INFN and The Future of Scientific Computing

Turin May 4, 2018

Using Hadoop ecosystem tools for distributed datacenters and the ALICE O2 farm monitoring

Gioacchino Vino (INFN Bari)

INFN BARI
Istituto Nazionale di Fisica Nucleare
Sezione di Bari

Tutor: Dott. Domenico Elia

Who am I?

- Postgraduate course in “Development and management of data centers for high performance scientific computing”, 2014-2015
  - Thesis title: “**Dashboard for the ALICE activity in Bari Tier-2 Site**”
  - Tutors: Domenico Elia and Antonio Franco
- Scholarship at GARR in “**Monitoring system for geographically distributed datacenters based on Openstack**”, 2016-2017
  - Tutors: Domenico Elia and Giacinto Donvito
- Scholarship at INFN, currently working on “**Monitoring of the ALICE O2 Facility**”, since Feb 2018
  - Tutor: Domenico Elia

INFN
BARI
Istituto Nazionale di Fisica Nucleare
Sezione di Bari

Index

- Monitoring of geographically distributed datacenter based on OpenStack: MonGARR
  - Motivations
  - Project Overview
  - Future works
- Monitoring of the ALICE O2 Facility @CERN: Modular Stack
  - Architecture
  - Future works

MonGARR: Motivations

The increasing of computation resource demand for scientific purposes is leading to:

- Datacenters increasing in complexity and size.
- Taking advantages of new technologies like virtualization and cloud computing.
- Datacenter cooperation needed in order to accomplish common goals.

⬇

Geographically distributed datacenters

- Goal: Increase the computation capability of overall system.
- Side effect: Increasing complexity from the monitoring and control system.

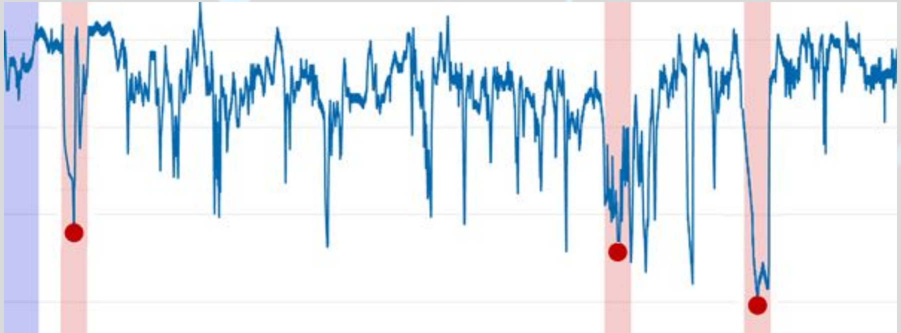
⬇

Project: Developing a monitoring system for geographically distributed datacenters.

MonGARR: Project Overview

Advanced features are required:

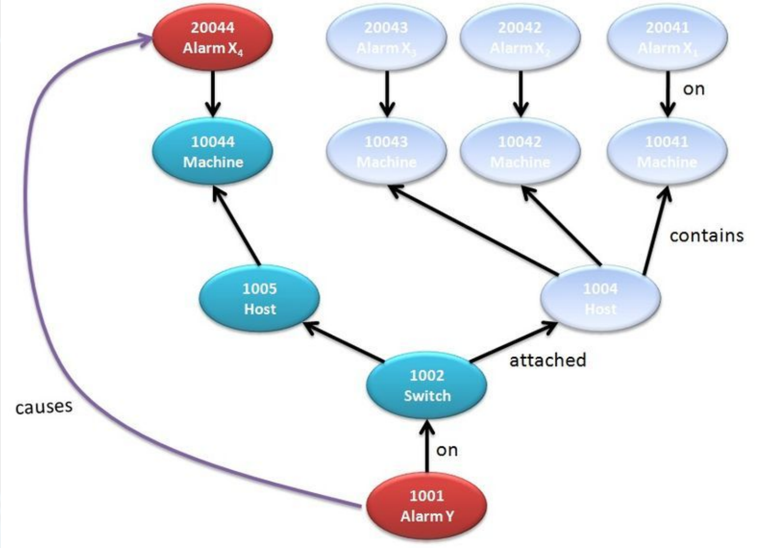
- Anomaly detector



MonGARR: Project Overview

Advanced features are required:

- Anomaly detector
- Root Cause Analysis



MonGARR: Project Overview

Advanced features are required:

- Anomaly detector
- Root Cause Analysis

Fully informative monitoring data are collected:

- Service monitoring (HTTP server, DBs, …)
- Openstack and middleware monitoring
- Hardware monitoring (physical servers, disks, disk controllers, network devices, PDU, …)

MonGARR: Project Overview

Testbed

ReCaS Bari Datacenter:

- More than 13.000 cores
- 7.1 PB Disk Storage
- 2.5 PB Tape storage
- HPC Cluster composed of 20 servers
- Dedicated network link: 10Gbps x2 to GARR, 20Gbps to Naples and 20 Gbps to Bologna
- Cloud platform: OpenStack
- Batch system: HTCondor
  - 184 Worker Nodes
  - 350+ network connections
- Local Monitoring System: Zabbix
- Including ALICE and CMS Tier2s

MonGARR: Project Overview

Syslog

Zabbix

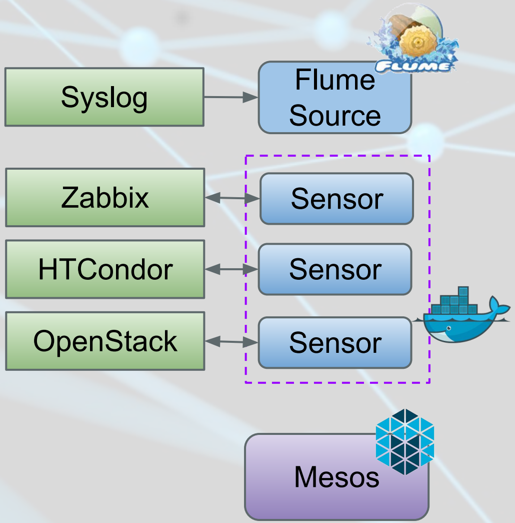
HTCondor

OpenStack

Data Sources:

- **Syslog**: System processes and service information.
- **Zabbix**: Computation resource usage, service and Openstack monitoring.
- **HTCondor**: Scheduler, completed and running job state
- **OpenStack**: Information on server, images, flavors, volumes, network devices, ….

MonGARR: Project Overview

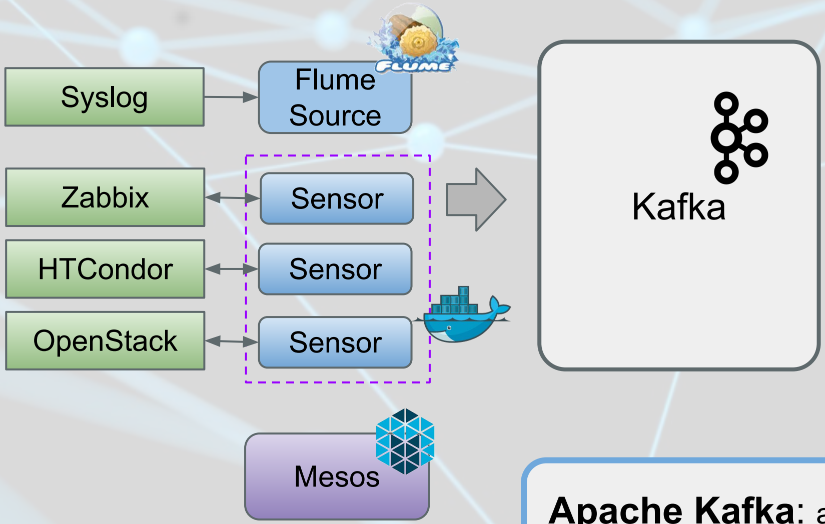


Metric collectors:

- **Apache Flume** Syslog Source.
- Python code inserted in **Docker**-container and executed periodically using **Apache Mesos.**

Apache Flume: a distributed and highly-reliable service for collecting, aggregating and moving large amounts of data in a very efficient way.
Apache Mesos: an open-source project to manage computer clusters.
Docker: a computer program that performs operating-system-level virtualization also known as containerization.

MonGARR: Project Overview

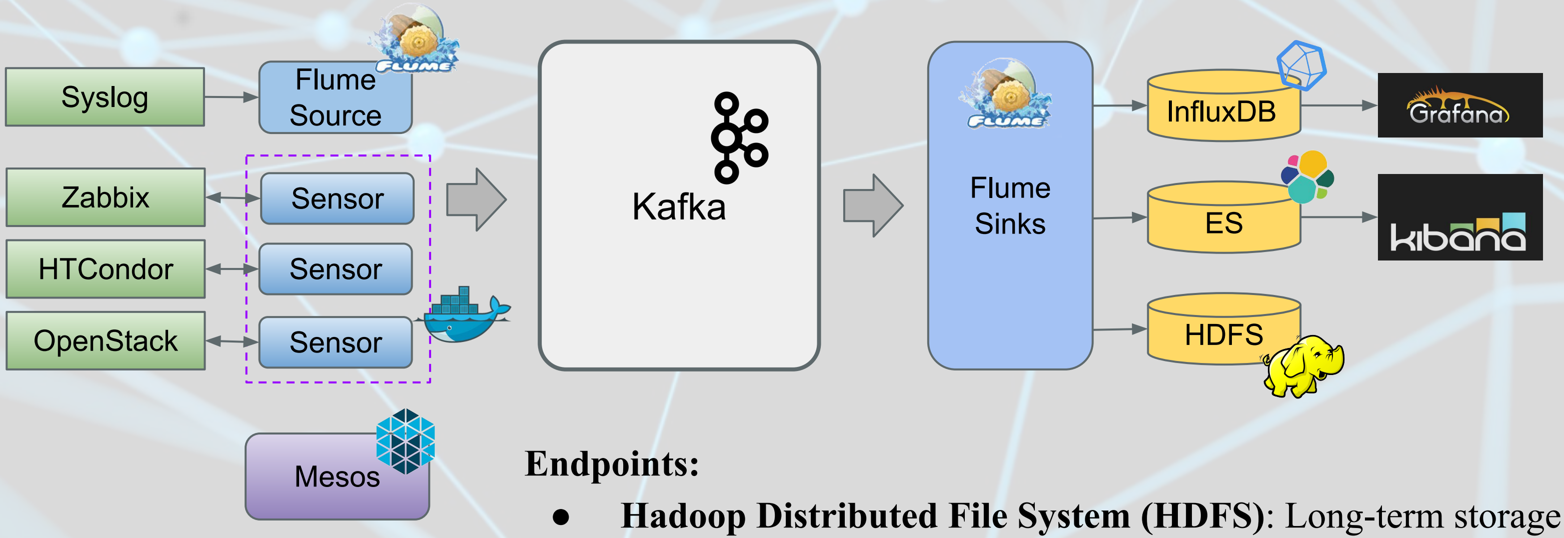


Transport Layer:

- **Apache Kafka.**
- Decouple all components.
- Increase the High Availability of system.

Apache Kafka: an open-source stream-processing software platform, provides a unified, high-throughput, low-latency platform for handling real-time data feeds.

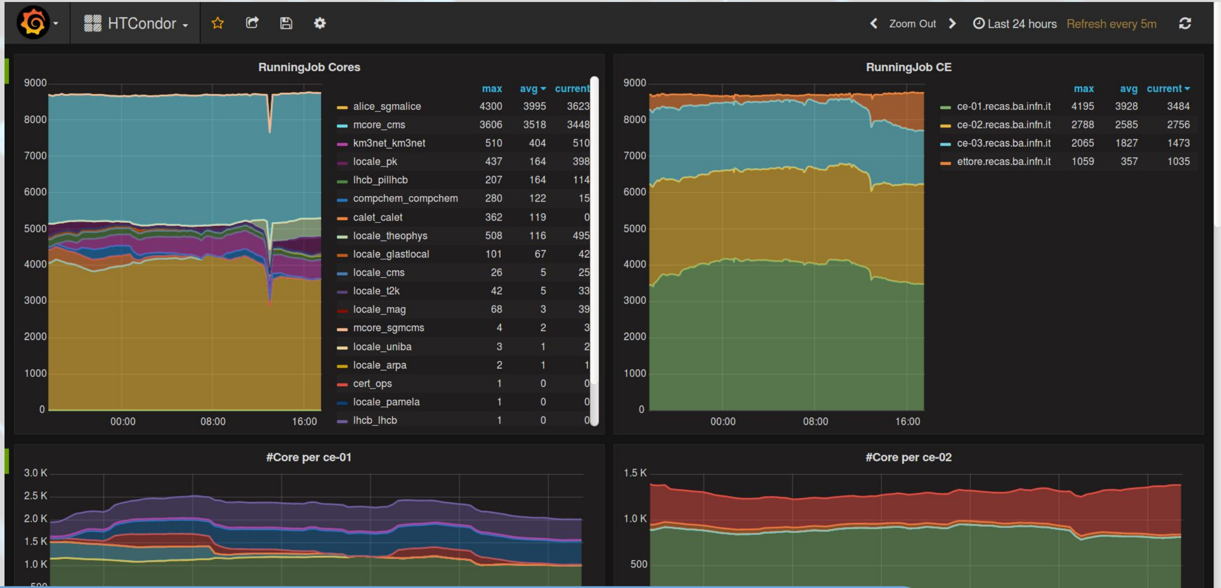
MonGARR: Project Overview



Endpoints:

- **Hadoop Distributed File System (HDFS)**: Long-term storage.
- **InfluxDB-Grafana**: Timeseries Dashboards.
- **ElasticSearch-Kibana**: Log Dashboards.

MonGARR: Project Overview



RunningJob Cores

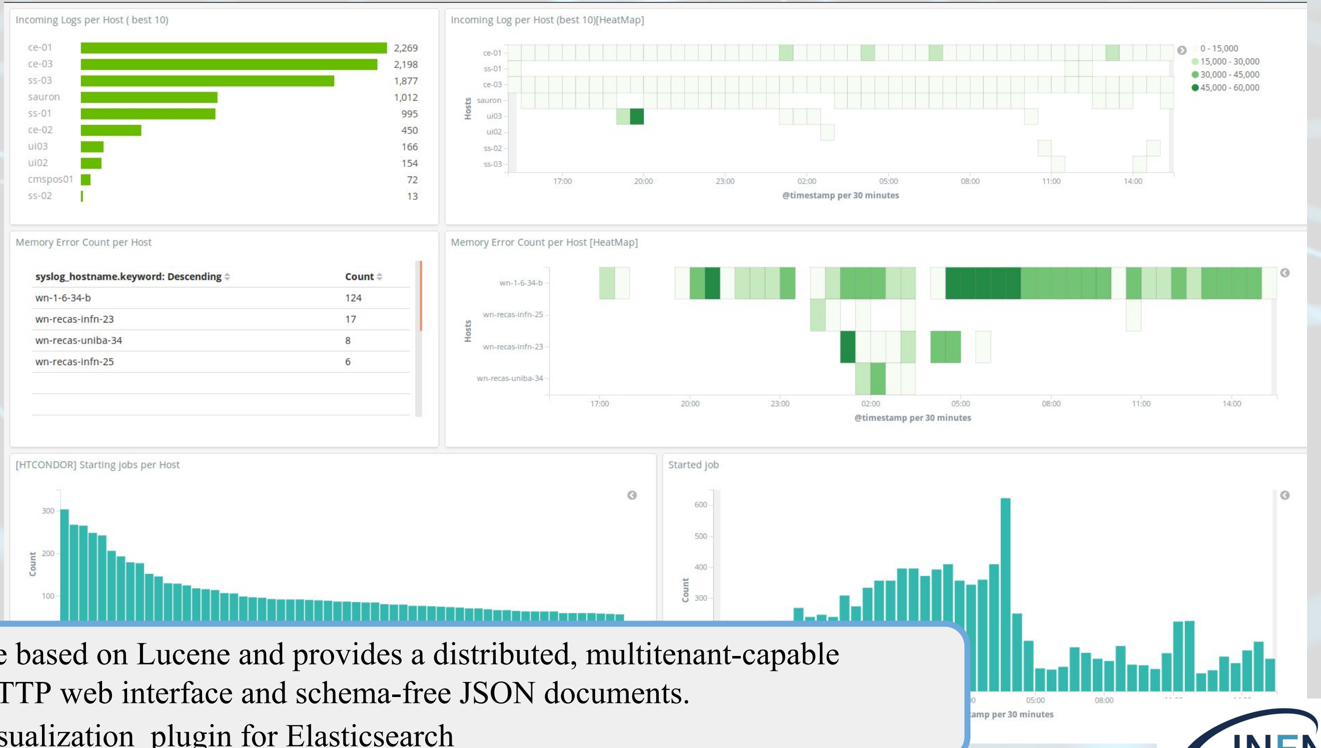
	max	avg ▾	current
alice_sgmalice	4300	3995	3623
mcore_cms	3606	3518	3448
km3net_km3net	510	404	510
locale_pk	437	164	398
lhcb_pillhcb	207	164	114
compchem_compchem	280	122	15
calet_calet	362	119	0
locale_theophys	508	116	495
locale_glastlocal	101	67	42
locale_cms	26	5	25
locale_t2k	42	5	33
locale_mag	68	3	39
mcore_sgmcms	4	2	3
locale_uniba	3	1	2
locale_arpa	2	1	1
cert_ops	1	0	0
locale_pamela	1	0	0
lhcb_lhcb	1	0	0

RunningJob CE

	max	avg	current ▾
ce-01.recas.ba.infn.it	4195	3928	3484
ce-02.recas.ba.infn.it	2788	2585	2756
ce-03.recas.ba.infn.it	2065	1827	1473
ettore.recas.ba.infn.it	1059	357	1035

InfluxDB: a custom high-performance data store written specifically for time series data.
Grafana: Dashboards’ builder for time-series data.

MonGARR: Project Overview



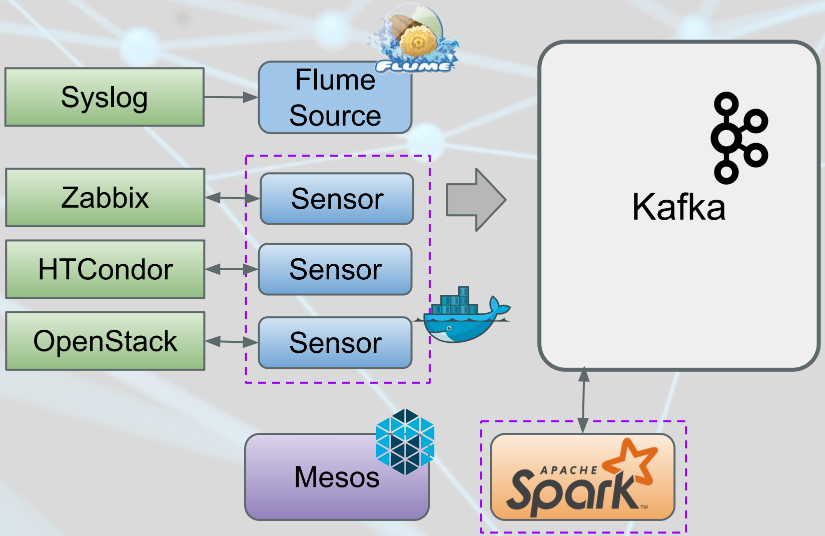
Memory Error Count per Host

syslog_hostname.keyword: Descending ⇕	Count ⇕
wn-1-6-34-b	124
wn-recas-infn-23	17
wn-recas-uniba-34	8
wn-recas-infn-25	6

ElasticSearch: a search engine based on Lucene and provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

Kibana: an open source data visualization plugin for Elasticsearch

MonGARR: Project Overview

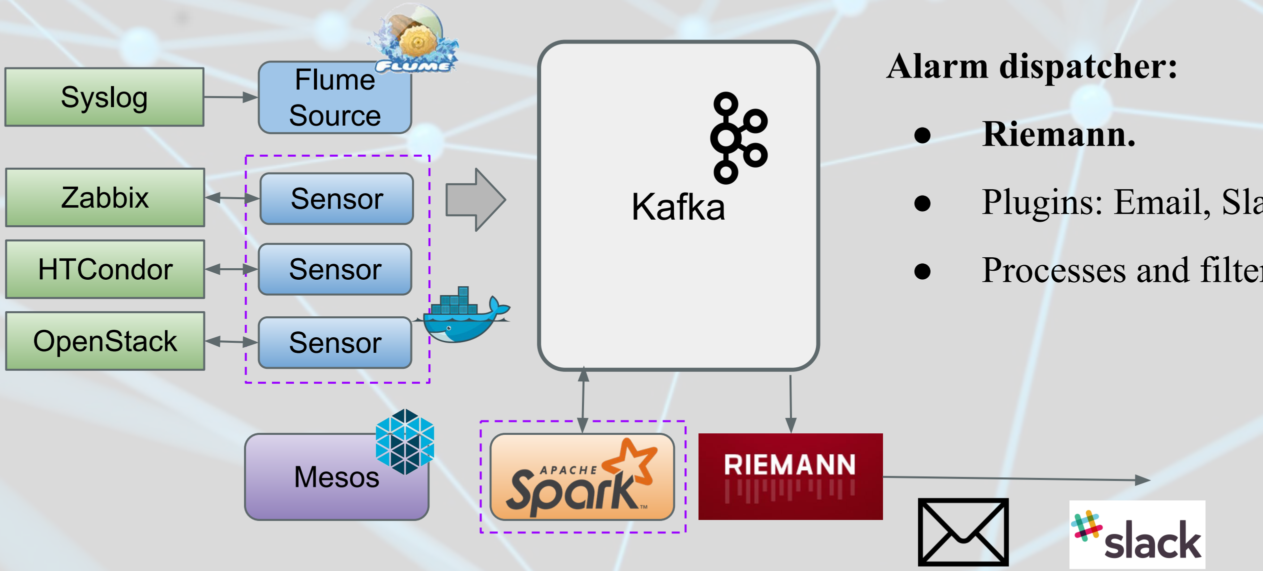


Processing Unit:

- **Apache Spark.**
- Log Analyzer.
- Anomaly Detector.
- Data Correlation.
- Root Cause Analysis.

Apache Spark: a fast and general engine for large-scale data processing.

MonGARR: Project Overview

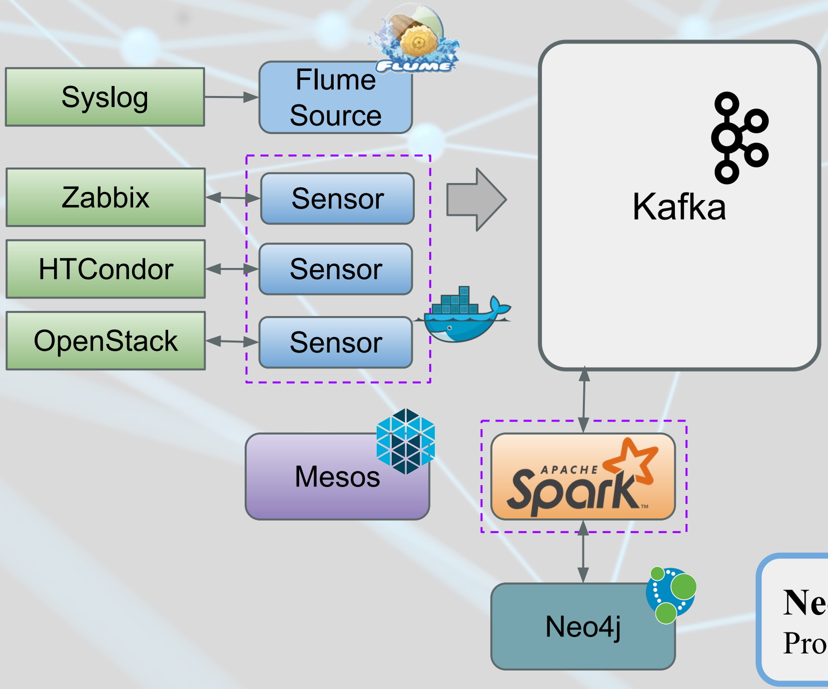


Alarm dispatcher:

