

UNIVERSIDAD EAFIT

Tópicos Especiales En Telemática

Laboratorio 3-3 PYSPARK

Miguel Angel Hoyos Velasquez

Medellín, 18 de Noviembre 2024

Recibe notificaciones en tu ordenador sobre archivos compartidos y eventos importantes

Activar X

Untitled6.ipynb

```
[7]   version
      v3.5.3
[8]   Mas...
      local[*]
[x]   AppName
      data_processing

[10]  1 df = spark.read.csv('/content/gdrive/MyDrive/content/gdrive/Casos positivos de COVID-19 en Colombia-100K.csv', inferSchema=True, header=True)

[11]  1 df.columns
      [
        'fecha reporte web',
        'ID de caso',
        'Fecha de notificación',
        'Código DIVIRPA departamento',
        'Nombre departamento',
        'Código DIVIRPA municipio',
        'Nombre municipio',
        'Edad',
        'Unidad de medida de edad',
        'Sexo',
        'Tipo de contagio',
        'Ubicación del caso',
        'Estado',
        'Codigo ISO del país',
        'Nombre del país',
        'Recuperado',
        'Fecha de inicio de síntomas',
        'Fecha de alta',
        'Fecha de diagnóstico',
        'Fecha de recuperación',
        'Tipo de recuperación',
        'Pertenencia étnica',
        'Nombre del grupo étnico'
      ]
```

0 s se ejecutó 6:03 p.m.

Inicialmente se puede apreciar el uso de Google Colab con una fuente de datos en Google drive. El dataset utilizado es el de datos de covid, encontrado en el repositorio del curso.

```

[6] 14 clasificar_edad_udf = udf(clasificar_edad, StringType())
[6] 15 df = df.withColumn("categoria_edad", clasificar_edad_udf(df.edad))
[6] 16 df.show()

```

Aquí la primera exploración de las columnas para validar el correcto funcionamiento. Toda la exploración del dataset se encuentra en el notebook adjunto.

Posteriormente se procedió a trabajar con los datos desde S3.

Nombre	Carpeta	Tipo	Tamaño	Estado	Error
Casos_posit... -		text/csv	16.3 MB	Realizado corre -	

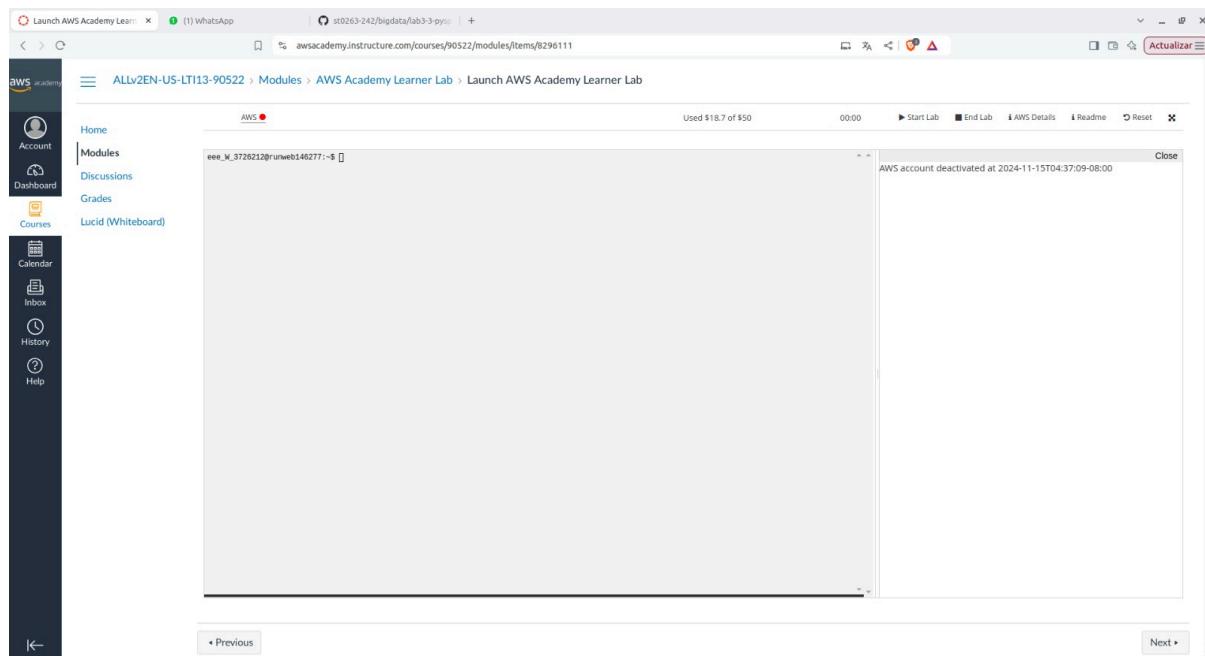
Las preguntas de negocio se responden así mismo en el notebook.

Aquí se adjuntan los resultados para cada pregunta. Cambié un poco la forma de cargar los datos a S3 por una que me resultaba más intuitiva. En el notebook se puede evidenciar.

Se ve la carga de manera efectiva para todos los directorios de resultados cargados. Se tomó un directorio aleatorio de todos los cargados para evidencia, para no hacer muy extensa esta sección.

El segundo componente de este laboratorio se realizó con el cluster EMR de aws, desde jupyterhub, para el cual también se adjuntan todos los notebooks de respuesta de preguntas de negocio y exploración de los datos.

Tuve un pequeño inconveniente, y los clusters EMR no se estaban creando.



Hubo un momento, en donde la cuenta quedó completamente suspendida. Por lo cual los buckets quedaron inaccesibles. Sin embargo, se adjunta la evidencia tanto del error, como de los buckets cuando estaban funcionales. (Este inconveniente también se reportó vía correo electrónico)

ID del clúster	Nombre del clúster	Estado	Hora de creación (UTC-05:00)	Tiempo transcurrido	Horas de instancia normalizadas
j-2U16CNLIKUH4	My cluster MAHOYOS	Terminado con errores Error de instancia	13 de noviembre de 2024 17:35	52 segundos	0
j-2N561GDOBCPCR	My cluster MAHOYOS	Terminado con errores Error interno	13 de noviembre de 2024 17:15	48 minutos, 57 segundos	0
j-ST059C4CXXJBU	My cluster Labs	Terminado con errores Error de instancia	13 de noviembre de 2024 17:12	54 segundos	0
j-18VC1EUX3N3M7	My cluster MAHOYOS	Terminado con errores Error interno	13 de noviembre de 2024 13:16	48 minutos, 53 segundos	0
j-3HDJ931KOMIOA	My cluster MAHOYOS	Terminado con errores Error de instancia	13 de noviembre de 2024 13:12	52 segundos	0
j-59UE7QF2YAOIGP	My cluster MAHOYOS	Terminado con errores Error interno	13 de noviembre de 2024 12:26	49 minutos, 8 segundos	0
j-5PQEYMP4TNLE	My cluster MAHOYOS	Terminado con errores Error de instancia	13 de noviembre de 2024 12:22	53 segundos	0
j-2FDNTZ6RZFL8I	My cluster MAHOYOS	Terminado Solicitud de usuario	1 de noviembre de 2024 10:45	23 minutos, 18 segundos	24
j-1R650NLGRY9IW	My cluster MAHOYOS	Terminado Solicitud de usuario	25 de octubre de 2024 10:31	1 hora, 4 minutos	24
j-O22TPMWPY5HS	My cluster MAHOYOS	Terminado Solicitud de usuario	18 de octubre de 2024 15:47	29 minutos, 2 segundos	24
j-1THAHAMWMVFKOE	My cluster MAHOYOS	Terminado con errores Error de instancia	18 de octubre de 2024 15:39	4 minutos, 25 segundos	0

Obviando estos detalles, procedo a adjuntar las evidencias restantes del laboratorio.

```

In [34]: # Dataframe
top_cities_df = df.groupby("Nombre municipio") \
    .agg(count("ID de caso").alias("total_casos")) \
    .orderBy(desc("total_casos")) \
    .limit(10)
top_cities_df.show()

```

Nombre municipio	total_casos
BOGOTÁ	30616
BARRANQUILLA	13065
CARTAGENA	8333
CALI	7747
SOLEDAD	6233
LETICIA	2194
MEDELLÍN	2137
TUMACO	1501
BUENAVENTURA	1453
QUIBDO	1367


```

In [36]: spark.sql("""
SELECT 'Nombre municipio', COUNT('ID de caso') AS total_casos
FROM covid_data
GROUP BY 'Nombre municipio'
ORDER BY total_casos DESC
LIMIT 10
""").show()

```

Nombre municipio	total_casos
BOGOTÁ	30616
BARRANQUILLA	13065
CARTAGENA	8333
CALI	7747
SOLEDAD	6233
LETICIA	2194
MEDELLÍN	2137
TUMACO	1501
BUENAVENTURA	1453
QUIBDO	1367

In []:

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
age_distribution/	Carpetas	-	-	-
gender_distribution/	Carpetas	-	-	-
top_cities/	Carpetas	-	-	-
top_days/	Carpetas	-	-	-
top_departments/	Carpetas	-	-	-

Aquí podemos apreciar las carpetas de outputs luego de responder las preguntas de negocio.

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with various AWS services like CloudWatch Metrics, Lambda, and Organizations. The main area displays a list of objects in a bucket named 'mahoyosv1'. The objects listed are 'SUCCESS' and 'part-00000-e59ada78-4e1d-4540-0564-fcb3b9786270-0000.snappy.parquet'. The 'part-' object is identified as a parquet file.

Adicional, la creación satisfactoria del cluster EMR.

The screenshot shows the AWS EMR console. It displays the details of a cluster named 'Mi cluster'. The 'Resumen' section provides an overview of the cluster's status. The 'Información del clúster' section lists the cluster ID (j-LBEXVIGT4BLZ), configuration, and capacity (1 Primary (Principal) - 1 Principal - 2 Tarea). The 'Aplicaciones' section lists installed applications like HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Spark 3.5.1, Sqoop 1.4.7, Zeppelin 0.11.1, and ZooKeeper 3.9.1. The 'Administración de clústeres' section shows log storage in Amazon S3 (aws-logs-744115299053-us-east-1/elasticmapreduce). The 'Estado y hora' section shows the cluster is in 'Esperando' state, last updated 1 minute ago, with a public DNS of ec2-54-146-219-132.compute-1.amazonaws.com. The 'Terminación del clúster y reemplazo de nodos' section has a 'Tiempo de inactividad' of 1 hora and 'Reemplazo de nodos en mal estado' set to Activado. Below the summary, there are tabs for 'Propiedades', 'Acciones de arranque', 'Instancias (hardware)', 'Pasos', 'Aplicaciones', 'Configuraciones', 'Monitorización', 'Eventos', and 'Etiquetas (0)'.