

Activity: Choosing a Hyperscaler for AI Deployment

You are an **AI Solutions Consultant**, advising a Fortune 500 retail client, **GlobalMart**, on deploying an AI-driven recommendation system. The client wants to personalize product suggestions for millions of customers across multiple regions.

Business Requirements:

- **Scalability:** Must handle seasonal spikes in online traffic (e.g., holiday sales, Black Friday).
- **Performance:** Low-latency inference is critical to ensure seamless user experience.
- **Cost Optimization:** The client is looking for a balance between infrastructure investment and operational costs.
- **Integration:** The system must integrate with their existing enterprise IT stack, which includes a mix of on-premises data centers and cloud-based solutions.
- **Compliance & Security:** The solution must adhere to stringent industry regulations (e.g., GDPR, CCPA) and enterprise security policies.

Your task is to **evaluate AWS, Azure, and GCP as potential cloud providers** and determine which one best aligns with GlobalMart’s needs.

Step 1: Review Hyperscaler Offerings

Below is a **summary of AI/ML deployment offerings from each hyperscaler**:

Cloud Provider	AI/ML Deployment Services	Key Strengths	Key Limitations
AWS	SageMaker, Lambda for inference, EKS for containerized models	Broadest set of AI/ML services, global scalability, strong security and compliance support	Pricing complexity, can be expensive for high compute workloads
Azure	Azure Machine Learning, AKS for model hosting, Functions for serverless inference	Strong enterprise integration (Active Directory, DevOps), hybrid cloud capabilities, trusted by large enterprises	Slightly less mature AI tooling compared to AWS
GCP	Vertex AI, Cloud Run, TensorFlow Serving, TPU accelerators	Best for AI-first workloads, strong in model training and scaling, competitive pricing for ML workloads	Fewer general enterprise integrations compared to AWS and Azure

Step 2: Make a Decision

Based on the **business requirements** and the information above, answer the following questions:

1. **Which cloud provider would you recommend for GlobalMart's AI deployment?**
2. **What are the top 2 reasons for your choice?**
3. **What trade-offs or risks does your choice involve?**
4. **If GlobalMart wanted a multi-cloud strategy, how would that affect your recommendation?**

Step 3: Group Discussion & Justification

- Share your selection and reasoning with the class.
- Debate alternative choices with peers.
- Discuss how different industries might prioritize different hyperscaler strengths.