data
iku

EBOOK

3 Keys to a Modern Data Architecture Strategy Fit for Scaling AI

Introduction

Let's face it: architecture frameworks start to decay as soon as someone puts them on a PowerPoint slide. If there's one thing we've learned at Dataiku after talking to thousands of prospects and customers about their data architecture it's that they also tend to be more aspirational than realistic because, at the enterprise level, data architecture is both complex and constantly changing.

So when it comes to a modern data architecture strategy, the most important factor is not actually the what but the how:

Is your architecture agile enough to be able to easily adapt to changing needs and technology requirements?

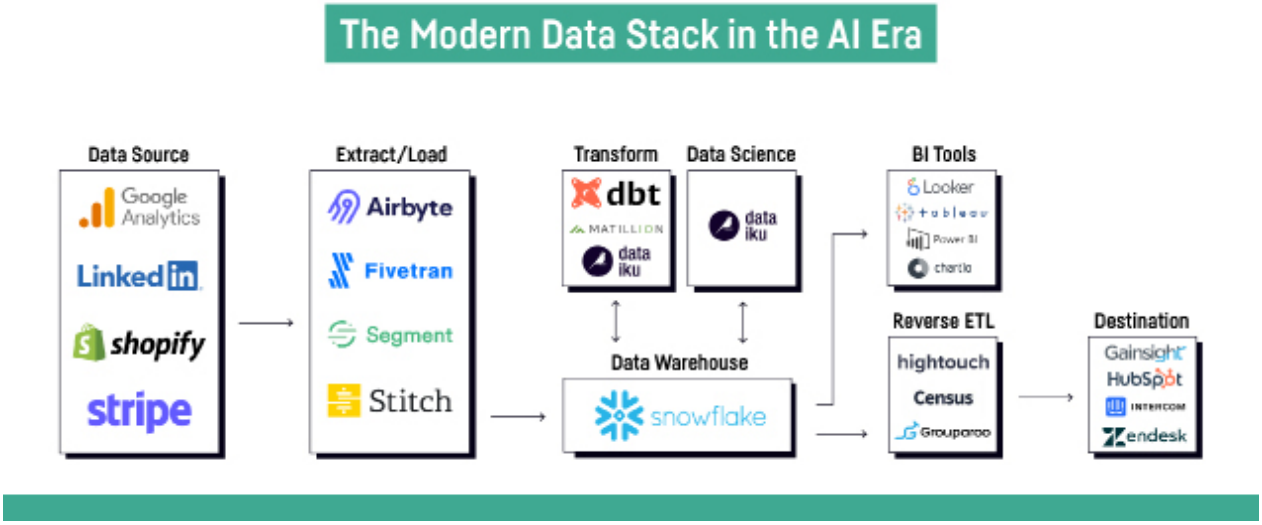
As data ambitions across the organization evolve in the next year, five years, 10 years, will the data architecture be able to support it?

Is the team thinking about the business strategy around data and architecture, not just for downstream consumption, but also upstream (i.e., how data will be made available to other services like applications, APIs, website, etc.)?

Aside from a small detour into what's referred to today as the modern data stack, this ebook won't detail exactly what the ideal architecture is for scaling AI, because the bottom line is… there isn't one (at least, there isn't one that works for every enterprise). Instead, it will focus on three key recommendations that will help teams determine and build the data architecture that's right for them — and, more importantly, for the organization.

The Modern Data Stack

Any small or midsize business that's serious about making the use of data, analytics, and AI everyday behavior for everyone is using a version of the modern data stack architecture. It can even make sense in the enterprise context for teams just getting started on their AI journey.



Some of the key buzzwords associated with the modern data stack are managed, serverless, and low-technical expertise required. Because storage and compute are independent in the modern data stack (and because cloud data warehouses can store massive amounts of data for cheap), data transformation can be done more on-demand, which places less of a burden on IT.

Ultimately, the modern data stack is about providing a seamless experience for all users, no matter what their data needs are. It:

1. Allows coders to do advanced data science on top of cloud data warehouses (including pushing down data processing tasks but also having the ability to operationalize data science projects quickly, to be leveraged by consumers on the business side) and
2. Allows non-coders (like analysts) to do their own data transformation plus advanced data work (e.g., predictive use cases) and
3. Automates and orchestrates the operationalization piece, including pushing the results of multi-tool analysis back to the SaaS tools business users are leveraging.

Even for organizations that have a much more complex existing, legacy setup and therefore can't fully leverage the simplicity of the modern data stack, the goal of providing a seamless experience for all users to work with data is a valuable takeaway.

3 Keys to a Modern Data Architecture Strategy

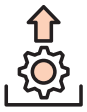
1. Don't Over-Centralize

When it comes to data architecture, for the past five to 10 years, centralization has been the name of the game — potentially to a fault. In fact, the overwhelming majority of IT teams and leaders today have probably tried to centralize too much and too many times.

Here at Dataiku, we talk to a lot of teams (generally within large multinational enterprises across a range of industries) about their data architecture. Usually, when the question "how many single source-of-truth data warehouses do you have?" comes up, the answer is not one. It's two, three, four, and sometimes even more. Sound familiar?

When efforts to centralize have failed over and over (and over and over) again, the answer isn't to double down on centralization, but that's often the reality. Today's efforts to centralize may sound something like:

"We need to get governance sorted out before we can start getting value from data and AI initiatives," or



"We need to solve data quality before we can start investing in tools to start doing any serious data science, machine learning, or AI projects," or

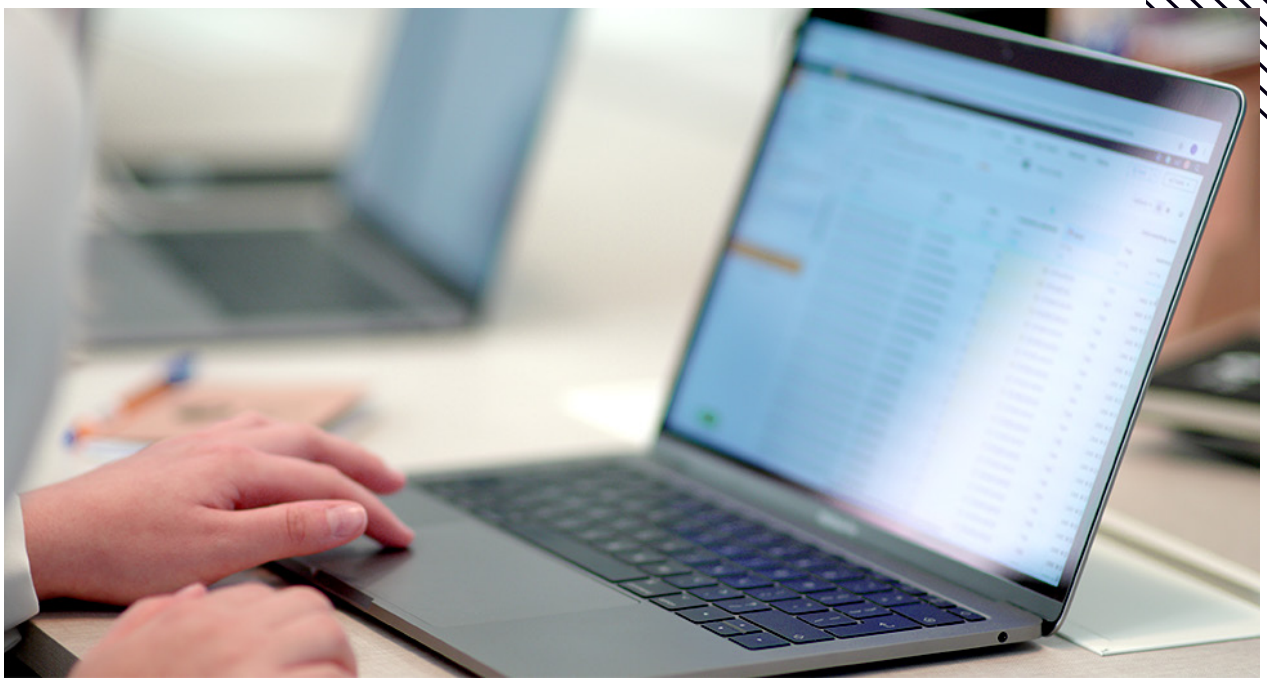
"We just need to finish our cloud migration, then we can start getting return on investment from data."

Why Migrating to the Cloud Isn't an Architecture Strategy

(and Why It Probably Won't Solve Your Problems)

Imagine you have tens of little Allen keys sitting around all over your house. You know, the kind that comes with furniture and other items you need to self-assemble. After assembly, you figure you may eventually need the Allen key someday, so now, you've accumulated a few in the kitchen, in the closet, tucked away in a desk drawer — who knows if you even own the associated items anymore with which they originally came. In this scenario, does putting all the Allen keys in one place resolve the problem?

Well, sort of, in that next time you need one, at least you know where all of them are and potentially save marginal time in digging around in multiple locations. But you still don't know which ones you need and which ones you can get rid of, what works with which item, and how many duplicates of the same size you're holding onto, etc. You haven't been able to assign meaning to each Allen key.



When it comes to cloud migration, many (if not most) people and teams think putting the data all in one bucket is the end of the journey. In our experience at Dataiku working with hundreds of multinational organizations, often with the IT teams, they rarely have thought about completing this sentence:

"We're migrating to the cloud, and because of that, we'll be able to … "

In other words, the value of the initiative is often an afterthought (if it's a thought at all). However, it's worth noting that:

1. **Migrating all data to the cloud is not totally risk-free** — yes, cloud storage can be cheap, but for some organizations (especially ones that are 100+ years old and have incredible amounts of historical data), not cheap enough to put every datapoint that's ever been collected. There's some data that's valuable on the day it's collected, some a week later, some three to four years later. But what about after seven years? It's worth putting some thought around what data really needs to be in the cloud, because after all...

2. ***All* data will never be in the cloud.** Most IT teams don't consider the fact that business people plan for a world where data will pretty much never all be in one place. At Dataiku, we talk to people every week who might have 60%-80% of their data in some big data platform, but inevitably some extremely important thing — like, for example, a list of product codes — comes out of some other business process. It's in peoples' inboxes, it's in XYZ SaaS tool, etc.

The bottom line: cloud migration in and of itself doesn't mean data is getting more meaningful or useful from a business perspective, so for it to be a strategic move with positive outcomes, there should ideally be a larger goal. In other words, cloud migration can (and should) be part of that goal, but it shouldn't be the goal.

Governance, data quality, and the move to the cloud are undoubtedly critical topics. But the point is that while irresistibly tempting to undertake such projects from an IT perspective, more centralization without a larger, use case-based goal or purpose doesn't actually generate any business value and can often mean that efforts in these areas ultimately fall flat.

It's nearly impossible to have a conversation about centralized vs. decentralized data architecture without mentioning today's trendiest term: the data mesh. The concept of the data mesh is less about architecture in the technical sense (while certainly there is something to be said about the tooling possibilities for data mesh architecture, it's outside the scope of this ebook) and more about data architecture from an organizational point of view.

> *What Is a Data Mesh Anyway?*
>
> *"The data mesh platform is an intentionally designed distributed data architecture, under centralized governance and standardization for interoperability, enabled by a shared and harmonized self-serve data infrastructure. I hope it is clear that it is far from a landscape of fragmented silos of inaccessible data."*
>
> — Zhamak Dehghani, Principal Technology Consultant at Thoughtworks and original architect of the term "Data Mesh"

In a nutshell, the data mesh is about decentralization and business ownership of business data assets. That means instead of central teams like IT controlling the source of truth for data across business lines, that responsibility falls on the business itself.

The advantage of this approach is that it puts the onus on the business to maintain, use, and create value from their data. After all, if IT owns the source of truth, but no one agrees with it, what use is that centralization? Having every department, team, or even individual employees creating (and re-creating) their own "single view" of the customer is inefficient, on top of undermining the work IT puts into centralizing in the first place.

The disadvantage, of course, is that the data mesh approach is extraordinarily challenging to achieve in practice. AI platforms (like Dataiku) can lower the barrier to making this switch and put more ownership in the hands of lines of business. But as always, technology isn't a magic bullet — the shift is also largely cultural and will take some serious change management.

There is a natural, underlying tension between IT and business, the desire to centralize and to decentralize efforts. However, technology alone can't (and doesn't) resolve this tension — what does resolve it is aligning business needs as closely as possible with owners of the data. That means getting domain experts to decide what the data means, who should use it, how it should be used, and more.

2. Rethink the Role of IT as One of Creating & Delivering Value

For data practitioners, the best-case scenario is when all data is available and accessible on the same technology and it's all joined together. Unfortunately, there is no product from any of the major cloud vendors that you can turn on and that joins all the data you have to multiply its business value.

What comes out of your operational or business systems is exhaust wherever you put it (cloud/on-prem, data lake/warehouse), and making that exhaust into a solid (i.e., valuable insights) requires human intelligence plus technology to gather all the exhaust from different systems, crystallize it, condense it into a liquid so that it can flow around more easily, then take the relevant liquids and freeze them into a solid — or a square table of numbers, usable by business teams.

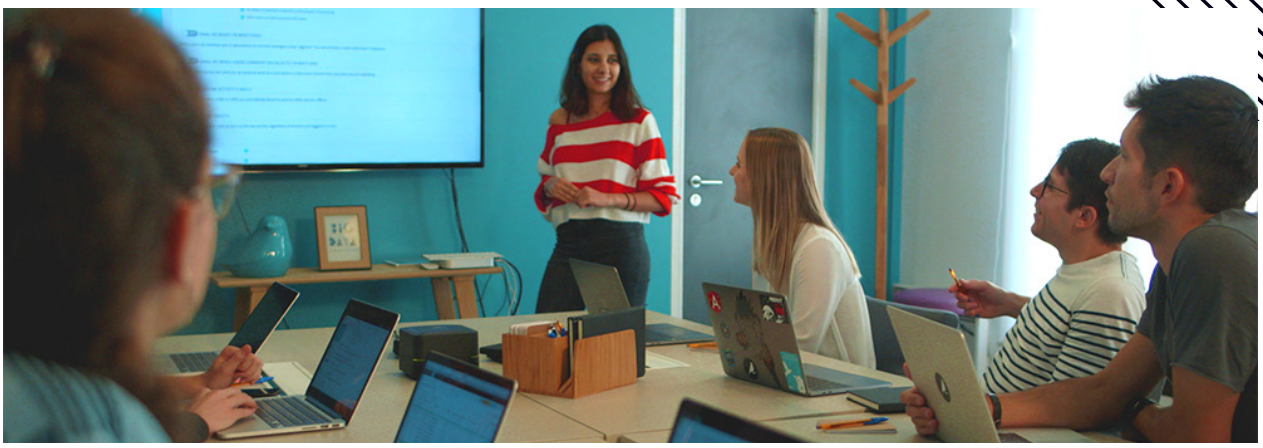
Providing different teams — whether technical or not — with the skills to do this is the root of IT's role in the modern enterprise. According to a *survey of 200 IT executives*, most companies have orders of magnitude more analysts than data scientists, a finding that demonstrates the need for data and the ability to build insights from that data to be accessible to (and over time, adopted by) a wider population within the enterprise.

> *On average, 41% of employees outside of IT customize or build data or technology solutions.*
>
> — Gartner, "Rethinking IT-Business Engagement," 22 February 2021

IT's Reconstructed Identity to Enable Technology Producers, Not Just End Users

Supporting Technology End Users Outside IT	Enabling Technology Producers Outside IT
IT plans, produces and delivers the data and technology solutions employees need to perform their work.	IT provides the digital foundations and platforms that employees need to build architecturally sound, secure solutions themselves.
Standardization — IT owns and provides support for data and technology standards and assets.	**Orchestration —** IT stewards the work of technology producers outside IT for security and integration.
Linear End-User Support — IT creates solutions and pushes them out to enterprise users.	**Producer Network Effects —** IT curates the assets and best practices of business technologists.
Solution Requirements Identification — IT solicits employee feedback on solutions requirements.	**Skill Requirements Identification —** Employees self-assess on the skills the enterprise needs.
Tools Training — Employees learn how to use the solutions provided by IT.	**Digital Skills Training —** Employees develop the skills to produce their own data and solutions.

Source: Gartner, "Rethinking IT-Business Engagement"



The good news is that AI platforms like Dataiku can help create this value, allowing people from around the organization to gather data themselves (even from multiple sources) and merge it without IT intervention. To avoid falling behind or becoming burdened with data processing and integration jobs, AI platforms like Dataiku can further ease the burden on IT teams by:

Enabling a rapid understanding of who is using data, as its size and overall expense continues to ramp up across today's enterprise

Handling the proliferation of data and analytics, the demand for more data, and ease at which data-driven projects can be reused

Offloading data prep and ETL to the teams regularly using data, such as data scientists

Supporting elasticity and resource optimization, allowing organizations to process massive amounts of data, large numbers of concurrent usage, and services deployed

The bottom line is that even if your organization has successfully put "all data" (remember, it's hardly ever truly *all* data) in one place, the human work needed to transformation that data, join it together, and get it into a state where it has the potential to provide business value is not entirely automatable — at least not for now. It's critical to develop a data architecture strategy that considers this factor and facilitates the value part of the pipeline as well.

3. Let Business Objectives Inform Your Architecture (Not the Other Way Around)

Making technical choices before considering business goals (or without considering them at all) can manifest itself as:

- Architecture diagrams and plans that are more well-defined and iterated on than the business case itself.
- Conversations about how to leverage the latest architecture trend that come before finding a use case on the business side.
- Initiatives around driving value from data that start with IT or architecture discussions.

Any way you slice it, designing architecture first and considering the business needs after is problematic — and not just from the business perspective. It's often the root cause of shadow IT because people will try to do things with data that the architecture just doesn't support, but if that initiative has enough value, they'll find a way to do it anyway.

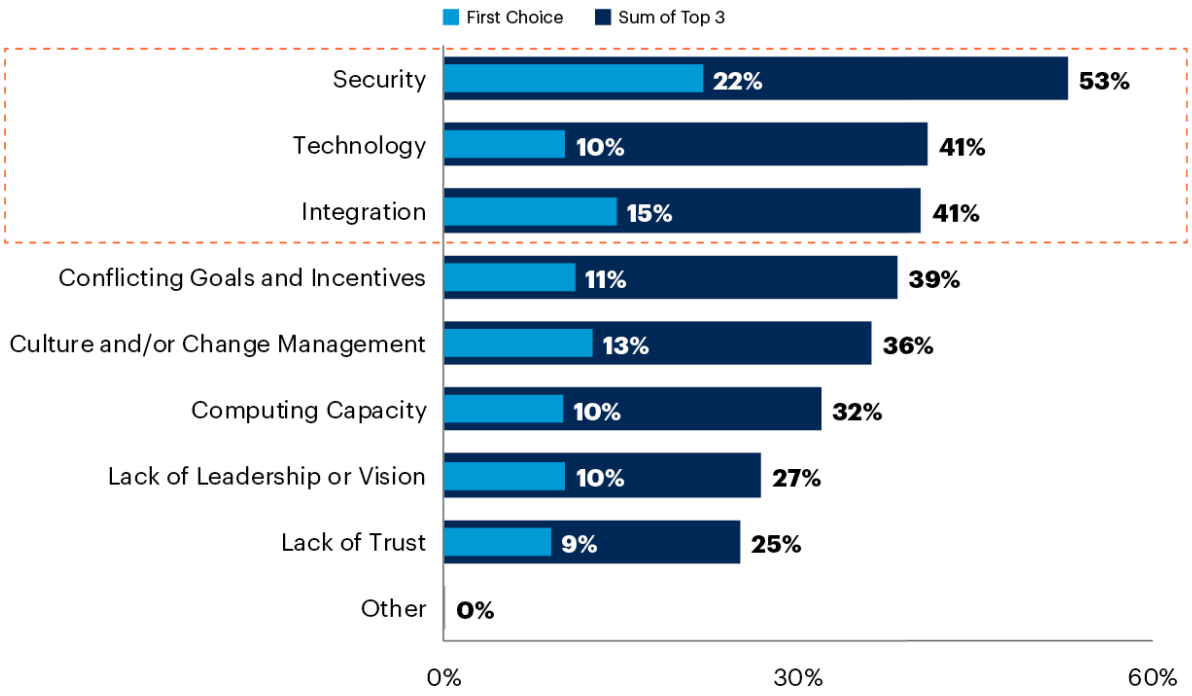
But what does it mean in practice to let business objectives inform architecture? It means answering questions like:

- Given your business and the data you have, what kinds of things do you want to do with it? Sell it? Use it? If the latter, for what, specifically?
- What's the business objective for collecting and storing all this data? How will it ultimately be used?

Bonus: if architecture choices are aligned with the business, the technical choices themselves become much easier. There's a focal point and an end goal — it's not just about subjective discussions around what architecture would be the coolest or most innovative, but what will accomplish that goal.

> *By 2022, organizations with diverse IT-business collaborations will deliver business outcomes 25% faster than their competitors.*
>
> — Gartner, "The Future of Applications Depends on IT-Business Collaboration," 20 October 2020

Top Three Obstacles to IT-Business Collaboration



n = 397; Base: Answering "Comprehensive/Moderate/Limited" for IT, LOB collaboration, excludes don't know

Q: What are the top 3 obstacles to the lines of business (LOB) and central IT (CIT) collaboration on application delivery in your organization?

Source: Gartner 2020 IoT Implementation Trends Survey

732996_C

Gartner

Of course, aligning closely with business objectives isn't without challenges (see figure from Gartner above). But the potential payoff, according to Gartner, is worth it, with 25% faster business outcomes when IT and business collaborate closely.

Conclusion: Dataiku's Role in Modern Data Architecture Strategy for Scaling AI

In this ebook, we explored how scaling AI from an architectural perspective requires rethinking efforts to continually centralize, rethinking the role of IT itself more broadly, and letting business objectives inform architecture. We touched, at times, on how AI platforms like Dataiku can help teams achieve these objectives, but more specifically, what other advantages does Dataiku bring and how can it help drive results with AI at scale?

First and most importantly, Dataiku was built from the ground up to be one central, controlled environment used by a range of profiles — including low-code analysts and no-code contributors on the business side — which, by nature, addresses and facilitates many of the three keys to modern architecture discussed in this ebook. But it's not just a low- and no-code solution and offers robust features to give IT teams maximum flexibility yet control over architecture:

Dataiku can run on-premise or in the cloud — with supported instances on Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure — integrating with storage and various computational layers for each cloud.

Dataiku uses a pushdown architecture to allow organizations to take advantage of existing, elastic, and highly scalable computing systems, including SQL databases, Spark, Kubernetes, and more.

Dataiku provides a fully managed Kubernetes solution that is compatible with all of the major cloud container services — Amazon EKS, Google Kubernetes Engine (GKE), and Azure Kubernetes Service (AKS) — as well as with on-premises Kubernetes/Docker clusters.

Dataiku supports the use of both CPUs and GPUs for model training. If multiple GPUs are available, Dataiku can distribute model training workloads across the GPUs to dramatically decrease training time.

In the Dataiku project flow, all visual components are reusable and portable. Individual preparation steps or entire sections of a flow (datasets and recipes together) can also be shared externally to other projects, allowing users to rename and re-tag objects in the process.

Organizations can extend the power of Dataiku with custom plugins. The Dataiku plugin library includes over 100 plugins that enhance existing Dataiku instances, including access to new data sources, charts, programming languages, algorithms and modeling techniques, partner integrations, and more.

Excellent tool to accelerate, from day-to-day data analysis to fast AI deployments!

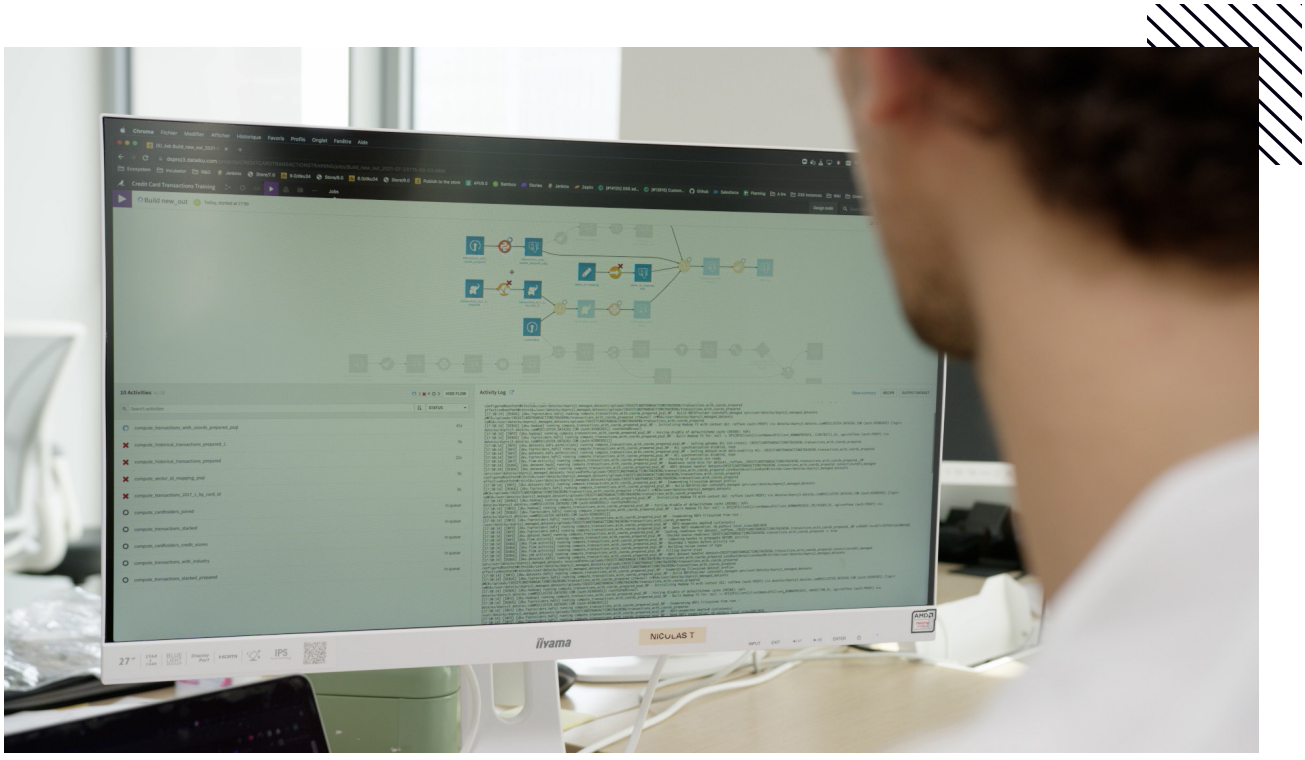
"Excellent support from Dataiku to accelerate the adoption of the tool, very good ratio 'time spent/added-value; for the company. Allowed us to bring a wide range of profiles to work with data and to be ready to deploy basic services in production (80% of our requirements) very quickly."

— Enterprise Architect in the Communications Industry[1]

Gartner peerinsights customers' choice 2021™

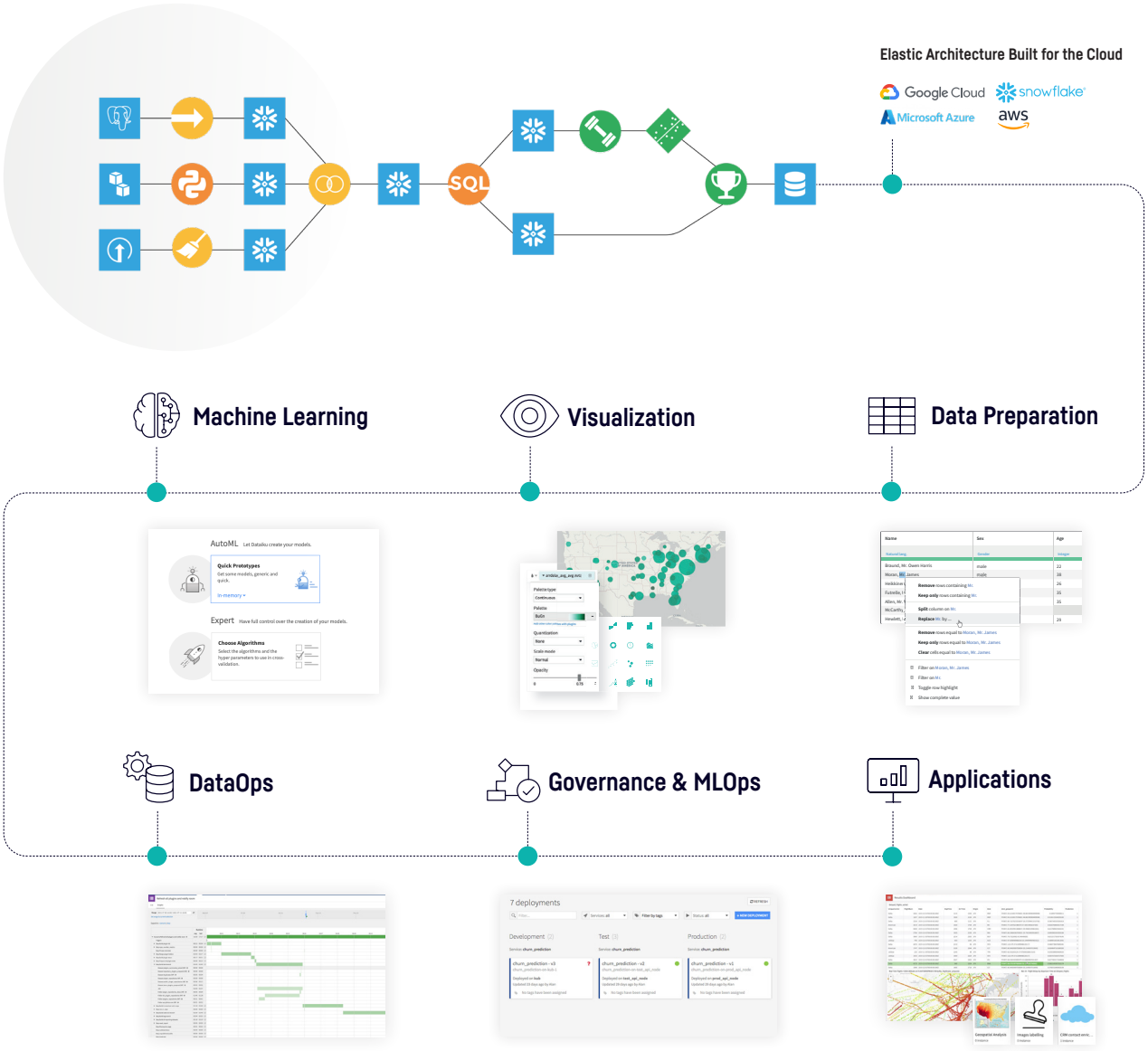
For those on the technical side — like data scientists, but also data engineers, architectus, and more — Dataiku facilitates quick experimentation and operationalization for machine learning at scale.

Discover how Dataiku can fit into the modern data architecture strategy, scaling AI across the entire organization (securely).



[1]https://www.gartner.com/reviews/market/data-science-machine-learning-platforms/vendor/dataiku/product/dataiku-dss/review/view/3636768

Everyday AI, Extraordinary People



Dataiku is the platform for Everyday AI, enabling data experts and domain experts to work together to build AI into their daily operations. Together, they design, develop and deploy new AI capabilities, at all scales and in all industries.