Course Syllabus: AI Data Engineer

Course Title: AI Data Engineering: Building Data Pipelines for Intelligent Systems

Target Audience: This course is for data engineers, data scientists, and software developers with a strong foundation in SQL and Python. It's for those who want to specialize in the data infrastructure layer that supports AI and machine learning.

Course Level: Advanced Intermediate to Expert.

Duration: 12 Weeks

Course Description: This curriculum provides a hands-on, project-based path to becoming an AI Data Engineer. The course goes beyond traditional ETL to cover the specific data challenges of AI: building pipelines for unstructured data, working with vector databases for Retrieval-Augmented Generation (RAG), and ensuring data quality for model training. By the end, you will have the skills to design, build, and maintain the data systems that are the foundation of modern AI applications.

Learning Objectives

Upon successful completion of this course, students will be able to:

- Design and implement data architectures that support AI and machine learning workflows.
- Master the use of distributed computing frameworks like Spark for large-scale data processing.
- Build and manage data pipelines for both structured and unstructured data.
- Work with modern data storage solutions, including data lakes, data lakehouses, and vector databases.
- Apply MLOps principles to monitor and maintain the data quality and reliability of AI data pipelines.
- Understand the unique data needs of different AI models (e.g., LLMs, vision models).
- Develop a portfolio of data engineering projects for AI applications.

Course Structure: A Step-by-Step Learning Path

Part 1: Core Data Engineering for AI (Weeks 1-4)

This section builds the foundational data engineering skills required to handle the scale and complexity of AI data.

Week 1: Data Architecture for AI

- The modern data stack: from data warehouses to data lakes and lakehouses.
- Designing data architectures for AI: batch vs. streaming.
- Introduction to cloud data platforms (AWS, GCP, Azure).
- **Hands-on Lab:** Design a high-level data architecture for an AI-powered application.

Week 2: Advanced Data Processing with Spark

- Introduction to **Apache Spark** and its role in distributed computing.
- Using **PySpark** for data processing, cleaning, and transformation at scale.
- Spark SQL for querying massive datasets.
- **Hands-on Project:** Use PySpark to clean and transform a large, unstructured dataset.

Week 3: Data Ingestion and Pipelines

- The ETL/ELT paradigm for AI data pipelines.
- Building batch data pipelines with orchestration tools like **Apache Airflow**.
- Introduction to streaming data and real-time pipelines.
- **Hands-on Project:** Build an end-to-end data pipeline to ingest data from a source and prepare it for an AI model.

Week 4: Data Storage & Modeling

- Choosing the right data storage: S3, GCS, data warehouses, and data lakes.
- Data modeling for AI: designing schemas for easy access and model training.
- The medallion architecture (Bronze, Silver, Gold layers) for data quality.
- **Hands-on Lab:** Implement the medallion architecture on a cloud data lake.

Part 2: Specialized AI Data Engineering (Weeks 5-8)

This section dives into the unique data requirements and technologies for modern AI models, particularly generative AI.

Week 5: Unstructured Data & Feature Engineering

- Handling unstructured data: text, images, and audio.
- Feature engineering for machine learning and deep learning models.
- Working with APIs to process unstructured data.
- **Hands-on Project:** Build a pipeline that extracts text from documents and prepares it for a text classification model.

Week 6: Vector Databases & Embeddings

- What are vector databases and why are they crucial for modern AI?
- Understanding embeddings and how they represent data.
- Using a vector database (e.g., Pinecone, ChromaDB) to store and query embeddings.
- **Hands-on Lab:** Create a pipeline that generates embeddings for a dataset and stores them in a vector database.

Week 7: Data for Retrieval-Augmented Generation (RAG)

- The RAG paradigm from a data perspective.
- Building a complete data pipeline to support a RAG system.
- Strategies for chunking data and managing metadata for effective retrieval.
- **Hands-on Project:** Create a complete RAG data pipeline that ingests data and prepares it for an LLM to answer questions.

Week 8: Data Governance & MLOps

- The importance of data quality, observability, and monitoring for AI models.
- Implementing data quality checks and automated tests.
- Introduction to **MLflow** for experiment tracking and model management.
- **Hands-on Lab:** Implement data quality checks on your data pipeline and use MLflow to track model metrics.

Part 3: Advanced Concepts & Professional Practice (Weeks 9-12)

This final section focuses on the real-world deployment, maintenance, and professional skills required for an AI Data Engineer.

Week 9: Containerization & Deployment

- The role of **Docker** in creating consistent and reproducible data environments.
- Using **Kubernetes** to orchestrate and scale data pipelines.
- **Hands-on Project:** Dockerize your entire RAG data pipeline and deploy it to a container orchestration platform.

Week 10: Cloud Deployment & Scaling

- Deploying your data pipelines on a major cloud platform (AWS, GCP, or Azure).
- Optimizing data pipelines for cost and performance.
- **Hands-on Lab:** Deploy your containerized pipeline to a cloud service and set up monitoring and logging.

Week 11: Real-Time Data Pipelines

- Introduction to streaming platforms like **Apache Kafka**.
- Building a real-time data pipeline for an AI application.
- **Hands-on Project:** Build a simple streaming pipeline to process a real-time data stream for an AI dashboard.

Week 12: Final Capstone Project & Career Skills

- **Capstone Project:** Design, build, and deploy a complete AI data engineering solution from scratch.
- Building a professional portfolio and resume tailored for AI Data Engineer roles.
- Interview preparation and understanding the industry landscape.