**Day 7 – Assignment 1 (Deep Gupta)**

Text, letter

Description automatically generated

***Place the dataset in the bucket***

A screenshot of a computer

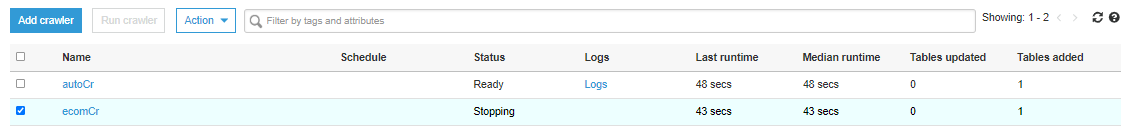
Description automatically generated

***Selecting Up the Crawler with proper Path***

Graphical user interface, text, application

Description automatically generated

***Crawler Run Complete***



***Table Schema Successfully Created***

Graphical user interface, text, application

Description automatically generated

***Cluster Creation For Redshift***

Graphical user interface, text, application, email

Description automatically generated

***Cluster Role Creation with Desired Roles***

Graphical user interface, text, application, email

Description automatically generated

***Cluster IAM Permission Association***

Graphical user interface, text, application, email

Description automatically generated

***Connect to a new database for table (Database Name – dev)***

Graphical user interface, text, application

Description automatically generated

**Since Data Not Mapping Correctly because of newline and ‘,’ characteers.**

**Using a different strategy**

**Steps: Preprocess the data using pyspark -> Create a cluster in EMR -> Read the data using DF - > Store in the bucket the dataset as Parquet -> Use the glue crawler**

**Creation of EMR cluster + notebook**

Graphical user interface, text, application

Description automatically generated

**Source Code for Notebook:**

*spark*

*ecomDF = spark.read.option("header", True).option("inferSchema", True).option("multiline", True).csv("s3://day7bucket/ecom/Ecommerce Purchases")*

*ecomDF.printSchema()*

*ecomDF.show(10)*

*ecomDFnew = ecomDF.withColumnRenamed("AM or PM", "AMorPM").withColumnRenamed("Browser Info", "BrowserInfo").\*

*withColumnRenamed("Credit Card", "CreditCard").withColumnRenamed("CC Exp Date", "CCExpDate")\*

*.withColumnRenamed("CC Security Code", "CCSecurityCode").withColumnRenamed("CC Provider", "CCProvider")\*

*.withColumnRenamed("IP Address", "IPAddress").withColumnRenamed("Purchase Price", "PurchasePrice")*

*ecomDFnew.write.option("header", True).parquet("s3://day7bucket/ecom/parquetEcom")*

Text

Description automatically generated with medium confidence

**Updated the crawler with the correct parquet file path**

Graphical user interface, text, application

Description automatically generated

**The code successfully wrote the dataset in the parquet file format**

Graphical user interface, text, application, email

Description automatically generated

**Again after successful run of the crawler a new table was added**

Graphical user interface, text, application, email

Description automatically generated

**Table Schema & Properties**

Graphical user interface, application

Description automatically generated with medium confidence

**Source Code for creating new schema and fetching the table in Redshift**

create external schema ecomschema from data catalog

database 'ecommerce'

iam\_role 'arn:aws:iam::991766072808:role/redshiftClusterRole'

region 'us-east-1';

Text

Description automatically generated

**Data coming in properly in the table**

A screenshot of a computer

Description automatically generated

**Queries:**

1. What is the average Purchase Price?

select avg(purchaseprice) from ecomschema.parquetecom;  
Graphical user interface, text, application, email

Description automatically generated

--2. What are the highest and lowest purchase prices?

select max(purchaseprice) as highest\_pp, min(purchaseprice) as lowest\_pp from ecomschema.parquetecom;

Graphical user interface, text, application, email, Teams

Description automatically generated

--3. How many people have English ‘en’ as their language of choice on the website?

select count(\*) from ecomschema.parquetecom where language like 'en';

Graphical user interface, text, application, email

Description automatically generated

--4. How many people have the job title of “Lawyer”?

select count(\*) from ecomschema.parquetecom where lower(job) like '%lawyer%'

Graphical user interface, text, application, email

Description automatically generated

--5. How many people made the purchase during the AM and how many people made the purchase during the PM?

select amorpm, count(\*) from ecomschema.parquetecom

group by amorpm;

Graphical user interface, text, application, email

Description automatically generated

--6. What are the 5 most common Job Titles?

select count(job) as jobTitlesCount, job from

ecomschema.parquetecom group by job

order by jobTitlesCount desc limit 5;

A screenshot of a computer

Description automatically generated

--7. Someone made a purchase that came from Lot: “90 WT”, what was the Purchase Price of this transaction?

select purchaseprice,lot from ecomschema.parquetecom

where lot like '%90 WT%';

Graphical user interface, text, application, email

Description automatically generated

--8. What is the email of the person with the following credit card number: 4926535242672853?

select email, creditcard from ecomschema.parquetecom

where creditcard = 4926535242672853;

Graphical user interface, text, application, email

Description automatically generated

--9. How many people have AMEX as their CC provider and made a purchase above $95?

select count(creditcard) from ecomschema.parquetecom

where purchaseprice > 95;

Graphical user interface, text, application, email

Description automatically generated