Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 6 years
- **Location:** Hyderabad
- **Work mode:** Hybrid
- **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
Data Structures & Algorithms

1. Remove Duplicates from Employee IDs

- Given a list of employee IDs with duplicates, remove duplicates and return sorted IDs.
- Follow-up: How would you handle very large datasets (10M+ IDs)? Discuss time/space complexity and streaming deduplication with external sorting.

2. Reverse a Linked List

- Problem: Reverse a singly linked list and print elements.
- Follow-ups:
- What's the time and space complexity? (O(n), O(1)).
- How would you reverse in groups of k nodes?

Round 2
Advanced SQL & Data Modeling

SQL Questions:

1. Third Highest Transaction per Branch

- Write a query to find the 3rd highest transaction amount per branch (considering ties).
- Follow-up: How do you handle cases where some branches have fewer than 3 transactions?

2. Query Optimization Discussion

- How would you optimize queries with billions of rows?
- (Partitioning, indexing, avoiding cross joins, materialized views).

Ankita Gulati

Shubh Goyal

Data Modeling Question:

- Design a Banking Schema for Accounts and Transactions
- Accounts Table: account_id, customer_id, branch_id, open_date, status
- Transactions Table: txn_id, account_id, txn_type, amount, date, status
- Expected: Indexing (account_id, date), Partitioning (by date or branch_id).
- Follow-ups:
- Extend schema for loans and credit cards.
- When would you use denormalization (e.g., for reporting dashboards)?

Ankita Gulati Shubh Goyal

Round 3
Data Engineering Concepts & ETL

Scenario 1 – Real-Time Fraud Detection Pipeline
How would you design a near real-time fraud detection system for credit card transactions?

- Ingestion: Kafka.
- Processing: Spark Structured Streaming / Flink with windowed aggregations.
- Model Scoring: Fraud ML model via REST / TensorFlow Serving.
- Serving Layer: Alerts to monitoring dashboards or fraud team.
- Reliability: Dead-letter queues for failed events.
- Follow-ups:
- Handling late-arriving events.
- Latency guarantees (<5 sec end-to-end).

Scenario 2 – Schema Evolution & Backward Compatibility

Pipelines often break when schemas change. How do you handle this?

- Use Schema Registry (Avro/Protobuf).
- Ensure backward compatibility (new fields nullable).
- Data contracts between producer/consumer.
- Follow-up: What if downstream breaks? → versioned topics or views.

ETL & Cloud Optimization Discussion

- Explain incremental ETL vs full reloads.
- How do you handle partial failure retries?
- How do you optimize cloud costs (cluster auto-scaling, spot instances, partition pruning)?

Round 4
HR & Managerial Discussion

Behavioral & Leadership Questions:

1. Why do you want to join Kotak Mahindra?
2. Where do you see yourself in the next 3–5 years?
3. Tell me about a time you resolved a critical production P1 issue.
4. How do you collaborate with business analysts, fraud teams, and product managers?
5. How do you maintain work-life balance under tight deadlines?
6. How do you handle mistakes in production?
7. Explain the architecture of your current project and your contributions.

Thank You

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

Ankita Gulati

Shubh Goyal