**Daniel Wellington AWS Data Engineer Interview Guide – Experienced 5+**

**Round 1: Soft Skills (Recruiter Assessment)**

**Key Topics:**

 Background discussion, highlighting skills and expertise relevant to the Senior Data

Engineer role.

 Questions focused on technical and non-technical challenges faced in previous projects.

 Insights into career goals and alignment with the company’s vision.

**Round 2: Take-Home Assignment (Data Modeling Case Study)**

**Duration:** 5 Working Days

**Task:**

 Design a working data pipeline to efficiently store, process, and report data.

**Key Requirements:**

 Data Schema Selection: Chose between star schema, snowflake schema, or a normalized design.

 Git Integration: Version control of the codebase.

 CDC Implementation: Efficient handling of data changes.

 Workflow Orchestration: Tools like Apache Airflow or equivalent.

 Data Warehouse: Store BI-ready tables for analysis.

 Analytics & SQL: Create approximately 10 SQL-based reports answering provided case study questions.

Dataset: Provided in multiple Excel sheets.

**Round 3: Technical Interview + Home Assignment Discussion**

**Duration:** 1.5 Hours (50% Case Study, 50% Technical Q&A)

**Case Study Discussion:**

 **Schema Design:**

 Why star schema? Compared with snowflake schema and normalized approaches.

 Feedback on fact and dimension table design (additional optimizations suggested).

 **ETL Design Choices:**

 Tools like Apache Airflow were discussed.

 Why a batch process over real-time? Demonstrated an understanding of both approaches.

 **Cloud Integration:**

How to adapt the same pipeline to a cloud environment?

**SQL Q&A:**

 Solved 3 medium-level problems involving window functions and joins.

**Round 4: Technical Interview (Hiring Manager + Team Lead)**

**Duration:** 1 to 1.5 Hours

**Spark-Related Questions:**

 Write a Spark job to count word occurrences from an S3 dataset.

 Discuss stages and tasks in a Spark execution plan.

 Explain Spark's fault tolerance mechanisms.

 Persistence Storage Levels: When to use MEMORY\_ONLY, MEMORY\_AND\_DISK, etc.

 Partitioning & Bucketing: Impact on PySpark performance.

**AWS Redshift & Spectrum:**

 What is Redshift Spectrum, and how does it differ from standard Redshift queries?

**Apache Kafka:**

 1 basic Kafka question (details not recalled).

**Apache Airflow DAGs:**

 Explanation of Directed Acyclic Graphs (DAGs) in Airflow.

**AWS Pipeline-Specific Questions:**

 Data Security: How to manage AWS IAM roles and policies.

 Glue ETL Optimization: Performance improvement strategies.

 Securing AWS Lambda: IAM roles, VPC integration, and security measures.

**Current Projects Discussion:**

 Optimization strategies in current projects.

 Cross-questioning on AWS services like S3, Glue, Lambda, and Step Functions.

**Key Takeaways:**

1. Be thorough in schema design: Always validate your approach and explore alternatives.

2. Spark Fundamentals: Cover concepts like persistence, fault tolerance, and optimization techniques.

3. AWS Services: Deep-dive into Glue, Redshift, S3, Lambda, and IAM roles for data security.

4. SQL Mastery: Practice complex queries involving joins, window functions, and performance tuning.

5. Hands-On Projects: Be ready to discuss real-world experiences, challenges, and your technical decisions.