Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Bangalore Pune
- **Work mode:** Hybrid
- **Compensation:** ₹20-21 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming

Ankita Gulati

Shubh Goyal

Round 1
Data Engineering & Foundations

1. Could you provide a technical overview of your recent project? What technologies did you use, and what was the primary focus?
2. What file formats are commonly used for data synchronization in big data systems?
3. What is YARN, and how does it manage resources in Hadoop?
4. What file formats can be utilized in Spark and Hadoop, and when do you prefer one over another?
5. Compare Avro vs Parquet – when would you use each?
6. What is Apache Flink, and how does it differ from Apache Spark? In which scenarios would you prefer Flink?
7. Explain the Spark architecture (driver, executors, cluster manager).
8. What is the use of coalesce() in Spark, and how is it different from repartition()?

Ankita Gulati Shubh Goyal

9. What are the different types of joins in Spark? Give practical examples.
10. Explain the role of the Catalyst Optimizer in Spark SQL.
11. Compare DataFrame vs Dataset in Spark. When would you use each?
12. What are the different serialization techniques in Spark, and why are they important?
13. Consider two tables:
    --> Table1 (ID): 1, 2, 3, 4, 5
    -->Table2 (ID): 1, 2, 3

What will be the row counts for inner join, left join, right join, full outer join, cross join, semi join, and left semi join?
14. What is the time limit for AWS Lambda functions, and how can you handle long-running processes?
15. Have you worked with Scala? How does it integrate with Spark compared to Python?
16. Are you using Git for version control in your projects? If yes, what branching strategy do you follow?

Ankita Gulati Shubh Goyal

Round 2
Coding & Problem Solving

1. Write a SQL query to find the most frequently ordered products for each customer. (Table: orders(order_id, date, customer_id, product_id))
2. Write a query to find the Nth highest salary without using TOP or LIMIT.
3. Given a sales table with (date, sales_amount), write a query to calculate a 7-day rolling average of sales.
4. Write a query to find customers who placed orders in three consecutive months.
5. Write a query to pivot sales data by month (rows → columns).
6. Write a query to find the second most purchased product per customer.
7. From a transaction table, identify users with duplicate transactions made within 5 minutes.

Ankita Gulati

Shubh Goyal

8. Write a Python program to check if two words are anagrams (silent, listen).
9. Write Python code to find the first non-repeating character in a string.
10. Write Python code to flatten a nested dictionary/JSON into a single level.
11. In PySpark, write code to deduplicate records based on (user_id, event_time), keeping only the latest.
12. Given web logs (user_id, timestamp, url), write PySpark code to sessionize logs with a 30-minute inactivity window.
13. In PySpark, given a DataFrame of (customer_id, product_id, quantity), find the Top 3 products purchased per customer.
14. Write a PySpark job to read nested JSON data from S3, flatten it, and write it as partitioned Parquet by date.
15. Python Challenge: Write a function to generate the longest substring without repeating characters from a string.

Ankita Gulati Shubh Goyal

16. How do you decide cluster resource allocation for Spark jobs (executors, memory, cores)?
17. Explain a scenario where you had to optimize a slow ETL job – what steps did you take?
18. How do you ensure fault tolerance in data pipelines?
19. Describe how you would design a real-time streaming pipeline using Kafka + Spark/Flink.

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal