Goldman Sachs

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 4+ years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹25+ LPA
- **Total Rounds:** 6
- **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
Advanced Querying

1. Write a query to find the median salary of employees in a table.

2. Identify and remove duplicate records, keeping only the most recent record based on a timestamp column.

3. Write a query to compute the 7-day moving average of daily transactions.

4. How would you optimize a slow query running on a large table with billions of rows?

Ankita Gulati Shubh Goyal

Round 2
Programming & Algorithms

1. Write a Python script to parse a large JSON file, filter records based on conditions, and write results to a database.

2. Implement a function to find the longest increasing subsequence in an array.

3. Write a program to simulate a producer-consumer model using multithreading.

4. How would you process a 10TB dataset in Python on a single machine? What constraints would you face?

Ankita Gulati　　Shubh Goyal

Round 3
Data Engineering Fundamentals

1. Design an ETL pipeline to process real-time stock market data.

2. How would you handle schema evolution in an ETL pipeline?

3. Describe how you would design a fault-tolerant distributed data processing system.

4. Compare batch vs stream processing for financial data.

Ankita Gulati

Shubh Goyal

Round 4
Big Data & Cloud Technologies

1. How does Spark’s lazy evaluation improve performance?

2. Explain how you would use Kafka to build a real-time streaming pipeline.

3. Describe a scenario where partitioning and bucketing would significantly improve query performance.

4. Compare AWS Glue vs Apache Airflow for orchestrating ETL pipelines.

Round 5
Data Modeling & Database Design

1. Design a database schema for tracking stock trades in real-time.

2. Explain when you would choose a star schema over a snowflake schema.

3. How would you design a database to handle historical data storage for compliance purposes?

4. What are the trade-offs between using a relational database vs a NoSQL database for financial applications?

Ankita Gulati

Shubh Goyal

Round 6
Behavioral & Scenario-Based

1. Tell me about a time you handled a data pipeline failure during a critical business operation.

2. Describe a challenging project where you optimized a complex ETL process.

3. How do you ensure collaboration with cross-functional teams when handling time-sensitive financial data?

4. Give an example of when you had to prioritize cost vs performance trade-offs in data engineering.

Thank You

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

Ankita Gulati

Shubh Goyal