데이터 베이스와 R

박찬엽

2017년 6월 27일

목차

과제 확인

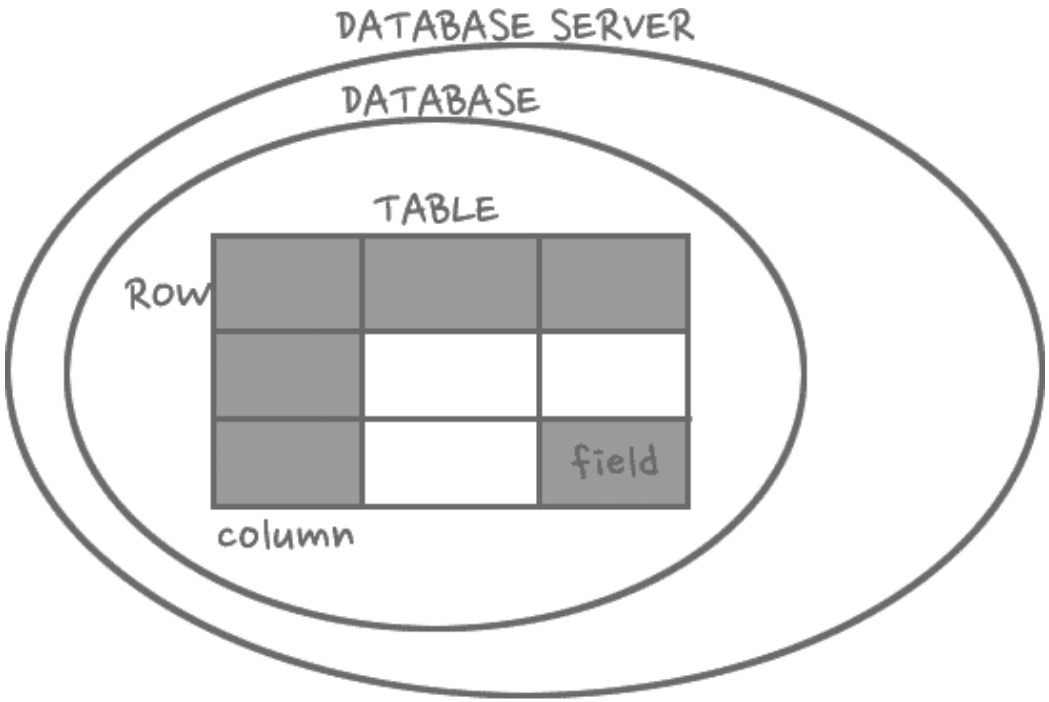
데이터베이스

데이터란

단순한 관찰이나 측정 등의 수단을 통해 현실 세계로부터 수집된 사실이나 값

의미있게 사용하기 위해서 구조화가 필요함
* 구조화: 체계적으로 조직하는 것

DBMS



* 이미지 출처: 생활코딩 MySQL 수업

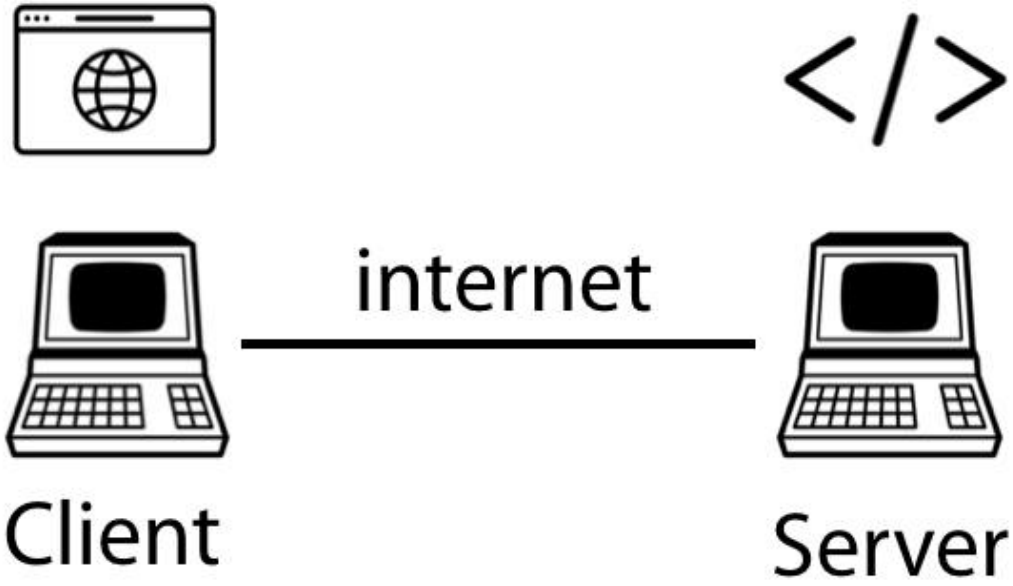
데이터베이스란

엑셀	DBMS
파일	데이터베이스
시트	테이블

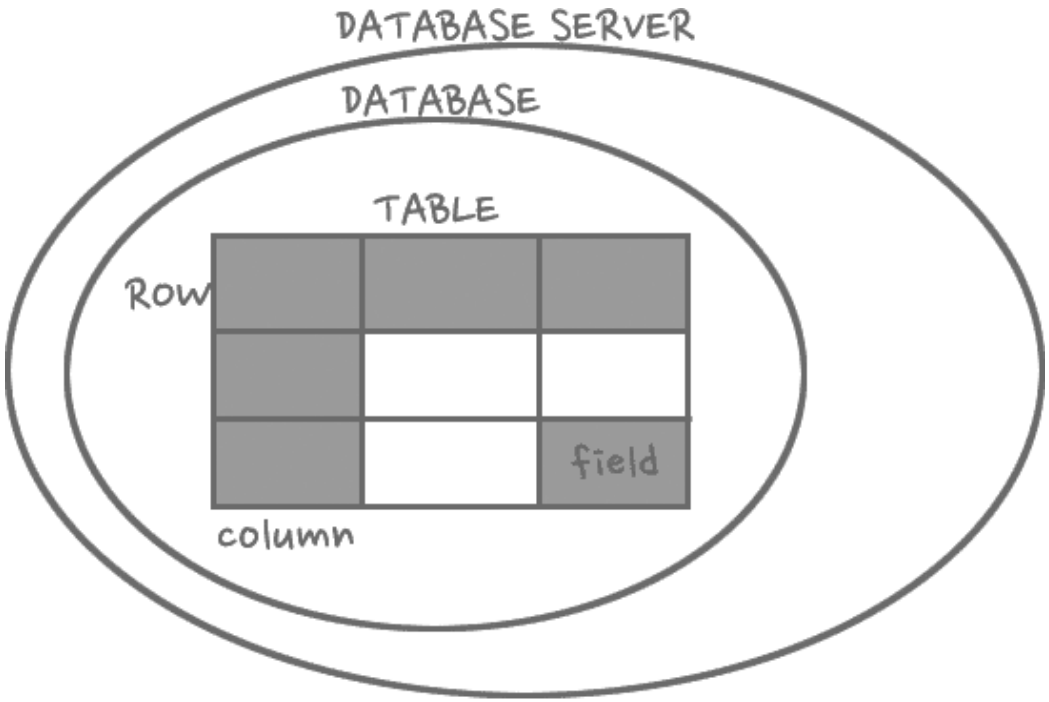
데이터베이스 클라이언트

- 대표적인 클라이언트
  - MySQL monitor
  - PHPmyAdmin
  - Navicat
  - HeidiSQL

서버와 클라이언트



테이블



* 이미지 출처: 생활코딩 MySQL 수업

data.frame

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

SQL

Structured Query Language

구조적 데이터 요청 언어

DBI

```
## Loading required package: devtools
```

```
## Loading required package: DBI
```

```
## Loading required package: RSQLite
```

SQLite

SQLite

SQLite is a self-contained, high-reliability, embedded, full-featured, public-domain, SQL database engine.

파일 하나로 구성하는 작고, 무료인 sql db

DBI

```
library(DBI)
library(RSQLite)
con <- dbConnect(RSQLite::SQLite(), dbname="class2.sqlite")

dbListTables(con)

dbWriteTable(con, "mtcars", mtcars, overwrite=T)
dbListTables(con)

dbReadTable(con, "mtcars")

dbRemoveTable(con,"mtcars")
dbListTables(con)

system.time(dbWriteTable(con, "member", "./recomen/membership.csv",row.names=F))
```

데이터

데이터 공유

Leek group에서 소개하는 데이터 공유 가이드

- 원시 데이터
- 정제후 데이터
- 코드북
- 변수 작성법
- 재현성

원시 데이터

최초 획득한 당시 그대로의 데이터

- 어떤 식으로든 수정을 가하지 않은 상태
- 수정을 가하는 과정을 함께 기록함으로써 신뢰성 확보
- 위 두 가지가 없는 경우 상황을 상상해야 함

정제후 데이터

해들리 위컴이 설명한 tidy data의 요건에 맞게 가공하여 데이터를 쉽게 다룰 수 있게 만든 상태

- 측정하는 각 변수는 하나의 열에 있어야 함
- 측정하는 각 관찰은 하나의 행에 있어야 함
- 각 종류의 변수에 대해 각 하나의 테이블이 있어야 함
- 여러 개의 테이블이있는 경우 테이블에 합치기 위한 기준 열을 포함해야 함

코드북

데이터셋에 대해 필요한 설명을 담은 문서

- 정제후 데이터에 대해 추가적으로 필요한 설명이나 정보(단위 등)
- 정제 과정에서 사용한 방법의 설명과 사용한 이유
- 데이터가 사용된 분석에 대한 정보

데이터 원본

확보할 당시의 원시 데이터나, 항상 최신 상태를 유지하여 신뢰할 수 있는 데이터

- 커뮤니케이션 비용 감소
- 의사결정 및 활동의 기준
- 가공된 데이터의 신뢰성 확보

데이터 훑어보기

- head: 최초 6행의 데이터를 보여줌(행갯수 조절 가능)
- tail: 마지막 6행의 데이터를 보여줌(행갯수 조절 가능)
- summary: 각 컬럼의 자료형과 숫자라면 대표값을 함께 보여줌
- str: 각 컬럼의 자료형과 초기 값을 보여줌
- length: 데이터의 길이 출력(vector)
- nrow: 행 갯수 출력(data.frame)
- is.na: NA 인지 확인
- complete.cases: 값이 모두 있는지 행단위로 검사
- tibble: 최근 기법으로 재구성된 data.frame

추천 데이터 - chennel

- cusID : 5자리 숫자조합으로 구성된 고객ID
- chennel: 접속 체널
- useCnt : 사용횟수(건)

```
##      cusID          chennel             useCnt      
##  Min.   :    7   Length:8824        Min.   :  1.00  
##  1st Qu.: 6107   Class :character   1st Qu.:  2.00  
##  Median : 9506   Mode  :character   Median :  7.00  
##  Mean   : 9835                      Mean   : 13.57  
##  3rd Qu.:13812                      3rd Qu.: 19.00  
##  Max.   :19382                      Max.   :240.00
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    8824 obs. of  3 variables:
##  $ cusID  : int  7 14 42 74 74 94 112 112 122 123 ...
##  $ chennel: chr  "A_MOBILE/APP" "A_MOBILE/APP" "B_MOBILE/APP" "A_MOBILE/APP" ...
##  $ useCnt : int  4 1 23 1 30 14 16 1 27 10 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - competitor

- cusID : 5자리 숫자조합으로 구성된 고객ID
- partner : 제휴사
- competitor: 경쟁사
- useDate : 이용년월(YYYYDD)

```
##     cusID              partner            competitor          useDate
##  Length:28159       Length:28159       Length:28159       Min.   :201501
##  Class :character   Class :character   Class :character   1st Qu.:201504
##  Mode  :character   Mode  :character   Mode  :character   Median :201507
##                                                           Mean   :201507
##                                                           3rd Qu.:201510
##                                                           Max.   :201512
```

추천 데이터 - competitor

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    28159 obs. of  4 variables:
##  $ cusID     : chr  "00002" "00051" "00077" "00077" ...
##  $ partner   : chr  "D" "D" "D" "D" ...
##  $ competitor: chr  "D02" "D01" "D02" "D02" ...
##  $ useDate   : int  201507 201504 201503 201506 201507 201508 201511 201510 201511 201508 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - customer

- cusID: 5자리 숫자조합으로 구성된 고객ID
- sex : 성별
- age : 연령
  - 19세이하, 20세~24세, 25세~29세, 30세~34세, 35세~39세, 40세~44세, 45세~49세, 50세~54세, 55세~59세, 60세이상
- area : 거주지역

추천 데이터 - customer

```
##     cusID               sex                 age           
##  Length:19383       Length:19383       Length:19383      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##      area          
##  Length:19383      
##  Class :character  
##  Mode  :character
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    19383 obs. of  4 variables:
##  $ cusID: chr  "00001" "00002" "00003" "00004" ...
##  $ sex  : chr  "M" "M" "M" "F" ...
##  $ age  : chr  "60세이상" "60세이상" "60세이상" "60세이상" ...
##  $ area : chr  "060" "100" "033" "016" ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - item

- partner : 재휴사
- cate_1 : 대분류
- cate_2 : 중분류
- cate_3 : 소분류
- cate_2_name: 중분류명
- cate_3_name: 소분류명

추천 데이터 - item

```
##     partner             cate_1             cate_2
##  Length:4386        Length:4386        Length:4386
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##     cate_3          cate_2_name        cate_3_name
##  Length:4386        Length:4386        Length:4386
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

추천 데이터 - item

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    4386 obs. of  6 variables:
##  $ partner    : chr  "A" "A" "A" "A" ...
##  $ cate_1     : chr  "01" "01" "01" "01" ...
##  $ cate_2     : chr  "0101" "0101" "0101" "0101" ...
##  $ cate_3     : chr  "A010101" "A010102" "A010103" "A010104" ...
##  $ cate_2_name: chr  "일용잡화" "일용잡화" "일용잡화" "일용잡화" ...
##  $ cate_3_name: chr  "위생세제" "휴지류" "뷰티상품" "일용잡화" ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - membership

- cusID : 5자리 숫자조합으로 구성된 고객ID
- memberShip: 멤버십명
- regDate : 가입년월

```
##     cusID           memberShip          regDate     
##  Length:7456        Length:7456        Min.   :201210
##  Class :character   Class :character   1st Qu.:201311
##  Mode  :character   Mode  :character   Median :201407
##                                        Mean   :201412
##                                        3rd Qu.:201504
##                                        Max.   :201512
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7456 obs. of  3 variables:
##  $ cusID     : chr  "00011" "00021" "00037" "00043" ...
##  $ memberShip: chr  "하이마트" "하이마트" "하이마트" "하이마트" ...
##  $ regDate   : int  201512 201506 201306 201403 201411 201312 201506 201404 201406 201311 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - tran

```
##    partner            receiptNum          cate_1          cate_2
##  Length:28593030    Min.   :       1   Min.   : 1.00   Min.   : 101
##  Class :character   1st Qu.: 3922474   1st Qu.: 4.00   1st Qu.: 401
##  Mode  :character   Median : 7167787   Median :11.00   Median :1102
##                     Mean   : 6447881   Mean   :18.37   Mean   :1840
##                     3rd Qu.: 9116336   3rd Qu.:18.00   3rd Qu.:1808
##                     Max.   :11096601   Max.   :92.00   Max.   :9206
##     cate_3              cusID         storeCode           date
##  Length:28593030    Min.   :    1   Min.   :  1.00   Min.   :20140101
##  Class :character   1st Qu.: 5206   1st Qu.: 16.00   1st Qu.:20140711
##  Mode  :character   Median :10104   Median : 44.00   Median :20150110
##                     Mean   : 9904   Mean   : 92.26   Mean   :20145817
##                     3rd Qu.:14638   3rd Qu.:110.00   3rd Qu.:20150703
##                     Max.   :19383   Max.   :593.00   Max.   :20151231
##       time            amount
##  Min.   : 0.00   Min.   :        1
##  1st Qu.:14.00   1st Qu.:     2050
##  Median :17.00   Median :     4290
##  Mean   :16.71   Mean   :    23678
##  3rd Qu.:19.00   3rd Qu.:     9900
##  Max.   :23.00   Max.   :101330000
```

추천 데이터 - tran

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    28593030 obs. of  10 variables:
##  $ partner   : chr  "B" "B" "B" "B" ...
##  $ receiptNum: int  8664000 8664000 8664000 8664000 8664001 8664001 8664002 8664002 8664002 8664003 ...
##  $ cate_1    : int  15 16 16 18 5 15 10 43 54 5 ...
##  $ cate_2    : int  1504 1601 1602 1803 509 1501 1003 4301 5403 504 ...
##  $ cate_3    : chr  "B150401" "B160101" "B160201" "B180301" ...
##  $ cusID     : int  17218 17218 17218 17218 17674 17674 14388 14388 14388 15773 ...
##  $ storeCode : int  44 44 44 44 44 44 44 44 44 44 ...
##  $ date      : int  20140222 20140222 20140222 20140222 20140222 20140222 20140222 20140222 20140222 20140222 ...
##  $ time      : int  20 20 20 20 22 22 23 23 23 21 ...
##  $ amount    : int  2420 1070 8060 6000 1120 1200 5290 5960 9900 970 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

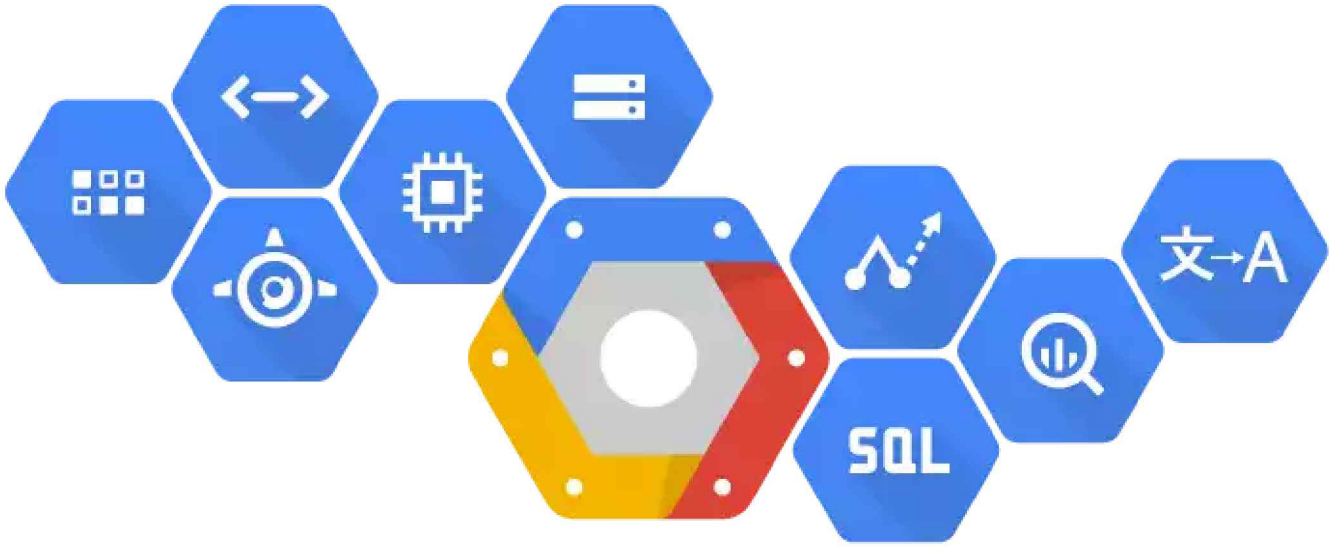
클라우드 서비스

클라우드 서비스 소개

클라우딩 컴퓨팅은 사용자의 환경 밖에서 서비스로서 제공된 확장 가능한 컴퓨팅 자원을 사용한 양에 따라 비용을 지불하고 사용하는 것

출처: P. Changanti, 가상 인프라용 클라우드 서비스, Part 1: IaaS(Infrastructure as a Service) 및 Eucalyptus

구글 클라우드

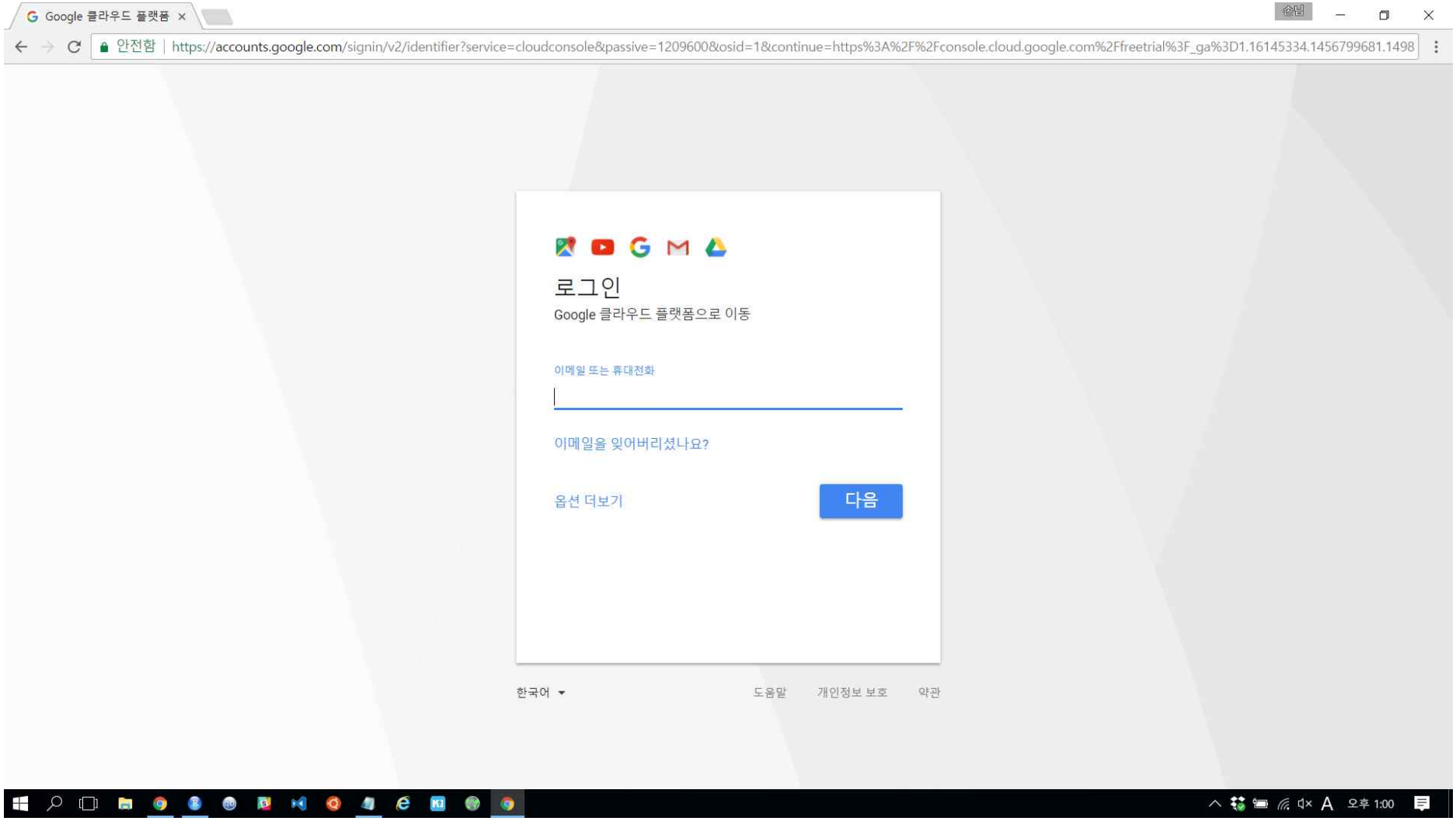


Google Cloud Platform

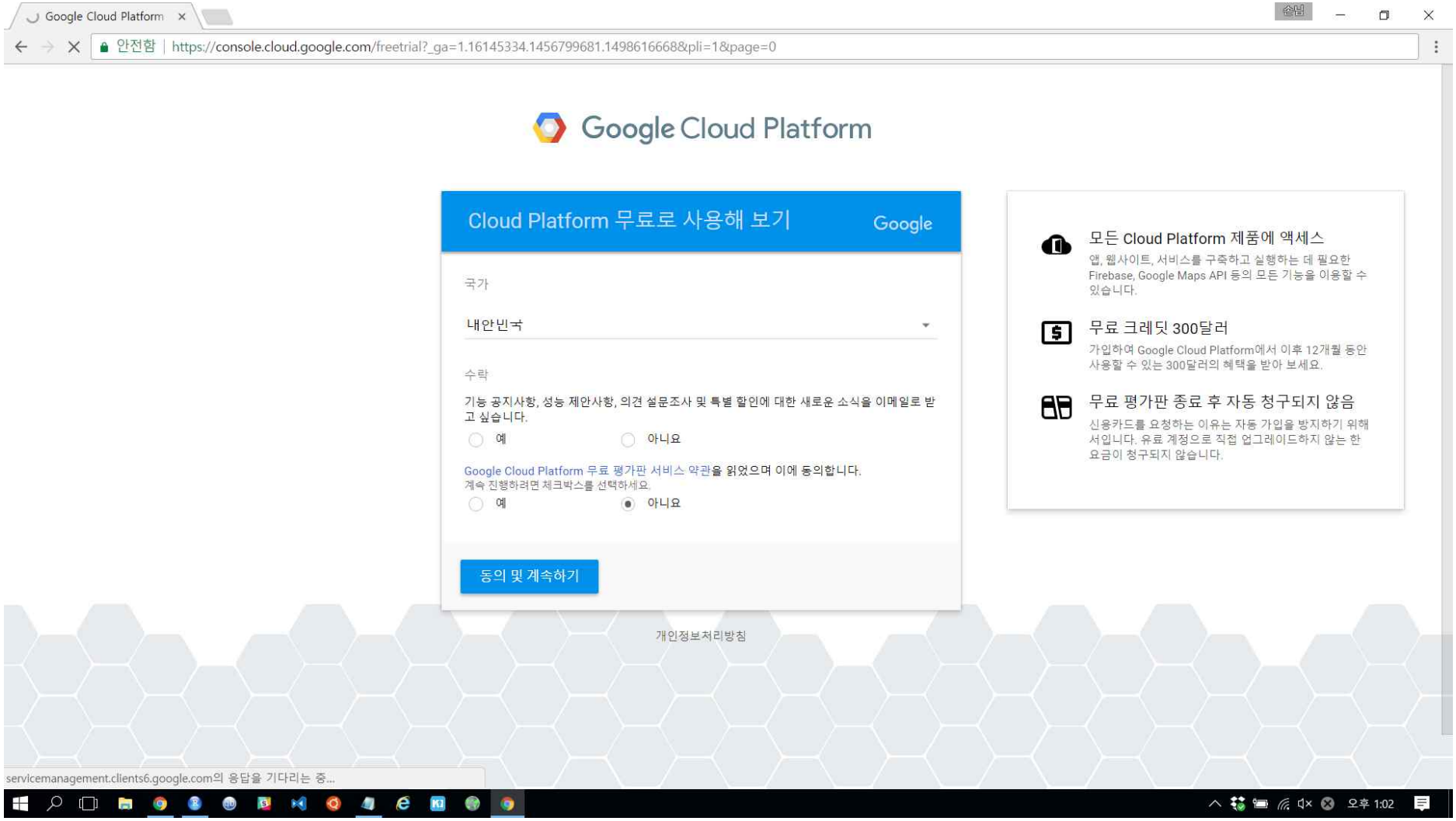
구글 클라우드 시작하기

https://cloud.google.com/

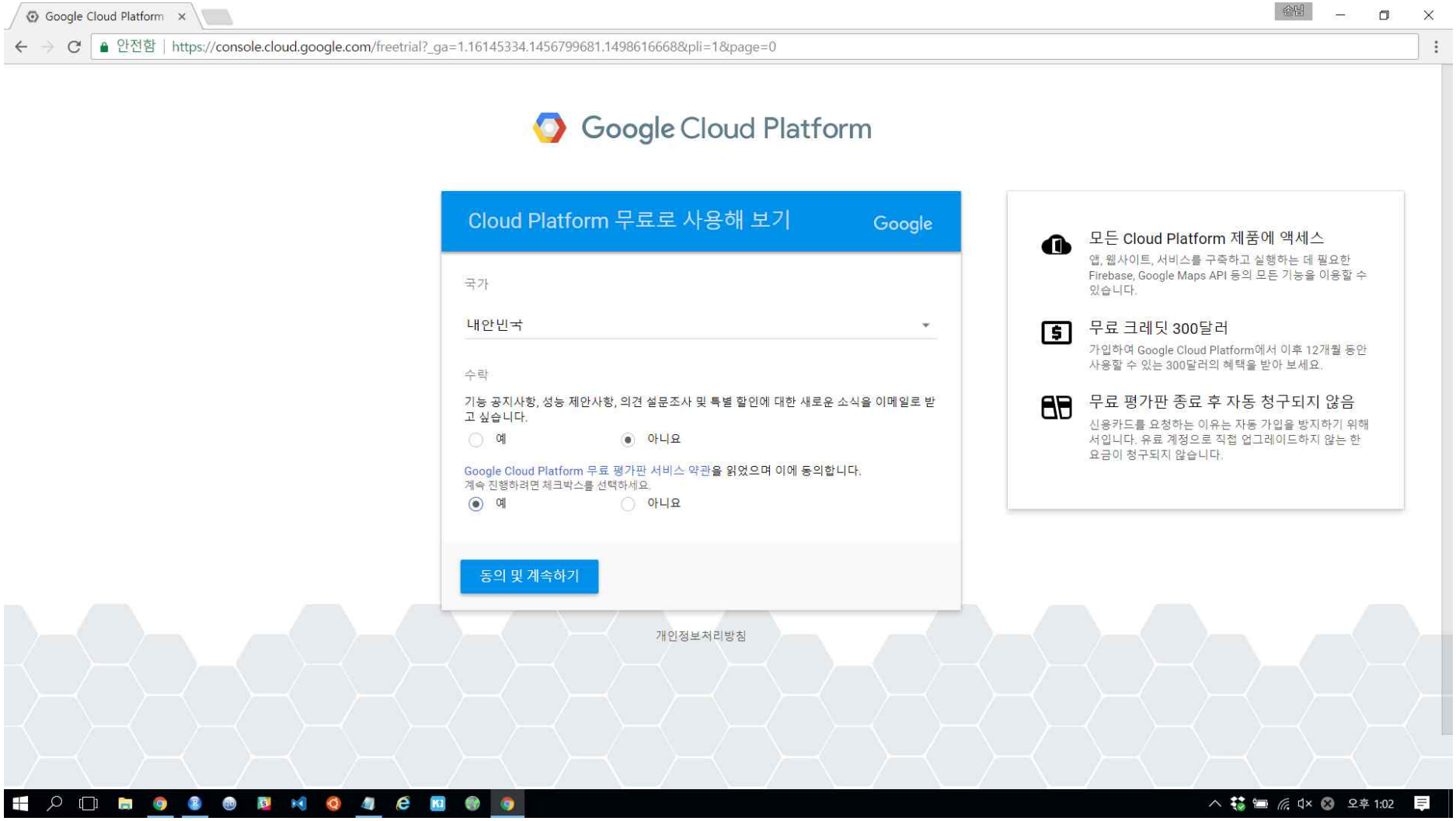
구글 클라우드 로그인


Google 클라우드 플랫폼
손님
안전함 | https://accounts.google.com/signin/v2/identifier?service=cloudconsole&passive=1209600&osid=1&continue=https%3A%2F%2Fconsole.cloud.google.com%2Ffreetrial%3F_ga%3D1.16145334.1456799681.1498
로그인
Google 클라우드 플랫폼으로 이동
이메일 또는 휴대전화
이메일을 잊어버리셨나요?
옵션 더보기
다음
한국어
도움말
개인정보 보호
약관
오후 1:00

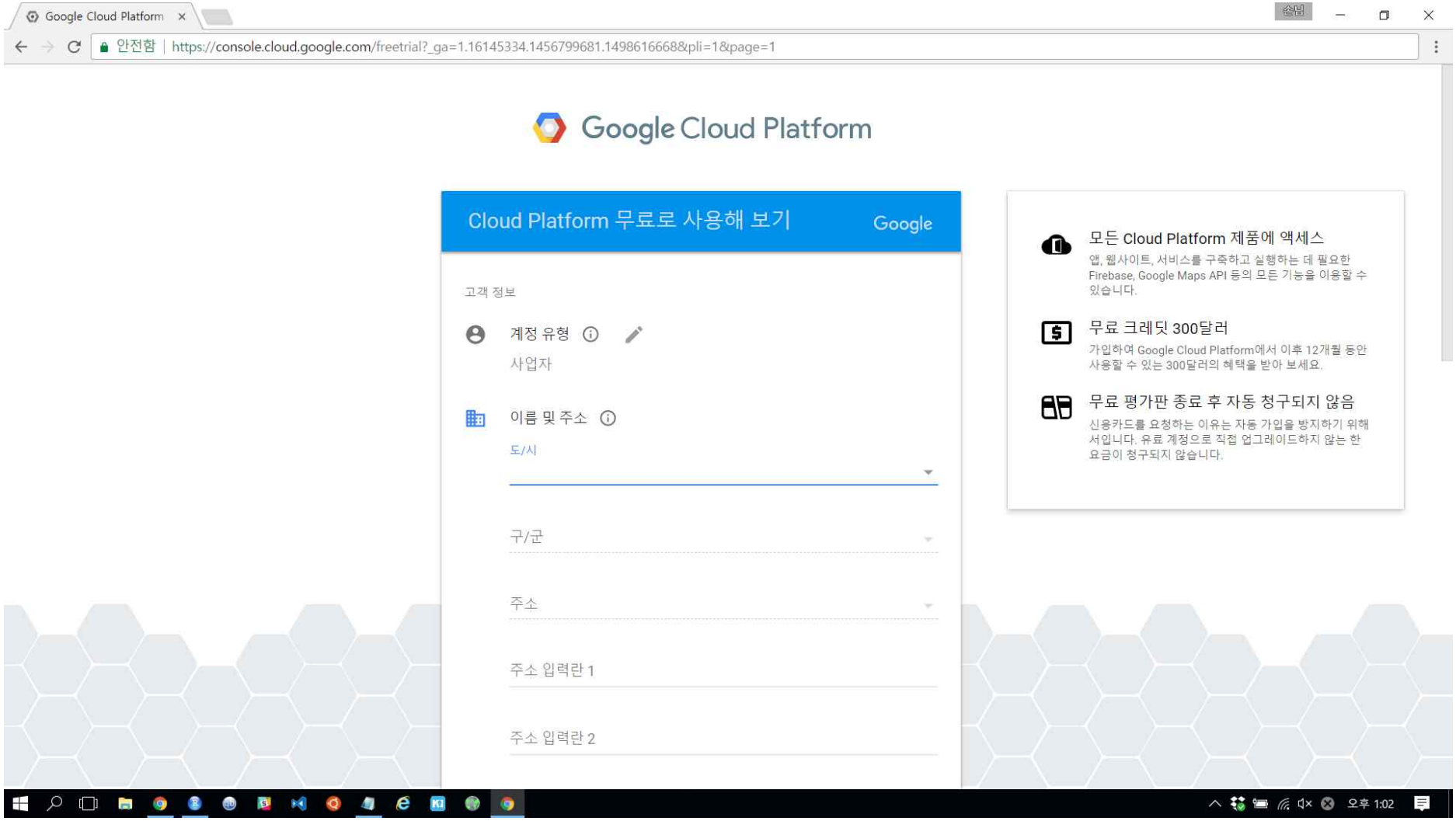
구글 클라우드 설문


Google Cloud Platform
안전함 | https://console.cloud.google.com/freetrial?_ga=1.16145334.1456799681.1498616668&pli=1&page=0
Google Cloud Platform
Cloud Platform 무료로 사용해 보기
Google
국가
내안민국
수락
기능 공지사항, 성능 제안사항, 의견 설문조사 및 특별 할인에 대한 새로운 소식을 이메일로 받고 싶습니다.
예
아니요
Google Cloud Platform 무료 평가판 서비스 약관을 읽었으며 이에 동의합니다.
계속 진행하려면 체크박스를 선택하세요.
예
아니요
동의 및 계속하기
모든 Cloud Platform 제품에 액세스
앱, 웹사이트, 서비스를 구축하고 실행하는 데 필요한 Firebase, Google Maps API 등의 모든 기능을 이용할 수 있습니다.
무료 크레딧 300달러
가입하여 Google Cloud Platform에서 이후 12개월 동안 사용할 수 있는 300달러의 혜택을 받아 보세요.
무료 평가판 종료 후 자동 청구되지 않음
신용카드를 요청하는 이유는 자동 가입을 방지하기 위해서입니다. 유료 계정으로 직접 업그레이드하지 않는 한 요금이 청구되지 않습니다.
개인정보처리방침
servicemanagement.clients6.google.com의 응답을 기다리는 중...
오후 1:02

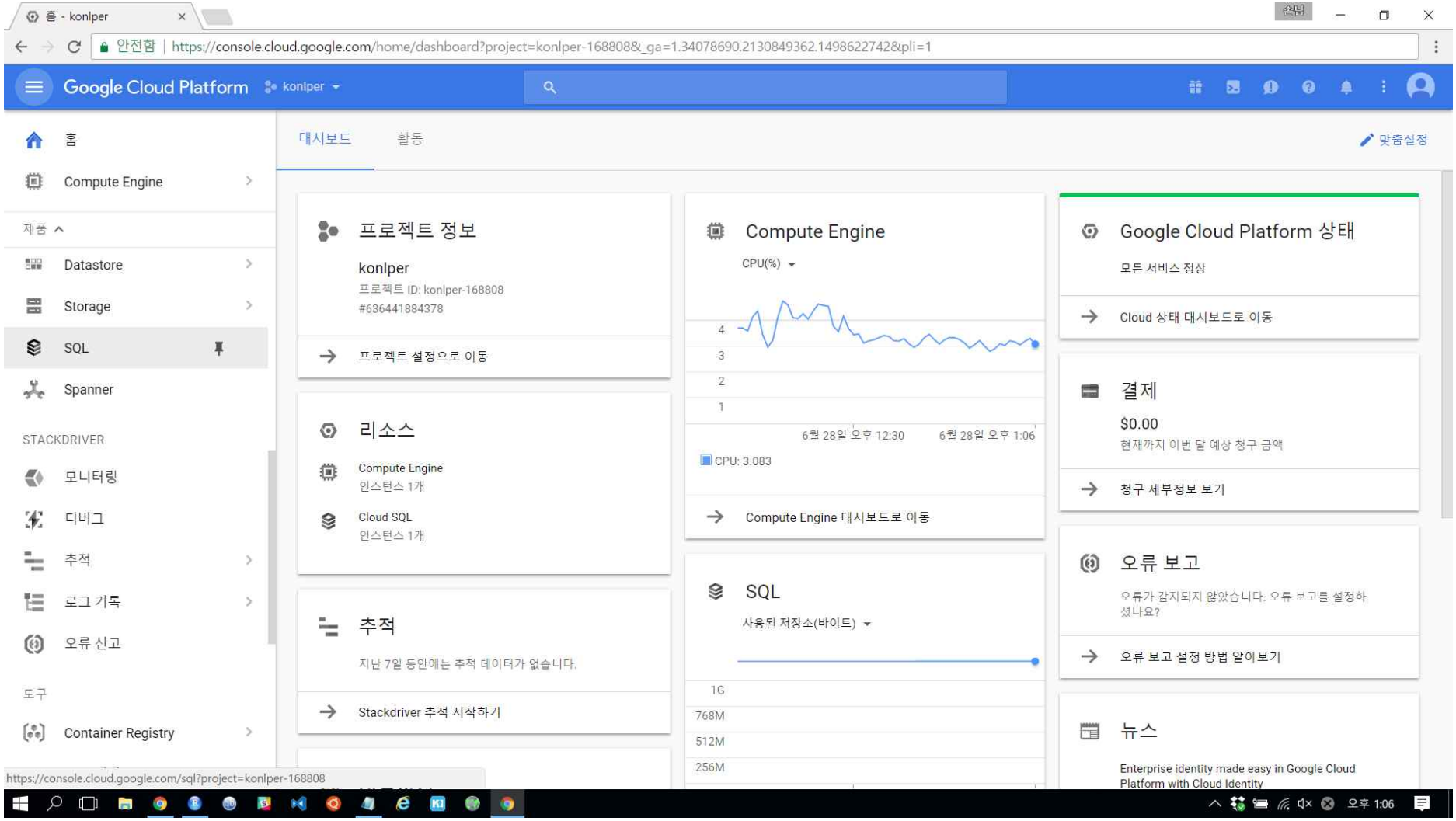
구글 클라우드 설문 선택


Google Cloud Platform
안전함 | https://console.cloud.google.com/freetrial?_ga=1.16145334.1456799681.1498616668&pli=1&page=0
Google Cloud Platform
Cloud Platform 무료로 사용해 보기
Google
국가
내안민국
수락
기능 공지사항, 성능 제안사항, 의견 설문조사 및 특별 할인에 대한 새로운 소식을 이메일로 받고 싶습니다.
예
아니요
Google Cloud Platform 무료 평가판 서비스 약관을 읽었으며 이에 동의합니다.
계속 진행하려면 체크박스를 선택하세요.
예
아니요
동의 및 계속하기
개인정보처리방침
모든 Cloud Platform 제품에 액세스
앱, 웹사이트, 서비스를 구축하고 실행하는 데 필요한 Firebase, Google Maps API 등의 모든 기능을 이용할 수 있습니다.
무료 크레딧 300달러
가입하여 Google Cloud Platform에서 이후 12개월 동안 사용할 수 있는 300달러의 혜택을 받아 보세요.
무료 평가판 종료 후 자동 청구되지 않음
신용카드를 요청하는 이유는 자동 가입을 방지하기 위해서입니다. 유료 계정으로 직접 업그레이드하지 않는 한 요금이 청구되지 않습니다.
오후 1:02

구글 클라우드 카드 등록


Google Cloud Platform
손님
안전함 | https://console.cloud.google.com/freetrial?_ga=1.16145334.1456799681.1498616668&pli=1&page=1
Google Cloud Platform
Cloud Platform 무료로 사용해 보기
Google
고객 정보
계정 유형
사업자
이름 및 주소
도/시
구/군
주소
주소 입력란 1
주소 입력란 2
모든 Cloud Platform 제품에 액세스
앱, 웹사이트, 서비스를 구축하고 실행하는 데 필요한 Firebase, Google Maps API 등의 모든 기능을 이용할 수 있습니다.
무료 크레딧 300달러
가입하여 Google Cloud Platform에서 이후 12개월 동안 사용할 수 있는 300달러의 혜택을 받아 보세요.
무료 평가판 종료 후 자동 청구되지 않음
신용카드를 요청하는 이유는 자동 가입을 방지하기 위해서입니다. 유료 계정으로 직접 업그레이드하지 않는 한 요금이 청구되지 않습니다.
오후 1:02

구글 클라우드 프로젝트 생성

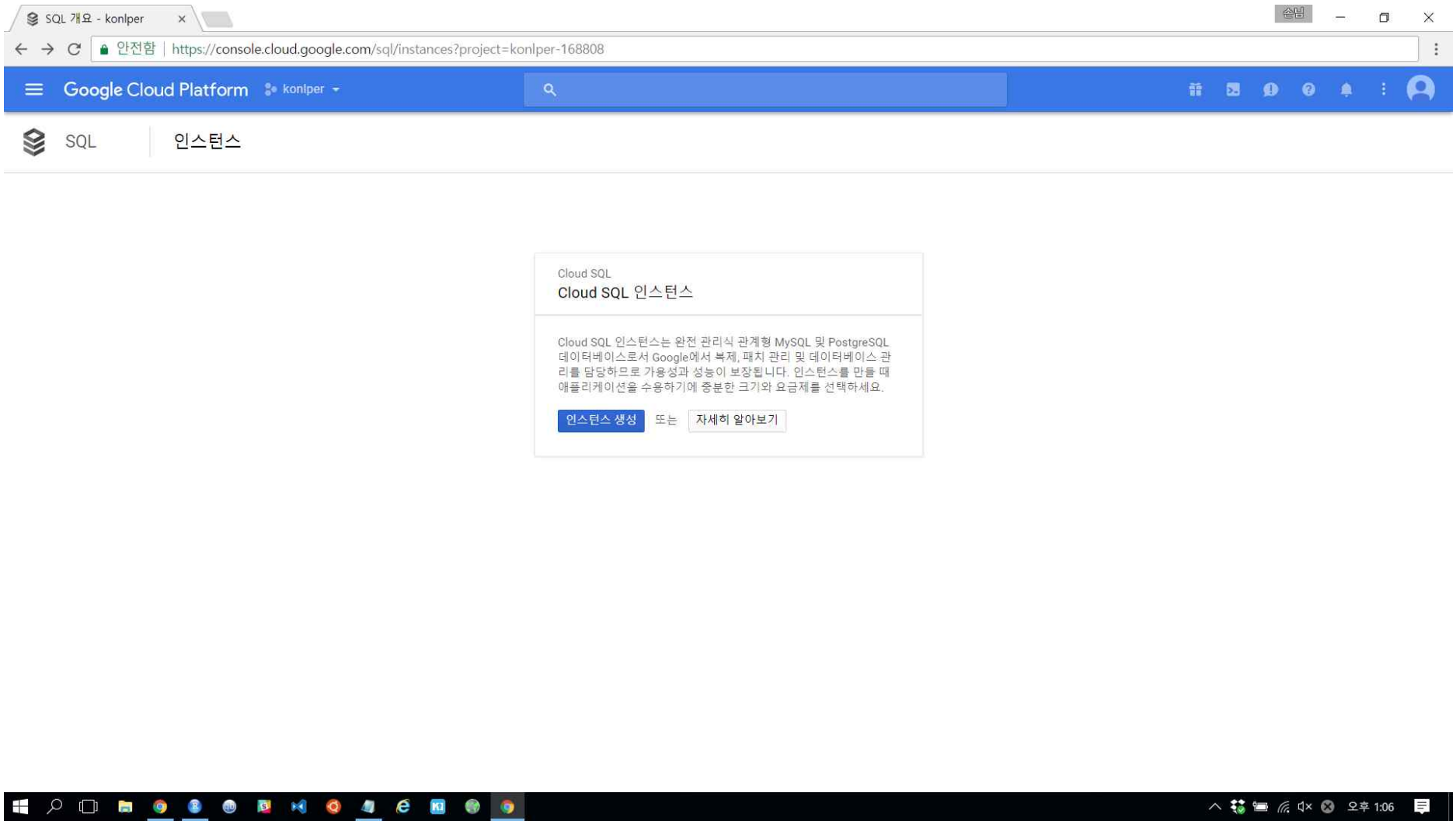


R MySQL 연결

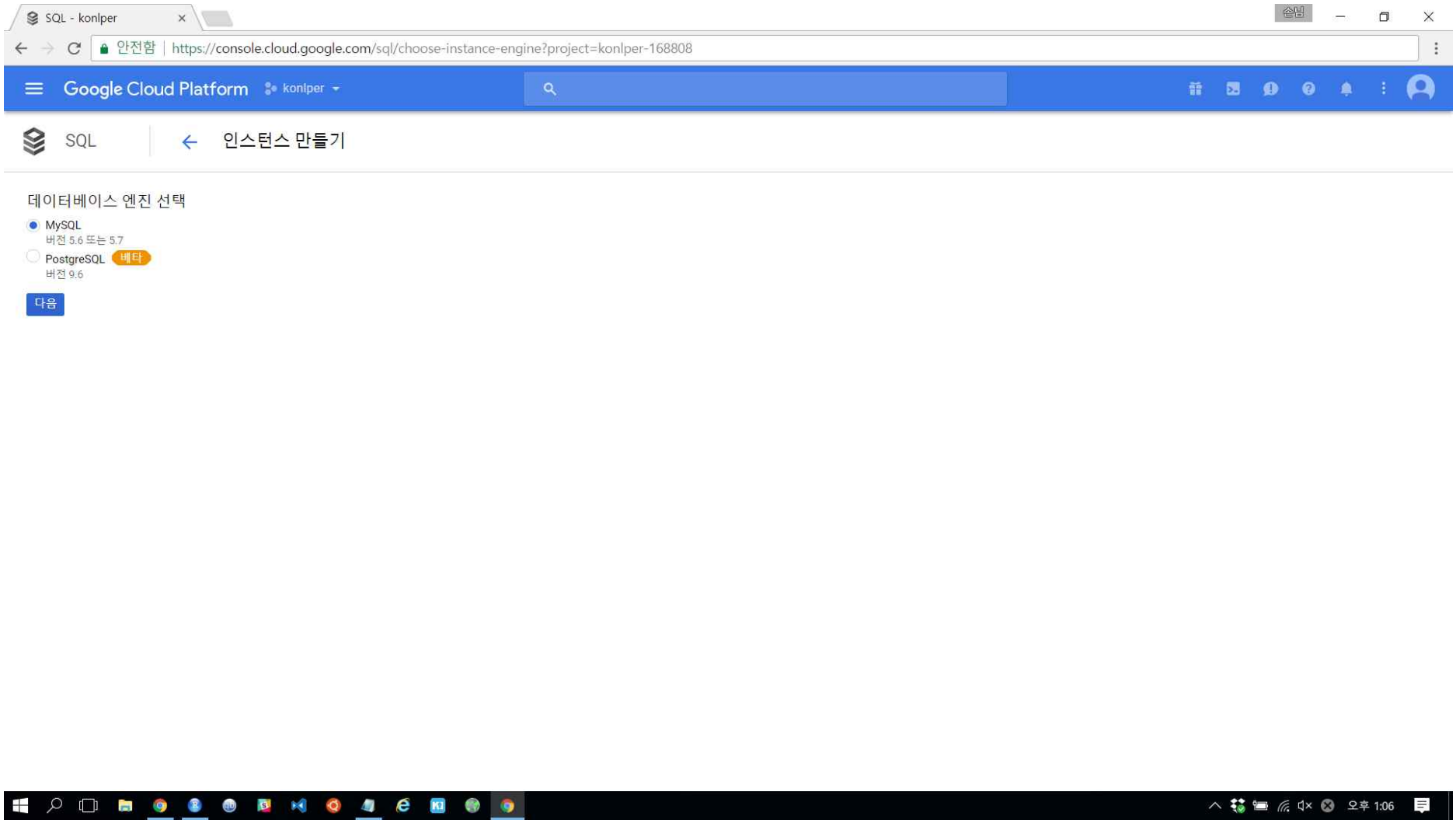
DBI로 MySQL을 연결하는 패키지와 사용

```
library(RMySQL)
con <- dbConnect(MySQL(),
                 user = user,
                 password = pw,
                 host = host,
                 dbname = "recom")
dbListTables(conn = con)
dbWriteTable(conn = con, name = 'tran', value = "./recomen/tran.csv")
dbReadTable(conn = con, name = "tran")
```

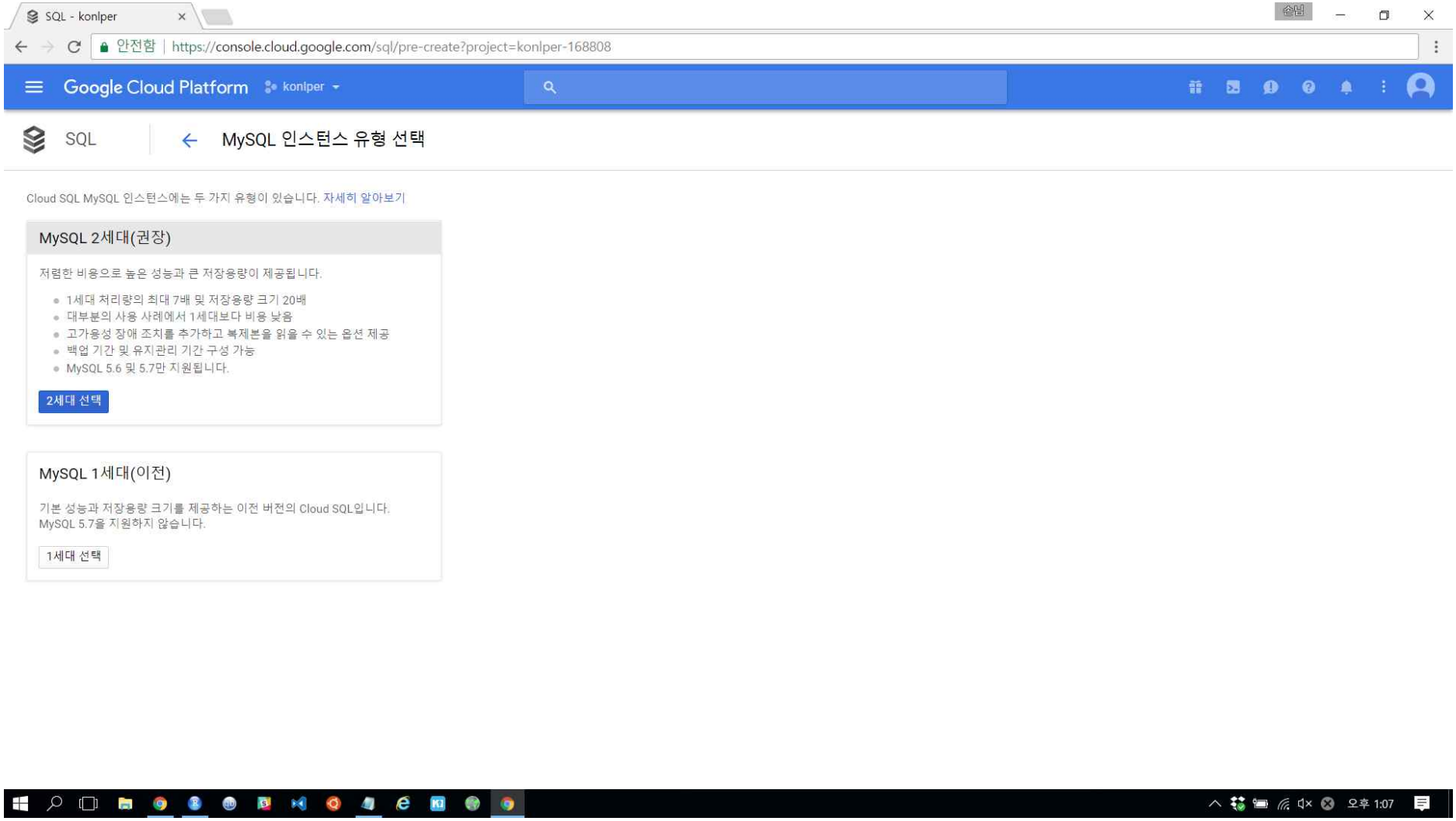
MySQL 인스턴스 생성



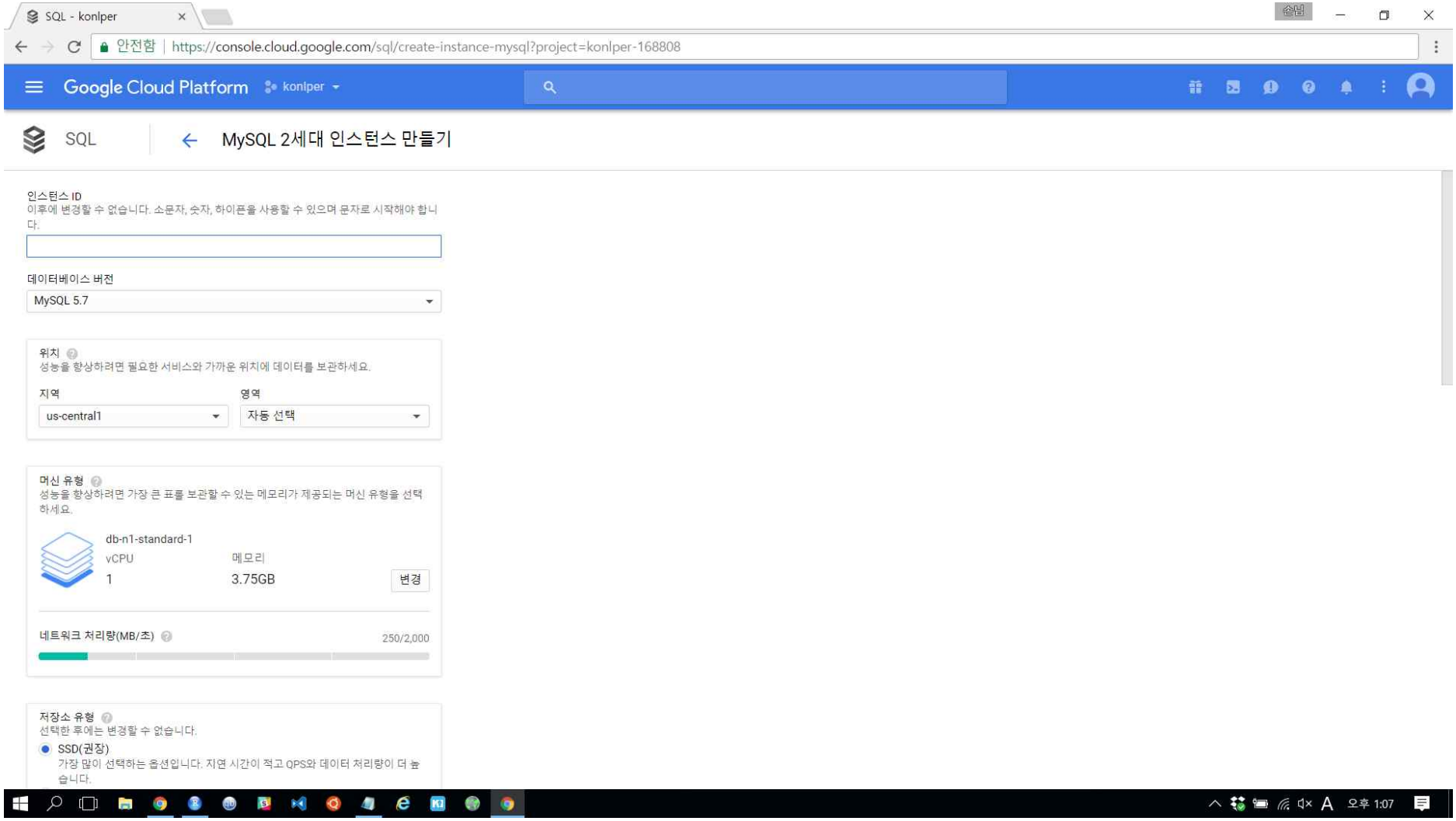
엔진선택


SQL - konlper
손님
안전함 | https://console.cloud.google.com/sql/choose-instance-engine?project=konlper-168808
Google Cloud Platform
konlper
SQL
인스턴스 만들기
데이터베이스 엔진 선택
MySQL
버전 5.6 또는 5.7
PostgreSQL
베타
버전 9.6
다음
오후 1:06

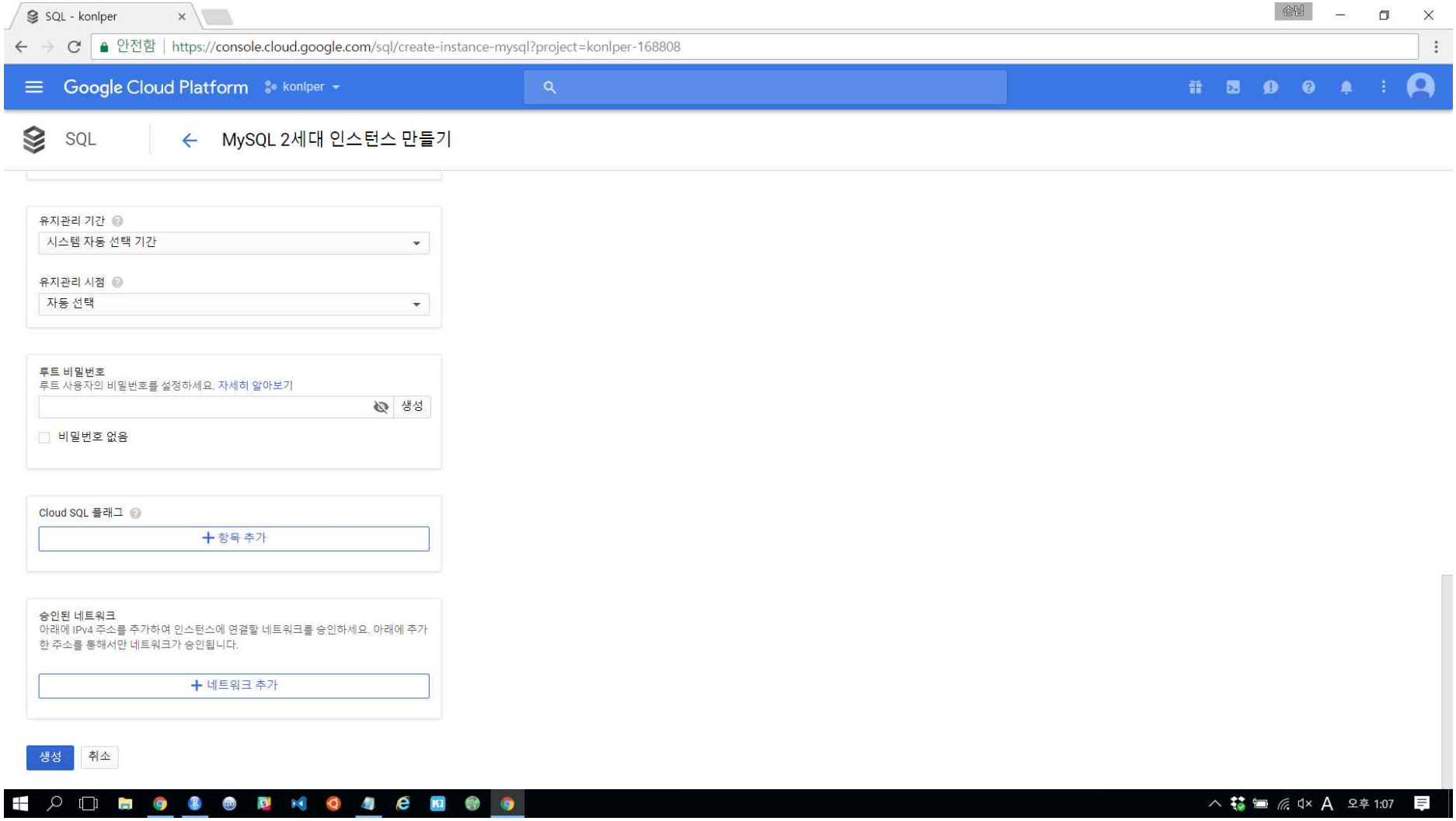
MySQL 2세대 선택


SQL - konlper
손님
안전함 | https://console.cloud.google.com/sql/pre-create?project=konlper-168808
Google Cloud Platform
konlper
SQL
MySQL 인스턴스 유형 선택
Cloud SQL MySQL 인스턴스에는 두 가지 유형이 있습니다. 자세히 알아보기
MySQL 2세대(권장)
저렴한 비용으로 높은 성능과 큰 저장용량이 제공됩니다.
1세대 처리량의 최대 7배 및 저장용량 크기 20배
대부분의 사용 사례에서 1세대보다 비용 낮음
고가용성 장애 조치를 추가하고 복제본을 읽을 수 있는 옵션 제공
백업 기간 및 유지관리 기간 구성 가능
MySQL 5.6 및 5.7만 지원됩니다.
2세대 선택
MySQL 1세대(이전)
기본 성능과 저장용량 크기를 제공하는 이전 버전의 Cloud SQL입니다.
MySQL 5.7을 지원하지 않습니다.
1세대 선택
오후 1:07

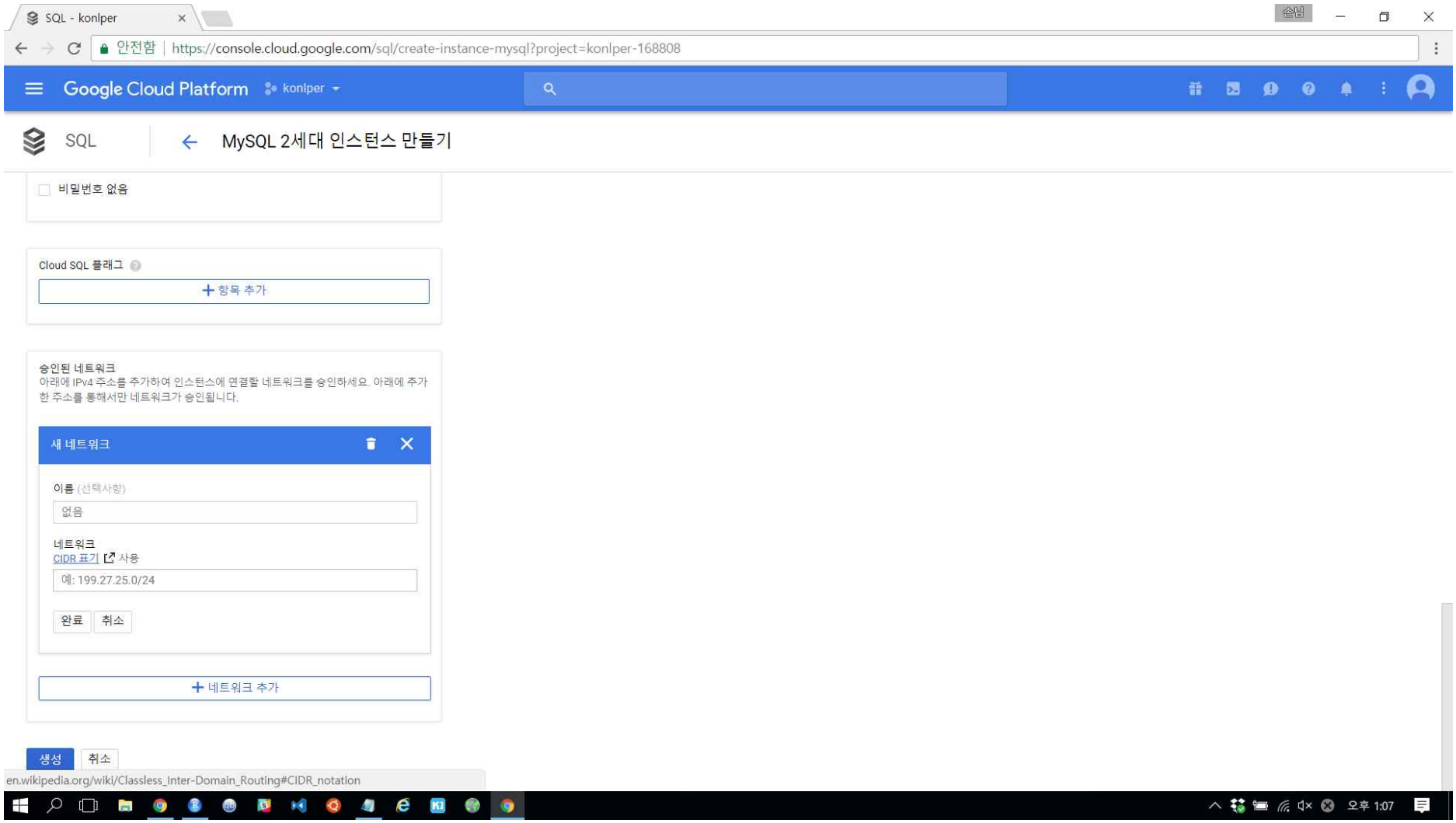
인스턴스 설정


SQL - konlper
손님
안전함 | https://console.cloud.google.com/sql/create-instance-mysql?project=konlper-168808
Google Cloud Platform
konlper
SQL
MySQL 2세대 인스턴스 만들기
인스턴스 ID
이후에 변경할 수 없습니다. 소문자, 숫자, 하이픈을 사용할 수 있으며 문자로 시작해야 합니다.
데이터베이스 버전
MySQL 5.7
위치
성능을 향상하려면 필요한 서비스와 가까운 위치에 데이터를 보관하세요.
지역
us-central1
영역
자동 선택
머신 유형
성능을 향상하려면 가장 큰 표를 보관할 수 있는 메모리가 제공되는 머신 유형을 선택하세요.
db-n1-standard-1
vCPU
1
메모리
3.75GB
변경
네트워크 처리량(MB/초)
250/2,000
저장소 유형
선택한 후에는 변경할 수 없습니다.
SSD(권장)
가장 많이 선택하는 옵션입니다. 지연 시간이 적고 QPS와 데이터 처리량이 더 높습니다.
오후 1:07

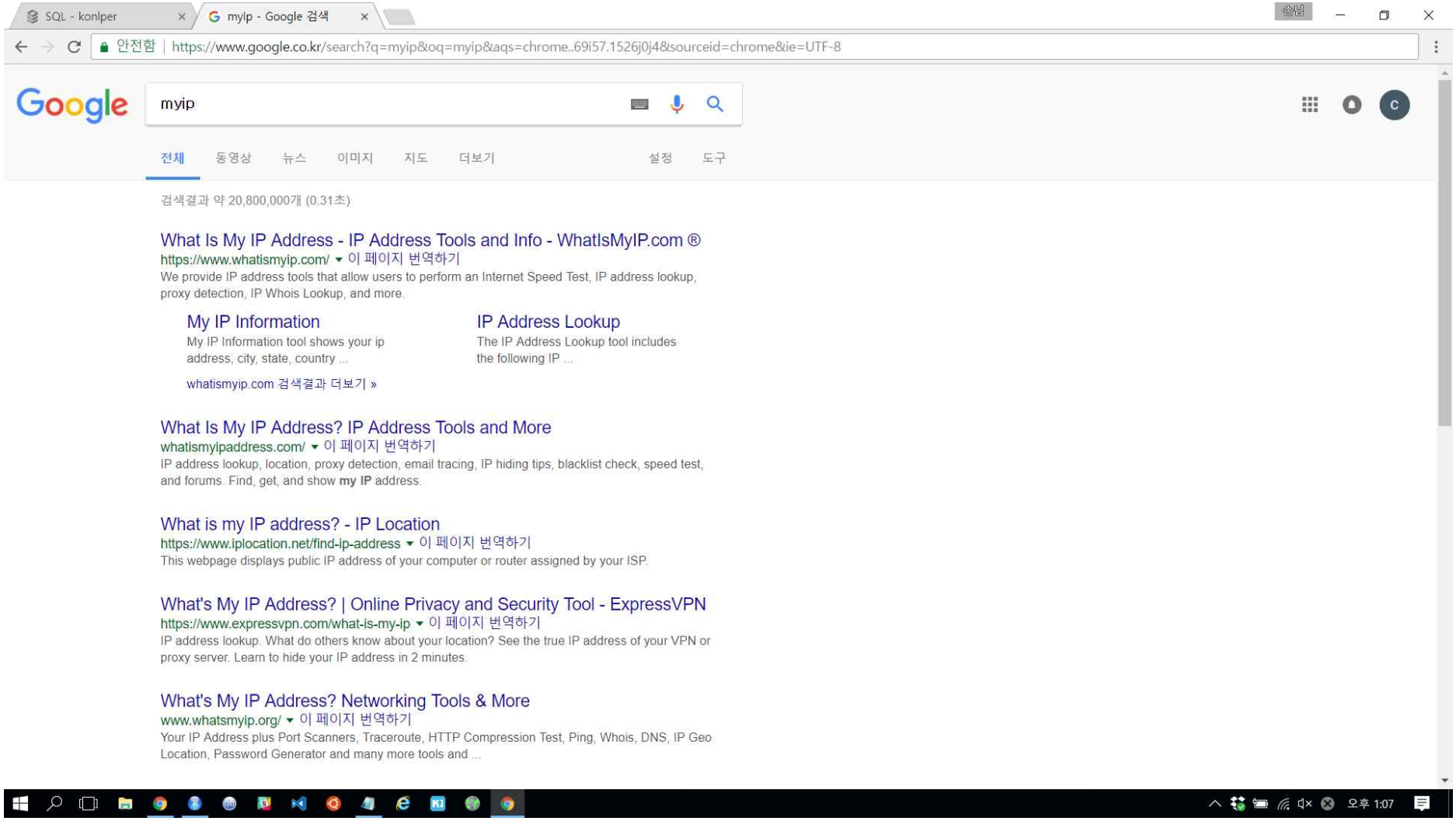
접속 허용 IP 설정


SQL - konlper
손님
안전함 | https://console.cloud.google.com/sql/create-instance-mysql?project=konlper-168808
Google Cloud Platform
konlper
SQL
MySQL 2세대 인스턴스 만들기
유지관리 기간
시스템 자동 선택 기간
유지관리 시점
자동 선택
루트 비밀번호
루트 사용자의 비밀번호를 설정하세요. 자세히 알아보기
생성
비밀번호 없음
Cloud SQL 플래그
+ 항목 추가
승인된 네트워크
아래에 IPv4 주소를 추가하여 인스턴스에 연결할 네트워크를 승인하세요. 아래에 추가한 주소를 통해서만 네트워크가 승인됩니다.
+ 네트워크 추가
생성
취소
오후 1:07

접속 허용 IP 설정


SQL - konlper
손님
안전함 | https://console.cloud.google.com/sql/create-instance-mysql?project=konlper-168808
Google Cloud Platform
konlper
SQL
MySQL 2세대 인스턴스 만들기
비밀번호 없음
Cloud SQL 플래그
+ 항목 추가
승인된 네트워크
아래에 IPv4 주소를 추가하여 인스턴스에 연결할 네트워크를 승인하세요. 아래에 주가한 주소를 통해서만 네트워크가 승인됩니다.
새 네트워크
이름 (선택사항)
없음
네트워크
CIDR 표기 사용
예: 199.27.25.0/24
완료
취소
+ 네트워크 추가
생성
취소
en.wikipedia.org/wiki/Classless_Inter-Domain_Routing#CIDR_notation
오후 1:07

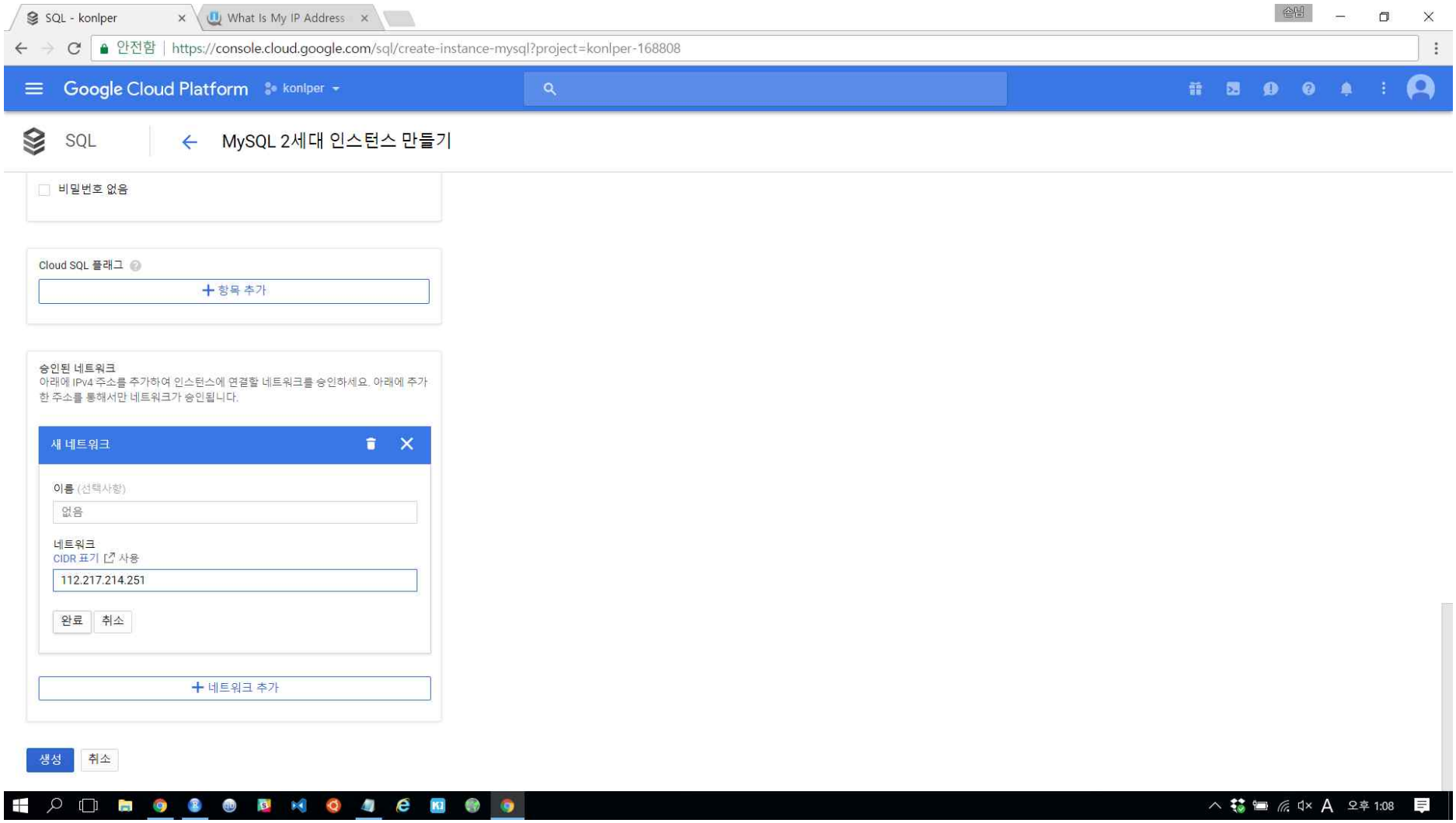
현재 IP 확인하기


SQL - konlper
myip - Google 검색
손님
안전함 | https://www.google.co.kr/search?q=myip&oq=myip&aqs=chrome..69i57.1526j0j4&sourceid=chrome&ie=UTF-8
Google
myip
전체 동영상 뉴스 이미지 지도 더보기 설정 도구
검색결과 약 20,800,000개 (0.31초)
What Is My IP Address - IP Address Tools and Info - WhatIsMyIP.com ®
https://www.whatismyip.com/ 이 페이지 번역하기
We provide IP address tools that allow users to perform an Internet Speed Test, IP address lookup, proxy detection, IP Whois Lookup, and more.
My IP Information
My IP Information tool shows your ip address, city, state, country ...
IP Address Lookup
The IP Address Lookup tool includes the following IP ...
whatismyip.com 검색결과 더보기 »
What Is My IP Address? IP Address Tools and More
whatismyipaddress.com/ 이 페이지 번역하기
IP address lookup, location, proxy detection, email tracing, IP hiding tips, blacklist check, speed test, and forums. Find, get, and show my IP address.
What is my IP address? - IP Location
https://www.iplocation.net/find-ip-address 이 페이지 번역하기
This webpage displays public IP address of your computer or router assigned by your ISP.
What's My IP Address? | Online Privacy and Security Tool - ExpressVPN
https://www.expressvpn.com/what-is-my-ip 이 페이지 번역하기
IP address lookup. What do others know about your location? See the true IP address of your VPN or proxy server. Learn to hide your IP address in 2 minutes.
What's My IP Address? Networking Tools & More
www.whatsmyip.org/ 이 페이지 번역하기
Your IP Address plus Port Scanners, Traceroute, HTTP Compression Test, Ping, Whois, DNS, IP Geo Location, Password Generator and many more tools and ...
오후 1:07

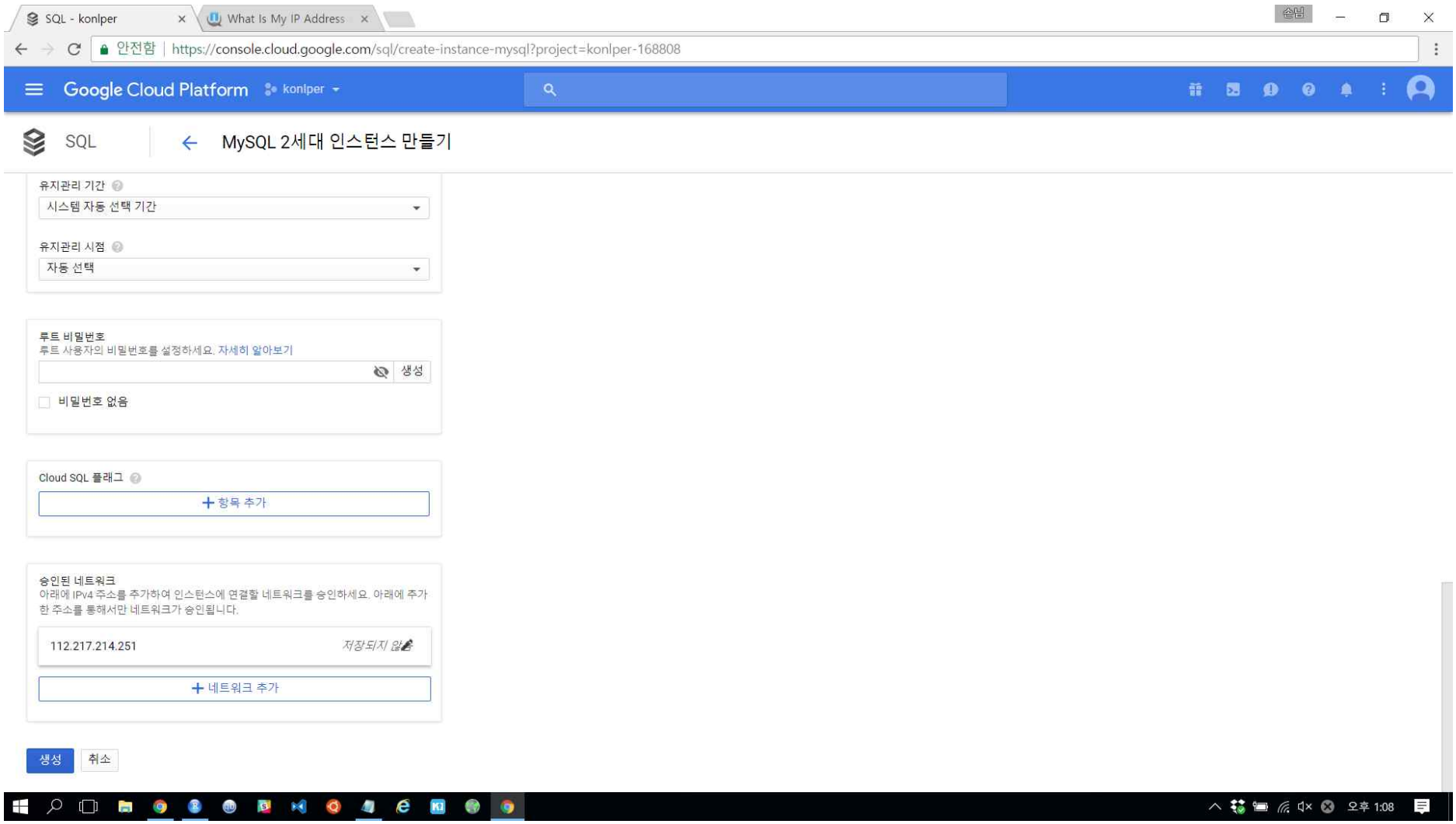
whatismyip

https://www.whatismyip.com/

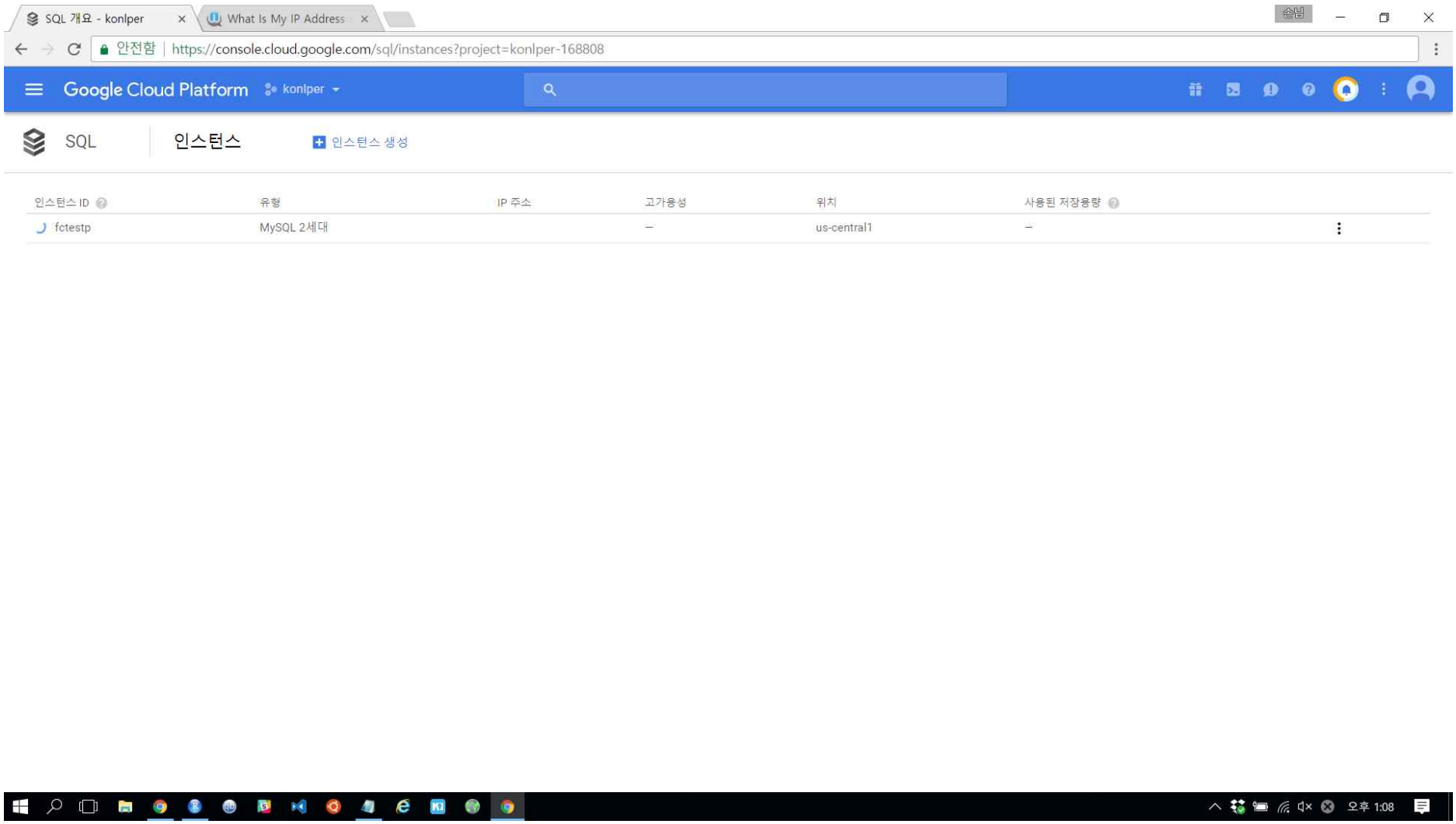
ip 입력


SQL - konlper
What Is My IP Address
손님
안전함 | https://console.cloud.google.com/sql/create-instance-mysql?project=konlper-168808
Google Cloud Platform
konlper
SQL
MySQL 2세대 인스턴스 만들기
비밀번호 없음
Cloud SQL 플래그
+ 항목 추가
승인된 네트워크
아래에 IPv4 주소를 추가하여 인스턴스에 연결할 네트워크를 승인하세요. 아래에 추가한 주소를 통해서만 네트워크가 승인됩니다.
새 네트워크
이름 (선택사항)
없음
네트워크
CIDR 표기 사용
112.217.214.251
완료
취소
+ 네트워크 추가
생성
취소
오후 1:08

root 계정 비밀번호 생성

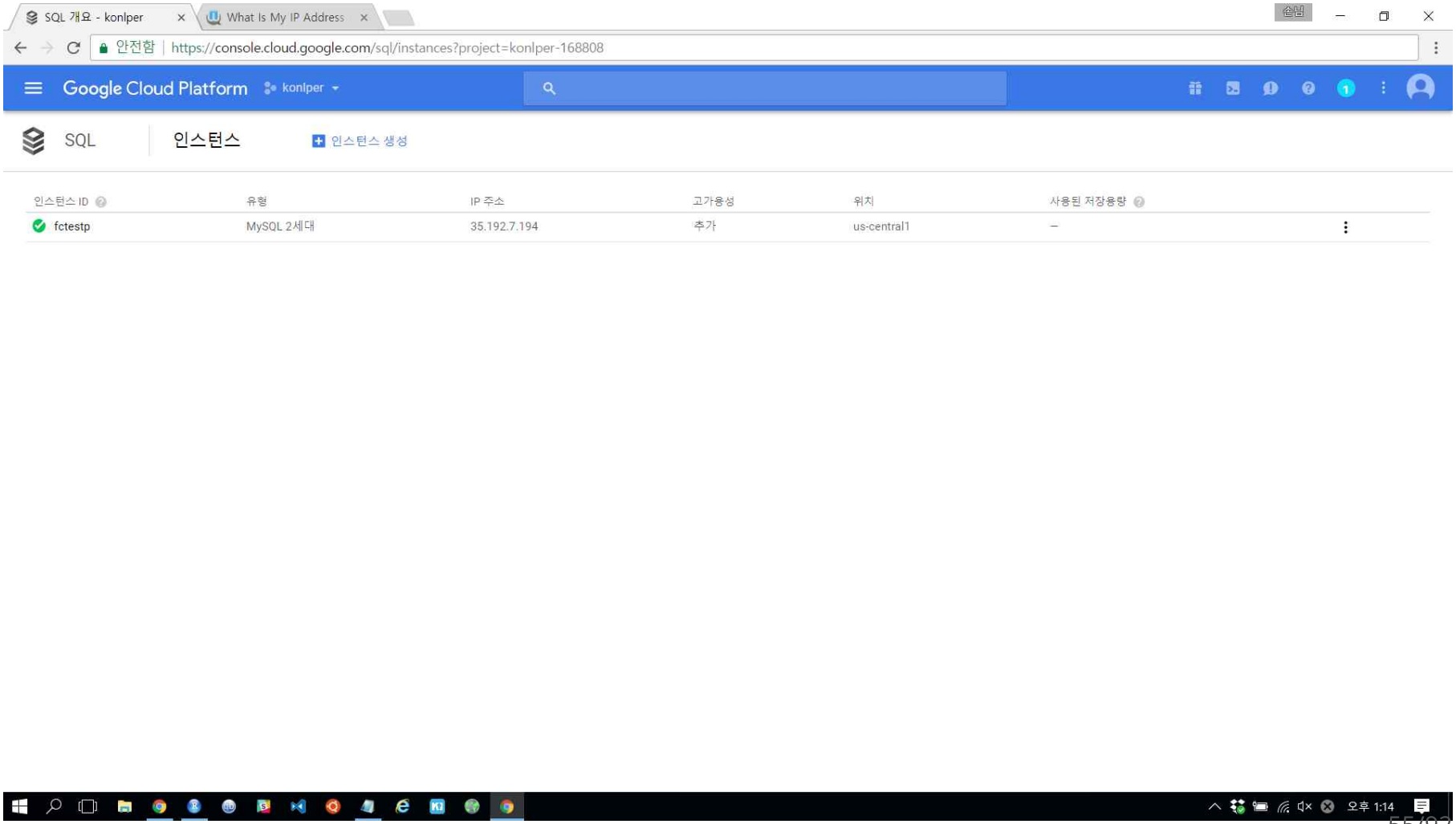


SQL 인스턴스 생성중

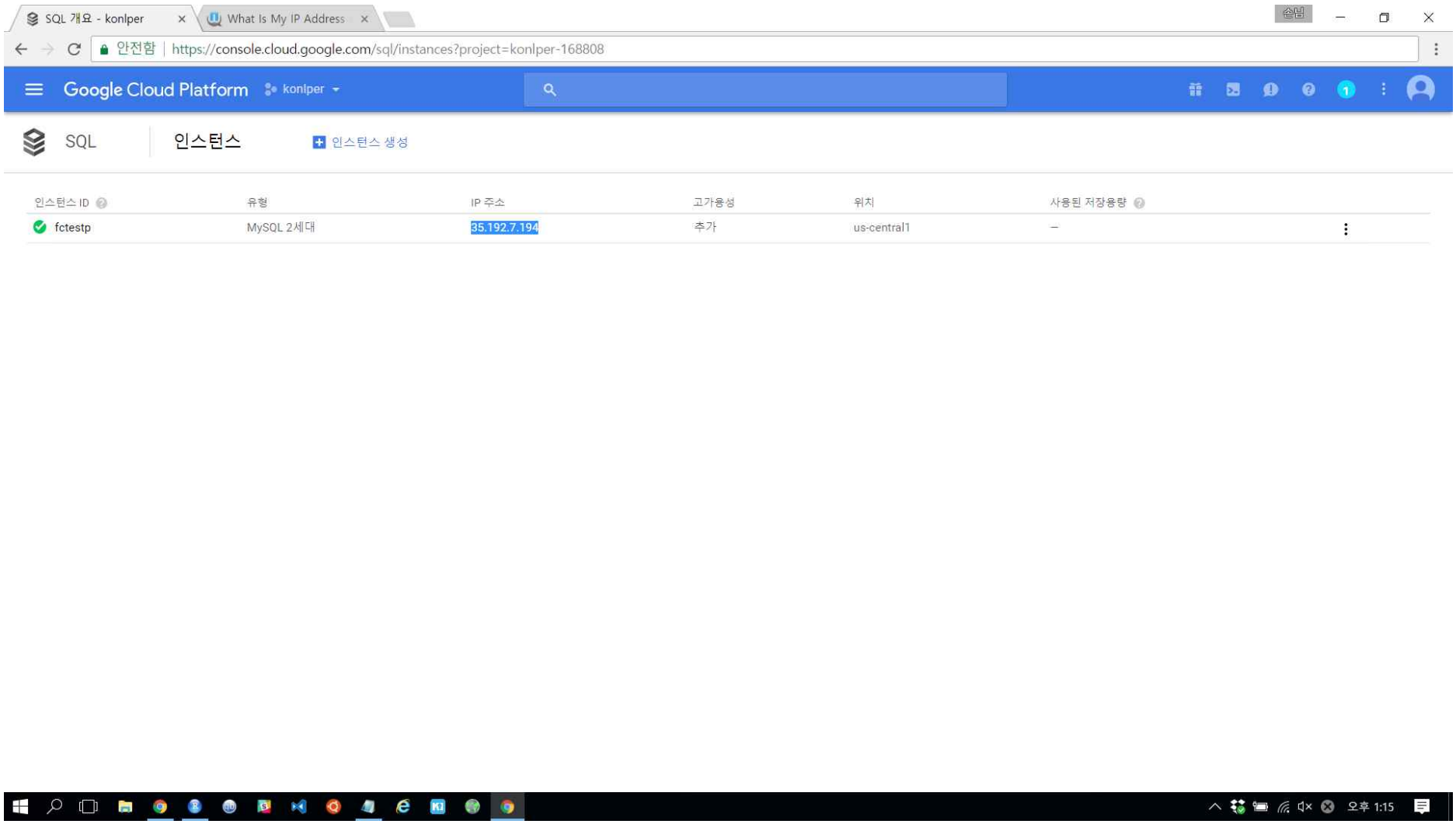

SQL 개요 - konlper
What Is My IP Address
손님
안전함 | https://console.cloud.google.com/sql/instances?project=konlper-168808
Google Cloud Platform
konlper
SQL
인스턴스
인스턴스 생성
인스턴스 ID
유형
IP 주소
고가용성
위치
사용된 저장용량
fctestp
MySQL 2세대
–
us-central1
–
오후 1:08

RMySQL로 연결하기

- host 주소 파악

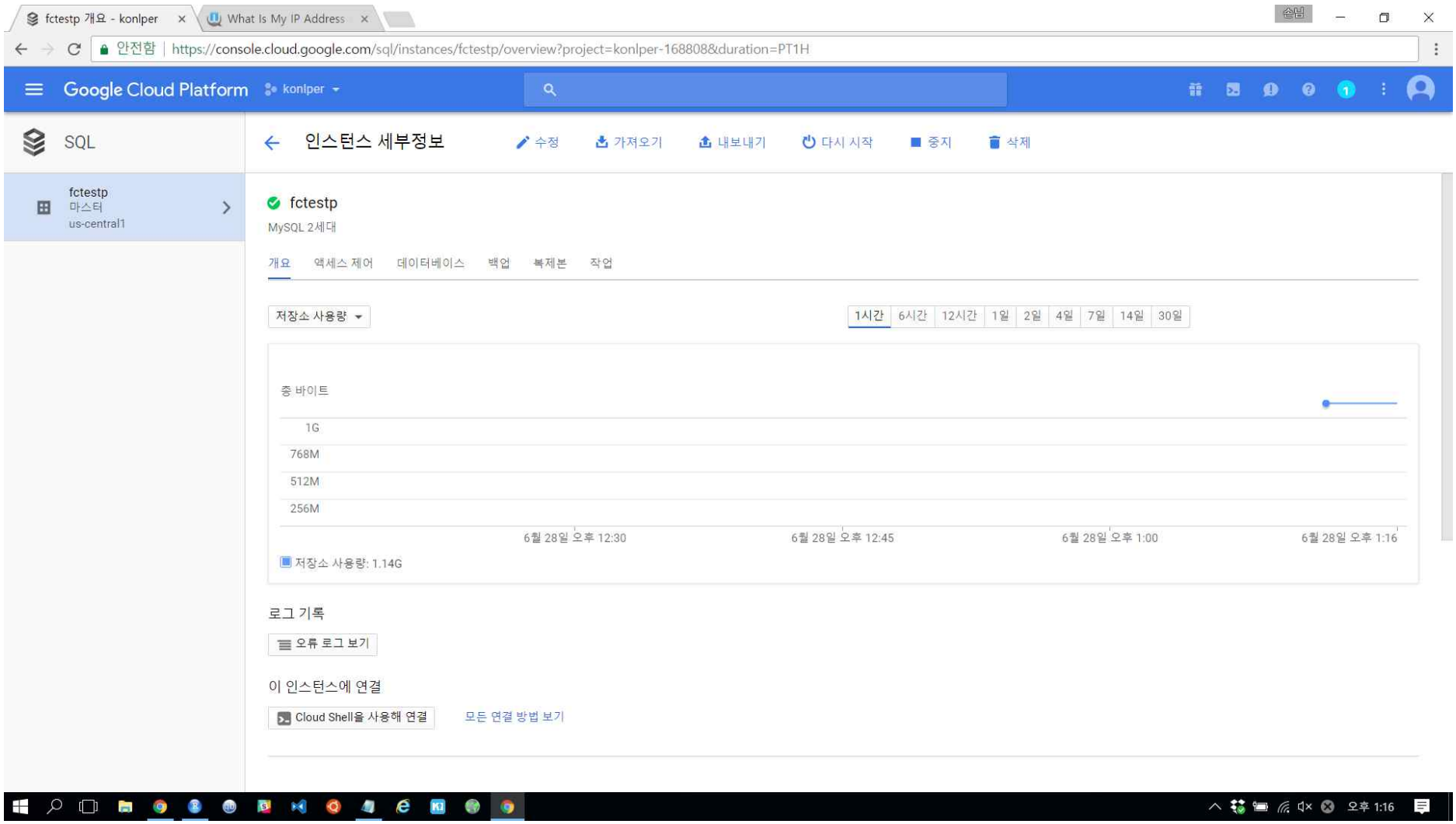


RMySQL로 연결하기

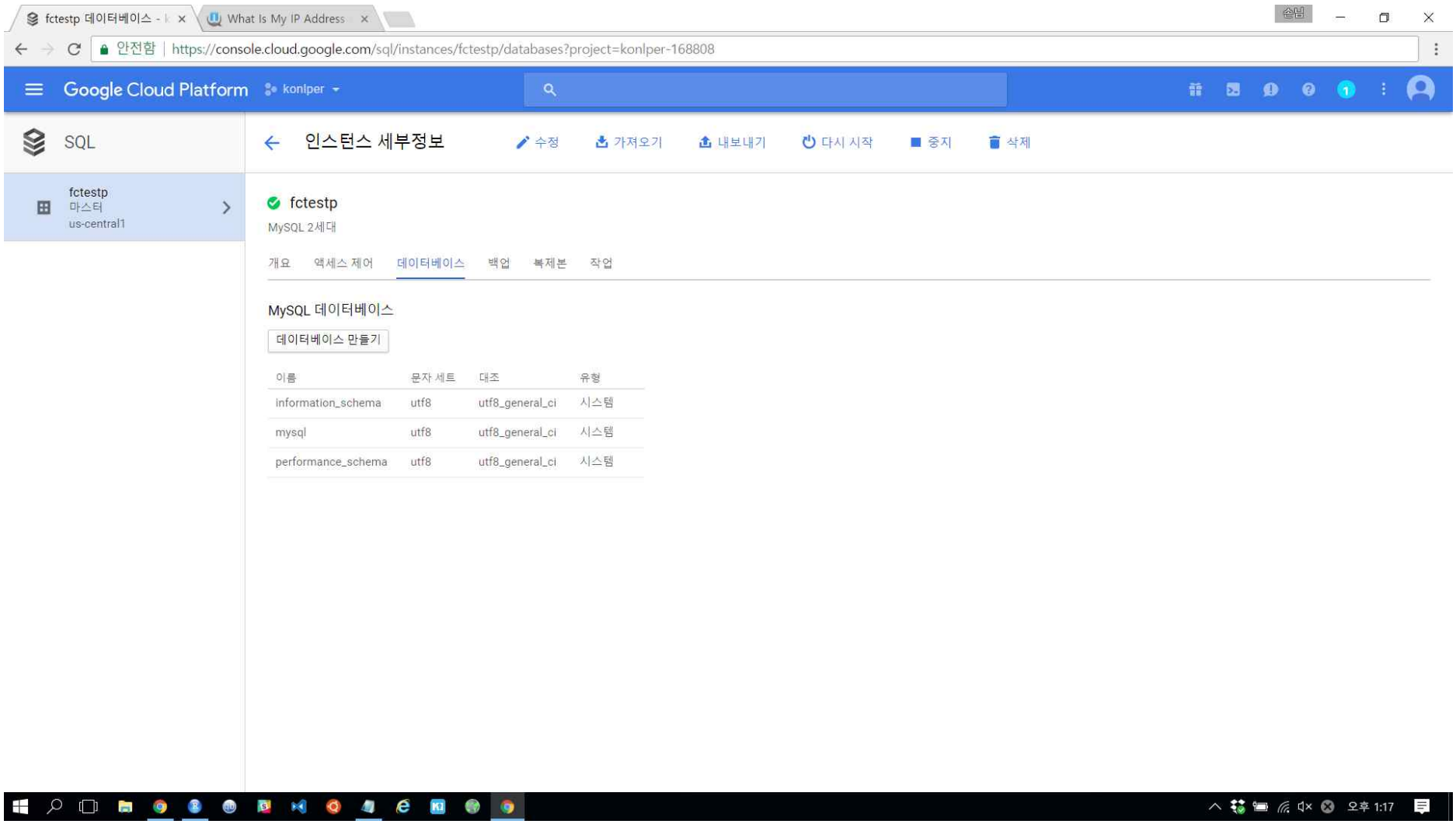


데이터베이스 만들기

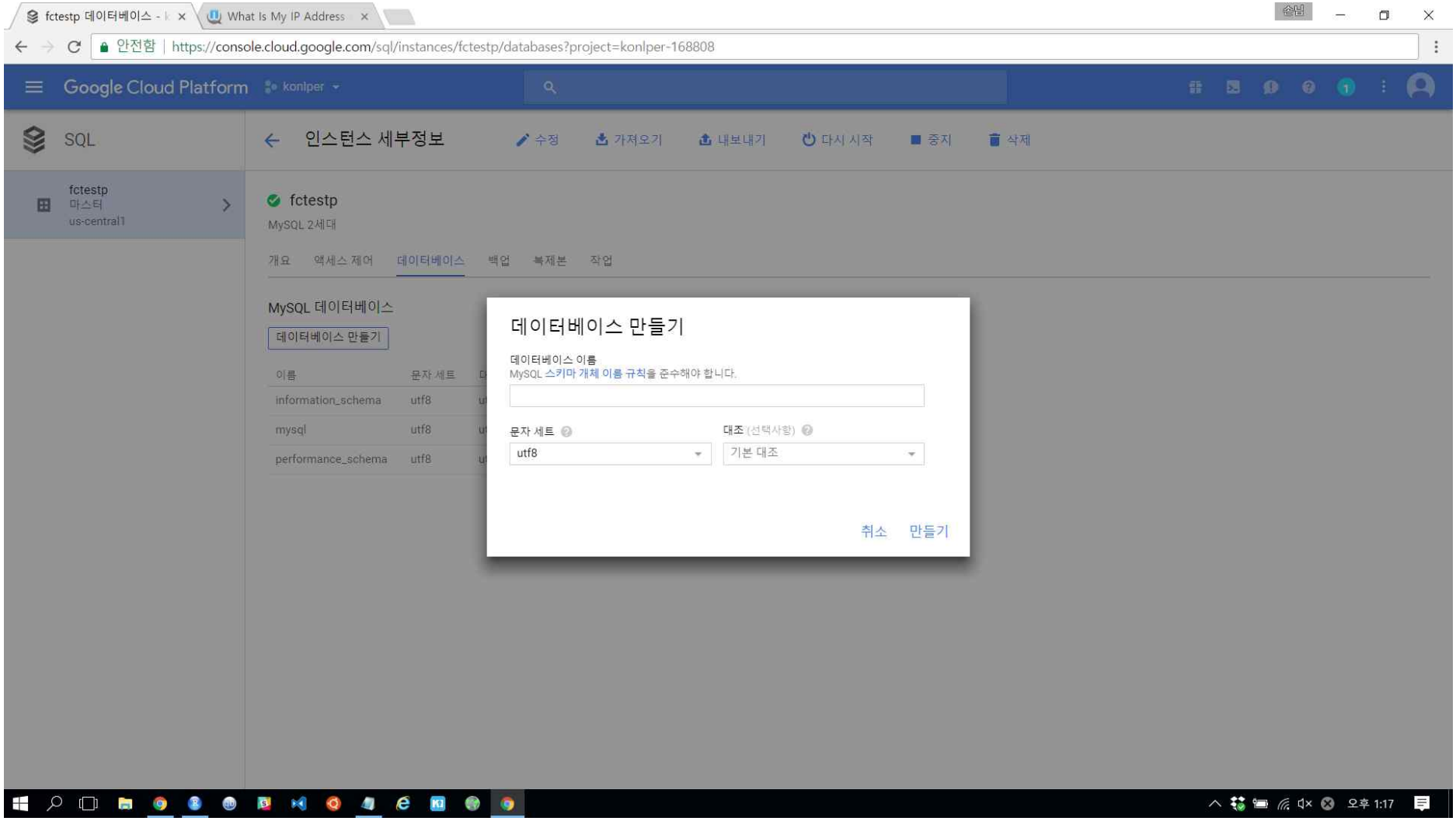
데이터베이스 만들기


fctestp 개요 - konlper
What Is My IP Address
손님
안전함 | https://console.cloud.google.com/sql/instances/fctestp/overview?project=konlper-168808&duration=PT1H
Google Cloud Platform
konlper
SQL
인스턴스 세부정보
수정
가져오기
내보내기
다시 시작
중지
삭제
fctestp
마스터
us-central1
fctestp
MySQL 2세대
개요
액세스 제어
데이터베이스
백업
복제본
작업
저장소 사용량
1시간
6시간
12시간
1일
2일
4일
7일
14일
30일
총 바이트
1G
768M
512M
256M
6월 28일 오후 12:30
6월 28일 오후 12:45
6월 28일 오후 1:00
6월 28일 오후 1:16
저장소 사용량: 1.14G
로그 기록
오류 로그 보기
이 인스턴스에 연결
Cloud Shell을 사용해 연결
모든 연결 방법 보기
오후 1:16

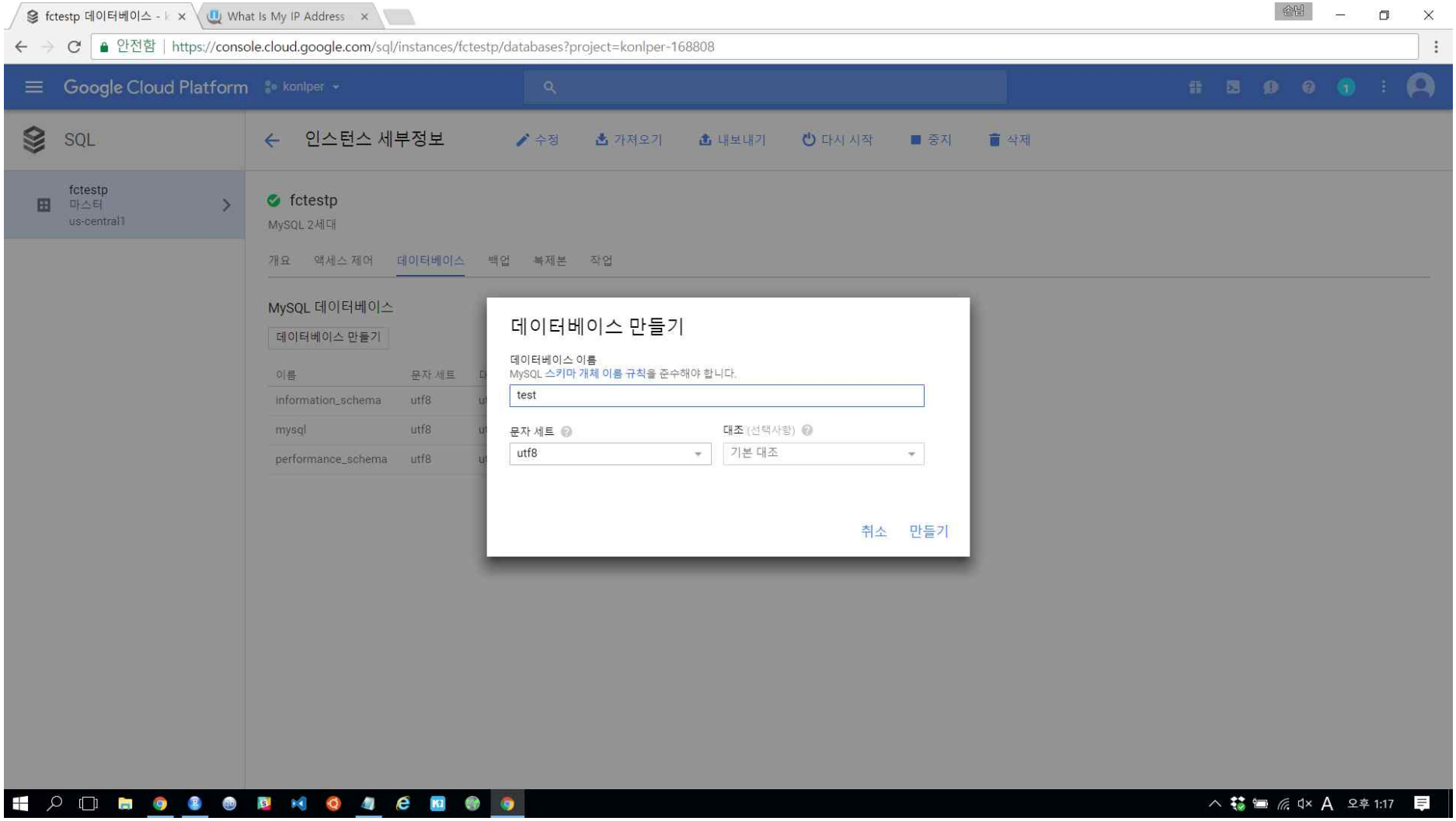
데이터베이스 만들기


fctestp 데이터베이스 -
What Is My IP Address
손님
안전함 | https://console.cloud.google.com/sql/instances/fctestp/databases?project=konlper-168808
Google Cloud Platform
konlper
SQL
인스턴스 세부정보
수정
가져오기
내보내기
다시 시작
중지
삭제
fctestp
마스터
us-central1
fctestp
MySQL 2세대
개요
액세스 제어
데이터베이스
백업
복제본
작업
MySQL 데이터베이스
데이터베이스 만들기
이름
문자 세트
대조
유형
information_schema
utf8
utf8_general_ci
시스템
mysql
utf8
utf8_general_ci
시스템
performance_schema
utf8
utf8_general_ci
시스템
오후 1:17

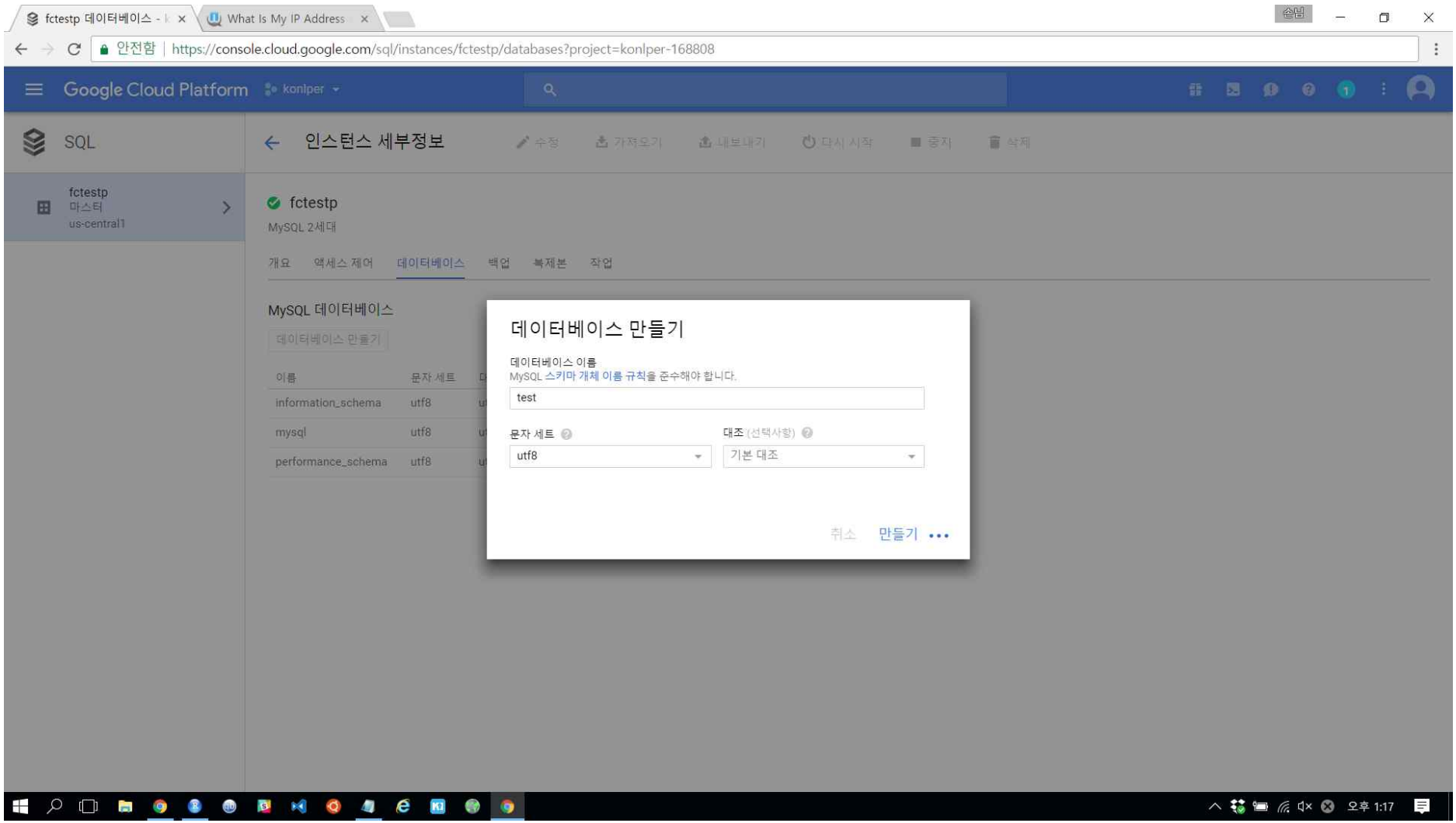
데이터베이스 만들기


fctestp 데이터베이스 -
What Is My IP Address
손님
안전함 | https://console.cloud.google.com/sql/instances/fctestp/databases?project=konlper-168808
Google Cloud Platform
konlper
SQL
인스턴스 세부정보
수정
가져오기
내보내기
다시 시작
중지
삭제
fctestp
마스터
us-central1
fctestp
MySQL 2세대
개요
액세스 제어
데이터베이스
백업
복제본
작업
MySQL 데이터베이스
데이터베이스 만들기
이름
문자 세트
information_schema
utf8
mysql
utf8
performance_schema
utf8
데이터베이스 만들기
데이터베이스 이름
MySQL 스키마 개체 이름 규칙을 준수해야 합니다.
문자 세트
utf8
대조 (선택사항)
기본 대조
취소
만들기
오후 1:17

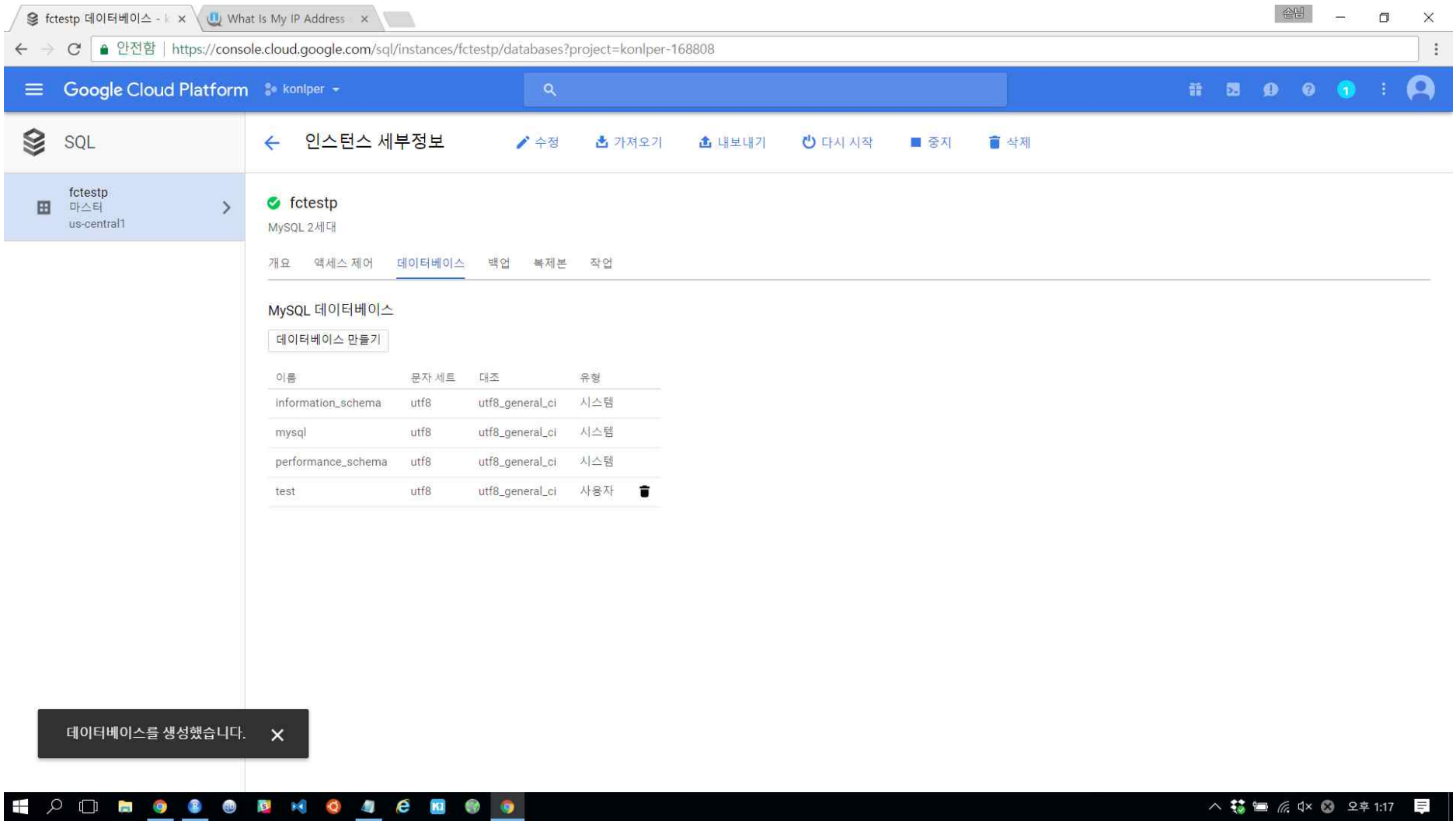
데이터베이스 만들기


fctestp 데이터베이스 -
What Is My IP Address
안전함 | https://console.cloud.google.com/sql/instances/fctestp/databases?project=konlper-168808
Google Cloud Platform
konlper
SQL
인스턴스 세부정보
수정
가져오기
내보내기
다시 시작
중지
삭제
fctestp
마스터
us-central1
fctestp
MySQL 2세대
개요
액세스 제어
데이터베이스
백업
복제본
작업
MySQL 데이터베이스
데이터베이스 만들기
이름
문자 세트
information_schema
utf8
mysql
utf8
performance_schema
utf8
데이터베이스 만들기
데이터베이스 이름
MySQL 스키마 개체 이름 규칙을 준수해야 합니다.
test
문자 세트
utf8
대조 (선택사항)
기본 대조
취소
만들기
오후 1:17

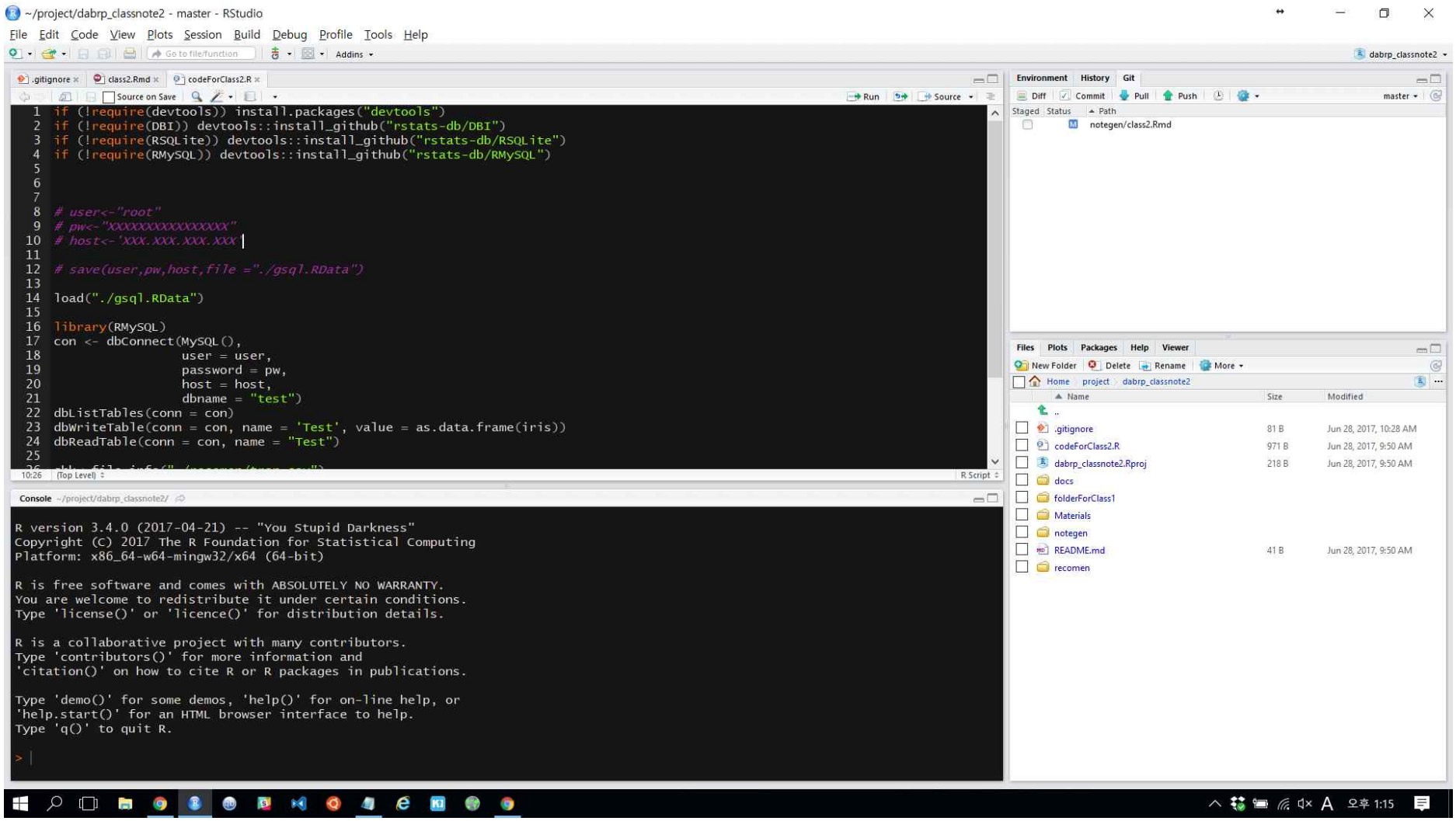
데이터베이스 만들기


fctestp 데이터베이스 -
What Is My IP Address
손님
안전함 | https://console.cloud.google.com/sql/instances/fctestp/databases?project=konlper-168808
Google Cloud Platform
konlper
SQL
인스턴스 세부정보
수정
가져오기
내보내기
다시 시작
중지
삭제
fctestp
마스터
us-central1
fctestp
MySQL 2세대
개요
액세스 제어
데이터베이스
백업
복제본
작업
MySQL 데이터베이스
데이터베이스 만들기
이름
문자 세트
information_schema
utf8
mysql
utf8
performance_schema
utf8
데이터베이스 만들기
데이터베이스 이름
MySQL 스키마 개체 이름 규칙을 준수해야 합니다.
test
문자 세트
utf8
대조 (선택사항)
기본 대조
취소
만들기
오후 1:17

host와 dbname 설정



실습 진행

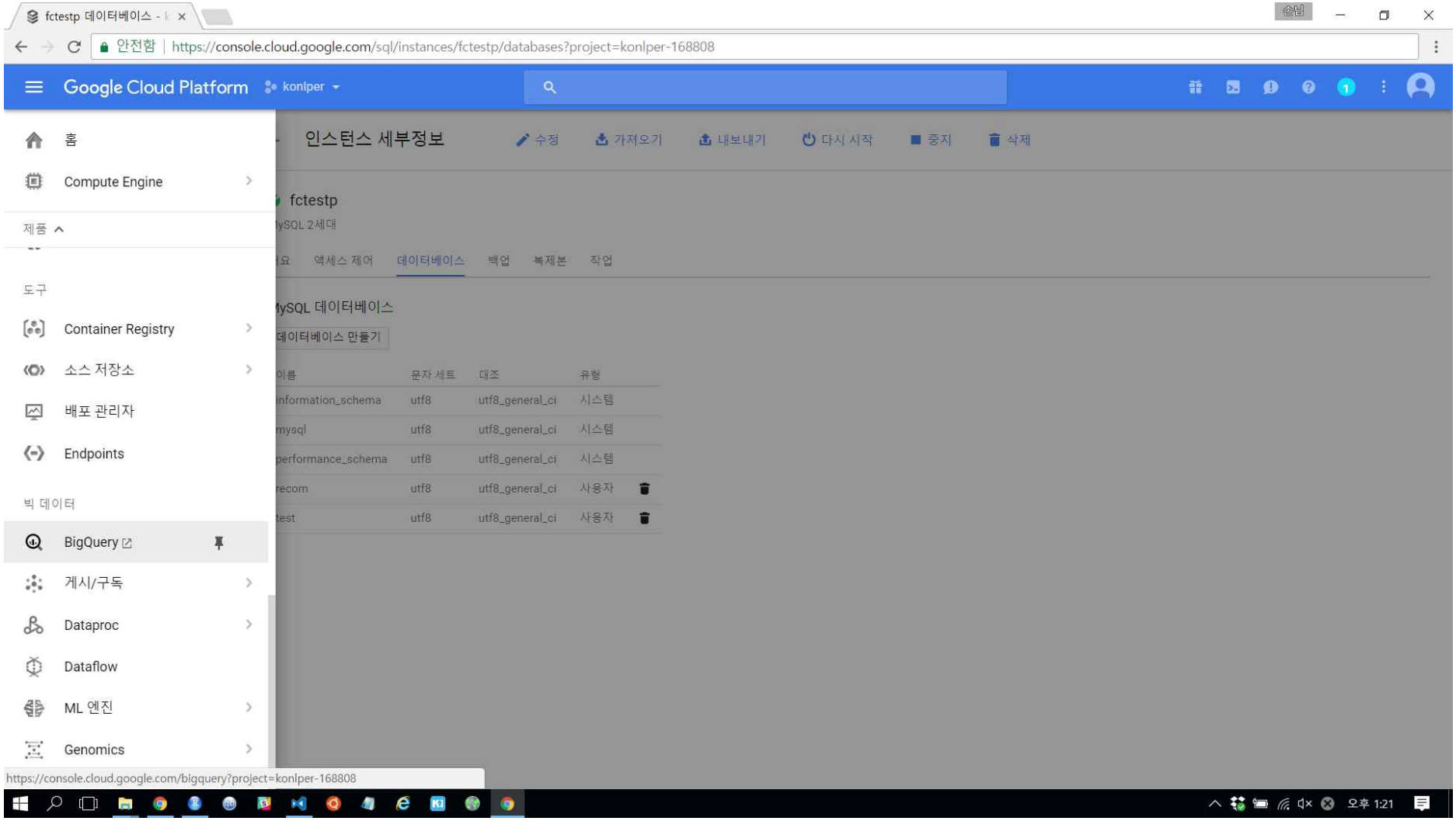

~/project/dabrp_classnote2 - master - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
.gitignore
class2.Rmd
codeForClass2.R
if (!require(devtools)) install.packages("devtools")
if (!require(DBI)) devtools::install_github("rstats-db/DBI")
if (!require(RSQLite)) devtools::install_github("rstats-db/RSQLite")
if (!require(RMySQL)) devtools::install_github("rstats-db/RMySQL")
# user<-"root"
# pw<-"XXXXXXXXXXXXXXX"
# host<-'XXX.XXX.XXX.XXX'
# save(user,pw,host,file ="./gsql.RData")
load("./gsql.RData")
library(RMySQL)
con <- dbConnect(MySQL(),
user = user,
password = pw,
host = host,
dbname = "test")
dbListTables(conn = con)
dbWriteTable(conn = con, name = 'Test', value = as.data.frame(iris))
dbReadTable(conn = con, name = "Test")
Console ~/project/dabrp_classnote2/
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
Environment History Git
notegen/class2.Rmd
Files Plots Packages Help Viewer
.gitignore 81 B Jun 28, 2017, 10:28 AM
codeForClass2.R 971 B Jun 28, 2017, 9:50 AM
dabrp_classnote2.Rproj 218 B Jun 28, 2017, 9:50 AM
docs
folderForClass1
Materials
notegen
README.md 41 B Jun 28, 2017, 9:50 AM
recomen
오후 1:15

R bigquery 연결

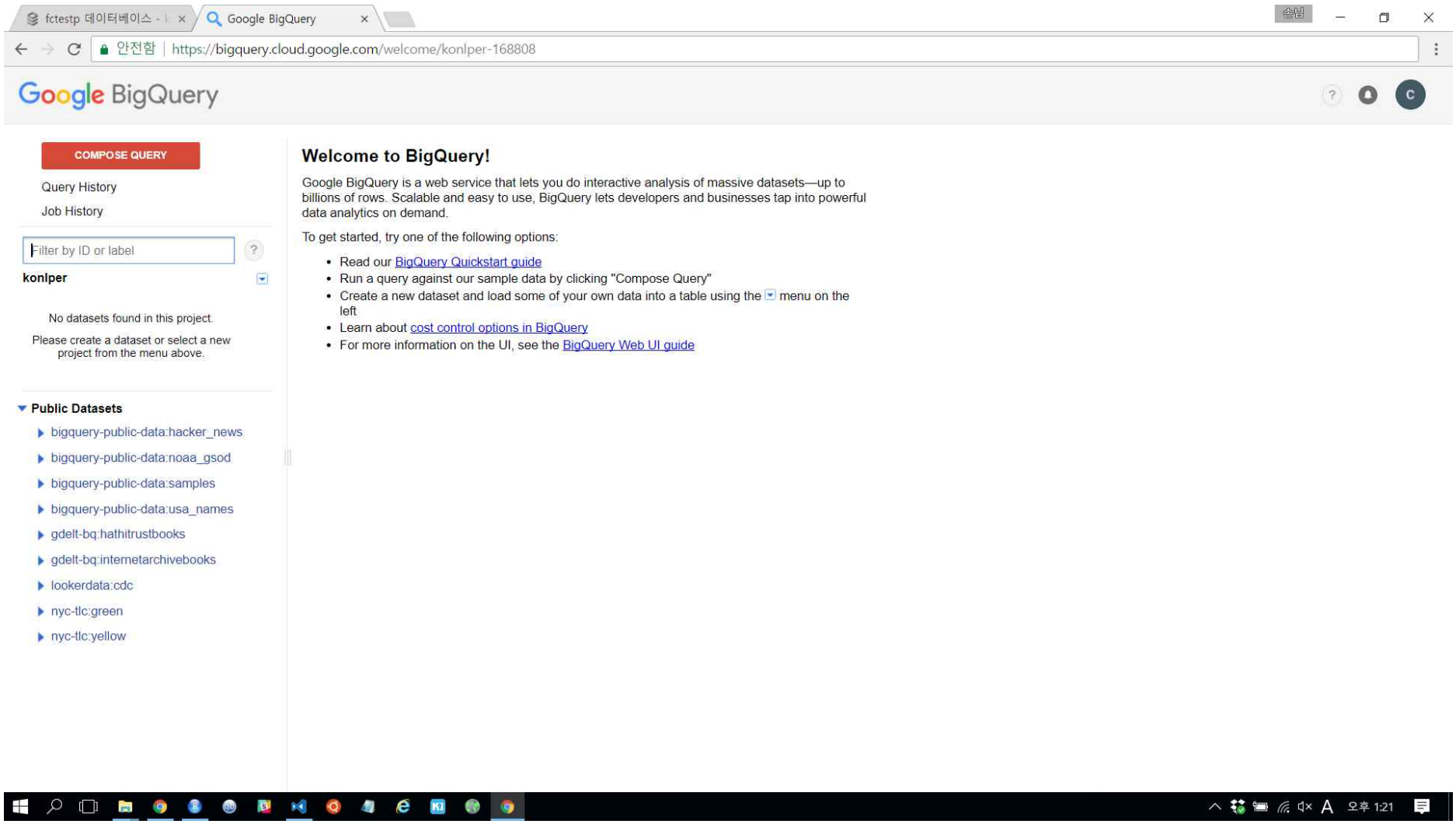
query_exec가 첫 실행이면 브라우저에서 권한 확인을 합니다.

```
library(bigrquery)
project <- "konlper-168808"
sql <- "SELECT * FROM [konlper-168808:recom.chennel] LIMIT 5"
query_exec(sql, project = project)
```

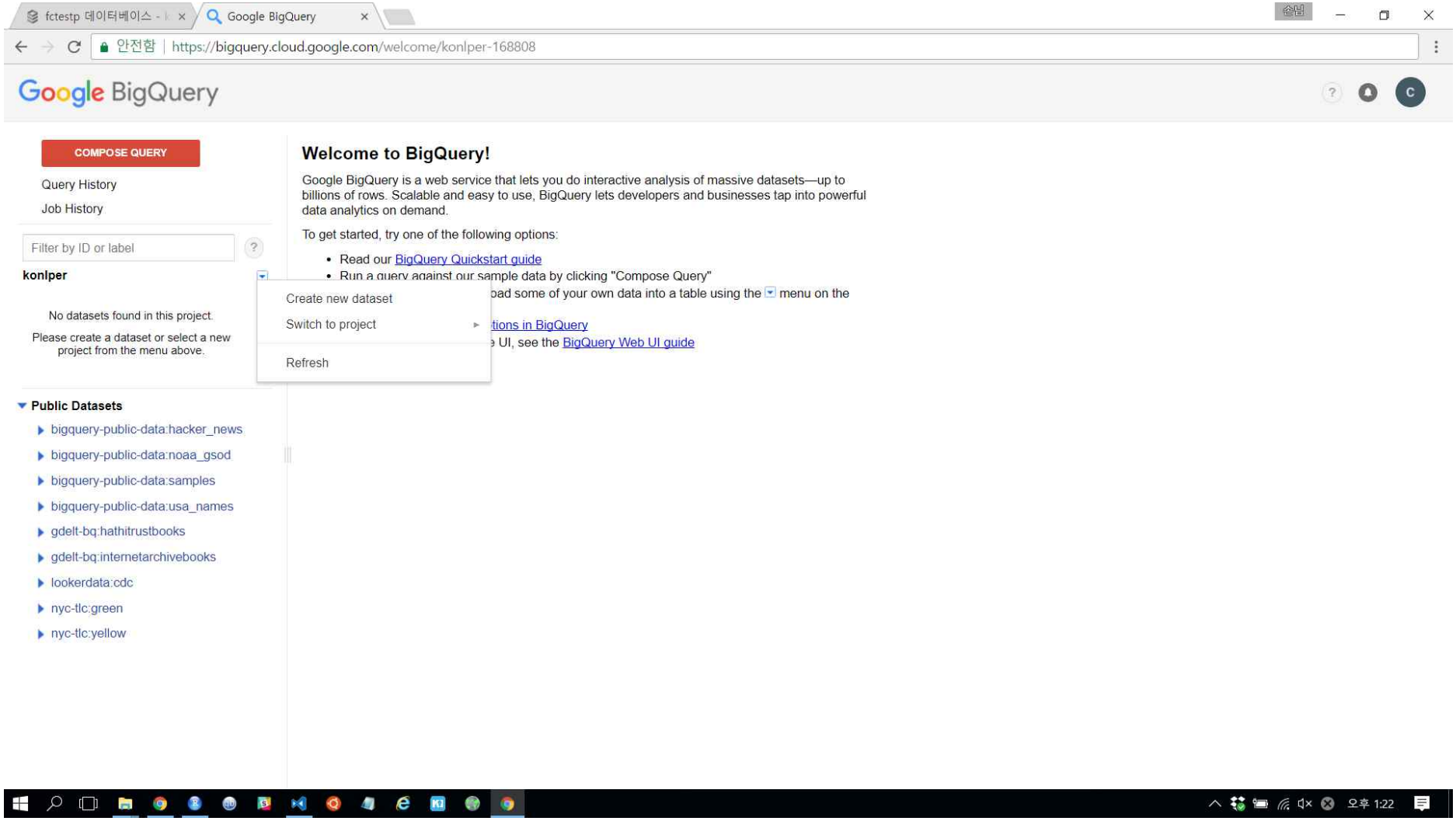
bigquery 인스턴스 생성



bigquery 첫 화면



dataset 만들기


fctestp 데이터베이스 -
Google BigQuery
손님
안전함 | https://bigquery.cloud.google.com/welcome/konlper-168808
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
No datasets found in this project.
Please create a dataset or select a new project from the menu above.
Create new dataset
Switch to project
Refresh
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
Welcome to BigQuery!
Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.
To get started, try one of the following options:
Read our BigQuery Quickstart guide
menu on the
tions in BigQuery
UI, see the BigQuery Web UI guide
오후 1:22

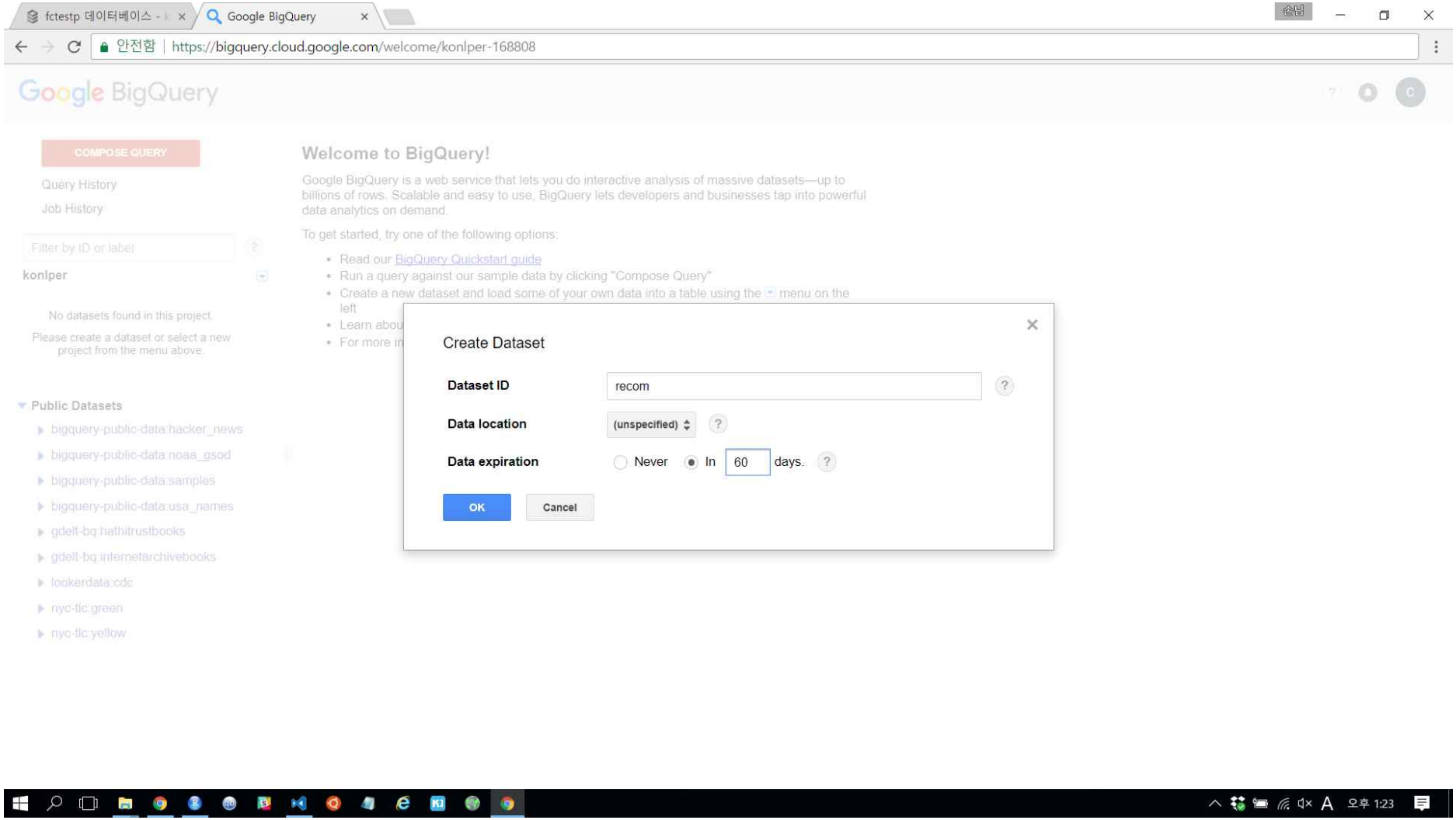
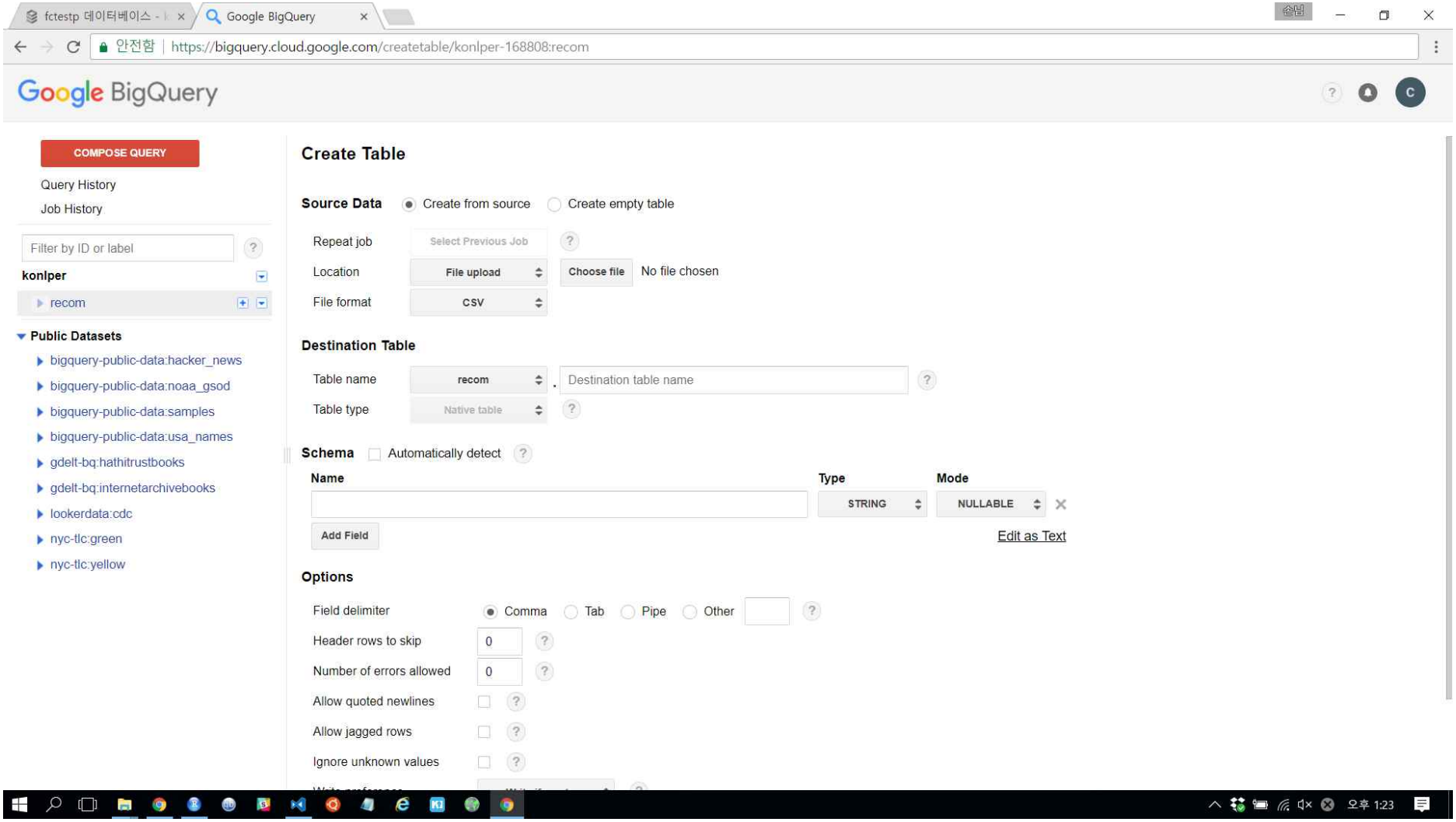
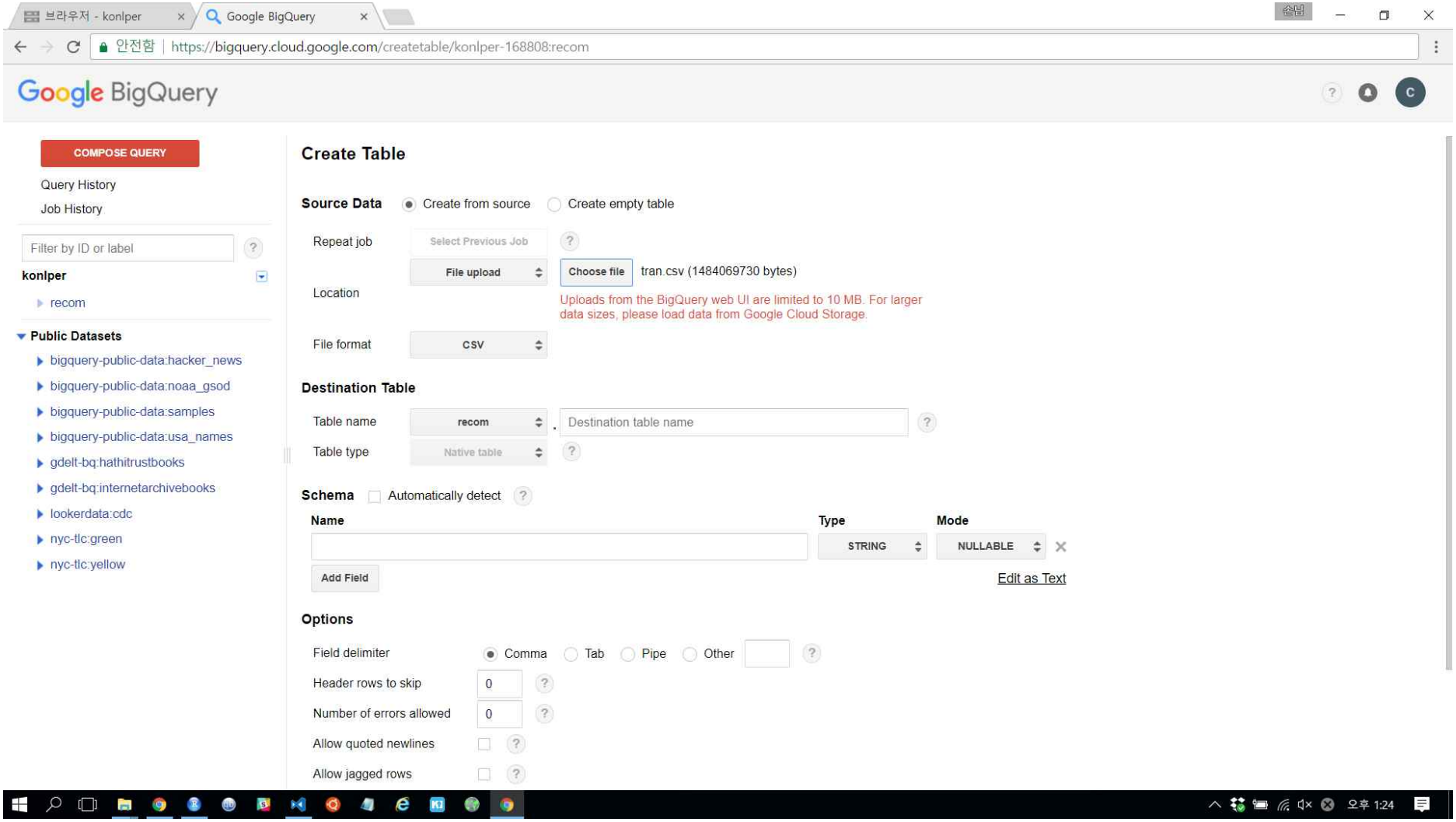
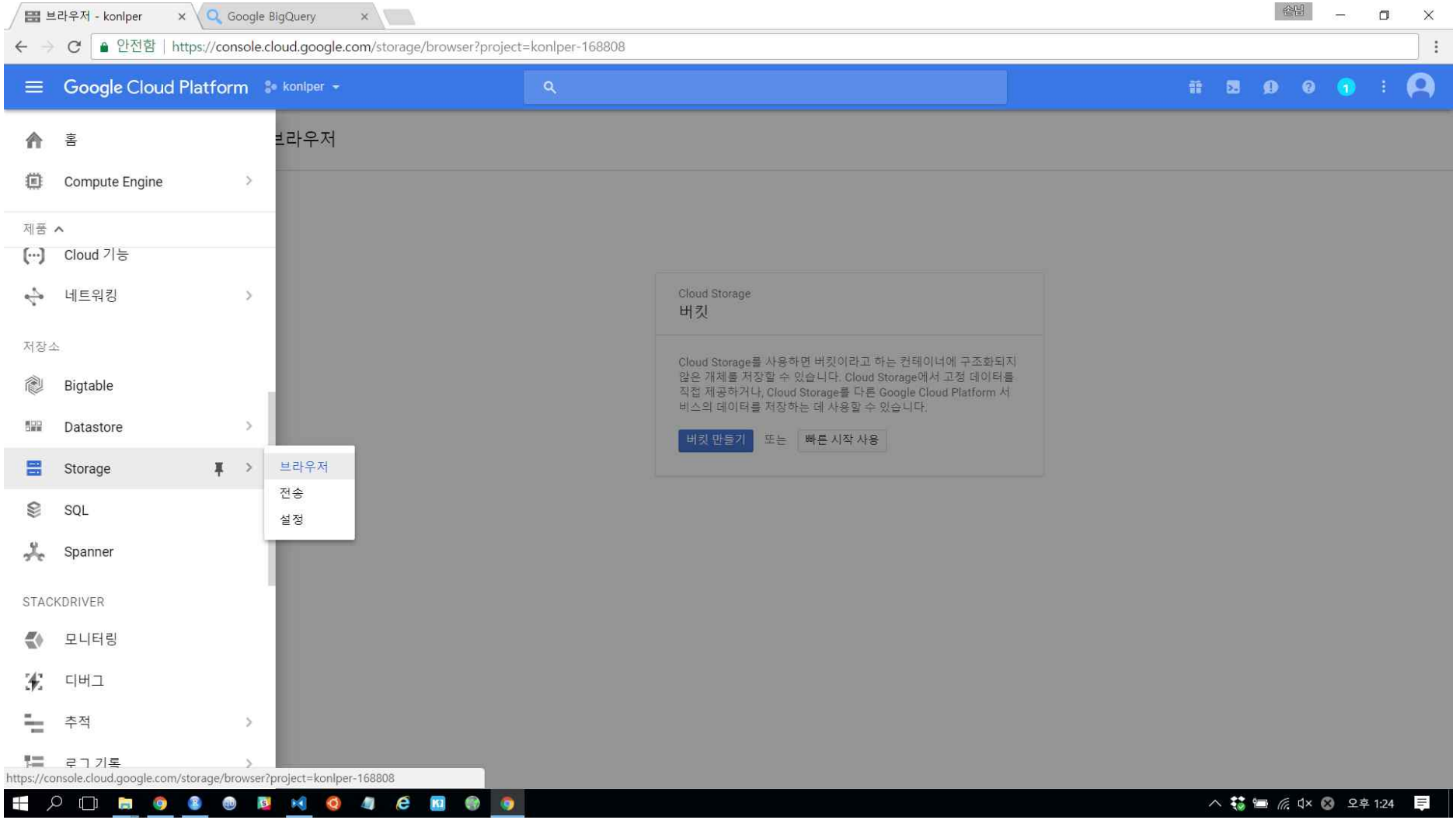

fctestp 데이터베이스 -
Google BigQuery
손님
안전함 | https://bigquery.cloud.google.com/welcome/konlper-168808
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
No datasets found in this project.
Please create a dataset or select a new project from the menu above.
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
Welcome to BigQuery!
Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.
To get started, try one of the following options:
Read our BigQuery Quickstart guide
Run a query against our sample data by clicking "Compose Query"
Create a new dataset and load some of your own data into a table using the menu on the left
Learn abou
For more in
Create Dataset
Dataset ID
recom
Data location
(unspecified)
Data expiration
Never
In
60
days.
OK
Cancel
오후 1:23

table 생성

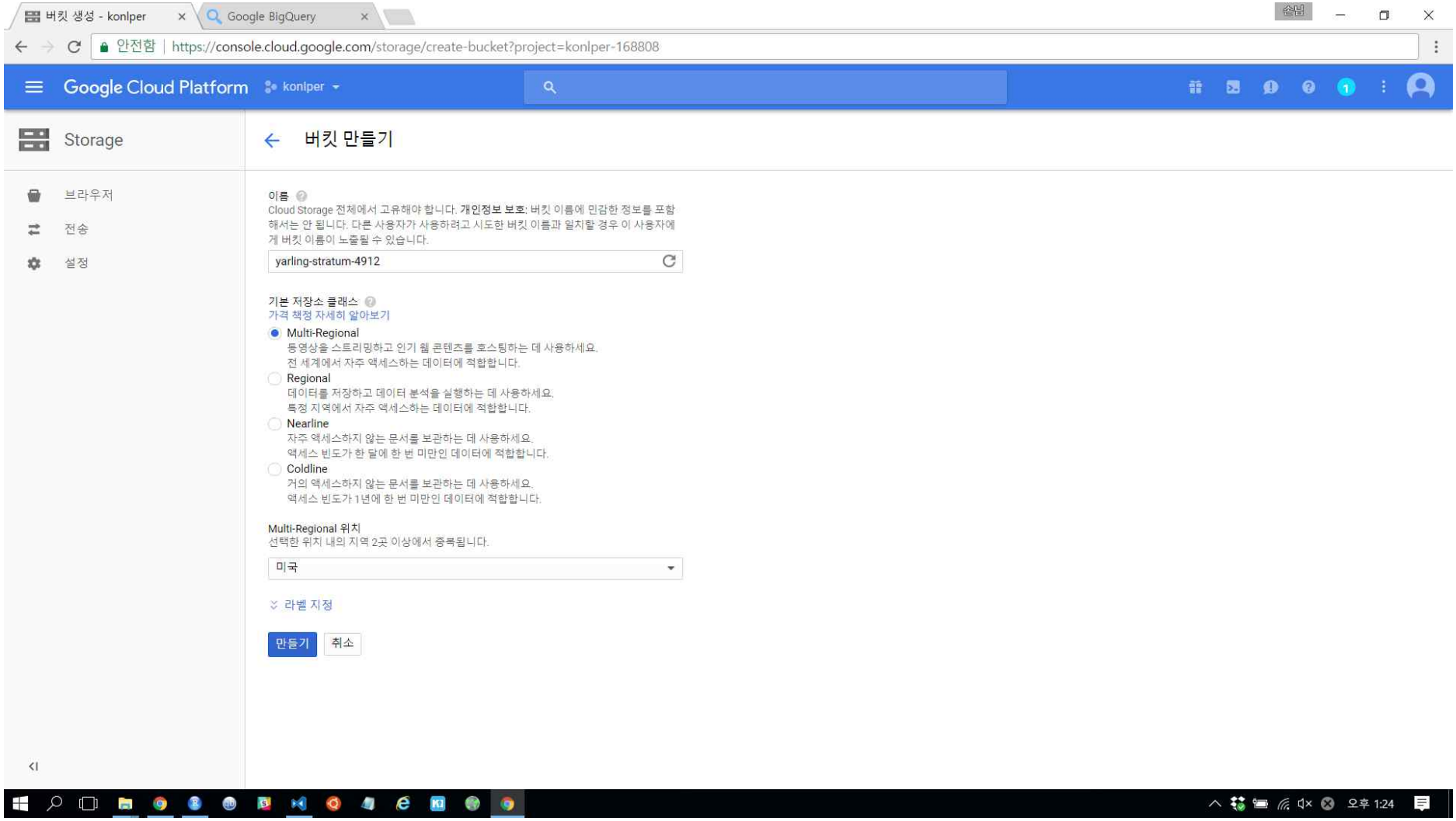

fctestp 데이터베이스 -
Google BigQuery
손님
안전함 | https://bigquery.cloud.google.com/createtable/konlper-168808:recom
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
Create Table
Source Data
Create from source
Create empty table
Repeat job
Select Previous Job
Location
File upload
Choose file
No file chosen
File format
CSV
Destination Table
Table name
recom
Destination table name
Table type
Native table
Schema
Automatically detect
Name
Type
Mode
STRING
NULLABLE
Add Field
Edit as Text
Options
Field delimiter
Comma
Tab
Pipe
Other
Header rows to skip
0
Number of errors allowed
0
Allow quoted newlines
Allow jagged rows
Ignore unknown values
오후 1:23

