CHUBB®

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 4+ years
- **Location:** Hyderabad
- **Work mode:** Hybrid
- **Compensation:** ₹20+ 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

Round 1
Technical Discussion

1. Walk me through a recent data pipeline you built. How did you ensure data quality and compliance with regulatory requirements?
2. What challenges have you faced while handling sensitive data (like PII), and how did you secure it?
3. Write a SQL query to fetch the second highest claim amount from an insurance claims table.
4. Explain the difference between ROW_NUMBER(), RANK(), and DENSE_RANK() with use cases.
5. Given a table claims(policy_id, claim_id, claim_date, claim_amount), write a query to calculate the average monthly claim amount per policyholder.
6. How would you detect and remove duplicate claim records while keeping the most recent entry?
7. Write a Python function to return the first non-repeating character in a string (O(n) solution).

8. How would you parse, validate, and flatten nested JSON claim data before loading it into Redshift?
9. What are Python generators? Give an insurance-related example where they help (e.g., processing a large claims feed).
10. Explain the difference between RDD, DataFrame, and Dataset. Which would you use for processing policy data?
11. How do you optimize Spark jobs that suffer from small files issue when writing to S3 in Parquet format?
12. What are wide vs. narrow transformations in Spark? Why does it matter for cost and performance?
13. How would you design an S3 → Glue → Redshift ETL pipeline for claims ingestion?
14. What measures do you take to secure S3 buckets containing customer data (IAM, encryption, VPC endpoints)?
15. Write a PySpark query to find the top 3 fraud-prone regions by claim amount using window functions.

Round 2
Advanced Technical Discussion

1. Design a near real-time pipeline for ingesting claim events from multiple regions. Which AWS services do you use (Kinesis, Glue, Lambda, Redshift)?
2. How would you partition claims data in Redshift or S3 to support fast queries by policy_id and claim_date?
3. Your Spark ETL for claims aggregation has heavy shuffle. How do you diagnose and reduce shuffle cost (partitioning, bucketing, broadcast joins)?
4. How would you handle data skew if one policy_id generates 80% of total claims?
5. How would you build a streaming fraud detection pipeline on AWS with exactly-once semantics?
6. How do you handle late-arriving claims data in streaming pipelines?
7. Compare Glue, EMR, and Snowflake on AWS for Chubb’s use case (compliance-heavy, insurance analytics).

8. What steps would you take to implement data lineage and audit tracking across S3, Glue Catalog, and Redshift?
9. How do you enforce role-based access control for sensitive policyholder data across AWS services?
10. How would you monitor and alert on a pipeline that processes daily policy and claims data? What KPIs are critical?
11. How would you reprocess one year of historical claims data without blocking production pipelines?
12. How would you design a PII masking strategy for policyholder addresses in Redshift?
13. Compare Athena vs Redshift for ad-hoc actuarial queries on claims data.
14. How would you test and validate a claims ingestion ETL pipeline before production release?

Round 3
HR Discussion

1. Why do you want to join Chubb and work in the insurance domain?
2. Give an example of how you handled regulatory or compliance requirements in your previous role.
3. Describe a time when you had to explain a complex technical pipeline to business or compliance stakeholders.
4. How do you balance fast delivery with data security and quality in projects?
5. Share a situation where you worked under tight deadlines with sensitive data at risk.

Thank You

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

Ankita Gulati

Shubh Goyal