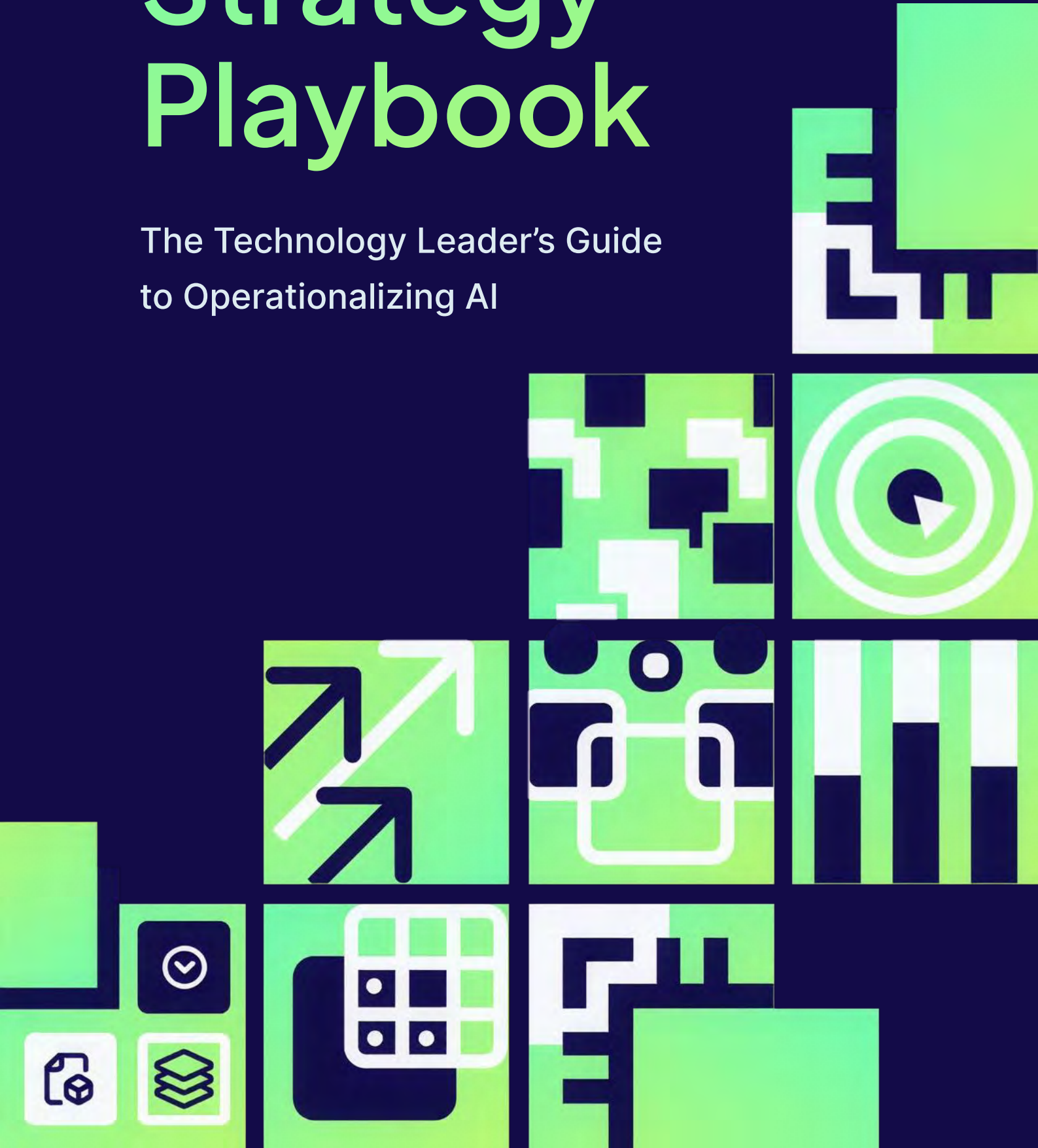


# The AI Strategy Playbook



The Technology Leader's Guide  
to Operationalizing AI



# Table of contents

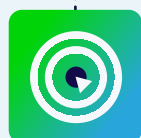


## Introduction.....02

Who is this guide for?.....03

Purpose.....03

How to use this guide.....03



## Part 1: Defining the Business Problem.....04

Audience and impact.....05

Identifying the problem.....11



## Part 2: Organizational Readiness.....14

Team composition and cross-functional readiness.....15

Dedicated team and ownership.....17

Prior experience with AI or GenAI projects.....18

PoC or long-term strategy?.....19

Time to production.....20



## Part 3: Calculating Return on Investment (ROI).....21

Common ROI metrics for AI projects.....22

Balancing immediate returns with strategic value.....23



## Part 4: Data Assessment.....25

Structured vs unstructured data.....26

Data locality and accessibility.....28

Scale of the dataset.....29

Single vs. multi-tenancy.....30

Data update frequency.....31

Handling sensitive data and compliance.....32



## Part 5: Technical Readiness.....33

Development languages and skills.....34

Existing tools and platforms.....35



## Part 6: Scalability and Growth.....36

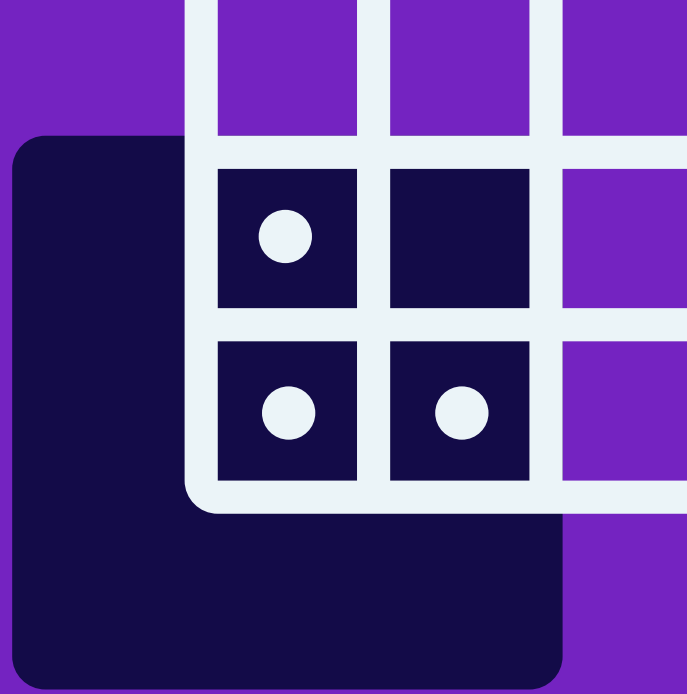
Anticipate data and usage growth.....37

Infrastructure readiness.....39

Cost and resource planning.....41



## Final Thoughts.....42



## INTRODUCTION

# The AI Strategy Playbook

The Technology Leader's Guide to Operationalizing AI

Byron Voorbach  
Field CTO  
Weaviate

## Who is this guide for?



Startup founders



Enterprise technical leaders



Individuals who want to bring AI to their workplace

## How to use this guide

Just getting started or refining an existing project?



Read from start to finish to build your understanding from the ground up



Only reference sections relevant to your current challenges

## Purpose

This guide is designed to help you:

It offers practical insights to make informed decisions, ensuring your project is well-defined and strategically sound.



Clearly define and articulate the business value of your AI project



Navigate internal conversations



Secure the necessary backing and resources to get your project off the ground



## PART 1

# Defining the Business Problem

The most successful AI projects solve well-defined business problems.

In this chapter, you'll learn how to frame that problem clearly, so you can make smart decisions, align your team, and get the support and resources you need.

# Audience and impact

When defining your business problem, determine if it's internal or external. Each type of challenge requires different approaches, stakeholders, and metrics to measure success.

## Internal audience / business problem



### What is it?

Internal problems typically involve issues with manual processes, data accessibility, or capabilities within your organization.



### What's the impact?

These problems often directly affect the productivity and efficiency of employees, operations, or internal systems.

**Let's look  
at some  
examples**



## Repetitive tasks slow teams down and burn out talent

Support teams can drown in repetitive, manual work.

You'll see slow service, rising costs, and high churn among your frontline teams who feel like their time is being wasted. And that's on top of compromised customer experience and resolution times.

### THIS INCLUDES:

- Context switching to hunt down past interactions
- Copying boilerplate answers
- Escalating simple tickets that should have been automated



## Poor internal search paralyzes productivity

When employees can't easily find what they need, they lose momentum.

Critical decisions get delayed. Work gets duplicated. Time gets wasted in Slack threads or waiting on subject matter experts (SMEs).

### THIS INCLUDES TIME WASTED SEARCHING FOR:

- Policy details
- Technical documentation
- Answers from past tickets



## HR drowns in questions that should answer themselves

HR teams spend hours fielding repetitive questions about benefits, policies, or onboarding steps – time that should be spent on building culture and strategy.

New hires get stuck waiting for access or answers, setting a tone of inefficiency from day one. When internal operations lag, employee experience and retention suffer.

### ✓ What results can you achieve with the right AI solution?

- Substantial improvements in productivity.
- Reduced manual labor.
- Enhanced accuracy.
- Increased employee and user satisfaction.



PRODUCTION-READY SELF-SERVE RAG PIPELINES

## Morningstar delivers a trustworthy, AI-driven financial data platform

Morningstar implemented an internal AI-powered knowledge base search system. This allowed even non-technical employees to access fast, accurate answers, significantly reducing time spent on manual searches.



“Through our Corpus API connected to Weaviate, users can build very powerful, low latency search engines in minutes with little to no code. Users can then also test different search algorithms without having to worry about re-indexing their data or that infrastructure at all.”

**Aisis Julian**, Senior Software Engineer, Morningstar

# External audience / business problem



## What is it?

External problems primarily impact your customers or end-users.



## What's the impact?

These problems usually involve issues with user experience, service delivery or converting and retaining customers.

**Let's look at some examples**



**Slow or inconsistent customer service drives churn**

**Customers today expect instant answers.**

When support teams can't scale effectively, response times lag, and customer experiences become inconsistent. Every delayed or subpar response results in frustration and an open invitation for your users to complain in public reviews or try a competitor.



**Irrelevant search and recommendations are killing conversions**

**When users visit your site, they're showing you what they want.**

If your recommendations or search can't keep up, they bounce, churn, or miss your best content and products. Poor UX leads to lost revenue, missed growth opportunities, and falling behind in your market.



**One-size-fits-all experiences are making you forgettable**

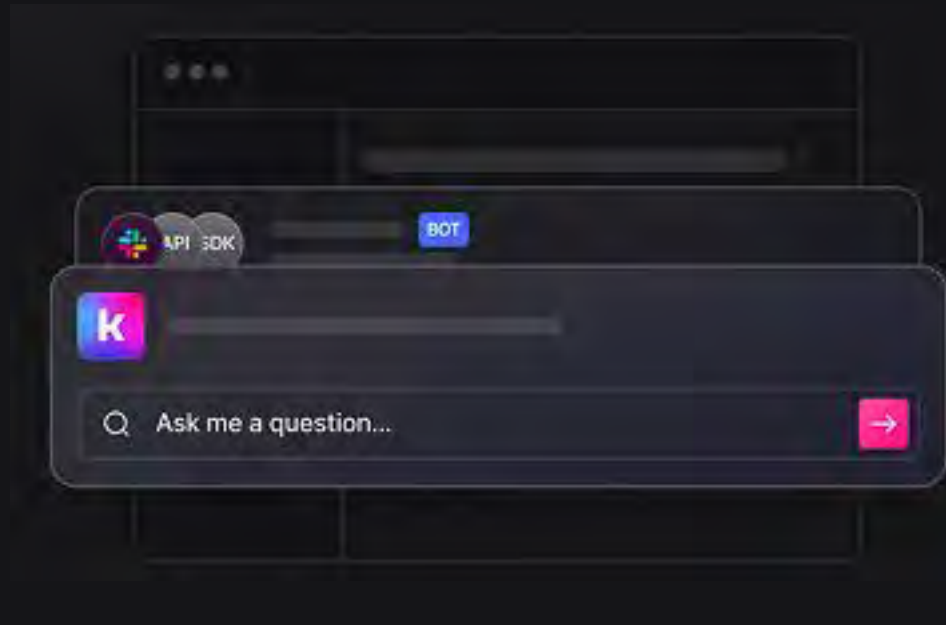
**Generic experiences don't win loyalty.**

If every customer gets the same homepage, email, or onboarding flow, you're just another vendor. AI makes it possible to personalize at scale, and if you're not doing it, your competitors are. In a world where everything is personalized, falling behind makes you easy to replace.



**What results can you achieve with the right AI solution?**

- Higher customer satisfaction and retention.
- Improved brand loyalty and purchase or trial conversions.
- A memorable user experience that makes your brand stand out.



## CHATBOT FOR TECHNICAL DOCUMENTATION

# Kapa.ai takes the pain out of finding accurate technical answers

Kapa AI developed AI-driven chatbots to convert complex, customer-facing technical documentation into easy to use chatbots.

They've helped support tens of millions of users, with CircleCI improving support response times by 28% and Monday.com increasing user engagement 10X.

# 28%

Faster response times

# 10x

Higher engagement

# Identifying the problem

Identifying the right business problem requires a thoughtful, structured approach. Begin by clearly stating the specific challenge your organization faces and break down the issue into concrete components.

## 1 Clarify the challenge

Define precisely what the problem is, making sure to have clear descriptions.

## 2 Determine your target audience

State who is impacted by this issue and how frequently it occurs.

## 3 Identify the impact

Evaluate and articulate the impact of this problem on your business operations, employees, or customers. Identify any measurable consequences such as increased cost, decreased efficiency, or declining customer satisfaction.

## 4 Explain the current limitations

Assess if and how your organization is addressing this problem and why existing solutions fall short. Are current methods overly manual, time-consuming, error-prone, or unable to scale?

## 5 Forecast what happens if the problem is not addressed

Consider the level of urgency solving this problem requires. Is it a foundational part of the business, a major risk to productivity or user happiness? Are competitors likely to outpace you?

## 6 Define your solution

Imagine the ideal state if the problem is solved.

## 7 Develop a measurement plan

Describe clearly how success will be measured. Whether through productivity gains, cost reductions, higher customer satisfaction, or improved competitive advantage.

# Scope out your target business problem using our framework

At \_\_\_\_\_  
we have an issue with

---

---

The people who are  
impacted by this  
problem are

---

Right now this challenge is causing the  
following problems:

---

---

---

---

Currently we are/n't addressing  
this problem. This looks like:

---

---

---

---

If we don't solve this problem in  
xx months, we run the risk of:

---

---

---

---

We propose to solve  
this problem with

---

---

---

We will gauge the success of the solution by:

---

---

---

---

# Example:

At Acme Corp we have an issue with poor internal documentation search.

The people who are impacted by this problem are our data analysts.

## Right now this challenge is causing the following problems:

- Reporting is taking up too much of their time and preventing them from working on new projects to enhance efficiency.
- We're missing big picture data patterns because our information comes from too many disparate places.
- Analysts are getting burned out.

## Currently we are *not* addressing this problem. This looks like:

- Analysts need to check 5+ different sources for relevant information.
- Catching multiple missed data points every week.
- Team morale is dropping.

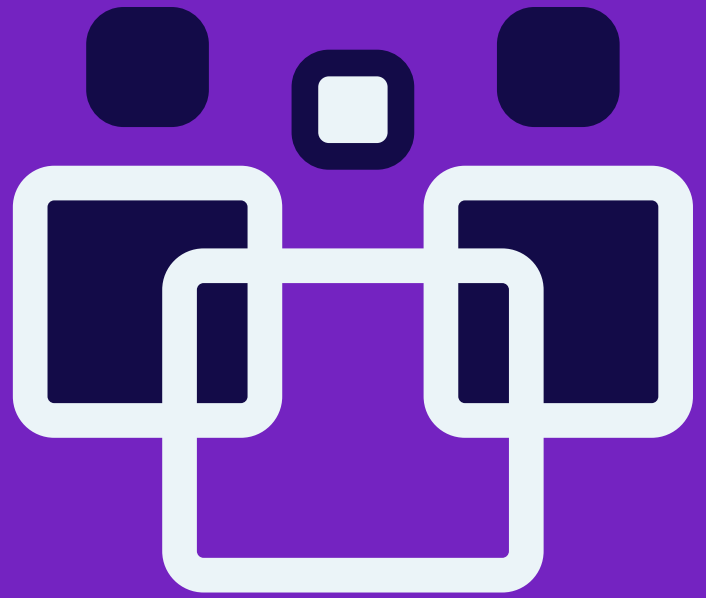
## If we don't solve this problem in 3–6 months, we run the risk of:

- Increased reporting delays.
- Slowing down other teams like finance, sales, and marketing.
- Team attrition from frustration and knowing our competitors already have internal AI search tools.

We propose to solve this problem with an internal AI chatbot that can bring all our investment data into one place, with citations.

## We will gauge the success of the solution by:

- Measuring amount of time saved for each financial report and the total time per week.
- Comparing the number of average report errors before and after chatbot implementation.
- Surveying analyst team on their job satisfaction.



## PART 2

# Organizational Readiness

AI projects run best when the right people and processes are looped in from the start.

In this chapter, we'll assess your organization's AI readiness and build alignment across roles, responsibilities, and workflows.

# Team composition and cross-functional readiness

At larger organizations, AI projects require collaboration across multiple teams like engineering, infrastructure, data, and security. In growing companies, one or a few people may hold multiple roles.

You'll need to identify if the right people are in place and whether they have a shared understanding of goals and responsibilities.



Do your developers understand how to work with embeddings, indexes, large language models (LLMs), and inference APIs?



Is your infrastructure team ready to manage services like AI-native vector databases?



Is your security team equipped to evaluate privacy, access control, and compliance implications?



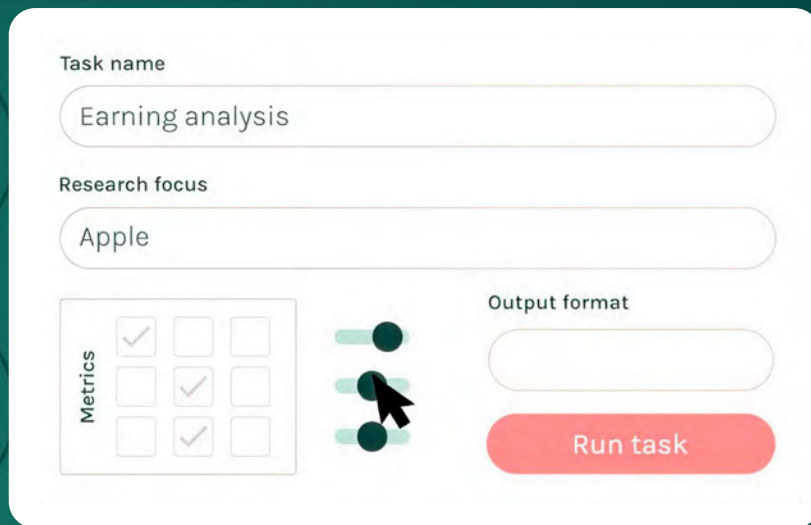
A lack of coordination across these groups often leads to project delays or incomplete solutions, while early alignment is a strong indicator of continued success.

If you are working with a growing team, consider how managed services or experienced vendors can help alleviate the pressure of operational tasks and enable you to build, test, and ship quickly.



## Take action

- Identify which stakeholders and influencers you need to consult to get your project approved.
- Map out the people and teams involved in building, deploying, and maintaining your AI solution.
- Plan to get alignment between these groups.



The screenshot shows a web interface for configuring a task. It includes a 'Task name' field with 'Earning analysis', a 'Research focus' field with 'Apple', a 'Metrics' section with a 3x3 grid of checkboxes (the first column is checked), a vertical slider with three green circles and a mouse cursor, an 'Output format' field, and a red 'Run task' button.

## AUTOMATED FINANCIAL DATA GATHERING AND ANALYSIS

# Finster Reimagines Investment Banking and Research with an AI-Native Platform

Our customer Finster, an AI platform for investment banking and research, was started in 2023 with a three-person founding team. Although they had considerable experience in AI, they reached out for additional guidance on how to optimize the structure of their application.



“Very early on we were fortunate to speak with Byron Voorbach, Weaviate’s Field CTO. He saved us several weeks of iterating on various retrieval methods and was able to guide us toward a specific solution that worked really well. It helped us deliver on the accuracy and consistency our users expect, and we still use the framework of that architecture now.”

**Seán Kilgarriff**, Product Lead, Founding Team at Finster

# Dedicated team and ownership



Do you have a dedicated team that can drive the project forward?

## ✓ Take action

- List the core team members responsible for the AI project.
- Agree who will be responsible for sourcing the tech stack, funding the project, delivery, success metrics, and ongoing iteration.

AI projects require sustained effort across product, engineering, data, and business functions. Without clear ownership, these projects can stall due to competing priorities or resource gaps.

While you don't need a large team to start, you should identify who will lead the effort, build the solution, ensure it delivers value to the business, and owns the budget.



# Prior experience with AI or GenAI projects



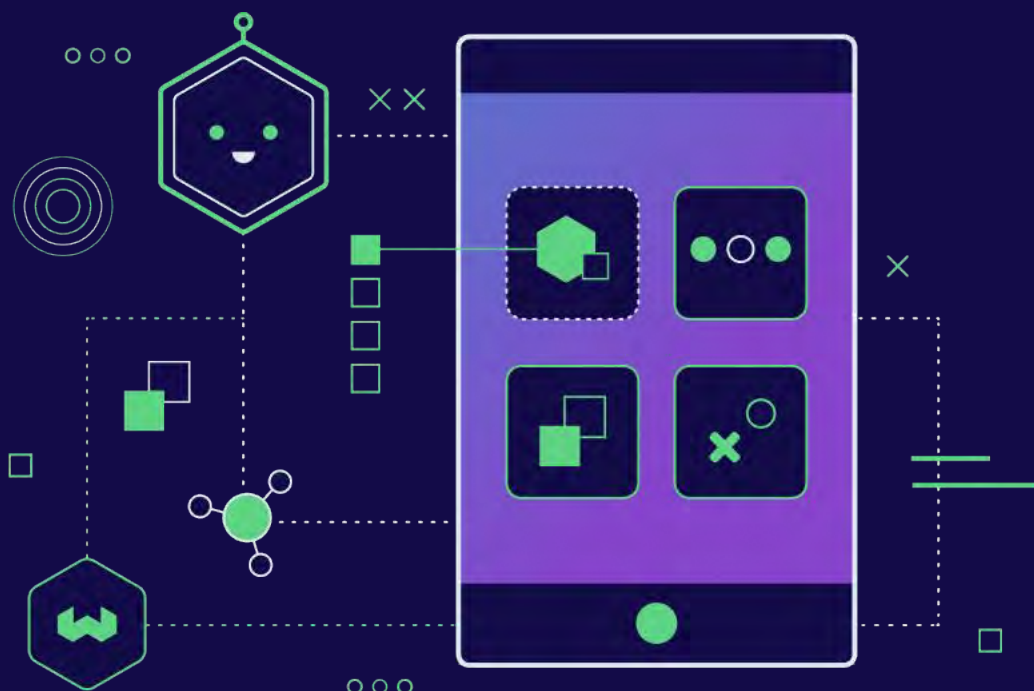
Have you already experimented with AI?

## Take action

- Document prior AI efforts within your organization.
- List what worked, what didn't, and what you could apply to the new project.

Successful or not, previous projects can provide important lessons about internal workflows, data challenges, or stakeholder dynamics.

Organizations that are new to AI might benefit from starting with a focused proof of concept (PoC) to build internal understanding, while more experienced teams may be ready for a larger strategic rollout.



# PoC or long-term strategy?



Are you approaching this as a short-term experiment or as part of a broader AI strategy?

## Take action

- Decide if your goal is to quickly validate an idea or to operationalize a long-term solution.
- Get clear on whether you plan to move this project into production.

Your answer will help clarify scope, resource investment, and stakeholder alignment.

PoCs are useful for testing technical feasibility and user value with minimal upfront investment.

Long-term projects require more structured planning around scalability, maintenance, and ongoing model or data updates.



# Time to production



Do you need to prove impact within weeks due to budget, competitive advantage, or resource availability constraints?



Do you have more flexibility to build toward strategic outcomes over months?

## ✓ Take action

- Establish a clear project timeline: what is the required date to ship this project?
- Align technical planning and stakeholders' expectations accordingly.
- Agree on a shared definition of what "production ready" means for your organization.

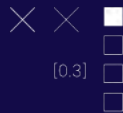
Time to value is often a key driver in AI initiatives.

Take the time to map out expectations to align teams and stakeholders.

It's also helpful to consider how this fits into your organization's roadmap. Is there a product release or customer pilot tied to the delivery timeline? Internal deadlines may introduce fixed constraints that affect what's possible within the projected time frame.

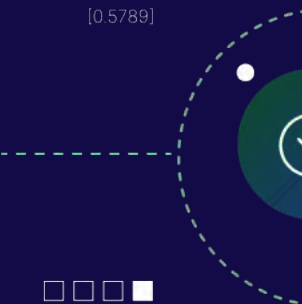
Finally, production-readiness involves more than just code. Think about documentation, observability, and support procedures. These often get overlooked, but can become critical blockers at the last mile.

[0.5789]



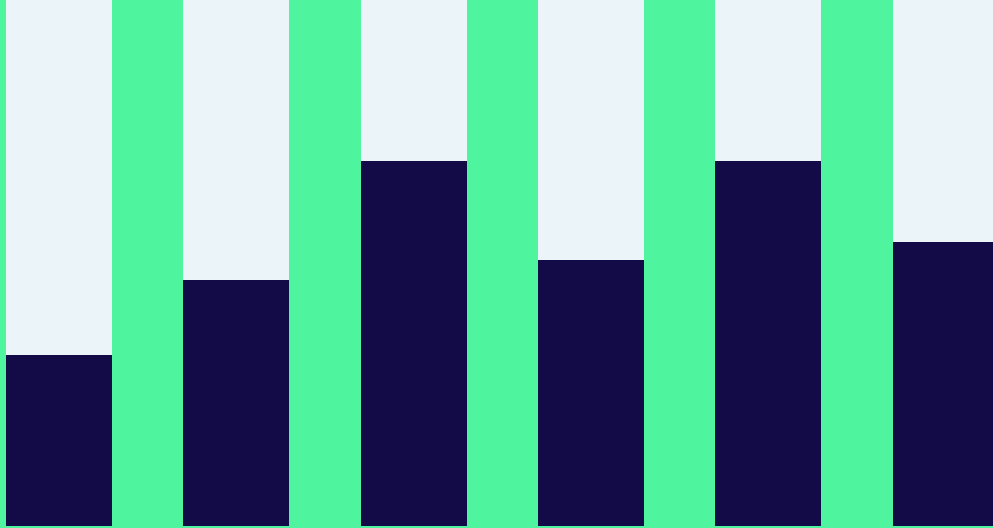
[0.3]

[0.5789]



[0.1023]





## PART 3

# Calculating Return on Investment (ROI)

Clearly articulating how you'll measure ROI will help you gain stakeholder buy-in and evaluate the success of your AI project.

In this chapter, we'll explore how you can assess and measure the impact of your use case.

# Common ROI Metrics for AI Projects

These are some of the most commonly tracked metrics among our customers, though your specific ROI metrics will ultimately depend on your unique use case. Identifying the right success criteria early helps justify the investment, and ensures your AI initiatives stay aligned with measurable business outcomes.



## Operational Efficiency

The right AI application can greatly reduce the need for humans to spend their time on repetitive and tedious processes. Measure time saved in key operations and multiply by the cost of employee time.



## Revenue Impact

AI capabilities can directly drive sales. Consider the benefits of increased conversions, average order value, or upsell and cross-sell strategies.



## User Experience Improvements

Measure increases in user engagement, satisfaction, and retention that result from enhanced AI capabilities.



## Scaling Capability

Calculate the value of expanded capabilities, like being able to support more customers faster, without proportional cost increases.



## Development Velocity

AI-native solutions reduce friction in the development process, allowing teams to innovate on projects faster. Track the reduction in time-to-market for new features.



## Infrastructure Cost Optimization

Companies often prefer managed AI solutions when in-house maintenance costs exceed subscription fees. Evaluate the total cost of ownership compared to maintaining self-hosted infrastructure.

## ✓ Take action

- Select and define the ROI metrics that align with your business priorities (efficiency, scalability, cost, user experience, etc.).
- Establish baseline data and measure post-implementation changes to demonstrate impact.
- Communicate results in terms that stakeholders value, such as cost savings or improved user engagement.

# Balancing Immediate Returns with Strategic Value

While quantifiable metrics are important, it's smart to consider the strategic value of your AI project.



**Competitive Positioning:** How does AI implementation position your company against competitors who may be investing in similar capabilities?



**Resource Reallocation:** Measure the value of redirecting talent from repeatable tasks to innovation and core business initiatives.

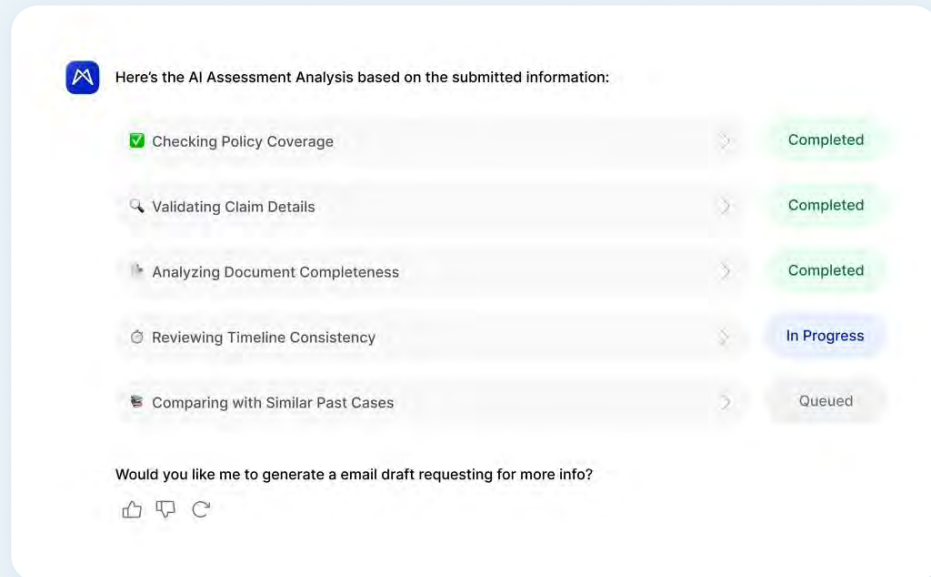


**Risk Mitigation:** Calculate the potential cost of falling behind in AI adoption compared to the investment required now.

Remember that ROI calculations should be tailored to your specific business context and use case. The most compelling ROI narratives connect technical improvements directly to business outcomes that matter to key stakeholders.

## ✓ Take action

- Assess how AI adoption improves your competitive position and enables resource reallocation.
- Compare the risks and benefits of immediate investment versus delayed adoption.
- Share both quantitative and strategic benefits with key decision makers and get their feedback.



## AUTOMATED INSURANCE OPERATIONS

# MarvelX is Scaling Insurance Processing at the Speed of AI

MarvelX solves operational inefficiencies in insurance and adjacent financial services industries. They've helped their customers 10x their claims processing output while maintaining the same team size, with 90%+ accuracy and 99.9% faster turnaround.

# 99.9%

Faster turnaround

# 10x

Improved processing output

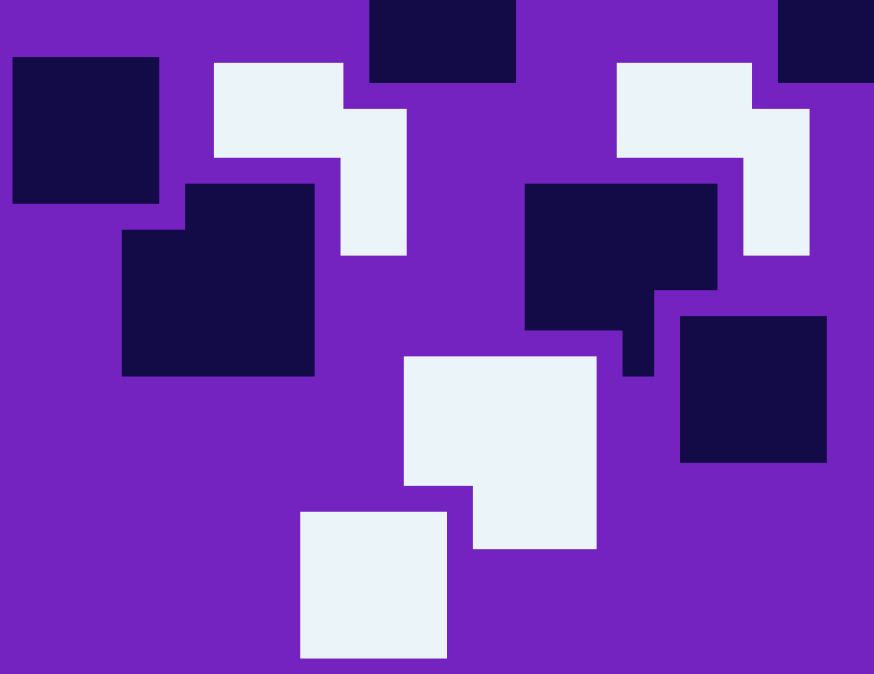
# 90%+

Accuracy



“Before MarvelX, I implemented Weaviate in banking... being able to have trust in something battle tested is very important. We also benchmarked multiple options and Weaviate’s speed blew the others out of the water. It wasn’t even close. As both our workload and customer base grew, the benefits just compounded.”

**Ali el Hassouni**, Founder, MarvelX



## PART 4

# Data Assessment

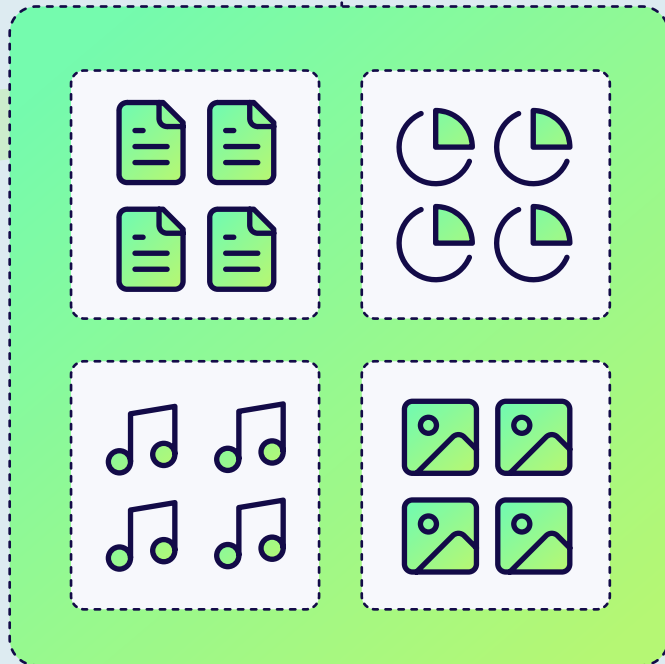
A comprehensive data assessment helps you identify the suitability, accessibility, and management needs of your datasets.

In this chapter, you'll review your data so you can leverage it effectively and responsibly.

# Structured vs unstructured data

Understanding whether your data is structured or unstructured significantly influences how you approach your AI project.

## Structured data



Includes clearly defined, easily searchable formats like relational databases, product catalogs, or CRM records. Structured data might involve product metadata in an e-commerce setting or user interaction logs from a digital service platform.

## Unstructured data



Includes information that doesn't follow a predefined model, like emails, user-generated content, customer reviews, multimedia files, or documents.

Different industries and companies use different kinds of structured and unstructured data. Here are some examples:



Organizations in the finance, insurance, or legal industries often process large volumes of documents (think contracts or compliance reports) extracted directly from PDFs or scanned files, to facilitate efficient retrieval and compliance monitoring.



A healthcare company might deal with images and many sources of structured and unstructured diagnostic documentation.



One of our customers offers an AI workspace platform that processes billions of emails for nearly 100K customers.

#### **Take action**

- Make a list of your key data sources and label them as structured or unstructured.
- Consider what data you would want to search across or make more accessible in your organization.

# Data locality and accessibility

Identifying where your data resides will help you determine the complexity and cost of accessing and managing it.



Data might be stored in various systems such as cloud storage services (AWS, Google Cloud, Azure), local servers, legacy databases, or third-party platforms. It's common for organizations to have data scattered across multiple databases, cloud environments, geographical regions, or legacy systems, which can complicate integration efforts.

## ✓ Take action

- Map out where your data currently lives.
- Identify any systems that would be hard to connect, require extra effort to extract data from, or would need special data isolation for security or compliance reasons.

# Dataset scale

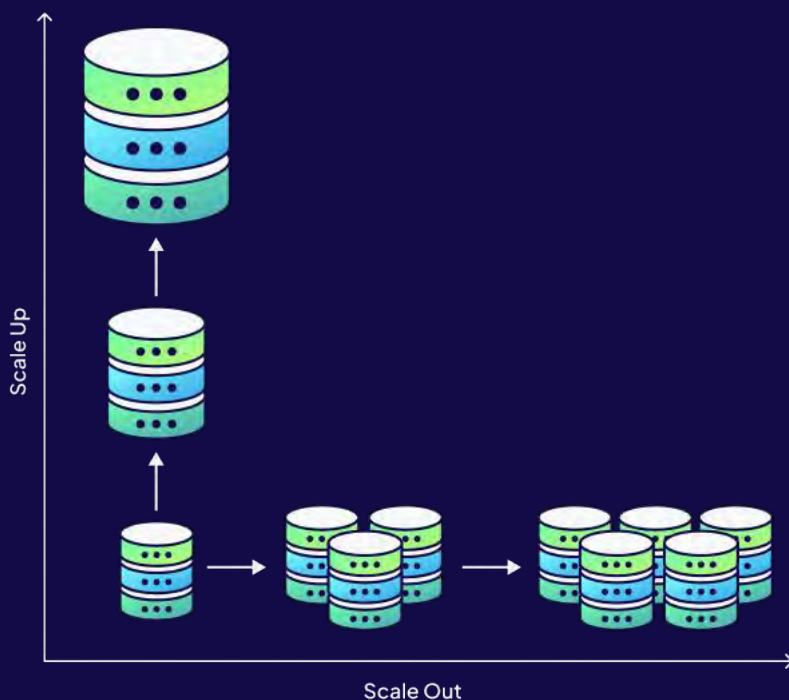
Whether you're working with thousands, millions, or even billions of records, the size of your dataset will determine the complexity, cost, and scalability of your project. Let's evaluate your capacity to process and manage this data effectively.



Consider a small retail business. They might have thousands of customer records that can be easily managed with a simple data infrastructure.



One of our enterprise eCommerce customers processes description data for **300M+** products to create personalized shopping feeds for **40M** monthly shoppers. Since they handle millions of records daily, they require a robust, scalable infrastructure capable of real-time data processing and analytics.



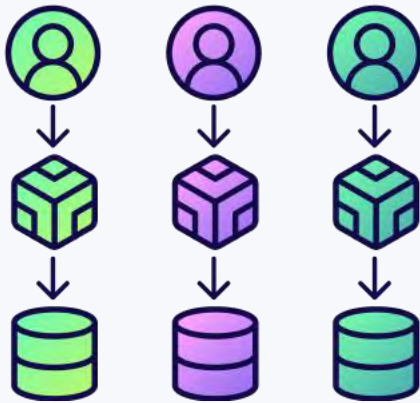
## Take action

- Estimate how many documents, records, or files you're planning to use.
- Determine if this number is expected to stay constant or grow over time.

# Single vs. multi-tenant architecture

Let's figure out if you should manage your data in a single, unified collection or segment it across multiple tenants.

## Single-tenancy



Allows you to manage all data within a single collection or index, where objects are globally accessible by your application logic.

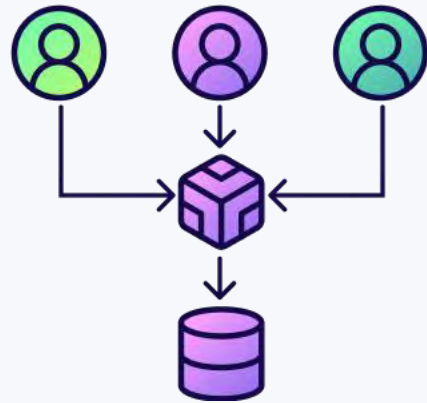
### Use it when...

- Data is centrally owned
- There is no need for isolation between different users or clients

### Use case example

The eCommerce customer I mentioned operates over a single dataset containing all product and user interaction data.

## Multi-tenancy



Enables you to separate data across distinct user groups, business units, or customers.

### Use it when...

- You're building SaaS applications
- Your platform serves multiple clients
- Your internal systems must adhere to strict data isolation and compliance requirements

### Use case example

The AI workspace platform I mentioned has tens of thousands of tenants to isolate customer data.

Choosing between these models has significant impact on scalability, security, query performance, and ease of management.

It's not just about scale. Your application *structure* will look very different if you want to search across 10M documents in a single collection as opposed to 100K isolated tenants that each require their own scoped results and access control.

### ▼ Take action

- Determine whether your application needs to isolate data between different users, customers, or teams.
- If so, consider the number of users you have and how you would segment data for these individual users.

## Data update frequency

How frequently your data changes or updates influences your data management strategy and the technologies you'll add to your AI stack.

Understanding update frequency helps you select suitable technologies to ensure timely and relevant insights from your AI solutions.



### Low data update frequency:

Relatively static datasets like quarterly sales reports or compliance documentation.



### High data update frequency:

Dynamic data streams requiring near-instantaneous processing, like real-time financial transactions or social media interactions.

- ### ▼ Take action
- Figure out how often your most important data sources change.
  - Identify if updates need to be reflected in your system in real time, hourly, daily, or less frequently.

# Handling sensitive data and compliance

Managing personally identifiable information (PII) or sensitive data requires adherence to strict compliance and security protocols, such as **GDPR**, **HIPAA**, **SOC 2**, or other industry or location-specific standards. Let's review some examples.



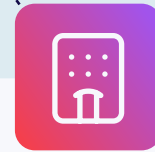
## Enhanced Network Isolation

Require VPC peering, PrivateLink mechanisms, or deployment entirely on-premises. These configurations ensure that data never traverses the public internet and that sensitive operations remain fully under enterprise control.



### Healthcare Providers

Need to handle patient data according to strict regulations, requiring robust encryption, secure data management practices, and advanced audit trails.

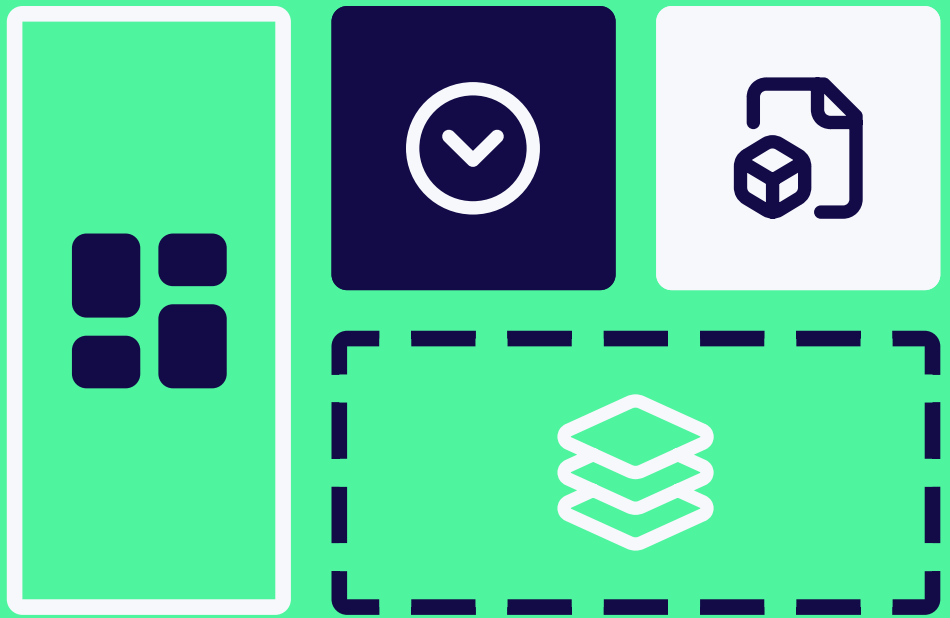


### Financial Institutions

Have to secure customer information to comply with privacy laws and regulations, requiring highly secure and traceable data processes.

#### Take action

- Review your datasets and identify any that include PII or regulated data.
- Check which compliance requirements apply to you (GDPR, HIPAA, etc).
- Figure out if your architecture requires private connectivity or isolated deployments.

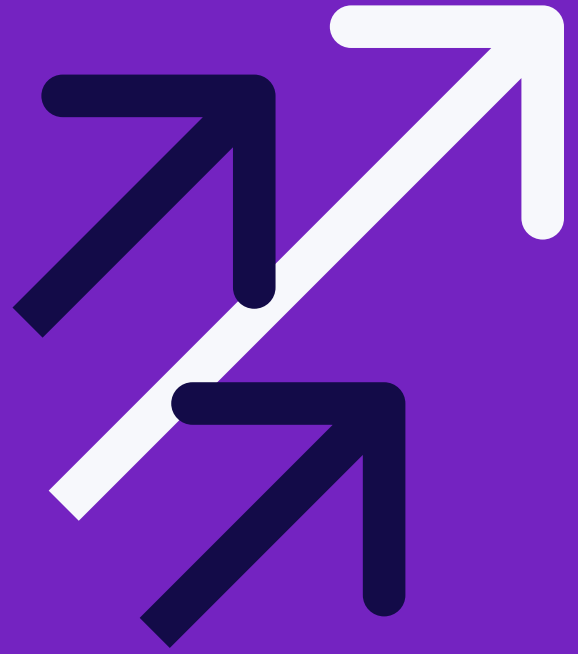


## PART 5

# Technical Readiness

Your tech environment can make or break your AI project. Is your team set up to build, launch, and maintain it successfully?

In this chapter, we'll assess your technical readiness so you can move faster knowing your choices align with the resources available to you.



## PART 6

# Scalability and Growth

The best AI projects are built with scale in mind. If your project can't keep up with demand, you risk outgrowing it before it delivers real value.

In this chapter, we'll explore what success looks like and how to build for that future.

# Anticipate data and usage growth



Start by asking yourself, "How much data do we have today, and how quickly are we expecting it to grow?"



Teams often underestimate how quickly data volumes expand once a solution proves valuable. Planning for scale prevents the collapse of systems that work perfectly during a pilot when they hit enterprise-wide adoption.

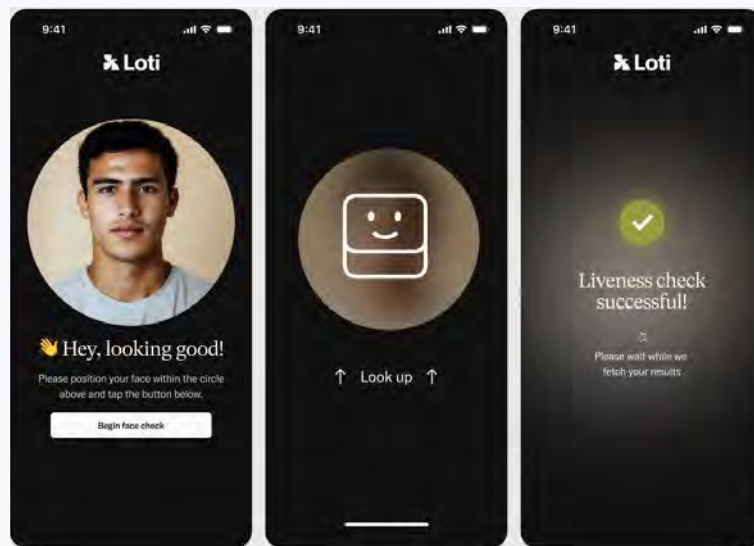
## Consider future scenarios:

- Will you add more data, customers, or data sources regularly?
- Will success lead to additional use cases or drive increased user access?
- Will you use new data types?
- Are there seasonal volume fluctuations?
- Do you anticipate increased query complexity or concurrent user growth?



## ✓ Take action

- Forecast your growth across some of the key dimensions of your use case: volume of data, number of users, frequency of queries, and expected latency requirements.



## DIGITAL IDENTITY PROTECTION

# Loti AI Fights Likeness Infringement and Digital Impersonation

Loti AI protects digital identities by finding and automatically removing unauthorized content like deepfakes and impersonations across the internet. They partnered with us to power their platform with one of the largest vector database deployments in the world.

**9**  
**BILLION**  
Vectors in production

**200+**  
Hours saved on  
database maintenance



“Weaviate's team was really great and solved every bottleneck we faced during our migration. Now, I never worry about our vector database, even with **9 billion** vectors. They've delivered maintenance and cost advantages, allowing us to focus on scaling our systems without concern.”

**Dr. Hirak Chhatbar**, Co-founder and CTO, Loti AI

# Infrastructure readiness



Scalability is about predictable, reliable, and cost efficient expansion. That means your infrastructure needs to handle initial demand and future growth. The last thing you want to do is a complete re-design six months down the road!

Start by assessing storage, compute capacity, and observability across services. From here, you'll need to decide if you want to manage infrastructure yourself or partner with a managed service provider.

If you plan to use a vector database, employing a third-party SaaS solution like [Weaviate Cloud](#) can offload much of the operational complexity around scaling, failover, and performance tuning. This allows your team to focus on building features, rather than maintaining infrastructure.

## ✓ Take action

- Sketch out two versions of your future state architecture: one where you scale in-house, and one where you rely on managed vendors.
- Identify pros, cons, and ownership responsibilities for each scenario.
- Use this to guide early architectural decisions and conversations with platform and/or procurement teams.

# neople



CUSTOMER SUPPORT AGENTS

## Neople Transforms Customer Service with Agentic AI

Neople agents are integrated with the tools businesses already use—like Slack, Microsoft Teams, and Zendesk—to work hand-in-hand with humans processing customer support cases. With Weaviate, they achieved massive performance improvements compared to their original database, all while freeing up engineers to work on new features.

**90%**

Faster search

**1000x**

More data objects stored



“We're really eager to learn from other people and to use services like Weaviate that specialize in vector databases to improve our offering. In the end, our main goal isn't to maintain a database, it's to deliver the best results and user experience for our customers.”

**Job Nijenhuis**, Co-founder and CTO, Neople

# Cost and resource planning

Scaling an AI project often requires more people, monitoring, security, governance, and platform tooling. Let's explore additional cost considerations to set you up for success.



## Human and operational costs

- Who will monitor the system?
- Who supports it when it breaks?
- Will you need new hires or training for existing teams?
- Do your security and compliance teams have the resources to evaluate and sign off on AI workloads?



## Cost control strategies

As an example, cost control strategies related to vector databases might include:

- Embedding model dimensionality evaluation
- Applying vector compression/quantization techniques
- Offloading cold data to cheaper storage tiers
- Embedding caching
- Using a managed service to free up your team's time to work on new features



## Inference costs

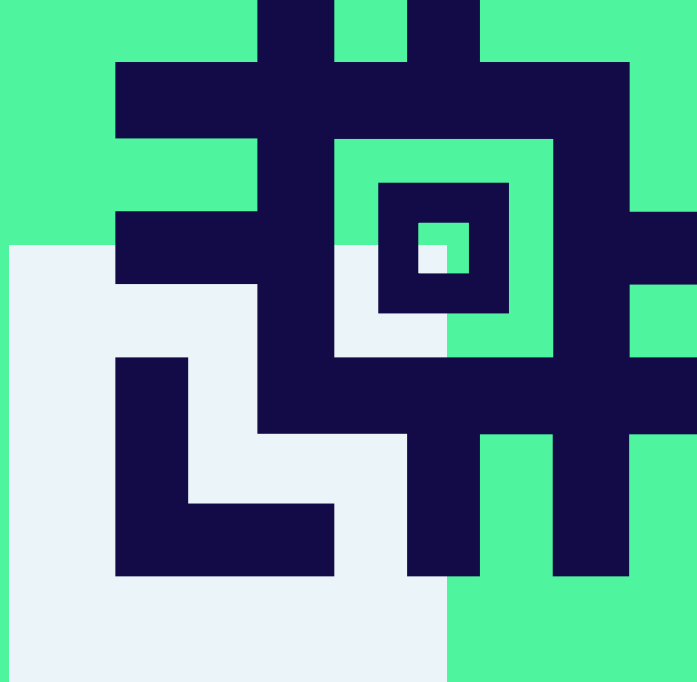
Not all growth is linear. Inference costs on embedding models or LLMs might rise disproportionately relative to the increased usage. Or, they may come down as LLMs become more efficient. Costs can spike if you don't monitor collection sizes, or if data isn't de-duped before ingestion.

Identifying these variables can help you mitigate risks and stay within budget. Decide if it's more resource-efficient to have your team build and run certain components in-house, or if offloading work to a managed solution makes more sense.



## Take action

- Identify who owns, and if possible, what the budget is for your AI use case.
- Build a rough cost model for the project at 1X and 10X scale.
- Review this plan with your finance and operations stakeholders.



# Final Thoughts

AI is no longer just an experiment. It's what users expect, and it's becoming central to how companies drive impact and stay competitive.

Succeeding with AI starts with building a smart strategy that incorporates a clear business problem, your people, your resources, and your business requirements.

Throughout this guide, we've explored the major components that influence the feasibility and impact of an AI project:



Defining a clear, tangible business problem that urgently needs to be solved.



Knowing your internal and/or external users.



Planning on how you'll measure and define success.



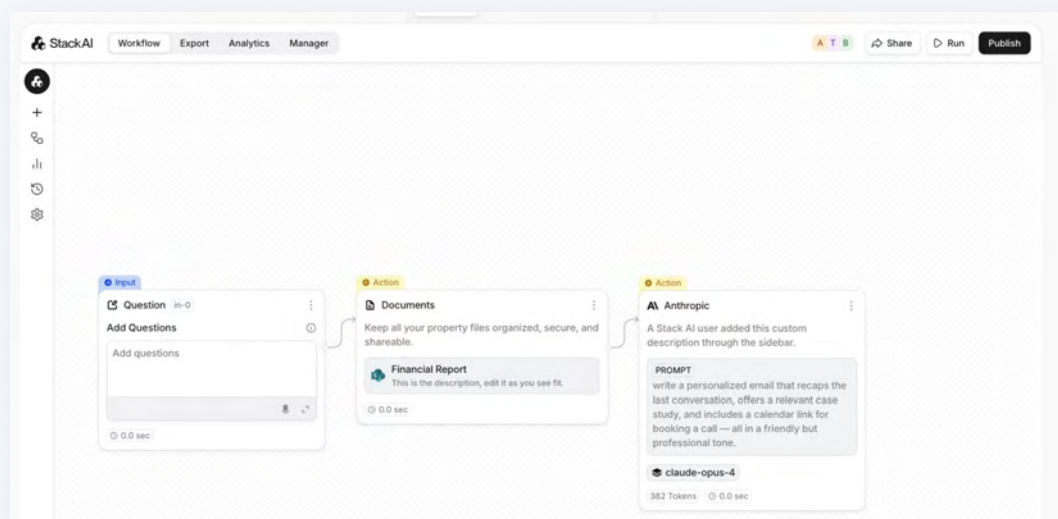
Understanding your data landscape and requirements.



Qualifying your technical and organization readiness.



Planning for scalability, cost, and operational sustainability.



NO-CODE PLATFORM FOR AGENTIC WORKFLOWS

## Stack AI Delivers Lighting-Fast Agentic AI for Enterprises

Stack AI chose Weaviate to power their enterprise AI orchestration platform. This decision resulted in significant cost savings, enhanced customer retention, and the ability to serve security-conscious customers through flexible deployment capabilities.



“The biggest benefit of using Weaviate isn't just the technology—it's the team behind it. The level of support we receive through their engineering team and support channels has been company-saving help.”

**Antoni Rosinol**, Co-founder, Stack AI

Having advised hundreds of companies on AI readiness and implementation, I've learned that most successful projects don't begin with answers, but rather with good questions. The questions in this guide are designed to help you avoid ambiguity and surface the realities of what it takes to go from prototype to production.

If you work through one or more of these sections, you'll make more informed decisions, and build the trust and internal alignment you need to move quickly and effectively.

Whether you're a startup looking to embed AI features in your product, or an enterprise modernizing a legacy system, the path forward is about making smart, thoughtful progress.

**Ready to explore  
your AI infrastructure  
options?**

[Reach out to our team](#)

## About the author:

Byron Voorbach has spent over a decade in the search domain, consulting hundreds of companies and aiding in implementing large-scale search systems. As the Field CTO at Weaviate, he collaborates with customers globally to harness the power of semantic search in their operations. He's also a regular conference speaker and active contributor to open-source projects. When he's not helping companies build and scale their AI projects, he enjoys spending time with his two cats, scuba diving, and traveling the world.

