Freeform Dynamics

February 2025

The Private AI Infrastructure Imperative: A Practical Perspective

By Dale Vile

Introduction

Early experimentation with generative AI ('GenAI') through generic cloud services has left many organisations questioning where to go next. While services like ChatGPT or AI features embedded in business software provide a taste of the potential, delivering real business value often demands something more substantial and controlled.

The key realisation emerging from early pilots is that effective use of GenAI typically relies heavily on incorporating enterprise data into the mix. The generic knowledge encoded in public large language models (LLMs) may be impressive, but it's inadequate for driving business decision-making and customer interactions. Whether it's systems to support marketing, sales, logistics, service, finance, audit, R&D or any other function, responses need to be grounded in accurate, up-to-date business information.

This creates an immediate set of requirements around data security, compliance and governance – especially when sending sensitive data to public cloud services. Add in lingering uncertainty about which use cases will deliver tangible ROI, and it's not surprising that adoption of GenAI for critical business operations is progressing relatively slowly, with many early initiatives paused or slowed down while organisations reset their thinking.

The question of infrastructure

Mike Hoy, CTO at the hosting company Pulsant, sees this hesitation first-hand but believes it stems partly from organisations finding it hard to work through some of the infrastructure and data implications. During a recent discussion on the topic of GenAI, he began by saying:

"What interests me about GenAI is the need to tap into enterprise data sources. You might be using an LLM up in the cloud, but your business data will often still be sitting on-premises or hosted in a data centre. How are you going to make sure these things have the right access at the right time?"

He argues that while the public cloud is fine for a range of use cases, many others will require a more controlled private environment, especially when dealing with sensitive data:

"From looking across our own customer base, I can see there's an infrastructure requirement that's going to play heavily into implementing GenAI in the right way. People aren't necessarily considering that as much as they should at the moment because AI is being handed to them on an app on a phone or bundled with their office subscription, which is creating a false sense of simplicity."

Hoy's observations here gel with our own at Freeform Dynamics. Public cloud based AI services certainly allow you to gain a sense of GenAI possibilities and 'do something' with minimal implementation effort. However, once you start looking at practicalities in the context of your own business, it's not long before you realise that we are really looking at just another set of data-driven applications.

Contrary to the mantra 'AI changes everything', you still need to consider data sourcing, staging, transformation, quality, transport, management and so on, not to mention security, privacy, compliance, cost and overall system reliability and performance. Put simply, it's business as usual in many respects, especially when it comes to infrastructure.

But how does this sit with the obvious need to experiment and learn in the early days?

Evangelist often give the impression that AI changes everthing, but in systems terms, it really doesn't

A safe space to play

Rather than rushing headlong into production deployments, Hoy advocates creating controlled environments where organisations can experiment safely with their own data:

"What we need is to give business teams and data scientists a proper sandbox environment where they can experiment with AI in a safe, controlled way, without

having to worry about public cloud risks. If we make this easy from an infrastructure perspective, then they can focus on how to get value from the technology and their data. Remove the distractions and constraints, and that's when they'll start to see the benefits and how to drive meaningful outcomes."

Reinforcing the value of this, we ourselves have picked up through our research and discussions with both early adopters and industry experts that it's important to be purposeful in your GenAI adoption. We've already seen pushback against the indiscriminate rollouts often associated with Microsoft Copilot, for example, which underlines the need to identify and focus on specific use cases that will deliver value. At this early stage in the market, when relatively few repeatable use cases and application patterns have so far been defined, this inevitably means taking the kind of exploratory approach Hoy refers to.

Once you have identified an opportunity to exploit GenAI and have shaped a solution, however, another range of considerations kicks in around production deployment and scale-up, which brings us back to the infrastructure question.

The network matters

Beyond the need for adequate compute and storage, an often overlooked aspect of AI infrastructure is the networking requirement. As Hoy explains:

"If you're dependent on AI capabilities to make decisions and run business operations, speed and latency become key. If there's a network related delay, or local data and private feeds can't be reached at all, then the impact can be very serious."

He argues that the conversation needs to move beyond simple bandwidth:

"It's not just about bandwidth and gigabytes. It's actually going to be about latency and how quickly that data can move. The volume of data being created is growing – you've got to be able to move and consume it quickly and reliably. AI-driven systems are not immune to this."

As AI increasingly gets used in a process automation or self-service context, addressing requirements here becomes particularly important.

Sovereignty considerations

Hoy also highlights data sovereignty as a critical consideration that's often overlooked in early GenAI experiments:

"There's a realisation that if you put my data up into a hyperscaler somewhere, there's no guarantee that foreign entities outside your local jurisdiction can't get access to it".

To be fair, this concern is not unique to public AI service providers; it's a familiar discussion in the context of cloud hyperscalers in general. Perhaps Hoy's comment is best taken as a reminder that AI is no different to any other type of application when it comes to controlling how enterprise data is used. It's then down to some physical practicalities as Hoy illustrates:

"Whether it's the far edge within the client's own workspace, or a regional data centre, it matters where the fiber in the ground sits, and how data physically moves around. We need to avoid complacency in areas like this when using AI, and people don't always realise the consequences of their decisions."

AI is no different when it comes to controlling how enterprise data is used

Need for a balanced approach

A lot of what we've touched in relation to private infrastructure to support GenAI initiatives reflects the pragmatic way in which Hoy's company, Pulsant, works with clients throughout the AI adoption and scale-up cycle, as do other MSPs active in this space. Hoy was keen to point out, however, that he expects customers to take a blended approach, saying:

"My message isn't that public cloud services have no role to play – they clearly do, particularly for initial experimentation and certain types of cloud-native applications. However, as organisations move towards more cost and privacy sensitive enterprise AI deployments, private infrastructure will make a lot more sense for many use cases."

We agree, which is why our advice to IT leaders is to consider private AI infrastructure requirements sooner rather than later, whether installed in your own data centre or in your MSP's hosting environment.

For more on CIO views of AI, take a look at our recent report entitled "Generative AI Checkpoint", available for download here.

About Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we help busy IT and business professionals get up to speed on the latest technology developments and make better-informed investment decisions.

For more information, please visit freeformdynamics.com.

Terms of Service