<epam>

# Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

# Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹22-25 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

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

# Round 1
# SQL/Python Fundamentals

1. Explain the difference between Normalization and Denormalization.
2. What are the different Slowly Changing Dimensions (SCD) types?
3. Suppose an employee was working in Bangalore and moved to Chennai. How would you design the table so that both old and new records are maintained (keeping history + latest record)?
4. What is an Index in SQL? Explain different types of indexes.
5. What is the difference between Materialized View and Non-Materialized View?
6. What are Window Functions in SQL? Provide examples.
7. Explain the difference between INNER JOIN, LEFT JOIN, RIGHT JOIN, FULL OUTER JOIN, UNION, UNION ALL.
8. Write a query to find the 3rd highest salary from the Employee table.

Ankita Gulati

Shubh Goyal

9. Write a query to find duplicate records in a table.
10. Write a query to retrieve customers who placed an order on the same date as another customer (self-join scenario).
11. Write a query to find the top 3 customers with the highest number of orders.
12. Write a query to calculate the running total of sales per month.
13. What is a Decorator in Python?
14. What is a Class in Python? Provide an example.
15. Write a Python class with a constructor initializing a and b, and a method that returns their sum.
16. Write Python code to reverse a list without using built-in reverse methods.
17. Write a Python script to connect to a database and fetch records.
18. How would you handle exceptions in a Python-based ETL pipeline?
19. What are the differences between groupByKey() and reduceByKey() in Spark?

20. Explain the difference between Dataset and DataFrame in Spark.

21. Explain the PySpark architecture.

22. Explain the difference between Partitioning and Bucketing in Spark.

23. What are UDFs in PySpark? Provide an example.

24. What is Serialization in Spark? How does Spark use serialization for performance?

25. Explain Data Skewness problem in Spark. How do you handle it?

26. Write PySpark code to:

→ Read two CSV files: orders.csv and customers.csv.
→ Print record counts.
→ Join both DataFrames on customer_id.
→ Save the joined DataFrame to a file.

27. Suppose you have a dataset with:

user_id | page_url | timestamp

→ Filter data by user_id.

→ Convert timestamp column into proper timestamp format.

→ Group by user_id and count page visits.

→ Find the Top 10 users by activity.

→ Assign rank per user using a window function.

# Round 2
# Advanced Data Engineering, Spark & Cloud

1. Explain the Spark Architecture.
2. Explain different types of Transformations in Spark.
3. What are the file formats you have used (CSV, JSON, Parquet, ORC, Delta)? Which one is best for performance, and why is Parquet/Delta preferred?
4. In Spark, when joining two large transactional tables, what steps would you take to minimize shuffling?
5. Based on Spark execution, explain how Jobs, Stages, and Tasks are created.
6. Explain the role of Serialization in Spark performance optimization.
7. What is ETL? Explain its phases and tools you have worked with.
8. How do you ensure data quality in ETL pipelines?

Ankita Gulati Shubh Goyal

9. Which orchestration tools (Airflow, Oozie, Azure Data Factory, AWS Step Functions) have you used? How do you schedule and monitor jobs?
10. Have you used Databricks Workflows? If yes, explain how you set up workflows for pipeline automation.
11. Explain the architecture of Snowflake or BigQuery.
12. What is the difference between OLAP and OLTP databases?
13. Explain Schema Evolution in a Data Lake.
14. How do you handle late-arriving data in streaming pipelines?
15. What is the role of Apache Kafka in a Data Engineering ecosystem?
16. How would you design a real-time streaming pipeline to push logs into a data warehouse?
17. What were the functional objectives and client requirements in your last project?
18. What were your roles and responsibilities on a day-to-day basis?

19. Suppose you need to ingest terabytes of log data daily into a warehouse like Snowflake/BigQuery — how would you design the pipeline?
20. How do you ensure fault tolerance and scalability in your data pipelines?
21. Write SQL to pivot a table (convert rows to columns).
22. Write SQL to find employees earning more than their manager.
23. Write PySpark code to read JSON files from S3 and store them in Parquet format.
24. Write Python code to implement a Producer-Consumer pattern.

# Thank You

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

Ankita Gulati

Shubh Goyal