Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Azure Data Engineer
- **Experience:** 3+ years
- **Location:** Pune / Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹10-15LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. Data Engineering Concepts

Ankita Gulati

Shubh Goyal

Round 1
Technical Foundations

1. Walk me through your past projects. What were the key challenges you faced and how did you solve them?
2. Can you explain the end-to-end architecture of one of your recent data engineering projects?
3. How does Databricks architecture work, and how is it integrated with Azure services?
4. What performance optimization techniques have you implemented in Databricks?
5. How would you fetch data from Azure Data Lake Storage (ADLS) and load it into a SQL database?
6. Write a PySpark script to read data from ADLS and write it into a SQL database.
7. Explain different types of SQL Joins: Inner, Left, Right, Full Outer, Cross, Semi.
8. With sample data, demonstrate the difference between Inner Join vs Left Join.

Ankita Gulati Shubh Goyal

9. Write a PySpark script to perform joins between two DataFrames.
10. Write a PySpark script to drop specific columns from a DataFrame.
11. Write a PySpark script to filter rows based on given conditions.
12. How do you partition and bucket tables in Spark? Why are they important?
13. Explain the difference between narrow vs wide transformations in Spark.
14. Write an SQL query to find the second highest salary per department using window functions.
15. Write a Python function to check if a string is a valid palindrome ignoring spaces and special characters.

Round 2
Advanced Techno-Managerial

1. Deep dive into your project design — what scalability challenges did you face and how did you optimize performance?
2. What real-world issues (like skewed data, late-arriving data, or job failures) did you face, and how did you handle them?
3. How do you handle failed or late records using Azure Stream Analytics?
4. What are the best practices for checkpointing, event retention, and reprocessing in a streaming system?
5. You need to find the top-selling product for each location — write a PySpark script using window functions.
6. Write a PySpark script to identify duplicate records in a DataFrame.
7. Explain the Spark Architecture: Driver, Executors, Cluster Manager.

Ankita Gulati

Shubh Goyal

8. Walk me through the internal working of Spark — DAG creation, Stages, Tasks, Execution Plan.

9. What types of activities are available in Azure Data Factory (ADF) (Data Flow, Copy, Lookup, ForEach, etc.)?

10. How do you schedule and monitor jobs in ADF?

11. How do you optimize queries in Big Data pipelines to handle billions of records efficiently?

12. Write an SQL query to get the moving average of sales per day for the last 7 days.

13. Write a Python program to implement a custom context manager (without using contextlib).

14. What is the difference between exactly-once, at-least-once, and at-most-once delivery in streaming systems?

15. How do you design a data lake vs data warehouse architecture for an analytics-heavy system?

Ankita Gulati

Shubh Goyal

Round 3
HR Discussion

1. Tell us about yourself and your long-term career goals.
2. Why are you looking to move from your current company?
3. What are your salary expectations?
4. How do you see your career growing at EY?
5. What excites you about EY’s data engineering projects and culture?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal