TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2021/2022.

VORLESUNG 1

14.09.2021

MÜNCHEN

STUDIENGANG DIGITAL MANAGEMENT.

AGENDA

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

1. ALLGEMEINES

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!!

Was sind Ihre Erwartungen?

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

1. Wahl je 1 Data Science- sowie 1 Artificial Intelligence-Themas und Umsetzung in Gruppenarbeit (3 oder 4 Personen).

2. Schulterblick-Termin: Aufzeigen aktueller Status. Abzugeben ist ein Word-/ PDF-Dokument mit ca. 4 Seiten je Gruppe
   - Detaillierung Problem statement und Problemdomäne: „Was ist das Problem? Was ist der Nutzen der Lösung?"
   - Metriken zur Evaluation Ergebnisse
   - Vorgehensweise Lösungsansatz

Template wird bereitgestellt

3. Präsentationstermin (beide Themen):
   - Powerpoint-Präsentation: Gruppe präsentiert ihre Ergebnisse mit gesamthaft 30 Minuten (jeder ca. 10 Minuten)
   - Schriftliche Ausarbeitung je Teilnehmer (Arbeitsumfänge müssen individuell zuordenbar sein):
     - 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

WAS HABEN WIR BIS JETZT GEMACHT?

ROADMAP	WAS HABEN WIR GEMACHT?
Vorlesung 1	Workflow Data Management, Datentypen und Datenqualität
Vorlesung 2	Einführung Data Science und Data Science Workflow, Grundlagen Data Management
Vorlesung 3	Grundlagen Stochastik: Wahrscheinlichkeitsrechnung, deskriptive und explorative Statistik
Vorlesung 4	Statistische Inferenz, lineare Regression
Vorlesung 5	Einführung Machine Learning, Unüberwachtes Lernen
Vorlesung 6	Überwachtes Lernen
Vorlesung 7	Neuronale Netze und Convolutional Neural Networks (CNN)
Vorlesung 8	Aufgabenstellung Projektarbeit Data Science, Case Study CNN: Malaria
Vorlesung 9	Aufgabenstellung Projektarbeit AI, Recurrent Neural Networks (RNN), Case Study RNN
Vorlesung 10	Status Projektarbeit Data Science und Fragen, Aufgabenstellung AI
Vorlesung 11	Status Projektarbeit AI, Ausblick

DATEN UND DATENQUALITÄT

ROADMAP

1. Datenbasierte Geschäftsmodelle: Daten haben Wert
2. Übersicht Workflow Datenmanagement
3. Datenformate: welche Daten werden benötigt?
4. Datenablage: wo werden die verarbeiteten Dateien gespeichert?
5. Datenerfassung und -transformation: wie werden die Daten verarbeitet?
6. Datenqualität: wie müssen die Daten sein, damit das Geschäftsmodell funktioniert?

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.

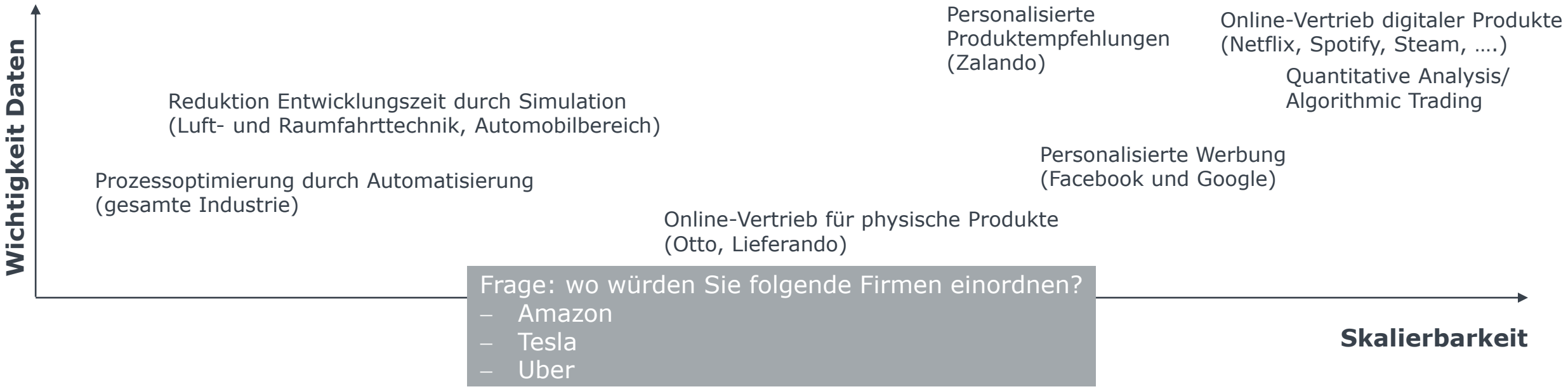
- **Data-informed[1] Geschäftsmodelle:** Optimierung bestehender Wertschöpfungsprozesse durch Daten.
  - Prozessoptimierung durch Automatisierung (gesamte Industrie).
  - Reduktion Entwicklungszeit/-kosten durch Simulation (Luft- und Raumfahrttechnik, Automobilbereich).
  - Online-Vertrieb für physische 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

1. DATENBASIERTE GESCHÄFTSMODELLE.



Stufe 1 Data-informed :
Optimierung bestehender Wertschöpfungsprozesse/ Geschäftsmodelle

Stufe 2 Data-infused:
Wertschöpfungsprozesse/ Geschäftsmodell wesentlich abhängig von Daten

Stufe 3 Data-Driven:
vollständig digitale Wertschöpfung/ Geschäftsmodell

Datenbasierte Geschäftsmodelle ermöglichen per Skalierung mehr Effizienz/ Profit bei gleichbleibender Kostenstruktur
„Data is the world's most valuable resource"[1]

Quelle: 1 Economist, 06.05.2017. Link

2. ÜBERSICHT WORKFLOW DATENMANAGEMENT

Rohdaten

...

Datenerfassung und –management/ Extract, Transform, Load (ETL)

Discover, Update or Initiate (Trigger) → Ingest/ Store raw data → Process/ Enrich Data → Store prepared data

Veröffentlichte Daten

Mainframes/ Server

Datenbanken, Data Warehouses

MySQL
ORACLE
mongoDB

Cloud Systeme

Spark
APACHE hadoop
Azure aws Google Cloud

Daten analysieren, Erkenntnisse generieren und visualisieren

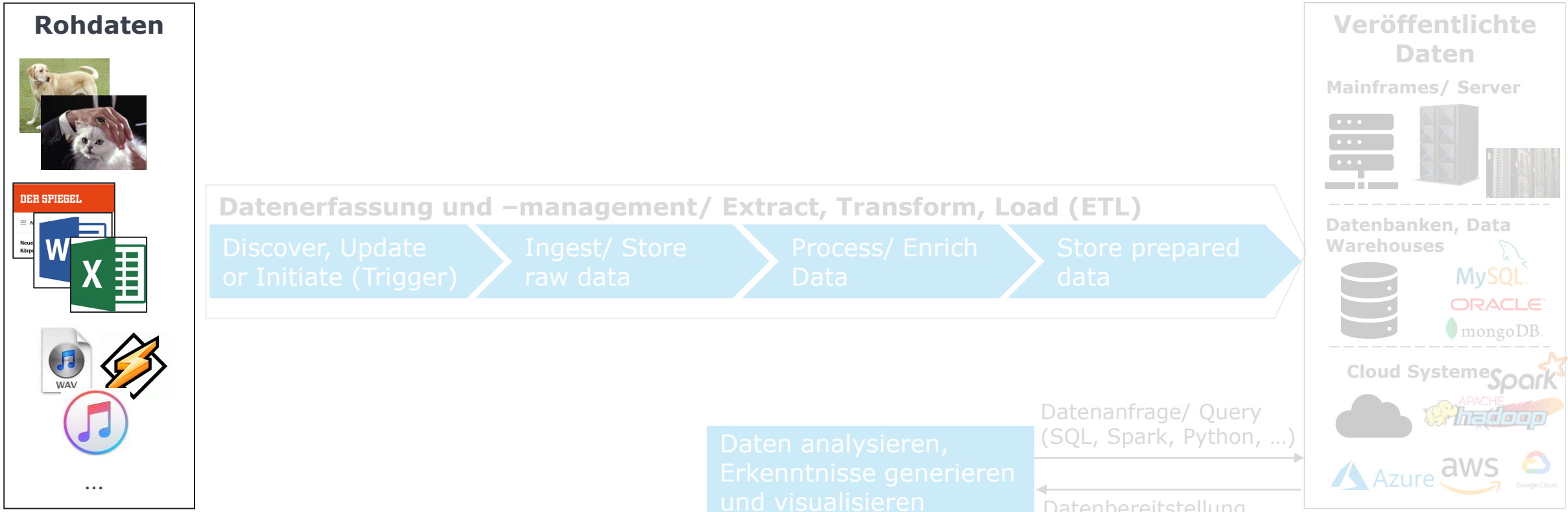
Datenanfrage/ Query (SQL, Spark, Python, …)

Datenbereitstellung (Tabellen, CSV, JSON, …)

Inhalt der nächsten Vorlesungen ☺

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

3. DATENFORMATE



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

3.1 AUDIODATEIEN.

Datentypen:

- Sprache
- Musik (iTunes, MP3, OGG, WAV, …)
- Geräusche

Datenstruktur:

- Binäre (nicht direkt lesbar) Datei,
- Größe abhängig von Sample rate (Frequenzen pro Sekunde) und bitrate (Abtastung in Bits pro Sekunde)

Typische Anwendungsgebiete:

- Spracherkennung (Siri, GoogleNow, Cortana)
- Computersprache (Amazon Polly)
- Automatisches Übersetzen/ Untertitel

Nicht im Fokus Vorlesung

3.2 BILDER/ VIDEO

Datentypen:

- Einzelne Bilder (RAW, JPEG)
- Video (MPEG)

Typische Anwendungsgebiete:

- Erkennen Inhalte eines Bildes
- Industrieroboter
- Autonomes Fahren

Beispiel JPEG-Datei: Speichern als Matrix Größe Anzahl x-/y-Punkte[1] * Farbtiefe[2]

Dog

Red channel

```
[[162 157 152 ... 132 125 123]
 [159 156 153 ... 135 133 133]
 [153 154 154 ... 136 139 140]
 ...
 [ 75  78 100 ...  64  84 135]
 [139 118 155 ...  72  92 141]
 [164 186 159 ...  74  91 121]]
```

Green channel

```
[[167 162 157 ... 137 130 129]
 [164 161 158 ... 140 138 139]
 [158 159 160 ... 144 147 148]
 ...
 [ 61  66  90 ...  71  90 139]
 [123 104 142 ...  78  98 142]
 [165 187 163 ...  78  87 115]]
```

Blue channel

```
[[101  96  93 ...  79  72  69]
 [ 98  95  94 ...  82  80  79]
 [ 92  95  96 ...  85  87  88]
 ...
 [ 24  28  54 ...   4  18  63]
 [ 87  67 107 ...  16  28  74]
 [133 155 128 ...   0  23  67]]
```


1 Anzahl Punkte gleichbedingt mit Begriff Pixel. Kamera mit 9 Mio. Pixel also 3000 x 3000 Punkte.
2 Häufig eingesetzt wird RGB mit 8 Bit/ 256 Farben je Farbkanal = 16,7 Mio. Farben
Quelle: Bild von Wikipedia-Artikel Hund (Link), Zerlegung in Farbkanäle per Python-Skript.

3.3 TEXTE

Datentypen:

- Strukturierte Texte (MS Office, Webseiten in XML/HTML, Social Media)
- Messdateien
- Unstrukturierte Texte
- …

Typische Anwendungsgebiete:

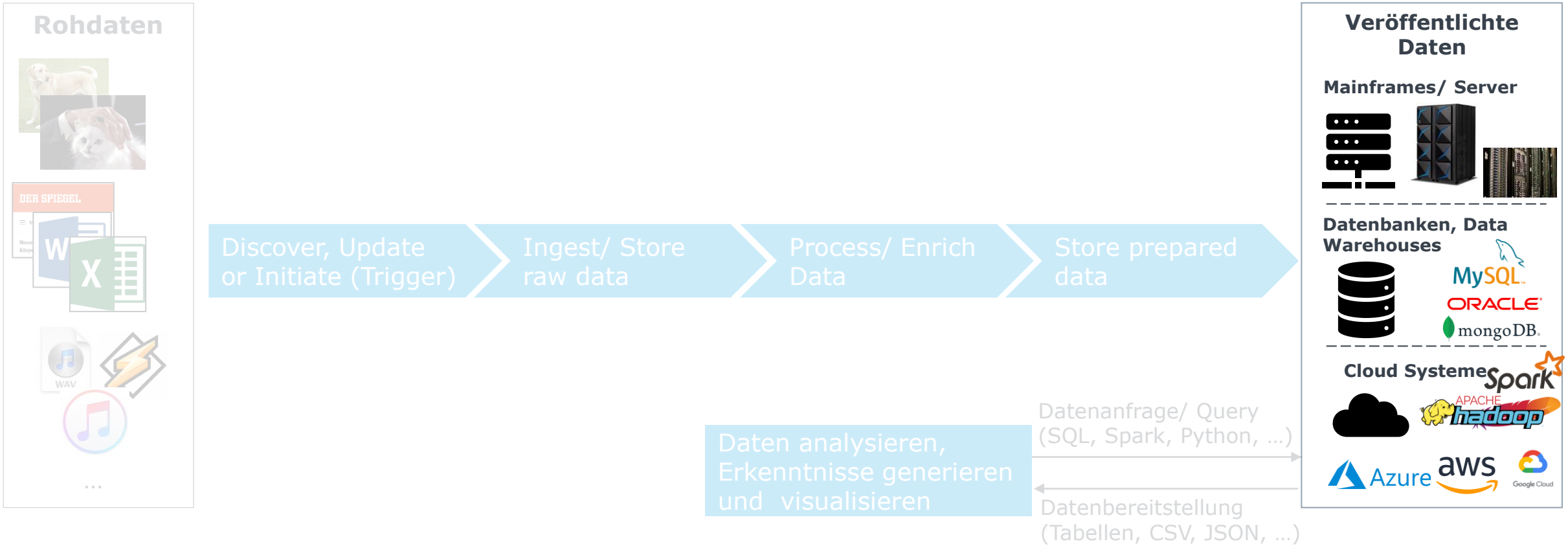
- Spamfilter
- Übersetzungen
- Suchanfragen

Beispiel Unicode UTF-8[1] (häufigste Codierung im Web[2])

Unicode-Zeichen	**Binäre Codierung**	**Was?**	**Beispiel**
U-00000000 – U-0000007F:	0*xxxxxxx*	Lateinisches Alphabet mit Satzzeichen ohne Umlaute	U+0021 → 100001 → ! U+0041 → 1000001 → A
U-00000080 – U-000007FF:	110*xxxxx* 10*xxxxxx*	Erweiterung um Sprachen mit Akzenten, Umlaute, ….	U+00A9 → 1100001010101001 → ©
U-00000800 – U-0000FFFF:	1110*xxxx* 10*xxxxxx* 10*xxxxxx*	Weitere Sprachen z.b. Chinesisch oder Japanisch	U+3231 → 11100011 10001000 10110001 → ㈱ U+4E76 → 111001001011100110110110 → 亶

Quellen: 1 https://www.utf8-chartable.de/, abgerufen am 07.09.2020
2 https://w3techs.com/technologies/details/en-utf8, abgerufen am 07.09.2020

4. DATENABLAGE



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

4. DATENABLAGE.

Hauptsächliches Unterscheidungskriterium: on-prem (vor Ort bei Person/Firma) oder Cloud.

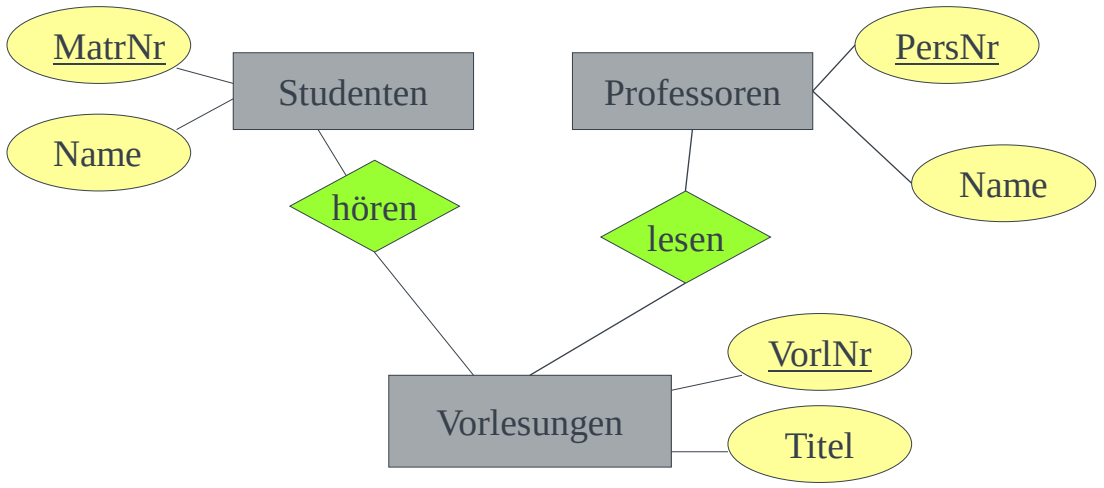
- [PC oder lokale Speichermedien]
- Mainframes
- Server im privaten oder Firmennetzwerk
- **relationale und nicht-relationale Datenbanken**
- **Data Warehouses** (Amazon RedShift, Snowflake, Google BigQuery, SAP, …)
- Cloud Systeme:
  - Buckets (Amazon S3, Azure Storage, …): enthält Objekte bis zu 5 TB Größe, Zugriff Web-Interface
  - Distributed Datasets (Spark, Hadoop, …): Daten (meist Tabellen) verteilt auf mehrere Systeme

Cloud Systeme ermöglichen (beliebige) Anpassung bereitgestellte Ressourcen an Nachfrage (elastic/ Skalierbarkeit), stellen aber höhere Anforderungen an Datensicherheit (DSGVO) und Datenhaltung (bspw. China)

Quelle Bilder: Stockpictures sowie Homepages der jeweiligen Hersteller.

4.1 DATENBANKEN: RELATIONALE DATENBANKEN.

Konzeptuelle Modellierung in Form von Beziehungen (Relationen)



Darstellung im sogenannten „Entity-Relationship“-Diagramm[1]

Speicherung der Daten in Form von Tabellen

Studenten	
MatrNr	Name
26120	Fichte
25403	Jonas
…	…

hören	
MatrNr	VorlNr
25403	5022
26120	5001
…	…

Vorlesungen	
VorlNr	Titel
5001	Grundzüge
5022	Glaube & Wissen
…	…

Bearbeiten der Daten mit SQL (Structured Query Language)

```
Select Name
From Studenten, hören, Vorlesungen
Where Studenten.MatrNr = hören.MatrNr
and hören.VorlNr = Vorlesungen.VorlNr
and Vorlesungen.Titel = `Grundzüge´;
```

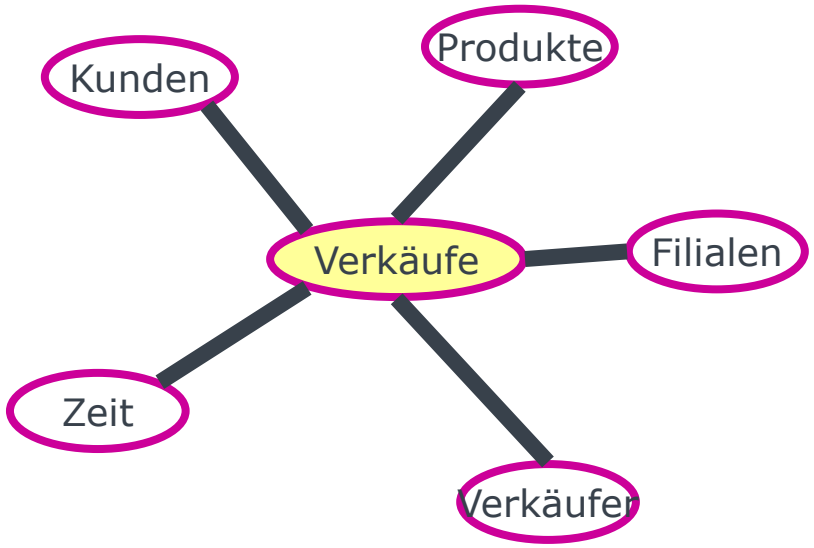
```
Update Vorlesungen
Set Titel = `Logik´
Where VorlNr = 5001;
```

Vorteile: weit verbreitet, robust, garantiert konsistente Daten und benötigt wenig Speicherplatz.
Nachteile: hoher Aufwand beim Speichern, Ändern und Abfragen von Daten → schlecht skalierbar.

Quellen: 1 Chen, „The Entity-Relationship Model“, 1976
Bilder, Tabellen und SQL-Statement aus: Kemper, Eickler. Datenbanksysteme - Eine Einführung, 2015.
MySQL und Oracle Logo von der jeweiligen Homepage entnommen.

4.2 DATENBANKEN: DATA WAREHOUSES.

Logische Struktur Data Warehouse (Star schema)



Speicherung der Daten im Data Warehouse

Produkte

ProduktNr	Produkttyp	Produktgruppe	Produkthauptgruppe	Hersteller	...
1347	Handy	Mobiltelekom	Telekom	Siemens	...
...	...	...	...	...	...

Kunden

KundenNr	Name	wiealt	...
4711	Kemper	43	...
...	...	...	...

Filialen

Filialenkennung	Land	Bezirk	...
Passau	D	Bayern	...
...	...	...	...

Verkäufe

VerkDatum	Filiale	Produkt	Anzahl	Kunde	Verkäufer
25-Jul-00	Passau	1347	1	4711	825
...	...	...	...	...	...

Zeit

Datum	Tag	Monat	Jahr	Quartal	KW	Wochentag	Saison	...
...	...	...	...	...	...	...	...	...
25-Jul-00	25	Juli	2000	3	30	Dienstag	Hochsommer	...
...	...	...	...	...	...	...	...	...
18-Dec-01	18	Dezember	2001	4	52	Dienstag	Weihnachten	...
...	...	...	...	...	...	...	...	...

Sehr große **Faktentabelle** enthält Daten des Geschäftsprozesses, verlinkt auf einzelne, kleine **Dimensionstabellen** mit den Daten.

Vorteile: schnelles Auswerten verschiedenster, zusammenhängender Daten.
Nachteile: Aufbereiten und Aktualisieren von Daten aufwendig. Hohe Wartungskosten. Struktur nur schwer änderbar.

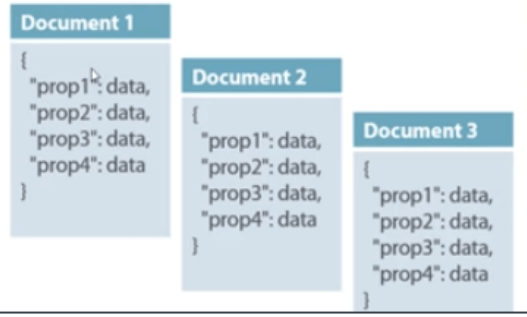
Quelle Bilder und Tabellen: Kemper, Eickler. Datenbanksysteme - Eine Einführung, 2015.
Amazon Redshift und Google BigQuery Logo von der jeweiligen Homepage des Herstellers entnommen.

4.3 DATENBANKEN: NICHT-RELATIONALE DATENBANKEN.

Key-Value

Key	Value
Name	Joe Bloggs
Age	42
Occupation	Stunt Double
Height	175cm
Weight	77kg

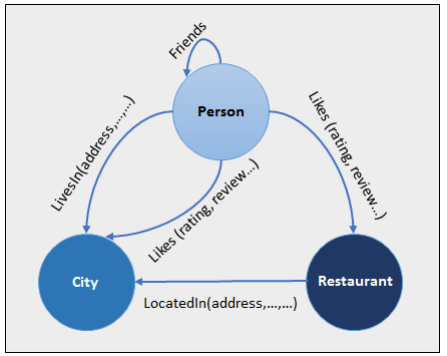
Document-oriented



Wide-Column Store

ColumnFamily			
Row Key	Column Name		
	Key	Key	Key
	Value	Value	Value
	Column Name		
	Key	Key	Key
	Value	Value	Value

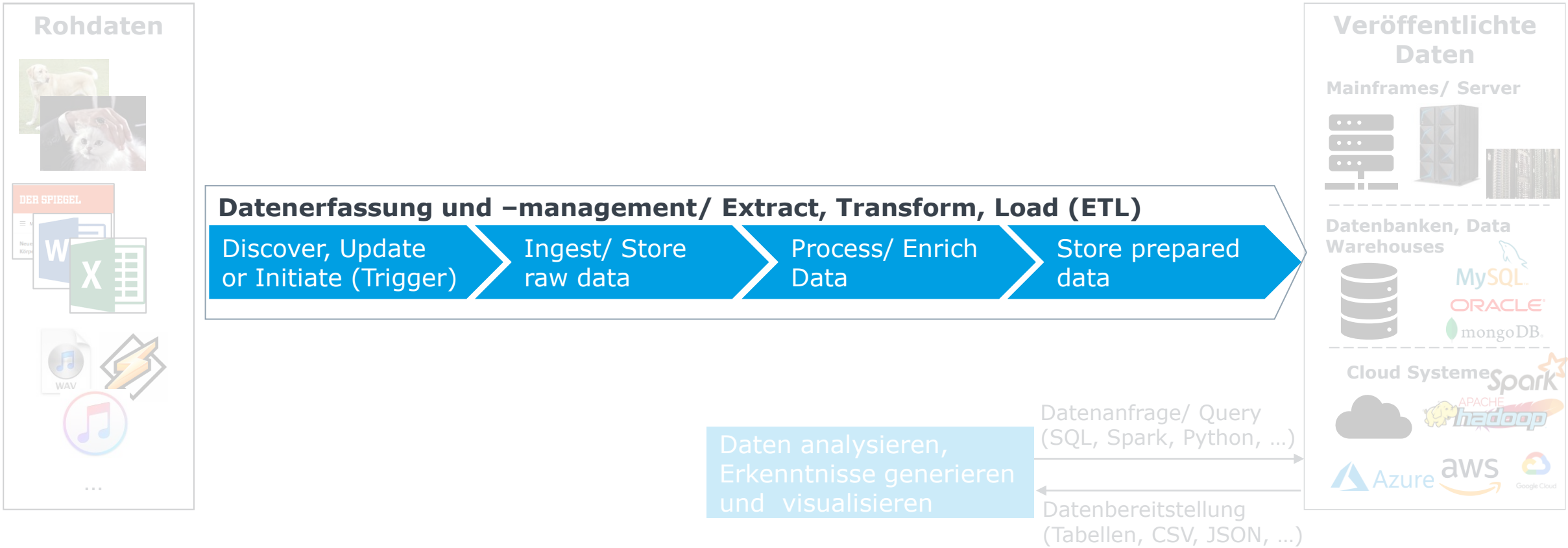
Graph-Based



Vorteile: sehr schneller Lese- und Schreibzugriff auch bei sehr großen Datenmengen. Ausfallsicher durch Replikationen.
Nachteile: Daten können inkonsistent oder veraltet sein. Fixes Datenschema, keine Verknüpfungen Daten möglich!

Quelle Bilder: NoSQL Tutorial, verfügbar unter: Link, abgerufen am 12.09.2021

5. WORKFLOW DATENERFASSUNG- UND MANAGEMENT/ ETL (EXTRACT, TRANSFORM, LOAD).



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

ETL IM DETAIL.

Rohdaten

...

Datenerfassung und –management/ Extract, Transform, Load (ETL)

Discover, Update or Initiate (Trigger) → Ingest/ Store raw data → Process/ Enrich Data → Store prepared data

periodisch, ereignis- oder anfragegesteuert

Daten temporär zwischenspeichern

- auf Fehler prüfen
- **Datenqualität sicher-stellen**
  - Duplikate/ Daten löschen
  - Fehlende Daten fixen
  - Daten anpassen (einheitliche Benennung, Währung, anonymisieren, ….)
- Daten verknüpfen
- Metadaten hinzufügen
- Logfiles oder Audit reports (Compliance!) schreiben

Aufbereitete Daten in Zielquelle(n) speichern

Veröffentlichte Daten

Mainframes/ Server

Datenbanken, Data Warehouses

MySQL ORACLE mongoDB

Cloud Systeme

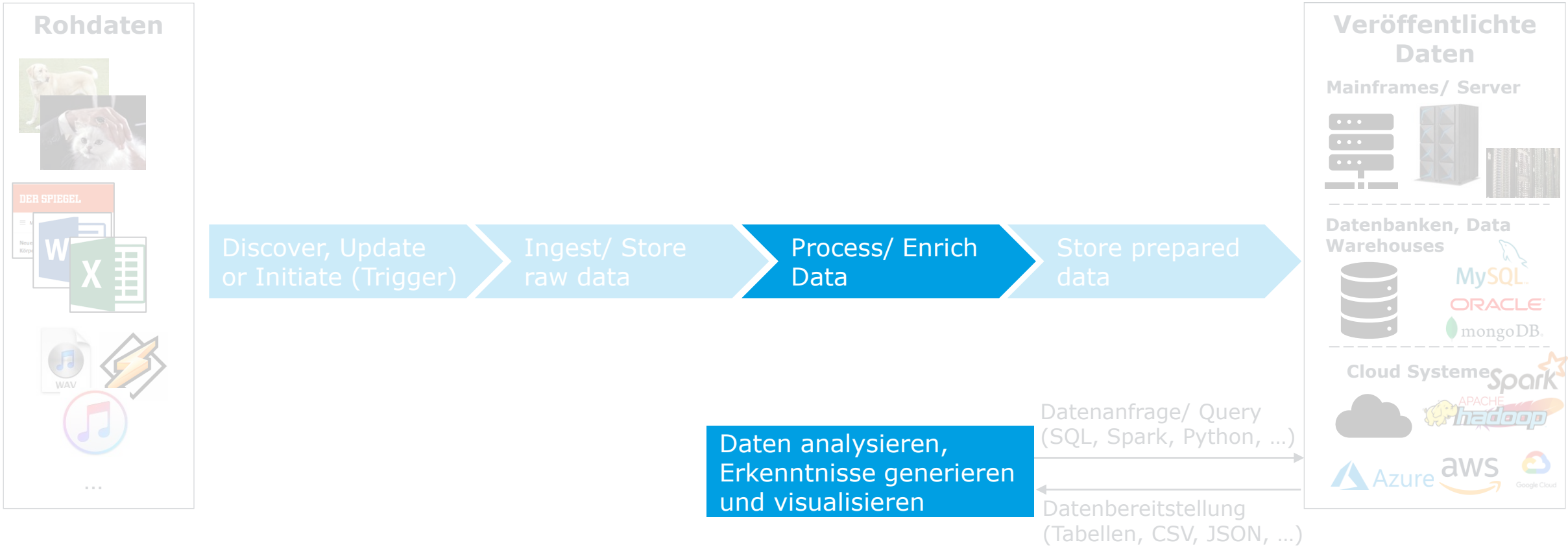
Spark APACHE hadoop

Azure aws Google Cloud

Daten aus mehreren Datenquellen extrahieren, an (Geschäfts-)Bedürfnisse anpassen und in neuer Quelle ablegen

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.

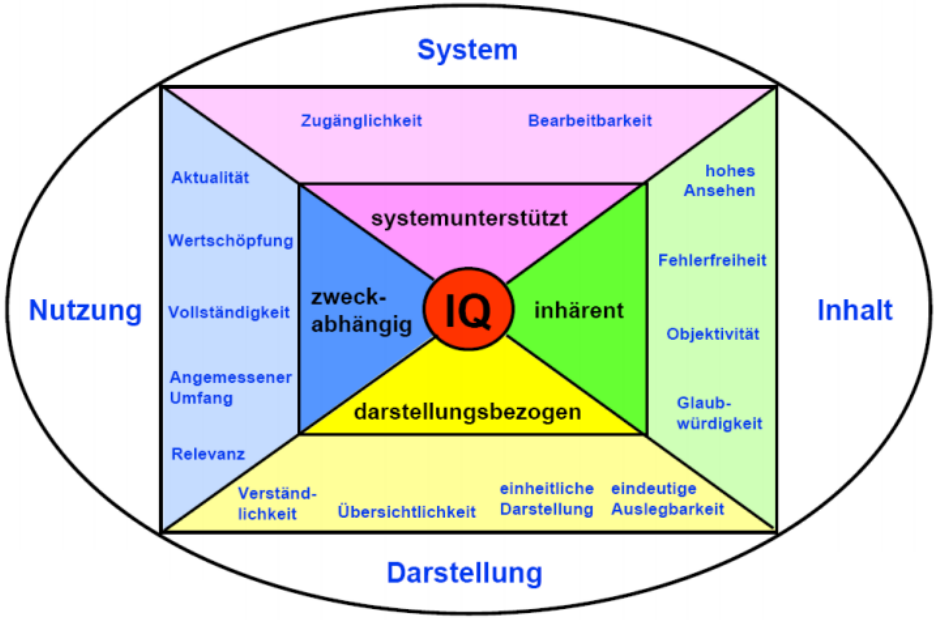
6. DATENQUALITÄT – EINFÜHRUNG.

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 IN GRUPPENARBEIT.

Wählen Sie für eine beliebige 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

Einheitl. 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 ein 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.

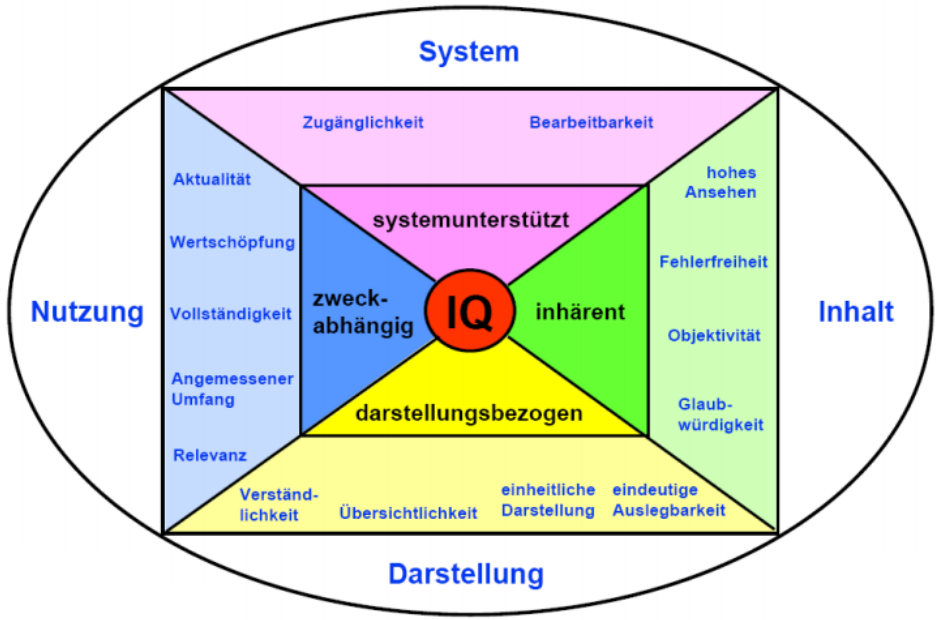
	Kunden-ID	Name	Geboren	Alter	Adresse	Kreditkartennummer	Einkäufe 2020	Umsätze 2020
Regel für Sicherstellen Datenqualität	ID definiert und eindeutig (d.h. darf max. 1 mal vorkommen)	Liegt vor	Geburtsdatum in europäischem Format: TT.MM.YY., sonst umwandeln	Alter < 120	muß vorliegen	1. 12 ≤ Anzahl Ziffern ≤ 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.



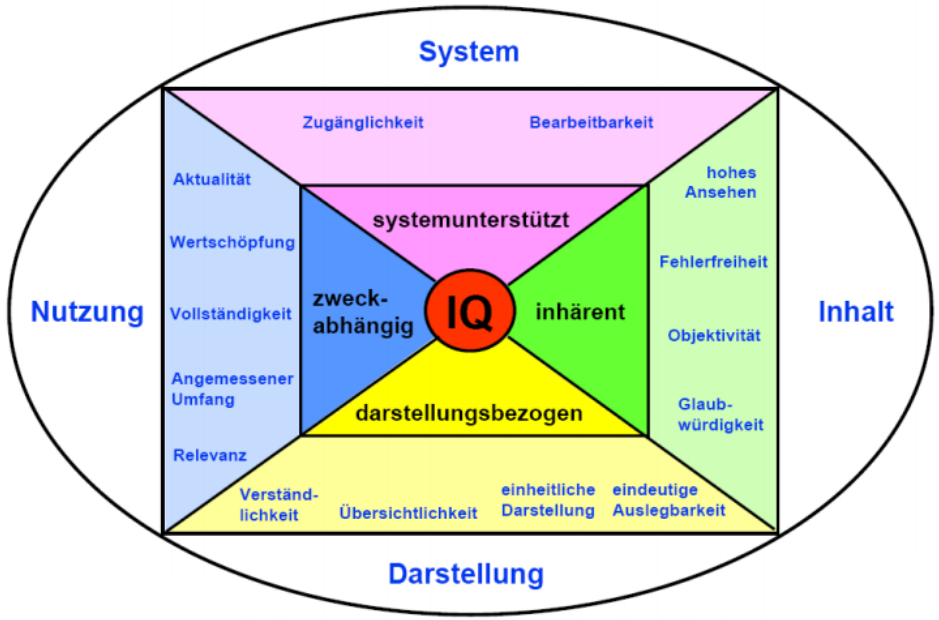
Informationen haben…

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.

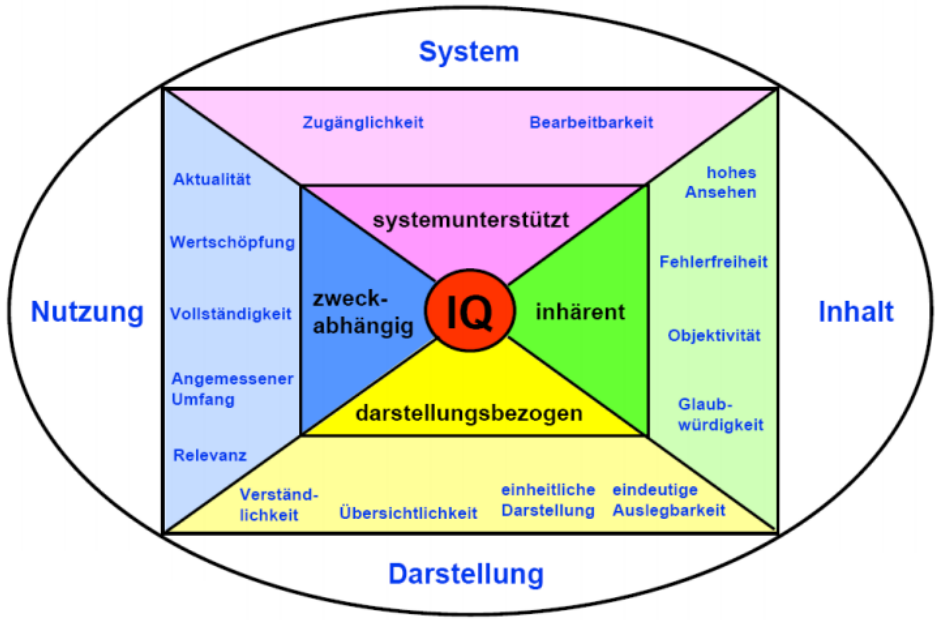
Fehlerfreiheit (free of error): wenn sie mit der Realität übereinstimmen.

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

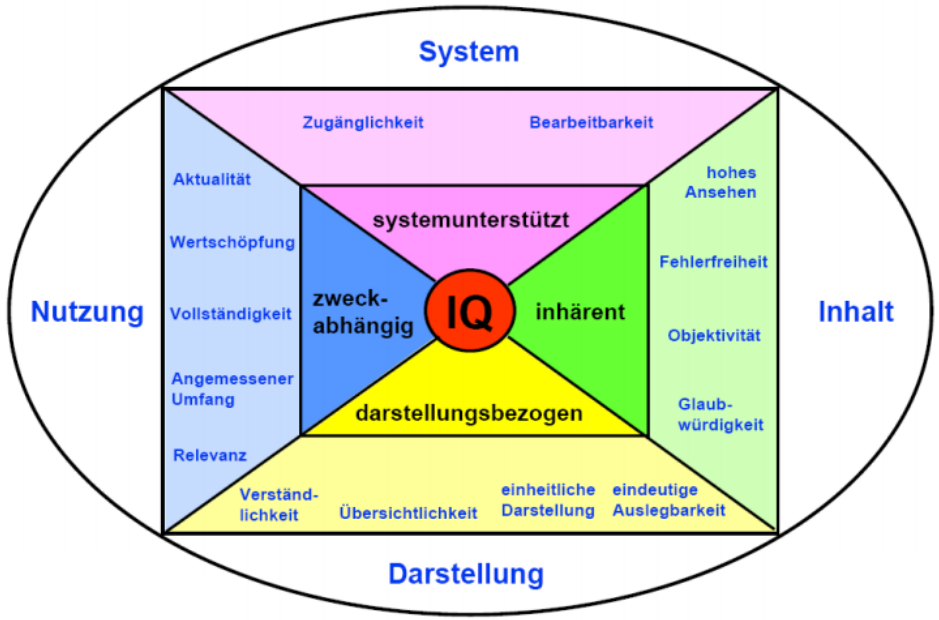
Einheitl. Darstellung (consistent representation): wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden.

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

Die Business Canvas[1] ist Bestandteil der Lean Startup

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?
- Welche Kriterien für Datenqualität sind wichtig?

→ Erster Schritt Richtung Durchführung Data Science/ AI Use Case