Nielsen

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 2+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹10-12 LPA
- **Total Rounds:** 3
- **Top Required Skills:**

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

Ankita Gulati

Shubh Goyal

Round 1
Core Technical (DSA + SQL)

1. Walk me through your previous tech stack.

2. What was the data volume you worked with, and what impact did your work have?

3. Longest Substring with Unique Characters
→ Input: "aabcdeeefijklmno"
→ Output: "fijklmno"
→ Expected: Solve using the sliding window technique with a hash set.

4. Check if Two Strings are Anagrams (O(n), no sorting allowed)
→ Input: s1 = "tan", s2 = "ant" → Output: Yes
→ Expected: Use a hash map/dictionary to count character frequencies.

5. Given a product pricing table with schema (product_name, product_id, price, price_change_month), write a query to return products with strictly increasing prices over months.
→ Approach: Use LAG / ROW_NUMBER window functions or self-joins to compare rows.

6. Write a query to find the 2nd highest salary per department from an Employee table.

7. Write a query to calculate the cumulative monthly sales amount per customer.

8. Python: Write a function to merge two sorted arrays into one sorted array in O(n).

9. Python: Implement a function to validate a balanced string of brackets {}, [], ().

Round 2
Data Engineering & Spark

1. Explain your previous project, its architecture, and your contributions.

2. How do you handle data skewness in Spark?

3. What are your strategies for code optimization and partitioning in Spark?

4. Design a ride-booking app like Uber/Ola.
Build a Galaxy schema with fact and dimension tables for:

→ Users
→ Drivers
→ Rides
→ Payments
→ Locations
→ Ratings

Ankita Gulati

Shubh Goyal

5. Design a simplified Ride Booking OOP system with classes for:

- → Users (request ride)
- → Driver (location, vehicle type)
- → Driver Availability
- → Fare Calculation
- → Status (accept/reject ride)
- → Payments

6. Explain partitioning vs bucketing in Spark.

7. What’s the difference between coalesce() and repartition()?

8. Write PySpark code to:

- → Drop specific columns from a DataFrame.
- → Filter rows based on a condition.
- → Identify duplicate rows using Window Fns.

9. How would you implement SCD Type 2 in a data pipeline?

Round 3
Situational + Hiring Manager

1. Give an introduction to your role at Nielsen.
2. What were your responsibilities and impact areas?
3. What are your future expectations from this role?
4. How can Amazon Prime detect if the same user logs in from different accounts and locations (e.g., India and US) to ensure consistent movie recommendations?
5. How would you design a real-time fraud detection system using Spark + Kafka + Cloud?
6. How would you build a data pipeline to handle schema evolution in production?
7. What steps do you take to ensure data governance and compliance (GDPR, PII masking, auditing)?
8. How do you measure data quality in pipelines (validations, deduplication, missing values)?
9. How do you monitor pipelines for failures and SLAs?

Thank You

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

Ankita Gulati

Shubh Goyal