**Data Engineering Roadmap**

**Phase 1: Basics of Data & Programming**

1. **Understanding Data Engineering**
   * Role of a Data Engineer vs Data Analyst vs Data Scientist
   * Data lifecycle: Ingestion → Processing → Storage → Analysis
2. **Programming Fundamentals**
   * **Python**: Data structures, file handling, OOP, error handling
   * **SQL**: Joins, Window functions, CTEs, performance tuning
3. **Linux & Command Line Basics**
   * Navigating file systems, running scripts, setting up cron jobs
4. **Git & Version Control**
   * git init, git commit, git push, branches, merges

**Phase 2: Data Modeling & Warehousing**

1. **Data Modeling Concepts**
   * OLTP vs OLAP
   * Star & Snowflake schema
   * Normalization/Denormalization
2. **Relational Databases (RDBMS)**
   * PostgreSQL, MySQL, SQL Server basics
3. **Data Warehouses**
   * Snowflake, Google BigQuery, Amazon Redshift
   * Partitioning, Clustering, Sharding

**Phase 3: ETL & ELT Pipelines**

1. **ETL Tools & Frameworks**
   * Apache Airflow
   * dbt (Data Build Tool)
   * Luigi, Prefect
2. **Data Ingestion**
   * Batch processing (CSV, Excel, JSON files)
   * Streaming ingestion (Apache Kafka, Amazon Kinesis)
3. **Data Transformation**
   * Data Cleaning (null handling, type casting)
   * Enrichment and Aggregations

**Phase 4: Big Data Technologies**

1. **Distributed Computing Basics**
   * What is Big Data? Hadoop vs Spark
2. **Apache Hadoop Ecosystem**
   * HDFS, MapReduce, Hive, Pig
3. **Apache Spark**
   * RDDs, DataFrames, SparkSQL
   * PySpark for transformation jobs

**Phase 5: Cloud Platforms**

1. **Cloud Basics**
   * IAM, Virtual Machines, Networking fundamentals
2. **Cloud Data Services**
   * **AWS**: S3, Redshift, Glue, Athena
   * **GCP**: BigQuery, Dataflow, Cloud Storage
   * **Azure**: Synapse, Data Lake, Data Factory
3. **Deployment & Automation**
   * CI/CD pipelines (GitHub Actions, Jenkins)
   * Terraform for infrastructure-as-code

**Phase 6: Orchestration & Workflow Automation**

* **Apache Airflow**
  + DAGs, scheduling, task dependencies
* **Prefect**
  + Modern alternative to Airflow
* **Cron Jobs**
  + Lightweight automation on Unix systems

**Phase 7: Monitoring & Data Quality**

* **Monitoring Tools**
  + Prometheus, Grafana, AWS CloudWatch
* **Data Quality Tools**
  + Great Expectations, Deequ
* **Logging**
  + ELK Stack (Elasticsearch, Logstash, Kibana)

**Phase 8: Real-time Data Processing**

* **Apache Kafka**
  + Topics, Producers, Consumers, Kafka Streams
* **Apache Flink**
  + Real-time stream processing

**Phase 9: Advanced Topics**

1. **Data Governance & Security**
   * GDPR, Data masking, encryption
2. **Data Catalogs**
   * Amundsen, DataHub, Apache Atlas
3. **Scalability & Optimization**
   * Partitioning, Indexing, Query tuning
4. **Data Lake vs Data Warehouse vs Lakehouse**
   * Concepts, architecture, and use-cases

**Phase 10: Projects & Portfolio**

* Build end-to-end pipelines:
  + Ingest data → Transform with Spark/SQL → Store in DW → Visualize in BI tools
* Host projects on GitHub
* Document and blog your architecture

**✅ Bonus: Tools to Master**

| **Category** | **Tools** |
| --- | --- |
| Databases | PostgreSQL, MySQL, MongoDB |
| Big Data | Apache Spark, Hadoop |
| Cloud | AWS, GCP, Azure |
| Workflow | Airflow, dbt, Prefect |
| DevOps | Docker, GitHub Actions, Terraform |
| Visualization | Looker, Power BI, Tableau (light use for testing outputs) |