Fallstudie

Alternative: A

Big Data Management & Analytics
MBIGDM

eingereicht am 03.11.2024 über den eCampus
Modulverantwortlicher Hochschullehrer:
Dr. Christian Prinz
SRH Fernhochschule
The Mobile University

Modul: Big Data Management & Analytics
Studiengang: (M.Sc.) Data Science & Analytics

von
Dagmar Rostek
Klosterstr 12
52511 Geilenkirchen
Studiengang: (M.Sc.) Data Science & Analytics
Matrikelnummer: 2004169

ABBILDUNGSVERZEICHNIS 3

GENDER-ERKLÄRUNG 4

1. EINLEITUNG 5

1.1 PROBLEMSTELLUNG 5

1.2 ZIELSETZUNG 6

1.3 AUFBAU DER ARBEIT 6

2. DEFINITIONEN UND CHARAKTERISTIKA VON BIG DATA 6

2.1 HISTORISCHE ENTWICKLUNG UND AKTUELLE ANSÄTZE 7

2.2 ANWENDUNGSFÄLLE 9

3. HERAUSFORDERUNGEN IM UMGANG MIT BIG DATA 10

3.1 DATENSCHUTZ UND DATENSICHERHEIT 10

3.2 TECHNISCHE UND ORGANISATORISCHE HERAUSFORDERUNGEN 11

4. NICHT-FUNKTIONALE ANFORDERUNGEN, BIG DATA ARCHITEKTUREN UND TECHNOLOGIEN 13

4.1 NICHT-FUNKTIONALE ANFORDERUNGEN AN EINE BIG DATA ARCHITEKTUR 14

4.2 FÜNF ANSÄTZE ZUM UMGANG MIT BIG DATA 15

5. DAS CAP-THEOREM IN BIG DATA 18

5.1 THEORETISCHE GRUNDLAGEN 18

6. DATAPIPELINE UND TECHNOLOGIEN 21

6.1 DATAPIPELINE-ANSÄTZE IN SERVERLESS COMPUTING 21

7. DISKUSSION 23

8. FAZIT UND AUSBLICK 25

LITERATURVERZEICHNIS 27

ANHANG – VERSICHERUNG 31

Abbildungsverzeichnis

Abbildung 1 Wichtigste nicht-funktionale Anforderungen für Big Data Systeme (Visualisierung der im Text angegebenen Prozentzahlen, eigene Darstellung. Quelle: Rahman & Reza, 2020, S. 028) 14
Abbildung 2 Lambda-Architektur (Quelle: Berle (2017)) 16
Abbildung 3 Mögliche Optionen des CAP-Theorems (Kaufmann & Meier, 2023, S. 163, direkt entnommen). 19
Abbildung 4 Taxonomie der Datenpipeline-Ansätze im serverlosen Computing (erstellt in Anlehnung an Shojaee & Ghobaei-Arani, 2024, S. 12) 22
Abbildung 5 KI Data Pipeline (erstellt in Anlehnung an Blum, 2023, S. 256) 23

Gender-Erklärung

In dieser Hausarbeit wird aus Gründen der besseren Lesbarkeit die Sprachform des generischen Maskulinums verwendet. Es wird an dieser Stelle darauf hingewiesen, das Weibliche und anderweitige Geschlechteridentitäten dabei ausdrücklich mitgemeint werden, soweit es für die Aussage erforderlich ist.

Integration des Beispiels

Das Beispiel des Unternehmens ***Game your Life*** ist frei erfunden, wobei die Problematik im Rahmen von Kundengesprächen identifiziert wurde. Das Beispielunternehmen wird für die Fallstudie verwendet, um die unterschiedlichen Aspekte direkt beispielhaft darzulegen.

1. Einleitung

„***The world's most valuable resource is no longer oil, but data*** "(Economist, 2017)
Das weltweite Datenvolumen stieg von 4,4 Zettabyte im Jahr 2013 auf 64 Zettabyte im Jahr 2020, und laut einer Prognose der International Data Corporation (IDC) aus dem Jahr 2022 wird es bis 2026 voraussichtlich rund 221 Zettabyte erreichen (Vorndran, 2024, S. 62).

1.1 Problemstellung

„*Wir sind als mittelständiges Unternehmen nicht in der Lage Big Data zu händeln. Das macht für uns auch keinen Sinn, die Erkenntnisse, die wir daraus gewinnen helfen uns nicht weiter. Wir haben die Kompetenz nicht und können diese auch nicht aufbauen. Wir machen alles wie bisher, ein paar Nutzerumfragen müssen reichen*". Das war die Aussage des Inhabergeführten Geschäftsführers des Unternehmens ***Game your Life***, der in einer heftigen Diskussion mit der Mitarbeiterin gefangen war. Die Mitarbeiterin Tabea Inno ist verantwortlich für das innovative Lernformat, bei dem die Teilnehmenden überfachliche Kompetenzen via Planspiel vermittelt bekommen. Die Teilnehmenden spielen ein online Game, das real-world Probleme in einen game based Learning Ansatz transferiert, und so eine horizontale und vertikale Ich-Entwicklung[1] der Teilnehmenden unterstützt. Das Game wird in einer App (native oder Webapp) gespielt. Während die Teilnehmenden das Game spielen, werden pro Gamer über 1000 Datenpunkte pro Tag erhoben. Die Mitarbeiterin Tabea möchte genau diese Datenpunkte auswerten, Korrelationen aufdecken und auch Maschine Learning für den Aufbau eines adaptiven und prozeduralen Games[2] nutzen, um so Zeit bezüglich der Gamekonzeptionierung einzusparen und Teilnehmende so lange wie möglich im Spiel zu halten, da jede Gameminute Geld für das Unternehmen einbringt.
Die dargelegte Situation ist eine Sachlage wie sie derzeit in eher kleinen und mittelständigen Unternehmen anzufinden ist. Daten könnten erhoben werden, aber es fehlen die Grundlagen, um dies strategisch sinnvoll zu vollziehen, oder wenn Daten erhoben werden, werden diese nicht in Entscheidungsfindungsprozessen der

[1] Bei der horizontalen Ich-Entwicklung bauen die Teilnehmenden ihr Expertenwissen aus, bei vertikalen Ich-Entwicklung handelt es sich um die „Transformation des Weltverständnisses in Richtung höherer Komplexität und Reflexionsfähigkeit, was auch Auswirkung auf die emotionale Intelligenz hat" (Steps & Milestone, n.d.).

[2] Bei einem prozeduralen Game baut sich das Game entsprechend der Spielehistorie, während des Spiels selbständig auf, so dass ein nie endendes individualisiertes adaptives Game für die Teilnehmenden existiert.

Unternehmen einbezogen. In größeren Unternehmen sind Datenspeicherung und Datenanalyse verbreiteter und haben einen Einfluss auf strategische und operative Prozesse, allerdings zählen zu den Herausforderungen von Big Data der fehlende geschäftliche Kontext, organisationale Hürden, eine komplexe Big-Data-Architektur, das Data Engineering, der rasche technologische Wandel sowie ein Mangel an Kompetenzen (Ataei & Litchfield, 2022, S. 113789).

1.2 Zielsetzung

In dieser Fallstudie werden grundsätzliche Herausforderungen erörtert, die bei der Arbeit mit Big Data entstehen. Darüber hinaus werden wesentliche nicht-funktionale Anforderungen an Big Data-Architektur erläutert und fünf Ansätze im Umgang mit Big Data skizziert. Das CAP-Theorem wird mit praktischen Beispielen für alle drei Kategorien AP, CA und CP dargelegt. Zusätzlich wird das Thema der Datapipeline und relevanter Technologien aufgezeigt.

1.3 Aufbau der Arbeit

Die Fallstudie folgt einer klaren Struktur, um die Lesenden durch die komplexen Aspekte von Big Data in einem schrittweisen Prozess zu führen. Es wird zunächst ein Überblick über die Definitionen und Charakteristika von Big Data gegeben. Das nachfolgende dritte Kapitel spricht die verschiedenen Herausforderungen an, wobei der Datenschutz und die Datensicherheit eine entscheidende Rolle spielen. Das vierte Kapitel thematisiert die verschiedenen Architekturen und Technologien, die zur Verarbeitung von großen Datenmengen verwendet werden. Im fünften Kapitel wird das CAP-Theorem erläutert. Das sechste Kapitel schließlich beschreibt die verschiedenen Schritte einer Datapipeline und führt relevante Technologien auf. Die Arbeit wird mit einer Reflektion der Ergebnisse und eines Ausblicks auf zukünftige Entwicklungen im Bereich der Big Data abgeschlossen.

2. Definitionen und Charakteristika von Big Data

Im nächsten Kapitel werden zunächst die grundlegenden Definitionen und charakteristischen Merkmale von Big Data betrachtet. Dieses Kapitel schafft eine Basis für das Verständnis der technischen und organisatorischen Herausforderungen, die in den folgenden Abschnitten der Fallstudie vertieft werden.

2.1 Historische Entwicklung und aktuelle Ansätze

Die technologische Entwicklung im Bereich Big Data ist gekennzeichnet durch den Übergang von traditionellen Datenbankmanagementsystemen hin zu neuen Systemen, die speziell für die Verarbeitung großer und variabler Datenmengen[3] entwickelt wurden. Ein bedeutender Meilenstein in dieser Entwicklung war die Einführung von Hadoop. Ursprünglich entwickelt, um die Einschränkungen herkömmlicher Datenbanksysteme zu überwinden, bot Hadoop ein skalierbares und flexibles Framework für die verteilte Speicherung und Analyse großer Datenmengen (Kambatla et al., 2014). Hadoop besteht aus dem Hadoop Distributed File System (HDFS) für die Speicherung und dem MapReduce-Programmiermodell für die Datenverarbeitung. Es wurde durch Projekte wie YARN[4], das Ressourcenmanagement übernimmt, und Tools wie Apache Hive und Pig ergänzt, die die Datenverarbeitung vereinfachen und den Umgang mit großen Datenmengen effizienter gestalten. Heutzutage wird Hadoop oft in Kombination mit moderneren Verarbeitungstools wie Apache Spark eingesetzt, die zusätzliche Flexibilität und Leistung bieten.

Parallel zur Entwicklung von Hadoop wurden NoSQL-Datenbanken entwickelt, um die Herausforderungen bei der Speicherung und Verarbeitung von semi-strukturierten und unstrukturierten Daten zu bewältigen. Diese Datenbanken zeichnen sich durch ihre verteilte Architektur und ihre Fähigkeit zur horizontalen Skalierung[5] aus, was sie äußerst anpassungsfähig macht. Es gibt verschiedene Arten von NoSQL-Datenbanken, darunter dokumentenorientierte, spaltenbasierte, key-value-basierte und Graphdatenbanken, die unterschiedliche Anwendungsfälle unterstützen (Fasel, & Meier, 2016, S. 12).

---

[3] Variable Datenmengen im Zusammenhang mit Big Data bedeuten, das mit vielfältigen unterschiedlichen Datenquellen, -typen und -formaten gearbeitet wird, die jeweils unterschiedliche Anforderungen an die Speicherung, Analyse und Verarbeitung mit sich bringen.

[4] Yet Another Resource Negotiator: YARN wurde entwickelt, um das Ressourcenmanagement und die Aufgabenverteilung in einem Hadoop-Cluster zu optimieren. Es bildet die Grundlage für die Verarbeitung großer Datenmengen und ermöglicht die parallele Ausführung von Anwendungen in verteilten Umgebungen.

[5] Bei der horizontalen Skalierung wird die Last auf mehrere Maschinen verteilt, bei der vertikalen Skalierung wird die Hardware eines einzelnen Servers verbessert (Grasl, 2022, S. 19).

Angesichts der exponentiellen Zunahme an Daten entwickelten sich diese Technologien kontinuierlich weiter, um den Anforderungen der '3 V's'[6] – Volume (Datenmenge), Velocity (Verarbeitungsgeschwindigkeit), Variety (Datenvielfalt), gerecht zu werden (Katal, Wazid, & Goudar, 2013). Tam und Van Halderen erweitern die „Big V's" um Veracity (Wahrhaftigkeit) und Vulnerability (Verletzlichkeit), wobei sich die Wahrhaftigkeit auf die Qualität und Genauigkeit der Daten bezieht und die Verletzlichkeit auf Sicherheitsrisiken und ethische Fragen verweist. Dazu gehört auch die Wahrung der Vertraulichkeit und die Einhaltung von Datenschutzrichtlinien (Tam & Van Halderen, 2020). Darüber hinaus fügen Experten ein weiteres „V" für Value hinzu, das den Fokus auf die Steigerung des Unternehmenswertes durch Big Data legt (Fasel & Meier, 2016, S. 6).

Moderne Technologien wie Apache Spark ergänzen das Hadoop-Ökosystem durch ihre In-Memory-Verarbeitung, die im Vergleich zu MapReduce eine schnellere Datenanalyse ermöglicht und für viele Anwendungen heute die bevorzugte Wahl ist. Spark bietet außerdem eine benutzerfreundlichere API und unterstützt sowohl Batch- als auch Streaming-Verarbeitung, was es zu einer vielseitigen Lösung für Big Data macht (Fasel & Meier, 2016, S. 308).

Der Trend hin zur Echtzeitdatenverarbeitung führte zu einem Paradigmenwechsel, der die batchorientierte Datenverarbeitung ergänzte, insbesondere in Bereichen, die schnelle Reaktionen erfordern. Technologien wie Apache Storm und Kafka haben die Datenverarbeitung revolutioniert, indem sie die Verarbeitung von Datenströmen nahezu in Echtzeit ermöglichten. Diese Entwicklung ist besonders in der Interaktionsanalyse und im Event-Tracking auf Social-Media-Plattformen sowie in der Steuerung von IoT-Geräten von entscheidender Bedeutung (Marjani et al., 2017). Kafka dient für moderne Dateninfrastrukturen, indem es große Datenmengen effizient und nahezu verzögerungsfrei streamt. In Kombination mit Verarbeitungs-Frameworks wie Apache Flink wird die Echtzeitverarbeitung in Big Data-Analysen weiter optimiert, was sowohl die Effizienz als auch die Flexibilität dieser Infrastrukturen erheblich erhöht (Angbera & Chan, 2022).

[6] „Das Modell der „drei V" geht ursprünglich auf einen Beitrag des Analysten Doug Laney aus dem Jahr 2001 zurück, in dem dieser die Herausforderungen des (damaligen) Datenmanagements in drei Dimensionen beschrieb" (Hornung & Herfurth, 2018, S. 151).

Mit dem Fortschritt im Bereich des maschinellen Lernens hat sich auch das Verständnis für deren Implementierung in Big Data-Architekturen vertieft. Machine-Learning-Algorithmen leisten einen wesentlichen Beitrag zur Erkennung von Mustern und zur Vorhersage von Trends in großen Datenmengen. Der frühzeitige Einbezug dieser Algorithmen in die Datenverarbeitungsarchitektur ermöglicht es, dynamische Datenströme in Echtzeit zu verarbeiten und schnellere sowie präzisere Geschäftsentscheidungen zu treffen (Azarmi, 2016). Innovative datenbasierte Lösungen, die auf maschinellem Lernen basieren, erfordern jedoch ein Umdenken bezüglich der Infrastruktur und stellen neue Anforderungen an die Performanz und Skalierbarkeit von Datenverarbeitungssystemen (Dhaouadi et al., 2024).

Die zunehmenden Datenschutzbedenken haben auch die Notwendigkeit einer adaptiven Sicherheitsarchitektur von Big Data-Plattformen betont. Mit der steigenden Menge an sensiblen Daten sind Unternehmen dazu angehalten, fortlaufend Datenschutz und Datensicherheit zu gewährleisten. Am 21. Mai 2024 wurde der AI Act mit einem einheitlichen Rahmen für den Einsatz von künstlicher Intelligenz in der Europäischen Union verabschiedet (Bundesregierung, 2024).

Die Betrachtung der historischen Entwicklung und der aktuellen Ansätze verdeutlicht die Tragweite, die Big Data in der heutigen Gesellschaft einnimmt. Es zeigt sich, dass die Dynamik und die Komplexität von Big Data nicht nur eine technologische, sondern auch eine kulturelle und organisatorische Herausforderung darstellen, die stetige Anpassungen und Weiterentwicklungen erfordert. Hornung und Herfurth postulieren, dass es zu eingeschränkt sei, Big Data als rein technische Herausforderung zu begreifen, es sich vielmehr um eine technisch-soziale Innovationen handle, „die nicht allein unter technischen Paradigmen der Machbarkeit gestaltet werden können“ (Hornung & Herfurth, 2018). Das folgende Kapitel beleuchtet die Herausforderungen im Umgang mit Big Data.

2.2 Anwendungsfälle

Die Möglichkeiten, Big Data einzusetzen, sind nicht auf einen bestimmten Unternehmensbereich beschränkt. Zwar liegen derzeit die Schwerpunkte häufig auf den Bereichen Marketing, Vertrieb, Finanzen und Controlling, jedoch kann Big Data auch in Bereichen wie IT, Logistik und Personalmanagement zum Einsatz kommen, um beispielsweise den Personaleinsatz gezielter zu steuern und damit Kosten zu senken (Gluchowski, 2024, S. 32). Zudem wird Big Data branchenübergreifend genutzt, etwa im Handel, der Automobilindustrie, Landwirtschaft, dem Gesundheitswesen sowie im

öffentlichen Sektor, um auf Grundlage von Verkehrsdaten maßgeschneiderte Angebote zu erstellen (Gluchowski, 2024, S. 32). Echtzeitdaten spielen auch eine bedeutende Rolle im Katastrophenschutz, indem sie helfen, Einsatzkräfte effizienter zu koordinieren. Auch in der öffentlichen Verwaltung wird Big Data eingesetzt, etwa zur Aufdeckung von Steuerhinterziehungen oder Sozialmissbrauch durch Datenmustererkennung. Insgesamt sind die Einsatzmöglichkeiten von Big Data nahezu unbegrenzt (Gluchowski, 2024, S. 33).

3. Herausforderungen im Umgang mit Big Data

Das Kapitel behandelt die zentralen Herausforderungen im Umgang mit Big Data, wobei insbesondere Datenschutz und Datensicherheit sowie technische und organisatorische Hindernisse im Fokus stehen. Denn „Big Data was not conceived initially as a secure environment and therefore, the main security problems are related to the specific architecture of Big Data itself“ (Moreno et al., 2018, S. 1). Ziel ist es, die wesentlichen Schwierigkeiten zu identifizieren, die Unternehmen bei der Umsetzung von Big Data-Initiativen bewältigen müssen.

3.1 Datenschutz und Datensicherheit

Die Verarbeitung und Analyse von Big Data birgt immense Potenziale für die Wirtschaft, steht jedoch gleichzeitig im Kontext von bedeutenden Herausforderungen bezüglich Datenschutzes und Datensicherheit. In Europa setzt die Datenschutz-Grundverordnung (DSGVO) einen stringenten rechtlichen Rahmen für die Verarbeitung personenbezogener Daten. Dieser regulatorische Kontext erfordert die Implementierung von Prozessen, welche die Einhaltung von Datenschutzprinzipien wie Rechtmäßigkeit, Transparenz und Zweckbindung sicherstellen. „Es dürfen nur so viele Daten verarbeitet werden, wie es für das Erreichen des Zwecks notwendig ist (Datenminimierung), und Daten dürfen nur so lange gespeichert werden, wie es für das Erreichen des Zwecks notwendig ist (Speicherbegrenzung)“ (Hornung, 2016, S. 1). Technisch gesehen stellen Verschlüsselungstechniken und Datenminimierungsprinzipien zwar effektive Mittel dar, doch in Hinblick auf Big Data ist die Anonymisierung eine komplexe Herausforderung, häufig bedingt durch die Verknüpfung verschiedenster Datensätze, die Rückschlüsse auf Individuen zulassen können. „Noch fundamentaler erscheinen die Probleme der Datenminimierung, Zweckbindung und Sprecherbegrenzung. In der Logik vieler Visionen von Big Data liegt es, möglichst viel Daten zu neuen, oft unbekannten Zwecken zu erheben, zu analysieren und aufzubewahren, um sie in der Zukunft mit weiteren Daten kombinieren zu können“ (Hornung, 2016, S. 2). Am 1. August 2024 ist die KI-Verordnung (KI-VO) verabschiedet worden die auch die Grundsätze der DSGVO wie Rechtmäßigkeit, Zweckbindung, Transparenz, Datenminimierung und Richtigkeit beim

Einsatz von KI beinhaltet (Industrie- und Handelskammer für München und Oberbayern, n.d.).

Das Spannungsfeld zwischen dem unternehmerischen Verlangen nach umfassender Datennutzung und den gesetzlich verankerten Datenschutzbestimmungen resultiert in komplexen Entscheidungsprozessen, welche den Einsatz von Big Data-Lösungen maßgeblich beeinflussen können.

Neben Datenschutz besteht eine weitere signifikante Hürde in der Sicherung von Big Data gegenüber Cyberrisiken, die aufgrund der enormen Datenmengen und Vernetzung entstehen. Laut Bitkom sehen sich zwei Drittel der Unternehmen durch Cyberattacken in ihrer Existenz bedroht (Bitkom, 2024). Unternehmen müssen proaktive Sicherheitsmaßnahmen ergreifen, um sich vor internen und externen Bedrohungen zu schützen und somit die Integrität ihrer Big Data-Systeme aufrechtzuerhalten. Das Erreichen von Sicherheitsstandards und das Erwirken von Zertifikationen sind dabei nicht nur für die Risikominimierung von Bedeutung, sondern spielen auch in der Vertrauensbildung mit Kunden und anderen Stakeholdern eine tragende Rolle.

Die Akzeptanz und Nutzung von Big Data in Unternehmen wird maßgeblich durch Datenschutz und Datensicherheit beeinflusst. Es gilt, ethische Aspekte in die Analyse und Nutzung von Daten einzubeziehen, insbesondere in Hinblick auf die informierte Einwilligung und die Transparenz der Datenverarbeitung. Durch konsequente datenschutzkonforme Lösungen können Unternehmen das Vertrauen ihrer Kunden drastisch steigern, was wiederum die Einsatzmöglichkeiten und Akzeptanz von Big-Data-Lösungen stärkt.

Abschließend repräsentiert die korrekte Handhabung von Datenschutz und Datensicherheit sowohl eine rechtliche Notwendigkeit als auch einen kritischen Faktor für den Erfolg von Big Data in unternehmerischen Kontexten.

3.2 Technische und organisatorische Herausforderungen

Es gibt heute eine Vielzahl an Technologien, Konzepten und Systemen, die im Zusammenhang mit Big Data stehen. Diese Technologien lassen sich in mehrere Kategorien unterteilen: Erzeugung, Verarbeitung und Integration von Daten, Speicherung sowie deren Analyse und Bereitstellung (Vorndran, 2024, S. 68). „Die Grundlage für Anwendungen zur Beschaffung, Speicherung, Aufbereitung und Bereitstellung von Big Data bilden häufig verbreitete Ansätze oder Technologien wie

Data Lakes[7], Not only Structured Query Language-Datenbanken (NoSQL-Datenbanken), oder In-Memory-Technologie“ (Vorndran, 2024, S. 69).

Zicari identifiziert folgende Herausforderungen im Zusammenhang mit Big Data:
Datenmengen und Vielfalt: Die enorme Datenmenge und die Vielfalt der Datenquellen und -formate stellen Unternehmen vor große Schwierigkeiten (Zicari, 2015). Skalierungsprobleme in Big Data-Architekturen sind ein zentrales Thema im Kontext der Verarbeitung exponentiell wachsender Datenmengen. Hadoop, als etablierte Lösung für Big Data-Herausforderungen, stößt in Bezug auf Skalierbarkeit und Verarbeitungsgeschwindigkeit zunehmend an Grenzen (Kambatla et al., 2014). Das System muss in der Lage sein, mit der Dynamik von Volume, Velocity und Variety Schritt zu halten. Doch die verteilte Datenverarbeitung und das Speichermanagement verlangen nach fortschrittlichen Strategien, um Daten effizient zu teilen und zu parallelisieren. Die Anpassung an Datenwachstum, ohne dabei die Leistung zu beeinträchtigen, stellt Entwickler und Architekten vor komplexe Herausforderungen (Oussous et al., 2018).

Prozesse: Die Auswahl geeigneter Analysemodelle und die Integration von Daten aus verschiedenen Quellen können zeitaufwendig sein (Zicari, 2015). Gleichwohl „die eigentliche Herausforderung nicht die Arbeit der einzelnen Maschine selbst, sondern die Arbeit sinnvoll zu verteilen und zu koordinieren, ohne dass die Koordinationsinstanz selbst zum Engpass würde“ ist (Grasl, 2022, S. 19). Die richtige Verarbeitung, Visualisierung und Interpretation der Daten sind komplexe und oft zeitintensive Prozesse. „Eine der Herausforderungen in der Praxis ist, Maschinen nicht zu „übertrainieren“: In solchen Fällen können Maschinen die Trainingsdaten perfekt vorhersagen, machen dann aber bei neuen Daten Fehler“ (Grasl, 2022, S. 41).

Management: Datenschutz und Sicherheit sind große Herausforderungen. Sensible Daten, insbesondere personenbezogene, müssen rechtlich und ethisch korrekt behandelt werden. Der Schutz der Daten und die Sicherstellung der korrekten Nutzung sind entscheidend (Zicari, 2015).

Hinzu kommt die **Gefahr der Cyberbedrohungen** und der nicht zu unterschätzende **Mangel an qualifizierten Fachleuten** in diesem Bereich. Die Implementierung und

[7] Data Lakes sind Datenspeicherlösungen, die große Mengen an Rohdaten für spätere Verwendungszwecke im Originalformat speichern (Vorndran, 2024, S. 69).

Verwaltung von Big Data-Technologien und die dafür erforderlichen Kompetenzen betonen eine wachsende Diskrepanz zwischen aktuellen Anforderungen und verfügbarem Fachwissen. Die Weiterbildung von Mitarbeitenden ist hierbei essentiell, um mit den Entwicklungen Schritt halten zu können. Besonders Maschine-Learning-Techniken und komplexe Datenanalysen erfordern umfassendes Know-how, welches in vielen Unternehmen noch nicht ausreichend vorhanden ist. Die Integration von Big Data-Systemen in bestehende IT-Infrastrukturen stellt häufig eine technische und organisatorische Hürde dar. Legacy-Systeme und Datensilos erschweren den nahtlosen Datenaustausch und behindern somit die ganzheitliche Nutzung von Big Data. Insbesondere die Kompatibilität und Interoperabilität von neuen Systemen mit vorhandenen Strukturen erfordern eine sorgfältige Planung und Implementierung, um die Datenintegration zu gewährleisten (Ataei & Litchfield, 2022). Fachkräfte müssen in interdisziplinären Kompetenzbereichen geschult werden, um den vielschichtigen Big Data-Problemstellungen gerecht zu werden (Azarmi, 2016). „Einfach ausgedrückt, werden Organisationen dazu verurteilt, bei ihren Projekten für große Datenmengen zu scheitern oder stecken zu bleiben. (…) Ich sehe oft die zweiten oder dritten Auswirkungen eines fehlenden Teams“ (Anderson, 2024, S. 20). Für Anderson ist für den Erfolg von Big Data entscheidend, ob das verantwortliche Team aus Repräsentanten aus Data Science, Data Engineering und Operations vertreten ist.

Mit Blick auf zukünftige Entwicklungen bleibt festzustellen, dass die genannten Herausforderungen einer kontinuierlichen und kritischen Betrachtung unterliegen müssen, um das Potenzial von Big Data vollends auszuschöpfen und nachhaltige Lösungen zu etablieren.

4. Nicht-funktionale Anforderungen, Big Data Architekturen und Technologien

In diesem Kapitel werden nicht-funktionalen Anforderungen und zentrale Big Data Architekturen und Technologien[8] betrachtet, die maßgeblich zur Bewältigung der Herausforderungen im Umgang mit großen Datenmengen beitragen. Es werden fünf Ansätze erläutert, die strukturelle und technologische Lösungen bieten, um die Verarbeitung, Speicherung und Analyse großer Datenmengen effizient zu gestalten.

[8] Die Architektur beschreibt den strukturellen Plan, oder das Konzept, wie das Big Data-System funktioniert. Bei der Technologie handelt es sich um die konkreten Werkzeuge, die verwendet werden, um die Architektur zu realisieren.

