Leveraging Cloud Abstractions to Drive Generative AI:

Scalability, Efficiency, and Diverse Applications

Karan Khanna

TABLE OF CONTENTS

Introduction to Gen AI

- Generative Artificial Intelligence is subset of deep learning that employs training on supervised, semi supervised or even unsupervised learnings.

- Differs from discriminative in the manner it operates i.e. it creates new content based on the data it was trained on.

- Popular examples are ChatGPT, Dall-E, Gemini and many more.

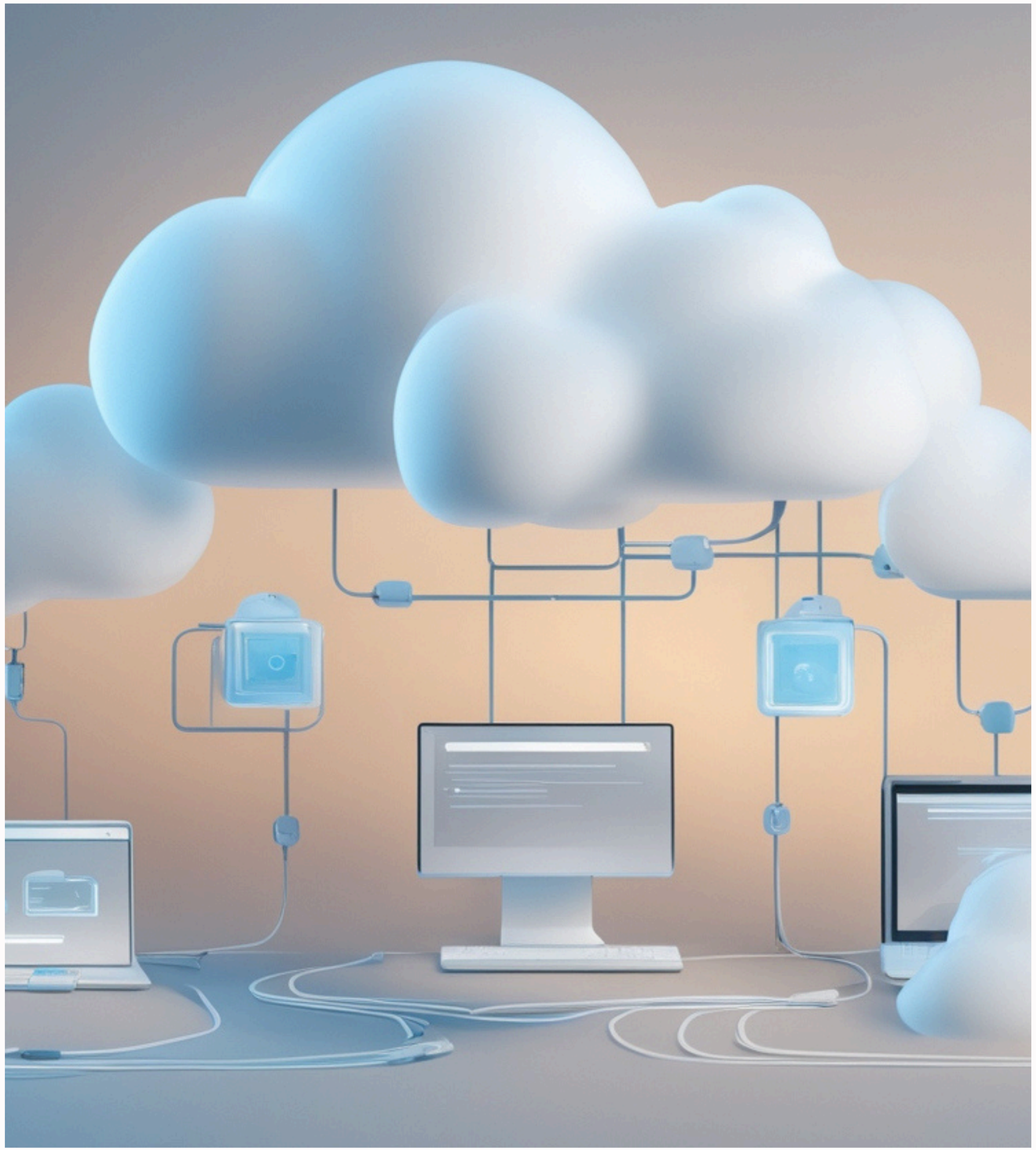
Building Blocks of Gen AI

- FMs (Foundational Models)
  - LLMs (Large language Models)
  - Trained Model
- Inference Code (Bridge between generative model & applications)
- Data Pipelines (preprocessing, formatting)
- APIs (Send & Receive)

Introduction to Cloud Abstractions

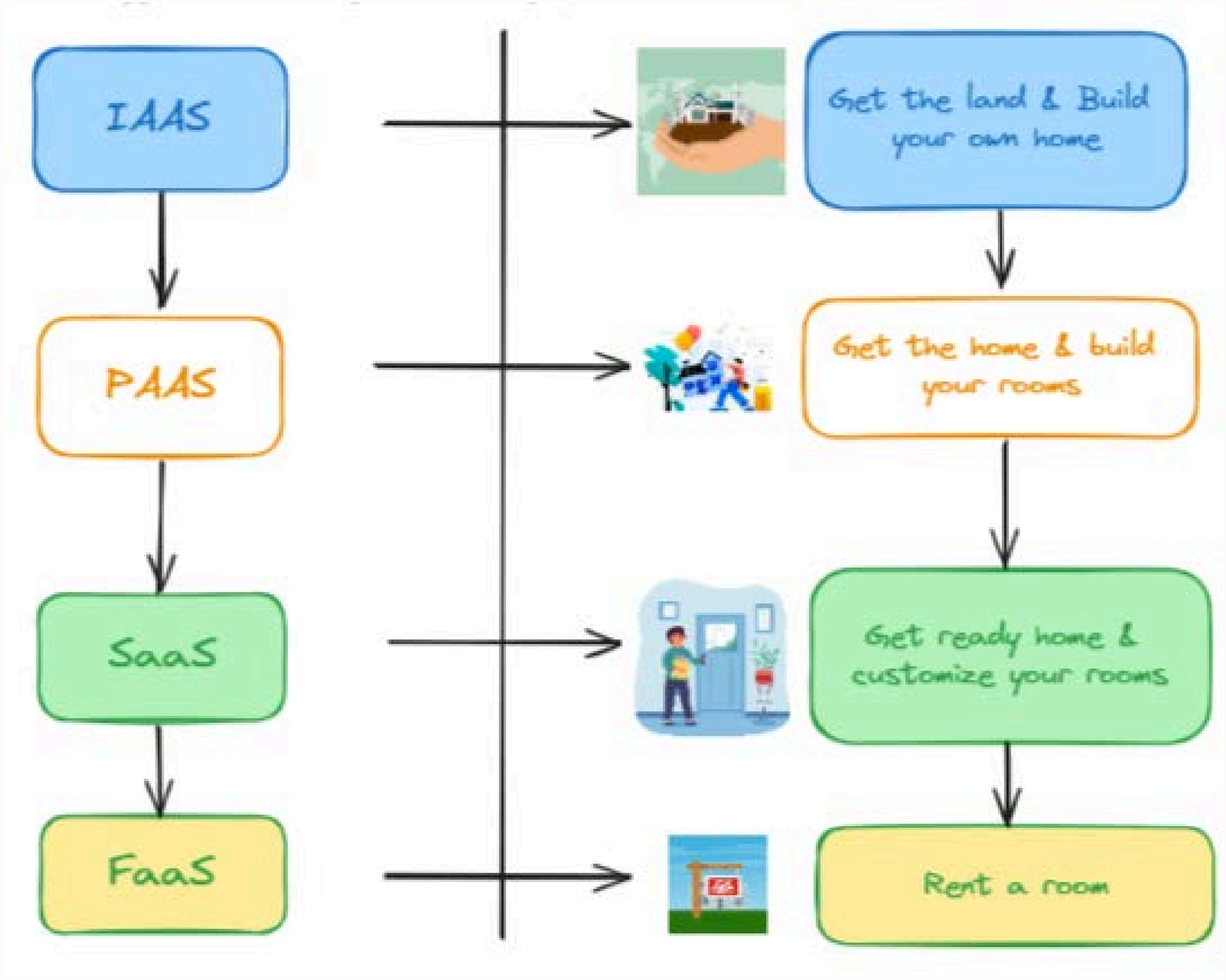
- Cloud abstractions are the mechanisms to hide internal and complex details of the infrastructure running somewhere, somehow.

- Types of Cloud Abstractions:
  - Infrastructure as a Service (IaaS)
  - Platform as a Service (PaaS)
  - Software as a Service (SaaS)
  - Function as a Service (FaaS)



Cloud Abstractions by examples

- Infrastructure as a Service (IaaS)
  - AWS EC2, Azure VM
- Platform as a Service (PaaS)
  - Heroku, Cloud Foundry
- Software as a Service (SaaS)
  - Hubspot, Salesforce, Shopify
- Function as a Service (FaaS)
  - AWS lambda, Google Cloud functions



Gen AI Deployment Story

- Model Architecture and Models
- Inference Code layer
- Data Pipelines
- Web Service or APIs
- And others:
  - Monitoring & alerting
  - Auto Scale groups
  - Load balancers
  - API gateways
  - Integration services
  - CICD
  - Regulatory & Security gates

Gen AI Deployment Strategies

- In house/ Private Deployment
  - End to end control
  - Involves managing everything on our own
  - Defense, HealthCare and numerous use cases
- Hybrid Deployment
  - Some building blocks in cloud
    - FMs
    - LLMs
    - Model Garden or VertexAI or SageMaker
  - Application code/Inference code lives in DC
    - Applies to most use cases
- SaaS Deployment
  - Everything on cloud with right controls
    - AI Studio
  - Low/No Code Deployment Model
    - Gen AI app builder

How cloud abstractions benefits Gen AI

- Efficient resource allocation & utilization
- Streamlined processes
  - CI
  - CD
  - Integrations
- Scalability

> Enables organizations to efficiently train and utilize AI models

Cloud Adoption and Kubernetes

- Cloud abstractions play a pivotal role in:
  - Operational scalability
  - Resource management

92% of organizations deploy cloud-native technologies

78% utilize Kubernetes for orchestration

Essential for managing and scaling Gen-AI workloads

Gen AI : Inspiring Industry Transformations

- Gen-AI drives transformations across various sectors:
  - Healthcare
  - Entertainment
  - Finance
  - And more
- This leads to increased innovation, efficiency, and new growth opportunities
- A diverse range of industries benefiting from Gen-AI integration
- Significant impact on operations and outcomes

Healthcare Transformation

- Gen-AI revolutionizes patient interaction and operational efficiency
- Klara: A Gen-AI tool in healthcare
  Reduces phone-based patient interactions by 50%
  Increases patient satisfaction by 30%
- This leads to improved patient outcomes and streamlined healthcare processes
- The transformative impact of Gen-AI in the healthcare sector

Creative Industries Impact

- AI-driven platforms democratize creative expression
  - Adobe Firefly
  - DALL-E

- Facilitate the generation of digital artwork
- Empower individuals to explore creativity
- Expand possibilities of artistic expression
- Making advanced AI tools accessible to a wider audience

Future Potential

- Continued evolution of Gen-AI and integration with cloud technologies
- Drives further innovation and accessibility
- The symbiotic relationship between Gen-AI and cloud abstractions
- Unlocks new possibilities and increases AI impact across industries
- Promising future for Gen-AI growth in synergy with cloud technologies

Conclusion

- A symbiotic relationship between Gen-AI and cloud abstractions
- Cloud abstractions : Essential enablers for deploying Gen-AI applications
  - Provide necessary scalability
  - Ensure efficiency
  - Support diverse applications
- Drives transformative impact across industries
- The future of Gen-AI is closely tied to the continued evolution and adoption of cloud technologies

Claudia Alves Everest Cantu

THANK YOU