Data Engineering Roadmap 2025

From Zero to First Job

No fluff. Just what works.

The Reality Check

❌ You don't need 20 tools

❌ You don't need a Master's degree

✅ You need: SQL + Python + One Cloud

✅ You need: 2-3 strong projects

←

→

Must-Master Concepts

(In this order)

1. SQL (Window functions, CTEs, query optimization)
2. Python/PySpark (intermediate level)
3. Databricks/Snowflake
4. Data Modeling (Star Schema, SCD Type 2)
5. ETL/ELT pipelines (Airflow, dbt)

←
→

Must-Master Concepts

(Continued)

6. Cloud (AWS or GCP or Azure - pick ONE)
7. Kafka basics (producer/consumer)
8. Spark internals (partitioning, shuffling)
9. Data Quality (Great Expectations, dbt tests)
10. Git + CI/CD for data pipelines

SQL Cheat Sheet

Practice more. Revise more

- JOINs (INNER, LEFT, RIGHT, FULL)
- GROUP BY + Aggregations
- WINDOW functions (RANK, ROW_NUMBER, LEAD)
- CTEs (WITH clause)
- Subqueries vs JOINs

Asked in EVERY interview

Python Essentials

Learn only what you need


Pandas (read, filter, transform, write)


File handling (CSV, JSON, Parquet)


APIs (requests library)


Error handling (try/except)


Functions + Classes (basics)

Keep it simple and hands-on

Tools You Must Know

Pick ONE from each row

Cloud	AWS or GCP or Azure
Compute	Spark or Snowflake
Orchestration	Airflow or Prefect
Storage	S3 or ADLS or GCS
Streaming	Kafka or Kinesis

Don't overthink. Just pick and learn.

Projects That Get Callbacks

End-to-End ETL Pipeline

API → S3 → Spark → Redshift/BigQuery

Python Airflow SQL

Real-time Dashboard

Kafka → Spark Streaming → Dashboard

Kafka PySpark Streamlit

Data Quality Framework

Automated data validation + alerts

Great Expectations Airflow

Resume One-Liners

Copy these formats

- Built end-to-end ETL pipelines using Spark and Python
- Optimized pipelines that cut runtime by 30%
- Designed data models used in production dashboards
- Automated ingestion from 3+ data sources
- Reduced query cost by 60% using partitioning
- Implemented SCD Type 2 for 2TB customer table

Interview Questions

What they actually ask

1. Design a SCD Type 2 table
2. Optimize a 10-min query to <10 seconds
3. Handle 100GB skewed data in Spark
4. Build API → Bronze → Silver → Gold pipeline
5. Explain partitioning vs bucketing vs Z-ordering
6. "Your Airflow DAG failed at 3 AM - debug it"

Your 90-Day Plan

Week 1-3
SQL mastery (LeetCode, HackerRank)

Week 4-6
Python + Pandas projects

Week 7-9
Pick a cloud + complete tutorial

Week 10-12
Build Project #1 (ETL pipeline)

Week 13+
Apply to jobs + keep building

Final Advice

✅ Learn ONE thing at a time

✅ Build small, daily

✅ Publish your learnings (LinkedIn/Medium)

✅ Reach out for referrals (DM people!)

❌ Don't get stuck in tutorial hell

❌ Don't wait to feel "ready"

Start today. Not tomorrow.

Learn from the Best

Ankit Bansal (YouTube)

SQL interview questions & patterns

Darshil Parmar (YouTube)

Real project tutorials

Zach Wilson (LinkedIn)

DE career & Bootcamps

Ansh Lamba (Youtube)

DE projects & Interview prep

Sumit Mittal (Youtube)

Interview prep & Spark

Shashank Mishra (LinkedIn)

Big Data and DE

Krish Naik (YouTube)

Python & ML pipelines