TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2022.

VORLESUNG 1

13.09.2022

MÜNCHEN

STUDIENGANG SUSTAINABILITY MANAGEMENT & LEADERSHIP SOWIE MEDIEN & KOMMUNIKATION.

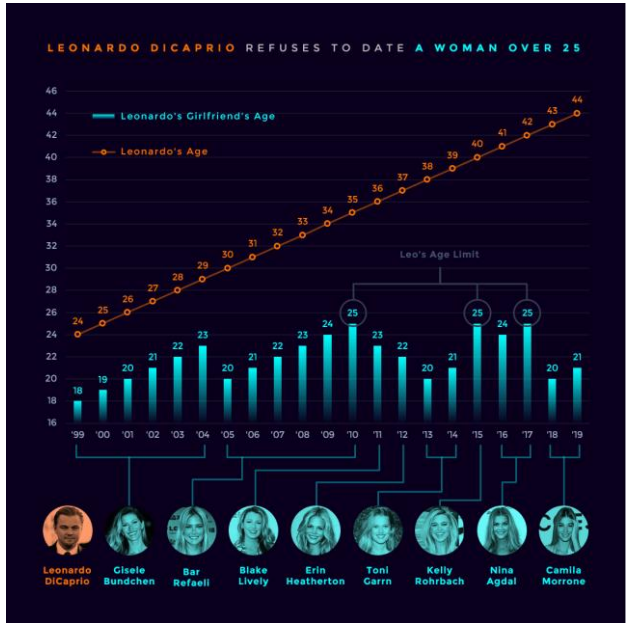
AGENDA

1. Allgemeines
2. Roadmap Vorlesung
3. Daten und Datenqualität

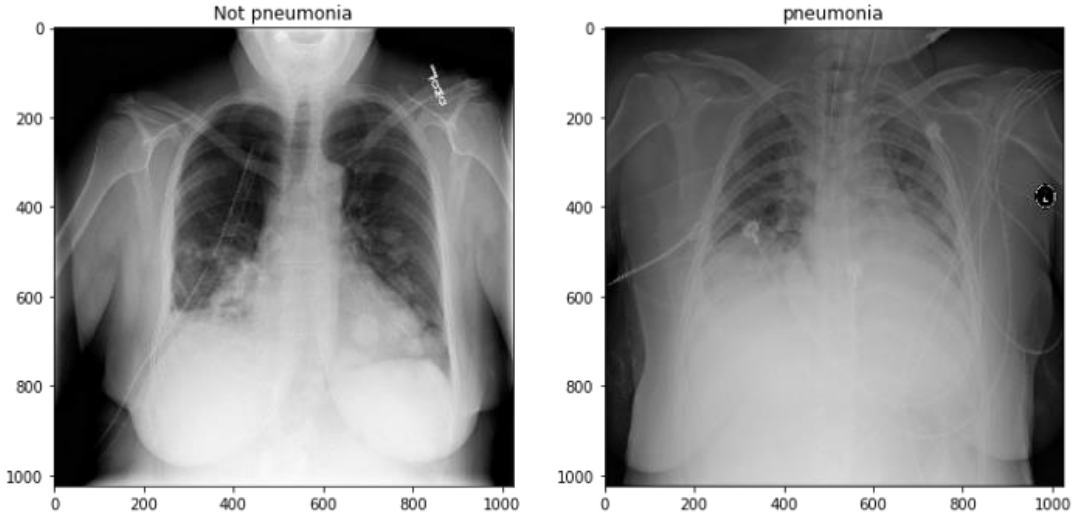
1. ALLGEMEINES

WAS MACHEN WIR HIER EIGENTLICH SO?

Data Science: Daten analysieren und daraus Erkenntnisse gewinnen und visualisieren



Machine Learning: aus Daten ein Modell lernen, das eigenständig Erkenntnisse gewinnt.



Wir versuchen, aus Daten einen Mehrwert zu schaffen

Quellen: Data Science Chart: Link: https://i.redd.it/jtmyrtyqmcl21.png
Pneumonia-Bild: aus Datensatz verfügbar unter Link

EXPECTATIONS EXCHANGE: WAS IST MIR WICHTIG?

- Reduktion zweier großer Themenfelder auf wesentliche Inhalte
- Verstehen der Grundlagen und praktisches Anwenden
- Sammeln von Hands-on Experience an praxisnahen Aufgabenstellungen/ Themen
- FRAGEN, FRAGEN, FRAGEN!!
- Angebot einer zweiwöchentlichen Sprechstunde

Was sind Ihre Erwartungen?

DIE BENOTUNG/ CREDITS-VERGABE ERFOLGT AUF BASIS VON GRUPPENARBEIT.

1. Wahl je 1 Data Science- sowie 1 Artificial Intelligence-Themas.

2. Zwei Schulterblick-Termine entlang gesamten Semesters anhand Word-/ Powerpoint-Dokument mit max. 4 Seiten:
   - Detaillierung Problem statement und Problemdomäne: „Was ist das Problem? Was ist der Nutzen der Lösung?"
   - Metriken zur Evaluation Ergebnisse
   - Vorgehensweise Lösungsansatz sowie aktueller Status

Template wird bereitgestellt

3. Präsentationstermin (beide Themen):
   - Schriftliche Ausarbeitung je Teilnehmer:
     - Vorgehensweise: Detaillieren und Erklären der eingesetzten Verfahren sowie der Implementation
     - Ergebnisse: Visualisierung Ergebnisse, Bewertung Ergebnisse anhand Metriken
     - Reflektion und Ausblick

Template wird bereitgestellt, Aufbau auf vorigem Dokument

Prüfungsleistung je Student: je Thema 1 Präsentation (~10 Min.), 1 Ausarbeitung (~7 Seiten) und dokumentierter Code

ROADMAP VORLESUNG

GEPLANTE ROADMAP VORLESUNG.

ROADMAP	WAS HABEN WIR VOR?
Vorlesung 1	Workflow Data Management, Datentypen und Datenqualität
Vorlesung 2	Einführung Data Science und Data Science Workflow, Grundlagen Data Management
Vorlesung 3	Deskriptive und explorative Datenanalyse
Vorlesung 4	Vertiefung Datenanalyse anhand Case Study
Vorlesung 5	Aufgabenstellung Data Science, Übersicht und Einführung Machine Learning, unüberwachtes Lernen
Vorlesung 6	Überwachtes Lernen
Vorlesung 7	Vertiefung überwachtes Lernen anhand Case Study
Vorlesung 8	Neuronale Netze und Convolutional Neural Networks (CNN)
Vorlesung 9	Vertiefung CNN anhand Case Study, Aufgabenstellung AI
Vorlesung 10	Schulterblick 1 Data Science
Vorlesung 11	Übersicht Rekurrente Neuronale Netze
Vorlesung 12	Schulterblick 2 AI
Vorlesung 13	Ausblick zukünftige AI-Themen, „Fragestunde“
Vorlesung 14	Präsentation Ergebnisse

Folien der bisherigen Vorlesung verfügbar unter Link

Semesterende: Januar 2023
Abgabedatum Hausarbeit: 13.02.2023

Angebot für Sprechstunden zwischen Vorlesungen (nach Bedarf)

DATEN UND DATENQUALITÄT

WAS SCHAUEN WIR UNS HEUTE AN?

- Datenbasierte Geschäftsmodelle: Daten haben Wert
- Übersicht Workflow Datenmanagement
- Datenqualität: wie müssen die Daten sein, damit Geschäftsmodelle funktionieren?

1. DATENBASIERTE GESCHÄFTSMODELLE.

“**Uber**, the world’s largest **taxi company**, **owns no vehicles**.

Facebook, the world’s most popular **media owner**, **creates no content**.

Alibaba, the most valuable **retailer, has no inventory.**

And **Airbnb**, the world’s largest **accommodation provider**, **owns no real estate**.

Something interesting is happening.”

Tom Goodwin (2015)

Geschäftsmodelle heutiger Tech-Firmen basieren auf der Sammlung, Verknüpfung und Auswertung von Daten

Quellen: Tom Goodwin, „The Battle is for the Customer Interface“, 2015, abgerufen am 11.09.2021 unter: Link
Bilder sind Firmenlogos, entnommen der jeweiligen Homepages.

1. DATENBASIERTE GESCHÄFTSMODELLE.

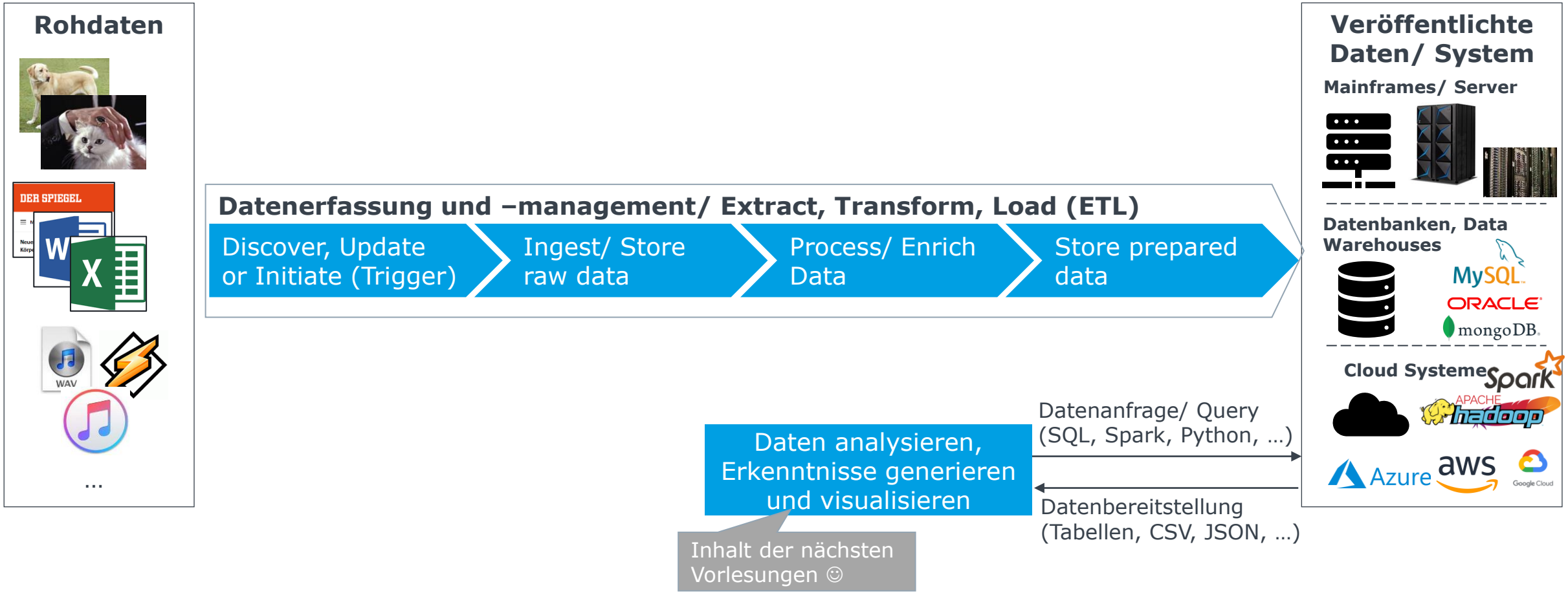
- **Data-informed[1] Geschäftsmodelle:** Optimierung bestehender Wertschöpfungsprozesse durch Daten.
  - Prozessoptimierung durch Automatisierung (gesamte Industrie).
  - Reduktion Entwicklungszeit/-kosten durch Simulation (Luft- und Raumfahrttechnik, Automobilbereich).
  - Online-Vertrieb für physische Produkte (Otto, Lieferando, Zalando, Amazon).
  - Mobility Dienste (Uber, Lyft).

- **Data-infused[1] Geschäftsmodelle**: Wertschöpfungsprozesse hängen wesentlich von Daten ab.
  - Personalisierte Werbung (Facebook und Google).
  - Personalisierte Produktempfehlungen (Amazon).
  - Quantitative Analysis/ Algorithmic Trading.

- **Data driven[1] Geschäftsmodelle:** Wertschöpfung vollständig digital.
  - Online-Vertrieb digitaler Produkte (Netflix, Spotify, Steam, ….).
  - Software-Geschäftsmodelle (Werbebasiert, Freeware, Freemium, Shareware, Mieten, Kauf).

Quelle: 1 Begriffe wurden so zuerst auf der UX Week 2010 durch A. Mosseri verwendet. Link

2. ÜBERSICHT WORKFLOW DATENMANAGEMENT



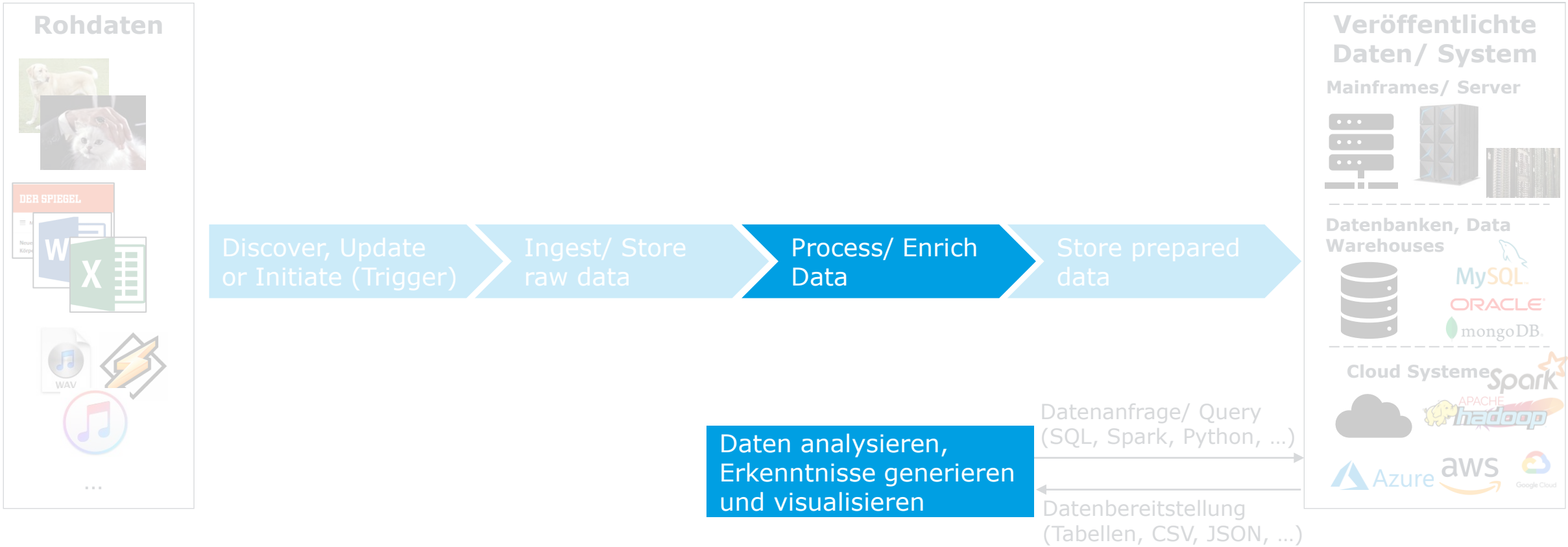
Quellen: Workflow basierend auf Google Dataflow, abgerufen am 07.09.2020 unter: Link
Bilder: Stockpictures sowie Homepages der jeweiligen Hersteller.

2. ÜBERSICHT WORKFLOW DATENMANAGEMENT

Detaillierungsfolien zu den einzelnen Schritten im Backup

Quellen: Workflow basierend auf Google Dataflow, abgerufen am 07.09.2020 unter: Link
Bilder: Stockpictures sowie Homepages der jeweiligen Hersteller.

6. DATENQUALITÄT



Quellen: Workflow basierend auf Google Dataflow, abgerufen am 07.09.2020 unter: Link
Bilder: Stockpictures sowie Homepages der jeweiligen Hersteller.

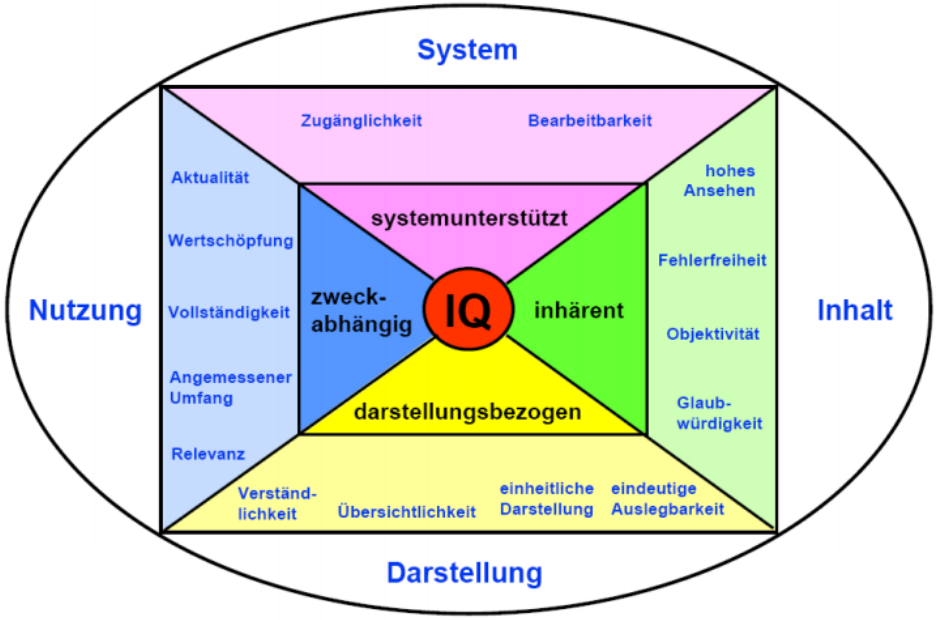
6. DATENQUALITÄT – EINFÜHRUNG.

Fallbeispiel: Kundenliste eines Online-Shops in einer Datenbank.

Kunden-Nr.	Name	Geburts-datum	Alter	Geschlecht	Email	PLZ	Stadt	Letzter Kontakt	T645fet	Umsatz 2015
20456	Tina Huber	10.01.2010	21	W		8000	München	01.08.2021	Ja	100€
20456	Teddy Test	6.8.1490	20	M	test@test.de	80797	Freising	05.03.2008	Nein	
23578	B. Trüger	08.07.1979	41	D	trueger@gmx.de	D-80793	Muenchen	01.07.2020	bald	10000
28903	Amy Doe	03/12/2003		F	amyd@yahoo.com		Düsseldoof	15.07.2020	ja	4000$

Welche Fehler/ Probleme sehen Sie?

6. ÜBERSICHT DATENQUALITÄT



Detaillierung Kriterien im Backup

Es gibt viele verschiedene Kriterien für Datenqualität, die o.a. Kriterien sind bekannte Beispiele.
Es werden auch nicht immer alle verwendet.

Quelle: Deutsche Gesellschaft für Informations- und Datenqualität Projekt, abgerufen am 06.09.2020 unter: Link.
Basiert auf MIT Total Data Quality Management Program (R. Wang)

FALLBEISPIEL DATENQUALITÄT

FALLBEISPIEL DATENQUALITÄT.

Wählen Sie eine beliebige Tech-Firma (Facebook, Google, Amazon, ...).

Prüfen Sie für die gewählte Firma folgendes:

- Kundenhypothesen: Wie generiert die gewählte Firma mit Daten Mehrwert für den Kunden?
- Geschäftsmodell: Wie generiert die gewählte Firma mit Daten Einnahmen?
- Leiten Sie aus der Kundenhypothese und dem Geschäftsmodell die Datenarchitektur ab:
  - Welche Daten benötigt die gewählte Firma hierfür?
  - Wie müssen die Daten dann sein? Welche Kriterien für Datenqualität sind dann wichtig?

- Skalieren: Nehmen Sie an, Sie haben 100 000 oder mehr Kunden/ User.
  - Können Sie Regeln für das Erfassen, Prüfen, Auswerten der Daten definieren?
  - Wie können Sie –bspw. auf Basis der definierten Regeln – die Vorgänge automatisieren?

BEISPIELHAFTE KRITERIEN FÜR DATENQUALITÄT.

Fehlerfreiheit: ... wenn sie mit der Realität übereinstimmen

Eindeutig. Auslegbarkeit: …wenn sie in gleicher, fachlich korrekter Art und Weise begriffen werden

Einheitliche Darstellung: …wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden

Übersichtlichkeit: …wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.

Vollständigkeit: ….wenn sie nicht fehlen & zu festgelegten Zeitpunkten in den jeweiligen Prozessschritten zur Verfügung stehen

Verständlichkeit: …wenn sie unmittelbar von den Anwendern verstanden und für deren Zwecke eingesetzt werden können

Relevanz: …wenn sie für den Anwender notwendige Informationen liefern.

Glaubwürdigkeit: wenn Zertifikate einen hohen Qualitätsstandard ausweisen oder die Informationsgewinnung und –verbreitung mit hohem Aufwand betrieben werden.

Aktualität: wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden.

Wertschöpfung: wenn ihre Nutzung zu quantifizierbaren Steigerung einer monetären Zielfunktion führen kann.

Datenaufbereitung und –bearbeitung beträgt ca. 70-80% der Zeit eines Use Case Data Science oder AI!

PERSONALISIERTE KAUFEMPFEHLUNGEN ONLINE-SHOP - PRÄMISSEN.

Ziel: Generieren Einnahmen für einen Online-Shop durch personalisierte Kaufempfehlungen (Was kauften ähnliche Kunden?).

Dazu benötigen wir (Auszug…):

- Für jeden Kunden eine Liste seiner Einkäufe, aus der wir per Abgleich mit ähnlichen Kunden Empfehlungen generieren.
- (viele) soziographische Daten je Kunde. Durch aggregieren dieser Kundendaten, lernen wir ein Modell für Bestimmen:
  - Wie solvent ein individueller Kunde ist (bspw. anhand Wohnviertel, Umsatz in den letzten Jahren, ….)
  - Ähnlicher Kunden zu einem individuellen Kunden („Was für Kunde A relevant ist, ist es vielleicht auch für Kunde B…“)
- Unser Geschäftsmodell funktioniert nur mit qualitativ guten Daten, da sonst die Kaufempfehlungen nicht überzeugen.
- Da wir viele Kunden haben, brauchen wir automatisiert auswertbare Regeln für das Prüfen der Daten (übernächste Folie).

Wie solche Regeln sowie Empfehlungsmodell programmiert wird, schauen wir uns in den weiteren Vorlesungen noch an..

PERSONALISIERTE KAUFEMPFEHLUNGEN ONLINE-SHOP – ANWENDEN DER AUSGEWÄHLTE KRITIERIEN FÜR DATENQUALITÄT.

Fehlerfreiheit: für jeden Eintrag/ Zeile ergeben die definierten Prüfkriterien keinen Fehler.

Einheitl. Darstellung: Geldsummen immer in Euro, Telefonnummern immer mit internationaler Vorwahl, …

Übersichtlichkeit: genau die für Betreuung relev. Eigenschaften in leicht fassbarem Format (z.B.: Adresse liegt vor, nicht zu viele Infos)

Verständlichkeit: die Attribute und Werte des Kunden sind für jeweilige Bearbeiter der Firma verständlich (Support, Werbeabteilung, …)

Vollständigkeit: für jeden Kunden sind alle Attribute befüllt.

Relevanz: die für die Anwendungsfälle (bspw. Betreuung, Kaufempfehlung, …) notwendigen Eigenschaften des Kunden sind vorhanden. Das ist das Zweckbindungsprinzip aus der Datenschutzgrundverordnung rein (Art. 5-1b[1]).

Angemessener Umfang: nur die für die Anwendungsfälle notwendigen Daten werden erfaßt (Minimalprinzip aus der DSGVO, Art. 5-1c[1])

Glaubwürdigkeit: die Daten sind vertrauenswürdig. Dieses Kriterium ist oft schwammig. In der Praxis geht man oft davon aus, daß falls die Postadresse existiert, Kreditkarte gültig ist (bspw. per Minibuchung 0,01€), die Daten des Kunden glaubwürdig sind.

Aktualität: Kundendaten sind auf dem letzten Stand (bspw. seiner letzten Transaktionen/ Interaktionen mit der Firma)

Wertschöpfung: siehe vorige Seite

Quelle: DSGVO, verfügbar bspw. unter Link, abgerufen am 19.09.2021

PERSONALISIERTE KAUF-EMPFEHLUNGEN ONLINE-SHOP – DATENARCHITEKTUR UND REGELN ZUR SICHERSTELLUNG DATENQUALITÄT.

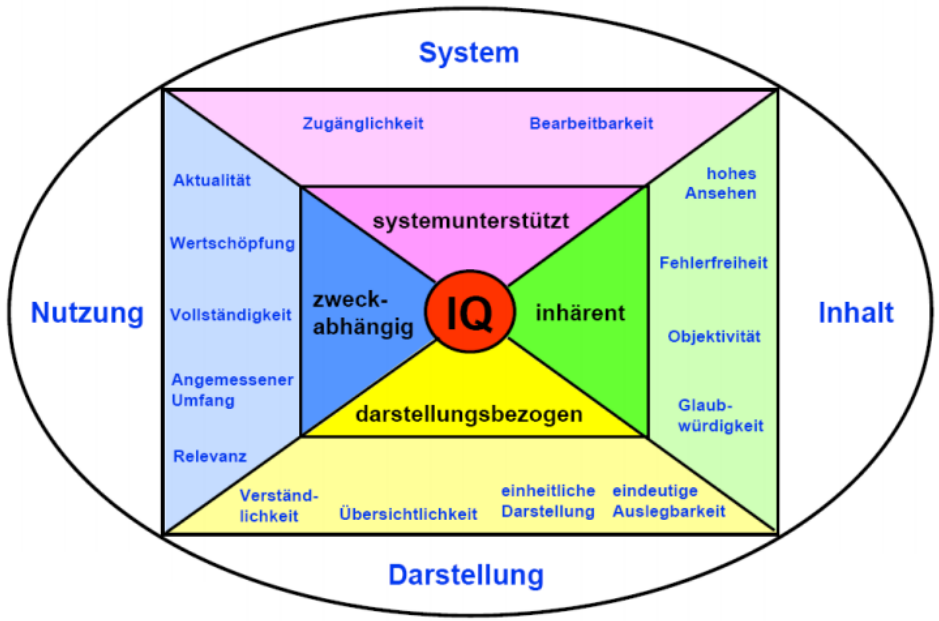
	Kunden-ID	Name	Geboren	Alter	Adresse	Kreditkartennummer	Einkäufe 2020	Umsätze 2020
Regel für Sicherstellen Datenqualität	ID definiert und eindeutig (d.h. darf max. 1 mal vorkommen)	Liegt vor	Geburtsdatum in europäischem Format: TT.MM.YY., sonst umwandeln	Alter < 120	muß vorliegen	1. 12 ≤ Anzahl Ziffern ≤ 16 2. Korrekte Prüfsumme (bspw. Luhn-Algorithmus[1])		Währung in EUR, sonst umwandeln
Relevant für Wertschöpfung per Service/ Empfehlung	-	-	Altersgruppen	Ja, für Empfehlungen Aber bspw. auch für Ansprache Kunde	Ja, bspw. Wohnort		Ja, für Empfehlungen	Ja, für Empfehlungen

Es gibt für Anzahl, Art und Umfang der Features kein richtig oder falsch.
Art und Umfang entwickelt sich über die Jahre, bspw. aufgrund gesetzlicher Anforderungen, Business Logic, …

1 Luhn Algorithmus von 1960, Patent verfügbar unter: Link, abgerufen am 19.09.2021

BACKUP

6. ÜBERSICHT DATENQUALITÄT. KATEGORIE SYSTEM.



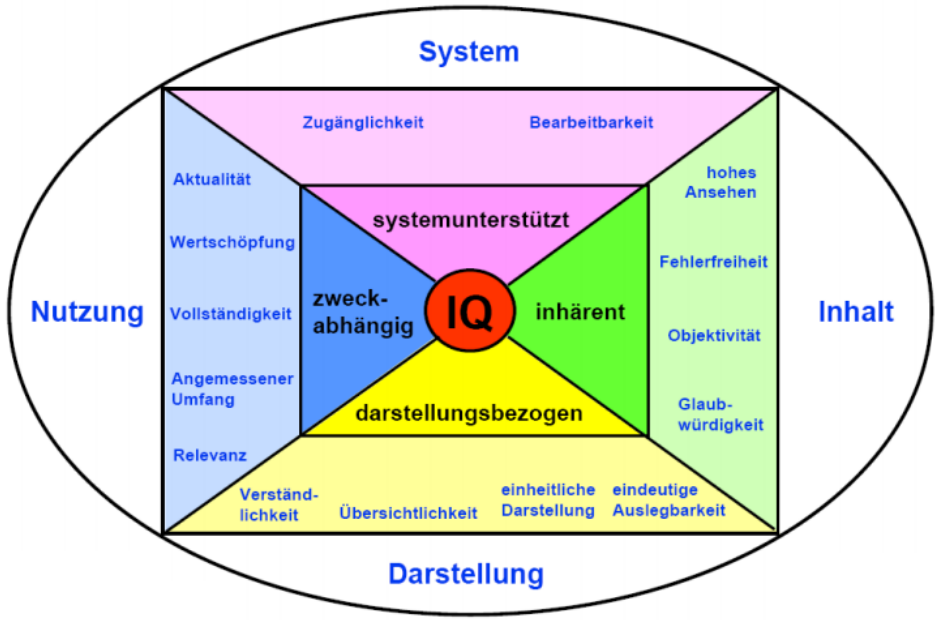
Informationen haben...

Zugänglichkeit (accessibilty): wenn sie anhand einfacher Verfahren auf direktem Weg für den Anwender abrufbar sind.

(leicht) Bearbeitbarkeit (ease of manipulation): wenn sie leicht zu ändern/ für unterschiedliche Zwecke zu verwenden sind.

Quelle: Deutsche Gesellschaft für Informations- und Datenqualität Projekt, abgerufen am 06.09.2020 unter: Link.
Basiert auf MIT Total Data Quality Management Program (R. Wang)

6. ÜBERSICHT DATENQUALITÄT. KATEGORIE INHALT.



Informationen haben…

Hohes Ansehen: (reputation): wenn die Informationsquelle, das Transport-medium und das verarbeitende System im Ruf einer hohen Vertrauenswürdigkeit und Kompetenz stehen.

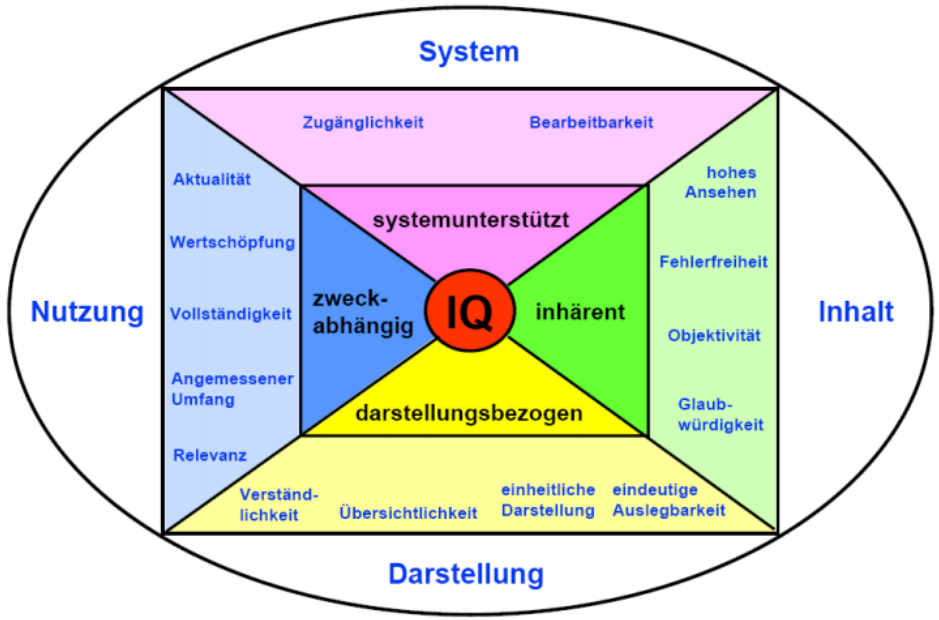
Fehlerfreiheit (free of error): wenn sie mit der Realität übereinstimmen.

Objektivität (objectivity): wenn sie streng sachlich und wertfrei sind

Glaubwürdigkeit (believability): wenn Zertifikate einen hohen Qualitäts-standard ausweisen oder die Informationsgewinnung und –verbreitung mit hohem Aufwand betrieben werden.

Quelle: Deutsche Gesellschaft für Informations- und Datenqualität Projekt, abgerufen am 06.09.2020 unter: Link.
Basiert auf MIT Total Data Quality Management Program (R. Wang)

6. ÜBERSICHT DATENQUALITÄT. KATEGORIE DARSTELLUNG.



Informationen haben…

Eindeutig. Auslegbarkeit (interpretability): wenn sie in gleicher, fachlich korrekter Art und Weise begriffen werden

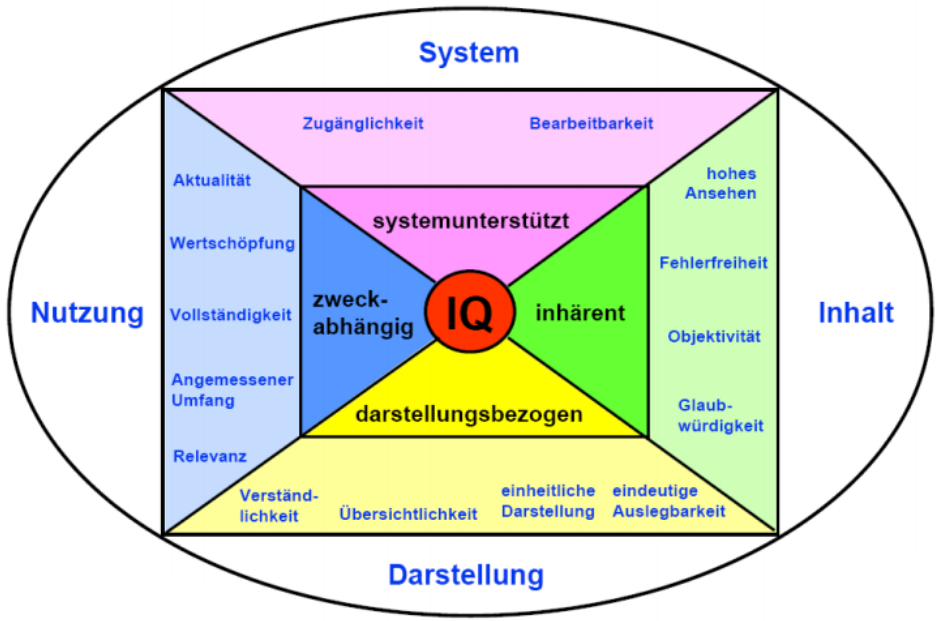
Einheitl. Darstellung (consistent representation): wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden.

Übersichtlichkeit (concise representation): wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.

Verständlichkeit (understandability): wenn sie unmittelbar von den Anwendern verstanden und für deren Zwecke eingesetzt werden können.

Quelle: Deutsche Gesellschaft für Informations- und Datenqualität Projekt, abgerufen am 06.09.2020 unter: Link.
Basiert auf MIT Total Data Quality Management Program (R. Wang)

6. ÜBERSICHT DATENQUALITÄT KATEGORIE NUTZUNG.



Informationen haben...

Aktualität (timeliness): wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden.

Wertschöpfung (value-added): wenn ihre Nutzung zu quantifizierbaren Steigerung einer monetären Zielfunktion führen kann.

Vollständigkeit (completeness): wenn sie nicht fehlen und zu den festgelegten Zeitpunkten in den jeweiligen Prozessschritten zur Verfügung stehen.

Angemessener Umfang (appropriate amount of data): wenn die Menge der verfügbaren Information den gestellten Anforderungen genügt.

Relevanz (relecvancy): wenn sie für den Anwender notwendige Informationen liefern.

Quelle: Deutsche Gesellschaft für Informations- und Datenqualität Projekt, abgerufen am 06.09.2020 unter: Link.
Basiert auf MIT Total Data Quality Management Program (R. Wang)

7. FALLBEISPIELE ANHAND BUSINESS CANVAS[1]

Key Partners

- Who are our Key Partners?
- Who are our key suppliers?
- Which Key Resources are we acquiring from partners?
- Which Key Activities do partners perform?

Key Activities

- What Key Activities do our Value Propositions require?
- Our Distribution Channels?
- Customer Relationships?
- Revenue streams?

Key Resources

What Key Resources do our Value Propositions require?
Our Distribution Channels?
Customer Relationships?
Revenue Streams?

Value Proposition

- What value do we deliver to the customer?
- Which one of our customer's problems are we helping to solve?
- What bundles of products and services are we offering to each Customer Segment?
- Which customer needs are we satisfying?

Customer Relationships

- What type of relationship does each of our Customer Segments expect us to establish and maintain with them?
- Which ones have we established?
- How are they integrated with the rest of our business model?
- How costly are they?

Channels

- Through which Channels do our Customer Segments want to be reached? How are we reaching them now? How are our Channels integrated? Which channels work best? Which ones are most cost-efficient? How are we integrating them with customer routines?

Customer Segments

- For whom are we creating value?
- Who are our most important customers?

Cost Structure

- What are the most important costs inherent in our business mode
- Which Key Resources are most expensive?
- Which Key Activities are most expensive?

Revenue

Fragen für die Einbindung von Daten:
- Wie funktioniert das datenbasierte Geschäftsmodell genau?
- Wie generiert die Firma (mit Daten) Value für den Kunden?
- Wie generiert die Firma (mit Daten) Einkommen?
- Welche Daten werden hierfür aus Ihrer Sicht benötigt?

Business Canvas[1] ist Bestandteil der Lean Startup Methode[2] und wird häufig im Umfeld Startups eingesetzt.

→ Erster Schritt Richtung Durchführung Data Science/ AI Use Case

FALLBEISPIEL DATENINTENSIVE VERTEILTE ANWENDUNG

WIE SCHAUT SO EIN GESAMTES SYSTEM AUS?

- Wir schauen uns so ein Gesamtsystem näher anhand des Fallbeispiels Instagram an.
- Dabei fokussieren wir auf das Design des Gesamtsystems und der Datenperspektive
- Wir gehen vor wie im „richtigen Leben" solche Systeme designt werden, aber aufgrund Zeit machen wir keinen Deep-Dive.

Wichtig: im weiteren Verlauf der Vorlesung werden wir uns mit vorhandenen Daten befassen.
Es ist aber interessant zu sehen, wie solche großen Systeme entwickelt werden.

FALLBEISPIEL INSTAGRAM (REDUZIERT). SCHRITT 1: KLÄRUNG BETRACHTUNGSUMFANG.

Funktionale Anforderungen:

- Bilder/ Videos hochladen, anschauen, liken und kommentieren
- Anderen Usern folgen
- Newsfeed
- **Datenanalyse – und auswertung** (im Hintergrund)
- Monetarisierung/ Werbung

Nicht im aktuellen Fokus:

- User Management

Nicht-funktionale Anforderungen:

- Hohe Verfügbarkeit
- Geringe Latenz (Wartezeit)
- Hohe Verläßlichkeit (Daten gehen nicht verloren)

FALLBEISPIEL INSTAGRAM (REDUZIERT). SCHRITT 2: ABSCHÄTZUNG LAST - WAS MUSS DAS SYSTEM ABKÖNNEN?

Speicherplatz:

Wie viele Daten kommen pro Tag/ Jahr zusammen?

Wie lange sollen die Daten gespeichert werden?

Abschätzung:

- 1 Mrd. aktive Anwender *
- Upload 3 Photos pro Tag je User *
- 300 KB Größe je Bild *

= 10 MB je Sekunde = 900 GB pro Tag = 328 TB pro Jahr

Leselast:

Wie viel der gespeicherten Daten wird je Sekunde abgerufen?

Abschätzung:

- 1 Mrd. aktive Anwender *
- 10 Bilder pro Tag je User abgerufen *
- 300 KB Bildgröße

= 3 PB pro Tag und ~34 GB je Sekunde

FALLBEISPIEL INSTAGRAM (REDUZIERT). SCHRITT 3: SCHNITTSTELLEN UND DATENSTRUKTUR.

Programmierschnittstellen System (API)

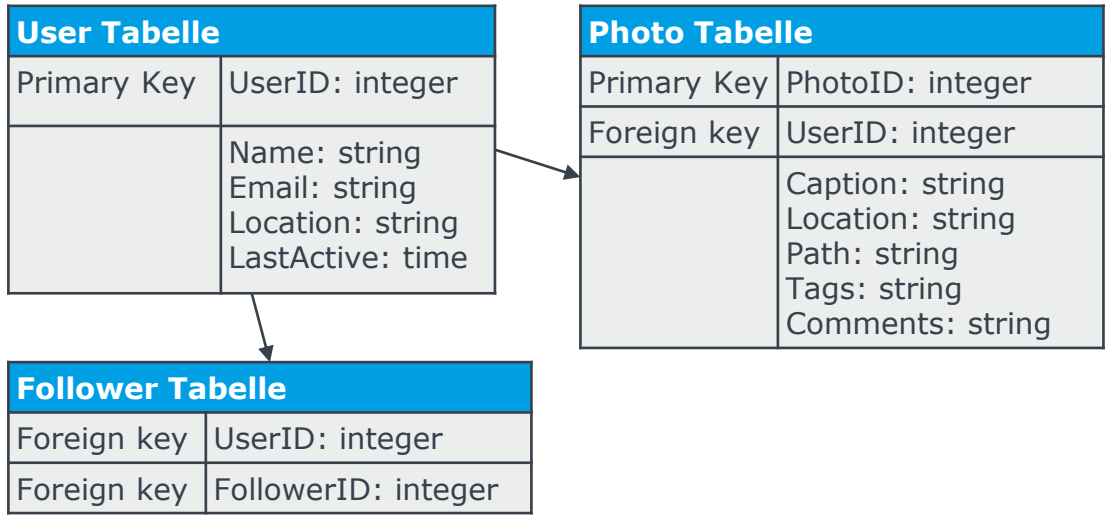
Daten speichern (POST API):

- Bilder hochladen: UploadImage(myUserID, Image, Caption, Location, Tags, Comment)
- Bilder liken: LikeImage(myUserID, PhotoID)
- Bilder kommentieren: CommentImage(myUserID, photoID, comment)
- User folgen: FollowUser (myUserId, UserToFollow)

Daten erhalten (GET API):

- Bilder anschauen: ViewImage(myUserId, PhotoID)
- News Feed: GetFeed(myUserID, FollowerID)
- Werbung ausspielen: GetAdverts(userID, LinkToSpot)

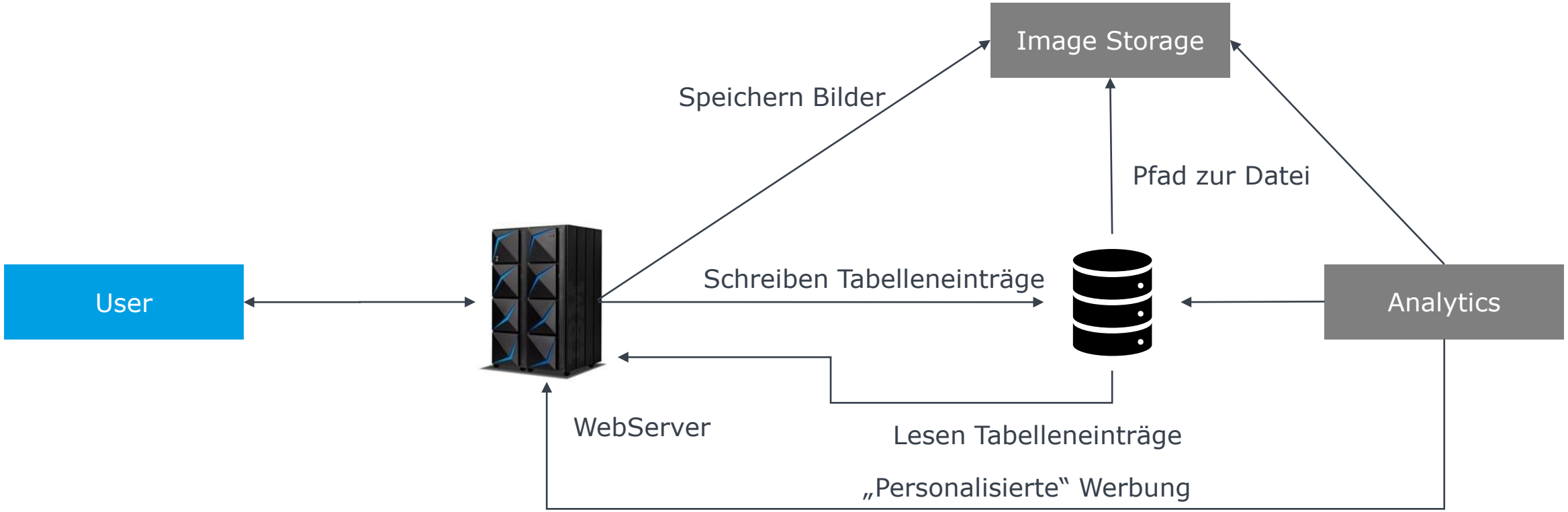
Datenstruktur



Welche weiteren Daten wären interessant, um mit Werbung (oder anderen Diensten) Geld zu verdienen?

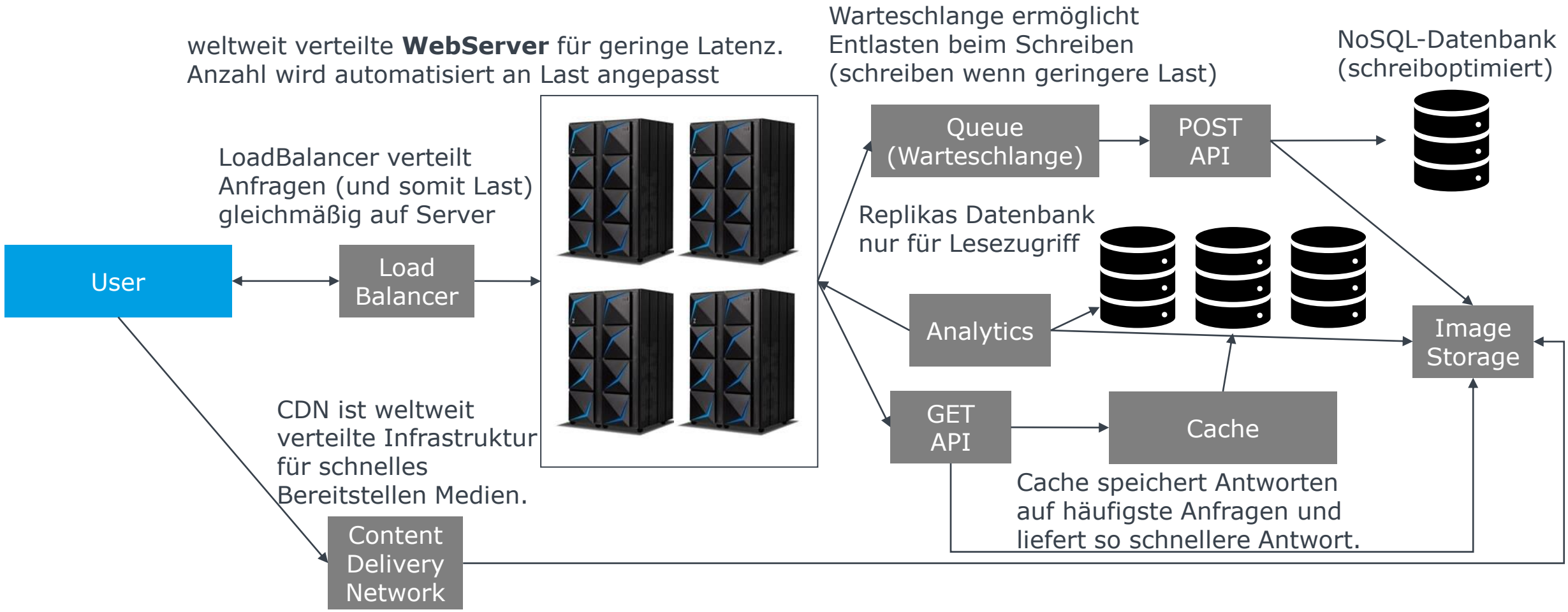
Schnittstellen und Datenstruktur werden im Laufe des Lebenszyklus kontinuierlich angepaßt.

FALLBEISPIEL INSTAGRAM (REDUZIERT). SCHRITT 4: GROB/HIGH LEVEL DESIGN.



Dies ist eine grundlegende Architektur, die aber nicht für eine hohe Anzahl User und Last geeignet ist

FALLBEISPIEL INSTAGRAM (REDUZIERT). SCHRITT 5: SKALIERBARES DESIGN.



Dies ist nur eine beispielhafte Lösung und unterscheidet sich je nach Fokus auf Verfügbarkeit, Latenz, Kosten, …