Data24Zone Academy

+91 94899 31327 (call / WhatsApp)
hello@data24zone.com · data24zone.com/courses/data-engineering

CAREER PROGRAM · SYLLABUS · DATA & AI

Data Engineering

Learn to design, build, and operate production-grade data pipelines — from ingestion and warehousing to distributed batch processing, streaming, and cloud lakehouse platforms.

DURATION	COMMITMENT	LEVEL	FORMATS
14 Weeks	10 hrs/week	Intermediate	Online Live · Offline (Coimbatore) · Self-Paced

Tools & tech stack

Python · Advanced SQL · Apache Spark · Kafka · Cloud Data Warehousing · SQL · dbt · Apache Airflow · Apache Kafka · Snowflake / BigQuery · Great Expectations

Who it's for

- Developers and analysts who want to design and operate reliable, production-grade data pipelines

What you'll learn

- Building ELT/ETL pipelines with Python, SQL, and dbt
- Batch processing at scale with Spark and workflow orchestration with Airflow
- Streaming pipelines with Kafka and cloud data warehousing on Snowflake or BigQuery
- Using AI-assisted tooling for schema inference, anomaly detection, and pipeline documentation

Curriculum

01 Getting Started & Environment Setup

- Installing Python, dbt, and setting up a free Snowflake or BigQuery trial
- Installing Apache Spark (PySpark) locally
- Setting up Apache Airflow and Apache Kafka via Docker
- Verifying your setup by running a trivial PySpark job

02 Basic

- Dev environment setup, Linux, and Git workflows
- Advanced SQL: window functions and CTEs
- API ingestion, pagination, and idempotent script design
- Working with Parquet, JSON, and CSV formats

03 Intermediate

- Star and snowflake schema design
- Dimensional modeling with facts and dimensions
- Slowly changing dimensions
- Building and testing dbt models

04 Advanced

- Batch vs. streaming ETL patterns
- Incremental loads, CDC, and backfills
- PySpark DataFrame API and joins at scale
- Performance tuning: partitioning, shuffling, and caching

05 Pro

- Airflow DAG design, dependencies, and retries
- Scheduling, backfilling, and failure alerting
- Kafka producers, consumers, and consumer groups
- Spark Structured Streaming and windowed aggregations
- Snowflake, BigQuery, and Databricks lakehouse architecture
- Query optimization, partitioning, and cost control
- Data validation suites, freshness monitoring, and lineage tracking

06 AI Pro

- Schema inference from raw files with an LLM
- AI-generated transformation code and SQL review
- Auto-generated pipeline documentation
- Guardrails for reviewing AI-generated code

Projects

- E-commerce data warehouse with a dimensional star schema and a Type 2 slowly changing dimension, built in dbt
- Real-time clickstream pipeline streaming windowed session aggregates into a lakehouse table with Kafka and Spark
- Batch ELT pipeline with automated data-quality gates that halt on schema violations or null-rate spikes
- CDC replication pipeline capturing change events from Postgres and syncing them to a cloud warehouse in near real time
- Cost-optimized cloud warehouse migration with before/after query-latency and cost benchmarks

---