

Relatório Day_2 – Fase 2 (Ana Letícia #53)

Link do repositório do GitHub: https://github.com/anaoliveira07/Desempate_Estadual

1- Ec2

Criei a instância Ec2 como foi solicitado.

Instances (1/1) Info		Last updated 4 minutes ago	Refresh	Connect	Instance state	Actions	Launch instances
<input type="text" value="Find Instance by attribute or tag (case-sensitive)"/>		All states		< 1 >		Settings	
<input checked="" type="checkbox"/>	Name ↗	Instance ID	Instance state	Instance type	Status check	Alarm	
<input checked="" type="checkbox"/>	Day_2-instance	i-00f88f5f9436f1ccd	Running	t2.micro	2/2 checks passed	View details	

i-00f88f5f9436f1ccd (Day_2-instance)

[Details](#) | [Status and alarms](#) | [Monitoring](#) | [Security](#) | [Networking](#) | [Storage](#) | [Tags](#)

▼ Instance summary [Info](#)

Instance ID
[i-00f88f5f9436f1ccd](#)

IPv6 address
–

Hostname type
IP name: ip-172-31-93-226.ec2.internal

Answer private resource DNS name
IPv4 (A)

Auto-assigned IP address
[3.83.202.50](#) [Public IP]

Public IPv4 address
[3.83.202.50](#) | [open address](#)

Instance state
Running

Private IP DNS name (IPv4 only)
[ip-172-31-93-226.ec2.internal](#)

Instance type
t2.micro

VPC ID
[vpc-0cce722bb94b82dfd](#)

Private IPv4 addresses
[172.31.93.226](#)

Public IPv4 DNS
[ec2-3-83-202-50.compute-1.amazonaws.com](#)
| [open address](#)

Elastic IP addresses
–

AWS Compute Optimizer finding
[Opt-in to AWS Compute Optimizer for recommendations.](#)

2- Conectar a instância via terminal

Usei o comando: `sudo yum install python3 git aws-cli -y`

```
#
#
~\##### Amazon Linux 2
~\#####
~\####| AL2 End of Life is 2026-06-30.
~\#/
~V~' '->
~~~
~-.-
~/m/'

Amazon Linux 2023, GA and supported until 2028-03-15.
https://aws.amazon.com/linux/amazon-linux-2023/

[ec2-user@ip-172-31-93-226 ~]$ sudo yum update -y
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
amzn2-core | 3.6 kB 00:00:00

No packages marked for update
[ec2-user@ip-172-31-93-226 ~]$ sudo yum install python3 git aws-cli -y
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
Package python3-3.7.16-1.amzn2.0.8.x86_64 already installed and latest version
Package awscli-1.18.147-1.amzn2.0.2.noarch already installed and latest version
Resolving Dependencies
--> Running transaction check
--> Package git.x86_64 0:2.47.1-1.amzn2.0.2 will be installed
--> Processing Dependency: git-core = 2.47.1-1.amzn2.0.2 for package: git-2.47.1-1.amzn2.0.2.x86_64
--> Processing Dependency: git-core-doc = 2.47.1-1.amzn2.0.2 for package: git-2.47.1-1.amzn2.0.2.x86_64
--> Processing Dependency: perl-Git = 2.47.1-1.amzn2.0.2 for package: git-2.47.1-1.amzn2.0.2.x86_64
--> Processing Dependency: perl(Git) for package: git-2.47.1-1.amzn2.0.2.x86_64
--> Processing Dependency: perl(Term::ReadKey) for package: git-2.47.1-1.amzn2.0.2.x86_64
--> Running transaction check
--> Package git-core.x86_64 0:2.47.1-1.amzn2.0.2 will be installed
--> Package git-core-doc.noarch 0:2.47.1-1.amzn2.0.2 will be installed
--> Package perl-Git.noarch 0:2.47.1-1.amzn2.0.2 will be installed
--> Processing Dependency: perl(Error) for package: perl-Git-2.47.1-1.amzn2.0.2.noarch
--> Package perl-TermReadKey.x86_64 0:2.30-20.amzn2.0.2 will be installed
--> Running transaction check
--> Package perl-Error.noarch 1:0.17020-2.amzn2 will be installed
--> Finished Dependency Resolution

Dependencies Resolved
```

i-00f88f5f9436f1ccd (Day_2-instance)

PublicIPs: 3.83.202.50 PrivateIPs: 172.31.93.226

3- S3

crie o Bucket no S3

datax-raw-storage-ana [Info](#)

[Objects](#)
[Metadata](#)
[Properties](#)
[Permissions](#)
[Metrics](#)
[Management](#)
[Access Points](#)

Objects (0)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

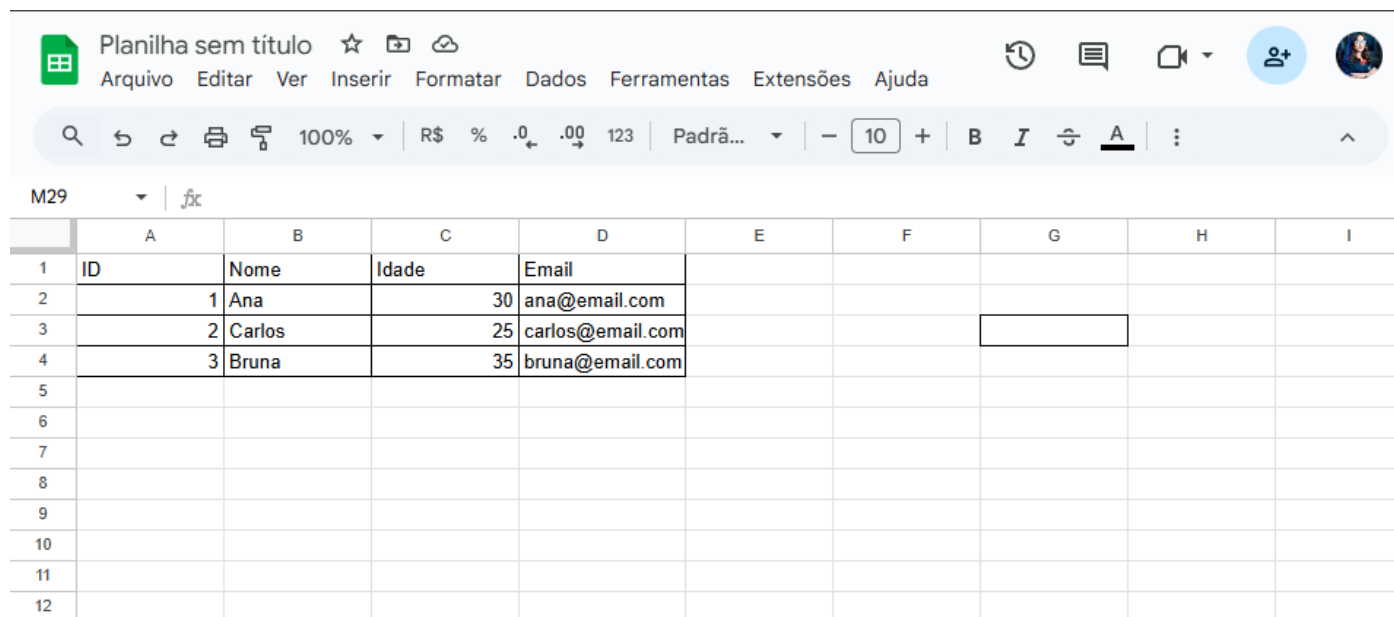
☒ Show versions

< 1 >

	Name	Type	Last modified
<div>No objects</div> <div>You don't have any objects in this bucket.</div> <div> Upload </div>			

4- Criação do arquivo .csv

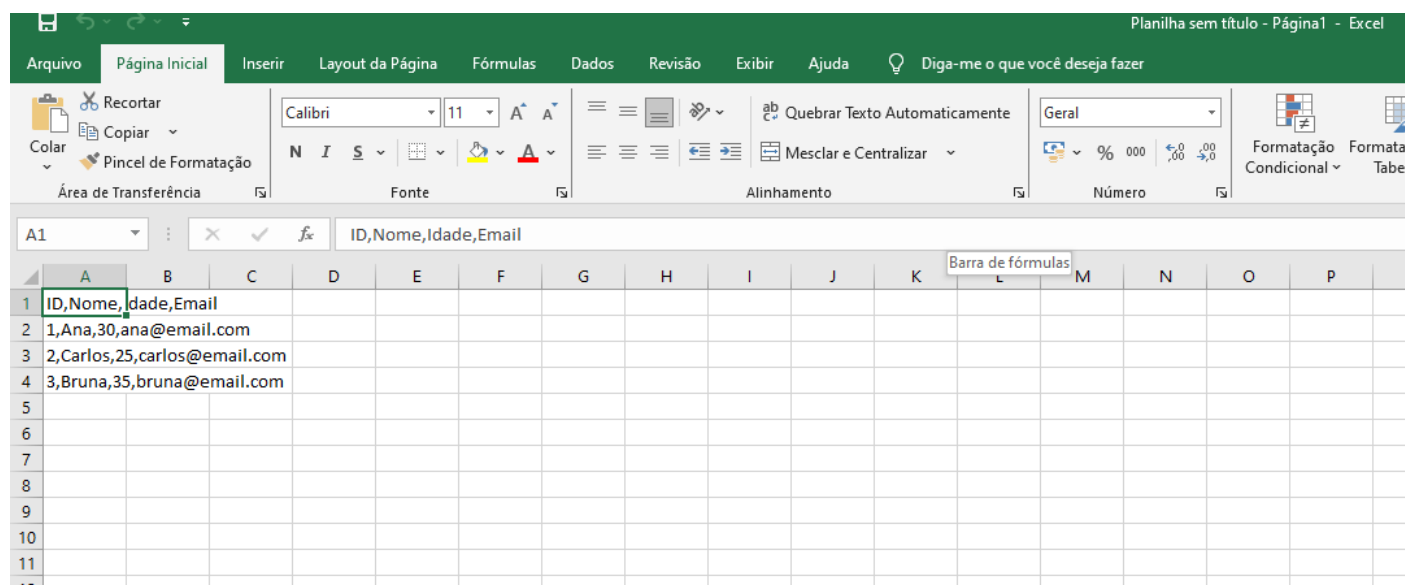
Criei no Sheets no Google uma tabela como foi solicitada no PDF e salvei como arquivo .csv



The screenshot shows the Google Sheets interface. The title bar says "Planilha sem título". The menu bar includes "Arquivo", "Editar", "Ver", "Inserir", "Formatar", "Dados", "Ferramentas", "Extensões", and "Ajuda". The toolbar shows various icons for undo, redo, print, and zoom. The main area displays a table with the following data:

	A	B	C	D	E	F	G	H	I
1	ID	Nome	Idade	Email					
2	1	Ana	30	ana@email.com					
3	2	Carlos	25	carlos@email.com					
4	3	Bruna	35	bruna@email.com					
5									
6									
7									
8									
9									
10									
11									
12									

Planilha depois de salva:



The screenshot shows the Microsoft Excel interface. The title bar says "Planilha sem título - Página1 - Excel". The ribbon includes "Arquivo", "Página Inicial", "Inserir", "Layout da Página", "Fórmulas", "Dados", "Revisão", "Exibir", and "Ajuda". The "Página Inicial" ribbon is active, showing options for "Recortar", "Copiar", "Colar", "Pincel de Formatação", "Fonte", "Alinhamento", "Número", and "Formatação Condicional". The main area displays the same table data as the Google Sheets screenshot:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ID, Nome, Idade, Email															
2	1, Ana, 30, ana@email.com															
3	2, Carlos, 25, carlos@email.com															
4	3, Bruna, 35, bruna@email.com															
5																
6																
7																
8																
9																
10																
11																
12																

5- Subir o arquivo no Bucket do S3

datax-raw-storage-ana [Info](#)

[Objects](#) | [Metadata](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (1)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

☐ Show versions < 1 > ⚙

<input type="checkbox"/>	Name	Type	Last modified
<input type="checkbox"/>	Arquivo Day_2 (Ana Letícia #53).csv	csv	January 28, 2025, 10:59:44 (UTC-03:00)

*OBS: Eu não soube o arquivo do S3 pela instância usando o AWS CLI

FASE 3

1- AWS Glue

Criação do Crawler:

Definir propriedades do rastreador

Detalhes do rastreador

Informações

Nome

datax-crawler

O nome pode ter até 255 caracteres. Alguns conjuntos de caracteres, incluindo caracteres de controle, são proibidos.

Escolha fontes de dados e classificadores

Configuração da fonte de dados

Seus dados já estão mapeados para tabelas do Glue?

☒ Ainda não

Selecione uma ou mais fontes de dados a serem rastreadas.

☐ Sim

Selecione tabelas existentes no seu Catálogo de Dados do Glue.

Fontes de dados (1)

Informações

A lista de fontes de dados a serem escaneadas pelo rastreador.

Editar

Remover

Adicionar uma fonte de dados

Tipo	Fonte de dados	Parâmetros
<input type="radio"/> S3	s3://datax-armazenamento-bruto-ana	Rastrear tudo novamente

Set output and scheduling

Output configuration

Info

Target database

datax-db

Clear selection

Add database

Crawler criado:

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1)

Info

Last updated (UTC)

January 28, 2025 at 14:13:55

Action

Run

Create crawler

View and manage all available crawlers.

Filter crawlers

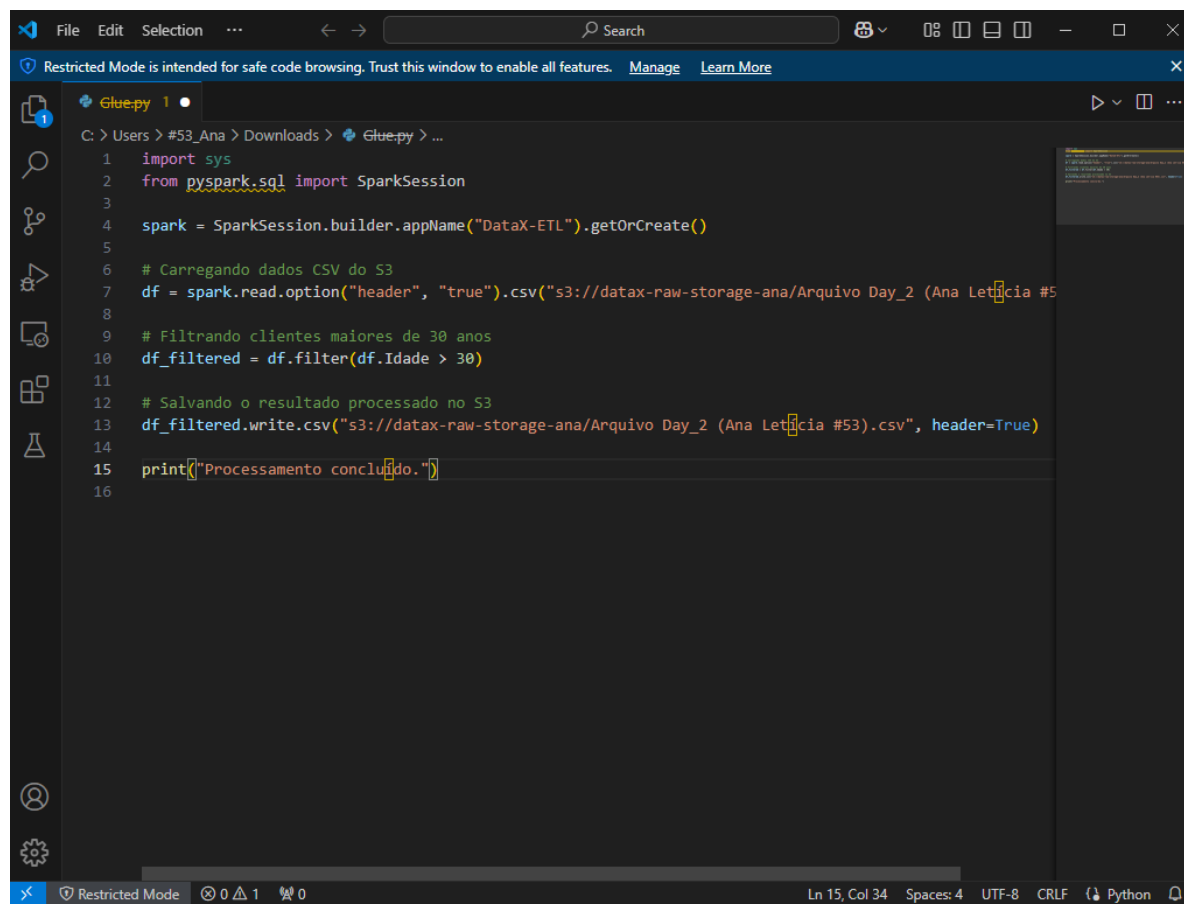
< 1 >

⚙

<input type="checkbox"/>	Name	State	Schedule
<input type="checkbox"/>	datax-crawler	🟢 Ready	

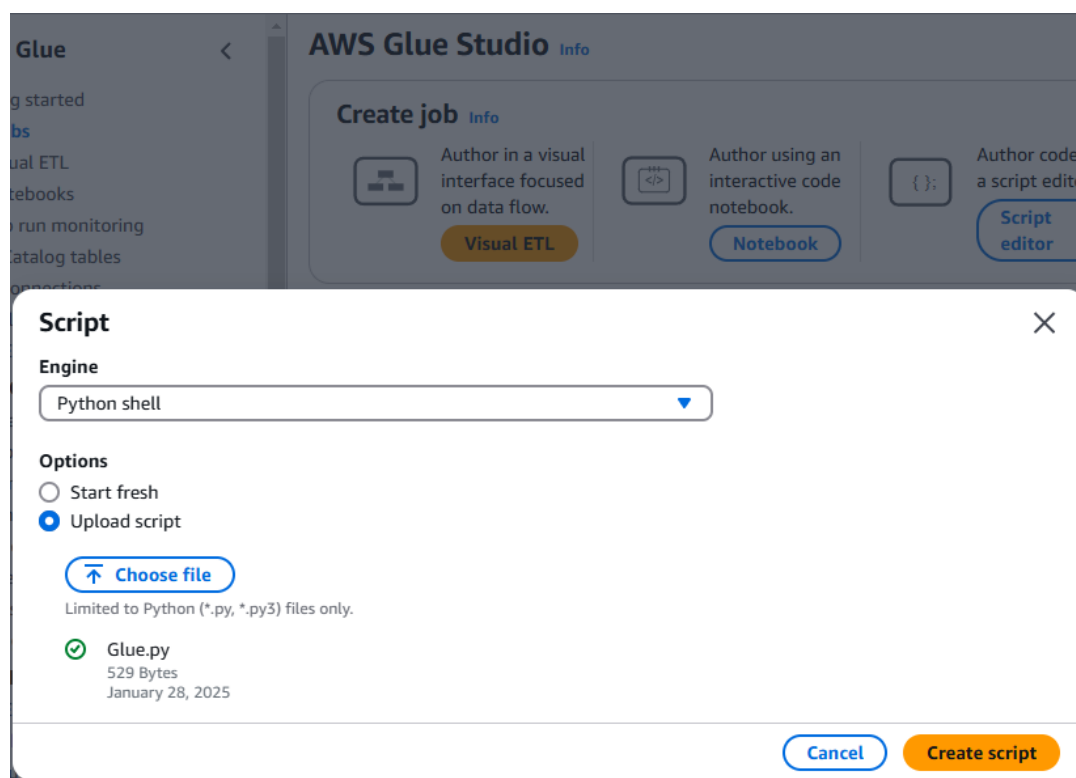
2- Código Python:

Baixei o arquivo .py do drive e modifiquei o que precisava



```
1 import sys
2 from pyspark.sql import SparkSession
3
4 spark = SparkSession.builder.appName("DataX-ETL").getOrCreate()
5
6 # Carregando dados CSV do S3
7 df = spark.read.option("header", "true").csv("s3://datax-raw-storage-ana/Arquivo Day_2 (Ana Letícia #53).csv", header=True)
8
9 # Filtrando clientes maiores de 30 anos
10 df_filtered = df.filter(df.Idade > 30)
11
12 # Salvando o resultado processado no S3
13 df_filtered.write.csv("s3://datax-raw-storage-ana/Arquivo Day_2 (Ana Letícia #53).csv", header=True)
14
15 print("Processamento concluído.")
16
```

3- Criação do Job



Ana Letícia #53

4- Executar Job

☑ **Successfully started job**
Successfully started job Day_2-Job(Ana). Navigate to [Run details](#) for more details.

Day_2-Job(Ana)

Last modified on 28/01/2025, 11:26:29

Actions

Save

Run

< **Script** | Job details | Runs | Data quality | Schedules | Version Control | Upgrade analysis | >

Script Info

```
1 import sys
2 from pyspark.sql import SparkSession
3
4 spark = SparkSession.builder.appName("DataX-ETL").getOrCreate()
5
6 # Carregando dados CSV do S3
7 df = spark.read.option("header", "true").csv("s3://datax-raw-storage-ana/Arquivo Day_2 (Ana Leticia #53).csv"
8       )
9
10 # Filtrando clientes maiores de 30 anos
11 df_filtered = df.filter(df.Idade > 30)
12
13 # Salvando o resultado processado no S3
14 df_filtered.write.csv("s3://datax-raw-storage-ana/Arquivo Day_2 (Ana Leticia #53).csv", header=True)
15
16 print("Processamento concluído.")
```

5- Depois de executar o job

Amazon S3

General purpose buckets

Directory buckets

Table buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

▼ Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight 10

► **Account snapshot - updated every 24 hours** All AWS Regions

[View Storage Lens dashboard](#)

Storage lens provides visibility into storage usage and activity trends. Metrics don't include directory buckets. [Learn more](#)

General purpose buckets

Directory buckets

General purpose buckets (2) Info All AWS Regions



Copy ARN

Empty

Delete

Create bucket

Buckets are containers for data stored in S3.

Find buckets by name

< 1 >



Name



AWS Region



[aws-glue-assets-211125364911-us-east-1](#)

US East (N. Virginia) us-east-



[datax-raw-storage-ana](#)

US East (N. Virginia) us-east-

6- Dentro do bucket criado pelo Job do AWS Glue:

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	scripts/	Folder	-	-	-

7- SNS

Tive que criar um SNS:

New Feature

Amazon SNS now supports High Throughput FIFO topics. [Learn more](#)

Topics (2)

EditDeletePublish messageCreate topic

Q Search

< 1 >

Name

▲

Type

Day_2-SNS

Standard

Usei um email criado no site que foi fornecido no Teams: pamawe1050@andinews.com



Simple Notification Service

Subscription confirmed!

You have successfully subscribed.

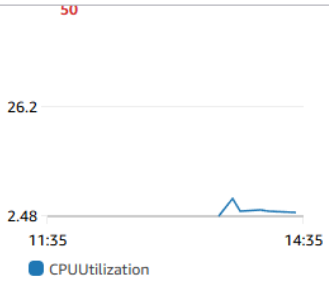
Your subscription's id is:

arn:aws:sns:us-east-1:211125364911:Day_2-SNS:b163408d-58e2-4e73-8949-6e9f6a0f488e

If it was not your intention to subscribe, [click here to unsubscribe](#).

8- Criação do CloudWatch

CloudWatch > Alarms > Create alarm



50

26.2

2.48

11:35 14:35

CPUUtilization

Metric name

CPUUtilization

InstanceId

i-00f88f5f9436f1ccd

Instance name

Day_2-instance

Statistic

Average

Period

5 minutes

Conditions

Threshold type

☒ Static

Use a value as a threshold

☐ Anomaly detection

Use a band as a threshold

Whenever CPUUtilization is...

Define the alarm condition.

☒ Greater

> threshold

☐ Greater/Equal

>= threshold

☐ Lower/Equal

<= threshold

☐ Lower

< threshold

than...

Define the threshold value.

50

Must be a number

Configure actions

Notification

Alarm state trigger

Define the alarm state that will trigger this action.

Remove

☒ In alarm

The metric or expression is outside of the defined threshold.

☐ OK

The metric or expression is within the defined threshold.

☐ Insufficient data

The alarm has just started or not enough data is available.

Send a notification to the following SNS topic

Define the SNS (Simple Notification Service) topic that will receive the notification.

☒ Select an existing SNS topic

☐ Create new topic

☐ Use topic ARN to notify other accounts

Send a notification to...

Day_2-SNS

Only topics belonging to this account are listed here. All persons and applications subscribed to the selected topic will receive notifications.

Email (endpoints)

pamawe1050@andinews.com - [View in SNS Console](#)

Specify metric and conditions

Step 2

Configure actions

Step 3

Add name and description

Step 4

Preview and create

9- Alarme criado

Alarms (1)

☐ Hide Auto Scaling alarms

Clear selection

Create composite alarm

Actions ▾

Create alarm

Alarm state: Any ▾

Alarm type: Any ▾

Actions status: Any ▾

< 1 > | ⚙

<input type="checkbox"/>	Name	State	Last sta
<input type="checkbox"/>	Day_2-Alarm	✔ OK	2025-0

Serviços usados:

- S3
- AWS Glue
- EC2
- SNS
- CloudWatch