

Lumiq.ai Data Engineer Interview Guide – Experienced 3+

Round 1: Project Explanation and Technical Discussion

Duration: 60 minutes

Project Explanation

The first section of the round involved a detailed walkthrough of my recent project.

1. Cloud-Based Data Pipeline on AWS

I explained how I built a data pipeline on AWS for ingesting large volumes of transactional data. The solution integrated AWS S3, Lambda functions, and EMR for batch processing. Data was processed using PySpark, partitioned, and stored in Parquet format to optimize querying. Glue Data Catalog was used for metadata management, and Athena for running queries directly on S3-stored data.

2. Scalable Data Models

Designing scalable models involved leveraging dimensional modeling for analytics. Fact tables captured daily transactions with dimensions for customers, products, and regions. The design minimized redundancy, optimized lookups, and supported time-series analyses.

3. Data Storage and Retrieval Optimization

Techniques included partitioning by time for S3 storage, applying predicate pushdown to reduce scanned data, and using broadcast joins in Spark for small reference tables. Query performance improved through compression (Parquet/Gzip) and reducing I/O overhead.

4. Data Security in BFSI

- Encryption: Implemented SSE-S3 and SSE-KMS for S3 buckets.
- IAM Policies: Restricted access to only necessary users using least privilege principles.
- Auditing: Enabled CloudTrail to track API calls.
- Compliance: Followed GDPR and PCI-DSS standards, ensuring PII data was masked or tokenized as needed.

5. Data-Related Issues Encountered

One significant challenge was handling skewed data in Spark jobs. The solution involved salting keys to distribute data more evenly and using repartitioning strategies to balance load.

6. Data Lake vs. Data Warehouse

- Data Lake: Stores raw, semi-structured, or unstructured data for flexible exploration.
- Data Warehouse: Designed for structured, curated data with schema enforcement for analytical queries.

7. Query Performance in Redshift

Optimization was achieved using:

- Distribution keys to avoid data shuffling.
- Sort keys for range-based queries.
- Vacuuming to reclaim space and improve performance.

8. Resolving Pipeline Failures

A major latency bottleneck was resolved by implementing buffering in Kafka and batch writes to downstream systems. Monitoring with AWS CloudWatch and Datadog provided alerts for lag or throughput issues.

9. Monitoring Tools for Pipelines

- Airflow DAGs were tracked via its UI and logs.
- AWS CloudWatch for EMR and S3 event tracking.
- Prometheus with Grafana for metrics visualization.

Round 2: Big Data and Cloud Technical Round

Duration: 75 minutes

This round tested knowledge across Big Data ecosystems, streaming data, and practical implementation questions on SQL, Kafka, and Data Modeling.

Big Data and Cloud Technical Questions

1. Spark Optimization

- Techniques like broadcast joins, caching, coalescing, and predicate pushdown were discussed.
- Explained the role of Adaptive Query Execution (AQE) in dynamic query plan optimization.

2. Spark Coding: Using explode() Function

```
from pyspark.sql.functions import explode, col

df = spark.createDataFrame([
    (1, ["a", "b", "c"]),
    (2, ["d", "e"])
], ["id", "values"])

result = df.select(col("id"), explode(col("values")).alias("value"))
result.show()
```

This task demonstrated my ability to **flatten nested arrays** using PySpark.

3. SQL Problem

Solved a query involving multiple table joins and window functions to compute ranked results without using suboptimal nested queries.

4. Kafka Basics

- Explained Kafka architecture, highlighting topics, partitions, producers, and consumers.
- Described the role of Zookeeper and strategies to handle offset management in consumer groups.

5. Data Modeling and Airflow Scheduling

- For data modeling, I discussed a star schema design and why it's preferred for analytics.
- Airflow scheduling required explaining cron expressions and backfill handling. I also elaborated on monitoring task failures with custom alerting.

Glassdoor Lumiq.ai Review –

<https://www.glassdoor.co.in/Reviews/Accenture-Reviews-E4138.htm>

Lumiq.ai Careers –

<https://lumiq.ai/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

© Shubham Wadekar