Distributed Computing Operations in HL-LHC era with **Operational Intelligence**

Alessandro Di Girolamo, Daniele Bonacorsi, David Hohn, Domenico Giordano, Federica Legger, Jaroslava Schovancová, Leticia Decker de Sousa, Lorenzo Rinaldi, Luca Clissa, Luca Giommi, Maria Grigorieva, Mario Lassnig, Matteo Paltenghi, Mayank Sharma, Michael Boehler, Micol Olocco, Nikodemas Tuckus Panos Paparrigopoulos, Siarhei Padolski, Simone Rossi Tisbeni, Stephane Jezequel, Thomas Beermann, Tomáš Javůrek, Tommaso Diotalevi, Valentin Kuznetsov, Vasilis Mageirakos

Outline

- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

Outline

- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

https://operational-intelligence.web.cern.ch/

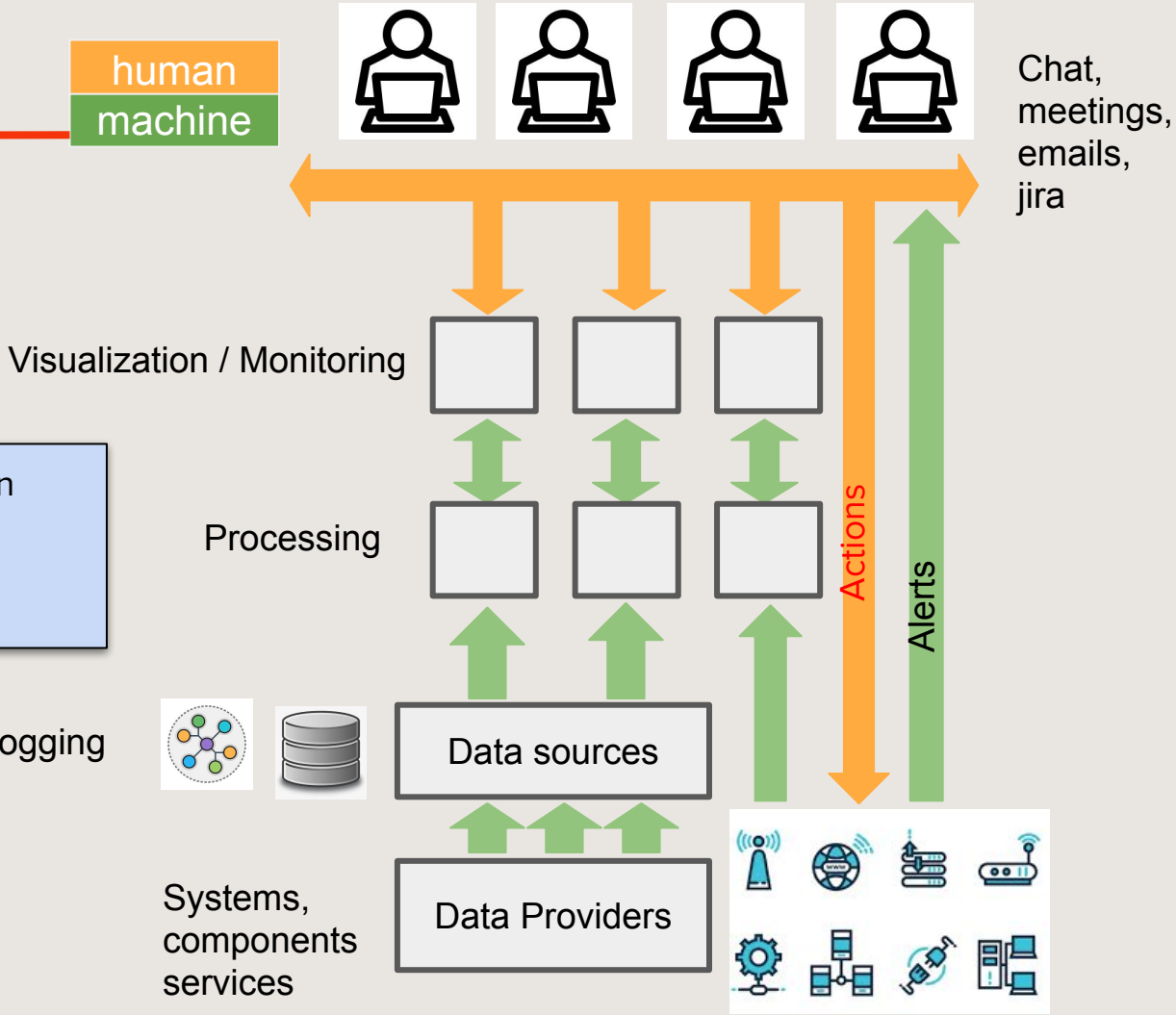
Our Mission

- 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)

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 the workflows and our services and use this modelling to make predictions?
  - Up to now we monitored to debug in near-time.
    - Can we analyse and learn from the past to design and build tools that will help with operations?
- HL-LHC: one order of magnitude more resources than today
  - We are not going to have one order of magnitude more personpower to operate
- **However**: computing operations (meta-)data is all archived
  - We have logs for transfers, job submissions, site performances, infrastructure and services behaviours, storage accesses, ..
  - All this knowledge should be exploited!

