



Data Engineering Interview Questions



Ankita Gulati

Shubh Goyal



Job Details

- **Position:** DataEngineer
- **Experience:** 3–4 years
- **Location:** Bangalore, Gurgaon
- **Work mode:** Office
- **Compensation:** ₹22–₹30 LPA
- **Total Rounds:** 4
- **Top Required Skills:**
 1. PySpark
 2. SQL
 - 3.DSA (Python)
 - 4.System Design
 - 5.Business Communication
 - 6.Data Engineering Best Practices

Round 1

PySpark Coding Challenge (HackerEarth CodePair)

1. Apply DataFrame operations:
 - filter, groupBy, agg, join, withColumn.
2. Handle nulls & missing values:
Use ◦ dropna(), na.fill().
Decide based on business logic.
3. Perform deduplication with distinct keys.
4. Show how to optimize performance:
 - Use partitioning.
 - Cache/persist when reused multiple times.
 - Explain difference between narrow vs wide transformations.
5. Debug execution bottlenecks:
 - Use explain() to analyze DAG.
 - Spot shuffles.
6. Use broadcast joins for small lookup tables.

Expectation: Candidate should demonstrate hands-on Spark fluency + performance tuning mindset.

Ankita Gulati

Shubh Goyal

Round 2

Problem Solving & Data Structures (Python + DSA)

Python Concepts

- Iterators vs Generators:
 - Iterators use `__iter__()` & `__next__()`.
 - Generators use `yield` → memory-efficient.
 - Use case in large-scale ETL pipelines.

DSA Questions

1. Anagram Grouping

- Group words into anagram sets.
- Approach: Hashmap with sorted string as key.
- Complexity: $O(n \cdot k \log k)$.

2. Word Formation

- Can a word be formed from a given character map?
- Use `collections.Counter`.
- Check frequency sufficiency.

3. Coin Change Problem (Min Coins)

- Dynamic Programming (bottom-up).
- Complexity: $O(n \cdot \text{amount})$.

Round 3

Process & Business Communication

1. Past Projects: Deep dive into 1–2 ETL pipelines.

- Problem statement.
- Tech stack (Spark, Kafka, AWS).
- Measurable business impact.

2. Travel Data Domain Fit:

- Handling cross-timezone datasets.
- Schema evolution for dynamic booking/payment systems.

3. Openness:

- Relocation, timezone collaboration, ambiguity handling.

Python Best Practices Discussion:

- Modular coding (functions, classes).
- Use of *args, **kwargs.
- Config-driven ETL logic.
- Logging & exception handling in pipelines.
- Writing reusable utilities.

Behavioral Expectation:

- STARmethod (Situation, Task, Action, Result).
- Tie answers to business ROI.
- Example: “Improved SLA from 4h → 30m with Spark tuning.”

Round 4

System Design & SQL Challenges

System Design Problem:

Design Uber-like Ride-Hailing Backend

- Core Entities: Riders, Drivers, Trips, Payments, Locations.
- Event Streaming: Kafka (driver updates, trip events).
- Geo-Spatial Matching: H3 indexing or R-Tree.
- Real-Time Store: Redis/Cassandra for driver availability.
- Scalability: Auto-scaling, partitioned topics, retries with DLQs.

Expected Discussion Points:

- Draw ER diagrams first.
- Technology trade-offs: Kafka vs Kinesis, H3 vs GeoHash.
- Explain read/write trade-offs & hotspot handling.

Ankita Gulati

Shubh Goyal

SQL Questions:

1.5thHighestSalary

- Use ROW_NUMBER() or DENSE_RANK().
- Must explain difference between them (ties).

2.Trip History Query

- Query on Trips table with filters (driver_id, trip_date, status).
- Discuss use of indexes.
- Avoid functions on indexed columns.
- Show ability to read query plans.

Final Takeaways from Expedia Process

- PySpark: Performance tuning (lazy eval, joins, caching, partitioning).
- DSA in Python: LeetCode Medium, arrays/hashmaps/DP.
- System Design: Data engineers must think like backend engineers → scalable, event-driven, real-time pipelines.
- SQL: Expect simple but optimized solutions.
- Business Impact: Always connect technical work to SLA, ROI, or customer experience.

Prep Strategy

Hands-on: Build batch + streaming pipelines with Spark + Kafka.

Mock interviews: Practice explaining trade-offs clearly.

Projects: Prepare 2–3 STAR stories with quantifiable impact.

SQL/Python: Practice on HackerRank, StrataScratch, LeetCode.

Ankita Gulati

Shubh Goyal

Thank You

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