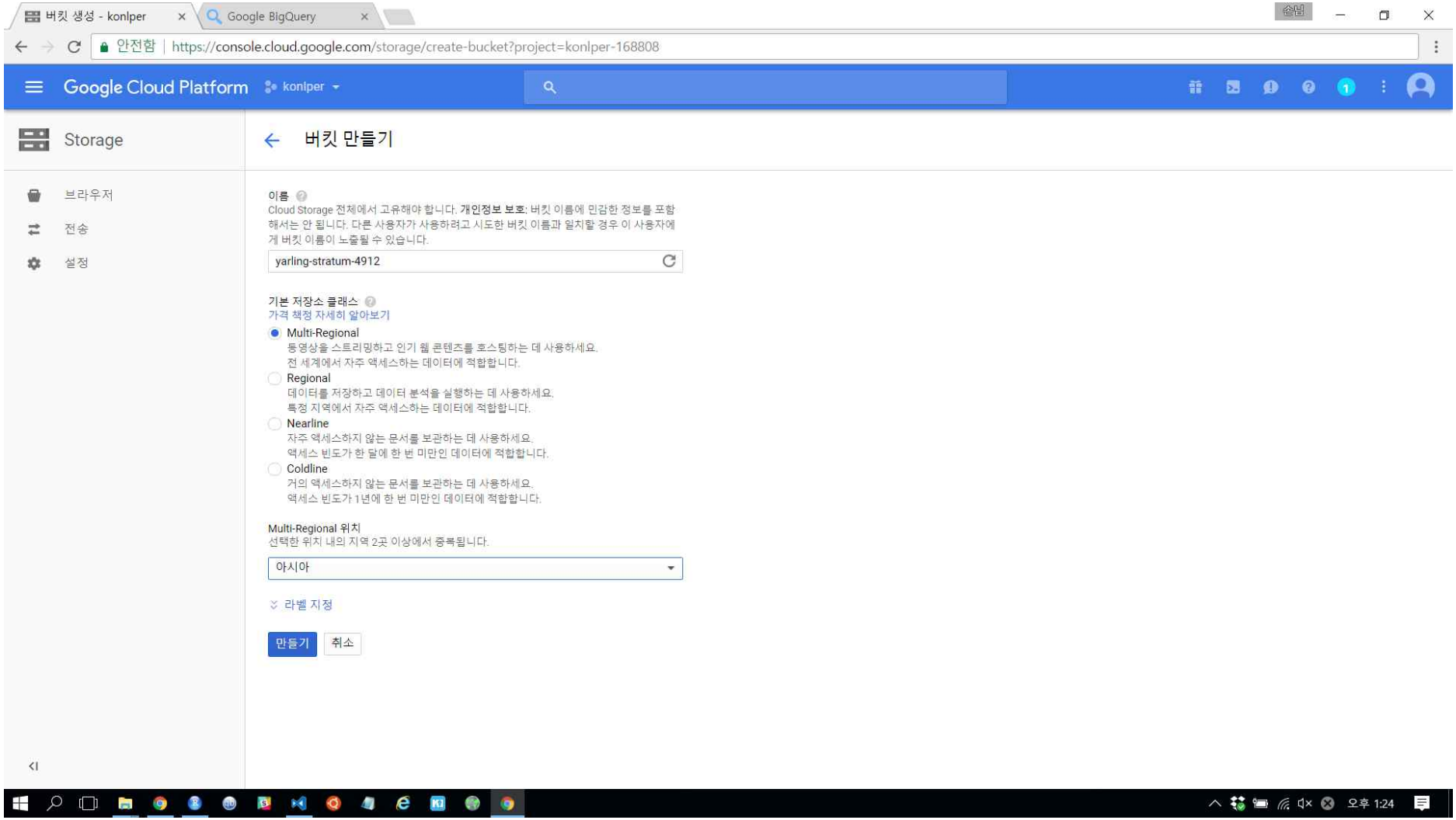
큰 데이터는 storage를 통해 업로드



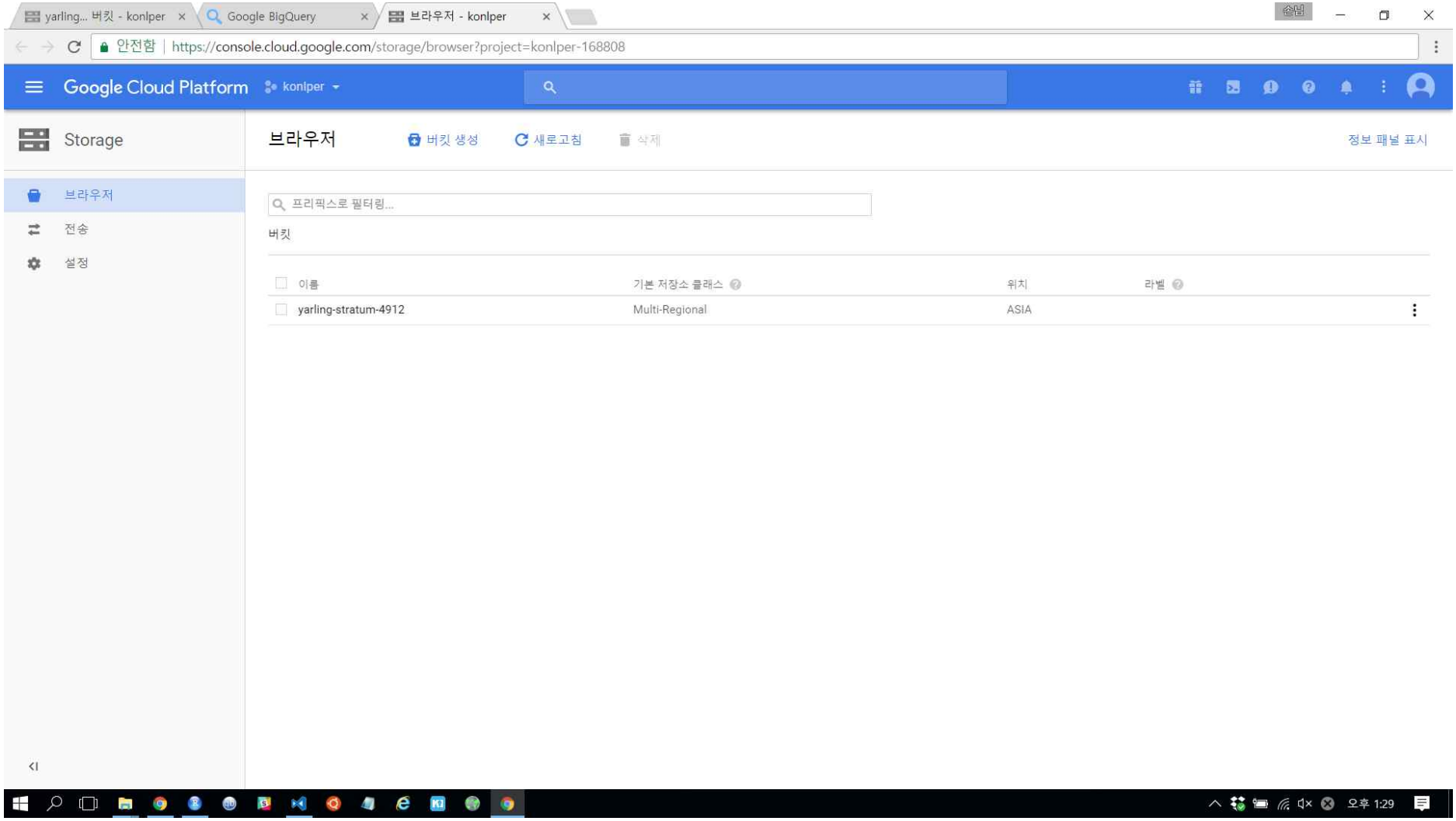

브라우저 - konlper
Google BigQuery
손님
안전함 | https://console.cloud.google.com/storage/browser?project=konlper-168808
Google Cloud Platform
konlper
홈
Compute Engine
제품
Cloud 기능
네트워킹
저장소
Bigtable
Datastore
Storage
SQL
Spanner
STACKDRIVER
모니터링
디버그
추적
로그 기록
브라우저
Cloud Storage
버킷
Cloud Storage를 사용하면 버킷이라고 하는 컨테이너에 구조화되지 않은 개체를 저장할 수 있습니다. Cloud Storage에서 고정 데이터를 직접 제공하거나, Cloud Storage를 다른 Google Cloud Platform 서비스의 데이터를 저장하는 데 사용할 수 있습니다.
버킷 만들기
또는
빠른 시작 사용
브라우저
전송
설정
https://console.cloud.google.com/storage/browser?project=konlper-168808
오후 1:24

버킷 만들기

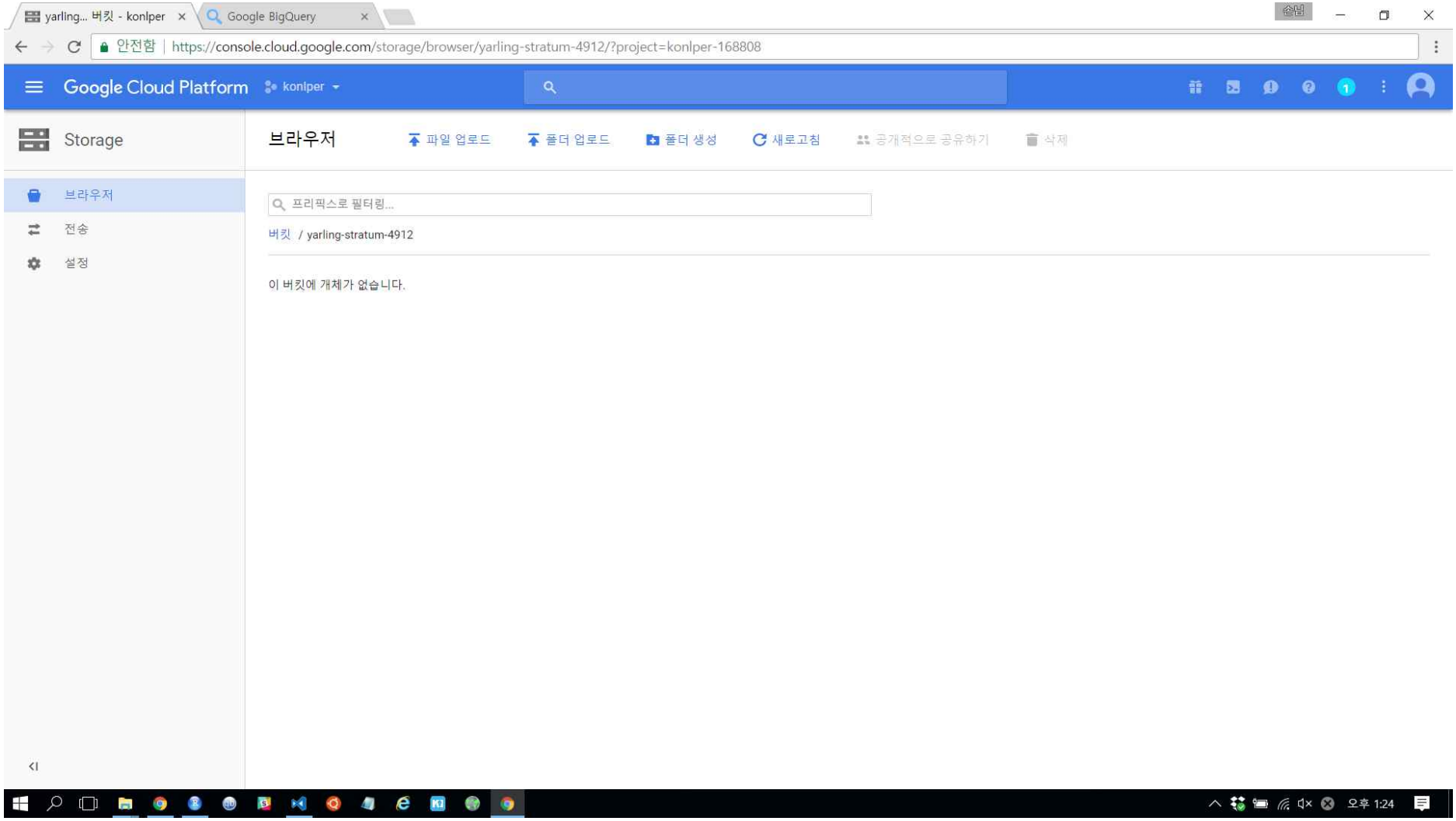

버킷 생성 - konlper
Google BigQuery
손님
안전함 | https://console.cloud.google.com/storage/create-bucket?project=konlper-168808
Google Cloud Platform
konlper
Storage
브라우저
전송
설정
버킷 만들기
이름
Cloud Storage 전체에서 고유해야 합니다. 개인정보 보호: 버킷 이름에 민감한 정보를 포함해서는 안 됩니다. 다른 사용자가 사용하려고 시도한 버킷 이름과 일치할 경우 이 사용자에게 버킷 이름이 노출될 수 있습니다.
yarling-stratum-4912
기본 저장소 클래스
가격 책정 자세히 알아보기
Multi-Regional
동영상을 스트리밍하고 인기 웹 콘텐츠를 호스팅하는 데 사용하세요.
전 세계에서 자주 액세스하는 데이터에 적합합니다.
Regional
데이터를 저장하고 데이터 분석을 실행하는 데 사용하세요.
특정 지역에서 자주 액세스하는 데이터에 적합합니다.
Nearline
자주 액세스하지 않는 문서를 보관하는 데 사용하세요.
액세스 빈도가 한 달에 한 번 미만인 데이터에 적합합니다.
Coldline
거의 액세스하지 않는 문서를 보관하는 데 사용하세요.
액세스 빈도가 1년에 한 번 미만인 데이터에 적합합니다.
Multi-Regional 위치
선택한 위치 내의 지역 2곳 이상에서 중복됩니다.
미국
라벨 지정
만들기
취소
오후 1:24


버킷 생성 - konlper
Google BigQuery
손님
안전함 | https://console.cloud.google.com/storage/create-bucket?project=konlper-168808
Google Cloud Platform
konlper
Storage
버킷 만들기
브라우저
전송
설정
이름
Cloud Storage 전체에서 고유해야 합니다. 개인정보 보호: 버킷 이름에 민감한 정보를 포함해서는 안 됩니다. 다른 사용자가 사용하려고 시도한 버킷 이름과 일치할 경우 이 사용자에게 버킷 이름이 노출될 수 있습니다.
yarling-stratum-4912
기본 저장소 클래스
가격 책정 자세히 알아보기
Multi-Regional
동영상을 스트리밍하고 인기 웹 콘텐츠를 호스팅하는 데 사용하세요.
전 세계에서 자주 액세스하는 데이터에 적합합니다.
Regional
데이터를 저장하고 데이터 분석을 실행하는 데 사용하세요.
특정 지역에서 자주 액세스하는 데이터에 적합합니다.
Nearline
자주 액세스하지 않는 문서를 보관하는 데 사용하세요.
액세스 빈도가 한 달에 한 번 미만인 데이터에 적합합니다.
Coldline
거의 액세스하지 않는 문서를 보관하는 데 사용하세요.
액세스 빈도가 1년에 한 번 미만인 데이터에 적합합니다.
Multi-Regional 위치
선택한 위치 내의 지역 2곳 이상에서 중복됩니다.
아시아
라벨 지정
만들기
취소
오후 1:24

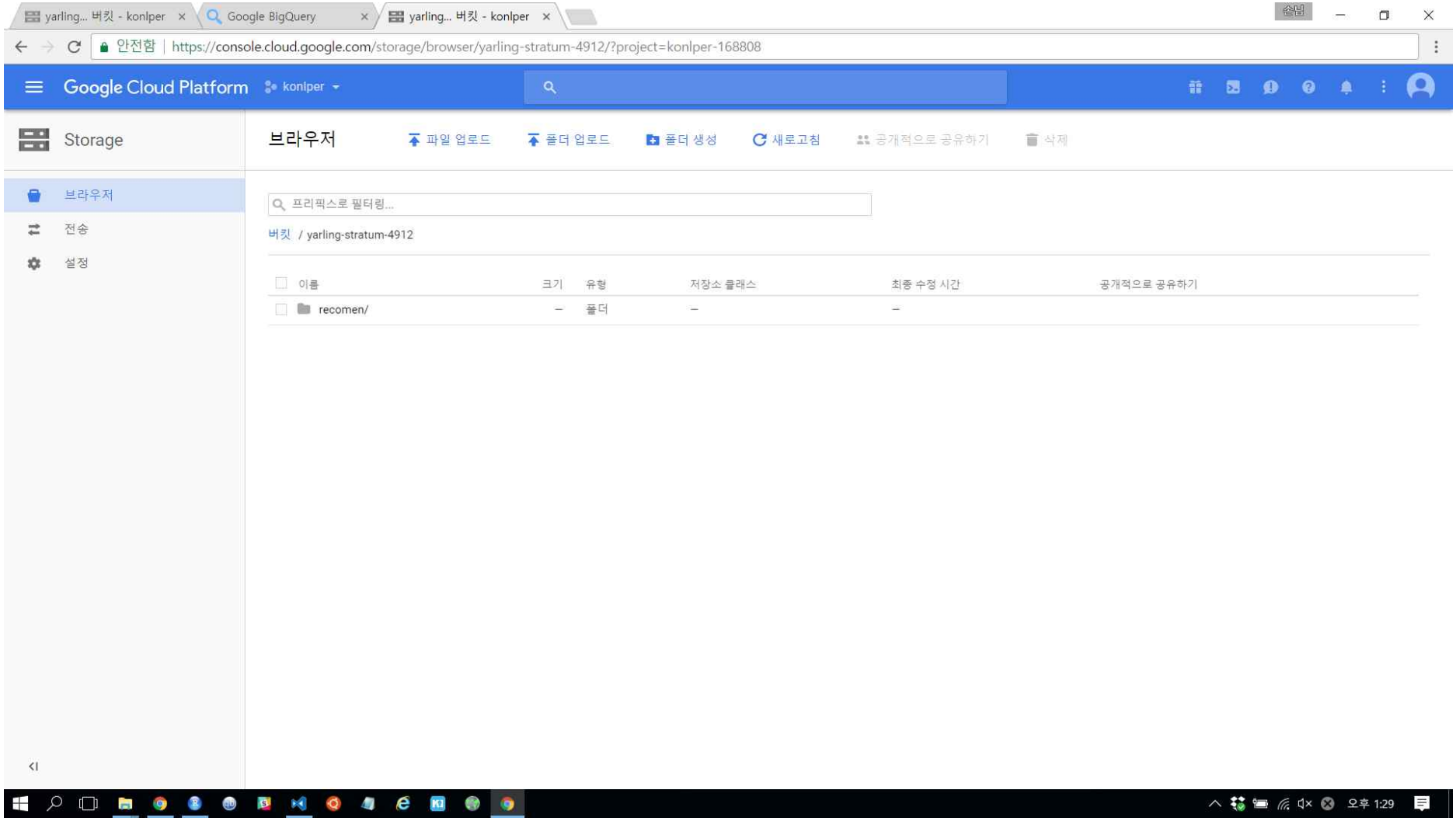
버킷 확인


yarling... 버킷 - konlper
Google BigQuery
브라우저 - konlper
손님
안전함 | https://console.cloud.google.com/storage/browser?project=konlper-168808
Google Cloud Platform
konlper
Storage
브라우저
버킷 생성
새로고침
삭제
정보 패널 표시
브라우저
전송
설정
프리픽스로 필터링...
버킷
이름
기본 저장소 클래스
위치
라벨
yarling-stratum-4912
Multi-Regional
ASIA
오후 1:29

폴더 업로드

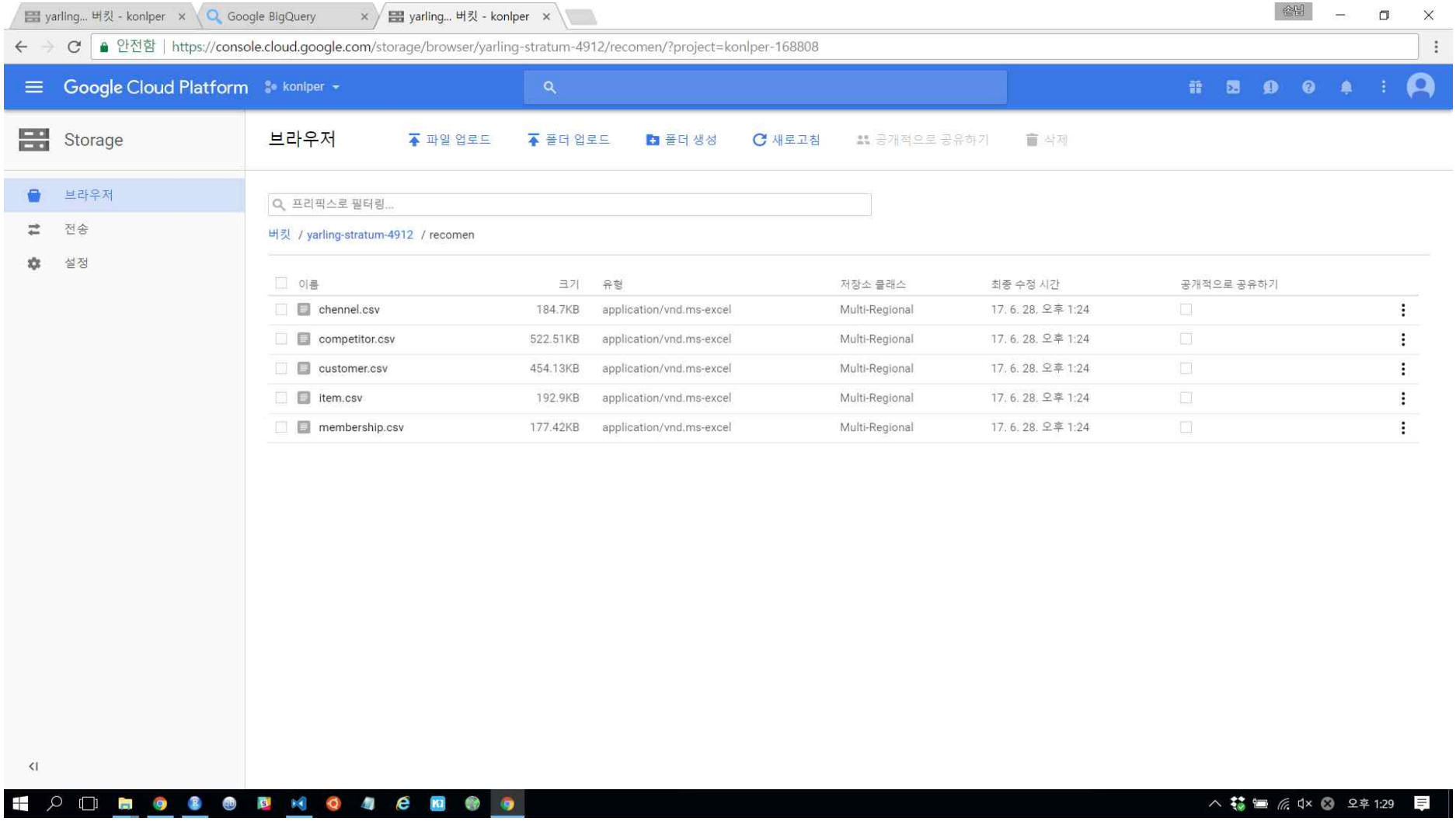

yarling... 버킷 - konlper
Google BigQuery
손님
안전함 | https://console.cloud.google.com/storage/browser/yarling-stratum-4912/?project=konlper-168808
Google Cloud Platform
konlper
Storage
브라우저
파일 업로드
폴더 업로드
폴더 생성
새로고침
공개적으로 공유하기
삭제
브라우저
전송
설정
프리픽스로 필터링...
버킷 / yarling-stratum-4912
이 버킷에 개체가 없습니다.
오후 1:24

업로드 확인


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷 - konlper
손님
안전함 | https://console.cloud.google.com/storage/browser/yarling-stratum-4912/?project=konlper-168808
Google Cloud Platform
konlper
Storage
브라우저
파일 업로드
폴더 업로드
폴더 생성
새로고침
공개적으로 공유하기
삭제
브라우저
전송
설정
프리픽스로 필터링...
버킷 / yarling-stratum-4912
이름
크기
유형
저장소 클래스
최종 수정 시간
공개적으로 공유하기
recomen/
–
폴더
–
–
오후 1:29

