media.net

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 4+ years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹25+ 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
SQL & DSA

1. Write a SQL query to evaluate the liability of customers to banks, given their existing loans and bank account balances. Use JOINS and CASE statements effectively.
   • (Scenario: customers may have multiple loans and accounts; query must compute net liability).

2. Explain how you would optimize queries that involve multi-table joins and conditional logic for performance on large datasets.

3. Solve an array-based problem (similar to Two-Sum).
   • Start with the brute-force solution.
   • Then, provide the optimized solution and explain the time and space complexity trade-offs.

Ankita Gulati Shubh Goyal

4. Write a SQL query to find the second highest transaction amount per customer using window functions.

5. Given a dataset of user activity, write SQL to calculate retention cohorts (weekly active users returning in later weeks).

6. Write a Python function that checks if an array contains any subarray with sum equal to a target k (use hashing for efficiency).

Ankita Gulati

Shubh Goyal

Round 2
Distributed Systems

1. How would you store huge datasets in a distributed system?
2. How would the system handle failures of nodes?
3. How would clients read data reliably from the system?
4. How would mappers and reducers work to solve the Two-Sum problem at scale? (Explain approach, even if incomplete).
5. Explain the architecture of HDFS — how are blocks stored, replicated, and recovered?
6. Describe speculative execution in MapReduce and why it is useful.
7. How does consistent hashing help in distributed systems

Ankita Gulati

Shubh Goyal

Round 3
Data Modeling & SQL

1. Design the data models for a music company to store:
   - Users
   - Artists
   - Songs
   - Other related entities (e.g., playlists, interactions).
2. How would you handle slowly changing dimensions (SCD)? Which type (Type 1, Type 2, etc.) would you choose?
3. Write a SQL query to find the relevant users to recommend a new artist's songs.
4. Explain star schema vs. snowflake schema. When would you use each?
5. How do you design a schema that supports both ad-hoc analytics and ML model training?
6. Explain how to implement data partitioning and bucketing to improve query performance.
7. Discuss data governance practices you'd apply when modeling user and artist data

Ankita Gulati Shubh Goyal

Round 4
Real-Time Streaming

1. Design a system that processes a huge real-time stream of messages.
   • How would it run in a distributed manner?
   • How would you ensure fault tolerance if nodes fail?
   • How would you handle reading from the stream efficiently?
2. How does Kafka handle partitioning, consumer groups, and fault tolerance?
3. How does repartitioning among consumers work in Kafka?
4. Compare Kafka, Kinesis, and Pulsar — when would you pick each for ingestion pipelines?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal