Data Engineering Interview Preparation Plan

Overview

This comprehensive guide is designed to help data engineering candidates crack interviews at product-based companies. It includes round-wise preparation strategies, topic breakdowns, important resources, and scenario-based questions with practical and theoretical insights.

Round 1: Online Coding Round (SQL, Python, Data Structures)

1. SQL Mastery

Topics to Focus:

- Window Functions
- GROUP BY, GROUP_CONCAT
- CTE - Recursive CTE
- Complex Joins (Self, Inner, Left Joins)
- SQL Execution Order
- Date Functions (DATEDIFF, TO_DATE)
- SQL Optimization Techniques
- HAVING vs WHERE, Ranking Queries

Resources:

- W3Schools SQL Tutorial
- GeeksforGeeks SQL Tutorial
- Practice Platforms: LeetCode → Datalemur → StrataScratch
- YouTube Tutorial Playlist

2. Python Proficiency

Topics to Cover:

- OOP Concepts, Data Structures (List, Tuple, Set, Dict, NamedTuples)
- File Handling, Exception Handling
- Lambda Functions, Decorators, Generators, Iterators
- Memory Management
- Multithreading vs Multiprocessing

Common Interview Questions:

- Tuple vs List
- Generators and Concurrency

Resources:

- GeeksforGeeks Python Tutorial

3. Data Structures Mastery

DSA Topics & Techniques for Interviews

Level 1 - Most Common (70% of Companies Focus Here):

- **Arrays**
    - Searching: Linear & Binary Search
    - Sorting Techniques
    - Two Pointer Approach
    - Sliding Window Technique
    - Kadane's Algorithm (Maximum Subarray)
    - Dutch National Flag Problem (3-way Partition)
    - **Practice:** Top 50 Array Coding Problems (Easy/Medium)
- **Strings**
    - Palindromes, Anagrams, Substrings/Subsequences
    - String Manipulation and Pattern Matching
    - **Practice:** Top 50 String Coding Problems (Easy/Medium)
- **HashMap / HashSet**
    - Frequency Counting
    - Finding Duplicates / Unique Elements
    - Subarray Sum problems
    - **Practice:** Top 50 HashMap Problems (Easy/Medium)

Level 2 - Intermediate (Asked by Selected Companies):

- **Recursion & Backtracking**
    - Permutations, Combinations
    - Subsets, N-Queens, Sudoku Solver
- **Dynamic Programming (DP)**
    - 1D DP: Fibonacci, Climbing Stairs, House Robber
    - 2D DP: Grid Problems, Knapsack, DP on Strings
    - Memoization vs Tabulation

Level 3 - Advanced (Few Companies, Mostly Later Rounds):

- **Graphs**
  - BFS & DFS
  - Union-Find
  - Topological Sorting
  - Dijkstra's / Bellman-Ford / Floyd-Warshall

Additional Tips:

- Always Know: Time & Space Complexity of every approach
- Understand Trade-offs between brute-force and optimized solutions

Practice Platform:

- LeetCode Algorithms Problems (Easy/Medium)

Round 2: Big Data Concepts

1. Hadoop

Topics:

- Architecture (HDFS, YARN)
- High Availability in HDFS
- MapReduce v1 vs v2
- HDFS Block Size Example: 10GB file, 128MB block
- Core Config Files: mapred.xml, core-site.xml

2. Apache Spark

Core Concepts:

- RDD, DataFrames, Datasets (Type Safety)
- Spark Job Lifecycle (Job → Stage → Task)
- Caching, Repartition, Coalesce
- Storage: Row vs Columnar, File Formats (CSV, ORC, Parquet, JSON, Avro)
- SparkSession vs SparkContext
- AQE (Adaptive Query Execution - Spark 3)
- Spark Web UI
- Spark memory management
- Spark join strategies
- Resource allocation in spark

- Spark plans
- Optimization of jobs
- File formats

Optimization Topics:

- Skewness, Salting
- Join Strategies
- Resource Allocation (Thin/Thick Executors)

Resources:

- Theory - https://www.youtube.com/watch?v=FNMoeE849Yw&list=PLTsNSGeIpGnGkpfKMf7ilFmzfx6AjMKyT
- Practical - https://www.youtube.com/watch?v=FNMoeE849Yw&list=PLTsNSGeIpGnGjaMSYVlidqVWSjKWoBhbr
- Interview Questions - https://www.youtube.com/@DataSavvy

  https://www.youtube.com/watch?v=ZirbI1355B8&list=PL9sbKmQTkW05mXqnq1vrrT8pCsEa53std
- Scenario Base Questions - https://medium.com/@goyalarchana17

  https://medium.com/@shrutighoradkar101/spark-scenario-based-interview-questions-1fd3485c2911

  https://medium.com/@shrutighoradkar101/spark-scenario-based-interview-questions-part-ii-cebc145e32c2

  https://www.youtube.com/@AzarudeenShahul/playlists

3. Data Warehousing Concepts

Topics:

- Star vs Snowflake Schema
- OLTP vs OLAP
- SCD, Fact vs Dimension
- Normalization vs Denormalization
- Indexing, Clustering

- Physical, Logical, Conceptual Design

Resources:

- https://www.javatpoint.com/what-is-database
- https://www.tutorialspoint.com/dwh/index.htm

Round 3: Design Round

1. Schema Design

Practice System Examples:

- Instagram/Facebook, LinkedIn
- OTT Platform, Food Delivery App
- Music Streaming App, Airbnb

Checklist:

- Entities → Attributes → Data Types → Relationships
- PK, FK, Indexing, Partitioning
- Nested Schemas, Audit Trails

Resources:

Schema Design - https://www.youtube.com/@DatabaseStar/featured

Database Design for a system like LinkedIn

https://medium.com/towards-data-engineering/database-design-for-a-system-like-linkedin-3c52a5ab28c0

Database Design for a food delivery app like Zomato/Swiggy

https://medium.com/towards-data-engineering/database-design-for-a-food-delivery-app-like-zomato-swiggy-86c16319b5c5

Data Modeling: Design a data model for a hotel booking system like Airbnb.

https://medium.com/towards-data-engineering/data-modelling-design-a-data-model-for-a-hotel-booking-system-like-airbnb-2110a6d079c6

Data Modeling: Design the data model for a taxi service like Uber

https://medium.com/towards-data-engineering/data-modelling-design-the-data-model-for-a-taxi-service-like-uber-eaedfa0e25f4

Design a database to store large amounts of historical data that needs to be regularly analyzed.

https://medium.com/towards-data-engineering/design-a-database-to-store-large-amounts-of-historical-data-that-needs-to-be-regularly-analyzed-dd6558e1bd02

3. ETL Design

ETL Pipeline Stages:

- Requirements Gathering
- Data Extraction, Transformation
- Data Quality, GDPR, Loading
- Monitoring, Logging, Testing
- Incremental Loads, Idempotency

Scenario-Based Questions:

- Real-Time Streaming ETL
- Incremental Load Optimization
- Handling Schema Evolution
- Data Quality & Error Handling
- Dependency Management
- Delta Lake & CDC
- Data Security (Encryption at rest, in-transit)
- Data Lineage and Auditing
- Windowed Aggregations

ETL Design Scenarios-

Scenario: Real-time Data Streaming

Problem: You are tasked with designing an ETL pipeline for a system that receives real-time streaming data from multiple sources. How would you approach the design to ensure efficient extraction, transformation, and loading of this streaming data into a data warehouse? Consider factors such as data consistency, latency, and scalability.

Scenario: Incremental Load Optimization

Problem: Assume you have a large dataset, and the source data is updated frequently. Design an ETL strategy that optimizes the incremental load process to update only the changed or new records. Discuss techniques and technologies you would employ to achieve this while minimizing the impact on performance.

Scenario: Handling Schema Changes

Problem: In a dynamic environment, source systems may undergo schema changes over time. How would you design an ETL system that gracefully handles these changes without causing disruptions to downstream processes? Consider scenarios where both the structure and semantics of the data may evolve.

Scenario: Data Quality and Error Handling

Problem: Ensuring data quality is crucial in ETL processes. Design an approach to identify and handle data quality issues during the transformation phase. Discuss strategies for detecting anomalies, handling missing values, and implementing error handling mechanisms to maintain data integrity.

Scenario: Dependency Management

Problem: ETL processes often have dependencies on external systems, services, or APIs. How would you design a robust ETL system that manages dependencies effectively, considering potential failures, retries, and ensuring data consistency across multiple interconnected systems? Discuss strategies for handling dependency failures and ensuring data completeness.

Scenario: Delta Lake and Change Data Capture (CDC)

Problem: Explain how Delta Lake and Change Data Capture (CDC) can be utilized in an ETL pipeline. Discuss the benefits of using Delta Lake for handling data versioning, transactional capabilities, and how CDC techniques can be employed for capturing and propagating changes in the source data.

Scenario: Data Encryption and Security

Problem: In the context of ETL, sensitive data may need to be protected. Discuss strategies for implementing data encryption and ensuring security throughout the ETL process. Consider encryption at rest, in transit, and access controls to safeguard sensitive information.

Scenario: Data Lineage and Auditing

Problem: Design a system for tracking data lineage in an ETL pipeline. How would you implement auditing mechanisms to trace the origin and transformations applied to each piece of data? Discuss the importance of data lineage for compliance, troubleshooting, and maintaining a transparent data flow.

Scenario: Windowed Aggregations(Streaming)

Problem: Suppose you need to perform windowed aggregations on streaming data (e.g., calculating hourly averages). How would you design an ETL pipeline to efficiently handle these windowed aggregations? Discuss the considerations for defining windows, managing state, and handling late-arriving data.

Blogs & Resources:

- MakeMyTrip Engineering Blog
- Netflix Tech Blog
- Airbnb Engineering Blog

Round 4: Hiring Manager Round

1. Project Demonstration

- Emphasize real-world impact, collaboration
- Discuss architecture, tools used, challenges faced
- Showcase metrics and outcomes

2. Behavioral Questions

- Use STAR Format (Situation, Task, Action, Result)
- Leadership, conflict resolution, team collaboration

3. Values and Culture Fit

- Research company culture and values
- Align personal motivations with the company's mission

Modern Data Engineering Tech Stack

- Databricks
- Snowflake
- DBT (Data Build Tool)
- Apache Airflow
- Generative AI & Prompt Engineering

Final Preparation Tips

Review Plan

- Focus on weak topics
- Revisit tricky concepts
- Build a revision schedule

Mock Interviews

- Practice with peers
- Record and review answers
- Refine articulation and technical depth

Mindset & Confidence

- Reflect on your journey and progress
- Stay calm, confident, and curious
- Acknowledge the efforts you've put in

Best of luck with your data engineering interview!
Curated and Compiled by Ankita Gulati
Connect with me: linkedin.com/in/ankita-gulati-de