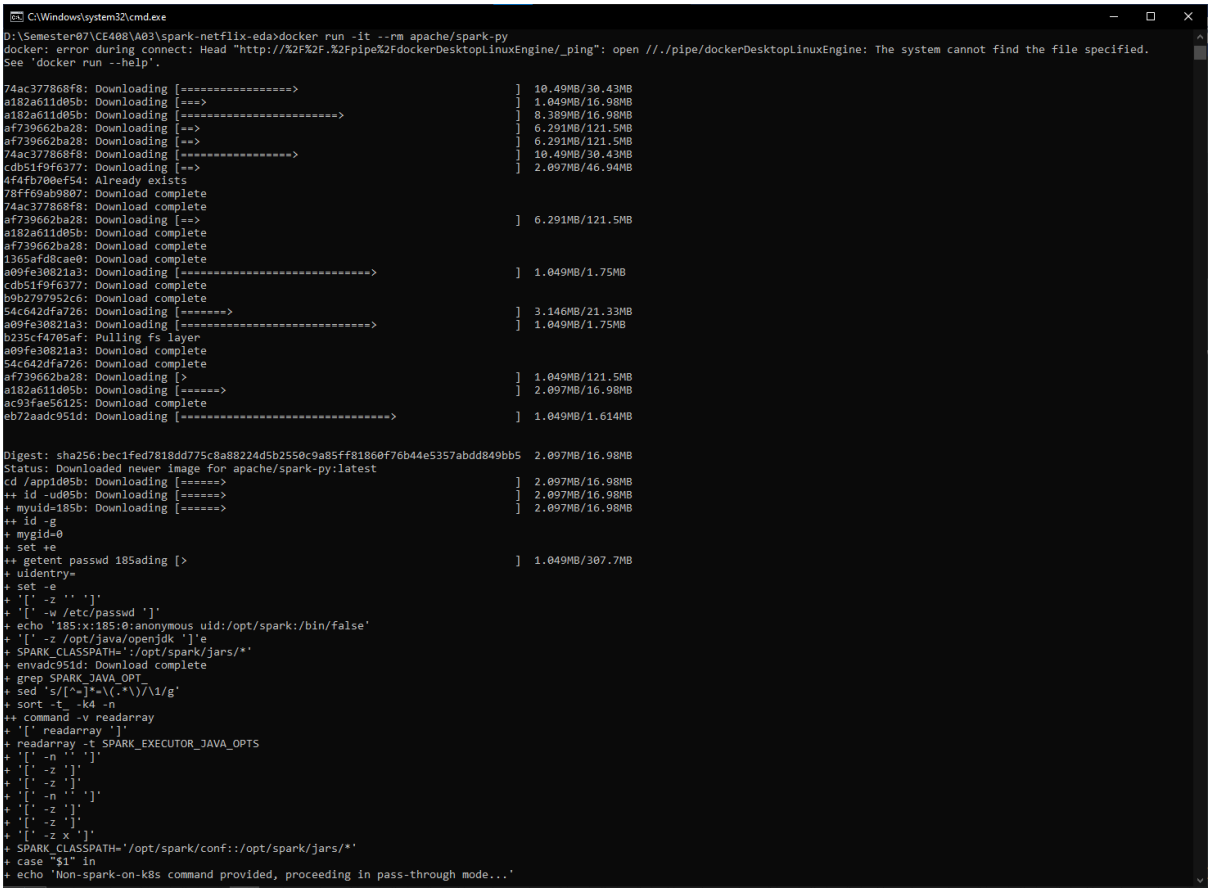
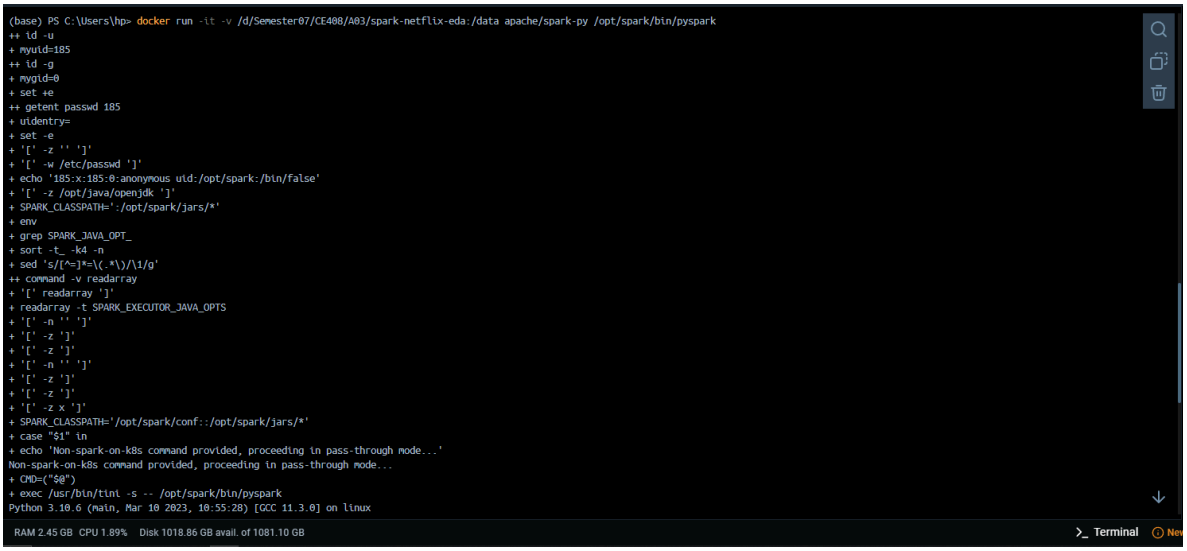
2021538
Romaisa
Assignment 03
CE408

Downloading the Spark Container:



Running the Script in Spark



Terminal

```
+ CMD=("$@")
+ exec /usr/bin/tini -s -- /opt/spark/bin/pyspark
Python 3.10.6 (main, Mar 10 2023, 10:55:28) [GCC 11.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/12/15 22:35:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Python version 3.10.6 (main, Mar 10 2023 10:55:28)
Spark context Web UI available at http://eb9f4e7be7f7:4040
Spark context available as 'sc' (master = local[*], app id = local-1734302133502).
SparkSession available as 'spark'.
>>> # Load the dataset
>>> df = spark.read.csv("/data/netflix_titles.csv", header=True, inferSchema=True)
>>>
>>> # Show the first 5 rows
>>> df.show(5)
+-------+-------+--------------------+---------------+--------------------+-------------+------------------+------------+------+---------+--------------------+--------------------+
|show_id|   type|               title|       director|                cast|      country|        date_added|release_year|rating| duration|           listed_in|         description|
+-------+-------+--------------------+---------------+--------------------+-------------+------------------+------------+------+---------+--------------------+--------------------+
|     s1|  Movie|Dick Johnson Is Dead|Kirsten Johnson|                null|United States|September 25, 2021|        2020| PG-13|   90 min|       Documentaries|As her father nea...|
|     s2|TV Show|       Blood & Water|           null|Ama Qamata, Khosi...| South Africa|September 24, 2021|        2021| TV-MA|2 Seasons|International TV ...|After crossing pa...|
|     s3|TV Show|           Ganglands|Julien Leclercq|Sami Bouajila, Tr...|         null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|To protect his fa...|
|     s4|TV Show|Jailbirds New Orl...|           null|                null|         null|September 24, 2021|        2021| TV-MA| 1 Season|Docuseries, Reali...|Feuds, flirtation...|
|     s5|TV Show|        Kota Factory|           null|Mayur More, Jiten...|        India|September 24, 2021|        2021| TV-MA|2 Seasons|International TV ...|In a city of coac...|
+-------+-------+--------------------+---------------+--------------------+-------------+------------------+------------+------+---------+--------------------+--------------------+
only showing top 5 rows

>>> 
```

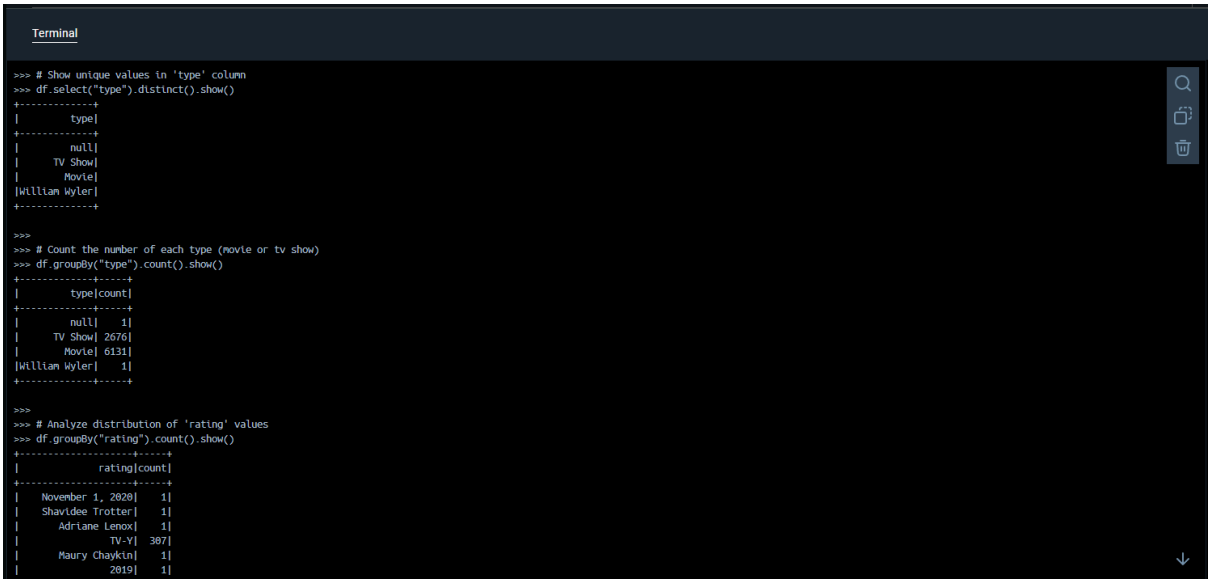
Terminal

```
>>> df.printSchema()
root
 |-- show_id: string (nullable = true)
 |-- type: string (nullable = true)
 |-- title: string (nullable = true)
 |-- director: string (nullable = true)
 |-- cast: string (nullable = true)
 |-- country: string (nullable = true)
 |-- date_added: string (nullable = true)
 |-- release_year: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- duration: string (nullable = true)
 |-- listed_in: string (nullable = true)
 |-- description: string (nullable = true)

>>>
>>> num_rows = df.count()
>>> num_cols = len(df.columns)
>>> print(f"Number of Rows: {num_rows}")
Number of Rows: 8809
>>> print(f"Number of Columns: {num_cols}")
Number of Columns: 12
>>> print(f"Columns: {df.columns}")
Columns: ['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description']
>>> df.describe().show()
24/12/15 22:38:30 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
+-------+------------------+-----------+------------------------------+------------------+------------------+---------------+--------------+----------------+----------------+------------+
--------------------+-------------------+
|summary|           show_id|       type|                         title|          director|              cast|        country|    date_added|    release_year|          rating|    duration|
          listed_in|        description|
+-------+------------------+-----------+------------------------------+------------------+------------------+---------------+--------------+----------------+----------------+------------+
--------------------+-------------------+
|  count|              8809|       8808|                          8807|              6173|              7983|           7977|          8796|            8807|            8803|        8804|
               8806|               8806|
|   mean|              null|       null|            1124.7692307692307|              null|              null|         1944.0|          null|2014.189598270172|          2016.8|      1994.0|
```


```
>>> num_rows = df.count()
>>> num_cols = len(df.columns)
>>> print(f"Number of Rows: {num_rows}")
Number of Rows: 8809
>>> print(f"Number of Columns: {num_cols}")
Number of Columns: 12
>>> print(f"Columns: {df.columns}")
Columns: ['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description']
>>> df.describe().show()
24/12/15 22:38:30 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
+-------+------------------+-----------+------------------------------+------------------+------------------+---------------+--------------+----------------+----------------+------------+
--------------------+-------------------+
|summary|           show_id|       type|                         title|          director|              cast|        country|    date_added|    release_year|          rating|    duration|
          listed_in|        description|
+-------+------------------+-----------+------------------------------+------------------+------------------+---------------+--------------+----------------+----------------+------------+
--------------------+-------------------+
|  count|              8809|       8808|                          8807|              6173|              7983|           7977|          8796|            8807|            8803|        8804|
               8806|               8806|
```


```
|   mean|              null|       null|            1124.7692307692307|              null|              null|         1944.0|          null|2014.189598270172|          2016.8|      1994.0|
               null|  2014.6666666666667|
| stddev|              null|       null|            1042.1800991068478|              null|              null|           null|          null|8.786618147271653|6.260990336999442|        null|
               null|   4.509249752822894|
|    min| and probably will."|      Movie|           "Behind ""The Cov...|"Sam ""Blitz"" Ba...|"Black Deniro, By...| Ama K. Abebrese| April 15, 2018|   Charles Rocket|    Adriane Lenox| Alan Cumming|
         Akin Lewis|      Alicia Sánchez|
|    max|              s999|William Wyler|최강전사 미니특공대 : 영웅의 탄생|      Şenol Sönmez|Ṣọpẹ́ Dìrísù, Wun...|       Zimbabwe|  United States|   United States|               UR|United States|
United States, Un...|“Last Chance U” h...|
+-------+------------------+-----------+------------------------------+------------------+------------------+---------------+--------------+----------------+----------------+------------+
--------------------+-------------------+
```


Terminal
>>> # Show unique values in 'type' column
>>> df.select("type").distinct().show()
+------------+
| type|
+------------+
| null|
| TV Show|
| Movie|
|William Wyler|
+------------+
>>>
>>> # Count the number of each type (movie or tv show)
>>> df.groupBy("type").count().show()
+------------+-----+
| type|count|
+------------+-----+
null	1
TV Show	2676
Movie	6131
William Wyler	1
+------------+-----+
>>>
>>> # Analyze distribution of 'rating' values
>>> df.groupBy("rating").count().show()
+-------------------+-----+
| rating|count|
+-------------------+-----+
November 1, 2020	1
Shavidee Trotter	1
Adriane Lenox	1
TV-Y	307
Maury Chaykin	1
2019	1