EXL

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹23–30LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark
3. ETL Development
4. AWS
5. Data Modeling
6. Data Warehousing

Ankita Gulati Shubh Goyal

Round 1
Python, SQL & Fundamentals

Python

1. Describe your current project and explain your role specifically related to Python.
2. What are Python generators? In what scenarios would you use them?
3. What are Python decorators? Provide an example use case.
4. Explain the concept of iterators and list comprehensions in Python.
5. What is the difference between static methods and class methods in Python?
6. What are magic (dunder) methods in Python?
7. What are the advantages of using context managers in Python?
8. Compare inheritance and composition in Python with an example.
9. How do you handle errors and exceptions when developing Python APIs?

SQL

10. Write an SQL query to find the department with the 3rd highest salary.
11. Write an SQL query to delete duplicate records from a table.
12. What is the difference between RANK() and DENSE_RANK() functions?
13. What are Common Table Expressions (CTEs), and when would you use them?
14. What is the difference between Clustered and Non-Clustered indexes? When should each be used?
15. What role do data types play in SQL performance and data integrity?

Miscellaneous

16. Do you use Git as a version control system? How?
17. What major challenges have you faced in your projects, and how did you overcome them?

Round 2
Spark + Advanced SQL + AWS

Spark

1. Explain the architecture of Apache Spark.
2. What file formats are supported in Spark/Hadoop?
3. Compare Avro and Parquet file formats.
4. What is the role of the Catalyst Optimizer in Spark SQL?
5. Compare DataFrame and Dataset APIs in Spark.
6. Explain the difference between reduceByKey and groupByKey in Spark.
7. What is the difference between partitioning and bucketing? When would you use each?
8. What is checkpointing in Spark, and why is it used?
9. What is data skewness in Spark, and how can it be handled?
10. What is the difference between orderBy and sortBy in Spark?

SQL

11. Write an SQL query to find the most frequently ordered product per customer.
12. Write an SQL query to update the gender column such that all values of “male” become “female” and vice versa.
13. How would you optimize a slow-running SQL query?
14. How can you use joins to eliminate duplicate rows? What are alternative approaches?

AWS / Cloud

15. What are the maximum execution time limits for AWS Lambda?
16. What happens if an AWS Lambda function runs asynchronously and one execution fails?
17. Provide a real-world use case for AWS Elastic Beanstalk.
18. How would you manage large-scale backups in AWS?

Round 3
System Design & Coding

System Design

1. Walk me through the design and architecture of your last two projects, particularly focusing on AWS Glue jobs.
2. How would you design a pipeline to join a 50 GB dataset with a 1.5 TB dataset and load the results into Redshift or Snowflake?
3. If the fact table arrives before the dimension table, how would you handle the situation?
4. In data warehousing, do you load fact tables or dimension tables first? Why?
5. How would you automatically handle schema evolution (e.g., extra or missing columns) in a data pipeline?
6 .If data from multiple sources is conflicting, how would you resolve it?
7. For processing 1 TB of data in EMR, how would you decide the amount of memory to allocate?
8. How would you design a partitioning strategy to handle 10,000 streaming records per second?

Spark

9. Explain the use of COALESCE in Spark.
10. How does serialization work in Spark?

Python - Coding

11. Write a Python program to reverse the words in a sentence.
Example: "God is great" → "great is God" (without using built-in reverse functions).
12. Write a Python program to duplicate each character in a string.
13. Write a Python program using list comprehension:
Input: ['a', 'b', 'c', 'd']
Output: ['a', 'bb', 'ccc', 'dddd']

SQL - Scenario & AWS

15. How would you optimise join performance when working with very large datasets?
16. What happens to external tables in Snowflake if the underlying S3 files are deleted?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal