Operational Intelligence

Optimizing computing operations

*Daniele Bonacorsi, Alessandro De Salvo, Alessandro Di Girolamo , **Federica Legger**, Lorenzo Rinaldi*

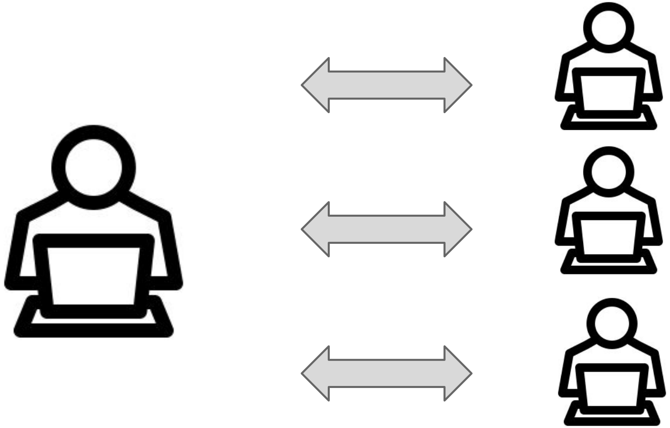
Istituto Nazionale di Fisica Nucleare

WLCG

- 200 sites
- >40 countries
- 750000 cores
- 2 million jobs/day
- 600 PB storage
- 10-100 GB links

Operations in a nutshell

Dashboards, logs, elogs, tickets



Email, JIRA, GGUS, chat, meeting

- Process information
- Triage
- Take decisions
- Escalate/fix
- Record actions

ATLAS/CMS report 100 people involved in computing operations (**50 FTEs/experiment**)!
In 1 year, > 1k GGUS tickets for ATLAS, > 2k for CMS

Can we do better?

- The LHC experiments built a computing system that **worked** in 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; ..)
  - For long, we monitored to debug in near-time, not to analyse and learn from the past to design and build what’s next.
- 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, ..

SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

Full Automation

0	1	2	3	4	5
No Automation	**Driver Assistance**	**Partial Automation**	**Conditional Automation**	**High Automation**	**Full Automation**
Zero autonomy; the driver performs all driving tasks.	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

LHC Run 1 Run 2 Run 3?

Operational intelligence

- A cross-experiment effort aiming to:
  - Streamline computing operations
  - **save manpower & improve resources utilization**
    - Increase level of automation in operation tasks
    - Cost reduction metrics: needed number of operators

- By:
  - Identifying common projects
  - leveraging **common** tools/infrastructure
  - Collaborate, share expertise, tools & approaches
    - Across experiments
    - Across teams (operations, monitoring, analytics)
  - **Bottom-up** approach

Why?

- Computing systems mature and well-understood
- Clear request from funding agencies: push on commonalities
- Easier to interest students/engineers (with background different from HEP) to work on topics using industry standard tools
- ATLAS/CMS use common analytics infrastructure@CERN
- More experiments starting (or considering) using LHC-developed tools
  - for example Rucio, and FTS
  - Share efforts with wider (than LHC) community

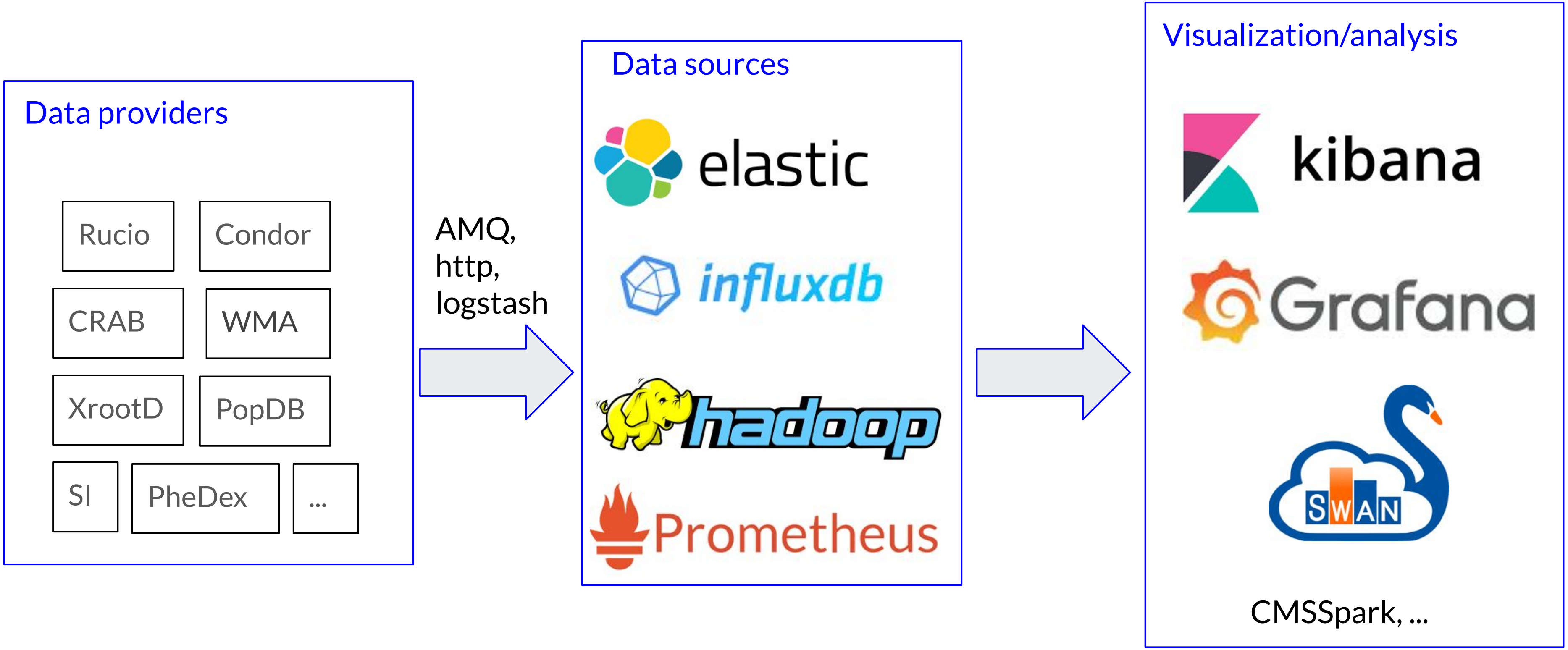
How?

- We successfully started OpInt activities
  - **Kickoff meeting** at HOW19
  - Regular biweekly meetings
  - For now **CMS, ATLAS, HammerCloud, Rucio, MONIT, DUNE/FNAL, LHCb**

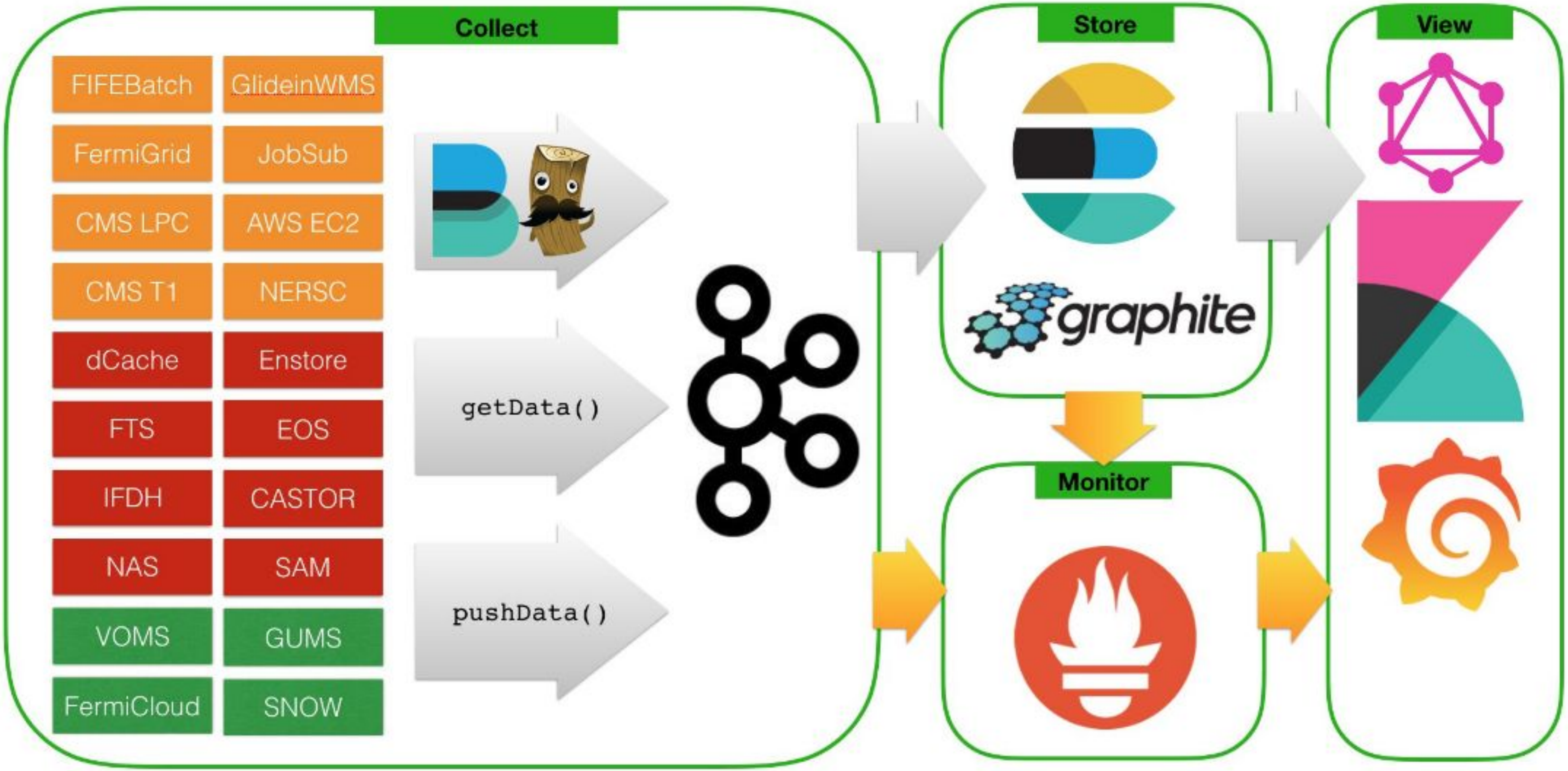
- **Start "simple"**
  - targeting well-identified projects with precise goals/metrics
    - Must show operational cost reduction
    - Must have **operation people** on board

  - Guinea pig: **Rucio**
    - But not limited to (i.e. Data Management all round)
    - Common analytics projects: understand issues with transfers, predict latencies, popularity prediction, ...

The analytics infrastructure @CERN



Landscape @FNAL



CMS Operator Console

Hamed Bakhshiansohi et al.

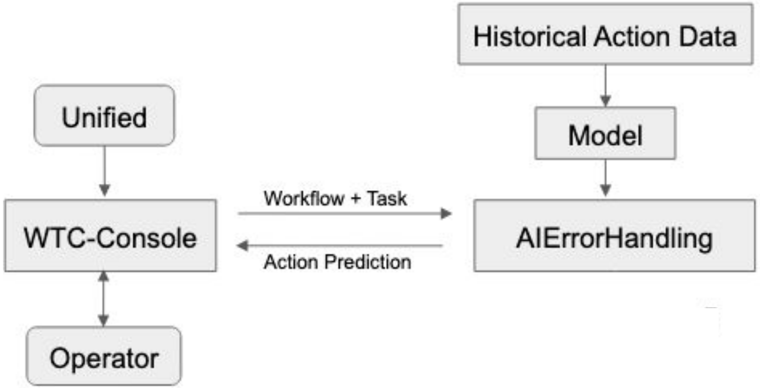
DNNs to predict the action of the operator

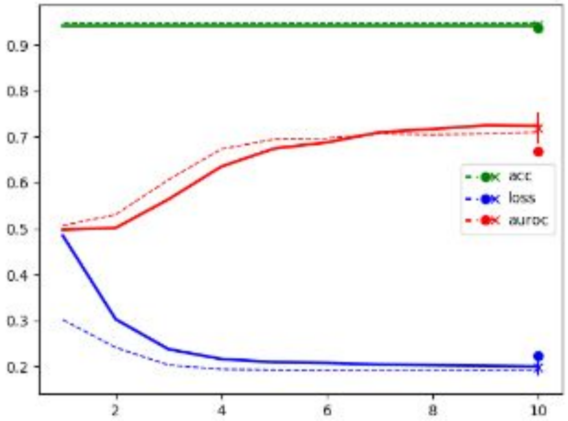
- Multi class and Binary (Retry/Non-Retry) classifications
- Imbalanced class distribution
  - Rate of Retry is dominant
  - Effect of resampling method studied
- **AUROC ~ 70% is achieved**

Possible Actions

- Retry (only failed)
- Kill and Clone
  - With new splitting
  - New settings for memory and cores
- Recovery

	T0_CH_CERN	T1_DE_KIT	T1_ES_PIC	T1_FR_CCIN2P3	T1_IT_CNAF	T1_RU_JINR	T1_UK_RAL	T1_US_FNAL	T2_CH_CERN	T2_CH_CERNBOX	T2_CH_CERN_HLT	T2_DE_DESY	T2_ES_IFCA	T2_FR_GRIF_IRFU	T2_FR_GRIF_LLR	T2_IT_Legnaro	T2_UK_London_Brunel	T2_UK_London_IC	T2_UK_SGrid_RALPP	T2_US_Florida	T2_US_MIT
-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
84	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
85	0	2	0	0	0	1	6	0	0	0	8	1	0	0	0	0	0	0	0	0	0
92	0	0	1	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	2	0	0
134	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
139	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8001	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
8004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
50110	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
50660	0	0	3	2	4	1	1	1	6	0	63	1	0	1	0	1	0	1	0	0	1
50664	0	0	2	7	0	0	2	18	0	0	32	0	0	7	0	0	4	0	2	0	0
71304	0	0	0	0	0	0	0	0	0	0	102	0	0	0	0	0	0	0	0	0	0
99305	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

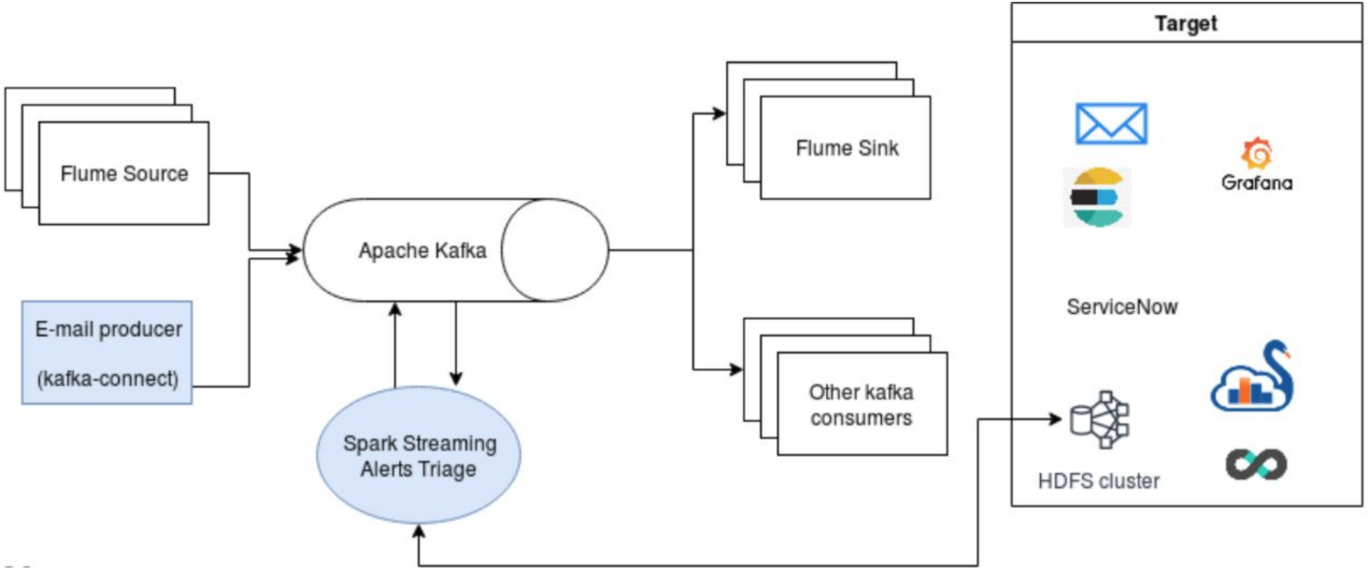




Alert Triage

- First project completely experiment agnostic
- Several data streams from multiple subsystems going through MONIT with a common messaging service: **Kafka**

Complex patterns	E.g.: • Timeout message has been repeated 5 times in a 1 hour window for a given node. • Memory increse alert but there is not a request increase alert.



Data Transfer Alert

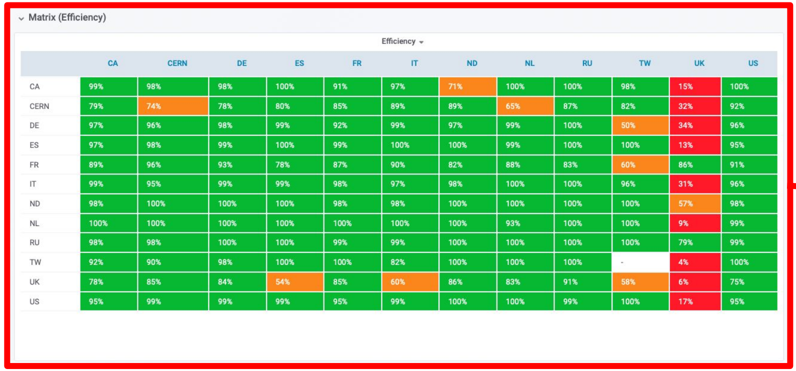
Goal: design a predictive intelligent algorithm to send alarms in case of steady state violations (i.e. degraded transfer performances)

- Harmonize and analyze the data transfer metrics
  - Many already collected in Analytics Platforms (ES, …)
  - Extract new metrics from log files (Rucio, FTS)
- Look for correlations among metrics, using a reinforced learning approach
  - Start simple and then go deep
  - Use correlations to send alarms if potential anomaly are detected
    - shifter/expert will validate

Data Transfer Alert

What we have **now**

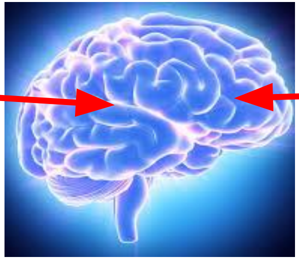
- Very reliable dashboards and metric collectors for Data Transfers
- Few systems send notifications (based on “hard-coded” rules)



Matrix (Efficiency)

Efficiency

	CA	CERN	DE	ES	FR	IT	ND	NL	RU	TW	UK	US
CA	99%	98%	98%	100%	91%	97%	71%	100%	100%	98%	15%	100%
CERN	79%	74%	78%	80%	85%	89%	89%	65%	87%	82%	32%	92%
DE	97%	96%	98%	99%	92%	99%	97%	99%	100%	50%	34%	96%
ES	97%	98%	99%	100%	99%	100%	100%	99%	100%	100%	13%	95%
FR	89%	96%	93%	78%	87%	90%	82%	88%	83%	60%	86%	91%
IT	99%	95%	99%	99%	98%	97%	98%	100%	100%	96%	31%	96%
ND	98%	100%	100%	100%	98%	98%	100%	100%	100%	100%	57%	98%
NL	100%	100%	100%	100%	100%	100%	100%	93%	100%	100%	9%	99%
RU	98%	98%	100%	100%	99%	99%	100%	100%	100%	100%	79%	99%
TW	92%	90%	98%	100%	100%	82%	100%	100%	100%	-	4%	100%
UK	78%	85%	84%	54%	85%	60%	86%	83%	91%	58%	6%	75%
US	95%	99%	99%	99%	95%	99%	100%	100%	99%	100%	17%	95%

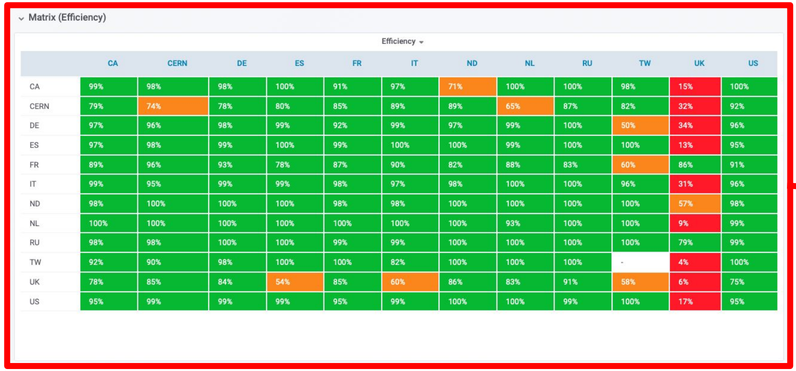


ATLAS Site	DDMEndpoint	Status																
		Delete	DELETE LAN	DELETE WAN	Fetch	Pilot Log Write	Pilot Read	Pilot Write	Read	READ LAN	READ WAN	Subscription	default SRM	Third party copy	Upload	Write	WRITE LAN	WRITE WAN
CERN-PROD	CERN-PROD_DERIVED														manual	manual		
	CERN-PROD_PERF-EGAMMA														QUOTASPACE	QUOTASPACE		
	CERN-PROD_PERF-JETS														QUOTASPACE	QUOTASPACE		
	CERN-PROD_PPSTAPE	manual			manual				manual			manual			manual	manual		
	CERN-PROD_RAW														manual	manual		
	CERN-PROD_TRIG-DAQ														QUOTASPACE	QUOTASPACE		
	CERN-PROD_TZDISK														manual	manual		
CSCS-LCG2	CSCS-LCG2_LOCALGROUPDISK														DISKSPACE	DISKSPACE		
CYFRONET-LCG2	CYFRONET-LCG2_PHYS-HI														QUOTASPACE	QUOTASPACE		
FZK-LCG2	FZK-LCG2_PERF-IDTRACKING														QUOTASPACE	QUOTASPACE		
IN2P3-CC	IN2P3-CC_PERF-MUONS														QUOTASPACE	QUOTASPACE		
INFN-COSENZA	INFN-COSENZA_DATADISK	AGIS			AGIS				AGIS						AGIS	AGIS		
	INFN-COSENZA_LOCALGROUPDISK	AGIS			AGIS				AGIS						AGIS	AGIS		
	INFN-COSENZA_SCRATCHDISK	AGIS			AGIS				AGIS						AGIS	AGIS		
NDGF-T1	NDGF-T1_GROUPTAPE_PHYS-TOP														QUOTASPACE	QUOTASPACE		
NIKHEF-ELPROD	NIKHEF-ELPROD_LOCALGROUPDISK														DISKSPACE	DISKSPACE		
RO-02-NIPNE	RO-02-NIPNE_LOCALGROUPDISK	AGIS			AGIS				AGIS						AGIS	AGIS		
SE-SNIC-T2	SE-SNIC-T2_LUND_LOCALGROUPDISK														DISKSPACE	DISKSPACE		
TOKYO-LCG2	TOKYO-LCG2_PERF-MUONS														QUOTASPACE	QUOTASPACE		
TUDresden-ZIH	TUDRESDEN-ZIH_LOCALGROUPDISK	manual			manual				manual			manual			manual +	manual +		
WT2	SLACXRD_DATADISK	manual			manual				manual			manual			manual	manual		
	SLACXRD_SCRATCHDISK	manual			manual				manual			manual			manual	manual		
praguelcg2	PRAGUELCG2_LOCALGROUPTAPE	manual			manual				manual			manual			manual	manual		

- NOTIFICATION not (necessary) a PROBLEM
- Many clicks to understand the problem (and take actions)
- Correlations spotted out by Human Intelligence

Data Transfer Alert

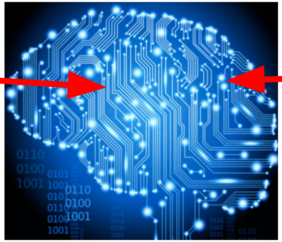
An Operational Intelligence system will spot correlations among the collected metrics and learn the thresholds above which send a notification



Matrix (Efficiency)

Efficiency

	CA	CERN	DE	ES	FR	IT	ND	NL	RU	TW	UK	US
CA	99%	98%	98%	100%	91%	97%	71%	100%	100%	98%	15%	100%
CERN	79%	74%	78%	80%	85%	89%	89%	65%	87%	82%	32%	92%
DE	97%	96%	98%	99%	92%	99%	97%	99%	100%	50%	34%	96%
ES	97%	98%	99%	100%	99%	100%	100%	99%	100%	100%	13%	95%
FR	89%	96%	93%	78%	87%	90%	82%	88%	83%	60%	86%	91%
IT	99%	95%	99%	99%	98%	97%	98%	100%	100%	96%	31%	96%
ND	98%	100%	100%	100%	98%	98%	100%	100%	100%	100%	57%	98%
NL	100%	100%	100%	100%	100%	100%	100%	93%	100%	100%	9%	99%
RU	98%	98%	100%	100%	99%	99%	100%	100%	100%	100%	79%	99%
TW	92%	90%	98%	100%	100%	82%	100%	100%	100%	-	4%	100%
UK	78%	85%	84%	54%	85%	60%	86%	83%	91%	58%	6%	75%
US	95%	99%	99%	99%	95%	99%	100%	100%	99%	100%	17%	95%



ATLAS Site	DDMEndpoint	Status																
		Delete	DELETE LAN	DELETE WAN	Fetch	Pilot Log Write	Pilot Read	Pilot Write	Read	READ LAN	READ WAN	Subscription	default SRM	Third party copy	Upload	Write	WRITE LAN	WRITE WAN
CERN-PROD	CERN-PROD_DERIVED														manual	manual		
	CERN-PROD_PERF-EGAMMA														QUOTASPACE	QUOTASPACE		
	CERN-PROD_PERF-JETS														QUOTASPACE	QUOTASPACE		
	CERN-PROD_PPSTAPE	manual			manual				manual			manual			manual	manual		
	CERN-PROD_RAW														manual	manual		
	CERN-PROD_TRIG-DAQ														QUOTASPACE	QUOTASPACE		
	CERN-PROD_TZDISK														manual	manual		
CSCS-LCG2	CSCS-LCG2_LOCALGROUPDISK														DISKSPACE	DISKSPACE		
CYFRONET-LCG2	CYFRONET-LCG2_PHYS-HI														QUOTASPACE	QUOTASPACE		
FZK-LCG2	FZK-LCG2_PERF-IDTRACKING														QUOTASPACE	QUOTASPACE		
IN2P3-CC	IN2P3-CC_PERF-MUONS														QUOTASPACE	QUOTASPACE		
INFN-COSENZA	INFN-COSENZA_DATADISK	AGIS			AGIS				AGIS						AGIS	AGIS		
	INFN-COSENZA_LOCALGROUPDISK	AGIS			AGIS				AGIS						AGIS	AGIS		
	INFN-COSENZA_SCRATCHDISK	AGIS			AGIS				AGIS						AGIS	AGIS		
NDGF-T1	NDGF-T1_GROUPTAPE_PHYS-TOP														QUOTASPACE	QUOTASPACE		
NIKHEF-ELPROD	NIKHEF-ELPROD_LOCALGROUPDISK														DISKSPACE	DISKSPACE		
RO-02-NIPNE	RO-02-NIPNE_LOCALGROUPDISK	AGIS			AGIS				AGIS						AGIS	AGIS		
SE-SNIC-T2	SE-SNIC-T2_LUND_LOCALGROUPDISK														DISKSPACE	DISKSPACE		
TOKYO-LCG2	TOKYO-LCG2_PERF-MUONS														QUOTASPACE	QUOTASPACE		
TUDresden-ZIH	TUDRESDEN-ZIH_LOCALGROUPDISK	manual			manual				manual			manual			manual +	manual +		
WT2	SLACXRD_DATADISK	manual			manual				manual			manual			manual	manual		
	SLACXRD_SCRATCHDISK	manual			manual				manual			manual			manual	manual		
praguelcg2	PRAGUELCG2_LOCALGROUPTAPE	manual			manual				manual			manual			manual	manual		

NOTIFICATION would eventually be a real ALARM

(with a relative high degree of confidence)

GATES

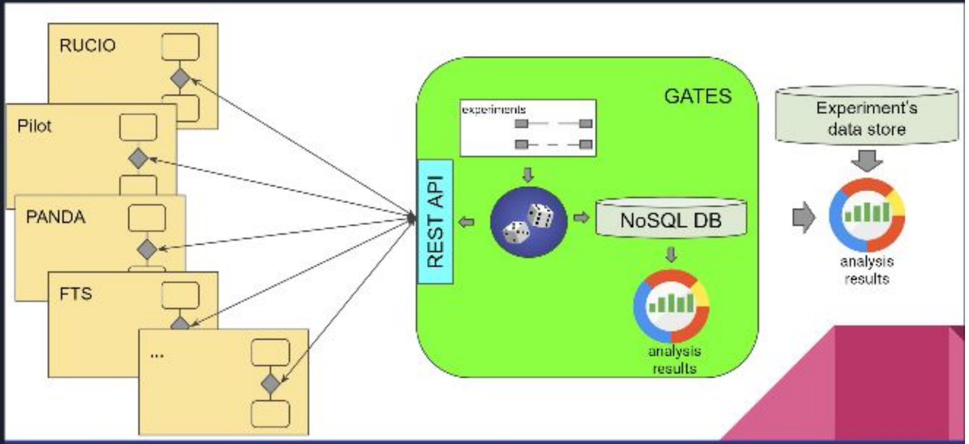
Ilija Vukotic et al.

Experiments have large data stores collecting data from different computing systems: job scheduling, data distribution, FTS, PerfSONAR, etc.

While that is great for monitoring, accounting, and finding issues, it is not sufficient for the system optimization.

One can try to guess what kind of effect a change will made, but without validation it does not mean much.

We need a way to quickly test different options and get actionable answers.

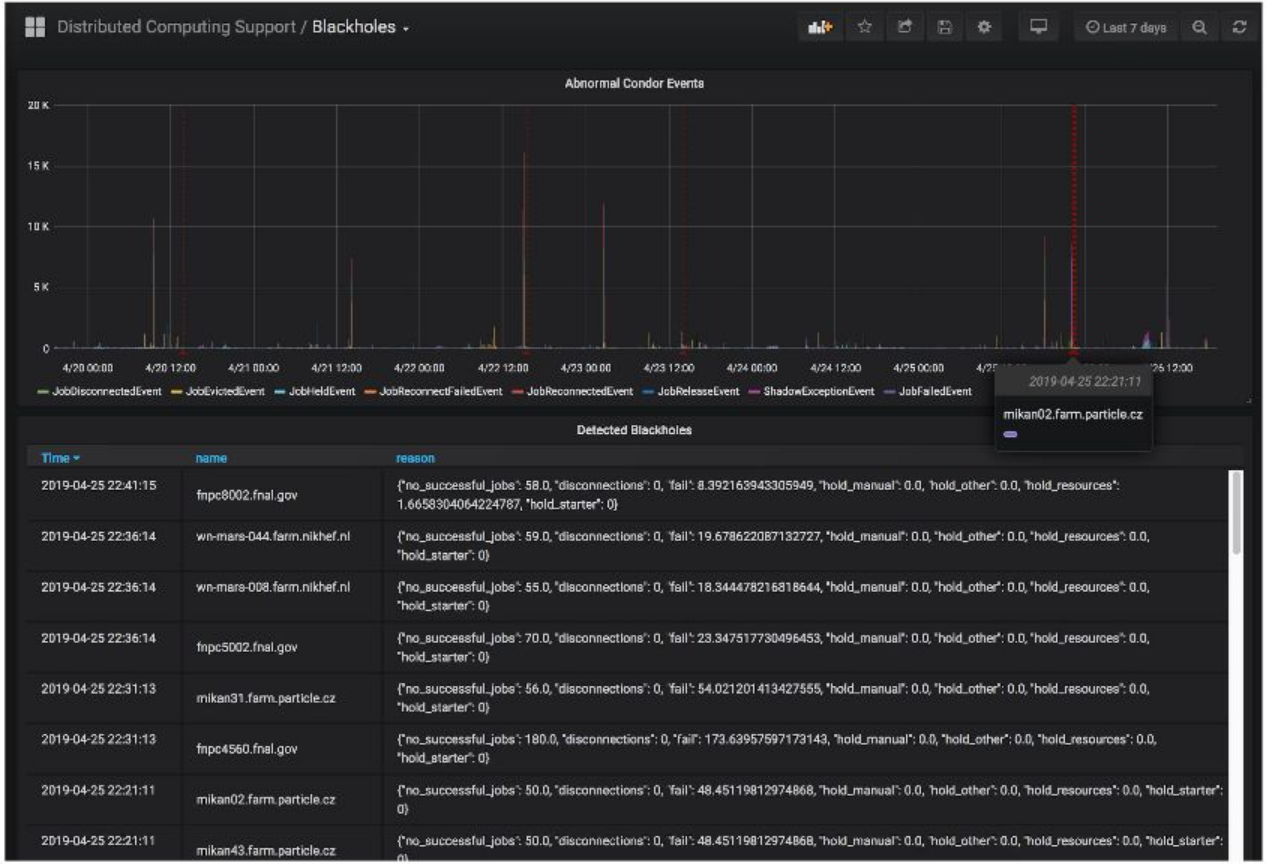


In development.
Completely experiment agnostic.
Contributions welcome.

A/B testing service

Blackhole Node Detection

Kevin Retzke et al.



- Occasionally we encounter “bad” nodes on the grid: hardware issues, missing/broken CVMFS mounts, etc.
- Analyze recent job events to look for telltale signs of bad node
- Challenges:
  - bad node vs. bad user code
  - Transient events (note and move on) vs. persistent issues (blacklist)

Predictive site maintenance

Smooth operations at the (multi-)experiment level rely on reliable underlying infrastructures and services

So far:

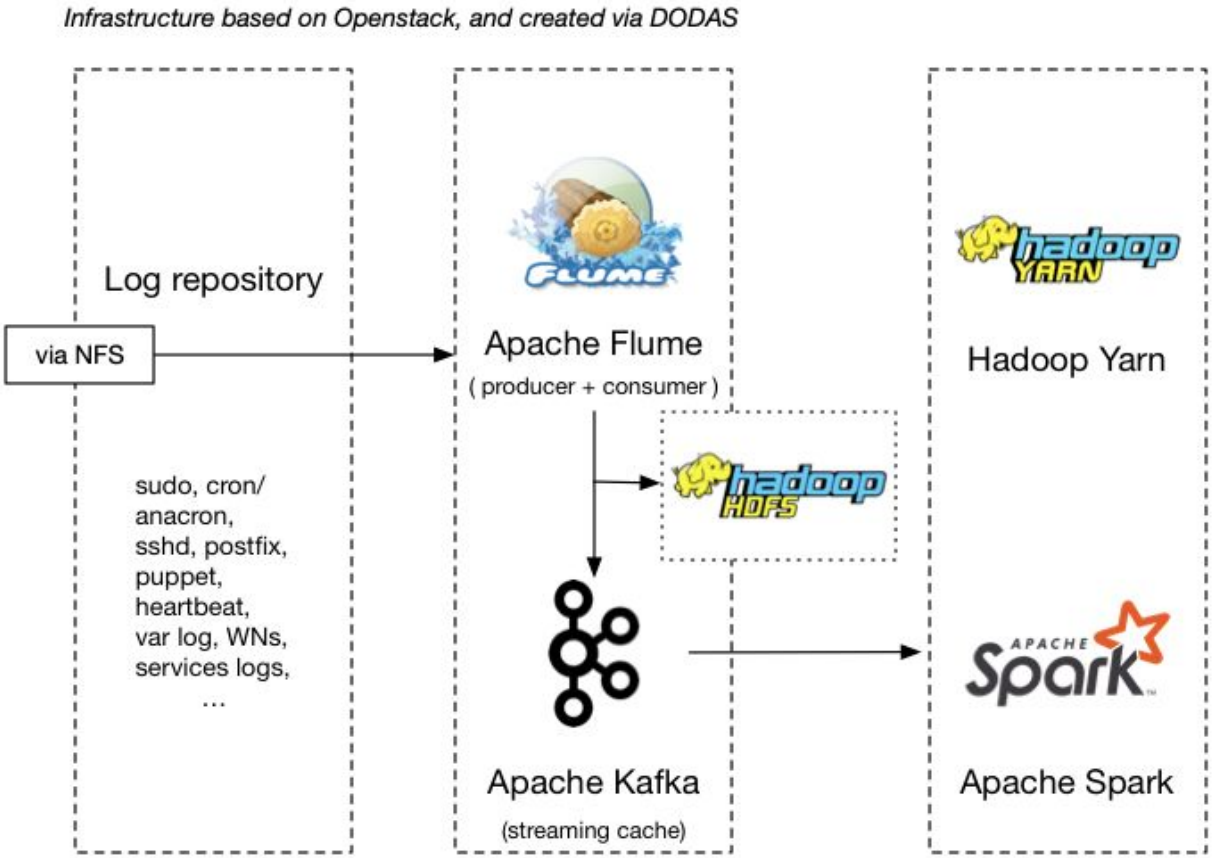
- **reactive-only** approach to problems after they show up
- rely on (whenever possible) **prompt reaction** to attack and solve issues

Work in progress, e.g. at INFN-CNAF

- Ongoing effort to rationalise the *collection of logs* from machines / services asynchronous log analyses (by summer students, service experts, external collaborators)
- first infrastructure work for a *long-term predictive maintenance* approach at INFN-CNAF, potentially exportable to a generic WLCG computing center

Predictive site maintenance

D. Bonacorsi et al.



- L.Giommi et al, "*Towards Predictive Maintenance with Machine Learning at the INFN-CNAF computing centre*", ISGC 2019, Taipei
- T. Diotalevi et al., "*Collection and harmonization of system logs and prototypal Analytics services with the Elastic (ELK) suite at the INFN-CNAF computing centre*", ISGC 2019, Taipei

Site Operation Anomaly Detection

A. De Salvo

The Goal: create a global, distributed anomaly detection system, based on ES/Beats and DL data analysis, to monitor site activities

- Prototyping in the ATLAS Italian Tier-2 infrastructure
  - All Tier-2 sites currently sending **auditbeat** data to a global collector in Roma since april 2019 (average ~ 500 Hz of collected data, 10 GB/day)
    - start/stop processes
    - Open sockets
    - User events such as nis calls
  - Central (expandable) infrastructure currently running on an ES cluster with 7 data nodes + dual queue buffer and a totale space ~30TB, sufficient to keep several months of data online
  - Shared infrastructure, currently also collecting the ATLAS global node description data

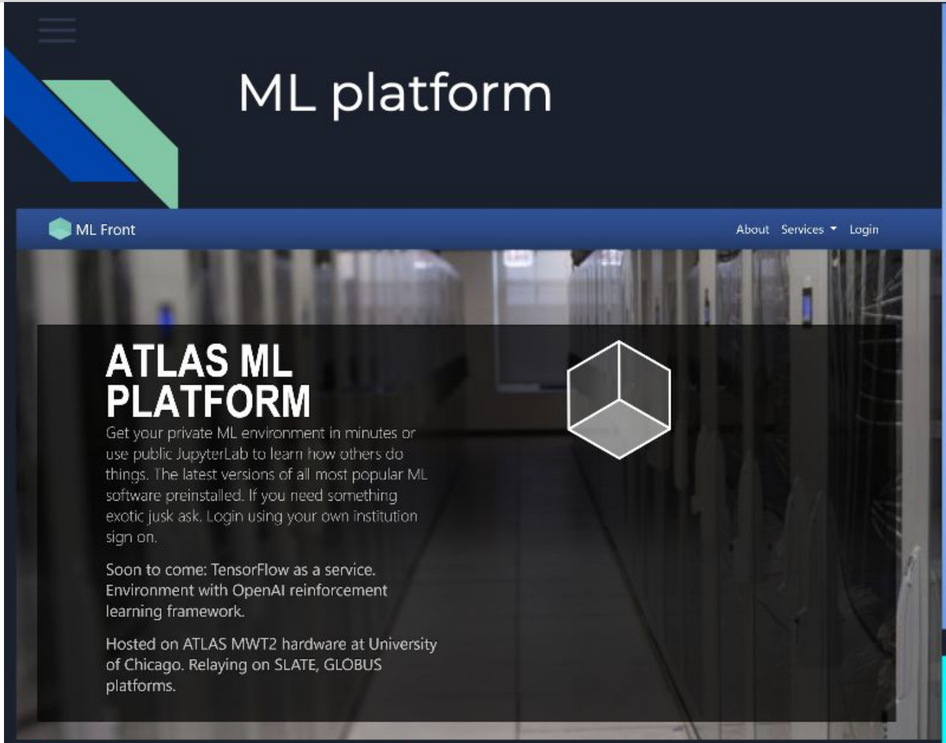
Site Operation Anomaly Detection

A. De Salvo

- Data analysis and goals
  - Anomaly detection based on autoencoders being prepared, to detect problems or intrusions in the sites
  - Generally useful for many different purposes
    - Transparent distributed firewall, intrusion detections, hardware failures, process misbehaviour or malicious attacks (useful also during security challenges), etc
  - DL training and analysis can also be performed via a dedicated GPGPU nVidia facility being deployed in Roma

ATLAS ML platform

Ilija Vukotic et al.



