

Data Engineering & ETL Pipeline Internship

(15 Days | Practical-Oriented | Industry Ready)

INTERNSHIP SYLLABUS

Internship Objectives

- Understand **end-to-end Data Engineering lifecycle**
- Build **ETL pipelines from scratch**
- Work with **real datasets & databases**
- Implement **cloud-ready data pipelines**
- Gain confidence to do **final-year / real-time projects**

TOOLS & TECHNOLOGIES

- Python, Pandas
- SQL (MySQL / PostgreSQL)
- MongoDB
- Apache Spark
- Apache Airflow
- Cloud Storage (AWS/Azure)
- Git & GitHub

DAY 1 – Introduction to Data Engineering

- What is Data Engineering vs Data Science
- Role of Data Engineer

- OLTP vs OLAP
- Data lifecycle & pipeline architecture

Practical

Install Python, VS Code, Git

Dataset exploration (CSV, Excel, JSON)

Outcome

Clear understanding of data engineering workflow

DAY 2 – Data Sources & Ingestion

- Structured, Semi-structured, Unstructured data
- Batch vs Streaming data
- APIs, Logs, Databases

Practical

Read data from CSV, Excel, JSON

Fetch data from REST API using Python

Outcome

Ability to ingest data from multiple sources

DAY 3 – Databases for Data Engineers

- SQL vs NoSQL
- MySQL, PostgreSQL, MongoDB
- Schema design basics

Practical

Create MySQL & MongoDB databases

Insert & query sample datasets

Outcome

Working knowledge of databases

DAY 4 – Data Cleaning & Transformation

- Data quality issues
- Missing values, duplicates, outliers
- Normalization & transformation rules

Practical

Pandas-based data cleaning

Transform raw data to clean format

Outcome

Clean, reliable datasets for pipelines

DAY 5 – ETL Fundamentals

- ETL vs ELT
- Pipeline components
- Error handling & logging

Practical

Build first ETL pipeline using Python

Source → Transform → Load to DB

Outcome

First working ETL pipeline completed

DAY 6 – Advanced SQL for Data Engineering

- Joins, Subqueries, Indexing
- Aggregations & Window functions

Practical

Optimize queries for large datasets

Load transformed data into SQL warehouse

Outcome

Efficient data querying skills

DAY 7 – Data Warehousing Concepts

- Star & Snowflake schema
- Fact & Dimension tables
- Data marts

Practical

Design data warehouse schema

Load fact & dimension tables

Outcome

Warehouse-ready data modeling skills

DAY 8 – Big Data Basics

- What is Big Data?
- Hadoop & Spark overview
- Distributed processing

Practical

Intro to PySpark

Data processing using Spark DataFrames

Outcome

Understanding scalable data processing

DAY 9 – Workflow Orchestration

- What is workflow scheduling?
- Apache Airflow basics
- DAG concepts

Practical

Create ETL DAG in Airflow

Schedule automated pipeline

Outcome

Automated pipelines using Airflow

DAY 10 – Cloud Data Engineering

- Cloud data architecture
- AWS / Azure / GCP overview
- Storage & compute services

Practical

Upload data to cloud storage

Run ETL on cloud environment

Outcome

Cloud-ready data pipeline skills

DAY 11 – Data Streaming Basics

- Real-time vs batch processing
- Kafka overview

Practical

Simulate streaming data

Process live data stream

Outcome

Introduction to real-time pipelines

DAY 12 – Data Validation & Monitoring

- Data consistency checks
- Logging & alerting
- Pipeline monitoring

Practical

Implement data validation rules

Error handling in ETL jobs

Outcome

Reliable & monitored pipelines

DAY 13 – Security & Optimization

- Data privacy & access control
- Performance tuning
- Cost optimization

Practical

Secure DB access

Optimize ETL performance

Outcome

Industry-level data pipeline practices

DAY 14 – Mini Project Development

- Sales Data ETL Pipeline
- Healthcare Data Warehouse
- E-commerce Analytics Pipeline

Practical

Build complete end-to-end pipeline

Source → Transform → Warehouse → Report

Outcome

Real-world project experience

DAY 15 – Final Project Demo & Career Guidance

- Project presentation
- Interview Q&A
- Resume & portfolio guidance

Outcome

Confidence to attend interviews & build projects independently

