**Data Engineering Roadmap For Freshers & Resources**

***1) Programming Language :***

* 1. Python
  2. Scala
  3. Java

***2) Operating Systems & Scripting:***

* 1. Linux
  2. Unix
  3. Shell Scripting

***3) Data Structures & Algorithms (No Hard level):***

* 1. Arrays
  2. Strings
  3. Linked List
  4. Stack
  5. Queue
  6. Tree
  7. Graph (Basics)
  8. Dynamic Programming
  9. Searching
  10. Sorting

***4) Core Basics of DBMS :***

* 1. DDL
  2. DCL
  3. DML
  4. Integrity Constraints
  5. Data Schema
  6. Basic Operations
  7. ACID Properties
  8. Transactions
  9. Concurrency Control
  10. Deadlock
  11. Indexing
  12. Hashing
  13. Normalization forms
  14. Views
  15. Stored Procedures
  16. ER Diagrams

***5) SQL Scripting :***

* 1. Transactional Databases : MySQL, PostgreSQL
  2. All types of joins
  3. Nested Queries
  4. Group By
  5. Use of Case When Statements
  6. Window Functions

***6) Basic Terminologies In BigData :***

* 1. What is BigData?
  2. 5 V’s of BigData
  3. Distributed Computation
  4. Distributed Storage
  5. Vertical vs Horizontal Scaling
  6. Commodity Hardwares
  7. Clusters
  8. File formats

CSV

JSON

AVRO

Parquet

ORC

i Type of Data

Structured

Unstructured

Semi-structured

***7) Data Exploration Libraries :***

* 1. Pandas
  2. NumPy

***8) Data Warehousing Concepts:***

* 1. OLAP vs OLTP
  2. Dimension Tables
  3. Fact Tables
  4. Star Schema
  5. Snowflake Schema
  6. Warehouse Designing Questions
  7. Many more topics

***9) BigData Frameworks :***

* 1. Apache Hadoop (Architecture Understanding Most Imp)
     1. HDFS
     2. Mar-Reduce
     3. Yarn
  2. Apache Hive
     1. How to load data in different file formats
     2. Internal Tables
     3. External Tables
     4. Querying table data stored in HDFS
     5. Partitioning
     6. Bucketing
     7. Map-Side Join
     8. Sorted-Merge Join
     9. UDF’s in Hive
     10. SerDe in Hive
  3. Apache Spark (Most Important)
     1. Spark Core
     2. Spark SQL
     3. Spark Streaming
  4. Apache SQOOP
  5. Apache NIFI
  6. Apache FLUME

***10) Workflow Schedulers, Dependency Management :***

* 1. Apache Airflow
  2. Azkaban

***11) NoSQL Databases :***

* 1. HBase
  2. DataStax Cassandra (Recommended)
  3. ElasticSearch
  4. MongoDB

***12) Messaging Queue Frameworks :***

* 1. Apache KAFKA

***13) Dashboarding*** ***Tools :***

* 1. Tableau
  2. PowerBI
  3. Grafana
  4. Kibana (Part of ELK (ElasticSearch - Logstash - Kibana)

***14) BigData Services in Cloud (AWS) :***

* 1. Ondemand Machines
     1. AWS EC2
  2. Access Management
     1. AWS IAM
  3. For Storing and Accessing Credentials
     1. AWS Secret Manager
  4. Distributed File Storage
     1. AWS S3
  5. Transactional Database Services
     1. AWS RDS
     2. AWS Athena
     3. AWS Redshift
  6. NoSQL Database Services
     1. AWS Dynamo
  7. Serverless
     1. AWS Lambda
  8. ETL Services
     1. AWS Glue
  9. Scheduler
     1. AWS Cloudwatch
  10. Distributed Data Computation
      1. AWS EMR
  11. Messaging Queue
      1. AWS SNS
      2. AWS SQS
  12. Real Time Data Processing
      1. AWS Kinesis