Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 2+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹15-22 LPA
- **Total Rounds:** 3
- **Top Required Skills:**
  - SQL
  - Big Data File Formats & Storage
  - Hive & Spark Fundamentals
  - Data Modeling
  - Query optimization

Ankita Gulati

Shubh Goyal

Round 1
SQL Deep Dive

1. Year-on-Year Growth
   - Table growth(customer, year, sales) was given.
   - Write a query to calculate YoY % growth in sales for each customer.
   - Follow-up: How would you solve using window functions (LAG) vs self-join? Which one is more efficient?
2. Department with Highest Salary Expenditure
   - From employee and department tables, write a query to return the department with maximum total salary.
   - Follow-up: How would the query change if multiple departments tie with the same max salary?
3. Employees with the Same Salary
   - Find employees who have the same salary.
   - Additional twist: Return all possible employee pairs with the same salary (excluding duplicates).
4. Employee with the 4th Highest Salary
   - Write an SQL query to return employee(s) having the 4th highest salary.
   - Interviewer asked to explain difference between RANK(), DENSE_RANK(), and ROW_NUMBER() in this case.

Ankita Gulati Shubh Goyal

5. Departments with More Than One Employee
   - Identify departments with more than one employee.
   - Follow-up: How would you handle if the interviewer wants only departments with exactly 2 employees?
6. Departments with No Employees
   - Return all department names where there are no employees.
   - Follow-up: Would NOT IN vs LEFT JOIN … IS NULL make any difference in performance?
7. Departments and Their Employees
   - Write query to return all employees grouped by department in a comma-separated format.
   - Follow-up: If dataset is huge (millions of employees), what performance issues may arise?

Round 2
Data Engineering & Big Data Concept

1. File Formats: CSV vs ORC vs Parquet vs Avro
   - Explain the characteristics of each format: storage, compression, schema evolution, query performance.
   - Scenario asked: If you are building a reporting system vs a streaming ingestion system, which format would you choose and why?
2. Hive Partitioning & Bucketing
   - How does partitioning improve performance in Hive?
   - What problems arise if you over-partition your table?
   - When should bucketing be preferred over partitioning?
   - Example: If you have a sales dataset partitioned by date, and queries filter by both date and region, how would you optimize it?

3. Repartition vs Coalesce (in Spark)
   - What is repartitioning, and when is it needed?
   - What is coalescing, and how does it differ?
   - Scenario: If a Spark job creates 200 tiny partitions, which function would you use and why?
   - Another scenario: If you want to evenly distribute data before a heavy join, which method is better?
4. Data Modeling – Library Database
   - Design a schema for a library management system.
   - Identify dimensions (Books, Customers, Branches) and facts (Loans, Reservations).
   - Follow-up: How would you model if one book has multiple authors?
   - Follow-up: How would you capture overdue penalties?

Round 3
HR & Managerial Round

1.Project Experience

- Explain your most recent project in detail: architecture, tech stack, challenges faced.
- What optimizations did you introduce in your pipeline?

2. **Behavioral Questions**

- Tell me about a time when you had conflicting priorities. How did you manage?
- Describe a time you had to debug a production issue at midnight. What steps did you take?

3. **Career Motivation**

- Why Flipkart? What excites you about the scale of Flipkart's data?
- What are your long-term career goals in Data Engineering?

4. **Teamwork & Leadership**

- How do you collaborate with analysts, product managers, and business teams?
- Have you mentored junior engineers? Share an example.

5. **Workplace Scenarios**

- Suppose a critical sales dashboard is broken during festival season (high load). How would you handle immediate resolution vs long-term fix?
- How would you prioritize work if business and engineering teams have conflicting demands?

Ankita Gulati Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal