databricks

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ Years
- **Location:** Bangalore
- **Work mode:** Hybrid
- **Compensation:** ₹22-30 LPA
- **Total Rounds:** 5
- **Top Required Skills:**

1.Apache Spark&Scala
2. Apache Kafka
3. SQL
4. ETL Design & Data Modeling
5. DevOps Tools
6. Behavioral Skills

Ankita Gulati

Shubh Goyal

Round 1
Hiring Manager Interview

Walk me through your current and past projects.

- What business problem were you solving?
- Which tools and technologies did you use?
- What role did you play in the team?
-

Which orchestration tools have you worked with (e.g., Airflow, Oozie, Azkaban)?

- How did they improve efficiency?
- How do you handle task failures or retries?

Describe your experience with CI/CD pipelines in data projects.

- What tools (Jenkins, GitHub Actions, GitLab CI) have you used?
- How do you ensure faster rollouts with minimal downtime?

Explain your usage of Jenkins and Docker.

- How did Jenkins automate workflows?
- Have you containerized Spark jobs with Docker?

Why do you want to join Databricks?

- What excites you about its platform?
- If you could improve Databricks, what feature would you suggest?

Ankita Gulati Shubh Goyal

Round 2
Home Assessment/Technical Round

Option 1: Home Assessment

- Solve SparkScala coding challenges (file reads, transformations, aggregations).
- Write SQL queries covering joins, aggregations, and window functions.

Option 2: Live Technical Round

1.Scala& Spark Coding:
   - Read a text file into Spark.
   - Demonstrate usage of high-order functions (map, flatMap, reduceByKey).

2. Kafka:
   - Explain Kafka architecture (brokers, topics, partitions, producers, consumers).
   - Difference between a topic and a partition.
   - What is Kafka Mirror Maker, and why is it used?

3.Spark Optimizations:
   - How do you reduce shuffle operations?
   - Compare Parquet vs CSV vs ORC formats. Which format is best for analytics and why?
   - How does Spark decide the number of partitions?

Round 3
Technical Data Design

1. Design an ETL pipeline for ingesting raw transactional data into a warehouse.
2. Which steps would you include (ingestion, staging, transformation, loading)?
3. What scheduling/orchestration would you use?
4. Which database would you choose for the pipeline (RDBMS, NoSQL, Data Lake, Delta Lake)? Why?
5. How would you decide partition keys in your tables?
6. Example: Partition by date or region? What's the trade-off?
7. How do you ensure data quality?
8. Write validation checks (null values, duplicates, referential integrity).
9. How do you handle schema evolution?
10. Explain your approach to GDPR compliance.
11. How would you anonymize or delete user data on request?
12. How do you manage data retention policies?
13. How would you monitor and report pipeline performance?
14. What metrics would you track (latency, throughput, error rates)?
15. Which monitoring tools have you used (Prometheus, Datadog, CloudWatch)?

Ankita Gulati

Shubh Goyal

Round 4
Technical Interview

Spark Scala Task:

1. YouhavemultipleCSV files with file names like ankita_gulati.csv.
2. Requirement: Merge them into one DataFrame.
3. Add a new column containing the formatted file name (e.g., "Ankita Gulati") for all rows from that file.
4. Explain how you would handle inconsistent file naming.

Kafka Processing:

1. How do you ensure sequential message processing in Kafka?
2. What happens when multiple consumers read from a single partition?
3. How do you guarantee ordering when consuming from multiple partitions?

Ankita Gulati Shubh Goyal

Round 5

Leadership Interview(Senior Director)

- What new skills have you self-learned in data engineering?
  → Give examples (e.g., learning Spark Structured Streaming on your own).
- How do you stay updated with the latest technology?
  → Do you follow blogs, courses, conferences, or communities?
- Share an impactful project you worked on.
  → What was your role?
  → How did it benefit the business?
- Why are you a good fit for Databricks? ecosystem.
  → Align your skills with Databricks' product

- How do you handle team conflicts?
  → Share a real example where you resolved disagreements constructively.

Thank You

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

Ankita Gulati

Shubh Goyal