Building Audi's enterprise big data platform

DataWorks Summit 2018 - Berlin

Matthias Graunitz (AUDI AG, Germany)
Carsten Herbe (Audi Business Innovation GmbH, Germany)
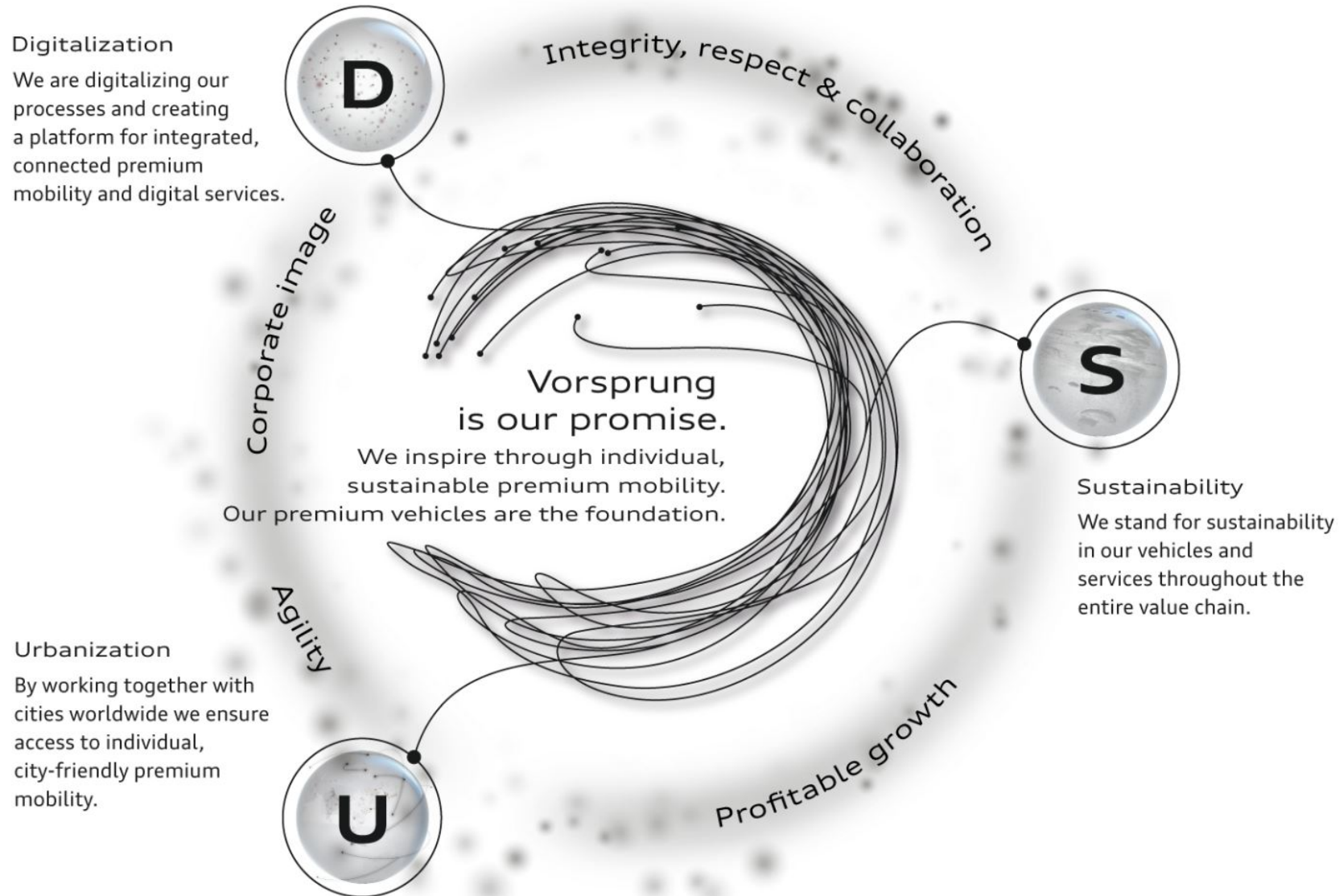
WHO ARE WE?

Audi Group
Audi, Lamborghini, Ducati and Italdesign

Vorsprung is our promise
Strategy 2025

Digitalization

We are digitalizing our processes and creating a platform for integrated, connected premium mobility and digital services.

Integrity, respect & collaboration

Corporate image

Vorsprung is our promise.

We inspire through individual, sustainable premium mobility. Our premium vehicles are the foundation.

Sustainability

We stand for sustainability in our vehicles and services throughout the entire value chain.

Agility

Urbanization

By working together with cities worldwide we ensure access to individual, city-friendly premium mobility.

Profitable growth

Audi Business Innovation GmbH

...is the <u>development, establishment, sales and operation</u> of innovative concepts, products and services, as well the <u>holding of shares</u> in the field of future mobility.

Audi mobility innovations

Audi on demand

Audi balanced technologies

Audi e-gas

Audi customer IT solutions

About us

Matthias Graunitz
AUDI AG

» Center of Competence **Big Data & BI**
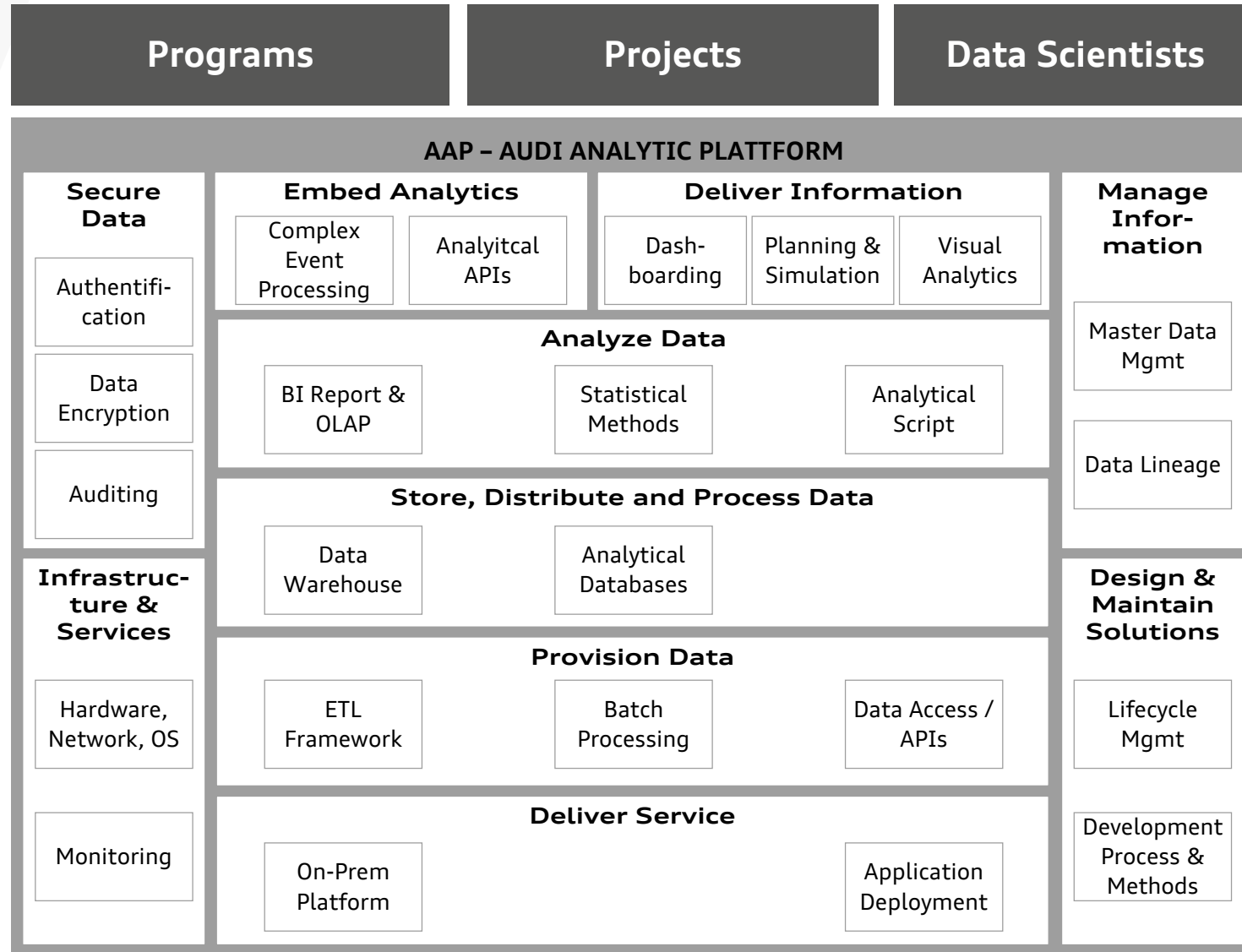» **Big Data** Architect
» **10+ years** Data Warehousing & BI

Carsten Herbe
Audi Business Innovation GmbH

» Data **Platform & Solution** Architecture
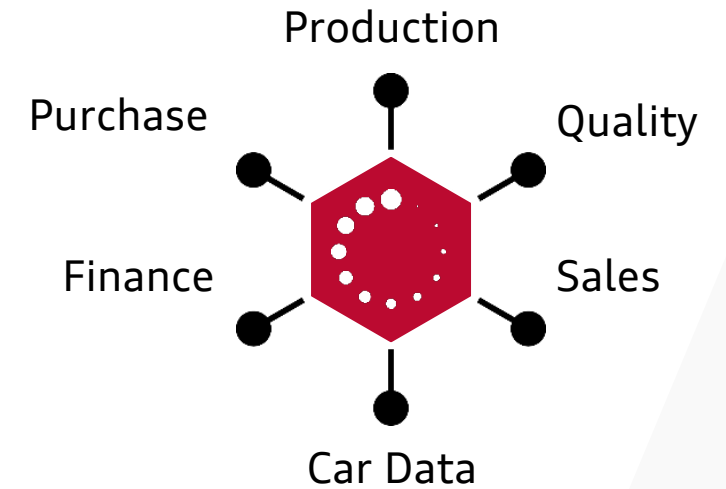» **Hadoop** since **2013**
» **10+ years** Data Warehousing & BI
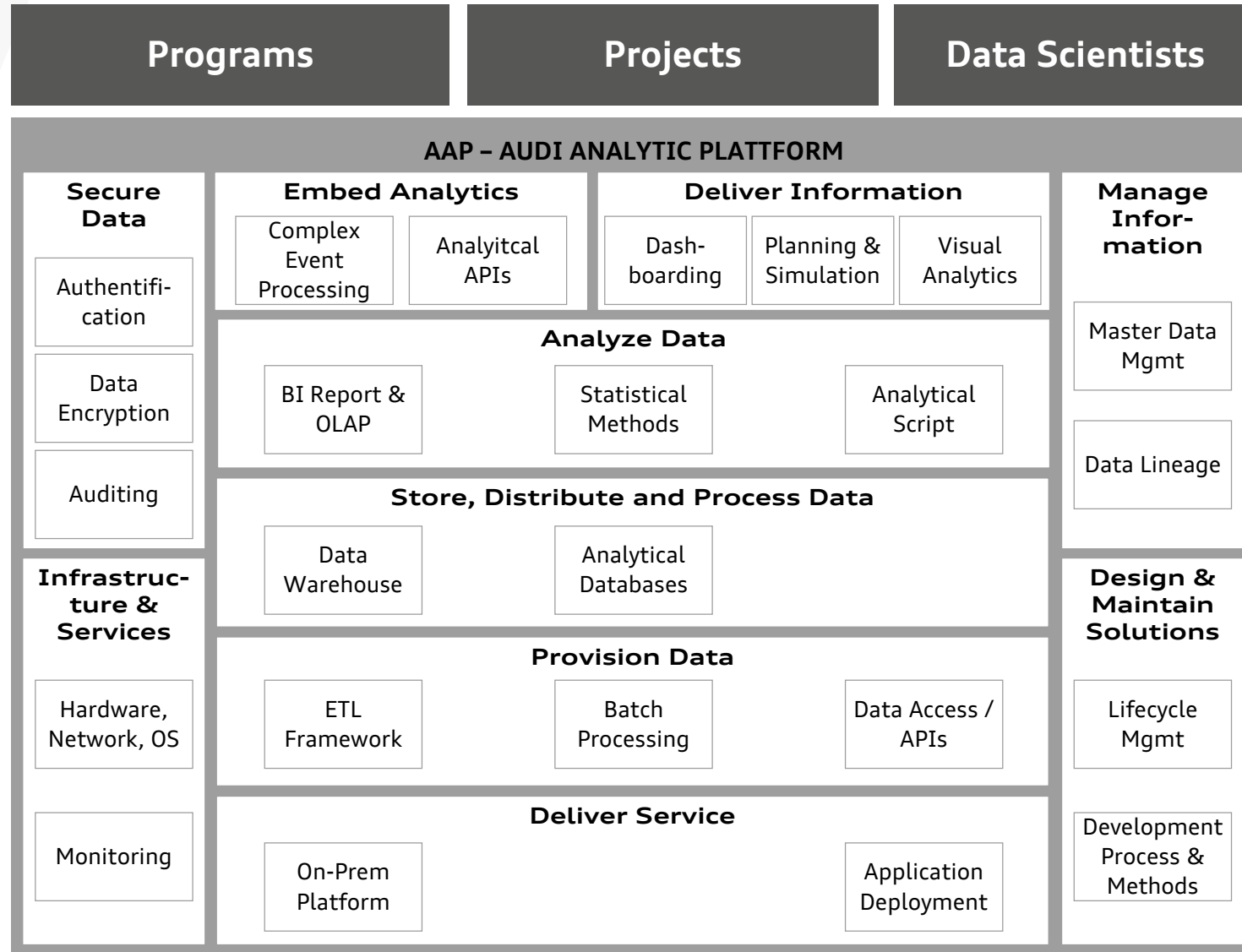
2 YEARS AGO...
STARTING BIG DATA AT AUDI
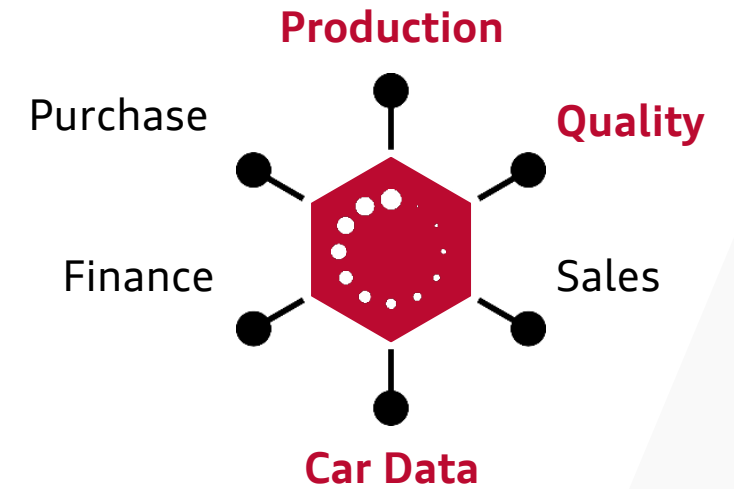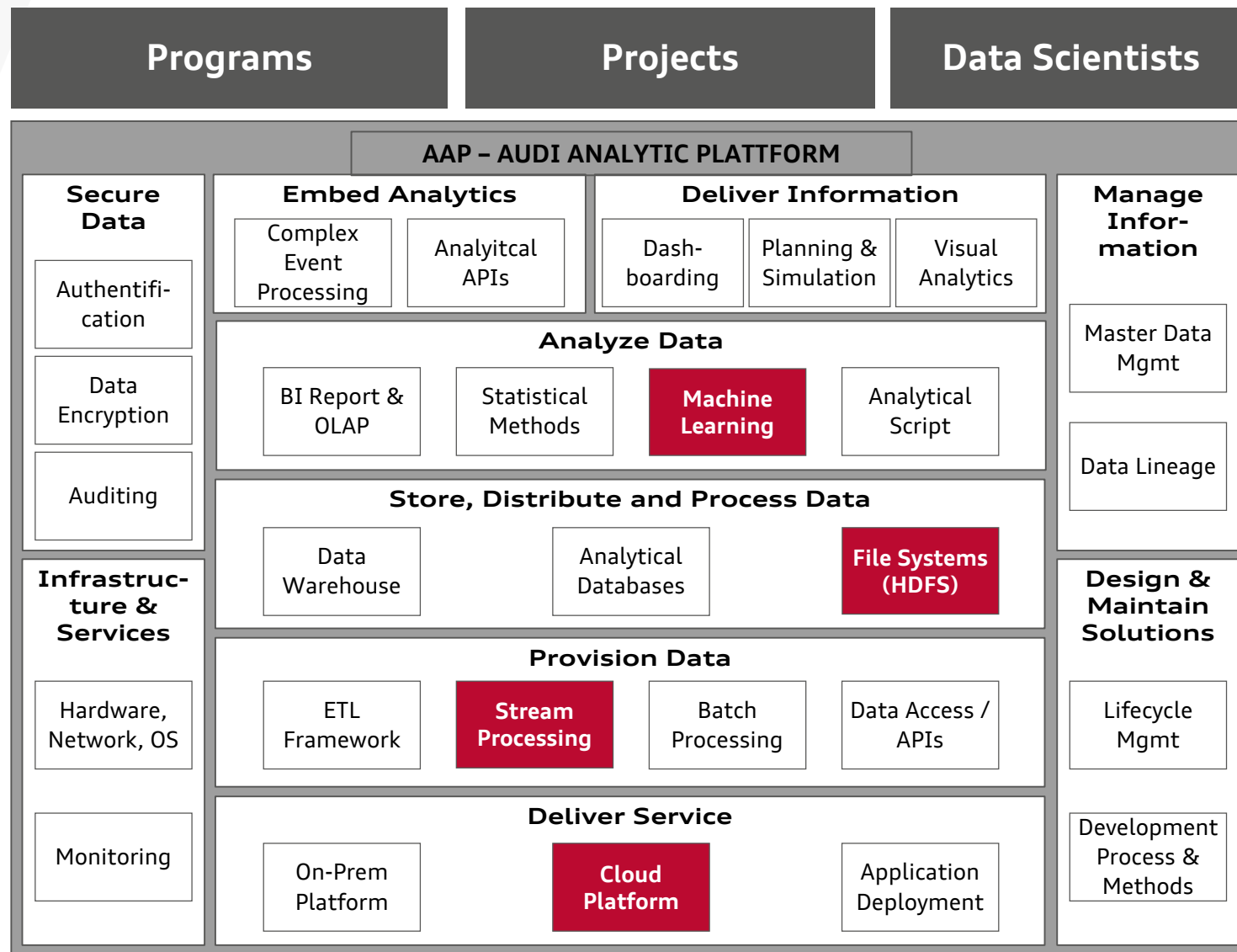
Analytical Capabilities by 2015

Programs	Projects	Data Scientists

AAP – AUDI ANALYTIC PLATTFORM

Secure Data

- Authentifi-cation
- Data Encryption
- Auditing

Infrastruc-ture & Services

- Hardware, Network, OS
- Monitoring

Embed Analytics

Complex Event Processing	Analyitcal APIs

Deliver Information

Dash-boarding	Planning & Simulation	Visual Analytics

Analyze Data

BI Report & OLAP	Statistical Methods	Analytical Script

Store, Distribute and Process Data

Data Warehouse	Analytical Databases

Provision Data

ETL Framework	Batch Processing	Data Access / APIs

Deliver Service

On-Prem Platform	Application Deployment

Manage Infor-mation

- Master Data Mgmt
- Data Lineage

Design & Maintain Solutions

- Lifecycle Mgmt
- Development Process & Methods

! Data Domains

Production
Quality
Sales
Car Data
Finance
Purchase

Analytical Capabilities by 2015

Programs	Projects	Data Scientists

! Data Domains

AAP – AUDI ANALYTIC PLATTFORM

Secure Data
- Authentifi-cation
- Data Encryption
- Auditing

Infrastruc-ture & Services
- Hardware, Network, OS
- Monitoring

Embed Analytics
- Complex Event Processing
- Analyitcal APIs

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

Analyze Data
- BI Report & OLAP
- Statistical Methods
- Analytical Script

Store, Distribute and Process Data
- Data Warehouse
- Analytical Databases

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

Deliver Service
- On-Prem Platform
- Application Deployment

Manage Infor-mation
- Master Data Mgmt
- Data Lineage

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

Production

Quality

Purchase

Finance

Sales

Car Data

Analytical Capabilities by 2015

Programs	Projects	Data Scientists

AAP – AUDI ANALYTIC PLATTFORM

Secure Data

- Authentification
- Data Encryption
- Auditing

Infrastructure & Services

- Hardware, Network, OS
- Monitoring

Embed Analytics

- Complex Event Processing
- Analyitcal 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

Manage Information

- Master Data Mgmt
- Data Lineage

Design & Maintain Solutions

- Lifecycle Mgmt
- Development Process & Methods

! Data Domains

- Production
- Quality
- Purchase
- Finance
- Sales
- Car Data

Our first Hadoop Cluster 2015

DEV

Hadoop	per node	Sum
# data nodes	1	4
RAM	128 GB	0,5 TB
Cores	24	96
HDD*	40 TB	160 TB

* Raw Capacity without replication and FS overhead!

Our first attempt to walk
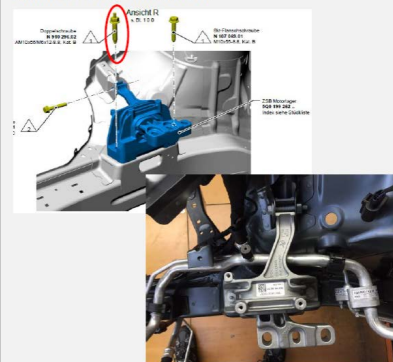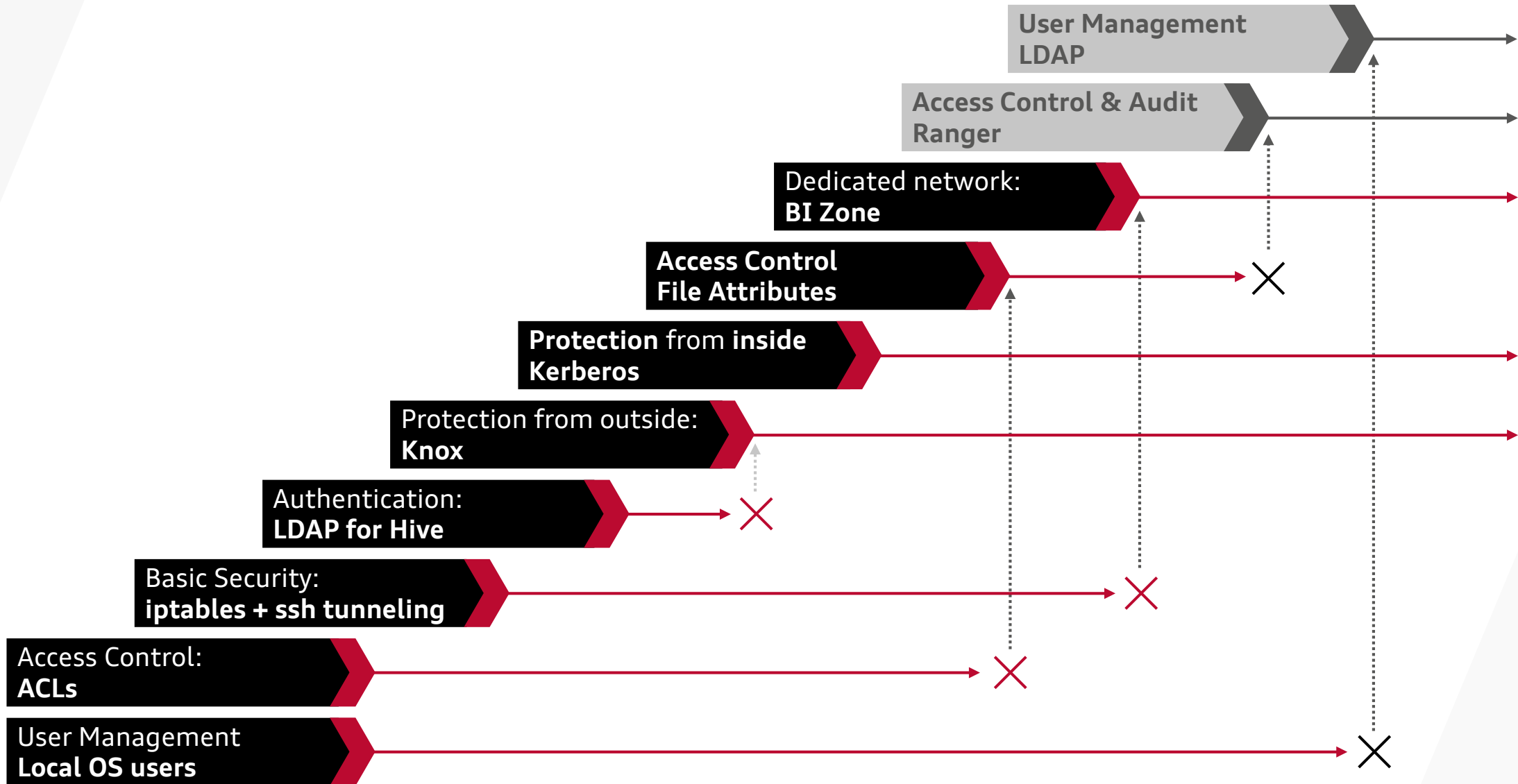## with Big Data Technologies

SCREWDRIVER ANALYSIS
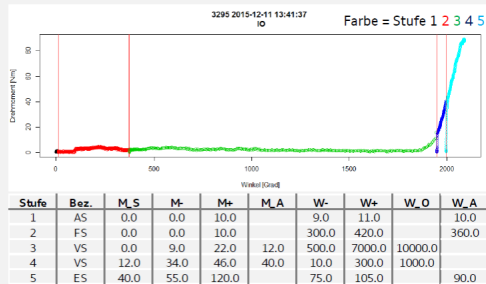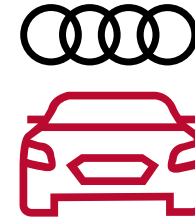
COMPANY CAR ANALYSIS

ENTERPRISE INTEGRATION VS
SPEED OF DELIVERY

Securing the Cluster as multi-tenant environment
Step by step by step towards our target architecture ...

User Management
LDAP

Access Control & Audit
Ranger

Dedicated network:
BI Zone

Access Control
File Attributes

Protection from **inside**
Kerberos

Protection from outside:
Knox

Authentication:
LDAP for Hive

Basic Security:
iptables + ssh tunneling

Access Control:
ACLs

User Management
Local OS users

Password Hell

DATA NODE 1 - X

NAME NODE 1 - 2

EDGE NODE 1 - 2

OS Level
[Local User]
OS Named User
Technical Hive User
Technical Project User
Hadoop User

Audi Active Directory:
[AD User]
Named User
Technical Hive User

Knox	Hive WebHDFS SparkUI	HDFS/YARN	kinit	SSH 2 EdgeNode

Hadoop KDC:
[Kerberos Principal]
Name User
Technical Hive User
Technical Project User
Hadoop User

Legend: ———— password required   ———— no password required   ----▶ next step

DATA
INGESTION

Data ingestion:
technical requirements from projects, security and ops

INGESTION
- » Streaming data
- » Batch data
- » easy writing to HDFS/DWH

DECOUPLING
- » Data Sources should not directly be coupled to analytical backend jobs
- » This allows adding new analytical jobs without changing the source

HA & BUFFERING
- » Data ingestion must be available 24x7
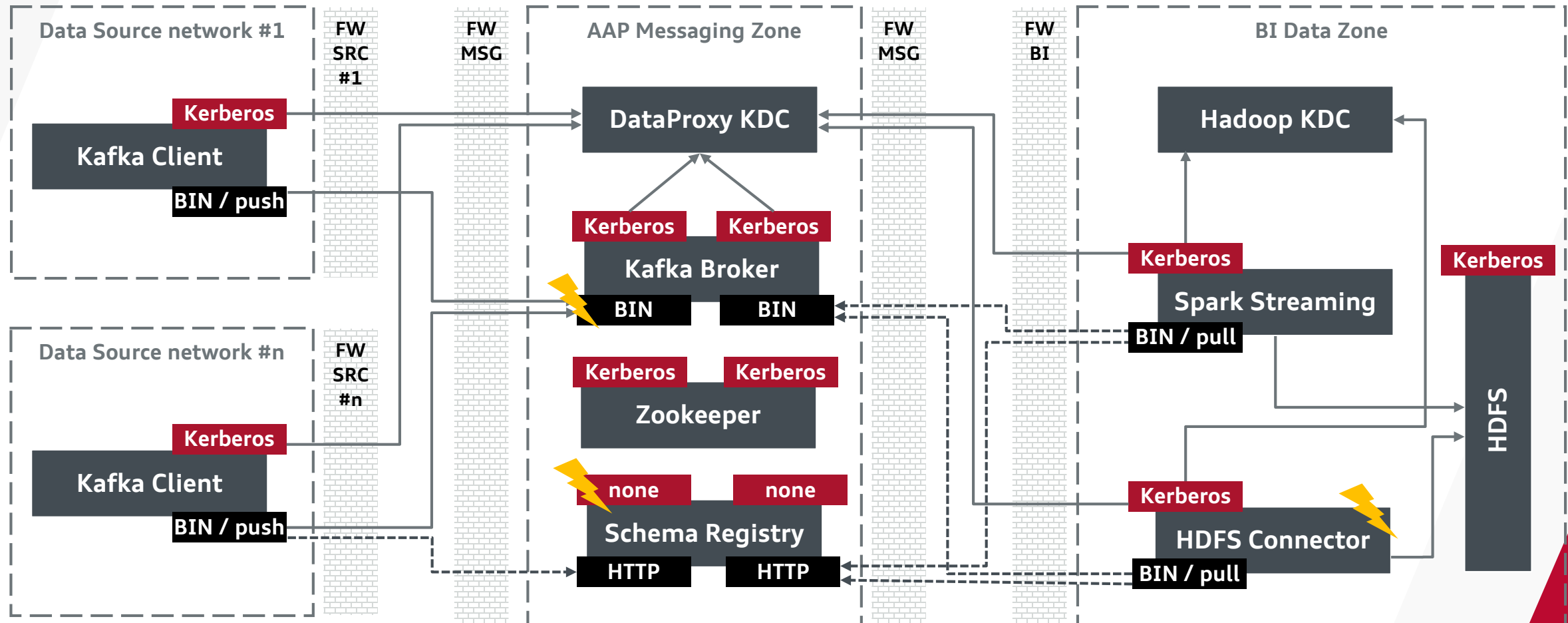- » Data must be buffered (persisted) in case backend or backend job is not available

SECURITY
- » Source systems must not connect directly to the data zone (Hadoop, DWH) – by IT Sec
- » Authentication + Data in motion encryption (multi tenancy)
- » Protocol must be auditable
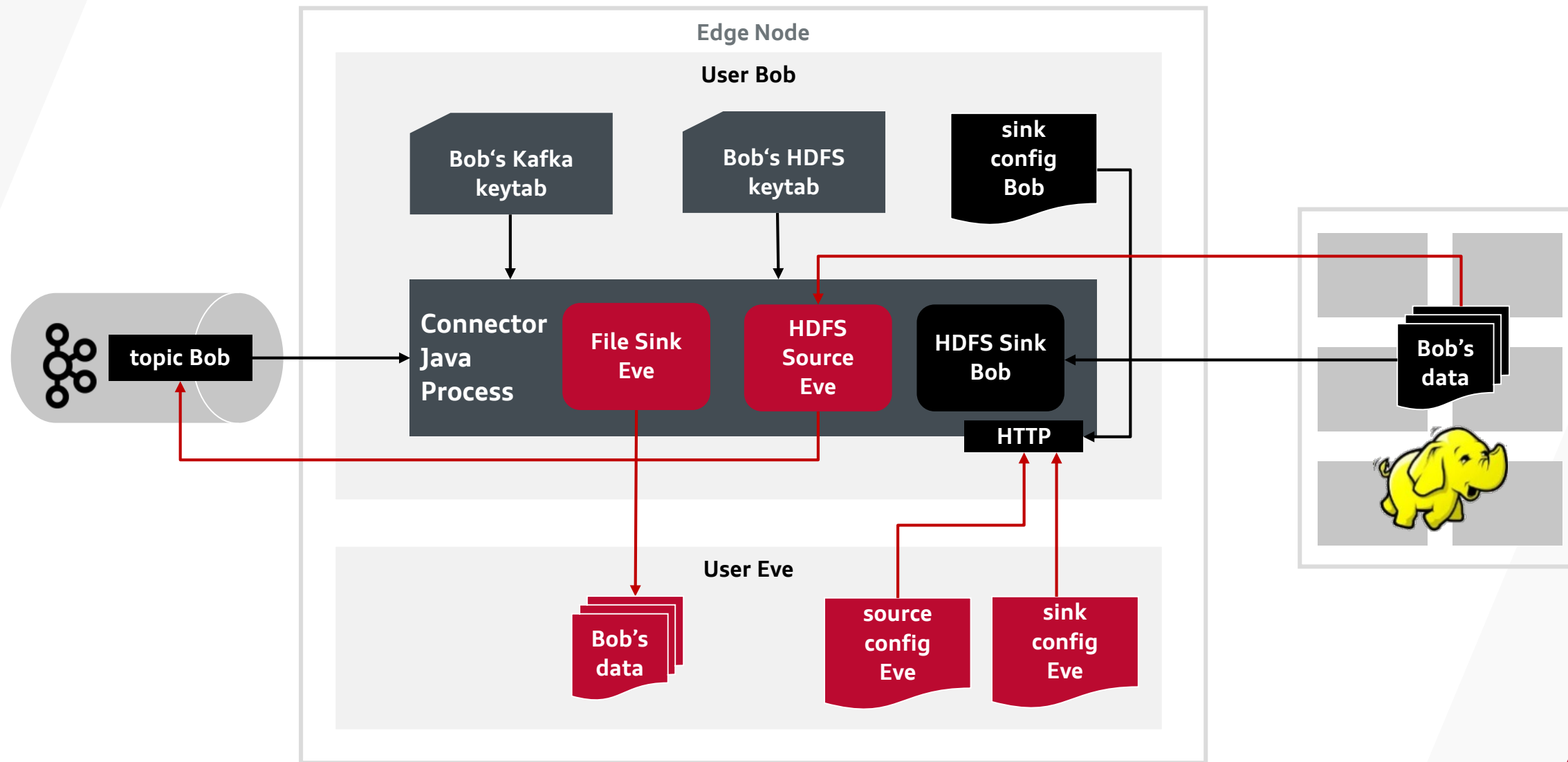- » Some data sources run in the cloud

SCALABILITY
- » Amount of data will increase over time for most projects
- » Number of projects will increase

Solution: Kerberized Confluent Kafka Platform

Data Source network #1 | **FW SRC #1** | **FW MSG** | **AAP Messaging Zone** | **FW MSG** | **FW BI** | **BI Data Zone**

Kafka Client — Kerberos — BIN / push

DataProxy KDC

Kerberos | Kerberos
Kafka Broker
BIN | BIN

Kerberos | Kerberos
Zookeeper

none | none
Schema Registry
HTTP | HTTP

Data Source network #n | **FW SRC #n**

Kafka Client — Kerberos — BIN / push

Hadoop KDC

Kerberos
Spark Streaming
BIN / pull

Kerberos
HDFS Connector
BIN / pull

Kerberos
HDFS

Legend: — encrypted (SSL)   ----- not encrypted   **authentication**   **protocol / direction**   firewall   pain point

Kafka Distributed Connector: unsecured REST API
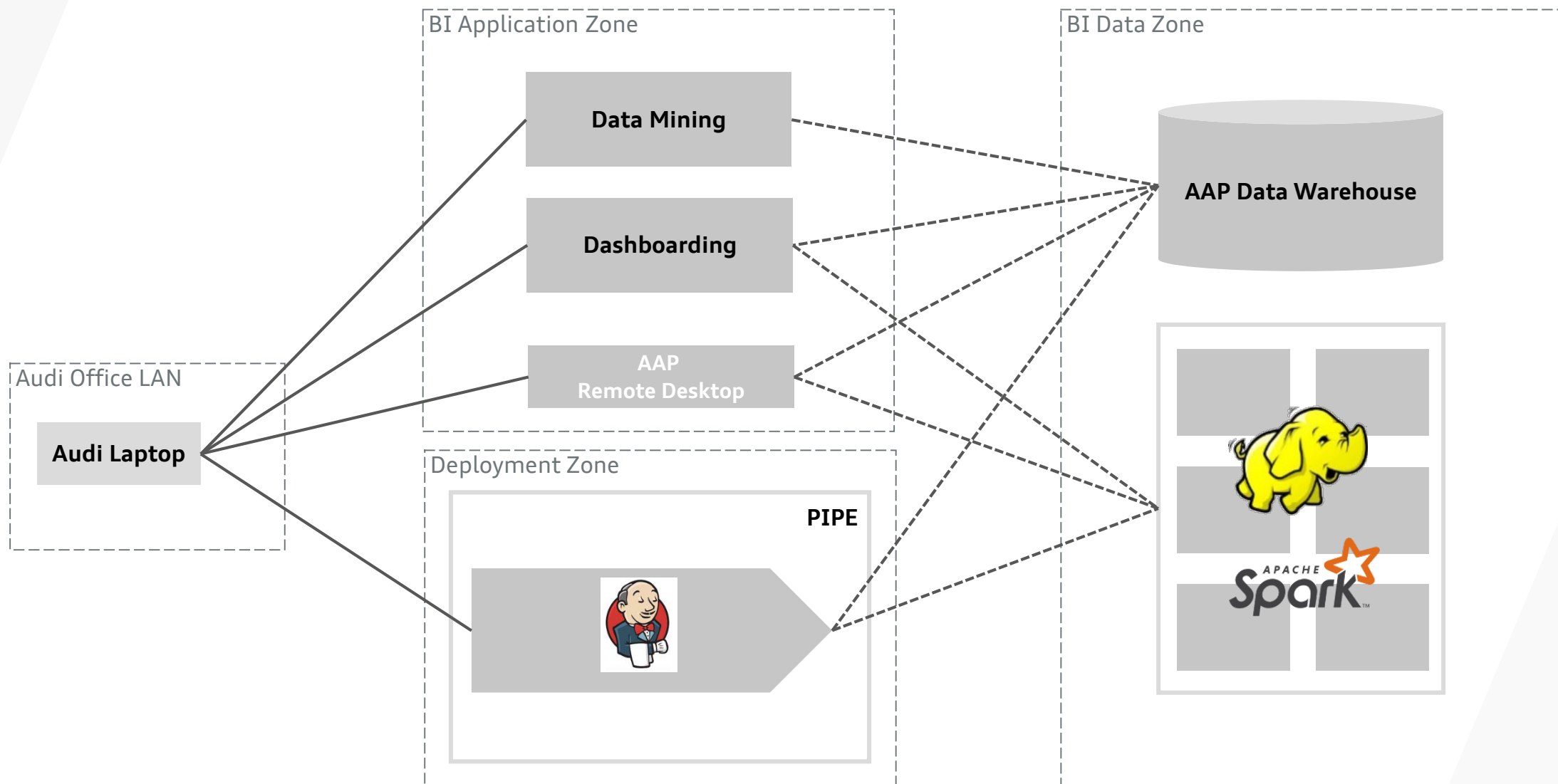
TODAY
CURRENT STATE

Architecture & Network Zones – Data Ingestion
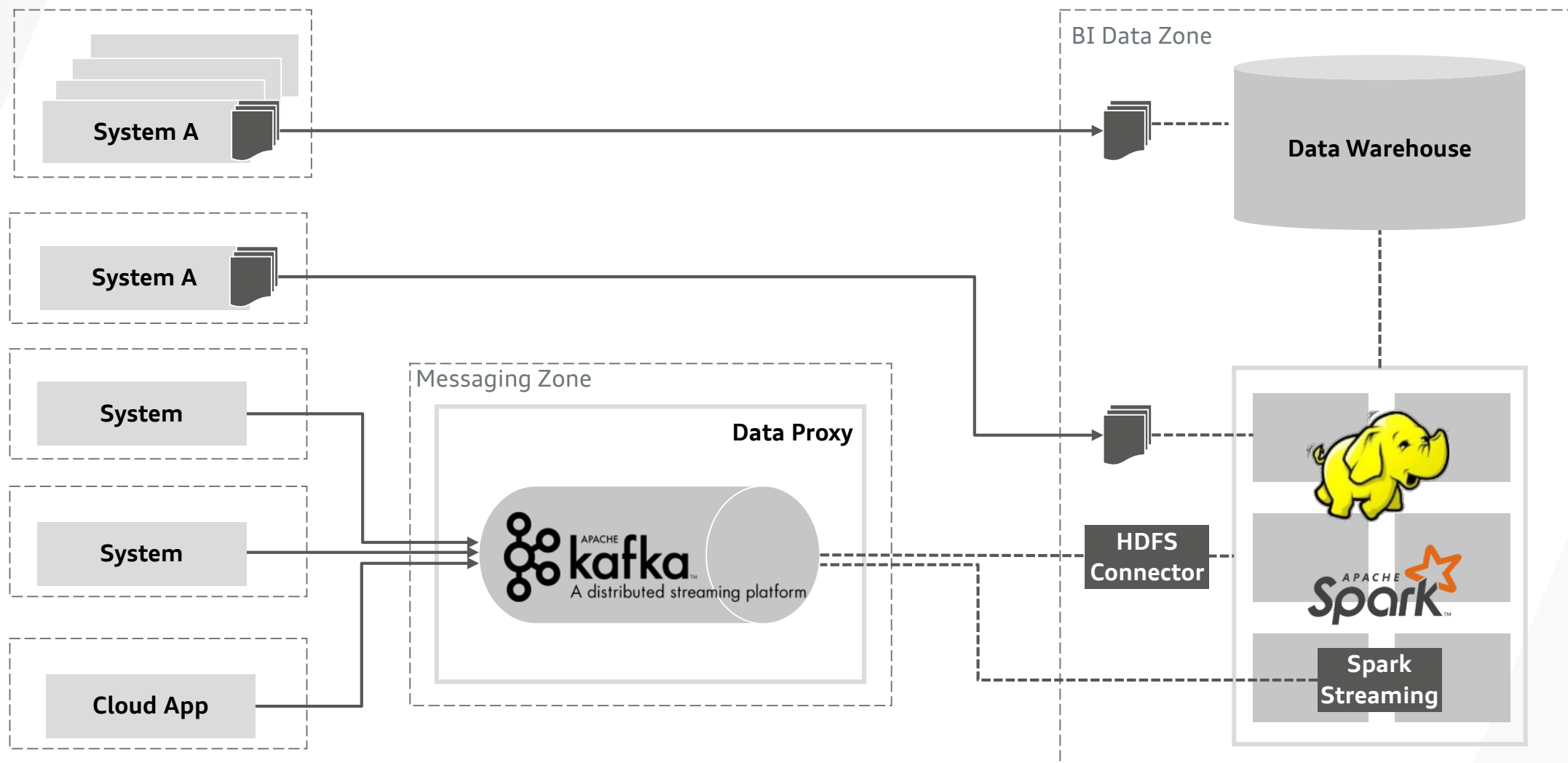
Architecture & Network Zones – User & Developer Access

Hadoop Cluster Sizing Production 2017

Hadoop		per node	Sum
PROD	# data nodes	1	12
	RAM	512 GB	6 TB
	Cores	24	288
	HDD*	96 TB	9.216 TB

Kafka		per node	Sum
PROD	# broker nodes	1	4
	RAM	32 GB	128 GB
	Cores	6	24
	HDD*	4 TB	16 TB

* Raw Capacity without replication and FS overhead!
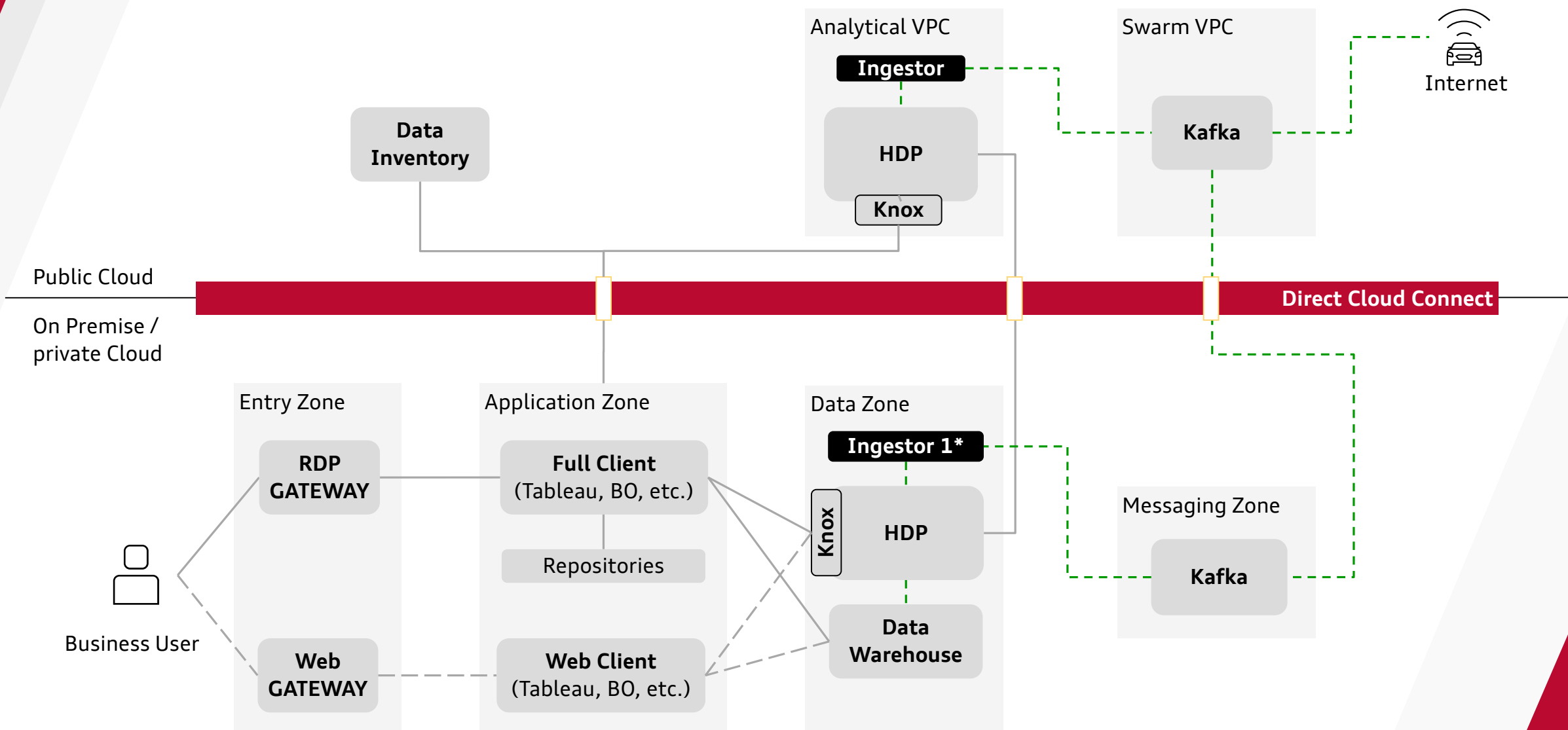
Organisational Tasks

Current state

Organisational Tasks

- ➤ Data Ownership & Data Governance
  (Data Domain Modell with clear responsibility in each domain)

- ➤ Lifecycle Management for each Shared Service in strong collaboration with the projects and programs

- ➤ Defined SLAs for each Shared Service based on general availability, data loss, confidentiality and verifiability

- ➤ Different Development Lifecycle between car and backend systems

- ➤ Use of Open Source Software and Support requirements from IT continuity

- ➤ Balance between multi tenant environment and flexibility

- ➤ Very long lifecycle of cars > 10 years with various built in software versions

TOMORROW
WHAT'S UP NEXT

Hybrid Approach for the AAP

WE ARE HIRING

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

https://karriere.audibusinessinnovation.com/