eClerx

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹20-22 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. Python
3. AWS
4. System Design & Problem-Solving

Ankita Gulati

Shubh Goyal

Round 1
Core Eng.& Problem-Solving

1. Every project has one critical issue that consumes significant resources to identify and resolve. What was the most challenging issue you faced in your project, and how did you handle it?
2. When making a POST API call, what should be the grant_type value?
3. Can you briefly explain all the different HTTP methods and their use cases?
4. What does a 401 status code mean in HTTP response?
5. Given an input list [A, B, C, D], create another list such that the output is [A, BB, CCC, D].
6. You have a list containing both numeric and string values. Write logic to print only the string values.
7. If Table A has 4 rows with value 1 and Table B also has 4 rows with value 1, how many rows will you get for INNER JOIN, LEFT JOIN, RIGHT JOIN, and FULL JOIN?

Ankita Gulati

Shubh Goyal

8. Demonstrate real-time use cases of List Comprehension in Python.

9. How can you optimize the performance of a Python function?

10. What is the difference between cache and persist in PySpark, and when would you use one over the other?

11. Given coin denominations [1, 2, 5] and a target amount 11, write a program to find the minimum number of coins required. Output should be 3 (coins used: [5, 5, 1]).

12. You are getting an Index Out of Range error when solving the coin problem with [1, 2, 5]. How would you debug and fix it?

13. Compare CSV, JSON, and Parquet file formats. Which one do you generally prefer and why?

14. Have you implemented Slowly Changing Dimensions (SCD) Types using Python, PySpark, or SQL? Explain the approach.

15. Explain Normalization and Denormalization in databases with examples.

Ankita Gulati Shubh Goyal

16. What are Indexes in SQL? How do they improve performance?
17. Suppose you have two tables with 800 million records each. How would you optimize the join between them?
18. Given an employee table with duplicate rows, write SQL logic to delete duplicates while keeping only one record.
19. You are given two tables with NULL and duplicate values. How would LEFT, RIGHT, INNER, FULL, and CROSS JOIN behave in this case?
20. Write an SQL query to find the complete employee hierarchy in an organization.
21. Given Month 1 sales = 1000 and Month 2 sales = 1500, calculate the percentage growth Month-over-Month.

Round 2
Foundations + Engineering Depth

1. What is the difference between AWS Glue and AWS Lambda? When would you use one over the other?
2. How is caching useful in an ETL setup, apart from reporting use cases?
3. In large-scale systems where data cannot be reduced and joins cannot be avoided, what strategies would you use to optimize performance?
4. Queries sometimes work efficiently in the beginning but later slow down. What could be the reasons, and how would you troubleshoot?
5. Share a real-world example where you used query optimization techniques. What was the issue, and how did you resolve it?
6. How would you design a solution to handle slow-performing queries caused by missing indexes, inefficient joins, or data skew?
7. You are using Airflow for orchestration. Explain your workflow design and the destination of processed data.

Ankita Gulati Shubh Goyal

8. How would you design a scalable data pipeline that handles both batch and real-time streaming data?

9. Explain how you would implement incremental data loading for large datasets in Redshift or Snowflake.

10. How would you handle schema evolution in a data lake when working with Parquet/Avro formats?

11. Describe a scenario where you applied partitioning and bucketing in Spark/SQL to improve query performance.

12. In a distributed setup, how do you identify and fix data skew issues?

13. How would you implement data validation and reconciliation in your ETL pipeline?

14. Explain how you would design an idempotent ETL job to avoid duplicates in target tables.

15. What strategies would you use to reduce AWS Glue job costs while processing TBs of data daily?

Thank You

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

Ankita Gulati

Shubh Goyal