Metadata Curation and Integration in High Energy and Nuclear Physics

Maria Grigorieva, Marina Golosova , Alexander Alexeev, Maxim Gubin, Vasiliy Aulov, Alexei Klimentov and Torre Wenaus

National Research Center "Kurchatov Institute"
Tomsk Polytechnic University
Brookhaven National Laboratory

Data Knowledge Base Highlights

DKB Basic Consideration

- Organizing metadata in ATLAS, so as to provide a holistic view on physics topics, including integrated representation of all ATLAS documents (papers, drafts, supporting documents, conference notes, Indico meetings, Twiki pages, etc) and corresponding data samples.

DKB Evolution

- The most important reports and summaries are made in ***Twiki pages*** (collaborative documentation) in semi-manual mode

- The metainformation in Twiki doesn't provide mechanisms for synchronization with database back-ends.
  - Provide *fully automatic* metadata search and aggregation by arbitrary set of parameters, synchronized with the existing database backends.
  - Index metadata and provide a *quick and flexible* google-like metadata *search*, categorization and aggregation.

ATLAS Metadata Sources

- Public Results
  - CERN Document Server
  - CERN Twiki
  - Indico
  - GLANCE (Papers and ConfNotes)
- Data Management and Analysis
  - Production System Database
  - ATLAS Metadata Interface (AMI)
  - Rucio DDM
  - Google Docs
  - JIRA
  - ATLAS SVN Repository
  - BigPanDA Monitoring System
- Physics and Experiment Environment
  - ATLAS Geometry database
  - ATLAS Conditions Database
  - ATLAS Twiki Pages

ATLAS Metadata Issue

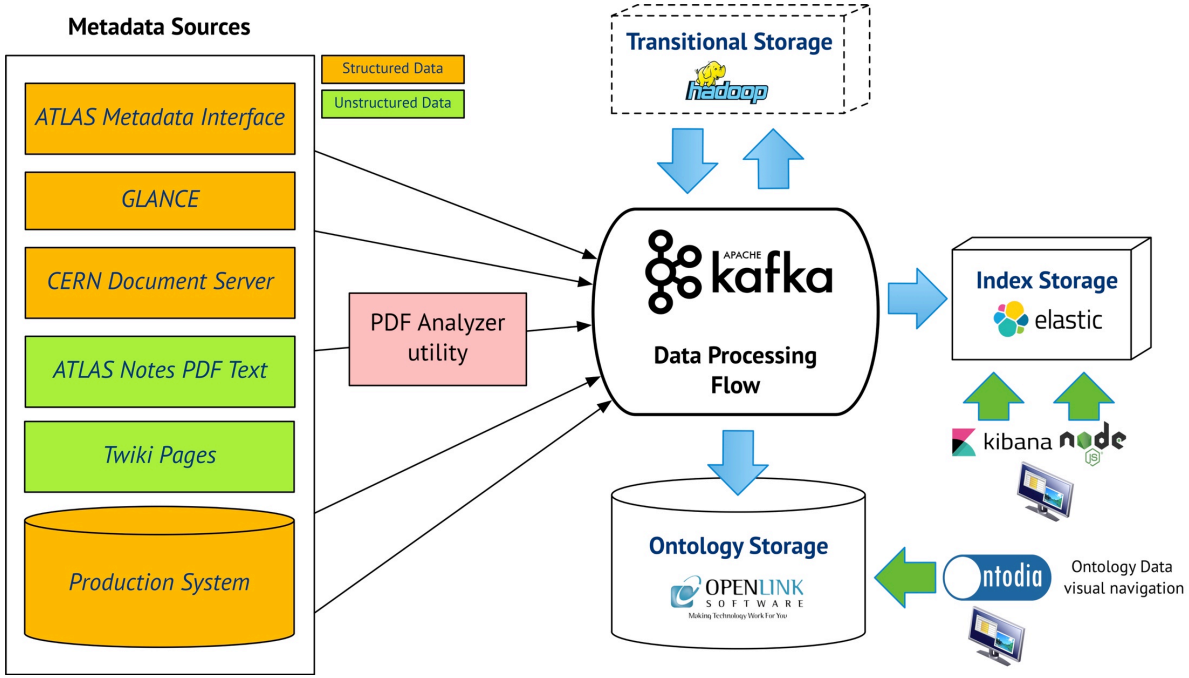
Most of these metadata sources exist *autonomously*. To gain the comprehensive information of research study, including experimental environment, data samples, and available results, scientists need to obtain *intersections among metadata* from different sources by themselves, as there are *no automatic tools* providing data integration.

Due to the complexity of modern HENP experiments, this task becomes more and more challenging.

DKB is aiming to fill this gap in meta-data integration.

DKB program code on GitHub:
https://github.com/PanDAWMS/dkb

DKB Architecture Prototype

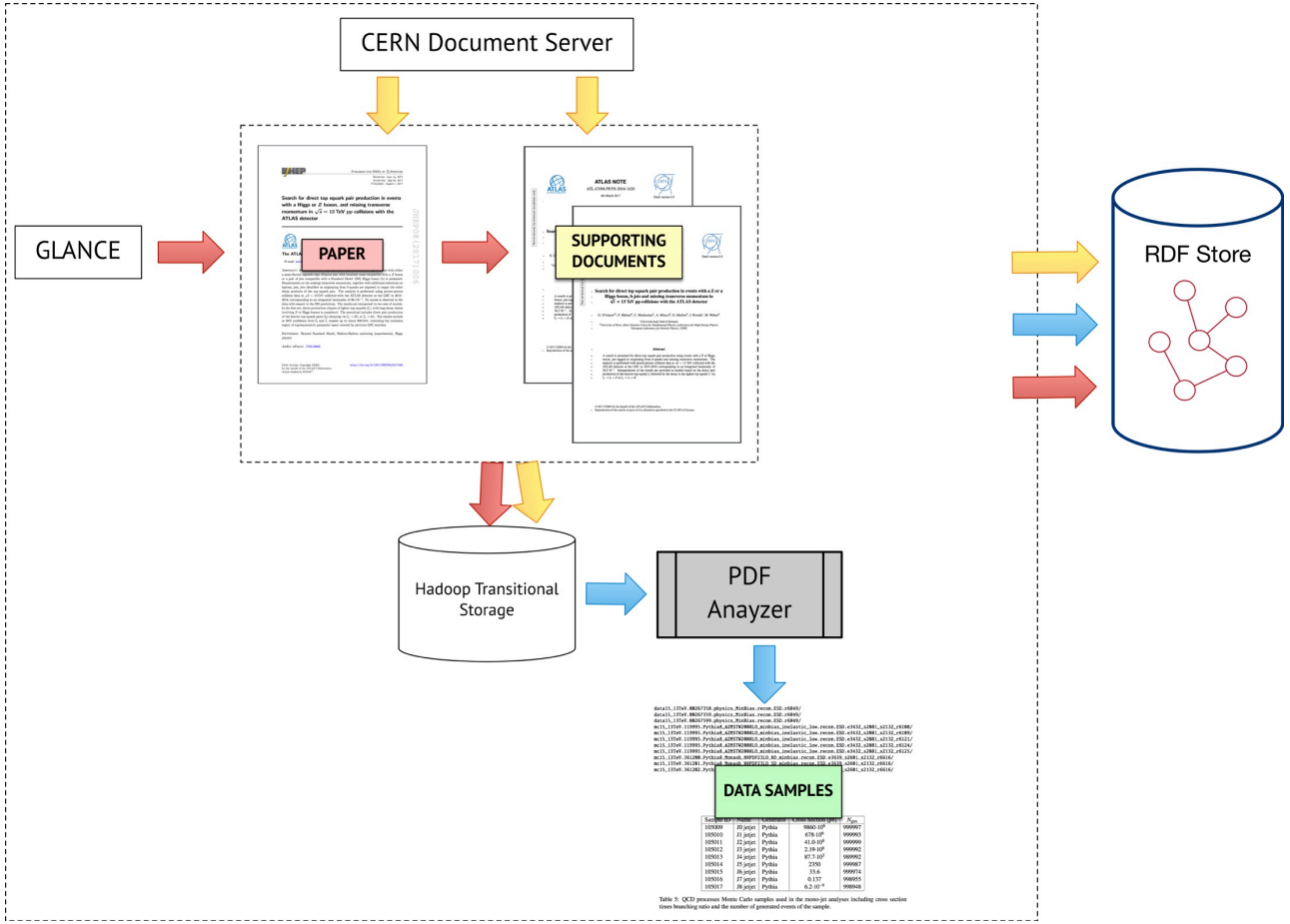


Technology evaluation and system prototype architecture:

- ✓ RDF storage: **Virtuoso**
- ✓ Transitional storage: **Hadoop**
- ✓ Metadata streaming&processing: **Apache Kafka**
- ✓ Knowledge Base navigation (Web GUI): **Ontodia**
- ✓ Index storage: **ElasticSearch**
- ✓ Index search web-interfaces: **Kibana, NodeJS**
- ✓ Documents metadata mining: **PdfMiner utility and implemented PDF Analyzer tool**

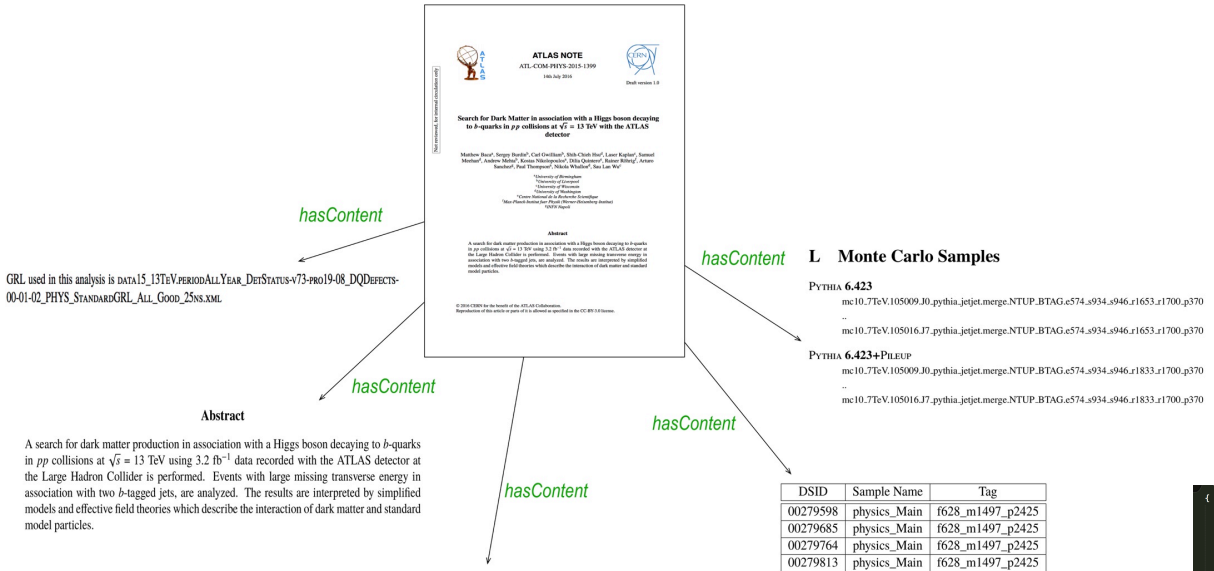
Basic programming language - python™

Initial Data Processing Flow



PDFAnalyzer tool extracts metadata from TXT and XML representation of PDF text, by regular expressions and context analysis.

Data Extraction from PDF Documents



Abstract

A search for dark matter production in association with a Higgs boson decaying to b-quarks in pp collisions at $\sqrt{s} = 13$ TeV using 3.2 fb^{-1} data recorded with the ATLAS detector at the Large Hadron Collider is performed. Events with large missing transverse energy in association with two b-tagged jets, are analyzed. The results are interpreted by simplified models and effective field theories which describe the interaction of dark matter and standard model particles.

DSID	Sample Name	Tag
341100	Pythia8EvtGen_A14NNPDF23LO_WlvH125_bb	e3885_s2608_s2183_r6869_r6282_p2419
341101	Pythia8EvtGen_A14NNPDF23LO_ZvvH125_bb	e3885_s2608_s2183_r6869_r6282_p2419
341102	Pythia8EvtGen_A14NNPDF23LO_ZllH125_bb	e3885_s2608_s2183_r6869_r6282_p2419

Table 19: Monte Carlo samples used as baseline for Standard Model VH($\rightarrow bb$).

DSID	Sample Name	Tag
00279598	physics_Main	f628_m1497_p2425
00279685	physics_Main	f628_m1497_p2425
00279764	physics_Main	f628_m1497_p2425
00279813	physics_Main	f628_m1497_p2425
00279867	physics_Main	f628_m1497_p2425
00279928	physics_Main	f628_m1497_p2425
00279932	physics_Main	f629_m1504_p2425
00279984	physics_Main	f629_m1504_p2425
00280231	physics_Main	f630_m1504_p2425
00280319	physics_Main	f629_m1504_p2425
00280368	physics_Main	f629_m1504_p2425
00280423	physics_Main	f629_m1504_p2425
00280464	physics_Main	f629_m1504_p2425
00280500	physics_Main	f631_m1504_p2425
00280520	physics_Main	f632_m1504_p2425
00280614	physics_Main	f629_m1504_p2425

PDFAnalyzer extracts:

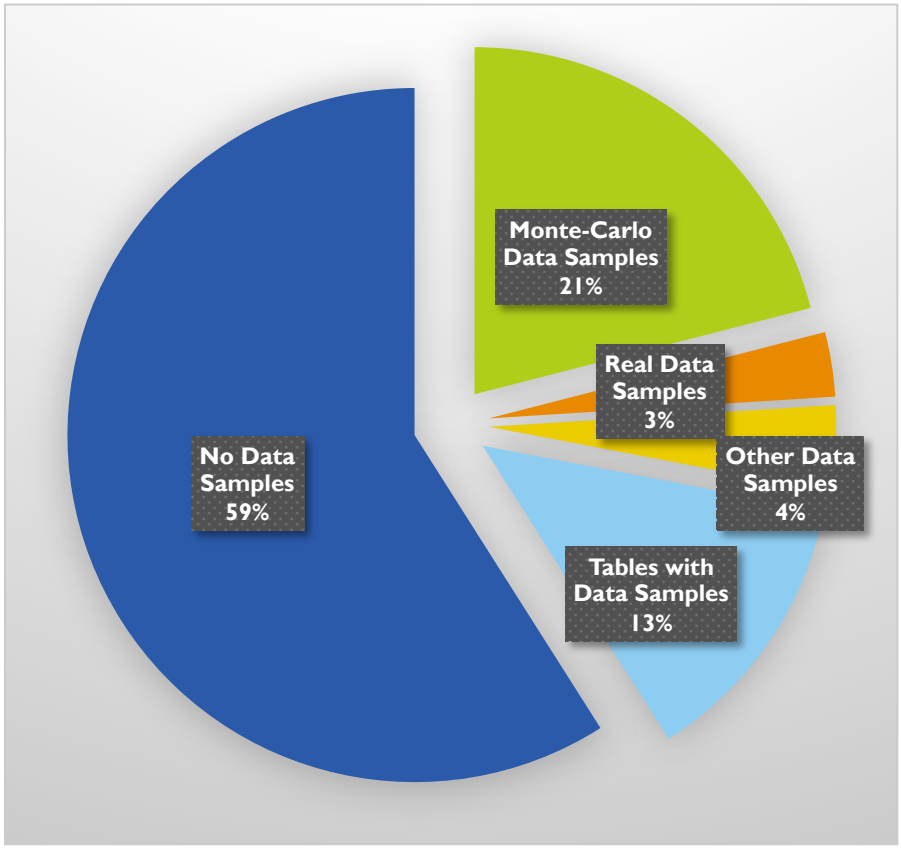
- ❑ dataset names by regular expression
- ❑ datasets metadata from tables
- ❑ experiment-specific metadata from text

Returns structured metadata in ***JSON.***

Has GUI interface, providing manual correction of analysis results


```
{
    "content": {
        "real_datasets": [
            "data10_hi.00169175.physics_bulk.RDOT.NTUP_HI.r2114_r2768_p753",
            "data13_2p76TeV.00219257.physics_MinBias.merge.NTUP_HI.f519_m1313",
            "data13_2p76TeV.00219263.physics_MinBias.merge.NTUP_HI.f519_m1313",
            "data13_2p76TeV.00219296.physics_MinBias.merge.NTUP_HI.f519_m1313",
            "data13_2p76TeV.00219305.physics_MinBias.merge.NTUP_HI.f519_m1313",
            "data13_2p76TeV.00219322.physics_MinBias.merge.NTUP_HI.f519_m1313",
            "data13_2p76TeV.00219364.physics_MinBias.merge.NTUP_HI.f519_m1313"
        ],
        "plain_text": {
            "atlas_name": "ATL-COM-PHYS-2016-XXX",
            "links": {},
            "data taking year": false,
            "luminosity": false,
            "energy": "2.76 TeV",
            "campaigns": [
                "mc12",
                "mc11"
            ],
            "collisions": "proton-proton",
        },
        "mc_datasets": [
            "mc10_2TeV.119114.Hijing_PbPb_2p75TeV_MinBias_Flow_JJFV6.recon.NTUP_HI.e1046_s1160_s1161_d490_r2134",
            "mc12_2TeV.209036.Pythia_AUET2BCTEQ6L1_minbias_inelastic.recon.NTUP_HI.e2812_s1647_s1586_r6309_tid06163221_00"
        ]
    },
    "fname": "CDS_CERN-ATL-COM-PHYS-2016-257"
}
```

PDF Analyzer Statistics

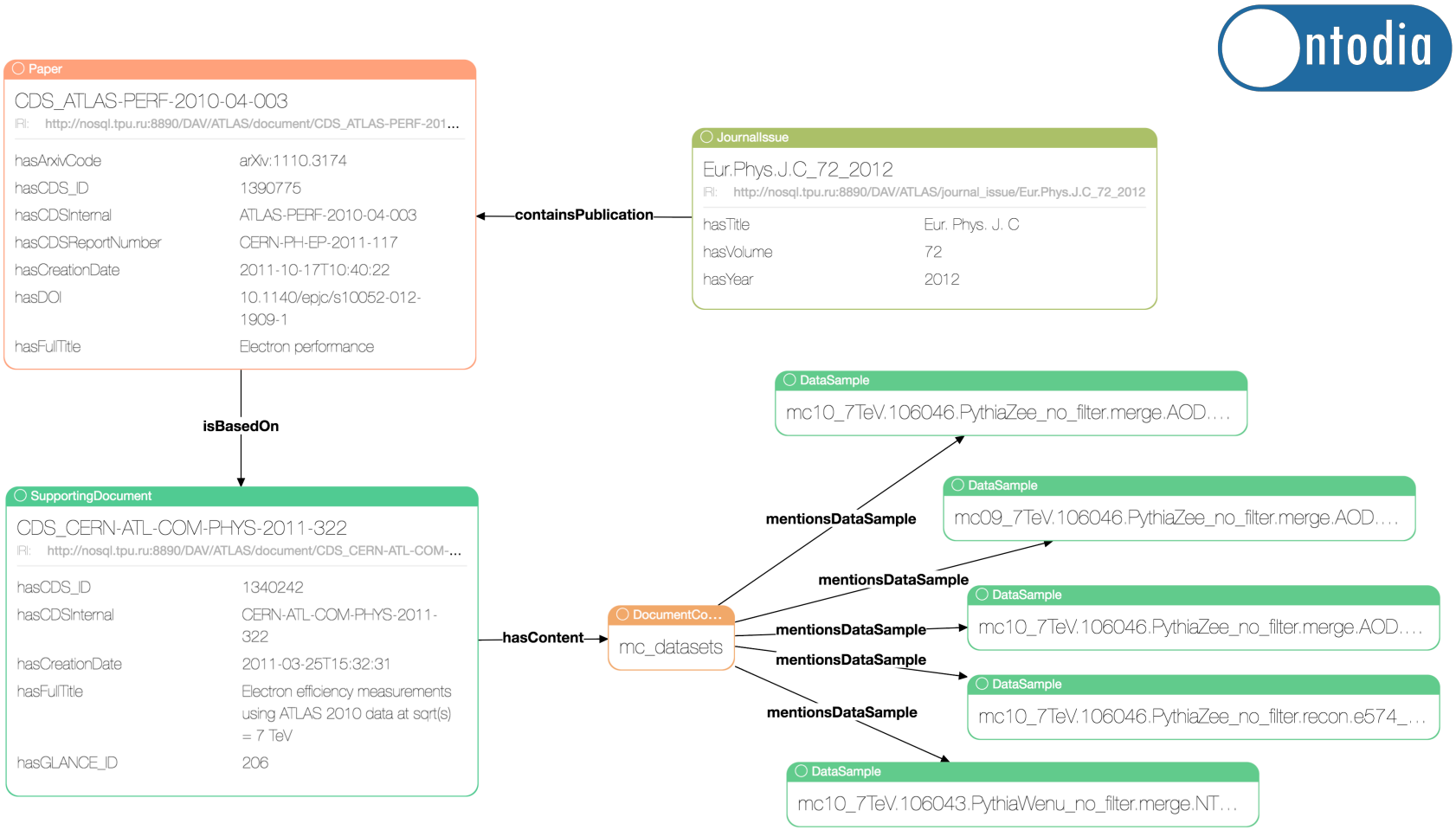


The statistics of ATLAS Internal Notes (500 documents) analysis showed that ~40% of documents allow to extract data samples automatically.

Why we can't extract data samples for for 60% of documents?

- ✓ No datasets in the text.
- ✓ Information about datasets is presented as instructions for human reader.
- ✓ Not all Papers have properly defined Supporting Documents.

Ontodia - JavaScript library for Virtuoso navigation



Metadata Indexing and Search Facilities

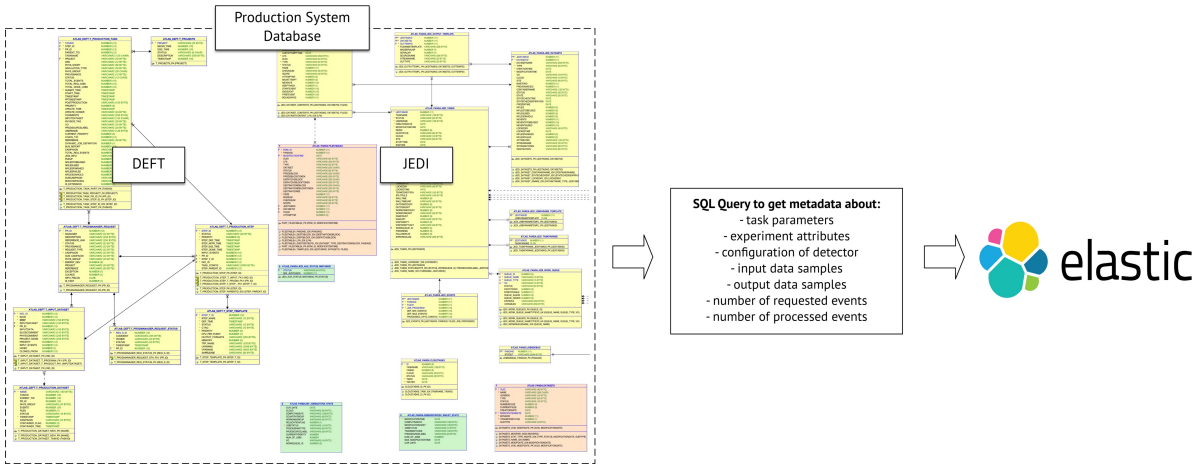
- To provide fast and flexible data categorization, search and aggregation

1. Reproduce Event Summary report,
   but with physics category
   breakdown (like in Twiki Event Summary).

&MC16c_CP

Step Name % Processed/Input events running/pending/finished	Evgen 99.8%	Evgen Merge 96.3%	Simul 97.26%	Merge 100.84%	Reco 99.76%	Rec Merge 99.91%
	1,913,235,610 / 1,956,298,240 0.0%/0.0%/97.8%	1,236,857,070 / 1,284,022,270 0.0%/0.0%/96.33%	1,934,968,180 / 1,989,539,330 0.49%/0.0%/97.26%	313,326,050 / 310,709,550 0.08%/0.0%/100.84%	1,925,870,230 / 1,930,593,230 0.06%/0.0%/99.76%	1,914,226,230 / 1,915,887,230 0.03%/0.0%/99.91%

2. Implement google-like search of tasks and data samples by the arbitrary set of attributes, like campaign, project, ATLAS geometry, Condition Tags, hashtags, physics category, and others.



Metadata in ElasticSearch Index

Metadata categories:

- **Task Parameters**
  - Taskid
  - Taskname
  - Status
  - Timestamp
  - Start time
  - End time
  - Request ID
  - Ticket ID
  - User Name
- **Experiment parameters**
  - Energy
  - Campaign/Subcampaign
  - Project
  - Physics group
  - Physics category
  - Hashtags
  - Run number
- **Configuration**
  - ALTAS geometry
  - Conditions tags
  - SW Release
  - Trigger Config
- **Events**
  - Requested
  - Processed
- **Data Samples**
  - Input
  - Output

t	_id	12115125
t	_index	prodsys
#	_score	-
t	_type	MC16
t	architecture	x86_64-slc6-gcc62-op
t	campaign	MC16
t	conditions_tags	OFLCOND-MC16-SDR-20
t	core_count	8
t	description	MC16c PA sheet 1
⊙	end_time	September 12th 2017, 20:32:57.000
#	energy_gev	13,000
t	geometry_version	ATLAS-R2-2016-01-00-01
t	hashtag_list	MC16c_PA
t	input_datasets	mc16_13TeV:mc16_13TeV.341080.PowhegPythia8EvtGen_CT10_AZNLOCTEQ6L1_VBFH125_WWlvlv_EF_15_5.simul.HITS.e3871_e5984_s3126_tid12023009_00, mc16_13TeV:mc16_13TeV.361238.Pythia8EvtGen_A3NNPDF23LO_minbias_inelastic_low.simul.HITS.e4981_s3087_s3111/, mc16_13TeV:mc16_13TeV.361239.Pythia8EvtGen_A3NNPDF23LO_minbias_inelastic_high.simul.HITS.e4981_s3087_s3111/
t	output_datasets	mc16_13TeV.341080.PowhegPythia8EvtGen_CT10_AZNLOCTEQ6L1_VBFH125_WWlvlv_EF_15_5.recon.AOD.e3871_e5984_s3126_r9781_tid12115125_00
t	phys_category	Higgs
t	phys_group	MCGN
t	pr_id	13326
#	processed_events	16,750,000
t	project	mc16_13TeV
#	requested_events	40,242,000
t	run_number	341080
⊙	start_time	September 8th 2017, 21:56:56.000
t	status	done
t	step_name	Reco
t	subcampaign	MC16c
⊙	task_timestamp	September 14th 2017, 15:35:07.000
t	taskid	12115125
t	taskname	mc16_13TeV.341080.PowhegPythia8EvtGen_CT10_AZNLOCTEQ6L1_VBFH125_WWlvlv_EF_15_5.recon.e3871_e5984_s3126_r9781
t	ticket_id	ATLPSTASKS-1144711
t	trans_home	Athena-21.0.32
t	trans_path	Reco_tf.py
t	trans_uses	Atlas-21.0.3
t	trigger_config	RDOtoRDOTrigger=MCRECO:DBF:TRIGGERDBMC:2170,46,199
t	user_name	dsouth
t	vo	atlas

Google-like keyword search

Summaries in Kibana

TTbarX: Category

Step	Requested Events	Processed Events
Simul	3,200,000	3,200,000
Reco	442,512,000	170,637,000
Rec Merge	2,600,000	2,600,000
Merge	200,000	200,000
Evgen Merge	1,800,000	1,800,000

DrellYan: Category

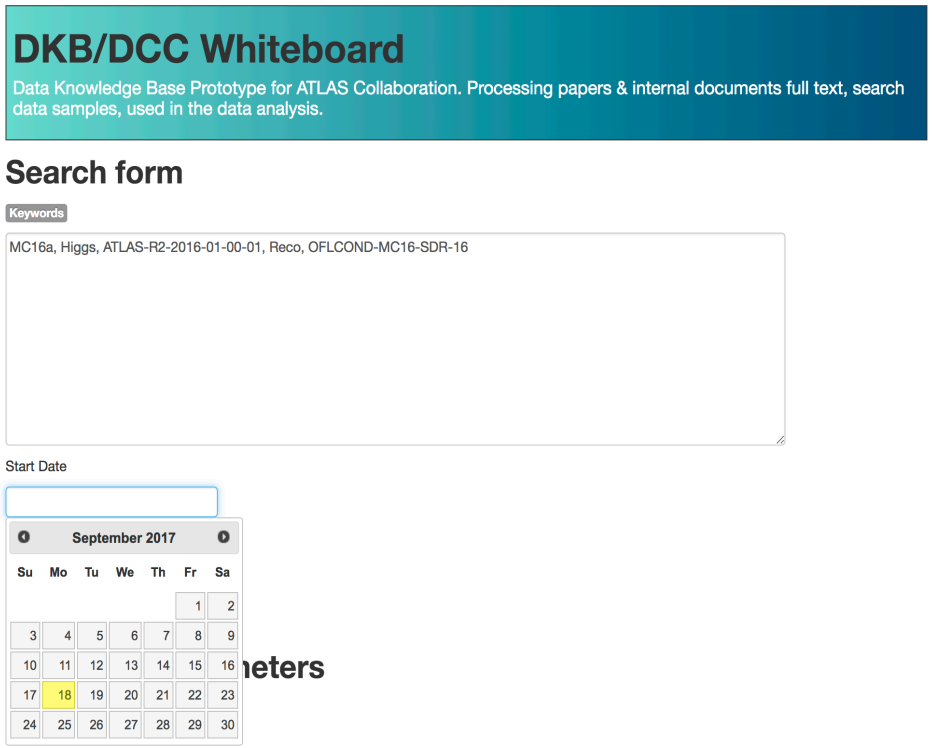
Step	Requested Events	Processed Events
Simul	194,913,900	194,913,900
Reco	12,444,403,000	11,579,977,000
Rec Merge	161,408,000	161,408,000
Merge	449,800	449,800
Evgen	2,202,000	2,202,000
Deriv Merge	176,500,000	176,500,000
Deriv	161,810,000	161,810,000

Higgs: Category

Step	Requested Events	Processed Events
Simul	106,328,100	106,328,100
Reco	9,628,753,000	5,850,766,000
Rec Merge	86,838,000	86,838,000
Evgen Merge	138,399,700	138,399,700
Evgen	38,570,400	38,570,400
Deriv Merge	82,000,000	82,000,000
Deriv	77,510,000	77,510,000

```
"query": {
  "bool": {
    "must": [
      { "term": { "subcampaign.keyword": "MC16a" } },
      { "term": { "status": "done" } }
    ],
    "should": [
      { "term": { "hashtag_list": "MC16a"} },
      { "term": { "hashtag_list": "MC16a_CP"} }
    ]
  }
},
"aggs": {
  "category": {
    "terms": {"field": "phys_category"},
    "aggs": {
      "step": {
        "terms": {"field": "step_name.keyword"},
        "aggs": {
          "requested": {
            "sum": {"field": "requested_events"}
          },
          "processed": {
            "sum": {"field": "processed_events"}
          }
        }
      }
    }
  }
}
```

Web-interface prototype




```
GET prodsys/MC16/_search
{
  "query": {
    "bool": {
      "must": {
        "query_string": {
          "query":“"MC16a” AND “Higgs” AND  “ATLAS-R2-2016-01-00-01” AND “Reco” AND “OFLCOND-MC16-SDR-16””,
          "analyze_wildcard": true
        }
      },
      "filter": {
        "range": {
          "task_timestamp": {
            "gte": "01-05-2017 00:00:00",
            "lt": "10-08-2017 00:00:00",
          }
        }
      }
    }
  }
}
```

DKB/DCC Whiteboard

Data Knowledge Base Prototype for ATLAS Collaboration. Processing papers & internal documents full text, search data samples, used in the data analysis.

Dataset Search

"MC16a" AND "Higgs" AND "ATLAS-R2-2016-01-00-01" AND "Reco" AND "OFLCOND-MC16-SDR-16"

mc16_13TeV.345318.PowhegPythia8EvtGen_NNPDF30_AZNLO_WpH125J_Hyy_Wincl_MINLO.recon.e5734_e5984_s3126_r9364

TASK		EXPERIMENT		CONFIGURATION	
taskID	11828859	**Campaign**	MC16	**Step Name**	Reco
taskName	mc16_13TeV.345318.PowhegPythia8EvtGen_NNPDF30_AZNLO_WpH125J_Hyy_Wincl_MINLO.recon.e5734_e5984_s3126_r9364	**Subcampaign**	MC16a	**ticket_id**	ATLPSTASKS-1108307
status	running	**Project**	mc16_13TeV	**Architecture**	x86_64-slc6-gcc62-op
Description	MC16a PA sheet 0	**Energy**		**Core Number**	
timestamp	31-08-2017 07:17:56	**Physics Group**	MCGN	**ATLAS Geometry**	ATLAS-R2-2016-01-00-01
start time	06-08-2017 21:24:08	**Physics Category**	Higgs	**Conditions Tags**	OFLCOND-MC16-SDR-16
end time		**Hashtags**	MC16a_PA	**trigger_config**	RDOtoRDOTrigger=MCRECO:DBF:TRIGGERDBMC:2136,35,16
Duration				**trans_path**	Reco_tf.py
EVENTS Requested / Processed / Done(%) **40242000 / 10530000 / 26%**				**trans_path**	AtlasOffline-21.0.20
				run_number	345318

DATASETS	
Input Datasets	mc16_13TeV:mc16_13TeV.345318.PowhegPythia8EvtGen_NNPDF30_AZNLO_WpH125J_Hyy_Wincl_MINLO.simul.HITS.e5734_e5984_s3126_tid11828856_00 mc16_13TeV:mc16_13TeV.361238.Pythia8EvtGen_A3NNPDF23LO_minbias_inelastic_low.simul.HITS.e4981_s3087_s3111/ mc16_13TeV:mc16_13TeV.361239.Pythia8EvtGen_A3NNPDF23LO_minbias_inelastic_high.simul.HITS.e4981_s3087_s3111/
Output Datasets	mc16_13TeV.345318.PowhegPythia8EvtGen_NNPDF30_AZNLO_WpH125J_Hyy_Wincl_MINLO.recon.AOD.e5734_e5984_s3126_r9364_tid11828859_00

mc16_13TeV.345433.PowhegPythia8EvtGen_NNPDF3_AZNLO_WmH125J_MINLO_qqWWlvlv.recon.e5811_e5984_s3126_r9364

mc16_13TeV.345326.PowhegPythia8EvtGen_NNPDF3_AZNLO_WmH125J_MINLO_lvWWlvlv.recon.e5823_e5984_s3126_r9364

mc16_13TeV.345337.PowhegPythia8EvtGen_NNPDF3_AZNLO_ZH125J_MINLO_llWWlvlv.recon.e5810_e5984_s3126_r9364

mc16_13TeV.341429.PowhegPythia8EvtGen_CT10_AZNLO_WpH125J_MINLO_eveWWlvlv.recon.e4210_e5984_s3126_r9364

mc16_13TeV.345319.PowhegPythia8EvtGen_NNPDF30_AZNLO_ZH125J_Hyy_Zincl_MINLO.recon.e5743_e5984_s3126_r9364

mc16_13TeV.341433.PowhegPythia8EvtGen_CT10_AZNLO_WpH125J_MINLO_qqWWlvlv.recon.e3938_e5984_s3126_r9364

What's been done:

- Developed the method and implemented modules, providing automatic extraction of the data samples from ATLAS documents
  - Ontological data model for ATLAS documents, data samples and Experiment Attributes
  - Tools for metadata extraction/processing/convertation/importing
  - PDF Analyzer, providing metadata extraction from PDF documents
  - Kafka-based automation of dataflows execution
  - Virtuoso database was filled with metadata
  - Ontodia as GUI for Virtuoso navigation
- Provided fast and flexible data categorization and search
  - Production System database partly (MC16 campaign) indexed in ElasticSearch
  - ElasticSearch and Kibana infrastructure is installed at CERN
  - Kibana dashboard with Event Summary report, and a set of diagrams
  - Implemented NodeJS-based web-inteface for ElasticSearch Storage

Future Plans

- DKB software transferring to CERN machines
- Synchronization between Production System database and ElasticSearch storage
- Performance tests of Production System and ElasticSearch storage:
  - Aggregations
  - Search by arbitrary set of parameters
- Web Interface for ElasticSearch storage
- Improve the mechanism of the Supporting Notes search, using CERN Document Server Publications history
- Improved Papers&Supporting Documents search results will be tested on neo4j or OrientDB graph database

Acknowledgment

- The work was supported by Russian Foundation for Basic Research (RFBR) under contract No. 16-37-00246.

Thanks

- Siarhei Padolski,
- Michail Borodin,
- Dimitry Krasnopevtsev,
- Dmitry Golubkov,
- Anastasia Kaida