Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ 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

1. Walk me through your past experiences and key projects in data engineering.
2. What challenges have you faced in your work, and how did you diagnose and resolve them?
• Example: using Spark plans and explain() for debugging.
3. Explain different data loading strategies and the file formats you used. Why did you choose them?
4. What is the Medallion architecture? Which languages and tools did you use to implement it?
5. What data quality (DQ) measures do you apply at each stage of a pipeline? What transformations are typically required?
6. Compare and contrast:
• REPARTITION vs. COALESCE
• OPTIMIZE command in Delta Lake
• Delta vs. Parquet
7. How would you implement incremental loading using Delta file formats?

Ankita Gulati Shubh Goyal

Round 2
Technical MCQs

Apache Spark — Core Concepts (1–7)

1. In Spark, which transformation is narrow?
a) reduceByKey
b) map
c) groupByKey
d) join
Answer: b) map

2. Which of the following triggers the execution of a Spark job?
a) map
b) filter
c) reduceByKey
d) count
Answer: d) count

Ankita Gulati

Shubh Goyal

3. Spark lazy evaluation means:
a) Data is processed immediately after each transformation.
b) Transformations are stored and executed only when an action is called.
c) Spark does not store execution plans.
d) Spark jobs never fail.
Answer: b)

4. Which operation is more efficient for aggregations?
a) groupByKey
b) reduceByKey
c) map
d) flatMap
Answer: b)

5. What is a wide transformation in Spark?
a) Transformation that requires shuffling of data between executors.
b) Transformation that works on a single partition only.
c) Any transformation performed on DataFrames.
d) Operations that do not trigger execution.
Answer: a)

6. Which Spark component creates the logical and physical execution plan?
a) Catalyst Optimizer
b) DAG Scheduler
c) Task Scheduler
d) Cluster Manager
Answer: a)

7. In Spark SQL, which file format is most efficient for analytical queries?
a) CSV
b) JSON
c) Parquet
d) TXT
Answer: c)

Joins in Spark (8–11)

8. Which join in Spark can cause data skew most often?
a) Broadcast join
b) Shuffle sort merge join
c) Bucketed join
d) Partitioned join
Answer: b)

9. For small lookup tables, which join strategy is recommended in Spark?
a) Shuffle hash join
b) Broadcast join
c) Sort merge join
d) Nested loop join
Answer: b)

10. Which join ensures null-safe equality check?
a) INNER JOIN
b) LEFT JOIN
c) RIGHT JOIN
d) join(..., Seq("colA"), "inner").na.fill()
Answer: a trick → requires <=> operator in Spark SQL.

11. Which Spark property controls the number of shuffle partitions?
a) spark.executor.instances
b) spark.sql.shuffle.partitions
c) spark.sql.autoBroadcastJoinThreshold
d) spark.dynamicAllocation.enabled
Answer: b)

Delta Lake (12–18)

12. Which feature makes Delta Lake ACID-compliant?
a) File formats
b) Transaction logs
c) Spark SQL optimizer
d) Z-order clustering
Answer: b)

13. What does the OPTIMIZE command in Delta Lake do?
a) Deletes old snapshots
b) Compacts small files into larger ones
c) Deletes null values
d) Repartitions data randomly
Answer: b)

14. Which Delta command physically deletes old data files no longer needed?
a) VACUUM
b) OPTIMIZE
c) DELETE
d) MERGE
Answer: a)

15. By default, Delta's VACUUM retention period is:
a) 1 hour
b) 7 days
c) 30 days
d) 90 days
Answer: b) 7 days

16. In Delta Lake, MERGE INTO is used for:
a) Schema evolution
b) Upserts (insert + update)
c) Z-ordering
d) Vacuuming files
Answer: b)

17. Which file format underlies Delta Lake?
a) ORC
b) Parquet
c) Avro
d) JSON
Answer: b)

18. Delta Lake time travel feature allows:
a) Restoring deleted files
b) Querying older snapshots of data by version/timestamp
c) Skipping schema checks
d) Bypassing transactions
Answer: b)

Performance Tuning (19–22)

19. When should you use REPARTITION over COALESCE?
a) When increasing partitions
b) When reducing partitions
c) When file sizes are fixed
d) When caching data
Answer: a)

20. Which Spark config controls broadcast join threshold?
a) spark.sql.shuffle.partitions
b) spark.sql.autoBroadcastJoinThreshold
c) spark.executor.memory
d) spark.sql.files.maxPartitionBytes
Answer: b)

21. What does Z-Ordering in Delta Lake optimize?
a) Write speed
b) Skew handling
c) Data skipping for multi-dimensional queries
d) File compaction
Answer: c)

22. In Spark, how do you identify data skew?
a) By checking executor logs
b) By analyzing skew metrics in Spark UI
c) By counting partitions manually
d) By enabling caching
Answer: b)

Miscellaneous (23–25)
23. Which Spark component breaks a job into stages and tasks?
a) DAG Scheduler
b) Catalyst Optimizer
c) Block Manager
d) Task Scheduler
Answer: a)

24. In Spark Streaming, the checkpointing mechanism is used for:
a) Storing raw data
b) Recovering from failures
c) Reducing latency
d) Avoiding shuffles
Answer: b)

25. Which Delta Lake feature handles schema evolution automatically?
a) AUTO MERGE ON
b) OPTIMIZE
c) VACUUM
d) REPARTITION
Answer: a)

Round 3
HR Discussion

1. Why are you looking to switch roles?

2. What specifically attracts you to Puma as your next employer?

3. Where do you see your career growth over the next 5 years?

Ankita Gulati

Shubh Goyal

Round 4
Group Case Study

Format: Team exercise (4 members: 3 Data Engineers + 1 DevOps).

Task:

- Design an end-to-end pipeline architecture based on a given problem statement.
- Ensure that you follow best practices.
- Present the solution on a whiteboard, covering:
  - Architecture design
  - Data ingestion, processing, storage, and analytics
  - Edge cases and scalability considerations
  - Monitoring and reliability

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal