BIG DATA APLICADO - Examen 1ª Evaluación

Instrucciones generales

1. Todas las sentencias deben ejecutarse desde la línea de comandos en las celdas que hay después del enunciado. No debes realizar ninguna tarea desde fuera de Jupyter.
2. Puedes **añadir** todas las celdas que necesites siempre y cuando estén antes del siguiente enunciado.
3. Todas las celdas **deben estar ejecutadas** y debe visualizarse el resultado de salida.
4. **No es necesario documentar** las respuestas, simplemente debes hacer lo que se pide en el enunciado.
5. Si un comando falla, explica la causa del error y cómo lo has solucionado.
6. Debes entregar tanto el **notebook** (fichero `.ipynb`) como el mismo fichero convertido a **PDF** (es muy probable que si intentas convertirlo en el propio contenedor te falle por no tener instalado `pandoc`, si es así descargalo en formato `.md` o `html` y conviértelo en tu máquina física).

NOMBRE:

Ejercicio 2: Uso de HDFS (5.5 puntos de RA1)

Gestión básica y estructura (1.5 puntos)

Preparación del entorno

- Crea un archivo local en tu máquina llamado `datos_alumno.txt`.
- El contenido del archivo debe ser tu nombre completo y tu DNI, repetido en 10 líneas.

```
!touch datos_alumno.txt
```

Creación de directorios en HDFS

- Crea la siguiente estructura de directorios dentro de HDFS:
  - `/examen/{tus_iniciales}/entradas`
  - `/examen/{tus_iniciales}/salidas`
  - `/examen/{tus_iniciales}/logs`

```
!hdfs dfs -mkdir -p /examen/abs/entradas
!hdfs dfs -mkdir  -p /examen/abs/salidas
!hdfs dfs -mkdir - p /examen/abs/logs
```

```
!hdfs dfs -ls /examen/abs/
```

```
Found 3 items
drwxr-xr-x   - root supergroup          0 2025-12-04 08:55 /examen/abs/entradas
drwxr-xr-x   - root supergroup          0 2025-12-04 08:55 /examen/abs/logs
drwxr-xr-x   - root supergroup          0 2025-12-04 08:55 /examen/abs/salidas
```

Ingesta de datos

- Sube el archivo local `datos_alumno.txt` al directorio HDFS `/examen/{tus_iniciales}/entradas`
- Verifica que el archivo se ha subido correctamente listando el contenido del directorio
- Verifica que el archivo se ha subido correctamente listando el contenido del archivo `datos_alumno.txt`

```
!ls /media/notebooks
```

```
PR0401.ipynb            mapper_ejercicio3.py
PR0402.ipynb            mapper_ejercicio4.py
Prueba2.ipynb           mapper_ejercicio_practicar.py
Untitled.ipynb          mapper_ejercicio_practicar2.py
Untitled1.ipynb         mapper_indice.py
Untitled2.ipynb         mapperdia_antes.py
Untitled3.ipynb         mapperdia_antes1.py
Untitled4.ipynb         practicar_dia_antes_examen.ipynb
city_temperature.csv        practicar_examen_ampreduce.ipynb
clean_file.csv          quijote.txt
dataset_31_credit-g.csv     quijote.txt.1
datos_alumno.txt        quijote.txt.2
doc1.txt            reducerQuijote.py
doc2.txt            reducerQuijote2.py
doc3.txt            reducer_ejercicio.py
enunciado_examen_ev1.ipynb  reducer_ejercicio2.py
mapper.py           reducer_ejercicio3.py
mapperQuijote.py        reducer_ejercicio4.py
mapperQuijote2.py       reducer_ejercicio_practicar.py
mapperQuijote3.py       reducer_ejercicio_practicar2.py
mapperQuijote3_2.py     reducer_indice.py
mapper_ejercicio.py     reducerdia_antes.py
mapper_ejercicio2.py        reducerdia_antes1.py
```

```
!hdfs dfs -put /media/notebooks/datos_alumno.txt /examen/abs/entradas
```

```
!hdfs dfs -ls /examen/abs/entradas
```

```
Found 1 items
-rw-r--r--   3 root supergroup        339 2025-12-04 09:00
/examen/abs/entradas/datos_alumno.txt
```

```
!hdfs dfs -cat /examen/abs/entradas/datos_alumno.txt
```

```
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
```

Manipulación y exploración (1.5 puntos)

Duplicación y renombrado

- Realiza una copia del archivo que acabas de subir (`datos_alumno.txt`) dentro de HDFS y colócala en la carpeta `/examen/{tus_iniciales}/salidas`.
- Renombra esta copia en HDFS para que se llame `backup_datos.txt`.

```
!hdfs dfs -cp /examen/abs/entradas/datos_alumno.txt /examen/abs/salidas
```

```
!hdfs dfs -ls /examen/abs/salidas
```

```
Found 1 items
-rw-r--r--   3 root supergroup        339 2025-12-04 09:02
/examen/abs/salidas/datos_alumno.txt
```

```
!hdfs dfs -mv /examen/abs/salidas/datos_alumno.txt
/examen/abs/salidas/backup_datos.txt
```

```
!hdfs dfs -ls /examen/abs/salidas/
```

```
Found 1 items
-rw-r--r--   3 root supergroup        339 2025-12-04 09:02
/examen/abs/salidas/backup_datos.txt
```

```
!hdfs dfs -cat /examen/abs/salidas/backup_datos.txt
```

```
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
Ángel Barrientos Simó 20533741P
```

Inspección de contenido

- Muestra por consola las últimas 3 líneas del archivo `backup_datos.txt` que reside en HDFS
- Muestra el tamaño total (en formato legible para los humanos) del directorio `/examen/{tus_iniciales}`

```
!hdfs dfs -ls /examen/abs |wc -c
print("Bytes")
```

```
249
Bytes
```

Movimiento de datos

- Mueve el archivo original `/examen/{tus_iniciales}/entradas/datos_alumno.txt` a la carpeta `/examen/{tus_iniciales}/logs`.

```
!hdfs dfs -mv /examen/abs/entradas/datos_alumno.txt /examen/abs/logs
```

```
!hdfs dfs -ls /examen/abs/logs
```

```
Found 1 items
-rw-r--r--   3 root supergroup        339 2025-12-04 09:00
/examen/abs/logs/datos_alumno.txt
```

Administración avanzada (2.5 puntos)

Factor de replicación

- Cambia el factor de replicación del archivo `/examen/{tus_iniciales}/salidas/backup_datos.txt` a **1**.
- Comprueba que el cambio se ha efectuado correctamente utilizando el comando `fsck` o `ls` con los parámetros adecuados.

```

```

Permisos

- Cambia los permisos del directorio `/examen/{tus_iniciales}/logs` para que solo el propietario tenga permisos de lectura, escritura y ejecución. El resto de los usuarios no debe tener acceso.

```
!hdfs dfs -chmod 700 /examen/abs/logs
```

```
!hdfs dfs -ls /examen/abs
```

```
Found 3 items
drwxr-xr-x   - root supergroup          0 2025-12-04 09:05 /examen/abs/entradas
drwx------   - root supergroup          0 2025-12-04 09:05 /examen/abs/logs
drwxr-xr-x   - root supergroup          0 2025-12-04 09:02 /examen/abs/salidas
```

Gestión de cuotas

- Asigna una cuota de espacio al directorio `/examen/{tus_iniciales}/entradas` limitada a 1 MB.
- Intenta subir un archivo (o varios) que superen en total 1 MB a ese directorio para demostrar que la cuota funciona.
- Elimina la cuota de espacio asignada al directorio `/examen/{tus_iniciales}/entradas`.

```

```

```
!hdfs dfsadmin -setSpaceQuota 1M /examen/abs/entradas
```

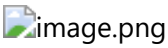
```
!hdfs dfs -count -q /examen/abs/entradas
```

```
        none             inf         1048576         1048576            1
0                  0 /examen/abs/entradas
```

Creamos un archivo grande para ver s funciona la cuota de espacio

```
!dd if=/dev/zero of=/tmp/archivo10mb.dat bs=1M count=10
!hdfs dfs -put /tmp/archivo10mb.dat /examen/abs/entradas
```

```
10+0 records in
10+0 records out
10485760 bytes (10 MB, 10 MiB) copied, 0.00340459 s, 3.1 GB/s
put: The DiskSpace quota of /examen/abs/entradas is exceeded: quota = 1048576 B =
1 MB but diskspace consumed = 402653184 B = 384 MB
```



Snapshots y recuperación

- Habilita la funcionalidad de snapshots en el directorio `/examen/{tus_iniciales}/salidas`.
- Crea un snapshot del directorio `/examen/{tus_iniciales}/salidas` llamado `snap_seguridad_v1`.
- Simula un error humano borrando el archivo `/examen/{tus_iniciales}/salidas/backup_datos.txt`.
- Recupera el archivo borrado restaurándolo desde el snapshot creado anteriormente.
- Comprueba que el archivo vuelve a aparecer en su ubicación original.

```
!hdfs dfsadmin -help |grep snap*
```

```
    [-allowSnapshot <snapshotDir>]
    [-disallowSnapshot <snapshotDir>]
    [-provisionSnapshotTrash <snapshotDir> [-all]]
    measures raw space used by replication, checksums, snapshots
-allowSnapshot <snapshotDir>:
    Allow snapshots to be taken on a directory.
-disallowSnapshot <snapshotDir>:
```

```
    Do not allow snapshots to be taken on a directory any more.
-provisionSnapshotTrash <snapshotDir> [-all]:
    Provision trash root in one or all snapshottable directories.   Trash
permission is rwxrwxrwt.
```

```
!hdfs dfs -help |grep snap*
```

```
    [-createSnapshot <snapshotDir> [<snapshotName>]]
    [-deleteSnapshot <snapshotDir> <snapshotName>]
    [-renameSnapshot <snapshotDir> <oldName> <newName>]
  The -x option excludes snapshots from being calculated.
  shows the erasure coding policy.The -s option shows snapshot counts.
-createSnapshot <snapshotDir> [<snapshotName>] :
  Create a snapshot on a directory
-deleteSnapshot <snapshotDir> <snapshotName> :
  Delete a snapshot from a directory
  -x  Excludes snapshots from being counted.
-renameSnapshot <snapshotDir> <oldName> <newName> :
  Rename a snapshot from oldName to newName
```

```
!hdfs dfsadmin -allowsnapshot /examen/abs/salidas
```

```
Allowing snapshot on /examen/abs/salidas succeeded
```

```
print("creo otro snapshot")
```

```
creo otro snapshot
```

```
!hdfs dfs -createSnapshot /examen/abs/salidas snap_seguridad_v2
```

```
Created snapshot /examen/abs/salidas/.snapshot/snap_seguridad_v2
```

```
!hdfs dfs -rm /examen/abs/salidas/backup_datos.txt
```

```
Deleted /examen/abs/salidas/backup_datos.txt
```

Recurperamos el documento con el snapshot mediante el cp

```
!hdfs dfs -ls /examen/abs/salidas/.snapshot
```

```
Found 2 items
drwxr-xr-x   - root supergroup          0 2025-12-04 09:16
/examen/abs/salidas/.snapshot/snap_seguridad_v1
drwxr-xr-x   - root supergroup          0 2025-12-04 09:26
/examen/abs/salidas/.snapshot/snap_seguridad_v2
```

```
!hdfs dfs -cp /examen/abs/salidas/.snapshot/snap_seguridad_v2 /examen/abs/salidas/
```

```
!hdfs dfs -ls /examen/abs/salidas/snap_seguridad_v2
```

```
Found 1 items
-rw-r--r--   3 root supergroup        339 2025-12-04 09:27
/examen/abs/salidas/snap_seguridad_v2/backup_datos.txt
```

Ejercicio 3: Computación distribuida con MapReduce (10 puntos de RA2)

Esta parte del examen la vamos a hacer con el *dataset* que puedes encontrar en https://www.kaggle.com/datasets/ashpalsingh1525/imdb-movies-dataset y que contiene datos sobre más de 10000 películas de IMDB. El fichero del *dataset* te lo habrá facilitado el profesor junto con el examen.

Número de películas por género

Número de películas de cada género

Queremos saber **cuántas películas hay en cada uno de los géneros**. Ten en cuenta que muchas películas pertenecen a más de un género. Consejo: antes de empezar observa y familiarízate con la estructura de los datos del fichero.

```
!hdfs dfs -put /media/notebooks/clean_file.csv /
```

```
%%writefile mapperExamen.py
#!/usr/bin/env python3

import sys


for line in sys.stdin:
    #line =     names   date_x  score   genre   overview    crew    orig_title
status  orig_lang   budget_x    revenue country
    line = line.strip()

names,date_x,score,genre,overview,crew,orig_title,status,orig_lang,budget_x,revenu
e,*country = line.split(',')
    for generos in genre.split(';'):
        if generos == "genre":
            continue
        else:
            print(f"{generos}\t1")
```

```
Overwriting mapperExamen.py
```

```
%%writefile reducerExamen.py
#!/usr/bin/env python3

import sys
genero_aux = None
contador = 0

for line in sys.stdin:
    line = line.strip()
    generos, valor = line.split('\t')
    #line = print(f"{generos}\t1")
    if generos is None:
        genero_aux = generos
    if genero_aux == generos:
        contador += 0
    else:
        print(f"{genero_aux}\t{contador}")
        genero_aux = generos
        contador = 1
if generos is not None:
    print(f"{genero_aux}\t{contador}")
```

```
Overwriting reducerExamen.py
```

```
!head clean_file.csv | python3 mapperExamen.py |sort| python3 reducerExamen.py
```

```
None    0
Action  1
Adventure   1
Animation   1
Comedy  1
Crime   1
Drama   1
Family  1
Fantasy 1
Science Fiction 1
Thriller    1
```

```
!hdfs dfs -rmdir /Examen_1
```

```
!hadoop jar \
/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
-file mapperExamen.py \
-file reducerExamen.py \
-mapper mapperExamen.py \
-reducer reducerExamen.py \
-input /clean_file.csv \
-output /Examen_1
```

```
2025-12-04 09:57:13,592 WARN streaming.StreamJob: -file option is deprecated,
please use generic option -files instead.
packageJobJar: [mapperExamen.py, reducerExamen.py, /tmp/hadoop-
unjar8804643804794167055/] [] /tmp/streamjob7877495906199875073.jar tmpDir=null
2025-12-04 09:57:14,106 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
to ResourceManager at yarnmanager/172.19.0.6:8032
2025-12-04 09:57:14,179 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
```

```
to ResourceManager at yarnmanager/172.19.0.6:8032
2025-12-04 09:57:14,347 INFO mapreduce.JobResourceUploader: Disabling Erasure
Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1764837993493_0002
2025-12-04 09:57:14,611 INFO mapred.FileInputFormat: Total input files to process
: 1
2025-12-04 09:57:14,675 INFO mapreduce.JobSubmitter: number of splits:2
2025-12-04 09:57:14,757 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1764837993493_0002
2025-12-04 09:57:14,757 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-12-04 09:57:14,880 INFO conf.Configuration: resource-types.xml not found
2025-12-04 09:57:14,880 INFO resource.ResourceUtils: Unable to find 'resource-
types.xml'.
2025-12-04 09:57:14,940 INFO impl.YarnClientImpl: Submitted application
application_1764837993493_0002
2025-12-04 09:57:14,966 INFO mapreduce.Job: The url to track the job:
http://yarnmanager:8088/proxy/application_1764837993493_0002/
2025-12-04 09:57:14,967 INFO mapreduce.Job: Running job: job_1764837993493_0002
2025-12-04 09:57:19,012 INFO mapreduce.Job: Job job_1764837993493_0002 running in
uber mode : false
2025-12-04 09:57:19,012 INFO mapreduce.Job:  map 0% reduce 0%
2025-12-04 09:57:22,063 INFO mapreduce.Job: Task Id :
attempt_1764837993493_0002_m_000000_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess
failed with code 1
    at
org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:129)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:350)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:178)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:195
3)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:172)

2025-12-04 09:57:22,077 INFO mapreduce.Job: Task Id :
attempt_1764837993493_0002_m_000001_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess
failed with code 1
    at
org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:129)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
```

```
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:350)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:178)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:195
3)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:172)

2025-12-04 09:57:24,090 INFO mapreduce.Job: Task Id :
attempt_1764837993493_0002_m_000001_1, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess
failed with code 1
    at
org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:129)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:350)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:178)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:195
3)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:172)

2025-12-04 09:57:24,091 INFO mapreduce.Job: Task Id :
attempt_1764837993493_0002_m_000000_1, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess
failed with code 1
    at
org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:129)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:350)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:178)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:195
3)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:172)

2025-12-04 09:57:27,113 INFO mapreduce.Job: Task Id :
attempt_1764837993493_0002_m_000001_2, Status : FAILED
```

```
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess
failed with code 1
    at
org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:129)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:350)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:178)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:195
3)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:172)

2025-12-04 09:57:27,115 INFO mapreduce.Job: Task Id :
attempt_1764837993493_0002_m_000000_2, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess
failed with code 1
    at
org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:539)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:129)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:350)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:178)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:195
3)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:172)

2025-12-04 09:57:31,141 INFO mapreduce.Job:  map 100% reduce 100%
2025-12-04 09:57:31,153 INFO mapreduce.Job: Job job_1764837993493_0002 failed with
state FAILED due to: Task failed task_1764837993493_0002_m_000001
Job failed as tasks failed. failedMaps:1 failedReduces:0 killedMaps:0
killedReduces: 0

2025-12-04 09:57:31,202 INFO mapreduce.Job: Counters: 14
    Job Counters
        Failed map tasks=7
        Killed map tasks=1
        Killed reduce tasks=1
        Launched map tasks=8
        Other local map tasks=6
```

```
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=11069
        Total time spent by all reduces in occupied slots (ms)=0
        Total time spent by all map tasks (ms)=11069
        Total vcore-milliseconds taken by all map tasks=11069
        Total megabyte-milliseconds taken by all map tasks=11334656
    Map-Reduce Framework
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
2025-12-04 09:57:31,202 ERROR streaming.StreamJob: Job not successful!
Streaming Command Failed!
```

```
generos = ["Animation","Adventure,","Family","Fantasy","Comedy"]
for genero in generos:
    print(genero)
```

```
Animation
Adventure,
Family
Fantasy
Comedy
```

```
!hdfs dfs -ls /
```

```
Found 23 items
drwxr-xr-x   - root supergroup          0 2025-11-26 16:43 /Quijote
drwxr-xr-x   - root supergroup          0 2025-11-26 18:18 /Quijote2
drwxr-xr-x   - root supergroup          0 2025-11-27 18:03 /Quijote3
drwxr-xr-x   - root supergroup          0 2025-11-27 18:06 /Quijote3_2
drwxr-xr-x   - root supergroup          0 2025-11-24 11:29 /Temperatura_30_ciudad
drwxr-xr-x   - root supergroup          0 2025-11-24 11:45 /Temperatura_MinMax
drwxr-xr-x   - root supergroup          0 2025-11-24 11:18 /Temperatura_maxima
drwxr-xr-x   - root supergroup          0 2025-11-24 11:19 /Temperatura_media_pais
drwxr-xr-x   - root supergroup          0 2025-12-03 17:20 /angel
drwxr-xr-x   - root supergroup          0 2025-12-03 17:52 /backup
-rw-r--r--   3 root supergroup    6622610 2025-12-04 09:31 /clean_file.csv
drwxr-xr-x   - root supergroup          0 2025-12-03 19:29 /dia_antes
drwxr-xr-x   - root supergroup          0 2025-12-04 08:55 /examen
drwxr-xr-x   - root supergroup          0 2025-11-18 09:05 /indice_invertido
drwxr-xr-x   - root supergroup          0 2025-12-03 17:04 /practica
drwxr-xr-x   - root supergroup          0 2025-11-26 15:30 /practica_401
drwxr-xr-x   - root supergroup          0 2025-12-01 11:45 /practicar_examen
```

```
drwxr-xr-x   - root supergroup          0 2025-12-03 17:51 /proyectos
drwxr-xr-x   - root supergroup          0 2025-12-03 17:51 /proyectoss
drwxr-xr-x   - root supergroup          0 2025-12-01 13:20 /prueba_practicar_1
drwxr-xr-x   - root supergroup          0 2025-11-18 09:22 /salida_indice
drwxrwx---   - root supergroup          0 2025-11-12 11:33 /tmp
drwxrwxrwt   - root root                0 2025-12-03 19:04 /yarn
```

Género más popular

Utilizando MapReduce, averigua cuál es el género más popular. Debes utilizar un segundo proceso MapReduce para procesar la salida del anterior.

```
%%writefile mapperExamen2.py
#!/usr/bin/env python3

import sys


for line in sys.stdin:
    #line =     names   date_x  score   genre   overview    crew    orig_title
status  orig_lang   budget_x    revenue country
    line = line.strip()

names,date_x,score,genre,overview,crew,orig_title,status,orig_lang,budget_x,revenu
e,*country = line.split(',')
    for generos in genre.split(';'):
        if generos == "genre":
            continue
        else:
            print(f"{generos}\t1")
```

```
Writing mapperExamen2.py
```

```
!head clean_file.csv | python3 mapperExamen2.py # |sort| python3 reducerExamen.py
```

```
Drama   1
Action  1
Science Fiction 1
Adventure   1
Action  1
Animation   1
Adventure   1
Family  1
Fantasy 1
```

```
Comedy  1
Animation   1
Comedy  1
Family  1
Adventure   1
Fantasy 1
Action  1
Thriller    1
Comedy  1
Crime   1
Action  1
Thriller    1
Crime   1
Animation   1
Family  1
Fantasy 1
Adventure   1
Comedy  1
Action  1
Science Fiction 1
```

```
%%writefile mapperExamen2_1.py
#!/usr/bin/env python3
import sys

import sys
genero_aux = None
contador = 0

for line in sys.stdin:
    line = line.strip()
    generos, valor = line.split('\t')
    #line = print(f"{generos}\t1")
    if generos is None:
        genero_aux = generos
    if genero_aux == generos:
        contador += 0
    else:
        print(f"{genero_aux}\t{contador}")
        genero_aux = generos
        contador = 1
if generos is not None:
    print(f"{genero_aux}\t{contador}")

for key, value in sorted(word_dict.items(), key=lambda x: x[1], reverse=True):
    print(f"{value}\t{key}")
```

País con películas más rentables

Queremos saber qué país tiene una filmografía más rentable (ten en cuenta que *budget*=presupuesto, *revenue*=ingresos), así que tienes que obtener un listado de países y beneficios promedio por película ((total ingresos - total presupuestos) / número películas de ese país)

```
%%writefile mapperExamen3.py
#!/usr/bin/env python3

import sys


for line in sys.stdin:
    #line =     names   date_x  score   genre   overview    crew    orig_title
status  orig_lang   budget_x    revenue country
    line = line.strip()

names,date_x,score,genre,overview,crew,orig_title,status,orig_lang,budget_x,revenu
e,*country = line.split(',')
    print(f"{country}\t{revenue}\t{budget_x}")
```

```
Overwriting mapperExamen3.py
```

```
!head clean_file.csv | python3 mapperExamen3.py
```

```
['country'] revenue budget_x
['AU']  271616668.0 75000000.0
['AU']  2316794914.0    460000000.0
['AU']  724459031.0 100000000.0
['AU']  34200000.0  12300000.0
['US']  340941958.6 77000000.0
['AU']  80000000.0  35000000.0
['AU']  351349364.0 100000000.0
['AU']  483480577.0 90000000.0
['US']  254946484.2 71000000.0
```