GCP기반의 금융데이터 분석

Day 1

INSTITUTE FOR INFORMATION

Chapter 01 GCP 소개

Chapter 02 GCE를 활용한 분석 환경 구축

Chapter 03 Cloud 환경에서의 Python 사용

Day 2

INSTITUTE FOR INFORMATION

Chapter 01 Cloud 환경에서의 금융 데이터 분석

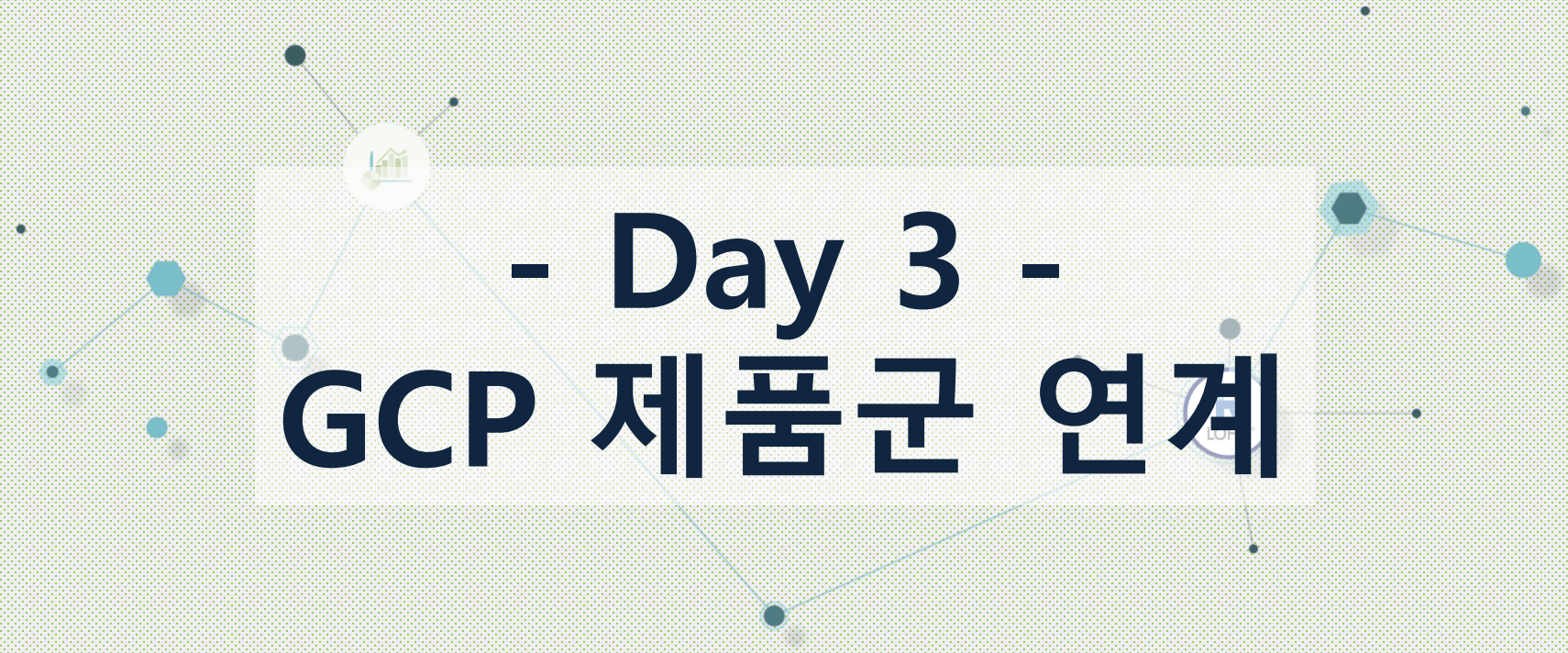
Chapter 02 Python 기반의 BigQuery 접근과 데이터 조작

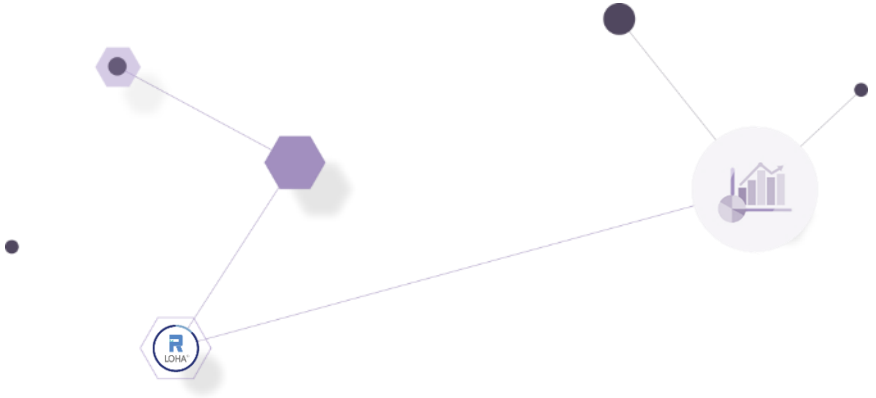
Day 3

INSTITUTE FOR INFORMATION

Chapter 01 크롤링을 통한 외부 데이터 수집과 적재

Chapter 02 Vertex AI를 활용한 머신러닝 모델 구현


- Day 3 -
GCP 제품군 연계



1) 크롤링을 통한 외부데이터 수집과 적재

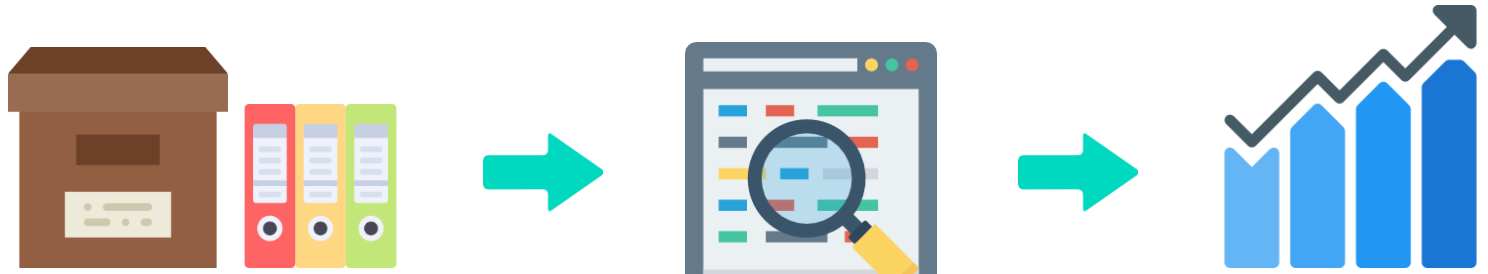
01 크롤링 시작하기

[Day 03] GCP 제품군 연계

크롤링(Crawling)?

- 컴퓨터 소프트웨어 기술로 웹 사이트들에서 원하는 정보를 추출하고 저장하는 일련의 과정
- 크롤링(Crawling), 스크레이핑(Scraping) 등 다양한 용어 존재
- 커뮤니티 게시글/댓글 수집, 온라인 컨텐츠 큐레이팅 등 다양한 곳에 활용
- 수백 수천번의 수작업 대신 자동화를 통하여 빠른 자료수집 가능

크롤러(Crawler)의 활용

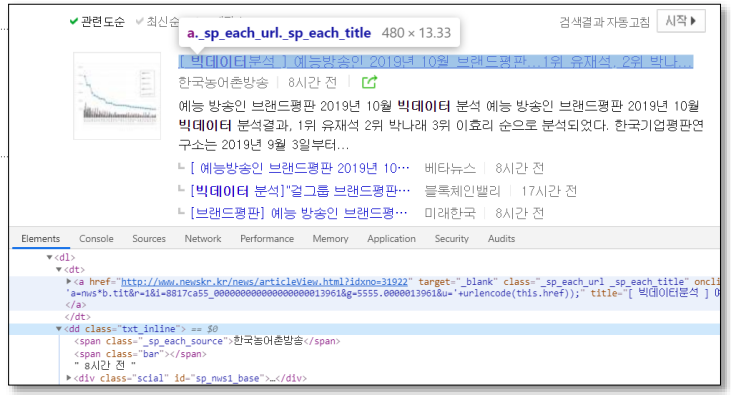


01 크롤링 시작하기

[Day 03] GCP 제품군 연계

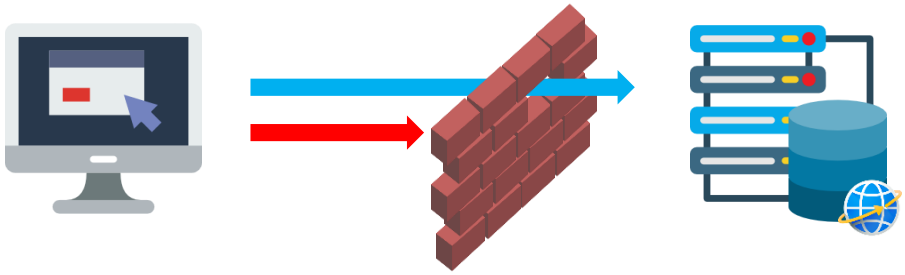
크롬 개발자 도구

- 웹 개발 및 디버깅을 위해 제공하는 크롬의 도구
- 웹의 각 요소의 위치와 코드를 보다 쉽게 확인할 수 있다.



주의사항

DDoS 오인



robots.txt

01 크롤링 시작하기

[Day 03] GCP 제품군 연계

법정 공방

▶ 타사의 데이터를 크롤링하여 상업적 목적으로 무단 사용할 경우 관련법에 의거 처벌받을 수 있음

매일경제

`야놀자` 숙박정보 복제한 `여기어때` 무죄 확정

야놀자` 숙박정보 복제한 `여기어때` 무죄 확정 - 매일경제, 작성자-김형주, 섹션-society, 요약-숙박업체 예약 중개 서비스 `야놀자`의 서버에...

3일 전

지디넷코리아

대법원, 야놀자 정보 크롤링 한 여기어때 창업주 '무죄'

숙박, 여가 플랫폼 선두 사업자 야놀자의 영업 정보를 무단으로 빼돌린 혐의로 기소된 여기어때 관계자들이 최종 무죄 판결받았다.

3일 전

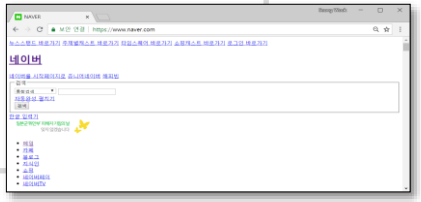
01 크롤링 시작하기

[Day 03] GCP 제품군 연계

웹페이지 구성

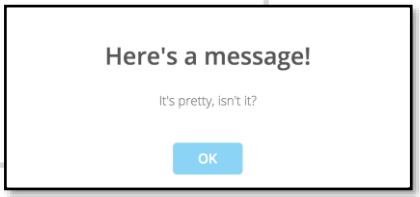
▶ 기본적으로 HTML, CSS, JavaScript를 말하며 추가로 PHP, JSP 등이 있다.

웹 문서의 전체적인 구조 담당



웹문서의 스타일(색상, 모양 등) 담당

웹문서의 각종 동적 동작 담당



02 데이터 수집 준비

[Day 03] GCP 제품군 연계

HTML이란?

- **H**yper**T**ext **M**arkup **L**anguage. 즉, 문서의 활자 및 조판을 지정하는 언어
- 웹 페이지의 골격을 구성하는 기본 언어
- 다양한 태그(tag)가 있으며 열리는 태그와 닫히는 태그가 쌍으로 되어있는 경우가 많음

02 데이터 수집 준비

[Day 03] GCP 제품군 연계

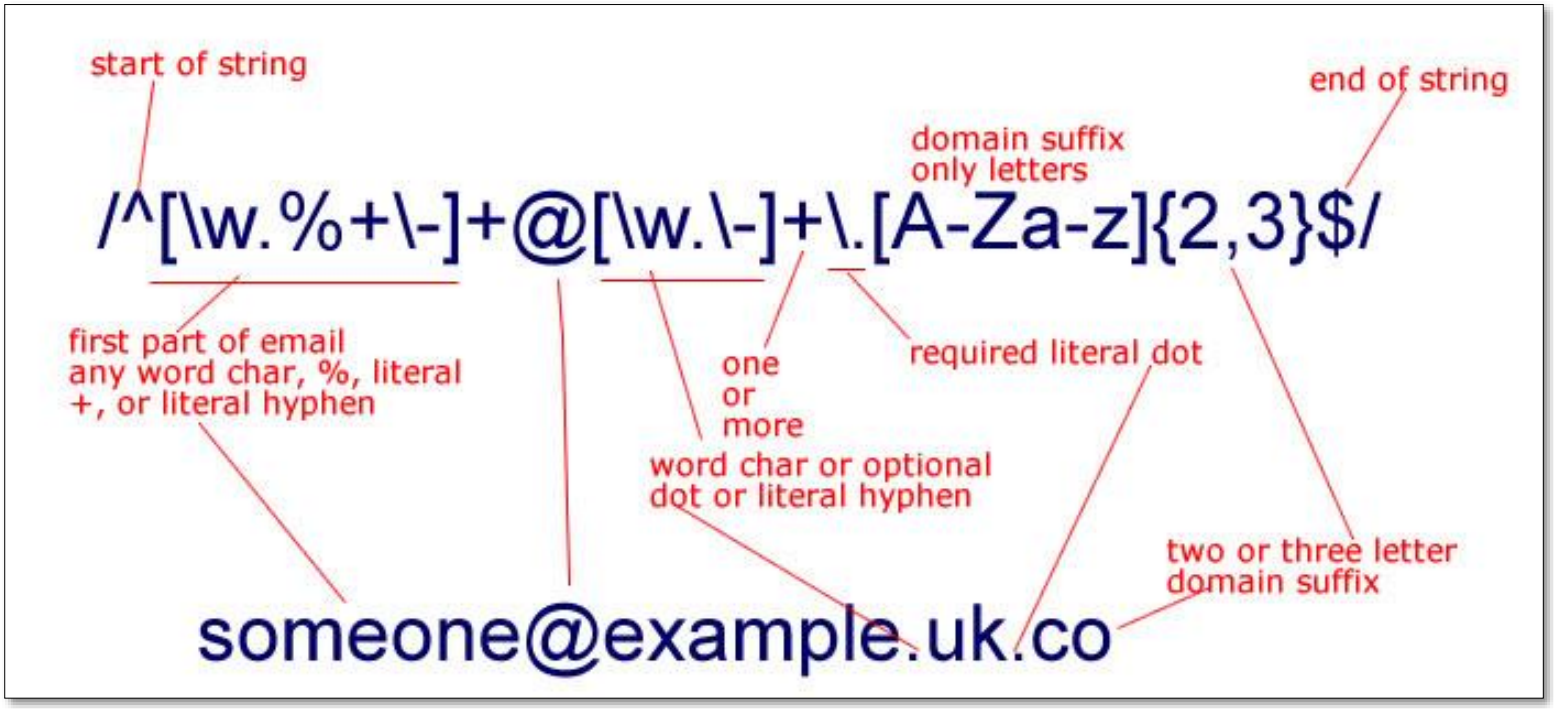
URL 이해하기

▶ URL에는 다양한 정보가 들어있으며 이를 구조적으로 해석해야 함

[Day 03] GCP 제품군 연계

정규표현식(regular expression)

- ▶ 특정한 규칙을 가진 문자열의 집합을 표현하는 데 사용하는 형식 언어
- ▶ 정규 표현식은 많은 텍스트 편집기와 프로그래밍 언어에서 문자열의 검색과 치환을 위해 지원하고 있음



02 데이터 수집 준비

[Day 03] GCP 제품군 연계

정규표현식 주요 문법

문법	설명
^	시작
$	끝
.	모든 문자열
₩.	마침표
[0-9]	숫자
[a-z]	모든 영문 소문자
[A-Z]	모든 영문 대문자
[a-zA-Z]	모든 영문자
[ㄱ-ㅎ]	한글 초성
[가-힣]	한글 완성형
[^0-9]	숫자가 아닌 모든 문자
ab|cd	ab 또는 cd
<.*?>	<로 시작하고 >로 끝나는 모든 패턴(<>, <a>, 등)

정규표현식 주요 문법

문법	설명
{n}	직전 문자를 n번 반복
ab{4}	a 다음에 b가 위치하며 b가 4번 반복됨(abbbb)
{n,}	직전 문자가 n번 이상 반복
{n,m}	직전 문자가 n번 이상 m번 이하 반복
ab(0|1)	ab0 또는 ab1
₩t	탭(tab)

02 데이터 수집 준비

[Day 03] GCP 제품군 연계

requests

- ▶ Python에서 HTTP 요청을 보내는 모듈
- ▶ GET요청과 POST요청을 하는 함수가 있다.
- ▶ 필요시 파라미터, 헤더, 쿠키를 추가할 수 있다.

사용 예시

```python
import requests
URL = 'https://www.naver.com'
res = requests.get(URL)
res.status_code
res.text
```

```
'<!doctype html>₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n<html lang="ko">₩n
n">₩n<meta http-equiv="Content-Script-Type" content="text/javascript
http-equiv="X-UA-Compatible" content="IE=edge">₩n<meta name="viewpor
ntent="NAVER" />₩n<meta name="robots" content="index,nofollow"/>₩n<m
컨텐츠를 만나 보세요"/>₩n<meta property="og:title" content="네이버">
```

02 데이터 수집 준비

[Day 03] GCP 제품군 연계

urllib

- Python에서 HTTP 요청을 보내는 모듈
- get요청과 post요청을 하는 함수가 있다.
- 필요시 파라미터, 헤더, 쿠키를 추가할 수 있다.

사용 예시

```
1  import urllib
2  url = "https://www.naver.com"
3  text = urllib.request.urlopen(url)
4  text.read(500)
```

```
b'<!doctype html>₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n₩n<html lang="ko">₩n<
n">₩n<meta http-equiv="Content-Script-Type" content="text/javascript">
tp-equiv="X-UA-Compatible" content="IE=edge">₩n<meta name="viewport" c
="NAVER" />₩n<meta name="robots" content="index,nofollow"/>₩n<meta nam
9₩x94₩xec₩x9d₩xb8₩xec₩x97₩x90'
```

02 데이터 수집 준비

[Day 03] GCP 제품군 연계

bs4

- Python 웹크롤링에서 가장 대중적으로 사용되는 라이브러리
- beautifulsoup이 정식명칭이며 최신 버전은 4이다.
- 수집된 데이터를 보다 효과적으로 처리할 수 있도록 해준다.

사용 예시

```
1  from bs4 import BeautifulSoup as bs
2  text = "<html><div>bs4!!</div></html>"
3  text_bs = bs(text, "html.parser")
4  text_bs.text
```

```
'bs4!!'
```

02 데이터 수집 준비

[Day 03] GCP 제품군 연계

GET vs. POST

- GET은 ?뒤에 파라미터를 붙여서 특정 자료를 요청하고 전달받는 것.
- GET의 경우 전달되는 데이터의 문자가 255자를 초과하면 문제가 발생할 수 있다.
- POST는 상대적으로 양이 많은 파라미터를 data 파라미터에 실어 전송하여 특정 동작을 수행하도록 하는 것
- POST의 경우 많은 양의 파라미터를 전달해야할 경우 json 모듈이 필요할 수 있다.

관련 함수/메소드

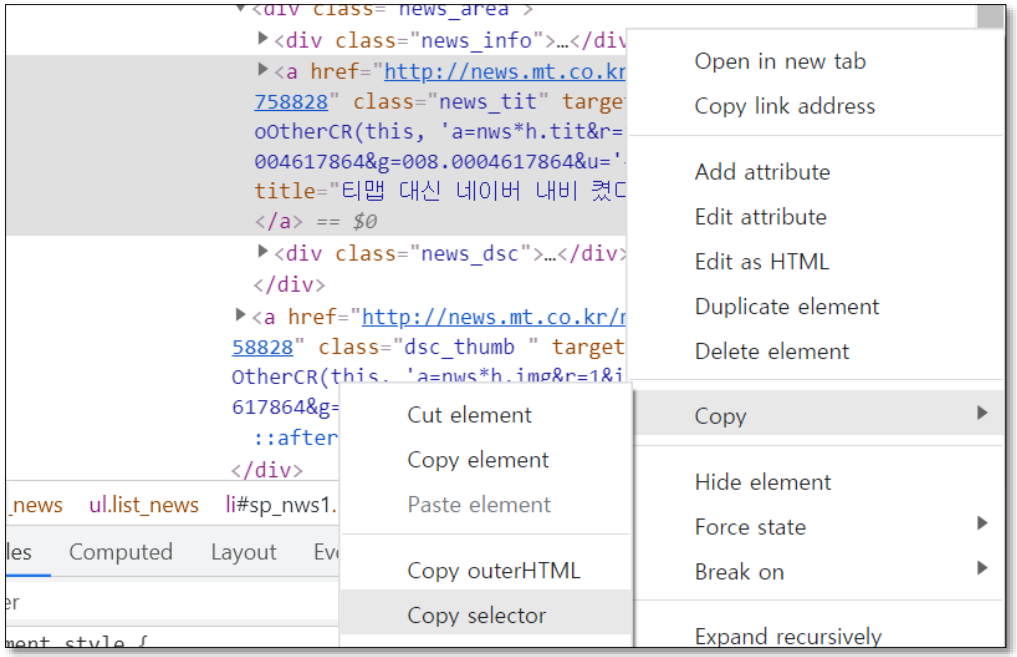
함수/메소드	모듈	주요 인자	설명
`get()`	requests	url headers cookies	HTTP 프로토콜 중 GET 방식으로 서버에 자료를 요청하는 방식
`post()`	requests	url data	HTTP 프로토콜 중 POST 방식으로 서버에 자료를 요청하는 방식

02 데이터 수집 준비

[Day 03] GCP 제품군 연계

CSS 선택자(Selector)

▶ HTML 문서의 특정 요소의 스타일 지정을 위해 사용하는 도구
▶ 웹 브라우저의 개발자 도구를 활용하면 특정 원소의 CSS Selector를 쉽게 추출할 수 있음
▶ 보다 정교한 원소 선택 및 접근을 위해서 다양한 문법을 활용할 수 있음
※ 추가 설명: https://www.nextree.co.kr/p8468



CSS 선택자 주요 문법

▶ Beautifulsoup() 함수로 변환한 객체의 .select() 메서드와 CSS Selector를 활용하여 특정 태그에 접근
▶ .select() 메서드를 통해 반환된 객체는 리스트 객체의 문법과 같이 하위 속성에 접근 가능

```
from bs4 import BeautifulSoup as bs
text = bs("<body><div></div></body>", features = "html.parser")
```

```
text.select("body") # body 태그
text.select("body div") # body 태그 하위의 div 태그
text.select("body > div") # body 태그 바로 아래의 div 태그
text.select(".news") # class명이 news인 태그
text.select(".news.box") # class명이 news이면서 box인 태그
text.select("div.news") # div태그 중 class명이 news인 태그
text.select("#pw") # ID가 pw인 태그
text.select("div#pw") # ID가 pw인 div태그
text.select("a[href]") # a태그 중에서 href 속성이 있는 태그
```

03 API의 활용

[Day 03] GCP 제품군 연계

API 문서

▶ 사용자로 하여금 제공되는 API를 보다 쉽게 이용할 수 있도록 작성된 문서
▶ 공공기관은 주로 문서(워드, 한글)파일로 제공된다.

DART 오픈API 개발가이드

- DART 공시 정보에 대한 검색 결과를 외부 개발자 및 사용자에게 XML 또는 JSON(P) 형식으로 제공하는 API서비스입니다.
- DART 오픈API를 이용해 DART의 공시 정보를 자사의 홈페이지나 앱에서 서비스할 수 있습니다.

검색API

요청 주소

http://dart.fss.or.kr/api/search.xml?auth=xxx	xml 응답
http://dart.fss.or.kr/api/search.json?auth=xxx	json 응답

요청 변수

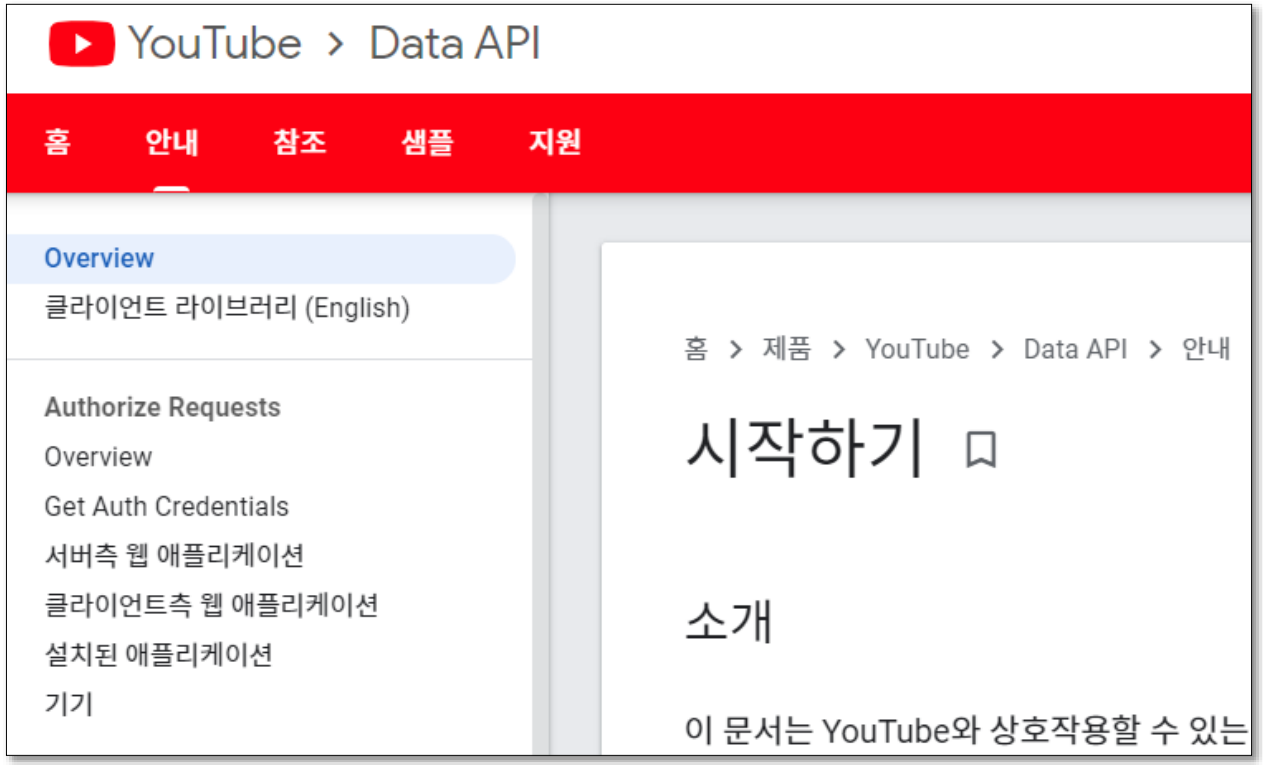
요청변수	설명
auth	발급받은 인증키(40자리) (필수)
crp_cd	공시대상회사의 종목코드(상장사)(6자리숫자) 또는 고유번호(기타법인)(8자리숫자)
end_dt	검색종료 접수일자(YYYYMMDD) : 없으면 당일
start_dt	검색시작 접수일자(YYYYMMDD) : 없으면 end_dt crp_cd가 없는 경우 검색기간은 3개월로 제한

04 GCP - youtube

[Day 03] GCP 제품군 연계

개요

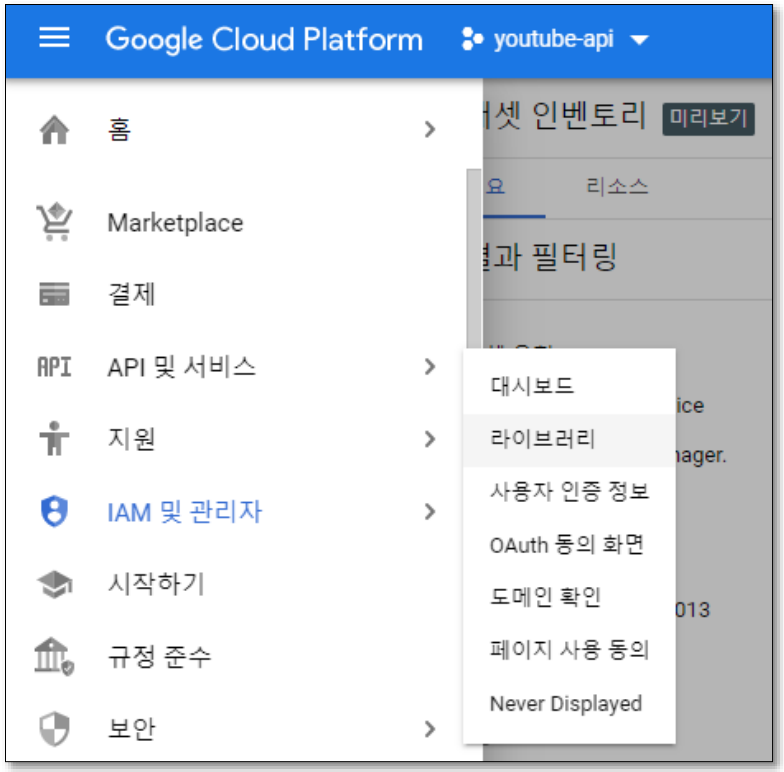
▶ Youtube 영상과 관련한 각종 정보를 API로 제공
▶ 영상 통계, 댓글, 사용자 정보 등 다양한 정보 호출 가능

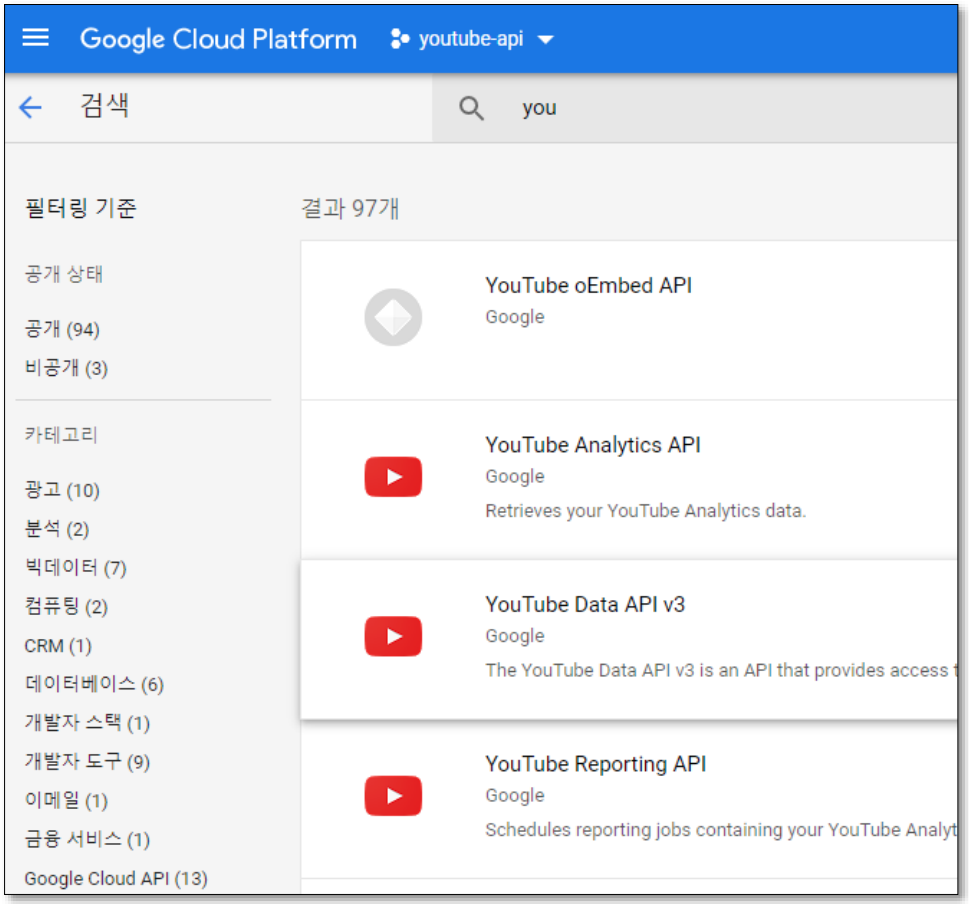


04 GCP - youtube

[Day 03] GCP 제품군 연계

신청 절차

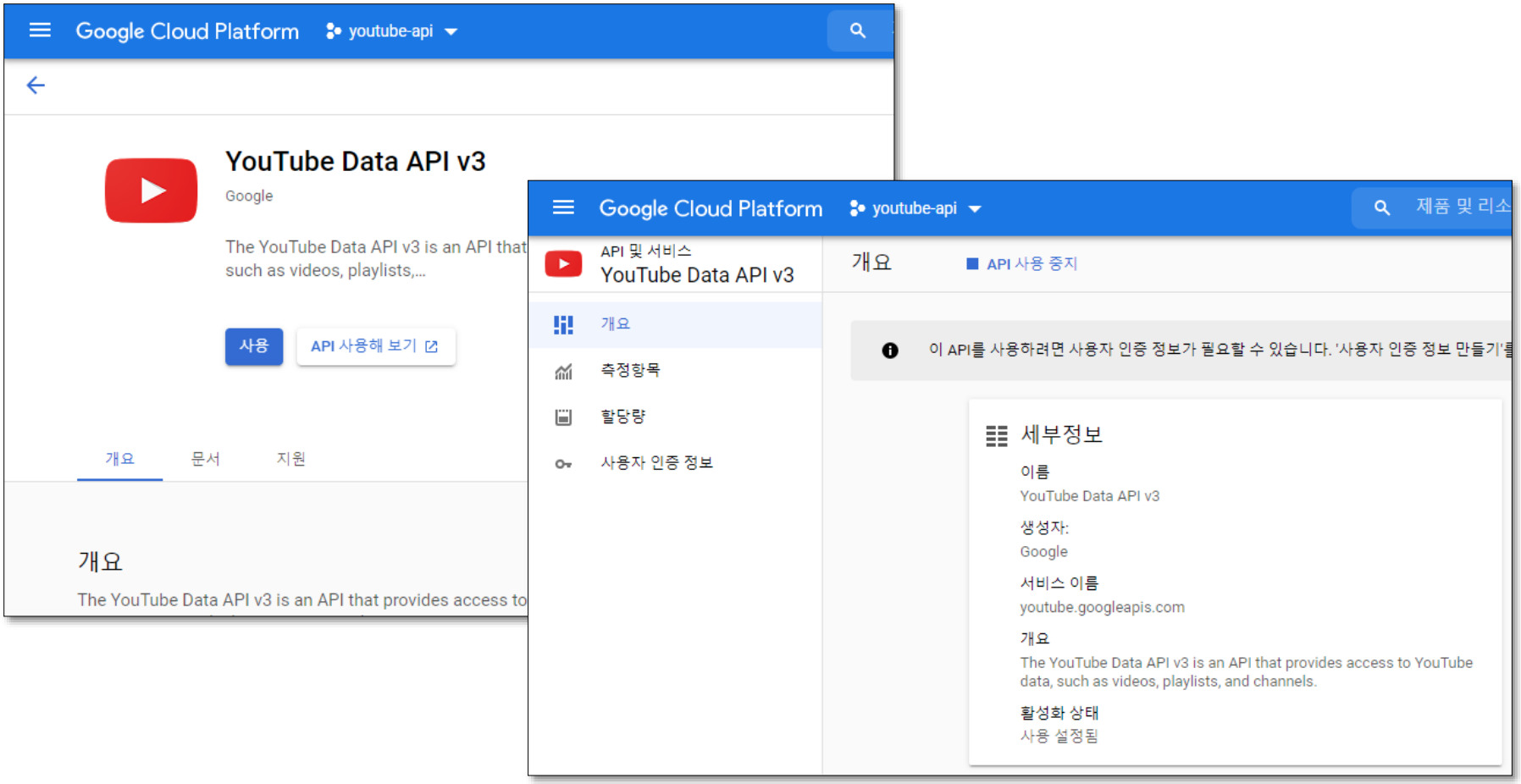




04 GCP - youtube

[Day 03] GCP 제품군 연계

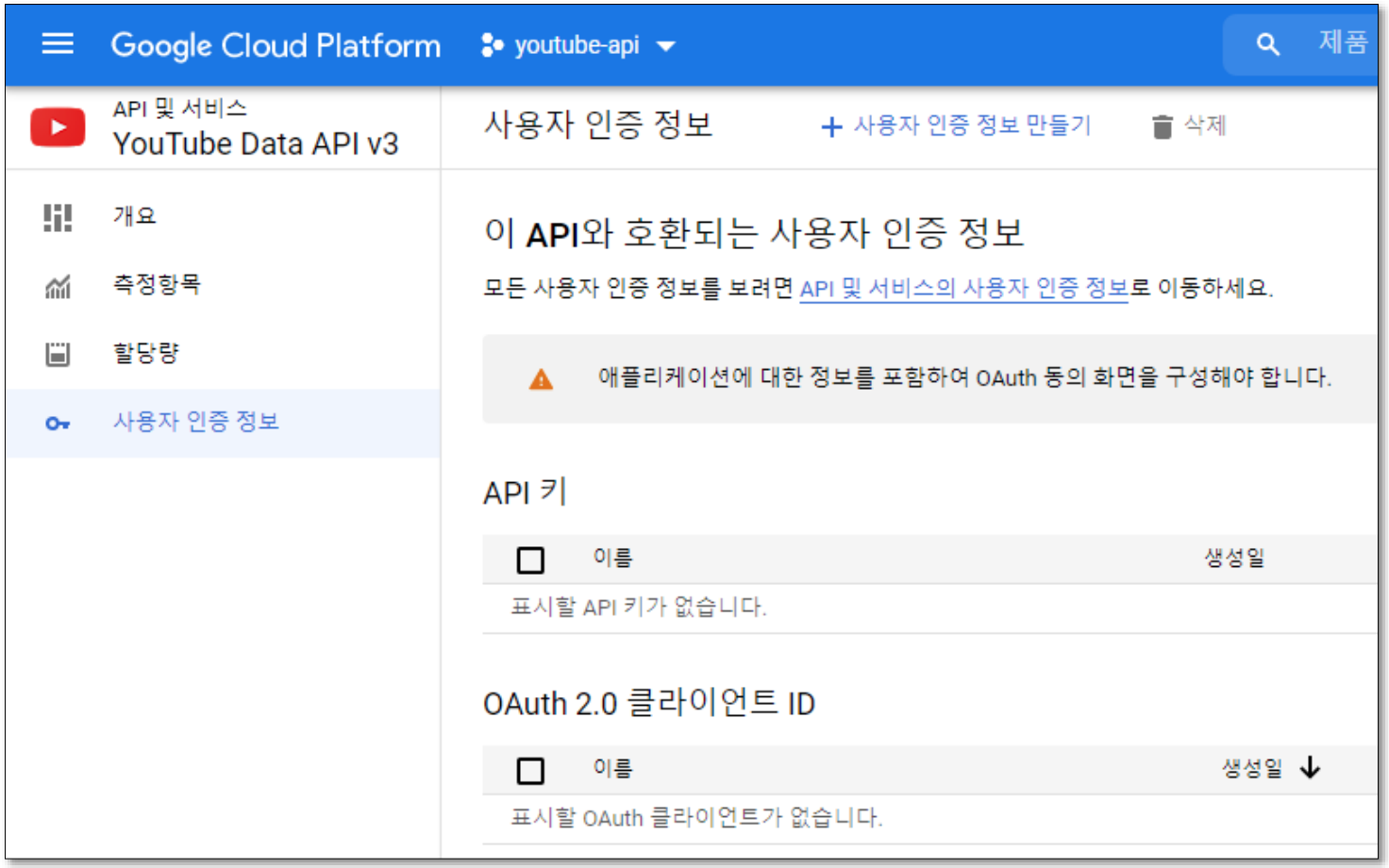
신청 절차



04 GCP - youtube

[Day 03] GCP 제품군 연계

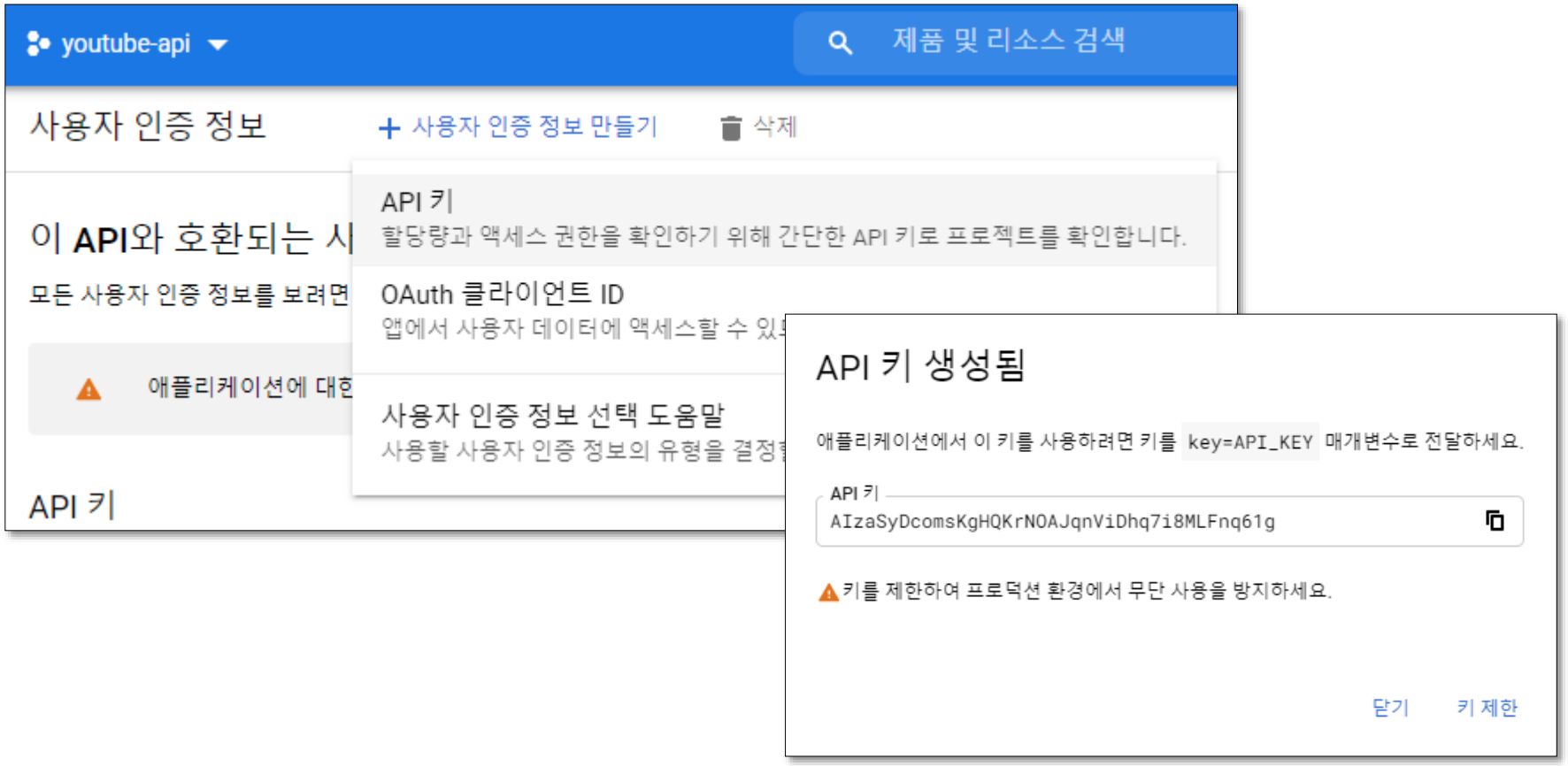
API key 설정



04 GCP - youtube

[Day 03] GCP 제품군 연계

API key 설정

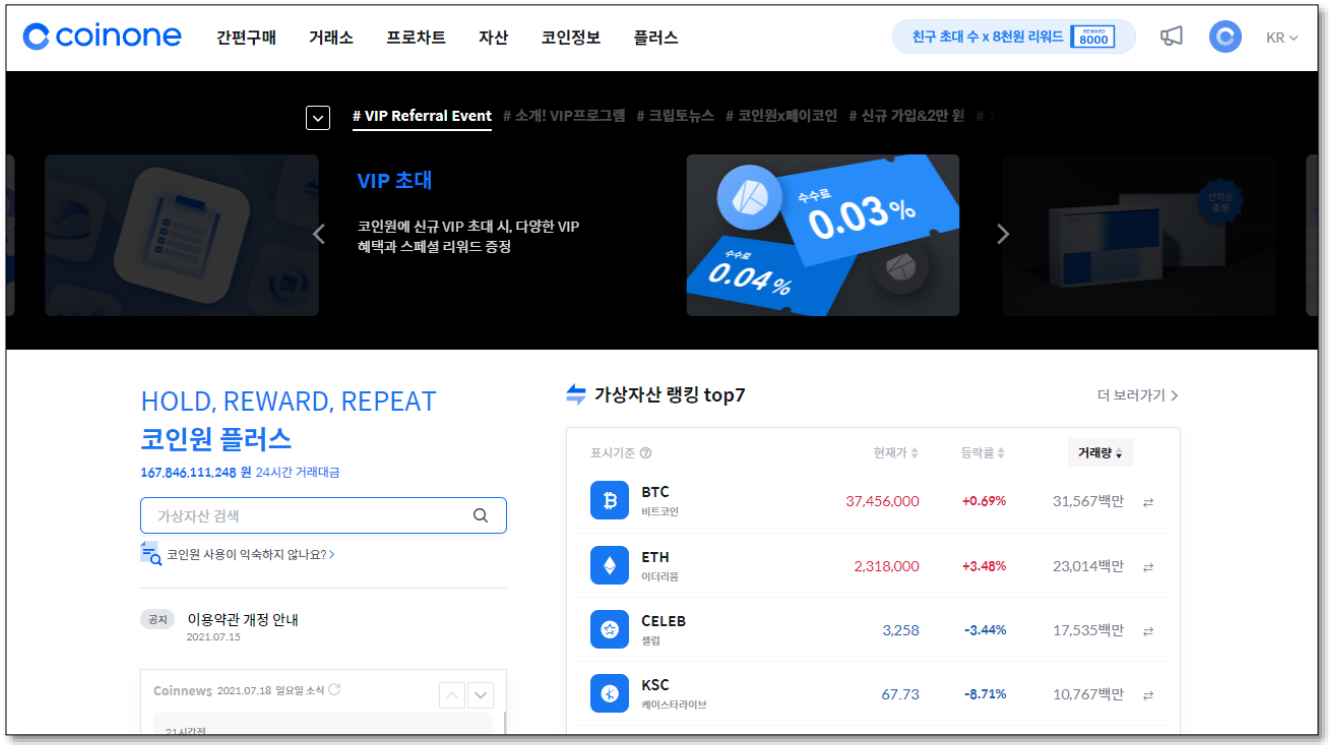


05 Coinone API

[Day 03] GCP 제품군 연계

개요

▶ 암호화폐 거래소의 각종 정보를 API로 제공
▶ 각 암호화폐별 시세 확인 및 매매 가능

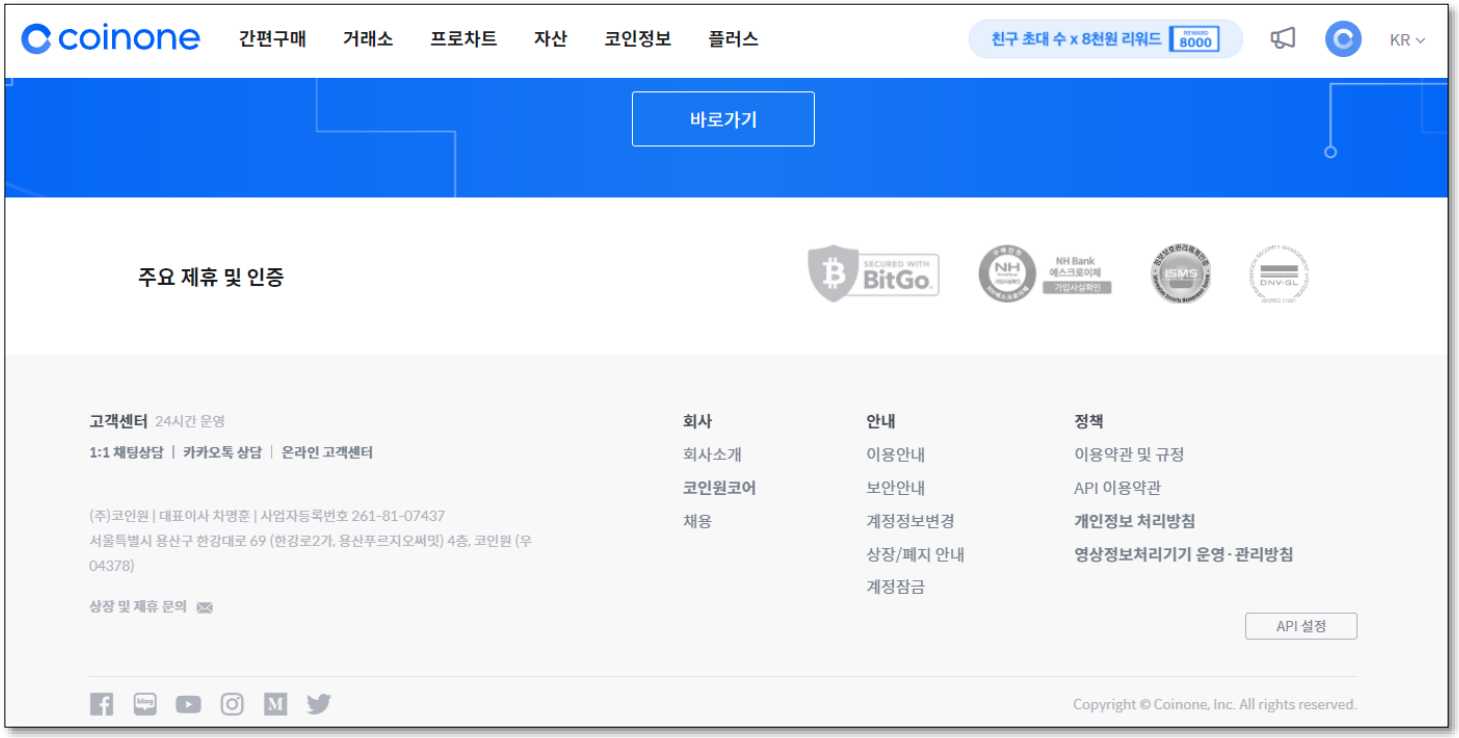


05 Coinone API

[Day 03] GCP 제품군 연계

API 문서 조회

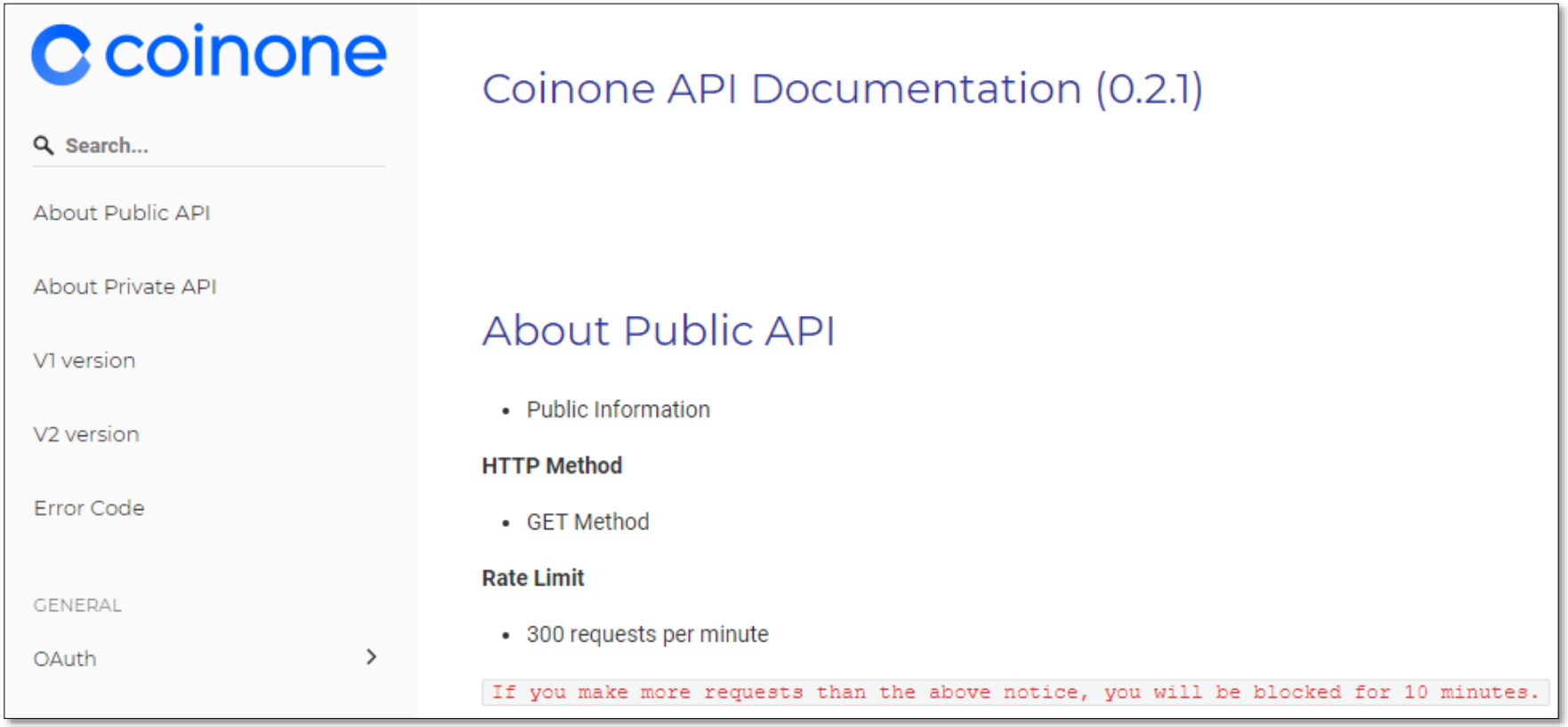
▶ 홈페이지의 footer(웹페이지의 바닥)의 왼쪽 아래에 [API 설정] 버튼 확인 필요
▶ Coinone의 경우 별도의 웹페이지에서 영문으로 API 문서 제공



05 Coinone API

[Day 03] GCP 제품군 연계

API 문서 조회

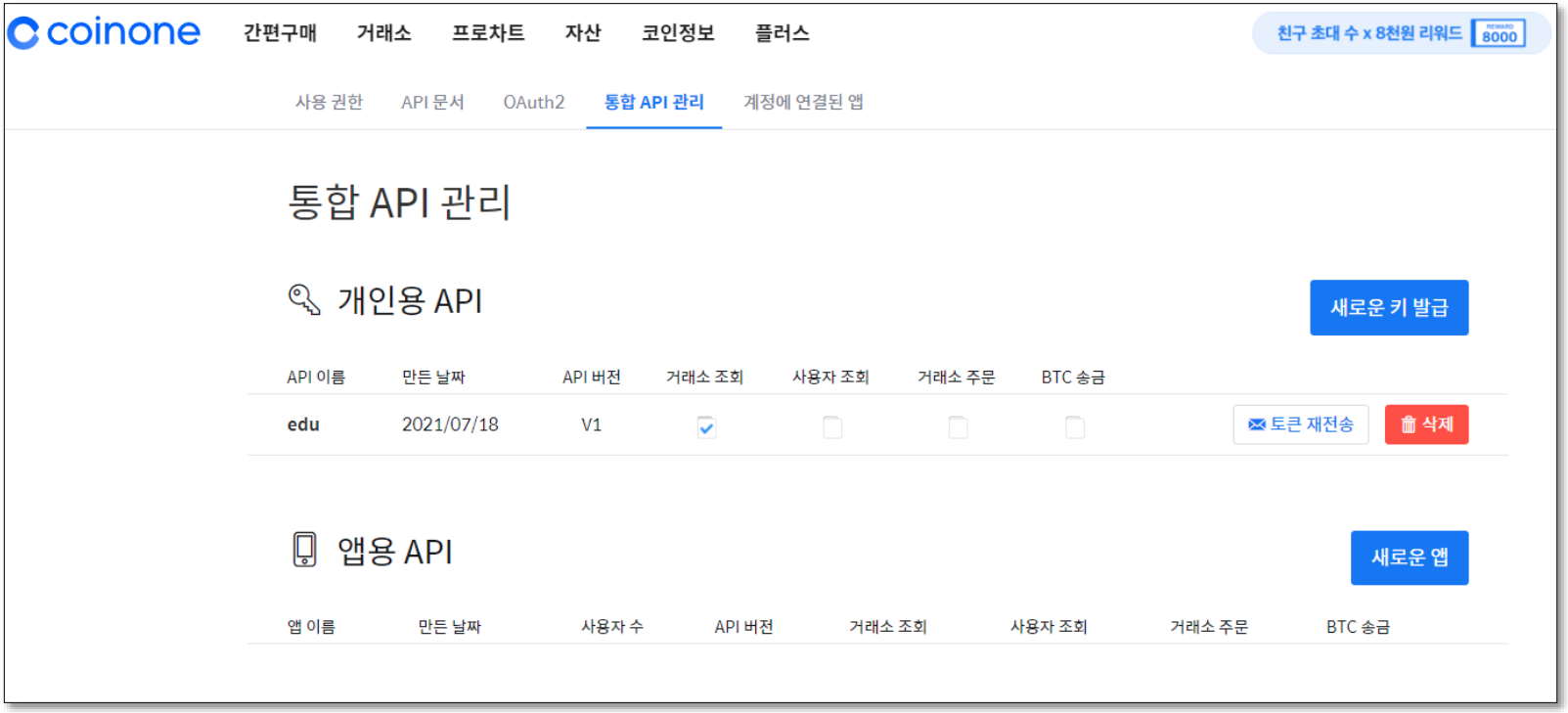


05 Coinone API

[Day 03] GCP 제품군 연계

API key 발급

▶ 목적에 맞는 API key 발급 가능하며, 간편한 사용을 위해서는 V1 권장
▶ API key는 등록된 email으로 전동되며 필요시 [토큰 재전송] 버튼을 눌러 email 재발송 가능



06 GCS 적재

수집 데이터의 적재

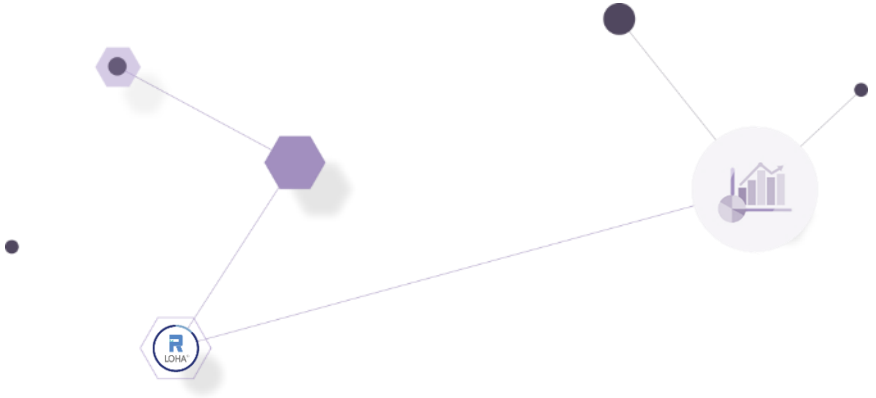
▶ "google-cloud-storage" 라이브러리를 활용한 GCS 접근 및 데이터 적재
▶ 데이터를 시간 또는 별도 항목 기준으로 파일 분리 후 차례대로 적재


```
!pip install google-cloud-storage --user
```

```
from google.cloud import storage
import os

os.environ["GOOGLE_APPLICATION_CREDENTIALS"] = "json 파일 경로"

storage_client = storage.Client()
buckets = list(storage_client.list_buckets())
buckets
```



2) Vertex AI를 활용한 머신러닝 모델 구현

01 Vertex AI 시작하기

[Day 03] GCP 제품군 연계

Vertex AI 개요

▶ GCP의 AI 관련 통합 플랫폼
▶ 코드 없이 이미지, 동영상, 테이블, 텍스트 데이터 학습 가능

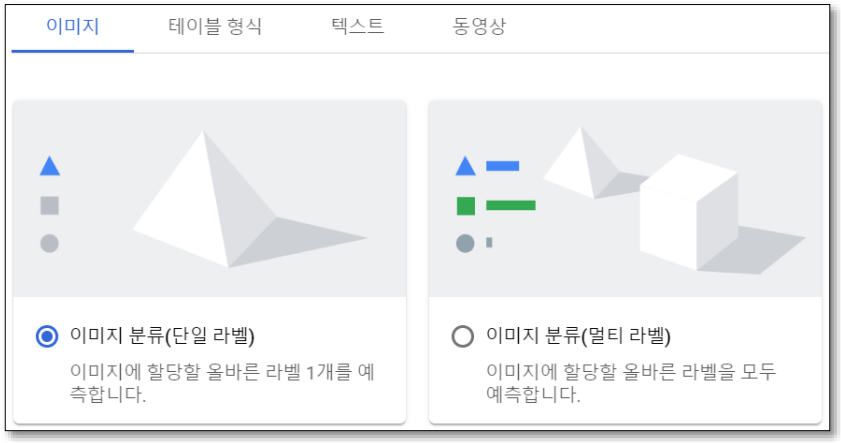
Vertex AI

01 Vertex AI 시작하기

[Day 03] GCP 제품군 연계

데이터 준비하기

- ▶ 업로드 데이터 세트 이름 설정 필수
- ▶ 데이터 특성에 따른 학습 모델의 종류 설정 필수

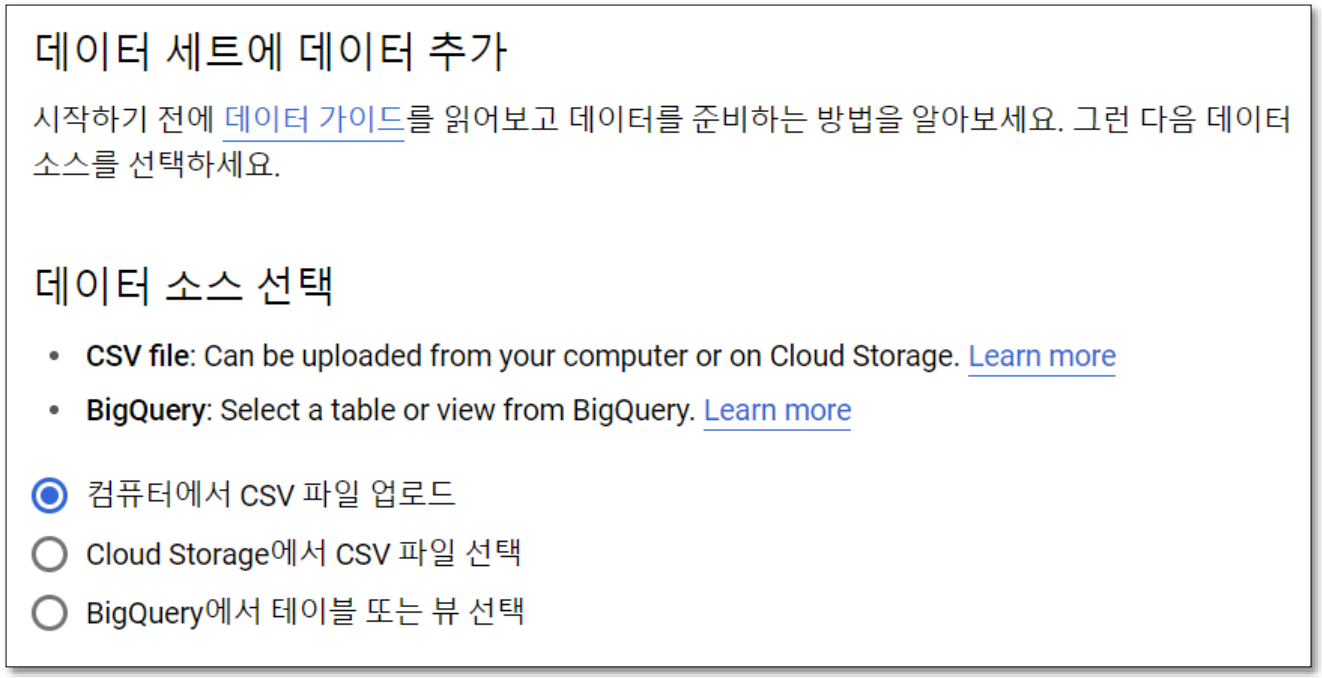


01 Vertex AI 시작하기

[Day 03] GCP 제품군 연계

데이터 준비하기

▶ 로컬 파일 업로드, GCS, BigQuery 중 하나를 선택하여 데이터를 서비스에 업로드(또는 연결)
▶ 파일형식의 경우 csv(comma separated values) 형식만 지원하니 주의

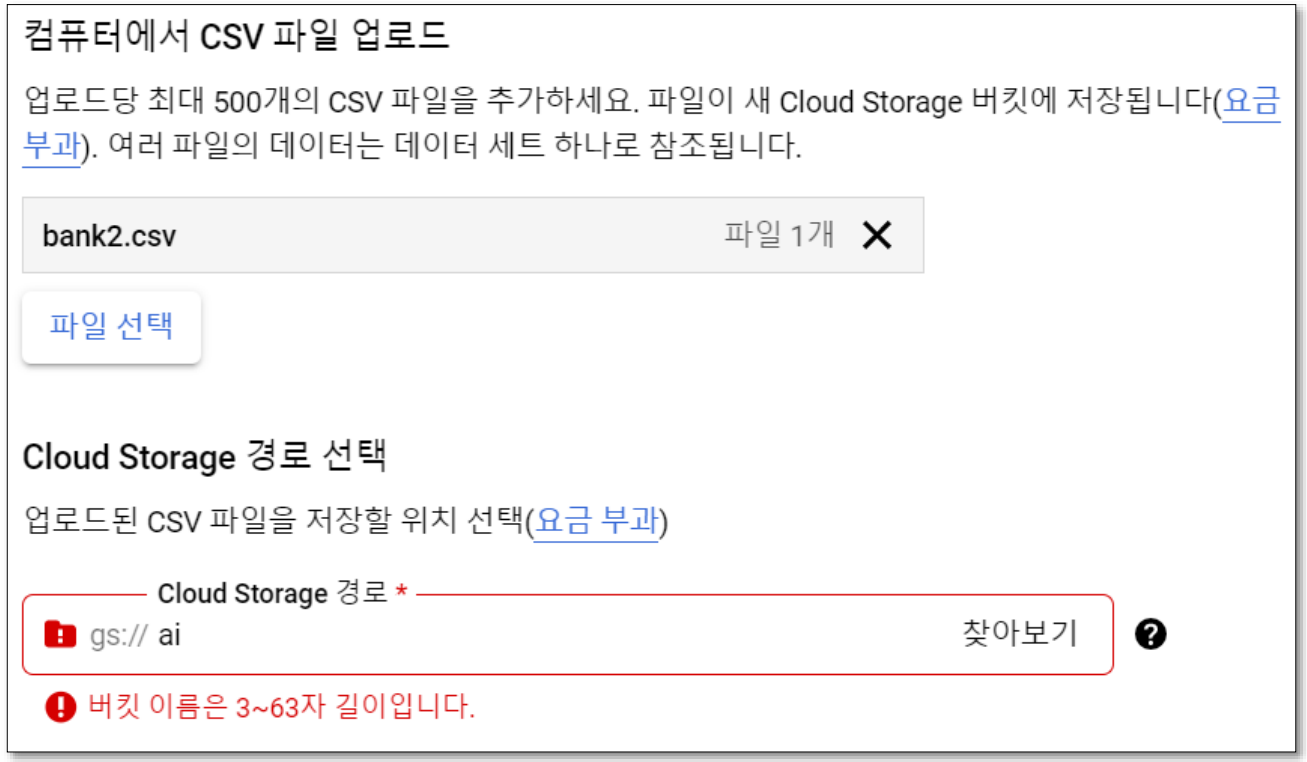


01 Vertex AI 시작하기

[Day 03] GCP 제품군 연계

데이터 준비하기

▶ 파일 업로드의 경우 GCS 경로(버킷) 지정 필수
▶ 별도의 GCS 버킷이 없는 경우 경로 입력칸 우측의 [찾아보기] 버튼을 눌러보면 자동생성 버킷 확인가능

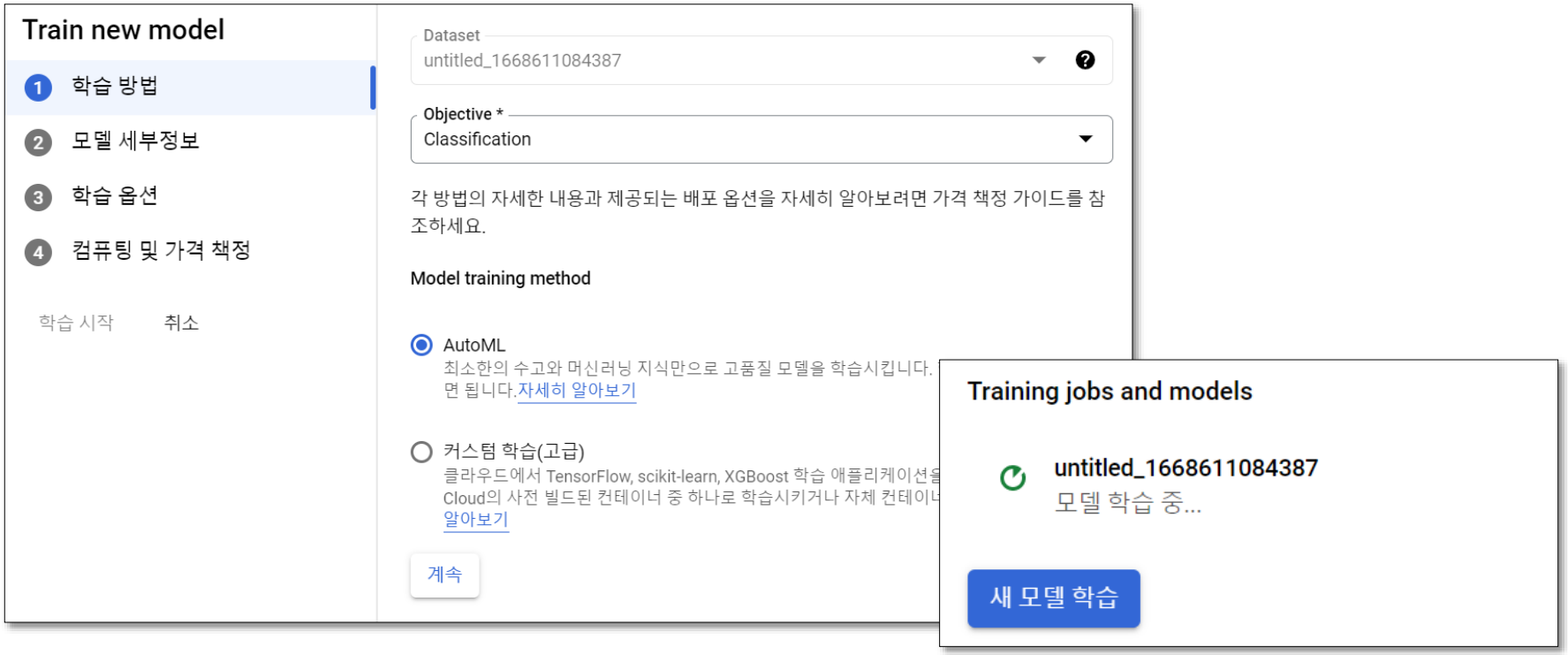


02 모델 학습

[Day 03] GCP 제품군 연계

모델 학습 설정

▶ 학습 방법부터 다양한 설정을 할 수 있으며 완전 자동으로 하려면 "AutoML"을 선택
▶ 학습은 최소 1시간 부터 가능



02 모델 학습

[Day 03] GCP 제품군 연계

모델 학습 확인

▶ 학습 중인 모델을 확인 가능하며 완료시 이메일로 알람 확인 가능