Preparing distributed computing operations for HL-LHC era with Operational Intelligence

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

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

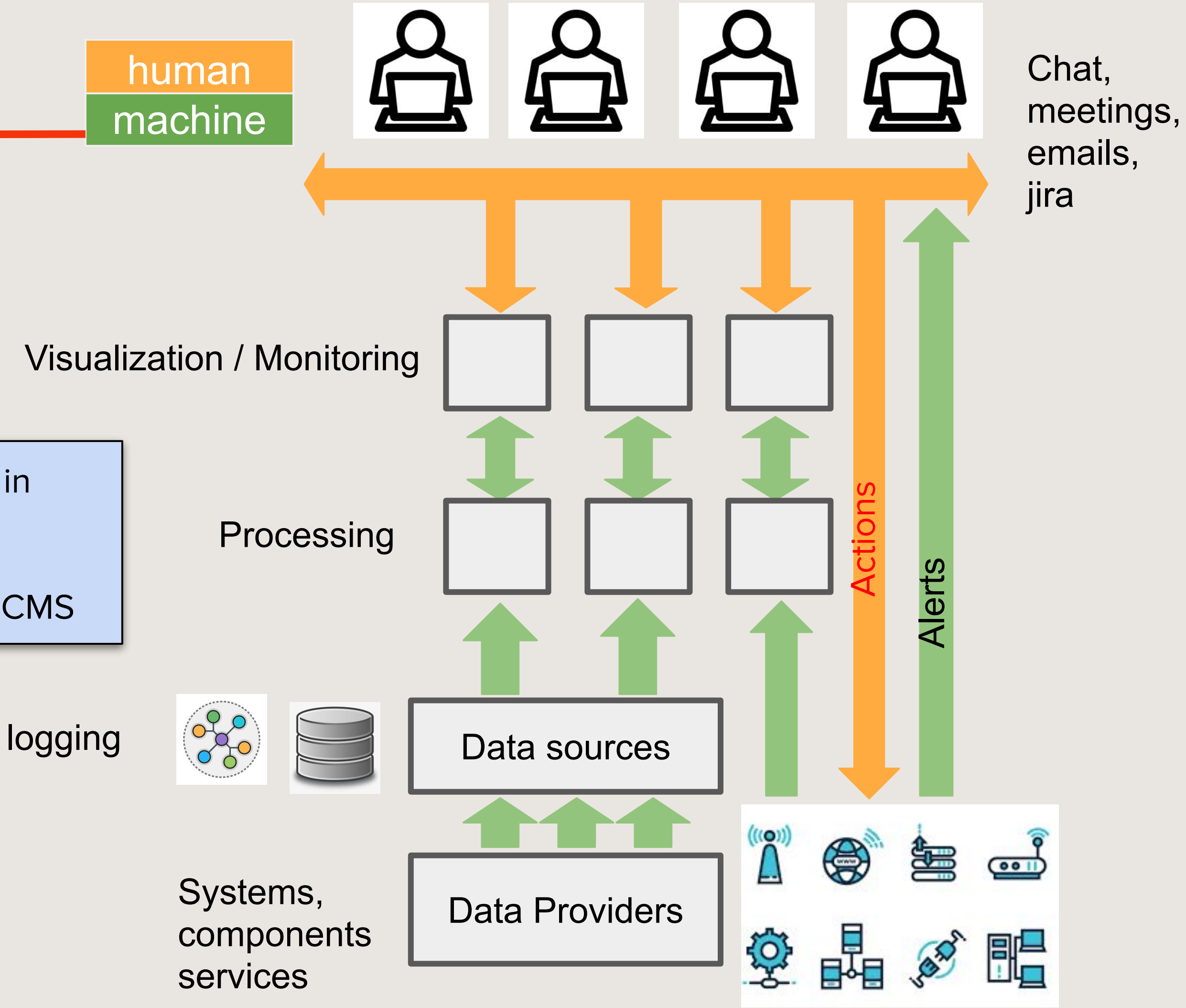
- A **cross-experiment** effort aiming to streamline computing operations:
  - **Improve resource utilization** by reducing the time needed to address operational issues
  - **Minimize human effort** for repetitive tasks by increasing the level of automation
  - **Build a community** of technical experts: 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?

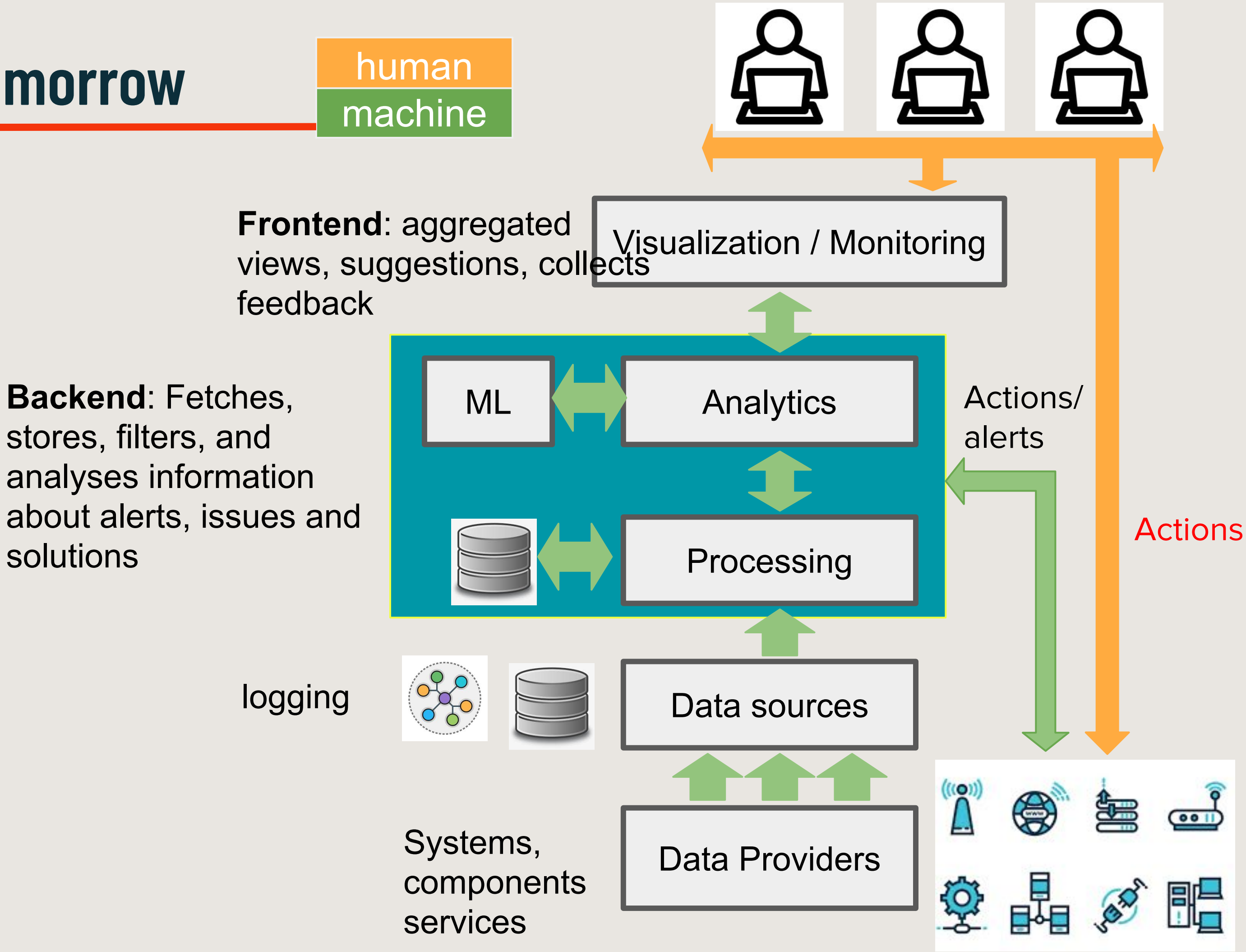
- **However**: computing operations (meta-)data is all archived.
    - We have logs for transfers, job submissions, site performances, infrastructure and services behaviours, storage accesses, ..
    - All this knowledge should be exploited!

Operations Today



ATLAS/CMS: A lot of people involved in Computing Operations
In 1 year:
> 1k GGUS 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

logging

Systems, components services

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

- We leverage open-source products for **Monitoring** and Operational Intelligence tasks
- The Kubernetes infrastructure is the de-facto standard for **deploying and scaling services**
- HTTP and AMQ are the main protocols for data injection
- Prometheus and ElasticSearch are main platforms 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
- Operations become easy with robust infrastructure and solid CLI tools
- Automation is a key to success
  - Data annotation, alerting, notifications, tagging, etc.

kubernetes

elasticsearch

Apache ACTIVEMQ

VICTORIA METRICS

Grafana

Apache Sqoop

Prometheus

kibana

hadoop

influxdb

NATS

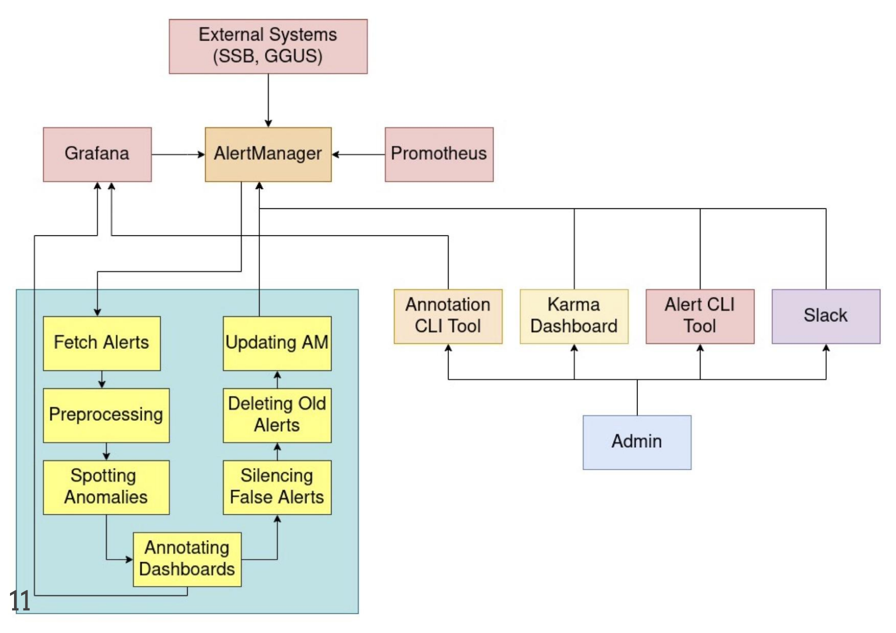
kafka

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.
- Using **open source tools** makes this effort experiment-agnostic

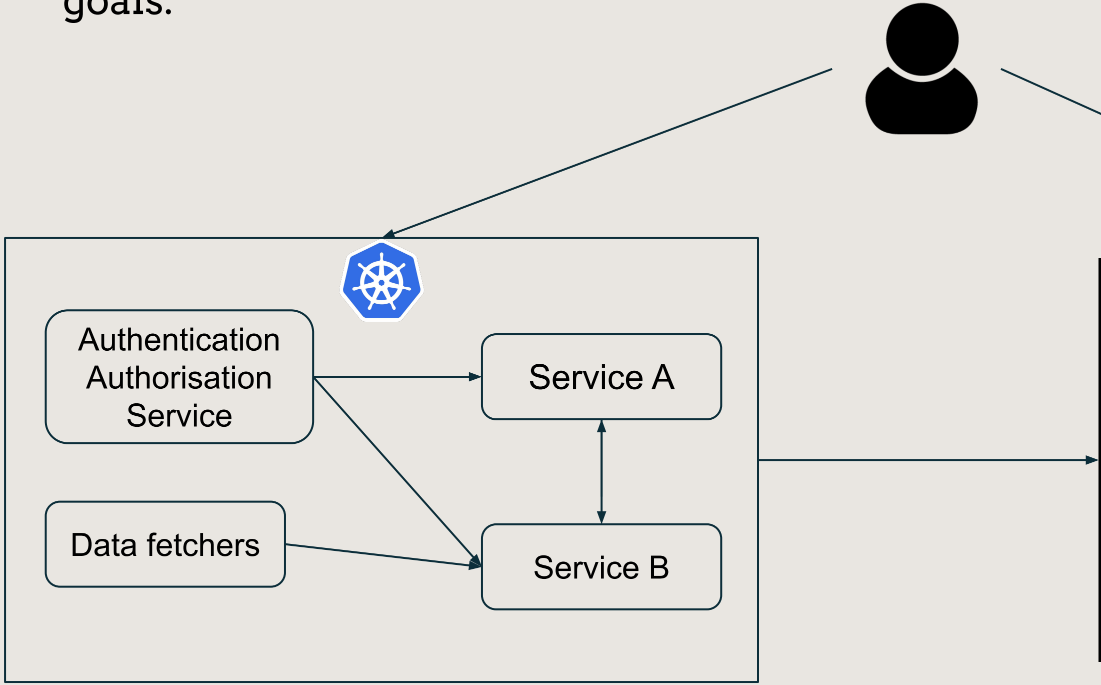


- SSB and GGUS are integrated into the Alert Manager.
- The alert manager fetches the existing alerts, filters them, and annotates Grafana dashboards based on the alert tag.
- Users can **add annotations** directly from the dashboard.
- Provides useful insights about when outages happen and how they affect the productivity reported by various systems in CMS dashboards.

http://cern.ch/go/cxg8

The shared k8s cluster

- Having a common space to deploy our applications is in line with our cross-experiment goals.



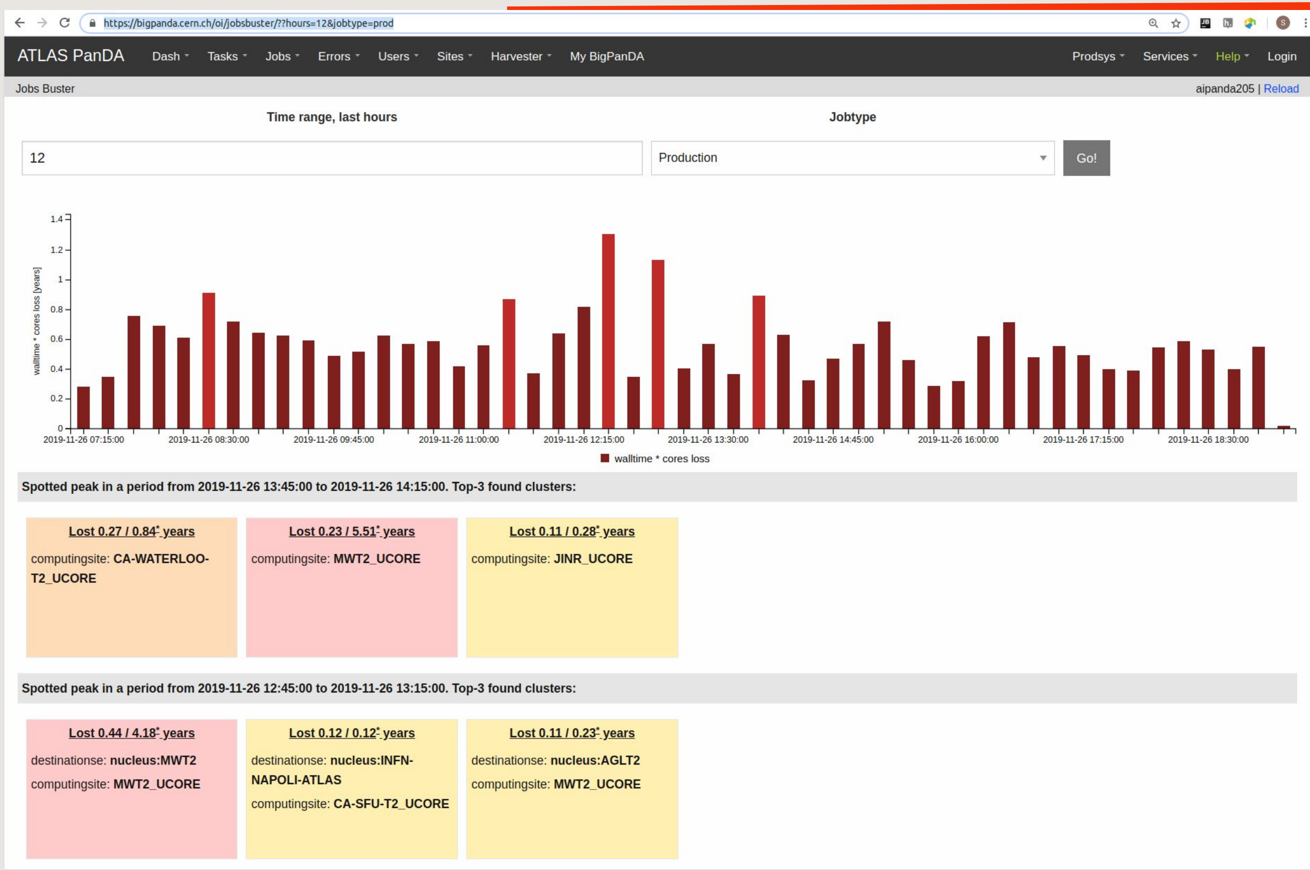
Outline

- 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.

- **NLP** 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)



Time range, last hours: 12

Jobtype: Production

Go!

Spotted peak in a period from 2019-11-26 13:45:00 to 2019-11-26 14:15:00. Top-3 found clusters:

Lost 0.27 / 0.84* years
computingsite: **CA-WATERLOO-T2_UCORE**

Lost 0.23 / 5.51* years
computingsite: **MWT2_UCORE**

Lost 0.11 / 0.28* years
computingsite: **JINR_UCORE**

Spotted peak in a period from 2019-11-26 12:45:00 to 2019-11-26 13:15:00. Top-3 found clusters:

Lost 0.44 / 4.18* years
destinationse: **nucleus:MWT2**
computingsite: **MWT2_UCORE**

Lost 0.12 / 0.12* years
destinationse: **nucleus:INFN-NAPOLI-ATLAS**
computingsite: **CA-SFU-T2_UCORE**

Lost 0.11 / 0.23* years
destinationse: **nucleus:AGLT2**
computingsite: **MWT2_UCORE**

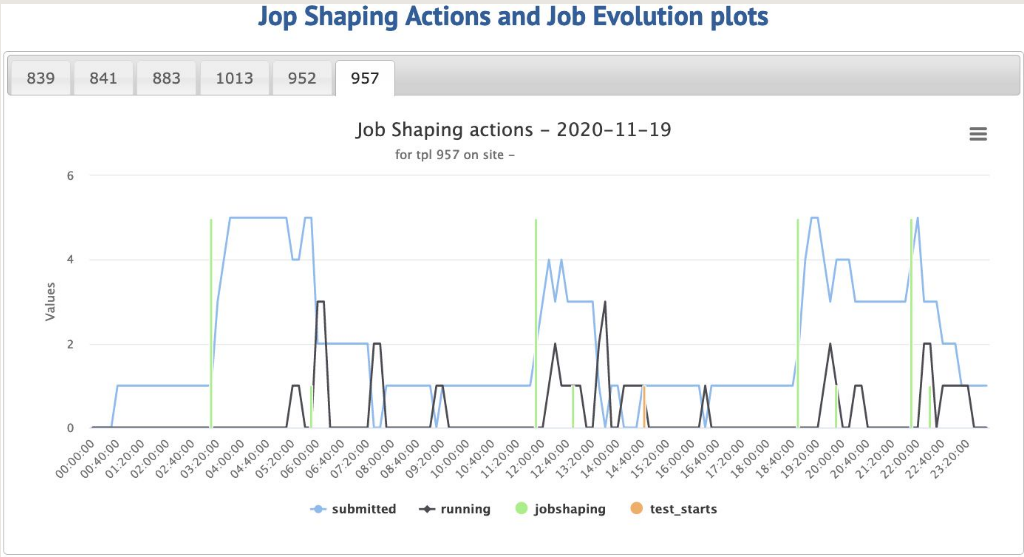
http://cern.ch/go/8qwC

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

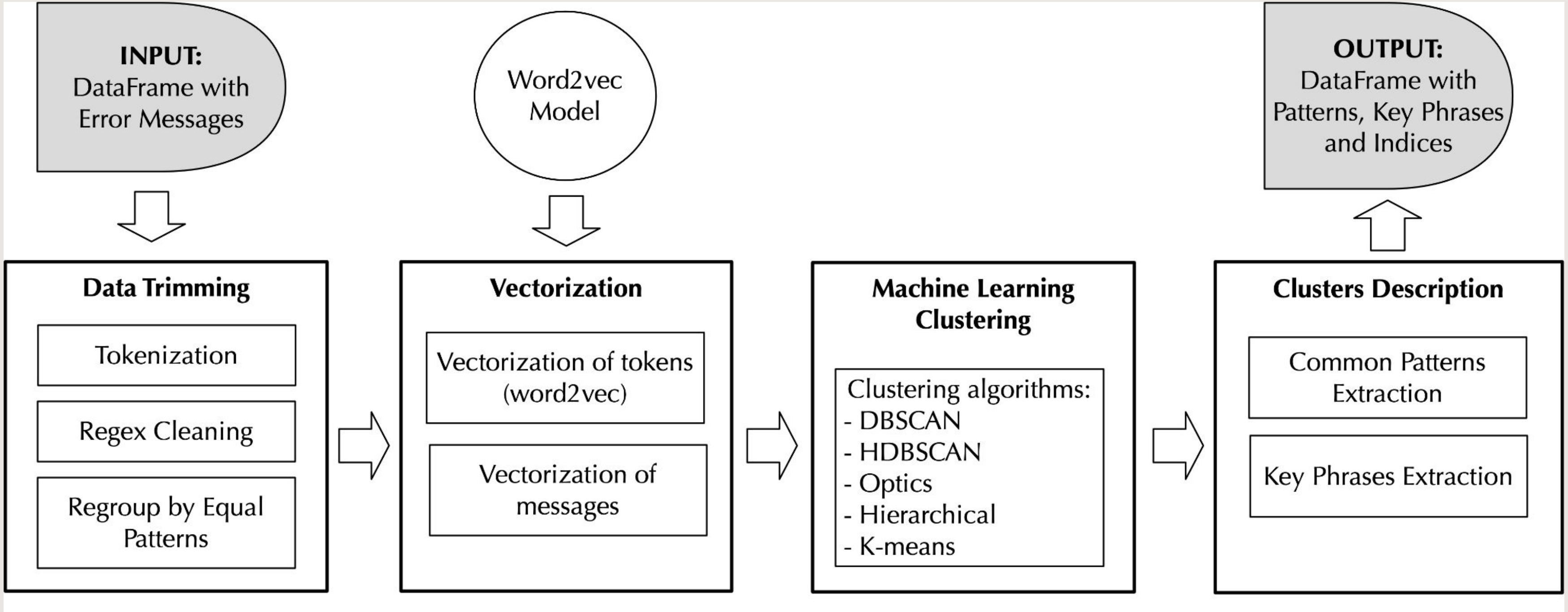
- Every day operation teams must deal with multiple data transfer errors.
- The monitoring systems help users to detect anomalies, to identify duplicated issues, to diagnose failures and to analyze failures retrospectively.
- Clustering of error messages is a possible way to simplify the analysis:
  - 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

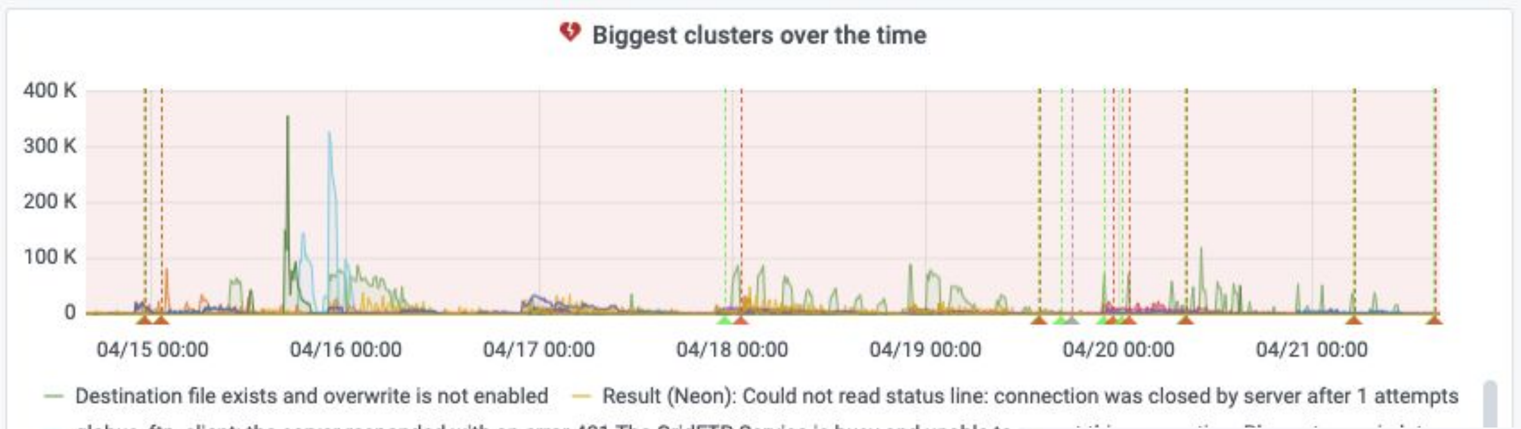
- ClusterLogs is one of the frameworks that was developed within OI to cluster error messages.

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



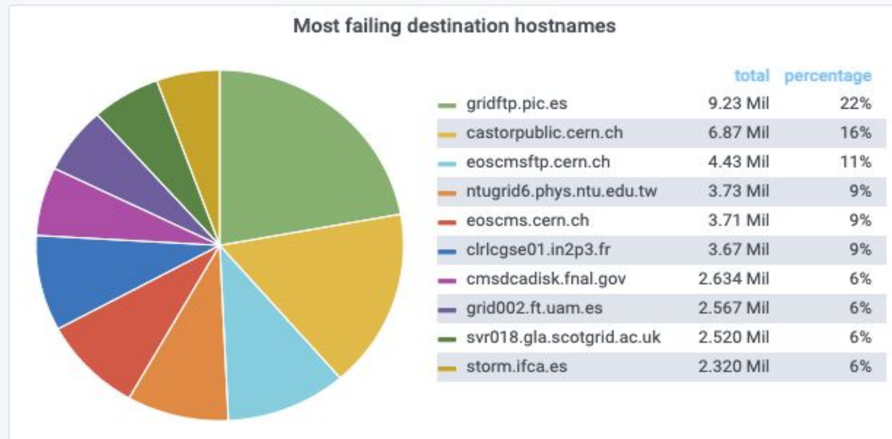
INPUT:
DataFrame with Error Messages
Word2vec Model
OUTPUT:
DataFrame with Patterns, Key Phrases and Indices
Data Trimming
Tokenization
Regex Cleaning
Regroup by Equal Patterns
Vectorization
Vectorization of tokens (word2vec)
Vectorization of messages
Machine Learning Clustering
Clustering algorithms:
- DBSCAN
- HDBSCAN
- Optics
- Hierarchical
- K-means
Clusters Description
Common Patterns Extraction
Key Phrases Extraction

Data Management: FTS log analysis



- Destination file exists and overwrite is not enabled
- Result (Neon): Could not read status line: connection was closed by server after 1 attempts
- globus_ftp_client: the server responded with an error 421 The GridFTP Service is busy and unable to
- Error reported from srm_ifce : 2 [SE][Ls][SRM_INVALID_PATH] No such file or directory
- [gfal2_stat][gfal_plugin_statG][davix2gliberr] Result HTTP 404 : File not found after 1 attempts
- R

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


Most failing destination hostnames

	total	percentage
gridftp.pic.es	9.23 Mil	22%
castorpublic.cern.ch	6.87 Mil	16%
eoscmsftp.cern.ch	4.43 Mil	11%
ntugrid6.phys.ntu.edu.tw	3.73 Mil	9%
eoscms.cern.ch	3.71 Mil	9%
clrlcgse01.in2p3.fr	3.67 Mil	9%
cmsdcadisk.fnal.gov	2.634 Mil	6%
grid002.ft.uam.es	2.567 Mil	6%
svr018.gla.scotgrid.ac.uk	2.520 Mil	6%
storm.ifca.es	2.320 Mil	6%

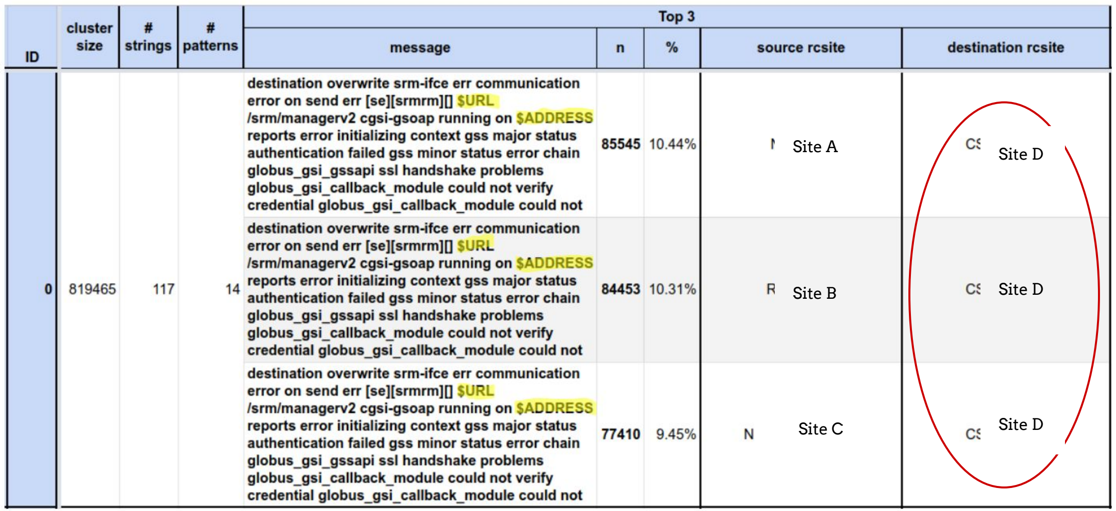
https://monit-grafana.cern.ch/d/Zx_bXneWz/fts-log-clustering?from=now-7d&orgId=11&to=now

Data Management: FTS log Analysis

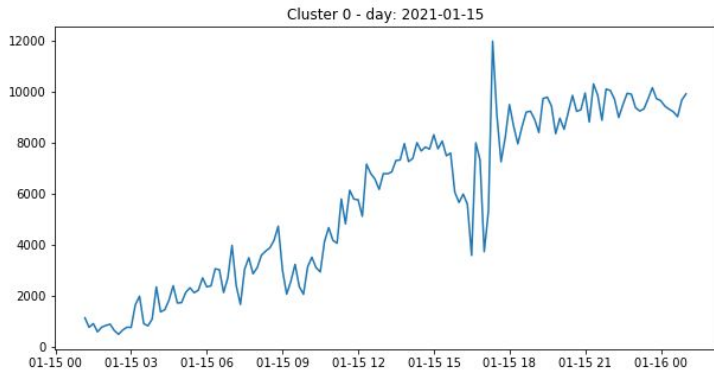
- Similar effort to clusterize FTS error messages, and validate the results using GGUS tickets

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%	N Site A	CS 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%	R Site B	CS 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%	N Site C	CS Site D



Time/# of errors for selected cluster

Anomaly detection on FTS transfers

- Google showed interest in trying to help us 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

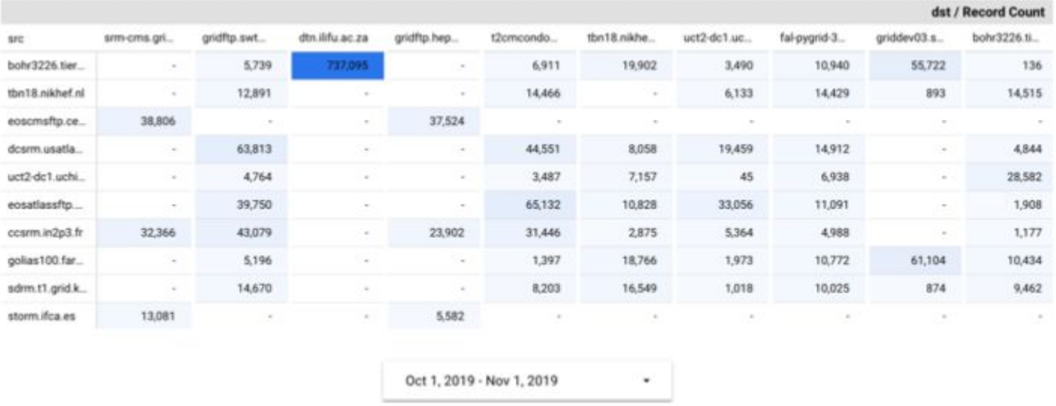
MIDAS (MIcrocluster-based Detector of Anomalies in Streams) :

- 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

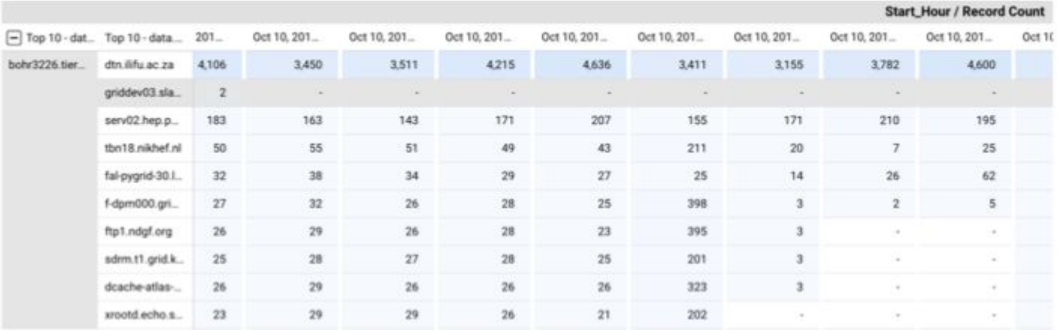


dst / Record Count

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



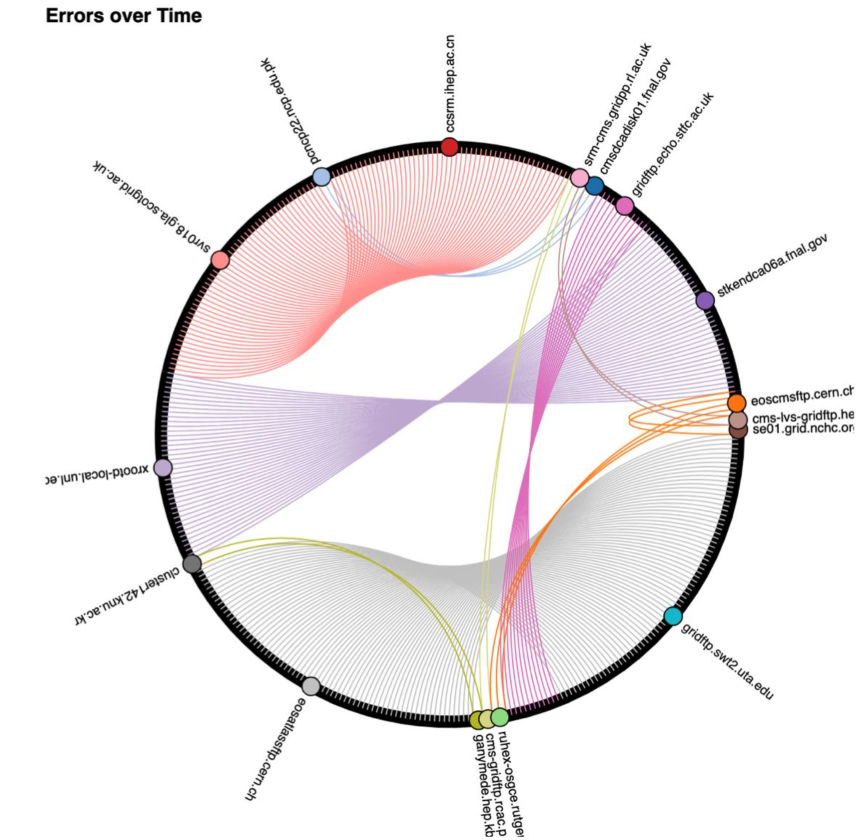
Start_Hour / Record Count

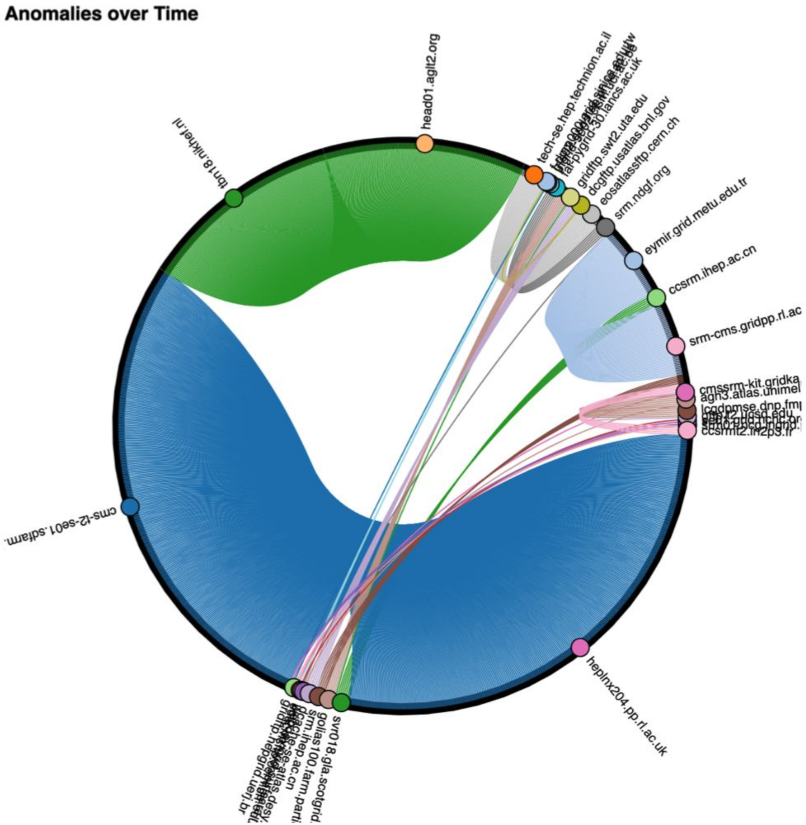
⊟ Top 10 - dat...	Top 10 - data...	201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 1
bohr3226.tier...	dtn.ilifu.ac.za	4,106	3,450	3,511	4,215	4,636	3,411	3,155	3,782	4,600	
	griddev03.sla...	2	-	-	-	-	-	-	-	-	
	serv02.hep.p...	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

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

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.

Outline

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

Cloud anomaly detection

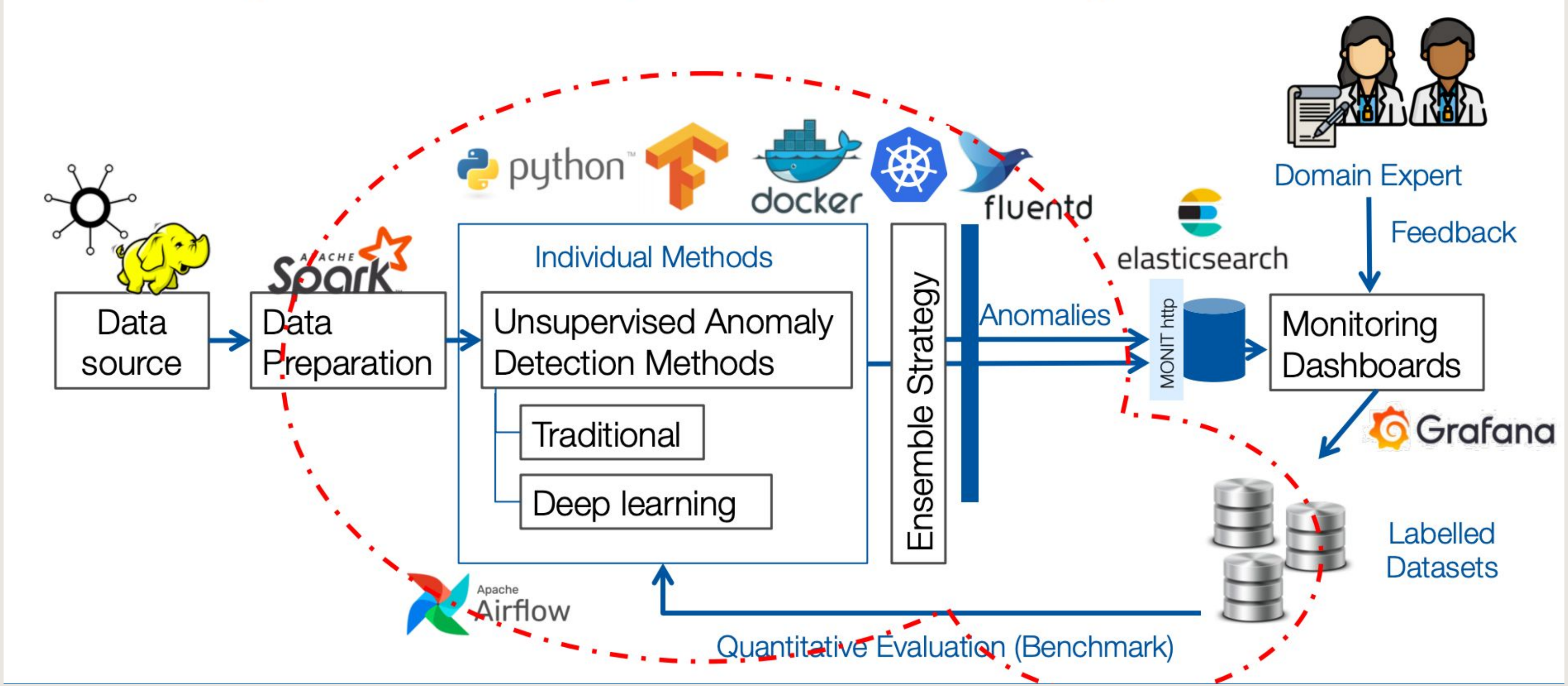
- A CERN based project to reliably detect anomalies in the CERN Cloud and help service managers to:
  - Identify operational issues
  - Get a comprehensive understanding of the cloud performance.

- A grafana annotations enhancement has alsobeen 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"

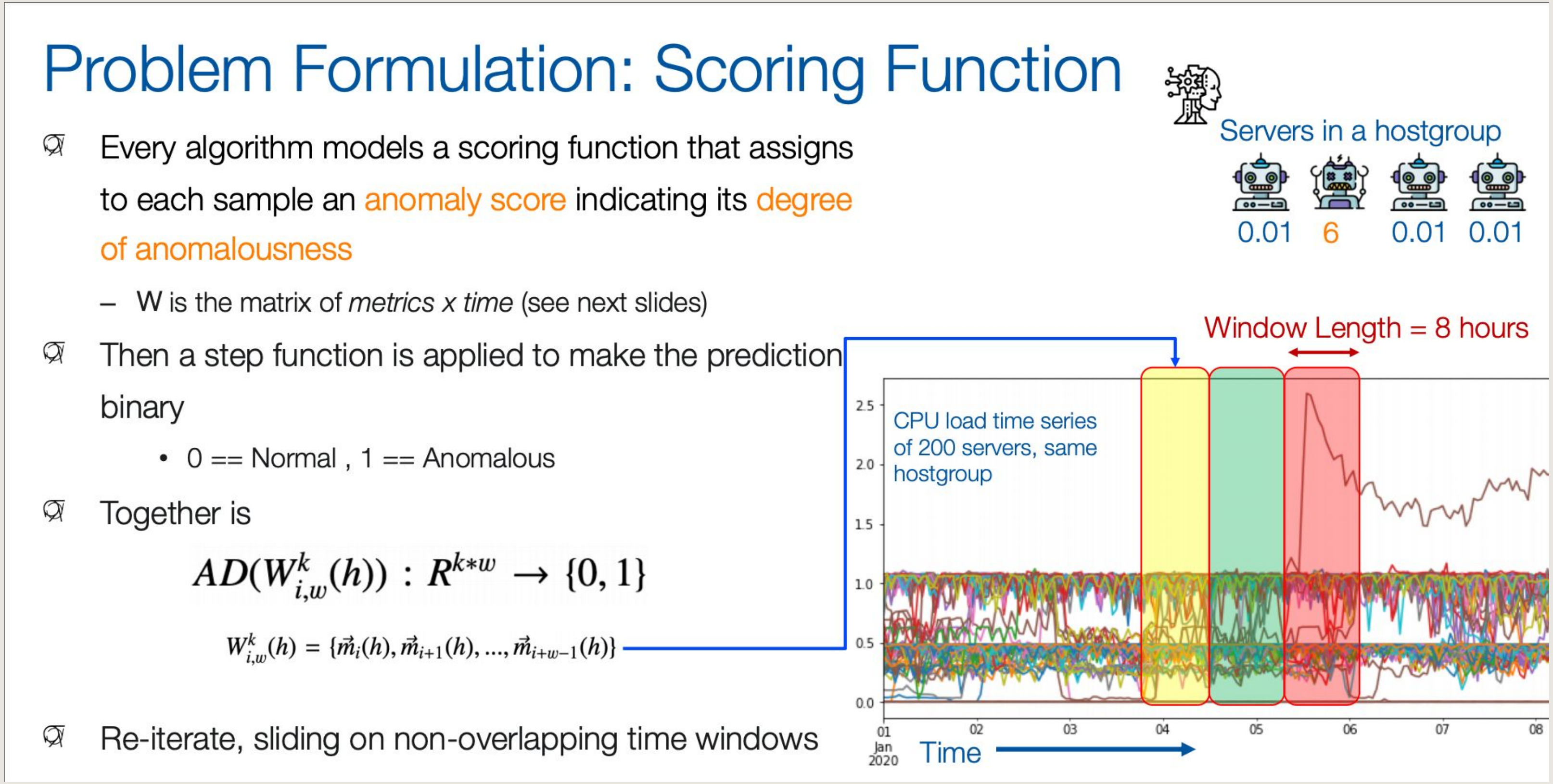
http://cern.ch/go/grf9

Cloud anomaly detection



http://cern.ch/go/grf9

Cloud anomaly detection



Problem Formulation: Scoring Function

- Every algorithm models a scoring function that assigns to each sample an anomaly score indicating its degree of anomalousness
  - W is the matrix of *metrics x time* (see next slides)
- Then a step function is applied to make the prediction binary
  - 0 == Normal , 1 == Anomalous
- Together is

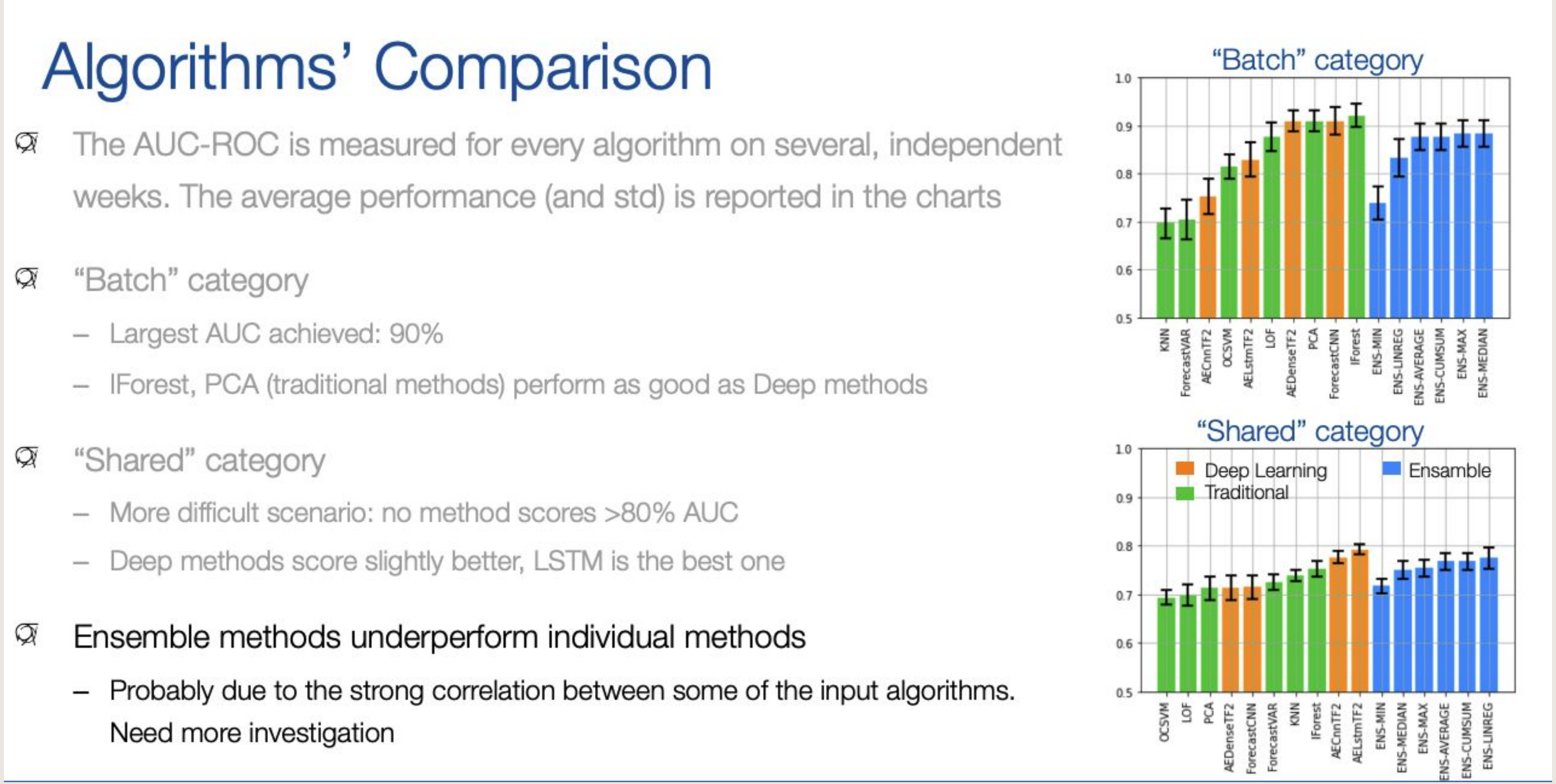
$$AD(W^k_{i,w}(h)) \; : \; R^{k*w} \;\to\; \{0, 1\}$$

$$W^k_{i,w}(h) = \{\vec{m}_i(h), \vec{m}_{i+1}(h), ..., \vec{m}_{i+w-1}(h)\}$$

- Re-iterate, sliding on non-overlapping time windows

http://cern.ch/go/grf9

Cloud anomaly detection



Algorithms’ Comparison

- The AUC-ROC is measured for every algorithm on several, independent weeks. The average performance (and std) is reported in the charts
- “Batch” category
  - Largest AUC achieved: 90%
  - IForest, PCA (traditional methods) perform as good as Deep methods
- “Shared” category
  - More difficult scenario: no method scores >80% AUC
  - Deep methods score slightly better, LSTM is the best one
- Ensemble methods underperform individual methods
  - Probably due to the strong correlation between some of the input algorithms. Need more investigation

http://cern.ch/go/grf9

Sites Optimisation

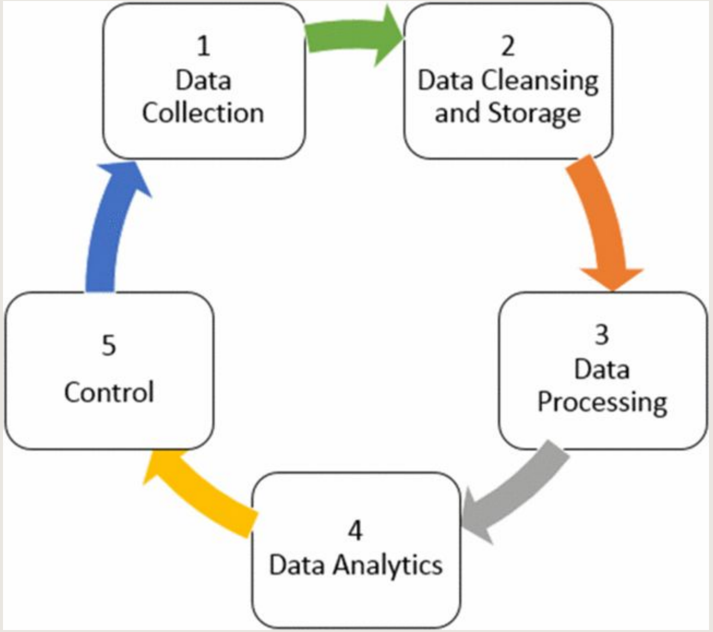
- We are also keeping an eye into what big companies from the industry do to automate their computing centers and reduce operational costs and environmental impact.

- 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

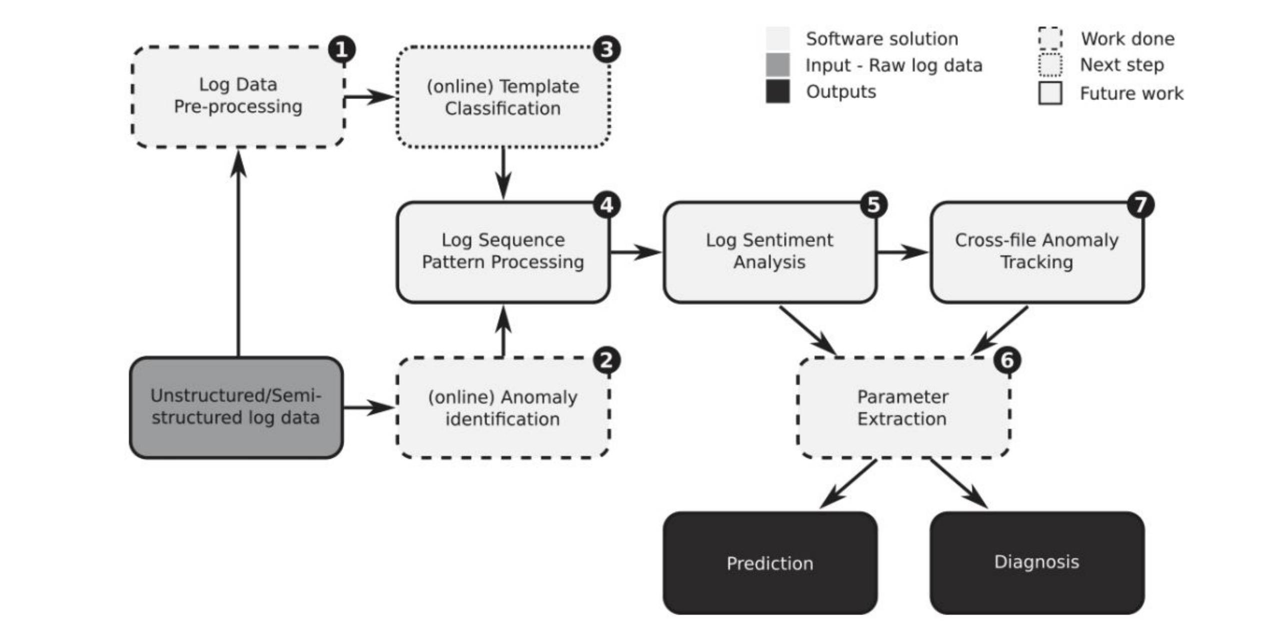
- A lot of interesting hardware related work:
  - Building sensors throughout their networks so that they can redirect workload to offload overloaded nodes
  - Using SMART (Self-Monitoring, Analysis and Reporting Technology) to derive disk failure predictions and replace hardware proactively
  - Using 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.



Software solution
Input - Raw log data
Outputs
Work done
Next step
Future work
1 Log Data Pre-processing
3 (online) Template Classification
4 Log Sequence Pattern Processing
5 Log Sentiment Analysis
7 Cross-file Anomaly Tracking
Unstructured/Semi-structured log data
2 (online) Anomaly identification
6 Parameter Extraction
Prediction
Diagnosis

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

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