publicis sapient

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹23-25 LPA
- **Total Rounds:** 3
- **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
Coding Assessment

1. Perform complex data transformations using PySpark DataFrame APIs.
2. Apply joins (inner, left, semi, anti) and window functions (ROW_NUMBER, RANK).
3. Handle nested structures, null values, and schema enforcement.
4. Implement aggregations (groupBy, agg) and filtering using conditions.
5. Write PySpark code to detect and handle schema evolution in incoming datasets.
6. Optimize performance by using repartition, coalesce, and broadcast joins.
7. Write a PySpark job to calculate the running total of sales by customer using window functions.
8. Given two DataFrames with different schemas, merge them into a single schema while handling missing columns.

Ankita Gulati

Shubh Goyal

9. Write a PySpark job to calculate the running total of sales by customer using window functions.
10. Given two DataFrames with different schemas, merge them into a single schema while handling missing columns.
11. Write PySpark logic to flatten a nested JSON structure and load it into a DataFrame.
12. Explain how Delta Lake or Apache Hudi helps with data versioning and ACID compliance in big data pipelines.
13. How do you implement idempotent pipelines to ensure no duplicate processing occurs?
14. Compare Athena vs Redshift Spectrum vs Presto for querying S3 data.

Round 2
Project & Coding Round

1. Write an SQL query to find the nth highest salary department-wise using multiple approaches (ROW_NUMBER, DENSE_RANK, subqueries).
2. Write a query using CTEs to identify employees who earn more than the department average.
3. Use window functions (LAG, LEAD, NTILE) to detect performance trends in employee data.
4. Write an SQL query to pivot and unpivot data dynamically.
5. Write a query to detect duplicate records and retain only the latest one using timestamps.
6. Write a PySpark job to calculate customer churn rate using conditional logic (when, otherwise).
7. Given a dataset of orders, calculate the average order value by month using groupBy and window.
8. Handle null values in a DataFrame by applying imputation (e.g., replace nulls with median).

Ankita Gulati Shubh Goyal

9. Write code to handle schema mismatch between two incoming datasets and unify them.
10. What are the most common performance tuning techniques in Spark? (e.g., caching, partitioning, broadcast joins).
11. Explain narrow vs wide transformations with examples.
12. How does the Spark execution plan (DAG) work, and how do you analyze it using explain()?
13. How do you handle data skew in Spark jobs? Provide approaches.
14. Walk me through a recent project architecture where you used Spark and cloud (AWS/Azure/GCP).
15. How do you design a data lake architecture with batch + streaming ingestion?
16. What orchestration tools have you used (Airflow, AWS Step Functions, Azure Data Factory)?
17. How do you implement data quality checks in your ETL pipelines?

Ankita Gulati Shubh Goyal

18. Explain how you would build a CDC (Change Data Capture) pipeline in Spark or AWS Glue.
19. Discuss your role in data ingestion, transformation, and performance tuning in past projects.

Ankita Gulati

Shubh Goyal

Round 3
HR & Behavioral

1. Walkthrough of your resume and project highlights.
2. What are you looking for in your next role at Publicis Sapient?
3. Discussion on work culture fit and collaboration style.
4. Expected salary, location preference, and joining timeline.
5. Career aspirations: individual contributor vs leadership path.

Thank You

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

Ankita Gulati

Shubh Goyal