

# 좌충우돌 Data Engineering 학습기

데이터 문맹에서 GCP 데이터 엔지니어 자격증 취득까지

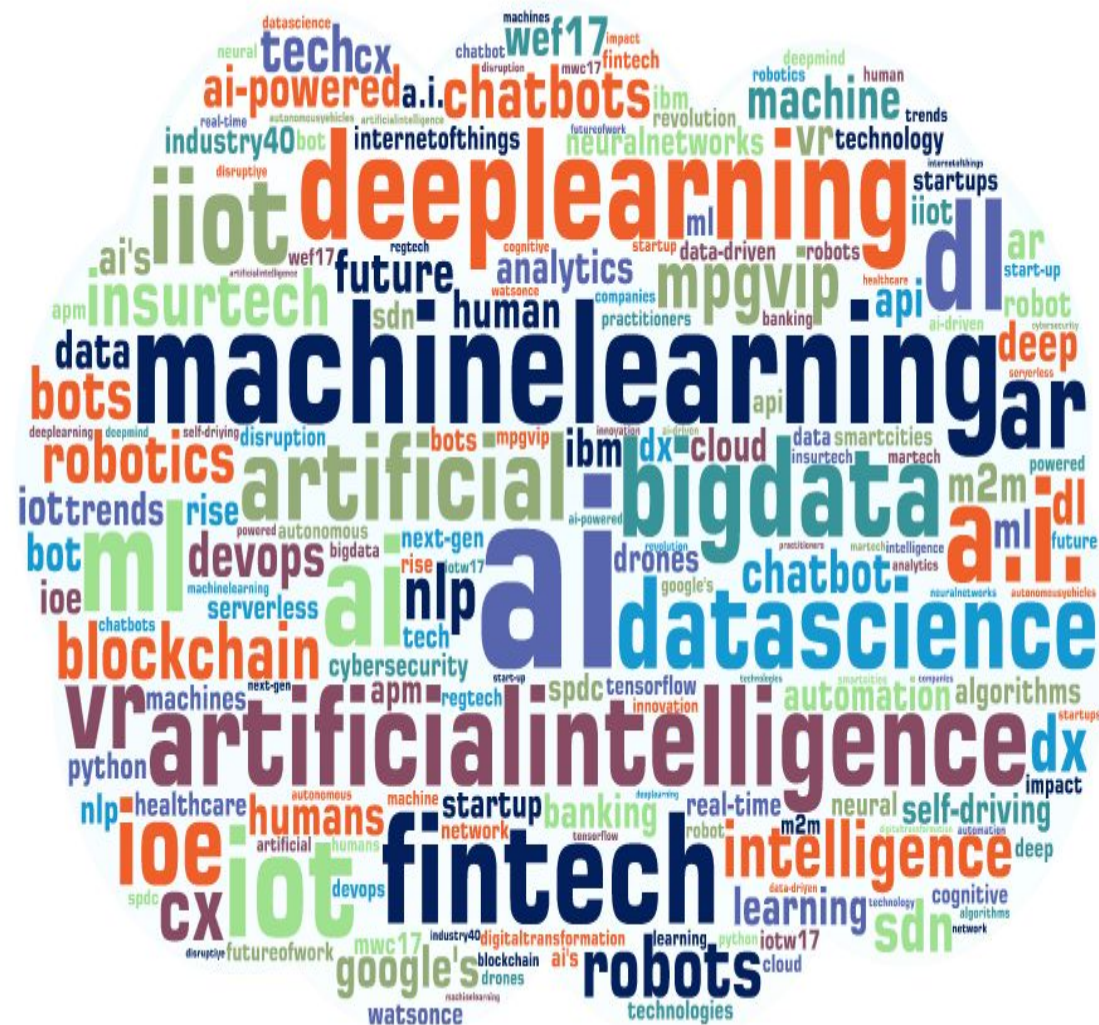
이동민



이 동 민

**Data 공부 꼭 해야하나요?**

**제 경우엔 그랬습니다**



# 나의 현실

SELECT \* FROM CUSTOMER WHERE CID = :cid

간단한 C/R/U/D SQL Query

거의 데이터 문맹 수준

# 빅데이터, 머신러닝 맛보기

더 늦기 전에  $\pi\pi$

# 어떤 맛을 볼래?

Data 2종 세트

Data Science

Data Engineering



# Data Science

# 님 수학 잘함?

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

**Law of Total Probability:** If  $A_1, \dots, A_n$  partitions  $S$ , then  $P(B) = P(B|A_1)P(A_1) + \dots + P(B|A_n)P(A_n)$

**Bayes Theorem:** (con't from above)  $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$

$$E[X] = \sum_k k \cdot P(X = k) \text{ or } \int_{-\infty}^{\infty} x \cdot f(x) dx, \quad \text{Var}(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$$

$$E[aX + b] = aE[X] + b, \quad E[X + Y] = E[X] + E[Y], \quad \text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$X \sim \text{Ber}(p) \quad P(X = 0) = 1 - p, \quad P(X = 1) = p, \quad E[X] = p, \quad \text{Var}(X) = p(1 - p)$$

$$X \sim \text{Bin}(n, p) \quad P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n, \quad E[X] = np, \quad \text{Var}(X) = np(1 - p)$$

$$X \sim \text{Geo}(p) \quad P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots, \quad E[X] = 1/p, \quad \text{Var}(X) = (1 - p)/p^2$$

$$X \sim \text{Poi}(\lambda) \quad P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots, \quad E[X] = \lambda, \quad \text{Var}(X) = \lambda$$

$$X \sim \text{U}(a, b) \quad f(x) = \frac{1}{b-a}, \quad a < x < b, \quad E[X] = \frac{b+a}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

$$X \sim \text{Expo}(\lambda) \quad f(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad E[X] = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

$$X \sim N(\mu, \sigma^2) \quad f(x) = \text{not needed}, \quad E[X] = \mu, \quad \text{Var}(X) = \sigma^2$$

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$\text{Sample statistics: } \bar{X}_n = \frac{1}{n} \sum X_i, \quad S^2 = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$$

For  $X_i$ 's i.i.d.  $\sim N(\mu, \sigma^2)$ ,  $\sigma$  known,  $\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  is a  $100(1 - \alpha)\%$  CI for  $\mu$

$$\text{ANOVA: } S_i^2 = \frac{1}{J-1} \sum (X_{ij} - \bar{X}_{i.})^2, \quad \text{MSTr} = \frac{J}{I-1} \sum (\bar{X}_{i.} - \bar{X}_{..})^2, \quad \text{MSE} = \frac{1}{I} \sum S_i^2, \quad F = \frac{\text{MSTr}}{\text{MSE}}$$



어려운 맛

무서운 맛

# Data Engineering 맛보기

Data Science는 다음 기회에.. 아디오스..

# Data Engineering 란?

- 다양한 형태의 데이터를 수집, 변환, 적재하는 시스템을 설계, 구현, 운영
- 많은 양의 데이터를 처리 할 수 있도록 확장성, 유연성, 효율성, 보안, 모니터링을 제공하는 데이터 프로세스 시스템 구축
- 만들어진 머신러닝 모델을 활용하고 학습시키고 배포

# Data Engineering 의 진화

## BEFORE

- 데이터 파이프라인 구축, 운영에 필요한 인프라까지 관리
- 넓은 학습 범위, 가파른 Learning curve
- 데이터 분석을 위한 서포팅 역할에 중점

## NOW

- Cloud Engineer + Data Engineer = Cloud Data Engineer
- Serverless, Auto Scailing, Managed Services
- 다양한 서비스를 이용한 Data-flow 구축하여 분석 목적뿐 아니라 여러 운영 환경에 데이터 제공

# Cloud + Data Engineering 학습 기왕 공부하는거 자격증도 취득하자



업계 1위 AWS, 업계 2위 Azure도 있는데

# Why Google Cloud Platform?



2018년 11월 AWS Korea region 장애

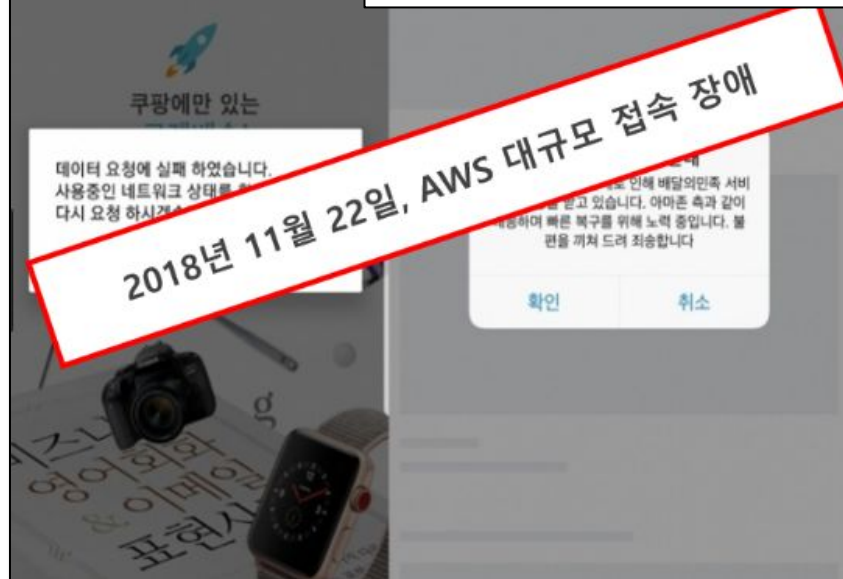
Multi-vendor 클라우드 구성의  
중요성

502 Bad Gateway

nginx

AWS코리아, 장애 고객에 11월 EC2 요금중 10% 환불  
조치

2018.12.11 10:01:34 / 백지영 jvp@ddaily.co.kr



서비스 정기 점검 안내

한국 아마존 서버 장애로 인해  
배달의민족 서비스도 영향을 받고  
있습니다. 아마존 측과 같이 대응하며  
빠른 복구를 위해 노력 중입니다. 불편을  
끼쳐 드려 죄송합니다

취소

확인

무이한행제를

**AWS + Azure ?**

**AWS + G C P ?**

영어는 기본이고 제 2 외국어로 중국어 할래 스페인어 할래 ?

이런 느낌 ..

희소성 높아보이는 GCP(스페인어 ?) 할게요 .





세계에서 2번째로 많은양의 데이터를 검색하는 Youtube  
Daily uploads in 2017 > 1 Petabyte(13.3 years of HDTV)

Google 꺼임

그런 Google이 사용하는 Cloud Infra



**세계에서 가장 많은양의 데이터를 검색하는 검색엔진**

**daily processing in 2008 > 20 Petabytes**

갓 구글..

그런 Google이 사용하는 Cloud Infra



**강력한 기계학습 라이브러리**  
**Google이 만들어서 오픈소스로 공개**  
**갓 구글..**

8주 후 ..

# 취 to the 득!!

## Google Cloud Certified



## Professional Data Engineer



**어떻게 준비했나요?**

# Outline 기반으로 학습계획 세우기

<https://cloud.google.com/certification/guides/data-engineer-2/>

- Section 1: 데이터 처리 시스템 디자인
- Section 2: 데이터 처리 시스템 구축 및 운영
- Section 3: 머신러닝 모델 운영
- Section 4: 솔루션 품질 보장

지난 3월 29일 시험 Outline 및 문제 유형 변경  
머신러닝, 모니터링 비중 높아짐

다양한 동영상 강의 플랫폼 활용

**coursera**



**Linux Academy**





# Data Engineering on Google Cloud Platform Specialization

(<https://www.coursera.org/specializations/gcp-data-machine-learning?>)

- Google Cloud 에서 만든 공식 강의
- Data Engineering 자체에 초점이 맞춰진 강의
- 총 5개 Course
  - Google Cloud Platform Big Data and Machine Learning Fundamentals
  - Leveraging Unstructured Data with Cloud Dataproc on Google Cloud Platform
  - Serverless Data Analysis with Google BigQuery and Cloud Dataflow
  - Serverless Machine Learning with Tensorflow on Google Cloud Platform
  - Building Resilient Streaming Systems on Google Cloud Platform



# Preparing for the Google Cloud Professional Data Engineer

(<https://www.coursera.org/learn/preparing-cloud-professional-data-engineer-exam>)

- Google Cloud 에서 만든 공식 강의
- GCP Data Engineering Exam 준비에 초점이 맞춰진 강의
- 약 13시간 분량, 모의시험 제공
- 요약 정리 강의. 이것만으로도 시험 준비하기엔 조금 빈약함.



Linux Academy

# Google Cloud Certified Professional Data Engineer

(<https://linuxacademy.com/google-cloud-platform/training/course/name/google-cloud-data-engineer>)

- 시험 Outline을 모두 커버하는 이론 + 실습 강의  
3/29 변경 전 시험 Outline 인 것이 함정
- 자체 리눅스 클라우드 서버 제공  
Google Cloud SDK 설치부터 Terminal 환경에서 다양한 GCP command-line 실습 가능
- 모의고사 제공
- Google Cloud Platform SandBox 제공 (4시간 뒤 초기화)



# Linux Academy

## Flash Card 기능 제공

Studying

**Google Cloud Certified Associate Cloud Engineer (108 cards)**  
by: Ben

Total Cards: 108  
Cards Learned: 0

**108**  
New

**0**  
Familiar

**0**  
Mastering

**0**  
Learned

Key-value pairs of configuration data that are accessible from code running in a Cloud Function.

Wrong

Right

Memorized

Reveal Back

Rate this deck of flashcards:

Exceeded my Expectations

Room for Improvement

시험 전에 30분정도 훑어보기에 좋다.



Credit 걱정없이 GCP를 마음껏 사용해보자

Google Cloud Training

Home

Catalog

My Learning

Help Center

We give you temporary credentials to Google Cloud Platform, so you can learn the cloud using the real thing – no simulations. From 30-minute individual labs to multi-day courses, from introductory level to expert, instructor-led or self-paced, with topics like machine learning, security, infrastructure, app dev, and more, we've got you covered.

In Progress

QUEST

Challenge: GCP Architecture

Advanced

QUEST

Kubernetes Solutions

Expert

QUEST

Google Cloud Solutions I: Scaling Your Infrastructure

Expert

QUEST

Cloud Architecture

Fundamental

QUEST

Google Cloud Solutions II: Data and Machine Learning

Expert

Your Favorites

HANDS-ON LAB

Continuous Delivery with Jenkins in Kubernetes Engine

Expert

★★★★☆

HANDS-ON LAB

Setting up Jenkins on Kubernetes Engine

Advanced

★★★★☆

HANDS-ON LAB

HTTP Load Balancer

Advanced

★★★★☆

HANDS-ON LAB

Creating Cross-region Load Balancing

Advanced

★★★★☆

HANDS-ON LAB

Internal Load Balancer

Fundamental

★★★★☆

## Data Engineering

Advanced 5 Steps 4h 51m 37 Credits

This advanced-level quest is unique amongst the other Qwiklabs offerings. The labs have been curated to give IT professionals hands-on practice with topics and services that appear in the [Google Cloud Certified Professional Data Engineer Certification](#). From Big Query, to Dataproc, to Tensorflow, this quest is composed of specific labs that will put your GCP data engineering knowledge to the test. Be aware that while practice with these labs will increase your skills and abilities, you will need other preparation too. The exam is quite challenging and external studying, experience, and/or background in cloud data engineering is recommended.

Data Machine Learning Business Transformation

### Prerequisites

This Quest requires proficiency with GCP Services, particularly those relating to working with large datasets. It is recommended that the student have at least earned a Badge by completing the hands-on labs in the [Baseline: Data, ML, and AI](#) and/or the [GCP Essentials](#) Quests before beginning. Additional lab experience with the [Scientific Data Processing](#) and the [Machine Learning APIs](#) Quests will be useful.

### Quest Outline

✓

HANDS-ON LAB

Creating a Data Transformation Pipeline with Cloud Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. In this lab you will explore the Cloud Dataprep UI to build a data transformation pipeline.

★★★★☆ 1h 15m Advanced 7 Credits

OR

HANDS-ON LAB

Run a Big Data Text Processing Pipeline in Cloud Dataflow

In this lab you will use Google Cloud Dataflow to create a Maven project with the Cloud Dataflow SDK, and run a distributed word count pipeline using the Google Cloud Platform Console.

★★★★☆ 40m Advanced 7 Credits

✓

HANDS-ON LAB

Building an IoT Analytics Pipeline on Google Cloud Platform

### Quest Complete!

Congrats! You completed this quest and earned a badge. Become a cloud expert and start another.

다양한 퀘스트와 Hands-on lab을 제공

## 추천 퀘스트 목록

- [Data Engineering](#)
- [Google Cloud Solutions II: Data and Machine Learning](#)
- [NCAA® March Madness®: Bracketology with Google Cloud](#)
- [Machine Learning APIs](#)
- [Scientific Data Processing](#)

# 학습자료가 너무 비싸요

강의 3개(각 \$49) + 퀵랩(\$55) = \$202(monthly)



# Coursera Financial Aid 신청하기

(<https://reoim.tistory.com/entry/Coursera-%EC%9C%A0%EB%A3%8C%EA%B0%95%EC%9D%98-financial-aid-%EC%8B%A0%EC%B2%AD%ED%95%98%EA%B8%B0>)

# Google Cloud 강의 한달 무료 이용 쿠폰




## Start Right Now

Get ahead with a select Google Cloud Specialization on Coursera. Your first month of learning is free.

Discount applied at checkout.  
One-time use only.  
Offer valid until 01/01/20, while supplies last.



### Google Cloud Specializations




1 MONTH FREE

Advanced Machine Learning with TensorFlow o...

Google Cloud

SPECIALIZATION (5 COURSES)




1 MONTH FREE

Machine Learning with TensorFlow on Google Cloud Platform

Google Cloud

SPECIALIZATION (5 COURSES)




1 MONTH FREE

Architecting with Google Cloud Platform

Google Cloud

SPECIALIZATION (6 COURSES)



1 MONTH FREE

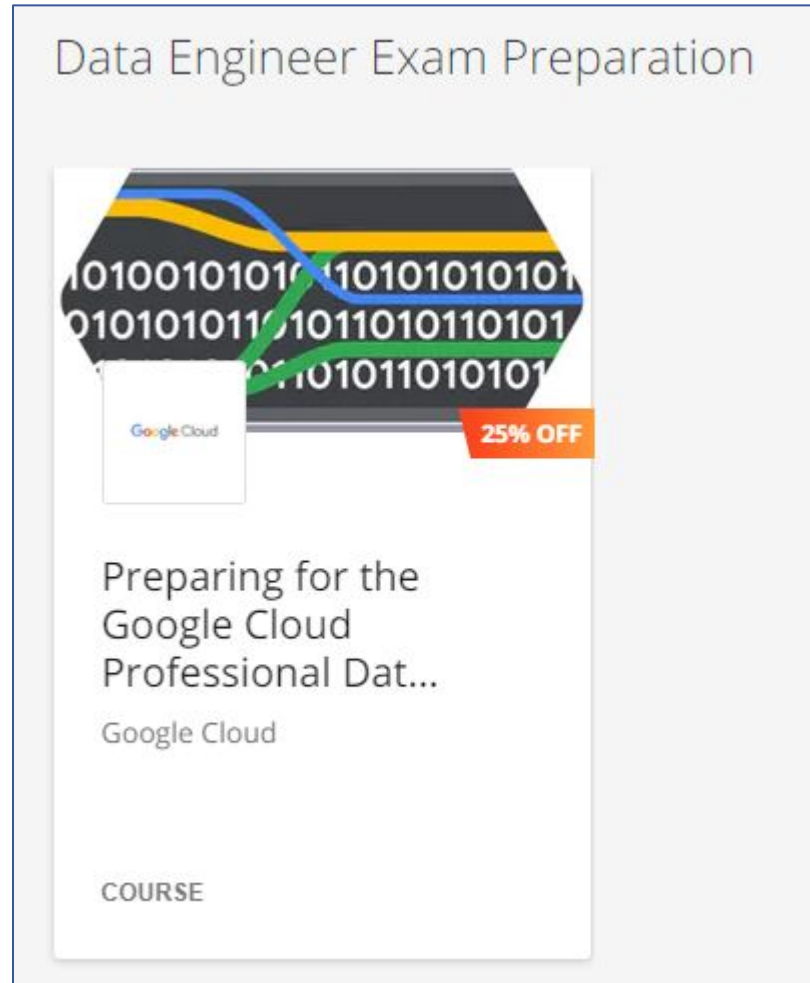
Data Engineering on Google Cloud Platform

Google Cloud

SPECIALIZATION (5 COURSES)

<http://bit.ly/2REhaSB>

# Data Engineer 시험 준비 강의 25% 할인 쿠폰



<http://bit.ly/2WgYtHg>

# 퀵랩 Data Engineering 퀘스트 무료 쿠폰

## Data Engineering

Advanced 5 Steps 4h 51m 37 Credits

This advanced-level quest is unique amongst the other Qwiklabs offerings. The labs have been curated to give IT professionals hands-on practice with topics and services that appear in the [Google Cloud Certified Professional Data Engineer Certification](#). From Big Query, to Dataproc, to Tensorflow, this quest is composed of specific labs that will put your GCP data engineering knowledge to the test. Be aware that while practice with these labs will increase your skills and abilities, you will need other preparation too. The exam is quite challenging and external studying, experience, and/or background in cloud data engineering is recommended.

Data Machine Learning Business Transformation

### Prerequisites

This Quest requires proficiency with GCP Services, particularly those relating to working with large datasets. It is recommended that the student have at least earned a Badge by completing the hands-on labs in the [Baseline: Data, ML, and AI](#) and/or the [GCP Essentials](#) Quests before beginning. Additional lab experience with the [Scientific Data Processing](#) and the [Machine Learning APIs](#) Quests will be useful.

### Quest Outline

HANDS-ON LAB

#### Creating a Data Transformation Pipeline with Cloud Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. In this lab you will explore the Cloud Dataprep UI to build a data transformation pipeline.

★★★★☆ 1h 15m Advanced 7 Credits

### Enroll Now

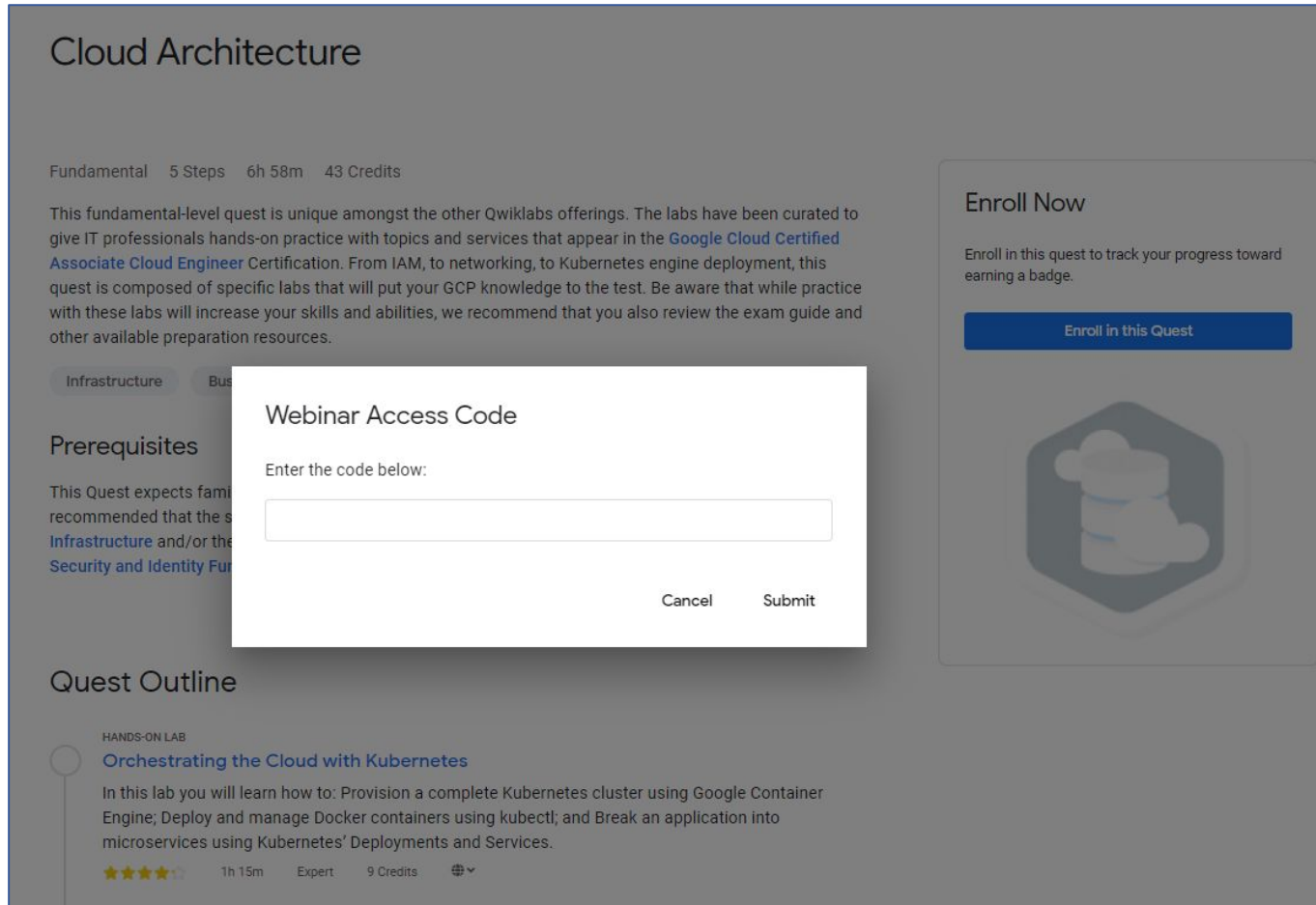
Enroll in this quest to track your progress toward earning a badge.

Enroll in this Quest



<http://bit.ly/2umWmpp>

# 퀵랩 1달 쿠폰 + Cloud Architecture 퀘스트



The screenshot shows the Google Cloud Architecture Quest page. A modal window titled "Webinar Access Code" is open, prompting the user to "Enter the code below:" with a text input field and "Cancel" and "Submit" buttons. The background page includes the following sections:

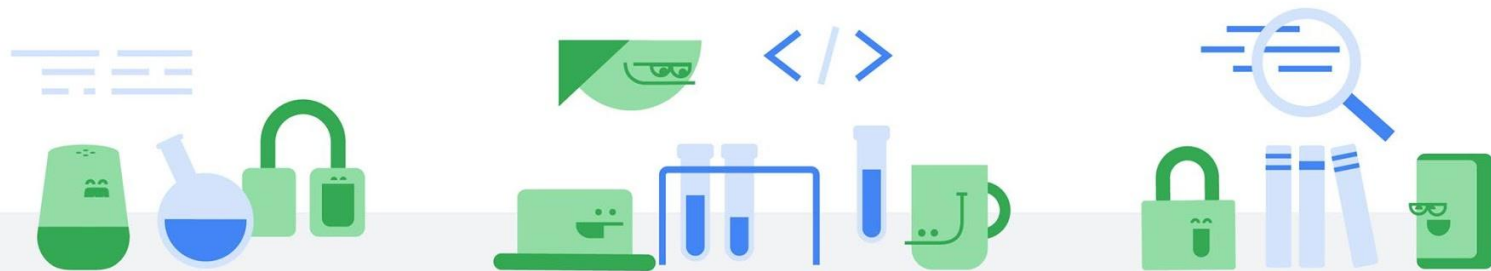
- Cloud Architecture**
- Fundamental** 5 Steps 6h 58m 43 Credits
- Description: "This fundamental-level quest is unique amongst the other Qwiklabs offerings. The labs have been curated to give IT professionals hands-on practice with topics and services that appear in the **Google Cloud Certified Associate Cloud Engineer** Certification. From IAM, to networking, to Kubernetes engine deployment, this quest is composed of specific labs that will put your GCP knowledge to the test. Be aware that while practice with these labs will increase your skills and abilities, we recommend that you also review the exam guide and other available preparation resources."
- Prerequisites**: "This Quest expects familiarity with Google Cloud Platform. We recommend that the student has completed the **Infrastructure** and/or the **Security and Identity Fundamentals** quests."
- Quest Outline**:
  - HANDS-ON LAB**
  - Orchestrating the Cloud with Kubernetes**
  - Description: "In this lab you will learn how to: Provision a complete Kubernetes cluster using Google Container Engine; Deploy and manage Docker containers using kubectl; and Break an application into microservices using Kubernetes' Deployments and Services."
  - Rating: 4 stars, 1h 15m, Expert, 9 Credits

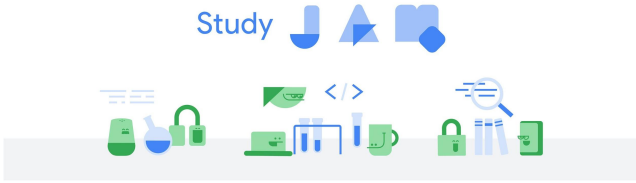
<http://bit.ly/cj-ace>

Access Code  
**3b-onair-94**

# Google Cloud 스테디잼

Study J A B





# Google 에서 지원하는 스터디 모임

그룹장 1, 그룹원 4명 이상 스터디 그룹 신청

퀵랩 1달 이용권, Coursera 강의등 학습자료 무료제공

쿠버네티스, 머신러닝 스터디잼

저는 2가지 모두 진행 하였습니다.

2달 Qwiklabs 무료이용



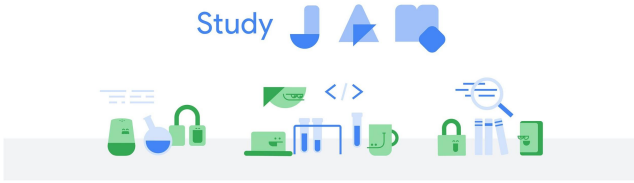
Study



# 구글 스터디 잼 상반기 일정







수료하면 선물도 줍니다.



# 자격증 취득. 드디어 끝!?

끝이 아닌 시작

**Next 19에서 발표된 내용만 122+ 개**

**Data 직접 관련 내용만 9개**

기술발전 속도가 내 학습 속도보다 빠르다

# 빠르게 발전하는 데이터 기술

## 점점 더 다양해지는 클라우드 서비스

폭 넓은 시야와 빠른 학습 능력, 유연성을 위한 노오력

엔지니어로 계속 먹고 살려면 ..  $\pi\pi$

# 정리하자면

- Cloud + Data Engineering 맛보기검 자격증 공부 추천
- 시험 OutLine으로 계획 세우기
- 동영상 강의 활용하기
- Qwiklabs 활용하기
- 스터디잼등 스터디 그룹으로 공부하기
- 엔지니어는 은퇴할때까지 평생 공부  $\pi\pi$

클라우드, 빅데이터, 머신러닝 공부할 게 너무 많지만  
이런 기술의 황금기에 엔지니어로 일하고 있어 행복합니다

**감사합니다.**