Expedia

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer III
- **Experience:** 5+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹28-32 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
Data Pipeline & System Design

1. Walk me through your past projects and professional experience as a Data Engineer.

→ Follow-up: What were the major challenges you faced and how did you resolve them?

2. How would you design a real-time streaming data pipeline for Expedia?

→ Which technologies would you use for ingestion, processing, and storage?

3. Can you provide a deep dive into Apache Kafka concepts, specifically covering offset management, synchronous vs. asynchronous commits, partition assignment strategies, consumer groups, and how Kafka handles backpressure?

4. How do you use Docker to scale real-time data streaming applications?

Ankita Gulati Shubh Goyal

5. What is your approach to deployment using CI/CD pipelines? Which tools have you used (e.g., Jenkins, GitHub Actions, GitLab, ArgoCD)?

6. Explain the difference between batch and streaming pipelines. When would you use each?

7. How do you ensure fault tolerance and high availability in streaming applications?

8. What is idempotency in data pipelines and why is it important?

Ankita Gulati

Shubh Goyal

Round 2
Coding – Python & SQL

1. Write efficient code to calculate the power of a given number with minimum time complexity.

→ Expected solution: Recursion + Dynamic Programming = O(log n) complexity.

2. Given an expression in Infix, Postfix, or Prefix form, write code to evaluate its final result.

→ Hint: Use a stack-based approach.

3. SQL Question: Write a query to fetch the second highest salary in each department.

4. SQL Question: Given a table orders(order_id, customer_id, order_date, amount):

→ Find the top 3 customers by total spending.

→ Compute the running total of amount per customer, ordered by date.

Ankita Gulati

Shubh Goyal

5. Write a Python function to find the longest substring without repeating characters.

6. Write PySpark code to perform an optimized join between two large datasets.

7. Given a huge log file, how would you write Python or Spark code to find the most frequently occurring IP addresses?

Ankita Gulati

Shubh Goyal

Round 3
Big Data, Spark & Cloud

1. Explain Apache Spark fundamentals.
2. How do you handle Out Of Memory (OOM) errors in Spark?
   → Partition tuning, caching, broadcast variables.
3. How do you optimize Spark applications for performance?
4. Explain optimized joins in Spark (Broadcast joins, Skew joins).
5. What is data skewness in Spark and how do you fix it?
   → Key salting technique.
6. Discuss Apache Kafka fundamentals and how it integrates with Spark streaming.
7. Write an SQL query involving Joins and Group By together with real-world use case (e.g., total bookings by user across different platforms).

Ankita Gulati Shubh Goyal

8. How do you manage schema evolution in data lakes?

9. What strategies do you follow to ensure data quality, lineage, and observability in your pipelines?

10. In AWS, how would you design a scalable pipeline using S3, Glue, Kinesis, and Redshift?

Round 4
HR & Behavioral

1. Walk me through your career journey so far and highlight your most impactful projects.
2. What were your best and worst experiences in past companies?
3. How do you manage tight deadlines and still ensure quality delivery?
4. Why did you leave McKinsey & Company in just 4 months?
5. What are your career expectations from this role at Expedia?
6. How do you collaborate with cross-functional teams (Product, Analysts, Data Scientists)?

Thank You

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

Ankita Gulati

Shubh Goyal