Monitoring Infrastructure for the CERN Data Centre

Asier Aguado

17/05/2018

Monitoring Mission

- Provide Monitoring as a Service for CERN Data Centre and the WLCG collaboration
  - e.g. Dashboards, Alarms, Search, Archive
- Collect, transport, store and process metrics and logs for applications and infrastructure

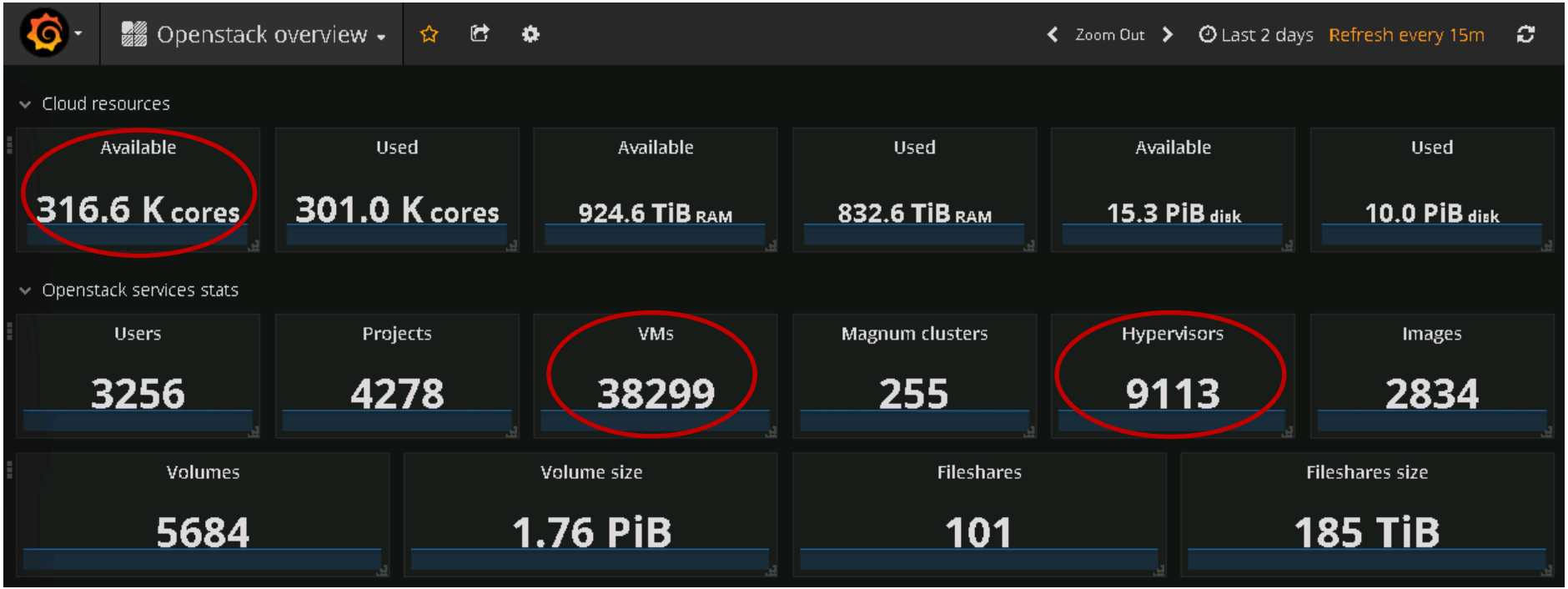
History

Two use cases at CERN had their own monitoring solution

- LEMON and SLS for Data Centre and IT services
- Monitoring tools for WLCG jobs and transfers

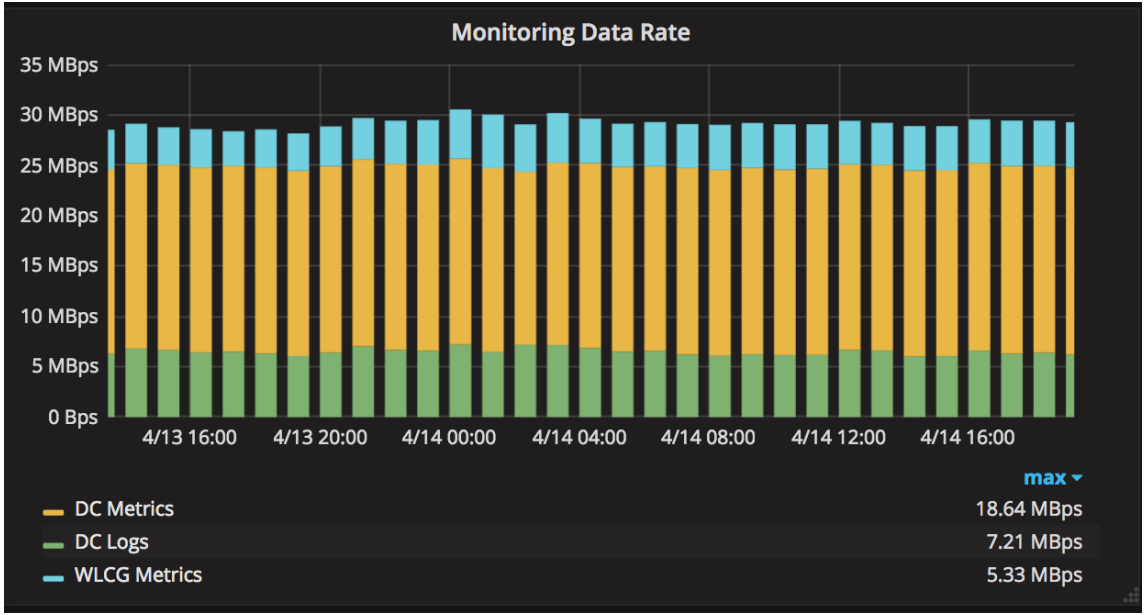
- Mission: integrating all of them into a single monitoring solution (MONIT)

Data Centre


Openstack overview
Zoom Out
Last 2 days
Refresh every 15m
Cloud resources
Available
316.6 K cores
Used
301.0 K cores
Available
924.6 TiB RAM
Used
832.6 TiB RAM
Available
15.3 PiB disk
Used
10.0 PiB disk
Openstack services stats
Users
3256
Projects
4278
VMs
38299
Magnum clusters
255
Hypervisors
9113
Images
2834
Volumes
5684
Volume size
1.76 PiB
Fileshares
101
Fileshares size
185 TiB

