VERTICAL-DRIVEN ARCHITECTURES · ENERGY & UTILITIES · DESIGNED WITH PRAXIS · NOVEMBER 2026

Sovereign AI Architecture for Distribution Loss and Theft Reduction in Pakistan

Energy & Utilities Watch: One Living Picture of the Operation

Muhammad Umar, Cofounder, CodeNinja

For the leaders accountable for energy & utilities 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 sovereign AI architecture for loss and theft reduction at an electricity distribution company in Pakistan: open weight models on the operator's own hardware in the country, reading its systems of record, with nothing about its customers leaving the country or reaching a foreign model service.

The design reads 7 source systems through 1 adapter family into one object model of 15 objects, runs 8 services over it and opens 5 surfaces to the people who decide, with 3 models placed where the data is. A sovereign, operator-owned platform that joins seven systems of record into one governed ontology of the network and its customers, computes each transformer's energy in against energy billed, and lets a loss agent, a tamper agent and a billing agent draft ranked causes, case files and corrections that subdivision engineers, inspection officers and revenue officers approve, with every confirmed outcome kept as a decision record that improves the next ranking, and every theft case ending in a court-ready Grade 17 complaint. Phase one runs one circle of feeders for one quarter against a measured baseline

with unchanged control feeders, on GLM 5.3 self hosted in an in-country data center, gated on four numbers before any other circle is touched.

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.

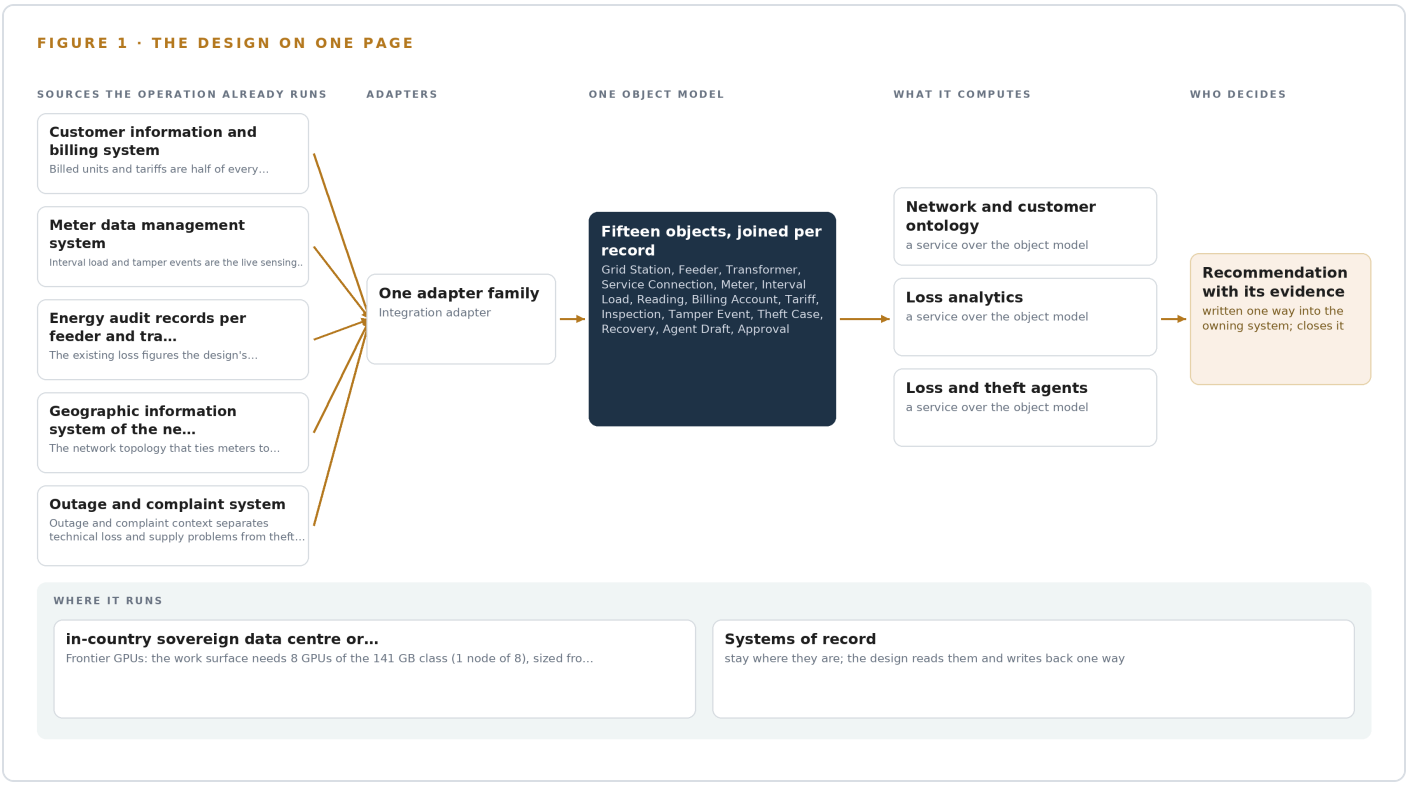


Figure 1. Energy & Utilities Watch on one page: the sources the operation already runs, one object model, what it computes, and the person who decides.

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	Fifteen 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 posed the question and promised a sovereign AI architecture that answers it inside the operator's boundary. This chapter states the question precisely, prices the cost of leaving it unanswered, and fixes the scenario the design must serve.

1.1 The Question and the Data It Requires

The operation needs one question answered every day for every transformer: why does the energy that entered leave a gap between what was measured in and what was billed downstream, and what should the inspection team do about it today. The data that answers it sits in four places at once. The energy entering each transformer comes from the feeder-head and transformer metering; the energy billed to every service connection downstream sits in the billing system; the topology that ties a meter to a transformer to a feeder sits in the geographic information system; and the human record of tamper events, inspections and closed cases sits in three further systems. The answer must also satisfy several regimes at once: the national program replacing every conventional meter by the end of 2026, a

government-ordered technical audit of billing systems now under way, privatisation preparations that place every process under external scrutiny, and theft prosecutions brought under the Penal Code section for electricity theft, filed as a Grade 17 complaint before the Electricity Utilities Court. Those regimes constrain the means as well as the ends, because the operator's requirement is explicit that nothing about its consumers may leave the country or reach a foreign model service.

1.2 The Documented Cost of the Problem

The direct cost is stated in the operator's own requirement: companies of this class lose a large share of the energy they buy to technical losses, theft and unbilled supply, and the sector's distribution losses ran to hundreds of billions of rupees in the last financial year. The cost of mishandling the data that would fix it is documented publicly. A utility in North America paid a 2.7 million dollar penalty after a vendor exposed its data (Natlawreview 2018), and a Japanese utility was issued business improvement recommendations by its market surveillance commission after consumer information obtained through a third party was leaked and accessed (Chuden 2023). Energy sector operators are explicit targets of state-sponsored cyber actors (CISA 2022), and a regulator elsewhere has opened a board inquiry into a cybersecurity incident at a power company (Nserbt 2025). Any architecture that sends consumption records abroad to a hosted model service carries exactly this class of exposure, which is why the boundary is a cost line here and not a slogan.

1.3 The Operation as a Scenario

The operator is an electricity distribution company in Pakistan, described by class only. It distributes power through hundreds of grid stations and thousands of feeders to several million consumers, and it knows its losses by feeder a month late and by transformer almost never. The people in the loop are subdivision engineers who own feeders, inspection officers and field teams who visit meters and transformers, revenue officers who own billing corrections, and the Grade 17 officer who signs each theft complaint; the loss reduction team works from a command center, and all compute runs in an in-country data center or colocation on the operator's own hardware. The physical environments are the field, where inspections happen and photographs are taken, and that data center, where every model runs. The regulatory ground combines the meter replacement program, the billing audit, privatisation conditions and the Penal Code theft route through the Electricity Utilities Court. The design counts seven named systems of record, the roles listed above, and places counted as grid stations, feeders, transformers, circles, subdivisions and courts; phase one runs one circle of feeders for one quarter before any other circle is touched.

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 fixed the question and the regimes it must satisfy. This chapter walks the seven systems of record one by one and shows that each holds one slice of the answer and none holds the whole.

2.1 What Each System Sees

The consumer information and billing system sees every account with its billed units, tariff and arrears, but it cannot say which transformer feeds the premises or whether the billed units match the energy that reached it. The meter data management system sees interval load and tamper events for every smart meter, but not the billing history of the account behind that meter or what an inspection later found there. The energy audit records hold the loss percent per feeder each month and almost never per transformer, and they carry no tamper events, no inspections and no case outcomes. The geographic information system holds the topology from meter to transformer to feeder to grid station, which is the skeleton of every energy balance, but it holds no measurement at all. The outage and complaint system records outages and complaints with times and places, which is what separates a supply problem from theft, yet it cannot join either to the meters and accounts affected. The meter reading and inspection app holds field findings and photographs, but the team using it arrives without the tamper alarms or the billing history that would tell it where to look. The theft case register holds case numbers, penal sections and recovery amounts, while the evidence that would have made each case defensible was never joined to it. Figure 2 sets these seven views side by side, converging on the question none of them can answer alone.



Figure 2. Seven 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

No system joins energy in against energy billed at transformer level with the tamper record, the inspection record and the topology in one argument, and that join is the whole answer. In practice a feeder losing a third of its energy has hundreds of possible causes: a tampered meter, an illegal connection, an overloaded transformer, a wrong meter multiplier, a reading never taken. Because no system ranks those causes, field teams are sent on hunches, tampering events reported by the meters are never joined to billing history or the inspection record, and case files are assembled by hand and fall apart on paperwork. The engineers who know which neighbourhoods and which hours the losses come from are retiring or leaving ahead of privatisation, so the tacit knowledge that once substituted for the join is walking out of the building.

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: seven systems, seven slices, no shared argument. This chapter fixes the goals and the exclusions that every component choice downstream had to satisfy.

3.1 Sovereignty Is a Boundary, Not a Preference

The first goal is that the operator owns the intelligence it produces: open-weight models served on its own hardware inside the country, reading its systems of record, with nothing about its consumers leaving the country or reaching a foreign model service. This is hard in this industry because the hardware itself is governed: export controls turn on total processing performance, so the GPU class a frontier model needs requires a license determination before any card is committed for Pakistan, with a non-US accelerator class held open as the alternative. Hosted model services are ruled out twice over, by the operator's own requirement and by the documented cost of utility data exposure abroad (Natlawreview 2018).

3.2 One Governed Model of the Network and Its Consumers

The second goal is a single governed ontology joining all seven systems of record, so that a query can walk from a feeder through its transformers to a meter, its readings, its tamper events and its closed cases. This is hard because no packaged connector set exists for the seven systems of a distribution company of this class: each connector is built as real work, and billing and network records change on their own cycles, so the design must choose carefully which systems are worth streaming and which are worth batching.

3.3 Every Decision Carries a Name, and Every Outcome Feeds the Next Ranking

The third goal is that agents draft and people decide: subdivision engineers, inspection officers and revenue officers approve every ranked cause, case file and correction, and every confirmed outcome is kept as a decision record that the next ranking reads, without training a model on the fly. This is hard because approvals degrade into rubber-stamping as trust builds, and rubber-stamped approvals logged as ground truth poison what the next ranking learns from; the design therefore polices its own approvers with known-answer probes and steps an agent class back to shadow mode when the upheld rate falls.

3.4 Proof Before Scale

The fourth goal is a measured claim before expansion: phase one runs one circle of feeders for one quarter against a measured baseline, with unchanged control feeders, and gates any extension on four numbers. This is hard because privatisation pressure rewards fast numbers, while only untouched control feeders make the numbers defensible before the billing audit; the discipline of not touching the other circles is the price of a causal claim.

3.5 Scoping Decisions

Three decisions fixed the design's shape before any component was chosen, and each bought something at a stated cost. Table 1 records them.

Table 1 · Scoping Decisions

DECISION	WHAT IT BUYS	WHAT IT COSTS
Run phase one on one circle of feeders for one quarter, against a measured baseline with unchanged control feeders	A causal claim that survives the billing audit and privatisation scrutiny	No other circle is touched until four gate numbers clear
Serve open-weight models in country on the operator's own hardware	A boundary no foreign model service ever crosses	An export-license checkpoint before any GPU is committed, with a non-US accelerator class held open
Keep every confirmed cause and closed case as a decision record the next ranking reads	Improvement without training on the fly, as the requirement demands	Rankings inherit the quality of past approvals, so probes must police the approvers

3.6 What the Design Chose Against

Each major choice excluded an alternative for a stated reason, and several whole areas were excluded outright. Table 2 records the rejections and closes with what is out of scope and why: no control path is touched, meters come from the national program, filing and prosecution rest with the authorized officer and the Electricity Utilities Court, and the hardware and data center belong to the operator.

Table 2 · What the Design Chose Against

WHERE	WHAT WAS PICKED	INSTEAD OF, AND WHY
Ingestion	Event streams for meter and tamper data, batch extracts for billing, energy audit, GIS and outage records	Live event streams on all seven systems; billing and network records change on their own cycles, and streaming them adds connector risk without improving a monthly transformer balance
Improvement	A kept record of confirmed causes and closed cases, with consumer identifiers minimized	Fine-tuning the work-surface model or training a theft classifier on historical labels; the requirement rules out training on the fly, and past labels are biased by where inspections happened to be sent
Model serving	GLM 5.3 at FP8 on one node of the 141 GB GPU class	BF16 serving on two nodes; FP8 halves the weights to about 756 GB so one node holds them, and doubling the hardware doubles export-license exposure for an accuracy gain the ranking and drafting tasks have not been shown to need
Sensing and edge compute	None in the field; the platform reads the meters and their head end from the data center	Cameras and edge nodes; the sensing already exists at the meters, and every added device widens the maintenance and export surface
Out of scope	SCADA integration, meter installation, model training, prosecution and recovery, transaction advisory, hardware and infrastructure, third-party license fees, historic data migration, estate-wide training, round the clock support	Any of those inclusions; no control loop is entered, meters arrive under the national program, the complaint pack ends at the Grade 17 signature, and the operator owns the machines the platform runs on

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: the country's boundary around every byte, the join before the analytics, a named officer behind every action, and a measured gate before any extension. This chapter turns those constraints into a stack of named components, each of which the operator can hold, inspect and replace.

4.1 Why the Pattern Fits

The pattern is the three layers with one direction of dependence that a sovereign AI architecture for a utility converges on. Below sit the seven systems of record exactly as they run today, named in full in Table 3: the billing, meter data, energy audit, geographic network, outage, field inspection and theft case systems. None is replaced and none is bypassed. In the middle sits one object model, fifteen objects with typed links, each anchored in exactly one system of record, holding the join no single system can make. Above it sit the applications and agents: eight services that compute each transformer's energy in against energy billed, rank causes, draft case files and corrections and record approvals, and five surfaces that put those drafts in front of the subdivision engineers, inspection officers and revenue officers who decide.

The pattern fits because the operator's problem is a join failure, not a sensing failure: the meters already report tampering and the registers already record cases, but no system can answer which transformer, which cause, which case in one question. A middle layer that owns the join lets every system keep its role while the intelligence compounds above it, and because the agents only draft and the officers approve, nothing above the object model ever writes to a system of record directly. Figure 3 shows the layered stack with its counts: seven sources, one adapter family, fifteen objects, three models, eight services and five surfaces.

FIGURE 3 · THE LAYERED STACK: 5 SURFACES, 8 SERVICES, 15 OBJECT MODEL, 3 INFERENCE, 1 ADAPTERS, 7 SOURCES

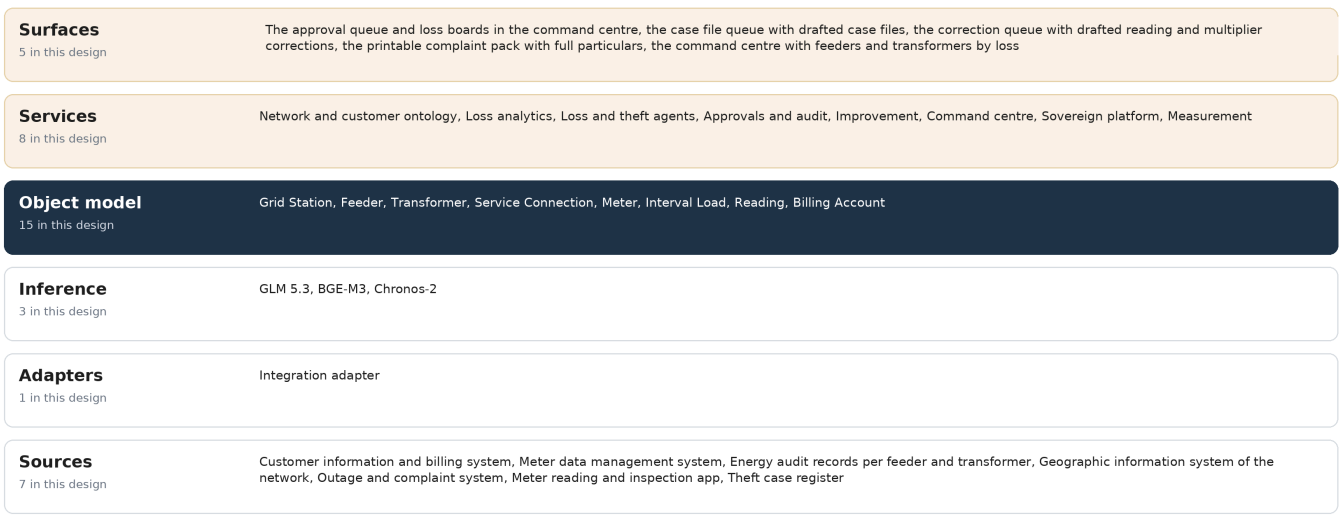


Figure 3. The layered stack: 7 sources, 1 adapter family, 15 objects, 8 services and 5 surfaces.

Table 3 · The Stack, Stage by Stage

STAGE	WHAT IT IS RESPONSIBLE FOR	HOW
Sources	Holding the billing, metering, audit, network, outage, field	Seven systems of record: the customer information and billing system, the meter data management system, the energy audit records per feeder and transformer, the geographic information

STAGE	WHAT IT IS RESPONSIBLE FOR	HOW
	and case records the platform reads	system of the network, the outage and complaint system, the meter reading and inspection app, the theft case register
Sensing	Producing the live signals the design runs on	Smart meters and feeder-head meters delivering interval load and tamper events into the meter data management system
Adapters	Owning every contract between a source and the object model; nothing enters any other way	One integration adapter family: a Kafka subscription for meter data, scheduled batch extracts for billing, energy audit, GIS and outage records, mapped connectors for the field app and the theft case register
Object model	Holding the join as fifteen objects with typed links, each anchored in one system of record	The network and customer ontology service on an ontology or graph store, with interval load in the time series store beside it
Inference	Running the three models inside the sovereign boundary	GLM 5.3 at FP8 on vLLM across one node of the frontier inference cluster class (GPU nodes for 300B to 1T parameter open weight models); BGE-M3 and Chronos-2 on the on site inference server class
Services	Computing, ranking, drafting, recording and measuring	Eight services: network and customer ontology, loss analytics, loss and theft agents, approvals and audit, improvement, the command center, sovereign platform, measurement
Surfaces	Putting drafts, evidence and boards in front of the people who decide	Five surfaces: the approval queue and loss boards, the case file queue, the correction queue, the printable complaint pack, the command center

4.2 The Stack Stage by Stage

Table 3 walks the stack stage by stage, from the records the operator already holds to the surfaces its people already work at, naming the component responsible at each stage.

4.3 A Reusable Pattern

A team carrying a similar brief reuses the shape and swaps the contents. Keep the direction of dependence: sources untouched below, one adapter tier owning every contract, one governed object model holding the join, agents that draft and never act, approvals recorded as objects carrying a hash and a written reason, and all inference inside a boundary the organization owns. Swap the seven sources for the industry's own records, the fifteen objects for its assets and measures, the energy-balance ranking for its own decision, and the models for whatever the license terms and hardware permit. What is never swapped is the rule the stack stands on: nothing above the adapters may know a source's schema, and nothing below the object model may know an agent exists.

Table 7 · Architecture, Layer by Layer

LAYER	COMPONENT	CHOICE	WHY
Event streaming and messaging	by class, no product named	Product	Apache Kafka
The ontology or graph store	by class, no product named	Class	An ontology or graph store class, fixed by what the fifteen objects and their links demand
Time series and telemetry storage	TimescaleDB on PostgreSQL	Class	Interval load volume for several hundred thousand meters is the bulk of the design's storage and its rate is not yet stated; the shelf holds the InfluxDB 3, TimescaleDB and QuestDB comparison for exactly this OT telemetry case, and the choice is committed at design against the measured interval rate.
Video ingest, recording and management	by class, no product named	Not needed	No cameras anywhere in this design; the sensing is smart meter interval data, tamper events and the records of seven systems of record.
Edge model serving	by class, no product named	Not needed	No edge compute; all inference runs in the sovereign data center on the operator's own hardware.
On site language model serving	by class, no product named	Product	VLLM
Object tracking and trajectory forecasting	by class, no product named	Not needed	No video and no moving object tracking; the trajectories here are loss trends over time, served by the time series model.
Edge orchestration of containers	by class, no product named	Not needed	No edge nodes to orchestrate; the platform runs as standard containers in the sovereign data center.
An air gapped artifact and model registry	by class, no product named	Product	Harbor
Labelling and retraining	by class, no product named	Not needed	Under system of context nothing is trained, fine tuned or distilled into weights: confirmed causes and closed cases become the record the next ranking reads, and the requirement asks for exactly that ('without training a model on the fly').
Time synchronisation across sensors and systems	by class, no product named	Product	Chrony

LAYER	COMPONENT	CHOICE	WHY
Observability of the running system	by class, no product named	Product	Prometheus

PART II · CHAPTER 5

Fifteen Objects Turn Operational Data Into One Argument

The object model is the join the systems never made.

Chapter 4 placed one object model at the center of the stack and gave every other layer its reason: the adapters feed it, the services read it, the surfaces show it. This chapter opens the model itself, object by object and link by link.

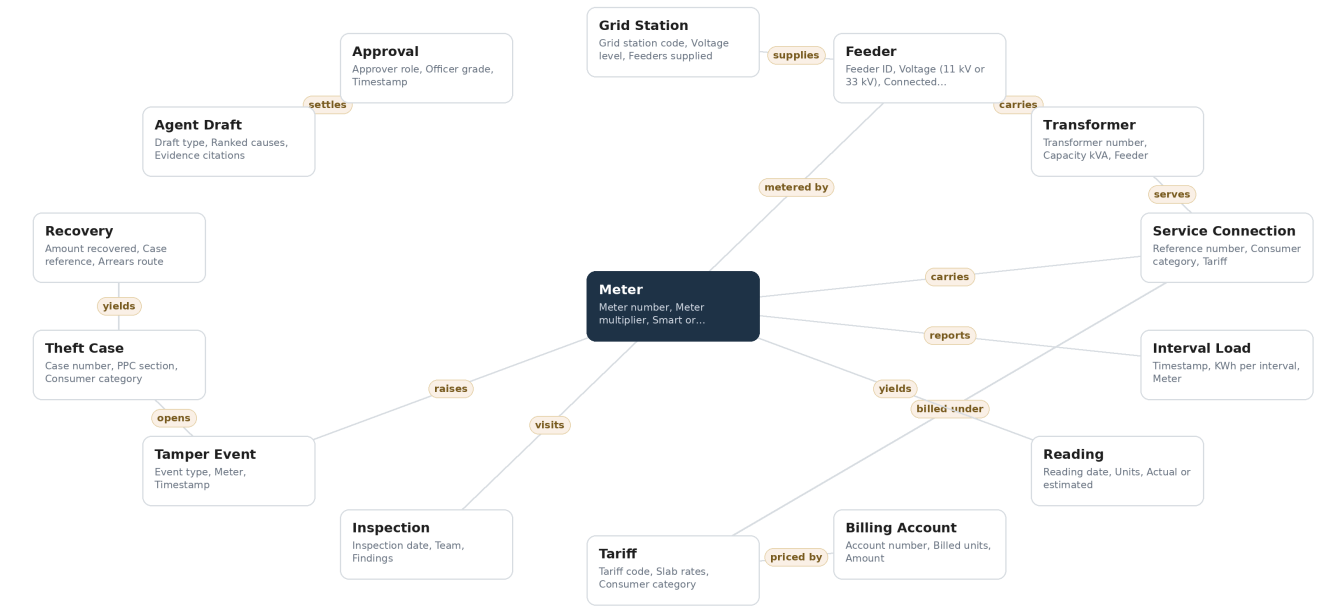
5.1 Every Object and Its Typed Links

The model holds fifteen objects: the grid station as its one site; the feeder, the distribution transformer, the service connection and the meter as assets; the interval load as its one measure; the reading, the billing account, the tariff, the theft case and the recovery as records; the inspection, the tamper event and the approval as events; and the agent draft as the document the whole loop turns on. Each is anchored in exactly one system of record, so provenance sits on every object's face: the transformer's energy in and energy billed come from the energy audit records, the meter's seal state from the meter data management system, the theft case's court and PPC section from the theft case register.

The typed links are what make the model an argument rather than a filing system. A grid station supplies feeders; a feeder is metered at its head and carries transformers; each transformer serves service connections; each connection carries a meter and is billed under a tariff; a meter reports interval load, yields readings and raises tamper events; an inspection visits a meter; a tamper event opens a theft case; a theft case yields a recovery; and an approval settles an agent draft. Figure 4 shows every object and every link, with the transformer at the center in navy.

One traversal across those links answers the question field teams actually ask: every transformer on a feeder whose monthly loss percent crossed its threshold, the tamper events its meters raised in the same window, what was billed to the connections it serves, and which inspections already visited. That question crosses four systems of record, and a document store can answer it only if the words co-occur in one document, which they never do, because the tamper event lives in one system, the billing history in another and the inspection findings in a third. Table 8, printed by the platform, records each object with the system it is anchored in and its typed links.

FIGURE 4 · THE OBJECT MODEL: 15 OBJECTS AND 14 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 fifteen 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
Grid Station	The geographic information system of the network	supplies: Feeder
Feeder	The geographic information system of the network	carries: Transformer; metered by: Meter
Transformer	The energy audit records per feeder and transformer	serves: Service Connection
Service Connection	The customer information and billing system	carries: Meter; billed under: Tariff
Meter	The meter data management system	reports: Interval Load; yields: Reading; raises: Tamper Event
Interval Load	The meter data management system	no typed link recorded
Reading	The customer information and billing system	no typed link recorded
Billing Account	The customer information and billing system	priced by: Tariff

OBJECT	ANCHORED IN	TYPED LINKS
Tariff	The customer information and billing system	no typed link recorded
Inspection	The meter reading and inspection app	visits: Meter
Tamper Event	The meter data management system	opens: Theft Case
Theft Case	The theft case register	yields: Recovery

5.2 The Human Loop and the Hosting Posture

The loop lives inside the model, not around it. Every agent draft is an object carrying its draft type, ranked causes, evidence citations, model version and confidence, and its status moves through drafted, approved, amended, rejected and expired; every approval is an object carrying the approver role, the officer grade, the timestamp, the artifact hash and the written reason, and the hash ties the approval to the exact draft text it settled. The surfaces are where the loop is worked: the command center where the loss reduction team sees feeders and transformers by loss, the approval queue, the case file queue, the correction queue and the printable complaint pack.

The hosting posture follows the requirement. The model and every service above it run in an in-country sovereign data center or colocation on the operator's own hardware; identity is held by Keycloak inside the boundary; there are no external links; and no consumer record leaves the country or reaches a foreign model service. The single write path out of the platform is the approved correction: reading and multiplier corrections land in the customer information and billing system, and approved theft cases land in the theft case register, each through the adapter tier and each traceable to the approval object that authorized it.

5.3 One Object in Its Recorded Form

The grid station record below shows the id, label, kind, anchor, properties and links the platform prints for every object; the tamper event, which carries the design's sharpest signal, the meter's own report that something was done to it, is recorded the same way.


```
{ "id": "grid-station", "label": "Grid station", "kind": "site", "anchored_in": "The geographic information system of the network", "properties": [ "Grid station code", "Voltage level", "Feeders supplied", "Circle", "Subdivision" ], "status_vocabulary": [], "links": [ { "to": "feeder", "label": "supplies" } ] }
```


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 what the object model holds. This chapter describes how data gets into it: seven sources, one adapter family, and a backbone that keeps every event ordered, buffered and inside the country.

6.1 Seven Sources and Their Provenance

All seven sources are the operator's own commercial systems of record, provenance class operator, read direction only: the customer information and billing system, the meter data management system, the energy audit records per feeder and transformer, the geographic information system of the network, the outage and complaint system, the meter reading and inspection app and the theft case register. Each earns its place: billed units and tariffs are half of every transformer's energy balance, interval load and tamper events are the live sensing, the audit records are what the computed balances are validated against, the geographic system ties meters to transformers to feeders to stations, outage and complaint context separates supply problems from theft in the rankings, the field app is where approved inspections land and findings and photographs return as evidence, and closed cases are what the next ranking learns from.

The transport is split deliberately, as Figure 5 shows. The meter data management system is read through a Kafka subscription, because interval load and tamper events lose their value if they wait for a batch; billing, energy audit, geographic and outage records arrive as scheduled batch extracts, because they change on their own cycles and streaming them would add connector risk without improving a monthly transformer balance; and the field app and the theft case register are reached through mapped connectors built for their interfaces, because no packaged connector set for a distribution operator's records exists on any shelf.

FIGURE 5 · 7 NAMED SYSTEMS, 1 ADAPTER FAMILY, ONE OBJECT MODEL

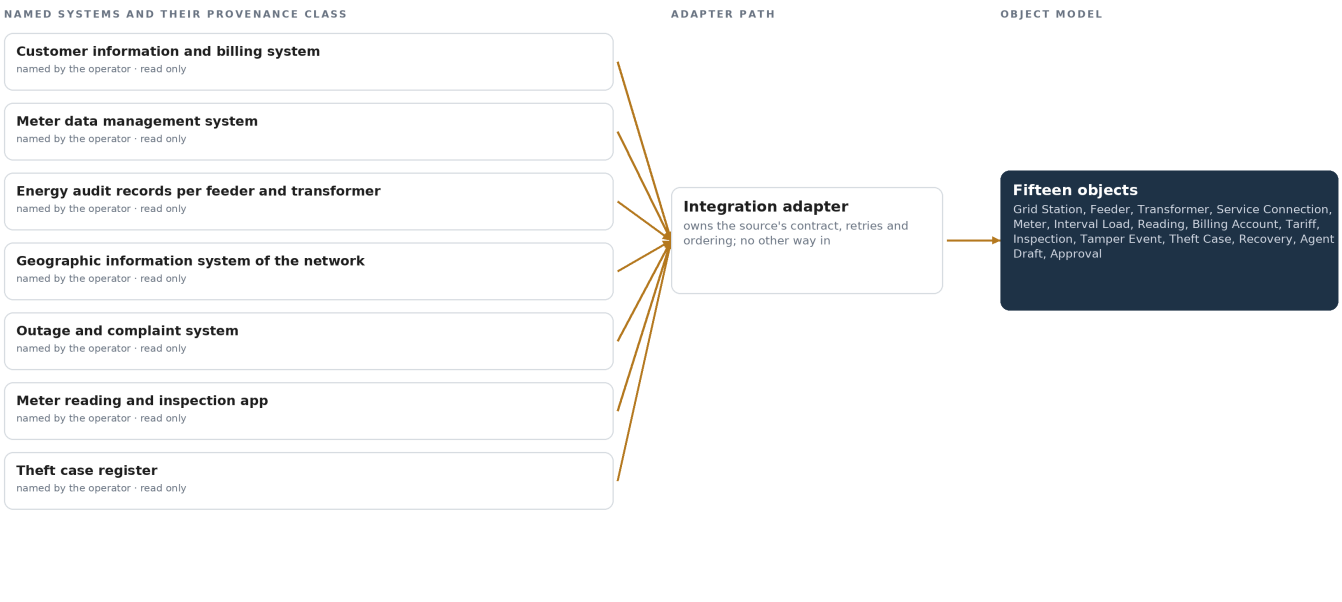


Figure 5. The 7 named systems, the adapter path each one takes, and the object model they all map into.

6.2 What the Adapter Tier Guarantees

The adapter tier owns every contract, and the guarantee is absolute: no service reads a source directly, and no source is written except through an approved correction path. Each adapter maps the source's schema onto the object model, normalizes units and timestamps, and carries the meter multiplier through untouched, because a wrong multiplier is itself one of the ranked causes, not something to silently correct. Every extract is minimized to the agreed fields, stamped with its provenance, and written idempotently so a re-handed over event collapses rather than duplicates. Each feed carries a freshness window with stale and flatline detection: a stale feed flags its drafts and holds them, and the staleness is visible on the command center, so no ranking rides on data nobody noticed had stopped.

Minimization is defense, not courtesy. A leak of consumer information through an operational system drew business improvement recommendations from a market surveillance commission in another utility (Chuden 2023), and a vendor data exposure cost a utility a seven-figure penalty (Natlawreview 2018); both were records leaving through a path nobody owned. Here every path is owned by one adapter, every extract stays inside the country, and every field in it is agreed.

6.3 The Event Backbone

Apache Kafka is the backbone. Interval load and tamper events travel on topics partitioned by meter number, so ordering holds per meter, which is the ordering a balance and a tamper join need; delivery is at least once, and the adapters' idempotent writes make re-delivery harmless. Buffering absorbs the bursts a head end produces after an outage, and replication inside the in-country data center means a broker loss does not lose a tamper event. Updates arrive the same way everything else does, as versioned images through the Harbor registry and never over an open link; Chrony keeps the clocks aligned so an event joins to the right billing period; and Prometheus watches consumer lag, so a feed falling behind is seen before it goes stale.

PART II · CHAPTER 7

Inference Belongs Where the Data Is

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

Chapter 6 described how seven systems of record enter through one adapter family, as event streams for the meter data and as scheduled batch extracts for the billing, audit, geographic, outage and case records. This chapter fixes where the inference over that data runs, and what the arithmetic of that placement obliges.

7.1 Each Tier and Its Arithmetic

All inference runs in one place: an in-country sovereign data center or colocation, on hardware the operator owns. There is no edge tier, because the smart meters and their head end already exist and the platform reads them from the data center rather than adding compute beside them; there is no hosted tier, because the operator's own requirement is that nothing about its consumers leaves the country or reaches a foreign model service. What the design carries instead is three serving classes inside that one boundary, and Figure 6 shows each tier, what runs on it and what crosses the boundary to reach it: a frontier node for the work surface model, and a dual-socket server with a 48 GB PCIe inference GPU that hosts the embedder and the forecaster together.

The frontier tier carries GLM 5.3, whose filed parameter count is 753 billion. At FP8, one byte per parameter, that is 753 GB of weights; the 1.2 planning factor for KV cache and activations lifts what the GPUs must hold to 904 GB. One node of eight GPUs of the 141 GB class provides 1,128 GB, which holds the weights with 375 GB left beside them against the 151 GB the planning factor reserves, so the KV cache ceiling is comfortable rather than marginal. The FP8 choice is the reason one node suffices: at BF16 the same model is roughly 1.5 TB of weights, which demands a second node and doubles the export-control exposure for an accuracy gain the ranking and drafting tasks have not been shown to need. Export control is itself a placement constraint, not just a procurement one: US ECCN 3A090 and 4A090 controls turn on total processing performance, so a license determination precedes any card commitment for Pakistan, and a non-US accelerator class stays open as the alternative. That determination may change which accelerator the node carries; it never changes which country the node sits in.

The second tier is deliberately small. The embedding model at FP16 occupies about 1.1 GB and the forecaster at FP32 about 0.48 GB, so both fit on the 48 GB card with two orders of magnitude of headroom, which is why the design gives them one server rather than two.

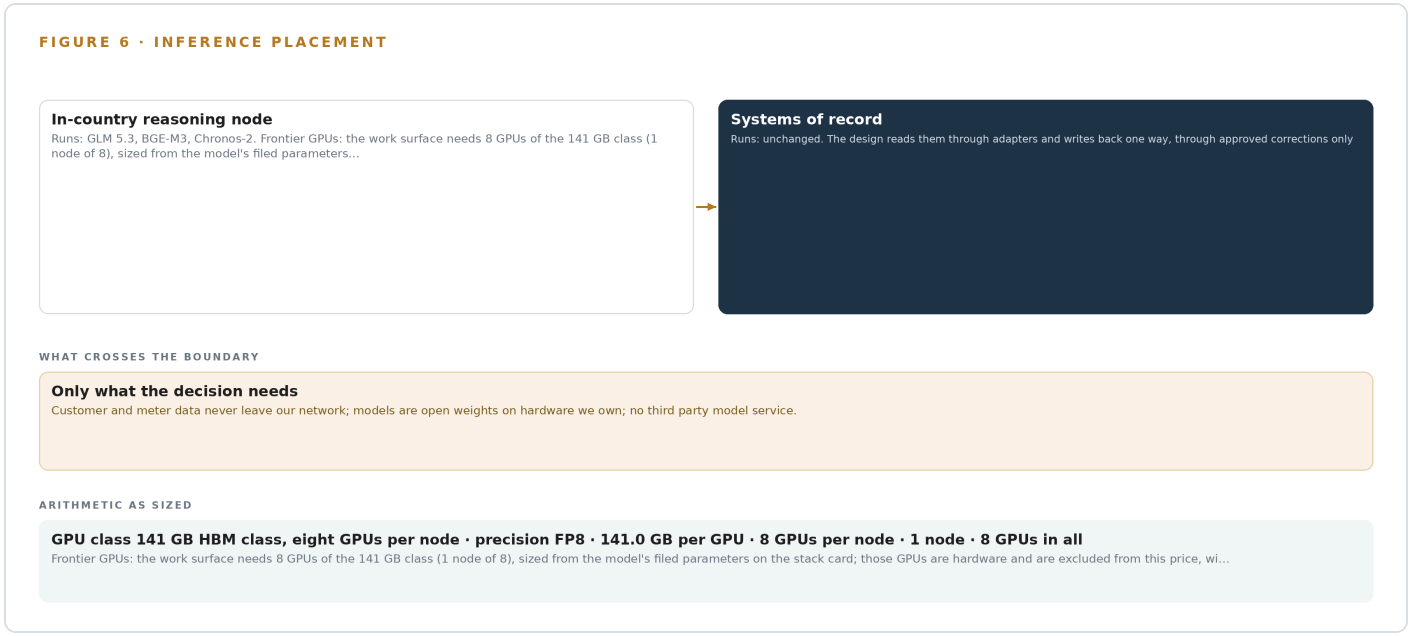


Figure 6. Where each tier runs, what runs there, and the narrow set of outputs that cross the boundary.

7.2 The Latency Budget

The design has no control loop, and nothing in it is allowed to touch one, so its latency budget buys interaction rather than reaction. The outer bound is set by the work itself: a transformer's monthly energy balance is computed on the billing cycle, and a feeder's loss percent moves monthly. Inside that outer bound the budget tightens in three steps. Tamper events must be joined to billing history and drafted as case files within a working day, so an inspection officer opens a queue that is already populated. The loss boards must refresh on the cadence of their feeds, not on the cadence of anyone remembering to run a report. And the command center's conversational interface must answer with citations at reading pace, which local serving on the operator's own hardware holds without any inference call leaving the network. Every hop, from Kafka through the object model to the surfaces, stays inside the operator's boundary, which is what makes the budget credible: there is no wide-area round trip in any path an officer waits on.

7.3 The Boundary and Its Failures

What crosses the network boundary is reads, and reads only. The meter data management system is read through a Kafka subscription; billing, energy audit, geographic information, outage and case records arrive as scheduled batch extracts. The single inbound crossing is the artifact path: model weights and container images enter through the Harbor registry as versioned artifacts, never as a call to a foreign service. This posture answers documented threats rather than hypothetical ones. State-sponsored actors have been indicted for targeting the energy sector specifically (CISA 2022), and a vendor data exposure at a utility drew a seven-figure reliability penalty (Natlawreview 2018), so a design that keeps consumer data off every external path is answering the sector's recorded history, not a general caution.

Each failure path has a stated behavior. If the link to a system of record fails, Kafka buffers the meter event stream and the batch extracts simply wait for their next schedule; per-feed freshness windows with stale and flatline detection flag any feed whose age exceeds its window, drafts built on stale feeds are held rather than sent, and staleness is visible on the command center so no one acts on a silent gap. If power fails, the data center's own resilience carries the platform and Kafka's replication preserves undelivered events for replay. If the update path fails, the artifact registry means the running versions continue to serve; a failed weight or image update leaves the previous version in place rather than leaving the platform down, and the retry is another artifact pull inside the same controlled path.

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 three serving classes inside one in-country data center and showed the arithmetic that fits them there. This chapter names the three models that occupy those classes, and shows why each

one's license was weighed before its accuracy: a license decides whether the operator can hold the weights at all. The model stack is small by design, and Figure 7 shows it whole: one frontier model for the work surface, one embedding model for cited retrieval, one forecaster for the energy balances that trigger every ranking. Three models, three licenses, one country. GLM 5.3 is the frontier model. Its filed parameter count is 753 billion, served at FP8, which is about 756 GB of weights and roughly 1.5 TB if served at BF16; it runs on one node of eight GPUs of the 141 GB class, in the central tier. Its role is the agentic work surface and ontology maintenance: the command center's conversational interface, the drafting of ranked causes, case files and corrections, and the upkeep of the governed object model all run on it. Its license is the GLM-5.3 bespoke license, whose terms permit commercial use, fine-tuning and redistribution, and those three permissions are exactly what an operator-owned system needs: the operator can hold the weights, serve them for its own commercial purpose, and keep serving them regardless of any vendor's fortunes.



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

8.1 Three Models, Three Licenses

BGE-M3 is the embedding model. It occupies about 2.27 GB at float32 as published, about 1.1 GB at the FP16 serving precision chosen here, and about 0.6 GB at 8 bit, and it runs on the 48 GB PCIe inference GPU beside the forecaster. Its role is retrieval embedding for cited answers: the command center answers with citations, which means tariff codes, meter numbers, PPC sections and inspection findings must survive retrieval exactly, and BGE-M3 provides dense, sparse and multi-vector retrieval in a single pass over an 8,192-token window, long enough to embed a full inspection finding rather than a truncated one. It is MIT licensed. It was chosen over multilingual E5, whose 512-token window cannot hold a long finding, and over Nomic v2, whose Matryoshka truncation this design has no use for.

Chronos-2 is the forecaster. It occupies about 0.48 GB at 32 bit, about 0.24 GB at 16 bit, and runs at FP32 on the same 48 GB card. Its role is transformer and feeder energy balance forecasting: it forecasts multivariate series with covariates zero-shot, so the residual between each transformer's forecast and its measured balance is the anomaly signal, with no training run anywhere in the loop. That last clause is a

requirement, not a convenience: the operator asked that the next ranking be better without training a model on the fly, so improvement runs on the kept record of confirmed causes and closed cases, never on the weights. Chronos-2 is Apache-2.0 licensed, and it was chosen over TimesFM 3.0 and Moirai 2.0, whose non-commercial terms cannot ship in an operator-owned system no matter how well they forecast.

The license test is the same for all three, and it is the chapter's argument. A non-commercial or service-only term would convert a sovereign system into a rented one, and a utility that leaked consumer information through a third party has already shown what such dependencies cost in regulatory response (Chuden 2023). Every model here is held as weights inside the country, on the operator's own hardware, under terms that let the operator keep them.

8.2 The Model and Equipment Register

Table 4 gathers the register: each model with its serving class, the hardware classes and sizing rules the placement stands on, the sensing the design reads, the patterns it stands on, and the ground everything runs on. Each row is a decision already argued in this part of the paper; the table exists so an engineer can check the whole stack against one view before quoting hardware or pulling weights.

Table 4 · Model and Equipment Register

THE CHOICE	WHAT WAS PICKED	WHY HERE
Frontier model	GLM 5.3 at FP8 on one node of the 141 GB HBM class	753 GB of weights inside 1,128 GB with the KV cache ceiling covered; the bespoke license permits commercial use, fine-tuning and redistribution, so the operator holds the weights
Embedding model	BGE-M3 at FP16 on the 48 GB PCIe inference class	Dense, sparse and multi-vector retrieval over an 8,192-token window under MIT; tariff codes and case particulars stay retrievable and citable
Time series forecaster	Chronos-2 at FP32 beside the embedder	Zero-shot multivariate forecasting with covariates under Apache-2.0; the residual against each transformer's own baseline is the anomaly signal, with no training run
Hardware classes	a frontier inference cluster class (GPU nodes for 300B to 1T parameter open weight models); an on site inference server class	One frontier node and one dual-socket inference server carry all three models, subject to the export license determination that precedes any card commitment
Sizing rules	Frontier Model GPU Footprint Sizing Rule (vLLM, SGLang); GLM 5.3 Node Count by GPU Class	Weights times the 1.2 planning factor, checked per GPU class before any card is committed, with a non-US accelerator class held open
Sensing	Smart meter interval load and tamper events; the records of seven systems of	No cameras and no field sensors; the meters and their head end already exist and the platform

THE CHOICE	WHAT WAS PICKED	WHY HERE
	record	reads them
Patterns	Asset to alert to work order; improvement through the kept record of decisions, never weights; event streams for meters, batch extracts for records	An anomaly becomes an object linked to its asset and is resolved by approved work, and rankings improve from confirmed outcomes officers uphold
Ground	In-country sovereign data center or colocation on the operator's own hardware	Nothing about consumers leaves the country or reaches a foreign model service; weights and images enter only as artifacts through the registry

PART III · CHAPTER 9

Shadow Mode Comes Before Any Flag Is Trusted

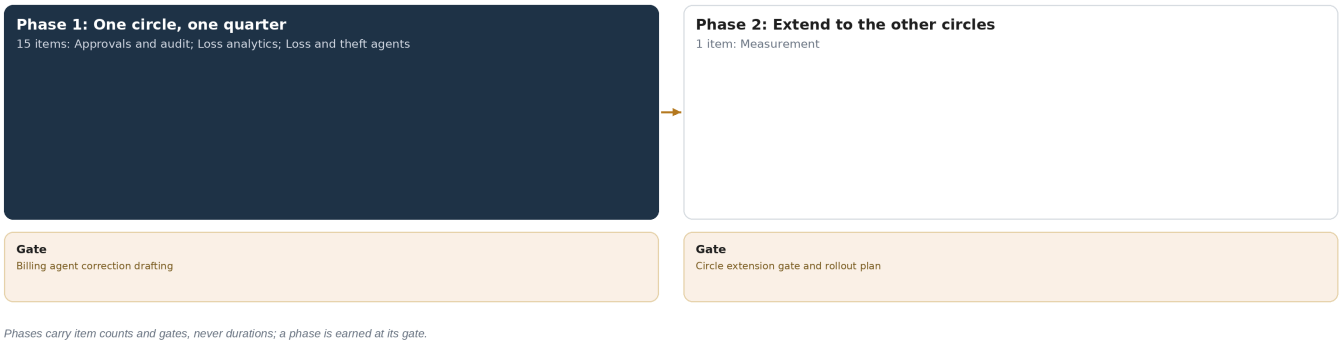
The rollout earns each phase at a gate the operator can read.

Chapter 8 fixed the model stack and the licenses that let the operator hold every weight it runs. This chapter turns from what the system is to how it earns the right to run: one circle of feeders first, a measured baseline, unchanged control feeders, and gates written so the operator can read them without taking the platform's own account of itself on trust.

9.1 Phases, Workstreams and Gates

The rollout is two phases, and neither is a schedule: each phase is a set of items with an exit gate, and the gate, not a date, moves the work forward. Phase 1, one circle and one quarter, carries 15 items across six workstreams: the network and customer ontology, loss analytics, the loss and theft agents, approvals and audit, the sovereign platform, and measurement. Its items include the governed network and customer ontology, the meter and tamper event stream ingest, and the billing and network record batch extracts. Its exit gate is the billing agent drafting corrections: reading and multiplier corrections that a subdivision engineer approves and that land back in the customer information and billing system only through the recorded approval path. The phase runs on GLM 5.3 self hosted in an in-country data center, across one circle of feeders measured against a baseline, with control feeders left unchanged so the comparison stays honest. Phase 2, extending to the other circles, is a single item in the measurement workstream: the circle extension gate and the rollout plan. No other circle is touched until the four numbers that gate reads are met. Requirement coverage holds the design to the same discipline: 29 requirements trace through the two phases, so an item that serves no requirement does not enter the plan, and a requirement with no item is visible as a gap rather than silent. Figure 8 sets the phases, their item counts, their workstreams, their gates and the requirement coverage on one page.

FIGURE 8 · 2 PHASES, 16 SCOPE ITEMS, 0 OF 0 REQUIREMENTS COVERED



Phases carry item counts and gates, never durations; a phase is earned at its gate.

Figure 8. The two phases and their gates, and coverage of the 29 requirements across them.

9.2 What the Rollout Measures

Phase 1 measures against a baseline the operator already trusts, and the control feeders carry the counterfactual. The computed transformer balances, energy in against energy billed, are validated against the existing energy audit records per feeder and transformer before any agent ranking is read. Beyond the balances, the phase measures the confirmation rate of agent-ranked causes at inspection, the acceptance rate of case files at the Grade 17 gate, and the outcome of billing corrections in the system of record itself. It measures the meters actually delivering interval load and tamper events, reported separately from meters installed, because the two populations diverge exactly when the rollout outruns the interface. It measures approval quality through known-answer probes in the normal queue, and feed health through per-feed freshness windows. These measurements feed the four numbers the circle extension gate reads, and the gate is the single path from one circle to the next.

9.3 Failure Modes

Table 5 records what fails and what the design does about it, and every row is a mode the recorded reading or the operator's own situation already produced.

Table 5 · Failure Modes

WHAT FAILS	WHAT THE DESIGN DOES
Approvals degrade into rubber-stamping, and rubber-stamped approvals poison the record the next ranking learns from	Known-answer probes, fabricated positive drafts on historical cases with no violation, run in the normal approval queue; the upheld rate is measured, approvers are re-briefed that the draft is preliminary, and an agent class steps back to shadow mode when the rate falls
Export controls block or delay the GPU class GLM 5.3 needs	The license determination is a phase 1 checkpoint before any GPU class is quoted, and a non-US accelerator class is held open as the alternative

WHAT FAILS	WHAT THE DESIGN DOES
The smart meter rollout outruns the interface, as recorded in the Great Britain rollout where installations ran far ahead of the interface	Meters actually delivering interval and tamper events are reported separately from meters installed, and the connector is tested per meter generation before each circle extension
Case files are rejected for missing evidence, the exact failures the court recorded in acquittals	The completeness check maps every particular the complaint must carry and blocks the pack from the Grade 17 gate until each is present
Theft on conventional meters and unmetered direct taps gives no tamper event and no interval load	No cause attribution is claimed where no meter exists; the transformer balance still bounds the loss, and the blind spot closes as the national meter replacement program completes
Stale feeds silently drive rankings and dispatches	Per-feed freshness windows with stale and flatline detection hold drafts, and staleness is visible on the command center

9.4 Lessons

Shadow the agents before anyone approves them. The trust that matters is the approvers', not the platform's, and trust built without measurement becomes rubber-stamping. The probe mechanism exists because an approval queue that is never tested cannot show that it has stopped judging, and a poisoned record of confirmed causes cannot be un-poisoned later. Report meters delivering data, not meters installed. An installed meter that has dropped out of smart mode is indistinguishable from a conventional one at the data center, and a loss ranking built on the installed count silently overstates its own coverage. The delivery count is the honest denominator for every coverage claim the command center makes. Treat the export license as a design input, not a procurement step. The controls turn on total processing performance, so the determination changes which accelerator class the sizing arithmetic runs on. Making it a checkpoint before any card is committed keeps the design portable between US and non-US classes instead of re-planned after a refusal.

9.5 What Is Still Open

Four questions stay open, and each has a named consequence. The measured interval rate is not yet stated, so the time series store is committed at design against that rate rather than now; settling it fixes the store and the retention arithmetic. The export license determination may move the node between accelerator classes without moving it out of the country; settling it fixes the hardware class and the node count. Interface charges levied by the operator's own vendors on the seven systems of record are payable by the operator; settling them fixes the connector cost line. The blind spot on conventional meters and unmetered taps closes only as the national program completes; until then the design bounds that loss without attributing it.

The Intelligence Should Stay with the Operation That Produced It

What the design builds belongs to the operator.

Chapter 9 ended at the circle extension gate, where the operator reads four numbers and decides whether the work widens. This chapter states what the operator holds by the time that gate is read, and what stays with the operation no matter how the platform grows.

10.1 What the Operator Holds

The object model is the operator's from the first item of phase 1: 15 objects, from the grid station to the approval, each anchored in one of the seven systems of record the operation already runs and joined by typed links that no document store reproduces. The weights are the operator's too. GLM 5.3 runs under a license whose terms permit commercial use, fine tuning and redistribution; BGE-M3 runs under MIT; Chronos-2 runs under Apache 2.0. All three run on hardware the operator owns, in an in-country data center, and no fine-tune exists anywhere in the design: the requirement asks that the next ranking improve without training a model on the fly, so confirmed causes and closed cases are kept as a record the next ranking reads and the weights stay as issued. The decision record is the operator's audit trail and its improvement engine at once. Every approval carries the approver role, the officer grade, the timestamp, the artifact hash and a written reason, and a recorded approval is the single write path into the systems of record. The boundary is the operator's to hold and to defend, and the recorded reading says why it must be defended: energy utilities are named targets of state-sponsored cyber actors (CISA 2022), a leak of customer data drawn from a utility's own systems drew business improvement recommendations from a sector regulator (Chuden 2023), and a vendor data exposure at a utility cost a seven-figure penalty (Natlawreview 2018). In-country serving on owned hardware, with nothing about customers leaving the country or reaching a foreign model service, is the single control that answers all three precedents.

10.2 The Offer Behind the Design

CodeNinja produced this design on Praxis, its platform for designing physical AI systems, and each element of it maps to CodeNinja's offer. The governed model of the network and its customers, 15 objects anchored in the operator's own systems of record, is Hyper Ontology. The ranked causes, case files and corrections that a named officer decides on, with every decision recorded against an artifact hash, are Decision Systems. The loss reduction command center, where the operator's engineers and officers approve agent drafts and build their own agents, is Hyper Pragma. The kept record of confirmed causes and closed cases that the next ranking reads before it acts is Hyper Engram, which does not update model weights. The operator-owned hardware, the open-weight licenses and the in-country, boundary-held operation are Sovereign Infrastructure. What the design builds therefore compounds inside the operation that produced it, which is the point.

PART IV · CONCLUSION

The Pattern Repeats Wherever Systems of Record Do Not Meet

Seen whole, the design is one object model over the systems an electricity distribution 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 closed the ownership argument, and every element it named, from the object model to the boundary, was placed there by a recorded decision rather than a default. This chapter shows where those decisions came from: every design in this series is produced on Praxis, the design team platform for designing physical AI systems, and this chapter lets a reader trace any choice above back to the reading that justified it (Figure 9).

11.1 Contextualizing the Ask

The ask reached Praxis in the operator's own words: a sovereign AI architecture for loss and theft reduction at an electricity distribution company in Pakistan, with open-weight models on the operator's own hardware in the country, reading the operator's systems of record, and nothing about its customers leaving the country or reaching a foreign model service. Praxis assigned the design to the system of context family in the Energy and Utilities industry. What Praxis read: the operator's requirement document and the pinned energy and utilities brief, together with the stack shelf holding the tooling, model and hardware cards the design draws on. Nothing read was skimmed, and nothing outside it was assumed.

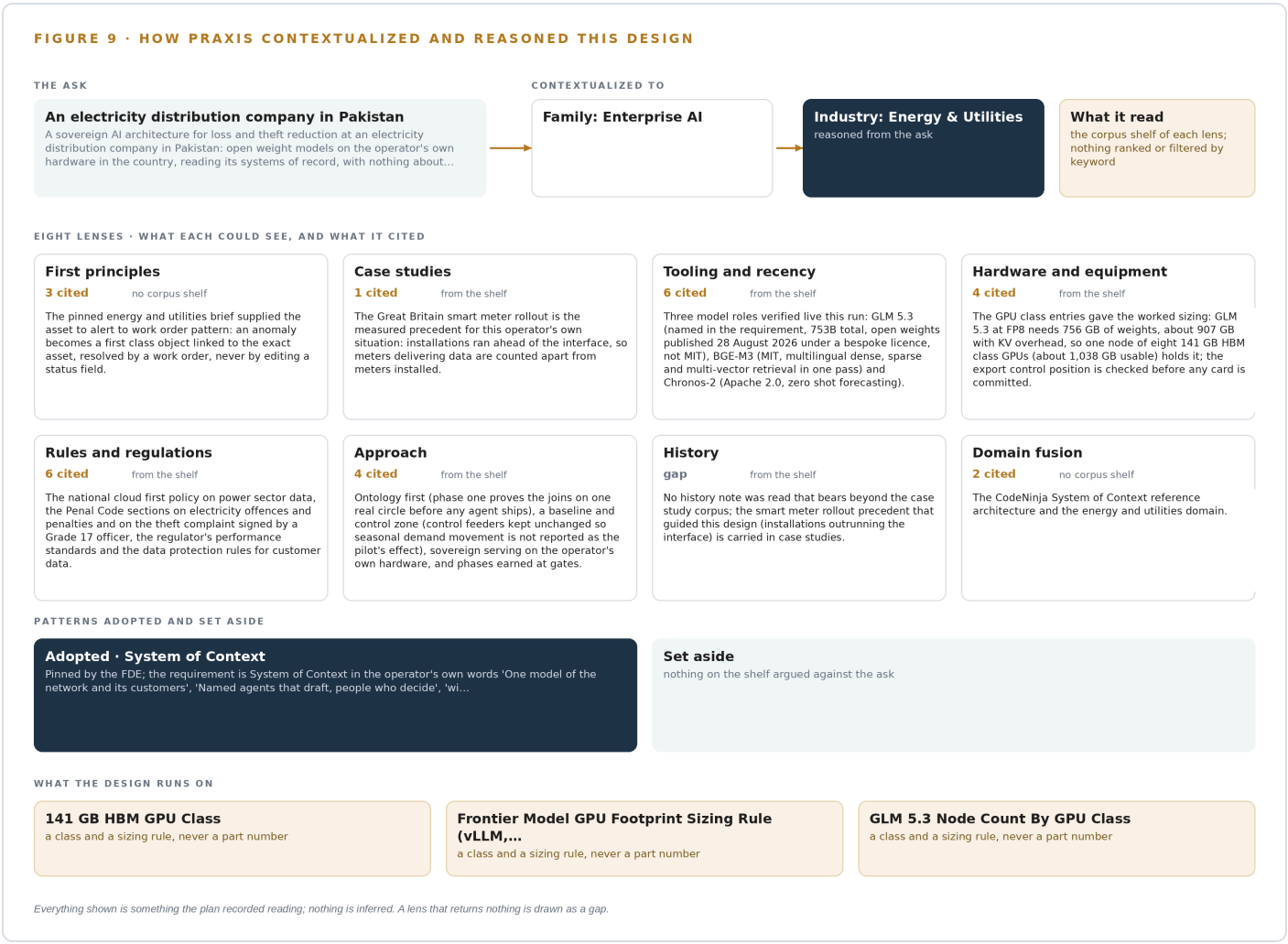


Figure 9. From the ask to the design: the family and industry Praxis assigned, the eight lenses and what each cited, the patterns adopted and set aside, and the equipment the design lands on.

11.2 The Lenses

Praxis read through eight lenses, and Table 6 records what each could see, what it cited and what it contributed to the design. Two lenses returned nothing the design could use, and they are shown as gaps rather than folded into neighbouring rows.

Table 6 · The Lenses and What They Contributed

LENS	COULD SEE	CITED	WHAT IT CONTRIBUTED
First principles	no corpus shelf	3	The pinned energy and utilities brief supplied the asset to alert to work order pattern: an anomaly becomes a first class object linked to the exact asset, resolved by a work order, never by editing a status field.
Case studies	the case study shelf	1	The Great Britain smart meter rollout is the measured precedent for this operator's own situation: installations ran ahead of the interface, so meters delivering data are counted apart from meters installed.

LENS	COULD SEE	CITED	WHAT IT CONTRIBUTED
Tooling and recency	the tooling and recency shelf	6	Three model roles verified live this run: GLM 5.3 (named in the requirement, 753B total, open weights published 28 August 2026 under a bespoke license, not MIT), BGE-M3 (MIT, multilingual dense, sparse and multi-vector retrieval in one pass) and Chronos-2 (Apache 2.0, zero shot forecasting with covariates).
Hardware and equipment	the hardware and equipment shelf	4	The GPU class entries gave the worked sizing: GLM 5.3 at FP8 needs 756 GB of weights, about 907 GB with KV overhead, so one node of eight 141 GB HBM class GPUs holds it, with the export control position checked before any card is committed.
Rules and regulations	the rules and regulations shelf	6	The national cloud first policy on power sector data, the Penal Code sections on electricity offences and penalties and on the theft complaint signed by a Grade 17 officer, the regulator's performance standards, and the data protection rules for customer data.
Approach	the approach shelf	4	Ontology first (phase one proves the joins on one real circle before any agent ships), a baseline and control zone (control feeders kept unchanged so seasonal demand movement is not reported as the pilot's effect), sovereign serving on the operator's own hardware, and phases earned at gates.
History	the history shelf	0	No history note was read that bears beyond the case study corpus; the smart meter rollout precedent that guided this design (installations outrunning the interface) is carried in case studies.
Domain fusion	no corpus shelf	2	The CodeNinja System of Context reference architecture and the energy and utilities domain.

11.3 Patterns Adopted and Set Aside

Three patterns were adopted. The asset to alert to work order pattern shapes the whole object model: a crossing threshold or a tamper alarm becomes a tamper event or an agent draft linked to the exact meter or transformer, and it is resolved by an inspection or an approval, never by editing a status field. The energy-balance pattern drives the analytics: each transformer's energy in against energy billed, forecast by Chronos-2 so the residual against the transformer's own baseline is the anomaly signal with no training run. The recorded-approval pattern guards every write: the single path into a system of record is an approval carrying an artifact hash and a written reason. Three were set aside. Fine-tuning on historical case labels was rejected because the labels are biased by where inspections happened to be sent on hunches, and because the requirement forbids training on the fly. Streaming all seven systems of record was rejected in favour of event streams for meter data and batch extracts for billing, energy audit, GIS and outage records, since billing and network records change on their own cycles and streaming them adds connector risk without improving a monthly balance. BF16 serving on two nodes was rejected in

favour of FP8 on one node of the 141 GB class, because the accuracy gain has not been shown to be needed and the doubling carries export-license exposure.

11.4 Where the Reasoning Lands

The reasoning lands on four equipment classes and their sizing rules: the frontier inference cluster class (GPU nodes for 300B to 1T parameter open weight models) for the frontier work surface model, sized from its filed parameter count at FP8 with the planning factor for KV cache and activations; the frontier model GPU footprint sizing rule under vLLM; the GLM 5.3 node count by GPU class, one node of eight GPUs holding the weights with room beside them; and the on site inference server class hosting the embedder and the forecaster in the same sovereign data center. Every choice traced in this chapter, and every figure in it, was recorded reading: what the lenses saw, what the shelf held and what the operator's own requirement stated. Nothing is inferred, and nothing enters the design that the reading did not contain.

SOURCES

Source Register

Nserbt. 2025. Nova Scotia Energy Board. https://nserbt.ca/sites/default/files/M12273%20-%20Incident%20Report%20%28redacted%29.pdf

Chuden. 2023. Chubu Electric Power Co., Inc. Report on the business improvement recommendations from the Electricity and Gas Market Surveillance Commission and guidance from. https://www.chuden.co.jp/english/corporate/releases/pressreleases/1210773_5163.html

CISA. 2022. AA22-083A: TTPs of Indicted State-Sponsored Russian Cyber Actors Targeting the Energy Sector. https://www.cisa.gov/sites/default/files/publications/AA22-083A_TTPs_of_Indicted_State-Sponsored_Russian_Cyber_Actors_Targeting_the_Energy_Sector.pdf

Natlawreview. 2018. $2.7 Million NERC Penalty Due to Data Exposure. https://natlawreview.com/article/data-exposure-vendor-leads-to-27-million-nerc-penalty-utility

---

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.