KPMG

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹25-30 LPA
- **Total Rounds:** 3
- **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
Foundations & Problem-Solving

1. Can you introduce yourself and describe the projects you have worked on, including the technologies used?
2. What does your typical day-to-day work look like as a Data Engineer?
3. Why did you choose the current tech stack you are working with, and what are its advantages?
4. What alternatives exist to the Medallion Architecture?
5. What kind and size of data do you handle on a daily basis?
6. Suppose the business is using JSON as the file format. How would you convince them to move to Parquet, and what benefits would you highlight?
7. Given a DataFrame, how would you split the data into two columns ("Even" and "Odd"), ensuring even numbers are populated into one column and odd numbers into the other?

8. Given an array, write a program to find the minimum and maximum elements.

9. You are given a table Matches(Country) with values: India, Australia, Pakistan. Write a query to generate the following output:

→ India vs Australia
→ India vs Pakistan
→ Australia vs Pakistan

10. You are given two tables:

→ Table A: 1, 1, 1, 1
→ Table B: 1, 1, 1

Find the count of records returned by a Left Outer Join and an Inner Join.

11. Consider the following data: 85, 85, 80, 75, 75, 70. Show the difference in output between DenseRank() and Rank().

Round 2
Advanced Concepts & Hands-On

1. Can you explain the overall architecture of Apache Spark?
2. Walk me through the process of how jobs are executed in Spark.
3. What role does the Catalyst Optimizer play in Spark query execution?
4. What is the difference between a Logical Plan and a Physical Plan in Spark?
5. What are the key differences between ORC and Parquet file formats?
6. Write code to read a CSV file and create a DataFrame with appropriate schema properties.
7. Write code to create a DataFrame with two columns:
→ one defaulting to String and the other to Integer.

Ankita Gulati

Shubh Goyal

8. Given a string :-

→ s = "aaabbbccddeeeee"

write a program to output the count of each character in dictionary format:

→ { "a": 3, "b": 3, "c": 2, "d": 2, "e": 5 }

9. Write a Python program to count the number of occurrences of a given word in a text file.
→ Example: Count occurrences of the word “The” in:
"The lazy fox jumps over the sleeping rabbit. The lazy rabbit doesn’t wake up."

10. You are given two tables:

→ Table 1 = col 1: 1, 1

→ Table 2 = col 1: b, a, 1

Write queries to demonstrate the results of Inner Join, Left Join, Right Join, and Full Join.

11. How would you design a schema for handling slowly changing dimensions (SCD Type 2)?

Round 3
HR & Behavioral

1. What is the difference between a Data Lake and a Delta Lake?
2. Why did you quit your previous job?
3. Even though you have an offer in hand, why did you apply again?
4. If we provide you the same compensation as your current offer, would you still consider joining KPMG?
5. You are settled in Hyderabad. Why are you willing to relocate to Bangalore?

Thank You

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

Ankita Gulati

Shubh Goyal