PYSPARK ASSIGNMENT

Group 13 - 20BDA47

ETL using Pyspark

Extract

- Read input from different datasource - json,csv,parquet
- from data structures - dataframe, array, others

Transform

- Data preprocessing/cleaning
- RDD - transformation - map,filter
- Dataframe fuction - groupby,filter and so on

Load

- Store as file
- store to DB

In [1]:
```
from random import random
import os
from pyspark.sql import SparkSession
import pandas as pd
```

Load the data : Reading data as pandas dataframe

In [2]:
```
datapd = pd.read_csv("D:/ADATA/netflix_titles.csv")
```

In [3]:
```
datapd
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

Creating spark dataframe

In [4]:
```
spark_session = SparkSession.builder.master("local").\
        appName("SparkApplication").\
        config("spark.driver.bindAddress","localhost").\
        config("spark.ui.port","4041").\
        getOrCreate()
sc = spark_session.sparkContext
```

In [5]:
```
data = spark_session.read.csv("D:/ADATA/netflix_titles.csv" ,sep=',',inferSchema=True, header=True ,nullValue='')
```

In [6]:
```
print(data)
```

```
DataFrame[show_id: string, type: string, title: string, director: string, cast: string, country: string, date_added: string, release_year: string, rating: string, duration: string,
listed_in: string, description: string]
```

In [7]:
```
print(data.show())
```

```
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+---------+--------------------+---------
-----------+
|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 fa
ther nea...|
|     s2|TV Show|       Blood & Water|                null|Ama Qamata, Khosi...|        South Africa|September 24, 2021|        2021| TV-MA|2 Seasons|International TV ...|After cro
ssing pa...|
|     s3|TV Show|           Ganglands|     Julien Leclercq|Sami Bouajila, Tr...|                null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|To protec
t his fa...|
|     s4|TV Show|Jailbirds New Orl...|                null|                null|                null|September 24, 2021|        2021| TV-MA| 1 Season|Docuseries, Reali...|Feuds, fl
irtation...|
|     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...|
|     s6|TV Show|       Midnight Mass|       Mike Flanagan|Kate Siegel, Zach...|                null|September 24, 2021|        2021| TV-MA| 1 Season|TV Dramas, TV Hor...|The arriv
al of a ...|
|     s7|  Movie|My Little Pony: A...|Robert Cullen, Jo...|Vanessa Hudgens, ...|                null|September 24, 2021|        2021|    PG|   91 min|Children & Family...|Equestri
a's divid...|
|     s8|  Movie|             Sankofa|        Haile Gerima|Kofi Ghanaba, Oya...|United States, Gh...|September 24, 2021|        1993| TV-MA|  125 min|Dramas, Independe...|On a phot
o shoot ...|
|     s9|TV Show|The Great British...|     Andy Devonshire|Mel Giedroyc, Sue...|      United Kingdom|September 24, 2021|        2021| TV-14|9 Seasons|British TV Shows,...|A talente
d batch ...|
|    s10|  Movie|         The Starling|      Theodore Melfi|Melissa McCarthy,...|       United States|September 24, 2021|        2021| PG-13|  104 min|    Comedies, Dramas|A woman a
djusting...|
|    s11|TV Show|Vendetta: Truth, ...|                null|                null|                null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, D...|"Sicily b
oasts a ...|
|    s12|TV Show|    Bangkok Breaking|   Kongkiat Komesiri|Sukollawat Kanaro...|                null|September 23, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|Strugglin
g to ear...|
|    s13|  Movie|        Je Suis Karl| Christian Schwochow|Luna Wedler, Jann...|Germany, Czech Re...|September 23, 2021|        2021| TV-MA|  127 min|Dramas, Internati...|After mos
t of her...|
|    s14|  Movie|Confessions of an...|       Bruno Garotti|Klara Castanho, L...|                null|September 22, 2021|        2021| TV-PG|   91 min|Children & Family...|When the 
clever b...|
|    s15|TV Show|Crime Stories: In...|                null|                null|                null|September 22, 2021|        2021| TV-MA| 1 Season|British TV Shows,...|Cameras f
ollowing...|
|    s16|TV Show|   Dear White People|                null|Logan Browning, B...|       United States|September 22, 2021|        2021| TV-MA|4 Seasons|TV Comedies, TV D...|"Students
 of colo...|
|    s17|  Movie|Europe's Most Dan...|Pedro de Echave G...|                null|                null|September 22, 2021|        2020| TV-MA|   67 min|Documentaries, In...|Declassif
ied docu...|
|    s18|TV Show|     Falsa identidad|                null|Luis Ernesto Fran...|              Mexico|September 22, 2021|        2020| TV-MA|2 Seasons|Crime TV Shows, S...|Strangers
 Diego a...|
|    s19|  Movie|           Intrusion|          Adam Salky|Freida Pinto, Log...|                null|September 22, 2021|        2021| TV-14|   94 min|           Thrillers|After a d
eadly ho...|
|    s20|TV Show|              Jaguar|                null|Blanca Suárez, Iv...|                null|September 22, 2021|        2021| TV-MA| 1 Season|International TV ...|In the 19
60s, a H...|
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+---------+--------------------+---------
-----------+
only showing top 20 rows

None
```

In [8]:
```
data.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)
```

In [9]:
```
print(data.columns)
```

```
['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description']
```

In [10]:
```
data.count()
```

Out[10]:
```
8809
```

In [11]:
```
header = data.first()
```

In [12]:
```
print(header)
```

```
Row(show_id='s1', type='Movie', title='Dick Johnson Is Dead', director='Kirsten Johnson', cast=None, country='United States', date_added='September 25, 2021', release_year='2020',
rating='PG-13', duration='90 min', listed_in='Documentaries', description='As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comi
cal ways to help them both face the inevitable.')
```

In [13]:
```
print(data.dtypes)
```

```
[('show_id', 'string'), ('type', 'string'), ('title', 'string'), ('director', 'string'), ('cast', 'string'), ('country', 'string'), ('date_added', 'string'), ('release_year', 'stri
ng'), ('rating', 'string'), ('duration', 'string'), ('listed_in', 'string'), ('description', 'string')]
```

In [14]:
```
print(data.head())
```

```
Row(show_id='s1', type='Movie', title='Dick Johnson Is Dead', director='Kirsten Johnson', cast=None, country='United States', date_added='September 25, 2021', release_year='2020',
rating='PG-13', duration='90 min', listed_in='Documentaries', description='As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comi
cal ways to help them both face the inevitable.')
```

In [15]:
```
data.describe().show()
```

```
+-------+--------------------+-------------+--------------------------------+-------------------+--------------------+----------------+--------------+-----------------+---------
--------+------------+--------------------+--------------------+
|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|6.2609903
36999442|        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|    Adria
ne 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...|
+-------+--------------------+-------------+--------------------------------+-------------------+--------------------+----------------+--------------+-----------------+---------
--------+------------+--------------------+--------------------+
```

In [16]:
```
print(data.distinct().count())
```

```
8809
```

In [17]:
```
from pyspark.sql import functions as F

data.filter(F.col("show_id").cast("int").isNull()).count()
```

Out[17]:
```
8809
```

In [18]:
```
data.describe().show()
```

```
+-------+--------------------+-------------+--------------------------------+-------------------+--------------------+----------------+--------------+-----------------+---------
--------+------------+--------------------+--------------------+
|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|6.2609903
36999442|        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|    Adria
ne 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...|
+-------+--------------------+-------------+--------------------------------+-------------------+--------------------+----------------+--------------+-----------------+---------
--------+------------+--------------------+--------------------+
```

In [19]:
```
print(data.columns)
```

```
['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description']
```

In [20]:
```
print(data.distinct().count())
```

```
8809
```

Identifying Null Values

In [21]:
```
from pyspark.sql.functions import col,isnan, when, count
data.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in data.columns]).show()
```

```
+-------+----+-----+--------+----+-------+----------+------------+------+--------+---------+-----------+
|show_id|type|title|director|cast|country|date_added|release_year|rating|duration|listed_in|description|
+-------+----+-----+--------+----+-------+----------+------------+------+--------+---------+-----------+
|      0|   1|    2|    2636| 826|    832|        13|           2|     6|       5|        3|          3|
+-------+----+-----+--------+----+-------+----------+------------+------+--------+---------+-----------+
```

Creating a table view "netflix" as spark SQL

In [22]:
```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
sqlContext.registerDataFrameAsTable(data, "netflix_table")
```

```
D:\Hadoop\spark-3.2.0-bin-hadoop2.7\python\pyspark\sql\context.py:77: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
  warnings.warn(
```

In [23]:
```
netflix_raw_data = sqlContext.sql("SELECT * FROM netflix_table ")
netflix_raw_data.show()
```

```
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+---------+--------------------+---------
-----------+
|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 fa
ther nea...|
|     s2|TV Show|       Blood & Water|                null|Ama Qamata, Khosi...|        South Africa|September 24, 2021|        2021| TV-MA|2 Seasons|International TV ...|After cro
ssing pa...|
|     s3|TV Show|           Ganglands|     Julien Leclercq|Sami Bouajila, Tr...|                null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|To protec
t his fa...|
|     s4|TV Show|Jailbirds New Orl...|                null|                null|                null|September 24, 2021|        2021| TV-MA| 1 Season|Docuseries, Reali...|Feuds, fl
irtation...|
|     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...|
|     s6|TV Show|       Midnight Mass|       Mike Flanagan|Kate Siegel, Zach...|                null|September 24, 2021|        2021| TV-MA| 1 Season|TV Dramas, TV Hor...|The arriv
al of a ...|
|     s7|  Movie|My Little Pony: A...|Robert Cullen, Jo...|Vanessa Hudgens, ...|                null|September 24, 2021|        2021|    PG|   91 min|Children & Family...|Equestri
a's divid...|
|     s8|  Movie|             Sankofa|        Haile Gerima|Kofi Ghanaba, Oya...|United States, Gh...|September 24, 2021|        1993| TV-MA|  125 min|Dramas, Independe...|On a phot
o shoot ...|
|     s9|TV Show|The Great British...|     Andy Devonshire|Mel Giedroyc, Sue...|      United Kingdom|September 24, 2021|        2021| TV-14|9 Seasons|British TV Shows,...|A talente
d batch ...|
|    s10|  Movie|         The Starling|      Theodore Melfi|Melissa McCarthy,...|       United States|September 24, 2021|        2021| PG-13|  104 min|    Comedies, Dramas|A woman a
djusting...|
|    s11|TV Show|Vendetta: Truth, ...|                null|                null|                null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, D...|"Sicily b
oasts a ...|
|    s12|TV Show|    Bangkok Breaking|   Kongkiat Komesiri|Sukollawat Kanaro...|                null|September 23, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|Strugglin
g to ear...|
|    s13|  Movie|        Je Suis Karl| Christian Schwochow|Luna Wedler, Jann...|Germany, Czech Re...|September 23, 2021|        2021| TV-MA|  127 min|Dramas, Internati...|After mos
t of her...|
|    s14|  Movie|Confessions of an...|       Bruno Garotti|Klara Castanho, L...|                null|September 22, 2021|        2021| TV-PG|   91 min|Children & Family...|When the 
clever b...|
|    s15|TV Show|Crime Stories: In...|                null|                null|                null|September 22, 2021|        2021| TV-MA| 1 Season|British TV Shows,...|Cameras f
ollowing...|
|    s16|TV Show|   Dear White People|                null|Logan Browning, B...|       United States|September 22, 2021|        2021| TV-MA|4 Seasons|TV Comedies, TV D...|"Students
 of colo...|
|    s17|  Movie|Europe's Most Dan...|Pedro de Echave G...|                null|                null|September 22, 2021|        2020| TV-MA|   67 min|Documentaries, In...|Declassif
ied docu...|
|    s18|TV Show|     Falsa identidad|                null|Luis Ernesto Fran...|              Mexico|September 22, 2021|        2020| TV-MA|2 Seasons|Crime TV Shows, S...|Strangers
 Diego a...|
|    s19|  Movie|           Intrusion|          Adam Salky|Freida Pinto, Log...|                null|September 22, 2021|        2021| TV-14|   94 min|           Thrillers|After a d
eadly ho...|
|    s20|TV Show|              Jaguar|                null|Blanca Suárez, Iv...|                null|September 22, 2021|        2021| TV-MA| 1 Season|International TV ...|In the 19
60s, a H...|
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+---------+--------------------+---------
-----------+
only showing top 20 rows
```

Exploratory data analysis using spark df : Unique showId count

In [24]:
```
data.select('show_id').distinct().count()
```

Out[24]:
```
8809
```

In [25]:
```
from pyspark.sql.functions import countDistinct

gr = data.agg(countDistinct("show_id"))
gr.show()
```

```
+--------------+
|count(show_id)|
+--------------+
|          8809|
+--------------+
```

GroupBy type and count of showId

In [26]:
```
gr = data.groupBy("type").agg(countDistinct("show_id"))
gr.show()
```

```
+-------------+--------------+
|         type|count(show_id)|
+-------------+--------------+
|      TV Show|          2676|
|        Movie|          6131|
|William Wyler|             1|
|         null|             1|
+-------------+--------------+
```

GroupBy type,release_year and count of showId

In [27]:
```
gr = data.groupBy("type","release_year").agg(countDistinct("show_id"))
gr.show()
```

```
+-------+---------------+--------------+
|   type|   release_year|count(show_id)|
+-------+---------------+--------------+
|  Movie|  June 12, 2021|             1|
|  Movie|           1963|             1|
|TV Show|           1981|             1|
|  Movie|           1971|             5|
|TV Show|           1988|             2|
|TV Show|           1972|             1|
|TV Show|  Nse Ikpe-Etim|             1|
|  Movie|           1956|             2|
|  Movie| Charles Rocket|             1|
|  Movie|           1997|            33|
|  Movie|           2015|           397|
|  Movie|           1969|             2|
|  Movie|           2010|           153|
|  Movie|           1993|            24|
|  Movie|           1977|             6|
|TV Show|           2020|           436|
|TV Show|           1997|             4|
|  Movie|           2016|           657|
|  Movie|           1992|            20|
|TV Show|           1945|             1|
+-------+---------------+--------------+
only showing top 20 rows
```

GroupBy type,release_year and count of showId

In [28]:
```
from pyspark.sql.functions import when
df = data.withColumn("duration", when(data.duration == "90 min","90") \
      .when(data.duration == "2 Seasons","2") \
      .otherwise(data.duration))
df.show()
```

```
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+---------+--------------------+---------
-----------+
|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|       Documentaries|As her fa
ther nea...|
|     s2|TV Show|       Blood & Water|                null|Ama Qamata, Khosi...|        South Africa|September 24, 2021|        2021| TV-MA|        2|International TV ...|After cro
ssing pa...|
|     s3|TV Show|           Ganglands|     Julien Leclercq|Sami Bouajila, Tr...|                null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|To protec
t his fa...|
|     s4|TV Show|Jailbirds New Orl...|                null|                null|                null|September 24, 2021|        2021| TV-MA| 1 Season|Docuseries, Reali...|Feuds, fl
irtation...|
|     s5|TV Show|        Kota Factory|                null|Mayur More, Jiten...|               India|September 24, 2021|        2021| TV-MA|        2|International TV ...|In a city
 of coac...|
|     s6|TV Show|       Midnight Mass|       Mike Flanagan|Kate Siegel, Zach...|                null|September 24, 2021|        2021| TV-MA| 1 Season|TV Dramas, TV Hor...|The arriv
al of a ...|
|     s7|  Movie|My Little Pony: A...|Robert Cullen, Jo...|Vanessa Hudgens, ...|                null|September 24, 2021|        2021|    PG|   91 min|Children & Family...|Equestri
a's divid...|
|     s8|  Movie|             Sankofa|        Haile Gerima|Kofi Ghanaba, Oya...|United States, Gh...|September 24, 2021|        1993| TV-MA|  125 min|Dramas, Independe...|On a phot
o shoot ...|
|     s9|TV Show|The Great British...|     Andy Devonshire|Mel Giedroyc, Sue...|      United Kingdom|September 24, 2021|        2021| TV-14|9 Seasons|British TV Shows,...|A talente
d batch ...|
|    s10|  Movie|         The Starling|      Theodore Melfi|Melissa McCarthy,...|       United States|September 24, 2021|        2021| PG-13|  104 min|    Comedies, Dramas|A woman a
djusting...|
|    s11|TV Show|Vendetta: Truth, ...|                null|                null|                null|September 24, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, D...|"Sicily b
oasts a ...|
|    s12|TV Show|    Bangkok Breaking|   Kongkiat Komesiri|Sukollawat Kanaro...|                null|September 23, 2021|        2021| TV-MA| 1 Season|Crime TV Shows, I...|Strugglin
g to ear...|
|    s13|  Movie|        Je Suis Karl| Christian Schwochow|Luna Wedler, Jann...|Germany, Czech Re...|September 23, 2021|        2021| TV-MA|  127 min|Dramas, Internati...|After mos
t of her...|
|    s14|  Movie|Confessions of an...|       Bruno Garotti|Klara Castanho, L...|                null|September 22, 2021|        2021| TV-PG|   91 min|Children & Family...|When the 
clever b...|
|    s15|TV Show|Crime Stories: In...|                null|                null|                null|September 22, 2021|        2021| TV-MA| 1 Season|British TV Shows,...|Cameras f
ollowing...|
|    s16|TV Show|   Dear White People|                null|Logan Browning, B...|       United States|September 22, 2021|        2021| TV-MA|4 Seasons|TV Comedies, TV D...|"Students
 of colo...|
|    s17|  Movie|Europe's Most Dan...|Pedro de Echave G...|                null|                null|September 22, 2021|        2020| TV-MA|   67 min|Documentaries, In...|Declassif
ied docu...|
|    s18|TV Show|     Falsa identidad|                null|Luis Ernesto Fran...|              Mexico|September 22, 2021|        2020| TV-MA|        2|Crime TV Shows, S...|Strangers
 Diego a...|
|    s19|  Movie|           Intrusion|          Adam Salky|Freida Pinto, Log...|                null|September 22, 2021|        2021| TV-14|   94 min|           Thrillers|After a d
eadly ho...|
|    s20|TV Show|              Jaguar|                null|Blanca Suárez, Iv...|                null|September 22, 2021|        2021| TV-MA| 1 Season|International TV ...|In the 19
60s, a H...|
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+---------+--------------------+---------
-----------+
only showing top 20 rows
```

In [29]:
```
from pyspark.sql.functions import *

newDf = df.withColumn('duration', regexp_replace('duration', 'Seasons', ''))
newDf = newDf.withColumn('duration', regexp_replace('duration', 'Season', ''))
Df = newDf.withColumn('duration', regexp_replace('duration', 'min', ''))
```

In [30]:
```
Df.show()
```

```
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+--------+--------------------+----------
----------+
|show_id|   type|               title|            director|                cast|             country|        date_added|release_year|rating|duration|           listed_in|         d
escription|
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+--------+--------------------+----------
----------+
|     s1|  Movie|Dick Johnson Is Dead|     Kirsten Johnson|                null|       United States|September 25, 2021|        2020| PG-13|      90|       Documentaries|As her fat
her nea...|
|     s2|TV Show|       Blood & Water|                null|Ama Qamata, Khosi...|        South Africa|September 24, 2021|        2021| TV-MA|       2|International TV ...|After cros
sing pa...|
|     s3|TV Show|           Ganglands|     Julien Leclercq|Sami Bouajila, Tr...|                null|September 24, 2021|        2021| TV-MA|      1 |Crime TV Shows, I...|To protect
 his fa...|
|     s4|TV Show|Jailbirds New Orl...|                null|                null|                null|September 24, 2021|        2021| TV-MA|      1 |Docuseries, Reali...|Feuds, fli
rtation...|
|     s5|TV Show|        Kota Factory|                null|Mayur More, Jiten...|               India|September 24, 2021|        2021| TV-MA|       2|International TV ...|In a city 
of coac...|
|     s6|TV Show|       Midnight Mass|       Mike Flanagan|Kate Siegel, Zach...|                null|September 24, 2021|        2021| TV-MA|      1 |TV Dramas, TV Hor...|The arriva
l of a ...|
|     s7|  Movie|My Little Pony: A...|Robert Cullen, Jo...|Vanessa Hudgens, ...|                null|September 24, 2021|        2021|    PG|     91 |Children & Family...|Equestri
a's divid...|
|     s8|  Movie|             Sankofa|        Haile Gerima|Kofi Ghanaba, Oya...|United States, Gh...|September 24, 2021|        1993| TV-MA|    125 |Dramas, Independe...|On a photo
 shoot ...|
|     s9|TV Show|The Great British...|     Andy Devonshire|Mel Giedroyc, Sue...|      United Kingdom|September 24, 2021|        2021| TV-14|      9 |British TV Shows,...|A talented
 batch ...|
|    s10|  Movie|         The Starling|      Theodore Melfi|Melissa McCarthy,...|       United States|September 24, 2021|        2021| PG-13|    104 |    Comedies, Dramas|A woman ad
justing...|
|    s11|TV Show|Vendetta: Truth, ...|                null|                null|                null|September 24, 2021|        2021| TV-MA|      1 |Crime TV Shows, D...|"Sicily bo
asts a ...|
|    s12|TV Show|    Bangkok Breaking|   Kongkiat Komesiri|Sukollawat Kanaro...|                null|September 23, 2021|        2021| TV-MA|      1 |Crime TV Shows, I...|Struggling
 to ear...|
|    s13|  Movie|        Je Suis Karl| Christian Schwochow|Luna Wedler, Jann...|Germany, Czech Re...|September 23, 2021|        2021| TV-MA|    127 |Dramas, Internati...|After most
 of her...|
|    s14|  Movie|Confessions of an...|       Bruno Garotti|Klara Castanho, L...|                null|September 22, 2021|        2021| TV-PG|     91 |Children & Family...|When the c
lever b...|
|    s15|TV Show|Crime Stories: In...|                null|                null|                null|September 22, 2021|        2021| TV-MA|      1 |British TV Shows,...|Cameras fo
llowing...|
|    s16|TV Show|   Dear White People|                null|Logan Browning, B...|       United States|September 22, 2021|        2021| TV-MA|      4 |TV Comedies, TV D...|"Students 
of colo...|
|    s17|  Movie|Europe's Most Dan...|Pedro de Echave G...|                null|                null|September 22, 2021|        2020| TV-MA|     67 |Documentaries, In...|Declassifi
ed docu...|
|    s18|TV Show|     Falsa identidad|                null|Luis Ernesto Fran...|              Mexico|September 22, 2021|        2020| TV-MA|       2|Crime TV Shows, S...|Strangers 
Diego a...|
|    s19|  Movie|           Intrusion|          Adam Salky|Freida Pinto, Log...|                null|September 22, 2021|        2021| TV-14|     94 |           Thrillers|After a de
adly ho...|
|    s20|TV Show|              Jaguar|                null|Blanca Suárez, Iv...|                null|September 22, 2021|        2021| TV-MA|      1 |International TV ...|In the 196
0s, a H...|
+-------+-------+--------------------+--------------------+--------------------+--------------------+------------------+------------+------+--------+--------------------+----------
----------+
only showing top 20 rows
```

Groupby type and avg durations

In [31]:
```
from pyspark.sql.functions import col,sum,avg,max
Df.groupBy("type") \
    .agg(avg("duration").alias("avg durations") \
        ) \
    .show(truncate=False)
```

```
+-------------+------------------+
|type         |avg durations     |
+-------------+------------------+
|TV Show      |1.7654320987654322|
|Movie        |99.88907068062828 |
|William Wyler|null              |
|null         |null              |
+-------------+------------------+
```

Saving analysis report- save as tables - partitionby type

In [34]:
```
Df.write.option("header",True) \
        .partitionBy("type") \
        .mode("overwrite") \
        .csv("D:/ADATA/netflix")
```