Desafío - Procesamiento Distribuido y Paralelo

- 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/Grupal.

Ejercicio 1 - Preparación del ambiente de trabajo

- Levante una instancia de trabajo AWS EMR.
- Levante o vuelva a utilizar el bucket anteriormente implementado.
- Desde su instancia de trabajo, descarge los archivos `.txt` definidos en la siguiente tabla:

Autor	Libro	http	Nombre a guardar
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 1	http://www.gutenberg.org/cache/epub/9296/pg9296.txt	clarissa_1.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 2	http://www.gutenberg.org/cache/epub/9798/pg9798.txt	clarissa_2.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 3	http://www.gutenberg.org/cache/epub/9881/pg9881.txt	clarissa_3.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 4	http://www.gutenberg.org/cache/epub/10462/pg10462.txt	clarissa_4.txt

Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 5	http://www.gutenberg.org/cache/epub/10799/pg10799.txt	clarissa_5.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 6	http://www.gutenberg.org/cache/epub/11364/pg11364.txt	clarissa_6.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 7	http://www.gutenberg.org/cache/epub/11889/pg11889.txt	clarissa_7.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 8	http://www.gutenberg.org/cache/epub/12180/pg12180.txt	clarissa_8.txt
Samuel Richardson	Clarissa, or, the History of a Young Lady. Volume 9	http://www.gutenberg.org/cache/epub/12398/pg12398.txt	clarissa_9.txt

- Guarde cada archivo con el nombre sugerido en la tabla.
- Mueva los `txt` a una carpeta con el nombre `richardson_clarissa` dentro de su bucket S3. (Adjunte un screenshot de esta etapa)

Ejercicio 2 - Implementación de un WordCount con MapReduce primitivo

- Para cada uno de los archivos txt descargados, implemente una tarea WordCount con MapReduce primitivo en su instancia de trabajo AWS EMR.
- Preserve cada resultado en un archivo con la siguiente extensión `word_count_clarissa_1.txt`, `word_count_clarissa_2.txt`, etc.
- Mueva los `txt` a una carpeta con el nombre `richardson_clarissa_wordcount` dentro de su bucket S3. (Adjunte un screenshot de esta etapa)

Ejercicio 3 - Implementación de un WordCount para los nombres ocurrentes en los textos

- Implemente una tarea WordCount con MapReduce primitivo que cuente la cantidad de ocurrencias de los nombres `'Clarissa'`, `'Arabella'`, `'Robert'` y `'James'`.
- Preserve cada resultado en un archivo con la siguiente extensión `name_count_clarissa_1.txt`, `name_count_clarissa_2.txt`, etc.
- Mueva los `txt` a una carpeta con el nombre `richardson_clarissa_namecount` dentro de su bucket S3.

Ejercicio 4 - Survey Data

- Leer los datos desde el bucket de Desafío Latam `s3://bigdata-desafio/challenges/u2act2/`.

- A continuación se presentan los datos de una encuesta de satisfacción de usuarios con un servicio.

- Los primeros 50 campos corresponden a una batería de preguntas sobre el servicio, y el último campo corresponde al plan suscrito por el usuario.

```
1,0,0,0,1,1,0,0,0,0,1,0,1,0,0,1,0,0,0,0,1,1,0,0,1,1,1,0,0,0,0,1,1,0,0,0,
1,1,0,0,1,1,1,0,0,0,0,1,1,0,E
1,0,1,1,1,1,1,1,0,1,0,1,1,1,1,0,0,1,1,1,0,0,0,1,1,1,0,0,1,0,0,1,1,1,1,0,
1,0,0,1,0,0,1,1,0,0,1,1,1,0,E
0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,1,1,1,1,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,1,
1,1,1,0,1,0,0,1,0,1,0,1,0,1,E
1,1,0,1,0,1,1,0,0,1,1,0,0,1,0,0,1,0,1,0,0,1,0,0,0,1,1,1,1,1,0,0,1,1,1,1,
1,0,1,0,1,0,1,1,0,1,1,0,1,0,A
1,1,0,1,0,0,1,1,0,1,0,1,1,1,0,0,1,1,0,0,0,0,1,0,1,0,1,1,1,0,1,0,1,0,1,0,
1,0,0,0,1,0,1,1,0,1,0,0,0,1,D
1,0,0,1,0,0,1,1,0,0,0,1,1,0,1,1,0,1,0,0,1,0,0,0,0,0,1,1,0,1,0,1,0,1,1,0,
1,0,0,0,0,0,0,1,1,0,1,1,0,0,B
0,1,1,1,1,0,1,1,0,0,1,1,1,0,1,0,1,1,1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,1,1,1,
0,0,1,0,1,0,1,0,0,0,1,0,0,1,E
0,0,1,0,0,0,1,0,1,0,1,0,1,1,1,0,0,0,0,1,0,0,0,1,0,1,0,0,0,1,1,1,1,0,1,1,
0,1,0,0,1,0,0,1,0,1,1,0,1,0,E
0,1,1,0,0,0,1,1,1,0,0,0,1,0,1,1,0,1,1,1,0,1,1,1,0,1,1,0,0,1,0,1,1,1,1,1,
1,1,1,1,1,0,1,1,1,1,1,1,1,1,E
1,1,1,0,0,1,0,1,0,1,1,1,0,1,1,0,0,0,1,0,1,1,1,1,0,1,0,0,1,0,1,1,1,1,1,1,
1,1,0,0,1,1,1,1,1,0,0,1,0,0,E
```

- **Primer MapReduce:** Cuente la cantidad de ocurrencias de cada plan suscrito. Guarde los archivos como `mapper_4_1.py` y `reducer_4_1.py`.

- **Segundo MapReduce:** Calcule el promedio de respuestas positivas (1) por plan suscrito. Guarde los archivos como `mapper_4_2.py` y `reducer_4_2.py`.

- **Tercer MapReduce:** Calcule el promedio de respuestas positivas (1) para los usuarios con por lo menos un 60% de respuestas positivas, por grupo. Guarde los archivos como `mapper_4_3.py` y `reducer_4_3.py`.

Ejercicio 5

- (Desde su instancia de trabajo) Suba los resultados de sus `.txt` y sus archivos `.py` al bucket S3 creado en el punto 1. Haga público su bucket y envíe la ruta de este en el `zip`.

- Termine la instancia de trabajo. (Adjunte screenshot de esta etapa)

- En un `.zip`, adjunte los siguientes archivos:
    - Los scripts del ejercicio 4.
    - Los screenshots de la actividad 1, 2 y 3.
    - El screenshot de la terminación de la instancia.