INTUIT

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer 2
- **Experience:** 3+ years
- **Location:** Bengaluru
- **Work mode:** Hybrid
- **Compensation:** ₹40+ LPA
- **Total Rounds:** 5
- **Top Required Skills:**

1. Data Structures & Algorithms
2. PySpark
3. Streaming Pipelines
4. Big Data Concepts
5. System Design
6. Performance Tuning
7. Behavioral & Project Deep-Dive

Ankita Gulati

Shubh Goyal

Round 1

Data Structures, Algorithms & Big Data

DSA (Python)

a. Write a function to compute LCM of two numbers using both:

b. Brute-force approach

c. Efficient GCD-based approach

d. Find the Longest Palindromic Subsequence in a string. Optimize using Dynamic Programming ($O(n^2)$).

Big Data Concepts

a. What challenges arise when migrating Spark 2 to Spark 3?

b. Explain Adaptive Query Execution (AQE) and its benefits for skew joins.

c. What shuffle improvements were introduced in Spark 3?

d. How does columnar processing help query execution?

e. Share your approach to performance tuning in Spark jobs.

Ankita Gulati

Shubh Goyal

Round 2

Craft Demo (Hands-on POC, 2–3 days)

Focus Areas: Architecture, Pipeline Design, Code Implementation

Task: Build and present a streaming + batch pipeline.

Key Deliverables:

- Architecture & Design Presentation (pros/cons, scalability, performance).
- Data Models & Test Cases for raw, processed, and final tables.
- Repository Implementation with code + documentation.
- Assumptions & Trade-offs: Why batch vs. streaming? Partitioning choices?

Expected Design:

- Streaming: Kafka → Spark Structured Streaming → Delta Lake → Query Layer.
- Batch: S3 ingestion → PySpark ETL → Snowflake.

Performance Optimizations:

- Partition pruning
- Caching
- Indexing
- Parallel execution

Round 3
Assessor Round

1. SCD Type 2 in PySpark:
    a. Implement Slowly Changing Dimension Type 2.
    b. Handle inserts, updates, and history tracking with window functions & joins.
    c. Optimize using Delta Lake merge statements.
2. RDD-based Question:
    a. Write transformations & actions using RDD API.
    b. How would you minimize data shuffling and manage parallelism efficiently?

Ankita Gulati

Shubh Goyal

Round 4

Team Member Round

1. Delta Lake Internals:
   a. Explain Time Travel and ACID transactions.
   b. Why is compaction important?
2. HDFS Storage:
   a. How does HDFS store files in blocks?
   b. Explain block replication strategies for fault tolerance.
3. Kafka Streaming Concepts:
   a. How is offset management handled in Kafka?
   b. Explain retention policies for topics.
   c. How does Kafka maintain message ordering?

Round 5

Hiring Manager Round

1. Project Deep Dive:
    a. Explain an end-to-end pipeline you built.
    b. What optimizations did you apply for performance & cost?
2. Challenges & Scenarios:
    a. How do you handle a pipeline failure in production?
    b. When would you choose batch vs. streaming for ingestion?
3. Behavioral & Values:
    a. Share a time you had to resolve conflict in the team.
    b. How do you prioritize conflicting stakeholder requirements?
    c. Example of working under tight deadlines.

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal