Morgan Stanley

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ years
- **Location:** Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹30+ LPA
- **Total Rounds:** 5
- **Top Required Skills:**

1. Advanced SQL
2. Python
3. Big Data
4. Cloud
5. Data Modeling
6. DevOps & Agile basics
7. Data Structures
8. Behavioral

Ankita Gulati

Shubh Goyal

Round 1
Preliminary Online Assessment

1. SQL Coding Questions
    a. Write SQL queries involving aggregates, joins, subqueries, unions, group by, having.
    b. Focus on window functions: ROW_NUMBER(), RANK(), DENSE_RANK(), LEAD(), LAG().
    c. Example: Find the 3rd highest salary per department using window functions.
2. Python & SQL MCQs
    a. Python: Lists, Sets, Tuples, Dicts, comprehension.
    b. SQL: difference between INNER JOIN vs LEFT JOIN, UNION vs UNION ALL, clustered vs non-clustered index.
3. One Coding DSA Question
    a. Medium-level problem (Array / String / Stack / Queue / Linked List / BST).
    b. Example: Given a string, remove adjacent duplicates recursively.
4. Relational DB Transactions & Admin MCQs
    a. ACID properties, isolation levels, transactions.
    b. Referential integrity and normalization.
5. Unix & OS Questions
    a. Basic shell commands (grep, awk, sed, find).
    b. File permissions (chmod 755).
    c. Process management (kill, ps, top).

Ankita Gulati Shubh Goyal

Round 2
Technical Interview

1. SQL & Joins
   a. Given two tables A and B (id column): Find number of rows in output for INNER, LEFT, RIGHT, FULL OUTER JOIN.
   b. Find highest salary in each department using window functions. → Explain why DENSE_RANK() is better than RANK() for handling ties.
   c. Queries on LEAD, LAG, NTILE with practical scenarios.
2. Hive & Sqoop
   a. Explain Hive architecture and default metadata database → (Ans: Derby).
   b. Differences between partitioning vs bucketing with use cases.
   c. Difference between external vs managed tables (ACID, metadata).
   d. If an external table is dropped, how can data still be accessed?
   e. Sqoop incremental import → Write command for importing only new rows from MySQL into HDFS.
   f. Sqoop import all tables command with example.
3. Spark / Hadoop
   a. Explain rack awareness in HDFS.
   b. What happens when you submit a Spark job? Walk through DAG creation, stages, tasks.
   c. Difference between narrow vs wide transformations with examples (map vs groupByKey).
   d. Coalesce vs Repartition: performance, number of partitions after operation.
   e. How to schedule Spark jobs in Databricks.
   f. How Hadoop achieves high availability.

Ankita Gulati

Shubh Goyal

4. Cloud & AWS
    a. Write Python boto3 code to upload CSV to S3 bucket. (pseudo-code acceptable).
    b. Discuss IAM roles & policies for access control.
    c. Explain S3 storage mechanism.
5. DevOps & Agile
    a. CI/CD: expand acronym and explain difference between Continuous Delivery vs Deployment.
    b. Explain GitLab CI/CD workflow and runners.
    c. Tools familiarity: Jira, Jenkins.
    d. Agile: sprint, scrum roles, iterative vs waterfall.

Ankita Gulati

Shubh Goyal

Round 3
Technical Interview

1. Design a relational schema for employee & department data. → Apply normalization and denormalization techniques.
2. Explain Databricks Lakehouse architecture (ingestion, storage, transformation, query layers).
3. Design an ETL pipeline: what preliminary checks must be considered (schema validation, deduplication, null handling, partition strategy)?
4. Batch vs Stream processing → when to choose Spark Streaming.
5. What is the purpose of a staging layer in a pipeline?
6. How do you write unit test cases for SQL & PySpark? Example scenarios.
7. Write pseudo-code for a dummy ETL pipeline in Python: read → clean → transform → write to CSV.
8. Spark monitoring: how do you optimize long-running Spark jobs (caching, partition tuning, AQE)?

Ankita Gulati Shubh Goyal

Round 4
Techno-Managerial Round

1. AWS Glue: How does it fetch metadata of parquet/CSV files stored on S3?
2. What is Parquet format? Why is it preferred in Delta tables?
3. Advantages of Delta Lake tables (ACID, schema evolution, time travel).
4. Spark cluster config in your project: executors, cores, memory tuning.
5. Explain internal working of Hadoop (Namenode, Datanode, fault tolerance).
6. Questions on Redshift vs relational DBs.
7. Team Fitment: How do you overcome challenges in your team?
8. Leadership values: Explain a situation when you inspired your team to deliver under tight deadlines.

Ankita Gulati

Shubh Goyal

Round 5
Hiring Manager Round

1. Tell me about your strengths and weaknesses.
2. Walk me through your family background.
3. Share your career aspirations.
4. Why do you want to join Morgan Stanley?
5. Why should we hire you over other candidates?
6. Salary discussion + negotiation.

Thank You

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

Ankita Gulati

Shubh Goyal