AWS Glue Concepts & Interview Q&A; — Illustrated Guide

A Light■Theme PDF covering architecture, hands■on examples, best practices, and 25+ interview questions for data engineering roles.

■ AWS Glue Architecture Overview

AWS Glue is a serverless ETL service for discovering, preparing, and integrating data from multiple sources. It automates schema inference, job orchestration, and metadata management.

• Data Catalog — Central metadata repository for databases, tables, schemas, partitions, and job definitions.

• Crawlers — Scan and infer schema from data sources; populate the Data Catalog.

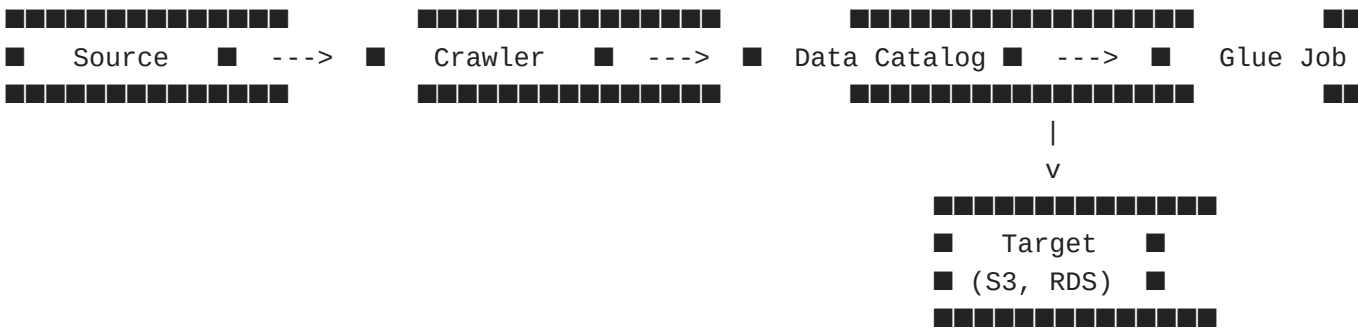
• Jobs — ETL scripts (Python or Scala) that extract, transform, and load data using Apache Spark.

• Triggers & Workflows — Automate job execution on schedule or events and chain multiple jobs together.

• Serverless Execution — AWS Glue provisions compute (DPUs) automatically and scales for workload size.

• Integration — Works seamlessly with S3, Redshift, RDS, Athena, Kinesis, and Lake Formation.

■■ Simplified Glue ETL Flow



■ Hands-On Glue Examples (PySpark)

Example: Initialize GlueContext, read from S3, transform using DynamicFrame, and write to target.

```
from awsglue.context import GlueContext
from awsglue.job import Job
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

# Read from Data Catalog
datasource = glueContext.create_dynamic_frame.from_catalog(
    database="sales_db",
    table_name="transactions_raw"
)

# Apply Transformations
mapped = datasource.apply_mapping([
    ("cust_id", "string", "customer_id", "string"),
    ("amount", "double", "amount_usd", "double")
])

# Write to S3 in Parquet format
glueContext.write_dynamic_frame.from_options(
    frame=mapped,
    connection_type="s3",
    connection_options={"path": "s3://analytics-zone/sales/"},
    format="parquet"
)
```

■ Incremental Load with Job Bookmarking

```
glueContext.create_dynamic_frame.from_catalog(
    database="sales_db",
    table_name="transactions_raw",
    transformation_ctx="datasource",
    additional_options={"jobBookmarkKeys": ["transaction_id"], "jobBookmarkKeysSortOrder":
)
```

■ AWS Glue Interview Q&A;

Q: What is AWS Glue?

A: A fully managed, serverless ETL service for discovering, preparing, and integrating data across AWS services.

Q: What is the Glue Data Catalog?

A: A persistent metadata store for databases, tables, and job definitions used by Glue, Athena, and Redshift Spectrum.

Q: What is the role of Crawlers?

A: Crawlers connect to data sources, infer schema using classifiers, and populate the Data Catalog automatically.

Q: Difference between DynamicFrame and DataFrame?

A: DynamicFrame is Glue's abstraction with schema flexibility and built-in transformation methods; DataFrame is Spark's native structure.

Q: What are Glue Triggers?

A: Mechanisms to run jobs on schedule, on-demand, or based on other job events.

Q: Explain Glue Workflows.

A: A workflow is a collection of jobs and triggers that manage complex ETL dependencies and execution order.

Q: What is AWS Glue Studio?

A: A visual interface for building, running, and monitoring ETL jobs without writing code manually.

Q: What are Classifiers?

A: Components used by Crawlers to identify file formats and infer schema for CSV, JSON, Parquet, etc.

Q: How does Glue handle schema evolution?

A: Glue can detect schema changes through Crawlers and update the Data Catalog automatically.

Q: Explain Job Bookmarking.

A: A mechanism that tracks previously processed data to ensure only new data is processed in subsequent runs.

Q: How can Glue integrate with Athena?

A: Glue Data Catalog acts as Athena's metastore, allowing direct querying of Glue tables.

Q: Explain Glue Streaming Jobs.

A: Jobs using Spark Structured Streaming for continuous ingestion from Kinesis or Kafka.

Q: What are common Glue job failure causes?

A: Incorrect IAM permissions, schema mismatch, memory/DPU limits, or transformation errors.

Q: Best practices for Glue security?

A: Use IAM roles with least privilege, enable encryption (KMS), and run inside VPC for private data sources.

Q: Glue vs EMR vs Data Pipeline?

A: Glue is serverless ETL, EMR is managed Hadoop/Spark cluster, and Data Pipeline is an orchestration service.

■ AWS Glue vs EMR vs Data Pipeline

Service	Type	Use Case	Management Level
AWS Glue	Serverless ETL	Automated data integration and transformation	Fully managed
Amazon EMR	Managed Cluster	Custom Spark/Hadoop jobs with fine-grained control	Semi-managed
Data Pipeline	Workflow Orchestration	Scheduling and dependency management	Managed control-plane only