Bitwise®

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ years
- **Location:** Pan India
- **Work mode:** Remote
- **Compensation:** ₹15-18 LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming
6. System Design

Ankita Gulati

Shubh Goyal

Round 1
Core Skills & Hands On

1. Can you introduce yourself and walk me through the projects you've worked on, including the technologies you used?
2. Can you explain one project in detail, including the architecture, data flow, and your specific contributions?
3. What were the biggest challenges in your projects, and how did you overcome them?
4. Given an Employee table with columns (name, dept, salary), write a query to find the 2nd highest salary in each department.
5. If you solved the above using DENSE_RANK(), why not use ROW_NUMBER()? What are the differences between them?
6. Given a Transaction table (trans_id, trans_date, trans_amt), write a query to add a cumulative monthly transaction amount column.

7. Describe the projects where you've used Spark/PySpark. What problem did Spark solve for you?
8. How do you manage and monitor Spark clusters in production?
9. Solve the 2nd highest salary per department problem using PySpark DataFrame API.
10. Explain partitioning and bucketing in Hive/Spark. When would you use each?
11. What are narrow vs wide transformations in Spark? Give examples.
12. How do you optimize PySpark jobs to handle skewed data?

Ankita Gulati

Shubh Goyal

Round 2
Client Interview

1. Provide a brief introduction and describe your projects, focusing on architecture and design choices.
2. What were the trade-offs you had to make between performance, cost, and scalability in your pipelines?
3. How do you ensure data quality, reliability, and consistency in your projects?
4. What is the difference between AWS Lambda and AWS Glue? When would you choose one over the other?
5. Explain the different AWS S3 storage classes and their use cases.
6. How do you use AWS Step Functions in a data pipeline?
7. If a PySpark script is running for a long time on EMR, how would you identify bottlenecks and optimize the job?

Ankita Gulati Shubh Goyal

8. Given two tables (Table1 & Table2), predict the output for Inner, Left, Right, and Cross Joins.

- Table 1: ID → 1,1,1,1,2,2,2,NULL
- Table 2: ID → 1,1,2,2,2,2,3,3

9. Given an Employee table with schema (Id, Name, Salary, ManagerId), write a query to find employees who earn more than their managers.

10. Write a function to check if two given strings are anagrams.

11. Write a function to merge two strings alternating characters.

→ Example: word1 = "abc", word2 = "pqr" → Output: "apbqcr"

→ Example: word1 = "ab", word2 = "pqrs" → Output: "apbqrs"

Round 3
Managerial & Behavioral

1. Describe yourself and summarize your professional experience in data engineering.
2. Walk me through one project where you had a significant impact on delivery.
3. Why are you looking for a change?
4. What challenges have you faced in your projects, and how did you overcome them?
5. Describe a failure you experienced and the lessons you learned.
6. How do you handle conflicts and disagreements within a team?
7. Explain the CAP theorem and how it applies to distributed databases.
8. What is eventual consistency? Can you give an example from cloud storage systems?
9. How do you design a data pipeline that handles both batch and streaming workloads?

Ankita Gulati Shubh Goyal

10. What is schema evolution in Spark, and how do you handle it in production?

11. How would you implement Change Data Capture (CDC) in a cloud-based data pipeline?

12. Compare Redshift vs Snowflake vs BigQuery. Which one would you recommend for Moonfare-like use cases?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal