✦ Member-only story

The Ultimate Roadmap for Data Engineers in 2024

Navigating the Data Engineering Landscape in 2024: A Comprehensive Roadmap

Vishal Barvaliya · Follow

4 min read · Dec 21, 2023

Listen Share

Welcome to the dynamic world of data engineering, where each line of code shapes the future of information. As we are going to step out on the journey of 2024, the demand for skilled data engineers is growing faster than ever. In this blog, we'll reveal the layers of the ultimate roadmap for eager newcomers through the essential skills that define the data engineering.

In this blog we will see all the important topics to learn about below skills in detail:

1. SQL
2. Python
3. DBMS
4. Big Data Terminologies

5. Data Warehousing
6. Big Data Frameworks
7. NoSQL Databases
8. Cloud Services

We will breakdown each skills in detail to make it more easy for everyone!

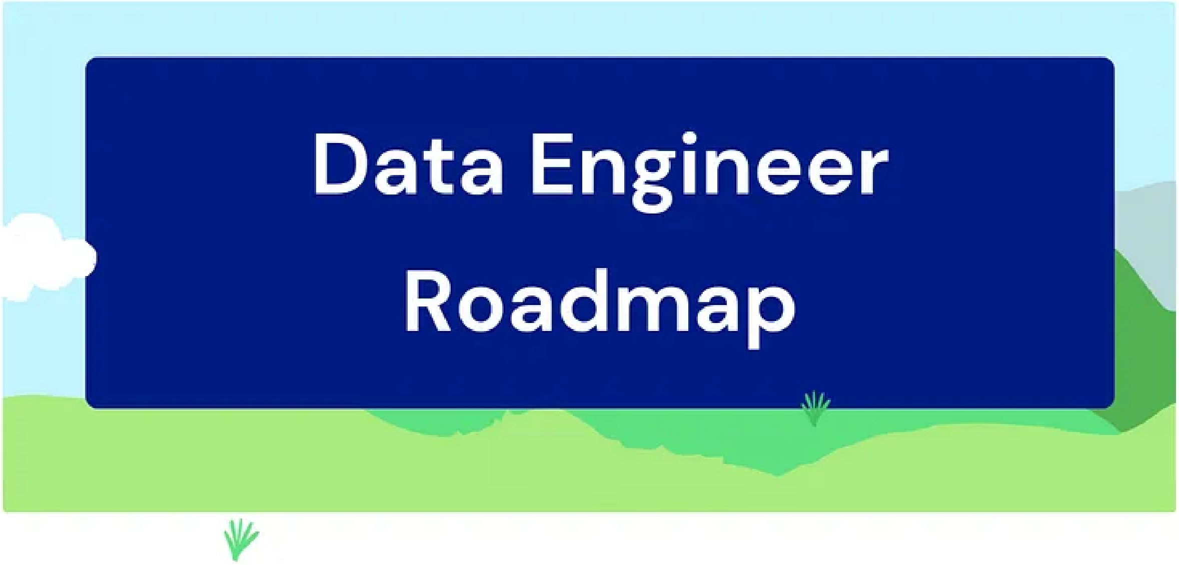

Image by Author

1. SQL:

one of the most important skills for data engineer is SQL, so once should be master at SQL to become a good data engineer and crack interviews. Below are the important topics one should prepare for.

- **DDL Commands:** Create, Alter, Drop, Truncate, Rename.
- **DML Commands**: Select, Insert, Update, Delete.
- **DCL Commands**: Grant, Revoke.
- **TCL Commands**: Commit, Rollback, Save Point.

- **All types of joins:** Cross Join, Inner Join, Left Join, Right Join, Full Outer Join.
- **Where** Clause and **Order by** Clause.
- **Group by** and **Having** Clause.
- **Case When Statement** (with group by as well)
- **SubQueries** and **Nested SubQueries**.
- In, Not In, Any, All, Exists, Not Exists.
- **Aggregation & Date** Functions.
- **Common Table Expressions:** Iterative, Recursive.

Window Functions:

- Over Clause, Partition by & order by
- Count, Sum, Min, Max, Avg.
- Row_number, Rank, Dense_rank.
- Lead, Lag.
- Nth_Value.
- Frame Clause: Range Between, Rows Between.

Most asked SQL interview questions in Data Engineering Interviews (Part II)

This is the Second part of my interview questions series. In this part, we will discuss another 15 most...

medium.com

2. Python:

- Input/ Output
- Command Line Arguments (Sys Library).
- Data Types: String, List, Tuple, Set, Dictionary, List Comprehension, Dictionary Comprehensions.
- Loops: for loop, while loop.
- If condition.
- Logical and Mathematical operators.
- Functions: passing and accessing arguments, KWARGS, return single and multiple values.
- Lambda Function
- Exception Handling
- File Handling

Data Structures: List, Tuple, Dictionary, Set, Linked List, Stack, Queues, Tree (Basics), Graph (Basics)

Algorithms: Searching, Sorting

Important Note:

Make sure to Practice only **Easy to medium (Mostly Easy)** level of coding problems on **Leetcode** or **HackerRank**. (Practice only string, List and dictionary related problems) Please don’t waste your time doing hard problems or many Medium Problems.
Solving about 10 easy and 3–5 medium problems for each data structure is more than sufficient.

Must-do Python topics for Data Engineers

Python is the most commonly used programming language for data engineering. Python has a large and…

medium.com

3. Basics of DBMS:

- Integrity Constraints
- Data Schema
- Basic Operations
- ACID Properties
- Transactions
- Concurrency Control
- Deadlock
- Indexing
- Hashing
- Normalization forms
- Views
- Stored Procedures
- ER Diagrams

. . .

4. Big Data Terminologies:

- What is Big Data?
- 5 V’s of Big Data
- Distributed Computation
- Distributed Storage
- Vertical vs Horizontal Scaling
- Commodity Hardware
- Clusters
- File formats : CSV, JSON, AVRO, Parquet, ORC.
- Types of Data: Structured, Unstructured, Semi-structured

. . .

5. Data Warehousing Concepts:

- OLTP
- OLAP
- Dimension tables
- Fact Tables
- Star Schema
- Snowflake Schema
- Data Warehouse designing.

“Data Modelling: Design the data model for a food delivery service like Uber.”

Introduction

medium.com

6. Big Data Frameworks:

Apache Hadoop: (Only need to understand Architecture)

- **HDFS**
- **Map-Reduce** (don’t waste time in map-reduce code as it’s outdated)
- **YARN**

Apache Hive:

- Loading data from various file formats
- Internal vs external tables
- partitioning
- bucketing
- map-side join
- sorted-merge join
- UDF
- SerDe

Apache Spark (Most Most Important for Data Engineer):

- RDDs (Resilient Distributed Datasets)

- Transformations and Actions
- Spark Context
- Shared Variables
- Caching
- DataFrame and Dataset API
- Spark SQL Functions
- Joins and Aggregations
- Window Functions
- Partitioning and Shuffling
- Broadcast Variables and Accumulators
- Understanding Spark Cluster & Cluster Modes
- How Spark Executes Program on the Cluster
- Stages in Spark
- Accumulators & Broadcast Variables
- Repartition Vs Coalesce
- Lazy Evaluation
- Narrow Vs Wide Transformations
- Spark Storage Levels
- Cache Vs Persist
- Spark Optimization Techniques
- File Formats — Parquet | ORC | Avro

- Compression Techniques
- Understanding Cluster Configurations
- How to Submit Spark Job Scheduling and Running Spark Jobs
- Sort Vs Hash Aggregate
- Spark Catalyst Optimizer

Must-Do Apache Spark Topics for Data Engineering Interviews

Apache Spark is an open-source big data processing framework that provides a flexible and powerful...

medium.com

7. NoSQL Databases :

- you need to learn NoSQL Database like HBase, Cassandra(Recommended), Cosmos DB, and many more.

8. Cloud Services:

Azure:

Azure is the most preferred cloud by data engineers. below are some most important services you should learn.

- Azure Data Factory
- Azure Synapse
- Azure Databricks
- Azure Blob Storage

- Azure Data Lake Storage Gen2
- Azure SQL Database

5 Essential Azure Services for Data Engineers in 2023

Data engineering is a field that deals with managing, processing, and storing data. With the advent of big...

medium.com

AWS

- Amazon S3
- AWS Glue
- Amazon RedShift
- Amazon EMR
- AWS Lambda.

GCP

- Google Cloud Data Fusion
- Google BigQuery
- Google Cloud Dataproc
- Google Cloud Storage
- Google Cloud SQL

. . .

> **Best of luck with your journey!!!**