SUMATOSOFT

AIoT Architecture

How To Choose

There are 3 architecture types:

- edge AI;
- cloud AI;
- hybrid AI.

This choice is about picking the most suitable architecture for your specific business case rather than picking the best one.

Start with your constraints: target latency, bandwidth budget, data sensitivity, offline tolerance, and model size. If you have a technical specialist on site, ask him to answer the following questions:

Target latency

- What is the action the model triggers, and what is the max acceptable end-to-end delay (p95) from signal to action? [ms]
- Is there a hard safety limit (cutoff) for this action? [ms]
- Are humans in the loop (UI/voice)? What delay is acceptable without hurting UX? [ms]
- How much jitter can the system tolerate? [ms]

Bandwidth budget

- What is your available uplink/downlink per site (typical/peak)? [Mbps]
- What's the cost per GB (and any monthly caps/fair-use limits)? [$ / GB, cap GB]
- How much data does one device produce (raw vs. filtered)? [KB/s, events/s]
- What's the monthly cost ceiling per site you can allocate to telemetry/AI? [$]

Data sensitivity

- Which fields are PII/PHI/IP or otherwise sensitive? [list]
- Any data residency constraints (on-prem/in-country/region)? [yes/no + region]
- Required retention period and minimization rules? [days/months]
- Mandatory encryption/audit standards (e.g., GDPR, HIPAA, ISO 27001, FIPS)? [list]

Offline tolerance

- How often and how long do WAN outages occur? [events/month, minutes/event]
- How long can you operate without WAN before service degrades? [minutes/hours/days]
- What is the required degraded mode during outage (full/limited/stop)? [choose]
- How much data must be buffered locally, and for how long? [GB, hours]

Model size / compute

- What is the edge hardware budget per device (RAM/flash/CPU/NPU, power)? [MB/GB, ops/s, W]
- Max model binary size and runtime memory you can afford on-device? [MB]
- Input characteristics: frame/sensor rate and input shape (e.g., 640×480@15fps).
- Expected model update cadence and OTA capability (staged/rollback)? [per week/month, yes/no]

SUMATOSOFT

Thank you for your time!

Any questions? Drop us a line!

Headquarters

One Boston Place, Suite 2602
Boston, MA 02108, United States

Other ways to get in touch

info@sumatosoft.com
sumatosoft.com