Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹20-24 LPA
- **Total Rounds:** 2
- **Top Required Skills:**

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

Ankita Gulati

Shubh Goyal

Round 1
Data Ops

1. What is the source of your data, and how do you retrieve it?
2. When data lands in S3, what is the frequency of ingestion?
3. Once the data is received in the S3 folder, what are the next steps you perform?
4. What is your role in the project? Can you explain your day-to-day activities?
5. What types of errors do you commonly face in pipelines? Apart from code errors, what other failures have you seen, and how do you handle them?
6. How do you manage schema evolution in your pipelines?
7. Can you explain what custom data types are and where you have used them?
8. If you receive a dataset with nested fields (e.g., Customer → Address → SalesInfo), how would you approach flattening it theoretically (without writing Spark code)?

Ankita Gulati

Shubh Goyal

9. Why would you use explode() in combination with DataFrames in Spark?

10. How do you optimize memory usage in Spark applications? What does it mean when memory gets “stuck” during processing?

11. Can garbage collection be triggered manually in Spark?

12. Once the data is ingested into Kinesis, how do you process it further?

13. How do you connect and read data from a source into Kinesis streams?

14. What are DISTKEY and SORTKEY in Redshift, and how do they affect performance?

15. After transforming data, do you store the final processed layer in Redshift? Why or why not?

16. Suppose you have a 5-year dataset in Redshift that needs correction (multiplying some fields by 2). How would you delete and reload it efficiently?

17. If Amazon sales data needs an additional 0.5% charge applied to every transaction, how would you implement this update at scale?

18. Can you run an UPDATE command directly in S3? Why or why not?
19. If the application can no longer regenerate past data, and only your side has it stored, how would you manage corrections and updates?

20. When you receive large files daily (with overlapping one-year ranges causing duplication), how would you:
-->Remove duplicates without using DELETE or dropDuplicates()?
-->Design a folder structure in the reporting layer to manage clean data?

21. What is ETL vs ELT? In what situations would you prefer one over the other?
22. What is BigQuery, and which company provides it?
23. Can you explain the difference between Data Layout, Data Mart, and Data Lake?

24. Write a SQL/PySpark solution to classify nodes in a tree as root, inner, or leaf based on id and parent_id.

25. Which AWS services have you used, and for what purposes?

26. Have you built end-to-end pipelines from scratch? Can you walk me through one?

27. Given a table with teams [A, B, C, D].
How would you generate all unique matchups (pairs like A-B, A-C, B-D, etc.)?

28. You have two tables:
-->Table A: [1, 0, 1, NULL]
--> Table B: [1, 1, 0, 0, NULL, NULL]
How many rows would be returned by each type of join (INNER, LEFT, RIGHT, FULL, CROSS)?

29. You have a vendor visits table with vendor_id, office_id, date_of_visit. How would you identify consecutive visits (streaks) for each vendor-office pair, outputting start_date, end_date, and streak length?

Round 2
SQL & Modeling

1. When designing a system, which tools (e.g., Draw.io, Miro) do you use for visualization, and what is your approach?
2. Design a system that collects, processes, and stores logs from millions of devices, with support for querying and real-time alerting. Illustrate your design.
3. Looking at your pipeline flow (Source → Kinesis → Stream Processing → Transformation → Storage), how would you handle unexpected changes or errors in real time?
4. In system design, why is it important to choose a single storage solution (e.g., S3, Redshift, PostgreSQL) instead of storing data everywhere? Which one would you choose for your use case, and why?
5. If you had to redesign your architecture, what cleaner approach would you take to make it more efficient and purposeful?
6. How does Apache Flink store data internally?

Ankita Gulati

Shubh Goyal

7. If you had to process AWS datasets both historically and in real time, how would you design the processing and reporting system?

8. You mentioned intermediate storage in S3 before persisting data into memory. If the dataset size is 1 PB, would you store it all in memory? Why or why not?

9. You have a daily incoming dataset and need to implement SCD Type 2. How would you design the ETL pipeline for it?

10. For an employee dataset (with employee ID and multiple attributes), how would you implement SCD Type 2 to track changes over time? Specifically:

--> How will you handle the initial load?
--> How will you detect changes in incremental loads?
--> What approach ensures accurate change capture?

Thank You

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

Ankita Gulati

Shubh Goyal