ALTAIR

A HUMAN'S GUIDE TO DATA ARCHITECTURE FOR DATA SCIENCE

INTRODUCTION

Data is at the core of business today. How do you manage data so that it gets to where it needs to go and can generate value for your organization? What are the best tactics—both technically and culturally—to harness data? There are also a lot of nitty-gritty questions surrounding implementation: Should you move to the cloud? Store and process everything locally? Use edge computing devices to process data where it's generated? These questions are central to any data management strategy, but have become increasingly important as more organizations look to create long-term value through data science and machine learning.

Ultimately, all these questions boil down to something a lot simpler: How do you build a data architecture that facilitates data science that has an impact on your business's bottom line?

> “For companies to build a competitive edge—or even to maintain parity, they will need a new approach to defining, implementing, and integrating their data stacks.”
>
> MCKINSEY & COMPANY

Table of Contents

In this eGuide, we'll explore some of the key technical aspects of data architecture that you should be thinking about as you're making decisions about how to store and access your data. This will help you understand what your options are so you can figure out what's most appropriate for specific use cases and your broader efforts across the organization.

We'll also address the cultural challenges that organizations often face as they start thinking through comprehensive data architectures so you can ensure that everyone, from your data wranglers to your subject-matter experts, are on board and involved in making data-first decisions that benefit your organization.

THE TECHNICAL SIDE

As mentioned, building a data science-ready architecture requires you to make both technical and human choices. In this section, we'll walk you through the key technical aspects that you'll need to consider to prepare your organization to tackle impactful data science projects.

Scalability

One of the most pressing challenges you'll face when building your data science-ready architecture is creating an architecture that's both appropriate right now and scalable in the future as you involve more users and rely on a growing volume of data.

"Scaling Up" Vs. "Scaling Out"

When thinking about scale, it's important to distinguish between "horizontal" and "vertical" scaling, as each has its own implications.

Vertical scaling, or "scaling up," describes the process of storing data on a single node and adding computing power as needed. Scaling up involves spreading the computational load through the CPU and RAM resources of a single machine and allows you to increase its capacity—to a point.

When scaling vertically, it's important to remember that you'll still reach an upper limit when you hit a server's capacity. Once that happens, you'll need to provision additional infrastructure to support your strategy, which often means purchasing newer, more powerful machines that can handle greater workloads.

Horizontal scaling, or "scaling out," refers to the practice of partitioning data so it can live across multiple nodes, with each node containing only a piece of said data. This approach allows multiple servers to each take on a fraction of the overall computational load in parallel with one another.

"Scaling out" doesn't require that you add more power to individual machines, but it does require you to combine the computing power of multiple devices to increase overall capacity.

Which Method Is Right for You?

	SCALING UP	SCALING OUT
BENEFITS	• Simpler and easier to manage • Works with existing code • Lower up-front cost	• Less prone to failure/downtime • No upper limit to scale • Easier to maintain and troubleshoot
DRAWBACKS	• Doesn't handle large data volumes or applications well • Less fault tolerant • Higher overall equipment investment	• Greater up-front investment • Requires adaptation of your code • More complex architectural design
WHEN TO USE	Vertical scaling is best used for smaller businesses whose capacity needs won't drastically change over time and are looking for a lower-cost, simpler solution.	Horizontal scaling is best for organizations who expect data volumes to significantly increase as they grow, and have in-house resources who can adapt code and manage their environments.

Portability

Portability refers to the ability to use the same piece of software in multiple operating environments. It's an important aspect of any data architecture, but it's especially crucial for organizations that want to set themselves up for a true digital transformation.

When you think about how much work goes into getting models ready for production, the last thing your analytics teams will want to do is throw out a perfectly good solution due to a lack of flexibility.

Start "Future-Proofing"

As your company grows, your data architecture will need to evolve too. More systems will be added, analytics goals will change, and more employees will need access to business-critical data to make informed decisions and drive desirable outcomes.

This creates an interesting challenge, which is that you'll need to create models that can be deployed in different environments depending on your strategy. If your teams have been working hard to create models that run well in certain environments, they'll want to know that those models can still provide useful insight when deployed elsewhere.

Support Iteration

You may not know this if your organization hasn't taken on many data science projects yet, so it's worth noting: it's highly unlikely that your teams will be able to envision, create, test, and deploy impactful models on the first try. Data science projects are iterative by nature, as there are so many different sources of data to consider, algorithms that can be used, and ways to embed models within the business so that they can create the most value possible. For example, even in scenarios where you create and train a model using a certain dataset, you may realize that its predictions would be more valuable with additional data from different parts of your company—having the ability to go back, gather that data, and improve the model is critical to optimize long-term results.

Iteration is especially important when it comes time to deploy a model, because the problem you're trying to solve with data science will dictate how a model should be operationalized.

For example, if your marketing team is planning a cross-sell campaign that primarily relies on past purchase history, it's okay for your model to process data in large batches periodically, as opposed to doing so in real-time. On the flip side, if you suspect credit card fraud is eating into your bottom line and want to identify illegitimate transactions as they're happening, your model will need access to a constant stream of data so it can make timely predictions.

Prevent Vendor Lock-In

Just about any software application your organization uses has access to important company data. Your CRM provides a comprehensive view of customers, ERP software helps to fulfill orders, and your HR platform contains everything you need to know about your employees.

Machine learning models can leverage all that information to provide valuable predictions that help your business run more smoothly—but what happens when the vendors storing that information change their offering, raise prices, or otherwise drop the ball?

From a data science standpoint, you'll want to know that the models you've created won't become useless if you decide to switch out a key vendor. While it's reasonable to expect a slight disruption during any software migration, you shouldn't have to create, train, and deploy a brand-new model from scratch if you already have a working solution.

By prioritizing portability and flexibility, you can provide your company with some much-needed leverage during contract negotiations and renewals while knowing that your teams' work will still drive long-term value.

Operationalization

As you may have noticed, the technical aspects of your data architecture have a common thread—they're designed to ensure you can operationalize the models you're building. Operationalization, also commonly referred to as ModelOps, refers to the process of embedding data science models within your business so they can develop predictions based on actual data, enabling your teams to make more informed decisions.

Effective operationalization is often what separates the data science projects that drive real business impact from the ones that never create value, making it a crucial technical element to consider in your architecture. Unfortunately, many enterprises (and data science vendors) prioritize model creation over ModelOps, which typically creates headaches during the data science-to-IT handoff.

But it doesn't have to be that way. In fact, your data architecture should facilitate model operationalization. Here are a few key factors to consider when you're building it.

Systems, Systems, Systems

It's no secret that as organizations have undertaken large-scale digital transformation efforts, the number of software applications they rely on has skyrocketed. To create efficiencies, boost revenue, and/or lower costs, companies often turn to software vendors to support key business functions like engineering, customer support, and sales. According to Okta, a leading enterprise identity-management firm, businesses with over 2,000 employees deploy an average of 175 applications across business units.

Because you're reading this, we'd guess you want to leverage the data that's stored in those apps for detailed analysis, which presents a unique challenge: creating models that can be deployed anywhere that valuable insight resides.

Let's say you want to create a model that can identify which of your customers are most likely to churn. It's one thing to train and test performance on a sample dataset, but eventually, that model will need to provide useful predictions based on live, fast-changing data stored in multiple business applications.

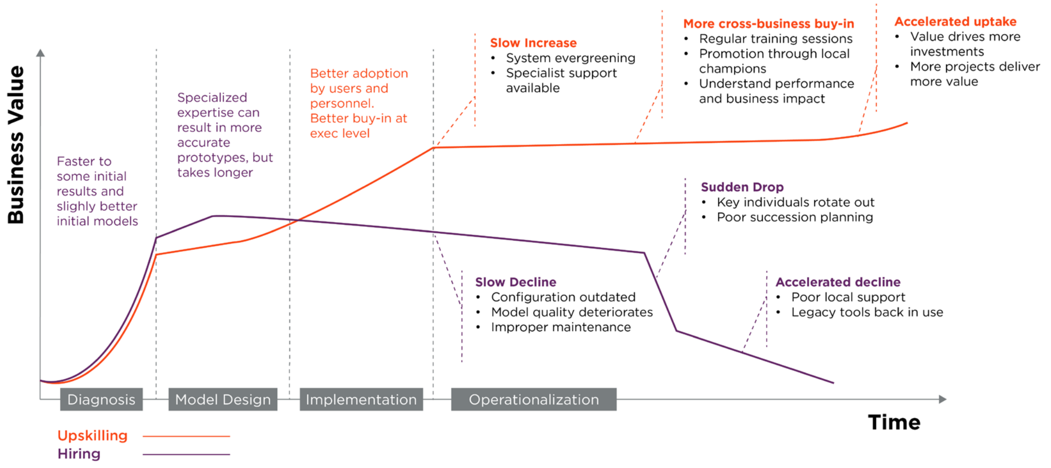
In the case of churn, that means the model would need to access CRM data to determine how much a given customer is spending, whether they're receiving a discount, and any other variables that could impact their willingness to continue doing business with you. But in many cases, CRM data alone won't provide the full picture. For example, a model may also need to account for relevant information that's logged in a support or ticketing system to see if a customer has outstanding issues, and, for software vendors, effective churn modeling should also account for customers' product usage.

All this is to say that even for a relatively common data science use case, a model's ability to integrate with existing enterprise systems is the difference between impacting business outcomes and not even being able to bring a model into production.

Long-Term Value Creation

Another major consideration concerning operationalization is the ability to maintain models over time. This is a sticking point for many organizations who gravitate toward taking on new use cases once they've done the work of building, training, testing, and deploying a model. According to Altair's Frictionless AI Global Survey Report, 35% of respondents cited the iterative nature of data science projects – i.e. not being able to set it and forget it and not having dedicated and knowledgeable staff to keep them running smoothly – as a common challenge in leveraging the value of their financial investments in AI. It's especially true of data scientists, who are understandably more inclined to apply the latest developments in data science to new problems than maintaining existing solutions. A failure to address this problem will likely lead to negative consequences like model drift, when a model trained on historical data can't provide useful predictions in non-static, fast-changing environments.

The solution to this problem isn't to simply bring on more data scientists—both for the reason mentioned above and because they're extremely in-demand and difficult to hire. Instead, consider developing an architecture that allows non-data scientists (business experts, analysts, etc.) to access, manage, and provide automated analysis to ensure that models create long-term value. Focus on creating an environment that's well governed and prevents valuable data from getting stuck in silos where it can't be monetized.



Looking Ahead

While the technical aspects of your data architecture provide the structure to extract value from your data, getting them right is only part of the battle. In truth, many enterprises struggle to implement impactful machine learning models because they neglect a crucial resource—their people.

CULTURAL ASPECTS

A lack of data science talent is prevalent in organizations; **75%** of the aforementioned Altair Frictionless AI Survey respondents said they struggle to find sufficient data science talent. Further, the combination of this lack of talent plus the challenge of upskilling the workforce prevents AI adoption. While we believe it's critical to turn towards your domain experts and analysts to optimize your work, this requires you to think hard about the cultural aspects of your data science initiatives.

No matter how much technical expertise you bring to your data architecture and digital transformation projects, it won't be enough to guarantee your success. In addition to making changes to your technical workflows, you also need to change how your employees think about and use data. If you don't consider the cultural changes that need to happen within your organization to make sure everyone's on board and can use the architecture you're building, you're setting yourself up for failure.

To get everyone on the same page, you need to create a data-first culture, break down silos, and iterate on the process of digital transformation. Let's examine these aspects in more detail.

Create a Data-First Culture

One of the key aspects of making data work for your organization is to ensure that data scientists and IT aren't the sole arbiters of your data. Sure, IT is going to have a say in where data is stored and how it's secured, and data scientists might have specific access needs, but it's critical for your operations that every data-loving person in your organization be upskilled so that they're not only able to access the data they need, but also that they understand what the data is and how it they can use it to make decisions and support their work.

As you push towards digital transformation and enterprise-level AI and machine learning strategies, more and more of your decisions are going to be based on the data you have and the models you build from that data. If you don't have employees who understand what decisions are being made and why, this creates friction and slows down change. Ideally, implementing data-driven processes should be something that anyone in the organization—from the C-suite to domain experts to data scientists—can understand.

Break Down Siloes to Promote Collaboration

In addition to not having enough data-fluent employees, modern organizations – by virtue of their size and structure – also struggle with siloed teams who work in isolation, who are often unaware of what other teams are doing. In some cases, this is a result of not promoting data fluency—teams with less experience and exposure to data can be left out of the work that data-savvy teams are doing, potentially blocking the ROI and the creation of value that comes from close collaboration between, for example, data scientists and domain experts.

But data fluency alone isn't sufficient to break down barriers. You can still end up with siloed teams working on their own data—or even shared data—without communicating with others about what they're doing and how different projects might interact.

Thus, once you've got data fluency in place, you can start encouraging your employees to think of data management and interpretation as being part of their jobs, not something they need to pull in a technical resource for. They might still seek consultations with machine learning experts, but they should be able to understand how they can harness the data that's available to them to work collaboratively.

To break down silos, look for systems and tools that let people with different skillsets collaborate, using a single ground truth of data in lots of different ways. The tooling that you're using should be accessible to people whether they can write code or not, and regardless of their level of domain expertise. Altair® RapidMiner®, Altair's data analytics and AI platform, for example, offers self-service data preparation and data visualization tools, as well as code-free to code-friendly machine learning solutions.

Successful data science projects rely on the involvement of data experts, business experts, and executives, so your data architecture should cater to all these personas to make the sharing of information and insight simple and efficient.

DATA FLUENCY ALONE ISN'T SUFFICIENT TO BREAK DOWN BARRIERS.

Iterate, Iterate, Iterate

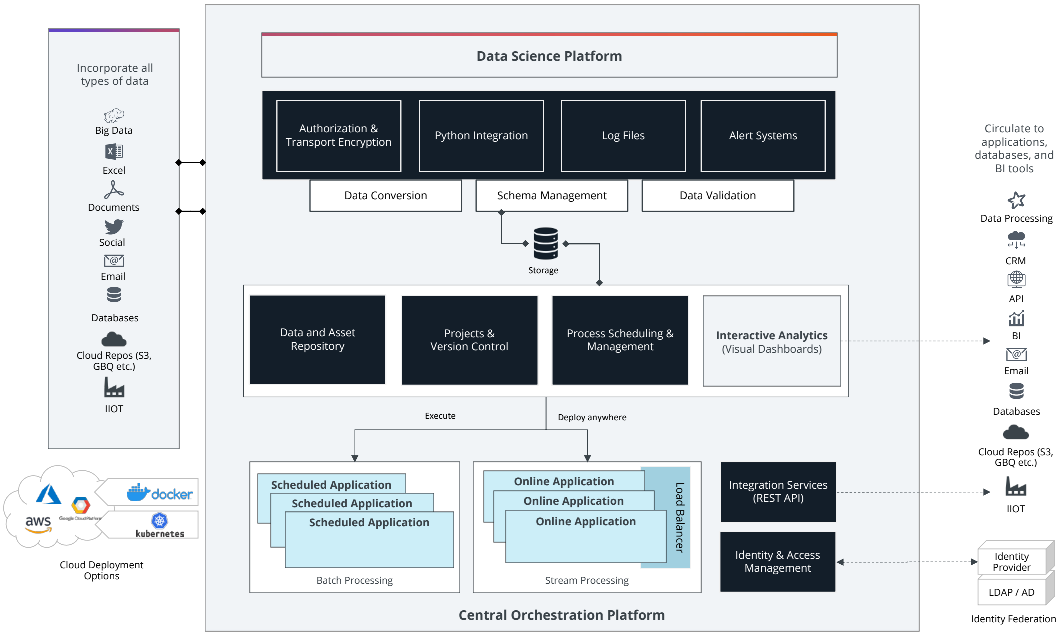
Just like with models, you're unlikely to hit the ball out of the park on your first attempt to create a data-centric, digitally transformed culture. Just as with model training and deployment, you'll need to constantly assess how your organization is handling these challenges and be ready to change your strategy if necessary.

What does that look like? It might mean regularly evaluating new tools that can do some or all the cultural (and technical) work described above. It might mean providing training opportunities to employees so they can learn more about how to access and use data to become citizen data scientists.

Ultimately, as you begin to integrate the data more and more into your everyday processes and decision-making, you're going to find areas in which you can improve. That doesn't mean you've made a mistake or done something wrong! It just means that you've identified a new area to continue pushing for data fluency and digital transformation.

Example Data Architecture Diagram

The sample data architecture diagram below shows how data, tooling, and employees can work together to drive ROI and promote digital transformation.



KEY TECHNOLOGIES TO NOTE

Docker

In a data science context, Docker can be an extremely helpful tool when it comes time to operationalize your models. It's one thing to create a model that runs well on a given machine, but another thing entirely to successfully run it in multiple complex environments.

Docker is a technology that can simplify that task by providing a single environment that contains everything needed to run a software application—not just the code itself, but also any frameworks, libraries, or dependencies that must be accounted for. Appropriately, these environments are called containers, and are a crucial part of model operationalization.

By simplifying the process of packaging and reproducing code, Docker removes one of the most common barriers to successful data science projects, which is the disjointed handoff between those who create models (usually data scientists and/or business experts) and those who operationalize them (data/machine learning engineers and IT professionals).

Kubernetes

Whereas Docker provides a strong solution for environment management and the reuse of code through containerization, Kubernetes offers users a way to manage those containers at scale in production environments. For example, if one container goes down, another container would need to start to avoid unexpected downtime—the problem is that it's not feasible to manage this process manually, especially at scale.

Kubernetes offers an all-encompassing framework that allows users to balance traffic to containers to maintain stability, manage each container's storage requirements, and automatically restart, replace, or kill unresponsive containers.

Using Kubernetes doesn't just bring productivity benefits, but can also help to attract top talent, future-proof your applications, and ensure you're only paying for the computing resources you need.

Cloud Platforms (AWS, Azure, Google Cloud)

As businesses have sought to reduce infrastructure costs, become more agile, and improve scalability, cloud computing has exploded in popularity. In turn, this had led to the emergence of the Infrastructure-as-a-Service (IaaS) model, in which an outside vendor hosts and provides key infrastructure like servers and storage while charging organizations for what they need.

This creates numerous advantages. For one, relying on a vendor that's dedicated to managing and provisioning cloud services usually means that your infrastructure will be more robust and reliable than you could design in-house. It also means that you won't be responsible for purchasing and maintaining hardware, which can not only present significant CapEx, but also distract internal IT resources from furthering business goals since they'll need to spend significant time ensuring your infrastructure is up to date.

The three highest-profile and most used Cloud Service Providers (CSPs) are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud. All three have worked to address concerns about enterprise data security, and have seen widespread adoption as a result.

HOW ALTAIR® RAPIDMINER® FITS INTO YOUR DATA SCIENCE ARCHITECTURE

If your team is focused on creating and operationalizing data science solutions throughout your business, it's worth mentioning that Altair RapidMiner is flexible enough to support different strategies while aligning with the concepts described earlier in this eGuide. If you're some distance from truly building out your data science initiatives, we encourage you to check out the 2023 Altair Frictionless AI Global Survey Report, an in-depth analysis of where current data and AI strategies are falling short and how organizations can eliminate friction for good.

If you're already bought in on the many benefits data science can bring, here's a brief overview of how Altair RapidMiner can help you create a data science-ready architecture.

Scalability

Altair RapidMiner supports both horizontal and vertical scaling, depending on an organization's deployment strategy. For single-node deployments, we typically recommend a vertical scaling approach—by contrast, we'd typically suggest that customers scale horizontally in multi-node deployments. Either way, Altair RapidMiner users can count on having high-availability configurations that ensure reliability.

It's also worth mentioning that Altair RapidMiner's machine learning solutions reflect a typical data science workflow, which means they offer support for processes that can be scheduled as well as those that require a real-time response (e.g., banks monitoring a constant stream of transactions to detect credit card fraud). To help with scalability in either scenario, we offer both Job Agents and Real-Time Scoring Agents. As you can probably guess, the former helps increase processing power for batch applications, while the latter increases processing power for streaming and edge use-cases.

Portability

Altair RapidMiner solutions are designed to provide enough flexibility to deploy solutions on-premises, using your preferred cloud (public or private), or at the edge in situations where data needs to be processed in real-time on specific devices. The platform can also be installed and run on Docker or Kubernetes, and even comes with pre-built templates to assist your users with that process. By providing organizations with options, our goal is to help them future proof—as they grow and their data strategies change, they can continue to see value from the models that they've spent time and effort creating.

Altair RapidMiner also aims to prevent vendor or strategic lock-in by supporting both batch and stream processing, allowing you to visualize results in your platform of choice, and by helping users access a wide range of data sources (which we'll cover in more detail below.)

Operationalization

To ensure organizations can operationalize models effectively, Altair RapidMiner can connect to your data just about anywhere it resides. This includes support for unstructured data in the form of web pages and documents (even difficult PDFs and complex excel files) data stored in popular cloud storage applications like Dropbox, as well as information that lives in big data technologies like Hadoop. This level of access can help you make sure promising models live up to their potential when your employees are relying on their predictions to make important decisions.

To help your teams create long-term value with their models, Altair RapidMiner also provides you with several ways to monitor them. To ensure models are staying accurate, users can measure and compare their latest performance against their expected error rates. Your teams can also identify early indicators for concept drift, which occurs when the relationship between inputs and predictions fundamentally changes. Lastly, Altair RapidMiner will also alert users when it detects the use of a data column that could be suspicious, which can help identify and avoid the creation of biased models.

Data Visualization

Data visualization is an important aspect of just about any analytics project, both because it can help users make sense of raw data and because it provides a user-friendly way to interpret results. The right data visualization software will make data analysis more accessible across your organization, which empowers your employees to make more informed decisions.

Altair RapidMiner allows business users, analysts, and engineers to spot anomalies, trends, and outliers in seconds. Your team can easily build, modify, and deploy sophisticated data visualization and stream processing applications with a drag-and-drop interface. Altair's stream processing and data visualization solutions are built for teams who need to make fast, informed decisions based on massive amounts of fast-changing telemetry, sensor, and trading data.

WRAPPING UP

As we've discussed throughout this eGuide, having a data science-ready architecture will save you a lot of headaches when you're working on high impact use-cases.

On a technical level, you'll need to ensure your systems scale effectively as your workloads grow, provide flexibility so that your solutions can run in different types of environments, and allow those solutions to be embedded within your business processes so they can create the most value possible.

Culturally speaking, you should focus on promoting data fluency across your organization so that analytics professionals, business experts, and executives have a shared language for analysis and decision-making. That doesn't just mean upskilling employees so they have a better understanding of the data they're working with and how to make it useful–it also involves breaking down silos and encouraging people with different backgrounds and responsibilities to collaborate when developing solutions.

If you can successfully address these areas, you'll lay a strong foundation for your data science projects, which in turn will help your organization make more informed, more impactful decisions.