Globant

# Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

# Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹22-24 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS, Azure
4. Airflow
5. System Design
6. API Error Handling
7. Testing Frameworks

Ankita Gulati

Shubh Goyal

# Round 1
# Python & Data Eng Concepts

1. Can you explain some challenges you faced in your projects and how you overcame them?
2. Walk me through the architecture of your current or a previous project you worked on.
3. What is the difference between composition and inheritance in Python? Can you give an example of when you would use each?
4. How do you handle multiple inheritance in Python? What potential issues can arise, and how do you resolve them?
5. Can you describe a scenario where you would use a Trie data structure? How would you implement it in Python?
6. How do you handle errors and exceptions in a Python API?
7. How would you approach testing and validation in Python? What testing frameworks and tools do you commonly use?

Ankita Gulati Shubh Goyal

8. What are Python generators and decorators, and when would you use them?
9. Can you describe a time when you optimized a piece of code or script for better performance? What was the challenge, what changes did you make, and what was the outcome?
10. What is the purpose of creating custom data structures in Python? Can you explain how you would implement one?
11. Can you explain the different data types and data structures available in Python?
12. Is it always possible to avoid wide transformations in data processing (e.g., PySpark)? Or are there scenarios where they are necessary?
13. If incoming data must be stored in sorted order automatically, which data structure would you choose and why?
14. Write a Python function to duplicate each character in a given string.
15. You are given a string containing alphabets, numbers, and special characters. Write a program to separate them into three different strings.

16. Given a mapping of Managers and Reportees, write a function that checks whether a Manager and Reportee are connected directly or indirectly in the reporting chain.

17. Merge two sorted lists into one sorted list without using the sort() function.

18. You are given a list of numbers and a target value. Write a function to return the indices of the two numbers that add up to the target.

# Round 2
# SQL & Cloud/Data Systems

1. How would you optimize join performance when working with large datasets?
2. How would you use joins to eliminate duplicate rows in a result set? What are alternative approaches?
3. What is the difference between clustered and non-clustered indexes? How do you decide which one to use?
4. Can you explain what a CTE (Common Table Expression) is and when it should be used?
5. What role do data types play in SQL performance and data integrity?
6. How would you manage large-scale data backups in AWS?
7. Can you give a real-time use case where AWS Elastic Beanstalk improves deployment?
8. When would you choose Parquet over CSV or JSON formats, and how do you make that decision?

Ankita Gulati

Shubh Goyal

9. Can you explain the difference between Star Schema and Snowflake Schema in data modeling?

10. In what scenarios would you use AWS Lambda vs. AWS Glue?

11. Given two tables:

--> Table A → IDs [1, 2, 1, 3]

--> Table B → IDs [1, 2, 3, 4]

Show the results of INNER JOIN, LEFT JOIN, RIGHT JOIN, and FULL OUTER JOIN.

12. Write a SQL query to delete duplicate records in the Employee table (based on name, email, etc.) while keeping only one copy.

13. Given a transaction table with columns (TransactionID, Product, Spend, Year), how would you aggregate data at the Year and Product level to calculate the current year spend?

14. If you have two tables with no common key, how would you join them?

15. Write a SQL query to fetch all employees’ names who belong to the Engineering department and are working on any project.

16. In an Employee table with fields (EmployeeID, FirstName, LastName, Place, DateOfBirth), how would you ensure that:

--> EmployeeID is auto-incremented for each new record.
--> Duplicate entries (same FirstName, LastName, DateOfBirth) are not inserted?

# Thank You

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

Ankita Gulati

Shubh Goyal