Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹25+ LPA
- **Total Rounds:** 2
- **Top Required Skills:**

1. SQL
2. PySpark / Python / Databricks
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming
6. System Design

Ankita Gulati

Shubh Goyal

Round 1
Technical Discussion

1. How would you design a strategy to ingest semi-structured data into a Redshift data warehouse?
2. In Spark, what are the trade-offs between using broadcast joins vs. partitioned joins, and when would you prefer one over the other?
3. How do you enforce data quality checks in ETL pipelines? Can you give examples of frameworks/tools you have used?
4. Write a SQL query to calculate the 90th percentile transaction amount per customer per month from a payments table.
5. In Python, how would you implement a function to detect anomalies in a time-series dataset without using third-party libraries?
6. What are concurrency challenges in Redshift when multiple ETL jobs and reporting queries run simultaneously? How would you handle them?

Ankita Gulati

Shubh Goyal

8. How do you efficiently process and flatten large-scale XML data in AWS Glue?
9. How would you design an event-driven data pipeline in AWS to process transactions in near real time? Which services would you choose ?
10. What approach would you take to enforce data governance and lineage tracking across multiple AWS accounts?
11. How do you set up observability and monitoring for a large-scale Spark job running on EMR ?
12. How would you implement data versioning in a data lake so analysts can query both current and historical versions?
13. Cognizant often deals with cost-sensitive clients. What are some AWS cost optimization strategies you apply when building pipelines?
14. How would you design a cross-account data sharing architecture in AWS while ensuring security and compliance?

Ankita Gulati Shubh Goyal

Round 2
HR Discussion

1. Cognizant projects often require working directly with clients. How do you handle situations where business stakeholders request something technically infeasible?
2. Give an example where you implemented a cost optimization strategy in a data pipeline that significantly saved expenses.
3. Tell me about a time when you had to work with distributed teams (onsite-offshore model). How did you ensure collaboration and delivery?
4. How do you prioritize multiple data requests coming from different business teams with conflicting deadlines?
5. Cognizant values innovation. Can you share an example of how you introduced a new tool, framework, or process that improved your team's productivity or delivery quality?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal