Palantir

# Palantir Edge AI

palantir.com



Palantir

# A Generational Shift in AI

Palantir Edge AI
Technical Whitepaper — 2022

**AI AT THE EDGE**



**KEY BENEFITS OF PALANTIR EDGE AI**

**Speed**
Accelerate operations with autonomous decision-making and by connecting devices across environments

**Scale**
Manage thousands of models at scale — federate development, training, and evaluation to internal teams and third-party vendors

**Security**
Lower risk and reduce single points of failure by distributing analysis and decision-making to the edge

**For Sensor Manufacturers**
Embed AI/ML alongside your sensors, enabling your customers to train, manage, and deploy models across the fleet of sensors

**PALANTIR EDGE AI OVERVIEW**

The world is seeing explosive growth in the amount of data being collected from IoT sensors at the edge — from drones to wind turbines to manufacturing robots. With billions of new sensors and zettabytes of additional data coming online in the near future, the edge presents incredible opportunities as well as challenges.

Historically, our customers have focused on acquiring the ability to integrate this data into a centralized platform that allows data-driven decision-making. Today, however, this strategy is becoming less optimal. Many of the most valuable possible actions — whether in combat or in manufacturing — are time dependent. Shuttling data back to the cloud for processing and analysis means a high-priority target may have changed location — or an operational failure may have occurred. The competitive advantage will accrue to those organizations who can make decisions on edge devices in seconds and in potentially disconnected environments.

Palantir Edge AI is Palantir's AI orchestration and sensor fusion engine that enables autonomous decision-making across edge devices and environments. Designed for situations where time and efficiency matter, it operates in low-bandwidth, low-power conditions — including on drones, aircraft, ships, robots, buildings, and satellites.

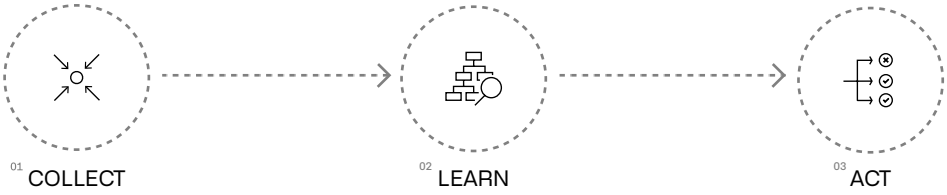
This paper explores how organizations typically achieve the goal of autonomous decisions at the edge, as well as the key technological pillars of Palantir Edge AI.

Palantir

# Achieving Edge AI Decision-Making

3 STAGES OF SOPHISTICATION

How do organizations begin to achieve real-time decisions and action at the edge? We have seen a pattern develop across major public and private institutions, where they progress through three stages of edge AI sophistication:




**01**

Collect
from the Edge →

Faced with massive-scale data from sensors, most organizations begin by collecting and attempting to organize it in a centralized location so that models and logic can be applied.

**02**

Learn
from the Edge →

Once data is collected in a central data and modeling environment, the organization can begin to learn. Teams train models, test hypotheses, and evaluate results. The environment should allow simulation and what-if hypothesis testing to promote confidence in models. Additionally, the platform should be designed for rapid experimentation, allowing you to easily manipulate data, build models, and iterate.

**03**

Act
at the Edge →

Once equipped with high-confidence models, the organization can transport them into the field. Operating effectively at the edge requires a range of technologies, including sensor fusion, model orchestration, and autonomous capabilities.

Palantir Edge AI is designed for every stage of this progression. The following pages describe the features that enable you to move from collection, to learning, to action — at the edge.

Palantir

# Palantir Edge AI Technology Pillars

TECHNOLOGY OVERVIEW

Palantir Edge AI is Palantir's AI orchestration and sensor fusion engine that embeds AI on disconnected, remote endpoints. The platform enables autonomous decision-making for models consuming real-time sensor data, including video, images, radar, acoustic, radio data, and more.

Extremely lightweight and power-efficient to deploy, Palantir Edge AI minimizes the data that needs to be stored and transmitted — enabling low-latency, real-time decision-making at the device level if necessary. It runs on cloud infrastructure, on-premise GPU servers, or Size, Weight, and Power (SWaP) optimized hardware.

Palantir Edge AI is built on the following technological pillars:

01

## Open model management infrastructure →

While most AI companies tend to focus on providing a single model out of the box, Palantir provides organizations with an end-to-end AI/ML solution. Palantir integrates data of any type or complexity — like high-volume sensor and streaming video data — then performs critical tasks around data quality, provenance, and cleaning to render a usable data asset for models. Complete AI/ML infrastructure offers the ability to independently version, test, release, and deploy models. Users can plug in models from any vendor and compare their performance side-by-side — avoiding the vendor lock-in that often comes with AI solutions.

02

## Dynamic model orchestration with Micro Models →

The biggest challenge in AI/ML today is defining valuable, solvable problems and then continuously deploying models against them. Palantir Edge AI addresses this challenge with a new technology called Micro Models. Micro Models are modular, operation-specific models designed around a measurable objective. They can be homegrown, open-source, or third-party algorithms.

Micro Models are environment agnostic and are able to run in isolation or consume outputs from upstream models. By allowing you to separate the logical pieces of your AI pipeline in this way, Palantir Edge AI helps organizations federate model development, evaluation, and training to both internal teams and external vendors — enhancing productivity, quality, iteration speed, and release cadence.

Palantir

# Palantir Edge AI Technology Pillars

**03**

## Adaptive runtime configuration →

To provide flexibility at the edge, models can be run as processors in a series or in parallel. Running them in a series allows models to utilize information from upstream models. Running them in parallel helps to reduce latency and more efficiently utilize hardware. Critically, users can hot swap models in real time without breaking the flow of sensor data through the system. This also means that if a model crashes, it does not impact downstream users who rely on that sensor output.

Palantir Edge AI containerizes models and their dependencies, allowing organizations to package necessary libraries and drivers directly with models rather than rely on coordination with Palantir or third-party models in the system.

**04**

## Lightweight, modular interface to conduct critical compute processes in remote or resource-constrained environments →

Palantir Edge AI runs on an organization's specialized compute hardware or other low-SWaP form-factors. Deploying at the point of use enables optimum quality AI detections derived from the highest quality sensor, IoT, and video inputs.

The platform transmits information away from the device depending on resource constraints. In less constrained situations, it can transmit all raw inputs and enriched metadata from models. In more constrained situations, it can be configured to transmit a "metadata-only" stream (e.g., any detection data, along with positional lat/long/elevation data), which uses less available bandwidth. In recent operational testing, Palantir Edge AI was able to reduce bitrate 20x, creating less pressure on the downlink network system. Running models on a minute form factor and calibrating data transmission makes autonomous sensor work possible.

**05**

## Tooling to streamline and automate complex data engineering tasks →

Palantir Edge AI provides a configuration interface that allows all users — including non-technical ones — to configure highly-complex AI pipelines on real-time sensors. The interface assists with a wide variety of use cases, including video stabilization, object detection, geo-registration, and movement detection and prediction.

**06**

## Live portable retraining →

Algorithms in Palantir Edge AI can be updated and deployed on live feeds with little to no downtime, tightening the feedback loop in model retraining. Teams can optimize models on factors including quality of output, speed, and bandwidth.

**07**

## Continuous integration / continuous delivery →

Decisions and contextual metadata from edge-deployed models flow back into the core Palantir AI/ML infrastructure, accelerating retraining. The result is a virtuous cycle, offering continuous integration / continuous delivery (CI/CD) of models all the way to the edge.

Palantir

# Palantir Edge AI Technology Pillars

**UNMANNED AERIAL VEHICLE / COMPUTER VISION EXAMPLE**

**FIGURE 01**

All processors plug into Palantir Edge AI over a gRPC API using protobuf.

Multiple AI models run in parallel and outputs from each are combined into a single stream by an Ensemble.

A Tracker module adds past context and predicted future paths for detected moving objects.

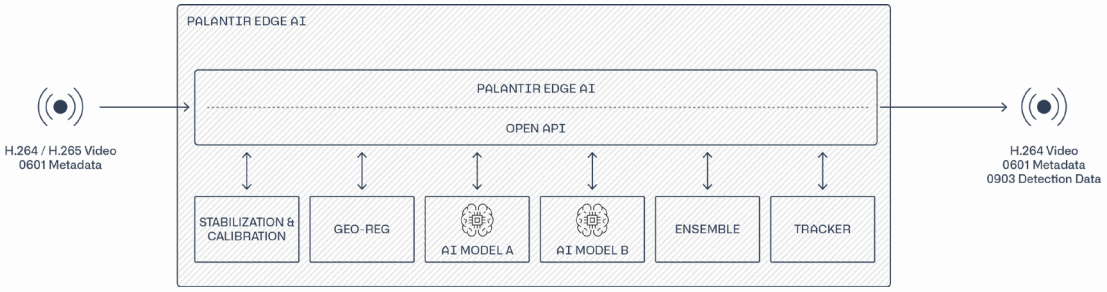
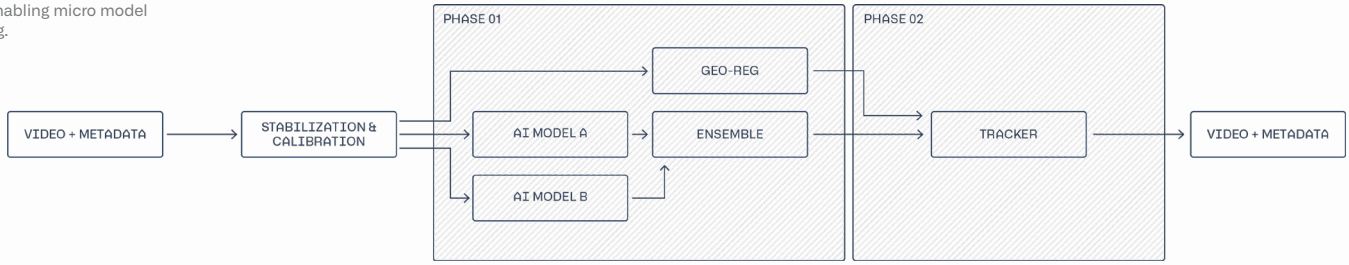



**FIGURE 02**

The order and latency of each processor is fully configurable, enabling micro model chaining.




**08**

## Modular & interoperable architecture →

Palantir Edge AI is designed based off of a Modular Open Systems Architecture (MOSA). This approach enables components of the system to be swapped in and out without any impact to the rest of the platform.

The architecture is implemented through utilizing standard, open-protocol interfaces that facilitate communication between Palantir Edge AI and processors. The solution handles a variety of sensor input formats, such as RTSP/RTP, NITF, GeoTIFF, and MPEG-TS. Finally, it supports outputs in open standard formats — such as Parquet, CoT, MISB 0601/0903 KLV, MPEG-TS, and GeoJSON — which enables data and insights to be sent downstream to other subsystems with little to no integration work required.

Palantir orchestrates all the pieces to the left and right of a given model, meaning teams can build models without knowledge of what other processors are running in the system. Authorized vendors (e.g., data labeling companies) can work directly in the platform or use Palantir's open APIs to plug in their own system.

Palantir

# Palantir Edge AI Technology Pillars

09

## Edge environment simulation →

Palantir integrates real-time sensor data from Palantir Edge AI with an enterprise's existing data foundation, giving teams a valuable corpus of historical data. This historical data can be used to simulate edge environments in Palantir and retrain models accordingly. Palantir Edge AI then extends your simulated environment into the physical world, where models autonomously run on sensors in a lower latency, offline, low bandwidth environment.

10

## Simplified computer vision →

When deploying computer vision models, Palantir Edge AI simplifies video preprocessing, allowing teams to focus on the models. The platform supports a range of cameras and sensors, including various video protocols and codecs. Specifically, Palantir Edge AI handles all video and metadata decoding and produces individual frames of imagery data and metadata for models to correct and analyze in order to augment live feeds.

11

## Sensor fusion →

Palantir Edge AI supports multi-sensor models for customers who need to fuse data across diverse payloads. For example, if a customer uses RF and EO collection, they can field AI models with Palantir, combining both modalities to achieve higher fidelity detections of entities of interest (e.g., military equipment).

These fusion models can be deployed to Palantir Edge AI and run on edge devices, such as spacecraft. Additionally, separate Palantir Edge AI instances can communicate with each other on a mesh network, enabling sensor fusion and teaming across edge equipment.

12

## Data governance →

Strong data governance practices are necessary to model performance and reliability. AI programs require sufficiently large sets of labeled, representative data for model training.

Palantir lets you define granular access control policies at the data integration stage, then propagate those policies intelligently across the system. This enables models to be promoted to collaboration confidently with granular data security and transparent data governance.

Palantir also captures a detailed history of integration, including time of ingestion, source, and revision history. This metadata is used to both track data and model provenance as well as manage compliance with data auditing and retention policies.

Palantir

# Learn More

Unparalleled technology, push to the edge →

Palantir Edge AI powers use cases from manufacturing quality to drug compound screening to autonomous navigation.

See Palantir Edge AI Use Cases ↗

Request a Demo ↗