Delivering a Continuous Enterprise Data Pipeline

Scott W. Ambler

Data Methodologist | Author

Ambysoft Inc.

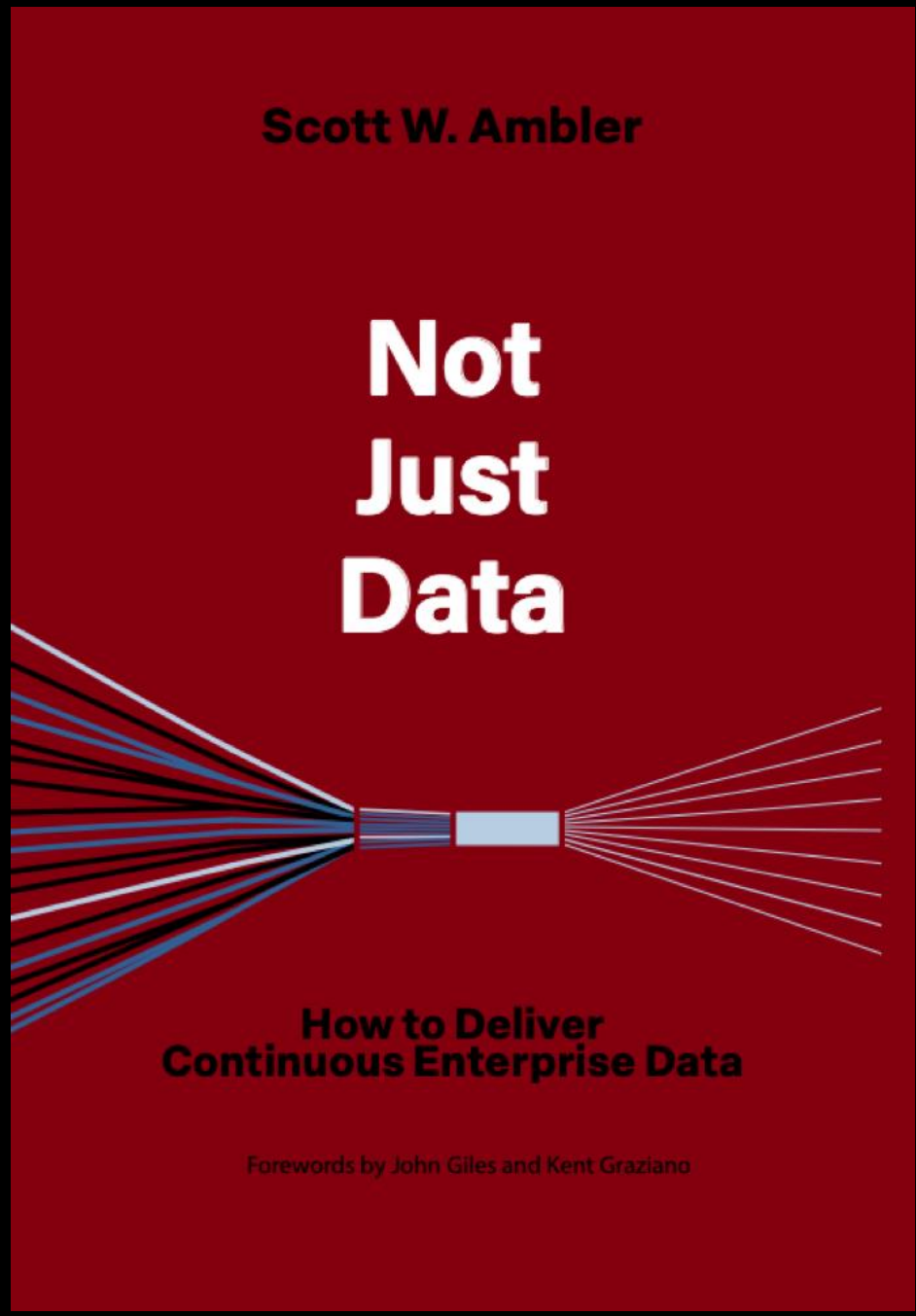
Scott Ambler

- scott@scottambler.com
- linkedin.com/in/sambler

Ambysoft	Data Methodologist, Board Advisor Ambysoft.com
Agile Data	Thought Leader AgileData.org
Agile Modeling	Thought Leader AgileModeling.com
PMI Disciplined Agile	Co-Creator pmi.org/disciplined-Agile

Source Material for this Presentation

ScottAmbler.com/not_just_data/



Agenda

- Today's takeaways
- Data debt
- Industry statistics
- Data is the new water
- A continuous enterprise data pipeline
- Fixing the f***ing data
- Parting thoughts

Today’s Takeaways

Data debt is a significant and underappreciated issue within many organizations

A continuous enterprise data pipeline is needed to deliver the information required for better decision making and AI

The best place to address DQ issues is at the source

Definition: Data Debt

Technical debt is the accumulation of defects, quality issues (such as difficult to read code or low data quality), poor architecture, and poor design in existing solutions

Data debt (data technical debt) refers to quality challenges associated with legacy data sources, including both mission-critical sources of record as well as “big data” sources of insight.

Source: AgileData.org/essays/dataTechnicalDebt.html

Causes of Data Debt

- Business prioritizing time to market over quality concerns
- Manual data entry
- Multiple siloed sources
- Lack of input validation
- Inconsistent data collection methods
- Inconsistent business rules across applications
- Ineffective data management
- Weak data literacy

Why Should We Care?

Clean data is required for:

- Better decision making
- Better business processes
- Building effective artificial intelligence (AI) solutions
- Running accurate AI solutions

The Impact of Data Debt

15-25%
of revenue lost due to bad data in most orgs (Thomas Redman, 2017)

$12.9 million
The average annual cost of poor-quality data (Gartner, 2021)

33%
of executives trust their data to derive value from it (Accenture, 2019)

46%
of organizations report that data quality and consistency is a challenge for their AI projects (K2View, 2024)

Source: AgileData.org/essays/impact-of-poor-data-quality.html

Data is the New Water

When water is dirty, we can:

1. Filter it just before we drink it
2. Filter it coming into our home
3. Filter it before it is dumped into our water supply
4. Filter it at the source
5. Clean and fix the actual source

Source: AgileData.org/essays/data-quality-metaphor.html

Data is the New Water

When data is dirty, we can:

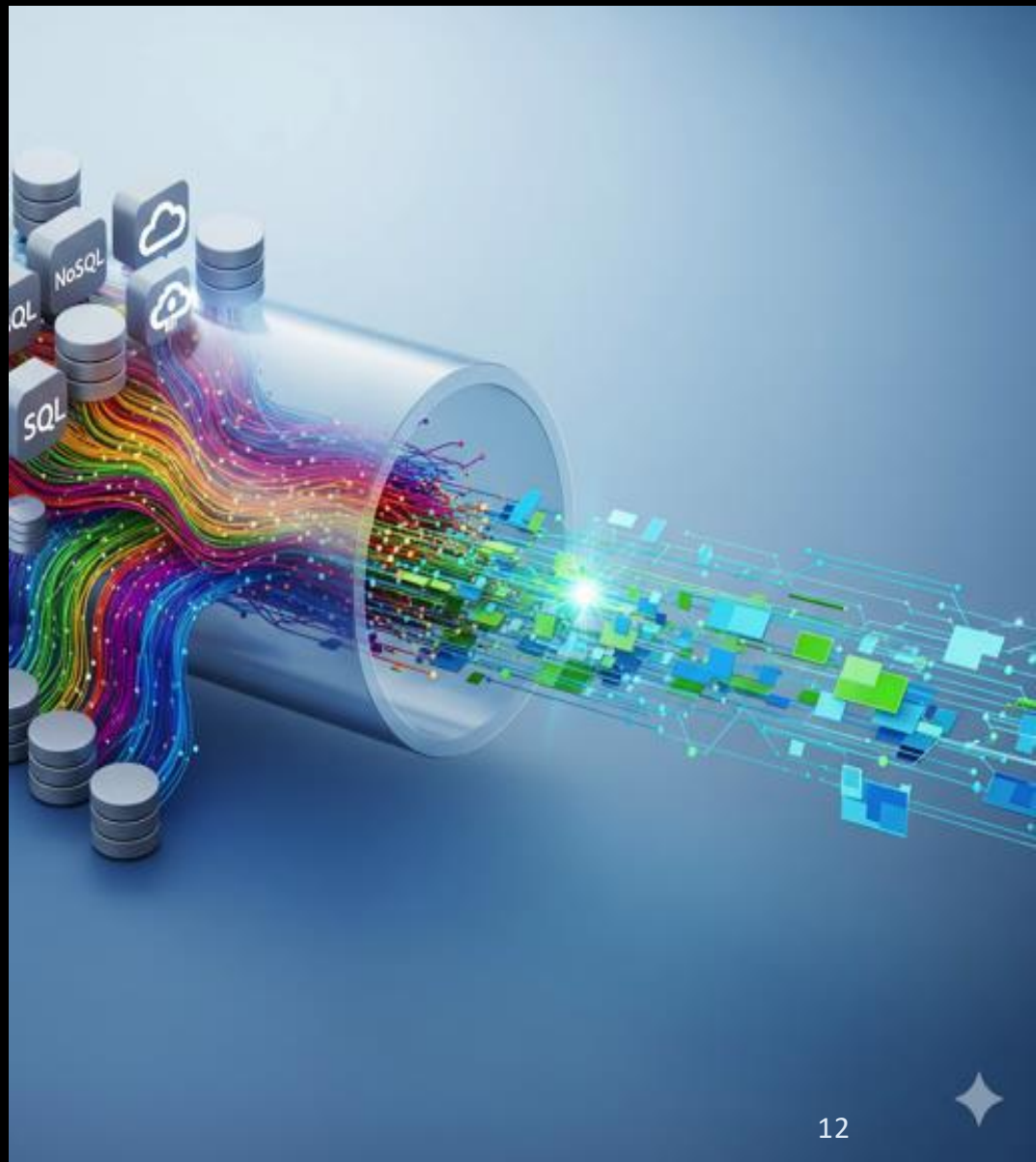
1. Cleanse data at point of use for a function
2. Cleanse data at point of use for a system
3. Create a copy of the data and cleanse that
4. Wrap the data source with a data contract or API
5. Refactor/repair the data source and any systems updating it

Source: AgileData.org/essays/data-quality-metaphor.html

Continuous Enterprise Data Pipeline

Continuous enterprise data is the delivery of high-quality data and information at scale as needed. Updates happen in real-time, or near real-time as changes are made available from data sources. The delivered data reflects the evolving needs of stakeholders as those needs emerge.

A *continuous enterprise data pipeline* gets the right data into the right hands at the right time. This provides end users with the data and information that they need to make data-informed decisions and AI-based systems with the data they require to make predictions.

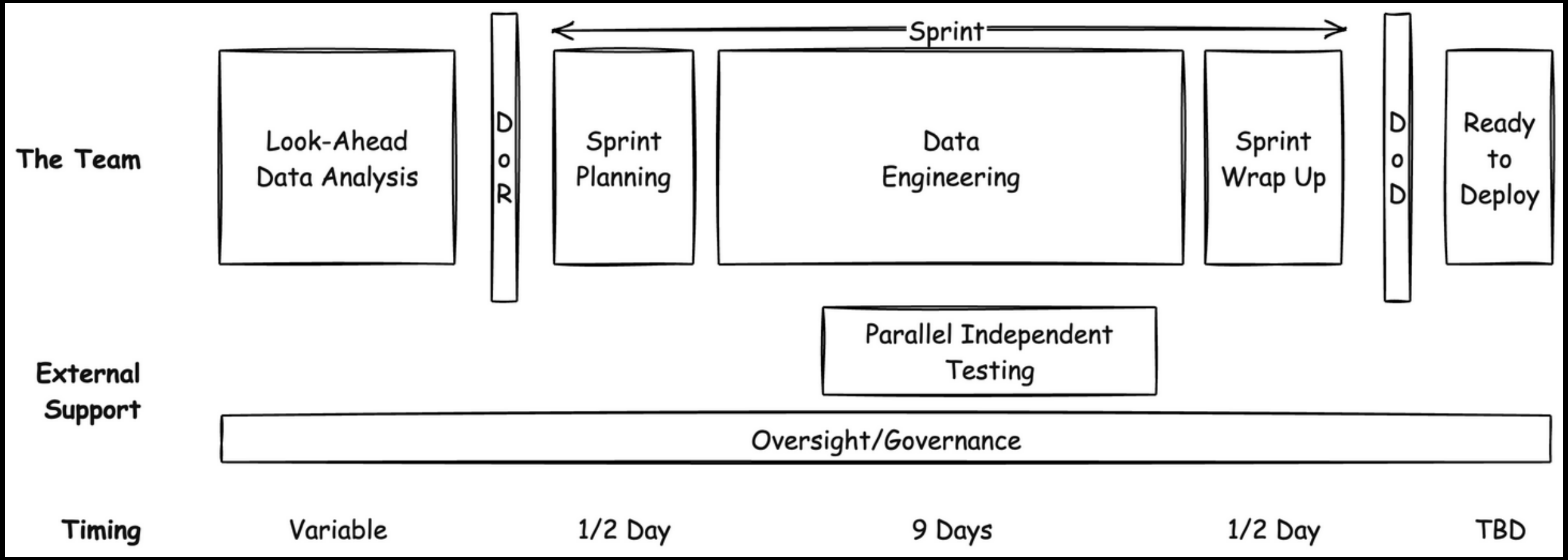


Continuous Enterprise Data Pipeline: Way of Thinking (WoT)

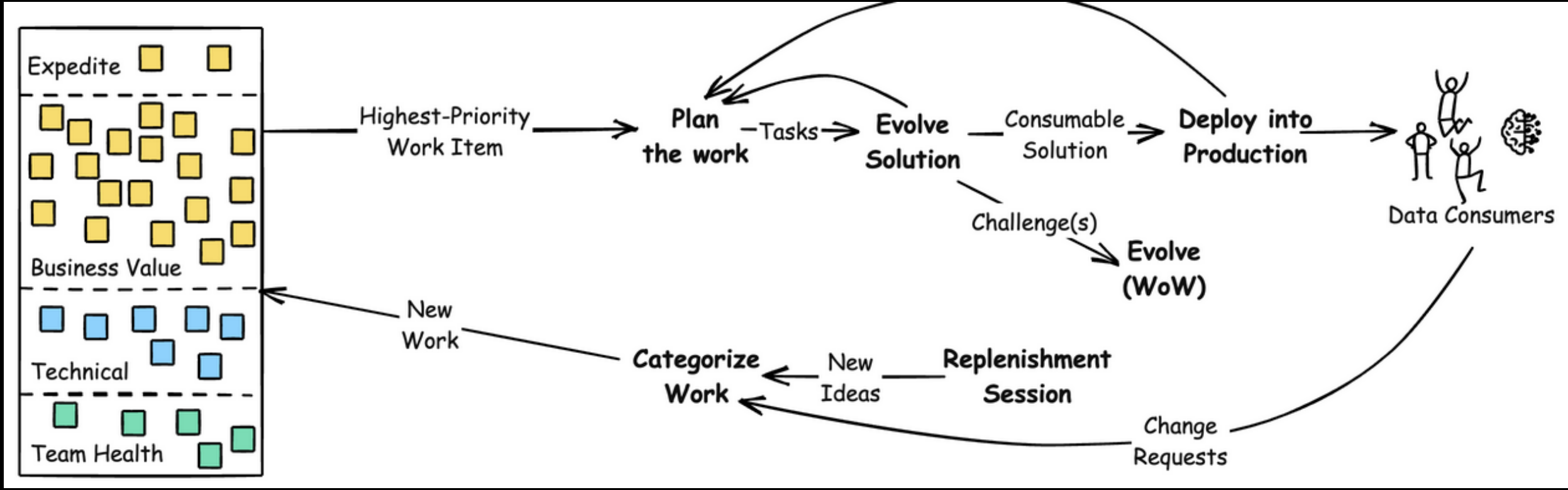
Principles:

1. Be enterprise aware
2. (Meta) data is an enterprise asset
3. Embrace evolution
4. Small things should be quick
5. Fit-for-purpose way of working (WoW)
6. Collaborate closely

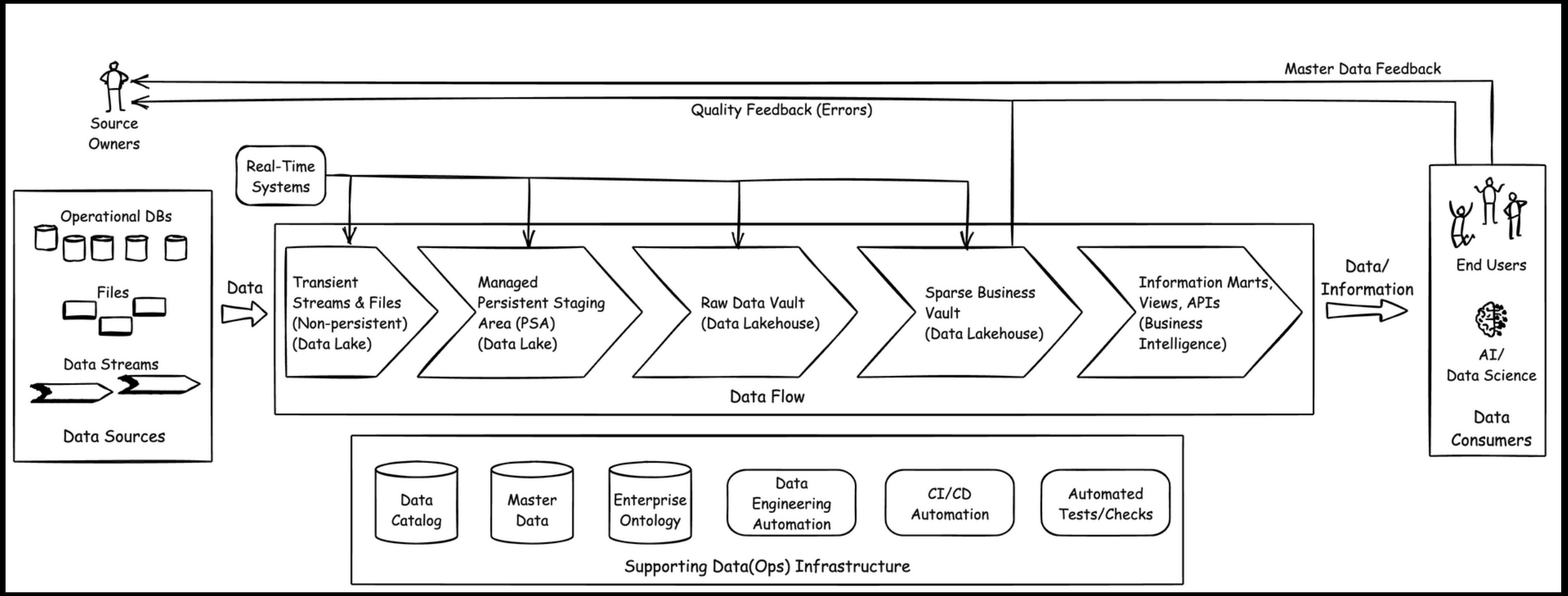
Agile Way of Working (WoW): A Great Start



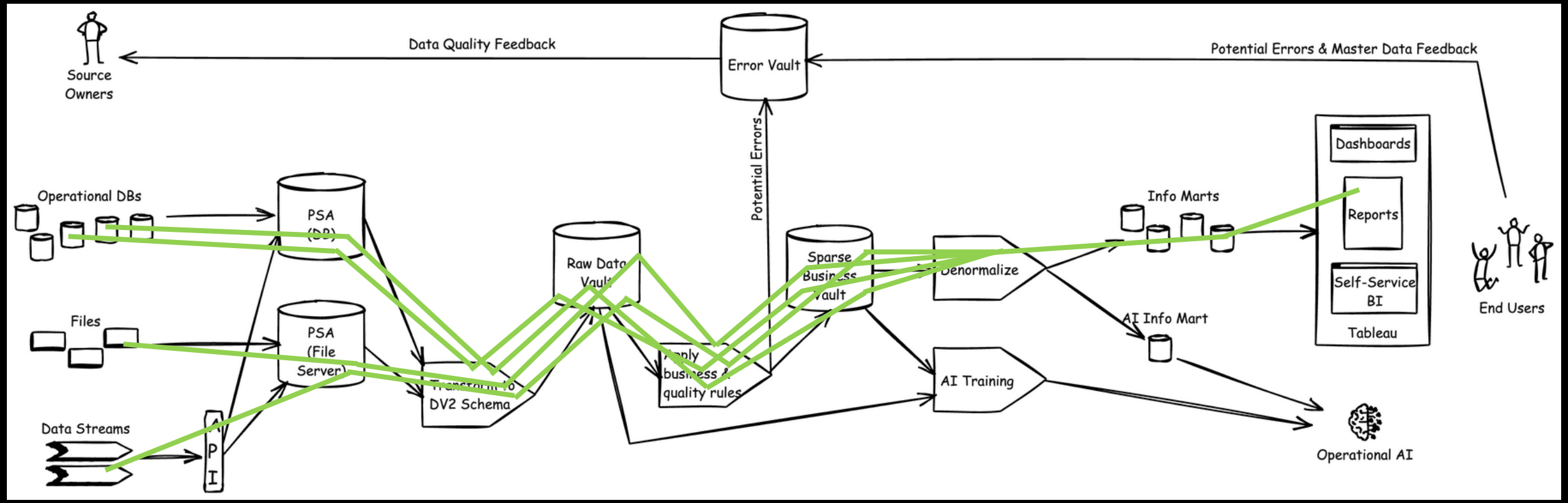
Beyond Agile: Continuous Delivery/DataOps



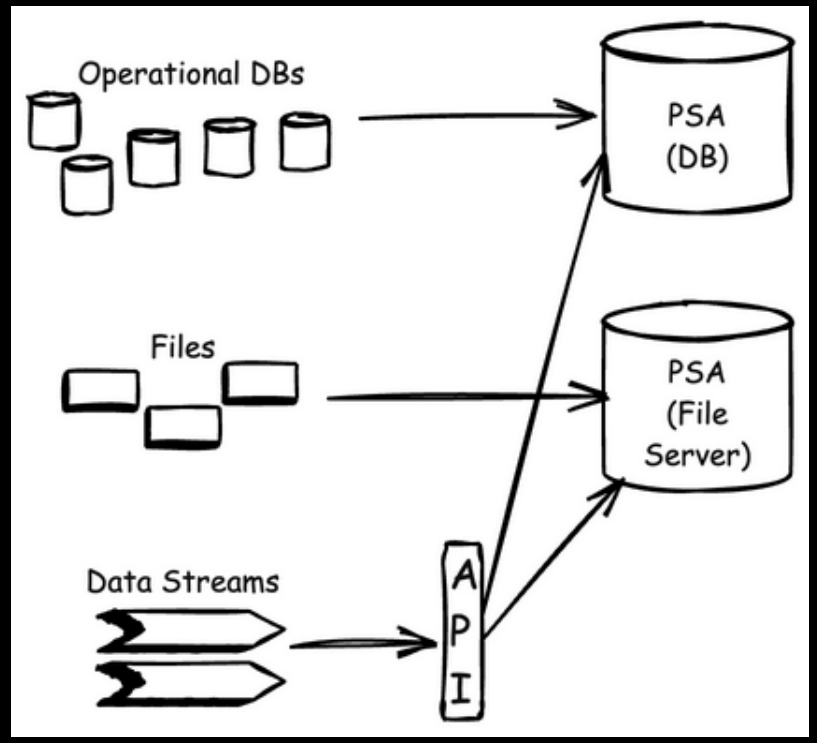
Continuous Data Pipeline: Logical Architecture



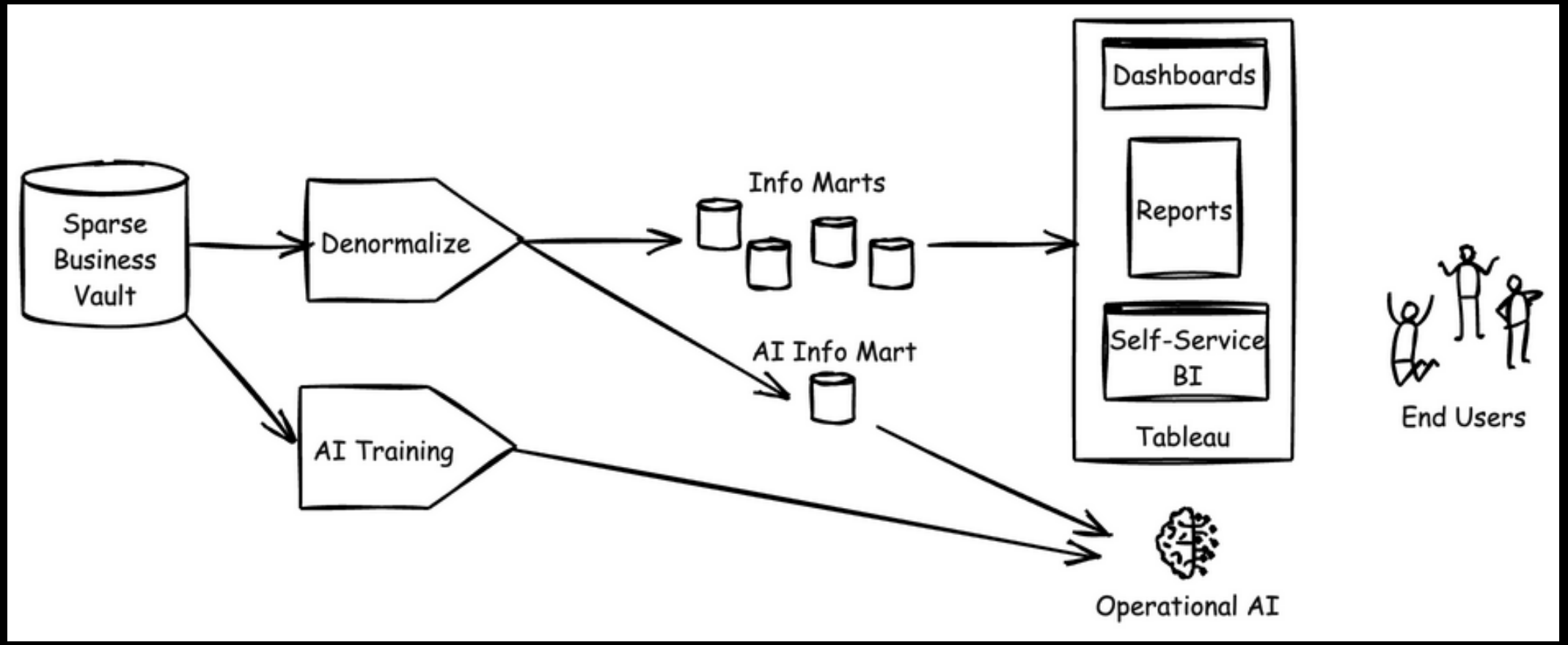
Critical Feature: Defensibility/Auditability



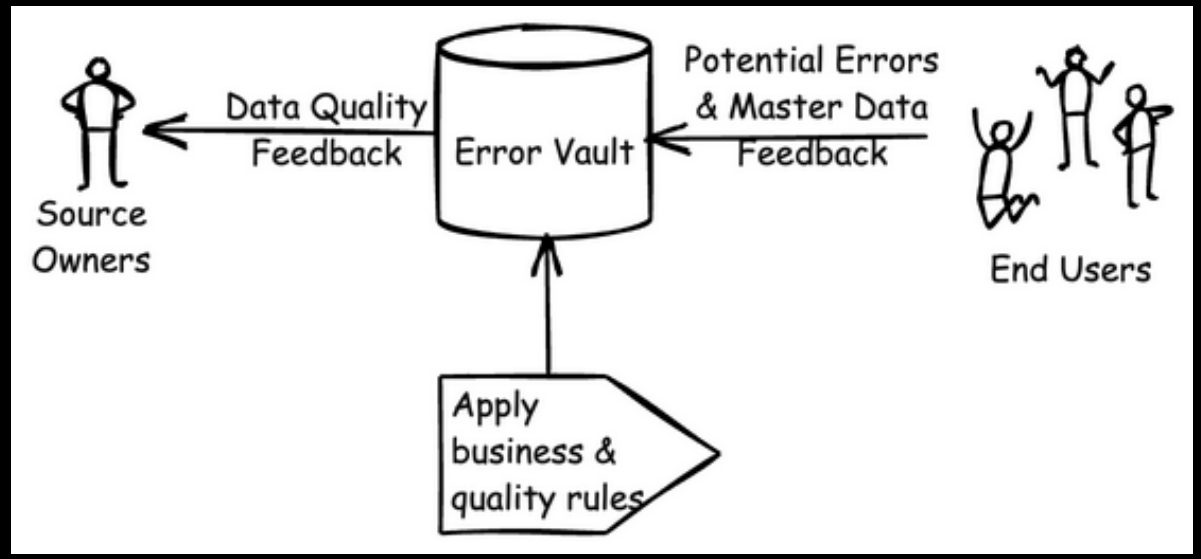
Critical Feature: Handling Disparate Data at Scale



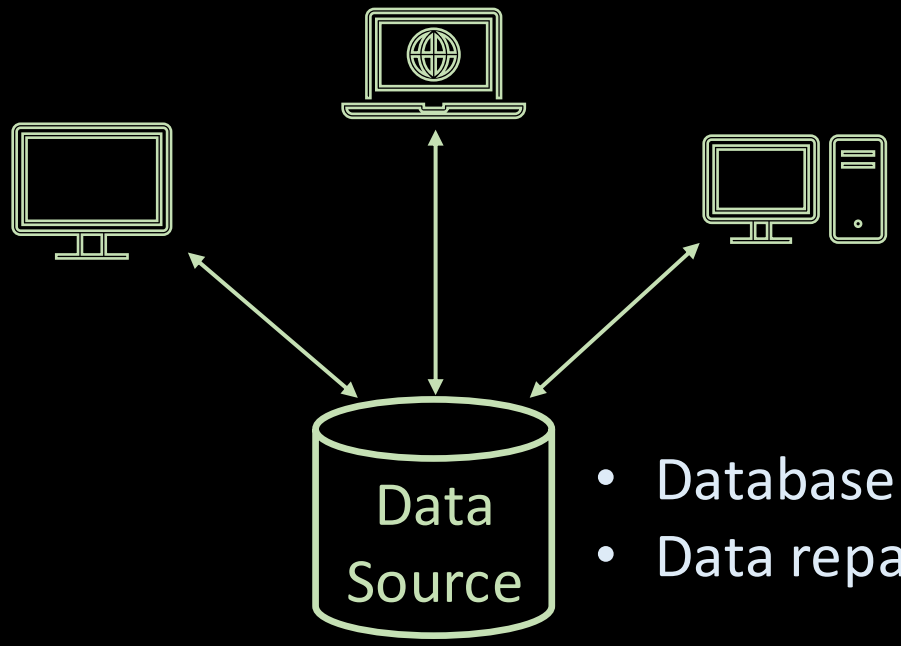
Critical Feature: Multiple Access Points



Critical Feature: Error Reporting



Fixing the F***ing Data: Technical Solutions



- Application/code refactoring

- Database refactoring
- Data repair

- AgileData.org/essays/databaseRefactoring.html
- AgileData.org/essays/data-repair.html
- Refactoring.com

Fixing the F***ing Data: Overcoming the Excuses

We don't have the time	➔ You're likely fixing the data already, why not do it once at the source?
We don't have the skills	➔ Then invest the time to learn them
We have more important things to do	➔ Do you?
My team doesn't own the data	➔ Another team does, work with them
Our company doesn't own the data	➔ Another organization does, work with them

Parting Thoughts

Data debt is a significant and underappreciated issue within many organizations

A continuous enterprise data pipeline is needed to deliver the information required for better decision making and AI

The best place to address DQ issues is at the source

Thank You!

- scott@scottambler.com
- linkedin.com/in/sambler/