LTIMindtree

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 5+ years
- **Location:** Pan India
- **Work mode:** Hybrid
- **Compensation:** ₹25-30LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark
3. ETL Development
4. AWS
5. Data Modeling
6. Data Warehousing
7. System Design / Architecture

Ankita Gulati

Shubh Goyal

Round 1
Data Eng. & Pipeline Design

Project & Pipeline Design

1. Can you share about your last two projects, especially your work with AWS Glue jobs and overall architecture?
2. In the current project, what is your role—development or operations?
3. You have many pipelines; how would you handle frequent component changes?
4. From a performance perspective, which is more efficient: UNION ALL with DISTINCT, UNION ALL with window functions or plain UNION ?
5. Migration task: multiple datasets/models integrated into warehouse — how to ensure consistency in unified platform?
6. What is cherry-pick conflict resolution in Git?

Processing & Schema Evolution

7. When working with ~15 GB of data, how much time does processing take?
8. If a job processes around 15 GB of data, including extraction, transformation, and loading into Redshift, how much time would the complete process typically take?
9. If you have two large datasets (50 GB and 1.5 TB in CSV, on-prem), how would you join, transform, and load into Redshift (or Snowflake) for optimal scalability?
10. If you receive 10,000 records/sec and partition by timestamp (per second), how would you handle the large number of partitions?
11. Why only use AWS Glue? Should we consider alternatives for performance and cost?
12. If we avoid EMR and use a standalone Spark job for heavy joins, what optimizations would you apply?
13. A customer has 1 TB of data — how do you decide EMR cluster memory allocation?

Ankita Gulati

Shubh Goyal

14. How do you decide executors, cores, memory, tasks, and nodes when tuning Spark on EMR?
15. If daily CSV (10 cols) suddenly has 12 cols, how would you handle schema evolution?
16. How to design a pipeline to automatically handle addition/removal of columns without code change?

Data Concepts & SQL

17. If data is ingested into Snowflake external tables pointing to S3, what happens if the S3 file is deleted?
18. Explain data lineage.
19. What is data locality? In data locality, does the data move to code or code to data?
20. Table A = 1,1,1, NULL & TABLE B = 1,1, NULL
→ What are the results for RIGHT, LEFT, INNER, FULL OUTER joins?

Ankita Gulati

Shubh Goyal

Round 2
Database & Orchestration

Orchestration, Glue & Spark

1. Best practices for optimizing performance and cost in AWS Step Functions.
2. How to customize AWS Glue Crawler for specific formats and schema evolution.
3. How does Glue Crawler handle inconsistent datatypes within a column?
4. Explain data lineage in Spark and its usage.
5. In Spark, what are the real-time challenges you've faced?
6. How do you handle long-running Spark jobs?
7. Suppose sources are in S3 with multiple file formats, ETL uses Glue/Airflow/RPSS, and warehouse is Snowflake → design end-to-end pipeline.
8. If one source is an API, how would you fetch and integrate it?
9. If multiple APIs/endpoints need to be ingested, how would you manage them efficiently?

Ankita Gulati

Shubh Goyal

Databases & SQL Concepts

10. What do you mean by clustering/cluster?
11. What are the types of indexing.
12. Difference between clustered vs non-clustered index.
13. Do you know about page offsets?

Airflow & Python Data Consistency

14. What is XCom in Airflow?
15. What is MT operator?
16. Difference between shallow vs deep copy in Python.
17. If multiple source models need integration, how would you ensure data consistency?

Thank You

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

Ankita Gulati

Shubh Goyal