EmHashLabs

Here’s a **detailed course content** for a training program on **Data Engineering, Pipeline Development, and ETL**:

Module 1: Introduction to Data Engineering

1. **What is Data Engineering?**
   - Definition and Role of a Data Engineer
   - Difference Between Data Engineering and Data Science
   - Data Engineering Workflow Overview
2. **Data Engineering Tools and Technologies**
   - Key Tools: Hadoop, Spark, Apache Kafka, Airflow
   - Cloud-Based Data Engineering: AWS, GCP, Azure
   - Introduction to NoSQL and SQL Databases
3. **Data Engineering Ecosystem**
   - Data Lakes vs. Data Warehouses
   - Data Streams and Batch Processing
   - Real-Time vs. Batch Processing

Module 2: Data Pipeline Basics

1. **What is a Data Pipeline?**
   - Definition and Purpose of Data Pipelines
   - Types of Data Pipelines: Batch, Streaming, Hybrid
   - Components of a Data Pipeline: Data Collection, Transformation, Storage, and Analysis
2. **Building Data Pipelines**
   - Design Considerations: Scalability, Fault Tolerance, Latency
   - Steps to Build a Simple Data Pipeline
   - Best Practices for Data Pipeline Development
3. **Data Flow Management and Scheduling**
   - Data Ingestion Methods: APIs, File Systems, Databases
   - Workflow Orchestration with Apache Airflow and Luigi
   - Scheduling and Monitoring Data Pipelines

Module 3: Extract, Transform, Load (ETL) Fundamentals

1. **What is ETL?**
   - Definition and Importance of ETL in Data Engineering
   - ETL vs. ELT: Key Differences
   - Common ETL Tools: Apache NiFi, Talend, Informatica
2. **Extracting Data (E)**
   - Data Extraction Methods: APIs, Database Queries, File Transfers
   - Handling Different Data Formats: CSV, JSON, Parquet, Avro
   - Dealing with Structured and Unstructured Data
3. **Transforming Data (T)**
   - Data Cleaning: Handling Missing Values, Duplicates, and Outliers
   - Data Transformation Techniques: Aggregation, Normalization, Parsing
   - Using Python, Spark, and SQL for Data Transformation
4. **Loading Data (L)**
   - Loading Data into Data Warehouses, Databases, or Data Lakes
   - Batch vs. Real-Time Data Loading
   - Optimizing Data Loading for Performance

Module 4: Advanced ETL Concepts

1. **Data Quality and Validation**
   - Importance of Data Quality in ETL Processes
   - Techniques for Data Validation and Integrity Checks
   - Implementing Error Handling and Data Cleansing
2. **Incremental and Delta Loads**
   - What is Incremental Loading?
   - Techniques for Handling Delta Loads
   - Using Change Data Capture (CDC) in ETL Pipelines
3. **Performance Optimization in ETL**
   - Parallelism and Distributed Computing with Spark
   - Caching and Partitioning in ETL Pipelines
   - Optimizing SQL Queries and Data Transformations

4. **ETL Best Practices and Challenges**
   - Designing Robust ETL Pipelines for Scalability
   - Managing Dependencies and Scheduling
   - Handling Large-Scale Data Volume

Module 5: Data Pipeline Automation and Orchestration

1. **Workflow Automation with Apache Airflow**
   - Introduction to Apache Airflow: DAGs and Tasks
   - Creating and Managing Data Pipelines in Airflow
   - Handling Task Failures and Retries
2. **Data Pipeline Orchestration**
   - Managing Data Workflow Dependencies
   - Orchestrating Complex Data Pipelines with Airflow or Luigi
   - Automating Data Pipelines for Continuous Integration
3. **Monitoring and Logging Pipelines**
   - Setting Up Logging and Alerts for Data Pipelines
   - Monitoring ETL Pipeline Performance
   - Using Prometheus and Grafana for Real-Time Monitoring

Module 6: Cloud-Based Data Engineering

1. **Data Engineering in the Cloud**
   - Advantages of Cloud-Based Data Pipelines
   - Overview of AWS, Azure, and Google Cloud Data Engineering Services
   - Cloud Storage Solutions: S3, GCS, Blob Storage
2. **Managed Data Services**
   - AWS Glue, Azure Data Factory, Google Cloud Dataflow
   - Using Managed Services for ETL and Data Integration
   - Integrating Cloud Services with On-Premise Data Sources
3. **Cloud Data Warehousing**
   - Introduction to Data Warehouses: Snowflake, Redshift, BigQuery
   - Cloud-Based ETL Pipelines for Data Warehouses

- Scaling Data Pipelines with Cloud Resources

Module 7: Real-Time Data Pipelines and Streaming

1. **Introduction to Real-Time Data Pipelines**
   - Differences Between Batch and Streaming Pipelines
   - Use Cases for Real-Time Data Pipelines
   - Challenges of Real-Time Data Processing
2. **Streaming with Apache Kafka**
   - Overview of Apache Kafka Architecture
   - Setting Up Kafka Producers and Consumers
   - Real-Time Data Processing with Kafka Streams
3. **Stream Processing Frameworks**
   - Using Apache Flink for Stream Processing
   - Stream Processing with Spark Streaming
   - Building a Real-Time ETL Pipeline

Module 8: Data Pipeline Security and Governance

1. **Data Security in Pipelines**
   - Securing Sensitive Data in Transit and at Rest
   - Role-Based Access Control (RBAC) and Authentication
   - Encryption Techniques for Data in Pipelines
2. **Data Governance and Compliance**
   - Importance of Data Governance in Pipelines
   - Data Lineage and Metadata Management
   - Adhering to Data Privacy Regulations: GDPR, HIPAA
3. **Monitoring Pipeline Security**
   - Best Practices for Monitoring and Auditing Data Pipelines
   - Threat Detection and Response in ETL Processes

Module 9: Hands-On Projects and Case Studies

1. **Project 1: Building an ETL Pipeline Using Apache Airflow**

2. **Project 2: Designing a Scalable Data Pipeline on AWS**
3. **Project 3: Real-Time Data Streaming with Apache Kafka**
4. **Case Study: Migrating ETL Processes to the Cloud**
5. **Capstone Project: End-to-End Data Pipeline Implementation**

Module 10: Closing and Certification

1. **Final Assessment**
2. **Review and Feedback**
3. **Certification of Completion**
4. **Career Guidance and Further Learning Resources**