Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer-2
- **Experience:** 4–5 Years
- **Location:** Bengaluru
- **Work mode:** Office
- **Compensation:** ₹30-40 LPA
- **Total Rounds:** 5
- **Top Required Skills:**

1. SQL
2. DSA
3. System Design
4. Data Modeling
5. Real-Time Processing
6. Behavioral Fit

Ankita Gulati

Shubh Goyal

Round 1
SQL & Data Structures (DSA)

SQL

1. Given two tables – Loans and Account_Balances – write an SQL query to evaluate customer liability:
   - If total loan > account balance → mark as "High Liability".
   - Else "Low Liability".
   - Handle edge cases where customers may have multiple accounts or multiple loans.
   - Optimize using CASE statements and JOINs.
2. Interviewer follow-up: How would you optimize this query for very large datasets (indexes, partitions, joins strategy)?

DSA Problem (Arrays)

1. Solve an array-based problem (variation of Two-Sum / Subarray problem).

- Example: Given an array, check if there exists a subarray with a target sum.
- First solve using brute-force ($O(n^2)$).
- Then optimize with hashing / two-pointer sliding window.
- Explain time and space complexity of both approaches.

Ankita Gulati Shubh Goyal

Round 2
System Design

- How would you store large datasets (terabytes/petabytes)?
  → HDFS vs S3 vs DynamoDB.
- How to ensure fault tolerance when nodes fail?
  → Replication strategies, leader election.
- How would you enable fast retrieval for clients requesting results?
  → Indexing, caching strategies.
- How would you use MapReduce for Two-Sum?
  → Map phase: emit (number, 1).
  → Reduce phase: check complement existence.
- If the dataset grows 10x in one year, how would your design scale?

Ankita Gulati Shubh Goyal

Round 3
Data Modeling & SQL

1. Data Model Design:
   - Users (user_id, name, age, country).
   - Artists (artist_id, artist_name, genre).
   - Songs (song_id, artist_id, title, release_date).
   - Plays (user_id, song_id, play_timestamp).
   - Recommendations (user_id, artist_id, score).
2. How to handle Slowly Changing Dimensions (SCD) if:
   - An artist changes genre.
   - A user updates country.
   - Which SCD type would you use? (Type 1 vs Type 2).
3. SQL Challenge:
   - "Find users who should receive recommendations for a newly onboarded artist."
   - Logic: Users who played songs from similar genres in the last 30 days.
   - Write a query joining Users, Plays, Songs, Artists.
4. Follow-up: How would the query behave if the tables had billions of rows?

Ankita Gulati Shubh Goyal

Round 4
Real-Time Data Processing

1. Which streaming tool would you use (Kafka, Kinesis, Flink, Spark Streaming)? Why?
2. How to handle distributed processing at scale?
   - Partitioning strategies in Kafka.
   - Exactly-once vs At-least-once semantics.
3. Fault Tolerance:
   - If a consumer node crashes mid-batch, how do you ensure messages aren't lost?
   - Role of checkpoints and offsets.
4. Data Consumption:
   - Push vs Pull strategies.
   - Backpressure handling.
5. If 1 partition has skewed data, how would you balance the load?

Ankita Gulati Shubh Goyal

Round 5
Hiring Manager Discussion

1. Walk me through your team projects and your role in them.
2. How do you handle conflicts within a team? Example scenario?
3. What do you expect from work culture at Amazon?
4. How do you prioritize tasks when multiple deadlines collide?
5. Questions from candidate: growth, opportunities, and Amazon Leadership Principles.

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal