Data Engineer Complete Roadmap (2026) – Interview Preparation Topics

This roadmap is designed for **0–5+ years of experience** and covers the topics most frequently asked in interviews at companies like Amazon, Microsoft, Google, Deloitte, Accenture, Infosys, TCS, Cognizant, Capgemini, EY, KPMG, Wipro, IBM, and many product-based companies.

Module	Topics to Learn	Interview Priority
1	Data Engineering Fundamentals	⭐ ⭐ ⭐ ⭐ ⭐
2	SQL (Basic to Advanced)	⭐ ⭐ ⭐ ⭐ ⭐
3	Database Concepts	⭐ ⭐ ⭐ ⭐ ⭐
4	Data Warehousing	⭐ ⭐ ⭐ ⭐ ⭐
5	ETL & ELT Concepts	⭐ ⭐ ⭐ ⭐ ⭐
6	Data Modeling	⭐ ⭐ ⭐ ⭐ ⭐
7	Python for Data Engineering	⭐ ⭐ ⭐ ⭐ ⭐
8	Linux & Shell Scripting	⭐ ⭐ ⭐ ⭐
9	Git & Version Control	⭐ ⭐ ⭐ ⭐
10	Apache Spark	⭐ ⭐ ⭐ ⭐ ⭐
11	PySpark	⭐ ⭐ ⭐ ⭐ ⭐
12	Spark SQL	⭐ ⭐ ⭐ ⭐ ⭐
13	Delta Lake	⭐ ⭐ ⭐ ⭐ ⭐
14	Databricks	⭐ ⭐ ⭐ ⭐ ⭐
15	Azure Data Factory (ADF)	⭐ ⭐ ⭐ ⭐ ⭐
16	Azure Synapse Analytics	⭐ ⭐ ⭐ ⭐
17	Microsoft Fabric	⭐ ⭐ ⭐ ⭐
18	Snowflake	⭐ ⭐ ⭐ ⭐ ⭐
19	Apache Airflow	⭐ ⭐ ⭐ ⭐ ⭐
20	Kafka	⭐ ⭐ ⭐ ⭐
21	Streaming Concepts	⭐ ⭐ ⭐ ⭐
22	Hadoop Ecosystem	⭐ ⭐ ⭐
23	Cloud Fundamentals (Azure/AWS/GCP)	⭐ ⭐ ⭐ ⭐ ⭐
24	Azure Data Lake Storage (ADLS Gen2)	⭐ ⭐ ⭐ ⭐ ⭐
25	Blob Storage	⭐ ⭐ ⭐ ⭐
26	Data Lake Concepts	⭐ ⭐ ⭐ ⭐ ⭐

27	Data Pipeline Design	⭐ ⭐ ⭐ ⭐ ⭐
28	Data Quality & Validation	⭐ ⭐ ⭐ ⭐
29	Data Governance	⭐ ⭐ ⭐
30	CI/CD for Data Engineering	⭐ ⭐ ⭐ ⭐
31	DevOps Basics	⭐ ⭐ ⭐
32	Performance Optimization	⭐ ⭐ ⭐ ⭐ ⭐
33	Partitioning & Indexing	⭐ ⭐ ⭐ ⭐ ⭐
34	Query Optimization	⭐ ⭐ ⭐ ⭐ ⭐
35	Scenario-Based Interview Questions	⭐ ⭐ ⭐ ⭐ ⭐
36	System Design for Data Engineers	⭐ ⭐ ⭐ ⭐ ⭐
37	Behavioral Interview Questions	⭐ ⭐ ⭐ ⭐
38	Real-Time Project Scenarios	⭐ ⭐ ⭐ ⭐ ⭐
39	Coding & Problem Solving	⭐ ⭐ ⭐ ⭐
40	Mock Interviews	⭐ ⭐ ⭐ ⭐ ⭐

Complete Topic Breakdown

Module 1: Data Engineering Fundamentals

- What is Data Engineering?
- OLTP vs OLAP
- ETL vs ELT
- Data Pipeline
- Batch vs Streaming
- Structured vs Semi-Structured vs Unstructured Data
- Data Lake vs Data Warehouse vs Data Lakehouse
- CAP Theorem
- ACID Properties
- Data Mesh vs Data Fabric (overview)

Module 2: SQL (Most Important)

Basic SQL

- SELECT
- WHERE
- ORDER BY
- GROUP BY
- HAVING
- DISTINCT
- LIMIT

Intermediate SQL

- INNER JOIN
- LEFT JOIN
- RIGHT JOIN
- FULL OUTER JOIN
- CROSS JOIN
- SELF JOIN

Advanced SQL

- Window Functions
- ROW_NUMBER()
- RANK()
- DENSE_RANK()
- LEAD()
- LAG()
- NTILE()
- CTEs
- Recursive CTE
- Views
- Stored Procedures
- Functions
- Dynamic SQL

Optimization

- Indexing
- Execution Plans
- Query Tuning
- Partitioning
- Clustering
- Materialized Views

Module 3: Database Concepts

- Normalization
- Denormalization
- Primary Key
- Foreign Key
- Candidate Key
- Composite Key
- Surrogate Key
- Transactions
- Locking
- Isolation Levels
- Deadlocks

Module 4: Data Warehousing

- Facts
- Dimensions
- Star Schema
- Snowflake Schema
- Factless Fact Tables
- Slowly Changing Dimensions (SCD 0–6)
- Surrogate Keys
- Data Marts

Module 5: ETL & ELT

- ETL Architecture
- ELT Architecture
- Incremental Load
- Full Load
- CDC (Change Data Capture)
- Watermarking
- Error Handling
- Retry Mechanisms
- Logging
- Audit Tables

Module 6: Python

- Variables
- Loops
- Functions
- OOP
- File Handling
- Exception Handling
- Collections
- List Comprehension
- Generators
- Decorators
- Pandas
- NumPy

Module 7: Apache Spark

- RDD
- DataFrames
- Transformations
- Actions
- Lazy Evaluation
- DAG
- Shuffle
- Broadcast Join

- Partitioning
- Caching
- Persistence

Module 8: PySpark

- SparkSession
- DataFrame API
- UDF
- Window Functions
- Joins
- Aggregations
- Performance Tuning

Module 9: Databricks

- Workspace
- Notebooks
- Jobs
- Clusters
- Unity Catalog
- Delta Live Tables
- Workflows
- Secrets
- Repos

Module 10: Delta Lake

- Delta Tables
- MERGE
- UPDATE
- DELETE
- Time Travel
- OPTIMIZE
- VACUUM
- Z-Ordering
- Schema Evolution

Module 11: Azure Data Factory

- Pipelines
- Activities
- Linked Services
- Datasets
- Triggers
- Parameters
- Variables
- Expressions
- Copy Activity

- Mapping Data Flow
- Lookup
- ForEach
- If Condition
- Until
- Execute Pipeline

Module 12: Snowflake

- Virtual Warehouses
- Time Travel
- Zero Copy Clone
- Streams
- Tasks
- Stages
- File Formats
- Snowpipe
- Secure Views

Module 13: Cloud

Azure

- ADLS
- Blob Storage
- Key Vault
- Synapse
- Microsoft Fabric

AWS

- S3
- Glue
- Redshift
- EMR
- Lambda
- Athena

GCP

- BigQuery
- Dataflow
- Cloud Storage
- Dataproc

Module 14: Apache Airflow

- DAG
- Operators

- Sensors
- Scheduling
- XCom
- Variables
- Connections
- Retry Policies
- Dynamic DAGs

Module 15: Kafka

- Producers
- Consumers
- Topics
- Partitions
- Offsets
- Consumer Groups
- Exactly Once Processing

Module 16: Performance Optimization

- Partitioning
- Bucketing
- Broadcast Join
- Predicate Pushdown
- Caching
- Adaptive Query Execution (AQE)
- Skew Handling

Module 17: System Design

- Design an ETL Pipeline
- Real-Time Streaming Pipeline
- Data Lake Architecture
- CDC Pipeline
- Incremental Load Pipeline
- Log Processing System
- Batch Processing System

Module 18: Real-Time Scenarios

- SCD Type 2 Implementation
- Incremental Load Design
- Late Arriving Data
- Duplicate Record Handling
- Corrupt File Processing
- Retry Logic
- Pipeline Failure Recovery
- Data Validation
- Slowly Changing Dimensions

- Schema Evolution

Most Asked Interview Topics

(Priority Order)

1. SQL (Advanced)
2. PySpark
3. Databricks
4. Azure Data Factory
5. Data Warehousing
6. Delta Lake
7. Performance Optimization
8. Data Modeling
9. System Design
10. Snowflake
11. Apache Airflow
12. Kafka
13. Cloud (Azure/AWS/GCP)
14. Python
15. Scenario-Based Questions
16. Real-Time Project Discussions

This roadmap covers the core knowledge expected for modern Data Engineer roles in 2026 and is structured to prepare you for both technical interviews and real-world project discussions.