MOD 6 - Infraestructura Big Data

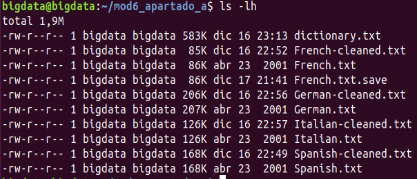
APARTADO A

**1. Descargar los diferentes diccionarios a una nueva carpeta**

**Se aplicó el siguiente código para crear una nueva carpeta 'mod6\_apartado\_a' y descargar los diccionarios en esa misma carpeta.**

****

**Además, los archivos descargados fueron tratados con el editor Nano para eliminar las líneas comentadas en el encabezado, y los cambios se guardaron en nuevos archivos, con el sufijo ‘-cleaned’.**



**2. Merge de ficheros**

**Para obtener un único fichero diccionario, se utilizó el comando CAT, de la siguiente manera:**

****

**3. Subir el fichero a HDFS**

**Para mover el fichero a HDFS en la ruta /user/bigdata/mapreduce, primero se creó la estructura de carpetas según se requiera, y luego se copió el archivo dictionary.txt a HDFS, usando los siguientes comandos:**

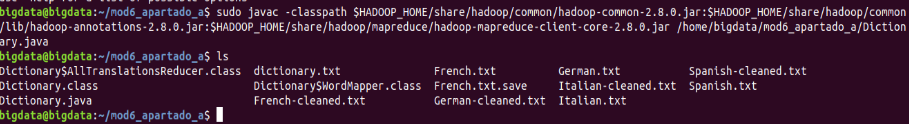
****



**4. Crear nuestro fichero Dictionary.java**

**Luego creé el archivo Dictionary.java con el código provisto para resolver el ejercicio, usando el editor Nano y luego se compiló el archivo java. A continuación un archivo jar**

****



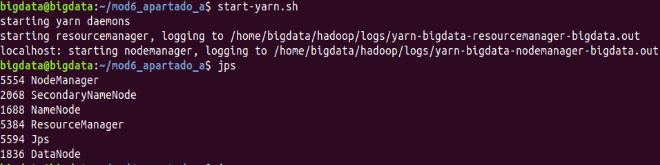
**Luego se creó un archivo jar:** *$ sudo jar cf Dictionary.jar Dictionary\*.class*

**Después fue ejecutado el proceso Map Reduce:** *$ hadoop jar Dictionary.jar Dictionary*

**Pero no pude conectarme al *ResourceManager* con el siguiente mensaje de error (extracto):**



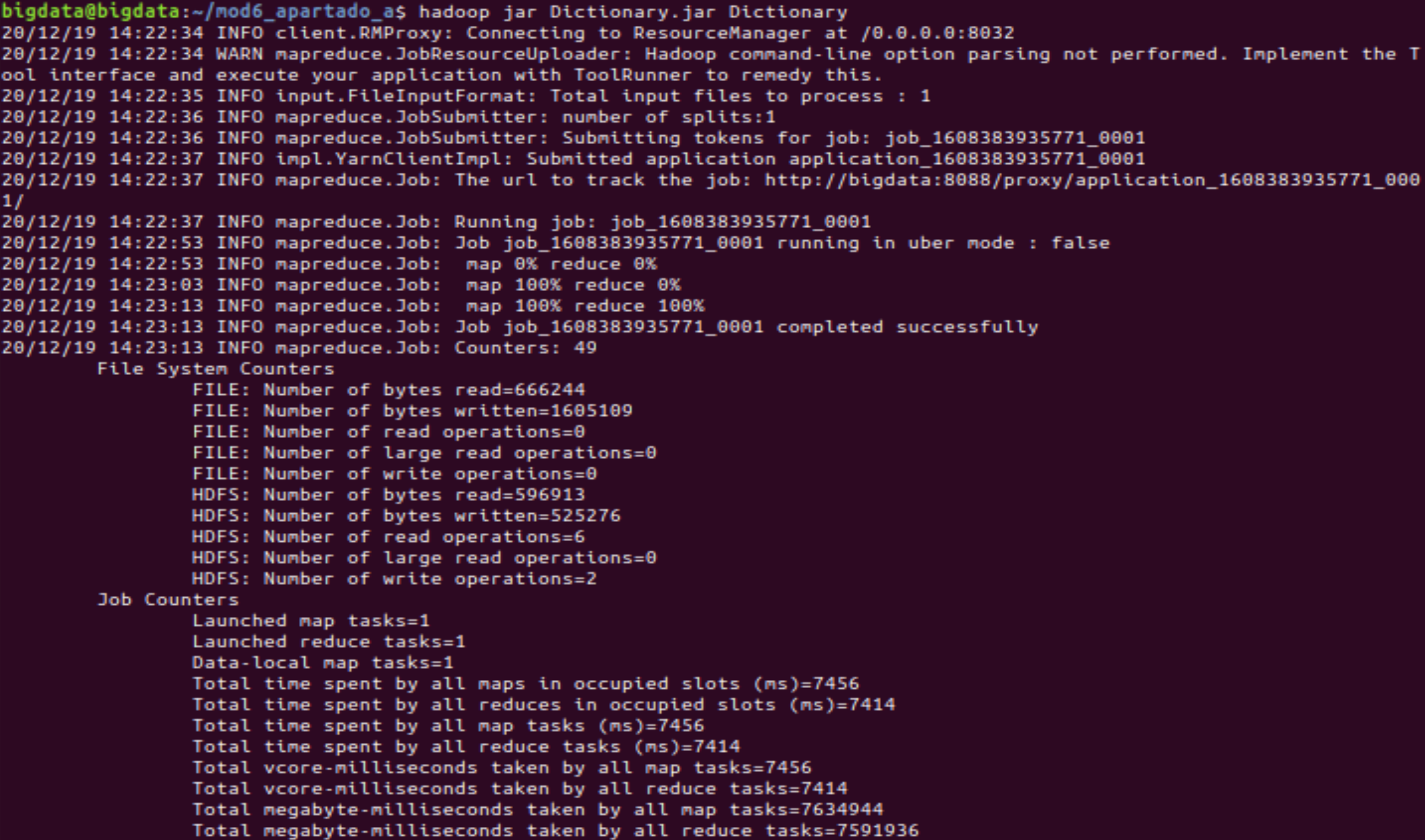
**Usando el comando j*ps* me di cuenta de que el *ResourceManager* no estaba activo. Entonces, tuve que iniciarlo, usando el comando:** *$ start-yarn.sh*

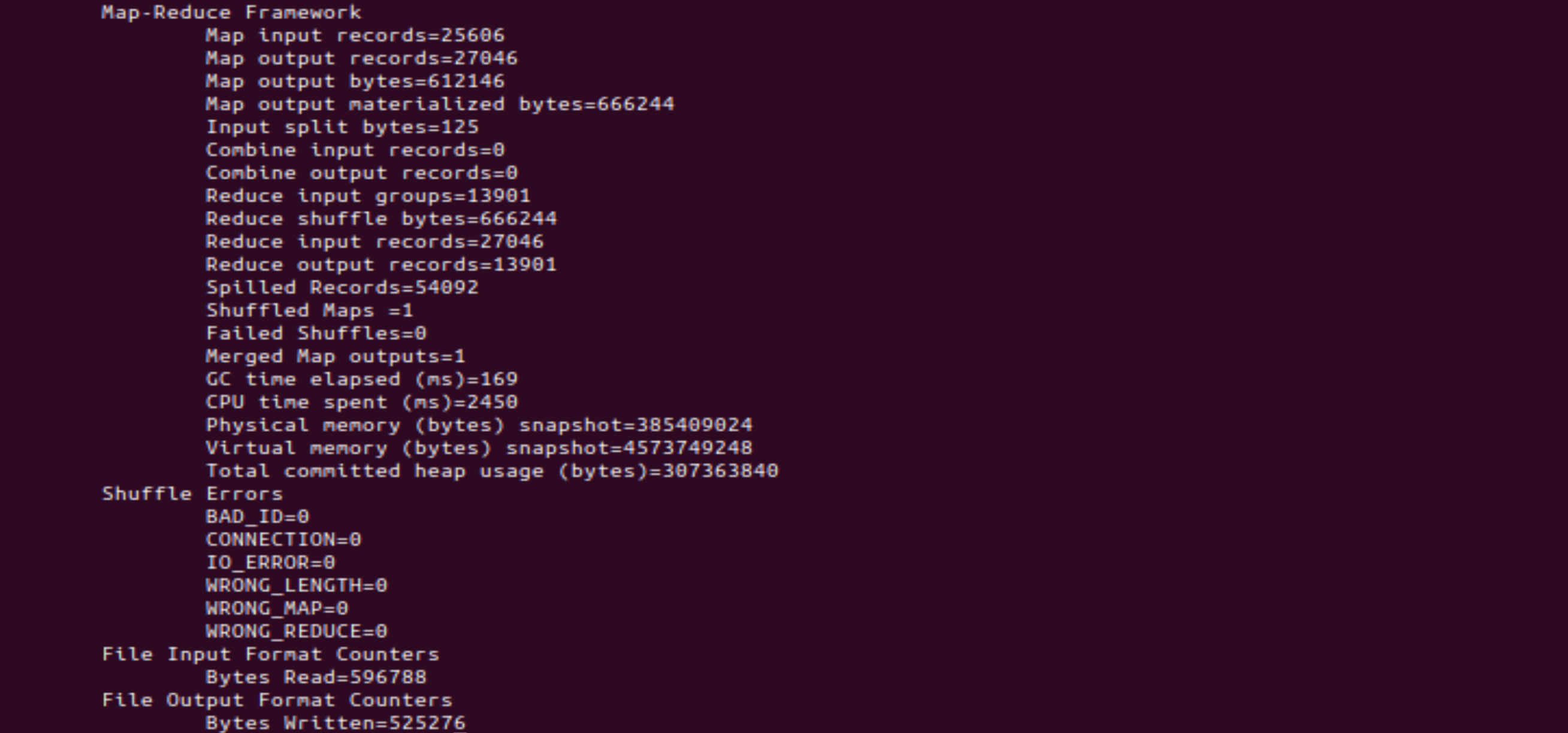


**Todavía tenía que superar otro error - *Container killed on request. exit code is 143* - relacionado con la falta de memoria disponible (memory overflow) para completar el proceso de Map Reduce. Para resolverlo, tuve que agregar las siguientes propiedades al archivo de configuración *yarn-site.xml y* reiniciar yarn:**

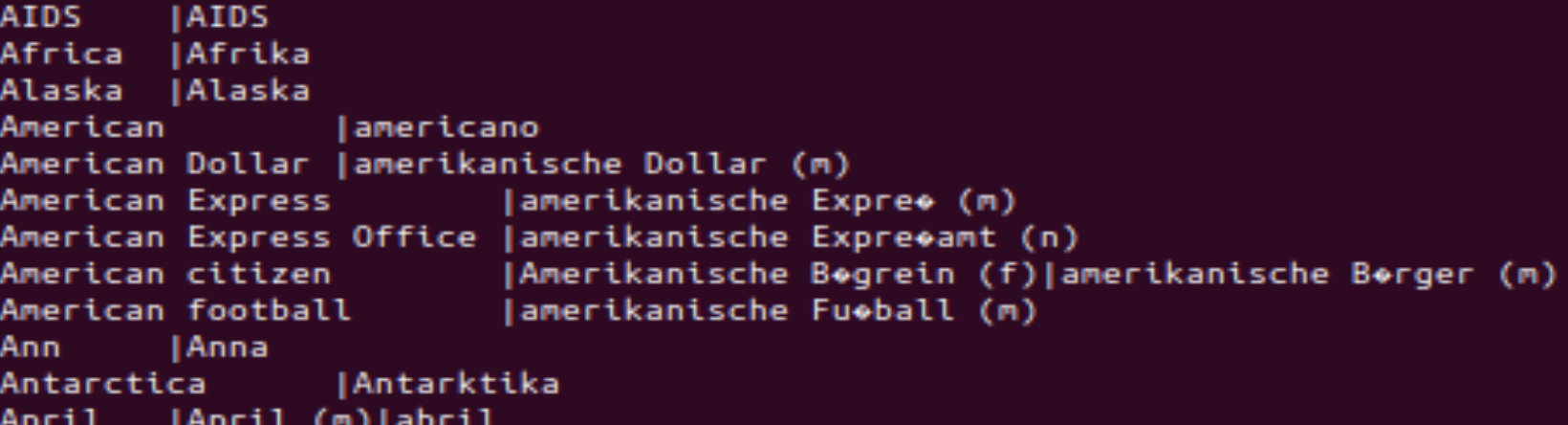
****

**Finalmente pudimos ejecutar el proceso Map Reduce sin más problemas:**





**Para confirmar los resultados obtenidos: *$hdfs dfs -cat /user/bigdata/output/part-r-00000***



**Para devolver la traducción de “pig”: *$ hdfs dfs -cat /user/bigdata/output/part-r-00000 | grep '\<pig\>'***

