Microsoft

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 4+ years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹30+ 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. Explain how jobs, stages, and tasks are executed in Spark. How can you analyze them using the Spark DAG?
2. Differentiate between narrow vs. wide transformations in Spark. Give examples and discuss their performance impact.
3. Compare broadcast joins vs. shuffle joins in Spark. When would you use each, and why?
4. How would you identify and optimize a Spark job suffering from skew or long-running stages?
   - Use .explain() to check physical plans.
   - Use Spark UI to identify bottlenecks.
5. Explain the architecture of HDFS (NameNode, DataNode, block size, replication factor).
6. How does YARN schedule applications and allocate resources?
7. Compare MapReduce vs. Spark. When is each preferable?

Ankita Gulati

Shubh Goyal

8. Differentiate between fact and dimension tables. Provide examples.

9. Explain star schema vs. snowflake schema. Which would you use in a read-heavy workload?

10. How does partitioning and indexing improve query performance?

11. Explain how you would implement Slowly Changing Dimensions (SCD Type 2) in a data warehouse.

12. Why are surrogate keys preferred over natural keys in dimensional modeling?

Ankita Gulati

Shubh Goyal

Round 2
Data Structures & Algorithms

Problem 1 — Dutch National Flag Problem

1. Solve the Dutch National Flag problem (sorting an array of 0s, 1s, and 2s).

- Explain the three-way partitioning approach with two pointers.
- Why is this more efficient than naive sorting?

Problem 2 — Interval Merging

2. Solve the interval merging problem.

- Approach 1: Sort intervals by start time and merge overlaps → O(n log n).
- Approach 2: Optimize using bucket sort or prefix-sum techniques when ranges are known.

3. Discuss time and space complexity trade-offs between naive and optimized approaches.

Ankita Gulati

Shubh Goyal

Round 3
System Design & Modeling

1. How would you design a data pipeline for large-scale data ingestion and processing?
• When would you use batch processing vs. streaming?
• Which tools would you choose for near real-time ingestion (e.g., Kafka + Spark Streaming)?
2. Discuss storage format choices:
• Parquet (columnar, analytical queries).
• Avro (row-based, schema evolution).
3. How would you optimize the compute & query layer (Spark, Snowflake, Redshift)?
• Partition pruning, clustering, caching, materialized views.
4. How do you handle Slowly Changing Dimensions (SCDs) in dimensional models? Which type would you choose in different scenarios?
5. Why are surrogate keys preferred in dimensional modeling?

Ankita Gulati Shubh Goyal

6. How would you partition a large fact table to improve query performance?

7. How can you minimize shuffles in Spark joins? (Partitioning strategies, co-location, broadcast joins).

Ankita Gulati

Shubh Goyal

Round 4
Techno-Managerial Fit

1. Walk through a past project's architecture and defend your design choices.
2. Describe a time you resolved a performance bottleneck in Spark (e.g., skewed data, inefficient joins).
3. How would you set up CI/CD pipelines for data workflows?
   • Tools: Airflow, GitHub Actions, Jenkins, Databricks Jobs.
   • Include testing, deployment, rollback, and validation.
4. Describe a time when you had to quickly adapt to a new tool or workflow.
5. Share an example of handling a production failure under tight deadlines.
6. How do you collaborate with cross-functional teams (PMs, data scientists, analysts) to deliver business impact?

Ankita Gulati Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal