Operational Intelligence

Federica Legger

on behalf of the Operational Intelligence forum

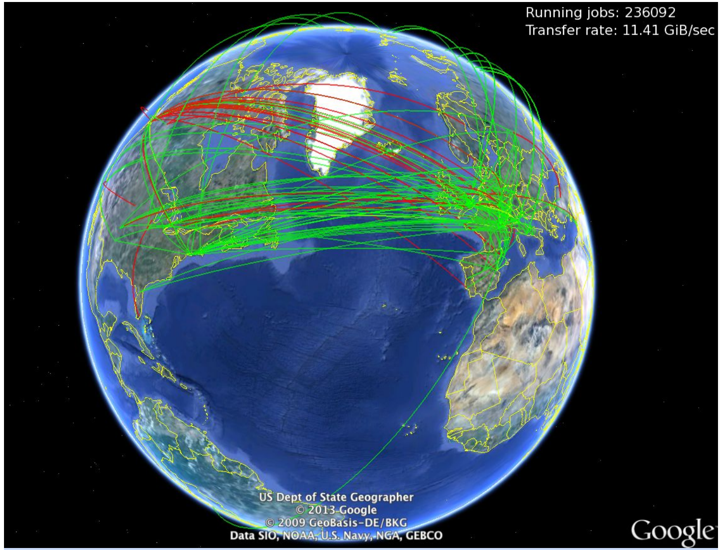
Operational Intelligence

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

- A **cross-experiment** effort aiming to streamline computing operations:
  - Minimize human effort in operations by increasing **automation**
  - **Improve resource utilization** by reducing wasted cycles
  - **Build a community** of technical experts: critical mass to have impact
- **Our mission**:
  - Identify **common projects**
  - Leverage **common tools/infrastructure**
  - **Collaborate**, share expertise, tools & approaches
    - Across experiments
    - Across teams (operations, monitoring, developers)

WLCG - Worldwide LHC computing grid

- Distributed computing infrastructure that provides computing services and storage resources to process and store LHC data
- WLCG is made of 900 000 computer cores from over 170 sites in 42 countries
- WLCG runs over 2 million tasks per day and, by the end of LHC Run 2, global transfer rates regularly exceeded 60 GB/s
- Close to an exabyte of LHC data already collected and stored



- Distributed workload management (WM)
- Distributed data management (DM)
- Sites and facilities

Preparing for High Lumi LHC

- LHC experiments built successful computing systems for LHC Run 1/2
  - We do not have a simulation of the Grid
  - Up to now we monitored to debug in near-time.
    - Can we do better?
- HL-LHC: one order of magnitude more resources than today
  - Personpower will not scale
- **However**: computing operations (meta-)data is all archived
  - We have logs for transfers, job submissions, site performances, infrastructure and service behaviours, storage access
  - All this knowledge should be exploited!

Ongoing Efforts

- 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:
  - **Workflow Management**
  - **Data Management**
  - **Computing facilities**

- We provide some **shared infrastructure**:
  - A common k8s cluster for services to be deployed.
  - A framework which can be used to develop new tools

ML areas of interest

- **Classification**: predict which transfers/jobs will fail
  - Decision trees, neural networks, ...
- **Regression:** predict time-to-complete for jobs, transfers, campaigns
- **Clusterization**: group alerts, errors, root cause analysis
  - K-means, DBscan, PCA, ...
- **Natural Language Processing (NLP):** log text analysis, tickets
  - Word2vec, transformers, …
- **Anomaly detection:** detect problematic links, hosts, storage end-points
  - Time series, multi-variate, ...

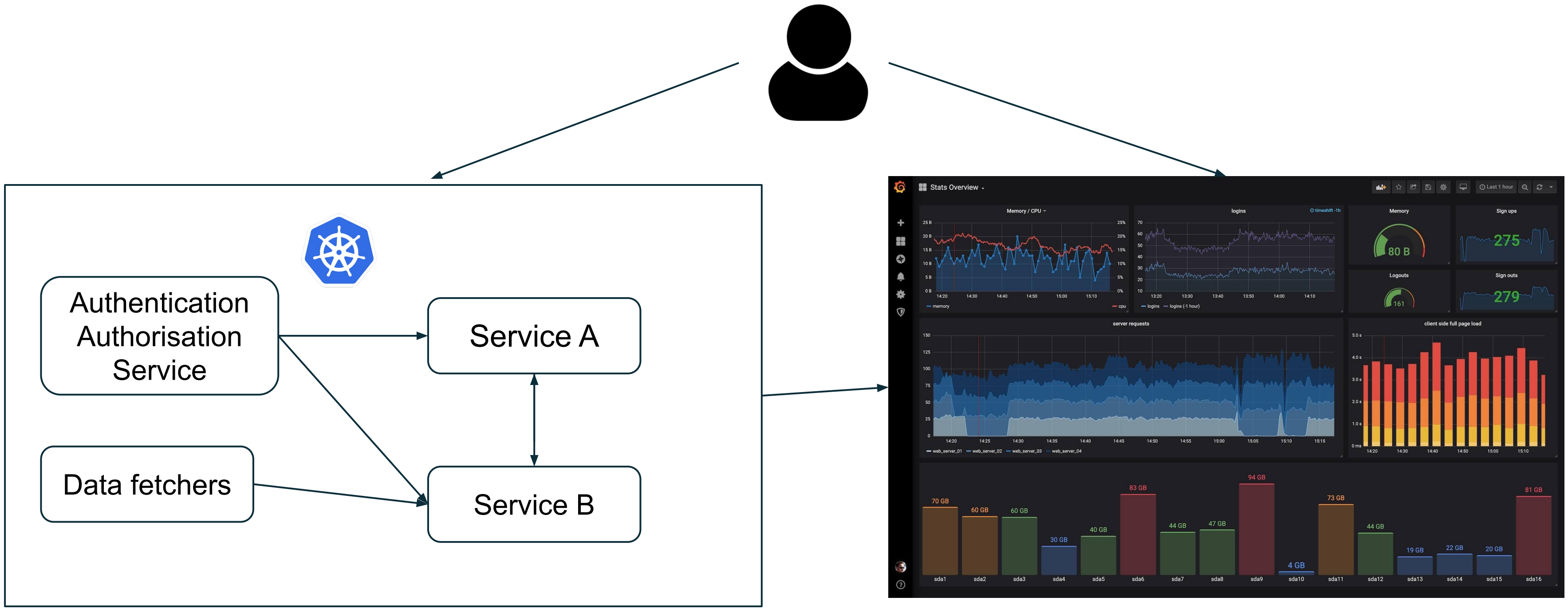
Infrastructure

- Based on **open-source** products
  - Kubernetes for **deploying and scaling services**
  - HTTP and AMQ for data injection
  - Prometheus and ElasticSearch for **managing metrics and meta-data**
- Features:
  - Clear separation of Data, Infrastructure, Visualization
  - Data standardization, common naming convention, data validation
- Automation:
  - Data annotation, alerting, notifications, tagging

kubernetes elasticsearch VICTORIA METRICS Grafana Prometheus Apache ACTIVEMQ influxdb Apache Sqoop kibana hadoop NATS kafka

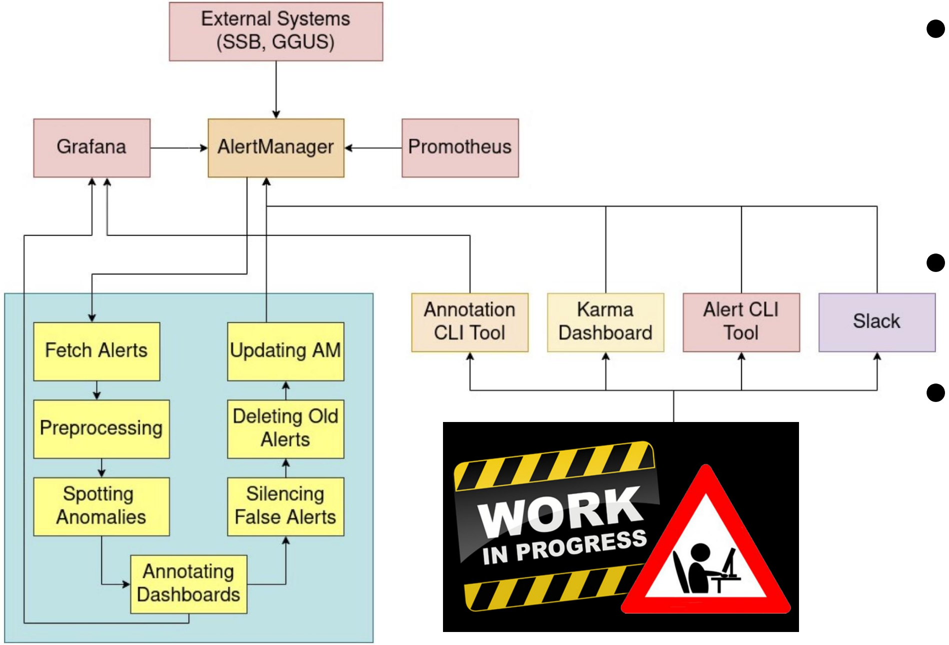
The shared k8s cluster

- common (not experiment specific) space to deploy applications



Intelligent Alert system

- We added an intelligent layer to CMS monitoring infrastructure to **detect, analyze and predict abnormal system behaviors** using **alerts**

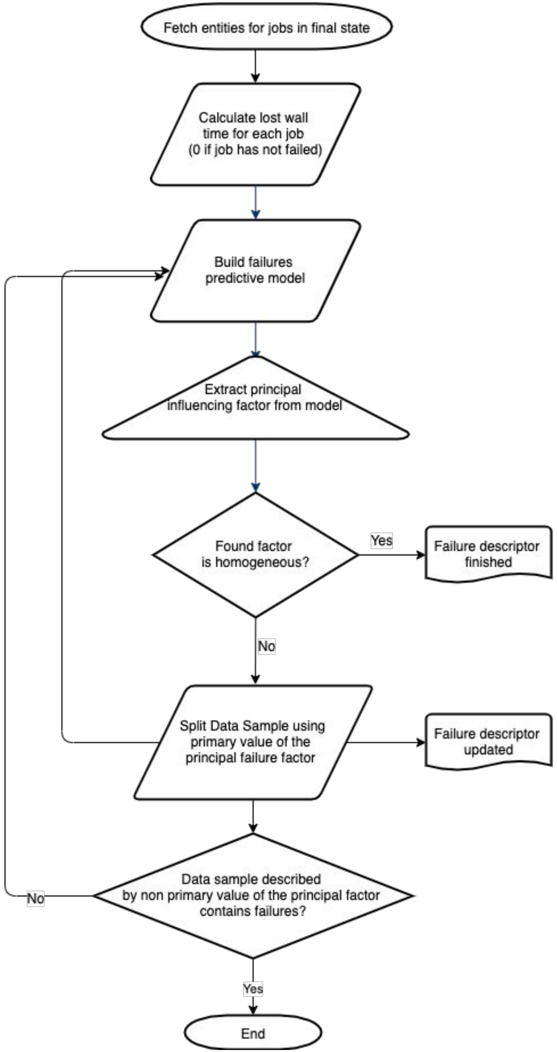


- The alert manager fetches the existing alerts, filters them, and **annotates Grafana** dashboards based on the alert tag
- SSB and GGUS are also integrated into the Alert Manager
- The system provides useful insights about when outages happen and how they affect the productivity reported by various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

More info: http://cern.ch/go/cxg8

Workload management: Jobs Buster

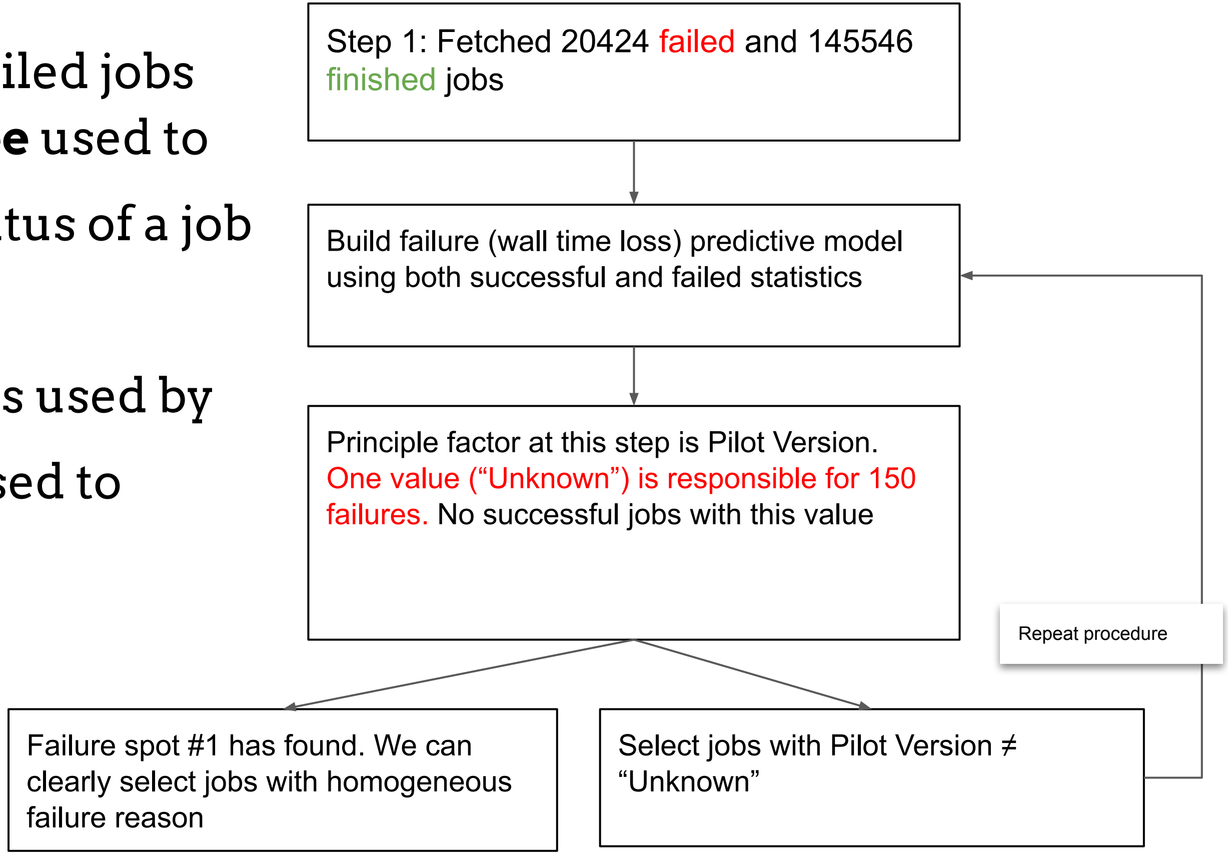
- ATLAS " Jobs Buster" tries to spot **operational problems** in submitted jobs
- Uses *catboost* library
  - Naturally supports both categorical and numerical features (walltime, CPU time, pilot version, software version, error message, site)
  - Supports GPU and CPU



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

Predictive model

- Input: successful and failed jobs
- **Gradient Boosting Tree** used to predict the outcome status of a job by its features
- Most important features used by GBTs are ranked and used to extract the root cause
- Iterative procedure



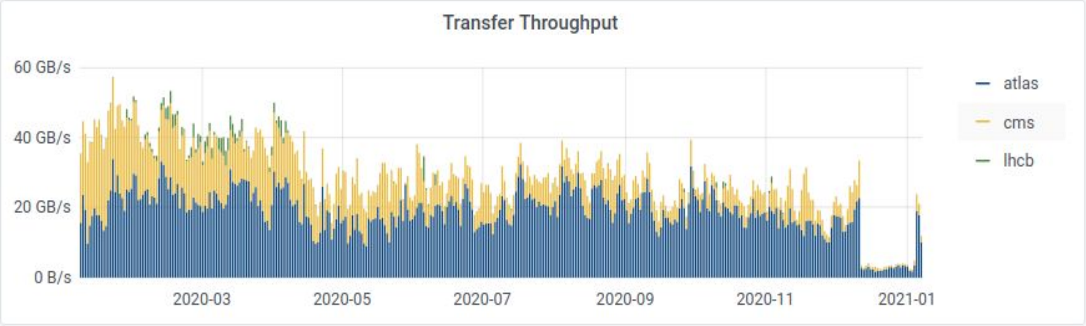
Results

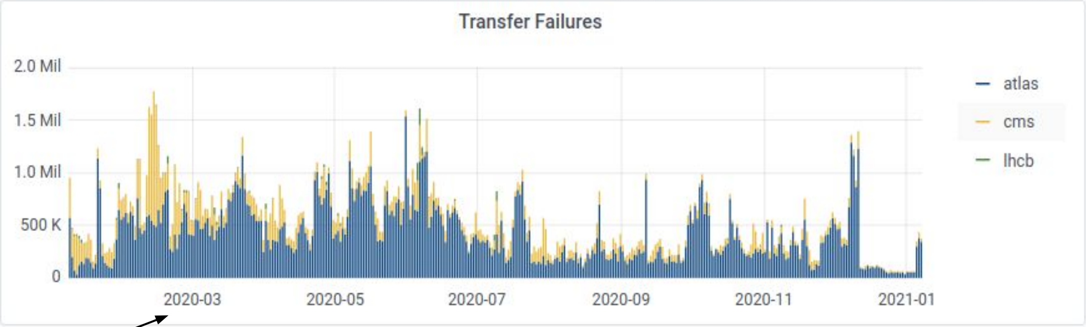

ATLAS PanDA Dash Tasks Jobs Errors Users Sites Harvester My BigPanDA
Job by ID
Enter...
Prodsys Services Help Siarhei
Jobs Buster
Time window: 2021-05-13T02:02:11 - 2021-05-13T14:02:11 UTC
Change
aipanda208 |Refresh
Order by lost walltime
Order by failures counts
GRID+HPC
walltime * cores loss [years]
N_0 N_1 N_2 N_3 N_4 N_5 N_6 N_7 N_8 N_9 N_10 N_11 N_12 N_13 N_14 N_15 N_16 N_17 N_18 N_19
N_0: lost 14.44 years / 856 jobs failed
REQID: 37255
show errors
N_1: lost 6.77 years / 536 jobs failed
COMPUTINGSITE: ARNES
COMPUTINGELEMENT: hpc.arnes.si
show errors
N_2: lost 3.01 years / 157 jobs failed
JEDITASKID: 25491855
REQID: 37108
show errors
N_3: lost 2.68 years / 449 jobs failed
JEDITASKID: 25446551
REQID: 37077
show errors
N_4: lost 1.69 years / 462 jobs failed
COMPUTINGSITE: UKI-SOUTHGRID-SUSX_UCORE
COMPUTINGELEMENT: grid-arc-01.hpc.susx.ac.uk:2811
show errors
N_5: lost 1.19 years / 517 jobs failed
NUCLEUS: TRIUMF-LCG2
show errors
N_6: lost 1.11 years / 290 jobs failed
COMPUTINGSITE: CERN-HPC
show errors
N_7: lost 1.08 years / 390 jobs failed
DESTINATIONSE: nucleus:IFIC-LCG2
show errors
N_8: lost 0.76 years / 258 jobs failed
COMPUTINGELEMENT: cet03.cern.ch:9619
show errors
N_9: lost 0.76 years / 230 jobs failed
NUCLEUS: NIKHEF
show errors
N_10: lost 0.61 years / 182 jobs failed
NUCLEUS: INFN-T1
show errors
N_11: lost 0.35 years / 4172 jobs failed
COMPUTINGELEMENT: ce01.tier2.hep.manchester.ac.uk:2811
show errors
N_12: lost 0.31 years / 57 jobs failed
COMPUTINGSITE: SWT2_CPB
REQID: 37041
show errors
N_13: lost 0.21 years / 349 jobs failed
COMPUTINGSITE: AGLT2_TEST
show errors
N_14: lost 0.21 years / 167 jobs failed
COMPUTINGSITE: Taiwan-LCG2-HPC_Unified
COMPUTINGELEMENT: arc-fdr1.grid.sinica.edu.tw:2811
show errors
N_15: lost 0.2 years / 12 jobs failed
JEDITASKID: 25492197
show errors
N_16: lost 0.18 years / 533 jobs failed
COMPUTINGSITE: RAL
DESTINATIONSE: nucleus:RAL-LCG2
show errors
N_17: lost 0.15 years / 88 jobs failed
COMPUTINGSITE: NET2
show errors

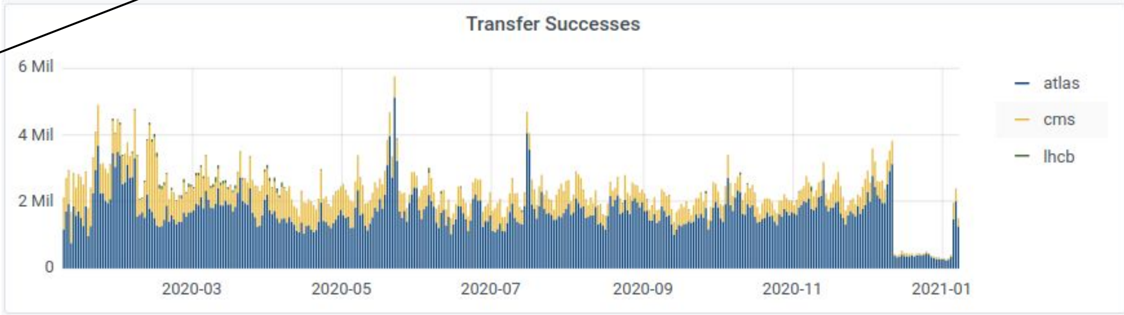
Data Management - FTS

FTS (FileTransferService) is the service responsible for globally distributing the **multiple petabytes of LHC data** across the WLCG infrastructure.

~10^5-10^6 error messages per day







Analysis of FTS error messages

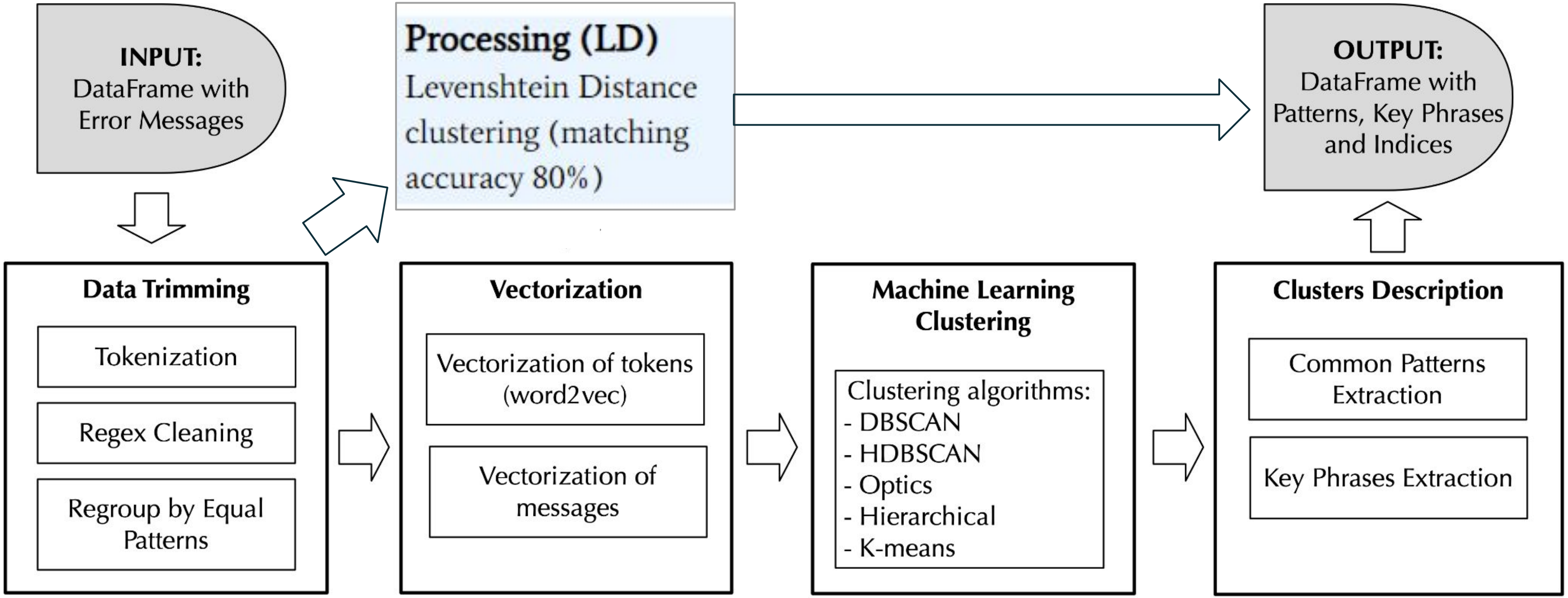
- Every day operation teams must deal with multiple data transfer errors
- The monitoring systems help users to detect anomalies, to identify duplicated issues, to diagnose failures and to analyze failures retrospectively
- Clustering of error messages is a possible way to simplify the analysis:
  -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

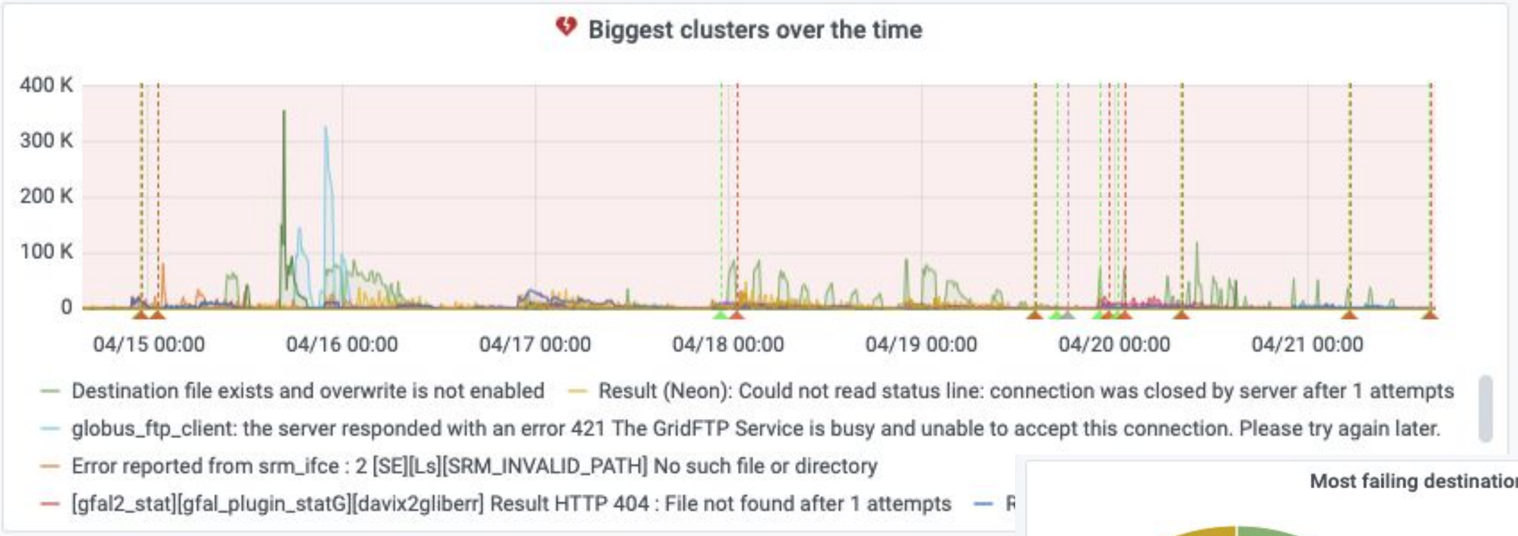
ClusterLogs

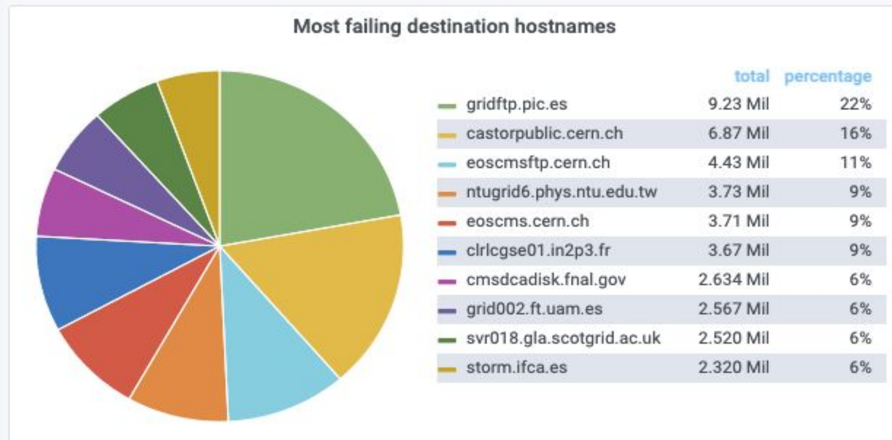
- ClusterLogs is a framework developed within OI to cluster error messages

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



CMS FTS log analysis with LD



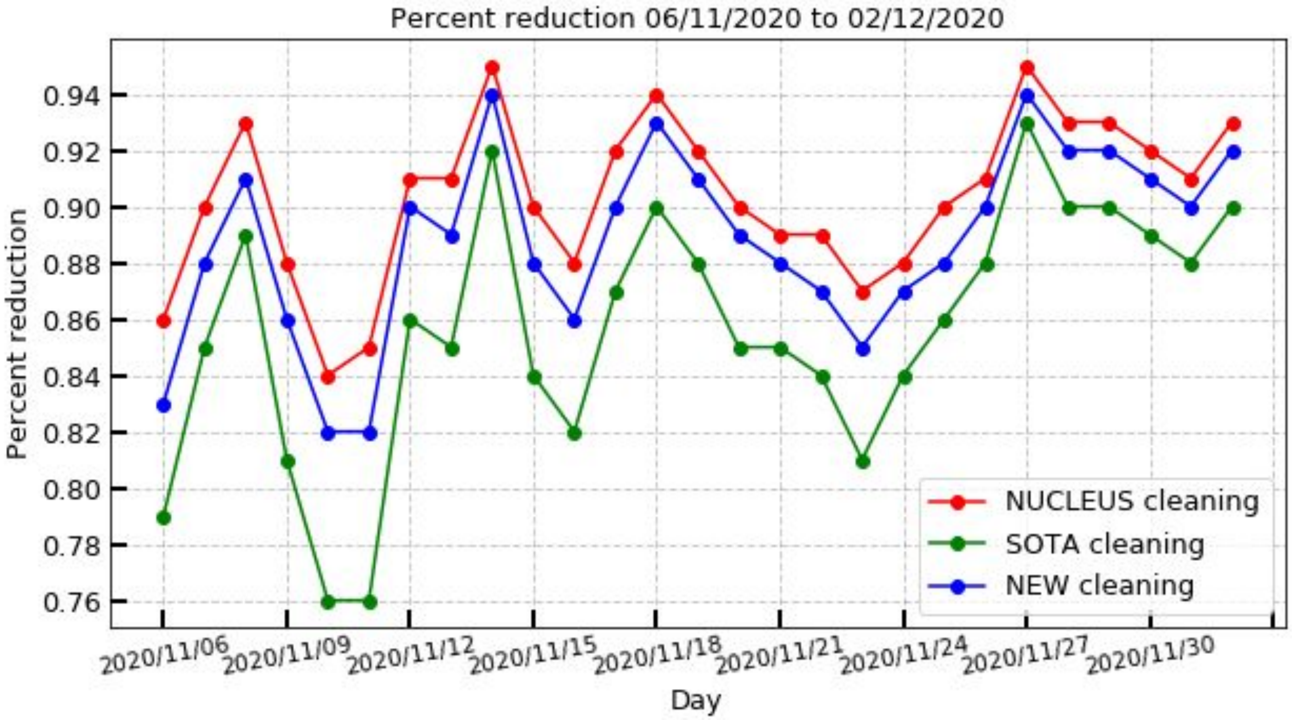


ClusterLogs is used to classify File Transfer Service (FTS) logs and results pushed back to the MONIT infrastructure where they can be browsed from a Grafana dashboard

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

CMS FTS log analysis with ML vs LD

- ML model (Word2vec+DBscan) requires more cleaning than LD approach
- Less clusters with ML, but top 10 clusters basically the same
  - Difference of 1-2 clusters/day



Clustering FTS Errors - Method

Vectorization (offline)

- Transform textual information into numerical
- Minimal pre-processing (split URLs and remove punctuation)

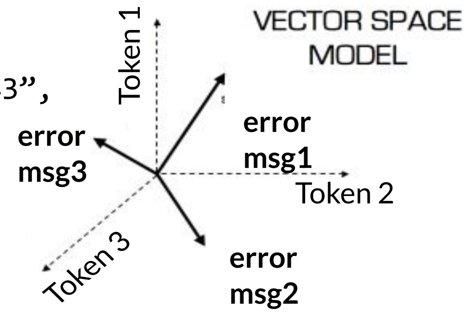
Raw message:

```
[se][statusofgetrequest][etimedout]
srm://uct2-dc1.uchicago.edu:8443/sr
m/managerv2: user timeout over.
```

Tokenization:

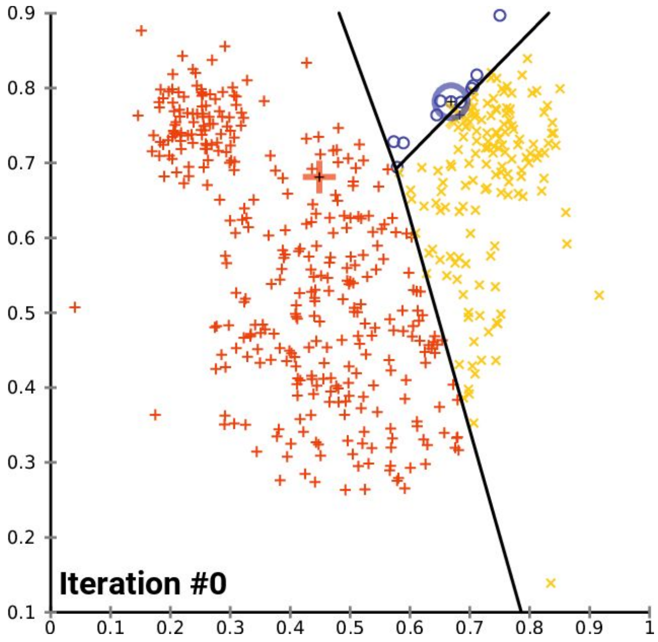
```
["[se]", "[statusofgetrequest]",
"[etimedout]",
"srm://uct2-dc1.uchicago.edu:8443",
"/srm/managerv2", "user",
"timeout", "over"]
```

Word2Vec:



Clustering (online - daily)

- Group message vectors using **KMeans** algorithm



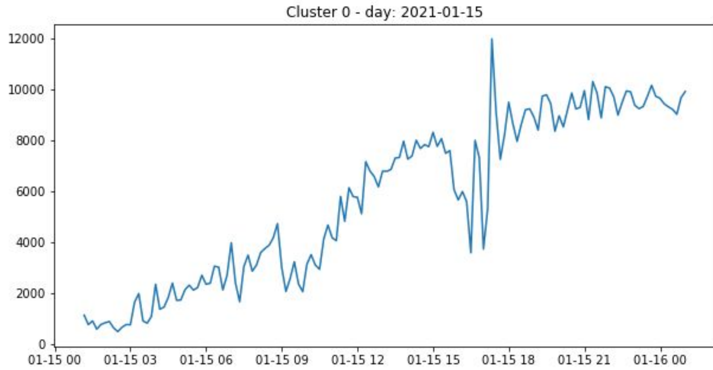
Clustering FTS Errors - Results

- Main results are: *i)* summary table and *ii)* time evolution plot

Highlights:

- Clusters contain similar messages, although allowing some variability
- The model learns to abstract message parameters as IPs, URLs, file paths
- Evolution plots give immediate indications on errors time trends (increasing, cyclical, transient, …)

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



\# of errors for selected cluster
vs
time

Clustering FTS Errors - Validation

- Clustering results compared to GGUS tickets opened within ± 3 days from the period of the analysis (15 January 2021)
- Promising results :
  - **Exact match:** GGUS site *AND* message match cluster summary
  - **Fuzzy match:** evident connection with more than one ticket
  - **Partial match:** GGUS site *OR* message match cluster summary

N. Clusters	Exact Match	Fuzzy Match	Partial Match	False Positives	False Negatives
15	7	3	2	3	1

Highlights:

- Some false positives were real problems, but were not reported in GGUS
- Few false negatives: unusual to have issues completing undetected

Anomaly detection on FTS transfers

- Ongoing collaboration with **Google** and **UCL** to develop a recommendation system to prioritise transfer errors
- Analysis of FTS data showed that we can study errors evolution not only over time but also over the interconnection between nodes (links)

										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

MIDAS (MIcrocluster-based Detector of Anomalies in Streams)

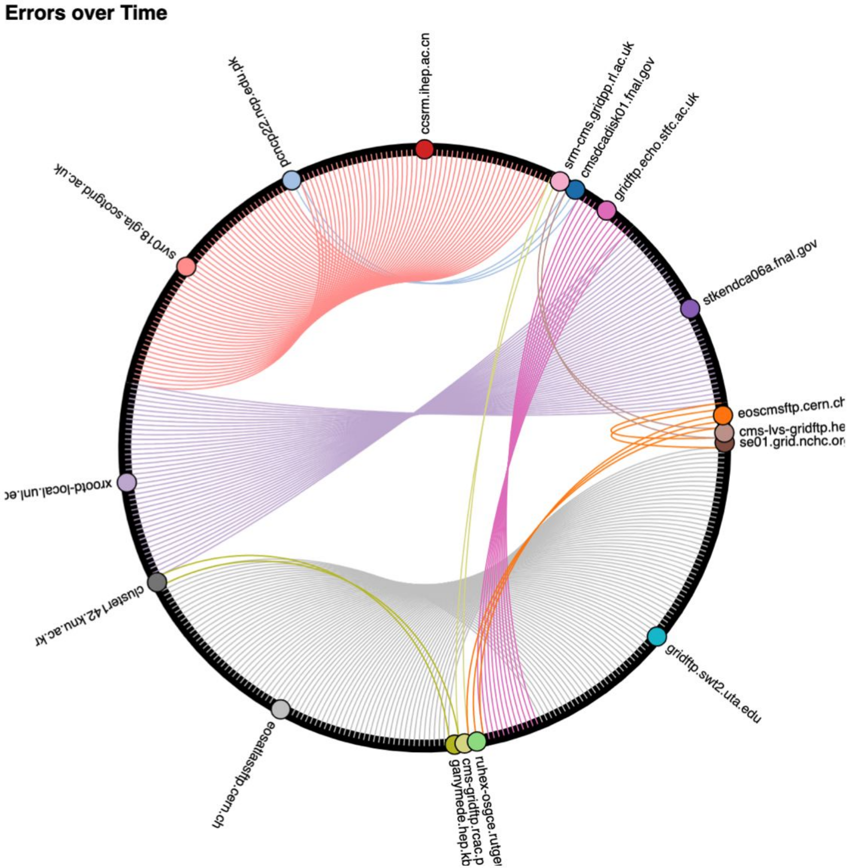
More info: https://arxiv.org/pdf/1911.04464.pdf

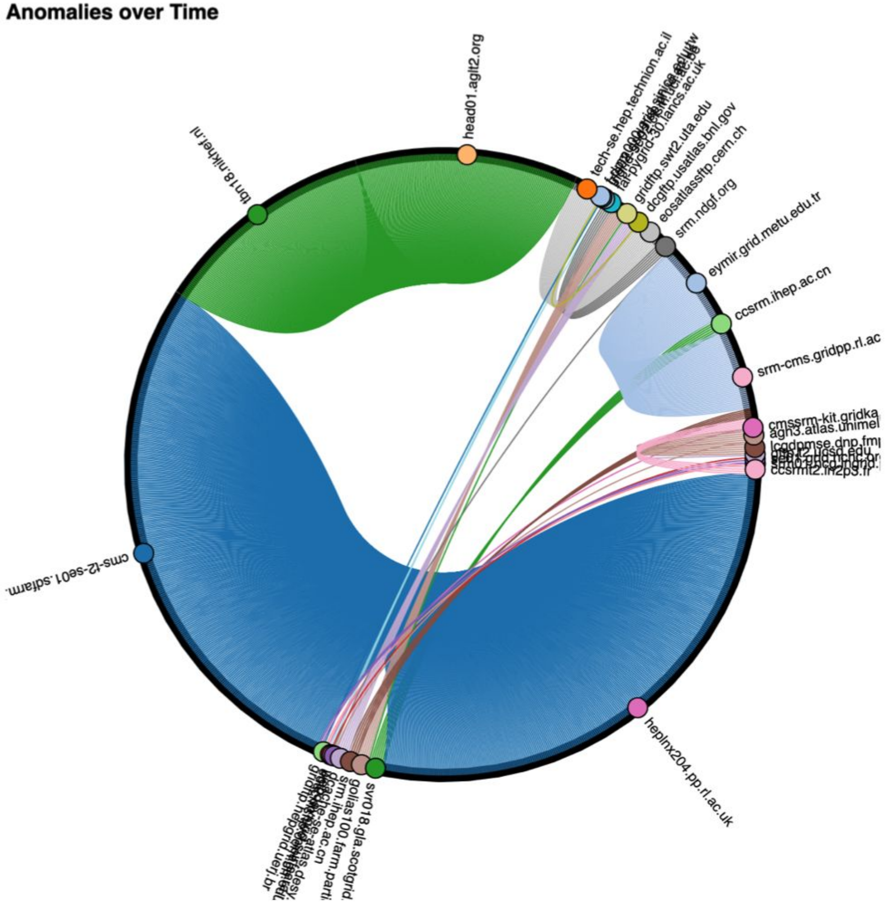
- Finds anomalies in dynamic graphs (file transfers, intrusions)
- Detects micro-clusters (sudden “burst” of connections between nodes (multiple retrials, DOS)
- Memory usage constant and independent of graph size
- Update time in streaming scenarios is also constant

										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

Results


Errors over Time
ccsrm.ihep.ac.cn
srm-cms.gridpp.rl.ac.uk
cmsdcadisk01.fnal.gov
gridftp.echo.stfc.ac.uk
stkendca06a.fnal.gov
eoscmsftp.cern.ch
cms-lvs-gridftp.he
se01.grid.nchc.or
gridftp.swt2.uta.edu
ruhex-osgce.rutger
cms-gridftp.rcac.p
ganymede.hep.kb
eosatlassftp.cern.ch
cluster142.knu.ac.kr
xrootd-local.unl.ec
svr018.gla.scotgrid.ac.uk
pcncp22.ncp.edu.pk


Anomalies over Time
head01.aglt2.org
tbn18.nikhef.nl
tech-se.hep.technion.ac.il
gridftp.swt2.uta.edu
dcgftp.usatlas.bnl.gov
eosatlassftp.cern.ch
srm.ndgf.org
eymir.grid.metu.edu.tr
ccsrm.ihep.ac.cn
srm-cms.gridpp.rl.ac
cmssrm-kit.gridka
ccsrmt2.in2p3.fr
heplnx204.pp.rl.ac.uk
svr018.gla.scotgrid
golias100.farm.parti
srm.ihep.ac.cn
cms-t2-se01.sdfarm.

Anomaly detection on transfers

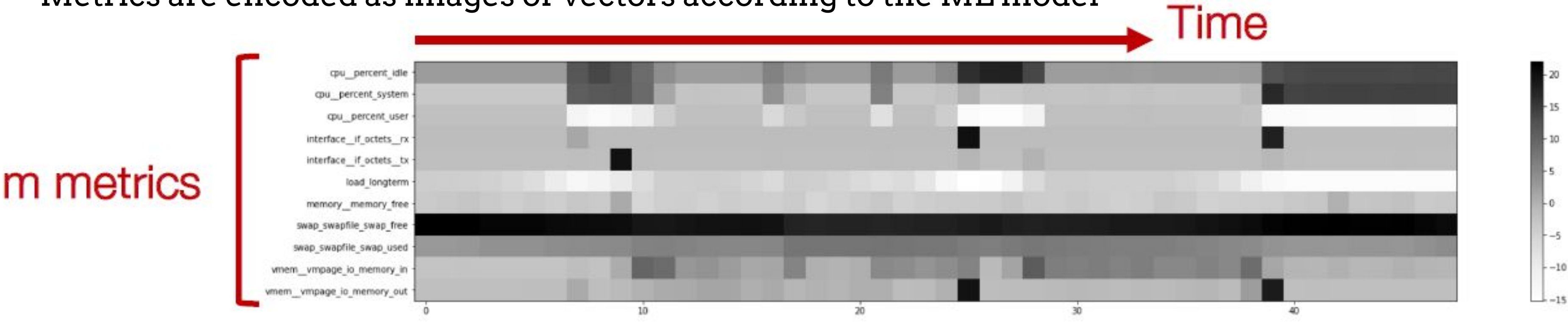
- Next steps:
  - Include text features in anomaly detection. We must consider not only the number, timing and location of links between nodes, but also the messages. Other metadata such as user, file size etc... may play a role too
  - Include data from GGUS tickets to validate the results
  - Build 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

Cloud anomaly detection

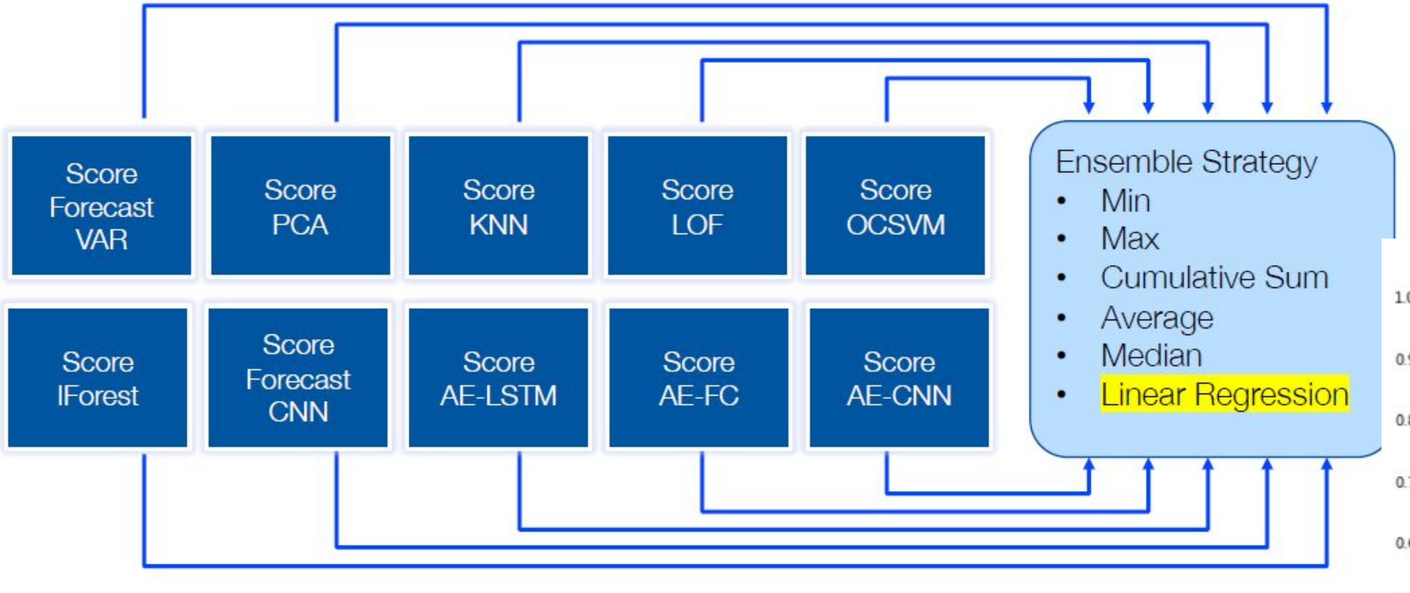
More info: http://cern.ch/go/grf9

- A CERN-IT project to detect anomalies in the CERN Cloud:
  - Identify operational issues
  - Get a comprehensive understanding of the cloud performance
- A grafana annotations enhancement has been developed in parallel to:
  - Allow experts easily give feedback on the results, directly from Grafana

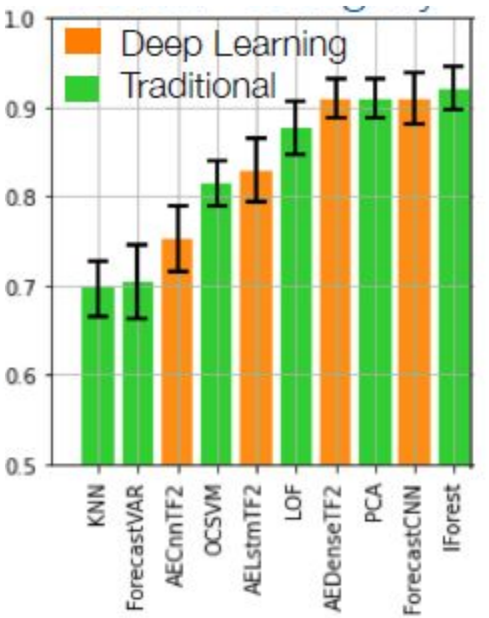
Metrics are encoded as images or vectors according to the ML model

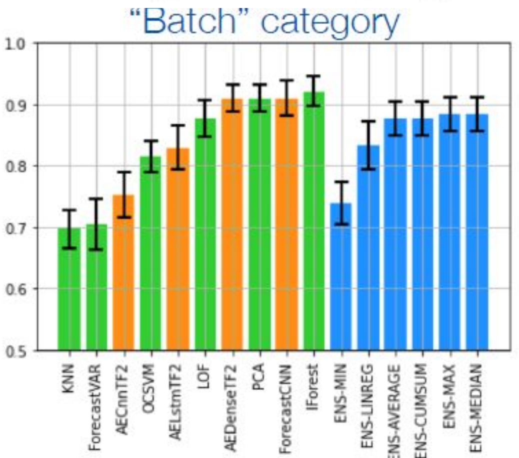


Cloud anomaly detection



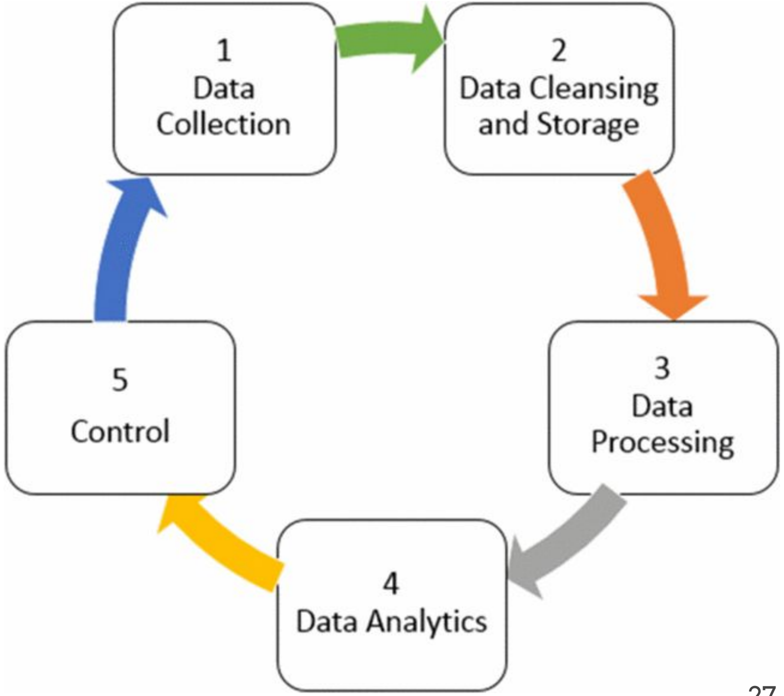
Anomaly Scores Single Methods





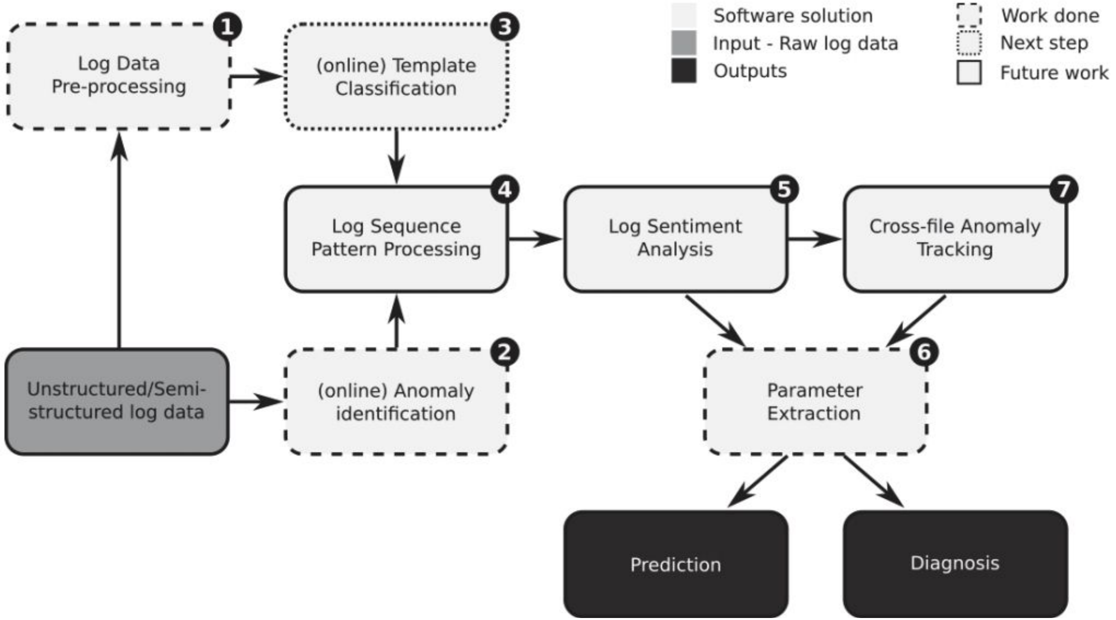
Facilities: Industry examples

- 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

- From reactive maintenance to predictive maintenance
- Parsing the logs of computing services



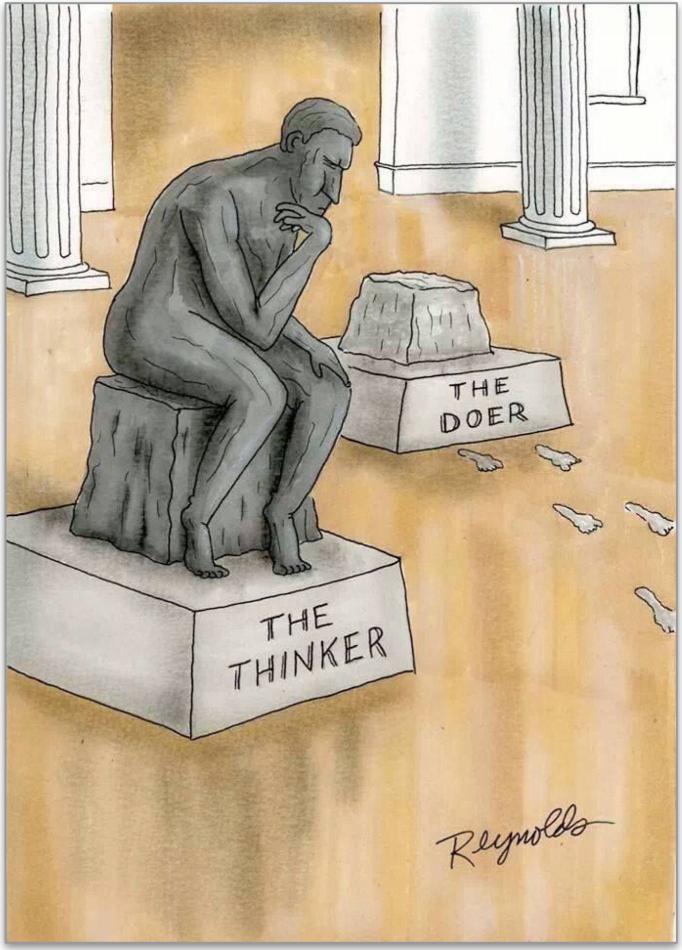
More info: http://cern.ch/go/B9KR

Project Status

Service	Status
Intelligent Alert System	In production
Shared k8s cluster	Work in progress
Jobs Buster	In production
HammerCloud JobShaping	In production
FTS log analysis (Levenshtein distance)	In production
FTS log analysis with ML	In testing phase
FTS Anomaly Detection with Google	Work in progress
Cloud anomaly detection	In production
INFN Bologna - Predictive Maintenance	Work in progress

Outlook

- We have in the past 2 years gathered expertise and an understanding of the various efforts
- Lot of opportunities to span new collaborations and work on exciting cross-experiment projects
- Challenges:
    - Anomaly detection/NLP
    - Lack of annotated datasets
    - Deploy to production

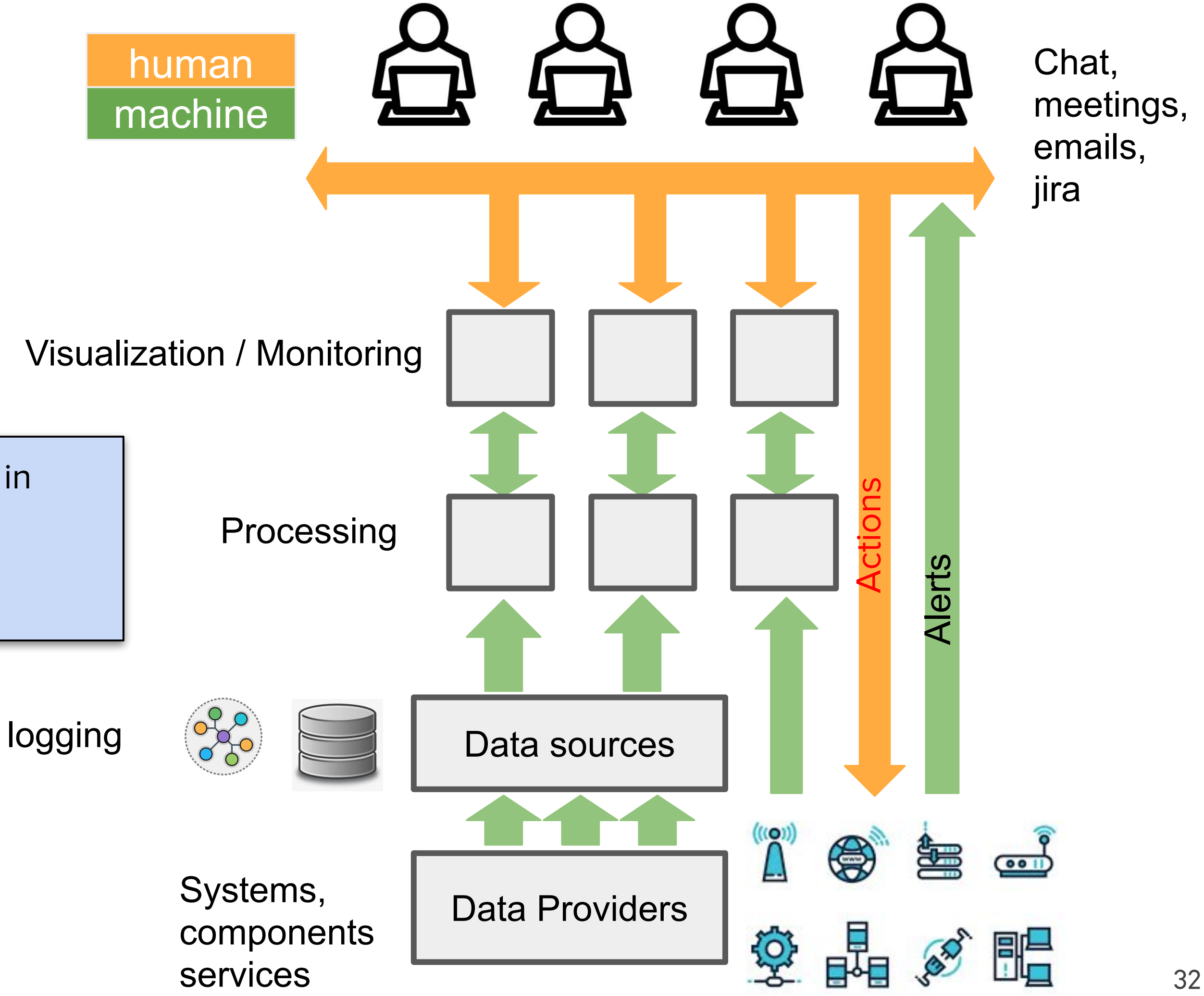


operational-intelligence@cern.ch
https://operational-intelligence.web.cern.ch/

Back up

Operations Today

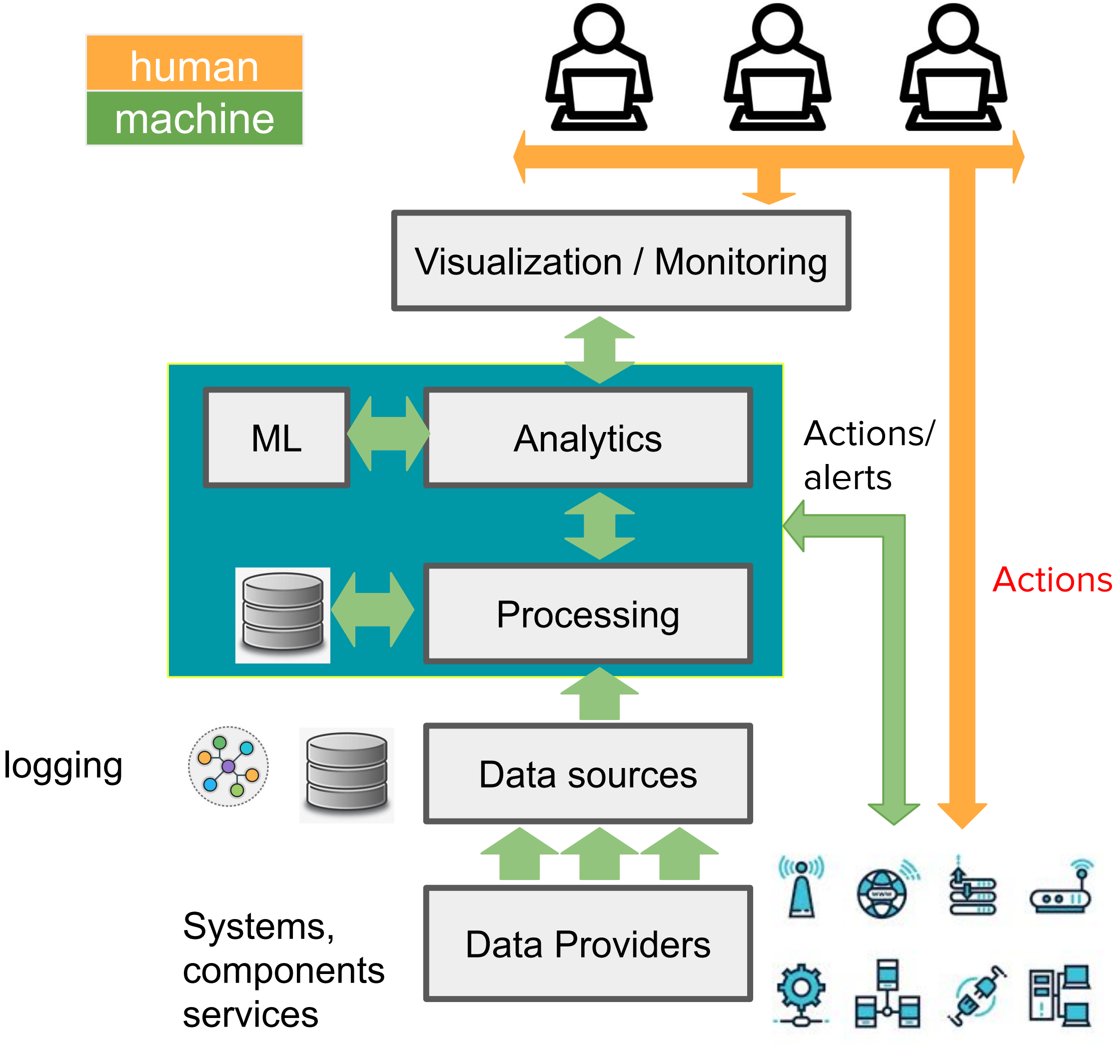
ATLAS/CMS: A lot of people involved in Computing Operations
In 1 year:
> 1k tickets for ATLAS, > 2k for CMS



Operations Tomorrow

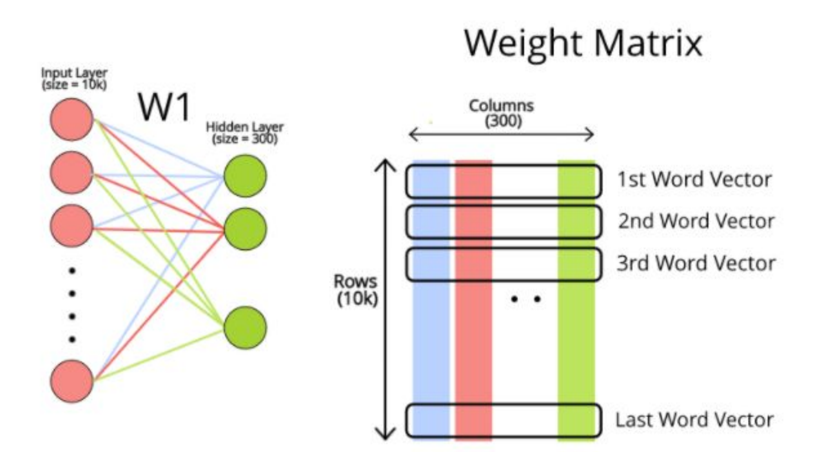
Frontend: aggregated views, suggestions, collects feedback

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

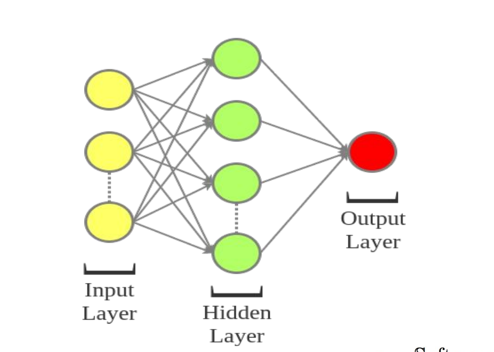


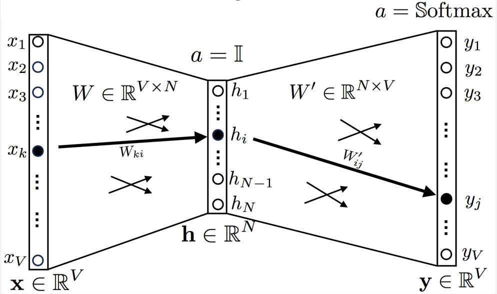
Word2Vec

- **Mapping of words into vectors**
- **Shallow Neural Network (NN)**: stack of hundred processing units (*neurons*) interconnected by nodes . Processing units are made up of input and output units. **Weight to be learnt for each interconnection**



(Ref. backup for Hyper-parameters values)





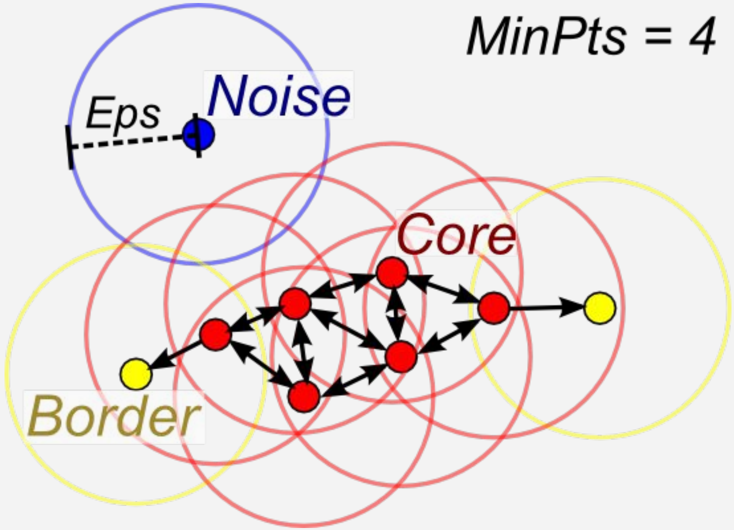
DBSCAN

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

Two parameters, **min_samples and eps**, which define formally the concept of *density* :

- **eps=** maximum distance between two samples for one to be considered as in the neighborhood of the other.

- **min_samples=** number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself.



Higher min_samples or lower eps indicate higher density necessary to form a cluster.