



Monitoring AI Platform

Deepak Vasthimal – MTS @ eBay

LinkedIn - <https://www.linkedin.com/in/whatisdeepak>

eBay is Hiring globally.

Overview of AI Platform

- Accelerated Training: High performance compute cluster powered by Nvidia GPUs and High-memory (up-to 1TB) bare metal nodes.
- Easy Access to Data: Data stored in Hadoop cluster, NFS volumes, proprietary storage easily and securely.
- Automate learning at scale : Scale and automate your training with learning pipelines. Use any language or framework (Caffe, Tensorflow, Keras, etc.) of your choice!
- Multi-Tenant architecture.
- eBay investing into Machine learning and deep learning. Use Cases: computer vision, recommendations, predictions, machine translation, etc. for model training and batch inferencing.

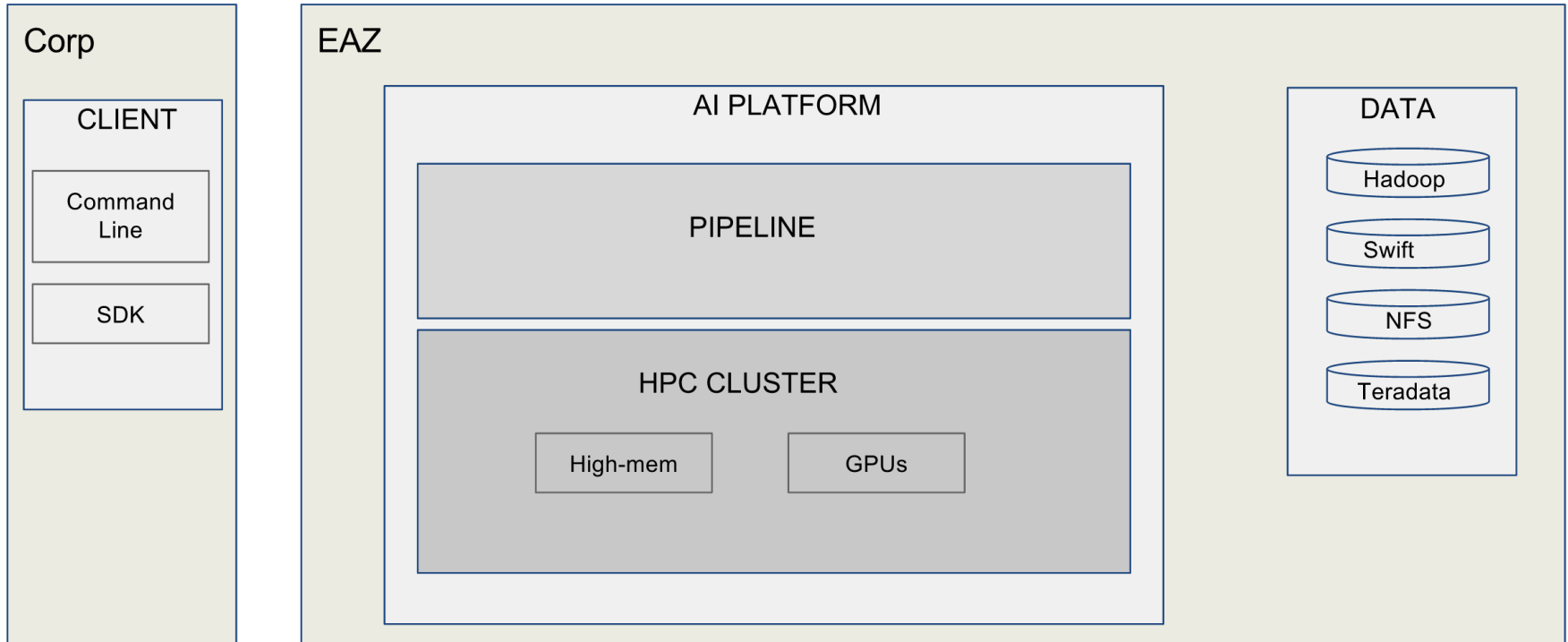
AI Platform

- 100+ of GPU Nodes with 4 GPU cards each (512Gig Memory).
- 100+ high memory bare metal nodes with 1TB memory.
- 100+ training jobs running daily.
- Command line utility to manage deep learning jobs written in Go.
- Create a project, test locally, submit to remote cluster, automate with service account.
- All external dependencies, frameworks, libraries are managed by Docker containers.

Working with pipelines

- Create project – A workflow with tasks. Local workspace to develop and customize ML application.
- Define Workflow and its task(s) through JSON.
- Create Artifact – Definition of application code and dependencies to run ML application on remote HPC cluster.
- Submit, Pause, Resume and Stop Job anytime during execution.

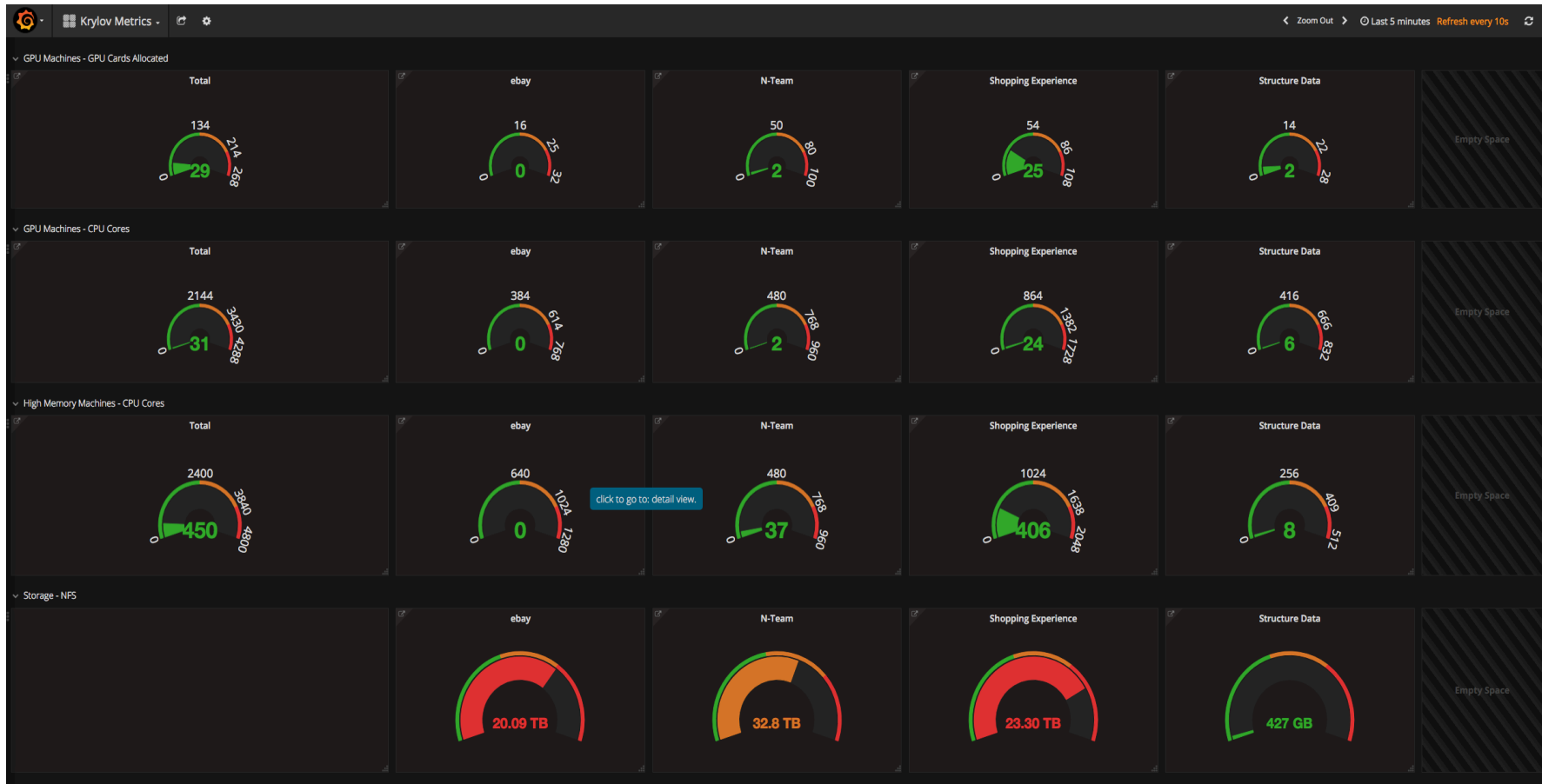
Architecture – Bird's eYe



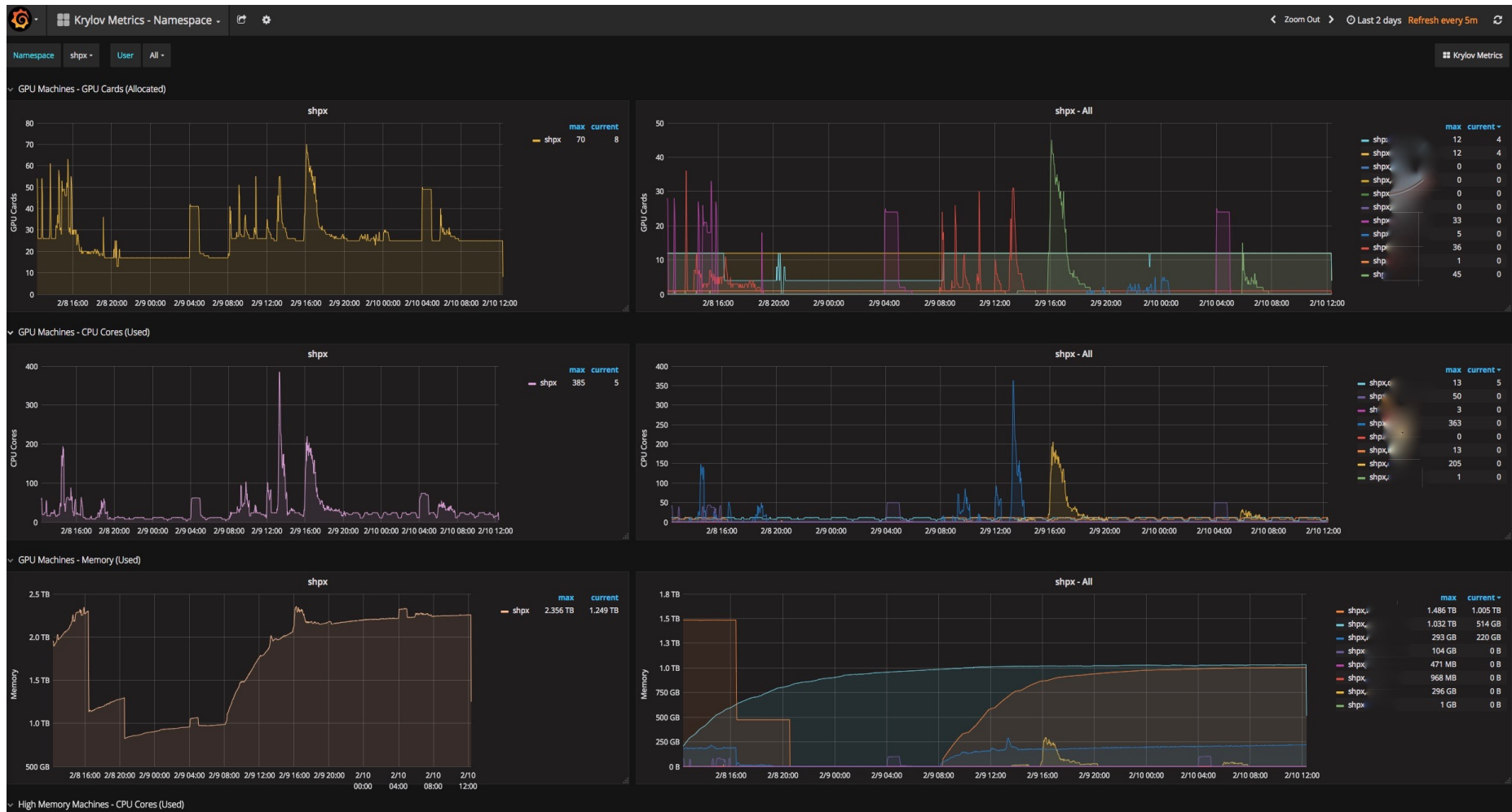
Monitoring Resources

- Monitor CPU, GPU, Memory usage of High Memory and GPU bare metal nodes.
- Track resource allocation & consumption at org, user and job level.
- Monitor 200+ Nodes for disk, memory and other system resources.
- Alerts – Zookeeper, Couchbase, Nodes DiskUsage, Availability.

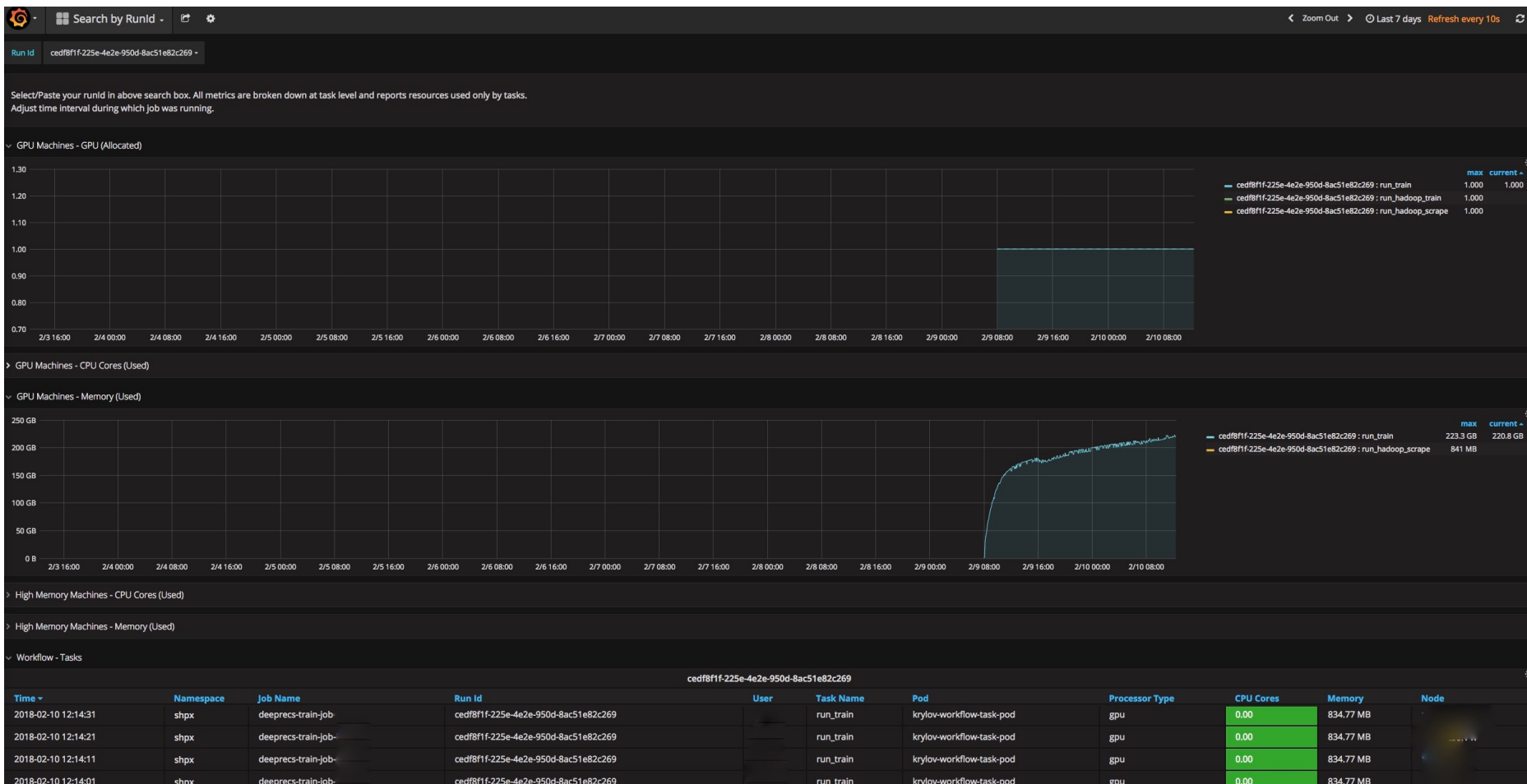
Home Page



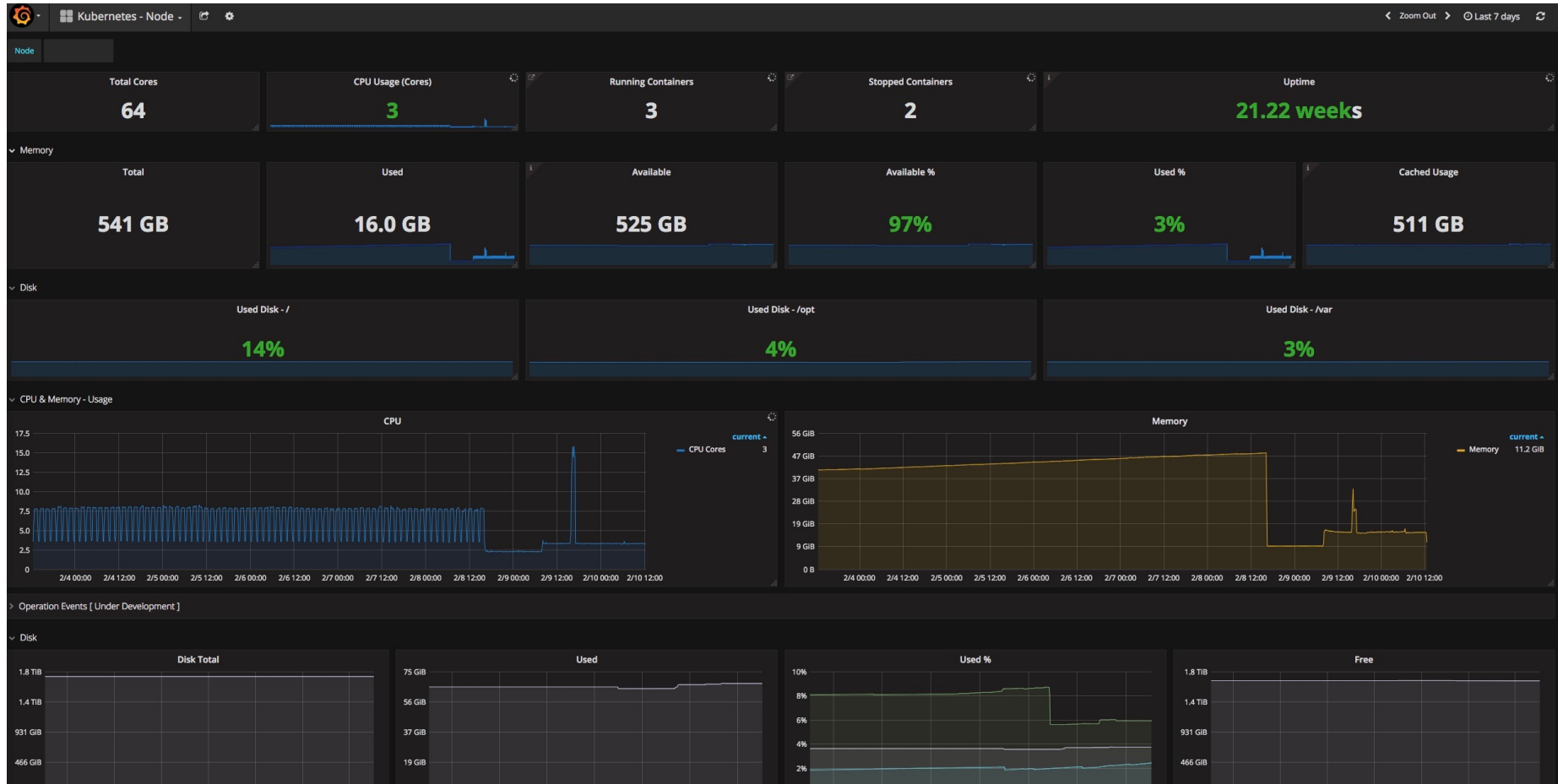
User Level – Resource Usage



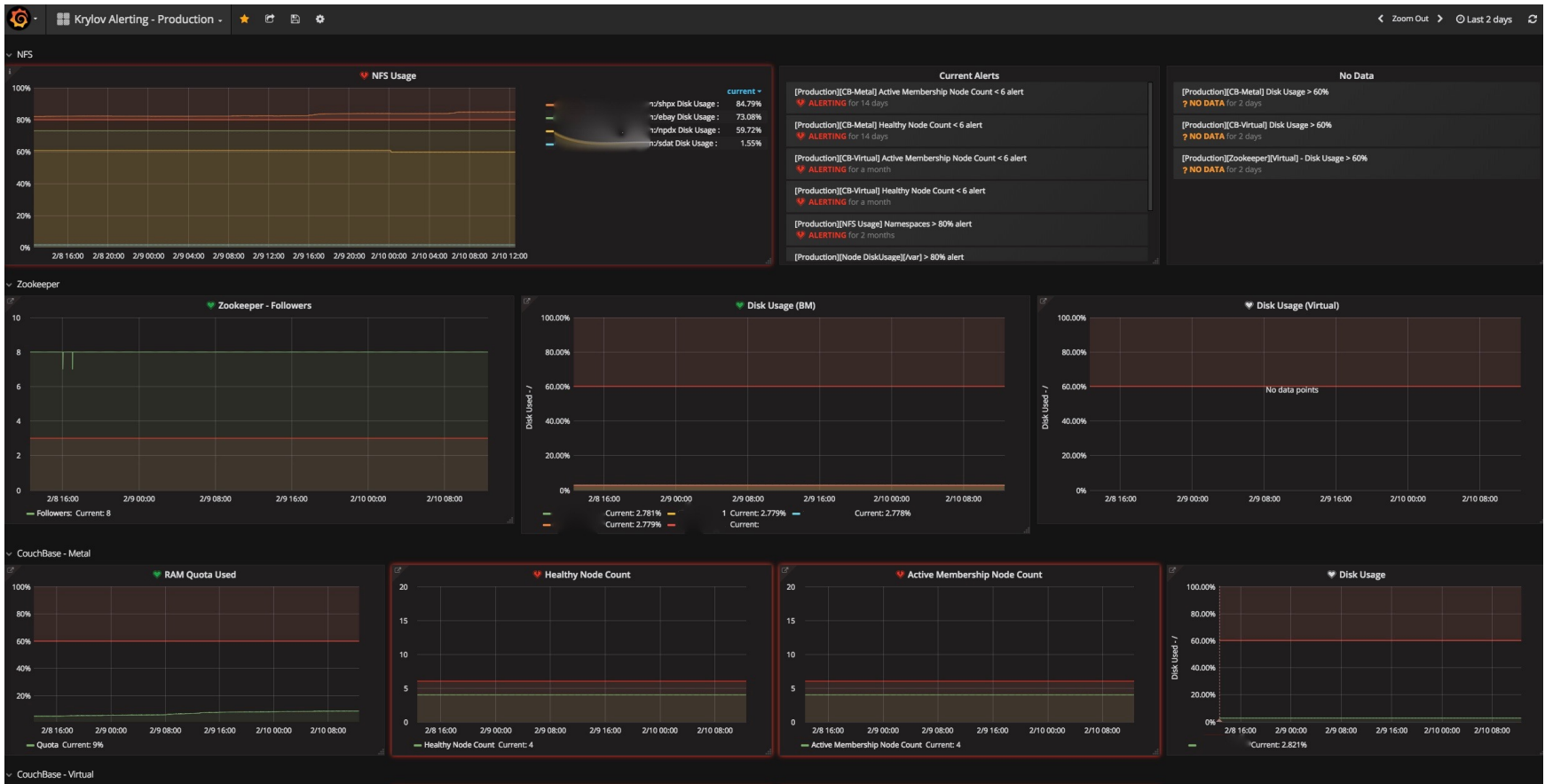
Search



Node - Metrics



Alerts



• Couchbase, Zookeeper, NFS, Nodes - Alerts

Scale

- Push metrics into influx via Telegraf running on 200+ nodes.
- InfluxDB and Grafana are installed as a service on Kubernetes cluster with 64 cores and 1TB memory.
- 100+ measurements each with 10-100s of metrics and dimensions.
- > 10 Million series.
- ~10k points collected every 10 seconds.
- Node and Job metrics are collected every 10 seconds.
- Rolling up : 10 second data with 1 month RP (Retention Policy), 1 min with 6 months, 10 mins with 3 years RP.

Grafana <3 is spreading.

- Performance analysis of MapReduce jobs by Experimentation platform with job counters.
- Monitor Elastic Search Cluster (10k+ points per second).
- Anomaly detection in Tracking data (10k+ points per minute).
- Monitoring Anomalies with Experimentation platform.
- Each use case stores data in InfluxDB.

Thank You