**Tài liệu hướng dẫn sử dụng**

**module crawl dữ liệu chứng khoán 24h**

|  |  |  |
| --- | --- | --- |
| lịch sử sửa đổi | |  |
| thời gian | mô tả | người cập nhật |
| 2017/09/13 | chỉnh sửa định dạng, thêm mục lục, cập nhật phần cấu hình scrapy | thanhdt |
| 2017/09/12 | cập nhật phần cấu hình socket, cập nhật hyperlink | thanhdt |
| 2017/09/11 | cập nhật phần cách thực thi & cơ sở dữ liệu – mối tương quan với command, dữ liệu hỗ trợ API, cấu hình – cấu hình socket. Sửa format | thanhdt |
| 2017/09/08 | tạo tài liệu, cập nhật phần mô tả, các thành phần và luồng xử lý, cấu trúc thư mục | thanhdt |

# Mô tả

Tài liệu này mô tả luồng xử lý của module crawl dữ liệu cho ứng dụng chứng khoán 24h, bao gồm các cài đặt, cấu hình và deploy server. Thông tin có thể được thay đổi, cập nhật theo thời gian, tùy theo giai đoạn phát triển.

# Nội dung

Đầu mục nội dung

[Mô tả 1](#_Toc493066981)

[Nội dung 1](#_Toc493066982)

[Các thành phần và luồng xử lý 2](#_Toc493066983)

[Bảng giá chứng khoán hàng ngày 2](#_Toc493066984)

[Hồ sơ doanh nghiệp 3](#_Toc493066985)

[Tin tức tổng hợp 4](#_Toc493066986)

[Cấu trúc thư mục 7](#_Toc493066987)

[Cách thực thi 9](#_Toc493066988)

[Thực thi thủ công 9](#_Toc493066989)

[Daemon 11](#_Toc493066990)

[Mô tả cơ sở dữ liệu 11](#_Toc493066991)

[Mối liên hệ giữa command & cơ sở dữ liệu 11](#_Toc493066992)

[Thông tin cơ sở dữ liệu 14](#_Toc493066993)

[Dữ liệu hỗ trợ API 14](#_Toc493066994)

[Cấu hình 14](#_Toc493066995)

[Cấu hình socket 15](#_Toc493066996)

[Cấu hình Scrapy 18](#_Toc493066997)

[Log 19](#_Toc493066998)

[Triển khai trên server 19](#_Toc493066999)

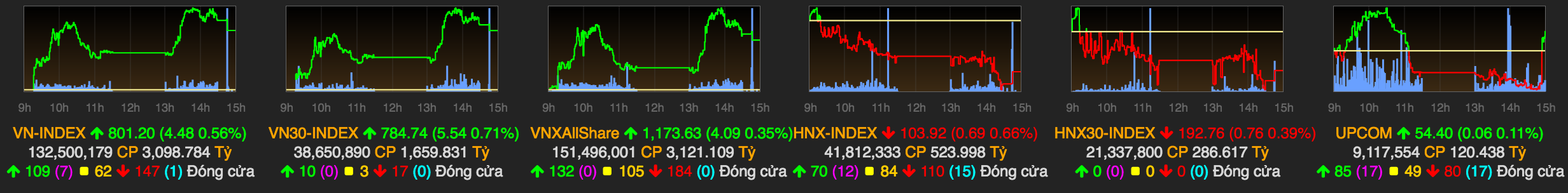
# 

# Các thành phần và luồng xử lý

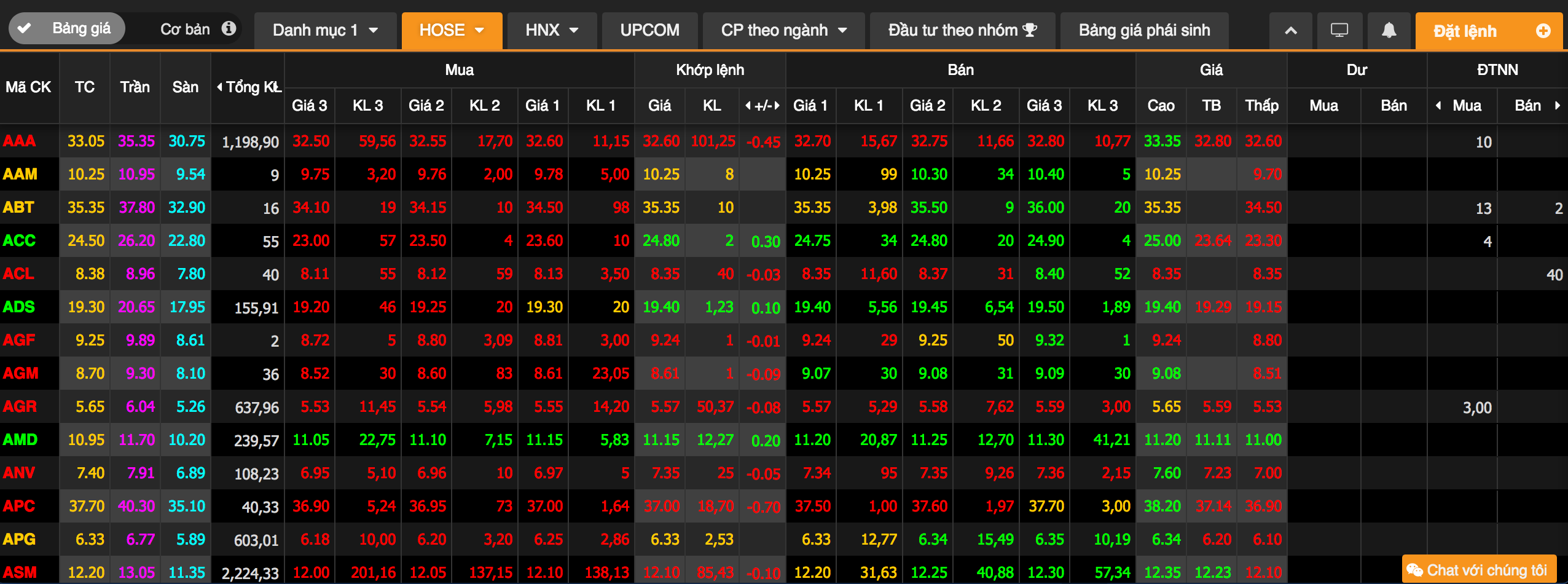
Dữ liệu được crawl bao gồm:

## Bảng giá chứng khoán hàng ngày

Thông tin được lấy tại <https://banggia.vndirect.com.vn>. Trong đó các thông tin được crawl là thông số dữ liệu các sàn HOSE, HNX, HNX-30, UPCOM, VN-30 & chỉ số thị trường tương ứng với: VN-INDEX, VN30-INDEX, VNXAllShare, HNX-INDEX, HNX30-INDEX, UPCOM



*chỉ số thị trường*



*giá mã chứng khoán theo sàn*

Thông tin các chỉ số này sẽ được cập nhật thời gian thực vào cơ sở dữ liệu. Các thông tin này sẽ được giữ lại trong cơ sở dữ liệu trong 1 khoảng thời gian nhằm phục vụ cho việc thống kê vẽ biểu đồ số liệu theo ngày/tháng/năm hoặc là API, sau đó sẽ bị xóa đi để tránh đầy dữ liệu (do dữ liệu cập nhật thời gian thực nên khối lượng rất lớn). Khoảng thời gian này có thể cấu hình được trong script. Mặc định đang để 1 ngày. Script sẽ chạy và thực hiện xóa dữ liệu cũ tại thời điểm bắt đầu (trước khi crawl) nếu thời gian xóa trùng theo cấu hình (xem thêm mục [Cấu hình](#_Cấu_hình_1)). Về thông tin cơ sở dữ liệu, xin xem thêm mục [Mô tả cơ sở dữ liệu](#_Mô_tả_cơ_1).

Ngoài ra hệ thống có sử dụng memcached để lưu thông tin hỗ trợ API, phần này xin xem thêm mục [Dữ liệu hỗ trợ API](#_Dữ_liệu_hỗ_1).

Script này cần được đặt lịch chạy 1 ngày một lần trong khoảng thời gian thị trường giao dịch (từ 9:00 đến 15:00 hàng ngày).

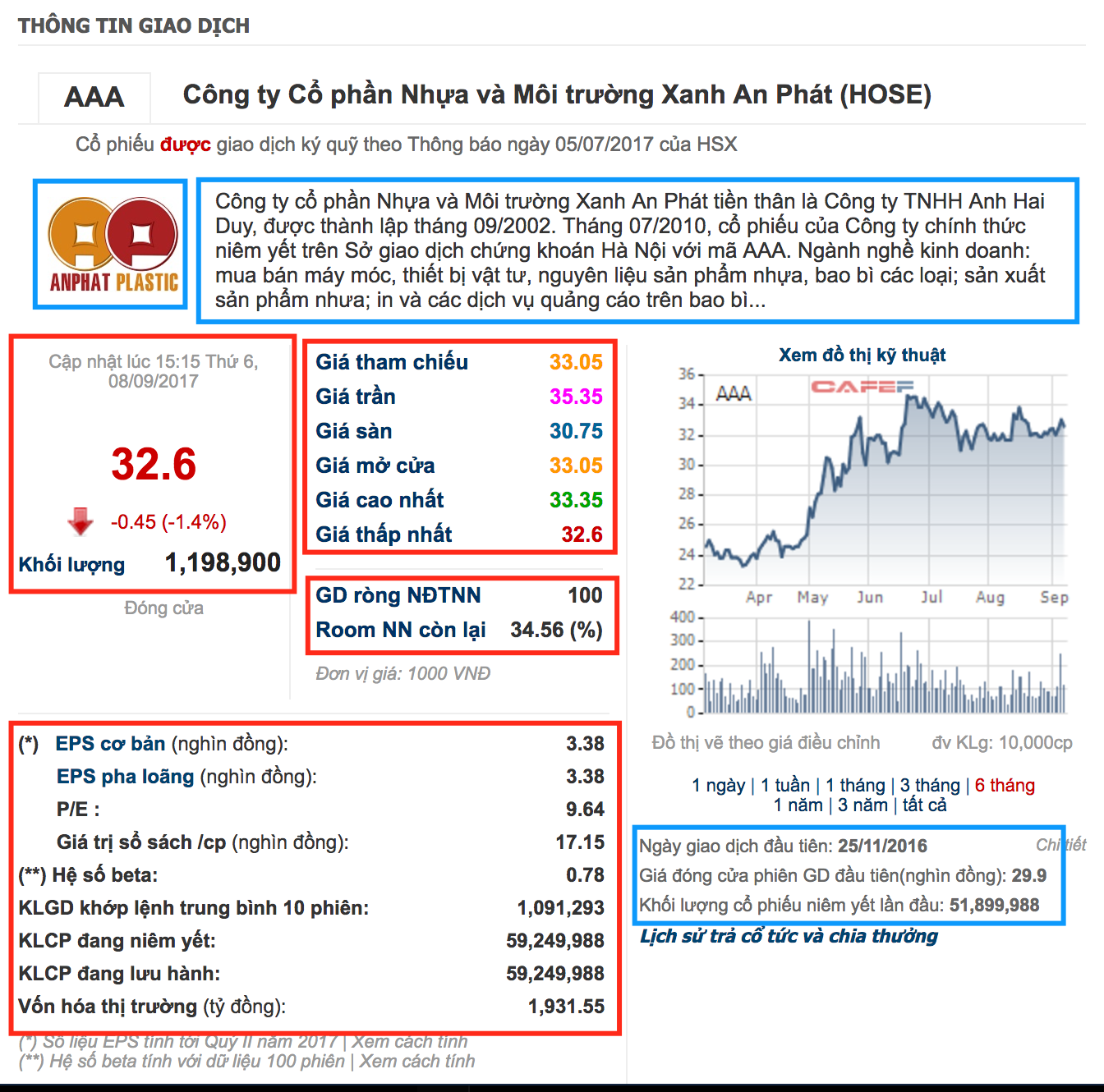
**Đặc tính của script**

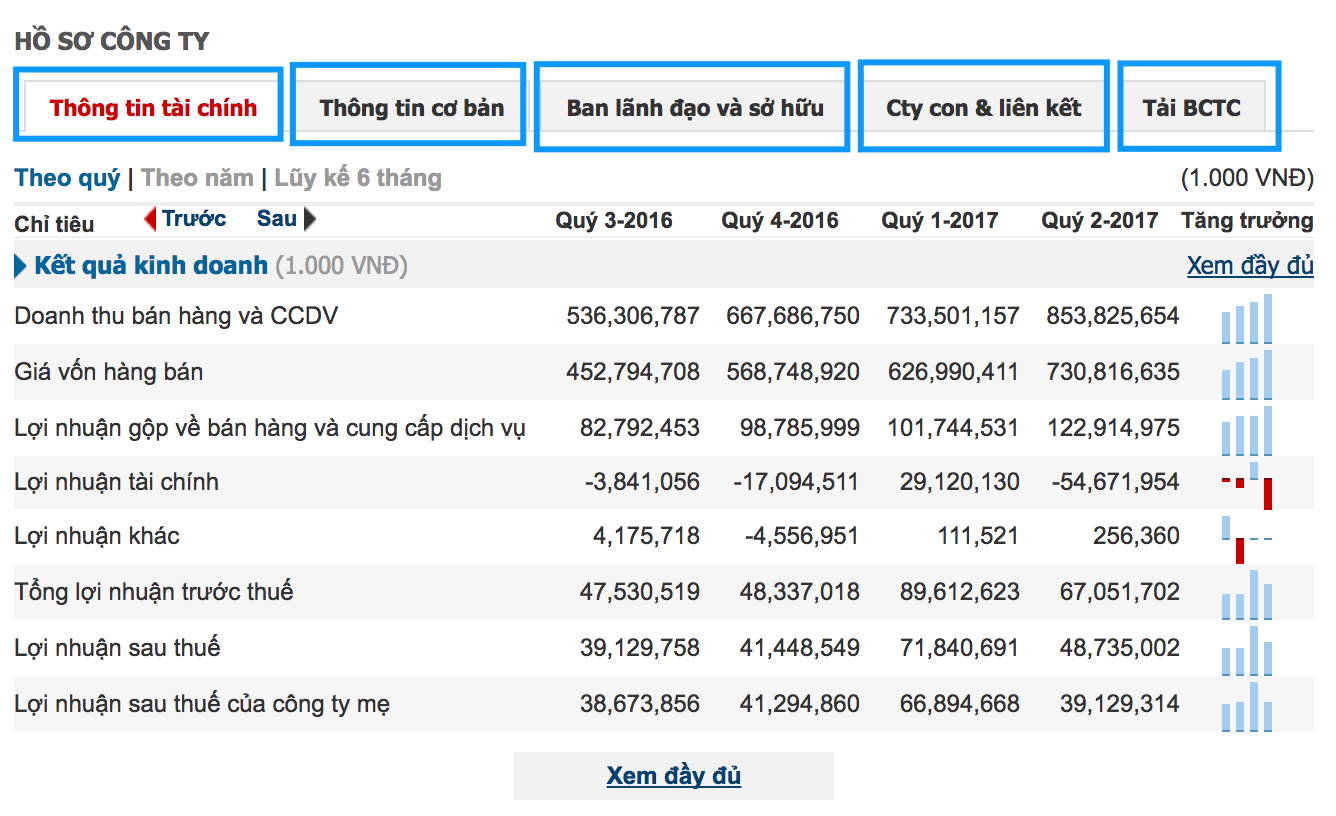
* + Tại thời điểm bắt đầu, kiểm tra xóa dữ liệu cũ nếu trùng ngày trong cấu hình (xem thêm mục “Cấu hình”)
  + Khi được thực thi sẽ kiểm tra thời gian chạy nằm trong khoảng thời gian thị trường mở cửa (có cấu hình khoảng thời gian này, xem thêm mục “Cấu hình”).
  + Nếu trước thời gian dịch, script sẽ đợi cho đến thời gian được cấu hình và mở socket để lấy dữ liệu cho đến khi hết thời gian giao dịch sẽ tự động dừng script
  + Nếu trong thời gian giao dịch, tiến hành ở socket lấy dữ liệu cho đến khi hết thời gian giao dịch sẽ tự động dừng script
  + Nếu quá thời gian giao dịch, sẽ tự động dừng script

## Hồ sơ doanh nghiệp

Danh sách các doanh nghiệp có trên sàn chứng khoán, dữ liệu được lấy từ <http://cafef.vn>. Danh sách các công ty có hồ sơ nằm tại: <http://s.cafef.vn/du-lieu-doanh-nghiep.chn>. Link ví dụ về một hồ sơ: <http://s.cafef.vn/hose/AAA-cong-ty-co-phan-nhua-va-moi-truong-xanh-an-phat.chn>.

Dữ liệu được crawl được phân chia theo 2 phần: dữ liệu được cập nhật thường xuyên và dữ liệu sẽ ít thay đổi. Thông tin cũng được lưu trữ ở 2 nơi khác nhau. Dữ liệu được mô tả bằng hình dưới đây:





Trong đó các dữ liệu được đánh dấu khung đỏ nằm trong nhóm dữ liệu thường xuyên thay đổi, nhóm khung xanh là dữ liệu ít thay đổi. Mục dữ liệu ít thay đổi ở thời điểm hiện tại chỉ crawl dữ liệu dạng html, chưa tách xuất theo từng mục cụ thể. Các dữ liệu thay đổi thường xuyên được thể hiện trên từng trường trong cơ sở dữ liệu.

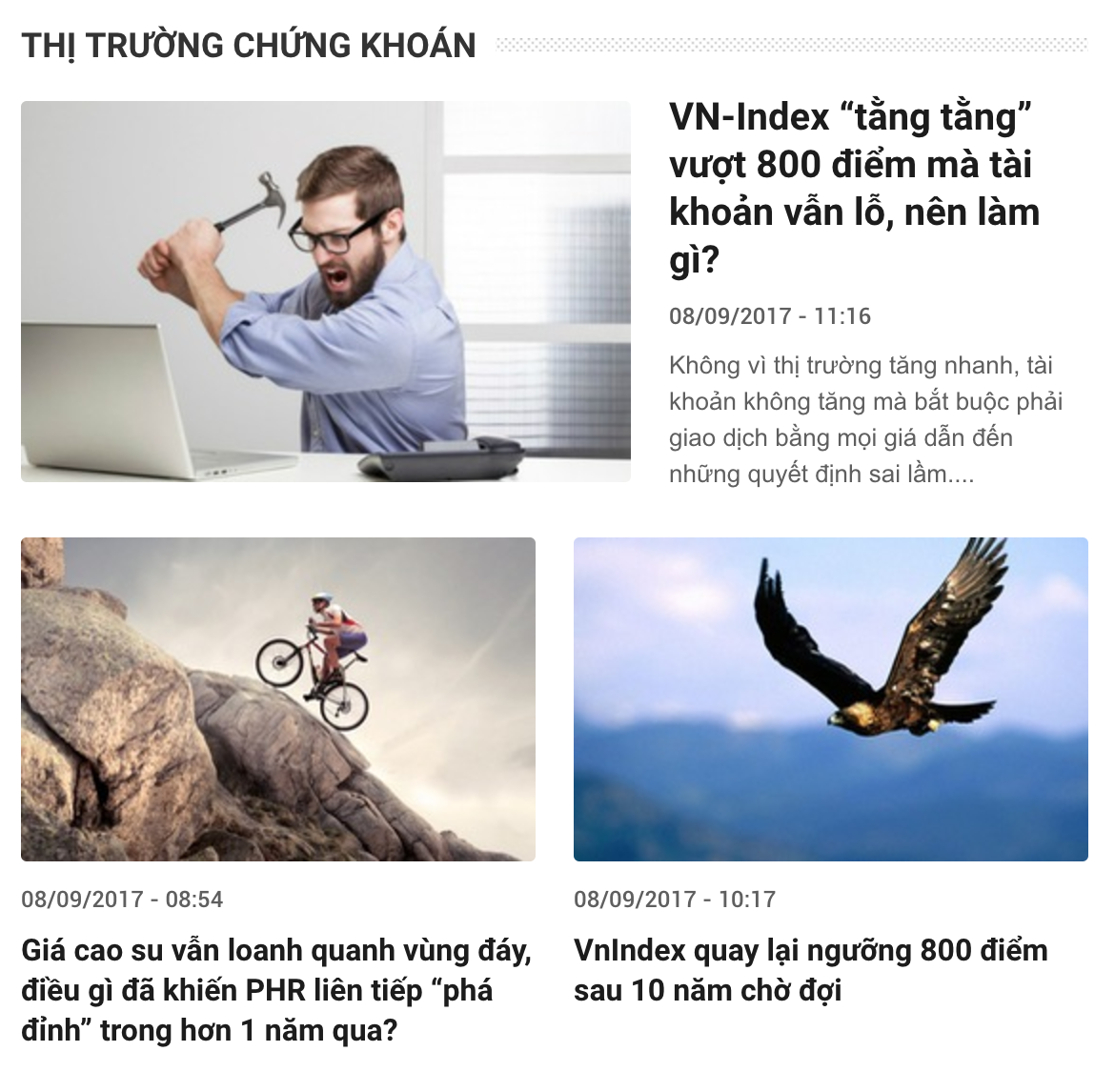
Có 2 script để thực hiện việc crawl dữ liệu hồ sơ: script lấy danh sách link hồ sơ của các công ty, lưu hoặc cập nhật dữ liệu trước đó vào cơ sở dữ liệu (chi tiết xem thêm mục [Mô tả cơ sở dữ liệu](#_Mô_tả_cơ_2)). Script này nên đặt lịch khoảng 10/15 ngày chạy lại một lần vì dữ liệu này sẽ ít thay đổi hoặc có thay đổi cũng không liên tục trong khoảng thời gian ngắn. Script thứ 2 sẽ dựa vào danh sách hồ sơ này tiến hành crawl chi tiết các dữ liệu đỏ/xanh và đổ vào 2 nơi tách biệt trong cơ sở dữ liệu. Script thứ 2 khi chạy sẽ tiến hành luôn việc crawl dữ liệu đỏ, dữ liệu xanh có được crawl hay không phụ thuộc vào cấu hình thời gian bắt đầu và chu kỳ thời gian crawl (chi tiết xem thêm phần [Cấu hình](#_Cấu_hình_2)). Script thứ 2 nên đặt chạy cuối giờ giao dịch hàng ngày (sau 15:00) để cập nhật các dữ liệu đỏ, còn dữ liệu xanh nên cấu hình 5/10 ngày cập nhật một lần (xem thêm phần [Cấu hình](#_Cấu_hình_3))

## Tin tức tổng hợp

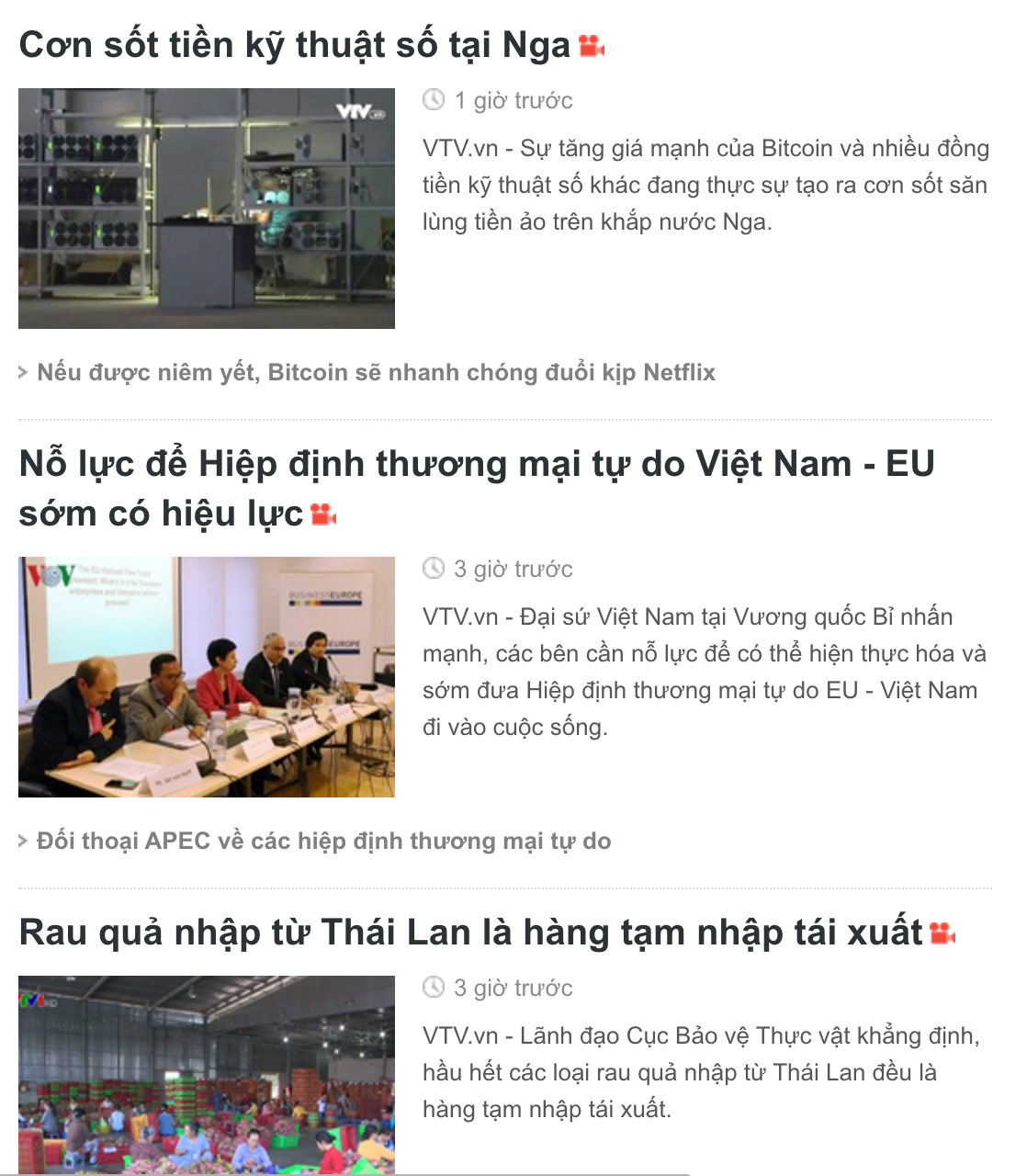
Bao gồm tin ở các link: <http://cafef.vn/thi-truong-chung-khoan.chn> (chứng khoán), <http://nhipcaudautu.vn/doanh-nghiep/tin-tuc/> (doanh nghiệp), <http://vtv.vn/kinh-te.htm> (kinh tế). Đây là 3 chuyên mục tin ứng với các trang cafef.vn, nhipcaudautu.vn & vtv.vn. Tương ứng với mỗi trang, có một script làm nhiệm vụ crawl danh sách các tin mới và cập nhật vào cơ sở dữ liệu. Tin sẽ không được lưu lại vào database cho lần kế tiếp nếu đã được crawl trước đó. Vì cafe.vn & vtv hỗ trợ kênh RSS nên script lấy dữ liệu từ 2 trang này sẽ lấy thông tin tại đây, nhipcaudautu.vn không hỗ trợ RSS nên lấy từ trang danh sách tin của họ.

Link RSS được sử dụng:

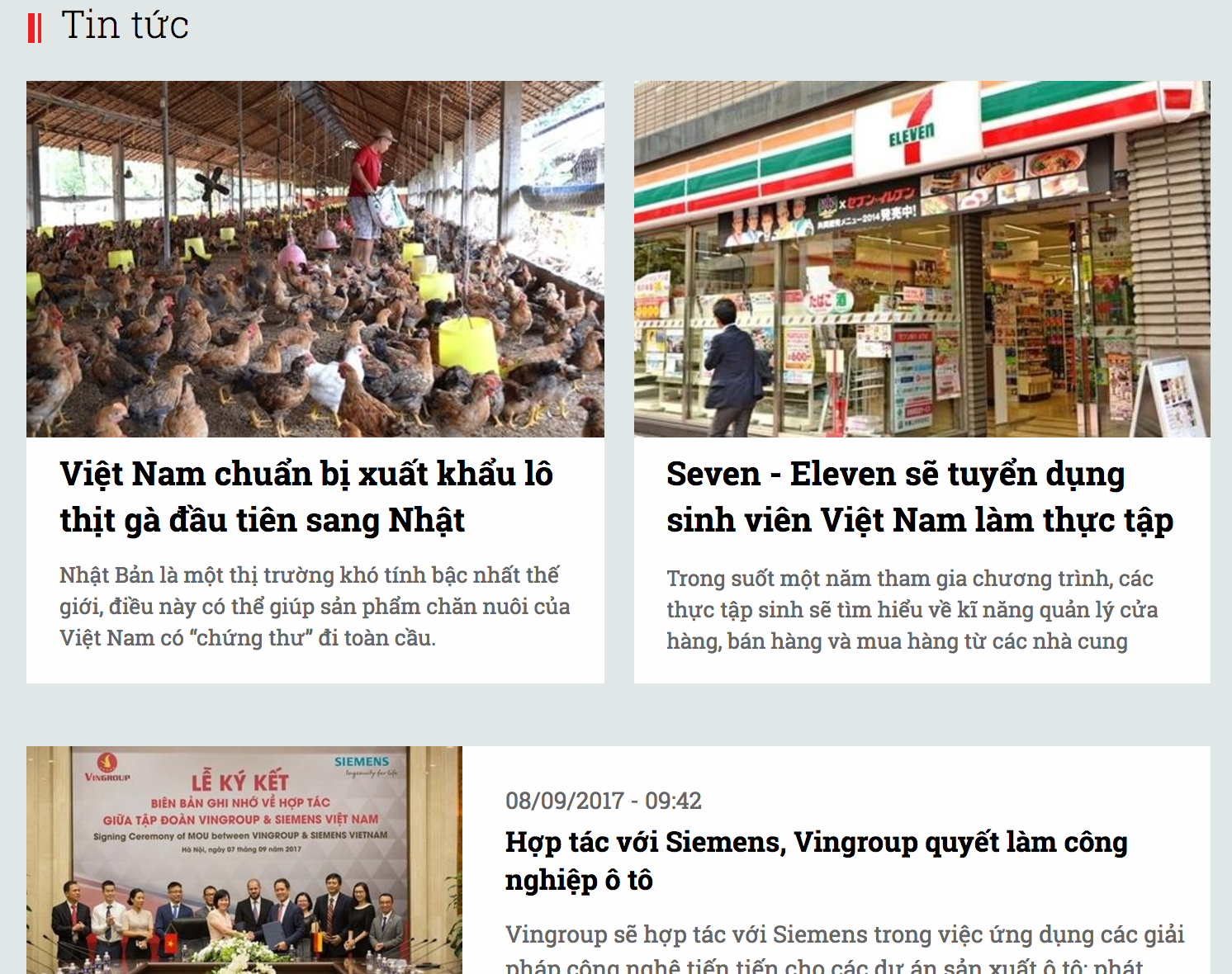
* + <http://cafef.vn/thi-truong-chung-khoan.rss>
  + <http://vtv.vn/kinh-te.rss>



*Danh mục tin trên cafef.vn*



*Danh mục tin trên vtv.vn*

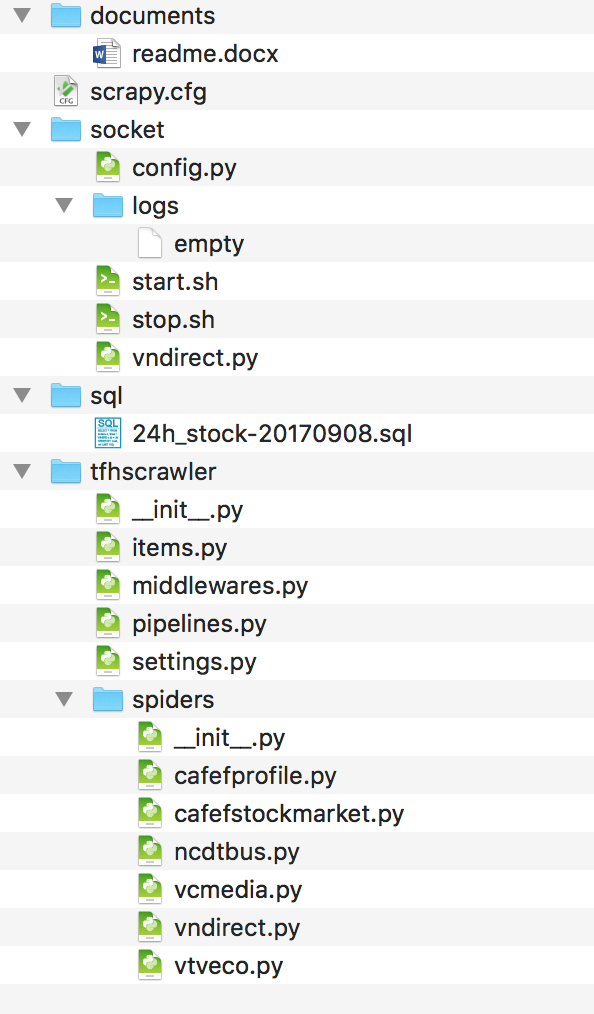


*Danh mục tin trên nhipcaudautu.vn*

3 Script cần được cấu hình crawl thường xuyên để cập nhật các bài viết mới, nên đặt khoảng 15-20 phút crawl 1 lần.

# Cấu trúc thư mục

Cấu trúc thư mục & file được thể hiện như hình dưới



|  |  |
| --- | --- |
| file/folder | mô tả |
| documents | thư mục chứa tài liệu hướng dẫn/cài đặt |
| readme.docx | file hướng dẫn này |
| scrapy.cfg | file cấu hình project của scrapy |
| socket | thư mục chứa script crawl dữ liệu chứng khoán từ vndirect (socket) |
| config.py | file cấu hình cho script crawl dữ liệu chứng khoán |
| logs | thư mục lưu log các thông tin khi crawl, mỗi sàn sẽ lưu một file log, thông tin thị trường cũng được log riêng. Cấu hình tên log được setting trong config.py. Log mặc định được cấu hình rotate theo ngày & lưu tối đa 30 ngày. |
| start.sh | sh script thực thi toàn bộ việc lấy dữ liệu chứng khoán (chỉ số các sàn + thị trường) |
| stop.sh | sh script thực thi việc dừng toàn bộ các script đang crawl dữ liệu chứng khoán nếu đang chạy |
| vndirect.py | script thực thi việc crawl dữ liệu chứng khoán |
| sql | thư mục chứa thông tin sql database |
| 24h\_stock-20170908.sql | file sql khởi tạo ban đầu (cấu trúc, không có dữ liệu) |
| tfhscrawler | thư mục project của scrapy, chứa các spider crawl dữ liệu hồ sơ công ty & tin tức tổng hợp |
| \_\_init\_\_.py | file init của python |
| items.py | file định nghĩa items của scrapy |
| middwares.py | file định nghĩa middware của scrapy |
| piplines.py | file định nghĩa pipline của scrapy |
| settings.py | file cấu hình project cho scrapy |
| spiders | thư mục chứa danh sách các spider thực hiện crwal dữ liệu hồ sơ công ty & tin tức tổng hợp |
| \_\_init\_\_.py | file init của python |
| cafefprofile.py | spider crawl dữ liệu hồ sơ công ty từ cafef.vn |
| cafefstockmarket.py | spider crawl tin tức chứng khoán từ cafef.vn |
| ncdtbus.py | spider crawl tin tức doanh nghiệp từ nhipcaudautu.vn |
| vcmedia.py | spider crawl danh sách công ty chứng khoán từ vcmedia (được cafef sử dụng). Danh sách này hỗ trợ việc crawl dữ liệu hồ sơ |
| vndirect.py | spider crawl danh sách công ty chứng khoán trên sàn được cung cấp bởi vndirect. Dữ liệu crawl được sẽ phục vụ cho API sau này |
| vtveco.py | spider crawl tin tức vtv kinh tế |

# Cách thực thi

Việc crawl dữ liệu được thực hiện bằng cách thực thi các script python, và đặt khoảng thời gian thích hợp với mục đích của từng script để chạy theo chu kỳ. Có 2 cách để thực thi: thủ công hoặc sử dụng daemon.

## Thực thi thủ công

Thực thi thủ công cần sử dụng một trình quản lý lịch như cron, jobber, anacron… đặt lịch cho từng script chạy theo chu kì thời gian.

Dưới đây là danh sách các script và command thực thi:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| stt | tên script | tham số | mô tả | đường dẫn | command | chức năng | đặt lịch |
| 1 | vndirect.py | arg[0] = market | lấy các chỉ số chung từ thị trường chứng khoán (realtime) (vn-index, vn30-index, vnxallshare, hnx-index, hnx30-index, upcom) từ vndirect | ./socket/vndirect.py | python vn-direct.py market | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 2 | vndirect.py | arg[0] = hose | lấy thông tin giá sàn hose (realtime) từ vndirect | ./socket/vndirect.py | python vn-direct.py hose | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 3 | vndirect.py | arg[0] = hnx | lấy thông tin giá sàn hnx (realtime) từ vndirect | ./socket/vndirect.py | python vn-direct.py hnx | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 4 | vndirect.py | arg[0] = hnx30 | lấy thông tin giá sàn hnx30 (realtime) từ vndirect | ./socket/vndirect.py | python vn-direct.py hnx30 | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 5 | vndirect.py | arg[0] = vn30 | lấy thông tin giá sàn vn30 (realtime) từ vndirect | ./socket/vndirect.py | python vn-direct.py vn30 | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 6 | vndirect.py | arg[0] = upcom | lấy thông tin giá sàn upcom (realtime) từ vndirect | ./socket/vndirect.py | python vn-direct.py upcom | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 7 | start.sh | null | lấy chỉ số thị trường (vn-index, vn30-index, vnxallshare, hnx-index, hnx30-index, upcom) + thông tin các sàn: hose, hnx, hnx30, vn30, upcom từ vndirect. **Script này là tổng hợp của vndirect.py. Khi chạy script này thì không cần chạy đơn từng script vndirect.py** | ./socket/start.sh | sh start.sh | lấy thông tin chứng khoán | chỉ cần thực thi 1 lần, nên đặt trước vào khoảng thời gian thị trường giao dịch (9:00 ⟶ 15:00 hàng ngày) |
| 8 | stop.sh | null | dừng các script được thự thi từ start và tất cả các script vndirect.py | ./socket/stop.sh | sh stop.sh | lấy thông tin chứng khoán | script này không cần đặt lịch vì các vndirect.py đã tự động stop sau khi kết thúc thời gian giao dịch. Script hữu ích trong trường hợp muốn stop thủ công toàn bộ vndirect.py |
| 9 | vndirect.py | null | không cần thiết cho các spider khác. Spider này lấy danh sách thông tin doanh nghiệp có trên sàn chứng khoán (được vndirect cung cấp). Dữ liệu được cập nhật, không lưu dạng version | ./ tfhscrawler/spiders/vndirec.py | scrapy runspider ./ tfhscrawler/spiders/vndirec.py hoặc scrapy crawl vndirect | hỗ trợ API | nên đặt lịch chạy khoảng 5 – 10 ngày một lần để cập nhật danh sách thông tin doanh nghiệp, nếu dùng hỗ trợ cho API |
| 10 | vcmedia.py | null | lấy danh sách thông tin các công ty chứng khoán và link profile các công ty trên cafef (mục đích giống vndirect.py nhưng khác là lấy thông tin từ vcmedia (liên kết với cafef) và bao gồm cả các công ty đang không hoạt động trên sàn). Dữ liệu được cập nhật, không lưu dạng version | ./ tfhscrawler/spiders/vcmedia.py | scrapy runspider ./ tfhscrawler/spiders/vcmedia.py hoặc scrapy crawl vcmedia | hỗ trợ spider cafefprofile lấy thông tin profile các công ty | nên đặt lịch chạy khoảng 5 – 10 ngày một lần để cập nhật danh sách thông tin doanh nghiệp. link profile & xác định link nào sẽ sử dụng crawl dữ liệu |
| 11 | cafefprofile.py | null | lấy thông tin profile các công ty và lưu cơ sở dữ liệu dạng version | ./ tfhscrawler/spiders/cafefprofile.py | scrapy runspider ./ tfhscrawler/spiders/cafefprofile.py hoặc scrapy crawl cafefprofile | lấy thông tin profile các công ty chứng khoán | nên đặt lịch chạy sau giờ kết thúc giao dịch hàng ngày (sau 15:00) để có số liệu tổng hợp cuối ngày & chạy 1 ngày 1 lần |
| 12 | cafefstockmarket.py | null | lấy tin tức mới nhất từ mục thị trường chứng khoán của cafef. Nếu tin đã lưu sẽ bỏ qua | ./ tfhscrawler/spiders/cafefstockmarket.py | scrapy runspider ./ tfhscrawler/spiders/cafefstockmarket.py hoặc scrapy crawl cafefstockmarket | lấy tin tức mục thị trường chứng khoán từ cafef | nên đặt lịch chạy 10 – 15 phút một lần |
| 13 | vtveco.py | null | lấy tin tức mới nhất từ mục kinh tế của vtv. Nếu tin đã lưu sẽ bỏ qua | ./ tfhscrawler/spiders/vtveco.py | scrapy runspider ./ tfhscrawler/spiders/vtveco.py hoặc scrapy crawl vtveco | lấy tin tức mục kinh tế từ vtv | nên đặt lịch chạy 10 – 15 phút một lần |
| 14 | ncdtbus.py | null | lấy tin tức mới nhất từ mục tin tức doanh nghiệp của nhipcaudautu. Nếu tin đã lưu sẽ bỏ qua | ./ tfhscrawler/spiders/ncdtbus.py | scrapy runspider ./ tfhscrawler/spiders/ncdtbus.py hoặc scrapy crawl ncdtbus | lấy tin tức mục doanh nghiệp từ nhipcaudautu | nên đặt lịch chạy 10 – 15 phút một lần |

## Daemon

mô tả daemon

# Mô tả cơ sở dữ liệu

## Mối liên hệ giữa command & cơ sở dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| stt | command | bảng tương tác | mô tả |
| 1 | python vn-direct.py market | * vndc\_urls (select) * vndc\_commands (select) * vndc\_cfc (insert) * vndc\_crawl\_cfc (insert) * vndc\_crawl\_markets (insert) * vndc\_markets (insert) | thực thi 3 socket command: info, cfc, market (cấu trúc lấy từ vndc\_commands), đổ dữ liệu vào các bảng vndc\_cfc, vndc\_markets. Dữ liệu các bảng vndc\_crawl\* lưu log thông tin crawl |
| 2 | python vndirect.py hose | * vndc\_urls (select) * vndc\_commands (select) * vndc\_hose (insert) * vndc\_crawl\_hose (insert) | thực thi socket command: hose (cấu trúc lấy từ vndc\_commands), đổ dữ liệu vào bảng vndc\_hose. bảng vndc\_crawl\_hose lưu log thông tin crawl |
| 3 | python vndirect.py hnx | * vndc\_urls (select) * vndc\_commands (select) * vndc\_hnx (insert) * vndc\_crawl\_hnx (insert) | thực thi socket command: hnx (cấu trúc lấy từ vndc\_commands), đổ dữ liệu vào bảng vndc\_hnx. bảng vndc\_crawl\_hnx lưu log thông tin crawl |
| 4 | python vndirect.py hnx30 | * vndc\_urls (select) * vndc\_commands (select) * vndc\_hnx30 (insert) * vndc\_crawl\_hnx30 (insert) | thực thi socket command: hnx30 (cấu trúc lấy từ vndc\_commands), đổ dữ liệu vào bảng vndc\_hnx30. bảng vndc\_crawl\_hnx30 lưu log thông tin crawl |
| 5 | python vndirect.py vn30 | * vndc\_urls (select) * vndc\_commands (select) * vndc\_vn30 (insert) * vndc\_crawl\_vn30 (insert) | thực thi socket command: hose (cấu trúc lấy từ vndc\_commands), đổ dữ liệu vào bảng vndc\_vn30. bảng vndc\_crawl\_vn30 lưu log thông tin crawl |
| 6 | python vndirect.py upcom | * vndc\_urls (select) * vndc\_commands (select) * vndc\_upcom (insert) * vndc\_crawl\_upcom (insert) | thực thi socket command: hose (cấu trúc lấy từ vndc\_commands), đổ dữ liệu vào bảng vndc\_upcom. bảng vndc\_crawl\_upcom lưu log thông tin crawl |
| 7 | scrapy crawl vcmedia | * vcme\_companies (insert/update) * vcme\_crawl\_companies (insert) | lấy thông tin doanh nghiệp từ vcmedia (liên kết với cafef), đổ dữ liệu vào vcme\_companies, nếu thông tin công ty chưa có sẽ tạo mới, ngược lại cập nhật. bảng vcme\_crawl\_companies lưu log thông tin crawl |
| 8 | scrapy crawl vndirect | * vndc\_companies (insert/update) * vndc\_crawl\_companies (insert) | lấy thông tin doanh nghiệp từ vndirect, đổ dữ liệu vào vndc\_companies, nếu thông tin công ty chưa có sẽ tạo mới, ngược lại cập nhật. bảng vndc\_crawl\_companies lưu log thông tin crawl |
| 9 | scrapy crawl cafefprofile | * vcme\_companies (select) * scff\_sprofiles (insert) * scff\_bprofiles (insert) | lấy danh sách link công ty từ bảng vcme\_companies, crawl thông tin hồ sơ từng công ty & đổ dữ liệu vào scff\_sprofiles (dữ liệu thay đổi thường xuyên: các thông số chứng khoán), scff\_bprofiles (dữ liệu thay đổi không thường xuyên: thông tin cơ bản, avatar, ban lãnh đạo, thông tin tài chính…) |
| 10 | scrapy crawl cafefstockmarket | * articles (insert) * crawl\_articles (insert) | lấy tin tức mới nhất từ mục thị trường chứng khoán của cafef, đổ dữ liệu vào bảng articles. Nếu tin đã có trong cơ sở dữ liệu rồi sẽ bỏ qua. Bảng crawl\_articles lưu log thông tin đã crawl được |
| 11 | scrapy crawl vtveco | * articles (insert) * crawl\_articles (insert) | lấy tin tức mới nhất từ mục kinh tế của vtv, đổ dữ liệu vào bảng articles. Nếu tin đã có trong cơ sở dữ liệu rồi sẽ bỏ qua. Bảng crawl\_articles lưu log thông tin đã crawl được |
| 12 | scrapy crawl ncdtbus | * articles (insert) * crawl\_articles (insert) | lấy tin tức mới nhất từ mục doanh nghiệp của nhipcaudautu, đổ dữ liệu vào bảng articles. Nếu tin đã có trong cơ sở dữ liệu rồi sẽ bỏ qua. Bảng crawl\_articles lưu log thông tin đã crawl được |

## Thông tin cơ sở dữ liệu

Thông tin chi tiết cơ sở dữ liệu xin xem thêm file <database.xlsx> cùng thư mục

# Dữ liệu hỗ trợ API

Các dữ liệu chứng khoán vì lấy dữ liệu từ socket nên các thông tin lưu trong cơ sở dữ liệu được cập nhật realtime. Tại thời điểm cập nhật, khối lượng dữ liệu & tần suất kết nối sẽ tương đối lớn nên để hỗ trợ API & giảm tải cho cơ sở dữ liệu khi API truy xuất/tính toán để lấy thông tin giá hiện tại, module crawl hỗ trợ lưu các thông tin realtime này ra memcache và API có thể truy xuất các dữ liệu này để phục vụ trả về cho client.

Mặc định, hỗ trợ này được tắt (False), để bật cấu hình, set lại key **APIDATA** trong file ./socket/config.py thành True, và ngược lại thành False nếu muốn tắt. Khi bật APIDATA, yêu cầu server cần cài memcached & cấu hình thông tin memcached trong key MEMCACHE của file cấu hình trên (xem thêm mục [Cấu hình](#_Cấu_hình_4) socket).

Các dữ liệu sau được cache:

* **thông tin giá thị trường**: giá thị trường được cập nhật realtime, mặc định được set vào key **vndc\_markets** (key có thể thay đổi trong file cấu hình, xem thêm mục [Cấu hình](#_Cấu_hình_4) socket) trong memcached dưới dạng json là một mảng danh sách với key là mã sàn tương ứng (mã sàn xin xem trong bảng vndc\_floors, field code. Lưu ý: ngoài các mã sàn được liệt kê trong vndc\_floors, socket server còn trả về một số mã sàn khác)
* **thông tin giá sàn hose**: các thông số sàn host được cập nhật realtime, mặc định được set vào key **vndc\_hose** (key có thể thay đổi trong file cấu hình, xem thêm mục [Cấu hình](#_Cấu_hình_4) socket) trong memcached dưới dạng json là một mảng danh sách với key là mã chứng khoán tương ứng (mã chứng khoán xin xem thêm trong bảng vndc\_companies, field symbol). Thông tin vị trí từng giá trị trong danh sách sẽ tương ứng với từng field trong bảng vndc\_hose từ floor\_code ⟶ sequence
* **thông tin giá sàn hnx**: các thông số sàn host được cập nhật realtime, mặc định được set vào key **vndc\_hnx** (key có thể thay đổi trong file cấu hình, xem thêm mục [Cấu hình](#_Cấu_hình_4) socket) trong memcached dưới dạng json là một mảng danh sách với key là mã chứng khoán tương ứng (mã chứng khoán xin xem thêm trong bảng vndc\_companies, field symbol). Thông tin vị trí từng giá trị trong danh sách sẽ tương ứng với từng field trong bảng vndc\_hnx từ floor\_code ⟶ sequence
* **thông tin giá sàn hnx30**: các thông số sàn host được cập nhật realtime, mặc định được set vào key **vndc\_hnx30** (key có thể thay đổi trong file cấu hình, xem thêm mục [Cấu hình](#_Cấu_hình_4) socket) trong memcached dưới dạng json là một mảng danh sách với key là mã chứng khoán tương ứng (mã chứng khoán xin xem thêm trong bảng vndc\_companies, field symbol). Thông tin vị trí từng giá trị trong danh sách sẽ tương ứng với từng field trong bảng vndc\_hnx30 từ floor\_code ⟶ sequence
* **thông tin giá sàn vn30**: các thông số sàn host được cập nhật realtime, mặc định được set vào key **vndc\_vn30** (key có thể thay đổi trong file cấu hình, xem thêm mục [Cấu hình](#_Cấu_hình_4) socket) trong memcached dưới dạng json là một mảng danh sách với key là mã chứng khoán tương ứng (mã chứng khoán xin xem thêm trong bảng vndc\_companies, field symbol). Thông tin vị trí từng giá trị trong danh sách sẽ tương ứng với từng field trong bảng vndc\_vn30 từ floor\_code ⟶ sequence
* **thông tin giá sàn upcom**: các thông số sàn host được cập nhật realtime, mặc định được set vào key **vndc\_upcom** (key có thể thay đổi trong file cấu hình, xem thêm mục [Cấu hình](#_Cấu_hình_4) socket) trong memcached dưới dạng json là một mảng danh sách với key là mã chứng khoán tương ứng (mã chứng khoán xin xem thêm trong bảng vndc\_companies, field symbol). Thông tin vị trí từng giá trị trong danh sách sẽ tương ứng với từng field trong bảng vndc\_upcom từ floor\_code ⟶ sequence

Các dữ liệu được cập nhật realtime theo từng field, chỉ khi có dữ liệu mới tương ứng với mã sàn hoặc mã chứng khoán được trả về từ socket server của vndirect.

Sau khi kết thúc giao dịch trong ngày, thông tin trên memcache sẽ phản giá giá/thông số giao dịch cuối cùng trong ngày trên thị trường + các sàn chứng khoán.

# Cấu hình

Cấu hình module được chia là 2 phần: cấu hình cho socket (lấy dữ liệu chứng khoán) & cấu hình cho scrapy (lấy dữ liệu hồ sơ công ty, tin tức tổng hợp). Cấu hình socket được đặt tại: ./socket/config.py, cấu hình scrapy được đặt tại: ./tfhscrawler/tfhscrawler/setting.py. Các key cấu hình được đặt dưới dạng dict hoặc variable của python. Dưới đây là bảng các tham sô cấu hình cho hai phần này.

## Cấu hình socket

|  |  |  |  |
| --- | --- | --- | --- |
| key | kiểu giá trị | mặc định | mô tả |
| DB | dict |  | thông tin truy cập cơ sở dữ liệu |
| HOST | string | localhost | database host |
| USER | string | user | database user |
| PWD | string | pass | user password |
| DB | string | db | database name |
| MEMCACHE | dict |  | thông tin memcached server. Thông tin này được sử dụng khi key APIDATA được set True. memcached server lưu thông tin giá/thông số chứng khoán hỗ trợ API |
| HOST | string | 127.0.0.1 | memcache host |
| PORT | string | 11211 | memcache port |
| DEBUG | boolean|int | 0 | đặt chế độ debug |
| APIDATA | boolean | False | bật/tắt chế độ hỗ trợ APIDATA. Khi bật yêu cầu server cần cài đặt memcached |
| TIMEZONE | string | Asia/Ho\_Chi\_Minh | giá trị timezone cho phần lưu log/tính toán thời gian cho script phần socket. Format theo timezone của python |
| DBTZ | string | +7:00 | giá trị timezone khi xử lý trong cơ sở dữ liệu. Forrmat theo timezone của mysql |
| FREDATA | dict |  | dữ liệu của các bảng lưu thông tin chứng khoán sẽ tự động được xóa sau một khoảng thời gian. Tham số này quy định danh sách bảng và khoảng thời gian (chu kì) xóa dữ liệu |
| ACTIVE | string | 1 days  KHÔNG CÒN SỬ DỤNG | thời gian dữ liệu đươc giữ lại. sau khoảng thời gian này, dữ liệu sẽ bị xóa. Format dạng time string của python (nên đặt theo ngày, các format dạng khác ngày không khuyến cáo). Việc xóa dữ liệu được thực hiện khi script bắt đầu, nếu thời gian xóa trùng cấu hình. Thời gian được đem ra so sánh là thời điểm gần nhất mà dữ liệu được crawl trong bảng cấu hình trong key DTABLE |
| TABLE | list | [  'vndc\_crawl\_hnx', 'vndc\_crawl\_hnx30', 'vndc\_crawl\_hose', 'vndc\_crawl\_upcom', 'vndc\_crawl\_vn30', 'vndc\_crawl\_markets',  'vndc\_hnx', 'vndc\_hnx30', 'vndc\_hose', 'vndc\_upcom', 'vndc\_vn30', 'vndc\_markets'  ] | danh sách các bảng sẽ được xóa dữ liệu (thao tác truncate) khi trùng thời điểm ACTIVE |
| DTABLE | dict |  | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu chứng khoán |
| NAME | string | vndc\_hose | tên bảng đối chiếu |
| FIELD | string | trading\_date | trường thời gian đối chiếu (milisecond unix timestamp) |
| CMNTABLES | dict |  | các bảng chung thường được sử dụng |
| VNDCCMD | string | vndc\_commands | bảng chứa danh sách socket command |
| VNDCURL | string | vndc\_urls | bảng chứa danh sách các socket server của vndirect |
| SOCKETS | dict |  | các cấu hình liên quan đến socket |
| CFC | dict |  | cấu hình lưu thông tin socket trả về liên quan đến giá sàn/giá trần (hiện tại không cần thiết được sử dụng) |
| CMD | list | ['cfc'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin giá sàn/giá trần |
| TABLE | string | vndc\_cfc | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_cfc | bảng lưu dữ liệu log khi crawl |
| ACTIVE | string | forever | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| MARKET | dict |  | cấu hình lưu thông tin socket trả về liên quan đến thị trường |
| CMD | list | ['info', 'cfc', 'market'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin thị trường. command ‘market’ là cần thiết. |
| TABLE | string | vndc\_markets | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_markets | bảng lưu dữ liệu log khi crawl |
| MEMKEY | string | vndc\_markets | key memcache lưu thông tin realtime |
| ACTIVE | string | 1 days | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| TABLET | list | ['vndc\_crawl\_markets', 'vndc\_markets'] | danh sách bảng dữ liệu được sử dụng khi xóa định kì (truncate) |
| FIELDT | list | ['vndc\_markets', 'trading\_date'] | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu |
| HOSE | dict |  | cấu hình lưu thông tin socket trả về liên quan đến sàn hose |
| CMD | list | ['info', 'hose'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin thị trường. command ‘hose’ là cần thiết. |
| TABLE | string | vndc\_hose | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_hose | bảng lưu dữ liệu log khi crawl |
| MEMKEY | string | vndc\_hose | key memcache lưu thông tin realtime |
| ACTIVE | string | 1 days | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| TABLET | list | ['vndc\_crawl\_hose', 'vndc\_hose'] | danh sách bảng dữ liệu được sử dụng khi xóa định kì (truncate) |
| FIELDT | list | ['vndc\_hose', 'trading\_date'] | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu |
| VN30 | dict |  | cấu hình lưu thông tin socket trả về liên quan đến sàn vn30 |
| CMD | list | ['info', 'vn30'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin thị trường. command ‘vn30’ là cần thiết. |
| TABLE | string | vndc\_vn30 | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_vn30 | bảng lưu dữ liệu log khi crawl |
| MEMKEY | string | vndc\_vn30 | key memcache lưu thông tin realtime |
| ACTIVE | string | 1 days | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| TABLET | list | ['vndc\_crawl\_vn30', 'vndc\_vn30'] | danh sách bảng dữ liệu được sử dụng khi xóa định kì (truncate) |
| FIELDT | list | ['vndc\_vn30', 'trading\_date'] | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu |
| HNX | dict |  | cấu hình lưu thông tin socket trả về liên quan đến sàn hnx |
| CMD | list | ['info', 'hnx'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin thị trường. command ‘hnx’ là cần thiết. |
| TABLE | string | vndc\_hnx | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_hnx | bảng lưu dữ liệu log khi crawl |
| MEMKEY | string | vndc\_hnx | key memcache lưu thông tin realtime |
| ACTIVE | string | 1 days | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| TABLET | list | ['vndc\_crawl\_hnx', 'vndc\_hnx'] | danh sách bảng dữ liệu được sử dụng khi xóa định kì (truncate) |
| FIELDT | list | ['vndc\_hnx', 'trading\_date'] | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu |
| HNX30 | dict |  | cấu hình lưu thông tin socket trả về liên quan đến sàn hnx30 |
| CMD | list | ['info', 'hnx30'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin thị trường. command ‘hnx30’ là cần thiết. |
| TABLE | string | vndc\_hnx30 | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_hnx30 | bảng lưu dữ liệu log khi crawl |
| MEMKEY | string | vndc\_hnx30 | key memcache lưu thông tin realtime |
| ACTIVE | string | 1 days | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| TABLET | list | ['vndc\_crawl\_hnx30', 'vndc\_hnx30'] | danh sách bảng dữ liệu được sử dụng khi xóa định kì (truncate) |
| FIELDT | list | ['vndc\_hnx30', 'trading\_date'] | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu |
| UPCOM | dict |  | cấu hình lưu thông tin socket trả về liên quan đến sàn upcom |
| CMD | list | ['info', 'upcom'] | danh sách các command được sử dụng gọi đến socket server để lấy thông tin thị trường. command ‘upcom’ là cần thiết. |
| TABLE | string | vndc\_upcom | bảng lưu dữ liệu được crawl |
| TABLEC | string | vndc\_crawl\_upcom | bảng lưu dữ liệu log khi crawl |
| MEMKEY | string | vndc\_upcom | key memcache lưu thông tin realtime |
| ACTIVE | string | 1 days | forever: không xóa, [x] days: tự động được xóa (truncate) sau khoảng thời gian (số ngày) được cấu hình |
| TABLET | list | ['vndc\_crawl\_upcom', 'vndc\_upcom'] | danh sách bảng dữ liệu được sử dụng khi xóa định kì (truncate) |
| FIELDT | list | ['vndc\_upcom', 'trading\_date'] | bảng và trường thời gian sẽ được đem so sánh với thời gian tại key ACTIVE để xác định thời điểm xóa dữ liệu |
| LOG | dict | {  'version': 1,  'disable\_existing\_loggers': False, # this fixes the problem  'formatters': {  'standard': {  'format': '%(asctime)s [%(levelname)s]: %(message)s'  },  },  'handlers': {  'default': {  'level':'INFO',  'class':'logging.handlers.TimedRotatingFileHandler',  'formatter': 'standard',  'filename': 'logs/vnds-%s.log',  'when': 'D',  'backupCount': 30,  'encoding': 'utf8'  },  },  'loggers': {  '': {  'handlers': ['default'],  'level': 'INFO',  'propagate': True  }  },  'levels': {  'SQ': 55,  'SA': 60  }  } | cấu hình log theo format của python. Mặc định file log dưới dạng 'logs/vnds-%s.log' (%s = tên cấu hình socket: thị trường hoặc tên sàn). Rotate theo ngày, backup 30 ngày. Thông tin chi tiết xem thêm <https://docs.python.org/2/library/logging.html> |

## Cấu hình Scrapy

Ngoài các cấu hình scrapy đã cung cấp, phần cấu hình được thêm cho các spider. Các cấu hình được sử dụng cho scrapy, xin xem thêm tại <https://doc.scrapy.org/en/latest/topics/settings.html>.

|  |  |  |  |
| --- | --- | --- | --- |
| key | kiểu giá trị | mặc định | mô tả |
| BOTNAME | string | tfhscrawler | tên bot tạo theo scrapy project |
| SPIDER\_MODULES | list | ['tfhscrawler.spiders'] | danh sách các spider module |
| NEWSPIDER\_MODULE | string | tfhscrawler.spiders | module được sử dụng khi tạo spider mới |
| ROBOTSTXT\_OBEY | boolean | False | crawl theo quy định trong robots.txt (nếu site tồn tại file này) |
| COOKIES\_ENABLED | boolean | False | sử dụng cookie trong quá trình crawl hay không |
| EXTENSIONS | dict | {  'scrapy.extensions.telnet.TelnetConsole': None,  } | bật/tắt các extension. Mặc định tắt telnet console. Các extension khác xin xem thêm tài liệu hướng dẫn cấu hình từ scrapy |
| DB | dict |  | thông tin kết nối cơ sở dữ liệu |
| HOST | string | localhost | database host |
| USER | string | user | database user |
| PWD | string | password | user password |
| DB | string | db | database name |
| TIMEZONE | string | Asia/Ho\_Chi\_Minh | giá trị timezone cho spider. Format theo timezone của python |
| DBTZ | string | +7:00 | giá trị timezone khi xử lý trong cơ sở dữ liệu. Forrmat theo timezone của mysql |
| CAFEFPROFILE | dict |  | cấu hình cho spider lấy thông tin hồ sơ doanh nghiệp từ cafef (cafefprofile spider) |
| COMTABLE | string | vcme\_companies | bảng chứa danh sách thông tin các công ty & link được crawl dữ liệu |
| STABLE | string | scff\_sprofiles | bảng sẽ lưu thông tin hồ sơ doanh nghiệp, dữ liệu này được xác định là thường xuyên thay đổi (xem thêm mục |
| BTABLE | string | scff\_bprofiles |  |
| DOWNLOAD\_DELAY | float | 1 |  |
| API\_DOWNLOAD\_DELAY | float | 0.25 |  |
| BURLS |  |  |  |
| V1 |  |  |  |
| FININF |  |  |  |
| QUARTER |  |  |  |
| YEAR |  |  |  |
| 6MONS |  |  |  |
| BASINF |  |  |  |
| AFFINF |  |  |  |
| FINRPT |  |  |  |
| V2 |  |  |  |
| FININF |  |  |  |
| QUARTER |  |  |  |
| YEAR |  |  |  |
| 6MONS |  |  |  |
| BASINF |  |  |  |
| LOINF |  |  |  |
| AFFINF |  |  |  |
| FINRPT |  |  |  |
| BTIME |  |  |  |
| START |  |  |  |
| INTERVAL |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Log

# Triển khai trên server