

The background of the entire image is a solid red color. Overlaid on this background is a large, stylized arch composed of numerous small, light-red dots. The dots are arranged in a way that creates a sense of depth and movement, with the arch curving from the left side towards the right. In the center of this arch, the letters "HUST" are displayed in a bold, white, sans-serif font.

HUST

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

PHÂN TÍCH DỮ LIỆU PHIM LẺ

Nhóm: 19

Giảng viên hướng dẫn: TS. Trần Việt Trung

ONE LOVE. ONE FUTURE.

- 1. Giới thiệu đề tài
- 2. Kiến trúc tổng quan
- 3. Các trải nghiệm



HUST

1. Giới thiệu đề tài

1. Giới thiệu đề tài

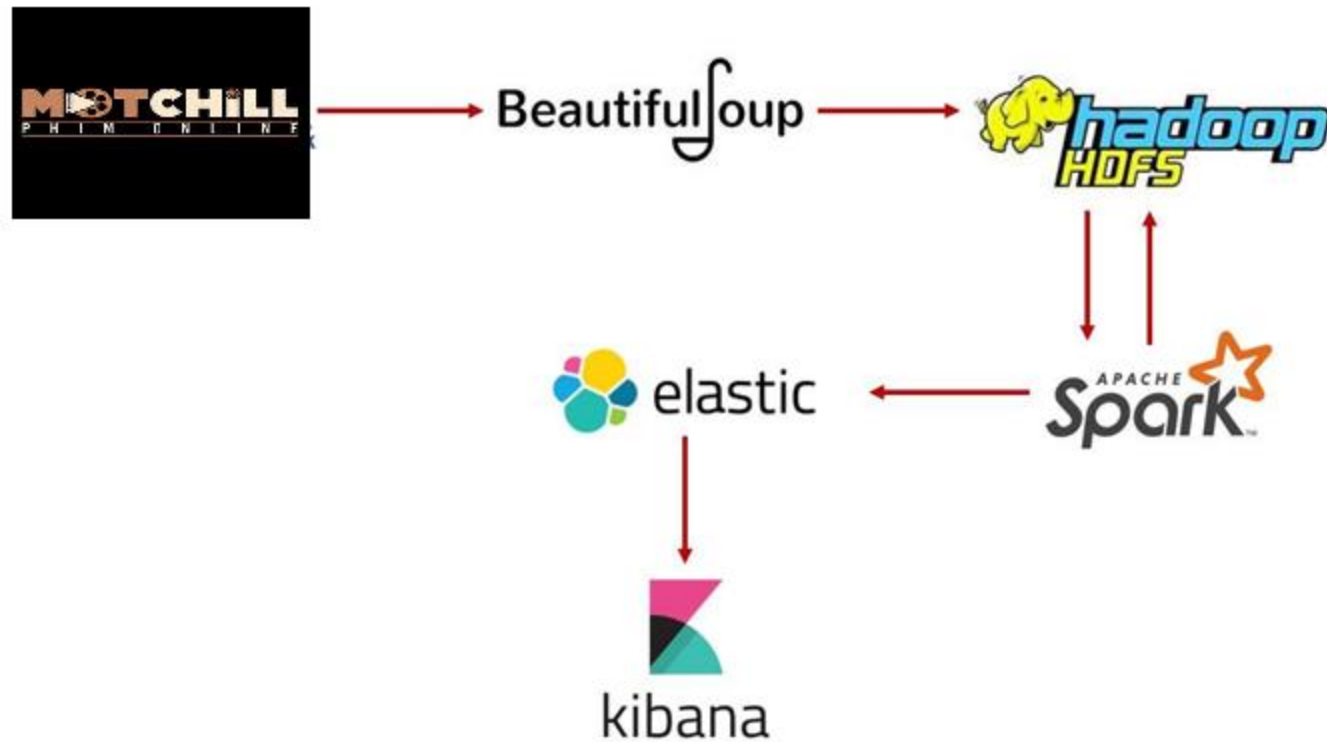
- Hiện nay, nhu cầu giải trí, đặc biệt là xem phim đã trở nên quá phổ biến trong cuộc sống. Mỗi ngày có hàng triệu người trên thế giới thưởng thức các bộ phim từ nhiều thể loại, phong cách khác nhau.
- Không chỉ là một hình thức giải trí, phim ảnh còn phản ánh văn hóa, xã hội và thị hiếu của khán giả. Việc hiểu rõ sở thích và hành vi của người xem đã trở thành một yếu tố quan trọng.
- Thông qua việc thu thập, xử lý và phân tích dữ liệu phim, chúng ta có thể có cái nhìn sâu sắc về chất lượng, mức độ phổ biến và thành công của các tác phẩm điện ảnh.



HUST

1. Kiến trúc tổng quan

1. Kiến trúc tổng quan





HUST

3. Các trải nghiệm

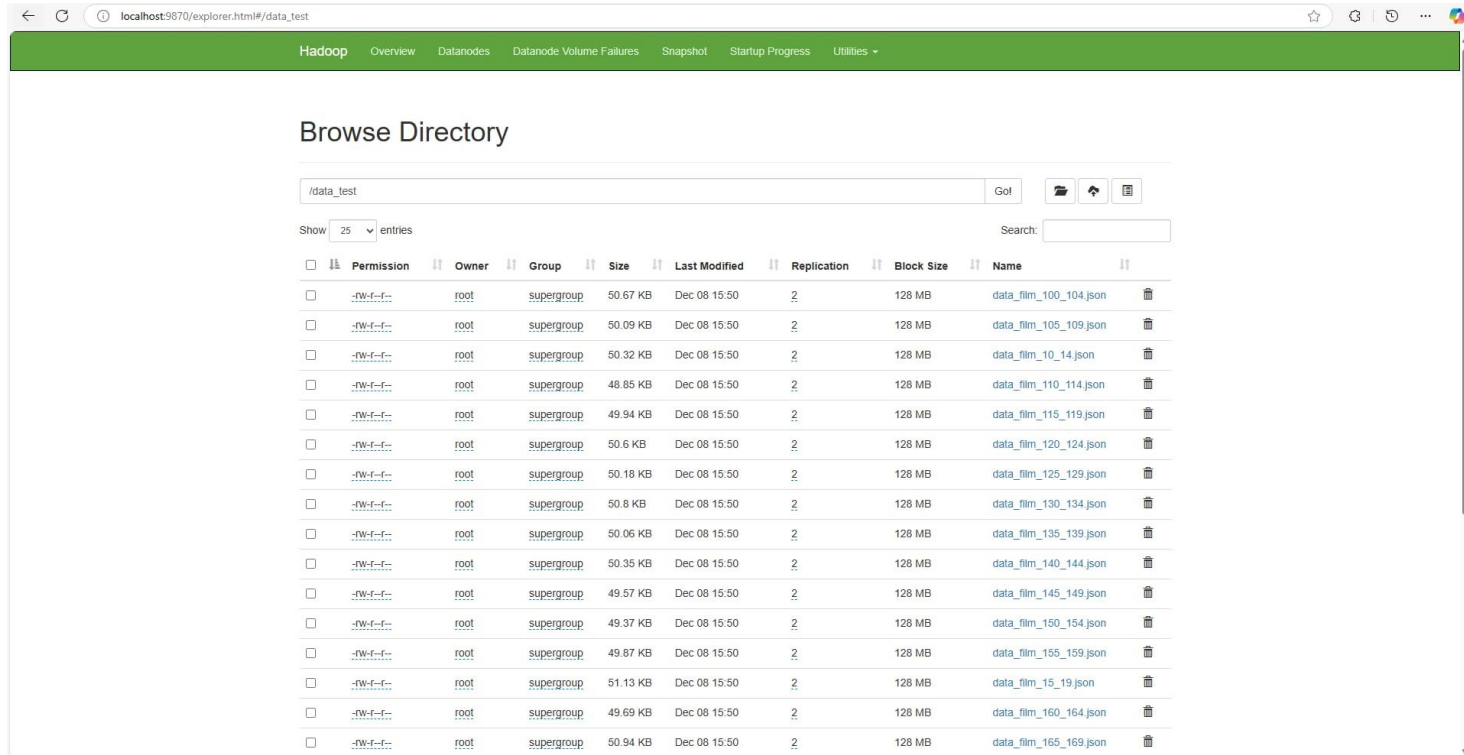
3. Các trải nghiệm

Trải nghiệm 1: Thu nhập dữ liệu

```
[{  
  "name": "Thiếu Chút Mỹ Vị",  
  "link": "https://motphim.ad/phim/thieu-chut-my-vi",  
  "type": "Phim bộ",  
  "Năm phát hành": "2024",  
  "Trạng thái": "Hoàn Tất (24/24) Vietsub",  
  "Số tập": "24 Tập",  
  "Tình trạng": "Đang chiếu",  
  "Thể loại": "Tình Cảm",  
  "Đạo diễn": "Xiu Xiao Nan",  
  "Diễn viên": "Lý Minh Tuấn",  
  "Đánh giá": "8.0"  
},
```

3. Các trải nghiệm

Trải nghiệm 2: Chuyển dữ liệu vào Hadoop để lưu trữ



The screenshot displays the Hadoop Distributed File System (HDFS) Explorer interface in a web browser. The address bar shows the URL `localhost:9870/explorer.html#/data_test`. The interface has a green header bar with navigation links: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. Below the header, the title "Browse Directory" is visible. A search bar and a "Go!" button are present. The main content area shows a table of files and directories. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The files listed are all JSON files with names like `data_film_100_104.json`, `data_film_105_109.json`, etc., up to `data_film_165_169.json`. Each file has a size of 50.67 KB to 51.13 KB, a last modified date of Dec 08 15:50, and a replication factor of 2. The block size for all files is 128 MB. The owner is `root` and the group is `supergroup`.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	50.67 KB	Dec 08 15:50	2	128 MB	data_film_100_104.json
-rw-r--r--	root	supergroup	50.09 KB	Dec 08 15:50	2	128 MB	data_film_105_109.json
-rw-r--r--	root	supergroup	50.32 KB	Dec 08 15:50	2	128 MB	data_film_10_14.json
-rw-r--r--	root	supergroup	48.85 KB	Dec 08 15:50	2	128 MB	data_film_110_114.json
-rw-r--r--	root	supergroup	49.94 KB	Dec 08 15:50	2	128 MB	data_film_115_119.json
-rw-r--r--	root	supergroup	50.6 KB	Dec 08 15:50	2	128 MB	data_film_120_124.json
-rw-r--r--	root	supergroup	50.18 KB	Dec 08 15:50	2	128 MB	data_film_125_129.json
-rw-r--r--	root	supergroup	50.8 KB	Dec 08 15:50	2	128 MB	data_film_130_134.json
-rw-r--r--	root	supergroup	50.06 KB	Dec 08 15:50	2	128 MB	data_film_135_139.json
-rw-r--r--	root	supergroup	50.35 KB	Dec 08 15:50	2	128 MB	data_film_140_144.json
-rw-r--r--	root	supergroup	49.57 KB	Dec 08 15:50	2	128 MB	data_film_145_149.json
-rw-r--r--	root	supergroup	49.37 KB	Dec 08 15:50	2	128 MB	data_film_150_154.json
-rw-r--r--	root	supergroup	49.87 KB	Dec 08 15:50	2	128 MB	data_film_155_159.json
-rw-r--r--	root	supergroup	51.13 KB	Dec 08 15:50	2	128 MB	data_film_15_19.json
-rw-r--r--	root	supergroup	49.69 KB	Dec 08 15:50	2	128 MB	data_film_160_164.json
-rw-r--r--	root	supergroup	50.94 KB	Dec 08 15:50	2	128 MB	data_film_165_169.json

3. Các trải nghiệm

Trải nghiệm 2: Chuyển dữ liệu vào Hadoop để lưu trữ

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview 'namenode:9000' (✓active)

Started:	Tue Dec 10 07:57:10 +0700 2024
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 22:56:00 +0700 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-963602b5-b9b4-42f0-8e31-680c21515f77
Block Pool ID:	BP-171222524-172.19.0.4-1733570200758

Summary

Security is off.

Safemode is off.

59 files and directories, 54 blocks (54 replicated blocks, 0 erasure coded block groups) = 113 total filesystem object(s).

Heap Memory used 133.17 MB of 328.5 MB Heap Memory. Max Heap Memory is 1.71 GB.

Non Heap Memory used 52.17 MB of 53.63 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	460.73 GB
Configured Remote Capacity:	0 B
DFS Used:	6.74 MB (0%)
Non DFS Used:	346.19 GB
DFS Remaining:	114.53 GB (24.86%)
Block Pool Used:	6.74 MB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)

3. Các trải nghiệm

Trải nghiệm 3: Sử dụng Spark để xử lý dữ liệu

```
schema = StructType([
    StructField("name", StringType(), True),
    StructField("link", StringType(), True),
    StructField("type", StringType(), True),
    StructField("Năm phát hành", StringType(), True),
    StructField("Trạng thái", StringType(), True),
    StructField("Số tập", StringType(), True),
    StructField("Tình trạng", StringType(), True),
    StructField("Thể loại", StringType(), True),
    StructField("Đạo diễn", StringType(), True),
    StructField("Diễn viên", StringType(), True),
    StructField("Đánh giá", StringType(), True)
])
```

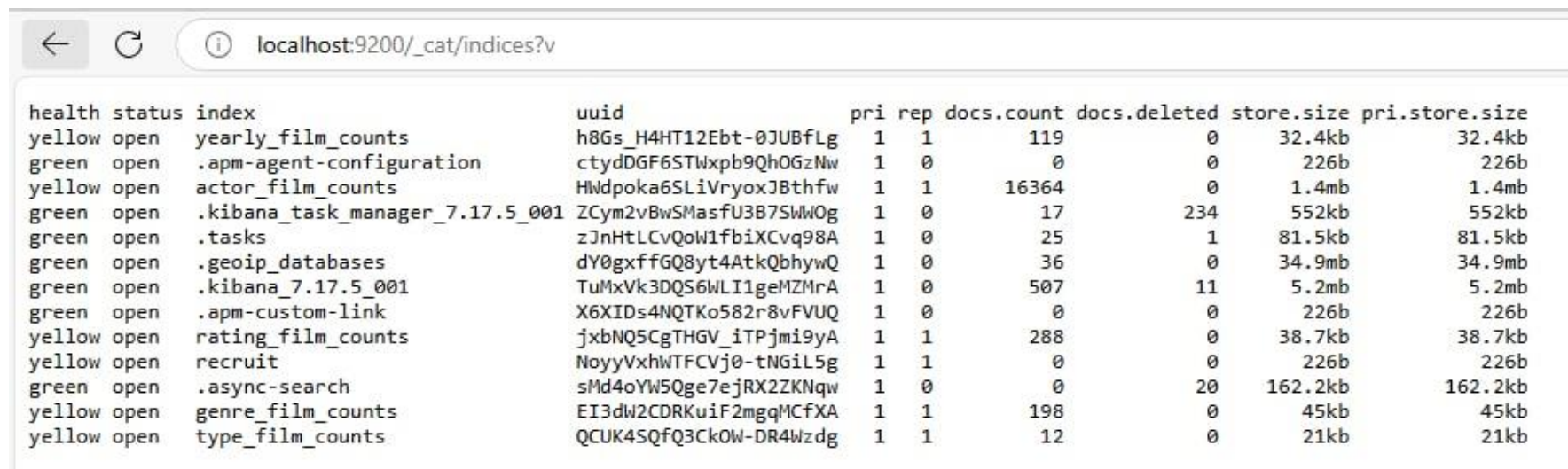
3. Các trải nghiệm

Trải nghiệm 3: Sử dụng Spark để xử lý dữ liệu

```
extracted_recruit_df = raw_recruit_df.select(raw_recruit_df["name"].alias("FilmName"),
                                             raw_recruit_df["link"].alias("LinkFilm"),
                                             udfs.extract_type(raw_recruit_df["type"]).alias("IsMovies"),
                                             udfs.extract_release_year(raw_recruit_df["Năm phát hành"]).alias("ReleaseYear"),
                                             udfs.extract_status(raw_recruit_df["Trạng thái"], raw_recruit_df["Số tập"]).alias("Status"),
                                             udfs.extract_episode_count(raw_recruit_df["Số tập"]).alias("EpisodeCount"),
                                             udfs.map_condition(raw_recruit_df["Tình trạng"]).alias("Condition"),
                                             udfs.extract_genres(raw_recruit_df["Thể loại"]).alias("Genres"),
                                             udfs.map_director(raw_recruit_df["Đạo diễn"]).alias("Director"),
                                             udfs.extract_actors(raw_recruit_df["Diễn viên"]).alias("Actors"),
                                             udfs.extract_rating(raw_recruit_df["Đánh giá"]).alias("Rating")
                                             )
extracted_recruit_df.cache()
extracted_recruit_df.show(5)
```

3. Các trải nghiệm

Trải nghiệm 4: Elasticsearch



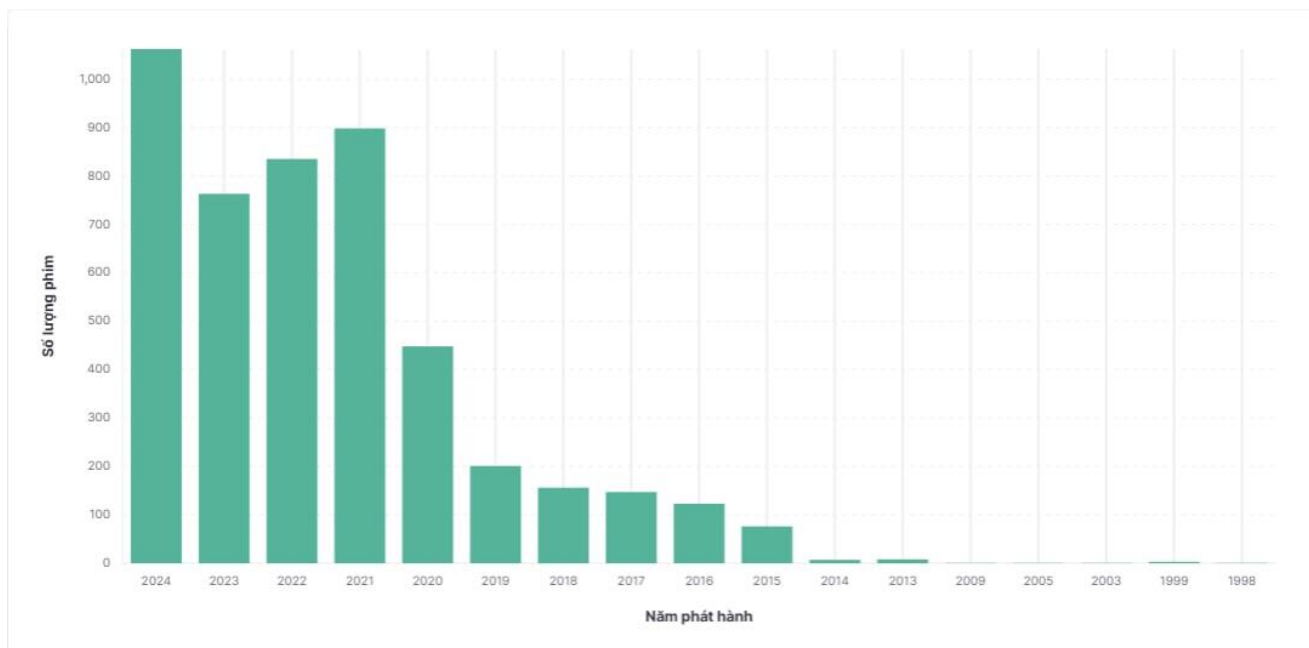
The screenshot shows a web browser window with the address bar displaying 'localhost:9200/_cat/indices?v'. The main content area displays a table of Elasticsearch indices. The table has columns for health, status, index, uuid, pri, rep, docs.count, docs.deleted, store.size, and pri.store.size. The data is as follows:

health	status	index	uuid	pri	rep	docs.count	docs.deleted	store.size	pri.store.size
yellow	open	yearly_film_counts	h8Gs_H4HT12Ebt-0JUBfLg	1	1	119	0	32.4kb	32.4kb
green	open	.apm-agent-configuration	ctydDGF6STWxpb9Qh0GzNw	1	0	0	0	226b	226b
yellow	open	actor_film_counts	HWdpoka6SLiVryoxJBthfw	1	1	16364	0	1.4mb	1.4mb
green	open	.kibana_task_manager_7.17.5_001	ZCym2vBwSMasfU3B7SWW0g	1	0	17	234	552kb	552kb
green	open	.tasks	zJnHtLCvQoW1fbiXCvq98A	1	0	25	1	81.5kb	81.5kb
green	open	.geoip_databases	dY0gxffGQ8yt4AtkQbhywQ	1	0	36	0	34.9mb	34.9mb
green	open	.kibana_7.17.5_001	TuMxVk3DQS6WLI1geMZMrA	1	0	507	11	5.2mb	5.2mb
green	open	.apm-custom-link	X6XIDs4NQTKo582r8vFVUQ	1	0	0	0	226b	226b
yellow	open	rating_film_counts	jxbNQ5CgTHGV_iTPjmi9yA	1	1	288	0	38.7kb	38.7kb
yellow	open	recruit	NoyyVxhWTFcvj0-tNGil5g	1	1	0	0	226b	226b
green	open	.async-search	sMd4oYW5Qge7ejRXZ2KNqw	1	0	0	20	162.2kb	162.2kb
yellow	open	genre_film_counts	EI3dw2CDRKuif2mgqMCFXA	1	1	198	0	45kb	45kb
yellow	open	type_film_counts	QCUK4SQfQ3CkOW-DR4Wzdg	1	1	12	0	21kb	21kb

3. Các trải nghiệm

Trải nghiệm 5: Biểu diễn dữ liệu bằng Kibana

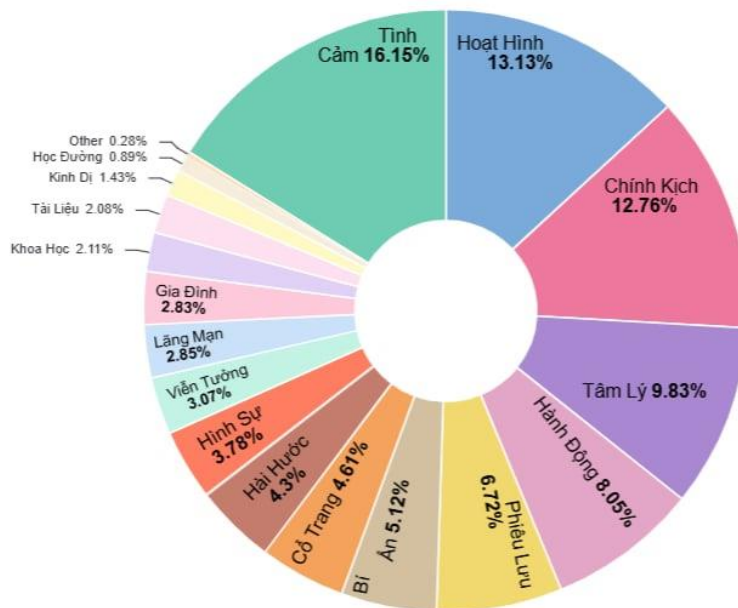
Số lượng phim theo năm phát hành



3. Các trải nghiệm

Trải nghiệm 5: Biểu diễn dữ liệu bằng Kibana

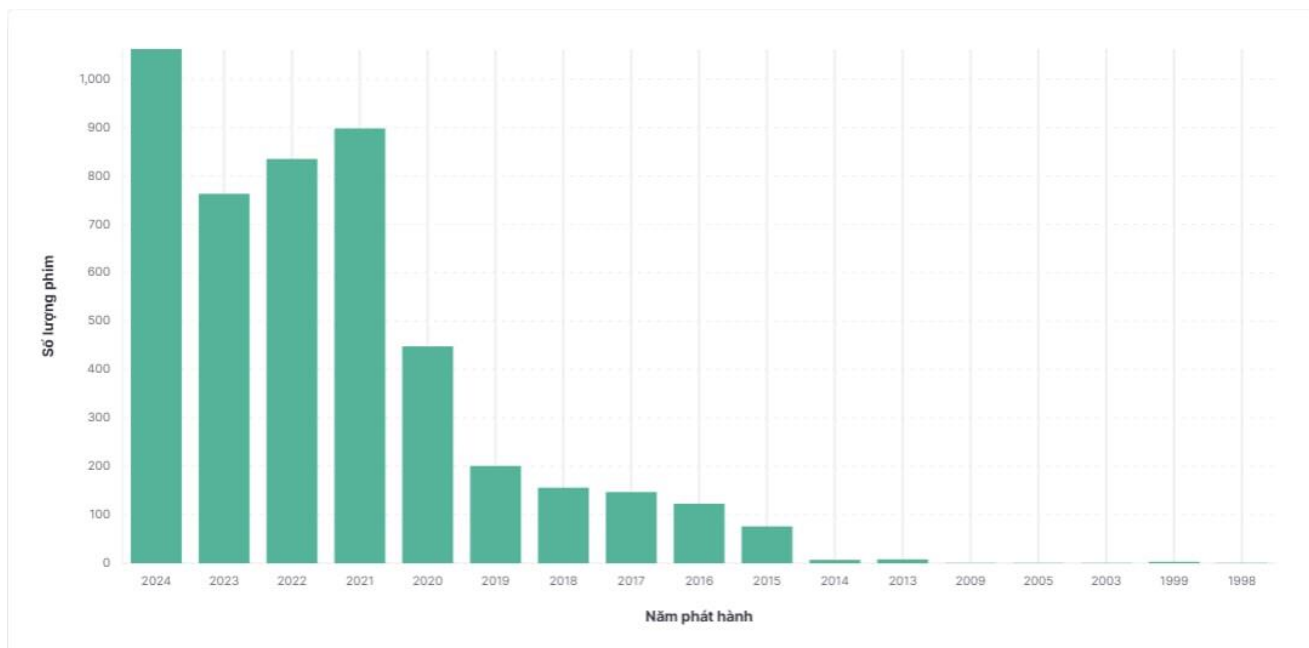
Các thể loại phim



3. Các trải nghiệm

Trải nghiệm 5: Biểu diễn dữ liệu bằng Kibana

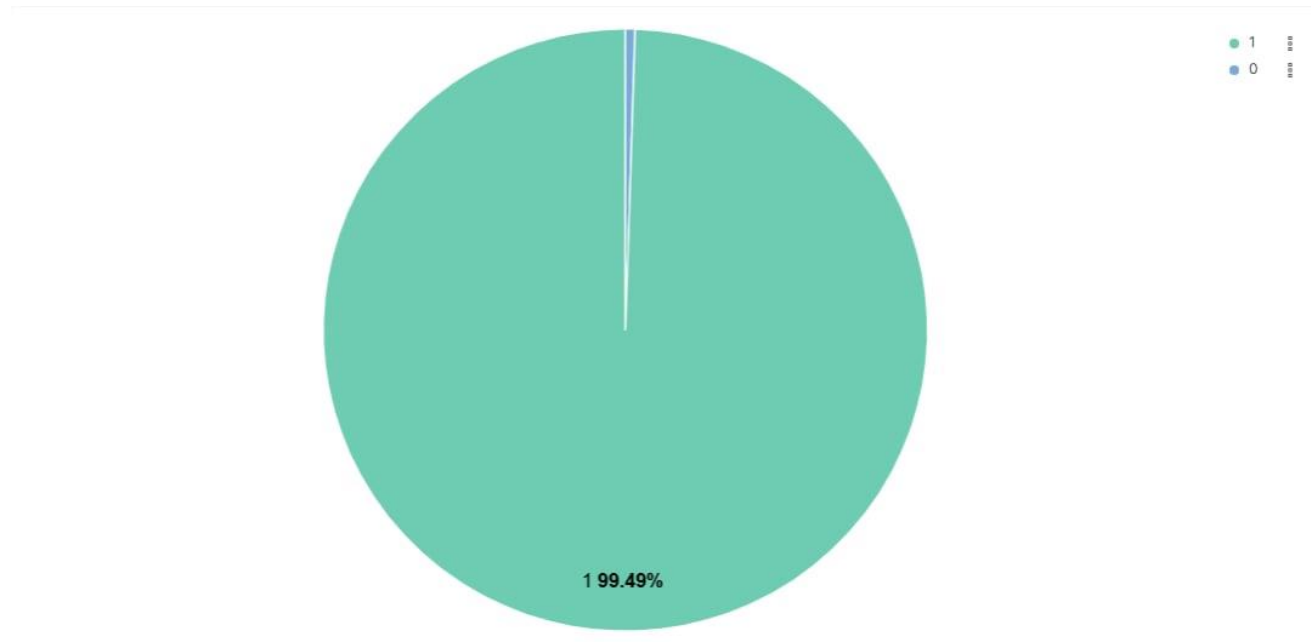
Số lượng phim theo năm phát hành



3. Các trải nghiệm

Trải nghiệm 5: Biểu diễn dữ liệu bằng Kibana

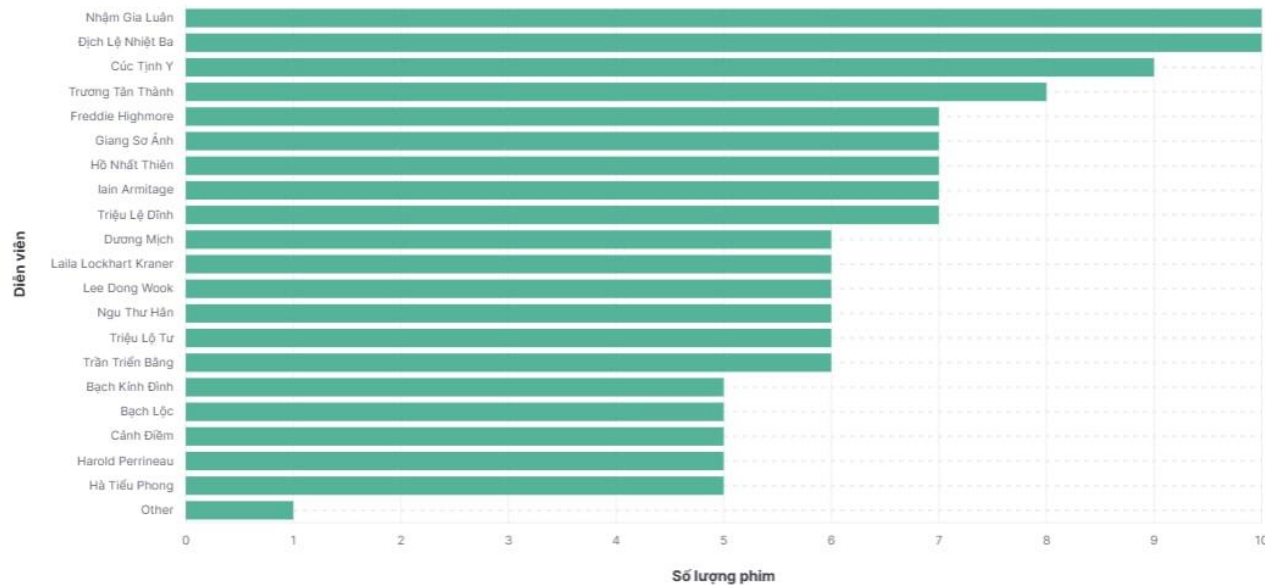
Tỉ lệ phim bộ và phim lẻ



3. Các trải nghiệm

Trải nghiệm 5: Biểu diễn dữ liệu bằng Kibana

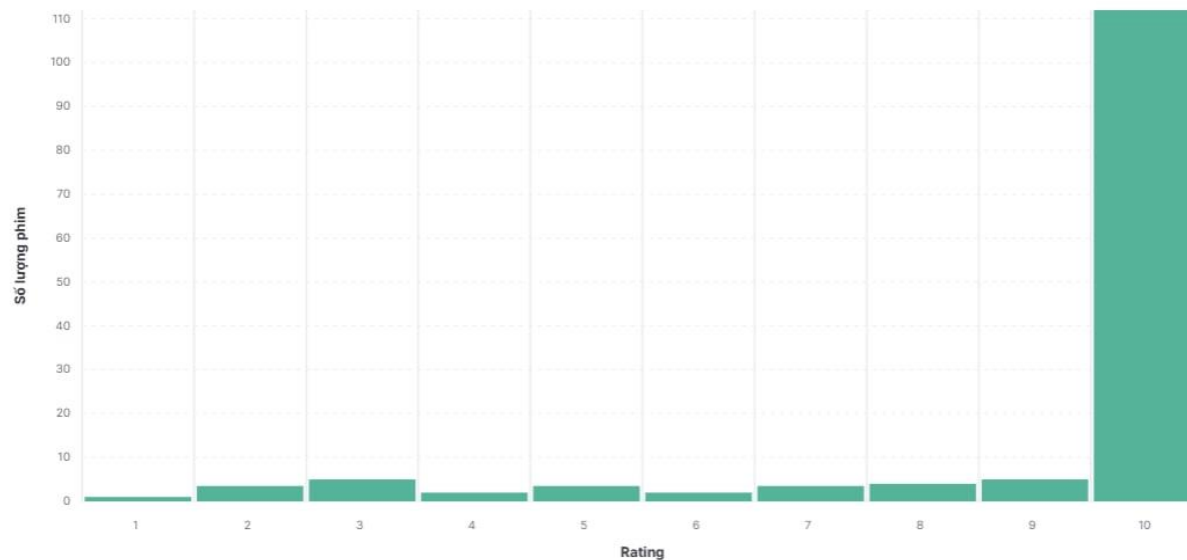
Số lượng phim của các diễn viên đóng



3. Các trải nghiệm

Trải nghiệm 5: Biểu diễn dữ liệu bằng Kibana

Số lượng phim theo đánh giá từ 1-10



A large, stylized graphic on the left side of the slide. It consists of a red background with a circular pattern of white dots of varying sizes, creating a sense of depth and movement. The word "HUST" is written in white, bold, sans-serif capital letters in the center of this graphic.

HUST

THANK YOU !