

 WEKA

# The AI Project Survival Guide



## KEY QUESTIONS TO

# Supercharge a Winning AI Strategy

**Are you actively planning to drive your business outcomes with AI?** With the recent advancements in technology, the business landscape is rapidly changing, and there is no doubt that AI is at the center of this tech revolution.

As organizations strive to stay at the forefront of their respective industries, the stakes for AI projects have never been higher. [Big tech is spending billions on AI](#), but mandates to incorporate AI into business operations are becoming commonplace in companies of all sizes. Overall, [AI is expected to account for 10% of IT budgets](#) in the tech sector, which is a radical increase from 2023 when AI only accounted for 1% of the IT spend. And as AI continues to infiltrate every sector of the economy, its impact only stands to become more pronounced and its adoption more critical.

Unfortunately the desire to incorporate AI into a business strategy and deploy an AI project does

not guarantee success. In fact, some sources estimate that [the failure rate of AI projects might be close to 80%](#). The data-heavy nature of AI requires that organizations rethink, not only their strategy, but also their goals, approach, and tech stack. Deploying a successful AI strategy requires asking the right questions and making sure that your organization has the infrastructure and capabilities to power a modern data pipeline.

In this rapidly changing environment, it can be challenging to know where to start with AI, but having an AI project guide can help set you up for success. WEKA's CTO, Shimon Ben David, has distilled the 5 key questions AI leaders need to have top of mind when embarking on an AI strategy.

So let's dive in and get started, not just dreaming about deploying AI, but actively planning an AI strategy built for the needs of an AI data pipeline and designed for success.

# 01

## Do You Have a Clear Vision and Goal for Your AI Project?



**While AI is undoubtedly a transformative technology** and changing the business landscape for almost every industry, casting a vision and defining clear outcomes for any AI project is key to meeting goals and ensuring a successful project.

Amazingly many companies do not have a clear vision of the goal they want to achieve from an AI project. "A lot of companies will start with 'We know that AI is a game changer, so let's see what we can do with it,'" says Shimon Ben David, CTO at WEKA.

Like an explorer preparing to reach a destination, a project leader needs to establish a final end point and then provide a map that includes specific directions to follow for each step of the journey. In addition to setting a goal, any successful roadmap needs to include the key questions that will help you plan out the project's path and reach your goal.

"Companies need to maintain a constant set of questions, and chances are good that they will morph with a project's progress, but it's imperative that you have an initial set of questions at the beginning to give yourself a place to start," says Ben David.

### PRO TIP

The key here is to create a good AI team with the ability to ask and answer these initial questions. This team could include software engineers, business leaders, subject matter experts, and possibly even customers within the environment.

Keep in mind that the questions a company creates to get to the final goal can change and evolve as more data is gathered. If you've chosen the right goal, it should stay the same, but the steps to get there might change as you encounter roadblocks and obstacles. If you haven't identified the right goal, asking questions will make that clear so you can pivot in the right direction.

# 02

## What Are Your Data Requirements?

**Data is the nexus of every AI and machine learning project,** and the model you will create will only be as good as the data that you train it on. Knowing what data you need, where it is coming from, and how to assure that you have quality data is key.

After an AI project team has identified the goal or specific problem that AI can solve, the team can begin to ask questions to determine the data or variables required to achieve the goal or resolve the specific problem. Most companies begin an AI project thinking they have enough data to answer the questions, but often a good portion of the data is missing, or the data that they have isn't useful to answer the question. In his experience, Ben David says that he has never encountered a company that has collected too much data.

### Where Will I Get the Data I Need?

If you find yourself needing more data, the next step would be to determine where you can get the data you need. Do you generate it, do you buy it, or do you rent it?

For example, a medical company embarking on an AI project involving genetics might look at data in a public genome database, but then the researchers might discover that they do not have the data needed for their particular AI model, in which case they might

#### PRO TIP

Keep everything! Companies tend to acquire a vast amount of data, distill it when creating the AI or ML model, and then either store the raw data somewhere never to be accessed again or, even worse, delete the unused data. Ben David says that data can be critical later when reassessing a particular model where raw data is needed again.

need to conduct their own experiments. Alternatively, perhaps they need only a single piece of data in an image versus looking at a complete set of labeled data.

"You want to make sure you know where you will acquire the data at the starting point of the journey, but also with the understanding that this could change along the way," Ben David says.

And data acquisition is not a one-and-done proposition. “You need to plan ahead for when and where you will get your next batch of data and take the steps to acquire it, often in parallel with your other work” says Ben David.

Sometimes companies have to come up with their own data to fill in for what’s missing. The tools that you would use to extract your data set would vary depending on the

type of data you need to collect. For example, Google Analytics provides website visitor data and metrics, but you might also have a customer or contact database through Hubspot, Salesforce, or numerous other services.

There are many places to find data to train your models. [Bright Data](#), [Apiscrapy](#) are a couple potential resources to get you started.

## How Do I Ensure Data Quality?

The conversation around data quality is rapidly becoming a central feature of the AI revolution. Feed your model bad data and you end up with a model that doesn’t work. When [Great Expectations, a leading open-source platform for data quality, conducted a survey about data quality](#), they found that 91% of the organizations they surveyed thought that data quality impacted their businesses performance.

Since data quality is directly tied to business outcomes, it is important to focus on not only collecting and procuring data, but also making sure that the data you are using is the right data for the job.

“

Every single company I’ve worked at and talked to has the same problem with a single exception so far—poor data quality.”

**RUSLAN BELKIN,**  
VP OF ENGINEERING, SALESFORCE.COM

Data quality is measured by a number of factors including the following:

- Accuracy
- Completeness
- Validity
- Consistency
- Timeliness
- Integrity
- Uniqueness

When people suggest that getting more data will help, they are often implying that they need to acquire a broader data set that meets high quality standards. Thomas C. Redman, president of Data Quality Solutions, believes companies often waste critical resources in dealing with bad data and [“bad data makes any sort of analytics or AI far more difficult.”](#)

Bad data can simply be that has not been “cleaned up,” or it may contain missing fields, duplications, or the data type is not in the correct format, such as dates written in text instead of the date format. But even data that has been cleaned may not truly be quality data if it is too specific or presents biases, such as problems generated in facial recognition or gender-bias.

One example was the data that was discovered in Amazon’s resume scanning application. The data may have appeared to be good initially but turned out to be bad after the algorithm kept eliminating female resumes because the model didn’t account for fewer female resumes in the historical data. This mistaken elimination indicated that the historical data was not broad enough.

In order to make sure you have quality data to train your model, it is important to be aware of the variety of dimensions that are required to have truly quality data. Taking the time to assess your training data based on the data quality elements listed above and making sure you have data governance processes and accurate documentation in place are key when planning your AI project.

# 03

## Is Your Infrastructure Up to the Task of Powering an AI Data Pipeline?

### Why Can't a Legacy Infrastructure Keep Up with my AI Pipelines?

One way to jeopardize the success of your AI project is to run it on a data infrastructure that is not up to the task of an AI pipeline. Traditional storage systems are ill-equipped to meet the demands of modern AI workloads, and although they have been improving, they still often lag behind the enormous leaps that have occurred in recent years in compute and networking.

The scale of data processed by AI has reached unprecedented levels, with models handling trillions of parameters and datasets at a petabyte scale. Storage systems, designed for conventional data processing tasks, struggle to efficiently manage and access such vast volumes of data. And the speed at which data needs to be processed has significantly increased with the advancement of AI algorithms and infrastructures. Despite orders of magnitude improvements in network speeds and GPU performance, legacy storage systems haven't kept up and lack the necessary throughput and latency characteristics to support the rapid data access requirements of AI applications.

As you plan for a cutting-edge AI project that leverages the fastest compute and latest networking capabilities, don't make the mistake of running your project on an infrastructure that is not designed to meet the scale, speed, and agility requirements of modern AI workloads.

**Jonathan Martin, President of WEKA**, unpacks the business benefit of a modern data platform like the WEKA Data Platform.

With the right infrastructure, organizations are "able to massively reduce the time it takes to do artificial intelligence. So the time it takes to do training of models or inference can be compressed by anywhere between 10 and 100 times. So if you can imagine being able to do 100 times more work in a single day—that's what the WEKA Data Platform gives you."

## What Is My Plan to Move and Access Data?

As they work to process the AI models, companies often discover they did not plan for how they would access and move their data. Imagine a company with divisions all over the world, generating petabytes worth of data in multiple locations across different continents. “Do I try to process it where it was

“There is no one right answer,” says Ben David, “but if you do not think about how you are going to move, store, and access your data when you are planning your project, then you are more than likely to have a problem.”

created, or do I try to move petabytes of data somehow between sites worldwide?” Ben David asks. “It’s one of the critical things that some teams may not consider when they are launching an AI project.”

One option is to centralize the data in a single data center, but moving data includes the possible need to compress data or physically ship it instead of transferring it across the cloud, which can get expensive quickly. Making sure the data is secured is also an issue, as some data cannot be moved due to local or federal regulations. Finally, by the time the data arrives at the site of AI processing, you might find that it’s already obsolete.

## Where Should I Deploy my AI Project?

AI and ML projects can be successfully deployed on-premises, in the cloud, or hybrid. The important thing is to make sure that your deployment strategy aligns with your company’s overall strategy and won’t conflict with changes or modifications down the road.

WEKA was designed to flip the script on legacy infrastructures and offer the ultimate in flexibility when it comes to deployment. Liran Zvibel, CEO of WEKA, explains that “We envisioned a world where customers wouldn’t be forced to make compromises in their

data environments. Where a single software-based solution that could be deployed anywhere would deliver a simple, seamless data management experience across edge, cloud and on-premises environments.”

Smaller companies might start with a cloud environment because they think it’s faster and less expensive, yet they may find that the costs become larger as the projects grow, and it may make more sense to move to an on-premises environment. When you’re starting out, it’s important to start where your data is.

### PRO TIP

As you plan for growth, keep in mind that the ability to scale on-prem, in the cloud, or even between the two environments may become an important feature of your ongoing strategy.

# 04

## How Do You Plan to Scale & Maintain Your Project?

### What Bottlenecks Should I Be Looking Out For?

Running AI workloads requires balancing each part of the AI pipeline, and as you scale up, a data pipeline that is out of balance can lead to costly bottlenecks that can slow outcomes. One of the primary potential bottlenecks in an AI workflow can be the ability of a

“

As you scale up the bottlenecks move from place to place and you always have a little bit of construction to do.”

**PATRICK BANGERT,**  
[WEKA'S SAMSUNG FIRESIDE CHAT](#)

GPU to process data. Research shows that GPUs can actually spend up to 70% of their time idling while they wait for data to be delivered. Because GPUs are fast, they require storage infrastructures that are capable of delivering all the data at the right speed. Having the right software to access the data more efficiently is essential for feeding GPUs. Data hungry GPUs sitting idle at the end of a congested data pipeline cost time, money, and lead to an unsustainable footprint.

There are many ways that a data pipeline can experience slowdowns that impact the utilization of GPUs. Copying data between infrastructure to optimize each part of a data pipeline can create data stalls during every copy operation. IO patterns also have the potential to create bottlenecks in the data pipeline. When data sources are brought into a training pipeline, the data often is in the form of LOTS of small files. This can be a scaling problem that many storage systems struggle to handle. Reads and writes early in the pipeline paired with lots of random reads from the model training, the need for low-latency writes for checkpointing, and metadata overhead creates a vicious cycle that puts immense pressure on the storage system to try and keep up with each stage of the data pipeline.

Whether it's data stalls, data copies, or even delays related to IO patterns and checkpointing, it is important to keep an eye on the places in your data pipeline where you might be experiencing bottlenecks and underutilized GPUs.



“We wanted to train a model on the 30 million files we had, but the models are fairly large, with 30-50 epochs, a timeline of up to four days, and a lot of random-access-file lookups. GPUs are quite fast and hungry for data—you want to feed them as much data as you can,” says Jon Sorenson, VP of Technology Development at Atomsie. With WEKA, “We could now consider experiments that earlier—because of all these headaches—might take us three months to figure out how to run. Now we can do this same experiment in less than a week.”

## How Do I Account for Scaling during Training vs. Inference?

When training an AI model, scale typically refers to the size of the data set you are using for training. Training a model may require the introduction of new data sets throughout the process. With this influx of data and the iterative nature of the training process, it is essential that your data pipelines be set up for success. Optimizing your data pipelines, reducing bottlenecks, and planning for sufficient compute, networking, and storage capabilities that can grow with your project are vital for the training phase.

“The greatest challenge in designing hardware for neural network training is scaling. Doubling the amount of training data doesn’t mean doubling the number of resources used to process it. It means expanding exponentially.” HPCWire

When it comes to inference, while the pressure on the data center may be reduced, performance and latency

become central to executing a successful inference strategy. Scaling during the inference phase is related to the number of requests that the system receives as the model is being used. Since inference is so closely tied to throughput, making sure that your infrastructure is designed to optimize I/O bandwidth is essential. It is also imperative that your storage system has enough memory to accommodate the training model(s) and the input data.

As inferencing is also increasingly being impacted by the introduction and acceleration of Retrieval-Augmented Generation (RAG), it is important to be sure that any infrastructure solution will be able to meet RAG requirements. Because RAG pipelines are heavily dependent on vector databases, the lower the latency for the vector database response to the LLM, the more efficiently the LLM can execute the query.

## How Does Fine Tuning my Model Impact my Strategy?

Because much of AI and ML is based in software, developers have previously been prone to adopt a “set it and forget it” approach, which can be disastrous. In order for models to continue operating effectively, they often need to be fine tuned. Fine-tuning not only involves being ready to change the model regularly but also understanding how practitioners can change different variables within the model to achieve different results. But there is no hard and fast rule for how often a model needs to be finetuned.

One essential component of making sure your model is operating optimally is accounting for model drift. As time goes by, all models face the risk of becoming

less accurate. Naturally the type of model and the data used to train the model will impact that timeline for a model to experience drift. Models that are trained on people’s views, perceptions, and culture may require more consistent finetuning, for example, than a model that is trained on relatively consistent, scientific data.

Regardless of the type of model, however, data engineers need to employ model observability and consistently monitor their models for accuracy so they know when they need to retrain. Fine tuning may also be more necessary if, at the beginning of the project, that data used to train the model is subpar.

### PRO TIP

At some point data scientists may decide to move to a different neural network for their AI model, which might require creating something new rather than simply fine tuning. AI project teams should consider at the beginning of their project about how they might deploy a new model and account for the potential infrastructure requirements should the need arise at a later date.

# 05

## What Does It Take to Future-Proof Your AI Strategy?



### What Will my AI Project Look Like a Year from Now?

Few advancements have moved as quickly as the development and adoption of AI, so it might seem like a crazy question to try to predict where your project will be in a year. But this question is less about predicting and more about having an eye on the future to make sure you are planning for growth and change.

"A company might begin their project on premises, with one or two data scientists running from their laptops with an external GPU," says Ben David. "If everything works out in a year, then they may have 20 data scientists, and they will then need a heavier infrastructure. It's important to think about potential growth when you begin so you can plan ahead for that growth."

As data grows, so does the need for a data platform that can keep data pipelines saturated. One frequent and expensive mistake companies make is not planning for the significant data growth over the course of the project. Using a modern data platform can dramatically alleviate the cost and management burden, helping you to keep productivity high as data increases.

Your project may change, where you store your data may change, and your model will likely need fine tuning or use RAG (Retrieval Augmentation) techniques. But without a doubt, regardless of how your project grows and changes over time, you will continue to need a modern data platform designed to support superior performance, smooth data pipelines, low latency, and the ability to store and access your data quickly and efficiently.

## How Can I Future-Proof my Project?

In light of the estimates that predict up to 80% of AI projects fail, driving toward success mandates that you future-proof your project when planning your deployment. Architecting an infrastructure built on the foundational principles of a modern data platform that can scale with your project ensures that you have room to grow without the risk of incurring additional bottlenecks, slowing your data management, or not having the resources to manage growing data sets.

**There are a few principles to keep in mind as you look to the future of a successful AI project.**

- **Keep an agile mindset.** By their very nature, AI projects do not run in a straight line. Training a model requires iterations, and you may need to procure additional quality data as you move forward. You will also need to adjust your strategy as your data set grows and as you transition from training to inference and then fine tune.
- **Plan for growth.** As your project matures, you will invariably be managing more data, so it's important to have scaling capabilities built into your plan. As more data is fed into your data pipelines, you also run the risk of data stalls and inefficiencies. Preparing for these expected changes across the life of your project will ensure that you do not have to re-architect your project in the middle of it.
- **Choose a data platform that can grow with you.** As AI projects grow over time, it is imperative that you have a truly modern data platform designed with the flexibility to grow with your project. Whether your growth leads to issues around scaling, changing where you manage your data, saturating increasingly growing data pipelines, or even maintaining low latency to power inference, you want to plan ahead so that you can avoid needing to upgrade your data infrastructure as you go.

Ben David says he sees many companies kicking off AI projects with high hopes for success, but the team has not taken a holistic view of the entire project, so down the line they run into trouble when it comes to growth. "We see projects that are starting with some environments that are adequate for one to five data scientists, but then the environment expands and suddenly they need additional infrastructure," he says.

Having a clearly defined strategy around how to effectively manage the growth and scale of the project can set you up for success, even as you move into the ever-changing environment of AI.


## Key Takeways

- ✓ Don't just implement an AI project because it is new and exciting. Have a clear vision and business goal that informs your project plan.
- ✓ Identify your data requirements at the outset of your project and ensure that you will be able to procure quality data.
- ✓ Use a data platform designed to meet the scale, speed, and agility requirements of modern AI workloads.
- ✓ Have a plan for optimizing your data pipelines during both the training and inference stages of your project.
- ✓ Think ahead. With the field of AI changing so rapidly, look into the future to make sure you have an infrastructure solution that can grow with you.

## Harness the Power of AI For Your Business

At WEKA we know it is possible to power AI workloads effectively and efficiently as you seek to supercharge your business because we have worked to help many, many customers streamline and accelerate their data pipelines. Keeping these key questions in mind will not only help you launch your plan, but they will also help you sustain your project as you move into the future.

But don't just take our word for it. You don't have to be one of the 80% that struggles to see their AI strategy succeed. Armed with a clear path and access to a data platform designed to power AI workloads, you can join the ranks of companies using AI to take their business to the next level with WEKA.

<p><b>Read more about how we make successes happen.</b></p>	<p><b>Genomics England</b></p> <p>"We needed something that's much more scalable than existing NAS solutions — an infrastructure that could grow to hundreds of petabytes. Our existing solution couldn't provide that scale and wasn't performing as well in these magnitudes — that's what drove us to WEKA."</p> <p><small>David Ardley, Director of Infrastructure Transformation</small></p>	<p><b>cerence</b></p> <p>"We looked at our legacy architecture, and instead of taking an evolutionary approach, upgrading every component, we took the revolutionary approach. WEKA cost-effectively enables the use of both POSIX and object storage with performance and latency that is far superior to any other solution."</p> <p><small>Erin Lee Collins, Chief Information Officer</small></p>	
---	---	---	---

**We would love to meet with you and have one of our experts show you how the AI-Native WEKA Data Platform was designed from the ground up to turbocharge AI data pipelines and help you launch a successful AI strategy.**

**Contact us to schedule a call today**



weka.io

844.392.0665

