ATLASSIAN

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 5+ years
- **Location:** Bengaluru
- **Work mode:** Remote
- **Compensation:** ₹40–55LPA
- **Total Rounds:** 5
- **Top Required Skills:**

1. Data Structures & Algorithms (interval problems, optimization techniques)
2. SQL (window functions, CTEs, ranking, deduplication)
3. ETL & Data Pipelines (batch + streaming, Delta Lake, schema evolution)
4. Cloud & Big Data Tools (Kafka, Spark, Flink, Presto, Snowflake)
5. System Design (scalability, fault tolerance, partitioning strategies)
6. Business Impact Analysis (metrics, CTR/CVR, recommendation systems)

Round 1
Data Structures, Algorithms & SQL

Duration: 90 minutes
Mode: Online Coding + SQL Challenge

Focus Areas: Problem-solving using DSA, interval merging logic, SQL with advanced window functions.

Questions:

1. You are given a list of intervals. Write an algorithm to merge all overlapping intervals and return a sorted list of non-overlapping intervals. Please explain how you would handle edge cases such as a single interval, fully overlapping intervals, and adjacent intervals. What would be the time and space complexity of your solution?
2. Suppose you are given a very large table of user activity with billions of rows. Write an SQL query using window functions (ROW_NUMBER, RANK, DENSE_RANK, LEAD, LAG) to identify the top three most active users per region. Explain how the PARTITION BY clause affects your query performance.
3. How would you compute running totals for daily sales data using SQL window functions? Walk through your approach with a Common Table Expression (CTE) and explain performance trade-offs when running such queries on a large dataset.

Round 2

Product Sense & Business Impact Analysis

Duration: 60 minutes
Mode: Case Study Discussion

Focus Areas: Product metrics interpretation, customer behavior analysis, business trade-offs.

Scenario Question:

1. Imagine you updated a recommendation algorithm for Atlassian products. After deployment, you notice that the Click-Through Rate (CTR) increased by 10%, but the Conversion Rate (CVR) decreased by 5%. How would you analyze this situation? Please describe the funnel (Impressions → Clicks → Conversions), explain what this result indicates, and propose an action plan to improve both CTR and CVR.
2. What additional metrics would you measure in this scenario (such as bounce rate, session duration, product match score, pre-click vs post-click behavior)? How would you run an A/B test to validate whether the updated algorithm should be rolled out globally?

Round 3
Techno-Managerial Discussion

Duration: 75 minutes
Mode: Video Call with Senior Engineer + Manager
Focus Areas: Data architecture, schema evolution, fault-tolerant pipeline design.

Questions:

1. Can you explain what schema enforcement and schema evolution mean in Delta Lake? How do these features ensure data consistency in production? Provide an example where schema evolution could introduce risks if not carefully managed.
2. Suppose a bad upstream file introduces unintended schema changes into your production pipeline. How would you detect this issue and safeguard downstream consumers such as BI dashboards?
3. Design a scalable, fault-tolerant real-time streaming pipeline for ingesting event data from multiple sources. Which technologies (e.g., Kafka, Spark Structured Streaming, Delta Lake) would you use, and why?
4. How would you ensure exactly-once semantics in such a streaming pipeline? Describe how you would use checkpoints, idempotent writes, and partitioning strategies.
5. Discuss the trade-offs between batch processing and streaming in the context of real-time analytics for Atlassian. When would you prefer one over the other, and why?

Round 4
Atlassian Values & Leadership

Duration: 45 minutes
Mode: Behavioral Interview

Focus Areas: Collaboration, ownership, openness, and learning from failure.

Behavioral Questions:

1. Tell me about a time when you failed while working on a data engineering project. What exactly happened, how did you respond, and what lessons did you learn from that failure?
2. Describe a project where you had to take ownership despite ambiguity. How did you prioritize tasks, communicate with stakeholders, and deliver results without waiting for explicit instructions?
3. Atlassian values openness and collaboration. Can you share an example where you worked across multiple teams (engineering, product, QA) and resolved conflicts to achieve a common goal?

Round 5
Data Architecture & ETL Design

Duration: 75 minutes
Mode: Whiteboard / Virtual System Design
Focus Areas: Designing large-scale ETL systems, data modeling, handling late-arriving data, deduplication, performance optimization.

Scenario Questions:

1. You are asked to design an ETL pipeline for analyzing ad click data with more than 1 million records per day. The raw ad logs come from Google DoubleClick and are stored in S3 or GCS. How would you design the ingestion, processing, and storage layers? Explain the concepts of Bronze (raw), Silver (processed), and Gold (analytics-ready) tables in your answer.
2. How would you handle late-arriving data in this pipeline? Please explain how watermarking in Spark or partition overwrites could be applied.
3. Deduplication is critical in ad click data. Describe how you would identify and remove duplicates using SQL window functions like ROW_NUMBER() or using event IDs.
4. What kind of schema design would you use for storing impressions and clicks data? Would you model them in separate tables and join later, or create a star schema for analytics? Provide detailed reasoning.
5. How would you monitor and optimize the performance of this ETL pipeline? What would you do if Spark jobs had long shuffles or skewed joins?

Thank You

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

Ankita Gulati

Shubh Goyal