Desafío - Ecosistema Hadoop

- Para realizar este desafío debes haber estudiado previamente todo el material disponibilizado correspondiente a la unidad.
- Una vez terminado el desafío, comprime la carpeta que contiene el desarrollo de los requerimientos solicitados y sube el `.zip` en el LMS.
- Desarrollo desafío:
    - El desafío se debe desarrollar de manera Individual.

Ejercicio 1 - Carga de archivos en HDFS

- Inicie una instancia de trabajo AWS EMR.
- Cree una carpeta en hdfs que se llama `movielens-20m`.
- Utilizando `s3-dist-cp`, copie los archivos del siguiente bucket `s3://bigdata-desafio/challenges/u3act1/` a su carpeta creada y liste su contenido.
- Saque un screenshot del listado con nombre `movielens_datos_hdfs`.
- Identifique el jar de streaming en su máquina. Saque un screenshot con el nombre `id_streaming_jar`.

Ejercicio 2 - Utilizando el archivo genome-scores.csv

- Desde su instancia de AWS EMR, implemente un trabajo de MapReduce en Hadoop Streaming donde reporte la media del índice de relevancia por cada uno de los tag id asociados.
- Escriba los scripts `mapper_1.py` y `reducer_1.py` con las instrucciones asociadas para cada etapa y súbelos a su instancia de AWS EMR:
    - No se olvide de hacer ejecutables los scripts.
- Guarde los resultados del trabajo en un archivo llamado `results_1.txt`.

Ejercicio 3 - Utilizando el archivo ratings.csv

- Desde su instancia de AWS EMR, implemente un trabajo de MapReduce en Hadoop Streaming donde reporte la cantidad de ratings **y** su promedio **a nivel de usuario**.

- Escriba los scripts `mapper_2.py` y `reducer_2.py` con las instrucciones asociadas para cada etapa y súbelos a su instancia de AWS EMR:
    - No se olvide de hacer ejecutables los scripts.

- Guarde los resultados del trabajo en un archivo llamado `results_2.txt`.

Ejercicio 4 - Utilizando el archivo ratings.csv

- Desde su instancia de AWS EMR, implemente un trabajo de MapReduce en Hadoop Streaming donde reporte el rating promedio por **película**.

- Escriba los scripts `mapper_3.py` y reducer_3.py con las instrucciones asociadas para cada etapa y súbelos a su instancia de AWS EMR:
    - No se olvide de hacer ejecutables los scripts.

- Guarde los resultados del trabajo en un archivo llamado `results_3.txt`.

Ejercicio 5 - Utilizando el archivo movies.csv

- Desde su instancia de AWS EMR, implemente un trabajo de MapReduce en Hadoop Streaming donde cuente la cantidad de géneros asociados a cada película y reporte la cantidad de películas con 2, 3, 4, 5, 6, 7, 8, 9 y 10 géneros.

- Escriba los scripts `mapper_4.py` y `reducer_4.py` con las instrucciones asociadas para cada etapa y súbelos a su instancia de AWS EMR:
    - No se olvide de hacer ejecutables los scripts.

- Guarde los resultados del trabajo en un archivo llamado `results_4.txt`.

Ejercicio 6 - Descarga de resultados y eliminación de Instancia

- Saque un screenshot del término de la instancia AWS EMR.

- En un `.zip`, adjunte los siguientes archivos:
  - Los Mappers y Reducers de los ejercicios 2, 3, 4 y 5.
  - Los resultados en texto plano de los ejercicios 2, 3, 4 y 5.
  - El screenshot del listado de archivos en HDFS.
  - El screenshot de la identificación de la ruta del jar Hadoop Streaming.
  - El screenshot de la terminación de la instancia.

(Opcional) Ejercicio 7 - Utilizando el archivo movies.csv

- Desde su instancia de AWS EMR, implemente un trabajo de MapReduce en Hadoop Streaming donde cuente la cantidad de películas por año.
  - tip: Puede implementar la expresión regular `\(\d+\)` para identificar el año asociado en el título de la película.

- Escriba los scripts `mapper_5.py` y `reducer_5.py` con las instrucciones asociadas para cada etapa y súbelos a su instancia de AWS EMR:
  - No se olvide de hacer ejecutables los scripts.

- Guarde los resultados del trabajo en un archivo llamado `results_5.txt`.