data.world
The Cloud-Native Data Catalog

Agile Data Governance

Why Modern Data Challenges Require a New Approach to Governance

Forward

We founded data.world to help the world adopt and improve data-driven decision-making and data literacy. We've seen firsthand how this substantially increases both corporate performance and democratization. And we've seen countless initiatives stall out or fail to launch, struggling with workflow between data producers and consumers, reuse and reproducibility, data literacy, security, and privacy and correctness concerns.

Enterprises waste millions of dollars on failed data initiatives because they apply outdated thinking to new data problems. This results in overly-complex, rigid processes that benefit the few, but make the rest of us less productive.

Fortunately, there's a new way to think about and run data analytics programs, and it starts with a modernized approach to data governance: Agile Data Governance.

Agile Data Governance is the process of creating and improving data assets by iteratively capturing knowledge as data producers and consumers work together so that everyone can benefit. It adapts the deeply proven best practices of Agile and Open software development to data and analytics.

In this report, we'll take you through the following:

- Lessons learned from software development history
- Defining Agile Data Governance
- Key roles and the process
- How Agile Data Governance removes five key barriers to data-driven culture
- Principles of Agile Data Governance

We believe this methodology is the fastest route to true, repeatable return on data investment. Please let us know what you think!
You can email me directly at jon@data.world.

Sincerely,

Jon Loyens

Jon Loyens
Chief Product Officer & Co-Founder
data.world

Lessons from the history of software development

People have touted the advantages of being data-driven, the "big data revolution," and ML/AI for years. Very few organizations make it to that promised land. Most end up stuck on a treadmill of building out new infrastructure or deploying the shiniest new self-serve BI or data science tool. But few big data initiatives earn adoption. This is because creating a data-driven culture is not a technology problem. It's a people problem. Many before us have made this important point.

The data and analytics industry today reminds us a lot of the time before the dot-com crash. Back then every company thought they had to transform into a tech company. Today, it's rare to meet a company that doesn't, at some level, consider themselves a big data company.

Twenty-five years ago "tech-native" companies were eating the world, and every other company raced to get online. They hired teams of engineers, contractors, and consultants. Vendors rushed in to "support" them with all types of expensive technology they hyped as silver bullets. But most software projects went nowhere. According to a 1995 report from The Standish Group, 31% of projects were canceled before completion, while only 16% of projects were "completed on-time and on-budget." These failures, wasting valuable time and money, were just as responsible for the dot-com crash as the hubris and mismanagement of the dot-coms themselves.

From the ashes emerged a new way of thinking about software engineering and project delivery. Two key movements changed everything: Agile and Open Source.

Rise of Agile Software Development

Agile took direct aim at the overruns that plagued the "waterfall", top-down style development that was the era's norm. The Manifesto for Agile Software Development is simple:

> We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:
>
> **Individuals and interactions** over processes and tools
>
> **Working software** over comprehensive documentation
>
> **Customer collaboration** over contract negotiation
>
> **Responding to change** over following a plan
>
> That is, while there is value in the items on the right, we value the items on the left more.

The 2001 manifesto and its companion principles spawned many official and formalized "processes." But whether you do Scrum, Kanban, or something else, the basics endure. Put people and iterative, inclusive delivery ahead of elusive concepts like "completeness." Harness the expertise of users and stakeholders in near real-time. Apply that collective knowledge with each iteration to deliver a better product that people actually use.

Rise of Open Source Software Development

Open source also played a major role in changing software delivery. It gave everyone access to incredible technology such as Linux and Postgres. But more subtly, and more significantly, open source changed how teams build software.

People were skeptical: was it as reliable and secure? Turns out, open source was and is more reliable and secure because of the rigor and inclusivity of its community-driven review processes, compared to the traditional top-down, closed models. Additionally, by operating in the open, software developers have become more skilled and literate in the craft. Open source projects spawned peer-review, test-driven development, continuous integration and deployment—techniques that every software engineering team worth their salt use today.

If you're trying to hire high-caliber software engineers today and you don't use agile processes or open-source best practices...good luck with that!

Data and analytics leaders must think about this as they try to build data-driven cultures. While waterfalls are a thing of the past in most software development operations, the top-down approach is still alive and well in data and analytics. The hard truth is that it's impossible to build data-driven cultures under waterfalls.

You won't gain adoption within your organization if you don't bring your community along for the ride. The best data-driven companies like Netflix, Lyft, and Airbnb started with culture and process and then, true to Agile and Open, built or adopted tools that support inclusive contribution and data literacy.

But many vendors are still thoroughly tops-down and closed. They're designed for a decades-old waterfall culture, but sold as modern solutions. These companies will never use words like "waterfall," and they'll spend millions on marketing to convince you they, too, are built for today. But when the sales cycle ends, the agile and open paint jobs start peeling. And you'll soon realize the expensive thing you've bought isn't a solution at all, but yet another ineffective, technology-before-people stab at fixing your business problems. All that, and you'll be no closer to the data-driven culture your company needed yesterday.

Enter Agile Data Governance.

What is Agile Data Governance?

Agile Data Governance is the process of creating and improving data assets by iteratively capturing knowledge as data producers and consumers work together so that everyone can benefit. It adapts the deeply proven best practices of Agile and Open software development to data and analytics.

How to Minimize Knowledge Debt and Make Agile Data Governance an Everyday Practice

In generating data assets, many companies have accrued what we call "knowledge debt." That's when data and analysis isn't documented, has no metadata, and isn't comprehensible. We can all understand why the many people tasked with creating data-driven cultures try to pay down this debt with a silver bullet.

Yet, a healthy data-driven culture minimizes knowledge debt as part of the process of doing the work. Capturing metadata and documentation in the flow of normal work fuels reproducibility and reuse. Adding roles like Data Stewards, Data Product Managers, and Knowledge Scientists makes this process easier because they act as Scrum Masters or Product Owners would if they were developing software, but instead they're building data assets. As with agile software development, Agile Data Governance needs tools that respect—and promote—the agile process and these roles. According to Gartner:

> Effective data management and governance are people-driven practices. They require consistent and high-quality interaction between a variety of roles, and these roles have grown more diverse and distributed over time. Maintaining communication and collaboration is even more critical in the current conditions, creating an opportunity for data and analytics teams to add value by furthering the adoption of new types of tools and approaches.

Agile Data Governance starts by identifying a business problem, then gathering stakeholders who know about the problem and are trying (or have tried) to solve it.

Stakeholders include:

- **Data producers:** data stewards, data engineers, data product managers
- **Data consumers:** business decision-makers, analysts, data scientists
- **Domain experts:** others with deep knowledge of the problem

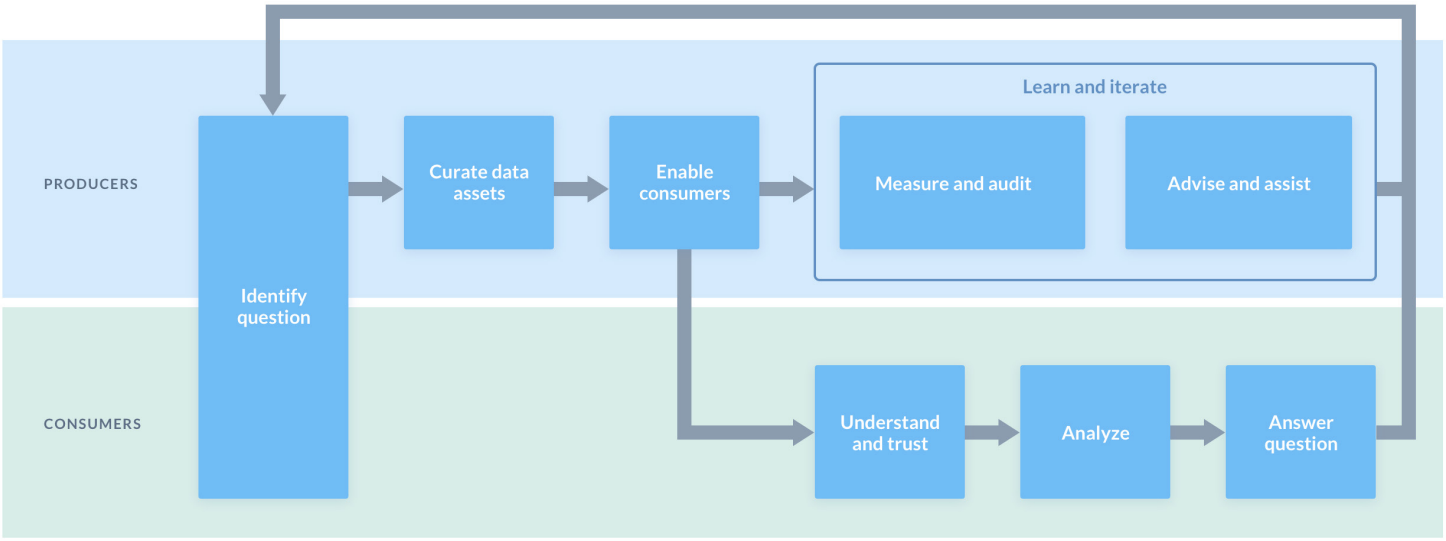
Identify question

Stakeholders should think about *classes of questions* they'd like to answer with data in order to progress toward solving the business problem. Then treat each class of questions as a potential data asset and each question within a class as a user story. From there, choose a question.

Curate data assets and enable consumers

Once the group establishes the question to answer or hypothesis to test, data producers gather data for data consumers to use. That curation has to happen in a durable place so its fruits can be preserved for others in the future. New knowledge and reusable assets will be created and captured with each iteration.

Tip: *Keep questions, clarifications, and modifications close to the data and work. Make sure they are easy to access so the next person with a related problem can find them.*



Learn and iterate

Within these cycles, data producers will quickly learn what's working and what's not about the data sources they're curating, and they can make improvements in real-time. Doing analytics with a living, evolving data asset focuses stakeholders and provides valuable insights at high frequency. That's why Agile Data Governance practitioners see ROI in days instead of months.

By cataloging the work as it happens, and not only the "finished" analysis, teams continuously learn from each other and elevate their data literacy. That's because people learn data skills and domain knowledge faster by doing the work and seeing their peers solve real problems.

Contribute and reuse

Transparency and iteration lead to progressively higher quality as teams refine analysis and data sources one step at a time. The completed reproducible output now gives people a jumping-off point.

When people document their analysis as part of the workflow—not as an afterthought, which is today's unfortunate norm—their coworkers can find, understand, reuse, and adapt it. As with software development (and everything else in life), it's easier to start a data project if you've got something to build on.

As teams build data assets together and watch each other solve real business problems, the community of data producers, data consumers, and domain experts within the organization grows. Useful, creative, once-rare data practices will spread from team to team and become true, widespread best practices. Anyone who wants to make data-driven decisions will finally find what they need without friction or fear.

Above all, there's one reason Agile Data Governance is the fastest, most reliable path to data-driven culture: *your people multiply your data's value, and their own power, just by doing their jobs.*

How Agile Data Governance advances data-driven cultures

If you try to build a data-driven culture with a top-down approach where every detail is planned far in advance, you will fail. But there's wisdom in the saying, "How do you eat an elephant? One bite at a time." Agile Data Governance gives us a way to build an efficient data supply chain and create a data-driven culture one bite at a time.

In his book, Winning with Data, Tomasz Tunguz describes five main challenges companies must overcome to create data-driven cultures. Here's how Agile Data Governance helps solve each of them.

1. Data breadlines

These are bottlenecks at the data producer/consumer threshold. Data producers can't keep up while servicing one ad-hoc request for data after another. Data consumers become frustrated with the delay in getting what they need. Analytics projects turn into endless email chains. By using agile principles, data producers, data consumers, and domain experts iterate together to build reusable assets that lower the frequency of ad-hoc requests. New ad-hoc requests will be preserved next to the cataloged data assets and analysis for the next person to find and use before asking data producers for help.

2. Data silos and rogue databases

Everyone has encountered that "one person" who gatekeeps a special dataset and is the only one who can create a necessary report. Perhaps this person built a one-off system to produce some analytics or scripts that only run on his or her laptop. With Agile Data Governance, data consumers have a direct, clear way to request and iterate on data assets. This reduces the prevalence of "emailed spreadsheets." Plus, data assets will be well-documented, so more people can find, understand, and use them.

3. Data obscurity and lack of understanding

In most organizations, those who try to understand the availability and use cases of data assets encounter inefficiencies, partial answers, and confusing systems. This is primarily a documentation problem, and disconnected tools that aren't built for agile processes make doing documentation both a chore and an afterthought. In Agile Data Governance, you do the documentation while you do the work. This near real-time documentation increases global knowledge about what data exists, what it means, and how to use it.

4. Data brawls

When data work isn't transparent, people don't trust it. People show up with different versions of the same analysis after months of work. They argue about small details, data sources, even project goals. With Agile Data Governance, transparency means course correction and peer review happen as analysis unfolds. This creates a shared understanding which can be poured into business glossaries and other alignment tools. No more tense meetings with three different answers to the same question.

5. Data literacy

You can't have a data-driven culture if your people don't understand the basic workings of statistics. They need to appreciate, and have simple ways to follow, the scientific method and other best practices that make analytics valid and useful. This may be the biggest long term benefit to practicing Agile Data Governance. Humans learn by copying and doing. An agile process encourages participation with, and observation of, talented people doing amazing work. This increases data literacy and skill across your entire company.

Why you need tools designed for the job

Only some tools are right for Agile Data Governance, in the same way the growth of Agile and Open-source software development demanded new tools. Agile software development meant throwing out heavyweight requirements docs and architecture diagrams that would take weeks to write. Sometimes they would consume entire office walls. It required tools that would help teams iterate and collaborate in real-time, like new issue tracking and documentation systems, distributed source control, and continuous test and integration tools. In data and analytics, this means data catalogs designed for inclusivity, crowdsourcing, exploration, access, iterative workflow, and peer review.

Principles of Agile Data Governance

No two Agile Data Governance programs are exactly alike, nor should they be. The histories of Agile and Open Source have taught us that the people who change the status quo can distill their vision down to a set of ideas. On this foundation they build and adapt practices, technologies, and skills to make it real. In that spirit, we offer these 10 principles.

1. Governance should increase transparency, trust, understanding, and speed—not obscurity, doubt, confusion, and delay.
2. Start with the business problems and analytics questions you have today.
3. Iterate quickly to build better habits and get to value faster.
4. One person's work should help everyone else's.
5. Give all stakeholders ways to add knowledge and improve data assets.
6. Keep people, data, docs, and analysis connected and accessible from the beginning.
7. Make documentation easy and iterative or it won't happen.
8. Promote good statistical and scientific methods.
9. Analytics is valuable while it's happening, not just when it's "done."
10. Make the user experience twice as good as the products and practices it competes with to earn adoption.

We want to hear from you. What do you think? Anything missing? How could they be better?

We hope that this exploration of Agile Data Governance gives you the knowledge, focus, and determination to take your first small steps on a faster path to data-driven culture. If you have feedback, questions, ideas, or want to learn how data.world makes Agile Data Governance possible, please email Jon Loyens, Chief Product Officer: jon@data.world.

About Us

data.world makes it easy for everyone—not just the "data people"—to get clear, accurate, fast answers to any business question. Our cloud-native data catalog maps your siloed, distributed data to familiar and consistent business concepts, creating a unified body of knowledge anyone can find, understand, and use. data.world is an Austin-based Certified B Corporation and public benefit corporation and home to the world's largest collaborative open data community.

Visit data.world for more information and expert guidance.

data.world