VERTICAL-DRIVEN ARCHITECTURES · DISCRETE MANUFACTURING & AUTOMOTIVE · DESIGNED WITH PRAXIS · NOVEMBER 2026

Physical AI System Design for Weld and Assembly Quality in a Vehicle Plant in Pakistan

Discrete Manufacturing & Automotive Watch: One Living Picture of the Operation

Muhammad Umar, Cofounder, CodeNinja

For the leaders accountable for discrete manufacturing & automotive operations, their supervisors, and the data and platform engineers who would build and run it.

Vertical-Driven Architectures is a CodeNinja series of system designs. Every design in the series is driven by a real-world problem and scenario in a single industry, and every one is designed on Praxis, CodeNinja's platform for designing physical AI systems. Operations are described by class, never by name.

ABSTRACT

The Answer Already Exists in the Systems the Operation Runs

A physical AI system design for weld and assembly quality in a vehicle plant in Pakistan starts from the auditor's question: when the parent's auditors ask why a body shop defect rate rose on a given day, the plant rebuilds the answer by hand across the robot logs, the torque records and the inspection sheets.

The design reads 6 source systems through 1 adapter family into one object model of 16 objects, runs 6 services over it and opens 5 surfaces to the people who decide, with 4 models placed where the data is. This design puts triggered cameras and edge units on the CodeNinja on device AI stack at the body shop, trim and underbody station exits, checks every body against the weld point list and the sealer bead with Apache-2.0 detectors the operator owns, and lands every result in a governed plant ontology built from the MES, controller logs, torque tools, end of line rigs, QMS and receiving records. Agents draft, people decide, memory improves the next check without training, and the whole thing, including GLM 5.3 on one 141 GB-class node in the plant's own server room, stays inside the plant network in Pakistan.

The paper covers the operation and the problem its systems cannot answer, the constraints, the stack, the object model, ingestion, inference placement, the models and their licenses, the rollout and its proof, ownership, and finally how Praxis contextualized and reasoned the design.

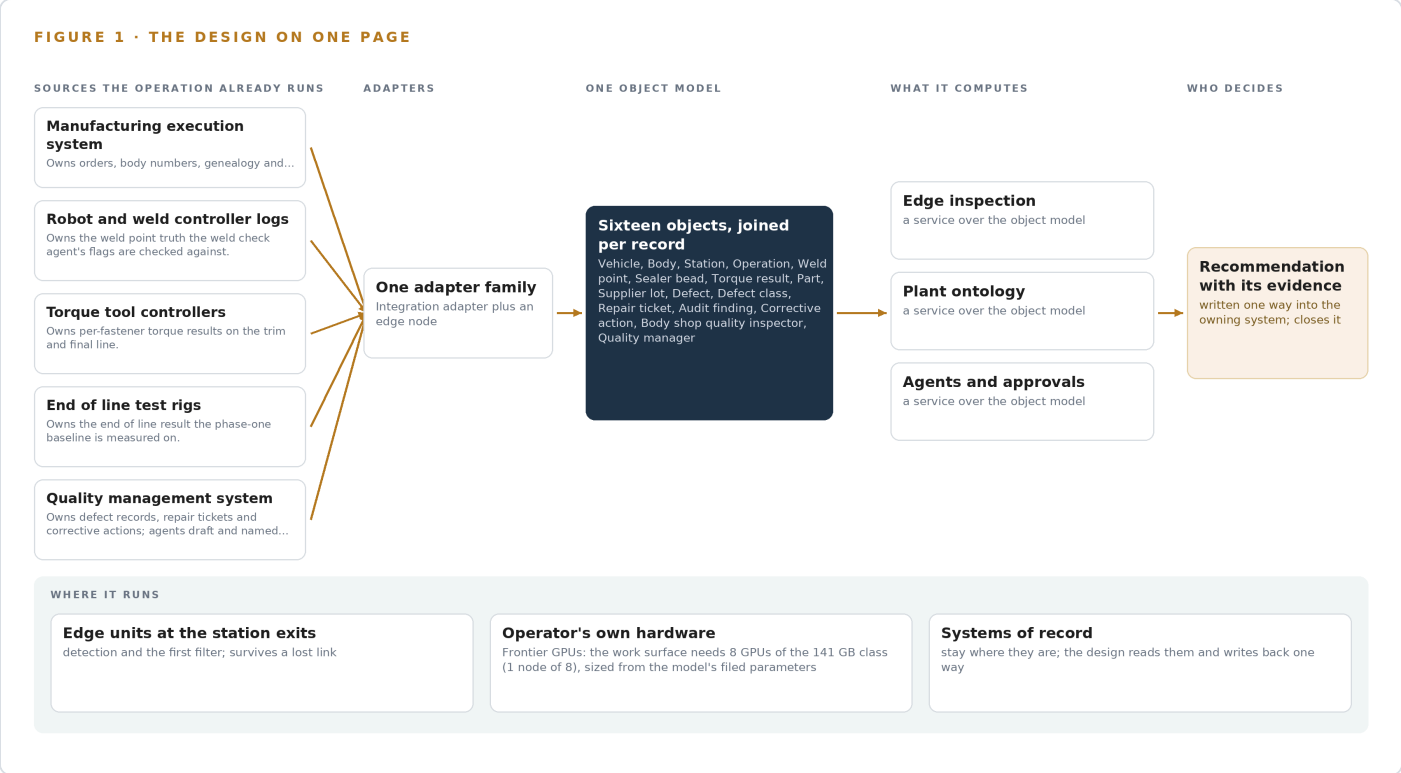


Figure 1. Discrete Manufacturing & Automotive Watch on one page: the sources the operation already runs, one object model, what it computes, and the person who decides.

Contents

Each chapter is tagged for the reader it serves most directly: Executive, Team Lead, FDE, Reference.

	Abstract · The Answer Already Exists in the Systems the Operation Runs	Executive
PART I · THE PROBLEM		
1	A Daily Question Goes Unanswered in the Systems That Hold the Data	Executive
2	Every System Sees One Slice of the Operation	Executive · Team Lead
PART II · THE DESIGN		
3	Four Constraints Shape Everything Downstream	Team Lead
4	One Stack Runs From the Systems of Record to the Surfaces	Team Lead · FDE
5	Sixteen Objects Turn Operational Data Into One Argument	FDE

6	Every Source Enters Through an Adapter, Never Directly	FDE
7	Inference Belongs Where the Data Is	FDE
8	The License Decides What the Operator Can Own	FDE Executive
PART III · THE ROLLOUT		
9	Shadow Mode Comes Before Any Flag Is Trusted	Team Lead Executive
10	The Intelligence Should Stay with the Operation That Produced It	Executive
PART IV · HOW IT WAS DESIGNED		
	Conclusion · The Pattern Repeats Wherever Systems of Record Do Not Meet	Executive
11	How Praxis Contextualized and Reasoned This Design	Team Lead FDE
	Sources	Reference

PART I · CHAPTER 1

A Daily Question Goes Unanswered in the Systems That Hold the Data

The operation holds every fact the question needs and no system answers it.

The abstract promised a physical AI system design that checks every body as it passes the station exits and keeps every byte of the reasoning inside the plant. This chapter starts where the operator's requirement starts, with physical AI at the station named as the ask: the question the plant needs answered every day, the data that question requires, the regimes that govern it, and the operation as it actually runs.

1.1 The Question and the Data It Requires

The question is stated in the operator's own framing: why did the body shop defect rate rise on a given day, and which supplier lot sits behind a defect pattern, answered while the shift is still running rather than weeks later in warranty. Answering it needs six kinds of fact at once: the weld point list and weld currents from the robot and weld controllers, the sealer bead outcome, the per-fastener torque readings, the presence of clips, connectors and fasteners on the trim line, the genealogy that ties a body number to a vehicle and a kit lot, and the lot identity that ties a defect cluster to a supplier. The answer must satisfy three regimes together: the parent company's global quality standards, which the plant passes through equipment validation, production trials and quality audits before commercial production; national data protection rules that govern cameras in a workplace; and a localisation policy that obliges the plant to

track each part's localisation status and local value addition, the end to end supplier traceability the International Automotive Federation frames as a requirement for its members' supply bases (IAF n.d.).

1.2 Defects Found Late Cost More Than Defects Caught at the Station

When a defect escapes the station, the cost jumps from a repair bay to a campaign. A US regulator's recall chronology for incomplete software remedies shows a concern group reconstructing an event history months after vehicles left the plant, the same hand-assembly of evidence across records that this plant performs by hand today whenever auditors ask a question (NHTSA 2024). Supplier-side blindness compounds the escape: a nonprofit study reports that many supplier factories of automobile brands misrepresent facts about work accidents, calling it a lost opportunity for improvement, and a plant that cannot trace a defect cluster to a lot cannot correct what it cannot see (Safeinindia 2024). The join itself is becoming a regulatory subject, because meeting the first Digital Product Passport deadlines depends not only on connecting data across systems but on proving how products were actually made (Qualitymag 2026).

1.3 The Operation as a Scenario

The operator is a vehicle assembly company in Pakistan starting production of new energy vehicles at a new plant, sized for about 25,000 vehicles a year on one shift, assembling kits shipped from the parent company with a rising share of locally made parts. Bodies pass through a body shop with robot and manual spot welding and sealer application, then paint, then a trim and final line where harnesses, interiors, battery packs and underbody parts are fitted and torqued, then end of line test and a final inspection booth: four physical zones in all, of which the body shop, with weld fume, spatter and summer heat, is the hardest environment for electronics. Six named systems of record hold the facts: the manufacturing execution system, the robot and weld controller logs, the torque tool controllers, the end of line test rigs, the quality management system and the supplier parts receiving records. Two named decision roles sit in the loop, the body shop quality inspector who confirms or dismisses flags at the station terminal and the quality manager who approves corrective actions, alongside plant engineers, the parent's auditors and a supplier base whose parts arrive through receiving records. The team is new to the parent's standards, and the experienced engineers who set up the line leave at handover, so the system must carry their knowledge rather than depend on their presence.

PART I · CHAPTER 2

Every System Sees One Slice of the Operation

Each system of record sees its own part of the answer and none sees the whole.

Chapter 1 stated the question, the regimes and the operation. This chapter shows why none of the six systems of record can answer that question alone, by walking through what each one sees.

2.1 Each System Sees One Slice

The manufacturing execution system owns orders, body numbers, genealogy and routing, so it knows which body is where and which vehicle it becomes, but it records nothing about the metal: no weld condition, no bead, no torque. The robot and weld controller logs own the weld point truth, the nominal positions and the weld currents, yet a controller reports what it commanded, not what the body shows, and nothing downstream of the body shop. The torque tool controllers own per-fastener torque readings with tool identity on the trim and final line, but not whether the clip or connector around that fastener is seated. The end of line test rigs own the verdict each vehicle leaves with, but not which station caused a failure. The quality management system owns defect records, repair tickets, corrective actions and audit findings, but they are written by hand, without evidence frames and without joins to lots. The supplier parts receiving records own lot identity, supplier and delivery date, but nothing about what happened to each lot inside the build. Figure 2 sets the six systems side by side, each with its sees and its misses, converging on the question none of them can answer alone.



Figure 2. Six systems, each seeing one part of the answer. The question needs all of them in one place at once.

2.2 What None of Them See Together

The auditor's question, why the body shop defect rate rose on a given day, requires joining weld point flags to body numbers to operations to supplier lots to confirmed defects in one argument, and today that argument is rebuilt by hand across the robot logs, the torque records and the inspection sheets. The same join failure delays the supplier trace: a defect pattern that points at one local lot surfaces weeks

later in warranty, because no system holds the path from a missing clip back to a delivery date. The US national standards body's recommendations on manufacturing data note that trustworthiness depends on traceable provenance across the data lifecycle, which is exactly the property that hand-built joins lack, since each rebuild depends on whoever happens to be on shift (NIST n.d.). The cost in practice is paid three times over: in repair bay minutes for defects found at end of line, in warranty exposure for defects found by the vehicle's owner, and in engineering hours spent reconstructing answers the data already holds.

PART II · CHAPTER 3

Four Constraints Shape Everything Downstream

The goals, the scoping decisions and the exclusions were fixed before any component was chosen.

Chapter 2 showed the join failure that no single system of record can repair. This chapter fixes the four goals that shape every component choice downstream, then the scoping decisions the design bought and the alternatives it set aside.

3.1 Every Body Is Checked, Not Sampled

The first goal replaces sample inspection with a check on every body, triggered once at each station exit, so a missing weld or an unseated clip is caught at the station rather than at end of line or in the field. It is hard in this industry because the body shop is hostile to electronics: weld fume, spatter and summer heat exceed the ambient ratings of the packaged smart camera class, and nuisance signatures such as dust, oil and lighting change produce false flags that make inspectors switch a system off. The design answers by hardening industrial-grade fanless edge units against the environment and by collecting and labelling nuisance images before any threshold is set, holding false flags under 10 per inspector per shift.

3.2 Nothing Leaves the Plant Network

The second goal keeps every frame, every model and every inference inside the plant network in Pakistan, on the operator's own hardware, with no third-party model service in the path. It is hard because the conversational work surface the requirement names runs on a frontier class model of 753 billion parameters at FP8 precision, where one byte per parameter means 753 GB of weights; the 1.2 planning factor for KV cache and activations raises what the GPUs must hold to about 904 GB, which one node of eight GPUs of the 141 GB class, 1,128 GB in total, carries in the plant's own server room. The same constraint forces an export license determination for the accelerator class before it is quoted.

3.3 A Named Person Decides

The third goal holds every decision with a named person: agents draft defect records, repair tickets and corrective actions, and only approved records are written to the quality management system, while no agent ever stops a line or rejects a supplier lot. It is hard because the value of the system is drafting speed, and the parent's audits demand named accountability at the same point, so the approval queue and the confirmation surface at the station terminal must be fast enough that approval is cheaper than working around it.

3.4 Improvement Never Touches the Weights

The fourth goal improves the next check through memory over confirmed outcomes, with no on the fly training or fine tuning anywhere in the design. It is hard because defects are rare and varied, so the temptation is always to retrain, yet changing weights under audit would invalidate the validation evidence the parent's standards rest on; confirmed and dismissed flags become the record each later check reads.

3.5 Scoping Decisions and Their Costs

The scoping decisions in Table 1 were fixed before any component was chosen, and each buys something at a stated cost. The object model is grown from the plant's own six systems of record rather than adapted from published industrial ontology examples, because the join keys are the plant's own body numbers and lot ids, and a borrowed model would have to be re-proven against them anyway.

Table 1 · Scoping Decisions

DECISION	WHAT IT BUYS	WHAT IT COSTS
Check every body at triggered station exits	Missing welds, bead gaps and missing clips caught on every body, not a sample, with evidence frames attached	Cameras, strobed lighting and edge units at every station exit, and a false-flag budget held under 10 per inspector per shift
A read-only projection over the six systems of record	No system of record is replaced or re-hosted, keeping validation risk and integration risk low	The design must prove every join itself, reconciling entity resolution against actual production before any agent runs on it
Evidence frames only, kept on site	Worker acceptance under national data protection rules, and a bounded retention cost	No continuous video, so events between station exits stay unseen by design

3.6 What the Design Chose Against

Table 2 records the rejected alternatives and the reasons, and closes with the scope boundary itself. Beyond that boundary, hardware and infrastructure, third-party license fees levied by the operator's own vendors, historic data migration, estate-wide end user training, round the clock support and regulatory sign-off remain the operator's, with the design supplying evidence rather than authority.

Table 2 · What the Design Chose Against

WHERE	WHAT WAS PICKED	INSTEAD OF, AND WHY
Station detection	Apache-2.0 D-FINE detectors and Anomalib anomaly models on edge units of the CodeNinja on device AI stack	Packaged smart cameras of the industrial vision vendor class, rated to 40 C ambient, below a body shop summer, with weights the operator would not own
Vehicle identity	One triggered capture per body, joined by body number and MES genealogy	Multi-object tracking, unneeded when each body is captured once at a station exit and identity comes from the record
Improvement	Memory over confirmed outcomes, never weights	On-the-fly fine tuning, which would change the model under audit and invalidate the validation evidence
Authority	Agents draft, named people approve	Agent line stop and lot rejection authority, which the operator's own constraint withholds
Scope boundary	Paint shop and end of line cameras held for the phase after the first quarter, gated on phase-one evidence	A single phase covering every zone, which would spend the budget before the first gate proves the body shop and trim checks

PART II · CHAPTER 4

One Stack Runs From the Systems of Record to the Surfaces

Systems of record below, one object model in the middle, services and surfaces above.

Chapter 3 fixed the constraints that shape everything downstream: data stays inside the plant network in Pakistan, the systems of record remain authoritative and unreplaced, no agent stops a line or blocks a supplier lot without a named approval, and improvement happens through memory over confirmed outcomes rather than changes to weights. This chapter turns those constraints into a stack, the layered shape a physical AI system design takes when checks must run at station speed and the argument must assemble itself across six systems of record.

4.1 Three Tiers and One Join

The stack has three tiers. Below sit the six systems of record: the manufacturing execution system, the robot and weld controller logs, the torque tool controllers, the end of line test rigs, the quality management system and the supplier parts receiving records, each owning its domain and keeping it. In the middle sits one object model, sixteen objects with typed links, that projects those records into a single argument a

person can query. Above it run six services and five surfaces, where agents draft findings and named people confirm, approve and decide. Figure 3 shows the layered stack with the count at each layer.

The pattern fits because the plant's failure is a join failure, not a sensing failure. A camera at a trim station can see a missing clip, but the question the quality manager actually asks crosses every system of record at once: which supplier lot, which body numbers, which weld points, which corrective action. The design therefore puts the join in one place, the object model, and refuses to let any application build its own private copy. The ontology is a projection over the manufacturing execution system, never a replacement for it, which is what keeps the parent's audits answerable from the records of record.

FIGURE 3 · THE LAYERED STACK: 5 SURFACES, 6 SERVICES, 16 OBJECT MODEL, 4 INFERENCE, 1 ADAPTERS, 12 SENSING, 6 SOURCES

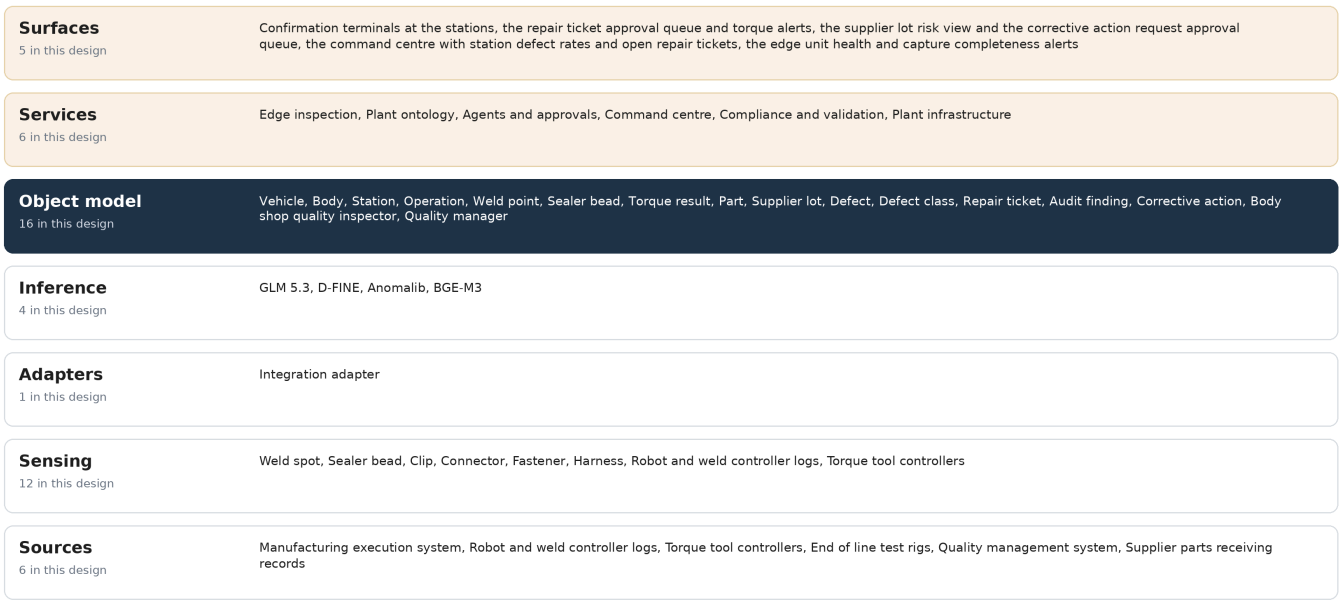


Figure 3. The layered stack: 6 sources, 1 adapter family, 16 objects, 6 services and 5 surfaces.

4.2 The Stack Stage by Stage

Table 3 walks the stack from sources to surfaces, naming what each stage is responsible for and the components that carry it, in the order the data flows: Sources, Sensing, Adapters, Object model, Inference, Services, Surfaces.

Table 3 · The Stack, Stage by Stage

STAGE	WHAT IT IS RESPONSIBLE FOR	HOW
Sources	Own the record of what happened, read-only	Six operator systems: the manufacturing execution system, robot and weld controller logs, torque tool controllers, end of line test rigs, quality management system and supplier parts receiving records

STAGE	WHAT IT IS RESPONSIBLE FOR	HOW
Sensing	Capture one frame per body at the station exits	5 MP global shutter GigE Vision cameras at five positions with strobed lighting, a body-present trigger from the station PLC, an OCP Time Card grandmaster with linuxptp
Adapters	Own the contract for every source	One integration adapter family: Debezium change data capture from the MES and QMS databases, collectors for controller and rig logs, one-way transfer out of a zone
Object model	Hold the join with lineage	the object model with sixteen objects, TimescaleDB for telemetry, Apache Kafka 4.4 as the event backbone
Inference	Run the checks and the reasoning	D-FINE and Anomalib at INT8 through ONNX Runtime on the edge NPU units; BGE-M3 embeddings on the site server; GLM 5.3 under vLLM on one 141 GB-class node
Services	Compose checks into decisions people own	Edge inspection, plant ontology, agents and approvals, command center, compliance and validation, plant infrastructure
Surfaces	Put the decision in a named person's hands	Confirmation terminals at the stations, the repair ticket and torque alert queue, the supplier lot risk view, the command center views, the edge health and capture completeness alerts

4.3 A Reusable Pattern

The reusable shape is this: triggered one capture per body checks at station exits, identity from the body number rather than tracking, an adapter-only ingestion tier, one object model that holds the join, agents that draft and named people who decide, and memory that improves the next check without touching weights. A team carrying a similar brief would swap the detections for their own domain, weld spot and sealer bead becoming whatever their stations must verify, swap the six sources for their own systems of record, and re-size the three inference tiers for their own evidence volumes. What they should keep is the invariant set: no source is read directly, no agent holds stop or rejection authority, and every improvement lands as a recorded, citable decision rather than a silent change to a model.

Table 7 · Architecture, Layer by Layer

LAYER	COMPONENT	CHOICE	WHY
Event streaming and messaging	by class, no product named	Product	Apache Kafka 4.4
The ontology or graph store	by class, no product named	Product	Hyper Ontology, the governed object model the whole argument runs through

LAYER	COMPONENT	CHOICE	WHY
Time series and telemetry storage	by class, no product named	Product	TimescaleDB (PostgreSQL extension)
Video ingest, recording and management	MediaMTX (for any RTSP overview camera)	Class	Station cameras are triggered GigE Vision machine vision cameras, so acquisition is by class with evidence frame only retention; the shelf holds RTSP ingest for IP cameras only.
Edge model serving	by class, no product named	Product	ONNX Runtime
On site language model serving	by class, no product named	Product	VLLM
Object tracking and trajectory forecasting	by class, no product named	Not needed	Each body is captured once at a triggered station exit; identity comes from the body number and MES genealogy, so no tracker is committed.
Edge orchestration of containers	by class, no product named	Product	K3s
An air gapped artifact and model registry	by class, no product named	Product	Harbor
Labelling and retraining	by class, no product named	Product	CVAT
Time synchronisation across sensors and systems	by class, no product named	Product	Linuxptp (ptp4l, phc2sys)
Observability of the running system	by class, no product named	Product	Prometheus, Grafana and Loki

PART II · CHAPTER 5

Sixteen Objects Turn Operational Data Into One Argument

The object model is the join the systems never made.

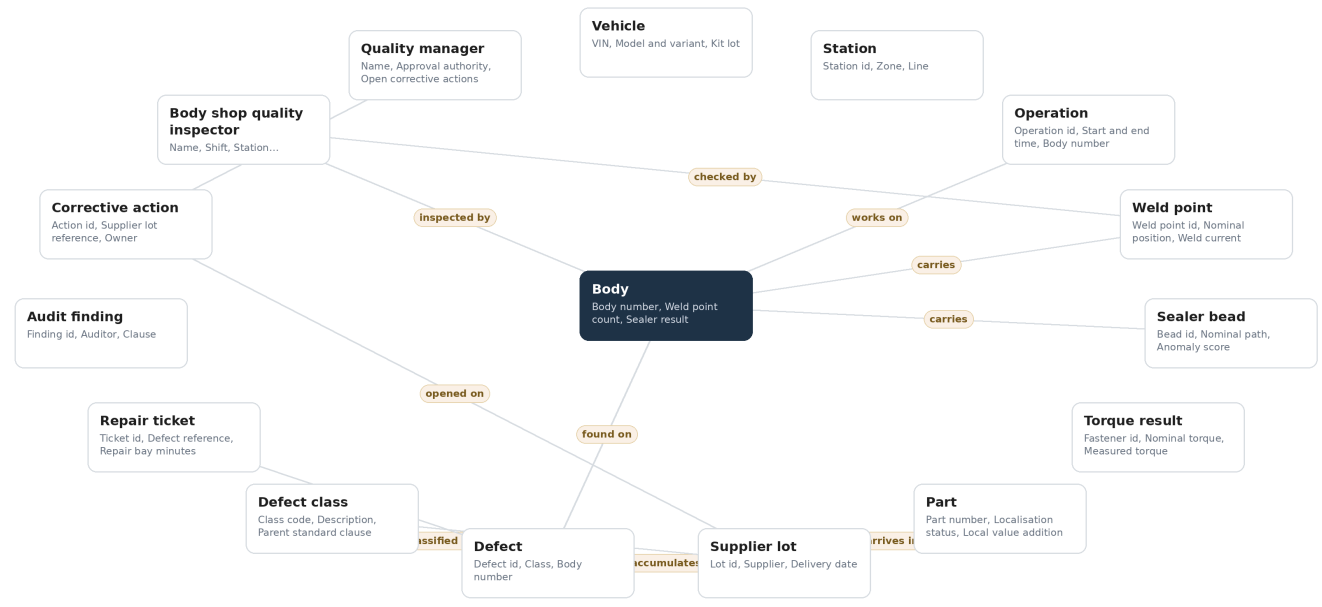
Chapter 4 described the stack that runs from the systems of record to the surfaces. This chapter describes what the middle layer holds: sixteen objects and the typed links between them, the join the plant's systems never made.

5.1 Sixteen Objects and Their Links

Sixteen objects carry the whole argument. Vehicle, Body, Station and Operation come from the manufacturing execution system, which owns orders, body numbers, genealogy and routing. Weld point and Sealer bead come from the robot and weld controller logs; Torque result from the torque tool controllers; Part and Supplier lot from the supplier parts receiving records; and Defect, Defect class, Repair ticket, Audit finding and Corrective action from the quality management system. Two objects are people: the body shop quality inspector, who confirms or dismisses what the cameras flag, and the quality manager, who approves repairs and corrective actions. Figure 4 shows every object and its typed links, with the defect object in focus.

The links are typed, which is what separates this governed plant model from the industrial ontology examples a search usually returns: every object here is anchored in a named system of record and every link is a stated relationship, not a co-occurrence. A defect links to the body it was found on; the body links to its weld points and sealer beads; a part links to the supplier lot it arrived in; the supplier lot links to the defects counted against it and onward to the corrective action opened on it. One query can walk from a flagged weld point to the body, the operation, the shift's inspector, the lots fitted that day and the open corrective actions, in a single traversal with lineage at every hop, where a document store would force a manual rebuild across exports. That manual rebuild is what the plant does today when the parent's auditors ask why a body shop defect rate rose on a given day. The platform prints Table 8 with this chapter, listing each object, the system it is anchored in and its typed links; the recorded form of one object follows in 5.3.

FIGURE 4 · THE OBJECT MODEL: 16 OBJECTS AND 13 TYPED LINKS



The navy card is the object the most links reach. A link label is the property the plan carries, or the verb the design names.

Figure 4. The sixteen objects of the model and the typed links that let a query reach across them.

Table 8 · The Object Model

OBJECT	ANCHORED IN	TYPED LINKS
Vehicle	Manufacturing execution system	no typed link recorded
Body	Manufacturing execution system	carries: Weld point; carries: Sealer bead; inspected by: Body shop quality inspector
Station	Manufacturing execution system	no typed link recorded
Operation	Manufacturing execution system	works on: Body
Weld point	Robot and weld controller logs	checked by: Body shop quality inspector
Sealer bead	Robot and weld controller logs	no typed link recorded
Torque result	Torque tool controllers	no typed link recorded
Part	Supplier parts receiving records	arrives in: Supplier lot
Supplier lot	Supplier parts receiving records	accumulates: Defect; traced to: Defect class
Defect	Quality management system	classified as: Defect class; found on: Body
Defect class	Quality management system	no typed link recorded
Repair ticket	Quality management system	resolves: Defect

5.2 The Human Loop and the Hosting Posture

Every judgment in the model has a named owner. A flagged weld point or a missing clip is a Defect in the state Flagged until the body shop quality inspector moves it to Confirmed or Dismissed at the station terminal; a repair ticket exists in Draft until the quality manager approves it; a corrective action opens only under the quality manager's authority. The single write path back into a system of record is the approved record: agents draft into the ontology, and confirmed findings flow to the quality management system through the adapter, never around it.

The hosting posture follows the boundary constraint. The ontology and its memory run on the operator's own hardware in the plant server room in Pakistan, and nothing leaves the plant network. Identity is Keycloak, on site. There are no external links in the running system: model artifacts arrive through the air-gapped Harbor registry, and any transfer out of a network zone passes a one-way data diode. Data residency, personal-data tagging and a recorded lawful basis per data class follow the draft national data protection rules, which are a planning signal rather than settled law.

5.3 One Object in Its Recorded Form

The vehicle record below shows the shape of all sixteen: an identifier, a label, a kind, the system it is anchored in, its properties and its status vocabulary. The defect is recorded the same way, with the body number, the evidence frame, the person who confirmed it and a three state vocabulary (Flagged, Confirmed, Dismissed) that keeps a machine flag and a human confirmed defect distinct.

```
{ "id": "vehicle", "label": "Vehicle", "kind": "asset", "anchored_in": "Manufacturing
execution system", "properties": [ "VIN", "Model and variant", "Kit lot", "End of line
result" ], "status_vocabulary": [ "In build", "Complete", "Hold", "Released" ], "links": [] }
```

PART II · CHAPTER 6

Every Source Enters Through an Adapter, Never Directly

Nothing reaches the object model except through an adapter that owns its contract.

Chapter 5 described the sixteen objects the middle layer holds. This chapter describes how data reaches them: every source enters through an adapter that owns its contract, and nothing reaches the object model directly.

6.1 Six Sources, One Adapter Family

Six named source systems feed the model, and all six are the operator's own systems of record, read-only, classed as operator provenance. The manufacturing execution system owns orders, body numbers, genealogy and routing. The robot and weld controller logs own the weld point truth the weld check agent's flags are checked against. The torque tool controllers own per-fastener torque readings on the trim and final line. The end of line test rigs own the outcome the phase-one baseline is measured on. The quality management system owns defect records, repair tickets and corrective actions. The supplier parts receiving records own lot identity, supplier and delivery date, which is what turns a defect cluster into a supplier lot trace, the end to end chain that supplier quality frameworks in the sector treat as the baseline obligation (IAF n.d.). Figure 5 sets out the integration map: each named system, its provenance class and its adapter path into the model.

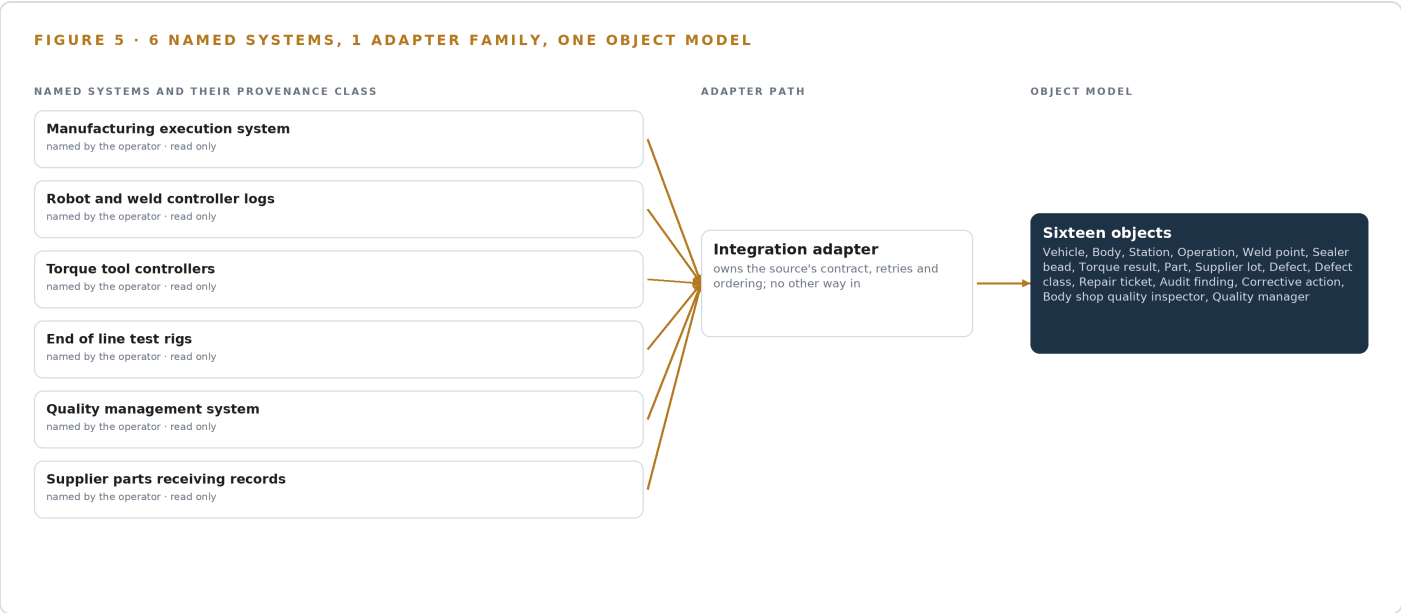


Figure 5. The 6 named systems, the adapter path each one takes, and the object model they all map into.

6.2 What the Adapter Tier Guarantees

One adapter family serves all six sources: the integration adapter. It guarantees a contract per source, so a schema change upstream surfaces as a broken contract rather than silent bad data. It guarantees identity resolution: controller, MES, QMS and receiving records are joined once, with lineage, inside the ontology rather than separately in each application, and the reconciliation is proved against yesterday's actual production before any agent runs on it. It guarantees read direction everywhere except the single approved write path to the quality management system. And it guarantees time discipline, with every event stamped from the PTP grandmaster so a camera frame, a torque reading and a weld log can be placed on one timeline. Change data capture through Debezium reads the MES and QMS databases without touching their transaction paths, and anything crossing a network zone passes the one-way transfer, so the boundary between office and plant networks stays one-directional. Lineage at every hop is also what national standards guidance asks of manufacturing data used for decisions: provenance, timestamping and a record of what was joined (NIST n.d.).

6.3 The Event Backbone

Apache Kafka 4.4 carries events between the adapters, the ontology and the services. Ordering is keyed by body number, so every record for one body arrives in sequence regardless of how the partitions are laid out. Delivery is at least once with idempotent writes in the ontology, which tolerates a replayed event without double-counting a defect. Buffering is on disk at the edge: if the link from a station unit to the backbone drops, capture continues locally and resynchronizes when the link returns, so a network blip never becomes a missed body. Replication runs inside the plant's Kafka cluster, so the loss of a broker does not lose an event. Telemetry and time series land in TimescaleDB, and Prometheus, Grafana and Loki watch the backbone itself, so capture completeness is an alert the command center sees rather than a gap the quality manager discovers at audit time.

Inference Belongs Where the Data Is

Each tier runs what its data and its link can carry.

Physical AI at the station only pays for itself when the model runs where the body is, at the pace the line already keeps. Chapter 6 described how all six source systems enter through the integration adapter and land in the plant ontology. This chapter fixes where each of the four models executes, which is a performance decision, a data-residency decision and an export-control decision at once.

7.1 Three Tiers and the Memory Arithmetic

The design runs three inference tiers, and Figure 6 sets them side by side with what crosses between them. The edge tier sits at the body shop, trim and underbody station exits: 5 MP global shutter GigE Vision cameras with strobed lighting capture each body once, on a body-present trigger from the station PLC, and a fanless industrial edge unit on an octa-core 6 TOPS NPU SoC rated -40 to 85 C runs the check on the device. The D-FINE detectors run there at INT8, where the N variant is about 8 MB at 16 bit and the X variant about 124 MB. The Anomalib models that watch sealer beads are model dependent, typically tens to hundreds of MB at 32 bit, and PatchCore memory banks grow with the number of reference images, so each station sizes its model on its own variant and its own reference set.

One air-cooled 48 GB PCIe inference-class server in the plant server room is the site tier. It serves BGE-M3 at FP16, about 1.1 GB, and re-scores the borderline detections the edge tier flags, so a station-level doubt becomes a full-precision second opinion without any frame leaving the plant. A US ECCN 3A090 license determination for Pakistan is required before this class is quoted, which is why the class carries its own checkpoint in the rollout.

The central tier is GLM 5.3 on one node of eight 141 GB HBM GPUs, and the memory arithmetic decides it. 753B parameters at FP8 is 1 byte per parameter, so 753 GB of weights, and the 1.2 planning factor for KV cache and activations raises the requirement to 904 GB; one node of 8 x 141 GB holds 1,128 GB, leaving 375 GB beside the weights against the 151 GB the factor reserves. At the roughly 756 GB the published weights occupy, the KV cache ceiling is about 280 GB, which bounds concurrent sessions on the conversational surface rather than forcing a second node. The printed architecture table (Layer, Component, Choice, Why) carries the full stack card, and these three tiers are its inference rows.

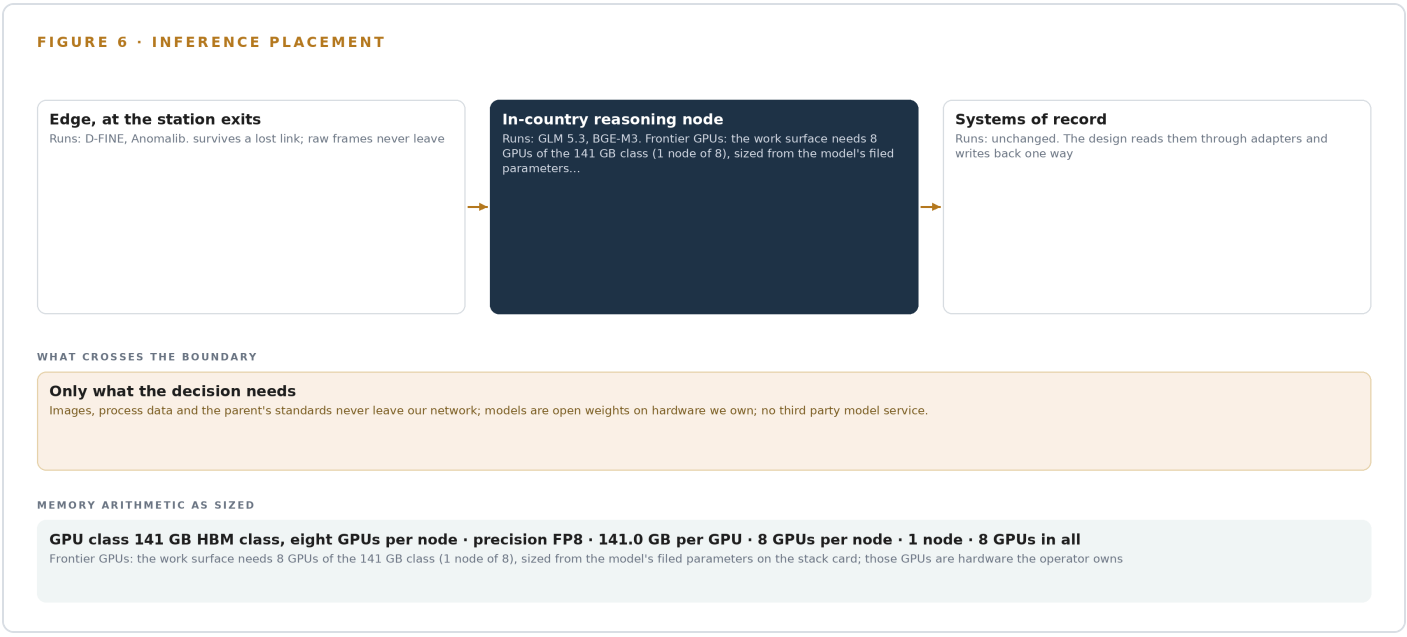


Figure 6. Where each tier runs, what runs there, and the narrow set of outputs that cross the boundary.

7.2 A Latency Budget Set by Takt

The budget is set by takt time, the property the ontology carries on every station object. Each presence check must finish inside the station's takt, so no body waits on a model and no inspector waits on a queue. The edge tier meets this by construction, because the detector runs on the unit beside the camera with no network round trip, and the published 2.12 to 12.89 ms single-image latency on a T4 leaves wide margin at station pace. A seconds-scale budget governs the site tier, for re-scoring and for embedding records into the retrieval index. The central tier works on a human-scale budget: conversational answers with citations are read and judged by named people, so the cost of a longer answer is attention, not throughput. Recall chronologies show how late a defect pattern is usually caught when the check is not at the station (NHTSA 2024); that delay is what this budget exists to remove.

7.3 What Crosses the Boundary, and What Fails

Nothing leaves the plant network in Pakistan. Detections, evidence frames and ontology writes move from the edge units across the plant backbone; model and container updates enter through the on-site Harbor registry fed by a one-way Lidi transfer; no third-party model service is ever called. Every record carries provenance back to its source system, the practice recommended for trustworthy manufacturing data (NIST n.d.). When the link fails, the edge units keep capturing and checking against the last loaded model, buffer locally, and resynchronize once the backbone returns. When power fails, the sealed fanless units ride the wide-input 24 V buck from the plant supply, and the specified universal-input supplies restart clean on the country's 230 V 50 Hz mains. When the update path fails, the one-way gate means an interrupted update leaves the running detector in place while the registry holds the last complete artifact. The number that decides whether any of this is trusted is the false-flag gate, held under 10 per inspector per shift.

PART II · CHAPTER 8

The License Decides What the Operator Can Own

Every model was chosen against its license before its accuracy.

Chapter 7 placed each of the four models on the tier its data and its link could carry. This chapter shows that the placement was the second decision, because every model was chosen against its license first: the license decides whether the operator can hold the weights, fine-tune them and keep them running when a vendor's attention moves on. Figure 7 stacks the four models with their specifications and their jobs.

FIGURE 7 · THE MODEL STACK: 4 MODELS, THEIR PLACEMENT AND THEIR WORK



Figure 7. The four models, their placement, and the work each one does.

8.1 Four Models and Their Terms

GLM 5.3 is the agentic work surface and the ontology's maintainer: it drafts weld checks, traces supplier lots and answers questions conversationally with citations. It is frontier class, 753B parameters as filed, at or above the 300B threshold that defines the class, with about 756 GB of weights at FP8 as published and roughly 1.5 TB at BF16, which means eight or more data center GPUs of the 140 GB class before any KV cache. Its bespoke license permits commercial use, fine-tuning and redistribution, and that is what lets the operator run it on its own node in its own server room; the operator names the model itself and fixes it to its own hardware.

D-FINE is the station detector for weld spots, parts, clips, connectors and fasteners, on the unconditional Apache-2.0 terms that cover code and weights alike. The N variant is about 8 MB at 16 bit and the X variant about 124 MB, with 42.8 to 55.8 COCO AP at 2.12 to 12.89 ms on a single T4 at batch size 1, and it runs INT8 on the edge NPU. The COCO checkpoints are chosen because they avoid the Objects365

terms that ship with other checkpoints, and site nuisance images are collected and labelled before any threshold is set.

Anomalib, version 2.7.0, is the sealer bead watcher on the same Apache-2.0 terms with edge export. Sealer gaps are rare and varied, so an unsupervised model fitted on the plant's own good beads beats any defect catalog; its footprint is model dependent, typically tens to hundreds of MB at 32 bit, and PatchCore memory banks grow with the reference set. The plant's own images replace the non commercial benchmark data the library ships with, which is the license's practical edge.

BGE-M3 is the retrieval model, MIT licensed, served at FP16 on the 48 GB site server: about 2.27 GB at FP32, about 1.1 GB at FP16 and about 0.6 GB at 8 bit, with an 8,192 token window and dense, sparse and multi-vector retrieval in one pass, so part numbers and error codes stay retrievable and every conversational answer can cite the record behind it.

8.2 The Register and the Ground It Runs On

Table 4 gathers the choices with the hardware classes and sizing rules, the sensing, the patterns and the ground, so that one row shows why each piece sits where it does. Each row is also a license record: what the operator holds, on what terms, and what it may keep doing when a term ends. The register matters most at handover, when the engineers who set up the line leave and the plant's own team runs the system from written ground, because supplier quality frameworks of the kind the international federation publishes make traceability to source a member obligation rather than a courtesy (IAF n.d.). The ontology underneath is not one of the catalogued industrial ontology examples a team adopts wholesale; it is a governed model built from the operator's own six systems of record, and unifying that data is now also a regulatory clock, because meeting digital product passport deadlines depends on proving how products were actually made (Qualitymag 2026).

Table 4 · Model and Equipment Register

THE CHOICE	WHAT WAS PICKED	WHY HERE
Station detection	D-FINE at INT8 on the 6 TOPS edge NPU	Apache-2.0 code and weights owned outright; COCO checkpoints avoid Objects365 terms
Sealer and unseen-defect watch	Anomalib 2.7.0 on the edge units	Apache-2.0 with edge export; the plant's own good beads replace the non commercial benchmark data the library ships with
Retrieval and citations	BGE-M3 at FP16 on the 48 GB site server	MIT license; dense, sparse and multi-vector retrieval in one pass, 8,192 token window

THE CHOICE	WHAT WAS PICKED	WHY HERE
Agentic work surface and ontology maintenance	GLM 5.3 at FP8 on one node of eight 141 GB GPUs	Bespoke license allowing commercial use, fine-tuning and redistribution; the operator names it and fixes it to its own hardware
Hardware classes and sizing rules	a frontier inference cluster class (GPU nodes for 300B to 1T parameter open weight models); GLM 5.3 node count by GPU class; plant server room GPU node ASHRAE class	One node of 1,128 GB holds the 904 GB the 1.2 planning factor requires, and the ASHRAE class fixes the server room envelope
Sensing	Triggered 5 MP global shutter GigE Vision cameras at five station exits; controller, torque and end of line feeds; the body-present PLC trigger	One capture per body; identity comes from the body number and MES genealogy, so no tracker is committed
Patterns	Triggered one capture per body checks; nuisance images collected before thresholds; agents draft and named people approve	Cameras check bodies and vehicles, never worker performance; the false-flag gate stays under 10 per inspector per shift
The ground it runs on	The plant network in Pakistan, the operator's own hardware, the on-site Harbor registry and the one-way Lidi transfer	Nothing leaves the boundary, no third-party model service is called, and updates enter one way

PART III · CHAPTER 9

Shadow Mode Comes Before Any Flag Is Trusted

The rollout earns each phase at a gate the operator can read.

Chapter 8 closed the model register with licenses that let the operator hold every weight it runs. This chapter sets out how the build earns its place on the line: one gated phase, measured against the plant's own production, with no flag trusted until it has run in shadow beside the people who would act on it.

9.1 One Phase, Fifteen Items, Six Workstreams

The rollout is a single phase, Phase One, carrying 15 items across six workstreams: edge inspection, plant ontology, agents and approvals, command center, compliance and validation, and plant infrastructure. The items run from the station camera and edge unit layout, through the weld and sealer check pipeline, to trim and underbody presence checking, and the phase's exit gate carries that last item's name: presence checking at the trim and underbody stations must run on every body and vehicle that passes before the phase closes. The phase covers the operator's 18 requirements, and coverage against

each requirement is read at the gate, so a requirement that is only partly met holds the phase open. Figure 8 shows the phase, its item count, its workstreams and its gate on one page.

Shadow mode governs the whole phase. Every check runs alongside the manual inspection it supports, flags are drafted for a named inspector to confirm or dismiss, and no agent stops a line or rejects a supplier lot, because that authority was scoped out in Chapter 3 and stays out.

FIGURE 8 · ONE PHASE, 15 SCOPE ITEMS, 18 REQUIREMENTS



Figure 8. The one phase and its gate, and coverage of the 18 requirements across it.

9.2 What the Rollout Measures

The phase measures itself on the plant's own numbers, not on laboratory ones. The end of line test rigs give the baseline: the defect classes the phase must catch earlier, counted per body before any camera is switched on. Against that baseline the phase tracks the false-flag count per inspector per shift, held under 10, because a budget above that is where inspectors stop trusting flags and switch the system off. It tracks capture completeness, meaning every body that passes a station has a frame, a verdict and a body number joined to its genealogy. It tracks the confirmation rate of flagged defects, the number of defect clusters that resolve to a supplier lot, and the time it takes to answer the parent's auditors when they ask why a body shop rate rose on a given day. Measurement data carries lineage and provenance so it can be defended under audit, following the recommendations on traceability and trustworthiness of manufacturing data published by the US national standards body (NIST n.d.).

9.3 Failure Modes

Table 5 lists what the design expects to fail and what it does about each, and every row is a risk the phase plans for rather than a surprise it absorbs.

Table 5 · Failure Modes

WHAT FAILS	WHAT THE DESIGN DOES
Nuisance signatures, such as dust, spatter, oil and lighting, drive false alarms and inspectors	Nuisance images are deliberately collected and labelled before thresholds are set, and the false-flag budget holds

WHAT FAILS	WHAT THE DESIGN DOES
switch the system off	under 10 per inspector per shift
Worker acceptance of cameras at manual stations stalls under the national data protection rules	Masking, signage and retention limits are built in phase one, with worker consultation before the first camera is switched on
The US export license determination for the 141 GB HBM frontier node delays or blocks plant hardware	The license checkpoint runs at phase-one start, and an approved data center or phased precision fallback stays on the table without quietly shrinking the model
Entity resolution across controller, MES, QMS and receiving records joins the wrong things	Identities are resolved once in the ontology with lineage, and the reconciliation is proved against yesterday's actual production before agents run on it
The parent's validation calendar slips and station checks are not accepted before auditors rely on them	Validation acceptance is a phase-one gate, and the evidence pack is aligned with the parent's audit terms from the start
Experienced engineers leave at handover and the system loses its operators	Runbooks, the ontology and decision memory capture their knowledge, and the plant's own team performs the integration with their own hands

9.4 Lessons

Shadow the flags before anyone trusts them. A flag that arrives unproven is a complaint, and inspectors treat it as one. Running every check in shadow, with confirmation and dismissal recorded per flag, turns the same hardware into evidence: the inspector who confirms ten flags and dismisses one has taught the system what this plant's bodies look like, and the record of those decisions is what the next check reads.

Collect the nuisance images before the thresholds are set. The case-study record is unambiguous on this: false alarms from ordinary plant grime are what switch inspection systems off, and thresholds tuned on clean images are tuned on a plant that does not exist. The nuisance set is a phase item with its own gate, not a tuning chore left to the end.

Resolve the body number once, in the ontology. Every join in this design rides on the body number and the MES genealogy, so a wrong resolution poisons weld truth, torque truth and defect records at once. Proving the reconciliation against yesterday's actual production, before any agent reads it, is cheaper than finding a wrong join inside an audit answer.

9.5 What Is Still Open

Four questions remain open, and settling each changes something specific. The US export license determination for the frontier node and every accelerator class is the first: settling it fixes the server room

hardware as designed, or forces the fallback path to be chosen openly rather than by shrinking the model in place. The plant network design is still being finalised, and settling it fixes where the IEC 62443 zone boundaries and PTP-capable switch hops sit relative to the station exits. The scale figures behind the layout come from public reporting on plants of this class in the country and are assumptions to confirm, so settling them fixes the camera count and the per-station capture budget against the real takt. The parent's validation calendar is the last: settling it fixes when auditors may rely on the station checks, and therefore when the phase's evidence pack stops being preparation and becomes the answer.

PART III · CHAPTER 10

The Intelligence Should Stay with the Operation That Produced It

What the design builds belongs to the operator.

Chapter 9 showed the phase earning trust at a gate the operator can read. This chapter settles who holds what the work produces once the gate is passed: the model of the plant, the weights, the record of decisions, and the boundary they run inside.

10.1 What the Operator Owns

The object model is the operator's. Its 16 objects are anchored in the operator's own systems of record, with the manufacturing execution system holding body numbers, genealogy and routing, and the ontology standing as a projection over it rather than a replacement. Published industrial ontology examples seldom anchor every object in a live system of record; this one does, and because the anchors are the operator's systems, the operator can extend the model as new parts are localised without asking anyone's permission. End to end traceability of this kind is what the International Automotive Federation's supplier requirements frame expects members to align against (IAF n.d.).

The weights are the operator's. The detectors carry Apache-2.0 terms and the embedding model an MIT license, both of which permit commercial use and modification outright, and the work surface model's bespoke license permits commercial use, fine tuning and redistribution. Every fine-tune, and every model hardened on the plant's own nuisance images, belongs to the operator and stays inside its boundary.

The decision record is the operator's. Every flag, confirmation, dismissal, approval and corrective action carries a named person, and that record is what the next check reads before it acts: memory improves the checks without any weight being trained, which was a scoping decision in Chapter 3 and holds here as ownership. Nothing about the operator's quality judgement lives anywhere it cannot be exported, audited or carried forward.

The boundary is the operator's. Everything runs inside the plant network in Pakistan, on hardware the operator owns, with no third-party model service in the path and evidence frames retained under the stated limits. Where the plant's data goes, and who may move it, is a decision the operator holds.

10.2 The Offer Behind the Design

CodeNinja built this design on Praxis, its platform for designing physical AI systems, and the design maps directly onto what CodeNinja offers: Adaptive Operations is the sensing, detection and forecasting of physical behavior at the station exits; Decision Systems is the ranking, attribution and recommendation a named person decides on, from weld flags to supplier lot traces; Hyper Ontology is the governed object model the whole argument runs through; Hyper Pragma is the work surface where the operator's own people build and run their agents; and Sovereign Infrastructure is the shape of the whole thing, own hardware, open-weight licenses and in-country operation, with the frontier node in the plant's own server room.

PART IV · CONCLUSION

The Pattern Repeats Wherever Systems of Record Do Not Meet

Seen whole, the design is one object model over the systems a vehicle assembly company in Pakistan already runs, with inference placed where the data is and every decision left to a named person.

Running the same shape elsewhere takes the same four things: the named systems, an object model anchored in them, a first gate that can stop the work cheaply, and the license posture that lets the operator own what it learns.

PART IV · CHAPTER 11

How Praxis Contextualized and Reasoned This Design

Every choice above traces back to something Praxis recorded reading.

Chapter 10 settled ownership with the operator. This chapter opens the record of how the design was reasoned, so any choice above can be traced to the reading that justified it. Every design in this series is produced on Praxis, the platform for designing physical AI systems, and this chapter exists so a reader can audit the reasoning rather than take it on trust: Figure 9 shows the ask, the family and industry assigned, what it read, the eight lenses, the patterns and the equipment classes the reasoning landed on.

11.1 Contextualizing the Ask

The ask arrived as a requirement written in the operator's voice: physical AI at the station, with edge units and cameras at the body shop, trim and underbody exits, checking every body against the weld point list and the sealer bead, and landing every result in a governed plant ontology built from the systems of record, all inside the plant network in Pakistan. Praxis assigned it to the physical AI family in the discrete manufacturing and automotive industry. What Praxis read was the corpus shelf of each lens: the case study, tooling and recency, hardware and equipment, rules and regulations, approach and history shelves, beside the first principles domain brief.

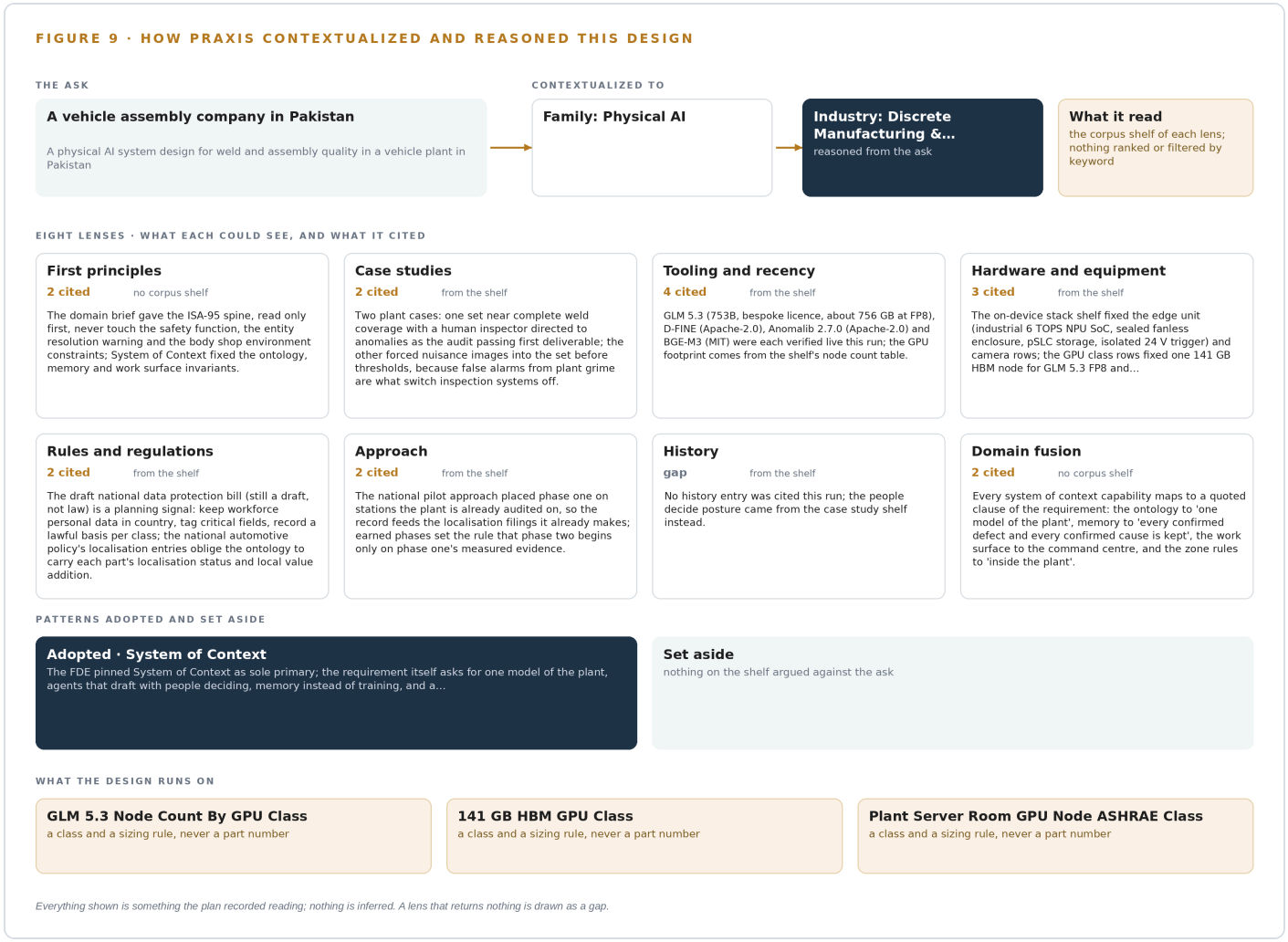


Figure 9. From the ask to the design: the family and industry Praxis assigned, the eight lenses and what each cited, the patterns adopted and set aside, and the equipment the design lands on.

11.2 The Lenses

Table 6 records each lens, how many records it could see, how many it cited, and what it contributed to the design. A lens that returned nothing is shown as a gap rather than smoothed over.

Table 6 · The Lenses and What They Contributed

LENS	COULD SEE	CITED	WHAT IT CONTRIBUTED
First principles	no corpus shelf	2	The domain brief gave the ISA-95 spine, read only first, never touch the safety function, the entity resolution warning and the body shop environment constraints; System of Context fixed the ontology, memory and work surface invariants.
Case studies	the case study shelf	2	Two plant cases: one set near complete weld coverage with a human inspector directed to anomalies as the audit passing first deliverable; the other forced nuisance images into the set before thresholds, because false alarms from plant grime are what switch inspection systems off.
Tooling and recency	the tooling and recency shelf	4	GLM 5.3 (753B, bespoke licence, about 756 GB at FP8), D-FINE (Apache-2.0), Anomalib 2.7.0 (Apache-2.0) and BGE-M3 (MIT) were each verified live this run; the GPU footprint comes from the shelf's node count table.
Hardware and equipment	the hardware and equipment shelf	3	The on device stack shelf fixed the edge unit (industrial 6 TOPS NPU SoC, sealed fanless enclosure, pSLC storage, isolated 24 V trigger) and the camera rows; the GPU class rows fixed one 141 GB HBM node for GLM 5.3 at FP8 and the ASHRAE class for the plant server room.
Rules and regulations	the rules and regulations shelf	2	The draft national data protection bill (still a draft, not law) is a planning signal: keep workforce personal data in country, tag critical fields, record a lawful basis per class; the national automotive policy's localisation entries oblige the ontology to carry each part's localisation status and local value addition.
Approach	the approach shelf	2	The national pilot approach placed phase one on stations the plant is already audited on, so the record feeds the localisation filings it already makes; earned phases set the rule that phase two begins only on phase one's measured evidence.
History	the history shelf	0	No history entry was cited this run; the people decide posture came from the case study shelf instead.
Domain fusion	no corpus shelf	2	Every system of context capability maps to a quoted clause of the requirement: the ontology to 'one model of the plant', memory to 'every confirmed defect and every confirmed cause is kept', the work surface to the command centre, and the zone rules to 'inside the plant'.

11.3 Patterns Adopted and Set Aside

Three patterns were adopted. Triggered one capture per body at each station exit, with identity from the body number and MES genealogy, was adopted over multi-object tracking because each body is captured

once and identity is already owned by a system of record. Nuisance-image collection before threshold setting was adopted from the case-study record as a gate, not a chore. Memory over confirmed outcomes was adopted over any form of on the fly training, which the requirement rules out and the architecture makes unnecessary.

Two were set aside. Packaged smart cameras were set aside for the Apache-2.0 detectors on the operator's own edge units, because the packaged class is rated for ambient temperatures below a body shop summer in this climate and its weights would not belong to the operator. Personnel and vehicle positioning was set aside because cameras watch bodies and vehicles, and worker location sensing would collide with the national data protection rules for cameras on workers.

11.4 Where the Reasoning Lands

The reasoning lands on equipment classes rather than named products where the plant environment decides: the node count by GPU class that sizes the work surface model's node; the frontier inference cluster class (GPU nodes for 300B to 1T parameter open weight models) itself; the plant server room GPU node ASHRAE class; the industrial fanless edge unit on an octa-core 6 TOPS NPU SoC rated for minus 40 to 85 degrees Celsius in a sealed aluminium enclosure; the 5 MP global shutter GigE Vision cameras with strobed lighting that freeze a moving body without rolling-shutter skew; and the export-control determination that turns on total processing performance rather than memory, which is why the license checkpoint sits at phase-one start. Everything shown in this chapter was recorded reading, and nothing is inferred: where the plan holds an assumption to confirm, it is carried as one.

SOURCES

Source Register

NIST. n.d. Recommendations on Ensuring Traceability and Trustworthiness of Manufacturing-Related Data. https://nvlpubs.nist.gov/nistpubs/ams/NIST.AMS.300-10.pdf

IAF. n.d. Supplier Quality & Traceability Requirements, Suppliers & Technology Providers, International Automotive Federation. https://www.iaf.com/membership/industry/suppliers-technology-providers/supplier-quality-traceability-requirements/

NHTSA. 2024. Chronology of Defect / Noncompliance Determination. https://static.nhtsa.gov/odi/rcl/2025/RMISC-25V169-3609.pdf

Safeinindia. 2024. Why do many supplier factories of auto mobile brands misrepresent facts about work-accidents?: A lost opportunity for improvement in Indian manufacturing. https://www.safeinindia.org/post/why-do-many-supplier-factories-of-auto-mobile-brands-misrepresent-facts-about-work-accidents-a-los

Qualitymag. 2026. Manufacturers Must Unify Data Now to Meet First Digital Product Passport Deadlines Quality Magazine. https://www.qualitymag.com/articles/99560-manufacturers-must-unify-data-now-to-meet-first-digital-product-passport-deadlines

About CodeNinja

CodeNinja is a Middle Eastern American artificial intelligence lab and system of context company. It builds sovereign self-improving systems that organizations own, run inside their own boundary and improve with every decision they make.

codeatoms.ai