

Lưu trữ xử lý dữ liệu lớn

ONE LOVE. ONE FUTURE.



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

Hệ thống crawl dữ liệu nhà đất Visualize, analyze dữ liệu thu thập

Nhóm Anh em ới:

Lường Hoàng Đức 20183712

Phùng Minh Hiếu 20183536

Phạm Văn Hạnh 20183525

Bùi Mạnh Trường 20183646

ONE LOVE. ONE FUTURE.

Nội dung

Giới thiệu

Công nghệ sử dụng

Tổng quan hệ thống

Triển khai hệ thống

Kết quả và Visualization

TOP 4 QUẬN

ĐĂNG BÁN NHIỀU NHÀ NHẤT QUÝ 4

Đồng Đa Hà Đông Thanh Xuân Hoàng Mai



TOP 3 QUẬN

CÓ LƯỢNG TÌM KIẾM CAO NHẤT QUÝ 4

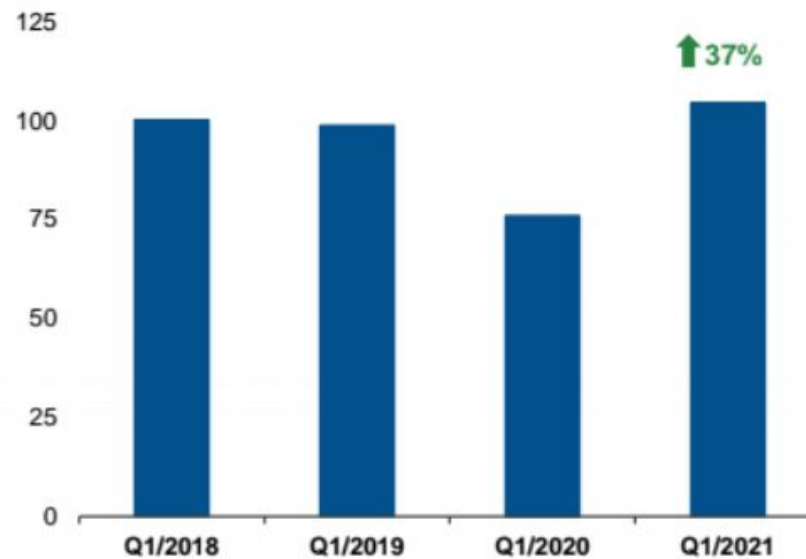




MỨC ĐỘ QUAN TÂM TỚI BẤT ĐỘNG SẢN TĂNG CAO NHẤT TRONG NHIỀU NĂM QUA

Tăng trưởng mức độ quan tâm

Đơn vị: index, Q1/2018 = 100 điểm



Alonhadat.com.vn

- Dữ liệu tin Hà Nội
- Gồm các thông tin giao bán của các loại bất động sản
- Giá, quận, diện tích, ...

Các source khác trong giai đoạn phát triển tiếp theo:

- Facebook.com
- Batdongsan.com
- Chotot.com

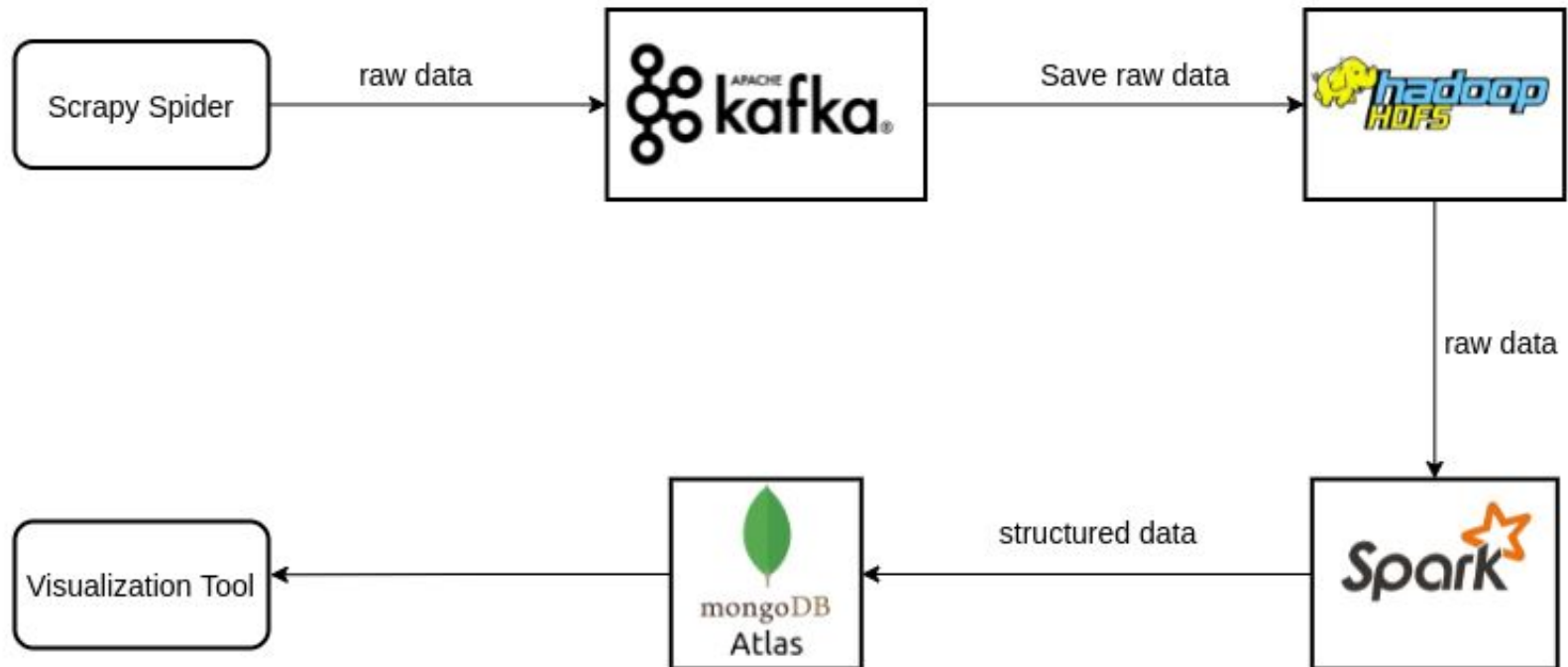




- Scrapy
- Kafka
- HDFS
- Spark
- MongoDB
- Charts Visualization



Tổng quan hệ thống



Chi tiết hệ thống

- 1 Scrapy Spider để crawl dữ liệu từ trang alonhadat.com
- Cụm Apache Kafka gồm: 1 Zookeeper node, 3 kafka nodes tương ứng với 3 broker
- Cụm HDFS gồm: 1 namenode, 1 datanode, 1 resource manager node, 1 node manager, 1 history node
- Cụm Spark gồm: 1 master node, 2 worker nodes (2 Cores, 3GB Ram/worker node)
- Cụm MongoDB Atlas được cấu hình Replica Set là 3 nodes
- Visualization Tool: Mongodb Charts

Item Pipeline - Scrapy

```
class AlonhadatPipeline:
    def process_item(self, item: AlonhadatNewsItem, spider: Spider):

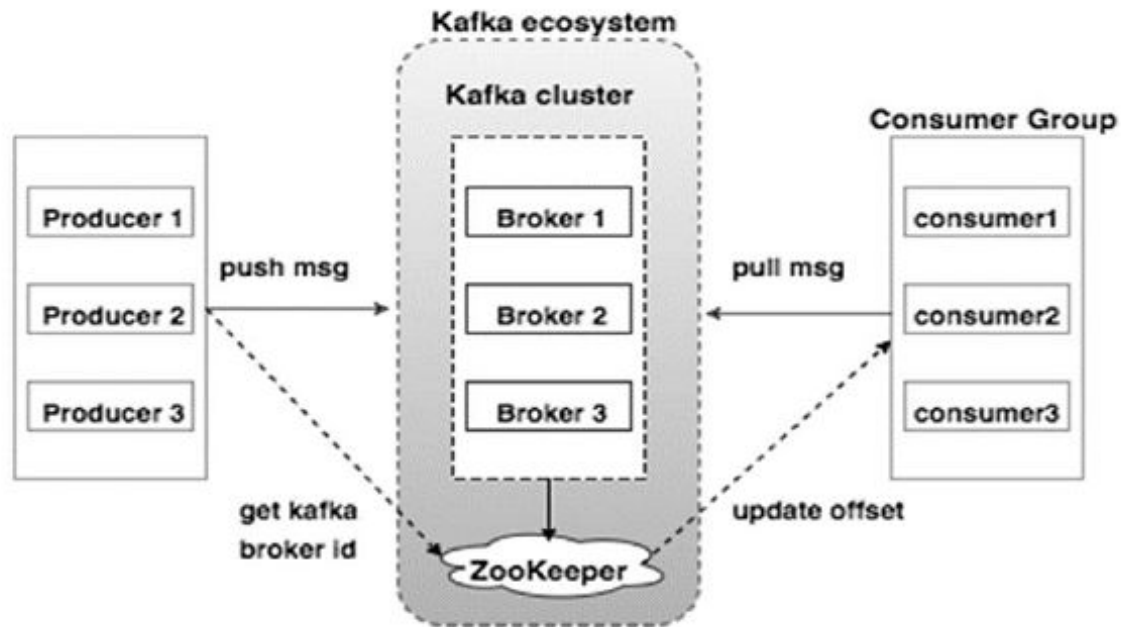
        def delivery_report(err, msg):
            """ Called once for each message produced to indicate delivery result.
                Triggered by poll() or flush(). """
            if err is not None:
                print('Message delivery failed: {}'.format(err))
            else:
                print('Message delivered to {} [{}]'.format(msg.topic(), msg.partition()))

        spider.producer.poll(0)
        spider.producer.produce(spider.topic,
                                json.dumps(dataclasses.asdict(item)).encode('utf-8'),
                                callback=delivery_report)
        spider.producer.flush()
```

Kafka

Kafka cluster: 1 Zookeeper node, 3 kafka nodes

Docker-compose file: [link](#)



Create **crawled_news** topic với cấu hình 3 partitions, 3 replicas

Kafka: Kafka connection

Topics

Consumers

Add producer

Add consumer

Name	Replicas	Partitions	In sync replicas	Replication factor	Under replicated partitions
crawled_news	9	3	9	3	0
mytopic	1	1	1	1	0

Partitions

Configuration

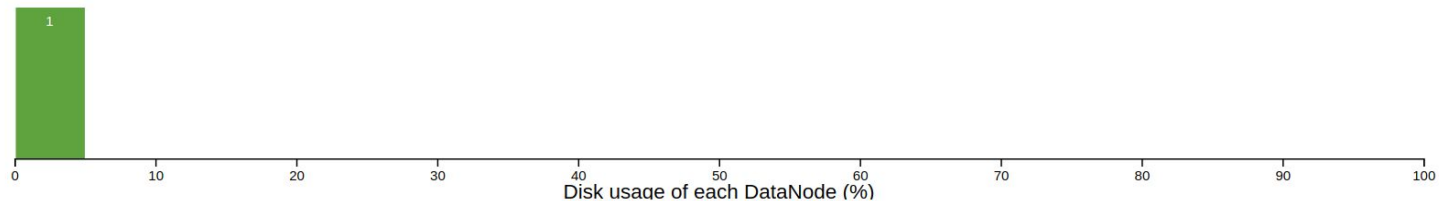
Partition id	Leader	In sync replicas count	Replicas count
0	3	3	3
1	1	3	3
2	2	3	3

HDFS Cluster

Xây dựng HDFS cluster với: **1 namenode**, **1 datanode**, 1 resource manager node, 1 node manager, 1 history node

Docker-compose file: [link](#) ; Monitoring: [link](#)

Datanode usage histogram



In operation

Show entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ 7f5946e6547b:9866 (172.22.0.3:9866)	http://7f5946e6547b:9864	0s	28m	468.44 GB <div><div></div></div>	64811	6.94 GB (1.48%)	3.2.1

Showing 1 to 1 of 1 entries

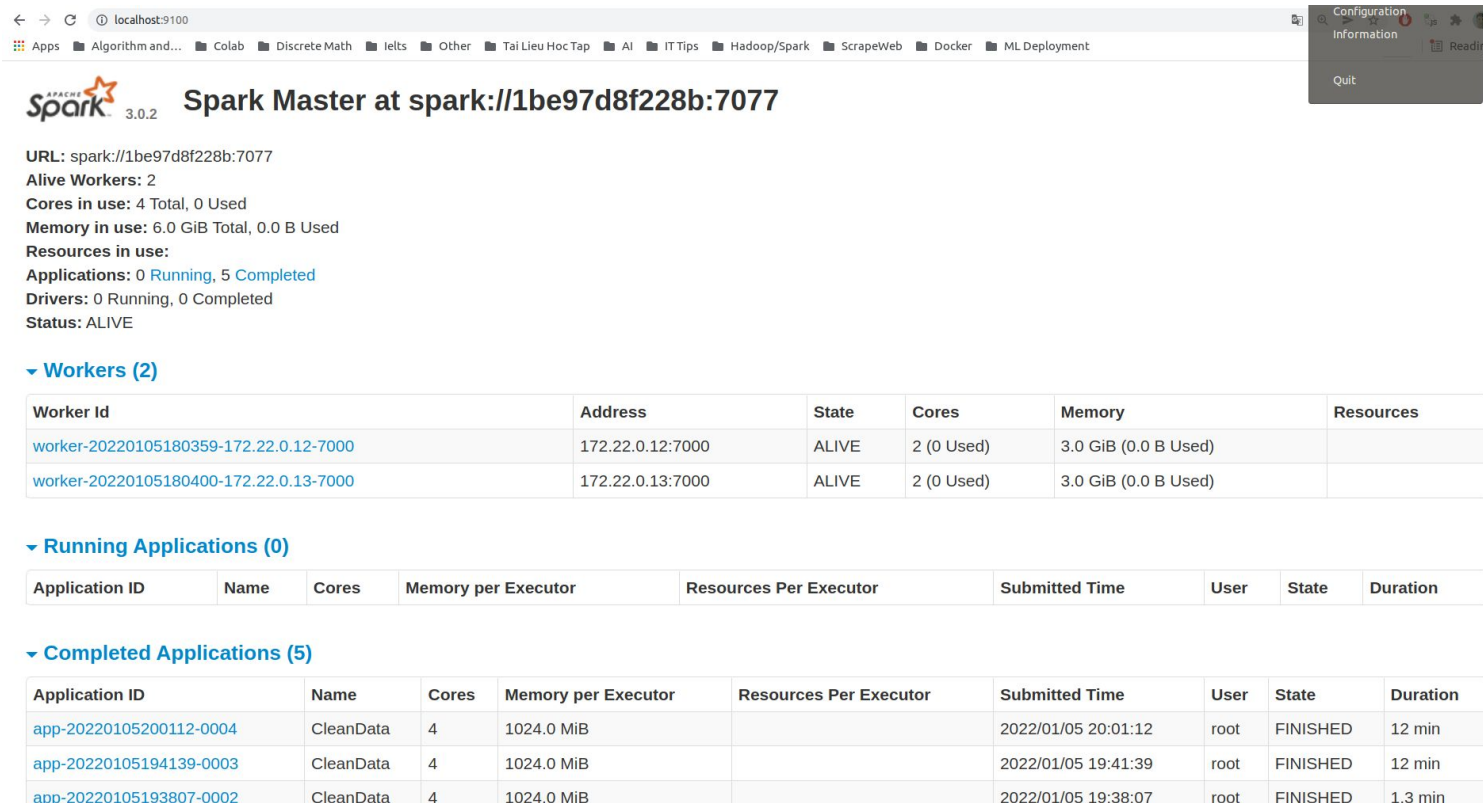
Previous **1** Next



Spark Cluster

Xây dựng Spark cluster với: **1 master node, 2 worker nodes** (2 Cores, 3GB Ram/ worker node)

Docker-compose file: [link](#); Monitoring: [link](#)



← → 📄 localhost:9100

Apps Algorithm and... Colab Discrete Math Ielts Other Tài Liệu Học Tập AI IT Tips Hadoop/Spark ScrapeWeb Docker ML Deployment

Spark Master at spark://1be97d8f228b:7077

URL: spark://1be97d8f228b:7077
Alive Workers: 2
Cores in use: 4 Total, 0 Used
Memory in use: 6.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 5 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Configuration Information Reading Quit

▼ Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20220105180359-172.22.0.12-7000	172.22.0.12:7000	ALIVE	2 (0 Used)	3.0 GiB (0.0 B Used)	
worker-20220105180400-172.22.0.13-7000	172.22.0.13:7000	ALIVE	2 (0 Used)	3.0 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 (5)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20220105200112-0004	CleanData	4	1024.0 MiB		2022/01/05 20:01:12	root	FINISHED	12 min
app-20220105194139-0003	CleanData	4	1024.0 MiB		2022/01/05 19:41:39	root	FINISHED	12 min
app-20220105193807-0002	CleanData	4	1024.0 MiB		2022/01/05 19:38:07	root	FINISHED	1.3 min

Clean Data Job

Code: [link](#)

Dữ liệu thô được trích xuất thành các trường: **Url, Title, Realestate_Type, Area (m2), Price (B), m/m2, Location, District, Province**

Trong đó trường **Realestate_type** có thể có các giá trị:

nha-mat-tien, nha-trong-hem, biet-thu-nha-lien-ke, can-ho-chung-cu, phong-tro-nha-tro, van-phong, kho-xuong, nha-hang-khach-san, shop-kiot-quan, trang-trai, mat-bang, dat-tho-cu-dat-o, dat-nen, dat-nong-lam-nghiep, cac-loai-khac

Sử dụng spark-submit để chạy job:

```
“/opt/spark/bin/spark-submit --master spark://spark-master:7077 --archives  
/opt/spark-data/bigdata_venv.tar.gz#enviroment /opt/spark-data/cleandata.py”
```

- **Crawling**

- Nhóm đã crawl được **48582** tin đăng bán bất động sản

```
Using Python version 3.7.3 (default, Jan 22 2021 20:04:44)
SparkSession available as 'spark'.
>>> raw_df = spark.read.json("hdfs://namenode:9000/alonhadatnews_json")
22/01/06 04:08:24 WARN SharedInMemoryCache: Evicting cached table partition metadata from memory d
leCacheSize = 262144000 bytes). This may impact query planning performance.
>>> raw_df.count()
48582
>>> 
```

- **Lưu trữ**

- Lượng dữ liệu tổng cộng lưu trong HDFS ~ **7GB** ([link](#))

- **Spark job**
 - Chương trình spark job được thực thi, dữ liệu được đưa vào MongoDB



3.0.2

Spark Master at spark://1be97d8f228b:7077

Quit

URL: spark://1be97d8f228b:7077

Alive Workers: 2

Cores in use: 4 Total, 0 Used

Memory in use: 6.0 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 [Running](#), 5 [Completed](#)

Drivers: 0 Running, 0 Completed

Status: ALIVE

▼ Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20220105180359-172.22.0.12-7000	172.22.0.12:7000	ALIVE	2 (0 Used)	3.0 GiB (0.0 B Used)	
worker-20220105180400-172.22.0.13-7000	172.22.0.13:7000	ALIVE	2 (0 Used)	3.0 GiB (0.0 B Used)	

▼ Running Applications (0)

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

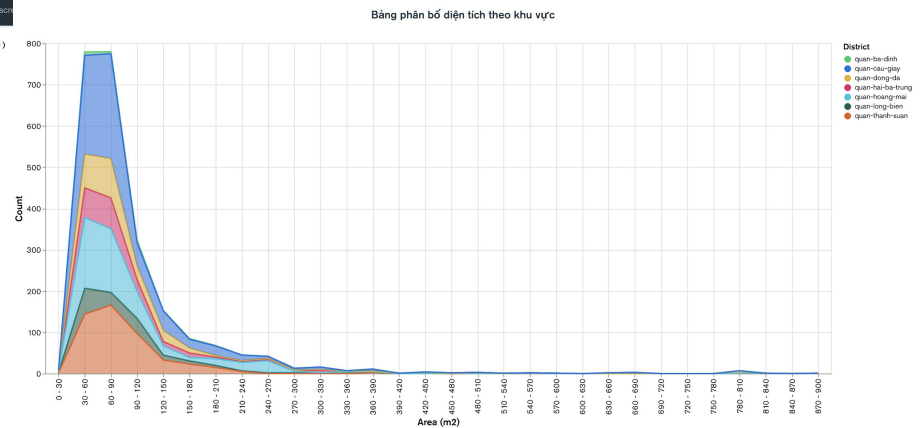
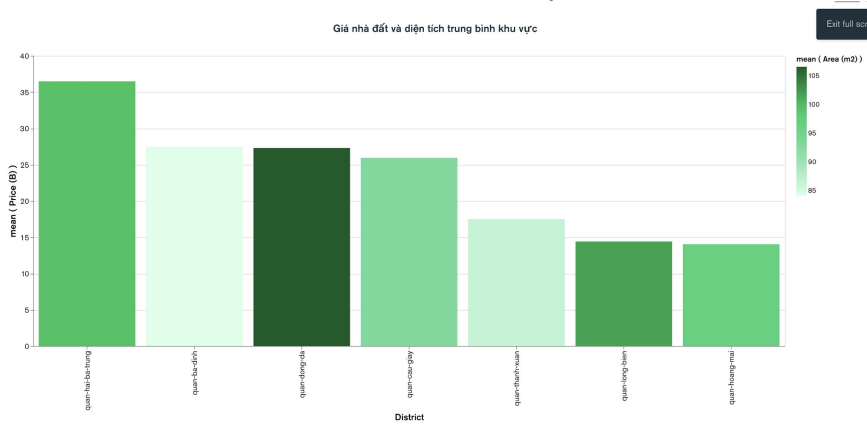
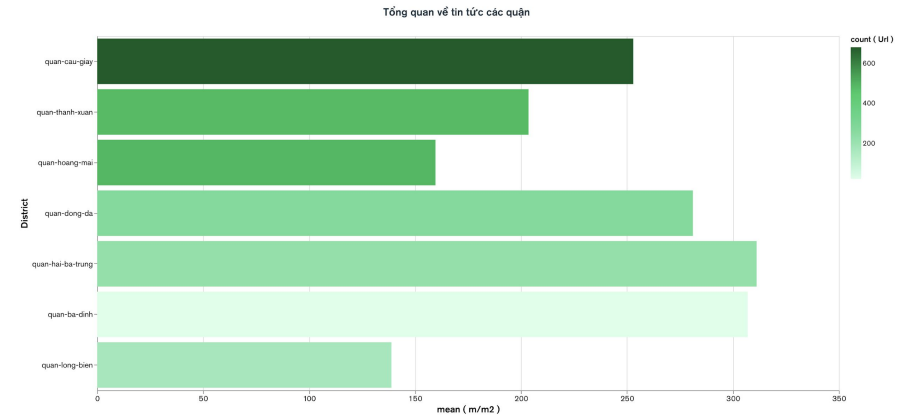
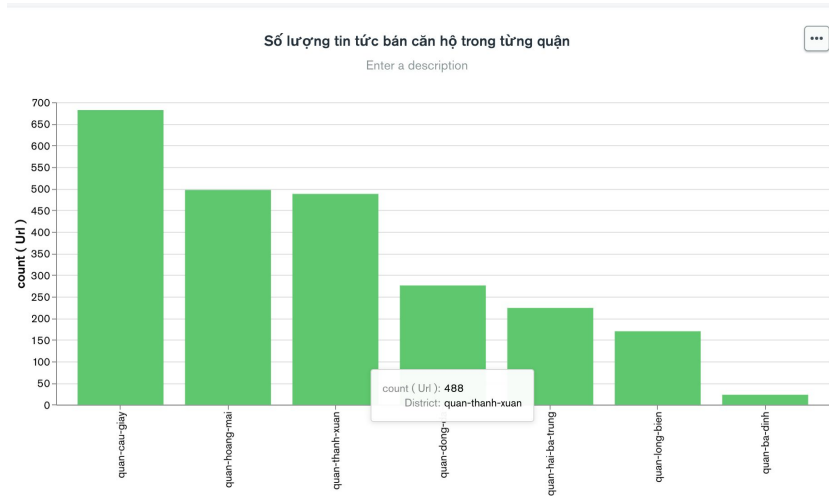
▼ Completed Applications (5)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20220105200112-0004	CleanData	4	1024.0 MiB		2022/01/05 20:01:12	root	FINISHED	12 min
app-20220105194139-0003	CleanData	4	1024.0 MiB		2022/01/05 19:41:39	root	FINISHED	12 min
app-20220105193807-0002	CleanData	4	1024.0 MiB		2022/01/05 19:38:07	root	FINISHED	1.3 min

Dữ liệu sạch trên MongoDB

🔍 cleaned_data											
	_id ObjectID	Url String	Title String	Type String	Area (m2) Double	Price (B) Double	m/m2 Double	Location String			
1	61d60c2afc7724b04c568d2f	"https://alohadat.com.vn/bai	"Bán \GẤP\ nhà Phố Vương Tl	"nha-mat-tien"	51	5.7	111.76470588235294	"Thanh Xuân Phố	🔍	🔍	🔍
2	61d60c2afc7724b04c568d30	"https://alohadat.com.vn/bai	"Bán nhà mới Hồ Tùng Mậu 125r	"nha-mat-tien"	125	17.5	140	"Đường Mai Dịch,	🔍	🔍	🔍
3	61d60c2afc7724b04c568d31	"https://alohadat.com.vn/bai	"Bán nhà phân lô 147 Tân Mai,	"nha-mat-tien"	250	5.8	23.2	"250 Phố Vong, 2	🔍	🔍	🔍
4	61d60c2afc7724b04c568d32	"https://alohadat.com.vn/hi	"HIỆM TOÀ VP MP HOÀNG CẦU VII	"nha-mat-tien"	96	30	312.5	"Phố Hoàng Cầu,	🔍	🔍	🔍
5	61d60c2afc7724b04c568d33	"https://alohadat.com.vn/ma	"MP VĨA HÈ RỘNG 10M TRUNG KỈ	"nha-mat-tien"	75	26	346.6666666666667	"Phố Trung Kinh,	🔍	🔍	🔍
6	61d60c2afc7724b04c568d34	"https://alohadat.com.vn/anl	"ẢNH THẬT Bán nhà 521 Trương	"nha-mat-tien"	250	5.8	23.2	"Ngõ 24 Kim Đồng	🔍	🔍	🔍
7	61d60c2afc7724b04c568d35	"https://alohadat.com.vn/bai	"Bán nhà phố Kim Đồng, Giáp I	"nha-mat-tien"	55	7.9	143.63636363636365	"Phố Kim Đồng, P	🔍	🔍	🔍
8	61d60c2afc7724b04c568d36	"https://alohadat.com.vn/lk	"LK SHOPHOUSE MẶT PHỐ TÔN TH	"nha-mat-tien"	120	63	525	"Phố Trần Thái T	🔍	🔍	🔍
9	61d60c2afc7724b04c568d37	"https://alohadat.com.vn/dai	"ĐẮC ĐỊA HIẾM CÓ MẶT PHỐ P.P	"nha-mat-tien"	315	138	438.0952380952381	"Đường Phạm Đình	🔍	🔍	🔍
10	61d60c2afc7724b04c568d38	"https://alohadat.com.vn/bi	"BIỆT THỰ CẦU GIẤY, PHÂN LÔ,	"nha-mat-tien"	85	10.8	127.05882352941178	"Đường Hồ Tùng M	🔍	🔍	🔍
11	61d60c2afc7724b04c568d39	"https://alohadat.com.vn/bai	"Bán nhà mặt phố Trần Đại Ng	"nha-mat-tien"	26	5.5	211.53846153846155	"Đường Sóng Sét,	🔍	🔍	🔍
12	61d60c2afc7724b04c568d3a	"https://alohadat.com.vn/bai	"Cần bán gấp nhà mặt phố kinl	"nha-mat-tien"	66	26	393.93939393939394	"Đường Trần Duy	🔍	🔍	🔍
13	61d60c2afc7724b04c568d3b	"https://alohadat.com.vn/bai	"BÁN NHÀ NGÃ TƯ SỐ 50M2x5T M	"nha-mat-tien"	50	4.3	86	"Đường Nguyễn Tr	🔍	🔍	🔍
14	61d60c2afc7724b04c568d3c	"https://alohadat.com.vn/si	"Siêu Giá Trị Mặt Phố Vành Đ	"nha-mat-tien"	250	62	248	"Phố Tân Mai, Ph	🔍	🔍	🔍
15	61d60c2afc7724b04c568d3d	"https://alohadat.com.vn/vi	"VỊ TRÍ VIP MẶT PHỐ TUỆ TỈNH	"nha-mat-tien"	145	69	475.86206896551727	"Phố Tuệ Tĩnh, P	🔍	🔍	🔍
16	61d60c2afc7724b04c568d3e	"https://alohadat.com.vn/vi	"ĐẦU TƯ VIP MẶT PHỐ MINH KH	"nha-mat-tien"	2.015	250	124069.47890818857	"Đường Minh Khai	🔍	🔍	🔍
17	61d60c2afc7724b04c568d3f	"https://alohadat.com.vn/hi	"HIẾM BÁN- NHÀ PHÂN LÔ QĐ - I	"nha-mat-tien"	105	14	133.33333333333334	"Phố Hoàng Văn T	🔍	🔍	🔍
18	61d60c2afc7724b04c568d40	"https://alohadat.com.vn/bai	"Bán Nhà Phố Hoàng Quốc Việt	"nha-mat-tien"	70	15	214.28571428571428	"Đường Hoàng Quố	🔍	🔍	🔍
19	61d60c2afc7724b04c568d41	"https://alohadat.com.vn/bai	"Bán nhà 48m2x5T Tân Mai, Hoi	"nha-mat-tien"	250	5.8	23.2	"250 Phố Vong, 2	🔍	🔍	🔍
20	61d60c2afc7724b04c568d42	"https://alohadat.com.vn/nh	"NHÀ MẶT PHỐ-ÔTÔ ĐỖ CỬA-KINH	"nha-mat-tien"	67	12	179.1044776119403	"2.9 . Phố Vong,	🔍	🔍	🔍

Charts Visualization



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 !