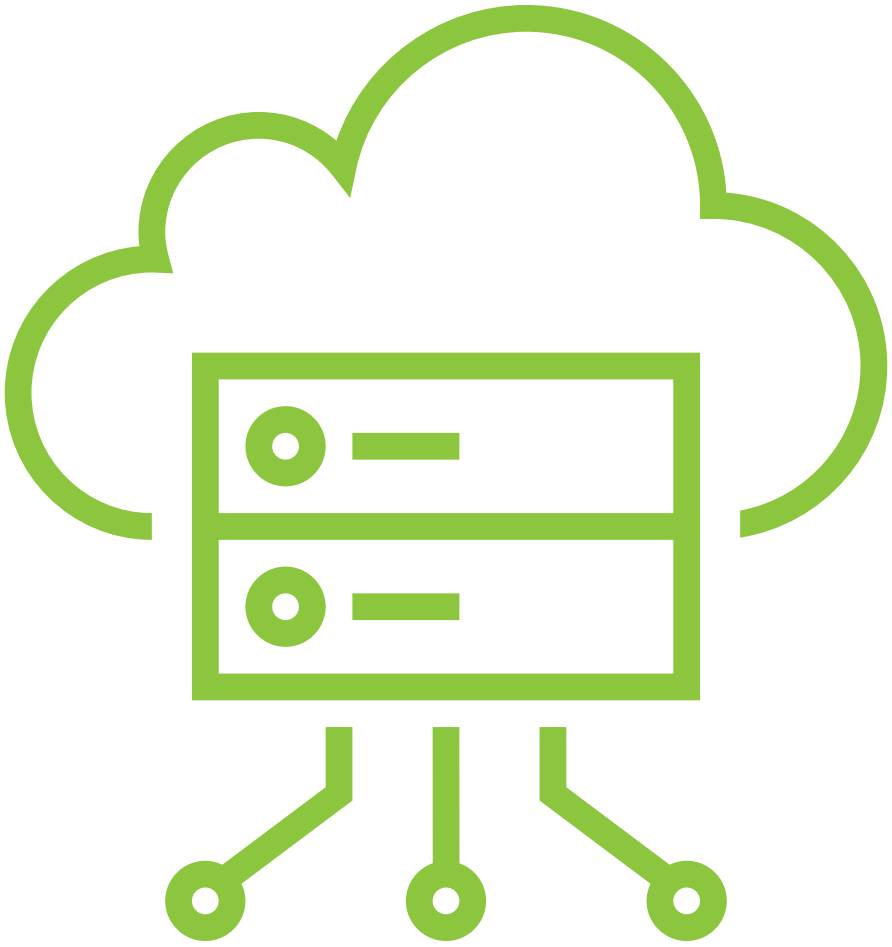
EDGE AI

5 STEPS FOR GETTING STARTED

NVIDIA



Getting Started with Edge AI

In a survey conducted by IBM, 94 percent of executives claim their organizations will implement edge computing in the next five years.

The demand for edge computing is higher than ever—driven by the pandemic and the need for more efficient business processes, as well as key advances in the Internet of Things (IoT), 5G, and AI.

Edge AI, the combination of edge computing and AI, is a critical piece of the software-defined business. From smart hospitals and cities to cashierless shops to self-driving cars, all are powered by AI applications running at the edge.

Transforming your business with intelligence driven by AI at the edge is just that, a transformation, which means it can be complicated. Whether you're starting your first AI project or looking at infrastructure blueprints and expansions, these five steps will help set your edge AI projects up for success.

01. Identify a Use Case

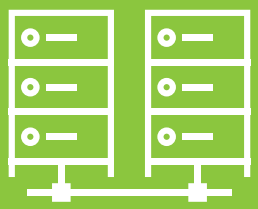
Shrinkage in retail is a $100 billion problem that can be mitigated by edge AI.

Even a 10 percent reduction in shrinkage represents billions in revenue. In the manufacturing space, time and money are wasted when machines go down for repairs. Predictive maintenance uses edge AI to anticipate when machines are about to fail, so they can be serviced or removed from production before they cause an outage. When it comes to getting started with edge AI, it's important to identify the right use case, whether it's to drive operational efficiency, financial growth, or social initiatives.

AiFi uses a combination of AI and computer vision that improves customer experience and brings new intelligence to brick-and-mortar stores.

Stakeholders

When developing an edge AI proof of concept (POC), both internal stakeholders and external partners are critical to its success.



Information Technology and Operational Technology

Security Operations

Business Owners

AI Developers and Data Scientists

Partners are instrumental in helping with best practices, implementation, and more. Engaging these stakeholders early in the process may take time but will ensure alignment and a smoother transition to production.

Value

Look to solve a problem that would be impactful enough to your organization that it justifies the resources and budget needed to prove the value of edge AI. If the use case has low value, the project may lose focus before a full solution can be rolled out. Safety, reduced cost, improved efficiency, and improved customer experiences are all high-value outcomes.

Once the initial AI applications have been rolled out, projects can expand to include more use cases that provide additional value.

Success Criteria

Defining what success looks like and how it will be measured is critical. Edge AI has a few different success factors, the first of which is the deployment and performance of the AI application. Whether it's reducing costs, waste, or downtime or improving efficiency, revenue, or accuracy, ensuring that the application addresses the initial challenge is a priority for the data scientist team or application vendor. Other stakeholders should also define success criteria. IT will want to make sure the AI application can scale to many locations, define tools for remote management, and verify that systems are protected at edge sites.

Time Frame

AI takes time. Depending on where you are in the process, the rollout to hundreds or thousands of locations at the edge can be very time consuming. Look for use cases that address long-term problems that don't have alternate solutions.

Most edge locations will end up running multiple AI applications. Start small with your first use case to create momentum, and then scale your edge AI platform to run the many applications that turn your edge site into an intelligent space.

02. Evaluate Your Data and Application Requirements

By 2025, more than 50 percent of enterprise-managed data will be created and processed outside the data center or cloud.*

Edge AI relies on the right data to get started. With billions of sensors located at the edge, it's generally a data-rich environment. Understanding what data is going to be used will help you plan for which AI models need to be trained, as well as which sensors are required to run AI inference at the edge.

Data requirements—including both the quantity and quality of data—depend on whether you're training or retraining models. Even when buying an AI application, most need to be retrained on labeled data from your environment to produce the accuracy required to show value.

AI can run on data processed at edge locations to help build intelligent spaces that improve efficiency, automation, and experience.

*Source: Gartner®, "Predicts 2022: The Distributed Enterprise Drives Computing to the Edge," G00757917, October 2021.

Data Strategy

AI applications rely on data for both training and inference. Data can come from a variety of locations, including raw data for sensors at the edge or synthetic data that's generated from simulations or algorithms.

Raw Data

Extract, Transform, Load

Trained Data

Internal Expertise

If you're trying to automate a process, use the experts who do the task manually to label data. As an example, for inspection, you can use your quality inspection team to identify defects in images or video. This data is then used to train a model that will have the same knowledge as your best inspectors.

Synthetic Data

Using annotated information that computer simulations or algorithms generate is a technique often used when there's limited training data or when the inference data will vary greatly from the original datasets. For example, retail packaging can change seasonally. Rather than retraining every season, synthetic data can create endless iterations so that the model will detect new packaging without having to be taught exactly what it looks like.

Crowd-Sourced Data

Leveraging your audience to help label large quantities of data has been effective for some companies. Examples include open-source datasets, social media content, or even self-checkout machines that collect information based on customer input. Crowd-sourced data does require quality checks to make sure there aren't outlier inputs that skew the AI model.

KION Group builds intelligent warehouse systems that increase the throughput and efficiency in retail distribution centers.

03. Understand Edge Infrastructure Requirements

AI inference infrastructure must be performant, efficient, and responsive.

Infrastructure can be one of the most important and costly expenses when rolling out an edge AI solution. Unlike data center infrastructure, edge computing has additional considerations around performance, bandwidth, latency, and security. Understanding the existing infrastructure and the requirements for your application are critical to building the right solution. The most common components of edge infrastructure are the sensors, compute systems, network, and management tools.

Edge Infrastructure

Edge environments often have existing infrastructure in place. Most projects start by looking at the current infrastructure to understand what can be repurposed and what needs to be added. Here are some of the infrastructure items to consider for your edge AI platform.

Sensors

Typically sensors are what you'll run your AI inference against. Most organizations today rely on camera streams, but sensors can include chatbots, radar, lidar, temperature sensors, and more. Make sure these sensors are providing the right type and quality of data that you need to run your AI application.

Compute Systems

When sizing compute systems, consider the performance of the application and the limitations at the edge location. Generally, the edge environment has some factors that cannot be changed, including space, power constraints, and heat. Start with these and then consider the performance requirements of your application.

Network

The main considerations for networking is how fast of a response you need for the use case to be viable and how much data there is and whether it needs to be transported across the network in real time. Many edge AI use cases require milliseconds or even sub-millisecond latency. Others generate too much data and saturate a network's bandwidth capacity. The benefit of edge computing is that processing is moved closer to where data is created, and as a result, latency and network bandwidth requirements are reduced compared to processing in the cloud. Wi-Fi is an option when wireless is required but can suffer from congestion that can impact latency. 5G is an option when guaranteed performance is required for wireless environments.

Management Tools

Edge computing presents unique challenges around management of environments. They're generally highly distributed, deployed in remote locations without trained IT staff, and lack the physical security expected from a data center. This means organizations need to consider solutions that solve the needs of edge AI, namely scalability, performance, remote management, resilience, and security. Some organizations will choose to build solutions internally based on tools they're familiar with. Often, organizations don't have the time to build custom management solutions and therefore can look to employ turnkey solutions that are purpose-built for edge management.

04. Roll out Your Edge AI Solution

Edge AI POCs can take anywhere from 3-12 months.

When it comes to rolling out an edge AI application, testing AI applications at the edge is critical for ensuring success. This POC-to-production rollout is done in waves to mitigate any risk from scaling too quickly. To ensure a smooth transition from POC to production, it's important to take into account what the end solution will look like.

Edge AI solutions must scale to support hundreds or even thousands of locations.

Design for Scale

Edge AI POCs are generally limited to one or a handful of locations, but if successful, they need to scale to hundreds or even thousands of locations. While remote management and automated provisioning may not be critical during the POC, using production tools to manage these environments, or at least considering how it will be done, will make scaling a solution much easier.

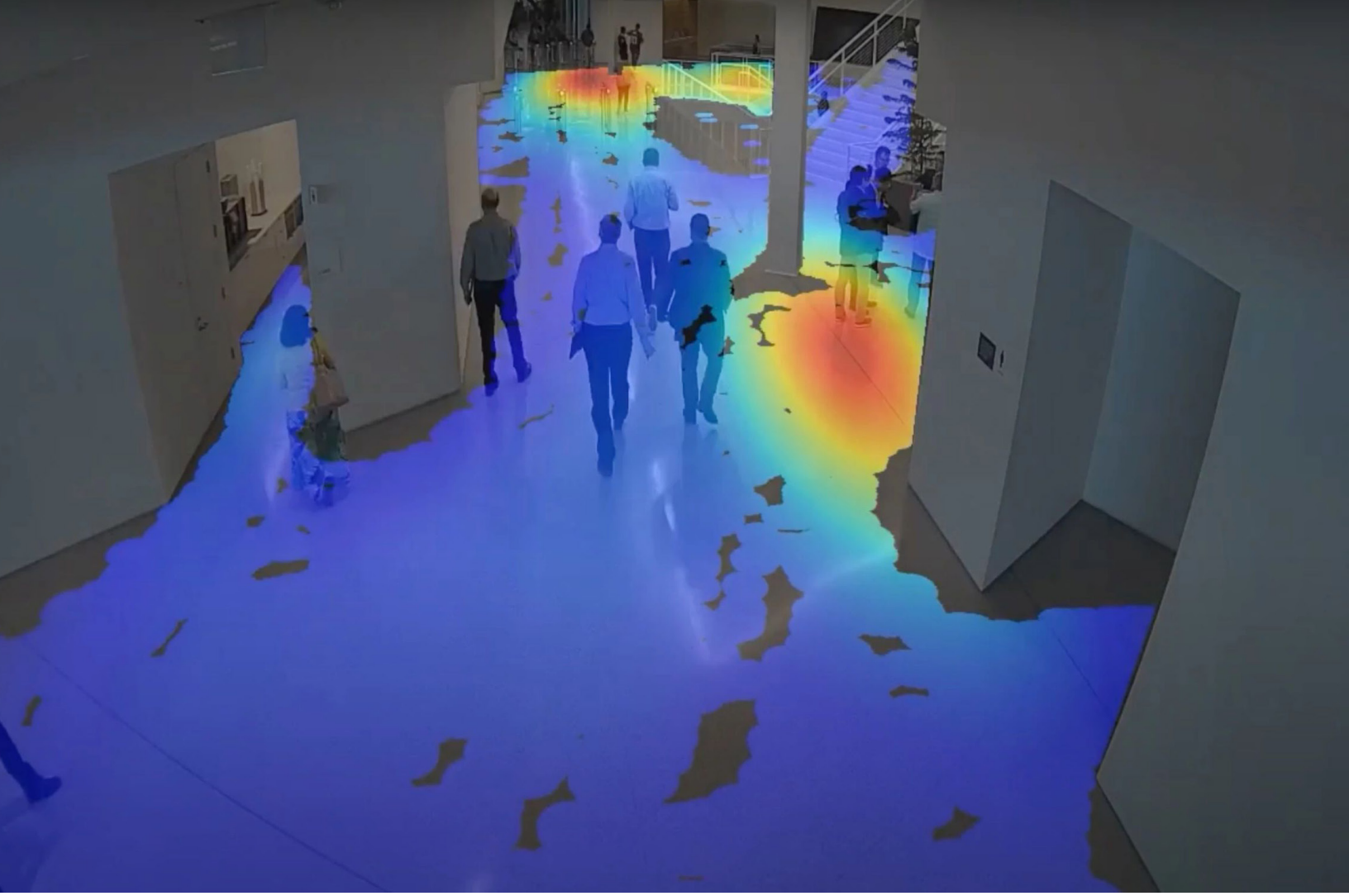
Constrain Scope

AI applications improve over time. Different use cases will have different accuracy requirements that can be defined in the success criteria. But if these goalposts begin to move or new features are requested, it can drag out a POC. Focus on providing value and expect to see improvements as solutions are rolled out at scale.

Prepare for Change

Edge AI has many variables, which means even the best-laid plans will change. Ensure the rollout is flexible without compromising the defined success criteria. Each location may have a unique set of characteristics, but the underlying edge AI platform should help accelerate the deployment process.

Kinetic Vision AI is used for inspection in factories to drive automation and efficiency.



05. Celebrate Your Success

Spark more green lights from stakeholders and guide future initiatives.

Edge AI is a transformational technology that helps businesses improve experiences, speed, and operational efficiency. Many organizations will have multiple edge use cases that they want to roll out, which is why celebrating success is so important. Companies that highlight success are more likely to drive interest, support, and funding for future edge AI projects.

In addition to sharing the results, best practices can be documented and shared across the organization to reduce the cycle time from project definition to production. Businesses that can improve their agility, flexibility, and intelligence with edge AI have a distinct advantage over their competitors.

Edge locations can run multiple AI applications that work together to create smart spaces.

NVIDIA is a Leader in Edge AI

Get deep technology and industry expertise with best-in class solutions.

The ability to glean faster insights can mean saving time, costs, and even lives. That's why modern enterprises are looking to tap into the data generated from the billions of IoT sensors found in retail stores, on city streets, in hospitals, and more to create smart spaces.

As a leader in AI, NVIDIA has worked with customers and partners to create edge computing solutions that deliver powerful, distributed compute, secure remote management, and compatibility with industry-leading technologies.

Compute Systems for the Edge

Optimized for the performance and security needed for edge computing, NVIDIA-Certified Systems™ simplify deployments with tested configurations available from all OEMs.

Management Hub for Edge AI

Purpose-built for AI, NVIDIA Fleet Command™ is a turnkey solution for AI lifecycle management. It removes the complexity of building and maintaining an edge software platform by offering streamlined deployments, over-the-air updates, and detailed monitoring capabilities. Layered security protocols protect intellectual property and application insights from cloud to edge.

The Foundation of Application Development

With NVIDIA application frameworks, companies can access a rich set of developer tools and a partner ecosystem for key edge use cases like vision AI, natural language processing, robotics, and more. With cutting-edge technology and an extensive developer ecosystem, you can quickly create, deploy, and scale AI and IoT applications—from the edge to the cloud.

Proofs of Concept

NVIDIA LaunchPad gives organizations immediate, short-term access to NVIDIA AI running on accelerated compute to speed up the development and deployment of modern, data-driven applications. Quickly test and prototype across your entire edge AI workflow on the same end-to-end stack that can be deployed at the edge.

Learn more about NVIDIA edge computing solutions at:

www.nvidia.com/edge

NVIDIA