**Job Title: Cloud Solutions Architect – Generative AI**  
**Location: [Specify Location] | Remote/Hybrid Options Available**  
**Job Type: Full-Time**

**About the Role:**

We are seeking an experienced **Cloud Solutions Architect – Generative AI** to join our team and help customers design, implement, and optimize cutting-edge AI-powered solutions on the cloud. You will be responsible for enabling customers to leverage **Generative AI models, AI/ML services, and cloud-native architectures** to build scalable, secure, and efficient AI-driven applications.

**Key Responsibilities:**

* **Architect and design** scalable, cost-effective, and high-performing **Generative AI solutions** leveraging cloud infrastructure and AI/ML services.
* **Engage with customers** to understand business challenges and translate them into AI/ML-powered solutions using **LLMs, embeddings, fine-tuning techniques, and model optimization**.
* **Develop and implement best practices** for deploying AI models, including **security, governance, and compliance**.
* **Collaborate with cross-functional teams** including data scientists, ML engineers, DevOps, and security teams to drive end-to-end AI solution deployment.
* **Optimize model performance** for real-time and batch inference workloads while ensuring low latency and high availability.
* **Provide thought leadership** on **Generative AI trends**, cloud AI services, and emerging technologies.
* **Create technical content, workshops, and proof-of-concepts** to educate customers and internal teams on AI capabilities.
* **Advise on AI ethics, bias mitigation, and responsible AI deployment** in production environments.

**Basic Qualifications:**

* **5+ years of experience** in **cloud architecture, AI/ML, or software development** roles.
* Deep knowledge of **Generative AI, Large Language Models (LLMs), and AI/ML frameworks** such as **TensorFlow, PyTorch, Hugging Face, and LangChain**.
* Hands-on experience with **AWS, Azure, or Google Cloud AI/ML services** (e.g., AWS Bedrock, Amazon SageMaker, Azure OpenAI, Vertex AI).
* Strong programming skills in **Python, Java, or Scala**, with experience in model fine-tuning, prompt engineering, and inference optimization.
* Experience deploying AI models in **serverless, Kubernetes, or containerized** environments.
* Understanding of **data engineering, MLOps, and CI/CD pipelines for AI model deployment**.
* Excellent communication and presentation skills, with the ability to explain complex AI concepts to non-technical stakeholders.

**Preferred Qualifications:**

* **Certifications in cloud architecture (AWS/Azure/GCP)** and AI/ML specializations.
* Experience integrating Generative AI into **chatbots, search, content generation, or automation workflows**.
* Familiarity with **vector databases, embeddings, and retrieval-augmented generation (RAG) techniques**.
* Background in **NLP, computer vision, or reinforcement learning** is a plus.
* Contributions to **open-source AI projects** or research in AI/ML.

**Why Join Us?**

* Work on cutting-edge **Generative AI solutions** that drive real-world impact.
* Collaborate with some of the brightest minds in AI and cloud computing.
* Access to top-tier cloud and AI technologies, certifications, and training.
* Competitive salary, stock options, and benefits package.

If you are passionate about **AI, cloud, and innovative problem-solving**, we want to hear from you!

**Apply Now** and be part of the future of AI-driven cloud solutions.