IAB NEMOPS Position paper

Projecting Data Mesh to Model-driven Telemetry: A Path to Data Ecosystem's Management Operations

by Parisa Foroughi and Laurent Ciavaglia, Nokia

I. Problem statement: Data Silos and Knowledge Gaps in Network Monitoring and Automation

- Heterogeneity: Network monitoring relies on a diverse range of data models, protocols, and schemas, often defined in isolated silos. This complexity is only understood by network engineers with specialized domain knowledge. [RFC 9232]
- Data Without Insight: Simply collecting data without connecting, processing, and understanding it is counterproductive.
- Bridging the Gap: To enable advanced automation in network management, AI scientists need to understand the data. This requires a mechanism for network engineers and data scientists to effectively share their expertise.
- New Producers, New Challenges: Emerging data-driven monitoring and automation solutions are being developed by data analysts and scientists who may lack domain knowledge of network operations.
- From monitoring to observability: Defining Observability as a consistent and systematic way towards gaining new knowledge about the network is beyond just monitoring by looking for patterns already known or monitoring alarms. For new knowledge to be gained, new tools and knowledge from other domains should be systematically introduced and the performance, quality and trustworthiness assessed.

Bridging the knowledge gap between network engineers and AI scientists is essential for advancing automation and moving towards automated observability Therefore, as a first step, the main question is: **How can we empower these two domains to leverage each other's expertise and create a seamless data ecosystem?**

II. Potential enablers

Data mesh is a data architecture that, while maintaining separate responsibilities between domains, enables federated and co-dependent data service support [DATA MESH]. This makes it a promising solution for the challenges outlined in the problem statement. However, the key question is: **How do the principles of data mesh map to the current NEMOPS enablers?**

a) The **domain ownership** principle emphasizes that domain teams are responsible for their data. This means that analytical data should be organized around domains, mirroring the team boundaries and the system's bounded context. In a domain-driven distributed architecture, ownership of both analytical and operational data shifts from a central data team to the domain teams.
To apply this principle, we need to define what constitutes a domain in our specific context. Several dimensions can be considered:
Data Model Domains: Each data model would be responsible for translating, connecting, packaging, and providing its data-to-data scientists.
Protocol Domains: Protocols would be responsible for translating their impact and connections with data models.
Additional Dimension: other relevant dimensions, such as functional areas, business units, or specific technology stacks.

This mapping necessitates a set of rules for each domain to follow. These rules should ensure that domains provide their insights in a composable manner, allowing for interoperability between them. These rules could be viewed as standards for new operations on top of "products" generated by each domain.

b) **Data as a Product: A Missing Link in Data Manifest**

The **"data as a product" principle** promotes a product-centric approach to analytical data. This means that data has consumers beyond the originating domain, and domain teams are responsible for providing high-quality data to meet those needs. Essentially, domain data should be treated like a public API. To address the problem of data silos, we need a unified way to share domain knowledge and connect data models.

Currently, the data manifest uses two YANG models to define context [DATA MANIFEST]:

- Platform Manifest: Provides information about the platform but lacks domain knowledge relevant to the data. It identifies the platform but doesn't offer insights into the specific domain the data belongs to.
- Data Collection Manifest: Focuses on collection configuration and subscriptions but doesn't offer context for data interpretation. It tells us how the data is collected but not what it means or how to use it.

To overcome these limitations, we need a new module or even a new language that is capable of not only introducing data but capable of carrying best of both worlds. Thanks to data scientists, the useful information on data's application and usefulness in AI-driven solutions is integrated in the model and owning to the network engineers, networking knowledge necessary to make correct decisions are captured. A construct as such can enable offering of new operations based on the two domains' characterization systematically. In a sense, data products can be viewed as an entity that marries the two worlds of data scientists and network engineers. The flexibility in the data composition can accommodate the requirements of data scientists and the semantics and operations introduced by network engineers can bring out the value from data most efficiently.

"Data as a product" could be implemented as a new YANG module, but it requires new protocols enabling new operations beyond traditional publish/subscribe mechanisms. Moreover, the constraints introduced by the hierarchical tree-like structure of YANG may need to be evaluated and further discussed. In another scenario, this new concept would require its own language, protocols and operations. The new solutions and enablers would provide a way to define and document the "data product" itself, including its domain context, meaning, intended use, etc.

c) **Self-Service Data Infrastructure: Enabling Interoperability**

The **"self-service data infrastructure platform" principle** promotes a platform-centric approach to data infrastructure. This means establishing a dedicated data platform team that provides domain-agnostic functionality, tools, and systems to build, execute, and maintain interoperable data products for all domains. This platform empowers domain teams to seamlessly consume and create data products, fostering a self-service environment.

To ensure interoperability across different data models and domains, the following considerations are crucial:

- New Language: If a new language is defined for data products, it needs a protocol for data product modules that allows for flexible operations between domains. Simple request/response mechanisms should be supported for customization. This ensures that different domains can interact with each other's data products in a standardized way.
- YANG Module: If a new YANG module is used for data products, it should work with a new unified protocol to support inter-data model and domain operations. The module should address the minimum requirements for data product interoperability. This ensures that data products defined using YANG can be seamlessly integrated and exchanged across different domains.

d) **Federated Governance: The Foundation for Interoperability**

The **federated governance principle** is crucial for achieving interoperability of all data products within a data mesh. It promotes standardization across the entire data ecosystem, ensuring adherence to organizational rules and industry regulations. This principle emphasizes the need for well-defined standards to achieve the vision of a cohesive and compliant data environment.

- New authority: the interoperability and federation of data products across multiple domains [if in definition of domains it is decided to go beyond network data] may need to adhere to a set of cross-industry standards. This matter is only of concern in IETF if we envision the solution's application scope is flexible and extensible to other industries. The domains are then defined and dimensioned to adapt and reach beyond IETF scope of data and operations.

Main Questions to Address:

To effectively implement federated governance, several key questions need to be addressed:

- YANG Extension as a Starting Point: Is extending YANG the most suitable approach for defining data product standards? Are there alternative options that might be more effective?
- Integration with Existing Models and Protocols: How can other data models and protocols be integrated with the new standards? What mechanisms are needed to ensure seamless interoperability?
- Domain Ownership and Responsibility: Are all domains willing to take responsibility for adhering to the new standards? Who should be involved in the initial discussions and decision-making process?
- Knowledge graph over data products or Knowledge Graph over all data: Is it more beneficial to create a knowledge graph/semantic layer over data products or over data abstracted on interfaces or on service level - What are the trade-offs in terms of scalability, responsiveness, and speed of adoption?

Final Words:

The final decision of how to tackle the above problem depends on several factors such as:

- How the domains are defined and dimensioned: this requires expertise and deeper understanding of both worlds (data science and networking) which can only be reached by collective discussions. The dimensions mentioned in this position paper are merely a starting point and should not be presumed as final.
- How big of a scope we aim to cover: Despite only highlighting the gap between the network engineers and data scientists, this gap is evident in almost all domains and bridging this gap requires a set of common enablers. This does not mean that the work should stop at common general aspects but, it could follow a layered approach where the first layer, lays down common enablers allowing for a level of interoperability and the other layer(s) will enrich the solution with network engineer's domain knowledge. After all, when it comes to data, the more unified and connected, the more the growth of the value generated.

The choice of opting for a new language, protocol and, likely, governance rules or going with YANG should be made after mulling over the pros and cons. Moreover, after clarifying and agreeing on the scope of adaptability of data mesh, a deeper study on requirements and their overlap with ongoing work in IETF can bring out additional insight that makes this decision easier. This insight allows us to assess if what YANG brings to the table are constraints or enablers.

IV. References

[RFC9232] Song, H., Qin, F., Martinez-Julia, P., Ciavaglia, L., and A. Wang, "Network Telemetry Framework", RFC 9232, DOI 10.17487/RFC9232, May 2022, https://www.rfc-editor.org/info/rfc9232

[DATA MESH] Inês Araújo Machado, Carlos Costa, Maribel Yasmina Santos, Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures, Procedia Computer Science, Volume 196, 2022, Pages 263-271, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.12.013

[DATA MANIFEST] B. Claise, J. Quilbeuf, D. Lopez, I. Dominguez, T. Graf, “A Data Manifest for Contextualized Telemetry Data”, draft-ietf-opsawg-collected-data-manifest, Work in Progress, October 2024, https://datatracker.ietf.org/doc/draft-ietf-opsawg-collected-data-manifest