

UNIVERSIDAD EAFIT

Tópicos Especiales en Telemática

Laboratorio 3-3

Juan Manuel Gómez - Ingeniería de Sistemas

(jmgomezp@eafit.edu.co)

Profesor: Edwin Nelson Montoya

Medellín, 7 de noviembre de 2024

Procesamiento del dataset Casos_positivos_de_COVID-19_en_Colombia-100K.csv usando PySpark en Google Colab

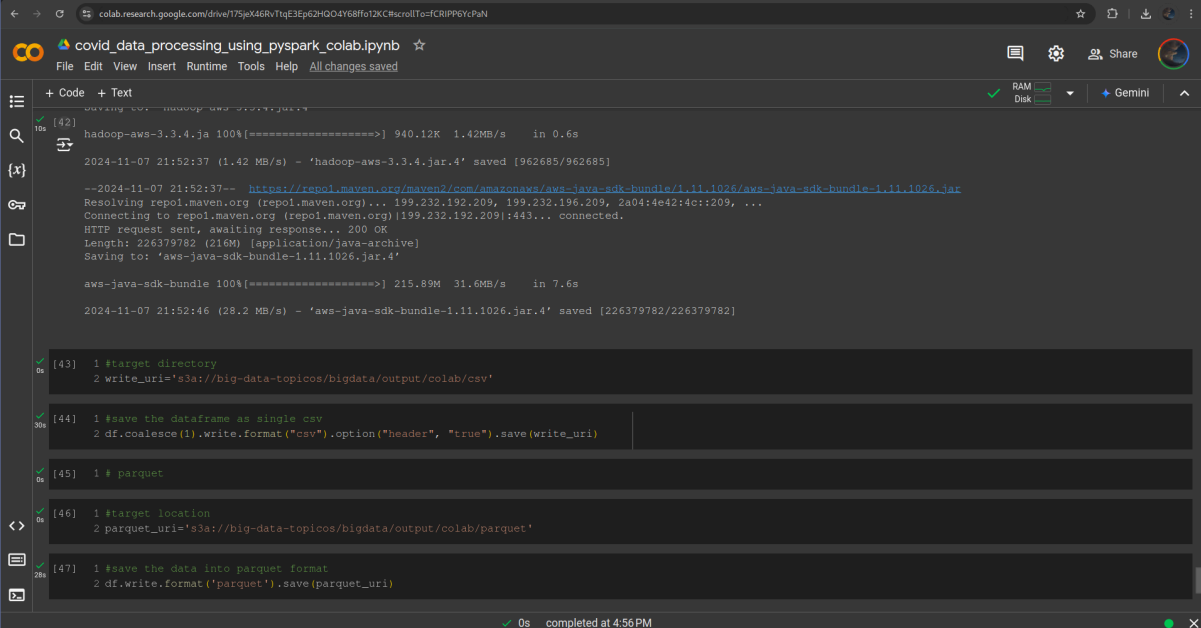
En esta ocasión se hizo una exploración de los datos referentes a los casos de COVID-19 en Colombia con los datos proporcionados por el Ministerio de Salud. Luego de haber hecho todo el procesamiento se subieron los archivos al Bucket de S3.

En este caso se desarrollaron los siguientes puntos:

- Columnas.
- Tipos de datos.
- Seleccionar algunas columnas.
- Renombrar columnas.
- Agregar columnas.
- Borrar columnas.
- Filtrar datos.
- Ejecutar alguna función UDF sobre alguna columna creando una nueva.
- Ejecutar una función lambda sobre alguna columna creando una nueva.

En este caso se utilizó un Notebook de Google Colab con PySpark. Dentro del repositorio de GitHub se encuentra el Notebook (covid_data_processing_using_pyspark_colab.ipynb) con los resultados de este laboratorio.

Output en el Bucket de S3



```
colab.research.google.com/drive/775jK46RvTtqE3Ep62HQO4Y68f012KC#scrollTo=ICRIPPOYcPaN
covid_data_processing_using_pyspark_colab.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
[42] hadoop-aws-3.3.4.jar 100%[=====] 940.12K 1.42MB/s in 0.6s
2024-11-07 21:52:37 (1.42 MB/s) - 'hadoop-aws-3.3.4.jar.4' saved [962685/962685]
--2024-11-07 21:52:37-- https://repol.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.1026/aws-java-sdk-bundle-1.11.1026.jar
Resolving repol.maven.org (repol.maven.org)... 199.232.192.209, 199.232.196.209, 2a04:4e42:4c::209, ...
Connecting to repol.maven.org (repol.maven.org)|199.232.192.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 226379782 (216M) [application/java-archive]
Saving to: 'aws-java-sdk-bundle-1.11.1026.jar.4'
aws-java-sdk-bundle 100%[=====] 215.89M 31.6MB/s in 7.6s
2024-11-07 21:52:46 (28.2 MB/s) - 'aws-java-sdk-bundle-1.11.1026.jar.4' saved [226379782/226379782]
[43] 1 #target directory
2 write_uri='s3a://big-data-topicos/bigdata/output/colab/csv'
[44] 1 #save the dataframe as single csv
2 df.coalesce(1).write.format("csv").option("header", "true").save(write_uri)
[45] 1 # parquet
[46] 1 #target location
2 parquet_uri='s3a://big-data-topicos/bigdata/output/colab/parquet'
[47] 1 #save the data into parquet format
2 df.write.format('parquet').save(parquet_uri)
0s completed at 4:56 PM
```

us-east-1.console.aws.amazon.com/s3/buckets/big-data-topicos?region=us-east-1&bucketType=general&prefix=bigdata/output/colab/&showversions=false

Services Search [Alt+S]

N. Virginia voclabs/user3070189-jmgomezp@eaflit.edu.co @ 2631-7354-0036

Amazon S3

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 7

AWS Marketplace for S3

Amazon S3 > Buckets > big-data-topicos > bigdata/ > output/ > colab/

colab/

Copy S3 URI

Objects Properties

Objects (2) Info

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

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

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

us-east-1.console.aws.amazon.com/s3/buckets/big-data-topicos?region=us-east-1&bucketType=general&prefix=bigdata/output/colab/csv/&showversions=false

Services Search [Alt+S]

N. Virginia voclabs/user3070189-jmgomezp@eaflit.edu.co @ 2631-7354-0036

Amazon S3

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 7

AWS Marketplace for S3

Amazon S3 > Buckets > big-data-topicos > bigdata/ > output/ > colab/ > csv/

csv/

Copy S3 URI

Objects Properties

Objects (2) Info

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	_SUCCESS	-	November 7, 2024, 16:53:16 (UTC-05:00)	0 B	Standard
<input type="checkbox"/>	part-00000-ead3114b-92f0-462f-bf83-f88b35c88ae4-c000.csv	csv	November 7, 2024, 16:53:12 (UTC-05:00)	17.5 MB	Standard

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

us-east-1.console.aws.amazon.com/s3/buckets/big-data-topicos?region=us-east-1&bucketType=general&prefix=bigdata/output/colab/parquet/&showversions=false

ServicesSearch[Alt+S]

N. Virginiavoclabs/user3070189-jngomezp@eafft.edu.co @ 2631-7354-0036

Amazon S3

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

AWS Marketplace for S3

Amazon S3 > Buckets > big-data-topicos > bigdata/ > output/ > colab/ > parquet/

parquet/

Copy S3 URI

Objects

Properties

Objects (3) info

Copy S3 URICopy URLDownloadOpenDeleteActionsCreate folderUpload

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](#)

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	_SUCCESS	-	November 7, 2024, 16:53:45 (UTC-05:00)	0 B	Standard
<input type="checkbox"/>	part-00000-79e18a78-9fe1-438e-af63-d17b573b670f-c000.snappy.parquet	parquet	November 7, 2024, 16:53:36 (UTC-05:00)	716.9 KB	Standard
<input type="checkbox"/>	part-00001-79e18a78-9fe1-438e-af63-d17b573b670f-c000.snappy.parquet	parquet	November 7, 2024, 16:53:41 (UTC-05:00)	431.8 KB	Standard

CloudShellFeedback

© 2024, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences