Data Engineering

TABLE OF CONTENTS

Comprehensive Generative AI Training for Data Engineers

Duration: 20 Hrs

Training Description

This course is designed to upskill data engineers with foundational and advanced knowledge of Generative AI (GenAI) technologies. Participants will learn how to integrate GenAI into modern data engineering pipelines, handle large-scale data preprocessing for GenAI models, and deploy AI-powered workflows in scalable production environments. The training includes hands-on demonstrations, real-world use cases, and best practices for optimizing data systems to serve GenAI requirements. By the end of the course, participants will be familiar with the tools, frameworks, and methodologies needed to operationalize Generative AI for enterprise use.

Training Duration

20 Hours (Delivered over 3 days)

Target Audience

- Junior, Mid-Level, and Senior Data Engineers looking to build expertise in GenAI tools and infrastructure.
- Professionals who manage data pipelines, ETL workflows, and data lakes and seek to optimize them for AI workloads.
- Engineers responsible for deploying scalable GenAI models within cloud environments like AWS, GCP, or Azure.
- Teams interested in enabling their organizations to handle Generative AI-specific infrastructure tasks like distributed data preparation or model integrations.

Tools and Frameworks Covered

- **Data Platforms**: Apache Spark, Databricks, Delta Lake.
- **GenAI Frameworks**: Hugging Face Transformers, OpenAI APIs, LangChain.
- **Cloud Technologies**: AWS S3, GCP BigQuery, Azure Data Factory, Snowflake.
- **AI Model Serving**: TensorFlow Serving, TorchServe, FastAPI, Gradio.
- **Data Orchestration**: Apache Airflow, Prefect.
- **Visualization & Monitoring**: MLFlow, Prometheus, Grafana.
- **Programming & Libraries**: Python (pandas, dask, PyTorch), SQL.

What Participants Can Expect After Completing This Training

- A solid understanding of how to operationalize Generative AI technologies within data engineering pipelines.
- Proficiency in building scalable, cloud-ready workflows for large-scale data generation, data preprocessing, and model serving.
- Knowledge of the best practices for integrating LLM-based solutions (e.g., GPT-4, DALL-E) into ETL workflows and creating AI-ready datasets.
- Hands-on experience with real-world data workflows related to GenAI, such as chunking data for LLMs or creating feature engineering pipelines.
- Ability to monitor, debug, and optimize AI-driven data pipelines for performance and accuracy.

Module 1: Introduction to Generative AI for Data Engineers (3 Hours)

Objective:

Understand the foundational concepts of Generative AI and the role of data engineers in supporting and optimizing AI workflows.

Topics Covered:

- Overview of Generative AI: LLMs, Diffusion Models, and Multimodal AI.
- Key challenges in Generative AI for data engineering (data scaling, preprocessing, and pipeline orchestration).
- Infrastructure concepts: high-performance data storage, distributed processing, and cloud optimizations.

Demo:

- Connecting an AI model (OpenAI API) to a database for text retrieval and summarization.

Hands-on Practice:

- Participants connect a cloud database (e.g., AWS RDS or GCP BigQuery) to GPT-4 to retrieve, summarize, and log queries into storage using Python.

Module 2: Preparing and Processing Data for Generative AI (4 Hours)

Objective:

Learn how to preprocess, transform, and scale data for GenAI models, ensuring optimal performance in training or inference tasks.

1. **Topics Covered**:
- Text Preprocessing: Tokenization, chunking for LLMs, cleaning noisy datasets.
- Image Preprocessing: Rescaling, augmenting for Stable Diffusion, or vision models.
- Optimized File Formats and Storage: Parquet, Delta Lake, and columnar storage for scalability.
- Distributed Frameworks: Using Apache Spark and Dask for batch processing large datasets.

2. **Demo**:
- Cleaning a public dataset of text data (e.g., customer reviews from Kaggle) and splitting it into chunks optimized for GPT token limits.

3.**Hands-on Practice**:
- Task: Participants preprocess a raw dataset (e.g., customer feedback logs or large JSON files) and transform it into chunks ready for LLM inference using Python.
- Tools: pandas, transformers (Hugging Face).

Module 3: Building Data Engineering Pipelines for AI Models (5 Hours)

Objective:
Learn how to orchestrate large-scale GenAI data pipelines using Airflow, Prefect, or equivalent tools.

1. Topics Covered:
- Building pipelines for GenAI workflows (e.g., create, preprocess, and serve datasets).
- Integrating with cloud storage systems like Amazon S3, GCP Cloud Storage, or Azure Blob Storage.
- Auto-scaling pipelines for text generation and batch processing.

2. Demo:
- Building an Apache Airflow pipeline for automating a weekly GPT-powered dataset summarization task.

3.Hands-on Practice:
- Task: Participants create an ETL pipeline that ingests raw data from cloud storage, preprocesses it for use with Hugging Face models, and stores results in a data lake (e.g., S3 or Delta Lake).

Module 4: Distributed Data Systems for Generative AI (4 Hours)

Objective:
Explore distributed systems for handling large GenAI datasets and scaling GenAI workflows.

1. **Topics Covered:**
- Using Spark or Dask to process distributed datasets for language models.
- Best practices for working with Delta Lake and Parquet for large-scale text/image datasets.
- Managing memory and optimizing query performance for real-time AI use cases.

2. **Demo:**
- Processing and cleaning a large-scale dataset using Apache Spark, optimized for LLM fine-tuning.
- Storing structured outputs in Delta Lake for downstream pipeline tasks.

3. **Hands-on Practice:**
- Task: Participants configure and process a 1GB sample dataset (e.g., Reddit comments or news data) using Apache Spark. They extract named entities (NER) and save them in an optimized Parquet format.

Module 5: Deploying GenAI Workflows and Monitoring Performance (4 Hours)

Objective:
Learn how to deploy Generative AI pipelines and monitor their performance in production environments.

1. Topics Covered

- Model Serving: Deploying Hugging Face models using TensorFlow Serving, TorchServe, or FastAPI.
- Automating Monitoring: Using Prometheus and Grafana to track pipeline performance and latency.
- Managing Deployments: Version control for models and workflows with MLFlow.

2. Demo:

- Deploying a fine-tuned GPT or Stable Diffusion model using a FastAPI endpoint and visualizing logs in Grafana.

3. Hands-on Practice:

- Task: Participants deploy a FastAPI application that serves a sentiment analysis model and tracks API performance metrics with Prometheus (e.g., requests/second, inference latency).

Summary of Training:

1. Participants will have working knowledge of prepping data, building pipelines, and deploying AI workflows at scale.
2. They will leave with clear, practical skills to collect, process, and deliver data efficiently for Generative AI tasks across real-world business environments.