

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG  
—o0o—



*Báo cáo bài tập lớn*  
*Lưu trữ và xử lý dữ liệu lớn*  
Phân tích dữ liệu bất động sản

Giáo viên hướng dẫn: PGS.TS.Nguyễn Bình Minh

Sinh viên thực hiện : Lương Cường Thịnh - 20183993  
Nguyễn Hữu Hiệp - 20193528  
Nguyễn Văn Duy - 20183514  
Vương Xuân Hoàng - 20183545

Lớp : CTTN CNTT K63

Hà Nội - 2021

# Mục lục

<b>1</b>	<b>Giới thiệu</b>	<b>1</b>
<b>2</b>	<b>Tổng quan về kiến trúc</b>	<b>1</b>
<b>3</b>	<b>Crawl và tiền xử lý dữ liệu</b>	<b>2</b>
3.1	Crawl data . . . . .	2
3.2	Tiền xử lý dữ liệu . . . . .	3
<b>4</b>	<b>Lưu trữ dữ liệu</b>	<b>4</b>
4.1	Cài đặt và cấu hình HDFS . . . . .	4
4.2	Cài đặt và cấu hình Elastic Search + Kibana . . . . .	5
<b>5</b>	<b>Phân tích dữ liệu</b>	<b>6</b>
5.1	Dựa theo phân tích trên Kibana . . . . .	6
5.2	Dựa theo phân tích trên Spark . . . . .	9
<b>6</b>	<b>Một số công cụ Spark nâng cao</b>	<b>13</b>
6.1	Áp dụng Structured Streaming xử lý dữ liệu stream . . . . .	13
6.2	Áp dụng MLlib vào dự đoán giá nhà . . . . .	13
<b>7</b>	<b>Kết luận</b>	<b>15</b>

## Danh sách hình vẽ

1	Kiến trúc của hệ thống . . . . .	1
2	Một số đặc trưng của dữ liệu . . . . .	2
3	Các trường dữ liệu thô . . . . .	3
4	Các trường dữ liệu sau khi tiền xử lý . . . . .	4
5	Hình ảnh dữ liệu trước và sau khi xử lý . . . . .	4
6	Cấu hình cơ bản của cụm HDFS . . . . .	5
7	Thông tin datanode . . . . .	5
8	Elastic Search + Kibana . . . . .	6
9	Thống kê chung về các giao dịch . . . . .	6
10	Thống kê về chiều dài mặt tiền và đường vào . . . . .	7
11	Thống kê về hướng ban công và hướng nhà . . . . .	7
12	Thống kê về số tầng, số phòng ngủ và số toilet . . . . .	8
13	Thống kê về loại tin được đăng . . . . .	9
14	Thống kê chung về các tỉnh . . . . .	10
15	Thống kê về giá đất trung bình trên $m^2$ của 6 tỉnh có diện tích trung bình lớn nhất . . . . .	10
16	Thống kê chi tiết về từng tỉnh được sắp xếp theo giá trị trung bình trên $m^2$ giảm dần . . . . .	11
17	Thống kê chi tiết về từng quận ở Hà Nội được sắp xếp theo giá trị trung bình trên $m^2$ giảm dần . . . . .	11
18	Thống kê về 5 quận của Hà Nội có giá trị trung bình trên $m^2$ lớn nhất . . . . .	11
19	Thống kê về các chủ đầu tư sắp xếp theo số dự án tham gia giảm dần . . . . .	12
20	Một số phân tích từ dữ liệu streaming . . . . .	13
21	Visualize phân bố dữ liệu . . . . .	14
22	Mô hình random forest . . . . .	14
23	Mô hình linear regression . . . . .	14

# Ứng dụng các công nghệ Hadoop, Spark, NoSQL vào phân tích dữ liệu bất động sản

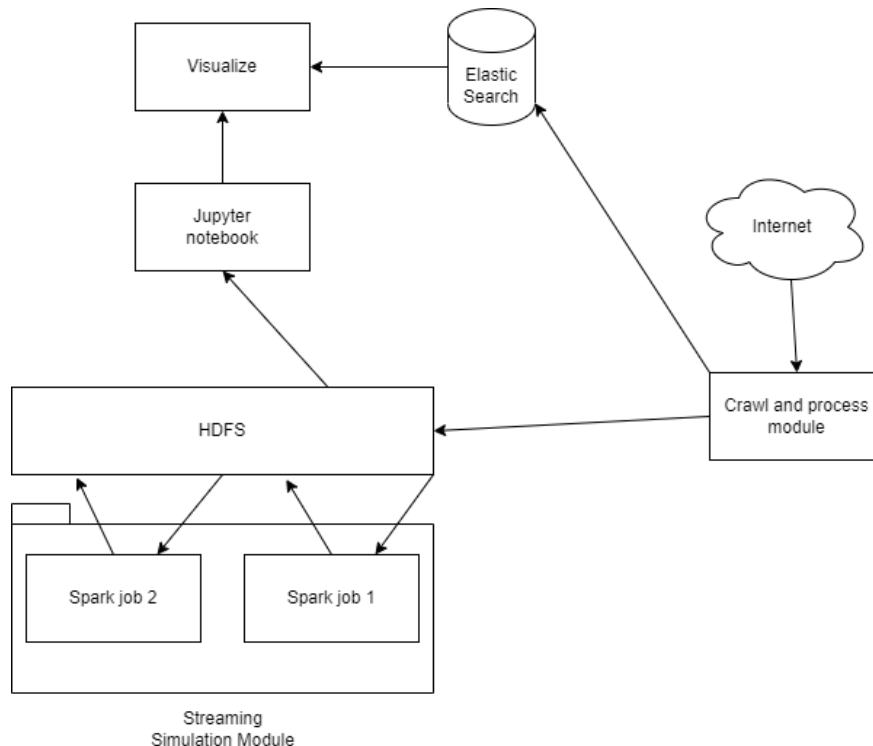
**Tóm tắt:** Project nhằm mục tiêu ứng dụng các công nghệ đã học vào bài toán phân tích dữ liệu bất động sản để có cái nhìn tổng quan về quy trình hoàn thiện một bài toán xử lý dữ liệu lớn và làm quen với các thư viện, công cụ xử lý dữ liệu lớn đã được học. Mã nguồn được lưu trữ trong link github: <https://github.com/vuonghoangbntt/BigDataProject>

## 1 Giới thiệu

Ngày nay, bất động sản đang là một tài sản có giá trị cao, tăng trưởng nhanh, giá cả biến động theo từng ngày, từng giờ. Kinh doanh, môi giới bất động sản cũng trở thành một ngành nghề có thu nhập cao trong xã hội. Với sự phát triển của internet, ngày nay các bài đăng về buôn bán bất động sản cũng xuất hiện ngày một nhiều trên internet. Các website về bất động sản cũng mọc lên ngày một nhiều mà nổi bật nhất trong đó là [batdongsan.com.vn](http://batdongsan.com.vn) với khoảng từ 100-500 bài đăng mới mỗi ngày. Từ đó, chỉ cần ở nhà chúng ta cũng có thể nắm bắt được các thông tin bất động sản trên cả nước. Với sự sôi động của thị trường bất động sản, dữ liệu bất động sản cũng trở nên ngày càng lớn, từ đó đặt ra bài toán phân tích, khai thác nguồn dữ liệu này

Trong bài tập lớn này, chúng em xin trình bày một quy trình thu thập, xử lý và phân tích dữ liệu bất động sản từ trang [batdongsan.com.vn](http://batdongsan.com.vn). Mục tiêu của project là nhằm hiểu rõ hơn về các quá trình thu thập dữ liệu, xử lý dữ liệu, phân tích dữ liệu và visualize dữ liệu sử dụng các công cụ đã được học trong môn Lưu trữ và xử lý dữ liệu lớn đồng thời đưa ra được một số thông tin hữu ích từ dữ liệu

## 2 Tổng quan về kiến trúc



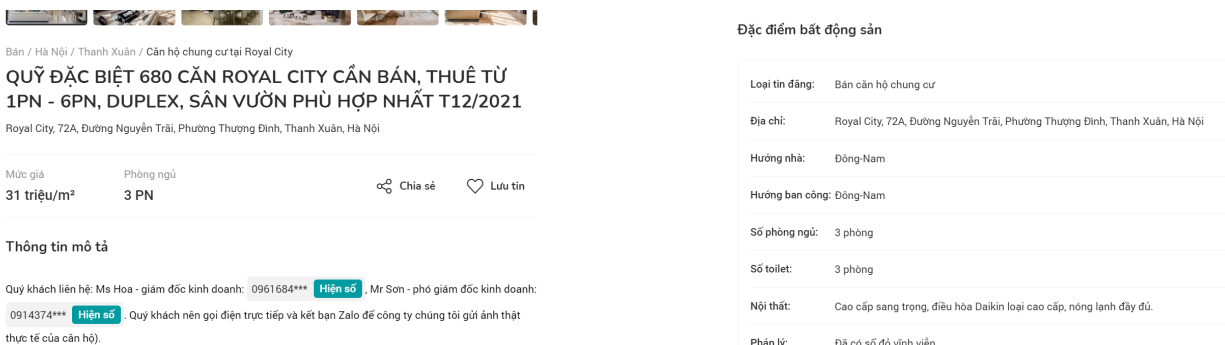
Hình 1: Kiến trúc của hệ thống

## \* Mô tả:

- Đầu tiên, dữ liệu từ trang [batdongsan.com.vn](http://batdongsan.com.vn) sẽ được crawl và tiền xử lý bởi python module. Dữ liệu xử lý và lưu dưới dạng file csv (kích thước dữ liệu vào khoảng 200Mb) trên hdfs để xử lý bằng spark. Ngoài ra, để phục vụ cho quá trình visualize data, dữ liệu sẽ được lưu trữ trên elastic search
- Dữ liệu trên elastic search được visualize thông qua Kibana UI và được tổng hợp lại thành một dashboard chứa một số thông tin cơ bản về dữ liệu
- Dữ liệu lưu trữ trên hdfs sẽ được xử lý bằng spark để phân tích các đặc trưng phức tạp, tận dụng khả năng xử lý phân tán để xử lý dữ liệu nhanh đồng thời đưa ra một số visualize trên các đặc trưng đã phân tích
- Cuối cùng, nhóm sử dụng một module để mô phỏng quá trình streaming dữ liệu trong thực tế. Quá trình mô phỏng thực hiện bởi 2 spark job, spark job 1 làm nhiệm vụ ghi dữ liệu vào thư mục trên hdfs và spark job 2 làm nhiệm vụ đọc dữ liệu streaming, xử lý và lưu trữ lại trên hdfs.

## 3 Crawl và tiền xử lý dữ liệu

### 3.1 Crawl data



Bán / Hà Nội / Thanh Xuân / Căn hộ chung cư tại Royal City

**QUỸ ĐẶC BIỆT 680 CĂN ROYAL CITY CẦN BÁN, THUÊ TỪ 1PN - 6PN, DUPLEX, SÂN VƯỜN PHÙ HỢP NHẤT T12/2021**

Royal City, 72A, Đường Nguyễn Trãi, Phường Thượng Đình, Thanh Xuân, Hà Nội

Mức giá: **31 triệu/m<sup>2</sup>** Phòng ngủ: **3 PN** Chia sẻ Lưu tin

**Thông tin mô tả**

Quý khách liên hệ: Ms Hoa - giám đốc kinh doanh: 0961684\*\*\* **Hiện số** Mr Sơn - phó giám đốc kinh doanh: 0914374\*\*\* **Hiện số** Quý khách nên gọi điện trực tiếp và kết bạn Zalo để công ty chúng tôi gửi ảnh thật thực tế của căn hộ).

**Đặc điểm bất động sản**

Loại tin đăng:	Bán căn hộ chung cư
Địa chỉ:	Royal City, 72A, Đường Nguyễn Trãi, Phường Thượng Đình, Thanh Xuân, Hà Nội
Hướng nhà:	Đông-Nam
Hướng ban công:	Đông-Nam
Số phòng ngủ:	3 phòng
Số toilet:	3 phòng
Nội thất:	Cao cấp sang trọng, điều hòa Daikin loại cao cấp, nóng lạnh đầy đủ.
Pháp lý:	Đã có sổ đỏ vĩnh viễn.

Hình 2: Một số đặc trưng của dữ liệu

Dữ liệu từ trang [batdongsan.com.vn](http://batdongsan.com.vn) được nhóm bắt đầu thực hiện crawl từ ngày 06/11/2021 với hơn 200000 samples tương ứng với 200000 link url. Quá trình crawl có gặp phải một số khó khăn:

- Khó khăn về vấn đề bảo mật: Trang [batdongsan.com.vn](http://batdongsan.com.vn) được bảo mật bởi hệ thống cloud flare không cho phép các công cụ thường dùng như: request, selenium,... có thể truy cập và lấy dữ liệu  
-> Vấn đề đã được xử lý với thư viện *cloudscraper*
- Khó khăn về vấn đề tài nguyên: Với thời gian trung bình cho một bản ghi là 4s, thời gian để crawl dữ liệu ước tính khoảng 223 tiếng nếu sử dụng một máy tính cá nhân  
-> Vấn đề được giải quyết bằng việc thuê một số máy ảo của Azure

Thời gian crawl một bản ghi trung bình mất khoảng 4s được thực hiện bằng 4 virtual machine trên *Microsoft Azure* trong khoảng thời gian là 2 ngày.

Mỗi bản ghi dữ liệu có chứa một số trường dữ liệu quan trọng như: 'Tên', 'Mô tả', 'Mức giá', 'Diện tích', 'Loại tin đăng', 'Địa chỉ', 'Mặt tiền', 'Đường vào', 'Hướng ban công', 'Số tầng', 'Số phòng ngủ', 'Số toilet', 'Nội thất', 'Pháp lý', 'Tên dự án', 'Chủ đầu tư', 'Quy mô', 'Ngày đăng', 'Ngày hết hạn', 'Mã tin', 'Phòng ngủ', 'Hướng nhà', 'Loại tin'. Tùy vào từng bản ghi mà có thể có các trường có giá trị null. => Thông tin dữ liệu thô còn cần một số phương pháp tiền xử lý để tách lọc dữ liệu

---	-----	-----	-----
0	Tên	72059 non-null	object
1	Mô tả	19742 non-null	object
2	Mức giá	9968 non-null	object
3	Diện tích	9810 non-null	object
4	Loại tin đăng	9968 non-null	object
5	Địa chỉ	9967 non-null	object
6	Mặt tiền	4649 non-null	object
7	Đường vào	4377 non-null	object
8	Hướng ban công	1472 non-null	object
9	Số tầng	3446 non-null	object
10	Số phòng ngủ	4768 non-null	object
11	Số toilet	4099 non-null	object
12	Nội thất	2648 non-null	object
13	Pháp lý	6091 non-null	object
14	Tên dự án	4087 non-null	object
15	Chủ đầu tư	3325 non-null	object
16	Quy mô	2880 non-null	object
17	Ngày đăng	9968 non-null	object
18	Ngày hết hạn	9968 non-null	object
19	Mã tin	9966 non-null	float64
20	Phòng ngủ	4769 non-null	object
21	Hướng nhà	3129 non-null	object
22	Loại tin	9968 non-null	object
23	url	9966 non-null	object

Hình 3: Các trường dữ liệu thô

### 3.2 Tiền xử lý dữ liệu

\* Một số bước tiền xử lý dữ liệu như sau:

- Tách thông tin dạng integer từ trường: *'Số tầng', 'Số phòng ngủ', 'Số toilet'*
- Tách thông tin và đưa về cùng đơn vị đo là m hoặc m2 đối với các trường: *'Diện tích', 'Mặt tiền', 'Quy mô'*
- Tách thông tin về giá và đưa về cùng đơn vị đo là **triệu** đối với trường *'Mức giá'* và tạo thêm trường *'Mức giá/m2'*
- Tách thông tin trong trường *'Địa chỉ'* về các trường đơn vị: *'Tỉnh', 'Quận/Huyện'*

*Sau khi đã được tiền xử lý, dữ liệu sẽ được lưu trữ trên hdfs và Elastic Search*

#	Column	Non-Null Count	Dtype
0	Tên	182320 non-null	object
1	Mô tả	182320 non-null	object
2	Mức giá	156887 non-null	float64
3	Diện tích	180251 non-null	float64
4	Loại tin đăng	182320 non-null	object
5	Địa chỉ	182320 non-null	object
6	Mặt tiền	84061 non-null	float64
7	Đường vào	79731 non-null	float64
8	Hướng ban công	24338 non-null	object
9	Số tầng	182320 non-null	int64
10	Số phòng ngủ	182320 non-null	int64
11	Số toilet	182320 non-null	int64
12	Nội thất	43763 non-null	object
13	Pháp lý	103416 non-null	object
14	Tên dự án	66543 non-null	object
15	Chủ đầu tư	53610 non-null	object
16	Quy mô	49047 non-null	object
17	Ngày đăng	182320 non-null	object
18	Ngày hết hạn	182320 non-null	object
19	Mã tin	182320 non-null	int64
20	Phòng ngủ	86037 non-null	object
21	Hướng nhà	56224 non-null	object
22	Loại tin	182320 non-null	object

Hình 4: Các trường dữ liệu sau khi tiền xử lý

Tên	Báo giá CH Hà Đô mùa Covid, 2PN 5.5 tỷ, 2PN + ...	Tên	Chính chủ cần bán gấp lô đất ngay gần chợ Sóng...
Mô tả	Cập nhật bảng giá tốt nhất căn hộ Hà Đô Q10!- ...	Mô tả	- Ngay chợ Sóng Trầu.- Dân cư đang phát triển...
Mức giá	4.65 tỷ	Mức giá	860
Diện tích	61 m²	Diện tích	90
Loại tin đăng	Bán căn hộ chung cư	Loại tin đăng	Bán đất
Địa chỉ	Dự án HaDo Centrosa Garden, Đường 3/2, Phường ...	Địa chỉ	Đường Sóng Trầu, Xã Sóng Trầu, Trảng Bom, Đồng...
Mặt tiền	NaN	Mặt tiền	NaN
Đường vào	NaN	Đường vào	8
Hướng ban công	NaN	Hướng ban công	NaN
Số tầng	NaN	Số tầng	-1
Số phòng ngủ	1 phòng	Số phòng ngủ	-1
Số toilet	1 phòng	Số toilet	-1
Nội thất	Đầy đủ nội thất cao cấp.	Nội thất	NaN
Pháp lý	NaN	Pháp lý	Số đó/ Số hồng
Tên dự án	HaDo Centrosa Garden	Tên dự án	NaN
Chủ đầu tư	Tập đoàn Hà Đô	Chủ đầu tư	NaN
Quy mô	NaN	Quy mô	NaN
Ngày đăng	08/11/2021	Ngày đăng	06/11/2021
Ngày hết hạn	18/11/2021	Ngày hết hạn	13/11/2021
Mã tin	3.11372e+07	Mã tin	31319056
Phòng ngủ	1 PN	Phòng ngủ	NaN
Hướng nhà	NaN	Hướng nhà	NaN
Loại tin	Tin VIP đặc biệt	Loại tin	Tin thường
url	<a href="https://batdongsan.com.vn/nha-dat-ban/ban-can...">https://batdongsan.com.vn/nha-dat-ban/ban-can...</a>		

Hình 5: Hình ảnh dữ liệu trước và sau khi xử lý

## 4 Lưu trữ dữ liệu

Dữ liệu sau khi xử lý sẽ được lưu trữ trên HDFS để tiến hành xử lý, phân tích bằng Spark. Ngoài ra, dữ liệu cũng được đẩy lên Elastic Search nhằm tạo ra các visualize cơ bản

### 4.1 Cài đặt và cấu hình HDFS

HDFS bao gồm có 1 namenode và 2 datanode. Namenode và các datanode được triển khai bằng dockere. Cấu hình cơ bản của một cụm HDFS như sau: hadoop 3.2.1, Block size có kích thước là 10Mb, số lượng replica là 2,... Các cấu hình khác xem thêm ở trong hình bên dưới. Chi tiết về file docker-compose và file cấu hình hadoop.env xem thêm trong link github của nhóm.

Safemode is off.

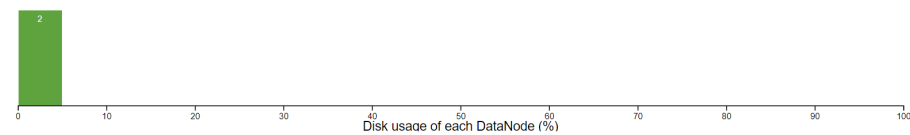
26,348 files and directories, 25,920 blocks (25,920 replicated blocks, 0 erasure coded block groups) = 52,268 total filesystem object(s).

Heap Memory used 129.64 MB of 407 MB Heap Memory. Max Heap Memory is 3.48 GB.

Non Heap Memory used 51.29 MB of 52.56 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	57.81 GB
Configured Remote Capacity:	0 B
DFS Used:	624.88 MB (1.06%)
Non DFS Used:	45.08 GB
DFS Remaining:	12.09 GB (20.91%)
Block Pool Used:	624.88 MB (1.06%)
DataNodes usages% (Min/Median/Max/stdDev):	1.06% / 1.06% / 1.06% / 0.00%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)

Hình 6: Cấu hình cơ bản của cụm HDFS



n operation

show

25

entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
<div><div>✓</div><div>32a67360c270:9866</div><div>(172.19.0.7:9866)</div></div>	<a href="http://32a67360c270:9866">http://32a67360c270:9866</a>	0s	34m	28.91 GB <div><div></div></div>	25920	312.44 MB (1.06%)	3.2.1
<div><div>✓</div><div>81b370276d63:9866</div><div>(172.19.0.8:9866)</div></div>	<a href="http://81b370276d63:9866">http://81b370276d63:9866</a>	0s	34m	28.91 GB <div><div></div></div>	25920	312.47 MB (1.06%)	3.2.1

showing 1 to 2 of 2 entries

Previous

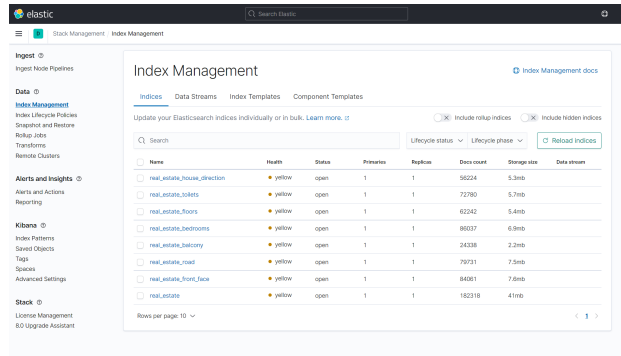
1

Next

Hình 7: Thông tin datanode

## 4.2 Cài đặt và cấu hình Elastic Search + Kibana

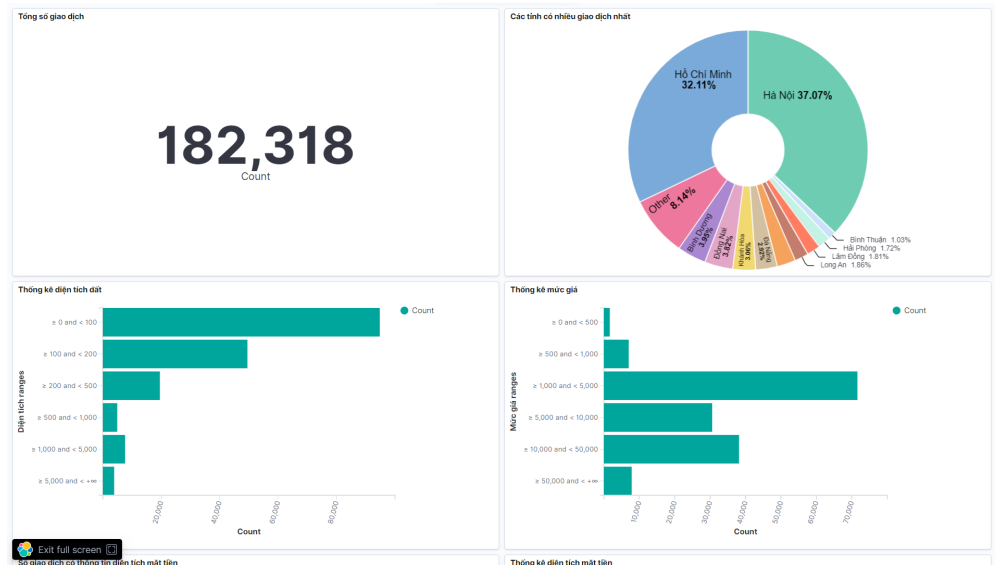
Để phục vụ cho việc visualize dữ liệu, nhóm sử dụng Elastic Search+Kibana. Nhóm triển khai Elastic Search + Kibana version 7.11 bằng docker. Dữ liệu lưu trữ trên Elastic Search đã được bỏ đi một số trường không mang lại ý nghĩa khi visualize đó là: *'Mô tả', 'Địa chỉ', 'Tên', 'url'*



Hình 8: Elastic Search + Kibana

## 5 Phân tích dữ liệu

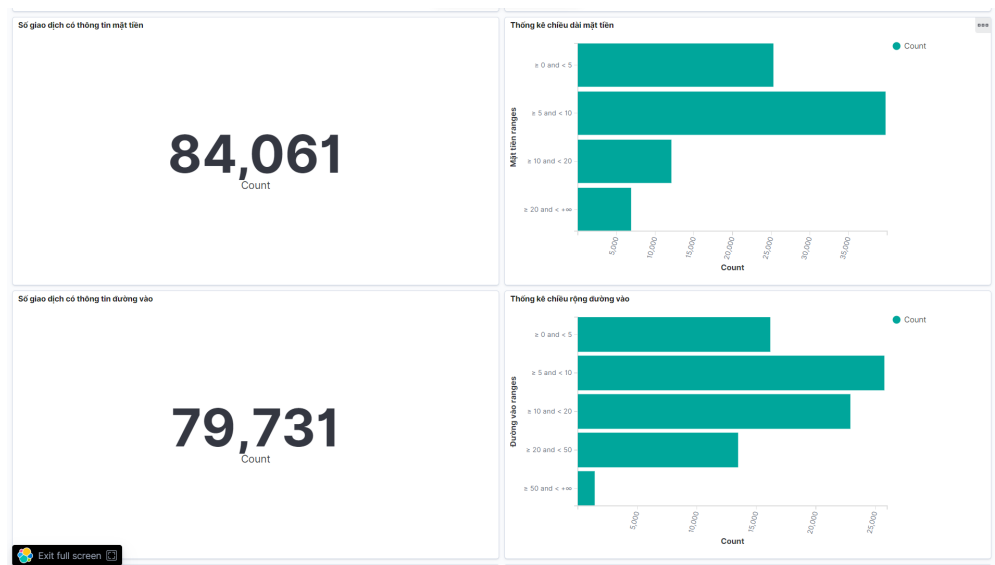
### 5.1 Dựa theo phân tích trên Kibana



Hình 9: Thống kê chung về các giao dịch

Từ bảng thống kê ta có thể thấy được có tổng cộng 182,318 giao dịch được trích xuất từ dữ liệu đã crawl, trong đó 2 tỉnh là Hà Nội và Thành phố Hồ Chí Minh chiếm tỉ lệ giao dịch mua bán nhà đất vượt trội so với các tỉnh còn lại. Diện tích đất được giao bán nhiều nhất là trong 2 khoảng từ 0-100  $m^2$  và từ 100-200  $m^2$  cho thấy đa số các giao dịch là mua bán đất để xây(hoặc mua bán) nhà ở hoặc các căn biệt thự, các căn hộ cao cấp. Và để củng cố cho luận điểm vừa nêu ra, mức giá bán trong các giao dịch cũng nằm chủ yếu trong các khoảng từ 1-5 tỉ đồng, 5-10 tỉ đồng và 10-50 tỉ đồng.



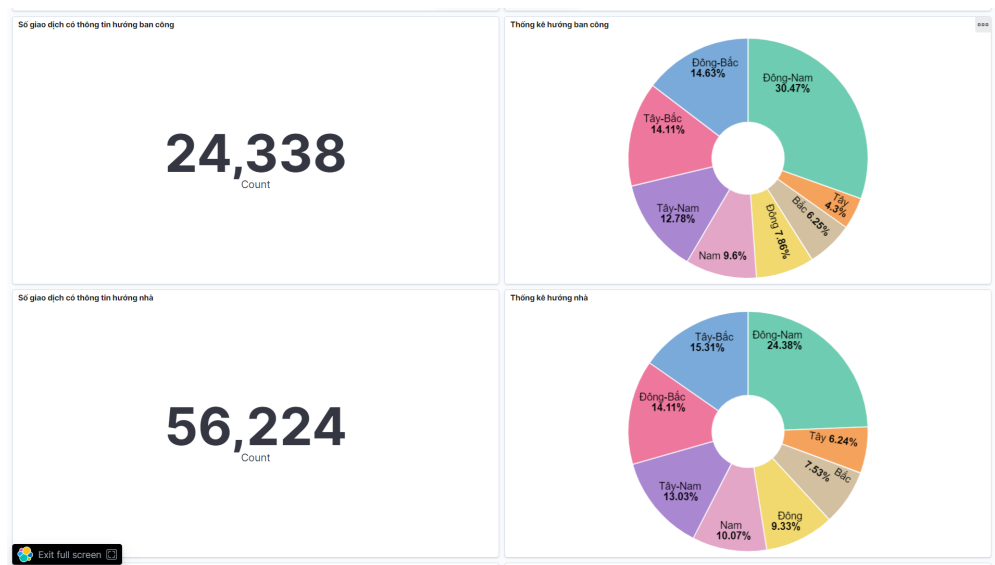


Hình 10: Thống kê về chiều dài mặt tiền và đường vào

Dù theo thống kê bên trên về diện tích và mức giá, số lượng giao dịch đất dùng để xây(hoặc mua bán) nhà ở, biệt thự, căn hộ cao cấp chiếm tới hơn 1 nửa số giao dịch thì bên dưới tổng số giao dịch có thông tin về mặt tiền hay đường vào thì lại chưa chiếm tới một nửa tổng số giao dịch cho thấy trong số giao dịch đất để xây(hoặc mua bán) vừa nhắc tới thì phần lớn dùng cho nhà ở và biệt thự (bởi các căn hộ sẽ không có 2 thông tin về mặt tiền hay đường vào).

Từ thống kê chiều dài mặt tiền, ta có thể nhận thấy các căn nhà được giao bán chủ yếu dành cho các gia đình từ tầng lớp vừa vừa cho tới trung lưu: 0-5m có 25,289 giao dịch và 5-10m có 39,789 giao dịch (2 khoảng chiều dài mặt tiền này chiếm tới hơn 75% số giao dịch có thông tin về mặt tiền)

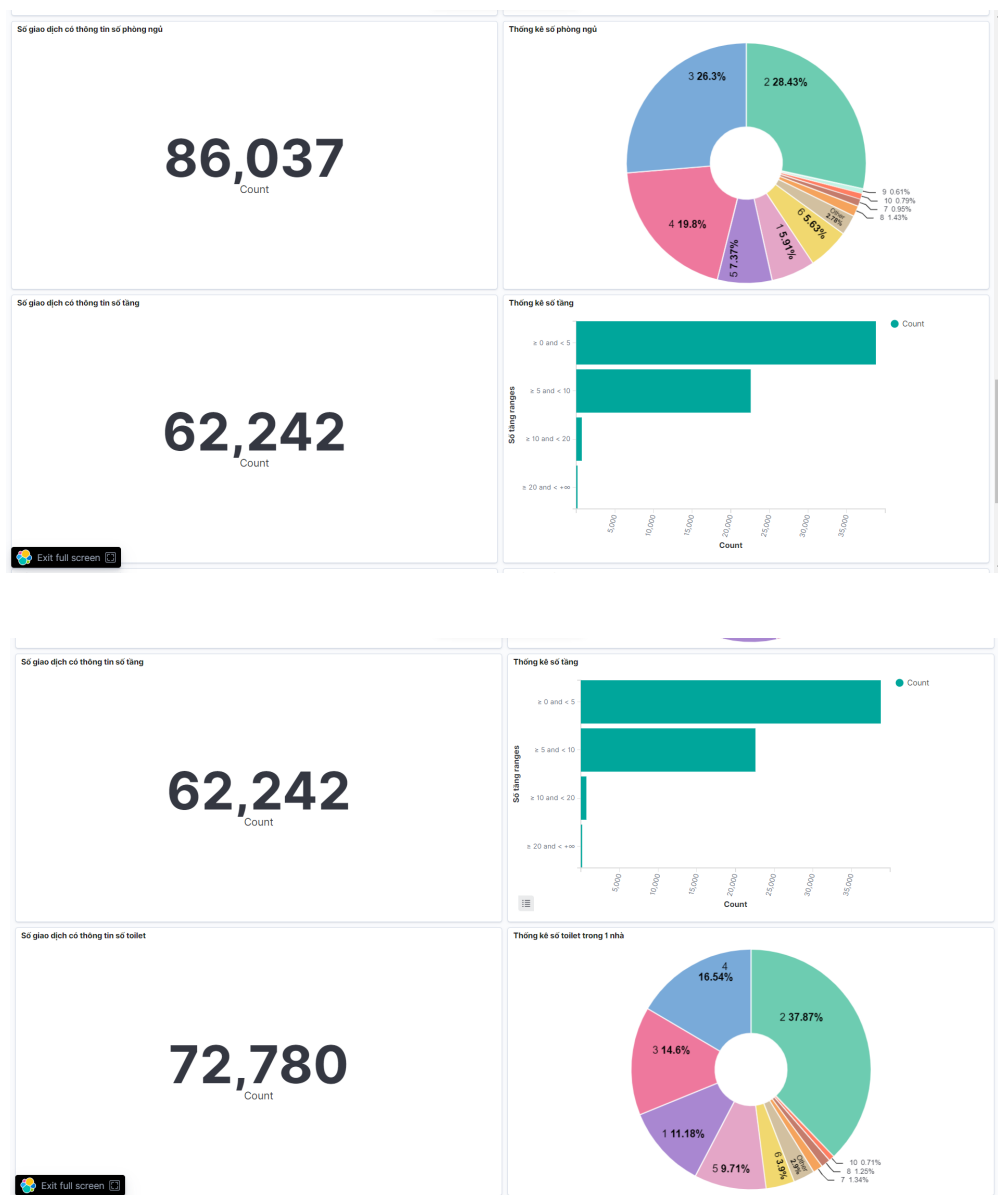
Về chiều rộng đường vào, các khoảng có sự chênh lệch không nhiều, 2 khoảng chiếm tỉ lệ lớn nhất là 5-10m và 10-20m cho thấy việc các căn nhà được giao bán đều nằm ở chỗ thoáng đàng, dễ dàng cho việc đi lại, ra vào nhà.



Hình 11: Thống kê về hướng ban công và hướng nhà

Thông qua bảng thống kê, ta có thể thấy số giao dịch có thông tin về hướng nhà nhiều hơn gấp đôi số

thông tin về hướng ban công. Đây cũng là một điều dễ hiểu khi theo phong thủy từ xưa tới nay, hướng nhà thường được coi là hướng quyết định đến vận may, tiền tài của gia đình. Còn trong những năm gần đây, hướng ban công cũng được cân nhắc khi xét đến các yếu tố phong thủy cũng có lẽ bởi các căn chung cư thường chỉ có hướng ban công nên cần có yếu tố phong thủy thích hợp khi xem xét mua các căn chung cư. Hơn thế nữa, số giao dịch có thông tin về hướng nhà cũng ít hơn số giao dịch có thông tin về mặt tiền hoặc đường vào cũng gợi ý rằng nhiều giao dịch có thông tin về mặt tiền hay đường vào cũng có thể chỉ là những dự án bán đất, chưa được xây dựng ở thửa đất đầy hoặc yếu tố phong thủy như hướng nhà hay hướng ban công không được người bán đưa vào vì yếu tố đó không thuận lợi, có thể khiến họ mất khách hàng tiềm năng.



Hình 12: Thống kê về số tầng, số phòng ngủ và số toilet

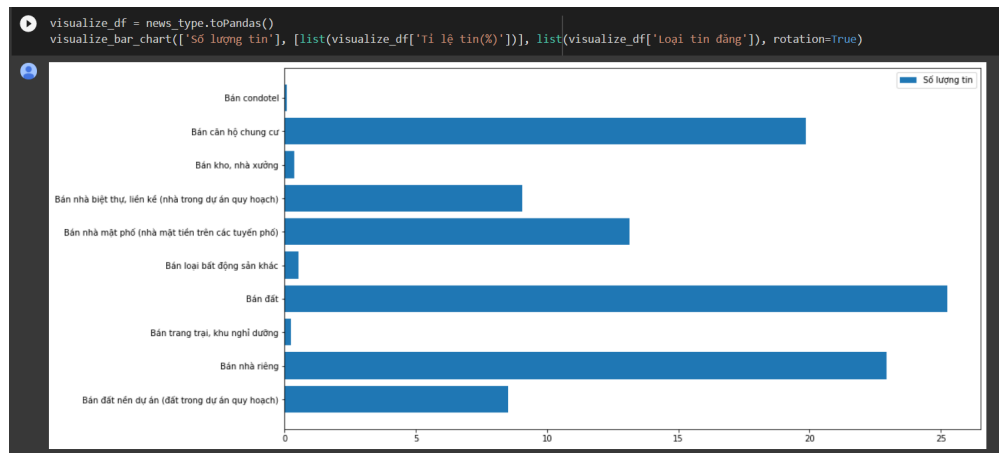
Thông tin về số tầng, số phòng ngủ và số toilet là những thông tin vô cùng cần thiết khi cân nhắc mua một ngôi nhà nên ta có thể thấy số giao dịch có các thông tin này có số lượng cũng gần tương đương số lượng số giao dịch có thông tin về mặt tiền hay đường vào.

Từ thông tin về số phòng ngủ và số toilet, ta thấy rằng đa số các căn nhà được giao bán, mỗi phòng ngủ sẽ

có 1 toilet riêng. Đồng thời, việc số lượng phòng ngủ và số lượng toilet chủ yếu nằm khoảng từ 1-5 cũng cho thấy số lượng thành viên trong gia đình muốn mua căn nhà cũng có từ 1-5 người.

## 5.2 Dựa theo phân tích trên Spark

Loại tin đăng	Số lượng tin	Tỉ lệ tin(%)
Bán đất	46048	25.26
Bán nhà riêng	41841	22.95
Bán căn hộ chung cư	36194	19.85
Bán nhà mặt phố (...)	23939	13.13
Bán nhà biệt thự,...	16523	9.06
Bán đất nền dự án...	15514	8.51
Bán loại bất động...	972	0.53
Bán kho, nhà xưởng	669	0.37
Bán trang trại, k...	477	0.26
Bán condotel	143	0.08



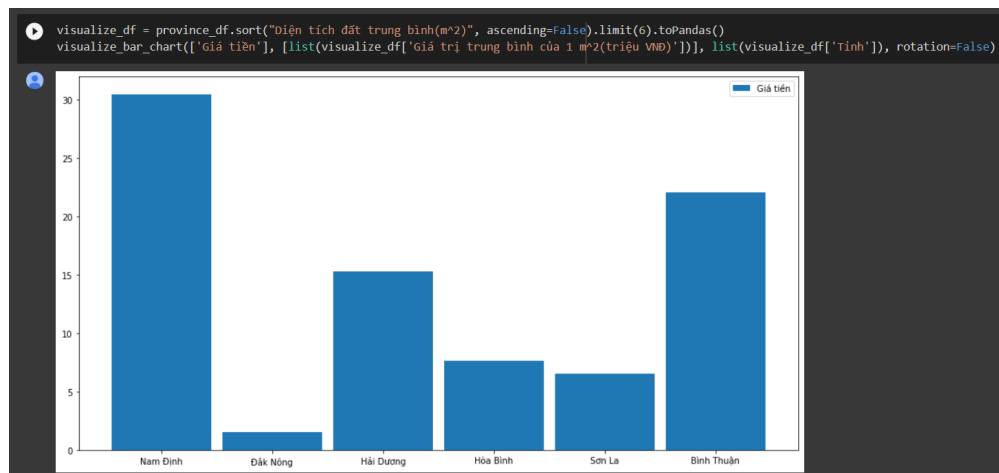
Hình 13: Thống kê về loại tin được đăng

Từ bảng thống kê, các loại tin được đăng chủ yếu là: bán đất, bán nhà riêng, bán căn hộ chung cư, bán nhà mặt phố, bán nhà biệt thự. Từ đó càng khẳng định phân tích đã được nêu ra ở phần phân tích ở Kibana.

Tỉnh	Diện tích đất trung bình(m <sup>2</sup> )	Mức giá đất trung bình(triệu VNĐ)	Giá trị trung bình của 1 m <sup>2</sup> (triệu VNĐ)	Số lượng tin	Tỉ lệ tin(%)
Hòa Bình	43609.57	5073.89	7.13	745	0.41
Hải Dương	41272.11	7947.81	14.96	89	0.05
Kiên Giang	34630.64	15594.8	68.19	953	0.53
Đắk Nông	18490.51	3700.18	1.24	35	0.02
Hà Nam	17838.4	5748.43	22.23	103	0.06
Nam Định	17145.57	18540.98	30.45	61	0.03
Thái Bình	7889.73	4351.27	20.0	64	0.04
Hà Tĩnh	7503.37	8497.82	6.23	971	0.54
Khánh Hòa	7313.53	1568471.72	203.97	5461	3.03
Bình Phước	7117.86	3484.29	4.97	1492	0.83
Sơn La	6665.99	3406.72	5.65	110	0.06
Bình Thuận	6566.06	202937.15	30.15	1848	1.03
Lào Cai	5355.09	20706.1	43.44	240	0.13
Quảng Ninh	4871.79	5660.1	34.02	435	0.24
Tiền Giang	4109.85	5663.9	13.46	150	0.08
Gia Lai	4060.78	2302.5	9.2	9	0.0
Vĩnh Long	3814.45	6569.52	11.48	84	0.05
Thanh Hóa	3679.3	5114.97	21.66	653	0.36
Lâm Đồng	3649.89	25110.19	19.4	3267	1.81
Hậu Giang	3228.65	37506.59	15.04	43	0.02

Hình 14: Thống kê chung về các tỉnh

Đứng đầu danh sách các tỉnh có diện tích đất trung bình được giao bán lớn nhất không phải là các tỉnh có lượng tin giao bán lớn nhất như đã thống kê ở biểu đồ kibana mà là các tỉnh có lượng tin giao bán vô cùng ít (chiếm dưới 1%) cho thấy các tỉnh này chủ yếu giao bán đất thuộc loại tin đất nền dự án, bán kho, nhà xưởng hoặc bán trang trại, khu nghỉ dưỡng. Trong số các tỉnh kể trên, một số tỉnh có giá trị đất trung bình trên  $m^2$  vô cùng lớn (Khánh Hòa - 203,97 triệu, Kiên Giang - 68,19 triệu, Lào Cai - 43,44 triệu) là minh chứng cho sức hút của những tỉnh này so với các tỉnh còn lại.



Hình 15: Thống kê về giá đất trung bình trên  $m^2$  của 6 tỉnh có diện tích trung bình lớn nhất

Tỉnh	Loại tin đăng	Diện tích đất trung bình(m <sup>2</sup> )	Diện tích đất lớn nhất(m <sup>2</sup> )	Diện tích đất nhỏ nhất(m <sup>2</sup> )	Mức giá đất trung bình(triệu VNĐ)	Mức giá đất lớn nhất(triệu VNĐ)
Khánh Hòa	Bán đất	15983.9	3.0E7	1.15	3226738.61	7.54
Hà Nội	Bán nhà mặt phố [...]	169.71	200000.0	7.6	54617.88	54000
Hồ Chí Minh	Bán nhà mặt phố [...]	172.24	7215.0	10.0	44838.76	40000
Đà Nẵng	Bán nhà riêng	92.49	1000.0	20.0	14294.1	28500
Đà Nẵng	Bán nhà mặt phố [...]	156.12	11237.0	36.0	31674.05	57000
Hà Nội	Bán nhà biệt thự [...]	168.82	2715.0	25.0	28018.19	1
Cần Thơ	Bán nhà mặt phố [...]	340.15	12000.0	36.7	31446.31	2641
Hồ Chí Minh	Bán nhà biệt thự [...]	215.95	7000.0	30.0	36517.96	31500
Lào Cai	Bán loại bất động [...]	90.0	90.0	90.0	14000.0	140
Hải Phòng	Bán nhà mặt phố [...]	97.44	644.0	19.0	12814.14	1250
Khánh Hòa	Bán nhà mặt phố [...]	133.09	1000.0	25.5	20249.12	3500
Hà Nội	Bán nhà riêng	57.45	12000.0	3.1	8026.29	64000
Lâm Đồng	Bán nhà mặt phố [...]	463.13	9000.0	60.0	30531.78	2400
Hồ Chí Minh	Bán nhà riêng	87.65	9548.7	7.2	10464.98	41500
Vĩnh Phúc	Bán loại bất động [...]	26.8	26.8	26.8	3200.0	32
Bà Rịa-Vũng Tàu	Bán nhà mặt phố [...]	255.68	8696.0	22.0	19005.27	2100
Hưng Yên	Bán nhà biệt thự [...]	256.35	1000.0	50.0	30413.42	1350
Kiên Giang	Bán nhà mặt phố [...]	160.5	557.0	81.0	17027.81	420
Lào Cai	Bán nhà mặt phố [...]	1094.21	6000.0	100.0	70920.0	2500
Hồ Chí Minh	Bán đất nền dự án [...]	171.10	10000.0	48.0	15581.32	48000
Kiên Giang	Bán condotel	32.48	87.0	29.0	3155.3	63
Đà Nẵng	Bán loại bất động [...]	122.48	443.4	8.0	13712.76	900
Đồng Nai	Bán loại bất động [...]	140.02	1300.0	30.0	8518.66	150
Lâm Đồng	Bán nhà biệt thự [...]	474.89	3150.0	70.0	35356.65	10100
Hưng Yên	Bán nhà mặt phố [...]	133.24	250.0	87.5	11274.44	355
Hải Phòng	Bán nhà riêng	70.25	2880.0	30.0	5548.48	12000
Thừa Thiên Huế	Bán nhà mặt phố [...]	122.21	305.0	60.0	11415.44	301
Hải Phòng	Bán nhà biệt thự [...]	143.99	644.0	42.0	12412.03	450
An Giang	Bán condotel	41.02	59.0	30.0	3391.6	41
Hồ Chí Minh	Bán loại bất động [...]	253.63	29000.0	13.5	15736.17	6600

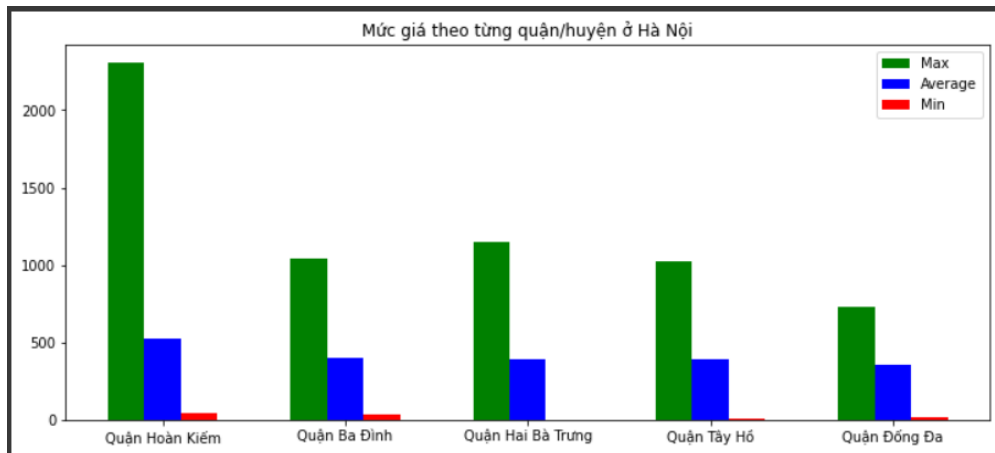
Hình 16: Thống kê chi tiết về từng tỉnh được sắp xếp theo giá trị trung bình trên  $m^2$  giảm dần

Từ bảng thống kê, loại bất động sản có giá trị đất trung bình trên  $m^2$  lớn nhất chủ yếu là nhà mặt phố, biệt thự với diện tích trung bình của các căn nhà trải dài từ 170-470  $m^2$ , với diện tích lớn nhất có thể lên tới 10000  $m^2$  và diện tích nhỏ nhất thì chưa tới 3  $m^2$

Quận/Huyện	Loại tin đăng	Diện tích đất trung bình(m <sup>2</sup> )	Diện tích đất lớn nhất(m <sup>2</sup> )	Diện tích đất nhỏ nhất(m <sup>2</sup> )	Mức giá đất trung bình(triệu VNĐ)	Mức giá đất lớn nhất(triệu VNĐ)
Quận Hoàn Kiếm	Bán nhà mặt phố [...]	230.9	35000.0	2.6	139000.0	540
Quận Ba Đình	Bán nhà mặt phố [...]	135.03	2300.0	16.0	54309.61	9
Quận Hai Bà Trưng	Bán nhà mặt phố [...]	122.02	1148.8	8.5	51801.17	50
Quận Tây Hồ	Bán nhà mặt phố [...]	174.81	3626.8	19.0	66898.24	70
Quận Đống Đa	Bán nhà mặt phố [...]	122.0	10000.0	19.2	43147.31	3
Quận Cầu Giấy	Bán nhà mặt phố [...]	140.59	3300.0	10.0	47793.13	9
Quận Thanh Xuân	Bán nhà mặt phố [...]	139.19	2250.0	28.0	38238.72	60
Quận Bắc Từ Liêm	Bán nhà mặt phố [...]	135.99	2500.0	24.0	33938.33	6
Quận Long Biên	Bán nhà mặt phố [...]	112.31	1050.0	20.0	20226.50	9
Quận Hà Đông	Bán nhà mặt phố [...]	385.0	200000.0	23.5	15425.29	1
Huyện Thanh Trì	Bán nhà mặt phố [...]	121.41	950.0	32.0	20250.11	1
Huyện Gia Lâm	Bán nhà mặt phố [...]	118.43	366.5	40.0	13732.34	1
Huyện Hoàn Kiếm	Bán nhà mặt phố [...]	92.08	286.0	28.0	8621.08	1
Huyện Thường Tín	Bán nhà mặt phố [...]	80.0	80.0	80.0	8000.0	1
Huyện Đông Anh	Bán nhà mặt phố [...]	87.73	230.0	40.0	7444.76	1
Huyện Ứng Hòa	Bán nhà mặt phố [...]	193.86	264.3	93.5	12310.0	1
Huyện Ba Vì	Bán nhà mặt phố [...]	352.5	460.0	90.0	6462.5	1
Huyện Chương Mỹ	Bán nhà mặt phố [...]	135.21	166.5	91.5	4300.0	1
Huyện Sóc Sơn	Bán nhà mặt phố [...]	400.0	400.0	400.0	11500.0	1
Huyện Quốc Oai	Bán nhà mặt phố [...]	105.9	110.0	101.8	2525.0	1

Hình 17: Thống kê chi tiết về từng quận ở Hà Nội được sắp xếp theo giá trị trung bình trên  $m^2$  giảm dần

Loại bất động sản có giá trị trung bình trên  $m^2$  lớn nhất cũng như phân tích ở trên là nhà mặt phố với 5 quận đứng đầu danh sách là Quận Hoàn Kiếm, Quận Ba Đình, Quận Hai Bà Trưng , Quận Tây Hồ và Quận Đống Đa.



Hình 18: Thống kê về 5 quận của Hà Nội có giá trị trung bình trên  $m^2$  lớn nhất

Từ hình vẽ ta có thể thấy giá trị trên  $m^2$  thấp nhất và giá trị trung bình trên  $m^2$  của các quận đa số là

như nhau, còn giá trị trên  $m^2$  lớn nhất thuộc về Quận Hoàn Kiếm với cách biệt gấp đôi so với 4 quận còn lại.

Chủ đầu tư	Tổng diện tích đất giao bán(m <sup>2</sup> )	Tổng giá trị đất giao bán(ngàn tỉ VNĐ)	Số dự án tham gia	Tỉ lệ tham gia dự án(%)
Tập đoàn Vingroup	687775.76	75.68	5025	9.37
Novaland Group	261552.11	17.95	2508	4.68
Công ty CP Tập đo...	192582.03	9.08	2155	4.02
Công ty TNHH Phát...	270770.07	37.9	1772	3.31
Tổng công ty Đầu ...	242525.85	13.14	1086	2.03
Công ty CP Đầu tư...	95906.63	5.55	960	1.79
Công ty Cổ phần T...	92542.59	4.06	737	1.37
Công ty CP Đầu tư...	62548.8	1.11	734	1.37
Tập đoàn Sun Group	121244.95	7.11	733	1.37
Công ty Cổ Phần Đ...	99315.71	8.78	724	1.35
Tập đoàn Capitaland	90340.57	5.62	671	1.25
Công ty CP Tập đo...	95807.49	6.71	625	1.17
Công ty TNHH Kepp...	72358.41	6.27	602	1.12
Công ty CP Him Lam	54907.55	6.46	466	0.87
Tổng công ty Đầu ...	83310.09	7.24	463	0.86
Tổng công ty Vigl...	164198.29	3.6	460	0.86
Công ty CP Phát t...	36593.31	1.33	454	0.85
Công ty TNHH Gamu...	49883.84	2.79	453	0.84
Công ty TNHH phát...	75884.54	8.68	444	0.83
Công ty CP Tập đo...	107863.21	2.74	422	0.79

Hình 19: Thống kê về các chủ đầu tư sắp xếp theo số dự án tham gia giảm dần

Từ bảng thống kê, ta có thể thấy được mặc dù số dự án tham gia là sắp xếp theo chiều giảm dần nhưng tổng diện tích đất giao bán và tổng giá trị đất giao bán lại không tuân theo quy luật giảm dần đó, điều đó cho thấy một vài công ty chú trọng vào việc đầu tư các dự án lớn hơn là đầu tư vào nhiều dự án nhỏ. Dễ dàng thấy rằng Tập đoàn Vingroup đứng đầu danh sách với cách biệt rất lớn so với các đối thủ còn lại về cả 3 tiêu chí: tổng diện tích đất giao bán, tổng giá trị đất giao bán và số dự án tham gia.

## 6 Một số công cụ Spark nâng cao

### 6.1 Áp dụng Structured Streaming xử lý dữ liệu stream

#### \* Mô tả

Phần mô phỏng dữ liệu streaming nhằm mô phỏng lại cách hệ thống hoạt động trong thực tế khi dữ liệu được cập nhật liên tục theo ngày với trung bình mỗi ngày có khoảng từ 100-500 bài viết mới được đăng tải lên trên website. Do đặc thù của dữ liệu bất động sản, số lượng tin đăng bất động sản là không lớn, không phải dữ liệu liên tục, update theo từng giây nên việc lựa chọn kiến trúc streaming như thế nào cũng cần lưu ý. Với các đặc điểm trên, nhóm đã xây dựng và mô phỏng việc truyền và nhận dữ liệu streaming sử dụng Spark Structured Streaming để xử lý dữ liệu stream

Spark Structured Streaming là một công cụ xử lý dữ liệu stream có khả năng mở rộng và khả năng chống chịu lỗi cao. Dữ liệu vào của Structured Streaming là các Dataset/DataFrame và các bước xử lý được thực hiện bằng Spark SQL. Với đặc trưng của dữ liệu và hệ thống khi dữ liệu là dạng bảng và hệ thống mỗi ngày chỉ nhận vào khoảng 100-500 bản ghi dữ liệu thì sử dụng Spark Structured Streaming là phù hợp.

Để mô phỏng quá trình streaming, nhóm đã sử dụng 2 spark job làm nhiệm vụ truyền và nhận dữ liệu:

- Spark job 1: Làm nhiệm vụ đóng gói và gửi dữ liệu là file .csv vào trong thư mục */bigdataproject/streaming\_data* trên HDFS. Dữ liệu được đóng thành các gói gồm 500 bản ghi(mô phỏng) và được gửi liên tục sau mỗi khoảng thời gian là 30s
- Spark job 2: Làm nhiệm vụ đọc dữ liệu streaming từ thư mục */bigdataproject/streaming\_data* và đưa ra các thông tin hữu ích như : Các dự án bất động sản đang hot, Các tỉnh thành đang có nhiều bất động sản giao bán, ... Xử lý streaming dữ liệu của Spark job 2 có thể hiển thị trong console hoặc lưu vào memory để thực hiện các truy xuất sau này.

#### \* Kết quả chạy

window	Tên dự án	Số lượng tin
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Lavila Kiến Á - Nhà Bè	9
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	TTTM và Phố chợ Đô Nghĩa	5
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Sim City	3
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Sarimi Sala	3
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Nine South Estates	3
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Khu đô thị Thanh Hà Mường Thanh	2
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	M-One Nam Sài Gòn	2
unscroll output: double click to hide {2021-06-03 12:00:00, 2021-06-05 12:00:00}	Saigon South Residences	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Sài Gòn Eco Lake	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	KDC Nam Long Phú Thuận	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Khu đô thị Minh Giang Đầm VÀ	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Khu đô thị mới Đại Kim - Định Công	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Tháp Mười Merita	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Lucky Palace	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Sky Garden I	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Riverside Residence	1
{2021-06-03 12:00:00, 2021-06-05 12:00:00}	Hưng Ngân Garden	1

Hình 20: Một số phân tích từ dữ liệu streaming

### 6.2 Áp dụng MLlib vào dự đoán giá nhà

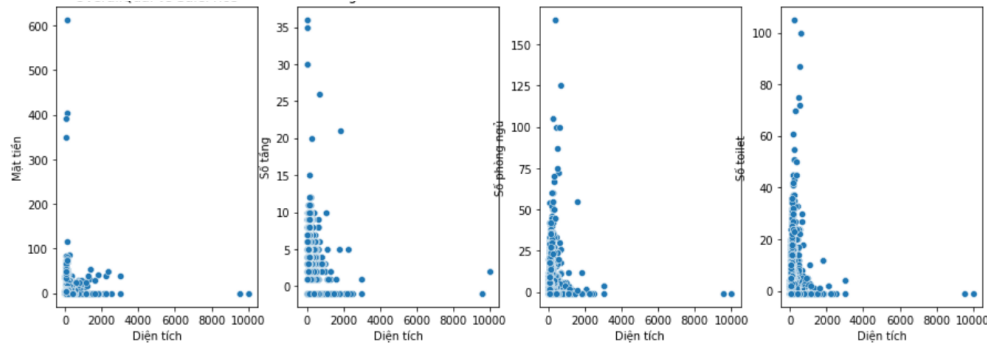
#### \* Mô tả

Với dữ liệu gồm hơn 200000 bản ghi với nhiều trường đặc trưng cho bất động sản, ta có thể áp dụng các mô hình regression để dự đoán giá bất động sản dựa trên các đặc trưng của bất động sản đó. Mô hình này có thể được áp dụng để gợi ý giá cho người dùng khi đăng tin lên website.

Để đơn giản, nhóm sẽ chỉ thực hiện mô hình dự đoán giá nhà riêng vì loại tin này hầu hết có đủ các trường quan trọng cho một bất động sản: *Diện tích, Số phòng ngủ, Số toilet, Số tầng, Quận/Huyện, Đường vào, Mặt tiền*. Chúng ta sẽ chỉ giữ lại những đặc trưng này cho bài toán dự đoán giá nhà riêng.

Để có một mô hình tốt, trước tiên nhóm đã sử dụng một số phương pháp tiền xử lý dữ liệu trước khi cho vào mô hình MLlib như sau:

- Lọc ra các Quận/Huyện có ít hơn 200 bản ghi, chỉ còn lại 23 quận/huyện với hơn 30000 bản ghi có thể dự đoán được.
- Lọc ra các outliers thông qua visualize trên đồ thị để thấy được đặc trưng của dữ liệu



Hình 21: Visualize phân bố dữ liệu

- Encode trường Quận/Huyện về dạng numeric, scale dữ liệu và chia thành hai tập train/test theo tỉ lệ 85:15

Sau khi xử lý dữ liệu, ta sẽ sử dụng hai mô hình Linear Regression, Random Forest Regression và dùng Root Mean Squared Error (RMSE) để đánh giá mô hình.

#### \* Kết quả chạy

```
+-----+
| prediction|Mức giá| scaledFeatures|
+-----+
| 3880.233530758681| 4150.0|[0.67881991248299...|
| 5044.895000000001| 4500.0|[0.98913758676093...|
| 9029.467620124973| 7600.0|[0.77579418569485...|
| 2375.8690476190477| 1800.0|[0.75639933105248...|
| 15774.916666666666| 14900.0|[1.84251119102527...|
+-----+
```

only showing top 5 rows

R Squared (R2) on val data = 0.625795

Hình 22: Mô hình random forest

```
+-----+
| prediction|Mức giá| scaledFeatures|
+-----+
| 6482.89550297895| 4150.0|[0.67881991248299...|
| 6640.44843632443| 4500.0|[0.98913758676093...|
| 7455.403147846682| 7600.0|[0.77579418569485...|
| 6247.414694413271| 1800.0|[0.75639933105248...|
| 14508.477990170839| 14900.0|[1.84251119102527...|
+-----+
```

only showing top 5 rows

R Squared (R2) on val data = 0.463612

Hình 23: Mô hình linear regression



## 7 Kết luận

Thông qua project này, bọn em đã có thể hiểu hơn các kiến thức đã được học và biết vận dụng những kiến thức đó để làm việc với một bộ dữ liệu trong thực tế là dữ liệu bất động sản, đưa ra các cách thức để thao tác, lưu trữ, xử lý, phân tích trên bộ dữ liệu đó:

- Lấy dữ liệu thô từ các website về để tiền xử lý trước khi được đem đi lưu trữ.
- Lưu trữ dữ liệu bất động sản trên cụm HDFS đảm bảo dữ liệu khi được sử dụng sẽ luôn sẵn sàng dù trong tình huống một vài node trong cụm có vấn đề, không thể kết nối đến để lấy dữ liệu được.
- Dùng Elastic Search và Kibana để trực quan hóa và phân tích dữ liệu theo những trường dữ liệu đơn giản, việc thực hiện tính toán không quá phức tạp.
- Tiếp tục sử dụng khả năng tính toán phân tán, song song của cụm Spark để phân tích dữ liệu theo các công thức, trường dữ liệu phức tạp hơn mà Elastic Search và Kibana không thể thực hiện được hoặc thời gian thực thi sẽ lâu.
- Thực hiện mô phỏng việc xử lý dữ liệu theo dòng thông qua Spark Streaming nhằm đánh giá, thử nghiệm hoạt động của cụm Spark nếu xử lý dữ liệu theo thời gian thực trong thực tế.
- Sử dụng thư viện MLlib để dự đoán giá nhà thông qua 2 mô hình Linear Regression và Random Forest Regression với mục đích thử nghiệm thư viện MLlib và giải quyết bài toán dự đoán giá nhà đất trong thực tế.

Để thực hiện những công việc đã nêu trên, chúng em xin cảm ơn thầy Nguyễn Bình Minh đã cung cấp những giờ học lý thuyết kèm với thực hành bổ ích, thú vị để từ đó bọn em dễ dàng tiếp thu và hứng thú áp dụng những công nghệ đã học vào bài toán thực tế này. Project này là sự nghiên cứu, đóng góp của tất cả các thành viên trong nhóm, và sau đây là công việc của từng thành viên:

- Vương Xuân Hoàng: MLlib + Structured Streaming
- Lương Cường Thịnh: Visualize Kibana + Phân tích Spark
- Nguyễn Hữu Hiệp: Tiền xử lý dữ liệu + Tổ chức lưu trữ
- Nguyễn Văn Duy: Crawl data + Phân tích Spark

*Source code tại: <https://github.com/vuonghoangbntt/BigDataProject>*

## Tài liệu

- [1] Các câu lệnh, ví dụ về sử dụng Spark:  
<https://sparkbyexamples.com/>
- [2] Spark Structured Streaming documents:  
<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html>
- [3] Spark MLlib documents:  
<https://spark.apache.org/docs/latest/ml-guide.html>
- [4] MLlib for house price prediction on Kaggle  
<https://www.kaggle.com/ilyapozdnyakov/house-prices-prediction-pyspark-guide>
- [5] Slides bài giảng môn Lưu trữ và xử lý dữ liệu lớn của thầy Nguyễn Bình Minh