



# 빅쿼리를 활용한 빅데이터 분석 (1)

서진호

# 제 소개를 하자면

- 서진호
- 20년 IT 컨설턴트 (데이터 과학/AI/ML 아키텍트), 마이크로소프트
- Coursera Contents Beta Tester – Google Cloud, TensorFlow, BigQuery
- Machine Learning on Google Cloud Specialization
- From Data to Insights with Google Cloud Specialization
- 책: Data Science at AWS (번역, 올 가을쯤)
- 상세정보
  - Brunch: <https://www.brunch.co.kr/@synabreu>
  - Medium: <https://medium.com/@synabreu>
  - GitHub: <https://github.com/synabreu>



Jinho Seo  
synabreu

# 전체 강의 목표

- 빅데이터에 대한 기술적인 배경과 지식 등을 이해한다.
- 빅데이터 분석에 필요한 구글 클라우드 플랫폼을 이해하고 직접 사용한다.
- 빅쿼리(BigQuery)를 통해 기본적인 SQL 사용법부터 최신 데이터웨어 하우스에서 빅데이터를 분석한다.
- 빅쿼리를 이용해 최신 인공지능 알고리즘 분석한다.

# 타겟 오디언스

- 데이터 분석가, 비즈니스 분석가, 비즈니스 인텔리전스  
프로페셔널이 되려는 분
- 데이터 분석가와 협력하여 구글 클라우드 플랫폼에서  
확장 가능한 데이터 솔루션을 구축할 데이터 엔지니어가  
되려는 분

# 선수 배경 지식

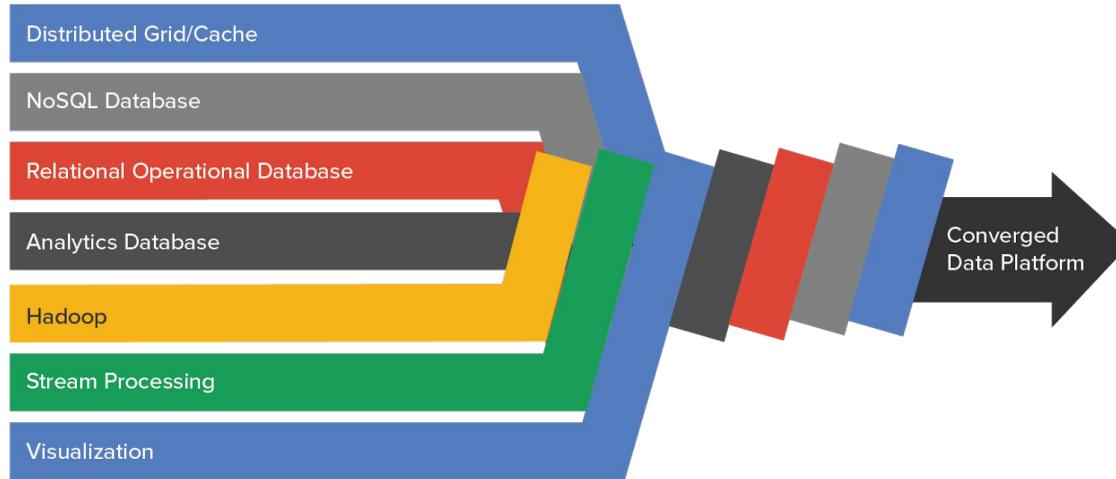
- 데이터 분석가, 비즈니스 분석가, 디지털 마케터, 비즈니스 인텔리전스 프로페셔널이 되려는 분
- 데이터 분석가와 협력하여 구글 클라우드 플랫폼에서 확장 가능한 데이터 솔루션을 구축할 데이터 엔지니어가 되려는 분

# 제 1 강 목표

- 
1. 전체 강의 목표
  2. 구글 클라우드 빅데이터 플랫폼
  3. 구글 빅데이터 서비스 소개

# 구글 클라우드 빅데이터 플랫폼

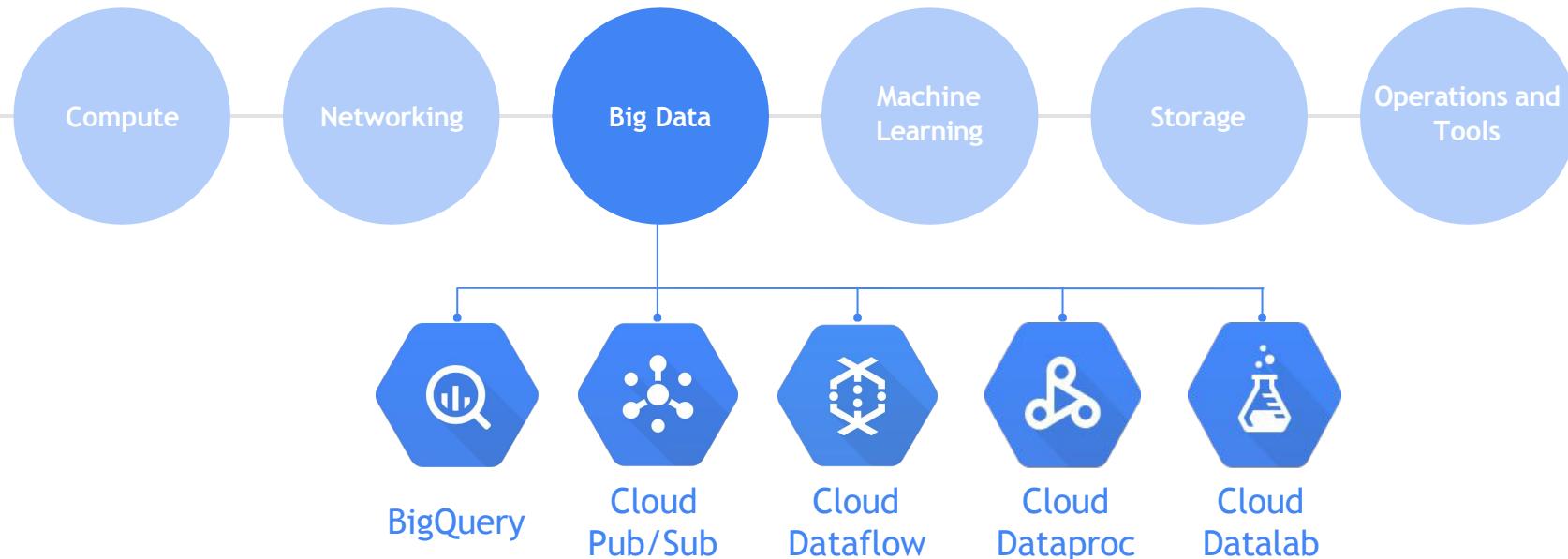
## 통합 위험 감소 및 가치 실현 시간 단축



비즈니스와 사용자 경험을 혁신하는 확장 가능하고 안전하며 안정적인 데이터 기반 애플리케이션을 구축하기 위한 통합 NoOps 클라우드 데이터 플랫폼

- 가치 실현 시간(time-to-value)
- 실시간 애플리케이션
- 머신 러닝을 포함한 혁신에 대한 접근
- 완전 통합성

# 구글 클라우드 플랫폼



# 빅데이터 서비스



BigQuery

데이터베이스 분석;  
초당 100,000  
행으로 데이터  
스트리밍



Pub/Sub

확장 가능하고  
유연한 엔터프라이즈  
메시징



Dataflow

스트림 & 배치  
프로세싱; 통합 및  
간편한 파이프라인



Dataproc

관리형 하둡  
맵리듀스, 스파크,  
피그 및 하이브  
서비스

완전 관리형, NoOps 서비스

# 빅쿼리 (1/2)

- 완전 관리형 데이터 분석 웨어하우스
  - 대규모 데이터셋 (수백 TB)에 대한 실시간에 가까운 대화형 분석 제공
- SQL과 비슷한 문법으로 쿼리
- 성능 및 규모에 대한 관리 없음



# 빅쿼리 (2/2)

- 구글의 완전 관리형 보안 고성능 인프라에서 실행
  - 컴퓨팅과 스토리지는 그 사이에 페타비트의 고속 네트워크로 분리함
  - 사용한 스토리지 및 프로세스 처리 비용만 지불
- 장기 데이터 저장에 대한 자동 할인



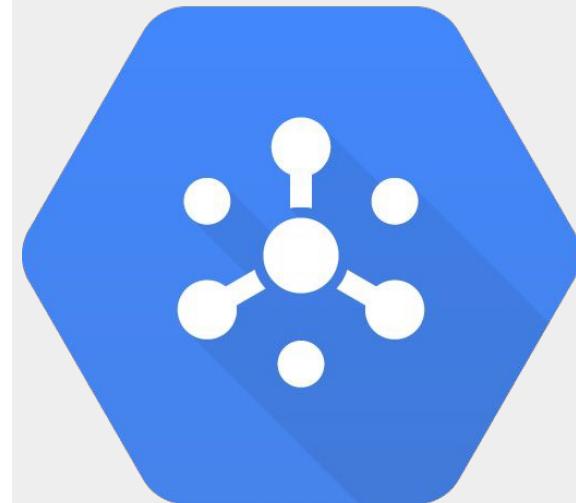
# 구글 클라우드 Pub/Sub (1/2)

- 구글 클라우드 플랫폼과 그 이상을 뛰어 넘은 확장 가능하고 안정적인 메시징
- 다대다(many-to-many) 비동기 메시징 지원
- 오프라인 컨슈머들을 위한 지원 포함
- 증명된 구글 첨단 기술 기반
- 데이터 처리 파이프라인을 위해 Cloud Dataflow와 통합



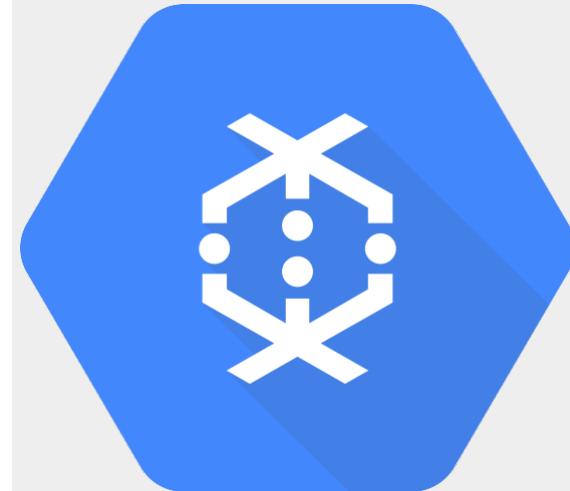
# 구글 클라우드 Pub/Sub (2/2)

- 토픽들에 대한 Push/Pull 구독을 사용
- 모범 사례:
  - 데이터플로, 사물 인터넷(IoT), 마케팅 분석의 데이터 수집을 위한 블록 생성
  - Dataflow 스트리밍을 위한 빠대
  - 클라우드 기반 앱용 푸시 알림
  - (Compute Engine과 App Engine 간 푸시/풀 할 수 있는) 구글 클라우드 플랫폼에서 애플리케이션 연결



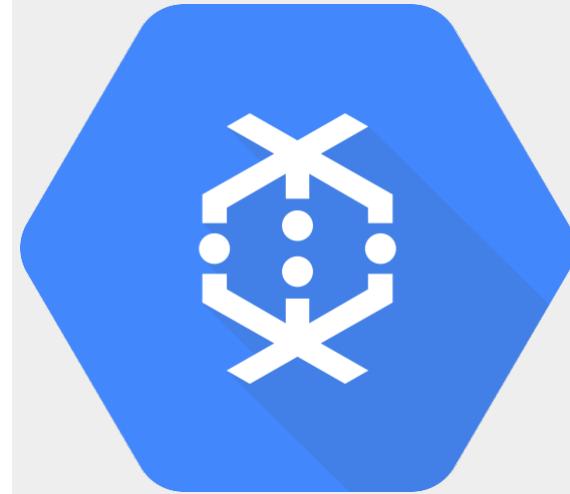
# 구글 클라우드 Dataflow (1/2)

- 확장 가능하고 안정적인 데이터 파이프라인을 실행하기 위한 관리형 서비스
- 코드를 한 번 작성하고 배치 처리 및 스트리밍 가져오기
  - 트랜스폼 기반 프로그래밍 모델
- 클러스터 크기 조정
- 컴퓨터 엔진 인스턴스를 사용하여 데이터 처리



# 구글 클라우드 Dataflow (2/2)

- Cloud Storage, Cloud Pub/Sub, BigQuery, Bigtable 와 같은 GCP 서비스와 통합
- 오픈 소스 [Java](#) 와 [Python](#) SDK
- 모범 사례:
  - 데이터를 이동, 필터링, 강화, 형성하기 위한 ETL (extract/transform/load) 파이프라인
  - [데이터 분석](#)-스트리밍을 사용한 배치 또는 연속 컴퓨테이션
  - [오케스트레이션](#)-외부 서비스를 포함한 서비스를 조정하는 파이프라인 생성



# 구글 클라우드 Dataproc (1/3)

- 구글 클라우드 플랫폼 상에서 Hadoop 과 Spark/Hive/Pig 을 실행하기 위해 빠르고 쉽고 관리형
- 클라우드 서비스 통합 장점
  - 클라우드 스토리지(Cloud Storage)
  - 스택드라이버(Stackdriver)
- initialization action 을 사용하여 클러스터 사용자 지정 및 구성



# 구글 클라우드 Dataproc (2/3)

- 90초 또는 그 이내에 클러스터 생성
- 분 단위로 청구되는 Dataproc 클러스터
  - 배치 처리용 선점형(preemptible) 인스턴스를 사용하여 비용 절감
- 잡(Job)이 실행 중일 때도 클러스터 확장 및 축소
- 개발 도구
  - RESTful API
  - [Google Cloud SDK](#) 와의 통합



# 구글 클라우드 Dataproc (3/3)

- 사용 사례:

- 온프레미스 하둡의 Job을 클라우드로 쉽게 마이그레이션
- 로그 데이터와 같은 클라우드 스토리지에 저장된 데이터를 빠르게 분석 - 2분 이내에 클러스터 생성 후 즉시 삭제
- 빠르게 데이터 마이닝과 분석하는 Spark/Spark SQL 사용
- 분류 알고리즘을 실행하는 머신러닝 라이브러리 사용 (MLlib)



# Vertex AI Workbench (1/3)

- 대규모 데이터 탐험, 변환, 분석, 시각화를 위한 대화형 도구
  - 완전 관리형 노트북(Managed notebooks)
  - 사용자 관리형 노트북(User-managed notebooks)
  - 구글 클라우드 데이터랩은 더 이상 서비스 하지 않음



# Vertex AI Workbench (2/3)

- 콘솔을 사용함으로써 관리형 노트북 인스턴스 생성
- JupyterLab 내에서 BigQuery 테이블의 데이터 쿼리
- 관리형 노트북 실행 예약 기능
- 관리형 노트북 인스턴스에 사용자 지정 컨테이너 추가
- Dataproc 클러스터에서 관리형 노트북 인스턴스 실행



# Vertex AI Workbench (3/3)

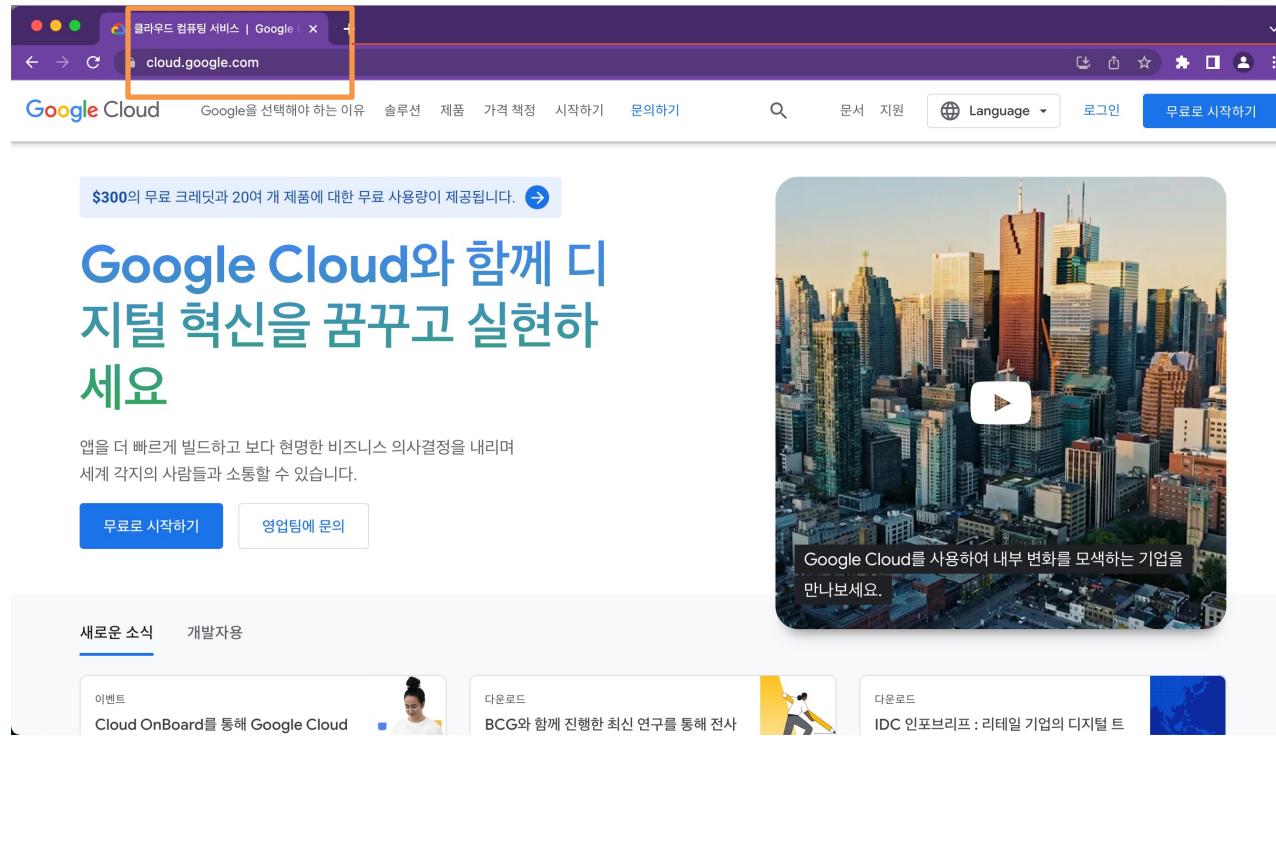
- 콘솔을 사용함으로써 관리형 노트북 인스턴스 생성
- JupyterLab 내에서 BigQuery 테이블의 데이터 쿼리
- 관리형 노트북 실행 예약 기능
- 관리형 노트북 인스턴스에 사용자 지정 컨테이너 추가
- Dataproc 클러스터에서 관리형 노트북 인스턴스 실행



실습 - 구글 클라우드 플랫폼  
회원 가입



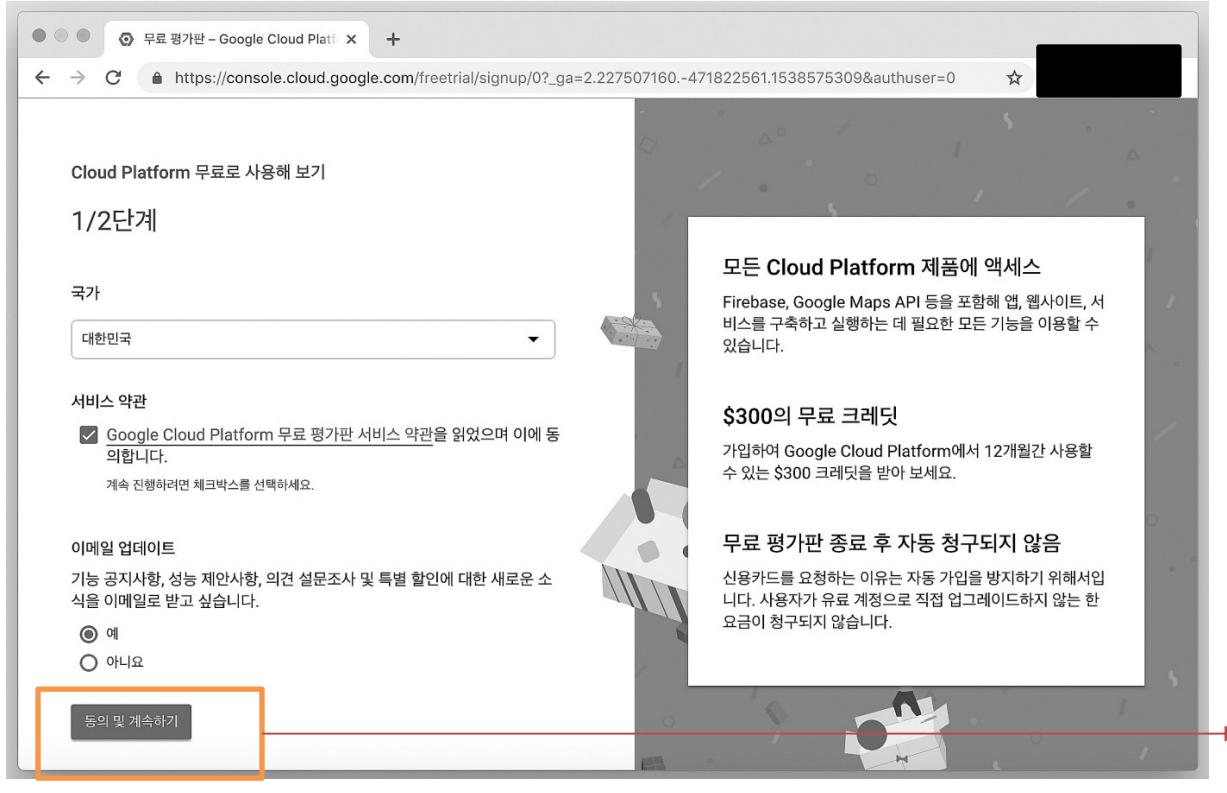
# 실습1 - 구글 클라우드 플랫폼 첫 화면



The screenshot shows the Google Cloud homepage. At the top, there's a navigation bar with icons for back, forward, refresh, and search, followed by the URL 'cloud.google.com' which is highlighted with a red box. To the right of the URL are language selection, login, and '무료로 시작하기' (Start for free) buttons. Below the navigation bar, there's a banner with the text '\$300의 무료 크레딧과 20여 개 제품에 대한 무료 사용량이 제공됩니다.' (Free \$300 credit and free usage limits for over 20 products) and a blue '→' button. The main headline reads 'Google Cloud와 함께 디지털 혁신을 꿈꾸고 실현하세요' (Achieve digital innovation with Google Cloud). Below the headline, there's a paragraph about building faster, more efficient applications and making informed business decisions. There are two buttons: '무료로 시작하기' (Start for free) and '영업팀에 문의' (Contact sales). A large video player in the center features a city skyline at sunset with a play button icon. Below the video, there's a caption: 'Google Cloud를 사용하여 내부 변화를 모색하는 기업을 만나보세요.' (Meet companies exploring internal changes using Google Cloud). At the bottom, there are sections for '새로운 소식' (New news) and '개발자용' (Developer), along with links to events like 'Cloud OnBoard를 통해 Google Cloud' and reports from BCG and IDC.

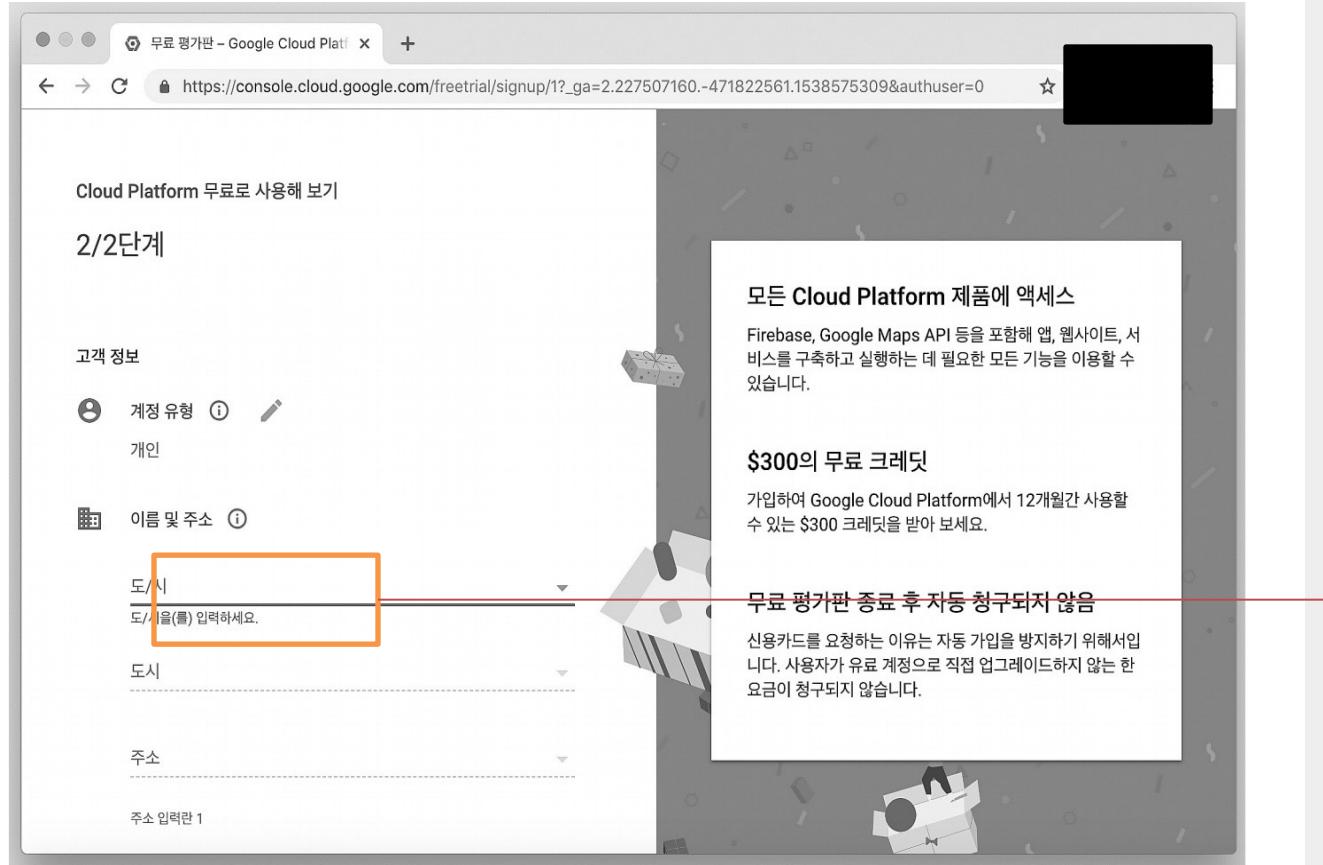
<https://cloud.google.com>  
접속하기

# 실습2 - 구글 클라우드 플랫폼 등록



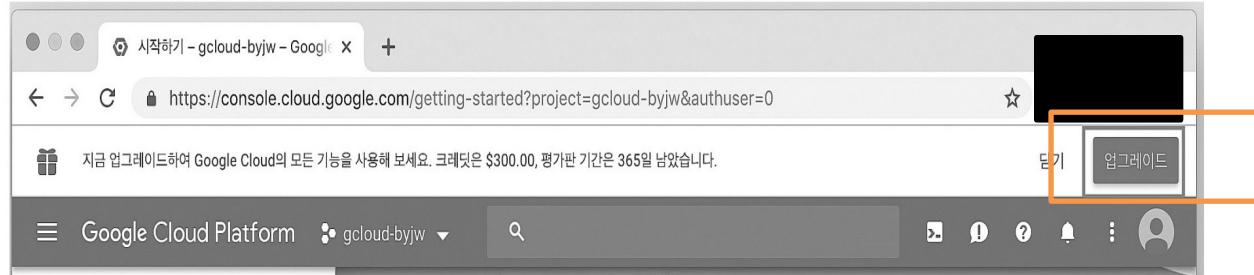
- 국가: 대한민국
- 서비스 약관: 체크
- 이메일 업데이트: 예
- 동의 및 계속하기

# 실습3 - 구글 클라우드 플랫폼 등록

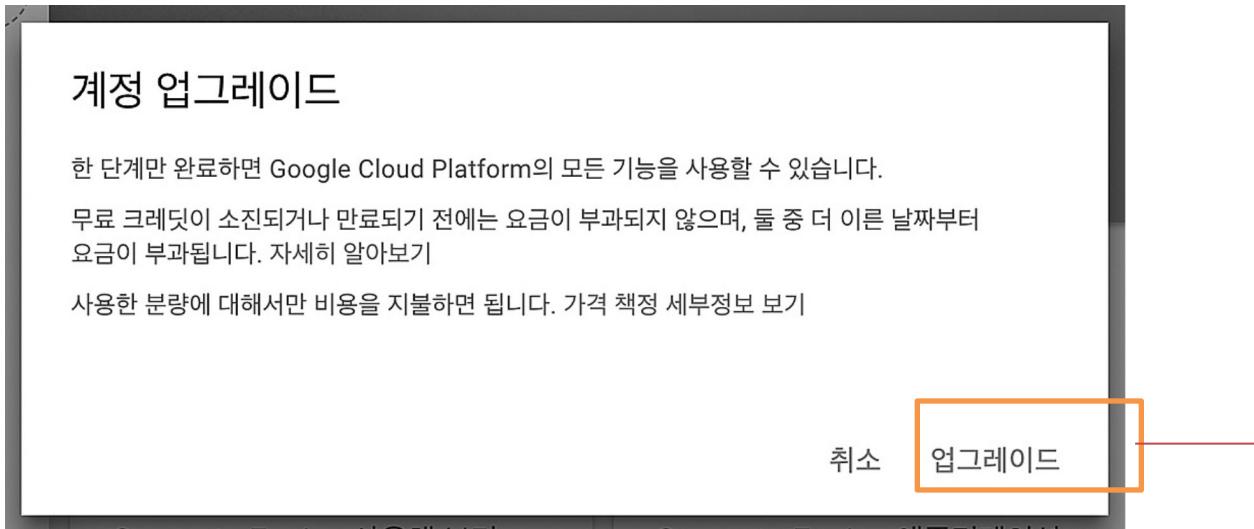


- 계정 유형: 개인
- 이름 및 주소: 본인 주소 입력
- 카드: 해외 Visa/Master 카드 입력

# 실습4 - 구글 클라우드 업그레이드



업그레이드 클릭



업그레이드 클릭