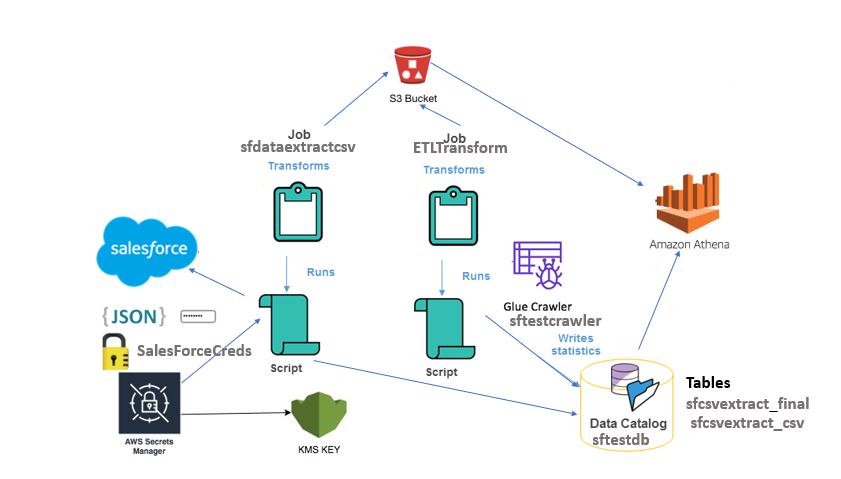
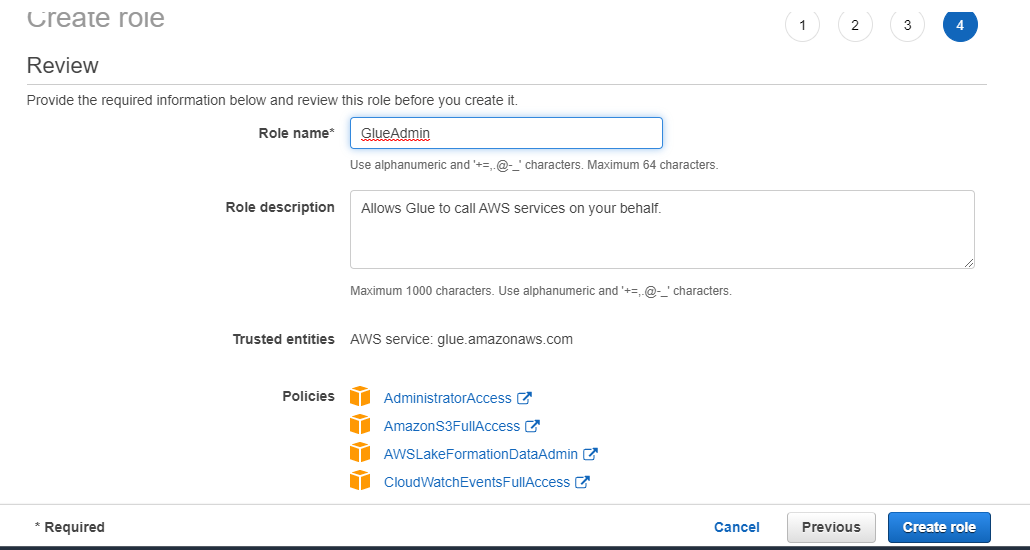
**SALESFORCE DATA EXTRACT IN TO S3 BUCKET WITH USING AWS GLUE ETL JOB AND VIEW WITH ATHENA**

