Microservice Architectures for Data Science Applications

Prof. Dr. Wilhelm (Willi) Hasselbring
Software Engineering Group, Kiel University
http://se.informatik.uni-kiel.de/
GEOMAR Data Science Symposium, April 19th, 2018
Agenda

1. A look at industrial, Internet-scale Software Systems
2. Software Engineering for Computational Science
3. OceanTEA: Platform for Repeatable Ocean Observation Data Processing
4. GeRDI: Generic Research Data Infrastructure
5. Summary & Outlook
Facts and figures about OTTO

<table>
<thead>
<tr>
<th>Profile</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Founded</td>
<td>August 17th, 1949</td>
</tr>
<tr>
<td>No. of employees</td>
<td>4,350</td>
</tr>
<tr>
<td>Revenue in 2015/16 FY</td>
<td>2.563 billion Euros</td>
</tr>
<tr>
<td>Online revenue share</td>
<td>round 90 percent</td>
</tr>
</tbody>
</table>

http://www.ottogroup.com/de/presse/material.php
Otto Web Shop until 2013

https://support.intershop.com/kb/index.php/Display/276B90
In 2011, Otto started a complete re-implementation of their ecommerce software from scratch. The drivers for this decision were diverse, but had mostly to do with non-functional requirements like scalability, performance and fault tolerance.

- Regarding scalability, they were not only thinking about technical scalability in terms of load or data.
- They needed a solution that was scaling with respect to the number of teams and/or developers working on the software at a given time.
- In addition to that, they planned to practice DevOps including continuous deployment, in order to deliver features quickly to the customer.

https://www.otto.de/unternehmen/de/newsroom/dossiers/lhotse.php
Modernization Strategy

• What they have found was in the first place a little bit unusual, but in the end highly successful:
  – Instead of setting up a single development team to create a new platform for the shop, they were actively employing Conway’s Law by starting development with initially four separate teams with four loosely coupled applications (a.k.a. microservices):
    • Product, being responsible for products and their presentation.
    • Order for shopping carts and the order process.
    • Promotion, serving product recommendations and promotions for assortments, brands, and so on.
    • Search and Navigation for search and navigation in the shop.

• In the following years, they founded several more teams and systems.
Verticals for Business Functions
Example: otto.de

[Hasselbring & Steinacker 2017]
Life Deployments @ Otto.de
Life Deployments & Incidents @ Otto.de

Live-Deployments and Prio 1 Incidents per Week 2014-2017

[Reliability]

[Agility]

[Reliability]

[Hasselbring & Steinacker 2017]
Agenda

1. A look at industrial, Internet-scale Software Systems
2. Software Engineering for Computational Science
3. OceanTEA: Platform for Repeatable Ocean Observation Data Processing
4. GeRDI: Generic Research Data Infrastructure
5. Summary & Outlook
Message:
Adapt existing software engineering techniques for computational science

https://doi.org/10.1109/MCSE.2018.108162940

[Johanson & Hasselbring 2018]
Agenda

1. A look at industrial, Internet-scale Software Systems
2. Software Engineering for Computational Science
3. OceanTEA: Platform for Repeatable Ocean Observation Data Processing
4. GeRDI: Generic Research Data Infrastructure
5. Summary & Outlook
Cloud-Based Platform for Repeatable Ocean Observation Data Processing

OceanTEA

[Future Ocean, Kiel Marine Sciences]

[Johanson et al. 2016]
4D Spatial Analysis with OceanTEA
Machine Learning on Ocean Observation Data with OceanTEA

- Paper: http://dx.doi.org/10.1016/j.ecoinf.2017.02.007
- Source code: https://github.com/a-johanson/oceantea
- Software service with data: http://maui.se.informatik.uni-kiel.de:9090/
  (URL will change, refer to the GitHub repository for updates)

References:

Modeling polyp activity of Paragorgia arborea using supervised learning

Arne N. Johanson, Sascha Flögel, Wolf-Christian Dullo, Peter Linke, Wilhelm Hasselbring

[19.04.2018 W. Hasselbring]
OceanTEA: Microservice-based Architecture

Agenda

1. A look at industrial, Internet-scale Software Systems
2. Software Engineering for Computational Science
3. OceanTEA: Platform for Repeatable Ocean Observation Data Processing
4. GeRDI: Generic Research Data Infrastructure
5. Summary & Outlook
Generic Research Data Infrastructure

GeRDI

http://www.gerdi.-project.de/

Gefördert durch
Deutsche Forschungsgemeinschaft

http://www.gerdi.-project.de/
Research Data Lifecycle

- Harvest
  - Search
  - Bookmark

- Archive
  - Store
  - Publish

- Giving Access to Data
- Preserving Data
- Analysing Data
- Processing Data
- Re-Using Data
- Creating Data

- Publish
- Preprocess
- Analyze

UK DATA ARCHIVE
Generic Research Data Infrastructure

[Tavares de Sousa et al. 2018]
Design Rationale

• Self-contained Systems (SCS)
  – Each SCS is an autonomous (web) application.
    • For the SCS's domain all data, the logic to process that data and all code to render the web interface is contained within the SCS.
  – Communication with other SCSs or 3rd party systems is asynchronous wherever possible.
    • This decouples the systems, reduces the effects of failure, and thus supports autonomy.
  – To avoid tight coupling an SCS should share no business code with other SCSs.

• Entry / Exit Options for users at various points in the workflow
Summary & Outlook

• Modularity is essential for sustainable software systems
  – Including data science applications

• Microservices provide a disruptive architectural style for modular systems
  – Successful for Internet-scale systems
  – However, non-trivial to achieve

• Transfer to research software
  – OceanTEA: Platform for Repeatable Ocean Observation Data Processing
  – GeRDI: Generic Research Data Infrastructure
References


19.04.2018

W. Hasselbring