Elastic & Fermilab

Joe Boyd, Tanya Levshina, Kevin Retzke
Landscape

- Provides a common monitoring and accounting platforms for Fermilab Scientific Computing Division services and users of HEP Cloud facilities.
- Exists for the last 4 years and is used by more than 450 users from 25 experiments, including NOvA, MicroBooNE, Mu2e, G-2, DUNE, DES and CMS.
- Allows:
  - monitor physics jobs and data transfers
  - monitor services
  - estimate resource usage and plan for future needs
- Delivers a comprehensive framework to monitor:
  - services health, resource utilization, jobs and data movement at HEP Cloud Facility
  - user-friendly dashboards, alarms, reports and several other important services.
High Level Architecture

- 13 billion documents in Elasticsearch
- using 20TiB without replication.
- Elasticsearch production installation:
  - 19 worker nodes
  - 5 nodes dedicated as masters and clients.
ML Usage Objectives

- **Improve users’ experience and decrease waste of resources:** Allowing us to detect “blackhole” nodes on time will prevent failure of users’ jobs and avoid data transfer to deficient nodes.
- **Enhance non-compliant resource usage detection:** Detection of abnormal resource utilisation will allow us to identify rogue jobs with little i/o and high CPU utilization.
- **Improve hardware acquisition planning:** Utilization of ML to improve the ability to identify trends and predict future resource requirements.
- **Smart resource utilization on HPC:** Improving prediction of resource utilization required by a particular workflow will allow to automatically find an appropriate HPC resource that is included in Fermilab HEP Cloud Facility.
Why Elastic ML Plugin?

- The usage of the native Elastic ML plugin allows to perform analysis with single install (X-Pack) and to mine data from the same cluster.
- We could rely on Elastic to provide scalability and fault tolerance by scheduling “ML job” that are distributed across ES nodes. ES is handling node failures as well.
Potential Benefits for Elastic

We have a reasonably big Elasticsearch cluster, huge amount of historical data and a lot of interesting use cases. Access to historic data and use cases will allow Elastic to improve ML plugin performance and create new visualization features.