Panel: Coordinating and Combining Data Processing, Movements, and Storage

David Mencin
Vice President
Data Services

Robert Casey
Software Section Manager

CI4MF 2024
Cloud Migration

- System foundation (CI/CD)
- Demonstration services
Decommission Boulder DC
GAGE services in the cloud, hybrid mode
Robust cloud GAGE data flow
SAGE system migration
SAGE archive complete

ShakeAlert Operations
Jan 2022
Mar 2023
Aug 2023
Sep 2023
Nov 2023
Feb 2024
June 2024

FULLY OPERATIONAL

NOW
Cloud Lift: Challenges

- Cost projection and cost control (FinOps) is very different from on-premise operations
- Where to find cost savings: contract deals, open data qualification
- Egress is now our major operational cost consideration
- Identity management and data usage tracking requirements add complexity, constraints on design, and additional starting cost
- On-premise software and systems are not cost effective as-is
- Attempting to avoid vendor lock-in whenever possible (might be more of a fear-of-change than real)
- Cloud migration project is concurrent with a corporate merger.
Cloud Lift: Messaging

Managing the perception that “the sky is falling!!”

*There are no plans to deprecate current capabilities and data formats.*

*The only significant change that impacts everyone, users and staff alike:*

*Identity Management*
**Cloud Lift: Egress of Data**

- **Data egress** to the internet costs roughly **$0.09/Gb**
- We currently export ~***1Pb per year*** = **$110,400.00**

- Can we transition power users to access data while **co-located in the same cloud region?** (little to no egress cost)

- Perhaps we train our users to do their work in AWS (On-Ramp!)
  - cloud training
  - account subsidies
  - limiting internet download in favor of local clusters
  - providing notebook resources on AWS

- Some data may qualify for Open Access at no cost
  - …but we might not be able to track its usage
Advances in Data Ingress
Ingress Processing - GNSS

- Bring data into Kafka
- Stateless: read bytes, send bytes
- Handles TCP and RabbitMQ feeds
- All config managed by API
  - Passed through Kafka
Many cloud systems are monitored and can be displayed in accessible dashboards.
Observability: Data Egress

We can track the geo-location of clients to chart demographics of usage.
GNSS processing center can be within the EarthScope cloud or external.

Dashboards
- Position time series
- Station maps
- Latency reports
- QC status
- Perform queries

Long-term data storage
- (raw, positions, QC)

Web service
- Retrieve
- Raw data
- Position data
- Latency & QC
- mseed/tank/geoJSON

Network Operator
- NTRIP
  - raw

GNSS Processing Center
- positions
- Kafka Connector

DATA HUB
- Kafka Connector
- QC analysis

SHAKEALERT DATA CENTER
- Kafka Connector
- ShakeAlert Analysis Center

Web service
Observability: ShakeAlert

Maps to view data spatially

Heat maps - easy to see the most common latencies, ~250ms & ~1 sec
Thank you!