ADASTRA

Von Data Warehouse über Data Lake zu Multi-Cloud Data Lake

ScaleUp 360° Big Data / 10. März 2021

Toma Buchinsky / Slavomir Krivak / Ankit Pandey

Presenters

Toma Buchinsky
CEO Adastra Germany & Data Architect
Toma.Buchinsky@adastragrp.com

Slavomir Krivak
Big Data & Cloud Architect
Slavomir.Krivak@adastragrp.com

Ankit Pandey
Senior Big Data & Cloud Architect
Ankit.Pandey@adastragrp.com

Agenda

Unser Solution Portfolio

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

Adastra Group weltweit

2000+

mehr als 2000 Experten

1000+

Projekte in 46 Ländern

20

Standorte in 10 Ländern

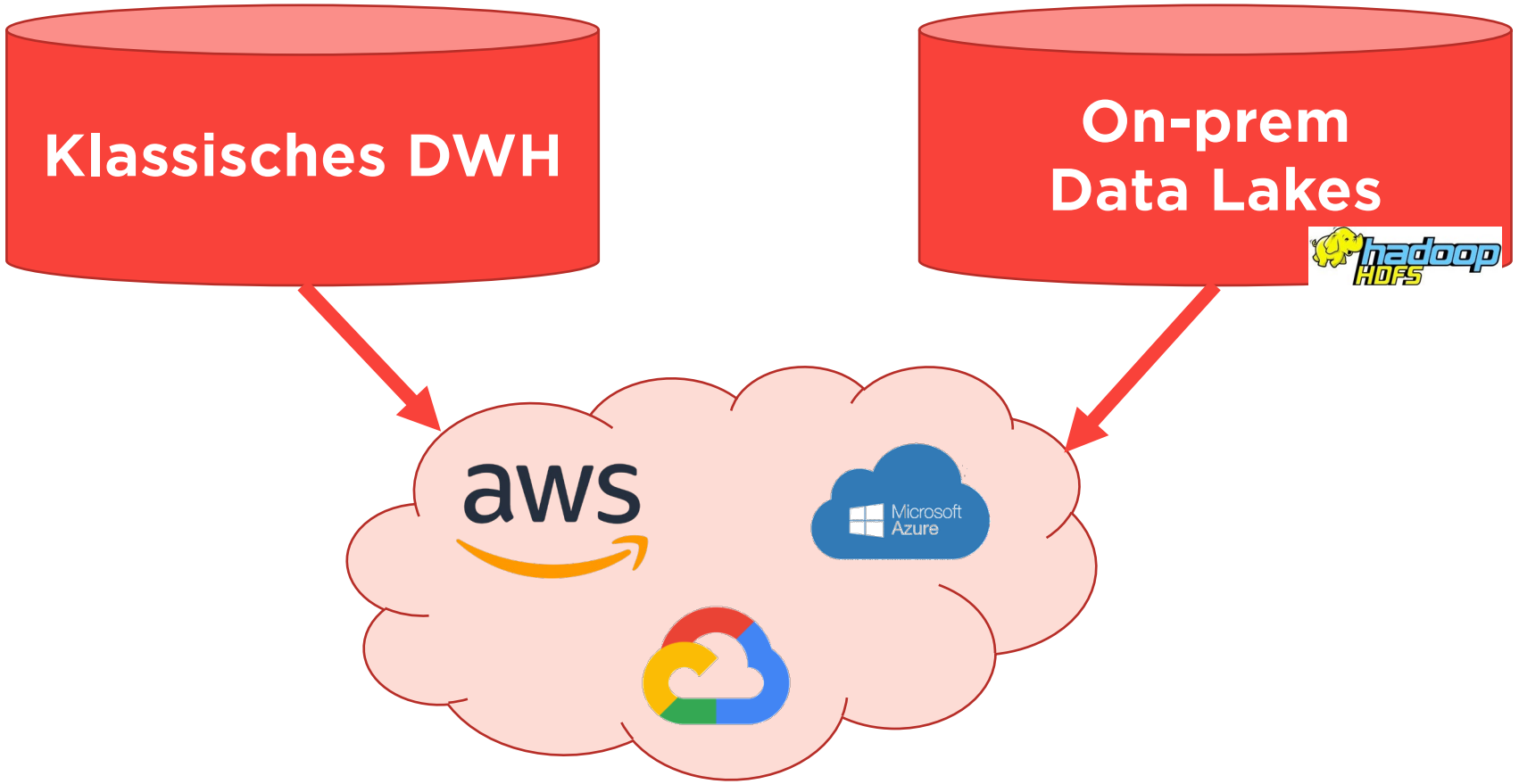
Die gute Nachricht

Deutsche Top Manager wollen ihre Unternehmen in

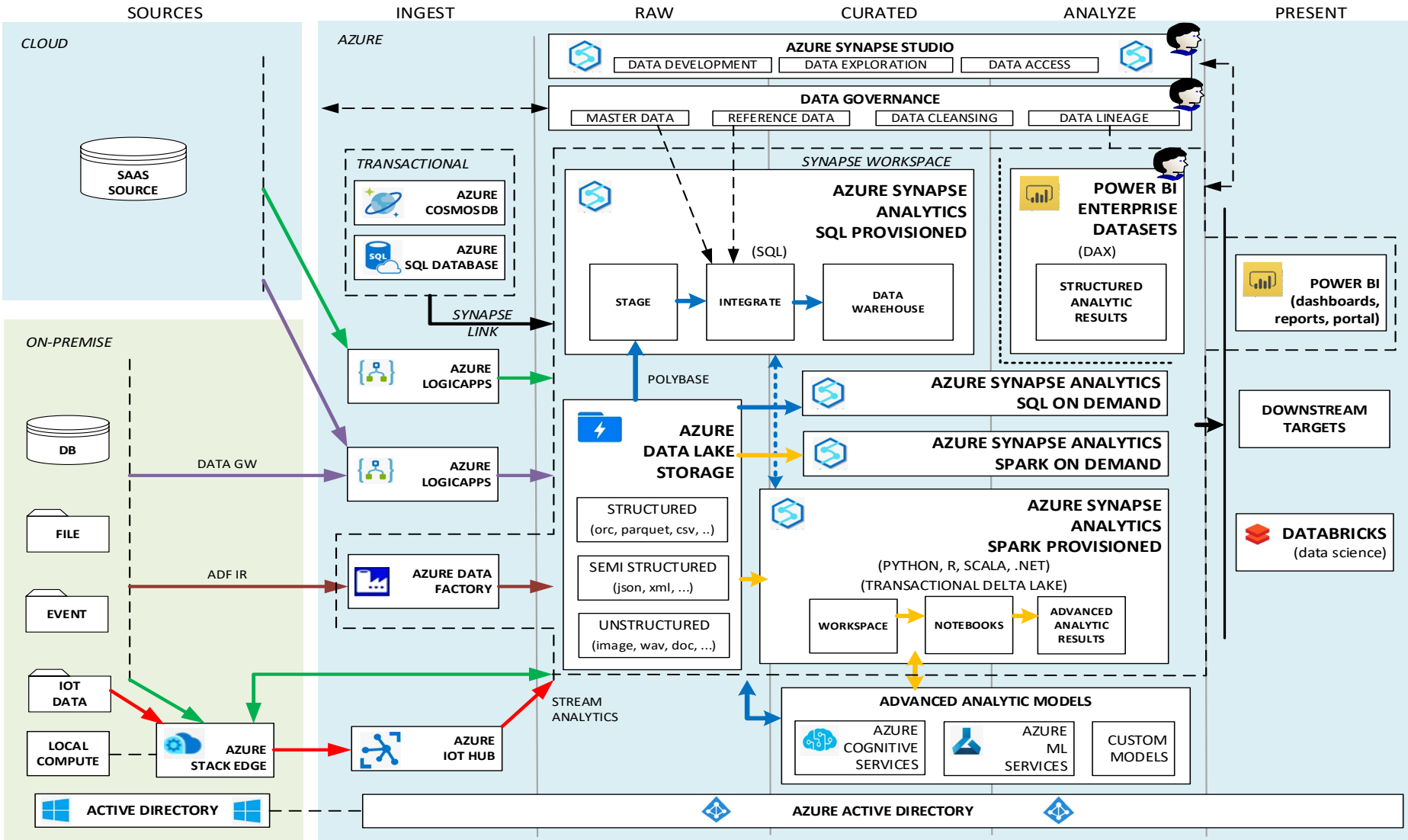
datengetriebene Unternehmen

transformieren.

Zeitalter der Cloud-First Programme



Was verkaufen große Cloud Anbieter?



...und erwarten euphorische Kunden

Kunden: Zurückhaltung und Skepsis

- Unsicherheit bezüglich IT-Sicherheit und Datenschutz
- Bedenken wegen US Cloud Act
- Fehlendes Know-how
- Komplexität und Kosten der Migration
- Große Investitionen in DWHs und Data Lakes in jüngster Vergangenheit machen es schwierig bei CFOs neue teuere Initiativen zu verkaufen
- Zurückhaltung wegen möglichem Vendor Lock-in

Cloud Einführung: Womit anfangen?

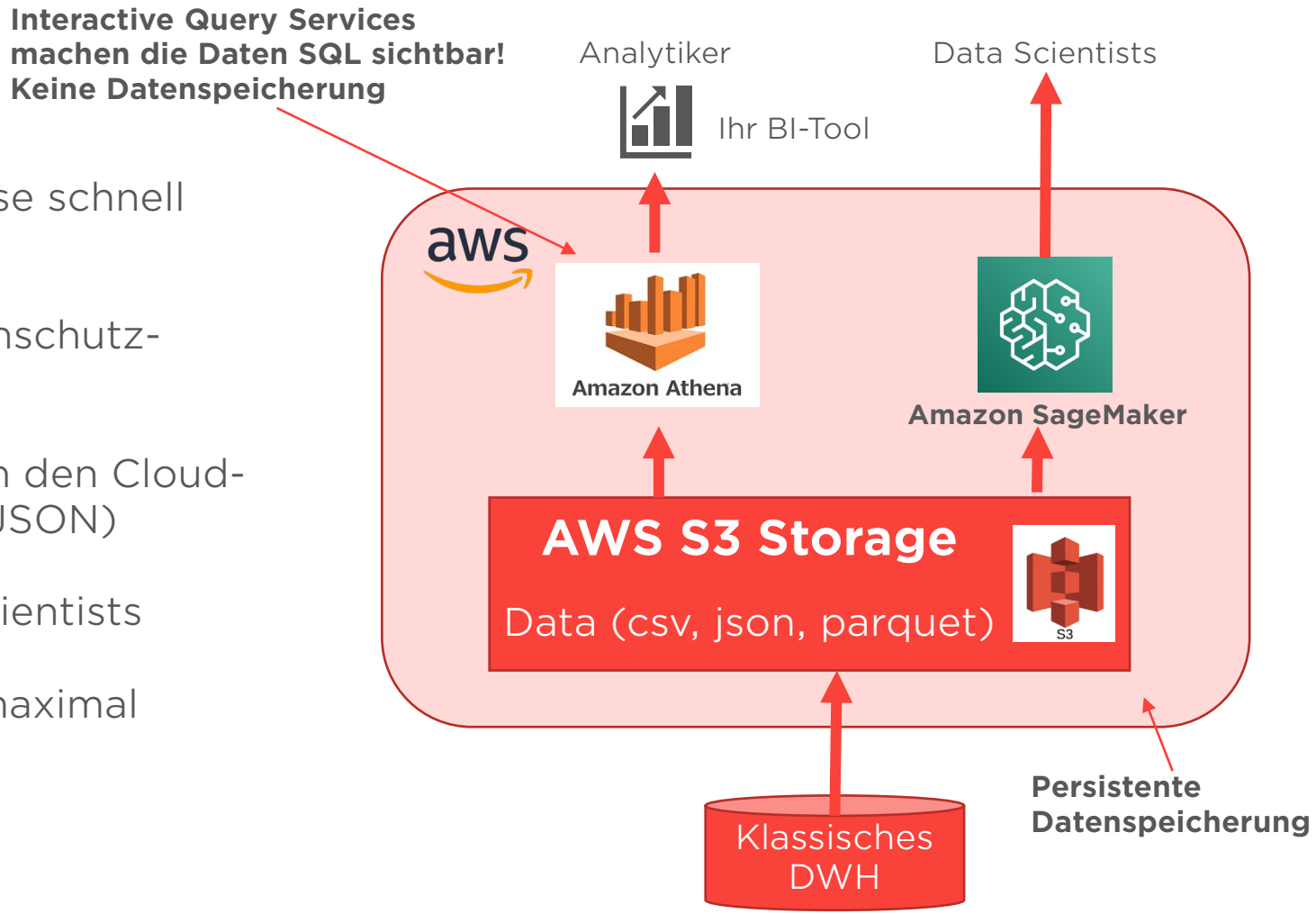
Don’ts:

- Big-Bang Approach: alles auf einmal oder nichts
- Langwierige Konzeptionsphase

Cloud Einführung: Womit anfangen?

Dos:

- Einen passenden, existierenden Use Case schnell identifizieren
- Freigabe durch IT-Sicherheit- und Datenschutz-Abteilungen
- Relevante Daten (Teilmengen) schnell in den Cloud-Storage hochladen (z. B. als CSV oder JSON)
- Datenzugriff für Analytiker und Data Scientists
- Inbetriebnahme des ersten MVP nach maximal 3 Monaten



Interactive Query Services

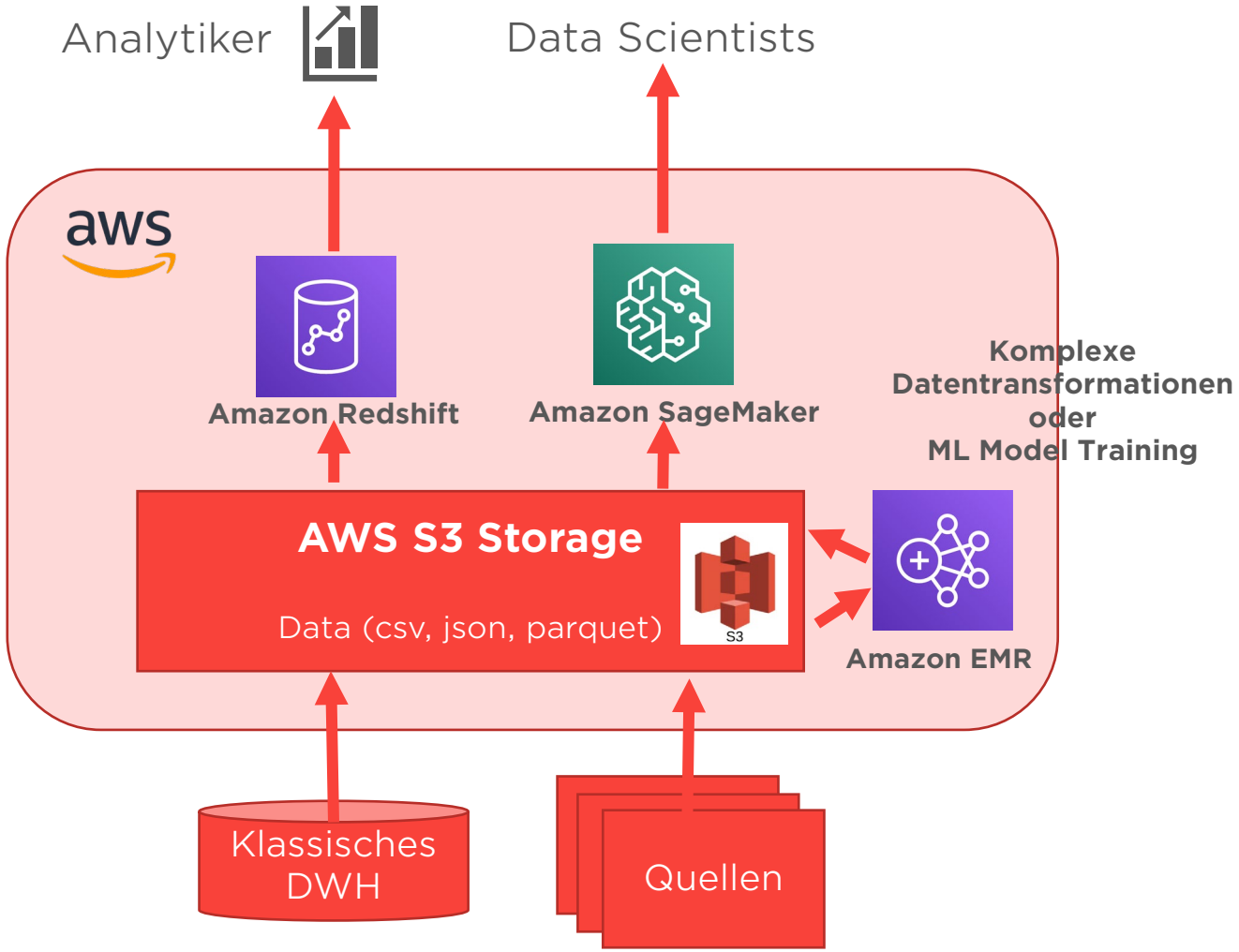
Vorteile

- Schnelle und unkomplizierte Einrichtung
- Datenzugriff über SQL auf Daten in den Object Storage
- Datenabfragen im native Format (csv, json etc.)
- Einfache Kostenstruktur – Preis basiert auf dem Volumen der abgefragten Daten

Nachteile

- Schlechte Performance bei komplexen Abfragen und großen Datenmengen
- Limitierung bei konkurrierenden Abfragen
- Keine User Defined Functions

Nächster Schritt: Cloud Data Processing & analytische Data Stores

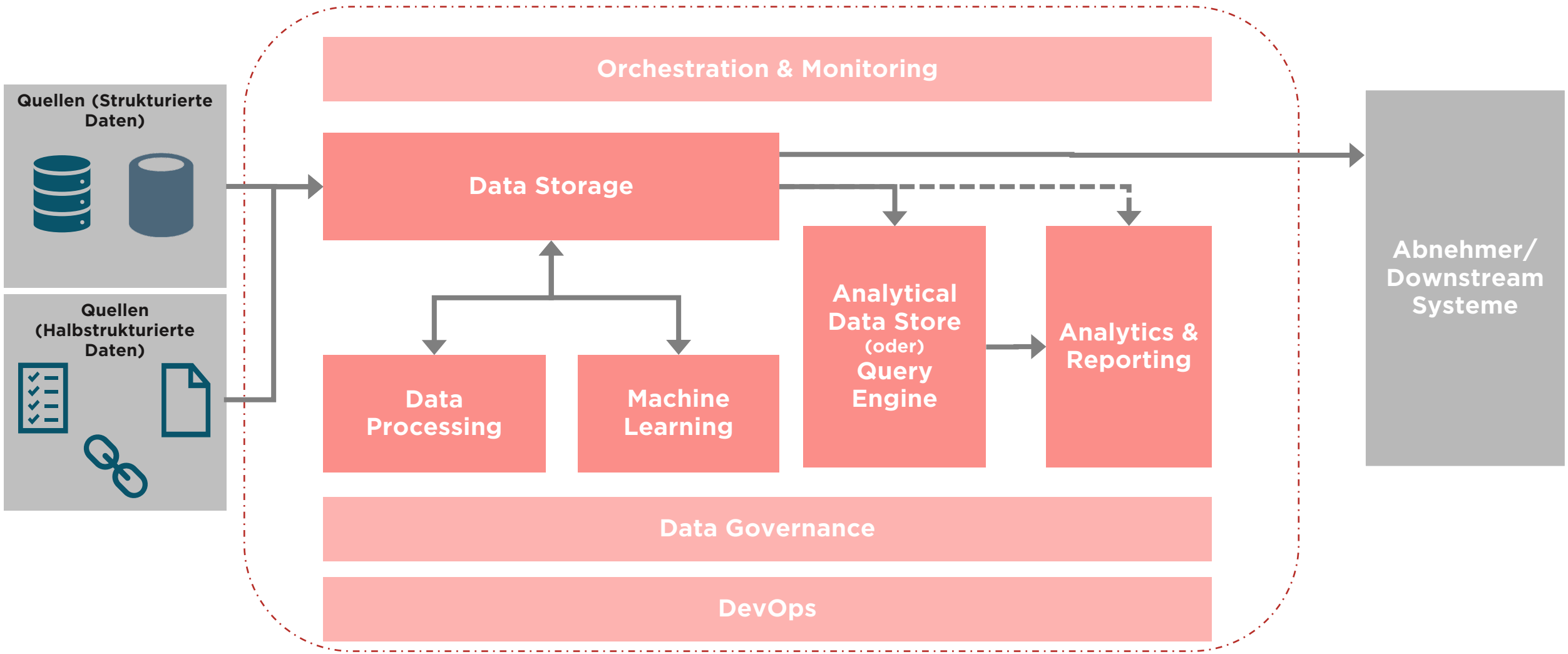


Was lerne ich von den ersten Cloud MVPs?

Erste Erfahrungen darüber...

- wie schnell neue Data & Analytics Anwendungen in die Cloud implementiert werden
- wie sich die Kosten entwickeln und wie ich sie in den Griff bekomme (deswegen MVP, nicht nur PoC)
- ggf. ob mein Legacy schnell, mit überschaubaren Kosten und am besten automatisch in die Cloud zu bringen ist?

Cloud Data Lake & Modern DWH



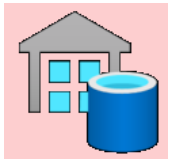
Muss man nur Cloud-native Technologien/Tools nutzen?

Nein!

Cloud Analytical Data Stores

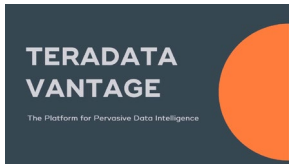
Cloud Native:

Amazon Redshift



MS Azure Provisioned SQL (SQLDW)

Alternativen:



Analytical Data Stores: Lock-in Vermeidung

- Betrachtung des Analytical Data Stores als eine SQL-Engine-Box, die analytische Abfragen ausführt…

  → mit akzeptabler Performanz
  → zu einem akzeptablen Preis
  → mit der notwendigen Elastizität

- Daten im Vorfeld transformieren/aggregieren (z. B. mit Apache SPARK oder ETL-Tools)

- Vermeidung von Plattform-native Funktionalitäten wie Stored Procedures etc.

Cloud Data Processing

Cloud Native:

AWS EMR

AWS Glue
Apache Spark
Jobs

Apache Spark
In Azure Synapse Analytics oder
Azure Databricks

Apache Spark
In Google Cloud
Dataproc

Alternativen:

Data Integration Tools
verfügbar in der Cloud

Vollintegrierte Data Management
Plattformen
Apache Spark / Hive / Apache NiFi
Jobs

Best-of-Breed Portable Data &
Analytics Plattform
als Kubernetes Cluster

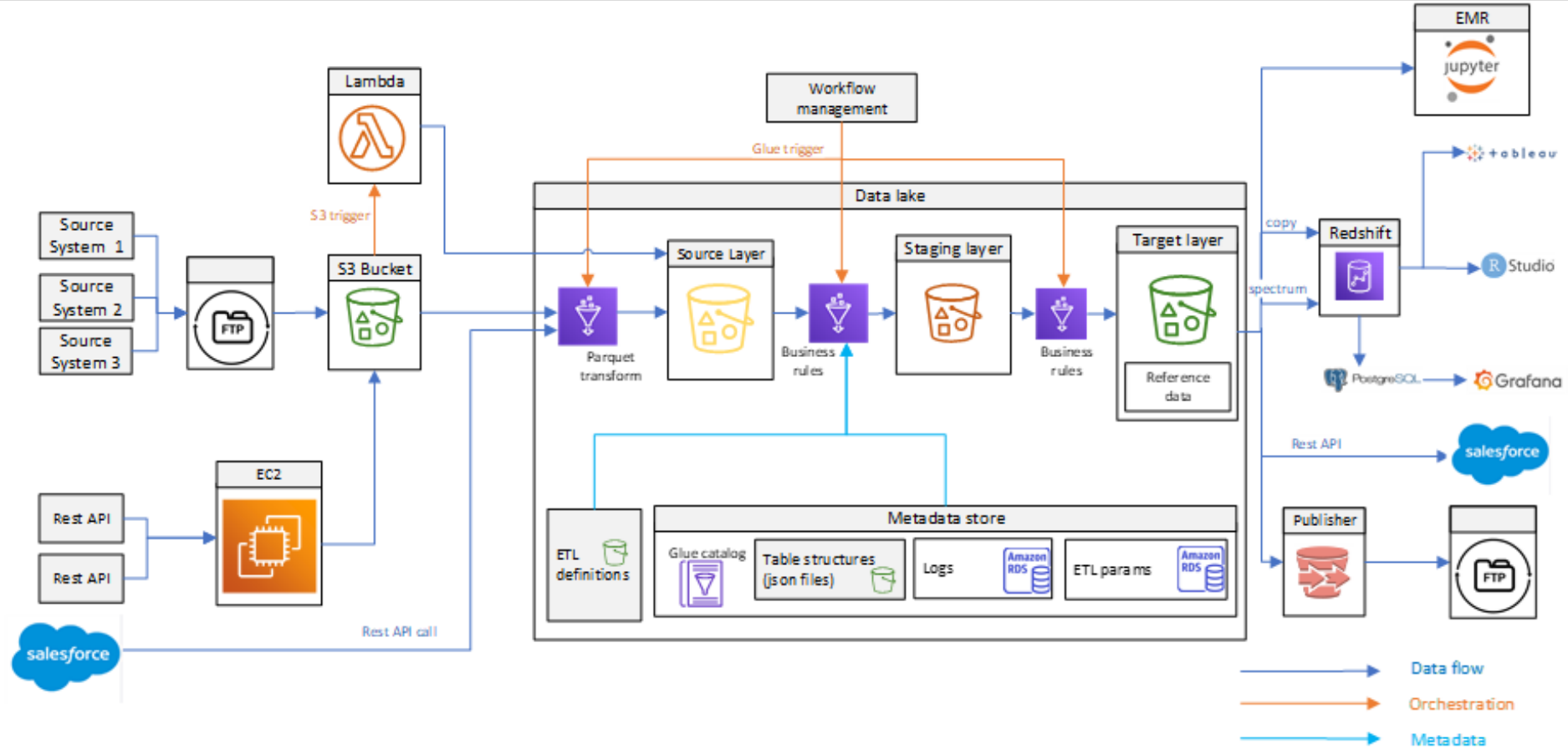
Cloud Data Processing: Lock-in Vermeidung

- Portabilität bereits während der Designphase beachten
- Definition von Entwicklungsrichtlinien, die dafür sorgen, portablen Code zu entwickeln
- Vermeidung von Technologie-spezifischen Bibliotheken und Abhängigkeiten (z. B. AWS Erweiterungen von Vanilla Spark)
- Möglichst viel Anwendung von wiederverwendbaren Komponenten, so dass bei einer Migration nur diese Komponenten anzupassen sind
- Test Automatisierung, damit nach einer Migration alles einfach zu validieren ist

Case Study 1: Multi-Cloud D&A with Azure DevOps

Automotive

CAP Data pipelines



amazon S3

S3 object storage leveraged for maintaining best possible value/price ratio

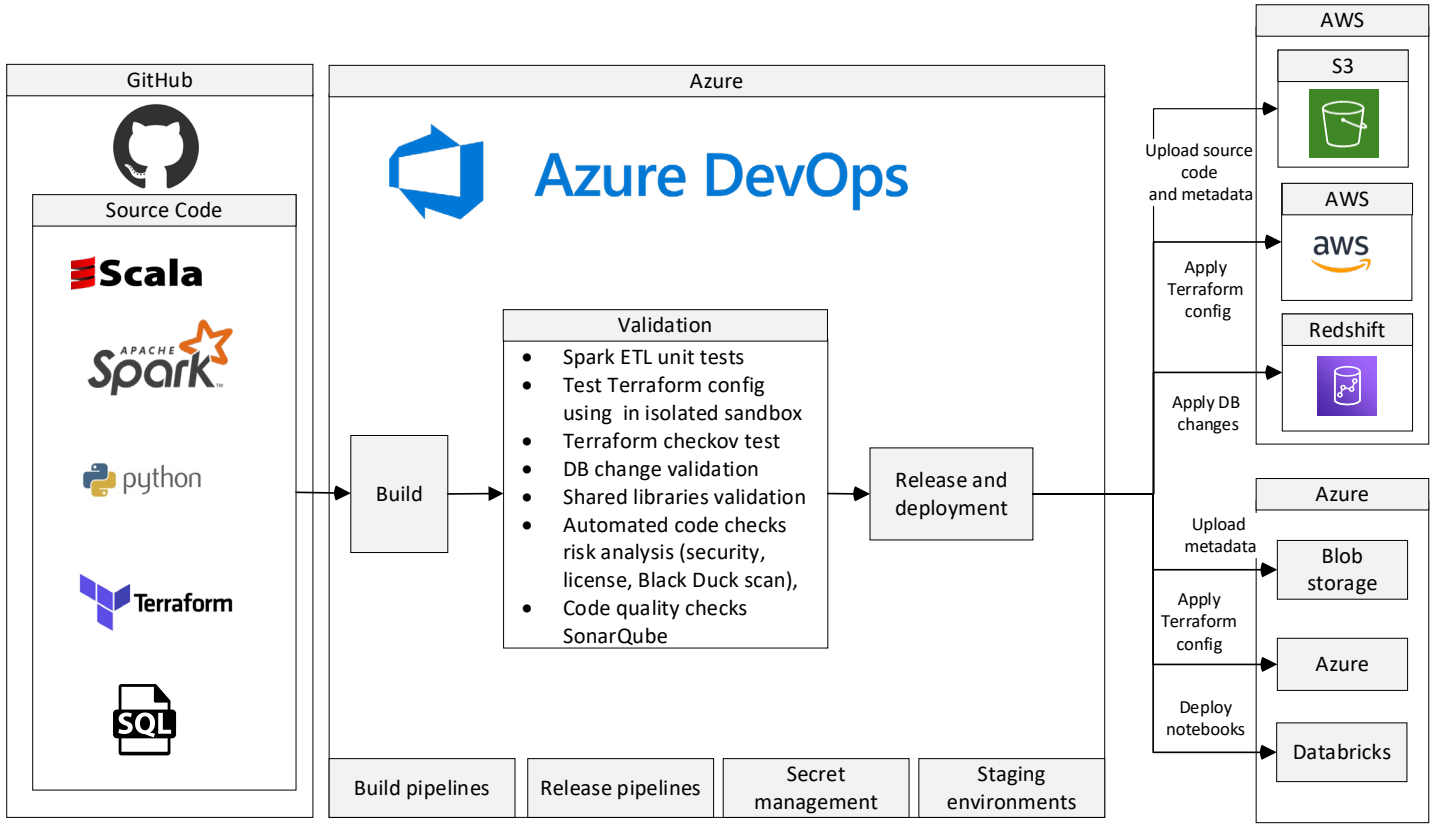
Amazon Glue

ETL processing based on SPARK ETLs (AWS Glue serverless service) implemented in Scala

HashiCorp Terraform | aws

Infrastructure as a Code used for maintaining and provisioning all cloud infrastructure (Terraform)

Multi-Cloud CI/CD pipeline design for CAP



- CI/CD pipeline implemented in Azure Cloud DevOps solution
- Integrated with AWS, Azure and GitHub
- Building and validation of Scala based Glue ETL spark jobs
- Automated unit testing – unit tests are executed in Azure DevOps agents
- Pull request validation pipeline supports code reviews
- Infrastructure as a code (Terraform) – all infrastructure changes are validated and released using CI/CD pipeline
- Pipeline configured including staging environments dev, test and prod
- Dedicated Database change management pipeline designed for validation and releases RedShift/PostgreSQL

Case Study 2: Portable D&A Platform Spark & Kubernetes

Automotive

Case Study: Portable Data & Analytics Platform Automotive

First Use Case

- Creation of a platform for Analytics on car configuration data, driven by new regulatory requirements
- The data must be processed locally in a unified way in the covered regions (EU, US, China)
- The platform should accommodate further Big Data & Analytics programs in the future
- The platform should be able to be deployed on-prem as well as in the Cloud

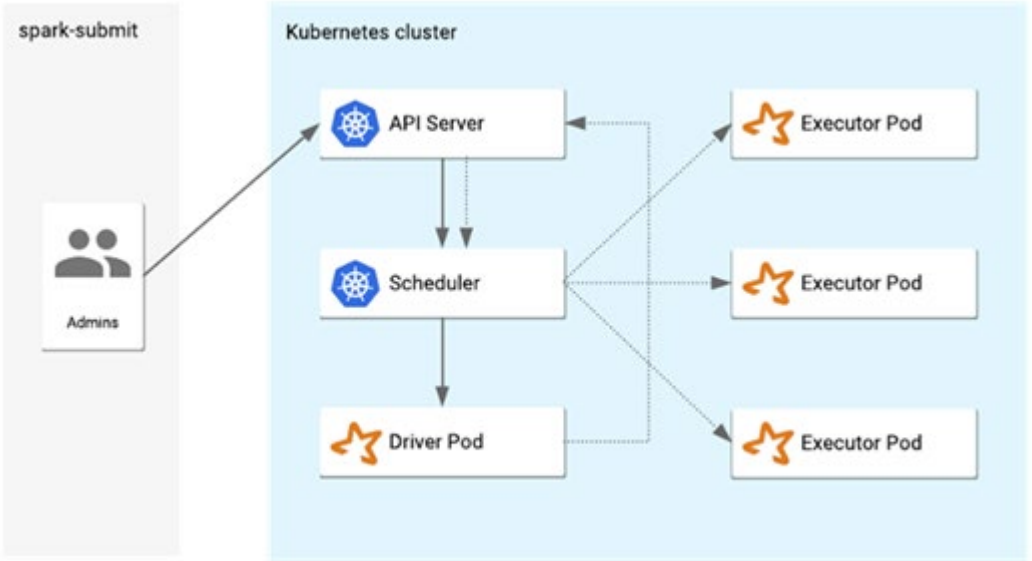
After ca. 3 months of assessing different possible solutions (incl. Cloud, Legacy DWH solutions) a Cloud-agnostic best-of-breed approach based on Kubernetes, Apache SPARK and Object Storage was selected.

Portable D&A Platform with Spark & Kubernetes

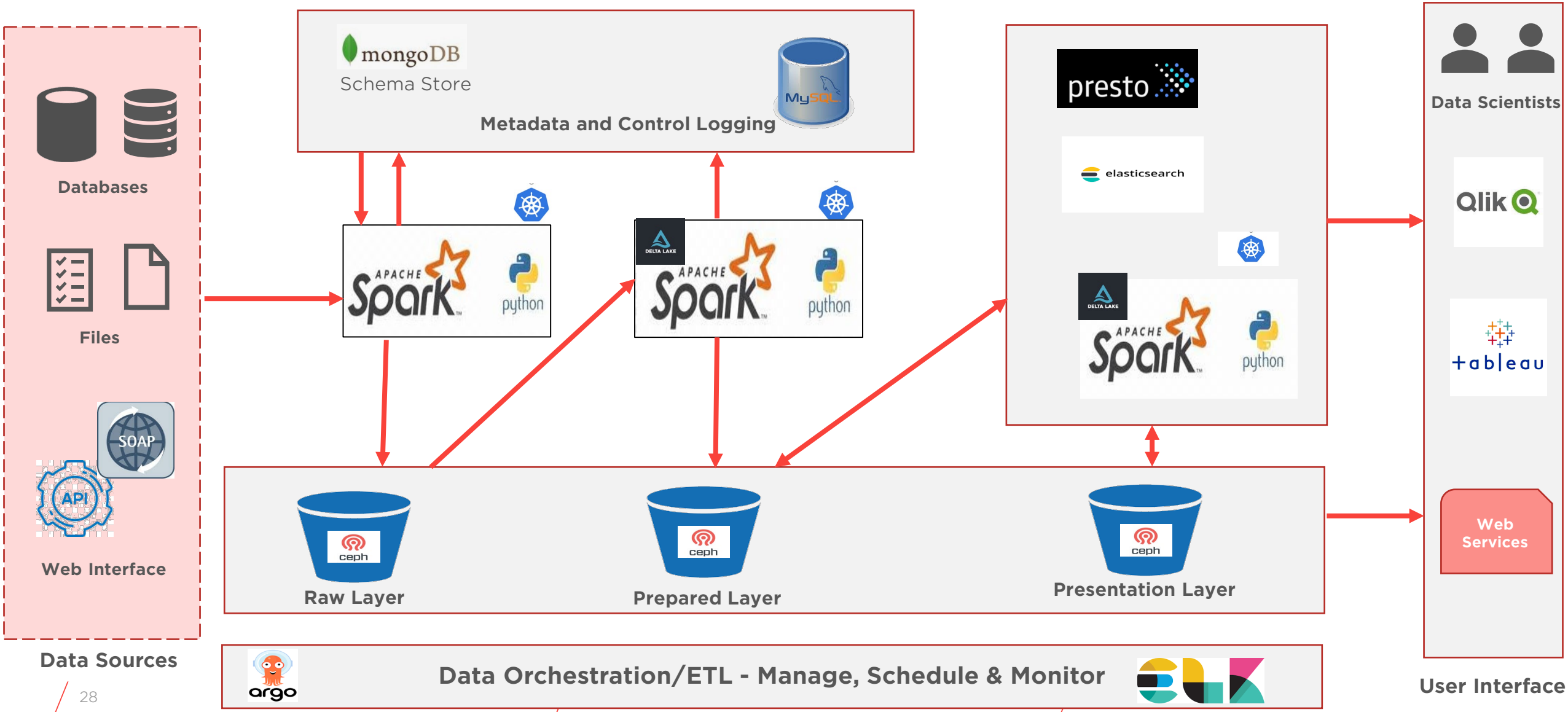
- **Kubernetes:** *„**Run K8s Anywhere** - Kubernetes is open source giving you the freedom to take advantage of on-premises, hybrid, or public cloud infrastructure, letting you effortlessly move workloads to where it matters to you.“* (https://kubernetes.io/de/)

- **Apache SPARK:** Already for the on-prem and now for the Cloud Data Lakes one of the most used platforms for parallel processing of Big Data

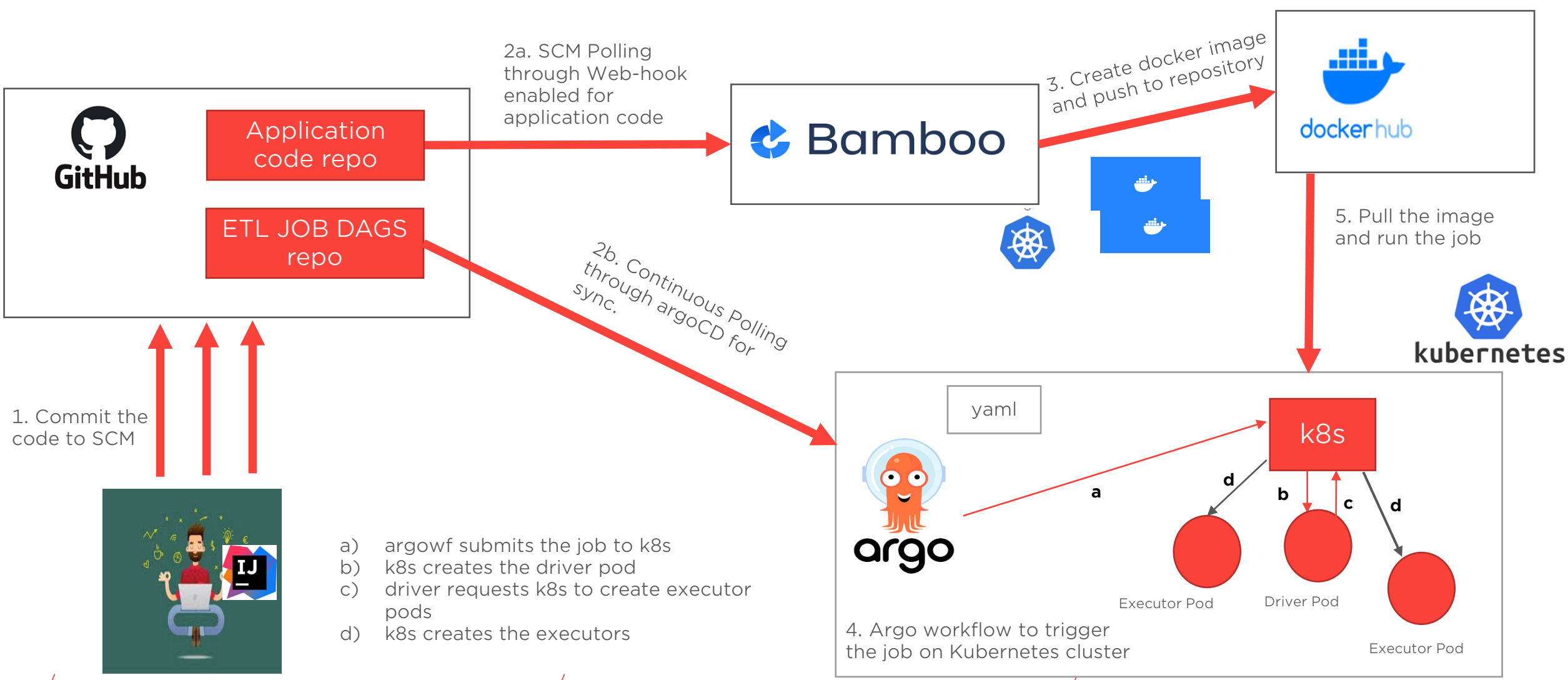
- Since 2018, **Kubernetes** is available as an alternative Apache Spark Cluster Manager (next to Mesos & Yarn)



Solution Overview



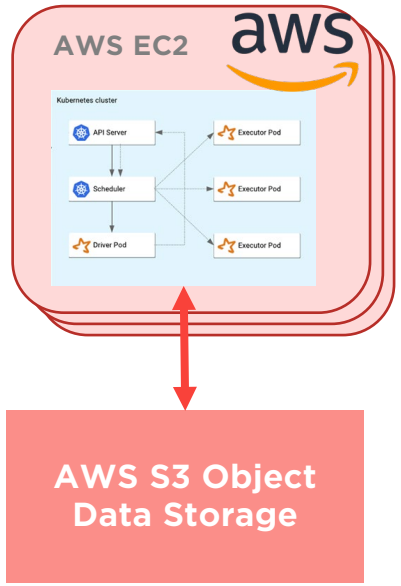
Agility with CI/CD Pipelines



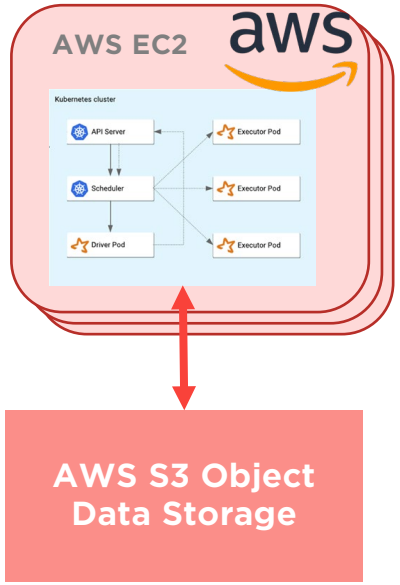
a) argowf submits the job to k8s
b) k8s creates the driver pod
c) driver requests k8s to create executor pods
d) k8s creates the executors

Portable from Day One

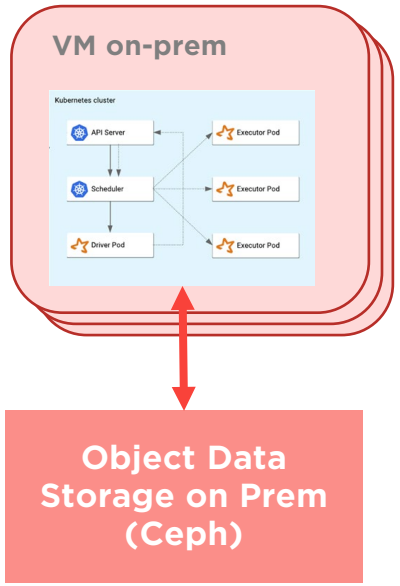
First Prototype developed in Adastra Germany Data Engineering Lab (AWS)



Following Sprints in AWS-Account of the Customer



Final Development on on-prem Infrastructure of the customer

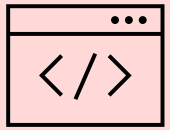


Adastra Modular Cloud D&A Framework Overview

Adastra Modular Cloud D&A Framework Overview

Infrastructure As a Code


Templates


Scripts


Policies

CI/CD Pipeline


Networking

Apps

Storage


Security

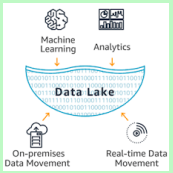

Cloud Infra

Design and Development


Reusable Modules


Ingestion


Historization and CDC



Data Lake best practices


Control and metadata capture


Generic Data Transformation


Coding Best practices

Danke

Adastra GmbH

Niedenau 36, 60325 Frankfurt am Main
.+49 (0)69 719 779 790 / infoDE@adastragrp.com

www.adastragrp.com