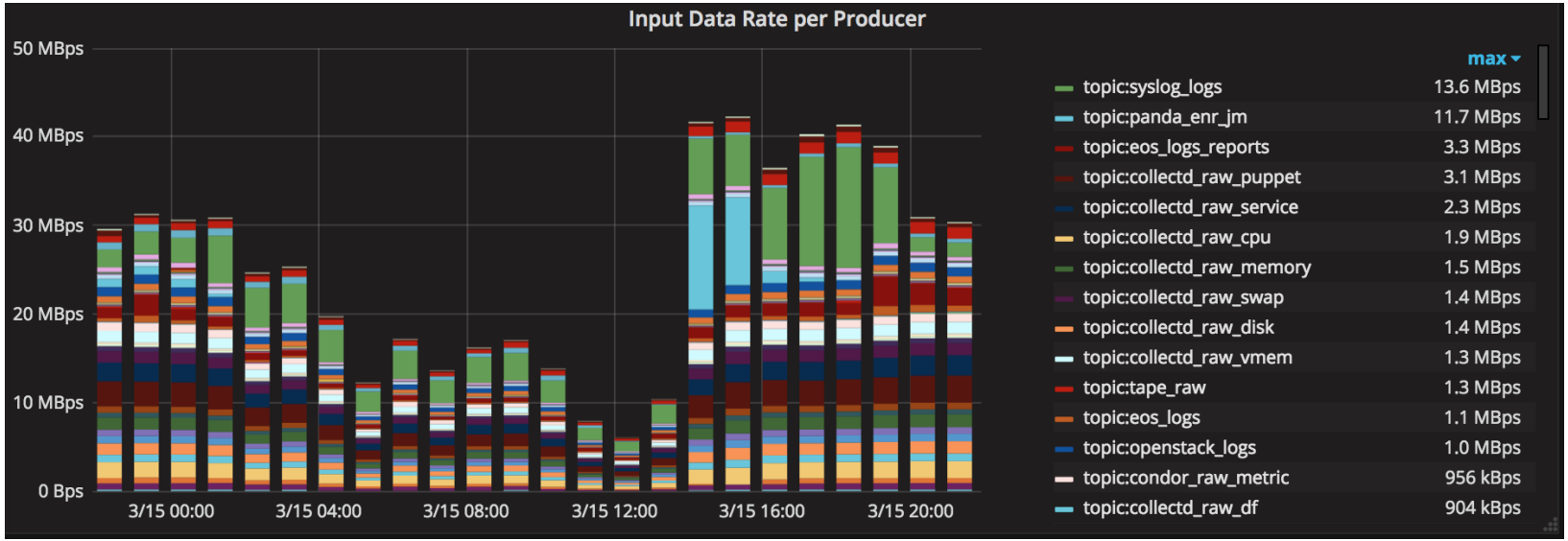
Monitoring Data

- DC Metrics include:
  - OS metrics
  - Hardware metrics
  - Application metrics
- from > 35k machines
- ~ 3 TB/day (compressed)
- > 80kHz rate



Monitoring Data - Workload

- Spikes in rate and volume
- Temporary outages (e.g. network outages) cause spikes

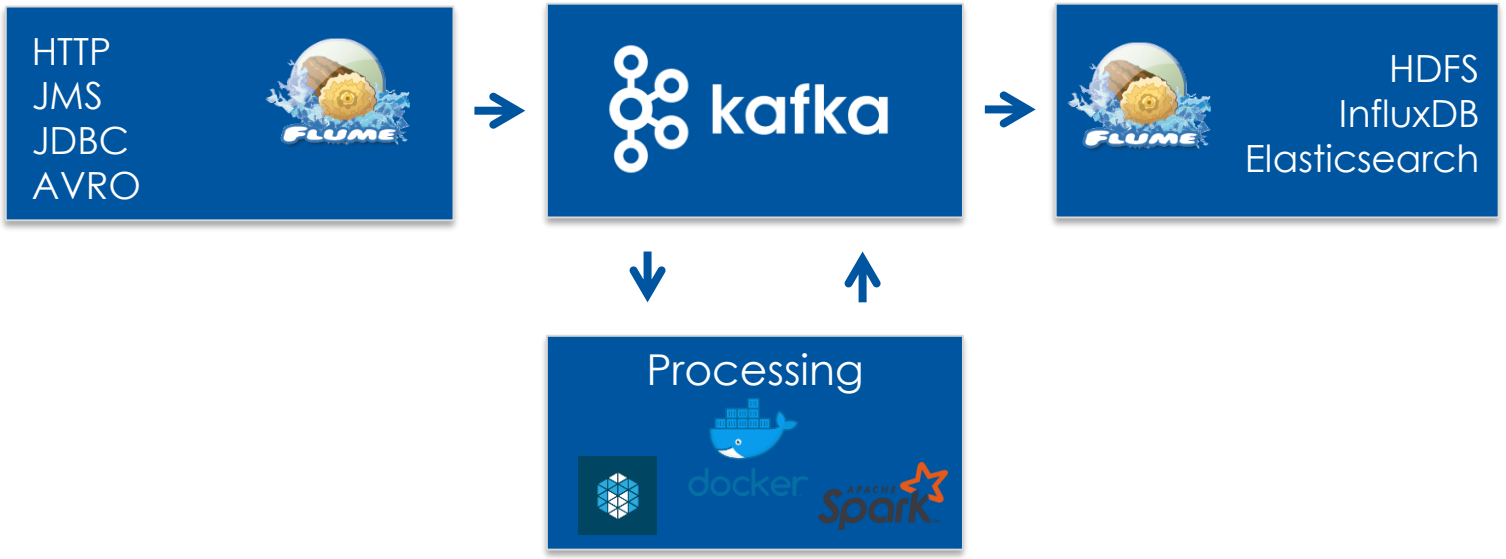


Architecture

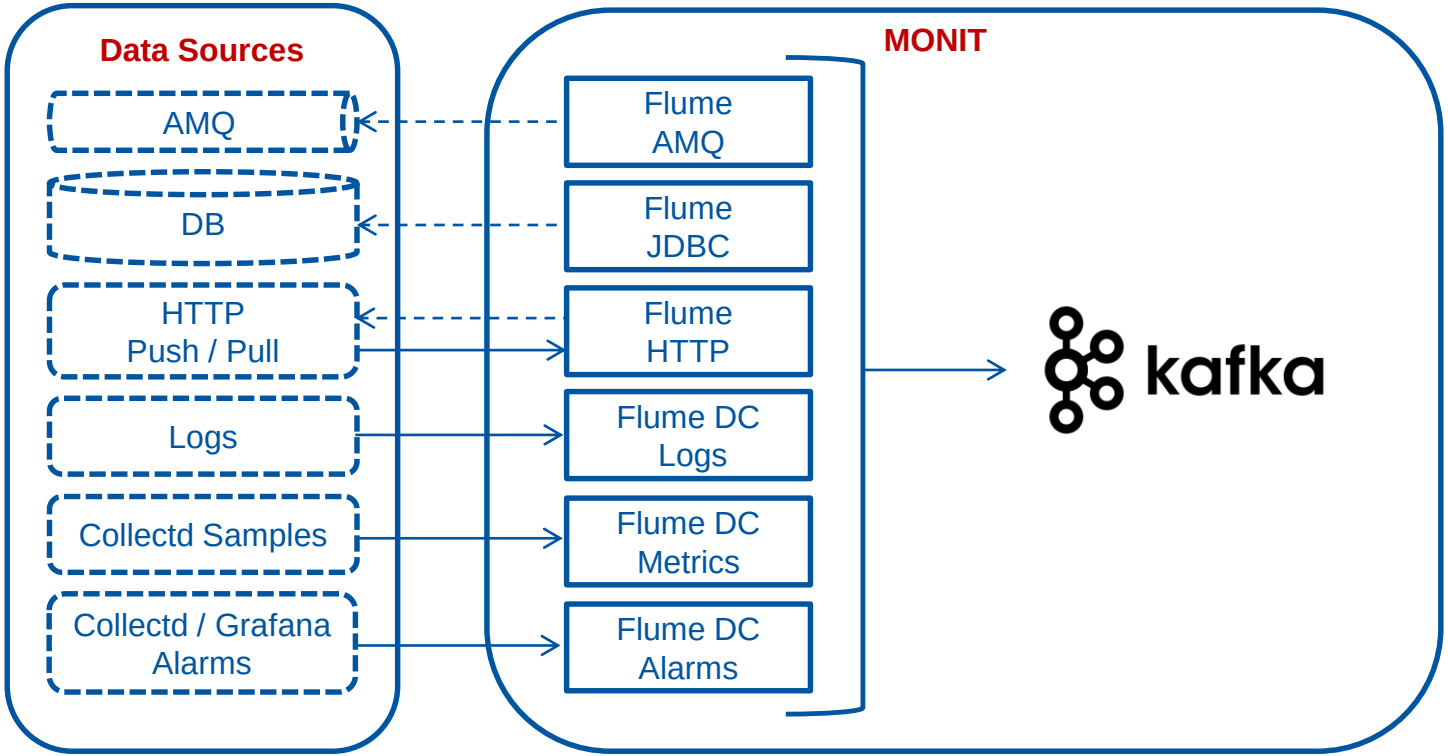
Technologies

- Moving to community built technologies
- Reduce in-house components when possible
- Collaboration with the community:
  - Development, bug reports…

Monitoring Architecture



Data Sources



Data Sources: collectd

- What is collectd

  *collectd is a **daemon** which **collects system and application performance metrics** periodically and provides mechanisms to store the values in a variety of ways*

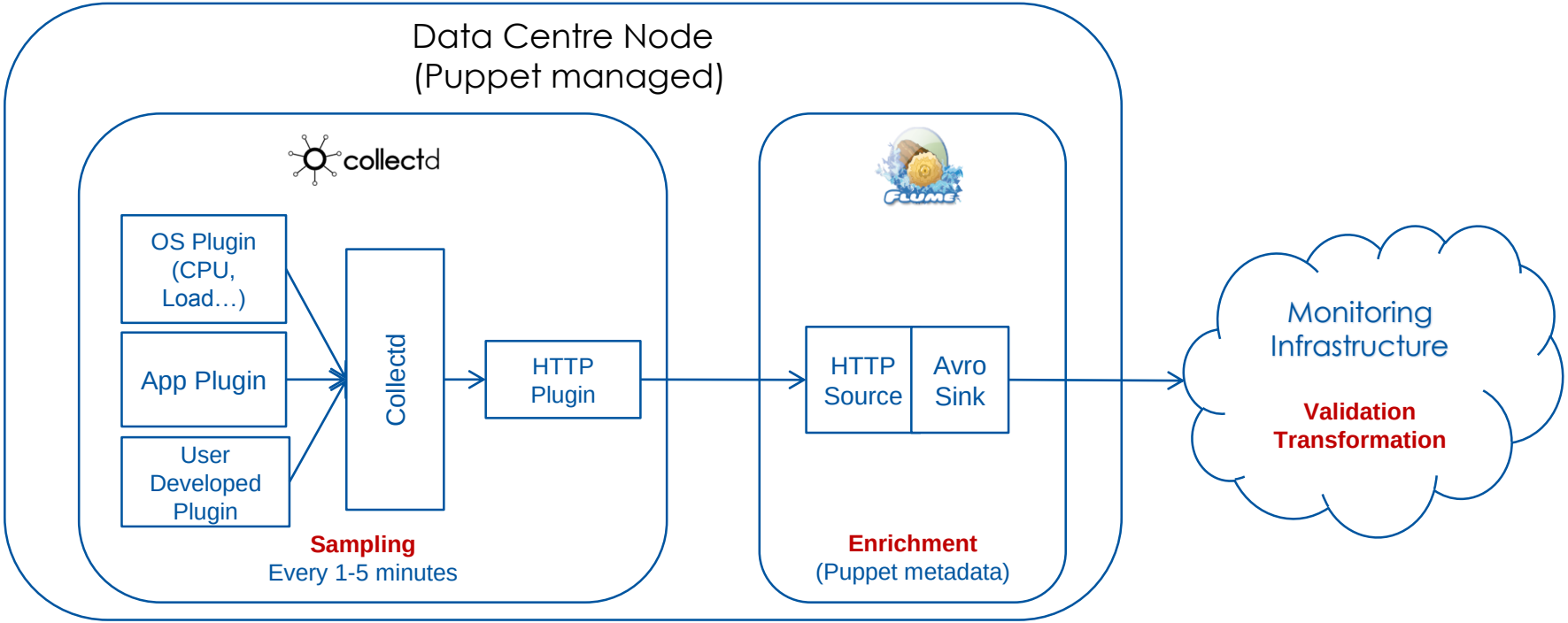
- Why we use collectd
  - Modular and easy to deploy
  - Community built plugins
  - Easy to develop new plugins
  - Continuously improving
    - (but documentation could be better)

Our case

Official support available for most of the metrics

Lemon Metric Classes	Lemon Sensors	Collectd Support	Collectd Plugin
file.filecount	file	Official	https://collectd.org/wiki/index.php/Plugin:FileCount
file.size	file	Official	https://collectd.org/wiki/index.php/Plugin:FileCount
file.spaceUsed	file	Official	https://collectd.org/wiki/index.php/Plugin:FileCount
file.sslmtime	file	?	
log.Parse	parseLog	Official	https://collectd.org/wiki/index.php/Plugin:Tail
log.Parse	parseLog	Official	https://collectd.org/wiki/index.php/Plugin:Tail
cmd.ParseCmd	parse-cmd	Official	https://collectd.org/wiki/index.php/Plugin:Exec
system.bootTime	linux	Official	https://collectd.org/wiki/index.php/Plugin:Uptime
system.contextSwitches	linux	Official	https://collectd.org/wiki/index.php/Plugin:ContextSwitch
system.CPUCount	linux	Official	https://collectd.org/wiki/index.php/Plugin:CPU
system.CPUInfo	linux	?	
system.CPUutil	linux	Official	https://collectd.org/wiki/index.php/Plugin:CPU
system.CPUutilization	linux	Official	https://collectd.org/wiki/index.php/Plugin:CPU
system.createdProcesses	linux	Official	https://collectd.org/wiki/index.php/Plugin:Processes
system.diskStats	linux	Official	https://collectd.org/wiki/index.php/Plugin:Disk
system.existingProcesses	linux	Official	https://collectd.org/wiki/index.php/Plugin:Processes
system.exitCode	linux	Official	https://collectd.org/wiki/index.php/Plugin:Exec
system.fullLoadAvg	linux	Official	https://collectd.org/wiki/index.php/Plugin:Load
system.interrupts	linux	Official	https://collectd.org/wiki/index.php/Plugin:IRQ
system.loadAvg	linux	Official	https://collectd.org/wiki/index.php/Plugin:Load
system.meminfo	linux	Official	https://collectd.org/wiki/index.php/Plugin:Memory
system.memoryShared	linux	Official	https://collectd.org/wiki/index.php/Plugin:Memory
system.memoryStats	linux	Official	https://collectd.org/wiki/index.php/Plugin:Memory
system.networkInterfaceDropped	linux	Official	https://collectd.org/wiki/index.php/Plugin:Interface
system.networkInterfaceInfo	linux	?	
system.networkInterfaceIO	linux	Official	https://collectd.org/wiki/index.php/Plugin:Interface
system.numberOfSockets	linux	Official	https://collectd.org/wiki/index.php/Plugin:TCPConns
system.numberOfUsers	linux	Official	https://collectd.org/wiki/index.php/Plugin:Users

Data Sources: collectd



Transport: Kafka

Kafka as rock-solid core of our pipeline

- Distributed message broker
- Enables stream processing
- 72h retention policy
- Gzip compressed
- Replica factor: 3

Processing

Stream processing

Data enrichment

- Join information from several sources (e.g. network topology)

Data aggregation

- Over time (e.g. summary statistics for a time bin)
- Over other dimensions (e.g. compute a cumulative metric for a set of machines hosting the same service)

Data correlation

- Advanced Alarming: detect anomalies and failures correlating data from multiple sources (e.g. data centre topology-aware alarms)

Batch processing

- Reprocessing, data compression, historical data, periodic reports

Storage

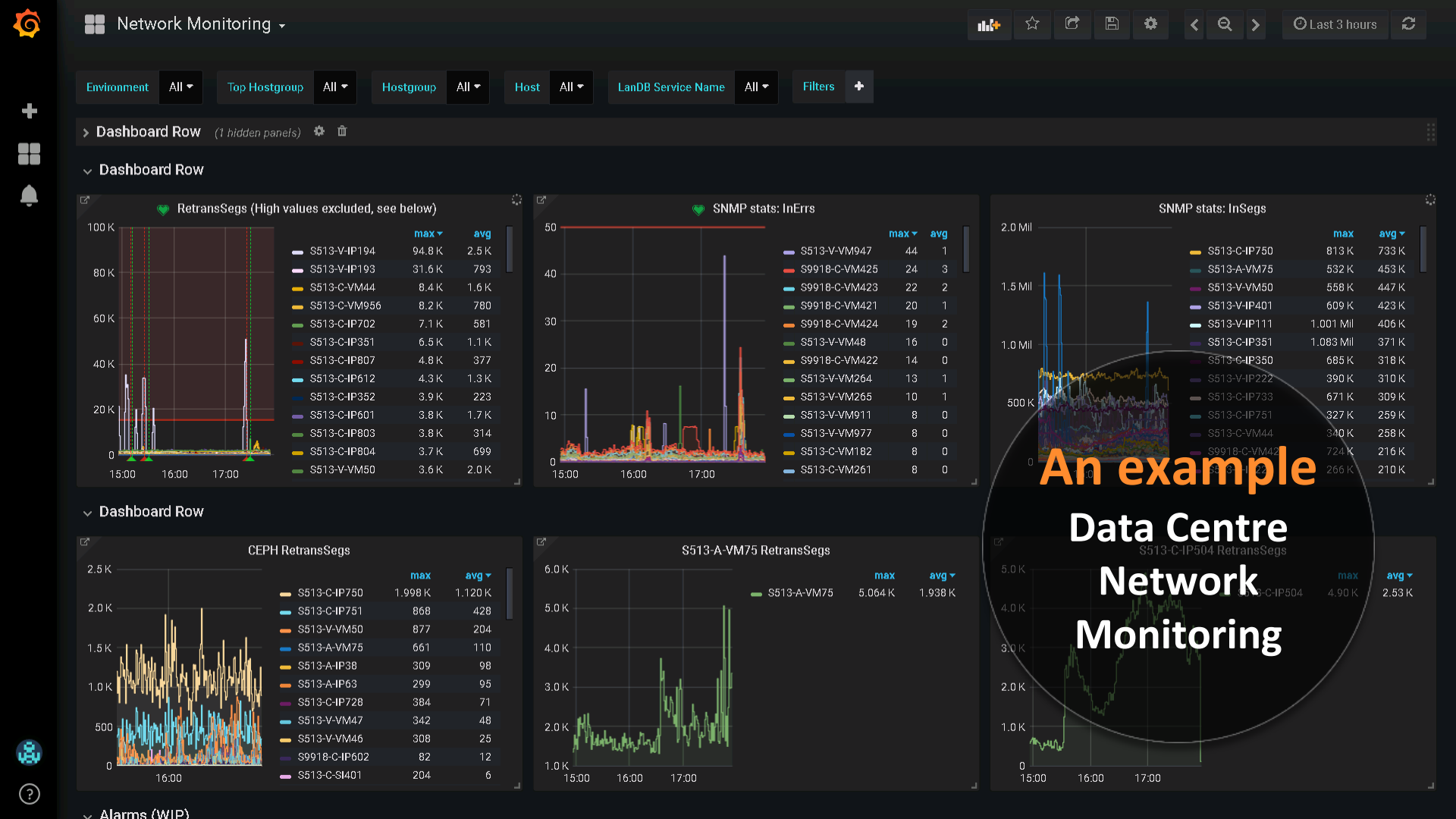
- HDFS for long-term archive
  - Compressed JSON or Parquet
  - Data kept ~ forever
- ES for data exploration & discovery
  - 2 large instances: 30 nodes
  - Data kept for 1 month
- InfluxDB for time-series dashboards
  - > 20 instances
  - From 8GB up to 128GB memory per instance
  - Data kept for ~5y, down-sampling
- Provided by other IT services

Visualization

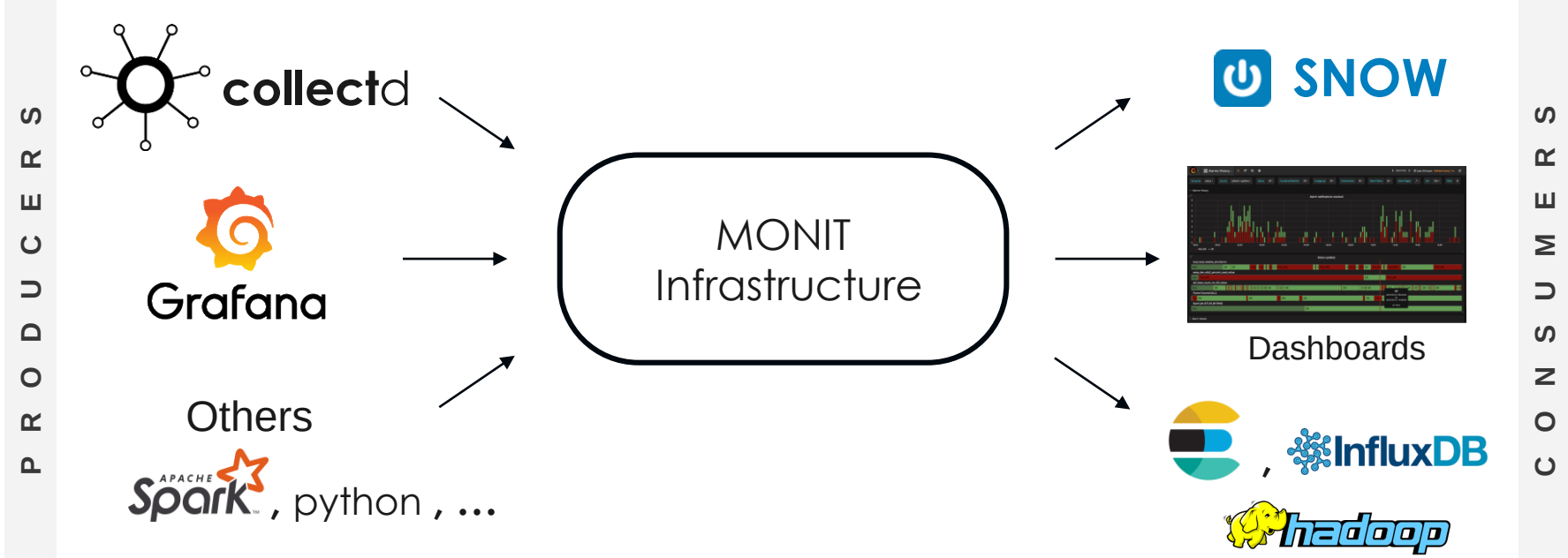
- Grafana for dashboards
  - Users can create their own
- Kibana for data exploration
  - Data discovery and logs
- Swan for analytics (notebooks)

Visualization: Grafana

- Open-source platform for dashboards
- Support multiple back ends: ES / InfluxDB
- Advanced features
  - Templates, ad-hoc filters, auto-completion
  - Advanced query syntax, organizations, ACLs
  - Alarms


Network Monitoring
Last 3 hours
Environment All
Top Hostgroup All
Hostgroup All
Host All
LanDB Service Name All
Filters
Dashboard Row (1 hidden panels)
Dashboard Row
RetransSegs (High values excluded, see below)
max avg
S513-V-IP194 94.8 K 2.5 K
S513-V-IP193 31.6 K 793
S513-C-VM44 8.4 K 1.6 K
S513-C-VM956 8.2 K 780
S513-C-IP702 7.1 K 581
S513-C-IP351 6.5 K 1.1 K
S513-C-IP807 4.8 K 377
S513-C-IP612 4.3 K 1.3 K
S513-C-IP352 3.9 K 223
S513-C-IP601 3.8 K 1.7 K
S513-C-IP803 3.8 K 314
S513-C-IP804 3.7 K 699
S513-V-VM50 3.6 K 2.0 K
SNMP stats: InErrs
max avg
S513-V-VM947 44 1
S9918-C-VM425 24 3
S9918-C-VM423 22 2
S9918-C-VM421 20 1
S9918-C-VM424 19 2
S513-V-VM48 16 0
S9918-C-VM422 14 0
S513-V-VM264 13 1
S513-V-VM265 10 1
S513-V-VM911 8 0
S513-V-VM977 8 0
S513-C-VM182 8 0
S513-C-VM261 8 0
SNMP stats: InSegs
max avg
S513-C-IP750 813 K 733 K
S513-A-VM75 532 K 453 K
S513-V-VM50 558 K 447 K
S513-V-IP401 609 K 423 K
S513-V-IP111 1.001 Mil 406 K
S513-C-IP351 1.083 Mil 371 K
S513-C-IP350 685 K 318 K
S513-V-IP222 390 K 310 K
S513-C-IP733 671 K 309 K
S513-C-IP751 327 K 259 K
S513-C-VM44 340 K 258 K
S9918-C-VM42 724 K 216 K
266 K 210 K
Dashboard Row
CEPH RetransSegs
max avg
S513-C-IP750 1.998 K 1.120 K
S513-C-IP751 868 428
S513-V-VM50 877 204
S513-A-VM75 661 110
S513-A-IP38 309 98
S513-A-IP63 299 95
S513-C-IP728 384 71
S513-V-VM47 342 48
S513-V-VM46 308 25
S9918-C-IP602 82 12
S513-C-SI401 204 6
S513-A-VM75 RetransSegs
max avg
S513-A-VM75 5.064 K 1.938 K
S513-C-IP504 RetransSegs
max avg
S513-C-IP504 4.90 K 2.53 K
An example
Data Centre
Network
Monitoring
Alarms (WIP)

Alarms: Overview

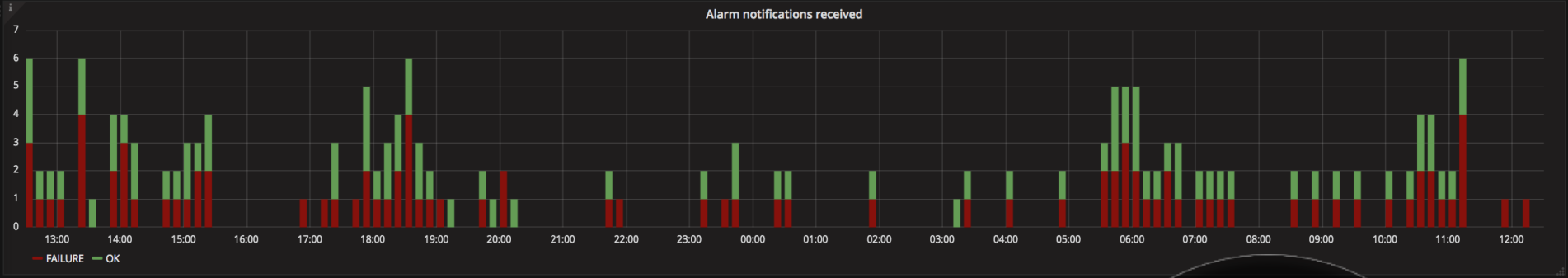


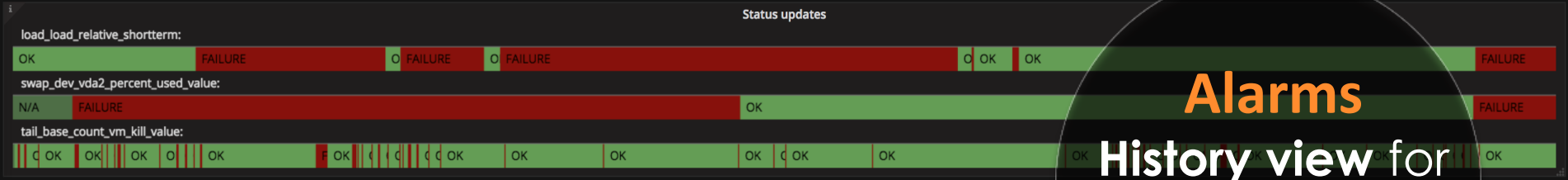
Alarms

- Standard representation as any other *MONIT document*
- Integrated notification system
  - All alarms can produce: support tickets, emails, etc.
- Collectd alarms
  - Use collectd local metrics with *threshold* plugin
  - *Low level,* created at the (virtual) machine
- Grafana alarms
  - Defined using the Grafana GUI
  - Integrated *automatically* into the MONIT workflow

Group by status | Source collectd + grafana | Status All | Functional Element LXBATCH | Hostgroup All | Environment All | Alarm Name All | Alarm Regex .* | Bin 10m | Filter +

Alarms History



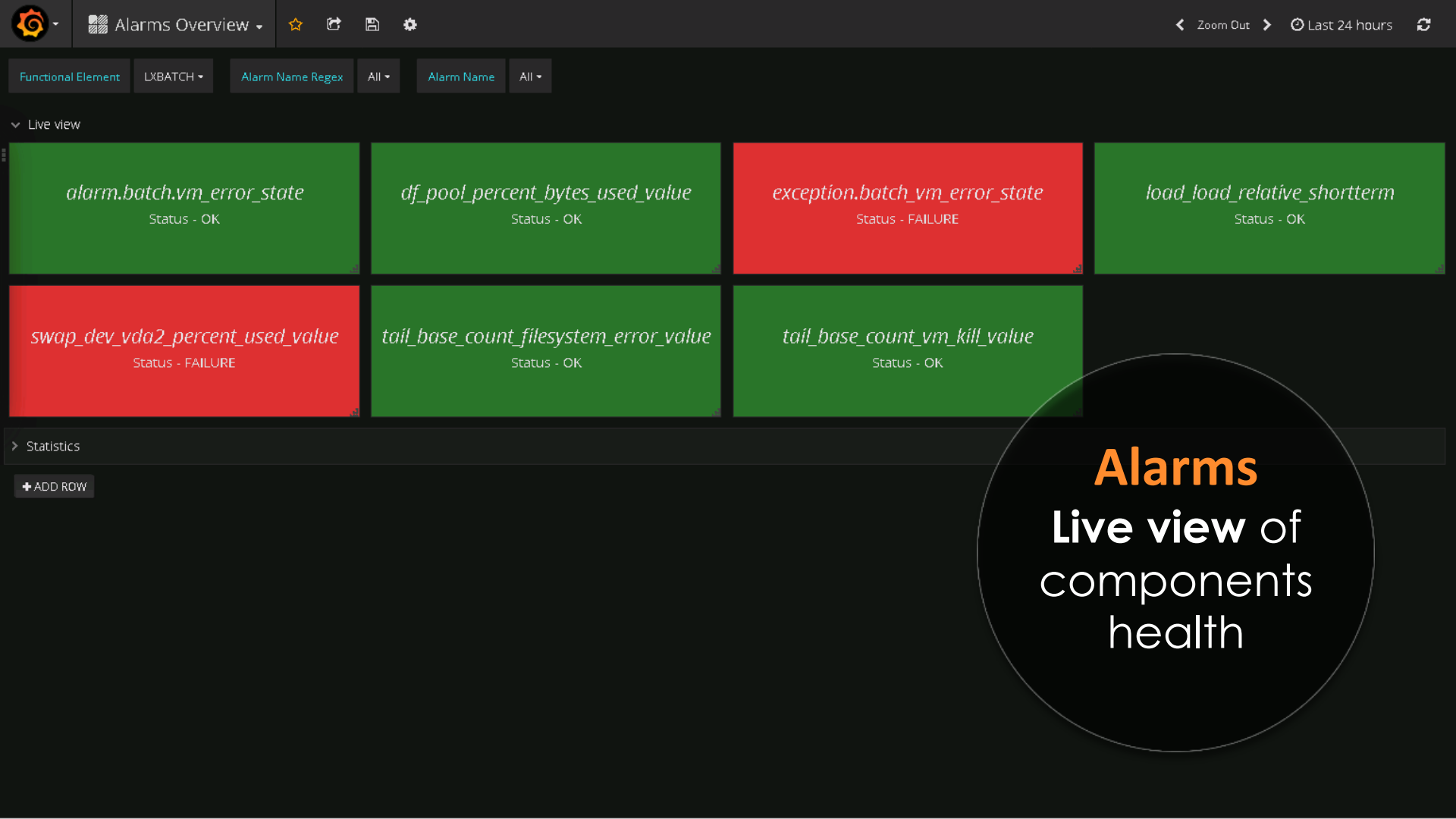


Alarms

History view for all hosts and services

Alarm Details

Time	alarm_id	alarm_name	entities	functional_element	metrics	source	status	submitter_environment	submitter_hostgroup	toplevel_hostgroup	correlation
2018-05-02 12:10:54		swap_dev_vda2_percent_used_value	b6d596ff20.cern.ch	LXBATCH	swap_percent	collectd	FAILURE	batchtest	bi/condor/gridworker/shareshort	bi	99.99936\ >\ 85.0
2018-05-02 11:57:18		load_load_relative_shortterm	b6894bb3d5.cern.ch	LXBATCH	load	collectd	FAILURE	batchtest	bi/condor/gridworker/shareshort	bi	12.0\ >\ 10.0
2018-05-02 11:19:51		tail_base_count_vm_kill_value	b6d596ff20.cern.ch	LXBATCH	tail_count	collectd	OK	batchtest	bi/condor/gridworker/shareshort	bi	nan\ >\ 0.0
2018-05-02 11:18:15		tail_base_count_vm_kill_value	b6d596ff20.cern.ch	LXBATCH	tail_count	collectd	FAILURE	batchtest	bi/condor/gridworker/shareshort	bi	1.0\ >\ 0.0
2018-05-02 11:15:45		load_load_relative_shortterm	b6d46d7d29.cern.ch	LXBATCH	load	collectd	FAILURE	batchtest	bi/condor/gridworker/shareshort	bi	10.87875\ >\ 10.0
2018-05-02 11:13:02		swap_dev_vda2_percent_used_value	b6d596ff20.cern.ch	LXBATCH	swap_percent	collectd	FAILURE	batchtest	bi/condor/gridworker/shareshort	bi	94.54682\ >\ 85.0
2018-05-02 11:11:07		tail_base_count_vm_kill_value	b6d596ff20.cern.ch	LXBATCH	tail_count	collectd	OK	batchtest	bi/condor/gridworker/shareshort	bi	nan\ >\ 0.0
2018-05-02 11:10:34		tail_base_count_vm_kill_value	b6d596ff20.cern.ch	LXBATCH	tail_count	collectd	FAILURE	batchtest	bi/condor/gridworker/shareshort	bi	1.0\ >\ 0.0

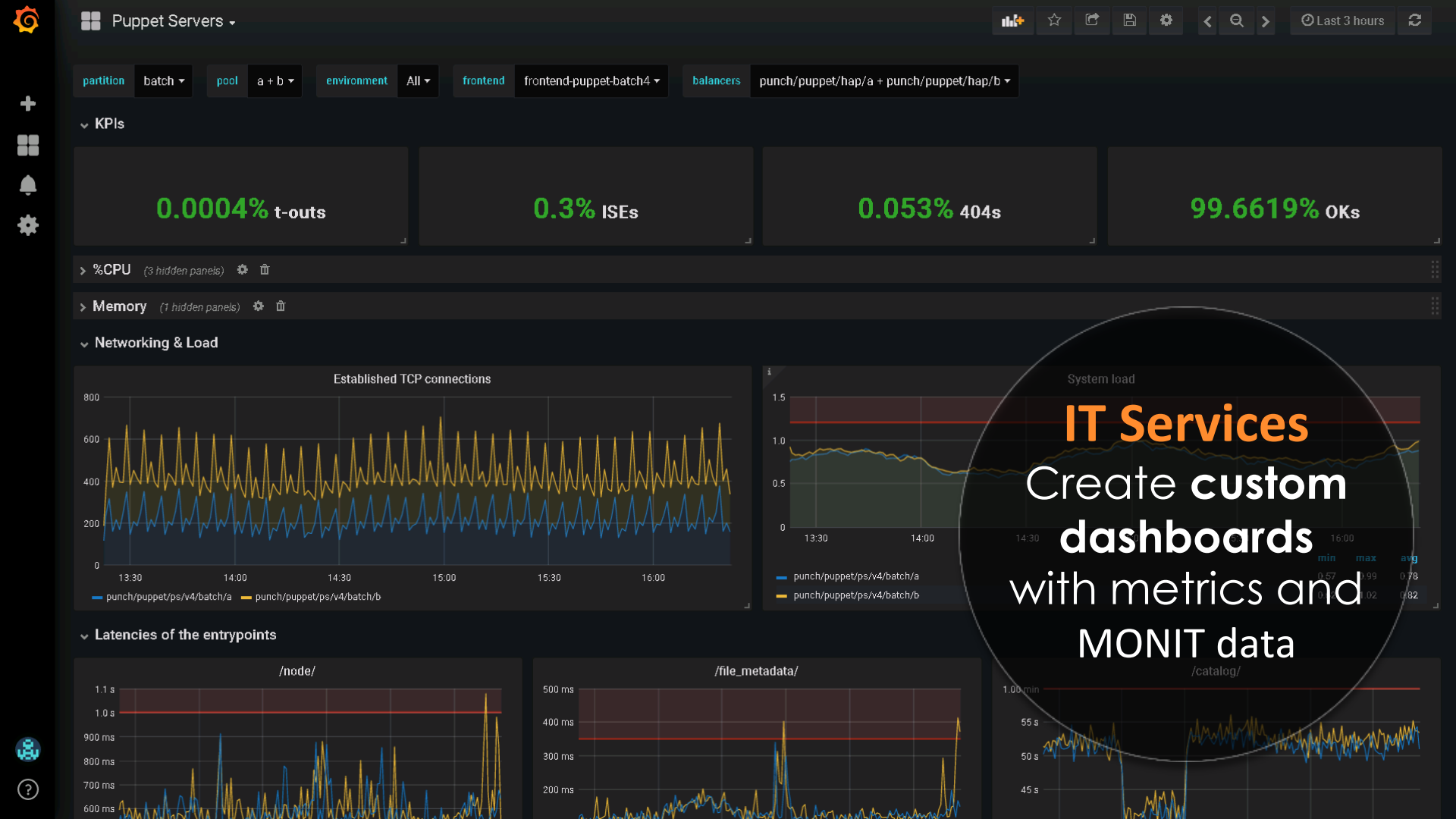

Alarms Overview
Zoom Out
Last 24 hours
Functional Element
LXBATCH
Alarm Name Regex
All
Alarm Name
All
Live view
alarm.batch.vm_error_state
Status - OK
df_pool_percent_bytes_used_value
Status - OK
exception.batch_vm_error_state
Status - FAILURE
load_load_relative_shortterm
Status - OK
swap_dev_vda2_percent_used_value
Status - FAILURE
tail_base_count_filesystem_error_value
Status - OK
tail_base_count_vm_kill_value
Status - OK
Statistics
ADD ROW
Alarms
Live view of components health

Operations

Using MONIT at CERN DC

1. Configure the service using the Agile Infrastructure (**Puppet**)
   - Basic monitoring configuration provided by MONIT
2. Configure service specific metrics and logs
   - Configure **collectd plugins** in Puppet
   - Configure local **log collector**
3. Monitor the service
   - General monitoring dashboards available in Grafana
   - Browse logs and discover service-specific data in Kibana
   - Create service-specific dashboards in Grafana
   - More advanced use cases


Host Metrics
Last 3 hours
Environment All
Hostgroup lxplus/nodes/login
Host All
Availability Zone All
Data Range one_week
Bin auto
Filters
CPU
CPU usage
100%
75%
50%
25%
0%
13:30 13:40 13:50 14:00 14:10 14:20 14:30 14:40 14:50 15:00 15:10 15:20 15:30 15:40 15:50 16:00 16:10 16:20
idle interrupt nice softirq steal system user wait
Network
Network usage
1.0 Gbps
750 Mbps
500 Mbps
250 Mbps
0 bps
13:30 13:40 13:50 14:00 14:10 14:20 14:30 14:40 14:50 15:00 15:10 15:20 15:30 16:20
rx tx
Memory
Memory usage
37 GiB
28 GiB
19 GiB
9 GiB
0 B
13:30 13:40 13:50 14:00 14:10 14:20 14:30 14:40 14:50 15:00 15:10 15:20 15:30 15:40 15:50 16:00 16:10 16:20
buffered cached free slab_recl slab_unrecl used
OS Metrics
Basic Monitoring
dashboard
for every host
in the DC


Puppet Servers
Last 3 hours
partition batch
pool a + b
environment All
frontend frontend-puppet-batch4
balancers punch/puppet/hap/a + punch/puppet/hap/b
KPIs
0.0004% t-outs
0.3% ISEs
0.053% 404s
99.6619% OKs
%CPU (3 hidden panels)
Memory (1 hidden panels)
Networking & Load
Established TCP connections
punch/puppet/ps/v4/batch/a
punch/puppet/ps/v4/batch/b
System load
punch/puppet/ps/v4/batch/a
punch/puppet/ps/v4/batch/b
IT Services
Create custom dashboards with metrics and MONIT data
Latencies of the entrypoints
/node/
/file_metadata/
/catalog/

MONIT Operations

- Based on Openstack VMs (CentOS7)
- Puppet for service configuration
- Separate production/QA clusters
- Collectd for infrastructure metrics
- MONIT alarms

Lessons learned

- Migration to open source technologies pays back
- Building a reliable pipeline
  - Decouple producers from back ends
  - Producers isolation (1 producer, 1 topic)
  - Rely on technology that scales (Kafka over Flume)

Lessons learned

- Component based approach over monolithic solutions
- Get to know your storage systems
  - InfluxDB
    - Works well with time-series data
    - SQL-like language with built-in aggregations
    - Requires care with high cardinality in number of series
      - Needs some schema design
  - Elasticsearch
    - Good for discovery and search of logs
    - Complex engine that requires special care with configuration
  - HDFS
    - Only for batch or high latency scenarios

Summary

- Production infrastructure in place
  - Including metrics, logs and alarms
- Based on mainstream and open source technologies
- We provide monitoring as a service

Thank you

- Our docs: http://monit-docs.web.cern.ch/monit-docs/
- Questions / comments are welcome!


CERN