

Programa AWS Big Data Analytics Specialist

Informe Arquitectura Final AWS

Integrantes:

Anthony Steve Brian Manosalva López
Julian Castiblanco P
Jennyfer Nereida Combariza Vanegas
Isel Paola Santos Rojas

Junio 2020

Introducción

Este proyecto constituye el informe final del curso AWS Big Data Analytics Specialist.

Comprender las emociones de las personas cada día se convierte en una de las herramientas más esenciales para las empresas, ya que les permite escuchar atentamente a sus clientes y adaptar sus productos y/o servicios para satisfacer sus necesidades.

En este breve proyecto vamos a explicar como hacer una arquitectura serverless para realizar análisis de sentimientos de tweets relacionados con el covid en Perú.

Para esa arquitectura serverless usaremos los siguientes servicios de AWS:

1. Kinesis Firehose: servicio de AWS que permite el procesar datos en tiempo real.
2. Lambda: servicio de AWS que ejecuta cómputo a demanda.
3. Comprehend: servicio de procesamiento de lenguaje natural.
4. Lambda: servicio de AWS que ejecuta cómputo a demanda.
5. Elastic Search: servicio de AWS facilita la implementación, la protección y la ejecución de Elasticsearch a escala de manera rentable
6. Kibana. Ventana al Elastic Stack. Kibana es una interfaz de usuario gratuita y abierta que te permite visualizar los datos de Elasticsearch

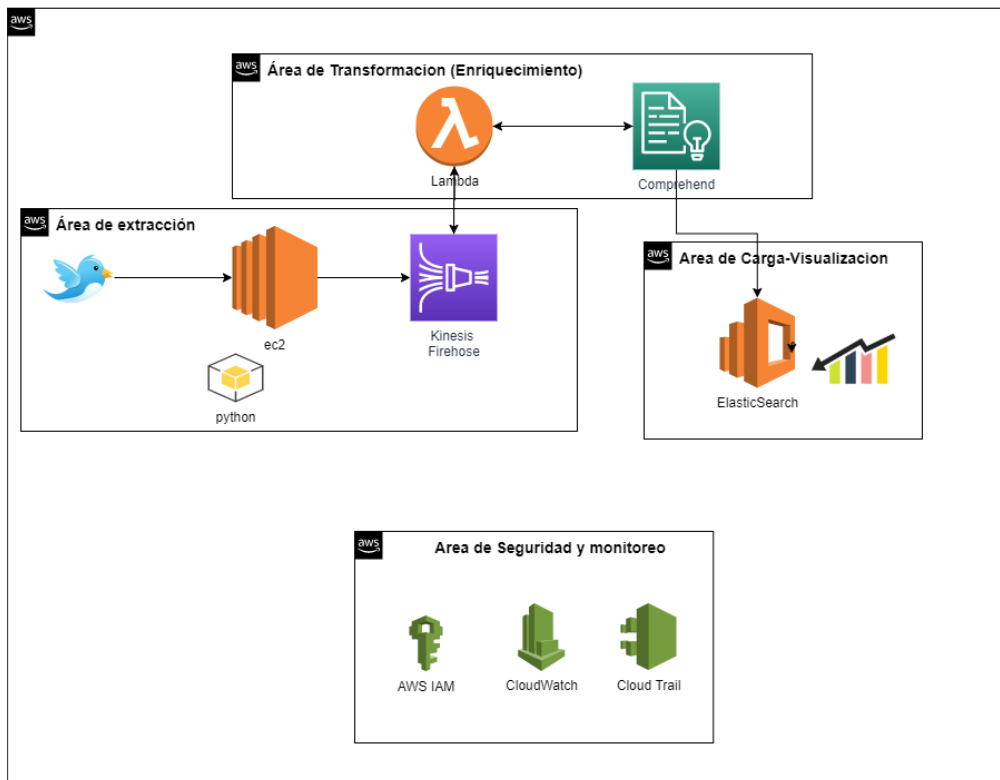
Para este análisis de sentimientos recopilamos los tweets con Kinesis Firehose, los procesamos con comprehend y visualizamos utilizando Elastic Search- Kibana.

1. Objetivo

Construir una arquitectura serverless que permita la visualización de análisis de sentimientos de tweets en tiempo real.

2. Arquitectura propuesta

Arquitectura propuesta End to End



3. Alcance

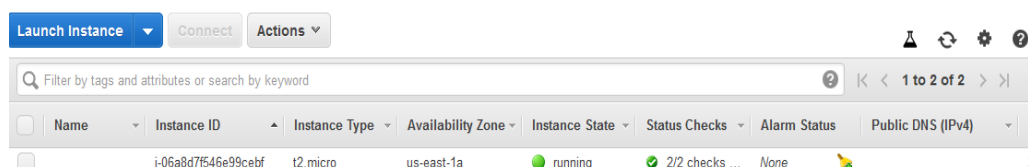
El presente documento describe los lineamientos llevados a cabo para la generación de la arquitectura propuesta desde la extracción de los tweets hasta la visualización de los resultados obtenidos del análisis de sentimientos.

4. Configuraciones generales

A continuación se listan los servicios que se usaron en el proceso de la generación de la arquitectura:

4.1. Configuración de EC2

Se crea una máquina EC2 que se encargará de ejecutar un código Python para conectarse con la cuenta de desarrollo de twitter y de esa manera extraer información acerca del covid en Lima Perú, con filtro de idioma en español.



Description	Status Checks	Monitoring	Tags
Instance ID	i-06a8d7f546e99cebf	Public DNS (IPv4)	-
Instance state	running	IPv4 Public IP	3.89.212.175
Instance type	t2.micro	IPv6 IPs	-
Finding	Opt-in to AWS Compute Optimizer for recommendations. Learn more	Elastic IPs	-
Private DNS	ip-172-30-0-146.ec2.internal	Availability zone	us-east-1a
Private IPs	172.30.0.146	Security groups	launch-wizard-1 , view inbound rules , view outbound rules
Secondary private IPs	-	Scheduled events	No scheduled events
VPC ID	vpc-0c269643596907e3d	AMI ID	ubuntu/images/hvm-ssd/ubuntu-bionic-18.04-amd64-server-20200611 (ami-0ac80df6eff0e70b5)
Subnet ID	subnet-07bda0fe1b38629b9	Platform details	Linux/UNIX
Network interfaces	eth0	Usage operation	RunInstances
IAM role	-	Source/dest. check	True
Termination protection	False	Root device	/dev/sda1
Lifecycle	normal	Block devices	/dev/sda1
Monitoring	basic	Elastic Graphics ID	-
Alarm status	None	Elastic Inference accelerator ID	-
Kernel ID	-	Capacity Reservation	-
RAM disk ID	-	Capacity Reservation Settings	Open
Placement group	-	Outpost Arn	-
Partition number	-		
Virtualization	hvm		
Reservation	r-04d74628cbaf91a18		
AMI launch index	0		
Tenancy	default		
Host ID	-		
ost resource group name	-		
Affinity	-		

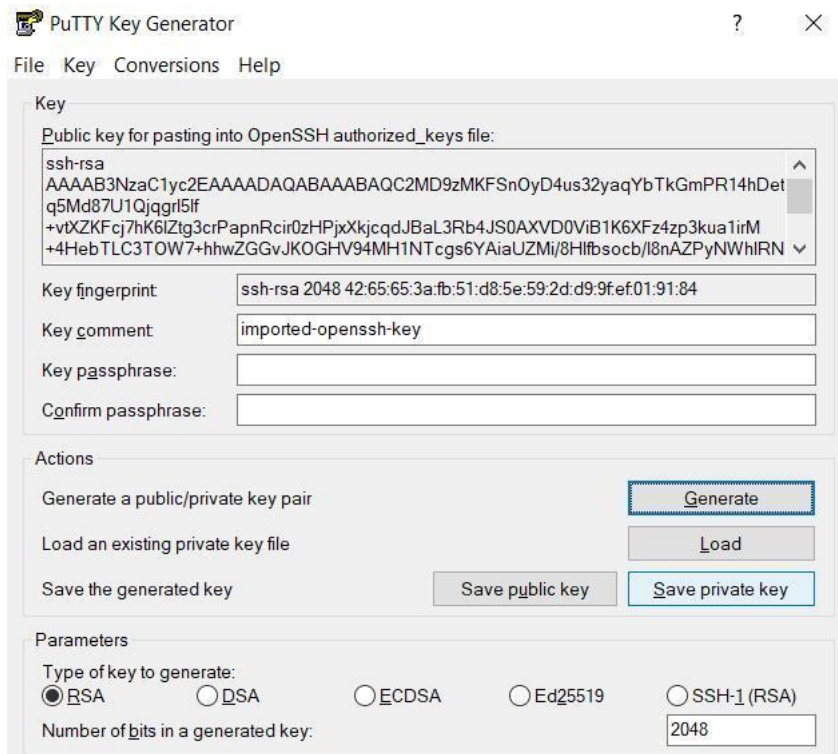
4.2. Conexión a EC2 por SSH a través de Putty

Objetivo: Cargar el código de python e instalar las librerías necesarias para una conexión exitosa.

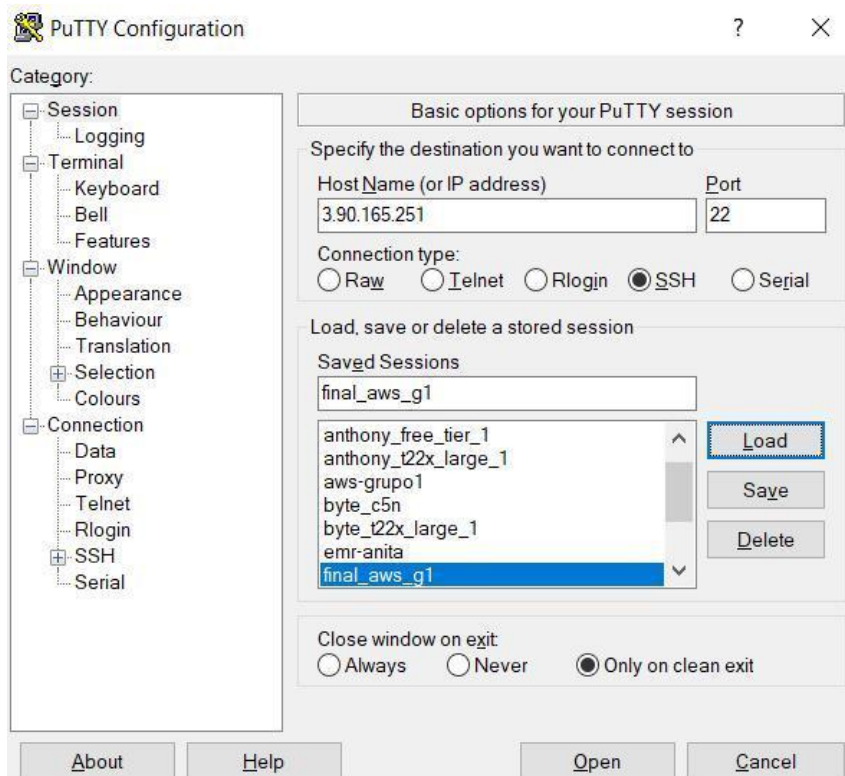
- Paso 1: Se descarga el archivo pem, donde viene la key privada.

aws-grupo1-pem.pem	6/27/2020 8:00 PM	PEM File	2 KB
elasticSearchAws.docx	6/28/2020 1:38 PM	Documento de Microsoft Word	648 KB

- Paso 2: Se abre la herramienta “putty generator” para convertir el archivo pem en ppk que es válido para la conexión desde putty.



- Paso 3: Se hace conexión a través de putty.



- Paso 5: Se ejecuta la función de extracción de tweets.

[illegible]

4.3. Configuración del servicio de Firehose

○	streamawsgrupo1	Active	2020-06-28T13:17-0500	Direct PUT and other sources	lambdaGrupo1 ↗	Amazon Elasticsearch Service elasticawsgrupo1
---	-----------------	--------	-----------------------	------------------------------	--------------------------------	--

Details	Monitoring	Tags	Encryption	Amazon Elasticsearch Service logs	Amazon S3 logs
---------	------------	------	------------	-----------------------------------	----------------

Delivery stream ARN

```
arn:aws:firehose:us-east-1:678761025749:deliverystream/streamawsgrupo1
```

Status

Active

Creation time

2020-06-28T13:17-0500

Permissions

IAM role

Kinesis Data Firehose uses this IAM role for all the permissions that the delivery stream needs. To specify different roles for the different permissions, use the API or the CLI. [Learn more](#)

[KinesisFirehoseServiceRole-streamawsgrp-us-east-1-1593368108763](#)

4.4. Configuración del servicio de Lambda, S3 y Elasticsearch

Source

To send records directly to the delivery stream use the [Amazon Kinesis Agent](#) or the [Kinesis Data Firehose API using the AWS SDK](#), or send records from AWS IoT, CloudWatch Logs, or CloudWatch Events. [Learn more](#)

Source
Direct PUT or other sources

Transform source records with AWS Lambda

Kinesis Data Firehose can transform source records before delivery. To return transformed source records to Kinesis Data Firehose, the Lambda function you invoke must be compliant with the required record transformation output model. [Learn more](#)

Source record transformation
Enabled

Lambda function
[lambdaGrupo1](#)

Lambda function version
\$LATEST

Description

Runtime
python3.8

Timeout
1 minute 30 seconds

S3 backup

Backup mode
All records

S3 bucket
[grupo1-bucket](#)

Prefix - optional
tweet/

Buffer conditions
5 MiB or 60 seconds

Compression
Disabled

Encryption
Disabled

4.5. Configuración del servicio Elasticsearch

Amazon Elasticsearch Service destination

Domain
[elasticawsgrupo1](#)

Index
tweet

Index rotation
Every day

Retry duration
300 seconds

Buffer conditions
5 MiB or 60 seconds

Settings

CloudWatch error logging
Enabled

4.6. Creación de la función Lambda

Se creó una función lambda que se encarga de utilizar el servicio de comprehend, para analizar los tweets.

lambdaGrupo1:\$LATEST

Version: \$LATEST Actions Select a test event Test Save

▼ Designer

lambdaGrupo1:\$LATEST

Layers (0)

+ Add trigger + Add destination

lambda_function

```
1 import json
2 import base64
3 import boto3
4
5 def lambda_handler(event, context):
6
7     comprehend = boto3.client("comprehend")
8     output = []
9     # s3 = boto3.resource("s3")
10    # bucket_name = 'test-firehose-to-lambda'
11
12    for record in event['records']:
13        payload = base64.b64decode(record['data']).decode("utf-8")
14        response = comprehend.detect_sentiment(Text = payload, LanguageCode = "es")
15        sentiment = response.get("Sentiment")
16        json_object = {'sentiment': sentiment}
17
18        ## To S3
19
20        # encoded_string = payload.encode("utf-8")
21        # file_name = "firehose.txt"
22        # lambda_path = "/tmp/" + file_name
23        # s3_path = file_name
24
25
26        # s3.Bucket(bucket_name).put_object(Key=s3_path, Body=encoded_string)
```

Se modificó la configuración de la función para que durara más tiempo y pueda utilizar más memoria en el procesamiento:

Basic settings Info

Description - optional

Runtime

Python 3.8

Handler Info

lambda_function.lambda_handler

Memory (MB)

Your function is allocated CPU proportional to the memory configured.

512 MB

Timeout

1 min 30 sec

Execution role

Choose a role that defines the permissions of your function. To create a custom role, go to the IAM console.

☒ Use an existing role

☐ Create a new role from AWS policy templates

Existing role

Choose an existing role that you've created to be used with this Lambda function. The role must have permission to upload logs to Amazon CloudWatch Logs.

service-role/lambdaGrupo1-role-qs6jnrq

4.7. Creación del elasticSearch

Step 1: Name and source

Step 2: Process records

Step 3: Choose a destination

Step 4: Configure settings

Step 5: Review

New delivery stream

Delivery streams load data, automatically and continuously, to the destinations that you specify. Kinesis Data Firehose resources are not covered under the [AWS Free Tier](#), and **usage-based charges apply**. For more information, see [Kinesis Data Firehose pricing](#). [Learn more](#)

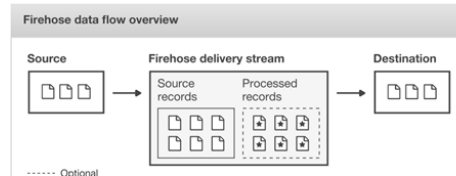
Delivery stream name

streamawsgrupo1

Acceptable characters are uppercase and lowercase letters, numbers, underscores, hyphens, and periods.

Choose a source

Choose how you would prefer to send records to the delivery stream.



Source

To learn about enabling server-side encryption (SSE), see [Data Protection in Amazon Kinesis Data Firehose](#).

☒ **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.

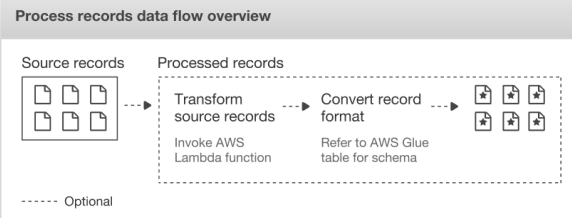
☐ **Kinesis Data Stream**

Server-side encryption for source records in the delivery stream

You can use AWS Key Management Service (KMS) to create and manage customer keys (CMKs) and to control the use of encryption across a wide range of AWS services and in your applications.

☐ Enable server-side encryption for source records in delivery stream

► **How to send source records to Kinesis Data Firehose**



Transform source records with AWS Lambda

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

Data transformation

☐ Disabled

☒ **Enabled**

Lambda function

lambdaGrupo1



Create new

View [lambdaGrupo1](#) in Lambda

longer in the **Advanced settings** section of your Lambda configuration.
[Go to Lambda configuration](#)

Timeout
3 seconds

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, Kinesis Data 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](#).

[Cancel](#)

[Previous](#)

[Next](#)

Select a destination

[Learn more](#)

Destination

- ☐ Amazon S3

Amazon S3 is an easy-to-use object storage, with a simple web service interface to store and retrieve any amount of data from anywhere on the web.

- ☐ Amazon Redshift

Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse that makes it simple and cost effective to analyze all your data using your existing business intelligence tools

- ☒ Amazon Elasticsearch Service

Elasticsearch is an open-source search and analytics engine for use cases such as log analytics, real-time application monitoring, and click stream analytics

- ☐ Splunk

Splunk is an operational intelligence tool for analyzing machine-generated data in real-time

delivery stream for VPC connectivity. [Learn more](#)

elasticawsgroup1



[Create new](#)

[View elasticawsgroup1 in Amazon Elasticsearch Service](#)

Index

tweet

A new index will be created if the the specified index name does not exist.

Index rotation

Every day

Select how often to rotate the Elasticsearch index. Kinesis Data Firehose appends a corresponding timestamp to the index and rotates it.

Type

A new type will be created if the specified type name does not exist.

Retry duration

Select how long a failed index request should be retried. Failed documents are delivered to the backup S3 bucket.

300

seconds

Enter a retry duration from 0 - 7200 seconds

S3 backup

To prevent against data loss, Kinesis Data Firehose can back up records to your S3 bucket while delivering it to your Elasticsearch cluster. [Learn more](#)

Backup mode

- ☐ Failed records only
- ☒ All records

Backup S3 bucket

grupo1-bucket

Create new

[View grupo1-bucket in S3 console](#)

Backup S3 bucket prefix - optional

tweet/

Kinesis Data Firehose automatically appends the "YYYY/MM/dd/HH/" UTC prefix to delivered S3 files. You can also specify an extra prefix in front of the time format and add "/" to the end to have it appear as a folder in the S3 console.

[Cancel](#)

[Previous](#)

[Next](#)

Elasticsearch buffer conditions

Kinesis Data Firehose buffers incoming records before delivering them to your Elasticsearch domain. Data delivery will be triggered when either of these conditions is satisfied. [Learn more](#)

Buffer size

5

MiB

Enter a buffer size between 1-100 MiB

Buffer interval

60

seconds

Enter a buffer interval between 60-900 seconds

S3 compression and encryption

Kinesis Data Firehose can compress records before delivering them to your S3 bucket. Compressed records can also be encrypted in the S3 bucket using a KMS master key. [Learn more](#)

S3 compression

- ☒ Disabled
- ☐ GZIP
- ☐ Snappy
- ☐ Zip
- ☐ Hadoop-Compatible Snappy

S3 encryption

- ☒ Disabled
- ☐ Enabled

Error logging

Kinesis Data Firehose can log record delivery errors to CloudWatch Logs. If enabled, a CloudWatch log group and corresponding log streams are created on your behalf. [Learn more](#)

Error logging

- ☐ Disabled
- ☒ Enabled

Tags - optional

You can add tags to organize your AWS resources, track costs, and control access. [Learn more](#)

Key

Enter key

Value - optional

Enter value

[Remove tag](#)

Permissions

IAM role

Kinesis Data Firehose uses this IAM role for all the permissions that the delivery stream needs. To specify different roles for the different permissions, use the API or the CLI. [Learn more](#)

- ☒ **Create or update IAM role `KinesisFirehoseServiceRole-streamawsgroup-us-east-1-1593368108763`**
This creates the role or updates it if it already exists, adds the required policies to it, and enables Kinesis Data Firehose to assume it.
- ☐ **Choose existing IAM role**
The role that you choose must have policies that include the permissions that Kinesis Data Firehose needs.

Cancel

Previous

Next

5 MiB or 60 seconds

S3 buffer conditions

5 MiB or 60 seconds

Compression

Disabled

Encryption

Disabled

Error logging

Enabled

Tags

no tags specified

IAM role

KinesisFirehoseServiceRole-streamawsgroup-us-east-1-1593368108763

Cancel

Previous

Create delivery stream

Se modificaron los permisos para acceder al Kibana desde cualquier lugar público

Status **Active**

Domain access policy

JSON defined access policy

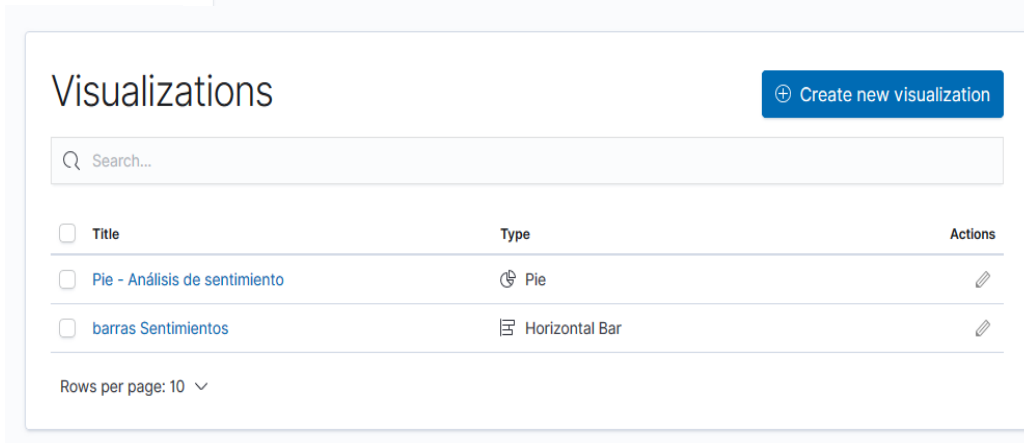
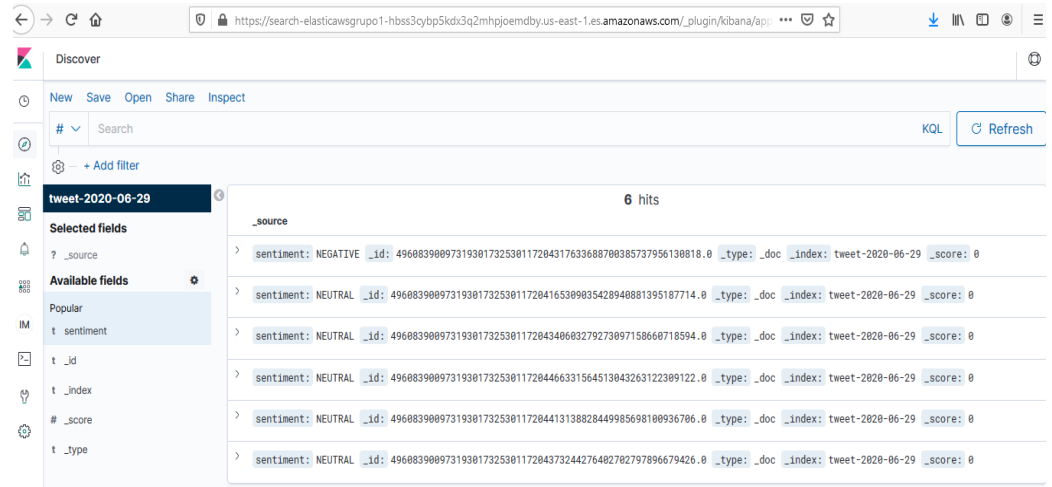
Allow or deny access by AWS account ID, account ARN, IAM user ARN, IAM role ARN, IPv4 address, or CIDR block.

Add or edit the access policy

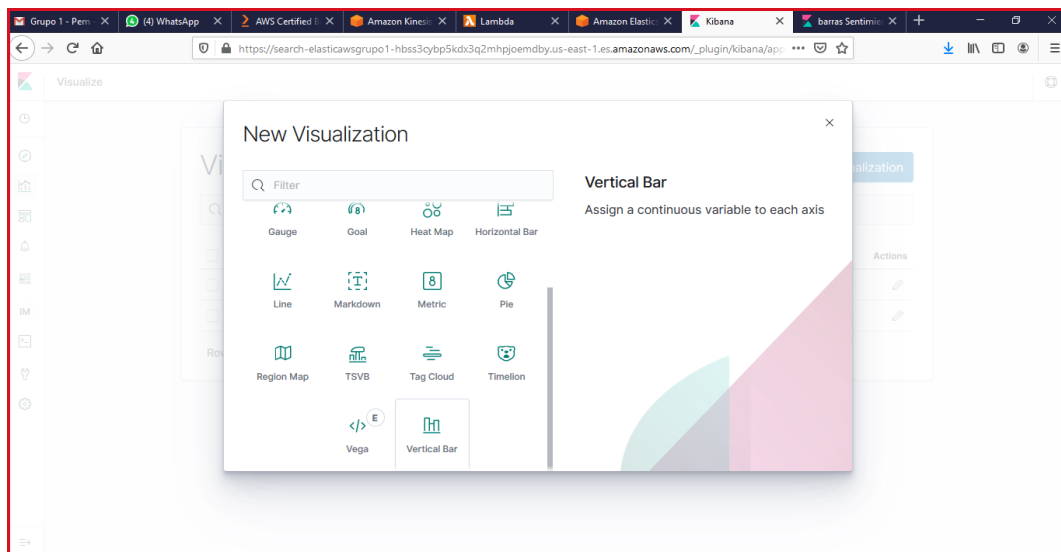
```
1 {  
2   "Version": "2012-10-17",  
3   "Statement": [  
4     {  
5       "Effect": "Allow",  
6       "Principal": {  
7         "AWS": "*"  
8       },  
9       "Action": "es:*",  
10      "Resource": "arn:aws:es:us-east-1:678761025749:domain/elasticawsgroup1/*"  
11    }  
12  ]  
13 }
```


A continuación, en el ícono de discover, podemos validar los datos disponibles para la creación de las gráficas.

Luego se da clic en el ícono de crear una nueva visualización



Seleccionamos nueva gráfica de barras



Seleccionamos el set de datos que queremos analizar

New Vertical Bar / Choose a source

Sort ▾

Types 2 ▾

 tweet-2020-06-29

En el eje Y se deja que cuente cuantos registros hay por tipo de respuesta sentimental

tweet-2020-06-29

Data Metrics & Axes Panel Settings ▶ ✕

Metrics

Y-axis

Aggregation Count help

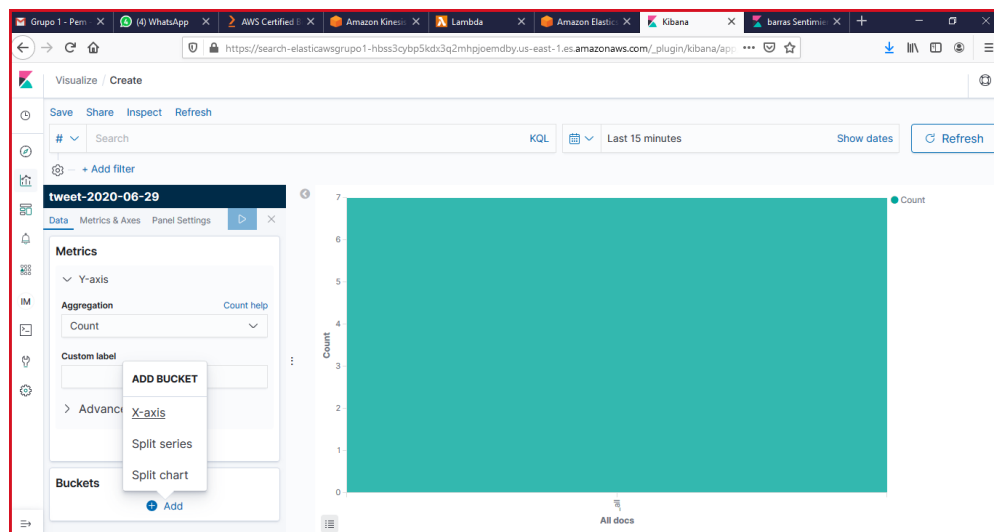
Count ▾

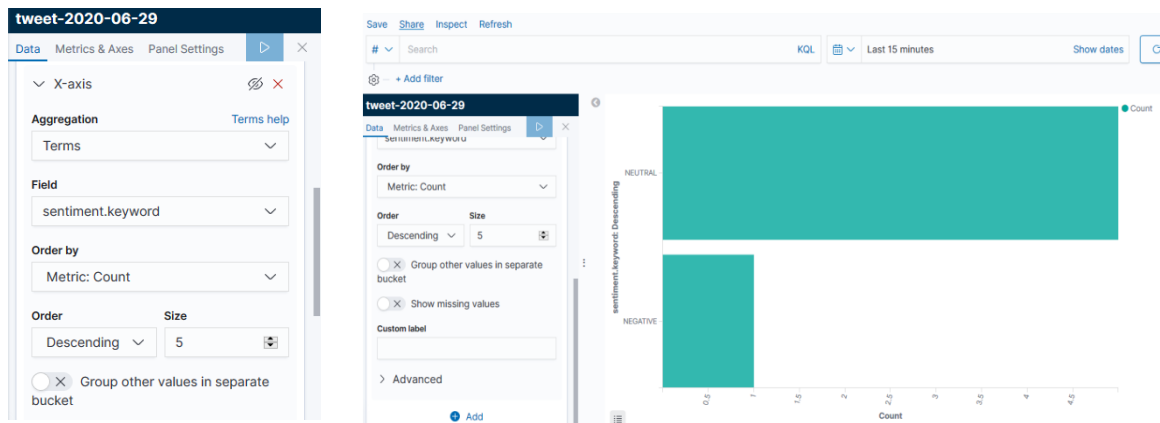
Custom label

Advanced

+ Add

A continuación le damos clic en el botón add bucket





De manera similar se pueden replicar los pasos para realizar un gráfico de tipo pie.

