SkadiHouse

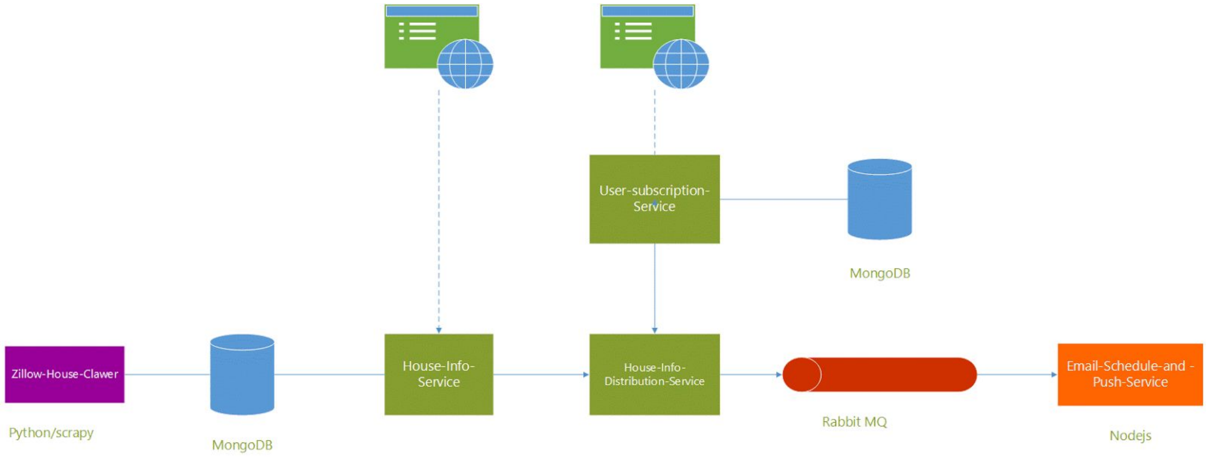
Introduction

A house information crawling and pushing project for real estate. House information will crawled from internet, and data will be periodic analyzed and push to subscribed user by email.

Use Case

Users are able to search real estate house information online or subscribe by setting some constraints and getting house information email periodically.

Project Architecture


User-subscription-Service
MongoDB
Zillow-House-Clawer
Python/scrapy
MongoDB
House-Info-Service
House-Info-Distribution-Service
Rabbit MQ
Email-Schedule-and -Push-Service
Nodejs

Features

- Python Scrapy Web Crawler
- SpringBoot, Spring Data, JPA Query
- MEAN Architecture
- Spring-Cloud Architecture
- Docker
- Separation of frontend and backend
- Message Queue: RabbitMQ
- Push and Subscribe Pattern

Infrastructure

Eureka

Eureka is a REST based service that is primarily used in the cloud for locating services for the purpose of load balancing and failover of middle-tier servers. In this project, Eureka was used for microservice registry and service discovery.

- Server Information: http://localhost:8761

Hystrix

Hystrix is a latency and fault tolerance library designed to isolate points of access to remote systems, services and 3rd party libraries, stop cascading failure and enable resilience in complex distributed systems where failure is inevitable.

- Server Information: http://localhost:7979

RabbitMQ

RabbitMQ is an open source message broker software (sometimes called message-oriented middleware) that implements the Advanced Message Queuing Protocol (AMQP).

- Server Information: http://localhost:5672
- Server management web portal: http://localhost:15672

Detailed design

Zillow-House-Crawler

- Python Scrapy Pipline
- Data Source,MongoDB
- Strategy to Anti-Bot: User Agent, Random Proxy, Disable Cookie, Set Header

House-Info-Service

- House information persistence
- Backend API for house information.
- Server Information: http://localhost:9000
- Rest API
    - GET:
        - /houseinfo/get/city/{city}
        - /houseinfo/get/zip/{zip}
        - /houseinfo/get/address/{address}
        - /houseinfo/get/days/{daysOnZillow}
        - /houseinfo/get/sellprice/{sellPrice}
        - /houseinfo/get/sellprice/{startPrice}/{endPrice}
        - /houseinfo/get/subscribed/{daysOnZillow}/{city}/{startPrice}/{endPrice}

User-Subscription-Service

- Suscribed user information persistence
- Backend API for user information
- Server Information: http://localhost:9001
- Rest API
    - POST: /userinfo/post
    - DELETE:/userinfo/delete/{email}
    - GET:
        - /userinfo/get
        - /userinfo/get/email/{email}/end

House-Info-Distribution-Service

This service will periodicity get subscribed userinfo list and search each user specified house information from house information service. Then pack and post these information as json to RabbitMQ.

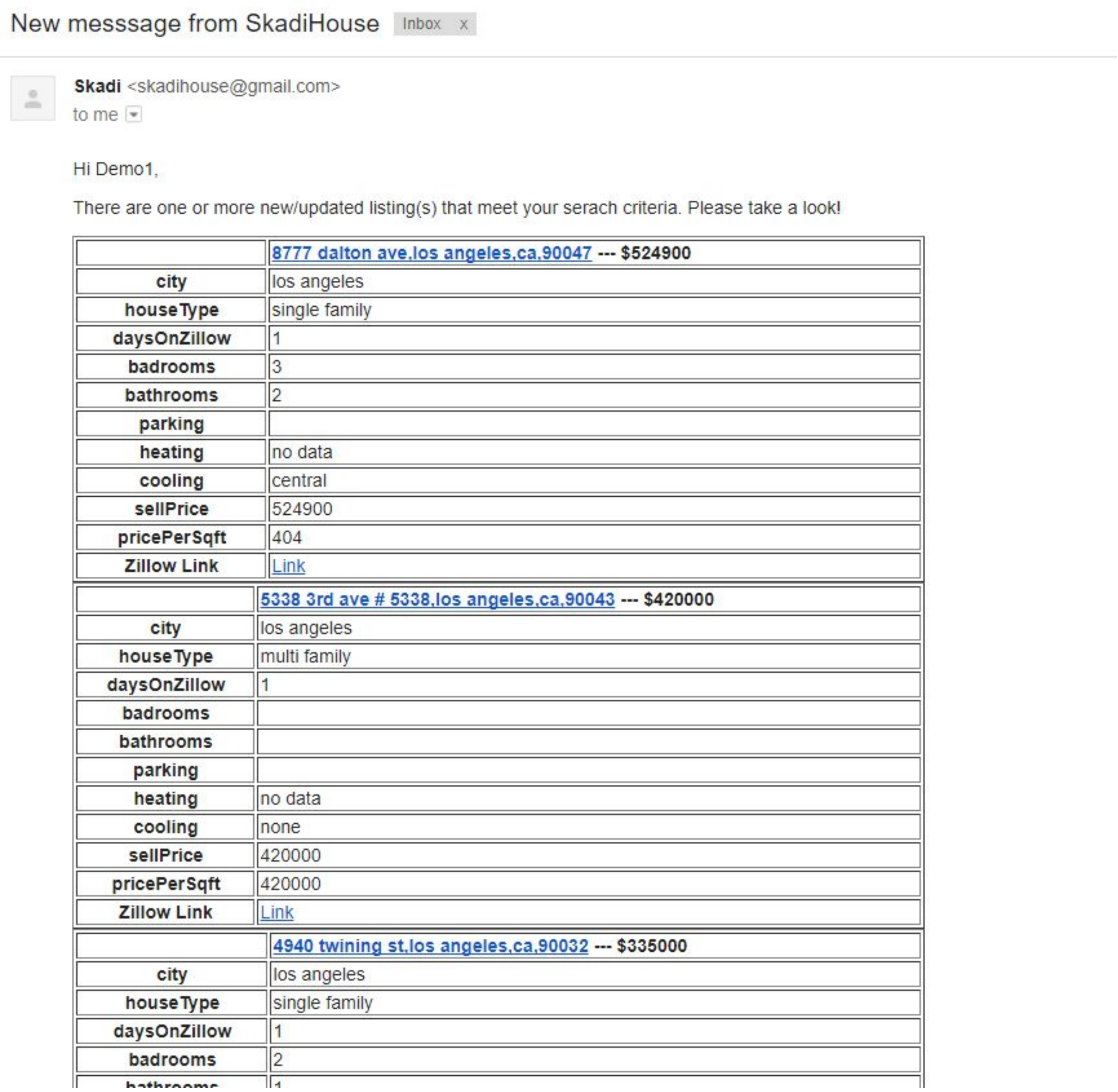
- Server Information: http://localhost:9002

Email-Consumer-Service

This Nodejs service will listen and consume RabbitMQ message. Parse object then sent email to each subscribed user.

- Middleware:
    - amqplib
    - nodemailer
    - node-json2h

- Screenshot



New messsage from SkadiHouse

Skadi <skadihouse@gmail.com>
to me

Hi Demo1,

There are one or more new/updated listing(s) that meet your serach criteria. Please take a look!

	8777 dalton ave,los angeles,ca,90047 --- $524900
city	los angeles
houseType	single family
daysOnZillow	1
badrooms	3
bathrooms	2
parking	
heating	no data
cooling	central
sellPrice	524900
pricePerSqft	404
Zillow Link	Link

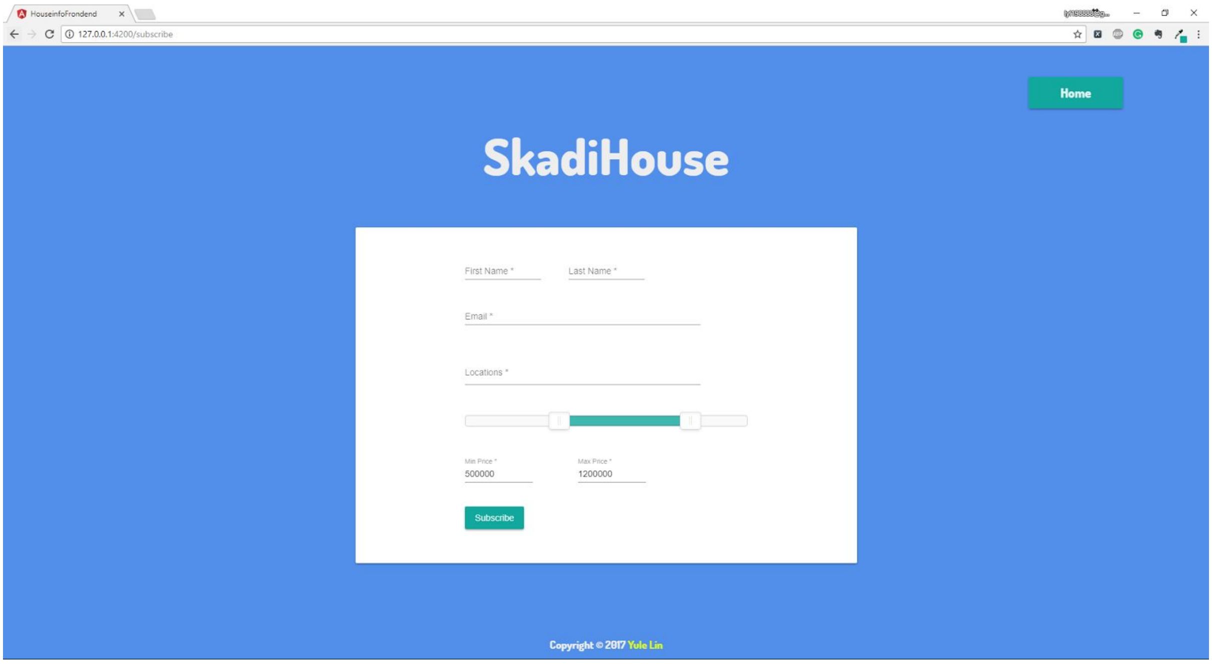
	5338 3rd ave # 5338,los angeles,ca,90043 --- $420000
city	los angeles
houseType	multi family
daysOnZillow	1
badrooms	
bathrooms	
parking	
heating	no data
cooling	none
sellPrice	420000
pricePerSqft	420000
Zillow Link	Link

	4940 twining st,los angeles,ca,90032 --- $335000
city	los angeles
houseType	single family
daysOnZillow	1
badrooms	2
bathrooms	1

Skadihoue-frontend

Angular frontend web for house information search and user suscription.

- Typescript, ES6, Promise
- Two-way binding
- Ng router
- Data validation
- Human Interactive Design
  - Auto Complate Input
  - Sidebar Binding Input Data
  - Message Snackbar
- Middleware:
  - angular material 2
  - ng2-nouislider
- Screenshot:



Min Price * 500000

Max Price * 1200000

Future work

- Crawler automation
- User Pushed History
- Mechine learning for housing data analysis
- Redis/Memcached LRU cache
- Scale to Distributed Architecture
- Deploy docker image