Architecture for this requirement

Data pipeline to Load the data into the data warehouse and for visualization purpose we are using Apache Superset.

A diagram of a diagram

Description automatically generated with medium confidence

As per the above Diagram I have used the following service to achieve this requirement.

1. **S3 Bucket**: The parquet files are stored in S3 bucket.

A screenshot of a computer

Description automatically generated

1. **Lambda Function**: When a new parquet file has been uploaded, the lambda function will be triggered for creating Glue Crawler.

newyork-taxi-data/lambda/ data\_pipeline\_trigger.py

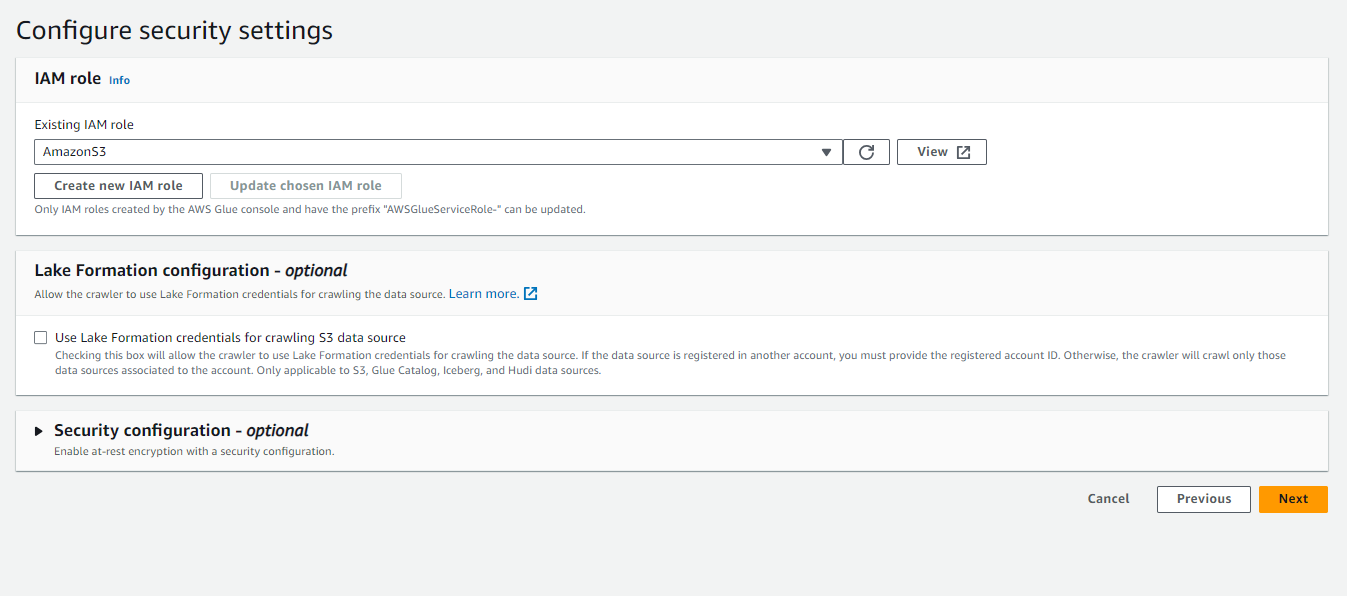
1. **Glue Crawler**: The Lambda function creates a new Glue crawler, which is configured to crawl the S3 bucket and the parquet file(s) that triggered the Lambda function. So, key advantages of using AWS Glue Crawler are that they can automatically create databases and tables in the AWS Glue Data Catalog without manual intervention.

A screenshot of a computer

Description automatically generated

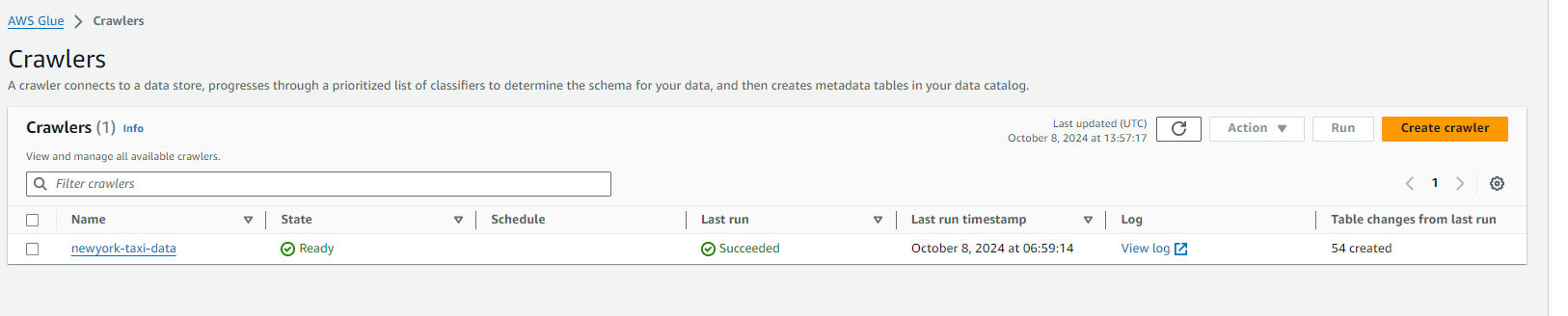
A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

1. **Glue Job**: After the crawler is created and has processed the parquet file(s), the Lambda function triggers a Glue job which will reads the data from the source (**S3 bucket**) based on the metadata in the Glue Data Catalog, performs necessary transformations, and loads the data into the target data store (**Amazon Redshift**). For connecting AWS Glue and Redshift cluster we need username and password so for making more secure manner we have used secret manager and passed an argument in Glue job.

**Glue Job** has been written at the following path in project directory

**ecommerce-datawarehouse/glue/etl.py**

1. **Redshift:** Once the data is loaded into Redshift then we can do further analysis and querying on the data using SQL queries.

Table Structure in Redshift

FHV Trip Data Tables

dev.public.fhv\_tripdata\_2022\_01

dev.public.fhv\_tripdata\_2022\_02

dev.public.fhv\_tripdata\_2022\_03

dev.public.fhv\_tripdata\_2022\_04

dev.public.fhv\_tripdata\_2022\_05

dev.public.fhv\_tripdata\_2022\_06

dev.public.fhv\_tripdata\_2023\_01

dev.public.fhv\_tripdata\_2023\_02

dev.public.fhv\_tripdata\_2023\_03

dev.public.fhv\_tripdata\_2023\_04

dev.public.fhv\_tripdata\_2023\_05

dev.public.fhv\_tripdata\_2023\_06

dev.public.fhv\_tripdata\_2024\_01

dev.public.fhv\_tripdata\_2024\_02

dev.public.fhv\_tripdata\_2024\_03

dev.public.fhv\_tripdata\_2024\_04

dev.public.fhv\_tripdata\_2024\_05

dev.public.fhv\_tripdata\_2024\_06

Green Trip Data Tables

dev.public.green\_tripdata\_2022\_01

dev.public.green\_tripdata\_2022\_02

dev.public.green\_tripdata\_2022\_03

dev.public.green\_tripdata\_2022\_04

dev.public.green\_tripdata\_2022\_05

dev.public.green\_tripdata\_2022\_06

dev.public.green\_tripdata\_2023\_01

dev.public.green\_tripdata\_2023\_02

dev.public.green\_tripdata\_2023\_03

dev.public.green\_tripdata\_2023\_04

dev.public.green\_tripdata\_2023\_05

dev.public.green\_tripdata\_2023\_06

dev.public.green\_tripdata\_2024\_01

dev.public.green\_tripdata\_2024\_02

dev.public.green\_tripdata\_2024\_03

dev.public.green\_tripdata\_2024\_04

dev.public.green\_tripdata\_2024\_05

dev.public.green\_tripdata\_2024\_06

Yellow Trip Data Tables

dev.public.yellow\_tripdata\_2022\_01

dev.public.yellow\_tripdata\_2022\_02

dev.public.yellow\_tripdata\_2022\_03

dev.public.yellow\_tripdata\_2022\_04

dev.public.yellow\_tripdata\_2022\_05

dev.public.yellow\_tripdata\_2022\_06

dev.public.yellow\_tripdata\_2023\_01

dev.public.yellow\_tripdata\_2023\_02

dev.public.yellow\_tripdata\_2023\_03

dev.public.yellow\_tripdata\_2023\_04

dev.public.yellow\_tripdata\_2023\_05

dev.public.yellow\_tripdata\_2023\_06

dev.public.yellow\_tripdata\_2024\_01

dev.public.yellow\_tripdata\_2024\_02

dev.public.yellow\_tripdata\_2024\_03

dev.public.yellow\_tripdata\_2024\_04

dev.public.yellow\_tripdata\_2024\_05

dev.public.yellow\_tripdata\_2024\_06