Facultad de Ingeniería y Ciencias

DEPARTAMENTO DE INFORMÁTICA Y TELECOMUNICACIONES

TAREA 3: HADOOP

SISTEMAS DISTRIBUIDOS

Sección 2 - Profesor Nicolás Hidalgo

Margarita Osorio
Isaac Riveros

Índice

1. Problema y solución

Se solicita entregar el trayecto profesional de los artistas, de tal forma que se pueda observar todos los trabajos relacionados a un actor, director o guionista en una tabla al momento de buscarlo por su nombre. Estos datos se sacan de la base de datos de IMDb, y se pueden encontrar utilizando índice invertido.
Específicamente para este trabajo se utiliza un ambiente HADOOP, que nace para satisfacer la necesidad de procesar grandes volúmenes de datos, facilitando la creación de clústeres de hardware de consumo para analizar conjuntos de datos masivos en paralelo. Para ello utilizamos HDFS, así almacenamos el dataset masivo de IMDb.

2. Códigos

Para esta tarea, el código se divide en cuatro categorías:

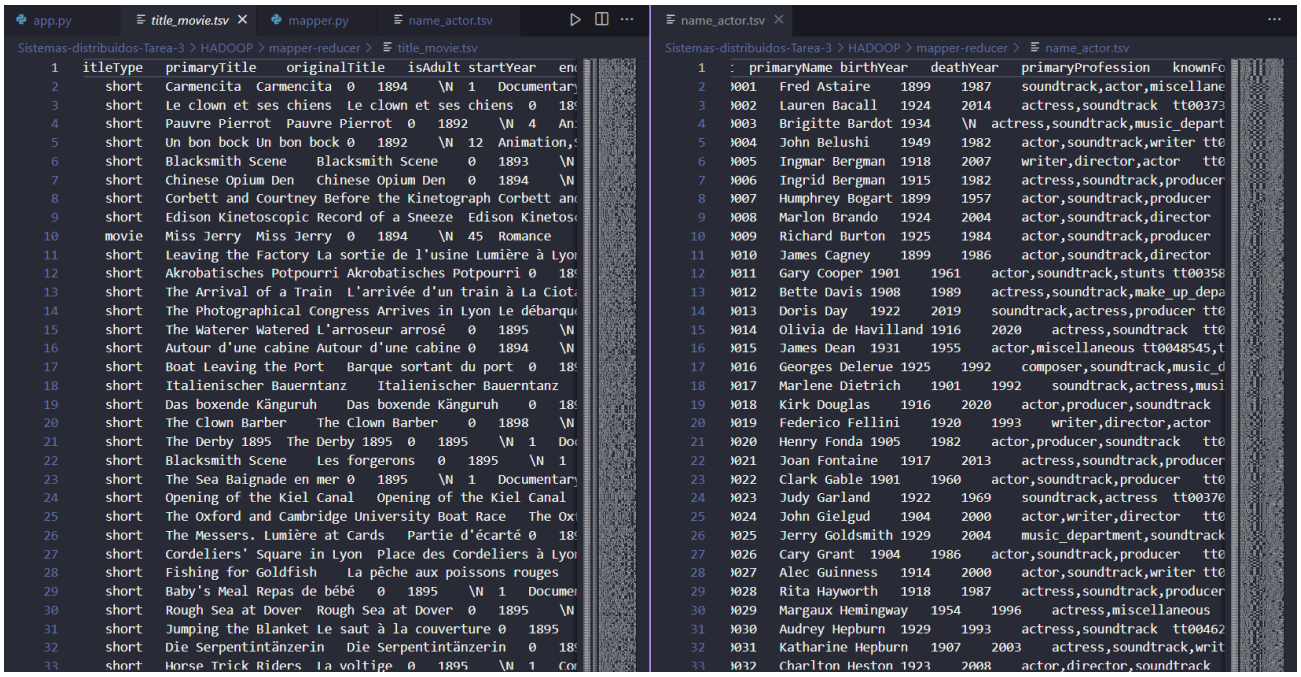
- Docker
- IMDb database
- MapReduce
- Buscador

2.1. Docker

Para esta tarea se realizan dos contenedores de Docker, uno que contiene Hadoop con MapReduce y otro donde se encuentra el buscador de palabras, ambos se encuentran conectados para generar el sistema distribuido.

2.2. IMDb database

En este caso, se trabaja con dos dataset para hacer el índice invertido, uno es `name.basics.tsv.gz` y el otro `title.basics.tsv.gz`.En name basics se tiene toda la información respectiva de los actores, guionistas o directores relacionados con los ids de los trabajos conocidos.



2.3. MapReduce

Para entender mejor MapReduce, se reduce en los siguientes puntos:

- Mapea los inputs para procesar gran cantidad de datos. Como lo son la cantidad de datos relacionados con los artistas y las películas que se presentan.
- Consiste de dos procesos. Uno es mapper, el cual hace los mapas de datos, el mapeo consiste en separar el input en cada id de pelìcula.
- La otra parte es reducer, quien recibe el output de mapper, es decir el mapa de los datos, y se realiza la relación de estos datos con los nombres y se guarda por cada una su título/id conocido. Las películas no las repite, sino que guarda esta una vez y en dónde es que se encontró esa palabra, teniéndola como llave.

Se solicita un diagrama de la utilización de MapReduce, la cual se presenta de la siguiente manera:

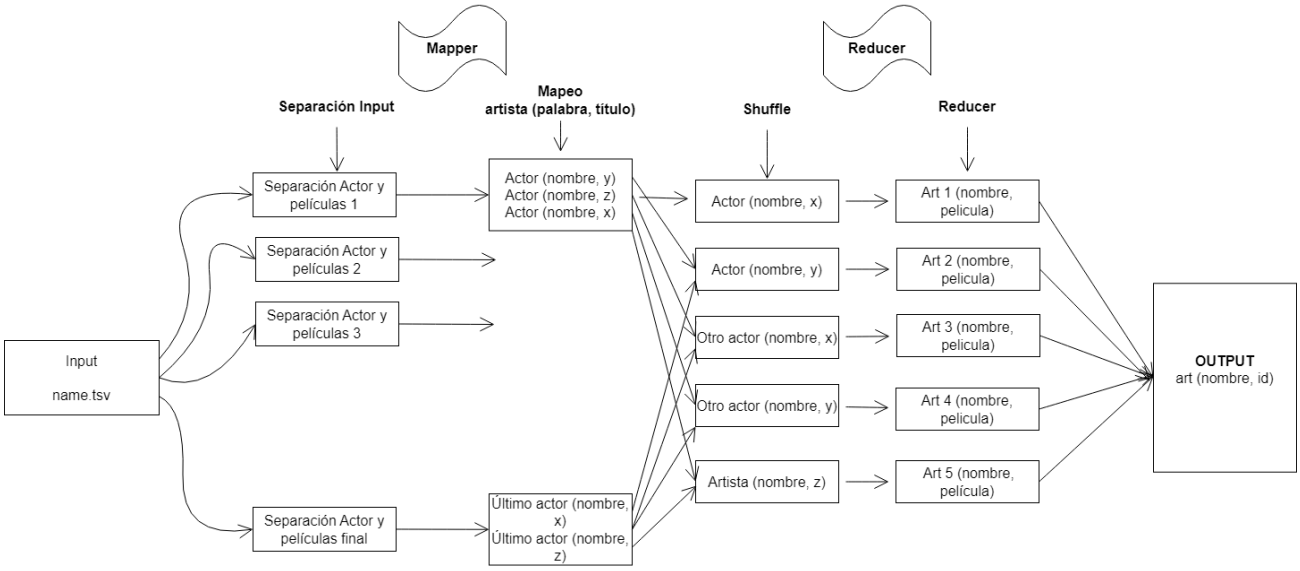


Figura 1: Diagrama de funcionamiento de MapReduce.

2.4. Base de datos

Para guardar los datos obtenidos por MapReduce, se pasan los datos del output del reducer a un archivo resultado tsv, que guarda el índice invertido que asocia el nombre de los artistas con el trabajo que han realizado.

2.5. Buscador

El buscador recibe dos archivos para entregar una respuesta al input del cliente. Estos son la base de datos de IMDb y el archivo que contiene la unión realizada con datos de artístas y las películas..
Se le pregunta al usuario qué artista busca, de tal forma que al recibir este input, la buscará en la base de datos y al encontrarla entregará el id de las películas en cual ha trabajado, haya sido actor, guionista o director.

3. Anexos

1. **Repositorio github**
   `https://github.com/motapod/Sistemas-distribuidos-Tarea-3/tree/main`