MetrANOVA

Updates and a demo of an interim federated analysis environment

Ed Balas, Andy Lake, David Ripley
December 11th 2024

ESnet GÉANT INDIANA UNIVERSITY INTERNET2 TACC

©2024 MetrANOVA Contributors • Licensed CC BY-SA 4.0 • https://www.metranova.org

MetrANOVA was created to…

1. Advocate for quality ubiquitous collections with appropriate access within all of R&E
    a. Provide training and policy guidance
    b. Create knowledge base articles and howtos
2. Lower the barriers through technical and policy collaboration
    a. Reduce need for bespoke solutions
    b. Amortize software sustainment costs through collaboration.
3. Retain Network Measurement as a core competency
    a. Requires ongoing care and feeding
    b. Deep domain knowledge in networking, systems, and to an extent stats
    c. Support next generation of R&E engineers

The Secret Sauce of Research and Education

- **Timeless design constructs**
  - Ubiquitous Access
  - Loose Coupling
  - Vendor Neutrality
  - Open Standards
  - Rough consensus and working code
- **Combined with community focus**
  - Our values differentiate us
    - not for profit
  - technologies applied to facilitate scientific and educational endeavors.

ESnet GÉANT INDIANA UNIVERSITY INTERNET2 TACC

Consortium Details

Goals

- Tools, Tactics and Techniques
- Develop and Share
  - Open Architectures
  - Technical Components
  - Design Patterns
  - Best Practices
  - Policy Recommendations.

Executive Committee

- Provides governance and oversight.
- Decides on new membership organizations.
- Representatives from each member.
  - Inder Monga - ESNet
  - Ivana Golub - PSNC/GÉANT
  - James Deaton - Internet2
  - Luke Fowler - Indiana University GlobalNOC
  - Nathaniel Mendoza - TACC
  - Ed Balas - Consortium Lead

Vision

- https://github.com/MetrANOVA/.github/blob/main/profile/vision.md
- A collaboratively developed ecosystem exists
- Open Source, loosely coupled, without cloud service dependence
- Solid foundation for production services and innovation
- Facilitate data driven design in engineering and operations

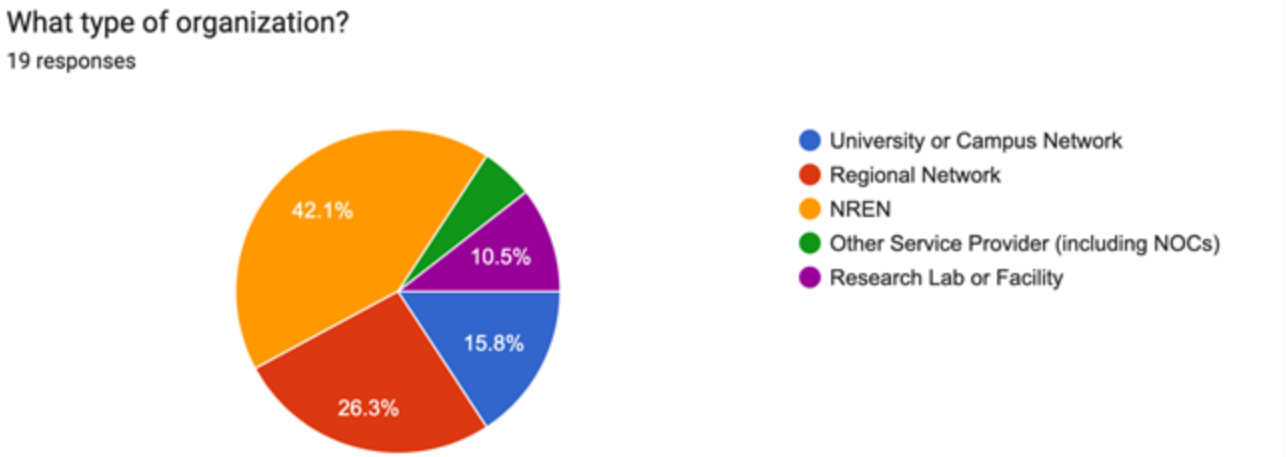
Participation Model

- Member Organizations
  - Requires >= 1 Full Time Staff Equivalent
  - Participates in governance process
- Affiliates
  - Any organization or individual able to contribute.
  - Lower bar to participate, more flexibility

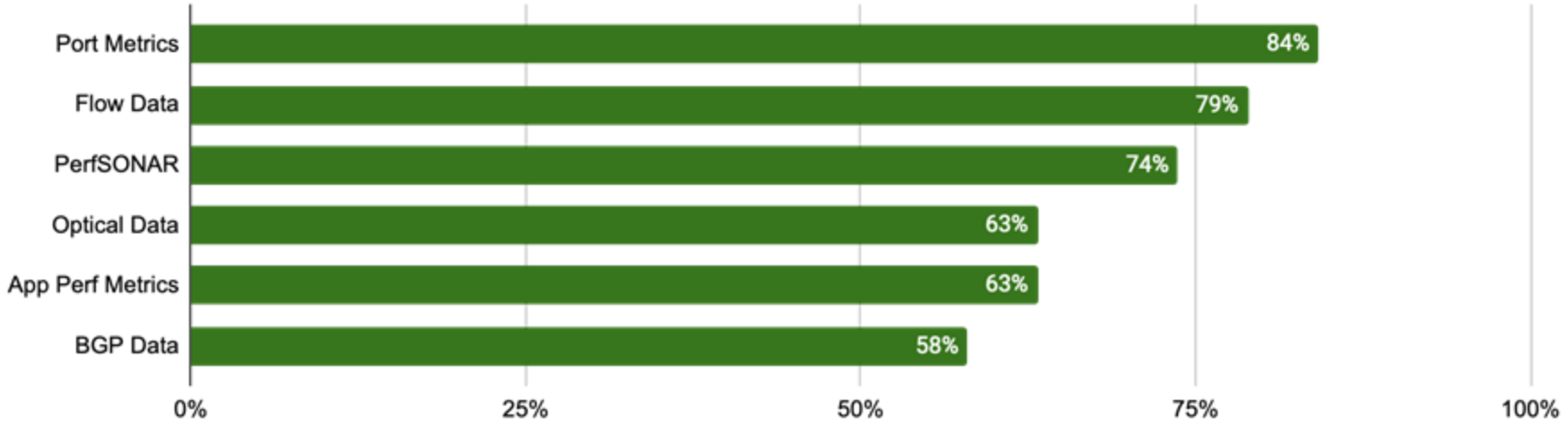
What you told us in 2024

2024 State of Community Survey

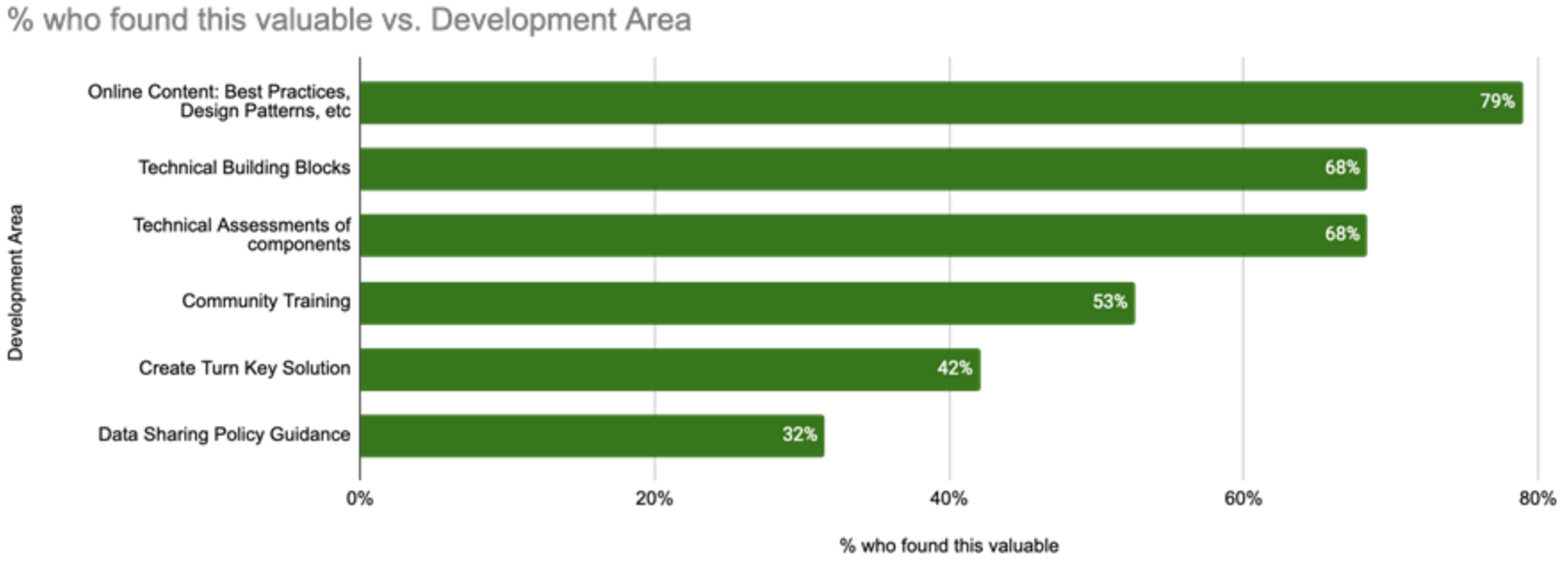
- 19 responses from 18 organizations
- 42% NREN, 26% Regional, 16% Campus, 10% Lab or Facility
- Majority identify as Network Engineers, 36% as leaders and 36% as Syseng



Survey: Data in use



Survey: What people need



What we have been up to

- Two work teams
  - Technical: Lead by Andy Lake
  - Policy: Lead by David Ripley
- Technical team has been building a preliminary technical stack
  - Low risk, based on past experience
  - Proving ground of sorts
  - Provide value quickly
- Policy team has been exploring data sharing models and policy constructs
  - Develop conceptual model for sharing based on what has worked within community
  - Guidance for the Technical team wrt Data Locality, RBAC and Authentication considerations

Designing the MetrANOVA interim Stack

- Let's focus on one aspect of the problem
  - **"Composite / End to End views of the R&E infrastructure will give stakeholders appropriate awareness"**
- Start with some initial questions:
  - **User Perspective:** Can I see my network traffic across multiple networks?
  - **Provider Perspective:** To answer the user's question, can I provide this information in a way that is within my institutions policy constraints, consistent with my peers and requires a minimal amount of work?
- What building blocks to we need to answer those questions
  - **Data:** We need enough information to identify per organization traffic i.e Flow Data but at a granularity that is closer to the **TLP:GREEN**
  - **Software:** Something to not only collect the data, but advertise its existence and display the results

Data Format

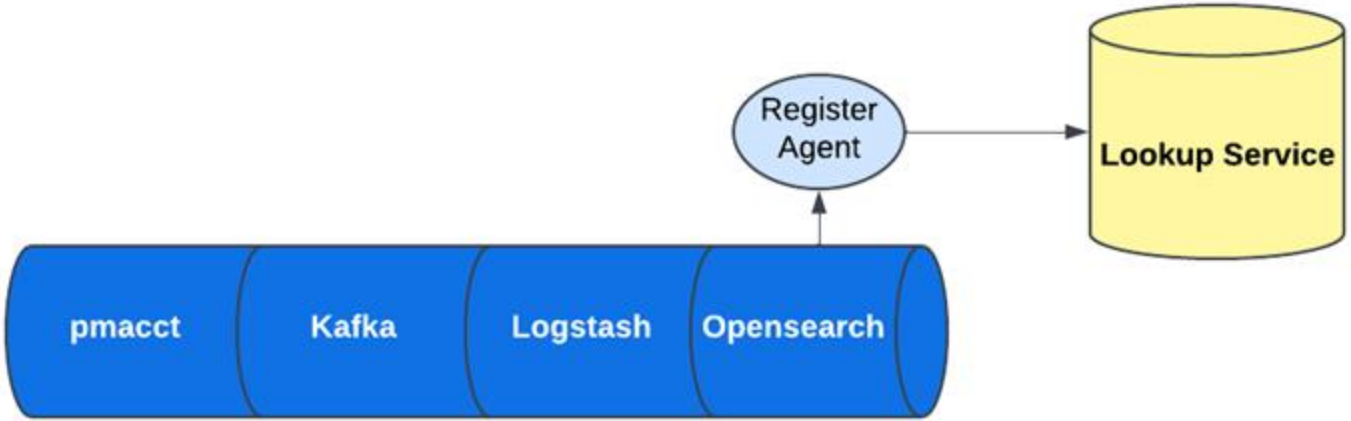
We're starting with Netflow/IPFIX Data but want it at a granularity that could be considered **TLP:GREEN**

- *Time Bucket Size*: 5 minutes
- *Fields*
  - Source AS number
  - Destination AS number
  - Protocol (e.g. TCP, UDP)
  - Application port (well-known src/dest ports kept, ephemeral/other zeroed-out)
  - Peer AS number
  - IP of the reporting router

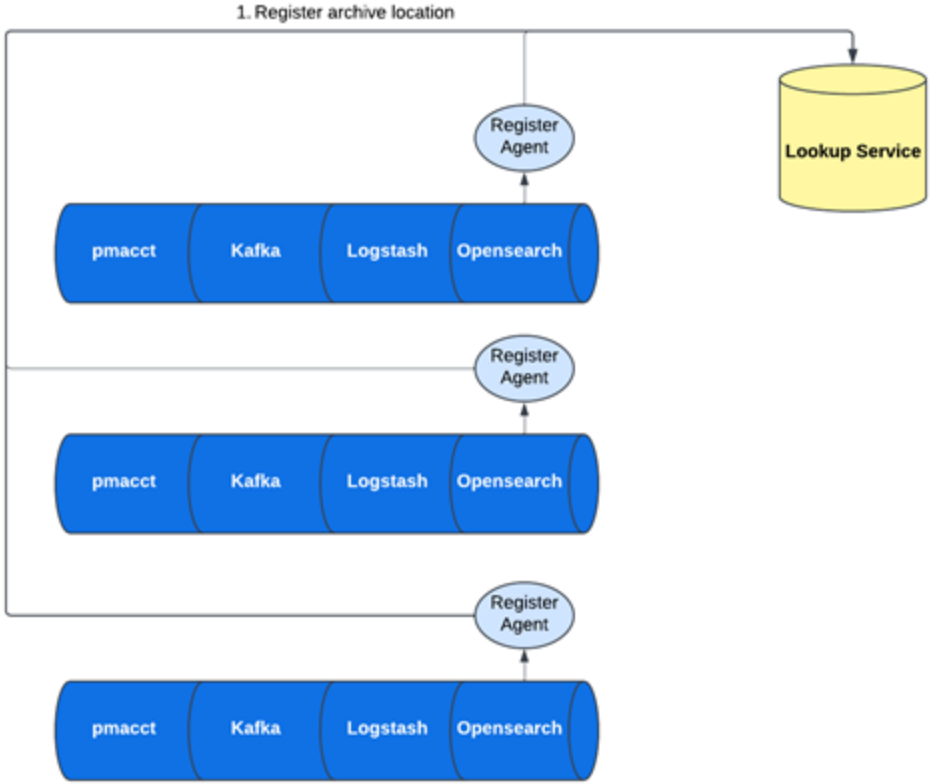
Building a Stack: The Pipeline

- **pmacct**: Collects flow data from routers and aggregates to ASN level
- **Kafka:** Message bus
- **Logstash:** Format message and add metadata
- **Opensearch:** Store data

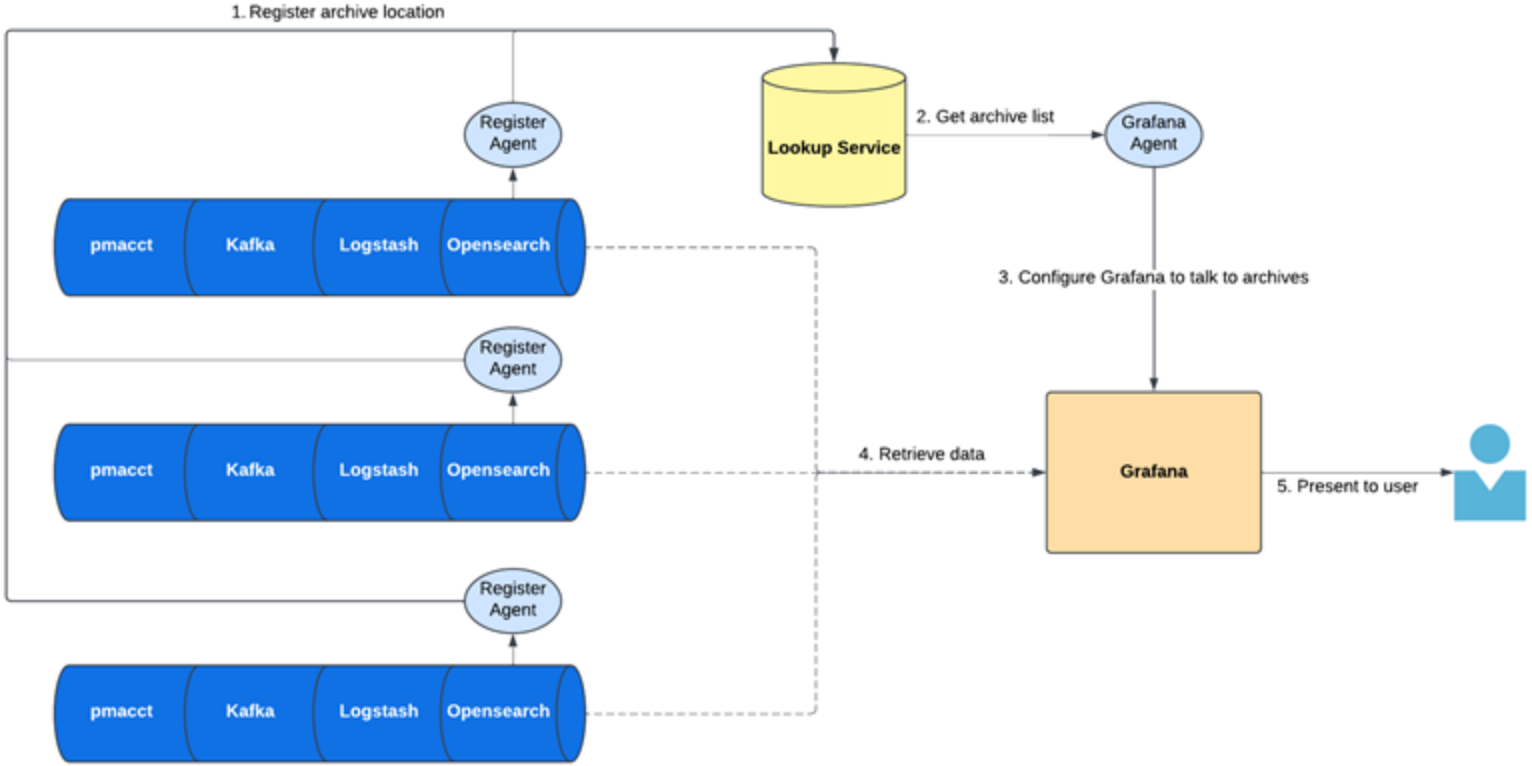
Building a Stack: Discovery



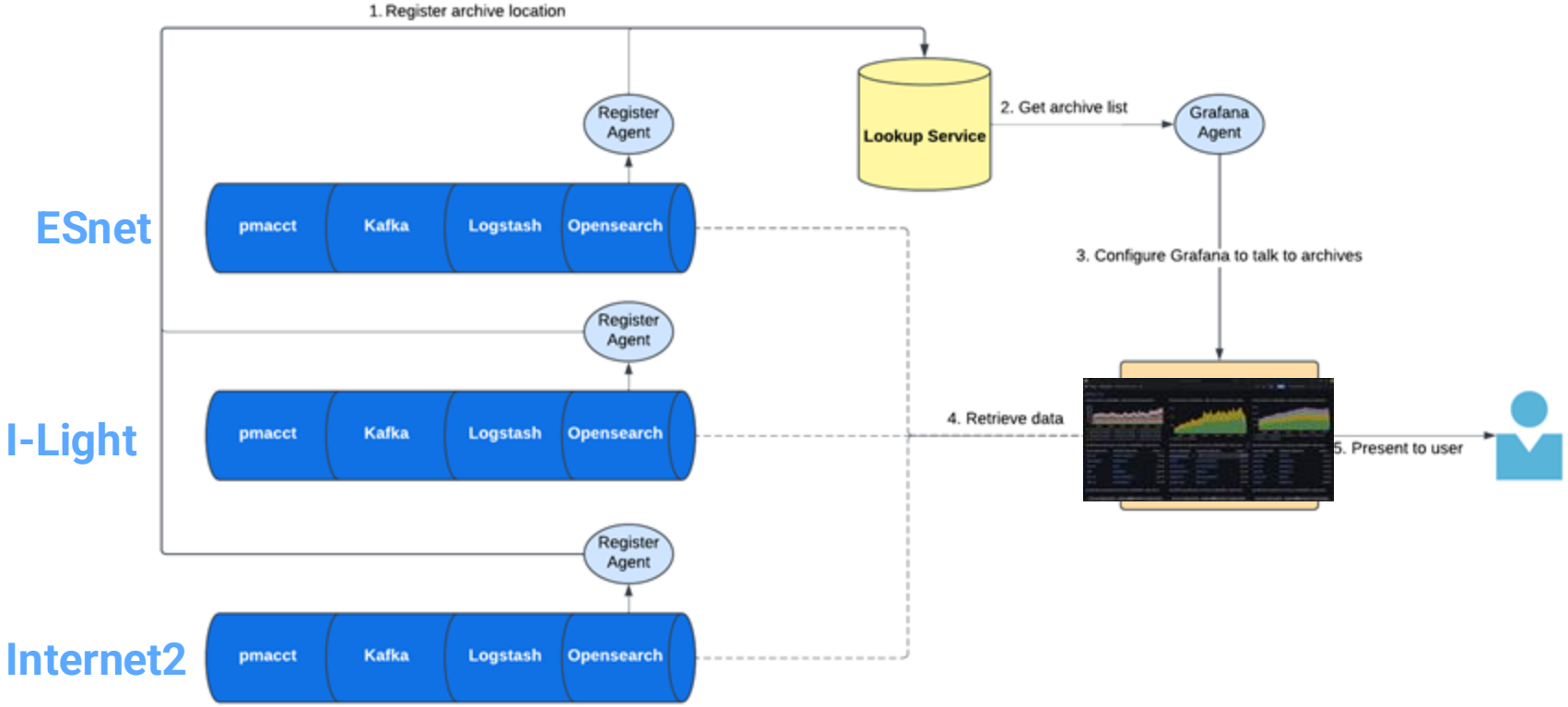
Building a Stack: Multiple Pipelines



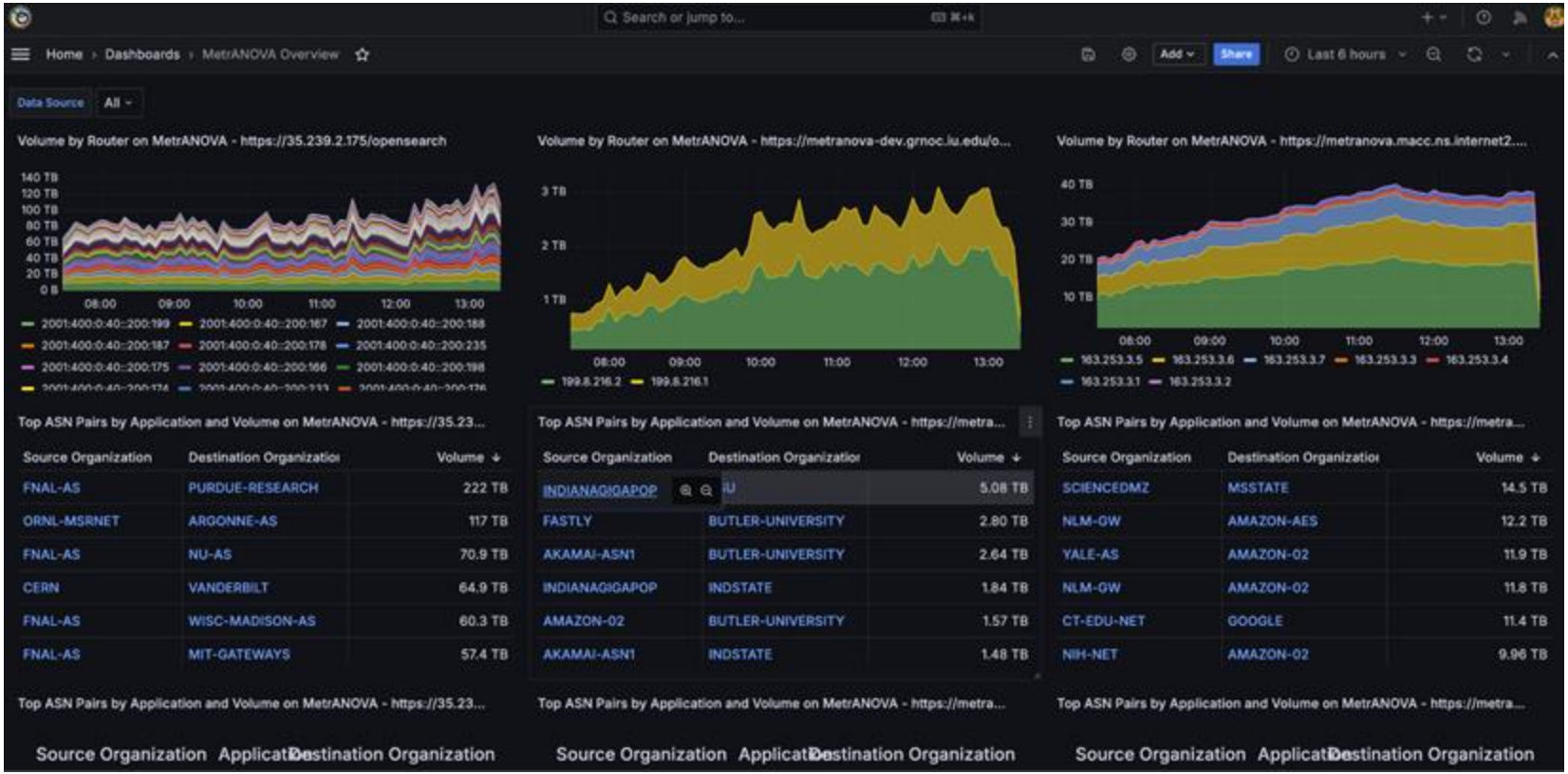
Building The Stack: Bringing it all together



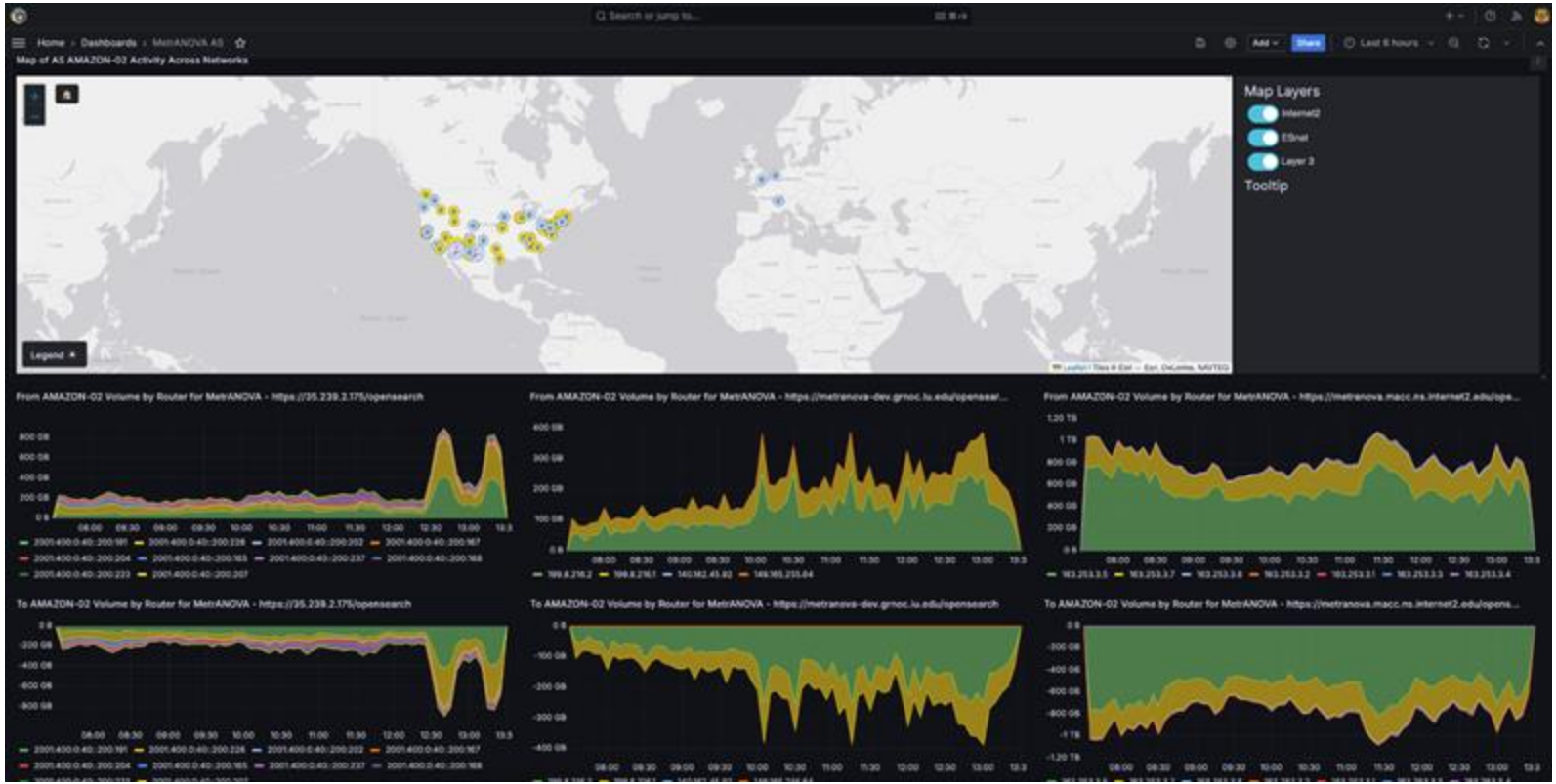
Building The Stack: The Demo



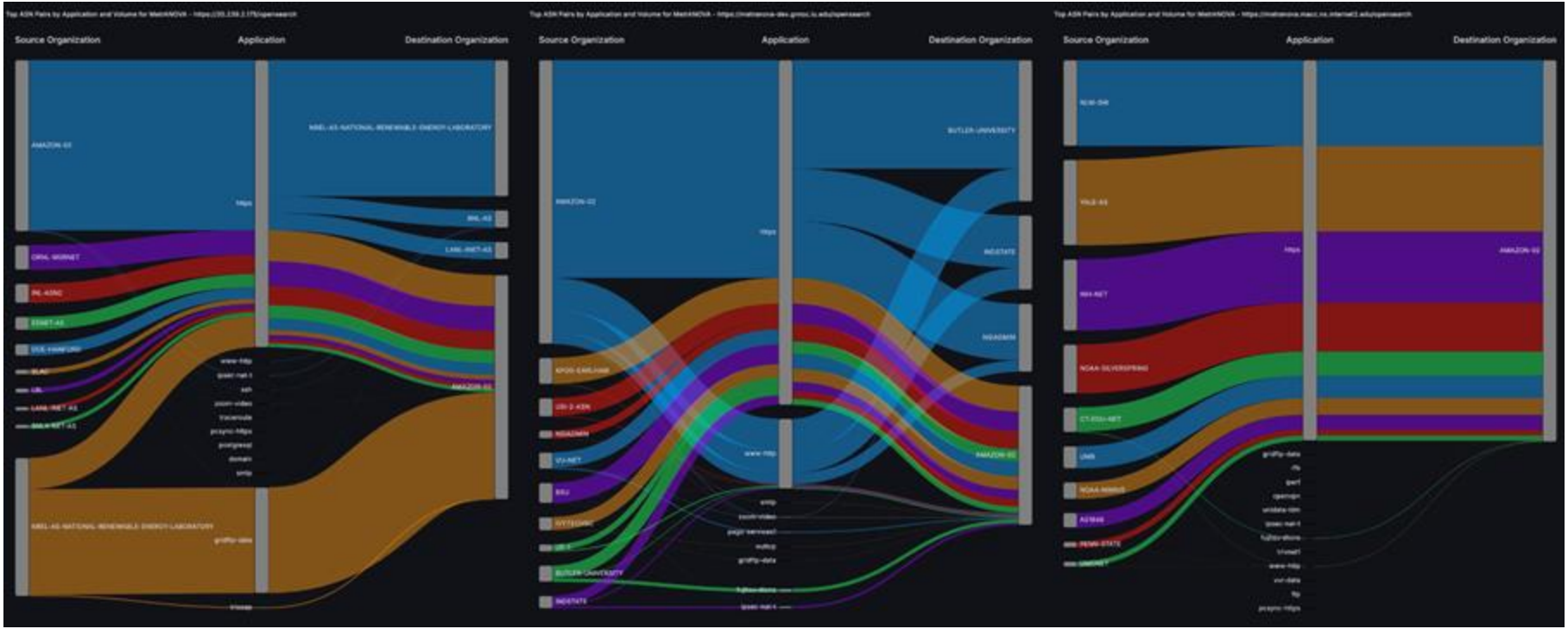
Demo: Who are the top talkers on each network?



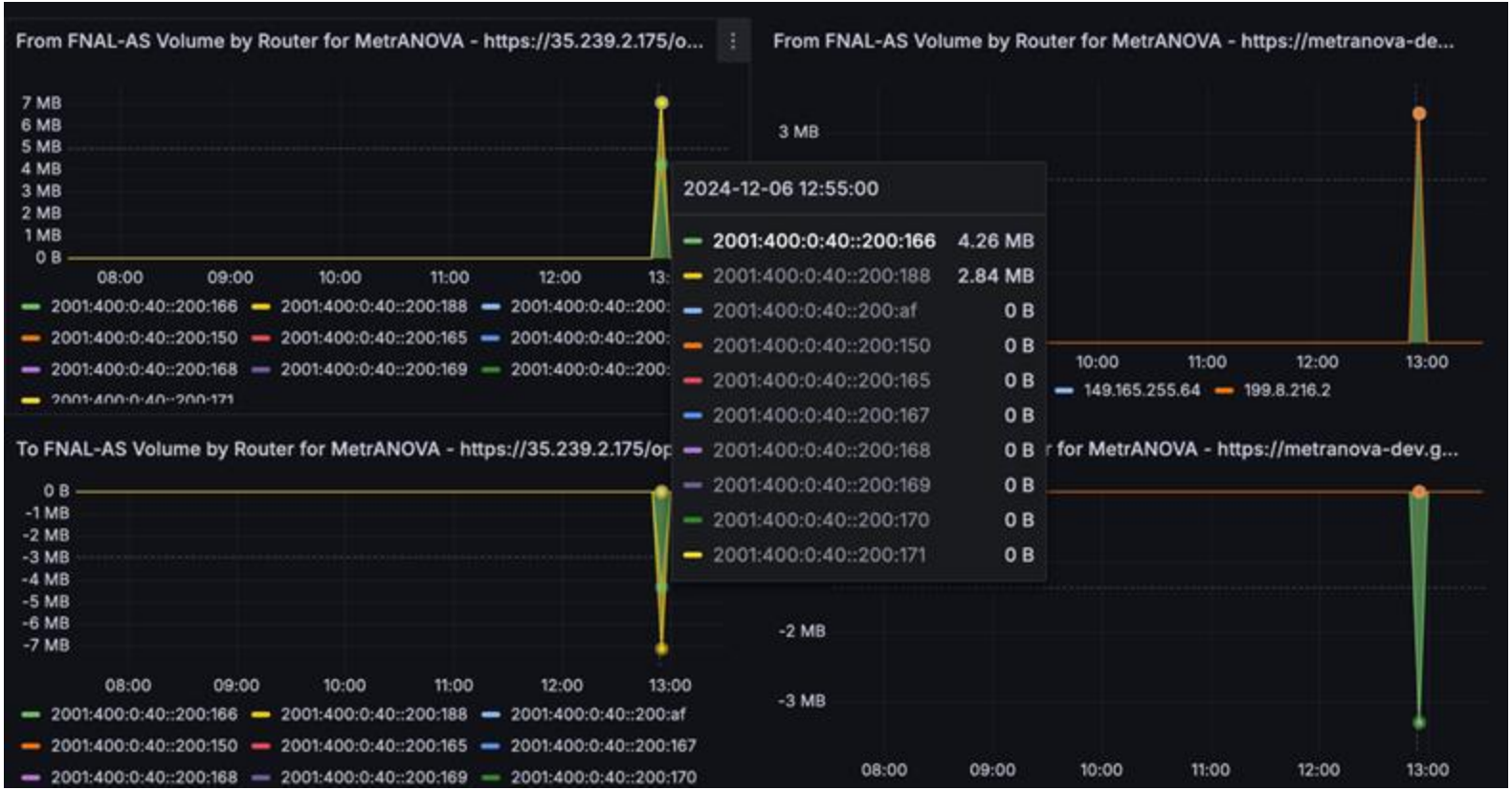
Demo: What is Amazon traffic like on each network?



Demo: Who is using Amazon on each network?



Demo: Was an event observed on multiple networks?



How do I actually run this?

This is not production software yet.

See GitHub project using QR code to right

Basic Process:

1. Set hostname, OpenSearch admin password and logstash password in .env file
2. Bring up stack with Docker Compose: `docker-compose --profile collector --profile archive up -d`
3. Send flow data from router to your collector.

All data stored locally, but will register URL where aggregated data can be queried.

https://github.com/MetrANOVA/techex-demo

Next Steps with MetrANOVA Technical Stack

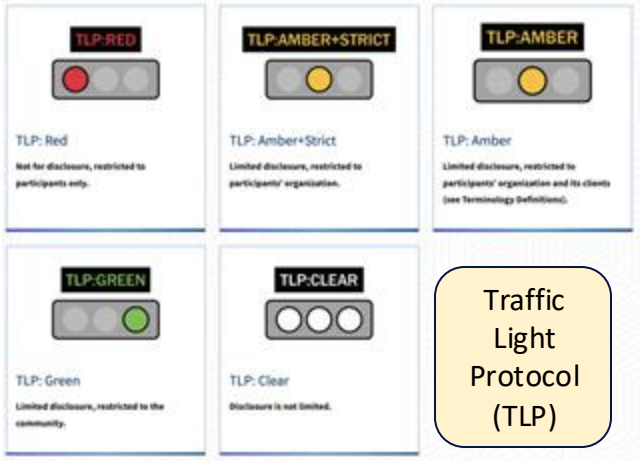
- This is not production software yet.
- Getting this ready for wider adoption is ultimate goal
- Datasets
  - What other datasets should we explore?
- Hardening stack
  - New perfSONAR Lookup Service (on which this relies) to be released soon
  - Datastore evaluation - is this the right stack
  - Better understand minimum requirements
  - Documentation

Data Sharing, Federation, Anonymization, Policy

- Federated services rely on data from multiple systems and domains
- Appropriate controls that respect each domains policies and constraints are a must for data sharing
- Having well defined policies is a precursor which today does not always exist
- Example of constraints you are facing:
  - GPDR, FERPA, HIPAA.
  - NDAs and customers who which to remain low profile
  - Institutional policies, funding bodies, etc.

Different kinds of Sharing

- Different data products have different levels of sensitivity
  - Raw measurements
  - API access to measurement repository with query language
  - Access to online dashboards and reports
- Clear policies helps set expectations
  - What is and what is not shared and at what level of access
- An example:
  - perfSONAR measurement dashboards
    - Performance monitoring data
    - Geopositioning data
    - Display in a single pane of glass.
    - Less policy constraints what is collected and what is shared



Many Thanks!

Edward Balas

MetrANOVA Consortium Lead

ebalas@es.net

For more information:

- Github: https://github.com/MetrANOVA
- Web: http://www.metranova.org/

Questions?

ESnet GÉANT INDIANA UNIVERSITY INTERNET2 TACC