Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹22-23 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. AWS, Azure
4. Airflow

Ankita Gulati

Shubh Goyal

Round 1
Data Engineering Concepts

1. Have you ever worked with Delta Live Tables? What were your learnings and challenges?

2. What was the actual usage of Databricks in your project? Could you explain the project details, how you used Databricks, and the business impact?

3. You are given a table with 700 rows containing user signups by date (YYYYMMDD format). Write a SQL query to calculate the month-over-month change in signups. Skip the first month since it has no preceding month.

4. Write a function letterCount(str) that returns the first word with the greatest number of repeated letters. If no word has repeated letters, return -1. Example: "today is the greatest day ever" should return "greatest".

Ankita Gulati   Shubh Goyal

5. Write a SQL query to identify employees who earn a higher salary than their manager. Display employee name, salary, and manager name. If the employee has no manager, display "No Manager" and treat manager’s salary as 0. Add a column "Promotion Opportunity" with Yes/No and order results by salary difference (descending).

6. What type of database is BigQuery?

7. How do you usually test SQL queries?

8. In your latest project, what were your roles and responsibilities? What type of data and processing were you handling?

9. What is OLAP? Is it read-intensive or write-intensive?

10. What is OLTP? Is it designed for read-intensive or write-intensive workloads?

11. Name some services or databases that fall under the OLTP category.

12. Which storage format is generally used by OLTP databases?

13. Is a Data Warehouse an OLAP or OLTP model? Why?

14. In a data processing pipeline, what usually comes first – Data Lake or Data Warehouse?

15. Explain the difference between Data Warehouse, Data Lake, and Delta Lake.

16. In Airflow, which operator would you use for implementing branching logic?

17. What is the purpose of the TriggerDagRunOperator in Airflow?

18. How can data be passed between different tasks in an Airflow DAG?

19. Can you name some commonly used Airflow operators?

20. In PySpark, apart from repartition and coalesce, what are the other ways to partition data while writing it?

21. How do you define a schema in Spark when reading raw data?

22. Explain the difference between Narrow vs Wide transformations in Spark. How do they work internally?

23. Can you list down transformations that fall under Narrow and those that fall under Wide?

24. Explain when to use REPARTITION vs COALESCE in PySpark.

25. What do you understand by salting and data skewness in Spark?

26. If you observe data skewness in your project, what techniques would you apply to fix it?

27. What are the different modes to write data in PySpark?

28. You are given two DataFrames:
---> Employee (EmployeeId, EmployeeName, DepartmentId)
---> Department (DepartmentId, DepartmentName)

Join them on DepartmentId and extract EmployeeId, Name, and DepartmentName. Filter for DepartmentId = 100 and write the result in PySpark.

29. Write a SQL query to find the second highest salary from an Employee table. Also compare results using ROW_NUMBER(), RANK(), and DENSE_RANK().

Round 2
Cloud & Architecture

1. How would you use a Lambda Layer in your Python code? Explain its purpose.

2. What are Lambda Layers in AWS, and why are they useful?

3. What AWS services are available to process data?

4. You have raw data stored in S3, and you need to transform it using AWS Glue. What steps would you follow?

5. Explain the general flow of processing data from S3 using PySpark in Glue Data Lake environment.

6. When would you use Redshift Spectrum?

7. What are the use cases where you would prefer Redshift Spectrum over Athena?

8. If data is stored in S3, what are the different query options provided by AWS to read it?

9. What is S3 Intelligent-Tiering, and when should it be used?

10. What are the different storage classes available in S3?

11. Explain the purpose of the VACUUM query in Redshift.

12. What is Workload Management (WLM) in Redshift?

13. Have you ever used the EXPLAIN clause in Redshift? What insights can it provide?

14. What type of storage is used by the RA3 node in Redshift?

15. What are some best practices for optimizing query performance in Redshift?

16. What types of nodes are available in Redshift?

17. Explain the architecture of Redshift.

18. What are distribution styles in Redshift? Can you explain EVEN, KEY, and ALL distribution?

19. How is data distributed in Redshift internally?

20. Could you explain the architecture of an ETL pipeline using AWS and Airflow?

Thank You

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

Ankita Gulati

Shubh Goyal