Data Analytics in ATLAS

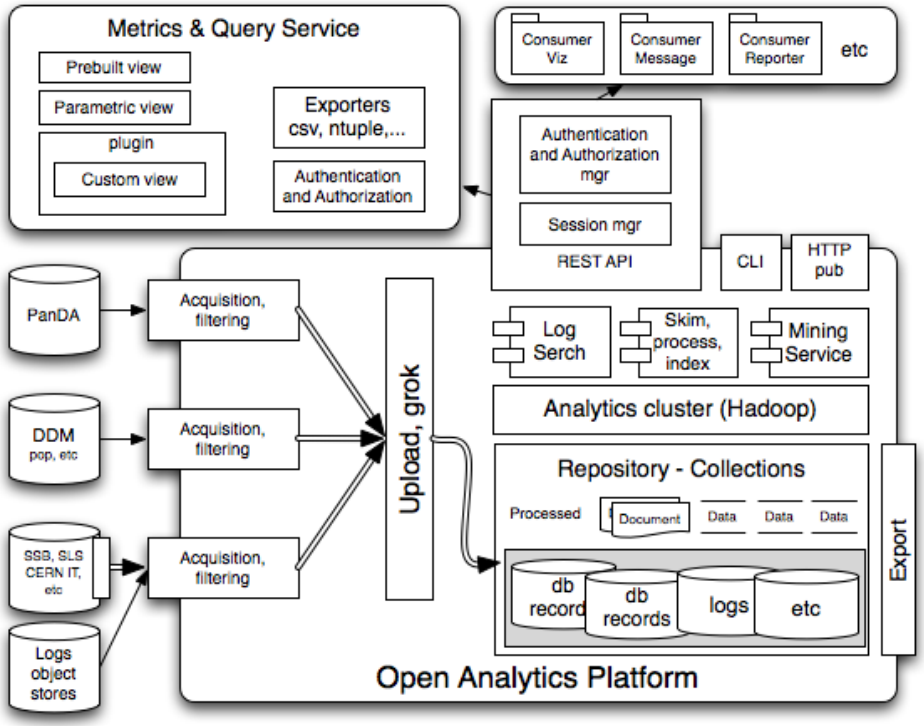
Ilija Vukotic • University of Chicago

NEC 2015, Bečići, Montenegro

Architecture

Main functions

- Acquisition, filtering and upload of data sources into a repository
- Hadoop cluster for analysis of multiple data sources to create reduced collections for higher level analytics
- Serve repository collections in multiple formats to external clients
- Makes collected sources available for export by external users
- Host analytics services on the platform such as ElasticSearch, Logstash, Kibana, etc.



Data sources

PanDA - a data-driven workload management system for production and distributed analysis processing

Rucio - a Distributed Data Management system used to manage accounts, files, datasets, and distributed storage systems.

FAX - Federated ATLAS storage system using XRootD protocol. Provides a global namespace, direct access to data from anywhere.

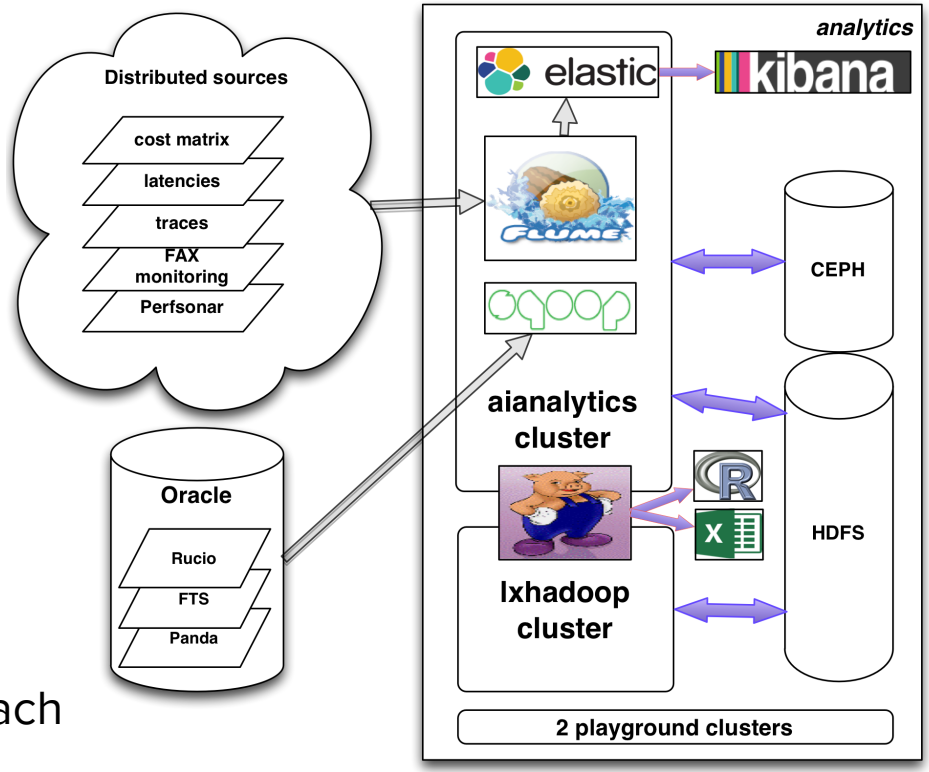
PerfSONAR - a widely-deployed test and measurement infrastructure that is used by science networks and facilities around the world to monitor and ensure network performance.

FTS - File Transfer Service - the lowest-level data movement service doing point-to-point file transfers.

xAOD - primary analysis data product.

Analytics Platform: Resources & Sources

- ***lxhadoop*** cluster
  - runs base load map-reduce jobs
- ***voatlasanalytics*** cluster
  - 5 VM nodes
  - sees the same data as lxhadoop
  - HDFS IO operations
  - runs Flume collectors
  - runs Sqoop jobs
  - runs ElasticSearch
  - runs Kibana
- "Playground" clusters: 3 small VMs each



Supporting Map Reduce & Search

Hadoop-based collections

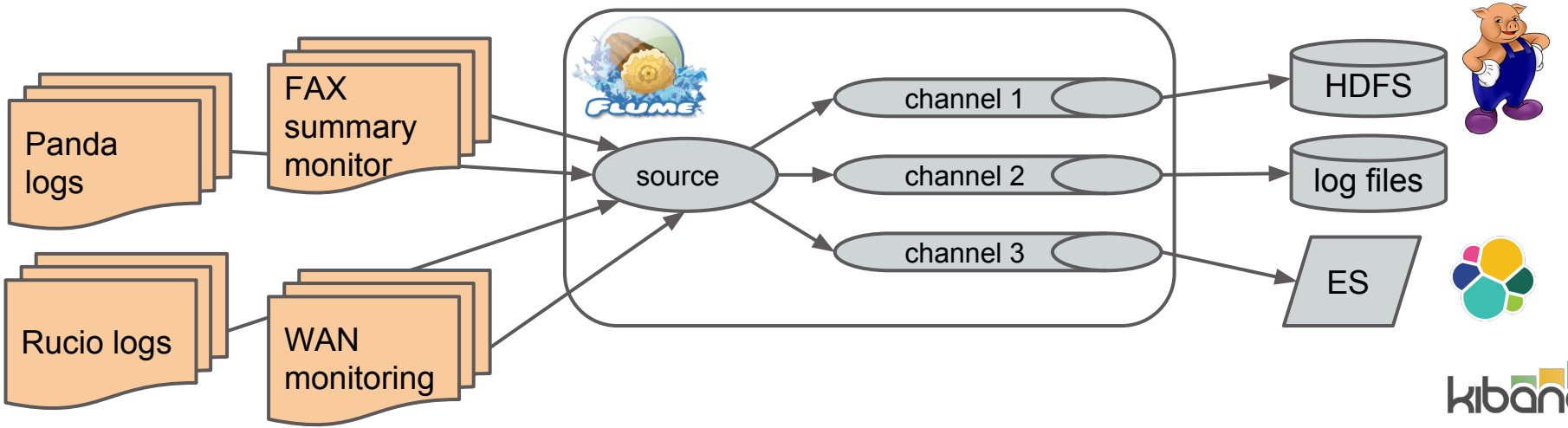
1. PanDA Job Archive (1TB)
2. PanDA State change logs (0.5TB)
3. PanDA Logs (16 GB)
4. FAX cost matrix, traces (2GB)
5. Rucio (42 TB)
6. Network data(2-3 GB/hour)

Elastic Search Indices

1. PanDA Job Archive (Sqoop import)
2. PanDA State change logs (pig reprocessing + import)
3. Fax redirectors monitoring (Flume)
4. Rucio logs imports (streamed using Logstash)
5. PanDA Logger (Flume)
6. Network Data (AMQ + Flume)

Central Flume Collector

- Listens for JSON messages (from multiple WLCG or ATLAS services)
- “events” multiplexed into different memory channels based on header content, and sent to log files and/or HDFS for analysis and/or ElasticSearch for indexing
- Currently collects from the CostMatrix service, traces, & FAX redirector summaries



elastic @ CERN

1 x head node:

- 2 core VM
- 4 GB RAM
- 1 TB storage

4 x data node:

- 8 cores VM
- 16 GB RAM
- 1 TB storage CEPH io1

Authentication through SSO

Runs on ***voatlasanalytics*** cluster:

- No in-box storage.
- CEPH duplicates data on top of ES sharding.
- Not enough memory/core.

New hardware expected.

ES will be offered as a service at CERN.

elastic @ THE UNIVERSITY OF CHICAGO

2 x head nodes:

- 1 core VM
- 4 GB RAM
- 1 TB storage

4 x data nodes:

- 48 GB RAM
- 24 cores (HT)
- 3 TB storage

authentication - username/password

- Also used for other MWT2 analytics.
- Currently more than 1.2 billion docs in 3300 shards.

Contains:

- PanDA job archive
- FAX costs
- FAX redirectors monitoring

elastic @ CloudLab

1 head node
5 data nodes

- 20 cores per node (2 CPUs)
- 256 GB RAM
- 2x10 Gb/s Ethernet card
- 40 Gb/s Infiniband
- 2x1 TB disk drives in each node.

Very easy to add more nodes.
Can changed for nodes with
8x1 TB disks + 12x4TB

simple firewall protection

- CloudLab is flexible, scientific infrastructure for research on the future of cloud computing.
- Provides control and visibility all the way down to the bare metal.
- Provisioning an entire cloud inside of CloudLab takes only minutes.

Sites:

- Utah / HP
- Wisconsin / Cisco
- Clemson / Dell
- GENI

Still testing a cluster at Clemson University. Great initial results

Analytics Infrastructure - lessons learned

- Need really good hardware - lot of RAM, SSD caching, large disks
- ES backups
  - while most of the data could be re-indexed, some exists only in ES.
  - the Shield - password protection, role-based access control, very expensive (6k$/host/year)
  - small but important indices (.kibana) easily backed up every night at Amazon S3.
- Very important to develop and well document pig UDFs for the different analysis needs.
- It is clear that Kibana accessible data have much more use. Try to index as much as possible of the hadoop data.
- Non-negligible learning curve (MR, pig, java, jython) - need a lot of documentation, support, education.

Covered use cases

- Rucio
  - Error monitoring
  - Activity tracking
  - Tracking a file/dataset
- Usage of beyond-pledge resources
- Per cloud performance metrics
- Data formats popularity
- xAOD usage monitoring / analysis
- FAX monitoring jobs accessing data over WAN
- FAX redirectors monitoring
- Monitoring local data storage resources (MWT2)

Use cases still to be covered

- WAN performance analysis
- Network weather service
- RTT jobs monitoring
- PanDA task duration analysis
- Geant production log analysis
- Addition to bigpanda (PanDA web frontend)- replace the slowest Oracle searches

This list grows faster and faster…

Immediate questions to answer

- is derivation production successful?

- do people run private filtering on AODs?

- how much private production the users run on their own?

- do majority of users use DAODs?

- does a small subset of users consume most of the cpus for analysis which is not in our computing model?

- how many cpus are needed to make 95% of users happy - i.e. to run analysis on DAODs?

- how many users would like to run parallel or complex jobs (eg, high memory, task chains...)?

- is the current system (PanDA) good enough to give a high throughput to DAOD analysis and prevent the users running non-standard jobs to take over the resources?

A network weather service for ATLAS

We will use ES as a Network weather service.

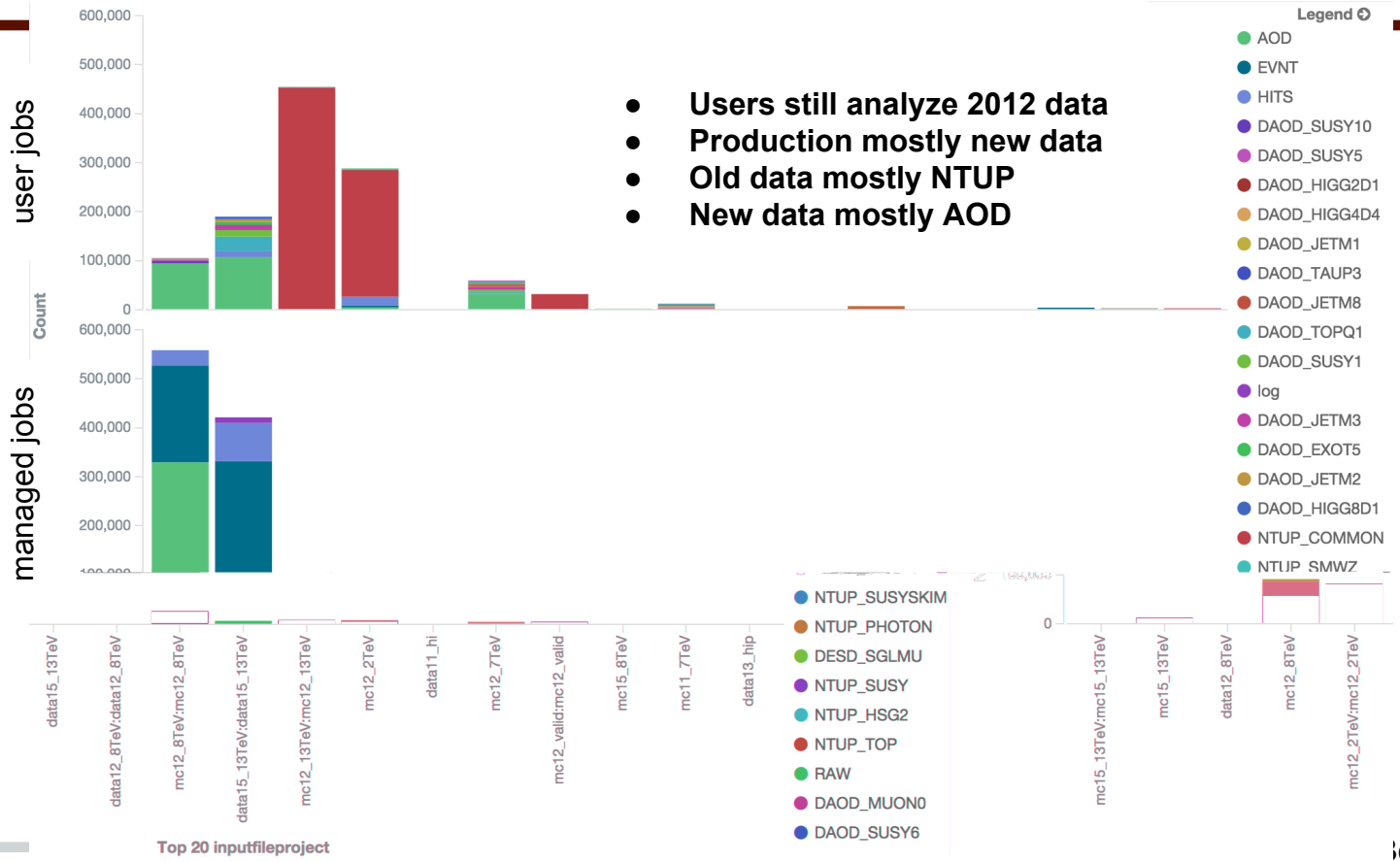
Data collection and warehousing

- throughput, latencies, and packet loss data come from OSG network datastore, Flum-ed from AMQ
- in-line prediction of future network performance
- FAX cost data already in Hadoop, will be indexed too

Serving the data:

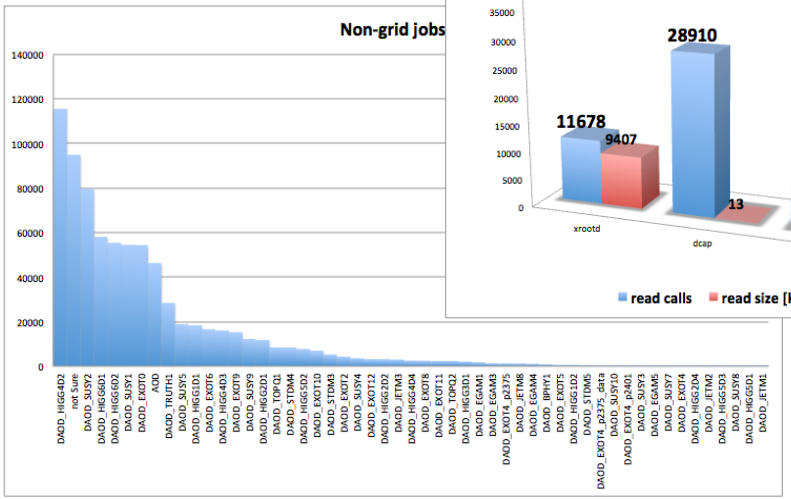
- very simple REST interface
- very simple python API
- ES should be able to deliver searches in <100ms @ 100Hz

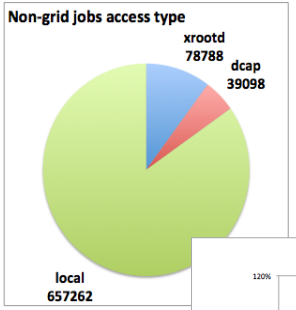
File formats popularity

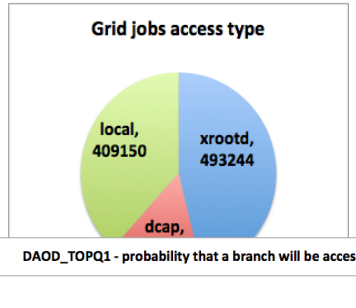


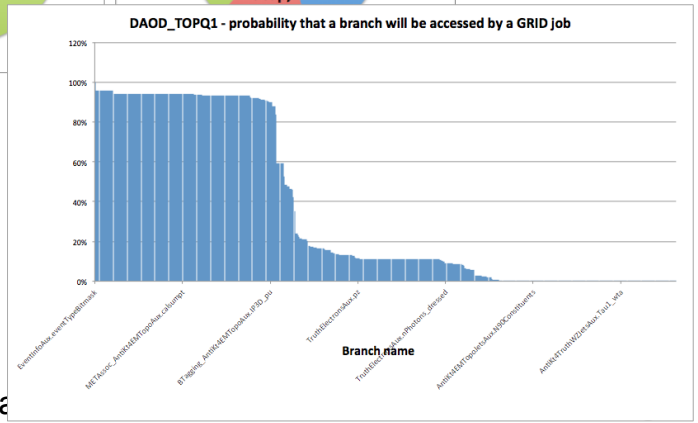
- Users still analyze 2012 data
- Production mostly new data
- Old data mostly NTUP
- New data mostly AOD

xAOD accesses analysis

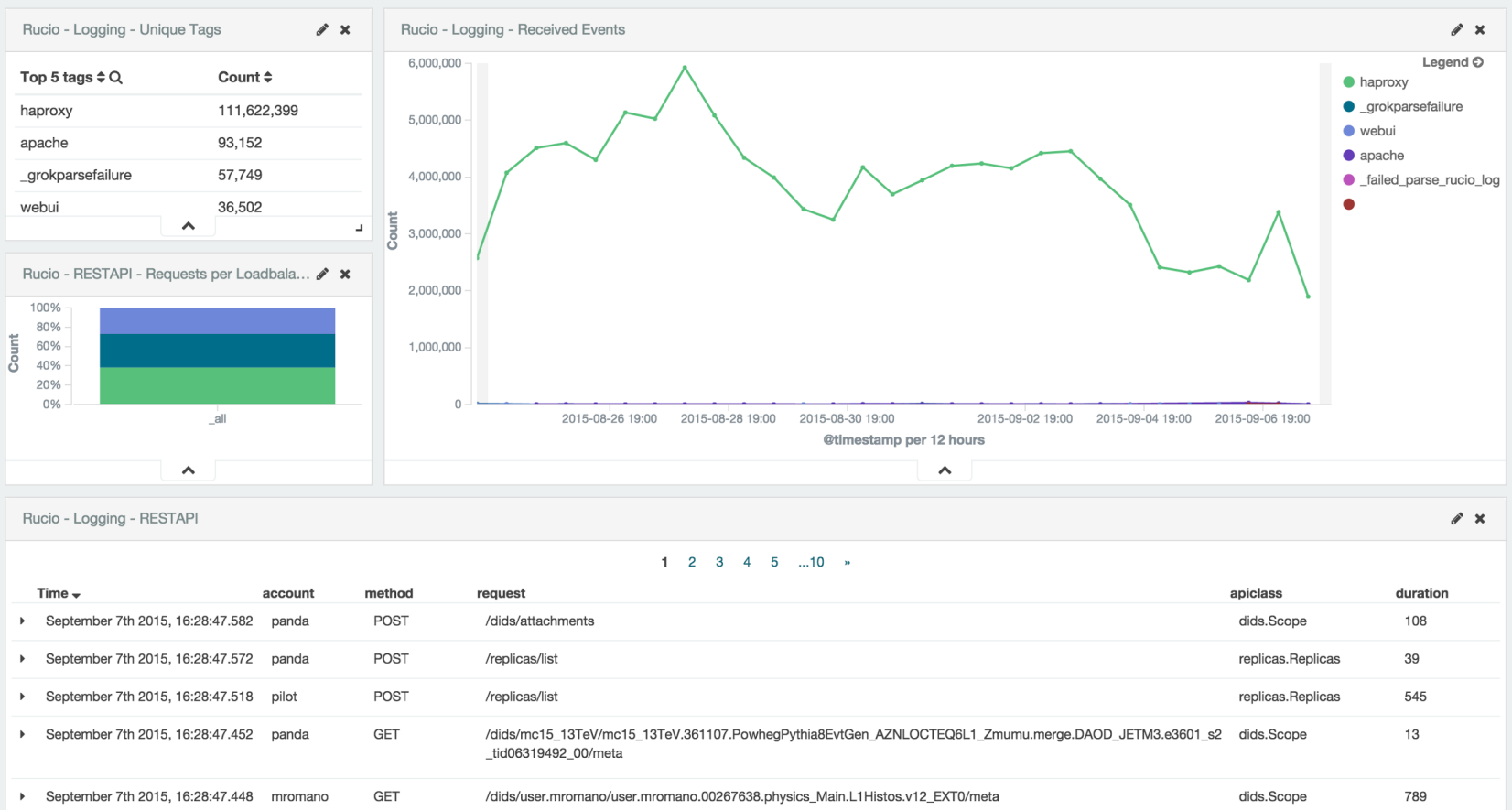




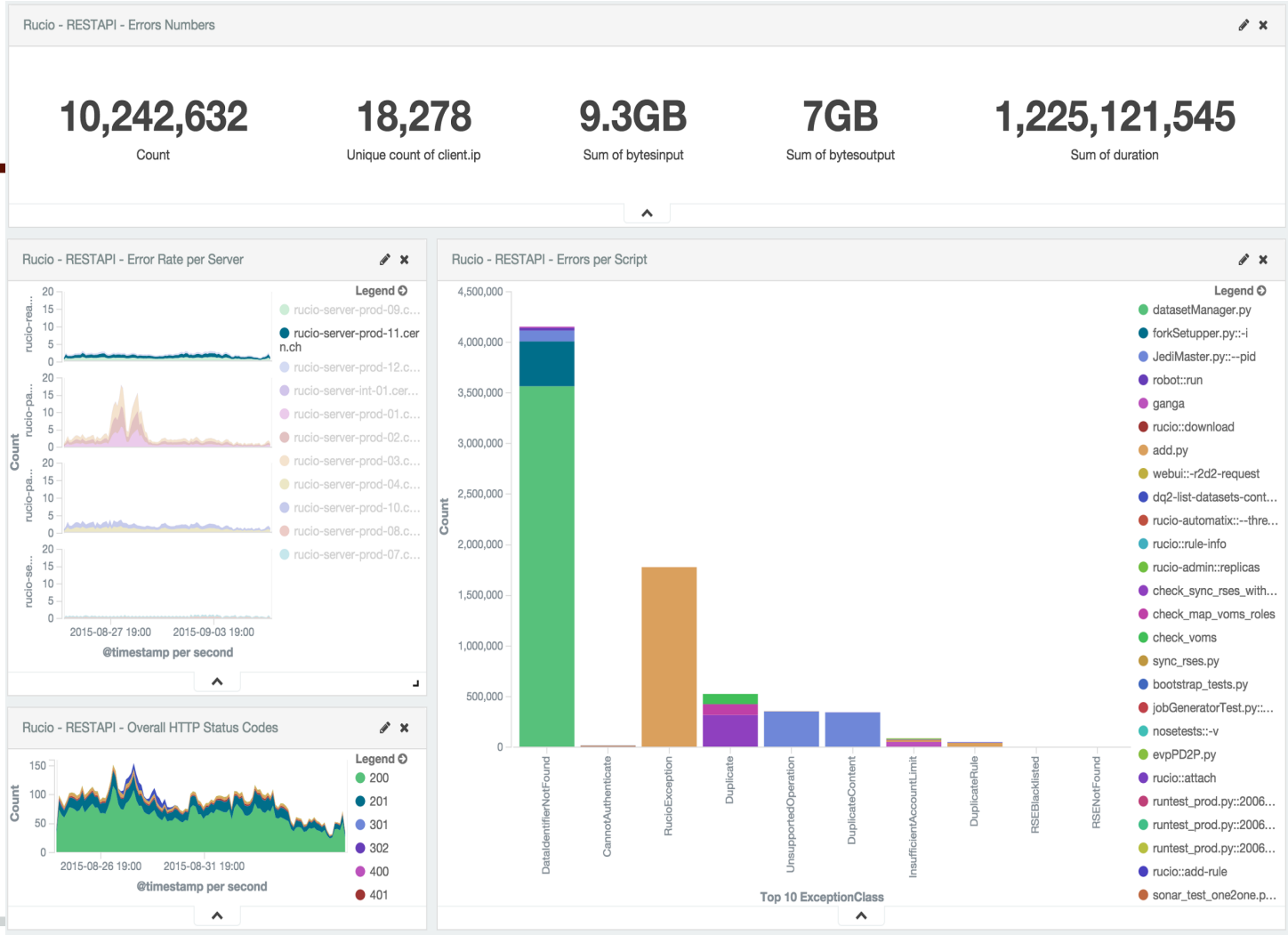




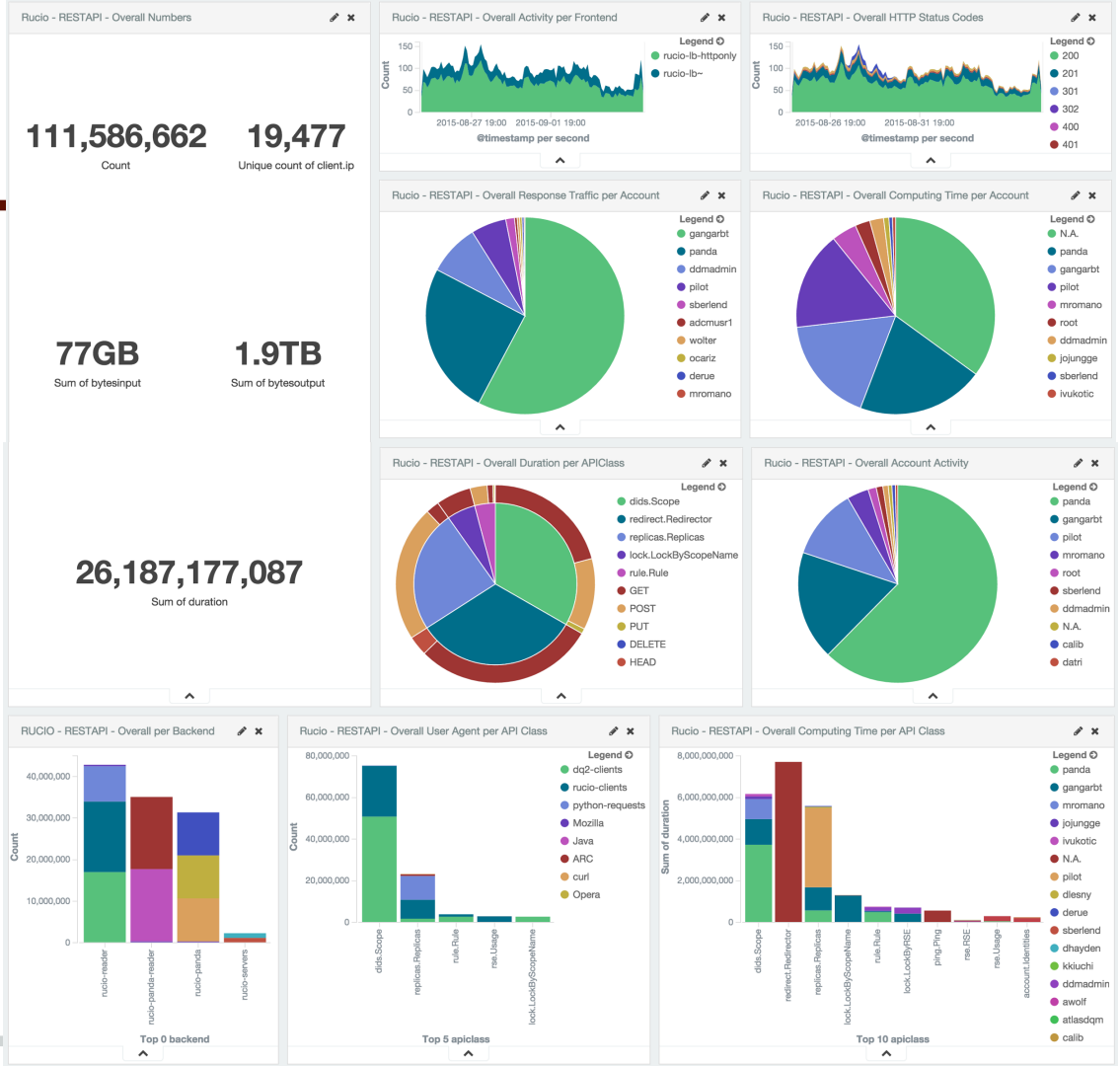
Rucio logging



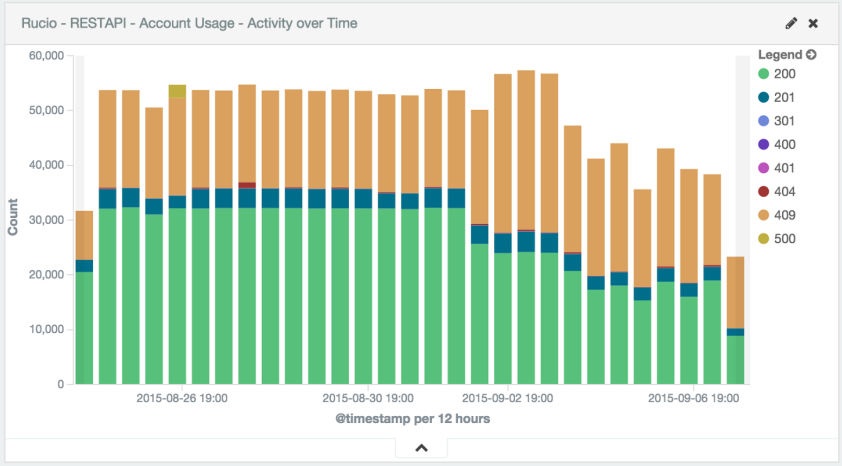
Rucio Error Tracking

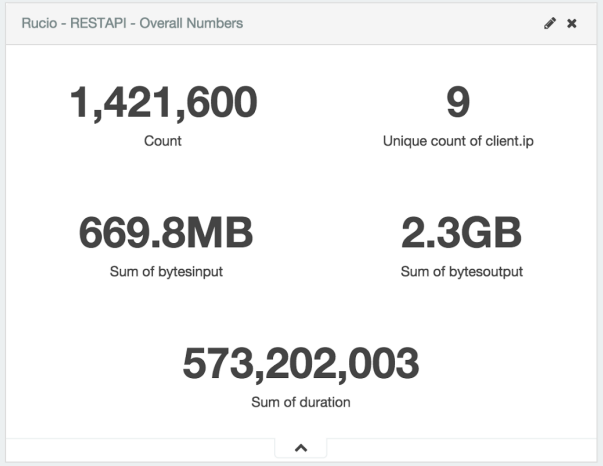


Rucio Account Activity



Rucio Account Usage

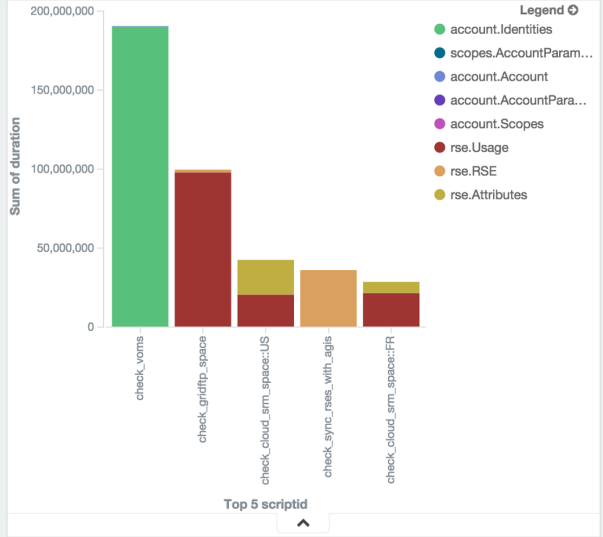




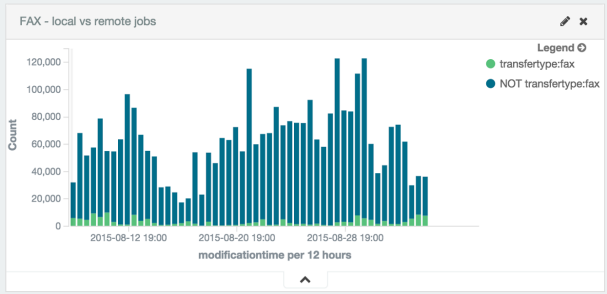
Rucio - RESTAPI - Account Usage - Resources

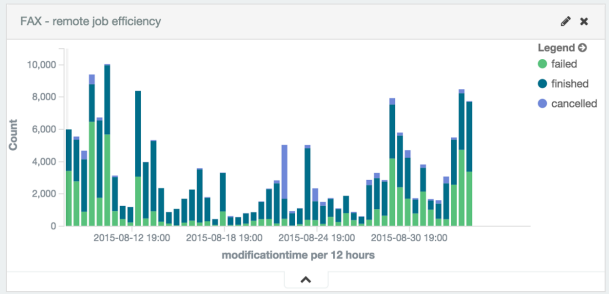
Top 10 request	Top 0 scriptid	Count	Sum of duration	Sum of bytesoutput	Sum of bytesinput
/accounts/pilot/identities	check_map_voms_roles	24,145	1,306,217	15.4MB	11.2MB
/accounts/phys-higgs/identities	check_map_voms_roles	15,082	694,607	9.5MB	7MB
/accounts/perf-muons/identities	check_map_voms_roles	11,786	543,156	7.6MB	5.6MB
/replicas/list	monitor_client::-s	6,229	3,316,796	34MB	35.9MB
/replicas/list	rucio::list-file-replicas	6	932	4.8KB	2.5KB
/replicas/list	nosetests::-v	2	184	10.6KB	808B
/replicas/list	rucio::list-dataset-replicas	1	133	25.7KB	480B
/accounts/phys-hi/identities	check_map_voms_roles	5,587	292,116	3.6MB	2.6MB
/accounts/phys-gener/identities	check_map_voms_roles	5,285	285,139	3.3MB	2.5MB
/accounts/perf-flavtag/identities	check_map_voms_roles	4,682	244,070	3MB	2.2MB

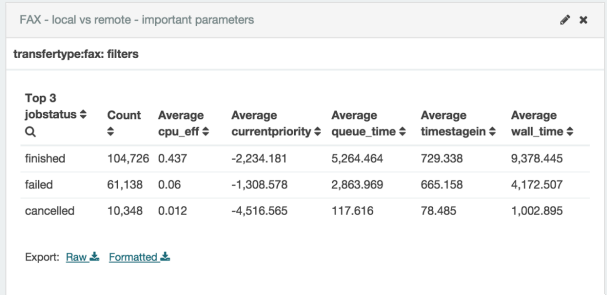
Export: Raw Formatted

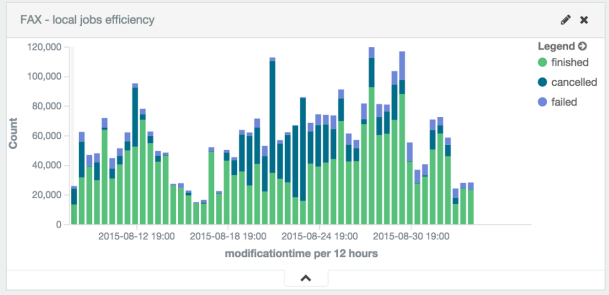


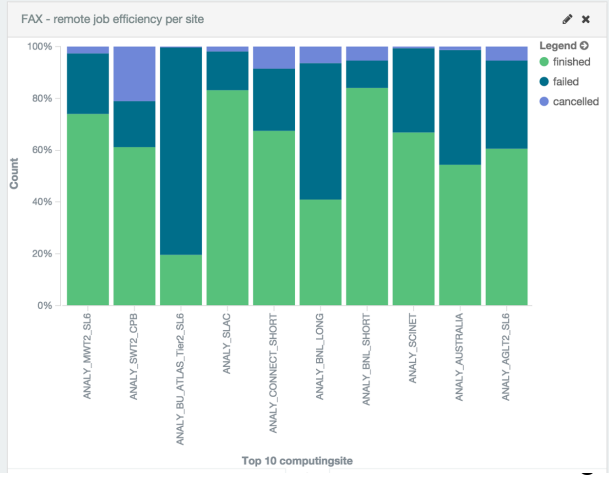
FAX overflow monitoring

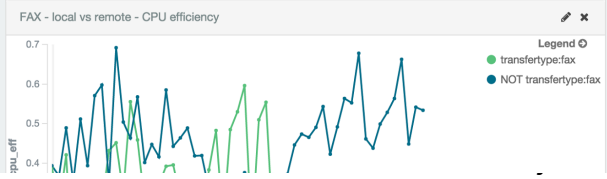




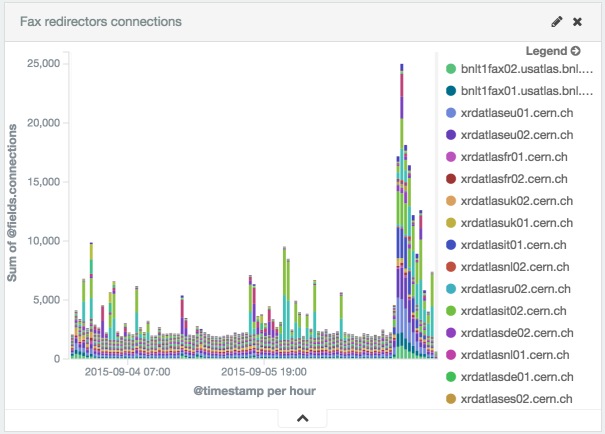


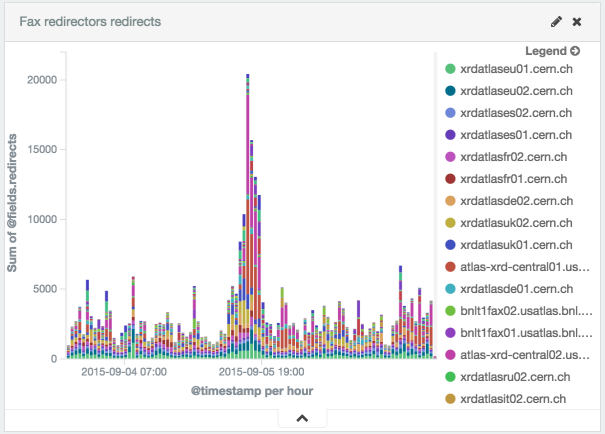


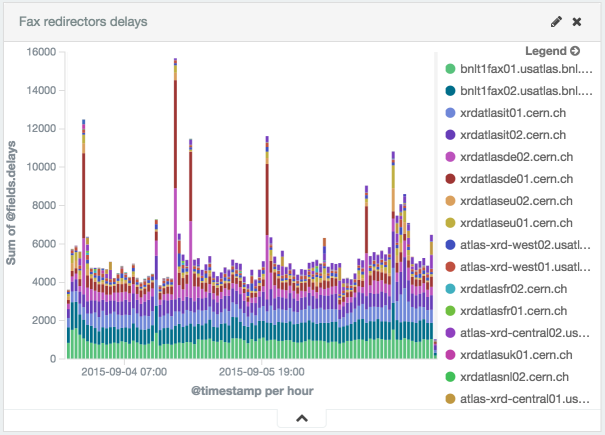


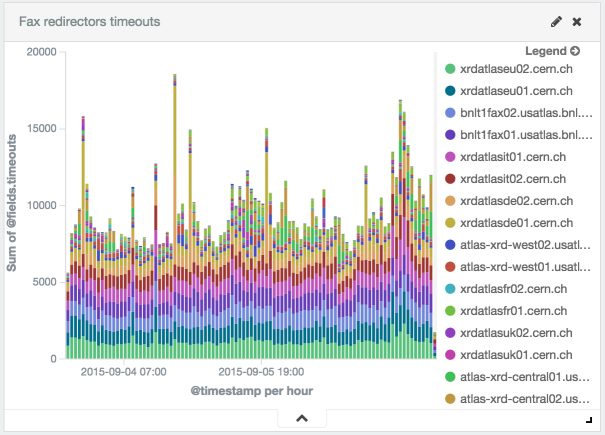


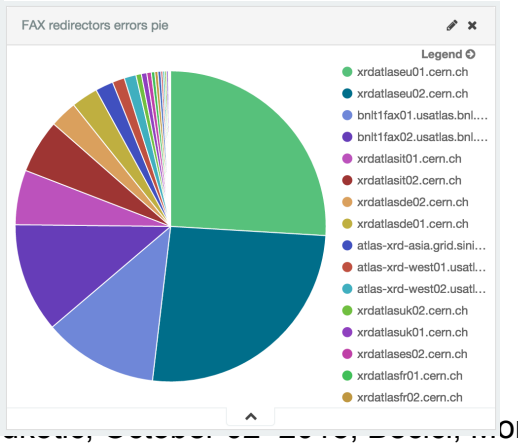
FAX redirector network monitoring











Conclusions

The new stack of Big Data tools (Hadoop, Flume, Sqoop, Logstash, pig, ES, Kibana) provide great platform for ATLAS ADC analytics tasks:

- horizontally scalable
- performant
- simple to develop for
- easy to use for an analyzer
- fast to make custom dashboards, GUIs
- can be used as a full search service