UBS

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹25-30 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

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

Ankita Gulati

Shubh Goyal

Round 1
Data Engineering & Optimization

1. Can you walk me through your professional experience, daily responsibilities, the technologies you have worked with, and how much time you typically spend on each technology?
2. Given a sales table, write queries/code to calculate country-wise total sales using:
   → SQL
   → Python (Pandas)
   → PySpark
3. Write a SQL query to retrieve the second highest salary from an employee table.
4. Write a SQL query to identify duplicate records in a table and display their counts.
5. Write a SQL query to calculate a running total of sales by date for each region.
6. Write a SQL query to get the top 3 highest transactions per customer.

Ankita Gulati Shubh Goyal

7. Given a transaction table, write SQL to find customers who had no transactions in the last 3 months.

8. Write a SQL query to calculate Year-over-Year (YoY) growth in sales for each product.

9. In PySpark, write a job that reads a JSON file with nested structure, flattens it, and writes output as Parquet.

10. In PySpark, how would you join a large fact table with a small dimension table efficiently? (Write sample code using broadcast join).

11. Write a PySpark job to calculate the moving average of stock prices from streaming data.

12. Write a Python program to read a CSV file, group by customer, and calculate total purchases and average purchase value.

13. In PySpark, how do you implement window functions (e.g., rank, row_number, lag/lead)? Write code to find the previous transaction amount for each customer.

14. You are given a large dataset of financial transactions in CSV format. Write a PySpark pipeline to:

→ Read raw data from S3/Azure Blob.
→ Clean null/invalid values.
→ Partition by transaction_date.
→ Write output in Delta format.

15. Write SQL to pivot transaction data such that months become columns and total sales are aggregated.

16. You have two datasets:

→ Customer Data (customer_id, name, country)
→ Transaction Data (transaction_id, customer_id, amount, date)

Write a PySpark job to calculate total amount spent by each customer in 2024.

17. Write SQL to find the highest spending customer in each country.

18. Write a SQL query to detect gaps in transaction dates for each customer (i.e., days without activity).

19. In PySpark, write code to handle schema evolution in Delta Lake (adding a new column).

20. Write Python code to parse a complex nested JSON, extract only required fields, and convert it into a flat CSV.

Ankita Gulati

Shubh Goyal

Round 2
Data Architecture & Cloud Engineering

1. How would you implement an incremental load in Azure Synapse Analytics?
2. If you need to implement data security and compliance policies in Synapse, how would you approach it?
3. Can you explain the procedure for migrating on-premise SQL Server data to Azure Data Factory (ADF)?
4. Suppose your Azure subscription is about to expire. How would you migrate to a new subscription while minimizing downtime?
5. What types of storage options are available in Azure, and in which scenarios would you use each?
6. How do you perform error handling in data pipelines, and what types of alerts/monitoring mechanisms do you use?
7. How do you ensure sensitive data security in Azure Data Factory?

Ankita Gulati

Shubh Goyal

8. How do you manage data lineage and data cataloging in Azure?
9. What is the role of Delta Lake in Databricks?
10. What are some common errors you have faced in Azure Databricks jobs, and how did you resolve them?
11. Describe the architecture of a cloud-based data warehouse (e.g., Snowflake, BigQuery, or Synapse).
12. What is the difference between ETL and ELT? Which approach do you prefer and why?
13. What is the role of Apache Kafka in data engineering pipelines?
14. How do you ensure data quality and governance in large-scale data pipelines?
15. How do you design a scalable, fault-tolerant ETL/ELT pipeline for financial data at an enterprise level like UBS?
16. Write a SQL query to calculate a running total of monthly transactions for each customer.
17. How would you handle schema evolution in a Delta Lake table?

Ankita Gulati Shubh Goyal

18. Given streaming data from Kafka (e.g., stock prices), how would you process and aggregate it using Spark Structured Streaming?

19. In PySpark, how would you join a large fact table with a small dimension table efficiently?

20. Design a pipeline in Azure Data Factory that ingests raw CSV data, transforms it using Databricks, and loads it into Synapse. Explain failure recovery steps.

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal