Course Syllabus: AI Infrastructure Engineer

Course Title: Building the Backbone: Scalable AI Infrastructure Design

Target Audience: Ideal for engineers, system administrators, and students interested in building robust AI systems. Basic programming (e.g., Python) and familiarity with cloud computing are helpful but not required.

Course Level: Comprehensive program covering Basic, Intermediate, and Advanced levels.

Duration: 12 weeks (flexible for selfpaced learning).

Course Description:

This course trains students to become AI Infrastructure Engineers, responsible for creating and maintaining the computing environments that power AI systems, like Zomato's DynamoDB-based billing platform. You'll learn to design scalable, highperformance infrastructure using cloud platforms, manage data pipelines, and optimize AI workloads for speed and cost. By mastering tools like AWS, Kubernetes, and GPUs, you'll ensure AI systems run smoothly in production.

Learning Objectives:

Upon completion, students will be able to:

- Understand the components of AI infrastructure (compute, storage, networking).
- Design and deploy scalable cloud-based AI systems.
- Optimize AI workloads for performance and cost efficiency.
- Manage data pipelines and storage for AI training and inference.
- Ensure security and reliability in AI infrastructure.
- Build a portfolio of infrastructure projects for AI applications.

Course Structure:

Part 1: Basic Foundations (Weeks 1-4)

This section introduces AI infrastructure and cloud computing basics.

- Week 1: Introduction to AI Infrastructure
  - Role of an AI Infrastructure Engineer.
  - Components: Compute (CPUs, GPUs), storage, networking.
  - Case Study: Zomato's DynamoDB infrastructure for billing.
  - Exercise: Explore AWS or Google Cloud console.
- Week 2: Cloud Computing Basics
  - Cloud platforms: AWS, Azure, Google Cloud.
  - Key services: EC2, S3, Lambda (serverless).
  - Handson: Set up a basic cloud instance for AI tasks.
- Week 3: AI Workloads and Hardware
  - AI workload types: Training vs. inference.
  - Hardware: CPUs, GPUs, TPUs for AI.
  - Exercise: Configure a GPU instance for model training.
- Week 4: Data Storage for AI
  - Storage systems: Relational (SQL), NoSQL (e.g., DynamoDB).
  - Data formats: CSV, Parquet, JSON.
  - Hands-on Project: Build a data pipeline for a sample AI dataset.

Part 2: Intermediate Concepts (Weeks 5-8)

This section focuses on scaling and optimizing AI infrastructure.

- Week 5: Scalable Compute Systems
  - Containerization: Docker for AI applications.
  - Orchestration: Kubernetes for managing AI workloads.
  - Hands-on: Deploy an AI model using Docker.
- Week 6: Data Pipelines for AI
  - Building data pipelines: ETL (Extract, Transform, Load).
  - Tools: Apache Airflow, AWS Glue.
  - Case Study: Zomato's data pipeline for billing data.
- Week 7: Performance Optimization
  - Optimizing compute: Batch processing, parallelization.
  - Cost management: Spot instances, autoscaling.
  - Hands-on: Optimize an AI workload for speed and cost.

- Week 8: Monitoring and Reliability
  - Monitoring tools: Prometheus, CloudWatch.
  - Ensuring uptime: Load balancing, fault tolerance.
  - Hands-on Project: Set up monitoring for an AI system.

Part 3: Advanced & ExpertLevel Application (Weeks 9-12)

This section prepares students for enterprise-grade AI infrastructure.

- Week 9: Advanced Cloud Architectures
  - Serverless AI: AWS Lambda, Google Cloud Functions.
  - Distributed training: Horovod, Ray.
  - Exercise: Deploy a distributed AI training job.
- Week 10: Security in AI Infrastructure
  - Securing data: Encryption, access controls.
  - Protecting models: Model versioning, secure APIs.
  - Hands-on: Implement security for an AI pipeline.
- Week 11: Infrastructure for Emerging AI
  - Supporting multimodal AI (vision, text, audio).
  - Edge AI: Deploying models on IoT devices.
  - Exercise: Design infrastructure for a Zomato chatbot.
- Week 12: Capstone Project & Trends
  - Capstone Project: Design and deploy a scalable AI infrastructure for a food delivery platform.
  - Trends: Quantum computing, federated learning infrastructure.
  - Career paths: Cloud engineering, AI ops, infrastructure consulting.

Assignments & Grading:

- Weekly Labs & Exercises: 25%
- Intermediate Projects (Weeks 4 & 8): 30%
- Capstone Project: 35%
- Class Participation & Peer Reviews: 10%