



Data Engineering Interview Questions



Ankita Gulati

Shubh Goyal



Job Details

- **Position:** Senior Data Engineer
- **Experience:** 3 years
- **Location:** Delhi
- **Work mode:** Hybrid
- **Compensation:** ₹20+ LPA
- **Total Rounds:** 5
- **Top Required Skills:**
 1. SQL
 2. PySpark / Python
 3. Cloud Data Engineering
 4. ETL / Data Modeling
 5. Big Data & Streaming
 6. System Design

Round 1

Online Assessment

1. Python Coding

- Write a function to calculate the sum of all even numbers in a list.

2. Regex

- Write a regular expression to validate email addresses.

3. SQL Queries

- Write a query to find the second-highest salary in each department.
- Write a query using GROUP BY and HAVING to aggregate sales per category.
- Write a query with a correlated subquery to identify employees earning above their department's average salary.

Round 2

Technical Telephonic Interview

1. Walk me through your past projects and your contributions.
2. **SQL**: Explain the difference between ROW_NUMBER(), RANK(), and DENSE_RANK() with examples.
3. **Spark**: How would you process and optimize large-scale datasets using Apache Spark?
4. **NoSQL**: What are the advantages of using Cassandra over traditional relational databases?
5. What challenges have you faced while working with Spark DataFrames or RDDs?

Round 3

Machine Coding

Format: Cloud-based environment with datasets

Task:

- Given an F1 race dataset in CSV format, perform:
 - 1. Data cleaning (handle missing/nulls, type casting).**
 - 2. Transformations (aggregate results by driver/team).**
 - 3. Analytics (find top drivers per season).**

Expected Approach:

Use PySpark DataFrame API for reading, filtering, grouping, and aggregating data.

Preparation Tip:

Practice PySpark data wrangling tasks and ensure efficiency with DataFrames.

Round 4

Technical Discussion

1. **SQL:** Given a user activity table, write a query to find the most active users.
2. **DSA:** Implement an algorithm to detect a cycle in a singly linked list.
3. **DSA/System Design:** Given a sorted array, write a function to construct a balanced binary search tree (BST).
4. Walk through your approach to SQL query optimization.
 - How would you reduce query runtime on billions of rows?

Round 5

Behavioral Interview

1. Describe a time when you led a team through a challenging project.
2. How do you approach problem-solving under ambiguity?
3. A client reports data quality issues in their pipeline. How would you investigate and resolve this?
4. How do you handle conflict resolution and cross-team collaboration?

Thank You

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

