MONTE CARLO

2024 Data Engineering Trends & Predictions

How will data engineering evolve in 2024?

Table of Contents

I. A Forward From Lior Gavish

- A renewed focus on data engineering

II. Organizational Trends

- Data contracts
- Cost optimization
- Data reliability engineering
- Data team changes

III. Industry Trends

- Return to the office
- Evolution of the modern data platform
- Data mesh
- Self-service

IV. Technology Trends

- Apache Iceberg
- Productization of internal data
- Data observability moves mainstream

V. Additional Resources

Introduction

2024 will be the year of the data engineer.

The data space moves fast. If you don’t stop and look around once in a while, you just might miss it.

Where 2023 saw teams scrambling to name drop their latest AI project, 2024 will see teams prioritizing real business problems — and the teams that will solve them.

That means a renewed focus on the priorities of data engineers, including data reliability and data quality.

When it comes to the future of data, a rising tide lifts all ships. And the value of data will continue to rise in 2024, raising the standards—and priorities—of the data industry right along with it.

In this eBook, we’ve analyzed the data landscape and put the spotlight on 11 of the biggest trends poised to impact data engineers in 2024.

Ready to see the future? Let’s dive in.

Reliably yours,

Lior Gavish

CTO & Co-founder
Monte Carlo

Organizational Trends

Data Contracts

For those unfamiliar with the hottest emerging data engineering concept of 2023, data contracts are designed not just to fix, but also prevent data quality issues that arise from unexpected schema changes and data swamps.

We wrote an introduction guide with more detail, but essentially data contracts involve working with data consumers to develop the schema and semantic requirements for production grade data pipelines and then delivering that data pre-modeled to the data warehouse.

Now a little wiser and more mature, data contracts have actually seen adoption at enterprise-scale, with companies like Whatnot and GoCardless leading the charge. As data and AI products become increasingly more dependent on high quality data to succeed, considering how to enforce governance standards upstream will be absolutely critical.

Practical insight:
One of the fears of implementing data contracts is that there will be a lot of upfront work to define schemas that will just change anyway. However, what we've found with our current user base is that their needs don't actually change frequently. This is an important finding because currently even one schema change would require a new contract and provision a new set of infrastructure creating the choice of going greenfield or migrating over the old environment.

Read: Data Contracts: 7 Lessons From GoCardless

Cost Optimization

Today's data leaders are faced with an impossible task. Use more data, create more impact— but lower those cloud costs.

As Harvard Business Review puts it, chief data and AI officers are set up to fail. As of Q1 2023, IDC reports that cloud infrastructure spending rose to $21.5 billion. According to McKinsey, many companies are seeing cloud spend grow up to 30% each year.

The good news? There are a variety of ways you can reduce cloud storage and compute costs using the the tools already in your data stack. See below for a few popular approaches:

- Leverage Data SLAs: Data SLAs (service-level agreements) are formalized commitments between data providers and consumers that define the expected levels of availability, performance, and quality of the data being accessed.
- Group Similar Workloads: Within your data warehouse, you can allocate separate virtual warehouses for managing different types of workloads, according to development, testing, production, or specific business units. Using separate warehouses ensures that resources are allocated according to the workload's importance and performance requirements.
- Right-size Utilization: Selecting the appropriate resources and sizes for your workloads is fundamental for controlling cloud costs. Typically, data warehouses and lakes allow you to configure resources like memory, storage, compute resources (CPU/GPU capacity), and concurrency settings (determining how many queries or users can access the warehouse).

And those just scratch the surface. Check out our new guide, 21 Ways to Reduce Your Cloud Data Warehouse Costs, to learn more.

Data Reliability Engineering

Included in our trends guide for the second year running, data reliability engineer is an increasingly important job role with the responsibility of helping an organization deliver high data availability and quality throughout the entire data life cycle, from ingestion to end products: dashboards, machine learning models, and production datasets.

Data reliability engineers often apply best practices from DevOps and site reliability engineering such as continuous monitoring, incident management, and observability to data systems.

All too often, bad data is first discovered downstream in dashboards and reports instead of in the pipeline – or even before.

Since data is rarely ever in its ideal, perfectly reliable state, data teams are hiring data reliability engineers to put the tooling (like data observability platforms and data testing) and processes (like CI/CD) in place to ensure that when issues happen, they're quickly resolved and impact is conveyed to those who need to know.

It's helpful for data reliability engineers to have a strong background in data engineering, data science, or even data analysis. The role requires a strong understanding of complex data systems, computer programming languages, and frameworks such as dbt, Airflow, Java, Python, and SQL.

As the data & AI space evolves, we anticipate this role (and others, like analytics engineer) evolving into "AIOps" or "MLOps" reliability engineer.

Read: What is a Data Reliability Engineer – And Do You Need One?

Data teams will become like software teams

The most sophisticated data teams are viewing their data assets as bonafide data products—complete with product requirements, documentation, sprints, and even SLAs for end-users.

So, as organizations begin mapping more and more value to their defined data products, more and more data teams will start looking—and being managed—like the critical product teams that they are.

When engineers try to build data products or GenAI initiatives without thinking about the data, it doesn't end well. Just ask United Healthcare.

As AI continues to eat the world, engineering and data will become one in the same. No major software development will enter the market without an eye toward AI—and no major AI will enter the market without some level of real enterprise data powering it.

That means that as engineers seek to elevate new AI products, they'll need to develop an eye toward the data—and how to work with it—in order to build models that add new and continued value.

Read: The Moat for Enterprise AI is RAG + Fine Tuning

Industry Trends

Return to Office

Nearly four years out from the beginning of the pandemic, the tide toward working from home is beginning to change. Many companies are requesting that employees return to the office at least a few days a week. According to a September 2023 report by Resume Builder, 90% of companies plan to enforce return-to-office policies by the end of 2024.

In fact, several powerful CEOs - including Amazon's Andy Jassy, OpenAI's Sam Altman, and Google's Sundar Pichai - have already enacted return-to-office policies over the past several months.

There do appear to be some benefits to working in an office (at least part-time) versus exclusively from home. A July 2023 study conducted by researchers at The Federal Reserve Bank of NYC, Harvard University, and the University of Virginia reported that while short-term output decreased for engineers working alongside their peers, those that worked from home were 22 % less likely to get valuable feedback from their coworkers. Further, the study suggests that women who work in an office do more mentoring and receive more mentorship than those who don't.

The good news for data engineering teams? Despite recent economic headwinds and its impact on the job market, your skills are in high demand. While some companies are mandating all employees return to the office regardless of role, other companies like Salesforce are requesting that non-remote engineers go in much less, for a total of 10 days per quarter.

Read: 5 Ways to Ensure High-Functioning Data Engineering Teams

The Evolution of the Modern Data

Since the phrase Modern Data Stack hit the scene in the late 2010s, it's largely been defined as a data platform that is: cloud-based, modular and customizable, best-of-breed first (choosing the best tool for a specific job, versus an all-in-one solution), metadata-driven, and runs on SQL.

Over the past few years, numerous debates have emerged about the future of the modern data stack. How far will the separation between storage and compute go? Will zero ETL actually become a thing? Will the semantic layer realize its full potential?

But today, these considerations have taken a backseat to discussions around how tech stacks will evolve to meet the needs of AI innovation. Of course, there will be the addition of vector databases, bespoke AI tooling to more seamlessly train and operationalize LLMs, and other shiny new toys, but what remains in the air is the role of existing solutions first bred for SQL pipelines.

In fact, various players in the space have already identified that the tools data engineering teams will use to train and fine-tune LLMs with unstructured data are similar to those found in a traditional data pipelines.

But regardless of how the modern data stack plays out, we'll likely experience some growing pains as we adapt to the needs of this rapidly unfolding shift in our day-to-day as data engineers.

Read: What's Next for the Modern Data + AI Stack?

The Data Mesh Matures

The data mesh made quite the splash when it landed in the data engineering lexicon in 2019. And by 2022, it was easily THE hottest trend on every CDOs radar.

Defined by its emphasis on domain-oriented ownership and self-service everything, data mesh is a philosophy of data platform architecture that embraces the ubiquity of data and seeks to elevate access for data consumers.

But like everything that finds itself caught in a hype cycle, the legend of the data mesh found itself bigger than the examples of teams actually doing it successfully. That's because while data mesh is no doubt a valuable theory, it relies on a mix of tooling, processes, and organizational change to be successful. And that latter portion especially doesn't happen overnight.

Fortunately, data mesh isn't all smoke and mirrors. The theory is sound, and the value proposition is spot on for many data teams. Where 2023 took a sober look at the real-world implications of reflection on the massive undertaking

Whereas 2023 stripped away much of the collective naivety surrounding the difficulty of implementing a data mesh, 2024 will see teams finally operationalizing it. And it's already well on its way.

And as data becomes more ubiquitous and the demands of data consumers continue to diversify, we anticipate that data mesh architectures will become increasingly common, especially for cloud-based companies with over 300 employees.

Read: What is a Data Mesh — And How Not to Mesh it Up

Self-Serve Data Platforms

One trend that goes hand-in-hand with the maturity of the data mesh theory is the proliferation of self-service as a priority for data teams. But with self-service, as with many things in data, it's about the definitions you create and where you draw the line.

In fact, it's rarely the case that self-service needs to be defined as every person in the company being able to answer every question they can think up without having to involve anyone in data engineering. If that's the case, where do you draw the line between those that need a higher level of self-service or data access from those who don't?

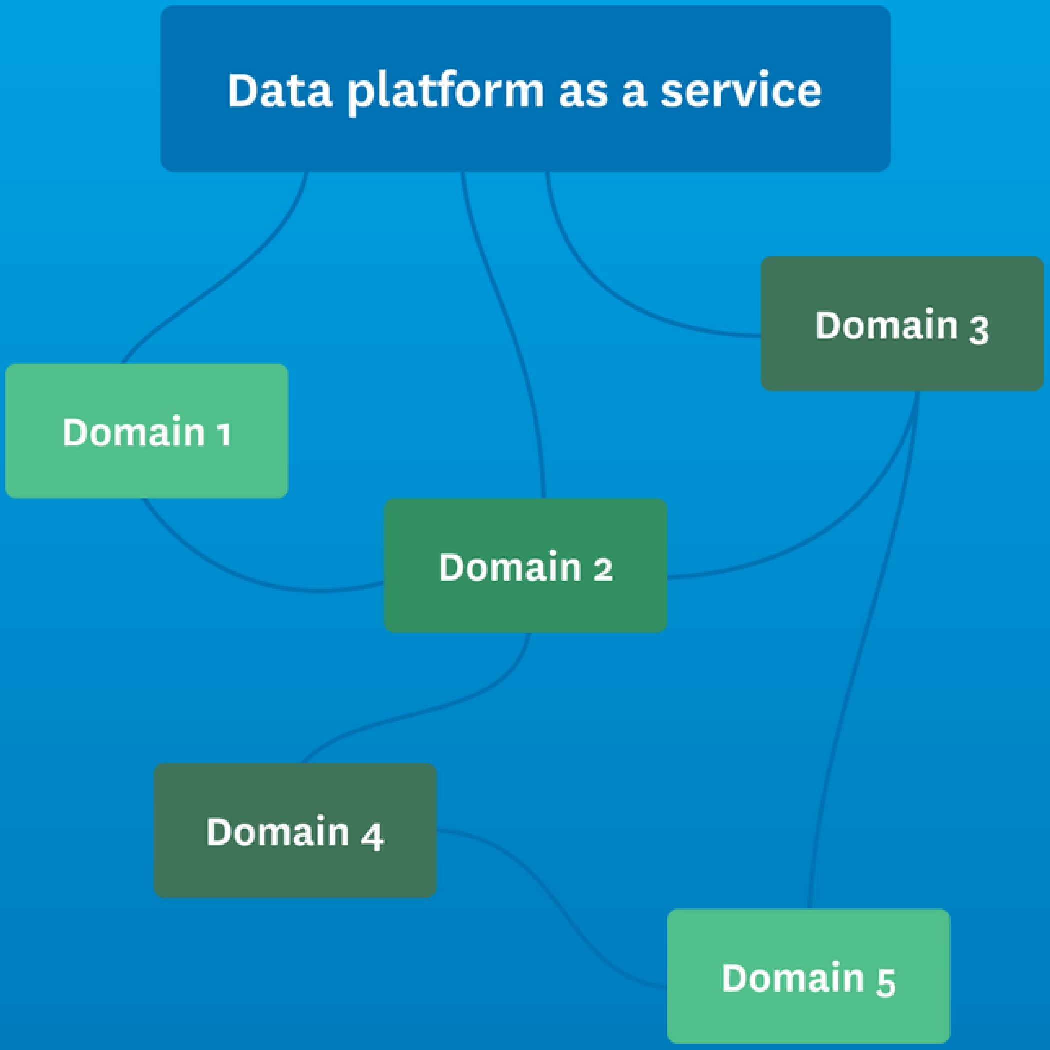
The general consensus is that data analysts are going to form the bedrock of your self-service and data democratization efforts. It's important they are or become fluent in writing SQL. They need to be both equipped and empowered to answer the questions they field as they sit closest to the business.

Unfortunately, there is often a wide gap between the engineering and analyst teams. Ironically, self-service initiatives can actually exacerbate that distance, so careful consideration should be given to ensuring data producers and engineers are still communicating and staying close to their customers.

One popular strategy for bridging this chasm is to leverage analytics engineers to serve as the layer between your messy, behind-the-scenes production database and the data consumer facing data warehouse or instance where the cleanest, most documented data lives for wider exploration.

Read: Enabling a Self Serve Data Culture at Whatnot

Technology Trends



Apache Iceberg Gains Popularity

Apache Iceberg is an open source data lakehouse table format developed by the data engineering team at Netflix to provide a faster and easier way to process large datasets at scale. It's designed to be easily queryable with SQL even for large analytic tables with petabytes of data.

Where modern data warehouses and lakehouses will offer both compute and storage, Iceberg focuses on providing cost effective, structured storage that can be accessed by the many different engines that may be leveraged across your organization at the same time, like Apache Spark, Trino, Apache Flink, Presto, Apache Hive, and Impala.

Recently, Databricks announced that Delta tables metadata will also be compatible with the Iceberg format, and Snowflake has also been moving aggressively to integrate with Iceberg.

Another thing that's made Iceberg so attractive is its time-travel functionality that simplifies versioning by allowing users to easily examine changes and correct problems by resetting tables to an approved state.

As the lakehouse becomes a de facto solution for many organizations, Apache Iceberg—and Iceberg alternatives—are likely to continue to grow in popularity as well.

Read: Is Apache Iceberg Right for Your Lakehouse?

Productization of Internal Data

The data engineering trend that keeps on trending—data products. And make no mistake, data is a product.

But what does data-as-a-product actually mean?

For the past few decades, most companies have kept data in an organizational silo.

Analytics teams served business units, and even as data became more crucial to decision-making and product roadmaps, the teams in charge of data pipelines were treated more like plumbers and less like partners.

In 2024, however, data is no longer a second-class citizen. With better tooling, more diverse data engineering and data governance roles, and a clearer understanding of data's full potential, businesses are embracing data as a product in ways they never have before.

Modern data teams that productize their internal data will need to abide by several key approaches, including gaining early stakeholder alignment, taking on a project management mindset, investing in self-service tooling, prioritizing data quality and reliability, and ensuring your structure supports your data organization.

Read: How to Treat Your Data as a Product

Data Observability Moves Mainstream

At the start of 2023, data observability hadn't been officially recognized as a category for the modern data stack.

Now, fast forward 12 months later to 2024, and key analyst firms and publications, including Gartner, GigaOm, Ventana Research, and G2 have all recognized data observability as a key technology.

Data quality has been a known issue for some time now. But as the data observability category develops and data quality becomes a critical priority for data engineering teams seeking to drive business value, data observability is quickly becoming an indispensable layer of the data stack.

For the third consecutive quarter, Monte Carlo was named the #1 Data Observability Platform by product review site G2. And since G2 is powered by real user feedback and ratings, based on their day-to-day experience, this recognition is especially gratifying.

Over the last few months, we've launched new features like Performance, which helps teams optimize data pipeline performance and cost, and our Data Product Dashboard, which enables organizations to manage the data quality of assets powering critical applications.

Read: Whatnot's Full-Circle Journey to Data Trust

Additional Resources

Check out more helpful resources on data and AI trends and best practices, including:

- Data Downtime Blog: Get fresh tips, how-tos, and expert advice on all things data.
- O'Reilly Data Quality Framework: The first several chapters of this practitioner's guide to building more trustworthy pipelines are free to access.
- Data Observability Product Tour: Check out this video tour showing just how a data observability platform works.
- Data Quality Value Calculator: Enter in a few specifics about your data environment and see how much you can save with data observability.