IBM

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 4+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹22+ LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark / Python / Databricks
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming
6. System Design

Ankita Gulati

Shubh Goyal

Round 1
Technical Discussion

1. Explain the difference between clustered and non-clustered indexes in a database. Provide examples of when to use each.
2. How would you design a schema for a system that requires high throughput and availability?
3. What strategies can you use to optimize query performance in a relational database?
4. How do you manage database transactions and maintain data integrity in high-concurrency environments?
5. Write a Python function to remove duplicates from a list without using library functions.
6. How would you debug a piece of Python code that is failing due to an unknown exception?
7. Demonstrate how you would use recursion in Python to solve a common algorithmic problem (e.g., factorial, Fibonacci).

Ankita Gulati

Shubh Goyal

8. Describe how to implement a queue using two stacks.
9. Which data structure would you use for efficiently searching a contact list and why?
10. How would you detect and remove a cycle in a linked list?
11. Explain the difference between a heap and a binary search tree (BST). When would you prefer one over the other?
12. Describe an algorithm to identify anomalies in a dataset using statistical methods.
13. How would you design an algorithm for an efficient text search across multiple documents?
14. Explain the process of balancing a binary search tree and why it is important.
15. Write a SQL query to find the top 3 customers by revenue in the last 90 days.
16. Python: Write a function to flatten a deeply nested JSON object into a key-value dictionary.
17. SQL: Given a table orders(order_id, customer_id, order_date, amount), write a query to calculate month-over-month revenue growth.

Ankita Gulati **Shubh Goyal**

Round 2
Advanced Technical Discussion

1. Explain how MapReduce works in Hadoop with an example use case.
2. What are the key advantages and disadvantages of Apache Spark compared to Hadoop MapReduce?
3. How would you design a system to process streaming data in real-time?
4. How would you build a system to analyze large volumes of streaming data for real-time insights (e.g., fraud detection, anomaly detection)?
5. Describe a scenario where cloud storage would be more beneficial than on-premises storage.
6. How would you ensure data security, compliance, and governance in a cloud environment?
7. Explain the concept of serverless computing. How can it be applied in data engineering (e.g., AWS Lambda, Azure Functions, GCP Cloud Functions)?
8. Walk me through the steps to migrate an on-premises data warehouse to the cloud.

Ankita Gulati

Shubh Goyal

9. How do you handle data skew in distributed processing frameworks like Spark?
10. What strategies would you use to optimize join operations in Spark or SQL?
11. How would you design a fault-tolerant ETL pipeline for mission-critical workloads?
12. Explain how you would implement exactly-once processing in a real-time data pipeline.
13. Explain the concept of schema evolution in a data lakehouse (e.g., Delta Lake, Apache Iceberg).
14. How would you design a multi-tenant data platform ensuring isolation, governance, and cost efficiency?
15. How would you monitor and scale a Kafka + Spark Streaming pipeline for high-throughput ingestion?
16. Describe how you would set up data quality checks (nulls, duplicates, schema drift) in an automated pipeline.

Ankita Gulati

Shubh Goyal

Round 3
HR Discussion

1. Can you walk me through your resume and highlight key projects that align with IBM's data engineering role?
2. Why do you want to join IBM as a Senior Data Engineer?
3. Tell me about a challenging project you worked on. How did you handle deadlines and pressure?
4. How do you approach cross-team collaboration (data scientists, product owners, DevOps engineers)?
5. What are your short-term and long-term career goals?
6. What is your notice period, preferred location, and salary expectation?

Thank You

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

Ankita Gulati

Shubh Goyal