Microsoft

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 3+ years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹25+ 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

Round 1
Behavioral & Project Discussions

1. Tell me about a time when you faced a particularly difficult challenge in a project. How did you approach and resolve it?

2. Describe a data quality (DQ) check you performed in your project.
   - What failed during file extraction?
   - Where did you add your pre-processing layer?

3. After the ETL and DQ checks, what happens next in your pipeline? Would you consider this step part of ETL or outside of it?

Ankita Gulati

Shubh Goyal

Round 2
SQL Problem-Solving

1. Write a SQL query to find the phone numbers of the top 2 highest salaried employees.
→ **Table schema:**
emp_id | emp_name | dept_id | salary | manager_id | emp_age | emp_phone

2. Write a SQL query to get the top 2 highest salaried employees within each department.
3. Explain your approach for the above queries and analyze their time complexity. Why is it O(n log n)?
4. Modify your query logic to return two employees with the same salary. Discuss how to handle ties using RANK(), ROW_NUMBER(), or DENSE_RANK().
5. Explain Slowly Changing Dimension (SCD) Type 2 and demonstrate how you would apply it to the given employee table.

- **Follow-up:** Extend this discussion to other SCD-related SQL scenarios.

Ankita Gulati   Shubh Goyal

Round 3
Python Coding & Logic

1. Write Python code to find the longest substring from the provided text.
→ text = "There was a long string provided, separated by ','"

2. Answer follow-up questions on handling edge cases (e.g., empty strings, duplicate longest values, performance on very large strings).

Round 4
Technical Deep Dive

1. Describe a complex data engineering task you worked on. What challenges did you face, and how did you solve them?
2. Have you encountered issues with Spark? List and explain a few examples.
3. What is inefficient serialization in Spark?
4. When is the byte stream conversion process initiated in Spark?
5. How have you used Broadcast Variables and Checkpointing in Spark applications?
6. How did you handle partition-based target locations when writing data to sinks?
7. Serialization efficiency: In which stage is it implemented — partitioning, reading, writing, or processing? If in processing, at what point exactly?
8. How do you handle node failures during operations like groupBy or during serialization in Spark?

9. What is granularity in a fact table, and how do you decide the appropriate level?
10. Explain your approach to implementing data lineage and auditing in an ETL pipeline.
11. What is data lineage?
12. If you had all the tools and data available, how would you implement end-to-end lineage?
13. How does data lineage benefit analytics teams and business stakeholders?
14. What is the Java Virtual Machine (JVM), and why is garbage collection important in Spark?
15. Explain how level caching is used in Spark for performance tuning.

Thank You

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

Ankita Gulati

Shubh Goyal