How to use AI on a cloud server

AI Overview

You can run AI on a cloud server by selecting the right cloud platform, leveraging GPU resources, deploying AI models or agents, and optimizing for scalability and performance.Choosing the Right Cloud InfrastructureTo use AI effectively, start by selecting a cloud provider that supports GPU-accelerated computing, such as AWS, Google Cloud, or Azure. GPU servers are essential for AI workloads because they can execute thousands of parallel operations, drastically reducing training and inference times compared to CPU-only servers . Cloud GPU hosting allows you to scale resources on demand, pay only for what you use, and avoid the high upfront costs of on-premises hardware .Deploying AI ModelsGPU Cloud Servers: Use dedicated GPU servers for training deep learning models with frameworks like TensorFlow, PyTorch, or Keras. High-end GPUs such as NVIDIA A100 or H100 accelerate matrix operations, reducing training time and improving inference speed .Virtual Private Servers (VPS): For smaller models or CPU-based inference, a VPS can suffice. It is suitable for lightweight AI applications or testing purposes .Serverless AI: Platforms like Azure Functions or Google Cloud Run allow you to deploy AI models or agents without managing the underlying infrastructure. These services automatically scale to handle multiple concurrent requests and integrate with AI tools and external APIs .Running AI AgentsAI agents are autonomous software entities that can perform tasks, make decisions, and interact with other services. On Google Cloud Run, you can deploy AI agents to orchestrate asynchronous tasks, connect to vector databases for retrieval-augmented generation (RAG), and execute code in secure sandboxed environments . Azure Functions also supports agentic workflows, enabling function calling to dynamically invoke AI tools and interact with multiple data sources .Optimizing PerformanceMulti-GPU setups: For high workloads, use multiple GPUs to parallelize training and inference, reducing processing time .Network architecture: Keep your...

Article Content

Red Hat OpenShift | comprehensive application platform

Managed cloud services Red Hat OpenShift is available as a managed application platform on the cloud provider of your choice. Make the most of your existing

Use connectors to extend Claude''s capabilities

Connectors work across Claude, Claude Desktop, Claude Code, and the API (via the MCP Connector). You can find available connectors in the Connectors Directory, where each connector

Running Your Own LLMs in the Cloud: A Practical Guide

We will be using Ollama, a tool for running LLMs, along with cost-effective cloud providers like RunPod and vast.ai. Here''s the basic process: RunPod (runpod.io) offers a streamlined

The AI Developer Cloud | Runpod

AI infrastructure with on-demand GPUs and serverless compute. Run training, inference, and batch workloads on the cloud with Runpod.

Welcome to Channel Dive | Channel Dive

Welcome to Channel Dive. We''re Informa TechTarget''s new publication, focused on delivering daily news and analysis for executives at North

Mac Mini M4 AI Server: Local LLM + Agent Setup (2026)

Turn your Mac Mini M4 into a local AI server. Ollama for LLMs, OpenClaw for AI agents, Claude Code for dev workflows. Hardware tiers

7 Best LLM Tools To Run Models Locally (July 2026)

This desktop platform lets you download popular AI models like Llama 3, Gemma, and Mistral to run on your own computer, or connect to cloud

Azure MCP Server documentation

Azure MCP Server documentation Learn how to use the Azure MCP Server to manage Azure resources through natural language commands. Connect from GitHub Copilot, custom AI agents, and MCP

Run AI solutions on Cloud Run | Google Cloud

This guide provides an overview of using Cloud Run to host apps, run inference, and build AI workflows. Cloud Run for hosting AI applications, agents,

Tech News | Today''s Latest Technology News | Reuters

Find latest technology news from every corner of the globe at Reuters , your online source for breaking international news coverage.

Deploying AI/ML Models on the Cloud: A Practical Guide

In this article, I will walk you through the process of deploying models on the cloud, discuss different deployment strategies, and compare various cloud platforms.

Collaborative Content Management, and Secure File

Create, share, and govern trusted knowledge with Microsoft SharePoint—powering collaboration, communication, automation, and AI experiences across Microsoft

How to Deploy AI Models on GPU Servers: A Beginner-Friendly Guide

Step-by-step guide to deploying AI models on GPU servers. Improve inference speed, optimize performance, and streamline your AI workflows.

Use Atlassian Rovo MCP Server

Atlassian Rovo MCP Server is a cloud-hosted Model Context Protocol (MCP) server that gives your AI tools secure, real-time access to your data across Jira, Confluence, Bitbucket, and other Atlassian

I quantized a local LLM on my home server and ditched cloud AI

On the AI side, I built a self-hosted llama.cpp server using hardware I''ve salvaged from outdated machines, and with the right set of tweaks, I managed to build a Home Assistant hub that ...

Service Cloud: : AI-powered Customer Service Agent

Service Cloud empowers service teams to manage cases, knowledge, and incidents collaboratively from a single, AI-powered workspace so you can boost

AI/ML orchestration on Cloud Run documentation | Google Cloud

Explore our tutorials and best practices to see how Cloud Run can optimize your AI/ML workloads. Develop with our latest Generative AI models and tools. Get free usage of 20+ popular

The Home Depot and Google Cloud Launch Agentic AI

Today at NRF 2026, The Home Depot and Google Cloud announced an expansion of their strategic partnership, further advancing the interconnected

How to Host Your Own Private AI on a Dedicated

In this guide, we will walk you through the exact hardware requirements and software steps to build your own private AI server using

Hosting AI And Machine Learning Web Apps: GPU

Learn how to host AI and machine learning web apps: when to use GPU servers, VPS or hybrid cloud architectures, plus sizing, security and cost tips.

From Local Dev to Production: How to Deploy AI Models in 2025

AI models are more accessible than ever, but taking one from your local machine to production — whether on-premises or in the cloud — requires thoughtful decisions.

Windows Server Automation: Key Trends in AI, Cloud, and Security

Discover how AI, automation, and cloud integration are transforming Windows Server administration and what skills you need to stay ahead in this evolving landscape.

How to Build an Affordable Custom AI Server for AI

In this overview, Jun Yamog guides you through the essentials of building a high-performance AI server, from selecting the right GPUs to

Getting started with the Atlassian Rovo MCP Server

Getting started with the Atlassian Rovo MCP Server In browser and desktop agents, Atlassian Rovo MCP connects your AI agent to your Atlassian apps and work, all accessible in one conversation –

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: https://www.sky-connect.co.za
Email: info@sky-connect.co.za
Phone: +44 20 7946 0958
Address: 1 Cornhill, London EC3V 3ND, United Kingdom

This document is for informational purposes only. Specifications subject to change without notice.