Deloitte.

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹22-25 LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS
4. ETL / Data Modeling
5. System Design & Optimization

Round 1
SQL Foundations & Querying

1. Write a SQL query to retrieve the top 3 highest salaries from an employee table.
2. Write a SQL query to find duplicate records in a table.
3. Write a SQL query to calculate the running total of sales for each month.
4. Explain the difference between INNER JOIN, LEFT JOIN, and FULL OUTER JOIN with examples.
5. Explain the difference between DELETE, TRUNCATE, and DROP in SQL.
6. How do you handle NULL values in SQL queries?
7. What are window functions in SQL? Provide real-time use cases.
8. What is a CTE (Common Table Expression) and how is it different from a subquery?
9. How would you optimize a SQL query that is taking too long to execute?

Ankita Gulati Shubh Goyal

10. Explain the difference between a clustered and non-clustered index.
11. You have two large tables (each with 800M+ rows). How would you optimize a join query between them?
12. How do you detect and remove duplicate records in a production table without impacting performance?
13. Given a sales table, write a query to find the second highest sale per region.
14. Explain how you would debug a slow query in production (step-by-step approach).

Round 2
Python & Data Processing

1. How would you use Python for data cleaning and transformation?
2. Write a Python script to connect to a database and fetch data using SQL queries.
3. Explain the difference between Pandas and PySpark for data manipulation.
4. How would you handle exceptions in a Python-based ETL pipeline?
5. What Python libraries have you used for data processing (e.g., Pandas, NumPy, PySpark)?
6. How can you optimize the performance of a Python function used in large-scale data processing?
7. Explain the concept of lazy evaluation in PySpark and why it is useful.
8. Write a Python snippet to filter only string values from a mixed list of numbers and strings.

9. Real-world: How would you use PySpark to implement Slowly Changing Dimensions (SCD Type 2)?

10. How do you handle memory issues when processing large datasets in Python?

11.How do you implement parallel processing in Python for ETL jobs?

12. Explain how you would implement a retry mechanism in an ETL pipeline.

13 How do you handle schema evolution in PySpark when processing data from different sources?

Round 3
Data Engineering Concepts & Cloud

1. Describe the architecture of a cloud-based data warehouse like Snowflake or BigQuery.
2. What is the difference between OLAP and OLTP databases? Provide examples.
3. What is ETL? Explain its phases and the tools you have worked with.
4. How do you ensure data quality during ETL processes?
5. What is the role of Apache Kafka in data engineering?
6. Explain the difference between AWS Glue and AWS Lambda.
7. How would you design a real-time streaming pipeline using Kafka and Spark?
8. What are partitions and bucketing in Hive/Spark, and how do they impact performance?

10. How do you design a scalable data lake architecture on AWS/Azure/GCP?
11.How would you design an ETL pipeline to ingest terabytes of daily log data into a data warehouse?
12. How do you ensure fault tolerance in an ETL/ELT pipeline?
13 What are best practices for partitioning data in S3/BigQuery/Redshift?
14. How do you handle late-arriving data in a streaming pipeline?

Thank You

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

Ankita Gulati

Shubh Goyal