Bigeye

The data observability team deployment guide

Design a data observability strategy that works for your unique data team structure

Contents

INTRODUCTION

How should your team approach data reliability?

Regardless of team structure, data reliability will become a challenge as your data volumes, use cases, and organization all expand.

If you're considering a platform for measuring and improving data reliability, take a moment to strategize. Your team will face unique challenges and opportunities, depending on how it's structured.

According to that structure, you can choose the best strategy for rolling out and managing observability over your data, pipelines, and assets like analytics dashboards and machine learning models.

In this guide, we'll walk you through the unique challenges, opportunities, and recommendations for data observability, according to the specifics of your data team.

The three types of data team structure

In our travels through data teams large and small, we've discovered that data teams tend to fall into a few basic shapes, which were well described in this excellent piece by Mikkel Densoe:

1. **Fully centralized**
   In a fully centralized model, there is a singular data team that typically rolls up to a VP of Data or CTO. This team includes data engineers, analysts, analytics engineers and data scientists under one umbrella. This singular team supports all data operations for the entire organization, including delivering data to the various functions or lines of business.

2. **Fully embedded**
   Data engineers, data scientists, and analysts are members of different teams throughout the business. Each team will support a different line of business or function (i.e., finance, growth, marketing, operations, etc.). These embedded teams typically have deep familiarity with the business use of the data they support.

3. **Hybrid:**
   There is a centralized team that owns the data infrastructure and provides tooling to help smaller embedded teams within the organization. The central team purely handles infrastructure and tools, or it may also centralize some of the data model, for example maintaining some "gold standard" or "core" tables, but still giving freedom to embedded teams to build their own models.

Most data teams use observability to **monitor millions of data points and fuel larger business strategies** with reliable data.

CENTRALIZED

Data observability for: centralized teams

Overview

It will be easy for centralized teams to ensure the basics across the entire organization (is data loading on time, is there too much or too little data being loaded, etc.). But that's just the tip of the iceberg when it comes to observability. Beyond basic freshness and volume monitoring that should be applied to every table in the warehouse, it's important to also monitor the critical tables that each line of business depends on. That involves deeper monitoring for duplication, completeness, distributions, and business logic that often rely on domain knowledge.

For centralized teams, it's often more challenging to gather this domain knowledge and apply the right in-depth monitoring.

When surveyed, 89.7% of participants indicated that data would be **"somewhat" or "significantly" more important** to their organization's decision-making over the next 12 months.

The challenge

With fully centralized data teams, domain knowledge about the data is often shared with the line of business, and not 100% in the heads of the data team members. Therefore, it can be hard to apply deep monitoring to track business logic-specific information without consulting with line of business owners, slowing down the rollout of data observability.

The advantage

Centralized teams have an easier time getting permissions and access to data and infrastructure, and can more easily assign responders to the data reliability issues that they identify, for example by sharing a single on-call rotation.

CENTRALIZED

Your recommended data observability rollout

If you are a centralized data team, first identify the lines of business (LOBs) that most heavily leverage your data.

Then, partner with them one at a time, taking time to understand where they're leveraging data (e.g. building a list of their key dashboards or ML models), what types of business-specific checks they need to have monitored, and who should be notified if problems are identified by the observability system.

Solve their needs before moving to the next line of business, making it easier to get stakeholders from other LOB's onboard.

This approach may feel repetitive at first, but it allows you to polish your team's data observability process, and build trust within the organization over time.

EMBEDDED

Data observability for: embedded teams

Overview

Embedded teams will have no problem applying in depth monitoring as they should already be deeply familiar with their data. If following a truly embedded model, there might not be a clear single owner of your data observability tools, so it's important that the tool be easy to adopt to prevent disparate silos from forming.

Embedded teams will have a harder time with widespread, shallow checks that reach across all parts of the organization. Is data being managed in the same way on one team as on another? It's difficult to tell.

Without one singular person responsible for that breadth of data quality, the diffusion of responsibility can cause haphazard levels of observability—and therefore reliability—for the organization's data, and slow down the ability for the org to trust data overall.

A recent report from Eckerson defines "quality" as having three main components: **accuracy, cleanliness, and usability.**

The challenge

Embedded teams often struggle to deploy uniform operational metrics across the entire data stack. For example, they might struggle to track freshness and volume for an entire pipeline, from raw data to the final dashboard or ML model. We often find that these types of teams duplicate efforts unnecessarily. It can be challenging to assign a responder to a data reliability issue, depending on where it occurred in the pipeline. Often, we don't see a central owner of data reliability issue, depending on where it occurred in the pipeline. Often, we don't see a central owner of data reliability and observability, so it can be an uphill battle to make data reliable for the organization as a whole.

The advantage

Embedded teams have an easier time enabling great observability for their own line of business. Data experts sit directly on the teams they're impacting, so they have deep expertise in their niche, and can operate data observability tooling without involving another team.

EMBEDDED

Your recommended data observability rollout

Each embedded team should be responsible for the operational quality of the data they support, but should share tools as much as possible with other embedded teams.

The more teams utilize the shared observability tooling, the fewer blind spots will occur from poor inter-team communication. Combining data lineage with SLAs can help teams identify where the hand-offs are when looking at an entire pipeline, and route responsibility for solving outages to the proper team's responder.

Data observability for: hybrid teams

Overview

In a hybrid model, the central team should own the observability tooling, making it available to their embedded counterparts as a self-service offering.

Before doing so, the centralized partners should deploy operational level checks—e.g.freshness and volume monitoring—for all tables in their data stack. That process will enable the embedded teams to implement deeper monitoring based on their domain knowledge.

The central team is responsible for ingestion, the transformation framework, and ETL/ELT orchestrator. A central control for these operational checks will prevent infrastructure issues from impacting the rest of the embedded teams' pipelines.

Once the centralized team has covered operational monitoring, the observability tool can open up to the embedded partners. They will then apply deeper monitoring for the specific parts of the data model that they own.

The challenge

The biggest challenge for hybrid teams is often getting data reliability on the roadmap – for both the central team and the embedded teams at the same time. Once both groups agree to staff the observability effort, they should have the easiest time rolling it out, thanks to the excellent division of labor.

> Data observability is a relatively new approach to tackling an age-old problem, but one that is likely to become increasingly critical to businesses as they become more data-driven and invest in DataOps approaches that automate previously manual data-operations processes.
>
> **Matt Aslett**
> Ventana Research

The advantage

Hybrid teams have a clear and straightforward division between operational and business-logic observability. When each team can focus on what they care about, and nothing more, the scope of data quality work is well-defined. Hybrid teams can be speedier and more efficient in achieving their objectives, as long as they both get data observability on their roadmap around the same time. The central team may need a few weeks to months to get started, with the embedded teams fast-following.

HYBRID

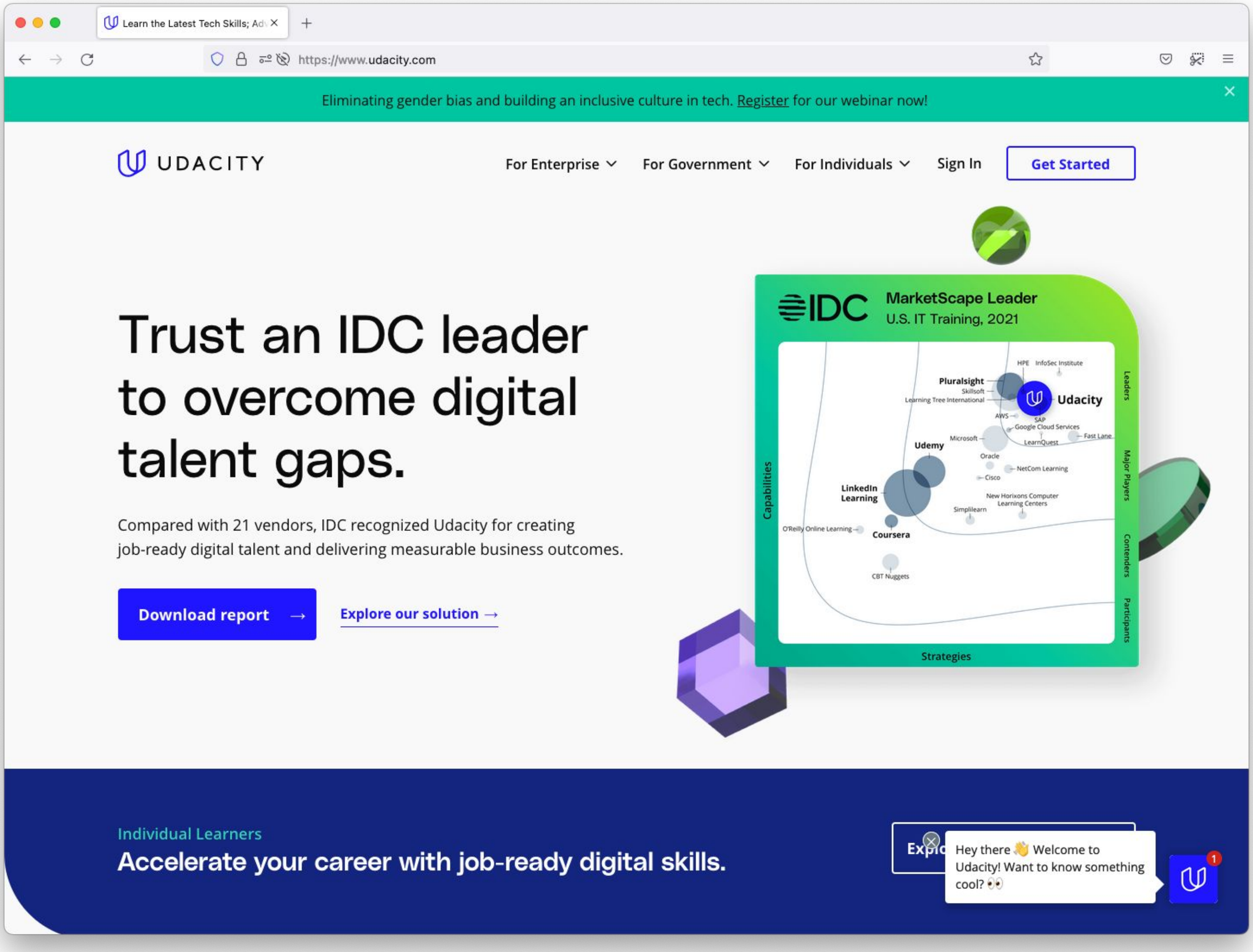
Your recommended data observability rollout

Roll out data quality tools to the central data team first, and then to just one or two embedded teams at a time. This ensures basic monitoring for ingestion, replication, etc. is in place before the embedded teams spend time enabling deeper monitoring.

DATA TEAMS IN ACTION

Udacity’s data team grows stronger with trusted data

When data teams roll out observability that complements their team structure, they get amazing results. Just ask Udacity.



Eliminating gender bias and building an inclusive culture in tech. Register for our webinar now!

Trust an IDC leader to overcome digital talent gaps.

Compared with 21 vendors, IDC recognized Udacity for creating job-ready digital talent and delivering measurable business outcomes.

Individual Learners

Accelerate your career with job-ready digital skills.

Hey there 👋 Welcome to Udacity! Want to know something cool? 👀

Udacity

<24 hrs

Average detection times using Bigeye

1.6 mil+

Number of students at Udacity

With Bigeye, Udacity has one place to understand data quality. Their data team reduced detection times from 3+ days to under 24 hours.

About Udacity: Better learning with data

More than 1.6 million people are advancing their careers with Udacity's programs. The e-learning school offers massive open online courses on everything from data science and AI to product management and cybersecurity. Supporting that growing customer base is a strong data culture that needs high-quality data for decision making and data science.

Challenge: Scattered testing

Maintaining the reliability of important data pipelines — including data from microservice-based systems and event-based data — is critical for supporting Udacity's business analysts and data scientists.

Before implementing Bigeye, Udacity's data engineering team tracked data quality with tools, like Airflow and Apache Spark. But the team wasn't satisfied with the scattered coverage provided by these tools, which also require a non-trivial amount of effort to configure and don't provide a way to detect the "unknown, unknowns" that can cause trouble for data pipelines. Solely relying on these tests made it difficult to catch subtle but important warning signs, like outliers.

Solution: Automating broad coverage

With Bigeye, detection times are down from 3+ days to under 24 hours, and the Udacity data engineering team has a single place to understand the quality of their data pipelines. They're able to monitor hundreds of datasets in a cloud-native data lake with better coverage and faster detection than the testing approach had achieved.

Now that Udacity has Bigeye monitoring data pipelines, data teams are the first to know about any issues in Udacity's data – not their customers. There's no fear of an undetected outlier slipping through and causing great damage.

> I've been in data for many years. As a data engineer, the biggest embarrassment is when the customer discovers that something has gone wrong and says, 'what has happened here?' With Bigeye, I'm confident I have the ultimate answer as to whether the data is good or not.

Simon Dong

Senior manager of engineering,
Udacity

UDACITY

Conclusion

Centralized, hybrid, or embedded, the constellation of your organization's data is unique. Data observability will help any team if it's properly rolled out, so why not roll it out with your team's unique advantages in mind?

As you play to your strengths, you'll find that your data observability capabilities are that much more powerful.

The data observability platform built by data people, for data people.

bigeye.com

Bigeye