

Streaming Data Into Teradata Vantage Using AWS Glue Streaming ETL

Many Teradata customers are interested in integrating Teradata Vantage with Amazon Web Services First Party Services. This guide will help you to stream data into Teradata Vantage using AWS Glue Streaming ETL.

The procedure offered in this guide has been implemented and tested by Teradata. However, it is offered on an as-is basis. Amazon does not provide validation of Teradata Vantage using Glue services.

We encourage your feedback. We want to understand what you found useful, and how we can improve this guide. Please send to [Wenjie.Tehan@teradata.com](mailto:Wenjie.Tehan@teradata.com) and [Shamira.Joshua@teradata.com](mailto:Shamira.Joshua@teradata.com).

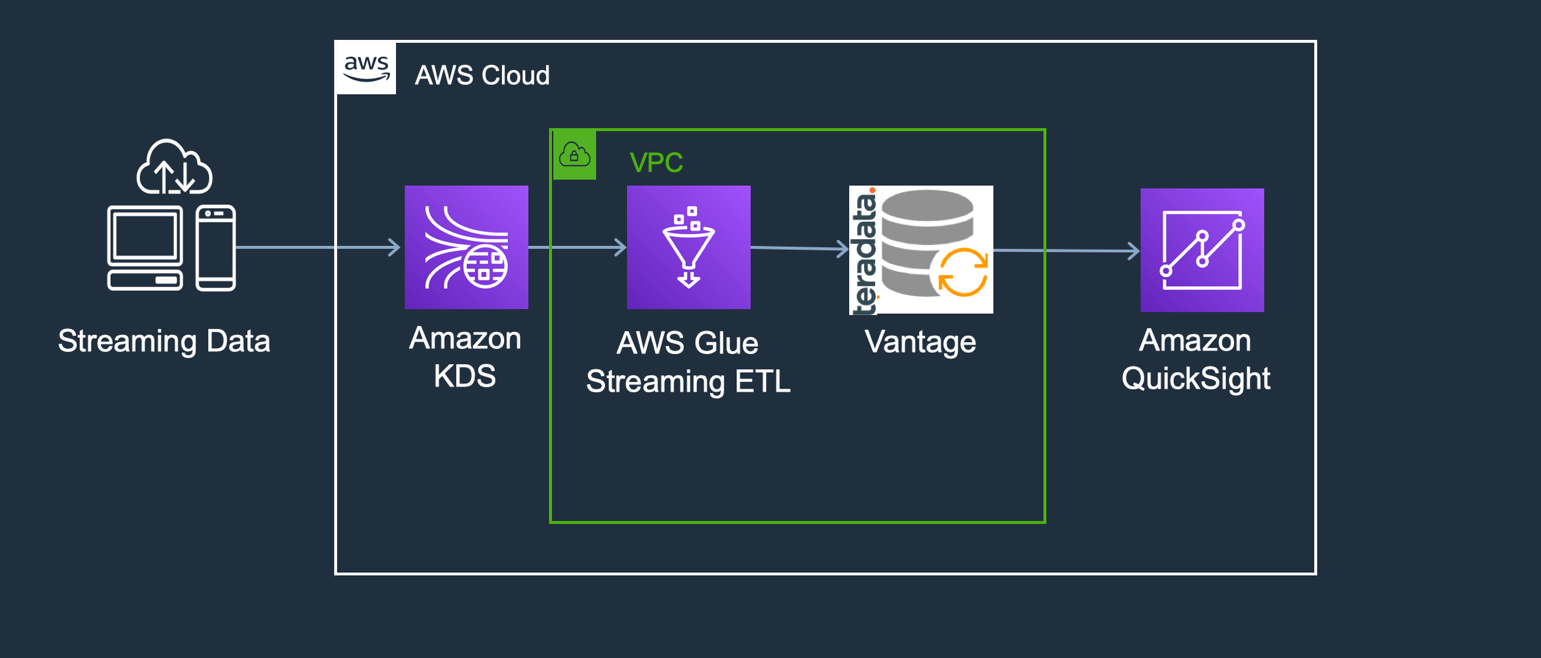
This guide was developed in collaboration with Jobin George, Sr. Partner Solutions Architect at AWS, and Vijay Pawar, Sr. Solutions Architect at AWS. This guide includes content from both Amazon and Teradata product documentation.

[Teradata Corporation](https://www.teradata.com/) is an [AWS Partner Network (APN)](https://aws.amazon.com/partners/) Advanced Technology Partner, specializing in cloud analytics, and has experience using these custom database connectors.

## Overview

This guide describes the procedure to stream data into Teradata Vantage on AWS with AWS Glue Streaming ETL jobs and Amazon Kinesis, and visualize the data with Amazon QuickSight.

The following architecture illustrates the flow of data from Amazon Kinesis, through which it is streamed by AWS Glue to Teradata Vantage where it’s analyzed, and finally to Amazon QuickSight, where it’s displayed. In this tutorial we will be using a simple Lambda function to stimulate a streaming source.



### About AWS Glue

[AWS Glue](https://aws.amazon.com/glue/) is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load your data for analytics. You can create and run an ETL job with a few clicks in the [AWS Management Console](https://aws.amazon.com/console/).

AWS Glue now supports streaming ETL. This feature makes it easy to set up continuous ingestion pipelines that prepare streaming data on the fly and make it available for analysis in seconds. Streaming ETL jobs in AWS Glue run on the Apache Spark Structured Streaming engine, so customers can use them to enrich, aggregate, and combine streaming data, as well as to run a variety of complex analytics and machine learning operations.

Previously, you had to manually construct and stitch together stream handling and monitoring systems to build streaming data ingestion pipelines. Streaming ETL jobs in AWS Glue leverage AWS Glue’s serverless infrastructure to simplify resource management, optimize cost, and enable you to set up continuous ingestion pipelines without writing code - reducing average implementation time from months to days.

### About Amazon Kinesis

[Amazon Kinesis Data Streams](https://aws.amazon.com/kinesis/data-streams/) (KDS) is a massively scalable and durable real-time data streaming service. KDS can continuously capture gigabytes of data per second from hundreds of thousands of sources such as website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events. The data collected is available in milliseconds to enable real-time analytics use cases such as real-time dashboards, real-time anomaly detection, dynamic pricing, and more.

### About Teradata Vantage

Vantage is the modern cloud platform that unifies data warehouses, data lakes, and analytics into a single connected ecosystem.

Vantage combines descriptive, predictive, prescriptive analytics, autonomous decision-making, ML functions, and visualization tools into a unified, integrated platform that uncovers real-time business intelligence at scale, no matter where the data resides.

Vantage enables companies to start small and elastically scale compute or storage, paying only for what they use, harnessing low-cost object stores and integrating their analytic workloads.

Vantage supports R, Python, Teradata Studio, and any other SQL-based tools. You can deploy Vantage across public clouds, on-premises, on optimized or commodity infrastructure, or as-a-service.

See the [documentation](https://docs.teradata.com/home) for more information on Teradata Vantage.

## Prerequisites

You should be familiar with AWS concepts, AWS Glue, Amazon Kinesis, Amazon QuickSight, and Teradata Vantage.

You will need the following accounts and systems:

* AWS account (you can [create a free account](https://aws.amazon.com/free/)),
* Amazon QuickSight account, which requires a subscription, and
* Teradata Vantage with the Advanced SQL Engine 17.0 or higher.

## Procedure

These are the steps to stream data into Teradata Vantage using AWS Glue:

* Launch Teradata Vantage on AWS
* Create a Kinesis table
* Author an AWS Glue streaming ETL job
* Generate streaming data
* Use Amazon QuickSight to visualize the data
* Clean up

## Launch Teradata Vantage on AWS

This step outlines subscribe and deploy Teradata Vantage in your AWS account.

### Create an EC2 key pair

The deployment of Teradata Vantage will require an EC2 key pair.

Create a key pair in your target region. You may call it what you want, but we will call it Teradata.pem in this guide.

### Subscribe to Teradata Vantage Developer Edition

Log into your [AWS account](https://aws.amazon.com/console/).

Find listing for [Teradata Vantage Developer (Free, DIY)](https://aws.amazon.com/marketplace/pp/B06Y4Z54R5?qid=1587421493636&sr=0-1&ref_=srh_res_product_title) in the AWS Marketplace.

Click Continue to Subscribe in the upper right.

Click Accept Terms.

Once you have agreed to the terms, you can now use this AWS Marketplace software in your AWS account.

### Deploy Teradata Vantage

AWS CloudFormation provides a common language for you to model and provision AWS and third-party application resources in your cloud environment.

Click Launch Stack to deploy the Teradata Vantage Developer Edition.

[Launch Stack](https://console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/create/review?stackName=VantageGlueStreamingKinesis&templateURL=https://cloudformation-template-repo.s3.amazonaws.com/Vantage_GlueStreamingKinesis.json)

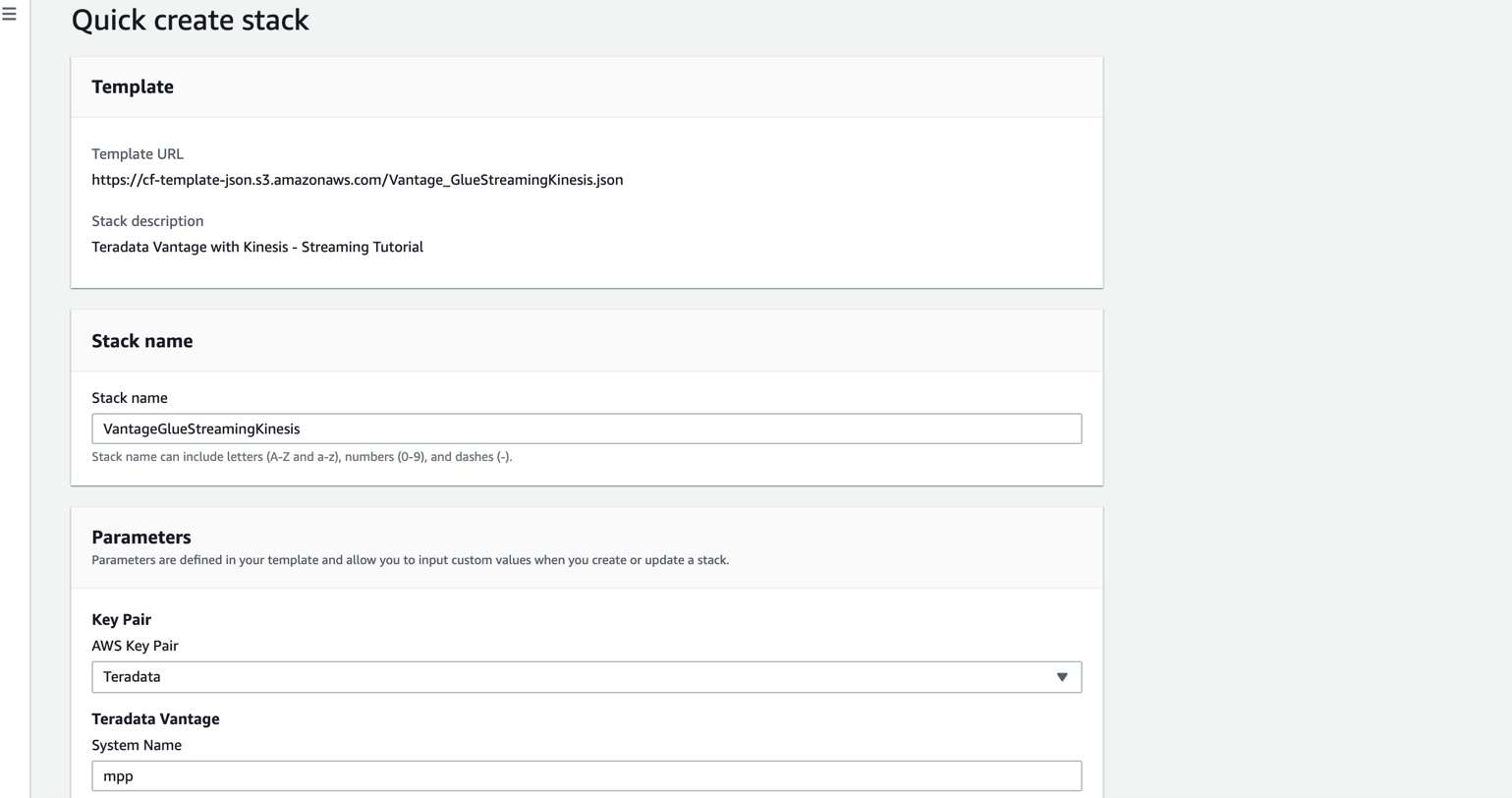
The CloudFormation console page will display.

Select the AWS Key Pair (which refer to as Teradata.pem) from the dropdown.

Leave the other parameters at their default.

Scroll down and acknowledge the IAM resource creation by clicking the checkbox.

Click Create Stack.



The deployment may take up to 20 minutes to complete.

Once the deployment is complete, navigate to the [Stack Output](https://console.aws.amazon.com/cloudformation/home?region=us-east-1#/stacks/outputs?filteringText=&filteringStatus=active&viewNested=true&hideStacks=false&stackId=Upsolver-DevDay-Kinesis) tab and note down the details listed there. These details are needed in future steps.

## Create a Kinesis Table

This step will create a Kinesis catalog table to use as a source for the AWS Glue Streaming Job ETL.

Open the AWS Glue console.

Click on Catalog Tables.

Click on the Add Tables button.

Select Add Tables Manually.

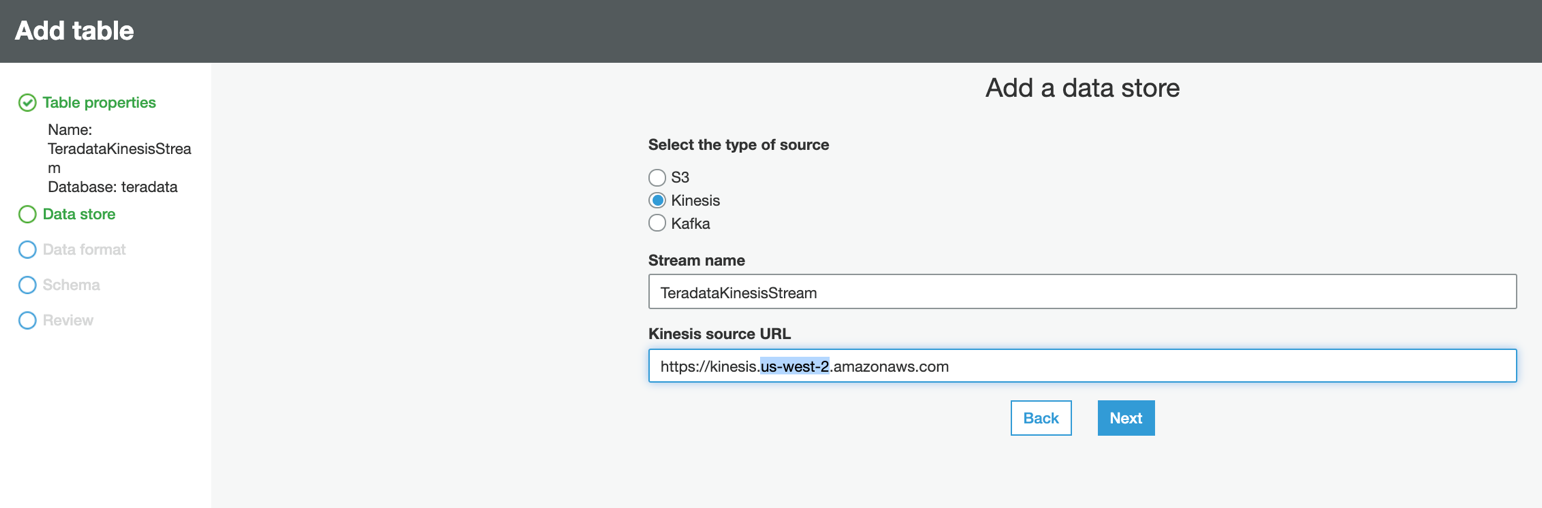
On the next screen, enter the nameTeradataKinesisStream**.**

Choose a database from dropdown. If you don’t have a database created already, refer to [Working with Glue Databases](https://docs.aws.amazon.com/glue/latest/dg/console-databases.html) to create one.

Click Next on the Add a Data Store page.

Select the type of source as Kinesis.

Enter the Stream Nameas TeradataKinesisStreamandKinesis source URLas https://kinesis.${AWS::Region}.amazonaws.com. Replace ${AWS::Region} with your region, such as us‑west‑2.



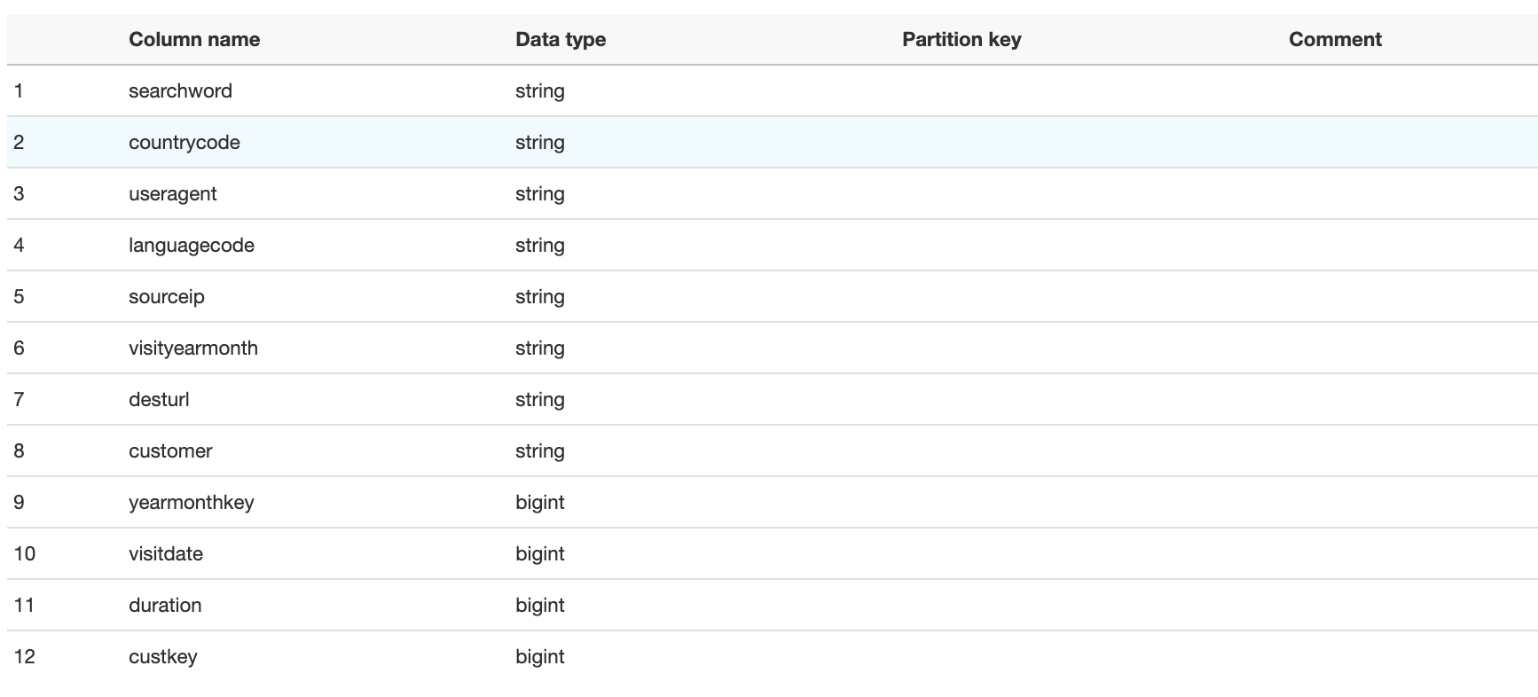
Click Next to continue.

On the following page, select Classification as JSON.

Click Next.

In the define schema screen, click Add Column for each of the following column names and associated data type.

|  |  |
| --- | --- |
| Type | Column Name |
| String | *searchword, countrycode, useragent,languagecode, sourceip, visityearmonth, desturl, customer* |
| BigInt | *yearmonthkey, visitdate, duration, custkey* |



Click Next and review.

Click Finish on next screen to complete Kinesis table creation.

## Author an AWS Glue Streaming ETL Job

### Install the Teradata JDBC driver

AWS Glue needs the Teradata JDBC driver to connect with Vantage. You can download the driver and place it into an Amazon S3 bucket where Glue can access it.

Download the latest [Teradata JDBC driver](https://downloads.teradata.com/download/connectivity/jdbc-driver) for free. If you do not have an account for the Developer section of Teradata.com, you can create an account for free.

Uncompress the tdjdbc4.jar from the downloaded file.

Create an [Amazon S3 bucket](https://docs.aws.amazon.com/AmazonS3/latest/user-guide/create-bucket.html) (or use an existing one).

Upload tdjdbc4.jar to the S3 bucket.

### Create the Glue job

Open the [AWS Glue ETL Jobs](https://console.aws.amazon.com/glue/home?region=us-east-1#etl:tab=jobs) tab.

From the left panel, click Jobs.

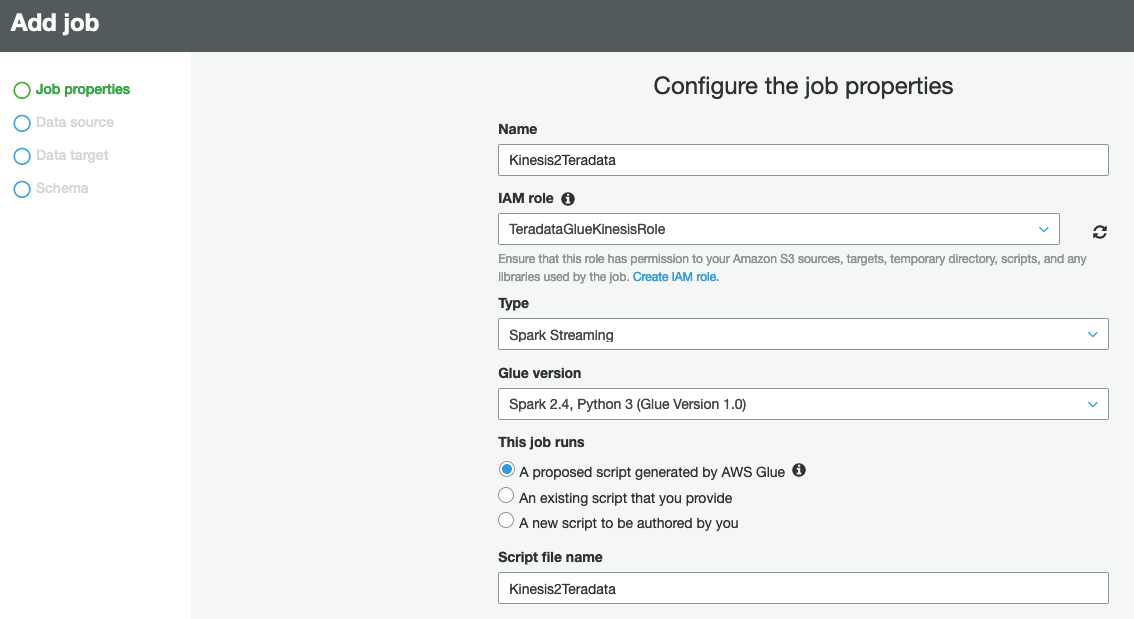
Click the Add Job button.

On the next page, in the Name text box enter Kinesis2Teradata**.**

In the IAM Role dropdown, select TeradataGlueKinesisRole*.*

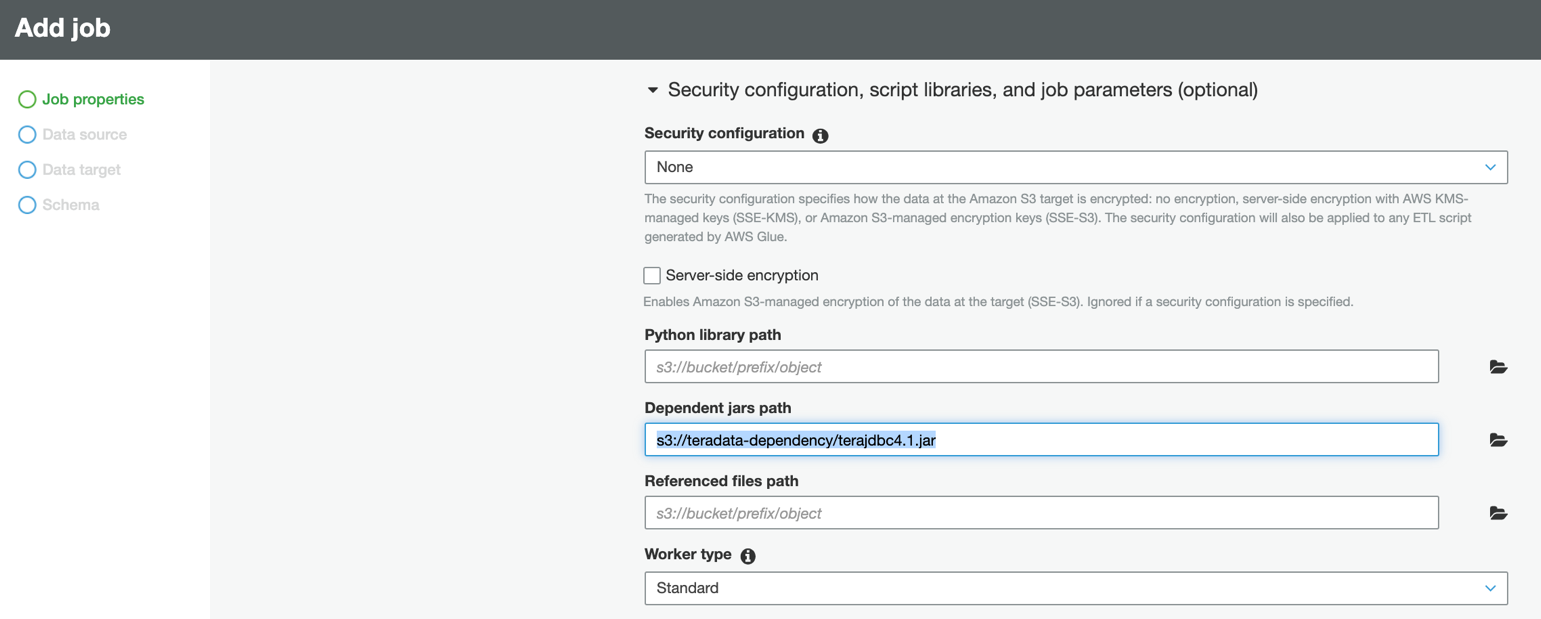
In the Type dropdown, select Spark Streaming.

Select A proposed script generated by AWS Glue for This job runs.



Scroll down to the Security Configuration, script libraries, and job parameters (optional) heading. Click the heading to expand the section.

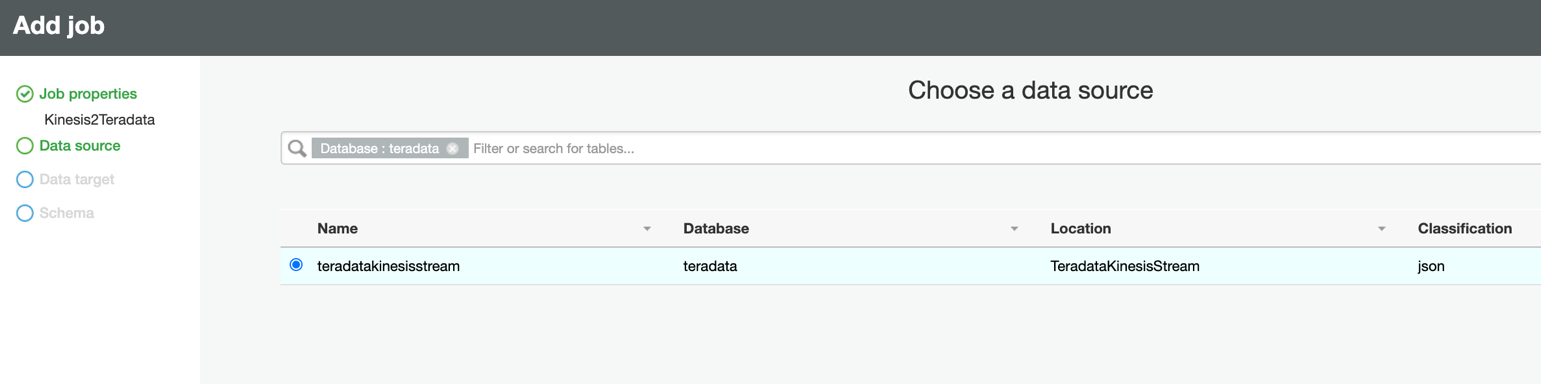
In the Dependent jars path field, enter the path of the S3 bucket and name of the Teradata JDBC driver. The format should be similar to s3://<your-bucket-name>/terajdbc4.jar.



Scroll down and click Next.

The Data Source pane will display.

Select the radio button for the table TeradataKinesisStream, whichyou created above.

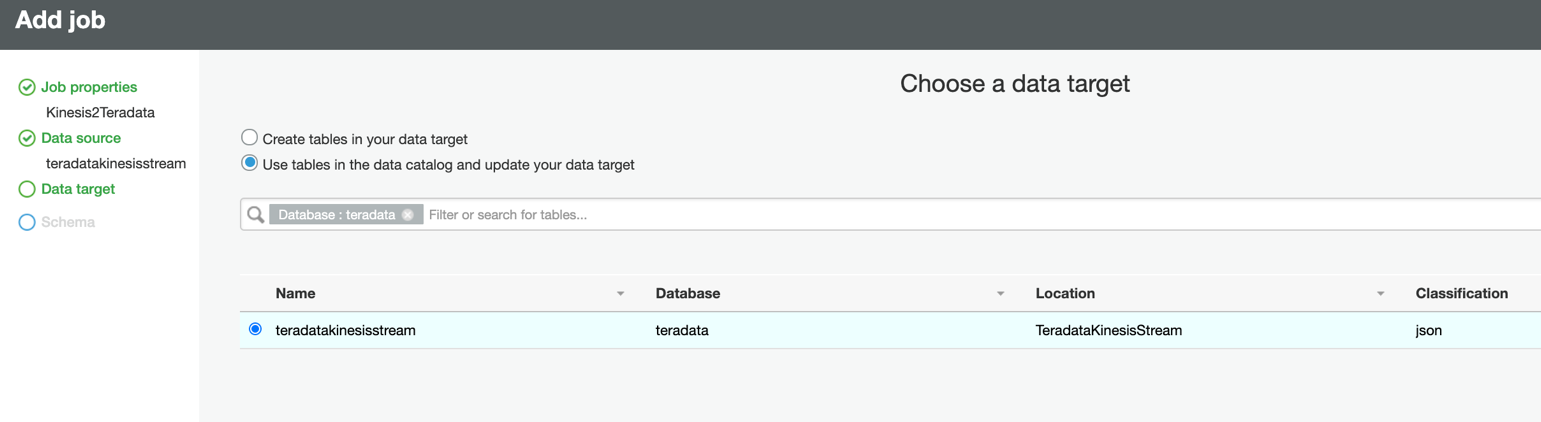


Click Next.

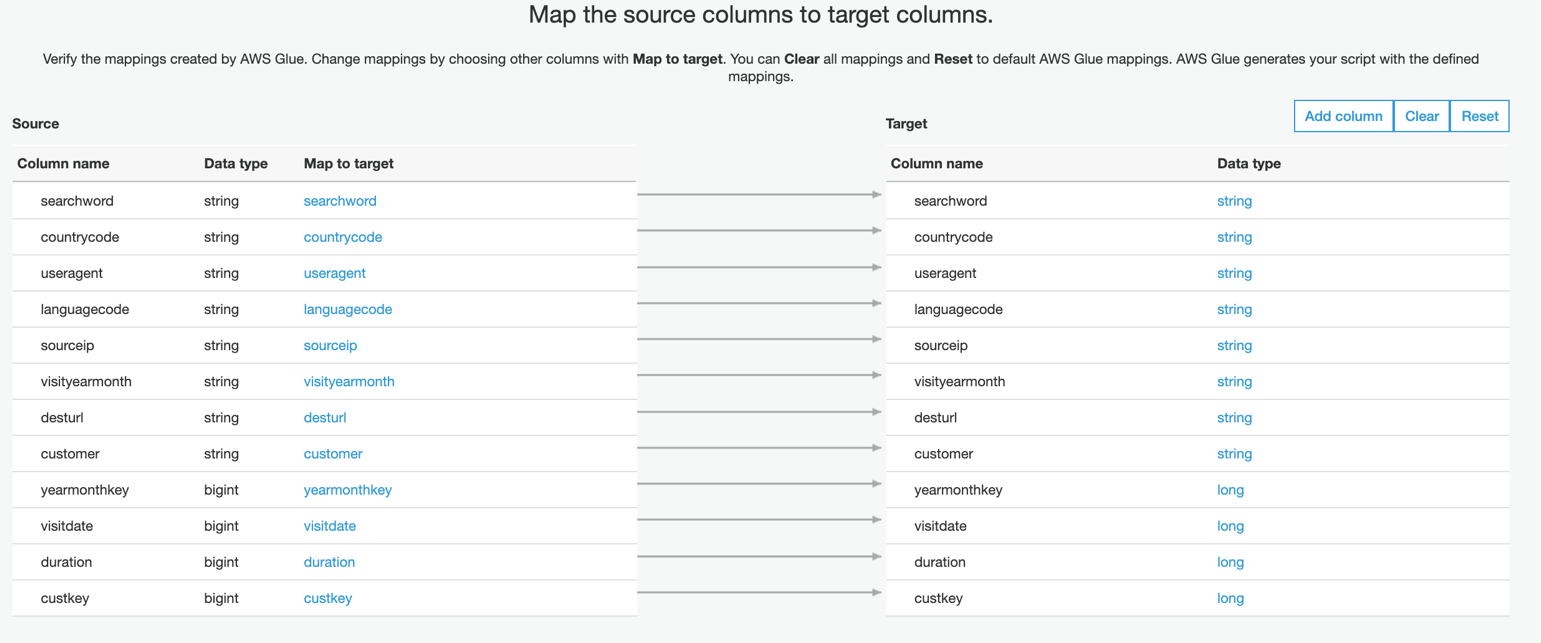
The Data Target pane will display.

Select the same TeradataKinesisStream.

Click Next.



The next window displays the mapping of source columns to target columns. No changes are needed.



Click **Save Job and edit script**.

We will edit the script directly.

On row 33, change windowSize from 100 seconds to 5 seconds.



On row 32, delete the datasink1 row and replace with the text below. Ensure you update your Vantage IP address (or hostname) in the row.

|  |
| --- |
| datasink1 = glueContext.write\_dynamic\_frame.from\_options(frame = apply\_mapping, connection\_type = "jdbc", connection\_options = {"url": "jdbc:teradata://ec2-XX-YYY-ZZZ-AA.us-west-2.compute.amazonaws.com/DATABASE=GlueDB,TMODE=ANSI","driver":"com.teradata.jdbc.TeraDriver","dbtable": "TeradataKinesisStream","database": "GlueDB","user": "GlueUser","password": "aws"}, transformation\_ctx = "datasink1") |

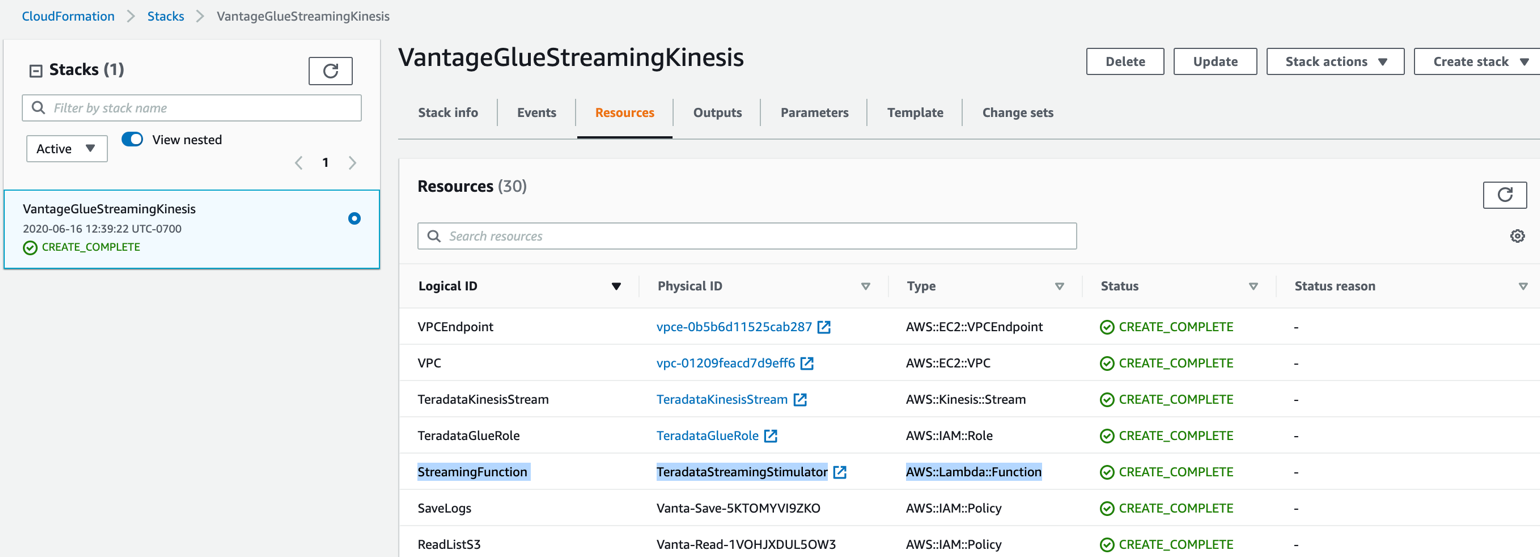
At the top of the page, click Save.

Click Run Job to begin streaming data from Kinesis to Vantage. The job will take a few minutes to start.

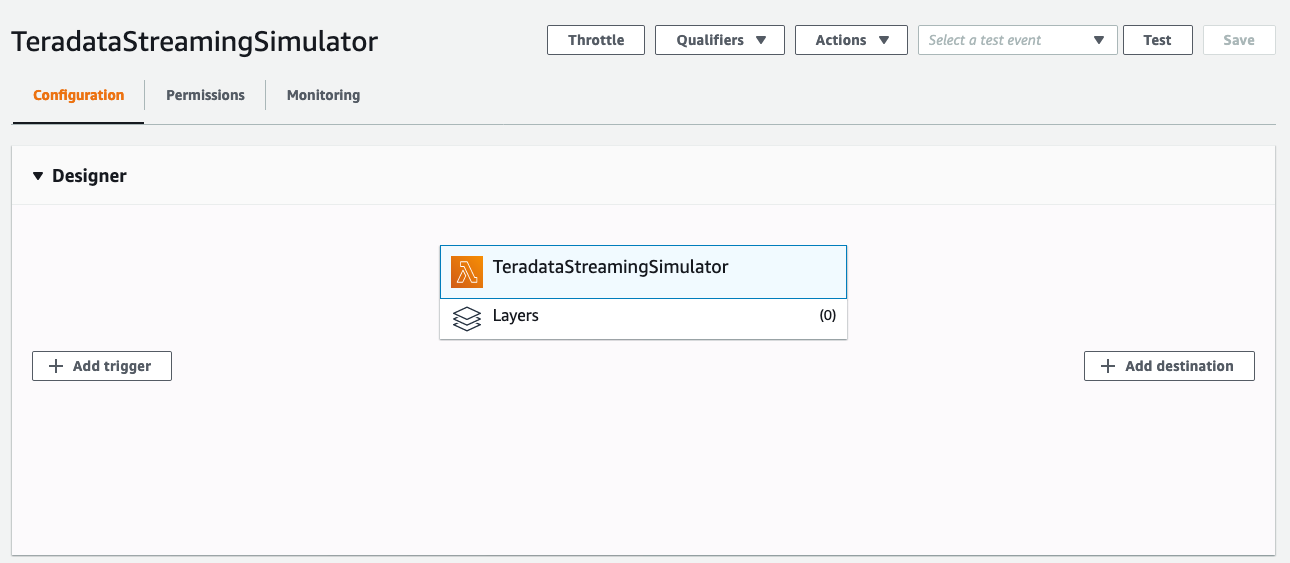
## enerate streaming data

This step will simulate a source data stream to Kinesis, which will forward it on to the Glue stream ETL job.

Navigate back to the [CloudFormation Resources](https://console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/resources?stackId=Vantage-StreamingETL-Blog) page to locate our Lambda Function name, or click on the TeradataStreamingStimulatorphysical ID link to launch the Lambda console.



In the Lambda console, click on the Test button in the upper right to simulate streaming data.



A configure test event pop-up will appear.

On the configure test event pop-up, provide the JSON record shown below, which is formatted with fields listed for simulator to run.

|  |
| --- |
| {  "bucket": "streaming-data-repo",  "stream\_name": "TeradataKinesisStream"  } |

Provide a name for the test event.



Click Save to create test event.

Click Save.

Click Test again to launch the similuator to stream data into the Kinesis Stream.

Once clicked, the simulator will run for two minutes before it times out with an error. (You can adjust the timeout in the Lambda configuration. The two minute threshold is to stop resource consumption.)

## Use Amazon QuickSight to Visualize the Data

In this step, we will connect Amazon QuickSight to Vantage and visualize the streamed data.

Open Amazon QuickSight.

Create a new dataset.

From the list of data sets, select Teradata.

A pop-up window will appear. Enter a name in the Data source name field.

In the Database server field, enter the DNS name of the Vantage instance.

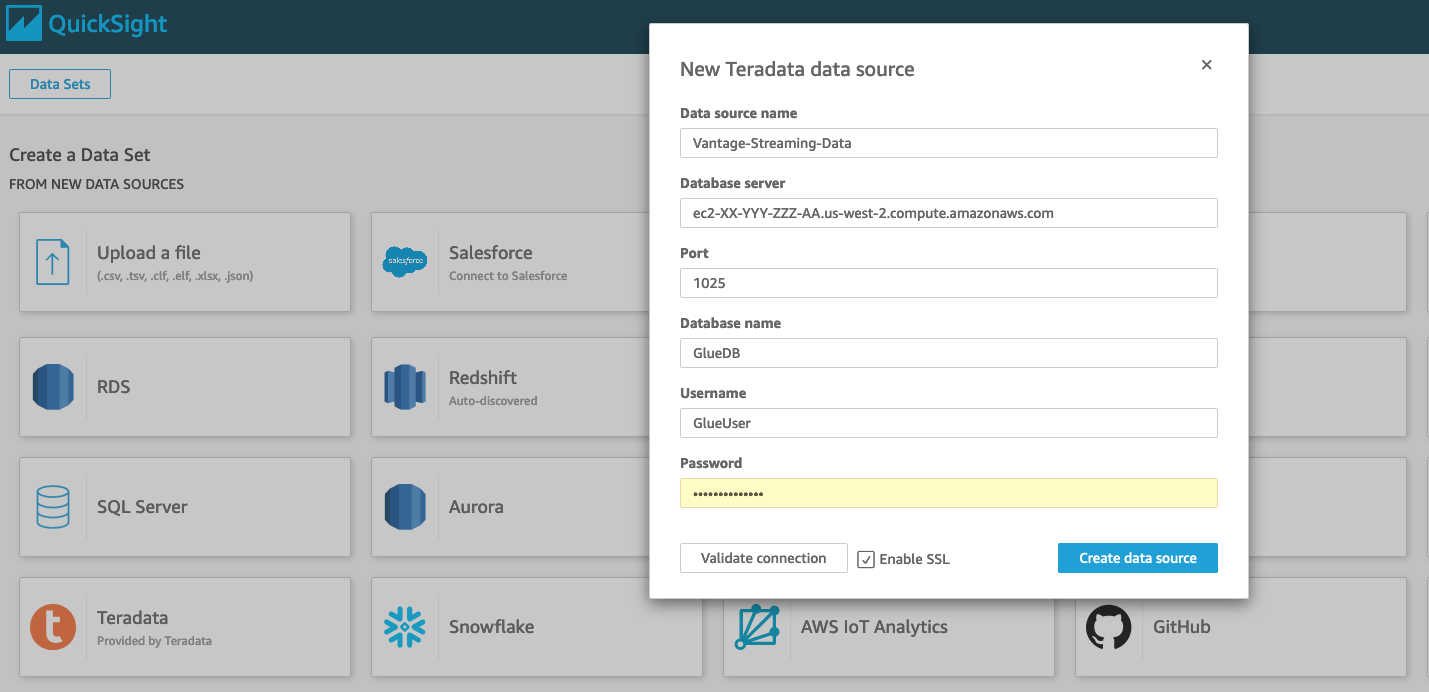
Enter 1025 as the Port.

Enter the database name, username, and password credentials in the following fields.

Click Validate Connection to check the correctness of the parameters.

A green checkmark will appear once the connection has been validated.

Click Create data source.



Amazon QuickSight will identify the tables in Vantage.

From the Choose Your Table, select TeraTopic.

A pop-up window will appear.

Select Use Custom SQL.

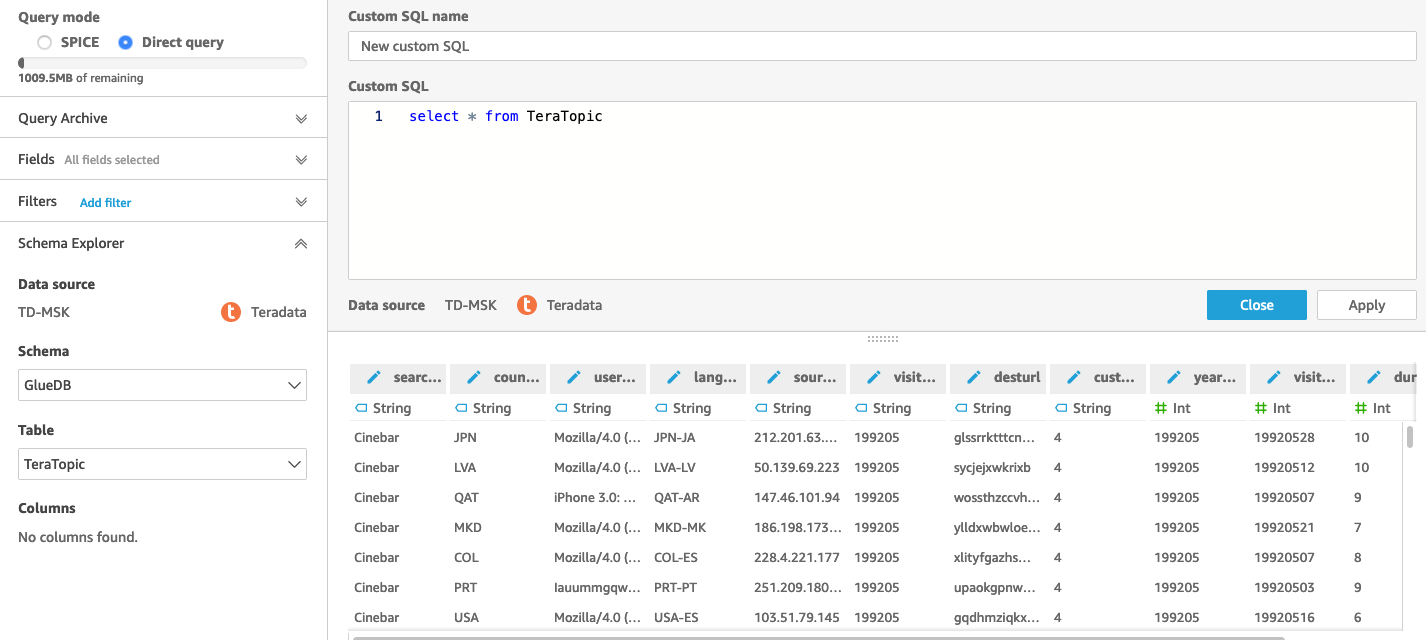
Enter a name for the query.

Enter the following as the query in the Custom SQL box.

|  |
| --- |
| select \* from TeraTopic |

Click Confirm query.

Click Edit/Preview data. The data will appear.



Change the data type of the Dates fields as required, or you may create calculated fields to start visualizing the data using QuickSight.

To learn more about [creating an AutoGraph visualization](https://docs.aws.amazon.com/quicksight/latest/user/example-create-an-analysis.html#example-create-autograph-visual) in the Amazon QuickSight, see the [documentation](https://docs.aws.amazon.com/quicksight/latest/user/autograph.html).

## Clean up

You can avoid incurring additional charges caused by resources created as part of this guide.

Delete the AWS CloudFormation stack by going to the CloudFormation console and deleting the stack that was created.

Stop the Glue jobs that were created and delete the connections, databases, tables, and jobs.

17095 Via Del Campo, San Diego, CA 92127      [Teradata.com](http://www.teradata.com)

Teradata and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions, and operations described herein may not be marketed in all parts of the world. Consult your Teradata representative or Teradata.com for more information.

Copyright © 2021 by Teradata Corporation    All Rights Reserved.    Produced in U.S.A.7.21 EBxxxx