INDIUM

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹10-14 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

Round 1
Python & SQL Fundamentals

1. Explain the difference between a Python list and a tuple. Give an example of each.
2. Create a tuple that contains only one string element. How is it different from a normal string?
3. What is a generator in Python, and when would you use it?
4. What does **kwargs mean in Python?
5. Is there any limit on the number of arguments we can pass using **kwargs?

6. Write a Python program to find the first non-repeating character in a string using a dictionary.
Example: Input = "abxabyz" → Output = x

7. What is the execution order of an SQL query? Explain with an example.
8. Write an SQL query to find the second-highest salary from an employee table.

Ankita Gulati Shubh Goyal

9. SQL Case Study: A phone call is considered international when the caller and receiver are in different countries. Write a query to calculate the percentage of international calls, rounded to 1 decimal.

Example: Tables → phone_calls, phone_info.

Round 2
Advanced Data Engineering

1. What do you mean by Schema on Read and Schema on Write? Which approach does Hive follow?
2. What are the key differences between Hadoop 1.0 and Hadoop 2.0?
3. Explain the difference between Hive Internal vs External Tables. When would you use each? (Added)
4. What are the stages in a typical ETL pipeline?
5. How would you design a data pipeline to process 1 TB of log data daily? (Added)
6. Explain the importance of partitioning and bucketing in Hive.
7. How would you orchestrate a Spark job using Airflow?
8. Compare AWS Glue and Databricks for ETL workloads.
9. How would you handle schema evolution in a data lake?

10. What is lazy evaluation in Spark? What actually happens during this stage?

11. Explain data skewness in Spark. What are the main causes, and how do you fix it?

→ You have a DataFrame with two columns:
→ student_id (millions of unique IDs)
→ students_joining_date (only 5–6 unique dates)

If you apply partitioning, which column would result in larger partition sizes and why?

12. Differentiate between repartition() and coalesce() in Spark.

13. How does the Catalyst Optimizer in Spark SQL work?

Round 3
Hiring Manager

1.Tell me about yourself and your career journey so far.

2. Why do you want to join Indium Software?

3. What challenges have you faced in your projects, and how did you overcome them?

4.How do you stay updated with the latest data engineering tools and technologies?

5. Salary expectations, notice period, and career aspirations.

Thank You

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

Ankita Gulati

Shubh Goyal