



# EDM Webinar

*2-part webinar series*

## Data Management in an AI-Driven Era: Why MDM is No Longer Sufficient

*A conversation with*



**Dr. Michael Stonebraker**  
Adjunct Professor  
MIT CSAIL



**Anthony Deighton**  
General Manager  
Tamr



Moderator



**Eric Bigelsen**  
Head of Industry  
Engagement  
EDM Council



**Dr. Michael Stonebraker**  
Adjunct Professor  
MIT CSAIL



**Anthony Deighton**  
General Manager  
Tamr





# Master Data in the age of AI and Cloud

---

# What we'll cover today

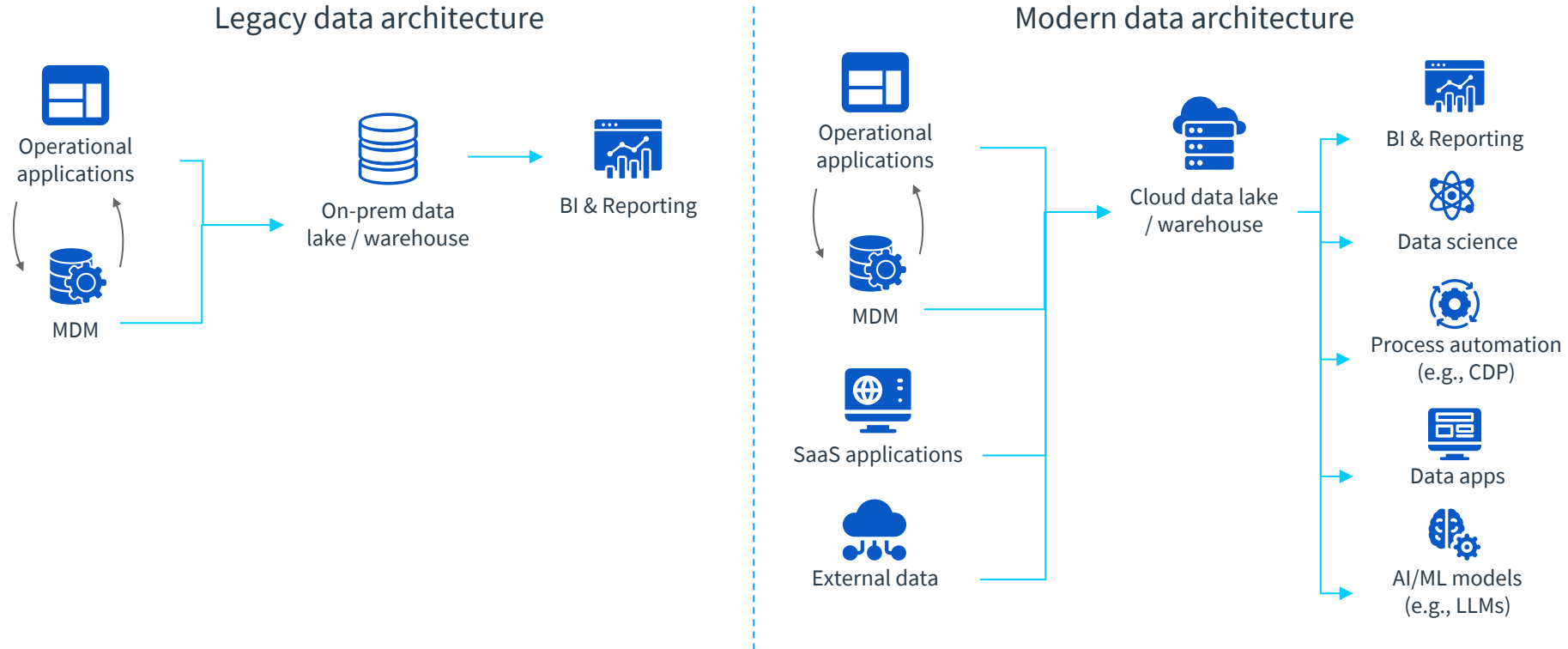
- What is the data mastering challenge (and opportunity!) in the enterprise?
- Why is data mastering hard?
- Best cloud architecture, and why move to SaaS
- Data product templates, and why they are useful

# Enterprises are full of “data silos”

## Why?

- Enterprises empower independent business units (IBUs), thereby creating silos
- Otherwise all decisions have to go through “God”
- And business agility goes out the window

# Sources & uses of data have exploded in cloud era

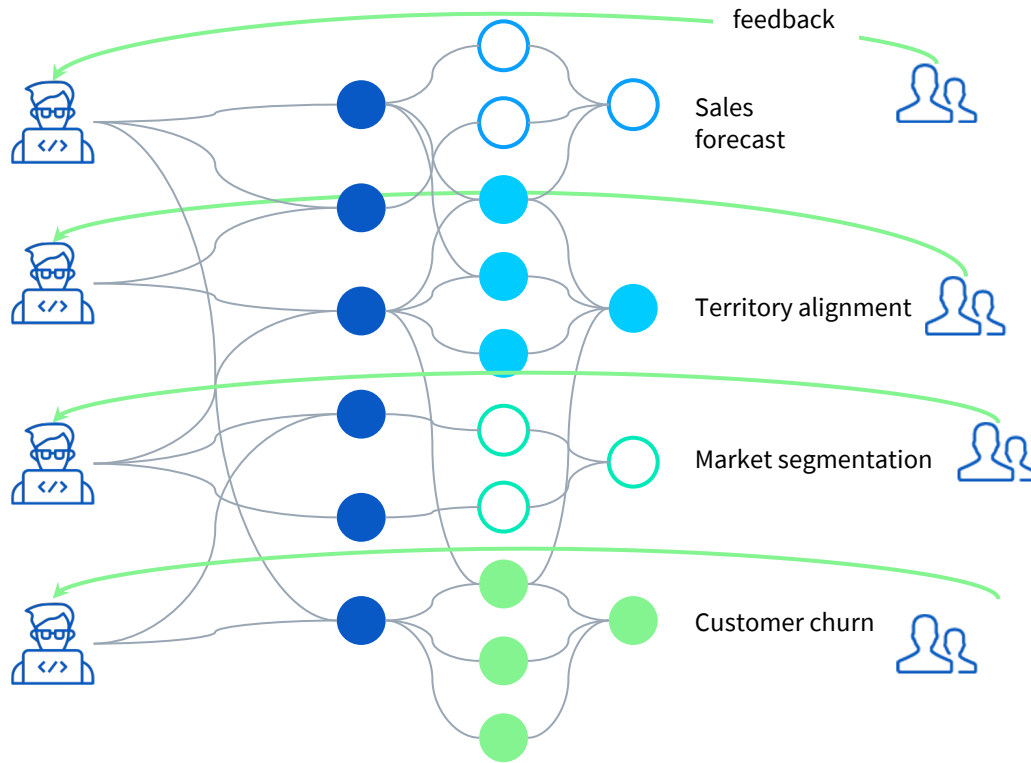


# There is HUGE business value in integrating the silos

- Cross selling between IBUs (integrate customers)
- Reduce regulatory risk (know your customer!)
- Get the best price across IBUs (integrate suppliers)
- Sharing spare parts between IBUs (integrate parts)



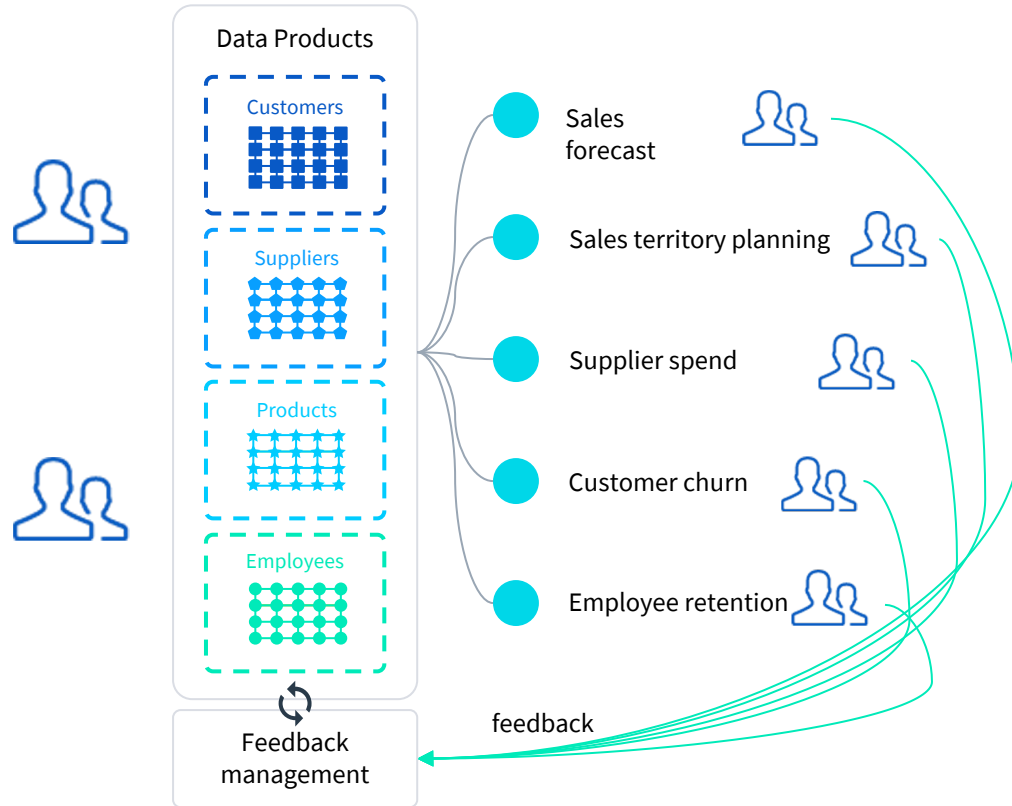
# Need to move away from 'use case' approach



- **Limited reuse**; difficult to increase volume of projects without increasing size of team
- **Unpredictable timelines** for delivering insights; reduces trust of stakeholders
- **Difficult to respond to feedback**, since few people understand logic and process & tooling for feedback are poorly defined

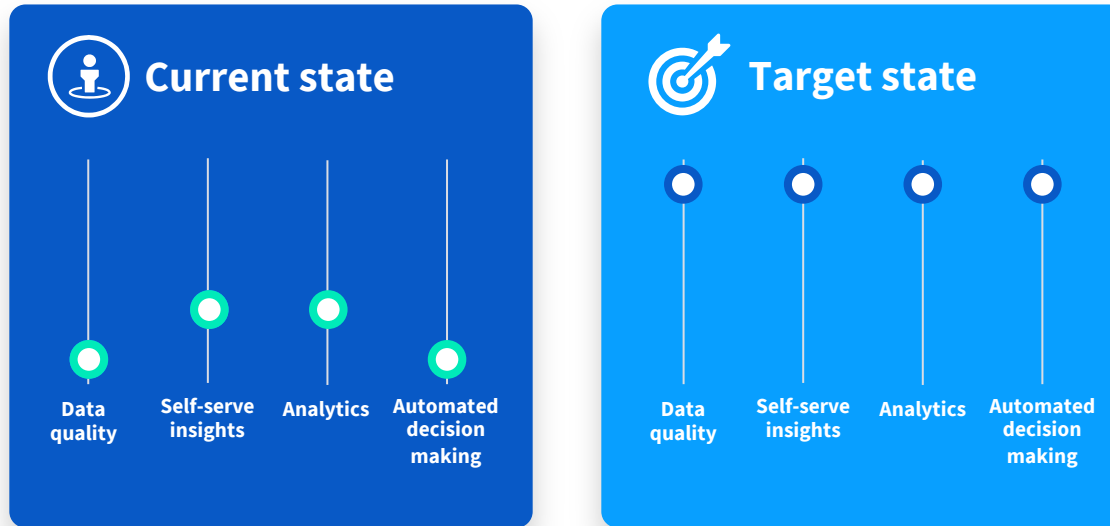


# Towards entities as key organizing principle



- **Drive reusability** by continuously monitoring & improving a set of foundational data assets that are frequently used (e.g., customer list)
- **Create predictability** by minimizing 'net new' work needed to answer each business question
- **Improve trust** with centralized, closed-loop feedback process to rapidly resolve issues

# How to decide when to maintain, improve, or migrate & sunset?



# Silo Integration

- Need (after the fact) to integrate independently constructed data sets for common entities
- Tattoo this on your brain:

**Independently constructed schemas  
are never plug compatible!!!!**



# What does silo integration entail?

- Move data sets to a common place
  - Think data warehouses ('90s), data lakes ('00s)
- Perform transformations
  - To get data into a common units and meaning
- Perform schema integration
  - Line up the various columns that mean the same thing
- Perform cleaning
  - E.g. -99 often means null
- Perform enrichment
  - Add more joining attributes to make next steps easier

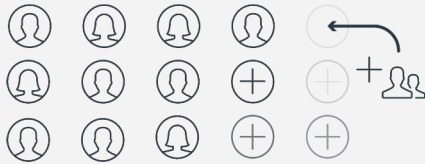
# What does silo integration entail?

- Perform entity consolidation
  - Consolidate duplicates
- Find golden values for clusters of records for each entity
  - E.g most frequent value
- Perform classification
  - E.g. classify suppliers as international or local
- Perform ongoing stewardship
  - On updates over time



# “Best practices” for solving the problem

## Throw people at it



Manual data curation

Integrating and maintaining an array of tools

Building custom solution

Limited sources/data

**Costly & slow**

## Lock it down



Rule-based and manual in nature

Limited sources

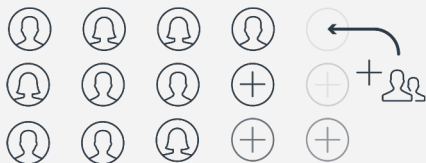
Static data

Platform play

**Costly & ineffective**

# AI enables a new modern approach

## Throw people at it



Manual data curation

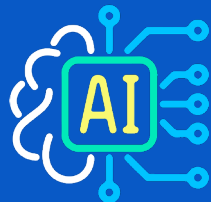
Integrating and maintaining an array of tools

Building custom solution

Limited sources/data

**Costly & slow**

## New approach



Trained models

Built-in Data quality & Enrichment

Millions of data records

Continuously updating

**Fast time-to-value**

## Lock it down



Rule-based and manual in nature

Limited sources

Static data

Platform play

**Costly & ineffective**

# Poll: How mature are your AI capabilities?

- A. We're still just learning about the technology
- B. We have some active POCs
- C. We have a few things in production that use AI but it's not critical to our business
- D. We have many business-critical production applications and/or a team focused on implementing AI in business-critical applications



# Why don't traditional solutions scale?

- Media company
  - Wrote 200,000 rules!!!!!!
  - In a home-brew rule system
  - Took 26 person years (think: a cost of \$5M)
  - Over 13 calendar years
  - Unmaintainable!!!!
    - Takes forever to add a new source
    - Or change anything

# Computation and the Cloud

- Essentially all (possible) applications will move to the cloud over time
  - Decision support first
  - Cobol later to never
- To get lower cost
  - Dewitt vignette
  - Hamilton vignette
- And elasticity!!!

# Data Mastering

- Is an ideal cloud application
  - Resource needs vary
  - Data intensive
  - Data may well already be on the cloud

# Data mastering cloud architecture

- Lift and shift
  - Please don't do this for anything!!!
  - You have a once in a generation opportunity to restructure your applications. Please fix the “sins of your predecessors”
  - Your successor will appreciate it!
- Platform as a service
  - Widely supported
  - But does not get you elasticity
- Software as a service
  - Gets you elasticity
  - Far and away – the best option

# Data product (templates)

- Most mastering projects entail common “entities”
  - Customers
  - Suppliers
  - Parts
  - Projects
  - ...
- Wouldn't it be nice to start with a “template” for your current entity?
  - Schema
  - Pre-build data cleaning routines
  - Pretrained model
  - ...

A “data product” is a consumption-ready set of high-quality, trustworthy, and accessible data that people across an organization can use to solve business challenges.

# Holistic view of key business entities

Healthcare  
Life Sciences  
Retail  
Financial  
Services  
Insurance  
CPG  
Software/Tech  
Manufacturing



## Customers



Google Inc.

1600 Amphitheatre  
Parkway  
Mountain View  
CA 94043  
+1 (650) 253-0000  
[www.about.google.com](http://www.about.google.com)

Contacts  
Purchase history  
Parent hierarchy

## People



Ben Green

Apartment 52  
125 Old Broad St,  
London  
EC2N 1DW  
+44 (750) 3690-430  
bgreen@gmail.com

Orders  
Touchpoints  
Household

## Suppliers

Google Inc.

1600 Amphitheatre  
Parkway  
Mountain View  
CA 94043  
+1 (650) 253-0000  
[www.about.google.com](http://www.about.google.com)

Contacts  
Purchase history  
Parent hierarchy

## Providers

Google Inc.

1600 Amphitheatre  
Parkway  
Mountain View  
CA 94043  
+1 (650) 253-0000  
[www.about.google.com](http://www.about.google.com)










Contacts  
Purchase history  
Parent hierarchy

demo.tamr.cloud

Home

### ADD DATA PRODUCT

What best describes your data?

-  B2B Customers
-  B2B Customers with D&B
-  B2B Customers with Firmographics
-  B2C Customers
-  Contacts
-  Healthcare Providers
-  Legal Entities
-  Patients
-  Suppliers with D&B

CANCEL NEXT

Your Access Level: Editor  
Last Run: 16 hours ago  
Owner: chad

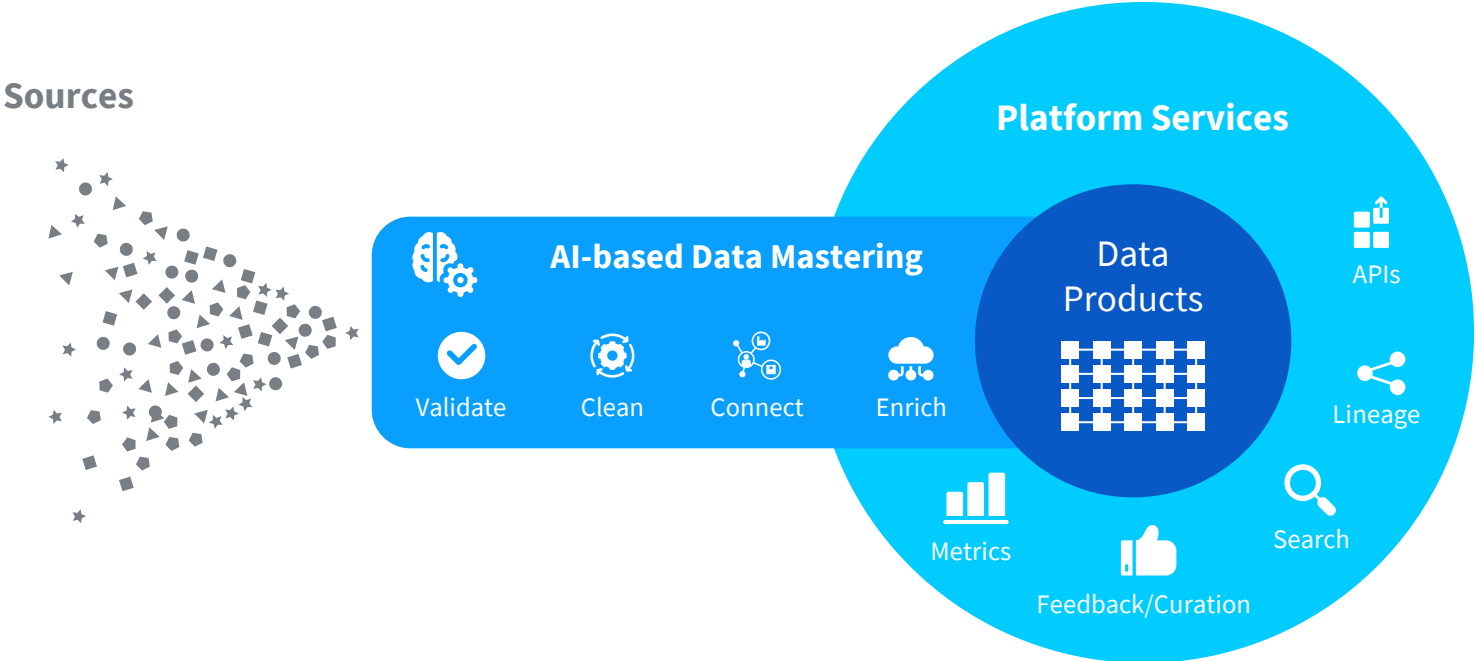
+ ADD DATA PRODUCT



# Poll: How familiar are you with data products?

- A. It's still a new concept
- B. Familiar with the concept but have not tried to establish one
- C. We are doing POCs now
- D. We have multiple data products in production

# Establishing a data product platform increases chances of success



# Takeaways and Lessons Learned

---

**Data silos are a pervasive and difficult problem in the enterprise**

---

**Traditional, rules based approaches to consolidating data fail**

---

**AI on the cloud is key to bringing data together in a scalable way**

---

**A data product templated approach allows you do get started quickly**

# Questions?



**EDM** Webinar 

---

WEBINAR SERIES

# Evolution of Data and AI



January 30th @ 11 - 12ET

## Part 2: Innovations in Data and AI in 2024 and Beyond



*SCAN ME*



Dr Micheal Stonebraker



Randy Bean



Salema Rice



Anthony Deighton



# Join EDM Council and our membership community of companies...



The screenshot shows the EDM Council website homepage. At the top, there is a navigation bar with links for Membership, Frameworks, Training, Engage, Innovation, About, Sign in, and Join now. The main header features the EDM Council logo and a large title: "Global Advocates for Data & Analytics Management". Below the title is a sub-header: "The leading global trade association providing best practices, standards and education to data and business professionals in our data-driven world." A "What we do" button is visible. On the right side, there is a "TODAY'S HIGHLIGHTS" section with three items: "Bank of Valletta becomes the newest member to join the EDM Council", "EDM Council welcomes Webber Wentzel as its newest member", and "Lion Group joins EDM Council as its newest member". At the bottom of the screenshot, there is a section titled "Join a vibrant community of 25,000+ business leaders, CDOs, and data and analytics professionals across all industries." followed by logos for Bank of England, NOVARTIS, HSBC, AWS, Schneider Electric, Microsoft, Google, and AEGON. Two buttons are present: "Explore membership" and "See all 350+ member organizations".



**350+ Member Firms**

Cross-industry,  
including Regulators



**25,000+**

Professionals



**Worldwide**

Americas, Europe,  
Africa, Asia, Australia

[edmcouncil.org](https://edmcouncil.org)



EDM Webinar 

Thank you!

