VULTR

USE CASE

AI-Powered Customer Service Agent with Document and Media Processing

Vultr Cloud GPUs, NVIDIA AI Blueprints, and NetApp storage power AI customer service agents to process and analyze structured and unstructured knowledge bases to provide real-time support, enabling the foundation towards agentic AI with improved decision-making through autonomous, insight-driven actions.

vultr.com

VULTR

Use Case

AI-Driven Content Processing to Improve Service Quality

Hunting for information delays customer interactions. AI-powered agents on Vultr Cloud GPUs enable fast, accurate search across PDFs, images, text, and videos, with agentic AI optimizing retrieval for faster support.

AI-powered customer service agents help organizations scale operations more effectively by enabling more efficient utilization of resources by processing various file types, including PDFs, images, and video recordings, that contain reports, policies, contracts, and other essential documents for use in customer service roles. While these formats present information in a structured way, they combine multiple structured and unstructured elements such as text, tables, charts, and visuals, making manual search and retrieval difficult. Their fixed layout forces employees to manually scan internal KBs with lengthy documents or recordings, which slows response times, reduces efficiency, and increases errors. With agentic AI on Vultr Cloud, workflows are executed autonomously and optimized as documents update, enabling businesses to deliver faster, more accurate customer support while streamlining operations.

Enhancing customer service with AI

Vultr, NVIDIA, and NetApp have partnered to improve customer service with AI-driven document processing.

Vultr provides high-performance compute, storage and cloud GPU for scalable and efficient data handling.

The NVIDIA AI Blueprint for retrieval-augmented generation (RAG), built with NVIDIA NeMo Retriever and NIM, can extract and analyze text, images, tables, and video content for precise search and retrieval.

NetApp enhances data management with optimized data movement and access while delivering compliance-assured, secure storage.

Together, this solution reduces manual effort, improves response times, and helps businesses deliver faster, accurate customer support by equipping customer service agents with autonomy to route tasks, integrate APIs, and optimize operations.

Key challenges and solutions

Manual search inefficiencies

Customer support teams spend too much time manually searching through large volumes of PDFs, images, and videos, leading to slow responses and inconsistent answers. AI-driven search with NVIDIA NeMo on Vultr Cloud GPUs, enables the foundation of agentic AI, automates retrieval and multi-step search operations for instant, accurate responses and improved efficiency.

High costs and ROI uncertainty

Traditional systems make processing various content formats costly and inefficient. Vultr's predictable pricing and compute costs—50%-90% lower than hyperscalers—ensure scalable AI processing without unpredictable expenses. AI-powered retrieval with Qdrant, NVIDIA AI Blueprints, and NetApp optimizes embedding, retrieval, and storage of multimodal content, while AI agents scale operations to reduce costs, boost performance, ensure compliance, and deliver measurable ROI.

Latency in AI processing

Legacy systems struggle with real-time AI inferencing, delaying customer interactions. Vultr Cloud GPUs, accelerated by NVIDIA, ensure high-speed AI processing, while NetApp's intelligent data management optimizes performance. Independent AI agents enhance this with autonomous decision-making for rapid adaptability and response enabling the foundation of agentic AI.

Implementation challenges

Deploying AI for multimodal content processing requires scalable infrastructure, efficient data management, and seamless integration. Businesses face challenges with costs, compliance, and workflow adaptation. With Vultr, NVIDIA, and NetApp, organizations get a flexible approach and support to implement AI-driven document and video processing effectively.

How it works

When a customer uploads a document, image, or video, an AI agent can instantly access key insights extracted from the file. Behind the scenes, Vultr's high-performance storage and compute resources process the content, enabling AI-driven search and retrieval. For PDFs and text-based documents, the system parses structured and unstructured data, identifying key text, tables, charts, and images to surface relevant details. For videos, the AI extracts metadata, generates speech-to-text transcriptions, and analyzes visual elements, converting content into vector representations that enhance searchability and provide contextual responses.

Meanwhile, AI agents autonomously select the best processing path, invoking tools and APIs as needed while refining retrieval strategies.

The extracted content is converted into vector representations and stored in Qdrant, a high-performance vector database. The Qdrant instance runs on NetApp ONTAP data volumes, ensuring performance, high availability, seamless data management, optimized data access with FlexCache and easy cloning for development and testing with FlexClones.

Using the NVIDIA AI Blueprint for Multimodal Content Processing, accelerated by NVIDIA NeMo Microservices (NIMs) on Vultr Cloud GPUs, the system processes and extracts relevant data. Uploaded files are structured into vector embeddings, enabling contextual understanding and semantic search across documents, images, and videos. For videos, this includes automated speech recognition (ASR), scene detection, and content tagging, making it easier to find insights, summarize information, and generate relevant responses based on video content.

NVIDIA NeMo Microservice for Deepseek R1 671B, enables AI-driven understanding and retrieval, ensuring fast and precise responses. While Qdrant and Deepseek R1 work together to build a real-time Retrieval-Augmented Generation (RAG) pipeline, ensuring that any added or deleted content is immediately updated in search results.

When a customer service agent queries a document, image, or video, Vultr converts the request into an embedding for similarity matching in Qdrant. NVIDIA NeMo reranking models refine search results, and Vultr Cloud GPUs, accelerated by NVIDIA HGX™ B200, perform real-time inference, generating structured responses instantly. This allows support teams to efficiently retrieve key information from transcripts, metadata, and insights, providing relevant, context-aware answers to customer inquiries.

NetApp ONTAP data management simplifies cloning, allowing businesses to create separate environments for development, testing, and production. NetApp snapshots enable compliance-assured backups, securing information while maintaining regulatory adherence with near instant data protection and recovery, and traceability. Further, NetApp enables efficient, and cost effective data movement into and out of Vultr cloud from geographically dispersed regions.

Integrating NVIDIA AI Blueprints on Vultr with NetApp's intelligent data management delivers a scalable, AI-driven multimodal intelligent content workflow, improving search accuracy, automation, and enhancing efficiency for customer service agents by reducing the time to respond to customer inquiries.

Why it's important

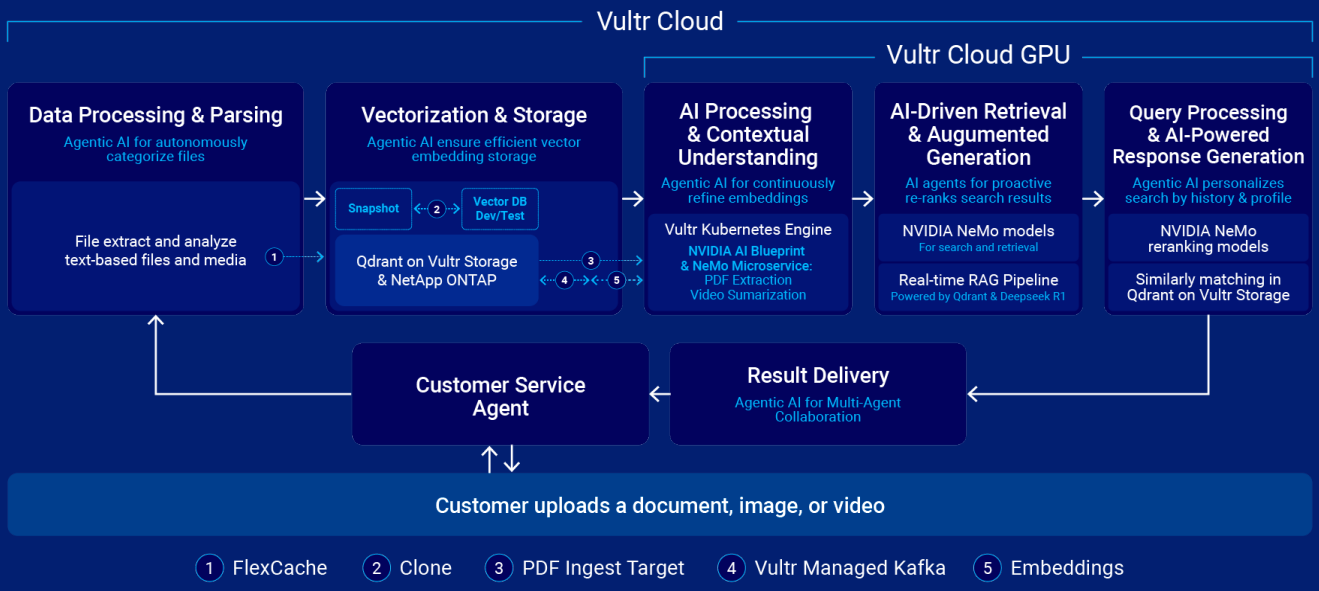
This is essential for businesses managing large volumes of content, enabling fast, accurate AI-powered extraction and retrieval. With Vultr's high-performance infrastructure, NVIDIA AI Blueprints, and NetApp's intelligent data management, organizations can efficiently process and search complex files.

This improves automation, enhances support, ensures compliance, and secures critical information, reducing manual effort and enabling real-time decision-making.

Benefits

- **Faster query resolution:** AI reduces agent workload and speeds up responses.
- **Enhanced compliance and data security:** Consistent snapshots ensure compliance across industry regulations.
- **Optimized AI model performance:** Secure integration for continuous model improvement.
- **Empowered AI agents:** Autonomous AI handles tasks, integrates systems, and optimizes in real time.

AI-Powered Customer Service Agent



Industry applications

Telecommunications

Managing service contracts – AI scans customer agreements, billing statements, and support documents, along with recorded service calls, to provide instant responses to plan, pricing, and policy inquiries.

Financial services

Client advisory and risk assessment – AI extracts key insights from financial documents, reports, and video briefings, enabling faster risk analysis and data-driven recommendations.

Energy

Interpreting compliance reports – AI extracts critical information from regulatory and safety documents, as well as visual inspections and recorded assessments, ensuring compliance teams quickly access accurate details.

Mobility and broadband

Optimizing network support – AI analyzes technical manuals, policies, support tickets, and network performance videos to streamline troubleshooting and improve broadband and mobility services.

High-frequency trading

Analyzing market data in real-time – AI scans PDFs of earnings reports, analyst forecasts, and economic briefs, ensuring traders quickly access critical insights for fast decision-making.

Retail

Handling supplier invoices and orders – AI processes purchase orders, invoices, and inventory records, enabling seamless vendor communication and faster order fulfillment.

Media and entertainment

Enhancing streaming and content discovery – AI automates metadata tagging, content indexing, and subtitle analysis to improve searchability, optimize recommendations, and enhance viewer engagement.

Healthcare and life sciences

Processing insurance claims – AI automates data extraction from claims, medical records, and policy documents, reducing manual effort and speeding up approvals.

Public sector and government

AI extracts key details from legislation, public records, and policy documents, including recorded proceedings, enabling faster decision-making and efficient response to citizen inquiries.

Learn more about documents and media processing

Demo: Simplifying International Call Charges with AI →

Read our white paper on Agentic AI and industry application examples →

Contact us at vultr.com to get started. →