***DataSet Review***

* The file name is netflix\_titles.csv
* There are 8000+ rows and 12 columns
* Columns names are

• show\_id

• type

• title

• director

• cast

• country

• date\_added

• release\_year

• rating

• duration

• listed\_in (genres)

• description

* There are many nulls values in this dataset

***Dataset Cleaning Process using PySpark***

* Correcting the dataType of the table

**df = df.withColumn('duration\_time' , col('duration\_time').cast('int')) df = df.withColumn('release\_year' , col('release\_year').cast('int'))**

* Splitting columns data into list to extract information

**df = df.withColumn('duration', split('duration', " "))df = df.withColumn('duration\_time' , col('duration')[0])**

**df = df.withColumn('duration\_unit' , col('duration')[1])**

* Correcting the date format

**df = df.withColumn('date\_added', regexp\_replace('date\_added', ",", " "))**

* Dropping Old columns which has replaced by new one’s

**df = df.drop(col('duration'))**

* Changing the sequece of columns

**df = df.select('show\_id', "type", "duration\_time", "duration\_unit", "title", "director",**  **"cast", "country", "date\_added", "release\_year", "rating","listed\_in", "description")**

* Handelings nulls and then dropping them

**df = df.fillna({'country': 'UNKNOWN', 'director': 'NOT KNOWN'})**

**df = df.withColumn('country', regexp\_replace('country', ",", ""))**

**df = df.dropna()**

**df = df.withColumnRenamed('listed\_in', 'genre')**

**df = df.withColumn(**

**'genre',**

**split(col('genre'), ","))**