



Solution Profile

Agentic AI Solution for Enterprise Summarization and Search

Power Your AI

GPU-enabled compute nodes deliver the processing power required for real-time large language model (LLM) inference, intent detection and agentic workflow execution

Derive Insight

Retrieve, correlate and synthesize insight across documents, images, video, audio, logs, records and machine-generated signals

Speed Time to Value

Validated architectures remove the guesswork and eliminate the piecemeal approach to summarization and search of diverse datasets

Enterprises need to enable fast, holistic Interpretation of diverse data.

Rapid accumulation of video, images, audio, documents, logs and other machine-generated data creates a need to search, correlate and summarize information across both structured and unstructured data at scale. However, most multimodal AI initiatives find it difficult to meaningfully deliver the insights hidden in the data due to fragmented architectures, disconnected data pipelines, unclear business alignment and operational risk. The outcome is a set of isolated demonstrations rather than an enterprise-grade, validated and scalable program, which can confidently deliver a capability to extract insights across multimodal dataset.

Deliver Real-Time Inference, Intent Detection and Agentic Workflows

Modern enterprises across all industries are awash in data from diverse sources. Video streams, industrial and robotics sensors, user devices and digital systems generate petabytes of videos, images, audio logs, text documents and machine data every day. The challenge is no longer capturing and storing the information. It is quickly and holistically interpreting it across all these forms. Human analysts and siloed tools struggle to piece together context from separate data streams. As a result, critical patterns or incidents can go undetected until it's too late.

Despite the maturity of AI technologies, enterprises face persistent challenges that hinder successful adoption. Many multi-modal search and summarization (MMSS) solutions are stitched together from point tools (LLMs, vector databases, APIs and storage systems) without a cohesive, validated architecture. This leads to unpredictable performance and operational fragility.

The solution to these enterprise challenges is grounded in a validated, purpose-built architecture. At its core, GPU-enabled compute nodes deliver the processing power required for real-time LLM inference, intent detection and agentic workflow execution — whether that means resolving a field service query for a utilities operator, routing a maintenance request on a manufacturing floor, or detecting a hazard on a rail to prevent mishaps and subsequent delays on a rail network.

Hitachi Vantara offers an enterprise-grade platform for real-time multi-modal summarization and search across diverse data types, leveraging accelerated compute and scalable storage to accelerate AI deployment and insight extraction. The pre-validated architecture eliminates guesswork and overcomes fragmented architectures, disconnected pipelines, and operational risk that hinder AI adoption enabling faster time to value with reliable AI deployment.

Solution Components: Hardware

- Hitachi iQ: An enterprise AI solution stack, the Hitachi end-to-end AI/ML solutions platform combines high-performance infrastructure and integrated AI software to accelerate enterprise AI deployments.
- Hitachi Virtual Storage Platform One Block (VSP One Block): High-performance block storage, this GPU-optimized block storage appliance provides ultra-low latency and high IOPS for mission-critical AI/ML workloads. It ensures fast retrieval of data and vector indexes needed for multimodal search, enabling the solution to instantly access operational knowledge (for example, maintenance logs or sensor readings) with minimal delay.
- VSP One Object: Scalable storage, this next-generation object storage system is designed to handle massive volumes of unstructured data (videos, images, audio, logs) at petabyte to exabyte scale. VSP One Object transforms object storage from a passive data archive into an active data foundation for AI and analytics, with intelligent data services like in-place search and query (via native Amazon S3® and metadata integration) to support multimodal retrieval across diverse content.
- Accelerated compute: GPU-enabled nodes, high-performance computing servers with NVIDIA® GPU acceleration, deliver the computational power necessary for real-time AI inference and complex multimedia processing.
- High-speed network: Unified fabric, a high-bandwidth Cisco® networking infrastructure (that is, advanced switching fabric), interconnects all MMSS components — compute nodes, storage systems and services — with low latency and massive throughput.

Solution Components: Software

- Hitachi iQ Studio: This turnkey AI integration and orchestration layer for the solution enables rapid composition of agentic workflows, retrieval pipelines and generative AI services. It offers a low-code environment with reusable blueprints to accelerate production deployment while maintaining enterprise-grade governance, observability and integration with underlying Hitachi infrastructure.
- NVIDIA software: NVIDIA AI Enterprise™ (including CUDA®, TensorRT® and Triton Inference Server™) accelerates multimodal model inference, vector processing and large-scale AI workloads on GPU-enabled infrastructure.
- Red Hat Enterprise Linux® (RHEL): This enterprise OS platform, a stable, secure Linux operating system, provides the foundation for the solution's servers and software components. RHEL offers enterprise-grade reliability, security enhancements and performance optimizations for AI workloads.
- Red Hat OpenShift®: Providing container orchestration and management, this Kubernetes®-based container platform is used to deploy and manage the MMSS microservices and AI components at scale.

Together, these components form a cohesive, enterprise-ready infrastructure foundation. The solution supports the full conversational and agentic AI stack, from real-time user interactions, search and summarization, through to data retrieval, model inference and workflow orchestration. It delivers measurable outcomes across some of the most demanding operational environments.

Use Cases: The Art of the Possible

Hitachi Vantara with its unique pedigree in industrial engineering combined with deep expertise in building data center solutions is uniquely positioned to guide customers towards successful implementation of multimodal search and summarization in industries such as:

Finance

- *Regulatory compliance and trade surveillance:* Financial institutions must continuously monitor for regulatory compliance, insider trading and other misconduct. An MMSS-powered compliance assistant can ingest and analyze diverse data — including emails, chat logs, recorded phone calls (transcribed audio) and trading records (structured data) — to spot potential policy violations or market abuse.
- *360° customer support assistant:* Large financial services firms often interact with customers across many channels — emails, chat, call centers and in-branch meetings — generating a wealth of multimodal data about each client. An MMSS customer support assistant can compile a holistic customer profile by retrieving long document and account statements along with call recordings. Using this profile, a bank's help desk can present a unified summary of the customer interactions, transactions and inquiries.
- *Investment research and risk intelligence:* Portfolio managers and analysts in finance need to stay on top of vast amounts of information — from market data to news and analyst reports. An MMSS agent can act as a market intelligence assistant, capable of combining textual data, numerical data and even multimedia information from news sources and research feeds to inform investment decisions.

Healthcare

- *Health monitoring and early warning:* In hospital operations, an MMSS agent can function as an AI-driven patient monitoring system that analyzes streaming data from bedside devices (heart rate and oxygen monitors), nurse notes and even patient-wearable sensor data, to detect early signs of patient deterioration.
- *Clinical trial and research insight aggregator:* Keeping up with the latest medical research and finding patients for clinical trials is a data-intensive task that spans research papers, patient records and registries. An MMSS assistant can serve as a research companion for clinicians or medical researchers by searching across scientific literature, trial databases and patient datasets.

- *Patient clinical summary and decision support:* Healthcare providers often need a holistic view of a patient's history across multiple data types. An MMSS-enabled clinical assistant can engage in multiturn conversations with doctors or nurses to provide on-demand patient summaries and support diagnosis.

Manufacturing

- *Real-time production floor assistant:* On the factory floor, line operators often need quick answers from manuals, schedules or IoT dashboards while keeping their hands free. An MMSS-powered shop floor assistant allows operators to ask questions via a voice interface (audio input) and receive immediate answers by combining data from various sources.
- *Predictive maintenance and anomaly detection:* Modern manufacturing equipment is equipped with sensors and cameras that generate massive data streams. An MMSS solution can act as an intelligent maintenance advisor, correlating sensor logs, equipment images and technical manuals to predict failures and suggest fixes.
- *Visual quality control and compliance:* Ensuring product quality and compliance with standards is a constant challenge. An MMSS system can ingest images from quality inspection cameras, sensor data from testing equipment and production records to identify defects or process deviations.

Extract Critical Business Insights

Multimodal summarization and search across documents, images, video, audio, logs, records and machine-generated data helps quickly interpret data, identify patterns and act with confidence. The Hitachi iQ solution helps by providing a validated, enterprise-ready foundation that combines GPU-enabled compute, scalable storage, high-speed networking and integrated AI orchestration to retrieve, correlate and summarize diverse data at scale. The result is faster time to value, more reliable AI deployment and actionable insight across mission-critical environments.

[Learn More](#)

About Hitachi Vantara

Hitachi Vantara is transforming the way data fuels innovation. A wholly owned subsidiary of Hitachi, Ltd., we're the data foundation the world's leading innovators rely on. Through data storage, infrastructure systems, cloud management and digital expertise, we build the foundation for sustainable business growth.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
hitachivantara.com | community.hitachivantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
hitachivantara.com/contact

© Hitachi Vantara LLC 2026. All Rights Reserved. All other trademarks, service marks and company names are properties of their respective owners.

NVIDIA, CUDA and TensorRT are registered trademarks and NVIDIA AI Enterprise and Triton Inference Server™ are trademarks of NVIDIA Corporation in the United States and other countries. Cisco is a registered trademark of Cisco, Inc. Red Hat Enterprise Linux, Red Hat and OpenShift are registered trademarks of Red Hat, Inc. Kubernetes is a registered trademark of the Linux Foundation. Amazon S3 is a registered trademark of Amazon Technologies, LLC.

HV-BTD-SP-Agent-ic-AI-Solution-for-Enterprise-Summarization-and-Search-29June26-A