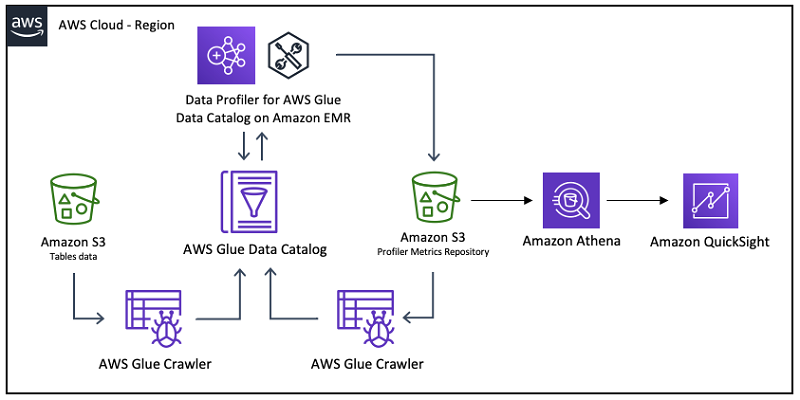
**Lab Guide**: Build an automatic data profiling and reporting solution with Amazon EMR, AWS Glue and Amazon Quicksight

Create an automatic data profiling repository, as an extension of AWS Glue Data Catalog metadata, and a reporting system.

**Overview and Architecture**

Data Profiler for **AWS Glue Data Catalog** is an **Apache Spark** application that profiles all the tables defined in the Data Catalog using the profiling capabilities of **Amazon Deequ library** and saves the results in the Data Catalog and an **Amazon S3** bucket in a partitioned Parquet format. Other analytics services such as **Amazon Athena** and **Amazon Quicksight** can be used to query and visualize the data.



**Technical Requirements**: Have adequate knowledge about AWS Glue, S3 and Apache Spark.

**Guidelines**:

**Reminder**: All works are conducted in us-west-2 (Oregon) region, remember to choose the region prior to creating resources.

**Step 1**: Create an EC2 key pair

This key pair will later be use to connect to EMR master node via SSH. **Remember** to keep your PPK/PEM file.

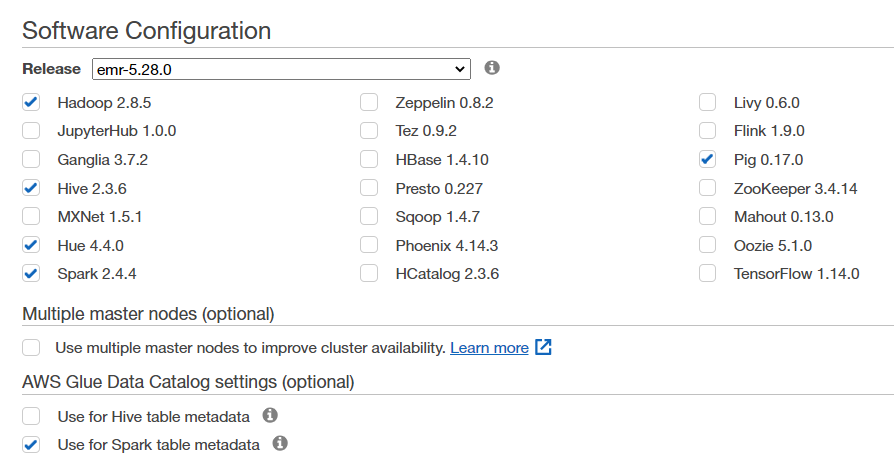
1. Open Amazon EC2 console at <https://console.aws.amazon.com/ec2/v2/>
2. Find Network & Security section on the left navigation pane, choose Key Pairs, Create key pair.
3. Enter key pair name <your-account>
4. Key pair type: RSA. Private key file format: PEM if you use OpenSSH (Mac/Linux), PPK if you use PuTTY (Windows). You can convert from PEM to PPK file using PuTTYgen.
5. Choose Create key pair and remember the location of your PEM/PPK file.

PuTTY and PuTTYgen download link: <https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>

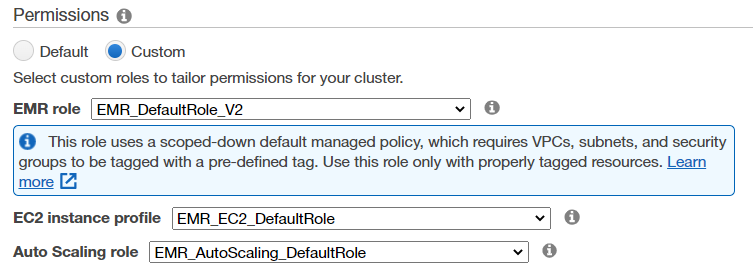
**Step 2**: Create an EMR cluster

Data Profiler for AWS Glue Data Catalog will be run on the cluster via spark-submit. Connect to the master node and execute code via spark-submit

1. Open Amazon EMR cluster console at: <https://console.aws.amazon.com/elasticmapreduce/>
2. Choose Create Cluster, choose Go to advanced options
3. Remember to choose cluster version 5.28 and above with Hadoop and Spark installed. Choose the option to use AWS Glue Data Catalog as table metadata for Spark. **Preferably** choose cluster version 5.x to prevent bugs or deprecated functions with Hadoop 3.x or Spark 3.x.



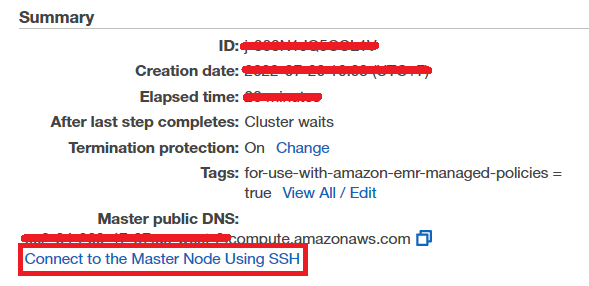
1. Choose next. No changes in hardware configuration
2. Choose next. For Cluster name, enter **<your-account>-dataprofiling-cluster**
3. Choose next. For EC2 key pair, choose the key pair you have created in the first step. In Permissions choose custom and roles as pictures below.



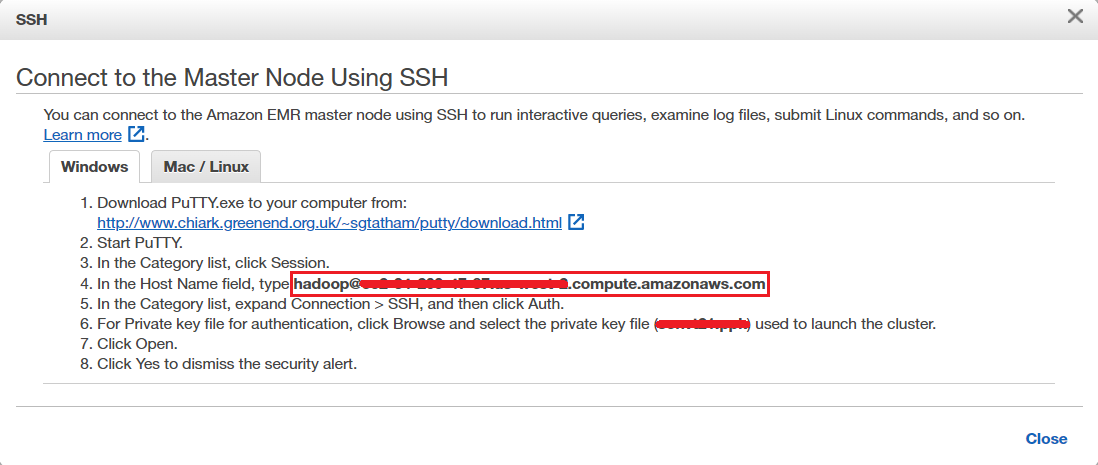
1. Create cluster.

**Step 3**: Connect with Amazon EMR master node via SSH.

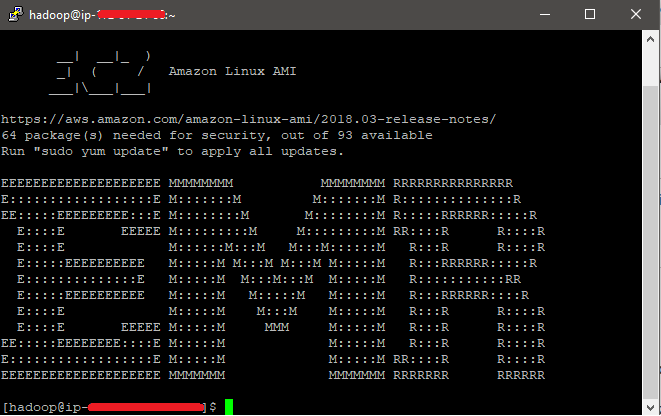
1. Start PuTTY.
2. Open your EMR cluster. Choose connect to the master node using ssh.



1. Do as the following steps. Choose for Windows or Mac/Linux accordingly



1. If you do your steps right, you will be presented with this screen. If it shows connection timed out, please try PuTTY again.



**Step 4**: Downloading the application

Source code can be accessed via this link: <https://github.com/aws-samples/data-profiler-for-aws-glue-data-catalog>.

1. Download Scala Build Tool (sbt) via <https://www.scala-sbt.org/1.x/docs/Setup.html>
2. Open the command line in the folder directory of your source code and type the following command: ***sbt assembly.***
3. By default, the .jar file file is created in the following path, relative to the project root directory: **./target/scala-2.11/data-profiler-for-aws-glue-data-catalog-assembly-1.0.jar**

**Step 5**: Setting up the S3 bucket and copy initial data.

1. Open Amazon S3 console via <https://s3.console.aws.amazon.com/s3/>
2. Choose create bucket. Bucket name: **<your-account>-dataprofiling-bucket**. Region: **us-west-2 (Oregon)**
3. Create a folder jar and upload your .jar file you just created in the previous step.
4. In this lab, we will use data from TLC Trip Record Data via link: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>. Download data files of Yellow taxi trip, Green taxi trip and For-hire vehicle trip in January 2022.
5. Create folder and subfolders **data/raw/nyc-tlc/trip-data-(yellow/green/fhv)** and upload the data files into the subfolders accordingly.
6. Create folder deequ-profiler for the output of the data profiler in the later steps.

**Step 6:** Upload your .jar file to EMR master node  
First, move the .jar file from S3 into HDFS, and then copy from HDFS into EMR master node. The better solution would be to move file directly from your computer into EMR master node via SSH.

1. Open the PuTTY console connected with EMR master node earlier.
2. To check if the .jar exists in the S3 bucket, use this command: **hadoop fs -ls s3://<your-account-name>-dataprofiling-bucket/jar/data-profiler-for-aws-glue-data-catalog-assembly-1.0.jar**
3. Copy file from your S3 bucket to HDFS using this command: **hadoop distcp s3a://<your-account>-dataprofiling-bucket/jar/data-profiler-for-aws-glue-data-catalog-assembly-1.0.jar /jar**
4. Move file from HDFS to EMR master node using this command: **hadoop fs -get /jar/data-profiler-for-aws-glue-data-catalog-assembly-1.0.jar /home/hadoop**

Step 7: Create AWS Glue crawler

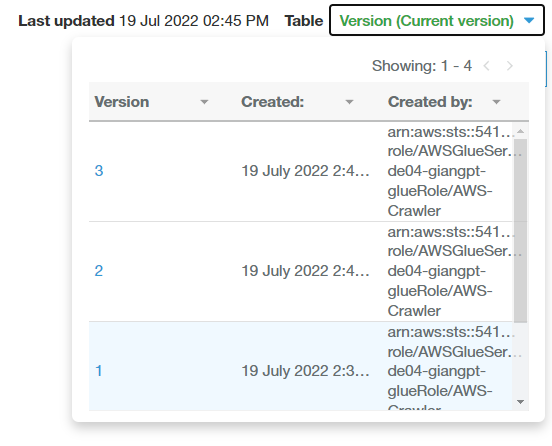
1. Open AWS Glue console via <https://console.aws.amazon.com/glue/>, choose crawlers on the left navigation panel.
2. Choose add crawler. Crawler name: <your-account>-nyc-tlc-db. Next
3. Crawler source type: Data stores. Repeat crawls of S3 data stores: Crawl all folders. Next
4. Data store: S3. Crawl data in specified path. Include path: **s3://<your-account>-dataprofiling-bucket/data/raw/nyc-tlc/**. Next
5. Choose an existing IAM role: **AWSGlueServiceRole-de04-giangpt-glueRole**. Next
6. Run on demand. Next
7. Add a database. Database name: **<your-account>nyctlc** (***Do not use “-“ in AWS Glue database name***). Create
8. Finish creating crawler.

Step 8: Run Data Profiler application using spark-submit.

1. Open the PuTTY console connected to EMR master node.
2. Copy and paste this command :

spark-submit --class awsdataprofiler.DataProfilerForAWSGlueDataCatalog --master yarn --deploy-mode cluster --name data-profiler-for-aws-glue-data-catalog /home/hadoop/data-profiler-for-aws-glue-data-catalog-assembly-1.0.jar --dbName **<your-account>nyctlc** --region us-west-2 --compExp true --statsPrefix DQP --s3BucketPrefix **<your-account>-dataprofiling-bucket /deequ-profiler** --profileUnsupportedTypes true --noOfBins 30 --quantiles 10

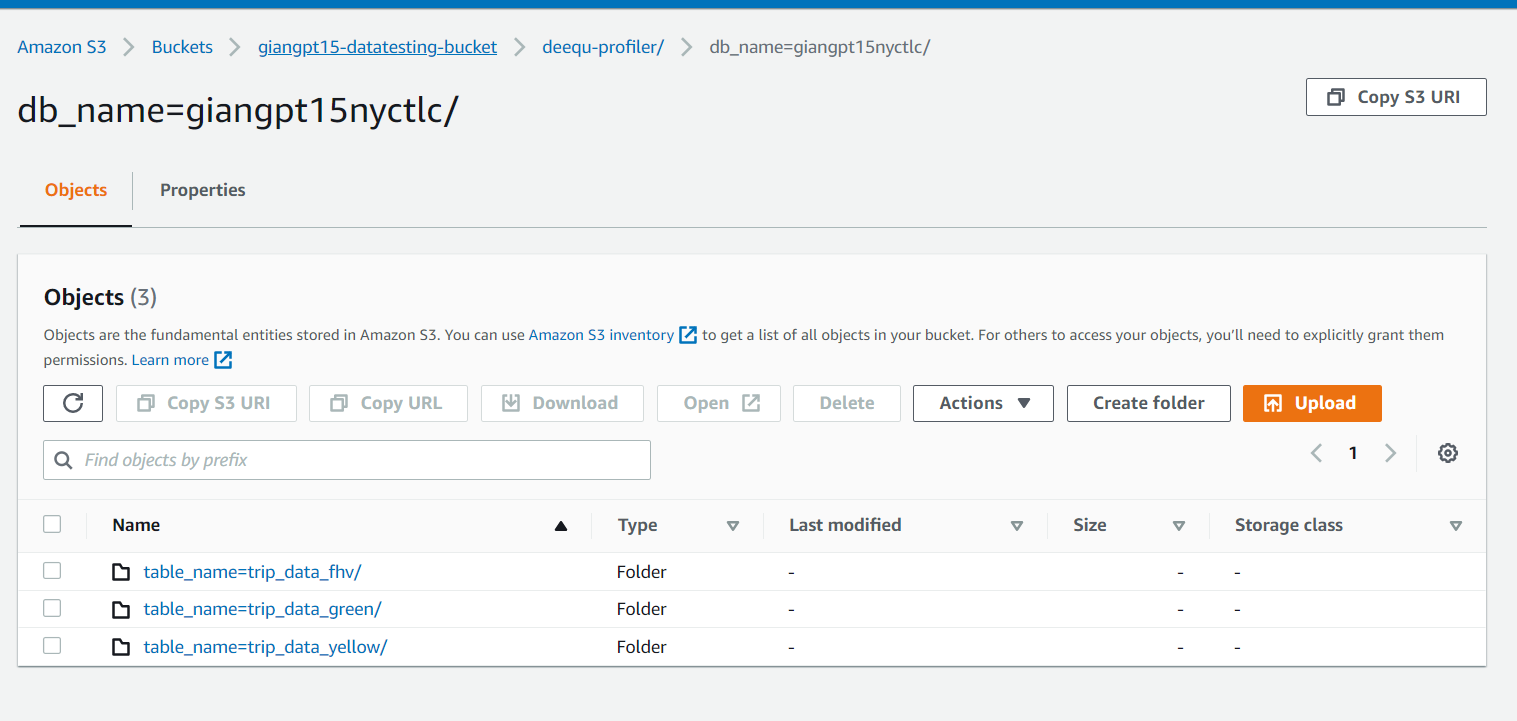
1. Go into databases and tables in the left navigation panel. Choose table trip\_data\_green/yellow/fhr to see if there are any new versions.



1. Keep adding new data from new months into s3 directory, run crawlers, run spark job to get more versions of the tables.

**Step 9** : Profiling information in AmazonS3

You can now also verify that the profiling information was saved in Parquet format in the S3 bucket you specified in the S3bucketPrefix application input parameter. The following screenshot shows the buckets via the Amazon S3 console .

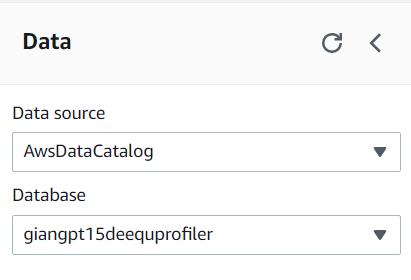


Create second Crawler with the include path s3://<your-account>-datatesting-bucket/deequ-profiler/db\_name=<your-account>nyctlc/ and Run on Command . Create a new database to save the result : <your-account>deequprofiler. Run crawler

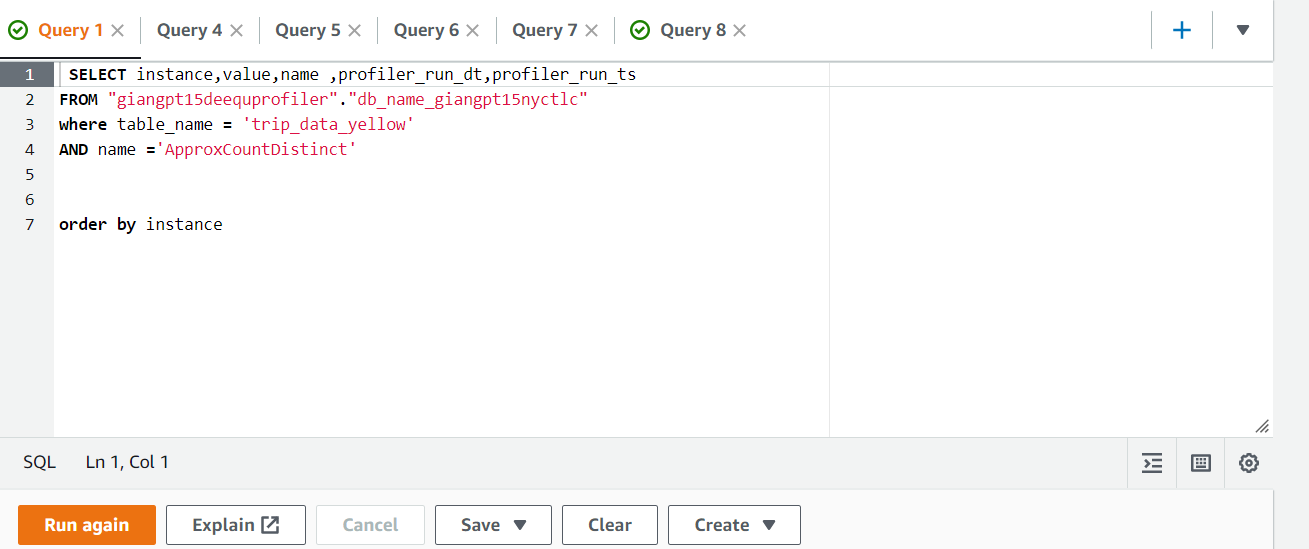
**Step 10** : Using Athena to query data and visualize on Quicksight

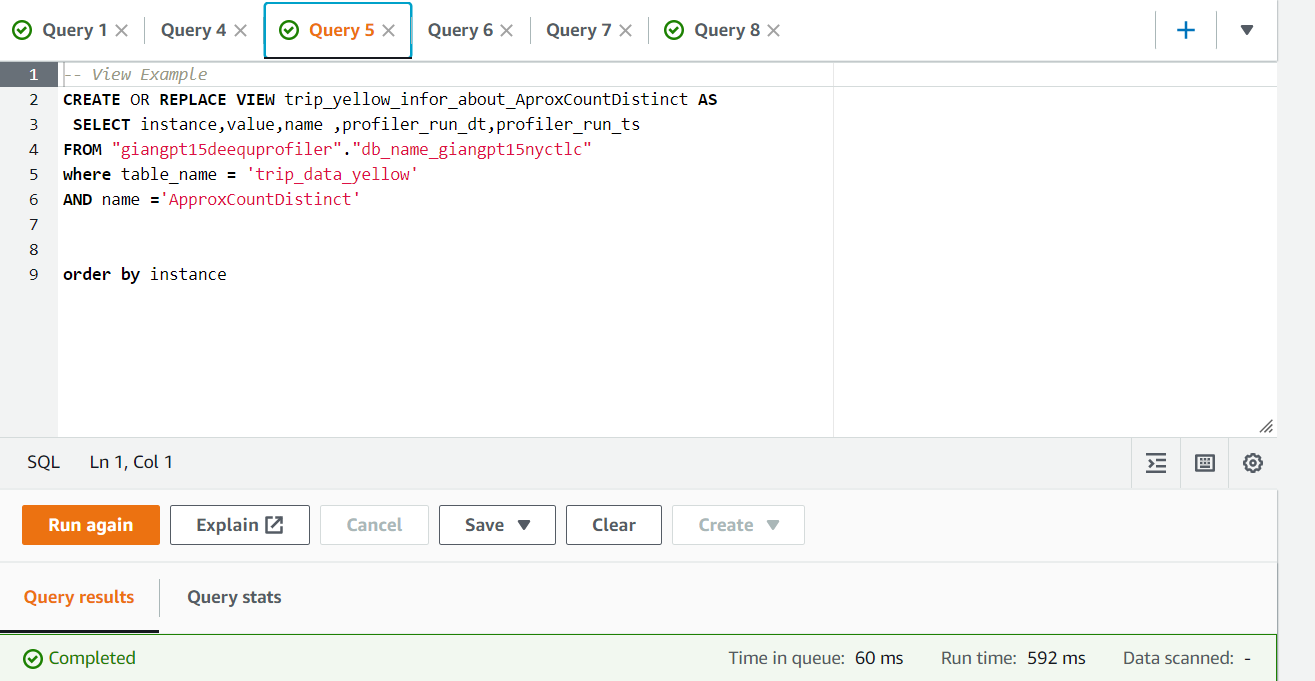
1. Open Athena : <https://us-west-1.console.aws.amazon.com/athena/home?region=us-west-1#/query-editor>

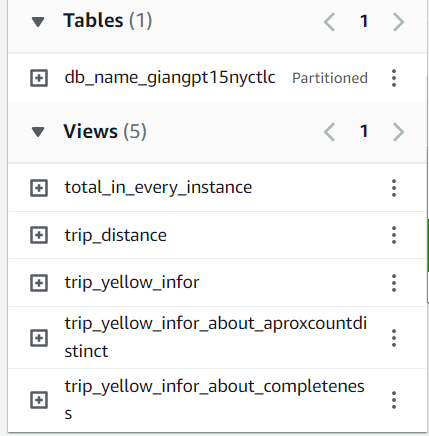
2. Choose Data source : AwsDataCatalog , Database : <your-account>deequprofiler



3. Run your query to extract data you need to visualize . Create a view for using in QuickSight in next step .

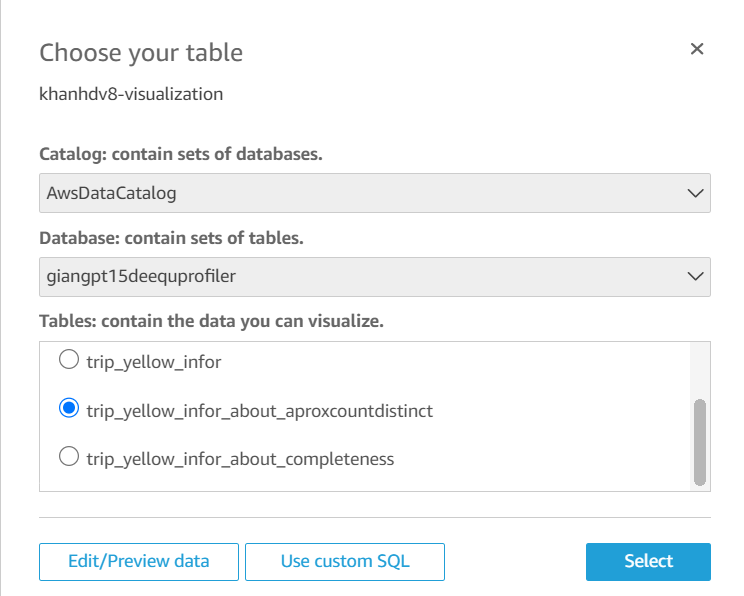


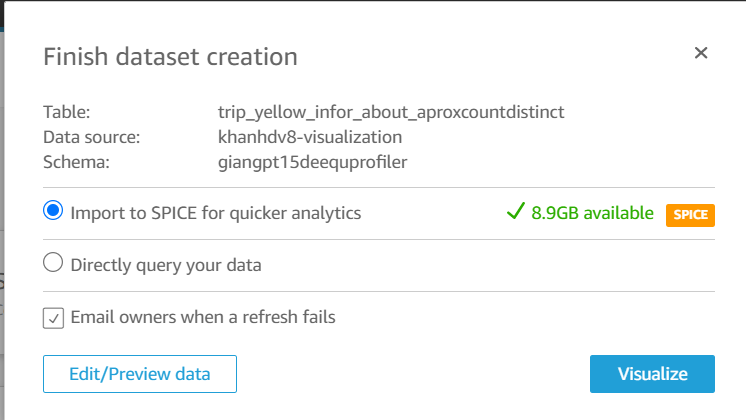




5. Open QuickSight : <https://us-west-2.quicksight.aws.amazon.com/sn/start/analyses>

6. Create New Analysis : Choose New Dataset - > Choose Athena -> Data source name : <your-account> -visualization -> Click Create data source -> In **Database: contain sets of tables** : Select <your-account>deequprofiler -> In **Tables: contain the data you can visualize :** Select 1 view in Athena that you have just created . Click Select - > Click Visualize .





7. Visualize your data :

