American Express Data Engineer Interview Guide – Experienced 2+

Round 1: Coding (SQL & Python) – 45 Minutes

This round focused on evaluating coding and optimization skills in SQL and Python.

Key Topics Covered

- **SQL**:
  - Window Functions
  - Recursive Queries
  - Joins
  - Query Optimization Techniques
- **Python**:
  - Data Structures: List, Set, Tuple, Dictionary, String
  - Decorators, Multiprocessing, and File Handling

Example Questions

1. Write a query to find the first number repeating consecutively three times in a sequence.
2. Explain and implement a recursive query to find a hierarchical structure (e.g., an employee-manager relationship).
3. Given a table of sales data, use window functions to calculate a running total.
4. How would you optimize a query with multiple joins and subqueries?
5. Implement a Python function to count unique words from a file and write them to another file.
6. Explain the differences between multiprocessing and multithreading.
7. Write a decorator function to log the execution time of a function.
8. Create a Python program to demonstrate the use of set operations (union, intersection).
9. Implement file handling in Python to read a CSV and store only specific columns in a dictionary.
10. Explain the difference between mutable and immutable objects in Python.

Round 2: Data Engineering Concepts (Big Data, PySpark, Databricks) – 1 Hour

This technical round tested the depth of PySpark knowledge, Databricks usage, and optimization techniques.

Key Topics Covered

- Spark Optimization
- PySpark Coding
- Databricks Architecture and Usage

Example Questions

1. What are broadcast variables in Spark? How do they improve performance?
2. Explain repartition vs. coalesce. Which one would you use to reduce shuffle operations?
3. What is the salting technique, and when would you use it?
4. Explain bloom filters in Spark and how they optimize join operations.
5. What are the differences between SparkContext and SparkSession?
6. Describe Spark's memory management model. How do you handle heap memory overhead issues?
7. Code a simple PySpark job to read a JSON file, filter records, and write output in Parquet format.
8. How does Spark's Catalyst Optimizer improve query performance?
9. Describe a scenario where you used Databricks for real-time data processing.
10. Explain a scenario-based question on Spark optimization and how you would troubleshoot performance issues.

Round 3: Techno-Managerial

The focus was on past projects, architecture design, and handling real-world challenges.

Example Questions

1. Describe the architecture of an ETL pipeline you built in your previous project.
2. How did you handle data ingestion and processing for large datasets?
3. What were the biggest infrastructure-level challenges you faced, and how did you resolve them?
4. Discuss the data size challenges in your previous projects. How did you optimize storage and processing?
5. How do you ensure data quality and consistency in your pipelines?
6. How do you handle schema evolution in data lakes or data warehouses?
7. Explain a situation where you had to coordinate with multiple teams to complete a project. How did you manage time and priorities?
8. What monitoring and logging strategies did you implement for your pipelines?
9. Why do you want to join American Express?
10. What are your strengths, and how do they align with the Data Engineer role?

Glassdoor American Express Review –

https://www.glassdoor.co.in/Reviews/American-Express-Reviews-E35.htm

American Express Careers –

https://www.americanexpress.com/en-us/careers/

Subscribe to my YouTube Channel for Free Data Engineering Content –

https://www.youtube.com/@shubhamwadekar27

Connect with me here –

https://bento.me/shubhamwadekar

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar