Synechron

Digital / Business Consulting / Technology

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 4+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹24-26 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS, Azure
4. Data Modeling & Governance
5. Big Data & Distributed Systems

Ankita Gulati

Shubh Goyal

Round 1
Core Data & Foundations

1. If a Spark job is taking longer than expected, what troubleshooting steps would you take to optimize its performance?

2. In which scenarios would you prefer using a Broadcast Join in Spark, and why?

3. Can you explain what shuffling is in Spark, and why it can impact performance?

4. What is the difference between RDDs and DataFrames in Spark? Provide practical use cases.

5. What do you understand about the Catalyst Optimizer in Spark?

6. Explain the difference between cache() and persist() in Spark. In which scenarios would you use each?

Ankita Gulati Shubh Goyal

7. Can you explain the Spark architecture and how execution happens internally?

8. How do you optimize PySpark transformations for large datasets in real-world projects?

9. Write PySpark code to join two DataFrames and return only matching records.

10. Write PySpark code to segregate a raw CSV dataset by bank name and save separate files for each bank.

11. Write PySpark code to clean a column containing inconsistent phone numbers (e.g., +91-, 91-, with/without prefix) and standardize them to a 10-digit format.

12. A Hive query is taking too long on a large dataset with multiple joins and aggregations. How would you optimize it?

13. What are the differences between an external table and an internal table in Hive?

14. Suppose a Hive table has missing records for certain columns. How would you handle this scenario while querying the data?

15. What is the default replication factor in HDFS?

16. What is the default block size in HDFS?

17. If the NameNode fails in a Hadoop cluster, what will be the impact and how is it handled?

18. You are working on a CSV file stored in HDFS, which is accessed frequently by multiple users but is taking too long to read. How would you optimize access?

19. Explain your understanding of HDFS architecture and how data is stored and retrieved.

20. What is your understanding of HBase? How does it differ from traditional RDBMS?

21. Explain the concept of NoSQL databases. When would you choose them over relational databases?

22. What is the role of a Lambda function in Python? Provide examples of when you would use it.

23. Write Python code to read a text file, count the frequency of each word, and return the results.

24. Suppose you have a text file containing city and country names along with word counts. Write Python code to search and return all occurrences of a particular keyword.

25. Given a list [7, 8, 1, 2, 3, 5, 8, 2], write Python code to find the median and count the occurrence of each element.

Ankita Gulati Shubh Goyal

Round 2
Cloud & Advanced

1. Suppose a Glue job must be triggered immediately after a file arrives in an S3 bucket. How would you design this solution?

2. Can you use multiple Python files in AWS Glue? If yes, how would you manage them?

3. What instance types have you used in AWS Glue, and why did you choose them?

4. What is a DynamicFrame in AWS Glue, and how does it differ from a DataFrame?

5. What are the prerequisites for reading a file from S3 in AWS Glue?

6. When running Glue jobs, certain IAM policies are required. Which policies would you attach and why?

Ankita Gulati Shubh Goyal

7. What is the AWS Glue Data Catalog, and how is it used in a pipeline?

8. In your raw dataset, customer names and bank names are stored in CSV format. Future data may include additional banks. How would you design your pipeline to dynamically segregate the data by bank name and write it back to S3?

9. What do you understand by horizontal scaling and vertical scaling in AWS? Provide real-time examples.

10. Why do we use AWS Step Functions? Can you provide a real-time use case from your projects?

11. How does AWS Lambda handle concurrency?

12. What is the difference between ECS and EKS, and which one would you choose in a real-time scenario?

13. What is Azure Data Factory, and how does it integrate with other services?

14. How would you implement incremental data loading in ADF from an on-premises SQL Server to Azure Data Lake?

15. What are the different types of triggers available in ADF?

16. What are the differences between Dedicated SQL Pool and Serverless SQL Pool in Azure Synapse Analytics?

17. Write an SQL query to find the top three highest salaries for each department.

18. Write an SQL query to fetch customers who placed an order exactly one month ago.

19. What are window functions in SQL? Provide an example.

20. What are common query optimization techniques for large datasets?

Ankita Gulati Shubh Goyal

21. You are working in a Hadoop environment and frequently face out-of-memory errors during execution. How would you handle and prevent this?

22. Suppose a Hive query is running on billions of rows and applying filters on top of aggregations. How would you design your query and cluster resources to optimize performance?

23. Describe how you would design a pipeline to read raw data from an S3 bucket, validate it, transform it, and load it into Redshift while handling schema evolution.

24. You are given a student table with columns (ID, Name, Phone). Some name values contain invalid UTF-8 or Chinese characters. How would you clean this dataset such that invalid names are replaced with NULL? (Write code in Python.)

Thank You

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

Ankita Gulati

Shubh Goyal