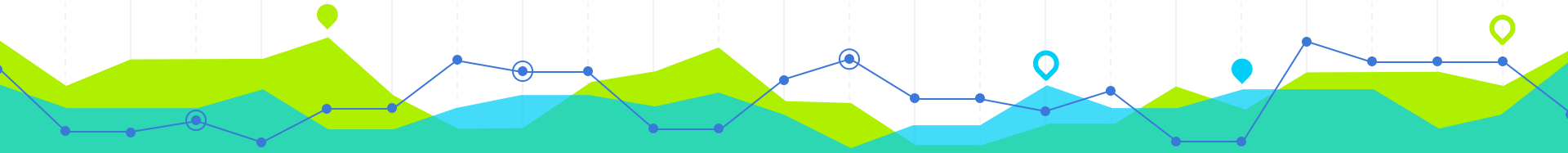
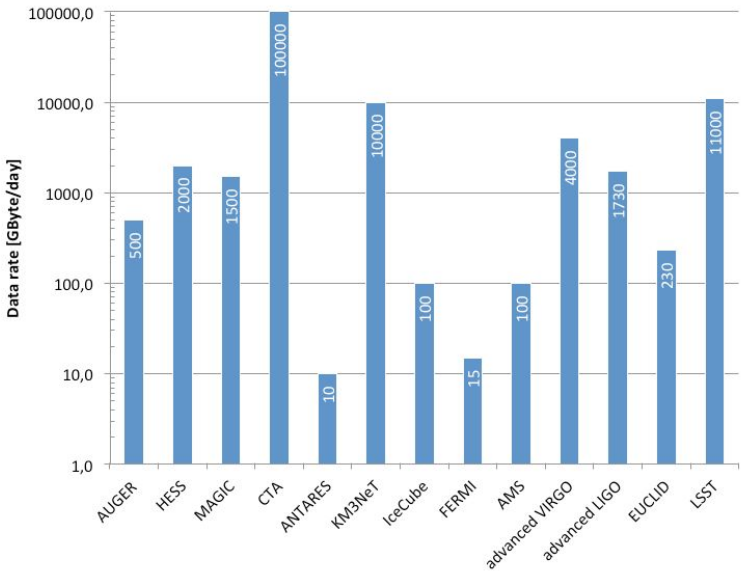
[Meta]data curation in astroparticle physics on KCDC use case

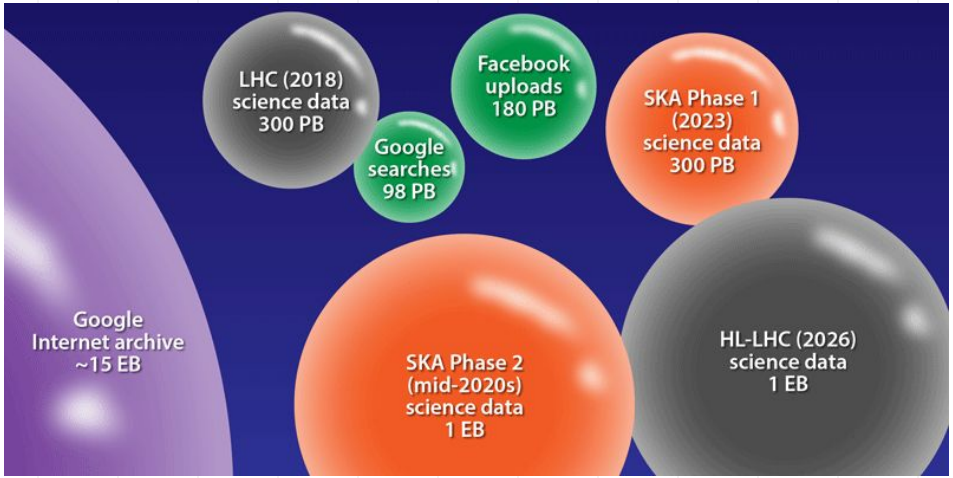


V. Tokareva, A. Haungs, D. Wochele, J. Wochele, D. Kang
Karlsruhe Institute of Technology, Institute for Astroparticle Physics
German Conference of Women in Physics
Karlsruhe, 24-27 November 2022

Data rates in modern [astro]particle physics projects



Data rates produced by the individual astroparticle physics experiments, 2015. [1]



A comparison of the yearly data volumes of current and future projects, where PB stands for petabyte (10^{15} bytes) and EB stands for exabyte (10^{18} bytes). [2]

[1] Berghöfer T. et al. Towards a model for computing in european astroparticle physics //arXiv preprint arXiv:1512.00988. – 2015.

