Infosys

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹20-24 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

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

Ankita Gulati

Shubh Goyal

Round 1
Data Engineering & Coding

1. Tell me about yourself and highlight your recent data engineering projects.
2. What types of data transformations have you worked on in ETL/ELT pipelines?
3. Given a CSV file, how would you read it into a DataFrame, remove duplicates, and write it back in Parquet format?
4. Write SQL/PySpark/Python code to calculate country-wise total sales from a sales table.
5. Write a query to fetch the top 3 highest salaries from an employee table.
6. Write a query to identify duplicate records in a table.
7. Write PySpark code to perform a broadcast join and explain when it is most efficient.
8. You are reading a Delta table, and it is taking 10 minutes. How would you optimize the read performance?

Ankita Gulati Shubh Goyal

9. What is the difference between cache() and persist() in Spark? Which one do you prefer and why?
10. Explain lazy evaluation in PySpark with an example.
11. Difference between deep copy and shallow copy in Python. Provide a coding example.
12. Do you use logging functionalities while writing production code? If yes, how?
13. Write a Python script to connect to a database and fetch data.
14. Write PySpark code to calculate the running total of sales by month.
15. Write SQL query to get nth highest salary without using TOP or LIMIT.

Ankita Gulati

Shubh Goyal

Round 2
Data Architecture & Cloud Engineering

1. What is Integration Runtime (IR) in ADF? What types are available and which one do you use most?
2. How do you implement incremental load in Azure Synapse pipelines?
3. If we want to implement data security or compliance in Synapse, how would you handle it?
4. What types of storage options are available in Azure? Which one is best suited for analytical workloads?
5. How do you secure sensitive data in Azure Data Factory pipelines?
6. How do you handle error logging & alerting in ADF or Databricks pipelines?
7. How do you implement data lineage and cataloging in Azure?

Ankita Gulati Shubh Goyal

8. What is the role of Delta Lake in Databricks? Why is it important for data engineering pipelines?
9. What kind of errors have you faced in Azure Databricks jobs, and how did you resolve them?
10. Explain RDD, DataFrame, and Dataset in Spark. How do they differ?
11. What do you consider while optimizing a PySpark job (partitioning, shuffling, skew, caching, etc.)?
12. Explain the difference between HAVING and WHERE clauses in SQL with examples.
13. What is the difference between DELETE and TRUNCATE? Which one can be rolled back?
14. What is the difference between Primary Key and Unique Key in a database?
15. What is a complex JSON? How do you handle it in PySpark or ADF pipelines?
16. Describe a scenario where you had to migrate data from on-premises SQL Server to Azure (ADF/Synapse).

Ankita Gulati Shubh Goyal

17. Suppose your Azure subscription is about to expire. How do you move resources to a new subscription?
18. Write SQL query to pivot sales data (months as columns, countries as rows).
19. Given a nested JSON file, write PySpark code to flatten it into a structured DataFrame.
20. Write Python code to generate 5 files from 1000 records, where each file contains a maximum of 200 records.
21. Write PySpark code to handle skewed data during joins.
22. Write SQL query to get employees with salary greater than department average.
23. Write PySpark code to calculate daily active users (DAU), weekly active users (WAU), and monthly active users (MAU) from a login dataset.

Thank You

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

Ankita Gulati

Shubh Goyal