



AGILE. IT. MANAGEMENT.



WEBINAR

DATA LAKE

EVENTS 2020 – HIER TREFFEN SIE UNS



Willkommen zum Webinar: Plattform für Industrielle KI. Es geht in Kürze los, bis dahin finden Sie eine Auswahl weiterer AIM-Events im April. – Mehr unter: www.agile-im.de/unternehmen/veranstaltungen/



Online

Webinar
Predictive Supply Chain

21. August 2020



Online

Webinar
Data Lake

27. August 2020



Online

Webinar
Intelligent Digital Asset
Management

28. August 2020



Online

Webinar
Plattform für Industrielle KI

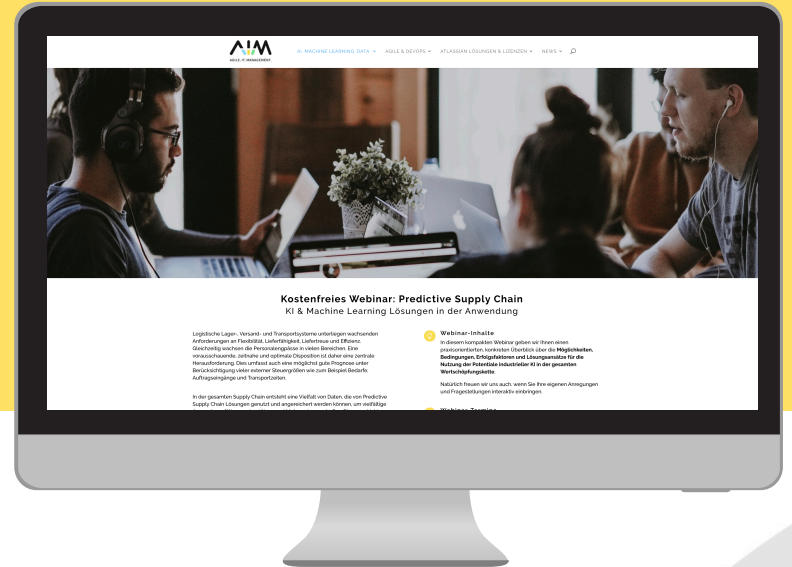
04. September 2020



ORGANISATORISCHES



- ✓ Entstehen Fragen während des Webinars? Dann schreiben Sie uns diese einfach in das F&A Fenster.
- ✓ Ihre Fragen werden nach dem Webinar persönlich per E-Mail beantwortet.
- ✓ Sie erhalten die Aufzeichnung des Webinars im Nachgang per E-Mail zugesendet.
- ✓ Die Folien des Webinars werden auf Nachfrage per E-Mail zugesendet. Anfrage an chilber@agile-im.de



STECKBRIEF & TEAM



Fast Facts

- Gründung Anfang 2016
- 25 Mitarbeiter in 2020



Kooperationen & Partner

- Enge Zusammenarbeit mit Schwesterfirma inserve
- Kooperation mit Uni Osnabrück und L3S in Hannover



VORSTELLUNG CARSTEN HILBER



Steckbrief

- Seit 19 Jahren Software Architekt und DevOps Engineer (delivery-boy)
- Erfahrung: Projekt- und Produktentwicklung / Aufbau von Entwicklungsprozessen
- Fokus Themen:
 - Automatisierung von Entwicklungsprozessen
 - Erstellung von industriell einsetzbaren Architekturen und Plattformen

INDUSTRIELLE KI - BAUSTEINE



Predictive Dynamics

Zeitreihen Vorhersage &
Modellierung dynamischer Systeme



Computer/Machine Vision

Visuelle Verarbeitung und
Informationsextraktion (Bildern, Videos)



Document/Language Processing

Klassifizierung, Strukturierung &
Extraktion sprachlicher Inhalte

AIM Industrial AI Plattform / Engines

Data Engineering – Data Lake



INDUSTRIELLE KI- ANWENDUNGEN



Predictive Dynamics



Computer/Machine Vision



Document/Language Processing

Predictive Supply Chain

Predictive Maintenance / Service

Predictive Marketing / Sales

Intelligent Quality Assurance

Intelligent Digital Asset Management

Object Detection & Tracking

Intelligent Document Management

Intelligent Content Management

Intelligent IT Service Management



**AIM Industrial AI
Plattform**

Data Lakes

Agenda

- **Motivation und Herausforderungen**
- **Abgrenzung zum Data Warehouse**
- **Extraktion von Daten aus Drittsystemen**
- **Architektur und Komponenten**
- **Datenablage**
- **Vorgehen**
 - Einstieg
 - Einbinden in Industrial AI Projekte



DATA LAKES

Motivation und Herausforderungen

EINORDNUNG



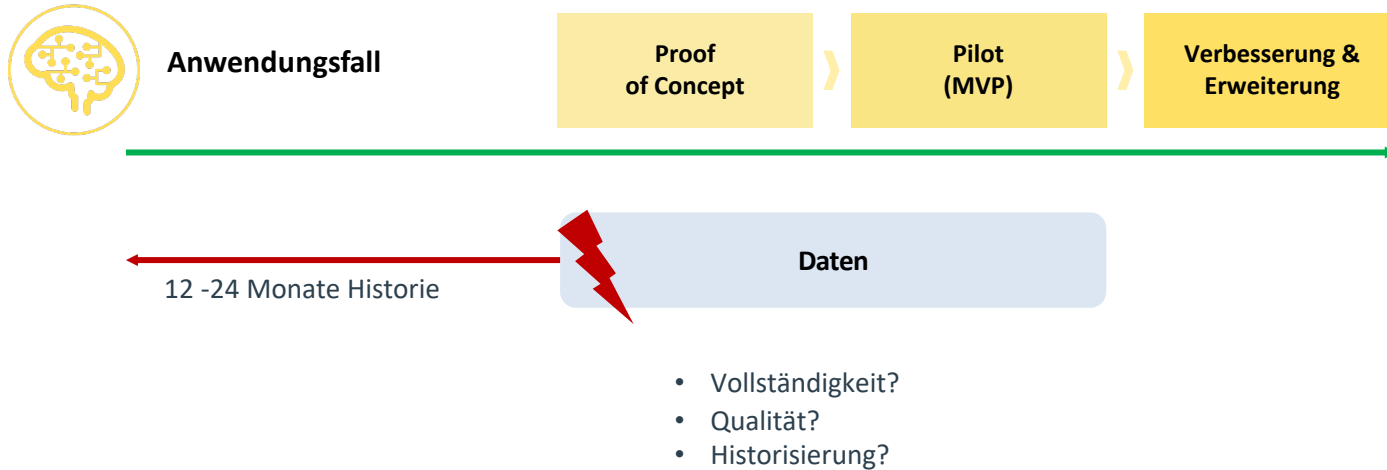
- **Zusatzdaten:** Einfache Bereitstellung und Nutzung relevanter Zusatzdaten (z.B. Wetter, Verkehr, Ferien, ...).
- **Extraktion:** Verbindung der Predictive Dynamics Lösung mit bestehenden Unternehmenssystemen.
- **Machine Learning:** Automatisierung von Training, Messung, Produktivsetzung, Monitoring leistungsfähiger Modelle.
- **Skalierung:** Verarbeitung großer Datenmengen und Generierung von Prognosen in Echtzeit.
- **Erklärung:** Anwenderfreundliche Benutzeroberflächen für die Praxis und Interpretationshilfen für die Akzeptanz.

MOTIVATION



- **Transaktionale Systeme** (z.B. ERP, CRM, Service / Helpdesks, Projektsteuerung, ...) sind in der Regel historisch gewachsen und auf die pragmatische Sicherstellung der Abwicklung des operativen Geschäfts ausgerichtet.
- **KI / Machine Learning Systeme** benötigen konsistente, häufig historische (z.B. für Predictive Analytics), weitergehende (Kontext) Daten – und können Mehrwerte aus unstrukturierten / binären Daten (Sprache, Bild, Video) schöpfen.

KI PROJEKTE - DATEN





Wodurch wird Ihr Interesse an einem DataLake getrieben?



→ Herausforderungen:

- (Präventive) Sicherung bisher nicht genutzter / nicht archivierter Daten (strukturiert, binär)
- Sicherung des historischen Verlaufes veränderlicher Dateninhalte
- Handhabung struktureller Veränderungen über die Zeit (Schema Evolution)
- Sicherstellung von Konsistenz, Korrektheit und Nachvollziehbarkeit
- Unterstützung der Navigation, Verknüpfung und Weiterverwendung
- Optimierung der Kosten für die Datenspeicherung
- Datenschutzaspekte müssen bedacht werden

AUFGABEN & FUNKTIONEN



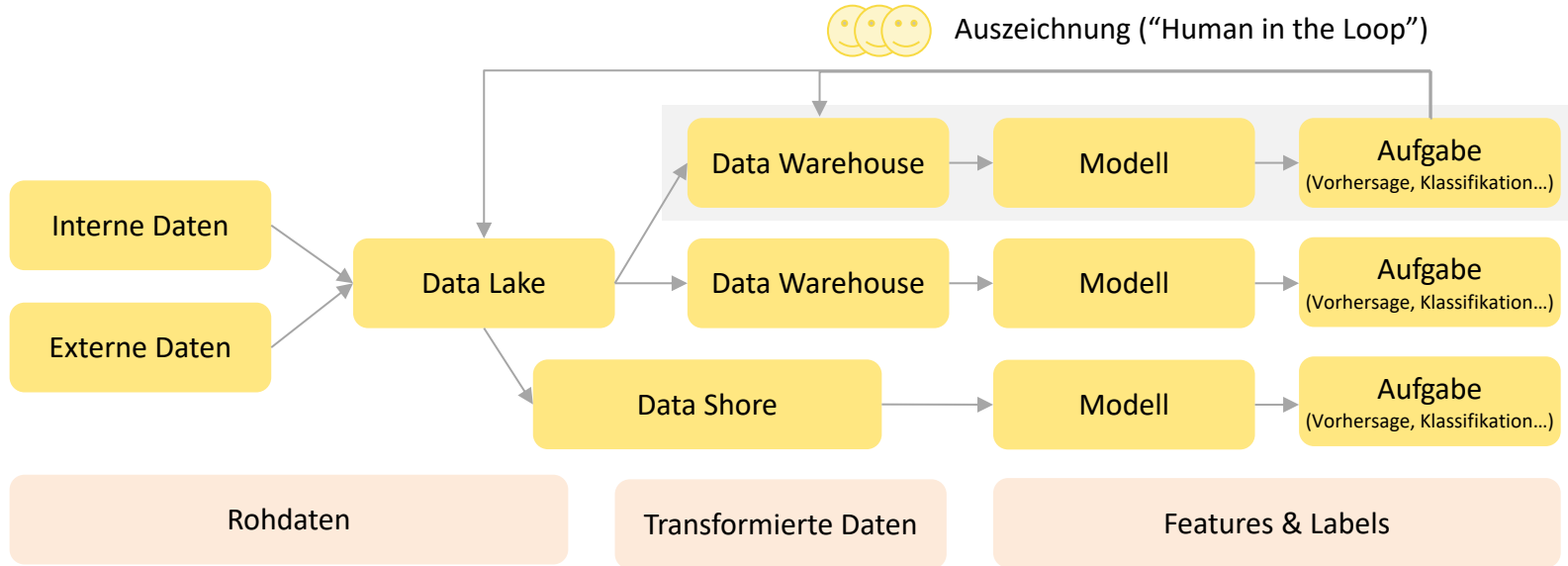
Bereich	Thema
Extraktion & Laden	Multiple Formate
	Anreicherung
Sammlung & Pflege	Historisierung - Daten
	Historisierung - Struktur
	Validierung / Bereinigung
	Verwerfen / Wiederholen
	Transformation
Bereitstellung	Durchsuchen
	Voranalyse
Übergreifend	Transaktionssicherheit
	Plattform Unabhängigkeit
	Kostenmanagement
	Staging
	Überwachung



DATA LAKES

Abgrenzung zum Data Warehouse

ABGRENZUNG / WEITERVERARBEITUNG

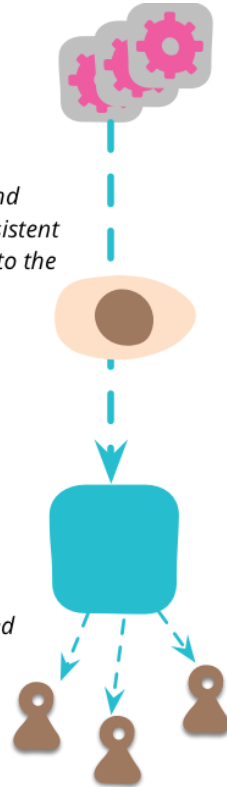


DATA LAKE VS. DATAWAREHOUSE



With a **data warehouse**, incoming data is cleaned and organized into a single consistent schema before being put into the warehouse...

... analysis is done directly on the curated warehouse data



With a **data lake**, incoming data goes into the lake in its raw form...

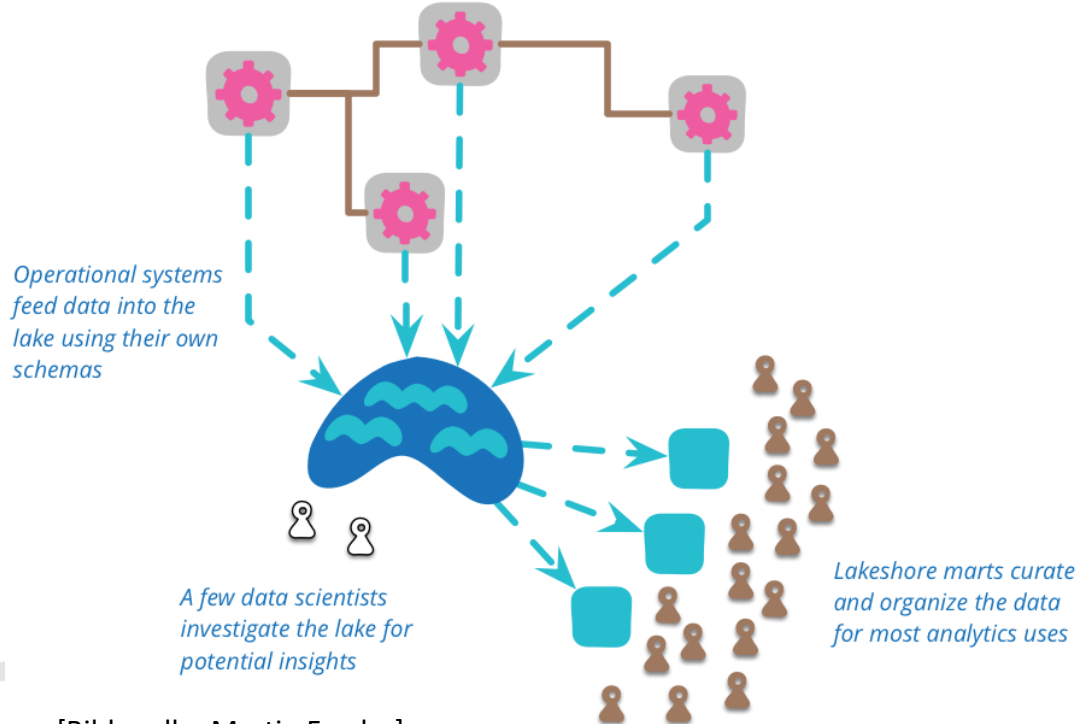
... we select and organize data for each need



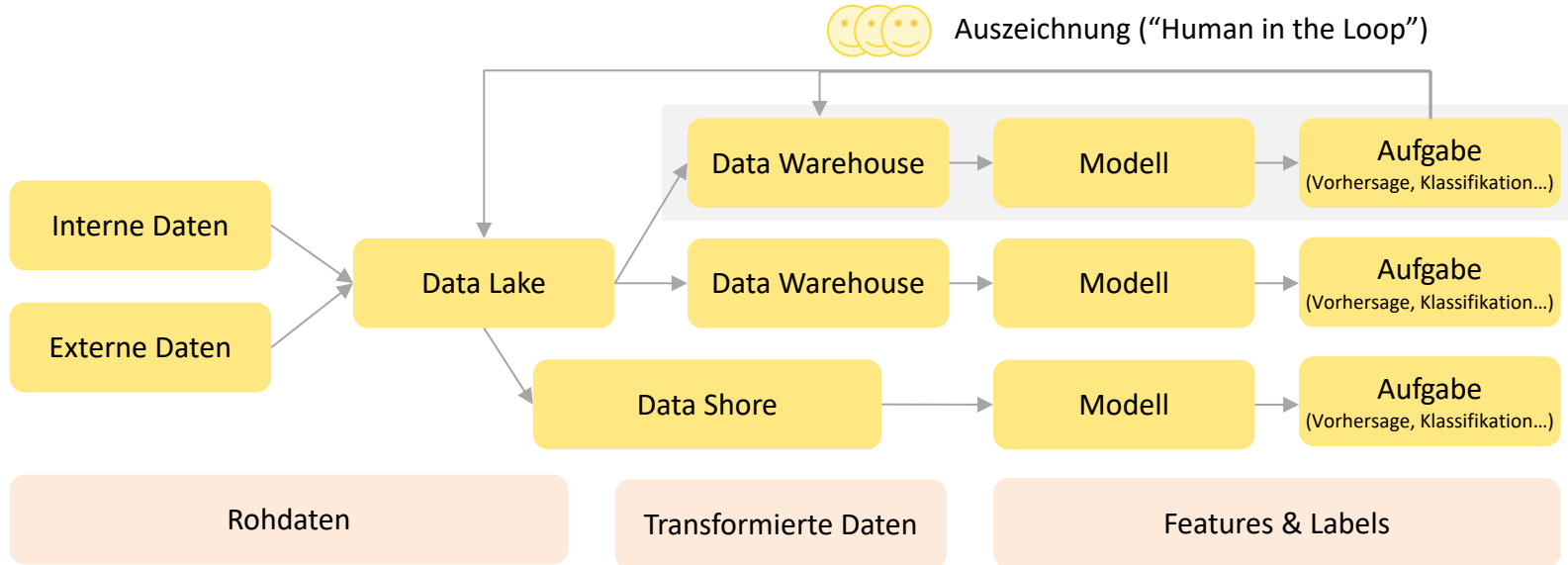
DATA SHORE – USE THE TRANSFORMED DATA



Operational systems communicate with each other via service connections, ignoring the data lake



ABGRENZUNG / WEITERVERARBEITUNG



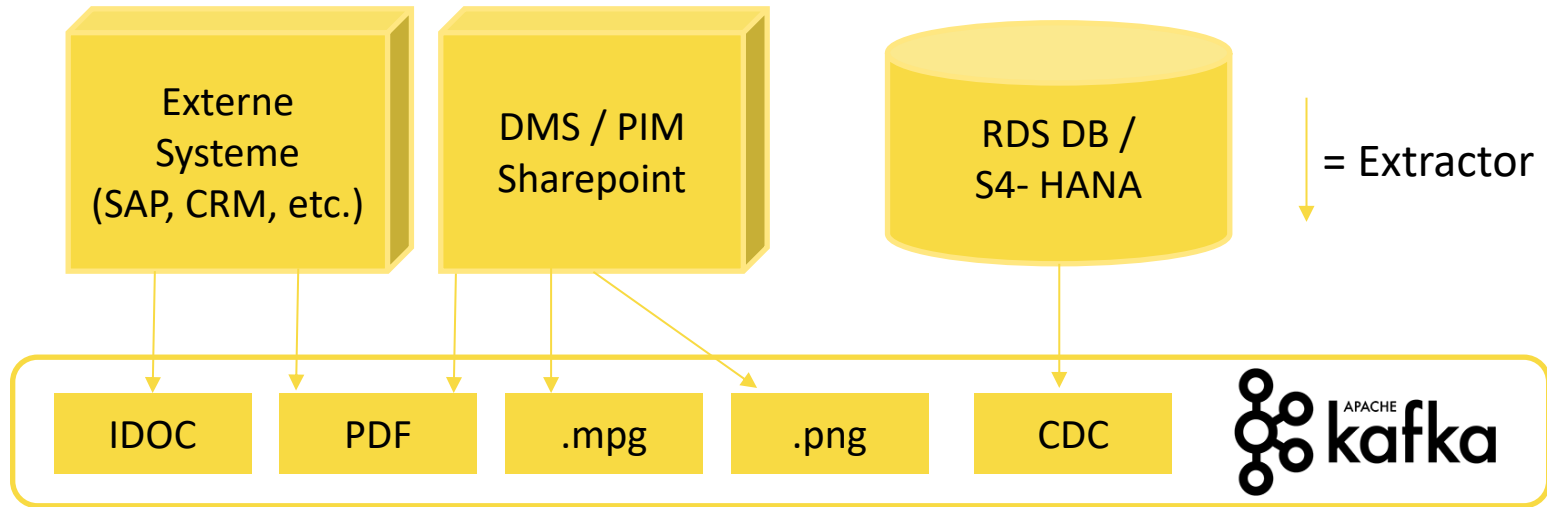


DATA LAKES

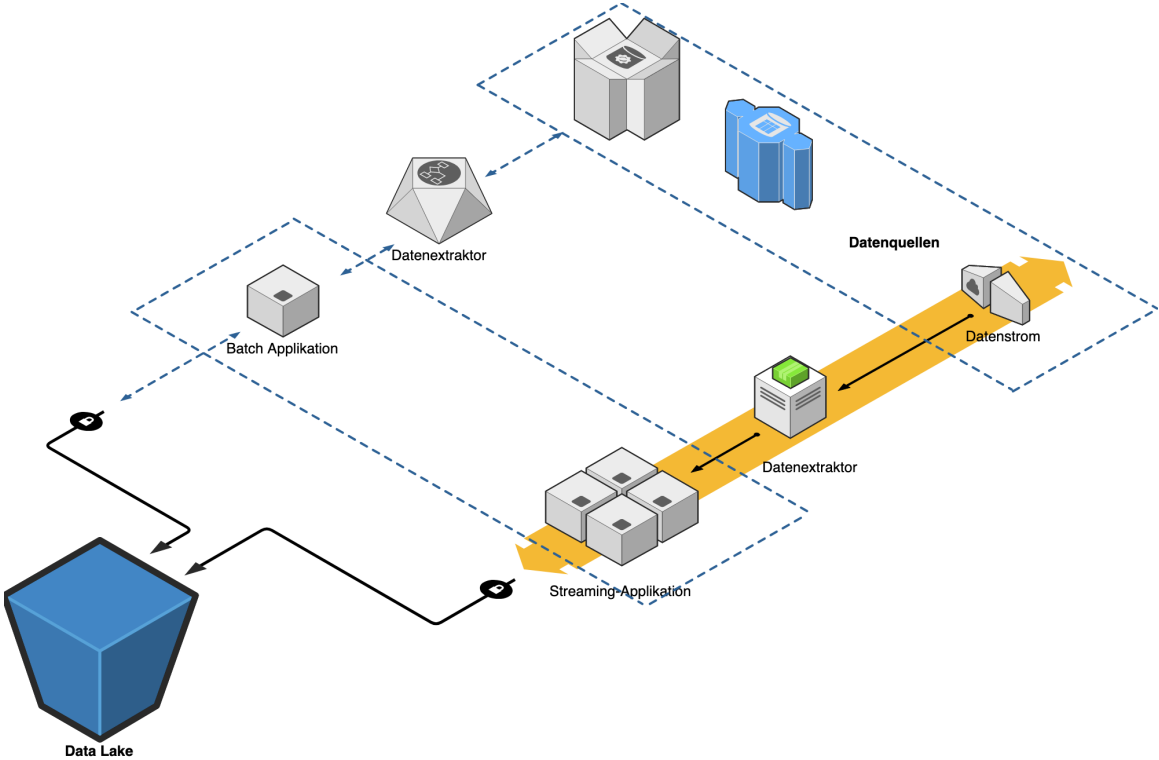
Extraktion



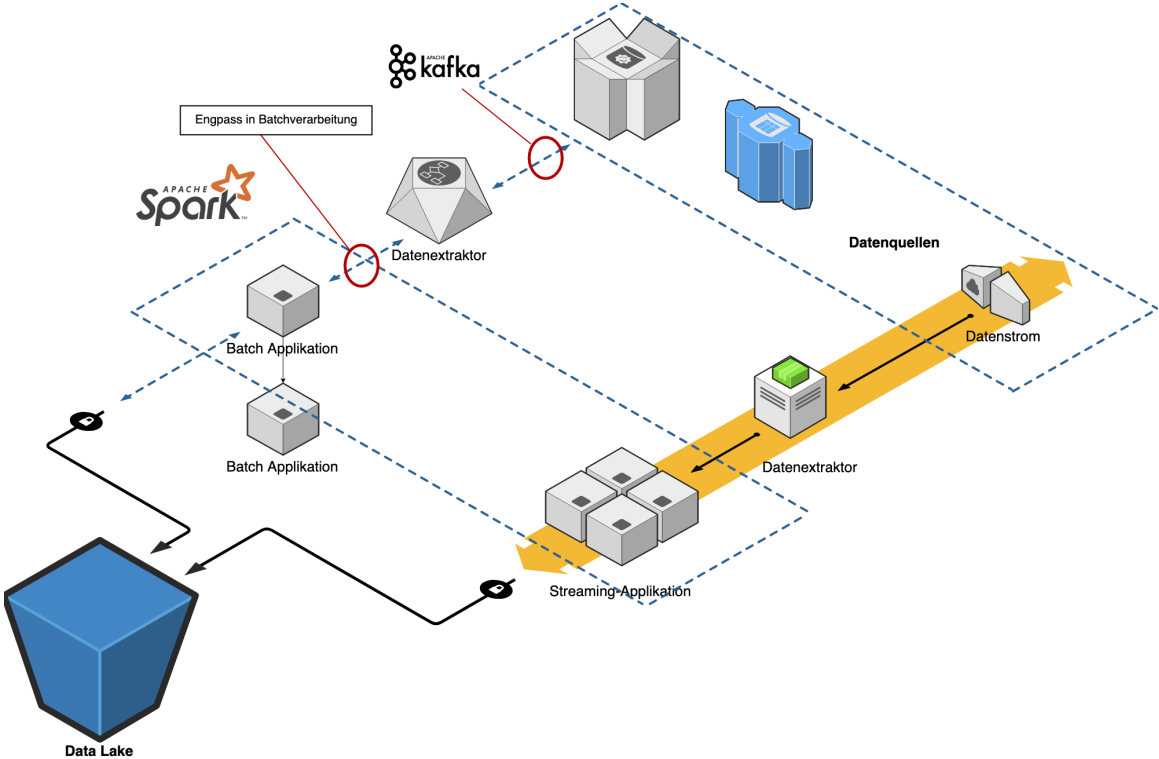
MULTIPLE QUELLEN / FORMATE



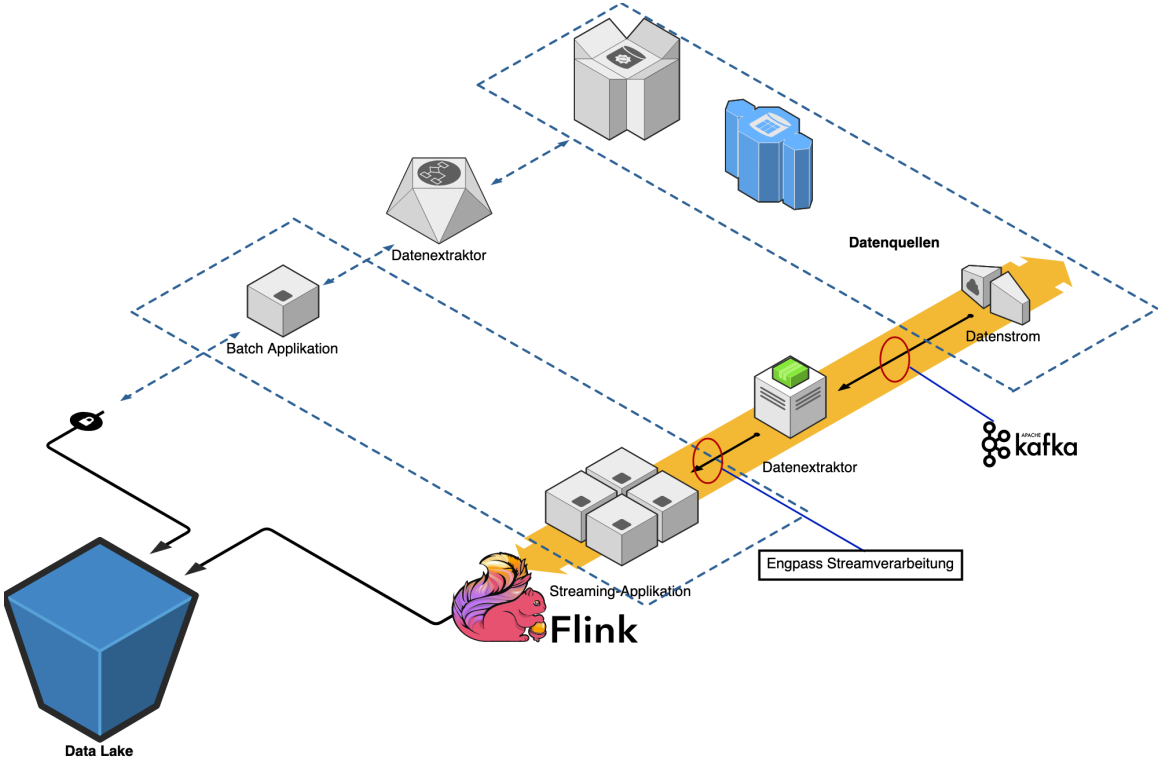
DATA LAKE – EXTRAKTION



DATA LAKE – EXTRAKTION BATCH ENGPASS



DATA LAKE – EXTRAKTION STREAM ENGPASS





DATA LAKES

Architektur und Komponenten



INDUSTRIELLE KI & BIG DATA KOMPETENZEN



Apache Flink



TensorFlow



Pandas



kubernetes



APACHE Spark



mxnet



Kubeflow



mlflow



scikit learn



docker



Grafana



python



Java



Scala

DATA LAKE KOMPONENTEN - AWS

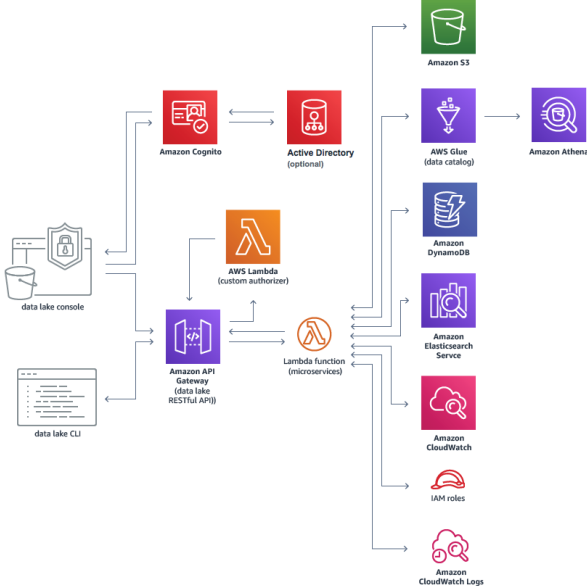


Extraktion & Laden

Sammlung & Pflege

Bereitstellung

Übergreifend



DATA LAKE KOMPONENTEN - AZURE

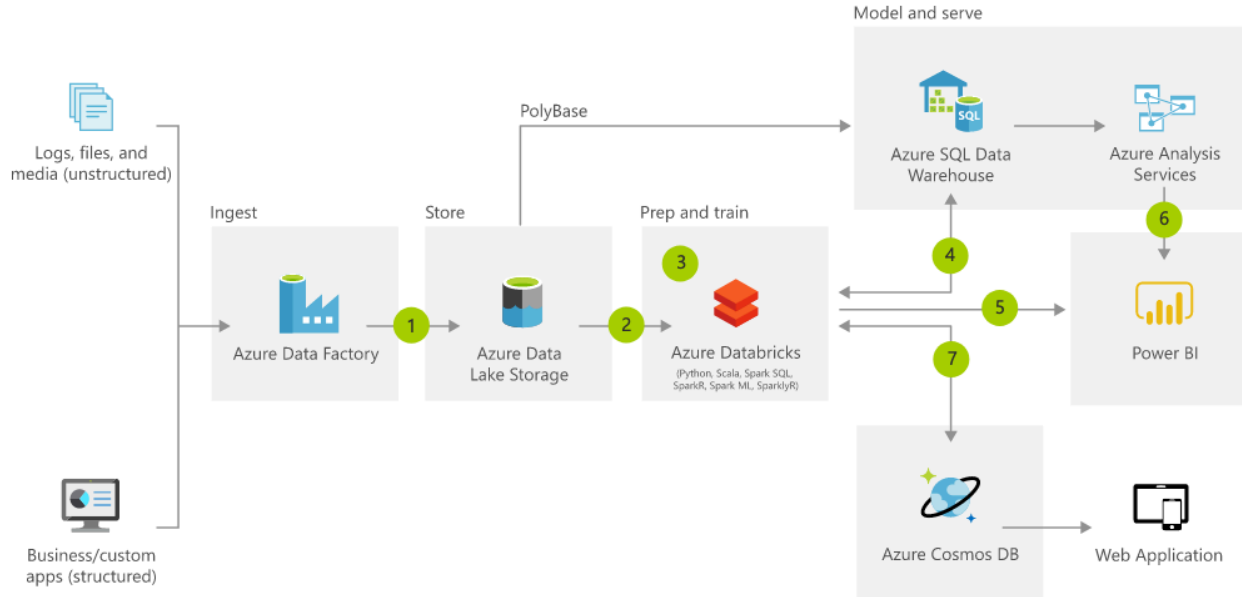


Extraktion & Laden

Sammlung & Pflege

Bereitstellung

Übergreifend



DATA LAKE KOMPONENTEN – GENERELL



Extraktion & Laden

Sammlung & Pflege

Bereitstellung

Übergreifend



UMFRAGE: CLOUD ANBIETER



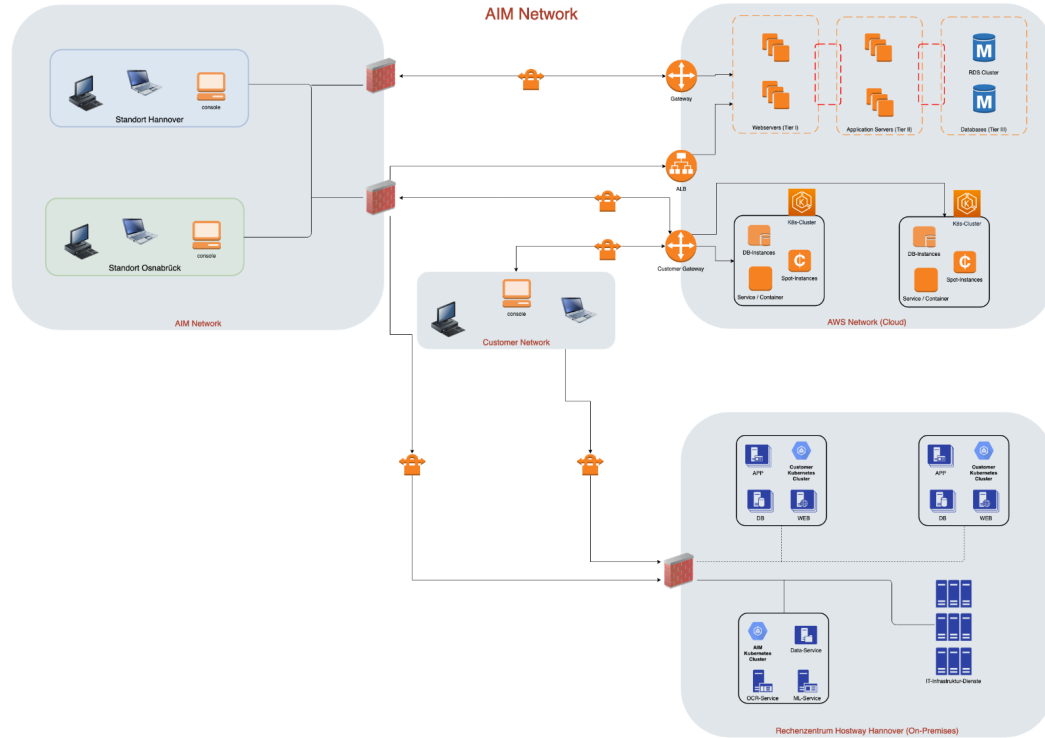
Welchen Cloud-Anbieter setzen Sie ein?



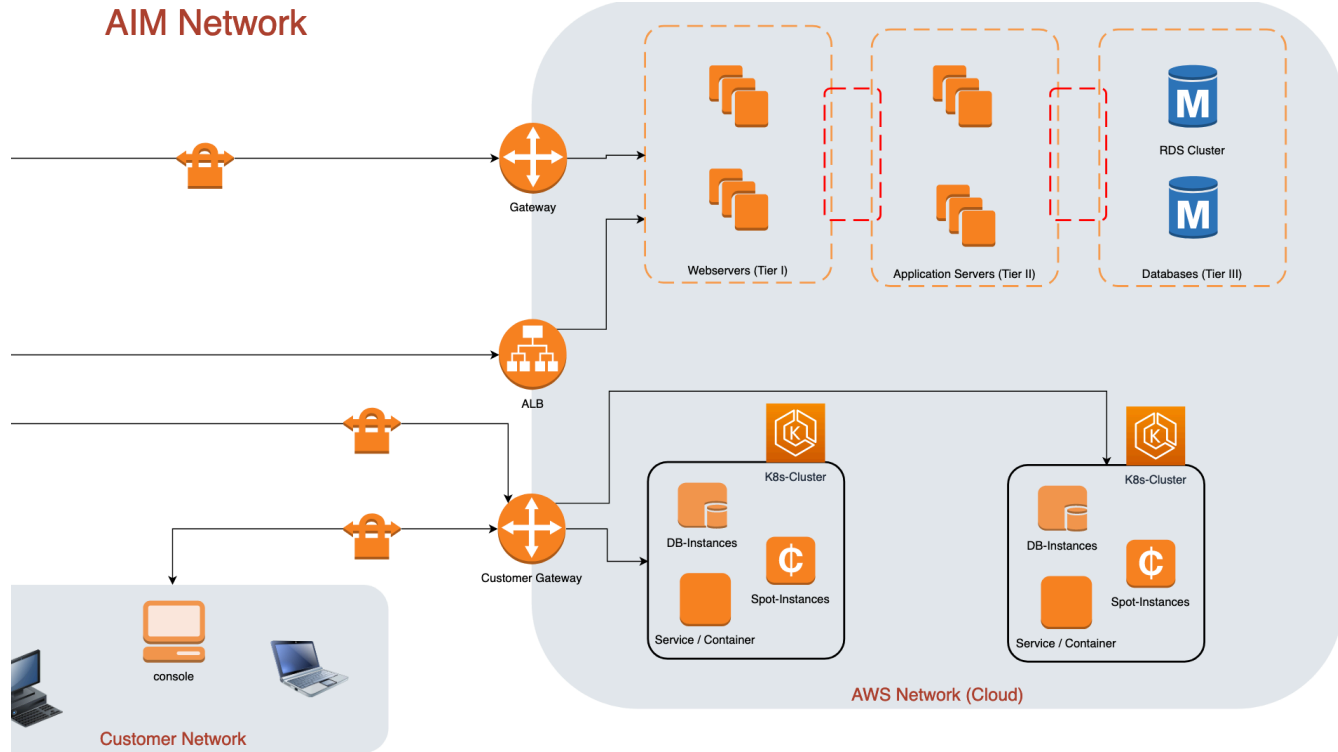
DATA LAKES

Datenablage

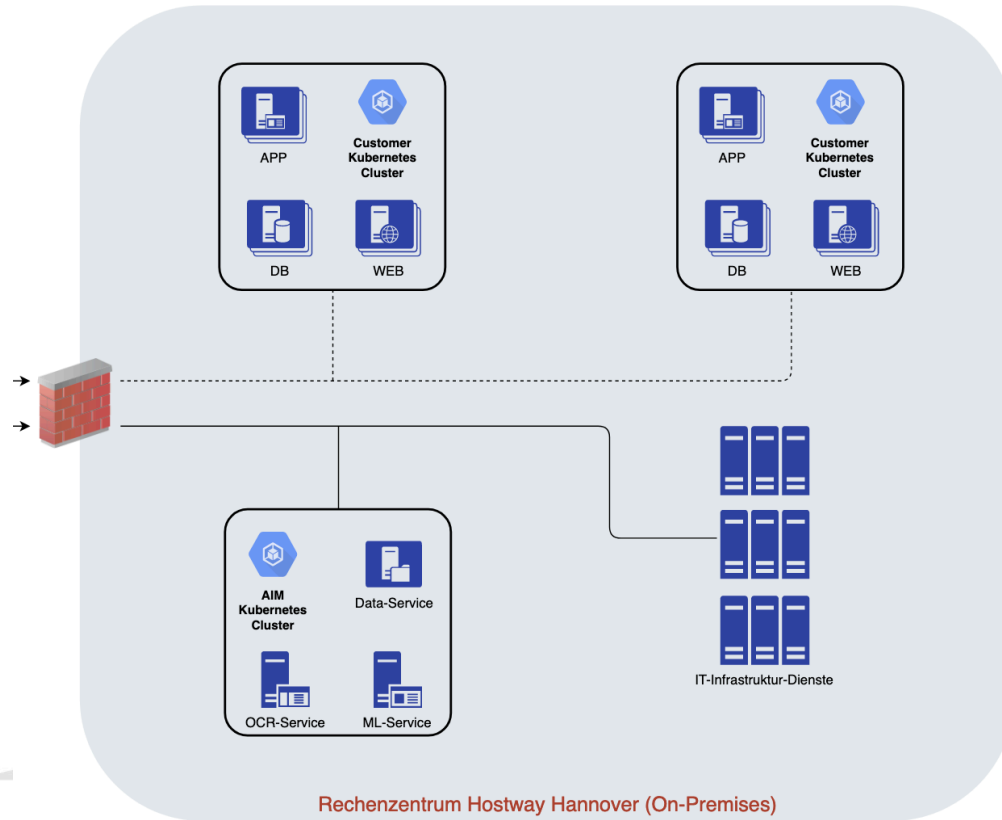
AIM NETZWERK – GESAMT



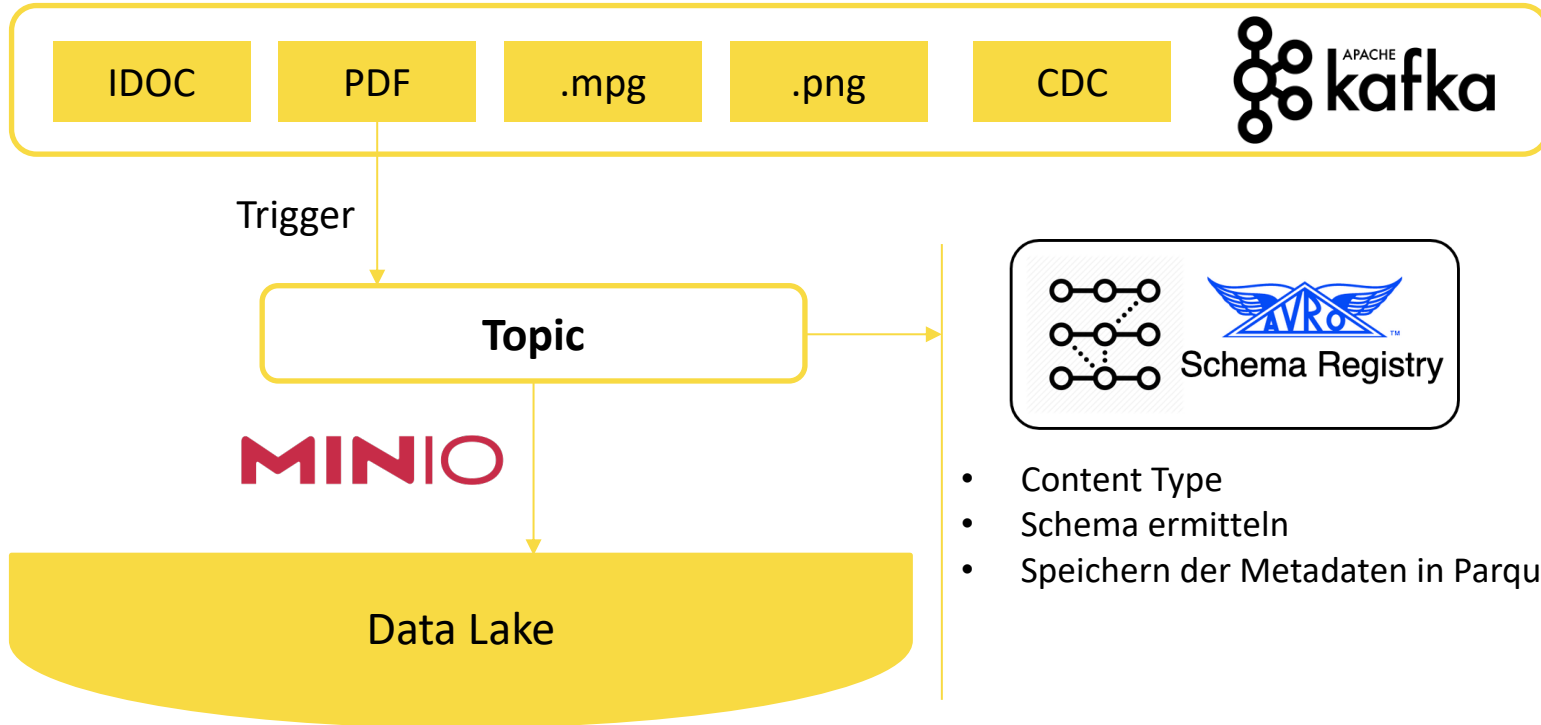
AIM NETZWERK – CLOUD (AWS)



AIM NETZWERK – RECHENZENTRUM HANNOVER

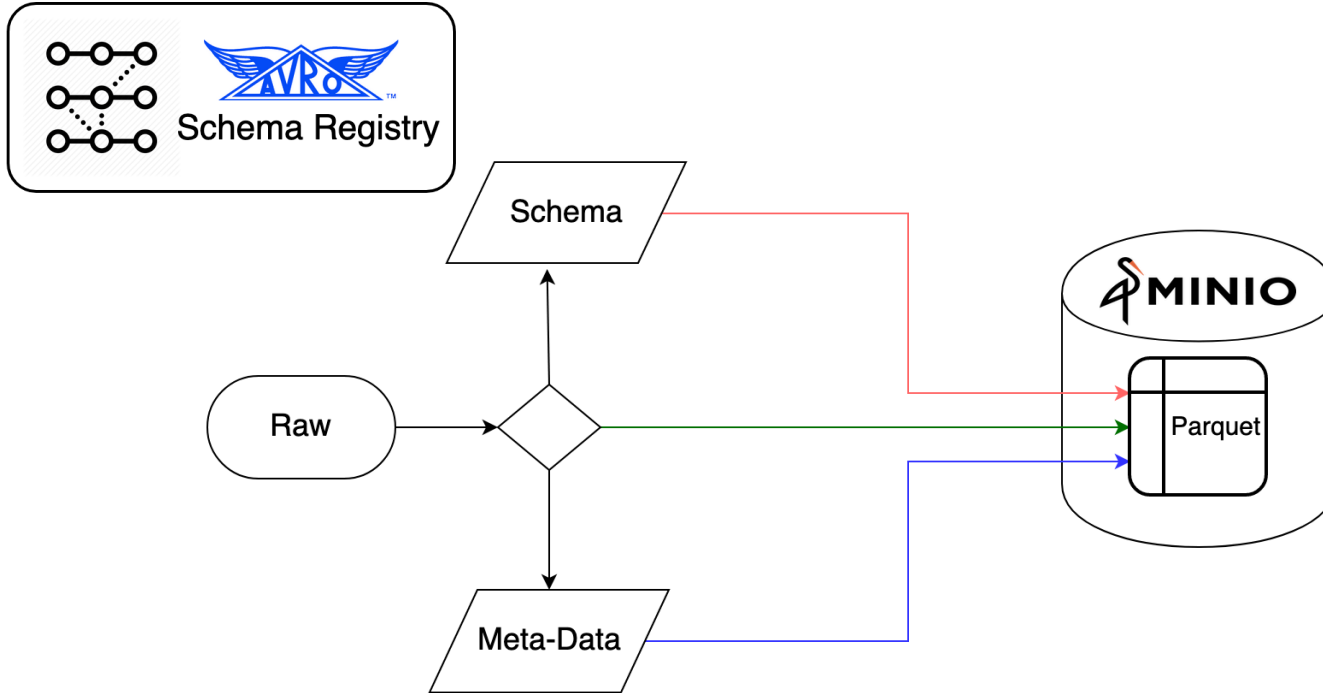


DATENABLAGE



- Content Type
- Schema ermitteln
- Speichern der Metadaten in Parquet

SCHEMA REGISTRY



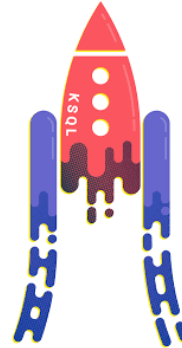
DATENBEREITSTELLUNG / -ANALYSE



- Content Type
- Metadaten
- Schema Veränderungen
- Datenbereitstellung zur Voranalyse



Grafana



MINIO

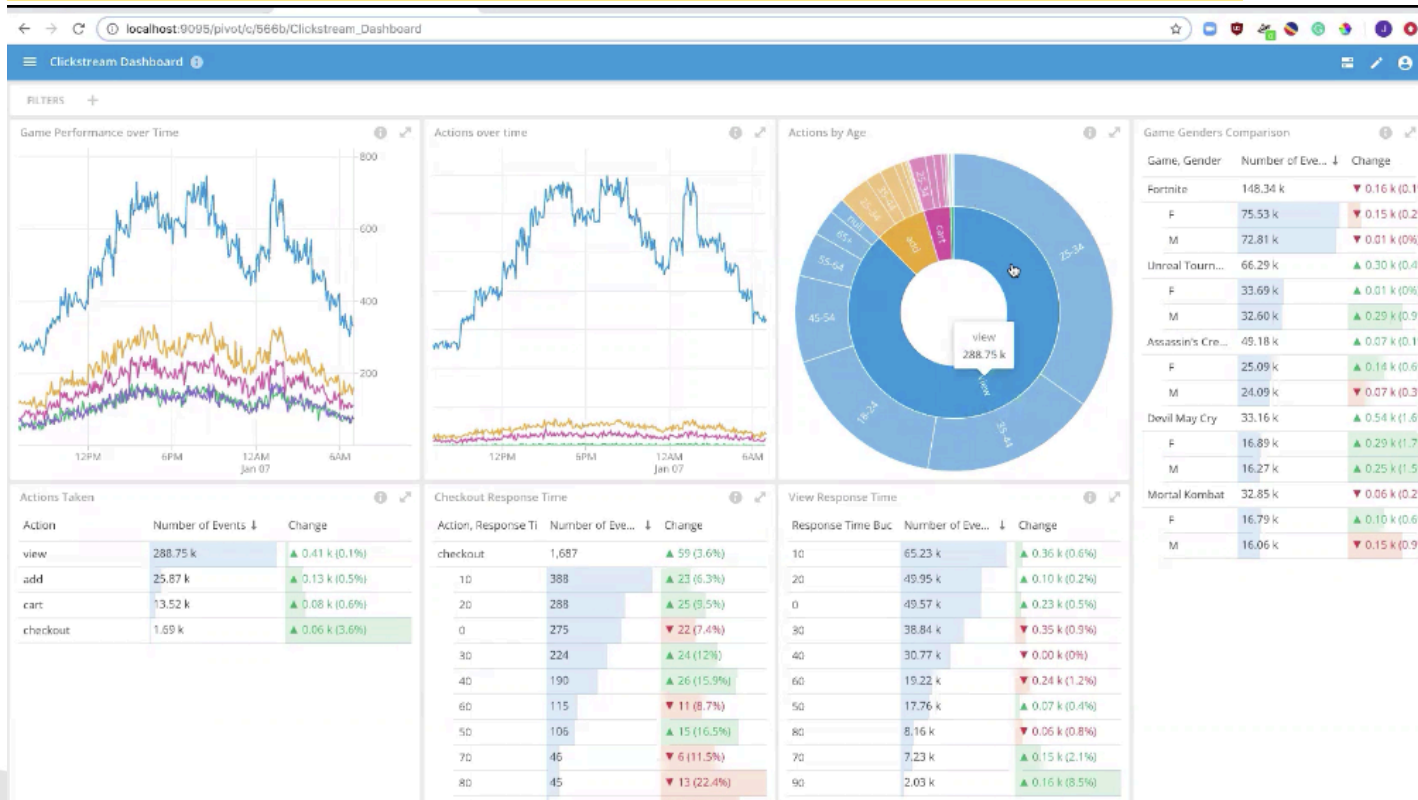
Data Lake

UMFRAGE: PROZESS- ODER ORGANISATIONSBEREICH

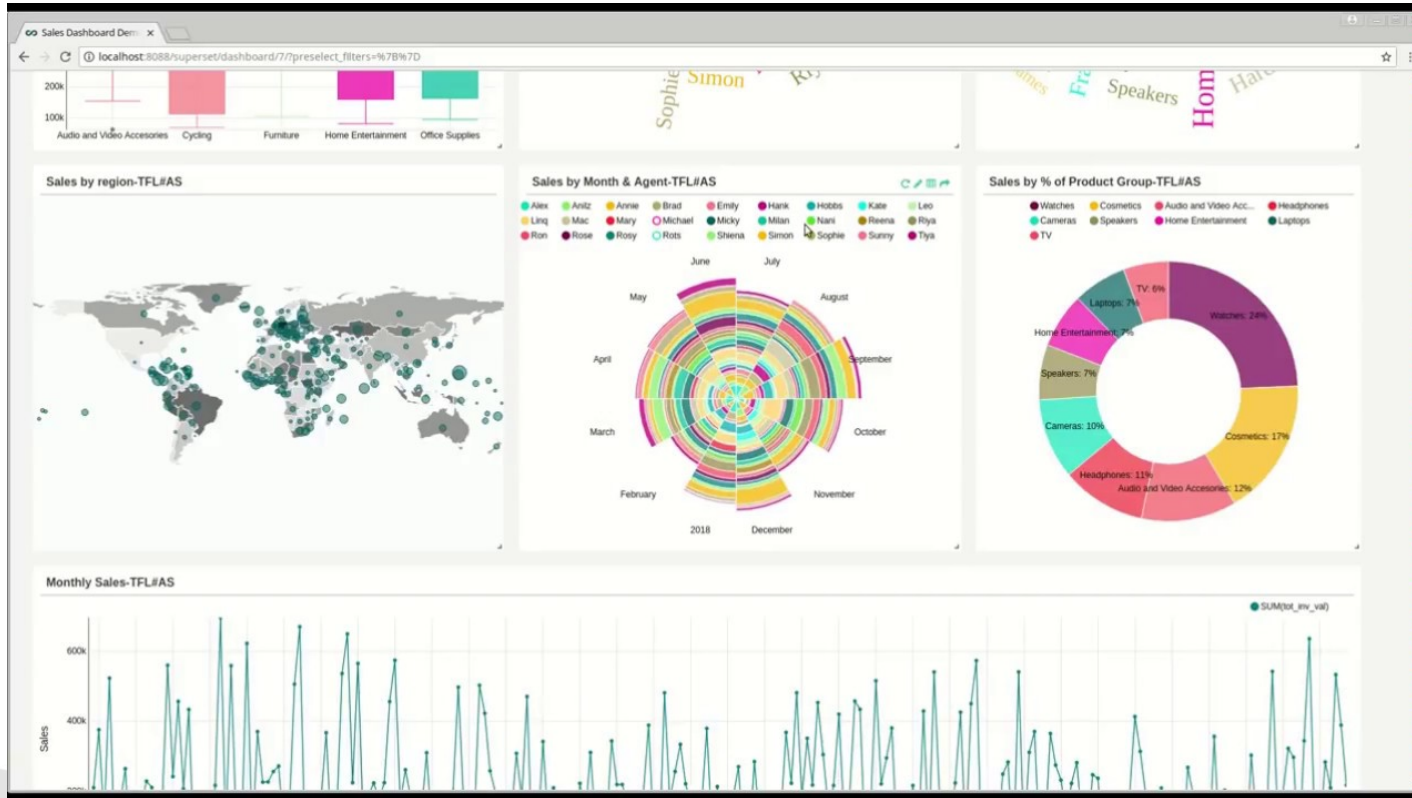


Für welchen Prozess- oder Organisationsbereich interessieren Sie sich bzw. sind Sie verantwortlich?

ANALYSE MIT DRUID



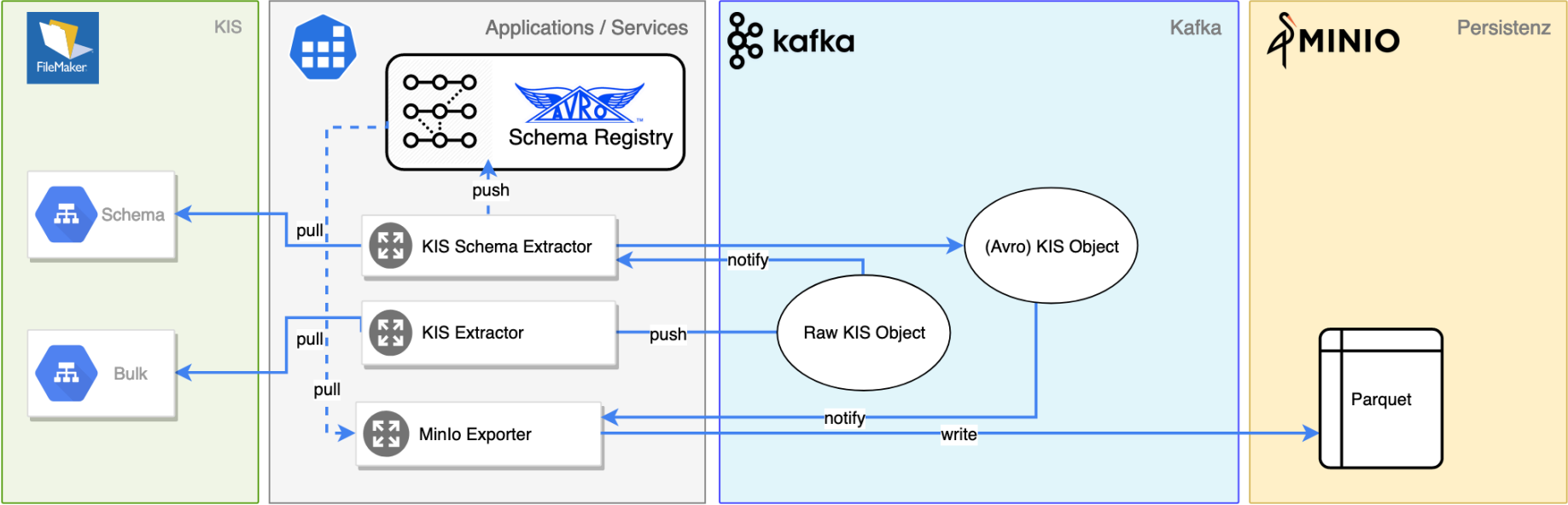
ANALYSE MIT APACHE SUPERSET



FALLBEISPIEL KUNDENINFORMATIONSSYSTEM



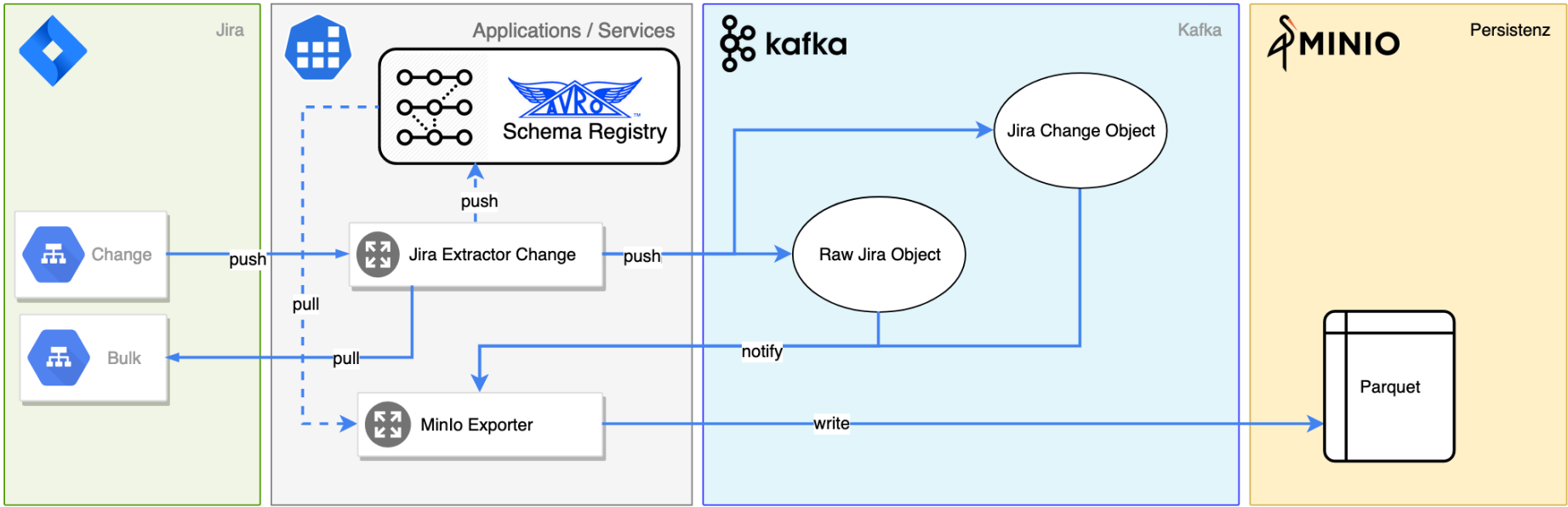
Data Lake Architecture: KIS Extractor Application



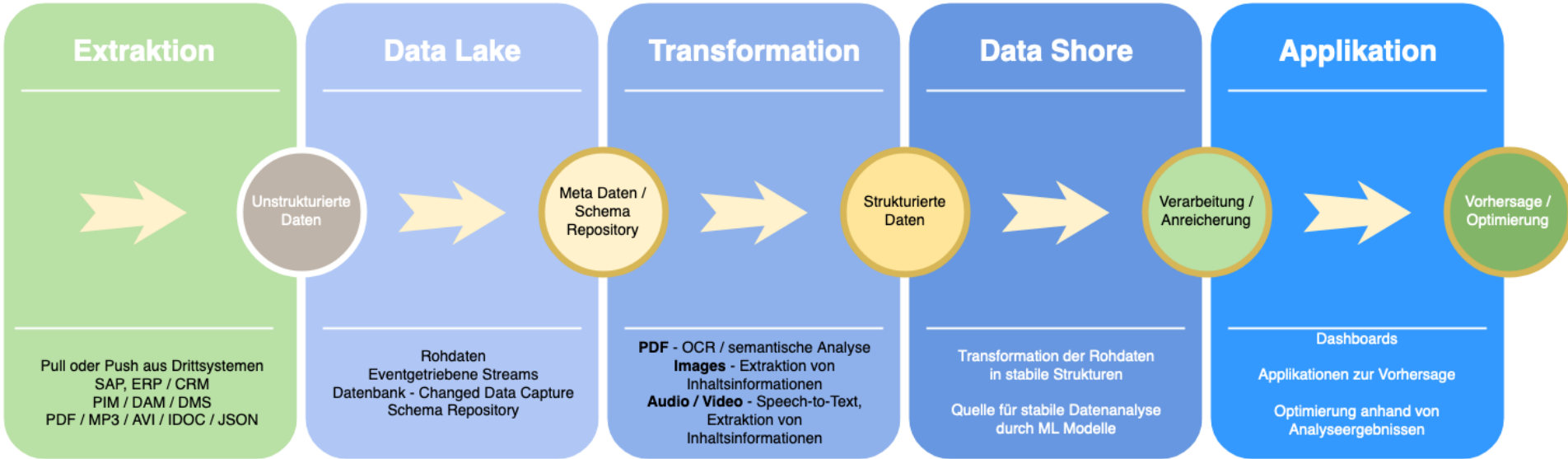
FALLBEISPIEL KUNDENINFORMATIONSSYSTEM



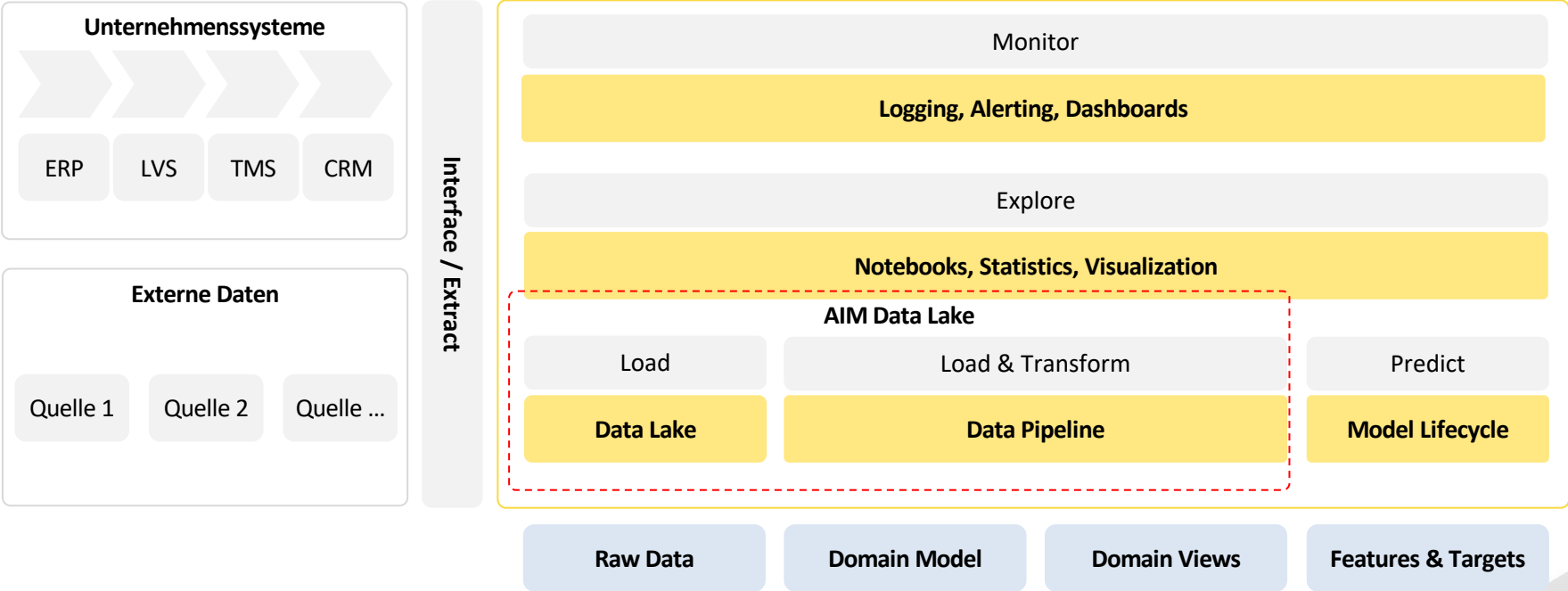
Data Lake Architecture: Jira Extractor Application



DER WEG ZUM KUNDENNUTZEN



KI PLATTFORM: PROZESSE & KOMPONENTEN



Vollständige KI Plattform: Automation, Qualität, Skalierbarkeit, Robustheit.



VORGEHEN

Proof of Concept, Projekte, Roadmap

DATA LAKE - EINSTIEG



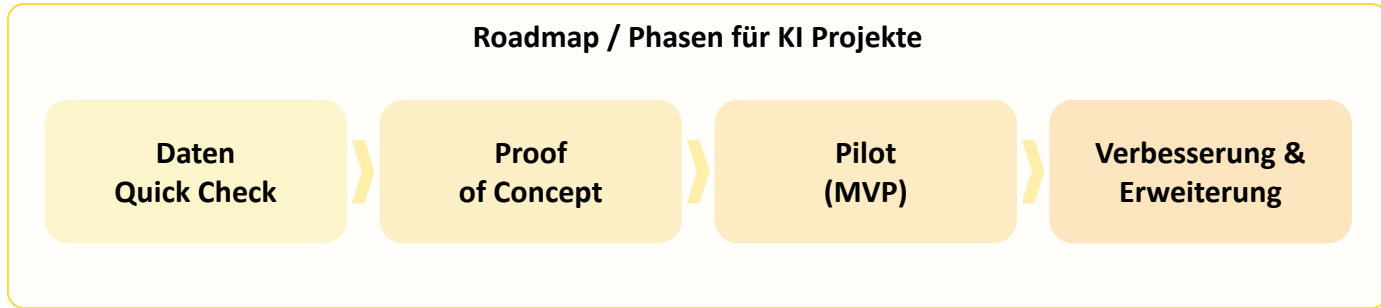
Vorbereitung:

- Identifikation (potentiell) relevanter Datenquellen
- Identifikation existierender / potentieller (AI) Anwendungsfälle
- Definition Gegenstand, Abgrenzung und Roadmap

Durchführung Pilot:

- Festlegung und Aufbau Pilot System
- Herstellung Schnittstellen und Integration
- „Trockenlauf“ (Offline“) und Start Pilotbetrieb

VORGEHEN: IAI PROJEKTE



VORGEHEN: IAI PROJEKTE



Entwicklung, Infrastruktur, Betrieb („DevOps“)

Data Science & Machine Learning

Problem
verstehen



Daten
akquirieren



Daten
explorieren



Modell
designen



Modell
optimieren



Modell
integrieren



Modell
einsetzen

Gesamt Anwendung realisieren

KI-PROJEKTE - ROLLEN



Rolle	Arbeitsweisen / Aufgaben	AIM Projektteam
Data Scientist	<u>Aufgaben</u> : Erkennung von Muster in Daten, Auswahl und Optimierung praxistaugliche KI Modelle, möglichst frühe produktivtaugliche Bereitstellung	✓
Data Engineer	<u>Aufgaben</u> : Verarbeitung & Bereitstellung großer, komplexer Datenbestände bzw. –ströme	✓
Software Engineer	<u>Aufgaben</u> : Realisierung & Betrieb robuster, skalierbarer, ergonomischer Anwendungen	✓
DevOps Engineer	<u>Aufgaben</u> : Automatisierung, Betrieb und Rollout	✓
UX Engineer	<u>Aufgaben</u> : Automatisierung, Betrieb und Rollout	✓

VORGEHEN: AI REISE



AI Strategie & Roadmap

Stoßrichtungen, Programm, Portfolio, Roadmap, Plattform in Gang bringen



AI Pilotprojekt(e)

Daten Integration & Sicherheit, Modelle & Methoden etablieren, erste Erfolge erzielen



AI Prototypen

Daten Exploration / Analyse, Modell Erprobung, Machbarkeit, Product / Project Visions



AI Erkundung

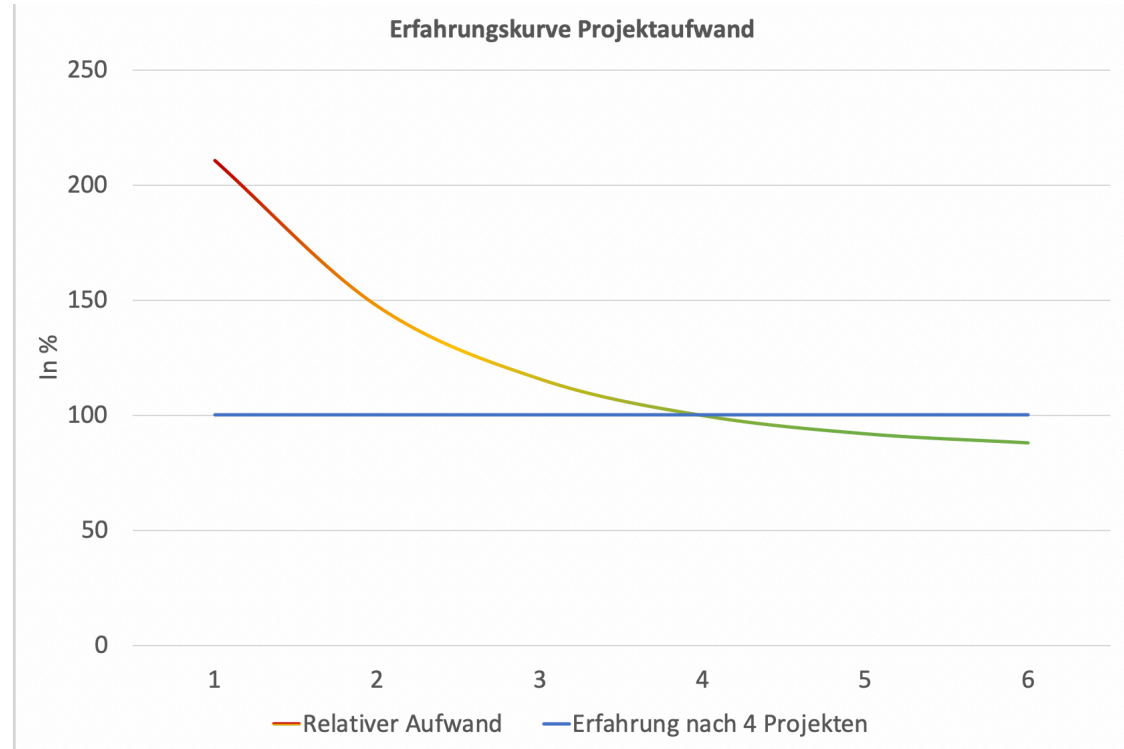
Inspiration, Verständnis, Ideen, Nutzen, Anwendungsfälle, Daten

AUFWANDSPARNIS DURCH ERFAHRUNG



Schneller zum Nutzen durch ein erfahrenes KI Projektteam.

Mit AIM als Partner mehr als 50% Aufwand sparen.



UMFRAGE: NÄCHSTE SCHRITTE



Wie können wir Sie am besten unterstützen?

PROJEKT BEISPIELE



- Verarbeitung > 10 Mio. Datensätze / Tag aus DWH > mehrere TByte
- Transformation aus Domänenmodell in Features für Predictive Analytics
- Nutzung von Flink und Kubernetes auf Kafka



- Management der DevOps und Data Management Infrastruktur
- Intelligente Verarbeitung / Analyse mehrerer Millionen Dokumente
- Nutzung von Kubernetes, Spark und Kafka auf AWS



- Kontinuierliche Verarbeitung von SAP Daten für Predictive Analytics
- Individuelle Architektur und Managed Services für On Premise System

KUNDEN



- Predictive Churn – Vorhersage der Abwanderungswahrscheinlichkeit
- Aufbau eines Data Lake für strukturierte Daten (KIS, Jira)
- Historisierung, Bereinigung und Verknüpfung der Daten



- Batch / Stream Verarbeitung mehrerer Millionen Assets (Multi Mandanten)
- Dynamisches Management der Daten (On Demand / Retention / Löschung)
- Einsatz von Kubernetes, Kafka und Managed Services in der Private Cloud



- Piloten für Batch / Stream Verarbeitung ~8 Mio. Dokumente / Jahr
- Intelligente Verarbeitung / Analyse vor Kernsystem



Carsten Hilber

“ Wir freuen uns mit Ihnen über einen Data Lake oder eine Industrial AI Anwendung zu sprechen.“

E-Mail: chilber@agile-im.de

Mobil: 0176/20901740

- chilber@agile-im.de
- <https://www.linkedin.com/in/carsten-hilber>
- https://www.xing.com/profile/Carsten_Hilber
- <https://twitter.com/chilb3r>