Onshore and private deployment options explained

Managed, private VPC and fully self-hosted: what each means for your data, your control and your cost.

What you are actually choosing between

"Onshore" gets used to mean at least four different things in AI procurement, and the confusion is expensive because the four have very different costs. Before comparing options it is worth separating the questions that are usually bundled together.

- **Where the application runs.** The service that holds your workflow logic, your configuration and your queue of work.
- **Where your data sits at rest.** Documents, extracted fields, audit logs, and any working copies.
- **Where inference happens.** The model call itself. This is a separate axis from the first two and it is the one most often assumed rather than checked.
- **Whose cloud account it all lives in.** The question that decides who holds the keys, who reads the logs, and who can revoke access unilaterally.

A deployment can be onshore on the second and third of these while the fourth still belongs to the vendor. That is a legitimate arrangement, and it is also the arrangement most buyers think they are avoiding when they ask for onshore. The three options below differ mainly on the fourth question.

Option one: managed

What it means. The application runs in our environment, on infrastructure we operate. Your data crosses into our boundary and is held there under the retention terms in the engagement. Inference runs in an Australian region. We patch it, monitor it, and restore it if it breaks.

What it means for your data. Your data is in a system somebody else administers. Your protection is contractual and architectural rather than physical: retention terms, access controls, an audit trail and a subprocessor list you have reviewed. That is a real form of protection and it is how most software you already use works, but it is worth naming plainly rather than describing as private.

What it means for control. You inherit our supplier list, our patching schedule and our recovery arrangements. You can inspect all three but you do not set them. Revoking our access means ending the service, because it is our service.

What it means for cost. Lowest to start and lowest to run at modest volume, because the fixed infrastructure cost is shared across clients and there is no environment for you to stand up. The cost is mostly variable and mostly proportional to use, so a quiet month is a cheap month.

The honest tradeoff. This is the right default for a pilot and for most production workloads that do not carry a specific residency or sovereignty obligation. It is the wrong choice if your obligations require that the data never enter a third party's account, or if your own security team needs to see the raw infrastructure logs.

Option two: private VPC

What it means. The same application, deployed into your own cloud account, in your own network, in an Australian region. We build and operate it, but the account is yours. Your identity provider, your network rules, your encryption keys, your logs.

What it means for your data. It never leaves your account. Inference still calls a model endpoint, so the question of what is sent to the model, and under what terms, remains live and has to be answered separately. Anyone who tells you a private VPC deployment means nothing is sent anywhere has not drawn the diagram.

What it means for control. This is where the real change is. You can revoke our access without ending the service. Your security team can read the infrastructure logs directly rather than requesting them. Your existing controls, alerting and compliance tooling apply because it is your account. If we disappeared tomorrow, the system would keep running until something needed changing.

What it means for cost. Adds a fixed floor. You are paying for dedicated infrastructure whether it is busy or not, plus a one off setup cost, plus our operating time. Below a certain volume you are paying for control rather than for capacity, which is a reasonable purchase as long as you know that is what you are buying.

The honest tradeoff. It requires you to have a cloud account and somebody who administers it. If your organisation has no platform capability, a private VPC deployment quietly transfers responsibility to a team that does not exist, and that is worse than managed hosting rather than better.

Option three: fully self-hosted

What it means. You run it. Your infrastructure, your operations, your on-call. We build it, hand it over with documentation, and support it under whatever arrangement you want. Depending on how far you take it, the model may also run on hardware you control.

What it means for your data. Complete containment is achievable, including inference, if you self-host the model as well as the application. This is the only configuration in which no third party is technically capable of seeing your data, and for a small number of obligations that distinction is the whole requirement.

What it means for control. Total, in both directions. You decide when to upgrade, and you also carry the consequence of not upgrading. You own availability, capacity, key rotation, model version management and the revalidation work when a model is replaced.

What it means for cost. This is the option most often mispriced, because the licence is not where the money goes. Self-hosting converts a variable cost into a fixed one. Hosted inference is billed by use, so it scales down to nearly nothing when idle. Self-hosted inference means paying for accelerator capacity continuously, whether the queue is full or empty, and it only wins on cost at sustained high utilisation.

> **Where a number would help, measure rather than assume.** The crossover point between hosted and self-hosted inference depends on your token volume, your utilisation profile through the day, and current hardware pricing. Publishing a general figure would be inventing one. What we do instead is size it with your own measured volume from the pilot, which is one of the reasons the pilot measures token usage as well as time saved.

The honest tradeoff. Choose this when containment is a genuine obligation rather than a preference, or when utilisation is high and sustained. Choosing it for a workload that runs a few hundred times a month means running a small AI platform team to service a workload that did not need one.

The three side by side

	MANAGED	PRIVATE VPC	SELF-HOSTED
WHOSE ACCOUNT	Ours	Yours	Yours
DATA AT REST	Our boundary, Australian region	Your account, Australian region	Wherever you put it
INFERENCE	Hosted, Australian region	Hosted, Australian region, or self-hosted	Your choice, including on your hardware
REVOKE VENDOR ACCESS	Ends the service	Unilateral, service continues	Unilateral, service continues
WHO PATCHES	Us	Us, in your account	You
COST SHAPE	Mostly variable	Fixed floor plus variable	Mostly fixed
TIME TO FIRST RUN	Fastest	Adds account and network setup	Longest
NEEDS A PLATFORM TEAM	No	Helpful	Yes

Residency is not the same as sovereignty

These two words are used interchangeably in vendor material and they are not the same requirement. Getting the distinction right before you write it into a policy saves an expensive correction later.

Data residency is a statement about geography: your data is stored and processed in a particular country. It is straightforward to satisfy and straightforward to evidence, because it is a configuration setting with a region name in it.

Data sovereignty is a statement about jurisdiction: which country's law can compel access to your data. An Australian region operated by a provider headquartered elsewhere gives you residency. It does not by itself remove the possibility of a foreign legal process reaching the parent company. Whether that matters depends entirely on what you hold and who you answer to, and for most commercial workloads it does not. For some government, defence and health workloads it is the requirement.

If your obligation is genuinely sovereignty rather than residency, say so at the start of a procurement, because it narrows the supplier set immediately and changes the architecture rather than the settings. If it is residency, say that too, because it is far easier to satisfy than a sovereignty requirement written by accident.

How to choose

Four questions, in order. The first one that returns a firm answer decides it.

1. **Is there a written obligation that data must not enter a third party's account?** If yes, you are choosing between private VPC and self-hosted, and the rest of the comparison is about cost and capability.
2. **Does your security team need direct access to infrastructure logs?** If yes, private VPC at minimum. This requirement is common and frequently discovered late.
3. **Do you have a platform team who will own the account?** If no, managed, regardless of preference. A private deployment nobody administers is a liability dressed as a control.
4. **Is utilisation high and sustained?** If yes, model the self-hosted inference case with your own measured volume. If no, do not.

Our own default recommendation is managed for the pilot and a deliberate decision before production, because the pilot produces the volume and utilisation data that the decision needs. Choosing the architecture first and measuring afterwards is how organisations end up paying for control they did not need or discovering a constraint they cannot meet.

NorthCape Technology, Victoria, Australia. northcape.tech