4.1 Nicht-Funktionale Anforderungen an eine Big Data Architektur

Nicht-Funktionale Anforderungen beschreiben Qualitätsmerkmale die sich auf die Leistung, Zuverlässigkeit, Nutzbarkeit und Architektur beziehen und keinen direkten Einfluss auf die Funktionalität haben. Rahman und Reza identifizieren über 40 Qualitätsmerkmale, die mit Big Data Systemen in Verbindung stehen und ordnen diese in 8 Merkmale des ISO/IEC 25010:2011-Modells[9] ein (Rahman & Reza, 2020, S. 025).

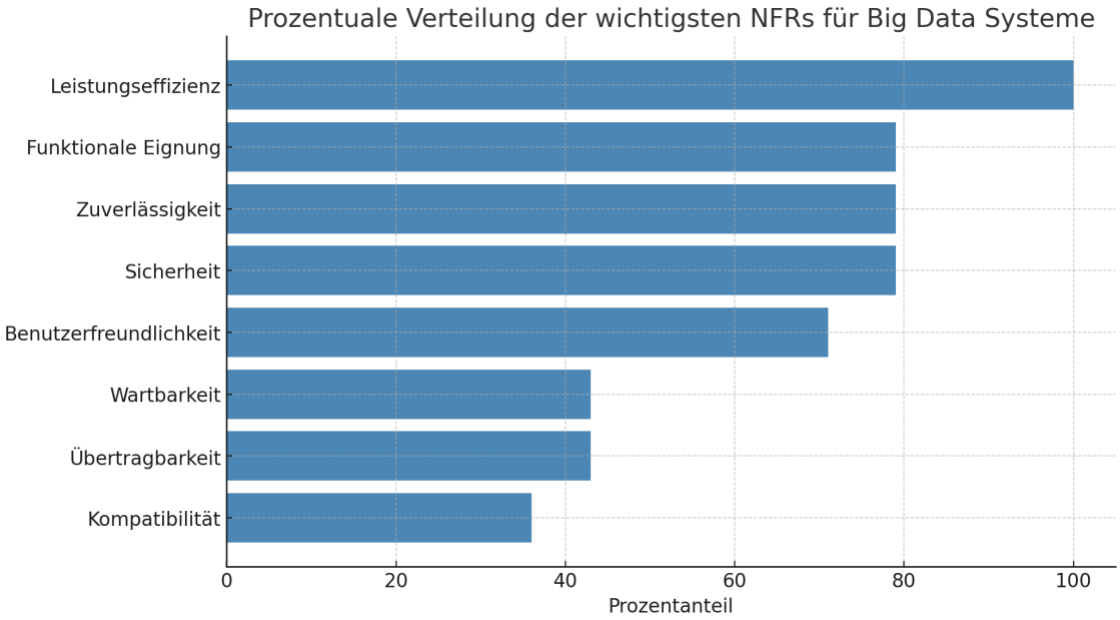

Abbildung 1 Wichtigste nicht-funktionale Anforderungen für Big Data Systeme (Visualisierung der im Text angegebenen Prozentzahlen, eigene Darstellung. Quelle: Rahman & Reza, 2020, S. 028)

Dabei bewertet die Leistungseffizienz die Software in Bezug auf Ressourcenverbrauch und Geschwindigkeit. Es werden die Reaktionszeiten, Verarbeitungszeiten und die Ressourcennutzung betrachtet. Die Funktionale Eignung misst, wie gut die Software die geforderten Funktionen bereitstellt und ob sie den gegebenen Anforderungen entspricht, das bedeutet die Vollständigkeit, Korrektheit und Angemessenheit. Bei der Zuverlässigkeit wird die Fähigkeit der Software gemessen unter bestimmten Bedingungen die Aufgaben stabil und zuverlässig zu erfüllen. Dazu gehört auch der Aspekt der Reife, Verfügbarkeit, Fehlertoleranz und Wiederherstellbarkeit nach einem

[9] Das ISO/IEC 25010:2011-Modell ist ein internationales Qualitätsmodell für Software und IT-Systeme, das von der International Organization for Standardization (ISO) und der International Electrotechnical Commission (IEC) entwickelt wurde. Es dient als Standard für die Bewertung und Sicherstellung der Softwarequalität und definiert spezifische Qualitätsmerkmale, die Softwareprodukte erfüllen sollten.

Ausfall. Bei der Sicherheit wird der Aspekt des Schutzes gegen unbefugte Zugriffe, Datenverlust und Manipulation betrachtet. Die Benutzerfreundlichkeit bezieht sich auf die Usability und Erlernbarkeit der Software für die Benutzer. Dabei betrachtet die Usability, ob die Software intuitiv, effizient und angenehm zu bedienen ist. Die Wartbarkeit umfasst Nachvollziehbarkeit von Änderungen und eine einfache Fehlerbehebung. Wie gut sich die Software in unterschiedliche Umgebungen und verschiedene Plattformen integrieren, lässt bezeichnet die Übertragbarkeit. Bei der Kompatibilität wird sichergestellt, dass die Software in einer Umgebung mit verschiedenen Systemen und Anwendungen zusammenarbeiten kann, was auch die Interoperabilität und Koexistenz mit anderen Programmen beinhaltet (Britton, 2021).

4.2 Fünf Ansätze zum Umgang mit Big Data

Bei Referenzarchitekturen handelt es sich um standardisierte Architekturvorlagen, die bewährte Strukturen, Komponenten und Muster bieten, die als Leitfaden und Best Practice dienen. Die Entwicklung und Popularisierung von Referenzarchitekturen für Big Data haben einen maßgeblichen Einfluss auf die Gestaltung heutiger Datenverarbeitungssysteme. Die **Lambda-Architektur** hat sich als verbreitetes Modell etabliert, das die gleichzeitige Verarbeitung von Batch- und Echtzeitdaten ermöglicht. „Lambda architecture is the most commonly discussed and studied“ (Ataei & Litchfield, 2022, S. 113798). Dadurch bietet sie eine Systemflexibilität und Agilität, insbesondere in Szenarien, in denen schnelle sowie tiefere Analysen erforderlich sind. Ataei und Litchfield (2022) unterstreichen die breite Adoption dieser Architektur in verschiedenen Branchen, wobei die Lambda-Architektur durch ihre konzeptionelle Dreiteilung in Batch-, Speed- und Serving-Layer sowohl die Langzeitanalyse als auch die unmittelbare Datenverarbeitung in Echtzeit unterstützt. Dennoch ergeben sich Herausforderungen bei der Implementierung, insbesondere was die Komplexität des Betriebs und das Management der parallellaufenden Systeme betrifft. Verwendete Technologien sind Hadoop für Batch-Verarbeitung und Apache Kafka oder Apache Storm für Echtzeit-Datenströme (Oussous, Benjelloun, Ait Lahcen, & Belfkih, 2018). Tabea Inno, die für das innovative Lernformat bei Game your Life verantwortlich ist, könnte die Lambda-Architektur nutzen, um die gesammelten Datenpunkte aus dem Game-basierten Lernansatz zu verarbeiten und auszuwerten. Ihr Ziel ist es, Korrelationen aufzudecken und maschinelles Lernen zu integrieren, um die Game-Entwicklung zu automatisieren und die Teilnehmenden so lange wie möglich im Spiel zu halten. Mittels der Lambda-Architektur integrierten Batch-Verarbeitung könnten langfristige Analysen und Mustererkennungen vollzogen werden. Dafür würden die historischen Daten der Gamer (über 1000 Datenpunkte pro Tag und Gamer) analysiert um Muster wie häufigsten Spielabbrüche, bevorzugte Spielstrategien oder Lernfortschritte zu identifizieren, um das

Spieldesign und die Learner Journey zu verbessern. Auch Schwachstellen könnten identifiziert werden. Die Echtzeit-Verarbeitung der Speed-Layer ist für die sofortige Feedbackgabe und direkte Anpassungen von Vorteil. Bezogen auf das Beispiel könnte das System sofort Anpassungen vornehmen, wenn ein Gamer an einer bestimmten Stelle Schwierigkeiten hat, z.B. durch die Bereitstellung von Hilfestellungen oder das Anpassen des Schwierigkeitsgrades der Challenges. Der Serving-Layer könnte die Ergebnisse aus den beiden Schichten (Speed und Batch) so kombinieren, um daraus eine individualisierte und einzigartige adaptive Learning Experience zu generieren. Dadurch würden die aktuellen Spieldaten mit den langfristigen Daten kombiniert. Durch die Verarbeitung von Echtzeitdaten und die Analyse historischer Daten könnten sowohl das Spiel als auch die Learning Experience kontinuierlich verbessert werden, wodurch der Lernfortschritt optimiert würde und durch die positiveren Erfahrungen der Gamer auch die Spielzeit verlängert würde.

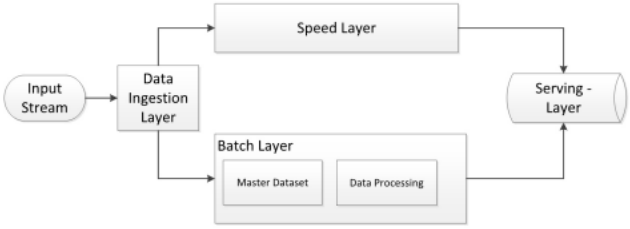

Abbildung 2 Lambda-Architektur (Quelle: Berle, 2017, direkt entnommen)

Die **Kappa-Architektur** ist eine vereinfachte Version der Lambda-Architektur, die sich hauptsächlich auf die Echtzeit- und nicht auf die Batch-Datenverarbeitung konzentriert. Verwendete Technologien sind zum Beispiel Apache Kafka als Datenstreaming-Plattform und Apache Flink oder Apache Spark Streaming für Echtzeit-Datenverarbeitung. Kappa-Architektur ist der ideale Fall für Anwendungen, in denen der Datenfluss die ganze Zeit in Echtzeit erfolgt.

Die **Cloud-native[10] Architektur** nutzt Cloud-Umgebungen für die flexible Skalierung und Speicherung großer Datenmengen. Die Microservice-Architektur ist ein Schlüsselmerkmal von Cloud-native Anwendungen, die es ermöglicht die Anwendungen aufzuteilen und die Microservices unabhängig voneinander zu entwickeln, zu aktualisieren und zu skalieren, was die Flexibilität und Agilität der Anwendung erhöht. Die Containerisierung sorgt dafür, dass eine reibungslose Bereitstellung und Skalierung

[10] Cloud-Native bezeichnet die Herangehensweise an die Softwareentwicklung und Bereitstellung von Anwendungen, die speziell für Cloud-Umgebungen konzipiert ist (Gecko, n.d.).

von Anwendungen, unabhängig von der zugrundeliegenden Infrastruktur ermöglicht wird. „Die automatische Verwaltung, Bereitstellung und Skalierung von Containern wird durch Orchestrierungstools wie Kubernetes ermöglicht“ (Gecko, n.d.). Automatisierungen insbesondere in den Bereichen Continuous Integration, Continuous Deployment ermöglichen eine schnellere und zuverlässigere Bereitstellung von Software. Die Skalierbarkeit wird dadurch gewährleistet, das Cloud-native Anwendungen sich automatisch an wechselnde Workloads anpassen können. Letztlich ist die Architektur so konzipiert, dass sie widerstandsfähig gegenüber Fehlern und Ausfällen ist (Gecko, n.d.).

Bei der **Datenvirtualisierung** wird auf verschiedene Datenquellen zugegriffen, ohne die Daten physisch zu replizieren. Der Zugriff auf die Daten erfolgt in Echtzeit, darüber hinaus können die Daten aus unterschiedlichen Quellen integriert werden. Dadurch entsteht ein geringerer Entwicklungsaufwand für Analysen und mehr Flexibilität im Vergleich zum klassischen Data Warehouse. Es besteht keine Notwendigkeit der Datenreplikation durch den Verbleib der Daten im originalen Quellsystem. Die Qualität und Integrität der Daten ist durch die Vermeidung der Datenreplikation erhöht, wohingegen die Kosten und Systemlast reduziert sind und gleichzeitig die Agilität der Analysemöglichkeiten bestehen bleiben. Zu beachten ist allerdings, das langsame Quellsysteme die Geschwindigkeit der virtuellen Datenintegration stark reduzieren können. Darüber hinaus kann das Fehlen einer Änderungshistorie der Daten und die mangelnde Homogenität des Datenmodells nachteilig sein (BigData Insider, n.d.).

Bei **In-Memory-Computing Technologien** erfolgt die Verarbeitung von Daten direkt im Hauptspeicher (RAM)[11], anstatt auf langsamen Festplattenspeichern. Dies führt zu erheblichen Geschwindigkeitsvorteilen bei der Datenanalyse. „In-Memory-Datenbanken – auch als Real-time-Datenbanken (RTDB) bezeichnet – gibt es seit den 1980er-Jahren. Sie wurden in Branchen eingesetzt, in denen schnelle Reaktionszeiten unverzichtbar sind, wie in der Telekommunikation, im Bankwesen, im Reiseverkehr und in der Spielebranche. Heute sind erschwingliche Speicher- und Systemhardware für alle Branchen und Anwendungen verfügbar, auch für ERP-Systeme“ (SAP, n.d.). Die Vorteile sind eine schnellere Verarbeitung, es können sowohl OLAP und OLTP

[11] Es werden nicht alle Daten im Arbeitsspeicher gehalten, der Teil der Informationen der im Arbeitsspeicher gehalten wird, beinhaltet die unternehmenskritischen und häufig abgefragten Daten, um diese schnell abrufen und ändern zu können (Hot Storage). Die daten, die seltener abgefragt oder geändert werden, werden auf der Festplatte gespeichert (Cold Storage) (SAP, n.d.)

unterstützt werden, „wodurch der Bedarf an redundanter Speicherung und Verzögerung zwischen den Datenübertragungen entfallen“ (SAP, n.d.).

Die gegenübergestellten Ansätze und Lösungen verdeutlichen die Bandbreite in der Gestaltung von Datenarchitekturen. In-Memory-Computing für Hochgeschwindigkeitsanalysen und -transaktionen im Speicher steht dabei im Kontrast zu Lösungen, die auf Massendatenverarbeitung ausgerichtet sind. Beide Ansätze reflektieren den signifikanten Einfluss, den Referenzarchitekturen auf die Standardisierung von Designprinzipien haben, und tragen zur Qualitätssteigerung und Kostenreduktion in Big Data Projekten bei (Ataei & Litchfield, 2022).

5. Das CAP-Theorem in Big Data

Das CAP[12]-Theorem ist ein grundlegendes Konzept in der Architektur verteilter Systeme und beschreibt die Herausforderung, wie Konsistenz, Verfügbarkeit und Partitionstoleranz in einem System balanciert werden können. Dieses Kapitel beleuchtet die theoretischen Grundlagen des CAP-Theorems und illustriert seine Implikationen anhand praxisbezogener Beispiele.

5.1 Theoretische Grundlagen

Zu unterscheiden sind relationale Datenbanksystem und webbasierte Anwendungen. Hierbei wird das ACID-Prinzip (Atomicity, Consitency, Isolation, Durability) und BASE-Prinzip (Basically Available, Soft State, Evenutally Consitent) bezüglich einer Transaktion unterschieden. So werden bei relationalen Datenbanksystemen die Transaktionen in der höchsten Isolationsstufe immer entsprechend dem ACID-Prinzip durchgeführt, wohingegen bei webbasierten Anwendungen die hohe Verfügbarkeit im Vordergrund steht, was beinhaltet, dass „replizierte Knoten zwischenzeitlich unterschiedliche Datenversionen halten und erst zeitlich verzögert aktualisiert werden“ (Kaufmann & Meier, 2023, S. 162). Im Jahr 2000 stellte Eric Brewer die Hypothese auf, dass die drei Eigenschaften Konsistenz, Verfügbarkeit und Ausfalltoleranz nicht gleichzeitig in einem verteilten System vollständig erreicht werden können (Kaufmann & Meier, 2023, S. 162). “Das CAP-Theorem (engl. “CAP theorem“) sagt aus, dass in einem massiv verteilten Datenhaltungssystem jeweils nur zwei Eigenschaften aus den drei der Konsistenz (C), Verfügbarkeit (A) und Ausfalltoleranz (P) garantiert werden können“ (Kaufmann & Meier, 2023, S. 163).

[12] CAP-Theorem (Consitency, Availability, Partition Tolerant) wird auch Brewers Theorem genannt, da es erstmals von Professor Eric A. Brewer in einem Vortrag über verteilte Datenverarbeitung im Jahr 2000 aufgestellt wurde (IBM, n.d.).

Die folgende Grafik zeigt die möglichen Optionen des CAP-Theorems.

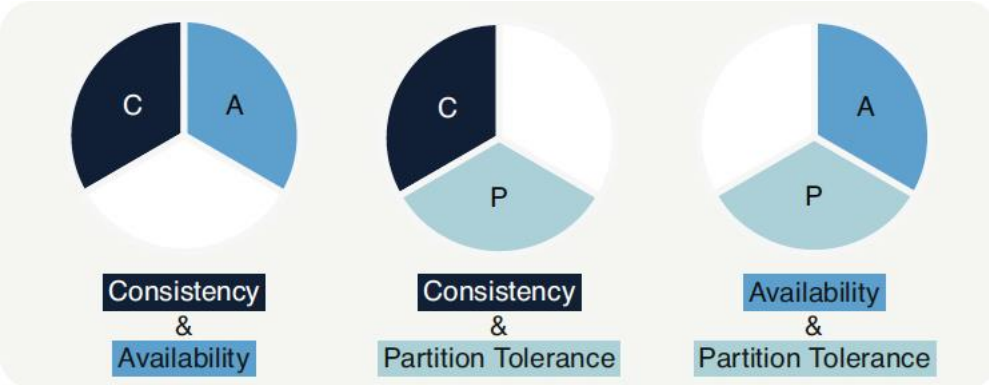

Abbildung 3 Mögliche Optionen des CAP-Theorems (Kaufmann & Meier, 2023, S. 163, direkt entnommen)

Ein Anwendungsbeispiel des CAP-Theorems findet sich an einem Börsenplatz, wo der Fokus auf Konsistenz und Verfügbarkeit gelegt wird, was bedeutet, dass CA bevorzugt wird. Hier werden relationale Datenbanksysteme verwendet, die den ACID-Prinzipien folgen. In einem landesweit verteilten Netzwerk von Geldautomaten bleibt Konsistenz entscheidend, während gleichzeitig Ausfalltoleranz gewährleistet sein muss. Leichte Verzögerungen in den Antwortzeiten werden in Kauf genommen, weshalb CP unterstützt wird. Diese Systeme setzen auf verteilte und replizierte relationale Datenbanken oder NoSQL-Systeme. Das Domain Name System (DNS) hingegen muss stets verfügbar und ausfalltolerant sein, da es für die Übersetzung von Webseitennamen in numerische IP-Adressen verantwortlich ist, die für die Kommunikation im TCP/IP-Protokoll erforderlich sind. Hier wird auf AP gesetzt, wobei NoSQL-Datenbanken bevorzugt werden, da eine weltweite Verfügbarkeit und Ausfalltoleranz mit einem relationalen Datenbanksystem schwer zu gewährleisten wäre (Kaufmann & Meier, 2023, S. 162).