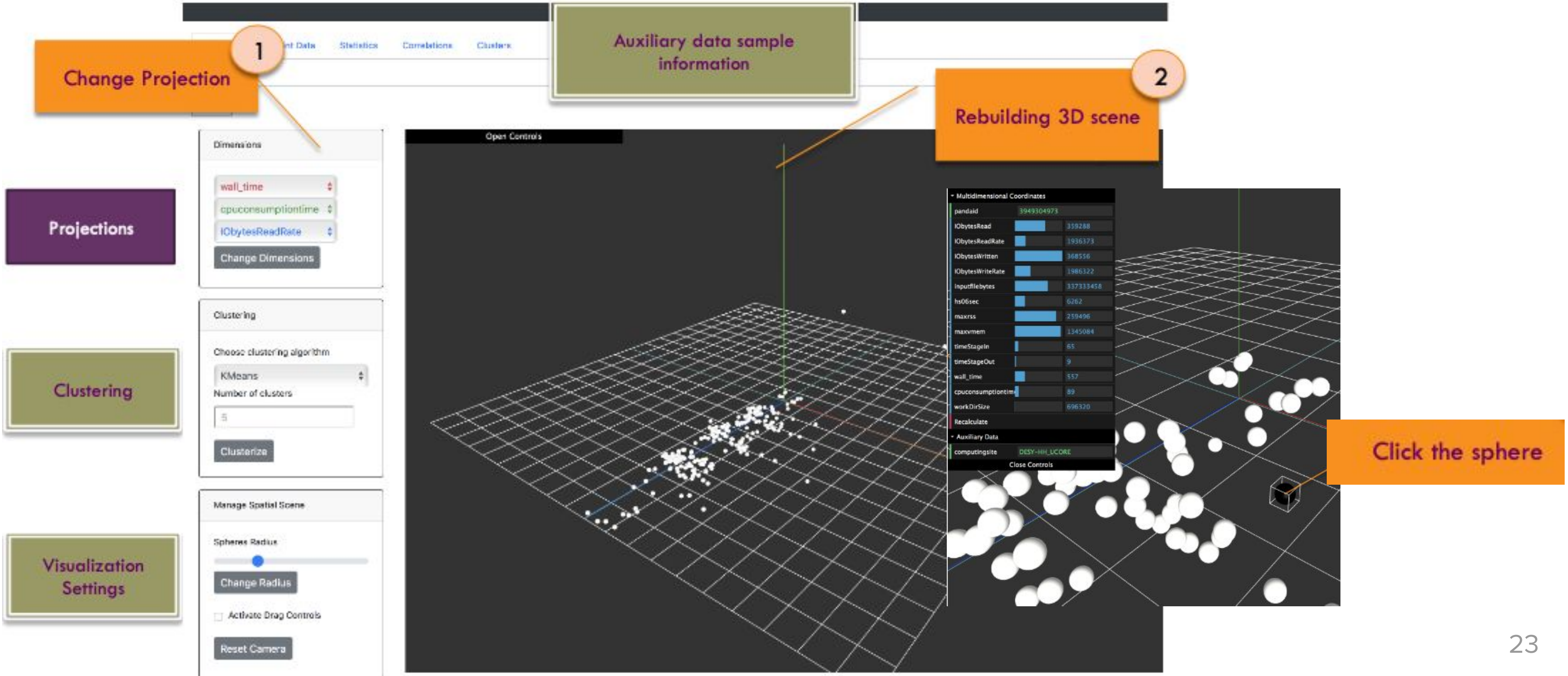
https://www.atlas-ml.org/

- Nodejs, expressjs site running on k8s. Has full control of k8s deployments.
- Uses Globus authentication.
- Can limit resources per instance.
- Extensible
  - Currently offers: private and shared JupyterLab instances and Spark Job submissions.
  - To come: TFAAS
- Easy to deploy elsewhere.
  - ML workshops
  - Teaching computational physics courses

Could improve:
- shared storage
- federate backend resources
- more options - CVMFS

Maria Grigoryeva et al.

Prototype of InVEx for ATLAS Computing http://vap-dev.tpu.ru/



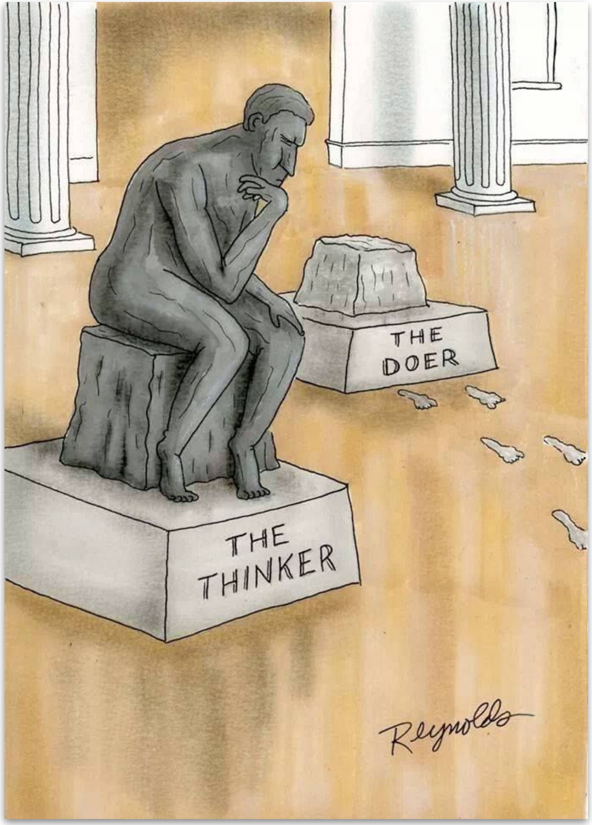
Challenges

- Anomaly detection in time series
  - Data quality, Network issues, Site performance
  - Despite importance, not many off-the-shelf tools
  - used or tried: simple hwm/lwm limits, ARIMA, plato detection, Bayesian simultaneous change point detection, ANN/BDT in time bins
  - Lack of well annotated data
- Classification
  - Data popularity prediction
  - Error classification:
    - Jobs - almost free style text - NLP?
    - FTS, Rucio, Frontier
- Need experiment-agnostic event annotation tool
  - Currently we only have tickets as a history of things that happened.
  - Not classified in any way that can be used to train any model.

Quick recap

- Operational Intelligence website
- Github repository: https://github.com/operationalintelligence
- E-group for communication: operational-intelligence@cern.ch
- JLAB session: https://indico.cern.ch/event/759388/sessions/295063/#20190321
- Google doc draft
- Regular meetings (2x/month): **3-4 pm on Mondays**
  - Indico category: https://indico.cern.ch/category/11205/

Today



Tomorrow