IRIS

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 5+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹19-22 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

Round 1
Data Systems & Big Data Core

1. Tell me about yourself and explain how you have used PySpark in one of your projects.
2. What is your understanding of MapReduce vs. Spark? How are they similar, and how are they different?
3. Can you explain the Spark architecture (driver, executor, cluster manager)?
4. What is the difference between RDD and DataFrame? If I say a DataFrame is made up of RDDs, is that correct?
5. What is the difference between narrow and wide transformations in Spark?
6. Can we reduce the number of partitions in Spark? Why would you prefer coalesce over repartition?
7. Suppose you have to process a 1024 MB file stored in HDFS with multiple executors (each one core). How many executors would be sufficient to minimize processing time?

8. Suppose you have a small parent table (1,000 rows) and a huge child table (5M rows). What join strategy would you use to optimize performance?
9. What happens internally when you use broadcast in Spark (for tables or variables)?
10. What are the different join strategies in Spark?
11. Give a use case where you would use sort-merge join and another where hash join would be better.
12. Explain SQL joins (INNER, LEFT, RIGHT, FULL OUTER, UNION) with example outputs for two sample tables.
13. Write a SQL query to display the top two highest salaries from each department.
14. Suppose you are applying a filter transformation in Spark in two steps (first on percentage, then on below-50%). Which transformations would you use?
15. Explain your experience with relational database systems and how you optimized queries for large datasets.
16. What challenges have you faced in handling large-scale pipelines, and how did you solve them?

Round 2
Python, SQL & Cloud Engineering

1. Tell me about yourself.
2. What is the difference between list and tuple in Python?
3. What is a class in Python? How do you define and create an object?
4. What is a constructor in Python? Provide a code example.
5. What is the difference between instance variable and class variable?
6. What are decorators in Python? Explain their functionality with an example.
7. How do you handle exceptions in Python? Explain try, except, finally, else blocks with an example.
8. What is an object in Python, and how do you create one?
9. Write a Python program to count the frequency of each non-space character in a given string and return the result as a dictionary.

Ankita Gulati

Shubh Goyal

10. Write a Python program to find the common elements between two lists:

→ list1 = [1, 2, 3, 4, 5]
→ list2 = [3, 5, 6, 7, 8]
# Expected Output → [3, 5]

11. Given the string:

→ text = "tuvxaaaajkluiammmmmmmm"
→ vowels = "aeiou"

Write Python code to count the number of repeated vowels.

12. What is lazy evaluation in Spark, and why is it important?
13. What is serialization in Spark, and why is it required?
14. What are the differences between RDD, DataFrame, and Dataset? When would you prefer RDD?
15. Advanced SQL: Write a query to find the most frequently ordered product per customer from an orders table with columns: order_id, date, customer_id, product_id.

16. Cloud-based: How would you integrate Spark with AWS or Azure for data ingestion and processing?

17. In cloud pipelines, how would you handle scalability and cost optimization while running Spark jobs?

Thank You

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

Ankita Gulati

Shubh Goyal