Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune/Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹27-30LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark
3. ETL Development
4. AWS
5. System Architecture

Ankita Gulati

Shubh Goyal

Round 1
Project & Data Eng. Fundamentals

Project & Responsibilites

1. Can you describe a recent project where you built a data pipeline?
2. In your latest project, what were your roles and responsibilities?
3. How was the data structured in your project and what type of processing did you perform?
4. What are the different data sources you worked with in your pipelines?
5. What tech stack have you worked with (AWS services, SQL, Python, etc.)?
6. Can you provide a high-level architecture of Airflow that you used in your projects?
7. What was the packet size or stream rate you were handling in your pipelines?

Ankita Gulati

Shubh Goyal

Data Concepts

8. Can you explain the difference between structured, semi-structured, and unstructured data with examples?
9. What is the difference between batch processing and real-time processing?
10. What are the steps to follow when creating a real-time data pipeline?
11. In a data pipeline, which comes first – Data Lake or Data Warehouse?
12. Can you explain the differences between Data Lake, Data Warehouse, and Delta Lake?

OLTP & OLAP

13. What is OLAP? Is it read-intensive or write-intensive?
14. What is OLTP? Is it read-intensive or write-intensive?
15. Can you name some services or databases that come under OLTP category?

Storage & Processing Formats

16. What is the Parquet format? Can you explain its purpose and advantages?
17. What type of storage format is used by OLTP systems?
18. Is a Data Warehouse based on OLAP or OLTP model?
19. What type of database is BigQuery?

Ankita Gulati

Shubh Goyal

Round 2
SQL & Query Optimization

SQL Scenarios & Challenges

1. Given a table of new user signups by date (YYYYMMDD), write a query to calculate the change in signups month-over-month. Skip the first month as it has no previous data.
2. Write a query to find employees whose salary is higher than their manager's. Display employee name, salary, manager name (use "No Manager" if none), order by salary difference (desc), and add a column Promotion Opportunity with Yes/No.
3. Write a query to find the second-highest salary in a table.
4. Given a table -
(Name, Salary → A-10, B-20, C-20, D-40).
Explain the difference in outputs of ROW_NUMBER, RANK, and DENSE_RANK.

Ankita Gulati Shubh Goyal

Joins & Query Types

5. Which join returns all values from the right table, plus matching values from the left table (or NULL if no match)?
6. Which of the following is not a type of join in SQL: Query, Copy, Update, Load?

SQL Testing & Tools

7. How do you test SQL queries? Which tools or platforms have you used (e.g., BigQuery, Redshift)?
8. Have you ever used the EXPLAIN clause to analyze query execution plans?

Query Optimization & Redshift Performance

9. Can you explain Workload Management (WLM) in Redshift?
10. What are some best practices for query performance optimization in Redshift?
11. Have you worked with VACUUM in Redshift? What is its purpose?

Ankita Gulati Shubh Goyal

Redshift Architecture & Nodes

12. Can you explain the architecture of Redshift?
13. What types of nodes are available in Redshift?
14. What type of storage is used by RA3 nodes?
15. What are the different distribution styles in Redshift? (You mentioned event distribution — what are the others?)
16. How is data distributed across tables in Redshift?

Ankita Gulati

Shubh Goyal

Round 3
Big Data, Cloud & Advanced Tools

Partitioning & Skewness

1. What is partitioning in data processing? Why is it important?
2. Is partitioning only relevant to real-time pipelines, or can it also be applied in batch pipelines?
3. Why does partitioning improve performance?
4. How do you handle queries on a partition column in Redshift?
5. When defining a sort key or partition column, do you do it at table creation or later?
6. If a table was created without a partition column but later needs one, how would you implement it?
7. When partitioning on a column like country, one partition (e.g., India) may have much larger data than others. How do you handle data skewness in this case?

Ankita Gulati Shubh Goyal

8. What are techniques like repartitioning, coalesce, and salting? When would you use each?
9. What are different options/functions to partition data in Spark write operations?

Spark Fundamentals

10. How do you define a schema for data in Spark?
11. Can you explain the difference between narrow and wide transformations in Spark?
12. Give examples of narrow and wide transformations.
13. Between repartition and coalesce, when should you use each?
14. What techniques can be applied in Spark if you notice data skewness?
15. Write PySpark code to:
• Create an Employee DataFrame (EmployeeId, EmployeeName, DepartmentId)
• Create a Department DataFrame (DepartmentId, DepartmentName)

- Join on DepartmentId
- Extract EmployeeId, Name, DepartmentName
- Filter where DepartmentId = 100
- Write the results.

Airflow

16. What are different types of operators in Airflow?
17. Which operator is used for branching logic in Airflow DAGs?
18. Can you explain the TriggerDagRunOperator?
19. How can you pass data between tasks in an Airflow DAG?

AWS Services

20. What are the different AWS services you've used for data processing?
21. Explain the concept of Lambda Layers and how you've used them in Python code.
22. Your raw data is stored in S3. How would you use AWS Glue to transform it?
23. If you want to process data from S3 using PySpark inside Glue, what are the steps?

Ankita Gulati Shubh Goyal

24. When would you use Redshift Spectrum instead of Athena?

25. What are the different query options provided by Amazon to read data from S3?

26. What is Intelligent-Tiering in S3?

27. What are the different storage classes in S3?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal