CONFLUENT

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5-6 years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹50+ LPA
- **Total Rounds:** 4
- **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
Recruiter Screen & Introductory

1. Can you walk me through your current role, projects, and the data stack you work with?
2. Why are you interested in Confluent, and what excites you about event streaming?
3. What experience do you have with Kafka or Confluent Cloud?
4. SQL: Write a query to fetch the top 5 customers by order amount.
5. Python: How would you remove duplicates from a list while preserving order?
6. What are the key differences between list, set, and dictionary in Python?
7. How would you explain event streaming to a non-technical stakeholder?

Round 2
SQL + Coding / DSA

SQL Problems:

1. Window Functions: Given a table (user_id, event_date, events), find users who performed more than 50 events in a single day.
   • Follow-up: Retrieve the top 3 busiest days per user.
2. Ranking: Find the 3rd highest transaction amount per customer, considering ties.
   • Hint: Use ROW_NUMBER(), RANK(), or DENSE_RANK().
3. Streaming-Style SQL: On a continuous stream of user activity, compute the rolling 1-hour count of logins per user.

Python / DSA Problems:

1. Implement an LRU Cache with O(1) time complexity for get() and put().
2. Given access logs (user_id, timestamp), write a function to check if a user accessed the system more than 3 times in any 5-minute window.

Ankita Gulati Shubh Goyal

3. Write a function to validate a partially filled Sudoku board (9×9).

Performance Optimization:

1. How would you optimize a SQL query with multiple joins and aggregations on billions of rows?
2. When would you prefer indexes, partitioning, or materialized views?

Ankita Gulati

Shubh Goyal

Round 3
System Design

Streaming Pipeline Design:

1. Design an end-to-end real-time pipeline for clickstream data → ingest (Kafka) → process (Spark/Flink) → store → dashboards with <5s latency.
   - How would you design retries if a consumer fails?
   - How do you ensure exactly-once processing?
   - How do you handle schema evolution in Avro/JSON?

Kafka Deep Dive:

1. How does offset management work in Kafka?
2. What's the difference between sync vs async commits?
3. What happens during a consumer group rebalance?
4. How do you decide partition count for a topic?
5. How do you prevent backpressure in consumers?

Ankita Gulati

Shubh Goyal

Big Data / Spark Concepts:

1. How do you handle OutOfMemory (OOM) issues in Spark executors?
2. Difference between map vs flatMap in Spark.
3. How do you handle data skewness? (e.g., salting, broadcast joins).
4. When would you use Parquet vs ORC vs Avro for storage?

Data Modeling:

1. Design a schema for storing event logs supporting both real-time querying and historical analysis.
2. Discuss trade-offs between Star schema vs Snowflake schema for analytics.

Round 4
Managerial

Project / Technical Deep Dive:

1. Walk through your most impactful pipeline project – from ingestion → processing → storage → consumption.
• What scaling challenges did you face, and how did you resolve them?

Behavioral Questions:

1. Tell me about a time you debugged a critical production issue under pressure.
2. How do you prioritize tasks when multiple stakeholders have urgent requests?
3. Give an example of solving a complex technical problem under tight deadlines (use STAR).

Team Fit & Leadership:

1. How do you ensure data quality across teams?
2. How do you mentor junior engineers in SQL/Python best practices?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal