Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 4+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹35+ LPA
- **Total Rounds:** 4
- **Top Required Skills:**
  - SQL
  - Big Data File Formats & Storage
  - Hive & Spark Fundamentals
  - Data Modeling
  - Query optimization

Ankita Gulati

Shubh Goyal

Round 1
Machine Coding / Spark Coding

1. Given 4 nested JSONs (players, teams, matches, scores), flatten them and write Spark code to join them into a single analytics-ready DataFrame.
2. How would you handle deeply nested arrays inside JSON fields? (e.g., players having multiple skills stored in arrays).
3. After joining, queries were asked such as:
    a. Find the top 5 players by total runs.
    b. Find the team with the highest win percentage.
    c. Return the stadium where maximum matches were played.
4. What caching strategy would you apply if these DataFrames are reused multiple times in the job? Explain why persist(MEMORY_ONLY) vs persist(MEMORY_AND_DISK) vs broadcast() matters.
5. If your Spark job runs into OOM (Out of Memory) errors while joining large tables:
    a. How would you debug the issue?
    b. Which Spark configs would you tune (spark.sql.shuffle.partitions, executor.memory, broadcastTimeout)?

Ankita Gulati Shubh Goyal

Round 2
Data Modeling

1. Draw an ER diagram for cricket tournament data.
    a. Entities: Players, Teams, Matches, Scores, Stadiums.
    b. Relationships: A player belongs to multiple teams, a team plays multiple matches, each match is in one stadium.
2. Implement 1-to-many and many-to-many relationships:
    a. How will you design schema to capture a player playing in both IPL and National Team?
    b. Where would you keep constraints? (junction table / association table).
3. Write SQL queries on top of the model:
    a. Calculate cumulative runs scored by each player across the tournament.
    b. List the top 3 bowlers by wickets taken per team.
    c. Find the stadium that hosted the maximum matches.
    d. Return the player with the highest batting average.
    e. Identify teams with more than 2 consecutive wins.
4. Follow-up: If data volume grows to millions of matches and players, how would you partition/federate the schema for performance at scale?

Round 3

Data Pipeline Handling & Spark Utilization

1. Data Pipeline Scenarios
    - How do you design an incremental data pipeline when only delta records arrive each day?
    - What would you do if late-arriving records are ingested after SLA cut-off?
    - How do you ensure idempotency of pipelines when reprocessing the same data multiple times?
2. Kafka & Streaming
    - Explain how you would design a pipeline consuming events from Kafka topics (e.g., clickstream or order events).
    - How do Kafka partitions, offsets, and consumer groups help scale throughput?
    - How do you handle backpressure when consumer is slower than producer?
    - How would you ensure exactly-once processing in Spark Structured Streaming with Kafka as source?

Ankita Gulati

Shubh Goyal

3. Spark Concepts
    - What is dynamic allocation in Spark? How does it help in resource optimization?
    - Explain scenarios where you'd use cache() vs persist() vs checkpoint().
    - Difference between shuffle join, broadcast join, sort-merge join – when do you use each?
    - How do you handle data skewness in Spark jobs (salting, repartitioning, etc.)?
4. Project Discussion
    - Walkthrough of a past end-to-end project: ingestion, transformations, warehousing.
    - What challenges did you face (OOM, skew, schema evolution)?
    - How did you optimize Spark pipelines to meet SLAs?

Round 4
Behavioral Interview

1. Past Experience & Role Fitment
    a. Walk me through your career so far.
    b. Which project are you most proud of, and why?
    c. What was the most challenging bug you solved in production?
2. Behavioral (STAR Method)
    a. Tell me about a time when you had a disagreement with your team on architecture. How did you resolve it?
    b. Share a situation where you worked under very tight deadlines. How did you manage stress?
    c. Describe a time when you introduced an optimization or improvement in a pipeline that had a measurable impact.
3. Collaboration & Stakeholder Management
    a. How do you handle conflicts between engineering teams and business stakeholders?
    b. If product demands real-time reporting but infra budget is limited, how would you negotiate trade-offs?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal