Moonfare

# Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

# Job Details

- **Position:** Senior Data Engineer
- **Experience:** 7+ years
- **Location:** Pan India
- **Work mode:** Remote
- **Compensation:** ₹30–40 LPA
- **Total Rounds:** 5
- **Top Required Skills:**

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

Ankita Gulati

Shubh Goyal

# Round 1
# Introductory / Cultural Fit

1. Can you briefly introduce yourself and summarize your journey as a Data Engineer so far?
2. What do you know about Moonfare as a company and its data-driven business model?
3. What services have you used in AWS, and what were the primary use cases?
4. How do you monitor and track the health of data pipelines in production?
5. Can you describe some of the biggest challenges you have faced in your projects and how you solved them?
6. How do you prioritize tasks in an Agile environment with competing deadlines?
7. How do you ensure continuous learning and keep yourself updated with new technologies?

Ankita Gulati

Shubh Goyal

# Round 2
# Managerial & Project Deep Dive

1. Walk me through some of the most complex projects you have worked on. What was your role, and which technologies did you use?
2. Can you explain the end-to-end architecture of your most recent data platform, including ingestion, storage, transformation, and consumption layers?
3. What are some best practices you follow when building scalable and fault-tolerant pipelines?
4. Moonfare uses a specific tech stack for data processing. Based on your experience, how would you adapt to a new stack quickly?
5. How do you collaborate with cross-functional teams (data scientists, analysts, business stakeholders) to ensure data requirements are met?
6. Describe a scenario where you had to make a trade-off between performance, cost, and scalability. What decision did you take, and why?

Ankita Gulati Shubh Goyal

# Round 3
# Technical

1. What is the AWS S3 storage cleanup command? Explain how you would automate S3 bucket cleanup for unused files.
2. How would you design a real-time Change Data Capture (CDC) pipeline during live database migration?
3. Explain how AWS Lambda, Kinesis, and DynamoDB can be integrated for building a real-time streaming pipeline.
4. You are given a healthcare dataset with missing values, duplicates, and inconsistent schema. Write a Spark/Scala program to clean the data and produce a standardized output.
5. How do you optimize Spark jobs to handle OOM (Out Of Memory) errors?

Ankita Gulati

Shubh Goyal

6. Implement a context manager class called SequenceGenerator that takes two integers, start and step. It should expose a method generate() that yields an infinite sequence of numbers starting from start, adding step in each iteration.

Ankita Gulati

Shubh Goyal

# Round 4
# Theoretical + Advanced Coding

1. Explain the difference between batch processing and real-time streaming. When would you use one over the other?
2. What is the CAP theorem, and how does it apply to distributed data systems?
3. Explain eventual consistency and provide an example from a real-world system.
4. How do you design a fault-tolerant pipeline that processes billions of events daily?
5. What are Slowly Changing Dimensions (SCDs)? Explain different types (Type 1, Type 2, Type 3) with examples.
6. When would you choose a Star schema over a Snowflake schema in a data warehouse?
7. Explain partitioning vs bucketing in data lake design.

Ankita Gulati

Shubh Goyal

8. Write an SQL query to return the top 3 customers with the highest total purchases in each country.
9. Write an SQL query to calculate a running total (cumulative sum) of sales for each product by date.
10. Given two tables Orders and Returns, write a query to find the percentage of returned orders per region.
11. Write a Python function to process a 10GB CSV file efficiently without running out of memory.
12. How would you design a Python-based ETL pipeline that fetches data from an API and stores it into AWS S3 in Parquet format?
13. Implement a Python function to detect anomalies in a time series dataset.
14. How do you manage CI/CD for data pipelines in AWS (or other cloud providers)?
15. Explain the use of Airflow vs Step Functions vs Prefect for orchestration. Which one would you choose for Moonfare, and why?
16. How do you monitor pipeline SLAs and detect failures proactively?

Ankita Gulati Shubh Goyal

# Round 5
# HR & Behavioral

1. Why do you want to join Moonfare?
2. What are your long-term career goals as a Senior Data Engineer?
3. Describe your best and worst experiences working in previous organizations.
4. How do you handle conflict in a team when working on tight deadlines?
5. Salary expectations, relocation preferences, and notice period.

Ankita Gulati

Shubh Goyal

# *Thank You*

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


Ankita Gulati

Shubh Goyal