TECH mahindra

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune / Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹22-26LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. Data Engineering Concepts

Ankita Gulati

Shubh Goyal

Round 1
SQL & Core Data Engineering

1. Can you write a SQL query to find the second highest salary from an Employee table?
2. Given an Employee table with two columns, how would you write a SQL query to find duplicate records based on those columns?
3. Can you tell me about the latest project you have worked on?
4. What were your roles and responsibilities in that project?
5. Which AWS services did you use in your project?
6. In which programming language did you write your data pipelines?
7. While building your pipelines, did you also focus on data quality aspects, or was it only about extraction and loading?
8. Can you explain Slowly Changing Dimensions (SCDs)?

9. What is the difference between a Fact table and a Dimension table?
10. Suppose you are receiving daily files in S3 which need to be transformed and loaded into Redshift. How would you design and implement such a pipeline?
11. Suppose a group job failed in a production environment. How would you troubleshoot and handle such a scenario?
12. Is there a way to set up alerts for such failures? If yes, how would you configure them?
13. How do you trigger a Lambda function?
14. If your data is present in an S3 bucket, can you create a Lambda trigger based on it?
15. Can a Lambda function be invoked asynchronously?
16. If a Lambda function is invoked asynchronously multiple times, what happens if one of the executions fails?
17. What are Dead Letter Queues (DLQs), and how are they used with Lambda?
18. How can you restrict a Lambda function's access to only a specific S3 bucket?

Ankita Gulati Shubh Goyal

Round 2
AWS & PySpark

1. Can you explain the difference between Redshift, Athena, and RDS, and when would you use each?
2. Have you worked with Amazon Kinesis? If yes, can you explain how you used it?
3. What is Amazon Aurora, and how does it differ from RDS?
4. How would you implement incremental data loads in Redshift?
5. Suppose you need to load terabytes of data into Redshift. How would you plan to do this efficiently?
6. What are the different distribution styles in Redshift, and when should each be used?
7. Can you explain Workload Management (WLM) in Redshift?
8. What is a Federated Query in Redshift, and when is it useful?

Ankita Gulati Shubh Goyal

9. What are Materialized Views, and how do they improve query performance?

10. How would you migrate data from an on-premises PostgreSQL database to Amazon Redshift?

11. Given a fact and a dimension table, how would you check whether the relationship between them is One-to-One, One-to-Many, Many-to-One, or Many-to-Many using PySpark

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal