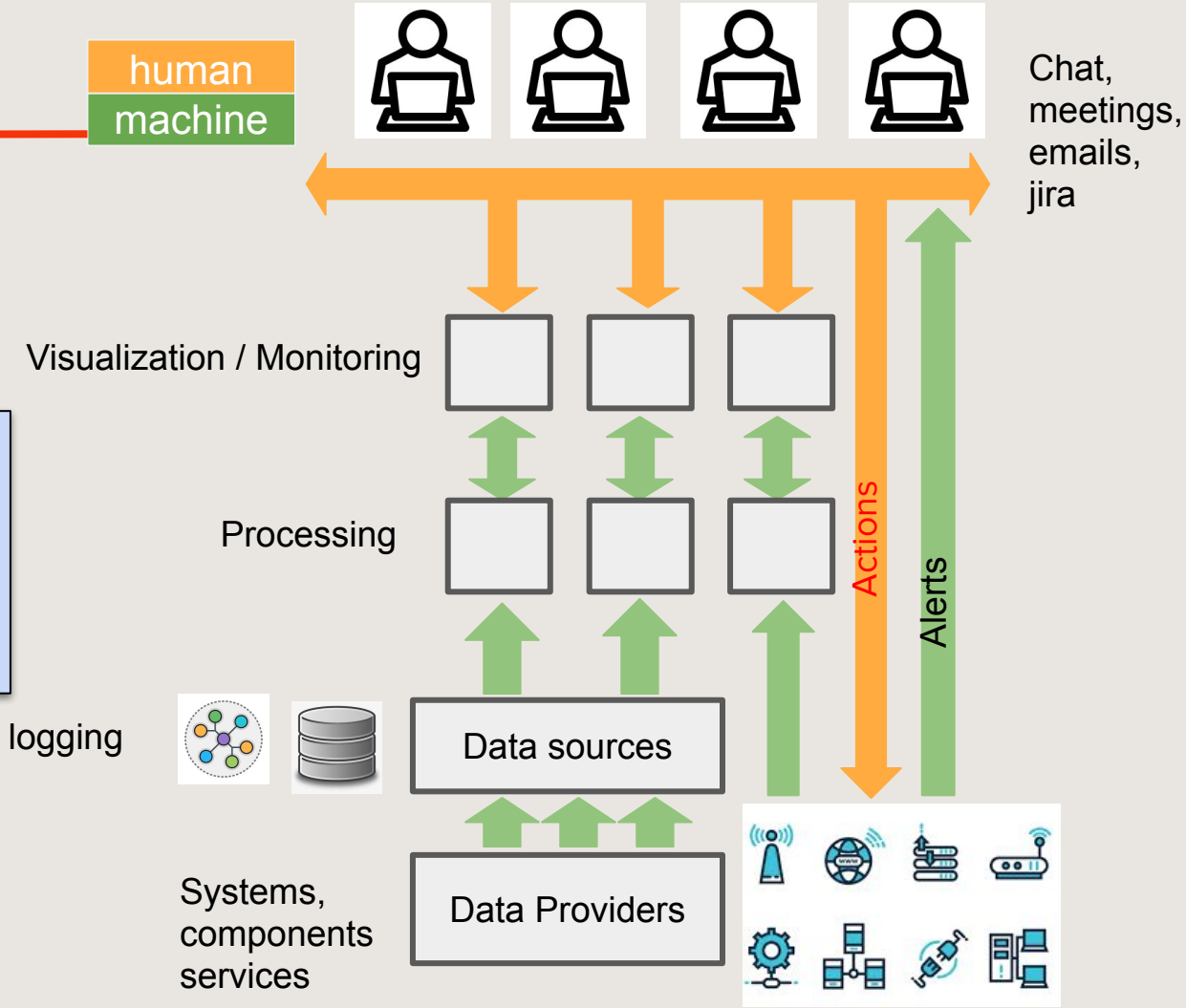
Operational Intelligence
status and future directions

Panos Paparrigopoulos on behalf of the Operational Intelligence initiative

Our Mission

- A cross-experiment effort aiming to streamline computing operations:
  - Improve resource utilization by reducing the time needed to address operational issues
  - Minimize human effort for repetitive tasks by increasing the level of automation
  - Build a community of technical experts: critical mass to have impact on concrete and common issues while setting up sustainable tools.
- Our mission:
  - Identify common projects
  - Leverage common tools/infrastructure
  - Collaborate, share expertise, tools & approaches
    - Across experiments
    - Across teams (operations, monitoring, developers)

Operations Today



ATLAS/CMS: 100+ people involved in Computing Operations (**50+ FTEs/experiment**)!
In 1 year:
> 1k GGUS tickets for ATLAS, > 2k for CMS

Can we do better?

- LHC experiments built a successful computing ecosystem for LHC Run-1/2
  - At which depth do we fully "**understand**" it?
    - Can we perform precise modelling of specific workflows / site behaviours / systems performances?
    - Can we use this modelling to make predictions (e.g. population vs pollution of Tier disks; TierX - Tier-Y data transfer patterns; smart data placement ...)
  - Up to now we monitored to debug in near-time,
    - not to analyse and learn from the past to design and build what's next.
- **However**: computing operations (meta-)data is all archived.
  - **Only recently started to be accessed**
  - e.g. transfers, job submissions, site performances, infrastructure and services behaviours, storage accesses, ..

Operations Tomorrow

human
machine

Frontend: aggregated views, suggestions, collects feedback

Visualization / Monitoring

Backend: Fetches, stores, filters, and analyses information about alerts, issues and solutions

ML

Analytics

Processing

Actions/ alerts

Actions

logging

Data sources

Systems, components services

Data Providers

Ongoing Efforts

What we are doing to succeed:

- Develop tools to automate computing operations exploiting state-of-the-art technology and tools
- Run a technical forum, experiment-agnostic to:
  - bring people together
  - discuss ideas, brainstorm together, share experience and code

We have identified the areas where shared development can occur:

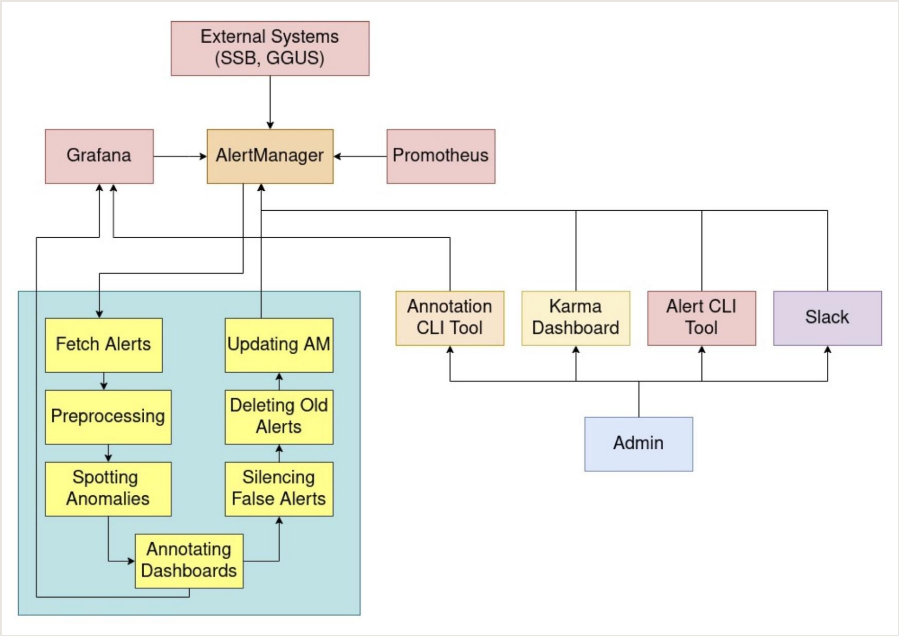
- Sites
- Workflow Management
- Data Management

And we are also trying to provide some shared infrastructure:

- A common k8s cluster for all those services to be deployed.
- A framework which can be used to develop new tools

Alert manager (AM)

- Of course there are activities that take place between all the aforementioned areas.
- CMS i.e. developed an intelligent layer in their infrastructure to detect, analyze and predict abnormal system behaviors using the alerts produced by the infrastructure. This allows the operation teams to focus more on finding solutions for the source of alerts rather than searching, filtering and collecting the alerts



- The alert manager fetches the existing alerts, pre-process them and try to spot any anomalies.
- It then annotates the corresponding Grafana dashboard, were false alerts can be silenced and automatically removed when resolved. This info is then being feeded back to the AM.
- SSB and GGUS are now being integrated to AM. Tickets will be feeded in AM, the same way as alerts do and will be automatically annotated to Grafana dashboard etc.
- This will provide useful insights about when outages happen and how they affect the productivity reported by various systems in CMS dashboards.

http://cern.ch/go/cxg8

Workflow Management

- Workflow management is a complex work that could benefit by optimisation and smart tools.

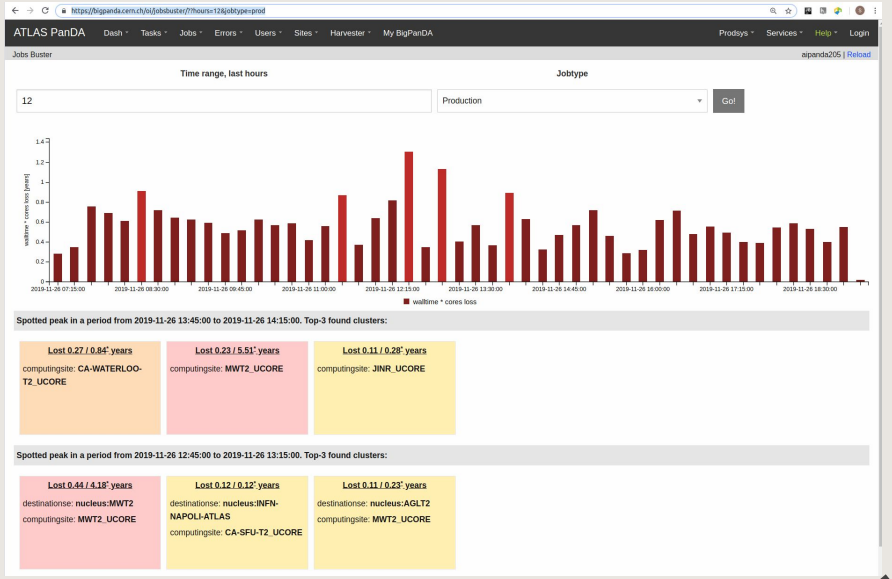
  - **ATLAS** has developed a tool called Jobs Buster which tries to spot operational problems errors in ATLAS jobs and display results in the BigPanDA monitoring page.

    http://cern.ch/go/8qwC

  - Using NLP Jobs Buster tries to find the common denominator between failed jobs

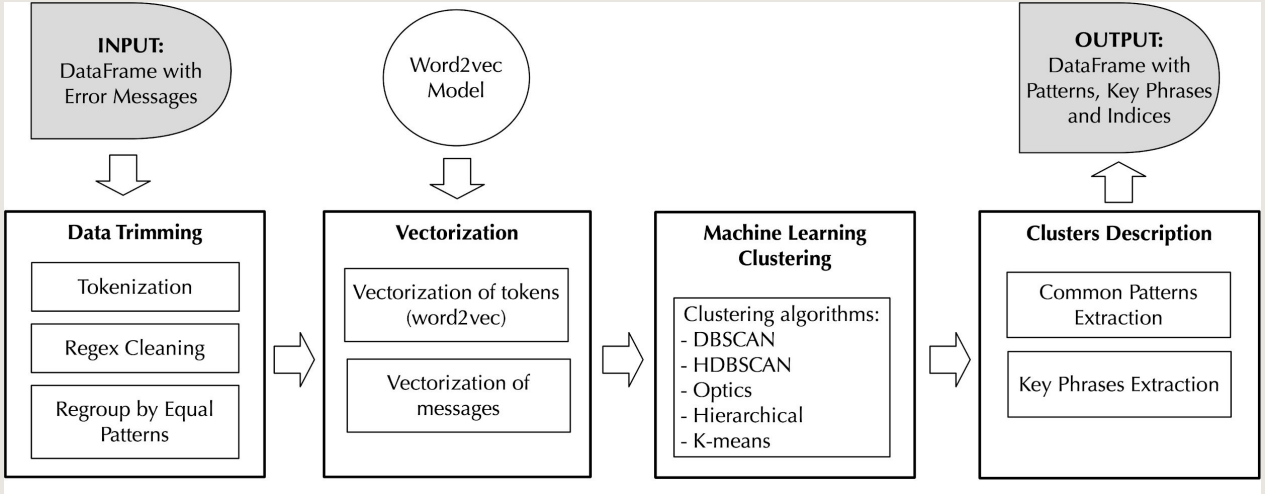
  - **CMS** has developed a similar system. The CMS Operator console.

    http://cern.ch/go/z76x



Data Management

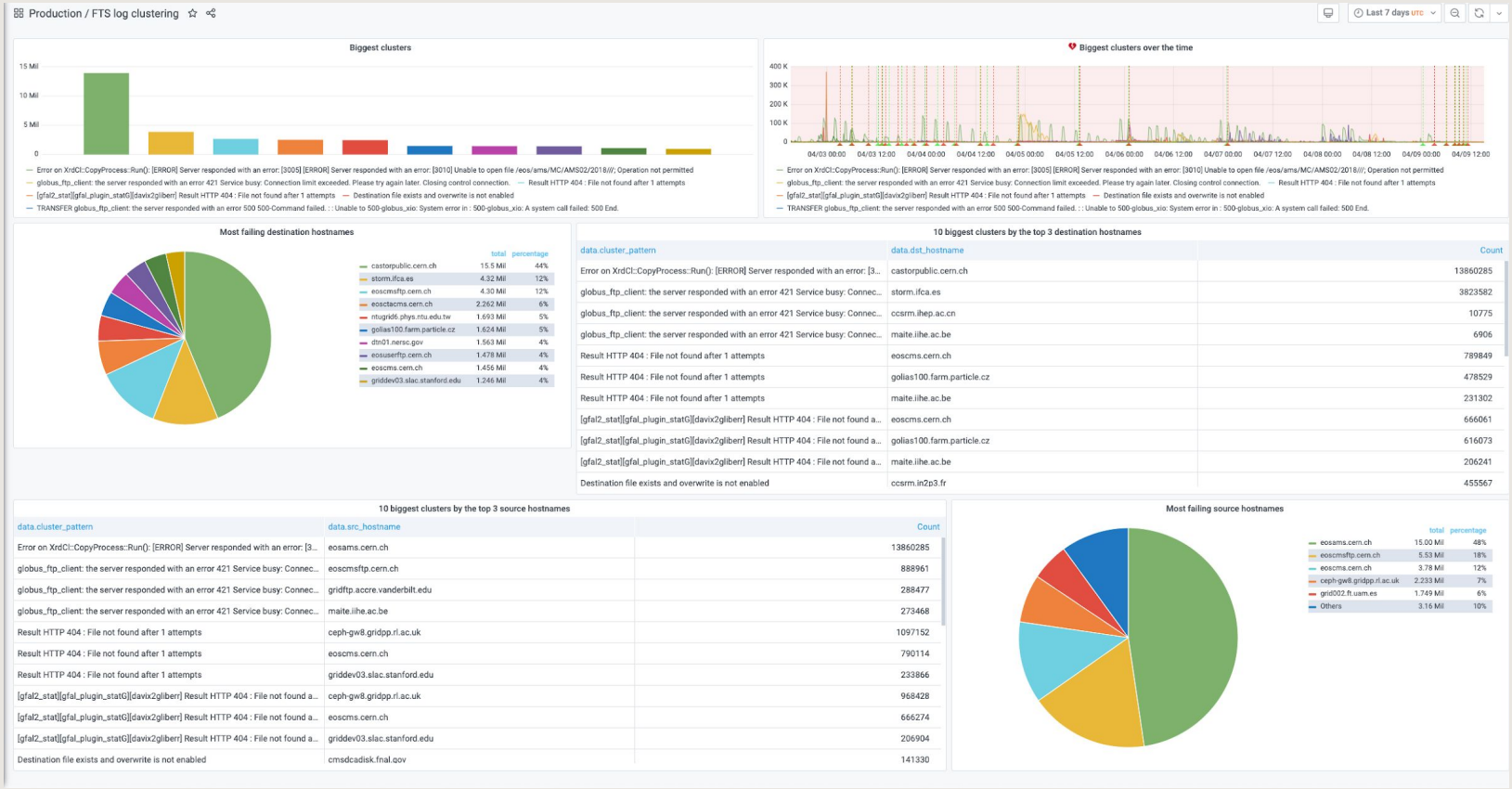
- There are currently two efforts in **CMS** and **ATLAS** to classify FTS error logs and try to provide useful information to DDM shifters.
- Currently the CMS version has moved to production. FTS logs are being analysed with NLP algorithms and the results are being displayed in a Grafana dashboard.



http://cern.ch/go/BX6l

http://cern.ch/go/7jMl

Data Management



https://monit-grafana.cern.ch/d/Zx_bXneWz/fts-log-clustering?from=now-7d&orgId=11&to=now

Anomaly detection on transfers

- In our search for experts and personpower we got in touch with Google which offered to help.

- The idea is to run anomaly detection algorithms on transfer logs and create a tool which will be able to spot anomalies in almost real time and alert the users.

- We managed to get promising results and we now look into validating them using data from the tickets that have already been created.

Anomaly detection on transfers

NEW

An interesting find was that error distribution not only varied over time, but also over the interconnections between nodes

Given the observed changes in error distribution across time, connection graph and content (as represented by the error categories), we investigated graph anomaly detection algorithms as a possible way to identify patterns in the logs.

MIDAS (MIcrocluster-based Detector of Anomalies in Streams) seemed a good fit:
- It finds anomalies in dynamic graphs (such as those generated by file transfers, but also
intrusions)
- It detects micro-clusters (sudden "burst" of connections between nodes, such as those
that may occur with multiple retrials, but also denials of service)
- Memory usage is constant and independent of graph size
- Update time in streaming scenarios is also constant

https://arxiv.org/pdf/1911.04464.pdf

dst / Record Count

src	srm-cms.gri...	gridftp.swt...	dtn.ilifu.ac.za	gridftp.hep...	t2cmcondo...	tbn18.nikhe...	uct2-dc1.uc...	fal-pygrid-3...	griddev03.s...	bohr3226.ti...
bohr3226.tier...	-	5,739	737,095	-	6,911	19,902	3,490	10,940	55,722	136
tbn18.nikhef.nl	-	12,891	-	-	14,466	-	6,133	14,429	893	14,515
eoscmsftp.ce...	38,806	-	-	37,524	-	-	-	-	-	-
dcsrm.usatla...	-	63,813	-	-	44,551	8,058	19,459	14,912	-	4,844
uct2-dc1.uchi...	-	4,764	-	-	3,487	7,157	45	6,938	-	28,582
eosatlassftp....	-	39,750	-	-	65,132	10,828	33,056	11,091	-	1,908
ccsrm.in2p3.fr	32,366	43,079	-	23,902	31,446	2,875	5,364	4,988	-	1,177
golias100.far...	-	5,196	-	-	1,397	18,766	1,973	10,772	61,104	10,434
sdrm.t1.grid.k...	-	14,670	-	-	8,203	16,549	1,018	10,025	874	9,462
storm.ifca.es	13,081	-	-	5,582	-	-	-	-	-	-

Oct 1, 2019 - Nov 1, 2019

Figure 3: Count of errors over connection pairs

Start_Hour / Record Count

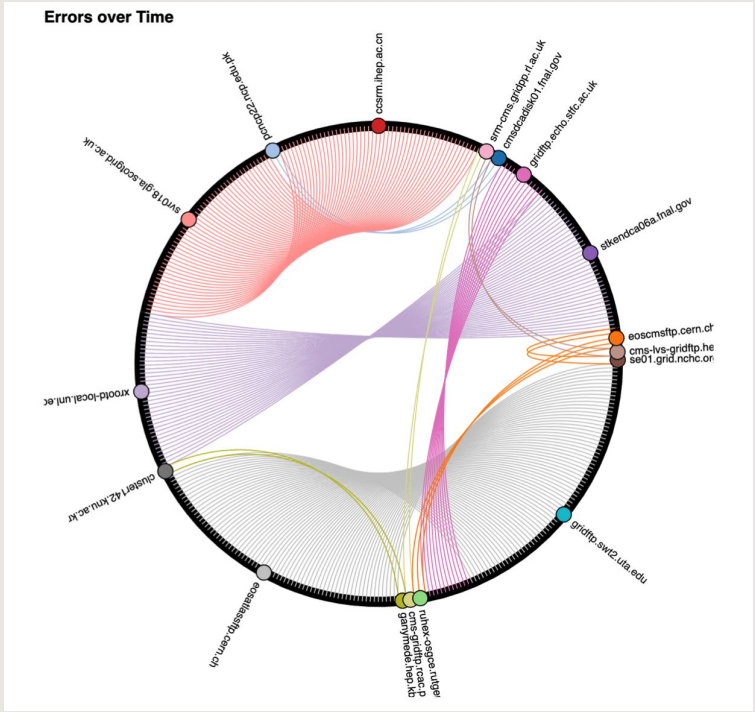
Top 10 - dat...	Top 10 - data...	201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 1(
bohr3226.tier...	dtn.ilifu.ac.za	4,106	3,450	3,511	4,215	4,636	3,411	3,155	3,782	4,600	
	griddev03.sla...	2	-	-	-	-	-	-	-	-	
	serv02.hep.p...	183	163	143	171	207	155	171	210	195	
	tbn18.nikhef.nl	50	55	51	49	43	211	20	7	25	
	fal-pygrid-30.l...	32	38	34	29	27	25	14	26	62	
	f-dpm000.gri...	27	32	26	28	25	398	3	2	5	
	ftp1.ndgf.org	26	29	26	28	23	395	3	-	-	
	sdrm.t1.grid.k...	25	28	27	28	25	201	3	-	-	
	dcache-atlas-...	26	29	26	26	26	323	3	-	-	
	xrootd.echo.s...	23	29	29	26	21	202	-	-	-	

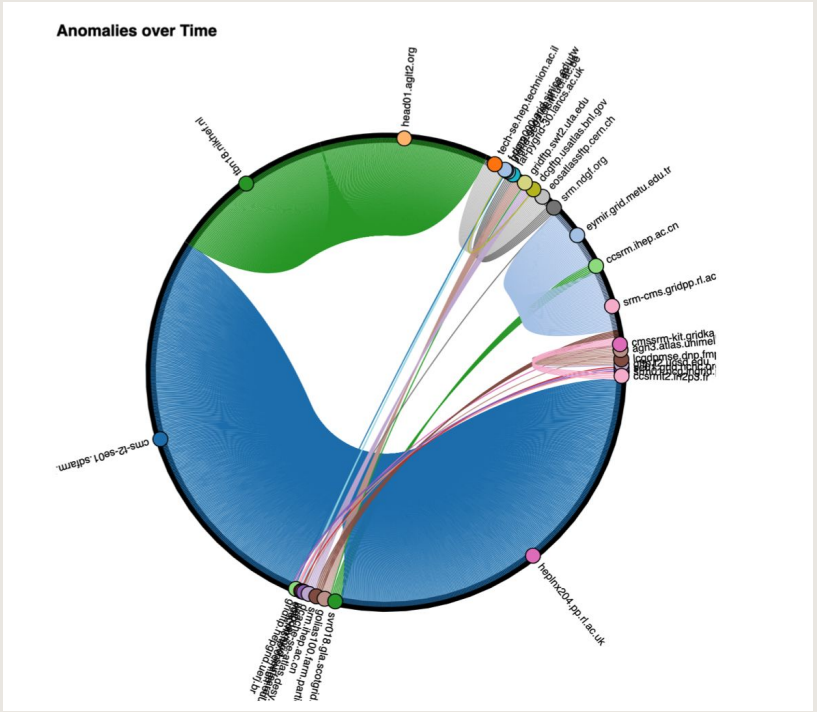
Figure 4: Variation over time for a given connection pair

Anomaly detection on transfers

NEW

With MIDAS we were able to find the most anomalous connections on a graph in a given time window and monitor the evolution of those connections over time





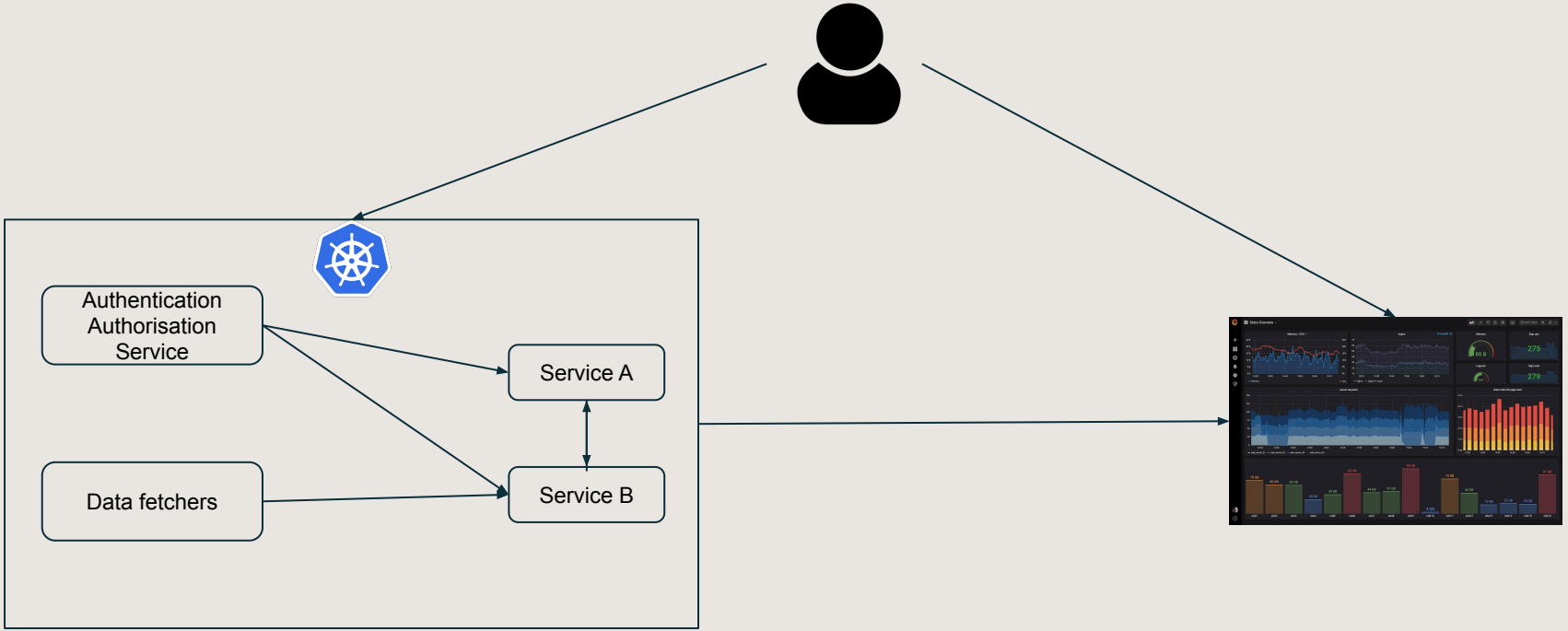
Anomaly detection on transfers

- There is of course a lot of work left to be done:
  - Include text features in anomaly detection. We must consider not only the number, timing and location of links between nodes, but also the messages. Other metadata such as user, file size etc... may play a role too.
  - Include data from tickets to strengthen and validate our results.
  - Build an interface for shifters to explore the results of this analysis.
- This effort is now a pilot project in the EU CloudBAnk.
- We hope to have more results in this front soon since collaboration with Google is moving fast.

The shared k8s cluster

NEW

- We have created a cluster in the WLCG space and we will start with deploying the FTS log monitoring project that is developed for CMS.



Sites Optimisation

- There have been some efforts by sites that are developing tools very much in line with the OpInt project.
- There is some very interesting work on cloud anomaly detection going on at CERN.
- The idea is to automatically detect anomalies in the Openstack infrastructure and alert users.
- An enchantment on Grafana annotations has been developed in order to automatically tag such anomalies at Grafana dashboards, helping people to correlate activity in graphs with anomalies.

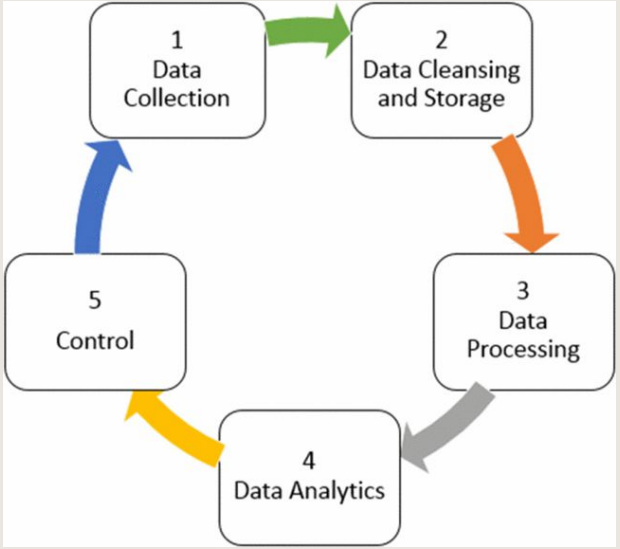
http://cern.ch/go/XnL6

Sites Optimisation

- We are also keeping an eye into what big companies from the industry do to automate their computing centers and reduce operational costs and environmental impact.
- Of course in a diversified environment like WLCG these holistic strategies cannot always apply.
- The past years we have moved into a more unified processing pipeline in our sites, something which creates possibilities for collaborative efforts.

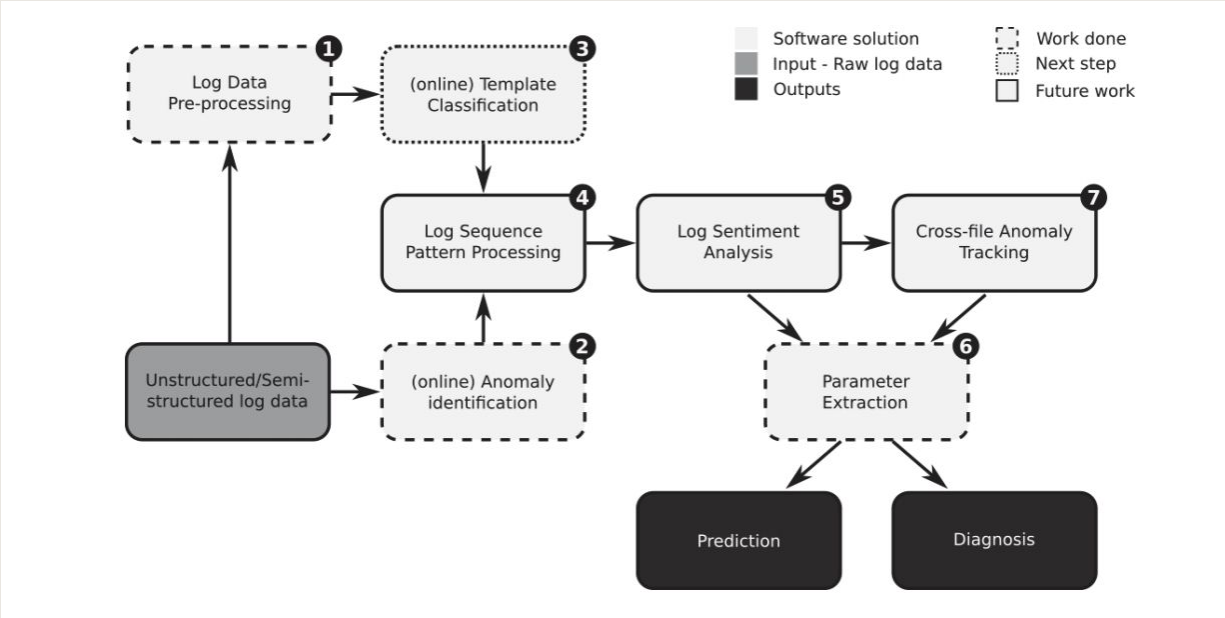
Industry examples

- A lot of interesting hardware related work:
  - Building sensors throughout their networks so that they can redirect workload to offload overloaded nodes
  - Using SMART (Self-Monitoring, Analysis and Reporting Technology) to derive disk failure predictions and replace hardware proactively
  - Using AI to manage the cooling and power management of the data center (advertising up to 5% gains in performance)
  - In general: **predictive maintenance** based on sensors and computing logs



INFN Bologna - Predictive maintenance

- INFN Bologna has started a very interesting project trying to switch from reactive maintenance to predictive maintenance.
- They are using the logs of the various services and through a pipeline of analysis they try to diagnose, or even better predict, errors.



http://cern.ch/go/B9KR

Site efforts

- We need more projects coming from sites that are trying to optimise their operations.
- Such initiatives cannot come from individual sites. Having a group of sites collaborate will produce results that are generic enough to be used by everyone.
- We are in a place that we can bring people together and help coordinating but we need ideas and volunteers from sites .

Conclusions

- We have in the past 2 years gathered expertise and an understanding of the various efforts.
- We can see there is room for improvement.
- Your feedback and your ideas are vital.

operational-intelligence@cern.ch