Data Engineering Roadmap

1. Programming Language :
    a. Python
    b. Scala
    c. Java

2. Operating Systems & Scripting:
    a. Linux
    b. Unix
    c. Shell Scripting

3. Data Structures & Algorithms (Average Level, No Hard level):
    a. Arrays
    b. Strings
    c. Linked List
    d. Stack
    e. Queue
    f. Tree (Basics)
    g. Graph (Basics)
    h. Dynamic Programming
    i. Searching
    j. Sorting

4. Core Basics of DBMS :
    a. DDL
    b. DCL
    c. DML
    d. Integrity Constraints
    e. Data Schema
    f. Basic Operations
    g. ACID Properties
    h. Transactions

i. Concurrency Control
j. Deadlock
k. Indexing
l. Hashing
m. Normalization forms
n. Views
o. Stored Procedures
p. ER Diagrams

5. SQL Scripting :
    a. Transactional Databases : MySQL, PostgreSQL
    b. All types of joins
    c. Nested Queries
    d. Group By
    e. Use of Case When Statements
    f. Window Functions

6. Basic Terminologies In BigData :
    a. What is BigData?
    b. 5 V's of BigData
    c. Distributed Computation
    d. Distributed Storage
    e. Vertical vs Horizontal Scaling
    f. Commodity Hardwares
    g. Clusters
    h. File formats
        i. CSV
        ii. JSON
        iii. AVRO
        iv. Parquet
        v. ORC
    i. Type of Data

    i. Structured
    ii. Unstructured
    iii. Semi-structured

7. Data Exploration Libraries :
    a. Pandas
    b. NumPy

8. Data Warehousing Concepts:
    a. OLAP vs OLTP
    b. Dimension Tables
    c. Fact Tables
    d. Star Schema
    e. Snowflake Schema
    f. Warehouse Designing Questions
    g. Many more topics

9. BigData Frameworks :
    a. Apache Hadoop (Architecture Understanding Most Imp)
        i. HDFS
        ii. Map-Reduce
        iii. Yarn
    b. Apache Hive
        i. How to load data in different file formats
        ii. Internal Tables
        iii. External Tables
        iv. Querying table data stored in HDFS
        v. Partitioning
        vi. Bucketing
        vii. Map-Side Join
        viii. Sorted-Merge Join
        ix. UDF’s in Hive

x. SerDe in Hive

c. Apache Spark (Most Important)
    i. Spark Core
    ii. Spark SQL
    iii. Spark Streaming
d. Apache Flink (Real Time Data Processing / Stream Processing)
e. Apache SQOOP
f. Apache FLUME

10. Workflow Schedulers, Dependency Management :
    a. Apache Airflow
    b. Azkaban
    c. Apache NIFI

11. NoSQL Databases :
    a. HBase
    b. DataStax Cassandra (Recommended)
    c. ElasticSearch
    d. MongoDB

12. Messaging Queue Frameworks :
    a. Apache KAFKA

13. Dashboarding Tools :
    a. Tableau
    b. PowerBI
    c. Grafana
    d. Kibana (Part of ELK (ElasticSearch - Logstash - Kibana)

14. BigData Services in Cloud (AWS) :
    a. Ondemand Machines

    i. AWS EC2
b. Access Management
    i. AWS IAM
c. For Storing and Accessing Credentials
    i. AWS Secret Manager
d. Distributed File Storage
    i. AWS S3
e. Transactional Database Services
    i. AWS RDS
    ii. AWS Athena
    iii. AWS Redshift (Data Warehousing)
f. NoSQL Database Services
    i. AWS Dynamo
g. Serverless
    i. AWS Lambda
h. ETL Services
    i. AWS Glue
i. Scheduler
    i. AWS Cloudwatch
j. Distributed Data Computation
    i. AWS EMR
k. Messaging Queue
    i. AWS SNS
    ii. AWS SQS
l. Real Time Data Processing
    i. AWS Kinesis