Im Zusammenhang mit dem Unternehmen Game your Life können die drei Varianten des CAP-Theorems (CA, CP, AP) in unterschiedlichen Szenarien eingesetzt werden, je nachdem, welche Anforderungen und Priorität vorliegt.
Bezüglich des Lernfortschritts und -feedbacks ist CA zu bevorzugen, da das Spiel sicherstellen muss, das die Teilnehmenden denselben Lernfortschritt und dieselben Ergebnisse sehen, nachdem ein Modul abgeschlossen ist, dabei ist die Konsistenz im Fokus. Ein Spieler hat eine Challenge abgeschlossen, die Punktzahl wird aktualisiert, dann muss das System sicherstellen, dass die Änderung sofort und korrekt auf allen

Endgeräten (native Apps, Webapp) synchronisiert wird. Die Verfügbarkeit ist auch entscheidend, da die Gamer zu jeder Zeit auf ihre Game-Lerninhalte zugreifen können müssen, ohne auf Netzwerkprobleme oder Ausfallzeiten zu stoßen. Das würde dazu führen, dass sehr schnell die Akzeptanz für das Game sinkt und das Unternehmen wertvoll akquirierte Kunden verlieren würde. Ein relationales Datenbanksystem, das auf ACID basiert ist, ideal, um Datenintegrität und Verfügbarkeit sicherzustellen, um beispielsweise Lernergebnisse, Ranglisten und Zertifizierungen zu speichern und zu synchronisieren. Nachdem ein Gamer mit einer bestimmten Punktzahl eine Challengewoche abgeschlossen hat, muss das System garantieren, dass die Punktzahl nicht nur korrekt angezeigt wird, sondern auch überall verfügbar ist.

Wenn es um kritische Lernprozesse und den prozeduralen Lernfortschritt geht, bei denen die genaue Abfolge und Konsistenz der Lernschritte entscheidend ist, wäre die CP-Variante optimal. Das System muss sicherstellen, dass der Fortschritt eines Teilnehmenden konsistent bleibt, auch wenn es temporäre Netzwerkprobleme gibt. Eine leichte Verzögerung in der Verfügbarkeit ist hierbei akzeptabel, solange die Lernpfade und Inhalte korrekt synchronisiert bleiben. Dies ist besonders wichtig, wenn die Lernplattform maschinelles Lernen verwendet, um die Lerninhalte individuell an die Fortschritte der Teilnehmenden anzupassen. Ein verteiltes, konsistentes System wie eine NoSQL-Datenbank, das CP unterstützt, wäre hier ideal, um sicherzustellen, dass der Lernprozess nicht unterbrochen wird und der Fortschritt der Spieler korrekt erfasst wird.

Für Echtzeit-Interaktionen im Game-basierten Lernformat, etwa bei der Erstellung von Ranglisten oder der Gamifizierung von Lernfortschritten, ist AP die beste Wahl. Hier ist es entscheidend, dass das System verfügbar ist und Ausfalltoleranz bietet, um eine nahtlose Spielerfahrung zu gewährleisten, auch bei Netzwerkproblemen. Die Konsistenz ist in diesen Szenarien weniger wichtig, da es nicht entscheidend ist, dass alle Spieler in Echtzeit exakt identische Daten sehen. Es reicht aus, wenn die Ranglisten und Punktestände so schnell wie möglich aktualisiert werden, ohne den Spielablauf zu stören. Ein verteiltes NoSQL-System, das hohe Verfügbarkeit und Ausfalltoleranz bietet, kann sicherstellen, dass die Spieler auch bei Netzwerkunterbrechungen auf das Spiel zugreifen und ihre Fortschritte einsehen können. Das Thema der Extraktion von Daten aus unterschiedlichen Quellen zur effizienten Datenverarbeitung spielt bezogen auf das Beispiel eine Rolle, wenn die Daten nicht nur Gamespezifisch, sondern auch bezüglich der Customer Journey[13] ausgewertet werden sollen. Beispielsweise möchte die

[13] „Die Customer Journey beschreibt den gesamten Kaufprozess in der Vorkaufphase, der Kaufphase und der Nachkaufphase“ (Kuenen, n.d.).

Geschäftsführung einsehen, welche Schritte intern vollzogen werden, zwischen der ersten Kontaktaufnahme mit dem Kunden und der Kündigung der Kunden, nachdem sie Gamer waren. Datapipelines umfassen Prozesse die diesbezüglich Anwendung finden. Im folgenden Kapitel wird das Thema der Datapipelines mit den Technologien vorgestellt.

6. Datapipeline und Technologien

Datapipelines sind in modernen datengetriebenen Systemen von zentraler Bedeutung, da sie Verarbeitungs-, Analyse- und Entscheidungsprozesse ermöglichen. Es werden Daten aus unterschiedlichen Quellen extrahiert und in spezifische Systeme überführt und erlauben eine effiziente Datenverarbeitung. Eine Datenpipeline umfasst eine Reihe von Prozessen wie die Extraktion, Transformation und das Laden von Daten. Dabei werden Schritte wie Datenaufnahme, -anreicherung und -laden durchgeführt. In einer serverlosen Umgebung können Datapipelines durch Funktionen abgebildet werden, die durch Ereignisse, wie beispielsweise das Eintreffen neuer Daten, ausgelöst werden. Jede Funktion übernimmt eine spezifische Aufgabe und gibt die Daten an die nächste Funktion weiter, was Flexibilität und Skalierbarkeit bietet.

Eine Datenpipeline besteht aus aufeinanderfolgenden Phasen, die die Daten von der Extraktion bis zur Speicherung händeln. Es gibt zwei geläufige Paradigmen für Pipelines: ETL (Extrahieren, Transformieren, Laden) und ELT (Extrahieren, Laden, Transformieren). In einer serverlosen Architektur kann jeder Schritt im ETL-Prozess als durch Ereignisse oder Zeitpläne ausgelöste Funktion implementiert werden. CI/CD-Pipelines, die serverlose Technologien wie AWS-Lambda, Azure Functions oder Google Cloud Functions verwenden, beschleunigen die Anwendungen durch die Zerlegung in einzelne Teil-Funktionen. Diese Dienste bieten die Infrastruktur, um serverlose Funktionen effizient zu erstellen, bereitzustellen und zu verwalten. Beim Entwurf einer Datapipeline sind Faktoren wie Komplexität, Leistung, Kosten und Skalierbarkeit sorgfältig abzuwägen (Shojaee & Ghobaei Arani, 2024, S. 1).

6.1 Datapipeline-Ansätze in Serverless Computing

Es werden heuristic-based, machine learning-based und framework-based Ansätze unterschieden. Die heuristic-based Ansätze basieren auf vordefinierten Regeln oder Schwellwerten, um die Datenpipeline zu verwalten und zu orchestrieren. Dabei werden Regelbasierte, schwellenbasierte und Expertensysteme unterschieden. Regelbasierte Systeme verwenden vordefinierte Regeln oder Heuristiken zur Verwaltung und Orchestrierung der Datapipeline. Schwellenwertbasierte nutzen Schwellenwerte oder Trigger, um zu bestimmen, wann skaliert werden soll. Expertensystem erfassen

menschliches Fachwissen und Entscheidungsprozesse in Regeln oder Modellen zur Automatisierung bestimmter Aufgaben.

Machine-Learning-basierte Ansätze verwenden prädiktive, adaptive und unsupervised Learning-Techniken, um die Leistung zu optimieren. Beim prädiktiven Ansatz wird maschinelles Lernen verwendet, um die Leistung der Pipeline vorherzusagen und zu optimieren. Beim adaptiven Ansatz kommen maschinelle Lerntechniken zum Einsatz, um die Pipeline dynamisch an sich entwickelnde Datenmuster anzupassen. Beim unsupervised Learning werden Muster und Beziehungen in den Daten identifiziert. Framework-basierte Ansätze nutzen serverlose Datapipeline-Frameworks oder Big-Data-Processing-Frameworks zur Orchestrierung der Pipeline und unterscheiden zwischen ereignisgesteuerter und workflow-basierter Orchestrierung. Big-Data-Processing-Frameworks werden in Batch und Stream unterteilt. Batch Processing nutzt batch-orientierte Big-Data-Frameworks wie Apache Spark oder Apache Flink, während Stream Processing stream-orientierte Big-Data-Frameworks wie Apache Kafka oder Apache Storm verwendet.

Die folgende Grafik zeigt eine Übersicht über die unterschiedlichen Data Pipeline Ansätze in Serverless Computing.

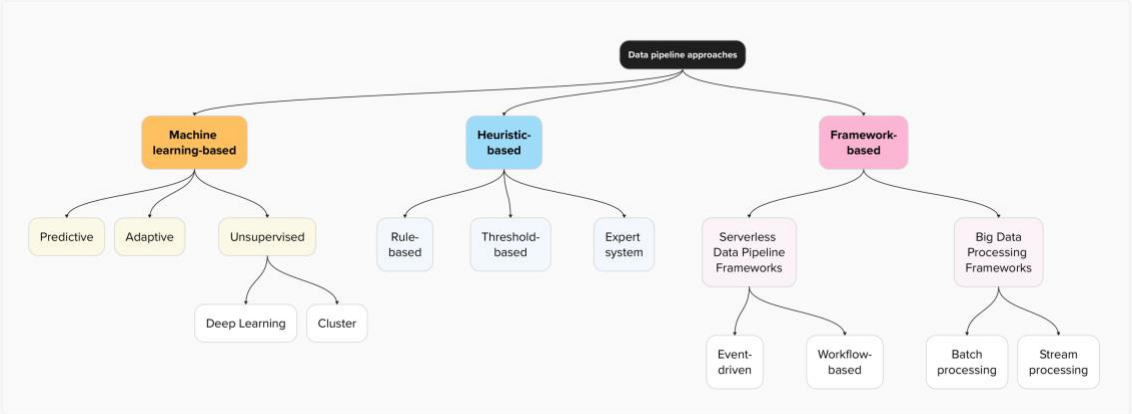

Abbildung 4 Taxonomie der Datenpipeline-Ansätze im serverlosen Computing (Quelle: erstellt in Anlehnung an Shojaee & Ghobaei-Arani, 2024, S. 12)

Durch den Einsatz analytischer Methoden und Werkzeuge werden nützliche Informationen gewonnen, die für KI-Anwendungen verwendet werden können. Diese Daten werden anschließend in Wissen umgewandelt, das in verschiedenen Anwendungen genutzt werden kann. Ein Beispiel hierfür ist die Verkehrsschilderkennung in autonomen Fahrzeugen oder die Implementierung eines

Chatbots, der mit Nutzern über ein Online-System interagiert (Blum, 2023, S. 256). Die folgende Grafik beinhaltet eine KI Data Pipeline.

