General purpose data streaming platform for log analysis, anomaly detection and security protection

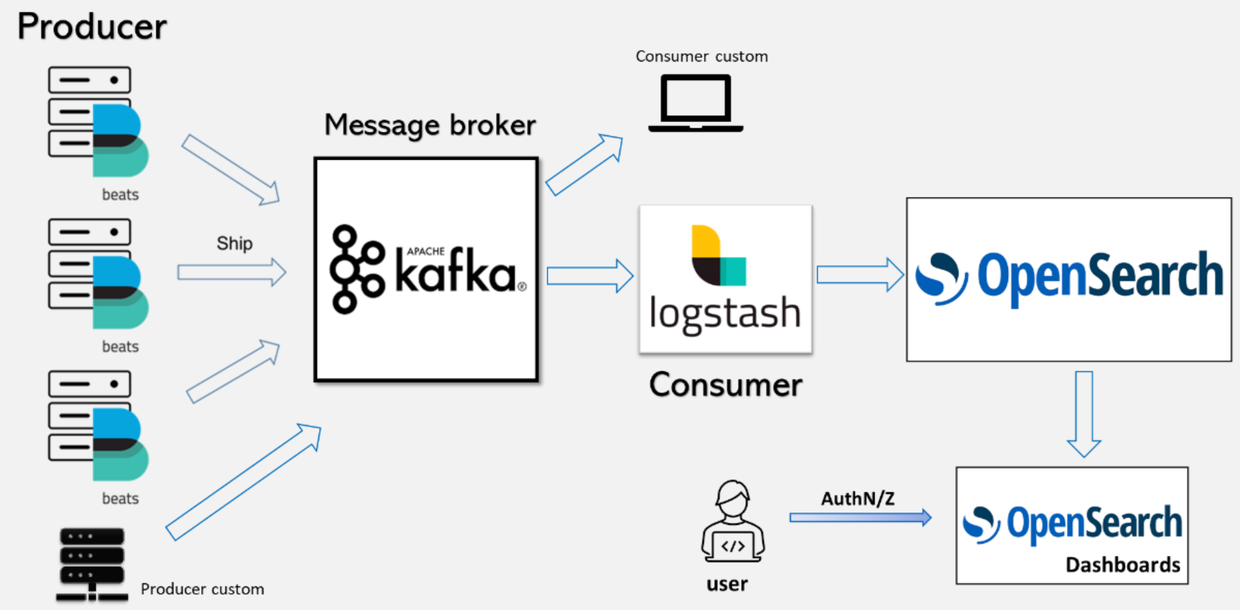
Motivation and design

General purpose platform

- ✓ CNAF is the main technological center of INFN (Italian Institute for Nuclear Physics), hosting computing and data resources for scientific communities
- ✓ Need for a data streaming infrastructure managing heterogeneous data (time series, logs, metrics, etc.) usable for:
  - ✓ Troubleshooting and monitoring
  - ✓ Security protection (threat prevention, detection and response)
  - ✓ Anomaly detection
  - ✓ Machine learning activities
- ✓ Data consumable by CNAF service administrators and multiple users/projects in a multi-tenant environment
- ✓ Log analysis platform integrated

Infrastructure based on open source components

- ✓ Data ingestion: multiple producers, i.e. Beats or custom
- ✓ Kafka message broker to manage topics
- ✓ Multiple consumer, i.e. Logstash or custom
- ✓ OpenSearch (open source fork of Elasticsearch) to search and visualize data



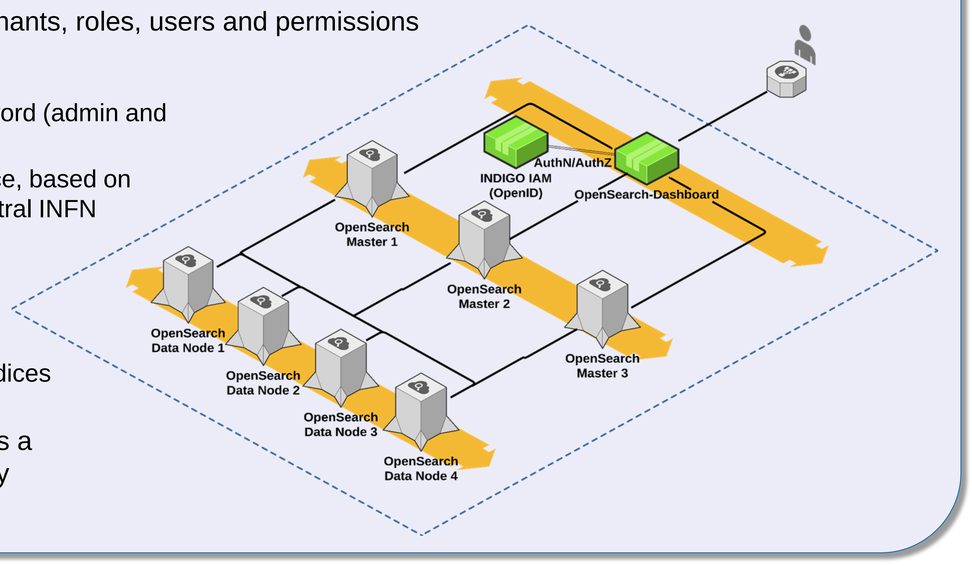
Access rules

Kafka

- ✓ Authorization managed via ACLs (on topics and consumer groups)
- ✓ Admin user can access all topics
- ✓ Each department / project can only write into their own topic

OpenSearch

- ✓ Security plugin to manage tenants, roles, users and permissions
- ✓ Double authentication
  - ✓ Local: username and password (admin and service users)
  - ✓ Through INDIGO IAM service, based on OIDC, connected to the central INFN Authentication service
- ✓ Authorization via IAM groups
  - ✓ Automatic mapping on OpenSearch backend roles
  - ✓ Permissions on particular indices associated to IAM groups
- ✓ Each department / project has a dedicated tenant and can only access their own indices



Production infrastructure

High availability and scalability

- ✓ All the components are deployed in clusters
- ✓ All the components can be scaled horizontally
  - ✓ Number of Kafka brokers
  - ✓ Number of Logstash instances
  - ✓ Number of OpenSearch master / data nodes
- ✓ Storage space can be increased

Data retention and archival

- ✓ Multiple copies of data
- ✓ Retention can be varied depending on the projects requirements and policies
- ✓ Old OpenSearch indices can be put offline via snapshots

Multi-tenancy and data segregation

- ✓ Public, private and department / project tenants
- ✓ Ad-hoc implementation of a quota management system

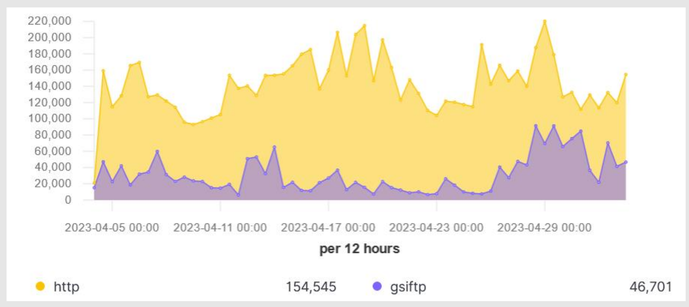
Integration with other services

- ✓ Provisioning platform based on Foreman and Puppet
- ✓ Monitoring based on Sensu, InfluxDB, Grafana

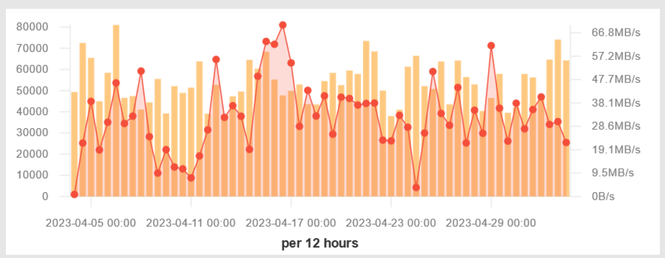
Use cases

A log analysis platform for the storage services

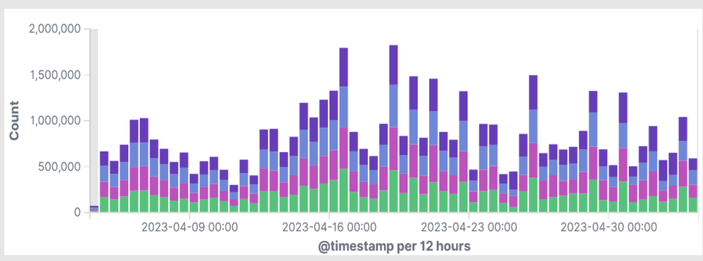
- ✓ Log files are collected from all data transfer and data management services in a continuous live-feed streaming
- ✓ Each record is parsed via appropriate *grok* filters based on regular expressions in Logstash, to generate structured data which can then be searched and visualized within OpenSearch
- ✓ Useful interface for CNAF administrators
  - ✓ powerful search engine to correlate many different log files
  - ✓ insightful visualizations and dashboards in support of day-by-day troubleshooting and operations



Number of files transferred with different protocols, monitoring the ongoing transition from gsiftp to https in WLCG (Worldwide LHC Computing Grid).



Number and throughput of http Third-Party-Copies for the ATLAS experiment.

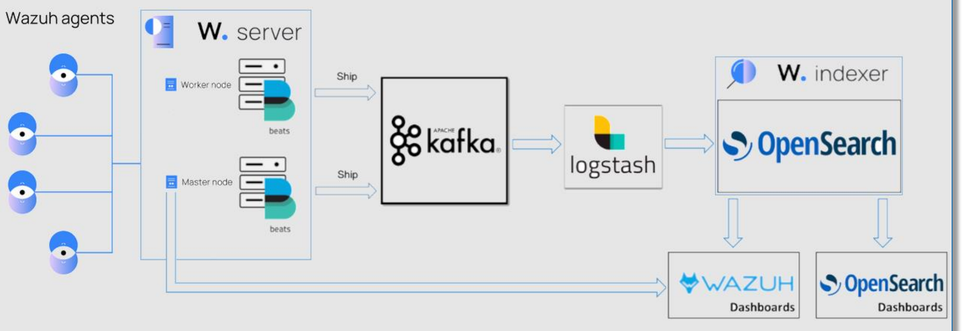


Distribution of the requests of file transfers among the 4 different StoRM WebDAV endpoints of the ATLAS experiment at INFN CNAF data center.

Future developments

Security protection via Wazuh

- ✓ Threat prevention, detection and response
- ✓ Unified XDR (eXtended Detection and Response) and SIEM (Security Information and Event Management) protection
- ✓ Auto-remediation with different modules
  - ✓ Host-based Intrusion Detection System (HIDS)
  - ✓ File Integrity Monitoring (FIM)
  - ✓ OSSEC Active Response
- ✓ Works with OpenSearch
- ✓ Data access and control through a custom dashboard



Anomaly detection

- ✓ Detectors based on data from indices can be used to create alarms

Machine learning

- ✓ Algorithms can be trained based on the log data or from a combination of time series

F. Amori, S. Antonelli, V. Ciaschini, A. Falabella, E. Fattibene, F. Fornari, D. Lattanzio, D. Michelotto, L. Morganti
INFN-CNAF Bologna (Italy)
enrico.fattibene@cnaf.infn.it