

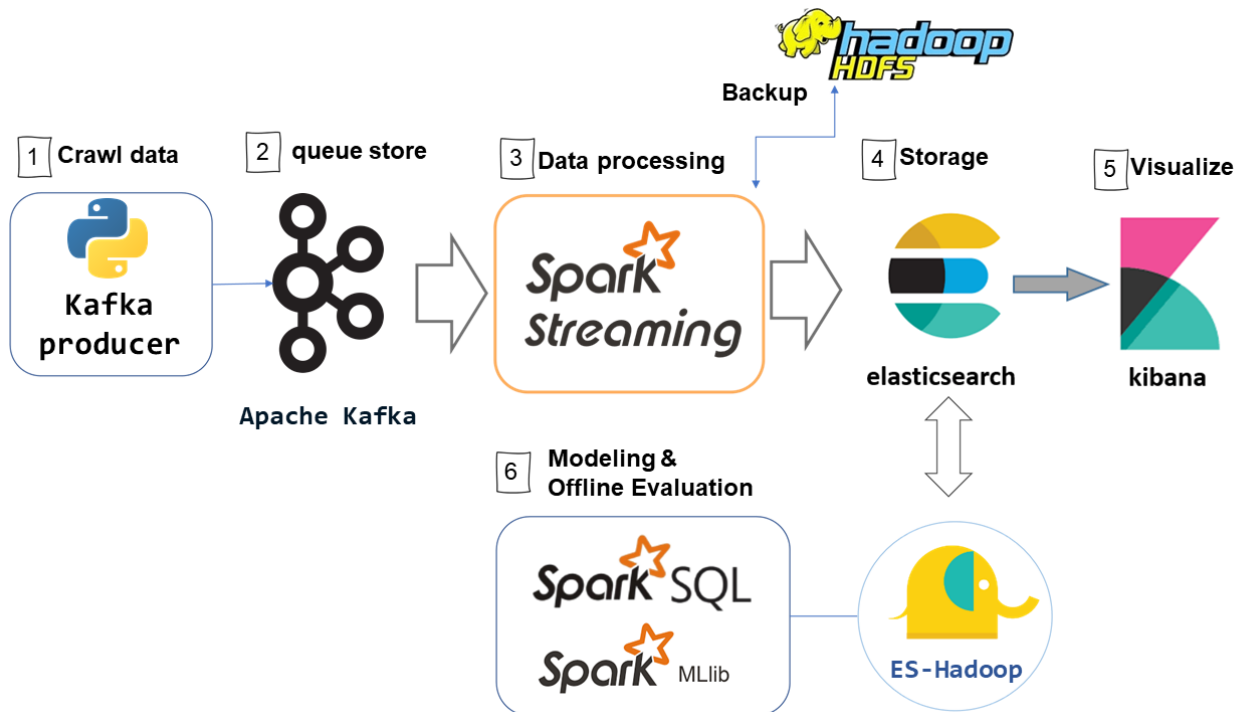
I. Pipeline

Bài toán: Thu thập dữ liệu về mua bán nhà ở Hà Nội, thực hiện phân tích, đánh giá, mô hình hóa dự đoán giá nhà.

Mô tả luồng hoạt động:

Luồng xử lý dữ liệu sẽ bao gồm các bước sau:

1. Dữ liệu sẽ được crawl từ trang <https://meeyland.com/> và đẩy vào hàng đợi kafka,
2. Spark streaming sẽ lấy dữ liệu trong kafka ra, tiến hành trích xuất, biến đổi dữ liệu thành đúng định dạng và đẩy vào Elasticsearch, đồng thời lưu backup vào HDFS.
3. Tiếp đó, dữ liệu trong elasticsearch sẽ được phân tích, đánh giá và trực quan hóa thông qua Kibana.
4. Cuối cùng, sử dụng ESHadoop để load dữ liệu từ Elasticsearch vào Spark, tiến xử lý bằng Spark SQL và tiến hành biến đổi, mô hình hóa dự đoán và đánh giá bằng Spark MLlib.



II. Crawl Data

Dữ liệu mua bán nhà ở tại Hà Nội được crawl trên trang <https://meeyland.com/> từ ngày 01/01/2020 – 31/12/2021.

The screenshot displays the Meey Land website interface. At the top, there's a navigation bar with the Meey Land logo, a search bar, and links for 'Mua bán', 'Cho thuê', 'Sang nhượng', and 'Tin tức'. Below this is a filter bar with checkboxes for 'Cần Bán' and 'Cần Mua', and dropdown menus for 'Khu vực' (Hà Nội), 'Loại nhà đất' (Nhà ở), 'Mức giá' (Tất cả), 'Khoảng diện tích' (Tất cả), and 'Dự án' (Tất cả). A 'Tìm kiếm' button is also present. The main content area shows a list of properties under the heading 'Loại nhà đất'. On the left, there's a sidebar with filters for 'Loại nhà đất' (Nhà ở, Chung cư, etc.) and 'Khu vực' (Thành phố Hà Nội, Quận Đống Đa, etc.). The main list displays three property cards, each with a photo, title, price, area, and location. For example, the first card is for a 35m² house in the Ba Đình district, priced at 2.15 billion VND.

Quá trình crawl sử dụng python **request** và **beautifulsoup** làm producer, thực hiện trên khoảng 1200 trang , mỗi trang 24 căn nhà. Sau khi crawl, các thẻ div quan trọng được đẩy vào hàng đợi Kafka topic “nha-hanoi”.

Tiếp đó, dữ liệu được lấy streaming từ kafka ra sử dụng gói **spark-streaming-kafka-0-8_211** sau đó tiến hành parse các trường quan trọng, transform dữ liệu về dạng có cấu trúc như dictionary hay documents thông qua tùy chỉnh map, reduce và cuối cùng tạo kết nối đẩy vào elasticsearch.

Ngoài ra, tại mỗi batch_interval, lưu các rdds vào hdfs (saveAsTextFiles) với thông tin về số trang, số thứ tự nhà crawl hiện tại để backup trong trường hợp lỗi khi lưu vào elasticsearch.

Kết quả chạy Spark Streaming Kafka:

```
docker exec -it e2bfa8253816950a570648ab65326b35dd1b898740c32101fb396b17eb18d66d /bin/sh
```

```
21/12/31 08:22:23 INFO storage.BlockManagerMasterEndpoint: Registering block manager 172.20.0.3:45265 with 93.3 MB RAM, BlockManagerId(driver, 172.20.0.3, 45265, None)
21/12/31 08:22:23 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 172.20.0.3, 45265, None)
21/12/31 08:22:23 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, 172.20.0.3, 45265, None)
21/12/31 08:22:23 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@60d3560a{/metrics/json,null,AVAILABLE,@Spark}
```

```
-----
RDDs RECEIVED : 0
-----
```

```
-----
RDDs RECEIVED : 0
-----
```

```
-----
RDDs RECEIVED : 7
-----
```

```
connected to elasticsearch status_code:200
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61c98a2369dd7400187b5d1b', '_type': 'nha', '_version': 2, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 1, '_index': 'nha_ha_noi_2'}
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61ceb9ec21ed040019136a9', '_type': 'nha', '_version': 2, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 3, '_index': 'nha_ha_noi_2'}
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61ceb9ec21ed040019136a9', '_type': 'nha', '_version': 2, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 3, '_index': 'nha_ha_noi_2'}
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61cb43ca662e110019672ca0', '_type': 'nha', '_version': 3, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 3, '_index': 'nha_ha_noi_2'}
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61ceb09421ed0400190de45d', '_type': 'nha', '_version': 3, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 3, '_index': 'nha_ha_noi_2'}
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61ceb2e8db87b00018dcf3ca', '_type': 'nha', '_version': 1, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 1, '_index': 'nha_ha_noi_2'}
```

```
TAKE(1) RDD TEST: [
```

```
[
  {
    "61c98a2369dd7400187b5d1b",
    {
      "location": {
        "street": null,

```

```
docker exec -it e2bfa8253816950a570648ab65326b35dd1b898740c32101fb396b17eb18d66d /bin/sh
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-2pn-gan-pho-co-gia-tu-2-5-ty-ck-7-lai-0-18-thang-co-qua-tan-gia-1640929199007
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-le-grand-jardin-sai-dong-o-ngay-htls-0-15-thang-ck-7-gtch-mien-lai-2-nam-dv-09345-98936-1640930731278
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/so-huu-can-ho-2pn-view-cong-vien-sai-dong-lake-view-htls-0-ck-6-giam-tan-180-trieu-1640926479789
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/nhan-dat-cho-phan-khu-vip-prime-jardin-2-2-toa-doc-ton-la-ke-view-tai-le-grand-jardin-sai-dong-1640925116526
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/can-ban-can-ho-cau-cap-le-grand-jardin-sai-dong-1-7-ty-54-m2-1640922336236Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-doi-can-ba-dinh-ha-noi-di-en-tich-54m2-gia-1-25-ty-1640920830850
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-toa-nha-x2-dai-kim-ha-noi-dien-tich-86m2-gia-2-8-ty-1640920436268
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-khuong-ha-thanh-xuan-ha-noi-dien-tich-60m2-gia-1-3-ty-1640920167351
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/chi-tu-2-1-ty-can-ho-3pn-duy-nhat-viet-hung-long-bien-0961491566-1640918570838
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-toa-chung-cu-mini-tai-me-tri-nam-tu-liem-ha-noi-dien-tich-65m2-gia-8-2-ty-1640917472310
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-hateco-laroma-ha-noi-dien-tich-112m2-gia-5-4-ty-1640916770551
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/gia-soc-ban-chung-cu-e1-ton-that-tung-dai-hoc-y-ha-noi-gia-tu-600tr-1can-ve-o-ngay-full-noi-that-1637198844952
```

```
docker exec -it e2bfa8253816950a570648ab65326b35dd1b898740c32101fb396b17eb18d66d /bin/sh
```

```
: 3, '_index': 'nha_ha_noi_2'
Pushing to Elasticsearch...
```

```
{'_id': '61ceb09421ed0400190de45d', '_type': 'nha', '_version': 2, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 3, '_index': 'nha_ha_noi_2'}
```

```
Pushing to Elasticsearch...
```

```
{'_id': '61ceb2e8db87b00018dcf3ca', '_type': 'nha', '_version': 2, '_shards': {'successful': 1, 'failed': 0, 'total': 2}, '_primary_term': 1, 'result': 'updated', '_seq_no': 1, '_index': 'nha_ha_noi_2'}
```

```
TAKE(1) RDD TEST: [
```

```
[
  {
    "61c98a2369dd7400187b5d1b",
    {
      "location": {
        "street": null,
        "district": "Qu\u01eadn Thanh Xu\u00e0n",
        "ward": "Ph\u01b0\u01edng Thanh Xu\u00e0n Trung",
        "city": "Th\u01b0\u01edng ph\u01b0\u01edng H\u01b0\u01edng",
        "project": "Bohemia Residence",
        "address": "S\u01edng 2 L\u01b0\u01edng Thi\u00eam, P
```

```
},
    "isInternalLink": false,
    "likes": 0,
    "publishedDate": "2021-12-27T09:41:20.000Z",
    "tags": null,
    "views": 145,
    "typeOfHouse": [
      "chung_cu"
    ],
    "metaTitle": "B\u01b0\u01edng chung c\u01b0 ch\u01b0\u01edng ch\u01b0\u01edng",
    "code": "101717262",
    "areaUse": 85.17,
    "filter": {
      "contactName": "Ng\u01b0\u01edc Xu\u00e0n Hi\u01b0\u01b0\u01edng",
      "need": "can ban",
      "typeOfRealEstate": "nha_du_an_cao_tang",
      "district": "Se5501caeb80a7245175de2f",
      "typeOfHouse": [
        "chung_cu"
      ],
      "contactPhone": "0967832939",
      "contactEmail": "",

```

```
docker exec -it e2bfa8253816950a570648ab65326b35dd1b898740c32101fb396b17eb18d66d /bin/sh
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-2pn-gan-pho-co-gia-tu-2-5-ty-ck-7-lai-0-18-thang-co-qua-tan-gia-1640929199007
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-le-grand-jardin-sai-dong-o-ngay-htls-0-15-thang-ck-7-gtch-mien-lai-2-nam-dv-09345-98936-1640930731278
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/so-huu-can-ho-2pn-view-cong-vien-sai-dong-lake-view-htls-0-ck-6-giam-tan-180-trieu-1640926479789
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/nhan-dat-cho-phan-khu-vip-prime-jardin-2-2-toa-doc-ton-la-ke-view-tai-le-grand-jardin-sai-dong-1640925116526
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/can-ban-can-ho-cau-cap-le-grand-jardin-sai-dong-1-7-ty-54-m2-1640922336236Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-doi-can-ba-dinh-ha-noi-di-en-tich-54m2-gia-1-25-ty-1640920830850
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-toa-nha-x2-dai-kim-ha-noi-dien-tich-86m2-gia-2-8-ty-1640920436268
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-khuong-ha-thanh-xuan-ha-noi-dien-tich-60m2-gia-1-3-ty-1640920167351
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/chi-tu-2-1-ty-can-ho-3pn-duy-nhat-viet-hung-long-bien-0961491566-1640918570838
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-toa-chung-cu-mini-tai-me-tri-nam-tu-liem-ha-noi-dien-tich-65m2-gia-8-2-ty-1640917472310
```

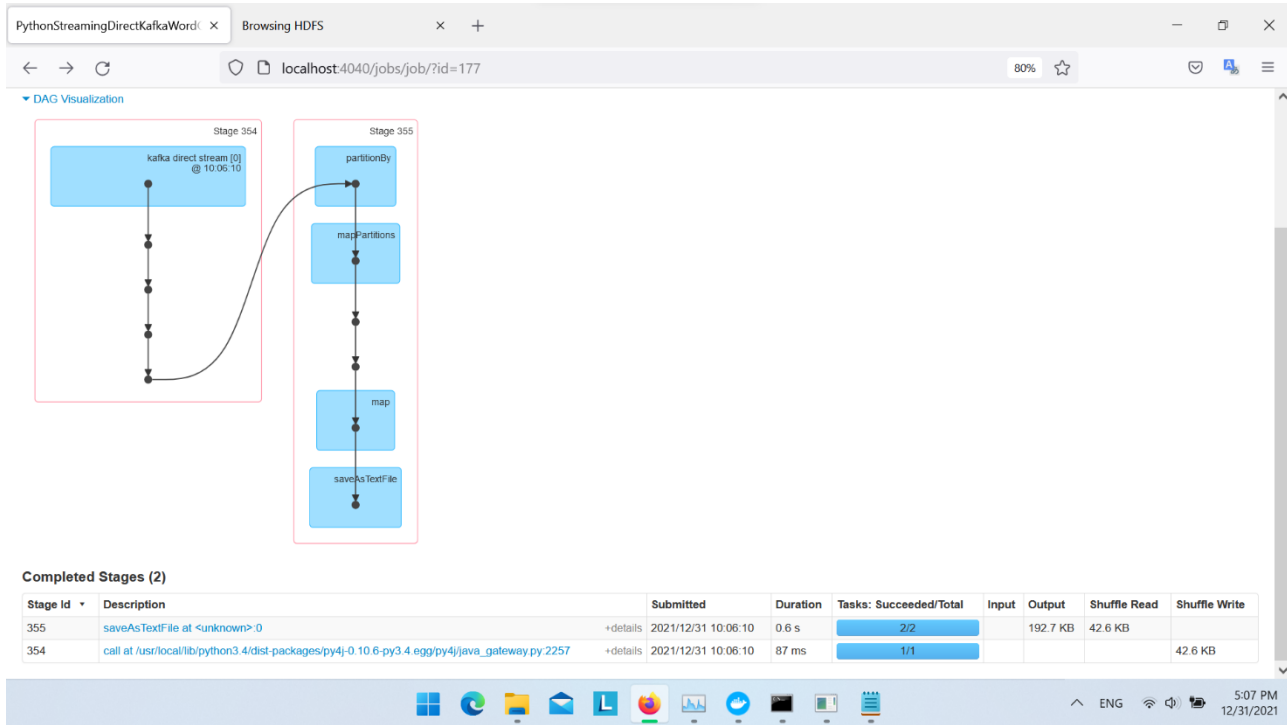
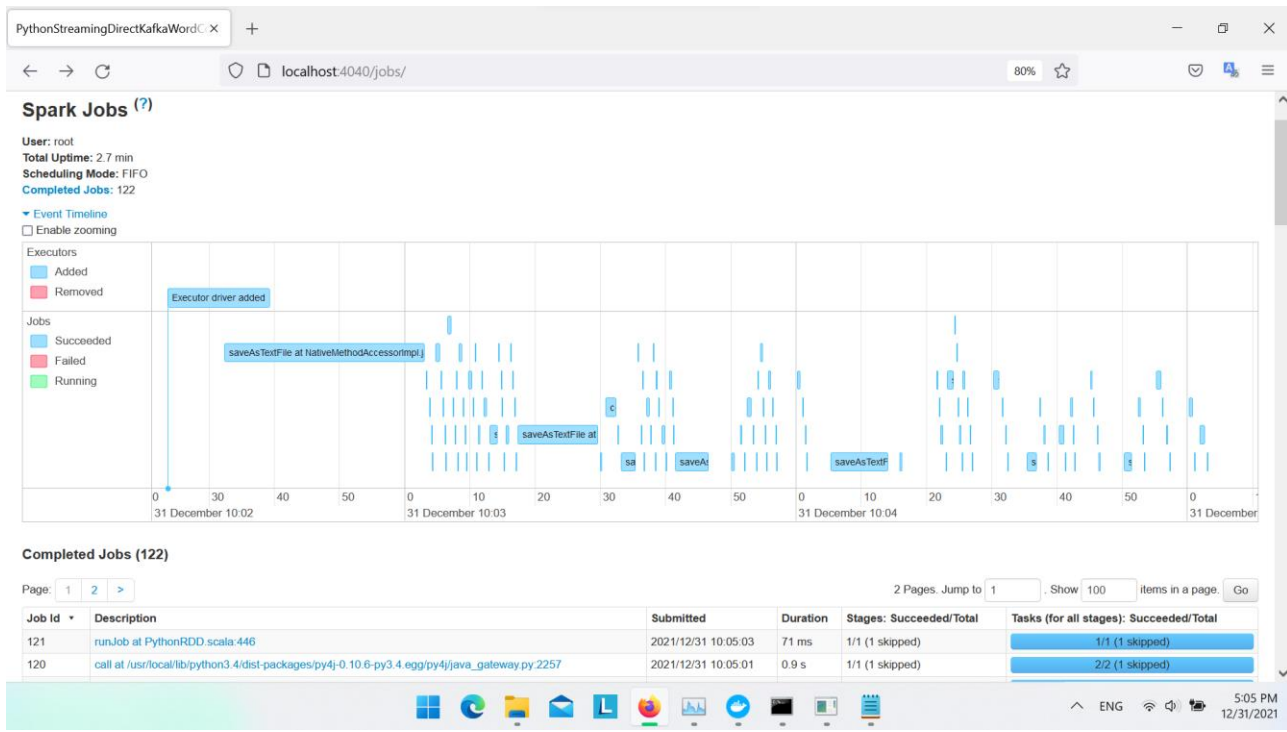
```
Crawling:https://meeyland.com/mua-ban-nha-dat/ban-can-ho-tai-hateco-laroma-ha-noi-dien-tich-112m2-gia-5-4-ty-1640916770551
```

```
Crawling:https://meeyland.com/mua-ban-nha-dat/gia-soc-ban-chung-cu-e1-ton-that-tung-dai-hoc-y-ha-noi-gia-tu-600tr-1can-ve-o-ngay-full-noi-that-1637198844952
```





Kết quả chạy job:

Chúng ta thấy rằng mỗi một time_interval là 5s, streaming sẽ xử lý được khoảng từ 5-6 căn nhà crawl được, tốc độ rất cân bằng với tốc độ của producer.



Chúng ta có thể thấy ngoài saveAsTextFile có một stage gọi đến lớp java_gateway để tạo kết nối và đẩy dữ liệu vào elasticsearch.

Kết quả lưu vào hdfs:

Browse Directory										
/Crawl_nha								Go!		
Show	25	entries		Search:						
<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:04	0	0 B	streamRDD.txt-1640945075000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:04	0	0 B	streamRDD.txt-1640945080000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:04	0	0 B	streamRDD.txt-1640945085000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:04	0	0 B	streamRDD.txt-1640945090000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:04	0	0 B	streamRDD.txt-1640945095000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:05	0	0 B	streamRDD.txt-1640945100000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:05	0	0 B	streamRDD.txt-1640945105000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:05	0	0 B	streamRDD.txt-1640945110000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:05	0	0 B	streamRDD.txt-1640945115000		
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Dec 31 17:05	0	0 B	streamRDD.txt-1640945120000		

Mỗi một rdd sẽ có kích thước khoảng từ 50-100 kb

PythonStreamingDirectKafkaWordC X

Browsing HDFS X

+


localhost:9870/explorer.html#/Crawl_nha/streamRDD.txt-1640945045000


Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities


Browse Directory

/Crawl_nha/streamRDD.txt-1640945045000

Go!







Show 25 entries


Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	root	supergroup	0 B	Dec 31 17:04	3	128 MB	_SUCCESS
<input type="checkbox"/>	-rw-r--r--	root	supergroup	47.02 KB	Dec 31 17:04	3	128 MB	part-00000
<input type="checkbox"/>	-rw-r--r--	root	supergroup	95.49 KB	Dec 31 17:04	3	128 MB	part-00001

Showing 1 to 3 of 3 entries

Previous 1 Next

Hadoop, 2019.



Kết quả lưu vào elasticsearch:

Dữ liệu thu thập được là **20780** bản ghi từ 1000 trang, dung lượng lưu trên 2 nodes **1.3gb**

Management / Elasticsearch / Index Management

Index management

Update your Elasticsearch indices individually or in bulk Include system indices

Search

<input type="checkbox"/>	Name	Health	Status	Primaries	Replicas	Docs count	Storage size	Primary storage size
<input type="checkbox"/>	data_nha_hanoi_final2	green	open	5	1	20780	1.3gb	678.5mb
<input type="checkbox"/>	data_nha_hanoi	green	open	5	1	1221	96.7mb	48.3mb
<input type="checkbox"/>	my-index-000001	green	open	5	1	18	178.2kb	89.1kb
<input type="checkbox"/>	test	red	open	5	1	0	1.5kb	783b
<input type="checkbox"/>	nha_ha_noi_2	green	open	5	1	12	1.7mb	903.4kb
<input type="checkbox"/>	shakespeare	green	open	5	1	111396	42.9mb	21.4mb
<input type="checkbox"/>	bank	red	open	5	1	809	767.8kb	383.9kb

Rows per page: 10

III. Phân tích & trực quan hóa dữ liệu

Dữ liệu bao gồm rất nhiều trường, tuy nhiên ta cũng không sử dụng hết cho bài toán dự đoán giá nhà.

Kibana Console localhost/ localhost/#/X pyspark.m Get API | Data sour Classificat Classificat Đường đn pyspark.sc +

localhost:9212/data_nha_hanoi_final2/data_nha/615b3079c2bb650012ebcdf1

JSON Dữ liệu thô Tiêu đề

Lưu Sao chép Thu gọn tất cả Mở rộng tất cả Bỏ lọc JSON

```

{
  "title": "Chính chủ bán nhà 4 tầng x 41m2 giá 2,2 tỷ Ngõ Hiệp, Thanh Trì",
  "location": {
    "city": "Thành phố Hà Nội",
    "district": "Huyện Thanh Trì",
    "project": null,
    "street": null,
    "ward": "Xã Ngõ Hiệp",
    "address": null
  },
  "metaTitle": "Chính chủ bán nhà 4 tầng x 41m2 giá 2,2 tỷ Ngõ Hiệp, Thanh Trì",
  "metaKeywords": "bán nhà, bán nhà 4 tầng, chính chủ bán nhà, ngõ hiệp, ngõ hiệp thanh trì",
  "metaDescription": "Bán nhà mới Ngõ Hiệp, Thanh Trì, cách QL 1A 100m. Nhà tự xây 4 tầng, hoàn tất nội thất 50 năm. Sổ đỏ chính chủ 41m2, mặt tiền 4m, sẵn sàng giao dịch. Khu dân trí cao, tiện ích đầy đủ, gần bệnh viện, bến xe, trung tâm thương mại. Giá 2,2 tỷ có th",
  "code": "181199628",
  "createdDate": "2021-10-04T16:46:00.000Z",
  "updatedDate": "2021-10-05T00:59:56.640Z",
  "subscription": {},
  "filter": {},
  "publishedDate": "2021-10-04T16:46:00.000Z",
  "expriedDate": "2024-06-29T16:46:00.000Z",
  "need": "can_ban",
  "typeOfHouse": {
    "0": "nha_o",
    "typeOfRealEstate": "Nha_tho_cu"
  },
  "price": {
    "total": 2100000000,
    "value": 2.1,
    "unit": "5df26a1737bdef347510cf8f",
    "mappingUnit": "ty",
    "n2": "51219512.19512195",
    "rentPrice": null
  }
}

```


20,780 hits

Search... (e.g. status:200 AND extension:PHP)

data_nha_hanoi_final2

Selected Fields

Available Fields

Popular

t feature

t ID

t _id

t _index

_score

t _type

access

t area.mappingUnit

t area.unit

area.value

areaUse

localhost:5601/app/kibana#/home

Một vài trường thuộc tính quan trọng:

20,777 hits

Search... (e.g. status:200 AND extension:PHP)

data_nha_hanoi_final

Selected Fields

Available Fields

t ID

area.value

bathroom

bedroom

t createdDate

floor

geoloc

t location.district

t location.street

t location.ward

price.total

t price.unit

price.value

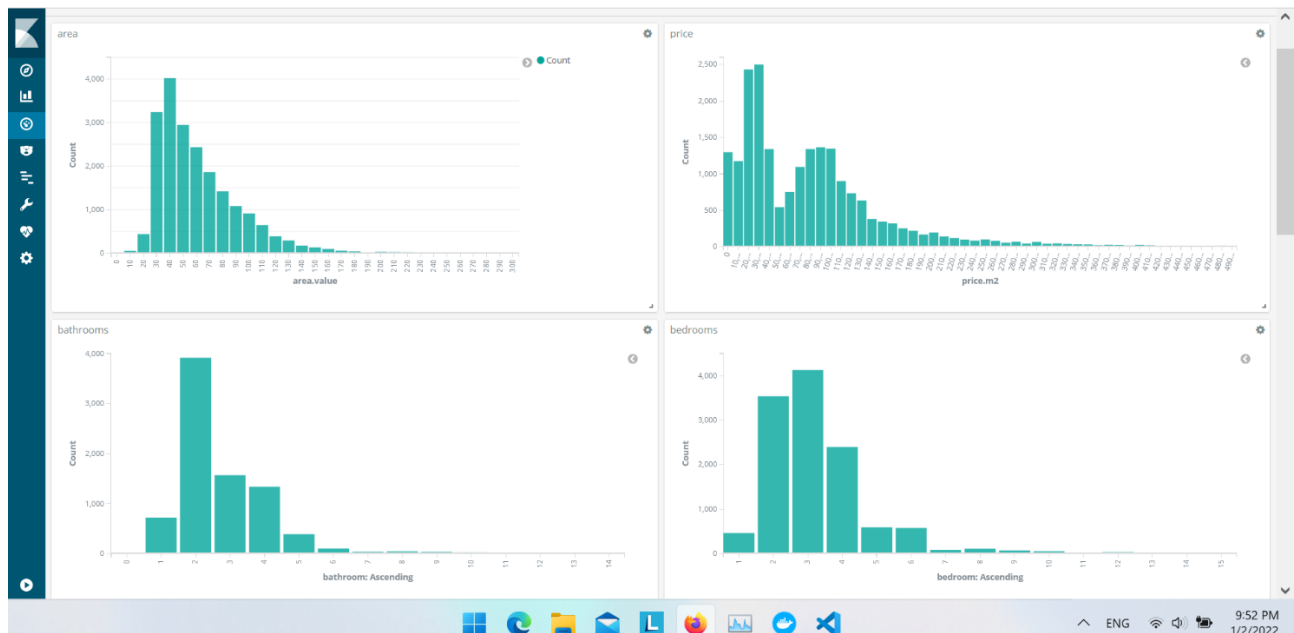
t priceData

reported

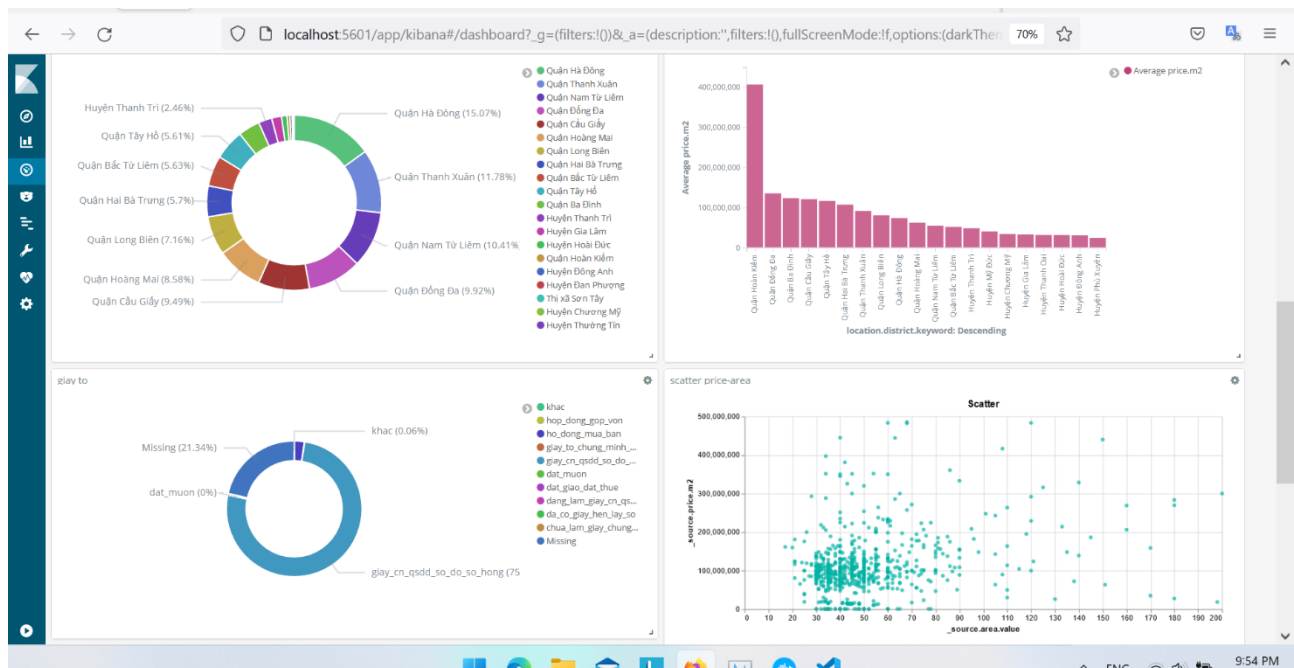
t typeOfHouse

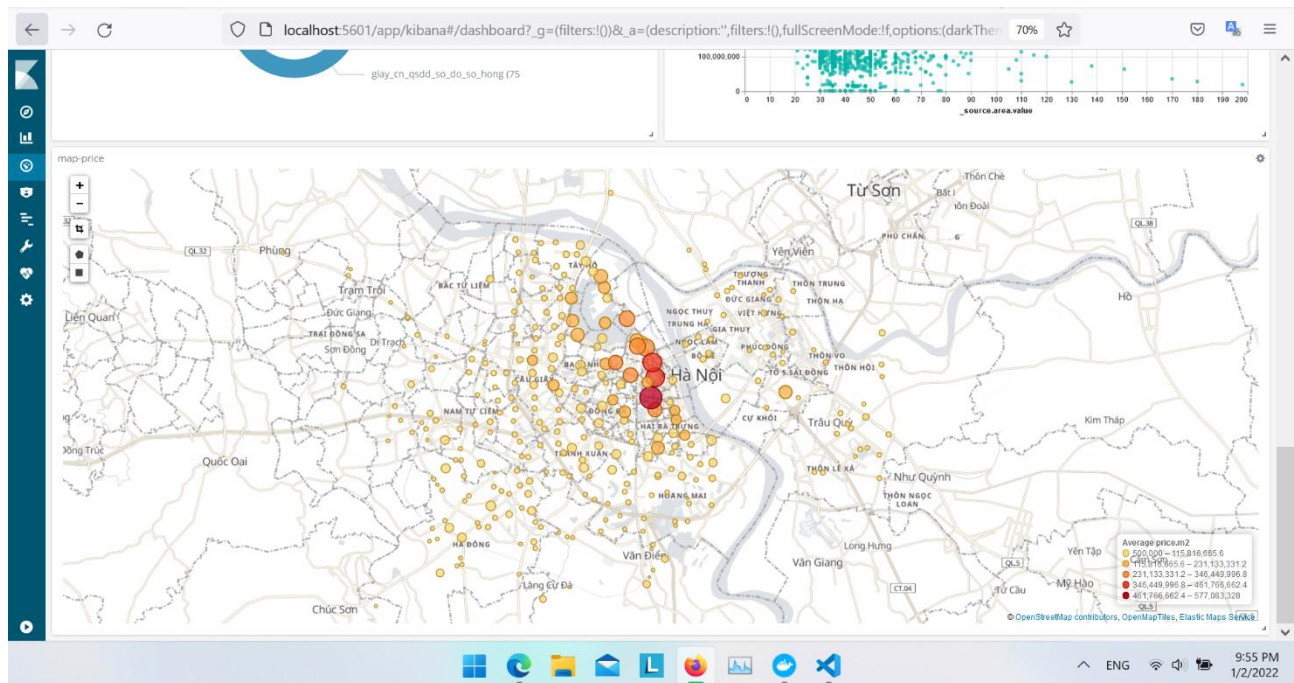
ID	createdDate	area.value	bathroom	bedroom	location.district	location.ward	location.street	price.total	geoloc	floor
6161b 6efac c2370 018a7 6635	October 9th 2021, 22:36:15.706	30	0	-	Quận Long Biên	Phường Việt Hưng	Đường Hoa Lâm	2,600,000,000	105.885, 21.019	-
615d4 ab8bb a2c50 0114b 0d5a	October 6th 2021, 14:04:00.000	45	1	1	Quận Ba Đình	Phường Vĩnh Phúc	Đường Văn Cao	650,000,000	-	-
61bd3 ab966 2e110 01988 c62b	December 18th 2021, 08:34:49.972	35	1	1	Quận Đống Đa	Phường Khâm Thiên	-	510,000,000	-	-
615fa da886 66c60 0186c f71d	October 8th 2021, 09:28:00.000	43	1	1	Quận Đống Đa	Phường Nam Đồng	Đường Nam Đồng	630,000,000	105.83, 21.018	-
61bcd f96d4 195d0 01878 c911	December 18th 2021, 02:05:58.283	138	1	1	Huyện Thanh Trì	Xã Vĩnh Quỳnh	Đường Vĩnh Quỳnh	10,000,000,000	-	1
6160f 672ac	October 9th 2021,	35	1	1	Quận Ba Đình	Phường Vĩnh Phúc	Đường Vĩnh Phúc	570,000,000	105.814, 21.034	-

Visualize dữ liệu thông qua Kibana dashboard:



Có thể thấy đa số giá sẽ rơi vào khoảng từ 20 – 300tr/m², diện tích sẽ từ 30-100m², các giá trị về số phòng ngủ, số phòng tắm, số tầng rơi vào khoảng từ 1-6,.. từ việc visualzie ta có thể thấy được khoảng giá trị cho các thuộc tính nhằm phát hiện và loại bỏ các ngoại lai.





Ta thấy giá tại các quận trung tâm hay khu vực quanh Hồ Gươm, Hồ Tây có giá trị rất cao so với các khu vực khác.

Ngoài ra, các yếu tố như về đặc điểm nổi trội(chỗ để oto, vị trí trung tâm, kinh doanh được,...), loại nhà (chung cư, nhà ở, tòa nhà văn phòng,..) , mặt tiền,.. cũng rất ảnh hưởng tới giá của nó :



IV. Mô hình hóa dự đoán giá nhà

Đầu tiên ta load data từ elasticsearch sử dụng ESHadoopAPI dưới dạng RDD, sau đó thực hiện map để lọc ra các trường cần thiết trong dữ liệu, đồng thời tiền xử lý về cấu trúc.

```
%pyspark

q = """{
  "query": {
    "match_all": {
    }
  }
}"""

es_read_conf = {
  "es.nodes" : "elasticsearch",
  "es.port" : "9200",
  "es.query" : q,
  "es.read.field.exclude" : "depth",
  "es.resource" : "data_nha_hanoi_final2/data_nha"
}

es_rdd = sc.newAPIHadoopRDD(
  inputFormatClass="org.elasticsearch.hadoop.mr.EsInputFormat",
  keyClass="org.apache.hadoop.io.NullWritable",
  valueClass="org.elasticsearch.hadoop.mr.LinkedMapWritable",
  conf=es_read_conf)
```

Took 2 sec. Last updated by anonymous at January 04 2022, 9:47:37 PM.

Tiếp đó ta thực hiện tạo Dataframe với Spark SQL:

```
%practicetest : + SQL(test.count())
test.show(7)
```

ID	area_value	bathroom	bedroom	creatorType	direction	feature	floor	geolocation_lat	geolocation_lon	legalPaper	location_district	location_street	location_ward	price_m2	wideRoad
[61cb099d9dd74001...	42	3	4	0	khac	mat_tien_rong dat...	4	null	null	giay_cn_qsdd_so_d...	Quận Hai Bà Trưng	null	Phường Thanh Lương	9.0476192E7	ngo_ngach
[61cb07da69dd74001...	47	4	4	0	khac	khac	4	null	null	giay_cn_qsdd_so_d...	Quận Long Biên	null	null	7.5531912E7	ngo_2_o_to_tranh
[61cb05a6db87b0001...	60	1	1	0	khac	gan_nhieu_tien_ic...	1	null	null	giay_cn_qsdd_so_d...	Quận Long Biên	null	Phường Ngọc Lâm	7.0E7	ngo_1_o_to
[61cb018e69dd74001...	52	null	4	0	khac	khac	4	null	null	giay_cn_qsdd_so_d...	Quận Đống Đa	Đường Láng	null	9.0384616E7	ngo_o_to_do_cua
[61cb09ab21ed94001...	55	3	5	0	khac	nhieu_mat_thoang ...	5	null	null	giay_cn_qsdd_so_d...	Quận Đống Đa	Đường Nguyễn Lươ...	Phường Nam Đồng	9.818181E7	ngo_ngach
[61cb093f69dd74001...	40	null	null	0	khac	khac	5	null	null	giay_cn_qsdd_so_d...	Quận Hai Bà Trưng	null	null	1.05E8	ngo_ngach
[61cb05e469dd74001...	37	4	3	0	khac	khac	5	null	null	giay_cn_qsdd_so_d...	Quận Nam Từ Liêm	Đường Lê Quang Đạo	null	7.8378376E7	ngo_o_to_do_cua

only showing top 7 rows

Took 3 min 10 sec. Last updated by anonymous at January 04 2022, 4:38:04 PM. (outdated)

```
%pyspark
test.printSchema()

root
 |-- ID: string (nullable = true)
 |-- area_value: long (nullable = true)
 |-- bathroom: long (nullable = true)
 |-- bedroom: long (nullable = true)
 |-- creatorType: long (nullable = true)
 |-- direction: string (nullable = true)
 |-- feature: string (nullable = true)
 |-- floor: long (nullable = true)
 |-- geolocation_lat: double (nullable = true)
 |-- geolocation_lon: double (nullable = true)
 |-- legalPaper: string (nullable = true)
 |-- location_district: string (nullable = true)
 |-- location_street: string (nullable = true)
 |-- location_ward: string (nullable = true)
 |-- price_m2: double (nullable = true)
 |-- wideRoad: string (nullable = true)
```

Took 0 sec. Last updated by anonymous at January 04 2022, 4:43:16 PM.

Thống kê dữ liệu sử dụng hàm describe():

%spark

test.describe().show()

FINISHED

summary	ID	area_value	bathroom	bedroom	creatorType	direction	feature	floor	geolocation_lat	geolocation_lon	legalPaper	location_district	location_street	location_ward	price_m2	wideRoad
count	28776	28514	8274	12228	28776	28780	28780	18619	11895	11897	28780	28776	14568	17595	28773	13547
mean	null	126.1178787224347	3.082054623958182	3.5793605364297982	0.186188028792225	null	null	5.175534419436858	105.80118441419512	21.081211415658943	null	null	null	null	8.9389181535423467	null
stddev	null	5625.648553698163	3.183145740485531	5.286437822918684	0.4956497376143683	null	null	3.5831864327792457	0.8334798183333795	0.5422649407186582	null	null	null	null	1.276582077227212668	null
min	5ef6c1d689496768c...	1	0	1	-1			1	18.4191794	-33.9192167		Huyện Ba Vì	Phố 72	Phường Biền Giang	2480.0	mat_pho_mat_duong
max	61ce62d21ed04081...	738393	108	432	6	tay_nam_tay_bac	xay_nha_chac_chan...	57	107.05737227639642	21.2573126	khac	Thị xã Sơn Tây	Đường Đồng Ngọc	Xã Đức Giang	1.0882353152E10	ngo_o_to_do_cua

Took 44 sec. Last updated by anonymous at January 04 2022, 4:48:25 PM. (outdated)

%spark

test.createOrReplaceTempView("nha")

sqlContext.sql("SELECT count(*) FROM nha").show()

FINISHED

count(1)	28780
----------	-------

Took 52 sec. Last updated by anonymous at January 04 2022, 4:10:24 PM.

Như ta thấy, có khá nhiều ngoại lệ khi min và max của các trường price, area, floor,.. đã vượt ngoài khoảng visualize trước đó.

Loại bỏ các ngoại lai sử dụng hàm filter:

test_filter = test_fill.filter(test_fill.area_value.between(10,300) & test_fill.bathroom.between(0,15) & test_fill.bedroom.between(0,15) & test_fill.floor.between(0,15) & test_fill.price_m2.between(500000,800000000))

test_filter.describe(["area_value","bathroom","bedroom","creatorType","direction","floor","geolocation_lat","geolocation_lon"]).show()

summary	area_value	bathroom	bedroom	creatorType	direction	floor	geolocation_lat	geolocation_lon
count	19123	19123	19123	19123	19123	19123	19123	19123
mean	64.94017675050986	3.4881033310673013	3.5848977670867543	0.1043769283062281	null	4.328923286095278	105.80901161262734	21.006148389736776
stddev	33.04934266250587	1.1271539713309884	1.2914180635038344	0.4396221463728421	null	1.0763421004728577	0.6329524146778739	0.39807750293930505
min	10	0	1	-1		1	18.4191794	-33.9192167
max	300	15	15	6	tay tay nam	15	106.1554873	21.2573126

Took 55 sec. Last updated by anonymous at January 04 2022, 9:49:07 PM.

⇒ Kết quả loại bỏ được khoảng hơn 1000 bản ghi (còn lại **19123**):

Tiếp đó, sử dụng fill missing các trường giá trị rời rạc bằng “khac”, các trường liên tục ra sẽ fill theo mode, mean, median,...(thực nghiệm trên đánh giá)

Cuối cùng, ta thực hiện cache dữ liệu vào RAM và Disk để thuận tiện tính toán:

%spark

#cache

test_filter.cache()

FINISHED

DataFrame[ID: string, area_value: bigint, bathroom: bigint, bedroom: bigint, creatorType: bigint, direction: string, feature: string, floor: bigint, geolocation_1 at: double, geolocation_lon: double, legalPaper: string, location_district: string, location_street: string, location_ward: string, price_m2: double, typeOfHouse: string, wideRoad: string]

Took 50 sec. Last updated by anonymous at January 04 2022, 9:49:08 PM. (outdated)

Tiền xử lý dữ liệu:

Ta chuyển đổi các thuộc tính rời rạc như tên phường, tên quận, loại nhà,... sang các chỉ số index sử dụng **StringIndexer** :

location_district	location_ward	location_street	direction	typeOfHouse	wideRoad
Quận Thanh Xuân	Phường Nhân Chính	Đường Vũ Trọng Phụng	dong dong nam nam...	nha_o	ngo_ngach
Quận Thanh Xuân	Phường Kim Giang	Đường Kim Giang	khac	nha_o	khac
Quận Đống Đa	Phường Phương Liên	Đường Đặng Văn Ngữ	khac	nha_o	khac
Quận Hoàng Mai	Phường Đại Kim	Đường Kim Giang	khac	nha_o	khac
Quận Ba Đình	Phường Ngọc Hà	Đường Hoàng Hoa Thám	khac	nha_o	khac
Quận Hà Đông	Phường Phú La	Đường Văn La	khac	nha_o	khac
Quận Đống Đa	Phường Cát Linh	Đường Cát Linh	khac	nha_o	mat_pho_mat_duong
Quận Thanh Xuân	khac	Đường Hoàng Văn Thái	khac	nha_o	ngo_2_o_to_tranh
Quận Hà Đông	Phường Văn Quán	Đường 19/5	khac	nha_o	khac
Quận Đống Đa	Phường Hàng Bột	Phố Tôn Đức Thắng	khac	nha_o	ngo_ngach

only showing top 10 rows

location_district_index	location_ward_index	location_street_index	direction_index	typeOfHouse_index	wideRoad_index
1.0	1.0	75.0	9.0	0.0	1.0
1.0	92.0	28.0	0.0	0.0	0.0
2.0	95.0	136.0	0.0	0.0	0.0
5.0	11.0	28.0	0.0	0.0	0.0
10.0	82.0	16.0	0.0	0.0	0.0
0.0	62.0	340.0	0.0	0.0	0.0
2.0	117.0	215.0	0.0	0.0	5.0
1.0	0.0	24.0	0.0	0.0	3.0
0.0	21.0	82.0	0.0	0.0	0.0
2.0	61.0	25.0	0.0	0.0	1.0

only showing top 10 rows

Tiếp đó ta tiến hành concat các thuộc tính giá trị số thành vector sử dụng **VectorAssembler**

area_value	bathroom	bedroom	floor	creatorType	geolocation_lat	geolocation_lon	features_number_concat
40	4	4	5	0	105.8097244	20.9959819	[40.0,4.0,4.0,5.0...
45	1	1	1	0	105.8184955468418	21.010621050323614	[45.0,1.0,1.0,1.0...
48	4	4	4	1	105.8299495	21.0180725	[48.0,4.0,4.0,4.0...
45	1	1	1	0	105.8184955468418	21.010621050323614	[45.0,1.0,1.0,1.0...
33	4	4	4	0	105.8140539	21.0337815	[33.0,4.0,4.0,4.0...
50	4	4	4	0	105.7563658	20.955835	[50.0,4.0,4.0,4.0...
36	4	4	6	0	105.8184955468418	21.010621050323614	[36.0,4.0,4.0,6.0...
57	4	4	3	0	105.8097244	20.9959819	[57.0,4.0,4.0,3.0...
31	4	4	4	1	105.7563658	20.955835	[31.0,4.0,4.0,4.0...
52	4	4	4	0	105.8184955468418	21.010621050323614	[52.0,4.0,4.0,4.0...

only showing top 10 rows

Trước khi vào mô hình, ta cần chuẩn hóa giá trị các trường trong dữ liệu của ta về khoảng nhỏ hơn sử dụng **StandardScaler**

```
%pyspark
model_trans.select(["features_number_concat","scaledFeatures"]).show(5)
```

```
+-----+-----+
|features_number_concat|scaledFeatures|
+-----+-----+
| [40.0,4.0,4.0,5.0...|[1.21031151537484...|
| [45.0,1.0,1.0,1.0...|[1.36160045479670...|
| [48.0,4.0,4.0,4.0...|[1.45237381844981...|
| [45.0,1.0,1.0,1.0...|[1.36160045479670...|
| [33.0,4.0,4.0,4.0...|[0.99850700018424...|
+-----+-----+
```

only showing top 5 rows

Took 0 sec. Last updated by anonymous at January 04 2022, 11:21:30 PM.

Ngoài ra, các trường text mang ý nghĩa như feature, legalpaper,.. ta sẽ mã hóa bằng

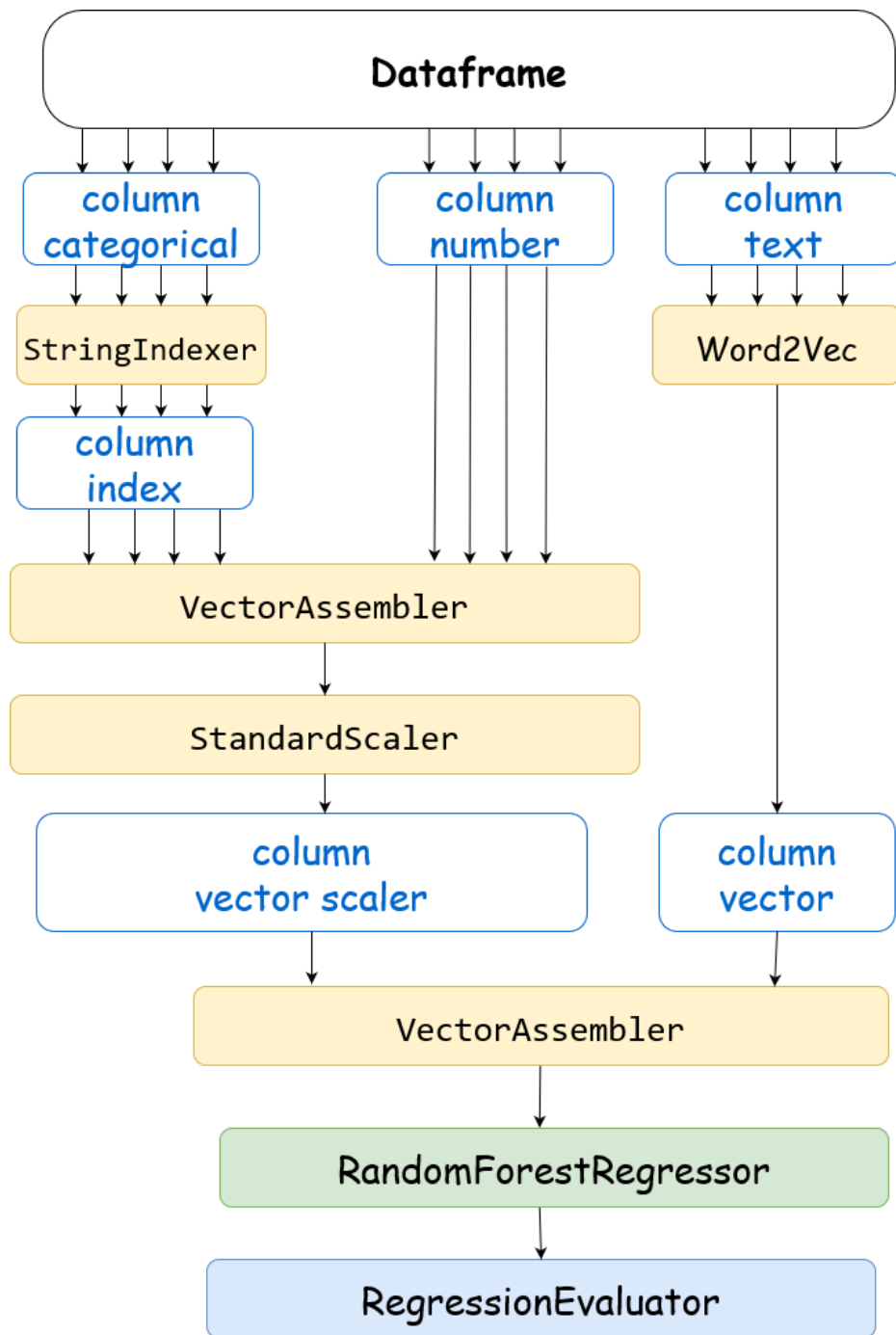
Word2Vec:

```
+-----+-----+
|feature_split|legalPaper_split|
+-----+-----+
|[mat, tien, rong,...|[giay, cn, qsdd, ...|
|[gan, nhieu, tien...|[giay, cn, qsdd, ...|
| [khac]| [khac]|
|[gan, nhieu, tien...|[giay, cn, qsdd, ...|
| [khac]| [khac]|
| [khac]| [khac]|
| [khac]| [giay, cn, qsdd, ...|
| [khac]| [giay, cn, qsdd, ...|
| [khac]| [khac]|
| [khac]| [giay, cn, qsdd, ...|
+-----+-----+
```

```
+-----+-----+
|feature_split_vector|legalPaper_split_vector|
+-----+-----+
|[0.03806417062878...|[0.64834641984530...|
|[-0.5463381069047...|[0.64834641984530...|
|[0.15623758733272...|[0.01179112959653...|
|[-0.5463381069047...|[0.64834641984530...|
|[0.15623758733272...|[0.01179112959653...|
|[0.15623758733272...|[0.01179112959653...|
|[0.15623758733272...|[0.64834641984530...|
|[0.15623758733272...|[0.64834641984530...|
|[0.15623758733272...|[0.01179112959653...|
|[0.15623758733272...|[0.64834641984530...|
+-----+-----+
```

only showing top 10 rows

Pipeline Transform & Model




```

from pyspark.ml.regression import RandomForestRegressor
from pyspark.ml.feature import VectorIndexer
from pyspark.ml.evaluation import RegressionEvaluator

stagelist=[]
#for number

#for string indexes
input_cols_number = ["area_value", "bathroom", "bedroom", "floor", "creatorType", "geolocation_lat", "geolocation_lon"]
input_cols_str_index = ["location_district", "location_ward", "location_street", "direction", "typeOfHouse", "wideRoad"]
output_cols_str_index = []

for inputCol in input_cols_str_index:
    output_cols_str_index.append(inputCol+"_index")
    input_cols_number.append(inputCol+"_index")
    indexer = StringIndexer(inputCol=inputCol, outputCol=inputCol+"_index", handleInvalid="keep")
    stagelist.append(indexer)

#for standarScaler

assembler_number = VectorAssembler(inputCols=input_cols_number, outputCol="features_number_concat")
stagelist.append(assembler_number)

scaler = StandardScaler(inputCol=assembler_number.getOutputCol(), outputCol="scaledFeatures", withStd=True, withMean=False)
stagelist.append(scaler)

#for string vector
input_cols_str_vector = ["feature_split", "legalPaper_split"]

for inputCol in input_cols_str_vector:
    outputs.append(inputCol+"_vector")
    input_cols_number.append(inputCol+"_vector")
    indexer = Word2Vec(vectorSize=3, minCount=0, inputCol=inputCol, outputCol=inputCol+"_vector")
    stagelist.append(indexer)

#concat

assembler_vec_all = VectorAssembler(inputCols=[scaler.getOutputCol(), "feature_split_vector", "legalPaper_split_vector"], outputCol="features_last_concat")
stagelist.append(assembler_vec_all)

rf = RandomForestRegressor(featuresCol="features_last_concat", labelCol="log2_price_m2")
stagelist.append(rf)

pipeline = Pipeline(stages=stagelist)

#split data
(trainingData, testData) = test_filter.randomSplit([0.7, 0.3])
# Fit the pipeline to training documents.
model = pipeline.fit(trainingData)

```

Đánh giá kết quả dự đoán:

```

+-----+-----+-----+
|features_last_concat|    log2_price_m2|    prediction|
+-----+-----+-----+
|[2.36481689453352...|25.262834519483164|25.007983762246774|
|[2.45695261769717...| 25.02157272720836|25.050307402379808|
|[3.31688603389118...|25.520976970637463|25.019991151271704|
|[1.81200255555166...|24.538335426537852| 24.96367043156377|
|[3.00976695667903...|25.246117193033427|25.019991151271704|
+-----+-----+-----+

```

only showing top 5 rows

Root Mean Squared Error (RMSE) on test data = 0.543146854897787

R Squared on test data = 0.7848733338449783

Mean Absolute Error(MAE) on test data = 0.39390784214697216

RandomForestRegressionModel (uid=RandomForestRegressor_45c489bcdddbc596682e) with 20 trees

Took 1 sec. Last updated by anonymous at January 05 2022, 1:22:03 AM.