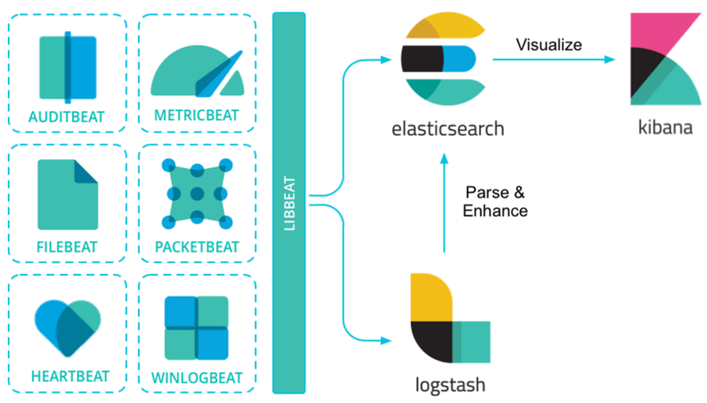
# Enhanced Data Analytics capabilities in the ELK Stack - A review of the premium features and their benefit to a Scientific Compute Facility

Michael Poat, Jerome Lauret
Brookhaven National Laboratory

## Introduction

- A challenge in real-time computing facilities is having adequate monitoring tools providing alerting and anomaly detection beyond the binary events
- Tools such as Ganglia, Grafana, & Nagios are noteworthy, but lack capabilities to ingest & index log messages, and detect irregular issues within.
- The ELK stack (Elasticsearch, Logstash, & Kibana) is the combination of three open-source projects to ingest, search, and visualize logs and data.
- The Basic license of the ELK stack enables you to setup a self-hosted ELK stack and begin monitoring your infra. but has some limitations
- By enabling the ELK premium license, you unlock enterprise features for authorization & authentication, Machine Learning features, and advanced alerting.
- Here we will describe what is the ELK stack, what are its basic and premium features, how it compares to other tools, and how one can benefit a scientific compute environment.

## The ELK Stack




- ELK is a group of opensource products from Elastic that is both distributed and scalable
- Four core components:
  - **Elasticsearch –** Search & Analytics Engine
  - **Logstash** – Log Data Ingestion tool
  - **Kibana** – Visualization & Exploration tool
  - **Beats** – Client Side Data Shipping Agent(s)

## Basic Vs. Premium License

- **Authentication:** LDAP, AD, SSO, OpenID, Kerberos authentication & authorization
  - Authorized “Spaces” allows separated Dashboards for individual groups
- **Authorization:** Enables you to setup 1 ELK stack for your entire enterprise while separating privileges
  - Without A/A you could get around by having an ELK stack per group
- **Machine Learning:** Anomaly detection (single/multi metric, rare analysis, & forecasting), data frame analysis, Model management, & AIOps (log rate spike)
  - Includes built-in ML jobs for beats data
  - Create your own ML jobs for non-standard indexed data
- **Alerts:** Anomaly & detection alerts from ML jobs (standard alerts are included in base license)

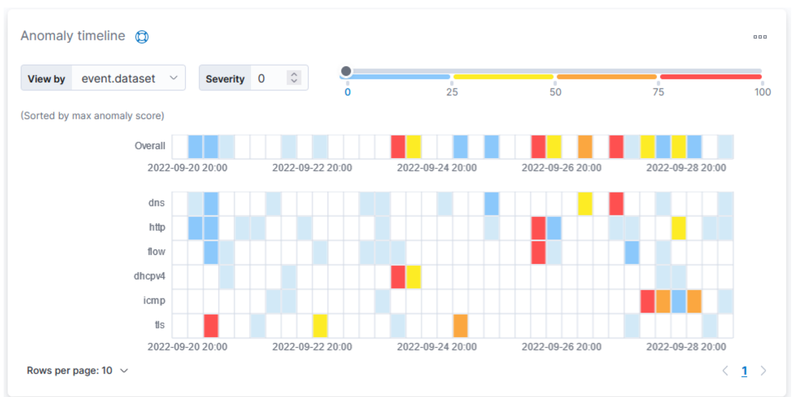
## Comparison of Alternative Tools

- Nagios & Ganglia are simple but lightweight tools that gather system metrics and scale easily
- Grafana has more in-depth features than Nagios & Ganglia but still lacks the ability to consider log context (indexing) and creating graphs from logs
- ELK covers all the listed features but has larger system requirements, to scale, it requires more hardware and added license cost

| Feature | Nagios | Ganglia | Grafana | ELK |
|---|---|---|---|---|
| Lightweight Clients | ☑ | ☑ | ☑ | ☑ |
| Server Side System Requirements | 1-CPU<br>2GB-RAM<br>40GB-HDD | 1-CPU<br>2GB-RAM<br>40GB-HDD | 2-CPU<br>4GB RAM<br>50GB-HDD | 4-CPU<br>16GB RAM (min)<br>64GB-HDD |
| Monitor Metrics | ☑ | ☑ | ☑ | ☑ |
| Visualize Metrics | ☒ | ☑ | ☑ | ☑ |
| Keyword log search | ☒ | ☒ | ☑ | ☑ |
| Considers Log Context | ☒ | ☒ | ☒ | ☑ |
| Creating graphs from context data | ☒ | ☒ | ☒ | ☑ |
| Visualize Logs | ☒ | ☒ | ☑ | ☑ |
| ML Capabilities | ☒ | ☒ | Yes (paid) | Yes (paid) |
| Price | Fee | Free | Free + paid option | Free + paid option |

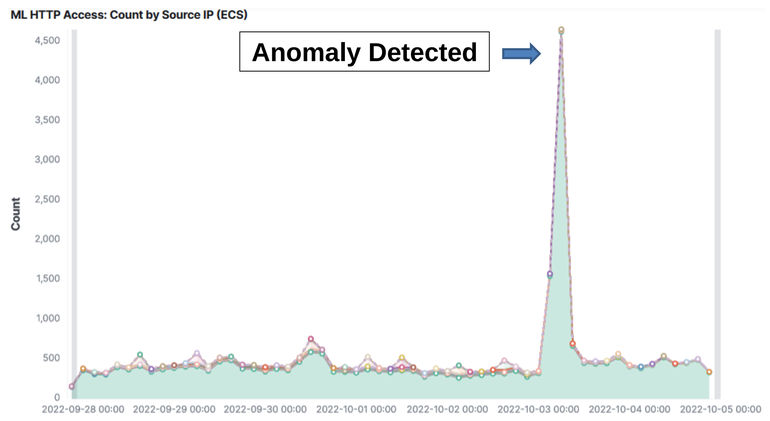
## ELK ML Examples

### Network Traffic Anomaly Timeline

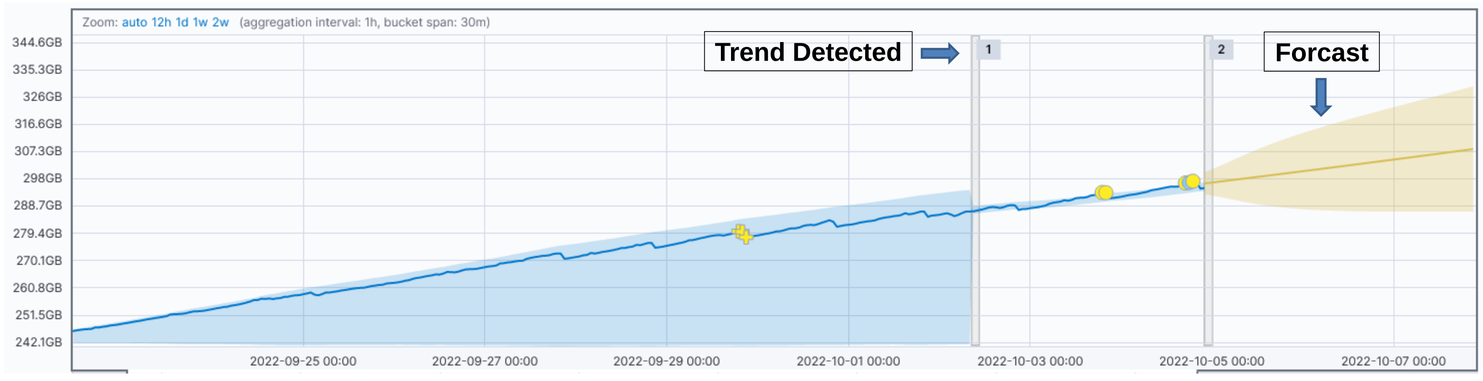



- Timeline showing DNS, HTTP, DHCPv4, ICMP, TLS, and network Flow anomalies
- **Red** squares are high anomalies
- **Blue** squares are low level anomalies

### Total HTTP Traffic Flow Anomaly Chart




- Aggregate chart of global client HTTP traffic to our webserver(s)
- Large spikes are detected as anomalies and can be pinpointed to a specific IP/client & geo-location.




- Total disk usage for a data storage cluster, the yellow portion showing a forecast of the potential usage
- The “1” marker is when ELK has detected a trend and sent an alert.

## Alerts

- The premium ELK license enables anomaly detection alerts from Machine Learning jobs run against the indexed data
- Anomalies are graded by severity; you can choose the severity level which constitutes an alert
- Alerts mechanisms: Email, Teams, Slack, ServiceNow, etc.

## Use Case

- *“A new storage service is implemented but has no monitoring, use ELK to index its log data + use of the ML tools to predict issues & detect anomalies”*
- *“Unusual user activity: User submits 10X jobs than normal + irregular CPU spike + error messages* – would be anomalous and would alert

## Conclusion

- The ELK premium features do offer benefits over the Basic license, but one would need to vet all features for their use case.
- ELK can offer a customized license (SSO auth or ML only) but the cost is based on the memory configuration for the Elasticsearch nodes.
- A simple ELK cluster can have low requirements, but as you scale and ingest more data, the cost of the license increases.
- For Machine Learning, the use case(s) can be very diverse. You can create simple anomaly detection jobs, multi-metric jobs, detect unusual activity, and rate values in time series data.
- Depending on the data you collect, what you want to visualize and inspect for oddities, then ELK with a premium license would be a benefit to you.

Brookhaven National Laboratory
U.S. DEPARTMENT OF ENERGY | Office of Science