

AI기법과 활용

Week-03. Search Engine Basic

2022-Summer 서중원



🔍 구글은 어떻게 만들까?



Google 검색

I'm Feeling Lucky

Google 제공 서비스: [English](#)

검색엔진

검색엔진에서의 고려사항

- Scalability
 - 콘텐츠적인 측면: 얼마나 많은 콘텐츠를 수집/저장/처리할 수 있는가?
 - 사용자적인 측면: 얼마나 많은 사용자에게 대응 할 수 있는가?
- High Quality Results
 - 관련된 콘텐츠인가?
 - 스팸인가?
- Dynamics
 - 하루에 생성되는 웹사이트의 수: *547,200
 - 추가적으로 기존의 웹사이트의 콘텐츠도 업데이트

* <https://siteefy.com/how-many-websites-are-there/>

검색엔진의 구성요소

1초 안에 양질의 검색결과가 나오기 위해서는?

- Crawling
 - Focused Crawling: 우선순위를 정해 크롤링
 - Deep Crawling: 페이지 안에.. 링크 안에.. 페이지 안에..
- Indexing
 - 웹 페이지를 등록하는 작업
 - 분산처리, Map Reduce?
- Ranking
 - 콘텐츠간의 순위를 매기는 작업
 - 어떤 결과가 더 좋은 결과인가?

검색엔진의 구성요소

크롤링과 스크래핑

- Web Scraping
 - 데이터를 추출(extracting)하는 행위
- Web Crawling
 - 반복적으로 링크를 찾고 데이터를 저장하는 행위
 - 링크를 찾는 과정 또는 데이터를 추출 하는 과정에서 웹 스크래핑 과정이 포함됨

검색엔진의 구성요소

Web Crawling

- 웹 페이지를 자동으로 찾고 다운받는 작업
- 웹이라는 것은 거대하고 지속적으로 성장
- 웹이라는 것은 검색엔진 제공자의 통제하에 있지 않음
- 웹 페이지는 지속적으로 변화함
- 크롤러는 다른 타입의 데이터를 활용하기도 해야함

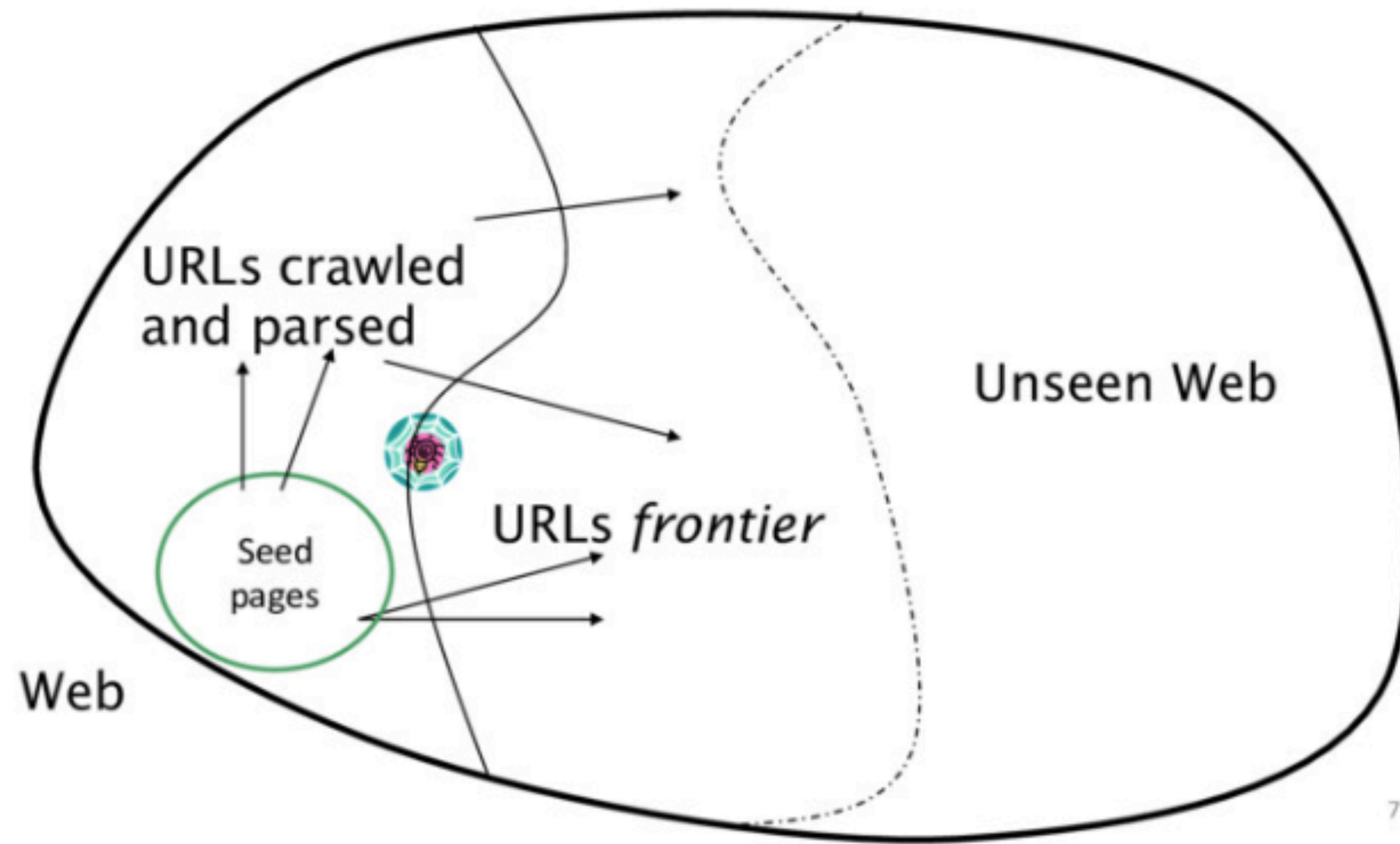
검색엔진의 구성요소

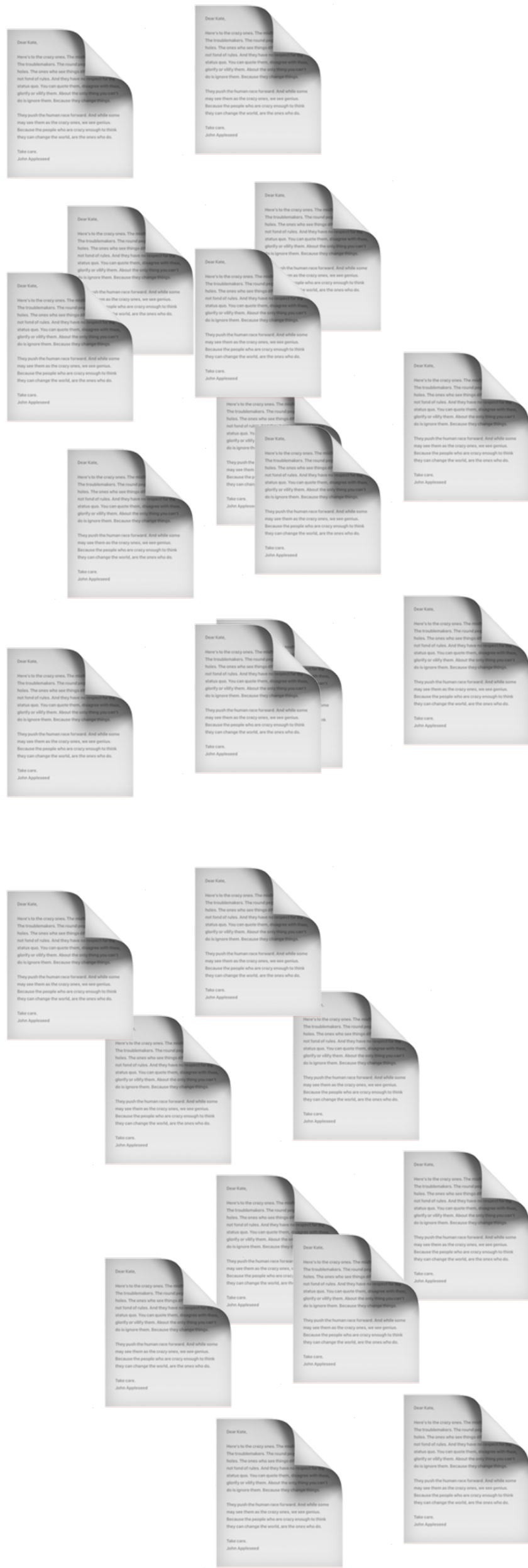
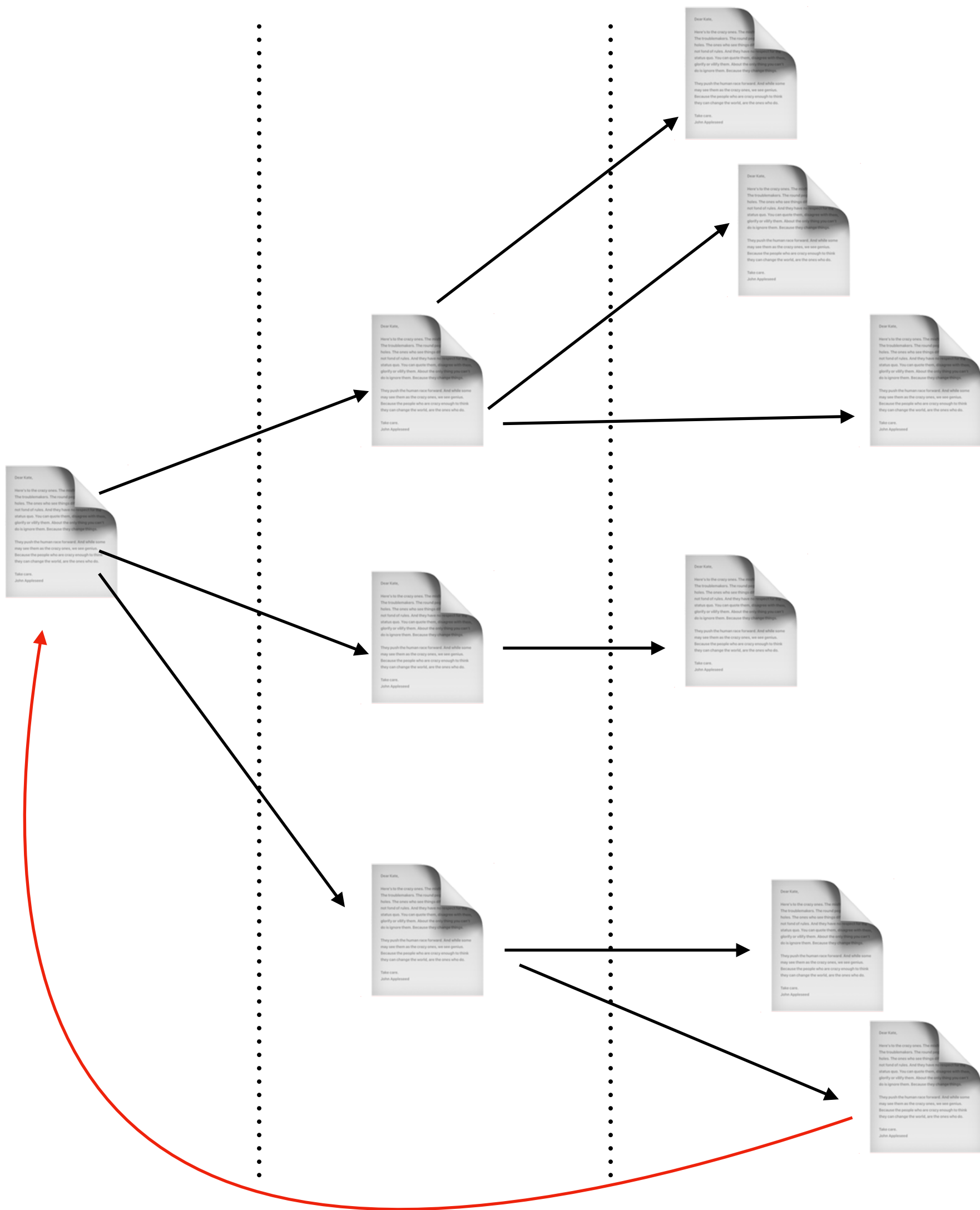
Web Crawling

- Seeds라고도 불리우는 몇몇 페이지를 기준으로 시작
- Seeds는 URL 요청 대기줄에 추가됨
- 크롤러는 페이지들을 URL 요청 대기줄에서 하나씩 빼와서 읽기 시작함
- 다운받아진 페이지는 페이지 내의 링크를 추출하기 위해 파싱
- 추출된 링크들을 URL 요청 대기줄에 추가
- 더 이상 새로운 URL이 없거나, 디스크의 용량이 가득 찼을 때까지 실행

검색엔진의 구성요소

Web Crawling

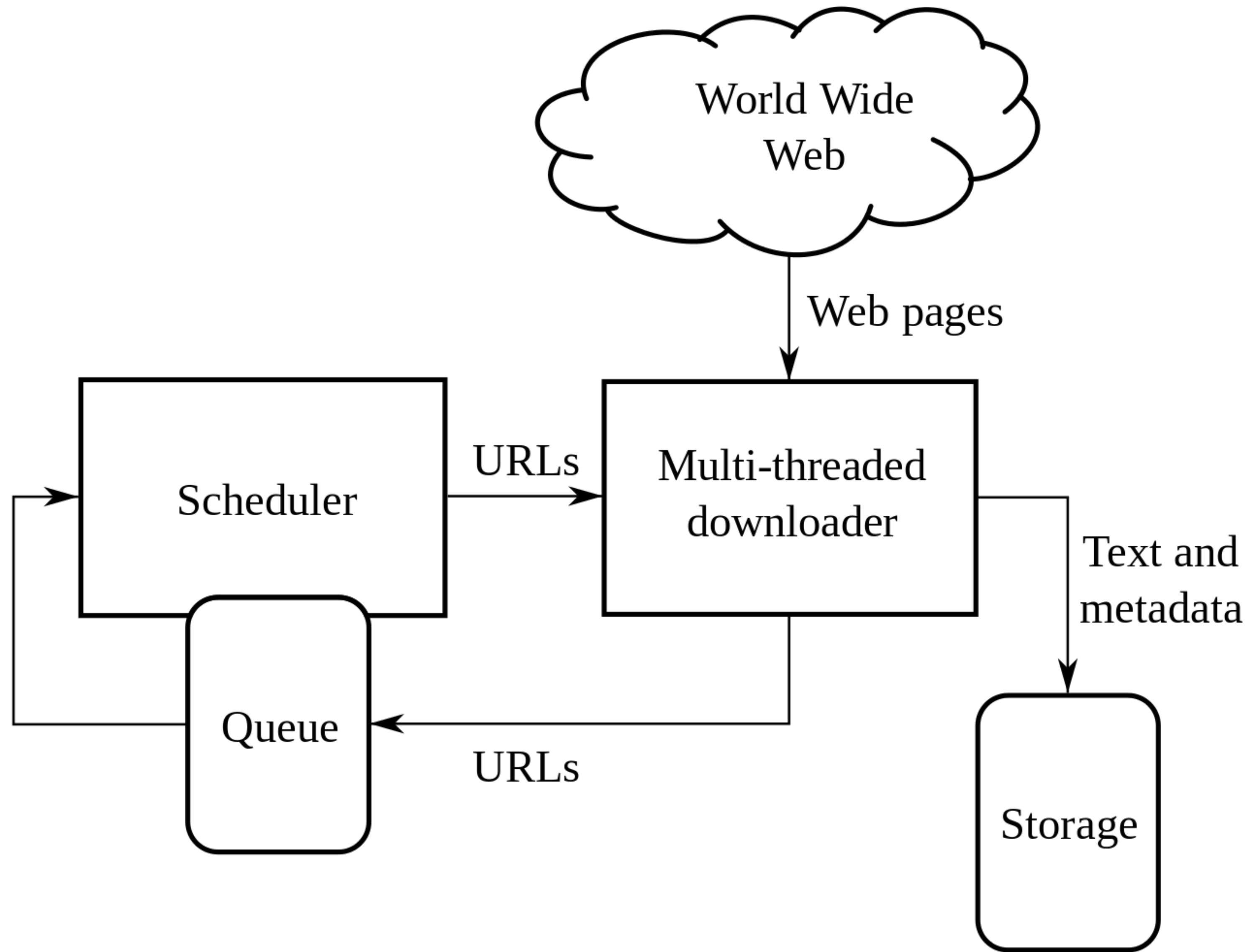




검색엔진의 구성요소

Web Crawling

- 요청을 주고 받는 과정은 많은 시간을 필요로 함
 - 한 페이지당 1초 정도 소요되지만, 검색엔진이 되기 위해서는 엄청난 양의 웹페이지를 인덱싱 하고 있어야 함
- 효율적으로 크롤링을 진행하기 위해, 멀티 쓰레딩 방식으로 동시에 요청을 보냄
- 이러한 방식은 특정 웹사이트를 마비 시킬 위험이 있음
- 피해를 주지 않기 위해, 동일한 웹사이트에 대한 요청은 인위적인 지연시간을 삽입



크롤링

링크 분석

- 링크는 웹에서 가장 중요한 성분 중 하나
- 이동을 위해서도, 검색을 위해서도 중요

```
<a href="https://yonsei.ac.kr">Yonsei University</a>
```

목적지의 주소

Anchor text

- Anchor Text와 링크 둘 다 Search Engine에서 사용됨

링크의 중요성

예제

Jungwon's Blog

CODETHIEF

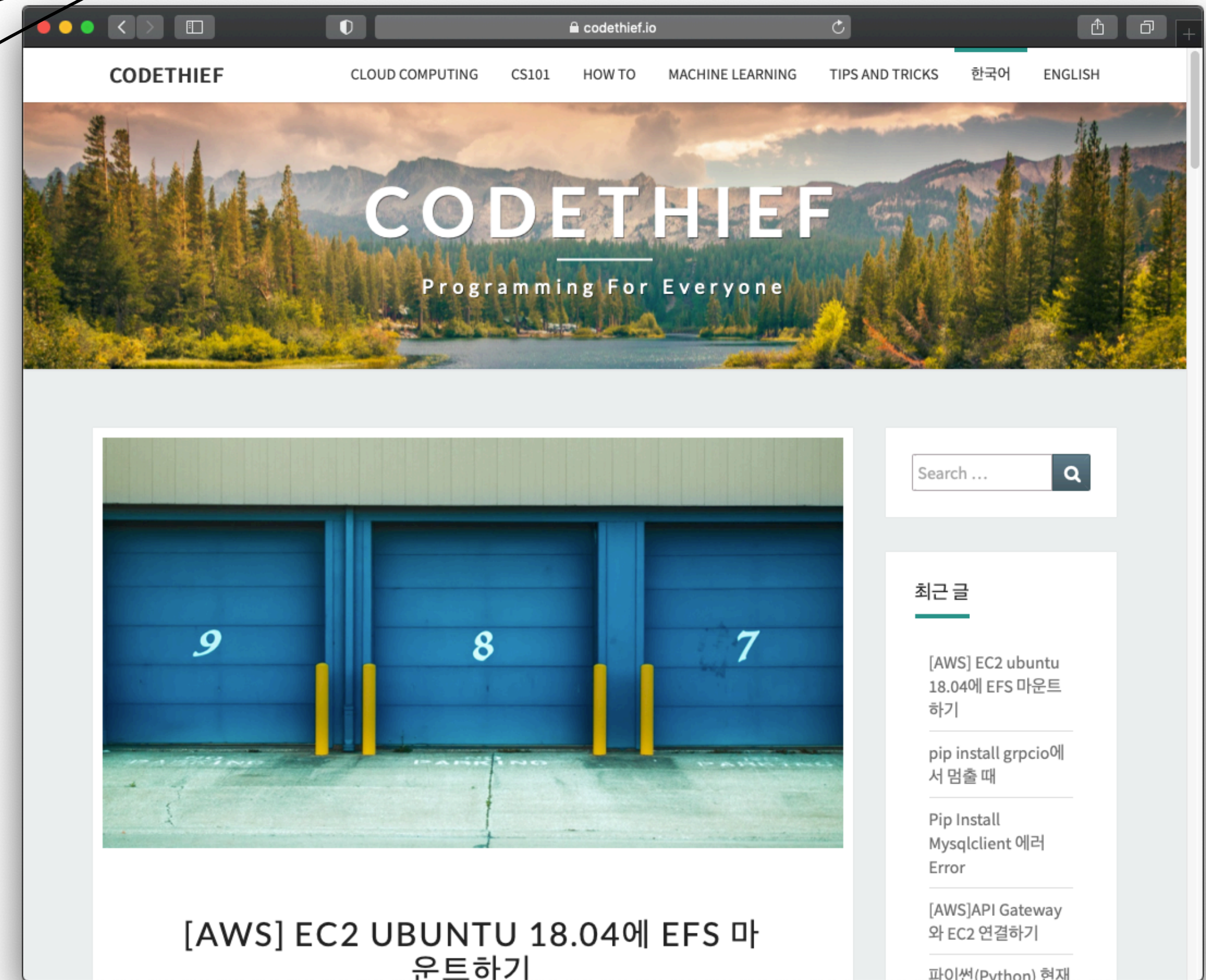
Coding Blog

Best Blog

```
<a href="https://codethief.io">Jungwon's Blog</a>
```

```
<a href="https://codethief.io">Coding Blog</a>
```

```
<a href="https://codethief.io">Best Blog</a>
```



Fielded Document Representation

문서를 저장할 때, 필드를 구별하여 저장 하는 방식

Title

- Yonsei Big Data 2022

Meta

- Yonsei, University, Bigdata, Machine Learning, 2022

Headings:

- Yonsei Big Data Course 2020 in Sinchon Campus

Body:

- Yonsei GSI opens new course called “Big Data Analytics Programming” ...

Anchors:

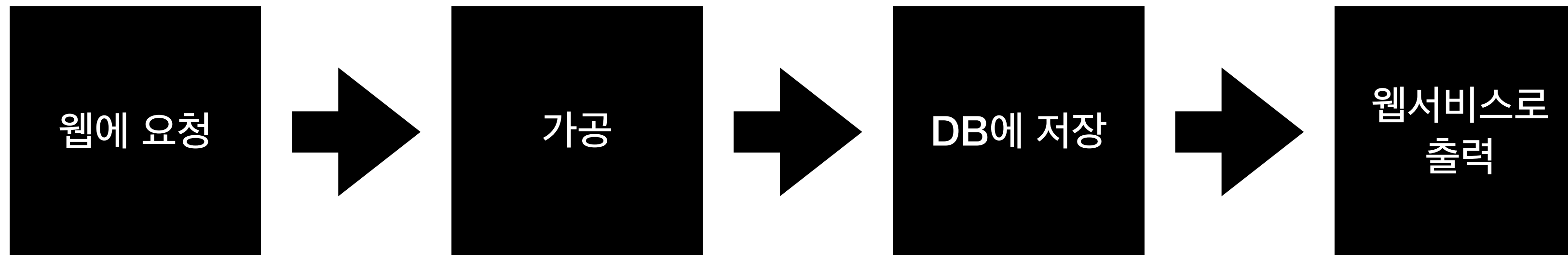
- Coding Blog
- Jungwon's Blog
- Best Blog

References

Krisztian Balog, DAT630, University of Stavanger, October 23, 2017,
<https://speakerdeck.com/kbalog/2017-web-search>

오늘의 실습!

데이터 확보-저장-출력



GCP 접속

클라우드 컴퓨팅 서비스 | Google

홈 - My First Project - Google

cloud.google.com/?authuser=2

Google Cloud

Google을 선택해야 하는 이유

솔루션

제품

가격 책정

시작하기

문의하기

문서

지원

한국어

콘솔

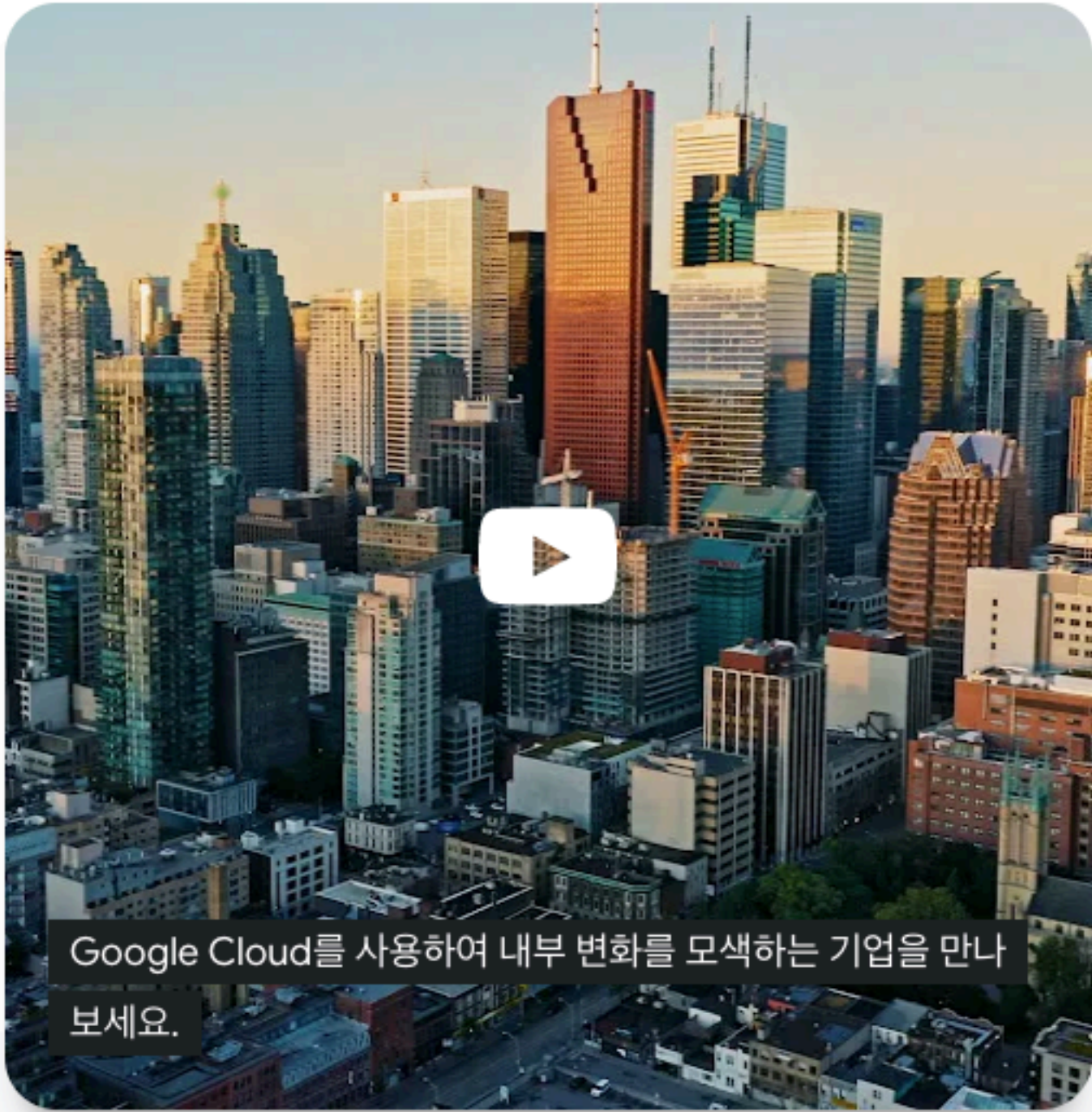
\$300의 무료 크레딧과 20여 개 제품에 대한 무료 사용량이 제공됩니다.

Google Cloud와 함께 디지털 혁신을 꿈꾸고 실현 하세요

앱을 더 빠르게 빌드하고 보다 현명한 비즈니스 의사결정을 내리며 세계
각지의 사람들과 소통할 수 있습니다.

Console로 이동

영업팀에 문의




Google Cloud를 사용하여 내부 변화를 모색하는 기업을 만나
보세요.

새로운 소식

개발자용


이벤트

5월 10일 Google Cloud Developer Summit




2분 퀴즈

기업 문화가 얼마나 데이터에 기반하느냐에 따라



보고서

글로벌 리더를 기준으로 귀사의 디지털 혁신을 평



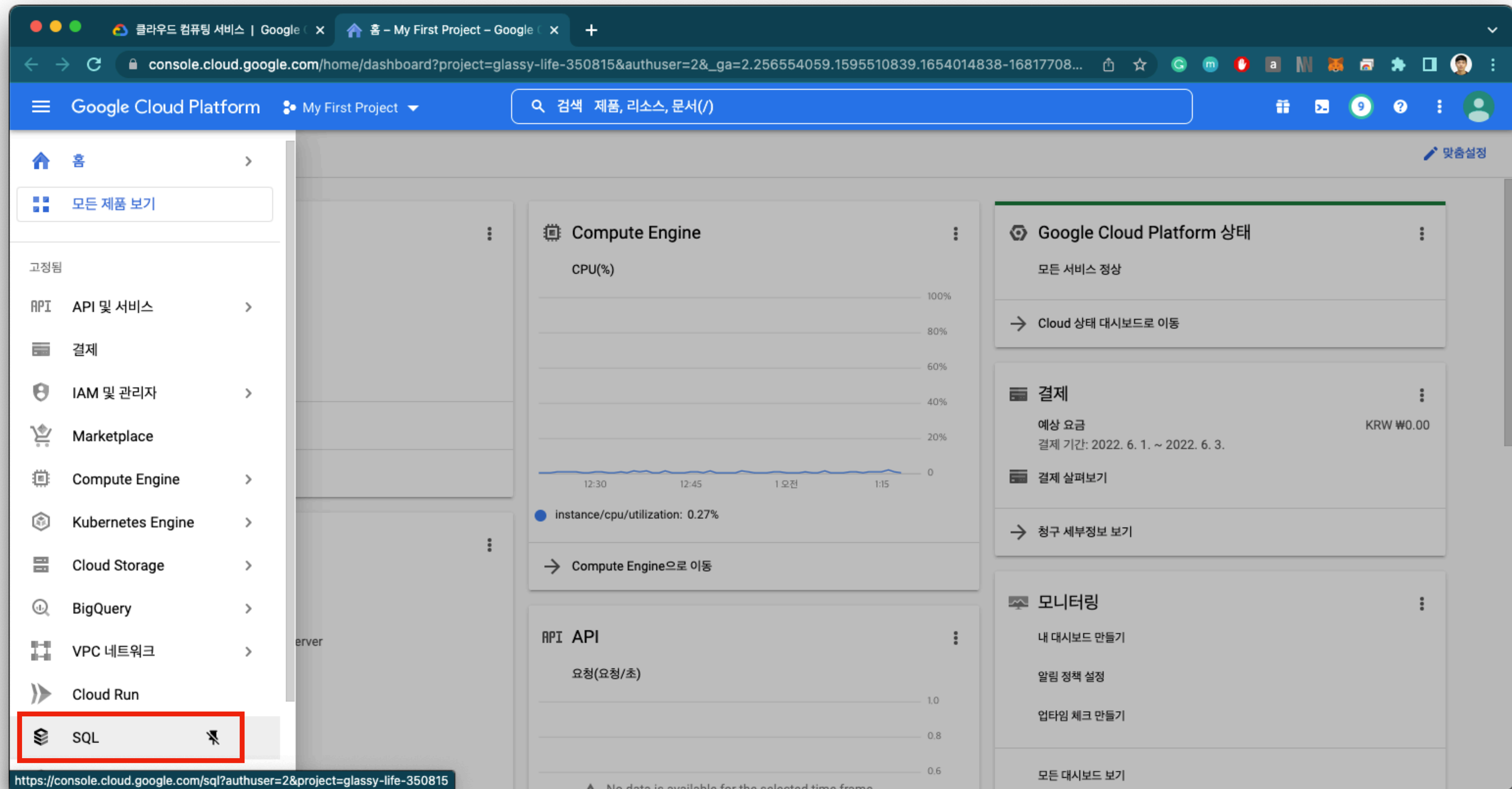
메뉴 클릭

The screenshot shows the Google Cloud Platform console interface. At the top, the browser address bar displays the URL `console.cloud.google.com/home/dashboard?project=glassy-life-350815&authuser=2&_ga=2.256554059.1595510839.1654014838-16817708...`. The main header bar is blue and contains the Google Cloud Platform logo, the project name 'My First Project', a search bar with the text '검색 제품, 리소스, 문서(/)', and user account icons. A red box highlights the hamburger menu icon (three horizontal lines) in the top left corner of the header bar.

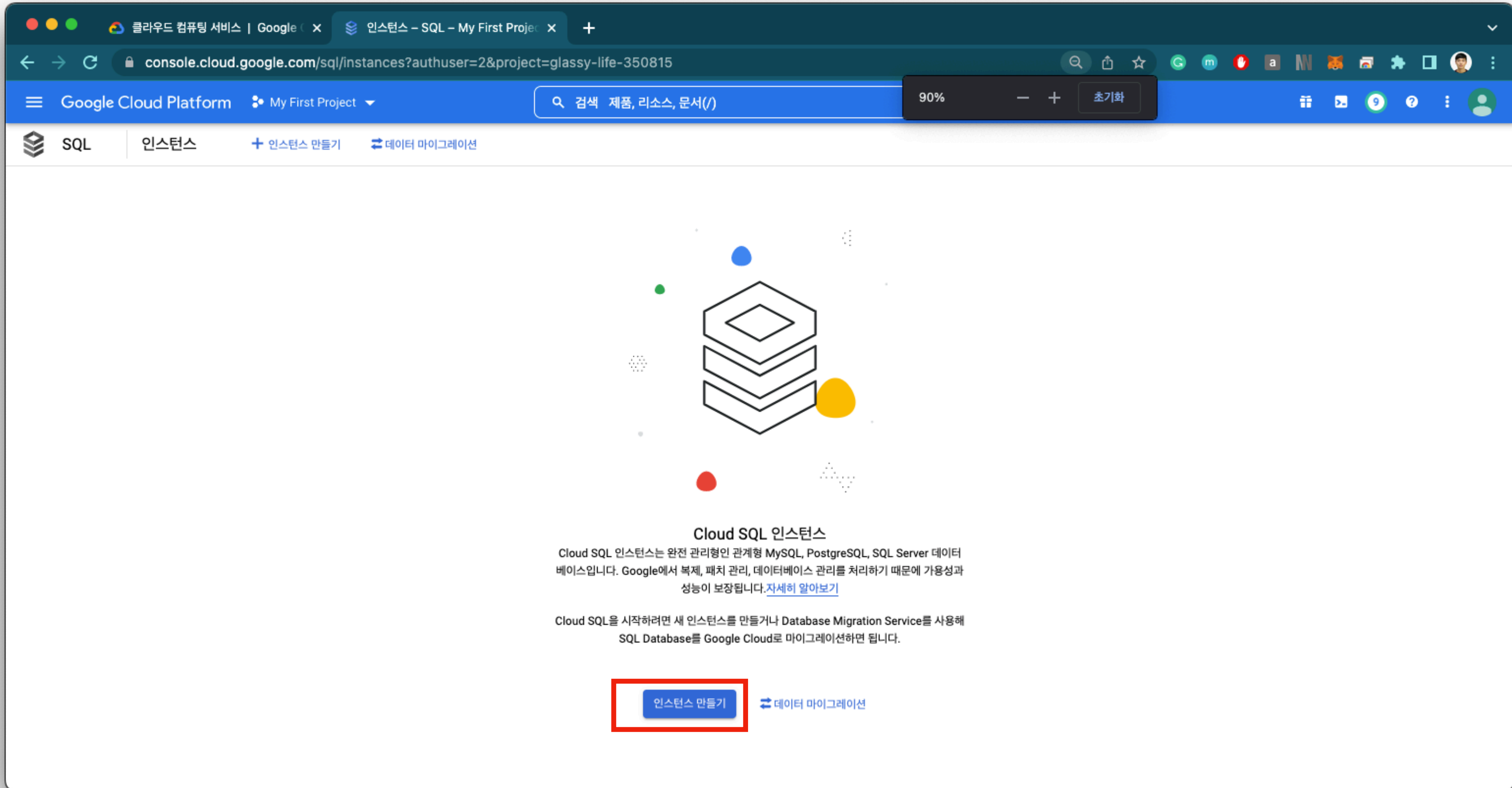
Below the header bar, the dashboard is organized into several sections:

- 대시보드 (Dashboard):** Includes tabs for '대시보드', '활동', and '권장사항', along with a '맞춤설정' (Customize) link.
- 프로젝트 정보 (Project Information):** Displays project details such as '프로젝트 이름' (My First Project), '프로젝트 번호' (768742373988), and '프로젝트 ID' (glassy-life-350815). It includes a link to '이 프로젝트에 사용자 추가' and a button to '프로젝트 설정으로 이동'.
- 리소스 (Resources):** Lists various Google Cloud services including BigQuery, SQL, Compute Engine, Storage, and Cloud Functions.
- Compute Engine:** Features a 'CPU(%)' graph showing utilization over time (12:30 to 1:15). The current utilization is 0.58%. A button 'Compute Engine으로 이동' is provided.
- API:** Shows a graph for '요청(요청/초)' (Requests per second) with a note: 'No data is available for the selected time frame'.
- Google Cloud Platform 상태 (Status):** Indicates '모든 서비스 정상' (All services healthy) and provides a link to 'Cloud 상태 대시보드로 이동'.
- 결제 (Billing):** Shows '예상 요금' (Estimated cost) as 'KRW ₩0.00' for the period '2022. 6. 1. ~ 2022. 6. 3.'. It includes a link to '청구 세부정보 보기'.
- 모니터링 (Monitoring):** Offers options to '내 대시보드 만들기' (Create my dashboard), '알림 정책 설정' (Set up alerts), '업타임 체크 만들기' (Create uptime checks), and '모든 대시보드 보기' (View all dashboards).

SQL 클릭



DB인스턴스 만들기



The screenshot shows the Google Cloud Platform console interface. The browser address bar displays the URL `console.cloud.google.com/sql/instances?authuser=2&project=glassy-life-350815`. The top navigation bar includes the Google Cloud Platform logo, the project name 'My First Project', a search bar, and a '초기화' (Reset) button. The main content area features a large Cloud SQL logo and the title 'Cloud SQL 인스턴스'. Below the logo, there is a paragraph explaining that Cloud SQL instances are fully managed relational databases for MySQL, PostgreSQL, and SQL Server. It mentions that Google handles backups, patches, and database management, ensuring availability and performance. A link '자세히 알아보기' (Learn more) is provided. Another paragraph states that users can create new instances or use Database Migration Service to migrate existing SQL databases to Google Cloud. At the bottom, there are two buttons: '인스턴스 만들기' (Create Instance) and '데이터 마이그레이션' (Database Migration). The '인스턴스 만들기' button is highlighted with a red rectangular box.

클라우드 컴퓨팅 서비스 | Google | 인스턴스 - SQL - My First Project

console.cloud.google.com/sql/instances?authuser=2&project=glassy-life-350815

Google Cloud Platform | My First Project

SQL | 인스턴스 | + 인스턴스 만들기 | 데이터 마이그레이션

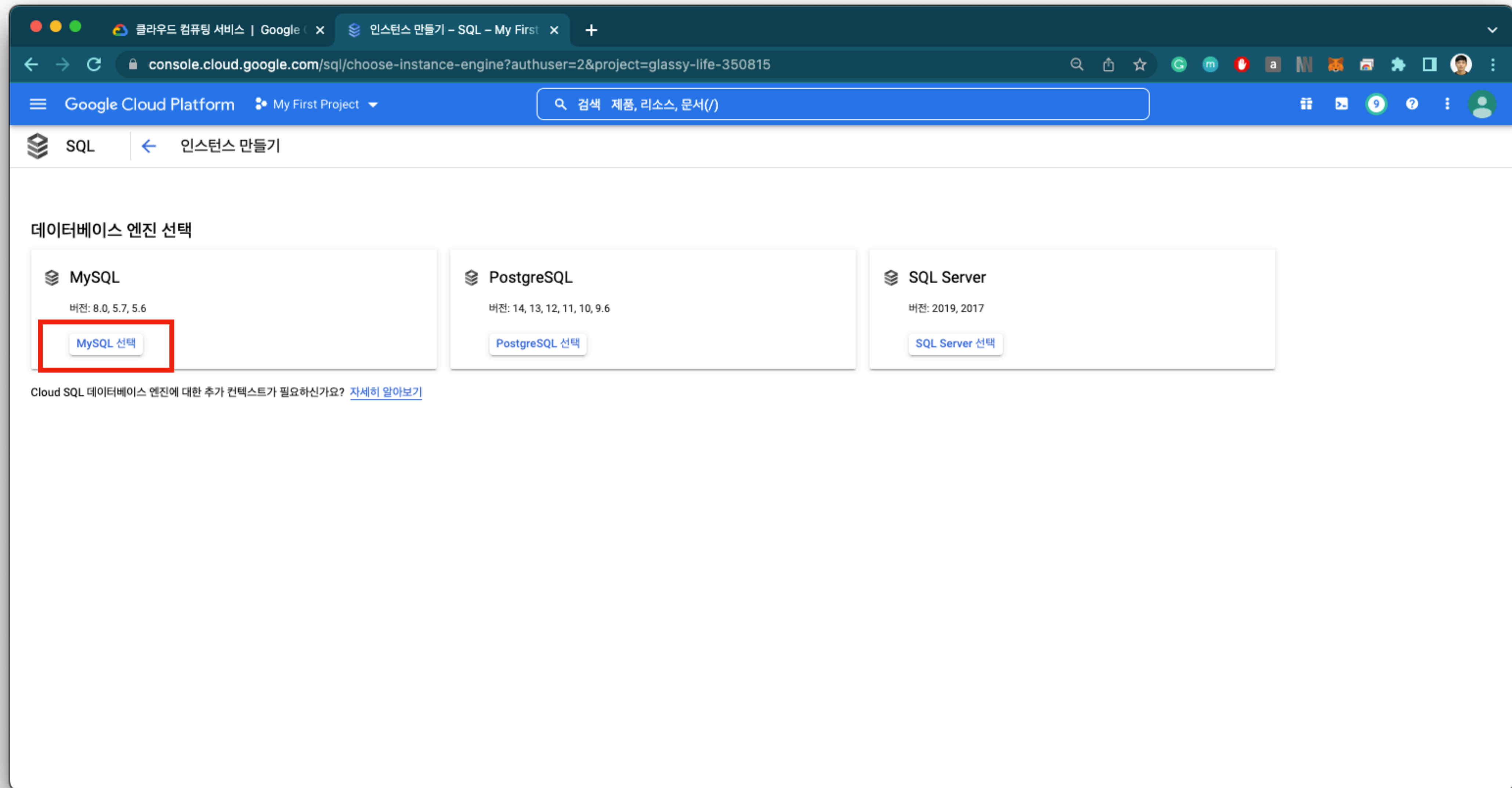
Cloud SQL 인스턴스

Cloud SQL 인스턴스는 완전 관리형인 관계형 MySQL, PostgreSQL, SQL Server 데이터베이스입니다. Google에서 복제, 패치 관리, 데이터베이스 관리를 처리하기 때문에 가용성과 성능이 보장됩니다. [자세히 알아보기](#)

Cloud SQL을 시작하려면 새 인스턴스를 만들거나 Database Migration Service를 사용해 SQL Database를 Google Cloud로 마이그레이션하면 됩니다.

인스턴스 만들기 | 데이터 마이그레이션

MySQL 선택



정보 입력

클라우드 컴퓨팅 서비스 | Google

MySQL 인스턴스 만들기 - SQL

+

← → ↺

console.cloud.google.com/sql/instances/create;engine=MySQL?authuser=2&project=glassy-life-350815

🔍 📄 ☆

🟢 🟡 🔴

📧 📺 📅 ⚙️ 🏠 👤

Google Cloud Platform My First Project 🔽

🔍 검색 제품, 리소스, 문서(/)

🏠 📄 9 ? ⋮ 👤

← MySQL 인스턴스 만들기

인스턴스 정보

인스턴스 ID *

mydatabase

인스턴스 ID는 소문자, 숫자, 하이픈을 사용할 수 있습니다. 영문자로 시작해야 합니다.

비밀번호 *

••••••••

👁️ 생성

로그인 사용자의 비밀번호를 선택하세요. 자세히 알아보기

☐ 비밀번호 없음

데이터베이스 버전 *

MySQL 8.0

리전 및 영역 가용성 선택

성능을 향상하려면 필요한 서비스와 가까운 위치에 데이터를 보관하세요. 리전은 영구적이지만 영역은 언제든지 변경할 수 있습니다.

리전

asia-northeast3 (서울)

영역 가용성

☒ 단일 영역
서비스 중단이 발생할 경우 장애 조치가 없습니다. 프로덕션에는 권장되지 않습니다.

☐ 여러 영역(고가용성)
선택한 리전 내의 다른 영역으로 자동으로 장애 조치가 적용됩니다. 프로덕션 인스턴스에 권장되며 비용이 증가합니다.

▼ 영역 지정

인스턴스 맞춤설정

나중에 인스턴스 구성을 맞춤설정할 수도 있습니다.

▼ 구성 옵션 표시

인스턴스 만들기

취소

요약

리전	asia-northeast3(서울)
DB 버전	MySQL 8.0
vCPU	4 vCPU
메모리	26GB
저장용량	100GB
네트워크 처리량(MB/초)	1,000/2,000
디스크 처리량(MB/초)	읽기: 48.0/240.0 쓰기: 48.0/240.0
IOPS	읽기: 3,000/15,000 쓰기: 3,000/15,000
연결	공개 IP
백업	자동
가용성	단일 영역
point-in-time recovery	사용 설정됨

조금 기다리기..

클라우드 컴퓨팅 서비스 | Google

mydatabase - 개요 - SQL - My

+

← → ↺

console.cloud.google.com/sql/instances/mydatabase/overview?authuser=2&project=glassy-life-350815

🔍 📄 ☆

🟢 🟡 🔴 🟠 🟤 🟦 🟩 🟪 🟨 🟧 🟥 🟣

👤

☰ Google Cloud Platform

My First Project

🔍 검색 제품, 리소스, 문서(/)

🏠 📄 9 ? ⋮

SQL

기본 인스턴스

개요

연결

사용자

데이터베이스

백업

복제본

작업

개요

수정

가져오기

내보내기

다시 시작

중지

삭제

클론

모든 인스턴스 > mydatabase

mydatabase

MySQL 8.0

인스턴스 생성 중. 완료되는 데 몇 분 정도 걸릴 수 있습니다. 작업이 실행되는 동안에도 인스턴스 관련 정보를 계속 확인할 수 있습니다.

1시간 6시간 1일 7일 30일 커스텀

차트 CPU 사용률

선택한 기간에는 데이터가 없습니다.

2022. 6. 3. 오전 7:47:55

이 인스턴스에 연결

연결 이름

glassy-life-350815:asia-northeast3:mydatabase

연결하는 데 도움이 필요하신가요?

문서를 검토하여 인스턴스에 연결하는 다양한 방법에 대해 알아보세요. 자세히 알아보기

gcloud를 사용하여 연결하려면 다음 안내를 따르세요.

CLOUD SHELL 열기

Compute Engine VM과 연결하는 방법을 알아보려면 다음 안내를 따르세요.

튜토리얼 시작

구성

vCPU 4

메모리 26GB

SSD 저장용량 100GB

데이터베이스 버전은 MySQL 8.0.26입니다.

저장용량 자동 증가가 사용 설정되었습니다.

자동 백업을 사용 설정했습니다.

point-in-time recovery 사용 설정됨

asia-northeast3-b에 있습니다.

가용성이 높지 않음(영역)

업로드 및 My First Project 작업

mydatabase 생성 중

0분 12초

인스턴스 런칭 성공!

클라우드 컴퓨팅 서비스 | Google

mydatabase - 개요 - SQL - My

+

← → ↺

console.cloud.google.com/sql/instances/mydatabase/overview?authuser=2&project=glassy-life-350815

🔍 📄 ☆

🟢 🟡 🔴 🟠 🟤 🟣 🟦 🟩 🟪 🟨 🟧 🟥 🟦 🟩 🟪 🟨 🟧 🟥

👤

☰ Google Cloud Platform

My First Project

🔍 검색 제품, 리소스, 문서(/)

🏠 📄 9 ? ⋮

SQL

기본 인스턴스

개요

연결

사용자

데이터베이스

백업

복제본

작업

개요

수정

가져오기

내보내기

다시 시작

중지

삭제

클론

모든 인스턴스 > mydatabase

✅ mydatabase

MySQL 8.0

1시간 6시간 1일 7일 30일 커스텀

차트 CPU 사용률

UTC+9 오전 4:00 오전 6:00 오전 8:00 오전 10:00 오전 12:00 오후 2:00 오후 4:00 오후 6:00 오후 8:00 오후 10:00 6월 4일

이 인스턴스에 연결

공개 IP 주소 34.64.120.112

연결 이름 glassy-life-350815:asia-northeast3:mydatabase

연결하는 데 도움이 필요하신가요?

문서를 검토하여 인스턴스에 연결하는 다양한 방법에 대해 알아보세요. 자세히 알아보기

gcloud를 사용하여 연결하려면 다음 안내를 따르세요. CLOUD SHELL 열기

Compute Engine VM과 연결하는 방법을 알아보려면 다음 안내를 따르세요. 튜토리얼 시작

구성

vCPU 4 메모리 26GB SSD 저장용량 100GB

데이터베이스 버전은 MySQL 8.0.26입니다.

저장용량 자동 증가가 사용 설정되었습니다.

자동 백업을 사용 설정했습니다.

point-in-time recovery 사용 설정됨

asia-northeast3-b에 있습니다.

가용성이 높지 않음(영역)

설정된 데이터베이스 플러그가 없습니다.

업로드 및 My First Project 작업

✅ mydatabase 생성됨 AM 1시 28분 21초 GMT+9

네트워크 연결 설정하기

클라우드 컴퓨팅 서비스 | Google | mydatabase - 연결 - SQL - My

console.cloud.google.com/sql/instances/mydatabase/connections/networking?authuser=2&project=glassy-life-350815

Google Cloud Platform | My First Project

SQL

기본 인스턴스

개요

연결

사용자

데이터베이스

백업

복제본

작업

연결

모든 인스턴스 > mydatabase

mydatabase

MySQL 8.0

네트워킹 | 보안 | 연결 테스트

소스를 이 인스턴스에 연결할 방법을 선택한 다음 연결하도록 승인된 네트워크를 정의합니다. [자세히 알아보기](#)

두 옵션 중 하나로 Cloud SQL 프록시를 사용하여 보안을 강화할 수 있습니다. [자세히 알아보기](#)

인스턴스 IP 할당

☐ 비공개 IP
Google에서 호스팅하는 내부 VPC IP 주소를 할당합니다. 추가 API 및 권한이 필요합니다. 사용 설정한 후에는 사용 중지할 수 없습니다. [자세히 알아보기](#)

☒ 공개 IP
인터넷에 액세스할 수 있는 외부 IP 주소를 할당합니다. 이 인스턴스에 연결하려면 승인된 네트워크 또는 Cloud SQL 프록시를 사용해야 합니다. [자세히 알아보기](#)

승인된 네트워크

CIDR 범위를 지정하여 해당 범위의 IP 주소가 인스턴스에 액세스하도록 허용할 수 있습니다. [자세히 알아보기](#)

i Cloud SQL 인스턴스에 연결하도록 승인된 외부 네트워크가 없습니다. Cloud SQL 프록시를 사용하면 외부 애플리케이션에서 인스턴스에 연결할 수 있습니다. [자세히 알아보기](#)

네트워크 추가

App Engine 승인

이 프로젝트의 모든 앱이 기본적으로 승인됩니다. [Cloud IAM](#)을 사용하여 다른 프로젝트의 앱을 승인할 수 있습니다. [자세히 알아보기](#)

저장 | 변경사항 취소

출시 노트

업로드 및 My First Project 작업

mydatabase 생성됨

AM 1시 28분 21초 GMT+9

모두 허용!

클라우드 컴퓨팅 서비스 | Google

mydatabase - 연결 - SQL - My

console.cloud.google.com/sql/instances/mydatabase/connections/networking?authuser=2&project=glassy-life-350815

Google Cloud Platform My First Project

검색 제품, 리소스, 문서(/)

9 ?

SQL

기본 인스턴스

개요

연결

사용자

데이터베이스

백업

복제본

작업

연결

☐ 비공개 IP
Google에서 호스팅하는 내부 VPC IP 주소를 할당합니다. 추가 API 및 권한이 필요합니다. 사용 설정한 후에는 사용 중지할 수 없습니다.[자세히 알아보기](#)

☒ 공개 IP
인터넷에 액세스할 수 있는 외부 IP 주소를 할당합니다. 이 인스턴스에 연결하려면 승인된 네트워크 또는 Cloud SQL 프록시를 사용해야 합니다.[자세히 알아보기](#)

승인된 네트워크
CIDR 범위를 지정하여 해당 범위의 IP 주소가 인스턴스에 액세스하도록 허용할 수 있습니다.[자세히 알아보기](#)

⚠️
허용되는 네트워크로 0.0.0.0/0을 추가했습니다. 이 프리픽스는 허용할 의도가 없었던 클라이언트를 포함하여 모든 IPv4 클라이언트에서 네트워크 방화벽을 통과하고 인스턴스 로그인에 시도할 수 있도록 허용합니다. 클라이언트에서 인스턴스에 로그인하려면 유효한 사용자 인증 정보가 필요합니다.

새 네트워크

이름
dbnetwork

[CIDR 표기법](#) 사용
네트워크 *
0.0.0.0/0
예: 199.27.25.0/24

취소

완료

네트워크 추가

App Engine 승인
이 프로젝트의 모든 앱이 기본적으로 승인됩니다. [Cloud IAM](#)을 사용하여 다른 프로젝트의 앱을 승인할 수 있습니다.[자세히 알아보기](#)

저장

변경사항 취소

업로드 및 My First Project 작업

✓ mydatabase 생성됨

AM 1시 28분 21초 GMT+9

조금 더 기다리기..

클라우드 컴퓨팅 서비스 | Google

mydatabase - 연결 - SQL - My

+

← → ↺

console.cloud.google.com/sql/instances/mydatabase/connections/networking?authuser=2&project=glassy-life-350815

🔍 📄 ☆

🟢 🟡 🔴

📧 📺 📅 ⚙️ 🏠 👤

Google Cloud Platform

My First Project

🔍 검색 제품, 리소스, 문서(/)

📦 📄 9 ? ⋮

SQL

기본 인스턴스

개요

연결

사용자

데이터베이스

백업

복제본

작업

연결

모든 인스턴스 > mydatabase

mydatabase

MySQL 8.0

🔔 인스턴스 업데이트 중. 완료되는 데 몇 분 정도 걸릴 수 있습니다. 작업이 실행되는 동안에도 인스턴스 관련 정보를 계속 확인할 수 있습니다.

네트워킹 보안 연결 테스트

소스를 이 인스턴스에 연결할 방법을 선택한 다음 연결하도록 승인된 네트워크를 정의합니다.[자세히 알아보기](#)

두 옵션 중 하나로 Cloud SQL 프록시를 사용하여 보안을 강화할 수 있습니다.[자세히 알아보기](#)

인스턴스 IP 할당

☐ 비공개 IP

Google에서 호스팅하는 내부 VPC IP 주소를 할당합니다. 추가 API 및 권한이 필요합니다. 사용 설정한 후에는 사용 중지할 수 없습니다.[자세히 알아보기](#)

☒ 공개 IP

인터넷에 액세스할 수 있는 외부 IP 주소를 할당합니다. 이 인스턴스에 연결하려면 승인된 네트워크 또는 Cloud SQL 프록시를 사용해야 합니다.[자세히 알아보기](#)

승인된 네트워크

CIDR 범위를 지정하여 해당 범위의 IP 주소가 인스턴스에 액세스하도록 허용할 수 있습니다.[자세히 알아보기](#)

⚠️ 허용되는 네트워크로 0.0.0.0/0을 추가했습니다. 이 프리픽스는 허용할 의도가 없었던 클라이언트를 포함하여 모든 IPv4 클라이언트에서 네트워크 방화벽을 통과하고 인스턴스 로그인을 시도할 수 있도록 허용합니다. 클라이언트에서 인스턴스에 로그인하려면 유효한 사용자 인증 정보가 필요합니다.

dbnetwork (0.0.0.0/0) (저장되지 않음) ▼

네트워크 추가

App Engine 승인

이 프로젝트의 모든 앱이 기본적으로 승인됩니다. [Cloud IAM](#)을 사용하여 다른 프로젝트의 앱을 승인할 수 있습니다.[자세히 알아보기](#)

출시 노트

◀

업로드 및 My First Project 작업

🔄 다음 인스턴스를 수정하는 중: 0분 9초

[mydatabase](#)

✅ mydatabase 생성됨 AM 1시 28분 21초 GMT+9

E.O.D