Decoding Data Quality with Data Products

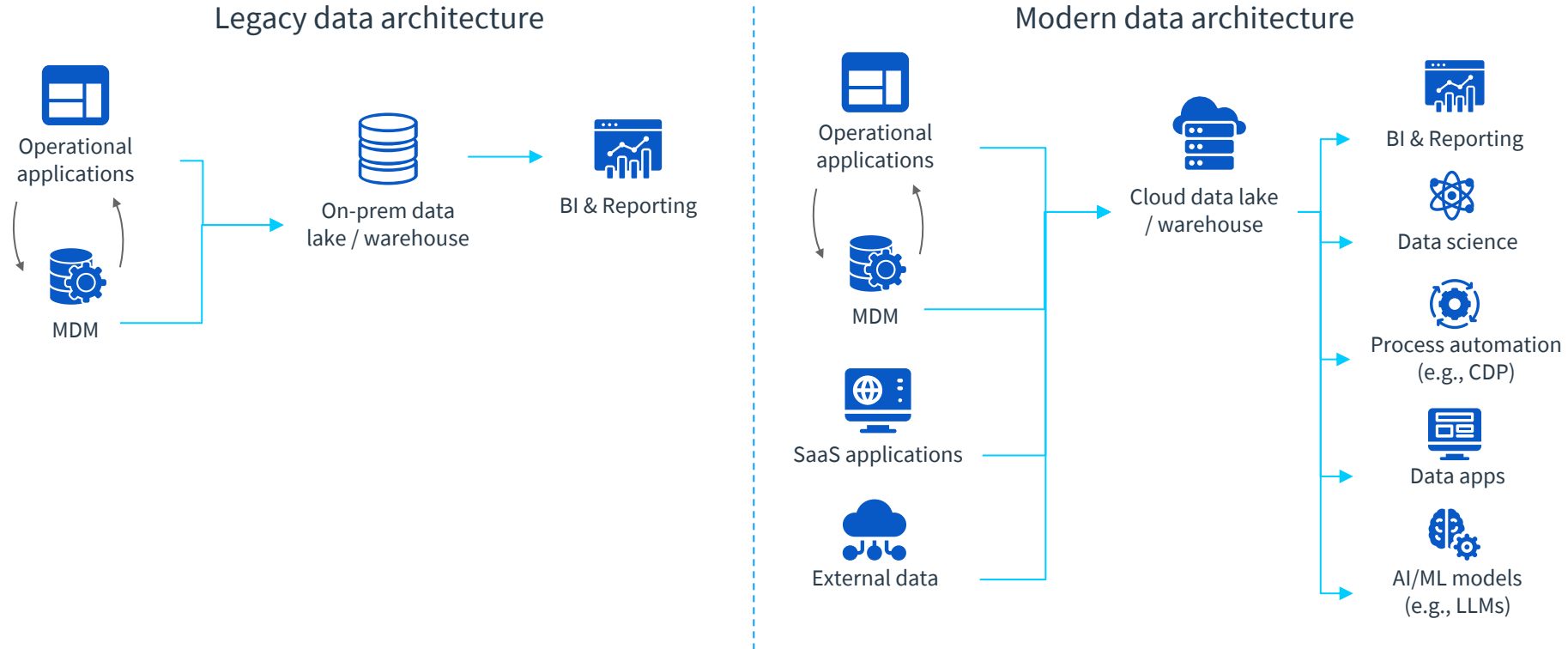# Poll: How familiar are you with data products?

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

Please enter your answers in the chat!

Tamr

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.

Sources & uses of data have exploded in cloud era

Legacy data architecture

Operational applications

MDM

On-prem data lake / warehouse

BI & Reporting

Modern data architecture

Operational applications

MDM

SaaS applications

External data

Cloud data lake / warehouse

BI & Reporting

Data science

Process automation (e.g., CDP)

Data apps

AI/ML models (e.g., LLMs)

Tamr

Being 'data-driven' requires more from data quality

BI & Reporting → **Data needs to be clean & standardized**

Data science → **It should have a lot of attributes**

Process automation (e.g., CDP) → **There should be a 'single view of a customer'**

Data apps → **Issues should be quickly fixed when flagged by users**

AI/ML models (e.g., LLMs) → **Garbage in, reputation damage out**

Tamr

Data teams can't keep up

Data leaders can't scale their teams quickly enough

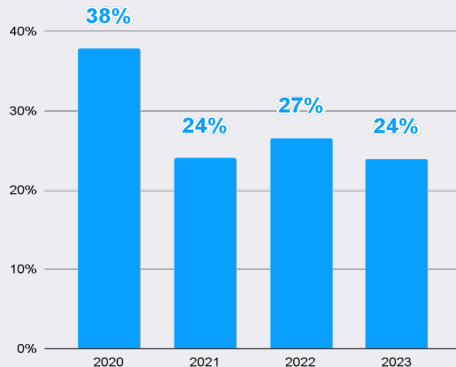Data teams describe the volume of their workload as exceeding their capacity — **84%**

Data teams report their company works with 50+ data sources — **69%**
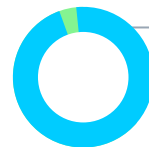
Source: Matillion Data Productivity Survey 2023

Decision makers aren't getting answers fast enough

Data leaders that say their company is data-driven

- 2020: 38%
- 2021: 24%
- 2022: 27%
- 2023: 24%

Source: NewVantage Partners Data and Analytics Leadership Annual Executive Survey 2023

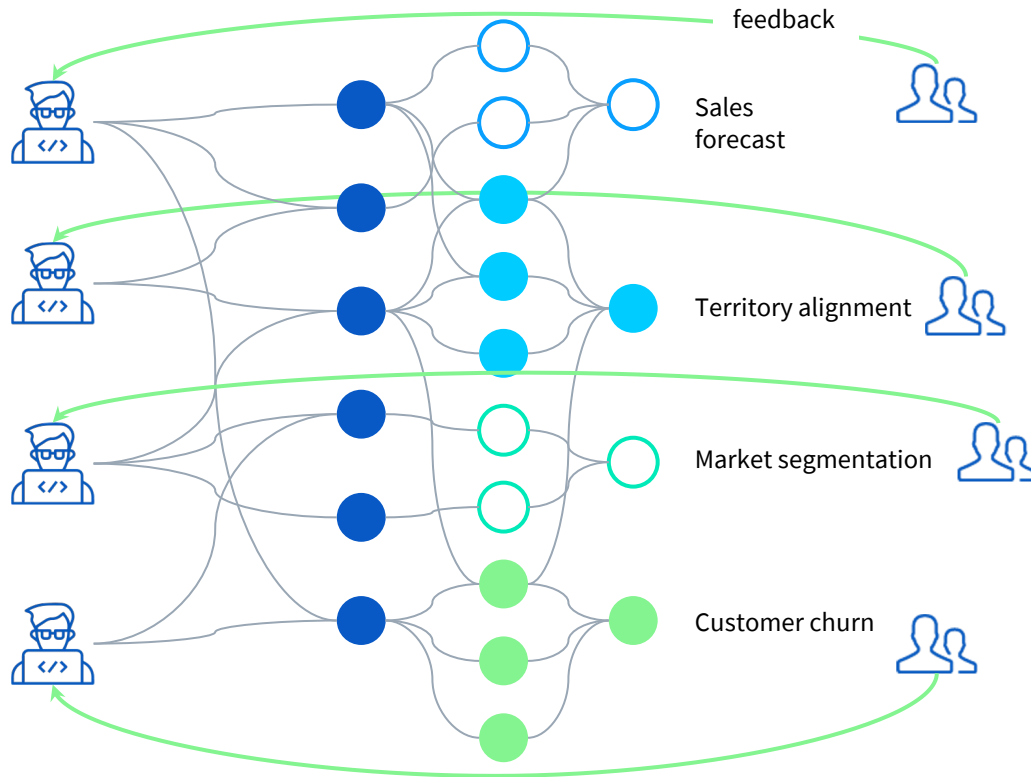### Business leaders want to use AI but don't trust their data enough

94% of business leaders believe AI is critical to their success over the next 5 years

67% of executives are "not comfortable" using data from advanced analytic systems
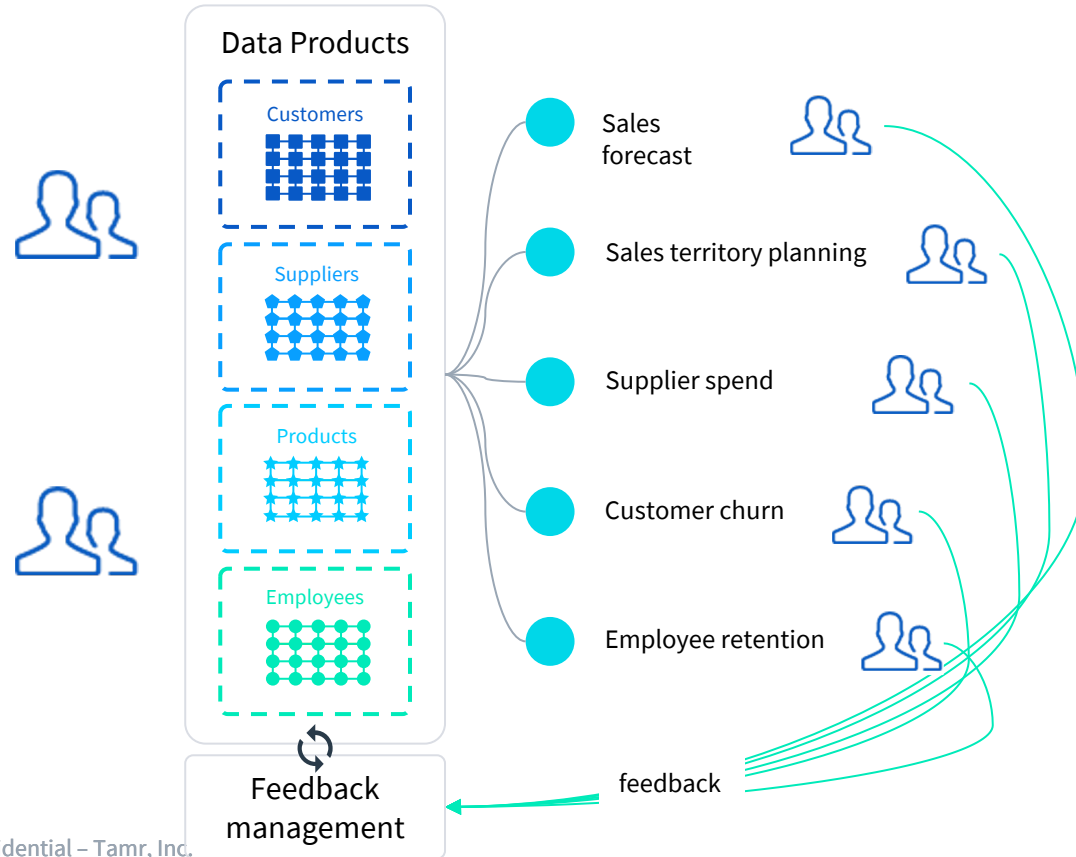
Sources: Deloitte State of AI in the Enterprise Survey, Deloitte Analytics and AI-driven Enterprise Survey

Tamr

Need to move away from 'use case' approach

feedback

Sales forecast

Territory alignment

Market segmentation

Customer churn

- **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 managing data as a product

Data Products

Customers

Suppliers

Products

Employees

Feedback management

Sales forecast

Sales territory planning

Supplier spend

Customer churn

Employee retention

feedback

- **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

What are the 'best practices' to build data products?
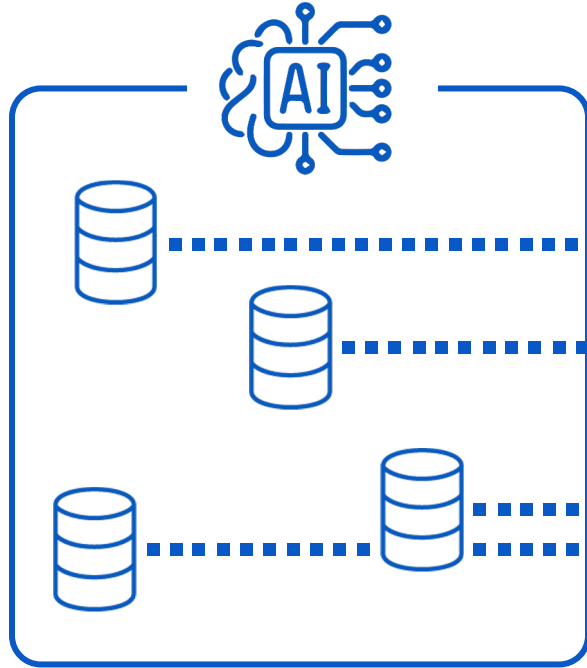
Hire more data engineers & stewards

**Reduce scope of data actively managed**

Add top-down "governance"

AI changes how data products are built & managed

- **Semantic validation and cleaning** reduces volume & complexity of transformations that need to be developed

- **Smart identification of relationships between records** enables data products to include a broad scope of sources

- **Enrichment of sparse data** elevates standard of data quality without heavyweight top-down governance

- **Human-machine feedback loop** builds confidence & trust

Tamr

Establishing a data product platform increases chances of success
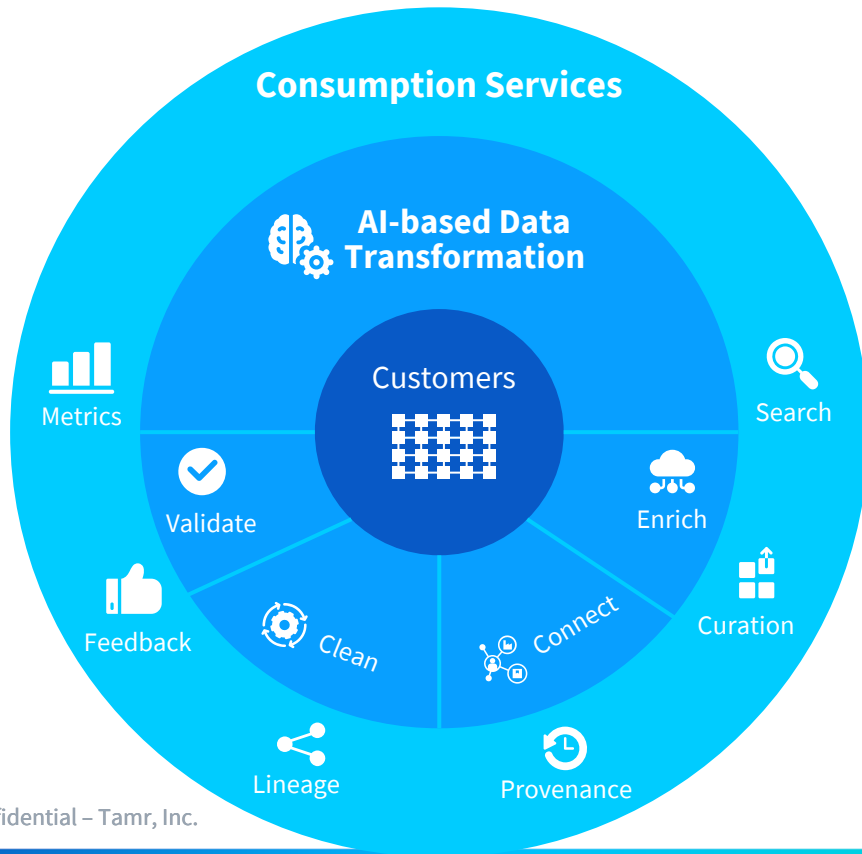
Platform must be designed for collaboration between data & business teams to build trust and ensure data products are fit for purpose

Enables faster value creation

Data teams

✓ Support more 'use cases'

✓ More control & predictability over roadmap

✓ Stay on the 'cutting edge' of data applications

Business teams

✓ Answer business questions faster

✓ Answer new types of questions

✓ Quickly create efficiencies by leveraging new technology (e.g., AI)

" *...revolutionizing our approach to data management - increasing speed, enhancing accuracy, and bolstering confidence in our data management processes and advance our mission of providing the best care possible for our patients".*

Marshall Worster, Senior Director Enterprise Transformation

novocure®

Tamr

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

Please enter your answers in the chat!

Getting started: Data product strategy checklist

Five steps to success

1. Know your why

2. Define your use case

3. Assess your data, your organization and your technology stack

4. Secure buy-in and budget

5. Develop a MVDP: Minimum Viable Data Product

Tamr

Challenge: Aggregating disparate sources

Source-specific standards and identifiers mean data isn't 'analytic-ready' after it has been centralized

SAP

Andy Wyatt
andy.wyatt@gmail.com

342 Assam, Apt 34
EC2B3L
United Kingdom

Monthly Sub: £59
YTD: £422

ID: x24x8310

salesforce

A.J. Wyatt
a.wyatt@tanger

342 Assam Street
Apartment 34
London, UK

Monthly: £42
Total Acc: £384

ID: AB221CC

Excel
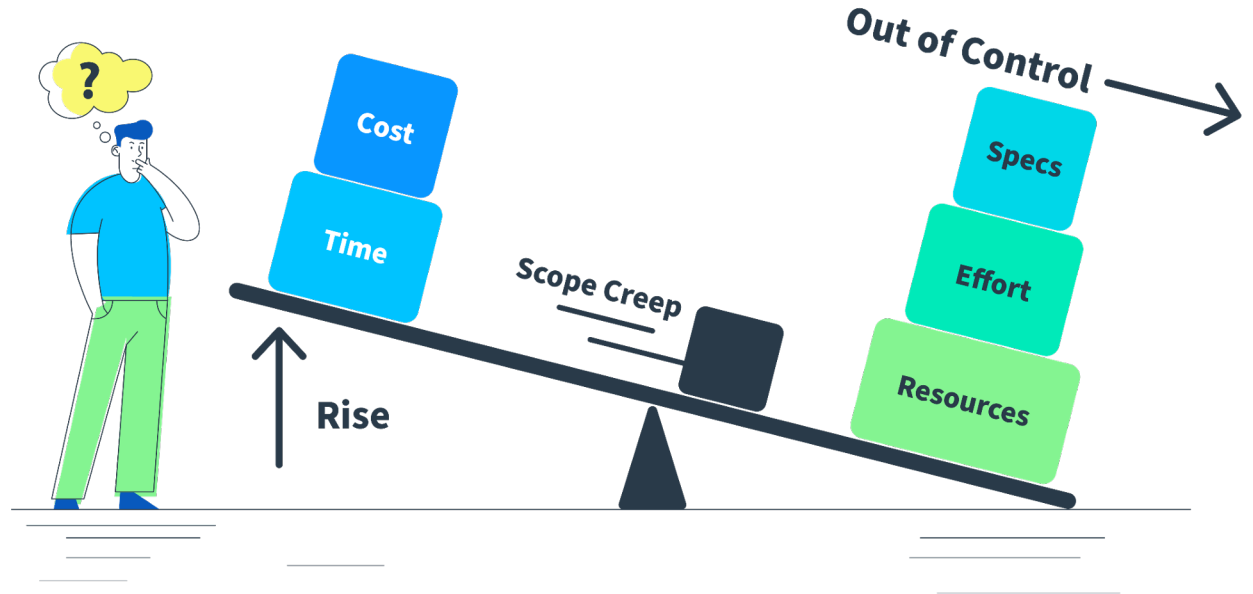
Andrew Wyatt
andy.wyatt@gmail.com

5 Chestnut Rd.
Dublin
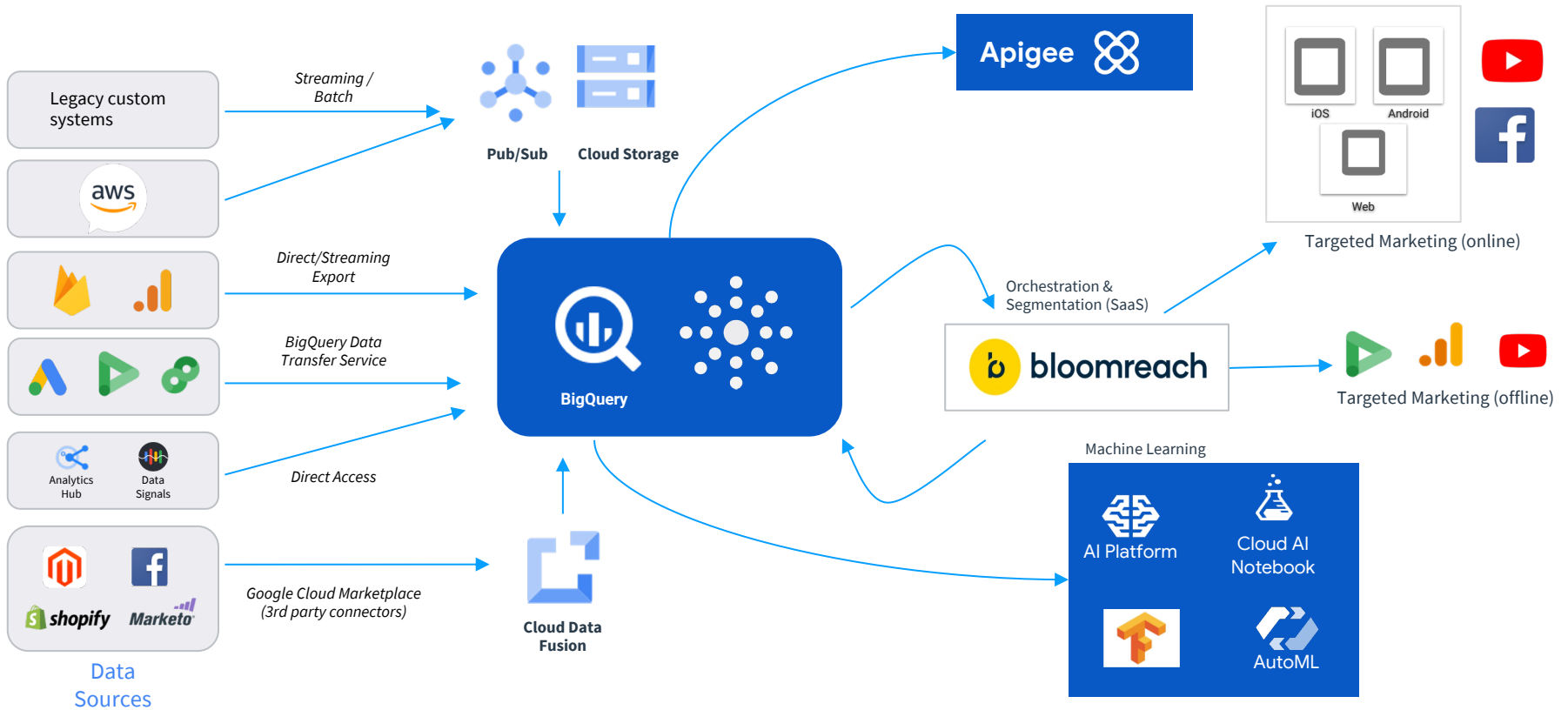Ireland

Products: €36
Total: €288

ID: 893173

Tamr

Challenge: Data vs system integration

Can feel like both are needed to realize the benefits of either

Example Customer: Solution Architecture

Data Sources (left column):
- Legacy custom systems
- aws
- Firebase / Google Analytics
- Google Ads / Display & Video / connectors
- Analytics Hub / Data Signals
- Magento / Facebook / Shopify / Marketo

Data Sources

Connections:
- Streaming / Batch
- Direct/Streaming Export
- BigQuery Data Transfer Service
- Direct Access
- Google Cloud Marketplace (3rd party connectors)

Pub/Sub **Cloud Storage**

BigQuery

Cloud Data Fusion

Apigee

bloomreach — Orchestration & Segmentation (SaaS)

Targeted Marketing (online) — iOS / Android / Web / YouTube / Facebook

Targeted Marketing (offline)

Machine Learning:
- AI Platform
- Cloud AI Notebook
- AutoML

Tamr

First project: Enterprise Datalake

novocure®

Novocure Systems	Google Cloud	SaaS Systems

SAP S/4HANA → Log Files → **Fivetran** → Secure Replication → **Google BigQuery**

Google BigQuery ← Secure Replication ← **Fivetran** ← Data Updates ← **Veeva CRM**

MuleSoft ← Cleaned HCP/HCO Data

IQVIA → **Tamr**

Cleaned Master Records → **Tamr** ↔ **MuleSoft**

HCP/HCO Data MDM

Confidential – Tamr, Inc.
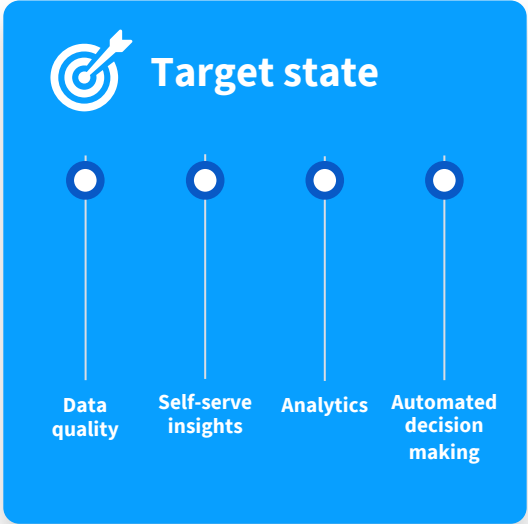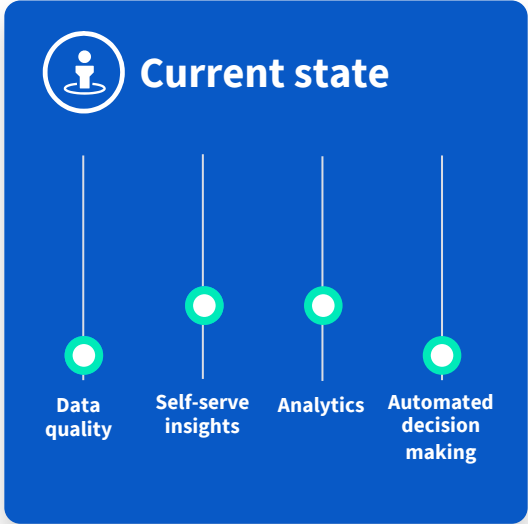
18

Tamr
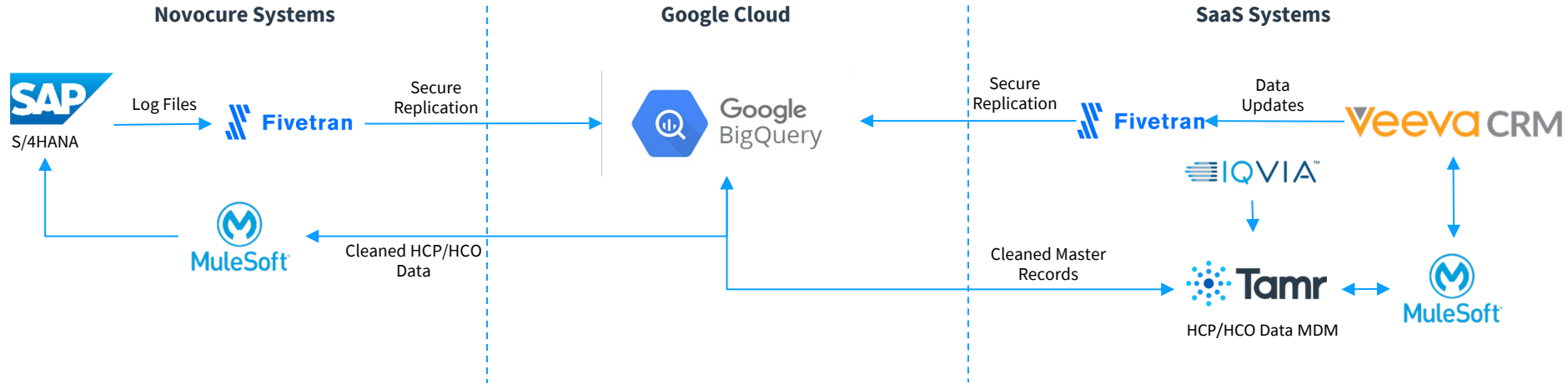
Challenge: Mix of legacy, homegrown, and modern

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

Current state

- Data quality
- Self-serve insights
- Analytics
- Automated decision making

Target state

- Data quality
- Self-serve insights
- Analytics
- Automated decision making

Old Mutual's Journey

1 Had existing own grown MDM system with parts still running on a mainframe
   a. Exported customer views from system to use for analytical workloads
   b. Built reports and business value on top of these

2 Leveraged API endpoints at point of consumption to check for existing entities
   a. Reduced the rate at which bad data was being generated at the the first point it enters the system

3 Switched over to using new SOR
   a. Over time switched systems over to using the new SOR leveraging APIs

OLD**MUTUAL**

Tamr

Real-time Reference Architecture

Data product platform processes
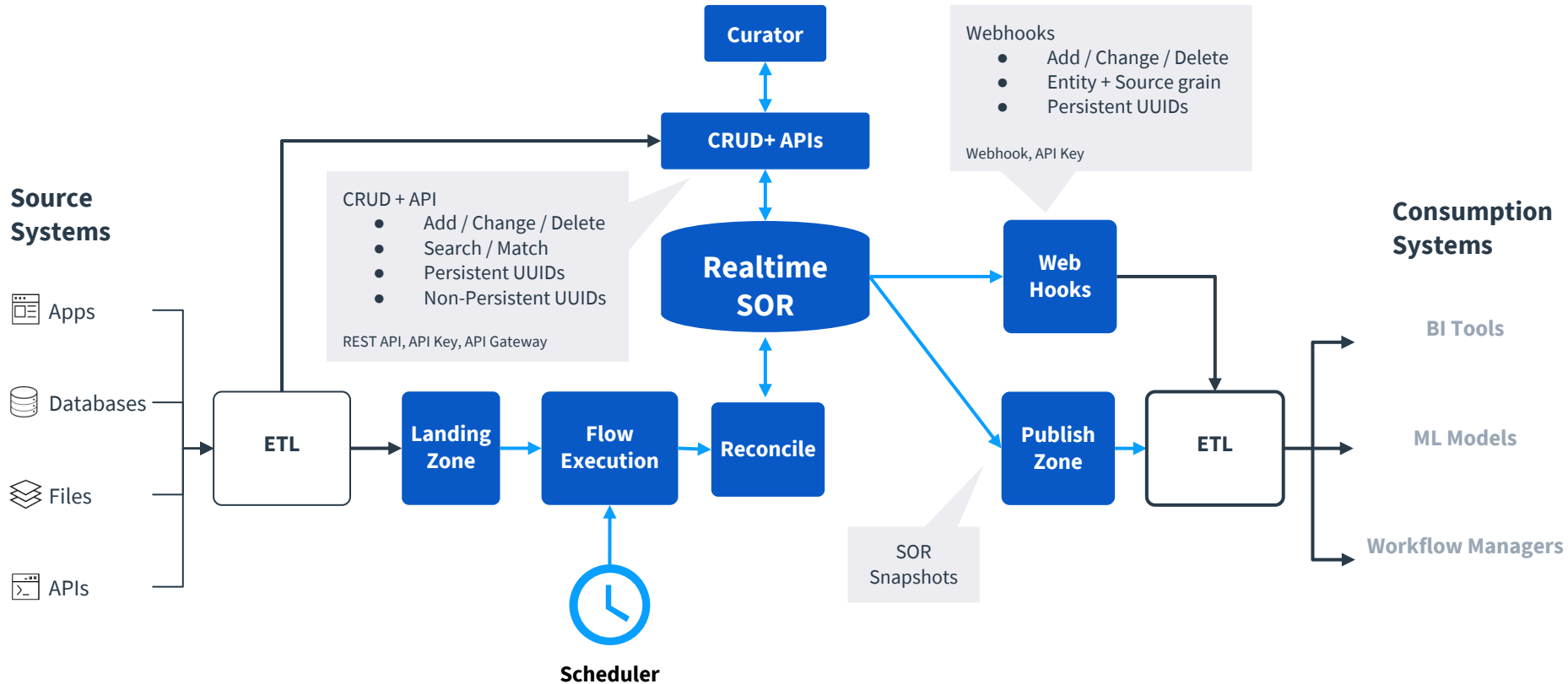
Questions to ask potential partners

Question(s) to ask	• What external data is integrated into the solution? • What capabilities are industry-specific?	• What AI features have they launched? How are customers using them? • What is their architecture for AI capabilities?	• What are the typical roles of people involved in setup? • Who typically uses the solution on an ongoing basis?
What you *really* want to know	• How much are they investing in domain-specific needs?	• Have their products been architected to take advantage of new technology?	• Does this solution require specific skills that you might not have in-house?

Takeaways and Lessons Learned

Prioritize **outcomes** over "stacks"

Human interfaces are just as important as the pipeline

Learning **how** to iterate is critical; get to a MVDP (minimum viable data product) to accelerate learnings

Build vs buy is a false dichotomy. You will be doing both; be deliberate in tradeoffs

Q&A

Tamr