GNARLY
Data_Waves

PRESENTED BY dremio

EPISODE 19

Data Mesh In Practice: Accelerating Cancer Research with Dremio's Data Lakehouse

Arfath Pasha
Sr. Software
Engineer, MSK

Tony Truong
Sr. Product Marketing
Manager, Dremio

May 31 at 8AM PST | 11AM EST | 4PM GMT

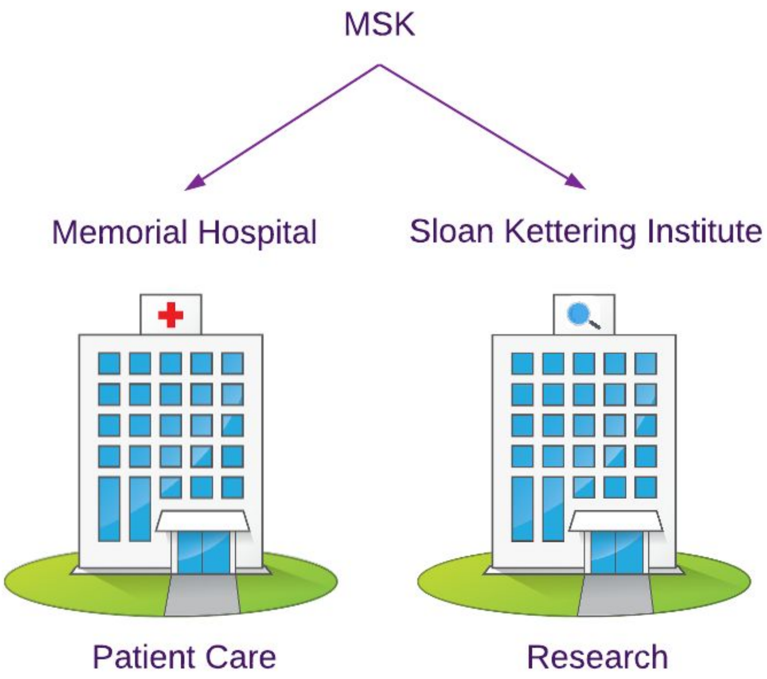
dremio | Memorial Sloan Kettering Cancer Center

Memorial Sloan Kettering Cancer Center

Founded in 1884

Treats more than 400 cancer subtypes

- 20k inpatient and 700k outpatient visits
- > 1800 research protocols



MSK

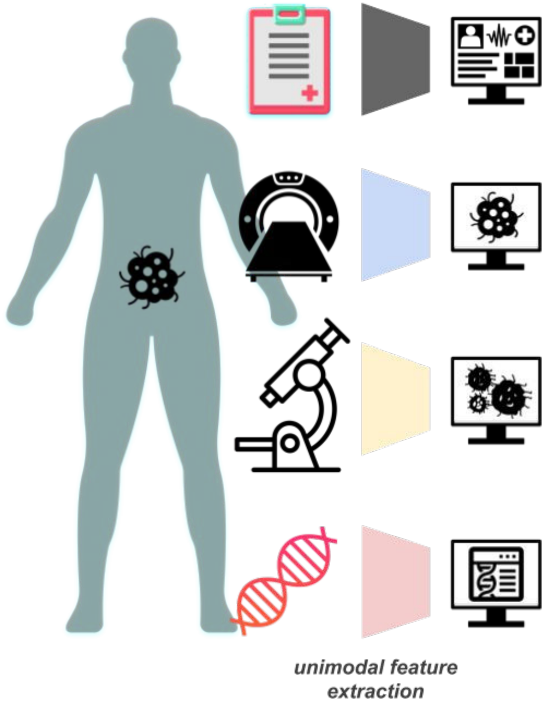
Memorial Hospital

Sloan Kettering Institute

Patient Care

Research

Research at MSK builds on top of existing resources in various departments at MSK

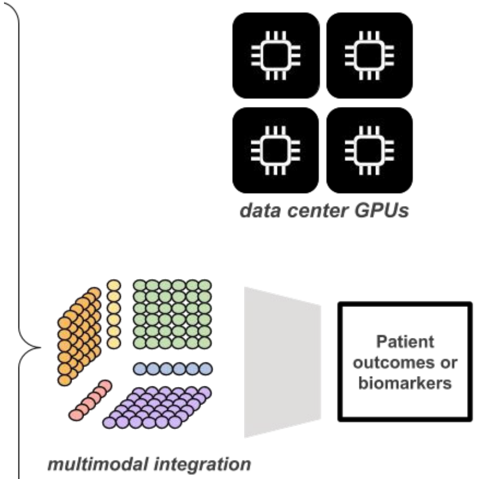


Clinical Informatics

Radiology Informatics

Pathology Informatics

Multiple sources



$$P(y|w,x) = \mathrm{Ber}(y|\sigma(wx))$$

biomarker status classification

$$l(w) = \sum_{c_i=1}\left[\theta_i' - \log\sum_{t_i \geq t_j}\exp(\theta_i')\right]$$

censored event modeling

Challenges: regulatory, data governance, and technical barriers

Engineering for Research

Our goal:

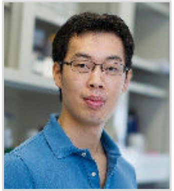
- Build the right (scientific data management+compute) system

We need to be:

- An infrastructure team
- A data products team
- An analysis team

We are:

- Research Software Engineers



Raymond Lim
Senior Engineer

Armaan Kohli
Engineer

Darin Moore
Engineer

Benjamin Gross
Lead Engineer

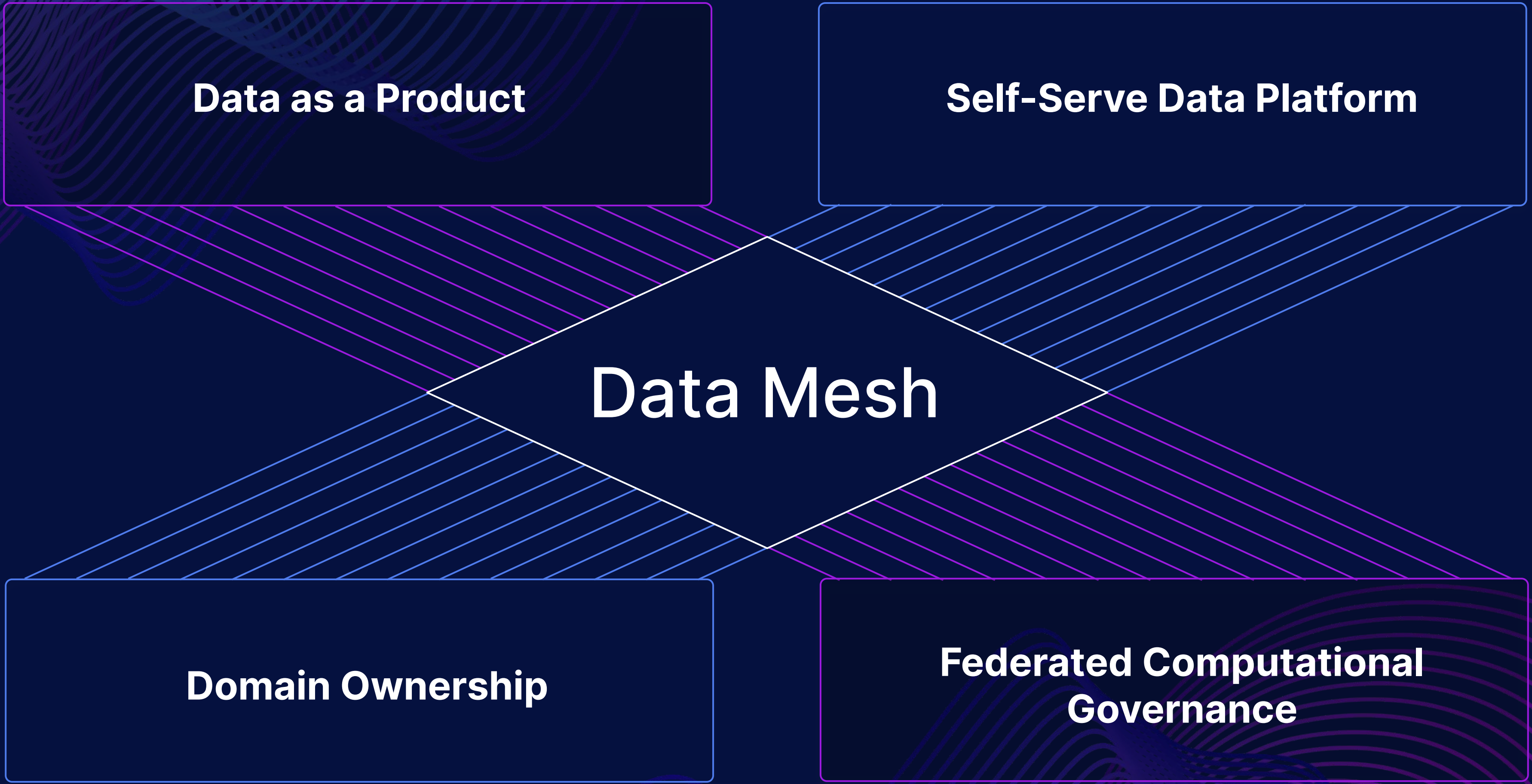
Anika Begum
Project Coordinator

Arfath Pasha
Senior Engineer

What is a Data Mesh?

A modern approach to data management that emphasizes distributed ownership and governance of data within domains, who then build, manage, and share data products across the organization.

“Dehghani, Z. (2022). Data Mesh: Delivering data-driven value at scale. O’Reilly Media, Incorporated.



Data as a Product

Self-Serve Data Platform

Data Mesh

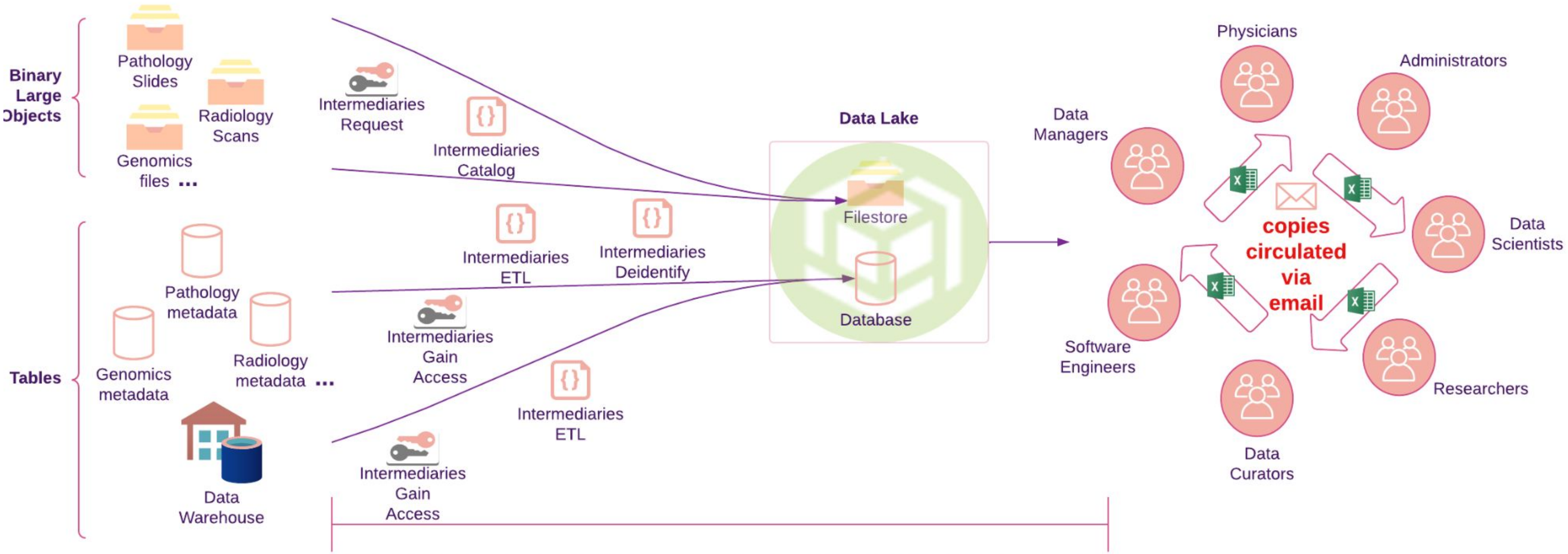
Domain Ownership

Federated Computational Governance

Data Challenges in Research

- High dimensional data (clinical, genomic, radiology, pathology, etc.)
- Teams with diverse skill-sets (engineers, scientists, pathologists, physicians, administrative staff, etc.)
- Highly iterative in nature (need for data versioning)
- Unstructured and structured data
    - Binary large objects and related tabular data
    - Long datasets (100-1 billion rows), wide datasets (10-1000 columns)
- Messy data (correctness, completeness issues)
- Siloed data (many data marts)
- Privacy

Data Management Before Dremio



Time (data is made available to researchers in weeks to months per project)

Solution Considerations

Architecture

- On-premises deployment
  - Data mesh
  - Query engine
- A no-copy data architecture

Decentralized Data Management

- Eliminate siloed ETL pipelines, provide faster access to data
- Documentation support - data sheets for data sets
- Simple, mature governance model

People

- Easy interface for all data consumers

Why Dremio

Easy barrier to entry (satisfied our 1 hour rule for evaluating a new technology)

Supports on-prem deployment with path to cloud

Access control (unified semantic layer)

Data democratization (almost spreadsheet like interface plus connection to Tableau)

No copies (no more emailing copies)

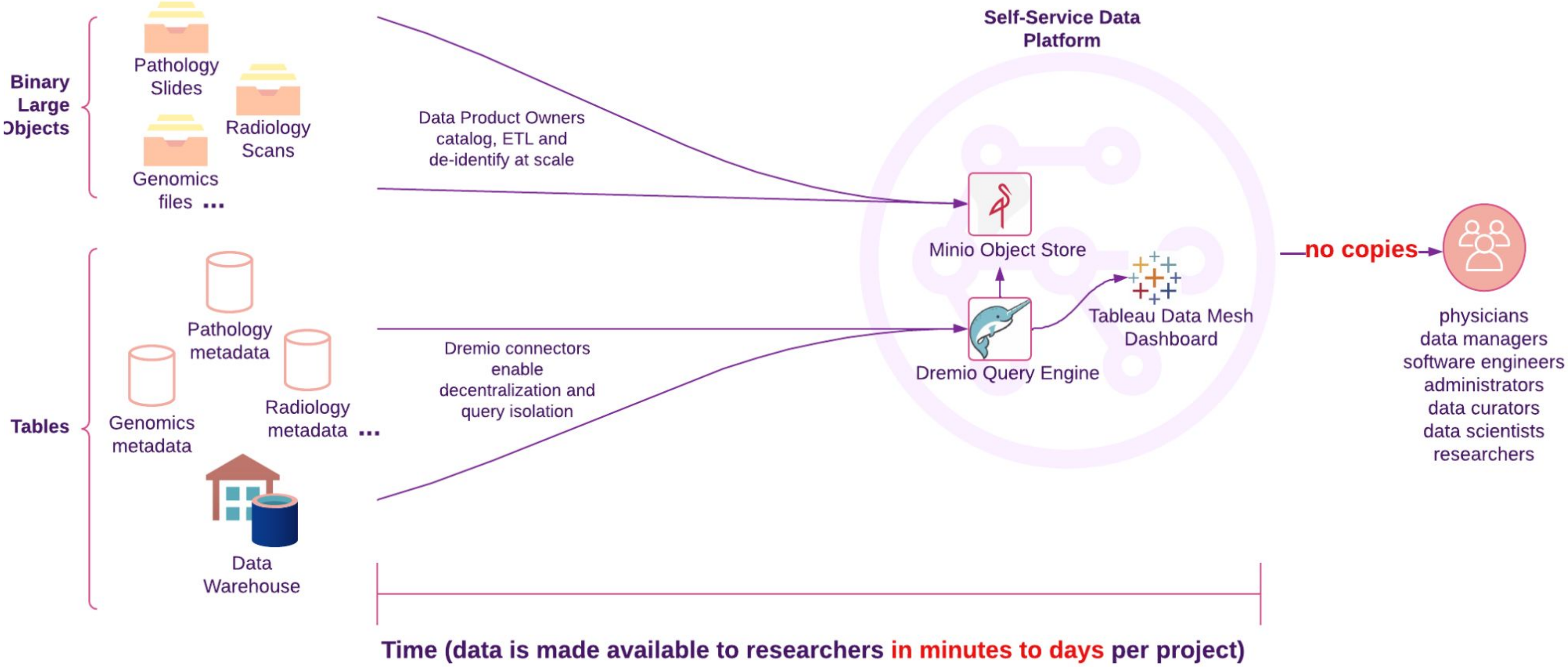
Low code / no code (easy data inspection/curation/integration without pandas code)

Performance (horizontally scalable; Arrow Flight access)

Datasheets for datasets (support for documentation through catalog wiki)

Data versioning (Iceberg and Nessie)

How We Are Decentralizing Data Management With Dremio



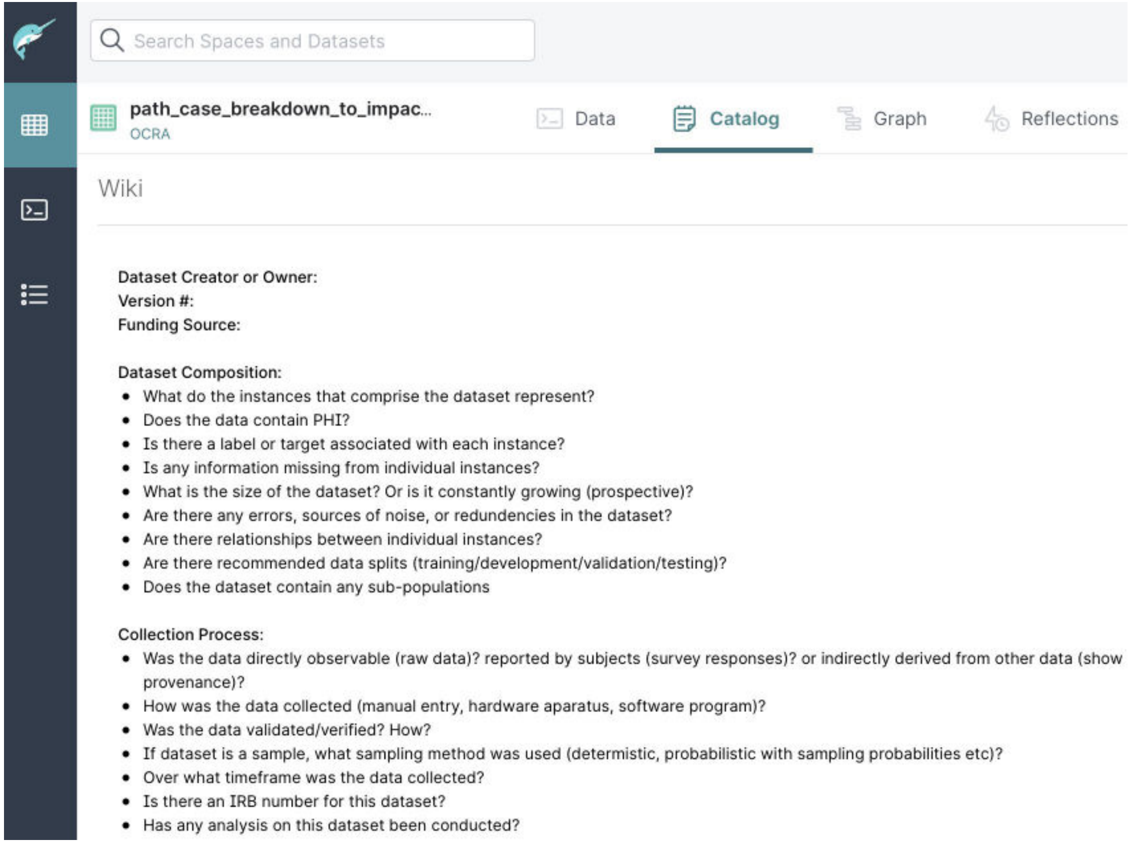
Lessons Learned Along the Way

✓ Building trust between data product owners and consumers with data mesh

✓ Reduced data product delivery time by eliminating siloed ETL with Dremio

✓ User data copies eliminated, now easier to track and share across domains

Datasheets for Datasets



Dataset Creator or Owner:
Version #:
Funding Source:

Dataset Composition:

- What do the instances that comprise the dataset represent?
- Does the data contain PHI?
- Is there a label or target associated with each instance?
- Is any information missing from individual instances?
- What is the size of the dataset? Or is it constantly growing (prospective)?
- Are there any errors, sources of noise, or redundencies in the dataset?
- Are there relationships between individual instances?
- Are there recommended data splits (training/development/validation/testing)?
- Does the dataset contain any sub-populations

Collection Process:

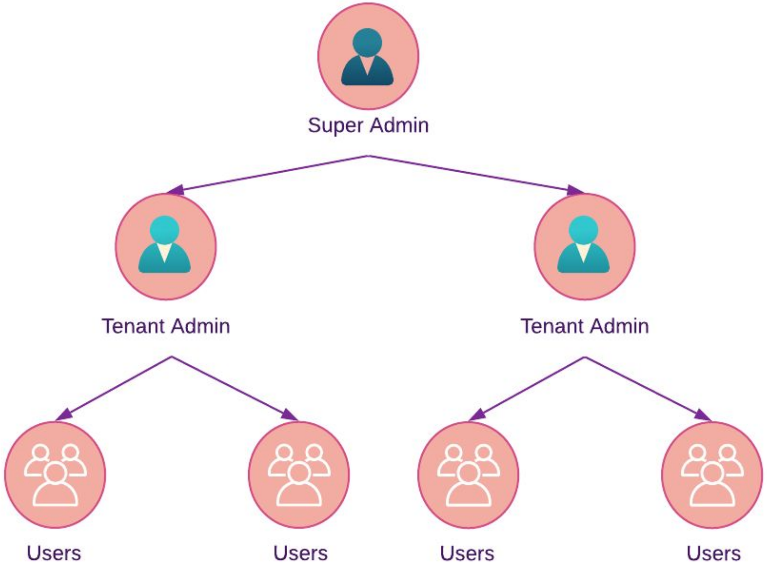
- Was the data directly observable (raw data)? reported by subjects (survey responses)? or indirectly derived from other data (show provenance)?
- How was the data collected (manual entry, hardware aparatus, software program)?
- Was the data validated/verified? How?
- If dataset is a sample, what sampling method was used (determistic, probabilistic with sampling probabilities etc)?
- Over what timeframe was the data collected?
- Is there an IRB number for this dataset?
- Has any analysis on this dataset been conducted?

Reference: https://arxiv.org/abs/1803.09010

What’s Next - Data Mesh Evangelism

Federated Computational Governance

- Standardizing data governance between domains and make decisions about how data is used and shared across the enterprise.
- Need for ‘tenant’ admins for multi-tenancy



GNARLY
Data_Waves

PRESENTED BY dremio

Thank you!