Apple

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 5+ years
- **Location:** Hyderabad
- **Work mode:** Hybrid
- **Compensation:** ₹27–40LPA
- **Total Rounds:** 4
- **Top Required Skills:**

1. Advanced SQL (joins, aggregations, window functions, optimization)
2. Python for data cleaning, ETL, and transformations
3. ETL and pipeline design (batch and streaming)
4. Big Data (Spark, Kafka, Snowflake)
5. Cloud data solutions

Ankita Gulati

Shubh Goyal

Round 1
Recruiter Screening

Focus Areas: Resume walkthrough, prior projects, role alignment, cultural fit.

Questions:

1. Can you walk me through your resume, highlighting the projects where you directly worked on data pipelines, infrastructure, or performance optimizations?
2. Please describe the most impactful data engineering project you have worked on. What was the business problem, what was your role, and what measurable impact did it create?
3. How do your skills and past experiences align with Apple's data engineering culture, which emphasizes scalability, performance, and simplicity?
4. Why do you want to work as a Data Engineer at Apple specifically, and how do you see yourself contributing to the company's mission?

Round 2
Technical Phone Screen

1. Write an SQL query to return the top 5 most-sold Apple products in the last 30 days. How would you ensure the query performs efficiently on a dataset with billions of records?
2. Suppose you are tasked with fetching sales data across multiple countries from fact and dimension tables. How would you optimize a query that involves multiple joins, aggregations, and filters? Please explain the use of indexes, partitions, and query plan analysis in your solution.
3. You are asked to design a simple pipeline to process daily sales logs and load the results into a reporting database. Walk me through your approach: which ingestion method would you choose, how would you transform the data, and what storage/query solution would you use?
4. When analyzing SQL queries, how do you read and interpret the execution plan? Can you describe common mistakes or anti-patterns that degrade performance, such as unnecessary subqueries or excessive sorting?

Ankita Gulati Shubh Goyal

Round 3
Onsite Interviews

SQL & Python Coding

1. Write an SQL query to find the top 3 customers by total purchase amount in each region during the last quarter. Please explain how you would handle ties in rankings and ensure the solution is scalable on very large datasets.
2. Given a dataset of daily sales transactions, write a Python function to detect anomalies in sales trends, such as sudden spikes or sharp drops. How would you use libraries like Pandas or NumPy to compute rolling averages or standard deviations for anomaly detection?
3. You are provided with a dataset containing null values, duplicate rows, inconsistent date formats, and timezone issues. How would you clean and standardize this dataset using Python or PySpark before loading it into a data warehouse?
4. Imagine you are working with a SQL query that includes multiple joins and aggregations on a table with billions of rows. How would you identify performance bottlenecks and optimize the query? What changes would you recommend in terms of partitioning, indexing, or rewriting the query logic?

System Design (ETL/Data Pipeline)

1. Design an end-to-end pipeline to collect, process, and analyze customer feedback data from Apple retail stores across the globe. Which technologies would you choose for ingestion, processing, storage, and querying, and why?
2. Suppose Apple wants to process millions of real-time events per second from its devices worldwide. How would you design a streaming pipeline using Kafka and Spark Streaming to handle ingestion, transformation, and storage? Please explain how you would address issues like scaling, monitoring, and fault tolerance.
3. What are the key trade-offs between batch and real-time pipelines? Can you provide an example of a use case where batch processing is more suitable, and another where streaming is essential?
4. Pipelines often fail due to transient issues like network outages or data corruption. How would you implement monitoring, retries, and idempotency mechanisms to ensure data reliability in your ETL pipelines?

Big Data & Cloud Technologies

1. Compare Parquet, Avro, and ORC file formats. For Apple’s analytical workloads, which format would you recommend and why? Please explain differences in compression, schema evolution, and query performance.
2. In Spark, transformations are evaluated lazily. Can you explain the concept of lazy evaluation and the difference between narrow and wide transformations? How does this affect shuffle operations and overall job performance?
3. Apple handles event data through messaging systems. Explain the delivery guarantees in Kafka: at-least-once, at-most-once, and exactly-once. How would you design a pipeline that ensures exactly-once delivery of messages?
4. Snowflake is often used as a cloud data warehouse. How does schema evolution work in Snowflake? Suppose Apple wants to track evolving attributes in user activity logs — how would you design the schema to accommodate new fields without breaking existing queries?

Behavioral Interview

1. Tell me about a time when you had to explain a complex technical issue to a non-technical stakeholder. How did you ensure they understood the core problem and its implications?
2. Describe a project where you optimized an existing data pipeline. What specific changes did you make (e.g., partitioning, memory tuning, algorithm redesign), and what measurable improvements did you observe?
3. You are assigned to multiple high-priority projects with tight deadlines. How would you manage competing priorities and ensure timely delivery without sacrificing quality?
4. Share an example of a failure in a past project. What was the root cause, how did you recover, and what lessons did you take forward to prevent similar issues?

Round 4
Final Behavioral & Cultural Fit

1. Why do you want to join Apple as a Data Engineer, and how do you see yourself growing within the company over the next five years?
2. What values or aspects of Apple's culture resonate most with you, and how do they align with your way of working?
3. How do you stay updated with rapidly evolving data engineering tools and frameworks? Please provide specific examples (e.g., courses, hands-on projects, blogs, or open-source contributions).
4. Collaboration is essential at Apple. Can you describe a situation where you ensured teamwork and ownership in a large-scale project?

Thank You

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

Ankita Gulati

Shubh Goyal