# Práctica Final Asignatura Herramientas Big Data

## Entrega: 13/05/2024

La práctica final de la asignatura consiste en resolver una serie de ejercicios de pyspark. Se debe adjuntar el script en formato .ipynb y debe ser íntegramente ejecutable.

El entorno de ejecución puede ser Google Colab o el local de vuestra máquina.

Se recomienda comentar el código explicando que se ha hecho en cada apartado y porque se ha recurrido a dicha solución.

La práctica está formada por 10 apartados cuya puntuación es de un punto cada uno. La práctica forma el 60% de la nota total, siendo el restante 40% los ejercicios pedidos en clase.

Para copiar el dataset lo descargaremos utilizando nuestro notebook con comandos Linux:

! wget -O /tmp/netflix\_titles\_dirty\_01.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_01.csv.gz?raw=True'

! wget -O /tmp/netflix\_titles\_dirty\_02.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_02.csv.gz?raw=True'

! wget -O /tmp/netflix\_titles\_dirty\_03.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_03.csv.gz?raw=True'

! wget -O /tmp/netflix\_titles\_dirty\_04.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_04.csv.gz?raw=True'

! wget -O /tmp/netflix\_titles\_dirty\_05.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_05.csv.gz?raw=True'

! wget -O /tmp/netflix\_titles\_dirty\_06.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_06.csv.gz?raw=True'

! wget -O /tmp/netflix\_titles\_dirty\_07.csv.gz 'https://github.com/datacamp/data-cleaning-with-pyspark-live-training/blob/master/data/netflix\_titles\_dirty\_07.csv.gz?raw=True'

A continuación, comprobamos que se han descargado con:

!ls /tmp/netflix\_titles\*

Y descomprimimos con:

!gunzip -c /tmp/netflix\_titles\_dirty\_03.csv.gz | head -20

1. Leer todos los csv descomprimidos guardados en la ruta de vuestro tmp en una sola línea de código (pista, usar wildcards para leer más de un fichero a la vez)
2. Analiza las columnas y renómbralos con un nombre que tenga sentido para cada una
3. Limpia el dataframe para que no existan nulos, adicionalmente elimina todos los valores que no se correspondan con el resto de datos de la columna
4. Revisa el tipo de dato de cada columna y parsealo según corresponda (la columna duración debe ser numérico)
5. Calcula la duración media en función del país
6. Filtra las películas y series que contengan la palabra music en su descripción y que su duración sea mayor a 90 minutos, ¿cuál es el actor que más películas y series ha realizado bajo estas condiciones?
7. Para el actor que más producciones ha realizado calcula cuantas semanas han pasado desde su primera producción hasta su última.
8. Transforma la columna de géneros para que su contenido sea un array con los valores de cada género por registro
9. ¿Cuántas producciones se han realizado en un único país y cuantas tienen 2 o más países?
10. Escribe el dataframe final como un fichero parquet