Audi‘s Hadoop Journey into the Hybrid Cloud

DataWorks Summit 2019 - Barcelona

Carsten Herbe (Audi Business Innovation GmbH, Germany)

About us

Audi AG

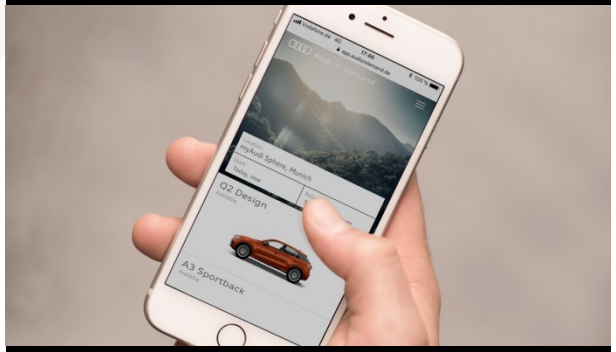
1,8 million cars per year*, 90.000 employees worldwide*

** source: https://www.audi.com/de/company.html*

Audi Business Innovation GmbH

Munich based subsidiary of Audi AG

Audi mobility innovations

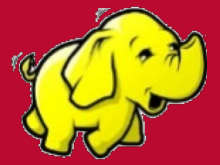


Audi on demand

Audi balanced technologies

Audi e-gas

Audi customer IT solutions



Carsten Herbe
Audi Business Innovation GmbH

- Data **Platform & Solution** Architecture
- Technical Product Owner & Architect for Cloud Hadoop
- 5 years **Hadoop**, 3 years **Kafka**, 1 year **AWS**
- 10+ years Data Warehousing & BI

HAAP – Hybrid Audi Analytic Platform

Big Data Capabilities & Focus data domains

Programs | **Projects** | **Data Scientists**

HAAP – HYBRID AUDI ANALYTIC PLATTFORM

Security
- Authentifi-cation
- Data Encryption
- Auditing

Embed Analytics
- Complex Event Processing
- Analytical APIs

Deliver Information
- Dash-boarding
- Planning & Simulation
- Visual Analytics

Analyze Data
- BI Report & OLAP
- Statistical Methods
- **Machine Learning**
- Analytical Script

Store, Distribute and Process Data
- Data Warehouse
- Analytical Databases
- **File Systems (HDFS)**

Provision Data
- ETL Framework
- **Stream Processing**
- **Batch Processing**
- Data Access / APIs

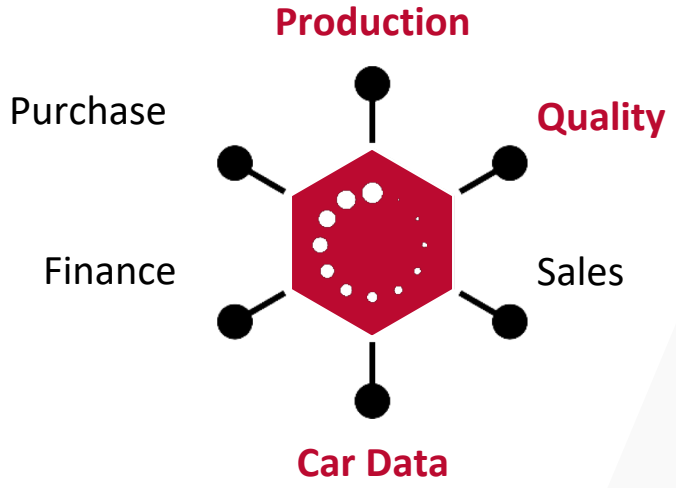
Deliver Service
- On-Prem Platform
- **Cloud Platform**
- Application Deployment

Infrastructure & Services
- Hardware, Network, OS
- Monitoring

Manage Information
- Master Data Mgmt
- Data Lineage

Design & Maintain Solutions
- Lifecycle Mgmt
- Development Process & Methods

! **Data Domains**



Why cloud?

Audi's motivation to extend its Hadoop platform to the cloud

Data "Locality"

- **Audi is moving many applications to the cloud**
- **Data of one important use case is already in the cloud**

Scalability

- **Scaling clusters:** number of nodes, node types, …
- **Scaling stages:** testing new features, upgrades, …

Functionality

- **Adding nodes with GPUs**
- **Use a more flexible staging process**
- **Cloud services:** S3, RDS, Docker Registry, …
  - **Reducing work on infrastructure**

Goals

One platform as a hybrid solution

Hybrid

- **Some related system are currently only on-premise:**
  - **DWH, Reporting Tool, …**
- **Some data sources remain on-premise (e.g. manufacturing)**

One platform

- **Write once, run everywhere:** identical tech stack
- **Single sign-on:** on-prem principals used for cloud
- **Data**: easy data movement & shared metadata

Project Setup

Team setup & project mode

Mixed Team

- **Companies:** internal **(**Audi + ABI) + external (2 partner + HWX)
- **Bases:** 4 cities in 2 countries
- **Nationalities:** 5 different nationalities

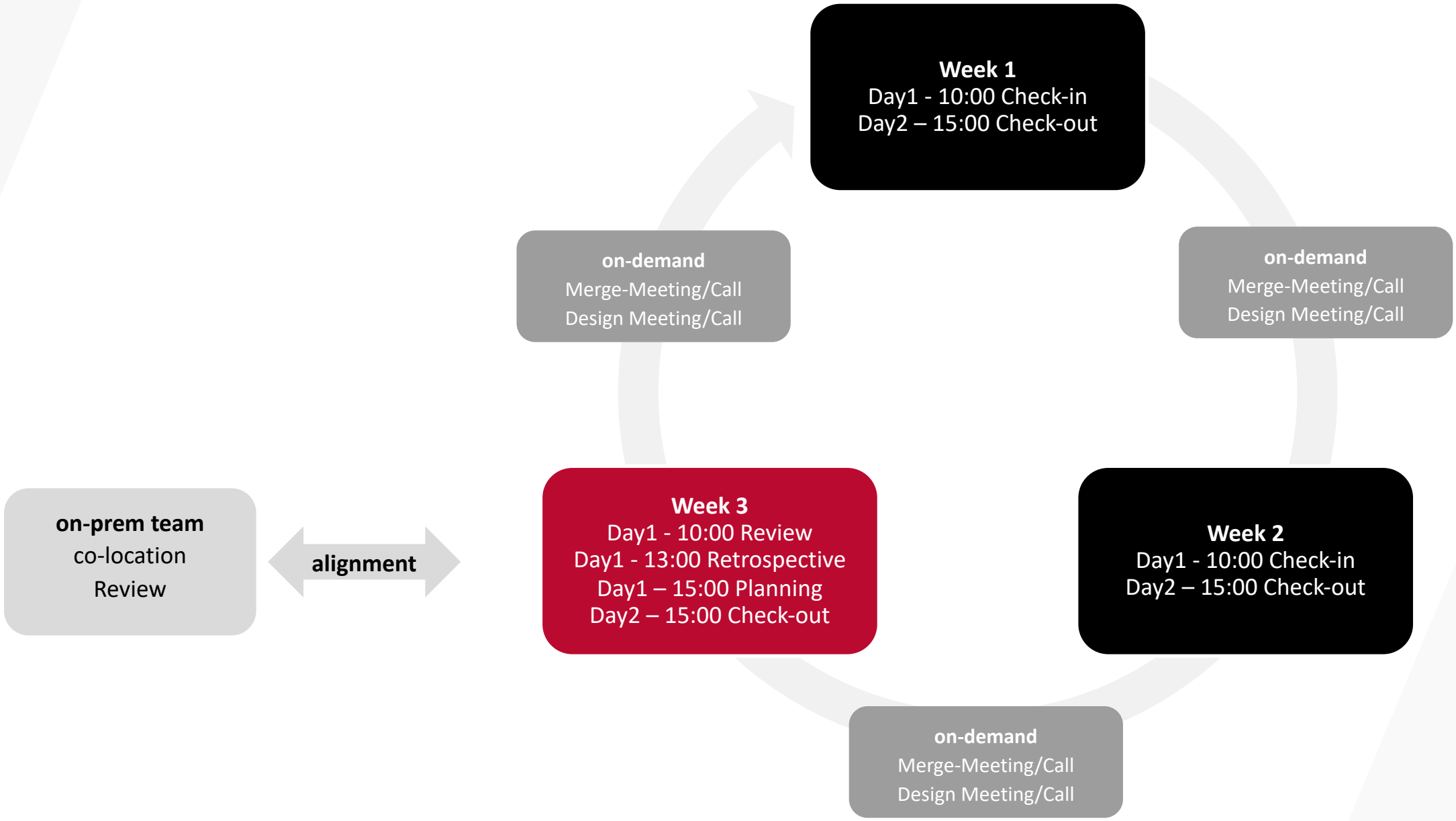
Collaboration

- **Scrum based**
- **Weekly 2 days on-site workshop at the Audi project office**
- **Tools: Jira, Bitbucket, RocketChat**

Goals

- **get experts on various topics (devops, Hadoop, AWS) together**
- **Knowledge transfer from external to internal**

Sprint Structure and on-site workshops



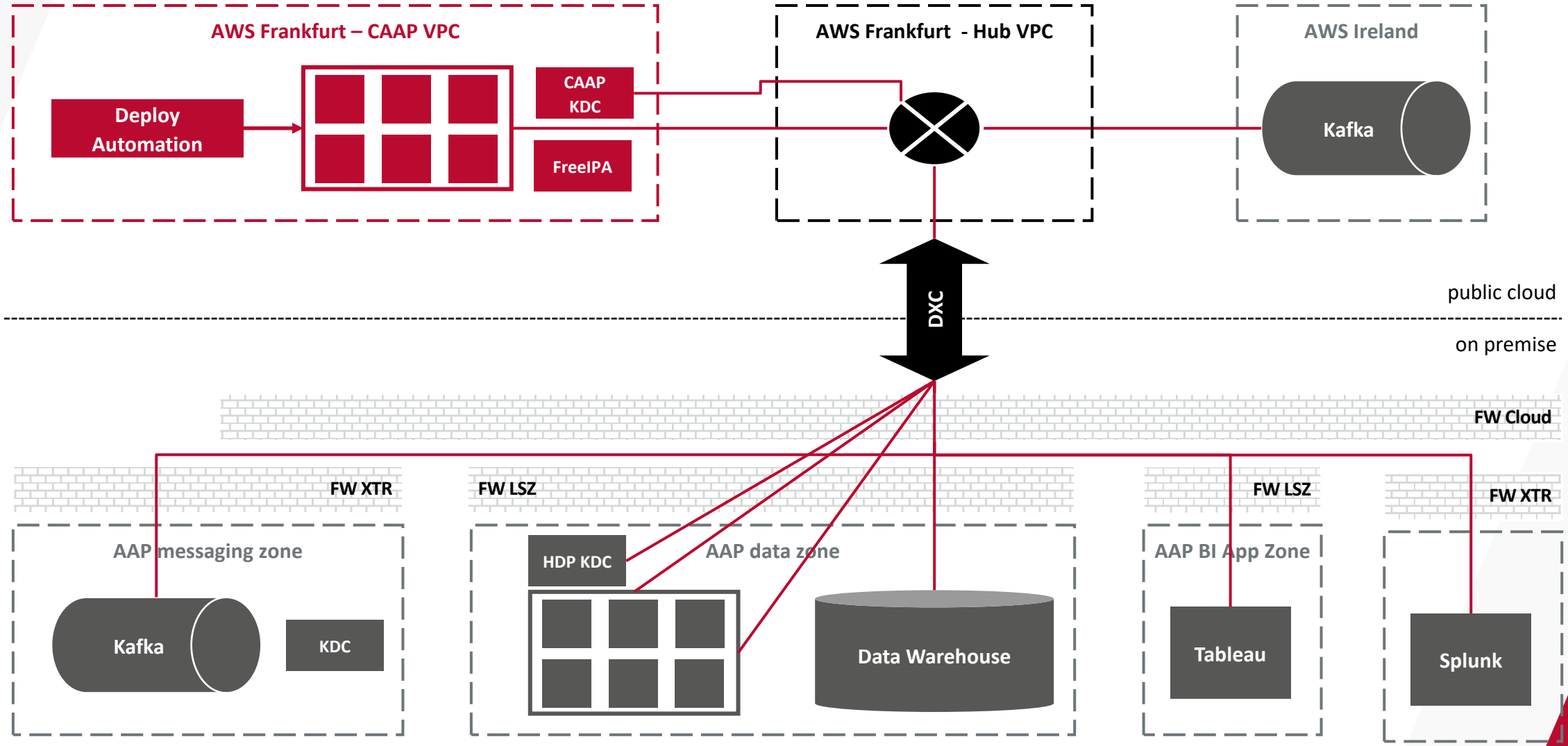
Choice of Technologies

Finding the best fitting tech stack for Audi

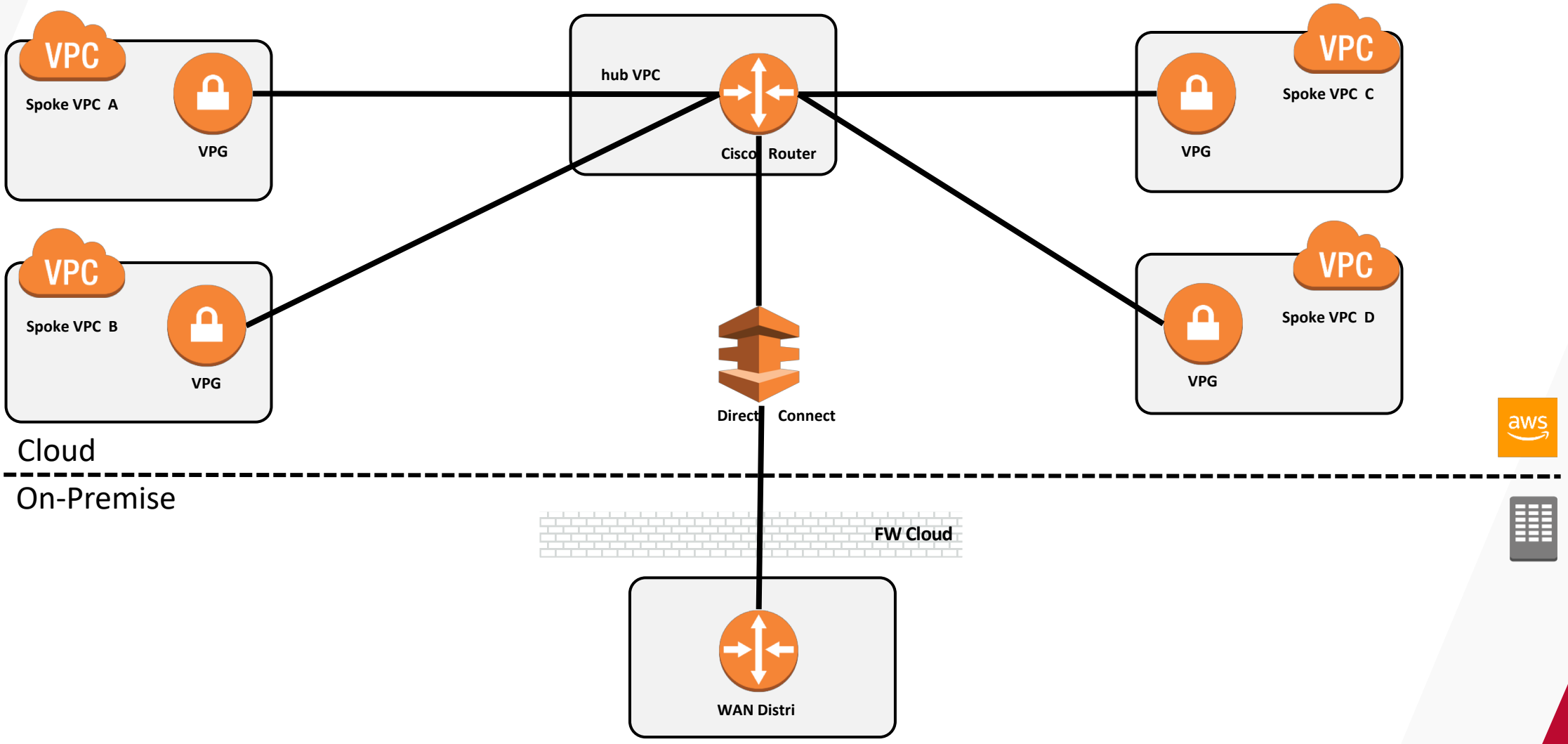
AWS Infrastructure setup	• CloudFormation • Terraform	**Terraform** • already used by other projects
Configuration Management	• Terraform + Bash • Ansible • …	**Ansible** • switched from Bash as complexity increased • already used by other projects
Hadoop Deployment	• Ambari Blueprints • Cloudbreak	**Ambari Blueprints** • Cloudbreak is difficult to integrate into existing environment • No versioning with Cloudbreak yet
User management	• Local users manually • Integrate with corporate AD/LDAP • Our own FreeIPA	**FreeIPA** • AD integration was not possible (yet) • Highest flexibility (+AD later) • DNS, Certificate Authority

Hybrid Architecture

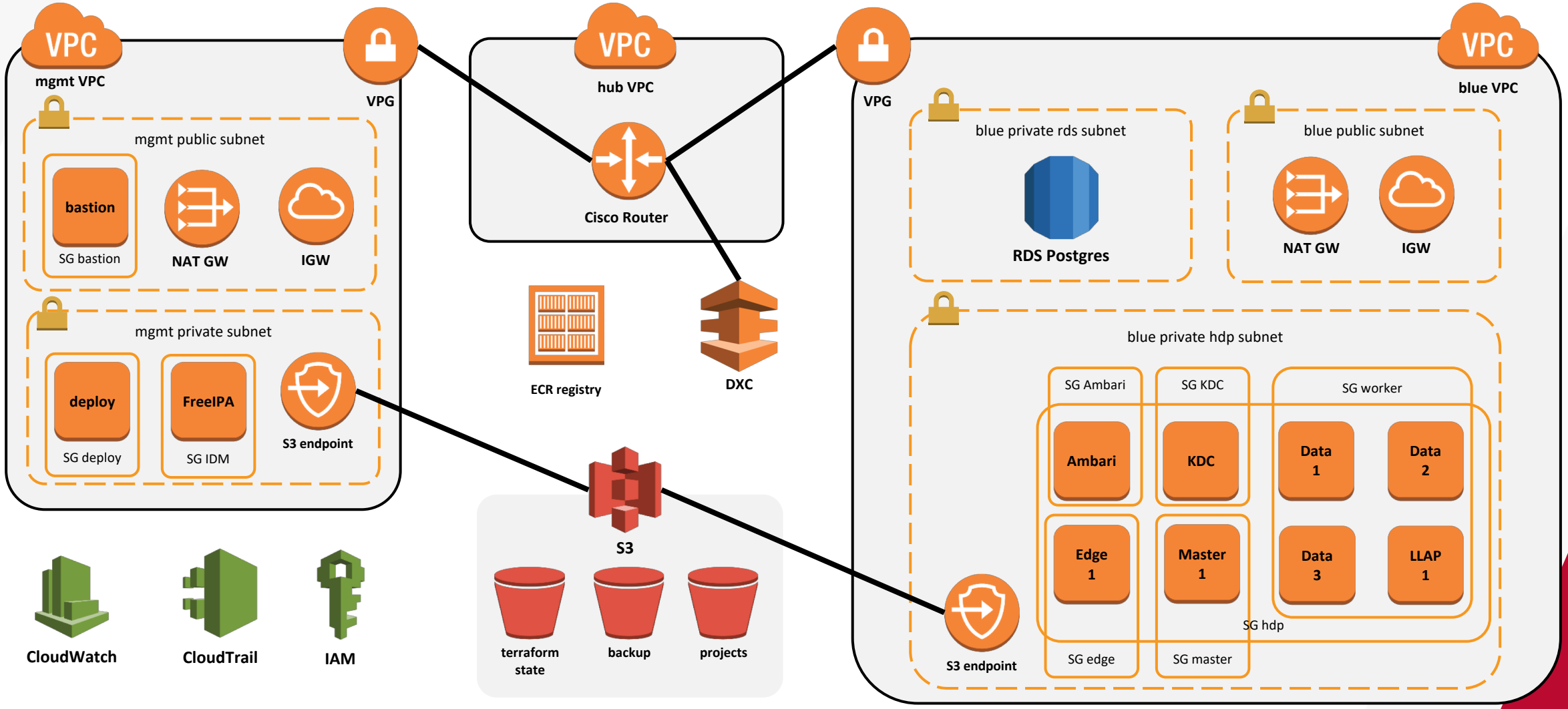
HAAP Architecture – Big Picture



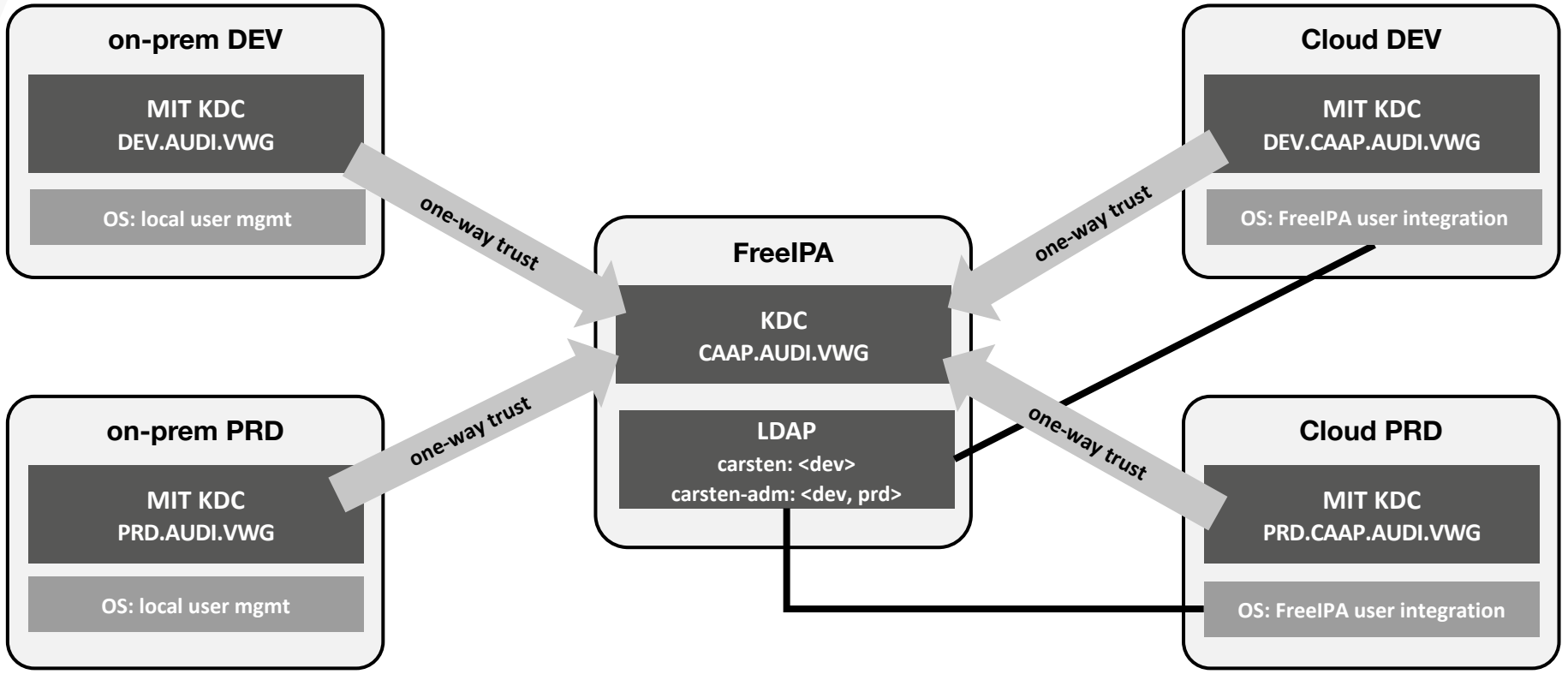
High-level AWS network architecture



Cloud Hadoop Platform: detailed view



User Management & Kerberos Trust




```
> kinit carsten@DEV.AUDI.VWG
> hdfs dfs -ls //ONPREMDEV:8020/user/carsten
> hdfs dfs -ls //CLOUDDEV:8020/user/carsten
```

```
> kinit carsten@CAAP.AUDI.VWG
> hdfs dfs -ls //CLOUDDEV:8020/user/carsten
> hdfs dfs -ls //CLOUDPRD:8020/user/carsten
> hdfs dfs -ls //ONPREMDEV:8020/user/carsten
```

Lessons learned

With great freedom come great responsibilities …

Cloud

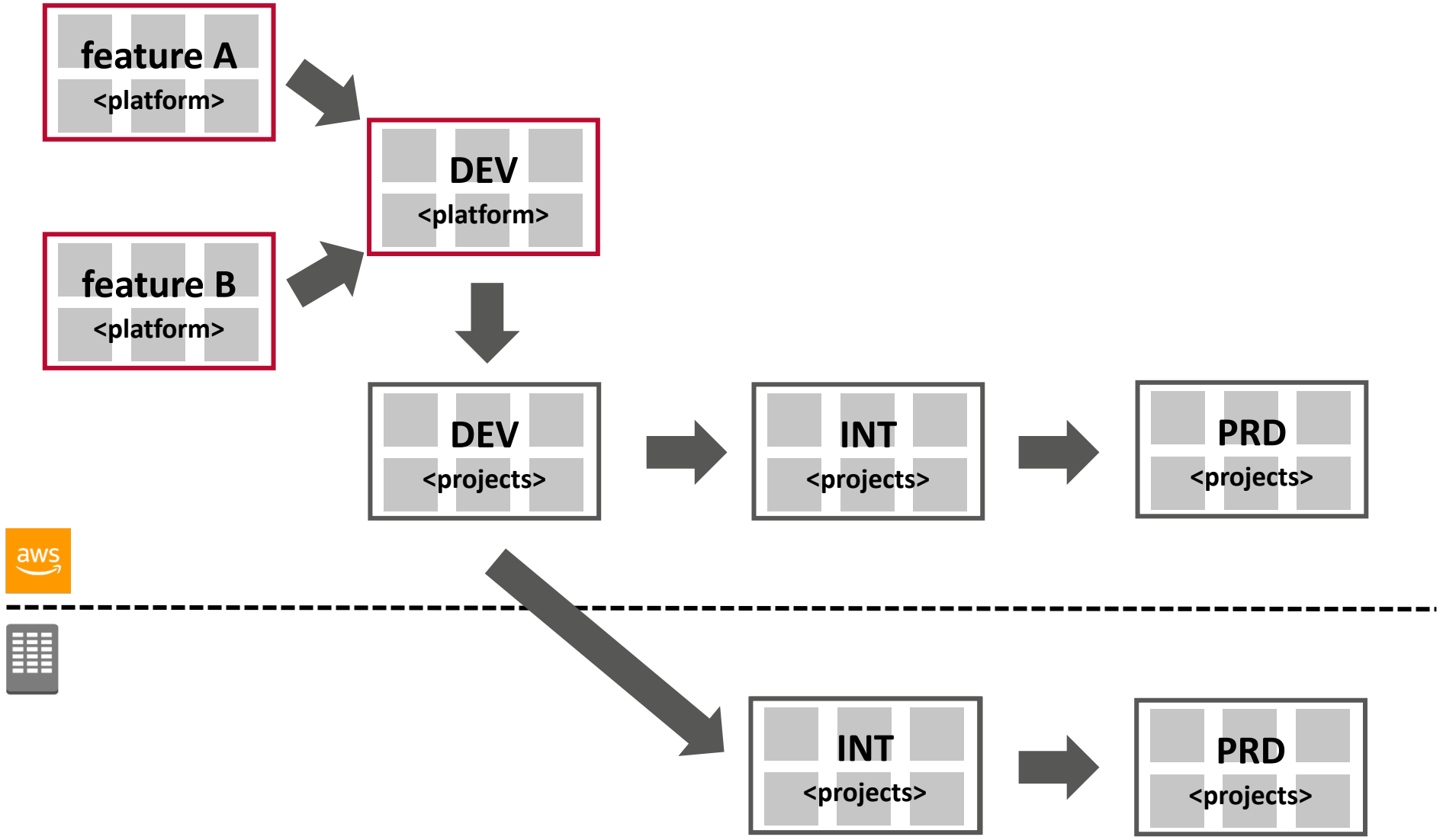
- **you can do anything you want right away!**
  - **but you have to do it yourself:** e.g. DNS, LDAP, …
- **Automation pays off but requires initial invest**
- **Security must be considered from the start**

Project setup

- **Agile**
  - **Strong involvement of product owner required**
  - **Distributed teams costs lot of travelling time**
- **Different experts required:** Cloud (AWS), Networking, DevOps, Hadoop, …
- **Fluctuation**: distribute knowledge

Looking into the Future

Staging process for projects and platform



Technologies on the road map

Cloud

- on demand nodes with GPU for machine learning
- S3/Glacier for „cold“ data
- Looking into Kafka as a Service (Confluent, AWS)

Machine Learing

- on demand nodes with GPU for machine learning
- Data Science Workbench

HDP3.x

- Using Docker under Yarn for more flexibility/functionality
- Hive3 Kafka Integration

Data Plane

- Data Steward Service for hybrid Data Governance
- Data Lifecycle Manager for data transfers and backup

WE ARE HIRING

https://www.audi.com/corporate/de/karriere/einstieg-bei-audi.html

https://karriere.audibusinessinnovation.com/