Data Knowledge Base for the ATLAS collaboration

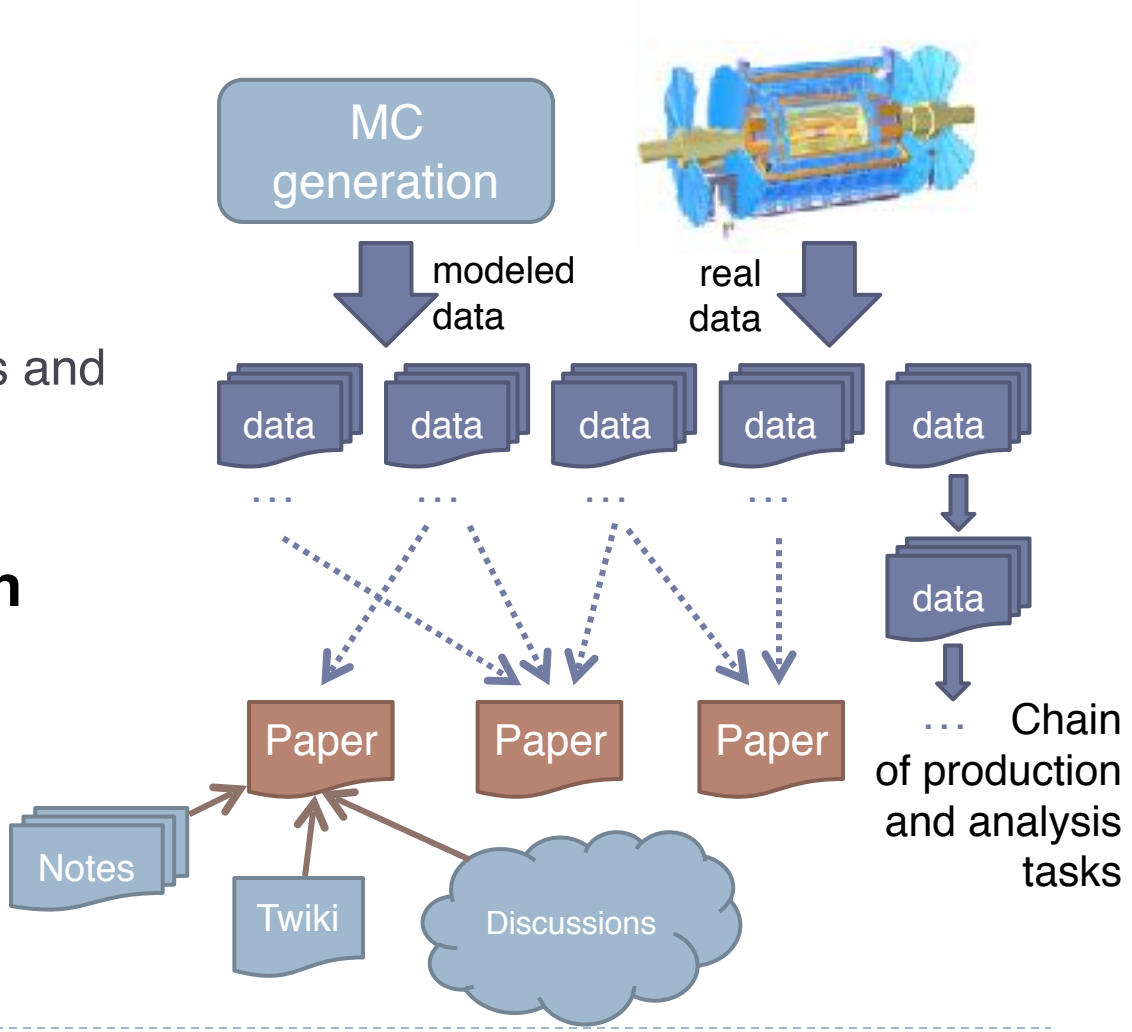
M Golosova[1], M Grigorieva[1], A Kaida[2]

[1]NRC "Kurchatov Institute"

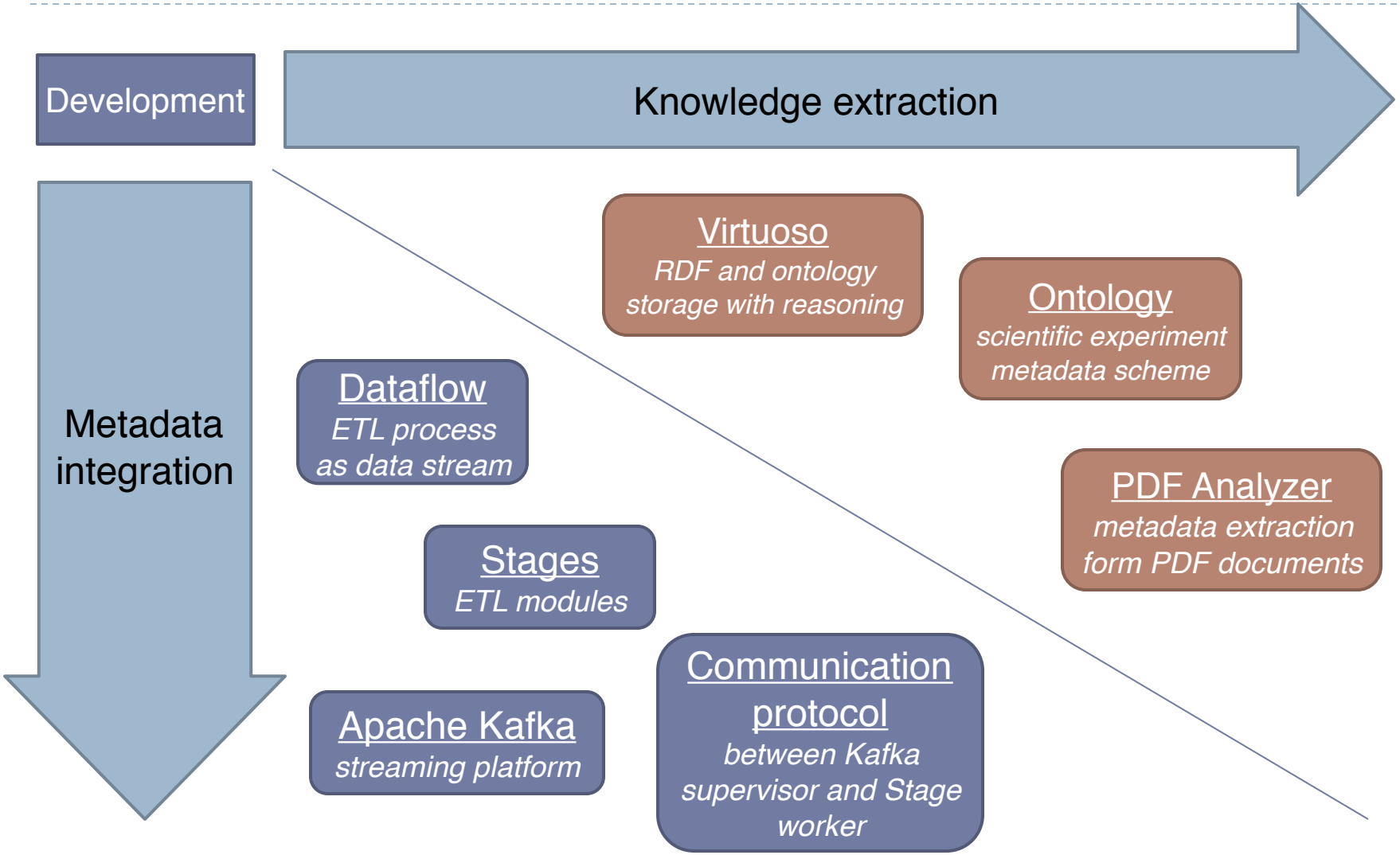
[2]NR Tomsk Polytechnic University

Data Knowledge Base

- **DKB R&D project**
  - was initiated in 2016
- **Initial goal**
  - reconstruct connections between research results and data samples
- **Suggested information sources**
  - PDF
  - wiki pages
  - Indico
  - …



Development routes



Concept evolution

- Started with *unstructured sources*:
  - proof-of-concept: DKB is able to integrate information from different sources and restore connections between information objects;
  - information from unstructured sources is not always reliable.
- Can be applied to *multi-source requests*:
  Request example:
  - *find production tasks by some attribute via BigPanDA monitoring web-interface;*
  - *get information from Rucio for related (input and output) data samples via CLI;*
  - *get information from AMI for same data samples via CLI.*
- New DKB concept: universal tool for multi-source queries:
  - choose storage and metadata scheme appropriate for the request;
  - extract metadata from multiple sources;
  - transform extracted metadata to fit data scheme;
  - load transformed metadata to the DKB internal storage;
  - provide interface to search and navigate integrated metadata as a consistent information field.

What`s done

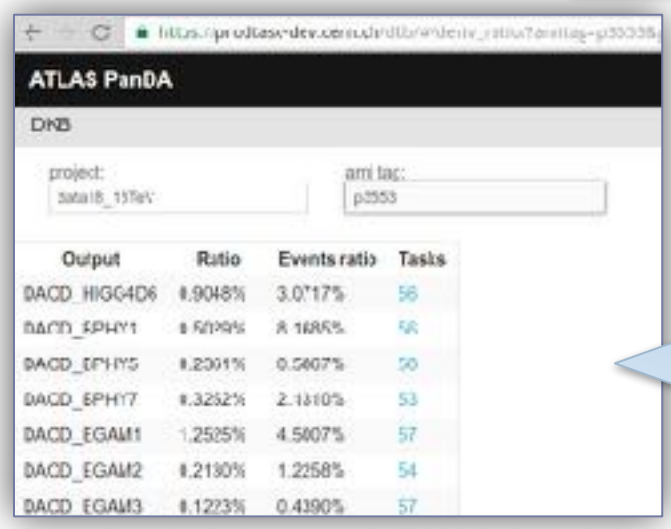
Request example:

- find production tasks by some attrib
- get information from Rucio for relat
- get information from AMI for

WEB-based interface:

- Production and analysis tasks search by keywords:
  (full-text search by all text information available for tasks: task name, user name, research group, campaign, project, description, ATLAS geometry, tags…)

- Derivation tasks statistics
  (project + AMI tag)

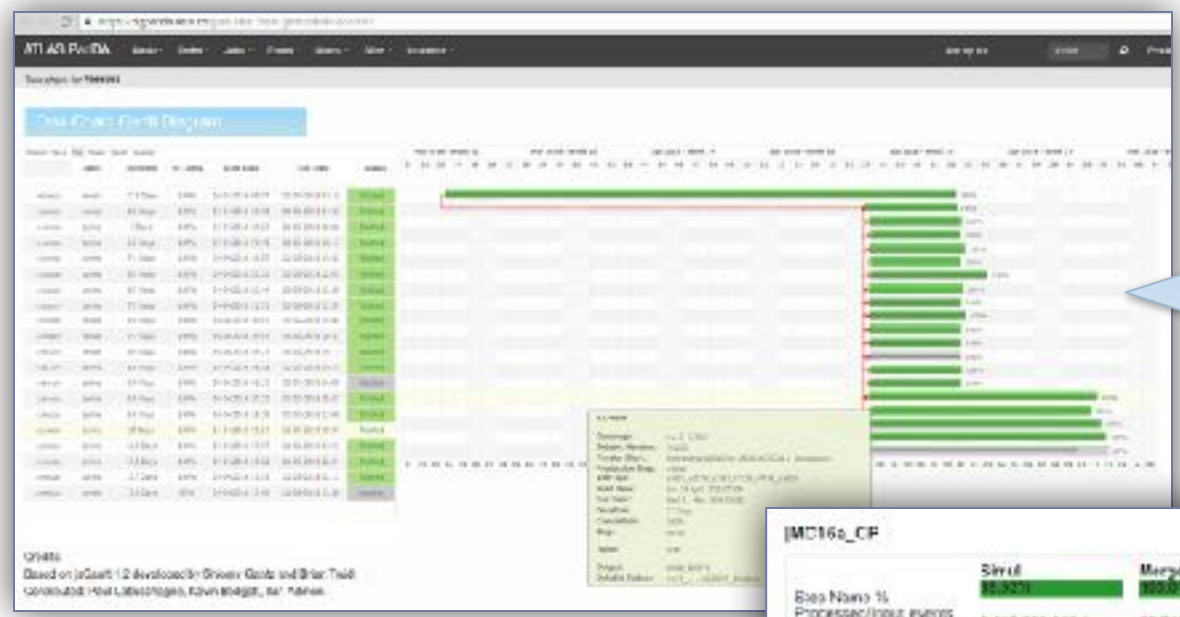


https://prodtask-dev.cern.ch/dkb/#/deriv_ratio/

http://prodtask-dev.cern.ch/dkb/#/task_keywords/

What`s coming next (WIP)

- New web interfaces:
  - Extended search interface
  - Task statistics by production steps
  - Task chain resource usage statistics
- REST API:
  - Isolate web interfaces from DKB internals (storage(s) and data schemas)



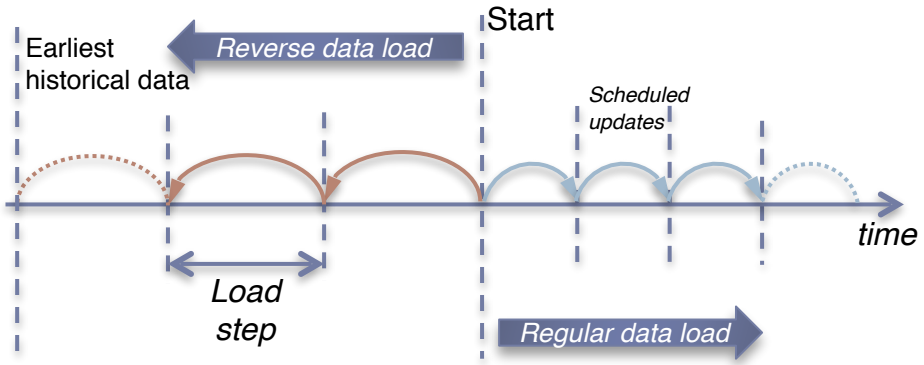
https://bigpanda.cern.ch/ganttTaskChain/

https://prodtask-dev.cern.ch/prodtask/request_hashtags_main/#/hashtags/

Under the hood (how it works)

- Dataflow task is run by schedule (every hour), integrating new/updated data from sources
- When dataflow changes (added extraction of new data or changed/extended logic of a processing step), whole set of data can be reprocessed without pause in the regular workflow

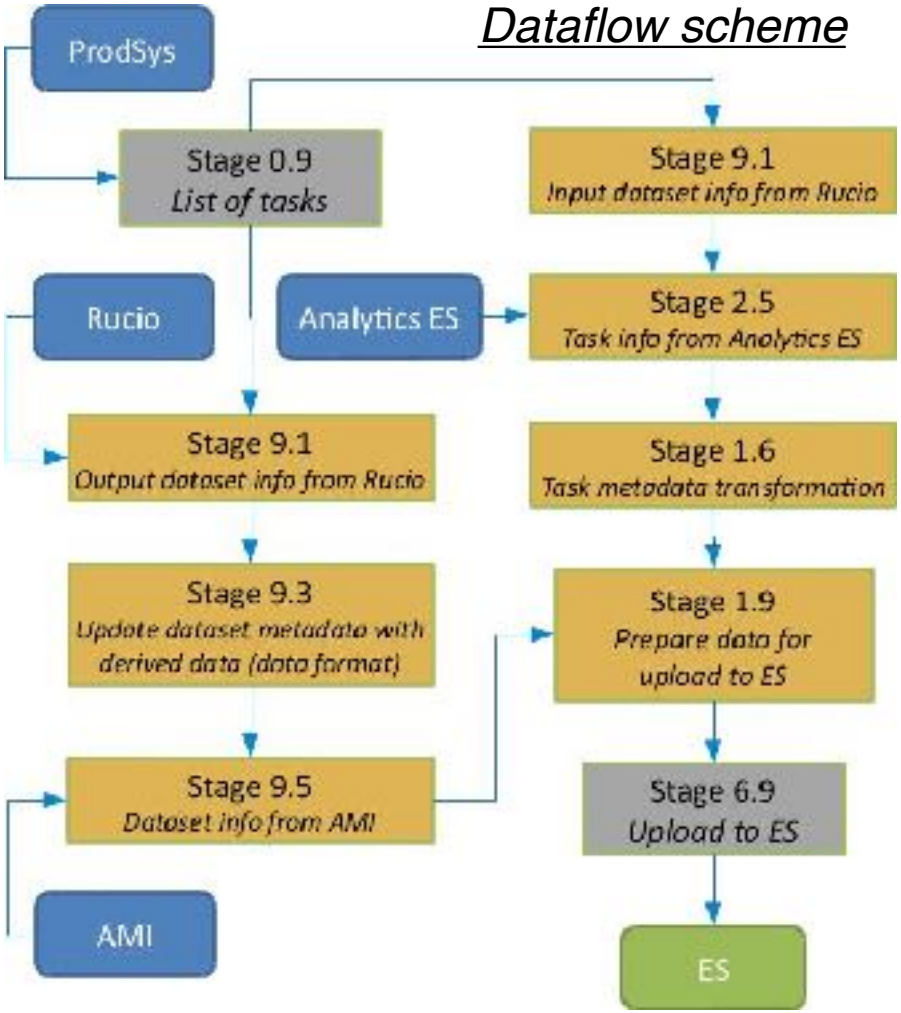
Storage populating / data reprocessing



Single thread processing:

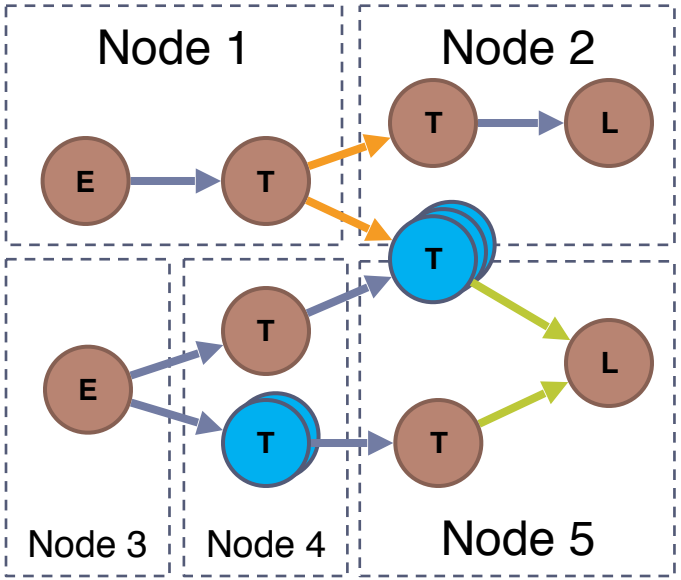
- hourly update: ~2 min
- reprocessing of 1 month: ~20 hrs

Dataflow scheme



Dataflow capabilities

- Arbitrary integration topology
  - **Splitting** in two identical data flows
  - **Joining** of similar data flows
- Scalability (provided by Apache Kafka)
  - Distributed processing
  - **Configurable parallelization**
- Multi-language modules support

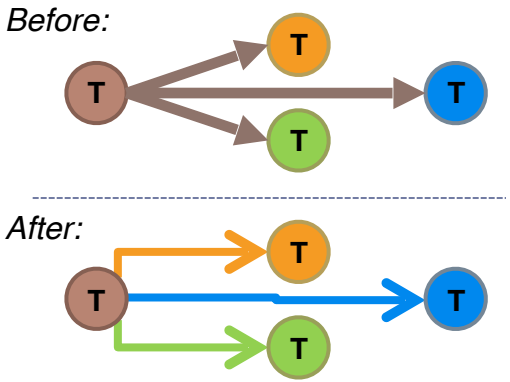


Summary

- ✓ DKB provides flexible mechanism for metadata integration, that allows to integrate data from sources of arbitrary number and type
- ✓ With this mechanism we have already addressed some user needs, providing information from Production System (DEFT, JEDI), Rucio, AMI and analytics cluster in the University of Chicago by single request to the DKB
- ✓ Extension and modification of the existing integration dataflow can be performed in the “continuous integration” way

Routes forward

- WEB-interface views:
  - ✓ pre-constructed
  - adjustable
- Improvement of the operating workflow mechanism
- New dataflow functionality
  - Data flow routing
    - Allow single processor to produce output messages of different types
    - Avoid reduplication of data flow on splitting



Glossary

- Indico: tool for conferences, workshops and meetings management (https://indico.cern.ch)
- Twiki: wiki pages (https://twiki.cern.ch)
- Rucio: Distributed Data Management system (https://rucio.cern.ch)
- AMI: ATLAS Metadata Interface (https://ami.in2p3.fr)
- Production System: top level workflow manager
  - **Related talk**: *The ATLAS Production System Predictive Analytics service: an approach for intelligent task analysis* by Mikhail Titov
- PanDA: workload management system (Production ANd Distributed Analysis)
  - **Related talk**: *BigPanDA Experience on Titan for the ATLAS Experiment at the LHC* by Alexei Klimentov
- BigPanDA: ATLAS PanDA monitor (https://bigpanda.cern.ch)
  - **Related talk**: *The BigPanDA monitoring system architecture* by Tatiana Korchuganova