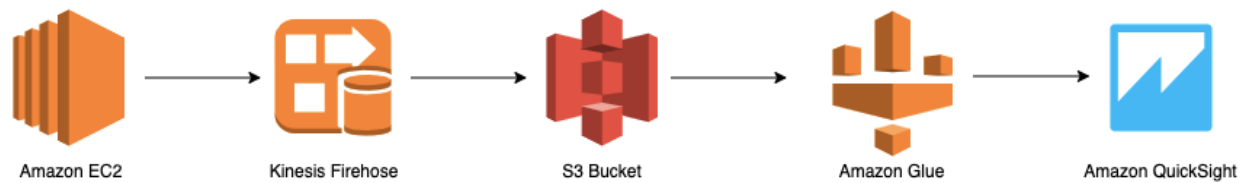


Hands On Lab Big Data Immersion Day



Ingestão, Processamento e Pesquisa de Dados com ferramentas AWS



Iremos criar uma instância EC2 e realizar a instalação do servidor Web Apache. Em seguida iremos realizar a instalação e configuração do Agente do Amazon Kinesis para direcionar logs de acesso do site para um bucket S3.

Em seguida iremos criar um Crawler do Glue para ler os dados do bucket S3 e gerar uma tabela, onde iremos realizar algumas consultas usando SQL.

E por fim, iremos criar um dashboard usando o Amazon QuickSight para visualização destes dados.

Criando a Instância EC2

Verifique antes de tudo se você está na região N. Virginia:



Se não estiver, altere para esta região.

Abra a console da EC2.

Clique em:



Em **Step 1: Choose an Amazon Machine Image (AMI)**, Selecione a primeira imagem do Amazon Linux e clique em **Select**:

Em **Step 2: Choose an Instance Type**, selecione **t2.micro** e clique em **Next: Configure Instance Details**.

Em **Step 3: Configure Instance Details**, não altere nada e clique em **Next: Add Storage**.

Em **Step 4: Add Storage**, não altere nada e clique em **Next: Add Tags**

Em **Step 5: Add Tags**, não altere nada e clique em **Next: Configure Security Group**

Em **Step 6: Configure Security Group**, na opção **Assign a security group**: deixe marcado a opção: **Create a new security group** marcado.

Em **Security Group Name**: coloque **web-secgroup**

E **Description**: adicione "**Security Group for Web Servers**"

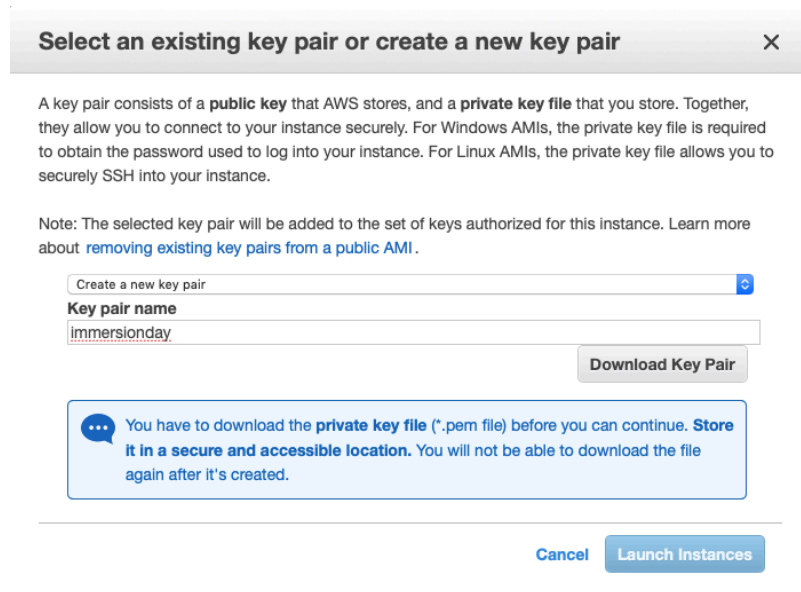
Mais abaixo clique no botão **Add Rule** e deixe preenchido conforme abaixo:

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ	Description ⓘ	
SSH	TCP	22	Custom 0.0.0.0/0	e.g. SSH for Admin Desktop	✕
HTTP	TCP	80	Custom 0.0.0.0/0	e.g. SSH for Admin Desktop	✕
Add Rule					

Clique no botão **Review and Launch**.

Em **Step 7: Review Instance Launch** apenas clique em **Launch**.

Na próxima tela selecione **Create a new key pair** e preencha conforme abaixo:



Select an existing key pair or create a new key pair ✕

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Create a new key pair

Key pair name
immersionday

Download Key Pair

You have to download the **private key file** (*.pem file) before you can continue. **Store it in a secure and accessible location.** You will not be able to download the file again after it's created.

Cancel Launch Instances

Clique em **Download Key Pair** e salve o arquivo immersionday.pem em algum diretório. Este arquivo será necessário para conectar na instância posteriormente.

E finalmente clique em **Launch Instances**.

Depois de lançar a instância clique em **View Instances** e selecione a instância que acabamos de lançar. Copie o campo Public DNS (IPv4)

No Microsoft Windows será necessário utilizar o puttygen para converter o formato da chave de .pem para .ppk e utilizar o putty para conectar à instância utilizando o DNS Público utilizando o usuário ec2-user e apontando a chave ppk.

Em sistemas *nix ajuste as permissões conforme abaixo:

```
$ chmod 600 immersionday.pem
```

E conecte na instância:

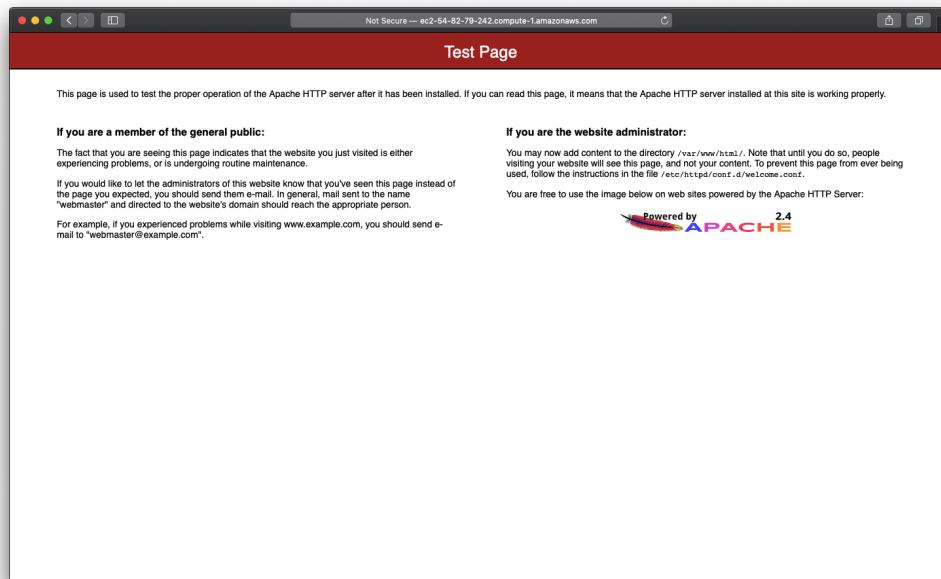
```
$ ssh ec2-user@ec2-54-82-79-242.compute-1.amazonaws.com -i immersionday.pem
```

Uma vez logado na instância, torne-se usuário root com “sudo su -“

Iremos instalar o servidor web, habilitar o início automático o boot e realizar um start no serviço com os comandos abaixo:

```
# yum install -y httpd
# for i in enable start status; do systemctl $i httpd ; done
```

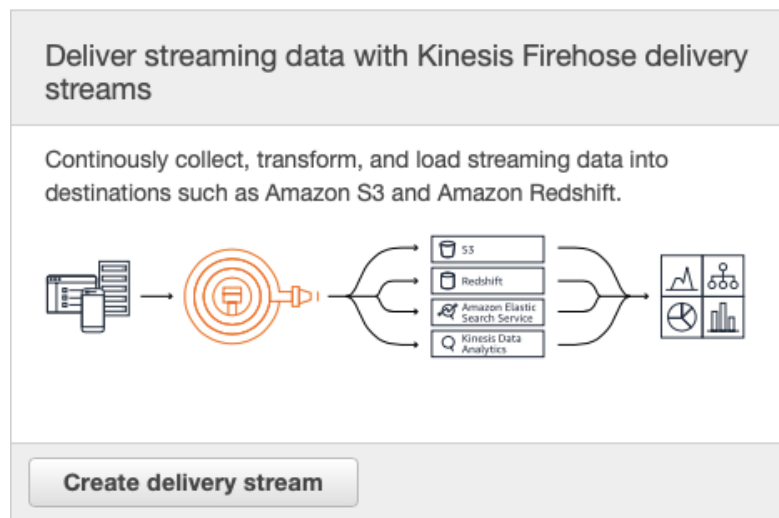
Verifique se o site default está acessível via browser utilizando o DNS público



Configurando o Kinesis Firehose

Na console da AWS abra o serviço do Kinesis Streams e clique no botão **Get started**.

Na próxima tela clique em **Create delivery stream**:



Em **Step 1: Name and source** no campo delivery stream name coloque:

log-webserver

Em Source deixe conforme abaixo:

- Source*** ☒ **Direct PUT or other sources**
Choose this option to send records directly to the delivery stream, or to send records from AWS IoT, CloudWatch Logs, or CloudWatch Events.
- ☐ **Kinesis stream**

E clique em **Next**.

Em **Step 2: Process Records** deixe conforme a imagem abaixo e clique em Next:

Transform source records with AWS Lambda

To return records from AWS Lambda to Kinesis Firehose after transformation, the Lambda function you invoke must be compliant with the required record transformation output model. [Learn more](#)

Record transformation* ☒ Disabled
☐ Enabled

Convert record format

Data in Apache parquet or Apache ORC format is typically more efficient to query than JSON. Kinesis Data Firehose can convert your JSON-formatted source records using a schema from a table defined in [AWS Glue](#). For records that aren't in JSON format, create a Lambda function that converts them to JSON in the [Transform source records with AWS Lambda](#) section above. [Learn more](#)

Record format conversion* ☒ Disabled

☐ Enabled

If record format conversion is enabled, Firehose can deliver data to Amazon S3 only. Record format conversion will be configured using the OpenX JSON SerDe. For other options use the [AWS CLI](#).

* Required

[Cancel](#)

[Previous](#)

[Next](#)

Em **Step 3: Choose destination**, Selecione S3.

Mais abaixo em **S3 destination** clique no botão **Create new**. Adicione um nome qualquer, lembrando que deve ser um nome único e deixe a região como **US East (N. Virginia)**:

Create S3 bucket

A bucket is a container for objects stored in Amazon S3. [Learn more](#)

S3 bucket name

logwebserver-im-tribanco

Region

US East (N. Virginia)

[Cancel](#)

[Create S3 bucket](#)

Clique em **Create S3 bucket**.

Clique em **Next**.

Em **Step 4: Configure settings**, deixe a seção **Buffer conditions** conforme abaixo:

S3 buffer conditions

Firehose buffers incoming records before delivering them to your S3 bucket. Record delivery will be triggered once either of these conditions has been satisfied. [Learn more](#)

Buffer size* MB

Specify a buffer size between 1-128 MB

Buffer interval* seconds

Specify a buffer interval between 60-900 seconds

Na seção **IAM Role** clique no botão **Create new or choose**, verifique se está conforme a imagem abaixo e clique em **Allow**.

▼ Hide Details

Role Summary ?

Role Description Provides access to AWS Services and Resources

IAM Role

Role Name

► View Policy Document

Em seguida clique em **Create a delivery stream**

<div>Create delivery stream Test with demo data Delete</div> <div>Filter Firehose delivery streams</div>						
< 1 >						
Name	Status	Created	Source	Record transformation	Destination	
<input type="radio"/> log-webserver	Active	2019-07-17T22:59-0300	Direct PUT and other sources	Disabled	Amazon S3 logwebserver-im-tribanco	

Configurando o agente do Kinesis Firehose

Logue na instância e instale o agente do Kinesis para que ele envie os dados de log do servidor web direto para o bucket S3 especificado no próprio Kinesis.

Instale o agente com o comando abaixo:

```
# yum install -y https://s3.amazonaws.com/streaming-data-agent/aws-kinesis-agent-latest.amzn1.noarch.rpm
```

Em seguida, apague o conteúdo do arquivo `/etc/aws-kinesis-agent` e adicione o conteúdo abaixo:

```
{
  "awsAccessKeyId": "AKIAXXXXXXXXXXXXXXXX",
  "awsSecretAccessKey": "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX",
```



```

"cloudwatch.emitMetrics": true,
"kinesis.endpoint": "",
"firehose.endpoint": "firehose.us-west-2.amazonaws.com",

"flows": [
  {
    "filePattern": "/var/log/httpd/access_log",
    "deliveryStream": "log-webserver"
  }
]
}

```

Substituindo o AccessKeyId e SecretAccessKey pelas chaves do usuário o qual possui permissão full no S3.

Certifique que no nome em deliveryStream é o mesmo nome criado no Kinesis Streams.

Em seguida ajuste as permissões para que o agente consiga ler o arquivo solicitado, neste caso o /var/log/httpd/access_log, caso contrário o serviço não conseguirá subir:

```

# chmod 755 /var/log/httpd
# setfacl -m u:aws-kinesis-agent-user:rwX /var/log/httpd/access_log

```

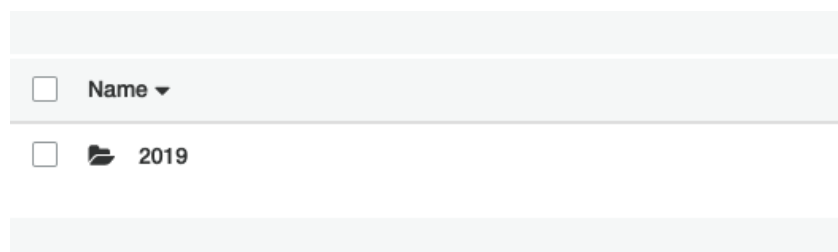
E inicie o agente do kinesis:

```
# systemctl start aws-kinesis-agent
```

Faça alguns acessos no site web para gerar tráfego e aguarde alguns segundo para que os dados sejam enviados para o S3.

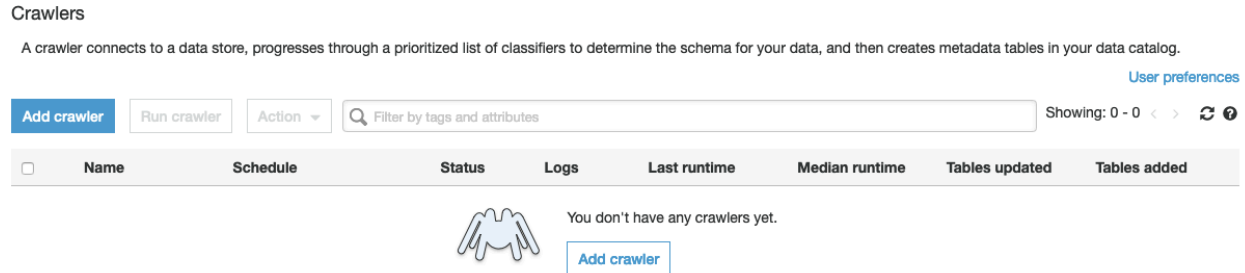
Pode acompanhar o log através de um tail no arquivo /var/log/aws-kinesis-agent/aws-kinesis-agent.log.

Depois de algum tempo verifique no bucket configurado os dados que o Kinesis enviou para o S3:



Configurando o Glue

Na console do Glue, vá em Crawlers e clique no botão **Add Crawler**:



Na tela de **Crawler info**, dê um nome qualquer para o crawler e clique em **Next**:

Em **Crawler source type**, selecione **Data stores** e clique em **Next**.

Na próxima tela deixe conforme abaixo, selecionando o Bucket S3 criado anteriormente:

The screenshot shows the 'Choose a data store' configuration screen. At the top, there's a dropdown menu with 'S3' selected. Below this, under 'Crawl data in', there are two radio buttons: 'Specified path in my account' (which is selected) and 'Specified path in another account'. Under 'Include path', there's a text input field containing 's3://logwebserver-im-tribanco'. To the right of the input field is a folder icon. Below the input field, there's a note: 'All folders and files contained in the include path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.' At the bottom, there's a section for 'Exclude patterns (optional)' with a right-pointing arrow. At the very bottom, there are 'Back' and 'Next' buttons.

Clique em **Next**:

Em **Add another data store**, deixe selecionado **No** e clique em **Next**.

Na próxima tela deixe conforme a imagem abaixo:

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. [Learn more](#)

☐ Update a policy in an IAM role
☐ Choose an existing IAM role
☒ Create an IAM role

IAM role ⓘ

AWSGlueServiceRole-

To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named **"AWSGlueServiceRole-rolename"** and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://logwebserver-im-tribanco

You can also create an IAM role on the [IAM console](#).

E clique em **Next**.

Em **Frequency** deixe **Run on demand** e clique em **Next**:

Em Database clique no botão **Add database**, e coloque o nome database_access_webserver e clique em **Create**.

Clique em **Next** e em seguida em **Finish**.

Assim que o crawler for criado, selecione o checkbox e clique no botão **Run crawler**.

<input type="button" value="Add crawler"/>	<input type="button" value="Run crawler"/>	<input type="button" value="Action ▼"/>	<input type="text" value="Filter by tags and attributes"/>
<input checked="" type="checkbox"/>	Name	Schedule	Status
<input checked="" type="checkbox"/>	mycrawler		Ready

Após a execução do crawler, verifique que foi criada uma nova database e uma nova tabela no menu à esquerda.

Usando o Athena

Vá para a console do Athena, e em **New Query 1** faça um select especificando a database e tabela criada no Glue:

```
SELECT * FROM "database_access_webserver"."logwebserver_im_tribanco" limit 10;
```

Clique no botão **Run query**:

Os resultados da consulta SQL irão aparecer abaixo:

Results											
	clientip	ident	auth	timestamp	verb	request	httpversion	response	bytes	referrer	agent
1	209.17.96.218	-	-	18/Jul/2019:02:08:37 +0000	GET	/	1.0	403	3630	-	Mozilla/5.0 (compatible; Ni
2	187.72.233.203	-	-	18/Jul/2019:02:01:20 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
3	187.72.233.203	-	-	18/Jul/2019:02:01:20 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
4	187.72.233.203	-	-	18/Jul/2019:02:01:20 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
5	187.72.233.203	-	-	18/Jul/2019:02:01:21 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
6	187.72.233.203	-	-	18/Jul/2019:02:01:21 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
7	187.72.233.203	-	-	18/Jul/2019:02:01:21 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
8	187.72.233.203	-	-	18/Jul/2019:02:01:21 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
9	187.72.233.203	-	-	18/Jul/2019:02:01:21 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
10	187.72.233.203	-	-	18/Jul/2019:02:01:22 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int
11	187.72.233.203	-	-	18/Jul/2019:02:01:22 +0000	GET	/	1.1	403	3630	-	Mozilla/5.0 (Macintosh; Int

QuickSight

Vá na console do QuickSight:



Your AWS Account is not signed up for QuickSight. Would you like to sign up now?

AWS Account

107100542058

Sign up for QuickSight

To access QuickSight with a different account, [log in](#) again.

Clique no botão **Sign up for Quicksight**, na próxima tela deixe preenchido conforme abaixo, no campo account name adicione seu nome em caixa baixa e sem espaços e em Notification email address coloque seu email:

QuickSight region

Select a region.

US East (N. Virginia)

US East (N. Virginia)

QuickSight account name

Enter a unique QuickSight account name

You will need this for you and others to sign in.

Notification email address

Enter account notification email address

For QuickSight to send important notifications.

☒ Enable autodiscovery of data and users in your Amazon Redshift, Amazon RDS, and AWS IAM services.

☒ Amazon Athena

Enables QuickSight access to Amazon Athena databases

Please ensure the right Amazon S3 buckets are also enabled for QuickSight.

☒ Amazon S3 (3 buckets selected)

Enables QuickSight to auto-discover your Amazon S3 buckets

[Choose S3 buckets](#)

☐ Amazon S3 Storage Analytics

Enables QuickSight to visualize your S3 Storage Analytics data

☐ AWS IoT Analytics

Enables QuickSight to visualize your IoT Analytics data

Finish

Na tela seguinte clique em **New analysis**, em depois em **New data set**.

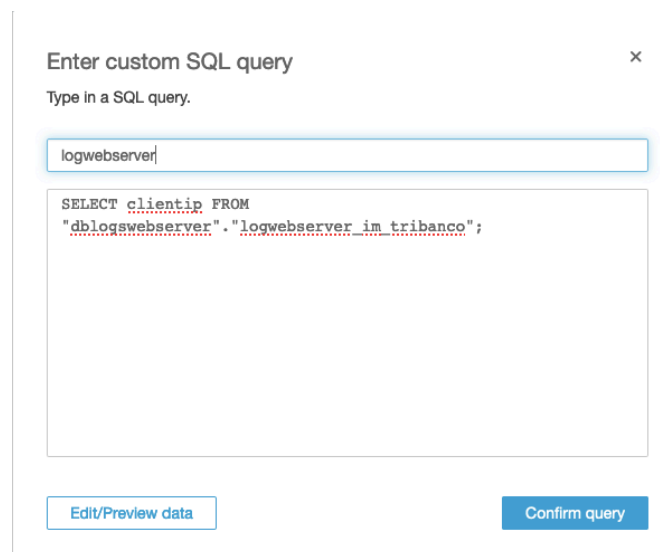
Na próxima tela selecione a origem dos dados para o gráfico, neste caso **Athena**.

Em seguida em **Data source name** coloque logwebserver e valide o SSL. Clique em seguida em **Create data source**.



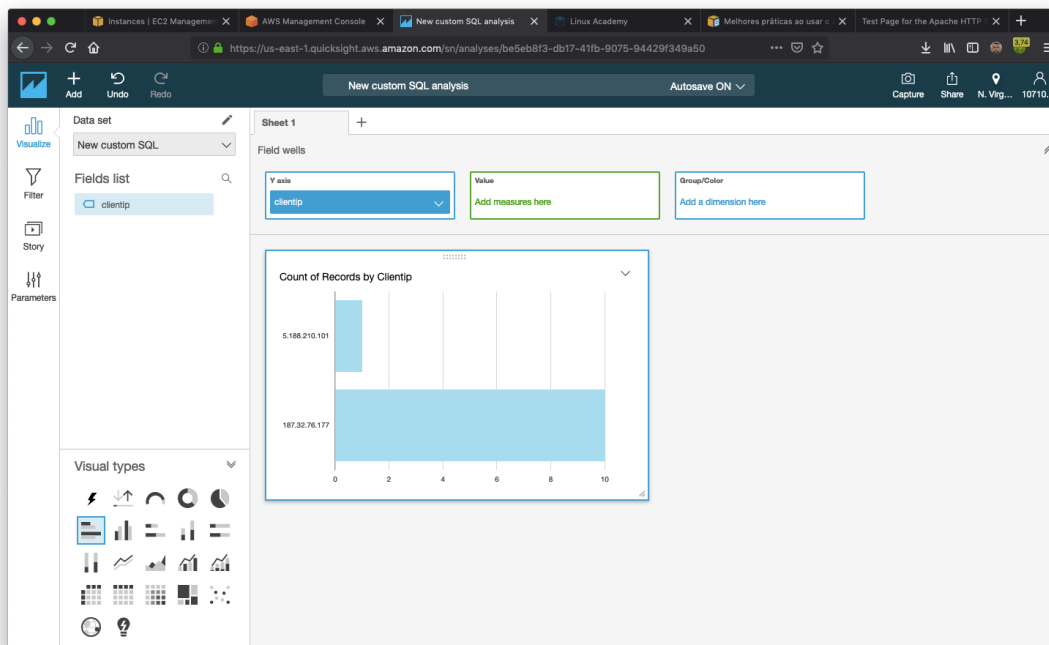
The screenshot shows a dialog box titled "New Athena data source" with a close button (X) in the top right corner. Below the title, there is a label "Data source name" followed by a text input field containing the text "logwebserver". At the bottom left, there is a green checkmark icon followed by the text "Validated". To the right of this, the text "SSL is enabled" is displayed. At the bottom right, there is a blue button labeled "Create data source".

Na tela seguinte, preencha conforme a image tomando cuidado no select SQL:



The screenshot shows a dialog box titled "Enter custom SQL query" with a close button (X) in the top right corner. Below the title, there is a label "Type in a SQL query." followed by a text input field containing the text "logwebserver". Below the input field, there is a large text area containing the SQL query: `SELECT clientip FROM "dblogswebserver"."logwebserver_im_tribanco";`. At the bottom left, there is a blue button labeled "Edit/Preview data". At the bottom right, there is a blue button labeled "Confirm query".

Na tela seguinte escolha o tipo de gráfico desejado e clique no campo clientip:



Vá no botão **Share**, e clique em **Publish dashboard**.

Na tela seguinte preencha conforme a imagem:

E clique no botão **Publish dashboard**.

Na próxima tela temos a opção de adicionar os usuários que terão acesso ao dashboard criado. Neste caso selecione a opção **Share with all users in this account**

Share dashboard with users

×

Select users in this account.

☐ Share with all users in this account

Name	Email	Permission	Role
------	-------	------------	------

Manage dashboard access

Share

Vá para a tela inicial do QuickSight e o dashboard já estará disponível:

