

GitHub Repository 추천 서비스

#OpenSource #Platform #RecSys #ChromeExtension

Rec-04
(마음에들조)



boostcamp aitech

목차

1. 프로젝트 소개
2. 시연 영상
3. 전체 서비스 아키텍쳐
4. 향후 계획
5. QnA

1. 프로젝트 소개

팀 소개



박기범
Data Engineer



조예진
FrontEnd
Developer



정인식
Modeler(CF)



조영하
Backend
Developer



최필규
Modeler
(인기도, 유사도)

프로젝트 소개

- GitHub Repository 추천 서비스
 - 크롬 익스텐션으로 구현
 - 인기도 기반, 사용자 선호도 기반, 유사도 기반 추천 리스트 제공



문제 정의

- Github Repository 추천이 필요한 이유
 - 현재 github엔 두 가지 repository 탐색 방법 존재
 1. 이름과 태그를 통한 직접 검색
 - 사용자가 자신의 관심사를 확실히 알아야 함 (Repository 이름까지)
 - 일일이 검색해야 하는 불편함
 2. 사용자 로그 데이터를 기반으로 한 'Explore' 추천 (메인페이지, Explore 탭)
 - 동적으로 갱신되지 않음
 - 따로 메인페이지나 Explore 탭을 방문해서 확인해야 함
 - 이와 같은 불편함을 개선할 새로운 서비스 필요

Here's what we found based on your interests...

Based on repositories you've starred

Beomi / KcBERT

[Code](#) [Issues](#) [Pull requests](#) [Discussions](#)

Pretrained BERT model & WordPiece tokenizer trained on Korean Comments 한국어 댓글로 프리트레이닝한 BERT 모델

korean-nlp bert-model

Updated on 24 Oct 2021

Based on repositories you've viewed

Kitura / HeliumLogger

[Code](#) [Issues](#) [Pull requests](#)

A lightweight logging framework for Swift

swift log logger logging

Updated on 2 Oct 2021 · Swift

Based on topics you've starred

prodigycheatmenu

There are 1 public repositories matching this topic

ProdigyAPI / ProdigyX

A cheat menu for Prodigy. The ultimate tool for Prodigy Hacking!

[See more matching repositories](#)

Trending repositories today

... / MarkovJunior ⭐ 2k

Probabilistic PL based on pattern matching and constraint propagation, 148 examples

Jo... / msdt-foll... ⭐ 871

Codebase to generate an msdt-follina payload

el... / RedditVid... ⭐ 1.1k

Create Reddit Videos with just '+' one command '+'

c... / New-Gra... ⭐ 1.8k

A collection of New Grad full time roles in SWE, Quant, and PM.

[See more trending repositories](#)

Trending developers

Mark Tyneway

tynes

Iman khoshabi

imaNNNeoFight

[fl_chart](#)

Joe Chen

unknwon

[the-way-to-go_ZH_CN](#)

smartcontracts

smartcontracts

[simple-optimism-node](#)

[See more trending developers](#)

개발 목표

- 사용자에게 편의를 제공하는 서비스

- 원하는 Repository를 일일이 검색할 필요가 없는,
- 초보자도 편하게 사용할 수 있는,
- 추천 리스트를 다른 창에서 확인할 필요가 없는,
- 한 번 깔아두면 신경쓰지 않아도 되는,

→ Chrome Extension 구현



2. 시연 영상

3. 전체 서비스 아키텍쳐

데이터셋 선정 및 데이터베이스 설계

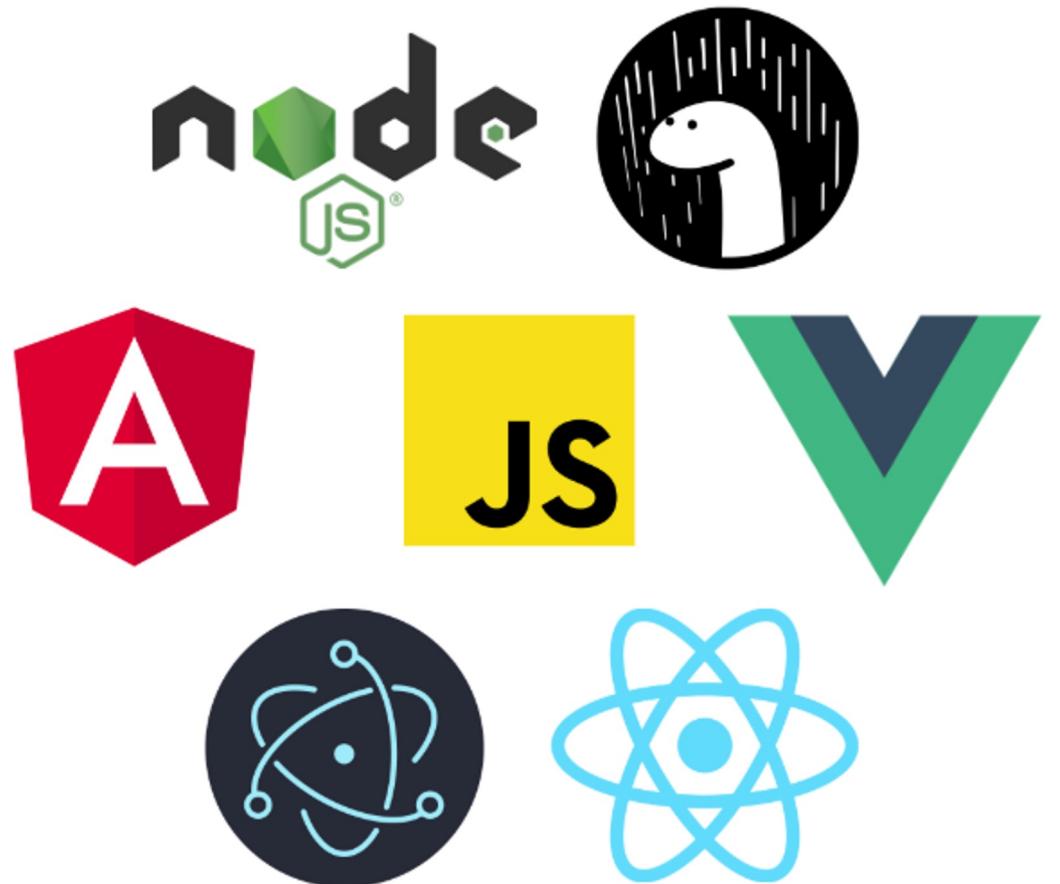
- Awesome Series - 학습 데이터

The screenshot shows the GitHub repository page for `sindresorhus/awesome`. The repository is public and has 21 issues, 24 pull requests, and 23.5k forks. The main branch is selected. The repository contains several files and folders, including `.github`, `media`, `.editorconfig`, `.gitattributes`, `awesome.md`, `code-of-conduct.md`, `contributing.md`, `create-list.md`, `license`, `pull_request_template.md`, and `readme.md`. The `awesome.md` file lists various topics such as lists, awesome, unicorns, and resources. The repository has 203k stars and 7.4k watchers. The `About` section describes it as a list of interesting topics and includes links to the Readme, license, code of conduct, and sponsor project.

- 다양한 직군, 소프트웨어의 주요 토픽 소개
- 약 20만명 Star repository

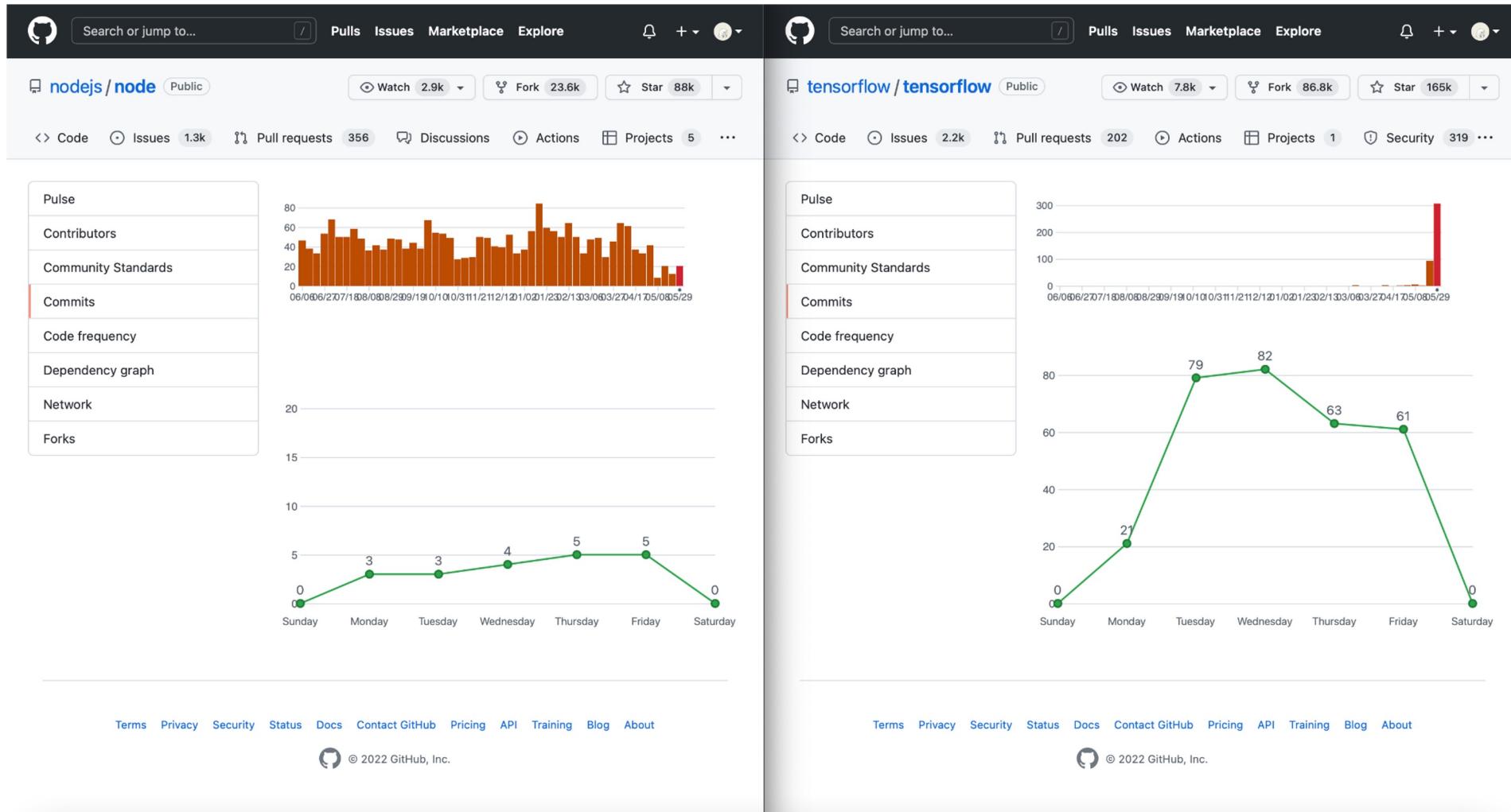
데이터셋 선정 및 데이터베이스 설계

- Awesome Series - 학습 데이터



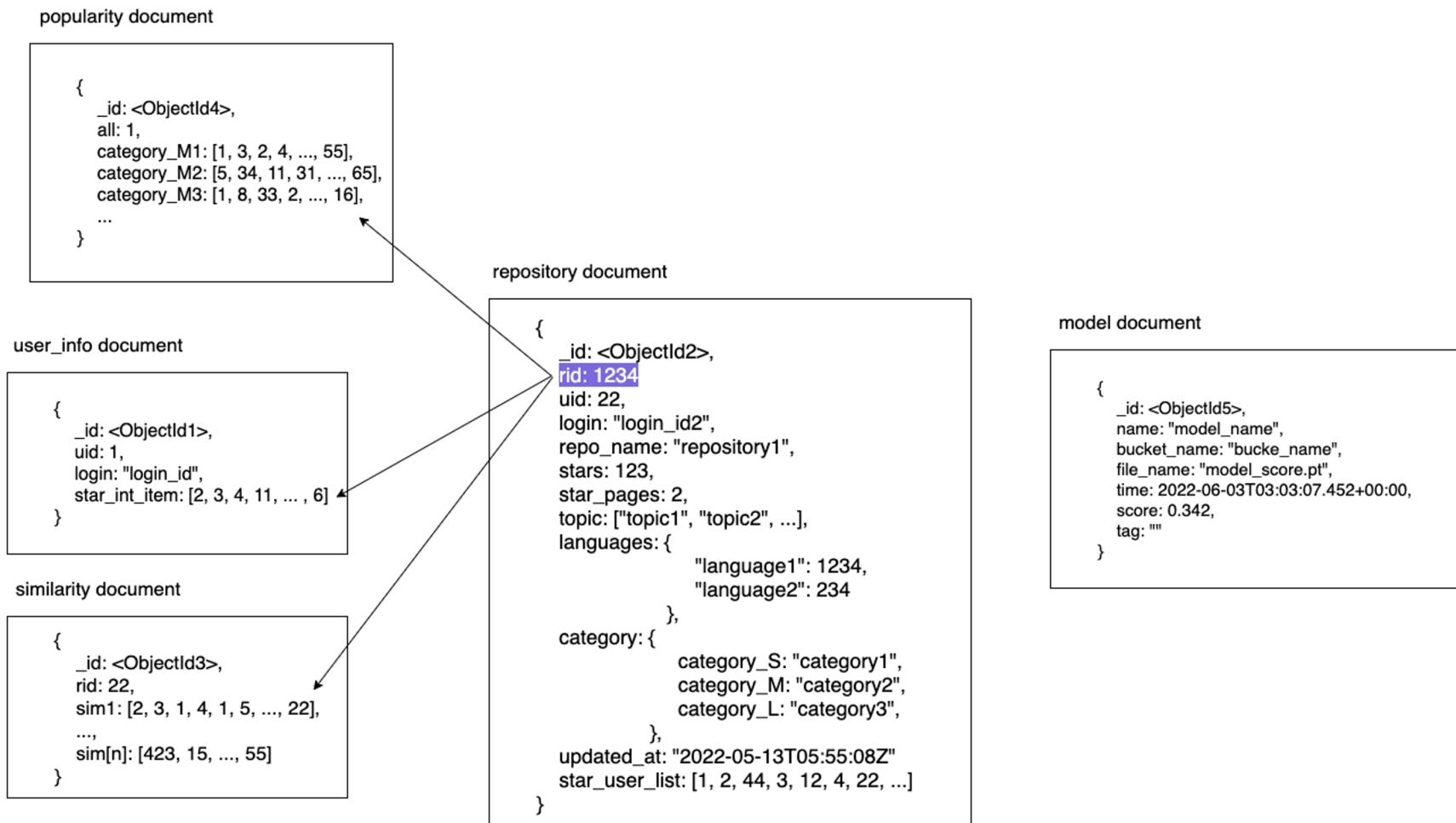
- 범용성이 넓고 사용자가 확보될 수 있는 데이터
- Github 활성화가 높은 Back-End, Front-End 관련기술
- 기술 간의 상호 연관성이 존재해야 실제 육안으로 검증이 가능
- JavaScript 기반의 Front / Back End 기술 6가지 선정

데이터셋 선정 및 데이터베이스 설계



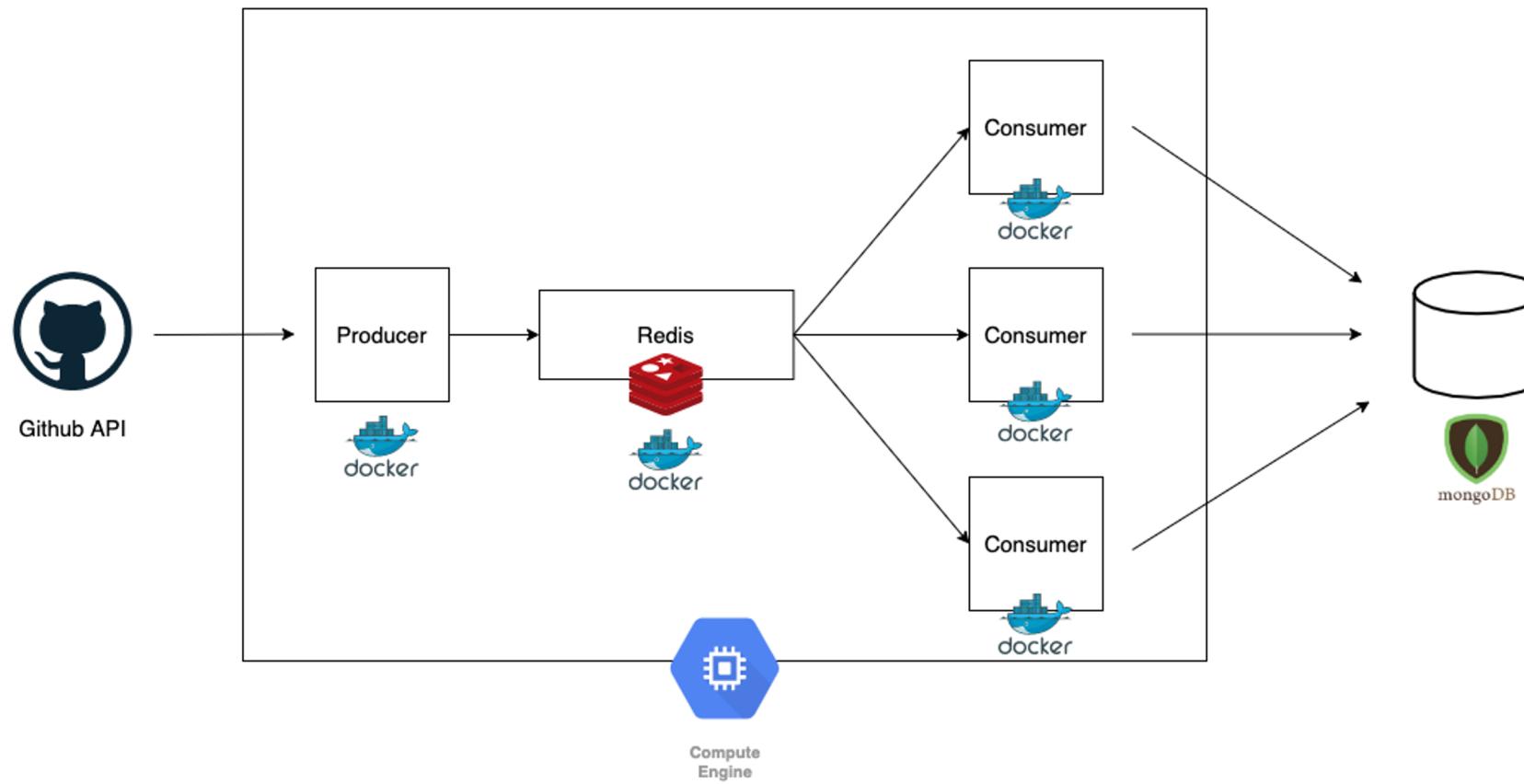
데이터셋 선정 및 데이터베이스 설계

- 데이터베이스 설계

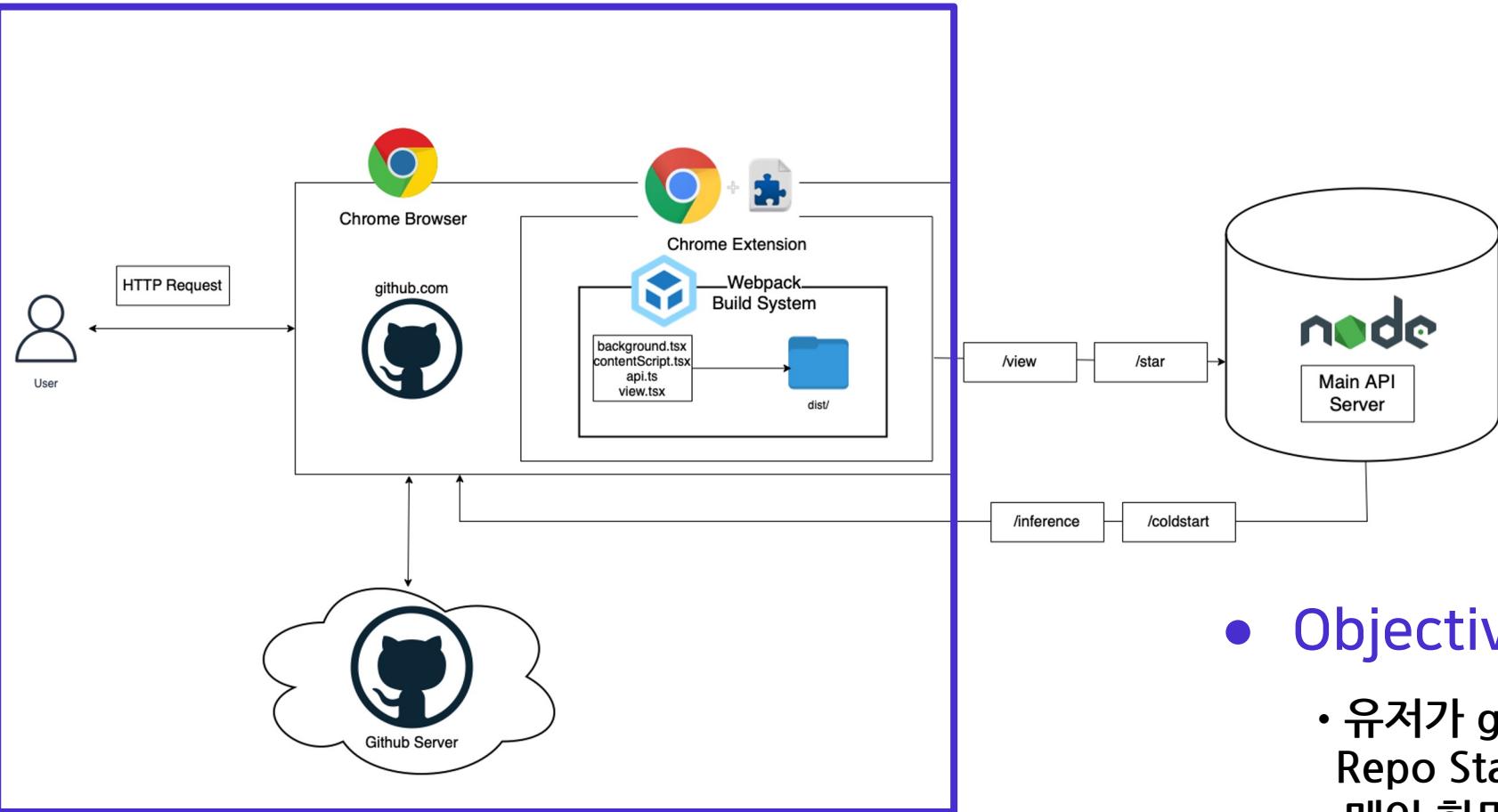


데이터 수집 방법

- Producer - Consumer



Chrome Extensions



• Objectives

- 유저가 github에서 하는 액션 중 Repo Visit, Repo Star를 server에 전송
- 메인 화면에서 coldstart repository 추천
- Repository 방문 시 비슷한 repository 추천

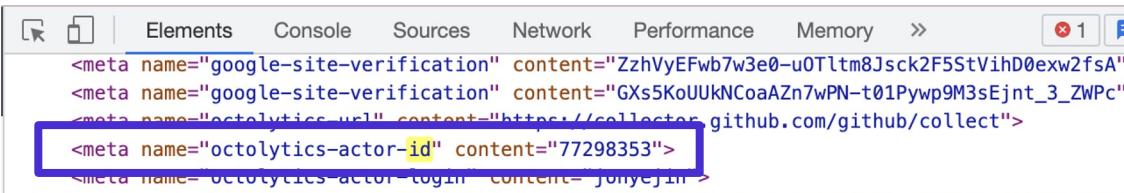
Chrome Extensions

- Chrome Extension → github.com 내부 HTTPS 리퀘스트
 - Repository 방문, 좋아요 등
 - Github Repository 방문 : URL로 확인
 - 좋아요 클릭: Button에 onClick액션 발생 시
- Extension을 처음 설치한 유저에 대해서는 init액션을 실행하고, Chrome 내 DB에 유저 정보를 저장
 - Coldstart: Coldstart Repository 선택하는 view
- 로그인 정보는 쿠키, Repo ID는 HTML스크립트에서 파싱
- 보안 이슈 주의!

* Login정보 (쿠키)

Application	Name	Value	D..	P..	E..	Size	H..	S..	S..	P..	P..	
Manifest	_gh_sess	DPuzVWyydtKL8z75h...	g...	/	S...	252	✓	✓	Lax	M...	M...	
Service Workers	dotcom_user	jonyejin	/	2...	19	✓	✓	Lax	M...	M...	
Storage	logged_in	yes	/	2...	12	✓	✓	Lax	M...	M...	
	user_session	WOQYLBGtXkJpslsQ...	g...	/	2...	60	✓	✓	Lax	M...	M...	
Local Storage	has_recent_a...	1	g...	/	2...	20	✓	✓	Lax	M...	M...	
Session Storage	__Host-user_...	WOQYLBGtXkJpslsQ...	g...	/	2...	77	✓	✓	S...	M...	M...	
IndexedDB	_locale	ko	/	2...	9	✓	Lax	M...	M...	M...	
Web SQL	_locale_expe...	ko	/	2...	20	✓	Lax	M...	M...	M...	
Cookies	tz	Asia%2FSeoul	/	S...	14	✓	Lax	M...	M...	M...	
https://github.com	_device_id	b6a925383785d4793b...	g...	/	2...	42	✓	✓	Lax	M...	M...	
	Trust Tokens	color_mode	%7B%22color_mode...	/	S...	215	✓	✓	Lax	M...	M...
	Interest Groups	_octo	GH1.1.1133967668.16...	/	2...	32	✓	Lax	M...	M...	
Cache												
		Cache Storage										

* RepoID 찾기 (HTML태그)



```
<meta name="google-site-verification" content="ZzhVyEFwb7w3e0-u0Tltm8Jsck2F5StVihD0exw2fsA">
<meta name="google-site-verification" content="GXs5KoUkNCoaAZn7wPN-t01Pywp9M3sEjnt_3_ZWPc">
<meta name="octolytics-url" content="https://collector.github.com/github/collect">
<meta name="octolytics-actor-id" content="77298353">
<meta name="octolytics-actor-login" content="jonyejin">
```

* Repository 방문 정보 (URL)

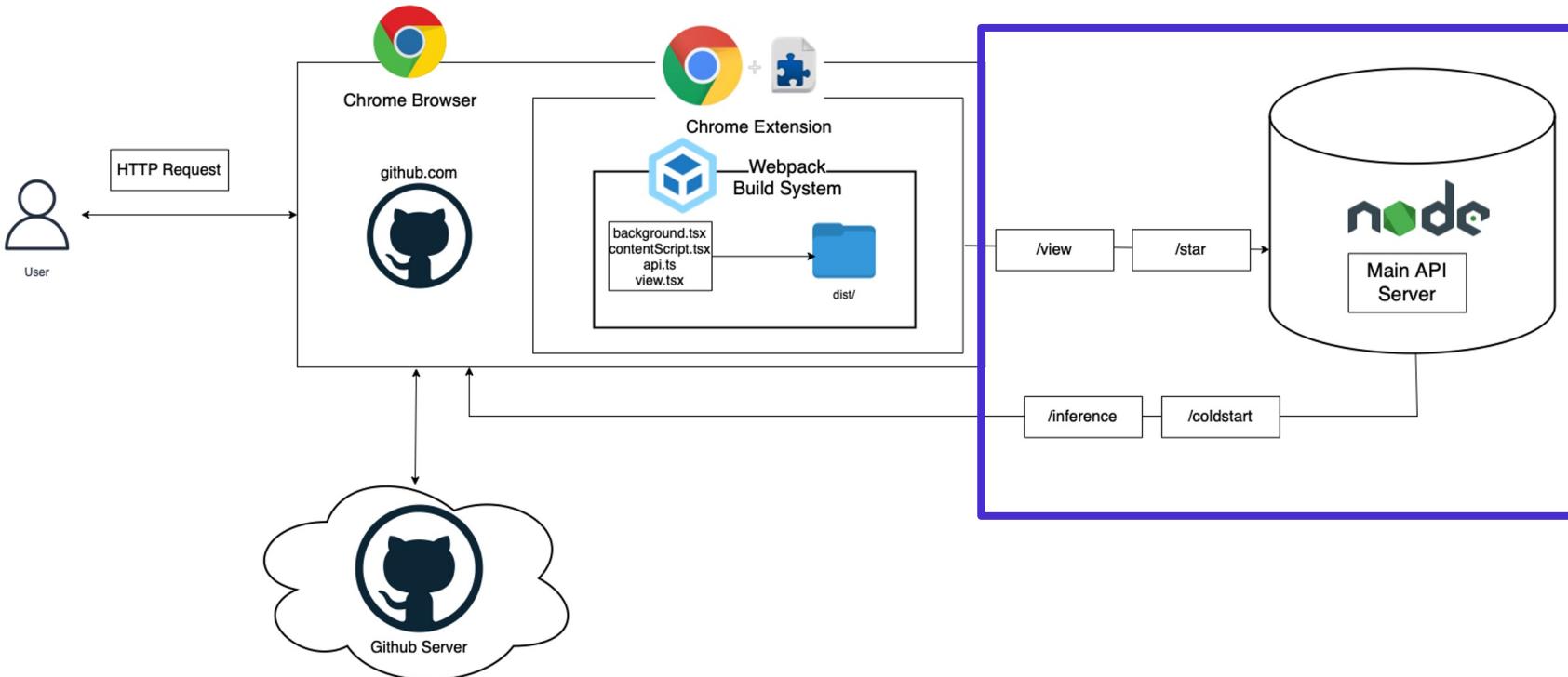
github.com/pyscript/pyscript

repoOwner

repoName

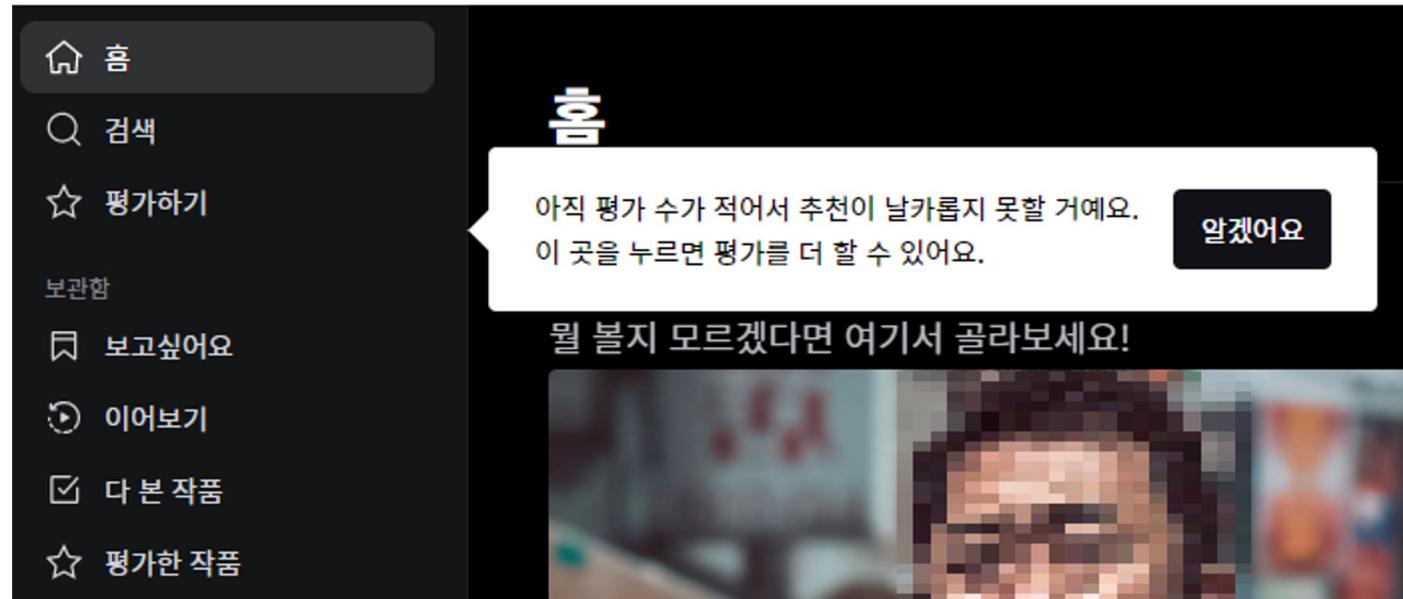
API 기능

• Objectives

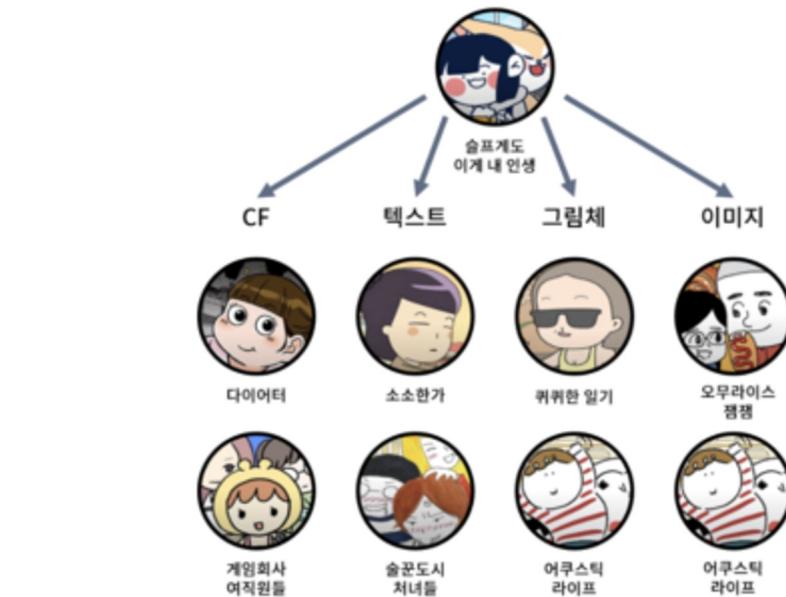


- 유저가 github에서 활동하는 동안 서버가 능동적으로 서비스 제공
 - Repository 방문할 때마다 관련 Repository 추천
- 유저의 기존 정보를 활용하는 동시에 click, star 활동을 추적해 DB 서버에 저장해야 함
- Repository 방문 케이스마다 달라지는 inference case를 처리해야 함

API 기능 요약



왓챠에서 새로 회원가입한 유저들에게 요구하는 '평가하기' 서비스



유저, 아이템 정보를 활용해 다양한 추천을 해주는 서비스

모델

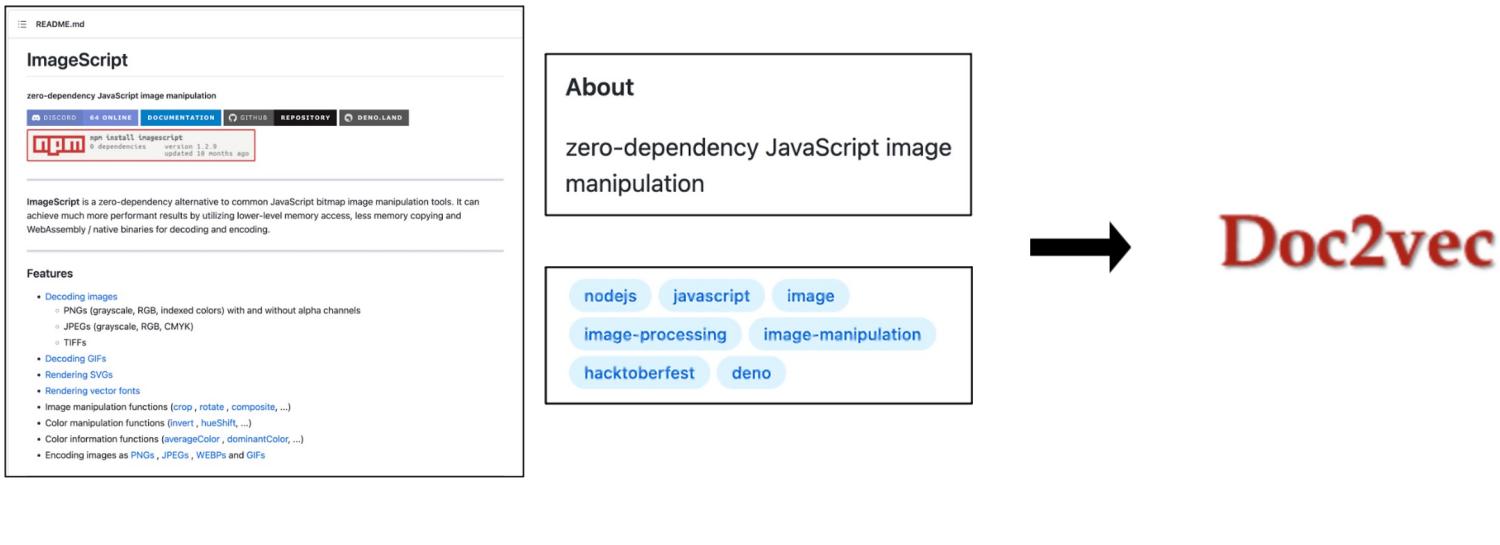
- 인기도 기반 추천
 - 카테고리별 Top N
 - 사용자가 현재 방문한 Repository가 속한 카테고리에서 인기도가 높은 상위 N개의 리포지토리들을 추천리스트에 포함(인기도 지표로 'Star 개수' 사용)
 - 이 때, 단순히 만들어진지 오래된 Repository가 star가 많을 수 있으니 '생성 경과시간' Term을 인기도 계산에 추가
 - 업데이트된 지 오래된 Repository에도 패널티를 주기 위해 '업데이트 경과시간' Term을 인기도 계산에 추가
 - 인기도 =
$$\frac{\text{Star 개수}}{(\text{생성 경과시간} + \text{업데이트 경과시간})^{\text{gravity}}}$$
 - 시간 Term의 영향을 조정하기 위해 분모에 gravity라는 상수를 사용

모델

- 유사도 기반 추천
 - 유저가 현재 방문한 Repository와 유사도가 높은 Repository를 다른 추천리스트와 중복되지 않도록 Top N개 추천
 - 유사도 계산을 위한 Repository 임베딩 방식은 두 가지
 - 문서 임베딩
 - 그래프 임베딩

모델

- 유사도 기반 추천
 - 문서 임베딩
 - 사용 데이터: 각 Repository의 About, Tags, README.md를 합친 문자열
 - 사용 모델: Gensim 라이브러리의 Doc2Vec



모델

- 유사도 기반 추천
 - 그래프 임베딩
 - 사용 데이터: User-Star-Repository 그래프
 - 사용 모델: PyTorch-BigGraph(PBG)

USER ID		REPO ID
86240	star	322936801
613956	star	322936801
13407106	star	322936801
11403655	star	322936801
1903667	star	322936801
69742611	star	322936801



 PyTorch BigGraph

The PyTorch BigGraph logo consists of three orange circles connected by lines, forming a triangular graph structure.

...

모델

- 유사도 기반 추천
 - 모델 학습된 임베딩값으로 Repository들 사이의 Angular 유사도를 계산
 - Angular 유사도는 두 벡터 간의 각도가 작으면 매우 비슷한 유사도를 가진다는 기준 Cosine 유사도의 단점을 개선
 - 빠른 계산 속도를 위해 Nearest Neighbor 알고리즘인 Annoy를 활용하여 Top N개 추출
 - Repository별로 유사한 Top N개 Repository 리스트를 DB에 저장해두고 사용

모델

- CF(Collaborative Filtering) 기반 추천
 - 데이터 및 모델
 - Data: User의 repository star list를 implicit data로 사용
 - Model: RecVAE, implicit data에 강하고 추론 속도가 빠름
 - 개인화된 실시간 추천
 - user의 star 및 조회 결과를 실시간 반영하여 추천목록 제공
 - 웹페이지와 로딩과 비슷한 속도 이내에 결과 제공
 - 사용자에게 지속적인 서비스 이용의 필요성을 제공

모델

- CF 모델 업데이트 자동화

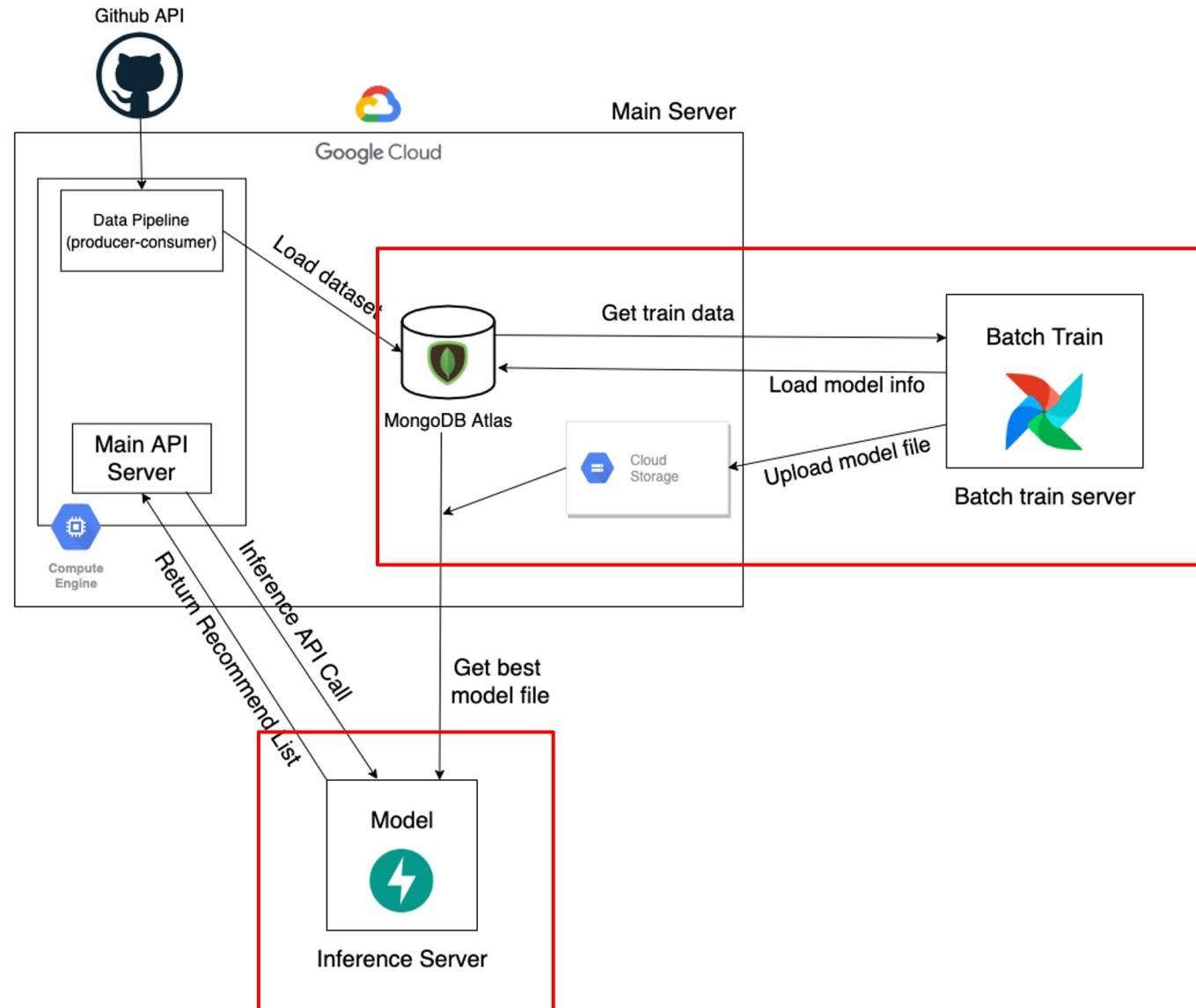
- 학습 서버

- Airflow로 구축하여 task의 안정성을 갖춤
 - 서비스를 통해 지속적으로 user data(방문 기록 및 star)를 수집
 - 훈련 완료시 model parameter를 model.pt에 저장하여 DB에 업로드

- 추론 서버

- fast-api를 end-point로 사용하여 추론 결과를 메인 서버에 제공
 - 정해진 시간에 DB에서 model parameter를 다운로드 하여 모델 업데이트

모델



모델

- 모델 특징 요약
 - 다양성
 - 인기도, 유사도, CF 총 3개의 방식의 추론 결과를 제공
 - 지속성
 - user의 사용기록을 기록
 - 서비스 제공과 동시에 훈련 데이터를 지속적으로 수집가능한 구조
 - 실시간
 - 사용자와 서비스의 상호작용을 사용자에게 인지시켜 서비스 사용을 독려

4. 향후 계획

A/B Testing

- Online Testing 기반의 성능 평가 필요
 - Offline 기반의 성능 테스트 (RMSE, Recall@10 등)만으로는 사용자들이 서비스를 얼마나 선호하는지 판단하기 어려움
 - 임의의 유저들에게는 무작위의 추천을 해주는 API를 설계해 A/B testing 진행할 예정
 - 나아가 위의 그룹과 원래 서비스를 받는 그룹 간의 click rate 차이를 분석해 online test 기반의 성능 평가를 진행할 예정

5. Q&A