

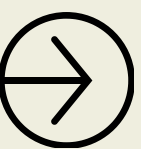


# Data Engineering Interview Questions



Ankita Gulati

Shubh Goyal



# Job Details

- **Position:** Data Engineer
- **Experience:** 3 years
- **Location:** Bengaluru
- **Work mode:** Hybrid
- **Compensation:** ₹18–22 LPA
- **Total Rounds:** 3
- **Top Required Skills:**
  - Advanced SQL
  - Python Programming
  - Big Data Technologies
  - Behavioral & Ownership

# Round 1

## Technical Coding / SQL / Python

1. SQL – Second Highest Salary
  - Write a query to get the 2nd highest employee salary from a Salaries table.
  - Follow-up: How would you handle if multiple employees share the same salary?
2. SQL – Employee-Manager Join
  - Question: Given Employee and Manager tables, write a query to show the employee → manager mapping.
  - Follow-up: Extend query to find employees without managers.
3. Python / DSA – Prime Number Check
  - Question: Write code to check if a number is prime.
  - Follow-up: Optimize it (e.g., loop till  $\sqrt{n}$ , skip even numbers).
4. Python Data Structures
  - Question: Compare list, set, dict, tuple.
  - Follow-up: In which cases would you use each? What are time complexities of lookup/insertion?
5. Spark / Big Data Concepts
  - Discussion:
    - What is Spark? Explain transformations vs actions.
    - What happens when an out-of-memory (OOM) error occurs in Spark? How would you fix it (partition tuning, caching, spill to disk, executor memory)?

# Round 2

## Advanced Technical / Deep Dive

### 1. Spark Optimization / Cluster Sizing

- You need to process multiple TBs of data in Spark. How do you decide number of executors, executor memory and partition count
- Follow-up: If the job is running slowly, what optimizations would you try (broadcast joins, caching, AQE, repartition)?

### 2. Advanced SQL – Window Functions

- Write a query to get each employee's salary difference compared to their department average.
- Follow-up: Use `ROW_NUMBER()` or `DENSE_RANK()` to get top 3 salaries per department.

### 3. Spark + Python Coding

- Write Python code to reverse a string without built-in reverse.
- Follow-up: Optimize for memory. Rewrite using Spark DataFrame API if data is large.

### 4. Python Data Structures – Performance

- Difference between list vs tuple vs set vs dict.
- Why are sets faster for membership checks?
- How dicts are implemented internally in Python?

### 5. DSA

- Given an array of integers, return indices of two numbers such that they add up to a target.
- Follow-up: Optimize from  $O(n^2)$  brute force  $\rightarrow O(n)$  using hashmap.

# Round 3

## Hiring Manager

### 1. Project Deep Dive

- “Tell me about a Spark job you worked on. What bottlenecks did you face? How did you optimize (e.g., repartition, broadcast join, caching)?”

### 2. Scenario / Trade-Off Questions

- When would you use PySpark vs SQL vs Hive?
- If cluster cost is too high, how would you balance cost vs performance?

### 3. Behavioral / Situational

- How do you communicate with cross-team members (product, business, infra)?
- If you forgot PySpark syntax in an interview/test, how would you handle it?
- How do you adapt when new tech (e.g., Databricks, Delta Lake) is introduced at work?

### 4. Other Technical Manager Questions

- Explain Spark cluster configuration: driver, executors, cores, memory.
- How to handle OOM exceptions at executor level.
- How do joins and window functions impact Spark performance?

*Thank You*

Best of luck with your  
upcoming interviews  
— you've got this!

