Data Engineering Course Content

1. Introduction to Data Engineering
2. Python for Data Engineering (Hands on)
3. SQL and Data Modeling (Hands on)
4. Data Warehousing and Ingestion (Hands on)
5. Big Data Processing with Apache Spark (Hands on)
6. Lakehouse and Open Table Formats
7. Workflow Orchestration with Apache Airflow (Hands on)
8. Transformation with dbt (Hands on)
9. Streaming with Apache Kafka (Hands on)
10. Cloud, DevOps and Data Quality
11. Capstone Projects

Introduction to Data Engineering

- Role of a Data Engineer vs Data Analyst and Data Scientist
- Data Lifecycle, Batch vs Streaming, ETL vs ELT
- The Modern Data Stack and Pipeline Architecture Patterns

Python for Data Engineering (Hands on)

- Python Basics, Functions, File Handling, Working with APIs and JSON
- Pandas, Polars and DuckDB for Local Processing
- Linux, Bash, Git and GitHub, Virtual Environments

SQL and Data Modeling (Hands on)

- Advanced SQL: Joins, CTEs, Window Functions, Query Optimization
- PostgreSQL, Indexing, Transactions
- OLTP vs OLAP, Star and Snowflake Schemas, Slowly Changing Dimensions

Data Warehousing and Ingestion (Hands on)

- Snowflake: Architecture, Stages, Loading Data, Cost Control
- BigQuery and Redshift Concepts
- Ingestion Tools: Airbyte and Fivetran Overview, Change Data Capture Concepts

Big Data Processing with Apache Spark (Hands on)

- Spark Architecture, PySpark DataFrames, Spark SQL
- Joins, Partitioning, Shuffles, Performance Tuning
- Databricks Workspace, Notebooks and Jobs

Lakehouse and Open Table Formats

- Data Lakes, Parquet and Columnar Storage
- Delta Lake and Apache Iceberg: ACID Tables, Schema Evolution, Time Travel
- Catalogs and Governance Overview

Workflow Orchestration with Apache Airflow (Hands on)

- DAGs, Tasks, Operators, Scheduling, Sensors
- Retries, Backfills, Alerts, Connections and Variables
- Dagster and Prefect Overview

Transformation with dbt (Hands on)

- Models, Sources, Refs, Materializations, Incremental Models
- Tests, Documentation, Snapshots, Macros
- dbt Core vs dbt Cloud, Project Structure and Best Practices

Streaming with Apache Kafka (Hands on)

- Topics, Partitions, Producers, Consumers, Consumer Groups
- Kafka Connect and Schema Management
- Spark Structured Streaming, Flink Overview, Real-Time Use Cases

Cloud, DevOps and Data Quality

- AWS (S3, Glue, Lambda, Redshift, IAM) or Azure (Data Factory, ADLS, Synapse) Overview
- Docker, GitHub Actions CI/CD, Terraform Basics
- Data Quality with Great Expectations and Soda, Lineage, Observability
- Data for AI: Vector Databases and Feeding RAG Pipelines Overview

Capstone Projects

- End-to-End Batch ELT Pipeline: Source to Lakehouse to dbt to Dashboard
- Real-Time Streaming Pipeline with Kafka and Spark
- GitHub Portfolio, Resume Building and Interview Preparation