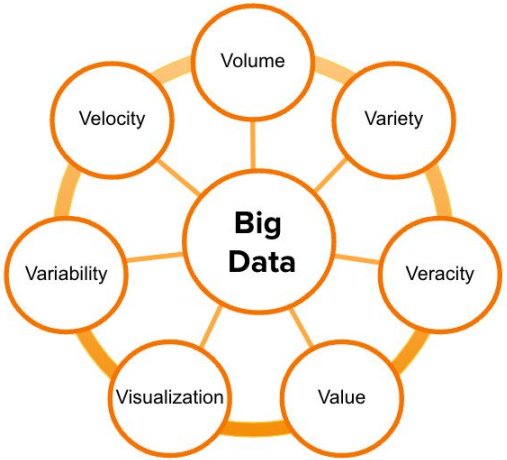
[2] Alan Stonebraker and V. Gülzow/DESY, Facing a Downpour of Data, Scientists Look to the Cloud, APS Physics, url: https://physics.aps.org/articles/v13/14, 2020.

[Meta]data curation challenges

Data is *potential* information, analogous to potential energy: work is required to release it [3].

Metadata record is itself a container for data about an object [3].

- Big and open data
- FAIR (Findable - Accessible - Interoperable - Reusable) data
- Highly collaborative globally distributed data management for big (10^3+) scientific communities
- Data irreversibility and reduction
- Harnessing heterogeneous resources

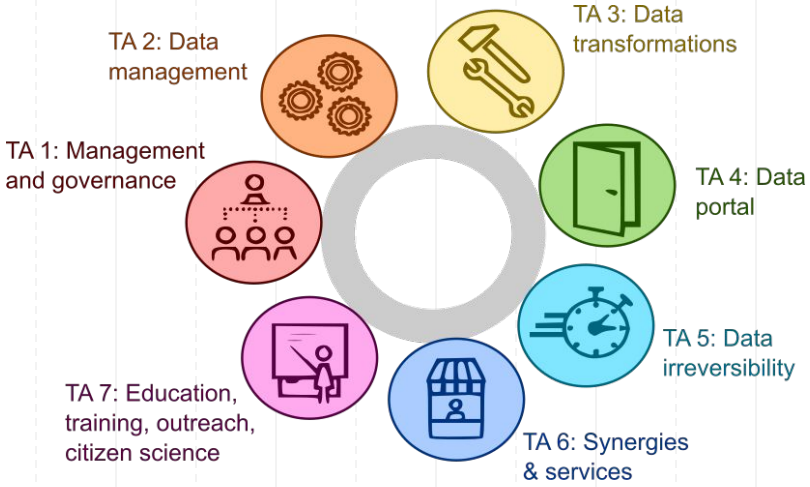


Source: Moore. M. The 7 V's of Big Data, url: https://shorturl.at/nzFN8, 2021.

[3] Pomerantz, J. Metadata. MIT Press, 2015.

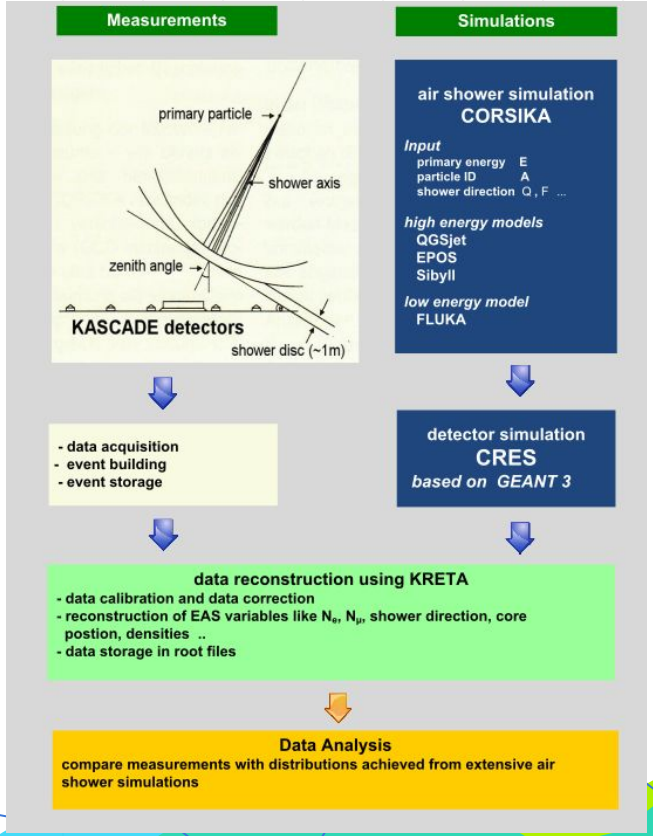
Particles, Universe, NuClei and Hadrons for Nationale Forschungs-Daten Infrastruktur (PUNCH4NFDI)

- PUNCH4NFDI is the NFDI consortium of particle, astro-, astroparticle, hadron and nuclear physics, which contributes in the objective of the NFDI is to systematically index, edit, interconnect and make available the valuable stock of data from science and research
- The prime goal of PUNCH4NFDI is the setup of a federated and "FAIR" science data platform, offering the infrastructures and interfaces necessary for the access to and use of data and computing resources of the involved communities and beyond
- https://www.punch4nfdi.de/



KCDC - KASCADE Cosmic Ray Data Centre

- KCDC is the public data centre for high-energy astroparticle physics
- Based on the data of the KASCADE experiment, contains as well data by KASCADE-Grande, LOPES, Maket-Ani, allows further extensions
- More than 433.000.000 events
- Established in 2013
- https://kcdc.iap.kit.edu/



KCDC's functionality

- Archive of KASCADE software and data
- Provides free, unlimited, reliable open access to KASCADE cosmic ray experiment
  - Selection of fully calibrated quantities and detector signals
  - Custom user data cuts
  - Allows data selection using both GUI and RESTless API
- Allows interactive analysis with integrated Jupyter Notebooks
- Information platform: physics and experiment backgrounds, tutorials, reference information

Data shops and formats at KCDC

The data sets are organised into so-called **data shops** (data marts):

- 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 quantities
- Maket-Ani - provides quantities from the Maket-Ani setup

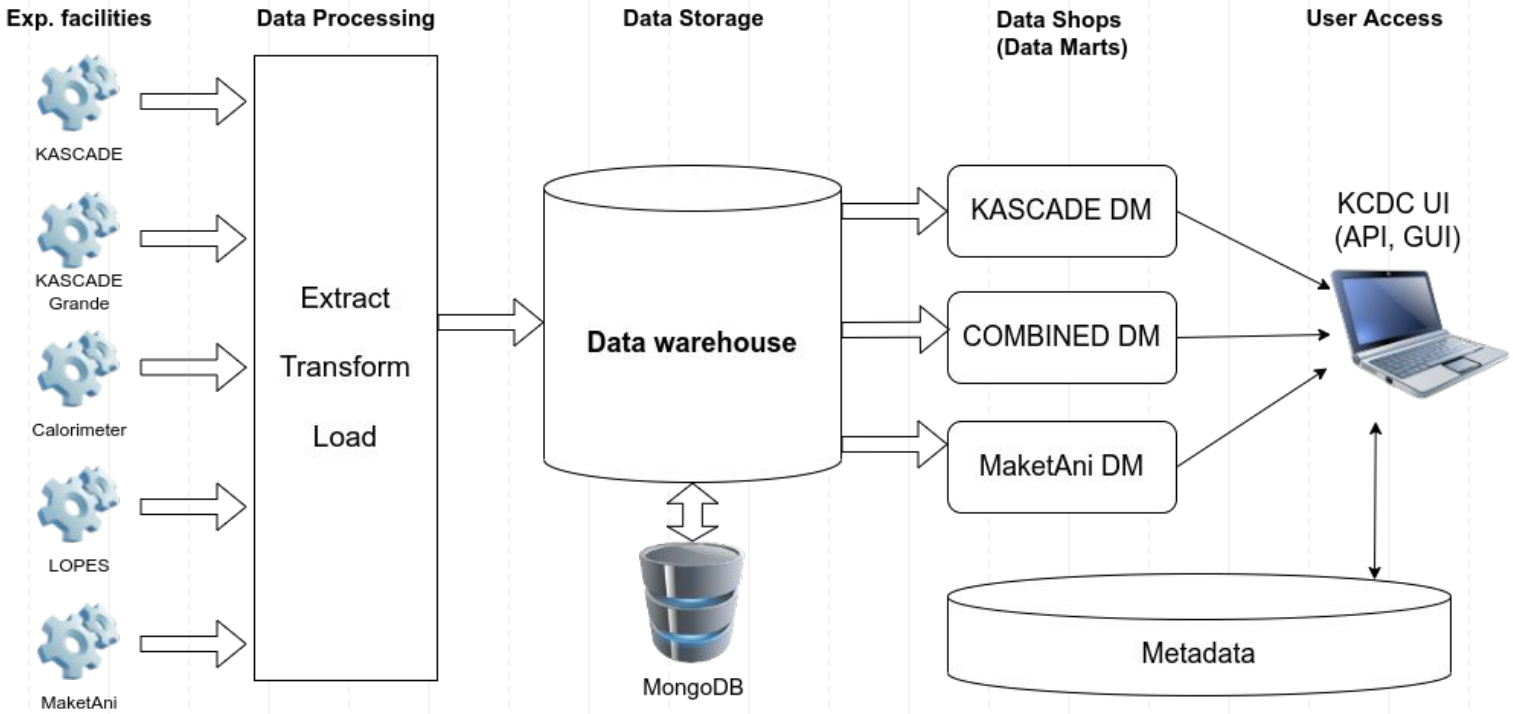
New data shops can be added.

The following **data formats*** are supported:

- 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

KCDC's software architecture

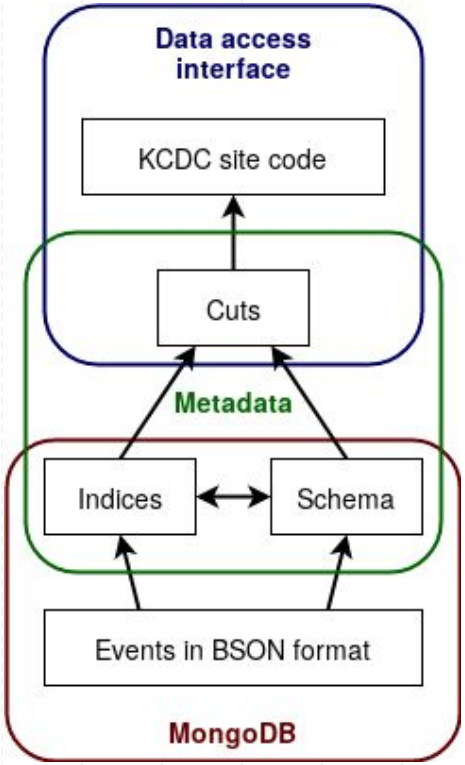


Metadata handling at KCDC

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 data acquisition



KASCADE Data Shop

Components Available	Components Selected	Quantities and Cuts			
GRANDE	General Info	Toggle all		*KASCADE*	
Calorimeter	KASCADE	Energy	range:	13 to 19 eV [log10]	Add Cut
LOPES		X Core Position	range:	-91 to 91 m	Add Cut
		Y Core Position	range:	-91 to 91 m	Add Cut
		Zenith Angle	range:	0 to 60 °	Add Cut
		Azimuth Angle	range:	0 to 360 °	Add Cut
		Electron Number	range:	2 to 8.7 [log10]	Add Cut
		Muon Number	range:	2 to 7.7 [log10]	Add Cut
		Shower Age	range:	0.1 to 1.48	Add Cut

Verify & Submit Request

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:

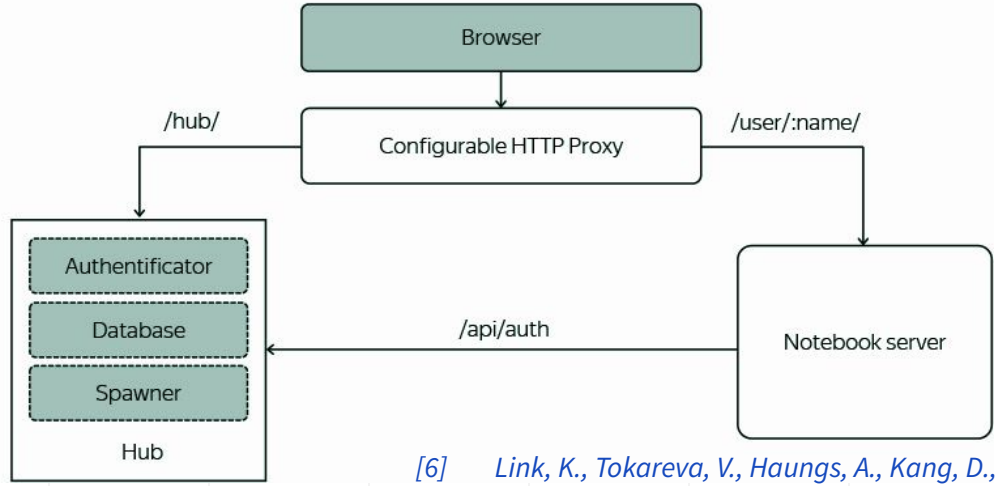
{"detail":"Invalid basic header. Credentials not correctly base64 encoded."}

[4] *Online API documentation: https://kcdc.iap.kit.edu/datashop/api/docs/index.html*

[5] *Wochele J. et al. KCDC User Manual: https://kcdc.iap.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf*

JupyterHub for data analysis

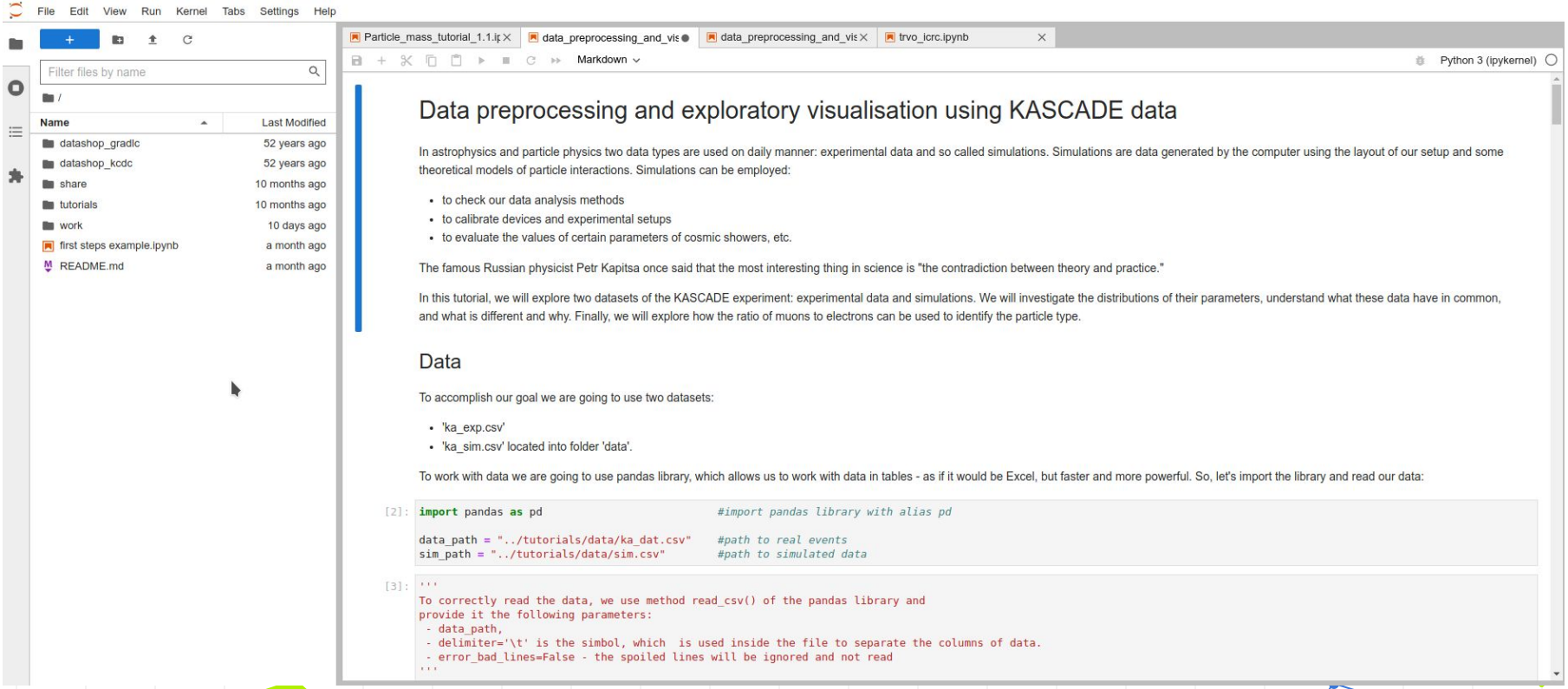
- Login via KCDC credentials
- Administration using Docker Swarm
- Tutorials by: KASCADE, IceCube, TRVO



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

[6] 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.

Usage of KCDC's JupyterHub



Open technologies in use:

- Django Web Framework
- Messaging RabbitMQ
- Celery Task Queue
- NoSQL (MongoDB) database
- JupyterHub for Jupyter Notebooks
- Docker/Singularity
- RESTful API
- Python, bash, JSON, HTML

Summary: Why is this use case valuable?

- Handling of heterogeneous [meta]data from different sources
- Support of multiple data formats and event-level data cuts (through data-on-read approach)
- Works with NoSQL database
- Only open source technologies
- Long-term experience of deployment and maintenance within a scientific organization
- Well-developed user interfaces
- Integration with data analysis services (Jupyter Notebooks)

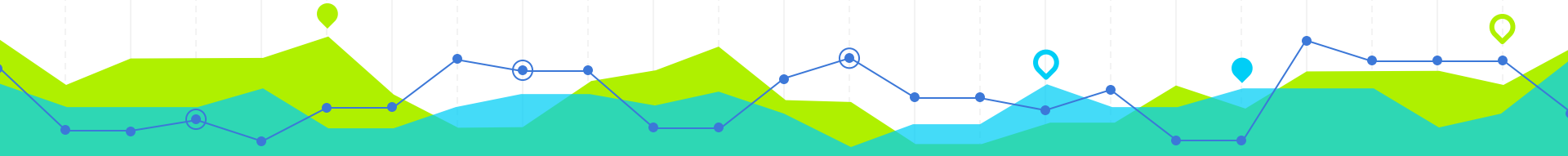
THANKS!

Contact me: victoria.tokareva@kit.edu

linkedin.com/in/victoria-tokareva-2a6999a2

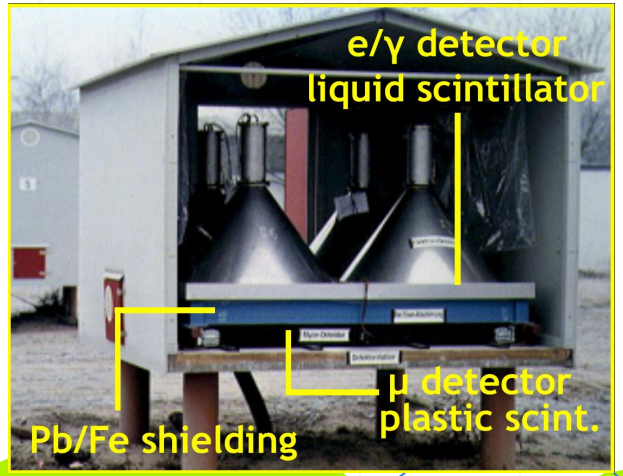
PUNCH4NFDI: https://punch4nfdi.de

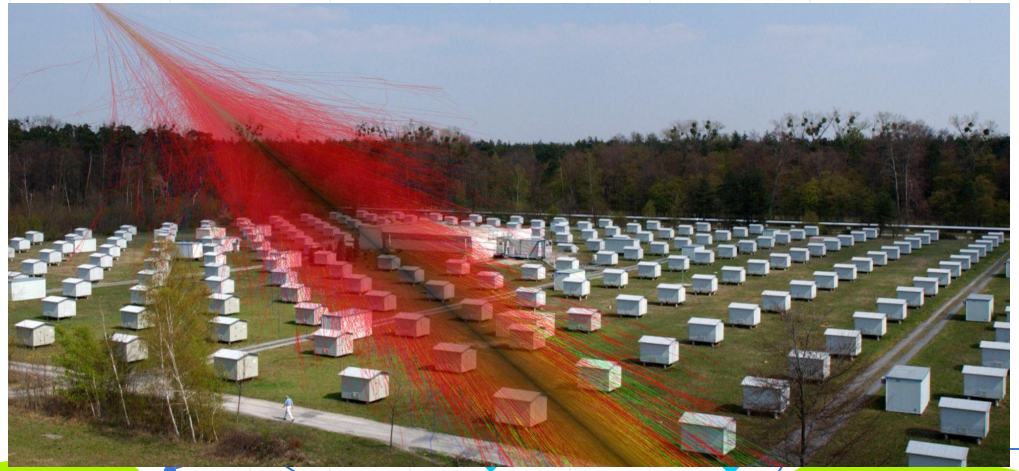
KCDC: https://kcdc.iap.kit.edu



KASCADE - KArlsruhe Shower Core and Array DEtector

- Location: 110 m a.s.l., 49◦ N, 8◦ E, KIT-Campus North, Karlsruhe, Germany
- Operation time: 1996 October – 2010 May ⇒ e/γ detector liquid scintillator effective time ~ 4223.6 days
- Area: 200 × 200 m^2, E = 100 TeV – 80 PeV
- 252 scintillator detectors
- KASCADE data are published in open access at KASCADE Cosmic Ray Data Centre since 2013



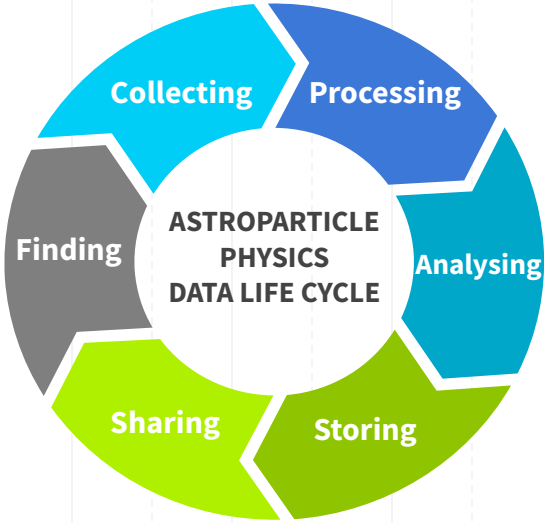


Use case studies for PUNCH4NFDI

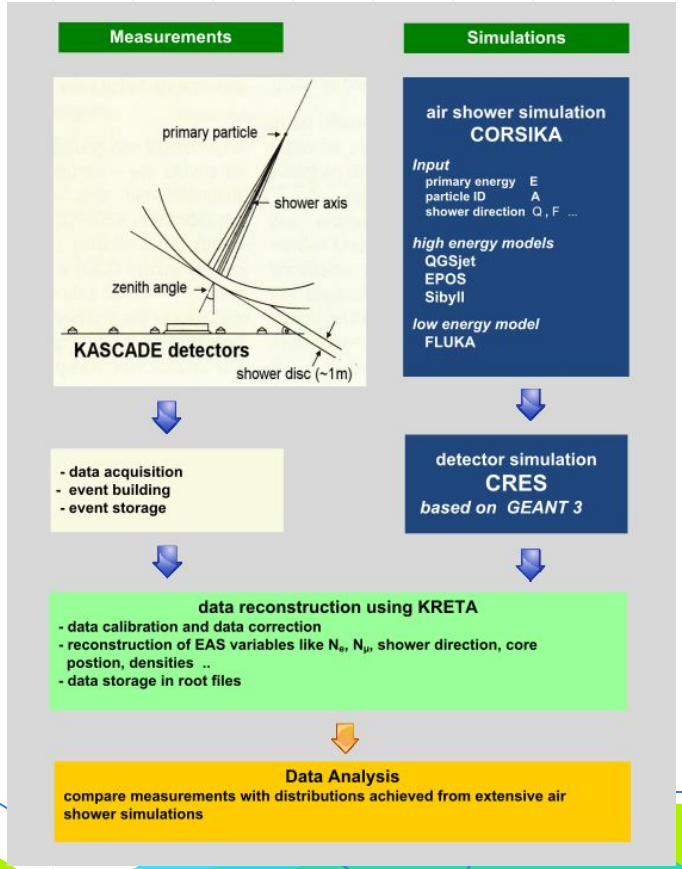
Use case studies allow definition of PUNCH-overarching tasks and deliverables and currently go in 6 classes:

- Validating and publishing scientific data collections
- Analysis of local or distributed data sets
- Execution of analysis of numerical simulations
- Community-overarching data challenges
- Real-time challenges & data irreversibility
- Use cases from external partners

Life cycles and workflows in astroparticle physics



Data analysis workflow in astroparticle physics, example of KASCADE experiment



Data quantities example

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

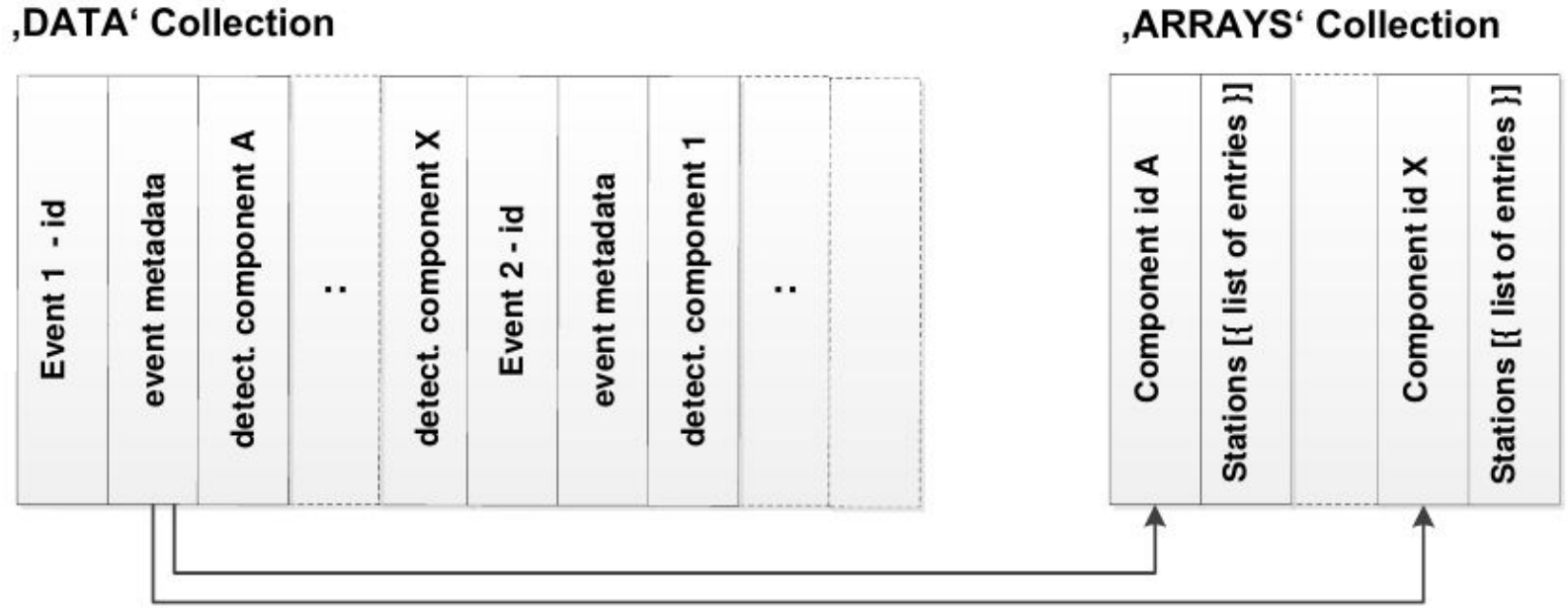
GRANDE Quantities

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

KCDC's data overview

Setup / Detector component	Experimental data		Simulations	
	Events	Size	Events	Size
KASCADE	433 209 340	3 200 GB	22 490 883	26.8 GB
GRANDE	35 310 393	260 GB	4 149 416	4.2 GB
COMBINED	15 635 550	120 GB	2 030 227	2.6 GB
LOPES	3 058	25 MB	—	—
MAKET-ANI	2 682 264	1 GB	—	—

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