HCLTech

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 4+ years
- **Location:** Pune
- **Work mode:** Hybrid
- **Compensation:** ₹20-22 LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark / Python / Databricks
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming
6. System Design

Round 1
Technical Discussion

1. Can you give a brief introduction about yourself and your experience as a Data Engineer?
2. Walk me through a recent data engineering project you worked on, including the challenges you faced and how you solved them.
3. How do you optimize a data pipeline for performance, scalability, and fault tolerance?
4. What is the difference between transformations and actions in Spark? Provide an example.
5. What are narrow and wide transformations in Spark? How do they impact performance?
6. How does Spark handle data shuffling, and what are some ways to minimize it?
7. What are the different persistence/storage levels in Spark, and when would you use them?
8. Explain the difference between repartition and coalesce in Spark with use cases.

9. What are the key differences between map and flatMap in PySpark? Provide an example.
10. How would you remove duplicate records in PySpark?
11. How do you detect and handle skewed data in PySpark?
12. Write a SQL query to find the second highest salary from an employee table.
13. What is the difference between RANK(), DENSE_RANK(), and ROW_NUMBER() in SQL? Provide examples.
14. What is the difference between internal and external tables in Hive?
15. Explain partitioning vs bucketing in Hive with an example.
16. How does HDFS store large datasets, and what are the advantages of using it?
17. Explain the roles of NameNode, DataNode, and Secondary NameNode in HDFS.

18. Explain the difference between batch processing and stream processing with examples.

19. How do you handle nested JSON data in PySpark?

20. What are some common ETL tools you’ve worked with, and how do you decide which one to use in a project?

Round 2
Advanced Technical Discussion

1. What are the different ways to load data into Amazon Redshift?
2. How does Redshift handle distribution styles (KEY, EVEN, ALL), and how do you choose one?
3. Explain how AWS Glue works and where it fits into a data pipeline.
4. How would you build a data pipeline in AWS using S3, Lambda, and Glue?
5. Compare EMR vs Glue for running Spark jobs. When would you prefer one over the other?
6. What is an efficient way to process large-scale nested JSON in AWS?
7. Suppose your Spark job is taking longer than expected, what steps would you take to debug and optimize it?
8. Have you worked on a data migration project (on-prem to cloud)? What challenges did you face, and how did you solve them?

Ankita Gulati Shubh Goyal

9. How do you store data using an AWS Lambda function? Which file formats are best for API data?
10. How would you build a beginner-friendly recommendation system using customer viewing data?
11. How would you monitor and handle backpressure in a streaming pipeline (e.g., Kafka + Spark Streaming)?
12. Explain schema evolution and MERGE operations in Delta Lake (Databricks).
13. How would you design a data model for a large-scale e-commerce system with millions of users and transactions?
14. How do you handle data governance, access control, and auditing in cloud data platforms?
15. How do you ensure data quality in an automated data pipeline?
16. SQL: Write a query to get the total revenue per customer for the last 30 days from an orders table.

Round 3
HR Discussion

1. Can you walk me through your resume and highlight key projects relevant to this role?
2. Why are you interested in joining HCL as a Senior Data Engineer?
3. What are your short-term and long-term career goals?
4. How do you handle tight deadlines and conflicting priorities?
5. Can you share an example of a time when you had to collaborate with cross-functional teams (data scientists, business analysts, DevOps)?
6. What are your salary expectations, notice period, and preferred location?

Thank You

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

Ankita Gulati

Shubh Goyal