A Pragmatic Data Mesh Introduction

Let's figure out if the data mesh is for you.

meltano
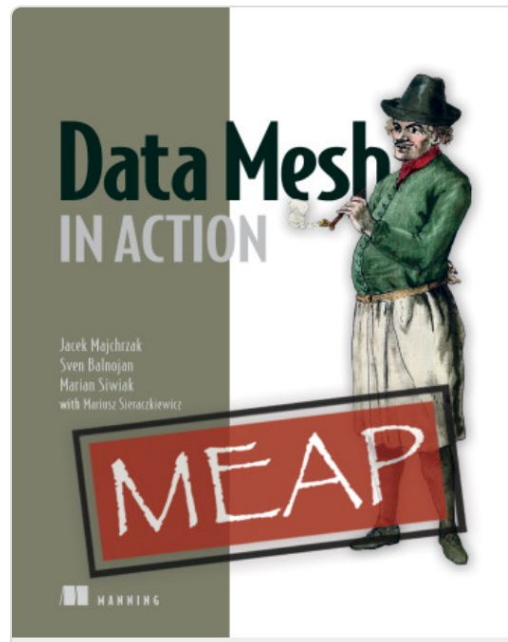
Sven & why you should listen

Sven Balnojan

DataOps Evangelist
@Meltano

Tweets at: @sbalnojan

What I do

- Coauthor of "Data Mesh in Action" (Book)

- Author of "Build a Small Dockerized Data Mesh" (LiveProject)

- Write a lot ("Three Data Point Thursday", Blogs,..)

Listener **Manifesto** - **What you get out of listening**

- **Goal**: Spike your interest, some guidance of whether this is something for you.
- **Method**: Real-world example taken from a medium-sized company (like a minimum viable example).
- **Prerequisites**: You should've heard the word data mesh before.
- **Not in scope**: technical things; completeness; details.

Important: You cannot "build" a data mesh.

But you can implement one, if you start talking to people about it, and then build an MVP platform to support it.

Definition?: "Data Mesh is a socio-technical paradigm shift,... modern distributed architecture, domains, platforms, data as a product,...
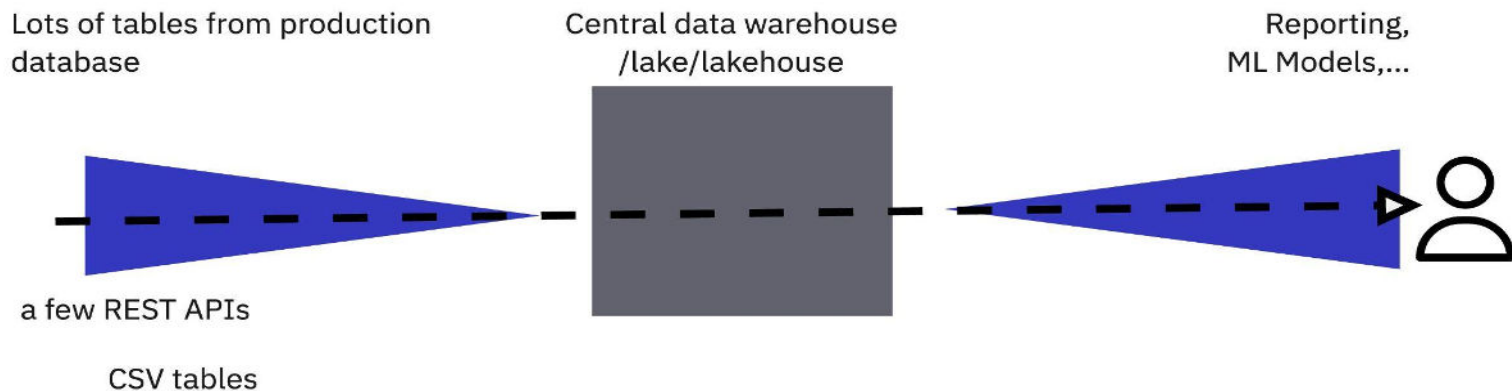
Definition! Data Mesh is about decentralization.

It aims to increase value/data.

It's guided by four "principles", implementations vary in their utilization.
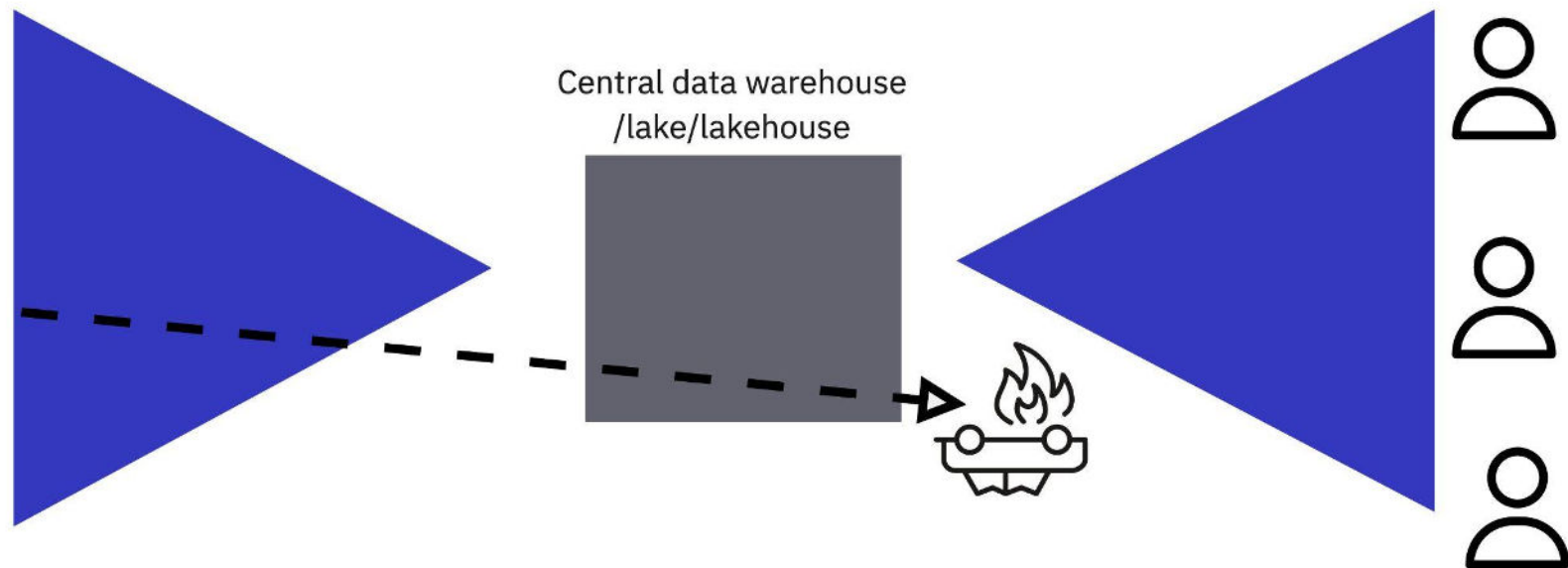
When **decentralize?** The Producer-Consumer Gap?

Lots of tables from production database

Central data warehouse /lake/lakehouse

Reporting, ML Models,...

a few REST APIs

CSV tables

The data world at "Unite", 2017.

When **decentralize?** The Producer-Consumer Gap?

Central data warehouse /lake/lakehouse

The data world at "Unite", 2020 (for a brief moment)

How the problem looked like @ Unite

Before these major shifts:

- Lead time for new use cases involving new data sources.

~ a couple of weeks

After these major shifts:

- Lead time for new use cases involving new data sources.

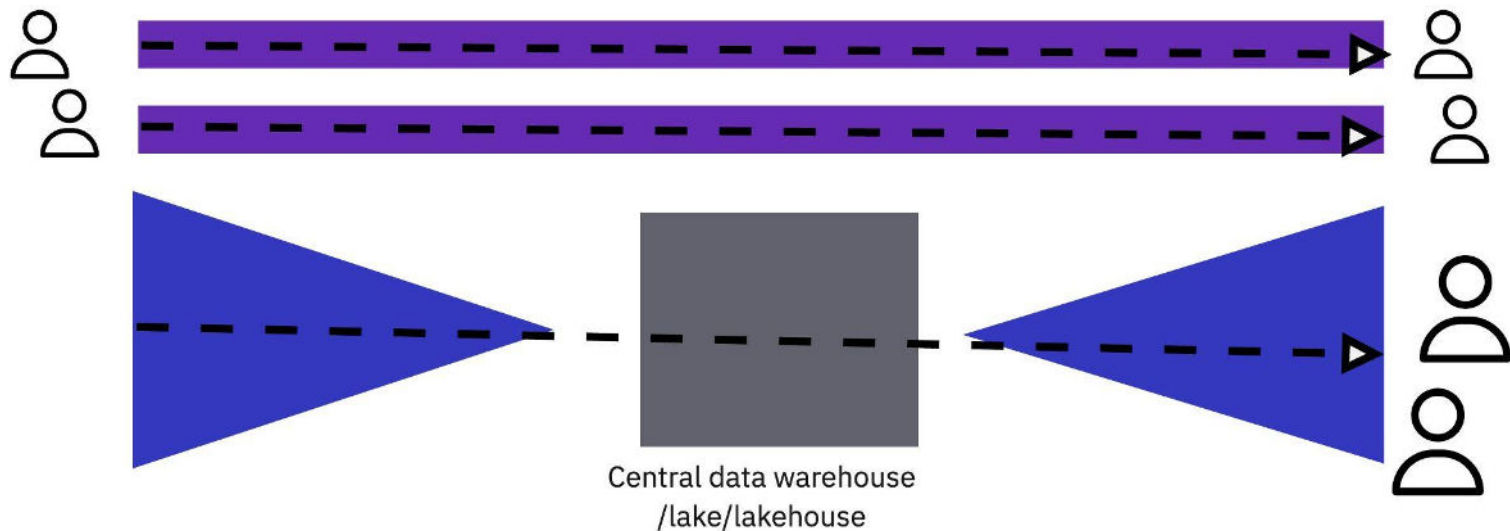*~ 3-6 months*

How to tackle new valuable use cases?

1. Operational (Automation)
2. Analytical (Product related reporting)

Identify the decentralized actors:

- **Data producers (new sources)**: a bunch of software engineering teams
- **Data consumers**: decentralized business analysts & some operational engineering teams

Example of the **Decentralization (Solution)**

Central data warehouse
/lake/lakehouse

Data world 2020: Key analytics
And key operational use cases handled NOT
through central pipe, but decentralized.

How the solution worked out @ Unite

Before data mesh:

- Lead time for new use cases involving new data sources.

~ 3-6 months

With data mesh:

- If involve no new platform functionality

~ a day

- With new platform functionality

~ 2-3 weeks

Isn't that a technical solution?

Let's compare what people now do.

What do people now do?
Shift in priorities.

Analyst: When a data point is missing in a report?

=> the data producing team (the dev team)

Dev team: When a consumer wants new data for operational use cases?

=> the PM for the data producing unit.

Dev team: When a consuming team needs new platform functionality?

=> the central data team.

The four principles

& where to start

Distributed Domain Ownership

Data is owned by a data-domain expert (team).

Applied: At Unite, the software engineering team producing the data also became the owner of the data.

Data as a Product

Data is not a "by-product" anymore.
It's wrapped into a "data-service",
subject to usual product management.

Applied: At Unite, the PM now has to weigh off feature dev & data dev.

Platform Thinking

Building data services is hard => duplicate effort should be abstracted away and put into a "platform".

Applied: At Unite, the central data team building the tracking solution became a partial platform team.

[Federated Computational Governance]

Data Mesh is iterative

And you don't need all principles, start with one.

"Yes we get it! We agree! BUT could you please talk to our PM to prioritize this kind of work?" - every dev I talked to about data mesh

How to start?

1. Consumer-Producer Gap? [Yes/No]
2. [Yes] Identify 1-2 use cases where $$$ is there.
3. [for use cases] have brainstorm/discussion with PMs & tech leads to see how this can be solved (better with the data mesh) - how **responsibilities** shift.
4. [If people agree] Start work on the kernel. *(Hint: If you don't have a data producer to test this with, you don't have buy-in yet)*
5. *(see "Data Mesh in Action" for a detailed walk-through - or the "Data Mesh learning community".)*

What I want you to **take away**

- You don't need all principles, you start with distributed domain ownership.
- If consumer-producer gap is not your problem, there likely is a better solution than the data mesh.

References:

- https://support.unite.eu/de_DE/blog/how-mercateo-is-rolling-out-a-modern-data-platform
- Jan Nitschke from 11 Owls, *Master Thesis on the Principles of the Data Mesh (ask him on LinkedIn or me)*.