Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹12-16 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 Screening

1. Walk me through a recent project where you built a data pipeline using Azure.
2. What were the challenges you faced, and how did you resolve them?
3. Write a query to find the second-highest salary from an Employee table without using TOP or LIMIT.
4. Write a query to find employees with duplicate salaries.
5. Explain window functions and write a query to calculate the running total of sales per month.
6. Difference between DELETE, TRUNCATE, and DROP.
7. Reverse a string in Python without using built-in functions.
8. Write a function to return the first non-repeating character in a string.
9. What are Python generators and when would you use them?

10. Difference between *args and **kwargs.
11. What is the difference between Azure IaaS and PaaS?
12. Key components of Azure Data Factory (ADF).
13. How does Linked Service work in ADF?
14. Explain different Integration Runtime types (Azure IR, Self-hosted IR, SSIS IR).

Ankita Gulati

Shubh Goyal

Round 2
Advanced Data Engineering

1. Explain the difference between Mapping Data Flow and Copy Activity in ADF.
2. How would you handle incremental loads in ADF (watermarking, control tables, LastModified filters)?
3. How do you debug and monitor failed ADF pipelines?
4. Difference between RDD, DataFrame, and Dataset in Spark.
5. Explain Broadcast joins vs Repartition vs Coalesce.
6. How do you handle data skew in Spark jobs?
7. What is the difference between left anti join and left semi join?
8. Difference between cache() and persist().
9. Why is Parquet preferred over CSV/JSON in big data pipelines?
10. Difference between ADLS Gen1 and Gen2.

Ankita Gulati

Shubh Goyal

11. Explain Delta Lake schema evolution and MERGE operations.
12. How would you optimize Azure Synapse/Databricks performance and cost?
13. Design a pipeline to handle both batch and streaming data using ADF, Event Hubs, and Databricks.
14. Design a retail analytics pipeline with historical + streaming datasets.
15. How do you ensure fault tolerance and high availability in Azure?
16. How do you implement RBAC (Role-Based Access Control) for secure data access?
17. What is Azure Purview, and how does it help with governance and lineage?

Ankita Gulati

Shubh Goyal

Round 3
HR Discussion

1. Tell me about a production issue you faced and how you resolved it.
2. Describe a situation where you had to manage conflicting stakeholder requirements.
3. Give an example of collaborating with data scientists or business analysts on a project.
4. How do you ensure data quality in your pipelines?
5. Why do you want to join TCS?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal