



Programa AWS Big Data – Analytics Specialist



“Big Data, el nuevo petróleo”

Kinesis Firehose

- Se gestiona completamente
- Solo va a S3, Redshift, Elasticsearch, Splunk
- Es solo una ingesta que almacenará en bufer y comprimirá los archivos en caso requieran, una vez procesada los datos desaparecerán.

Kinesis Data Stream

- Se gestiona manualmente
- Puede ir a otros servicios.
- Almacenarlos de 24 horas hasta 7 días (con costo adicional).
- Opción mas personalizada , es decir para desarrolladores que crean aplicaciones o transmiten datos especializados.

	Kinesis Data Streams	Kinesis Firehose
Propósito	Low latency streaming service for ingest at scale	Data transfer service to load streaming data into Amazon S3, Redshift, Elasticsearch & Splunk
Provisionamiento	Managed service but needs configuration for shards.	Fully managed service, no administration.
Procesamiento	Real Time (~200ms latency for classic, ~70ms for enhanced fan out)	Near real time (depends on buffer size OR buffer time min. 60 secs)
Escala	Must manage scaling (configure shards)	Automated Scaling – as per the demand
Almacenamiento de data	Configurable from 1 to 7 days	Does not provide data storage
Capacidad de reproducción	Supports replay capability	Does not support replay capability
Producer	Need to write code for producer. Supports SDK, Kinesis Agent, KPL, CloudWatch, IoT	Need to write code for producer. Supports KPL, Kinesis Agent, Data Streams, CloudWatch, IoT
Consumer	Open ended. Supports multiple consumers and destinations. Supports KCL and Spark.	Closed ended. Handled by Firehose. Does not support KCL or Spark.

Casos prácticos

1) Su organización necesita ingerir una gran secuencia de datos en su lago de datos en Amazon S3. Los datos pueden transmitir a una velocidad de cientos de megabytes por segundo. ¿Qué servicio de AWS logrará el objetivo con la menor cantidad de administración?

- 1.Amazon Kinesis Firehose
- 2.Amazon Kinesis Streams
- 3.Amazon CloudFront
- 4.Amazon SQS

2. Su organización está buscando una solución que pueda ayudar a la empresa con la transmisión de datos. Varios servicios requerirán acceso para leer y procesar la misma transmisión al mismo tiempo. ¿Qué servicio de AWS cumple con los requisitos comerciales?

- 1.Amazon Kinesis Firehose
- 2.Amazon Kinesis Streams
- 3.Amazon CloudFront
- 4.Amazon SQS

3. Su aplicación genera una carga útil JSON de 1 KB que debe ponerse en cola y entregarse a las instancias EC2 para las aplicaciones. Al final del día, la aplicación necesita reproducir los datos de las últimas 24 horas. En el futuro cercano, también necesitará la capacidad de que otras aplicaciones EC2 múltiples consuman la misma transmisión de manera simultánea. ¿Cuál es la mejor solución para esto?

- 1.Kinesis Data Streams
- 2.Kinesis Firehose
- 3.SNS
- 4.SQS



EC2



Kinesis Data Stream

Databricks Community Edition



Spark Structured Streaming



Spark SQL

Python libraries



NumPy



matplotlib

pandas

$My = (I^T x_d + j_0) + y_d$



ARQUITECTURA A IMPLEMENTAR



La arquitectura de la canalización de datos.



Servicios ^

Grupos de recursos v



Data Acad

Historial

VPC

EC2

Lambda

IAM

Kinesis

CloudFormation

ec2



Informática

EC2

Lightsail ↗

Lambda

Batch

Elastic Beanstalk

Serverless Application Repository



Cadena De Bloques

Amazon Managed Blockchain



Satélite

Ground Station



Análisis

Athena

EMR

CloudSearch

Elasticsearch Service

Kinesis

QuickSight ↗

Launch Instance ▼ **Connect** **Actions** ▼

Filter by tags and attributes or search by keyword 1 to 1 of 1

- Events **New**
- Tags
- Reports
- Limits

Instances

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Instance: i-0bc5f1f8e1af8fb81 Public DNS: ec2-34-201-161-78.compute-1.amazonaws.com

Description	Status Checks	Monitoring	Tags
-------------	---------------	------------	------

☒ New VPC Experience
Tell us what you think

VPC Dashboard

Filter by VPC:

▼ VIRTUAL PRIVATE CLOUD

Your VPCs

Subnets

Route Tables

Internet Gateways New

Egress Only Internet Gateways

DHCP Options Sets New

Elastic IPs New

Endpoints

Endpoint Services

NAT Gateways

Peering Connections

▼ SECURITY

Network ACLs

Create VPC

Actions ▾

1 to 1 of 1

<input type="checkbox"/>	Name ▾	VPC ID ▴	State ▾	IPv4 CIDR	IPv6 CIDR	DHCP options set	Main Route table	Main
<input type="checkbox"/>		vpc-06601b53fe9d1a37d	available	172.31.0...	-	dopt-f8df0482	rtb-0720fe0ca4c3b8cde	acl-0

VPC: vpc-06601b53fe9d1a37d

Description

CIDR Blocks

Flow Logs

Tags

IPv4 CIDR Blocks:

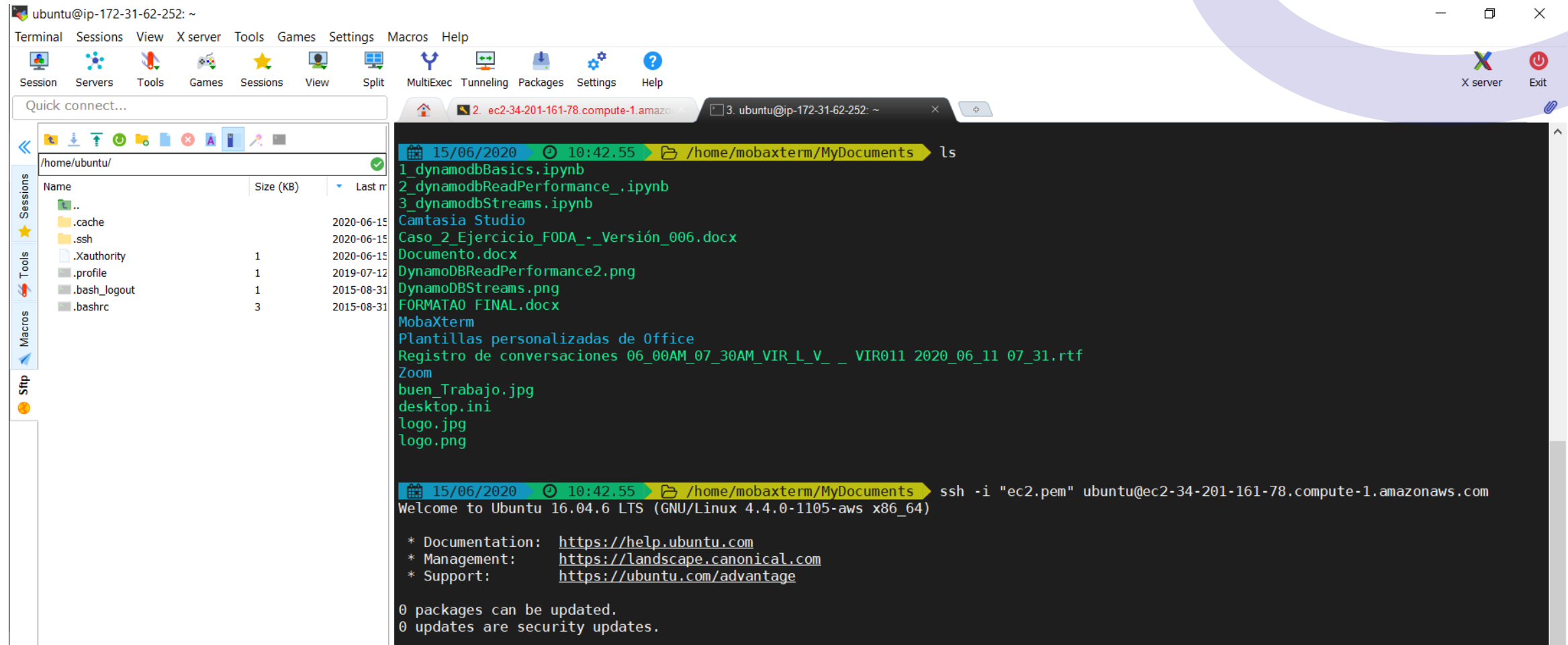
CIDR ⓘ	Status	Status reason
172.31.0.0/16	associated	-

IPv6 CIDR Blocks:

CIDR ⓘ	IPv6 Pool	Status	Status reason
--------	-----------	--------	---------------

Para ingresar :

ssh -i "ec2_2.pem" ubuntu@ec2-54-210-89-79.compute-1.amazonaws.com



The screenshot shows a MobaXterm window with a terminal session and a file explorer. The terminal window is titled "3. ubuntu@ip-172-31-62-252: ~" and displays the following output:

```
15/06/2020 10:42:55 /home/mobaxterm/MyDocuments ls
1_dynamodbBasics.ipynb
2_dynamodbReadPerformance_.ipynb
3_dynamodbStreams.ipynb
Camtasia Studio
Caso_2_Ejercicio_FODA_-_Versión_006.docx
Documento.docx
DynamoDBReadPerformance2.png
DynamoDBStreams.png
FORMATAO FINAL.docx
MobaXterm
Plantillas personalizadas de Office
Registro de conversaciones 06_00AM_07_30AM_VIR_L_V_ _ VIR011 2020_06_11 07_31.rtf
Zoom
buen_Trabajo.jpg
desktop.ini
logo.jpg
logo.png

15/06/2020 10:42:55 /home/mobaxterm/MyDocuments ssh -i "ec2.pem" ubuntu@ec2-34-201-161-78.compute-1.amazonaws.com
Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-1105-aws x86_64)

* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage

0 packages can be updated.
0 updates are security updates.
```

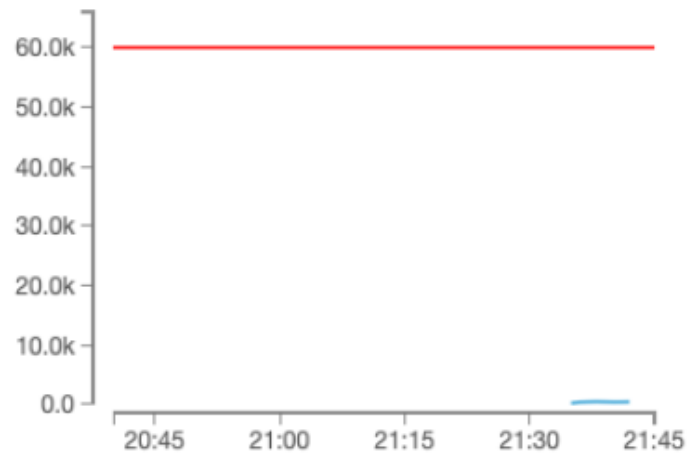
The file explorer on the left shows the contents of the /home/ubuntu/ directory:

Name	Size (KB)	Last m
..		
.cache		2020-06-15
.ssh		2020-06-15
.Xauthority	1	2020-06-15
.profile	1	2019-07-12
.bash_logout	1	2015-08-31
.bashrc	3	2015-08-31

- `sudo add-apt-repository universe`
 - `sudo apt update`
 - `sudo apt install python-pip`
 - `sudo apt install python3-pip`
 - `sudo apt install python3-boto3`
 - `sudo apt install python3-tweepy`
 - `sudo su`
 - `pip install --upgrade pip`
 - `sudo apt update && apt upgrade`
-
- `python3 scrapping_covid.py \#COVID_19 lima`

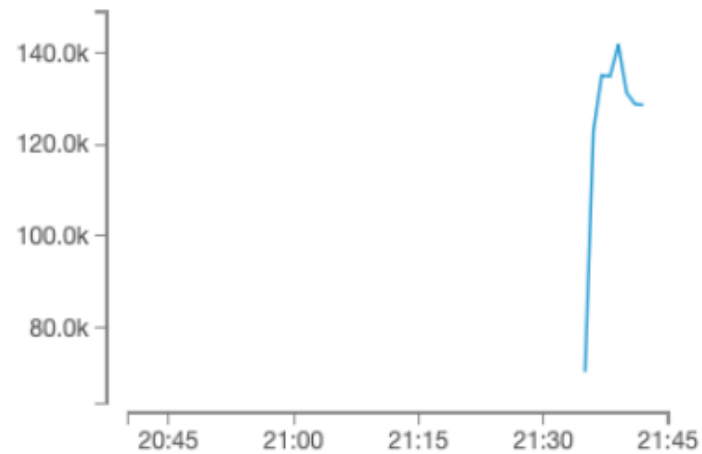
Incoming Data (Count) — Sum

IncomingRecords



Put Record (Bytes) — Sum

PutRecord.Bytes



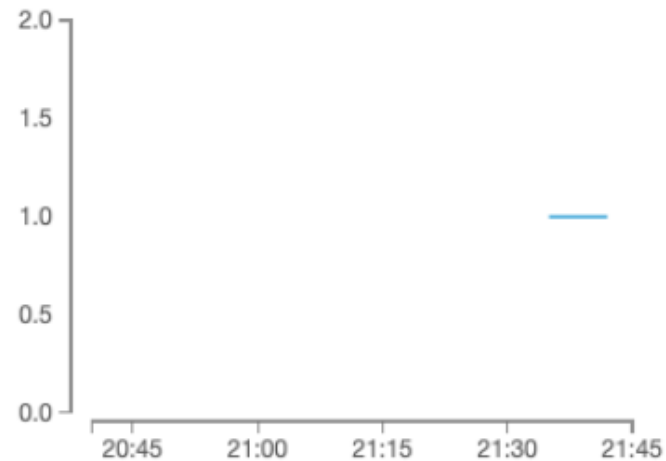
Put Record Latency (Milliseconds) — Average

PutRecord.Latency



Put Record Success (Percent) — Average

PutRecord.Success



Métricas reflejadas en Kinesis Data Stream

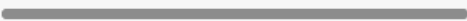


- <https://mvnrepository.com/artifact/com.amazonaws/amazon-kinesis-client/1.13.3>
- https://mvnrepository.com/artifact/org.apache.spark/spark-streaming-kinesis-asl_2.12/2.4.6

```
1 df = kinesisDF \
2   .writeStream \
3   .format("memory") \
4   .outputMode("append") \
5   .queryName("tweets") \
6   .start()
```

Cancel

▼ (1) Spark Jobs

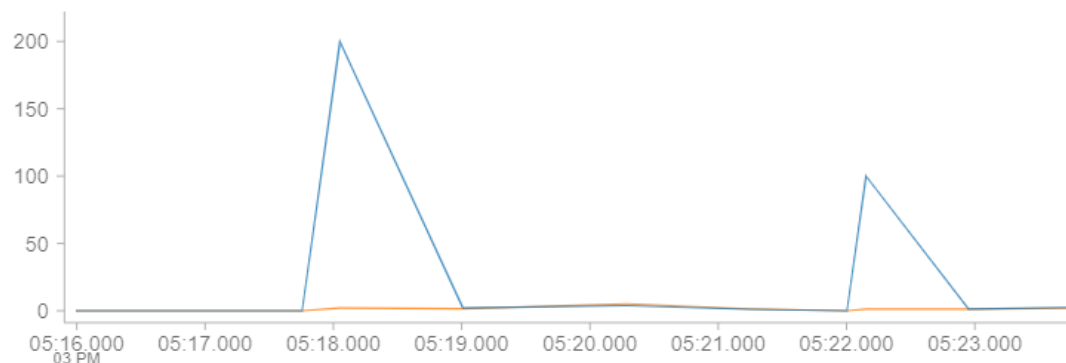
► Job 228  [View \(1 stages\)](#)

▼  tweets (id: 45e862f9-d301-40dc-bb24-3329715c3d0f) *Last updated: 5 seconds ago*

Dashboard [Raw Data](#)

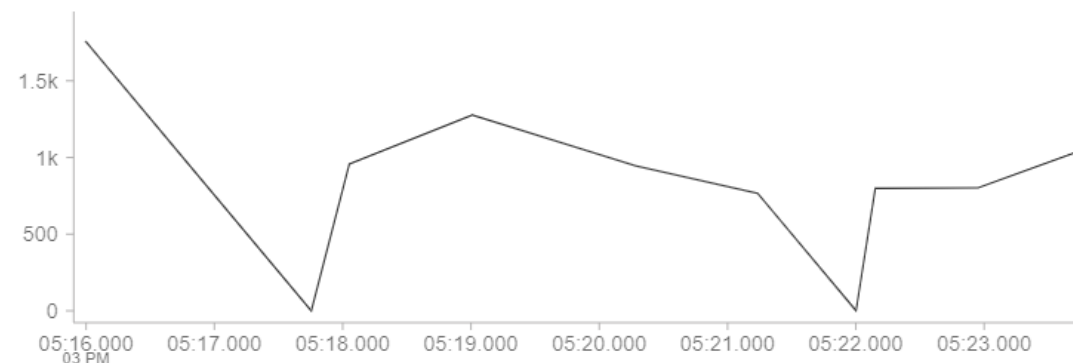
Input vs. Processing Rate
records per second

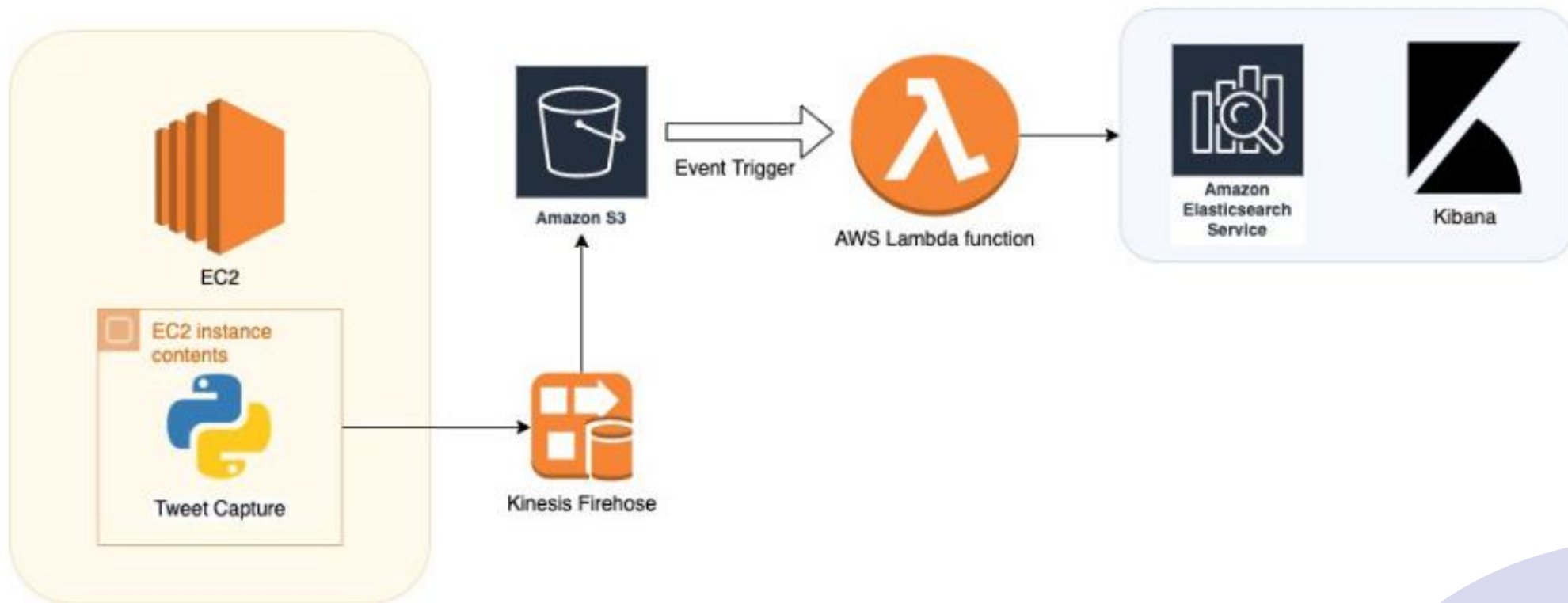
2.5 rec/s **1.9 rec/s**
Input rate Processing rate



Batch Duration
in milliseconds

835.2 ms **1047 ms**
Average Latest





La arquitectura de la tubería de datos.

ROLES A TENER EN CUENTA

- Kinesis Firehose necesita un rol de IAM con permisos otorgados para entregar datos de flujo, lo cual se discutirá en la sección de Kinesis y S3.
- AWS Lambda necesita permisos para acceder al desencadenador de eventos S3, agregar registros de CloudWatch e interactuar con Amazon Elasticsearch Service.