Straive

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ 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. Big Data
6. System Design & Problem-Solving

Round 1
Data Engineering Foundations

1. Can you elaborate on your experience with AWS Glue and Crawlers, and explain how you used them to improve the efficiency of ETL pipelines in your past projects?
2. Can you share an example of a complex data problem you solved using SQL and Python libraries such as Pandas, NumPy, Matplotlib, or Seaborn?
3. How do you handle situations where data is conflicting across multiple sources?
4. Can you explain your experience with real-time analytics on streaming data? Which tools or services did you use?
5. What are some of the basic SQL query optimization techniques you have applied to improve performance?
6. Write a SQL query to find the department with the 3rd highest salary.

Ankita Gulati

Shubh Goyal

7. Write a SQL query to delete duplicate records from an existing table.
8. Write a SQL query to swap gender values in a column so that “male” becomes “female” and “female” becomes “male.”
9. Explain how you would optimize a slow-running SQL query in a production database.
10. What is data warehousing, and how is it different from a traditional database?
11. What strategies do you use to ensure high-quality data pipelines with validation, metadata management, and monitoring?
12. How would you implement data encryption in Amazon S3 for sensitive data? What best practices should be followed?
13. For a Spark job running slower than expected, what are the key steps you would take for performance optimization?
14. Can you explain the concept of data skewness in distributed systems and how to handle it in Spark?

15. What is the difference between Order By vs Sort By in SQL or Spark?

16. What is the difference between ReduceByKey and GroupByKey in Spark? Which one would you prefer in large-scale data processing, and why?

Round 2
Data Engineering Depth & Design

1. Explain the differences between RDD, DataFrame, and Dataset in Spark. When would you use each?
2. What is checkpointing in Spark, and why is it important?
3. Explain partitioning vs bucketing in Spark. When would you prefer one over the other?
4. What is the role of the Catalyst Optimizer in Spark SQL?
5. What are the different types of joins in Spark SQL? Can you explain their use cases?
6. Explain the role of serialization in Spark and how it impacts performance.
7. What is the use of the coalesce() function in Spark, and how is it different from repartition()?
8. What are Python generators? In what scenarios would you prefer to use them?

Ankita Gulati Shubh Goyal

9. What are Python decorators, and how do they work? Can you give a real-world use case?
10. What is Python’s garbage collector, and how does it manage memory?
11. What is the difference between iterators and iterables in Python?
12. What are Python magic methods? Can you give examples of how they are used?
13. What are static methods and class methods in Python? How do you implement them?
14. What is a descriptor in Python, and how is it used?
15. What are the advantages of using context managers in Python (with statement)?
16. Write Python code to reverse the words in the string “God is great” to get "great is God", without using built-in reverse functions.
17. Write Python code to create an Employee class that accepts an employee’s name and salary, and then prints them.
18. Explain the difference between Rank() and DenseRank() in SQL. Provide an example.

17. What is serialization in Spark, and why is it important?
18. Could you provide a technical overview of your project? What technologies are you using, and what is the main focus?
19. Are you using Git as a repository? If yes, explain how.
20. What is the time limit for an AWS Lambda function?
21. Have you worked with Scala? If yes, in what context?
19. Write a Python script using Ansible to automate the deployment of a web application on AWS EC2 instances.
20. How would you design and implement a serverless data pipeline using AWS Lambda, Step Functions, and S3?
21. Suppose you are working on a large-scale streaming pipeline. How would you ensure fault tolerance, scalability, and real-time accuracy?

22. How would you design a near real-time pipeline for fraud detection using AWS Kinesis or Kafka?
23. Can you explain a data migration strategy you would follow when moving from on-premises to AWS Redshift?
24. How do you manage schema evolution in a data lake with Parquet or Avro formats?
25. Explain the trade-offs between batch vs streaming pipelines in terms of cost, latency, and scalability.

Thank You

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

Ankita Gulati

Shubh Goyal