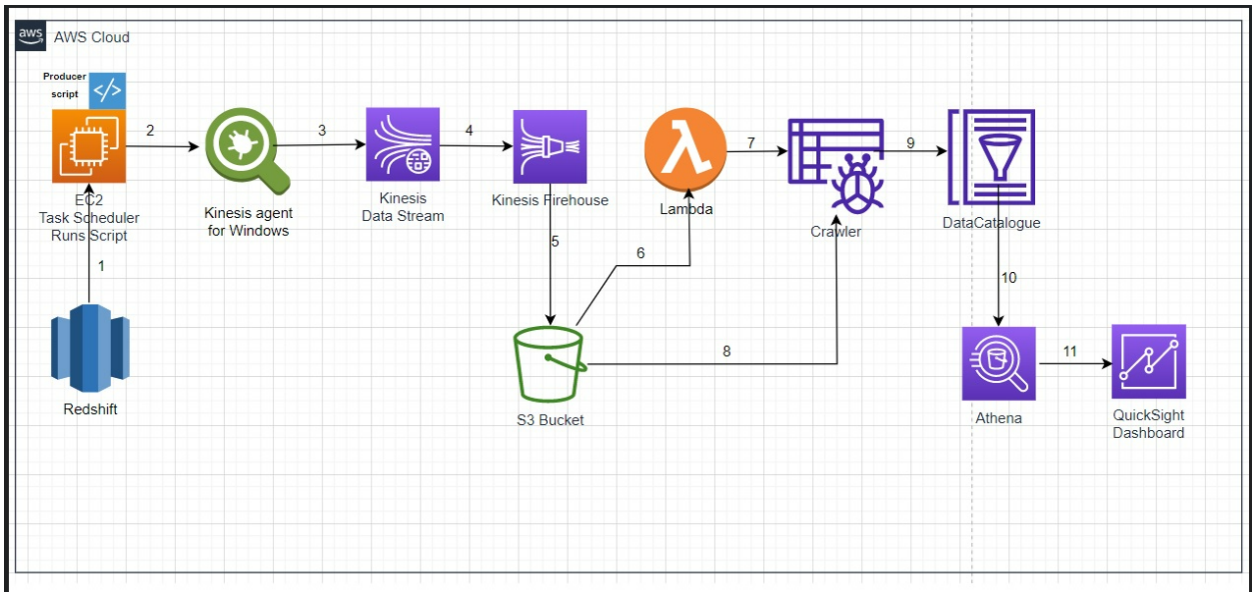
LIVE DATA STREAM WITH KINESIS FROM REDSHIFT AND DELIVERY TO QUICKSIGHT DASHBOARD

Warning: This project will pick up where Project 3 ended - Redshift table. To successfully complete this project you must have finished Project 3. This project is using EC2, Security Group, some IAM Roles, Redshift and S3 buckets that were created in Project 3.
Addition to 'enriched' glue job from Project 3 will be performed to produce trigger file for the next pipeline to start once pipeline from Project 3 completes.

Purpose: Design an aws cloud based data pipeline that will read a table from redshift database, simulate streaming data delivery into S3 bucket, from where crawlers will read the data and create metadata table called - catalog, and finally quicksight dashboard will pick up streaming data from data catalog though direct athena queries when an update is performed.

Take a look at the following Diagram to get a high level glimpse into the expected work awaiting us:



Prerequisites: AWS Account, Installed SQL Workbench, Python, Basic Network Configuration, json, SQL, RDC, Successful pipeline from Project 3, downloaded files from folder 'Project 4' in GitHub link provided: https://github.com/Myself1214/Upwork.git

Plan of Work (Pseudo Work):

1. Implement additional code to 'Enriched' glue job from Project 3 to produce trigger file in S3 for this pipeline to start
2. Program a data producer and load it on EC2 through a shared path that is mounted from local to EC2.
3. Create a role for producer to write data on Kinesis Data Stream and attach it to EC2
4. Create a folder in your S3 raw bucket and name it 'from-firehouse'
5. Create Kinesis Data Stream
6. Create Kinesis Firehose
7. Install and configure Kinesis Agent for Windows on your EC2 to stream data from EC2 server into kinesis data stream and then to firehose delivery stream which will deliver data to S3 bucket in 'from-firehouse' folder. Run the agent
8. Create crawler to crawl streaming data from raw S3 bucket and create table metadata as a catalog
9. Create Lambda function to trigger crawler every time when new data is streamed and loaded in S3 raw bucket
10. Create Quicksight dashboard to be updated through direct athena queries
11. Create an event based EventBridge rule to trigger execution of 'producer' file that in turn will trigger all data pipeline.

Actual Steps:

1. Add new code to 'enriched' glue job from Project 3 for trigger file
   - From github link provided, under 'Project 4' download 'producing_trgger_file.py' and add it to existing script of 'enriched' glue job from previous project at the end of code
2. Get producer and load it onto EC2 mounted shared path from local
   - From the github link provided, under 'Project 4' download file 'producer.py', edit it and add your redshift cluster connection strings where indicated and table name. Since you already came here after finishing Project 3, you already have a Redshift cluster with a loaded table - '911 calls' file in it. (If you have not gone through Project 3, you will not be able to continue). Once your edited producer, save it and copy it into path in your local machine that is mounted on your EC2 Windows instance that you created in Project 3 and it will be accessible on your remote server

3. Create a role for producer to write data on Kinesis Data Stream and attach it to EC2
   - From github link provided, under 'Project 4' download file 'ec2_kinesis_role_policy.json'
   - Log in to the IAM console on aws, select 'Roles' on the left pane. Click on 'Create role'. Under common use cases select 'EC2' and on the next screen select 'Create policy'. Now switch to the 'JSON' tab and paste the code from the file downloaded from github and click next twice. On review page give your role name 'kinesis_role' and click on 'Create policy'
   - Switch to EC2 console, select your windows instance created in Project 3, and from right top select 'Actions'. From the list select 'Security' and then 'Modify IAM role'. In new screen from drop-down list select 'kinesis_role' created earlier and click on 'Update role'
4. Create a folder in your S3 raw bucket and name it 'from-firehouse'
   - Switch to the S3 console, select the 'Raw' bucket created in Project 3, then click on 'Create folder'. Give it a name 'from-firehouse/' and create.
5. Create Kinesis Data Stream
   - Switch to Kinesis console, on the left pane select 'Data Streams', then 'Create data stream'. Give it a name of "MyFirstDataStream', scroll down and click on 'create data stream'
6. Create Kinesis Firehose
   - On the left pane select 'Delivery Streams', then click on 'create delivery stream'. For the 'Source' select 'Amazon Kinesis Data Stream' and for 'Destination' select 'Amazon S3'
   - For 'Source settings' select the data stream created in step 5, then give delivery steam a name of your choice.
   - For 'Destination Settings' select 'Raw' S3 bucket and for 'S3 bucket prefix' put name of folder 'from-firehouse/' created in step 4 and click on 'Create delivery steam'.
7. Install and configure Kinesis Agent for Windows on your EC2
   - To install kinesis agent for ms windows, go to: https://s3-us-west-2.amazonaws.com/kinesis-agent-windows/downloads/index.html, and follow instructions for 'Install using PowerShell'
   - Now we need to create configuration file for Kinesis Windows Agent that will help it find necessary file to read from - 'Source', and target destination where it should stream data to - 'Sink'
   - Once installation is done, open the notepad and create a Kinesis Agent for Windows configuration file called 'appsettings.txt'. Inside this file copy and paste content of file 'appsettings.txt' that you'll download from the github link provided

under Project 4. Replace following values in your ‘appsettings.txt’ file with your own:

- ‘id’ - pick any id name
- ‘Directory’ - path to source file from which kinesis will read data - this is same path where ‘proucer.py’ file will write ‘json_file.txt’ file in
- ‘SourceRef’ - id name that you assigned to ‘Sources’
- ‘SinkRef’ - id name that you assigned for ‘Sink’
- ‘StreamName’ - name of your Data Stream created in step 5
- ‘AccessKey - your account access key
- ‘SecretKey’ your account secret key
- ‘Region’ - you account region

- Once you are done with all changes to the ‘appsettings.txt’ file, save it in the following directory: *C:\Program Files\Amazon\AWSKinesisTap* . If such a file with the same name already exists in the above directory, replace it with yours.
- Now we need to start Kinesis agent by executing following command in PowerShell: *Start-Service -Name AWSKinesisTap*
(To stop Kinesis agent, use this command: *Stop-Service -Name AWSKinesisTap*)
Now the agent will be continuously running in the background and waiting for the ‘json_file.txt’ file in the directory that we specified in the ‘appsettings.txt’ file. The file ‘json_file.txt’ will be produced by our ‘producer.py’ file created in step 2.

Start-Service -Name AWSKinesisTap
Stop-Service -Name AWSKinesisTap

To check status of SSMAgent installation on ec2

```
Get-Service AmazonSSMAgent
```

```
Turned on eventbridge on properties of enriched bucket to allow eventbridge rule to get object created event when triger_file.txt is created

Created folder ‘for_trigger_file’ in enriched bucket
```

(gcm py).source -- to get python executable location on windows (for ssm agent for Systems Manager to be able to find python executable to run producer.py)