J.P.Morgan

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹24 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. ETL / Data Modeling
5. System Design & Optimization

Ankita Gulati

Shubh Goyal

Round 1
SQL & ETL Fundamentals

1. Suppose you have two tables with the following values:

- Table A: 1, 2, NULL, NULL, 1
- Table B: 1, NULL, NULL, 1, 5, 7

How would the result sets differ if you perform a LEFT JOIN, RIGHT JOIN, INNER JOIN, FULL OUTER JOIN, and a CROSS JOIN between these tables?

2. Given an employee table with columns: employee_id, salary, and department, where the first two rows are duplicates, how would you write a SQL query to delete duplicate rows while retaining exactly one unique record?

3. You have a sales table with columns:
---> order_id | customer_id | order_date | amount
Write a query to find the top 2 customers with the highest total sales amount in each month.

Ankita Gulati Shubh Goyal

4. Given a transactions table:
---> txn_id | account_id | txn_date | amount
Write a query to identify accounts that had 3 or more consecutive failed transactions.

5. You have a huge table with billions of rows. Your query uses GROUP BY and SUM, but it's running extremely slow. What techniques would you use to improve performance (partitioning, indexes, pre-aggregation, etc.)?

6. A query that worked fine earlier is now very slow. You can't change the schema or reduce joins. What real-world steps would you take to debug and optimize it? (EXPLAIN plan, indexes, statistics, partitioning strategy, etc.)

7. How would caching be beneficial in an ETL pipeline setup? While its benefits in reporting are obvious, how does caching help in ETL workflows?

Ankita Gulati Shubh Goyal

8. You have a table where data comes from multiple sources and sometimes contains dirty records (nulls, duplicates, mismatched formats). How would you build a SQL-based data validation framework before loading into a warehouse?

9. Suppose you need to design an ETL pipeline using AWS services:
--->A client uploads a CSV with 70–80 columns to an SFTP server once daily at an unpredictable time. The file must be split into 10–15 tables and support upserts along with basic validation.
Which AWS data warehouse would you choose, and which services would you include in the pipeline design?

10. In that ETL architecture, which orchestrator would you prefer—AWS Step Functions or Airflow? Additionally, how would you implement upsert operations—using AWS Glue or an alternate method?

11. If you are using Spark, would you write code to call a stored procedure, or directly implement the logic in Spark? Does Spark support the UPDATE command?

12. Window function challenge: You have a sales table with columns:
---> order_id | customer_id | order_date | amount
Write a query to get the running total of sales amount for each customer ordered by date.

13. Data skew question: In a distributed environment, one of your joins is failing because of data skew (a single key has too many rows). How would you handle it in SQL/Spark?

14. You are asked to create a process to generate files of fixed size (100 rows each) dynamically from a table of unknown row count. How would you implement this in SQL or ETL logic?

15. Explain how you would design a Slowly Changing Dimension (SCD) Type 2 process in SQL for tracking employee department changes. How would you handle large volumes efficiently?

Ankita Gulati

Shubh Goyal

Round 2
Python, AWS, Data Engineering

1. How would you restore a deleted object from an S3 bucket—what steps or AWS features would you use?

2. What are decorators in Python, and in what scenarios would you use them?

3. Can you explain the difference between exception handling and error handling in Python, with examples?

4. How do you connect Python code to AWS services programmatically? What tools, libraries, or IAM mechanisms would you use?

5. If a job runs automatically every day, how would you securely enable AWS connections from Python code using IAM roles?

Ankita Gulati

Shubh Goyal

6. How does AWS SNS support system automation in ETL pipelines, and what is its purpose in your workflow?

7. Imagine your data contains extra spaces or special characters. How would you clean and prepare it for processing?

8. What is the difference between cache and persist in Spark, and when would you choose one over the other?

9. Given a ‘coin change’ problem—list of denominations and target amount—how would you write code to find the minimum number of coins needed and the denominations used?
(e.g., denominations [1,2,5] with target 11 → 3 coins: [5, 5, 1])

10. Which file format do you generally prefer working with—CSV, JSON, or Parquet—and why, considering their advantages and limitations?

Ankita Gulati

Shubh Goyal

11. Have you implemented Slowly Changing Dimensions (SCD) Type 2 using Python, PySpark, or a database? Walk me through your approach—initial load, change detection, and incremental updates.

12. What is the difference between normalization and denormalization?

13. How would you define indexes in a database, and how do they help query performance?

14. Suppose you have two tables, each containing approximately 800 million records, and you need to join them—what strategy would you use to optimize the join?

15. Can you explain the difference between AWS Glue and AWS Lambda, and when you would use each?

16. How do you manage schema evolution, and what tools or practices do you use?

17. Write a SQL query to fetch the entire reporting hierarchy from an employee table structured as:

emp_id | emp_name | manager_id
1 | Alice | NULL
2 | Bob | 1
3 | Carol | 1
4 | Dave | 2
5 | Eve | 3

18. Describe a data engineering problem you solved at scale. What was the challenge, and how did you address it?

19. How do you ensure data quality and consistency in production ETL pipelines?

20. Explain your experience with distributed data processing frameworks like Spark, and optimization strategies you used.

21. Have you used infrastructure-as-code tools like Terraform or CloudFormation for data infrastructure?

Thank You

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

Ankita Gulati

Shubh Goyal