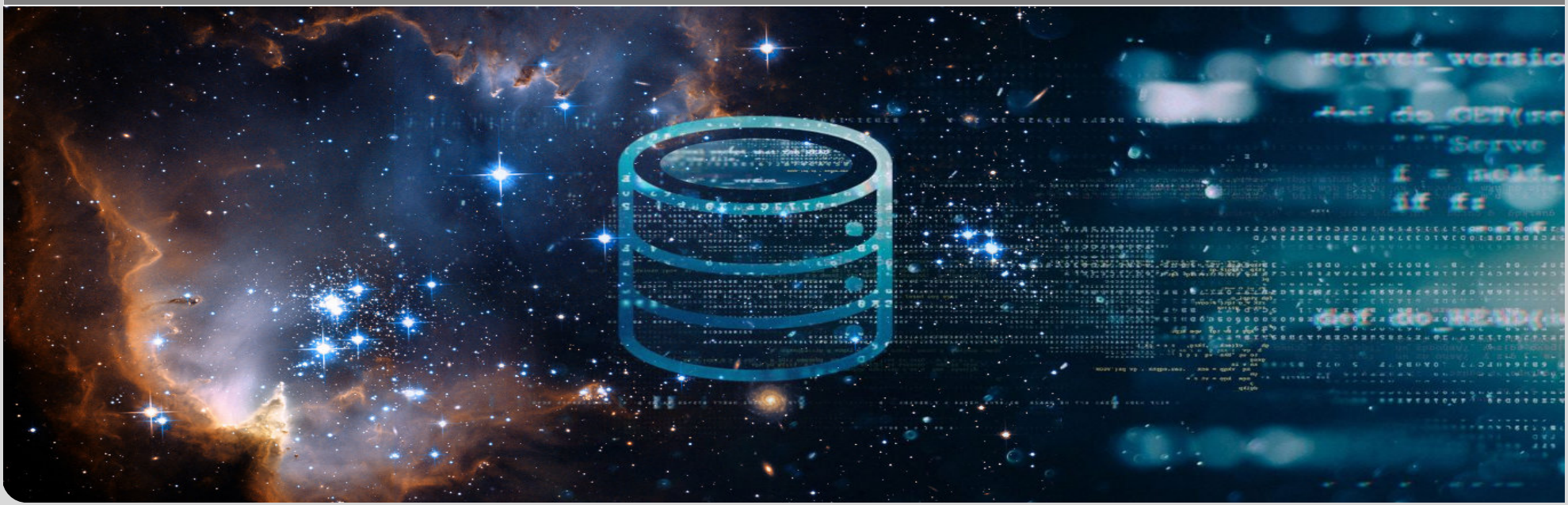
Karlsruhe Institute of Technology

Data aggregation of distributed setups in astroparticle physics

IV International Workshop ”Data life cycle in physics” (DLC-2020)

Victoria Tokareva | 8-10 June 2020

INSTITUTE FOR NUCLEAR PHYSICS (IKP)



German-Russian Astroparticle Data Life Cycle Initiative*

The international initiative aimed at automating the maintenance of astroparticle-physics data throughout their entire life cycle.

Matrosov Institute
for System Dynamics
and Control Theory

KASCADE - Grande
KArlsruhe Shower Core and Array DEtector - Grande

*Granted by RSF-Helmholtz Joint Research Groups

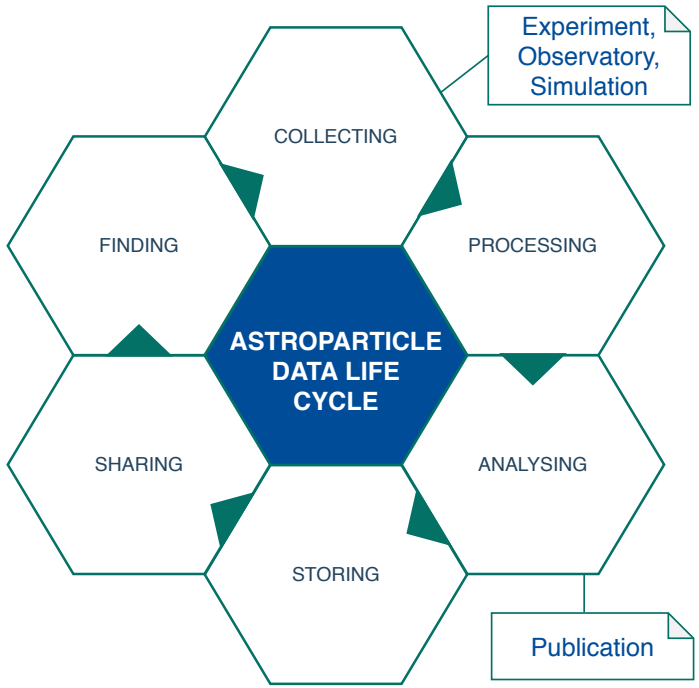
Astroparticle Data Life Cycle

Data Life Cycle (DLC)

The sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion.

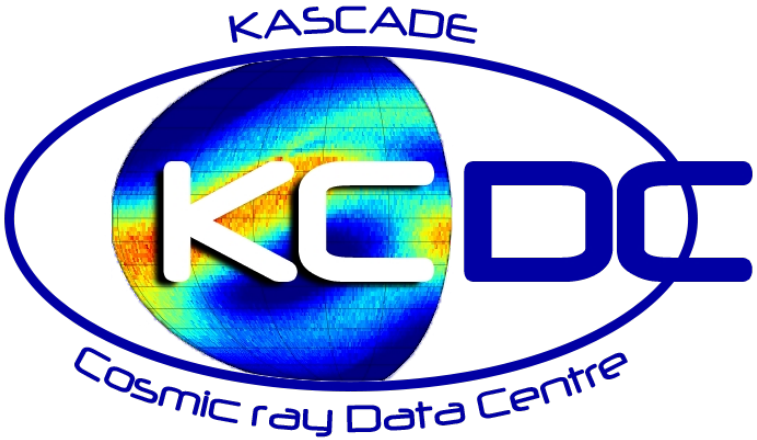
Features

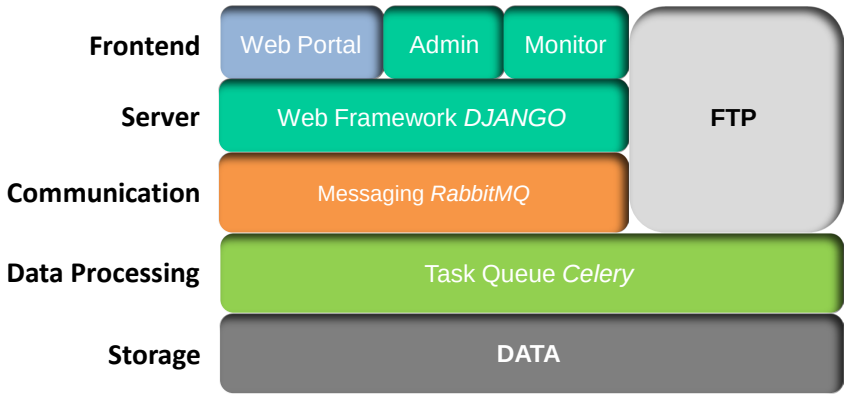
- The system of aggregated data selection and retrieval;
- Flexible horizontal expansion allowing connection of heterogeneous storages of astroparticle data;
- Usage of modern virtualization and machine-learning technologies;
- Online data analysis capability;
- Access to scientific data for the general public.



KASCADE Cosmic-ray Data Center (KCDC)

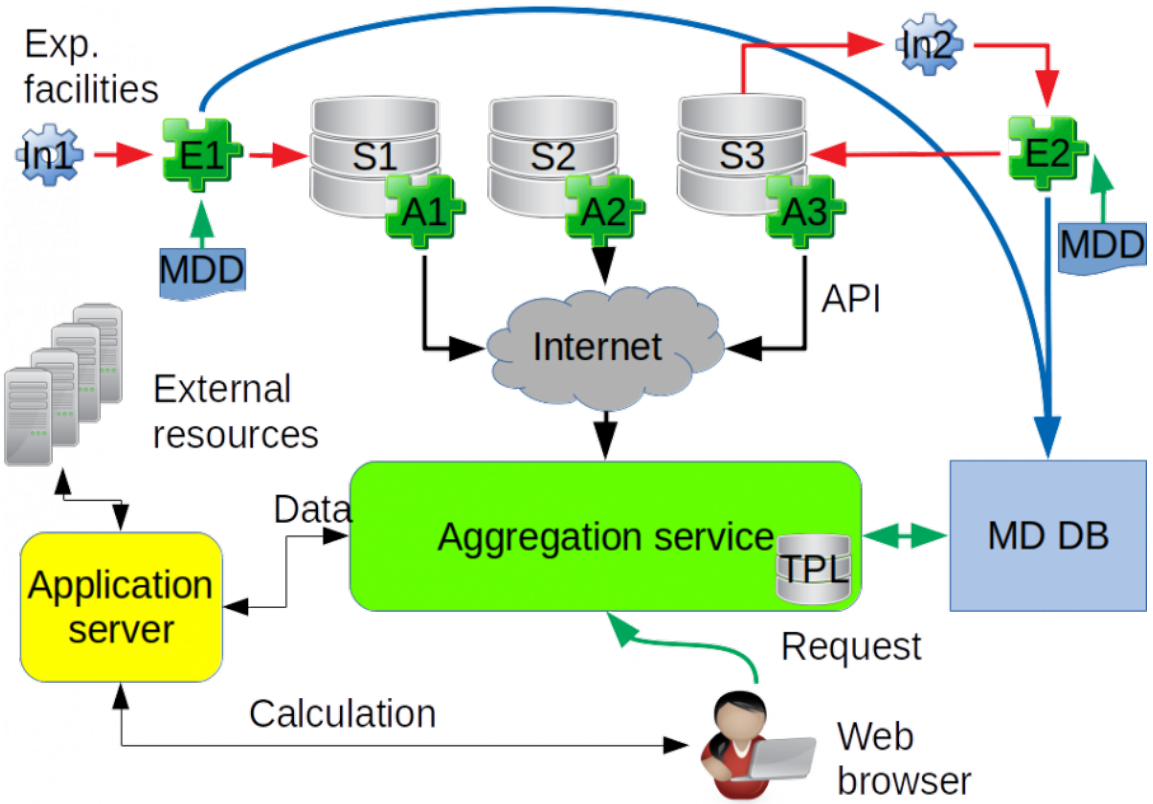
- providing free, unlimited, reliable open access to KASCADE cosmic ray data at https://kcdc.ikp.kit.edu;
- the new release, named PENTARUS was published just recently with many improvements (see J. Wochele's talk)
- keep being extended with new fruitfull sevices (see F. Polgart's talk)
- selection of fully calibrated quantities and detector signals;
- information platform and archive of KASCADE software and data;
- uses modern and open source web technologies.





DLC Architecture

See the talks by A.Kryukov and M.D.Nguyen



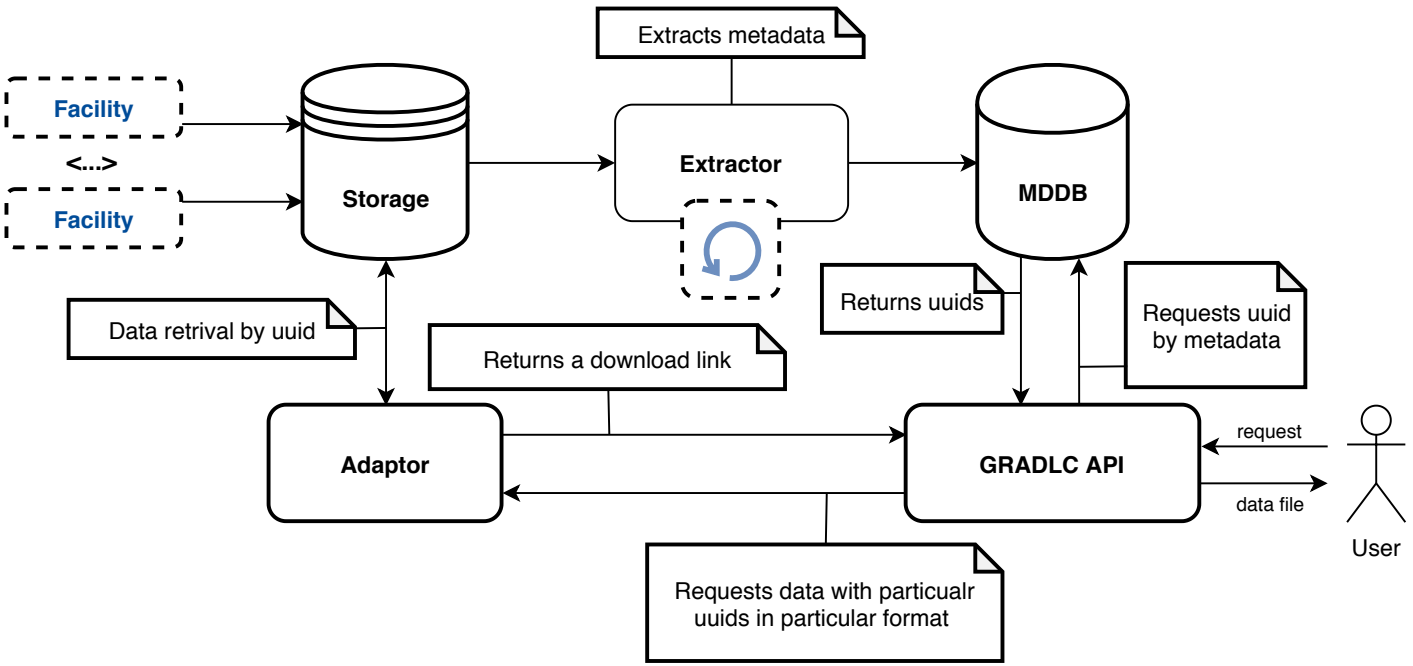
- **Si** — local data storages;
- **Ini** — data sources of different types;
- **MDD** — metadata description;
- **Ei** — metadata extractors;
- **Ai** — adapters, provide API for data access;
- **TPL** — template library;
- **MD DB** — metadata database.

Data aggregation

Metadata definition

We introduce 2 level metadata model:

1. physical level metadata: file size, file type, last changed, etc.
2. event level: event_id, datetime, setup, atmosphere, etc.



Setups and data - 1

KASCADE

- 252 scintillators
- 450 000 000 evnt.

KASCADE-GRANDE

- 37 scintillators
- 35 310 393 evnt.

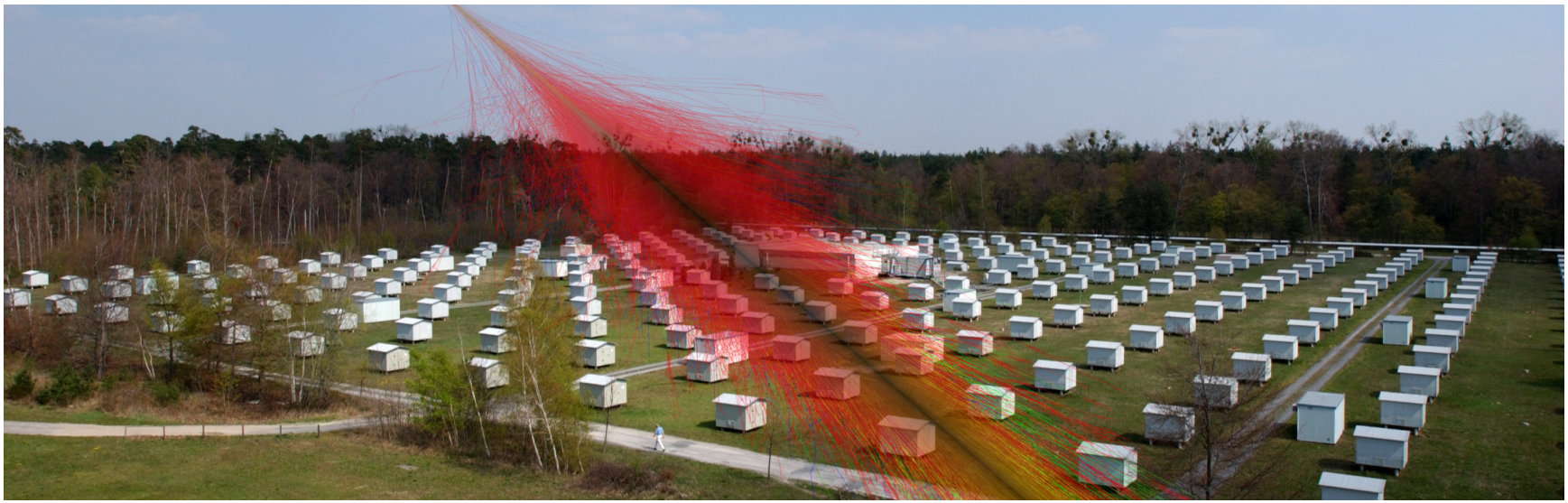
LOPES

- radio antennas array
- 3 058 evnt.

COMBINED

- All three datasets mapped for combined analysis

- MongoDB
- total data volume $\approx$ 20 TB



Setups and data - 2

Simulations

- KASCADE, KASCADE-GRANDE, COMBINED
- format: ROOT

Tunka-133

- 133 photomultipliers
- MySQL
- 7 421 630 events
- 0.5 GB

Tunka-Rex

- 63 radio antennas
- MySQL
- 107 360 524 events
- $\approx$ 3 TB

Connection and data retrieval

Connection details

IP: 141.52.67.147 -P 55000

Request type: JSON-RPC

Protocol: http

Authentication: HTTP Basic Auth, with KCDC or GRADLCI account

Possible requests:

- Data retrieval
- Request status
- List of requests
- Remove request
- Download file

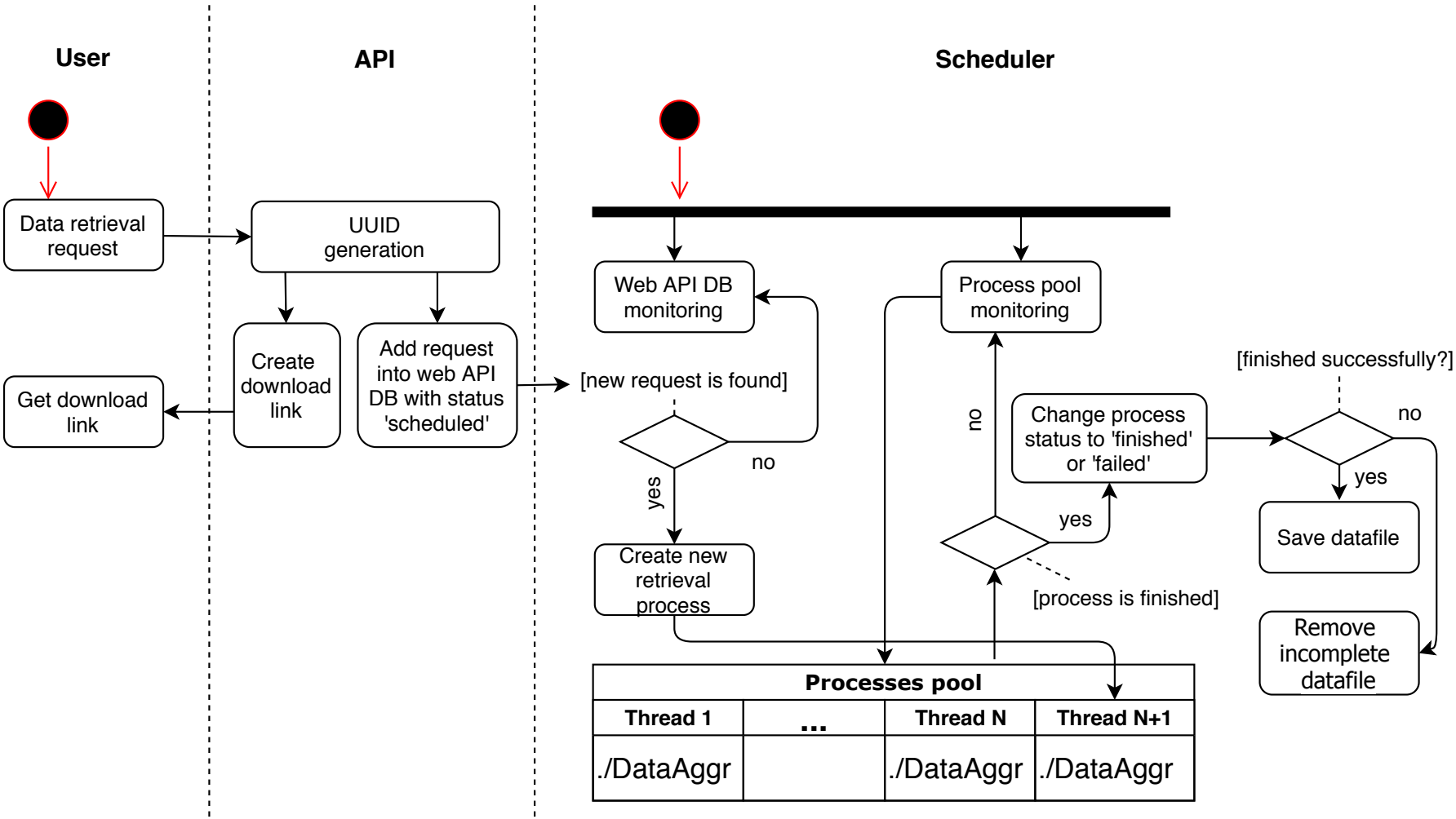
Data retrieval

Request: {"id": "930a254d-f6e9-478b-a589-494aff92f721", "jsonrpc": "2.0", "method": "new_task", "params": {"kascade_exp": {"datetime_min": "2010-10-10 00:00:00", "datetime_max": "2011-11-11 00:00:00", "energy_min": 14.0, "energy_max": 19.0, "zenith_min": 0.0, "zenith_max": 20.0}, "tunka133_exp": {"datetime_min": "2010-10-25 00:00:00", "datetime_max": "2012-12-10 00:00:00", "energy_min": 14.0, "energy_max": 19.0, "zenith_min": 0.0, "zenith_max": 20.0}}}

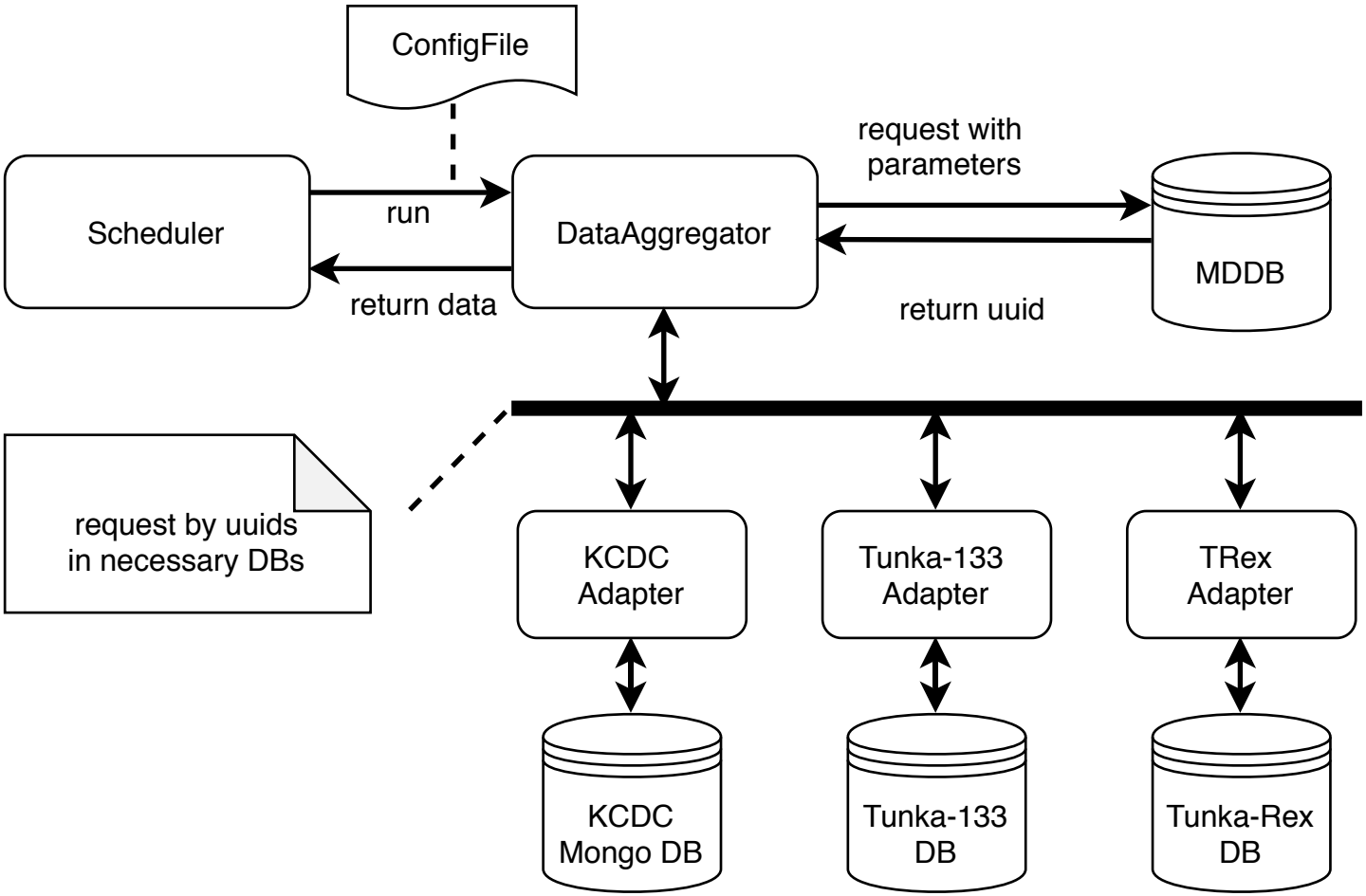
Response: { "jsonrpc": "2.0", "result": {"task_uuid": "89c70a67-28cb-4952-a3ed-250116a72363", "url": "http://141.52.67.147:55000/download/89c70a67-28cb-4952-a3ed-250116a72363.txt.bz2" }, "id": "cf919bb3-7028-4238-9438-2601cbfde3a6"}

Download: `wget http://141.52.67.147:55000/download/89c70a67-28cb-4952-a3ed-250116a72363.txt.bz2`

Async request processing



Data retrieval



Request query status

- Single query status:

```
{ "id": "e7cbcfdc-8bfb-440f-8fc9-3e8bbadaf99d",
"jsonrpc": "2.0", "method": "show_task", "parameters": {
"task_uuid": "89c70a67-28cb-4952-a3ed-250116a72363" } }
```

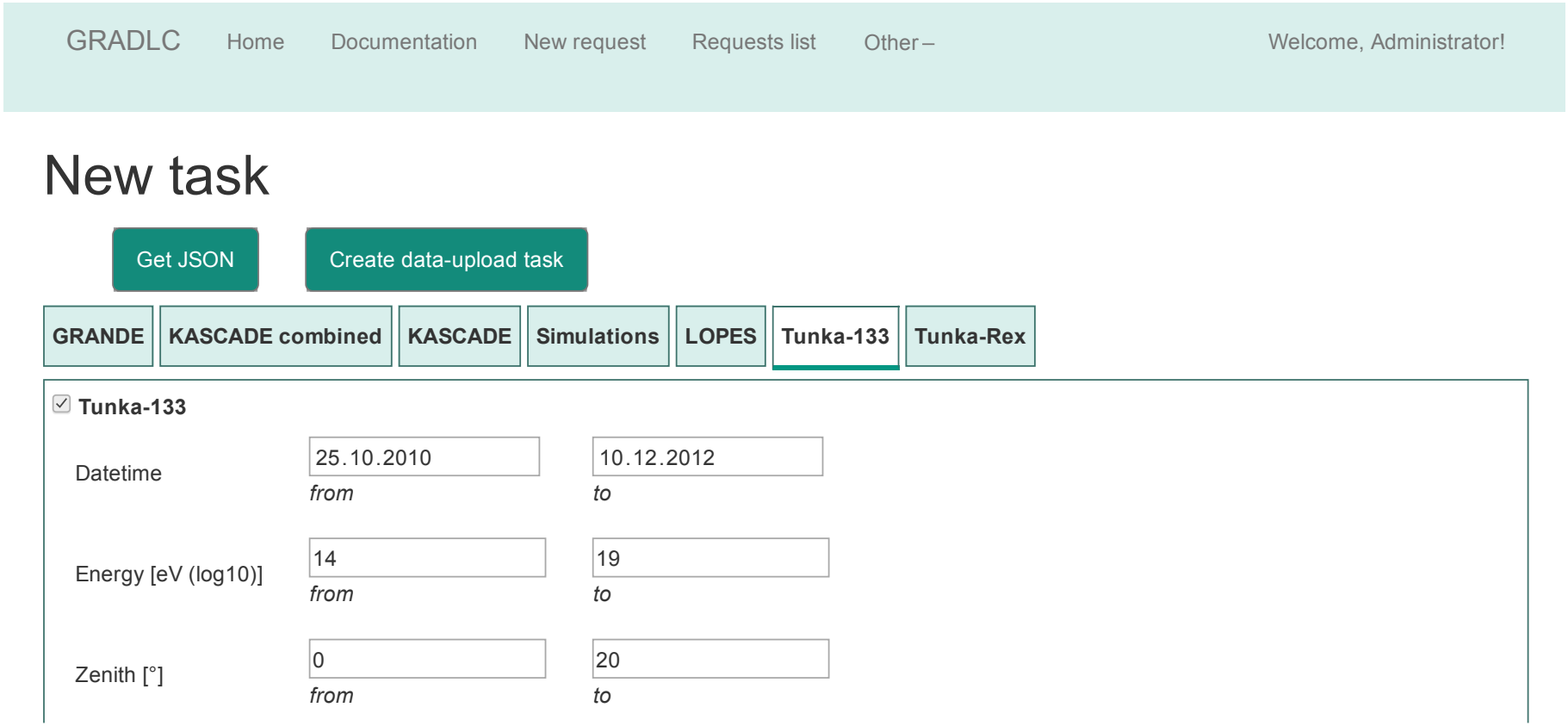
- All queries status:

```
{ "id": "3c325e96-c073-44de-abb6-3c84b2e21297",
"jsonrpc": "2.0", "method": "list_tasks" }
```

Possible statuses:

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

Web UI: new request



Web UI: request status

GRADLC Home Documentation New request Requests list Other – Welcome, Administrator!

List of data-upload tasks

ID	User	Requested at	Finished at	Status	Download	
34	Administrator	2020-06-07 04:41:24	2020-06-07 06:41:25	Finished	89c70a67-28cb-4952-a3ed-250116a72363.txt.bz2	Delete
33	Administrator	2020-06-06 14:10:22	2020-06-06 16:10:24	Finished	1546370b-23b8-404a-885a-1443433eebbc.txt.bz2	Delete
10	Administrator	2020-03-09 23:42:46	2020-03-09 23:44:00	Expired		Delete

Outlook

- In framework of GRADLCI an API for distributed data aggregation was developed
- It provides an access to the experimental data and simulations for such experiments as KASCADE, KASCADE-GRANDE, LOPES, Tunka-133 Tunka-Rex and can be extended with other setups
- It can be used as a service as well as a stand-alone application with web user interface
- Features: faster data search with metadata database (MDDB), asynchronous multithread data processing, user authentication with KCDC and GRADLCI accounts
- More information can be found in the documentation at http://141.52.67.147:55000/web/doc.