

**ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC BÁCH KHOA**  
**KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH**

-----oOo-----



**BÁO CÁO ĐỒ ÁN**  
**BUILD A SIMPLE WEB CRAWLER**  
**KỸ THUẬT LẬP TRÌNH**

**GVHD: THẦY LƯU QUANG HUÂN**

<b>NHÓM B2</b>	
<b>Sinh viên thực hiện</b>	<b>Mã số sinh viên</b>
Trần Phương Tĩnh	1927038
Lê Tất Thiện	1920058
Dương Quang Tuấn	1920068
Nguyễn Đức Tuấn	1927040

TP. HỒ CHÍ MINH, tháng 12 năm 2020

# Mục lục

1. <a href="#"><u>Giới thiệu đề tài</u></a>	<b>01</b>
2. <a href="#"><u>Giới thiệu công nghệ</u></a>	<b>03</b>
3. <a href="#"><u>Deploy</u></a>	<b>10</b>
4. <a href="#"><u>Hướng dẫn sử dụng</u></a>	<b>12</b>
5. <a href="#"><u>Báo cáo hàng tuần</u></a>	<b>14</b>
6. <a href="#"><u>Lịch sử commit code trên github</u></a>	<b>19</b>
7. <a href="#"><u>Tài liệu tham khảo</u></a>	<b>29</b>

# 1. Giới thiệu đề tài

## MỤC ĐÍCH - YÊU CẦU

Deep learning, Machine Learning hay AI đã và đang là một chủ đề đang được quan tâm với nhiều ứng dụng trong thực tiễn và các nghiên cứu mang tính học thuật cao. Tuy nhiên, việc triển khai các ứng dụng, các mô hình cũng như việc tiến hành nghiên cứu gặp nhiều khó khăn. Một mặt, do các bộ dữ liệu huấn luyện hạn chế, về cả số lượng và không đảm bảo về chất lượng. Các bộ dữ liệu hiện có cũng hạn chế, đa phần bằng các ngôn ngữ không phải tiếng Việt.

Xuất phát từ nhu cầu thực tế của việc thực hiện tác đề tài, đồ án môn học sau này hoặc khi làm LVTN liên quan đến dữ liệu, sinh viên gặp nhiều khó khăn trong việc thu thập dữ liệu mẫu (Dataset) để triển khai các mô hình Machine Learning, Deep Learning hoặc cơ bản hơn là thu thập dữ liệu làm dữ liệu mẫu cho một ứng dụng, hệ thống đã xây dựng. Trong bài tập lớn này, SV sẽ có cơ hội hiện thực việc thu thập các dữ liệu này, làm quen với các thao tác tiền xử lý, phân loại, lưu trữ và biểu diễn một tập dữ liệu (Data set).

Mỗi nhóm sẽ được phân công thu thập dữ liệu về một chủ đề từ các website khác nhau.

## CÁCH THỨC TRIỂN KHAI

Project này được thực hiện trong vòng 5 tuần với kế hoạch và kết quả dự kiến như sau:

- Task 2.0: Chốt danh sách đăng ký nhóm, khởi tạo các công cụ làm việc nhóm.
- Task 2.1: Tìm hiểu và hiện thực các công cụ Crawl dữ liệu theo yêu cầu. Thường mỗi nhóm sẽ được phân crawl từ nhiều website khác nhau, do đó các thành viên trong nhóm làm việc để tự phân chia công việc. Kết quả dự kiến tuần 1: source code, hướng dẫn sử dụng, sample dữ liệu đã thu được (~1000 items). Trong quá trình crawl chắc chắn gặp nhiều khó khăn, nhóm cần tự tìm hiểu về những cách sử dụng VPN, proxy, socks để có thể hoàn thành yêu cầu.
- Task 2.2. Triển khai việc Crawl dữ liệu, xây dựng Bộ dataset (~10.000.), cải tiến công cụ (source code đã xây dựng) để tạo thành công cụ hoàn chỉnh cho việc Crawl dữ liệu. Tìm hiểu cách làm sạch, chuẩn hóa dữ liệu, phân loại dữ liệu. Thiết kế CLDS, các định dạng file biểu diễn

dataset. Kết quả dự kiến: dataset, bản thiết kế CSDL, báo cáo tìm hiểu về cách làm sạch, source code nếu có.

- Task 2.3. Kết quả dự kiến: Báo cáo + source code làm sạch dữ liệu, source code lưu trữ dữ liệu raw (dataset raw) vào database, source code truy vấn dữ liệu, model data...
- Task 2.4. Thực hành việc phân tích dữ liệu. Xây dựng một số tính năng phân tích, thống kê, tìm kiếm, export....
- Task 2.5. Báo cáo tổng kết.

Trước 23h59 ngày chủ nhật hàng tuần, các nhóm nộp kết quả từng tuần lên BKEL (qua module Bài tập lớn). Trong đó, sv cần nêu báo cáo văn tắt (dưới 1) trang theo format

- Các công việc đã thực hiện
  - ✧ Tìm hiểu scrapy (Hung) đã có báo cáo tại link (hoặc đã push git)
  - ✧ Implement vnexpresscrawler.py (Huân) đã push code lên git.
  - ✧ ....
- Các khó khăn gặp phải
  - ✧ Session limit, đã khắc phục bằng cách abc... đã viết lại báo cáo hướng dẫn crawl tại link
- Dự kiến công việc tuần tới
  - ✧ a
  - ✧ b
  - ✧ c

Kèm theo đó là đường link chi tiết về báo cáo kĩ thuật, hoặc source code cần được push lên git. Những báo cáo này cần viết và quản lý để tuần thứ 5 tổng hợp thành báo cáo Đồ Án môn học.

## ĐÁNH GIÁ

Cách thức đánh giá chung: kết quả theo từng giai đoạn mà nhóm thu được thông qua source code, dataset và báo cáo. Kết quả đánh giá từng thành viên thông qua log git (do đó cần thường xuyên push code), các tài liệu cũng nên được commit lên git. Kết quả sẽ được GV đánh giá và update từng tuần trên file.

Các nhóm làm việc với nhau dưới sự hướng dẫn của GV chủ yếu thông qua các công cụ online như email, bkel... nếu trong trường hợp chưa rõ, cần hỗ trợ hướng dẫn, SV cần chủ động liên hệ với GV qua email để setup meeting (online hoặc offline) để hiểu rõ yêu cầu và thực hiện.

## 2. Giới thiệu công nghệ

### 2.1. Giới thiệu về Angular



Angular 2 được biết đến tên rộng rãi như hiện tại là Angular thôi nhé. Nó là một framework cho frontend và là bản tiếp theo của AngularJS. Angular là mã nguồn mở giúp chúng ta xây dựng một Single Page Applications (SPAs).

Angular là cung được xây dựng cả ứng dụng Mobile và Desktop. Nó được xây dựng sử dụng JavaScript. Bạn phải sử dụng nó để xây dựng ứng dụng hoàn chỉnh kết hợp với HTML, CSS và JavaScript.

Angular có nhiều cải tiến thông so với AngularJS. Nó có nhiều cải tiến làm dễ học và phát triển ứng dụng cho doanh nghiệp. Bạn có thể xây dựng một ứng dụng dễ dàng mở rộng, bảo trì, test.

## Tính năng của Angular

Angular được load với tính năng Power-packaged. Một số tính năng được liệt kê ra đây như sau:

- Cơ chế Two-Way Data Binding: Đây là tính năng cool nhất của Angular. Data binding tự động và rất nhanh tức là bất cứ thay đổi nào trên view đều được tự động cập nhật vào component class và ngược lại.
- Hỗ trợ cơ chế Routing mạnh mẽ: Angular có cơ chế routing tải trang một cách bắt đồng bộ trên cùng một trang cho phép chúng ta tạo SPA.
- Mở rộng HTML: Angular cho phép chúng ta sử dụng cấu trúc lập trình giống như điều kiện if, vòng lặp for...để render các control.
- Thiết kế module hoá: Angular tiếp cận theo hướng thiết kế module hoá. Bạn phải tạo các Angular Module để tổ chức tốt hơn và quản lý source code.
- Hỗ trợ làm việc với hệ thống Backend: Angular được xây dựng hỗ trợ làm việc với backend server và thực thi bất cứ logic nào và nhận dữ liệu về.
- Cộng đồng tốt: Angular được hỗ trợ bởi Google và cộng đồng.

Angular được thay đổi rất nhiều từ AngularJS. Angular đã thiết kế lại từ đầu nên có nhiều khái niệm đã thay đổi từ AngularJS.

## Ưu điểm

- Angular sử dụng Typescript, Typescript hỗ trợ các cú pháp và hàm chức năng mà javascript sẽ có trong tương lai nên lập trình viên không phải lo quá nhiều về sự thay đổi phiên bản của Javascript sau này.
- Có nhiều thư viện component hỗ trợ angular, giúp việc phát triển nhanh và dễ dàng hơn rất nhiều.
- Tài liệu rõ ràng và đầy đủ, cộng đồng sử dụng angular lớn.

- Hỗ trợ dependency injection là một công cụ quản lý các ràng buộc một cách đơn giản và tối ưu.
- Angular là một framework lớn, các thư viện cần thiết đã được tích hợp sẵn.
- Điều này giúp việc cài đặt angular đơn giản. Đồng thời kiến trúc của angular được phân chia rõ ràng, hỗ trợ tốt cho các ứng dụng phức tạp và yêu cầu khả năng mở rộng.

## Nhược điểm

- Angular là một framework phức tạp nên việc học và làm quen tốn nhiều thời gian.
- Tốc độ chậm hơn so với các framework SPA còn lại.

## 2.2. Hướng dẫn triển khai lên Cloud

Dưới đây là các bước để triển khai một ứng dụng angular lên server Firebase của google. Yêu cầu Git, Nodejs, npm, editor và một tài khoản Firebase.

Bước 1: Clone project về bằng lệnh:

```
git clone https://github.com/thien-lebk/DAKTLT-Frontend.git
```

Bước 2: Cài đặt các thư viện cho project bằng lệnh:

```
npm install
```

Bước 3: Build project bằng lệnh. Sau khi build thành công thì mã nguồn của web sẽ nằm trong thư mục dist/

```
ng build --prod
```

Bước 4: Cài đặt Firebase Cli:

```
npm install -g firebase-tools
```

Bước 5: Login vào tài khoản firebase bằng lệnh:

```
firebase login
```

Bước 6: Khởi tạo file config của firebase:

```
firebase init
```

```
$ firebase init

#####
##      ##      ##      ##      ##      ##      ##      ##      ##
##      ##      ##      ##      ##      ##      ##      ##      ##
#####      ##      #####      ##      #####      ##      #####      ##
##      ##      ##      ##      ##      ##      ##      ##      ##
##      ###### ##      ##      #####      ##      ##      ##      ##

You're about to initialize a Firebase project in this directory:
/home/tuanduong/Documents/VB2/201/KTLT/btl/2/frontend_Tuan/web

? Which Firebase CLI features do you want to set up for this folder? Press Space to select features, then
Enter to confirm your choices.
  o Database: Deploy Firebase Realtime Database Rules
  o Firestore: Deploy rules and create indexes for Firestore
  o Functions: Configure and deploy Cloud Functions
> o Hosting: Configure and deploy Firebase Hosting sites
  o Storage: Deploy Cloud Storage security rules
  o Emulators: Set up local emulators for Firebase features
```

Chọn mục Hosting

```

==> Project Setup

First, let's associate this project directory with a Firebase project.
You can create multiple project aliases by running firebase use --add,
but for now we'll just set up a default project.

? Please select an option:
  Use an existing project
> Create a new project
  Add Firebase to an existing Google Cloud Platform project
  Don't set up a default project
==> Project Setup

First, let's associate this project directory with a Firebase project.
You can create multiple project aliases by running firebase use --add,
but for now we'll just set up a default project.

? Please select an option: Create a new project
i  If you want to create a project in a Google Cloud organization or folder,
   please use "firebase projects:create" instead, and return to this command w
hen you've created the project.
? Please specify a unique project id (warning: cannot be modified afterward)
[6-30 characters]:
() ktlt-fe-angular
Your public directory is the folder (relative to your project directory) tha
t
will contain Hosting assets to be uploaded with firebase deploy. If you
have a build process for your assets, use your build's output directory.

? What do you want to use as your public directory? public
? Configure as a single-page app (rewrite all urls to /index.html)? No
✓  Wrote public/404.html
✓  Wrote public/index.html

i  Writing configuration info to firebase.json...
i  Writing project information to .firebaserc...

✓  Firebase initialization complete!

```

Bước 7: Chạy lệnh "firebase deploy" để deploy lên firebase

```

$ firebase deploy

==> Deploying to 'ktlt-fe-angular'...

i  deploying hosting
i  hosting[ktlt-fe-angular]: beginning deploy...
i  hosting[ktlt-fe-angular]: found 7 files in dist/web
✓  hosting[ktlt-fe-angular]: file upload complete
i  hosting[ktlt-fe-angular]: finalizing version...
✓  hosting[ktlt-fe-angular]: version finalized
i  hosting[ktlt-fe-angular]: releasing new version...
✓  hosting[ktlt-fe-angular]: release complete

✓  Deploy complete!

Project Console: https://console.firebaseio.google.com/project/ktlt-fe-angular/o
verview
Hosting URL: https://ktlt-fe-angular.web.app

```

## 2.3. Giới thiệu về NodeJS



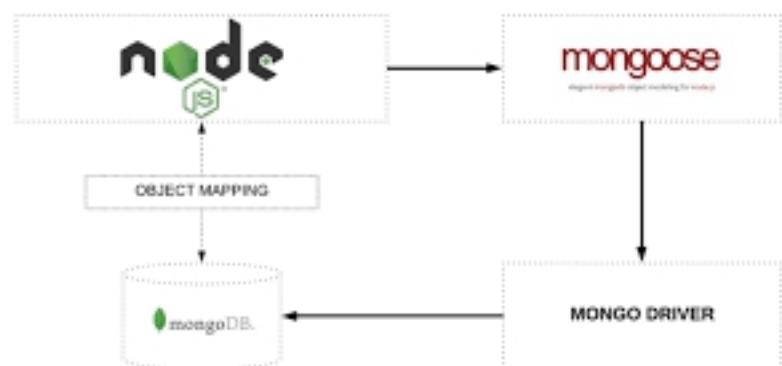
- **Node.js** là một hệ thống phần mềm được thiết kế để viết các ứng dụng internet có khả năng mở rộng, đặc biệt là máy chủ web. Chương trình được viết bằng JavaScript, sử dụng kỹ thuật điều khiển theo sự kiện, nhập/xuất không đồng bộ để tối thiểu hóa chi phí và tối đa hóa khả năng mở rộng. Node.js bao gồm có V8 JavaScript engine của Google, libUV, và vài thư viện khác.
- Node.js được tạo bởi Ryan Dahl từ năm 2009, và phát triển dưới sự bảo trợ của Joyent.
- Node.js là một JavaScript runtime được build dựa trên engine JavaScript V8 của Chrome. Node.js sử dụng kiến trúc hướng sự kiện event-driven, mô hình non-blocking I/O làm cho nó nhẹ và hiệu quả hơn. Hệ thống nén của Node.js, npm, là hệ thống thư viện nguồn mở lớn nhất thế giới.
- Node.js áp dụng cho các sản phẩm có lượng truy cập lớn, cần mở rộng nhanh, cần đổi mới công nghệ, hoặc tạo ra các dự án Startup nhanh nhất có thể.
- Node.js tạo ra được các ứng dụng có tốc độ xử lý nhanh, realtime thời gian thực.
- Phần Core bên dưới của Node.js được viết hầu hết bằng C++ nên cho tốc độ xử lý và hiệu năng khá cao.
- Trong đề tài lần này, nhóm chúng em sử dụng NodeJS và các thư viện liên quan để thiết kế web server cũng như kết nối với các thiết bị để thực hiện các chức năng đã đề ra.

## 2.4. Thư viện dùng để crawl với NodeJS: Cherrio và Axios



Cherrio: là một thư viện hỗ trợ parse DOM giống như Jquery cung cấp nhiều công cụ để việc crawl được hiệu quả  
Axios là một thư viện HTTP Client dựa trên Promise. Cơ bản thì nó cung cấp một API cho việc xử lý XHR (XMLHttpRequests).

## 2.5. Object Data Model: Mongoose



Mongoose là một thư viện mô hình hóa đối tượng (Object Data Model - ODM) cho MongoDB và Node.js. Nó quản lý mối quan hệ giữa dữ liệu, cung cấp sự xác nhận giản đơn và được sử dụng để dịch giữa các đối tượng trong mã và biểu diễn các đối tượng trong MongoDB. Nhóm nhận thấy sử dụng Mongoose rất thuận tiện cho việc tạo model nên đã đưa vào sử dụng.

### 3. Deploy

### 3.1. Backend

Nền tảng sử dụng: Heroku

Link api backend sau khi deploy: <http://dakslt.herokuapp.com/>

Truy cập api: getAll

## Truy cập api getThongKe

## 3.2. Front end

Nền tảng sử dụng: Firebase

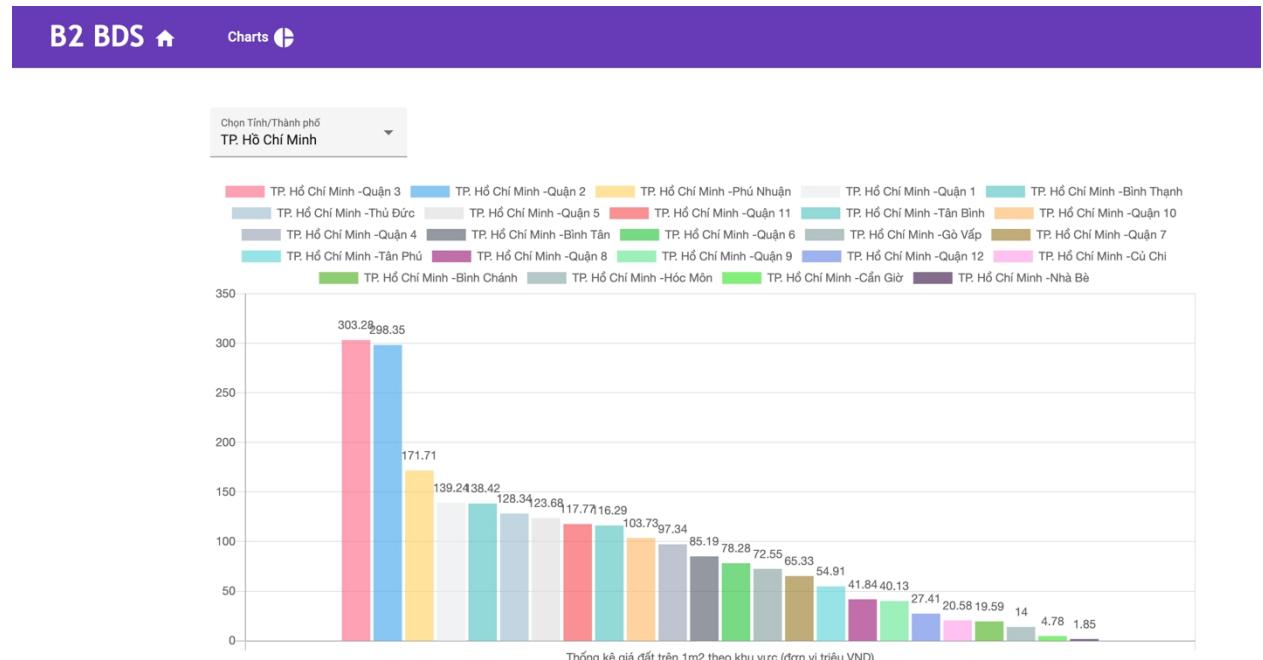
Url sau khi deploy: <https://ktlt-fe-angular.web.app/>

Giao diện trang chủ:

The screenshot shows the homepage of a real estate website. At the top, there is a purple header bar with the logo "B2 BDS" and a "Charts" button. Below the header, there are search filters: "Tim kiếm" (Search), "Tim theo tiêu đề" (Search by title), "Tim theo khu vực" (Search by area), and "Lọc theo diện tích" (Filter by area). There are also fields for "Từ" (From) and "Đến" (To). To the right, there are three property cards:

- Bán Gấp Căn Hộ 3 Phòng Quận 2**  
Vị trí: TP. Hồ Chí Minh - Quận 2  
Diện tích: **86m<sup>2</sup>**  
Giá: ₫3,100,000,000
- Bán Nhà Xuân Thới Thượng, Hóc Môn (Chính Chủ)**  
Vị trí: TP. Hồ Chí Minh - Hóc Môn  
Diện tích: **18m<sup>2</sup>**  
Giá: ₫650,000,000
- Cần Bán Lô Đất Vườn 10.000 Gần Nhánh Rè Cao Tốc Ph...**  
Vị trí: Bình Thuận - Bến Tre  
Diện tích: **11000m<sup>2</sup>**

Giao diện trang thống kê:



## 4. Hướng dẫn sử dụng

B2 BDS ↑ Charts 3

**Tim kiếm** 1

Tìm theo tiêu đề \_\_\_\_\_

Tìm theo khu vực \_\_\_\_\_

**Lọc theo diện tích**

Từ \_\_\_\_\_

Đến \_\_\_\_\_

Items per page: 10 1 – 10 of 9919 2 < >

**Vị trí: TP Hồ Chí Minh -Quận 2**

**Bán Gấp Căn Hộ 3 Phòng**  
**Quận 2**

Diện tích: **86m<sup>2</sup>**  
đ3,100,000,000

**Vị trí: TP Hồ Chí Minh -Hóc Môn**

**Bán Nhà Xuân Thới**  
**Thượng, Hóc Môn (**  
**Chính Chủ )**

Diện tích: **18m<sup>2</sup>**  
đ650,000,000







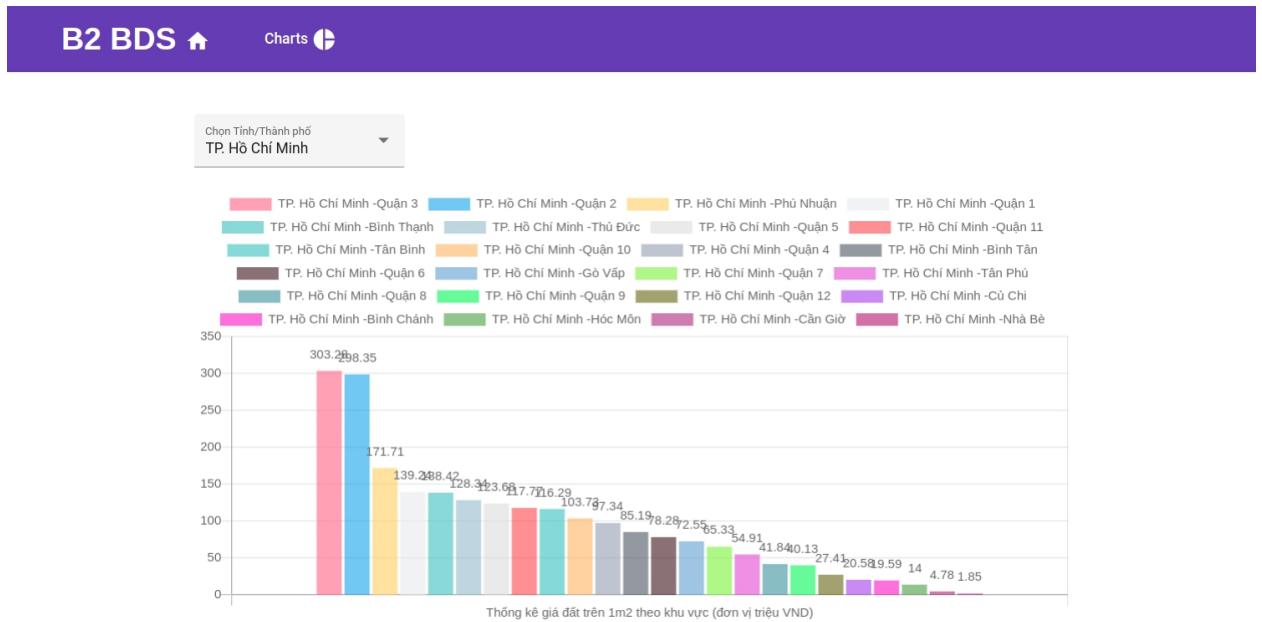


Chú thích:

Số 1: Bộ lọc và tìm kiếm, dùng để tìm kiếm bất động sản theo tiêu đề, khu vực và trong khoảng diện tích nhất định.

Số 2: Paging dùng để xem các trang tiếp theo.

Số 3: Xem biểu đồ.



Ở trang biểu đồ, ta có thể chọn Tỉnh/Thành phố tương ứng để xem các thông tin về bất động sản của các quận/huyện thuộc tỉnh đó.

## 5. Báo cáo hàng tuần

### BÁO CÁO TUẦN 1

Thời gian: 22/11 – 29/11/2020

#### Danh sách nhóm:

- 1920058 Lê Tất Thiện
- 1927038 Trần Phương Tĩnh
- 1920068 Dương Quang Tuấn
- 1927040 Nguyễn Đức Tuấn

#### Các công việc đã thực hiện

- Setup backend, hiện thực crawl dữ liệu bằng selenium và cheerio, làm module xuất dữ liệu ra file csv, crawl 400 record từ batdongsan.com (Thiện), đã push code lên repository DAKTLT-Backend.
- Crawl dữ liệu trên trang homedy.com, viết hàm convert thời gian sử dụng cho các trang không ghi rõ định dạng ngày (Tĩnh), đã push code lên repository DAKTLT-Backend.
- Crawl dữ liệu trên trang batdongsan321.com (Tuấn Dương), đã push code lên repository DAKTLT-Backend.
- Crawl dữ liệu trên trang 123nhadat.vn (Tuấn Nguyễn), đã push code lên repository DAKTLT-Backend.
- Viết report, hướng dẫn sử dụng (Thiện), đã push báo cáo lên folder report/tuan1 trong repository DAKTLT-Backend.

#### Các khó khăn gặp phải

- Ban đầu nhóm sử dụng Selenium để tiến hành crawl dữ liệu, nhưng khi sử dụng vòng lặp với selenium thì có vấn đề với bất đồng bộ, kết quả chạy không được như mong muốn.
- Nhóm chuyển qua sử dụng Cheerio và nhận thấy thời gian crawl tăng đáng kể và khắc phục được vấn đề trên.

#### Dự kiến công việc tuần tới

- Triển khai Frontend để thuận tiện cho việc crawl các dữ liệu cần thiết.
- Sử dụng MongoDB để lưu trữ và GraphQL để thao tác và truy vấn API.
- Hoàn thiện module xuất file csv.

# BÁO CÁO TUẦN 2

Thời gian: 30/11 – 6/12/2020

## Danh sách nhóm:

- 1920058 - Lê Tất Thiện
- 1927038 - Trần Phương Tĩnh
- 1920068 - Dương Quang Tuấn
- 1927040 - Nguyễn Đức Tuấn

## Các công việc đã thực hiện

- Viết prototype function crawlWeb (Thiện), đã update vào file module/tinh.js trong repository DAKLT-Bbackend.
- Hoàn thiện function crawlWeb và chạy thử (Tĩnh), đã update vào file module/tinh.js trong repository DAKLT-Bbackend.
- Tìm hiểu format chuẩn, cách làm sạch data set (Tuấn Nguyễn).
- Tuần 1 đã crawl 10k record và gửi cho Tĩnh (Tuấn Dương), tuần 2 làm sạch và đưa về format chuẩn mong muốn (Tĩnh, Thiện), đã push file 10kRecords.json lên folder samplerecord trong repository DAKLT-Bbackend.
- Tìm hiểu và setup frontend (Tuấn Dương, Tuấn Nguyễn), đã push lên repository DAKLT-Frontend.
- Kết nối backend với mongoDB, insert 10k record vào mongoDB và tạo các model cho web (Thiện), đã push lên repository DAKLT-Bbackend.
- Viết báo cáo tuần 2 (Tĩnh), đã push báo cáo lên folder report/tuan2 trong repository DAKLT-Bbackend.

## Các khó khăn gặp phải

- Web <https://www.muabannhadat.com.vn/> bị sập nên chuyển sang crawl web khác.
- 10k record được crawl về ở tuần 1 chưa đủ thuộc tính, chưa sạch và chưa đúng format chuẩn mong muốn, ở tuần 2 này nhóm đã viết thêm function crawlWeb để crawl trực tiếp từng mẫu tin của trang chính, địa chỉ để crawl lấy từ 10k record cũ. Từ đây crawl được 10k record mới đủ thuộc tính, sạch, và đúng format chuẩn mong muốn, lưu vào file json, sau đó insert vào mongoDB.

## Dự kiến công việc tuần tới

- Sử dụng MongoDB để lưu trữ và GraphQL để thao tác và truy vấn API.
- Backend: Viết API trả về dữ liệu.
- Frontend: Hiển thị dữ liệu.

# BÁO CÁO TUẦN 3

Thời gian: 7/12 – 13/12/2020

## Danh sách nhóm:

- 1920058 - Lê Tất Thiện
- 1927038 - Trần Phương Tĩnh
- 1920068 - Dương Quang Tuấn
- 1927040 - Nguyễn Đức Tuấn

## Các công việc đã thực hiện

- Hiện thực function searchLocation dùng để tìm kiếm theo địa điểm (Thiện), đã push code lên repository DAKTLT-Backend, branch master.
- Hiện thực function searchArea dùng để tìm kiếm diện tích trong đoạn mong muốn (Thiện), đã push code lên repository DAKTLT-Backend, branch master.
- Hiện thực function getall dùng để lấy tất cả tin (Tĩnh), đã push code lên repository DAKTLT-Backend, branch tinh.
- Hiện thực function searchTitle dùng để tìm kiếm theo tựa đề (Tĩnh), đã push code lên repository DAKTLT-Backend, branch tinh.
- Dùng api của Thiện và Tĩnh viết để render data ra Frontend (Tuấn Dương + Tuấn Nguyễn), đã push code lên repository DAKTLT-Frontend.
- Viết báo cáo tuần 3 (Tĩnh), đã push báo cáo lên folder report/tuan3 trong repository DAKTLT-Backend, branch master.

## Các khó khăn gặp phải

- Khó khăn trong việc xử lý hơn 10k dữ liệu một cách hợp lý và trả về data nhanh; khó khăn khi làm phân trang. Đã khắc phục bằng cách lấy hết về rồi đếm phần tử.

## Dự kiến công việc tuần tới

- Backend: Hoàn thiện các api.
- Frontend: Hoàn thiện render data.
- Thông kê dữ liệu.

## BÁO CÁO TUẦN 4

Thời gian: 14/12 – 20/12/2020

## Danh sách nhóm:

- 1920058 - Lê Tất Thiện
  - 1927038 - Trần Phương Tĩnh
  - 1920068 - Dương Quang Tuấn
  - 1927040 - Nguyễn Đức Tuấn

## Các công việc đã thực hiện

- Hiện thực api cleanAddress , fixAddress, findnotadd, batdongsan321, batdongsan (Tỉnh) để clean address, đã push code lên repository DAKTLT-Backend, branch tinh.
  - Hiện thực api fixpricebds, thongketheoquan, fixthongke (Thiện) để clean price và thông kê theo quận, đã push code lên repository DAKTLT-Backend, branch master.
  - Render thông kê ra Frontend (Tuấn Dương), đã push code lên repository DAKTLT-Frontend.
  - Viết báo cáo tuần 4 (Tỉnh), đã push báo cáo lên folder report/tuan4 trong repository DAKTLT-Backend, branch master.

## Các khó khăn gặp phải



## Dự kiến công việc tuần tới

- Hoàn thiện cũng như sửa các lỗi nhỏ.
  - Bổ sung tính năng cần thiết.

# BÁO CÁO TUẦN 5

Thời gian: 21/12 – 27/12/2020

## Danh sách nhóm:

- 1920058 - Lê Tất Thiện
- 1927038 - Trần Phương Tĩnh
- 1920068 - Dương Quang Tuấn
- 1927040 - Nguyễn Đức Tuấn

## Các công việc đã thực hiện

- Gộp api get all + tìm theo title, theo location, theo area vào làm một (Thiện, Tĩnh), đã push code lên repository DAKTLT-Backend, branch master.
- Fix paging (Tuấn Dương), đã push code lên repository DAKTLT-Backend, branch master.
- Display products (Tuấn Dương), đã push code lên repository DAKTLT-Frontend.
- Add search to side bar, Done filter (Tuấn Nguyễn), đã push code lên repository DAKTLT-Frontend.
- Viết báo cáo tuần 5 (Tĩnh), đã push báo cáo lên folder report/tuan5 trong repository DAKTLT-Backend, branch master.

## Các khó khăn gặp phải

- Các khó khăn đã được giải quyết.

## Dự kiến công việc tuần tới

- Đã hoàn thành project.

## 6. Lịch sử commit code trên github

<https://github.com/users/thien-lebk/projects/1>

### 6.1. Backend

<https://github.com/thien-lebk/DAKLT-Backend>

commit 47a19ca554539a0c3ad8600cd3d58eb9e89ef3cb (HEAD -> master, origin/master)

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 27 22:17:50 2020 +0700

add cors

commit bab0e2a3d3bca6fcf59edf3c98e79470e8e5c7e2

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 27 22:12:37 2020 +0700

disable x powered by

commit 3a6f430329714fc810bab0c9d667cc12c17de545

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 27 13:26:50 2020 +0700

add report tuan5.pdf

commit 56c6d50b74dcb4aa07b8cf4c268888bae76fe76

Merge: 8e03f23 90a3c3f

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 26 22:39:17 2020 +0700

Merge branch 'master' of  
<https://github.com/thien-lebk/DAKLT-Backend>

commit 8e03f2345bdea6c9c42f0405311fe01396409874

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 26 22:38:52 2020 +0700

add get thong ke  
commit 90a3c3f7b9760354085f2e8f4834f1499d179537  
Author: Tuan Duong <quangtuan9237@gmail.com>  
Date: Sat Dec 26 03:22:41 2020 +0700

fix paging

commit c851217e62c07fd7e8596e826594446eb85d2d81  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Fri Dec 25 20:45:10 2020 +0700

.

commit 263cedbe9c2981689efcbce8153c7a97a829870d  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Fri Dec 25 20:40:03 2020 +0700

update get all

commit 1ee6b78ff275e8a70cbeefbb97cf41e6fc1a1583  
Merge: 9017ab5 e475563  
Author: tinhptran <74109334+tinhptran@users.noreply.github.com>  
Date: Sun Dec 20 22:10:46 2020 +0700

Merge pull request #2 from thien-lebk/tinh

Tinh

commit 9017ab51057a699f91c4ba8985af98606e6a7135  
Author: tinhptran <74109334+tinhptran@users.noreply.github.com>  
Date: Sun Dec 20 15:28:06 2020 +0700

add report tuan4.pdf

commit e47556333b20888bf8e0e3c37e8eec38a65d5b (origin/tinh, tinh)  
Merge: 848d1c6 9faba01  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Sun Dec 20 15:07:19 2020 +0700

merge cfonlic

commit 848d1c6ed6216b9f07ecbfa3a95dfef8259ea31b  
Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 20 14:34:01 2020 +0700

add thong ke

commit 9faba0129af173e7f6fc86e3403e38ddc2379352

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 20 14:27:56 2020 +0700

clean data

commit 537176f7110a34c7d1edc5f1f8516dbb97ee8800

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 20 14:24:55 2020 +0700

clean data

commit 3e4877b003bec31f110b4b671f97bfd727bc808f

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 13 23:41:24 2020 +0700

add report tuan3.pdf

commit adeb4bb383cdf69615d220d8f455faa3ed26b1ab

Merge: eebcf42 9b0262c

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 13 22:43:33 2020 +0700

Merge branch 'tinh'

commit 9b0262c0e12b82176be347742b3cf4804a591bef

Merge: 0192509 630fded

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 13 22:43:16 2020 +0700

Merge branch 'tinh' of  
<https://github.com/thien-lebk/DAKLT-Backend> into tinh

commit eebcf4277537417a9940695eadb8f3a230483fea

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 13 22:32:03 2020 +0700

.

commit b8ab67c776e8a2f6f9e3f52e8f0870728b3dca36

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 13 22:29:56 2020 +0700

commit ef22505fbc9de8a3a49dc65ddf181b17a9c4d052

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 13 22:29:03 2020 +0700

add api serach area + location

commit 630fded1d6a460167c595c3f86e8fae658b87da7

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 13 22:27:09 2020 +0700

add api getall and searchtitle

commit 1c05670bd75cb099e9bdd550e686f4582bdf2e9d

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 13 22:24:24 2020 +0700

add api getall and searchtitle

commit 632ccbdc2e75d96a55976ffcf1166dfad578154f

Merge: 30f9020 fcflaa0

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 6 23:52:46 2020 +0700

Merge branch 'master' of  
<https://github.com/thien-lebk/DAKLT-Backend>

commit 30f9020791110f41341509ecb14effa14fb42e73

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 6 23:52:09 2020 +0700

commit fcflaa0212b2a0864064828cffb774022dae7375

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Dec 6 23:36:27 2020 +0700

Delete test.doc

commit b524015250efe33f8c38f6eefee161f8c4500d79  
Author: tinhptran <74109334+tinhptran@users.noreply.github.com>  
Date: Sun Dec 6 23:34:30 2020 +0700

add report tuan2.pdf

commit 7d12b05db52d93ab7257397b60fff1d286bb4da9  
Author: tinhptran <74109334+tinhptran@users.noreply.github.com>  
Date: Sun Dec 6 23:34:02 2020 +0700

Delete tuan2.pdf

commit 2d6f21daa39bd563bf641f1896971bc9b51d494b  
Author: tinhptran <74109334+tinhptran@users.noreply.github.com>  
Date: Sun Dec 6 23:29:55 2020 +0700

add 10kRecords.json

commit dff47cbdda43a28cb3f785054e2304be7030ed3b  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Sun Dec 6 23:12:25 2020 +0700

.

commit 98afb5c677e893a6e18416cfacaed0a88d84cea9  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Sun Dec 6 18:50:07 2020 +0700

.

commit 4ec1df0c21324c4d23b02c1826c7447c6a1eec12  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Sun Dec 6 18:47:38 2020 +0700

setup db

commit 1056014a209acca967c5fa67e6c917b3e4b7ce67  
Author: thien-lebk <bhnhock@gmail.com>  
Date: Sun Dec 6 01:34:12 2020 +0700

.

commit 52244348aa28675bff6855596d976b6de0fd467a

Merge: ad38c4b 75b0879

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 6 01:32:57 2020 +0700

.

commit ad38c4bfb847f1b131ce8ef63de870e9aa3164db

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Dec 6 01:32:03 2020 +0700

add db

commit 75b0879019dd17c85b3826435c0628530385e726

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sat Dec 5 23:33:50 2020 +0700

update function crawlWeb

commit 5880c4d3e444416fb78446893bd88988ab67e594

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sat Dec 5 23:21:47 2020 +0700

update function crawlWeb

commit 98044a0009b18aa357e36759f5f83e1d104bc41d

Merge: 0192509 b5b741c

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 22:39:26 2020 +0700

Merge branch 'master' of  
<https://github.com/thien-lebk/DAKLT-Backend>

commit 01925098c940b71b754ed76f30cf2fd948463c9b

Merge: 2b439fe 80e8662

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 22:36:43 2020 +0700

Merge branch 'master' into tinh

commit 2b439fe322716e7ba13c8a92f2b81f581306dbef

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 22:31:13 2020 +0700

add function cua tinh~

commit a753bf22f9784795a4ce6ca94987e621d78fe80c

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 22:18:58 2020 +0700

commit 92cc8cf068a6af172a36e16768126ac4534ad650

Merge: b284bce 71e7b5b

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 22:17:31 2020 +0700

commit 71e7b5bec66dfc991515e891253c842df68e6752

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sat Dec 5 22:16:08 2020 +0700

update function crawlWeb

commit b284bcef01fb068657397ffcaf3d88b05c1fc0a0

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 22:15:56 2020 +0700

commit 806773a48bde19fd9f175e5eefc939ff35ca543e

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sat Dec 5 21:21:27 2020 +0700

update function crawlWeb

commit a13d5e7b7ecd0014679b9485c5aaa1da6b854176

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 20:18:34 2020 +0700

add function crawlWeb

commit 3d8d5f93b18a7add87521c30303f8cfaa2020678

Merge: a25ff28 7d0643e

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 20:16:05 2020 +0700

Merge branch 'tinh' of  
<https://github.com/thien-lebk/DAKTLT-Backend> into tinh

commit 80e86627945f41afd34307087b8e5f2ecd453f43

Author: thien-lebk <bhnhock@gmail.com>

Date: Sat Dec 5 20:02:57 2020 +0700

>

commit b5b741cd32c0bfa7f7f1d70efddedfd98f213c8b

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Nov 29 23:50:05 2020 +0700

add sample data

commit 8719148dca4c1835421cc394d8b45cc5ec24e29f

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Nov 29 22:54:49 2020 +0700

add report

commit a25ff28075915aea33cbec975fe25e05c7132b56

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Nov 29 22:43:37 2020 +0700

.

commit 7d0643e61c4e987dc1ebdbbed93181168cbc0875

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Nov 29 01:01:23 2020 +0700

add api homedy.com

commit f0f3c1d47fa1aa56fb56c3098e74dcb383e57a86

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Nov 29 01:00:18 2020 +0700

add api homedy.com

commit 0da24e0983668254f1ae69a57ca84745cf1ca0de

Author: tinhptran <74109334+tinhptran@users.noreply.github.com>

Date: Sun Nov 29 00:50:45 2020 +0700

add api homedy.com

commit 5c69b90008a7b3852c719a9a103f713e1c0165ef

Author: thien-lebk <bhnhock@gmail.com>

Date: Mon Nov 23 01:07:39 2020 +0700

add promise

commit c726cb7f5ee40022033cc1eba285b1f558d87c61

Author: thien-lebk <bhnhock@gmail.com>

Date: Mon Nov 23 01:01:13 2020 +0700

add module

commit 86280f6df112bea777346ca348c4ed9c7e3a01a2

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Nov 22 19:55:06 2020 +0700

edit muabandat

commit 8129d5419292d7ff74ea2b5b0929248c2f585677

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Nov 22 19:49:21 2020 +0700

add api batdongsan.com

commit 3d1ceeb275ba11f990cfcc22f66c17fcfd79b3195

Author: thien-lebk <bhnhock@gmail.com>

Date: Sun Nov 22 15:17:08 2020 +0700

add cheerio

commit c29d548b5a4eb5721d1f52c5b5e01257388e0c6c

Author: thien-lebk <bhnhock@gmail.com>

Date: Fri Nov 20 21:08:25 2020 +0700

1st commit

## 6.2. Frontend

<https://github.com/thien-lebk/DAKLT-Frontend>

-o	Commits on Dec 27, 2020
	<b>add loading for charts and no display chart if no selected city</b> NguyenDucTuan92N1 committed 2 minutes ago
	<b>add firebase hosting configs</b> NguyenDucTuan92N1 committed 14 minutes ago
	<b>fix bug paging</b> quangtuan9237 committed 6 hours ago
	<b>add chart to web</b> quangtuan9237 committed 6 hours ago
-o	Commits on Dec 26, 2020
	<b>merge new code</b> quangtuan9237 committed 23 hours ago
	<b>done filter</b> NguyenDucTuan92N1 committed 2 days ago
-o	Commits on Dec 24, 2020
	<b>add search to side bar</b> NguyenDucTuan92N1 committed 4 days ago
	<b>display products</b> quangtuan9237 committed 4 days ago
-o	Commits on Dec 6, 2020
	<b>add web front end</b> quangtuan9237 committed 21 days ago
	<b>move crawler to a folder</b> quangtuan9237 committed 21 days ago
-o	Commits on Nov 28, 2020
	<b>Add files via upload</b> NguyenDucTuan92N1 committed 29 days ago
-o	Commits on Nov 27, 2020
	<b>first commit</b> NguyenDucTuan92N1 committed on Nov 27
	<b>first commit</b> NguyenDucTuan92N1 committed on Nov 27

## **7. Tài liệu tham khảo**

- Slide bài giảng Kỹ thuật lập trình trên BKel.
- <https://towardsdatascience.com/how-to-build-a-simple-web-crawler-6082fc82470>
- <https://scrapy.org/>
- <https://viblo.asia/p/lan-dau-tien-crawl-du-lieu-cua-toi-nhu-the-nao-RnB5p78JlPG>
- [https://www.youtube.com/watch?v=hkF\\_oIm3lU4&t=2197s](https://www.youtube.com/watch?v=hkF_oIm3lU4&t=2197s)