BI CONSULTING SERVICES · PRIVATE AI SERVERS

Every Answer From Your Own Files. In Ten Seconds.

An AI server inside your business that answers from your own documents, contracts, and policies — for every team, without customer or commercial data touching a cloud vendor.

- Customer data stays in-house
- Same answer for every team
- Unlimited seats

THE EXPOSURE

What happens when customer or vendor data leaves for a public AI

Not a hypothetical. This is the ordinary, documented behaviour of hosted AI platforms.

It leaves your control

Customer details and vendor terms transit third-party infrastructure and are stored under that vendor's retention policy — not yours.

It multiplies

Hosted platforms use subprocessors, request logging, and abuse-monitoring pipelines. Depending on plan and settings, human review may sit in that path.

It is regulated data

If you extend financing, consumer financial information carries obligations under the FTC Safeguards Rule that a free chatbot does not come close to satisfying.

It is your moat

Your cost structure, sourcing terms, and vendor agreements are what let you compete. That is the last thing you want sitting on someone else's server.

The uncomfortable version: every customer record or vendor term pasted into a public AI tool is an unlogged transfer to a company you have no agreement with — and it may carry exactly the sourcing and pricing detail that makes your business model work.

THE SOLUTION

The BICS Private AI Server

A dedicated GPU server at your headquarters, running an open-weight language model indexed against your own product data, vendor agreements, policies, and training material. Every location connects over your existing network. Nothing leaves the company.

1

Sovereign

The hardware is yours. The model weights are yours. The index sits on your disk. There is no outbound AI traffic — it will run fully air-gapped if you want it to.

2

Grounded

It answers from your own catalog, specifications, vendor agreements, and policies — with citations back to the source — instead of guessing from the open internet.

3

Unmetered

No per-seat licence. No per-token bill. Once it is installed, your whole team can use it all day and the cost does not move.

Deployed, indexed, and in your people's hands in two to three weeks.

HOW IT WORKS

One boundary. Nothing crosses it.

Every layer below runs on hardware you own, inside a network you control.

YOUR NETWORK PERIMETER

Access

Browser · Microsoft Teams · Desktop shortcut

Staff sign in with the account they already use.

Application

Chat interface · SSO / Entra ID · Role-based access · Full audit log

Who can ask what, and a record of every question and answer.

Retrieval (RAG)

Vector database · Document index · Permission-aware search · Citations

Finds the handful of passages that actually answer the question.

Model

Open-weight LLM on your GPU server · Weights stored locally, version-pinned

Reads those passages and writes the answer. Offline capable.

YOUR DATA SOURCES

- Product catalog and inventory
- Specifications and product data
- Vendor and brand agreements
- Warranty, returns, service policies
- Custom order and repair history
- Training material and sales playbooks
- Handbook and store procedures

Indexed on a schedule you set. Read-only connections. Existing permissions respected.

No inference call reaches the public internet. The server can sit on an isolated VLAN with no outbound route at all.

SIDE BY SIDE

Public cloud AI vs. your own AI server

	PUBLIC CLOUD AI	BICS PRIVATE AI SERVER
Where your data lives	A vendor's data centre, in regions and subprocessors you don't choose	**On a server you can physically walk up to**
Who can access it	Vendor staff, subprocessors, abuse-review pipelines, legal process	**Only accounts your admins provision**
Retention & deletion	Governed by vendor policy and plan tier	**Governed by your own retention schedule**
Used to train models	Depends on plan, settings, and vendor terms — and those change	**Never. The weights are static until you choose to update them**
Cost model	Per seat, per month, forever — and per token on API use	**One-time build, then support. Adding users costs nothing**
If the internet drops	AI stops	**AI keeps working**
Model changes underneath you	Vendors deprecate and silently swap models; your prompts break	**You pin a version and upgrade on your schedule**
Audit trail	Partial, vendor-side, hard to export	**Complete, in your systems, queryable by your team**
Knowledge of your organisation	None, unless you upload it every single time	**Indexed once, permanently available, cited**

Public cloud AI is the right tool for plenty of work. This comparison is specifically about work involving information you are obliged to protect.

PRECISION MATTERS

What we mean when we say "private"

Five commitments you can hand to your IT team, your auditor, or your client.

1 No outbound inference

The model runs locally. We can configure the server with no outbound internet route whatsoever — fully air-gapped — and it still works.

2 The weights are yours

We deploy open-weight models, downloaded once and stored on your disk. No licence check, no phone-home, no deprecation notice.

3 Your index, your storage

The vector database and every extracted document chunk sit on your drives, inside your backup and retention policy.

4 Your identity system

Entra ID / Active Directory single sign-on, role-based access, and document-level permissions that mirror what people can already see.

5 Your audit trail

Every prompt, response, and document retrieved — logged locally, exportable to your monitoring tools, retained as long as you require.

THE PART THAT MAKES IT USEFUL

A model without your files is a clever stranger

Retrieval-Augmented Generation (RAG) is what turns a general AI into one that knows your organisation.

WITHOUT RAG

- Staff re-upload the same documents every single session
- Long files blow past the context window and get silently truncated
- Answers degrade the more you paste in — the model loses the thread
- No citations, so nobody can verify anything
- Every user starts from zero. Nothing accumulates.

WITH RAG

- Your documents are indexed once into an AI-ready database
- The system retrieves only the passages that matter, then answers
- Every answer cites the source file, page, and section
- Accuracy holds as the library grows — millions of pages is fine
- Knowledge compounds. What one person indexes, everyone can use.

Plain English: RAG is the difference between an assistant who has read your files and one you have to re-brief from scratch every morning.

STRAIGHT TALK

Where this is the wrong tool — and we will say so

We would rather lose the sale than sell you a server that ends up unplugged in a closet.

Software development copilots	Coding assistants demand far more GPU capacity per user than document work does. The hardware bill would be several times this proposal and you would still be behind commercial coding tools. Buy those instead.
Absolute frontier reasoning	For the hardest novel reasoning tasks, the largest cloud models still lead. Private AI is excellent on your documents; it is not a claim to beat frontier models at everything.
Public-facing chatbots at scale	Serving thousands of anonymous internet users is a different architecture with different economics. Ask us and we will scope it separately.
Poor document hygiene	If your key documents live in six versions across three drives and nobody knows which is current, AI will confidently repeat the wrong one. We will scope a cleanup first — or advise you to wait.
Pricing and valuation decisions	This retrieves your own policies and records for a trained person to act on. It does not set prices or make valuation calls. Those stay with your people, where they belong.

Selling you something that quietly underperforms is worse for us than not selling you anything.

DELIVERY

Two to three weeks from kickoff to people using it

Hardware is built, burned in, and pre-loaded at BICS before it ever arrives at your site.

WEEK 1

Discovery & Build

- Kickoff, success criteria, and named pilot user group
- Data source inventory and permission mapping
- Security review with your IT team; network and VLAN plan
- Hardware procured, assembled, burned in at BICS
- Model selected, installed, and benchmarked before shipping

WEEK 2

Deploy & Index

- Rack, power, network — onsite or with remote hands
- SSO / Entra ID integration and role configuration
- First knowledge base indexed and citation-tested
- Pilot group onboarded and working the same day
- Audit logging wired into your existing tooling

WEEK 3

Enable & Tune

- Index expanded to the remaining agreed sources
- Evaluation set run against your real questions
- Prompt library built for your top workflows
- Administrator training and runbook handover
- All-staff enablement sessions, recorded for reuse

Week 3 is where adoption is won or lost. A server nobody knows how to use is a very expensive space heater. Enablement is not an add-on in our scope — it is half the engagement.

PREREQUISITES

What we need from you

None of it is exotic — but gaps here are the number one cause of slipped timelines.

Facilities & Network

- Rack space or a secured, ventilated location
- Dedicated circuit — we confirm draw at spec sign-off
- Adequate cooling for a sustained GPU load
- Wired network drop and an assigned static IP
- Firewall change window if rules must be edited

People

- An IT or infrastructure contact available during deployment week
- An identity admin who can configure SSO groups
- An executive sponsor who will use it visibly
- 5–10 named pilot users from one or two teams
- A content owner for each data source we index

Data & Governance

- Agreed list of sources for the first index
- Confirmation of which document set is authoritative
- Existing permission model documented at group level
- Retention and logging requirements stated up front
- Any regulatory constraints flagged before we design

We send this as a formal readiness checklist at kickoff. Everything is signed off before hardware ships, so deployment week is installation — not discovery.

FOR WHOEVER SIGNS OFF ON RISK

How this maps to the obligations you already carry

We are not claiming certification on your behalf. We are removing the third-party data transfer that makes AI hard to approve.

Data residency
Data never leaves the physical location you designate. Cross-border transfer is structurally impossible.

No third-party processor
Nothing to add to a subprocessor register, no agreement to negotiate, no vendor terms to review annually.

Access control
Federated to your existing identity provider. Joiner-mover-leaver runs through the process you already have.

Auditability
Complete prompt, response, and retrieval logs held locally, exportable, retained on your schedule.

What this commonly supports
FTC Safeguards Rule (GLBA) where financing is offered · PCI-DSS obligations · state consumer privacy statutes · authorized retailer agreements · minimum advertised price terms

Air-gap option
Deployable with no outbound route at all, for the environments where that is the only acceptable posture.

Change control
Model versions are pinned. Nothing changes underneath you because a vendor shipped an update on a Tuesday.

Data minimisation
You choose exactly which sources are indexed. Nothing is ingested by default.

We will complete your standard security questionnaire and join a call with whoever owns risk before you sign anything.

WHY US

Why BI Consulting Services

AI on top of a data foundation that was built properly. Most failed AI projects are actually failed data projects.

1

We are a data firm first

Since 2018 we have built the data platforms — Power BI, Azure, SQL, SharePoint — that AI has to sit on. We know what your data looks like before we point a model at it.

2

Microsoft Partner, MBE certified

Aligned to the Microsoft stack you already run, and a certified Minority Business Enterprise for organisations with supplier diversity requirements.

3

We build, we don't resell

Solution architects, developers, and project management in-house. You get engineers who will still be on the call in month nine.

4

We measure instead of promising

We build an evaluation set from your real questions and show you scored results. If it does not clear the bar, we say so.

5

We teach this for a living

Our team publishes Microsoft data-stack training to an audience of over 100,000. Enablement is not something we bolt on at the end.

6

We build for retail

We deliver data platforms, operational dashboards, and reporting for retail and consumer brands — and our sister agency Evolve runs digital marketing and analytics for multi-location operators.

BEFORE YOU ASK

The questions we get every time

? Who owns the hardware?

You do. It is your asset, on your balance sheet, in your building. If you ever stop working with us, the server keeps running.

? What if the hardware fails?

Manufacturer warranty passes to you, and we specify enterprise support tiers where available. For critical deployments we can spec a cold spare or a second node.

? Is it as good as ChatGPT?

On questions about your own documents, it is better — because ChatGPT has never seen them. On open-ended frontier reasoning, no, and we do not pretend otherwise.

? What about hallucinations?

Grounding answers in retrieved passages with visible citations reduces them substantially, and we measure it against your own evaluation set. It does not reach zero, so we teach people to check the citation.

? Does this touch our POS or card data?

No, and it should not. We index product, policy, and vendor content — never payment data. Keeping cardholder data entirely out of scope is deliberate, and it keeps this well away from your PCI boundary.

? Will associates actually use it?

They use it when the answer is faster than asking a person — which it is, on the floor, with a customer waiting. We build the first prompt library from the questions your own associates ask most, so day one is useful rather than theoretical.

NEXT STEPS

Three conversations, then a decision

1

Discovery workshop

90 minutes, no charge

90 minutes with a store leader, someone from merchandising, and your IT contact. We map your product and policy content, your vendor agreements, and the two or three workflows worth building first.

2

Scoped proposal

Within one week

Fixed-price services scope, hardware specification and current quote, timeline, and a cost model built with your actual headcount and current AI spend.

3

Risk review

Before signature

We complete your security questionnaire and take a call with whoever owns risk and compliance. Nothing is signed until they are satisfied.

Every answer.
Every team.
Ten seconds.

BI Consulting Services
Mooresville, North Carolina

Nichole McKinney
CEO, BI Consulting Services
nichole.mck@powerbiconsultingservices.com
704-997-3533