roadmap.sh

Data Engineering

Understanding of Databases
- Relational Databases (MySQL, Oracle)
- NoSQL Databases (MongoDB, Cassandra, Redis)
- Columnar Databases (Redshift, BigQuery)
- Graph Databases (Neo4j, TigerGraph)
- Time-series Databases (InfluxDB, TimescaleDB)
- Cloud based Databases (DynamoDB, Firestore)

Data Warehouse Concepts
- Designing Warehouse Schemas (Star Schema, Snowflake Schema)
- Data Modelling (ER Model, UML)
- ETL (Extract, Transform, Load) Process
- Data Warehousing Tools (Talend, Informatica)
- OLAP Operations
- MDM (Master Data Management)

Big Data Technologies
- Distributed Storage (HDFS, S3)
- Distributed Processing (MapReduce, Spark, Flink)
- Big Data Frameworks (Hadoop, Spark)
- Real-time Processing (Storm, Samza)
- Big Data Querying Tools (Hive, Pig)
- Resource Managers (YARN, Mesos)

Data Pipeline and Workflow Management
- ETL Tools (Talend, Pentaho)
- Pipeline Management Tools (Airflow, Luigi)
- Streaming Systems (Kafka, Kinesis)
- Scheduling Tools (Cron, Oozie)
- Monitoring Tools (Ganglia, Nagios)
- Error Handling Procedures

Data Lake Architecture
- Basics of Data Lake
- Storing Data in Data Lakes (Raw data, Processed Data)
- Data Lake vs Data Warehouse
- Managing Data Lakes (Cataloging, Security)
- On-premise vs Cloud-based Data Lakes
- Data Lake Tools (Azure Data Lake, AWS Lake Formation)

Data Visualization
- Knowledge of visualization tools (Tableau, PowerBI)
- Creating Interactive Dashboards
- Data Reporting Methods
- Understanding of Color Theory in Data Visualization
- Advanced Charting Techniques
- Data-driven Storytelling Techniques

Machine Learning
- Building Data Sets
- Supervised Learning Algorithms
- Unsupervised Learning Algorithms
- Reinforcement Learning Algorithms
- Evaluation Strategies and Testing
- ML Libraries and APIs (scikit-learn, TensorFlow)

Python Programming
- Core Python Concepts
- Advanced Python Concepts
- NumPy, SciPy, Pandas
- Web Scraping Libraries (BeautifulSoup, Selenium)
- Django and Flask Frameworks
- PyTest Testing Framework

Shell Scripting
- Basic Bash and Shell Scripts
- Command line tools (grep, awk, sed)
- Regular Expressions
- Process Monitoring and Management
- Shell Script Debugging
- Text Processing Tools

SQL
- SQL basics (SELECT, INSERT, UPDATE, DELETE)
- Advanced SQL (JOINs, VIEWs, INDEXes)
- Stored Procedures and Functions
- Database Normalization and Denormalization
- SQL Debugging
- SQL vs NoSQL

Cloud Computing
- Basics of Cloud Computing
- Understanding major cloud providers (AWS, Google Cloud, Azure)
- Cloud Storage and Computation Services
- Serverless Architecture
- Data Engineering on Cloud
- Cloud Security

Data Governance
- Data Quality Management
- Data Policy Management
- Data Lineage
- Data Cataloging
- Data Privacy and Compliance
- Metadata Management

DevOps
- Understanding of CI/CD Pipeline
- Knowledge of Tools (Jenkins, Docker, Kubernetes)
- Infrastructure as Code
- Monitoring and Log Management
- Cloud DevOps Practices
- Basics of Git and Version Control

Agile Methodologies
- Understanding Scrum and Kanban
- Agile Planning and Estimation
- Writing User Stories
- Sprints and Iterations
- Agile Metrics and Forecasting
- Agile Tools (Jira, Trello)

Security in Data Engineering
- Data Privacy and Compliance
- Encryption and Data Masking Techniques
- Secure Data Transmission and Storage
- Privacy by Design and by Default
- Data Access Control
- Security in Cloud Environments