Karlsruher Institut für Technologie

DATA CURATION IN KCDC

V. Tokareva[*], A. Haungs, D.Kang, F. Polgart, D. Wochele, J. Wochele

[*]victoria.tokareva@kit.edu

INSTITUT FÜR ASTROTEILCHENPHYSIK (IAP)

2 May 2022, TA4 WP2 meeting

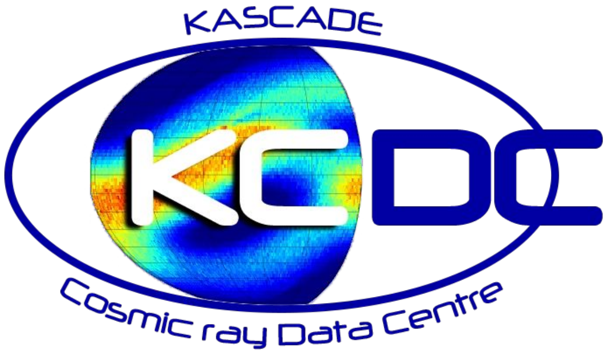
Content

- KCDC overview
- Data provided by KCDC
- Software architecture, data and metadata flows
- Jupyter Notebooks for data analysis

KASCADE Cosmic-ray Data Center (KCDC)

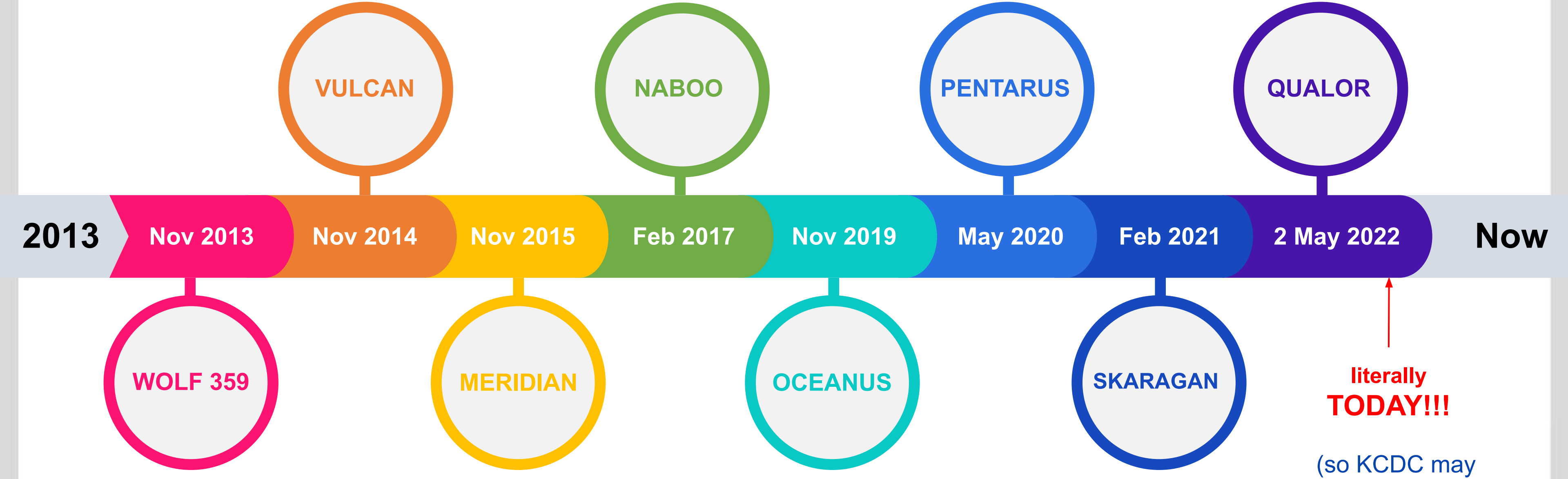
Karlsruher Institut für Technologie

- First released in 2013
- Aimed to provide free, unlimited, reliable open access to the data of various experiments measuring cosmic radiation by different methods and techniques both for scientists and the broad public
- Functions:
  - Data archive
  - Data analysis platform
  - Information center
  - Outreach platform
- Features:
  - Open data access
  - Allows custom data cuts
  - Ensures analysis reproducibility
  - Built on modern open source web technologies



https://kcdc.iap.kit.edu/

KCDC timeline



2013

Nov 2013 – WOLF 359

Nov 2014 – VULCAN

Nov 2015 – MERIDIAN

Feb 2017 – NABOO

Nov 2019 – OCEANUS

May 2020 – PENTARUS

Feb 2021 – SKARAGAN

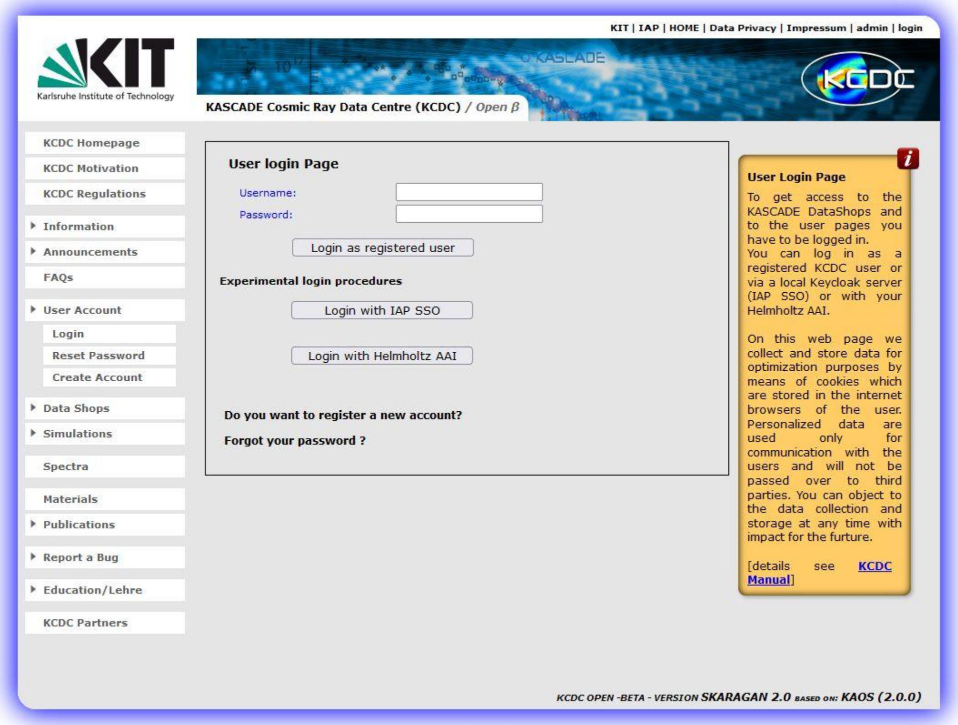
2 May 2022 – QUALOR

Now

literally
TODAY!!!

(so KCDC may
still down for
the moment)

What’s new in QUALOR?



KASCADE Cosmic Ray Data Centre (KCDC) / *Open β*

User login Page

Username:

Password:

Experimental login procedures

Do you want to register a new account?

Forgot your password ?

User Login Page

To get access to the KASCADE DataShops and to the user pages you have to be logged in.
You can log in as a registered KCDC user or via a local Keycloak server (IAP SSO) or with your Helmholtz AAI.

On this web page we collect and store data for optimization purposes by means of cookies which are stored in the internet browsers of the user. Personalized data are used only for communication with the users and will not be passed over to third parties. You can object to the data collection and storage at any time with impact for the furture.

[details see KCDC Manual]

KCDC OPEN -BETA - VERSION ***SKARAGAN 2.0*** *BASED ON:* ***KAOS (2.0.0)***

- Login via Helmholtz AAI and via a Keycloak server
- 4-shard database
- New QGSJet-II-04 Simulations

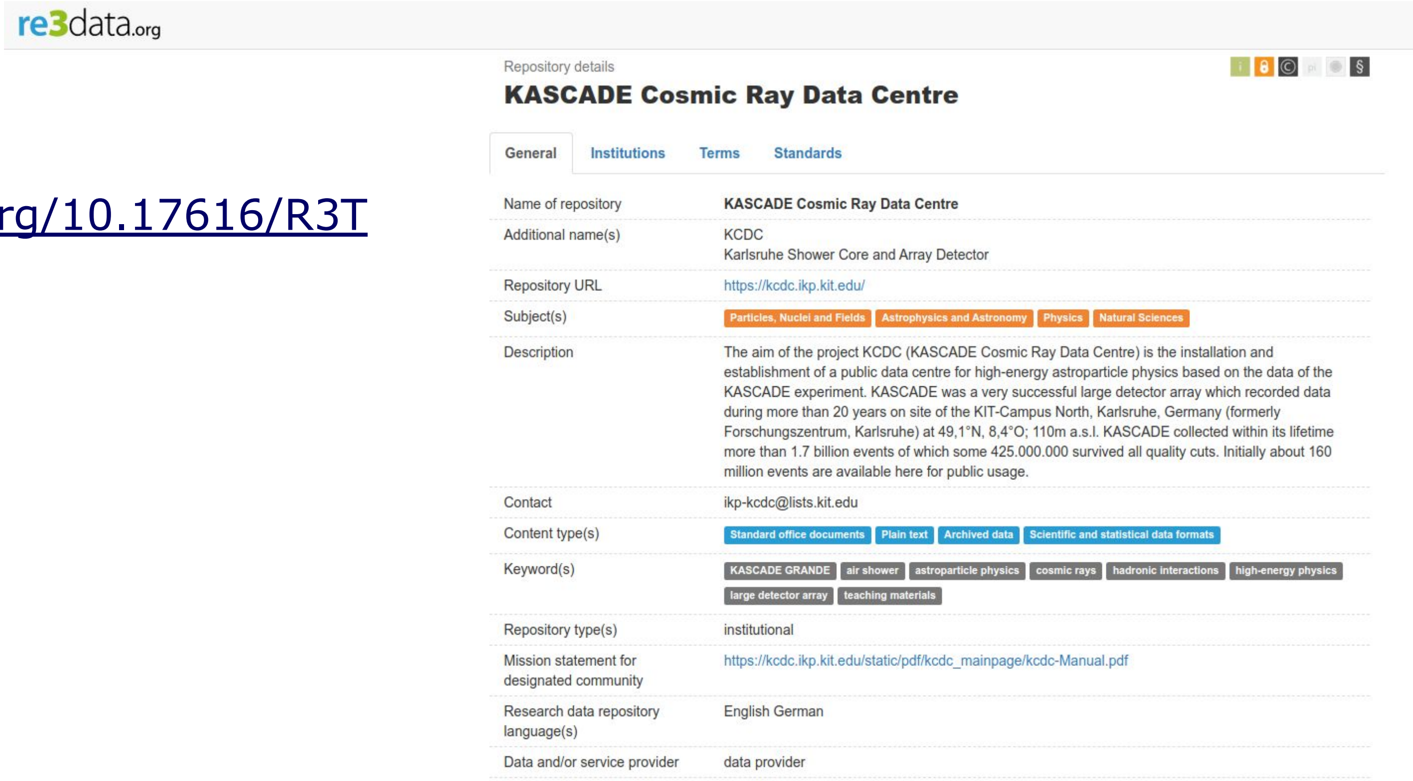
Data overview

Setup/ Detector component	Experimental data		Simulations	
	Events	Size	Events	Size
KASCADE	433 209 340	3 200 GB	21 388 000	25 GB
GRANDE	35 310 393	260 GB	3 545 000	4 GB
COMBINED	15 635 550	120 GB	1 590 000	2 GB
LOPES	3 058	25 MB	—	—
MAKET-ANI	2 682 264	1 GB	—	—

* This table doesn’t take into account the new QGSJet-II-04 simulations added in QUALOR

How published: DOI

doi.org/10.17616/R3TS4P



re3data.org

Repository details

KASCADE Cosmic Ray Data Centre

General | Institutions | Terms | Standards

Name of repository	**KASCADE Cosmic Ray Data Centre**
Additional name(s)	KCDC Karlsruhe Shower Core and Array Detector
Repository URL	https://kcdc.ikp.kit.edu/
Subject(s)	Particles, Nuclei and Fields · Astrophysics and Astronomy · Physics · Natural Sciences
Description	The aim of the project KCDC (KASCADE Cosmic Ray Data Centre) is the installation and establishment of a public data centre for high-energy astroparticle physics based on the data of the KASCADE experiment. KASCADE was a very successful large detector array which recorded data during more than 20 years on site of the KIT-Campus North, Karlsruhe, Germany (formerly Forschungszentrum, Karlsruhe) at 49,1°N, 8,4°O; 110m a.s.l. KASCADE collected within its lifetime more than 1.7 billion events of which some 425.000.000 survived all quality cuts. Initially about 160 million events are available here for public usage.
Contact	ikp-kcdc@lists.kit.edu
Content type(s)	Standard office documents · Plain text · Archived data · Scientific and statistical data formats
Keyword(s)	KASCADE GRANDE · air shower · astroparticle physics · cosmic rays · hadronic interactions · high-energy physics · large detector array · teaching materials
Repository type(s)	institutional
Mission statement for designated community	https://kcdc.ikp.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf
Research data repository language(s)	English German
Data and/or service provider	data provider

KCDC DataShops

The data sets are organised into so-called datashops:

- KASCADE - contains 'common data' and data from four detector components: KASCADE, GRANDE, CALORIMETER, LOPES
- COMBINED - includes 'common data', data from KASCADE and GRANDE detectors combined for joint analysis as well as data arrays from KASCADE and GRANDE and LOPES quntities
- Maket-Ani - provides quantities from the Maket-Ani setup

New data shops can be added.

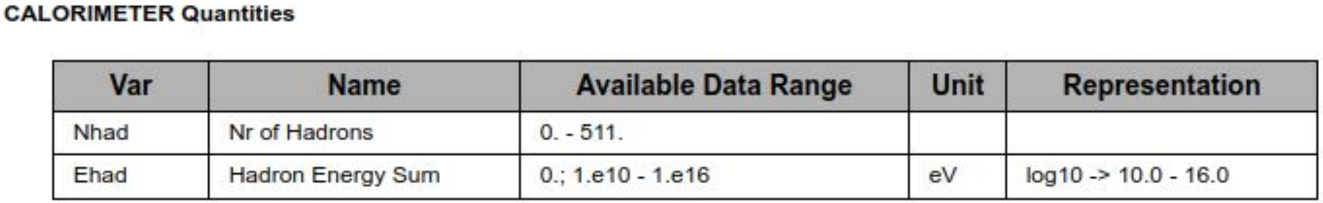
KCDC DataShops and data formats

They are supplied in the following file formats*:

- ASCII - plain text format
- ROOT - object oriented framework developed by CERN
- HDF5 - hierarchical data format

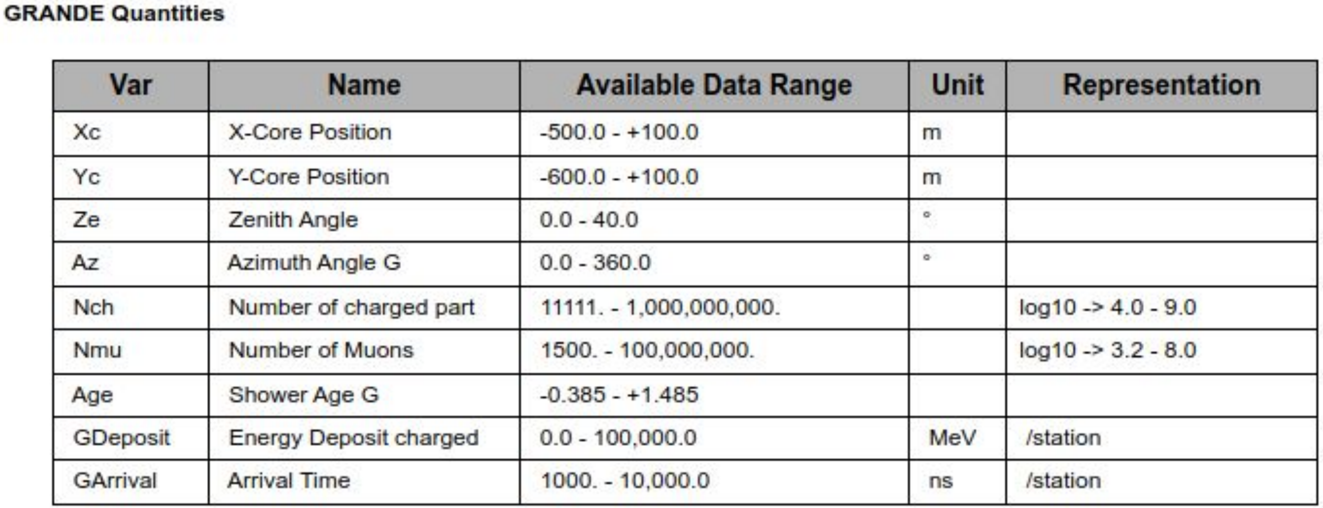
* Selectable by the user and depending on the quantities chosen

Data quantities examples



CALORIMETER Quantities

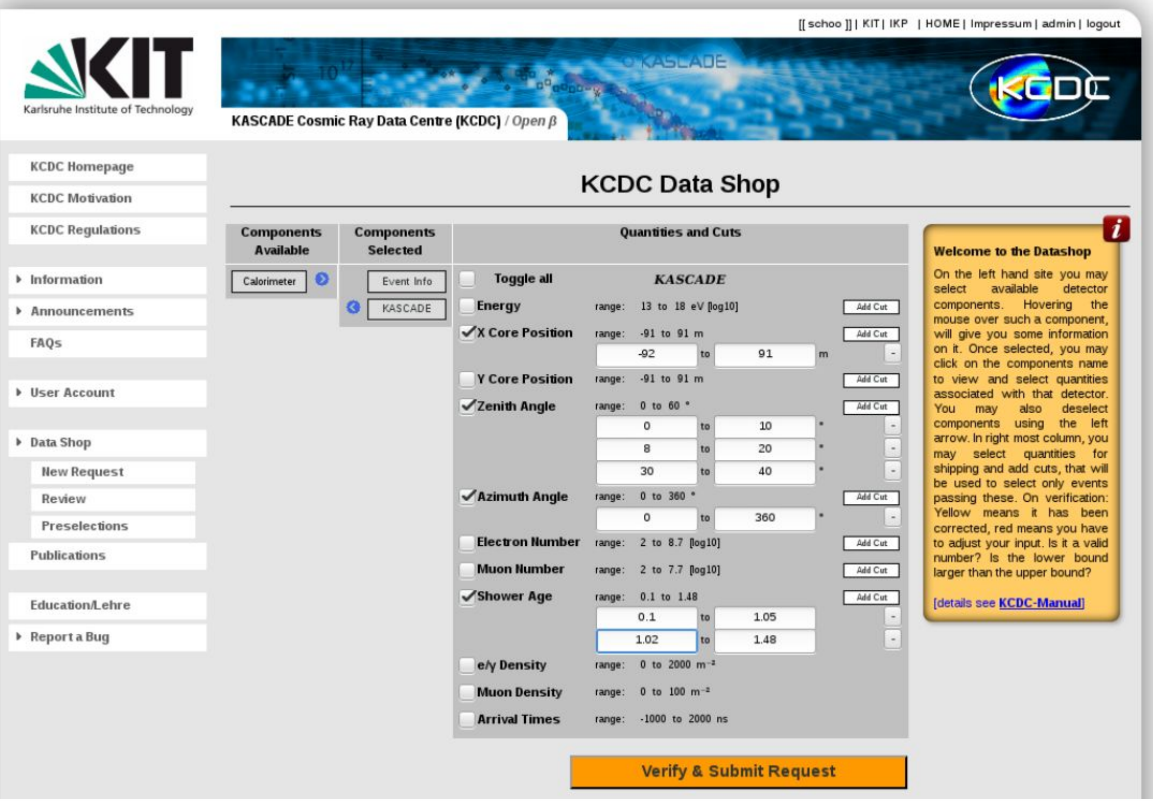
Var	Name	Available Data Range	Unit	Representation
Nhad	Nr of Hadrons	0. - 511.		
Ehad	Hadron Energy Sum	0.; 1.e10 - 1.e16	eV	log10 -> 10.0 - 16.0



GRANDE Quantities

Var	Name	Available Data Range	Unit	Representation
Xc	X-Core Position	-500.0 - +100.0	m	
Yc	Y-Core Position	-600.0 - +100.0	m	
Ze	Zenith Angle	0.0 - 40.0	°	
Az	Azimuth Angle G	0.0 - 360.0	°	
Nch	Number of charged part	11111. - 1,000,000,000.		log10 -> 4.0 - 9.0
Nmu	Number of Muons	1500. - 100,000,000.		log10 -> 3.2 - 8.0
Age	Shower Age G	-0.385 - +1.485		
GDeposit	Energy Deposit charged	0.0 - 100,000.0	MeV	/station
GArrival	Arrival Time	1000. - 10,000.0	ns	/station

Usage of KCDC datashop GUI



KCDC Data Shop

Quantities and Cuts

KASCADE

	Quantity	range			
□	Toggle all				
□	Energy	13 to 18 eV [log10]			
✓	X Core Position	-91 to 91 m	-92	to	91 m
□	Y Core Position	-91 to 91 m			
✓	Zenith Angle	0 to 60 °	0	to	10 °
			8	to	20 °
			30	to	40 °
✓	Azimuth Angle	0 to 360 °	0	to	360 °
□	Electron Number	2 to 8.7 [log10]			
□	Muon Number	2 to 7.7 [log10]			
✓	Shower Age	0.1 to 1.48	0.1	to	1.05
			1.02	to	1.48
□	e/γ Density	0 to 2000 m⁻²			
□	Muon Density	0 to 100 m⁻²			
□	Arrival Times	-1000 to 2000 ns			

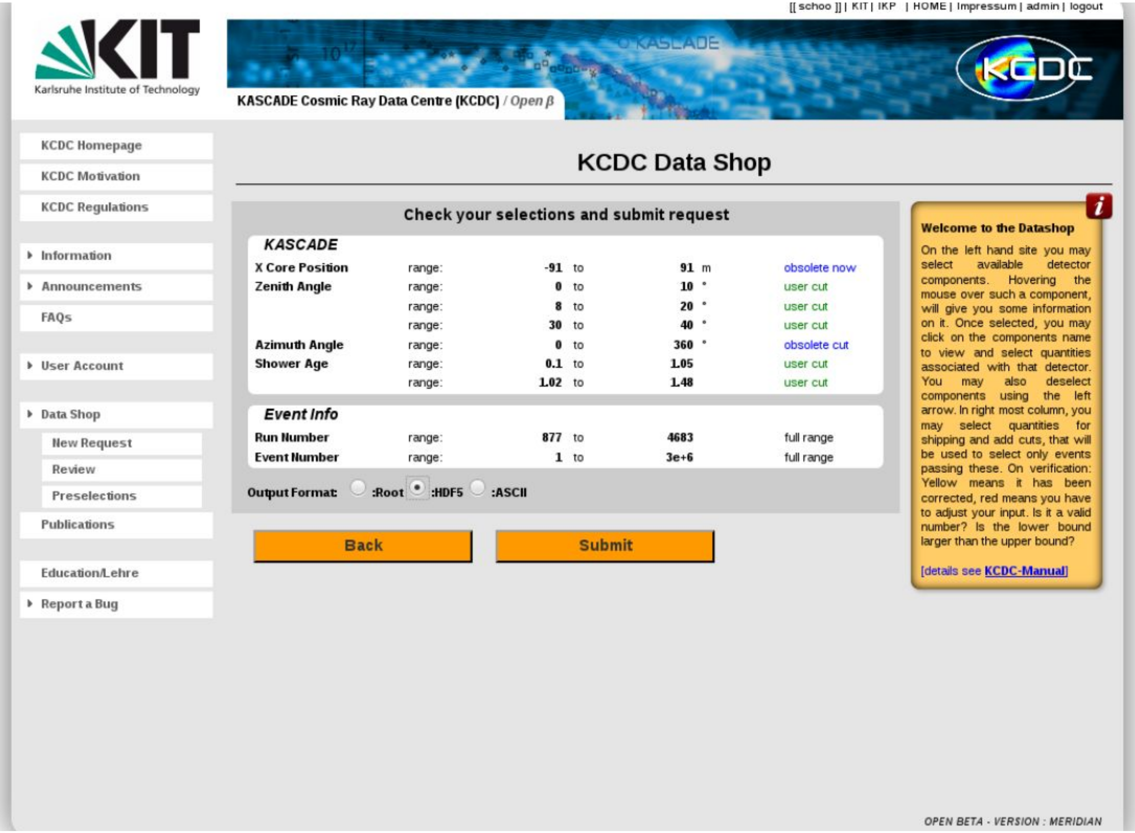
Welcome to the Datashop

On the left hand site you may select available detector components. Hovering the mouse over such a component, will give you some information on it. Once selected, you may click on the components name to view and select quantities associated with that detector. You may also deselect components using the left arrow. In right most column, you may select quantities for shipping and add cuts, that will be used to select only events passing these. On verification: Yellow means it has been corrected, red means you have to adjust your input. Is it a valid number? Is the lower bound larger than the upper bound?

[details see KCDC-Manual]

The entry page of the KCDC DataShop pages

Usage of KCDC datashop GUI



KCDC Data Shop

Check your selections and submit request

KASCADE

X Core Position	range:	**-91**	to	**91**	m	obsolete now
Zenith Angle	range:	**0**	to	**10**	°	user cut
	range:	**8**	to	**20**	°	user cut
	range:	**30**	to	**40**	°	user cut
Azimuth Angle	range:	**0**	to	**360**	°	obsolete cut
Shower Age	range:	**0.1**	to	**1.05**		user cut
	range:	**1.02**	to	**1.48**		user cut

Event Info

Run Number	range:	**877**	to	**4683**	full range
Event Number	range:	**1**	to	**3e+6**	full range

Output Format: ○ **:Root** ● **:HDF5** ○ **:ASCII**

Back **Submit**

Welcome to the Datashop

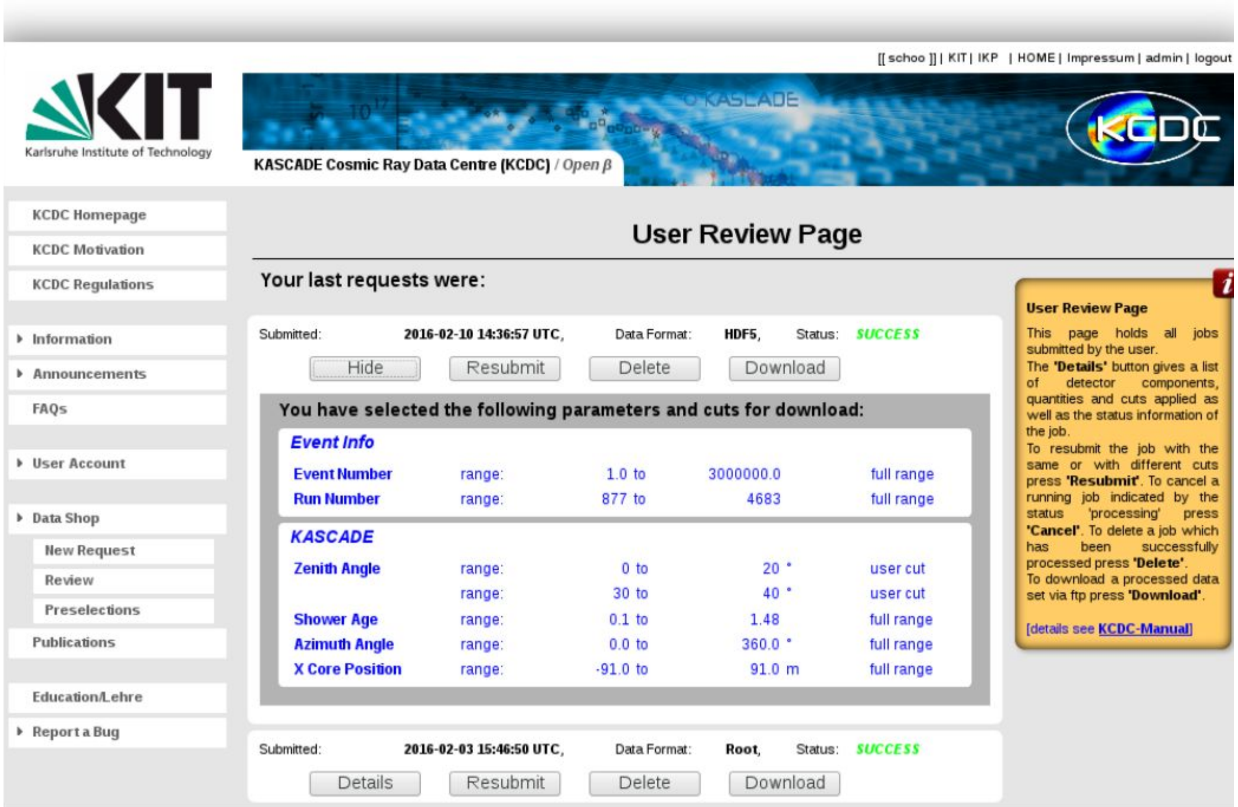
On the left hand site you may select available detector components. Hovering the mouse over such a component, will give you some information on it. Once selected, you may click on the components name to view and select quantities associated with that detector. You may also deselect components using the left arrow. In right most column, you may select quantities for shipping and add cuts, that will be used to select only events passing these. On verification: Yellow means it has been corrected, red means you have to adjust your input. Is it a valid number? Is the lower bound larger than the upper bound?

[details see KCDC-Manual]

OPEN BETA - VERSION : MERIDIAN

The confirmation page of the KCDC DataShop pages

Usage of KCDC datashop GUI



User Review Page

Your last requests were:

Submitted: **2016-02-10 14:36:57 UTC,** Data Format: **HDF5,** Status: ***SUCCESS***

You have selected the following parameters and cuts for download:

Event Info

Event Number	range:	1.0 to	3000000.0	full range
Run Number	range:	877 to	4683	full range

KASCADE

Zenith Angle	range:	0 to	20 °	user cut
	range:	30 to	40 °	user cut
Shower Age	range:	0.1 to	1.48	full range
Azimuth Angle	range:	0.0 to	360.0 °	full range
X Core Position	range:	-91.0 to	91.0 m	full range

Submitted: **2016-02-03 15:46:50 UTC,** Data Format: **Root,** Status: ***SUCCESS***

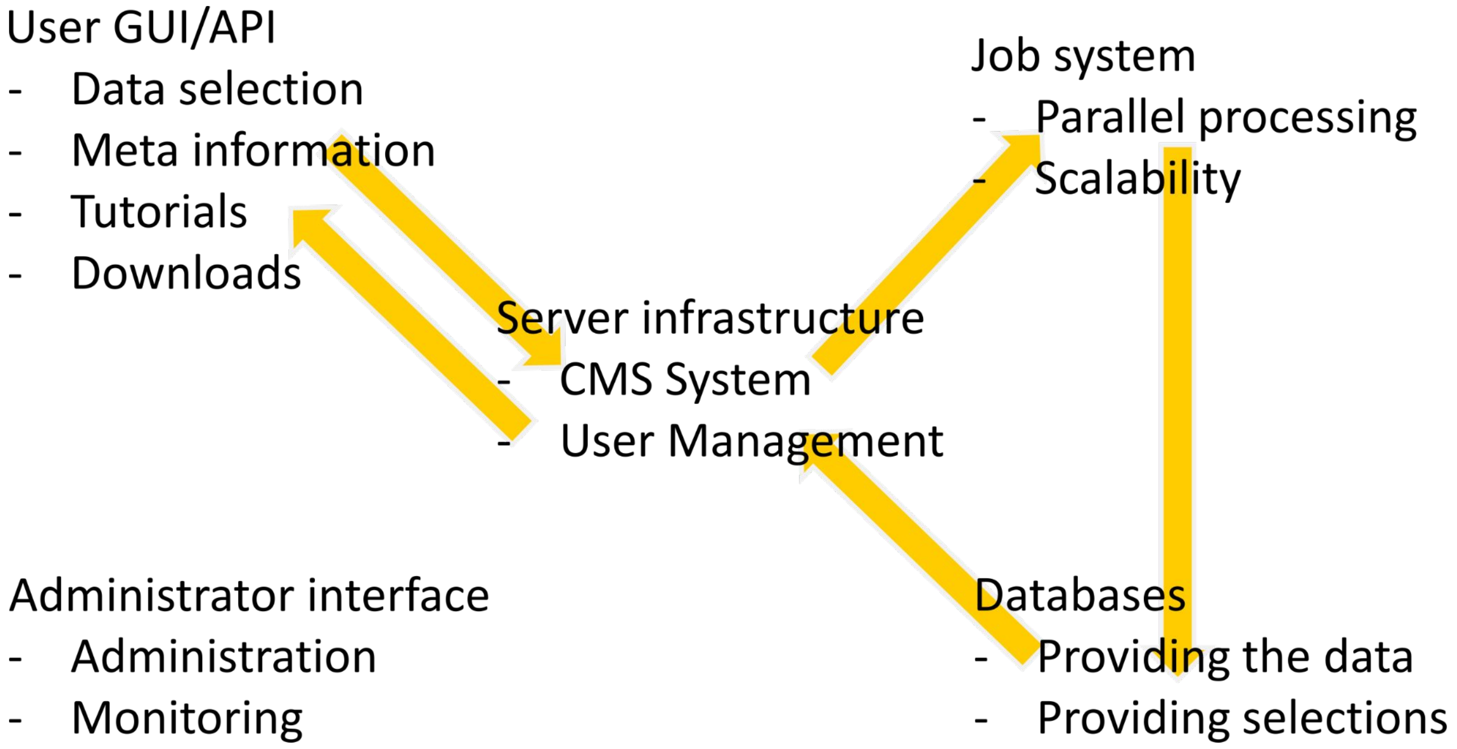
User Review Page

This page holds all jobs submitted by the user.
The **'Details'** button gives a list of detector components, quantities and cuts applied as well as the status information of the job.
To resubmit the job with the same or with different cuts press **'Resubmit'**. To cancel a running job indicated by the status 'processing' press **'Cancel'**. To delete a job which has been successfully processed press **'Delete'**.
To download a processed data set via ftp press **'Download'**.

[details see KCDC-Manual]

The review page of the KCDC DataShop pages

KCDC user’s job data workflow



User GUI/API

- Data selection
- Meta information
- Tutorials
- Downloads

Job system

- Parallel processing
- Scalability

Server infrastructure

- CMS System
- User Management

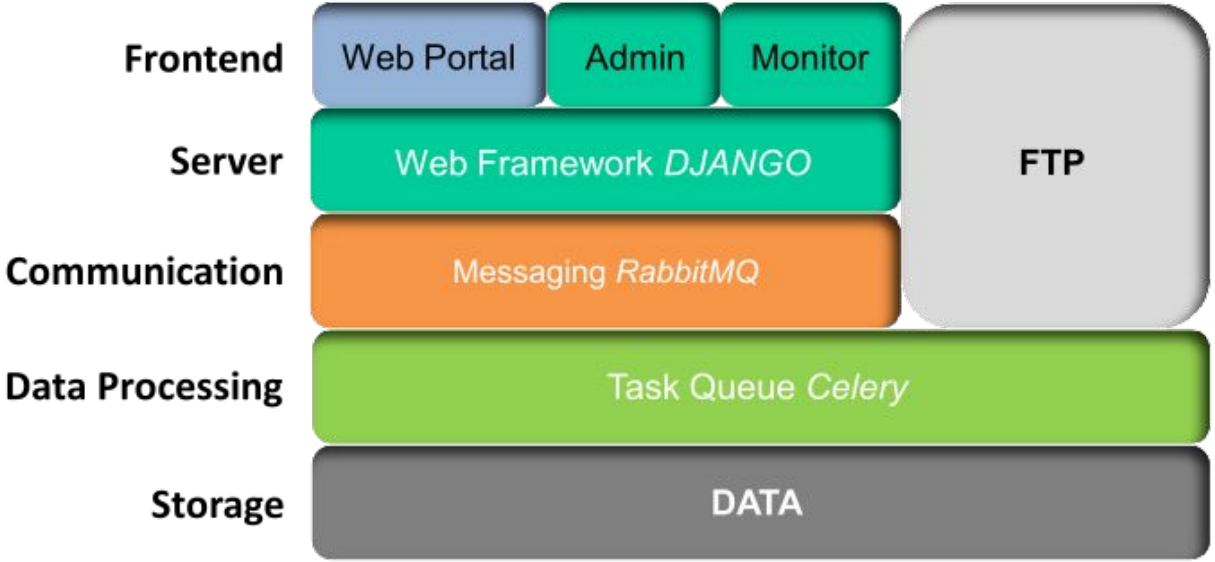
Administrator interface

- Administration
- Monitoring

Databases

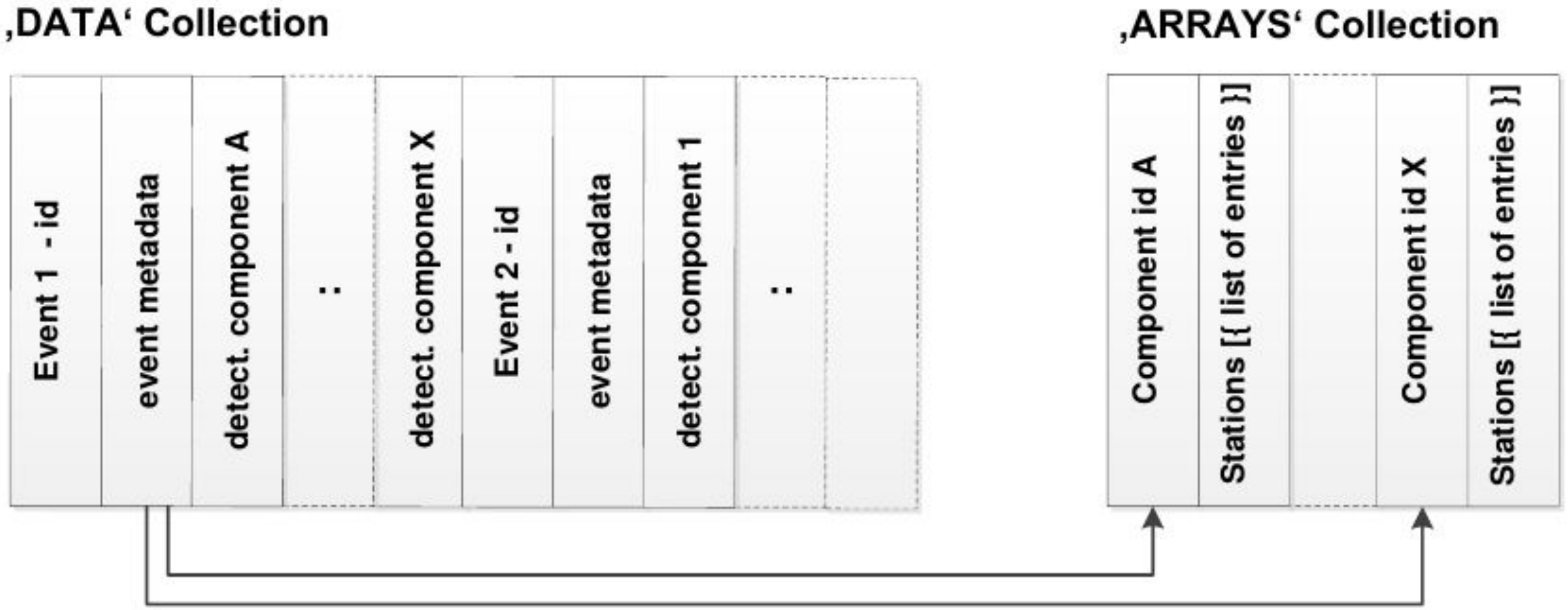
- Providing the data
- Providing selections

Architecture and technology stack



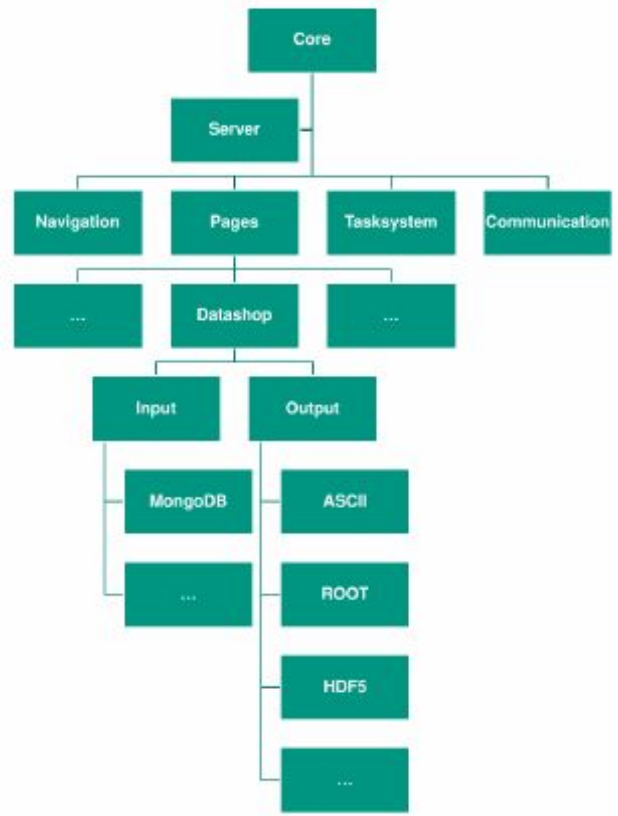
Database - **MongoDB**; **REST**full API (starting from SKARAGAN release), **JupyterLab** for data analysis

MongoDB data storage structure



Wochele, D., Wochele, J., Polgart, F., Tokareva, V., Kang, D., & Haungs, A. Data Structure Adaption from Large-Scale Experiment for Public Re-Use. CEUR-WS (2019) 2406, 114

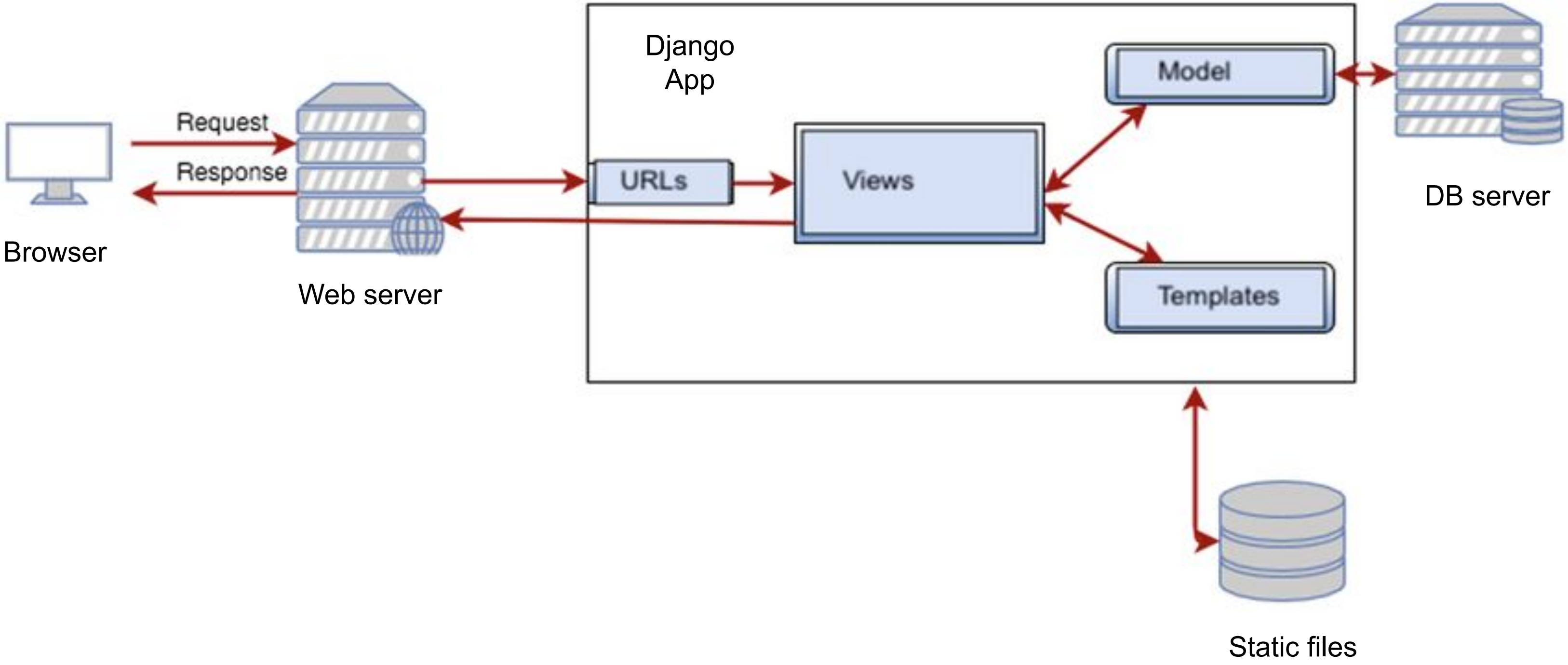
KAOS - KCDC’s backend



- **K**arlsruhe **A**stroparticlephysics **O**pen data **S**oftware (KAOS)
- Implemented using a plugin based design with a focus on easy extensibility and modifiability
- Can work as well outside the context of KCDC

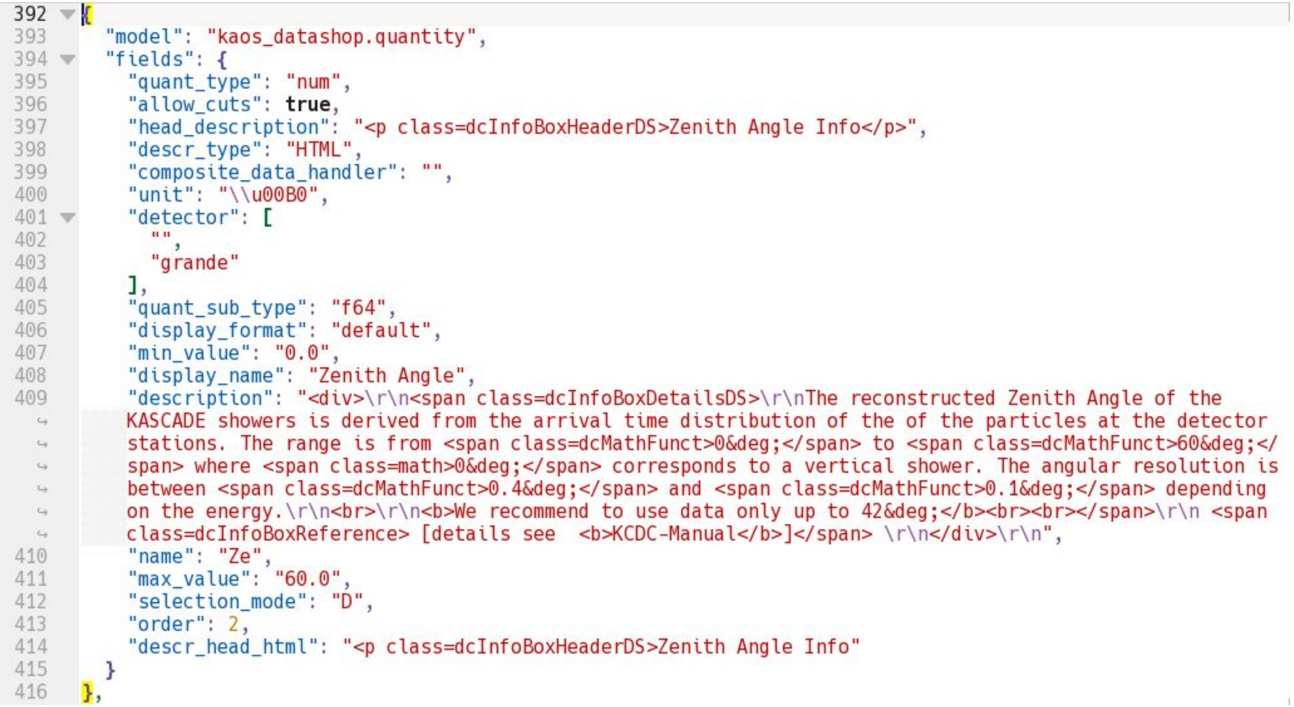
Schoo, S. Energy Spectrum and Mass Composition of Cosmic Rays and How to Publish Air-Shower Data. PhD thesis, 2016, link: https://publikationen.bibliothek.kit.edu/1000055797

Django application structure



Metadata on KCDC

JSON metadata schema, example of a record from KCDC

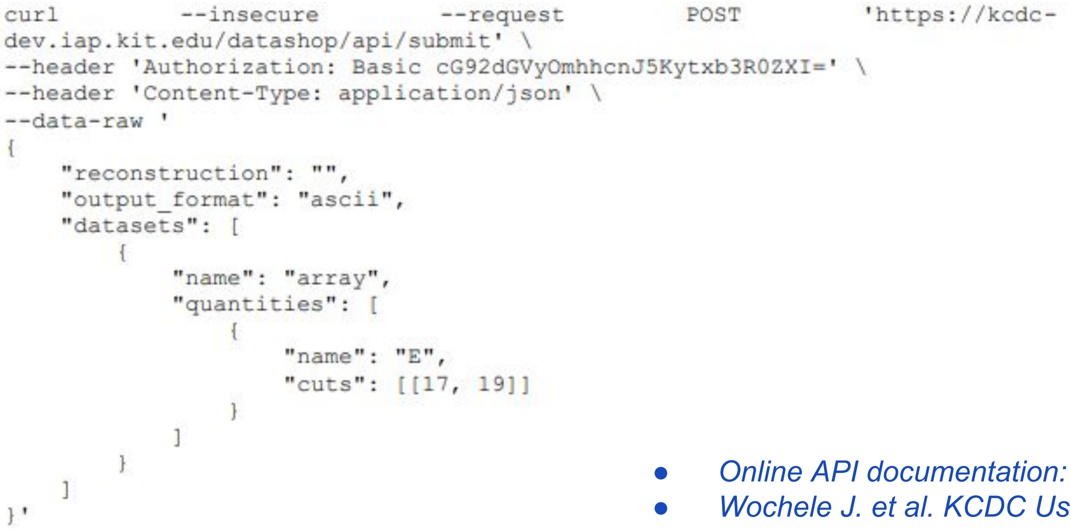


```
{
  "model": "kaos_datashop.quantity",
  "fields": {
    "quant_type": "num",
    "allow_cuts": true,
    "head_description": "<p class=dcInfoBoxHeaderDS>Zenith Angle Info</p>",
    "descr_type": "HTML",
    "composite_data_handler": "",
    "unit": "\\u00B0",
    "detector": [
      "",
      "grande"
    ],
    "quant_sub_type": "f64",
    "display_format": "default",
    "min_value": "0.0",
    "display_name": "Zenith Angle",
    "description": "<div>\r\n<span class=dcInfoBoxDetailsDS>\r\nThe reconstructed Zenith Angle of the
    KASCADE showers is derived from the arrival time distribution of the of the particles at the detector
    stations. The range is from <span class=dcMathFunct>0&deg;</span> to <span class=dcMathFunct>60&deg;</
    span> where <span class=math>0&deg;</span> corresponds to a vertical shower. The angular resolution is
    between <span class=dcMathFunct>0.4&deg;</span> and <span class=dcMathFunct>0.1&deg;</span> depending
    on the energy.\r\n<br>\r\n<b>We recommend to use data only up to 42&deg;</b><br><br></span>\r\n <span
    class=dcInfoBoxReference> [details see  <b>KCDC-Manual</b>]</span> \r\n</div>\r\n",
    "name": "Ze",
    "max_value": "60.0",
    "selection_mode": "D",
    "order": 2,
    "descr_head_html": "<p class=dcInfoBoxHeaderDS>Zenith Angle Info"
  }
},
```

KCDC APPLICATION PROGRAMMING INTERFACE (API)

Shell example: Extraction of the all data with an energy range from 17-19eV[log10]

Request:



```
curl          --insecure          --request          POST          'https://kcdc-
dev.iap.kit.edu/datashop/api/submit' \
--header 'Authorization: Basic cG92dGVyOmhhcnJ5Kytxb3R0ZXI=' \
--header 'Content-Type: application/json' \
--data-raw '
{
    "reconstruction": "",
    "output_format": "ascii",
    "datasets": [
        {
            "name": "array",
            "quantities": [
                {
                    "name": "E",
                    "cuts": [[17, 19]]
                }
            ]
        }
    ]
}'
```

Responce:

job id:

{"id":"dbf1e608b6044223afe472125c020d88"}

or error message:

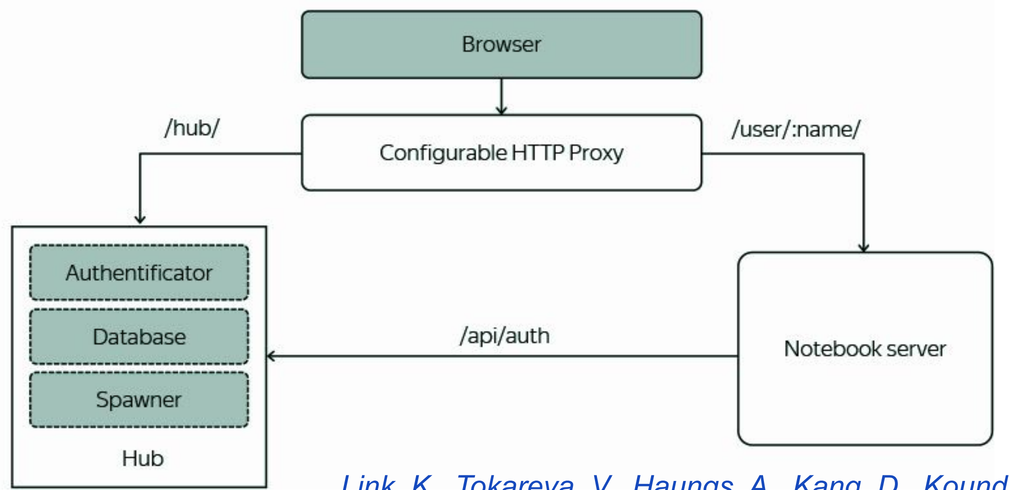
{"detail":"Invalid basic header. Credentials not correctly base64 encoded."}

- *Online API documentation: https://kcdc.iap.kit.edu/datashop/api/docs/index.html*
- *Wochele J. et al. KCDC User Manual: https://kcdc.iap.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf*

JupyterHub for data analysis

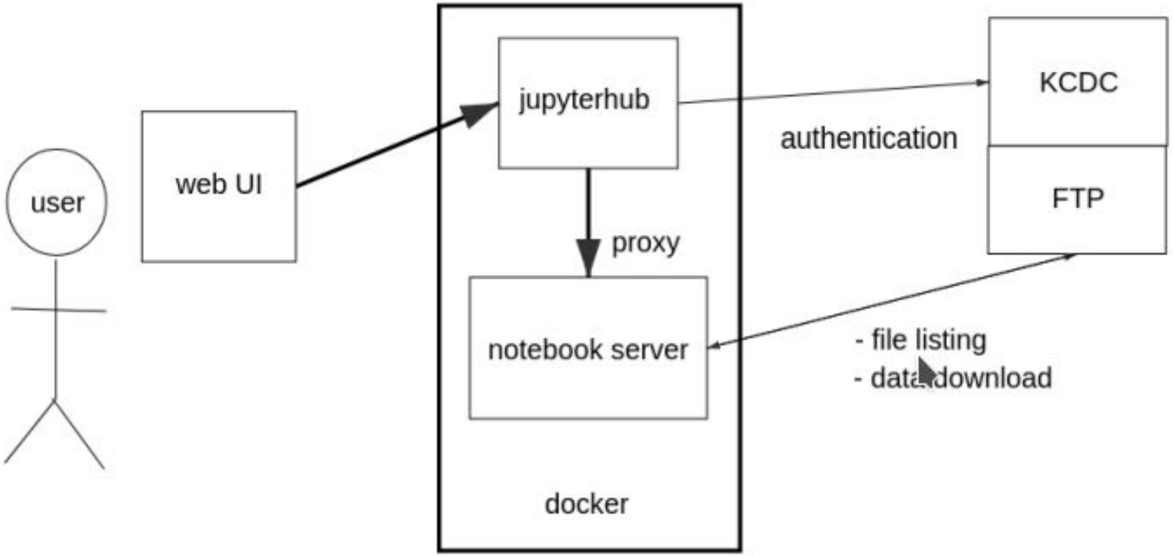
- Login via KCDC credentials
- Administration using Docker Swarm
- Tutorials by: KASCADE, IceCube, TRVO

Authorization	Run individual Jupyter servers
Proxy to Jupyter servers	System and database management



Link, K., Tokareva, V., Haungs, A., Kang, D., Koundal, P., Polgart, F., Tkachenko, O.,Wochele, D., Wochele, J. Online masterclass built on the KASCADE cosmic ray data centre. In 37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07.

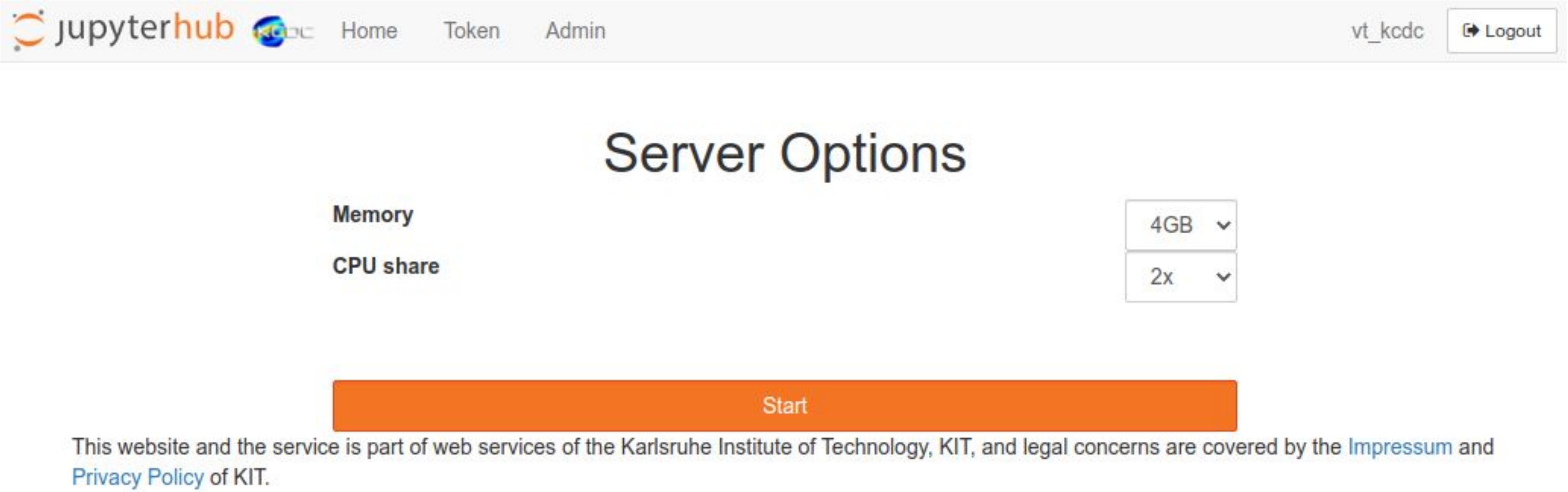
JupyterHub integration in KCDC



Polgart, F., Haungs, A., Kang, D., Wochele, D., Wochele, J., & Tokareva, V. (2020). An analysis framework for KCDC. In DLC 2020: Proceedings of the 4th International Workshop on Data Life Cycle in Physics. Ed.: A. Kryukov (p. 111).

Usage of KCDC’s JupyterHub

https://jupyter.iap.kit.edu/



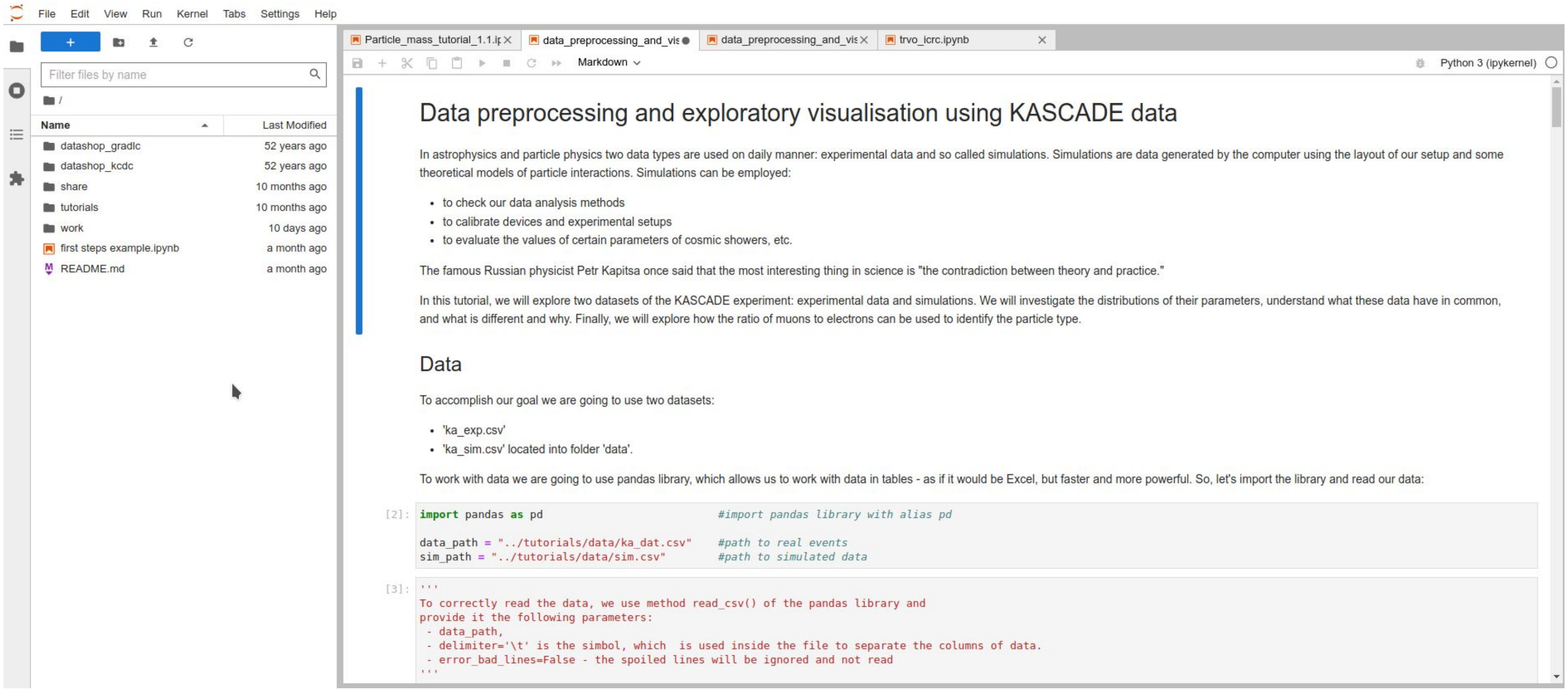
Server Options

Memory 4GB

CPU share 2x

This website and the service is part of web services of the Karlsruhe Institute of Technology, KIT, and legal concerns are covered by the Impressum and Privacy Policy of KIT.

Usage of KCDC’s JupyterHub



Name	Last Modified
datashop_gradlc	52 years ago
datashop_kcdc	52 years ago
share	10 months ago
tutorials	10 months ago
work	10 days ago
first steps example.ipynb	a month ago
README.md	a month ago

Data preprocessing and exploratory visualisation using KASCADE data

In astrophysics and particle physics two data types are used on daily manner: experimental data and so called simulations. Simulations are data generated by the computer using the layout of our setup and some theoretical models of particle interactions. Simulations can be employed:

- to check our data analysis methods
- to calibrate devices and experimental setups
- to evaluate the values of certain parameters of cosmic showers, etc.

The famous Russian physicist Petr Kapitsa once said that the most interesting thing in science is "the contradiction between theory and practice."

In this tutorial, we will explore two datasets of the KASCADE experiment: experimental data and simulations. We will investigate the distributions of their parameters, understand what these data have in common, and what is different and why. Finally, we will explore how the ratio of muons to electrons can be used to identify the particle type.

Data

To accomplish our goal we are going to use two datasets:

- 'ka_exp.csv'
- 'ka_sim.csv' located into folder 'data'.

To work with data we are going to use pandas library, which allows us to work with data in tables - as if it would be Excel, but faster and more powerful. So, let's import the library and read our data:

```
[2]: import pandas as pd                              #import pandas library with alias pd

     data_path = "../tutorials/data/ka_dat.csv"       #path to real events
     sim_path = "../tutorials/data/sim.csv"           #path to simulated data
```

```
[3]: '''
     To correctly read the data, we use method read_csv() of the pandas library and
     provide it the following parameters:
      - data_path,
      - delimiter='\t' is the simbol, which  is used inside the file to separate the columns of data.
      - error_bad_lines=False - the spoiled lines will be ignored and not read
     '''
```

Thank you for your attention!

victoria.tokareva@kit.edu, iap-kcdc@lists.kit.edu

Back up

GRADLCI and KCDC

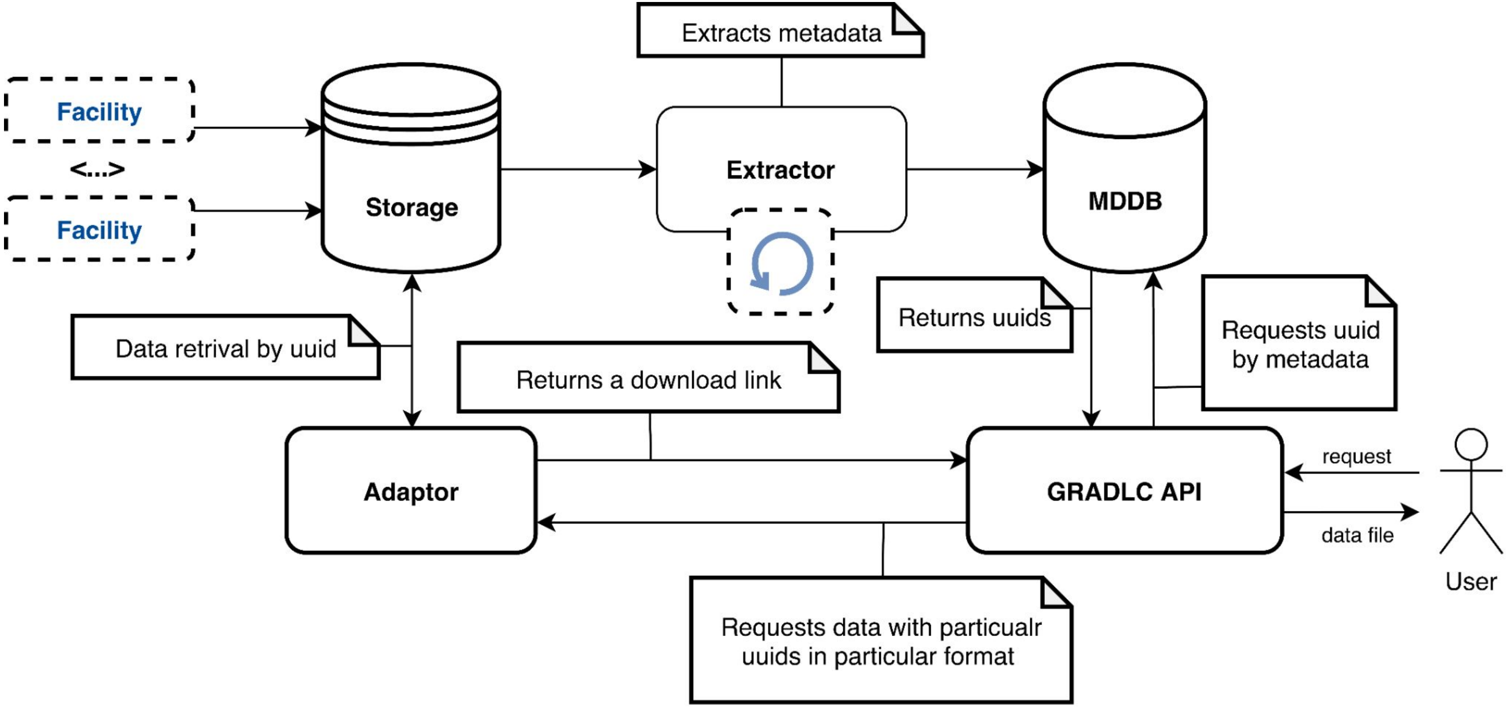
German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI)

- The international initiative aiming at automatisation the maintenance of astroparticle-physics data throughout their entire life cycle
- 2018 - 2021
- Task areas:
  - KCDC extension
  - Prototype analysis and data center
  - Machine learning for astroparticle physics
  - Outreach and education
- Aggregated data by KCDC, Tunka-133, TAIGA and Tunka-Rex Virtual Observatory (TrVO)
- Data throughput: 4.5 TB
- Features:
  - Metadata database (MDDB) as SQL DB
  - 2 level metadata model:

    (1) file level metadata: file size, file type, last changed, etc.;

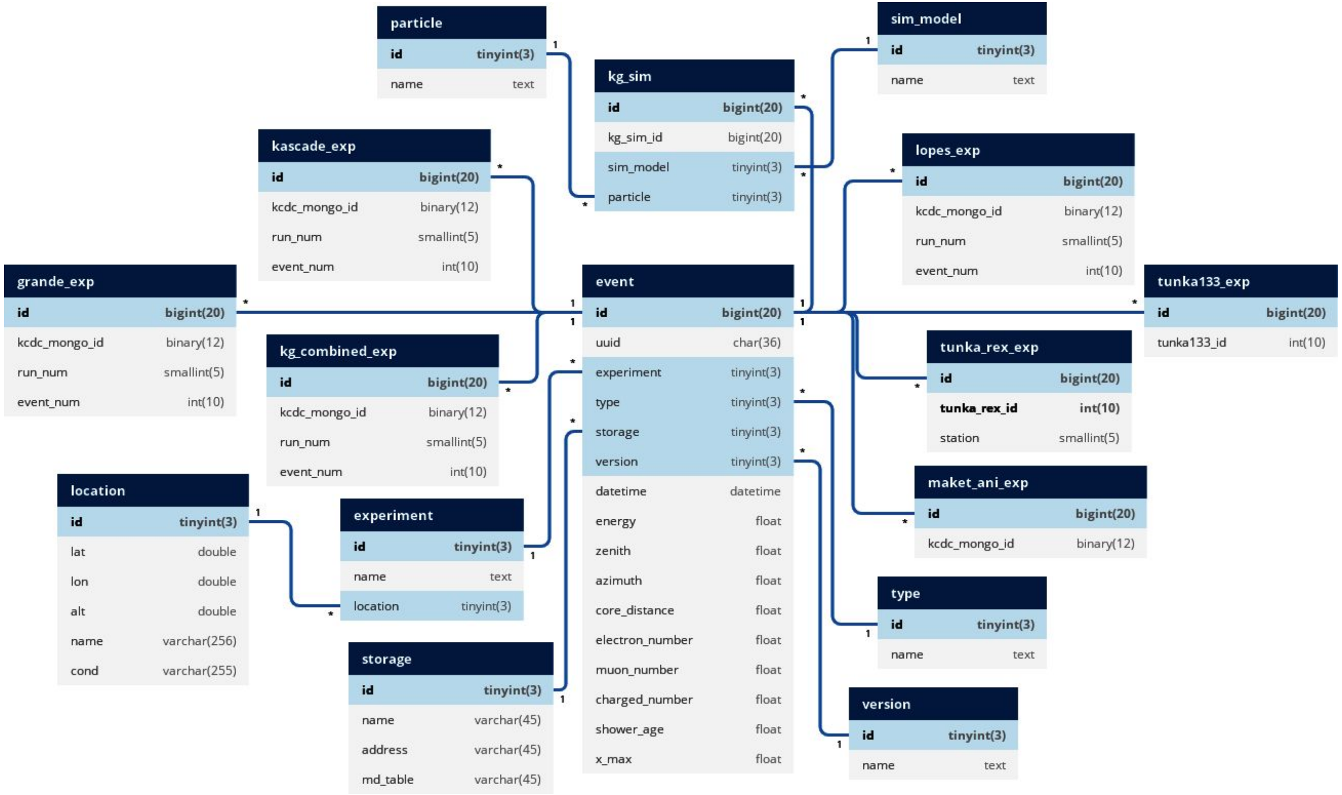
    (2) event parameter level: event id, datetime, setup, atmosphere, etc.

GRADLCI (meta)-data flow



* Implemented on dedicated server at KIT. Considered to be integrated into PUNCH4NDFI data portal

Metadata schemata GRADLCI



particle	
id	**tinyint(3)**
name	text

kg_sim	
id	**bigint(20)**
kg_sim_id	bigint(20)
sim_model	tinyint(3)
particle	tinyint(3)

sim_model	
id	**tinyint(3)**
name	text

kascade_exp	
id	**bigint(20)**
kcdc_mongo_id	binary(12)
run_num	smallint(5)
event_num	int(10)

lopes_exp	
id	**bigint(20)**
kcdc_mongo_id	binary(12)
run_num	smallint(5)
event_num	int(10)

grande_exp	
id	**bigint(20)**
kcdc_mongo_id	binary(12)
run_num	smallint(5)
event_num	int(10)

event	
id	**bigint(20)**
uuid	char(36)
experiment	tinyint(3)
type	tinyint(3)
storage	tinyint(3)
version	tinyint(3)
datetime	datetime
energy	float
zenith	float
azimuth	float
core_distance	float
electron_number	float
muon_number	float
charged_number	float
shower_age	float
x_max	float

tunka133_exp	
id	**bigint(20)**
tunka133_id	int(10)

kg_combined_exp	
id	**bigint(20)**
kcdc_mongo_id	binary(12)
run_num	smallint(5)
event_num	int(10)

tunka_rex_exp	
id	**bigint(20)**
tunka_rex_id	**int(10)**
station	smallint(5)

maket_ani_exp	
id	**bigint(20)**
kcdc_mongo_id	binary(12)

location	
id	**tinyint(3)**
lat	double
lon	double
alt	double
name	varchar(256)
cond	varchar(255)

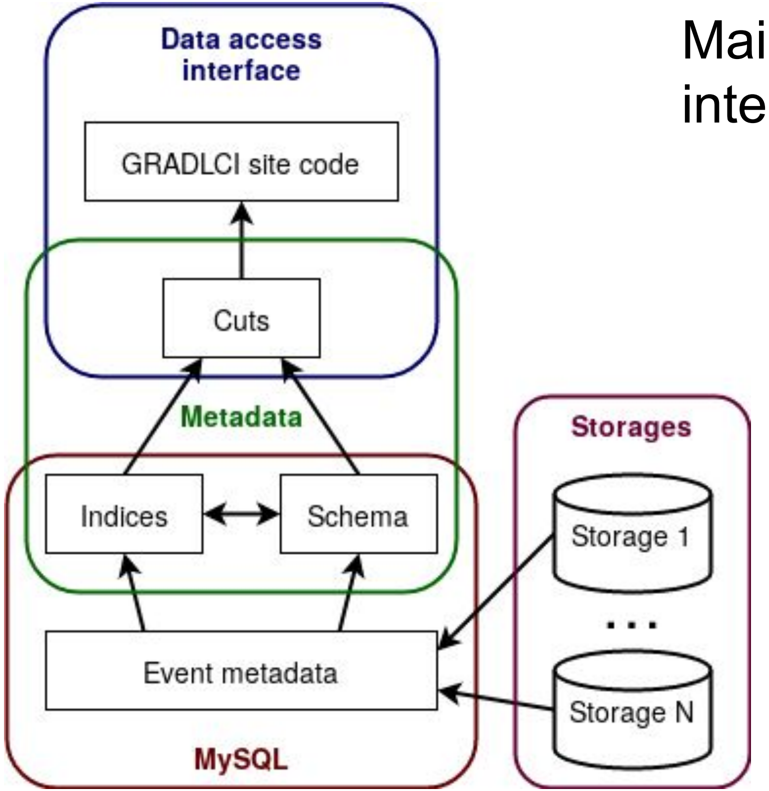
experiment	
id	**tinyint(3)**
name	text
location	tinyint(3)

storage	
id	**tinyint(3)**
name	varchar(45)
address	varchar(45)
md_table	varchar(45)

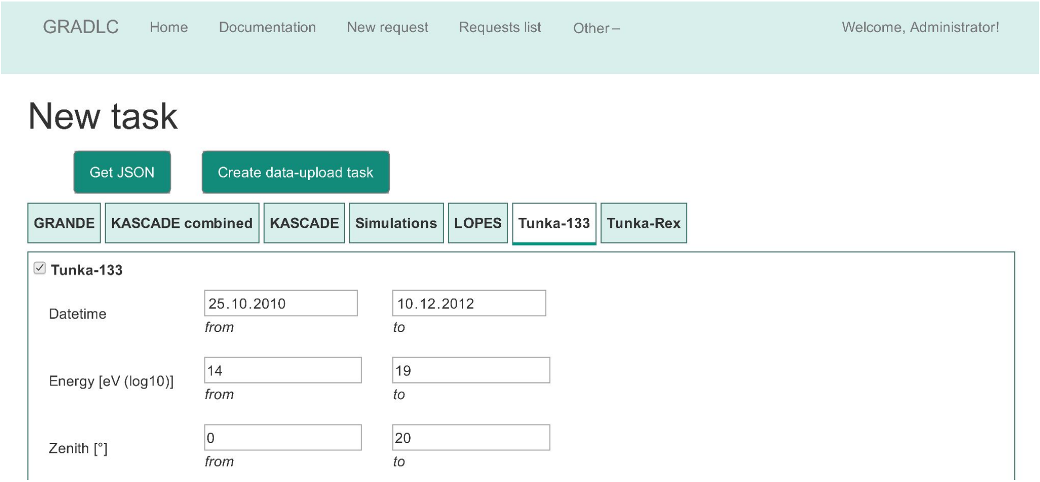
type	
id	**tinyint(3)**
name	text

version	
id	**tinyint(3)**
name	text

GRADLCI metadata acquisition



Main idea: provide a common data selection interface for various data sources



New task

☑ Tunka-133

Datetime	25.10.2010 *from*	10.12.2012 *to*
Energy [eV (log10)]	14 *from*	19 *to*
Zenith [°]	0 *from*	20 *to*

Aggregation server Web API*

Domain name: `gradlc-dc.iap.kit.edu`

Request type: JSON-RPC

Protocol: http

Authentication: HTTP Basic Auth

Possible requests

- Data requests
- Request status
- List of requests
- Remove request from the list
- Download file

Possible request status

- Running
- Scheduled
- Finished
- Failed
- Deleting
- Expired

Example:

Request:

```
{"id": "4998715b-cd5d-4c17-80fb-8139a74d66ea", "jsonrpc": "2.0", "method": "new_task", "params":
{"kascade_exp": {"datetime_max": "2011-10-10 00:00:00", "datetime_min": "2010-10-10 00:00:00",
"zenith_max": 20.0, "zenith_min": 0.0}}}
```

Response:

```
{"id": "4998715b-cd5d-4c17-80fb-8139a74d66ea", "jsonrpc": "2.0", "result": {"url":
"http://gradlc-dc.ikp.kit.edu/download/c3feaa45-b654-44d3-83e7-671b1ac0499c.7z  ", "uuid":
"c3feaa45-b654-44d3-83e7-671b1ac0499c"}}
```

* *Application Programming Interface*