amazon

# Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

# Job Details

- **Position:** Data Engineer
- **Experience:** 3 Years
- **Location:** Bengaluru
- **Work mode:** Office
- **Compensation:** ₹28-40 LPA
- **Total Rounds:** 7 (2 Online + 5 Onsite)
- **Top Required Skills:**

1. SQL Mastery
2. Python Coding
3. Big Data Tools
4. System Design
5. AWS Cloud
6. Behavioral Fit

Ankita Gulati Shubh Goyal

# Round 1

## Online Assessment (HackerRank)

1. Write a query to calculate the total outstanding loan liability of each customer by joining Loans and Accounts.
2. Using window functions, find the top 3 highest spending customers per region.
3. Identify customers who have missed payments in 2 consecutive months (use LAG).
4. From a transactions table, calculate running total balance per customer (use SUM() OVER (PARTITION BY ORDER BY)).
5. Debug a query with nested subqueries and rewrite it using joins for better performance.

# Round 2
# Technical (SQL + Python + Big Data)

## SQL Section

1. Given an orders table, write a query to return the 2nd highest order value per customer (handle case when only 1 order exists).
2. Optimize a query performing multiple self-joins on a large table – suggest indexing and partitioning strategies.
3. Write a query to calculate monthly active users where a user is active if they logged in at least 3 times.

## Python Section

4. Implement a function to:

- Read a list of customer transactions.
- Remove duplicates.
- Aggregate the total spend per customer using dictionaries.

5. Follow-up: Extend the function to process large datasets (handle memory efficiency).

**Big Data Concepts:**

6. What is partitioning in Spark? How does it help performance?
7. Difference between distributed computing in Hadoop vs Spark.
8. How would you handle data skew in joins?

# Round 3
# SQL + System Design

## SQL

1. You're given a table of product sales. Write a query to:
   - Find the top selling product in each category.
   - Optimize query for billions of rows.
2. Rewrite a poorly written query with subqueries into window functions for efficiency.

## System Design

3. Design a pipeline for processing clickstream data from an e-commerce app.

- How would you ingest the data?
- Which storage would you choose (RDS, Redshift, S3, DynamoDB)? Why?
- Batch vs Streaming approach → trade-offs.
- How do you ensure fault tolerance if a node fails?
- How do you scale this design when traffic spikes 10x?

Ankita Gulati Shubh Goyal

# Round 4
# Data Modeling + Python

## Data Modeling

1. Design a schema for an online food delivery platform with tables: Users, Orders, Restaurants, Payments.
    - ○ Show relationships between entities.
    - ○ Handle Slowly Changing Dimensions (SCD) for restaurant name changes.
    - ○ Normalize the schema but explain cases where denormalization might be better.

## Python

2. Given a large text file of user logs:
- Count the frequency of each action (login, purchase, logout).
- Use dictionaries for aggregation.
- Handle file I/O efficiently for large files.

3. Follow-up: Extend to stream processing instead of batch.

Ankita Gulati Shubh Goyal

# Round 5
# BAR RAISER (Behavioral)

1. Describe a project where you demonstrated ownership beyond your defined role.
2. Tell me about a time you faced a conflict with a teammate. How did you resolve it?
3. Share an example of delivering results under tight deadlines.
4. How do you learn new technologies quickly when asked to use something unfamiliar?
5. What motivates you to perform at your best?

# Round 6
# Big Data + Cloud (AWS)

**Big Data Section:**

1. How do you handle data skew in Spark when one key dominates a join?
2. Difference between Spark repartition() vs coalesce().
3. How would you evolve schemas in Hive without breaking queries?
4. Explain wide vs narrow transformations in Spark.

**Cloud Section (AWS):**

5. How do you build a fault-tolerant pipeline with AWS services?
   - Example: AWS Glue → Lambda → S3 → Redshift.
6. Given a dataset growing daily, how do you design partitioning in S3?
7. Cost optimization strategies when EMR clusters are underutilized.
8. How do you secure data in AWS S3? (IAM, encryption, access policies).

Ankita Gulati Shubh Goyal

# Round 7
# End-to-End Project Discussion

1. Explain a recent project you worked on:
   → Problem statement.
   → Tools & technologies used.
   → Architecture design.
   → Trade-offs you made.
2. How did you ensure data quality (validation, error handling, monitoring)?
3. Describe how you optimized your pipelines:
   a. Partitioning
   b. Caching
   c. Parallel processing
4. How did you handle schema changes mid-project?
5. What was the business impact of your project (time saved, cost reduction, improved reporting)?

Ankita Gulati Shubh Goyal

# Thank You

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

Ankita Gulati

Shubh Goyal