Data Life Cycle Lab Earth and Environment

LSDMA All-Hands Meeting Oct 2, 2015

Jörg Meyer

LSDMA

The Team

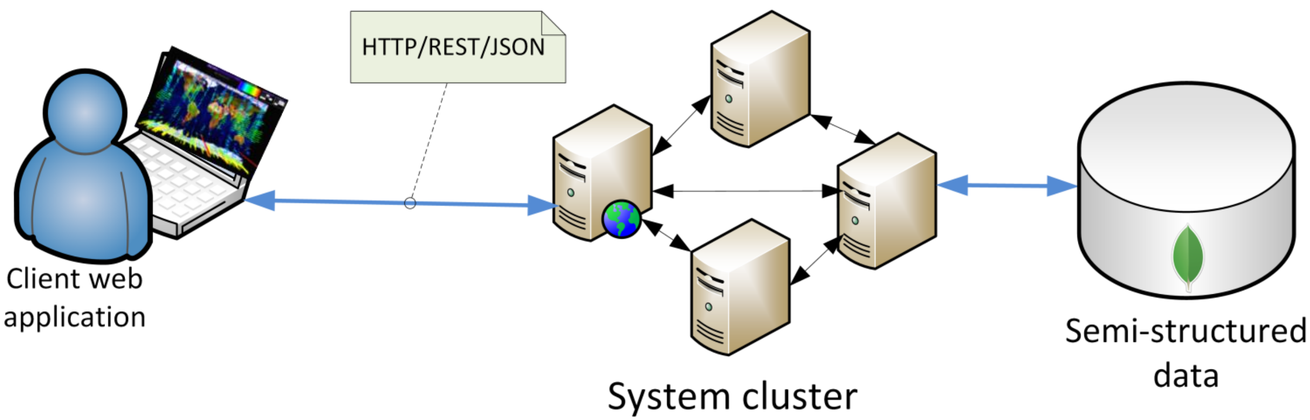
- DKRZ
  - Carsten Ehbrecht
  - Stephan Kindermann
  - Michael Lautenschlager

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

  - Ahmad Maatouki (conference presentation in August)
  - Intern: Cannon Kalra (Feb. – Jul.)
  - Students: Jiang Zhong Bo, Haipeng Guan, Florian Klemme

Climate Analysis with MEAN Stack

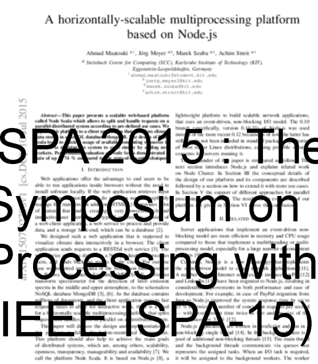
KAGLVis **Node Scala** **MongoDB**




A Distributed MEAN Stack-based System for Storing, Visualising and Analysing Data from Earth-observing Satellites

Marek Szuba, Steinbuch Centre for Computing (SCC)
Udo Grabowski, Institute of Meteorology and Climate Research (IMK)
Jörg Meyer, Steinbuch Centre for Computing (SCC)
Achim Streit, Steinbuch Centre for Computing (SCC)
Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

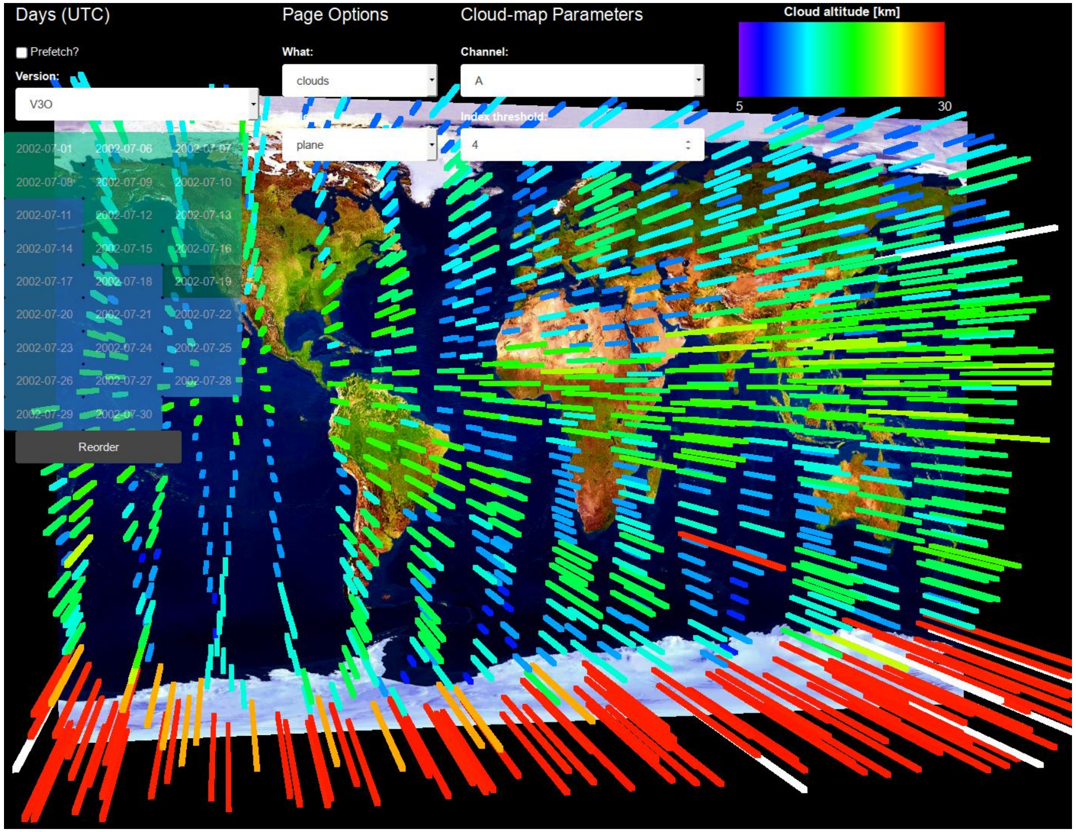
4th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (submitted)


A horizontally-scalable multiprocessing platform based on Node.js

ISPA 2015 : The 13th IEEE International Symposium on Parallel and Distributed Processing with Applications (IEEE ISPA-15)

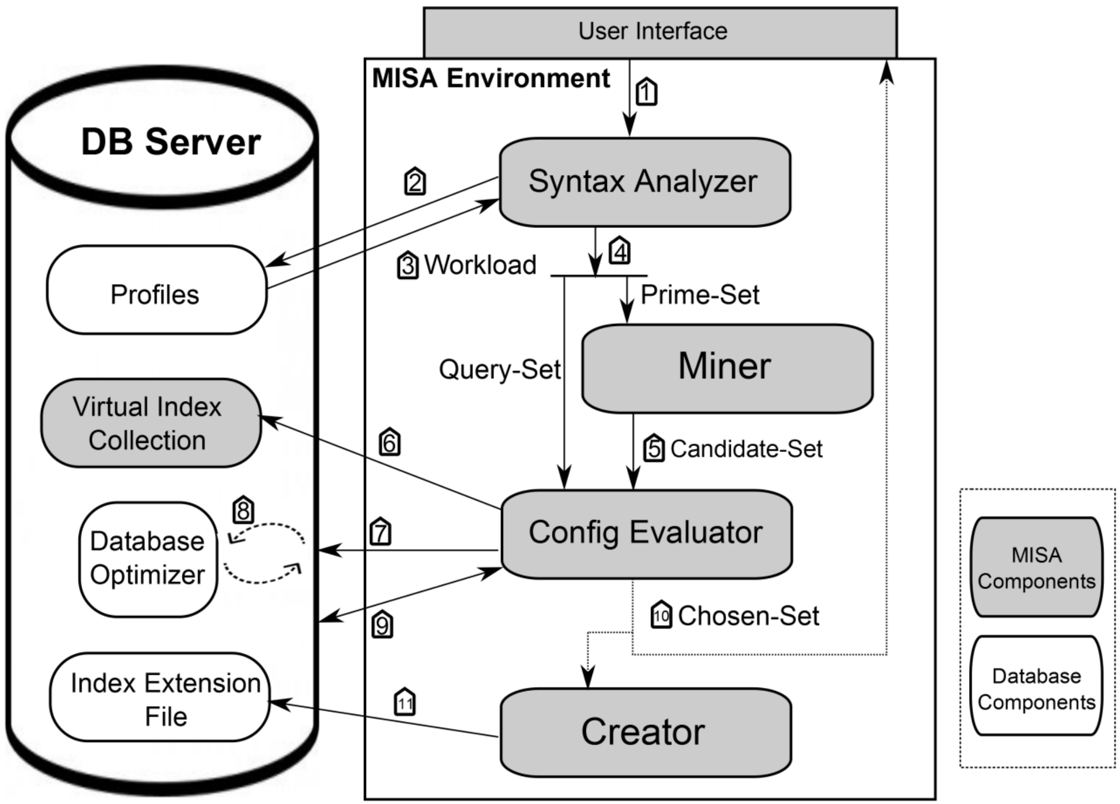
Real-time 3D Visualization of Earth-observing-satellite Data

- Visualization of climate data in a Web browser
- Cross-platform, including mobile devices
- Access to input data from MongoDB (via scalable Node.js cluster and REST API)
- Uses WebGL, AngularJS, Twitter Bootstrap
- Presented at the European Geoscience Union General Assembly 2015
- Paper submitted to BigSpatial2015



Mining Index Selection Approach

- Automatic index recommendation and management of db
- Dynamic adoption to workload changes
- Utilized on climate data as real-world use-case



On a New Approach to the Index Selection Problem using Mining Algorithms

Parinaz Ameri*, Jörg Meyer* and Achim Streit*
*Karlsruhe Institute of Technology (KIT), Steinbuch Centre for Computing (SCC), Eggenstein-Leopoldshafen, Germany, Emails: parinaz.ameri@kit.edu, joerg.meyer2@kit.edu, achim.streit@kit.edu

***Abstract*—Considering the wide usage of databases and their ever growing size, it is crucial to improve the query processing performance. Selection of an appropriate set of indexes for the workload processed by the database system is an important part of physical design and performance tuning. This selection is a non-trivial tasks, especially considering possible number of native indexes in modern databases.**

We introduce a new approach to the index selection problem using data mining. The method recommends the creation of indexes as well as the type of each index. This results in more precise index recommendations that allows not only to create ascending and descending indexes, but also special indexes supported by the database system. Mining of queries results in candidate indexes for which virtual indexes get created. As the approach does not require modifications of the database system, it is generically applicable. Evaluations of the scalability are given for different workloads for the document-based NoSQL database MongoDB.

***Keywords*-index selection; index type; frequent itemset; mining; performance tuning; query optimizer; MongodB; index recommendation; NoSQL database;**

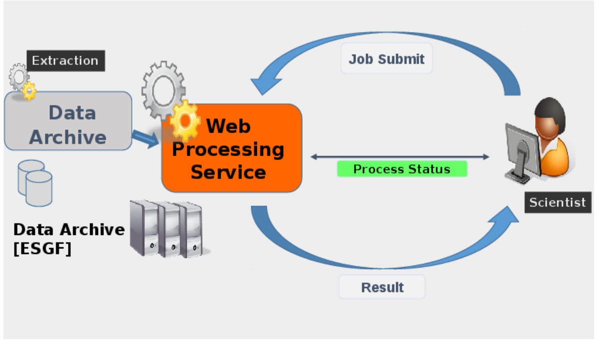
named Mining Index Selection Approach (MISA). Considering an increasing number of native specialized indexes that nowadays are provided by each database, it is very important to take into account not only the fields to be indexes, but also the type of the corresponding indexes. The novelty of our approach is to consider recommendation of the type of index as important as selecting index fields themselves. Index type selection is done by analyzing special operators that indicates the need for special corresponding indexes.

We utilize the query optimizer of the database itself to select best index. Each database usually has its own special query optimizer (either cost-based or rule-based) that analyzes queries, generates one or more query plans and eventually based on its criteria, determines the most efficient query execution mechanism [4]. We also take advantage of frequent itemsets mining algorithms [5] to build all combinations of fields for each frequent query.

IEEE Big Data conference
Workshop on Data-Centric Infrastructure for Big Data Science (accepted)

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:
  - Quality Assurance Tools (DKRZ) as WPS process:
    - checks of NetCDF files for compliance to the CF standard.
    - project specific checks for CORDEX, CMIP5, ...
  - LDAP Support in Phoenix web client (implemented by KIT).
  - Using Travis Continuous Integration for all Birdhouse components.
  - Data Access:
    - NetCDF files from Thredds catalogs.
    - Birdhouse Solr Index for Thredds catalogs and local files.
  - Deployment: automatic builds of Docker images on Docker Hub.

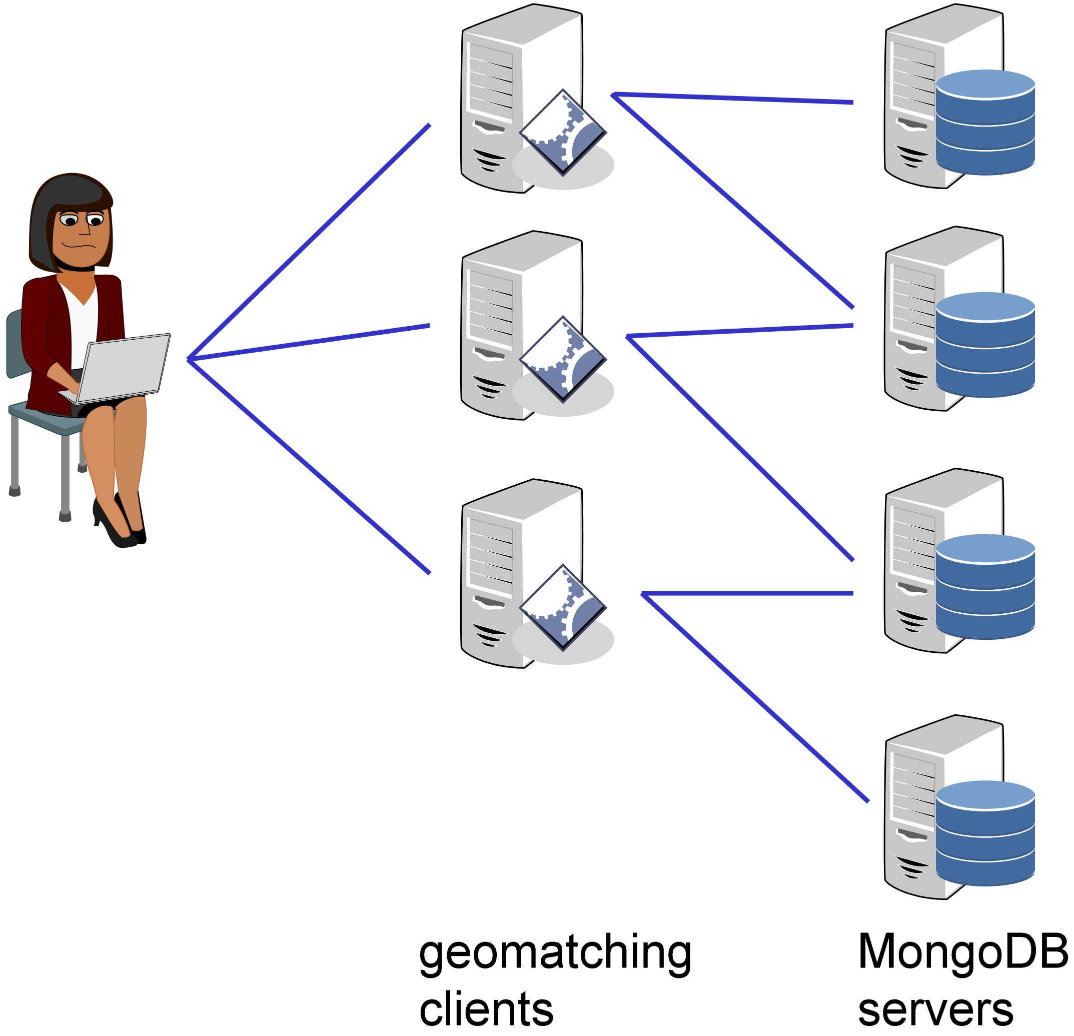
Data Management for Climate Research

PhD thesis on data management in climate research (SCC+IMK)

- Data discovery
  - Meta data catalogue
  - Meta data quality
- Dynamic transformation of data
  - Interpolation of gridded data
  - Conversion of formats
- Automation of workflows

Distributed Geomatching

- Matching of geo-coordinates and time
- New distributed architecture
  - CPU-bound → add clients
  - I/O-bound → add DB servers
- Added meta data for more instrument versions



Services for Climate Research

LSDMA

- GLORIA
  - MongoDB infrastructure on LSDF (7TB)
  - campaign will start in spring 2015
  - replication/redundancy required

- Satellite data
  - MongoDB with metadata (geolocations) of 22 instruments
  - improved geo-matcher

- EUDAT B2SAFE
  - Safe replication of ENES data
  - iRODs + PIDs (EPIC-handles)

B2SAFE
Replicate Research Data Safely

EUDAT2020

- KIT
  - Scientific communities environments and requirements
    - survey on data and computing landscapes, environments, and service requirements
  - B2SAFE (iRODS + PIDs)
    - New federations being created
      - GFZ Potsdam (seismology)
      - Institut für Anatomie Leipzig (medical data)
- DKRZ
  - B2FIND: meta data catalogue for research data

Proposals

- State of Baden-Württemberg: Virtual Research Environment
- BMBF: Establishment and development of innovative R&D networks with partners in the Danube States

Ongoing Projects / Services

- MongoDB for GLORIA project
- B2SAFE for ENES data