Audi's journey to an enterprise big data platform

Strata Data 2018 - London

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.

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

Integrity, respect & collaboration

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

Profitable growth

Corporate Image

Vorsprung is our promise.
We inspire through individual, sustainable premium mobility. Our premium vehicles are the foundation.
Audi Business Innovation GmbH

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

### Data Domains
- Production
- Quality
- Finance
- Sales
- Car Data

### Program Categories
- Secure Data
  - Authentication
  - Data Encryption
  - Auditing
- Infrastructure & Services
  - Hardware, Network, OS
  - Monitoring
- Embed Analytics
  - Complex Event Processing
  - Analytical APIs
- Analyze Data
  - BI Report & OLAP
  - Statistical Methods
  - Analytical Script
- Data Domains
  - Finance
  - Purchase
  - Production
  - Quality
  - Sales
- Car Data
- Projects
- Deliver Information
  - Dashboarding
  - Planning & Simulation
  - Visual Analytics
- Provision Data
  - ETL Framework
  - Batch Processing
  - Data Access / APIs
- Design & Maintain Solutions
  - Lifecycle Mgmt
  - Development Process & Methods
- Data Scientists
- Manage Information
  - Master Data Mgmt
  - Data Lineage
- Store, Distribute and Process Data
  - Data Warehouse
  - Analytical Databases
- Deliver Service
  - On-Prem Platform
  - Application Deployment
- Application Deployment

### AAP – AUDI ANALYTIC PLATFORM

<table>
<thead>
<tr>
<th>Programs</th>
<th>Projects</th>
<th>Data Scientists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secure Data</td>
<td>Embed Analytics</td>
<td>Manage Information</td>
</tr>
<tr>
<td>Authentication</td>
<td>Complex Event Processing</td>
<td>Master Data Mgmt</td>
</tr>
<tr>
<td>Data Encryption</td>
<td>Analytical APIs</td>
<td>Data Lineage</td>
</tr>
<tr>
<td>Auditing</td>
<td>Dashboarding</td>
<td>Analytical Script</td>
</tr>
</tbody>
</table>
| Infrastructure & Services | Planning & Simulation | Analytical 
| Hardware, Network, OS | Visual Analytics | Script |
| Monitoring | ETL Framework | Design & Maintain Solutions |
| | Batch Processing | Lifecycle Mgmt |
| | Data Access / APIs | Development Process & Methods |

### Projects
- Secure Data
- Embed Analytics
- Data Domains
- Analyze Data
- Store, Distribute and Process Data
- Provision Data
- Design & Maintain Solutions
- Deliver Service
## Analytical Capabilities by 2015

### Program Categories

#### Secure Data
- Authentication
- Data Encryption
- Auditing

#### Embed Analytics
- Complex Event Processing
- Analytical APIs

#### Deliver Information
- Dashboarding
- Planning & Simulation
- Visual Analytics

### Project Categories

#### 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

### Data Domains

- Finance
- Sales
- Quality
- Production
- Car Data

---

**AAP – AUDI ANALYTIC PLATFORM**

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

- Manage Information
  - Master Data Mgmt
  - Data Lineage

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

- Hardwar, Network, OS
- Monitoring
## Analytical Capabilities by 2015

### Data Domains
- Production
- Quality
- Finance
- Sales
- Car Data

### Data Scientists
- Embed Analytics
  - Complex Event Processing
  - Analytical APIs
  - BI Report & OLAP
  - Statistical Methods
  - Machine Learning

- Deliver Information
  - Dashboarding
  - Planning & Simulation
  - Visual Analytics
  - Analytical Script

- Analyze Data
  - Audit
  - Data Encryption
  - Auditing

### Programs
- Secure Data
  - Authentication
  - Data Encryption
  - Auditing

- Infrastructure & Services
  - Hardware, Network, OS
  - Monitoring

- Projects
  - Analytical Databases
  - File Systems (HDFS)
  - Batch Processing
  - ETL Framework
  - Stream Processing
  - Data Warehouse

- Provision Data
  - On-Prem Platform
  - Cloud Platform
  - Application Deployment

- Manage Information
  - Master Data Mgmt
  - Data Lineage
  - Lifecycle Mgmt
  - Development Process & Methods
Our first Hadoop Cluster 2015

<table>
<thead>
<tr>
<th>Hadoop</th>
<th>per node</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td># data nodes</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>RAM</td>
<td>128 GB</td>
<td>0,5 TB</td>
</tr>
<tr>
<td>Cores</td>
<td>24</td>
<td>96</td>
</tr>
<tr>
<td>HDD*</td>
<td>40 TB</td>
<td>160 TB</td>
</tr>
</tbody>
</table>

* Raw Capacity without replication and FS overhead!
Our first attempt to walk with Big Data Technologies
ENTERPRISE INTEGRATION
VS
SPEED OF DELIVERY
Securing the Cluster as multi-tenant environment
Step by step by step towards our target architecture ...

Access Control:
ACLs
User Management
Local OS users

Basic Security:
iptables + ssh tunneling

Authentication:
LDAP for Hive

Protection from inside:
Kerberos
Protection from outside:
Knox

Dedicated network:
BI Zone

Access Control & Audit
Ranger

User Management
LDAP
Password Hell

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
Kafka Distributed Connector: unsecured REST API

Legend:
- red: evil connection
- black: good connection

Graphical representation of the Kafka Distributed Connector with unsecured REST API connections.
TODAY
CURRENT STATE
Architecture & Network Zones – Data Ingestion

Legend:
- -------- encrypted (SSL)
- -------- not encrypted
Architecture & Network Zones – User & Developer Access

Legend:
- ——not encrypted
- ———encrypted (SSL)
### Hadoop Cluster Sizing Production 2017

#### PROD

<table>
<thead>
<tr>
<th>Component</th>
<th>Per Node</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td># data nodes</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>RAM</td>
<td>512 GB</td>
<td>6 TB</td>
</tr>
<tr>
<td>Cores</td>
<td>24</td>
<td>288</td>
</tr>
<tr>
<td>HDD*</td>
<td>96 TB</td>
<td>9.216 TB</td>
</tr>
</tbody>
</table>

#### Kafka

<table>
<thead>
<tr>
<th>Component</th>
<th>Per Node</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td># broker nodes</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>RAM</td>
<td>32 GB</td>
<td>128 GB</td>
</tr>
<tr>
<td>Cores</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>HDD*</td>
<td>4 TB</td>
<td>16 TB</td>
</tr>
</tbody>
</table>

* 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

Public Cloud

On Premise / private Cloud

Entry Zone
- RDP GATEWAY
- Web GATEWAY
- Business User

Application Zone
- Full Client (Tableau, BO, etc.)
- Repositories
- Web Client (Tableau, BO, etc.)

Data Zone
- Ingestor 1*
- Knox
- Ingestor
- HDP
- Data Warehouse

Analytical VPC
- HDP
- Knox

Swarm VPC
- Kafka

Internet

Direct Cloud Connect

Messaging Zone
- Kafka
WE ARE HIRING


https://karriere.audibusinessinnovation.com/