CAPCO

a wipro company

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 6+ years
- **Location:** Bangalore Pune
- **Work mode:** Hybrid
- **Compensation:** ₹22-25 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

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

Ankita Gulati

Shubh Goyal

Round 1
Data Foundations & Architecture

1. Can you explain the Hadoop architecture and its core components?
2. Walk me through your end-to-end data pipeline implementation in a recent project.
3. What was the largest volume of data you've handled, and how did you manage performance and scalability?
4. Describe a complex scenario or technical challenge you faced in your data engineering career and how you solved it.
5. What are the differences between batch processing and stream processing? When would you use each?
6. Explain the role of data partitioning and bucketing in improving query performance.
7. Compare Parquet, ORC, and Avro file formats. Which one do you prefer and why?

Ankita Gulati Shubh Goyal

8. Explain Spark architecture and the role of Catalyst Optimizer.
9. What are different types of joins in Spark/SQL and their use cases?
10. How do you ensure data quality, consistency, and lineage in large-scale pipelines?
11. What are the common cloud services for data engineering (AWS/GCP/Azure) you have worked with?
12. How does distributed storage (HDFS/S3) differ from a traditional file system?
13. Explain the concept of dynamic resource allocation in Spark.
14. How do you approach optimizing joins and aggregations in Hive/Spark SQL?
15. Can you explain the importance of checkpointing and fault tolerance in Spark Streaming?

Round 2
Applied Data Engineering

1. Write a Python function to find common elements between two lists:

→ list1 = [1,2,3,4,5]

→ list2 = [3,5,6,7,8]

Output: [3,5]

2. Given a string:

→ text = "tuvxaaaajkluiammmmmmmmm"

→ vowels = "aeiou"

Write Python code to count repeated vowels in the text.

3. Write a Python program to check if two strings are anagrams. Example: "listen" and "silent".

4. Implement a Python function to find the longest substring without repeating characters.

5. You are given a log file (large text) stored in S3. Write PySpark code to extract the top 10 most frequent error messages.

Ankita Gulati

Shubh Goyal

6. Write a Python function to generate a Fibonacci sequence using recursion and generators.
7. Given a list of numbers, write Python code to find pairs that sum up to a target value K.
8. Write a SQL query to display the top 2 highest salaries from each department.
9. You have an orders table with order_id, customer_id, order_date, amount. Write a query to find the customer with the highest total spend.
10. Write a SQL query to return duplicate records from a table.
11. Write a query to find the second highest salary in a table without using LIMIT or TOP.
12. You have a transactions table with transaction_id, account_id, amount, date. Write a query to find the moving average of the last 3 transactions per account.
13. Write a SQL query to find employees who earn more than their department's average salary.
14. Write a query to pivot sales data to show total sales by month for each product.

Ankita Gulati Shubh Goyal

15. How would you design a data lakehouse architecture on AWS using S3, Glue, Redshift/Snowflake?
16. How do you secure data pipelines on the cloud (IAM roles, encryption, VPC, etc.)?
17. If your Spark job on EMR keeps failing due to memory issues, how would you debug and optimize it?
18. Explain how you would handle schema evolution in Parquet files stored in S3.
19. How would you set up a real-time data streaming pipeline using Kafka + Spark Structured Streaming?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal