**Parcial II – Big\_data e Ing. De Datos**

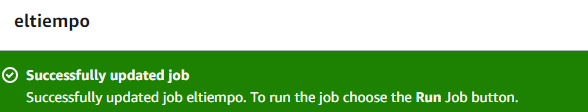
**Estudiantes:** Nestor Morales Calixto

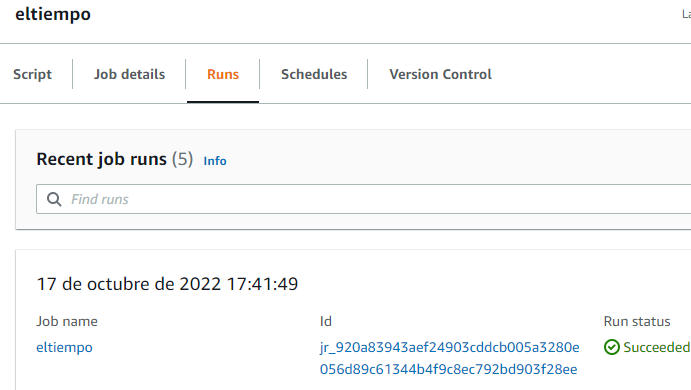
(máximo 2 estudiantes, los 2 puntos son obligatorios sin importar el número de estudiantes)

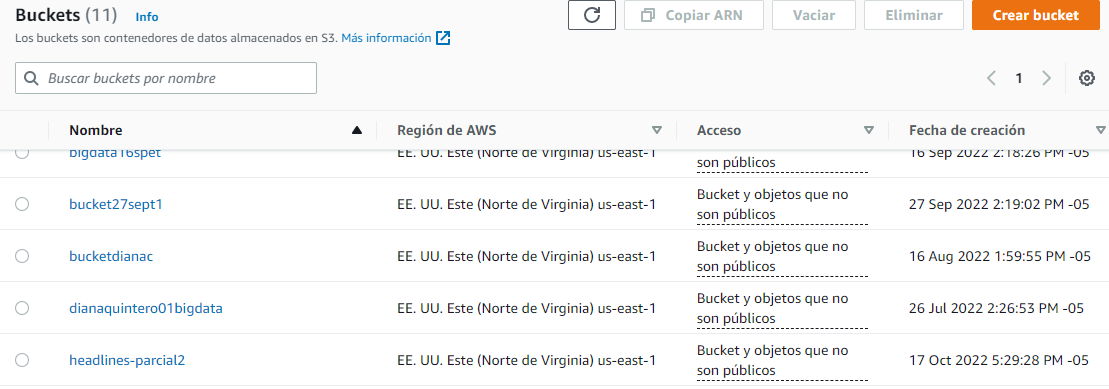
1. **Desarrollar los ETL Workflow en AWS Glue para la descarga de información de periódicos (eltiempo y el espectador):**
2. **Crear un Job en AWS Glue(con un trigger) que descargue cada día la página principal de El Tiempo y El Espectador(o publimetro).**

**La información debe quedar en S3 con la estructura:**

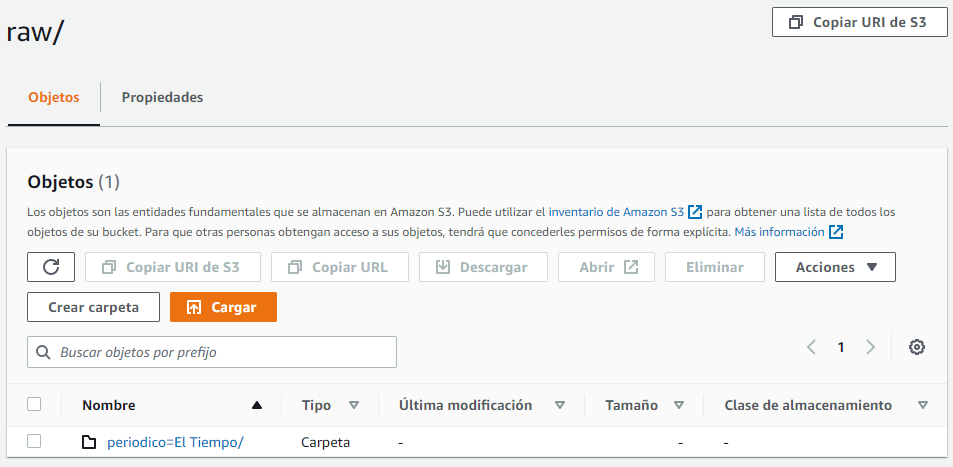
**• s3://bucket/headlines/raw/periodico=xxx/year=xxx/month=xxx/day=xxx**

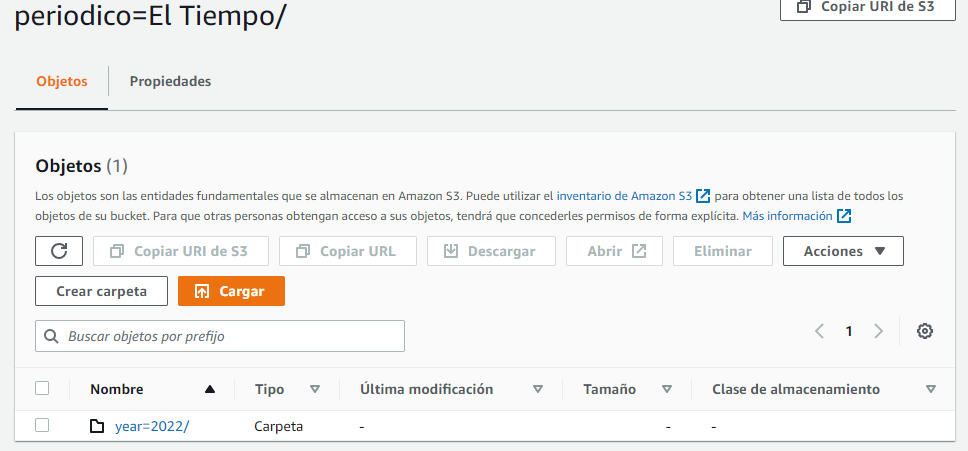


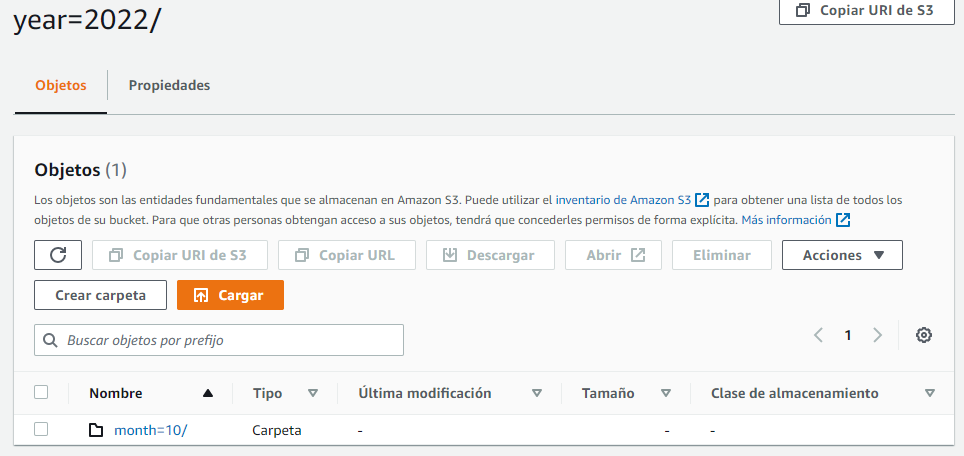


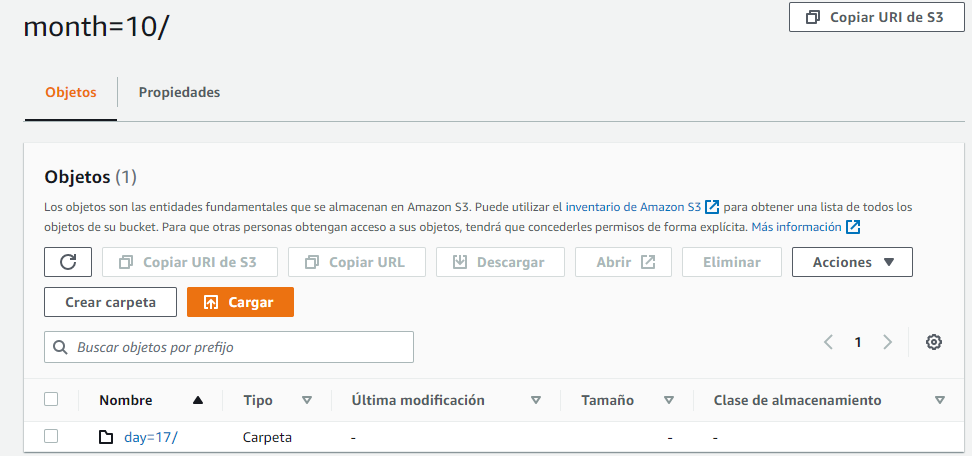


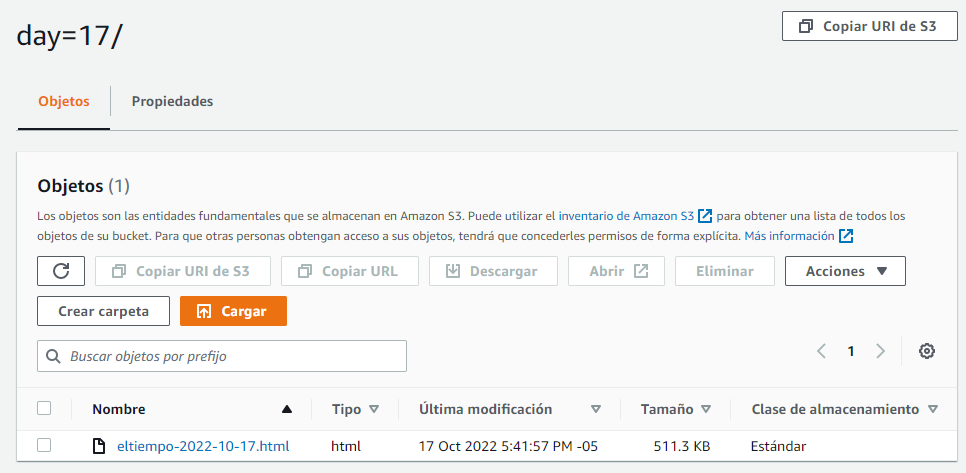




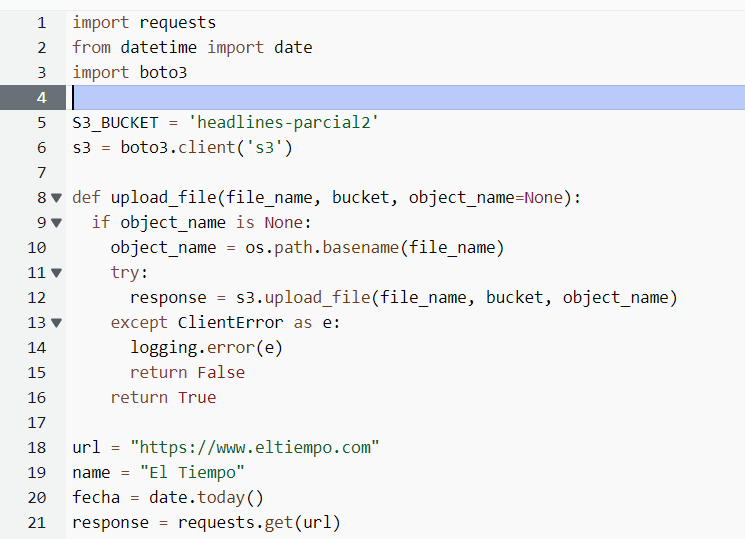


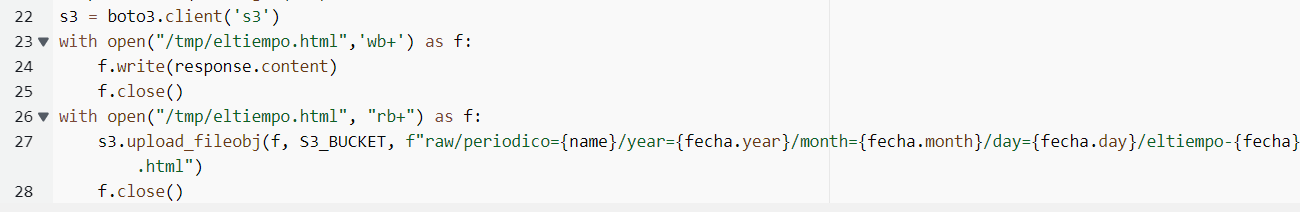










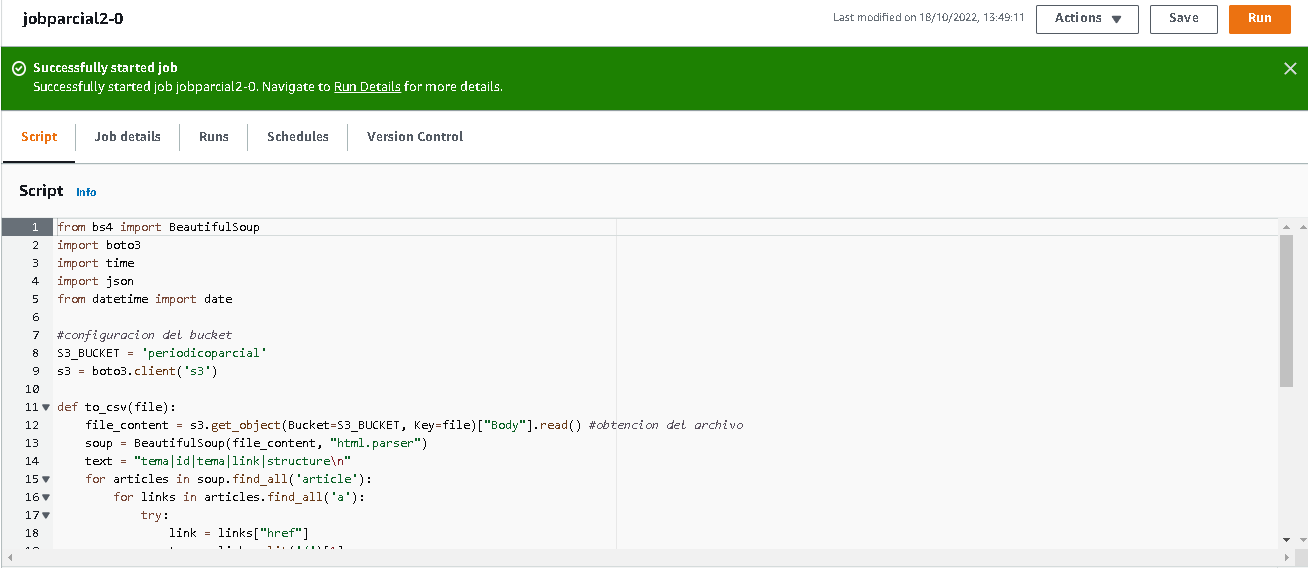


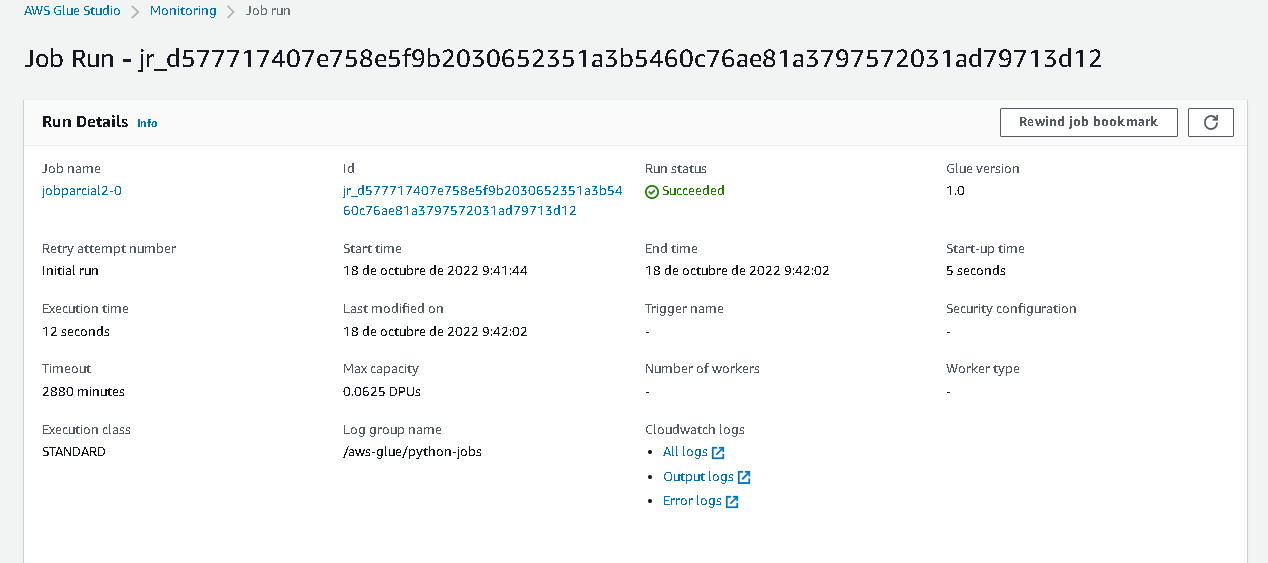
1. **Una vez llega el archivo a la carpeta raw, se debe activar un segundo job que procese los datos que llegaron utilizando Beautifulsoup. Este proceso debe extraer la categoría, el titular y el enlace para cada noticia. Estos datos se deben guardar en un csv en la siguiente ruta:**

**• s3://bucket/headlines/final/periodico=xxx/year=xxx/month=xxx/day=xxx**

**• Este archivo dispara un nuevo job que descarga la noticia y lo guarda en:**

**• s3://bucket/news/raw/periodico=xxx/year=xxx/month=xxx/day=xxx**





1. **Una vez terminados estos jobs, se debe activar un crawler que actualice el catálogo de AWS Glue y permita visualizar los datos en AWS Athena.**

**Para usar paquetes externos revisar:** [**https://aws.amazon.com/es/premiumsupport/knowledge-center/glue-version2-external-python-libraries/**](https://aws.amazon.com/es/premiumsupport/knowledge-center/glue-version2-external-python-libraries/)

**from** bs4 **import** BeautifulSoup

**def** to\_csv(file):

soup **=** BeautifulSoup(file, "html.parser")

text **=** "tema|id|tema|link|structure\n"

**for** articles **in** soup**.**find\_all('article'):

**for** links **in** articles**.**find\_all('a'):

**try**:

link **=** links["href"]

tema **=** link**.**split('/')[1]

id\_news **=** links["id"]

code **=** soup**.**select\_one(f"a#{id\_news}")**.**get\_text(strip**=True**, separator**=**" ")

text **+=** f"{code}|{id\_news}|{tema}|{link}|{links}\n"

**except**:

**...**

**return** text

**texto1** **=** """<article class="image-left video-icon" dtm-name="Contenido3" itemscope="" itemtype="https://schema.org/NewsArticle" id="m1246-1-1247"><div class="recurso" itemtype="https://schema.org/VideoObject" itemprop="video" itemscope="">

<figure class="image-container" width="170" height="113"><a id="m1251-1250-1252" class="image page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295"><picture><source srcset="/files/image\_400\_266/uploads/2022/10/16/634c3fa467652.png" media="(max-width:767px)" class="js-lazy-image--handled"><img style=" display: inline;" alt="" src="https://www.eltiempo.com/files/image\_170\_113/uploads/2022/10/16/634c3fa467652.png" itemprop="thumbnailUrl" class="js-lazy-image--handled" width="170px" height="113px"></picture></a></figure><div class="oculto videoDesc"><span itemprop="name">Avioneta se salió de pista del aeropuerto de Santa Marta y cayó al mar</span><span itemprop="description">En la aeronave se movilizaban dos tripulantes y cuatro pasajeros, entre los que se encuentran una niña.</span><meta itemprop="uploadDate" content="2022-10-16T20:08:02-0500"></div></div>

<div class="article-details">

<div class="category-published">

<a id="m1258-1-1259" class="category page-link colombia colombia--otras-ciudades" href="/colombia/otras-ciudades">Otras Ciudades</a><span id="m1264-2-1265" class="published-at">8:08 pm</span></div>

<h2 class="title-container" itemprop="headline">

<a class="multimediatag page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295">

<span id="m1270-3-1271">Video</span></a>

<a id="m1276-4-1277" class="title page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295">Así fue el accidente aéreo que causó la tragedia de una familia en Santa Marta</a></h2>

</div>

<h3 class="epigraph-container">

<a id="m1282-5-1283" class="epigraph page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295">Un niño murió y otra menor se encuentra en estado grave. Una adulta mayor también permanece en UCI.</a></h3>

<meta itemprop="datePublished" content="2022-10-16"><meta itemprop="dateModified" content="2022-10-16"><meta itemscope="" itemprop="mainEntityOfPage" itemtype="https://schema.org/WebPage" itemid="https://www.eltiempo.com/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295"><span class="oculto" itemprop="author" itemscope="" itemtype="https://schema.org/Person">

<span itemprop="name">Roger Urieles</span>

</span>

<div class="oculto" itemprop="publisher" itemscope="" itemtype="https://schema.org/Organization">

<div itemprop="logo" itemscope="" itemtype="https://schema.org/ImageObject">

<meta itemprop="url" content="https://www.eltiempo.com/bundles/eltiempocms/images/el-tiempo/logo-el-tiempo-azul.jpg"><meta itemprop="width" content="600"><meta itemprop="height" content="60"></div>

<meta itemprop="name" content="EL TIEMPO"><meta itemprop="url" content="https://www.eltiempo.com"></div>

<div class="oculto" id="foto-oculta-video"><figure class="foto-desktop-noticias-seccion" itemprop="image" itemscope="" itemtype="https://schema.org/ImageObject"><img alt="" src="/images/1x1.png" data-original="https://www.eltiempo.comhttps://www.eltiempo.com/files/image\_170\_113/uploads/2022/10/16/634c3fa467652.png" width="170" height="113"><meta itemprop="width" content="170"><meta itemprop="height" content="113"><meta itemprop="url" content="https://www.eltiempo.comhttps://www.eltiempo.com/files/image\_170\_113/uploads/2022/10/16/634c3fa467652.png"></figure></div></article>"""

**texto2** **=** """<article class="image-left" dtm-name="Contenido4" itemscope="" itemtype="https://schema.org/NewsArticle" id="m1290-2-1291"><div class="recurso" itemtype="https://schema.org/NewsArticle">

<figure class="image-container" width="170" height="113"><a id="m1295-1294-1296" class="image page-link" itemprop="image" itemscope="" itemtype="https://schema.org/ImageObject" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269"><picture><source srcset="/files/image\_400\_266/uploads/2022/10/16/634c3a10e5792.png" media="(max-width:767px)" class="js-lazy-image--handled"><img style=" display: inline;" alt="" src="/files/image\_170\_113/uploads/2022/10/16/634c3a10e5792.png" class="js-lazy-image--handled" width="170px" height="113px"></picture><meta itemprop="width" content="170"><meta itemprop="height" content="113"><meta itemprop="url" content="https://www.eltiempo.com/files/image\_170\_113/uploads/2022/10/16/634c3a10e5792.png"></a></figure></div>

<div class="article-details">

<div class="category-published">

<a id="m1302-1-1303" class="category page-link colombia colombia--otras-ciudades" href="/colombia/otras-ciudades">Otras Ciudades</a><span id="m1308-2-1309" class="published-at">3:18 pm</span></div>

<h2 class="title-container" itemprop="headline">

<a class="multimediatag page-link" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269">

<span id="m1314-3-1315">Video</span></a>

<a id="m1320-4-1321" class="title page-link" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269">Muere niño que fue embestido por avión que se salió de pista en Santa Marta</a></h2>

</div>

<h3 class="epigraph-container">

<a id="m1326-5-1327" class="epigraph page-link" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269">Otra menor está en delicado estado de salud. La tripulación resultó con golpes leves.</a></h3>

<meta itemprop="datePublished" content="2022-10-16"><meta itemprop="dateModified" content="2022-10-16"><meta itemscope="" itemprop="mainEntityOfPage" itemtype="https://schema.org/WebPage" itemid="https://www.eltiempo.com/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269"><span class="oculto" itemprop="author" itemscope="" itemtype="https://schema.org/Person">

<span itemprop="name">Roger Urieles</span>

</span>

<div class="oculto" itemprop="publisher" itemscope="" itemtype="https://schema.org/Organization">

<div itemprop="logo" itemscope="" itemtype="https://schema.org/ImageObject">

<meta itemprop="url" content="https://www.eltiempo.com/bundles/eltiempocms/images/el-tiempo/logo-el-tiempo-azul.jpg"><meta itemprop="width" content="600"><meta itemprop="height" content="60"></div>

<meta itemprop="name" content="EL TIEMPO"><meta itemprop="url" content="https://www.eltiempo.com"></div>

</article>"""

**texto3** **=**"""<article class="image-left" dtm-name="Contenido5" itemscope="" itemtype="https://schema.org/NewsArticle" id="m1334-3-1335"><div class="recurso" itemtype="https://schema.org/NewsArticle">

<figure class="image-container" width="170" height="113"><a id="m1339-1338-1340" class="image page-link" itemprop="image" itemscope="" itemtype="https://schema.org/ImageObject" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140"><picture><source srcset="/files/image\_400\_266/uploads/2022/02/27/621c269ca7f84.jpeg" media="(max-width:767px)" class="js-lazy-image--handled"><img style=" display: inline;" alt="" src="/files/image\_170\_113/uploads/2022/02/27/621c269ca7f84.jpeg" class="js-lazy-image--handled" width="170px" height="113px"></picture><meta itemprop="width" content="170"><meta itemprop="height" content="113"><meta itemprop="url" content="https://www.eltiempo.com/files/image\_170\_113/uploads/2022/02/27/621c269ca7f84.jpeg"></a></figure></div>

<div class="article-details">

<div class="category-published">

<a id="m1346-1-1347" class="category page-link justicia justicia--cortes" href="/justicia/cortes">Cortes</a><span id="m1352-2-1353" class="published-at">12:00 am</span></div>

<h2 class="title-container" itemprop="headline">

<a class="multimediatag page-link" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140">

<span id="m1358-3-1359">En Vivo</span></a>

<a id="m1364-4-1365" class="title page-link" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140">David Barguil: las interceptaciones que tienen en líos al exsenador en la Corte</a></h2>

</div>

<h3 class="epigraph-container">

<a id="m1370-5-1371" class="epigraph page-link" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140">EL TIEMPO revela el expediente en contra del exsenador quien fue llamado a indagatoria.</a></h3>

<meta itemprop="datePublished" content="2022-10-16"><meta itemprop="dateModified" content="2022-10-15"><meta itemscope="" itemprop="mainEntityOfPage" itemtype="https://schema.org/WebPage" itemid="https://www.eltiempo.com/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140"><span class="oculto" itemprop="author" itemscope="" itemtype="https://schema.org/Person">

<span itemprop="name">Alejandra Bonilla Mora</span>

</span>

<div class="oculto" itemprop="publisher" itemscope="" itemtype="https://schema.org/Organization">

<div itemprop="logo" itemscope="" itemtype="https://schema.org/ImageObject">

<meta itemprop="url" content="https://www.eltiempo.com/bundles/eltiempocms/images/el-tiempo/logo-el-tiempo-azul.jpg"><meta itemprop="width" content="600"><meta itemprop="height" content="60"></div>

<meta itemprop="name" content="EL TIEMPO"><meta itemprop="url" content="https://www.eltiempo.com"></div>

</article>"""

**sub1=**"""tema|id|tema|link|structure\n|m1251-1250-1252|colombia|/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295|<a class="image page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295" id="m1251-1250-1252"><picture><source class="js-lazy-image--handled" media="(max-width:767px)" srcset="/files/image\_400\_266/uploads/2022/10/16/634c3fa467652.png"/><img alt="" class="js-lazy-image--handled" height="113px" itemprop="thumbnailUrl" src="https://www.eltiempo.com/files/image\_170\_113/uploads/2022/10/16/634c3fa467652.png" style=" display: inline;" width="170px"/></picture></a>\nOtras Ciudades|m1258-1-1259|colombia|/colombia/otras-ciudades|<a class="category page-link colombia colombia--otras-ciudades" href="/colombia/otras-ciudades" id="m1258-1-1259">Otras Ciudades</a>\nAsí fue el accidente aéreo que causó la tragedia de una familia en Santa Marta|m1276-4-1277|colombia|/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295|<a class="title page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295" id="m1276-4-1277">Así fue el accidente aéreo que causó la tragedia de una familia en Santa Marta</a>\nUn niño murió y otra menor se encuentra en estado grave. Una adulta mayor también permanece en UCI.|m1282-5-1283|colombia|/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295|<a class="epigraph page-link" href="/colombia/otras-ciudades/asi-fue-accidente-aereo-que-causo-la-tragedia-en-una-familia-en-santa-marta-710295" id="m1282-5-1283">Un niño murió y otra menor se encuentra en estado grave. Una adulta mayor también permanece en UCI.</a>\n"""

**sub2=**"""tema|id|tema|link|structure\n|m1295-1294-1296|colombia|/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269|<a class="image page-link" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269" id="m1295-1294-1296" itemprop="image" itemscope="" itemtype="https://schema.org/ImageObject"><picture><source class="js-lazy-image--handled" media="(max-width:767px)" srcset="/files/image\_400\_266/uploads/2022/10/16/634c3a10e5792.png"/><img alt="" class="js-lazy-image--handled" height="113px" src="/files/image\_170\_113/uploads/2022/10/16/634c3a10e5792.png" style=" display: inline;" width="170px"/></picture><meta content="170" itemprop="width"/><meta content="113" itemprop="height"/><meta content="https://www.eltiempo.com/files/image\_170\_113/uploads/2022/10/16/634c3a10e5792.png" itemprop="url"/></a>\nOtras Ciudades|m1302-1-1303|colombia|/colombia/otras-ciudades|<a class="category page-link colombia colombia--otras-ciudades" href="/colombia/otras-ciudades" id="m1302-1-1303">Otras Ciudades</a>\nMuere niño que fue embestido por avión que se salió de pista en Santa Marta|m1320-4-1321|colombia|/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269|<a class="title page-link" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269" id="m1320-4-1321">Muere niño que fue embestido por avión que se salió de pista en Santa Marta</a>\nOtra menor está en delicado estado de salud. La tripulación resultó con golpes leves.|m1326-5-1327|colombia|/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269|<a class="epigraph page-link" href="/colombia/otras-ciudades/avioneta-en-santa-marta-se-accidento-muerte-nina-pasajera-710269" id="m1326-5-1327">Otra menor está en delicado estado de salud. La tripulación resultó con golpes leves.</a>\n"""

s3**=**"""tema|id|tema|link|structure\n|m1339-1338-1340|justicia|/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140|<a class="image page-link" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140" id="m1339-1338-1340" itemprop="image" itemscope="" itemtype="https://schema.org/ImageObject"><picture><source class="js-lazy-image--handled" media="(max-width:767px)" srcset="/files/image\_400\_266/uploads/2022/02/27/621c269ca7f84.jpeg"/><img alt="" class="js-lazy-image--handled" height="113px" src="/files/image\_170\_113/uploads/2022/02/27/621c269ca7f84.jpeg" style=" display: inline;" width="170px"/></picture><meta content="170" itemprop="width"/><meta content="113" itemprop="height"/><meta content="https://www.eltiempo.com/files/image\_170\_113/uploads/2022/02/27/621c269ca7f84.jpeg" itemprop="url"/></a>\nCortes|m1346-1-1347|justicia|/justicia/cortes|<a class="category page-link justicia justicia--cortes" href="/justicia/cortes" id="m1346-1-1347">Cortes</a>\nDavid Barguil: las interceptaciones que tienen en líos al exsenador en la Corte|m1364-4-1365|justicia|/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140|<a class="title page-link" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140" id="m1364-4-1365">David Barguil: las interceptaciones que tienen en líos al exsenador en la Corte</a>\nEL TIEMPO revela el expediente en contra del exsenador quien fue llamado a indagatoria.|m1370-5-1371|justicia|/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140|<a class="epigraph page-link" href="/justicia/cortes/david-barguil-las-interceptaciones-que-lo-tienen-el-lios-en-la-corte-710140" id="m1370-5-1371">EL TIEMPO revela el expediente en contra del exsenador quien fue llamado a indagatoria.</a>\n"""

**import** unittest

**class** MyTestCase(unittest**.**TestCase):

**ef** test\_function(self):

self**.**assertEqual(to\_csv(t1),s1)

self**.**assertEqual(to\_csv(t2),s2)

self**.**assertEqual(to\_csv(t3),s3)

unittest**.**main(argv**=**[''], verbosity**=**3, exit**=False**)

1. **Crear un Job que inserte la información en una base de datos MYSQL(usando aws glue connectors y aws job)**

import sys

from awsglue.transforms import \*

from awsglue.utils import getResolvedOptions

from pyspark.context import SparkContext

from awsglue.context import GlueContext

from awsglue.job import Job

args = getResolvedOptions(sys.argv, ["1d\_parcial"])

sc = SparkContext()

glueContext = GlueContext(sc)

spark = glueContext.spark\_session

job = Job(glueContext)

job.init(args["1d\_parcial"], args)

# Se genera un nodo de s3

S3bucket\_node1 = glueContext.create\_dynamic\_frame.from\_catalog(

database="news", table\_name="final", transformation\_ctx="s3\_node1"

)

# Script generated for node ApplyMapping

Mapping\_node2 = ApplyMapping.apply(

frame=s3\_node1,

mappings=[

("tema", "string", "tema", "string"),

("id", "string", "id", "string"),

("categoria", "string", "categoria", "string"),

("link", "string", "link", "string"),

("estructura", "string", "estructura", "string"),

("col5", "string", "col5", "string"),

("periodico", "string", "periodico", "string"),

("year", "bigint", "year", "bigint"),

("month", "bigint", "month", "string"),

("day", "bigint", "day", "string"),

],

transformation\_ctx="mapping\_node2",

)

# Script generated for node S3 bucket

node3 = glueContext.getSink(

path="s3://periodicoparcial/news/raw/",

connection\_type="s3",

updateBehavior="UPDATE\_IN\_DATABASE",

partitionKeys=["periodico", "year", "month", "day"],

compression="snappy",

enableUpdateCatalog=True,

transformation\_ctx="node3",

)

S3bucket\_node3.setCatalogInfo(catalogDatabase="news", catalogTableName="news\_daily")

node3.setFormat("glueparquet")

node3.writeFrame(Mapping\_node2)

job.commit()

**Nota:**

**\* Hay que crear un proyecto de zappa para cada función o jobs de Glue.**

**\*\* Se debe entregar el código en github(utilizar ramas, commits, código limpio, código comentado,**

**pruebas unitarias si aplica).**