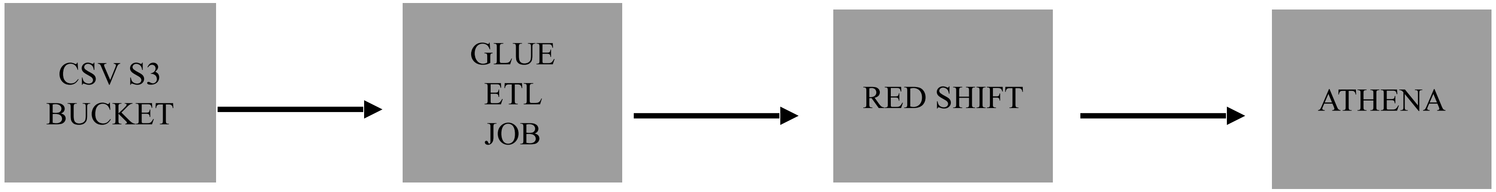
AWS ASSESSMENT

POLIREDDI GOVIND AS1552

Problem Statement:

Build a Batch Data Pipeline to Ingest CSV Data from S3 → Process with Glue → Load to Redshift → Query via Athena



STEP 1: SERVICE CREATION + IAM ROLES

- Create **three S3** buckets (**Source CSEV storage, Failed CSV storage, Processed CSV storage**) to store the CSV files.
- Create a **GlueRole** in IAM Roles with **AWSGlueServiceRole** and **AmazonS3FullAccess policies.**
- Create a **RedshiftRole** in IAM Roles with **AmazonS3ReadOnlyAccess** and **AWSGlueConsoleFullAccess** policies.
- Create a new Redshift serverless **namespace** and **workgroup**, attach the **RedshiftRole** to grant the necessary permissions to access the **S3** and **AWS Glue Data Catalog.**
- Create a **Glue Crawler** and add the **S3 bucket** as data source, attach **GlueRole** policy to grant the required permissions to access the data.

STEP 2: AUTOMATION WITH LAMBDA + STEP FUNCTIONS

- When ever a new CSV file is added to the **S3** bucket a **LAMBDA** function will get triggered.

- ➢ The **LAMBDA** function starts the **AWS Step Functions** workflow
- ➢ Step Functions **orchestrate** the pipeline:
  - Start Glue Job and Monitor job completion.
  - Trigger next steps automatically.
- ➢ Failed files will be directed to **Failed_CSV_S3 storage** and then by using **AWS SNS** a notification will be sent.

STEP 3: CRAWLER + DATA PROCESSING

- ➢ The raw CSV data stored in the Amazon S3 bucket will be crawled then metadata and schema will be stored in **GLUE CATALOG** and then ingested to an **AWS Glue ETL job.**
- ➢ The Glue job will extract, clean, and transform the data, then convert it into an **Parquet** format.
- ➢ The processed Parquet files will be stored in a **separate Amazon S3 bucket.**

STEP 4: LOADING PARQUT FILES INTO AMAZON REDSHIFT

- ➢ The processed Parquet files from the new Amazon S3 bucket are loaded into Amazon Redshift with(COPY/Auto COPY from processed Parquet).
- ➢ Use IAM role to allow Redshift to read processed bucket.

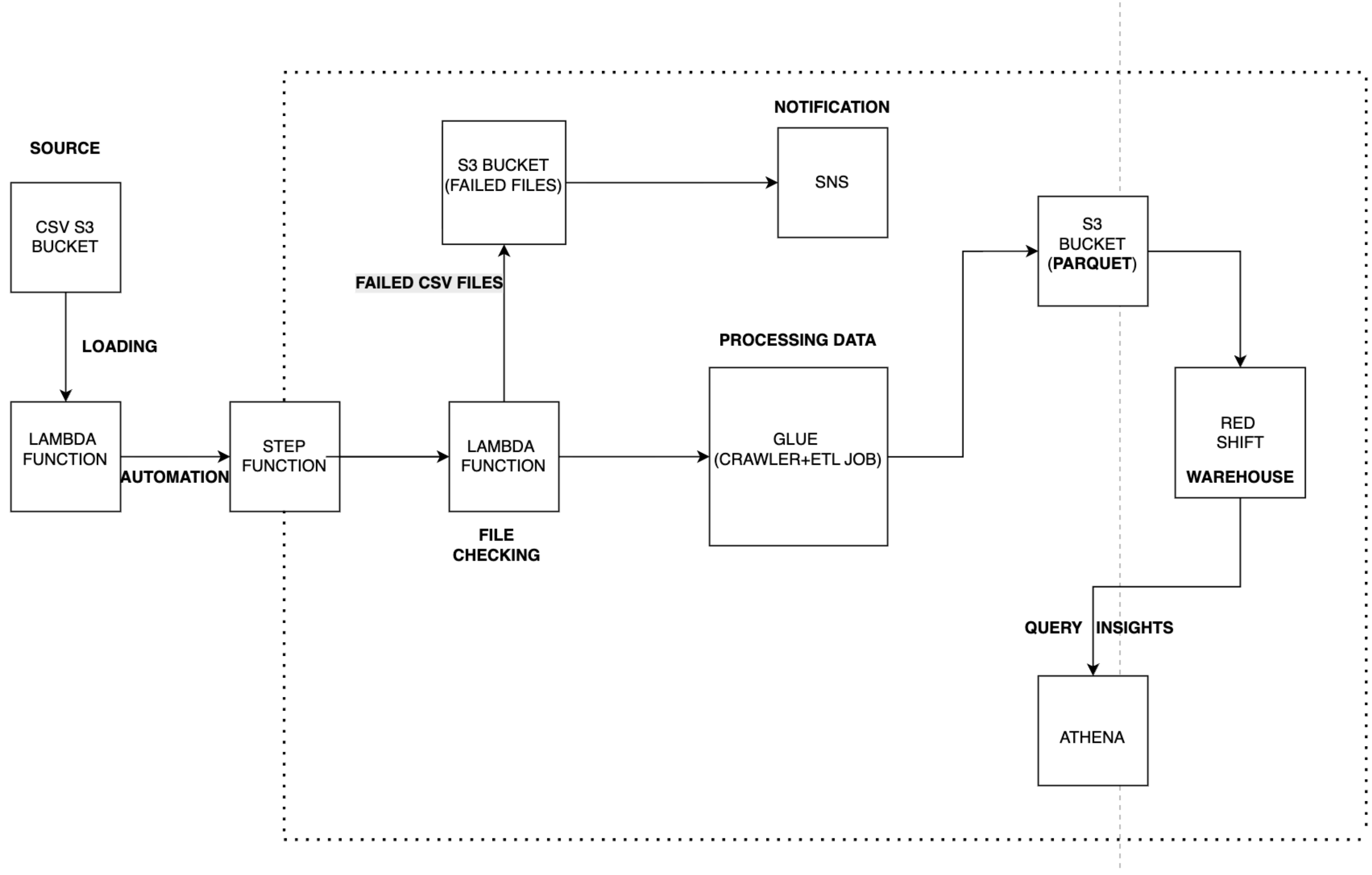
STEP 5: **CONNECTING TO AMAZON ATHENA**

- ➢ Define **Athena Workgroup** and connect it with Redshift Spectrum for querying processed data.

CloudFormation for the Pipeline:

This CloudFormation stack sets up an end-to-end data processing pipeline on AWS. It provisions the required S3 buckets, IAM roles, Glue crawlers, Glue ETL jobs, and Redshift Serverless resources. The pipeline automates ingestion of CSV files using Lambda and Step Functions, transforms them into Parquet format, and loads the processed data into Amazon Redshift and integrates with Amazon Athena for querying and analytics.

ARCHITECTURE


SOURCE
CSV S3 BUCKET
LOADING
LAMBDA FUNCTION
AUTOMATION
STEP FUNCTION
LAMBDA FUNCTION
FILE CHECKING
FAILED CSV FILES
S3 BUCKET (FAILED FILES)
NOTIFICATION
SNS
PROCESSING DATA
GLUE (CRAWLER+ETL JOB)
S3 BUCKET (PARQUET)
RED SHIFT
WAREHOUSE
QUERY INSIGHTS
ATHENA