Copernicus Data Infrastructure NRW

Arne de Wall, Albert Remke, Thore Fechner, Jan van Zadelhoff, Andreas Müterthies, Sönke Müller, Adrian Klink, Dirk Hinterlang, Matthias Herkt, Christoph Rath

GDI-NW Nordrhein-Westfalen GeoIT Round Table

Copernicus Relays

Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen

Information und Technik Nordrhein-Westfalen

EFTAS

con terra

52north exploring horizons

Introduction

The remote sensing competence center of the State Agency for Nature, Environment and Consumer Protection in North Rhine-Westphalia (LANUV) deals with the use of remote sensing methods to support nature conservation tasks. Particular attention is paid to the European earth observation program Copernicus, which has been operational since 2014 and has enormous potential for LANUV. However, the IT infrastructure currently used for the work of the competence center is not sufficiently dimensioned to handle the data volume and the computing-intensive data processing tasks. Other government agencies such as Geobasis.NRW and Forest & Wood see themselves in a similar situation.

IT.NRW supports the state authorities in its function as state computing center and is asked to set up a service offering for the use of remote sensing data. An agile methodology is used, which makes it possible to take into account the dynamic technological development in this area and to best meet the requirements of the users, i.e. very quickly and cost-effectively.

In a first step, a pilot project - CDI@IT.NRW - was carried out based on IT.NRW's IT infrastructure, in which specific information products were derived from Copernicus data sources and made available to users. As part of this project, architectural questions were answered and the first structures developed that contribute to the sustainable development of a powerful Copernicus Data Infrastructure for the state administration of NRW. The project has been inspired by the activities of the GeoIT Round Table NRW, which one of the Copernicus Relays in Germany.

In this display, the goals and results of the pilot project CDI@IT.NRW are presented with a focus on the technical aspects.

Overall Objectives

The goal of the CDI@IT.NRW project was to establish a software architecture for processing large amounts of EO data based on a powerful server infrastructure, which makes it possible to exploit the vast potentials of Copernicus Data for creating NRW-wide EO information products.

Use-Case

- Partially automated proof-of-concept data processing workflow for monitoring orchards (ingestion, pre-processing, processing, publication) in the CDI@IT.NRW environment

Aims

- Answer basic questions about data organization and efficient processing of Copernicus Data as an integral part of the GeoIT-infrastructure at IT.NRW

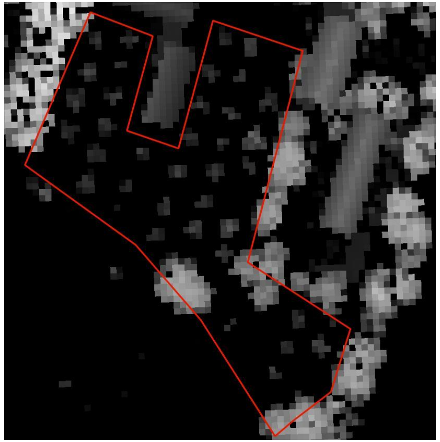
Use case & Requirements

Monitoring of Orchards

- Relevance: existing reporting obligations on the occurrence and condition of these biotopes / habitats
- New, high temporal resolution remote sensing products can be used for frequent and large scale change detection purposes
- In our case: use remote sensing data to re-evaluate an existing inventory of orchards, i.e. detect changes

Requirements

- Creating an information product for the entire area of North Rhine-Westphalia
- Complex and large amount of input data (Sentinel-2, nDOM, DOPs, …)
- (Partially) automated and efficient processing
- Chaining processes into workflows that can be automatically executed in a scalable and efficient processing environment
- Infrastructure / Architecture that is suitable to support similar use cases



Architectural Considerations

- Modular and reusable design of the infrastructure
- Horizontal and vertical scalability during operation
- Dynamic on-demand use of the hardware
- Technical foundation: The core of the processing environment is based on the concept of containerization, where the individual processing components are defined in form of container images
- Generic API for control and orchestration based on a Kubernetes environment
- Storage and computing resources, available to the pilot project
  - Server: RAM 768 GB, CPU: 24 cores, video Nvidia Tesla V100, RedHat Linux
  - Storage: 200 GB SSD, ~ 4000 GB SAN

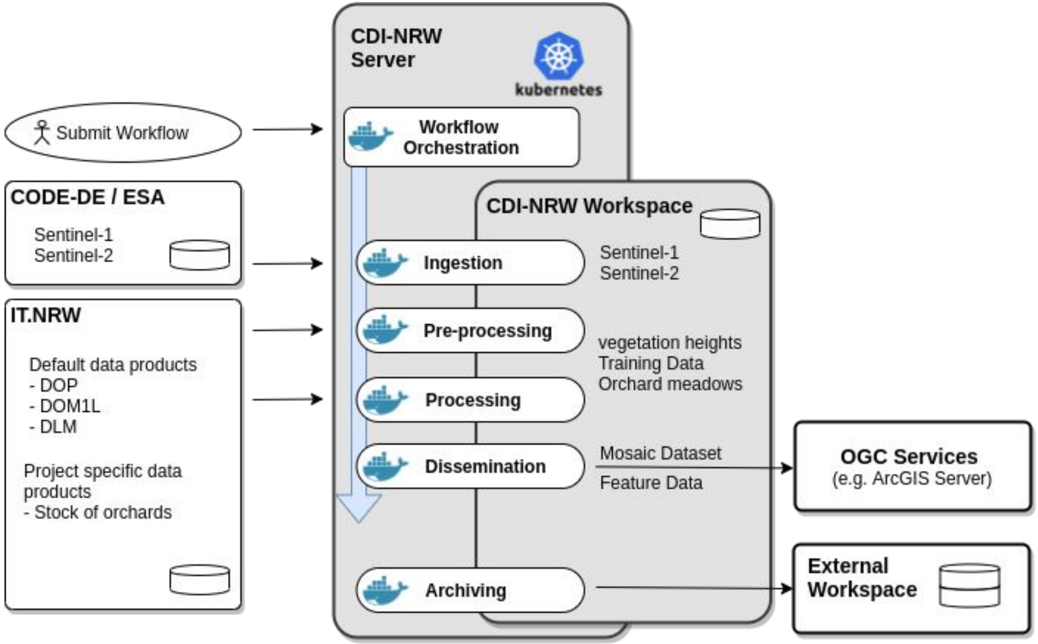
Architectural Approach

Reusability, modularization, automation and scaling are central aspects of the considerations for the architecture of the CDI. Cloud technologies are extremely helpful here.

- Breakdown of the overall processes into automatable sub-processes that can be configured and reused across different application contexts (e.g. ingestion, pre-processing, processing, dissemination)

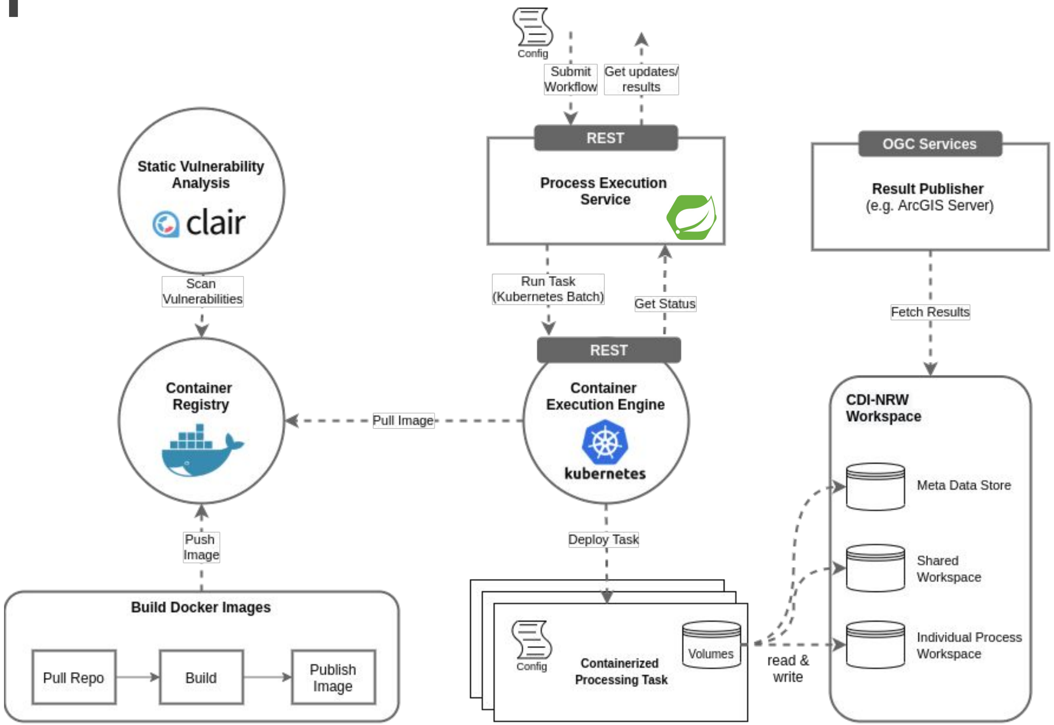
- Composition of individual modules to configurable workflows that can be automated and executed in parallel.

- Shared and process individual workspace utilization for reading and writing data products and intermediate results.



Architectural Approach

- The **Process Execution Service** extends the Docker and Kubernetes-driven architecture by an additional workflow orchestration and controlling service.
    - JSON-based workflow definition and submission via a REST interface.
    - Sequential execution of the workflow tasks as Kubernetes Batch jobs. Automatic start of next steps upon completion of previous steps.
    - Monitoring of running tasks as well as logging and exception handling
- Shared and process individual workspace utilization for reading and writing data products and intermediate results.



Security Aspects

An important aspect of the project is the level of security provided by docker containers. Docker containers only provide a very lightweight virtualization where it takes advantage of the Linux kernel's ability to create isolated environments in which the applications run. Therefore, they share the hosts' kernel, but work in separate runtime environments thanks to Cgroups and Namespaces. However, this level of container isolation is less tight and processes have access to hosts' kernel, which provides additional security concerns and are particularly crucial for official institutions.

- **KataContainers** - enables VM-isolation of Docker on a container level. It bridges the gap between traditional VM security and the advantages of Docker containers by running a lightweight kernel for each container. It seamlessly integrates with the OCI Runtime Specification.

- **CoreOS Clair** - is an Open Source scanning component that enables the static vulnerability analysis on the level of container images.

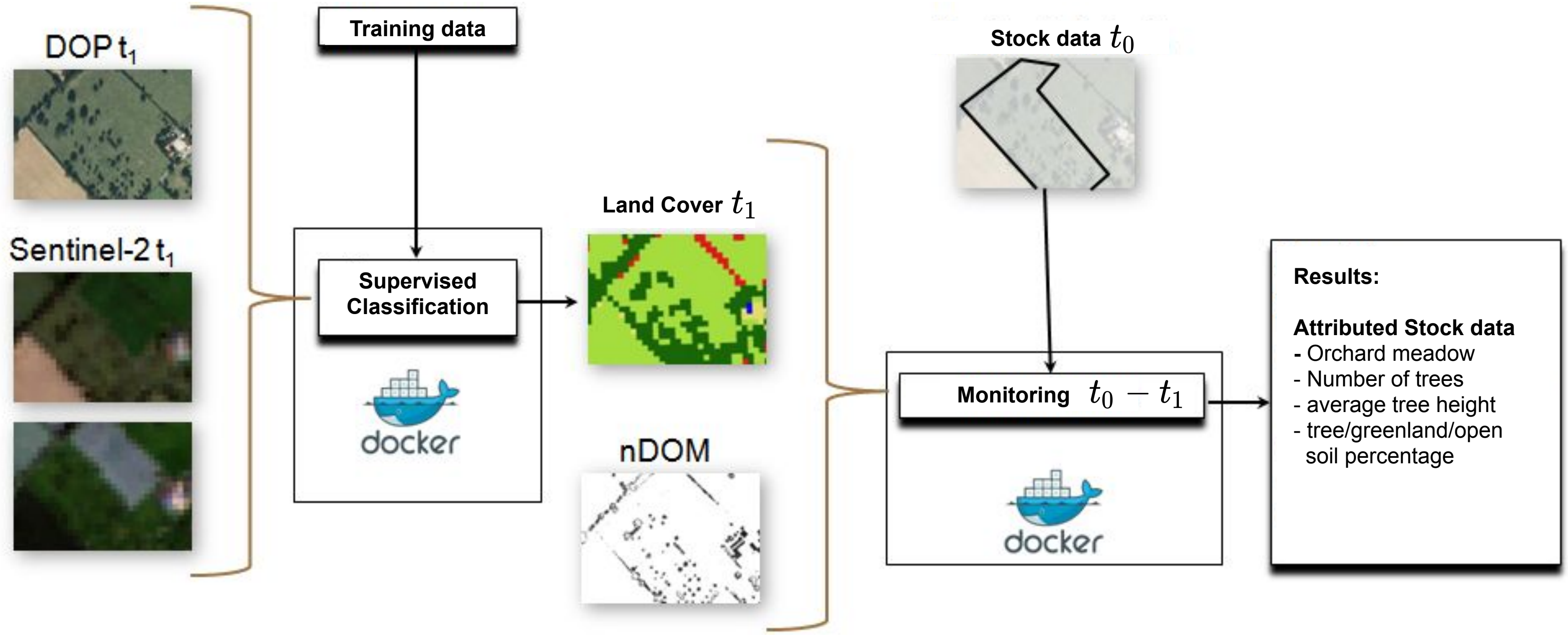
docker

katacontainers

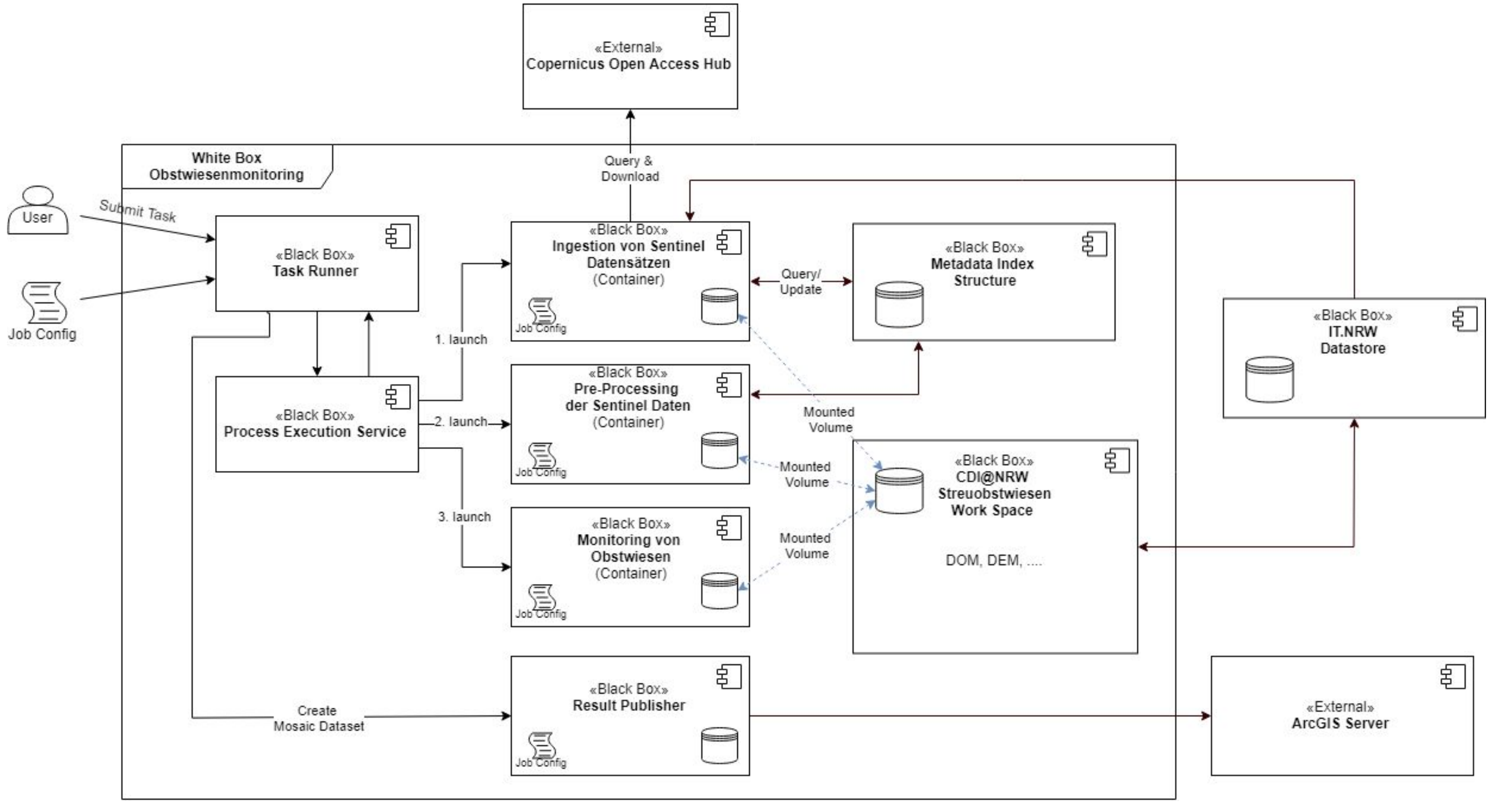
clair

A Container Image Security Analyzer by CoreOS

Use-Case: Monitoring Orchards Workflow



Architecture - Components Designed for the Use Case

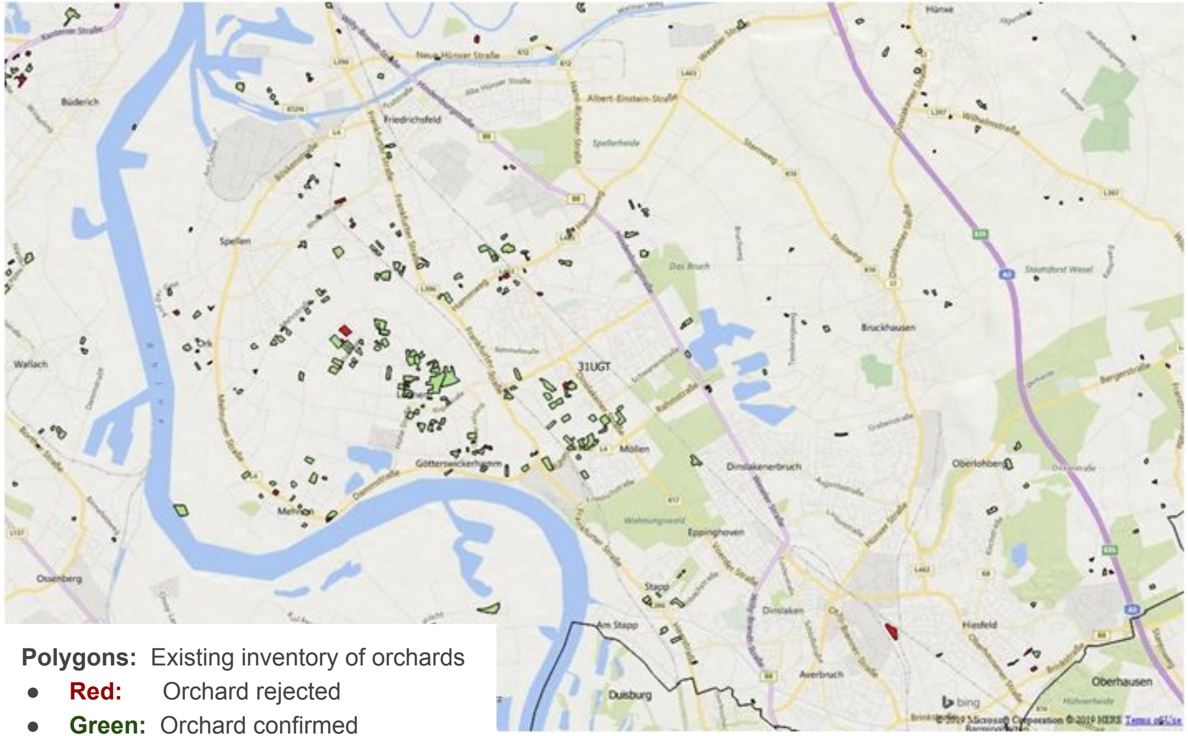


Results of the Data Analysis

Polygons: Existing inventory of orchards

- **Red:** Orchard rejected
- **Green:** Orchard confirmed

Results of the Monitoring



Polygons: Existing inventory of orchards
- **Red:** Orchard rejected
- **Green:** Orchard confirmed

Wesel district in total:

Monitoring result	Number of Objects	Ratio
confirmed	410	77 %
rejected	125	23 %

Discussed Development Paths for the Future

Technical Improvements

- Enhance the integration with DIAS (Data and Information Access Services) platforms such as CODE DE, Mundi Web Services. Evaluate hybrid approaches.
- Improve flexibility of the workflow execution by using established technology stacks and standards (e.g. Common Workflow Language (CWL), Argo (CNCF incubating project), Apache Airflow)
- Make use of Harbor (CNCF incubating project) for an improved container management including content signing and validation, security and vulnerability analysis.

Summary & Resumé

- **The CDI@IT.NRW pilot project was successfully conducted**
  - Based on a requirements analysis a software architecture for scalable semi-automated GeoIT processing chains has been developed.
  - The technical approach is based on strict containerization and workflow management using Docker & Kubernetes; the special security concerns of a public sector computing center have been taken into account.
  - The architecture has been evaluated successfully by implementing a process chain for detecting changes in an existing state wide orchard inventory (34,000 km²).
    Estimated processing time is about 6 hrs (ingestion, pre-processing, processing)

- **Resumé**
  - The use of satellite and in situ data for the creation of large-scale information products for applications in the field of nature conservation requires powerful IT infrastructures
  - By using cloud technologies (especially Docker, Kubernetes), conventional IT infrastructures can be enabled to support these tasks with reasonable efforts.
  - Improving the ability of existing computing centers to meet the requirements of using Copernicus data is an important step on the way of integrating Copernicus into Spatial Data Infrastructures such as the GDI NRW.

Thank you for your attention!

Contacts

- Arne de Wall, Prof. Dr. Albert Remke, 52°North
- Dr. Thore Fechner, Jan van Zadelhoff, con terra GmbH
- Dr. Andreas Müterthies, Dr. Sönke Müller, Adrian Klink, EFTAS GmbH
- Dr. Dirk Hinterlang, Matthias Herkt, LANUV NRW
- Christoph Rath, IT.NRW