A black screen with white text

Description automatically generatedA black and white text on a black background

Description automatically generatedA black screen with white text

Description automatically generated

Project Structure:

Create 2 crawler-

1. MyCrawlerFetchFromS3
2. MyCrawlerFetchFromRedshift

Create 2 ETL Job-

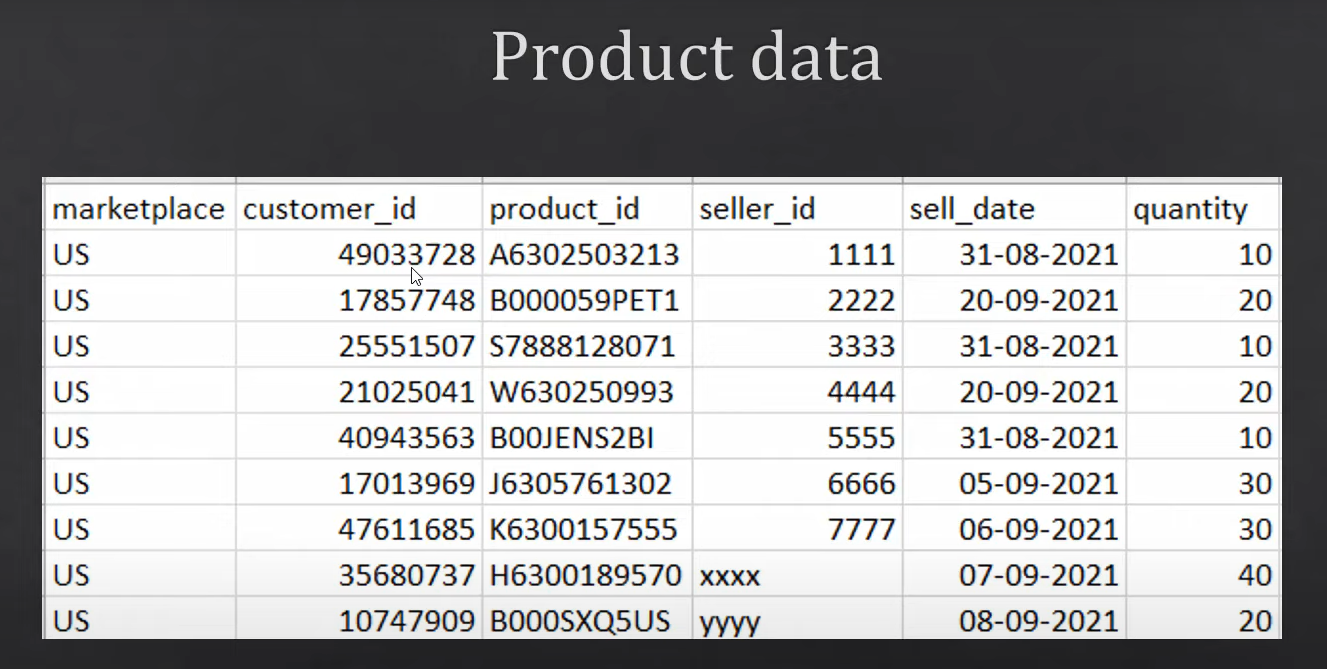
1. MyGlueJobReadFromS3
2. MyGlueInsertRedshift

AWS Lamba - can run max 15 minutes continuously

AWS Glue- can run 48 hours continuously

Both are serverless

We will read below data:

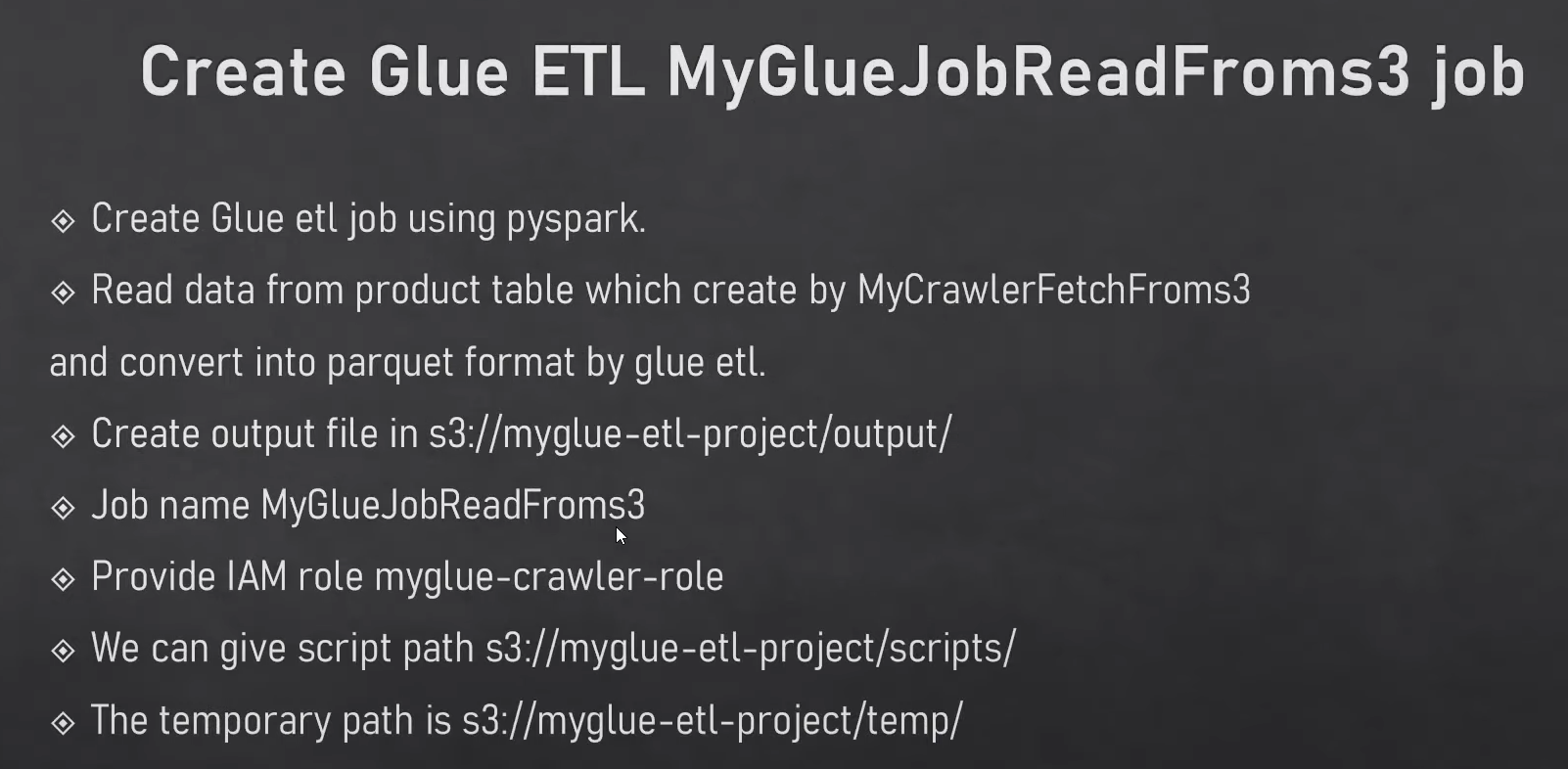


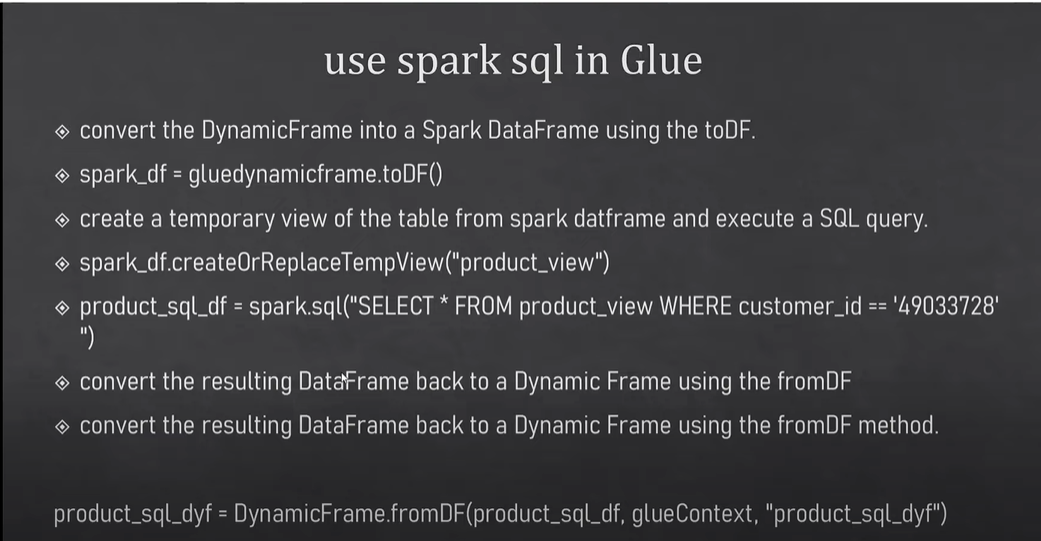
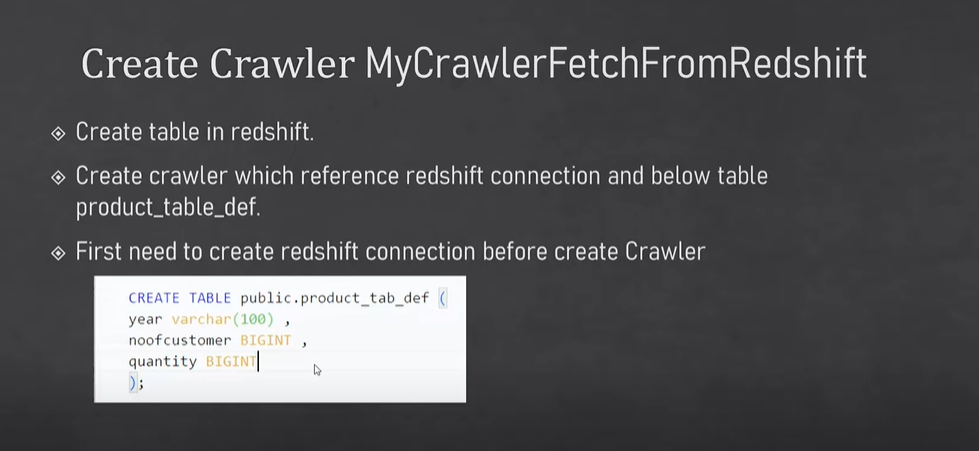
1. Create s3 bucket named - myglue-etl-project-sit
2. Create 4 folder inside above bucket
   1. Input -> product/year=2021
   2. Output
   3. Scripts
   4. Temp
3. Upload file- product\_data.csv file into year=2021 location
4. Create database named- mydatabase into glue
5. Once you create database into glue it will start reflecting into Athena
6. If we don’t have database, then it will create schema into default database
7. Create IAM role named- **myglue-crawler-role**
8. Create first crawler named- MyCrawlerFetchFromS3
9. If we upload another file into year=2021 then it will start reflecting at athena but if we create new folder named- year=2022 and add file into it then it will not reflect automatically, we need to run crawler again.

A black and white screen with white text

Description automatically generated

Now we will create our first ETL script that will fetch data from catalog and convert into parquete format and store into s3 bucket.

 A black sign with white text

Description automatically generated  

CREATE TABLE public.product\_tab\_def(

    year VARCHAR(100),

    noofcustomer BIGINT,

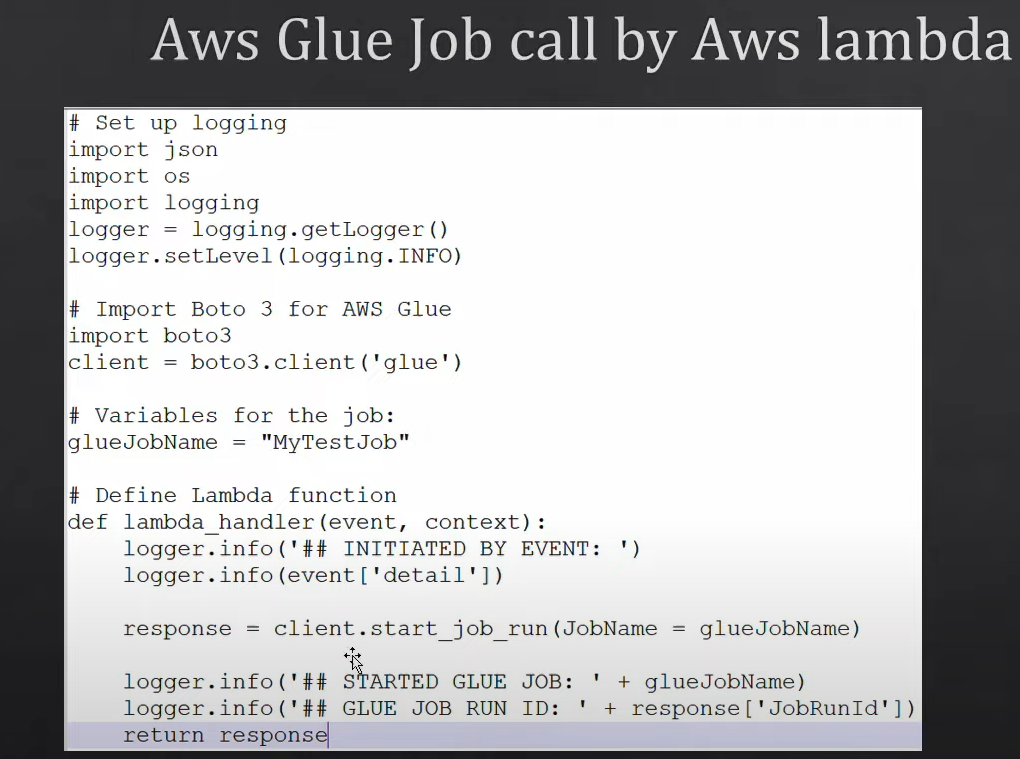
    quantity BIGINT

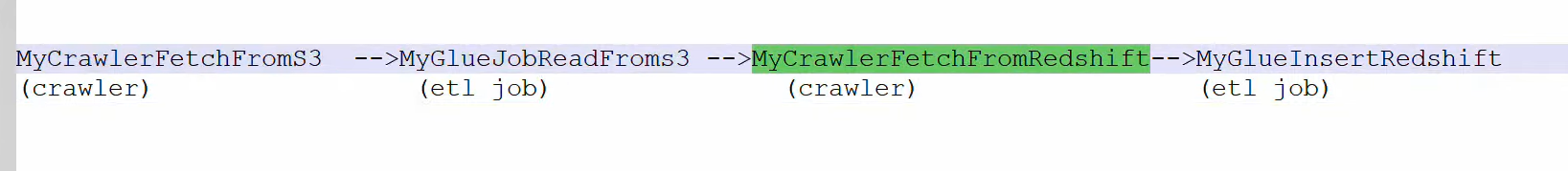
);

SELECT \* from dev.public.product\_tab\_def;

1. Create serverless redshift cluster named(namespace) – myredshiftcluster and create above table inside the redshift cluster.
2. Create username and password- awsuser /Admin123
3. Create connection in glue before creating crawler for redshift and test this connection
4. Redshift is a jdbc type of database- so while creating crawler we need to select jdbc type of data source.
5. Create another role named- **GlueRedshiftCrawlerRole** Which have s3,glue,redshift full access. It will attach with redshift.
6. Create crawler named – MyCrawlerFetchFromRedshift (source- redshift and ouput-mydatabase)
7. After running this crawler table is created in glue but not visible in athena
8. Create etl job named- MyGlueInsertRedshift (source s3 output location -> target redshift)
9. Post this run parquet file data inserted into redshift.

**We can run this etl job using aws lambda.**



* We can create [Workflows (orchestration)](https://ap-south-1.console.aws.amazon.com/glue/home?region=ap-south-1#/v2/etl-configuration/workflows) named-myworkflow in glue that will run etl job in sequence
* 
  1. First create new trigger named – Myeventtrigger that will run ondemand

Myeventtrigger-for-MyCrawlerFetchFromS3

Myeventtrigger-for-MyGlueJobReadFromS3

Myeventtrigger-for-MyCrawlerFetchFromRedshift

Myeventtrigger-for-MyGlueInsertRedshift

A screenshot of a computer screen

Description automatically generated