Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹23-25 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. Airflow

Ankita Gulati

Shubh Goyal

Round 1
SQL & Database Fundamentals

1. Write the SQL syntax to create the following table:
→ Team = A, B, C, D
2. What are some basic SQL query optimization techniques you follow?
3. Write a SQL query to retrieve the count of distinct non-null values in a specific column of a table.
4. Write a SQL query to find the second-largest word in the string: "I'm a good programmer". (Answer should be "good")
5. Write a SQL query to find the most frequently ordered products for each customer. The data is in the orders table with columns: order_id, date, customer_id, product_id.
6. Given the following dataset:
→ Name = A, B, C, D
→ Salary = 10, 20, 20, 40
• What will be the output for ROW_NUMBER, RANK, and DENSE_RANK functions?
• Write a query to find the 2nd highest salary.

Ankita Gulati Shubh Goyal

7. Suppose you have two tables:

→ Table 1 (ID) = 1, 2, 3, 4, 5

→ Table 2 (ID) = 1, 2, 3

How many rows will be returned for the following joins?

- Inner Join
- Left Join
- Right Join
- Cross Join
- Full Outer Join
- Semi Join
- Left Semi Join

8. You have an Employee table with columns (EmpID, Name, Salary, DeptID). Write a SQL query to find the highest-paid employee in each department.

9. In a sales table (OrderID, CustomerID, OrderDate, Amount), write a query to:

- Find the top 3 customers by total sales amount.
- Also return ties if multiple customers have the same sales amount.

10. Given a Transaction table (TxnID, UserID, Amount, TxnDate), write a query to detect customers who made consecutive transactions on the same date.

11. Suppose you have a table of website clicks (UserID, Page, ClickTime). Write a query to calculate the average session time for each user. (Assume a session ends after 30 minutes of inactivity.)

12. You have a Customer table with duplicate records (same CustomerID appearing multiple times). Write a query to remove duplicates while keeping only the most recent record (based on LastUpdated).

13. You have an Orders table with (OrderID, OrderDate, DeliveryDate). Write a query to calculate the average delivery time (in days) for each month.

14. A table has millions of rows. What steps would you take to improve query performance if:

- Queries are running slow?
- Indexes are already in place?
- The query involves multiple joins?

Round 2
Big Data, Cloud & Data Eng.

1. What AWS services have you used in your projects?
2. Imagine you're tasked with designing data models and optimizing queries for both relational (Redshift) and NoSQL (DynamoDB) databases within AWS. How would you approach this?
3. Suppose you need to build a serverless data processing pipeline using AWS Lambda. How would you design and implement it, including ingestion, processing, and storage?
4. How can AWS Step Functions be used to design and orchestrate automated data pipelines for large-scale processing?
5. How do you ensure that data pipelines meet data quality standards, include proper metadata, and are validated for completeness and accuracy using AWS services?

Ankita Gulati

Shubh Goyal

6. Suppose you need to implement data encryption for sensitive data stored in Amazon S3. What are the key considerations and best practices?
7. What file formats are commonly used for data synchronization in projects?
8. What file formats are supported in Spark and Hadoop, and when would you use them?
9. Compare Avro vs Parquet. In what scenarios would you prefer one over the other?
10. What is YARN, and what role does it play in Hadoop?
11. What is Apache Spark? Also, explain what Apache Link is and when it occurs.
12. Explain the Spark Architecture.
13. What is the role of the Catalyst Optimizer in Spark?
14. What are the different types of joins in Spark?
15. What is the use of COALESCE in Spark?
16. Difference between DataFrame and Dataset in Spark.

Ankita Gulati Shubh Goyal

17. What is serialization in Spark, and why is it important?
18. Could you provide a technical overview of your project? What technologies are you using, and what is the main focus?
19. Are you using Git as a repository? If yes, explain how.
20. What is the time limit for an AWS Lambda function?
21. Have you worked with Scala? If yes, in what context?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal