DATA PRODUCTS – FROM POC INTO PRODUCTION

Andreas Buckenhofer, DOAG K&A, Nuremberg 2018
ANDREAS BUCKENHOFER, DAIMLER TSS GMBH

“Forming good abstractions and avoiding complexity is an essential part of a successful data architecture”

Data has always been my main focus during my long-time occupation in the area of data integration. I work for Daimler TSS as Database Professional and Data Architect with over 20 years of experience in Data Warehouse projects. I am working with Hadoop and NoSQL since 2013. I keep my knowledge up-to-date - and I learn new things, experiment, and program every day.

I share my knowledge in internal presentations or as a speaker at international conferences. I'm regularly giving a full lecture on Data Warehousing and a seminar on modern data architectures at Baden-Wuerttemberg Cooperative State University DHBW. I also gained international experience through a two-year project in Greater London and several business trips to Asia.

I’m responsible for In-Memory DB Computing at the independent German Oracle User Group (DOAG) and was honored by Oracle as ACE Associate. I hold current certifications such as "Certified Data Vault 2.0 Practitioner (CDVP2)", "Big Data Architect", „Oracle Database 12c Administrator Certified Professional“, “IBM InfoSphere Change Data Capture Technical Professional”, etc.
INTERNAL IT PARTNER FOR DAIMLER

+ Holistic solutions according to the Daimler guidelines
  + IT strategy
  + Security
  + Architecture
+ Developing and securing know-how
+ TSS is a partner who can be trusted with sensitive data

As subsidiary: **maximum added value** for Daimler
+ Market closeness
+ Independence
+ Flexibility (short decision making process, ability to react quickly)
DAIMLER TSS. HOLISTIC, INNOVATIVE, CLOSE

Digital Retail
Information Security
Car IT & Mobility
Analytics
Strategic Initiatives
Digital Customer Experience
Daimler TSS Germany
4 locations
Ulm (headquarters)
Stuttgart
Berlin
Karlsruhe
1058 employees*

Daimler TSS China
Hub Beijing
10 employees*

Daimler TSS Malaysia
Hub Kuala Lumpur
48 employees*

* as of 31.12.2017
1. Data Products
2. Data Pipeline
3. Summary
Software is becoming more and more important
- 100Mio lines of code

Physical products
- are significantly enhanced with digital service capabilities, e.g. the value of the car comes increasingly from digital assets
- become digital services, e.g. car2go
- Fleet management

IOT, Robotics, etc.

Source image: https://www.linkedin.com/pulse/20140626152045-3625632-car-software-100m-lines-of-code-and-counting
AI FOR CAR DIAGNOSIS – PREDICTIVE REPAIR DIAGNOSIS

1.) Vehicle with a problem in a workshop
2.) Vehicle status is read and transmitted
3.) Patterns from previous repairs to identify problem and predict remedy
4.) Better and faster repairs in the workshop
PRODUCT VISION: REPAIR DIAGNOSIS

• Collect and use data from workshops
  • worldwide
• Use warranty claims
• Predict remedy
  • Work position (Arbeitsposition)
  • Damage type/code (Schadensschlüssel)
  • Parts (Teilenummer)
  • Feedback from workshops: 86% of forecasts are considered as helpful (600K forecasts per month)
  • Integration with existing diagnosis system
UP TO 3 REPAIR FORECASTS
agenda

1. Data Products

2. Data Pipeline

3. Summary
OVER 75%

- Of time is spent for
- Say they least enjoy

DATA PREPARATION
**DOES METADATA MANAGEMENT PROVIDE ANSWERS TO SUCH QUESTIONS ACROSS THE WHOLE WORKFLOW?**

<table>
<thead>
<tr>
<th>Find</th>
<th>Understand</th>
<th>Trust</th>
<th>Access</th>
<th>Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search for data</td>
<td></td>
<td></td>
<td></td>
<td>Work with data</td>
</tr>
<tr>
<td>What tables are important?</td>
<td></td>
<td></td>
<td>How to get access to the data?</td>
<td>Is FIN unique?</td>
</tr>
<tr>
<td>Who knows about the data?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What table contains diagnosis data?</th>
<th>How is this column calculated?</th>
<th>Is the data reliable?</th>
<th>How to join the tables?</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the difference between <code>production_date</code> and <code>prod_dt</code>?</td>
<td>Who knows about the data?</td>
<td>How to join the tables?</td>
<td>How to get access to the data?</td>
</tr>
</tbody>
</table>

Data Products – From PoC into Production | 11-2018 | 14
CATALOGS ARE EVERYWHERE ... GOOGLE, AMAZON
INVENTORY VS USER EXPERIENCE

Suppliers provide inventory

• A catalog should list everything that is actually available

Consumers require user experience

• A catalog should provide data usage statistics, ratings, data samples, statistical profiles, lineage, lists of users and stewards, and tips on how the data should be interpreted
DATA CATALOG – AMAZON FOR INFORMATION

Data Catalog

- Technical Metadata
- Business Metadata
- Automation
- Machine Learning
- Collective Intelligence
- Governance
- Expert Sourcing
- Data Access

Inventory

User experience & enrichment

Data Products – From PoC into Production | 11-2018

Daimler TSS
Most people in AI forget that the hardest part of building a new AI solution or product is not the AI or algorithms — it’s the data collection and labeling.

Source: https://medium.com/startup-grind/fueling-the-ai-gold-rush-7ae438505bc2#.ywjvuca6z (Luke de Oliveira)
XGBOOST BY TIANQI CHEN
VERSION 2

• open-source software library which provides a gradient boosting framework
• >50% winners in Kaggle competitions
• Based on Decision Trees: majority decision of several trees
  • trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree
• Sparse-aware
• Parallelization: an make use of multiple cores on CPU
• Good accuracy and speed
THE MACHINE LEARNING DATA PIPELINE

Data Products – From PoC into Production | 11-2018
MONITORING AND TESTING: HOW TO PROVIDE EVIDENCE THAT A SYSTEM IS WORKING AS INTENDED?

- Unit testing of individual components?
- End-to-end tests?

Correct behavior is unknown. Quality?
- Live monitoring of system behavior in real time, e.g. comprehensive logging
- Responses, e.g. user feedback is priceless if available
1. Data Products
2. Data Pipeline
3. Summary
POC IS FAR FROM BEING PRODUCTION-READY

“We made 20 models last year. Only two made it to production!” Chief Data Scientist at a major Wall Street Bank.

Data scientists rarely have strong engineering backgrounds and need help with the manual coding required to make their models API accessible to the rest of the business.

Source: 2018 Enterprise Almanah
A tiny fraction of the code in many ML systems is actually devoted to learning or prediction.
ML PROJECTS

One-time analysis using ML for one-time decision

conscious choice of the required quality of services

Automated, productive ML model handling e.g. data quality issues

TECHNICAL DEBT IN ML SYSTEMS

• Developing and deploying ML systems is relatively fast and cheap
• Maintaining them over time is difficult and expensive

Technical debt: introduced by Ward Cunningham in 1992 to help reason about the long term costs incurred by moving quickly in software engineering

Reduce technical debt by

• refactoring code
• improving unit tests
• deleting dead code
• reducing dependencies
• tightening APIs
• improving documentation

DATA SCIENCE VS SOFTWARE ENGINEERING
POC VS PRODUCTION-READY

Data Products – From PoC into Production | 11-2018
MULTI-LANGUAGE CHALLENGE

Multiple languages increase the cost of

• effective testing
• transferring ownership to other individuals
• establishing [repeatable & measurable] standards
DATA ENGINEERING CHALLENGE

- Standardisation
- joins
- sampling steps
- intermediate files output
- detecting and handling errors

Engineer for data collection and feature extraction from the beginning with proper tools. Refactoring: Redesigning from ground up is a worth while major investment

- reduce ongoing costs
- speed further innovation.
SUMMARY

PoC is far from being production ready

Data preparation and data quality are paramount

Infrastructure gets more and more important

Software Engineering
FURTHER VISION

New vehicle generations send constantly diagnosis data and not just during workshop visit

Damage forecast

Health monitor for a vehicle fleet as a data product