Project: Big Data

Julian M. Kunkel, Eugen Betke, Jakob Lüttgau,
Tobias Finn, Andrej Fast, Heinrich Widmann

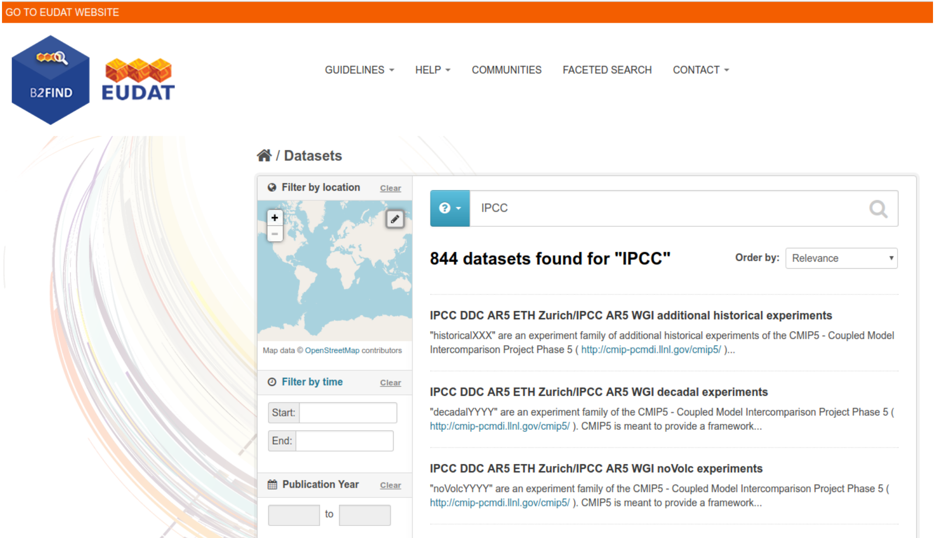
German Climate Computing Center (DKRZ)

2017-10-16

Semantic Search with Apache Solr

Motivation

Data services at DKRZ provide services to search through (dirty) metadata



844 datasets found for "IPCC"

IPCC DDC AR5 ETH Zurich/IPCC AR5 WGI additional historical experiments

"historicalXXX" are an experiment family of additional historical experiments of the CMIP5 - Coupled Model Intercomparison Project Phase 5 (http://cmip-pcmdi.llnl.gov/cmip5/)...

IPCC DDC AR5 ETH Zurich/IPCC AR5 WGI decadal experiments

"decadalYYYY" are an experiment family of the CMIP5 - Coupled Model Intercomparison Project Phase 5 (http://cmip-pcmdi.llnl.gov/cmip5/). CMIP5 is meant to provide a framework...

IPCC DDC AR5 ETH Zurich/IPCC AR5 WGI noVolc experiments

"noVolcYYYY" are an experiment family of the CMIP5 - Coupled Model Intercomparison Project Phase 5 (http://cmip-pcmdi.llnl.gov/cmip5/). CMIP5 is meant to provide a framework...

Semantic Search with Apache Solr

Goals

- Identify strategies to efficiently search in scientific metadata
- Optimize explorability / searchability for users

Tools/Knowledge

- Apache Solr; http://lucene.apache.org/solr/
- Linux command line
- Web page development (JavaScript)
- Text mining

Semantic Search with Apache Solr

Methodology

- Setup of Apache Solr (in a Virtual Machine)
- Integrating sample text (e.g., a database from DKRZ)
  - Identify best practices for schemas
- Investigating existing features
  - Extensions like Carrot2, user query statistics, cloud tags
- Developing a Webpage demonstrating the features
- Apply more detailed machine learning on the data:
  - Develop facets to quickly navigate data based on keywords
  - Collaboration with exploration of news team

Supervisors

- AnAndrej Fast Heinrich Widmanndrej Fast
- Heinrich Widmann
- Julian Kunkel

Apache Flink: Performance Analysis

Goals

- Explore the usage of Apache Flink for processing of big data workloads
- Utilization/development of benchmarks for understanding performance

Tools/Knowledge

- Apache Flink
- Linux command line
- Java
- Python
- Methodical performance analysis

Apache Flink: Performance Analysis

Methodology

- Installation/Setup of Apache Flink (in a Virtual Machine)
- Documentation of setup, will be installed then on the cluster...
- Exploring existing benchmarks / applications for benchmarking
- Systematic analysis of CPU, network, and I/O performance
- Compare performance of Java / Python

Supervisors

- Julian
- Eugen

Scientific Data Processing with Ophidia

Goals

- Explore the usage of Ophidia for processing of scientific BD workloads
- Utilization/development of benchmarks for understanding performance

Tools/Knowledge

- Ophidia https://github.com/OphidiaBigData/ophidia-analytics-framework
- Linux command line
- Parallel programming with MPI (a primer)
- Methodical performance analysis

Scientific Data Processing with Ophidia

Methodology

- Installation/Setup of Ophidia (in a Virtual Machine)
- Documentation of setup, will be installed then on our cluster...
- Exploring existing benchmarks / applications for benchmarking
- Systematic analysis of CPU, network, and I/O performance

Supervisors

- Julian
- Jakob

Workflow Processing Engines

Goals

- Understand pro/cons of existing workflows
- Understand usage of workflow engines

Tools/Knowledge

- Workflow processing, directed-acyclic-graphs
- Cylc: https://cylc.github.io/cylc/
- Camunda + Business Process Model and Notation 2.0: https://camunda.org
- Common Workflow Language for HPC: http://www.commonwl.org/
- Linux Command Line
- https://en.wikipedia.org/wiki/Workflow_engine

Workflow Processing Engines

Methodology

- Understanding the paradigm
- Setup of (the) tool/s (in a VM)
- Playing with workflows
- Some performance considerations / analysis

Supervisors

- Jakob Lüttgau
- Julian

Game AI

Goals

- Create an AI bot for RTS using genetic algorithms to train it

Tools/Knowledge

- Spring RTS https://springrts.com/
- C/C++
- AI strategies https://springrts.com/wiki/AI:Skirmish:List
- Reinforcement learning
- Genetic algorithms

Game AI

Methodology

- Setup Spring RTS
- Program an AI that can play again another AI
- Create a genetic algorithm to evolve the AI

Supervisors

- Julian
- Eugen Betke

Suicide Prevention

Goals

- Prediction of suicide rates based on news feeds

Tools/Knowledge

- Machine learning
- Python

Suicide Prevention

Methodology

- Explore news and model potential depressing occurrences
- We will request suicide rates from Hamburg
- Investigate temporal correlation between news and news...

Supervisors

- Julian
- This topic is also a collaboration between the UKE and UHAM

Exploration of News

Goals

- Investigate characteristics of online news
  - Changes between reposts of similar articles, time between reposting
  - Length of articles
  - Sentiment of articles
  - Differences between news’ sites/agencies

Tools/Knowledge

- Machine learning
- Python
- Text analysis methods

Exploration of News

Methodology

- Analyzing a subset of harvested data (CSV-format)
- Identify similar articles (copies) using text mining
- Create predictive models
- Clustering of results
- K-Cross validation of results
- Collaboration with the team Semantic Search (Solr)
- See the thesis of Jan Bilek...

Supervisors

- Julian
- Eugen

Chat Bot Guiding Users Support

Goals

- Building a chat bot that helps users to find the answer to typical question
- Support for data centers – embedded in the project PeCoH

Tools/Knowledge

- Web crawling, data ingestion
- Python
- Chat bot (e.g., https://github.com/gunthercox/ChatterBot)

Chat Bot Guiding Users Support

Methodology

- Developing a webpage Crawler and Indexer for data center pages
- Building an index
- Extracting features from the index
- Deriving chat responses to direct users to the index

Supervisors

- Julian Kunkel
- Andrej Fast

Analysis of the Linux Kernel Overhead

Goals

- Identify performance characteristics particularly for I/O
  - System call execution time
  - Mode change (user-space to OS) time
  - Block I/O overhead
  - File system overhead

Tools/Knowledge

- Kernel development
- Performance analysis tools
- C-Programming

Analysis of the Linux Kernel Overhead

Methodology

- Systemtap to instrument the kernel
- Oprofile
- Write a dummy block device/file system kernel driver
- See: http://www.brendangregg.com/linuxperf.html

Supervisors

- Jakob Lüttgau
- Julian

Efficient Management of Scientific Metadata

Goals

- Efficient storage and search of scientific metadata from applications
  - Experiment, date, model used, region, tags, ...
- Development of a configurable metadata view
- The results of this project are used for the ESiWACE middleware

Tools/Knowledge

- MongoDB
- FUSE
- Linux Command Line
- C-Programming

Efficient Management of Scientific Metadata

Methodology

- Setup of a VM with Linux and MongoDB
- Define a metadata schema
- Create dummy data/import into MongoDB
- Benchmarking of queries
- Development of the FUSE client
  - Define mapping of directory structure vs. metadata
  - Example: Experiment/Model/Date/Variable

Supervisors

- Jakob
- Julian

Analysis of Time Series Data

Goals

- Understanding and diagnosing behavior of time series data
- Use case: DKRZ system monitoring, what is I/O bound? Reasons?

Tools/Knowledge

- Grafana (visualization tool)
- OpenTSDB (time series database)
- Machine learning (simple)

Analysis of Time Series Data

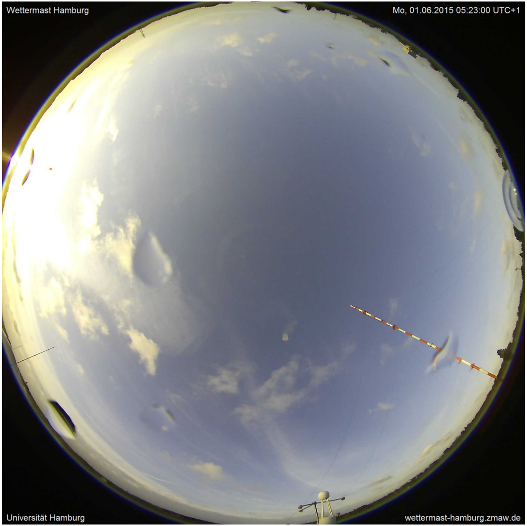
Methodology

- Setup of a Grafana / OpenTSDB system
- Identify access methods for data and its performance
- Write scripts to automatically assess system behavior
- Deploy scripts on DKRZ system

Supervisors

- Eugen Betke
- Julian

Cloud camera processing with deep learning



Cloud camera processing with deep learning

Goals

- Image segmentation for clouds
- Cloud base height determination
- Complete image segmentation (incl. sun/rain drops)

Tools/Knowledge

- Python
- Tensorflow /(pytorch)
- Deep learning / Machine learning
- Computer vision

Cloud camera processing with deep learning

Methodology

- Preprocessing of cloud camera images
- Creation of different deep learning / neural network models
  - Unsupervised learning with generative adversarial networks
  - Supervised learning with point measurement data
    (Usage of transfer learning)
  - Possible usage of recurrent neural networks
- Training / Finetuning of the models on GPUs
- Testing of different models

Supervisors

- Tobias Finn
- Julian