Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer III
- **Experience:** 2.5 years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹25+ 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
Technical

Introduction & Experience

1. Walk me through your previous tech stack and experiences.
2. What data volumes have you worked with?
3. Describe the type of work you've done and its business impact.

SQL Questions

4. Write a query to identify continuous date ranges having the same status.
   • Approach: Use ROW_NUMBER() with grouping to find gaps and identify ranges.

5. Given 2 tables with NULL values, determine the row count outputs for different joins:
   • INNER JOIN
   • LEFT JOIN
   • RIGHT JOIN
   • FULL OUTER JOIN

DSA Question

6. Solve the Group Anagrams problem:

• Input: ["eat", "tea", "tan", "ate", "nat", "bat"]

• Output: [["bat"], ["tan","nat"], ["eat","tea","ate"]]

• Approach: Use a dictionary with sorted strings as keys to group anagrams.

PySpark

7. Write a PySpark script to union two datasets with different schemas.

• Handle different row counts, column names, and schema mismatches.

• Approach: Use column alignment and unionByName(allowMissingColumns=True).

Round 2
PySpark and SQL

PySpark Coding Tasks

1. Write a PySpark job to compute word frequency from a text file.
- Approach: Use flatMap → map → reduceByKey.

2. Given Spark code snippets, identify:
- Narrow vs. wide transformations.
- Number of stages created.

3. Compute the yearly salary for each employee along with department name using PySpark.

Spark Internals

4. Explain the Spark architecture: jobs, stages, and tasks.

SQL Tasks

5. Write a query using LEFT ANTI JOIN to find the number of records present in table1 but not in table2.

Ankita Gulati Shubh Goyal

Data Modeling Task

6. Given Customer, Product, and Orders tables:

• Identify fact vs. dimension tables.

• Validate schema rules (e.g., fact PK should not be an FK in a dimension table).

Ankita Gulati

Shubh Goyal

Round 3
Spark & SQL - Advanced Concepts

Spark Deep Dive

1. Explain Spark architecture end-to-end.
2. Spark memory management for:
- Small clusters
- Small data
- Large data
3. What is hardcoded reserved memory in Spark?
- Can it be modified?
- When should it be changed?

Spark Versions

4. Spark 2 vs Spark 3 differences:
- AQE (Adaptive Query Execution)
- Dynamic Partition Pruning
- Pandas UDFs

Cloud Cost Optimization

5. How do you optimize storage and compute costs in the cloud?

Ankita Gulati Shubh Goyal

6. From sales and refund tables, find customers who:

- Cancelled an order (order_status = 'cancelled')
- Did not get a refund
- Placed another order after the cancellation
- Sales Table schema:

customer_id | order_id | order_plcd_ts | order_shipped_dt | order_amt | order_status

- Refund Table schema:

order_id | order_refund_dt | refund_amt

Git

7. Explain the conceptual and practical differences between git fetch and git pull.

Round 4
HR Discussion

Project & Role Discussion

1. Walk me through your past projects in detail.
- Data pipelines, tools used, challenges solved.

CI/CD & Automation

2. Explain how you've implemented CI/CD and pipeline automation.
3. How do you set up Airflow in AWS (MWAA)?

Spark Optimization

4. Compare Repartition vs Coalesce:
- When and why to use each.
- Pros & cons.

Team & Role Fit

5. Discussion on role responsibilities and team expectations at Walmart.

Ankita Gulati Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal