pwc

# Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

# Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹25+ 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
# Technical Discussion

1. Briefly describe your most recent data engineering project. What was the business problem, what pipeline did you build, and what was your role?
2. Tell me about a production issue you faced (data loss, late data, performance). How did you diagnose and resolve it?
3. Write a SQL query to get the second-highest salary from an employees table (explain at least two approaches).
4. Given orders(order_id, customer_id, order_date, amount), write an SQL query to compute total customer revenue for the last 30 days.
5. Explain ROW_NUMBER(), RANK() and DENSE_RANK() — give one real use case for each.
6. How would you detect and remove duplicate rows in a table while keeping one canonical record?
7. Write a Python function (O(n)) to return the first non-repeating character in a string.

Ankita Gulati

Shubh Goyal

8. What are Python generators and where would you use them in an ETL workflow? Give a short code sketch.

9. How would you parse and flatten a nested JSON payload from a partner API before landing it to S3?

10. Explain the difference between RDD, DataFrame and Dataset. Which do you use for production ETL and why?

11. What are narrow vs wide transformations? Give examples and explain why wide transformations cost more.

12. How do you avoid the small files problem when writing Parquet to S3?

13. Sketch a simple batch pipeline:
S3 → Glue (or EMR) → Redshift/Athena.
What IAM permissions and key configuration steps are required?

14. How do you secure S3 buckets containing PII? (principles: encryption, IAM, bucket policy, logging, VPC endpoints)

Ankita Gulati

Shubh Goyal

# Round 2
# Advanced Technical Discussion

1. Design: build an end-to-end pipeline for ingesting clickstream data from multiple regions in near-real time. Which AWS services would you choose and why?
2. How will you design the data model and partitioning strategy so both ad-hoc analytics and daily aggregates perform well?
3. Your Spark job has huge shuffle and runs slowly. Walk through how you would profile it and reduce shuffle cost.
4. How would you detect and fix data skew in joins? → show pseudo-code or config changes.
5. Explain cache() vs persist() and when to use which storage level for iterative machine-learning style jobs.
6. Design a streaming fraud-detection pipeline with exactly-once semantics on AWS. How do you ensure idempotency and deduplication?
7. How would you handle late arriving data and corrections (retractions) in a streaming pipeline? Describe replay/backfill strategies.

Ankita Gulati Shubh Goyal

8. Can you compare AWS Glue, Amazon EMR, and Spark on Amazon EKS for running Spark workloads? Explain the trade-offs in terms of cost, operational overhead, level of control, and whether the service is serverless or managed.

9. How would you implement data lineage, cataloging and access governance across S3 + Glue Data Catalog + Redshift + Athena (which metadata do you track and which AWS tools do you use)?

10. How would you implement CI/CD for data pipelines

11. Which logs and metrics would you monitor for pipeline health? Suggest alert thresholds and SLOs.

12. PwC works with regulated customers — how would you encrypt, mask and audit PII during ETL and in the data lake?

13. You need to backfill a year of data into Redshift without blocking production queries. Design a strategy to ingest, validate and swap tables safely.

14. A downstream BI dashboard is slow; explain how you would trace the bottleneck end-to-end.

Ankita Gulati

Shubh Goyal

15 PySpark: top-3 products by revenue per region using window functions.

16. SQL: find customers who made purchases in every month of the last 12 months.

17. Python DSA: longest substring without repeating characters (sliding window).

18. Given two sample tables with NULLs and duplicates, predict counts for INNER, LEFT, RIGHT and FULL joins.

Ankita Gulati

Shubh Goyal

# Round 3
# HR Discussion

1. Why PwC? Why consulting instead of an in-house product or startup?
2. Describe a time you led a technical decision under ambiguity. What tradeoffs did you make and how did you communicate them to stakeholders?
3. Give an example where you simplified a complex technical concept for a non-technical client.
4. How do you handle conflict inside a cross-functional team? Provide a concrete example.
5. Talk about a time you had to meet a tight deadline while maintaining compliance or quality. How did you prioritize?

Ankita Gulati

Shubh Goyal

# *Thank You*

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


Ankita Gulati

Shubh Goyal