

Data Life Cycle Lab Earth and Environment

LSDMA All-Hands Meeting Oct 04, 2016

Jörg Meyer



The Team



- Carsten Ehbrecht
- Stephan Kindermann
- Michael Lautenschlager



- Parinaz Ameri
- Uğur Çayoğlu
- Jörg Meyer
- Marek Szuba
- Marcus Strobl (V-FOR-WaTer)
- Elnaz Azmi (V-FOR-WaTer)



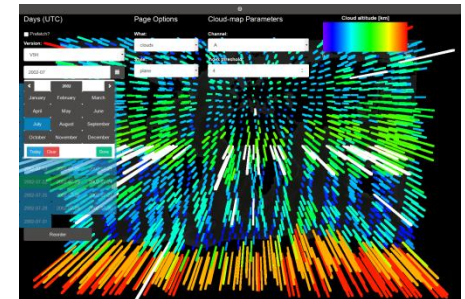
- Students: Stephan Holljes, Benedikt Riehm, Yannick Nesselhauf, Florian Klemme

Ongoing Projects

- GLORIA
 - MongoDB for GLORIA meta data
 - replica sets
- KAGLVis
 - browser application for visualization of large amount of unstructured data (MIPAS)

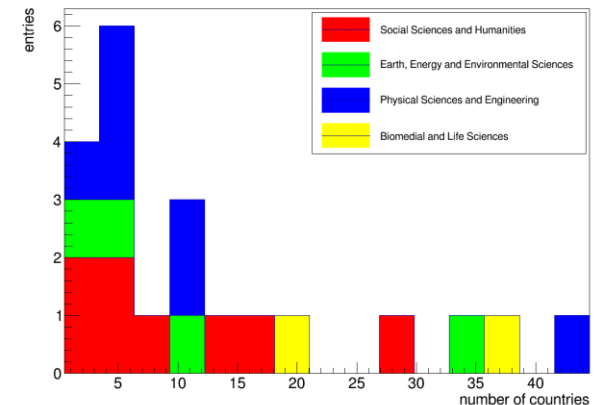
A Distributed System for Storing and Processing Data from Earth Observing Satellites: System Design and Performance Evaluation of the Visualisation Tool , CCGrid 2016, doi: 10.1109/CCGrid.2016.19

- Metadata catalogue
 - metadata of NetCDF files of IMK in MongoDB



Ongoing Projects

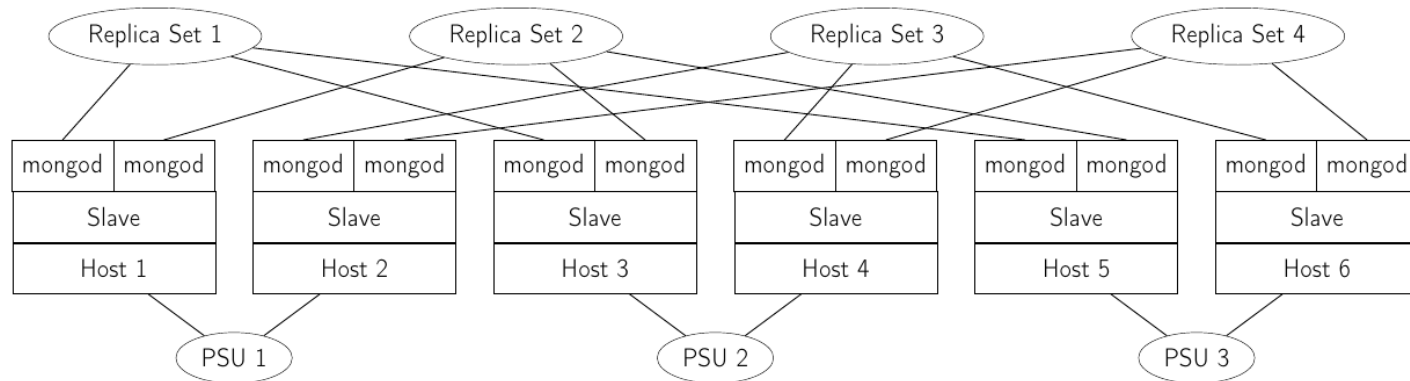
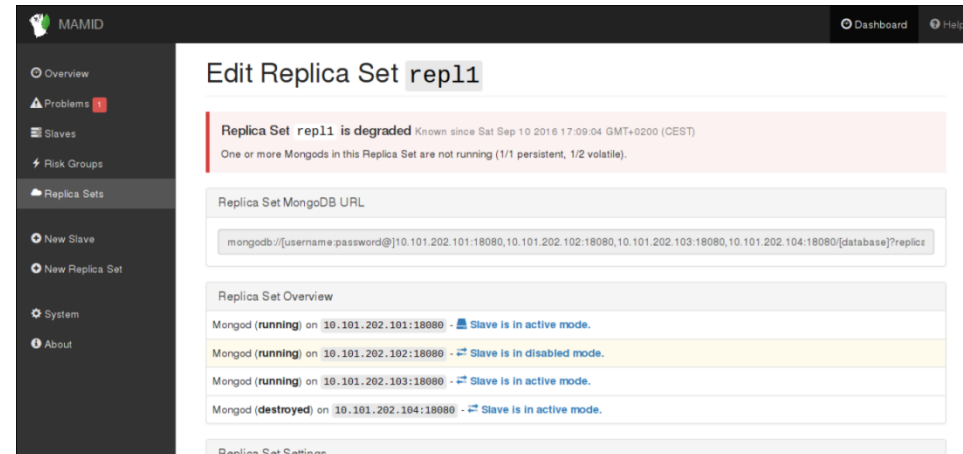
- EUDAT2020: B2SAFE
 - safe replication for
 - GFZ Potsdam (23 million files)
 - IST Data Repo
- EUDAT2020: Scientific Communities Environments and Requirements
 - detailed report (deliverable)
- geomatcher
 - MongoDB for geolocations
 - import of new data
 - new feature requests from IMK



In-memory MongoDB



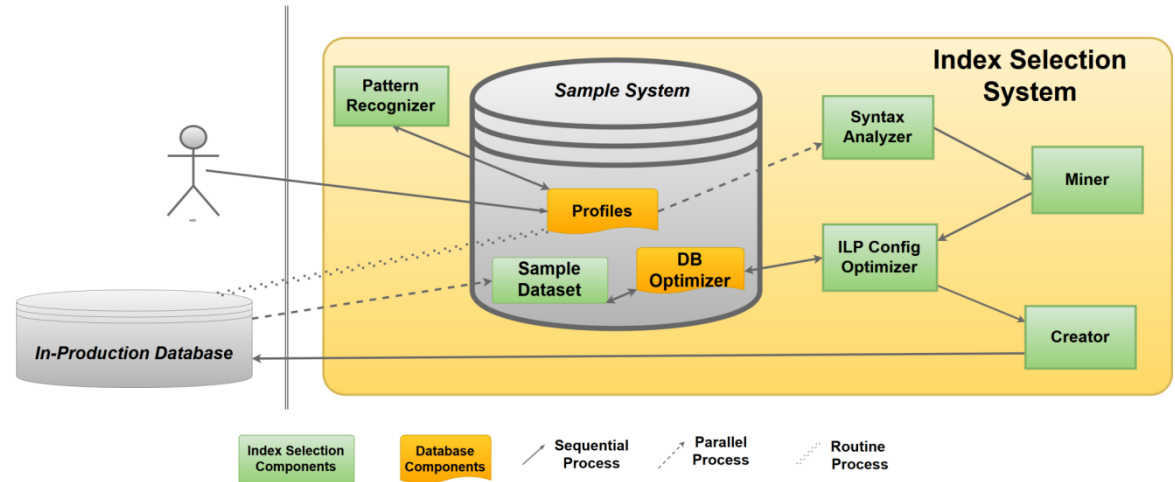
- 96+ TB in distributed MongoDB
- all data in memory
- tool to initialize database, add nodes, configure replica sets, sharding
- environment: blade cluster at IMK
- use case: MIPAS profile data



Niklas Fuhrberg, Anton Schirg, Christian Schwarz, Janis Streib, Bob Weinand

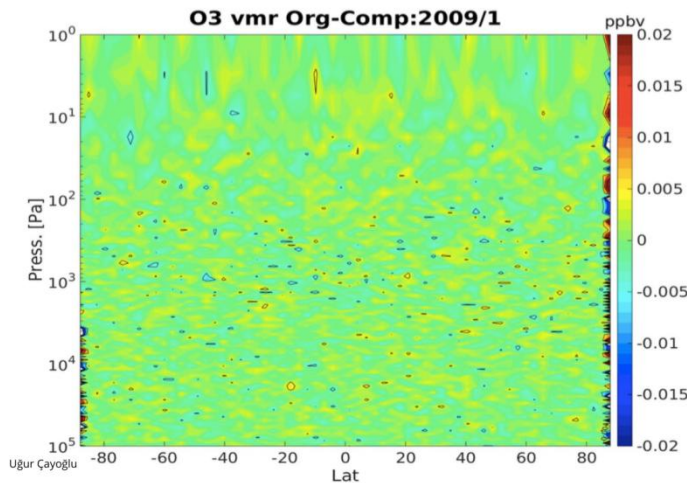
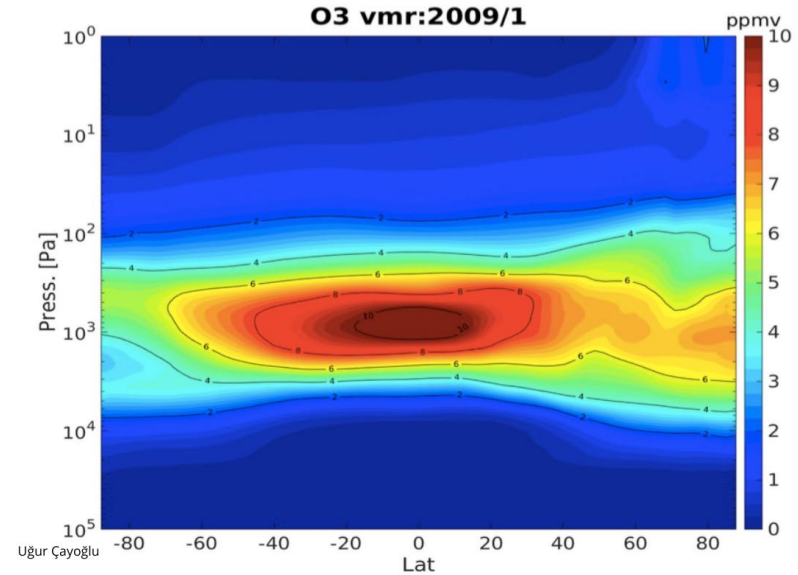
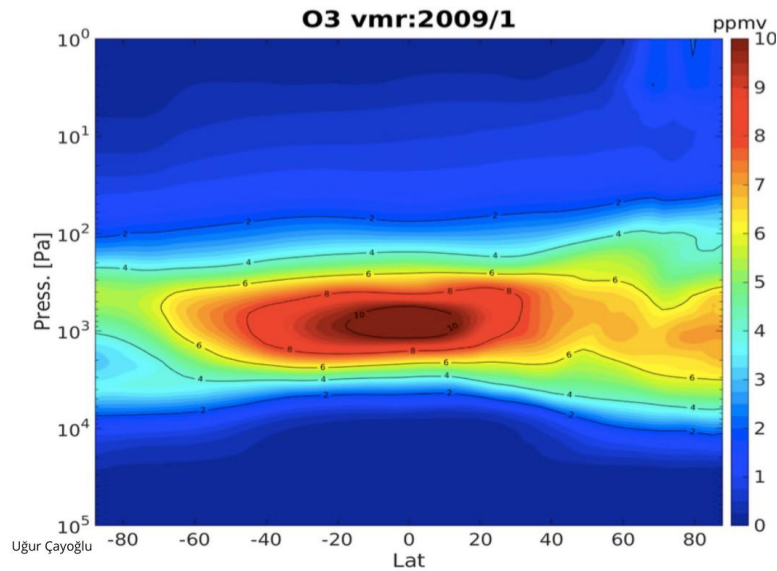
- Extend Automated Index Selection Framework (MISA):

- cost function
- statistical sampling
- evaluation



- On a New Approach to the Index Selection Problem using Mining Algorithms
doi: 10.1109/BigData.2015.7364084
- On a Self-Tuning Index Recommendation Approach for Databases
doi: 10.1109/ICDEW.2016.7495648
- Challenges of Index Recommendation for Databases, 28th GI-Workshop on Foundations of Databases <http://ceur-ws.org/Vol-1594/paper3.pdf>
- NoWog: A Workload Generator for Database Performance Benchmarking
doi:10.1109/DASC-PICom-DataCom-CyberSciTec.2016.120

Lossless Compression of Environmental Datasets



differences due to lossy
compression not acceptable for
scientists

➡ need lossless compression

Lossless Compression of Environmental Datasets

Prediction based compression

- use neighbouring data
- optimize x_1, x_2, x_3
- residuals:
diff between data and prediction

d_{11}	d_{12}	d_{13}	d_{14}	d_{15}
d_{21}	d_{22}	d_{23}	d_{24}	d_{25}
d_{31}	d_{32}	d_{33}	d_{34}	d_{35}
d_{41}	d_{42}	d_{43}	d_{44}	d_{45}
d_{51}	d_{52}	d_{53}	d_{54}	d_{55}

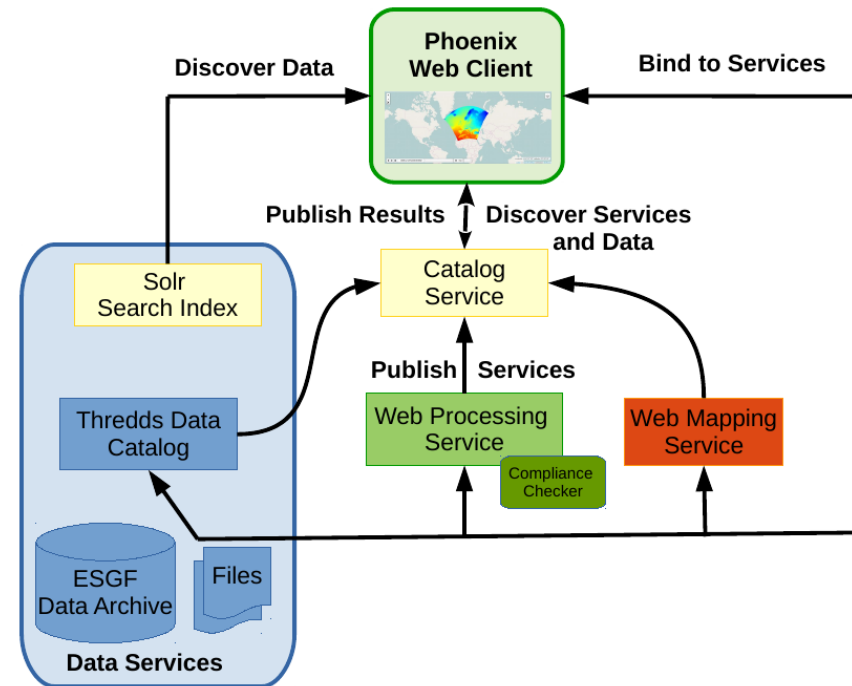
$$x_1 \cdot d_{22} + x_2 \cdot d_{23} + x_3 \cdot d_{32} \approx d_{33}$$

Geospatial Data Life Cycle Framework Birdhouse



latest changes

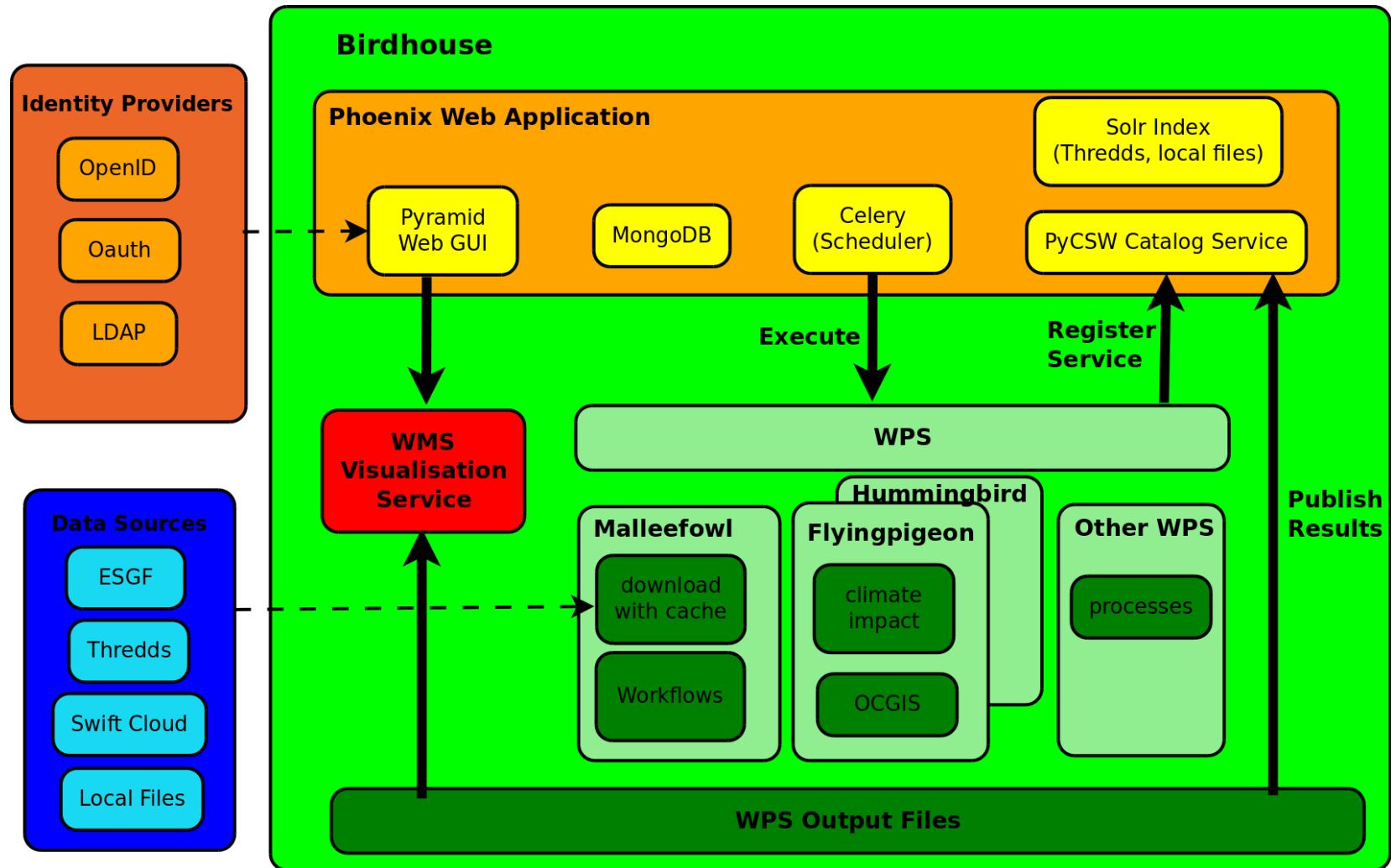
- improved Phoenix user interface
- Leaflet map to visualize climate data
- upload widget for user supplied files
- enhanced deployment with buildout and ansible
- using the new PyWPS-4 implementation (supports Python 2.7 and 3.x)

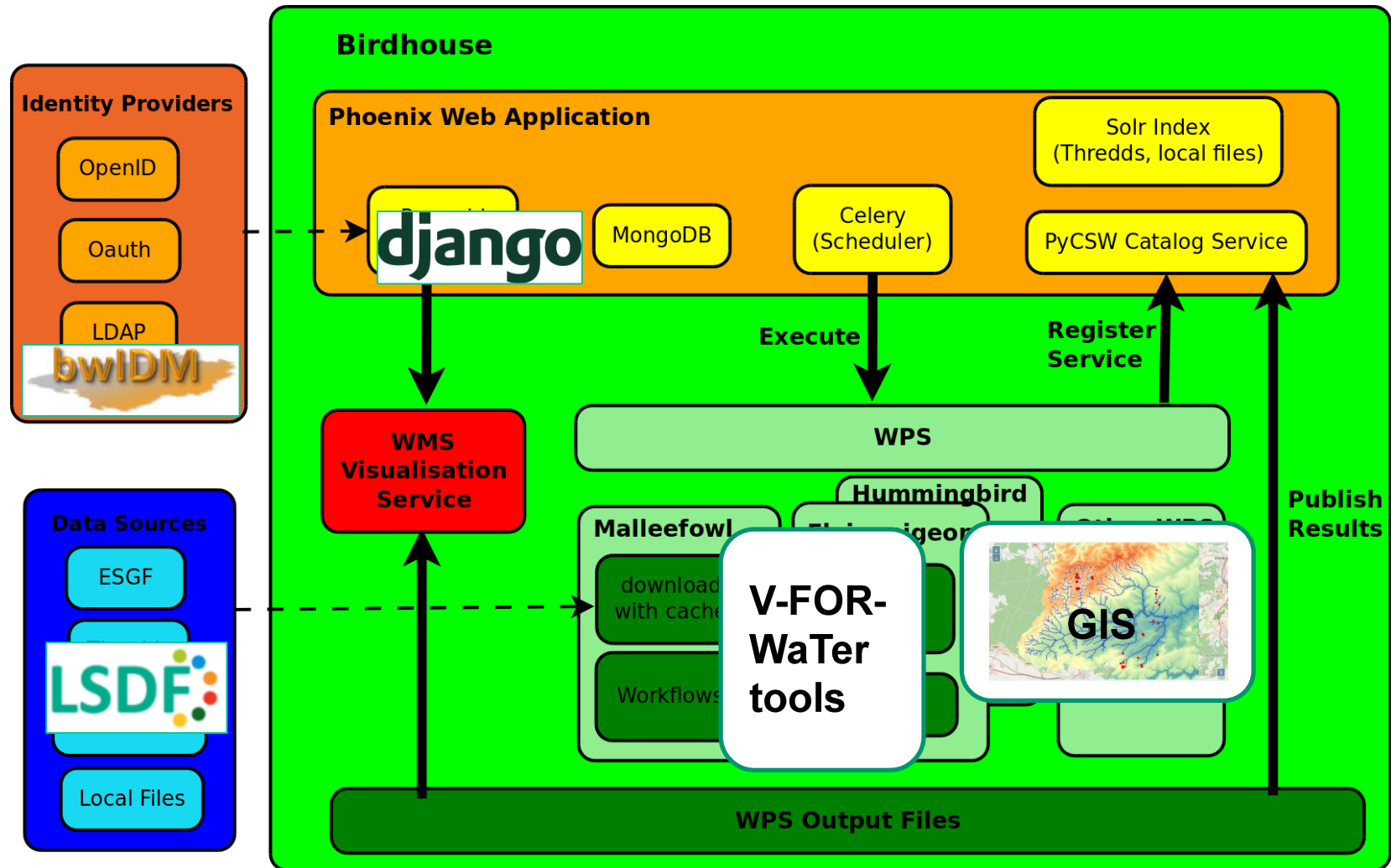


- Virtual research environment for handling of hydrology research data
- funded by state of Baden-Württemberg
- partners: IWG and SCC at KIT
- start: July 2016
- goals:
 - VRE for systematic treatment of hydrology research data
include data of Landesanstalt für Umwelt, Messungen und Naturschutz
Baden-Württemberg (LUBW)
 - direct access to analysis tools
 - provide web processing services (WPS, OGC standard)
 - based on framework **birdhouse**
- <http://www.vforwater.de>



Birdhouse Architecture





- continuation of most activities
- second half of EUDAT 2020
- MongoDB related
- Climate data compression
- V-FOR-WaTer

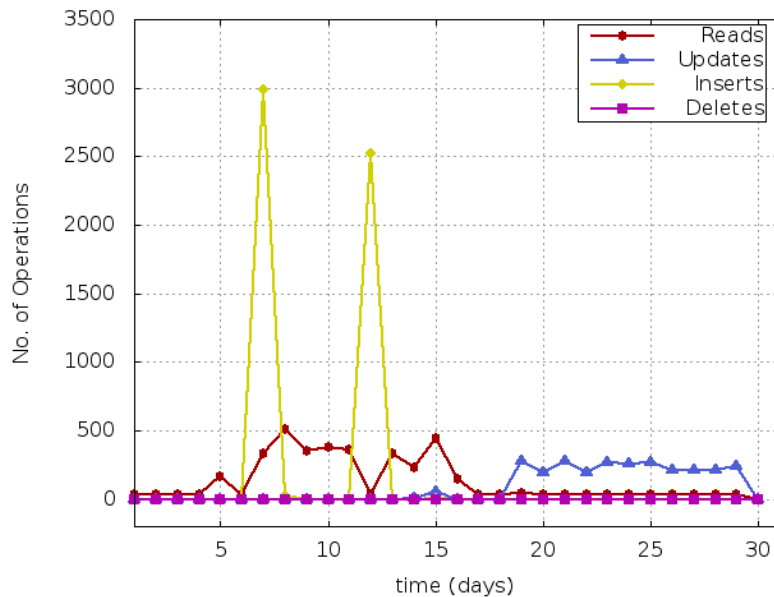
backup

NoWog: Workload Generator



Real MongoDB Read:

```
db.nowog.find({"cloud_index":true,  
              "zpd":{"$gte":200,  
                  "$lt":478  
              }).sort({"zpd":1})
```



NoWog Description:

```
find_SORT2:{  
  {(cloud_index:True)  
   (zpd:range_op)(zpd:range_op)},  
  },  
  {(zpd:1)},  
  14400 - 21600= normal(1000.6,160));
```