- **Riemann.**
- Plugins: Email, Slack.
- Processes and filters events.

Riemann: aggregates events from your servers and applications with a powerful stream processing language.

MonGARR: Project Overview



Information Structure:

- Classical monitoring is not enough.
- Relation information (Services, network, virtual-physical server, …)
  - Openstack data.
  - Open connections.
  - Other monitoring data.

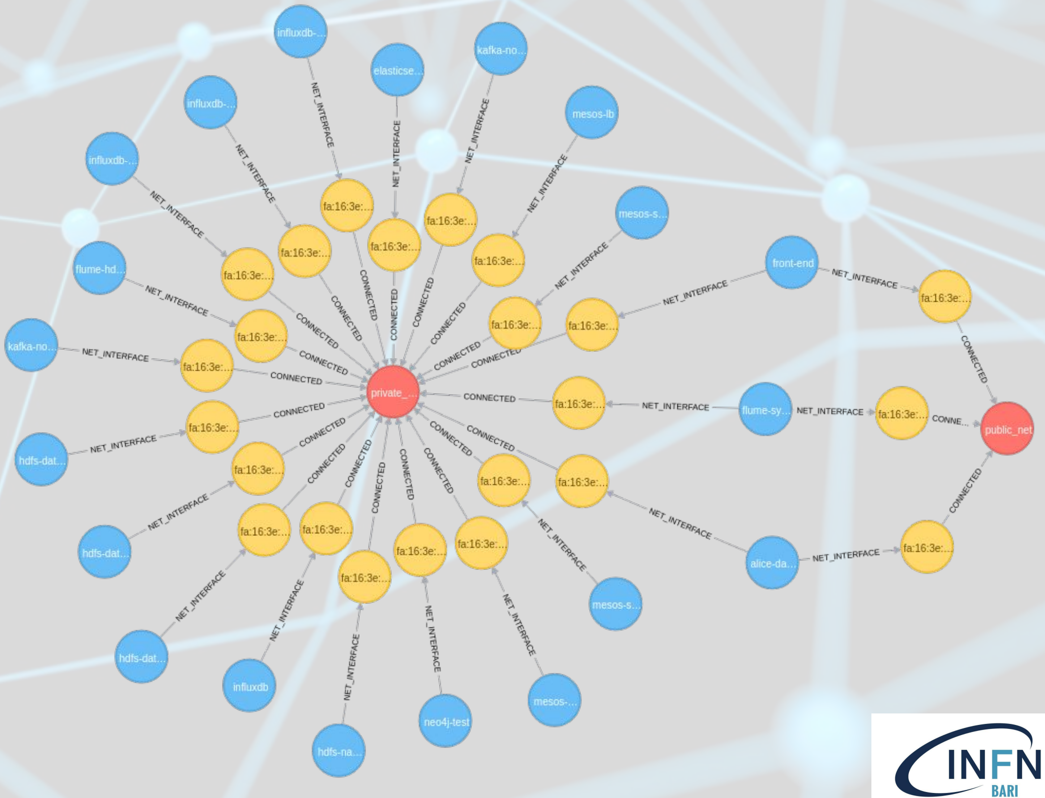
Neo4j: High Performance native Graph Storage & Processing.

MonGARR: Project Overview

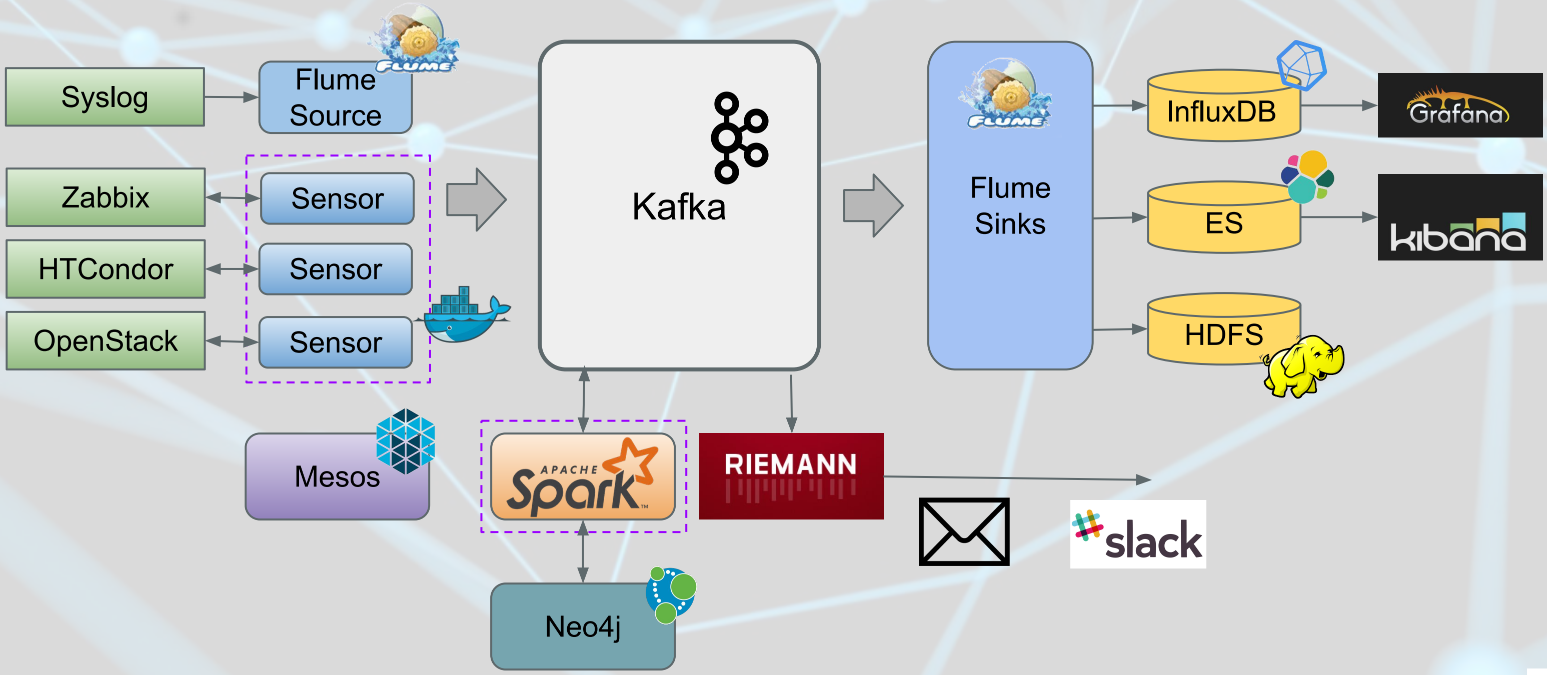
Information Structure:

Subgraph example:

- Blues nodes: virtual machines.
- Yellow nodes: network interfaces.
- Red nodes: networks.



MonGARR: Project Overview



MonGARR: Project Overview

Resource Usage for the monitoring system:

- 80 CPUs
- 150GB RAM
- 3 TB Disk
    - 1.5TB for HDFS in replica 3
    - 600 GB for Kafka nodes
    - No-volatile virtual machine volumes

MonGARR: Project Overview

Apache Mesos:

Cluster:

- 3x Master (2 CPUs, 4GB RAM, 20GB Disk)
- 2x Slaves (4 CPUs, 8GB RAM, 20 GB Disk)
- 1x Load Balancer (2 CPUs, 4GB RAM, 20GB Disk)

Frameworks:

- Chronos
- Marathon
- Spark

MonGARR: Future works

- Migrate all components in Mesos
- Improve the Machine Learning algorithms efficacy
- Root Cause Analysis algorithm
- Integration with project management systems (OpenProject, Trello, ….)

Modular Stack solution for ALICE O2 monitoring

- ALICE is a heavy-ion detector designed to study the physics of strongly interacting matter (the Quark–Gluon Plasma) at the CERN Large Hadron Collider (LHC).
- During the Long Shutdown 2 in the end of 2018, ALICE will start its upgrade to fully exploit the increase in luminosity.
- The current computer system (Data Acquisition, High-Level Trigger and Offline) will be replaced by a single, common O2 (Online-Offline) system.
- Some detectors will be read out continuously, without physics triggers.
- O2 Facility will compress the 3.4 TB/s of raw data to 100 GB/s of reconstructed data

⬇

- Development of a Monitoring System for ALICE O2 Facility:
  Modular Stack solution, with components and tools already used and tested in the MonGARR project (approved by the ALICE O2 TB last February)

Modular Stack solution for ALICE O2 monitoring

ALICE O2 Facility:

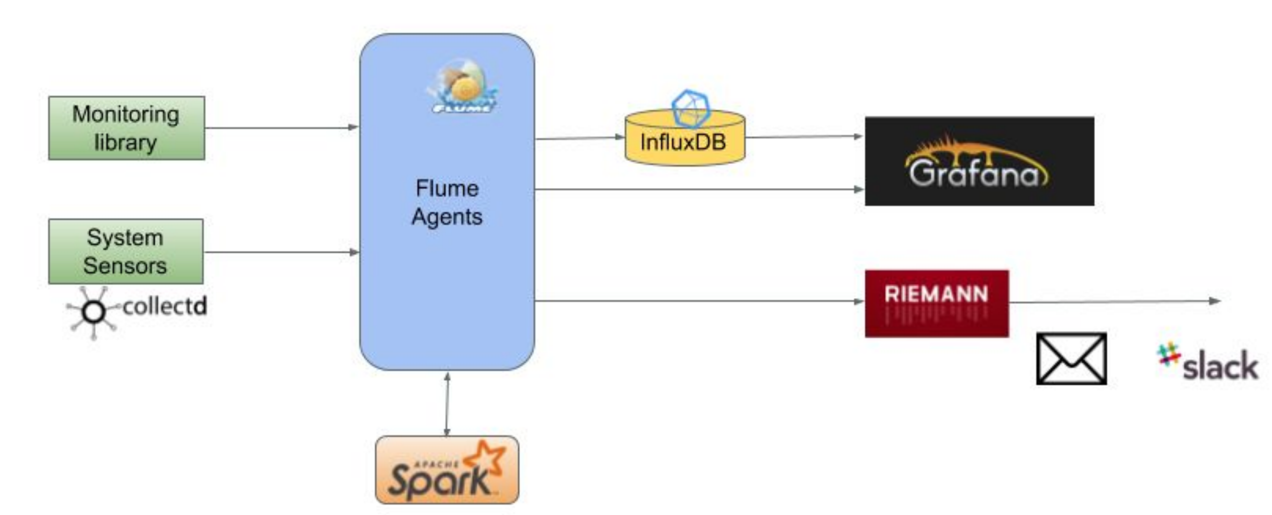
- 268 First Level Processors
- 1500 Event Processing Nodes

Requirements:

- Capable of handling O2 monitoring traffic – 600 kHz
- Scalable >> 600 kHz
- Low latency
- Compatible with CentOS 7
- Open Source, well documented, actively maintained and supported by developers
- Impose low storage size per measurement

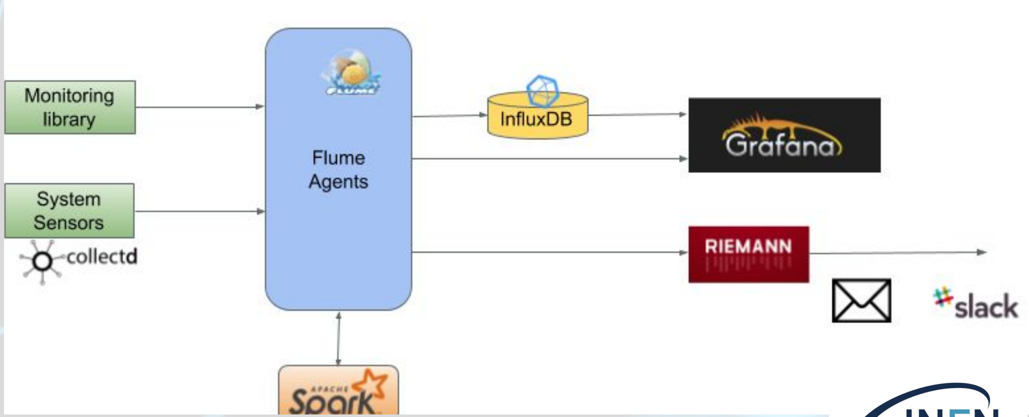
Modular Stack: Architecture

- Sensors:
  - **Monitoring Library**
  - **CollectD**
- Transport Layer:
  - **Apache Flume**
- Time-series Database:
  - **InfluxDB**
- Visualization interface:
  - **Grafana**
- Alarming component:
  - **Riemann**
- Processing component:
  - **Apache Spark**



Modular Stack: Architecture

- Sensors:
  - **Monitoring Library:** user defined metrics, monitoring process metrics
  - **CollectD:** CPU, network, memory, load, uptime, disk, log files,....
- Transport Layer:
  - **Apache Flume**: implemented custom components
- Time-series Database:
  - **InfluxDB**
- Visualization interface:
  - **Grafana:** users, teams, dashboard
- Alarming component:
  - **Riemann**: Slack alarm
- Processing component:
  - **Apache Spark:** aggregation jobs



Modular Stack: Future works

- I System Validation using the TPC monitoring data, May 2018
- New functionalities will be added (new streaming analysis, alarming, log analysis)
- II System Validation using ITS monitoring data, Dec 2018

THANKS FOR

YOUR

ATTENTION

INFN BARI
Istituto Nazionale di Fisica Nucleare
Sezione di Bari