Persistent

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹20-25 LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark / Python / Databricks
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming
6. System Design

Ankita Gulati

Shubh Goyal

Round 1
Core Data Engineering

1. What is the difference between cache() and persist() in Spark? In which scenarios would you prefer one over the other?
2. Explain the differences between map, flatMap, and mapPartitions in PySpark with real-world use cases.
3. What are wide and narrow transformations in Spark? Provide examples of each.
4. How does Spark handle data skew during shuffling? What strategies can you use to optimize skewed joins?
5. How does the HDFS architecture ensure fault tolerance and data reliability?
6. Explain the roles of NameNode, DataNode, and Secondary NameNode in HDFS.
7. What are the differences between internal and external tables in Hive?
8. Explain static vs dynamic partitioning in Hive with an example.

Ankita Gulati Shubh Goyal

9. Write a SQL query to find the second highest salary from an Employee table.

10. What is the difference between RANK(), DENSE_RANK(), and ROW_NUMBER()? Provide a use case for each.

11. What are broadcast joins in Spark? When would you use them?

12. Explain the concept of Catalyst Optimizer in Spark SQL.

13. What are the trade-offs between Parquet, ORC, and Avro file formats?

Ankita Gulati

Shubh Goyal

Round 2
Cloud & Advanced Coding

1. How would you build a batch data pipeline in AWS using S3, Glue, and Athena?
2. Compare EMR vs Glue for running Spark jobs. When would you prefer one over the other?
3. Explain how to use AWS Lambda with S3 to trigger a data transformation pipeline.
4. How would you implement incremental data loads in AWS Glue?
5. What strategies would you use for cost optimization in EMR?
6. Design a pipeline to handle real-time streaming + batch data in AWS (using Kinesis, Glue, S3, Redshift).
7. How would you monitor and retry failed jobs in Airflow or AWS Step Functions?
8. Explain how to enforce data governance and lineage in a cloud-based data lake.

Ankita Gulati

Shubh Goyal

9. Write a PySpark program to find the highest-paid employee in each department.

10. SQL Coding Question
Table: orders(order_id, customer_id, order_date, total_amount)
→ Write a SQL query to calculate total revenue per customer in the last 30 days

11. Hive Coding Question
Table: sales_data(product_id, region, sale_date, sale_amount)
→ Write a Hive query to partition this table by region and bucket by product_id into 5 buckets.

12. Python DSA: Write a function to find the longest substring without repeating characters.

13. SQL: Write a query to return the top 3 products by sales in each region (use window functions).

Ankita Gulati Shubh Goyal

Round 3
HR Discussion

1. Walk me through your current project responsibilities at a high level.
2. Describe a time when you faced a critical production issue. How did you resolve it?
3. How do you collaborate with cross-functional teams (business analysts, data scientists, DevOps)?
4. Tell me about a situation where you had to handle data quality issues in production.
5. Why are you interested in joining Persistent Systems?
6. What are your career aspirations over the next 3–5 years ?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal