MCP SERVER | NO CODE | CLOUD HOSTED

Vertex AI Search MCP

Ground answers in your private company knowledge.

Vertex AI Search connects your agent directly to Google's semantic search engine, allowing you to ask complex questions about vast amounts of private company data. Instead of generic answers, it grounds responses in your own documents and knowledge bases. Manage structured datasets, find specific internal policies, or get personalized product recommendations—all through natural conversation.

A+ Quality Score 100/100

enterprise-search | grounding | semantic-search | generative-ai | natural-language-processing | data-retrieval

MAINTAINED BY VINKIUS CLOUD FOR AI AGENTS · JUNE 2026
vinkius.com/mcp/vertex-ai-search

SCAN TO VIEW

The connectivity layer between AI and the world's software.

Vinkius sits between AI and every application. All communication passes through Vinkius Cloud via the Model Context Protocol (MCP) — with governance, observability, and security at every layer.

Your AI Connections Run Through Vinkius Cloud

The world's largest managed MCP catalog

Vinkius is the connectivity layer where AI connects to the software your business already runs. We handle the hosting, the security, the credentials, the uptime — you get agents that actually do things.

We operate the world's largest managed MCP catalog. Major SaaS platforms, CRMs, databases, and cloud providers — running, monitored, production-ready. This MCP server is hosted and maintained by the Vinkius Cloud for AI Agents.

> *The agent doesn't manage credentials, doesn't manage uptime, doesn't manage security. Vinkius does.*
>
> — Architecture principle

Four Pillars of the Vinkius Runtime

01 — Security by design

Credentials stay encrypted at rest via AES-256. The AI agent never touches raw keys — they're injected into a sandboxed V8 isolate at runtime. Actions are logged, and connections have an emergency kill switch.

02 — Built on MCP Fusion

This MCP server was built with **MCP Fusion**, the open-source framework (Apache 2.0) that powers the entire Vinkius catalog. Schema-as-firewall strips undeclared fields, compiled PII redaction runs at zero overhead, and cryptographic lockfiles produce git-diffable audit trails.

03 — Deterministic observability

Eight immutable metrics per endpoint: request volume, p95 latency, error rate, active connections, cost attribution. A live payload feed logs every tool call with mutation detection.

04 — Autonomous operations

Servers are deployed, monitored, and patched autonomously. New capabilities and security patches ship weekly. Zero-downtime deployments ensure continuous availability across all managed MCP servers.

AES-256	Ed25519	24h TTL	V8 Isolate
Encryption at rest	PKI vault signatures	Ephemeral session keys	Sandboxed execution

One Token. Instant Access.

Every MCP server on Vinkius is accessed through a **Connection Token**. Tokens are generated in the cloud dashboard and produce a unique MCP endpoint URL. Paste this URL into any MCP-compatible client — no SDK required.

A single token can serve **multiple AI clients simultaneously**, or you can issue separate tokens per client for granular access control. Each token tracks its own request count, last activity timestamp, and can be individually enabled or revoked.

MCP ENDPOINT

```
https://edge.vinkius.com/{token}/mcp
```

Claude · Cursor · VS Code · Windsurf · Grok · Gemini

Security Is the Architecture

Security in Vinkius is not a feature — it's the foundation of the runtime. The gateway enforces multiple independent protection layers between AI agents and third-party APIs.

01 — Ed25519 PKI Vault

Every workspace has an Ed25519 Master Key. Session keys are generated ephemerally (24h TTL) and signed by the Master Key. Credentials never leave the vault boundary.

02 — V8 Isolate Sandboxing

Tool code runs inside isolated-vm V8 isolates with 64 MB memory caps and per-request timeouts. No filesystem access, no network access except through the SSRF-guarded fetch bridge.

03 — SSRF Guard

All outbound HTTP requests are DNS-resolved and validated before execution. Private IP ranges (10.x, 172.16-31.x, 192.168.x, AWS metadata 169.254.x) are blocked at the network layer.

04 — DLP & PII Redaction

A ResponseGuard pipeline intercepts every tool response. Configurable redaction patterns strip sensitive fields (emails, SSNs, card numbers) before data reaches the AI agent.

05 — Cryptographic Audit Trail

Every request is signed into a SHA-256 hash chain with Ed25519 signatures. Events form a tamper-proof, SIEM-exportable forensic record.

06 — Honeytoken Trap System

Phantom credentials are injected into isolated environments. If a honeytoken is used outside Vinkius infrastructure, the server is quarantined instantly.

Emergency Kill Switch

EU AI Act Art. 14(1) Compliant

The kill switch is an **emergency halt** mechanism — not a simple toggle. When triggered, it executes three actions atomically:

01 — Server deactivated

The MCP server is immediately taken offline across the entire cluster.

02 — All tokens revoked

Every connection token is invalidated. Total lockout — reconnection blocked until new tokens are issued.

03 — WebSocket connections killed

Active connections terminated via Redis pubsub broadcast. Propagates to every runtime node in the cluster.

Full Visibility. Zero Guesswork.

The Vinkius cloud dashboard includes a full MCP Governance suite — real-time analytics and security controls for production AI operations.

Control Plane

KPI dashboard with request volume, latency, success rate, token consumption, and AI-generated operational briefings.

Agent Activity

Which AI clients are connecting, how often, and what they're doing — real-time session tracking.

FinOps

Cost tracking per tool, payload compression savings, budget optimization signals, and consumption trends.

Tool Health

Slowest and most error-prone tools, with actionable root-cause insights and performance baselines.

Firewall & DLP

PII redaction activity, sensitive data protection counters, and security event timeline.

Incident Log

Error trends, failure rates, status-code breakdowns, and forensic audit trail access.

> Get started at **cloud.vinkius.com** — connect your AI agent in under 60 seconds.

Vertex AI Search MCP

7 tools available
Cloud-hosted on Vinkius

This MCP lets your agent read and reason over your enterprise documentation like a human expert does. You connect it to any compatible AI client, and suddenly, your agent can stop hallucinating and start answering based on facts pulled from your own data stores. Need to know the current PTO policy? Or what the specs are for Product X? Instead of manually digging through shared drives or outdated wikis, you just ask your agent a question in plain language, and it pulls a direct, verifiable answer grounded in your internal documents. When you connect this MCP via Vinkius, you give your agent an entire knowledge layer built from scratch. You can even use the tool to list all available data sources so your agent knows exactly what information is accessible, making complex searches simple and repeatable.

Core Capabilities

01 — Ask questions using private documents

It generates a natural language answer by retrieving and citing specific passages from your designated company documents.

02 — Identify available data sources

You can list every searchable dataset or document collection you have configured within Google Cloud.

03 — Review data source configurations

It retrieves specific metadata and setup details for any given data store, letting you check its status.

04 — Search across documents by query

You perform a general search query against all indexed content within a specified document repository.

05 — Discover specific files and branches

It lists every individual file or branch contained inside a target data store, helping you pinpoint sources of information.

06 — Get personalized product suggestions

The agent retrieves recommendations by analyzing user interaction patterns against a specific dataset.

One Click on Vinkius — From Prompt to Execution

Available at **vinkius.com/mcp/vertex-ai-search** — connect your AI agent in three steps.

01 Subscribe to this MCP and provide your Google Cloud Project ID, Location, and Access Token.

02 Your AI client authenticates the connection and establishes access to all your enterprise data stores.

03 You prompt your agent with a question or command (e.g., 'What is our remote work policy?'), and it uses its tools to retrieve and format an answer based solely on your documents.

The bottom line is that you get reliable, fact-checked answers drawn directly from the knowledge base you already own.

Built For

Knowledge managers and technical writers who are constantly drowning in documents need this. If your team spends half a day just figuring out 'where to look' for an answer, this MCP is for you. It turns scattered documentation into one reliable source of truth.

Technical Writer

Using the system to list all search engines and data stores helps them map out how different company knowledge bases are structured.

Customer Support Lead

They use it to instantly get grounded answers about complex product rules or policies without having to guess which internal wiki page is correct.

Data Analyst

Analysts can list all data stores and then check the metadata for each one, ensuring they are querying the most up-to-date source of truth.

What Changes When You Connect

01 You eliminate guesswork. Instead of getting a generic answer from an LLM that might be wrong, you use the `get_grounded_answer` tool to ensure every piece of information comes directly from your verified internal documents.

02 Manage complex sources easily. Use `list_data_stores` and `get_datastore_details` to see exactly what datasets exist before you start querying them, saving time on failed searches.

03 Go deeper than keywords. The MCP performs semantic search across all content, meaning you don't have to know the exact terminology; just ask the question naturally.

04 Pinpoint sources of truth. If a document is misfiled or outdated, use `list_datastore_documents` to browse and see every indexed file inside a data store branch.

05 Understand user patterns. The `get_recommendations` tool lets your agent act like a personalized assistant by suggesting items based on past interactions.

06 View the full scope of search capabilities using `list_search_engines`, giving you an overview of all business-specific applications configured for searching.

Real-World Applications

The HR team needs to update policy documentation.

A manager asks their agent, 'What is the new parental leave policy?' The agent uses `get_grounded_answer` on the HR data store and replies with a direct quote and citation from the correct document version.

Product teams are launching a new feature.

An engineer asks, 'What features should we highlight for customers who bought Product A?' The agent uses `get_recommendations` on the product catalog data store and suggests related accessories or upgrades.

A support rep needs to find a specific error code.

The rep asks, 'Search for all mentions of error code 404b in our technical manual.' The agent uses `search_documents` and returns the exact document sections where that code is discussed.

Data architects need to audit data sources.

An architect asks, 'List all operational data stores.' The agent responds using `list_data_stores`, giving them a complete map of every available knowledge source for auditing purposes.

Patterns to Avoid

Assuming the LLM knows internal rules

✗ AVOID

Asking ChatGPT, 'How many vacation days do I have?' and getting a vague answer like, 'Check your HR portal.'

✓ INSTEAD

Instead, use this MCP. Ask your agent directly through `get_grounded_answer` to get the specific policy details from the designated HR data store.

Copy-pasting large documents for context

✗ AVOID

Pasting a 50-page PDF into an agent and hoping it remembers one small detail.

✓ INSTEAD

Use `search_documents` with the specific query text. The MCP searches across all indexed sources efficiently, without you having to feed it massive blocks of text.

Only searching by keyword

✗ AVOID

Searching for 'best phone' and getting results about phones that are technically correct but not what the user actually needs.

✓ INSTEAD

Use the MCP's semantic search. It understands the intent behind your query, giving you contextually relevant answers even if you don't use the exact keywords from the document.

The Right Fit

You should use this MCP when your core need is factual retrieval from a known, private corpus of documents. If you are an employee who needs to know 'What does Company Policy X say about Y?', this is exactly what you want. It's superior to general-purpose AI because it forces grounding in verifiable data.

Don't use this if your goal is creative generation—if you need the agent to write a poem, brainstorm product names, or draft an

entirely new contract from scratch, this isn't the tool. For those tasks, you need pure generative models without the knowledge base constraint. If your task involves complex multi-step coding logic, use a dedicated workflow automation MCP instead.

The Pain of Hunting for Answers in Corporate Documentation

Today, finding a simple answer means opening five different tabs: the HR wiki, the Product Spec sheet, the Legal guidelines, and maybe an old Confluence page. You copy a phrase here, paste it there, hoping you haven't missed a crucial detail or conflicting policy buried three clicks deep.

With this MCP, your agent handles that messy process for you. You ask one question, like 'What is the required lead time for international shipping?', and instead of giving you five links to click through, it gives you the single, definitive answer grounded in the correct, up-to-date source.

Vertex AI Search MCP: Grounding Answers in Your Private Data

You stop manually cross-referencing documents or guessing which data store holds the truth. The system handles the complex task of listing all available data stores and identifying the most relevant sources for your specific query.

The result is a reliable knowledge layer that behaves like an expert teammate who has read every manual, policy, and spec sheet in the company—instantly.

Vertex AI Search: 7 Tools for Knowledge Retrieval

These seven tools let your agent list, check, search, and retrieve highly specific information across all of your connected data sources.

#	TOOL	DESCRIPTION
01	`get_grounded_answer`	Generates an answer in natural language using only information from your private documents.
02	`get_datastore_details`	Pulls the setup configuration and technical details for a specific data store.
03	`list_data_stores`	Lists all searchable document collections available in your Google Cloud project.
04	`list_datastore_documents`	Shows every indexed file or branch within a specified data store for review.
05	`list_search_engines`	Retrieves a list of all high-level search applications configured in the collection.
06	`get_recommendations`	Analyzes user behavior data to suggest relevant items or next steps for the user.
07	`search_documents`	Executes a general text search query across all documents in a specific repository.

See It in Action

Real prompts you can use once this MCP is connected to your AI agent through Vinkius Cloud.

U List all my available data stores in Vertex AI Search.

I found 3 data stores: 'documentation-v1' (ID: doc-123), 'hr-policies' (ID: hr-456), and 'product-catalog' (ID: prod-789). Which one would you like to search through?

U Based on our documentation, what is our remote work policy?

Grounded in the 'hr-policies' data store: Our remote work policy allows for up to 3 days of work from home per week, provided there is prior alignment with the team manager. Employees must ensure a stable internet connection and maintain core working hours. Would you like me to pull the full document?

U Search the product catalog for 'blue wireless headphones'.

I found several matches in 'product-catalog': 1. 'CloudBass Pro Blue' (In Stock), 2. 'SkyBuds v2 Azure' (Limited Stock), and 3. 'Oceanic Beats Wireless' (Discontinued). I can provide more details on features or pricing for any of these.

Frequently Asked Questions

01 **How does Vertex AI Search MCP handle conflicting policies?**

The agent is designed to prioritize grounded answers from your specific data stores. If conflicts exist across sources, it presents the findings and cites the source for you to resolve.

02 **Do I need to pay extra for list_data_stores?**

No. Listing all available data stores is a foundational capability of this MCP and helps you map your entire knowledge footprint before you start querying.

03 **Can Vertex AI Search MCP search live websites?**

This MCP searches within the private documents and configured data stores you connect. It is not designed for general, real-time web crawling.

04 **What if I want to find all mentions of a product ID?**

You can use `search_documents` by providing the data store ID and the specific product ID query. This is far more effective than general searching.

05 **How do I know what documents are available before connecting?**

Use the `list_data_stores` tool first. It gives you a complete catalog of all searchable datasets, allowing you to understand your data scope.

Go Live in 60 Seconds

Get your connection token from **cloud.vinkius.com**, then paste the endpoint URL into any MCP-compatible client.

YOUR MCP ENDPOINT

```
https://edge.vinkius.com/[TOKEN]/mcp
```

CLIENT	WHERE TO CONFIGURE
Claude AI	Profile → Customize → Connectors → "+" → Add custom connector → Paste endpoint
Cursor	Settings → Features → MCP Servers → "+ Add New MCP Server" → Type: SSE → Paste endpoint
VS Code	Ctrl/Cmd+Shift+P → "MCP: Add Server" → add `"vertex-ai-search": { "url": "…" }`
Windsurf	MCP Settings → `mcp_settings.json` → Add endpoint URL
ChatGPT	Settings → Tools & plugins → Add MCP server → Paste endpoint
Gemini	Extensions → Add MCP Server → Paste endpoint URL

ASK AN AI ABOUT THIS

Let your preferred AI explain this MCP server

- Ask ChatGPT
- Ask Claude
- Ask Perplexity
- Ask Gemini
- Ask Grok

READY TO CONNECT

Vertex AI Search is live on Vinkius Cloud.

Get your connection token, paste it into your AI agent, and start building. No SDK. No deployment. Just results.

Start at cloud.vinkius.com →

vinkius.com · support@vinkius.com

INDEPENDENT PLATFORM DISCLAIMER

Vinkius is an independent platform and is not affiliated with, endorsed by, sponsored by, verified by, or otherwise authorized by Vertex AI Search. All third-party trademarks, logos, and brand names are the property of their respective owners. Their use in this document is strictly for informational purposes to identify service compatibility and interoperability.

DOCUMENT INFORMATION

Generated	June 2026
MCP Server	Vertex AI Search MCP
Server ID	`019d761c-1b01-70cb-b8ba-124cc3ce7604`
Platform	Vinkius Cloud for AI Agents
Endpoint	`https://edge.vinkius.com/{token}/mcp`

LICENSE & USAGE

This document is generated automatically by the Vinkius PDF Engine. Content reflects the MCP server configuration at the time of generation and may change as updates are deployed. For the most current information, visit vinkius.com/mcp/vertex-ai-search.