Data Life Cycle Lab Earth and Environment

LSDMA All-Hands Meeting Mar 10, 2016

Jörg Meyer

The Team

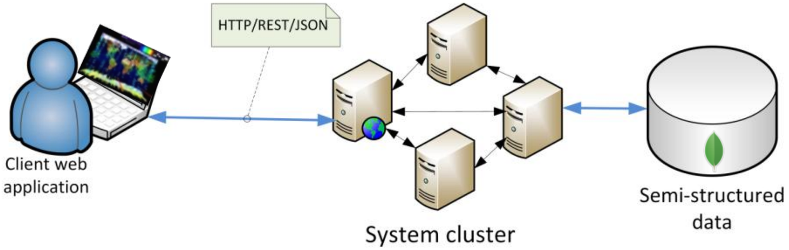
- DKRZ
  - Carsten Ehbrecht
  - Stephan Kindermann
  - Michael Lautenschlager

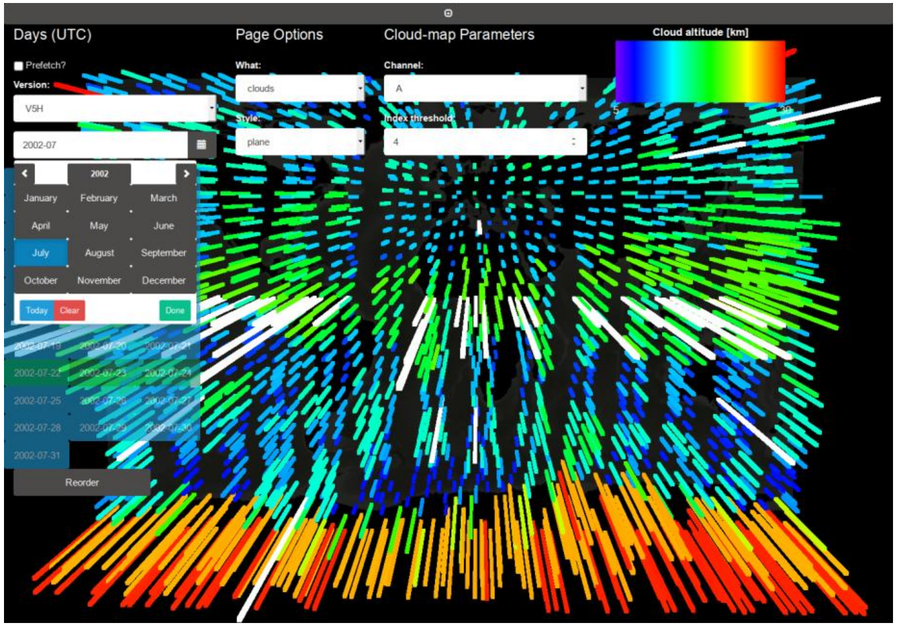
- KIT
  - Parinaz Ameri
  - Uğur Çayoğlu
  - Jörg Meyer
  - Marek Szuba

  - Students: Jiang Zhong Bo, Haipeng Guan, Florian Klemme

Visualization and Data Fusion

- KAGLVis: browser application for visualization of large amount of unstructured data (MIPAS)
- based on components developed in DLCL (Node Scala)
- continuation of development
- involved in Helmholtz initiative for environment visualisation „Komplexe Umweltdaten: Exploration, Interpretation, Synthese“ (KUDOS)
- data fusion
- planned proposal with HPI




Days (UTC)

□ Prefetch?

Version: V5H

2002-07

Page Options

What: clouds

Style: plane

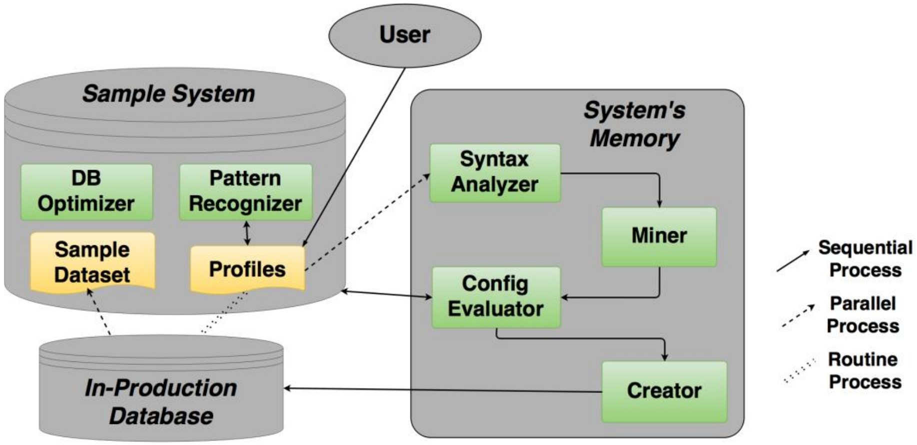
Cloud-map Parameters

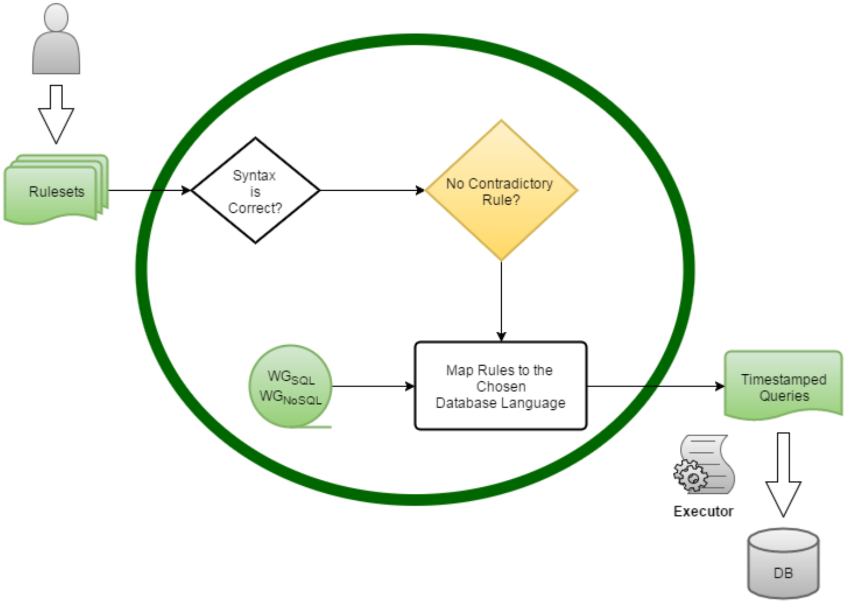
Channel: A

Index threshold: 4

Physical Database Design and Benchmark

- Extend Automated Index Selection Framework (MISA):
  - Model an optimized cost function considering:
    - RAM limitation, Different Workloads, Similarity of queries
  - Apply a Sample Model to Minimize Data Transfer



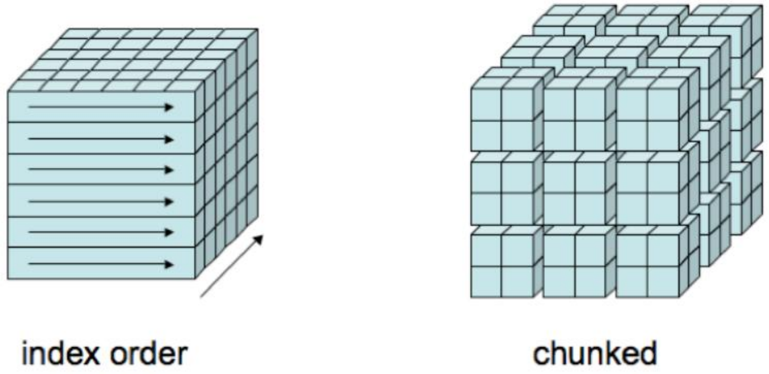


Release Generic Database Workload Generator (NoWog)

- Integrated Layer over Databases
- Generic Grammar
- Base for Application-Specific Benchmarking

Optimization of Climate Analses

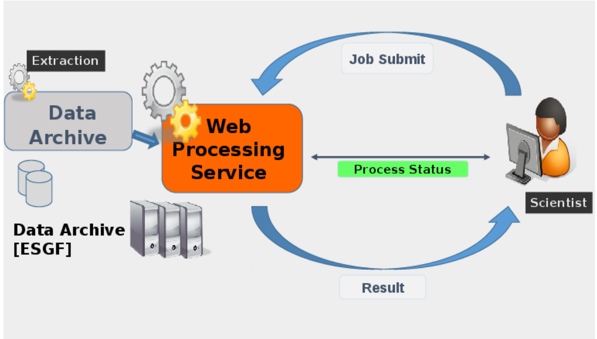
- Common PhD project of IMK-ASF and SCC
- Meta data catalogue for IMK data
- Improved analysis scripts
- Compression of climate data in NetCDF files
  - optimize data structure
  - increase compression rate
  - balance of size and access



http://www.unidata.ucar.edu/blog_content/images/2013/blog_rew_chunking.png

Geospatial Data Life Cycle Framework
Birdhouse

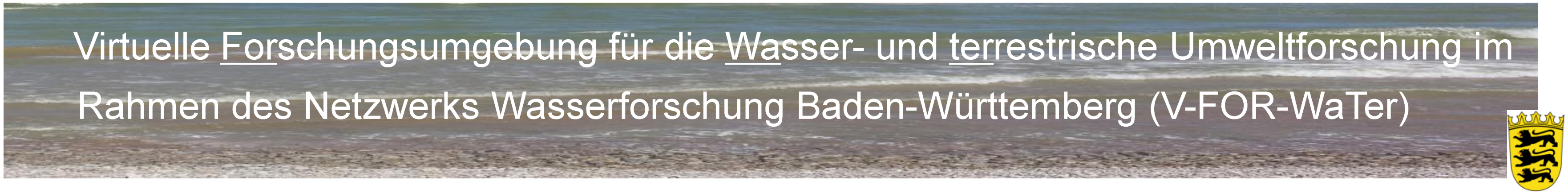
- Birdhouse: Web Processing Services for climate data
  - code: https://github.com/bird-house doc: http://bird-house.github.io/
  - based on:
    - Malleefowl: base processes and mandatory in a bird-house
    - Emu: a few test cases to try out
    - Hummingbird: provides CDOs and Quality Assurance tools as a service
    - Flyingpigeon: a collection of processes useful for the impact community
    - Phoenix: the simple web browser application for WPS



- Recent improvements:
  - New Twitcher component: a token based security proxy for WPS and other OGC services:
    - Implemented as a Python WSGI middleware.
    - Uses (short living) string tokens to access WPS processes securely.
    - Tokens can be part of the URL or header so that existing client and server WPS implementations can be used without modification.
  - Improved docker deployment for birdhouse components with docker compose.
  - Uploading of local files to the Phoenix web application to be used by WPS processes:
    - Example: run CF conventions checker on user uploaded NetCDF file.
    - Uploaded files are cached in a file storage and can be reused for processing.
    - Uploading to OpenStack Cloud planned.

V-FOR-WaTer

- Virtual research environment



Virtuelle Forschungsumgebung für die Wasser- und terrestrische Umweltforschung im Rahmen des Netzwerks Wasserforschung Baden-Württemberg (V-FOR-WaTer)

- BW project of IWG and SCC at KIT
- start: spring 2016
- goals:
  - VRE for systematic treatment of hydrology research data

    include data of Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW)
  - direct access to analysis tools
    - provide web processing services (WPS, OCG standard)
    - based on framework **birdhouse**

EUDAT2020

- Scientific communities environments and requirements
  - survey on data and computing landscapes, environments, and service requirements
  - interviews with technical community experts

- B2SAFE: safe replication of scientific data (iRODS + PIDs)
  - Technology
    - iRODS: rule-oriented data system
    - PIDs: persistent identifiers based on EPIC handles
- Users
  - GFZ Potsdam (seismology)
  - IST DataRep (repository for citable data)
  - Institut für Anatomie Leipzig (medical data)

Ongoing Projects

- GLORIA
  - MongoDB for GLORIA meta data
  - replica sets

- SAT
  - MongoDB for MIPAS geolocations
  - maintenance of geomatcher application
  - distributed database for climate data
    - profiles of trace gases in MongoDB shards