BARCLAYS

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹23-27 LPA
- **Total Rounds:** 3
- **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
Core ETL & Data Engineering

1. What are the different types of jobs available in AWS Glue, and in what scenarios would you use each?
2. What is the difference between a Spark ETL job and a Python Shell job in Glue?
3. When would you prefer to use Ray jobs in AWS Glue, and what are their advantages?
4. How would you design an ETL pipeline to support incremental data loads?
5. What key attributes would you consider when designing incremental loads in a data pipeline?
6. What are Glue bookmarks, and how do they simplify incremental processing?
7. Why can merging only on timestamps in ETL be risky? Provide an example.
8. How would you design a safe merge condition for incremental data ingestion?
9. What strategies would you use to handle late-arriving data in AWS Glue pipelines?

Ankita Gulati Shubh Goyal

10. Compare full load vs incremental load in ETL. What are the trade-offs of each?
11. Explain the difference between batch processing and stream processing with examples.
12. How would you ensure idempotency in ETL pipelines to avoid duplicate records?
13. What approaches do you use to handle slowly changing dimensions (SCDs) in data warehousing?
14. What are the most common causes of data skew in Spark, and how would you handle them in Glue?

Ankita Gulati

Shubh Goyal

Round 2
Coding & Monitoring

1. How can you run a Glue job manually, and how is it different from programmatic execution?
2. How do you run a Glue job using the AWS CLI? Provide an example command.
3. How do you trigger a Glue job programmatically from Python (Boto3)?
4. How can you automate Glue job execution using AWS Lambda or EventBridge?
5. What's the difference between deploying a Glue job vs. executing it?
6. How do orchestrators like Step Functions or Airflow integrate with Glue for pipeline execution?
7. If you deploy a pipeline, does it automatically run the Glue job? Explain with a real-world case.
8. How do you use CloudWatch Logs Insights to troubleshoot Glue jobs with frequent errors?
9. Write a CloudWatch Logs Insights query to retrieve the top error-heavy log streams.

Ankita Gulati Shubh Goyal

10. How would you configure alarms in CloudWatch to proactively notify you of Glue job failures?
11. How do you handle schema evolution in Glue jobs without breaking downstream pipelines?
12. What steps would you take if a Glue job runs slower than expected in production?
13. Python: Write a function to parse nested JSON data from an S3 bucket, flatten it, and store it in Parquet format.
14. SQL: Given a table transactions(user_id, txn_time, amount). Write a query to find the top 3 spenders per day.
15. SQL: Write a query to identify customers who made purchases in three consecutive months.

Ankita Gulati

Shubh Goyal

Round 3
System Design

1. How would you design a data lake architecture in AWS for Barclays that ingests data from multiple banking systems ?
2. How would you design a Glue-based pipeline to process real-time fraud detection events while also supporting daily batch aggregates?
3. What trade-offs would you consider between Glue vs EMR vs Spark on EKS for Spark workloads in terms of cost, operational overhead, and control?
4. How would you design a system to reprocess failed Glue jobs without duplicating records in downstream tables?
5. What steps would you take to ensure data security, encryption, and GDPR compliance for sensitive financial datasets in AWS?
6. How would you scale an ETL pipeline to handle 10x more data volume without significantly increasing costs?

Ankita Gulati

Shubh Goyal

7. Can you describe a complex ETL pipeline you built in your current role and how you optimized it?

8. Tell me about a time when you had to debug a failing pipeline under strict deadlines. What approach did you take?

9. How do you collaborate with data scientists and business teams to deliver end-to-end solutions?

10. Barclays handles highly sensitive financial data. How do you ensure data governance, lineage, and auditability in your projects?

Round 4
HR Discussion

1. Can you walk me through your professional journey and highlight the key data engineering projects you've worked on?
2. Why are you interested in joining Barclays, and how does this role align with your career goals?
3. How do you manage conflicts or disagreements within your team, especially on technical decisions?
4. Describe a situation where you took ownership of a project and ensured its success despite challenges.
5. How do you prioritize multiple urgent requests from different stakeholders while ensuring data quality?
6. What unique strengths or value do you believe you will bring to the Barclays data engineering team?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal