Thu thập, lưu trữ và xử lý dữ liệu báo điện tử sử dụng Spark, Hadoop và ELK Stack

Nhóm sinh viên: Zootopia
Đàm Trọng Tuyên 20173364
Ngô Việt Hoàng 20173142
Nguyễn Thị Nhung 20173294

Nội dung

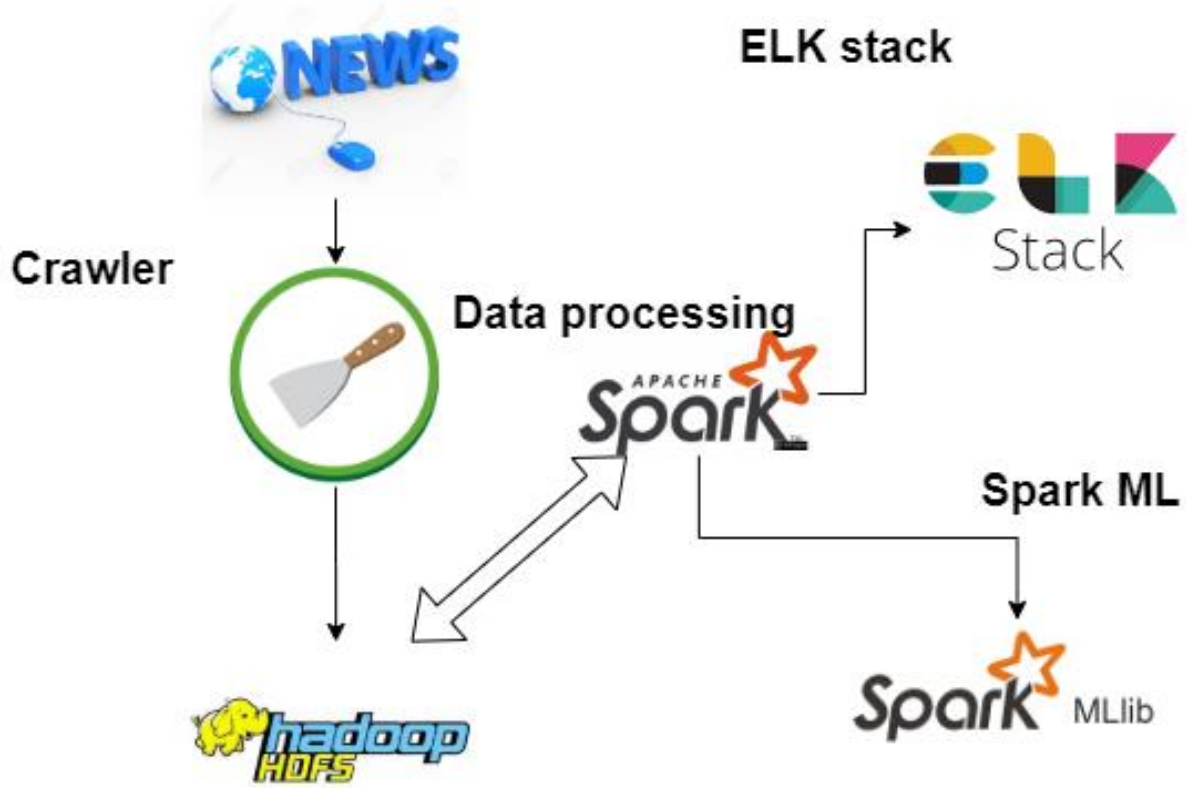
1. Giới thiệu
2. Tổng quan hệ thống
3. Thu thập, xử lý và lưu trữ dữ liệu
4. Quản lý dữ liệu với ELK Stack
5. Phân loại văn bản sử dụng SparkML
6. Tổng kết

1. Giới thiệu

- Dữ liệu là tài nguyên quý giá trong thời đại hiện nay
  - Nắm bắt và sử dụng dữ liệu tạo nên lợi thế cho các quốc gia, tổ chức và cá nhân
- Dự án của nhóm: Xây dựng hệ thống quy mô nhỏ thực hiện:
  - Thu thập dữ liệu trừ một số trang báo trực tuyến
  - Tích hợp dữ liệu
  - Lưu trữ phân tán
  - Xử lí và làm sạch dữ liệu
  - Quản lý dữ liệu đã thu thập
  - Xây dựng mô hình phân loại văn bản

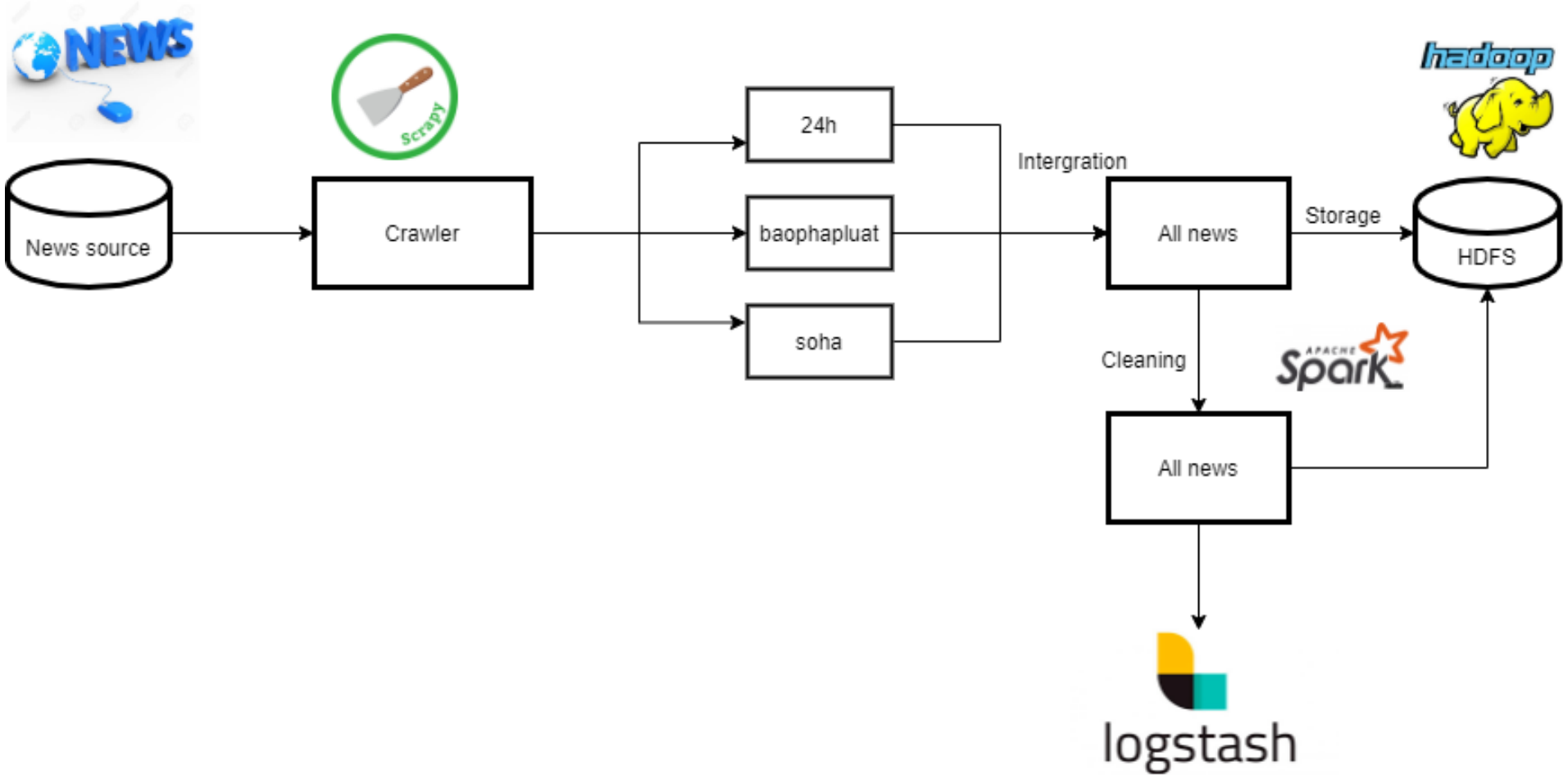
2. Tổng quan hệ thống

- Các thành phần hệ thống:
  - Thu thập dữ liệu: Scrapy
  - Lưu trữ dữ liệu: Hadoop
  - Xử lí và tích hợp dữ liệu: Spark
  - Biểu diễn dữ liệu: ELK Stack
- Đạt được các yêu cầu đối với hệ thống dữ liệu lớn về:
  - Khả năng mở rộng
  - Tính bền vững của dữ liệu
  - Tốc độ tính toán cao
  - Khả năng trực quan hóa dữ liệu



3. Thu thập, xử lý và lưu trữ dữ liệu

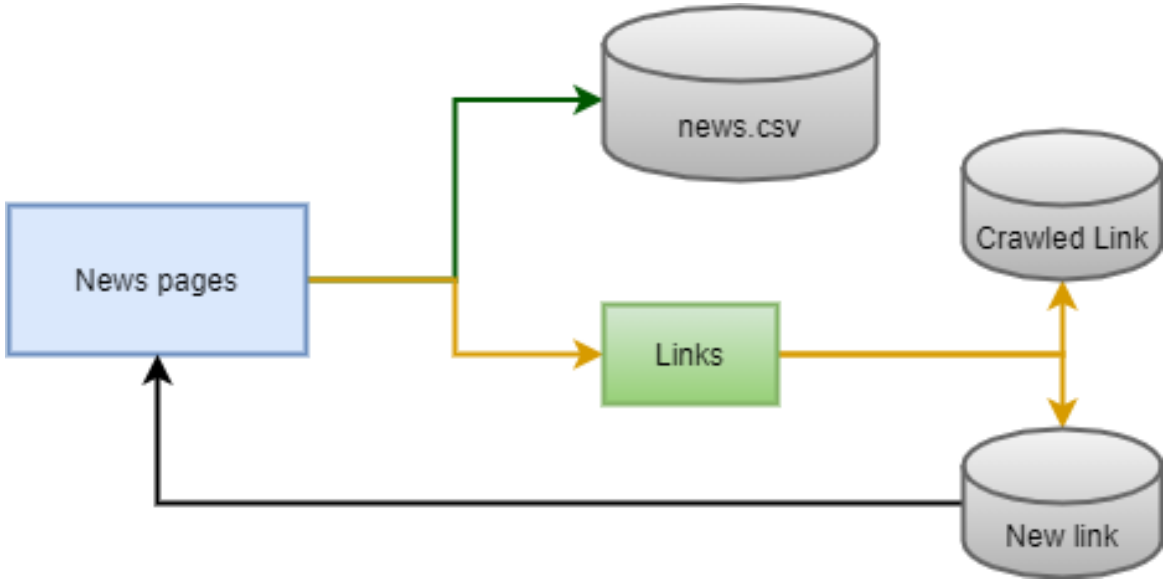
Luồng thu thập và xử lý dữ liệu



3.1. Thu thập dữ liệu

Thu thập dữ liệu

- Công cụ sử dụng: Scrapy (python).
- Đầu vào: link bài báo
- Đầu ra: các thông tin: tiêu đề, thể loại, nội dung, ngày đăng, link các bài báo liên quan.
- Phương pháp: duyệt theo chiều rộng .
- Nguồn: soha.vn, baophapluat.vn, 24h.com
- Thời gian: 8 giờ.



Chi tiết luồng xử lý thu thập dữ liệu

3.1. Thu thập dữ liệu (tiếp)

Kết quả thu thập dữ liệu

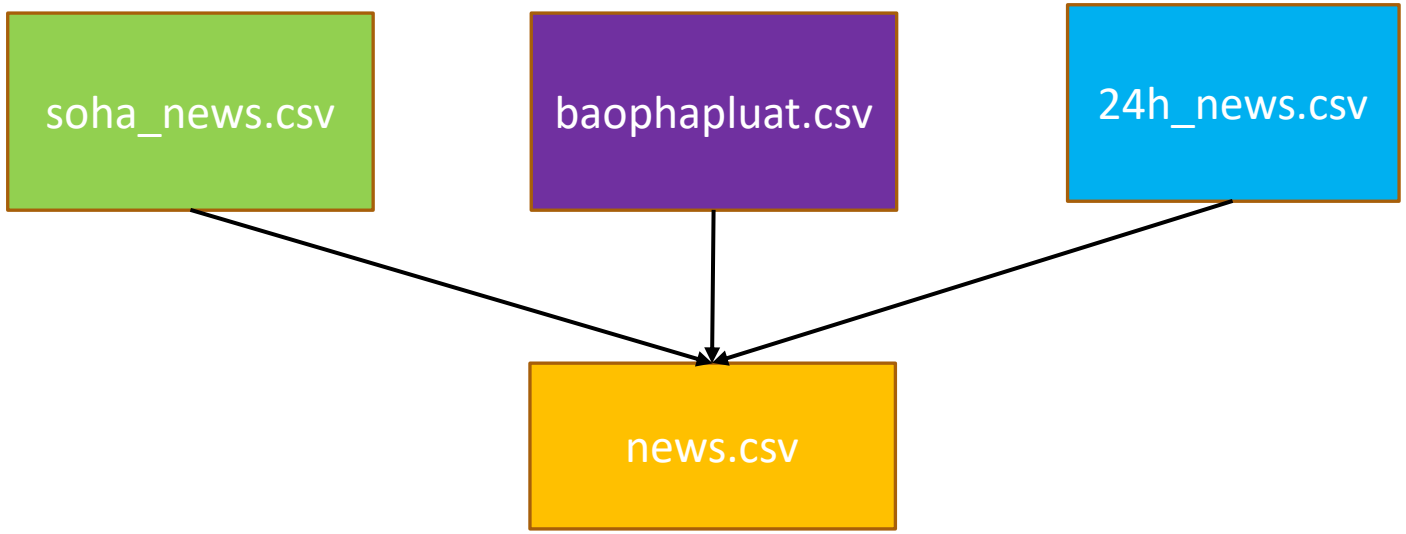
Nguồn	Link đã duyệt qua	Số lượng bài báo	Dung lượng (MB)
24h.com	~800.000	555143	802
Baophapluat.vn	~100.000	95690	314
Soha.vn	~300.000	196497	613

Kết quả thu thập dữ liệu

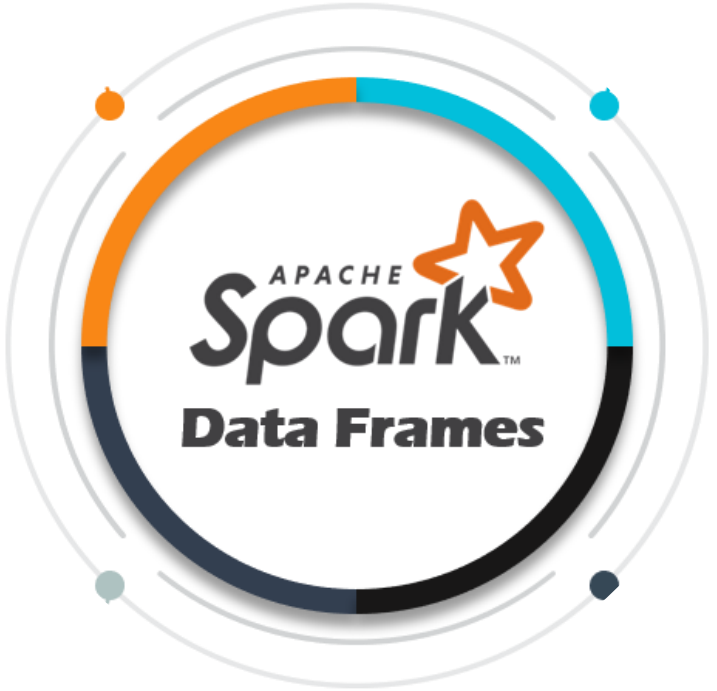
3.2. Xử lý dữ liệu

Xử lý dữ liệu:

- Công cụ: Spark
- Loại bỏ bản ghi bị lỗi: thiếu nội dung, tiêu đề, thể loại...
- Lưu lại thành một file duy nhất, thống nhất về các trường dữ liệu.



Tích hợp dữ liệu



3.2. Lưu trữ dữ liệu

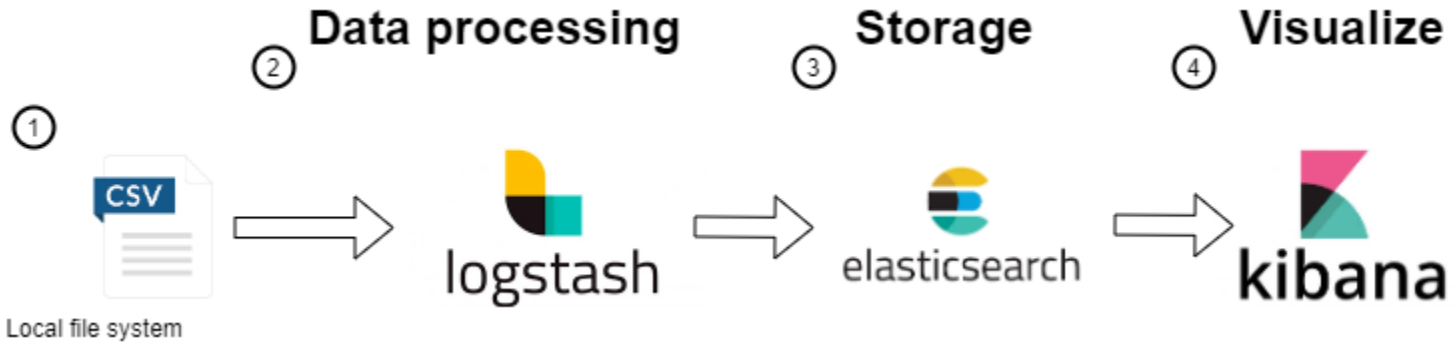
Kết quả lưu trữ dữ liệu

Owner	Group	Size	Last Modified	Replication	Block Size	Name
hadoop	supergroup	1.61 GB	Dec 21 16:08	2	64 MB	all_news_cleaned.csv
hadoop	supergroup	802.19 MB	Dec 21 15:56	2	64 MB	news_24h.csv
hadoop	supergroup	314.21 MB	Dec 21 16:00	2	64 MB	news_laws.csv
hadoop	supergroup	612.84 MB	Dec 21 15:59	2	64 MB	news_soha.csv

Dữ liệu lưu trữ trong cụm Hadoop

4. Quản lý dữ liệu với ELK Stack

- Lưu trữ dữ liệu vào Elasticsearch
  - Thông thường: Tạo file JSON, đưa vào bằng bulk API -> gặp vấn đề với file lớn hơn 2GB.
  - Sử dụng Logstash để tự động gửi dữ liệu vào Elasticsearch.
- Luồng dữ liệu trong ELK Stack



4. Quản lý dữ liệu với ELK Stack (tiếp)

Dữ liệu được lưu trong Elasticsearch

Pretty Raw Preview Visualize Text

```
health status index                uuid                   pri rep docs.count docs.deleted store.size pri.store.size
green  open   zootopia_news        vmYbKdCPTfyTE0Sgojx5QA   5   1     605551            0        4gb            2gb
green  open   .kibana_1            BccQ7kL_TuWj2AtPLqbIxw   1   1          5            0     42.4kb         21.2kb
green  open   zootopia             G0TVaOmcSOSwlZ1WMVUszA   5   1     386163            0    525.1mb        262.8mb
green  open   .kibana_task_manager VpeNX-cFR9iInk5vXpaXdg   1   1          2            0     25.1kb         12.5kb
```

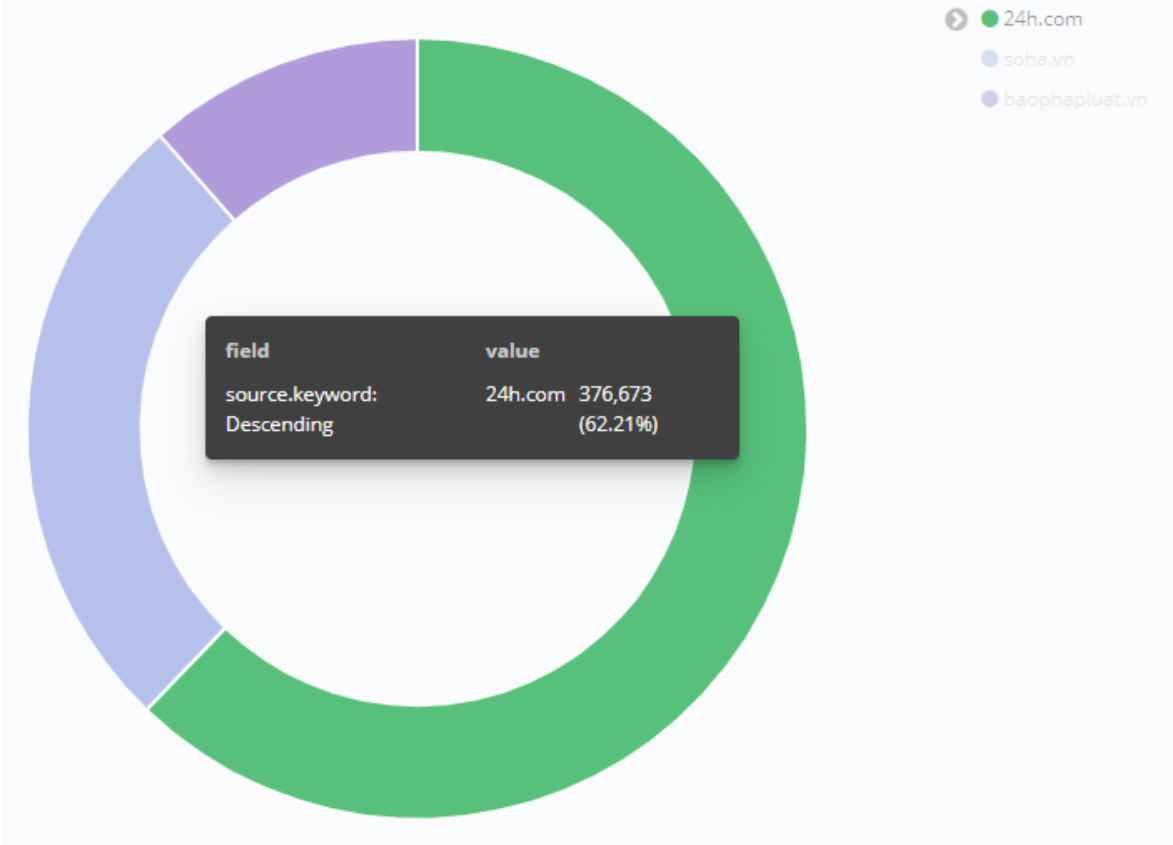
4. Quản lý dữ liệu với ELK Stack (tiếp)

Phân bố dữ liệu trên các máy

index	shard	prirep	state	docs	store	ip	node
zootopia	3	r	STARTED	77651	52.7mb	192.168.0.113	node3
zootopia	3	p	STARTED	77651	52.2mb	192.168.0.112	node2
zootopia	1	p	STARTED	76792	52.5mb	192.168.0.113	node3
zootopia	1	r	STARTED	76792	52.6mb	192.168.0.112	node2
zootopia	2	p	STARTED	77200	52.7mb	192.168.0.111	node1
zootopia	2	r	STARTED	77200	52.7mb	192.168.0.112	node2
zootopia	4	p	STARTED	77371	52.3mb	192.168.0.111	node1
zootopia	4	r	STARTED	77371	52.2mb	192.168.0.113	node3
zootopia	0	p	STARTED	77149	52.4mb	192.168.0.111	node1
zootopia	0	r	STARTED	77149	52.4mb	192.168.0.112	node2
zootopia_news	3	r	STARTED	120855	410.6mb	192.168.0.113	node3
zootopia_news	3	p	STARTED	120855	411.8mb	192.168.0.112	node2
zootopia_news	1	p	STARTED	121415	415.5mb	192.168.0.111	node1
zootopia_news	1	r	STARTED	121415	414.1mb	192.168.0.112	node2
zootopia_news	2	p	STARTED	120915	412.9mb	192.168.0.113	node3
zootopia_news	2	r	STARTED	120915	411.3mb	192.168.0.112	node2
zootopia_news	4	p	STARTED	121321	412.7mb	192.168.0.111	node1
zootopia_news	4	r	STARTED	121321	412.5mb	192.168.0.113	node3
zootopia_news	0	r	STARTED	121045	405.3mb	192.168.0.111	node1
zootopia_news	0	p	STARTED	121045	408.2mb	192.168.0.112	node2
.kibana_task_manager	0	r	STARTED	2	12.5kb	192.168.0.111	node1
.kibana_task_manager	0	p	STARTED	2	12.5kb	192.168.0.113	node3
.kibana_1	0	p	STARTED	5	21.2kb	192.168.0.111	node1
.kibana_1	0	r	STARTED	5	21.2kb	192.168.0.113	node3

4. Quản lý dữ liệu với ELK Stack (tiếp)

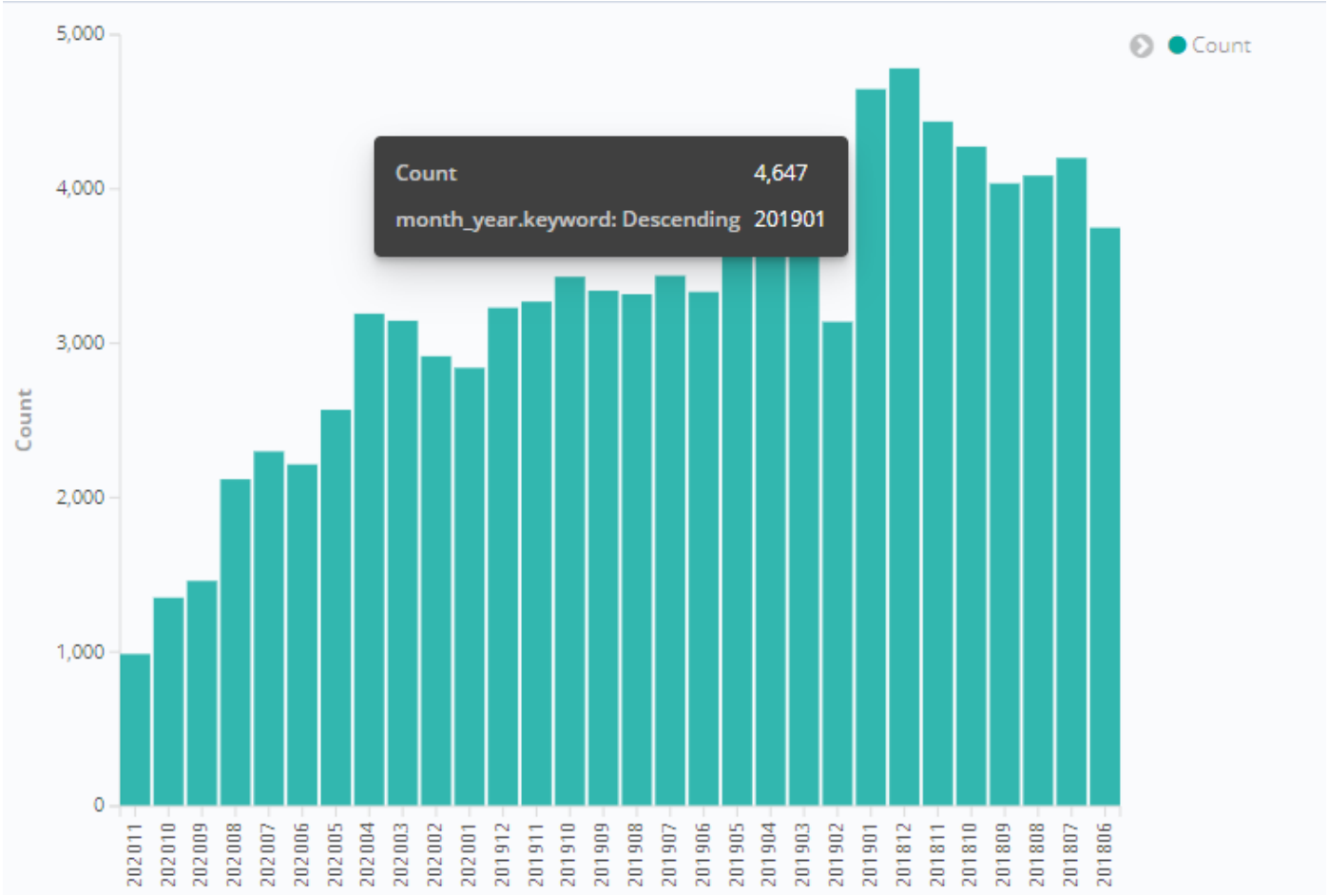
Biểu diễn dữ liệu bằng Kibana



Phân phối dữ liệu theo nguồn thu thập

4. Quản lý dữ liệu với ELK stack (tiếp)

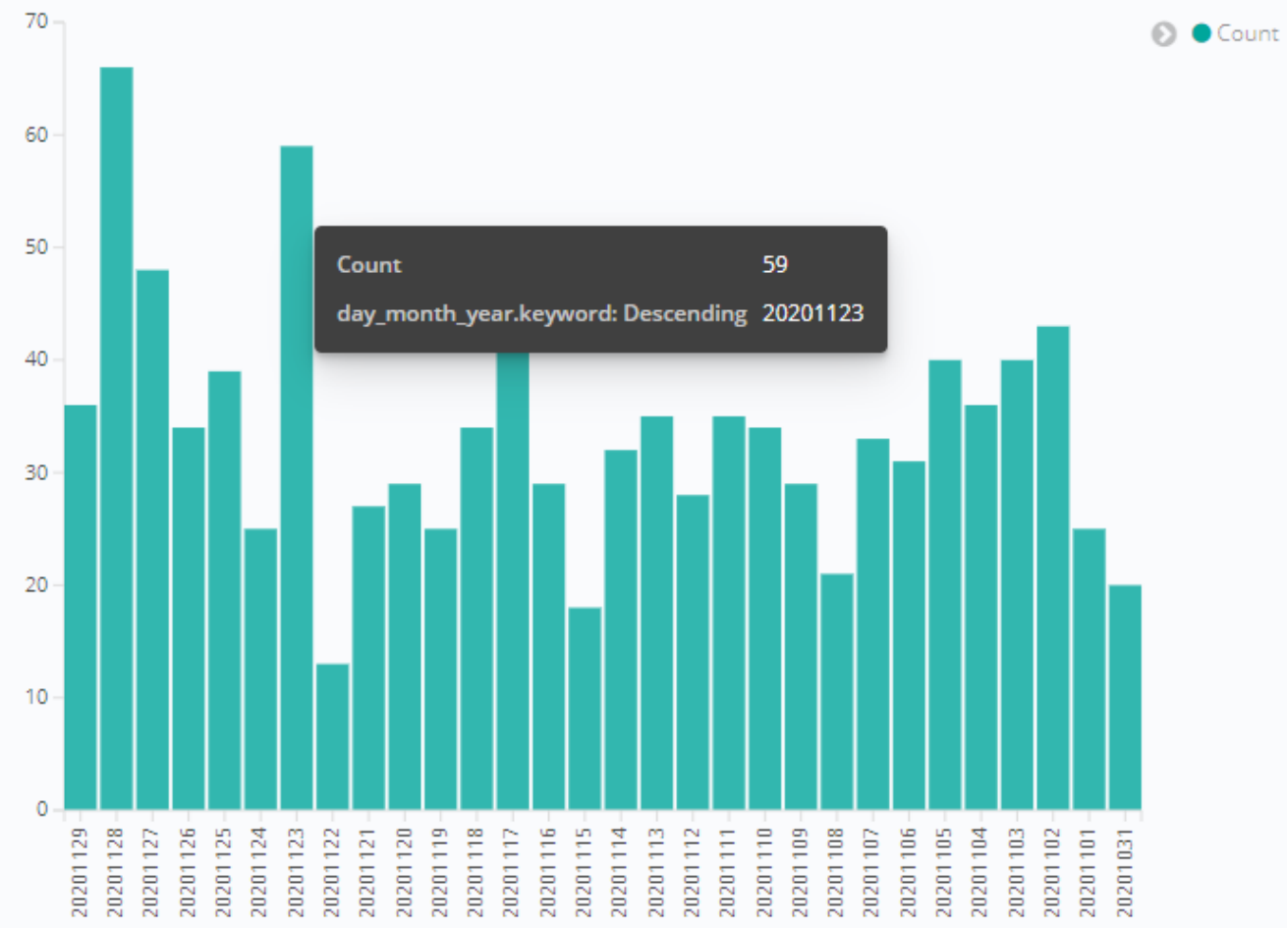
Biểu diễn dữ liệu bằng Kibana



Phân phối dữ liệu trang soha.vn theo tháng đăng

4. Quản lý dữ liệu với ELK stack (tiếp)

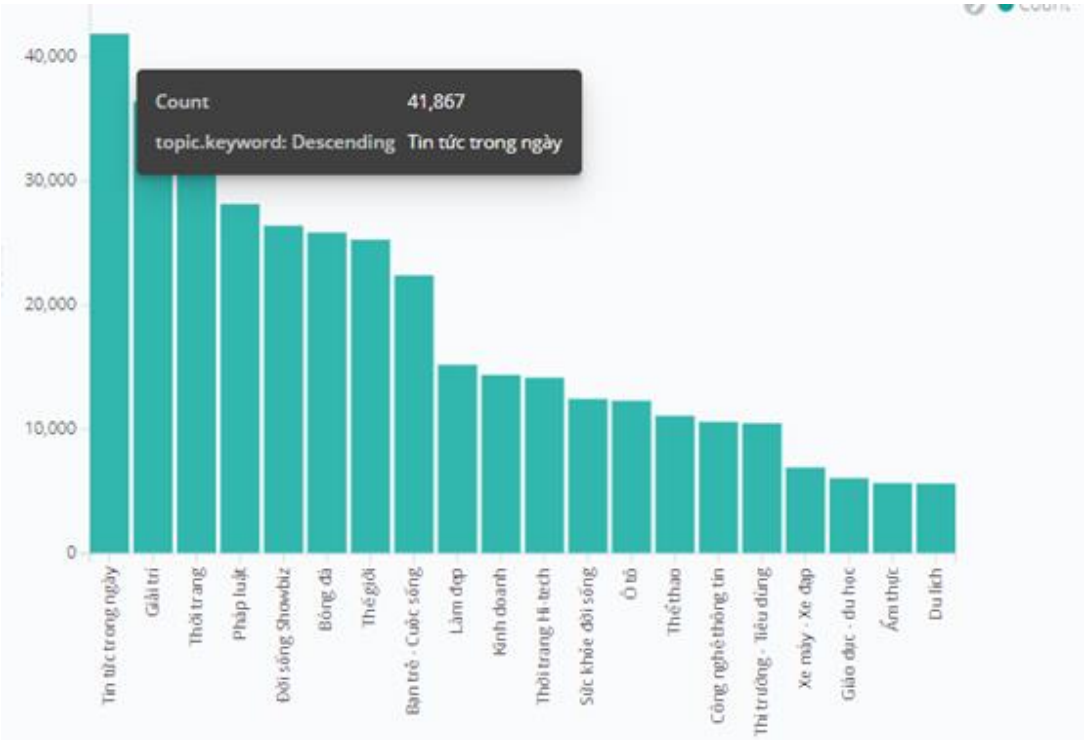
Biểu diễn dữ liệu bằng Kibana



Phân phối dữ liệu trang soha.vn theo ngày đăng trong tháng

5. Phân loại văn bản với Spark MLlib

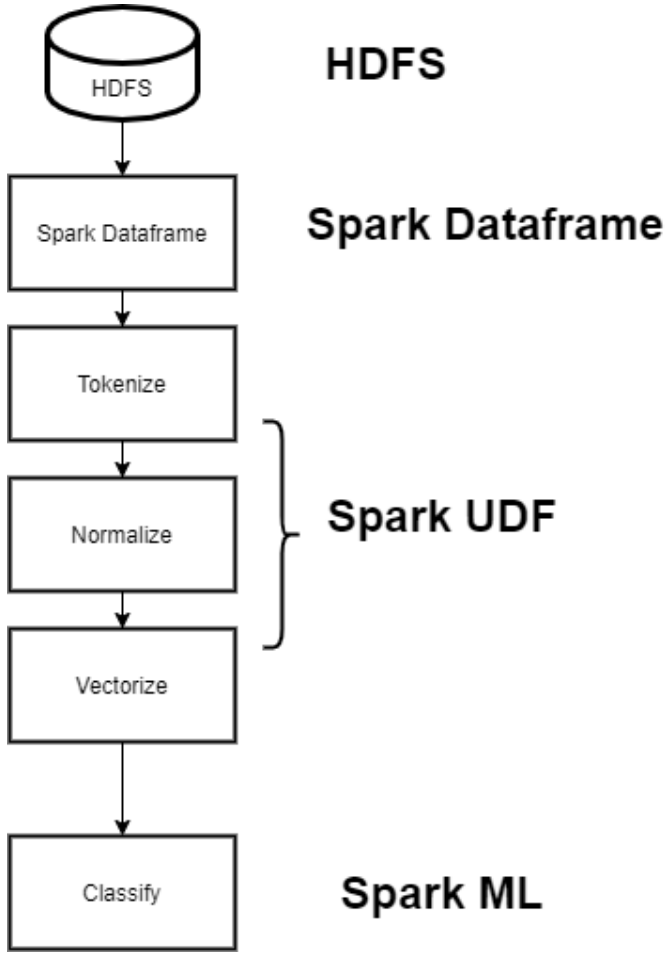
- Dữ liệu: Dữ liệu đã thu thập được từ trang 24h.com
- Bài toán: Phân loại bài báo theo chủ đề (thể loại)
  - Đầu vào: Nội dung của bài báo
  - Đầu ra: Chủ đề của bài báo



Thống kê phân bố dữ liệu theo thể loại

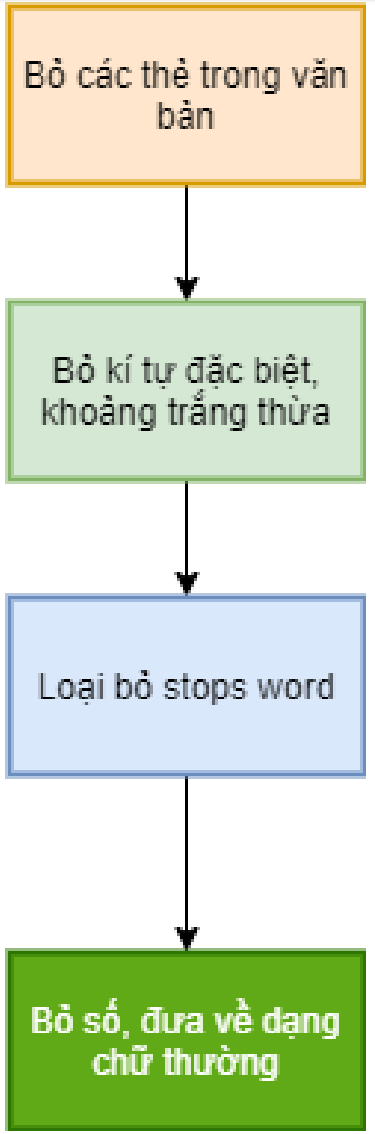
5.1. Xử lí và phân loại dữ liệu

- Luồng xử lí và phân loại văn bản
  - Chuẩn bị dữ liệu
  - Xây dựng mô hình phân loại văn bản



5.2. Chuẩn bị dữ liệu

- Tách từ
  - Công nghệ sử dụng: sử dụng thư viện Pyvi để tách từ
- Chuẩn hóa câu
  - Bỏ các thẻ trong văn bản
  - Loại bỏ các kí tự đặc biệt, khoảng trắng thừa
  - Bỏ các chữ số, chuyển chữ hoa về chữ thường
  - Loại bỏ stopword: sử dụng StopWordsRemover của Spark ML Feature
- Vector hóa
  - Phương pháp: sử dụng tf-idf để mã hóa văn bản
  - Công nghệ: sử dụng CountVectorizer trong Spark ML Feature
- Phân chia tập dữ liệu
  - Phân chia dữ liệu huấn luyện và dữ liệu đánh giá theo tỉ lệ 7-3
    - Dữ liệu huấn luyện: 160001 văn bản
    - Dữ liệu đánh giá: 69007



5.3. Mô hình phân loại văn bản

- Phương pháp:
  - Sử dụng mô hình phân loại xác suất cơ bản Naive Bayes
- Công nghệ:
  - Xây dựng mô hình phân loại bằng NaiveBayes trong SparkML Classification
  - Đánh giá mô hình bằng MulticlassClassificationEvaluator với độ đo accuracy trong SparkML evaluation

5.4. Huấn luyện mô hình phân loại văn bản

Job huấn luyện mô hình phân loại văn bản trên Spark

Spark Jobs (?)

User: deserteagle735
Total Uptime: 51 min
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 7

▸ Event Timeline

▾ **Active Jobs (1)**

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	collectAsMap at MulticlassMetrics.scala:61 collectAsMap at MulticlassMetrics.scala:61 (kill)	2020/12/21 21:08:38	1.2 min	0/2	0/14 (4 running)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

▾ **Completed Jobs (7)**

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
6	collect at NaiveBayes.scala:192 collect at NaiveBayes.scala:192	2020/12/21 20:33:17	35 min	2/2	207/207
5	collect at StringIndexer.scala:204 collect at StringIndexer.scala:204	2020/12/21 20:32:52	24 s	2/2	8/8
4	top at CountVectorizer.scala:236 top at CountVectorizer.scala:236	2020/12/21 20:32:51	0.3 s	1/1 (1 skipped)	7/7 (7 skipped)
3	count at CountVectorizer.scala:233 count at CountVectorizer.scala:233	2020/12/21 20:20:44	12 min	2/2	14/14

5.4. Huấn luyện mô hình phân loại văn bản (tiếp)

Các job đã chạy xong

URL: spark://192.168.0.111:7077
Alive Workers: 2
Cores in use: 8 Total, 0 Used
Memory in use: 11.3 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 3 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20201221200300-192.168.0.111-40751	192.168.0.111:40751	ALIVE	4 (0 Used)	6.6 GiB (0.0 B Used)	
worker-20201221200328-192.168.0.113-42885	192.168.0.113:42885	ALIVE	4 (0 Used)	4.6 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (3)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20201221201919-0002	Zootopia-ML	8	2.0 GiB		2020/12/21 20:19:19	deserteagle735	FINISHED	1.4 h
app-20201221201444-0001	Zootopia-clean-data	8	2.0 GiB		2020/12/21 20:14:44	deserteagle735	FINISHED	1.5 min
app-20201221200332-0000	Zootopia-clean-data	4	2.0 GiB		2020/12/21 20:03:32	deserteagle735	FAILED	0.4 s

5.5. Đánh giá mô hình và quá trình xử lí

- Thời gian chuẩn bị dữ liệu: 40 phút
- Thời gian huấn luyện: 35 phút
- Thời gian kiểm thử: 9 phút
- Chất lượng mô hình: 73.84%

6. Tổng kết

- Các kết quả đã đạt được
  - Xây dựng và cài đặt hệ thống thu thập, xử lí và lưu trữ dữ liệu
    - Lưu trữ và xử lí phân tán.
    - Đảm bảo một số tính chất của hệ thống dữ liệu lớn: có khả năng mở rộng, lưu trữ đảm bảo sẵn sàng dữ liệu, có khả năng biểu diễn dữ liệu.
    - Sử dụng dữ liệu để xây dựng mô hinh phân loại văn bản.
- Hướng phát triển
  - Thu thập dữ liệu thời gian thực, ứng dụng Spark Streaming và Kafka.
  - Sử dụng đầu vào của Logstash là file từ HDFS thay vì file local.

Cảm ơn thầy và các bạn đã lắng nghe