databricks API data

(https://databricks.com)

mount storage act

2

```
dbutils.fs.mount(
    source="wasbs://raw-api@casestudy1new.blob.core.windows.net",
    mount_point = "/mnt/raw-api",
    extra_configs={"fs.azure.account.key.casestudy1new.blob.core.windows.net": dbutils.secrets.get(scope =
    "casestudy", key = "storage")})
```

```
True
```

Read the users.parquet File

4

```
# Read the Parquet file into a DataFrame
file_path = "/mnt/raw-api/users.parquet"
df = spark.read.parquet(file_path)

# Display the DataFrame for inspection
df.show()
```

```
33263|-31.8129|  62.5342|       (254)954-1289| demarco.info|       Keebler LLC|User-centric faul...|revolutioni
ze end...|
|  6|Mrs. Dennis Schulist|Leopoldo_Corkery|Karley_Dach@jaspe...|Norberto Crossing| Apt. 950| South Christy|2350
5-1337|-71.4197|  71.7478|1-477-935-8478 x6430|      ola.org| Considine-Lockman|Synchronised bott...|e-enable i
nnovati...|
|  7|     Kurtis Weissnat|    Elwyn.Skiles|Telly.Hoeger@bill...|        Rex Trail|Suite 280|     Howemouth|5880
4-1099| 24.8918|  21.8984|        210.067.6132|     elvis.io|       Johns Group|Configurable mult...|generate e
nterpri...|
|  8|Nicholas Runolfsd...|   Maxime_Nienow|Sherwood@rosamond.me| Ellsworth Summit|Suite 729|     Aliyaview|
45169|-14.3990|-120.7677|   586.493.6943 x140| jacynthe.com|   Abernathy Group|Implemented secon...|e-enable ex
tensib...|
|  9|     Glenna Reichert|        Delphine|Chaim_McDermott@d...|       Dayna Park|Suite 449|Bartholomebury|7649
5-3109| 24.6463|-168.8889|(775)976-6794 x41206|   conrad.com|     Yost and Sons|Switchable contex...|aggregate 
real-ti...|
| 10|  Clementina DuBuque|  Moriah.Stanton|Rey.Padberg@karin...|  Kattie Turnpike|Suite 198|   Lebsackbury|3142
8-2261|-38.2386|  57.2232|        024-648-3804|  ambrose.net|       Hoeger LLC|Centralized empow...|target end
-to-end...|
+---+--------------------+----------------+--------------------+-----------------+---------+--------------+----
------+--------+---------+--------------------+-------------+------------------+--------------------+----------
----------+
```

5

```
# Get the total number of rows in the DataFrame
total_rows = df.count()

# Display the total number of rows
print(f"Total number of rows: {total_rows}")
```

```
Total number of rows: 10
```

Inspect the Data

7

```
# Print schema of the DataFrame
df.printSchema()

# Get a summary of the data
df.describe().show()
```

```
+-----------------+-------------------+-----------------+-----------------+-----------+---------------+--------
------------+--------------------+
|  count|                10|              10|       10|                  10|           10|       10|         10
|               10|                 10|               10|               10|         10|             10|        
10|                  10|
|   mean|               5.5|            NULL|     NULL|                NULL|         NULL|     NULL|       NULL
|          39216.0|-22.675190000000004|        -24.09295|             NULL|       NULL|           NULL|        
NULL|                NULL|
| stddev|3.0276503540974917|            NULL|     NULL|                NULL|         NULL|     NULL|       NULL
|8418.813336807034|  37.66371590289961|98.37461287753563|             NULL|       NULL|           NULL|        
NULL|                NULL|
|    min|                 1|Chelsey Dietrich|Antonette|Chaim_McDermott@d...|   Dayna Park| Apt. 556|  Aliyaview
|       23505-1337|           -14.3990|        -120.7677|    (254)954-1289|ambrose.net|Abernathy Group|Centrali
zed empow...|aggregate real-ti...|
|    max|                10|Patricia Lebsack| Samantha|Telly.Hoeger@bill...|Victor Plains|Suite 879|Wisokyburgh
|       92998-3874|            29.4572|          81.1496|586.493.6943 x140|ramiro.info|  Yost and Sons|User-cen
tric faul...|transition cuttin...|
+-------+------------------+----------------+---------+--------------------+-------------+---------+-----------
+-----------------+-------------------+-----------------+-----------------+-----------+---------------+--------
------------+--------------------+
```

data cleaning

""" delete the row where the username is "Samantha """

9

```
from pyspark.sql import functions as F
from pyspark.sql.window import Window

# Step 1: Filter out the row where username is 'Samantha'
df_cleaned = df.filter(df.username != 'Samantha')

# Step 2: Re-index the `id` column to make sure it is consecutive
# This assumes you have an existing 'id' column and want to reassign the values.
df_cleaned = df_cleaned.withColumn('id', F.row_number().over(Window.orderBy('username')))

# Step 3: Show the cleaned DataFrame
df_cleaned.show()
```

```
4-1099| 24.8918|  21.8984|        210.067.6132|     elvis.io|      Johns Group|Configurable mult...|generate en
terpri...|
|  5|    Chelsey Dietrich|          Kamren|Lucio_Hettinger@a...|     Skiles Walks|Suite 351|    Roscoeview|
33263|-31.8129|  62.5342|       (254)954-1289| demarco.info|      Keebler LLC|User-centric faul...|revolutioniz
e end...|
|  6|    Patricia Lebsack|        Karianne|Julianne.OConner@...|      Hoeger Mall| Apt. 692|   South Elvis|5391
9-4257| 29.4572|-164.2990|   493-170-9623 x156|     kale.biz|    Robel-Corkery|Multi-tiered zero...|transition 
cuttin...|
|  7|Mrs. Dennis Schulist|Leopoldo_Corkery|Karley_Dach@jaspe...|Norberto Crossing| Apt. 950| South Christy|2350
5-1337|-71.4197|  71.7478|1-477-935-8478 x6430|      ola.org|Considine-Lockman|Synchronised bott...|e-enable in
novati...|
|  8|Nicholas Runolfsd...|   Maxime_Nienow|Sherwood@rosamond.me| Ellsworth Summit|Suite 729|     Aliyaview|
45169|-14.3990|-120.7677|   586.493.6943 x140| jacynthe.com|  Abernathy Group|Implemented secon...|e-enable ext
ensib...|
|  9|  Clementina DuBuque|  Moriah.Stanton|Rey.Padberg@karin...|  Kattie Turnpike|Suite 198|   Lebsackbury|3142
8-2261|-38.2386|  57.2232|        024-648-3804|  ambrose.net|      Hoeger LLC|Centralized empow...|target end-
to-end...|
+---+--------------------+----------------+--------------------+-----------------+---------+--------------+----
------+--------+---------+--------------------+-------------+-----------------+--------------------+-----------
---------+
```

10

```
# Get the total number of rows in the DataFrame
total_rows = df.count()

# Display the total number of rows
print(f"Total number of rows: {total_rows}")
```

```
Total number of rows: 10
```

11

```
# Get the total number of rows in the cleaned DataFrame
total_rows = df_cleaned.count()

# Display the total number of rows
print(f"Total rows in cleaned DataFrame: {total_rows}")
```

```
Total rows in cleaned DataFrame: 9
```

""" The coalesce() function in PySpark is often used to reduce the number of partitions in a DataFrame. It can help optimize performance by limiting the number of partitions when performing actions like writing to disk. """

13

```
# Coalesce the DataFrame to a single partition
df_cleaned_coalesced = df_cleaned.coalesce(1)

# Show the DataFrame after coalescing
df_cleaned_coalesced.show()
```

```
4-1099| 24.8918|  21.8984|        210.067.6132|     elvis.io|      Johns Group|Configurable mult...|generate en
terpri...|
|  5|    Chelsey Dietrich|          Kamren|Lucio_Hettinger@a...|     Skiles Walks|Suite 351|    Roscoeview|
33263|-31.8129|  62.5342|       (254)954-1289| demarco.info|      Keebler LLC|User-centric faul...|revolutioniz
e end...|
|  6|    Patricia Lebsack|        Karianne|Julianne.OConner@...|      Hoeger Mall| Apt. 692|   South Elvis|5391
9-4257| 29.4572|-164.2990|   493-170-9623 x156|     kale.biz|    Robel-Corkery|Multi-tiered zero...|transition 
cuttin...|
|  7|Mrs. Dennis Schulist|Leopoldo_Corkery|Karley_Dach@jaspe...|Norberto Crossing| Apt. 950| South Christy|2350
5-1337|-71.4197|  71.7478|1-477-935-8478 x6430|      ola.org|Considine-Lockman|Synchronised bott...|e-enable in
novati...|
|  8|Nicholas Runolfsd...|   Maxime_Nienow|Sherwood@rosamond.me| Ellsworth Summit|Suite 729|     Aliyaview|
45169|-14.3990|-120.7677|   586.493.6943 x140| jacynthe.com|  Abernathy Group|Implemented secon...|e-enable ext
ensib...|
|  9|  Clementina DuBuque|  Moriah.Stanton|Rey.Padberg@karin...|  Kattie Turnpike|Suite 198|   Lebsackbury|3142
8-2261|-38.2386|  57.2232|        024-648-3804|  ambrose.net|       Hoeger LLC|Centralized empow...|target end-
to-end...|
+---+--------------------+----------------+--------------------+-----------------+---------+--------------+----
------+--------+---------+--------------------+-------------+-----------------+--------------------+-----------
---------+
```

14

```
# Number of partitions before coalescing
print("Number of partitions before coalescing:", df_cleaned.rdd.getNumPartitions())

# Number of partitions after coalescing
print("Number of partitions after coalescing:", df_cleaned_coalesced.rdd.getNumPartitions())
```

```
Number of partitions before coalescing: 1
Number of partitions after coalescing: 1
```

""" DataFrame (df_cleaned) was already in a single partition before applying coalesce(1) """

Save cleaned & coalesced DataFrame to processed-api

17

```
# Mount the processed-api container using secret scope for authentication
dbutils.fs.mount(
    source="wasbs://processed-api@casestudy1new.blob.core.windows.net",
    mount_point="/mnt/processed-api",
    extra_configs={"fs.azure.account.key.casestudy1new.blob.core.windows.net": dbutils.secrets.get(scope =
    "casestudy", key = "storage")}
)
```

```
True
```

18

```
# Defining path to save the DataFrame
output_path = "/mnt/processed-api/cleaned_users.parquet"

# Save the DataFrame to the processed-api container in Parquet format without overwriting existing data
df_cleaned_coalesced.write.mode("append").parquet(output_path)
```

For staging

20

```
dbutils.fs.mount(
    source="wasbs://staging-api@casestudy1new.blob.core.windows.net",
    mount_point="/mnt/staging-api",
    extra_configs={"fs.azure.account.key.casestudy1new.blob.core.windows.net": dbutils.secrets.get(scope =
    "casestudy", key = "storage")}
)
```

```
True
```

21

```
# Defining path to the 'staging-api' container
staging_container_path = "/mnt/staging-api/staging_users.parquet"

# Save the cleaned DataFrame to the 'staging-api' container using 'append' mode
df_cleaned_coalesced.write.mode('append').parquet(staging_container_path)
```