Capgemini

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune / Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹24-30LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. Big Data / Distributed Systems
5. ETL & Data Engineering Concepts

Ankita Gulati

Shubh Goyal

Round 1
SQL & Core Data Engineering

1. If an S3 object is deleted, is there any way to restore it?
2. What are decorators in Python, and when would you use them?
3. Can you explain exception handling in Python? How does it differ from general error handling?
4. What do you understand by stored procedures in SQL, and where are they typically used?
5. What are some of the key techniques you use to optimize SQL queries for better performance?
6. Consider an employee table with the columns: employee_id, employee_name, and manager_id, where each manager is also an employee (manager_id refers back to employee_id). A new column manager_name has been added. Can you write an UPDATE statement to populate the manager_name by referencing the appropriate employee_name based on the manager_id?

Ankita Gulati Shubh Goyal

7. How would you choose the appropriate DISTKEY and SORTKEY for a table in Amazon Redshift?

8. What are the benefits of using DISTKEY and SORTKEY in Redshift? Do they primarily help with data retrieval, storage optimization, or both?

9. If you have a lookup table with a small number of records (around 100), which distribution style would you use in Redshift to optimize performance?

10. Suppose you have two tables:

- Table A: 1, 0, 1, NULL
- Table B: 1, 1, 0, 0, NULL, NULL

How would you perform all types of joins (INNER, LEFT, RIGHT, FULL OUTER, CROSS) between these two tables and count the number of rows returned by each join?

11. You have an input table with one column containing four teams: A, B, C, and D. Write a query to generate all possible unique matchups in an output table with columns team1 and team2 (e.g., A vs B, A vs C, etc., without duplicates).

12. Consider a table with columns: vendor_id, office_id, and date_of_visit. Each vendor may visit different offices on different dates, and vendor IDs can repeat. Write a query to identify consecutive visits for each vendor and office, and return:

- vendor_id
- office_id
- start_date and end_date of each streak
- Total number of consecutive visits in that streak.

13. What is schema evolution, and how do you handle it in a data pipeline?

14. Can you explain the difference between ETL and ELT? In what scenarios would you prefer one over the other?

15. What is the difference between a data layout, data mart, and data lake?

16. Consider a table with two columns: id and parent_id. How would you categorize each record as a root, inner, or leaf node based on the relationships?
17. Write a solution to determine the type of each node (root, inner, leaf) in a tree structure using the above table.

Round 2
AWS, Python & Advanced Data Engineering

1. How do you programmatically connect Python code to AWS services? Can you explain the process?
2. If you have a backend job that runs daily, how would you configure it to connect securely to AWS from Python?
3. Once you have IAM roles, how do you use them in Python for connecting to AWS services?
4. How are you managing your ETL processes currently? Which tools or frameworks are being used in your organization?
5. You mentioned using SNS in your setup. What purpose does it serve, and how does it support automation?
6. If your data contains extra spaces or special characters, how would you clean and prepare it for processing?

Ankita Gulati

Shubh Goyal

7. Which AWS services are used for which purposes? Can you briefly name a few with their primary use cases?

8. Do you have experience building end-to-end data pipelines from scratch? Can you walk me through one pipeline you designed?

9. Given daily incoming data, how would you design an ETL pipeline to implement SCD Type 2? What are the key components involved?

10. Suppose you have an employee dataset with multiple fields including employee_id, and you need to track changes over time using SCD Type 2. How would you:

- Perform the initial load?
- Identify changes in incremental loads?
- Detect and capture changes accurately?

11. Whose product is BigQuery, and how does it fit into the modern data stack?

Thank You

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

Ankita Gulati

Shubh Goyal