amdocs

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 4+ years
- **Location:** Pune/ Gurgaon
- **Work mode:** Hybrid
- **Compensation:** ₹12–18LPA
- **Total Rounds:** 4
- **Top Required Skills:**

1. SQL Query Optimization
2. Python for ETL Development
3. Data Structures & Algorithms
4. Big Data Technologies (Hadoop, Hive, Kafka, Spark) Cloud
5. Data Warehousing & (Redshift, S3, Airflow)
6. Distributed Systems & Architecture
7. Behavioral & Team Collaboration

Ankita Gulati

Shubh Goyal

Round 1
Online Assessment

SQL Query Optimization

1.Youare givena query that fetches data from very large tables. How would you optimize the query to reduce execution time? Discuss the use of indexes, selecting only required columns instead of SELECT *, writing efficient JOINs, and considering multi-column indexes for faster filtering.

Python Scripting for ETL

2.Writea Python script that can perform an ETL process on JSON files. Specifically, you need to extract data from JSON files, transform them into a structured format, remove null values, and then load the transformed data into CSV format. Explain how you would ensure efficiency when handling very large datasets.

Data Structures & Algorithms

3. Given a list of integers, write an algorithm to identify duplicate elements and return the top N most frequent duplicates. Explain how you would use a hashmap (dictionary) to count frequencies and then sort the results by frequency.

Round 2
Technical Interview

Real-Time Data Processing

1.Howwould you design adata pipeline for real-time data processing? Provide a detailed answer that covers using Apache Kafka for ingesting streaming data, Spark Streaming for real-time transformations, and Spark SQL for performing aggregations. Also explain how you would ensure fault tolerance by using checkpointing and Kafka's replication mechanism.

ETL Pipeline with Hadoop

2. Suppose you need to build an ETL pipeline using Hadoop. How would you design this pipeline? Explain how HDFS can be used as the storage layer, Hive for managing and querying the datasets, and how partitioning Hive tables (e.g., date-based partitioning) can optimize queries for faster retrieval.

Data Warehousing

3. How would you design a scalable data warehouse that integrates data from multiple sources? Provide details on how you would use Amazon Redshift for warehousing, S3 buckets for storing raw data, and Airflow for orchestrating ETL jobs. Explain why you would transform and load only the necessary data into Redshift to ensure both cost-efficiency and performance.

Round 3
System Design & Architecture

Billing System Data Flow

1.Imagine you are designingAmdocs' billing system that handles millions of transactions per day. How would you architect the data flow? Discuss how Apache Kafka could be used to ingest real-time transaction events, how microservices could validate and enrich these events, and how you would persist data into Cassandra for its high write throughput and multi-datacenter replication.

Data Integrity & Consistency

2. How would you ensure data integrity and consistency across multiple distributed services? Explain how Kafka's exactly-once semantics and idempotency features prevent duplicate processing. Discuss eventual consistency in distributed systems and how the Two-Phase Commit protocol could be used to coordinate updates across services.

Data Security

3. How would you secure sensitive billing data flowing through this system? Explain how you would implement encryption both at rest and in transit, and how you would use AWS Key Management Service (KMS) for managing encryption keys securely.

Round 4
Behavioral & Managerial

Problem Solving in Production

1.Tell us about a time when youencountered a large-scale data issue in a production pipeline. What steps did you take to investigate the problem, what root cause did you identify, and how did you resolve it?

Cross-Functional Collaboration

2. Describe your experience working with cross-functional teams such as DevOps, product management, and QA. How did you manage dependencies across teams, and how did tools like JIRA or sprint planning help in aligning deliverables?

Team & Learning Approach

3.What is your approachtohandling critical issues during peak workloads while maintaining team communication? How do you balance short-term fixes with long-term stability?

4. How do you approach continuous learning and staying updated with evolving data engineering technologies, especially in a fast-moving industry like telecom?

Thank You

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

Ankita Gulati

Shubh Goyal