



**크롤링 프로젝트
최종발표
(2조)**

2021.03.19

임현수
최민권

목차

1. 소개
2. 시스템 구조
3. 프로세스별 상세 설명
4. 프로세스 예시 화면
5. 웹 서비스 실습
6. 추후 개선 방향





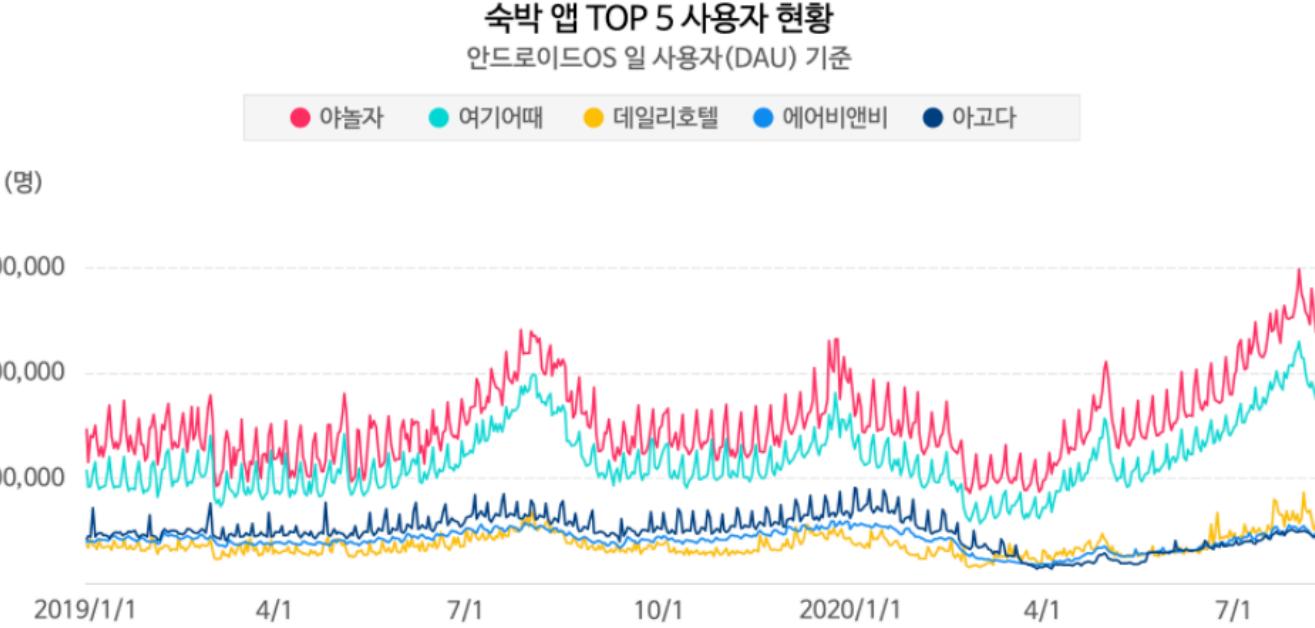
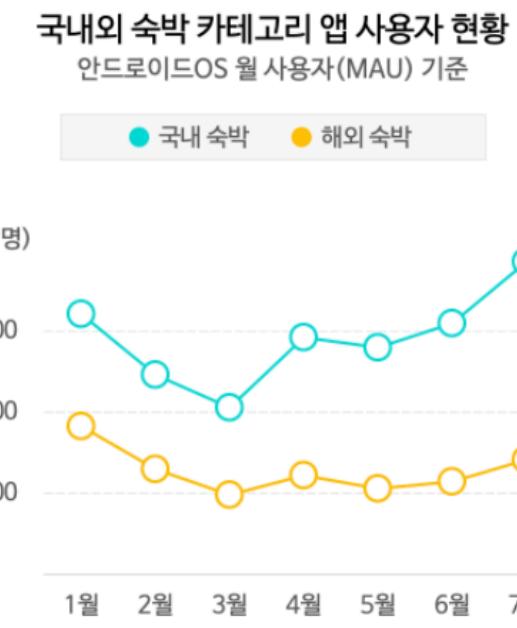
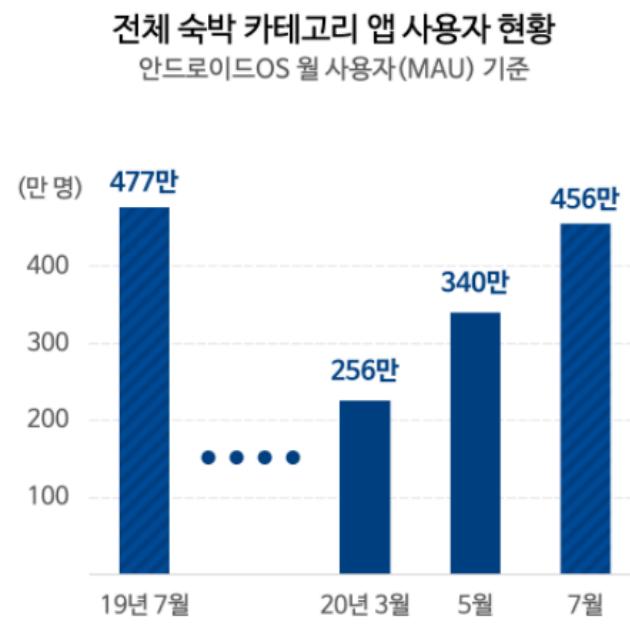
소개

시장 현황



코로나로 인해 해외 숙박 시장이 어려워진 반면, 국내 숙박 시장은 오히려 활발해짐

성장세를 보이는 국내숙박 앱 사용자 수



상위권을 모두 차지한 국내 숙박 앱



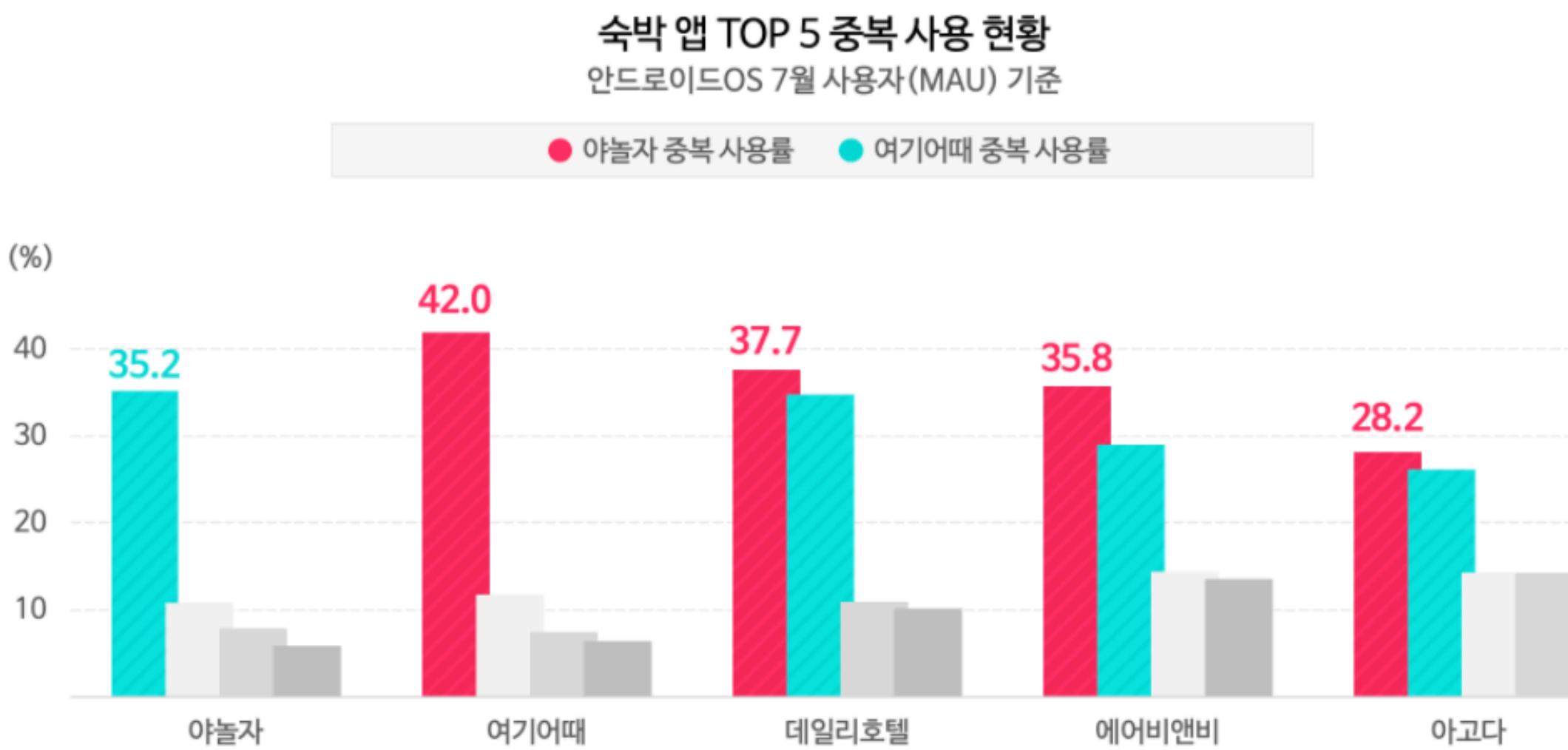
(출처: 모바일인덱스)

문제 인식



동일 상품이라도 플랫폼 간 가격이 상이하여, 비교하다가 상품을 놓치는 불편함 발생

숙박 앱 중복 사용 현황



(출처: 모바일인덱스)

동일 상품이지만 상이한 가격

파크 하얏트 서울
★ 4.9 / 5 후기 117개
3월 18일(목) ~ 3월 19일(금)

스탠다드 도심전망 1킹
단독★Good Evening PKG, 휴스불가
기준 2명 / 최대 2명
숙박 18:00 부터

디럭스 1킹
단독★Good Evening PKG, 휴스불가
기준 2명 / 최대 3명
숙박 18:00 부터

310,750
332,750
394,625

(21.03.18, 10:40 검색 기준)

목적 및 기대효과



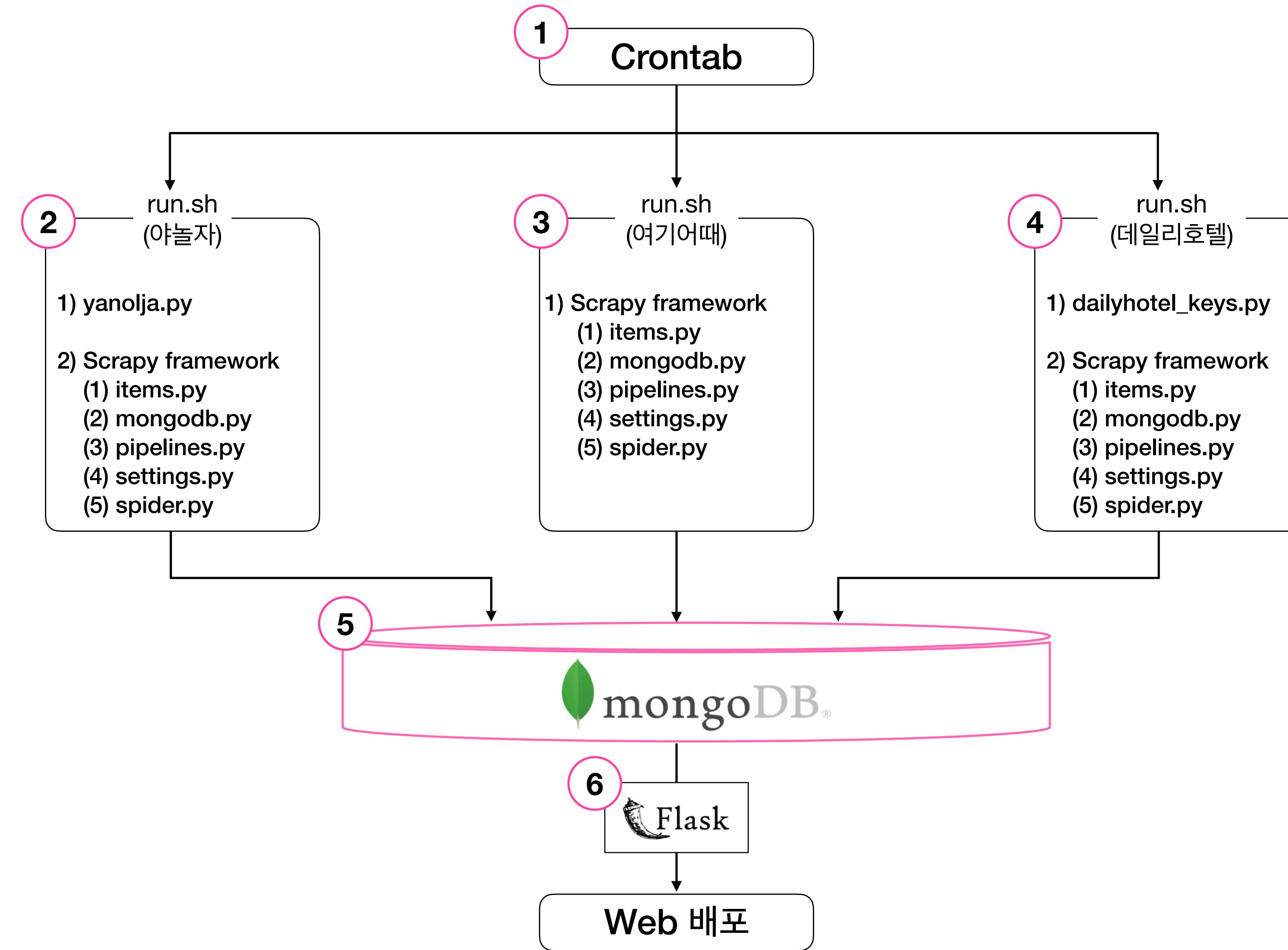
원하는 상품의 최저가를 한 곳에서 쉽게 비교, 결제까지 하여 불편을 해소하고자 함





시스템 구조

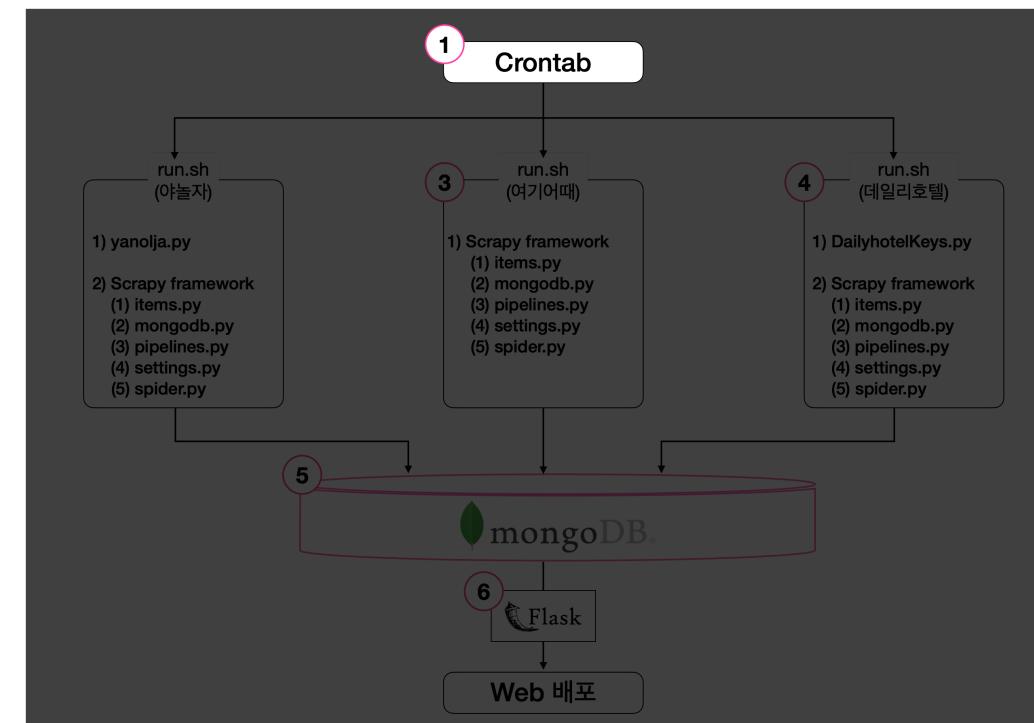
시스템 구조





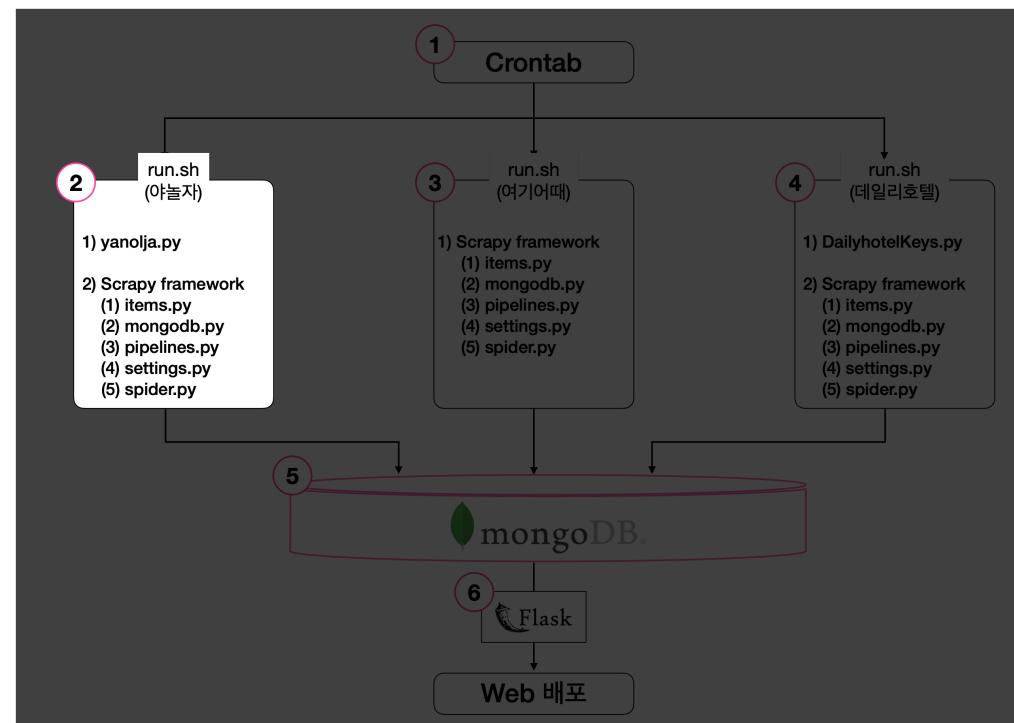
프로세스별 상세설명

1) Crontab



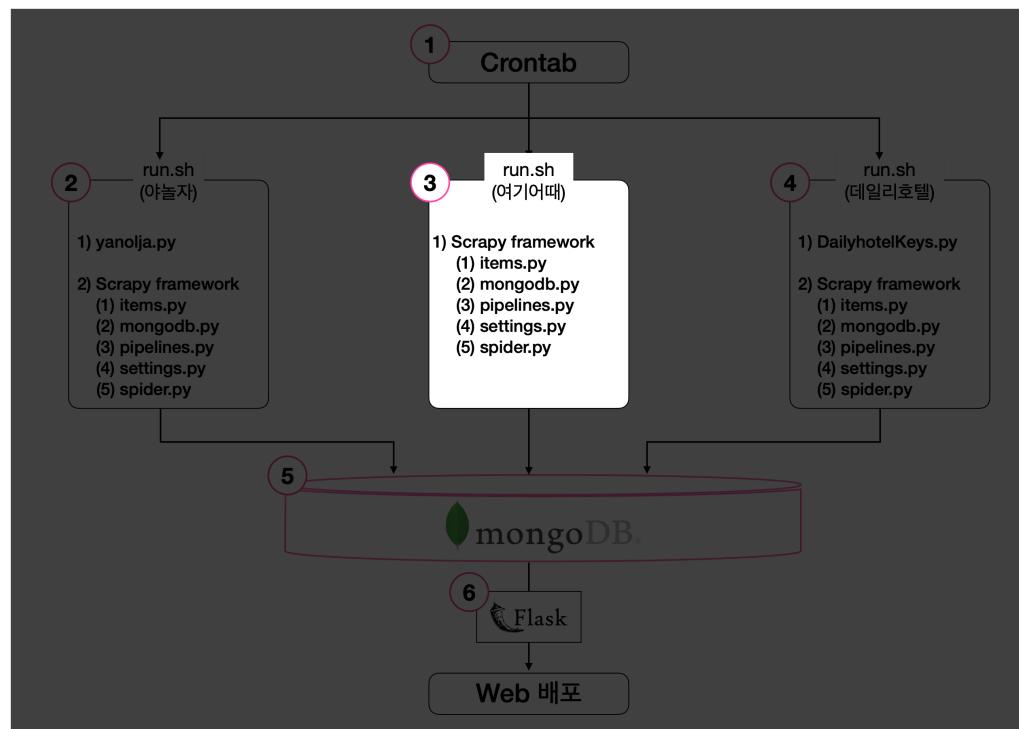
- * 3개의 쉘스크립트 (야놀자, 여기어때, 데일리호텔 각 1개)를 실행하는 역할 수행
- * 5분 주기로 실행되도록 세팅

2) 쉘스크립트 (야놀자)



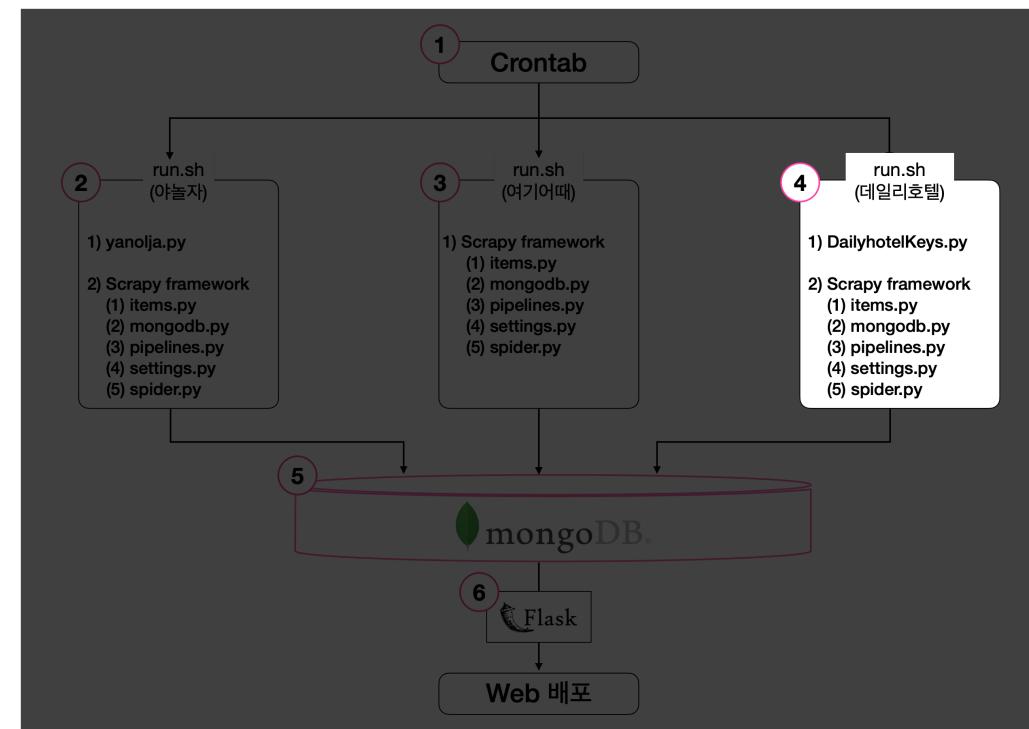
- * 야놀자 사이트에서 호텔별 상세정보를 크롤링하여 취합한 후, MongoDB에 데이터를 적재하도록 제작됨
- * yanolja.py 모듈은 호텔별 고유ID값을 가져오는 역할 수행, 호텔별 고유ID값은 각 호텔별 상세페이지 url 생성 시 필요
- * Scrapy framework (Spider)를 사용하여 크롤링을 진행, item.py에 데이터를 적재하고, 적재 완료 후 pipelines.py가 실행되며 MongoDB로 데이터 전송

3) 쉘스크립트 (여기어때)



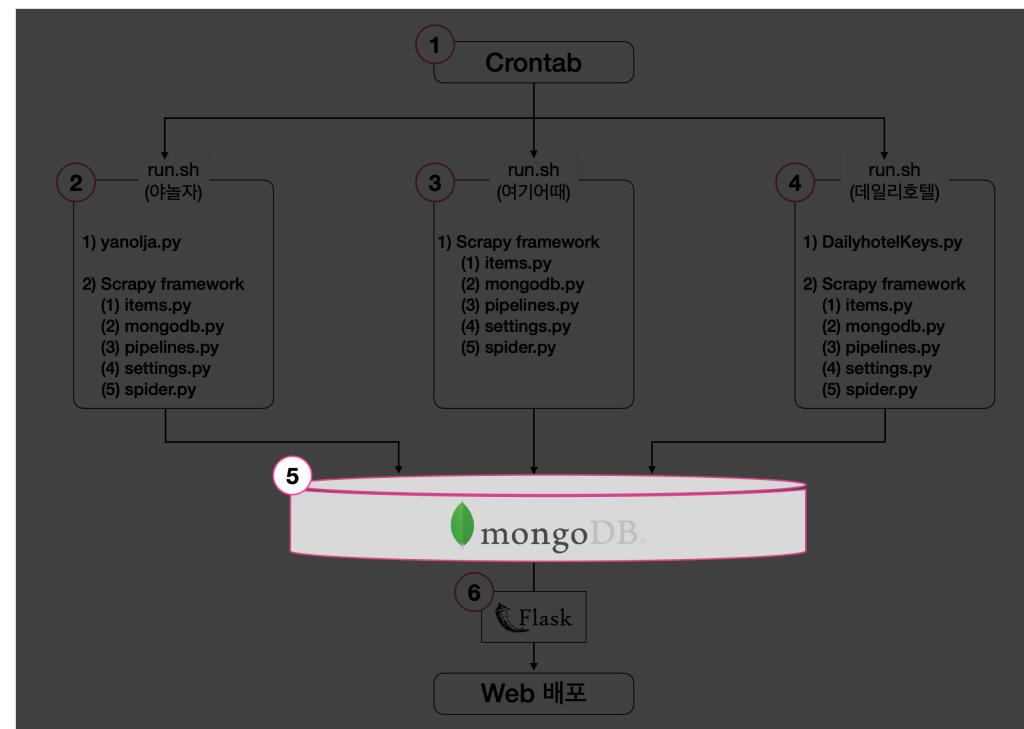
- * 여기어때 사이트에서 호텔별 상세정보를 크롤링하여 취합한 후, MongoDB에 데이터를 적재하도록 제작됨
- * 별도 모듈은 제작하지 않음
- * Scrapy framework (Spider)를 사용하여 크롤링을 진행, item.py에 데이터를 적재하고, 적재 완료 후 pipelines.py가 실행되며 MongoDB로 데이터 전송

4) 웰스크립트 (데일리호텔)



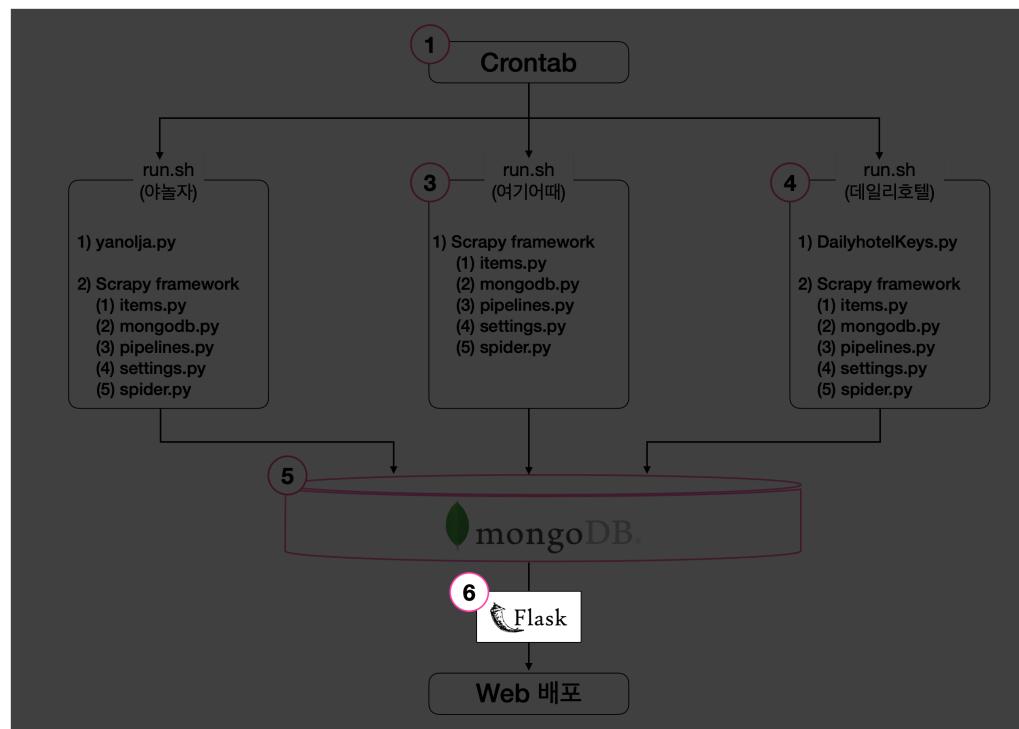
- * 데일리호텔 사이트에서 호텔별 상세정보를 크롤링하여 취합한 후, MongoDB에 데이터를 적재하도록 제작됨
- * dailyhotel_keys.py 모듈은 호텔별 고유ID값을 가져오는 역할 수행, 호텔별 고유ID값은 각 호텔별 상세페이지 url 생성 시 필요
- * Scrapy framework (Spider)를 사용하여 크롤링을 진행, item.py에 데이터를 적재하고, 적재 완료 후 pipelines.py가 실행되며 MongoDB로 데이터 전송

5) MongoDB



- * 실시간에 가깝게 데이터를 적재한다는 특성상 빠른 업로드가 중요
→ NoSQL 계열의 MongoDB를 사용
- * ‘date’ 열은 ‘년월일 - 시분’까지 반영
→ Flask를 통한 웹 배포 시, 최신 데이터 필터링에 활용하기 위함
- * 1사이클 당 데이터 7,000 ~ 8,000행 수준으로 업데이트
→ 1시간 정도의 분량인 100,000행으로 Collection 크기 설정
(100,000행이 초과되면 가장 오래된 데이터부터 순차적 삭제)

6) Flask



- * MongoDB에서 최신 1사이클의 데이터셋만을 전송받아 웹으로 배포하며, 구분 기준은 'date' 열을 사용
- * 데이터셋을 받을 때 'Price' 열 값이 0인 행들은 제외됨
- * 지역명이나 호텔명을 입력하고 'search' 버튼을 클릭
→ 조건에 맞는 호텔 리스트 출력
→ 'link' 열값 클릭하여 예약 페이지로 이동



프로세스 예시 화면

프로세스 예시 화면



플랫폼별 / 호텔별 상세페이지

나인트리 프리미어 호텔 인사동

★ 4.8/5 후기 1,491개 > 숙소 답변 (1,000+)

₩ 10,000원 할인 최대 할인은 초기화됩니다. 최대 할인 가능

객실 선택하기

나인트리 프리미어 호텔 인사동

69,000원-

신작은 구운 최대 1만원 할인

트루리뷰 ⚡ 95% (538명 평가)

MongoDB에 데이터 적재

db.getCollection('HotelInfo').find({})										
#	Date	Platform	Name	Level	Score	Review	Location	RoomType	Price	Link
0	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	MY OFFICE,숙박불가, 침대 없는 사무실형 객실/오픈 이용권	45000	Go
1	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	MY OFFICE,숙박불가, 침대 없는 사무실형 객실/오픈 이용권	55000	Go
2	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	내암대로 12시간STAY,숙박불가, 8-24시, 체크인시 갤럭시 배정	60500	Go
3	2021/03/19 11:46	여기어때	신라스테이 서대문	3성급	9.1	785	서울 서대문구 미근동 163	[반나절 호캉스-숙박불가] 체크인시 배정 (내...)	60500	Go
4	2021/03/19 11:50	데일리호텔	신라스테이 서대문구	1급	4.6	3651	서울특별시 서대문구 충정로 76	[숙박불가] 내 밤대로 12시간STAY (8-24 시) 체크인 시 배정	62500	Go
5	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	내암대로12시간STAY+스마트 스트리밍,숙박불가, 8-24시, 체크인시배정	66550	Go
6	2021/03/19 11:46	여기어때	신라스테이 서대문	3성급	9.1	785	서울 서대문구 미근동 163	[반나절 호캉스-숙박불가&스마트 스트리밍]	66550	Go
7	2021/03/19 11:50	데일리호텔	신라스테이 서대문구	1급	4.6	3651	서울특별시 서대문구 충정로 76	[숙박불가] 내 밤대로 12시간STAY (8-24 시) + 스마트 스트리밍 PKG	68800	Go

Flask 통한 웹 배포

신라스테이 서대문										
#	Date	Platform	Name	Level	Score	Review	Location	RoomType	Price	Link
0	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	MY OFFICE,숙박불가, 침대 없는 사무실형 객실/오픈 이용권	45000	Go
1	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	MY OFFICE,숙박불가, 침대 없는 사무실형 객실/오픈 이용권	55000	Go
2	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	내암대로 12시간STAY,숙박불가, 8-24시, 체크인시 갤럭시 배정	60500	Go
3	2021/03/19 11:46	여기어때	신라스테이 서대문	3성급	9.1	785	서울 서대문구 미근동 163	[반나절 호캉스-숙박불가] 체크인시 배정 (내...)	60500	Go
4	2021/03/19 11:50	데일리호텔	신라스테이 서대문구	1급	4.6	3651	서울특별시 서대문구 충정로 76	[숙박불가] 내 밤대로 12시간STAY (8-24 시) 체크인 시 배정	62500	Go
5	2021/03/19 11:50	야놀자	신라스테이 서대문구	3성급	4.5	2127	서울특별시 서대문구 충정로 76	내암대로12시간STAY+스마트 스트리밍,숙박불가, 8-24시, 체크인시배정	66550	Go
6	2021/03/19 11:46	여기어때	신라스테이 서대문	3성급	9.1	785	서울 서대문구 미근동 163	[반나절 호캉스-숙박불가&스마트 스트리밍]	66550	Go
7	2021/03/19 11:50	데일리호텔	신라스테이 서대문구	1급	4.6	3651	서울특별시 서대문구 충정로 76	[숙박불가] 내 밤대로 12시간STAY (8-24 시) + 스마트 스트리밍 PKG	68800	Go



웹 서비스
실습 :)



추후 개선 방향

1. 날짜 확장



AS - IS

* 현재 서비스는 당일예약 기준으로 하루치 데이터만을 웹을 통해 배포 중

Issue

1) 크롤링 툴의 변경 필요성

: 야놀자의 경우, 날짜 정보가 쿠키에 담겨있어 Scrapy나 BeautifulSoup을 통한 크롤링 불가능하다고 판단
→ Selenium의 사용 필수

2) 서버 메모리의 한계

: 데이터 규모 상 서버 1대로는 비교적 속도가 느린 Selenium을 사용하면서 실시간성을 충족시키기 어려워 추가 서버 확보 필요

TO - BE

* 서버를 추가 확보하여 Selenium으로도 충분히 실시간성을 충족시키도록 하여, 여러 날짜의 숙박 정보를 추가 크롤링

2. 웹페이지 기능/디자인 개선



AS - IS

* 현재 서비스는 형식이 단순한 편이며, 지역 혹은 호텔명만을 입력받아 그에 해당하는 객실 정보를 반환함

TO - BE

* HTML과 CSS에 대한 추가 학습 후, 필터 추가 및 반응형 홈페이지 / 동적페이지로 디벨롭



Q & A



E.O.D