Capgemini

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 6+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹25-28 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. Airflow
5. System Design & Problem-Solving

Ankita Gulati

Shubh Goyal

Round 1
Data Processing & Optimization

1. A file lands in S3. How would you ensure that only incremental files are processed instead of reprocessing the full dataset?
2. How can we implement incremental data processing from S3 architecturally?
3. Write a SQL query: Given a Customer table (CustomerID, CustomerName) and an Orders table (OrderID, OrderDetails, CustomerID), find all customers who never placed an order.
4. Write a SQL query to retrieve the second highest salary per department from an Employee table.
5. When someone asks for the “second highest” or “third highest” value, what is the most efficient SQL approach to achieve this?
6. You have two Parquet files in S3. You need to read them, perform a left join on a column, apply filters, reverse the column order, and write the output back to S3 in Parquet format. Write the PySpark code for this.

7. If DF1 is significantly larger than DF2 in a join operation, how would you optimize the join in PySpark?
8. Suppose your SQL query is running slower than expected. What techniques would you use to optimize query performance?
9. Architecturally, isn't using a ranking function for "second highest" queries expensive? As an architect, how would you handle this requirement efficiently?
10. Explain the difference between partitioning and bucketing in Spark, and when to use each.
11. How do you identify and handle data skewness in Spark?
12. How would you design a data validation framework to ensure data quality (standardizing departments, dates, ages, etc.) for all incoming files?
13. If this framework must be reusable and extensible for new rules, how would you design it to avoid updating code repeatedly?

Round 2
System Design & Architecture

1. When data ingestion happens, where would you land the data in AWS initially, and why?
2. After landing data in AWS, how would you design the rest of the process in the pipeline?
3. The requirement states the architecture should be serverless. How would you design a serverless data pipeline in AWS?
4. You have data in S3, and you want to move it to Athena for querying. How would you design the end-to-end pipeline? Which AWS components would you use?
5. Apart from AWS Glue, what other AWS components could you use for ETL orchestration?
6. Before pushing cleansed and validated data into EMR, how would you orchestrate data validation across different stages?
7. How does DynamoDB get the threshold information for validations?

Ankita Gulati Shubh Goyal

8. Suppose DynamoDB does not have the threshold declared. How would you design a process to determine thresholds dynamically instead of manual updates?
9. If a field such as a flight number has complex values (e.g., “Datum plus 7”), how would your standardization framework handle such cases in the background?
10. If function requirements state that users should be notified when values exceed a threshold, but ingestion happens via SFTP (batch), how often would you collect the data, and how would you handle delays in notifications?
11. Even though frequent batches are running, the system is still batch-oriented. How would you redesign it to meet near real-time processing requirements?
12. For performance optimization, suppose your Spark job is running slowly. What are the steps you would take to investigate and resolve the issue?

13. In a data warehouse environment, how would you ensure incremental data loading without reprocessing the full dataset?

14. Can you design a data lake architecture that supports ingestion, validation, standardization, and querying in a scalable, serverless, and cost-efficient way?

15. How would you handle a large-scale join between two datasets in AWS (one very large, one small) to ensure performance and cost optimization?

Thank You

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

Ankita Gulati

Shubh Goyal