Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer III
- **Experience:** 2.5 years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹25+ LPA
- **Total Rounds:** 6
- **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
Preliminary Screening

1. Walk me through your past projects and experience with:
- Mixpanel
- Kafka
- ETL concepts
- Spark Lineage in Datahub
- Apache Spark

2. Explain the data model you developed for A/B testing experimentation on Presto architecture.

3. Why do you want to work at Walmart?

Ankita Gulati

Shubh Goyal

Round 2
Technical Interview I

DSA Questions

1. Find the minimum number of coins required to make a given change.
2. Given a linked list and a value x, partition the list such that:
   - Nodes < x come first.
   - Nodes >= x come after.
   - If x is present, it belongs in the right partition.

SQL Questions

3. Write a query to find the nth highest salary per department, with and without using window functions.
4. Using employee and department tables, write a query to find the highest salary in each department with DENSE_RANK().
   - Why would you choose DENSE_RANK over RANK?

Ankita Gulati Shubh Goyal

Big Data & Spark

5. Explain how Airflow on Kubernetes works with pod concepts.

6. How does the Airflow scheduler interact with workers and the webserver?

7. Difference between container deployment vs. stateful deployment in Kubernetes.

8. How does Kubernetes handle fault tolerance?

9. A Spark job is running slower than expected:

• How would you diagnose bottlenecks?

• What steps would you take to optimize performance?

10. With limited cluster resources, how would you allocate and configure Spark optimally?

Python/AWS

11. Write Python code using boto3 to upload a Parquet file to S3.

12. How does Airflow store logs in S3, and what role does its backend DB play?

Ankita Gulati

Shubh Goyal

Other Concepts

13. Explain SDLC phases and Agile (Scrum).
14. What's Walmart's DevOps/CI-CD approach?
15. Discuss NoSQL databases and common AWS services (scenarios).

Ankita Gulati

Shubh Goyal

Round 3
System Design + Big Data

System Design

1. Design a simplified version of the Mixpanel system (event-driven analytics).
- How do events flow from apps (Android, iOS, Web)?
- How does the load balancer manage requests?
- What happens when a user accesses a Presto URL (DNS resolution → LB → target gateway → Presto Coordinator)?

2. Write a custom API (service + controller classes) using Spring Boot.

Spark & Delta Lake

3. Write Spark code to read data from Delta Lake (S3) and implement upsert logic
4. Explain Spark optimizations:
- Skewed joins
- Broadcast joins
- Cost-based optimization (CBO)
- Repartition vs. Coalesce

Ankita Gulati Shubh Goyal

5. What are Tungsten and Catalyst Optimizer in Spark?

Java & OOPs

6. Java collections: Interfaces, Maps, LinkedLists.
7. Write code to trigger garbage collection using a GC thread.
8. Write code demonstrating synchronized multithreading.
9. Serialization vs. Deserialization.
10. Use case of the transient keyword.

Concurrency & Synchronization

11. What is a Semaphore variable?
12. How do you prevent deadlocks?
13. Complete Java code for a Semaphore implementation (synchronization).

Ankita Gulati

Shubh Goyal

Data Modeling / Warehousing

14. Compare Snowflake schema vs. Star schema.
15. How would you design a data warehouse from scratch for new requirements?
16. Explain Normalization and SCD Type 2 with examples.
17. How to onboard Delta Lake catalog into Presto?

Agile Practices

18. Why is Agile preferred over the waterfall model?

Ankita Gulati

Shubh Goyal

Round 4
Techno-Managerial

1. Explain your projects on Mixpanel, Databricks, PySpark, Datahub.
2. Difference between batch vs. streaming processing in Spark.
3. Walk through a pipeline you created in Databricks (silver → aggregated tables).
4. How did you contribute to open-source projects (e.g., Datahub, Spark Lineage with Spline)?
5. Examples of cloud cost optimization strategies:
   - How you reduced storage/compute costs.
   - Real-world impact of your initiatives.
6. How do you monitor Spark jobs and capture event logs (clusters, job execution in Databricks)?
7. How do you manage multiple tasks in Agile (Scrum, JIRA)?

Ankita Gulati

Shubh Goyal

Round 5
Director Round

Behavioral & Leadership

1. Introduce yourself and highlight key projects (Meesho, Morgan Stanley).
2. Tell me about a time you faced a challenging situation and how you handled it.
3. Share a leadership example where you resolved a conflict or drove a project.

Technical Deep Dive

4. Compare Presto vs. Spark architectures.
5. Can Presto handle near real-time streaming data?
6. What is the Avro file format, and how is it used in Delta tables?
7. How do you develop Datahub using open-source projects (Spline, Datahub)?
8. What's your view on data uncertainty and how to manage it?
9. What are the core values of Walmart and how do they resonate with you?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal