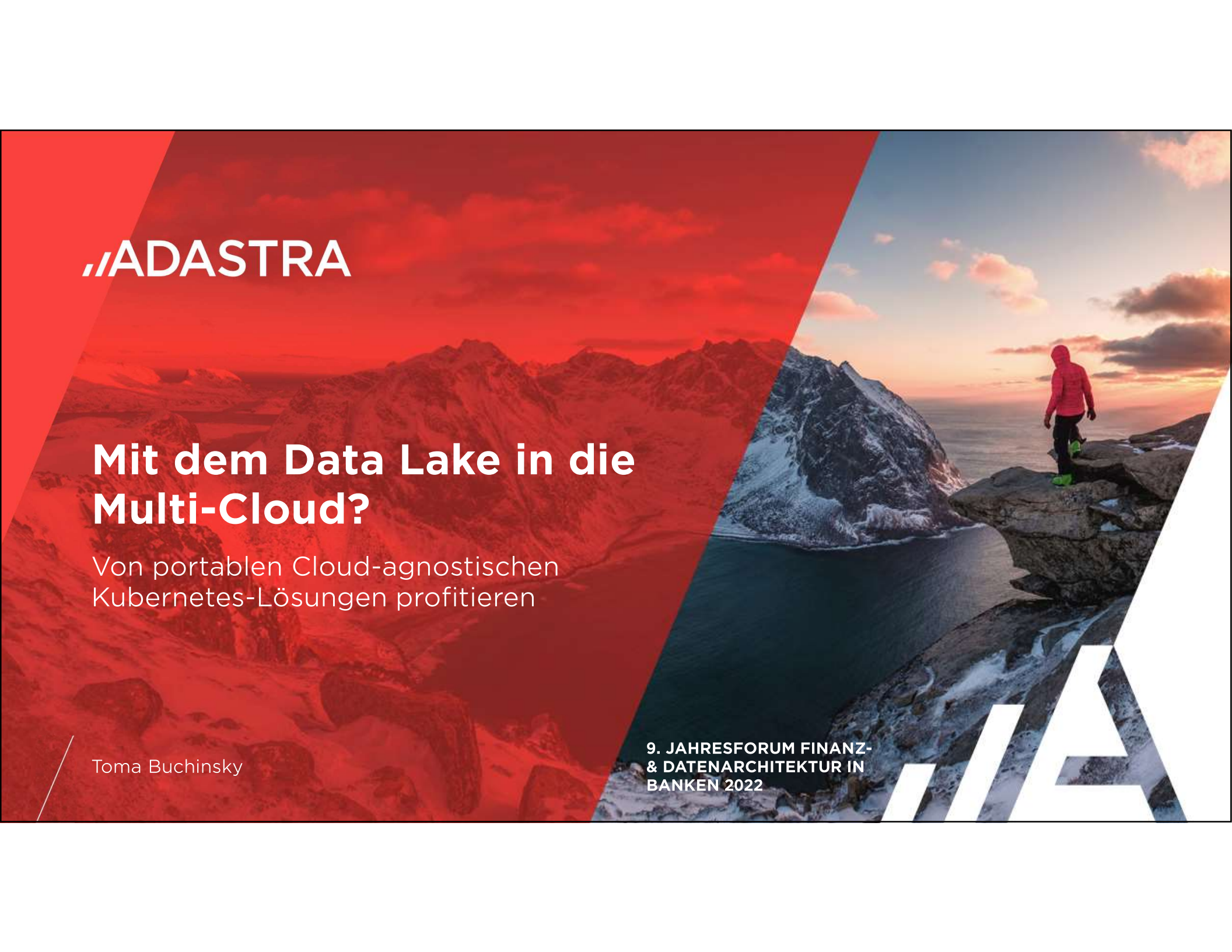




„ADASTRA

Mit dem Data Lake in die Multi-Cloud?

Von portablen Cloud-agnostischen Kubernetes-Lösungen profitieren

Toma Buchinsky

9. JAHRESFORUM FINANZ-
& DATENARCHITEKTUR IN
BANKEN 2022

./A

Über Adastra



Adastra Group weltweit

2000+

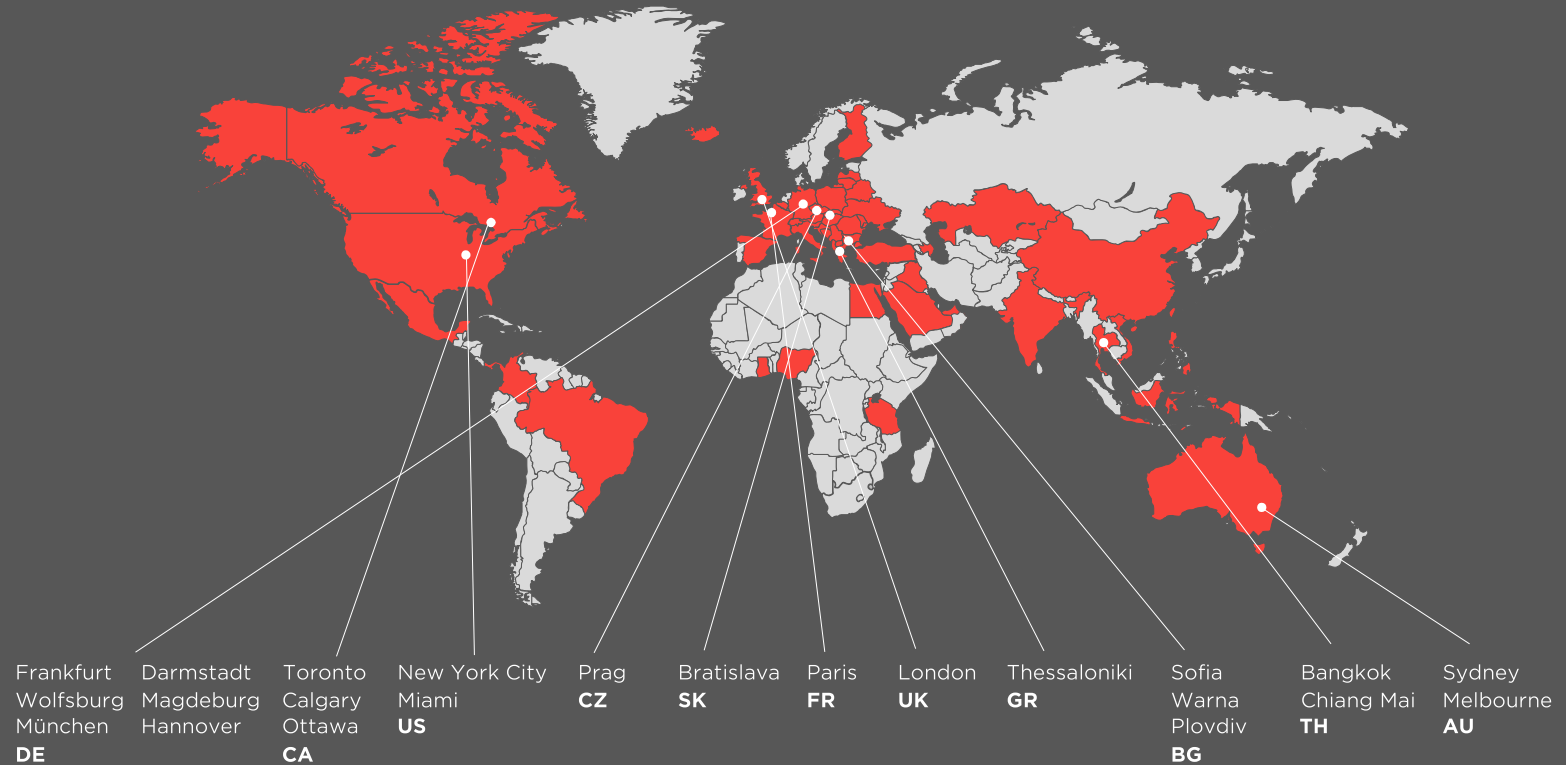
Experten auf der
ganzen Welt

5000+

Projekte in 42
Ländern

23

Standorte in 11
Ländern





Unser Leistungsportfolio



DATA ENGINEERING

Data Strategy
Modern Data
Warehousing
Data Lake
Data Integration
Data Visualization
Business Intelligence



ARTIFICIAL INTELLIGENCE

Machine Learning
Deep Learning
Statistical Analysis
Text Mining
Exploratory Data
Analysis
Visual Analytics
Feature Engineering



CLOUD SERVICES

Readiness
Assessment
Cloud Provider
Evaluation
Cloud Migration
Managed Services
Azure, AWS, GCP



DIGITAL BUSINESS

Digital
Transformation
Robotic Process
Automation (RPA)
Internet of Things
(IoT)
Blockchain
Mobile Apps



GOVERNANCE

Data Governance
Data Quality
Master/Reference
Data Management
Data Lineage
Metadata
Management



./A

Multi-Cloud Data Analytics



Gartners Vorhersage zu Multi-Cloud

Laut Gartner®s Report:

“Through 2026, **90% of data management tools** and platforms **that fail to support multicloud** and hybrid capabilities will be set for **decommissioning within three years.**”

(Predicts 2022: Data Management Solutions Embrace Automation and Unification,
Quelle: www.gartner.com, 01. Dezember 2021)



Aktuelle Entwicklungen bei der Cloud-Adoption

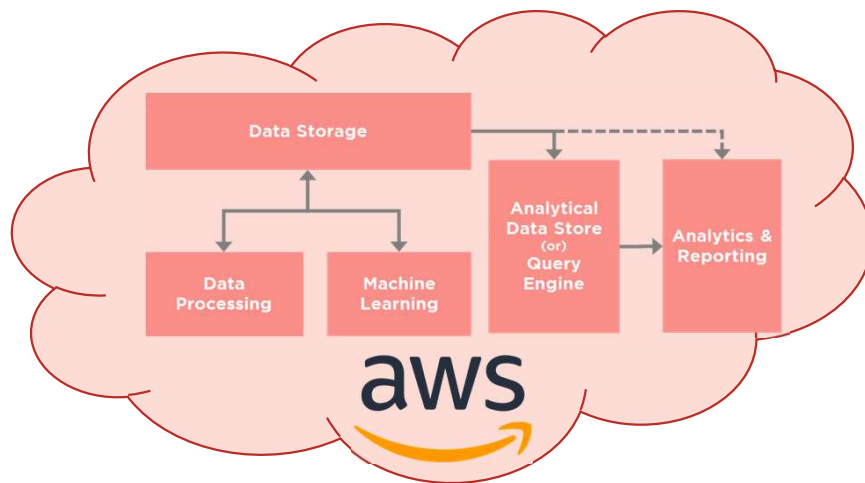
- / Die Nutzung von **mindestens einer öffentlichen Cloud-Plattform** ist heutzutage ein Standard
- / Auch die **Art der Nutzung weitet** sich immer mehr **aus**, von einfacher Daten- und Dateispeicherung und Webserver-Hosting bis hin zur **Implementierung komplexer, auf Microservices** basierender Anwendungen und Plattformen für Advanced Analytics und Machine Learning.
- / Viele große Unternehmen kündigen **strategische Partnerschaften mit** mindestens zwei **Hyperscalern** an





Beispiele aus der realen Welt: Globaler Autohersteller

Data Lake für Produktion & Logistik

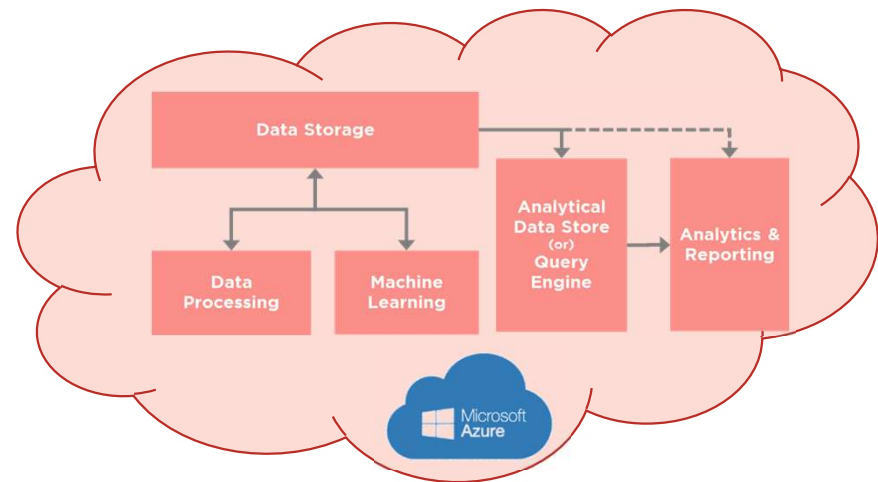


Storage: AWS S3

Data Processing & ML: AWS Glue, AWS EMR

Analytische Database: AWS Redshift

Sales Data & Analytics



Storage: Azure Data Lake Storage

Data Processing & ML: Azure Databricks, Azure Data Factory
Azure Synapse Analytics, SQL. Stored Procedures

Analytische Database: Azure Synapse SQL

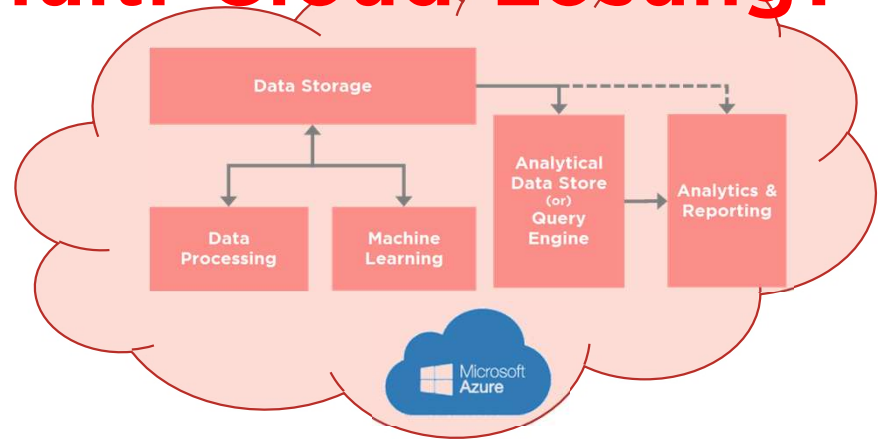
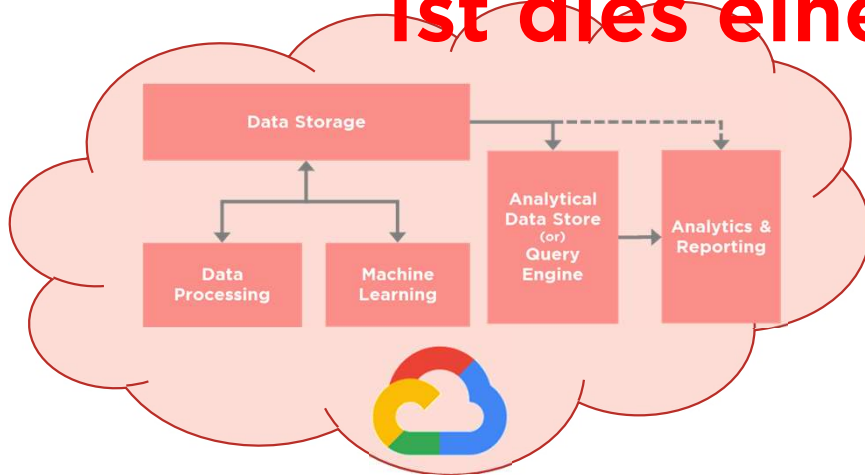


Beispiele aus der realen Welt: Große europäische Bank

Sales & Kundenanalyse

Kreditrisiko-Analyse

Ist dies eine Multi-Cloud-Lösung?



Storage: Google Cloud Storage

Data Processing & ML: Dataproc, GCP Dataflow

Analytische Database: Google BigQuery

Storage: Azure Data Lake Storage

Data Processing & ML: Azure Databricks, Azure Data Factory
Azure Synapse Analytics, SQL. Stored Procedures

Analytische Database: Azure Synapse SQL

Was ist
Multi-
Cloud?!



Laut Wikipedia:

- / Nutzung von **Cloud Computing und Speicherdiensten verschiedener Cloud-Plattformen** in einer einzigen heterogenen Architektur
- / ≠ Hybrid-Cloud bietet kombinierten Zugang zu abstrahierten IT-Infrastrukturen aus den Bereichen Public Clouds, Private Clouds und on-prem, nach den Bedürfnissen ihrer Nutzer

Quelle: <https://en.wikipedia.org/wiki/Multicloud>



Hauptfaktoren für die Multi-Cloud-Adoption



- / Vermeidung eines Vendor Lock-ins
- / Möglichkeit, dieselbe Anwendung gleichzeitig auf verschiedenen Cloud-Plattformen einzusetzen (z. B. aufgrund regionaler Beschränkungen)



Beispiele aus der realen Welt: Große europäische Bank

Keine Multi-Cloud-Anwendung?

~~Kreditrisiko-Analyse
(mit Azure-eigenen
Komponenten und Tools)~~



Kreditrisiko-Analyse
(mit Azure-eigenen
Komponenten und Tools)



Storage: Azure Data Lake Storage

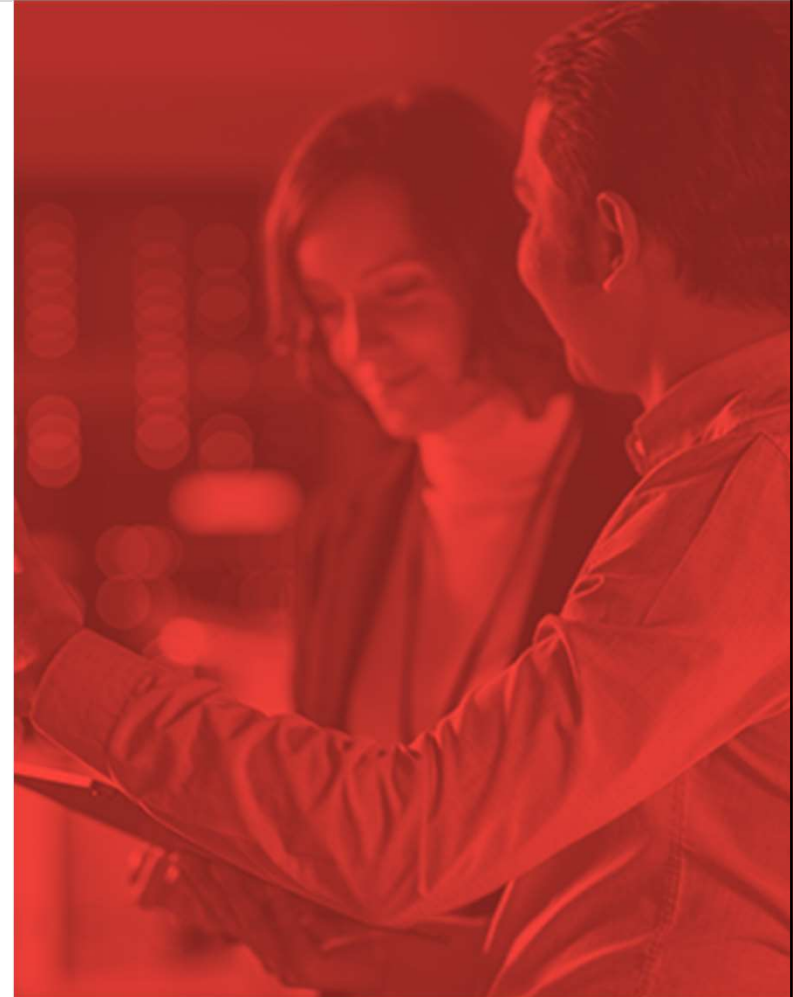
Data Processing & ML: Azure Databricks, Azure Data Factory
Azure Synapse Analytics, SQL. Stored Procedures

Analytische Database: Azure Synapse SQL



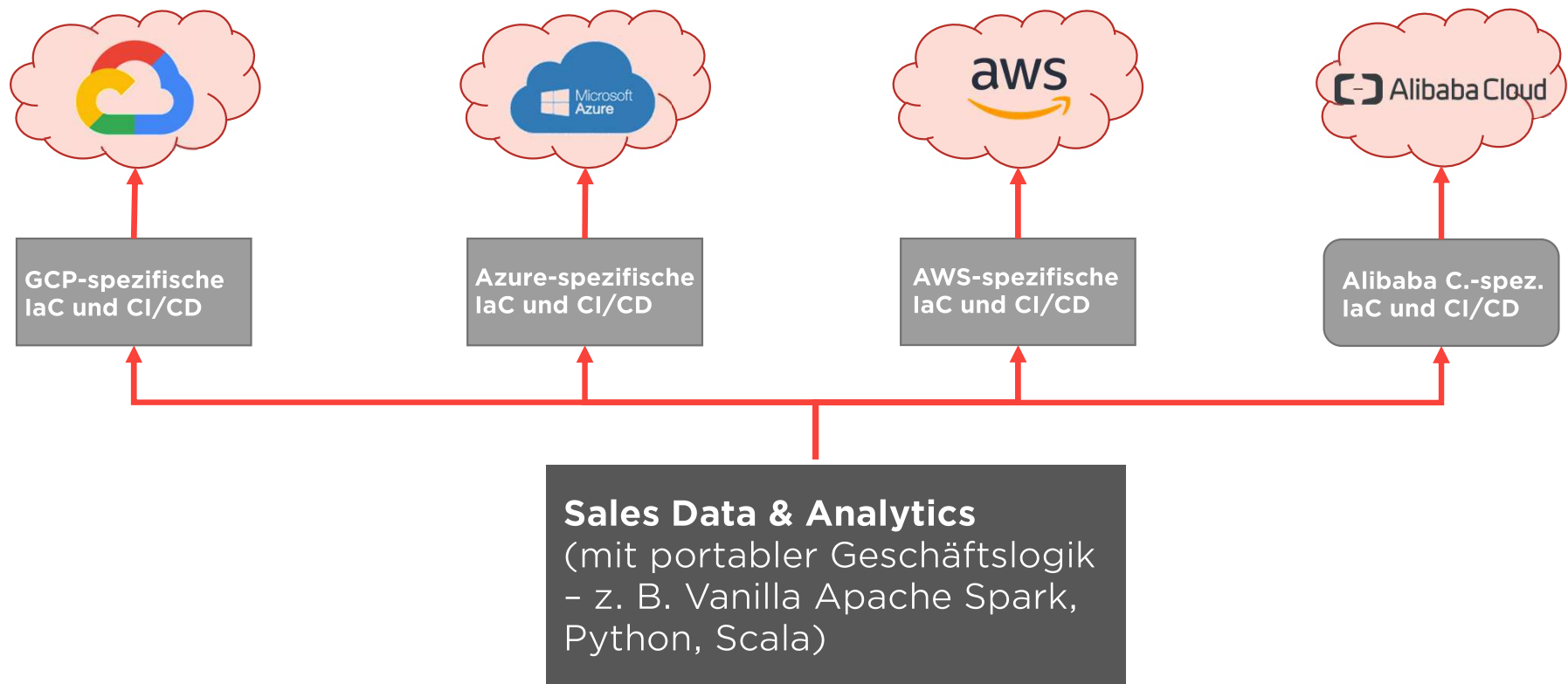
Guidelines für Multi-Cloud-Anwendungen

- / Berücksichtigung der **Portabilität** bereits in der Design und Entwurfsphase
- / Definition von **Entwicklungsrichtlinien**, die die Entwicklung von **portierbarem Code** sicherstellen
- / **Vermeidung** von **technologiespezifischen Bibliotheken** und **Abhängigkeiten** (z. B. Vermeidung von AWS/Azure/GCP-Erweiterungen von Spark und ausschließliche Verwendung von Vanilla Spark)
- / **Vermeidung** von **plattformspezifischen Funktionalitäten**, wie z. B. Stored Procedures etc.
- / Verwendung möglichst vieler **wiederverwendbarer Komponenten**, so dass bei einer Migration nur diese Komponenten angepasst werden müssen
- / **Testautomatisierung** zur Gewährleistung einer einfachen Validierung nach einer Migration





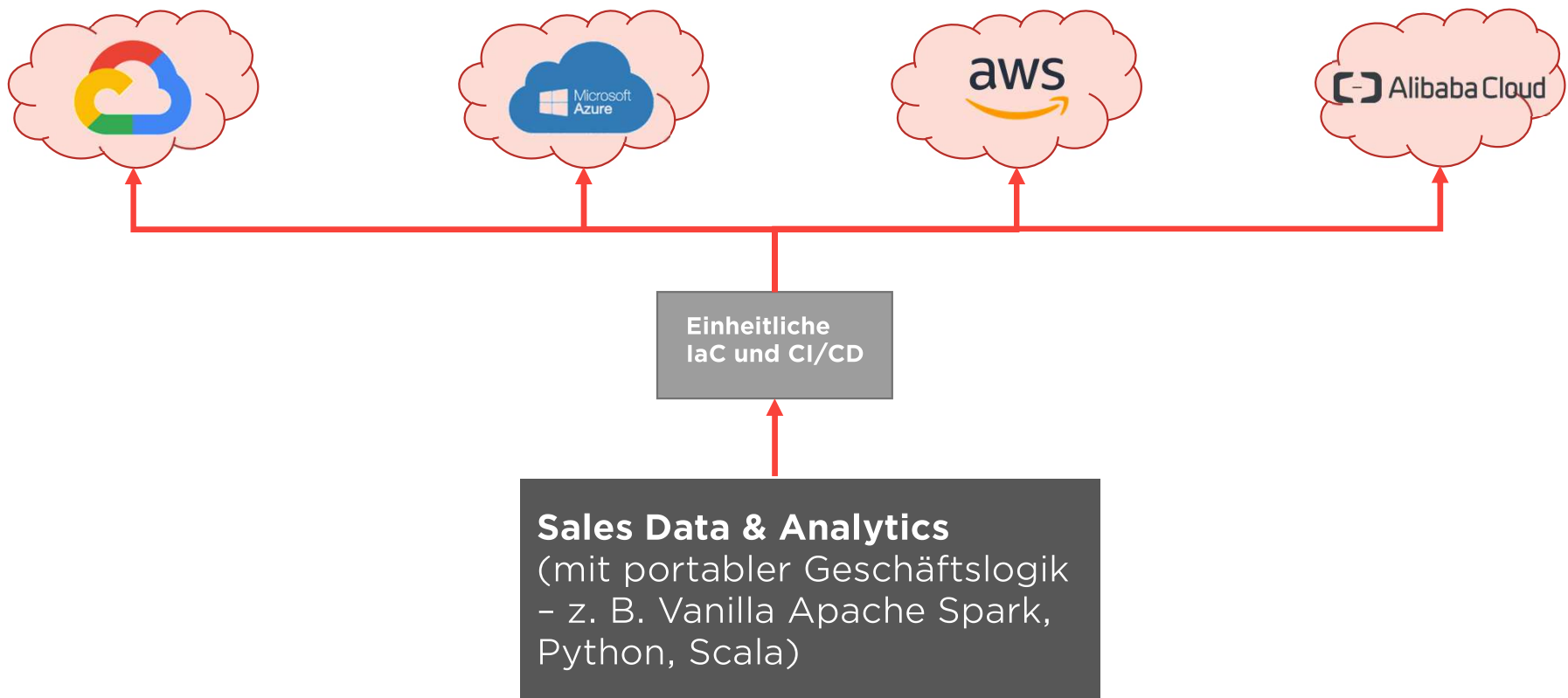
Portable Geschäftslogik & Cloud-native Basiskomponenten





Portable Geschäftslogik & einheitliche Basiskomponenten

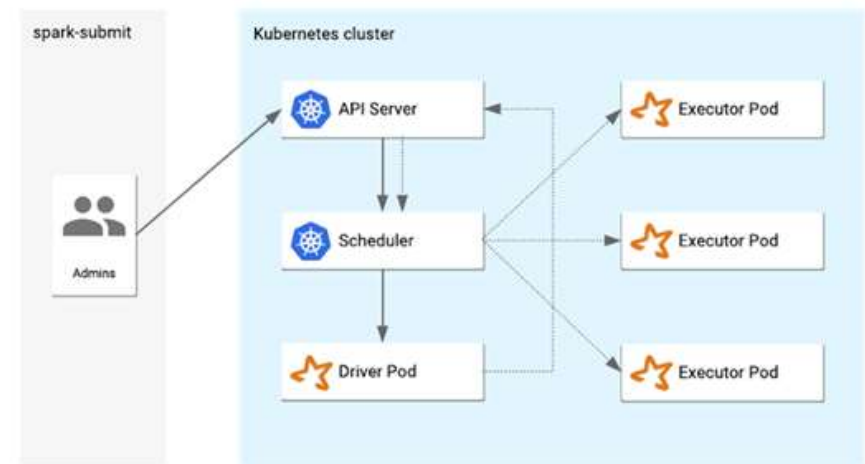
Ist eine solche Architektur möglich?





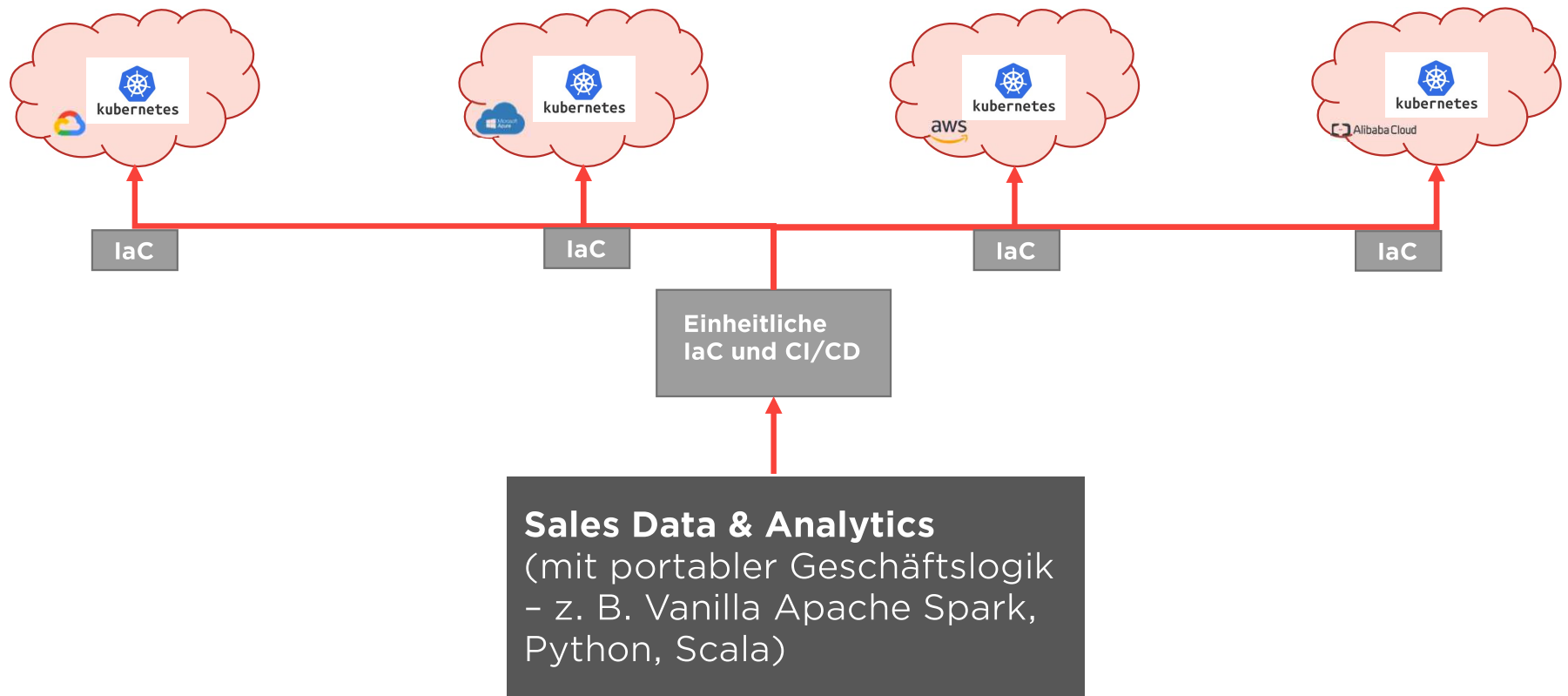
Portable D&A Plattform mit Spark & Kubernetes

- / **Kubernetes:** „*Run K8s Anywhere*“ - Kubernetes ist Open Source und bietet Dir die Freiheit, die Infrastruktur vor Ort, Hybrid oder Public Cloud zu nutzen. So kannst Du Workloads mühelos dorthin verschieben, wo es Dir wichtig ist.“ (<https://kubernetes.io/de/>)
- / **Apache SPARK:** Bereits für die on-prem und nun auch für die Cloud Data Lakes eine der meistgenutzten Plattformen zur parallelen Verarbeitung von großen Datenmengen
- / Seit 2018 ist **Kubernetes** als eine Alternative für Apache Spark Cluster Manager (neben Mesos & Yarn) verfügbar





Portable Geschäftslogik & einheitliche Basiskomponenten



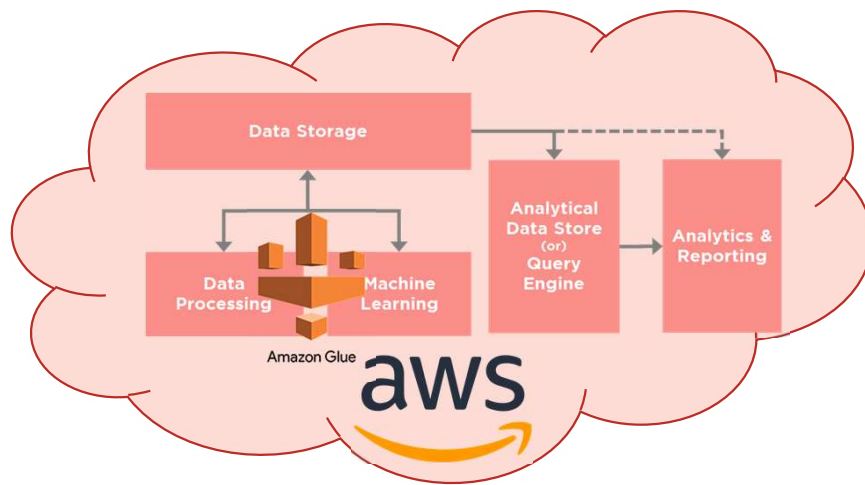


Zentralisierte Big Data Analytics Plattform

Vermeidung eines Vendor Lock-ins



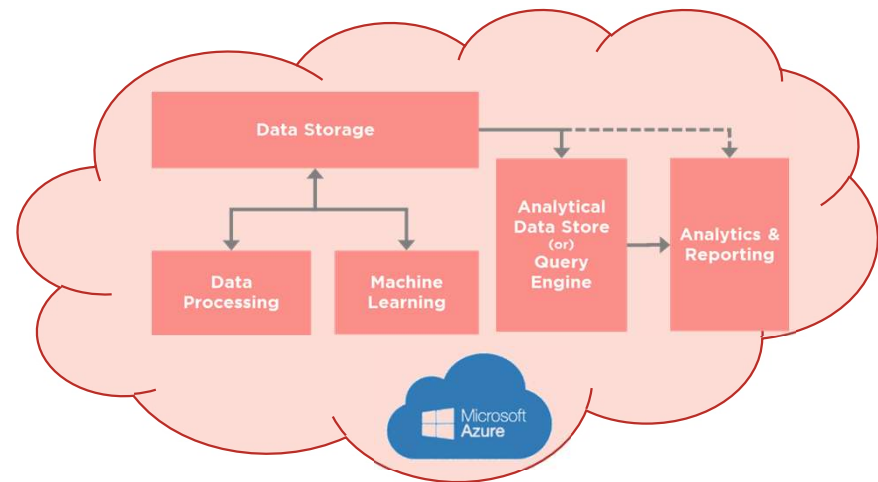
Data Lake mit Cloud-nativen Komponenten



Storage: AWS S3

Data Processing & ML: AWS Glue, AWS EMR

Analytische Database: AWS Redshift



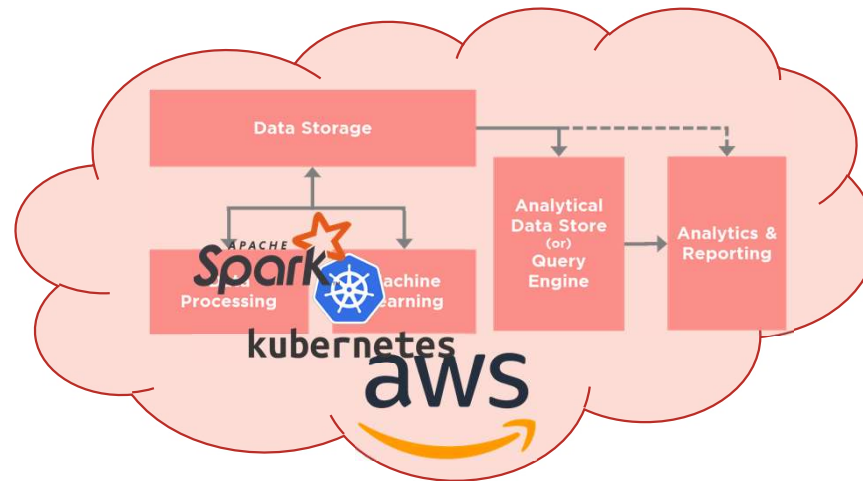
Storage: Azure Data Lake Storage

Data Processing & ML: Azure Databricks, Azure Data Factory
Azure Synapse Analytics, SQL. Stored Procedures

Analytische Database: Azure Synapse SQL



Portabler Data Lake mit Spark+K8s auf AWS



Storage: AWS S3

Data Processing & ML: Apache SPARK auf Amazon Elastic Kubernetes Service (EKS)

Analytische Database: AWS Redshift



Multi-Cloud-fähige Analytics Plattform



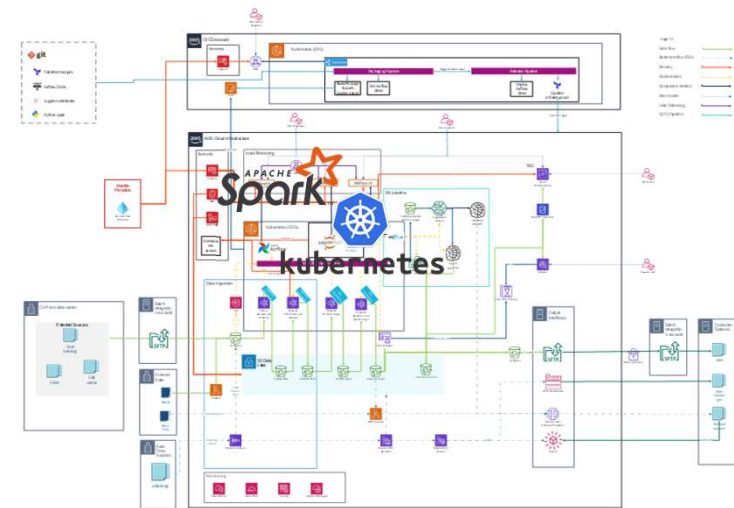
ZIELE

- Eine EU Bankengruppe errichtet eine brandneue Cloud-basierte Data Analytics Plattform
- Um einen Vendor Lock-in zu vermeiden, wurden Kubernetes und Containerisierung als Architektur-Konzept gewählt
- Verwendeter Tool-Stack: Spark auf Kubernetes, EKS, S3 Object store, Sagemaker, Jupyter, Airflow, Jenkins
- Die gesamte Infrastruktur der Plattform wird als Infrastructure as Code (Terraform) bereitgestellt.

Eine Multi-Cloud-fähige Lösung auf Kubernetes



LÖSUNG



NUTZEN

- Vermeidung eines Vendor Lock-ins durch die Verwendung von nicht ausschließlich AWS-eigener Services
- Verwendung von Spark auf Kubernetes als Haupttool zur Datenverarbeitung
- Alle Anwendungskomponenten werden mit Kubernetes bereitgestellt und gewartet
- Bei Bedarf Rollback zu on-prem oder zu einem anderen Cloud-Anbieter



Portable Big Data Lösung auf Kubernetes



ZIELE

Ein großer Automobilhersteller benötigt eine **portable** Data Analytics Plattform basierend auf Kubernetes-Clustern und Apache Spark.

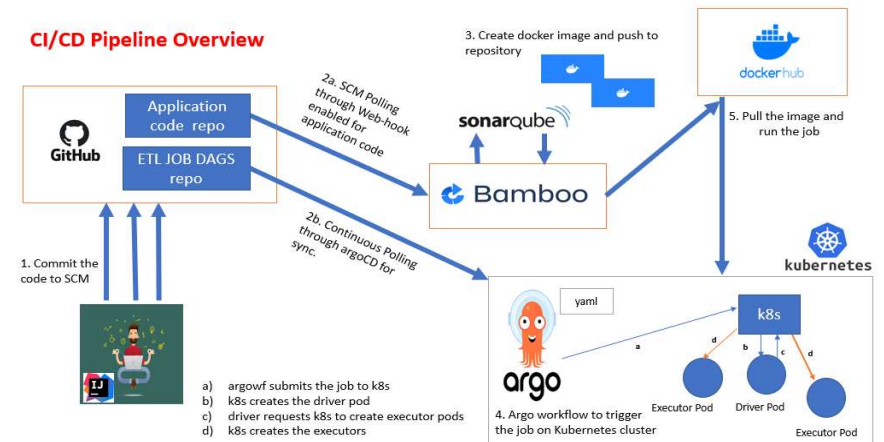
- Adastra implementierte das **Data Lake Framework** sowie die CI/CD-Pipelines Automatisierung
- Mehrere Systeme auf dem Cluster (FE/BE, Data Management, Analytics)
- Umgebung: Openshift, Bamboo, BitBucket, Presto, Argo, Helm, Nifi, S3, Prometheus, Grafana, ELK Stack etc.
- Das DevOps-Team wird zwischen den Projekten geteilt und sorgt für einheitliche DevOps-Prozesse

Eine portable Big Data Lösung auf Kubernetes



LÖSUNG

CI/CD Pipeline Overview



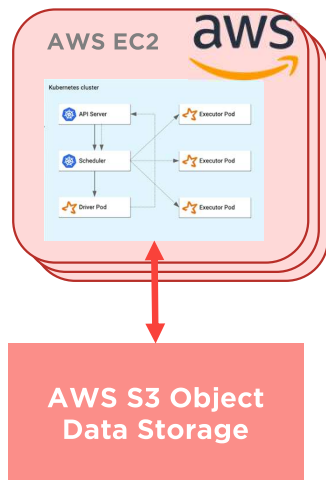
NUTZEN

- Ein vollständig automatisiertes Ökosystem für Entwicklungs-, Test- und Deployment-Pipelines
- Scheduled Workflows, die täglich Millionen von Datensätzen verarbeiten
- Einfache Releases/Rollback während Aktualisierungen, um high Availability und Zero down-time zu gewährleisten
- Zentralisiertes Konfigurationsmanagement
- Portable, selbstheilende, automatisch skalierbare Infrastruktur mit Cloud-agnostischen Funktionen
- Einfache Überwachung und Unterstützung mit Prometheus, Kibana und Grafana Dashboards



Portabel von Anfang an

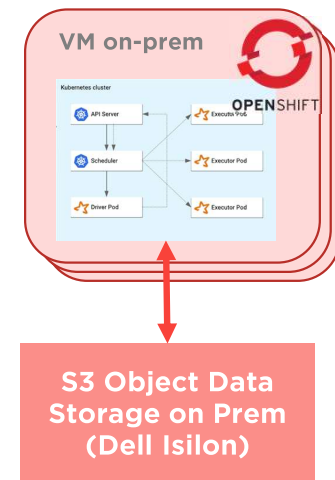
Erster Prototyp
entwickelt im Adastra
Germany Data
Engineering Lab (AWS)



Nachfolgende Sprints in
AWS - Account des
Kunden

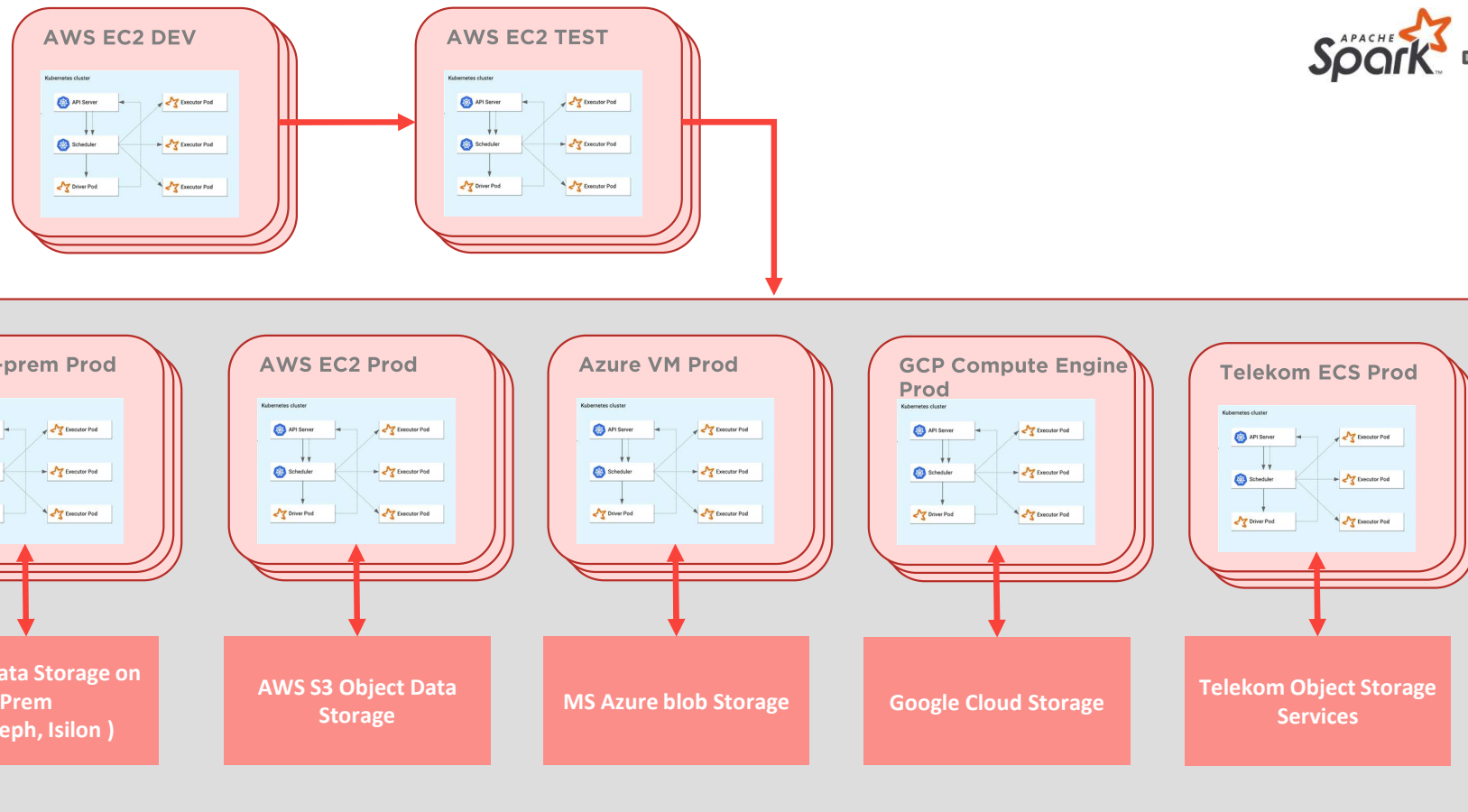


Endgültige Entwicklung
auf der on-prem
Infrastruktur des Kunden





Endziel: Ein Framework, überall einsetzbar



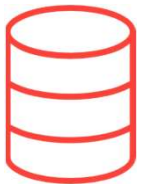


Nicht vergessen: Grundkonzepte



MULTI-CLOUDS

- Angesichts ihrer Komplexität und Verbreitung mehr Überlegungen zur Sicherheit und Governance erforderlich
- Vermeidung eines Vendor Lock-ins
- Haben nur dann einen Wert, wenn Sie die richtigen Anbieter auswählen, ob On-Demand oder privat, um Ihre Anforderungen zu erfüllen



CONTAINER (Kubernetes) sind das wichtigste Architekturkonzept, um Multi-Cloud realisierbar und bezahlbar zu gestalten



Cloud-agnostische Lösungen

Kubernetes Best Practices



Wie vermeidet man einen Vendor Lock-in?

Also was
muss ich
dafür tun?



VERWENDUNG EINER CONTAINER (KUBERNETES) ARCHITEKTUR

- / Dies stellt sicher, dass die **Anwendungen erweiterbar** und **wartungsfähig** sind, da alle Komponenten in kleinere und unabhängige Module unterteilt sind, um eine bestimmte Aufgabe zu erfüllen
- / Als Open-Source-Plattform übernimmt Kubernetes die Planung von **Containern** auf einem Compute-Cluster und verwaltet die Workloads, um sicherzustellen, dass sie in der Produktion **wie vorgesehen ablaufen**
- / Kubernetes ermöglicht eine **optimale Hardware-Nutzung**, es ist **portabel** und **auto-skalierbar**
- / Als Open-Source-Container-Orchestrierungstechnologie ermöglicht es Benutzern, ihre **Workloads in privaten, öffentlichen und hybriden Clouds auszuführen**





Zentralisierung des Managements

WIE KUBERNETES MULTI-CLOUD-HERAUSFORDERUNGEN MEISTERN KANN

Kubernetes kann Ihnen helfen, Ihre Multi-Cloud-Infrastrukturen zu verwalten und alle auftretenden Herausforderungen erfolgreich zu meistern.



**Einsatz in
Multi-Cloud-
Systemen**



Monitoring



Sicherheit



Kostenkontrolle



**EINE MULTI-CLOUD-
STRATEGIE IST KEIN MUST-HAVE,
ABER SIE HILFT IHNEN DABEI, VENDOR
LOCK-INS ZU VERMEIDEN.**



Wie geht es weiter?



Q&A

Haben Sie Fragen oder Anregungen?





Kontaktieren Sie uns!

Toma Buchinsky

CEO & Data Architect

Toma.Buchinsky@adastragrp.com



Adastra GmbH

Niedenau 36, 60325 Frankfurt am Main

+49 (0)69 719 779 790 / infoDE@adastragrp.com

www.adastragrp.com



Frankfurt / Wolfsburg / München / Hannover / Magdeburg / Darmstadt