Data Engineering Roadmap

A comprehensive guide on becoming a data engineer

Sairam Elangovan

Required Skillset

- SQL
- Linux basics and shell scripting
- Python/Java
- Cloud stack exposure (AWS, Azure, GCP)
- Distributed computing frameworks (Hadoop, Spark, Kafka etc)
- Data centric software architecture and terminologies
- NoSQL databases (dynamoDB, mongoDB, cassandra etc)
- Data orchestration tools (Apache Airflow, Oozie etc)

SQL

- Creating Databases and tables
- Basic commands - select, from and where clause in queries
- Grouping , having and aggregate functions
- Order by, with, in, in between
- Joining tables (left, right, inner, outer and self join of tables)
- Case statements, windowing, coalesce and sub query

Linux basics and shell scripting

- Linux file structure
- Shell terminal
- Basic commands for creating and navigating files
- Setting up cron jobs
- Network management
- Monitoring services and troubleshooting

Distributed frameworks

- HDFS
- Hadoop
- Apache Spark
- Apache Kafka

Hadoop

- HDFS file system
- Hadoop architecture
- MapReduce framework
- Apache Hive
- Apache Sqoop

Apache Spark

- Various installation modes
- Spark architecture
- Structured APIs - RDDs, dataset and dataframes
- SparkSQL
- Optimisations and tuning
- Spark Streaming
- Deploying Spark applications

Apache Kafka

- Installation for local and production grade deployments
- Architecture and producer consumer understanding
- Building a producer and consumer client
- Design considerations
- Security and authorisation
- Performance tests

Software Architecture and terminologies

- Lambda and kappa architecture
- Data lakes, warehouse, data mart
- ETL vs ELT
- Change data capture
- Data replication and Disaster recovery design
- Scaling Big data pipelines

NoSQL Databases

- NoSQL vs RDBMS
- Querying in NoSQL databases
- Scaling NoSQL databases
- Significance of CAP theorem and design considerations

Data Orchestration tools

- Apache Airflow/Oozie
- DAG
- Scheduling jobs
- Why orchestration tools are used instead of cron jobs?
- Checking logs and troubleshooting jobs