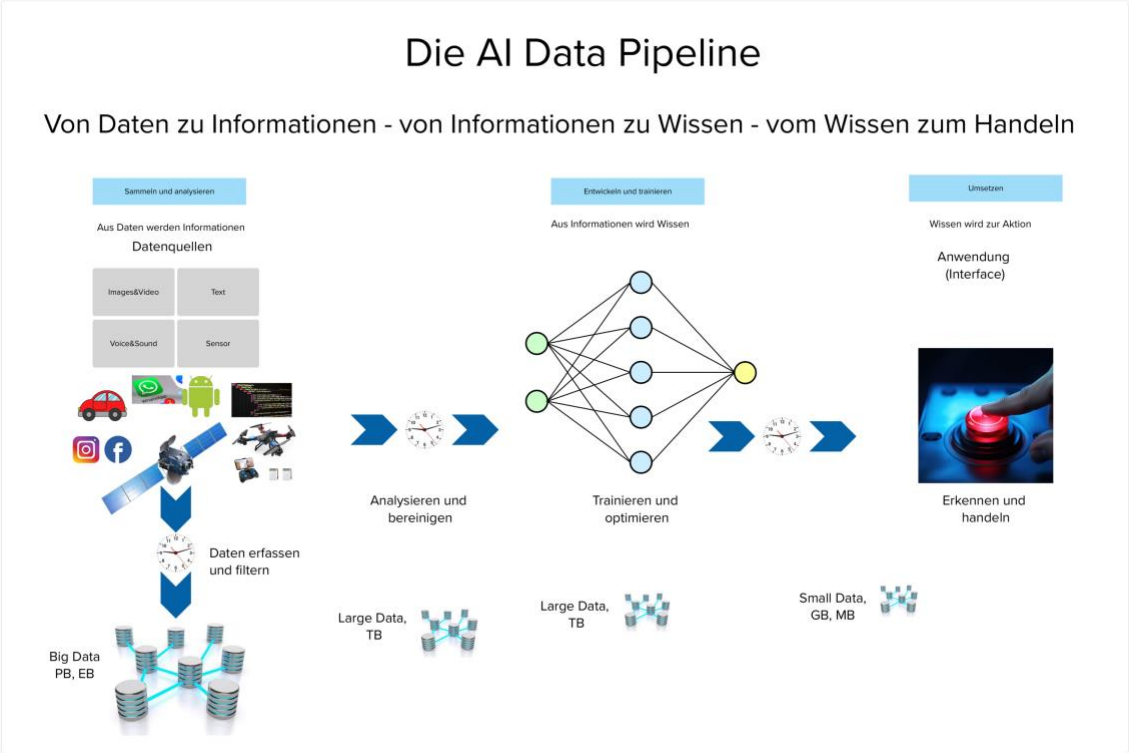

Abbildung 5 KI Data Pipeline (Quelle: erstellt in Anlehnung an Blum, 2023, S. 256)

7. Diskussion

Die wachsende Bedeutung von Big Data, wie sie durch das exponentielle Ansteigen des weltweiten Datenvolumens illustriert wird (Vorndran, 2024, S. 62), stellt Unternehmen vor immense Herausforderungen und bietet gleichzeitig enorme Chancen. Während große Unternehmen oft in der Lage sind, Big Data-Technologien erfolgreich zu integrieren und daraus strategische Vorteile zu ziehen, sehen sich insbesondere mittelständische Betriebe wie das Beispielunternehmen „Game Your Life“ (Kapitel 1.1) mit erheblichen Hürden konfrontiert. Die Aussage des Geschäftsführers, dass „Big Data für uns keinen Sinn macht“, spiegelt die verbreitete Unsicherheit und das mangelnde Vertrauen in die Nutzbarkeit von Big Data wider, obwohl das Potenzial für datengestützte Entscheidungen und Geschäftsoptimierungen offenkundig ist (Ataei & Litchfield, 2022, S. 113789).

Der Konflikt zwischen der Innovationskraft der Mitarbeiterin Tabea Inno und der ablehnenden Haltung des Geschäftsführers illustriert ein typisches Problem in kleinen und mittleren Unternehmen (KMUs). Der Wille, Daten systematisch zu nutzen, scheitert

oft an fehlender Infrastruktur und mangelndem Wissen über geeignete Big Data-Architekturen und -Strategien. Es zeigt sich, dass technologische Entwicklungen wie Hadoop oder Apache Spark, die in großen Unternehmen längst Einzug gehalten haben (Kapitel 2.1), in KMUs oft nicht implementiert werden, da entweder die finanziellen Mittel oder das notwendige Know-how fehlen (Fasel & Meier, 2016, S. 308).

Gleichzeitig wird deutlich, dass Unternehmen wie „Game Your Life“ durch die richtigen architektonischen Ansätze, etwa der Lambda-Architektur, erheblichen Nutzen aus ihren Daten ziehen könnten. Der Wunsch der Mitarbeiterin, maschinelles Lernen zu integrieren und das Spielerlebnis zu optimieren, lässt sich durch die Kombination von Echtzeit- und Batch-Verarbeitung verwirklichen (Kapitel 4.1). Insbesondere die Echtzeitanalysen könnten wertvolle Einblicke in das Spielverhalten liefern, während die Batch-Verarbeitung langfristige Muster und Trends aufdeckt, die das Spieldesign weiter verbessern könnten.

Die technische Herausforderung, große Datenmengen effizient zu verarbeiten und zu speichern, wird in der Fallstudie durch die Darlegung moderner Architekturen wie der Lambda- und Kappa-Architektur umfassend beleuchtet (Kapitel 4.1). Dabei zeigt sich, dass die Wahl der richtigen Architektur entscheidend ist, um die Daten effizient zu nutzen. Während große Unternehmen mit ausgeklügelten Architekturen wie der Lambda-Architektur operieren, könnte für KMUs die Kappa-Architektur eine weniger komplexe und dennoch leistungsfähige Alternative darstellen, um Echtzeitdaten zu verarbeiten und dabei die Komplexität des Systems zu verringern (Oussous et al., 2018).

Nicht zu vernachlässigen ist allerdings der Aspekt des Datenschutzes, insbesondere in Europa, wo die Datenschutz-Grundverordnung (DSGVO) strenge Vorgaben macht (Hornung, 2016, S. 1). Gerade für kleinere Unternehmen sind die Anforderungen, personenbezogene Daten zu schützen und die Datenminimierung konsequent umzusetzen, oft schwer zu erfüllen. Die Fallstudie zeigt auf, dass neben der technischen Implementierung auch die organisatorischen und rechtlichen Rahmenbedingungen beachtet werden müssen, um die Datensicherheit zu gewährleisten und gleichzeitig das Vertrauen der Kunden zu gewinnen. Neben den in der Fallstudie aufgezeigten Aspekten stellen auch kulturelle und organisatorische Barrieren ein Hindernis für die Anwendung von Big Data im Unternehmen dar. Zwar illustriert der Konflikt zwischen der innovationsfreudigen Mitarbeiterin und dem konservativen Geschäftsführer diese Problematik, jedoch fehlt eine vertiefte Diskussion über die Rolle des Managements und die Notwendigkeit einer datenorientierten Unternehmenskultur. Hier sind insbesondere

Change-Management-Ansätze sowie eine umfassende Schulung und Sensibilisierung der Mitarbeitenden von Bedeutung. Ein weiterer Aspekt, der in der Fallstudie nicht ausreichend berücksichtigt wurde, ist eine Kosten-Nutzen-Analyse, die kritisch hinterfragt, ab wann der Einsatz von Big Data für das Unternehmen tatsächlich gewinnbringend ist.

8. Fazit und Ausblick

Zusammenfassend zeigt die vorliegende Arbeit, dass Big Data weit mehr als eine technische Herausforderung ist. Es betrifft die gesamte Organisation eines Unternehmens und erfordert sowohl technisches Wissen als auch organisatorische Anpassungen. Für KMUs stellt sich dabei besonders die Frage, inwiefern die Implementierung von Big Data sinnvoll ist und welche Strategien verfolgt werden sollten, um mit den gegebenen Ressourcen den größtmöglichen Nutzen zu ziehen.

Die Fallstudie verdeutlicht, dass Big Data-Anwendungen, insbesondere maschinelles Lernen und Echtzeitanalysen, für KMUs wie „Game Your Life“ einen erheblichen Mehrwert bieten können, wenn sie strategisch und zielgerichtet eingesetzt werden. Innovative Technologien wie die Lambda-Architektur bieten flexible und skalierbare Lösungen, um sowohl Echtzeitdaten als auch historische Daten effizient zu verarbeiten und für strategische Entscheidungen zu nutzen (Kapitel 4.1). Gleichzeitig darf der Datenschutz nicht vernachlässigt werden. Die strengen Anforderungen der DSGVO erfordern eine datenschutzkonforme Nutzung von Big Data, was insbesondere in KMUs eine Herausforderung darstellt.

Für die Zukunft lässt sich festhalten, dass die Relevanz von Big Data weiter zunehmen wird. Prognosen wie die der IDC, die ein Ansteigen des weltweiten Datenvolumens auf 221 Zettabyte bis 2026 vorhersagen (Vorndran, 2024, S. 62), verdeutlichen, dass Unternehmen, die sich nicht frühzeitig mit Big Data beschäftigen, Gefahr laufen, ihre Wettbewerbsfähigkeit zu verlieren. In diesem Kontext müssen KMUs Wege finden, ihre Kompetenzen im Umgang mit Big Data auszubauen und ihre Daten effizient zu nutzen, um langfristig erfolgreich zu bleiben.

Ein spannender Ausblick zeigt sich in der weiteren Entwicklung von Cloud-basierten und Serverless-Technologien. Diese könnten es auch kleinen und mittelständischen Unternehmen ermöglichen, Big Data-Anwendungen ohne große Investitionen in Hardware und Infrastruktur zu nutzen (Kapitel 6). Hier könnten erste Schritte für KMUs die Nutzung Cloud-basierter Pilotprojekte oder die Zusammenarbeit mit externen Dienstleistern sein, um Big Data risikoarm zu testen. Cloud-native Architekturen bieten

vielversprechende Ansätze, um die Flexibilität und Skalierbarkeit der Systeme zu gewährleisten (Gecko, n.d.). Mit der zunehmenden Integration von maschinellem Lernen und Künstlicher Intelligenz könnten Big Data-Anwendungen in der nahen Zukunft noch automatisierter und effizienter werden. Dies könnte insbesondere für KMUs die entscheidende Veränderung sein, um den Sprung in die datengetriebene Unternehmensführung zu schaffen.

Damit bleibt abschließend zu betonen, dass neben den technologischen Anforderungen auch eine datenorientierte Unternehmenskultur entscheidend ist. Change-Management-Ansätze und eine umfassende Schulung und Sensibilisierung der Mitarbeitenden können helfen, diese Transformation nachhaltig zu gestalten und Big Data erfolgreich in die Unternehmensprozesse zu integrieren.

Literaturverzeichnis

Anderson, J., (2024). Daten-Teams. Ein einheitliches Managementmodell für erfolgreiche, datenorientierte Teams. Springer Nature.

Angbera, A., & Chan, H. Y. (2022). *A Novel True-Real-Time Spatiotemporal Data Stream Processing Framework*.

Ataei, P., & Litchfield, A. (2022). *The State of Big Data Reference Architectures: A Systematic Literature Review*. IEEE Access, 10, 113789-113804. https://doi.org/10.1109/ACCESS.2022.3217557

Azarmi, B. (2016). Scalable big data architecture. A Practitioner's Guide to Choosing Relevant Big Data Architecture. Apress.

Bundesregierung. (2024). *AI Act: Weichenstellung für den Einsatz künstlicher Intelligenz*. Abgerufen am 16.09.2024 von https://www.bundesregierung.de/breg-de/themen/digitalisierung/kuenstliche-intelligenz/ai-act-2285944

Berle, L. (2017). *Streamingarchitekturen in der Praxis: Lambda vs. Kappa*. JAXenter. Abgerufen am 20. September 2024 von https://jaxenter.de/streaming-lambda-kappa-64573

Big Data (n.d.): Herausforderungen bei der Datenanalyse. https://publicgovernance.de/media/Big_Data_Herausforderungen_bei_der_Datenanalyse.pdf

BigData Insider. (n.d.). *Was ist Datenvirtualisierung?* Abgerufen am 16.09.2024 von https://www.bigdata-insider.de/was-ist-datenvirtualisierung-a-841015/

Bitkom. (2024). *Wirtschaftsschutz und Cybercrime 2024: Unternehmensbefragung*. Abgerufen am 17. September 2024 von https://www.bitkom.org/sites/main/files/2024-08/240828-bitkom-charts-wirtschaftsschutz-cybercrime.pdf

Blum, Lothar., B., (2023). Angewandte Data Science. Projekte Methoden Prozesse. Springer Vieweg.

Britton, J., (2021). Was ist ISO 25010? PERFORCE. Abgerufen am 2. November 2024, von https://www.perforce.com/blog/qac/what-is-iso-25010

Dhaouadi, A., Bousselmi, K., Monnet, S., & Mohsen, M. (2024). A machine learning based decision support framework for big data pipeline modeling and design. Jordanian Journal of Computers and Information Technology (JJCIT), 10(03).

Economist, (2017). The world's most valuable resource is no longer oil, but data. Abgerufen am 05.03.2024 von https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

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

Gecko, (n.d.). *Individuelle Software-Entwicklung: Cloud-native Entwicklung*. Gecko. https://www.gecko.de/individuelle-software-entwicklung/cloud-native-entwicklung/

Gluchowski, P., (2024). Datenmanagement und Datenanalyse. Konzepte, Technologien und Methoden für die Organisation und Aufbereitung von Daten in Unternehmen. Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature.

Grasl, O., (2022). Big Data & Analytics. Ihr Leittext. SRH Fernhochschule – The Mobile University. Titel Nr. 1832-01. Riedlingen.

Hornung, G. (2016). *Big Data und Datenschutz: Zehn Thesen*. Statistisches Bundesamt (Destatis). Abgerufen am 17.09.2024 von https://www.destatis.de/DE/Ueber-uns/Kolloquien-Tagungen/Veranstaltungen/BigData/11Hornung.pdf?__blob=publicationFile

Hornung, G., & Herfurth, C. (2018). Datenschutz bei Big Data: Rechtliche und politische Implikationen. In C. König, J. Schröder, & E. Wiegand (Hrsg.), *Big Data: Chancen, Risiken, Entwicklungstendenzen* (S. 149ff). Springer.

IBM. (n.d.). *Das CAP-Theorem: Konsistenz, Verfügbarkeit und Partitionstoleranz in verteilten Systemen*. IBM. Abgerufen am 20.09.2024 von https://www.ibm.com/de-de/topics/cap-theorem

Industrie- und Handelskammer München und Oberbayern. (n.d.). Künstliche Intelligenz und Datenschutz. IHK München. Abgerufen am 17.09.2024 von https://www.ihk-muenchen.de/de/Service/Recht-und-Steuern/Datenschutz/ki/.

Kaufmann, M., Meier, A., (2023). SQL- & NoSQL-Datenbanken. 9. Auflage. Springer Vierweg.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. Journal of parallel and distributed computing, 74(7), 2561-2573.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: issues, challenges, tools and good practices. In 2013 Sixth International Conference on Contemporary Computing (IC3). IEEE. S. 404-409.

Kuenen, K. (n.d.). Customer Journey Prozess. Gabler Wirtschaftslexikon. Abgerufen am 3. November 2024, von https://wirtschaftslexikon.gabler.de/definition/customer-journey-prozess-100259

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT data analytics: Architecture, opportunities, and open research challenges. IEEE Access, 5, 5247-5261.

Moreno, J., Serrano, M. A., Fernandez-Medina, E., & Fernandez, E. B. (2018). *Towards a security reference architecture for big data*. Workshop Proceedings of the EDBT/ICDT 2018 Joint Conference, Vienna, Austria. CEUR-WS.org. https://ceur-ws.org

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences, 30*(4), 431-448. https://doi.org/10.1016/j.jksuci.2017.06.001

Rahman, M. S., & Reza, H. (2020). *Systematic mapping study of non-functional requirements in big data systems*. In *Proceedings of the IEEE* (pp. 025-031). IEEE. https://doi.org/10.1109/ISBN-978-1-7281-5317-9

SAP. (n.d.). *Was ist eine In-Memory-Datenbank?* SAP. https://www.sap.com/germany/insights/in-memory-database.html

Shojaee Rad, Z., & Ghobaei-Arani, M. (2024). Data pipeline approaches in serverless computing: A taxonomy, review, and research trends. *Journal of Big Data, 11*(82). https://doi.org/10.1186/s40537-024-00939-0

Steps & Milestones. (n.d.). *Horizontale vs. vertikale Entwicklung*. https://www.stepsandmilestones.ch/Portals/0/adam/Content/Z7w9Ci0kb06gAthJBXK4Fg/BetweenTheLines/Horizontale%20vs.%20Vertikale%20Entwicklung.pdf

Tam, S.-M., & Van Halderen, G. (2020). The five V's, seven virtues and ten rules of big data engagement for official statistics. *Statistical Journal of the IAOS, 36*(3), 423–433. https://doi.org/10.3233/SJI-190595

Vorndran, A., (2024). Big Data und Advanced Analytics im Controlling. Potenziale, Herausforderungen und Implikationen für die Praxis. Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024.

Zicari, R. V. (2015). Big Data – Chancen und Herausforderungen. **Aktuelles Goethe Universität Frankfurt**. https://aktuelles.uni-frankfurt.de/forschung/big-data-chancen-und-herausforderungen/

Anhang – Versicherung

Versicherung

Name: Dagmar Rostek
Matrikel-Nr.: 2004169
Studiengang: (M.Sc.) Data Science & Analytics

Hiermit versichere ich, dass ich die vorliegende Hausarbeit mit dem Thema

„Alternative A: Big Data Grundlagen.“

ohne fremde Hilfe erstellt habe. Alle verwendeten Quellen wurden angegeben. Ich versichere, dass ich bisher keine Hausarbeit mit gleichem oder ähnlichem Thema an einer anderen Hochschule eingereicht habe.

Datum: 03.11.2024 Unterschrift: