Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5-7 years
- **Location:** Bangalore
- **Work mode:** Remote
- **Compensation:** ₹50+ LPA
- **Total Rounds:** 4
- **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
Recruiter / Phone Screen

1. Introduce yourself and walk through major data engineering projects.

2. Why do you want to join NVIDIA? What excites you about AI/data infrastructure?

3. Which cloud or streaming tools have you worked with? (e.g., Kafka, Spark, AWS/GCP)

4. Light technical question: “In one of your pipelines, how did you handle late-arriving or out-of-order data?”

Ankita Gulati

Shubh Goyal

Round 2
Technical (Coding + SQL)

Coding / Algorithm Questions:

1. Missing Number Problem

• Given two arrays, find the integer missing in the second array.

• Example: A = [1,2,3,5], B = [2,3,5] → Missing = 1.

2. String Manipulation / Greedy Problem

• Example: Given a line of text and a max width, format it with even spacing (text justification).

SQL Questions:

3. Ranking Query

• Find the 5th highest salary from the employee table, handling ties.

• Discuss difference between DENSE_RANK() vs RANK().

4. Joins & Aggregations

Given - sales(product_id, sale_date, amount)
& product(product_id, region)

• Write a query for total sales per region in the last month, or top-selling product per region.

Ankita Gulati Shubh Goyal

Round 3
System / Pipeline Design

1. **Real-Time Streaming Pipeline**

• Design a pipeline to ingest clickstream/telemetry data and refresh dashboards every few seconds.

• Include ingestion (Kafka/Kinesis), processing (Spark/Flink), serving (Redis/Elasticsearch).

• Handle fault tolerance, retries, and late-arriving events.

2. **GPU Server Health Monitoring**

• How would you collect logs & metrics from thousands of servers?

• Discuss alerting on anomalies, scaling, and monitoring.

3. **SQL for Analytics**

• Given a logs table with timestamps/events, compute rolling averages using sliding windows.

Ankita Gulati

Shubh Goyal

4. **Data Modeling**

• Compare star schema vs normalized schema for analytics.

• How would you partition event data for fast queries and archival?

5. **Performance Optimization**

• Where do bottlenecks occur in Spark pipelines? (shuffles, I/O, network)

• How to optimize joins, transformations, and caching?

Ankita Gulati

Shubh Goyal

Round 4
Behavioral Interview

1. Describe your most technically complex project. What were the biggest challenges? What would you do differently?

2. Share a time you faced conflicting priorities or ambiguous requirements. How did you align stakeholders?

3. How do you make trade-offs between performance, cost, and speed?

4. Have you worked on incident response for broken pipelines? What did you learn?

5. How do you collaborate with ML engineers, infra teams, analysts, and product managers?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal