PATHAHEAD · FREE ROADMAP

Become a Data Engineer from Scratch

Build the pipelines that power every dashboard and ML model. Go from zero to job-ready in about six months, with real pipelines and a data platform capstone.

24	5	4	₹0
weeks · ~10–12 hrs/wk	hands-on projects	phases · to production	free tiers only

A free PathAhead learning resource · adjust the pace to your life — the order matters more than the dates.

THE ONE RULE

Ingest → transform → serve

Data engineering is plumbing at scale — you learn it by moving real data. Every project below builds a pipeline that ingests, transforms, and serves data reliably.

THE ROADMAP

Four phases, step by step

0 Foundations

Weeks 1–5

Python, SQL, and databases — deeply.

WEEK 1
Python for data
Scripting, files, APIs, virtualenvs.

WEEK 2
SQL deep dive
Joins, aggregation, window functions.

WEEK 3
Database design
Relational modeling, normalization, indexes.

WEEK 4
Linux & CLI
Shell, cron, file handling.

WEEK 5
Git & environments
Version control and reproducible setups.

1 Pipelines & ETL

Weeks 6–11

Move and clean data on a schedule.

WEEK 6
ETL/ELT concepts
Extract, transform, load patterns.

WEEK 7
Batch processing
pandas and data transformation at scale.

WEEK 8
Orchestration
Scheduling pipelines with Airflow.

WEEK 9
Ingestion
Pull from APIs, files, and databases.

WEEK 10
Data quality
Validation, tests, and monitoring.

WEEK 11 · PROJECT ▶
ETL pipeline
API → clean → load into a database.

2 Big Data & Warehouses

Weeks 12–18

Store and process data at real scale.

WEEK 12
Data warehouses
Snowflake or BigQuery fundamentals.

WEEK 13
Dimensional modeling
Star schemas, facts and dimensions.

WEEK 14
Apache Spark
Distributed processing basics.

WEEK 15
Scaling
Partitioning, parallelism, big datasets.

WEEK 16
Streaming
Real-time data with Kafka.

WEEK 17
dbt
Version-controlled SQL transformations.

WEEK 18 · PROJECT ▶
Warehouse + dbt
Model a warehouse and transform with dbt.

3 Cloud & Scale

Weeks 19–24

Run pipelines in production, in the cloud.

WEEK 19
Cloud data platforms
AWS or GCP data services.

WEEK 20
Docker for data
Containerize pipelines.

WEEK 21
Infrastructure basics
IaC and deployment for data.

WEEK 22
Production orchestration
Monitoring and alerting on pipelines.

WEEK 23
Performance & cost
Tuning queries and controlling spend.

WEEK 24 · CAPSTONE ▶
Data platform
An end-to-end ingest-to-serve platform.

BUILD A PORTFOLIO

Five projects that get you hired

Projects beat certificates. Do them in order — each builds on the last. Put every one on GitHub with a clear README: problem → approach → result.

Start here

1 · ETL Script

Pull from a public API, clean it, load it into Postgres.

Orchestration

2 · Airflow Pipeline

A scheduled pipeline that runs and monitors itself.

Warehouse

3 · Warehouse + dbt

Model data in a warehouse and transform it with dbt.

In demand

4 · Streaming Pipeline

Ingest real-time events with Kafka.

Capstone

5 · End-to-End Data Platform

Ingestion → warehouse → transformation → serving layer.

ROLES YOU CAN APPLY FOR

Data Engineer · Analytics Engineer · ETL Developer · Big Data Engineer

TYPICAL STARTING SALARY · INDIA

₹6–14 LPA

LEARN IT

Learning resources

Every item links to the real source (clickable in this PDF).

BOOKS

- Designing Data-Intensive Applications by Martin Kleppmann
- Fundamentals of Data Engineering by Joe Reis and Matt Housley
- The Data Warehouse Toolkit by Ralph Kimball and Margy Ross

COURSES

- Data Engineering Zoomcamp by DataTalks Club (free)
- Google Cloud Data Engineering by Coursera
- Big Data by NPTEL (free)
- IBM Data Engineering Certificate by Coursera

YOUTUBE

- Seattle Data Guy
- Data with Zach
- freeCodeCamp.org

PRACTICE

- Kaggle
- GitHub
- DataCamp

DO THIS TODAY

Your first step (takes 30 minutes)

1

Install Python + PostgreSQL

Free — your local data engineering workbench.

2

Write one SQL query

Load sample data and query it. The core skill starts here.

3

Build one ETL script

Pull from a public API and save it to a table today.

Job-ready checklist: 3–5 solid projects on GitHub (≥1 deployed or documented) · can explain your work in an interview · comfortable with the core tools of the field · a portfolio or profile recruiters can find.

PathAhead — India's premium career guidance platform. This roadmap is a free educational resource; timelines are a guide and everyone learns at a different pace. Referenced third-party resources (Data Engineering Zoomcamp, dbt Learn, DataCamp, freeCodeCamp, Mode SQL Tutorial) are for learning only and are not affiliated with PathAhead.