Operations Today



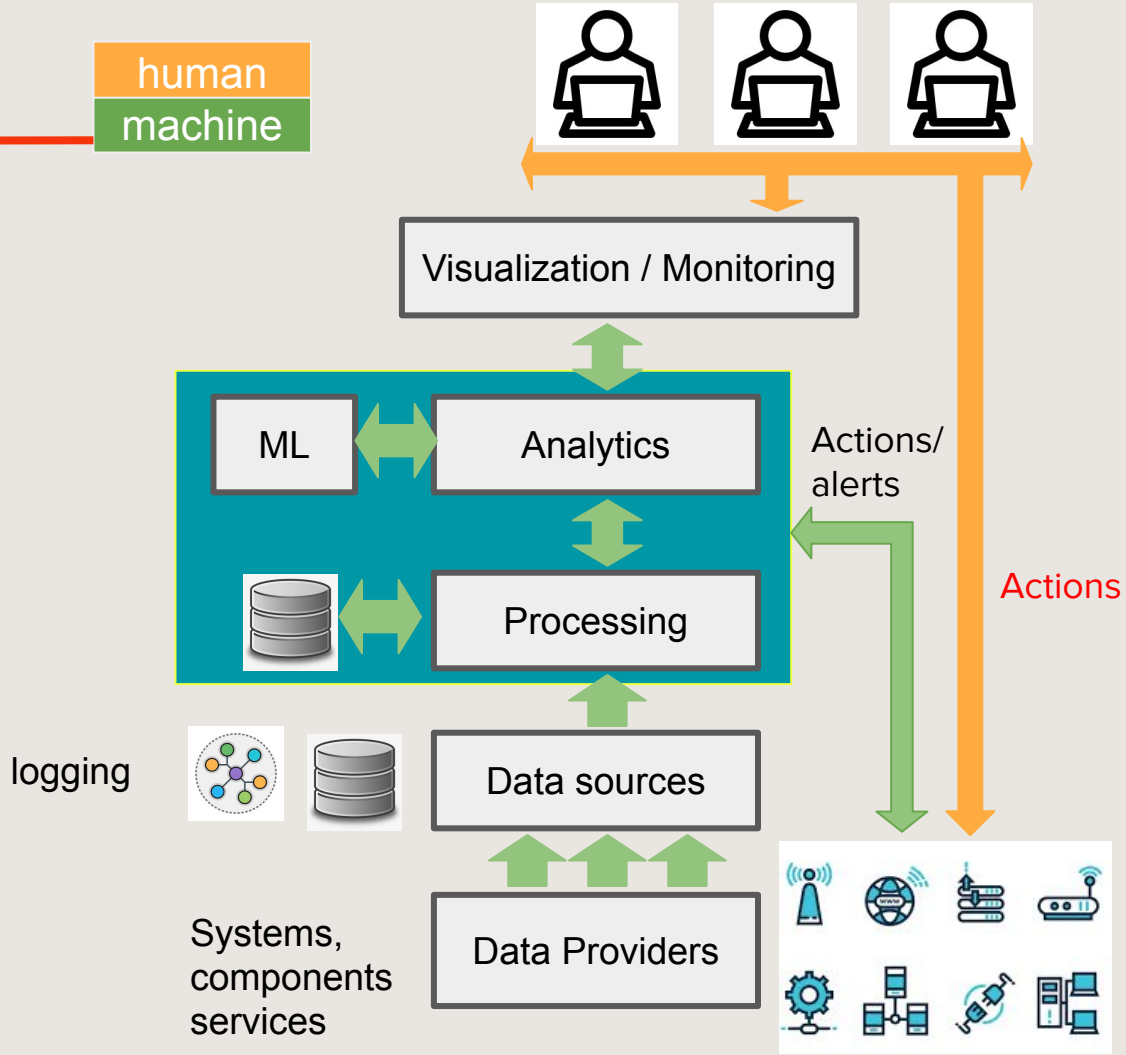
ATLAS/CMS: A lot of people involved in Computing Operations
In 1 year:
> 1k tickets for ATLAS, > 2k for CMS

Operations Tomorrow

human
machine

Frontend: aggregated views, suggestions, collects feedback

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



Ongoing Efforts

What we are doing:

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

We identified **areas where shared development can occur**:

- Computing facilities
- Workflow Management
- Data Management

And we provide some **shared infrastructure**:

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

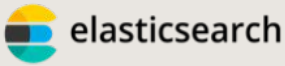
Outline

- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

Infrastructure

- Our infrastructure is based on **open-source** products
  - Kubernetes is the de-facto standard for **deploying and scaling services**
  - HTTP and AMQ for data injection
  - Prometheus and ElasticSearch for **managing metrics and meta-data**
- Clear separation of Data, Infrastructure, Visualization simplify operations
- Data standardization, common naming convention, data validation plays an important role

- Automation is a key to success
  - Data annotation, alerting, notifications, tagging, etc.

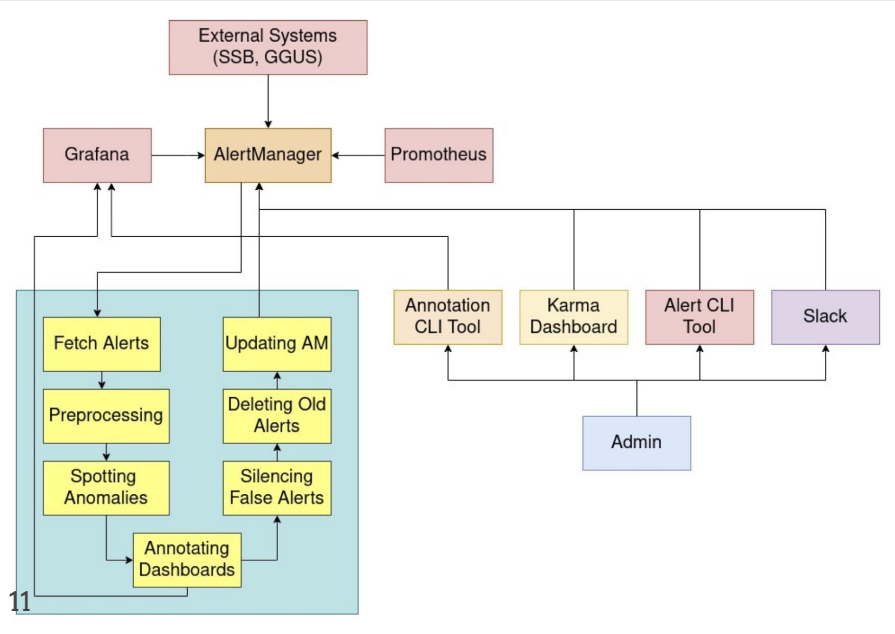


For more details see:

The Evolution of CMS Monitoring Infrastructure talk

Intelligent Alert system

- CMS developed an intelligent layer in their infrastructure to **detect, analyze and predict abnormal system behaviors** using the **alerts** produced by the infrastructure.

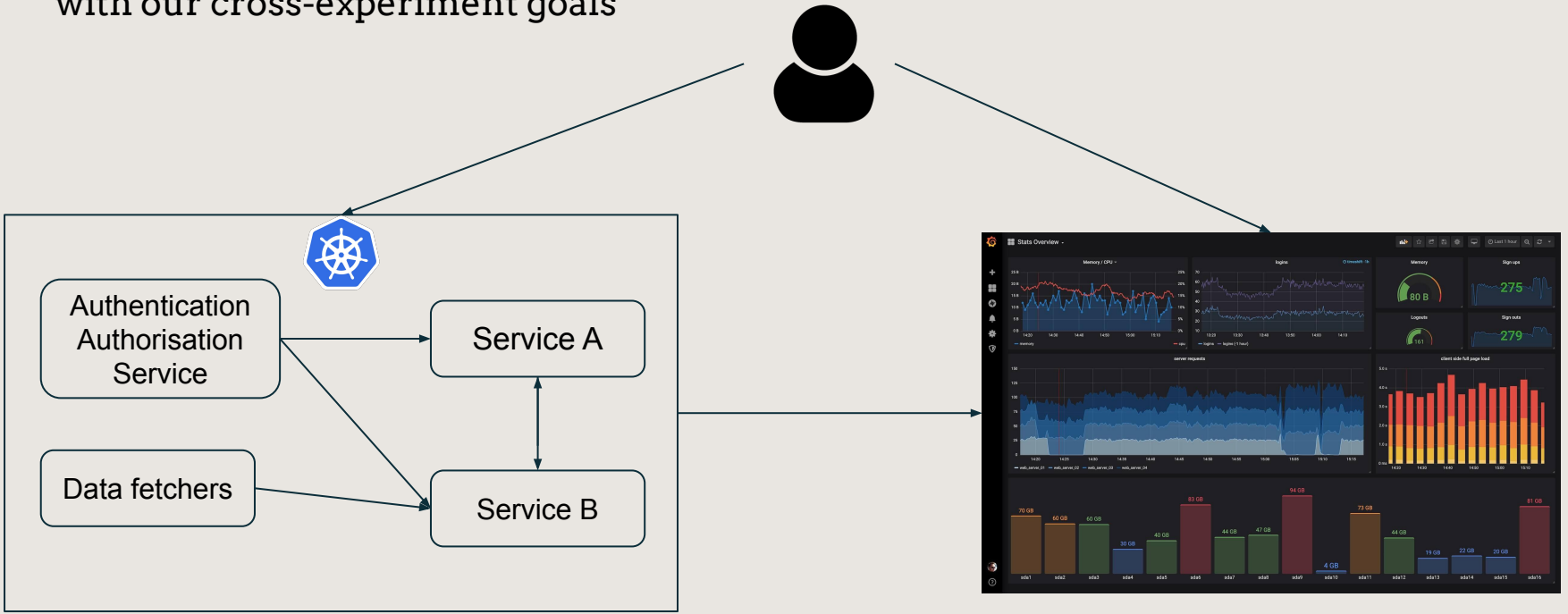


- The alert manager fetches the existing alerts, filters them, and **annotates Grafana** dashboards based on the alert tag
- SSB and GGUS are also integrated into the Alert Manager
- The system provides useful insights about when outages happen and how they affect the productivity reported by various systems in CMS dashboards
- Using **open source tools** makes this effort experiment-agnostic

More info: http://cern.ch/go/cxq8

The shared k8s cluster

- Having a common (not experiment specific) space to deploy our applications is in line with our cross-experiment goals



Outline

- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

Jobs Buster

- ATLAS " Jobs Buster" tries to spot **operational problems** in submitted jobs
- **Machine Learning** is used to cluster the errors and then find the **common denominator** between failed jobs in the cluster (could be software version, site name, transfer src/dst etc)

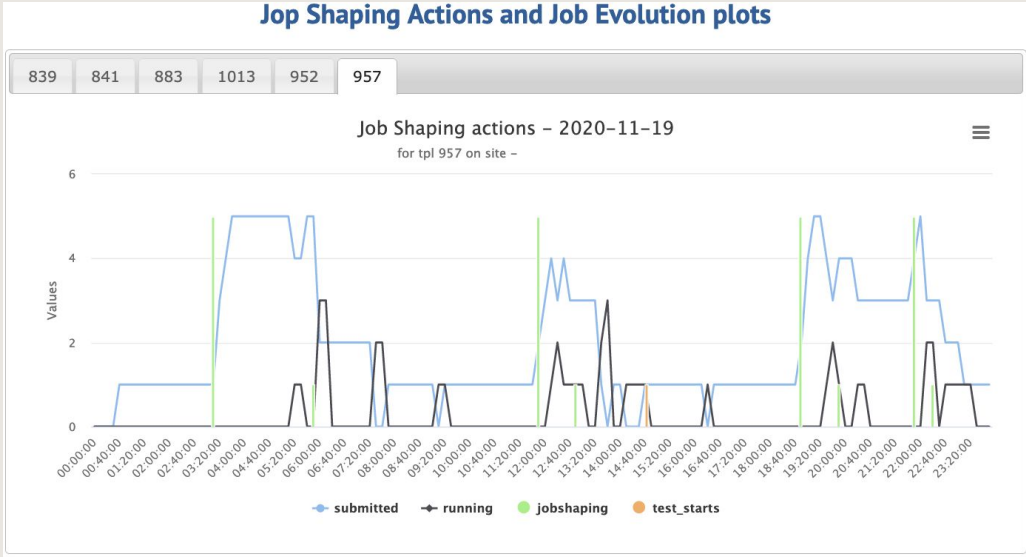
More info: http://cern.ch/go/8qwC

HammerCloud JobShaping

- HC checks functionality of each compute sites for ATLAS & CMS in WLCG

- ATLAS runs an **auto-exclusion mechanism**
  - Sets sites “offline” with failing functional tests
  - Re-includes succeeding sites automatically

- JobShaping aims to **speed up** the automatic exclusion and recovery decisions
  - Problem: test jobs might get stuck or run much longer than expected -> lacking fresh info for decision
  - Solution: adjusting the number of parallel running jobs per site and test type dynamically



Prototype view of jobShaping web interface

- Next steps: add specialised debug tests only sent to sites with failing test jobs
  - Help problem solving and identifying failure source

Outline

- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

Data Management - Analysis of error messages

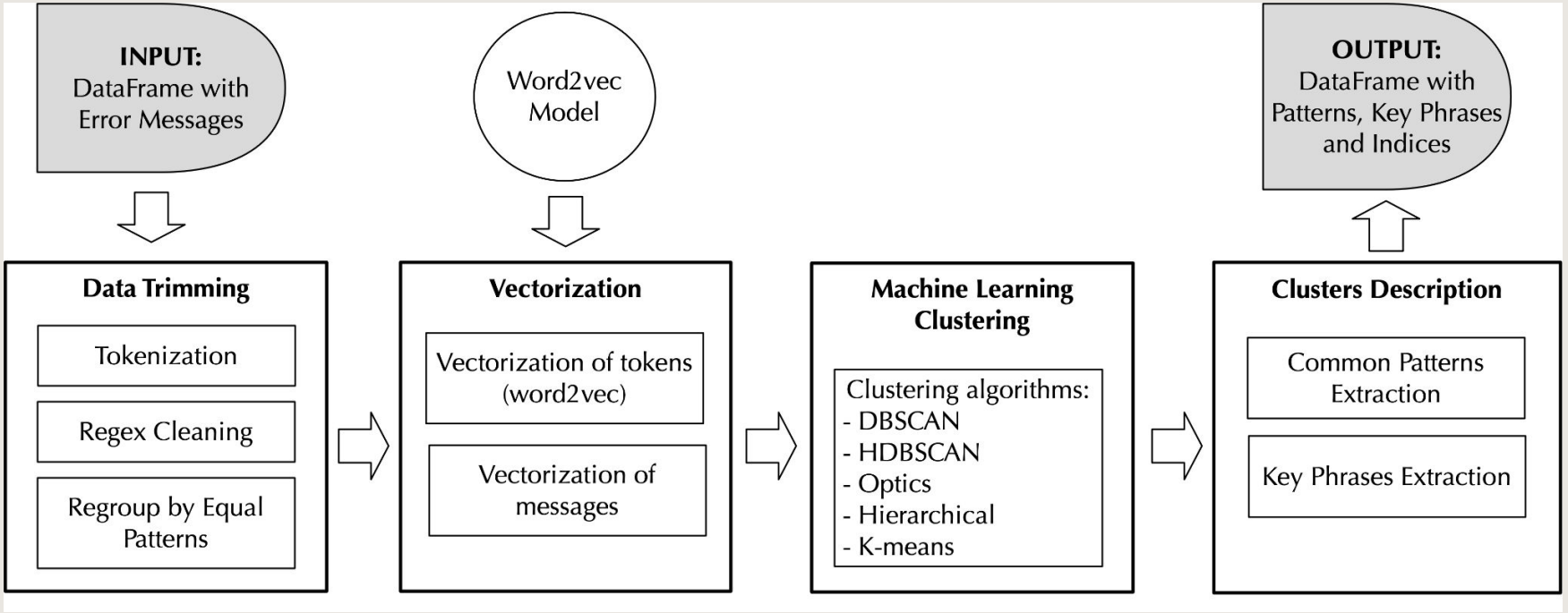
- Every day operation teams must deal with multiple data transfer errors
- The monitoring systems help users to detect anomalies, to identify duplicated issues, to diagnose failures and to analyze failures retrospectively
- Clustering of error messages is a possible way to simplify the analysis:
  - Messages having the similar text pattern and error conditions are grouped
  - Groups of similar messages are described by the common text pattern(s) and keywords
  - Messages encountered only once or several times are considered as anomalies

- There are currently multiple efforts trying to analyze the error messages and simplify operations

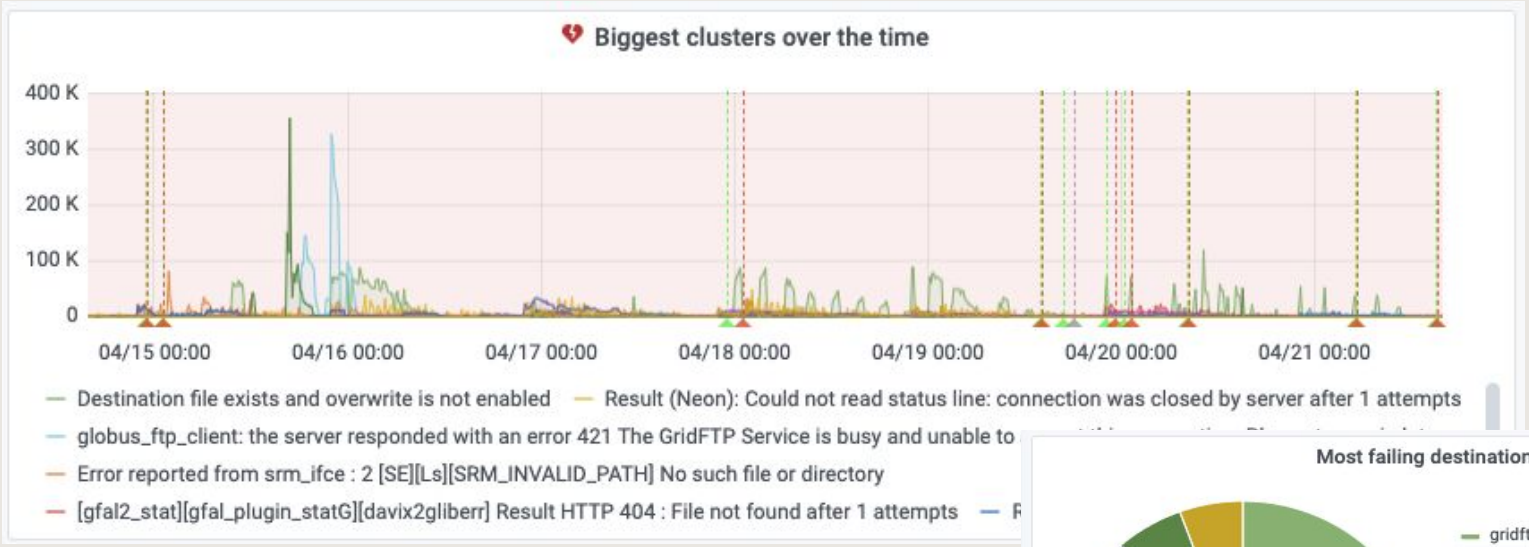
ClusterLogs

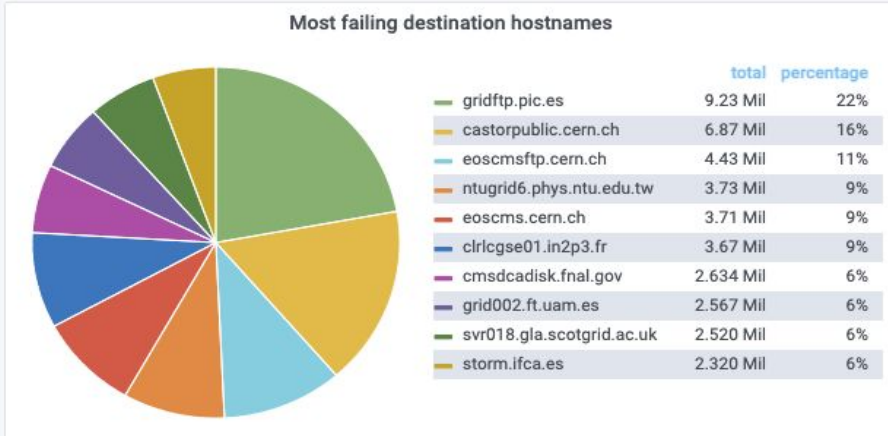
- ClusterLogs is one of the frameworks that was developed within OI to cluster error messages

More info: https://github.com/maria-grigorieva/ClusterLog



Data Management: FTS log analysis





ClusterLogs is used to classify File Transfer Service (FTS) logs and results pushed back to the MONIT infrastructure where they can be browsed from a Grafana dashboard

More info: http://cern.ch/go/qk7S

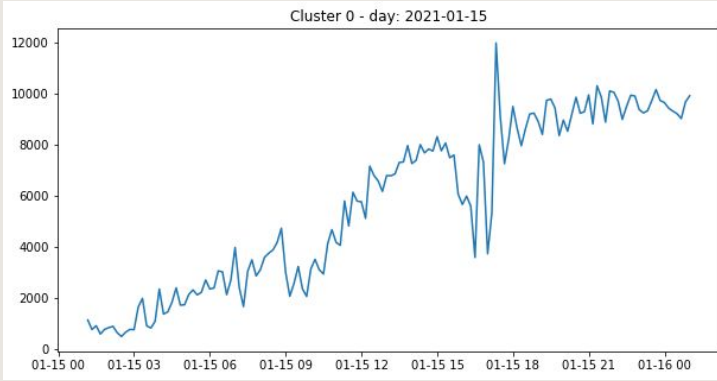
Data Management: FTS log Analysis

- Similar effort to clusterize FTS error messages, and validate the results using GGUS tickets

Preliminary results:

- The model learns to abstract message parameters as IPs, URLs, file paths, …
- Testing against GGUS tickets gives promising results:
  - Most problems recognized → exact match between cluster and GGUS ticket
  - Undetected/unreported issues → hints of real problems that were not reported on GGUS (under study)

ID	cluster size	# strings	# patterns	Top 3 message	n	%	source rcsite	destination rcsite
0	819465	117	14	destination overwrite srm-ifce err communication error on send err [se][srmrm][] $URL /srm/managerv2 cgsi-gsoap running on $ADDRESS reports error initializing context gss major status authentication failed gss minor status error chain globus_gsi_gssapi ssl handshake problems globus_gsi_callback_module could not verify credential globus_gsi_callback_module could not	85545	10.44%	Site A	Site D
				destination overwrite srm-ifce err communication error on send err [se][srmrm][] $URL /srm/managerv2 cgsi-gsoap running on $ADDRESS reports error initializing context gss major status authentication failed gss minor status error chain globus_gsi_gssapi ssl handshake problems globus_gsi_callback_module could not verify credential globus_gsi_callback_module could not	84453	10.31%	Site B	Site D
				destination overwrite srm-ifce err communication error on send err [se][srmrm][] $URL /srm/managerv2 cgsi-gsoap running on $ADDRESS reports error initializing context gss major status authentication failed gss minor status error chain globus_gsi_gssapi ssl handshake problems globus_gsi_callback_module could not verify credential globus_gsi_callback_module could not	77410	9.45%	Site C	Site D



Time/# of errors for selected cluster

Anomaly detection on FTS transfers

- Google is working with us to develop a recommendation system to help operation teams to prioritise transfer errors

- FTS logs analysis showed that we can study errors evolution not only over time but also over the interconnection between nodes (site endpoints)

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

Anomaly detection on FTS transfers

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

MIDAS (MIcrocluster-based Detector of Anomalies in Streams) :

- Finds anomalies in dynamic graphs (such as those generated by file transfers, but also intrusions)
- 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

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

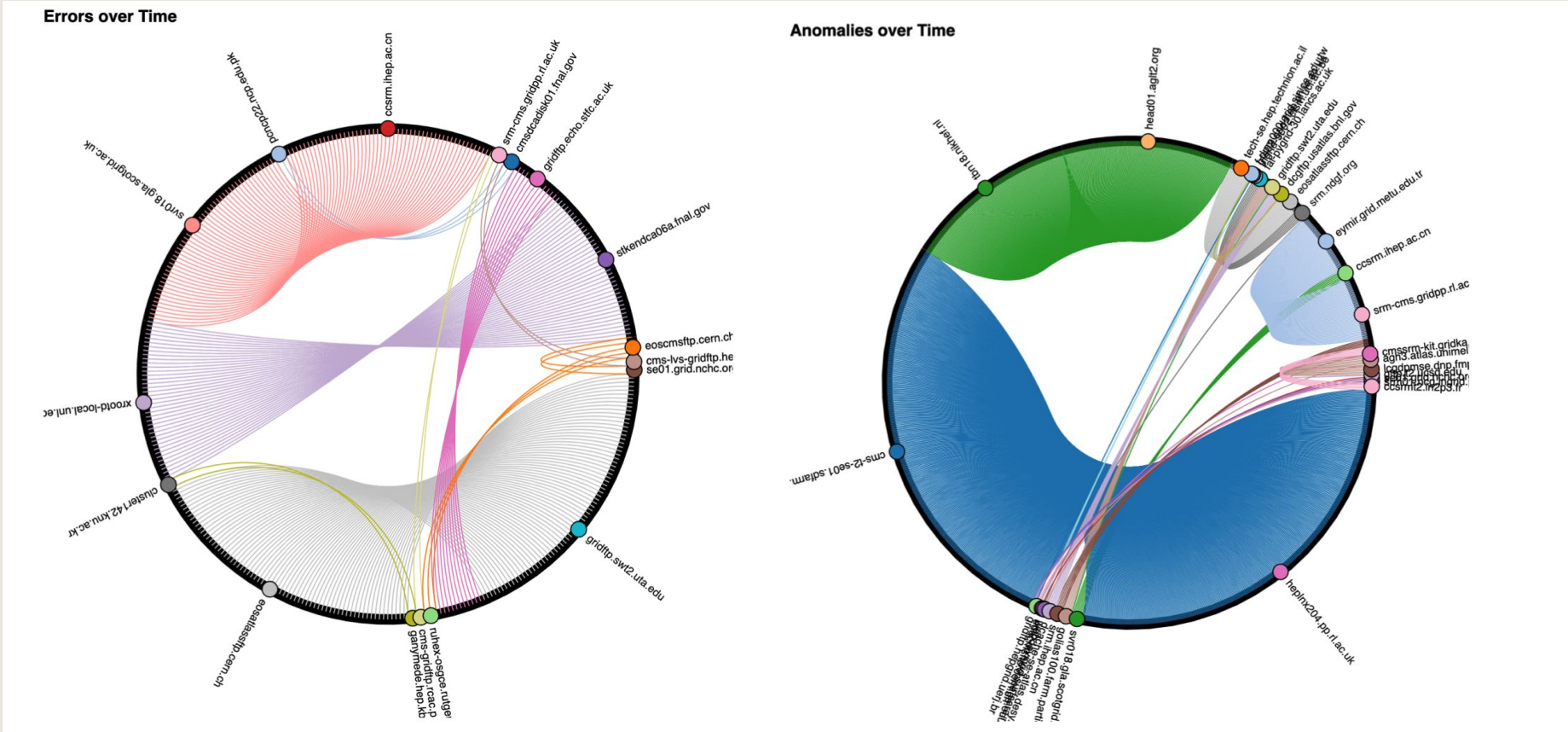
Figure 3: Count of errors over connection pairs

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 FTS transfers



Anomaly detection on transfers

- Next steps:
  - 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 GGUS tickets to validate the results
  - Build an interface for shifters to explore the results of this analysis
- This effort is now a pilot project in the EU CloudBank: https://ngiatlantic.eu/funded-experiments/cloudbank-eu-ngi

Outline

- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

Cloud anomaly detection

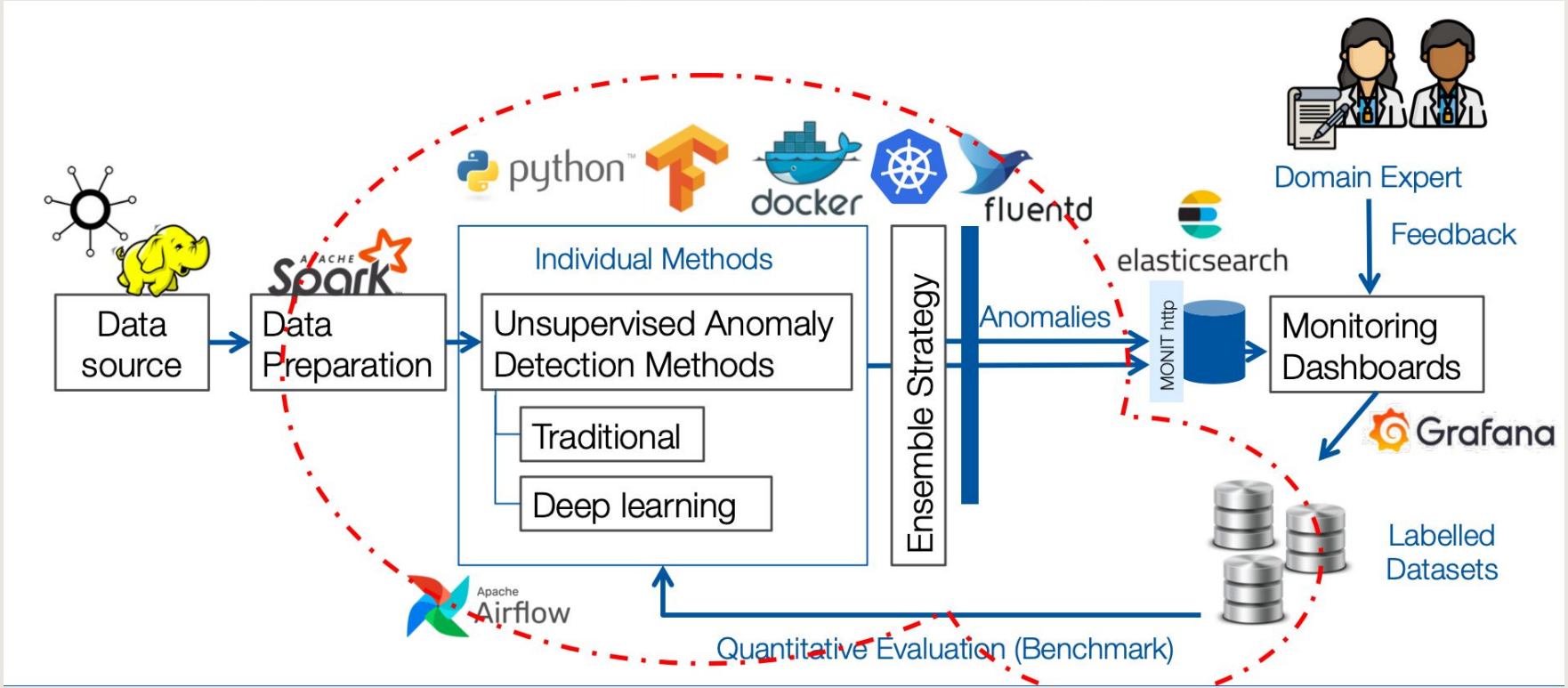
- A CERN based project to reliably detect anomalies in the CERN Cloud and help service managers to:
  - Identify operational issues
  - Get a comprehensive understanding of the cloud performance

- A grafana annotations enhancement has also been developed in parallel to:
  - Allow experts easily give feedback on the results, directly from Grafana.
  - Add the dashboard template variables as tags

- Don't miss the vCHEP talk "#20, Anomaly detection in the CERN cloud infrastructure"

More info: http://cern.ch/go/grf9

Cloud anomaly detection



Sites Optimisation

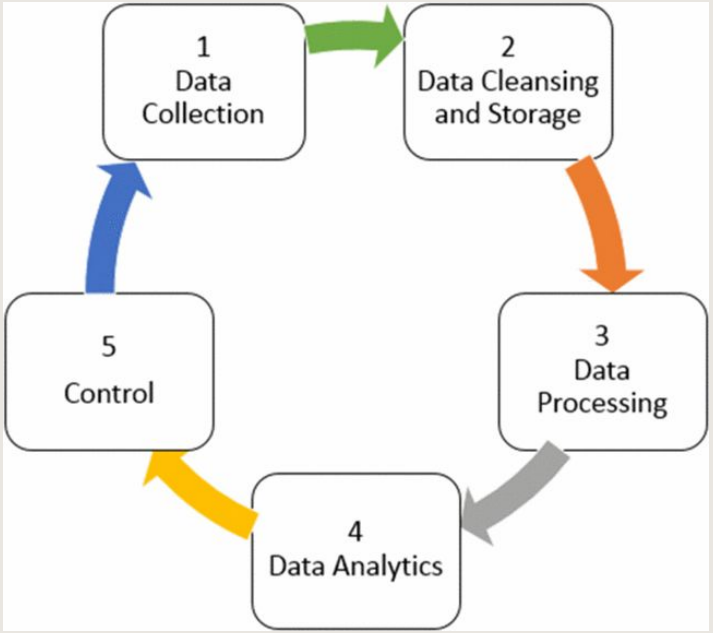
- We are also keeping an eye into what big companies from the industry do to automate their computing centers and reduce operational costs and environmental impact

- Of course in a diversified environment like WLCG these holistic strategies may not always apply

- The past years we have moved into a more unified processing pipeline in our sites, something which creates possibilities for collaborative efforts

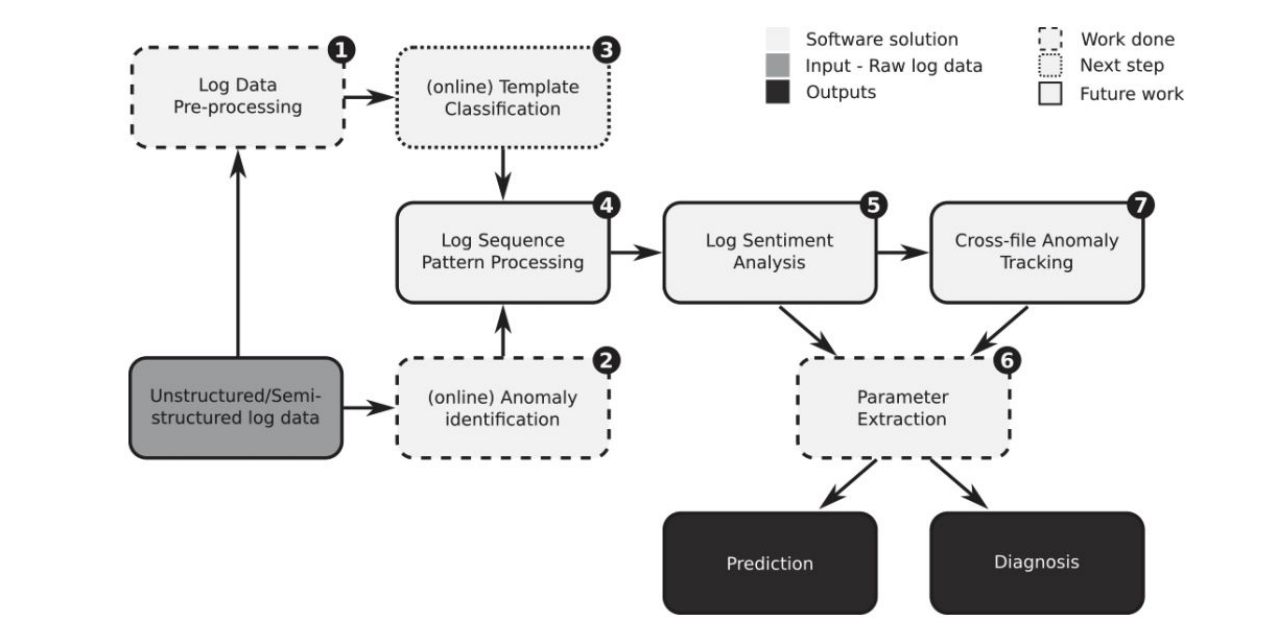
Industry examples

- 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



More info: http://cern.ch/go/B9KR

Outline

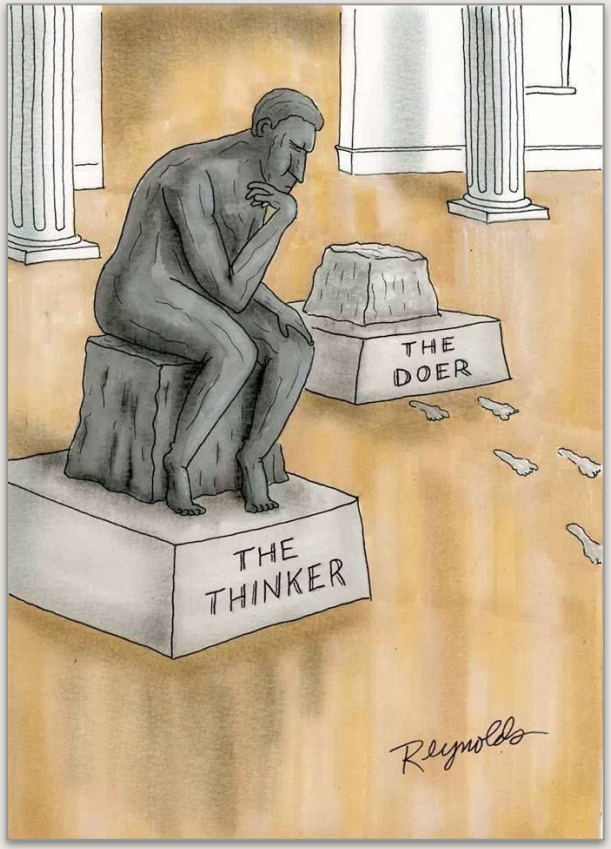
- About the Operational Intelligence initiative
- The infrastructure
- OpInt in workflow management
- OpInt in data management
- Computing center optimisation
- Conclusions

Status of our projects

Service	Status
Intelligent Alert System	In production
Shared k8s cluster	Work in progress
Jobs Buster	In production
HammerCloud JobShaping	In production
ClusterLogs / Grafana Dahsboard	In production
FTS log analysis	In testing phase
Anomaly Detection with Google	Work in progress
Cloud anomaly detection	In production
INFN Bologna - Predictive Maintenance	Work in progress

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 and there was already some progress done
- We will continue trying to span new collaborations
- Your feedback and your ideas are vital and always welcome



operational-intelligence@cern.ch
https://operational-intelligence.web.cern.ch/