Carelon Global Solutions AWS Data Engineer Interview Guide – Experienced 3+

Interview Rounds Overview

1. **Round 1 – Technical**
    - Focused on PySpark, SQL, and AWS S3 fundamentals and operations.
2. **Round 2 – Techno-Managerial**
    - Covered Spark, Snowflake, AWS Services, and end-to-end data pipeline architecture.
3. **Round 3 – HR Discussion**
    - Focused on cultural fit, company policies, and salary discussions.

Detailed Breakdown of Rounds

Round 1 – Technical

Key Topics & Questions:

1. **PySpark and AWS S3**:
    - Write PySpark code to save a DataFrame in Parquet format to an S3 bucket.
    - Explain how to overwrite a file stored in S3 using PySpark.

```
from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder \
    .appName("SaveToS3") \
    .getOrCreate()

# Sample DataFrame
data = [(1, "Alice"), (2, "Bob"), (3, "Carol")]
df = spark.createDataFrame(data, ["id", "name"])

# Save DataFrame to S3
df.write.mode("overwrite").parquet("s3://your-bucket-name/folder/")
```

2. **AWS S3 Versioning**:
    - Discuss how versioning works in S3 and its use cases, such as data recovery and auditing.
3. **SQL Query Challenge**:
    - Write a query to generate the specified output. This tested advanced SQL skills with joins, aggregations, and window functions.
4. **Running PySpark on AWS EC2**:
    - Steps to execute a Python file with PySpark code on an EC2 environment, emphasizing environment setup, JAR dependencies, and S3 integration.
5. **File Copying to S3**:
    - Methods to copy files to S3 without using the bucket upload feature, such as using the AWS CLI or Python's boto3 library.
6. **Snowflake vs. Spark**:
    - Time and cost comparisons for executing the same query in Snowflake and Spark.
    - Discussed cost optimization techniques for Snowflake, such as utilizing clustering keys and data compression while loading data from S3.

Round 2 – Techno-Managerial

This round was conducted in two phases and focused on assessing my expertise with AWS data engineering services and Spark optimization techniques.

Key Questions:

1. **Spark Memory Distribution**:
    - Explain how Spark processes a 500GB file, covering memory allocation, shuffles, and spillovers to disk.
2. **Architecture-Level Scenarios**:
    - Designing an end-to-end data pipeline using Glue, Lambda, EC2, S3, Redshift, and Athena.
3. **PySpark Program**:
    - Write a complete PySpark program from import statements to the stop statement, covering transformations and actions.
4. **SQL Challenges**:
    - Tested SQL skills using advanced window functions such as LAG, LEAD, and DENSE_RANK.
    - Practical examples focused on ranking and trend analysis.
5. **Spark Optimization**:
    - Discuss techniques such as partitioning, broadcast joins, and caching to enhance Spark job performance.

Round 3 – HR Discussion

Topics Covered:

1. **Team Culture**:
    - Discussion around the company's values, team dynamics, and expectations.
2. **Work Policies**:
    - Details on leave policies, holidays, and work flexibility (e.g., hybrid vs. remote).
3. **Compensation**:
    - Negotiated salary, including fixed and variable components.
4. **Work Culture**:
    - Insight into the company's work-life balance and career growth opportunities.

Glassdoor Carelon Global Solutions Review –

https://www.glassdoor.co.in/Reviews/Carelon-Global-Solutions-Reviews-E2036118.htm

Carelon Global Solutions Careers –

https://www.carelonglobal.in/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