenerierung. Informationen können in Echtzeit über Streaming-Daten analysiert und überwacht werden. Das bedeutet, dass die die Daten direkt aus den Quellsystemen zur Informationsgewinnung genutzt werden. Hierbei geht es hauptsächlich darum, Daten über Störungen bspw. so schnell wie möglich und direkt an die Empfänger und Verantwortlichen zu melden. Somit wird der Schritt der Datenbereitstellung übersprungen und Daten in Echtzeit mithilfe von Streaming übertragen. Darüber hinaus besteht die Möglichkeit, Informationen unmittelbar aus der Big-Data-Station zu generieren. Durch direkte Abfragen und Suchen können Daten gezielt in Relation zueinander gesetzt werden und so zu Informationen verarbeitet werden. An dieser Stelle wird das Predictive Maintenance eingesetzt. Hierzu werden Daten über verschiedene Fehlermeldungen mithilfe von Algorithmen gesammelt und gespeichert, um diese jederzeit auswerten zu können. Anschließend sollen innerhalb kürzester Zeit Prognosen über weitere Ausfälle oder eine effiziente Behebung der Fehlerquelle aus den vorhandenen Informationen generiert werden. Eine weitere Möglichkeit ist die interaktive Analyse und Visualisierung durch Visual Analytics. Hierzu werden die benötigten Daten aus dem anwendungsspezifischen Bereich der Datenbereitstellung herangezogen und zu aktuellen Dashboards verarbeitet. Für die Visualisierung kann ebenso Realtime Intelligence eingesetzt werden. Darunter versteht man eine kontinuierliche Echtzeitanimation, basierend auf Streaming-Daten.

Das Modell beinhaltet zudem auch Prämissen, die die Anwendungen der Big-Data-Station gewährleisten. Dazu zählen Administration, Metadata-Management und Data Quality. Diese drei Voraussetzungen müssen strikt eingehalten werden, damit das Modell zweckorientiert angewendet werden kann und die Informationsgewinnung eine hohe Reliabilität aufweist.

Die Besonderheit des Big Data Stores besteht darin, dass er eine Hybridlösung aus Batch- und Streaming-Verfahren darstellt, welche auch als Lambda-Architektur bezeichnet wird. Nun wird

diese Lambda-Architektur näher betrachtet. Abbildung 43 veranschaulicht graphisch den Aufbau dieser Hybridlösung.

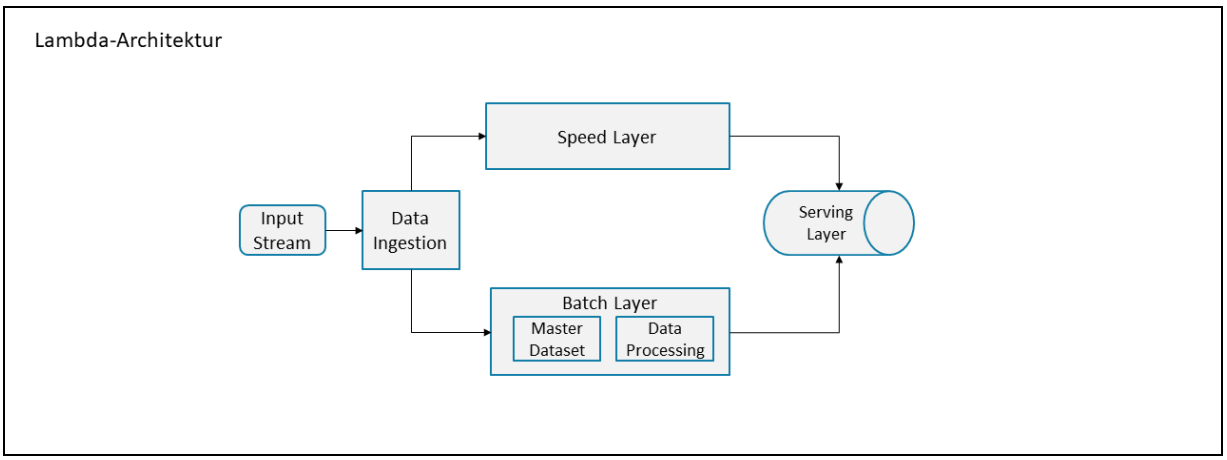


Abbildung 43: Lambda-Architektur nach Nathan Marz (eigene Darstellung)

Die Besonderheit der Lambda-Architektur ist, dass hierbei Daten einerseits über Batch und andererseits durch Streaming verarbeitet werden (vgl. Freiknecht & Papp 2018, S. 477). Die Kombination beider Verfahren ermöglicht die schnelle Datenausgabe sowohl von historischen Daten als auch von Echtzeitdaten (vgl. Freiknecht & Papp 2018, S. 478). Die historischen Daten werden dazu im Batch Layer aufbereitet und anschließend durch den Service Layer in das Datenbanksystem geladen. Die Aufbereitung erfolgt durch das ETL-Verfahren.

Im Speed Layer werden eingehende Daten im Cache gespeichert, welcher die schnelle Datenabfrage begünstigt (vgl. Freiknecht & Papp, 2018, S. 478). Ein Nachteil des Streamings ist jedoch, dass Daten bei einem unerwarteten Systemausfall im Speed Layer verloren gehen können. Dem wird versucht, entgegenzuwirken, indem alle eingehenden bzw. neuen Daten insgesamt zweimal verfügbar gemacht werden. Einerseits erfolgt die Archivierung der Daten im Batch Layer und andererseits sind diese im Speed Layer für die schnelle Verarbeitung vorhanden (vgl. Freiknecht & Papp, 2018, S. 478). Diese Hybridlösung ermöglicht mithilfe dieser Architektur die Verarbeitung von Daten in Echtzeit.

Hinsichtlich der Implementierung soll der Big Data Store nicht nur Daten in Realtime verarbeiten, sondern zusätzlich die Problematik der Insellösungen und deren Inkompatibilität beseitigen. Das Zielsystem für das Datenmanagement soll zukünftig nicht durch komplexe und systemübergreifende Insellösungen Daten erfassen und ex post aufbereiten, sondern gezielt durch den Big Data Store Echtzeitanalysen generieren und darüber hinaus durch das Predictive Maintenance mögliche kritische Situationen in der Produktion sowohl identifizieren als auch antizipieren. Der Vorteil der Implementierung dieser Technologie basiert auf der Big-Data-

Verarbeitung in Clustern (verteiltes System) und auf In-Memory- und Streaming-Daten. Insgesamt soll so die jetzige Anwendungssoftware im SCADA-System langfristig nicht mehr zur Datengewinnung eingesetzt werden, da sie nicht für die Datengenerierung im Bereich Datenanalyse und für Predictive Maintenance geeignet ist. Durch die Big-Data-Station wird stattdessen eine neue Schnittstelle erzeugt. Die Tools werden zusammen mit den Schnittstellen in einer Big-Data-Architektur untergebracht, um somit eine gesamte Struktur für das Datenmanagement zu gewährleisten. Aufgrund dieser technologischen Entwicklung können Daten temporär gespeichert werden und somit komplexe Serverstrukturen umgangen werden. Hierzu wird nicht mehr die Gesamtheit aller Daten in der Datenbank gespeichert, sondern nur noch relevante Produktionsdaten in den Zwischenspeicher geladen. Dies bietet den Vorteil, den Abruf der Daten in Echtzeit zu generieren.

Die Big-Data-Station kann durch ihre Beschaffenheit über verschiedene Ebenen hinweg, begonnen beim Quellsystem über die Datenbereitstellung bis hin zur Informationsgenerierung, Daten verarbeiten und direkt für die Visualisierung aufbereiten und somit fundierte Entscheidungen unterstützen.

In Bezug auf die Produktionsprozesse in der Fertigung des Automobilwerks soll die Big-Data-Station im Wesentlichen die Koordination und das Management der unterschiedlichen Daten übernehmen, um somit Insellösungen zu vermeiden. So bildet sie den zentralen Knotenpunkt zu weiteren vernetzten Systemen und fungiert als zentrales Organ in einem zukünftigen CPS-Ökosystem. In operativer Hinsicht sollen insbesondere Konzepte wie Machine Learning und Predictive Maintenance auf einer skalierbaren Plattform Anwendung finden und zu einer Effizienzsteigerung führen, da somit zukünftige Ausfälle und optimale Wartungsintervalle berücksichtigt werden. Langfristig soll der Karosseriebau dadurch so weit wie möglich digitalisiert werden und die Steuerung der Prozesse dezentral ermöglichen.

4.3 Datenmanagement

In der Zeit vor Big Data wurden umfangreiche Analysen nahezu ausnahmslos in Datenbanken durchgeführt. Grundsätzlich folgte jedes Data Warehouse den Prinzipien relationaler Datenbanken. Voraussetzung für den zielgerichteten Umgang mit Big Data ist die Schaffung einer strukturierten und einheitlichen Datengrundlage gemäß den Anforderungen der Endnutzer (vgl. Quaing 2010, S. 279). Der Einsatz von unterschiedlichen operativen Systemen in den jeweiligen Fachabteilungen führte zu einer unterschiedlichen Datenstruktur und -qualität.

Darüber hinaus können durch heterogene Systemlandschaften Inkompatibilitäten und hohe Datenredundanzen entstehen. Die Einführung von zentral verwalteten Datenbanken soll dem nun entgegenwirken. Mit dem Einsatz verschiedener operativer Systeme werden zwar umfangreiche Datenmengen gesammelt, aber nur ein kleiner Teil von ihnen wird ausgewertet. Ein wesentlicher Grund hierfür ist das Fehlen effizienter Analysesysteme (vgl. Samtleben, 2007, S. 26).

Datenquellen (Data Sources):

Die verschiedenen Arten von Quellen, die im großen Datenökosystem vorhanden sind, sind Sensoren, Standortverfolger, Transaktionsprotokolle, soziale Medien, E-Mail und Nachrichtenübermittlung. Dies sind die wesentlichen Quellen, welche die Lebenszyklusphase der Analytik kennzeichnen (siehe Kapitel 3, S. 63). Die Formate der aus den einzelnen Quellen produzierten Daten unterscheiden sich und müssen durch Transformation verarbeitet werden. So können bspw. die Sensoren und Standortverfolger Daten in strukturiertem Format produzieren, während das Format der in sozialen Medien erzeugten Daten unstrukturiert ist. Daher müssen verschiedene Datenquellen, die unterschiedliche Formate der Daten erzeugen, in einer Form aggregiert werden, die in Datenerfassungssystemen gespeichert werden kann. Die Anbindung an eine klassische Datenquelle wie das Data Warehouse wird nun zunächst untersucht, um dann auf den Verarbeitungsprozess mit weiteren unterschiedlichen Datenquellen (z.B. Data Lake, IoT) einzugehen.

Data Ingestion (Input Layer):

Die aus den Quellen generierten Daten können durch Systeme wie Hadoop, Hive, HBase, Kafka, Storm etc. erworben werden. Das Beispiel der hier dargestellten Systeme gehört zum Hadoop-Ökosystem. Obwohl für die Datenerfassung mehrere Plattformen zur Verfügung stehen, werden diese mehrheitlich für analytische Anwendungen genutzt. Der Hauptvorteil von Hadoop-basierten Systemen beruht auf einer verteilten Datenverarbeitung. Diese Systeme unterstützen die Anzeige der Daten in verschiedenen Formaten, z.B. kann eine in Hadoop gespeicherte Textdatei als Tabelle in Hive angezeigt werden. Auf diese Weise werden Datenerfassungssysteme zum Speichern der aus verschiedenen Quellen erfassten Daten verwendet.

Data Processing (Process Layer)

Die in den Systemen gespeicherten Daten können mit folgenden Techniken analysiert werden: Maschinelles Lernen, Tiefenlernen, Klassifikation, Clustering, Regression und andere. Die

Datenerfassungssysteme müssen die Datenanalyse in diesem Zusammenhang weiter unterstützen. Die im vorherigen Schritt betrachteten Systeme unterstützen die Datenanalyse durch verschiedene Programmiermodelle, bspw. stellen Hadoop und Amazon Ec2 das Programmiermodell MapReduce für die Analyse zur Verfügung. Hive bietet SQL-Unterstützung für die Analyse. Kafka und Storm unterstützen die Echtzeitanalyse von Daten, die aus verschiedenen Quellen über Tüllen und Bolzen gesammelt werden. Allerdings werden Techniken des Maschinellen Lernens wie Klassifikation, Regression, Clustering von Apache Spark unter Verwendung des MLib-Moduls unterstützt.

Applications (Service Layer):

Das Ökosystem von Big Data mit Datenerfassung und Datenanalyse führt zu intelligenten Anwendungen. Zu den Beispielen für intelligente Anwendungen gehören kontextbezogene Anwendungen, Warnungen und Benachrichtigungen, Genomanalyse, standortbezogene Anwendungen, Einzelhandelsanalysen und Fitnessanwendungen. Diese Anwendungen basieren auf den in der Datenanalysephase erstellten Modellen. Somit befinden sich die intelligenten Anwendungen an der letzten Linie des Ökosystems. Da sich das Ökosystem mit neueren Technologien und Datenerfassungssystemen weiterentwickelt, können intelligente Anwendungen nicht nur genutzt, sondern auch mit neuen Funktionen aktualisiert werden. Das Big-Data-Ökosystem kann auf folgende Weise mit dem Analyselebenszyklus verbunden werden. Die Datenquellen können auf Phase 1 abgebildet werden, d.h. mit der Identifizierung der Ziele. Die Datenerfassungssysteme können auf Phase 2, d.h. beim Verstehen der Anforderungen, abgebildet werden. Die Datenverarbeitung kann auf Phase 3 mit der Modellbildung und intelligente Anwendungen können auf Phase 4, der Durchführung von Analysen, abgebildet werden. Die Phase 5 des Analyselebenszyklus, d.h. die Visualisierung, ist nicht im Big-Data-Ökosystem aufgeführt, da sie von der für die Analyse betrachteten Domäne abhängt.

4.3.1 Datenarten

Im Allgemeinen findet in der Literatur eine Differenzierung zwischen strukturierten und unstrukturierten Daten statt. Unter diesem Aspekt sind die Rohdaten der Maschinen und Sensoriken zu nennen, um die es in diesem Kapitel geht. Strukturierte Daten haben eine normalisierte Form. Sie sind in zeilen- und spaltenorientierten Datenbanken gespeichert (vgl. Luber & Litzel, 2017). Die strukturierten Daten lassen sich u.a. auch in Stammdaten und

Bewegungsdaten aufteilen. Stammdaten wiederum werden auch in Grund- und Strukturdaten unterschieden. Unstrukturierte Daten besitzen hingegen eine nicht identifizierbare Struktur. Als Beispiel sind hier Bilder, Videodaten, Audiodaten oder auch Präsentationen zu nennen.

In der Tabelle 9 sind die Daten in ihrer Eigenschaft dargestellt.

Datenquelle	Herkunft	Eigenschaft	Geschwindigkeit
Operative Daten (ERP, MES, SPS, Maschinen etc.)	Unternehmensintern	Strukturiert	Grund- & Strukturdaten
Dokumenten-Management	Unternehmensintern	Unstrukturiert	Grund- & Strukturdaten
Web-Server-Logs	Unternehmensintern	Semi-strukturiert	Bewegungsdaten
Netwerk-Router-Logs	Unternehmensintern	Semi-strukturiert	Bewegungsdaten
Sensor-Zeitreihen	Unternehmensintern	Semi-strukturiert	Bewegungsdaten
Intelligente Stromzähler	Unternehmensintern	Semi-strukturiert	Bewegungsdaten
Internet-Foren	Unternehmensextern	Unstrukturiert	Grund- & Strukturdaten
Blogs	Unternehmensextern	Unstrukturiert	Grund- & Strukturdaten
Micro-Blogs	Unternehmensextern	Unstrukturiert	Bewegungsdaten

Tabelle 9: Eigenschaften von Daten (eigene Darstellung nach Lanquillon & Mallow, 2015b, S. 266)

Die Datenquelle ist die Beschreibung davon, welche Daten in welcher Anwendung erzeugt wurden (vgl. Lanquillon & Mallow, 2015, S. 266). Die Quelle kann zugleich auch das System sein, aus welchem die Daten generiert werden. Operative Daten sind somit aus einem ERP-, MES-, CRM-System etc. zu entnehmen. Auch können Daten direkt aus der Anlage oder der Maschine als operative Daten verstanden werden. Weitere Daten können aus der Unternehmensbetrachtung auch von außerhalb generiert und gewonnen werden. Das sind Daten aus dem Internet über soziale Netzwerke oder Videoplattformen, aber auch Daten, die käuflich zu erwerben sind. Zu nennen sind hier bspw. Wetterdaten. Dieser Aspekt wird in der „Herkunft“ der Daten als interne und externe Daten angedeutet. In Bezug auf die Datenqualität sowie deren Datenmanagement ist eine Differenzierung der Herkunft ebenfalls von Bedeutung (vgl. Lanquillon & Mallow, 2015, S. 266). Die Unterteilung der Daten in ihrer Geschwindigkeit ist notwendig für die Datenverarbeitung. Sind Grund- oder Strukturdaten vorhanden, werden diese im Batch Layer verarbeitet. Handelt es sich um Datenströme, die sogenannten Bewegungsdaten, so sind sie durch den Speed Layer abrufbar.

4.3.2 Datenbereitstellung

Im Kapitel 4.2. wurde die Architektur erläutert, auf der auch die Daten verarbeitet und die Bereitstellung der Daten unter zwei Aspekten betrachtet wurden. Die Integration einer Big-Data-Architektur in ein bestehendes System bzw. eine IT-Architektur bringt Zeit- und Kostenvorteile mit sich. Die Datenbereitstellung durch bestehende IT- und Produktionssysteme sowie dem dazugehörigen Data Warehouse ermöglicht eine breitgefasste Analyse von strategischen und auch operativen Geschäftsprozessen. Nun wird der Input Layer betrachtet und daraus geeignete Tools vorgestellt.

4.3.2.1 Apache Kafka

Damit der Betrieb unter mehreren IT-Systemen, in dieser Arbeit speziell Datenströmen, bestehen kann, ist eine Schnittstelle (Messaging-Instanz) zwischen der Hardware und der Software notwendig. Apache Kafka ist eine Open-Source-Software, welche eine vereinfachte und schnelle Verarbeitung von Datenströmen ermöglicht. Sie ist eine Lösung zur Speicherung, Übertragung und Weiterverarbeitung von Daten. Apache Kafka ist eine Anwendung der Apache Software Foundation und fügt sich nahtlos in das zuvor beschriebene Big-Data-Cluster ein. Mit Kafka werden Datenübertragungen zwischen Anlagen/Maschinen und anderen Produktionssystemen vorgenommen und verarbeitet. Sie ist die Direktverbindung zwischen Datenempfänger und Datenquelle. Dabei löst Kafka das Problem des Datenverlusts. Des Weiteren wird eine Überlastung des Datenempfängers verhindert. Dieses Problem tritt auf, wenn die Informationsgeschwindigkeit zu hoch ist für den Empfänger und dieser sie nicht verarbeiten kann. Apache Kafka arbeitet fehlertolerant. Bei Unterbrechungen bekommt Kafka ein Feedback und leitet die Informationen mit Datenfehlern nicht weiter, sodass ein Absturz vermieden wird. Der Vorteil ist auch die hohe Skalierbarkeit. Der Datentransport wird auf beliebig vielen Systemen verteilt. Damit ist eine schnelle Speicherung und Verarbeitung sowie hohe Verfügbarkeit möglich. Der Aufbau einer Kafka-Architektur wird als Cluster auf einem oder mehreren Servern ausgeführt. Kafka besteht aus Brokern, Topics und Partitionen (siehe Abbildung 44).

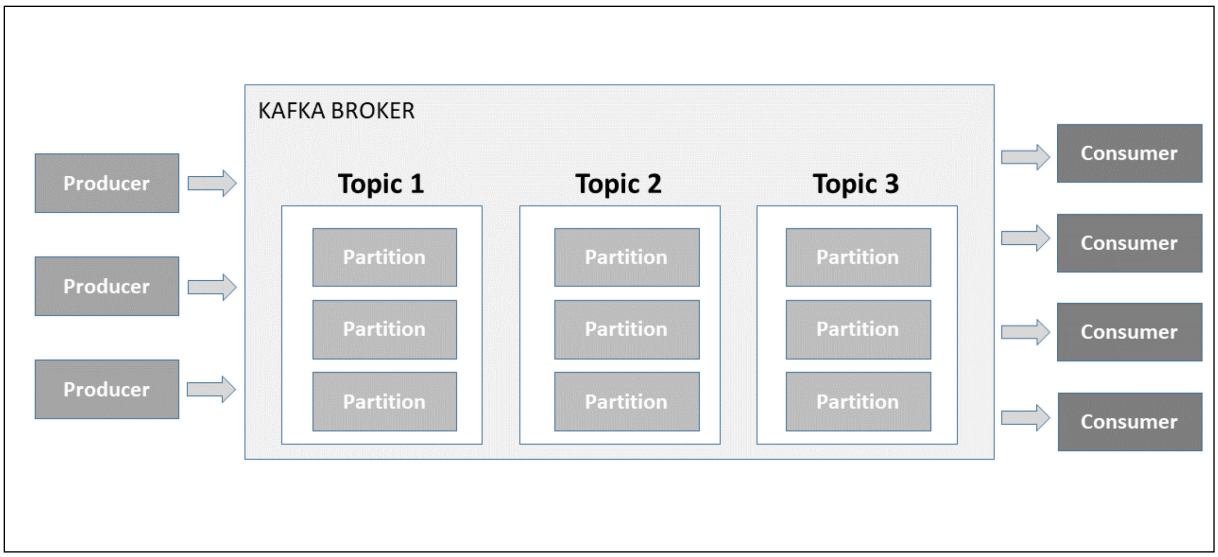


Abbildung 44: Kafka-Broker (eigene Darstellung in Anlehnung an Apache Software Foundation)

Broker sind einzelne Knoten im Cluster, welche die eingehenden Datenströme in Topics kategorisieren und speichern. Hier werden die Daten in Partitionen aufgeteilt und im Cluster repliziert und verteilt. Zusätzlich werden diese Partitionen mit einem Zeitstempel versehen. Das ist die Grundlage für die hohe Verfügbarkeit und erlaubt einen schnellen Lesezugriff. Topics werden in Normal Topics und Compacted Topics unterteilt. In Normal Topics können die Daten gelöscht werden, sobald Speicherzeitraum oder Speicherlimit überschritten werden. In Compacted Topics sind die Daten hingegen nicht abhängig von Zeit- und Platzrestriktionen. Es bestehen zwei Möglichkeiten, ein Kafka-Cluster aufzubauen. Ein Producer ist eine Anwendung, die Daten in den Kafka-Cluster schreibt. Ein Consumer ist eine Anwendung zum Lesen der Daten aus dem Cluster. In der Abbildung 45 sind zum einen das Producer-Consumer-Modell, zum anderen das Kafka-Connect-Modell dargestellt. Im Erstgenannten werden die Daten vom Producer (Quelle) als „push message“ über Kafka direkt als „pull message“ an den Consumer (Senke) gesendet. So wird ein Abbild der Quelle erzeugt, das uneingeschränkt weiterverarbeitet werden kann. Im Kafka-Connect-Modell werden zwischen der Datenquelle, Kafka und der Datensenke jeweils Kafka-Connectoren implementiert, um gezielt Daten abzugreifen. Diese werden dann wiederum durch einen weiteren Kafka-Connector selektiert und letztlich in die Datenquelle transportiert (vgl. Apache Foundation). Das zweite Modell ist

sehr gut für Streaming geeignet. Dort werden die Streaming-Daten kontinuierlich importiert und exportiert.

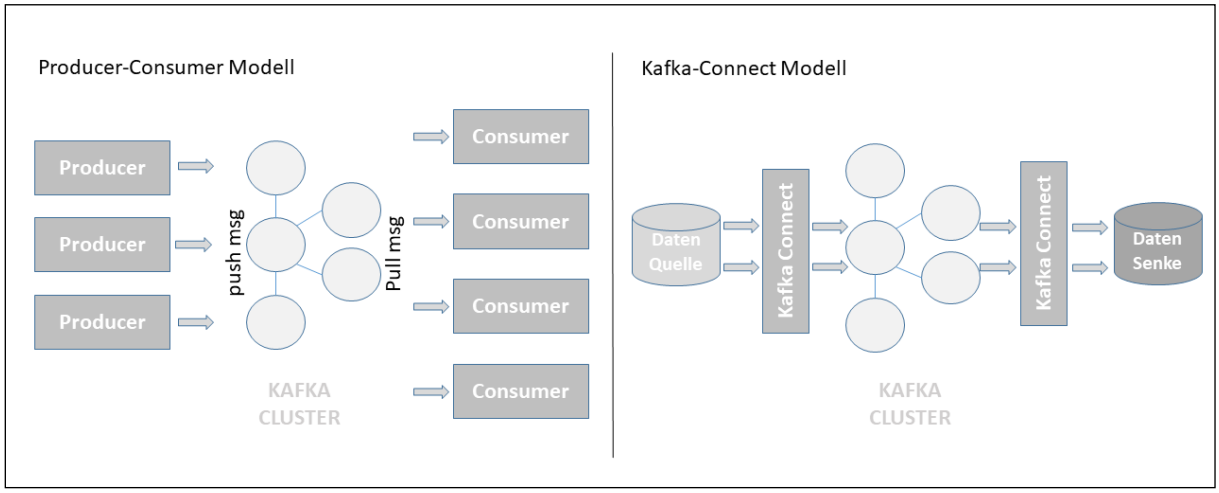


Abbildung 45: Kafka-Modelle (eigene Darstellung in Anlehnung an Siciliani, 2017)

4.3.2.2 Apache Nifi

Als weiteres Tool zur Datenbereitstellung bietet Apache Nifi seine Dienste an. Apache Nifi wurde wie auch das zuvor genannte Apache Kafka von der Apache Software Foundation entwickelt. Die eigentliche Technologie „Niagara Files“ wurde 2006 von der NASA entwickelt und im Jahr 2014 an die Apache Software Foundation als Open Source freigegeben. Damit beruht Apache Nifi auch auf dieser Grundlage. Seine Hauptfunktion ist der Datentransport zwischen Systemen, um einen Datenfluss zu gewährleisten. Darüber hinaus können Daten auch transformiert werden. NiFi kann in unternehmenskritischen Datenströmen mit strengen Sicherheits- und Compliance-Anforderungen eingesetzt werden, bei denen der gesamte Prozess visualisiert und Änderungen sofort und in Echtzeit vorgenommen werden können. Zum Zeitpunkt der Erstellung dieser Arbeit verfügt es über fast 200 sofort einsatzbereite Prozessoren (einschließlich Flume- und Kafka-Prozessoren), die per Drag & Drop gezogen, konfiguriert und sofort in Betrieb genommen werden können. Einige der Hauptmerkmale von NiFi sind die priorisierte Warteschlangenbildung, die Rückverfolgbarkeit von Daten und die Konfiguration von Gegendruckschwellenwerten pro Verbindung.

Obwohl es zur Erstellung fehlertoleranter Produktionspipelines verwendet wird, repliziert NiFi noch keine Daten wie Kafka. Wenn ein Knoten ausfällt, kann der Datenfluss zwar zu einem anderen Knoten geleitet werden, aber die am ausgefallenen Knoten in die Warteschlange

gestellten Daten müssen warten, bis der Knoten wieder hochkommt. NiFi ist weder ein vollwertiges ETL-Tool noch ideal für komplexe Berechnungen und Ereignisverarbeitung (CEP). Stattdessen sollte es eine Verbindung zu einem Streaming-Framework wie Apache Flink, Spark Streaming oder Storm herstellen.

4.3.2.3 Hadoop-Cluster

Inzwischen existieren verschiedenste IT-Lösungen zur Nutzung von Big Data. Die Anwendung Hadoop hat sich jedoch als Standard auf diesem Gebiet etabliert. Hadoop ist ein Open-Source-Projekt, das in der Sprache Java programmiert ist. Mit einer entsprechenden Apache-Lizenz ist es somit frei verfügbar. Es existieren verschiedene Software-Distributionen zur Ergänzung (vgl. Fels & Schinkel, 2015, S. 279) Hadoop hilft dort weiter, wo herkömmliche Data-Warehouse-Systeme an ihr Limit kommen. Der Grund dafür ist, dass Hadoop mit enormen Datenmengen umgehen und sowohl strukturierte als auch unstrukturierte Daten verarbeiten kann. Weiterhin kann Hadoop stark skalieren und eine performante Datenverarbeitung anbieten. Die Funktionen von Hadoop umfassen das Erfassen, das Speichern, die Organisation, das Suchen und das Analysieren von Daten auf einem Cluster von Rechnern. Dabei genügen Standardrechner, was zur Folge hat, dass sich die Kosten in Grenzen halten (vgl. Müller, 2016, S. 145). Hauptkomponenten von Hadoop sind das Hadoop Distributed File System (HDFS), das Hadoop MapReduce und der Yet Another Resource Negotiator (YARN). Diese werden nachfolgend erläutert.

Hadoop Distributed File System

Das HDFS ist ein verteiltes Dateisystem und arbeitet nach dem Master-Slave-Prinzip. Es ist vor allem für die Verarbeitung von hohen Datenvolumina sowie für eine hohe Verfügbarkeit in einem Cluster konzipiert (vgl. Fels & Schinkel, 2015, S. 279). Beim HDFS werden Dateien in binäre Blöcke aufgeteilt. Diese Blöcke werden repliziert, also kopiert, und auf mehreren Knoten verteilt, um bei Datenverlust korrekt weiterarbeiten zu können. Weiterhin ist somit auch eine parallele Bearbeitung einer bestimmten Datei durch verschiedene Maschinen möglich. Die Knoten, auf denen die Dateiblöcke gespeichert werden, heißen Data Nodes. Außerdem existiert ein Name Node bzw. Master-Knoten. Dieser verwaltet das komplette HDFS und koordiniert die Data Nodes. Ihm ist der Lagerort jeder Datei im HDFS bekannt. Er weiß, in welche Blöcke eine Datei unterteilt ist, über welche Zugriffsberechtigungen die Clients bzw. Benutzer

verfügen und kennt auch weitere Eigenschaften. Diese Eigenschaften werden in regelmäßigen Abständen vom RAM in Log-Dateien gewandelt. Diese Log-Dateien werden wiederum regelmäßig vom Secondary Name Node gelesen sowie zusammengefügt. Mit diesem Vorgang wird der jeweils aktuelle Dateibaum des HDFS abgebildet. Sobald ein Client Dateien auf das HDFS schreibt oder von diesem liest, findet als Erstes eine Kommunikation mit dem Name Node statt. Dies geschieht, damit der Name Node dem Client sagen kann, für welche Dateien er Lese- bzw. Schreibberechtigungen besitzt und wo sich die Dateiblöcke befinden bzw. auf welche Data Nodes sie geschrieben werden sollen (vgl. Fasel, 2016, S. 129). In der Abbildung 46 ist das HDFS schematisch dargestellt.

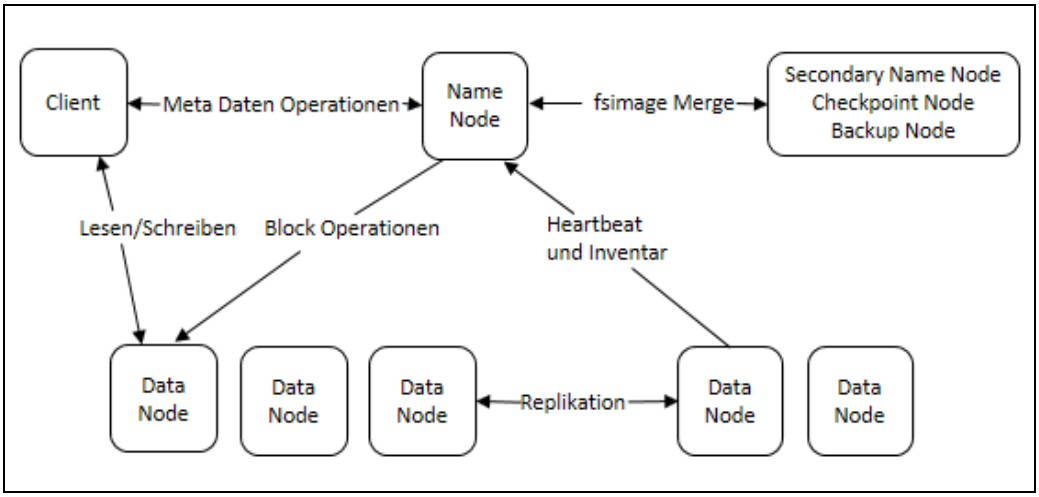


Abbildung 46: Schematische Darstellung der HDFS-Komponenten (eigene Darstellung in Anlehnung an Fasel, 2016, S. 130)

MapReduce-Engine

MapReduce ermöglicht die verteilte Verarbeitung von Daten, indem es sie in Teilaufgaben bzw. Map-Tasks aufteilt. Es ist nicht als eigenständiges Programm, sondern als Programmiermodell anzusehen. Grundsätzlich geht es darum, in der Map-Phase lokale Teilresultate zu berechnen und diese anschließend in der Reduce-Phase zusammenzuführen. Der MapReduce-Vorgang läuft in fünf Phasen ab, so wie in Abbildung 47 dargestellt. Bei der ersten Phase wird eine zu verarbeitende Datei innerhalb des HDFS aufgeteilt. In der zweiten Phase werden die Datensätze eingelesen und anschließend an die Map-Funktion übergeben. Die Map-Funktion generiert Schlüssel-Wert-Paare. Durch den Schlüssel wird definiert, zu welcher Partition bzw. zu welchem Bereich eines Datenträgers der Datensatz gehört. Der Partitioner führt eine Vorsortierung der Datensätze durch. Danach werden sie in einem intermediären Cache geschrieben. Der Cache befindet sich im lokalen Dateisystem der Maschine und nicht im HDFS. Die Phase 3 umfasst das Kopieren intermediärer Dateien auf die Maschine sowie das

Shuffling, also den Austausch von Zwischenergebnissen. Im abschließenden Reduce-Schritt können mehrere Partitionen gleichzeitig verarbeitet werden. Da Partitionen mit denselben Schlüsseln von verschiedenen Map-Tasks aus Phase 2 stammen können, muss in Phase 4 zunächst eine Sortierung nach gleichen Partitionen stattfinden. In der finalen Phase 5 werden die Ergebnisse der einzelnen Bearbeitungsschritte zusammengeführt und letztendlich als eine Datei in das HDFS übertragen (vgl. Fasel, 2016, S. 125–126).

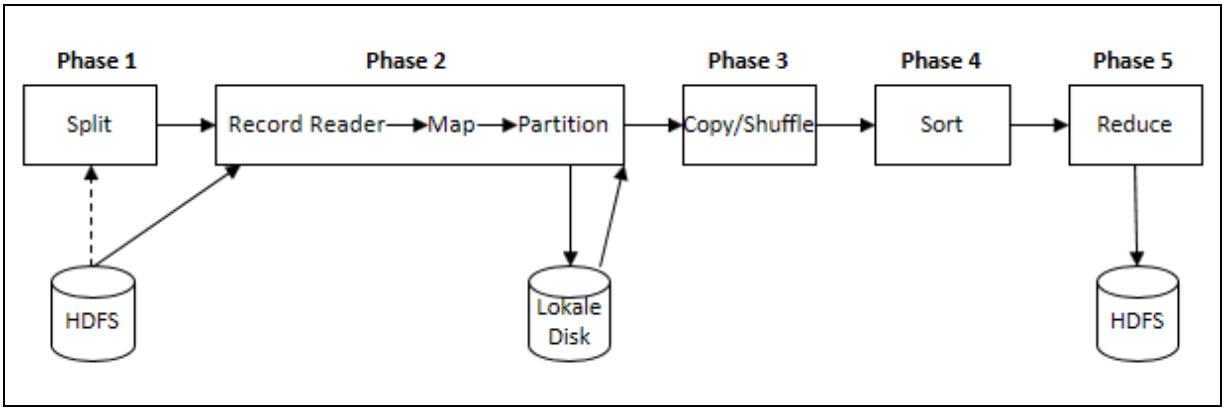


Abbildung 47: MapReduce-Prozess (eigene Darstellung in Anlehnung an Fasel, 2016, S. 126)

Yet Another Resource Negotiator

Der YARN kann als der Ressourcenverwalter von Hadoop definiert werden. Im Vergleich zum MapReduce, bei dem JobTracker und TaskTracker eigenständig funktionieren, ist das Modul YARN in drei Entitäten aufgeteilt:

- RessourceManager
- NodeManager
- ApplicationManager

Sobald eine Applikation in Hadoop startet, teilt der YARN dieser Applikation die notwendigen Ressourcen wie CPU, RAM, Festplatten und Netzwerk zu. Bis zur Beendigung einer Applikation wird diese vom YARN überwacht und verwaltet. Eine Einheit von Ressourcen wird als Container bezeichnet. Ein Container wird als ein CPU-Kern plus eine festgelegte Menge RAM definiert. Jeder Data Node verfügt über einen NodeManager, der die Container des Data Node verwaltet. Außerdem existiert ein RessourceManger, der mit dem NodeManager kommuniziert und im gesamten Hadoop die Ressourcen verwaltet (vgl. Fasel, 2016, S. 130–131).

Die Abbildung 48 stellt den Ablauf der YARN-Vorgänge dar. Sobald ein Client bzw. Benutzer eine Applikation startet, werden zunächst die nötigen Informationen an den RessourceManager

übergeben. Über einen NodeManager wird dann ein Container für das Überwachen der Applikation gestartet. Dies geschieht, um die Arbeitslast des RessourceManagers zu verteilen. Außerdem kann damit verhindert werden, dass beim Absturz des RessourceManagers ausgeführte Applikationen in Mitleidenschaft gezogen werden. Anschließend startet der ApplicationManager und analysiert die Informationen der Applikationen, um daraus abzuleiten, welche zusätzlichen Ressourcen vom RessourceManager angefordert werden müssen. Der RessourceManager wiederum sendet eine Liste mit verfügbaren Containern auf Data Nodes. Sobald diese Container von den NodeManagern freigegeben werden, starten die einzelnen Tasks der Applikationen bei den zugesicherten NodeManagern. Zwischenstände und der finale Stand werden an den ApplicationManager zurückgemeldet. Mit den beschriebenen Abläufen kann ein effizientes und ausfallsicheres Ausführen von Applikationen erreicht werden (vgl. Fasel, 2016, S. 130–131).

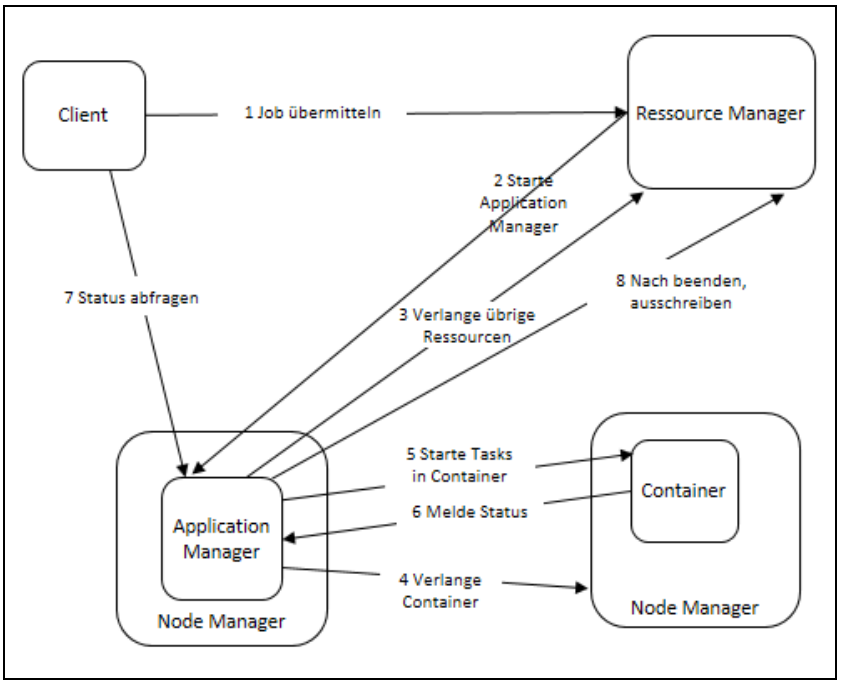


Abbildung 48: Schematische Darstellung der YARN-Komponenten (eigene Darstellung in Anlehnung an Fasel, 2016, S. 131)

Neben den zuvor genannten Hauptkomponenten von Hadoop existieren weitere Unterprojekte von Apache, mit denen Hadoop erweitert werden kann. Mit deren Hilfe kann Hadoop zu einer universell einsetzbaren Plattform gemacht werden, mit der verschiedenste Analyseanwendungen durchgeführt werden können (vgl. Fels & Schinkel 2015, S. 286). In Tabelle 10 sind einige Unterprojekte aufgeführt und knapp erläutert.

Unterprojekt	Funktion
Pig	Skriptsprache zur Analyse großer Datenmengen
Hive	Ad-hoc-Abfragen und Reports erstellen mit geringem Aufwand, sowie Datenimport aus HDFS
Sqoop	Datenaustausch zwischen relationalen Datenbanksystemen und Hadoop
Flume	Bereitstellung von Datenströmen aus z.B. Web-Logs und anderen Protokolldaten für die Analyse
Avro	Serialisierung strukturierter Daten, indem strukturierte Daten in Bitketten konvertiert und als kompaktes Format im HDFS abgelegt werden
Mahout	Bibliothek mit Modulen für das Maschinelle Lernen
ZooKeeper	Bibliothek mit Modulen zur Integration von Koordinations- und Synchronisierungsdiensten für ein Hadoop-Cluster, die von anderen Unterprojekten verwendet werden
Oozie	Beschreibung und Automatisierung von Abläufen
Chukwa	Überwachung und Visualisierung von Hadoop-Umgebungen
Ambari	Verwaltung von Hadoop-Umgebungen auf einem Web-Interface
Spark	Datenverarbeitungsplattform für parallele Verarbeitung großer Datenmengen mittels In-Memory-Technologie

Tabelle 10: Übersicht der Hadoop-Unterprojekte (eigene Darstellung in Anlehnung an Fels & Schinkel, 2015, S. 286–288)

Kommerzielle Software-Lösung

Nachfolgend wird zum Vergleich eine kommerzielle Software-Lösung hinsichtlich der Big-Data-Analyse vorgestellt. Vor dem Hintergrund der Kooperation des Projektpartners bzw. als Kunde von SAP ist eine kurze Erläuterung der Plattform auf Basis des Produkts S4/HANA interessant. Einerseits ist das Produkt aus Sicht der Datenanalyse aus einer zeitlichen Betrachtung heraus weniger relevant und andererseits ist die technologische Prozessbetrachtung aus Datenperspektive ähnlich zur aufgestellten Big-Data-Architektur. Aus der technischen Betrachtung existieren im S/4 weitaus weniger Tools zur Datenverarbeitung als es eine Big-Data-Architektur gewährleisten kann. Die Einführung eines neuen ERP-Systems soll ab 2022 stufenweise eingeleitet werden, sodass dieses Thema der kommerziellen Software-Lösung nicht weiter vertieft wird.

SAP HANA ist eine vielseitige Plattform, die für verschiedenste prozessspezifische Anwendungsfälle in einem Unternehmen eingesetzt werden kann. Bei SAP HANA handelt es sich um eine Kombination aus Hardware- und Software-Technologien. Es eignet sich besonders für die Datenanalyse im Rahmen von kaufmännischen Prozessen und Vorgängen bezüglich der Supply Chain. SAP HANA nutzt die Möglichkeiten aktueller Hardware-Systeme im vollen Umfang. Dadurch kann die Leistungsfähigkeit der laufenden Applikationen stets gesteigert werden. Separate Altsysteme können in SAP HANA zusammengeführt werden, wodurch die Komplexität der IT-Systeme eines Unternehmens sinkt und sich Bearbeitungszeiten deutlich verkürzen. SAP HANA verwaltet alle Daten zentral, weshalb keine Datenhaltung mehr auf verschiedenen Datenbanksystemen notwendig ist. Szenarien, welche bisher aus technischen oder finanziellen Gründen nicht umsetzbar waren, sind nun mit SAP HANA realisierbar. SAP HANA ermöglicht bspw. die Realisierung von Internet-of-Things-Projekten und die Umsetzung von Big-Data-Anwendungsfällen (vgl. Prassol, 2016, S. 198).

In Abbildung 49 ist die Architektur der SAP-HANA-Plattform dargestellt. Kernkomponente der SAP-HANA-Plattform ist die SAP-HANA-Datenbank. Diese ist ein spaltenorientiertes In-Memory-Datenbanksystem. Mittels Integrationsservices wird von der Plattform auf unterschiedliche Datenquellen zugegriffen. Die gesammelten Daten werden mittels verschiedener Algorithmen, Engines und Logiken ausgewertet und aufbereitet. An der Plattform angekoppelte Anwendungen können auf die Resultate in Echtzeit zugreifen (vgl. Prassol, 2016, S. 198).

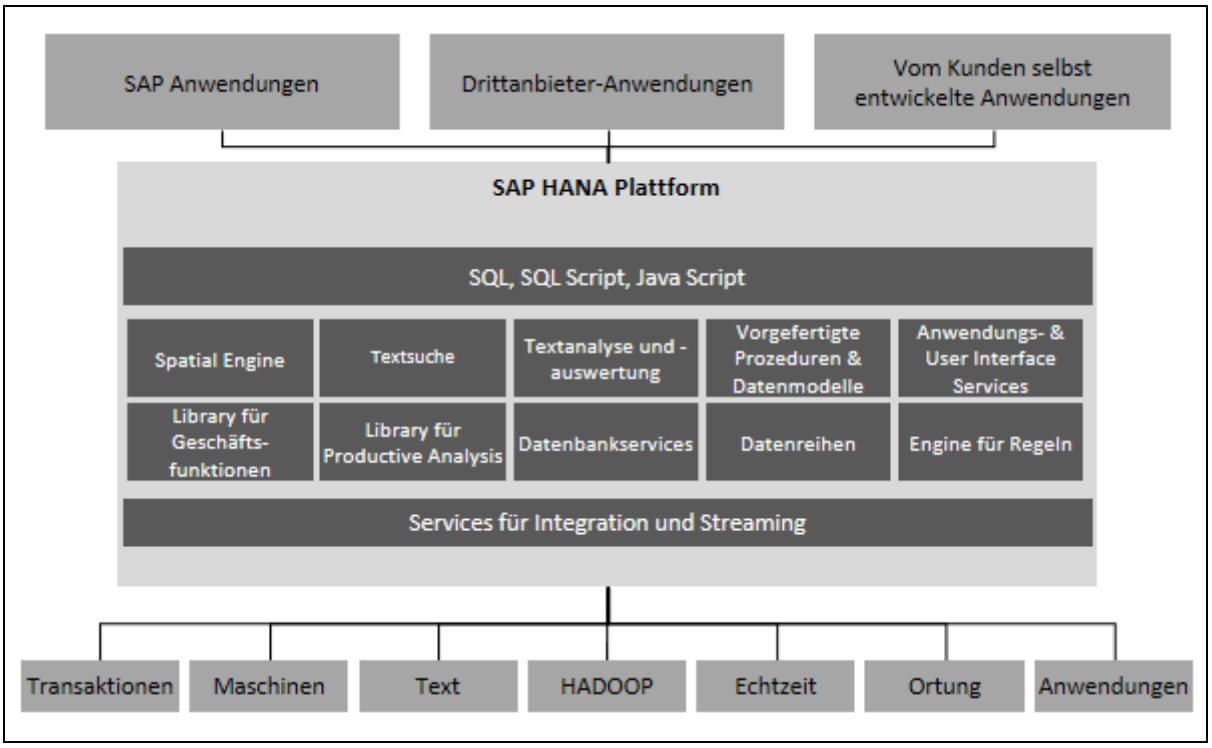


Abbildung 49: Architektur der SAP-HANA-Plattform (in Anlehnung an Prassol, 2016, S. 199)

Die SAP-HANA-Plattform kann in vielfältiger Weise genutzt werden. Eine Möglichkeit ist der Einsatz als flexibler Data Mart, um zeitnah Antworten auf unternehmenskritische Fragen zu generieren. Außerdem ist ein Einsatz als Accelerator möglich, damit die Performance von SAP-Applikationen gesteigert werden kann. Dafür werden Operationen mit intensiver Datenbanknutzung nicht über die primäre SAP-Datenbank abgewickelt, sondern über SAP HANA. Eine weitere Variante ist es, SAP HANA als komplette Datenbank von SAP- und Nicht-SAP-Applikationen zu nutzen. In SAP HANA können auch Anwendungen entwickelt und betrieben werden. Außerdem können auch alle zuvor beschriebenen Nutzungsmöglichkeiten kombiniert und somit eine umfassende Datennutzung realisiert werden (vgl. Prassol, 2016, S. 204–205).

4.4 Analysemöglichkeiten

Wenn Daten analysiert werden sollen, gibt es dafür einen bestimmten Grund bzw. bestehen bestimmte Ziele, die damit verfolgt werden. Die Analyse der Big Data kann auf verschiedensten Zielen beruhen. Je nach individuellem Ziel werden bestimmte Erkenntnisse aus der Datenanalyse erwartet. Generell wird angestrebt, besonders nutzbringende Informationen aus den riesigen Datenmengen zu filtern. Ein Beispiel dafür sind Vorhersagen. Ziel ist es hier,

Trendaussagen zu geben, die auf Vergangenheitsdaten beruhen. Ein weiteres Ziel kann es sein, *schwache Signale* aus den Datenmengen herauszufiltern. Gemeint sind damit bspw. spezielle Informationen, die sehr gewinnbringend für ein Unternehmen sein können, aber in herkömmlichen Suchmaschinen nicht wahrgenommen werden, da sie im Internet noch nicht intensiv genug vertreten, mit anderen Seiten verknüpft oder nicht in einem speziellen Kontext aufgetreten sind. Außerdem kann es ein Ziel sein, neue Erkenntnisse festzustellen. Gemeint ist damit das unbekannte Unwissen. Diese Idee besagt, dass man eine Information nicht kennt, sich aber nicht dessen bewusst ist, dass man diese Information nicht kennt.[4] Da das unbekannte Unwissen eine Gefahr für das Unternehmen darstellen kann, ist es sinnvoll, es aufzudecken. Weiterhin kann nach Relationen von Daten gesucht werden. Damit ist das Analysieren komplexer Inhalte und Zusammenhänge zwischen Daten gemeint. Als Resultat sollen Relationen, Ähnlichkeiten und Verknüpfungen zwischen Daten festgestellt werden (vgl. Schulmeyer, 2015, S. 323–328).

Entsprechend dem Ziel, das erreicht werden soll, müssen passende Methoden bzw. Analyseaufgaben ausgewählt werden. Die Analyseaufgaben werden nach prädiktiven und beschreibenden Aufgaben unterschieden. Diese umfassen jeweils verschiedene Varianten.[5] Prädiktiven Verfahren ist gemeinsam, dass eine Zielgröße bzw. ein Zielattribut vorliegt. Das Ziel ist bei allen Verfahren, unbekannte Werte für die Zielgröße bzw. das Zielattribut vorherzusagen. Ein anderes Ziel kann sein, Zusammenhänge von Daten zu bestimmen. Zur Umsetzung werden überwachte Lernverfahren genutzt. Für diese werden Lernbeispiele bzw. Trainingsdaten gebraucht, die das Programm als Grundlage nutzt (vgl. Lanquillon & Mallow, 2015a, S. 64). Die beschreibenden Analyseaufgaben umfassen alle Bereiche, die nicht durch die prädiktiven Analyseaufgaben abgedeckt werden. Mit ihnen sollen Ausprägungen eines Zielattributs bestimmt werden. Außerdem ist es möglich, mit bereits bekannten Zielattributen aus dem prädiktiven Bereich Zusammenhänge bewerten zu lassen. Beschreibende Analyseaufgaben können in Form überwachter und unüberwachter Methoden ausgeführt werden. Der Unterschied ist, dass unüberwachte Methoden ohne Nutzung eines speziellen Zielattributs arbeiten (vgl. Lanquillon & Mallow, 2015a, S. 66). Nachfolgend werden die Varianten kurz vorgestellt. Auf eine detaillierte Betrachtung der Umsetzung der Verfahren wird verzichtet, da diese teilweise auf komplexen mathematischen Funktionen beruhen. Diese

[4] Das bekannte Unwissen hingegen sagt aus, dass einem bewusst ist, eine bestimmte Information nicht zu kennen. Dieser Fall ist weniger kritisch, da man gezielt nach dieser Information suchen kann, um diese Wissenslücke zu schließen (Schulmeyer 2015, S. 327).

[5] Für die Datenanalyse an sich existiert eine Vielzahl an Analysemöglichkeiten. Nicht alle Verfahren werden als geeignet für Big Data bzw. sehr große Datenmengen angesehen. In diesem Kapitel wird die Auswahl von Lanquillon und Mallow behandelt, da sie als geeignet erweist.

intensiv zu betrachten ist für diese Arbeit nicht notwendig, weil einerseits Big-Data-Technologien die Umsetzung dieser Verfahren bereits grundlegend implementiert haben, andererseits das darauffolgende Machine Learning die Analyse unterstützt.

4.4.1 Machine Learning

Der nächste Schritt ist eine Implementierung von Machine-Learning-Ansätzen. Damit ist eine automatisierte Datenbeschaffung für die darauffolgende Analyse durch Maschinelles Lernen möglich. Daraufhin werden drei grundlegende Lernverfahren erörtert. Mit Machine Learning sollen das Big-Data-Konzept und die Analysen optimiert werden. Dabei soll der Schwerpunkt auf die Produktion, speziell in der Instandhaltung, gelegt werden.

Maschinelles Lernen wird nach MITCHEL wie folgt definiert: „study of computer algorithms that improve automatically through experience." (Mitchel, 1997, S. 2) Demnach ist Machine Learning ein Ansatz, bei dem ein Programm selbstständig anhand von Beispielen mit entsprechenden Regeln lernt. Als Beispiel ist die Bilderkennung zu nennen. Anhand von Bildern soll der Algorithmus Äpfel von Birnen unterscheiden. Das Training des Algorithmus besteht aus einer Sammlung von Bildern, auf denen Äpfel und Birnen zu erkennen sind. Diese Beispiele werden dem Algorithmus vorgelegt. Das Lernen der Struktur und der Form von Äpfeln und Birnen wird nun in Gang gesetzt. Im maschinellen Lernprozess wird ein Modell aus dem bestehenden Datensatz erstellt. Darin werden Wissen und Regeln aus den Merkmalen extrahiert. Diese werden Features genannt. Der Prozess wird als Lernen oder Training definiert. Der Algorithmus führt nun den Lernprozess aus (vgl. Zhou, 2012, S. 20). Anschließend ist das trainierte Modell in der Lage, das gewonnene Wissen auf neue Datensätze anzuwenden (vgl. Brink, Richards & Fetherolf, 2016, S. 2). Neben der Bilderkennung existieren auch andere erfolgreiche Anwendungen des Maschinellen Lernens in weiteren Bereichen. Kommerzielle Systeme werden zur Erkennung von Sprache und Handschrift eingesetzt. Handelsunternehmen analysieren ihre Verkaufsdaten, um das Verhalten ihrer Kunden zu verstehen und das Kundenbeziehungsmanagement zu verbessern. Finanzinstitute durchleuchten vergangene Transaktionen, um Kreditrisiken oder potenzielle Betrugsfälle von Kunden vorherzusagen (vgl. Alpaydin, 2010, S. 31). Darüber hinaus verwenden Industrieunternehmen Algorithmen des Maschinellen Lernens, um Fehler zu erkennen und Ausfälle ihrer Maschinen vorherzusagen. Maschinelle Lernverfahren lassen sich nun in vier verschiedene Kategorien einteilen:

überwachtes Lernen, unbeaufsichtigtes Lernen, teilüberwachtes Lernen und Verstärkungslernen, auch Reinforcement Learning genannt (vgl. Suthaharan, 2016, S. 126).

Überwachtes Lernen

Beim überwachten Lernen lernt ein Algorithmus eine Zielfunktion, die zur Vorhersage der Werte eines diskreten oder kontinuierlichen Attributs verwendet werden kann. Der Lernprozess basiert auf Eingabedaten (typischerweise Vektoren), deren Ergebnis bereits bekannt ist. Daher werden den Eingabedaten einfache Werte zugewiesen, die Labels genannt werden, die das wirklich gewünschte Ergebnis anzeigen (vgl. Kaur & Jindal, 2016, S. 7). Das Ziel ist die Schätzung der Ergebnisvariablen mit dem gegebenen Satz unabhängiger Eingangsvariablen. Während des Trainings versucht der Algorithmus, eine Beziehung oder ein Muster zwischen Input und Output zu finden und eine Funktion aufzubauen, die den Input auf dem gewünschten Output abbilden kann (vgl. Alpaydin, 2010, S. 9). Der Trainingsprozess setzt sich in iterativer Weise fort, während das Modell eine gewisse Genauigkeit der Trainingsdaten erreicht. Nach der Trainings- (oder Lern-)Phase folgt das Testen des Modells, wobei nicht offenbarte Testdaten zur Beurteilung der Modellgenauigkeit verwendet werden. Durch den kontinuierlichen Einsatz von Versuch und Irrtum lernt und verbessert sich die Maschine aus ihren bisherigen Erfahrungen und versucht, das bestmögliche Wissen zu erfassen (vgl. Kaur & Jindal, 2016, S. 7). Schließlich ist sie in der Lage, genaue Vorhersagen über neue, ungesehene Daten zu treffen. Je nach Charakteristik des Outputs wird das betreute Lernen im Allgemeinen in Regression und Klassifikation unterteilt. Erstere bestimmt eine diskrete Klasse als Output. Sie wird häufig zur Spam-Filterung, zur Erkennung von Herstellungsfehlern, zur Betrugserkennung oder zur Bilderkennung verwendet. Beliebte Algorithmen, die zur Klassifikation verwendet werden, sind logistische Regression, Entscheidungsbäume oder Support-Vektor-Maschinen. Im Gegensatz dazu ist die Regression die Vorhersage eines kontinuierlichen Wertes auf der Grundlage von Eingabedaten. Sie wird in verschiedenen Bereichen wie Börsenprognosen, Nachfrageprognosen, Preisschätzungen oder Ad-Bid-Optimierung eingesetzt und beruht auf Algorithmen wie der linearen oder multivariaten Regression (vgl. Brink et al., 2016, S. 4). Abbildung 15 veranschaulicht den Hauptunterschied zwischen Klassifizierung und Regression. Die gelernte Funktion während eines Klassifizierungsprozesses dient als Entscheidungsgrenze, die zwei Klassen trennt, während die Funktion während der Regression verwendet wird, um den Ausgabewert auf der Grundlage der Eingabevariablen zu extrapolieren oder zu interpolieren.

Unüberwachtes Lernen

Beim unüberwachten Lernen erhält die Maschine lediglich Eingabedaten, erhält aber weder überwachte Zielausgaben noch Belohnungen aus ihrer Umgebung (vgl. Ghahramani, 2003, S. 74). Es gibt keine Etiketten, die mit einem bestimmten Datenpunkt verbunden sind. Stattdessen besteht das Ziel darin, ein Modell zu entwickeln, das in der Lage ist, unbekannte Strukturen innerhalb der Daten zu entdecken und nützliche Informationen zu extrahieren. Es wird der Begriff des „unüberwachten“ Lernens verwendet, da es keinen „Supervisor"“ gibt, der den Output verifiziert. Daher liefern Algorithmen für unüberwachtes Lernen kein Feedback auf der Grundlage der Vorhersageergebnisse, was es subjektiver macht als das überwachte Lernen (vgl. Galar, 2017, S. 174). Es ist jedoch möglich, formale Rahmen für unbeaufsichtigtes Lernen zu entwickeln, die Darstellungen des Inputs aufbauen, die für die Entscheidungsfindung oder die Vorhersage künftiger Inputs verwendet werden können. Vorgefundene Muster oder Gruppen in den Daten werden vom menschlichen Auge als reines unstrukturiertes Rauschen betrachtet (vgl. Ghahramani, 2003, S. 74). Unüberwachte Lernalgorithmen werden in verschiedenen praktischen Anwendungen eingesetzt, z.B. bei der Suche nach auffälligen Genexpressionsmessungen von Brustkrebspatientinnen, bei der Charakterisierung von Kundengruppen durch ihre Browsing- und Kaufhistorie oder bei der Strukturierung gespeicherter Daten, um entbehrliche Informationen für die Kompression oder Visualisierung zu verwerfen (vgl. Galar & Kumar, 2017, S. 175). Generell lassen sich zwei Methoden des unüberwachten Lernens unterscheiden:

Reduktion der Dimensionalität:

Während komplexe Daten wie Bilder oder Spektrogramme durch Punkte in einem hochdimensionalen Vektorraum dargestellt werden können, haben sie in der Regel eine wesentlich kompaktere Beschreibung. Eine kohärente Struktur in Daten führt zu starken Korrelationen zwischen Eingaben (vgl. Roweis & Saul, 2000, S. 2323). Mit Nutzung dieser Beobachtung werden Techniken zur Reduzierung der Dimensionalität eingesetzt, um eine hilfreichere Darstellung von Informationen zu erhalten. Diese Methoden reduzieren die Anzahl der Dimensionen, um die Rechengeschwindigkeit zu erhöhen, Platz zu sparen oder die Visualisierung der Daten zu erleichtern. Gleichzeitig wird versucht, den während dieses Übergangs auftretenden Informationsverlust zu minimieren. Die Reduzierung der Dimensionalität befasst sich im Allgemeinen mit Skalierungsproblemen, die mit einem massiven Datensatz und einer großen Anzahl von Klassen verbunden sind (vgl. Suthaharan, 2016, S. 329). Beliebte Techniken zur Dimensionalitätsreduktion sind die

Hauptkomponentenanalyse (Principal Component Analysis, PCA) oder die lineare Diskriminanzanalyse (Linear Discriminant Analysis, LDA).

Clusterbildung:

Die formale Untersuchung von Algorithmen zur Gruppierung oder Gruppierung von Objekten nach gleichen Merkmalen wird als Clusteranalyse bezeichnet (vgl. Khanum, Mahboob, Imtiaz, Ghafoor & Sehar, 2015, S. 35). Während eines Clustering-Prozesses werden die Daten in verschiedene Gruppen unterteilt, sodass jede Gruppe durch eine Art Ähnlichkeit gekennzeichnet ist. Die Definition dessen, was „ähnlich“ oder „verschieden“ bedeutet, stellt daher einen entscheidenden Faktor für das Ergebnis dar (vgl. Wagner & Wagner, 2007, S. 1). Das Ähnlichkeitsmaß muss im Hinblick auf das zugrundeliegende Problem ausgewählt werden. Abbildung 50 veranschaulicht die Verwendung eines einfachen geradlinigen Abstands als Ähnlichkeitsmaß für zweidimensionale Datenpunkte mit dem populärsten und einfachsten Clustering-Algorithmus K-Mittel.

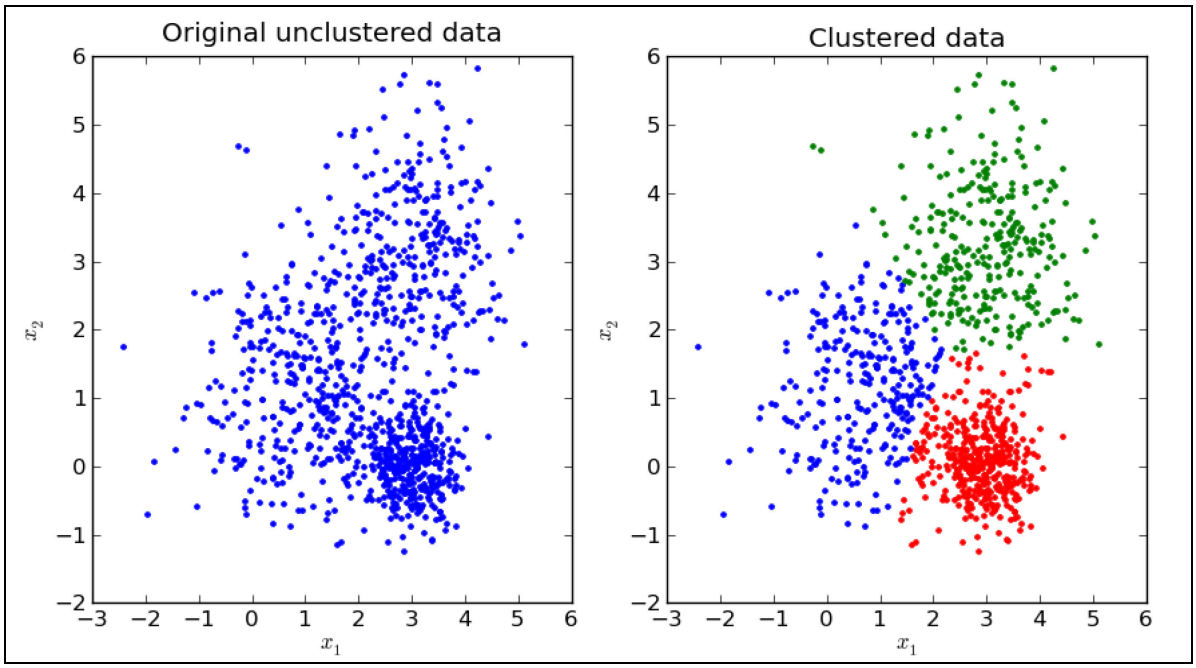


Abbildung 50: K-means Clustering (mubaris.com)

Semi-supervised Learning

Wenn es um Klassifizierungsprobleme geht, gehen die meisten Forschungsarbeiten von der Annahme aus, dass es große Mengen an beschrifteten Trainingsdaten gibt. In Wirklichkeit kann der hohe Aufwand für die Markierung der Daten eine solche Annahme jedoch ungültig machen (vgl. Chen et al., 2013, S. 2). So sind bspw. Zeit und Fachwissen eines Kardiologen erforderlich, um einzelne Herzschläge in einer EKG-Datenspur zu beschriften, aber ein einziger

Schlafstudientest kann bis zu 40.000 solcher Herzschläge erzeugen (vgl. Chazal, O'Dwyer & Reilly, 2004, S. 1196). Das Sammeln unmarkierter Daten ist quantitativ wertvoll, bspw. sind im Internet unzählige Bücher, Bilder, Karten und historische Manuskripte verfügbar, von denen viele als Trainingsbeispiele für überwachte Klassifikationsmodelle verwendet werden, wenn nur diese Daten beschriftet sind (vgl. Chen et al., 2013, S. 2). Diese ermöglichen den Einsatz von Algorithmen des semi-supervised Learning (SSL). SSL liegt zwischen überwachtem und unbewachtem Lernen und basiert auf dem Prinzip, sowohl aus gekennzeichneten als auch aus nicht gekennzeichneten Daten zu lernen (vgl. Wei & Keogh, 2006, S. 2). Der Ansatz lässt sich in fünf Klassen unterteilen: SSL mit generativen Modellen, SSL mit Low-Density-Separation, graphenbasierte Methoden, Co-Training-Methoden und Selbstkennzeichnungsmethoden (vgl. Chapelle, Schölkopf & Zien, 2006, S. 1).

- **Generative Modelle:**

Generative Modelle sind der älteste teilüberwachte Lernansatz. Sie gehen davon aus, dass die Daten in Mustern angeordnet sind, die auf einer Mischverteilung basieren, die durch große Mengen unmarkierter Daten identifiziert werden können. Ein Vorteil der generativen Methode besteht darin, dass das Wissen über die Struktur der Daten auf natürliche Weise in das Modell einfließen kann. Sie wurde in mehreren Bereichen erfolgreich eingesetzt, darunter Textklassifikation und Gesichtsorientierungsunterscheidung (vgl. Wei & Keogh, 2006, S. 2).

- **Trennmodelle mit niedriger Dichte:**

Low-Density-Trennungsansätze versuchen, die Annahme, dass die Entscheidungsgrenze in einer Region mit geringer Dichte liegen sollte, dadurch auszugleichen, dass sie die Entscheidungsgrenze von den unmarkierten Daten wegdrücken (vgl. Chapelle et al., 2006, S. 6). Dieses Ziel wird in der Regel durch die Verwendung eines Margenmaximierungsalgorithmus wie Support Vector Machines (SVM) erreicht.

- **Graphenbasierte Modelle:**

Graphenbasierte SSL-Verfahren gehen davon aus, dass „die (hochdimensionalen) Daten (grob) auf einer niedrigdimensionalen Mannigfaltigkeit liegen" (vgl. Wie & Keogh, 2006, S. 3). Solche Ansätze stellen die Daten nach Knoten in einem Graphen dar, dessen Kanten die Abstände zwischen den Knoten sind. Das Hauptproblem dieser Methode besteht darin, dass die Graphenkonstruktion für jeden Bereich spezifisch angepasst werden muss, da sie Vorwissen codiert (vgl. Blum & Chawla, 2001, S. 22).

- **Co-Training-Modelle:**

Die Idee des Co-Trainings wurde zuerst von Blum und Mitchell vorgeschlagen (vgl. Blum & Mitchell, 1998, S. 92–100). Dabei werden die Merkmale der Daten in zwei disjunkte Sätze unterteilt, wobei jeder Satz ausreicht, um einen guten Klassifikator zu trainieren. Zwei Klassifikatoren werden separat für jede Merkmalsuntermenge trainiert, und die Vorhersagen des einen Klassifikators werden zur Erweiterung der Trainingsmenge des anderen verwendet.

- **Selbstbeschriftung:**

Die Selbstetikettierung oder das Selbsttraining ist ein iterativer Prozess, bei dem ein Lernender die Etiketten der Beispiele eingibt, die im vorhergehenden Schritt mit dem höchsten Vertrauen klassifiziert wurden (vgl. Chapelle et al., 2006, S. 3). Zuerst wird ein Klassifikator mit einer kleinen Menge verfügbarer beschrifteter Daten trainiert. Dann klassifiziert er die ungekennzeichneten Daten und fügt die am sichersten klassifizierten Beispiele (zusammen mit ihren vorhergesagten Kennzeichnungen) in den Trainingssatz ein. Das Verfahren wird so lange wiederholt, bis das Hinzufügen neuer Objekte zum markierten Satz die Genauigkeit des Klassifikators nicht mehr erhöht oder einige andere Anhaltekriterien erfüllt sind (vgl. Chen et al., 2013, S. 2). Zusammengefasst: Der Klassifikator bedient sich seiner eigenen Vorhersagen, um sich selbst zu unterrichten. Im Bereich der Zeitreihendaten hat die Technik des Selbsttrainings die vielversprechendsten Ergebnisse gezeigt.

Reinforcement Learning

Es gibt Anwendungen, bei denen das Ergebnis eine Folge von Aktionen ist. In einem solchen Fall ist eine einzelne Aktion eher unbedeutend, während es auf die richtige Reihenfolge der Aktionen ankommt. Daher wird nicht das Zwischenergebnis einer einzelnen Handlung bewertet, sondern die Eignung für die aktuelle Politik der Handlungen. In einem solchen Fall kann maschinelles Lernen angewandt werden, um die optimale Politik zu finden. Lernmethoden, die zur Lösung dieser Aufgaben entwickelt werden, gehören in den Bereich des Reinforcement Learning (vgl. Alpaydin, 2010, S. 13). Reinforcement-Learning-Modelle können durch eine Interaktion mit ihrer Umgebung trainiert werden. Das Training erfolgt durch einen sogenannten Agenten, der aus den Folgen seiner Handlungen lernt. Die Auswahl der Handlungen basiert entweder auf früheren Erfahrungen (Ausbeutung) oder auf neuen Wahlmöglichkeiten (Erforschung). Dieser Prozess wird oft als Trial-and-Error-Learning bezeichnet (vgl. Ertel, 2016, S. 316). Nach einer Teilsequenz von Aktionen erhält der Agent ein Verstärkungssignal, d.h. eine numerische Belohnung, die den Erfolg einer angewandten

Politik bestimmt. Der Algorithmus versucht, die in einer gegebenen Situation am besten geeigneten Aktionen zu finden, um diese Belohnung zu maximieren (vgl. Bishop, 2006, S. 3). Die beiden häufigsten Bereiche des angewandten Verstärkungslernens sind Spiele (z.B. Schach oder Go) (vgl. Alpaydin, 2010, S. 13) und Roboternavigation (vgl. Ertel, 2016, S. 313).

4.4.2 Apache Spark

Apache Spark wird in der Arbeit als zentrales Werkzeug für die Datenanalyse und das Predictive Maintenance genutzt und wird daher in diesem Abschnitt näher erklärt. Apache Spark ist ein Framework, das auf Basis von Hadoop entwickelt wurde. Das Framework wurde speziell für die Verarbeitung von Big Data konzipiert. Spark arbeitet mit der In-Memory-Technologie [6] und kann daher kürzere Bearbeitungszeiten als Hadoop erreichen (siehe Abbildung 51).

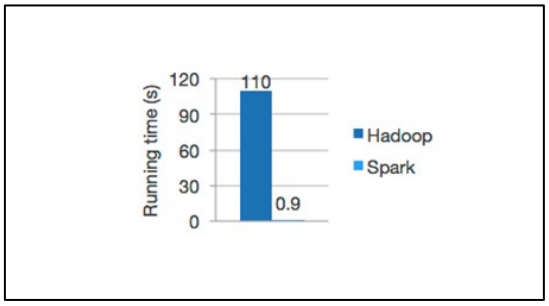


Abbildung 51: Geschwindigkeit Logistische Regression (Apache, 2018c, S. 1)

Dementsprechend höher ist dabei aber auch der hardwaretechnische bzw. finanzielle Aufwand. Dort, wo ein hoher Anspruch an Echtzeitanalysen gestellt wird, wäre die Nutzung von Spark zu empfehlen (Apache, 2018c, S. 1). Spark kann sowohl als Stand-Alone-Software genutzt werden als auch in einem Hadoop Cluster fungieren. Letzteres ist hervorzuheben, da Spark auf das Hadoop-HDFS zugreifen kann und daraus eine Datenlokalität [7] beim Lesen der Daten gewährleistet wird. Spark basiert auf gezieltem Caching von Zwischenergebnissen. Das Caching funktioniert im Speicher oder auf Festplatten mit verschiedenen Replikationsgraden. Für den Fall, dass der Speicher nicht ausreicht, wird das Caching über Spark nur zum Teil ausgeführt und ausgelagert. Bei Bedarf erfolgt eine Neuberechnung der Daten. Das bietet den

[6] Die In-Memory-Datenbank ist ein Datenbank-Managementsystem, das die Daten direkt im Arbeitsspeicher verarbeitet und somit höhere Zugriffsgeschwindigkeiten gewährleistet als herkömmliche Datenbanken (vgl. Litzel, 2017, S. 1).

[7] Datenlokalität heißt, dass einmal gespeicherte Daten in dieser Quelle bleiben, und durch Befehle oder Algorithmen darauf zugegriffen wird. Dadurch entfallen ETL-Jobs samt deren Verzögerungen und es bilden sich keine Datensilos zwischen den Abteilungen mehr. Stattdessen entsteht eine einzige Quelle.

Vorteil, dass iterative Prozesse im Machine Learning ausgeführt werden können (vgl. Aunkofer, 2016, S. 1). Die Apache-Spark-Architektur besteht aus folgenden Einzelkomponenten:

- Spark Core
- Spark SQL
- Spark Streaming
- MLib Machine Learning Library
- GraphX

Das Cluster basiert auf dem Spark Core. Die Aufgaben des Spark Core sind die Bereitstellung grundlegender Funktionalitäten, Verteilung von Aufgaben und Steuerung der Ein- und Ausgabeprozesse (vgl. Luber & Litzel, 2016). Die Daten werden im Spark-Prozess als Resilient[8] Distributed Datasets (RDD) verarbeitet und gespeichert. RDD ist eine Collection-Klasse, die im Cluster verteilt operiert. Zwischenergebnisse werden gespeichert und können an herkömmliche Datenbanken oder auch NoSQL-Datenbanken weitergeleitet werden. Ein RDD stellt eine unveränderliche, partitionierte Sammlung von Elementen dar, die parallel bearbeitet werden können. Diese Klasse enthält die grundlegenden Operationen, die auf allen RDDs verfügbar sind, z.B. Map, Filter und Persist.

Intern ist jedes RDD durch fünf Haupteigenschaften gekennzeichnet:

- Eine Liste von Partitionen,
- Eine Funktion zur Berechnung jeder Teilung,
- Eine Liste der Abhängigkeiten von anderen RDDs,
- Optional ein Partitionierer für Schlüsselwert-RDDs (z.B., um zu sagen, dass die RDD hash-partitioniert ist),
- Optional eine Liste bevorzugter Orte zur Berechnung jeder Teilung (z.B. Blockorte für eine HDFS-Datei).

In der Abbildung 52 ist ein Resilient Distributed Dataset dargestellt, woraus auch der Prozess der Transformation zu entnehmen ist.

[8] Resilient bedeutet „fehlertolerant“.

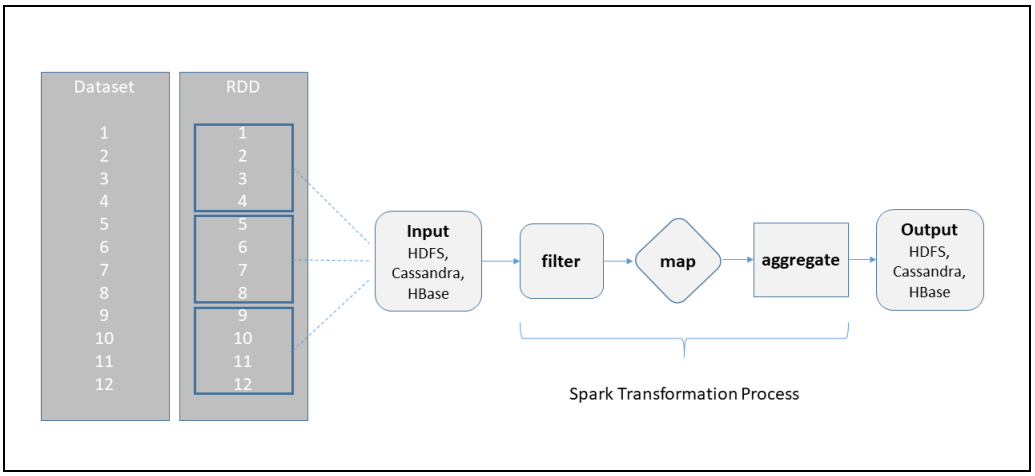


Abbildung 52: Resilient Distributed Dataset (eigene Darstellung in Anlehnung an Soutier, 2015)

Die gesamte Planung und Ausführung in Spark erfolgt auf der Grundlage dieser Methoden, sodass jedes RDD seine eigene Art der Selbstberechnung implementieren kann. Anwender können benutzerdefinierte RDDs implementieren, bspw. zum Lesen von Daten aus einem neuen Speichersystem (vgl. Apache Software Foundation, 2020). Mit Spark SQL besteht ein Tool, um die Datensätze (RDDs) zu transformieren, sodass SQL-Abfragen möglich sind. Die RDDs werden als temporäre Tabellen erzeugt. Für Streaming-Daten wird Spark Streaming in Anspruch genommen. Spark Streaming verarbeitet die kontinuierlichen Datenströme. Dabei entstehen einzelne Datenpakete, in denen gezielt Analysen ausführbar sind. Das Spark Streaming ist in dem Big Data Shop in der Lambda-Architektur im Speed Layer vorzufinden. Mit der vorhandenen Bibliothek MLlib (Machine Learning Library) sind Algorithmen verfügbar, die für das Machine Learning benötigt werden. Als letzte Instanz in dem Spark-Cluster ist GraphX zu nennen. Mit dieser Komponente lassen sich Graphen aus dem Spark-Cluster erzeugen (vgl. Luber & Litzel, 2016, S. 1).

Funktionsweise des Spark-Clusters

Im Spark-Cluster wird das Master-Worker-Prinzip ausgeführt. Eine Applikation von Spark wird daher zum Master geschickt und dort als Java-Instanz[9] gestartet. Folglich wird diese Applikation nun als Driver eingesetzt. Die Aufgabe eines Drivers ist die Koordination der Ausführung der Jobs. Dafür werden gerichtete Graphen aus den RDD-Operationen erstellt und danach werden die Operationen in Stages aufgeteilt. Aus den Stages werden Tasks zerlegt, damit das parallele Abarbeiten möglich ist. Jedem Knoten werden somit Tasks zugeteilt. Daraus

[9] Auch Java Virtual Machine Instanz. JVM dient hier als Schnittstelle zwischen Spark-Cluster und Zielsystem/Quellsystem.

folgen Knoten (Worker-Knoten), aus denen dann weitere Java-Instanzen gestartet werden. Innerhalb eines Executors (Java-Instanz) werden auch mehrere Tasks ausgeführt, um alle zugewiesen Ressourcen (Kerne) auszulasten (vgl. Soutier, 2015). Ein Task hat einen Teil-Input und erzeugt damit einen Teil-Output. Teil-Outputs werden auch als Shuffle bezeichnet. Voneinander abhängige Tasks laufen auf gleichen Worker, damit die Datenlokalität gewährleistet bleibt. Dabei werden die Tasks in Stages zusammengefasst. Innerhalb der Stages ist ein Netzwerktransfer möglich, um dann diese zu einem Task zusammenzufassen (siehe Abbildung 53).

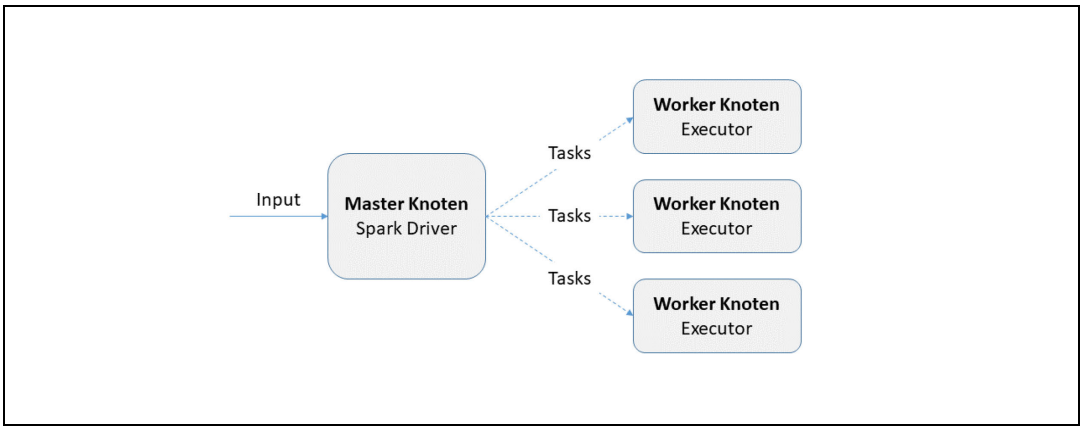


Abbildung 53: Spark-Knoten (eigene Darstellung in Anlehnung an Soutier, 2015)

4.5 NoSQL-Datenbanken

Aufgrund der immer größer werdenden Datenmengen und der daraus resultierenden Analysemöglichkeiten ist eine Speicherung dieser Vorgänge in einem Datenbank-Managementsystem essentiell. Um dies zu bewältigen, ist der Ansatz der NoSQL-Datenbanken von Bedeutung. NoSQL-Datenbanken[10] sind nichtrelationale Datenbanken (vgl. Han, E, Le & Du, 2011, S. 363). Sie sind alternative Datenbankmodelle im Vergleich zu den traditionellen relationalen SQL-Datenbanken. Eine Kombination aus SQL- und NoSQL-Datenbanken ist möglich. NoSQL-Datenbanken sind auch als strukturierte Datenspeicher bekannt. Ein NoSQL-Datenbanksystem basiert auf verteilten Datenhaltungsarchitekturen (vgl. Meier & Kaufmann, 2016, S. 18). Die Grundstruktur eines NoSQL-Datenbanksystems ist aus der Abbildung 54 zu entnehmen.

[10] NoSQL steht für „not only SQL“.

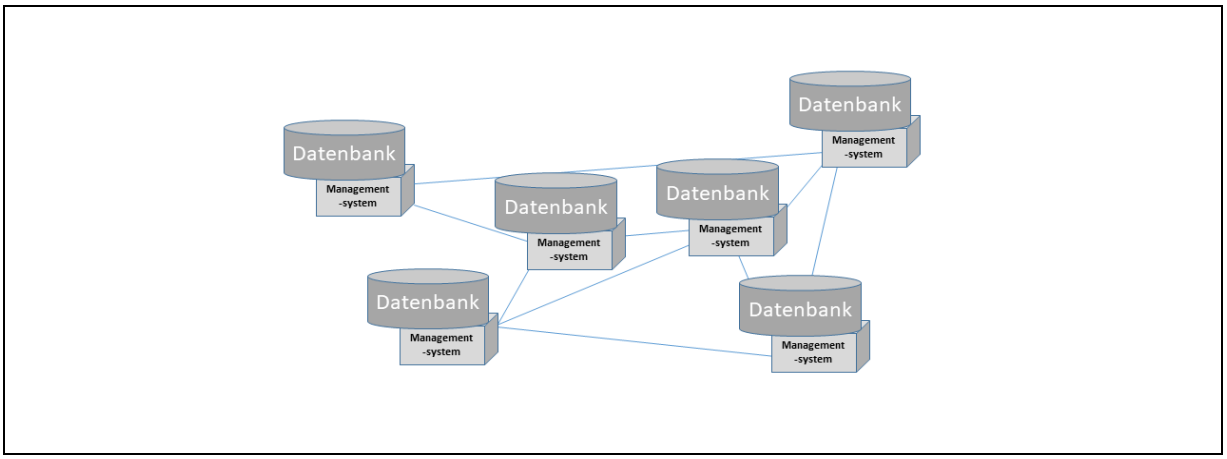


Abbildung 54: NoSQL-Grundstruktur (eigene Darstellung in Anlehnung an Meier & Kaufmann, 2016, S. 19)

Dort sind die Daten als Schlüssel-Wert-Paare, Spalten, Dokumente oder Graphen gespeichert und dementsprechend abgebildet. Damit nutzen NoSQL-Datenbanken keine klassischen Tabellen, die durch SQL-Abfragen abgerufen werden können. Eine hohe Verfügbarkeit wird erreicht, indem unterschiedliche Replikationen erstellt werden.

Graphendatenbanken

Durch Knoten und Kanten werden Beziehungen zwischen Daten abgebildet. Die Organisation der Daten erfolgt über Knotenpunkte und ihre Verbindung zueinander. Daten mit starkvernetzten Informationen erreichen eine hohe Verfügbarkeit und somit höhere Zugriffsgeschwindigkeit als relationale SQL-Datenbanken. Graphendatenbanken findet man primär dort, wo Daten in Netzwerken vorkommen (vgl. Meier & Kaufmann, 2016, S. 238). Dabei ist nicht der einzelne Datensatz von Bedeutung, sondern vielmehr die Verknüpfung. Als Beispiele sind soziale Medien, Wasser- und Energieversorgung oder auch die Analyse von Verlinkungen von Webseiten zu nennen.

Schlüssel-Wert-Datenbanken

Nach dem Schlüssel-Wert-Datenmodell entspricht ein Wert genau einem Schlüssel. Von der Struktur ist dieses Modell im Gegensatz zu relationalen Datenbanken einfacher gestaltet (vgl. Han et al., 2011, S. 364). Die Abfragegeschwindigkeit ist dadurch höher, außerdem ist ein paralleles Abrufen der Daten möglich. Abfragen und Änderungen der Daten werden über den Primärschlüssel ausgeführt.

Spaltenfamilien-Datenbanken

Das zuvor beschriebene Schlüssel-Wert-Modell ist einfach aufgebaut. Das Modell bedarf aber einer Strukturierung des Datenraums (vgl. Meier & Kaufmann, 2016, S. 226). Die Erweiterung ist das Konzept der Spaltenfamilien-Datenbanken. Diese geben dem Schlüssel-Wert-Konzept mehr Struktur. Eine spaltenweise Speicherung bringt zudem Vorteile beim Lesezugriff. Beim zeilenweisen Speichern werden durchaus nicht alle Spalten benötigt. Im Gegenzug bestehen aber Gruppen von Spalten, die häufig zusammen gelesen werden (vgl. Meier & Kaufman, 2016, S. 226). Für die Zugriffsoptimierung werden die Daten in Spaltenfamilien strukturiert und als Gruppe von Spalten erzeugt. Folgende Eigenschaften kommen nach HAN ET AL. hiermit zutrage:

- Daten werden spaltenweise in mehrdimensionalen Tabellen gespeichert,
- Die Spalten sind zu Spaltenfamilien zusammengeführt,
- Jede Datenspalte ist ein Datenbankindex,
- Gleichzeitige Prozessabfragen sind möglich, da jede Spalte von einem Prozess abgefragt werden kann.

Die Vorteile sind einerseits hohe Skalierbarkeit und andererseits hohe Verfügbarkeit. Diese Verfügbarkeit resultiert aus der massiven Verteilung (vgl. Meier & Kaufmann, 2016, S. 228).

Dokument-Datenbanken

Dokument-Datenbanken sind von ihrer Struktur den Schlüssel-Wert-Datenbanken ähnlich. Die Gemeinsamkeit liegt in der Schemafreiheit. Erweitert wird hierbei aber die Eigenschaft der Strukturierung der gespeicherten Daten. In einer Dokument-Datenbank werden strukturierte Daten in Datensätzen gespeichert. Ihr Einsatz ist für Webdienste prädestiniert. Datenformate wie JSON[11] oder HTTP[12] sind integrierbar (vgl. Meier & Kaufmann, 2016, S. 229). Dokument-Datenbanken sind zudem horizontal skalierbar. Wie eingangs in diesem Abschnitt erwähnt, arbeiten auch Dokument-Datenbanken in einem verteilten System. Dadurch können große Mengen heterogener Daten verarbeitet werden. Hinzu kommt, dass Dokument-Datenbanken schemafrei sind. Es existiert somit keine Vorgabe, welche Datenstruktur erzeugt werden muss. Daraus kann Flexibilität erreicht und unterschiedlichste Daten gespeichert werden. Durch das Map/Reduce-Verfahren werden Anfragen an eine Dokument-Datenbank parallelisiert und beschleunigt (siehe Abschnitt 4.3.2.3).

[11] JavaScript Object Notation.
[12] HyperText Transfer Protocoll.

Vorteile einer NoSQL-Datenbank

Zusammengefasst ergeben sich folgende Gründe für die Integration und Nutzung einer NoSQL-Datenbank. Einerseits ist die Flexibilität zu nennen. NoSQL-Datenbanken haben in der Regel, ausgenommen Apache Hive, flexible Schemata. Sie ermöglichen eine schnelle und iterative Entwicklung der Datenbank. Durch ihre Flexibilität sind NoSQL-Datenbanken für semistrukturierte sowie unstrukturierte Daten geeignet. Ein weiterer Grund ist die Skalierbarkeit und der sich daraus ergebene Lösungsansatz mit Cloudservern. Die NoSQL-Datenbanken können durch ihre Nutzung mit verteilten Hardware-Clustern beliebig hoch skaliert werden (horizontale Skalierung). Die hohe Zugriffsgeschwindigkeit ist gewährleistet, da bestimmte Datenmodelle und Zugriffsmuster darauf optimiert sind. NoSQL-Datenbanken haben eine Fülle von APIs[13] und Datentypen, die auf das jeweilige Datenmodell anwendbar sind.

4.5.1 ACID-, CAP- und BASE-Theorem

Für den Mehrbenutzerbetrieb und das daraus erläuterte verteilte System ist die Konsistenz oder auch Integrität einer Datenbank entscheidend. Technisch betrachtet können durch gleichzeitiges Zugreifen mehrerer Nutzer auf eine Datenbank eine Blockierung oder Konsistenzverletzungen auftreten (vgl. Meier & Kaufmann, 2016, S. 135 f.). Dadurch müssen Regeln eingeführt werden, um keine Deadlocks[14] entstehen zu lassen. Im Folgenden befasst sich der Autor mit den Integritätsregeln nach dem ACID-, CAP- und BASE-Theorem.

ACID-Theorem

Damit die Integrität gewährleistet werden kann, ist mit dem ACID-Theorem ein Standard aufgestellt worden, an dem sich die Datenbanken orientieren sollten. Dabei steht die Transaktion unter den Integritätsregeln. Eine Transaktion ist eine Folge von Operationen und soll demnach atomar, konsistent, isoliert und dauerhaft sein (vgl. Meier & Kaufmann, 2016, S. 136). Dafür steht das Akronym ACID:

[13] Application Programming Interface.
[14] Zustand in der Informationsverarbeitung, bei dem sich zwei oder mehrere ablaufende Prozesse gegenseitig blockieren (vgl. Lackes, 2020).

- Atomicity: Hierbei handelt es sich um das Prinzip: „All or nothing“, d.h. innerhalb einer Transaktion wird entweder die gesamte logische Folge einer Operation durchgeführt oder gar keine.
- Consistency: Damit wird sichergestellt, dass alle in eine Datenbank geschriebenen Daten immer gemäß allen in der Datenbank definierten Regeln gültig sind. Dabei werden bei der Ausführung einer Transaktion die Datenbankregeln oder die Datenbankbeschränkungen nicht verletzt.
- Isolation: Dies bedeutet, dass keine neue Transaktion Zugriff auf Daten hat, solange eine Transaktion nicht beendet oder aktuell bearbeitet wird. Jede Transaktion ist unabhängig und wird nicht von anderen Transaktionen beeinflusst.
- Durability: Die Haltbarkeit stellt sicher, dass die Ergebnisse einer Transaktion dauerhaft gespeichert werden, sobald die Transaktion abgeschlossen ist (vgl. Cattell, 2010, S. 12).

CAP- und BASE-Theorem

Das CAP-Theorem von BREWER besagt, dass eine verteilte Datenbank nur zwei der folgenden drei Eigenschaften garantieren muss:

- Consistence: Gewährleistung, dass jeder Knoten in einem verteilten System (Cluster) den letzten Schreibvorgang zurückgibt. Jeder Knoten gibt somit die gleichen Daten als Antwort zurück.
- Availability: Jeder aufgefallene Knoten gibt in einer angemessenen Zeitspanne eine Antwort zurück.
- Partition Tolerance: Das Cluster funktioniert auch ohne Netzwerkpartitionen weiter. Der Ausfall eines Knotens hat keinen Einfluss auf das Gesamtsystem.

NoSQL-Datenbanken verstoßen in vielen Gesichtspunkten gegen das ACID-Theorem. Gerade in den Punkten Fehlertoleranz und Verfügbarkeit bei Ausfällen eines Rechnerknotens sollen NoSQL-Datenbanken für den Anwender weiterhin vorhanden sein (vgl. Meier & Kaufmann, 2016, S. 148). Ausfall- oder fehlertolerante Systeme nutzen replizierte Rechnerknoten. Damit entstand das BASE-Theorem, um auch für nichtrelationale Datenbanken eine Regelung treffen zu können. BASE steht für:

- Basically Available: Gewährleistung der Verfügbarkeit im Sinne des CAP-Theorems, d.h. jeder nicht ausfallende Knoten gibt in angemessener Zeit eine Antwort zurück.
- Soft state: Der Zustand des Systems ändert sich im Laufe der Zeit auch ohne Eingabe.

- Eventual consistency: Das System wird im Laufe der Zeit aufgrund von Nichteingabe konsistent.

4.5.2 NoSQL-Open-Source-Datenbanksysteme

Nachfolgend werden zwei Open-Source-Datenbanken vorgestellt: Apache HBase und Apache Cassandra. Mit Apache Hive wird ein Data-Warehouse-Managementtool vorgestellt, welches mit dem HDFS in Verbindung gebracht wird, um Abfragen und Verwaltungsaufgaben auszuführen. Daraus sollen je nach Anwendungszweck ein, zwei oder auch alle drei Datenbanksysteme zum Einsatz kommen, denn die Auswahl „einer“ geeigneten Datenbank-Engine wäre sehr eingeschränkt in der Funktionalität. Daher wird hier die Empfehlung einer simultanen Nutzung aller vorgestellten Tools ausgesprochen.

4.5.2.1 Apache HBase

Apache HBase ist eine verteilte, nichtrelationale Datenbank. HBase ist eine Open-Source-NoSQL-Datenbank im Hadoop-Ökosystem (vgl. Apache Software Foundation). Dadurch ist eine Implementierung und die Nutzung von HBase ohne Weiteres möglich. Durch die Bereitstellung von Diensten wie dem BigTable-Konzept von Google kann HBase große Datenmengen mit Fehlertoleranz speichern und wird in vielen datengesteuerten Websites wie bspw. dem Messenger von Facebook eingesetzt. HBase speichert die relevanten Daten, die ursprünglich in mehreren SQL-Tabellen gespeichert sind, in einer gleichen Tabelle. Dieser Vorgang ermöglicht einen direkten Zugriff auf diese Daten, ohne zusätzlich Zeit damit verbringen zu müssen, verwandte Tabellen miteinander zu verbinden. Wenn eine relationale Datenbank in einer bestimmten Anwendung aber oft mehrere Tabellen verknüpft, sollten die Tabellen in HBase-Tabellen konvertiert werden (vgl. Chen, 2017, S. 170). Das wiederum gewährleistet und verbessert die Effizienz der Datenverarbeitung. Dann sind parallele Analysen ausführbar und durch die Integration auf dem Hadoop Distributed File System (HDFS), in Apache Hive oder auch in Apache Pig dementsprechend abrufbar. Dabei nutzt HBase das HDFS als fehlertoleranten Data Store. Durch die Datenmodelle, die Geschwindigkeit und die Fehlertoleranz im verteilten Hadoop-Ökosystem ist Apache HBase geeignet für Arbeitslasten bei Anzeigentechnologien, Webanalysen, Finanzdiensten, Zeitreihendaten-Anwendungen.

4.5.2.2 Apache Cassandra

Apache Cassandra ist ebenfalls eine verteilte nichtrelationale Datenbank wie das oben erwähnte Apache HBase. Cassandra wurde bei Facebook 2008 entwickelt und 2009 als Open-Source-

Speichersystem an die Apache Software Foundation übergeben (vgl. Drilling & Litzel, 2015, S. 1). Ein Merkmal von Cassandra ist das verteilte Speichersystem. Die Hauptaufgabe von Datenbanken sind allgemein das Speichern und Berechnen von Daten. In verteilten Systemen werden Speicherung und Berechnung im Gegensatz zu herkömmlichen relationalen Datenbanken von mehreren Knoten übernommen. Apache Cassandra arbeitet in einem Cluster, das aus mehreren Knoten besteht. Somit ist die Lösung eines Problems nicht von einem einzelnen Knoten abhängig. Ein weiteres Merkmal und zugleich Vorteil ist, dass Cassandra für den Betrieb auf handelsüblicher Hardware ausgelegt ist (vgl. Apache Software Foundation). Damit ist eine kostengünstige Skalierung möglich. Die lineare Skalierbarkeit ist ein großer Leistungsvorteil. Wird die Anzahl der Knoten bspw. verdoppelt, wird die Leistung ebenso auf das Doppelte gesteigert. Somit ist Cassandra in der Lage, Terabyte an Daten und Tausende gleichzeitige Operationen pro Sekunde zu verarbeiten (vgl. Lakshman & Malik, 2010, S. 35–40). Die Fehlertoleranz ist demnach bei der Ausführung einer Anwendung auf einem Cluster gegeben. Durch die Konzeption als verteiltes System ist Cassandra so konzipiert, dass die Nutzung über mehrere Rechenzentren hinweg eingesetzt werden kann. Eine optimale geographische Verteilung vermindert Redundanzen und Ausfälle. Darüber hinaus fungiert ein verteiltes System als Disaster Recovery.

4.5.2.3 Apache Hive

Apache Hive ist ein Data-Warehouse-Tool zur Verarbeitung strukturierter Daten in Hadoop. Die Abfragesprache ist angelehnt an die SQL-Abfrage. Hive erleichtert das Lesen, Schreiben und Verwalten großer Datensätze auf dem HDFS (vgl. Apache Software Foundation). Weitere Funktionen sind:

- Erstellung von ETL[15]-Prozessen, Berichterstattungen und Analyse,
- Strukturierung einer Vielzahl an Datenformaten,
- Zugriff auf Daten im Data Warehouse, auf Apache Hadoop HDFS oder auch Apache HBase,
- Abfrageausführung über Apache Spark oder auch MapReduce (HDFS),
- Sub-Sekunden-Abfrage über Apache YARN.

Das Speichern von Daten ist mit Hive schemafrei, solange die Daten strukturiert sind. Apache Hive besitzt zudem Connectors für Textdateien wie CSV/TSV[16] und andere Datenformaten.

[15] Extract Transform Load.
[16] CSV/TSV (Comma-Separated-Values/Tab-Separated-Values).

4.6 Visualisierung

Unter einer Visualisierung ist eine graphische Darstellung gemeint, die dazu beiträgt, dem Betrachter Informationen zu vermitteln. Mithilfe von Visualisierungen können dem Betrachter komplexe Sachverhalte in einer übersichtlichen und leicht verständlichen Form veranschaulicht sowie mögliche Zusammenhänge zwischen den Daten gezeigt und verdeutlicht werden (vgl. Ward et al., 2015, S. 1–2). Im Folgenden wird zunächst das Konzept der Visualisierung definiert und anschließend die menschliche Interaktion mit Visualisierungen beleuchtet. Auf dieser Basis werden Gestaltungsregeln für Visualisierungen erarbeitet. Die Aufgabe einer Visualisierung ist die Präsentation von komplexen Daten und Informationen, sodass sowohl Strukturen als auch Abhängigkeiten erkennbar werden.

McCormick definiert Visualisierungen als ein Berechnungsverfahren, das zur Transformation von Informationen durch die Umwandlung von symbolischen und geometrischen Darstellungen mittels graphischer Objekte dient. Hierbei werden Zusammenhänge und Strukturen der Informationsbasis identifiziert und für den Betrachter graphisch aufbereitet. Die Übertragung der Informationen von der Visualisierung auf den Betrachter folgt keiner einheitlichen Prozessstruktur, denn jeder Betrachter versteht und interpretiert die Visualisierung subjektiv. Allerdings kann er in diesem kreativen Entstehungsprozess durch die Anwendung von Interpretationsregeln für einzelne Objekte und Strukturen unterstützt werden. Darüber hinaus wird die Visualisierung in einen inhaltlichen Zusammenhang gestellt, um die gewünschte Interpretation anzuregen (vgl. McCormick, 1988, S. 15 ff.).

Die spezifischeren Visualisierungen von Daten unterscheiden sich von den herkömmlichen Visualisierungen dadurch, dass die darzustellenden Informationen nicht mit den bestehenden Daten in einer Datenbank identisch sind. Daher müssen diese Informationen zunächst aus den Daten extrahiert werden. Laut CARD, MACKINLEY und SHNEIDERMAN kann eine solche Datenvisualisierung durch die folgenden fünf Eigenschaften von anderen Visualisierungen unterschieden werden (vgl. Card et al., 1999):

- Computergestützte Darstellungen und Reports,
- Interaktive Auswahl und Filterung der Datenbasis sowie die Änderung des Abstraktionsgrades in der Visualisierung,

- Präsentation von Informationen in visueller Form durch die Verwendung von Eigenschaften und Attributen der Datenbasis in graphischen Elementen (Form, Farbe, Größe),
- Darstellung abstrakter Daten, die keiner physischen Form angehören,
- Nutzung und Förderung der menschlichen Wahrnehmungsfähigkeiten.

Ausgehend von der Datenbasis werden diese Datenvisualisierungen algorithmisch generiert und verändern sich in Abhängigkeit von der Datenbasis dynamisch. So grenzen sich Datenvisualisierungen maßgeblich von statischen Graphiken ab, die lediglich einmalig Probleme aufzeigen. Generell ermöglichen Datenvisualisierungen sowohl eine explorative als auch eine konfirmative Datenanalyse. Die konfirmative Datenanalyse zeigt in Form einer Datenvisualisierung die Ergebnisse definierter Sachverhalte auf. Im Gegensatz dazu bildet eine explorative Datenanalyse die Daten der Datenbank ab (Schumann & Müller 2013, S. 5–6). In dieser Darstellung müssen in der Folge Annahmen getroffen und formuliert werden.

4.6.1 Psychologische Wahrnehmung

Datenvisualisierungen sind ein Medium zur Übertragung von Informationen, die in Daten gespeichert sind. Damit diese Informationen gewonnen und interpretiert werden können, muss zunächst die Wahrnehmungspsychologie des Betrachters untersucht werden.

Als Betrachter verfügt der Mensch über unbewusste Fähigkeiten zur Mustererkennung und automatischen Abstraktion im Gehirn. Diese Automatismen werden unbewusst und automatisch vom menschlichen Gehirn ausgeführt. Dieser Prozess dient dazu, die Komplexität der zu verarbeitenden Informationen zu reduzieren. Die Ausprägung dieser Fähigkeiten hängen aber von der individuellen Entwicklung und der persönlichen Erfahrung des Betrachters ab. Um diese unterbewussten Prozesse durchführen zu können, sind eindeutige Entscheidungsregeln erforderlich, die festlegen, wie komplexe Beziehungen aufgelöst und klassifiziert werden. Diese Entscheidungsregeln werden nicht vom Betrachter aktiv beeinflusst, sondern vielmehr durch die menschliche und persönliche Entwicklung des Betrachters im Laufe seines Lebens. Dementsprechend ist der Prozess des Erkennens des Gesehenen und der Interpretation in Bezug auf die Bedeutung für jeden Menschen individuell und nie objektiv, sondern stets subjektiv. Diese Erkenntnis ist für die Erstellung von Datenvisualisierungen unerlässlich, da darauf unterschiedliche Auslegungen und Interpretationen basieren (vgl. Doherty et al., 2010).

4.6.2 Intuitive Wahrnehmung

Ein anderer Aspekt ist die intuitive Wahrnehmung, die von der historisch-kulturellen Entwicklung des Menschen geprägt ist. Diese intuitive Wahrnehmung ist im Hinblick auf die Datenanalyse in Datenvisualisierungen der derzeit größte Unterschied zwischen dem Menschen und dem Computer und charakterisiert somit die Grundlage für die Einbindung eines Menschen in den Prozess der explorativen Datenanalyse. In der Phase der menschlichen Entwicklung war es zum einen überlebenswichtig, schnell und präzise zu entscheiden, und zum anderen, vorhandenes Wissen an neue Probleme anzupassen. Diese Flexibilität kann nur durch eine hohe Abstraktionsfähigkeit erreicht werden. Nur so wird die Handlungsfähigkeit in Gefahrensituationen, die schnellere Entscheidungen erfordern, sichergestellt. Diese Fähigkeit des Menschen ist derzeit der größte Unterschied zwischen menschlicher Reaktion und Denken und jener eines computerstützten Systems. Daher können entsprechende serielle Berechnungen oder vordefinierte „True-False“-Regeln allein keine komplexen Muster erkennen und gleichzeitig Lösungsansätze an neue Probleme anpassen. Im Vergleich dazu bewerkstelligt das menschliche Gehirn dies durch parallele Prozesse der Abstraktion, Klassifizierung und Komplexitätsreduktion. Dies führt einerseits zu vermeidbaren Fehlern bei den durchgeführten Aktionen, andererseits ermöglicht es eine sehr schnelle Entscheidungsfindung. Ausgehend von diesem grundlegenden Unterschied haben Menschen die Möglichkeit, verschiedenste Informationen sinnvoll zu kombinieren und die daraus resultierende Bedeutung zu interpretieren. Diese Fähigkeit wird als „symbolisches Denken" bezeichnet. Nachdem die Sachverhalte gelöst wurden, können Muster so manipuliert werden, dass sie abstrakt auf neue Probleme übertragen werden können. Diese Fähigkeit bildet gleichzeitig die Voraussetzung dafür, semantisch wertvolle Entscheidungen auf der Grundlage unbekannter Daten zu treffen und die damit verbundene Bedeutung der Informationen innerhalb einer Datenvisualisierung zu verstehen (vgl. Fischer, 2014, S. 19–20).

4.6.3 Big-Data-Datenvisualisierung

Die Datenvisualisierung von Big-Data-Inhalten nimmt eine ganz besondere Rolle ein, weil sie den Wissensgewinn bei unbekanntem Nutzen hervorheben und nutzbar machen kann. Um Informationen aus den Daten zu extrahieren, sind diese vorab zu analysieren und in einen sinnvollen Zusammenhang zu bringen. Dabei unterstützen Datenvisualisierungen sowohl bei der Selektion als auch bei der Betrachtung von relevanten Daten aus großen und komplexen Datenmengen. Dieser Abschnitt befasst sich mit den möglichen Visualisierungen solcher großen Datenzusammensetzungen.

Datenvisualisierungen dienen dazu, die psychologischen Fähigkeiten der Betrachter auszuschöpfen, um netzwerkartige Strukturen, Muster oder Zusammenhänge innerhalb der Daten zu erfassen. Dieser Vorgang ist naturgemäß kompliziert und erfordert die Unterstützung des Betrachters. Dies kann durch die beschriebenen Gestaltungsregeln erleichtert werden, um die natürlichen Defizite der subjektiven Wahrnehmung zu überbrücken. Aufgrund des Datenvolumens großer Datenmengen ist es zudem erforderlich, Voraussetzungen zu definieren, die den Umfang der in der Datenvisualisierung zu berücksichtigenden Daten eingrenzen. Sonst wird der Betrachter durch die Datenmenge überfordert und ist nicht in der Lage, daraus eine sinnvolle Aussage zu treffen (vgl. Keim, 2001, S. 1).

Derartige Voraussetzungen können aus dem erwarteten Nutzen auf der Basis des Big-Data-Frameworks abgeleitet werden. Abhängig von der Orientierung der Big-Data-Anwendung und der zu erwartenden Aussage aus der Fragestellung sind solche Voraussetzungen nötig, um die Komplexität der Datenvisualisierung zu reduzieren. Andernfalls kann keine zielgerichtete Datenvisualisierung erstellt werden. Nach der Auswahl der richtigen und sinnvollsten Prämissen steht für verwertbare Aussagen die menschliche Fähigkeit der Mustererkennung innerhalb der dargestellten Daten im Vordergrund. Dies beruht auf einem möglichen Verfahren zur Massendatenanalyse durch KEIM.

„Never before in history data has been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of data becomes increasingly difficult. […] The advantage of visual data exploration is that the user is directly involved in the data mining process." (Keim, 2001)

Derzeit können rechnergestützte Systeme nur eine unterstützende Funktion in diesem Prozess übernehmen, weshalb die IT noch ein enormes Entwicklungspotenzial in diesem Bereich aufweist. Die Möglichkeit, Komplexität zu reduzieren und zu bewältigen, bildet den neuen

„Business Driver“, der einen erheblichen Wettbewerbseffekt hat. Eine vergleichbare Beurteilung findet sich in der Arbeit von SCHOENEBERG. Dieser bezeichnet die (erweiterte) Visualisierung von Daten und Informationen als den stärksten Trend in der Business Analytics, um Wissen und Wettbewerbsvorteile zu generieren (vgl. Schoeneberg & Pein 2014, S. 309 ff.).

Auch wenn diese Einschätzung und das damit verbundene Potenzial für die Softwareanbieter wie Tableau, Microsoft Power BI oder Qlik View in der rechnergestützten Datenanalyse liegen, müssen derartige Analysen in absehbarer Zeit weiterhin durch eine menschliche Datenvisualisierung durchgeführt werden. Aus diesem Grund ist es in der gegenwärtigen wie auch in der zukünftigen Software-Entwicklung wichtig, nicht nur die technologischen Beschränkungen zu berücksichtigen, sondern weiterhin die menschlichen Grenzen der Wahrnehmung zu erfassen und genau diese zu unterstützen.

Basierend auf der Visualisierungsklassifizierung von Keim formulieren SCHOENEBERG und PEIN die Aufgaben einer guten Softwarelösung wie folgt:

- Schaffung von Interaktionsmöglichkeiten zur Datenanalyse,
- Einfache Handhabung mit Fokus auf den Daten,
- Einfaches und intuitives Navigieren durch die verschiedenen Sichten eines Analyseprozesses und die Ebenen der Abstraktionsebenen (vgl. Schoeneberg & Pein 2014, S. 309 ff.).

Das Abspeichern von Analysen erfolgt in den implementierten NoSQL-Datenbanken (siehe Abschnitt 4.5)

4.7 Integration in bestehende Prozesse

Die Nutzung neuer Technologien in einem Unternehmen bringt eine gewisse Unsicherheit mit sich. Vor allem für Unternehmen oder zumindest Fachbereiche ohne Erfahrung in der Anwendung von Big Data kann dies eine große Herausforderung darstellen. Bei der Einführung von Big-Data-Technologien in einem Unternehmen sollte systematisch vorgegangen werden. Durch eine systematische Vorgehensweise steigt die Wahrscheinlichkeit für den Erfolg des Big-Data-Projekts (vgl. Henke et al., 2016, S. 3–4).

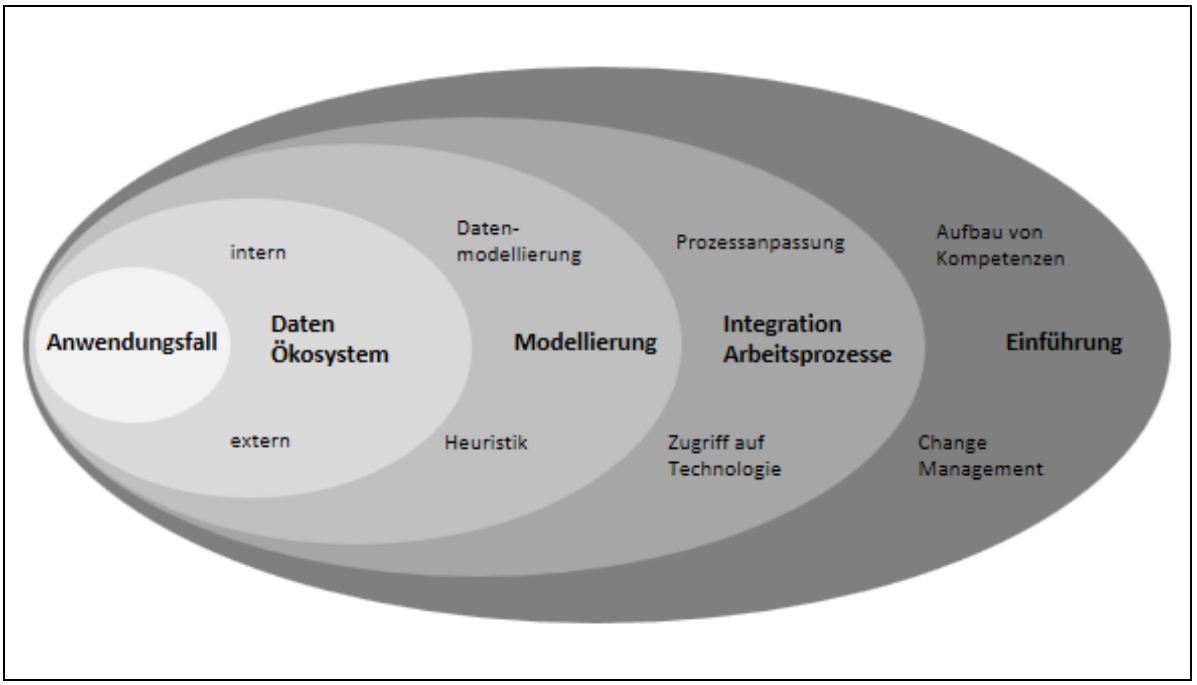


Abbildung 55: Vorgehensmodell für die Einführung der Datennutzung (Henke et al. 2016, S. 4)

In Abbildung 55 ist ein von der Unternehmensberatung McKinsey im Rahmen einer Studie entwickeltes Vorgehensmodell dargestellt. Dieses kann als Orientierung bei der Einführung von Big-Data-Technologien genutzt werden. Nach diesem Modell wird empfohlen, sich bei der Einführung an einem konkreten Anwendungsfall zu orientieren und Big Data nicht sofort auf dem breiten Feld einzuführen. Dabei ist es besonders wichtig, sich intensive Gedanken über den Anwendungsfall zu machen. Bei einem sorgfältig ausgewählten Anwendungsfall steigt die Wahrscheinlichkeit des Gelingens des Big-Data-Projekts. Daraufhin steht die Beschaffung von relevanten Daten im Fokus. Hier sollen sowohl interne als auch externe Quellen von Daten in das Big-Data-System einbezogen werden. Im nächsten Schritt sollen passende Analysemethoden bestimmt werden, mit denen aus den Daten relevante Informationen gewonnen werden können bzw. ein Nutzen generiert werden kann. Der nächste Schritt umfasst die Integration von Big-Data-Technologien in die bestehenden Arbeitsprozesse. Im letzten Schritt geht es um eine umfassende Einführung der neu gewonnenen Möglichkeiten aus der Big-Data-Anwendung. Das Management soll die Ergebnisse der Big-Data-Analysen für die Entscheidungsfindung nutzen. Außerdem sollte sich die Nutzung fachabteilungs- bzw. unternehmensübergreifend weiterentwickeln (Henke et al. 2016, S. 3–4).

5. Evaluierung der Big-Data-Architektur mit der Fallstudie: Predictive Maintenance

Zustandsüberwachung, Prognose und Diagnose sind die Essenz jeder vorausschauenden Instandhaltungsstrategie. Die Erstellung geeigneter Modelle für diese Aufgaben kann durch datengetriebene oder physikalisch basierte Ansätze erreicht werden. Algorithmen, die den datengetriebenen Ansatz verwenden, lernen Modelle direkt aus den Daten und gehören daher üblicherweise zum Bereich des maschinellen Lernens (siehe Abschnitt 4.4). Die massive Zunahme maschinell erzeugter Daten hat zu maschinellen Lernmodellen mit bisher unerreichter Leistungsfähigkeit geführt. Dieses Kapitel gibt einen Einblick, wie maschinelle Lernmethoden Daten verarbeiten, um solch erstaunliche Ergebnisse zu erzielen. Im ersten Abschnitts des Kapitels 5 werden die Instandhaltung und der spezifische Klebeprozess eingeordnet. Der zweite Abschnitt wird die Vorgehensweise der Datenbeschaffung und der Datenaufbereitung erläutern und den Zusammenhang mit den zuvor behandelten Kapiteln verstärken. Dabei liefert der Abschnitt 5.2 einen Beitrag zur Modellentwicklung für das Predictive-Maintenance-Konzept. In diesem Abschnitt werden unterschiedliche Ansätze im Bereich des Machine Learnings angewendet. Dabei werden sämtliche Herausforderungen, Probleme und Lösungsansätze beschrieben. Im Wesentlichen soll das Modell zur Verbesserung der Zustandsüberwachung beitragen und eine Grundlage für weitere Diagnose- und Prognosemodelle darstellen. Der letzte Abschnitt 5.3 widmet sich der Predictive-Maintenance-Strategie. Aus der erfolgten Analyse wird ein Konzept entwickelt, welches die sinnvolle Anwendung des Modells empfiehlt.

5.1 Instandhaltung in einem Karosseriewerk

Zunächst wird anhand von Kennzahlen die mögliche Dimension beschrieben, die ein Karosseriewerk in einem der führenden Automobilhersteller beherbergt. Damit schaffen wir die Grundlage für die Einordnung und Stellenwert der Instandhaltung im Karosseriebau. Die Produktion in dem betrachteten Karosseriewerk generiert in einem Jahr rund 88.500 Karossen. Auf Tage heruntergebrochen sind dies ungefähr 377 Karossen pro Tag. Das Werk selbst hat eine Produktionsfläche von rund 78.000 m². Die Anzahl der Beschäftigten beträgt insgesamt 642 Mitarbeiter, davon sind 601 direkt an den Stationen und 41 indirekt tätig. Zu den indirekten Mitarbeitern zählen das Management und die Instandhaltung im Werk. Um den hohen Stellenwert der Ingenieure hervorzuheben, ist die Betrachtung der automatisierten Prozesse

notwendig. In dem in der Arbeit analysierten Werk liegt der Automatisierungsgrad bei rund 70 %. Dort befinden sich 76 Hubgehänge, 440 Skids[17], 410 Roboter, 88 Kleberanlagen, 36 Heber, zehn Laserroboter und elf automatisierte Schweißapparate.

Die Instandhaltung in der Produktion und hier im Fallbeispiel in einem Karosseriewerk wird durch die präventive Wartung und die korrektive Instandhaltung (siehe Kapitel 2, S. 32–35) durchgeführt. Um einem Maschinenausfall vorzubeugen, werden täglich Wartungsmaßnahmen ausgeführt. Anhand der abgebildeten Tabelle 11 sind die Tätigkeiten der Wartung und die Dauer abzulesen.

Wartungsvorgang 1			Wartungsvorgang 2		
Prozess	**Dauer in min**	**Intervall**	**Prozess**	**Dauer in min.**	**Intervall**
Schmierung der Dosierkammer	10	14 Tage	Filterkontrole	5	täglich
Reinigung Seitenlager der Spiendel	15	6 Monate	Sicherheitsventilkontrolle	5	täglich
Schmierung Seitenlager der Spiendel	5	6 Monate	Betriebsdruckkontrole (Pumpen,Fass…)	5	täglich
Schmiern der Gewindespiendel	15	6 Monate	Kontrolle wichtiger Schraubverbindungen	5	täglich
Reinigung der Gewindespiendellager	15	6 Monate	Kontrolle wichtiger Steckverbindungen	5	täglich
Schmierung der Gewindespiendellager	5	12 Monate	Konrtolle Elektrisch. Und Pneumat. Leitungen	5	täglich
Uberprufung RCD	5	täglich	Prufen auf Undichtigkeiten	5	täglich

Tabelle 11: Wartungsvorgänge (eigene Darstellung)

Aus der Tabelle ist zu erkennen, dass allein an einer Kleberanlage der erste Wartungsvorgang 70 Minuten erfordert. Im Wartungsvorgang 2 sind es 35 Minuten. Hochgerechnet auf die Personenminuten für die Wartung aller Anlagen und Maschinen im gesamten Werk ist eine Überlastung durch unvorhersehbare Fehler oder auch Maschinenausfälle gegeben. Dadurch ist an der geforschten Anlage eine tägliche Wartung geplant, sodass eventuelle Ausfälle gegen 0 tendieren. Unerwartete Fehler, die auftreten, werden zunächst vom SCADA-System erkenntlich gemacht. Durch den Ersatz bestimmter Teile werden die Ausfälle und Fehler behoben. Dieser Vorgang ist bei der Klebedüse als Beispiel zu nennen. Ist die Düse verstopft, wird sie mechanisch ausgetauscht. Damit wird die Häufigkeit der Maschinenfehler oder auch Maschinenausfälle durch präventive und korrektive Maßnahmen zwar auf einen sehr geringen Anteil reduziert, jedoch ist eine Produktion gänzlich ohne Fehler utopisch. Also existieren durchaus geringe, aber unvorhersehbare Fehler. Infolgedessen wird anhand eines Predictive-Maintenance-Konzepts eine noch geringere Fehlerquote angestrebt. Um dieses Ziel zu erreichen, werden zum einen durch die in der Fallstudie erstellten Analysemethoden mit ihrer Visualisierung, bspw. Korrelationsmatrix, die von den Maschinen generierten Daten besser verstanden. Zum anderen können die Wartungs- und Instandhaltungsstrategien angepasst und verbessert werden, wenn Fehler vorhergesagt werden.

[17] Förderbänder.

5.2 Vorgehensweise und Vorbereitung des Datenmanagements

Die Vorgehensweise beruht auf dem im Kapitel 3 erwähnten ASUM-DM-Modell von IBM, das wiederum eine Weiterentwicklung des CRISP-DM-Modells ist. Der Vorteil des ASUM-DM-Modells ist die Eliminierung statistischer Schwachstellen des CRISP-DM-Modells. Beim ASUM-DM-Modell ist beim Prozess die Chronologie nicht entscheidend und jeder Prozess kann mehrmals durchlaufen werden. Dadurch können Schwachstellen schon am Anfang des Prozesses behoben werden, um Folgeprozesse zu optimieren.

Die Architektur, mit der gearbeitet wird, ist für die Entwicklung der Modelle entscheidend. Die aus der Arbeit entwickelte Big-Data-Architektur wird den Anforderungen zur Analyse gerecht. Die angewendeten Tools wurden nach der Nutzwertanalyse (Kapitel 4) selektiert. Die Arbeitsschritte spiegelt demnach auch die daraus entwickelte Architektur wider (Kapitel 4). Nochmals zur Veranschaulichung dient die Abbildung 56.

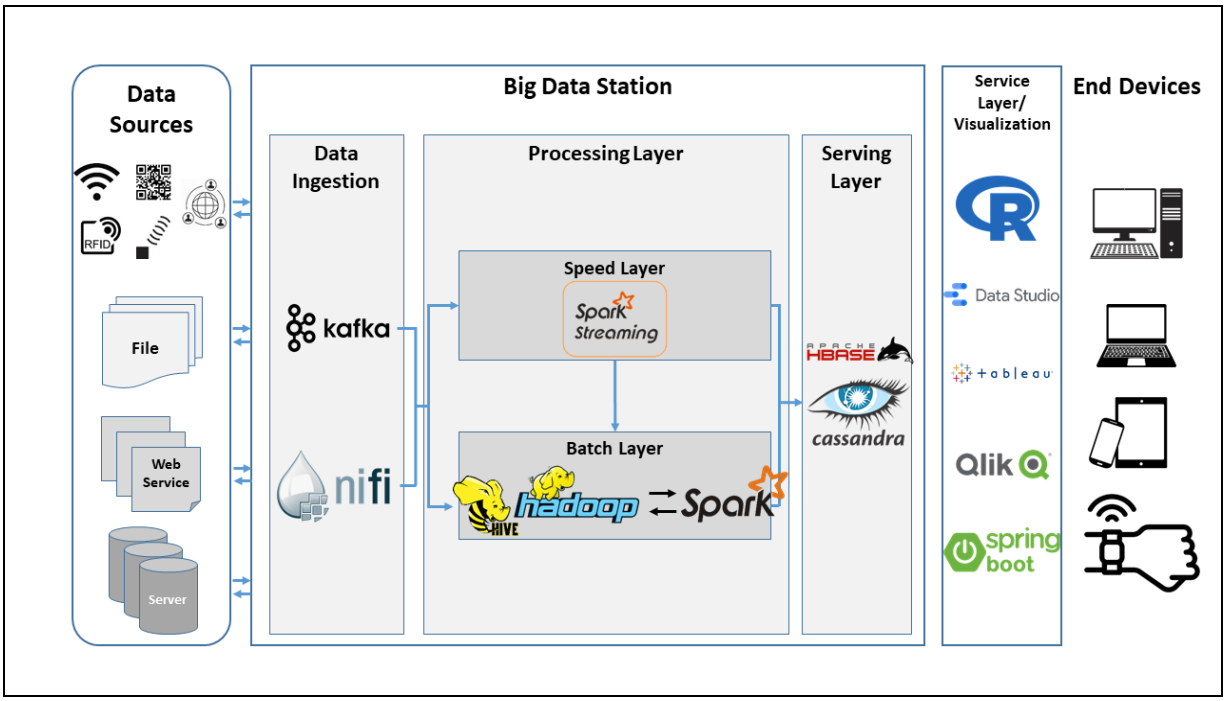


Abbildung 56: Referenzarchitektur basierend auf Lambda (eigene Darstellung)

Dabei wird die Lambda-Architektur genutzt, sodass eingehende Daten dupliziert und zur Berechnung an die Batch- und Speed Layer gesendet werden. Jeder Layer in der Lambda-Architektur erfüllt Teilaufgaben, die für verteilte Big-Data-Architekturen mit großen Datenmengen erforderlich sind. Die drei Teilaufgaben beinhalten: die Datenbeschaffung (Data Ingestion), die Datenverarbeitung/Datenanalyse (Processing Layer) und die Visualisierung. Für das Aufsetzen eines Predictive-Maintenance-Systems ist es wichtig, zu verstehen, wie der Datenfluss in dieser Architektur aussieht. Das Zusammenspiel der Big-Data-Tools bietet große

Vorteile für die Erreichung des Ziels, welches in den folgenden Abschnitten beispielhaft näher erläutert wird. In der Fallstudie wird der Großteil der Analysen im Batch Layer durchgeführt. Im Data Ingestion wird im Wesentlichen der komplette Datensatz von den Datenquellen mit Kafka in den Batch Layer kopiert. Wichtig hierbei ist es, dass die Daten unverändert in den Batch Layer übertragen werden. Es soll bestenfalls keine Vorverarbeitung der Daten durchgeführt werden, dies garantiert die Vollständigkeit der Daten, da System- und menschliche Fehler so nicht auftreten. Die Hauptaufgabe des Batch Layers ist die Verarbeitung der Daten mit hoher Genauigkeit. Maschine-Learning-Algorithmen benötigen in der Regel Zeit, bis ein Modell antrainiert ist, daher spielt die Verarbeitungsgeschwindigkeit hier nur eine untergeordnete Rolle. Ein wichtiger Bestandteil des Batch Layers ist das Hadoop-Framework, wovon hier nur das HDFS genutzt wird. Alle Daten werden im Hadoop File System (HDFS) gelagert. Das HDFS ist verteilt und fehlertolerant und folgt einem Append-Only-Ansatz, um die Anforderungen des Batch Layers der Lambda-Architektur zu erfüllen. Der Append Only ist eine Eigenschaft der Datenspeicherung, die besagt, dass der Datenspeicher nur als Datenspeicher genutzt wird und nicht für Verarbeitungen. Die Möglichkeit für die Nutzung von MapReduce ist ebenfalls gegeben, dennoch wurde in dieser Architektur darauf verzichtet. Gründe werden im Folgenden genannt. Ein Problem des Batch Layers ist die hohe Verarbeitungslatenz. Die Batch-Jobs müssen über den gesamten Datensatz ausgeführt werden und sind zeitaufwendig. Dies stellt eine ernsthafte Einschränkung für die Echtzeitdatenverarbeitung dar. Um diese Einschränkung zu überwinden, ist der Streaming Layer sehr wichtig.

Der Streaming Layer ist ebenfalls ein wichtiger Teil der Lambda-Architektur. Datenströme werden ohne Berücksichtigung von Vollständigkeit oder Korrekturen in Echtzeit verarbeitet. Der Streaming Layer kompensiert dadurch die hohe Latenz im Batch Layer. Ziel ist es, Echtzeitansichten der aktuellen Daten zu erstellen, dabei opfert der Streaming Layer den Durchsatz der Daten und verringert die Latenz erheblich. Die Echtzeitansichten werden unmittelbar nach dem Empfang der Daten generiert, sind jedoch nicht so vollständig und präzise wie im Batch Layer. Die Idee hinter diesem Design besteht darin, dass die genauen Ergebnisse des Batch Layers die Echtzeitansichten überschreiben, sobald sie eintreffen. Diese Trennung der Rollen in den verschiedenen Schichten zeichnet die Lambda-Architektur aus. Wie bereits erwähnt, nimmt der Batch Layer an einer ressourcenintensiven Operation teil, indem er über den gesamten Datensatz läuft. Daher muss der Streaming Layer einen anderen Ansatz verfolgen, um die Anforderungen an die geringe Latenz zu erfüllen.

Daher wird in dieser Arbeit auch sämtliche Machine-Learning-Algorithmen im Batch Layer und nur einige Echtzeitanalysen im Streaming Layer durchgeführt. Gerade in Bezug auf Predictive Maintenance ist dieser Ansatz sehr wichtig, da durchgehend Daten erzeugt werden, die für eine Vorhersage wichtig sind. So werden im Stream Layer neue Maschinendaten analysiert und parallel im Batch Layer das Maschine-Learning-Modell weiter trainiert.

Der Serving Layer ist u.a. ein Speicherort für die errechneten Ergebnisse. Dadurch ist die Aktualität des Serving Layers nach dem Abschluss des Batch-Jobs gegeben. Die Volatilität ist durch durchgängige Datenströme sehr hoch. Die Aktualität dieser Daten wird durch den Speed Layer aufrechterhalten. Der Speed Layer bearbeitet alle Daten, jedoch werden diese nicht auf einen festen Datenspeicher festgehalten. Die Daten werden in diesem Prozess „in-memory" verarbeitet und die Ergebnisse wiederum im Serving Layer abgespeichert. Dieser Vorgang beschreibt die Echtzeit und Aktualität der verarbeiteten Daten im Serving Layer.

Die Programmiersprache, die in der Fallstudie genutzt wird, ist Python. Möglich sind aber u.a. auch R, Java und Scala im Apache Spark. Apache Spark verfügt sowohl über Programmiersprachen-Schnittstellen, die sogenannten APIs, als auch über Schnittstellen zu den Sprachen R, Java und Scala. Grundsätzlich sind alle Sprachen für Datenanalysen geeignet. Trotz der schnelleren Arbeitsgeschwindigkeit der Programmiersprachen Scala und Java ist in dieser Arbeit die Wahl auf Python gefallen. Diese wird im Spark-Framework auch PySpark genannt. Die zahlreiche Unterstützung von Machine Learning und weiteren statistischen Bibliotheken durch Python war letztlich ausschlaggebend für die Entscheidung. Im Allgemeinen lässt sich sagen, dass Python mehr analytisch, während Scala mehr ingenieurwissenschaftlich orientiert ist. In Abbildung 57 ist das Ergebnis einer Umfrage aus dem Jahr 2019 von KDnuggets zu den am häufigsten eingesetzten Tools in den Bereichen Analytics, Data Science und Machine Learning zu sehen.

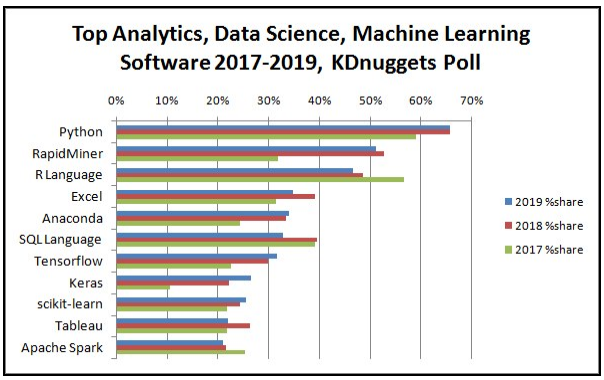


Abbildung 57: Ranking Programmiersprachen (eigene Darstellung nach KDnuggets, 2019)

5.3 Analyse zum Predictive-Maintenance-Ansatz

Nachdem im vorherigen Abschnitt Grundlagen gelegt wurden, folgt nun die Beschreibung des Datensatzes. Zunächst werden die Rohdaten beschrieben. Es wird erläutert, woher die Daten stammen und in welchem Format sie vorliegen. Dieser Abschnitt beinhaltet die komplette Aufbereitung der Daten sowie erste Analysen der Daten für das bessere Datenverständnis. Die folgenden Ergebnisse und einzelnen Schritte sind als iterative Ergebnisse anzusehen. Erst durch Erkenntnisse bzw. Probleme bei der Modellerstellung mussten weitere Bereinigungsmaßnahmen durchgeführt werden. Hier wird die Bereinigung dennoch als ein eigenständiger Prozess angesehen.

5.3.1 Datenbasis

Die in der Arbeit genutzten Daten basieren auf Echtzeitdaten aus fünf aktiven Maschinen der Kleberanlage aus dem Karosseriewerk. Zur Erstellung eines Vorhersagemodells wurden vom OEM monatliche Datensätze bereitgestellt, diese liegen in Form von Log-Dateien vor. Ausgegeben werden die Dateien als CSV-Daten und sind zu jeder Zeit abrufbar, sodass immer aktuelle Datensätze herangezogen werden können. Die bereitgestellten Daten umfassen die Daten von fünf Maschinen in einem Zeitraum von ca. acht Monaten. Diese wurden für diesen Zweck ein Jahr lang gesammelt, da sich die Daten nach einem Monat automatisch löschen. Wichtig dabei ist, dass das System im ursprünglichen Sinne nicht für die Verwendung im Rahmen eines Predictive-Maintenance-Systems gedacht war. Derzeit werden diese Daten lediglich für Monitoring-Zwecke genutzt. Hier handelt es sich detailliert um Dosierer-Statistiken der Kleberanlagen. Diese umfassen 35 Attribute, bei denen jeder Datensatz mit einem Zeitstempel versehen ist. Von den 35 Attributen sind nur bei zwölf bekannt, was diese Attribute aussagen. Unter den Attributen werden Zustandsdaten wie bspw. Fehlercodes, aktuelle Drehzahl oder der Mitteldruck gesammelt (siehe Tabelle 12). Hervorzuheben ist, dass die Maschinen täglich gewartet werden. Daher sind im Datensatz sehr wenige Fehler zu finden.

Zeitstempel	Startzeit des Klebeprozesses
Soll Klebemenge	Soll-Menge des zu beklebenden Teils in $ccm^3/100$
Ist Klebemenge	Ist-Menge des zu beklebenden Teils in $ccm^3/100$
Zeit	Vorgangszeit des Klebeprozesses. Nicht die Betriebszeit in Sekunden/100
Drehzahl	Das Drehmoment eines Bauteils in der Klebemaschine in %/10
Mitteldruck	Mitteldruck der Klebemaschine in bar

Maximaler Druck	Maximaler Druck der Klebemaschine in bar
Programmnummer	Auswahl der Klebeprogramme
Fehlernummer	Art des Fehlers
Dosierungszeit	Die Zeit, die benötigt wurde, um den Kleber zu dosieren, in Sekunden

Tabelle 12: Attribute Kleberanlage (eigene Darstellung)

Zusätzlich zu der Dosierer-Statistik gibt es noch eine weitere Tabelle, welche alle einzelnen Fehlercodes näher beschreibt. Insgesamt gibt es 243 verschiedene Fehlermeldungen, wobei manche sehr häufig, andere jedoch so gut wie nie auftreten. Anzumerken ist, dass nach Anforderung eines neuen Datensatzes eine neue CSV-Datei erstellt wird. Datensätze müssen somit von allen fünf Maschinen angefordert werden. Die Anforderung einer kumulierten Ansicht über alle Datenzeiträume hinweg ist nicht möglich. Auch in dieser Fallstudie wurde mit mehreren CSV-Dateien gearbeitet. Kumuliert existieren ca. 165900 Einträge.

Um die Daten an das HDFS zu senden, muss zunächst Kafka mit der Datenquelle verbunden werden. Die CSV-Dateien werden durch das SPS-System (über WinCC) direkt von der Maschine generiert. Dateien, die generiert werden, sind durch das interne Netzwerk im Karosseriewerk zu jeder Zeit durch TCP-IP erreichbar. Damit sich Kafka mit dieser Datenquelle verbinden kann, muss ein Kafka-Producer erstellt werden. Dazu wird ein Kafka-Producer geschrieben, der sich mit den bestimmten TCP-Port verbindet. Dies kann durch andere Tools geschehen. In dieser Arbeit kristallisierte sich Kafka Connect heraus, da dieser schon im Kafka-Framework integriert ist. Für die Verbindung von Kafka und HDFS wird ebenfalls Kafka Connect genutzt. Für diesen Zweck gibt es im Kafka-Framework einen speziellen HDFS-Connector. Hierfür wird ein Kafka-Consumer geschrieben, der die Dateien ins HDFS schreibt (siehe Abbildung 58).

Abbildung 58: Kafka Connect (eigene Darstellung)

Für den Producer wurde dabei folgende Config.-Datei genutzt:

```
name = socket-connector
connector.class = org.apache.kafka.connect.socket.SocketSourceConnector
tasks.max = 1
topic=topic
schema.name = socketschema
```

```
port = 12345
batch.size = 100
```

Für den Consumer wurde diese Config.-Datei verwendet:

```
name = hdfs-sink
connector.class = io.confluent.connect.hdfs.HdfsSinkConnector
tasks.max = 1
topic = hdfs
hdfs.url = hdfs://localhost:9000
flush.size = 100
```

5.3.2 Datenaufbereitung

Nach der Einrichtung von Kafka-Connect erfolgt nun die Datenaufbereitung. Im ersten Schritt müssen die zu verarbeitenden Dateien in Spark geladen werden. Der Leseprozess soll hierbei automatisch funktionieren. Um das zu ermöglichen, müssen alle Pfade der CSV-Dateien vor dem Laden zunächst in Arrays geschrieben werden. Auf dem HDFS-Cluster landen in den bestimmten Zeitabschnitten immer wieder neue, aktuelle CSV-Dateien. In dieser Zeit entsteht eine große Datenmenge, die gelesen werden muss. Durch die Abfrage der Dateipfade ist kein manuelles Hinzufügen der aktuellen Dateien notwendig. Diese Abfrage wurde durch drei einfache Funktionen in Python gelöst. Dieser Schritt ist im späteren Aufbereitungsverlauf auch wichtig, da aus den Pfaden bzw. Dateinamen die Maschinen-IP-Nummer ermittelt werden kann. In dieser Arbeit wurde mit fünf Maschinen gearbeitet, somit werden fünf Arrays für jede einzelne Maschine erstellt. Mit den Arrays der Dateipfade werden fünf Data-Frames erstellt. Zusätzlich wird zu jedem Data-Frame noch eine weitere Spalte mit den Maschinen-IP-Nummern hinzugefügt. Um eine aussagekräftige Analyse durchzuführen, werden nun alle Data-Frames miteinander verbunden. Im letzten Schritt wird noch die Kopfzeile für das Data-Frame hinzugefügt.

```
#Wende Funktion an, um Dateipfade zu erhalten
m150l = Erhaltepfade(150)
m195l = Erhaltepfade(195)
m226l = Erhaltepfade(226)
m200l = Erhaltepfade(200)
m214l = Erhaltepfade(214)

#Erstelle die Data-Frames
m150 = ErstelleAusPfadListeEinDataFrame(m150l)
m195 = ErstelleAusPfadListeEinDataFrame(m195l)
m226 = ErstelleAusPfadListeEinDataFrame(m226l)
m200 = ErstelleAusPfadListeEinDataFrame(m200l)
m214 = ErstelleAusPfadListeEinDataFrame(m214l)
```

```
#Führe sämtliche Bereinigung und Korrekturen durch + Füge Spalten
hinzu
m150 = Bereinigung(m150)
m200 = Bereinigung(m200)
m195 = Bereinigung(m195)
m226 = Bereinigung(m226)
m214 = Bereinigung(m214)

#Führe alle Data-Frames zusammen
m150["machine_ip"] = "M150"
m195["machine_ip"] = "M195"
m200["machine_ip"] = "M200"
m214["machine_ip"] = "M214"
m226["machine_ip"] = "M226"
fulldata = pd.concat([m150, m195, m200, m226, m214])
```

Um eine erfolgreiche Analyse der Daten durchzuführen, ist es existenziell, die eingelesenen Daten aufzubereiten. Die Abbildung 59 zeigt dabei einen Auszug aus den Rohdaten, die aus den Maschinen gewonnen werden.

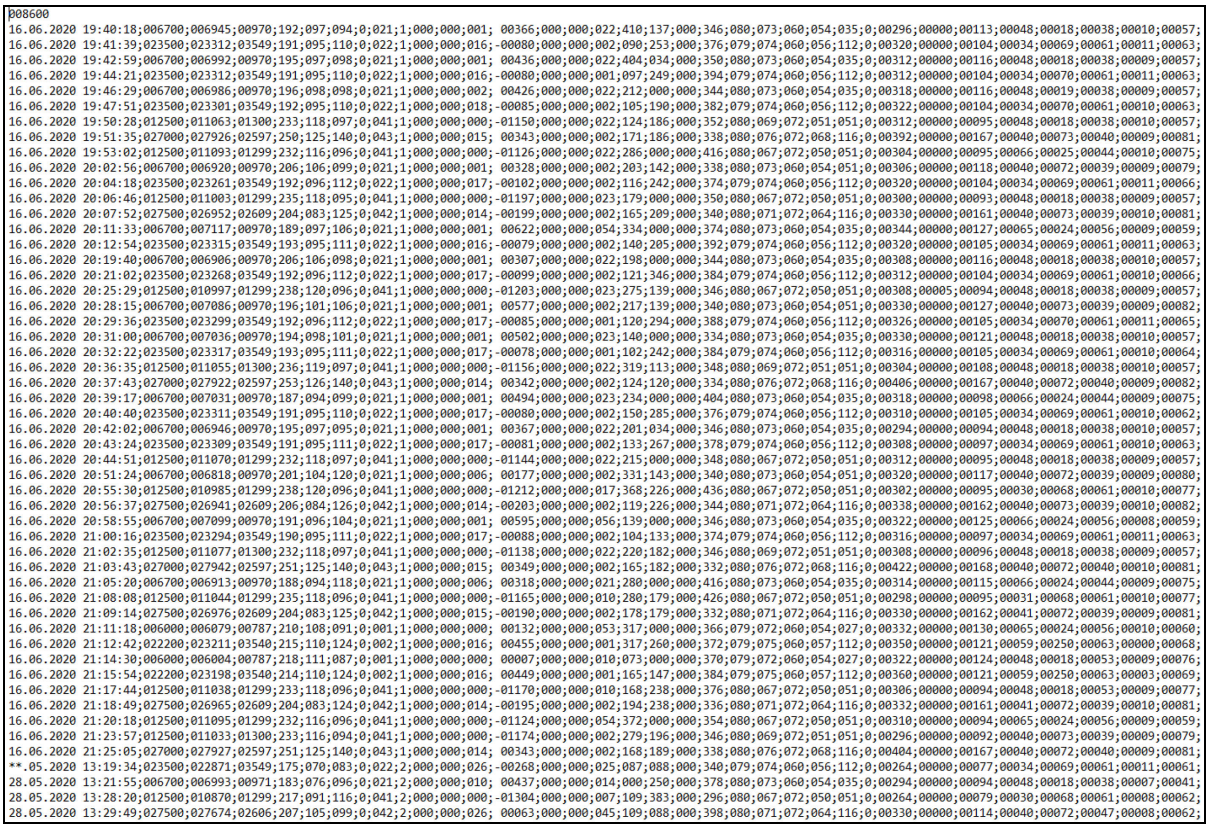


Abbildung 59: Auszug aus den generierten Rohdaten (eigene Darstellung)

Die Aufbereitung der Daten zählt dabei zu den wichtigsten Tätigkeiten jeder Datenanalyse. Bereits kleinste Veränderungen können starke Auswirkungen auf die Ergebnisse haben (vgl.

Molnár et al., 2017, S. 3). Das Gleiche gilt auch für Machine-Learning-Verfahren (vgl. Eid et al., 2017, S. 707).

Beim Laden der Dateien wurden in einigen Spalten falsche Datentypen ausgewählt. Diese werden zunächst in den richtigen Datentyp geändert. Bei einfachen Zahlen gab es dabei keine Probleme, den Datentypen zu ändern, dennoch tritt ein Fehler im Dateityp Date auf. Dieser muss im Zeitstempel erst in das richtige Format formatiert werden, da PySpark mit dem amerikanischen Zeitstempel 'd-m-Y H:M:S' arbeitet. Im nächsten Schritt wurden Duplikate und fehlerhafte Daten aus dem Data-Frame entfernt. Im Datensatz befinden sich mehrere Einträge, welche eine ungewöhnliche Zeichenfolge sowie leere Zellen aufweisen. Diese wurden ebenfalls entfernt. Auch dazu wurde eine Funktion geschrieben. Diese Funktion behandelt lediglich die oben genannten Fehler. Sollten neuartige Probleme auftauchen, muss die Funktion neu geschrieben werden. Wie in den oben genannten Kapiteln tauchen Fehler eher selten auf. Vereinzelte Fehler treten somit nie oder nur einmal auf. Um diesem Problem ein wenig entgegenzuwirken, wurden mehrere Fehlerklassen erstellt, anstatt mit den Fehlercodes zu arbeiten. Dadurch verringert sich die Zahl der Fehlerklassen erheblich und seltene Fehler können mit anderen gruppiert werden. Für die Erstellung der Fehlerklassen wurde eine Tabelle mit einzelnen Fehlerbeschreibungen genutzt. Nach intensiver Untersuchung der Fehlerbeschreibungen konnten folgende Fehlerklassen ermittelt werden: Temperatur, Hochdruckfehler, Dosierfehler, Referenzschalter, Pistolenmotor, Dosiermotor, Fatale Motorfehler, Behälterfehler, Füllhahndosierung, Proportionalventil, Pistole, Spannungsüberwachung, Düsenabstand, Mengenfehler, Netzwerkfehler, Fasspresse, Klebstoff/Material. Folgende Tabelle kann im Anhang entnommen werden. Durch die Erstellung konnten die Klassenmenge von 243 auf 17 reduziert werden. Da später auch mit der Überlebenszeit der Maschinen gearbeitet wird, wurde eine weitere Spalte erstellt, welche eine kumulierte Zeit der Betriebszeit berechnet. Diese wurde aus der Differenz der Zeitstempel berechnet. Da am Wochenende, Feiertagen sowie an anderen speziellen Tagen nicht gearbeitet wird, wurde zusätzlich eine Funktion geschrieben, welches dieses Problem löst. Normalerweise dauert ein Klebeprozess nur einige Minuten. Die Funktion filtert Prozesse heraus, welche länger als zehn Minuten dauern, und zählt diese nicht zur Betriebszeit hinzu.

TimeStamp	TargetAdhes	ActualAdhes	Time	Torque	Medium Pre	Maximum Pr	8	Program	RAM	ErrorCode	12	DosFullTime	Variance	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Errcateg	machine_ip
10/2/2019 2:56	6400	6849	810	219	120	179	0	41	2	0	0	9	701	0	0	34	0	0	0	504	79	72	60	54	33	0	416	0	174	26	12	50	13	74	normal	M150
10/2/2019 2:57	16800	18877	2070	233	113	155	0	42	2	0	0	0	1236	0	0	10	0	7	0	520	79	35	60	26	76	0	399	0	150	22	12	61	14	83	normal	M150
10/2/2019 2:58	6400	6883	810	220	119	180	0	41	2	0	0	8	755	0	0	29	0	0	0	496	79	72	60	54	76	0	407	10	176	26	13	50	14	74	normal	M150
10/2/2019 3:00	19500	19323	2078	259	105	144	0	82	2	0	0	15	-91	0	0	11	0	4	0	536	80	52	60	39	73	0	363	0	145	21	11	61	14	83	normal	M150
10/2/2019 3:01	14700	14784	1727	261	114	116	0	81	2	0	0	7	57	0	0	29	0	0	0	502	79	68	60	51	61	0	420	0	167	19	12	47	14	63	normal	M150
10/2/2019 3:03	16800	18861	2069	239	111	152	0	42	2	0	0	0	1227	0	0	0	0	45	0	704	79	35	60	26	76	0	372	0	148	45	72	61	13	99	normal	M150
10/2/2019 3:03	6400	6882	809	219	118	179	0	41	2	0	0	10	753	0	0	28	0	0	0	508	79	72	60	54	33	0	409	0	169	25	12	50	13	73	normal	M150
10/2/2019 3:09	16800	18772	2069	237	113	150	0	42	2	0	0	0	1174	0	0	10	0	0	0	530	79	35	60	26	76	0	374	0	152	22	12	61	13	82	normal	M150
10/2/2019 3:10	6400	6848	810	217	119	177	0	41	2	0	0	10	700	0	0	34	0	0	0	488	79	72	60	54	76	0	412	0	174	26	12	50	13	71	normal	M150
10/2/2019 3:30	16800	18776	2068	238	115	152	0	42	2	0	0	0	1176	0	0	10	0	0	0	496	79	35	60	26	76	0	386	0	150	21	12	61	14	82	normal	M150
10/2/2019 3:30	6400	6845	810	218	120	179	0	41	2	0	0	9	695	0	0	35	0	0	0	508	79	72	60	54	76	0	425	0	174	25	13	50	13	73	normal	M150
10/2/2019 3:33	19500	19276	2078	256	105	143	0	82	2	0	0	14	-115	0	0	10	0	1	0	544	80	52	60	39	73	0	352	0	141	21	12	61	14	83	normal	M150
10/2/2019 3:34	14700	14733	1727	259	114	114	0	81	2	0	0	6	23	0	0	31	0	0	0	516	79	68	60	51	61	0	396	0	160	20	12	47	14	63	normal	M150
10/2/2019 3:35	16800	18900	2069	239	112	154	0	42	2	0	0	0	1250	0	0	0	0	0	0	702	79	35	60	26	76	0	401	0	155	45	72	61	14	98	normal	M150
10/2/2019 3:36	6400	6894	810	223	119	180	0	41	2	0	0	8	772	0	0	27	0	0	0	482	79	72	60	54	76	0	418	0	173	26	12	50	13	75	normal	M150
10/2/2019 3:45	16800	18764	2068	236	114	151	0	42	2	0	0	0	1169	0	0	10	0	144	0	526	79	35	60	26	76	0	386	0	149	22	12	61	14	83	normal	M150
10/2/2019 3:46	6400	6859	810	221	120	179	0	41	2	0	0	9	717	0	0	35	0	0	0	494	79	72	60	54	33	0	423	0	176	25	12	50	14	72	normal	M150
10/2/2019 3:53	16800	18812	2069	240	114	154	0	42	2	0	0	0	1198	0	0	8	0	116	0	482	79	35	60	26	76	0	399	0	155	21	12	61	14	83	normal	M150
10/2/2019 3:53	6400	6862	810	222	120	180	0	41	2	0	0	9	722	0	0	33	0	0	0	542	79	72	60	54	76	0	419	0	179	26	12	50	14	73	normal	M150
10/2/2019 4:06	19500	19238	2078	260	107	142	0	82	2	0	0	18	-134	0	0	9	0	0	0	506	80	52	60	39	73	0	368	0	113	21	11	61	14	83	normal	M150
10/2/2019 4:06	14700	14709	1727	263	116	114	0	81	2	0	0	7	6	0	0	9	0	0	0	506	79	68	60	51	61	0	404	0	156	21	11	61	14	83	normal	M150
10/2/2019 4:10	16800	18842	2069	239	113	153	0	42	2	0	0	0	1216	0	0	0	0	0	0	498	79	35	60	26	76	0	388	0	148	45	73	47	13	99	normal	M150
10/2/2019 4:11	6400	6876	810	220	120	181	0	41	2	0	0	8	744	0	0	30	0	0	0	478	79	72	60	54	76	0	425	0	178	25	13	50	14	73	normal	M150
10/2/2019 4:13	16800	18828	2069	235	113	152	0	42	2	0	0	0	1207	0	0	8	0	131	0	508	79	35	60	26	76	0	368	0	147	22	12	61	14	83	normal	M150
10/2/2019 4:14	6400	6883	810	220	119	180	0	41	2	0	0	9	754	0	0	31	0	0	0	490	79	72	60	54	76	0	413	0	171	26	12	50	13	73	normal	M150
10/2/2019 4:19	19500	19284	2073	257	105	145	0	82	2	0	0	15	-111	0	0	8	0	2	0	540	80	52	60	39	73	0	371	0	148	21	11	61	14	83	normal	M150
10/2/2019 4:20	14700	14769	1727	263	115	116	0	81	2	0	0	6	47	0	0	30	0	0	0	512	79	68	60	51	61	0	388	0	167	20	12	47	14	65	normal	M150
10/2/2019 4:25	19500	19250	2073	258	106	143	0	82	2	0	0	15	-128	0	0	0	0	84	0	476	80	52	60	39	73	0	363	0	139	45	72	61	14	99	normal	M150
10/2/2019 4:26	14700	14752	1727	261	114	115	0	81	2	0	0	6	35	0	0	0	0	0	0	476	79	68	60	51	61	0	412	0	162	45	72	61	14	99	normal	M150
10/2/2019 4:30	16800	18819	2069	237	113	152	0	42	2	0	0	0	1202	0	0	0	0	19	0	556	79	35	60	26	76	0	382	0	158	45	73	47	14	99	normal	M150
10/2/2019 4:31	6400	6865	810	222	120	179	0	41	2	0	0	9	726	0	0	31	0	0	0	496	79	72	60	54	76	0	412	10	176	26	12	50	14	73	normal	M150
10/2/2019 4:33	16800	18842	2069	239	112	152	0	42	2	0	0	0	1216	0	0	8	0	1	0	508	79	35	60	26	76	0	376	0	148	22	12	61	13	82	normal	M150
10/2/2019 4:33	6400	6877	810	221	118	180	0	41	2	0	0	8	745	0	0	30	0	0	0	486	79	72	60	54	33	0	421	0	178	25	12	50	13	73	normal	M150
10/2/2019 4:42	16800	18762	2068	237	114	151	0	42	2	0	0	0	1168	0	0	9	0	0	0	532	79	35	60	26	76	0	380	0	150	21	11	61	14	83	normal	M150
10/2/2019 4:42	6400	6838	810	220	120	177	0	41	2	0	0	9	685	0	0	35	0	0	0	504	79	72	60	54	76	0	426	0	173	25	13	50	13	72	normal	M150

Abbildung 60: Auszug aus bereinigter Tabelle (eigene Darstellung)

Nach dem kompletten Datenbereinigungsprozess ergibt sich eine Gesamttabelle mit allen wichtigen Daten (siehe Abbildung 60). Aus dieser Gesamttabelle werden dann zusätzlich fünf einzelne Tabellen mit den sogenannten Überlebenszeit der Maschinen erstellt.

5.3.3 Datenexploration

Für einen erweiterten Einblick auf die Daten werden zunächst erste einfache Visualisierungen durchgeführt. Für die ersten Analysen und Visualisierungen werden die Python-Bibliotheken Pandas, Numpy und Matplotlib genutzt (siehe Tabelle 13).

Bibliothek	Beschreibung
matplotlib	Matplotlib ist eine Python-2D-Plottbibliothek, die plattformübergreifend Zahlen in einer Vielzahl von Diagrammen und interaktiven Umgebungen erstellen kann (Matplotlib Org., 2019).
numpy	Dient für die einfache Handhabung von Vektoren, Matrizen oder mehrdimensionalen Arrays (NumPy, 2017).
pandas	Dient zur Verwaltung von Daten und deren Analyse (Panda.PyData, 2019)

Tabelle 13: Python-Bibliotheken (eigene Darstellung)

Diese Bibliotheken werden im Laufe der Fallstudie auch wieder genutzt, um einfache Berechnungen durchzuführen. Später in der Fallstudie wird aber hauptsächlich mit Spark-Data-Frames anstatt mit Panda-Data-Frames gearbeitet, da später hauptsächlich Bibliotheken von Spark genutzt werden und diese nur mit Spark-Data-Frames funktionieren. Für erste

Erkenntnisse der Daten waren diese Bibliotheken dennoch ausreichend. In der ersten Graphik (Abb. X) wurde zunächst einfach dargestellt, wie der Klebeverbrauch der einzelnen Maschinen aussieht. Auffällig sind die unterschiedlichen verbrauchten Klebemengen der Maschinen. Das ist der erste Indikator, welcher darauf hinweist, dass verschiedene Maschinen verschiedene Teile bekleben. Diese Information ist sehr wichtig für die Erstellung eines Modells. Da der Verbrauch der Klebemenge verschieden ist und andere Programmnummern für verschiedene Teile verwendet werden, ist davon auszugehen, dass dies auch Auswirkungen auf die Fehler hat. Somit muss für jede Maschine ein eigenes Modell entwickelt werden, da jede Maschine anders arbeitet, obwohl es sich hier um das gleiche Kleberanlage-Modell handelt. In der Abbildung 61 sind die Abweichungen bezüglich der Soll- und Ist-Werte dargestellt.

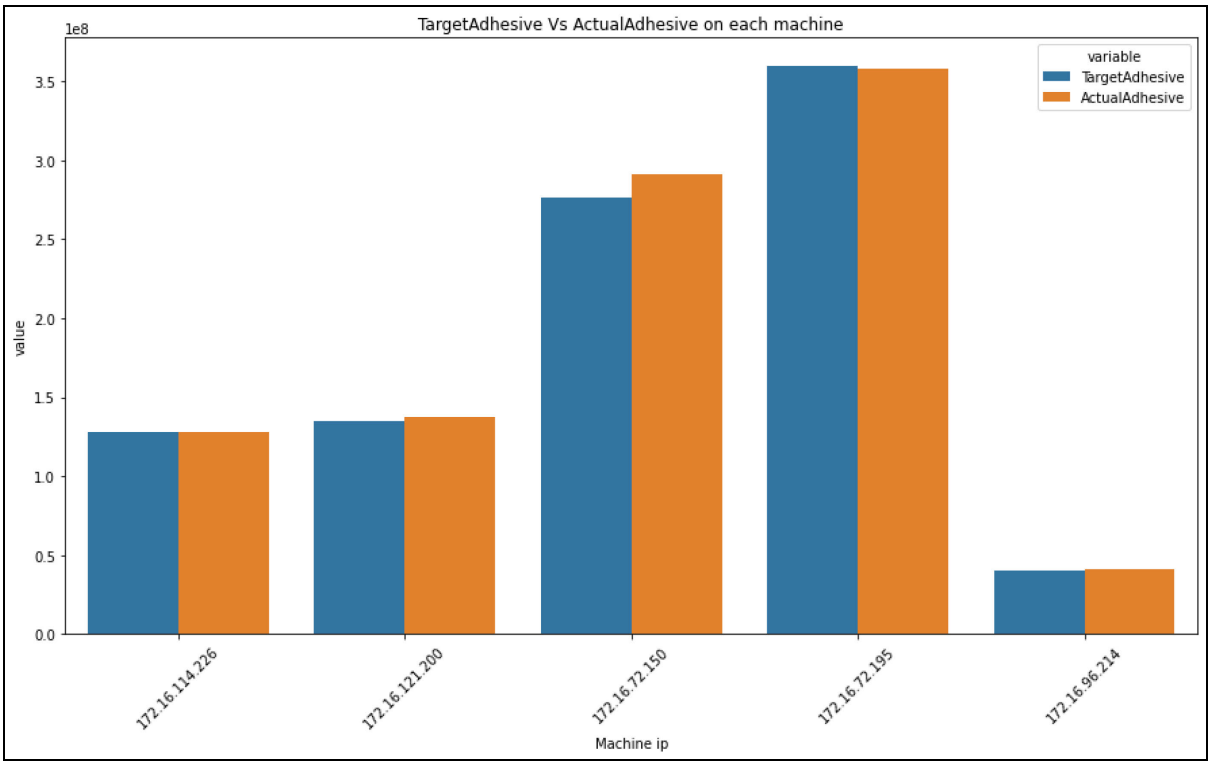


Abbildung 61: Soll/Ist-Wert nach Maschinen (eigene Darstellung)

In weiteren Analysen wurde ebenfalls herausgefunden, dass es enge Zusammenhänge zwischen Drehzahl und einigen Fehlercodes gibt. Das gilt auch für die durchschnittliche Dosierungszeit und einige Fehlercodes. Auffällig ist auch der enge Zusammenhang zwischen der Drehzahl und dem Mitteldruck (siehe Abbildung 62).

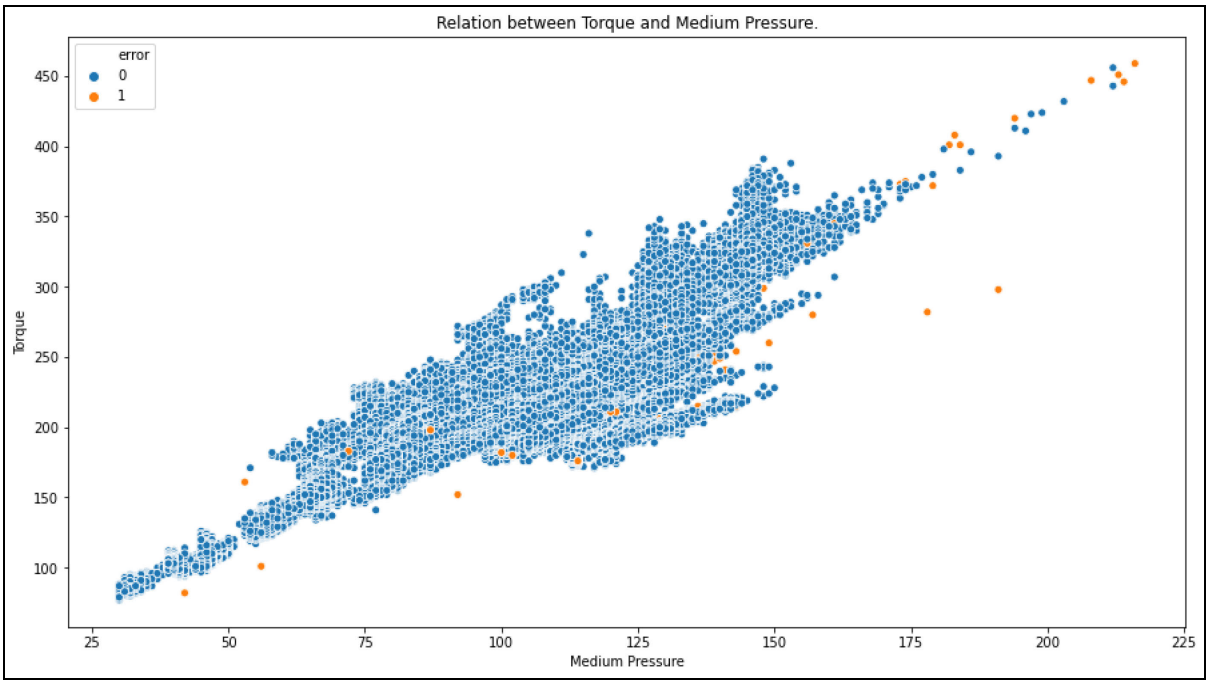


Abbildung 62: Relation Druck und Drehzahl (eigene Darstellung)

Eine genaue Korrelationsmatrix wird dann im folgenden Kapitel erstellt, um zu sehen, wie einzelne Parameter miteinander zusammenhängen.

Die in dem Abschnitt dargestellten Abbildungen dienen lediglich zur Informationsgewinnung der Daten und stellen noch keine Erkenntnisse zum Predictive-Maintenance-Konzept dar. Bei der Datenexploration wurden noch weitere Graphiken und Tabellen erstellt, welche aber für die weitere Bearbeitung der Daten nicht notwendig sind. All diese weiteren Tabellen und Diagramme befinden sich im Anhang.

Beim vorliegenden Datensatz besteht das Problem, dass sehr wenig Daten mit Fehlercodes vorhanden sind. Bei dem Fall, dass alle fünf Anlagen einzeln betrachtet werden müssen, reduzieren sich die Daten weiter. Es steht nur ein Datensatz von einem Jahr zur Verfügung. Dieser umfasst noch zu große Produktionspausen durch besondere unerwartete Ereignisse. Kumuliert stehen in dieser Fallstudie Produktionsdaten aus weniger als acht Monaten zur Verfügung. Ein weiteres Problem besteht auch bei der Fehlerklasse an sich. Im Werk werden alle Anlagen täglich gewartet und auf Fehler überprüft, um Produktionstopps zu vermeiden. Dadurch fällt die Fehlerklasse sehr klein aus. Infolgedessen lässt sich auch vermuten, dass es sich hier bei der Mehrheit der Fehler um unerwartete Fehler handelt. Es wird nicht davon ausgegangen, dass es einen generellen Verursacher für die Fehler gibt, es könnten lediglich Indikatoren für bestimmte Fehler gefunden werden. Durch die geringe Anzahl der Fehler ist es auch unmöglich, gute Modelle zu erstellen. Man spricht hier auch von einem Ungleichgewicht

der Klassen. Im Folgenden wird dennoch versucht, trotz dieser Einschränkungen ein Predictive-Maintenance-System aufzusetzen. Dabei gilt bereits die Erstellung eines Modells als ein Teilerfolg.

5.3.4 Implementierung von PM mit Apache Spark in die Big-Data-Architektur

Nachdem nun ein Großteil der Daten teilbereinigt wurde, wird in diesem Abschnitt näher auf die Datenanalyse und die Erstellung des Modells eingegangen. Bei diesem Prozess wird die Forschungsfrage aufgegriffen, inwiefern das Problem gelöst werden kann, ein Predictive-Maintenance-Modell trotz beschränkter Datenverfügbarkeit aufzusetzen. Um dieses Ziel zu erreichen, wird versucht, verschiedene Teilergebnisse zu einer Gesamtlösung zusammenzufassen. Dabei wird der iterative Prozess in mehreren Versuchsreihen durchgeführt. Zunächst werden Machine-Learning-Methoden angewendet, um weitere Erkenntnisse über die Daten zu erhalten. Erste Analyseergebnisse können hierbei als Teillösung angesehen werden. Daraufhin wird ein Predictive-Maintenance-Modell mithilfe von Ereigniszeitanalysen aufgesetzt.

5.3.4.1 Versuchsreihe 1: Anwendung von Standard-Machine-Learning-Methoden

In PySpark gibt es bereits eine Machine-Learning-Bibliothek, welche auch in dieser Fallstudie genutzt wird. Dabei handelt es sich um die bereits genannte Bibliothek MLlib. MLlib ist dabei Bestandteil der Spark-Machine-Learning-Bibliothek, welches das Ziel hat, Maschinelles Lernen skalierbar und einfach zu machen. Zu den wichtigsten Machine-Learning-Algorithmen von MLlib gehört die Erstellung von Klassifikationsmodellen (logistische Regression, Naiver Bayes etc.), Regressionsmodellen (lineares Modell, Cox-Regression etc.), Entscheidungsbaummodellen, Random-Forest-Modellen, Clustering-Modellen wie K-means, sowie Kollaborativen Filtermodellen. Für die ersten Versuche werden zunächst einfache Modelle angewendet, um zu testen, wie gut die Datenqualität ist. Bei diesen Datensatz handelt es sich grundsätzlich um einen strukturierten Datensatz. Die Klassen sind bereits definiert, daher werden überwachte Machine-Learning-Methoden gewählt. Im Folgenden werden Klassifikationsmodelle erstellt.

Lediglich die fehlenden Informationen über die Parameter könnten die Ergebnisse stark beeinflussen. In der ersten Versuchsreihe wurde eine klassische Korrelationsmatrix mit allen Daten erstellt, um zu erkennen, wie Parameter miteinander zusammenhängen können (siehe Abbildung 63).

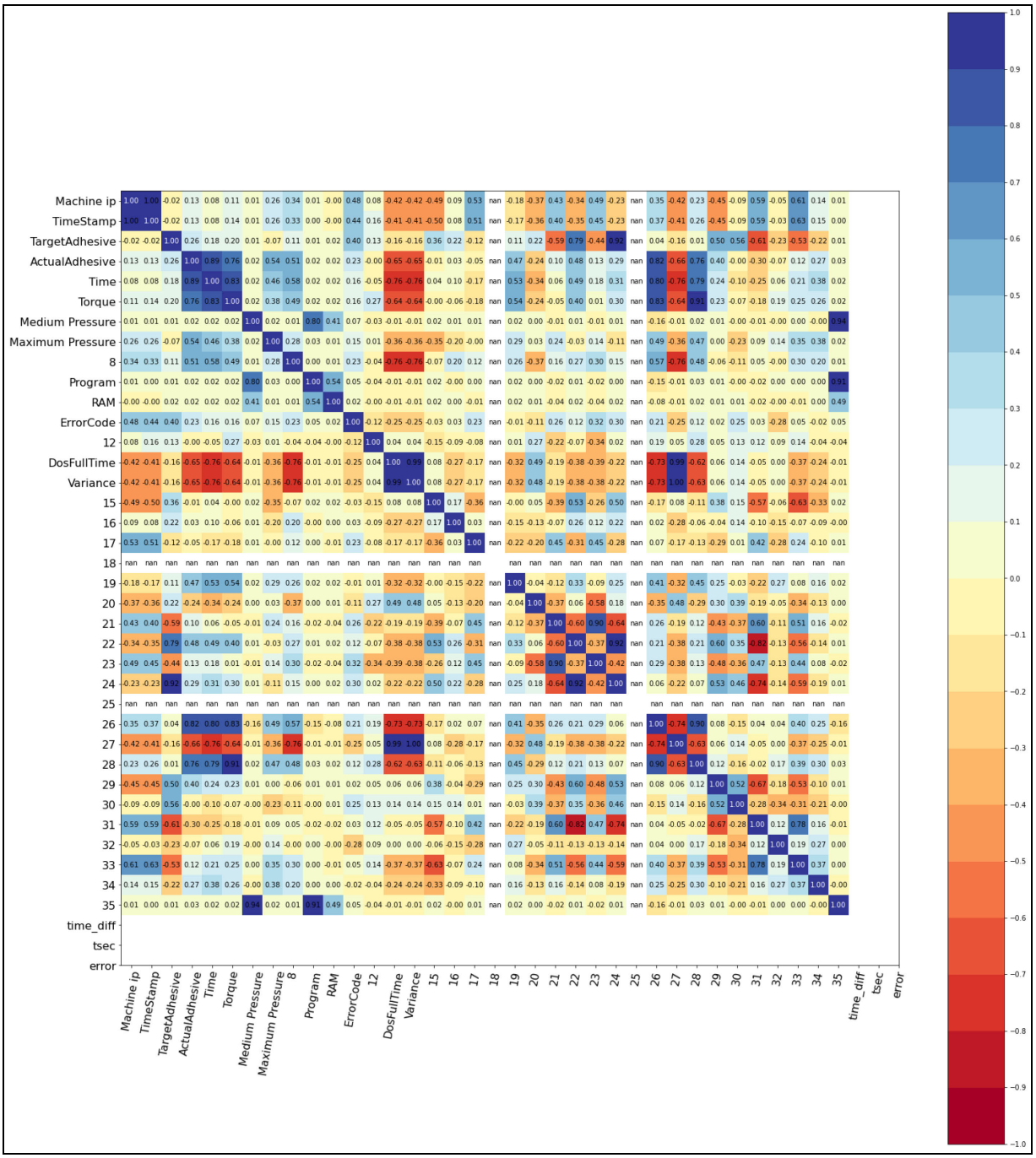


Abbildung 63: Korrelationsmatrix Kleberanlage (eigene Darstellung)

In dieser Korrelationsmatrix sind nicht alle Werte aussagekräftig, da darin auch boolesche Werte enthalten sind, welche für Korrelationen nicht aussagekräftig sind. Auch bei dem Parameter, bei dem keine Information gegeben ist, sind die Korrelationswerte als fraglich anzusehen. In dieser Matrix stehen verschiedene Korrelationskoeffizienten, welche Werte zwischen -1 und +1 annehmen können. Dabei steht +1 für einen starken positiven linearen Zusammenhang und -1 für einen starken negativen linearen Zusammenhang. Der Fokus in der Korrelationsmatrix liegt auf den Parametern, die bekannt sind. Gut zu erkennen ist, dass es

mehrere Korrelationen innerhalb dieser Parameter gibt, sowohl positive als auch negative Zusammenhänge. Diese erste Analyse gibt zunächst ein erstes Teilergebnis an, welches noch keine Aussagekraft über die Vorhersage der Fehler hat. Da es keine Aussagekraft für die Parameter gibt, bei denen die Informationen nicht bekannt sind, werden diese in den vielen Analysen nicht berücksichtigt. Lediglich bei einigen Modellen werden diese Werte miteinbezogen, um festzustellen, wie diese die Ergebnisse beeinflussen können. Gerade bei Modellen mit unüberwachtem Lernen können diese Werte genutzt werden. Da keine Informationen über diese Parameter gegeben sind, gibt es ebenso die Möglichkeit, dass diese Informationen für die Vorhersage von Fehlern irrelevant sind. So könnten diese Werte das Ergebnis beim überwachten Lernen entweder verbessern oder verschlechtern.

Im Folgenden wird ein Machine-Learning-Algorithmus angewendet, um zu sehen, wie gut sich Fehler voraussagen lassen. Daher wird eine logistische Regression durchgeführt. Zwar lassen sich so keine Fehler voraussagen, aber es kann analysiert werden, ob andere Parameter Fehler beeinflussen. Bevor mit der logistischen Regression begonnen werden kann, muss der Datensatz wieder vorbereitet werden. Zunächst muss die zu untersuchende Klasse ausgewählt werden. Diese erfordert eine One-Hot-Codierung, d.h. diese Klasse darf nur zwei Zustände besitzen. Diese zwei Zustände sind in diesem Fall: Fehler ist aufgetreten (1) und Fehler ist nicht aufgetreten (0). In dieser Fallstudie bedeutet das, dass immer jeweils nur ein Regressionsmodell für eine Fehlerklasse erstellt wird. Daher muss, bevor das Modell erstellt werden kann, der Datensatz gefiltert werden. Dazu wurde eine Funktion geschrieben, welche die Daten nach einer Fehlerklasse filtert und ihnen den Zustand 1 oder 0 gibt. Auch müssen Spalten ausgewählt werden, die für die Regression wichtig sind. Hierbei wurden nur die Spalten gewählt, welche bekannt sind. Diese werden zusammen in eine Feature-Spalte transformiert. Dies ermöglicht für Spark eine schnellere Verarbeitungszeit. Im letzten Schritt müssen die Daten noch in Trainingsdaten und Testdaten aufgeteilt werden. Hierbei wurden mehrere Versuche durchgeführt. Die logistische Regression wurde mit mehreren Fehlerklassen angewendet, z.B. wurden alle Fehlerklassen, nur bestimmte Fehlercodes oder nur bestimmte Fehlerklassen miteinbezogen. Das Ergebnis fiel jedoch immer ähnlich aus.

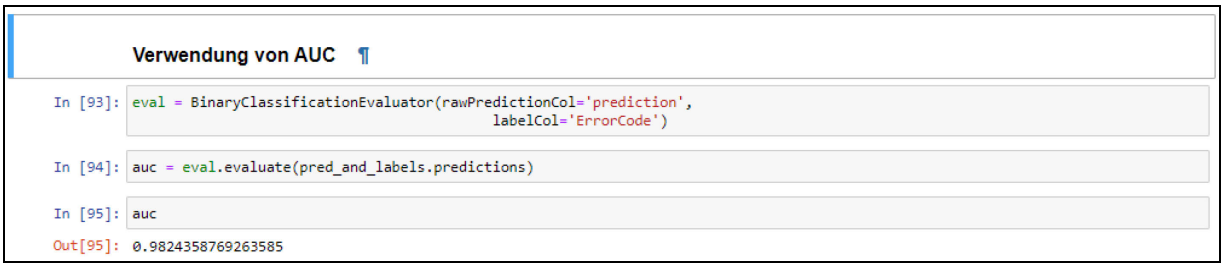


Abbildung 64: Anwendung AUC-Score (eigene Darstellung)

Die Anwendung des AUC-Score wird, wie in der Abbildung 64 dargestellt, implementiert. Bei diesem Versuch mit allen Maschinen und Fehlerklasse Hochdruck ergab das AUC einen Score von 0,98. AUC (Area under Curve) ist dabei das Maß für die Fähigkeit eines Klassifikators, zwischen Klassen zu unterscheiden, und wird als Zusammenfassung der ROC(Receiver-Operating-Characteristics)-Kurve verwendet. Je höher der AUC-Score, desto besser ist die Leistung des Modells bei der Unterscheidung zwischen den positiven und negativen Klassen.

Bei jedem Versuch gibt es eine sehr hohe Wahrscheinlichkeit, dass diese Klasse richtig vorausgesagt wird. Dies ergibt sich daraus, dass es sehr viele Einträge gibt, in denen keine Fehler auftauchen. Selbst wenn das Modell immer keinen Fehler voraussagen würde, gäbe es trotzdem eine hohe Trefferwahrscheinlichkeit, dass das Modell die richtige Vorhersage trifft. Im Datensatz selbst tauchen Fehler nur sehr selten auf. Wie in Abschnitt zur Datenexploration gesehen, gibt es in 165900 Einträgen insgesamt nur 334 Fehler. Daher ist dieses Ergebnis zunächst nicht aussagekräftig für die Erkennung von Fehlern. Es wird dadurch bestätigt, dass mit diesen teilbereinigten Datensatz keine guten Ergebnisse erzielt werden kann. Alle anderen überwachten Machine-Learning-Methoden erzielten ähnliche Ergebnisse. Daher werden diese Ergebnisse zunächst ignoriert. Im nächsten Schritt muss nun der Datensatz so angepasst werden, dass diese Methoden angewendet werden können. Dies ist eine Voraussetzung für das Predictive-Maintenance-Modell.

5.3.4.2 Versuchsreihe 2: Anwendung von ML mit manipulierten Daten

Da die Ergebnisse in der Versuchsreihe eher bedingt aussagekräftig sind, muss ein Weg gefunden werden, der den Datensatz so zu manipulieren versucht, dass Machine-Learning-Methoden besser funktionieren. Es existieren mehrere Methoden zum Resampling. Grundsätzlich wird aus dem gesamten Datensatz eine ausgewählte Stichprobe entnommen, die mehr Aussagekraft bei Analysen aufweist. In diesem Kapitel wird versucht, diejenigen Resampling-Methoden anzuwenden, welche die Machine-Learning-Ergebnisse verbessern sollen. Da später eine Ereigniszeitanalyse bzw. Überlebenszeitanalyse durchgeführt werden

soll, bietet sich hier kein Oversampling an. Beim Oversampling werden Einträge der schwachen Klasse dupliziert oder durch andere statistische Methoden randomisiert erstellt. Auch ist es schwierig, einen Zeitstempel zu randomisieren. Gerade diese Zeitdaten sind jedoch essenziell für Ereigniszeitanalysen. Es bietet sich daher nicht an, diese Daten zu manipulieren bzw. zu duplizieren. Beim Duplizieren der Daten würde ein Fehler zu demselben Zeitpunkt passieren. Daher wurde das Undersampling gewählt. Beim Undersampling wird der Datensatz so manipuliert, dass alle Klassen gleich groß sind. Dort wird die dominierende Klasse so manipuliert, dass sie die gleiche Anzahl an Einträgen hat wie die schwache Klasse. Die Einträge wurden hierbei zufällig ausgewählt. Dazu wurde eine Undersampling-Funktion geschrieben, welche den Datensatz beliebig nach Maschinennummer und Fehlerklasse reduzieren kann. Diese Funktion basiert hier ebenfalls auf einer Funktion von Spark. Demnach wird ein neuer Data-Frame aus dem veränderten Datensatz erstellt. Nun wird versucht, mit diesen verschiedenen Datensätzen Machine-Learning-Methoden anzuwenden. In der ersten Versuchsreihe gab es keine guten Ergebnisse für die logistische Regression sowie für die Entscheidungsbaummethode. Durch den manipulierten Datensatz werden aussagekräftigere Ergebnisse erwartet. Dennoch dienen diese Ergebnisse nur als Indikatoren für spätere Teillösungen, da beim Undersampling die Daten randomisiert ausgewählt werden. Somit entstehen beim Undersampling jeweils immer andere Ergebnisse, diese sind jedoch in dieser Versuchsreihe immer ähnlich ausgefallen. Der gesamte Datensatz enthält 165900 Einträge, wovon 334 Fehler sind. Das ergibt eine Relation von 1:497. Nach der Manipulation reduzierte sich der Datensatz auf 668 Einträge, mit einer Relation von 1:1 Fehler- zu Nichtfehler-einträgen.

Als Erstes wurden Korrelationsmatrizen mit manipulierten Datensätzen und unberührten Datensätzen erstellt und dahingehend miteinander verglichen, wie die Korrelation aussehen (siehe Abbildung 65).

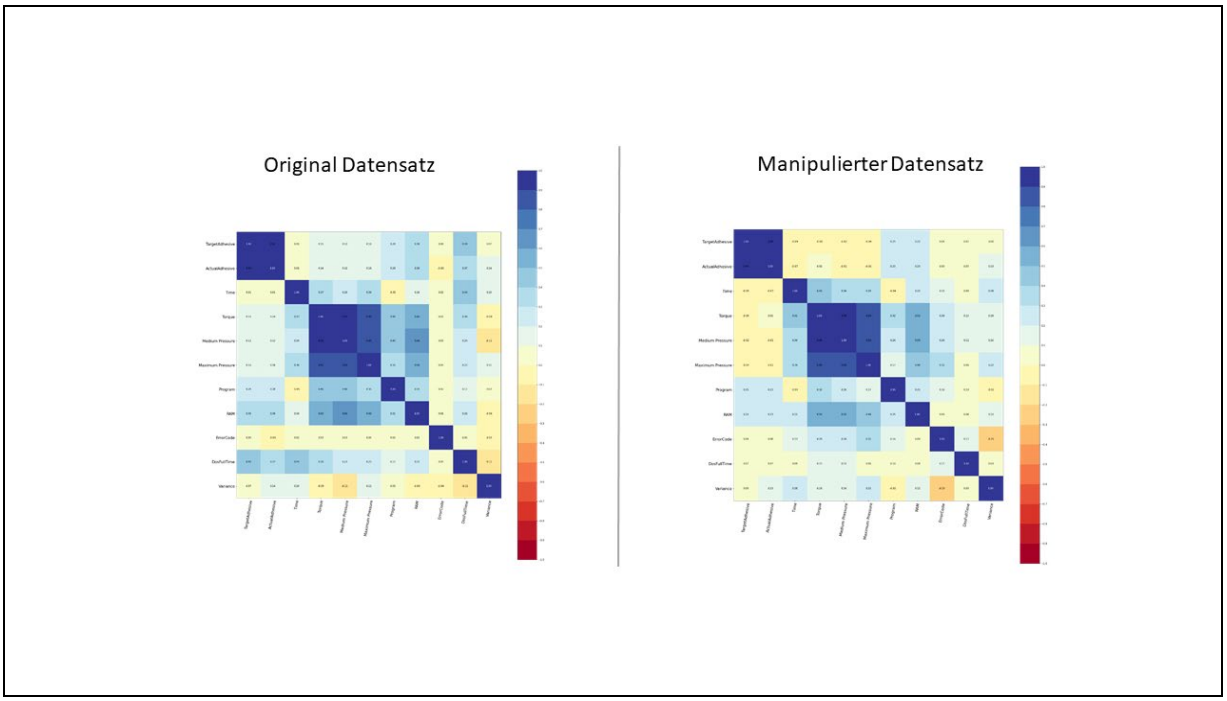


Abbildung 65: Vergleich Korrelationsmatrizen (eigene Darstellung)

Beide Korrelationsmatrizen zeigen ähnliche Ergebnisse auf. Auch gibt es keine starken Korrelationen zwischen den Parametern. Lediglich Klebemenge Soll und Klebemenge Ist zeigt eine starke Korrelation. Das ist ein Hinweis darauf, dass die Daten trotz der Manipulation in der Datenstruktur gleich geblieben sind. Im nächsten Schritt wird mit allen Maschinen und allen Fehlerklassen eine logistische Regression durchgeführt. Für diese wurden zunächst alle Fehler zusammengefasst, um herauszufinden, wie diese erkannt werden können. In Spark müssen für Machine-Learning-Methoden ausgewählte Spalten in einer Feature-Spalte zusammengefasst werden. Hierbei wurden wie in der ersten Versuchsreihe nur die bekannten Spalten ausgewählt, welche auch für die Korrelationsmatrix ausgewählt wurden. Danach muss der Datensatz in Trainingsdaten und Testdaten aufgeteilt werden. In dieser Versuchsreihe wurde ein 0.7/0.3-Split ausgewählt. Bei diesem Datensatz konnten die Fehlerklassen zu 71 % richtig erkannt werden, was für diesen Datensatz und die Korrelationen für ein gutes Ergebnis spricht.

Im nächsten Schritt wurde eine Entscheidungsbaummethode angewendet, um zu sehen, welche Variablen den Fehlercode am meisten beeinflussen. Dazu wurde der Decision Tree Classifier von MLlib genutzt. Wie auch schon in der logistischen Regression wird die gleiche Feature-Spalte genutzt. Dabei erkennt die Entscheidungsbaummethode für alle Variablen diese Abhängigkeiten. Wie auch schon in der Korrelationsmatrix steht 1 für eine starke und 0 für eine schwache Abhängigkeit. Die Ergebnisse zeigen leichte Abhängigkeiten beim maximalen Druck und bei der Dosierungszeit auf (siehe Tabelle 14). Diese Ergebnisse sind zunächst jedoch nur

kleine Teillösungen. Sie sollen das Predictive Maintenance später unterstützen. Diese Modelle können nun im Serving Layer abgespeichert und visualisiert werden. Diese Informationen sind demnach als ein erweitertes Monitoring-Tool anzusehen.

Soll Klebemenge	0.0924
Ist Klebemenge	0.1048
Zeit	0.0671
Drehzahl	0.0656
Mitteldruck	0.0055
Maximaler Druck	0.4091
Dosierungszeit	0.2555

Tabelle 14: Abhängigkeiten von Fehlern (eigene Darstellung)

5.3.4.3 Versuchsreihe 3: Echtzeitanalyse und Überlebenszeitanalyse

Nachdem die ersten Machine-Learning-Modelle entwickelt wurden, ist nun das Ziel die Erstellung eines Predictive-Maintenance-Modells mithilfe einer Ereigniszeitanalysen. Wichtig hierfür ist die Berechnung der Zeitabstände zwischen dem Zeitpunkt, zu dem die Maschine ohne Fehler funktioniert, bis zum Zeitpunkt des Fehlers. Hierfür wird die Zeitangabe genutzt, welche auch in der Datenaufbereitung herangezogen wurde. Hierbei gibt es zwei Spalten in der Tabelle: einmal eine Zeitspalte der normalen Zeitdifferenz zwischen zwei Einträgen sowie eine kumulierte Zeitangabe der Zeitdifferenz. Mit diesen Werten kann nun ermittelt werden, wie lange es gedauert hat, bis die Fehler aufgetreten sind. In der Abbildung 66 ist ein Auszug aus dem Datensatz der Maschine dargestellt.

```
In [10]: m200
```

Out[10]:

getAdhesive	ActualAdhesive	Time	Torque	Medium Pressure	Maximum Pressure	8	Program	RAM	...	30	31	32	33	34	35	Errcateg	machine_ip	time_diff	tsec
4700	4667	3246	303	142	195	0	25	2	...	193	201	63	15	10	42	normal	M200	00:00:00	0.000000
4700	4670	3246	305	142	197	0	25	2	...	196	202	63	15	10	39	normal	M200	00:01:47	0.001238
4700	4674	3246	305	142	197	0	25	2	...	196	201	63	15	10	40	normal	M200	00:02:31	0.002986
4700	4677	3246	305	141	197	0	25	2	...	196	202	63	15	10	39	normal	M200	00:02:31	0.004734
4700	4677	3246	298	140	196	0	25	2	...	195	202	63	15	10	40	normal	M200	00:02:42	0.006609

Abbildung 66: Auszug Datensatz Maschine (eigene Darstellung)

Um mit diesen Distanzen zu rechnen, wird nun in einem ersten Ansatz das Kaplan-Meier-Modell genutzt. Der Kaplan-Meier ist ein Schätzer und eignet sich als eine Methode, mit der Vorhersagemodelle geschätzt werden. Die Schätzung bedingter Wahrscheinlichkeiten zum Zeitpunkt eines auftretenden Events ist ein Bestandteil des Kaplan-Meier-Modells. Der andere Bestandteil basiert auf der Bildung des Produktgrenzwertes dieser Wahrscheinlichkeiten zur Schätzung der Überlebensrate zu jedem Zeitpunkt (vgl. Zwiener, Blettner & Hommel, 2011, S. 165). Das Kaplan-Meier-Modell arbeitet dabei mit diesen Zeitabständen. Alle anderen

Parameter haben auf das Modell keinen Einfluss. Für die Berechnung dieses Modell müssen die Daten auf eine Maschine und eine Fehlerklasse reduziert werden. Denn da alle Maschinen unterschiedlich arbeiten, ist es sinnvoll, diese Daten zu trennen. Ebenfalls ist es sinnvoll, die Modelle der Fehlerklasse zu reduzieren, damit auch genau die betroffene Klasse vorausgesagt werden kann. Würde der gesamte Datensatz genutzt werden, wüsste man nur, wann ein Fehler auftritt, aber nicht, welcher. Das Gleiche gilt für die Maschinen. Daher ist es sinnvoll, diese zu trennen.

Bevor wieder mit den manipulierten Daten gearbeitet wird, steht noch ein Versuch an, die Analyse mit dem gesamten Datensatz durchzuführen (siehe Abbildung 67).

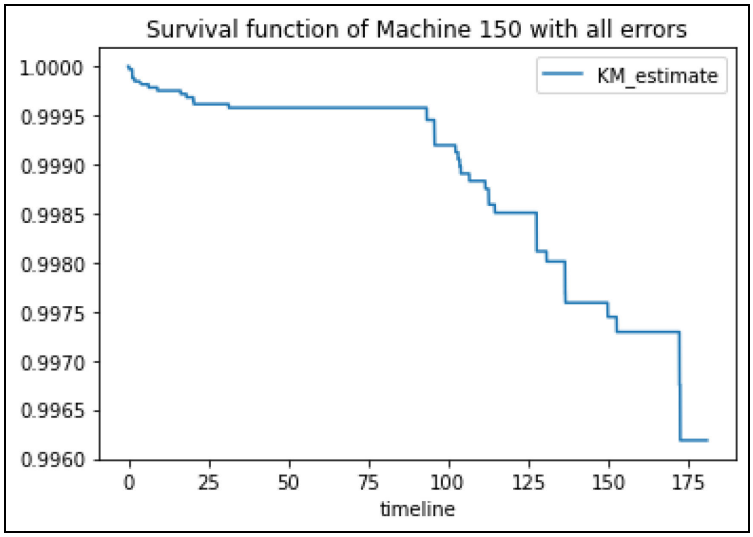


Abbildung 67: Überlebenszeitanalyse Maschine 150 mit allen Fehlern (eigene Darstellung)

Für diesen ersten Lösungsansatz wurde versucht, das Kaplan-Meier-Modell an Maschine 150 anzuwenden. Wie bereits im Grundlagenteil beschrieben, beziehen sich Ereigniszeitanalysen auf sogenannte Lebenslinien. An der Y-Achse befindet sich die Überlebenswahrscheinlichkeit. Auf der X-Achse befindet sich die Zeit in Tagen. Hier ist gut zu erkennen, dass dieser Graph keine große Aussagekraft hat, da die Wahrscheinlichkeit, dass die Maschine nicht ausfallen wird, hier zwischen 100 % und 99,6 % liegt,. Selbst nach 175 Tagen liegt die Wahrscheinlichkeit, dass die Maschine ausfallen wird, gerade mal bei 0,4 %. Das kann mit den vorhandenen Daten nicht bestätigt werden. Dieser Wert kommt vielmehr durch ein Ungleichgewicht der Klassen zustande. Speziell bei diesem Modell ist es wichtig, dass eine Ausgeglichenheit vorhanden ist, denn hier wird mit Überlebenszeiten gearbeitet. Daher muss jeder Fehler einzeln betrachtet werden. Bei dieser Funktion ist es wichtig, die Dauer zu bestimmen. Dazu sind immer ein Startpunkt und ein Endpunkt notwendig. Bei dem nichtmanipulierten Datensatz gibt es zu viele Startpunkte, aber keine Endpunkte. Im Datensatz

selbst steht jeder Eintrag, in dem kein Fehler auftrat, für den Anfang eines Lebenszyklus einer Maschine. Da es im nichtmanipulierten Datensatz weitaus mehr Einträge gibt, welche keine Fehler haben, erkennt der Algorithmus mehr Lebenszyklen, die noch leben, als beendete Lebenszyklen. Somit muss der Datensatz manipuliert werden, sodass die Fehlerklasse und die Nichtfehlerklasse gleich groß sind. Weitere Versuche mit anderen Maschinen und Datensätzen ergeben ähnliche Ergebnisse.

Im weiteren Schritt wird nun die Funktion aus dem vorherigen Abschnitt angewendet, um die Daten nach Maschine und Fehlerklassen zu manipulieren. Weithin werden alle Datensätze ignoriert bzw. erstmal ausgelassen, die weniger als zehn Fehler beinhalten. Ergebnisse dieser Untersuchungen haben gezeigt, dass sobald die Anzahl der Fehlerklassen unter 10 fällt, keine sinnvollen Ergebnisse generiert werden. Da es nicht genug Datensätze gibt, um diese in Trainingsdaten und Testdaten aufzuteilen, werden diese Modelle zunächst nicht beachtet. Diese Modelle können aber für zukünftige Anwendungen vorbereitet werden. Damit diese Modellerstellung automatisch erfolgen kann, wurde hier auch eine Funktion geschrieben, welche nur Modelle mit über zehn Fehlern in Betracht zieht. Sobald die Daten aktualisiert werden und die Fehleranzahl der jeweiligen Klassen ansteigt, werden diese Modelle einbezogen.

	Alle Maschinen	**M150**	**M195**	**M200**	**M214**	**M226**
Alle Fehlerklassen	334	41	68	142	33	50
Temperatur	49	7	**18**	**13**	1	**10**
Hochdruckfehler	128	3	**18**	**91**	**16**	0
Dosierfehler	40	7	6	9	3	**15**
Referenzschalter	0	0	0	0	0	0
Pistolenmotor	47	0	0	**24**	0	**23**
Dosiermotor	0	0	0	0	0	0
Fatale Motorfehler	5	0	5	0	0	0
Behälterfehler	5	1	4	0	0	0
Füllhahndosierung	2	1	1	0	0	0
Proportionalventil	0	0	0	0	0	0
Pistole	7	6	1	0	0	0
Spannungsüberwachung	0	0	0	0	0	0
Düsenabstand	0	0	0	0	0	0
Mengenfehler	48	**16**	**14**	3	**13**	2
Netzwerkfehler	2	0	1	1	0	0
Fasspresse	1	0	0	1	0	0
Klebstoff/Material	0	0	0	0	0	0

Tabelle 15: Fehlertabelle (eigene Darstellung)

In der Tabelle 15 wurden alle Fehler mit einer Anzahl über 10 markiert. Somit werden für diese Versuchsreihe zwölf Vorhersagemodelle erstellt, von denen alle mit der Kaplan-Meier-Methode erfolgreich aufgebaut werden konnten. Diese geben erste Informationen darüber, wie lange die Überlebenszeiten der jeweiligen Maschinen sind und mit welcher Wahrscheinlichkeit diese auftreten. Dabei zeigen diese einen ersten Ansatz eines Predictive Maintenance an (siehe Abbildung 68).

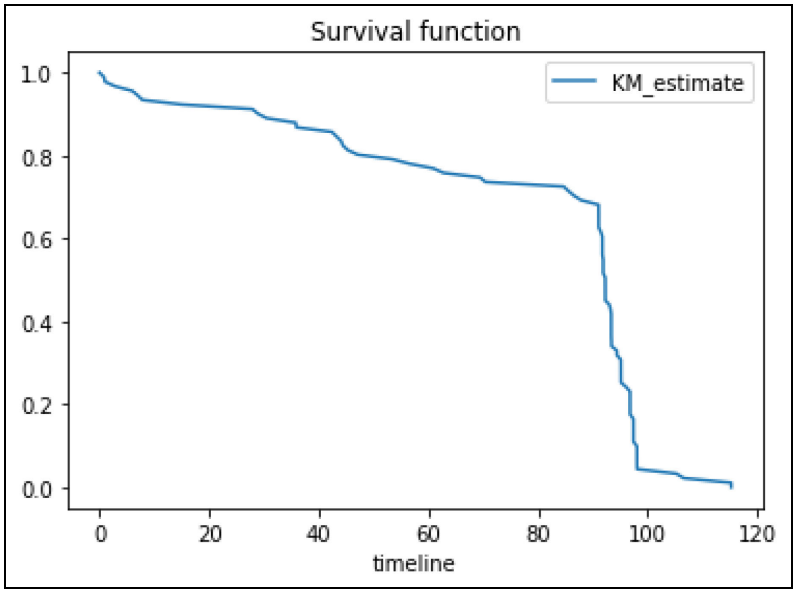


Abbildung 68: Überlebenszeit Maschine 200 (eigene Darstellung)

An dieser Abbildung wird deutlich, dass es zwischen Tag 90 und Tag 100 eine hohe Ausfallsquote mit der Fehlerklasse Druckfehler gibt. Im Regelfall sollte bei einer Ausfallwahrscheinlichkeit von 0,5 die Maschine auf dieser Fehlerklasse gewartet werden. Es besteht auch die Möglichkeit, die Produktion an diese Ausfallraten anzupassen. Speziell an diesem Beispiel ist die Empfehlung, die Maschine ab Tag 80 zu warten. Anschließend wird versucht, mit dem Cox-Proportional-Hazard-Modell zu arbeiten.

Mit der Cox-Regression wird ein Vorhersagemodell mithilfe von Daten erstellt, welches die Zeit bis zum Eintreten eines Ereignisses widergibt (vgl. Zwiener et al., 2011, S. 166). Hier wird eine Überlebensfunktion erstellt. Dabei wird die Wahrscheinlichkeit des Ereignisses bei vorgegebenen Daten zu einer gegebenen Zeit t berechnet. Sowohl die Überlebensfunktion als auch die Regressionskoeffizienten werden geschätzt. Das Vorhersagemodell kann immer angewendet werden, sobald Prädiktorvariablen durch Messungen vorhanden sind.

Auch hier wurde versucht, zwölf Modelle zu erstellen, dies ist jedoch nicht gelungen. Der Grund ist wahrscheinlich ein zu geringer Datensatz, da MLlib dort immer eine Fehlermeldung ausgegeben hat, welche auf eine zu geringe Datenmenge hinwies. Im nächsten Kapitel wurde

wieder mit dem Cox-Proportional-Hazard-Modell nach KATZMANN gearbeitet, jedoch diesmal mit einem anderem Framework als MLlib. In diesem Kapitel lässt sich indes festhalten, dass bei einer geringen Datenmenge das Kaplan-Meier-Modell zu bevorzugen ist.

5.3.4.4 Versuchsreihe 4: Kombination von Echtzeitanalyse und ML mithilfe von Deep Learning

In dieser Versuchsreihe wird nicht mit der Machine-Learning-Bibliothek von MLlib gearbeitet. Jared L. KATZMANN hat hierfür einen neuartigen Ansatz für das Cox-Proportional-Hazard-Modell entwickelt. Dazu gibt es eine freie Bibliothek namens PySurvival (PySurvival, 2020). Dort kombiniert KATZMANN Ereigniszeitanalysen mit Deep-Learning-Methoden. Wie auch im Kaplan-Meier-Modell müssen bei diesem Modell die manipulierten Daten verwendet werden. Dazu wird wieder die Resampling-Funktion genutzt. Ähnlich wie bei MLlib von Spark müssen bei dieser Methode auch Feature-Spalten genutzt werden. Aber anders als bei der Klassifikation werden bei dieser Methode alle Parameter genutzt. Durch das Deep Learning besteht die Hoffnung, dass das Modell von selbst erkennt, welche Parameter wichtig für das Ergebnis sind und welche nicht. Auch bei dieser Methode müssen die Dimensionen reduziert werden. Um dabei die optimale Dimensionsmenge zu ermitteln, wird hierbei eine Hauptkomponentenanalyse durchgeführt. Die Hauptkomponentenanalyse ist ein Verfahren der multivariaten Statistik. Um nun die optimale Komponentenmenge zu ermitteln, wird nach Korrelationen zwischen den Variablen gesucht. Das Maß für den Informationsgehalt ist der Anteil an der totalen Varianz zwischen den Variablen. Diese Varianzen wurden kumuliert und wie folgt in einem Plot dargestellt (siehe Abbildung 69).

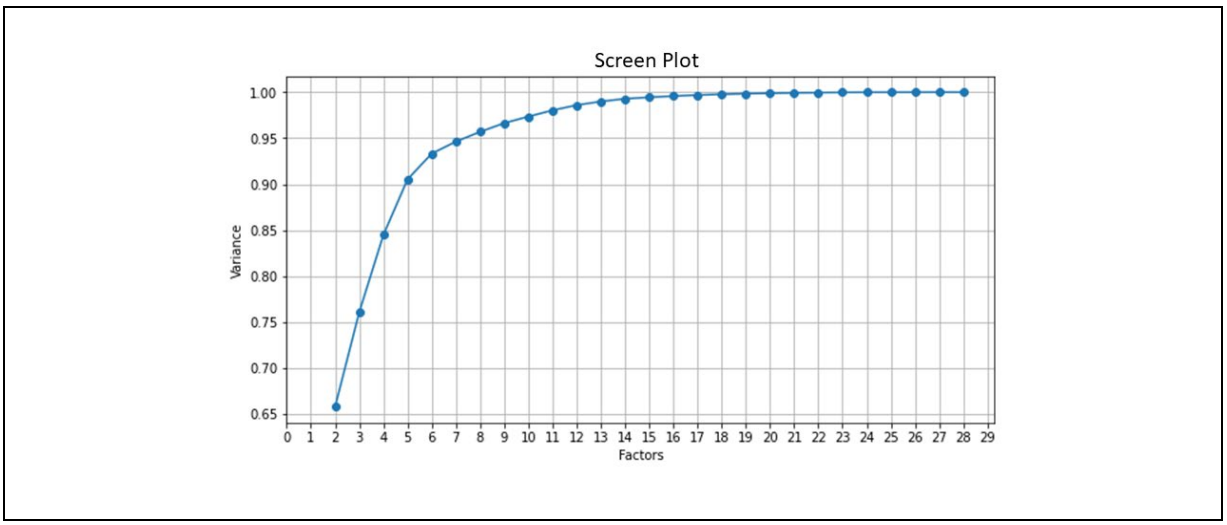


Abbildung 69: Hauptkomponentenanalyse (eigene Darstellung)

Oft wird die Dimension des Ausgangsproblems (= Anzahl Merkmale in X) auf diejenige Anzahl an HKs reduziert, die mindestens 95 % der Gesamtvariation repräsentieren. Somit wurde hier eine Hauptkomponentenanzahl von 7 ausgewählt. Auch hier wird der Datensatz in 0.7 Trainingsdaten und 0.3 Testdaten aufgeteilt.

Dennoch reichen die Daten aus der dritten Versuchsreihe nicht aus, um diese Methode anzuwenden. Durch die kleine Datenmenge treten auch bei dieser Methode Fehlermeldungen auf, dass die Datenmenge zu gering sei. Lediglich bei der Maschine 200 mit der Fehlerklasse Hochdruckfehler konnte diese Methode erfolgreich angewendet werden. Die Fehleranzahl beträgt hier 91.

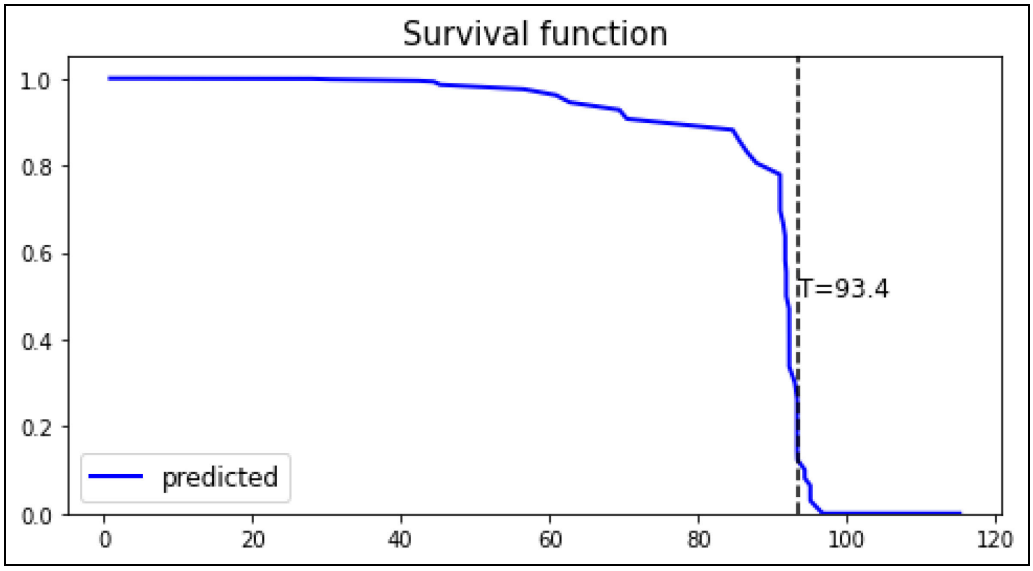


Abbildung 70: CoxPH nach Katzman (eigene Darstellung)

Aus der Abbildung 70 ist deutlich zu erkennen, dass eine Ähnlichkeit zum Kaplan-Meier-Modell vorliegt. Für die Bewertung dieses Modells wird eine Kreuzvalidierung mittels des Konkordanzindexes durchgeführt. Der Konkordanzindex (C-Index) ist ein „globaler“ Index zur Validierung der Prognosefähigkeit eines Überlebensmodells. Er entspricht einer Rangkorrelation. Wenn der C-Index nahe bei 1 liegt, hat das Modell im Allgemeinen ein nahezu perfektes Ergebnis, wenn er jedoch nahe bei 0,5 liegt, ist er nicht in der Lage, zwischen Ereignissen mit niedrigem und hohem Risiko zu unterscheiden. Da der C-Index (hier 0,69) relativ nahe bei 1 liegt, kann davon ausgegangen werden, dass das Modell einigermaßen gute Ergebnisse liefert.

Das theoretische Modell wird nun nach der Anpassung darauf hinterfragt, inwieweit die Realität abgebildet werden kann. Die Schwäche des Modells sind Verzerrungen, die durch falsche Annahmen, fehlerhafte Daten oder durch die Nichtberücksichtigung von Störgrößen entstehen.

Als Validierungswerkzeug für Überlebenszeitmodelle dient der Konkordanzindex C (vgl. Harrell et al., 1982). Dabei stellt der C-Index die relative Häufigkeit von konkordanten Datenpaaren unter allen möglichen Paaren für unzensierte Daten dar. Als konkordant wird das Paar bezeichnet, sobald das Individuum mit der geringeren Überlebenszeit ebenfalls ein höheres Risiko für ein Geschehen innehat (vgl. Gerds et al., 2013). Ist die prognostizierte Überlebenszeit der zwei ausgewählten Individuen nach Berechnung des Modells fast identisch, geht das Resultat zur Hälfte in die Prognose mit ein. Sind die realen Überlebenszeiten hingegen gleich, sind sie nicht für das Modell relevant. Der C-Index gibt Werte zwischen 0 und 1 wieder. Je näher der Wert gegen 1 tendiert, desto genauer ist das Vorhersagemodell. Bei einem Wert von 0,5 bedeutet dies hingegen eine ungenaue Vorhersage (vgl. Harrell et al., 1996).

Zur Berechnung der Güte eines Modells wird neben den verwendeten Daten ein weiterer Datensatz benötigt. Dieser ist unabhängig von den bisher genutzten Daten auszuwählen. Damit wird die Prognosefähigkeit des Modells getestet.

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

(In dieser Fallstudie wurden die Daten daher in Trainingsdaten und Testdaten aufgeteilt.)

Eine weitere Validierungsmethode ist der Brier-Score. Dieser misst die durchschnittlichen Abweichungen zwischen dem eigentlichen Wert und dem geschätzten Wert zu einem bestimmten Zeitpunkt. Je niedriger der Score (in der Regel unter 0,25), desto besser ist also die Vorhersageleistung. Der Brier-Score ist ein skalares Maß zur Verifikation probabilistischer Vorhersagen anhand der mittleren quadratischen Abweichung einer Wahrscheinlichkeitsvorhersage (vgl. Wilks, 2011):

$$BS = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2$$

Bei der Beobachtung o_k handelt es sich um eine dichotome Variable, d.h. die Beobachtung ist $o_k = 1$, wenn das Ereignis eintritt, und $o_k = 0$, wenn das Ereignis nicht eintritt. Die Vorhersage y_k kann alle Werte zwischen 0 und 1 annehmen. Der Brier-Score mittelt die quadratischen Differenzen zwischen n Paaren von Beobachtung und Vorhersage. Tritt das Ereignis ein, sollte die Wahrscheinlichkeit möglichst groß und nahe bei 1 sein. Umgekehrt sollte diese annähernd

0 sein, wenn das Ereignis nicht eintritt und daher also $o_k = 0$. Deshalb gilt für eine perfekte Vorhersage BS = 0. Da der Wertebereich von Beobachtung und Vorhersage auf das Intervall zwischen 0 und 1 beschränkt ist, liegt der Brier-Score auch bei weniger präzisen Vorhersagen immer zwischen 0 und 1 (vgl. Wilks, 2011).

Der Brier-Score liegt hier auf der gesamten Modellzeitachse sehr nahe bei 0,0. Dies deutet auf eine gute Vorhersagefähigkeit hin (siehe Abbildung 71).

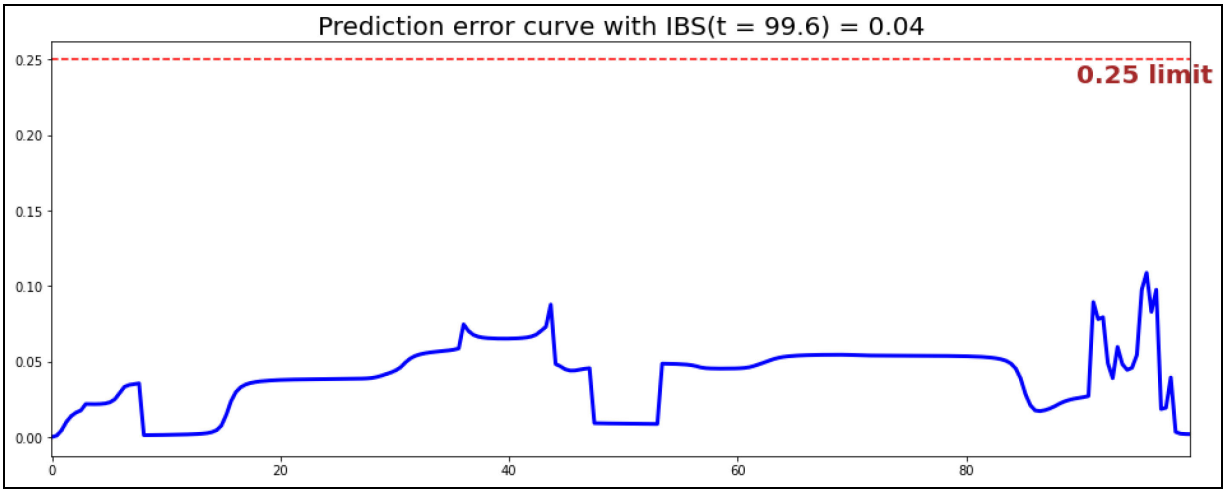


Abbildung 71: Brier-Score Fehlerkurve (eigene Darstellung)

5.3.4.5 Speicherung und Visualisierung

Nach der Fertigstellung der Modelle und der Analysen sollten die Ergebnisse im Serving Layer abgespeichert werden. Bevor aber dieser Schritt ausgeführt wird, sollten fertige funktionierende Modelle zurück ins HDFS gespeichert werden. Spark kann mit einem einfachen Save-Befehl fertige Modelle abspeichern. Da aber laufend Daten produziert werden, wird das Modell auch laufend weiter trainiert. Die Abspeicherung des Modells dient hier mehr als eine Sicherheitskopie. So könnten alte eventuell besser funktionierende Modelle genutzt werden.

In dieser Fallstudie wird mit den Datenbanken Hbase und Cassandra gearbeitet. Beide Datenbanken bieten hierbei verschiedene Vorteile. Gerade Cassandra bringt hier viele Vorteile für die Zusammenarbeit mit Spark-Streaming mit sich. Da es sich bei Streaming-Daten meistens um unstrukturierte Daten handelt, eignet sich am besten die verteilte NoSQL-Datenbank Cassandra. Diese bietet eine hohe Fehlertoleranz, hohe Verfügbarkeit sowie hohe Schreib- und Lesegeschwindigkeiten. Für diese Anwendung gibt es für Spark einen Cassandra-Connector, der eine Daten-Pipeline zwischen diesen beiden Tools erstellt. Für die Ergebnisse des Predictive Maintenance und der Analyse war die Datenbank HBase dennoch ausreichend.

Ein Ziel ist es, aus den generieten Daten dynamische Dashboards zu erstellen, die sich mit wandelnden Daten ständig ändern. Dort bietet HBase Vorteile mit Fokus auf schnellen und konsistenten Lesevorgängen, da nur auf einem Server geschrieben wird und Datenversionen nicht mit verschiedenen Knoten verglichen werden müssen. Ein HBase-Server hat auch nicht zu viele Datenstrukturen, die die Datenbank überprüfen muss. Die Ergebnisse der Fallstudie sind sehr einfach gehalten und bieten eine einfache Datenstruktur, daher sollten diese Ergebnisse auch im HBase gespeichert werden. Auch hier existiert ein HBase-Connector. Zusätzlich wurde ein spezieller Connector von Hortonworks genutzt. Nach der Installation des Connectors und Einrichten der Config.-Datei können nun fertige Data-Frames aus den Analysen in HBase abgespeichert werden. Die Ergebnisse der Kaplan-Meier-Methode und des CoxPH-Modells sollten in Data-Frames oder Data Tables gespeichert werden, damit diese in HBase abgespeichert werden können. Sind die Daten erstmal auf HBase, können verschiedene Tools für Visualisierungen genutzt werden, u.a. Tools wie Google Data Studio, Tableau oder Qlik. HBase ist eine weitverbreitete Datenbank, sodass es viele Connectoren zu den genannten Visualisierungstools gibt.

Das Beispiel Dashboard in der Abbildung 72 dient dafür als Veranschaulichung.

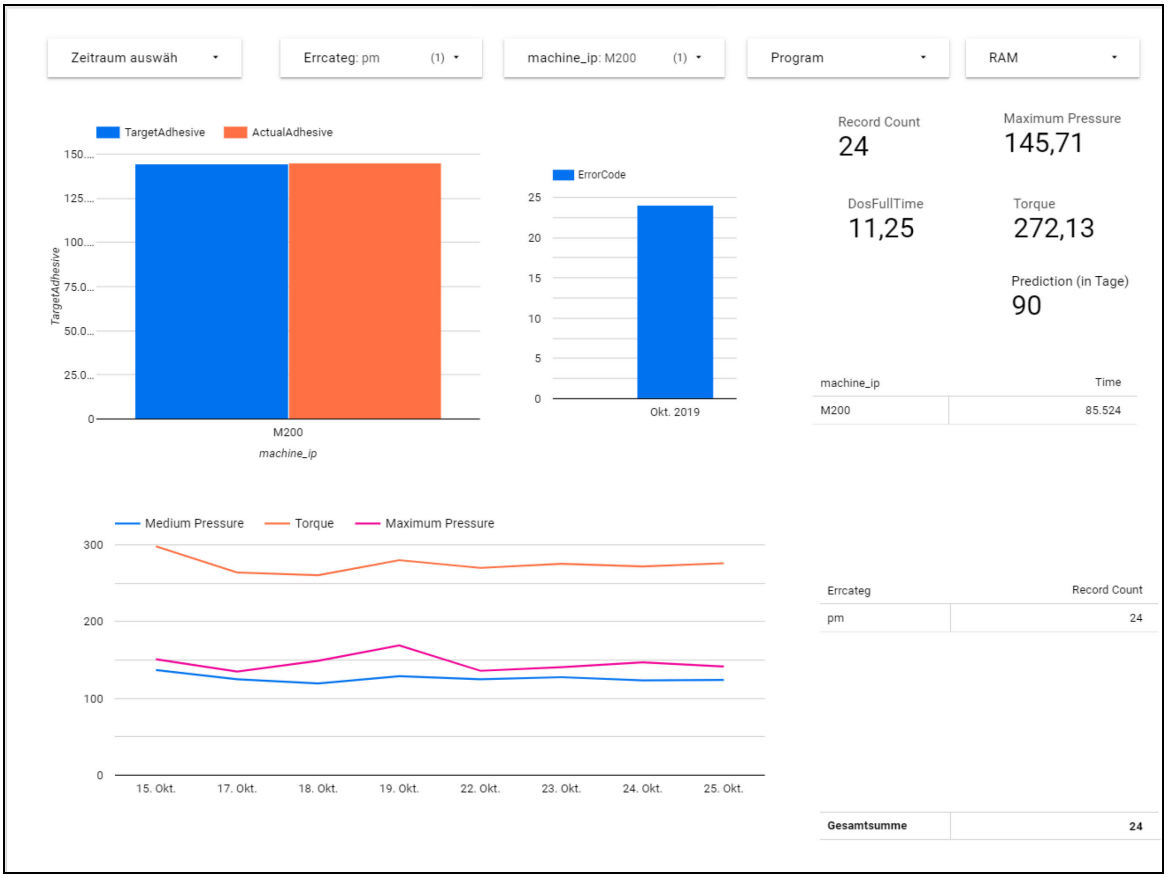


Abbildung 72: Beispielvisualisierung mit Google Data Studio (eigene Darstellung)

5.3.5 Zwischenfazit

Die Herausforderung dieser Versuchsreihe bestand darin, Ergebnisse mit dem gegebenen Datensatz zu erreichen. Mithilfe von Datenmanipulation mittels Resampling und klassischer Datenbereinigung konnten einige Vorhersagemodelle erfolgreich erstellt werden. Durch die gegebene Big-Data-Architektur ist eine Automatisierung aller Prozesse ebenfalls möglich. Dadurch verbessern sich alle Vorhersagemodelle automatisch, sobald neue Daten von den Anlagen generiert werden. Arbeitet man nur mit den Zeitangaben, ist es sogar möglich, mit einem geringen Datensatz zu arbeiten. So konnten mit dem Kaplan-Meier-Modell zwölf Modelle erstellt werden, wenn eine Fehleranzahl von über 10 vorlag. Ein Ergebnis des Kaplan-Meier-Modells konnte sogar mit dem Cox-Proportional-Hazard-Modell bestätigt werden, da dort die nahezu gleichen Ergebnisse erzielt wurden. Auch konnte das Cox-Proportional-Hazard-Modell durch die Validierungsmethoden Konkordanzindex und Brier-Score bestätigt werden. Auch wenn bei dieser Versuchsreihe nur ein Teilerfolg erzielt wurde, empfiehlt es sich, immer mit einer großen Datenmenge zu arbeiten anstatt mit einer kleinen. Ergebnisse dieser Versuchsreihe haben gezeigt, dass immer bessere Ergebnisse erzielt werden, wenn ein großer Datensatz vorhanden ist. Insgesamt lohnt es sich dennoch, in Predictive-Maintenance-Technologien zu investieren, da es, wie diese Fallstudie zeigt, durchaus möglich ist, auch mit einer geringen Datenmenge zu arbeiten. Die Überführung einer vorbeugenden Wartung in eine vorrausschauende Wartung sollte aber mit diesen Modellen noch nicht vorgenommen werden. Es muss bedacht werden, dass alle erstellten Modelle und Analysen auf Basis der vorbeugenden Wartung entstanden sind. Alle Daten, die genutzt worden sind, unterliegen täglichen Wartungsprozessen. Ohne diese Wartungsprozesse würden alle Modelle mit hoher Wahrscheinlichkeit andere Ergebnisse ergeben. Daher werden diese Ergebnisse bzw. Vorrausagen eher als erweitertes Monitoring zur Unterstützung der vorbeugenden Wartung angesehen. Auch werden alle Fehler bei dem Modell als unerwartete bzw. unvermeidbare Fehler angesehen. Das Modell dient nur der Voraussage des Zeitpunktes des Fehlers, sodass sich die Mitarbeiter darauf vorbereiten können. Alle erstellten Modelle sind skalierbar und können problemlos auch mit großen Datenmengen arbeiten. Das ist hauptsächlich möglich, da mit einer Big-Data-Architektur gearbeitet wurde.

5.4 Anwendungsempfehlung des Modells

Nachdem in den Abschnitten zuvor ein Predictive-Maintenance-Modell erstellt wurde, ist nun erstens herauszukristallisieren, wie die Interpretation der Visualisierungen zu deuten ist, und zweitens, welche Strategien in der Instandhaltung dadurch verwirklicht werden können. Die Interpretation des Modells läuft wie folgt ab: Nachdem die Analyse erste Ergebnisse hervorbringt, ist die Interpretation der Überlebenszeitgraphik essenziell. Die „Survival Function“, wie sie in der Abbildung 63 zu sehen ist, muss so interpretiert werden, dass der Zeitpunkt 0 das Ende des letzten Fehlers bedeutet. Somit ist Zeitpunkt 0 nicht dem Tag der Analyse gleichzusetzen. Daraufhin wird empfohlen, die Analyse täglich zu betreiben, sodass keine Fehlinterpretation entsteht. Des Weiteren ist die Betrachtung der Tage entscheidend. In dem verwendeten Algorithmus wurden alle produzierenden Tage berechnet. Damit sagt das Modell die nächsten Produktionstage voraus, an denen die Anlage oder die Maschine aktiv ist. Für eine geeignete Strategiefindung betrachten wir zunächst die Übersicht in Abbildung 73. Das Instandhaltungsmanagement wird unterteilt in das zentrale und dezentrale Management. Im betrachteten Werk wird die Instandhaltung zentral gesteuert. Durch das SCADA-System und die sich an den Anlagen befindenden SPS-Systeme können die Daten erhoben und somit nur eine reine Beobachtung der Sensoren vorgenommen werden. Erst eine Verknüpfung und somit die Implementierung der Produktionsprozesse mit der Big-Data-Architektur erlaubt eine vollautomatisierte Datenanalyse. Daraus können Instandhaltungsstrategien erarbeitet werden. Wird stattdessen eine Teilimplementierung vorgenommen, kann nur eine dezentrale Instandhaltungsstrategie vollzogen werden. Das bedeutet, dass das Predictive-Maintenance-Konzept mit einer Anlage/Maschine gekoppelt und daraus lokale Datenanalyse betrieben wird.

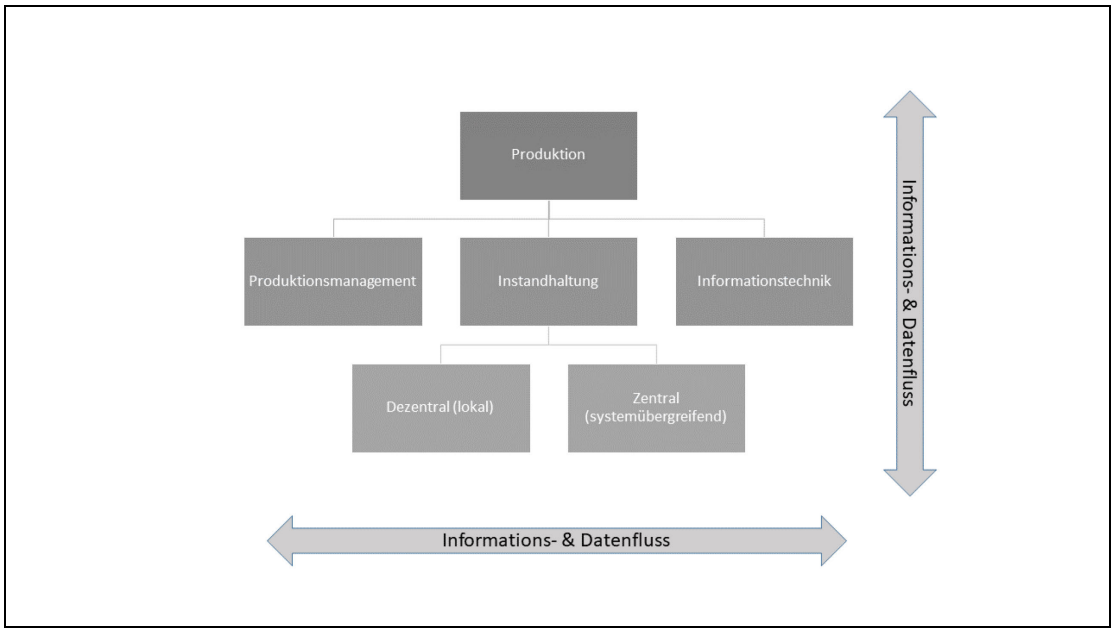


Abbildung 73: Konzept des Predictive Maintenance (eigene Darstellung)

Die Anwendung der Überlebenszeitanalyse hat zwei mögliche Ausprägungen der Durchführung der Instandhaltung zur Folge: beobachtende, vorbereitende Wartung und das Anpassen der Wartungs- und Instandhaltungsstrategie. Letzteres setzt eine hohe Datenquantität voraus, um eine sehr hohe Genauigkeit der Vorhersage zu erzielen. Strategisch betrachtet ist zunächst die beobachtende Wartung und Instandhaltung zu empfehlen. Aus dieser Sicht wird eine Fehlervorhersage beobachtet und mögliche Szenarien werden vorbereitet. Das hat den Vorteil, dass eine bevorstehende Instandhaltung schnellstmöglich durchgeführt werden kann, je nach Ausmaß des vorhergesehenen Fehlers. Nach mehrmaliger Bestätigung des Ergebnisses des Modells ist der weitere Schritt denkbar. In diesem kann der Wartungs- und Instandhaltungsplan angepasst werden. Die Wartungs- und Instandhaltungsprozesse an einer Anlage oder einer Maschine können dann nach ihrer Priorität den vorhersehbaren Fehlern zugeordnet werden. Auch hier werden wiederum mit der Wartung- und Instandhaltungsplanung Ressourcen optimiert. Zu beachten ist bei einem Eingriff vor Fehlerauslösung, dass sich das Ergebnis des Predictive Maintenance ändert. Daher ist zunächst die beobachtende und vorbereitende Methode sinnvoll.

6. Daten als Innovationstreiber der Automobilindustrie in der Digitalisierung

Im letzten Kapitel wird nun die Bedeutung durch die Einführung von Big-Data-Technologien adressiert. Dabei soll die aktuelle wirtschaftliche Situation der Automobilindustrie durchleuchtet werden. Die sich aus der Big-Data-Architektur ergebenden Handlungsfelder bedingen auch neue Möglichkeiten der Datenanalyse. Ein Mehrwehrt ist die Informationsgenerierung und die Informationsverarbeitung. Die Sichtweise auf Daten und deren Umgang verschafft der Automobilindustrie neue Erkenntnisse, die ihr zu einem Marktvorteil verhelfen sollen. Neben dem Verbesserungspotenzial durch Datenanalysen ist auch die Manifestierung der traditionellen Automobilproduzenten am Automobilmarkt von Bedeutung. Durch die rasante Entwicklung der IT-Branche und das aufsteigende Umweltbewusstsein ist die Elektromobilität gefragter denn je. Nun drängen junge Unternehmen wie Tesla Automotive oder auch Fisker Inc. hervor. Letztgenanntes Unternehmen hat mit einem neuen Vorstandsmitglied die Weichen im digitalen Zeitalter gestellt. Mit Bill McDermott unterstützt ein ehemaliger CEO von SAP den Automobilproduzenten. Damit ist eine klare Strategie erkennbar, dass im digitalen Zeitalter Daten und die damit einhergehenden Informationen zielführend für Innovationen sind. Innovationen werden zum einen gefördert durch Konkurrenz, zum anderen durch Technologie. Die Geschichte hat gezeigt, dass etablierte Unternehmen wie Nokia, Kodak oder auch IBM, die als Pioniere in ihrer Branche galten, durch entstandene disruptive Technologien Schwierigkeiten bekamen, auf dem Absatzmarkt zu performen. Einerseits können neue disruptive Produkte auf dem Markt diesem Effekt entgegenwirken. Andererseits können neue Erkenntnisse aus der Datenwissenschaft gewonnen werden, um Prozesse so innovativ zu gestalten, dass sie zu fortschrittlichen Produkten, Dienstleistungen oder auch Technologien führen.

Zunächst werden die Big-Data-Technologie und die daraus resultierende Auswirkung auf ein Unternehmen in der Automobilindustrie durchleuchtet. Darin wird Big Data selbst als disruptive Technologie betrachtet, um dem Datenverständnis und dem Datenmanagement eine neue Rolle im Unternehmen zukommen zu lassen. Der sich anschließende Abschnitt 6.2 befasst sich mit der Big-Data-Technologie und den daraus resultierenden datengetriebenen Geschäftsmodellen. Im letzten Abschnitt wird dann auf die Auswirkungen auf ein Unternehmen in der Automobilindustrie hinsichtlich ihrer Flexibilität gegenüber Markt- und

Umweltveränderungen eingegangen. Darunter fallen Aspekte wie bspw. Kundenbedürfnisse und umweltpolitische Einflüsse.

6.1 Disruptives Potenzial von Big Data

Was ist disruptive Technologie?

Für das Verständnis dessen, was „disruptiv“ bedeutet, gehen wir geschichtlich zurück in das 19. Jahrhundert. In dieser Zeit gab es eine Vielzahl an disruptiven Beispielen, die den Rückgang von großen sogenannten Pionieren verdeutlichen. Dies ist in der Automobil-, Elektronik- oder Chemiebranche, aber auch in anderen Branchen, die auf mechanische Technologien Wert legen, zu beobachten. Ein sehr bekanntes Beispiel ist das Segelschiff Thomas W. Lawson aus dem Jahre 1902. Es war das größte je gebaute Segelschiff ohne Hilfsantrieb und damit auch der größte jemals gebaute Schoner. Bei der ersten Atlantiküberquerung geriet das Schiff in einen Sturm und kenterte vor den Scilly-Inseln. Der Großteil der Besatzung kam dabei ums Leben. Kurze Zeit darauf wurde das Segelschiff vom Dampfschiff abgelöst. Schon Jahre zuvor war mit dem Dampfschiff eine konkurrierende Erfindung zu der Segelschifffahrt zum Vorschein gekommen. Die Dampfschiffe waren bis dato zwar nicht leistungsfähiger als die Segelschiffe, jedoch konnten sie in der Binnenschifffahrt ihre Vorteile gegenüber den Segelschiffen ausspielen und sich auf einem eigenen neuen Markt entwickeln. Sowohl ohne Wind als auch gegen den Wind konnten die Dampfer ohne Probleme fahren. Die Hersteller der Segelschiffe verpassten indessen den technologisch radikalen Sprung zum Dampfer, weil sie auf ihre Stammkunden hörten und die Segelschiffe weiterentwickelten. Der Untergang der Segelschifffahrt war nach der Ausreifung der Dampfschiffe auf Ozeanfahrten und dem Untergang der Thomas W. Lawson besiegelt. Ein weiteres Beispiel zeichnete sich Anfang der 60er Jahre ab. Das Handelsunternehmen Sears, Roebuck & Co. galt als eines der bestgeführten Unternehmen weltweit. Aus diesem Unternehmen stammen die Supply Chain Excellence und somit die Etablierung von Handelsmarken, der Versandhandel wie auch die Bereitstellung der Kundenkreditkarten. Anfang der 90er Jahre verpasste Sears Roebuck jedoch die Investition im Discounter-Markt und das Management wurde für seine starre Denkweise von der Presse und Wirtschaftsexperten kritisiert. Dennoch spezialisierte sich das Unternehmen in seiner Strategie weiter und riss das Kreditkartengeschäft an sich. Nach und nach wurden Anteile von Sears und die damit verbundenen Geschäftsmodelle verkauft. Das nächste Beispiel sollte nun für das heutige Zeitalter der Technologie als Warnhinweis dienen. Das Unternehmen IBM dominierte

den Markt für Großrechner. IBM fokussierte sich dabei auf sein aufgebautes Know-how und optimierte Technologie stetig weiter. Der Trend ging jedoch in eine weitaus simplere Richtung, nämlich die der Minicomputer. Dann trat Digital Equipment Corporation in diesem Markt für Minicomputer ein und feierte große Erfolge. Nichtsdestotrotz entwickelte sich DEC im Hinblick auf den Trend der Computerbranche nicht weiter und verpasste den Sprung zum Desktop-PC. In diesem Markt waren Apple, Commodore, Tandy und IBM die Pioniere. Anhand des Technologielebenszyklus ist aber ersichtlich, dass auch diese genannten Unternehmen mit dem nächsten simpleren Sprung zu tragbaren Computern ihre Probleme hatten. An diesen Beispielen wird deutlich, dass einfache, simple Innovationen, auch disruptive Technologien, zu einer Gefahr für Pioniere werden können. Das Problem der Unternehmen, die eine hervorragende Technologie hervorbringen und diese evolutionär weiterentwickeln, ist, dass sie oft den radikalen Sprung der disruptiven Technologie verpassen. Im Hinblick auf die traditionelle Automobilindustrie sind dabei zwei Aspekte zu beachten, die Umwelt- und Marktbetrachtung zum einen und die innerbetriebliche Technologie, die genutzt wird, zum anderen. Die Merkmale disruptiver Technologien werden nach CLAYTON folgendermaßen definiert:

1. kostengünstig (Anschaffung)

2. besser zugänglich

3. struktureller Kostenvorteil gegenüber bestehenden Lösungen.

Beim Nutzen einer disruptiven Technologie ist nicht zwingend vorausgesetzt, dass daraus ein neues innovatives Produkt entsteht. Disruptive Technologien haben, wie oben erwähnt, den Vorteil, dass sie kostengünstig in der Anschaffung sind. Die vom Autor integrierte Big-Data-Architektur besteht aus Open-Source-Software. Lediglich die Bereitstellung der Hardware, die Programmierung und das Verständnis für die Anwendung sind als Human Ressource einzusetzen. Die Zugänglichkeit ist unter dem Aspekt der Open-Source-Architektur bereits gegeben, sodass das Know-how nicht unternehmensspezifisch erlangt werden muss. Der Kostenvorteil gegenüber anderen, insbesondere Konkurrenzprodukten, im kommerziellen Markt ist ausschlaggebend. Prozesse der automatisierten Datenverwertung und die hohen Verarbeitungsgeschwindigkeiten sind weitere klare Indizien für disruptive Technologien. Aus dieser Perspektive bietet Big Data als Technologie die Möglichkeit, Kernprozesse zu stärken und somit die Qualität und Innovation eines bestehenden oder auch neuen Produktes zu erhöhen. Auf den Nutzen und weitere Auswirkungen wird nun im Abschnitt 6.2 näher eingegangen.

6.2 Der Nutzen und die Auswirkungen von Big Data

Big Data hat sowohl Auswirkungen auf die Betriebsebene als auch auf die Managementebene. So verändert z.B. das Ergebnis eines Predictive-Maintenance-Ansatzes operativ die Instandhaltung. Eine Betrachtung der Gesamtproduktion deutet darauf hin, dass sowohl in der Planung als auch in der Fertigung selbst neue Strategien angepasst werden. Durch die Vernetzung der innerbetrieblichen Prozesse ist ausgehend von einem Prozess die gesamte Wertschöpfungskette involviert. Die Wertschöpfungskette beginnt mit den Bereichen F&E, führt weiter zum Einkauf, Logistik und Qualität. Anschließend läuft die Wertschöpfungskette weiter zur Produktion, zum Marketing und Vertrieb. Aftersales und Financial Services runden die gesamte Wertschöpfungskette ab (siehe Abbildung 74).

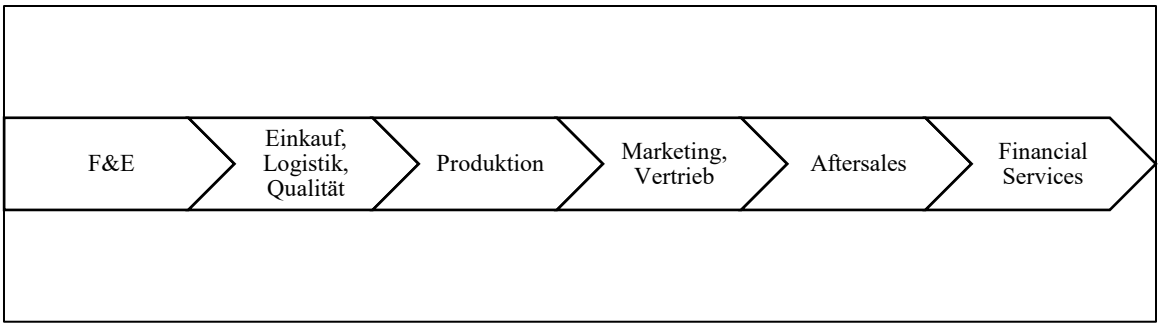


Abbildung 74: Wertschöpfungskette Automobilhersteller (eigene Darstellung)

In der folgenden Betrachtung werden nun die Hebelwirkungen durch die optimale Nutzung von Datenanalyse mittels Big-Data-Technologien erörtert. Dabei soll herauskristallisiert werden, welchen Einfluss das Ergebnis der Arbeit auf die Wertschöpfungskette im Unternehmen hat.

Die erarbeitete Big-Data-Architektur bildet die Grundlage für die darauf bezogenen Anwendungen. In der Arbeit wurde detailliert auf die Instandhaltung im Karosseriebau eingegangen, um die Instandhaltung und dahingehend auch die Produktion im weiteren Sinne zu optimieren. Durch mögliche Fehlerfrüherkennung der Anlage/Maschinen konnten Instandhaltungsstrategien verbessert werden, um den Produktionsstopp bzw. die Totzeit eines Prozesses zu verringern. Eine flexible Strategie zur Wartung, Instandsetzung und Inspektion wird durch das Wissen eines bestimmten Ausfallzeitpunktes ermöglicht. Zum einen wird dadurch ein verbessertes Ressourcenmanagement erreicht und zum anderen kann über den Produktionsprozess hinweg, speziell im Nachfolgeprozess der Kleberanlage, die Planung und Durchführung optimiert werden. Das hat u.a. Auswirkungen auf entstehende Kosten in der Produktion und Instandhaltung. Ein Nebeneffekt, der durch das Aufsetzen einer Big-Data-Architektur auftritt, ist das automatisierte Datenmanagement. Mit der automatisierten

Datenbeschaffung, Datengenerierung, Datenbereinigung und Datenanalyse können unternehmensweit analytische Betrachtungen durchgeführt werden. Aus der Wertschöpfungskette heraus können die Bereiche Marketing und Vertrieb Marktanalysen mit neuen Datensätzen ausführen, um flexibler auf Veränderungen reagieren zu können. Darauf wird im Abschnitt 6.2 näher eingegangen.

Daten und Informationen können somit eine Quelle von Wettbewerbsvorteilen und ein Treiber für erfolgreiche Geschäftsmodelle sein. Die Erhebung einer IBM-Studie bezüglich Innovation aus dem Jahr 2014 zeigt, dass „Organisationen, die Big Data und Analytics in ihren Innovationsprozessen verwenden, ihre Konkurrenten in Bezug auf Umsatzwachstum und Betriebseffizienz mit 36 % höherer Wahrscheinlichkeit schlagen werden“ (vgl. Marshall, Mueck, Shockley, 2015). Die Betrachtung ist umso interessanter, wenn ein traditioneller Produzent sich auf Daten einlässt und um die Daten herum in neue Geschäftsfelder eindringt. Das Big-Data-Konzept ist nach der Erkenntnis der Arbeit ein datengetriebenes Modell. Die Besonderheit ist, dass sowohl die historischen Daten als auch die aktuellen Daten in die Analyse einfließen und durch Visualisierung das Datenverständnis verbessert wird. Das automatisierte Datenmanagement ist ausschlaggebend für die erfolgreiche Umsetzung der Technologie. Im Kapitel 3 ist demnach das Feld der Datenwissenschaften als eine von vielen Möglichkeiten, ein neues Geschäftsmodell zu entwickeln, dargelegt. Bezogen auf die Automobilindustrie sind unter zwei Gesichtspunkten folgende Aspekte zu definieren, einerseits die Veränderung eines Produktes selbst und andererseits aus Sicht eines Produzenten eine Veränderung bezüglich der Geschäftsprozesse.

Das traditionelle Auto als Produkt mit seinen dazugehörigen Serviceleistungen verändert sich ständig und somit auch die Wertschöpfungskette. In der Abbildung 68 wurde die Wertschöpfungskette dargestellt. Daraus sind nun der Bereich Aftersales und (Financial-)Services hervorzuheben. In beiden Bereichen werden sich die Struktur und die strategische Ausrichtung ändern, sobald das Produkt sich in seinen Merkmalen verändert. Wie schon erwähnt, ist erstens Elektromobilität im Kommen und mit ihr zweitens datenbasierte Geschäftsmodelle als Produkt: Das Automobil wird sich in Zukunft mehr zum Software-Produkt entwickeln (siehe Tesla, Inc.). Drittens ist autonomes Fahren datenbasiert, sowohl, was die Steuerung anbelangt, als auch Wartung und Instandhaltung. Die Elektromobilität erfordert im Vergleich zu der herkömmlichen Mobilität mit Verbrennungsmotoren durch die spezifischen Bauteile gerade einmal 50 % Wartungskosten. Die Herausforderungen an den Produzenten sind klar ersichtlich. Einerseits muss ein Umdenken erfolgen, angefangen beim

Produkt selbst, bis hin zur Produktion und zur damit einhergehenden Wertschöpfungskette. Dabei geht es darum, das Aufbauen langfristiger Strategien und damit digitale Geschäftsprozesse in den Mittelpunkt zu stellen.

6.3 Flexibilität gegenüber Markt- und Umweltveränderungen

Im heutigen Technologiezeitalter entwickeln sich die technischen Möglichkeiten hinsichtlich der Mobilität, der Vernetzung und die damit verbundenen Anforderungen an die Technologien rasant. Der Markt und die Kundenbedürfnisse verändern sich fortlaufend. Aufgrund einer Vielzahl von Mitbewerbern und deren Produktlösung in der Automobilindustrie steigen auch die Kundenbedürfnisse. Gehen wir nun gezielt auf den Markt ein. Das Streben nach Weiterentwicklung und Effizienz beinhaltet zum einen die Betrachtung der Konkurrenten und eröffnet den technologischen Wettbewerb. Zum anderen sind reaktionsschnelle Entscheidungen bezüglich der Produktion und des daraus entstehenden Produkts unumgänglich. Dem quantitativ hohen Angebot steht zwar noch eine hohe Nachfrage gegenüber, dies ist aber gemessen an der Zeit variabel darzustellen. So zeichnet sich in den letzten Jahren ein leichter Rückgang der Antriebsart mit fossilen Kraftstoffen in der Automobilproduktion ab. Beträgt der Anteil der Benziner im Jahr 2018 noch 62,4 %, sind es im Jahr 2019 nur noch 59,6 %. Der Diesel verhält sich ähnlich zum Benziner. Im Jahr 2018 beträgt sein Anteil 32,3 % und im Jahr 2019 nur noch 29,8 %. Diesem negativen Verlauf stehen die alternativen Antriebsstränge mit Hybridtechnik oder dem Elektromotor gegenüber. So kann der hybride Antrieb seinen Marktanteil von 3,8 % auf 8,3 % und rein elektrisch angetriebene Autos von 1 % auf 2 % Marktanteil verbessern (KBA Center of Automotive Management, 2020).

Des Weiteren dürfen neue Wettbewerber nicht vernachlässigt werden. Durch das Eindringen neuer Mitbewerber in den Automobilmarkt stehen die traditionellen OEMs unter Zugzwang. Das bedeutet, dass sich das Produkt Automobil aus technologischer Sicht weiterentwickeln muss. Gleichzeitig sollten die aktuellen Prozesse und Strategien kritisch hinterfragt werden. Wie aus dem Abschnitt zuvor ersichtlich wurde, bietet Big Data die Möglichkeit, um die Daten herum neue Geschäftsmodelle zu bilden. Das Verständnis der Daten und die damit einhergehenden Analysen unterstützen die Ideenfindung als Innovationstreiber sowie die Prozessoptimierung. Der Vorteil traditioneller OEMs sind die Erfahrungen und Know-how der letzten Jahrzehnte, der große Kundenstamm und die erzeugten Daten. Eine Kombination aus den genannten Faktoren bringt enorme Vorteile gegenüber neuen Mitbewerbern. Sowohl im

Geschäftsprozess als auch am Produkt selbst sind die Entwicklungspotenziale bezüglich der Digitalisierung erkennbar (siehe Abbildung 75).

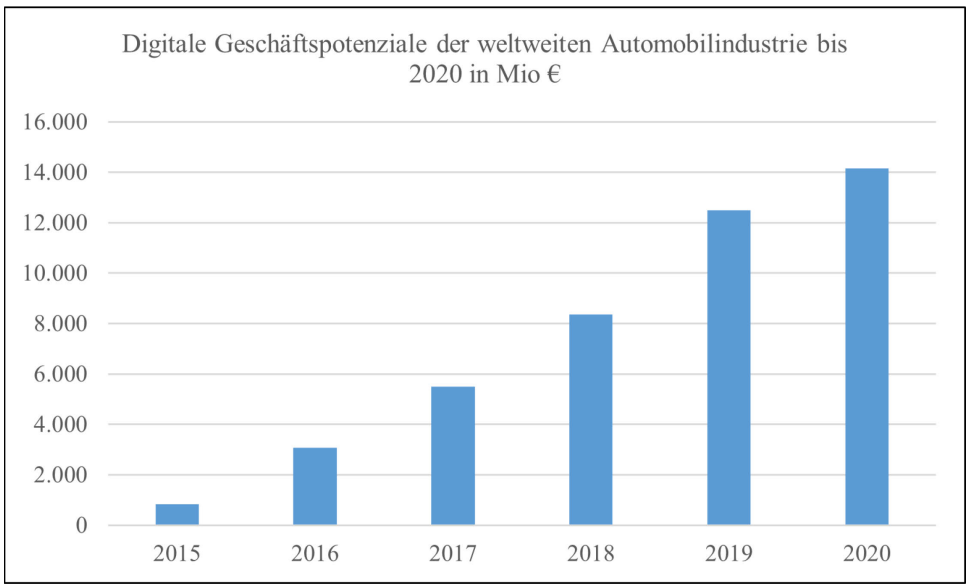


Abbildung 75: Befragung zu Digitalisierungspotenzialen (eigene Darstellung in Anlehnung an Strategy & Booz Company, 2014)

Auf diesem noch eher jungen Gebiet der zielgerichteten Datenanalyse wird durch die Anwendung geeigneter Werkzeuge ein neues Datenverständnis erzeugt. Mit diesen Erkenntnissen sind folglich neue Informationen bezüglich Produktion, Instandhaltung und der daraus folgenden Qualität zu entnehmen. Das belegt auch die Umfrage des Fraunhofer IPA. Sie kommt u.a. zu dem Ergebnis, dass ein Mehrwert im Unternehmen anhand der von ihnen produzierten Daten generiert wird. Dahinter folgen Logistik mit ca. 20 % und Produktentwicklung mit ca. 13 % (vgl. Fraunhofer IPA, 2016).

Die umweltpolitische Ausrichtung und die damit einhergehenden Restriktionen und Anforderungen verändern sich jedoch stetig. Das Produkt Auto muss sich also auch grundlegend ändern. Es muss nachhaltig und strategisch geplant werden, sodass im Gegensatz zum Verbrenner neue Produktions- und Software-Plattformen entwickelt werden, die den Herausforderungen gerecht werden und somit auf Veränderungen reagieren können. Unvorhersehbare Probleme wie Umweltkatastrophen, globale Erderwärmung und Pandemien zwingen einerseits zum Umdenken, andererseits helfen Daten aber auch dabei, die Umwelt zu analysieren und dadurch ein besseres Verständnis und nachhaltige Produkte zu erlangen. Ein Vorteil von zusätzlich generierten digitalen Geschäftsmodellen ist die Resistenz gegenüber Marktschwankungen bei nicht vorgesehenen Ereignissen. Zum einen wird das Produktportfolio erweitert, um somit eine höhere Flexibilität im Markt zu erzielen. Zum anderen reagieren

digitale Produkte anders auf Marktveränderungen. Verdeutlicht wird dies in der folgenden Chart-Abbildung der Aktienkurse. Aktienkurse spiegeln den Markt und die Art und Weise, wie er auf Veränderungen reagiert, sehr gut wider. In der Abbildung 76 ist ein Vergleich zwischen deutschen Automobilhersteller und US-Tech-Unternehmen zu sehen. Dieser Vergleich soll hervorheben, dass digitale Produkte sich von Marktveränderungen aus der Pandemie schnellstmöglich erholen und sogar davon profitieren. Ausgehend vom Zeitpunkt November 2019 ist eine Steigerung von mindestens 25 % bei Apple und ca. 65 % bei Alphabet und Amazon zu verzeichnen. Sie sind eindeutig die Profiteure der weltweiten Pandemie in 2020. Die OEMs hingegen ringen um eine Stabilisierung ihrer Ausgangsposition vor dem Ausbruch der Pandemie Anfang 2020. So können sich die BMW AG, Daimler AG und auch die Volkswagen AG, nachdem die Aktienkurse Anfang März 2020 eingebrochen waren, erst heute allmählich erholen.

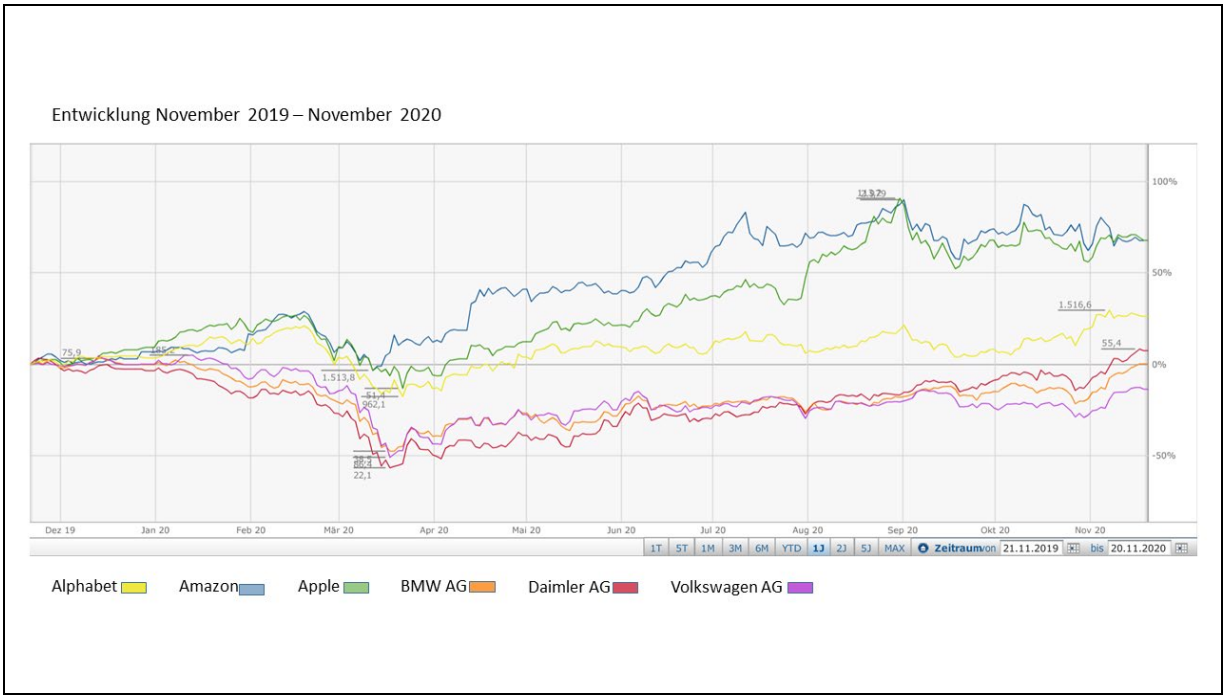


Abbildung 76: Vergleich Tech-Unternehmen und OEM (eigene Darstellung)

Wenn wir in den Vergleich nun einen weiteren Protagonisten in der digitalen und automobilen Welt einbeziehen, wird sehr deutlich, in welchen Sphären sich innovative Produkte bewegen können. In der Abbildung 77 ist Tesla mit abgebildet. Aus diesem Chart-Verlauf ist zu entnehmen, dass die Entwicklung des Aktienwertes von Tesla eine Steigerung von bis zu 600 % innerhalb eines Jahres verzeichnet.

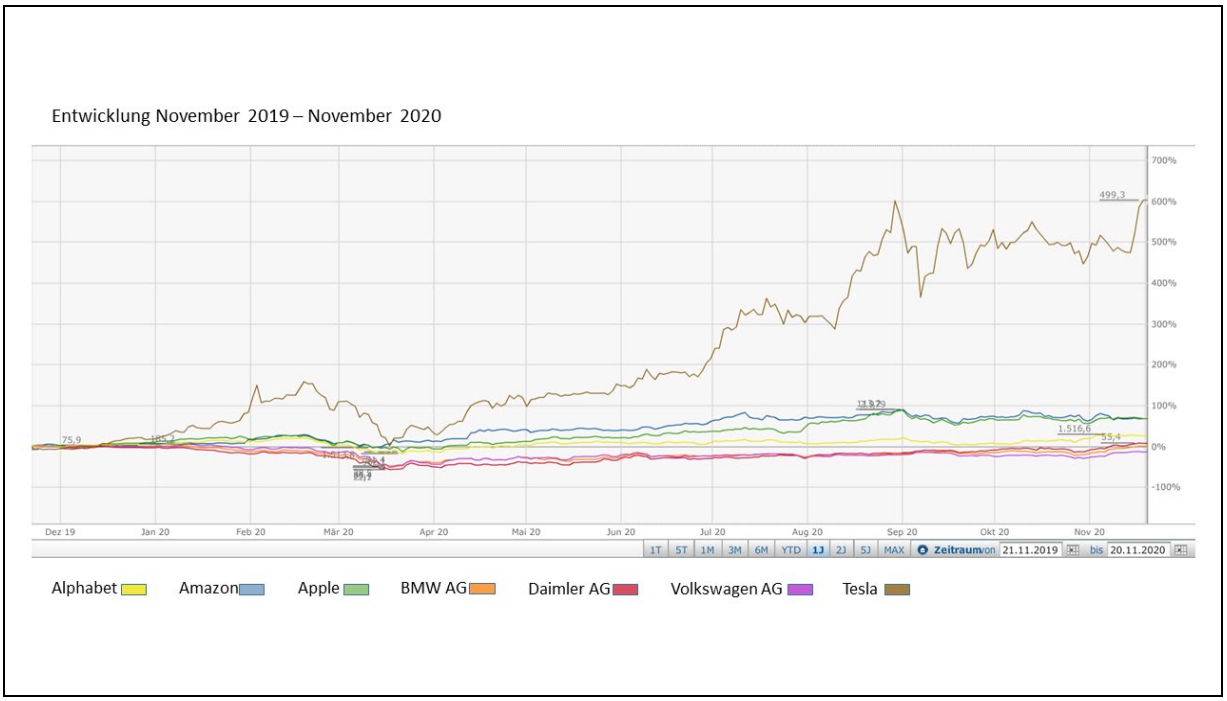


Abbildung 77: Vergleich Tech-Unternehmen, OEM und Tesla (eigene Darstellung)

Die Flexibilität ist ein großer Vorteil, den ein Automobilproduzent entlang der Wertschöpfungskette für sich nutzen sollte. Diesen Nutzen unterstreicht auch eine Umfrage zu Einsatzzwecken für die Nutzung von Big Data in Unternehmen aus dem Jahre 2017 (siehe Abbildung 78).

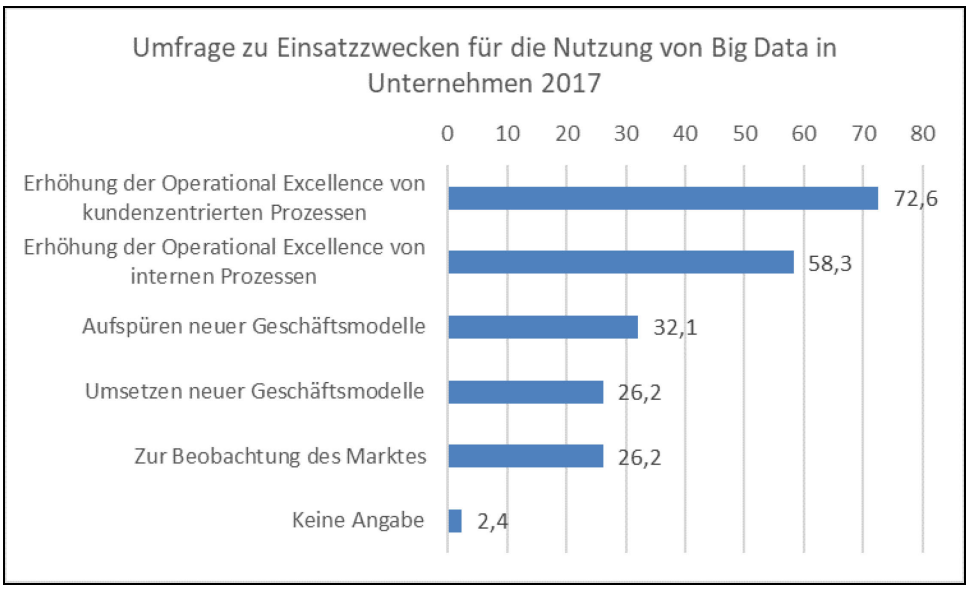


Abbildung 78: Einsatz von Big Data in Unternehmen (eigene Darstellung in Anlehnung an Capgemini, 2018)

Aus der Umfrage ist zu erkennen, dass der Großteil der befragten Unternehmen mit Big Data zunächst innerbetriebliche Prozesse optimiert. Der Trend geht aber auch dahin, dass Unternehmen Big-Data-Analysen zum Aufspüren neuer Geschäftsmodelle sowie zur Beobachtung in dem agierenden Markt nutzen. Mit den Ergebnissen der Arbeit kann die Anwendung des Datenmanagements hoch skaliert werden, um die Vernetzung innerhalb des Unternehmens zu gewährleisten. Datenanalysen können unternehmensweit sowohl horizontal als auch vertikal integriert werden. In der vertikalen Integration ist die Datenanalyse auf allen Unternehmensebenen möglich. Damit werden schnelle Entscheidungshilfen geschaffen, die zu schnellen Reaktionsgeschwindigkeiten hinsichtlich der Planung und Durchführung führen.

Damit schließt sich der Kreis der in der Arbeit aufgeführten Vorgehensweise, ausgehend von der strategischen in die operative Betrachtung und durch die Adaption von Technologien im Operativen zurück zur Verbesserung von Strategien und Planungen.

7. Zusammenfassung und Ausblick

In der Arbeit wurde eine Problematik der Produktion und des Datenmanagements in der Automobilindustrie erläutert. Aus Sicht des Autors wurde zunächst eine Betrachtung der allgemeinen Produktion angestrebt, um daraus detaillierter in das operative Geschehen einzusteigen. Zudem wurde eine Einordnung der Themenschwerpunkte in die Produktion und somit den Zusammenhang zur Instandhaltung und zur Informationstechnik vorgenommen. Basierend auf der Erkenntnis, dass durch eine Optimierung im operativen Management eine Verbesserung des strategischen Managements erfolgen kann, wurde das Thema spezifisch eingegrenzt, um daraus einen Use Case zu entwickeln. Dieser wurde in Zusammenarbeit mit einem der weltweit größten Automobilhersteller aus Europa erarbeitet. Die Problemstellung ließ sich in drei Kategorien einteilen: Prozess, Technologie und Anwendung. Die erste Problemstellung Prozess lautet dabei: Welche Probleme im operativen Management können identifiziert werden? In diesem Zusammenhang wurde ermittelt, welche Technologien derzeit genutzt werden. Bei der Problemstellung der Technologie galt es, herauszufinden, welche innovativen Technologien das Prozessproblem beheben könnten. In der letzten Kategorie wurde dann die Anwendung thematisiert. Der Schwerpunkt lag darauf, ein geeignetes Konzept zu kreieren, welches genutzt werden kann, um die Produktion flexibler zu gestalten. Zur Evaluierung des Konzepts wurden zwei Maßnahmen getroffen. Einerseits konnte das Modell durch statistische Methoden wie Brier-Score oder Kreuzvalidierung auf seine Richtigkeit hin überprüft werden. Andererseits wurde eine Implementierung des Konzepts angestrebt und die Tauglichkeit in der Praxis erprobt. Letzteres befindet sich in der Testphase, sodass einige Aspekte zum Zeitpunkt der Fertigstellung dieser Arbeit noch nicht validiert werden konnten. Dies betrifft die Ergebniskontrolle des Predictive-Maintenance-Modells an der Kleberanlage im laufenden Betrieb. Dennoch wird das Modell mit aktuell erhobenen Daten weiter antrainiert. Mit der virtuellen Validierung wurde eine Anwendungsempfehlung für den Projektpartner verfasst.

> Welche Vorteile konnten durch Big Data und Predictive Maintenance gewonnen werden?

Rückblickend auf die in der Einleitung gestellten Forschungsfragen werden alle Aspekte positiv beantwortet:

1. Ist eine Optimierung hochautomatisierter Produktionsprozesse mit einer sehr geringen Ausfallquote durch neue Ansätze der Informationstechnik möglich?
2. Ist eine Implementierung unter Berücksichtigung der technischen Restriktion möglich?
3. Welche Strategien resultieren aus den Ergebnissen der Arbeit?

Die Vorzüge des Big-Data-Konzepts mit seiner Architektur und dem daraus entwickelten Prozess des Predictive Maintenance haben Einfluss auf das gesamte Unternehmen. Strategisch gesehen wurde durch die Entwicklung einer Big-Data-Architektur für das Automobilunternehmen ein skalierbares Werkzeug zur Datengenerierung und Datenverarbeitung bereitgestellt. Mit diesem können in Zukunft beliebig oft und beliebig viele Analysen durchgeführt werden. Die vom Autor entwickelte Architektur kann dabei sowohl unternehmensweite als auch unternehmensübergreifende Analysen verwirklichen. Die Aufhebung des Schichtenmodells wird mit der Big-Data-Architektur vorangetrieben. Aus operativer Sicht hat das Ergebnis der Zeitreihenanalyse der Arbeit zu einem sofortigen Nutzen geführt. Das Modell kann sowohl lokal auf mehreren Anlagen adaptiert und getestet werden, bevor es unternehmensweit und zentral integriert wird. Dabei wurde auch eine Anwendungsempfehlung für den Predictive-Maintenance-Ansatz erarbeitet. Die Skalierbarkeit ist einer der Vorzüge der Architektur und des erstellten Modells. Neben dem durchgeführten Projekt im Karosseriebau hat das Unternehmen nun die Möglichkeit, weitere Gebiete mit Datenanalysen aus der Big-Data-Architektur zu bereichern.

Weitere Forschung hinsichtlich Technologie in Bezug auf andere Sektoren/Bereiche:

Die vom Autor verfasste Dissertation beinhaltet viele Berührungspunkte im Management, in der Informationstechnik und im Anlagen- bzw. Maschinenbau. Das Datenverständnis wurde in der Arbeit hervorgehoben, um geeignete Modelle für die Automatisierung des Datenflusses zu entwickeln und daraus die Analysen betreiben zu können. Als Schwerpunkt ist u.a. die Orchestrierung vieler Einzelwerkzeuge anzusehen. Der Nebeneffekt beinhaltet dabei auch das Prozess- und das Maschinenverständnis. Auch konnten weitreichende Erkenntnisse durch den Prozess der Datenanalyse erlangt werden. Dennoch sieht sich diese Arbeit als Grundlage für weitere Forschungen. Sowohl die in der Arbeit aufgeführte Big-Data-Architektur als auch das Predictive-Maintenance-Modell für eine Fehlervorhersage sind je nach Anwendung flexibel auszugestalten. In der Arbeit wurde eine innerbetriebliche Analyse auf der Grundlage von Maschinen- und Sensorikdaten durchgeführt. Mit dieser Grundlage sieht der Autor Potenzial für weitere wissenschaftliche Arbeiten hinsichtlich der Kombination innerbetrieblicher Daten mit außerbetrieblichen Daten. Insbesondere für den Logistikbereich ist eine gesamte Supply-

Chain-Analyse demnach sehr reizvoll. Von der offenen Architektur können aber auch andere Bereiche wie Marketing, Vertrieb oder Aftersales profitieren. Der Aspekt der autonomen Systeme bekommt dadurch einen wichtigen Stellenwert. Autonome Fahrzeuge sind datengetriebene Systeme. Dabei wird zukünftig eine Flut von Daten von den Produkten erfasst. Um diese beherrschen zu können, ist ein umfassendes Konzept zum Datenmanagement unabdingbar. Unter diesem Gesichtspunkt ist das Thema Big Data für die Produktentwicklung sehr interessant und zugleich mit Herausforderungen verbunden. Ausgehend von der umfassenden Markt- und Produktanalyse bis hin zur Planung des Produktes, der Produktion und den daraus folgenden Managementprozessen in der gesamten Wertschöpfungskette können mit der Big-Data-Architektur sehr große Datenmengen erfasst, verarbeitet und genutzt werden. Dahingehend ist Big Data aus Konsumenten- und Produzentensicht zentral und wird es auch weiterhin bleiben.

Wirtschaftswissenschaftlicher Mehrwert (Gegenwirkung zu neuen Konkurrenten/Produkten, Erforschung neuer Geschäftsfelder, Erprobung neuer Technologien)

Diese Arbeit hat ihre wirtschaftswissenschaftliche Relevanz durch die Teilintegration des Konzepts in der Industrie untermauert. So bietet das Thema mit seiner Kombination aus Big-Data-Architektur und Predictive Maintenance einen neuen Ansatz zur Skalierung mehrerer Anwendungen wie Machine Learning oder Künstlicher Intelligenz in größerer Dimension. Ein weiterer wissenschaftlicher Mehrwert ist das Aufzeigen disruptiver Potenziale von Open-Source-Software, worauf im Ausblick eingegangen wird. Ein besonderes Merkmal dieser wissenschaftlichen Arbeit ist die interdisziplinäre Betrachtungsweise. Sowohl Methoden aus der Wirtschaftswissenschaft als auch Methoden aus der Medizin wurden miteinander kombiniert, um zu einem Ergebnis zu gelangen.

Ausblick

Entwicklung von Big Data, Internet of Things und Data Analytics.

Die Entwicklung der Datenanalyse und das daraus abgeleitete Verständnis der Schlagwörter Big Data, Internet of Things und Data Analytics zeigen, dass speziell in Deutschland weitere Forschung betrieben werden muss. Im internationalen Vergleich befindet sich Deutschland sowohl in der Digitalisierung als auch in der Technologieanwendung/-forschung im europäischen Vergleich im Mittelfeld (vgl. Europäische Kommission, 2020).

Herausforderung für Softwareindustrie bezüglich Kundenbindung (SAP, Oracle, Microsoft)

Technologisch betrachtet haben nicht nur OEM als Pioniere in der Automobilindustrie im Hinblick auf Digitalisierung und Innovation Nachholbedarf, sondern auch Software-Anbieter im industriellen Markt rund um Oracle, SAP, Microsoft oder IBM. Auch sie stehen unter Zugzwang, sich strategisch mit ihren Industrielösungen bezüglich ihrer Produkte neu aufzustellen. Die Abbildung 79 zeigt eine eindeutige Verteilung der Marktanteile an Big-Data-Produkten und Analytics-Software. Mit rund 40 % Marktanteilen dominieren Open Source und andere Anbieter den Analytics-Software-Markt.

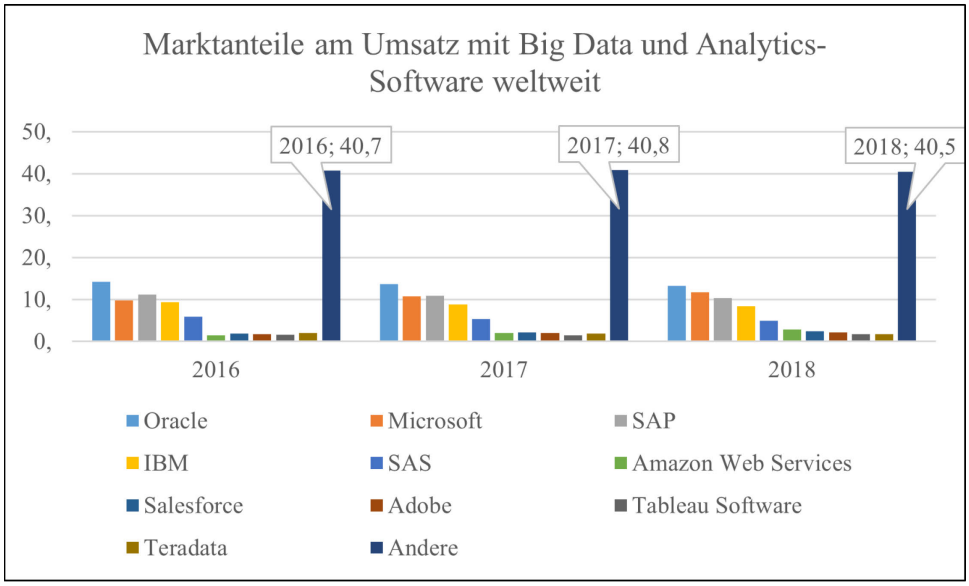


Abbildung 79: Marktanteile Big Data und Analytics-Software (eigene Darstellung in Anlehnung an SAS Institute, 2019)

Die Arbeit hat gezeigt, dass durch Open-Source-Lösungen innovative, unkonventionelle Lösungen der Datenanalyse gewonnen werden können. Dabei sind die Kunden aus der Industrie nicht abhängig von ihrem IT-Systemanbieter. Insbesondere die Funktionalität und die einfache Ausführung einzelner Software oder ganzer Ökosysteme können einfach darin integriert werden, ohne dass die vorhandene IT-Architektur umgestellt werden muss. Die Geschichte der disruptiven Technologien hat gezeigt, dass das Eindringen neuer innovativer und insbesondere disruptiver Produkte in ein bestehendes Ökosystem zur Folge haben kann, dass selbst Pioniere scheitern.

IV. Literaturverzeichnis

Abts, D., & Mülder, W. (2017). *Grundkurs Wirtschaftsinformatik: Eine kompakte und praxisorientierte Einführung* (9.,erweiterte und aktualisierte Ausg.). Wiesbaden: Springer Vieweg.

Adam, D. (1993). *Produktions-Management.* Wiesbdaden: Springer Gabler.

Akca, N., & Ilas, A. (2005). *Produktionsstrategien: Überblick und Systematisierung.* Essen: Universität Duisburg-Essen.

Alpaydin, E. (2010). *Introduction to Machine Learning* (2. Ausg.). Cambridge: MIT Press.

Apache Foundation. (2019a). *Apache Kafka*. Abgerufen am 20. September 2019 von https://kafka.apache.org/intro.html

Apache Foundation. (September 2019b). *Apache Nifi*. Abgerufen am 20. September 2019 von https://nifi.apache.org/

Apache Foundation. (September 2019c). *Apache Spark*. Abgerufen am 20. September 2019 von https://spark.apache.org/

Apache Foundation. (September 2019d). *Apache Hadoop*. Abgerufen am 21. September 2019 von https://hadoop.apache.org/

Apache Foundation. (September 2019e). *Apache Hive*. Abgerufen am 21. September 2019 von https://hive.apache.org/

Apache Foundation. (25. September 2019f). *Apache Cassandra*. Abgerufen am 25. September 2019 von https://cassandra.apache.org/

Apache Foundation. (Oktober 2019g). *Apache HBase*. Abgerufen am 1. Oktober 2019 von https://hbase.apache.org/

Bauer, A., & Günzel, H. (2013). *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung.* Heidelberg: dpunkt.Verlag.

Beierle, C., & Kern-Isberner, G. (2014). *Methoden wissensbasierter Systeme: Grundlagen, Algorithmen, Anwendungen* (5., überarb. und erw. Ausg.). Wiesbaden: Springer Vieweg.

Ben-Daya, M., Duffuaa, S. O., & Raouf, A. (2000). *Maintenance, Modeling and Optimization.* Boston, MA; s.l.: Springer US.

Berner, W. (2012). *Culture Change: Unternehmenskultur als Wettbewerbsvorteil.* Stuttgart: Schäffer-Poeschel.

Bischoff, J., Taphorn, C., Wolter, D., Braun, N., Fellbaum, M., Goloverov, A., & Ludwig, S. (2015). *Studie: Erschließen der Potenziale der Anwendung von Industrie 4.0 im Mittelstand.* Bundesministerium für Wirtschaft und Energie (BMWi). Abgerufen am 9. Oktober 2019 von https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/erschliessen-der-potenziale-der-anwendung-von-industrie-4-0-im-mittelstand.pdf %3F__blob %3DpublicationFile %26v %3D5

Bishop. (2006). *Pattern Recognition and Machine Learning.* New York: Springer-Verlag.

Blecker, T., & Kaluza, B. (2003). *Forschung zu Produktionsstrategien – Ergebnisse und Entwicklungsperspektiven.* Klagenfurt: Universität Klagenfurt.

Bleicher, K. (1991). *Organisation: Strategien - Strukturen - Kulturen* (2., vollständig neu bearbeitete und erweiterte Ausg.). Wiesbaden: Gabler Verlag.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT' 98 Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 92-100.

Brauchlin, E., & Pichler, J. H. (2000). *Unternehmer und Unternehmensperspektiven für Klein- und Mittelunternehmen.* Berlin: Duncker & Humboldt.

Brink, H., Richards, J. W., & Fetherolf, M. (2017). *Real-world machine learning.* Shelter Island, NY: Manning Publications Co.

Brossardt, B. (2014). Dienstleistungspotenziale im Rahmen von Industrie 4.0. *vbw - Die Bayrische Wirtschaft*. Abgerufen am 8. Januar 2018 von https://docplayer.org/3440151-Information-dienstleistungspotenziale-im-rahmen-von-industrie-4-0-stand-maerz-2014-www-vbw-bayern-de.html

Capgemini. (2018). *Studie IT-Trends 2018.*

Cattell, R. (Dezember 2010). Scalable SQL and NoSQL Data Stores. *Sigmod Record, 4*(39), 12-27.

Chapelle, O., Schölkopf, B., & Zien, A. (2010). *Semi-supervised learning.* Cambridge: MIT Press.

Chazal, P. d., O'Dwyer, M., & Reilly, R. B. (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE transactions on bio-medical engineering, 51*(7), 1196–1206.

Chen, Y., Hu, B., Keogh, E., & Batista, G. E. (2013). DTW-D. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD*, 383-391.

Cleve, J., & Lämmel, U. (2016). *Data Mining* (2. Ausg.). Berlin: De Gruyter.

Clutterbuck, D., & Ragins, B. R. (2002). *Mentoring and diversity: An international perspective.* London, New York: Routledge - Taylor & Francis Group.

Collins, J. C., & Porras, J. I. (1. October 1996). Building Your Companies Vision. *Havard Business Review, 74*(5), S. 65-77.

Conway, D. (2010). *drewconway.com*. Abgerufen am 21. April 2017 von http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Corsten, H., Gössinger, R., & Spengler, T. S. (2018). *Handbuch Produktions- und Logistikmanagement in Wertschöpfungsnetzwerken.* Berlin, Boston: Walter de Gruyter GmbH.

Daugherty, P., Banerjee, P., Negm, W., & Alter, A. (2015). Driving unconventional growth through the Industrial Internet of Things. (Accenture, Hrsg.) *Accenture Technology*.

Davenport, T. H., & Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review, 90*(10), 70–76.

Demchenko, Y., Ngo, C., & Membrey, P. (2013). *Architecture Framework and Components for the Big Data Ecosystem.* System and Network Engineering. Amsterdam: System and Network Engineering Group, UvA. Abgerufen am 11. Mai 2018 von http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.670.8078&rep=rep1&type=pdf

DIN Deutsches Institut für Normung. (2012). Grundlagen der Instandhaltung. (DIN e.V., Hrsg.) *DIN 31051: 2012-09*.

DIN Deutsches Institut für Normung. (2015). Maintenance terminology. (DIN e.V., Hrsg.) *DIN EN 13306:2015-09*.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics, 4*(26), 745–766. doi:10.1080/10618600.2017.1384734

Dorschel, J. (2015). *Praxishandbuch Big Data: Wirtschaft - Recht - Technik.* Wiesbaden: Springer Gabler.

Dyckhoff, H., & Spengler, T. S. (2010). *Produktionswirtschaft.* Berlin, Heidelberg: Springer-Verlag.

Ehrenmann, F. (2015). *Kosten- und zeiteffizienter Wandel von Produktionssystemen.* Wiesbaden: Springer Gabler.

Endrenyi, J., Aboresheid, S., Allan, R. N., Anders, G. J., Asgarpoor, S., Billinton, R., . . . Dialynas, E. N. (2001). The present status of maintenance strategies and the impact of maintenance on reliability. *IEEE Transactions on Power Systems, 16*(4), S. 638-646.

Ertel, W. (2016). *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung.* Wiesbaden: Springer Verlag.

Fallenbeck, N., & Eckert, C. (2014). IT-Sicherheit und Cloud Computing. In Bauernhansl, Hompel et al., *Industrie 4.0 in Produktion, Automatisierung und Logistik* (S. 397–431). Wiesbaden: Springer Vieweg.

Farkisch, K. (2011). *Data-Warehouse-Systeme kompakt: Aufbau, Architektur, Grundfunktionen.* Berlin, Heidelberg: Springer Verlag.

Fasel, D., & Meier, A. (2016). *Big Data: Grundlagen, Systeme und Nutzungspotenziale.* Wiesbaden: Springer Vieweg.

Fels, G., & Schinkel, F. (2015). IT-Infrastrukturen für Big Data. In J. Dorschel, *Praxishandbuch Big Data* (S. 278-306). Wiesbaden: Springer Gabler.

Fraunhofer IPA. (2016). *Entwicklungsfelder für den Mittelstand 2016.*

Freiknecht, J. (2014). *Big Data in der Praxis: Lösungen mit Hadoop, HBase und Hive : Daten speichern, aufbereiten, visualisieren.* München: Hanser-Verlag.

Freiknecht, J., & Papp, S. (2018). *Big Data in der Praxis: Lösungen mit Hadoop, Spark, HBase und Hive. Daten speichern, aufbereiten, visualisieren* (2. Ausg.). München: Hanser-Verlag.

Freytag, J.-C. (2014). Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data. *Informatik Spektrum, 37*(2), 97-104. doi:10.1007/s00287-014-0771-y

Fritz, A. H., & Schulze, G. (2015). Fügen. In A. H. Fritz, & G. Schulze, *Fertigungstechnik* (11., neu bearbeitete und ergänzte Ausg., S. 117-256). Berlin, Heidelberg: Springer Vieweg.

Fu, C., Ye, L., Liu, Y., Yu, R., Iung, B., Cheng, Y., & Zeng, Y. (2004). Predictive Maintenance in Intelligent-Control-Maintenance-Management System for Hydroelectric Generating Unit. *IEEE Transactions on Energy Conversion, 19*(1), 179–186.

Gadatsch, A. (2012). *Grundkurs Geschäftsprozess-Management: Methoden und Werkzeuge für die IT-Praxis: Eine Einführung für Studenten und Praktiker* (7. Ausg.). Wiesbaden: Imprint Vieweg+Teubner Verlag.

Galar, D., & Kumar, U. (2017). *eMaintenance: Essential electronic tools for efficiency.* Cambridge: Academic Press.

Gao, J., Koronios, A., & Selle, S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. *AMCIS*.

Ghahramani, Z. (2003). Unsupervised Learning. In S. Mendelson, *Advanced Lectures on Machine Learning* (S. 72-112). Berlin: Springer-Verlag.

Gluchowski, P., Dittmar, C., & Gabriel, R. (2008). *Management Support Systeme und Business Intelligence: Computergestützte Informationssysteme für Fach- und Führungskräfte* (2, vollst. überarb. Ausg.). Heidelberg: Springer Verlag.

Goldschmidt, A. (2009). Was bedeutet eigentlich ... Data Science? (Berufsverband Medizinischer Informatiker e.V., Hrsg.) *DATA SCIENCE: Auswirkungen auf med. Dokumentation und Informatik, 2*.

Gopalkrishnan, V., Steier, D., Lewis, H., & Guszcza, J. (2012). Big data, big business: Bridging the gap. (Fan, Hrsg.) *BigMine '12*, 7–11.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques* (Bd. Intelligent Systems Reference Library). Berlin, Heidelberg: Springer-Verlag.

Gutenberg, E. (1983). *Grundlagen der Betriebswirtschaftslehre.* Berlin, Heidelberg: Springer Verlag.

Hachtel, G., & Holzbaur, U. (2010). *Management für Ingenieure.* Wiesbaden: Vieweg + Teubner.

Han, J., E, H., Guan, L., & Jian, D. (2011). Survey on NoSQL database. *6th International Conference on Pervasive Computing and Applications*, 363–366.

Haneke, U., Trahasch, S., & Zimmer, M. (2019). *Data Science: Grundlagen, Architekturen und Anwendungen* (1. Ausg.). Heidelberg: dpunkt.Verlag.

Hansen, H. R., & Neumann, G. (2005). *Informationstechnik* (9. Ausg.). Stuttgart: Lucius & Lucius.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama, 287*(18), 2543-2546.

Hartmann, H. (2002). *Materialwirtschaft: Organisation, Planung, Durchführung, Kontrolle* (8., überarb. und erw. Ausg.). Gernsbach: Dt. Betriebswirte Verlag.

Hutzschenreuter, T. (2015). *Allgemeine Betriebswirtschaftslehre.* Wiesbaden: Springer Gabler.

Inmon, W. H. (1996). *Building the data warehouse* (2. Ausg.). New York: Wiley.

Jahn, M. (2016). *Ein Weg zu Industrie 4.0: Geschäftsmodell für Produktion und After Sales.* Berlin, Boston: De Gruyter.

Kaiser, K. A., & Gebraeel, N. Z. (2009). Predictive Maintenance Management Using Sensor-Based Degradation Models. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 39*(4), 840–849.

Kaur, S., & Jindal, S. (2016). A Survey on Machine Learning Algorithms. *International Journal of Innovative Research in Advanced Engineering (IJIRAE), 3*(11), S. 6-14.

Kellner, F., Lienland, B., & Lukesch, M. (2020). *Produktionswirtschaft: Planung, Steuerung und Industrie 4.0* (2. Ausg.). Wiesbaden: Springer Gabler.

Kern, W. (1996). *Handwörterbuch der Produktionswirtschaft* (2., völlig neu gestaltete Ausg.). Stuttgart: Schäffer-Poeschel.

Khanum, M., Mahboob, T., Imtiaz, W., Abdul Ghafoor, H., & Sehar. (2015). A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance. *IJCA (International Journal of Computer Applications), 113*(15), 34–39.

Kiener, S., Meier-Scheubeck, N., Obermaier, R., & Weiß, M. (2017). *Produktionsmanagement.* Berlin: De Gruyter.

Kieser, A., & Walgenbach, P. (2003). *Organisation* (4., überarb. und erw. Ausg.). Stuttgart: Schäffer-Poeschel.

Kletti, J. (2007). *Konzeption und Einführung von MES-Systemen.* Berlin, Heidelberg: Springer-Verlag.

König, C., Schröder, J., & Wiegand, E. (2018). *Big Data: Chancen, Risiken, Entwicklungstendenzen.* Wiesbaden: Springer VS.

Krüger, W. (1994). *Organisation der Unternehmung* (3., verb. Ausg.). Stuttgart: Kohlhammer.

Kunath, M. (2018). In R. Obermeier, *Handbuch Industrie 4.0 und Digitale Transformation.* Wiesbaden: Springer Fachmedien Wiesbaden.

Lanquillon, C., & Mallow, H. (2015a). Advanced Analytics mit Big Data. In J. Dorschel (Hrsg.), *Praxishandbuch Big Data : Wirtschaft - Recht - Technik* (S. 55-89). Wiesbaden: Springer Gabler.

Lanquillon, C., & Mallow, H. (2015b). Big Data-Lösungen. In J. Dorschel, *Praxishandbuch Big Data: Wirtschaft - Recht - Technik* (S. 263-277). Wiesbaden: Springer Gabler Verlag.

Laudon, K. C., Laudon, J. P., & Schoder, D. (2016). *Wirtschaftsinformatik: Eine Einführung* (3., vollständig überarbeitete Ausg.). Hallbergmoos: Pearson.

Leidinger. (2017). *Wertorientierte Instandhaltung* (2. Ausg.). Wiesbaden: Springer Verlag.

Luber, S. (2018). Was ist ein Data Lake? *BigData Insider*. Abgerufen am 19. März 2019 von http://www.bigdata-insider.de/was-ist-ein-data-lake-a-686778/

Luber, S., & Litzel, N. (2016). Was ist Spark? *Big Data Insider*. Abgerufen am 14. Januar 2018 von https://www.bigdata-insider.de/was-ist-spark-a-572706/

Luber, S., & Litzel, N. (2017). Was ist Digitalisierung? *BigData Insider*. Abgerufen am 28. Januar 2020 von https://www.bigdata-insider.de/was-ist-digitalisierung-a-626489/

Lucke, D., Defranceski, M., & Adolf, T. (2017). Cyberphysische Systeme für die prädiktive Instandhaltung. In B. Vogel-Heuser, T. Bauernhansl, & M. Ten Hompel, *Handbuch Industrie 4.0* (2., erweiterte und bearbeitete Ausg., Bd. 1. : Produktion). Berlin: Springer Vieweg.

Luczak, H., & Eversheim, W. (1999). *Produktionsplanung und -steuerung: Grundlagen, Gestaltung und Konzepte.* Berlin, Heidelberg: Springer-Verlag.

Matyas, K. (2002). Ganzheitliche Optimierung durch individuelle Instandhaltungsstrategien. *Industrie Management, 18*(2), 13-16.

Meier, A., & Kaufmann, M. (2016). *SQL- & NoSQL-Datenbanken* (8., überarbeitete und erweiterte Ausg.). Berlin, Heidelberg: Springer Vieweg.

Mertens, P. (1995). Wirtschaftsinformatik — Von den Moden zum Trend. In *König (Hg.) – Wirtschaftsinformatik '95* (S. 25–64).

Mertens, P., Bodendorf, F., König, W., Schumann, M., Hess, T., & Buxmann, P. (2017). *Grundzüge der Wirtschaftsinformatik* (12., grundlegend überarbeitete Ausg.). Berlin: Springer Gabler.

Mitchel, T. (1997). *Machine Learning.* New York: McGraw-Hill Education.

Müller, S. (2016). Erweiterung des Data Warehouse um Hadoop, NoSQL & Co. In M. Fasel, *Big Data: Grundlagen, Systeme und Nutzungspotenziale* (S. 139–158). Wiesbaden: Springer Vieweg.

Müller-Wiegand, M. (2019). Integrale Unternehmensführung und Zukunftsfähigkeit. In M. Groß, M. Müller-Wiegand, & D. F. Pinnow, *Zukunftsfähige Unternehmensführung* (S. 33-44). Berlin: Springer Gabler.

Nagl, A. (2015). *Der Businessplan: Geschäftspläne professionell erstellen Mit Checklisten und Fallbeispielen.* Wiesbaden: Springer Fachmedien Wiesbaden.

Nyhuis, P. (2008). *Beiträge zu einer Theorie der Logistik.* Berlin, Heidelberg: Springer-Verlag.

Omri, F. (2015). Big Data-Analysen: Anwendungsszenarien und Trends. In J. Dorschel, *Praxisbuch Big Data: Wirtschaft - Recht - Technik* (S. 104-112). Wiesbaden: Springer Gabler Verlag.

Prassol, P. (2016). In-Memory-Platform SAP HANA als Big Data-Anwendungsplattform. In D. Fasel, & A. Meier, *Big Data: Grundlagen, Systeme und Nutzungspotenziale* (S. 195-210). Wiesbaden: Springer Vieweg.

Quaing, B. (2010). *Einsatz und Nutzen von Business Intelligence in Unternehmen: Ergebnisse einer kausalanalytischen Untersuchung* (Bd. 51). Hamburg: Kovač Verlag.

Rouse, M. (2016a). What is a Data Lake? *TechTarget: The Evolution of Data Center Storage Architecture*. Abgerufen am 1. Februar 2020 von https://searchaws.techtarget.com/definition/data-lake

Rouse, M. (2016b). What is a Data Warehouse? *TechTarget (An admin's guide to AWS data management)*. Abgerufen am 1. Februar 2020 von https://searchdatamanagement.techtarget.com/definition/data-warehouse

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science Magazin, 290*(5500), 2323–2326.

Rudolph, T., & Linzmajer, M. (2014). Big Data im Handel. *Marketing Review St. Gallen, 31*(1), S. 12–25.

Samtleben, M. (2007). *Wirkung von Business Intelligence auf die Allokation controllingspezifischer Aufgaben* (Bd. 42). Hamburg: Kovač Verlag.

SAS. (2013). *Five Big Data Challenges and how to overcome them with visual analytics.* SAS. Abgerufen am 3. August 2018 von https://de.slideshare.net/AllAnalytics/five-big-data-challenges-and-how-to-overcome-them-with-visual-analytics

Sathi, A. (2012). *Big Data Analytics: Disruptive technologies for changing the game.* Chicago: MC Press.

Schenk, M. (2013). *Instandhaltung technischer Systeme.* Heidelberg: Springer.

Schöning, H., & Dorchain, M. (2014). Data Mining und Analyse. In Bauernhansl, Hompel et al., *Industrie 4.0 in Produktion, Automatisierung und Logistik* (S. 543–554). Wiesbaden: Springer Vieweg.

Schuh, G., & Schmidt, C. (2014). *Produktionsmanagement.* Berlin, Heidelberg: Springer Verlag.

Schulmeyer, C. (2015). Herausforderungen an Big Data-Analyse. In J. Dorschel, *Praxishandbuch Big Data: Wirtschaft - Recht - Technik* (S. 307-328). Wiesbaden: Springer Gabler Verlag.

Schwabacher, M., & Goebel, K. (November 2007). A Survey of Artificial Intelligence for Prognostics. *AAAI Fall Symposium: Artificial Intelligence for Prognostics*, 108-115.

Seibt, D. (1991). Informationssystem: Architekturen–Überlegungen zur Gestaltung von technikgestützten Informationssystemen für Unternehmungen. In *Innovations-und Technologiemanagement* (S. 251-280). Stuttgart: Schäffer-Poeschl.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*(5), 13-22.

Shin, J.-H., & Jun, H.-B. (2015). On condition based maintenance policy. *Journal of Computational Design and Engineering, 2*(2), 119-127.

Sicular, S. (2012). No Data Scientist Is an Island in the Ocean of Big Data. *Gartner*.

Siepmann, D., & Graef, N. (2016). Industrie 4.0 – Grundlagen und Gesamtzusammenhang. In A. Roth, *Einführung und Umsetzung von Industrie 4.0.* Berlin, Heidelberg: Springer Gabler.

Soutier, M. (18. Januar 2015). *Marius Soutier, Softwareentwicklung und -beratung*. Abgerufen am 9. Februar 2020 von http://www.soutier.de/blog/2015/03/11/apache-spark-intro/

Srinivasa, K. G., Siddesh, G., & Srinidi, H. (2018). *Network Data Analytics: A Hands-On Approach for Application Development.* Cham: Springer International Publishing.

Strategy & Booz Company. (2014). *Connected Car Studie 2014.* PwC.

Strunz, M. (2012). *Instandhaltung: Grundlagen - Strategien - Werkstätten.* Berlin: Springer Vieweg.

Sullivan, G., Pugh, R., Melendez, A., & Hunt, W. (2010). Operations & Maintenance Best Practices. (O. o. (OSTI), Hrsg.) *U.S. Department of Energy, Federal Energy Management Program*.

Susto, G. A., Beghi, A., & Luca, C. d. (2012). A Predictive Maintenance System for Epitaxy Processes Based on Filtering and Prediction Techniques. *IEEE Transactions on Semiconductor Manufacturing, 25*(4), 638–649.

Suthaharan, S. (2016). *Machine learning models and algorithms for Big Data classification: Thinking with examples for effective learning.* New York: Springer.

Teubner, R. A. (1999). *Organisations- und Informationssystemgestaltung: Theoretische Grundlagen und integrierte Methoden.* Wiesbaden; s.l.: Deutscher Universitätsverlag.

Thommen, J.-P., Achleitner, A.-K., Gilbert, D. U., Hachmeister, D., & Kaiser. (2007). *Allgemeine Betriebswirtschaftslehre* (8., vollständig überarbeitete Auflage Ausg.). Wiesbaden: Springer Gabler.

Thomson, R., Edwards, M., Britton, E., & Rabenau, B. (2014). Predictive Maintenance: Is the timing right for predictive maintenance in the manufacturing sector? *THINK ACT - Roland Berger*.

Thonemann, U. (2015). *Operations Management: Konzepte, Methoden und Anwendungen* (3., aktualisierte Auflage Ausg.). Hallbergmoos: Pearson.

Tran Anh, D., Dabrowski, K., & Skrzypek, K. (2018). The Predictive Maintenance Concept in the Maintenance Department of the Industry 4.0 Production Enterprise. *Foundations of Management*(Vol.10), 283-292.

Umbeck, T. (2009). *Musterbrüche in Geschäftsmodellen.* Wiesbaden: Gabler Verlag / GWV Fachverlage GmbH Wiesbaden.

Voigt, K.-I. (2008). *Industrielles Management: Industriebetriebslehre aus prozessorientierter Sicht.* Berlin, Heidelberg: Springer Verlag.

Wagner, S., & Wagner, D. (2007). *Comparing Clusterings - An Overview.* Karlsruhe.

Wannenwetsch, H. (2014). *Integrierte Materialwirtschaft, Logistik und Beschaffung* (5., neu bearb. Ausg.). Berlin: Springer Vieweg.

Wei, L., & Keogh, E. (2006). Semi-supervised time series classification. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD*, 748-753.

Wheel Wright, S. C. (1984). Manufacturing strategy: Defining the missing link. *Strat. Mgmt. J. (Strategic Management Journal)*, S. 77–91.

Wien, A., & Franzke, N. (2014). *Unternehmenskultur: Zielorientierte Unternehmensethik als entscheidender Erfolgsfaktor.* Wiesbaden: Springer Gabler.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (3. Ausg.). Amsterdam: Elsevier/Academic Press.

Winkelhake, U. (2017). *Die Digitale Transformation der Automobilindustrie: Treiber - Roadmap - Praxis.* Berlin, Heidelberg: Springer.

Witte, T. (1979). *Heuristisches Planen.* Wiesbaden: Gabler-Verlag.

Zahn, E. (1988). Produktionsstrategie. In H. A. Henzler, *Handbuch Strategische Führung* (S. 515–542). Wiesbaden: Springer Fachmedien Wiesbaden GmbH.

Zäpfel, G. (2000). *Taktisches Produktions-Management.* München: Oldenburg.

Zäpfel, G. (2001). *Grundzüge des Produktions- und Logistikmanagement.* München: Oldenburg.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms.* Boca Raton: CRC Press.

ZVEI, Zentralverband Elektroindustrie und Elektrotechnik e.V. (2017). Industrie 4.0: MES – Voraussetzung für das digitale Betriebs- und Produktionsmanagement. *ZVEI: Die Elektroindustrie*, 8-22.

Zwiener, I., Blettner, M., & Hommel, G. (2011). Survival Analysis. *Deutsches Aerzteblatt Online*, 163-169.

V. Anhang

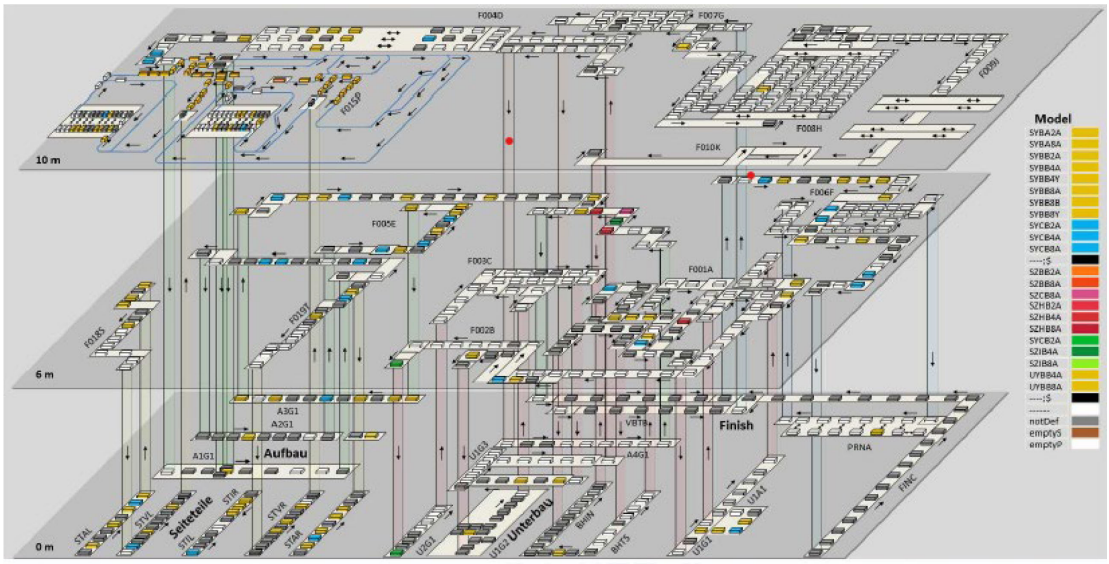
Rohdaten Kleberanlage (Auszüge)

	A	B	C	D	E
1	630209				
2	08.04.2017 00:07	144	0	1	0
3	08.04.2017 00:10	144	0	1	0
4	08.04.2017 02:15	43	0	1	82
5	08.04.2017 03:19	43	0	2	82
6	08.04.2017 12:54	1	0	1	81
7	08.04.2017 13:50	1	0	2	81
8	08.04.2017 19:35	168	0	1	81
9	08.04.2017 19:37	168	0	2	81
10	09.04.2017 01:23	204	0	1	81
11	09.04.2017 01:25	204	0	2	81
12	09.04.2017 01:46	207	0	1	2
13	09.04.2017 01:47	207	0	2	2
14	09.04.2017 14:08	1	0	1	3
15	10.04.2017 02:11	1	0	2	3
16	10.04.2017 06:04	144	0	1	0
17	10.04.2017 06:06	144	0	1	0
18	10.04.2017 16:47	207	0	1	2
19	10.04.2017 16:49	207	0	2	2
20	10.04.2017 17:19	204	0	1	21

	A	B	C	D	E	F	G	H	I	J
1	937252									
2	02.07.2020 00:43	12500	10823	1300	256	129	109	0	41	1
3	02.07.2020 00:45	27000	27894	2597	286	143	157	0	43	1
4	02.07.2020 00:46	6700	6961	985	199	104	114	0	21	1
5	02.07.2020 00:47	23500	23325	3550	214	109	127	0	22	1
6	02.07.2020 00:49	12500	10872	1300	257	130	109	0	41	1
7	02.07.2020 00:50	27000	27936	2597	283	142	159	0	43	1
8	02.07.2020 00:51	12500	10916	1299	257	130	107	0	41	1
9	02.07.2020 00:54	6700	6791	985	210	109	105	0	21	1
10	02.07.2020 00:55	23500	23275	3550	219	110	128	0	22	1
11	02.07.2020 00:57	6700	6854	985	211	109	110	0	21	1
12	02.07.2020 00:58	23500	23299	3550	214	109	128	0	22	1
13	02.07.2020 00:59	12500	10882	1300	259	131	109	0	41	1
14	02.07.2020 01:01	27000	27760	2597	285	143	158	0	43	1
15	02.07.2020 01:02	6700	6820	986	200	104	134	0	21	1
16	02.07.2020 01:05	6700	6671	986	211	111	135	0	21	1
17	02.07.2020 01:08	6700	6742	986	218	114	138	0	21	1
18	02.07.2020 01:10	14500	13527	1535	247	109	110	0	81	1
19	02.07.2020 01:12	30000	30883	3266	266	134	148	0	82	1
20	02.07.2020 01:35	6700	6807	985	232	122	115	0	21	1

	A	B
1	9033	
2	25.02.2016 12:28	16
3	09.03.2016 11:01	16
4	10.03.2016 20:13	1
5	11.03.2016 09:18	8
6	15.03.2016 08:56	16
7	15.03.2016 08:56	8
8	15.03.2016 10:30	1
9	15.03.2016 14:24	16
10	16.03.2016 08:49	1
11	17.03.2016 09:46	8
12	21.03.2016 15:55	16
13	30.03.2016 12:27	1
14	30.03.2016 15:10	16
15	02.04.2016 07:09	4
16	05.04.2016 14:48	64
17	06.04.2016 16:46	16
18	26.04.2016 11:46	4
19	02.05.2016 09:03	1
20	04.05.2016 15:05	8

Karosseriewerk



Kleberanlage Seitenteile