CGI

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Bangalore / Pune
- **Work mode:** Hybrid
- **Compensation:** ₹22-24 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

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

Round 1
Data Engineering & Cloud

1. Can you describe some Agile ceremonies you were a part of in your previous projects?
2. What are the different roles in Agile methodology?
3. What is the difference between a sprint backlog and a product backlog?
4. Who is responsible for owning the product backlog?
5. What is Azure Data Lake, and how is it typically used in data engineering?
6. What are the common file formats used in data lakes? Suppose you have a file in a data lake or container that you cannot download — how would you read it?
7. How do you connect to Azure Data Lake using Databricks?
8. In Azure, if you need to read a parquet file from Databricks and extract secrets required for connection, how would you securely fetch these secrets without exposing them to the team?

9. What are the limitations of Azure Key Vault?
10. What is the difference between Azure containers and file storage within a data lake? Can you provide an example?
11. If you have a CSV file, can it be stored in both an Azure container and a file share?
12. Have you worked on data warehousing concepts?
13. What is the difference between a fact table and a dimension table?
14. How do you design and implement a slowly changing dimension (SCD) in a data warehouse?
15. What are the key differences between batch processing and real-time streaming in data engineering?
16. How would you optimize storage and query performance in Azure Synapse or Databricks?
17. Can you explain how partitioning and bucketing improve query performance in Spark/Databricks?
18. What are different methods to implement data versioning in data lakes?

Round 2
Advanced Engineering

1. Why do we use Delta tables in Databricks? What is the purpose of time travel in Delta, and why can’t the same functionality be achieved using CSV files?
2. You are asked to create a pipeline that pulls data from a REST API using ADF, and the output should be in a format other than JSON. How would you design this pipeline?
3. What is the use of the Lookup activity in ADF?
4. Suppose you have created a pipeline, and in the destination, you need to make the sink file dynamic. How would you implement this?
5. Coding Question:

-->Input: data = ["ABC", "ABC", "DEF", "XYZ", "XYZ", "XYZ"]
-->Output: Return the strings sorted in ascending order based on their frequency count. Write code in PySpark.

6. What are Logic Apps in Azure, and when would you use them in a data pipeline?
7. Suppose you created a pipeline in ADF that loads data into ADLS Gen2, and it worked fine for 6 months. Later, the Power BI user consuming this data reports that dashboards are taking too long to refresh. How would you troubleshoot and improve the performance?
8. What are the main differences between AWS and Azure in terms of data engineering services? Which one do you find more comfortable and why?
9. SQL Coding: Given a Sales table, write a query to find the top 3 products by total revenue for each country.
10. PySpark Coding: Given a DataFrame of transactions, write code to implement window functions to calculate the cumulative sales per customer.

11. System Design: Design a real-time pipeline that ingests streaming data from Kafka into Azure Data Lake, applies transformations in Databricks, and makes it available for reporting in Power BI.

12. Error Handling: How would you implement retry policies and failure alerts in ADF pipelines?

13. Data Governance: How would you manage data lineage and metadata for an enterprise-wide data platform in Azure?

14. You have a JSON log dataset in Azure Data Lake containing user events with nested metadata. Write PySpark code to flatten the data, calculate each user’s average daily session length, find their top 3 event types by frequency, and save the result as a Parquet file partitioned by event_date.

Thank You

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

Ankita Gulati

Shubh Goyal