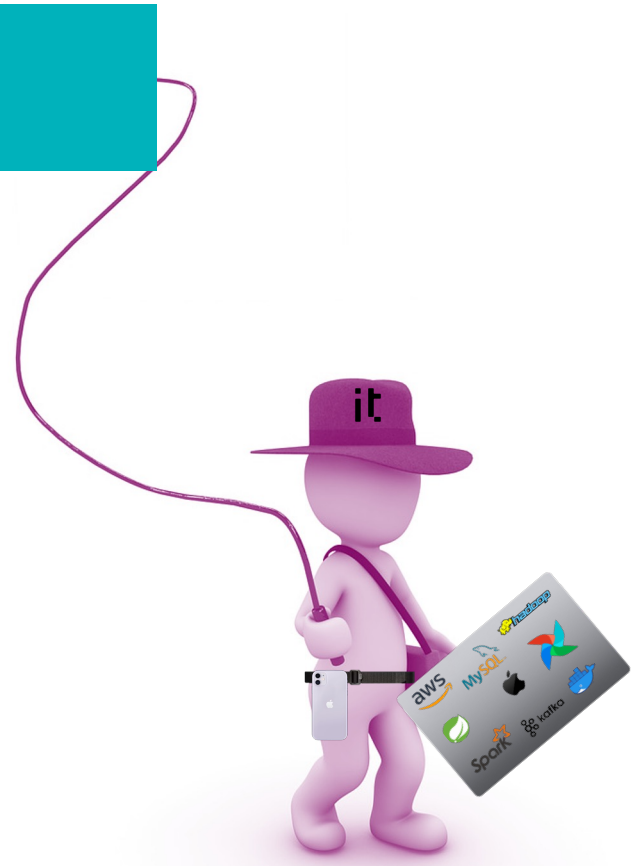
Datenzähmen leicht gemacht

Warum verlässliche Data-Architekturen auf solidem Software-Engineering basieren

Dr. Masud Afschar

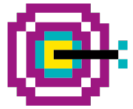


iteratec

iteratec

Agenda

Kurzvorstellung iteratec



Projekt und Technische Herausforderungen

Wichtige Erkenntnisse


iteratec

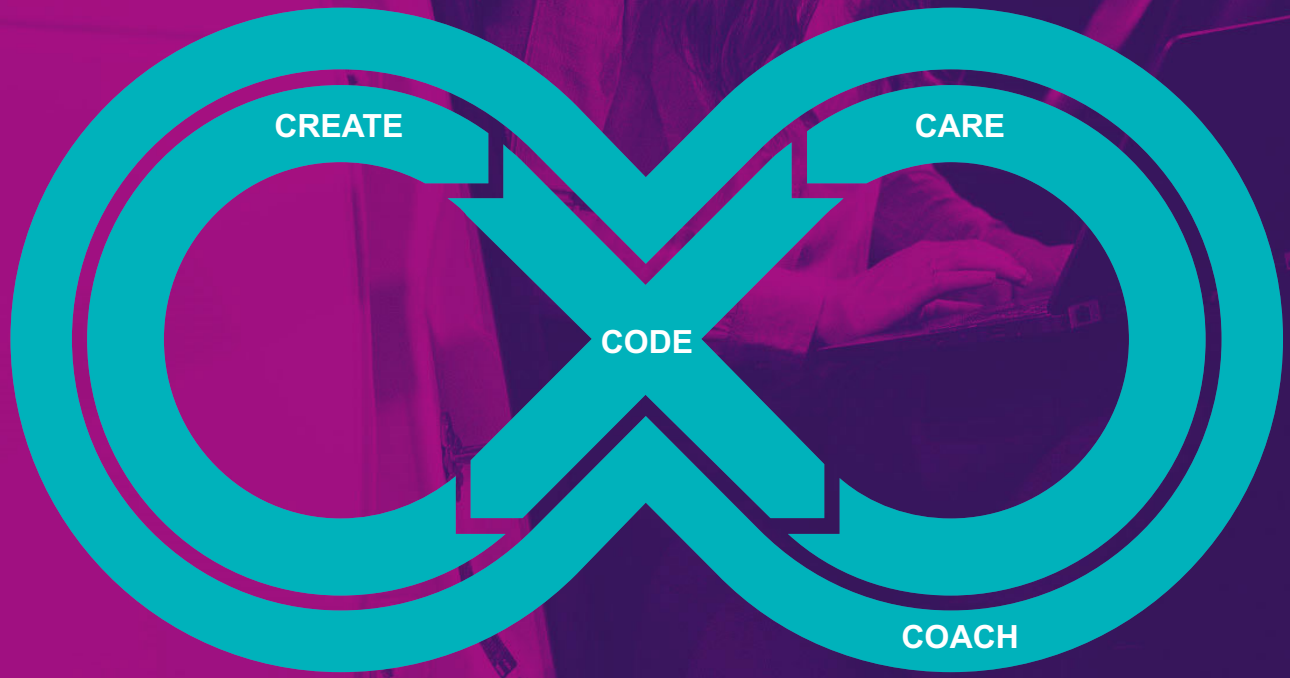
iteratec

Wie wir digitale Champions entwickeln

Innovationspotenziale erkunden, erproben und umsetzen

Digitale Lösungen & Leistungen entwickeln

Funktion, Performance und Sicherheit gewährleisten



Technologie und Wissen vermitteln, Zusammenarbeit und Strukturen verbessern

iteratec

Gut aufgestellt

470
Kolleg*Innen

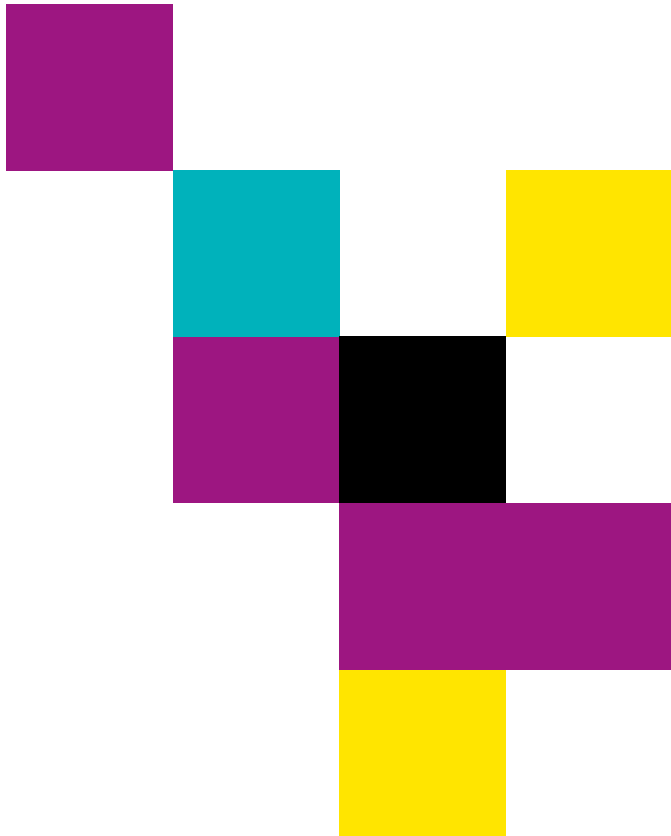
100
Studierende

27
Jahre Erfahrung

64
Mio. Euro Umsatz

7
Standorte

iteratec



Einstieg

Data Engineering und das Projekt

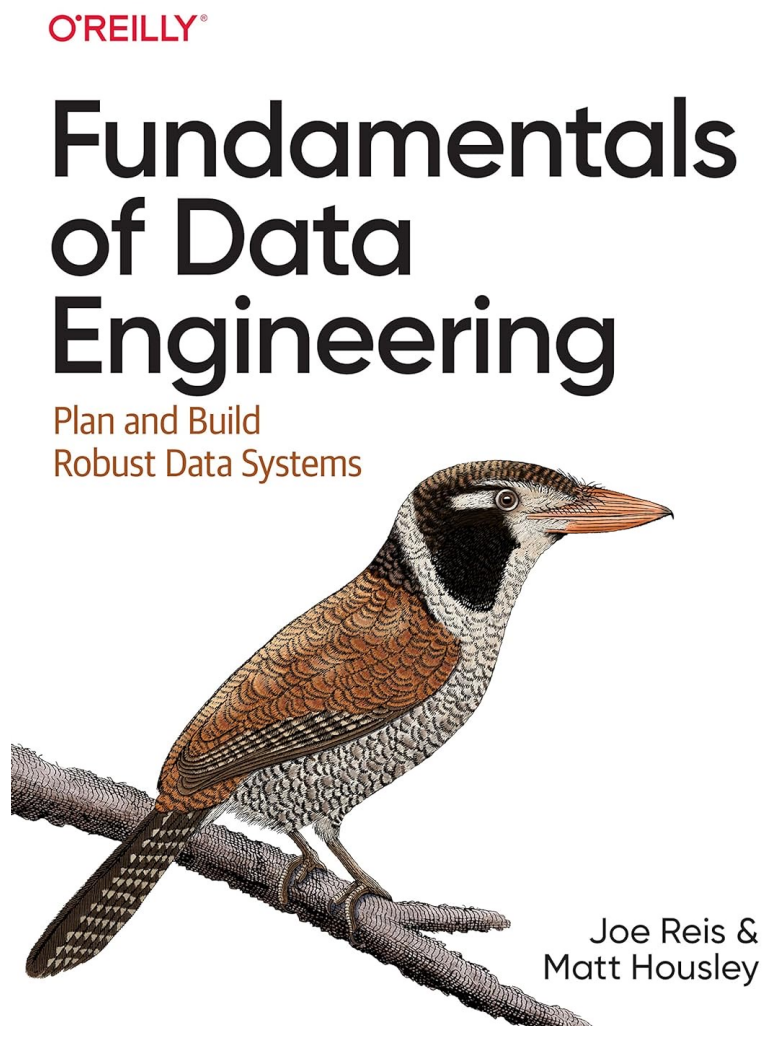
Was ist Data Engineering?

iteratec

Und was macht ein Data Engineer?

- „Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream uses, such as analysis and machine learning.“

- „A data engineer manages the data engineering lifecycle, beginning with getting data from source systems and ending with serving data for uses cases, such as analysis and machine learning“



Datenarchitektur

Layer und Tech-Stack

iteratec

Data flow

Data Sources

Data Ingestion

kafka Streams

kafka

Data Transformation

Raw | **Refined** | **Business**

amazon web services S3

Apache Spark

Apache hudi

Data Consumption

Power BI

MySQL

jupyter

Supporting Systems

Grafana

Apache Airflow

kibana

influxdb

Online-Vertriebsplattform

Herausforderungen und Lösungsansätze

iteratec

Pionierarbeit

- Erste Cloud-basierte Big Data-Plattftorm innerhalb des Unternehmens
- Seit mehr als 5 Jahren in Produktion

Datenquellen

Herausforderung: Extrahierung aus den operationalen Systemen
Lösung: Verringerung der Einstiegsbarrieren durch selbstentwickelte Librarys
Ergebnis : Alle relevanten Business-Domänen sind angebunden

Real-time Daten

Herausforderung : Anforderungen shiften in Richtung (near-)real-time
Lösung : Event-Driven Architektur von Anfang an
Ergebnis : Zukünftige near-real-time Anwendungsfälle können realisiert werden

Data-Ingestion

Herausforderung : Sicherstellung hoher Qualitätsstandards
Lösung : Einbindung von Domain Experts
Ergebnis: Hohe Datenqualität in der gesamten Transformationskette

Datenprozessierung

Herausforderung : Beherrschung komplexer Transformationsgraphen
Lösung : Einführung aufbauender Datenreifegrade und Smart-Scheduling
Ergebnis : bis zu 2.500 Transformationen pro Tag laufen unter kontrollierten Bedingungen

Data-Services

Herausforderung : Auslieferung der korrekten Daten an verschiedene Consumer-Rollen
Lösung : Individuelle Datenprodukte für Anwendungsfälle
Ergebnis : Diverse, aktive Datennutzerbasis im Unternehmen

Herausforderung: Datenschutz

Datenschutz “at scale”

Harmonisierung von Consumer-Anforderungen und Data-Governance

iteratec

Data-Konnektoren

- Standardisierte API
- Einfache Integration

Data-Realms

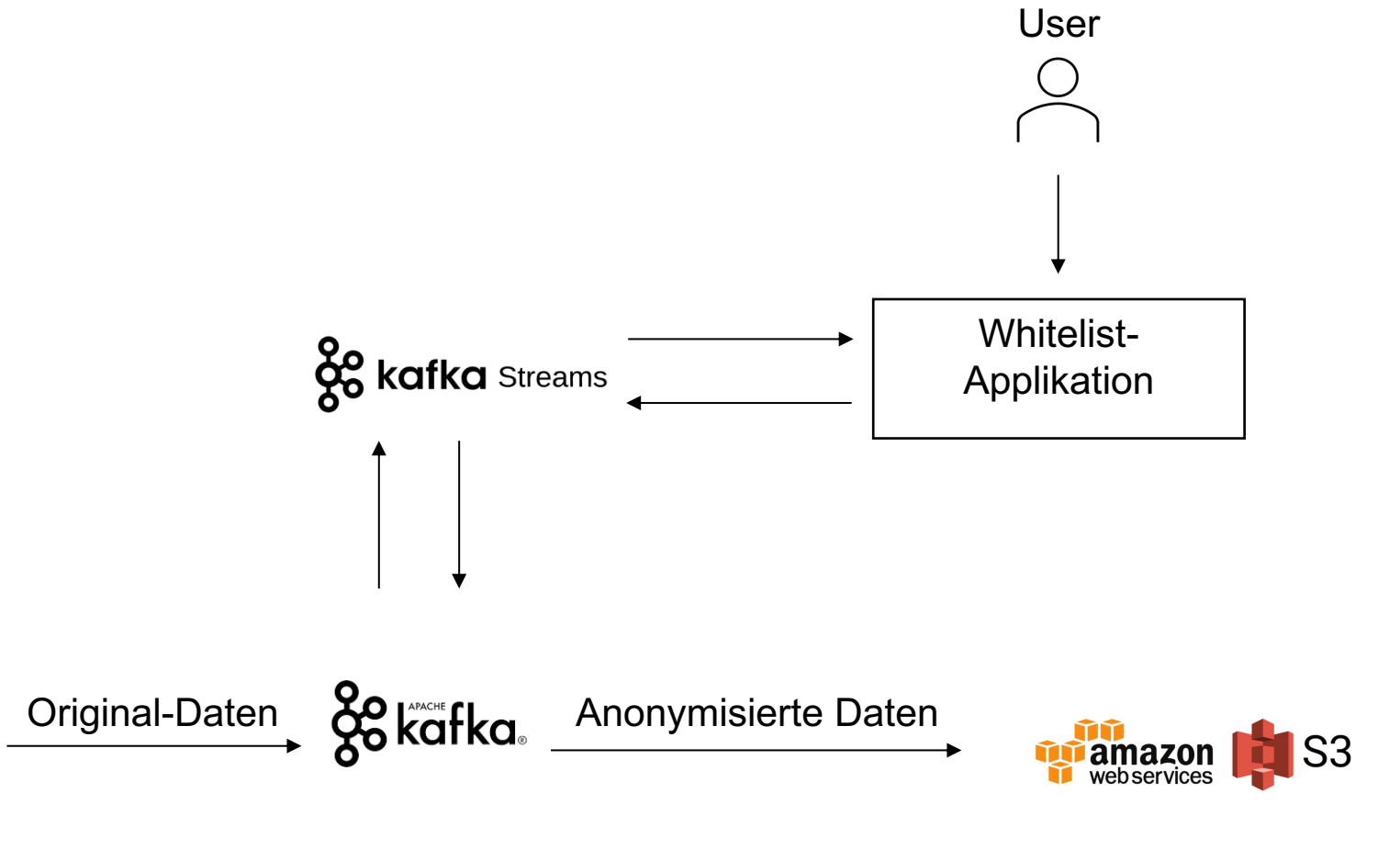
- Aufteilung gemäß Anforderungen
- Einstellbares Routing

Quality-Gates

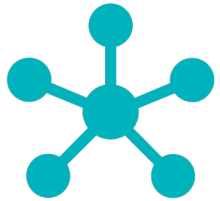
- Individuelle UI
- Verschiedene Modi

Datenanonymisierung

Technische Lösung



- Daten werden von Domänen an Kafka geschickt
- Separate Kafka Streams-Applikation greift Daten sofort ab
- Erhält aus eigens entwickeltem System sog. Whitelist
- User können Whitelist über UI pflegen und Anonymisierung konfigurieren



iteratec

Herausforderung: Komplexität

Datentransformation

Komplexität beherrschen

iteratec

- Über 2.000 Transformationen pro Tag
- Unterschiedliche Granularität und Load
- Abhängigkeiten bilden komplexe Verarbeitungsgraphen

Datentransformation

Komplexität beherrschen

iteratec

- Über 2.000 Transformationen pro Tag
- Unterschiedliche Granularität und Load
- Abhängigkeiten bilden komplexe Verarbeitungsgraphen

- Smartes Scheduling und Monitoring

Datentransformation

Komplexität beherrschen

iteratec

- Über 2.000 Transformationen pro Tag
- Unterschiedliche Granularität und Load
- Abhängigkeiten bilden komplexe Verarbeitungsgraphen

- Smartes Scheduling und Monitoring
- Eigenentwickeltes Test-Framework für anwendungsfallorientierte End-To-End-Tests

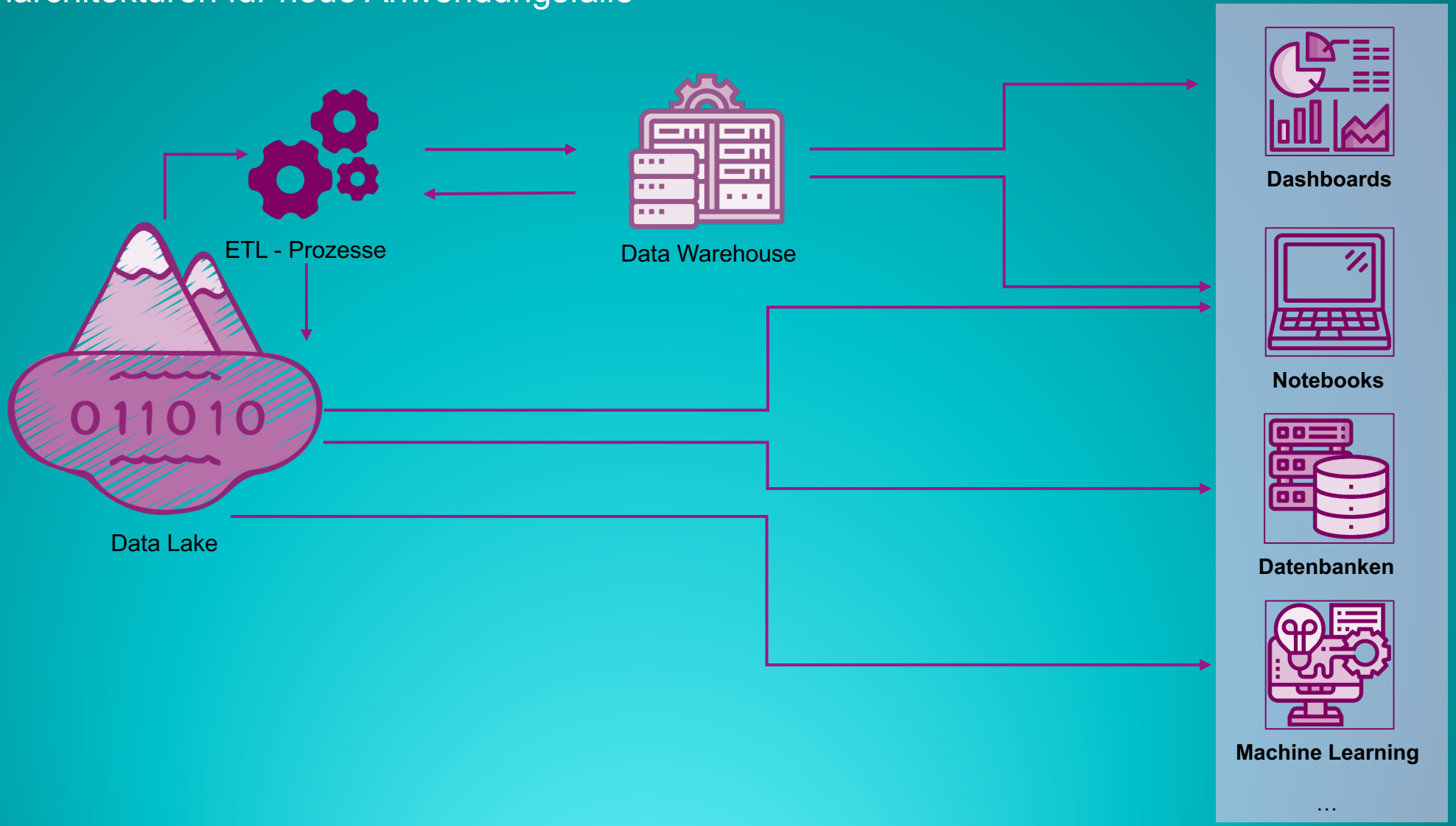
iteratec

Herausforderung: Austauschbar- und Erweiterbarkeit

Data Services

Flexible Datenarchitekturen für neue Anwendungsfälle

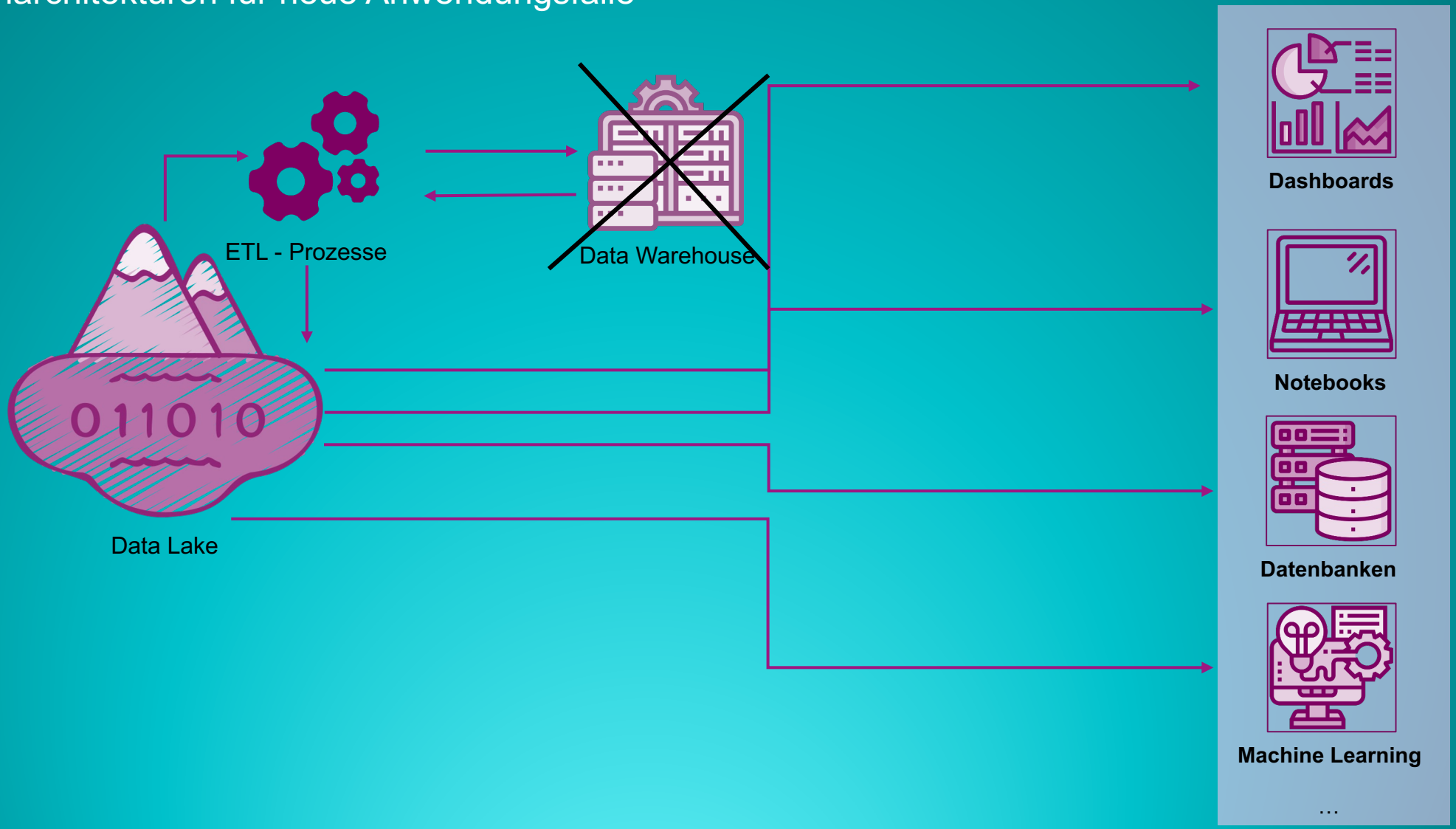
iteratec



Data Services

Flexible Datenarchitekturen für neue Anwendungsfälle

iteratec



iteratec

Abschluss

Key Take-Aways

iteratec

Warum gutes Data-Engineering gutes Software-Engineering voraussetzt

- Datenarchitekturen sind weiter auf individuelle Software-Lösungen angewiesen
  - Vorhandene Tools müssen ggf. ergänzt, erweitert oder verbaut werden
  - Lösungen haben nur marginal mit „Data“ zu tun

- Datenarchitekturen haben wie jedes andere Software-System
  hohe Anforderungen an Security, DevOps, Logging, Modularität, etc.
  - Data-Engineers müssen Systeme hoch-verfügbar, robust, resilient, fehlertolerant, sicher halten

- Integratives Testen im Data-Bereich benötigt bis heute viel „Handarbeit“

- Cutting-Edge-Technologien brauchen zur Integration umfangreiches Engineering-Wissen
  - Je neuer desto weniger Erfahrung, Dokumentation, …

iteratec

Dr. Masud Afschar

Senior AI & Data Engineer

iteratec GmbH
Westhafenplatz 1
60327 Frankfurt

+49 170 3748731
masud.afschar@iteratec.com