**Objective**: Salesforce Data Extract into S3 Bucket with Using Aws Glue ETL Job and View with Athena.

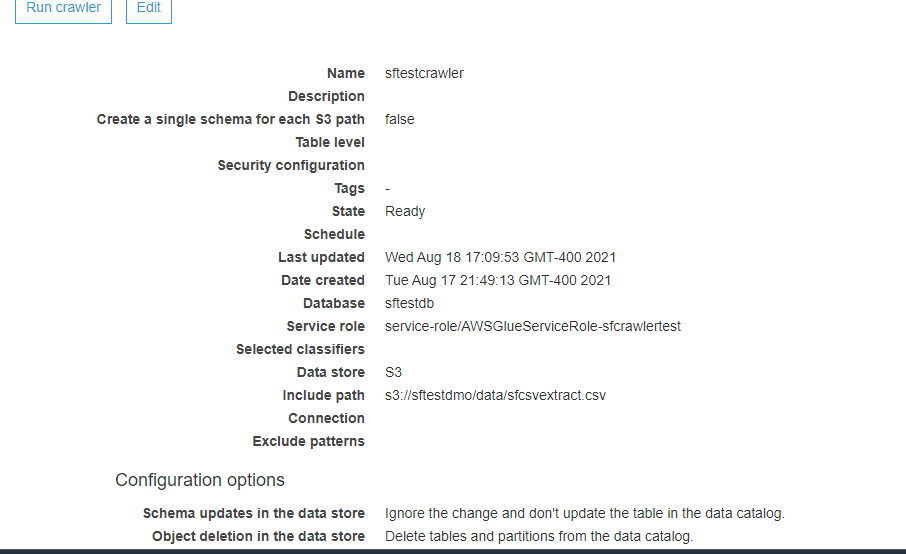
1. **AWS architecture with data flow diagram:**

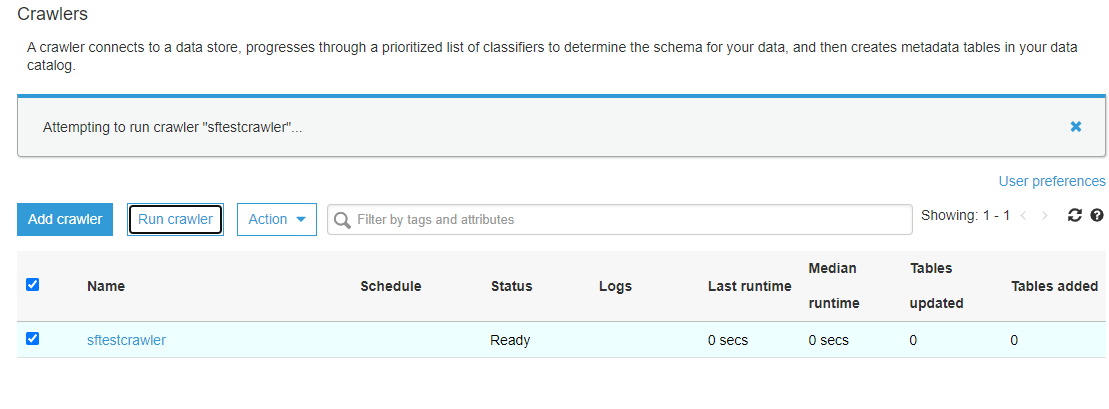


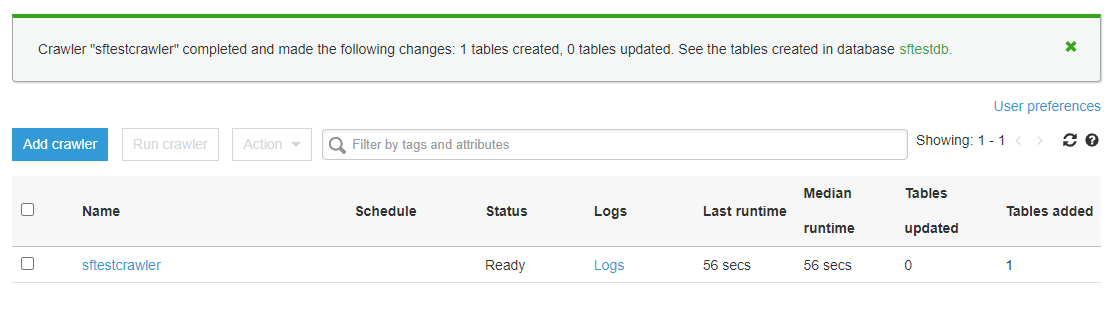
1. Create Glue role used in the Glue jobs.



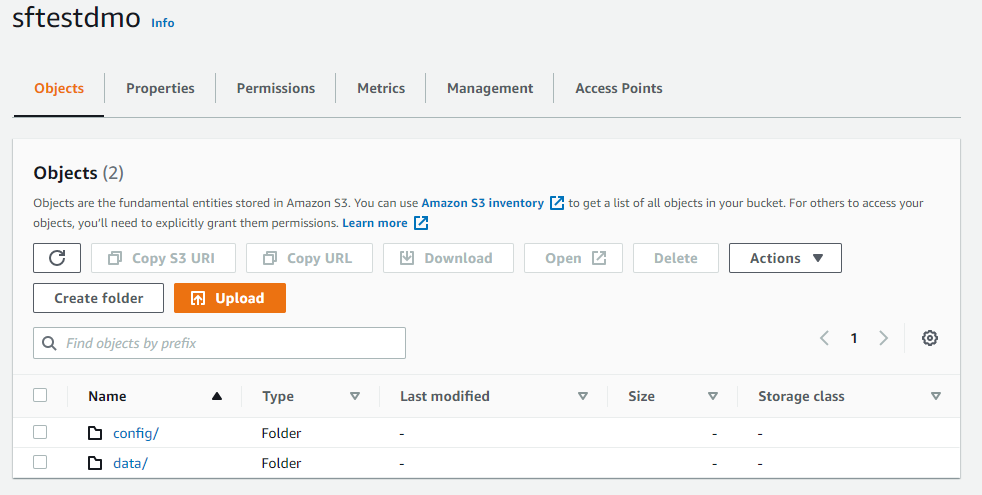
1. Create a crawler in the Glue job for database catalog.



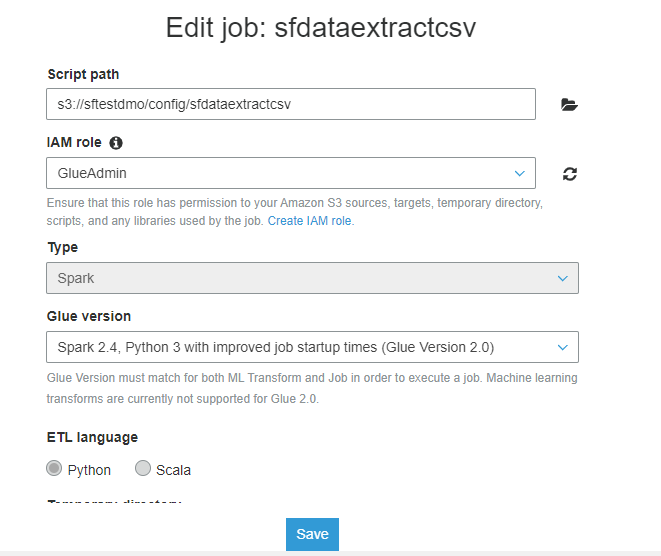


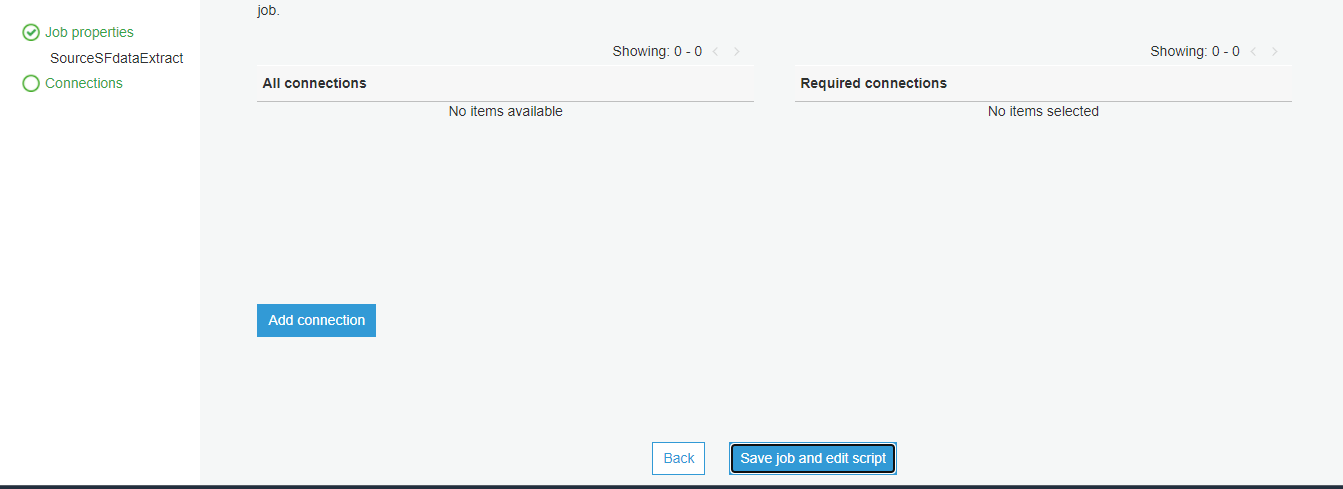


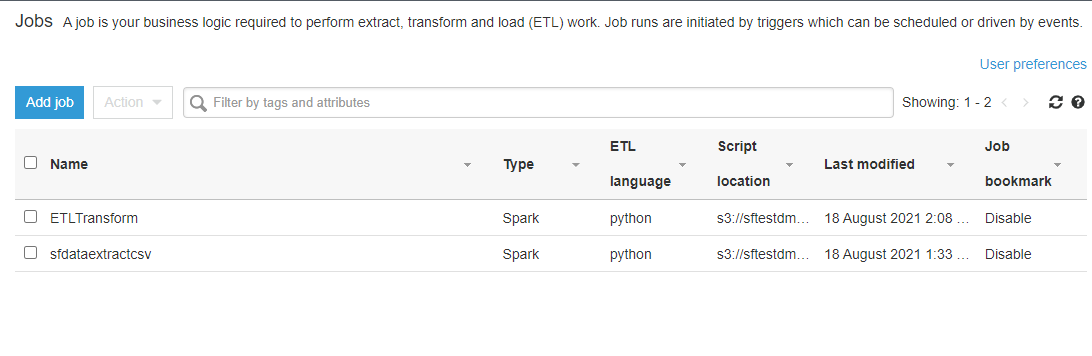
1. Create S3 bucket with different folders to keep CSV, Config and Temp files generated by the Glue jobs while processing data.



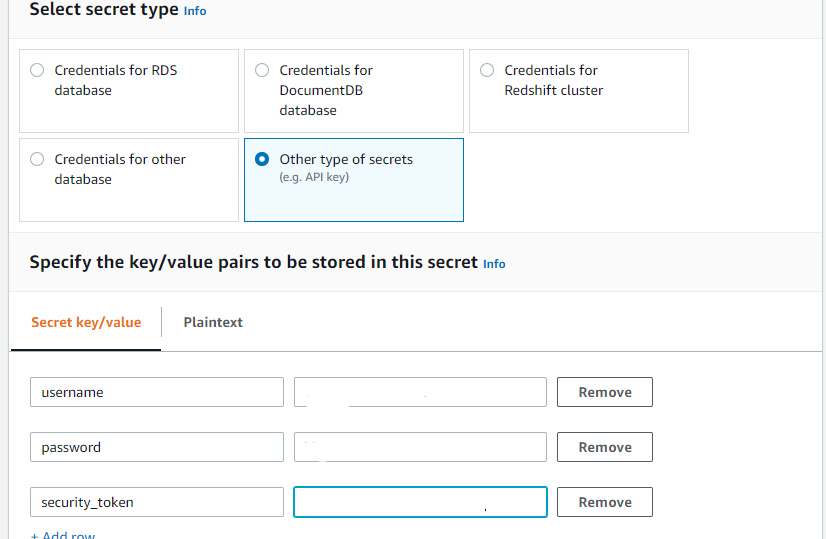
1. Create glue job -1 to connect Salesforce and extract data from the table and configure as needed.

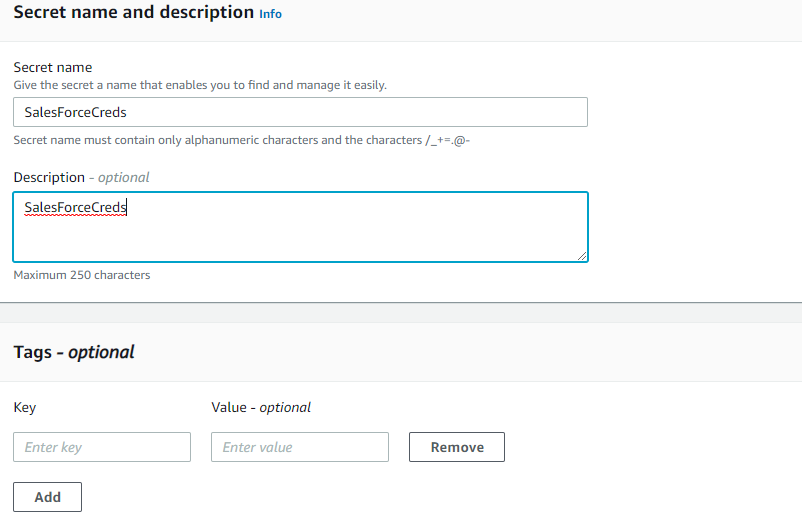


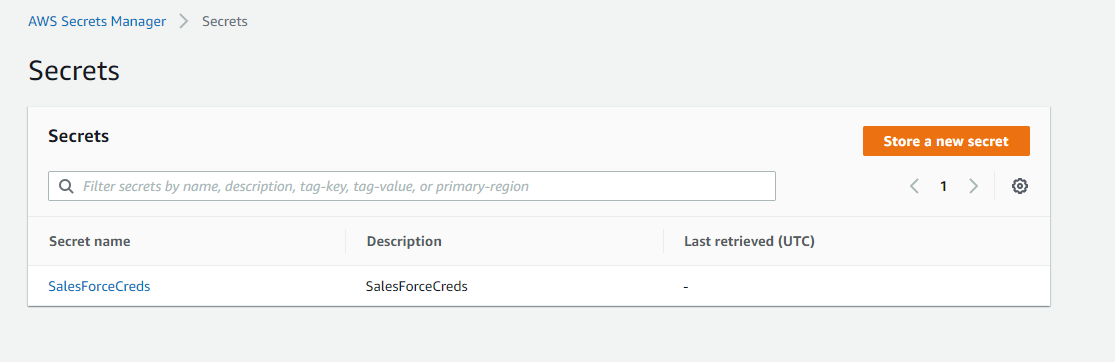




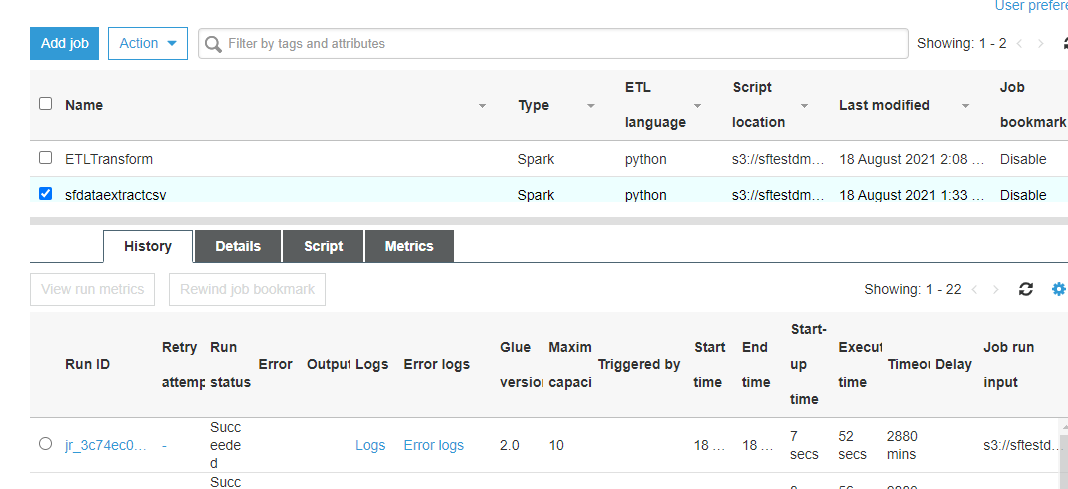
1. Store Salesforce credentials in SecretsManager.

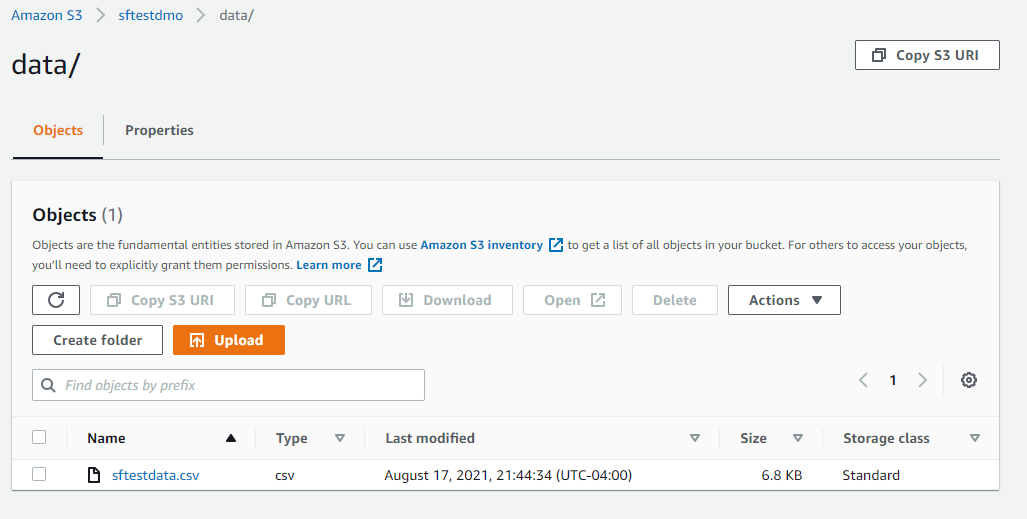






1. Run the SalesForce data extract job for below CSV raw data.





1. Sfdataextractcsv job python script. Which imports credentials from secretManagers from aws and connect to SalesForce and extract data into S3 bucket and removes null rows.

import boto3

import base64

from botocore.exceptions import ClientError

import pandas as pd

import numpy as np

import json

import pandas

from pandas import DataFrame, Series

from simple\_salesforce import Salesforce #imported salesforce

client = boto3.client('secretsmanager')

response = client.get\_secret\_value(

SecretId='SalesForceCreds'

)

secretDict = json.loads(response['SecretString'])

s3 = boto3.resource("s3")

obj = s3.Object("sftestdmo", "/data/sfcsvextract.csv")

obj.delete()

sf = Salesforce(username=secretDict['username'], password=secretDict['password'], security\_token=secretDict['security\_token'])

s3 = boto3.resource('s3')

#BUCKET = "sfdatatestdemo"

a\_query= pd.DataFrame(sf.query("SELECT Salutation, FirstName, LastName, Email, Title, Department FROM Contact")['records'])

filter = a\_query["Email"]!=""

a\_query.where(filter, inplace = True)

a\_query=a\_query.dropna()

a\_query.replace({'\"': ''}, regex=True)

a\_query.to\_csv('s3://sftestdmo/data/sfcsvextract.csv',index=False)

print('Completed')

1. Create a table with the below ddl script for Athena data view.

CREATE EXTERNAL TABLE `sfcsvextract\_final`(

`salutation` string,

`firstname` string,

`lastname` string,

`email` string,

`title` string,

`department` string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

WITH SERDEPROPERTIES (

'quoteChar'='\"')

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

's3://sftestdmo/data/sfcsvextract.csv'

TBLPROPERTIES (

'CrawlerSchemaDeserializerVersion'='1.0',

'CrawlerSchemaSerializerVersion'='1.0',

'UPDATED\_BY\_CRAWLER'='sftestcrawler',

'areColumnsQuoted'='false',

'averageRecordSize'='97',

'classification'='csv',

'columnsOrdered'='true',

'compressionType'='none',

'delimiter'=',',

'objectCount'='1',

'recordCount'='57',

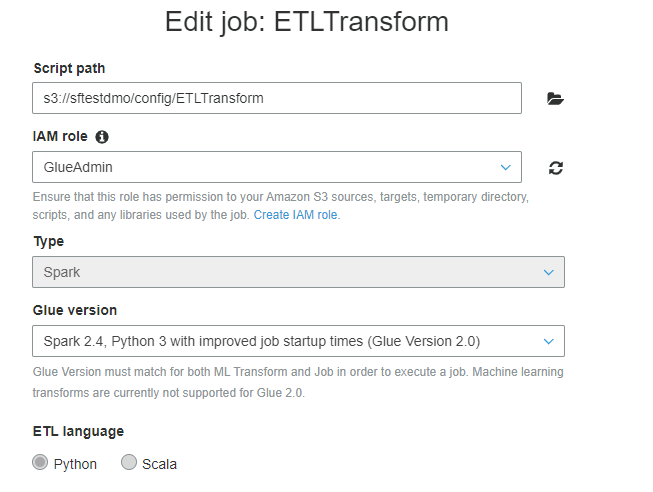
'sizeKey'='5554',

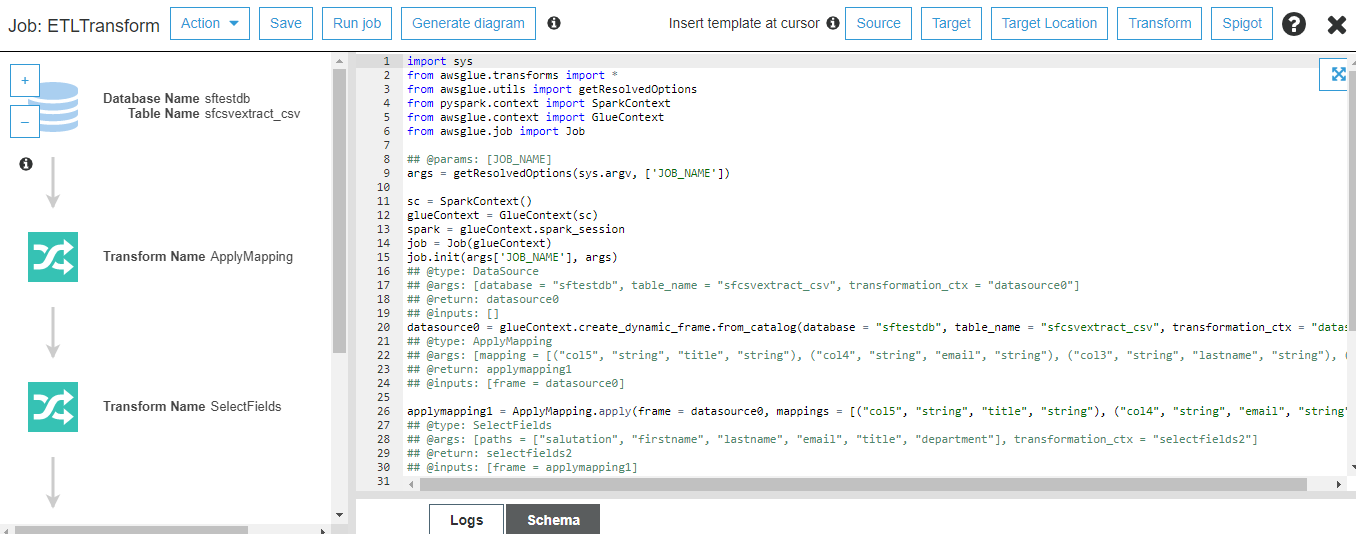
'skip.header.line.count'='2',

'transient\_lastDdlTime'='1629330051',

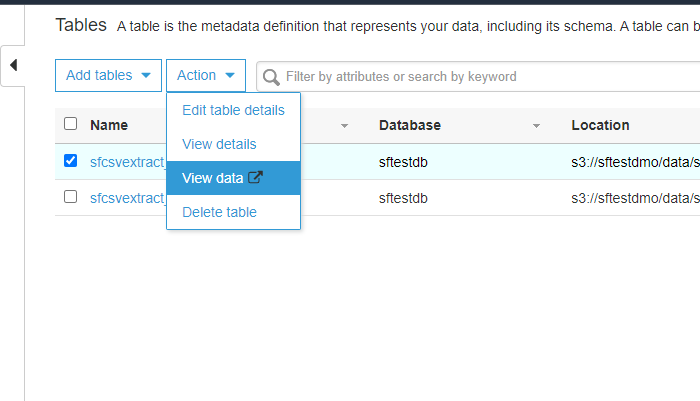
'typeOfData'='file')

1. Job-2 to extract csv file data for transformation for viewing data in Athena.

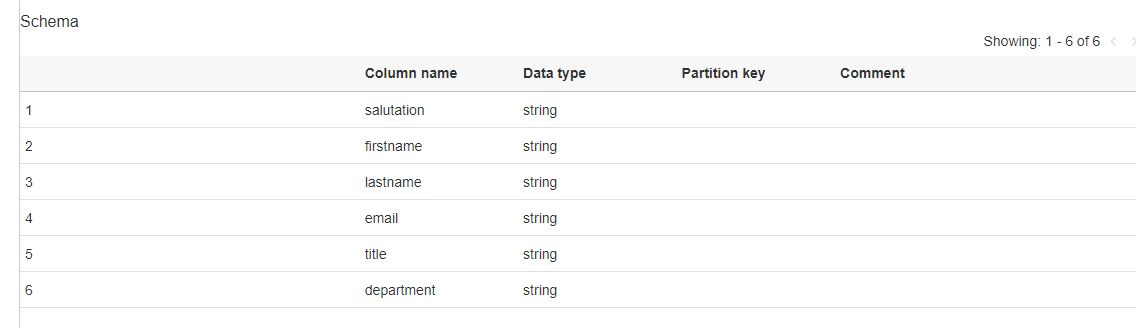


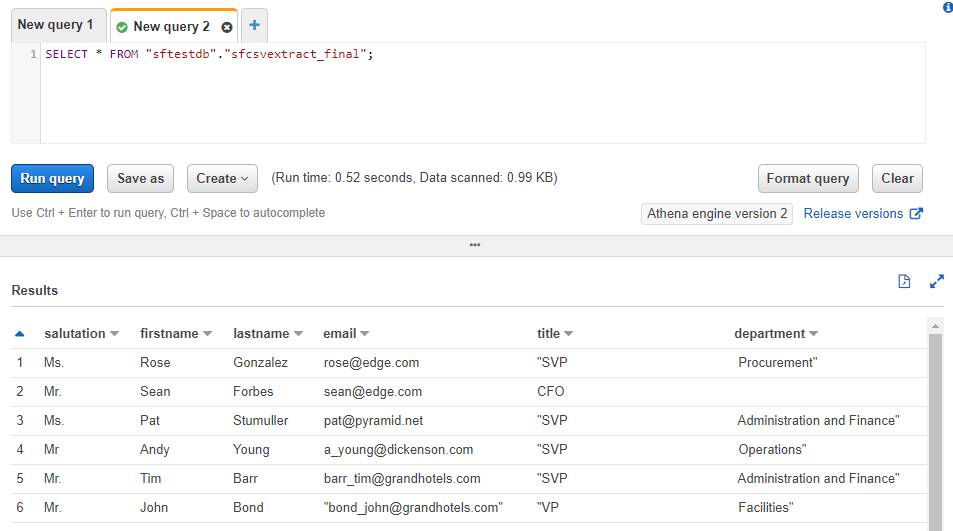


1. View the data using athena.



1. Data catalog table columns details:





Thank you..

Vijay S (DEA Team)