



udp UNIVERSIDAD
DIEGO PORTALES

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	2
2. Códigos	2
2.1. Docker	2
2.2. IMDb database	2
2.3. MapReduce	3
2.4. Base de datos	4
2.5. Buscador	4
3. Anexos	5

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.

1	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear
2	short	Carmencita	Carmencita	0	1894	\N
3	short	Le clown et ses chiens	Le clown et ses chiens	0	1894	\N
4	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N
5	short	Un bon bock	Un bon bock	0	1892	\N
6	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N
7	short	Chinese Opium Den	Chinese Opium Den	0	1894	\N
8	short	Corbett and Courtney Before the Kinetograph	Corbett and Courtney Before the Kinetograph	0	1894	\N
9	short	Edison Kinetoscopic Record of a Sneeze	Edison Kinetoscopic Record of a Sneeze	0	1894	\N
10	movie	Miss Jerry	Miss Jerry	0	1894	\N
11	short	Leaving the Factory	La sortie de l'usine	0	1894	\N
12	short	Akrobatisches Potpourri	Akrobatisches Potpourri	0	1894	\N
13	short	The Arrival of a Train	L'arrivée d'un train à La Ciotat	0	1895	\N
14	short	The Photographical Congress Arrives in Lyon	Le débarquement du congrès photographique à Lyon	0	1895	\N
15	short	The Waterer Watered	L'arroseur arrosé	0	1895	\N
16	short	Autour d'une cabine	Autour d'une cabine	0	1894	\N
17	short	Boat Leaving the Port	Barque sortant du port	0	1894	\N
18	short	Italienischer Bauerntanz	Italienischer Bauerntanz	0	1894	\N
19	short	Das boxende Känguruh	Das boxende Känguruh	0	1894	\N
20	short	The Clown Barber	The Clown Barber	0	1898	\N
21	short	The Derby 1895	The Derby 1895	0	1895	\N
22	short	Blacksmith Scene	Les forgerons	0	1895	\N
23	short	The Sea Baignade en mer	The Sea Baignade en mer	0	1895	\N
24	short	Opening of the Kiel Canal	Opening of the Kiel Canal	0	1895	\N
25	short	The Oxford and Cambridge University Boat Race	The Oxford and Cambridge University Boat Race	0	1895	\N
26	short	The Messers.	Lumière at Cards	0	1895	\N
27	short	Cordeliers' Square in Lyon	Place des Cordeliers à Lyon	0	1895	\N
28	short	Fishing for Goldfish	La pêche aux poissons rouges	0	1895	\N
29	short	Baby's Meal	Repas de bébé	0	1895	\N
30	short	Rough Sea at Dover	Rough Sea at Dover	0	1895	\N
31	short	Jumping the Blanket	Le saut à la couverture	0	1895	\N
32	short	Die Serpentina	Die Serpentina	0	1895	\N
33	short	Horse Trick Riders	La voltige	0	1895	\N

1	primaryName	birthYear	deathYear	primaryProfession	knownFor
2	0001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous
3	0002	Lauren Bacall	1924	2014	actress,soundtrack,make_up_department
4	0003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department
5	0004	John Belushi	1949	1982	actor,soundtrack,writer
6	0005	Ingmar Bergman	1918	2007	writer,director,actor
7	0006	Ingrid Bergman	1915	1982	actress,soundtrack,producer
8	0007	Humphrey Bogart	1899	1957	actor,soundtrack,producer
9	0008	Marlon Brando	1924	2004	actor,soundtrack,director
10	0009	Richard Burton	1925	1984	actor,soundtrack,producer
11	0010	James Cagney	1899	1986	actor,soundtrack,director
12	0011	Gary Cooper	1901	1961	actor,soundtrack,stunts
13	0012	Bette Davis	1908	1989	actress,soundtrack,make_up_department
14	0013	Doris Day	1922	2019	soundtrack,actress,producer
15	0014	Olivia de Havilland	1916	2020	actress,soundtrack
16	0015	James Dean	1931	1955	actor,miscellaneous
17	0016	Georges Delerue	1925	1992	composer,soundtrack,music_department
18	0017	Marlene Dietrich	1901	1992	soundtrack,actress,music_department
19	0018	Kirk Douglas	1916	2020	actor,producer,soundtrack
20	0019	Federico Fellini	1920	1993	writer,director,actor
21	0020	Henry Fonda	1905	1982	actor,producer,soundtrack
22	0021	Joan Fontaine	1917	2013	actress,soundtrack,producer
23	0022	Clark Gable	1901	1960	actor,soundtrack,producer
24	0023	Judy Garland	1922	1969	soundtrack,actress
25	0024	John Gielgud	1904	2000	actor,writer,director
26	0025	Jerry Goldsmith	1929	2004	music_department,soundtrack
27	0026	Cary Grant	1904	1986	actor,soundtrack,producer
28	0027	Alec Guinness	1914	2000	actor,soundtrack,writer
29	0028	Rita Hayworth	1918	1987	actress,soundtrack,producer
30	0029	Margaux Hemingway	1954	1996	actress,miscellaneous
31	0030	Audrey Hepburn	1929	1993	actress,soundtrack
32	0031	Katharine Hepburn	1907	2003	actress,soundtrack,writer
33	0032	Charlton Heston	1923	2008	actor,director,soundtrack

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 película.
- 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:

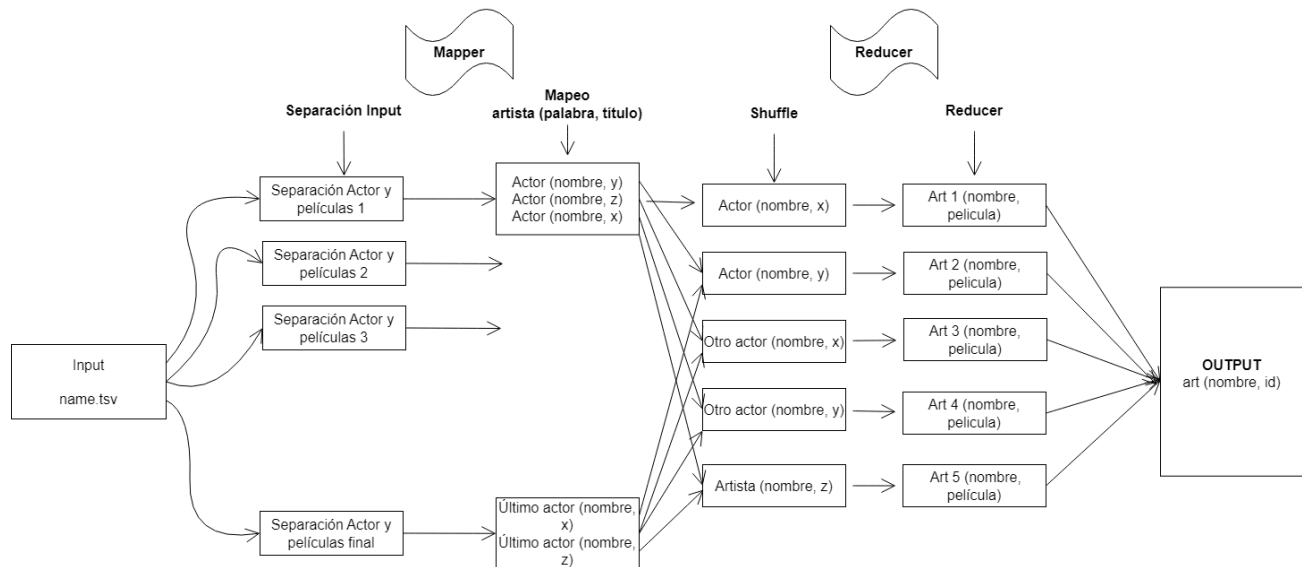


Figura 1: Diagrama de funcionamiento de MapReduce.

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 artistas 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>