Building a Data Mesh using the Lake House Approach

Roy Hasson – Principal Product Manager, AWS

Nivas Shankar – Principal Data Architect, AWS

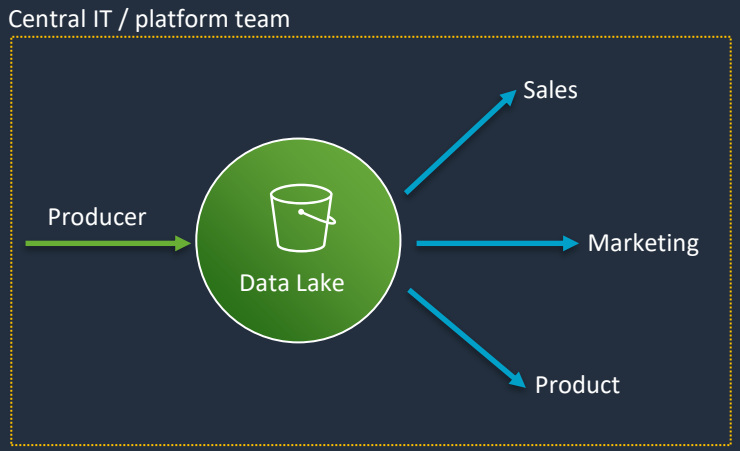
What is a data lake?

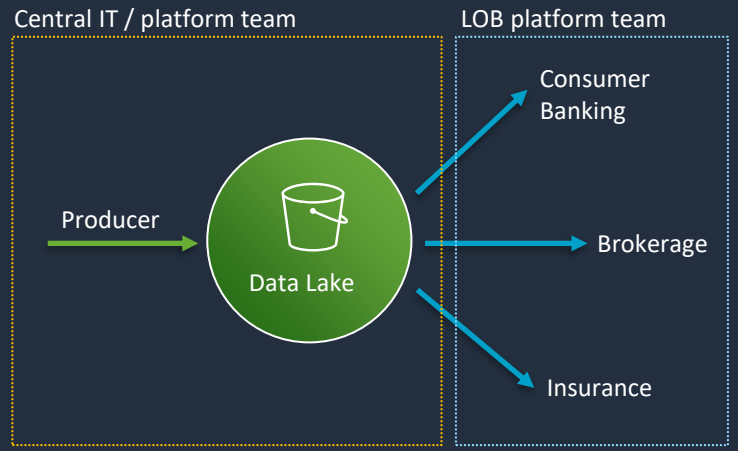
A data lake is a **centralized**, curated, and secured **repository** that stores **all your data**, both in its original form and prepared for analysis.

A data lake enables you to **break** down data **silos** and combine different types of analytics and ML to **gain insights** and guide **better business decisions**.

aws

Organizing around a central data lake





aws

Challenges with central data lake *management*

Key Challenges:

- Misalignment between producer and consumer needs
- Lack of consumer autonomy
- Lack of data ownership and accountability
- Difficult adopting to multi-regional and conglomerate structures

Results:

- Slower pace of innovation
- Reduced collaboration
- Duplication of effort, personnel and data
- Diverging tech stacks, increasing costs and tech debt
- Emphasis on using data, rather than making data more useable

aws

The Data Mesh pattern

A **Data Mesh** is a paradigm shift in how we think about building data platforms. The architecture is the convergence of *Distributed Domain Driven Architecture*, *Self-serve Platform Design* and *Product Thinking with Data*. *Zhamak Dehghani, Thoughtworks*

Key pillars of a data mesh

- Domain-oriented, decentralized data ownership and architecture
  - Organizational autonomy
- Creating data products with true data owners
  - Single-threaded owner of data, treated as a product
- A self-service data infrastructure platform
  - Simple to use data tools using a common infrastructure framework
- Federated computational governance
  - Virtualize access to data in a secure and governed way

* https://martinfowler.com/articles/data-monolith-to-mesh.html
** https://martinfowler.com/articles/data-mesh-principles.html

aws

Pros and Cons of a Data Mesh pattern

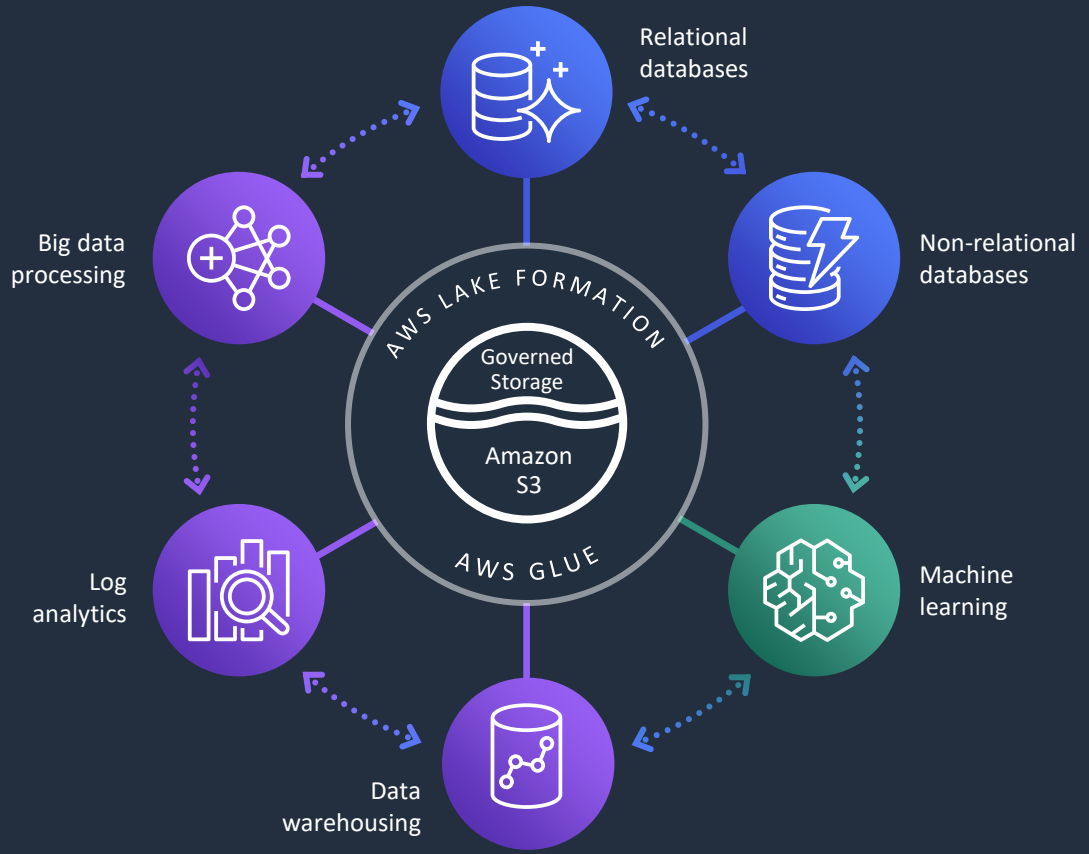
Pros

- Distributed control
- Solutions aligned with business needs
- Data ownership and accountability
- Increase adoption of data
- Scales to meet evolving org structure
- Encourages product thinking

Cons

- Distributed knowledge and skillsets
- Potential for diverging tech stacks
- Difficult to secure and audit data
- Product thinking is not for everyone
- Building data as product is not simple
- Not always a good fit for small orgs

The Lake House Approach



SCALABLE DATA LAKES

PURPOSE-BUILT DATA SERVICES

AUTOMATED DATA MOVEMENT

CENTRAL GOVERNANCE

PERFORMANT AND COST-EFFECTIVE

aws

AWS Lake Formation

Build a secure data lake in days

Build data lakes quickly

Move, store, update, and catalog your data faster
Automatically organize and optimize your data

Simplify security management

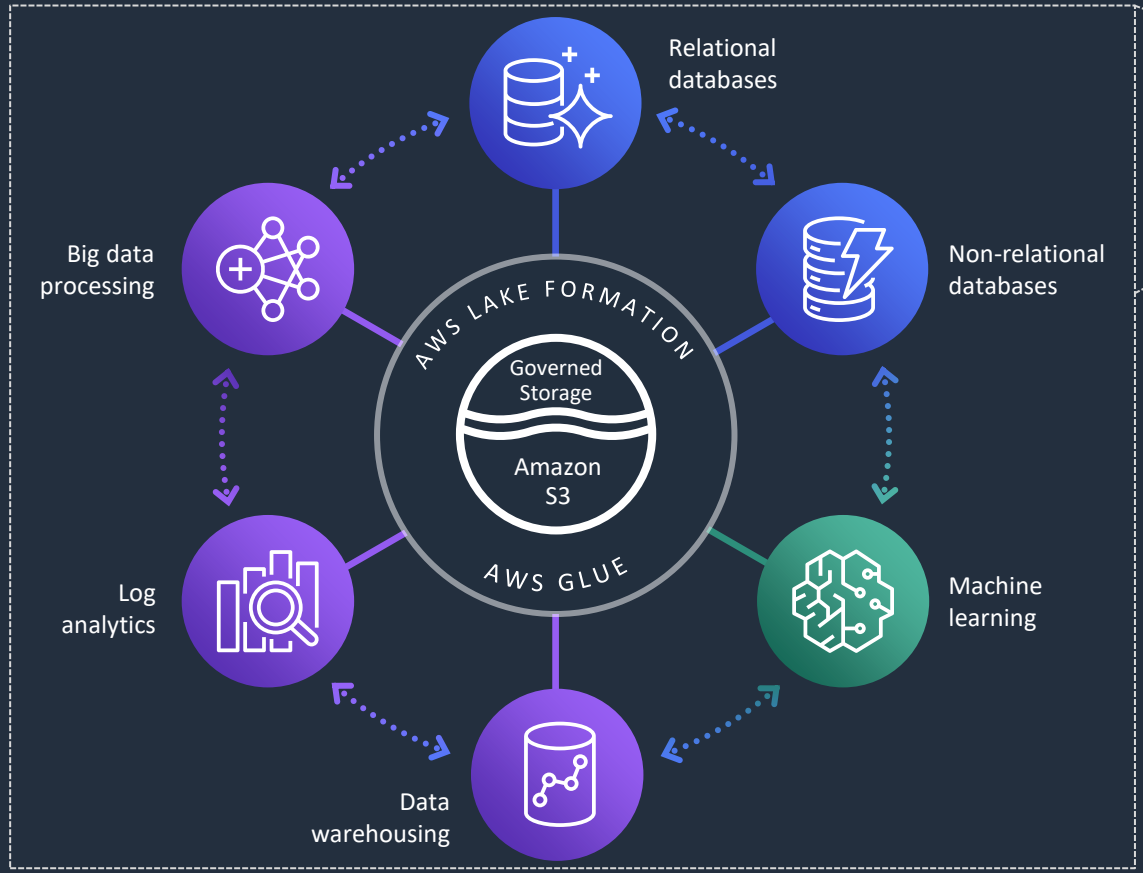
Centrally define and enforce security, governance, and auditing policies

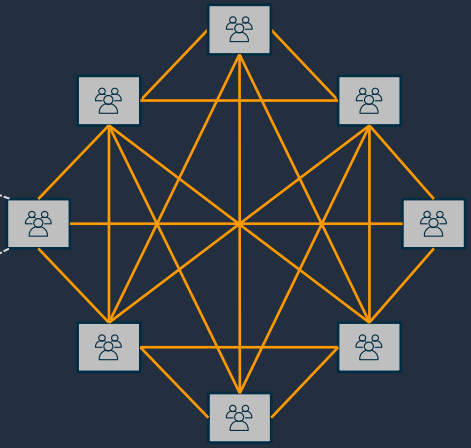
Easily discover and share data

Catalog all of your data assets and easily share datasets between consumers

aws

Implementing Data Mesh with Lake House





- Common tech stack
- Scalable, durable and available
- Secure and compliant
- Simple to manage, standard ops
- Common skillset / quicker ramp up
- Cost effective

aws

Lets get technical

aws

Data Mesh - Core Concepts

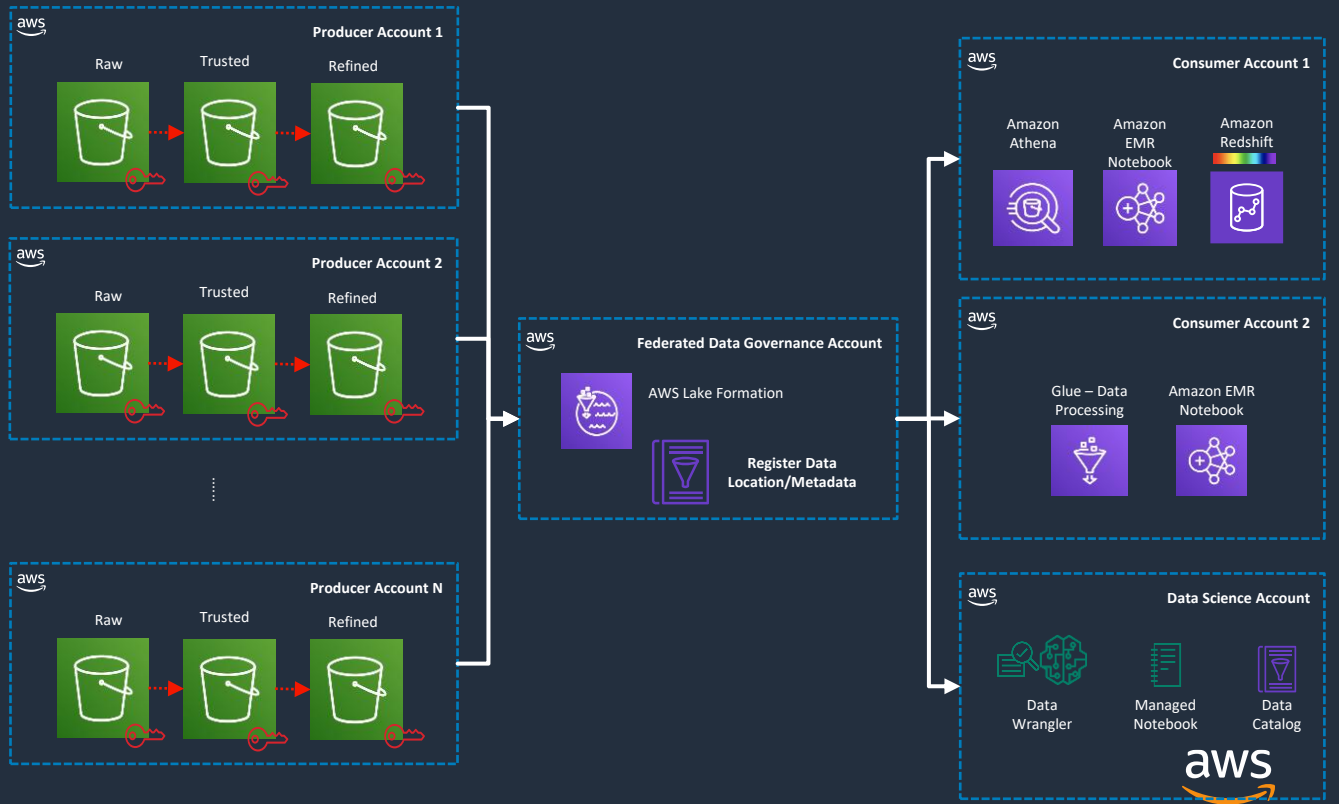
A decentralized, domain-oriented data architecture to enable governed sharing across data lakes

Data Domains are nodes that make up a data mesh.

Data producers shares one or more Data Products by making them discoverable through a common catalog

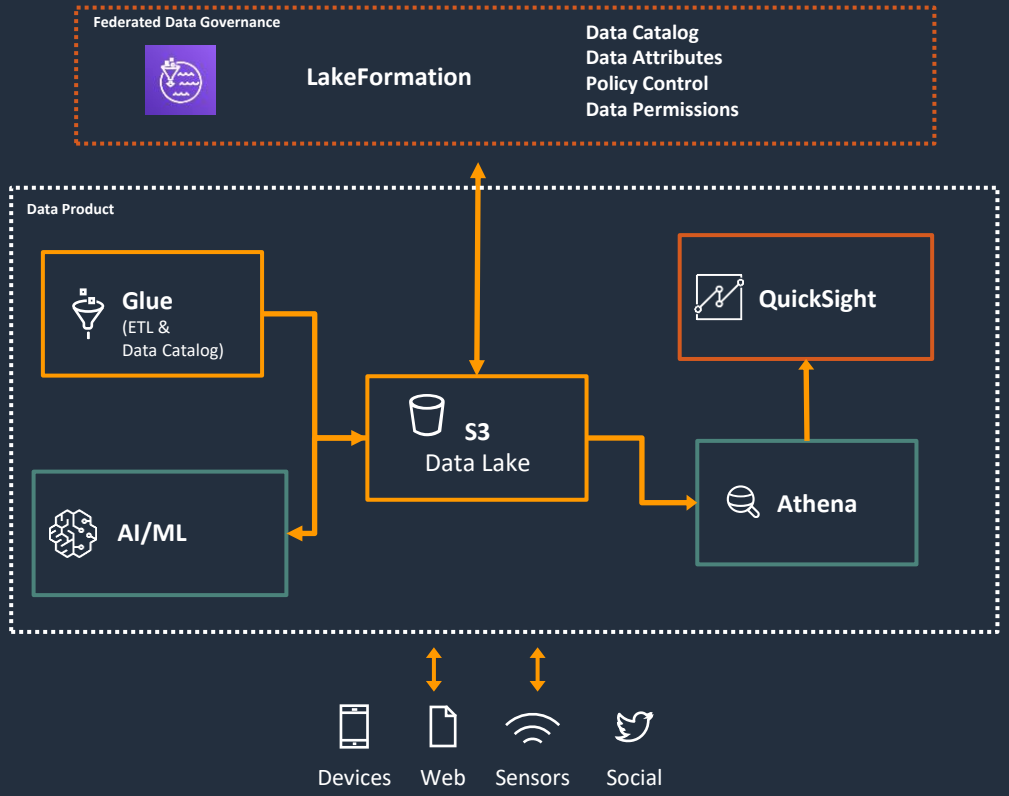
Federated data governance enables security and compliance across data domains

Data consumers easily discover and access data using Resource Shares or APIs



Build Data Products

Single Account design using Data Mesh pattern



Data domain producers ingest data into their respective S3 buckets through a set of pipelines that they own and operate

Producers are data owners responsible for the full lifecycle of the data under their control

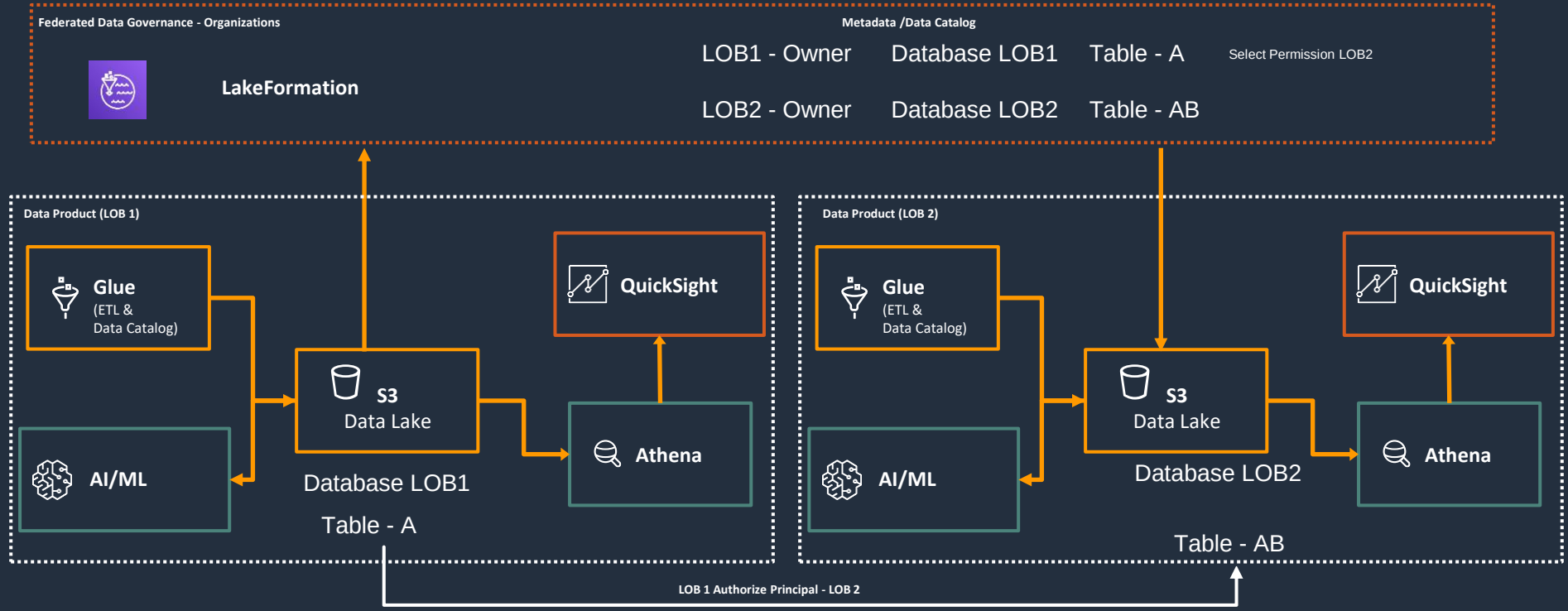
Producers catalog datasets to make them discoverable and accessible by consumers

Producers control access to data through Federated data governance model to enforce fine-grained permissions

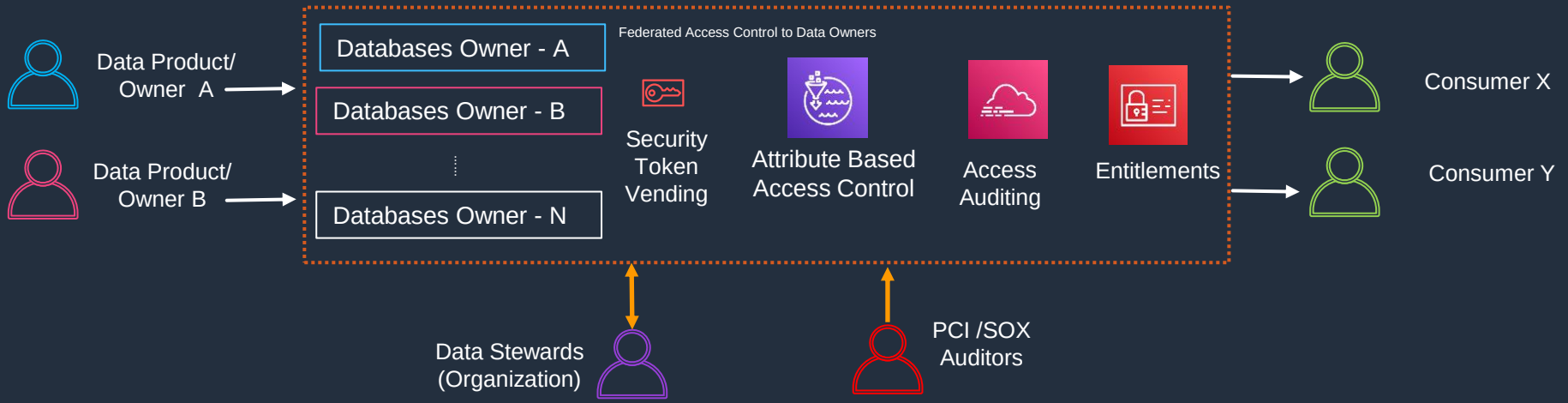
aws

Build and Share Data Products

Securely share AWS Accounts & Organizations



Federated Data Governance



Common security features are core to a governed data mesh

Common identity provider across Producers and Consumers

Credential vending to simplify service integration

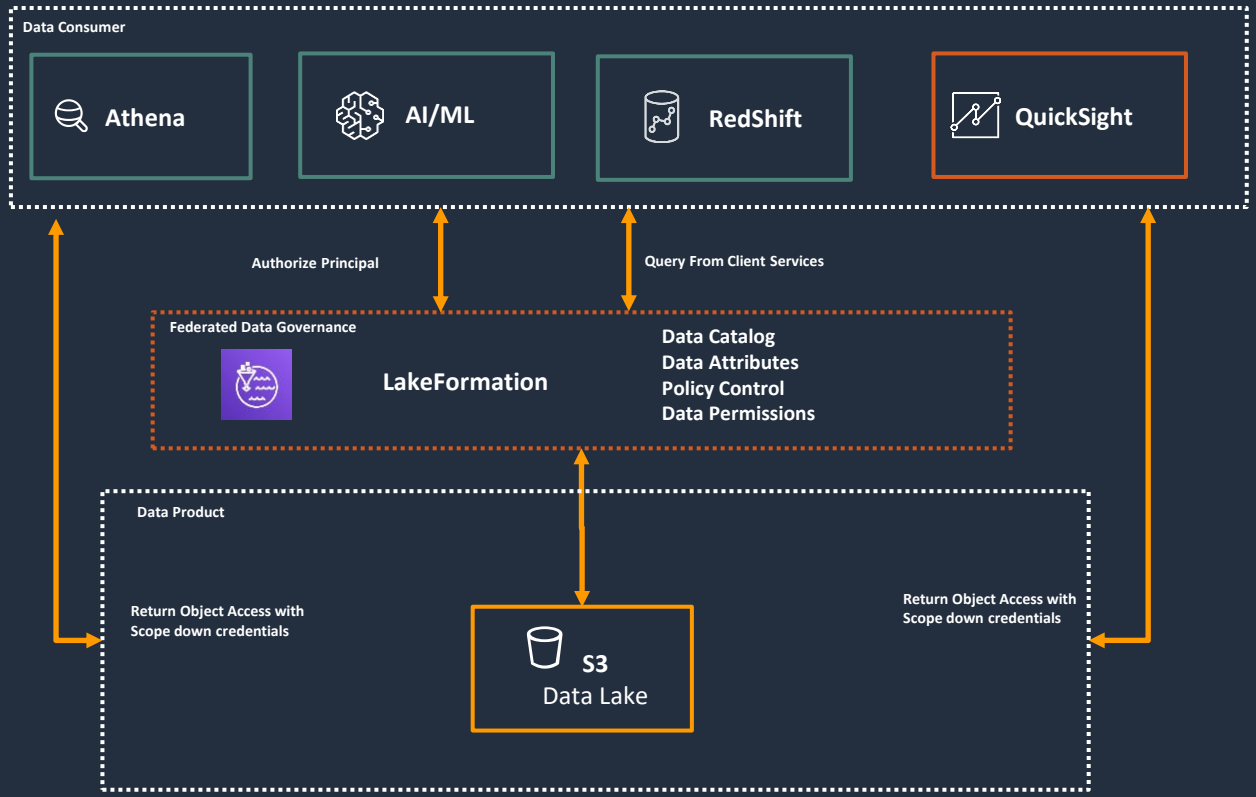
Federated metadata search

Fine-grained entitlements & ABAC

Central audit and compliance

aws

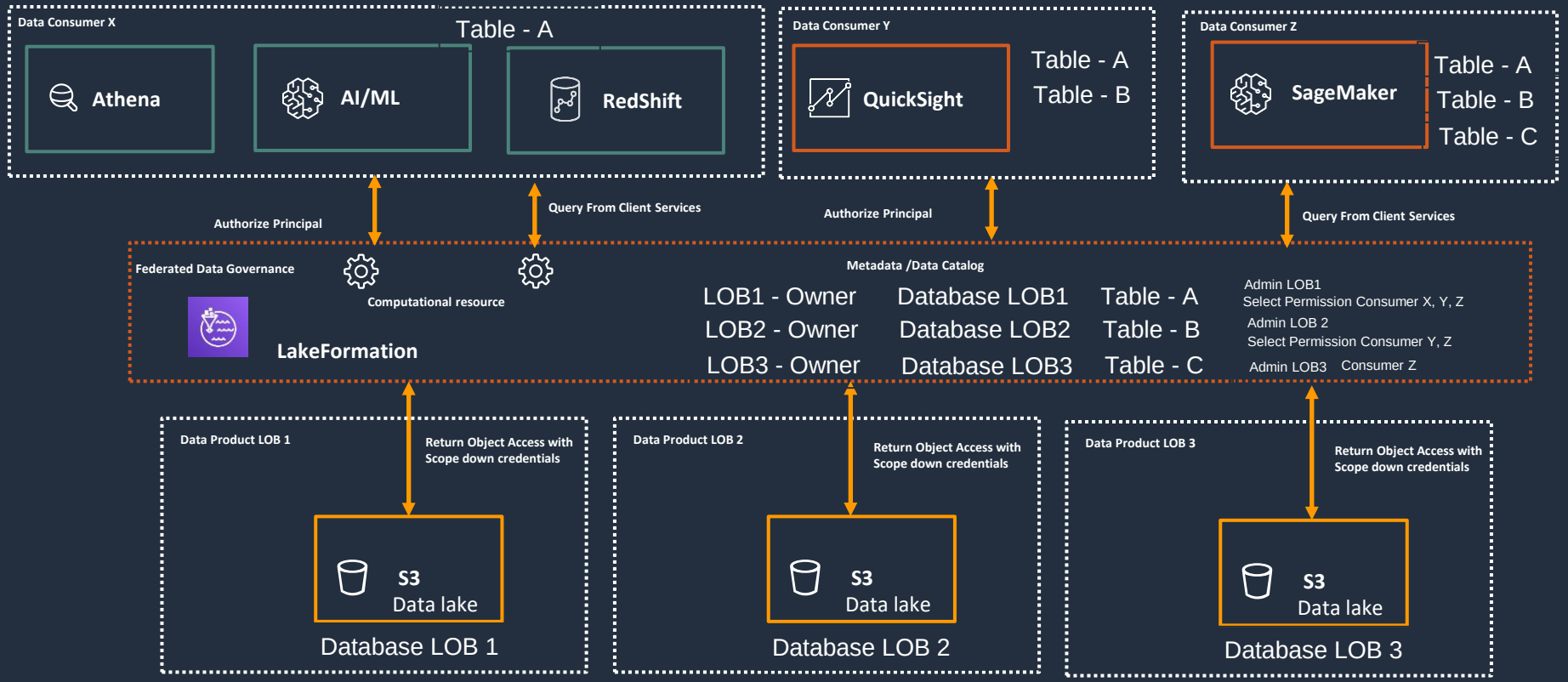
Data Consumer - Common Access (Single Account)



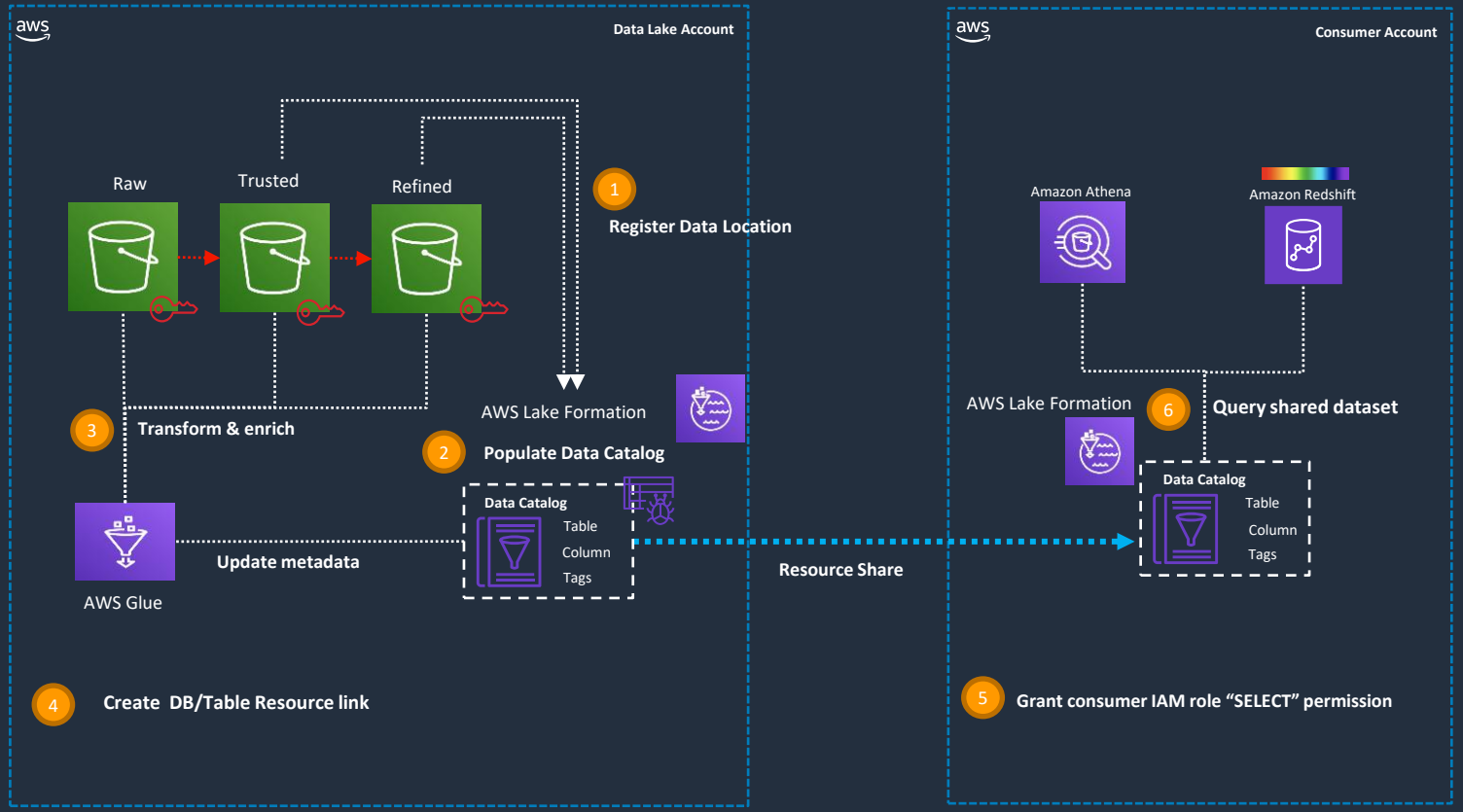
- Data Consumer finds and requests access to a data product
- Data Producer approves request and initiates a resource share with consumer account.
- Data Producer grants fine-grained permissions to dataset shared with the consumer.
- Data access can be validated and audited through federated data governance

aws

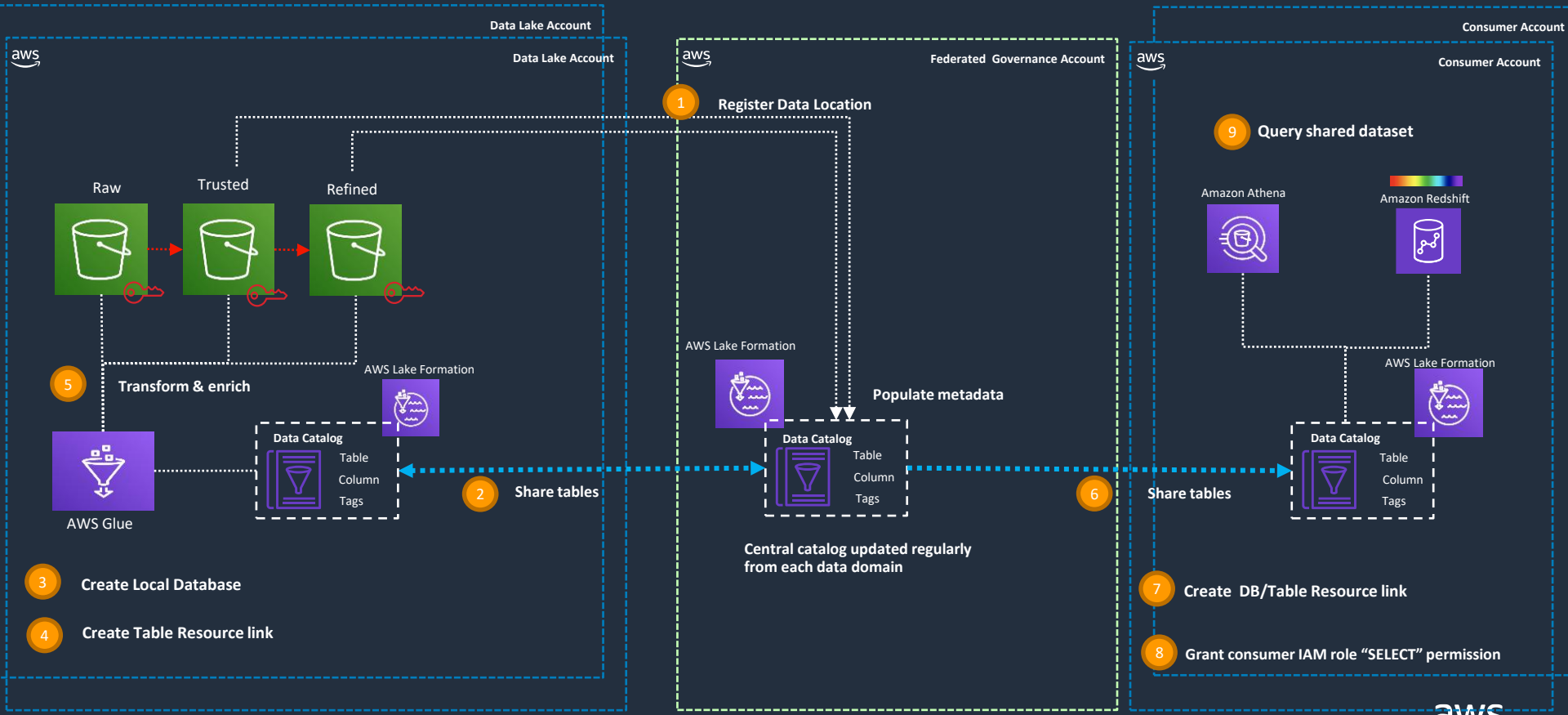
Data Consumer - Federated Computational Governance



Data Mesh on AWS – Peer to Peer



Data Mesh on AWS – Central Governance



Summary

- Data Mesh enables organizations to be autonomous, increase pace of innovation
- Data Mesh is not for everyone
- Lake House approach offers a common tech stack to simplify deploying a data mesh
- AWS Lake Formation simplify building a data product (catalog, security, access, sharing)
- AWS native services and partner solutions enable self-service analytics and ML

aws

Thank You !

Roy Hasson - /in/royhasson
Nivas Shankar - /in/nivasshankar/

Learn more:

https://martinfowler.com/articles/data-monolith-to-mesh.html

https://datameshlearning.com/

https://aws.amazon.com/blogs/big-data/design-a-data-mesh-architecture-using-aws-lake-formation-and-aws-glue/

https://aws.amazon.com/blogs/big-data/how-jpmorgan-chase-built-a-data-mesh-architecture-to-drive-significant-value-to-enhance-their-enterprise-data-platform/

aws