Von Chaos zu Erfolg

Datenqualität beherrschen

Sebastian Löfflath & Tim Bossenmaier

Sebastian Löfflath

Cloud Architect

@

Sebastian Löfflath

sebastianloefflath

Tim Bossenmaier

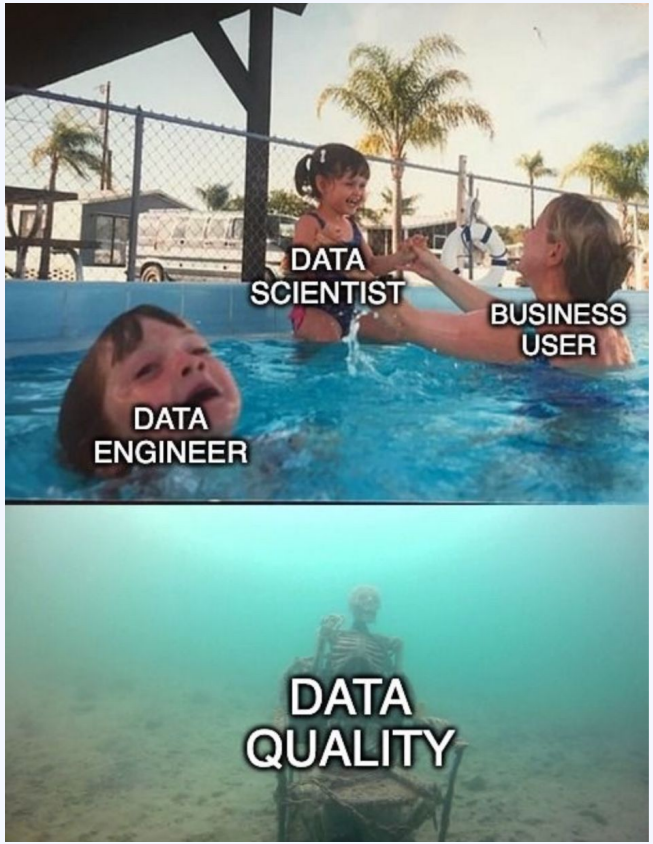
Data & Software Engineer

@ Bytefabrik.AI

Tim Bossenmaier

bossenti

Motivation



Nur drei Prozent der Unternehmensdaten erfüllen Qualitäts-standards[1]

[1] Nagle,T., Redman, T., Sammon, D. (2017). Only 3% of companies' data meets basic quality standards. Harvard Business Review. https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards

Agenda

- Was ist Datenqualität?
- Datenqualität sicherstellen: Cui Bono?
- Maßnahmen & Strategien

Was ist Datenqualität?

Keine einheitliche Definition

“conformance to requirements”
(Crosby 1979)

“We define high-quality data as data that is fit for use by data consumers”
(Strong 1997)

“data is [...] of high quality [...] if they are fit [...] in operations, decision making and planning”
(Redman 2016)

Unser Verständnis

“**Datenqualität** bezieht sich auf die **Fähigkeit** von Daten, die r**eale Welt widerz-uspiegeln** und die **Entscheidungs-findung zu unterstützen**.”

Dimensionen von Datenqualität nach DAMA

Vollständigkeit

Grad des Vorhandenseins der Attribute einer Datenentität (bspw. Tabelle).

- Wie viele der Attribute sind gefüllt (!= NULL)?
- in Abhängigkeit des konkreten Anwendungsfalls

ID	Name	Industrie	Umsatz (Mio USD)
1	Walmart	Retail	611 289
2	Amazon	null	513 983
3	Exxon	null	null

Name	100 %
Industrie	33,3 %
Umsatz	66,7 %

Kardinalität...

beschreibt den Grad an existierenden Duplikaten.

verschiedene Grade existieren nebeneinander:

- **Hoch**: Attribute haben einen hohen Anteil an einmaligen Werten
  Beispiele: IDs, E-Mail Adressen, Handy-Nummern, ...
- **Mittel**: Attribute mit seltenen Duplikaten
  Beispiele: Postleitzahl, Produktgruppen, Kostenstelle, ...
- **Gering**: Attribute mit häufigen Duplikaten
  Beispiele: Statuscodes, Flags, ...

Sicherheit

Data Security beinhaltet den Schutz von Daten
vor zerstörerischen Akteuren.

- inkls. Schutz vor unautorisierten Zugriffen durch Benutzer & Systeme
- betrifft den kompletten Lebenszyklus der Daten
- Bedrohungen wie Hackerangriffe, Betrug oder Malware
- Daten stehen jedem im Unternehmen zur Verfügung, der Zugang zu ihnen benötigt

Rechtzeitigkeit...

bezieht sich auf die Aktualität der Daten und ob die aktuellste Version verfügbar ist, wenn sie gebraucht wird

- Grad zu dem Daten mit der realen Welt übereinstimmen
- hängt von der Kritikalität des Geschäfts und den Auswirkungen ab
- Gewinnt zunehmend an Bedeutung, entscheidend für echtzeitnahe Entscheidungsfindung

Wieso gibt es Probleme mit der Datenqualität?

Ursachen für mangelnde Datenqualität I

Manuelle Fehler

Größte Quelle für schlechte Datenqualität: manuelles Pflegen/ Eintragen von Daten

Fehlende Anreize

kaum Anreize, die Verantwortung für die Datenqualität zu übernehmen

Analyse-Fokus

Daten-Teams: SQL und Python

Icons von freeicons.io. Creator: Raj Dev, Aficons, ColourByteDesigns

Ursachen für mangelnde Datenqualität II

Organisations-Silos führen zu Daten-Silos

fehlender abteilungsübergreifender Austausch

Unterschiedliche Nutzung & Interpretation von Daten

Verwendung entgegen des ursprünglichen Erhebungszwecks

Datenqualität nimmt über Zeit ab

Unternehmen ändern und entwickeln sich über die Zeit und mit ihnen die produzierten Daten

Icons von freeicons.io. Creator: manshagraphics, Hilmy Abiyyu Asad, MD Badsha Meah


Shadow Data Teams
Duplicate Data Warehouse
Marketing buying their own data analytics solutions
etc
Data Governance

Ursachen für fehlende Datenqualität - Takeaway

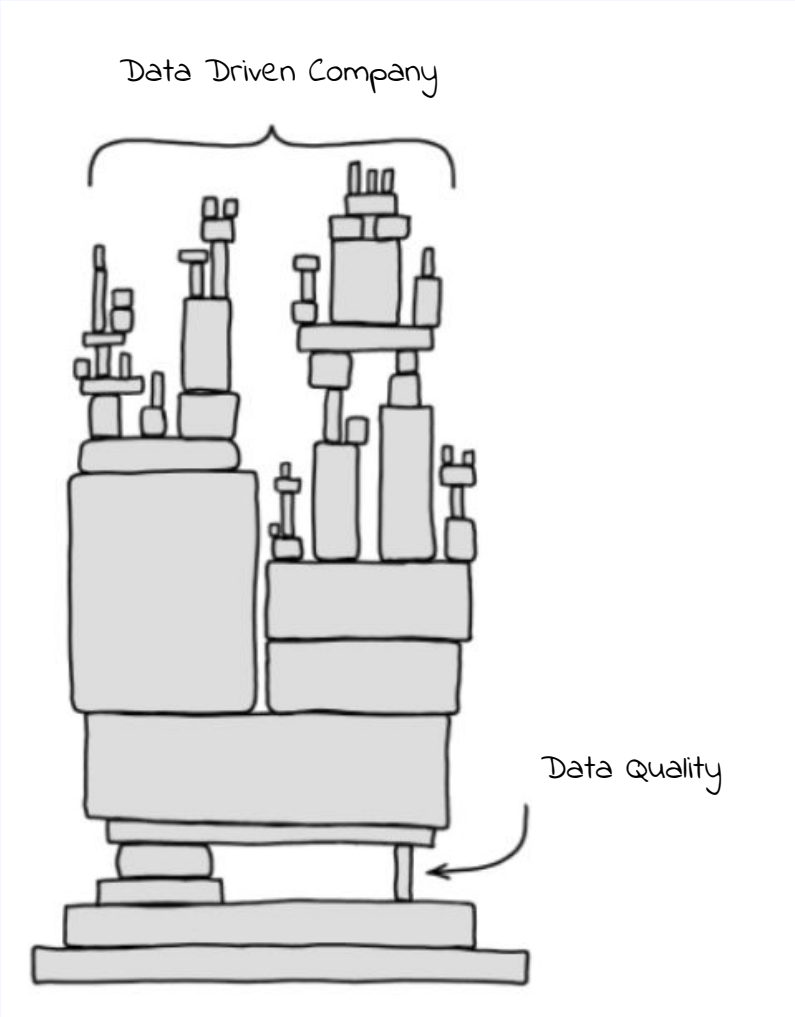
Ursachen sind vielfältig

Es gibt nicht die eine Stelle an der nachgebessert werden kann.

Datenqualität ist eine Unternehmensaufgabe

... und keine IT Angelegenheit. Datenqualität erfordert, dass Unternehmen Verantwortung übernehmen und Verbesserungen vorantreiben.

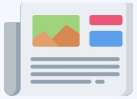
Datenqualität sicherstellen: Cui Bono?


Data Driven Company
Data Quality

Auswirkungen mangelnder Datenqualität

Finanzielle Auswirkungen
erhöhte operative Kosten, verlorener Umsatz, verpasste Möglichkeiten

Auswirkungen auf den Geschäftsbereich
Verfehlen von Erwartungen, Rückgang von Vertrauen, Fehlentscheidungen



Auswirkungen auf die Produktivität
Effizienz leidet, erhöhte Zykluszeiten

Compliance Auswirkungen
erhöhte finanzielle und compliance Risiken, Verstoß gegen Rahmenbedingungen

Icons von freeicons.io. Creator: manshagraphics, Hilmy Abiyyu Asad, MD Badsha Meah

Kosten von Datenqualität - Prävention lohnt sich

1 - 10 - 100 Regel

Maßnahmen & Strategien

Datenqualität beherrschen


sichtbare Qualitätsprobleme
verdeckte Qualitätsprobleme

Data Profiling

sichtbare und verdeckte Qualitätsprobleme finden

Data Profiling

OSS Tools wie Deequ und GreatExpectations unterstützen beim Profiling

Basic Daten KPIs

#Zeilen, #Spalten, Dateigröße,

%NULL-Werte, %Leerwerte,

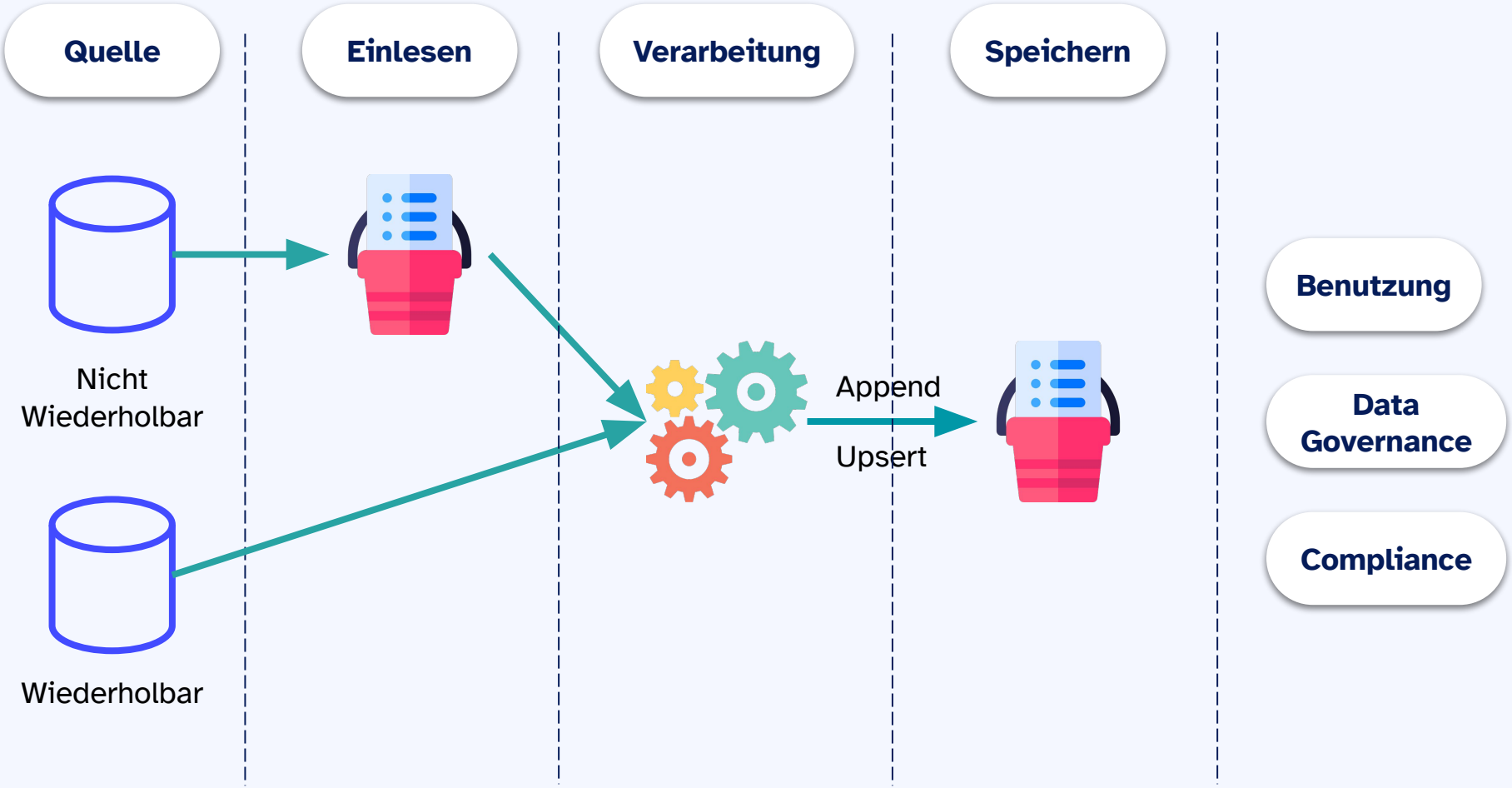
#einzigartigeElemente, Duplikate des Primary Keys, ...

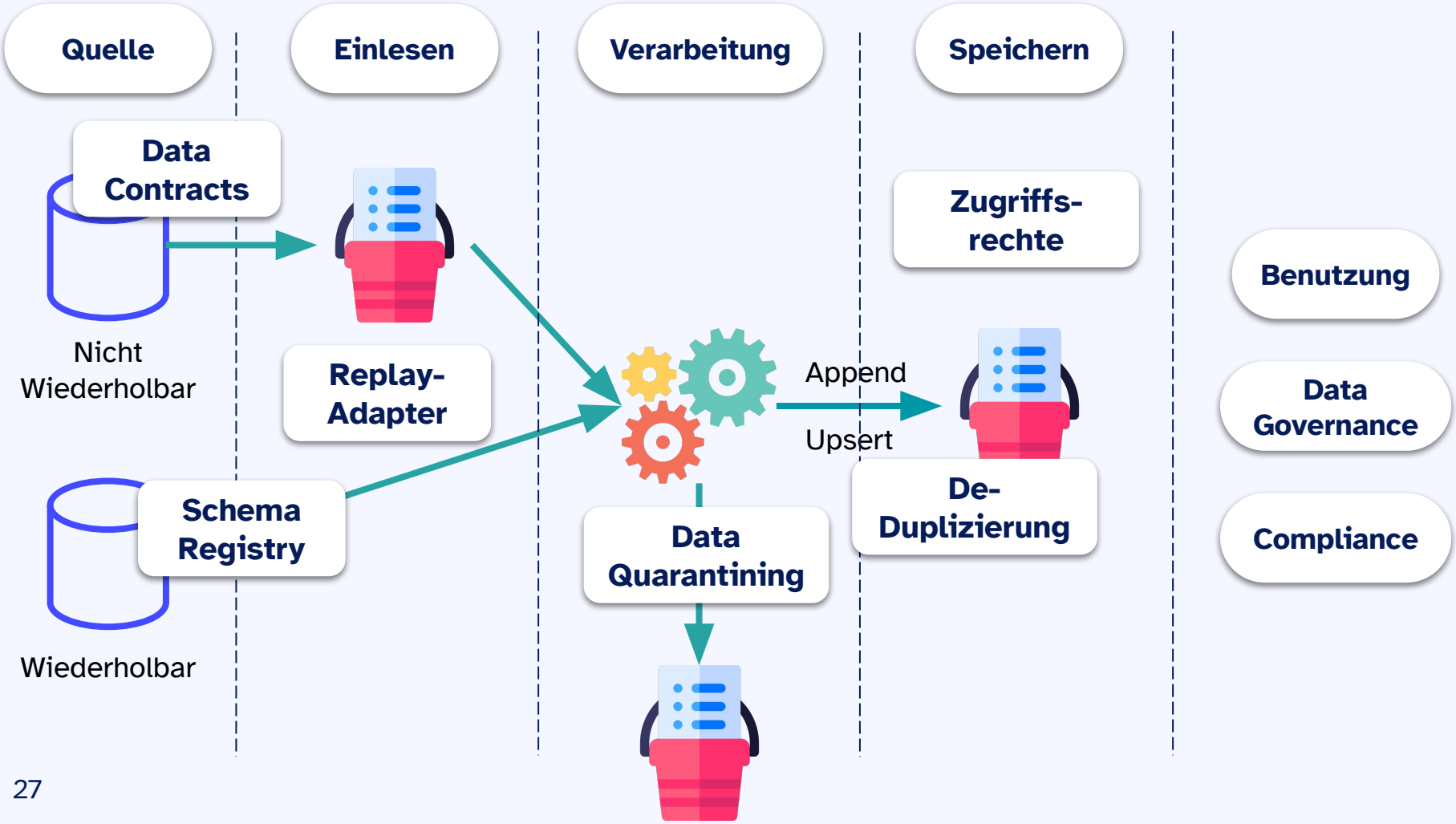
Kategorische Daten KPIs

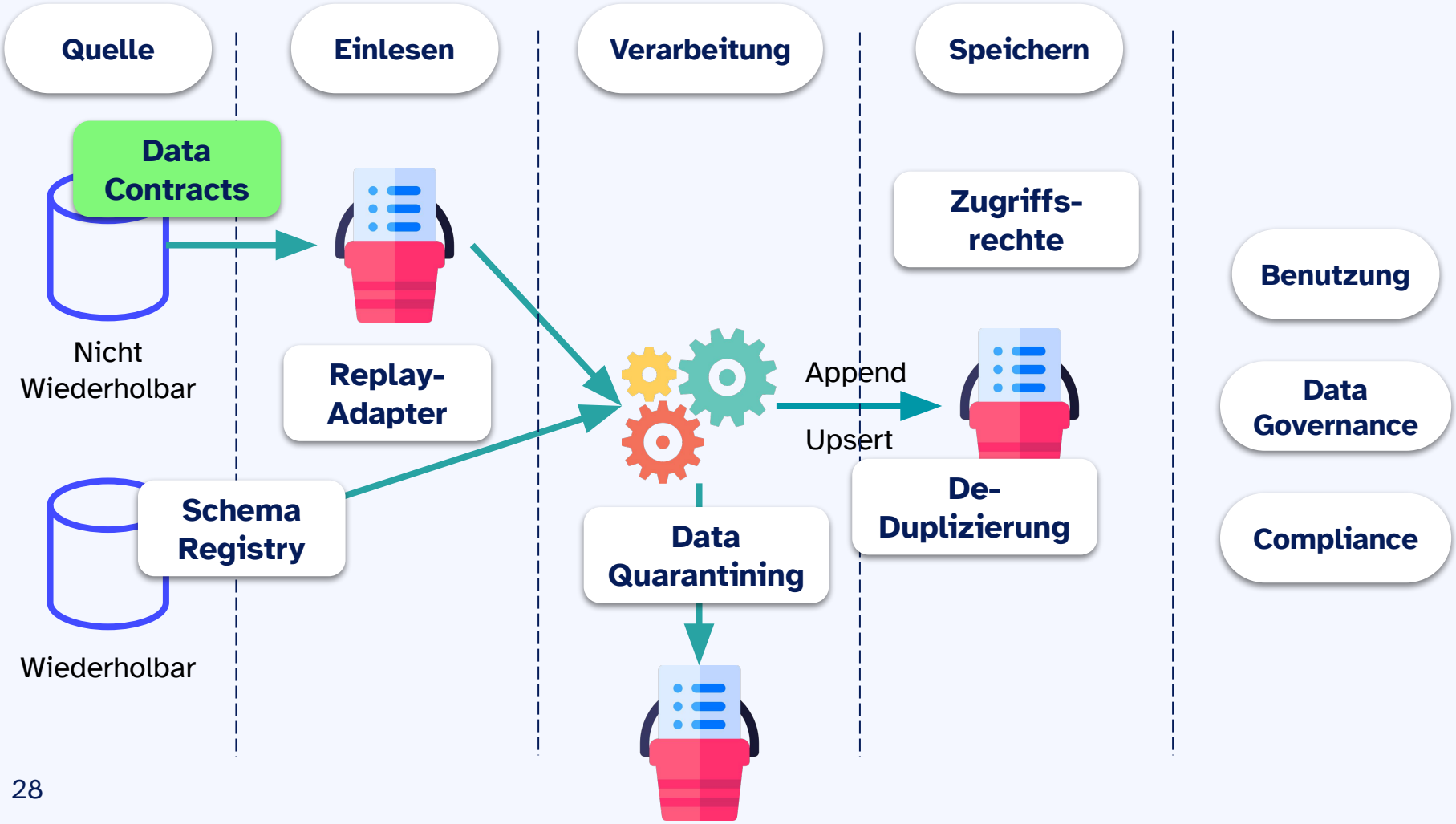
#Kategorien, Histogram, ...

Numerische Daten KPIs

Mittel, Min, Max, Varianz, ...


Quelle
Einlesen
Verarbeitung
Speichern
Nicht Wiederholbar
Wiederholbar
Append
Upsert
Benutzung
Data Governance
Compliance


Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Zugriffs-rechte
De-Duplizierung
Data Quarantining
Benutzung
Data Governance
Compliance


Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Zugriffs-rechte
De-Duplizierung
Data Quarantining
Benutzung
Data Governance
Compliance

Data Contracts

Data Contracts sind Vereinbarungen zwischen Softwareingenieuren, die über eigene Dienste verfügen, und Datenkonsumenten, die wissen, wie das Unternehmen funktioniert, um gut modellierte, hochwertige und vertrauenswürdige Daten zu generieren.

- Aufbrechen von Silos zwischen Datenproduzenten und Datenkonsumenten
- Konsumenten definieren Schema gemäß ihren Anforderungen
  - losere Kopplung
- kann sich im Laufe der Zeit ändern unter Maßgabe der Kompatibilität
- sind in erster Linie ein kultureller Wandel hin zu einer datenzentrierten Zusammenarbeit

Data Contracts

backward-compatible
Schema update zuerst im Konsument

erlaubte Änderungen des Data Contracts:

- hinzufügen von optionalen Feldern
- löschen eines Feldes

Data Contracts

forward-compatible
Schema update zuerst im Produzent

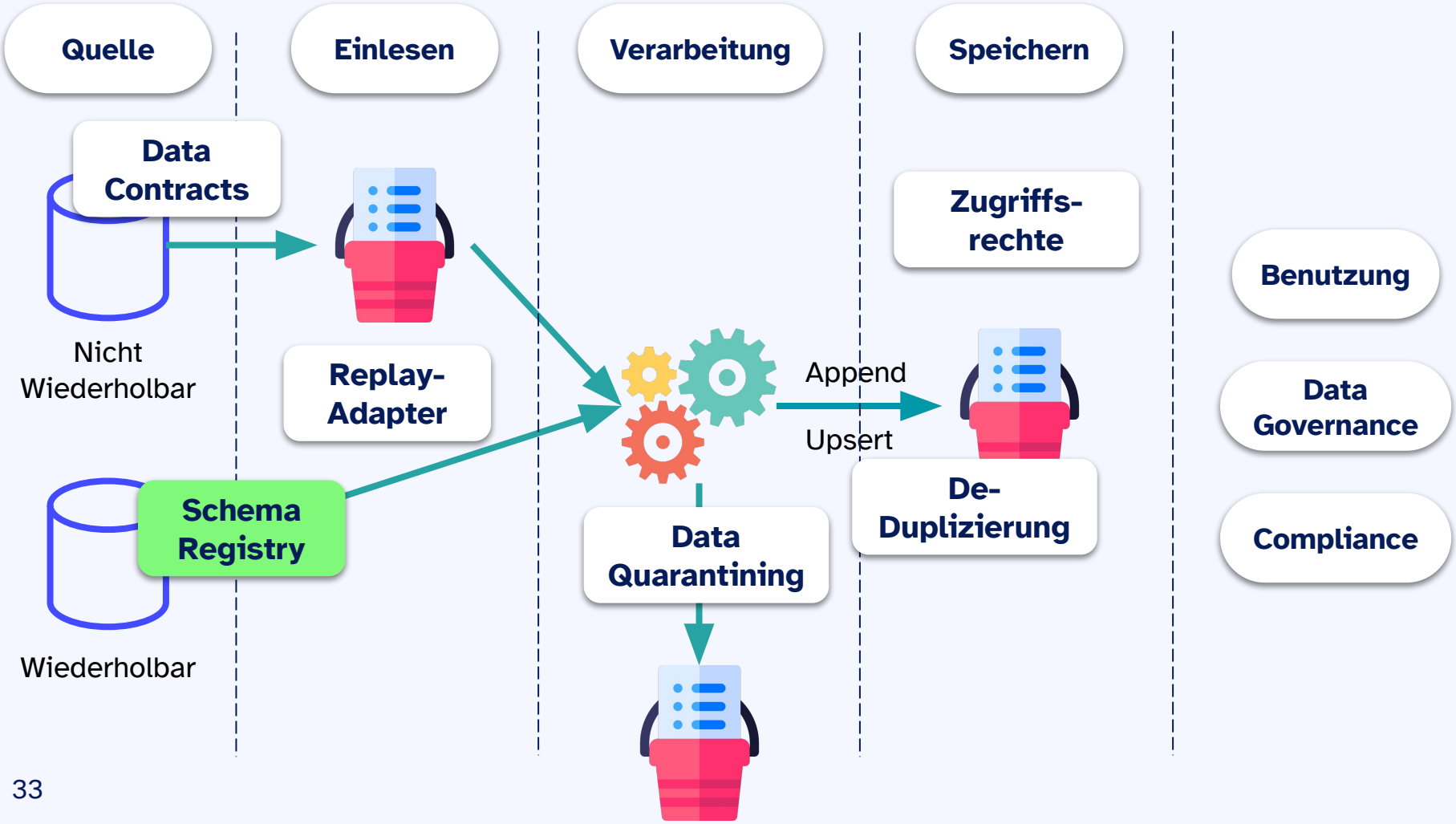
erlaubte Änderungen des Data Contracts:

- hinzufügen von Feldern
- löschen von optionalen Feldern

Data Contracts - Tools

Eigener Ansatz

Bestehendes Produkt


Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Zugriffs-rechte
De-Duplizierung
Data Quarantining
Benutzung
Data Governance
Compliance

Schema Registry...

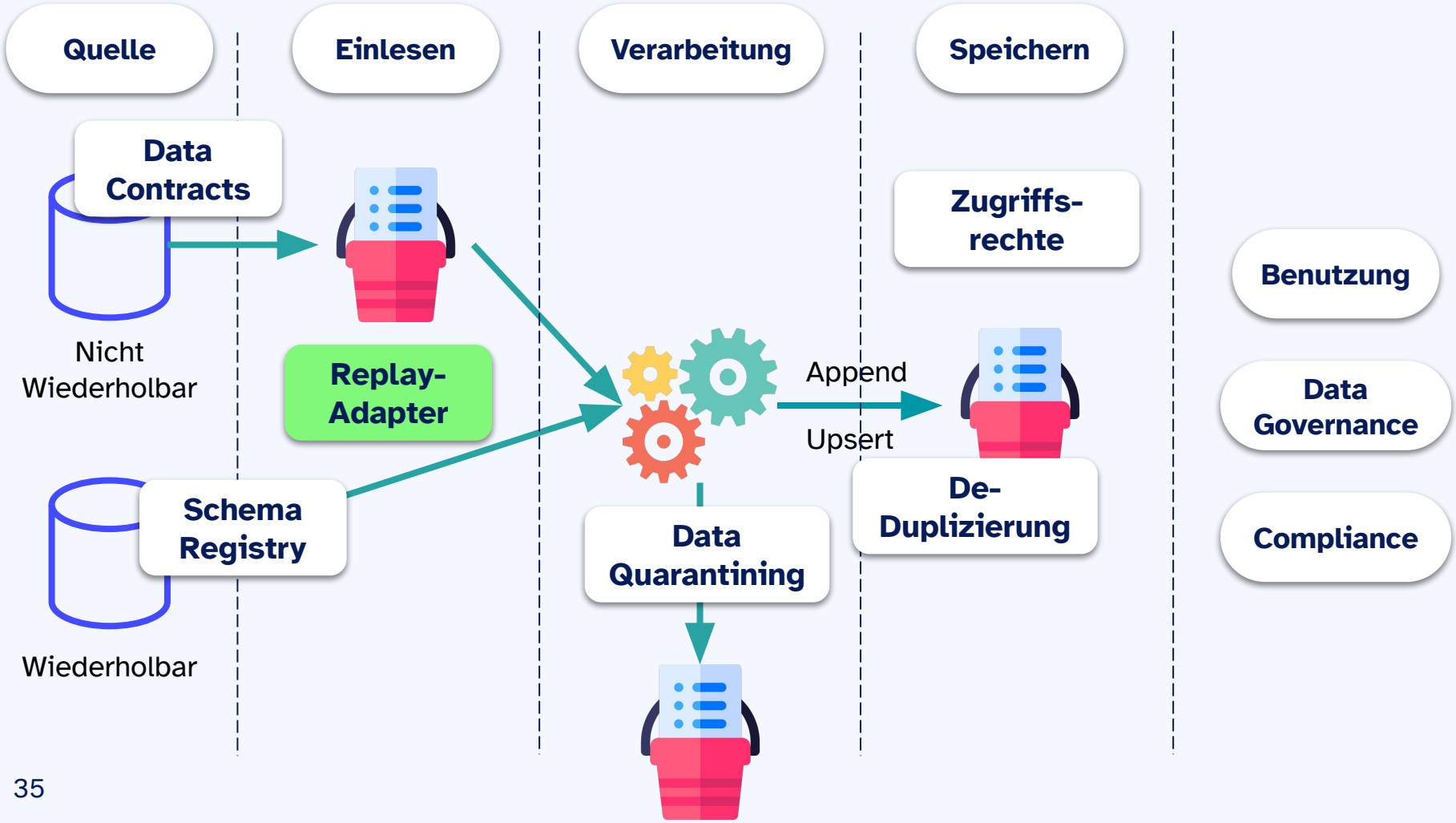
ist zentraler Speicher für die Verwaltung und Validierung von Schemata für thematische Nachrichtendaten und für die Serialisierung und Deserialisierung der Daten.

- festgelegtes Schema auf das sich Konsumenten und Produzenten verlassen
- Konsument kennt immer das dazugehörige Schema
- sorgt für Konsistenz der Daten
- Versionierung der Schemata ermöglicht Kompatibilität
- Validierung der Schemata beim Schreiben

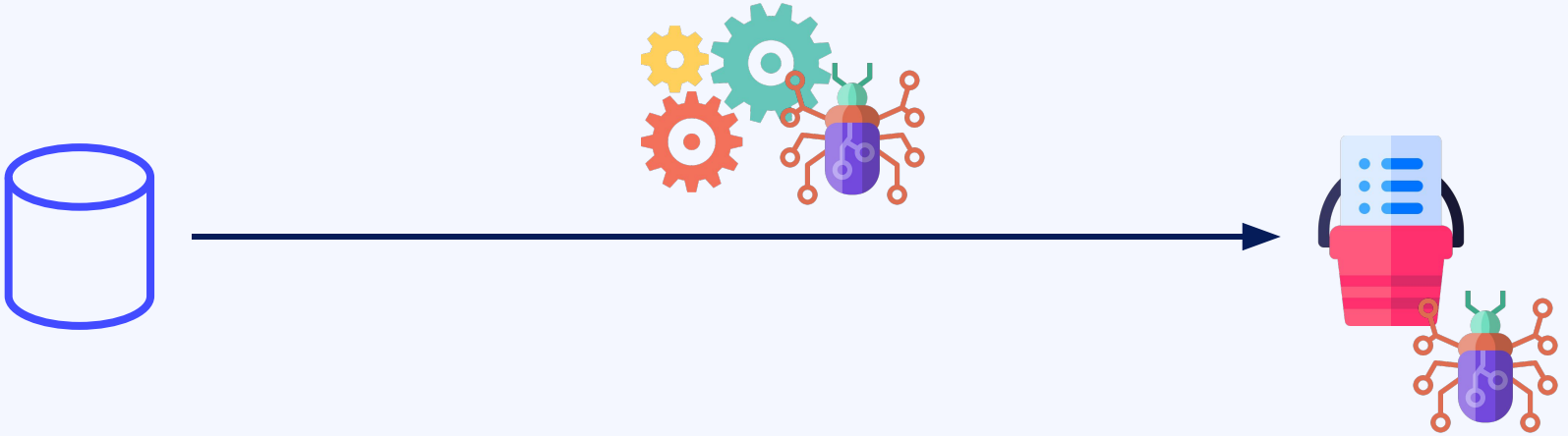
Confluent Schema Registry

AWS Glue Registry

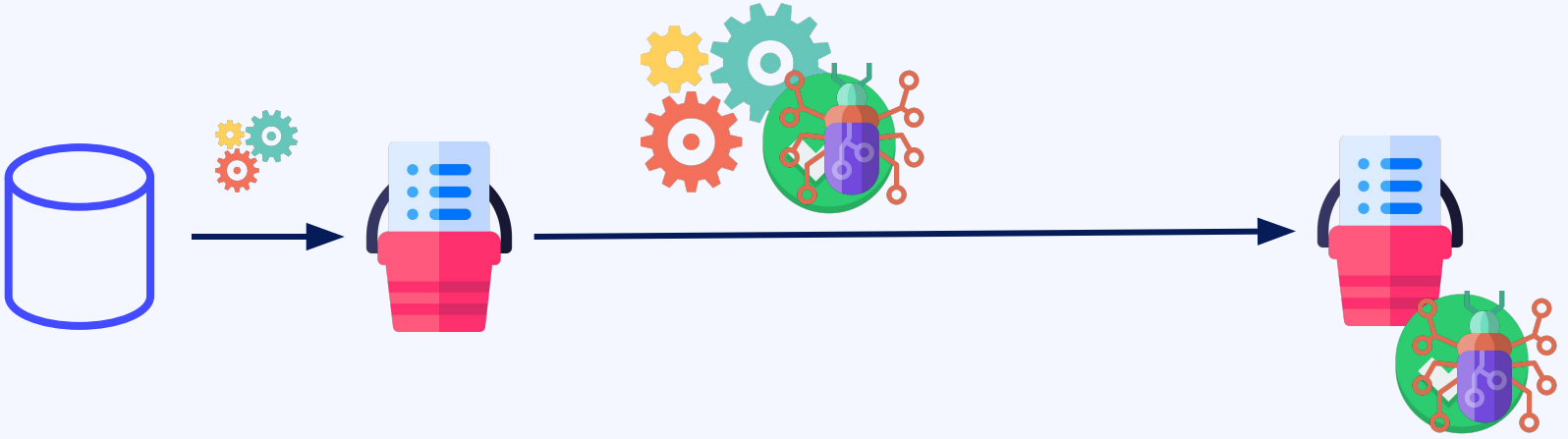
GCP Pub Sub

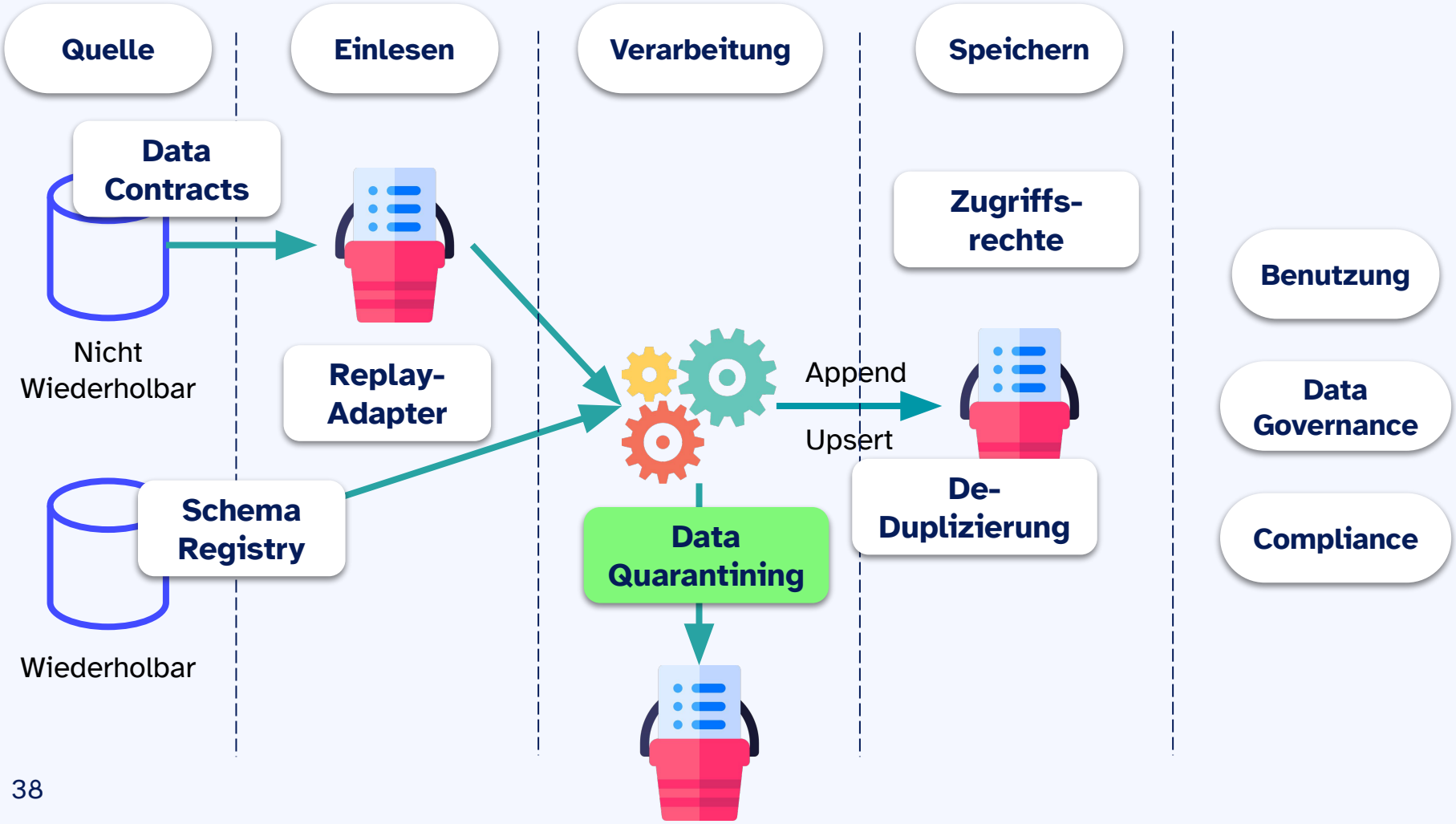

Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Zugriffs-rechte
De-Duplizierung
Data Quarantining
Benutzung
Data Governance
Compliance

Replay - Adapter

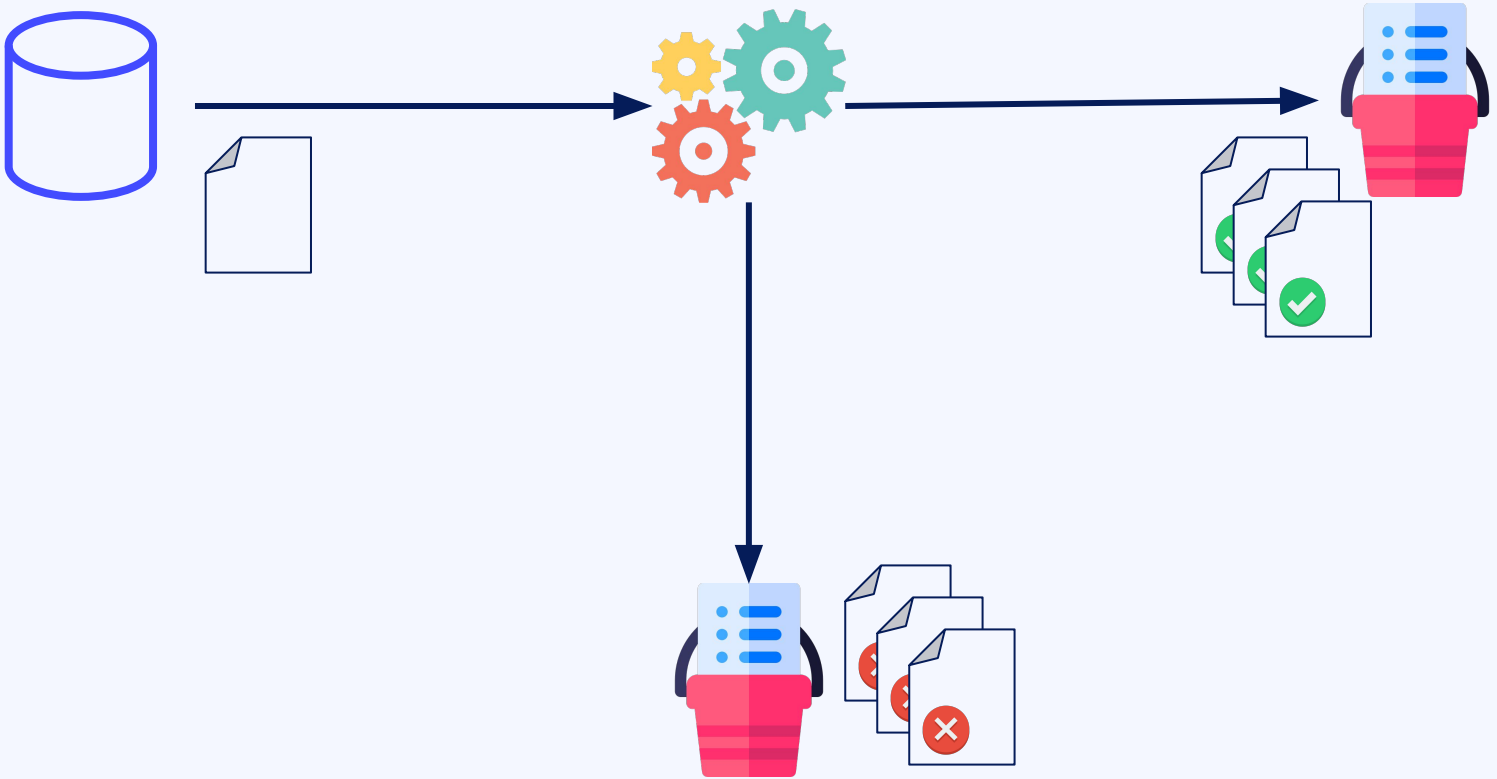


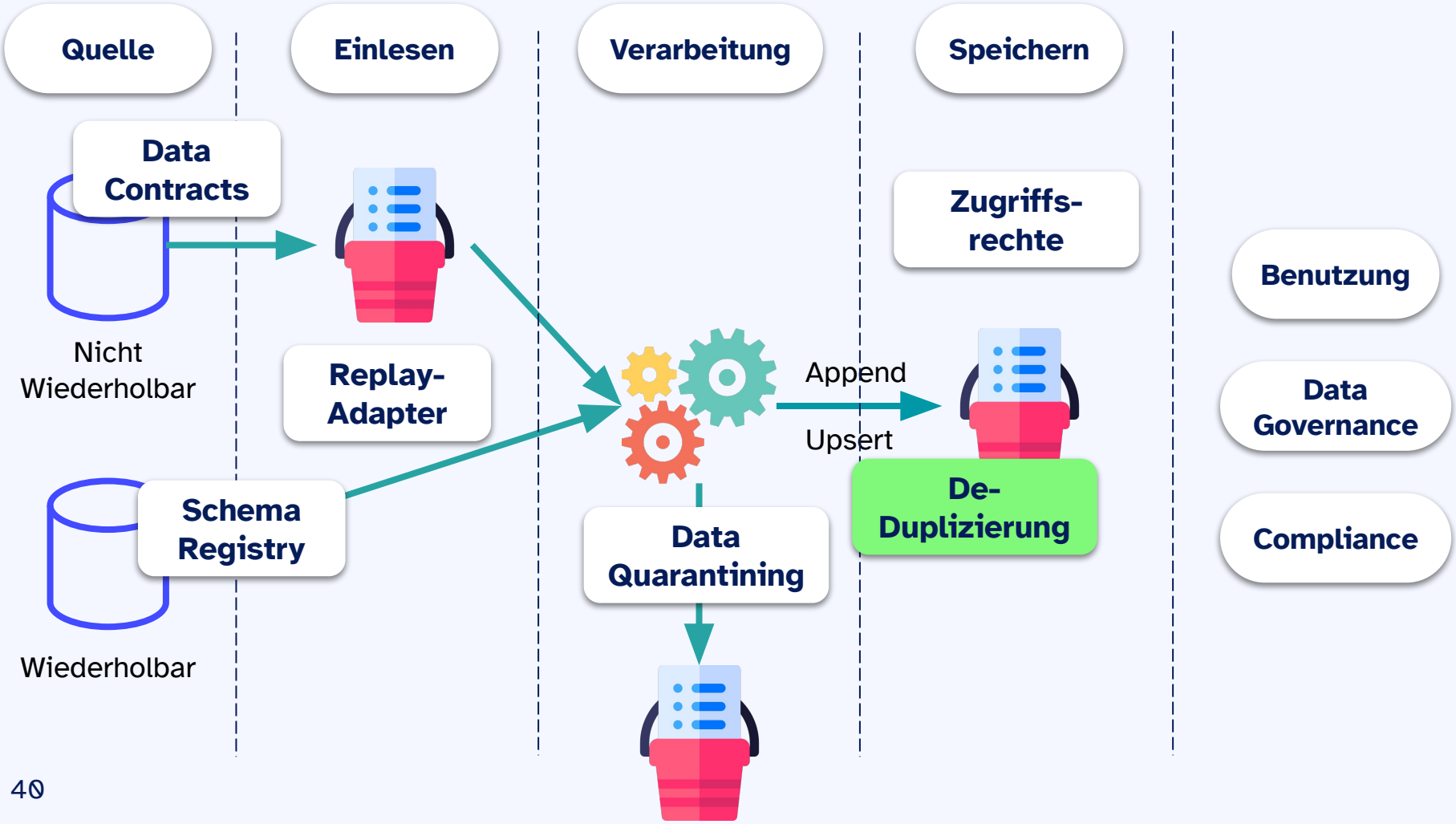
Replay - Adapter




Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Data Quarantining
Zugriffs-rechte
De-Duplizierung
Benutzung
Data Governance
Compliance

Data Quarantining




Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Zugriffs-rechte
De-Duplizierung
Data Quarantining
Benutzung
Data Governance
Compliance

De-Deduplizierung

Häufiges Problem bei event-basierten Datenquellen: at-least-once Garantie

Methoden für De-Duplizierung:

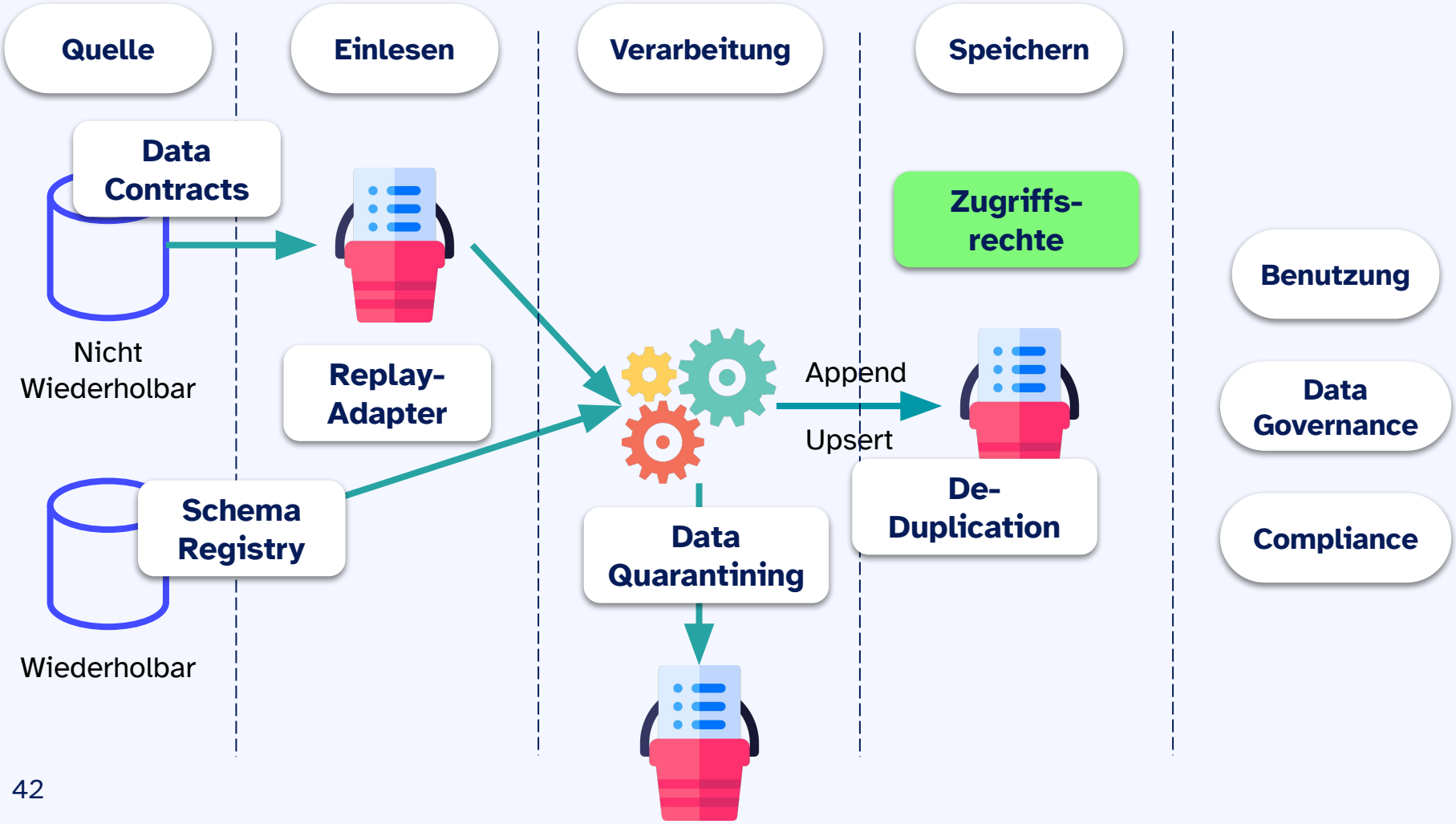
- Stateful dropDuplicates bei Spark Streaming
- Upsert als Schreibstrategie verwenden

💡 Falls kein eindeutiger Schlüssel vorhanden
-> Hash-Wert über eine Menge an identifizierenden Spalten berechnen

Datum	Land	Appname	Downloads
01.10.23	DE	DataQuality+	520
02.10.23	DE	DataQuality+	310

identifizierenden Spalten

Key	Datum	...
c325c10	01.10.23	...
38ec90b	02.10.23	...


Quelle
Einlesen
Verarbeitung
Speichern
Data Contracts
Nicht Wiederholbar
Replay-Adapter
Schema Registry
Wiederholbar
Append
Upsert
Data Quarantining
De-Duplication
Zugriffs-rechte
Benutzung
Data Governance
Compliance

Zugriffsrechte

- Zugriffsrechte auf Schemata, Tabellen und Views sollte wohl überlegt sein
- Welches Aggregationslevel wird auf Konsumentenseite benötigt?
- Logging von Zugriffen
- Zusätzlich: Spalten-basierte Zugriffsrechte

AWS Lake Formation

GCP Big Query

Apache Superset

Fazit

Ende zu Ende

Tools

Schritt für Schritt

Vielen Dank!

Sebastian Löfflath
Cloud Architect

sebastian.loefflath@inovex.de

Ludwig-Erhard-Allee 6
76131 Karlsruhe

Tim Bossenmaier
Data & Software Engineer

tim.bossenmaier@bytefabrik.ai
+49 176 8779 0585

Haid-und-Neu Straße 10-14
76131 Karlsruhe

Materialien

Southekal, Prashanth (2023): Data Quality - Empowering Businesses with Analytics and AI. Wiley, New Jersey.

Meier, Kolja & Spitzer Marcel (2023): Hohe Qualität vom Anfang bis zum Ende - Digital Future, Folge 17. https://digital-future.podigee.io/17-neue-episode

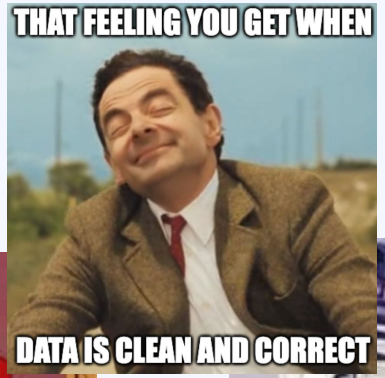
Estuary (2023): What Is Data Quality? Dimensions, Standards, & Examples. https://estuary.dev/data-quality

PayPal (2023): Template for Data Contract. https://github.com/paypal/data-contract-template

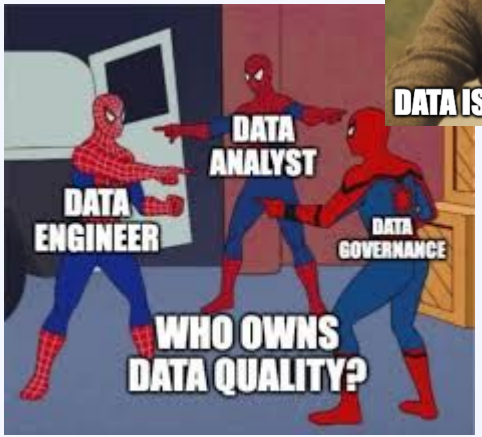
Sanderson, Chad (2022): The Rise of Data Contracts. https://dataproducts.substack.com/p/the-rise-of-data-contracts

SeattleDataGuy (2023): Is Everyone's Data Infrastructure A Mess? https://seattledataguy.substack.com/p/is-everyones-data-infrastructure

Fragen ?


THAT FEELING YOU GET WHEN
DATA IS CLEAN AND CORRECT

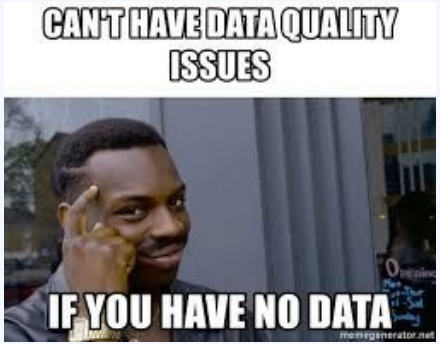

QUALITY IS EVERYONE'S RESPONSIBILITY


DATA ENGINEER
DATA ANALYST
DATA GOVERNANCE
WHO OWNS DATA QUALITY?


YOUR DATA WITH DATA QUALITY


YOUR DATA WITHOUT DATA QUALITY


LET ME INTRODUCE YOU TO
DATA QUALITY


CAN'T HAVE DATA QUALITY ISSUES
IF YOU HAVE NO DATA


DUPLICATES, YOU HAVE.
CLEAN UP, YOU MUST.


WHAT IF I TOLD YOU
YOU CAN IMPROVE QUALITY *AND* LOWER COST

BackUp: Data Contract in DataHub