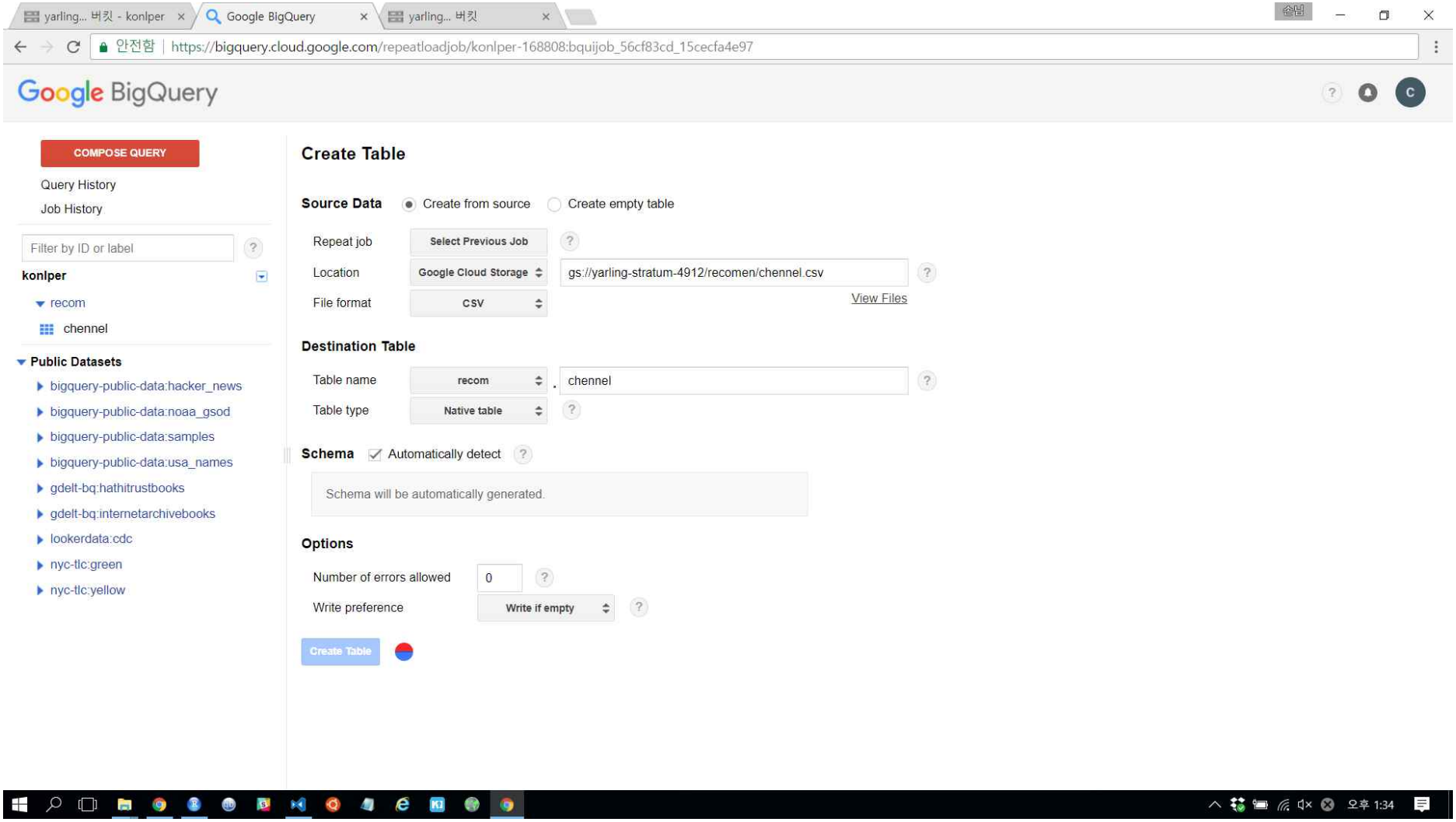
77/92

업로드 확인

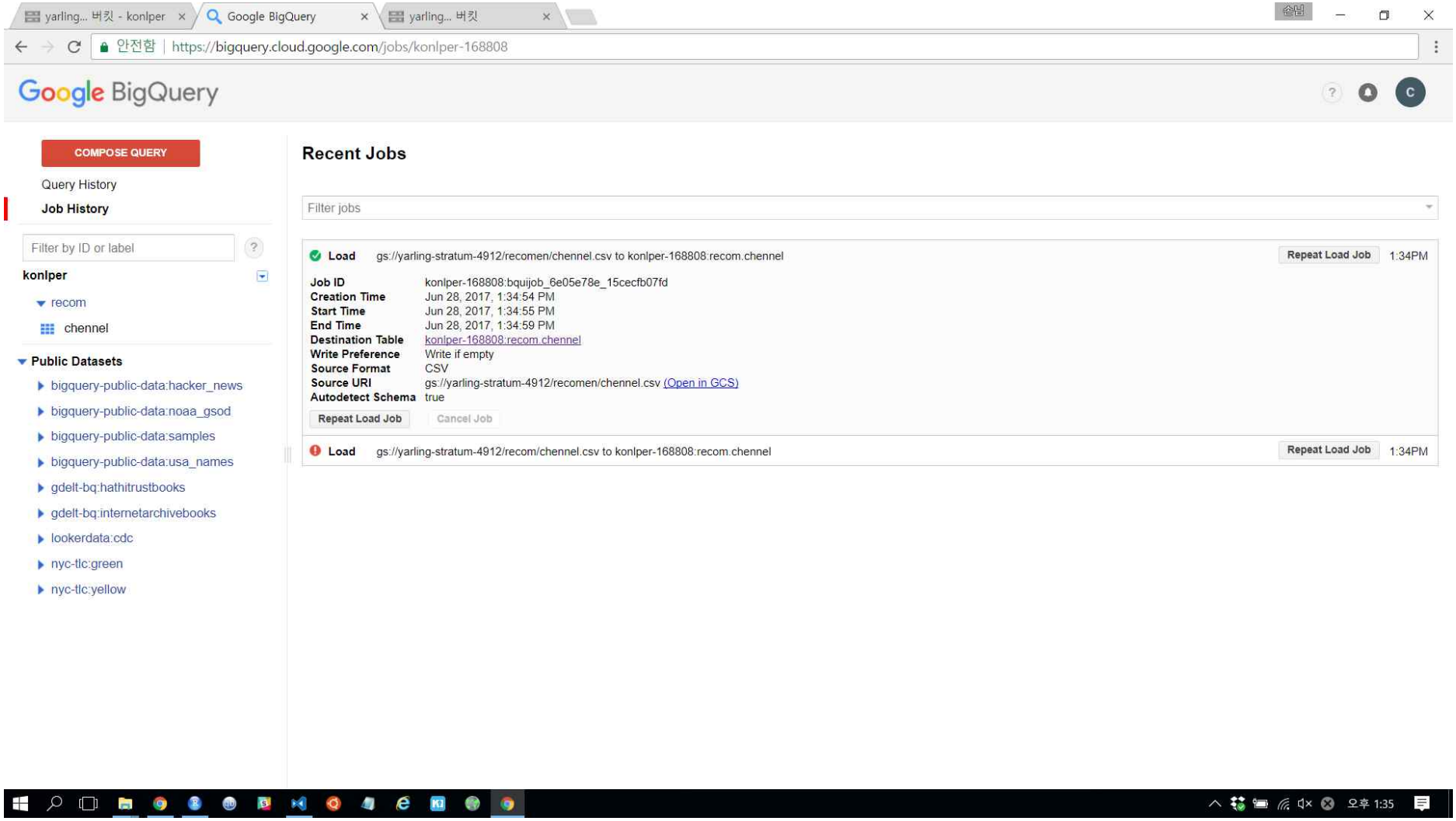

yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷 - konlper
손님
안전함 | https://console.cloud.google.com/storage/browser/yarling-stratum-4912/recomen/?project=konlper-168808
Google Cloud Platform
konlper
Storage
브라우저
전송
설정
브라우저
파일 업로드
폴더 업로드
폴더 생성
새로고침
공개적으로 공유하기
삭제
프리픽스로 필터링...
버킷 / yarling-stratum-4912 / recomen
이름 크기 유형 저장소 클래스 최종 수정 시간 공개적으로 공유하기
chennel.csv 184.7KB application/vnd.ms-excel Multi-Regional 17. 6. 28. 오후 1:24
competitor.csv 522.51KB application/vnd.ms-excel Multi-Regional 17. 6. 28. 오후 1:24
customer.csv 454.13KB application/vnd.ms-excel Multi-Regional 17. 6. 28. 오후 1:24
item.csv 192.9KB application/vnd.ms-excel Multi-Regional 17. 6. 28. 오후 1:24
membership.csv 177.42KB application/vnd.ms-excel Multi-Regional 17. 6. 28. 오후 1:24
오후 1:29

78/92

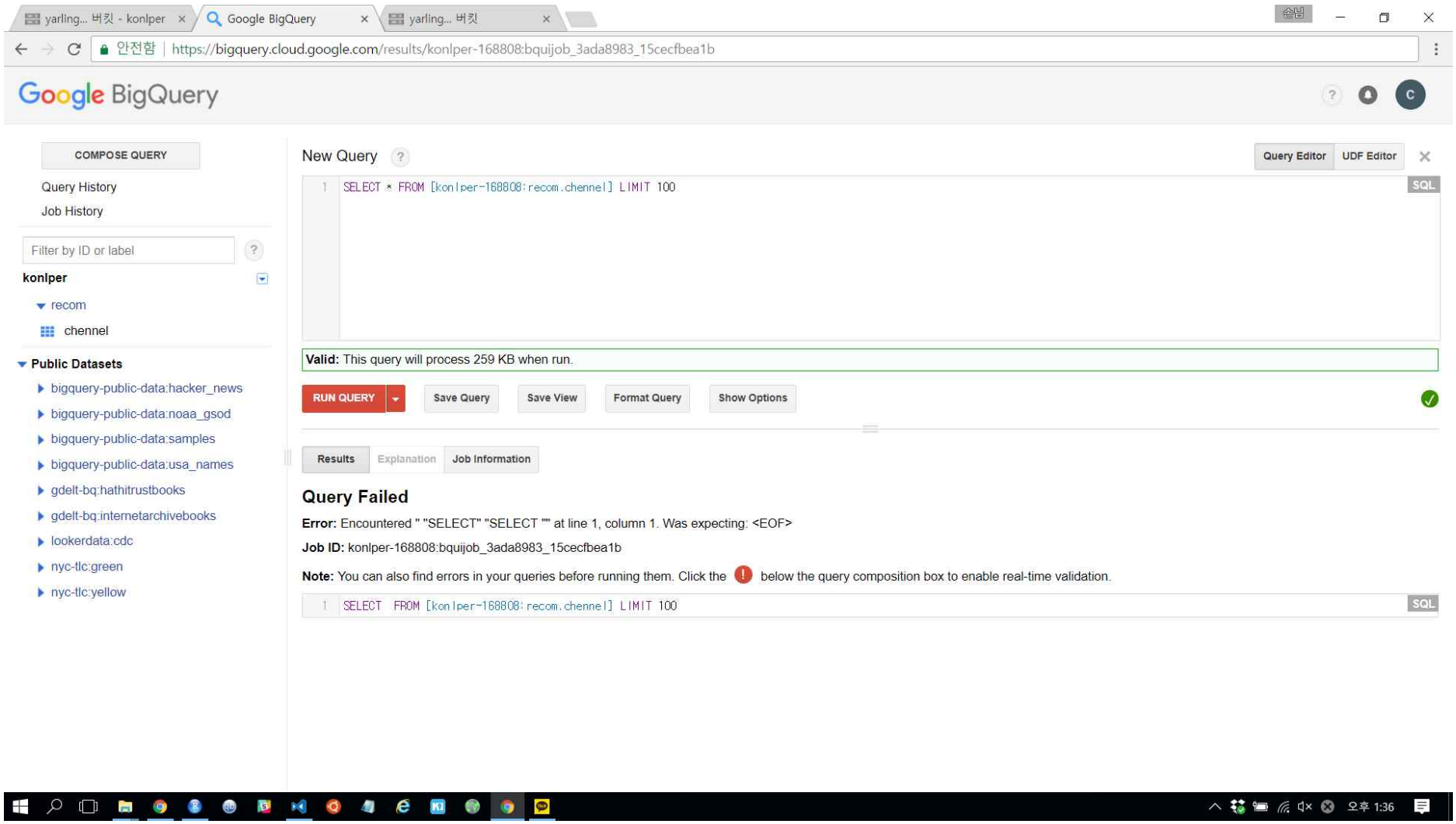
gs:// 경로확인


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
손님
안전함 | https://bigquery.cloud.google.com/repeatloadjob/konlper-168808:bquijob_56cf83cd_15cecfa4e97
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
Create Table
Source Data
Create from source
Create empty table
Repeat job
Select Previous Job
Location
Google Cloud Storage
gs://yarling-stratum-4912/recomen/chennel.csv
File format
CSV
View Files
Destination Table
Table name
recom
chennel
Table type
Native table
Schema
Automatically detect
Schema will be automatically generated.
Options
Number of errors allowed
0
Write preference
Write if empty
Create Table
오후 1:34

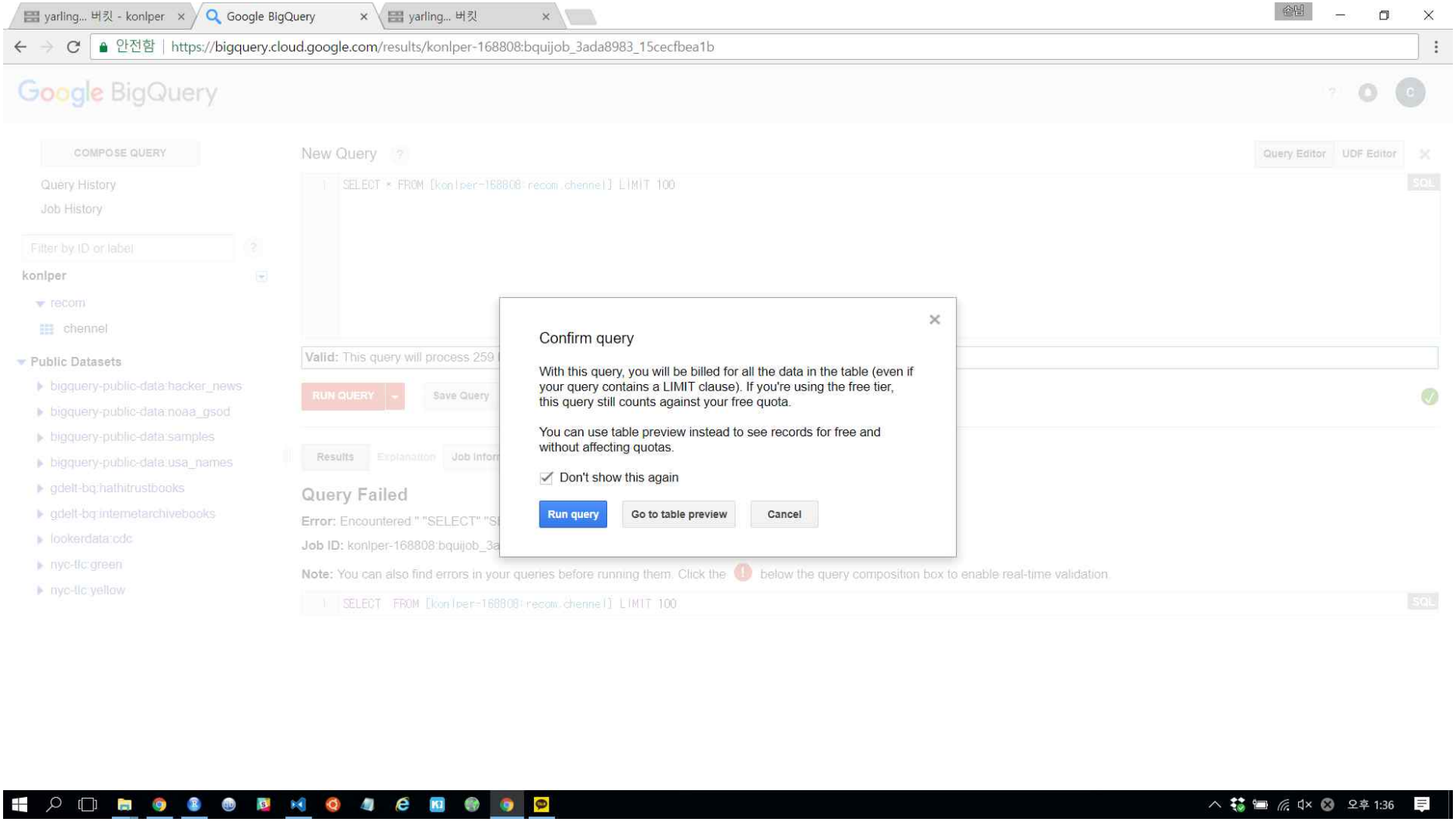
작업 결과 확인


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
손님
안전함 | https://bigquery.cloud.google.com/jobs/konlper-168808
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
Recent Jobs
Filter jobs
Load gs://yarling-stratum-4912/recomen/chennel.csv to konlper-168808:recom.chennel
Repeat Load Job 1:34PM
Job ID konlper-168808:bquijob_6e05e78e_15cecfb07fd
Creation Time Jun 28, 2017, 1:34:54 PM
Start Time Jun 28, 2017, 1:34:55 PM
End Time Jun 28, 2017, 1:34:59 PM
Destination Table konlper-168808:recom.chennel
Write Preference Write if empty
Source Format CSV
Source URI gs://yarling-stratum-4912/recomen/chennel.csv (Open in GCS)
Autodetect Schema true
Repeat Load Job
Cancel Job
Load gs://yarling-stratum-4912/recom/chennel.csv to konlper-168808:recom.chennel
Repeat Load Job 1:34PM
오후 1:35

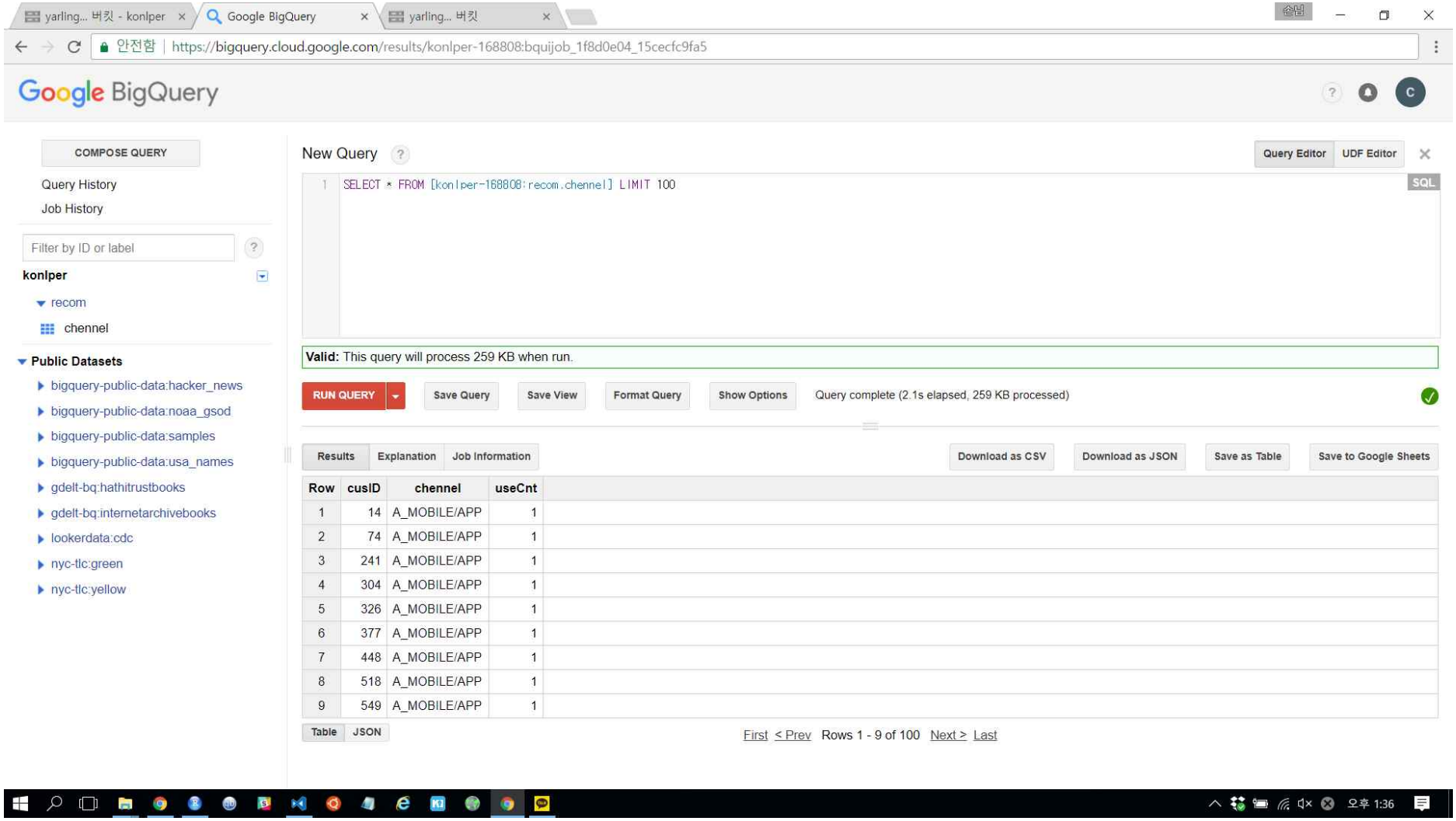
query 실행화면


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
손님
안전함 | https://bigquery.cloud.google.com/results/konlper-168808:bquijob_3ada8983_15cecfbea1b
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
New Query
Query Editor
UDF Editor
SELECT * FROM [konlper-168808:recom.chennel] LIMIT 100
SQL
Valid: This query will process 259 KB when run.
RUN QUERY
Save Query
Save View
Format Query
Show Options
Results
Explanation
Job Information
Query Failed
Error: Encountered " "SELECT" "SELECT "" at line 1, column 1. Was expecting: <EOF>
Job ID: konlper-168808:bquijob_3ada8983_15cecfbea1b
Note: You can also find errors in your queries before running them. Click the below the query composition box to enable real-time validation.
SELECT FROM [konlper-168808:recom.chennel] LIMIT 100
SQL
오후 1:36

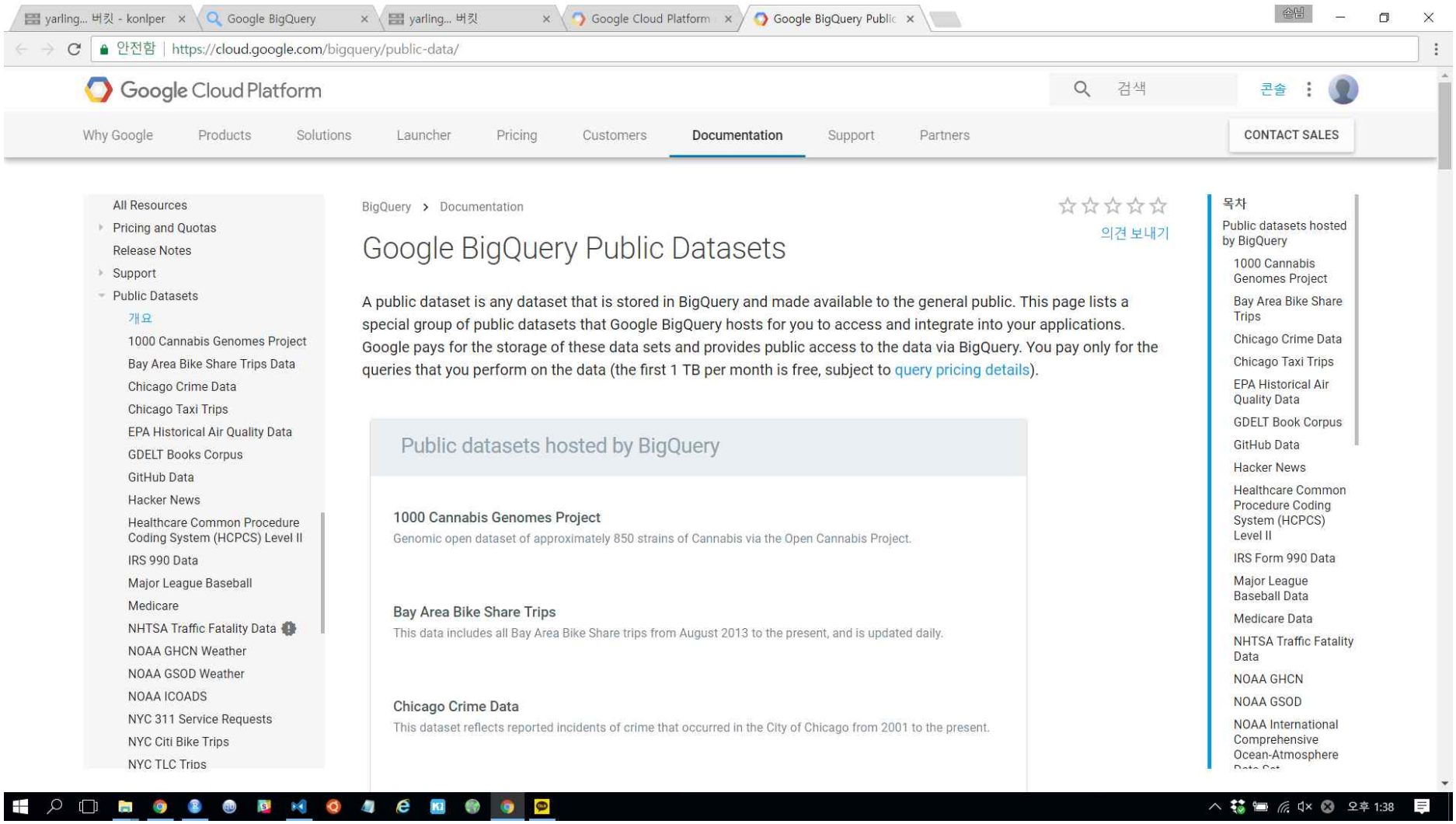
비용 발생 경고


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
안전함 | https://bigquery.cloud.google.com/results/konlper-168808:bquijob_3ada8983_15cecfbea1b
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
New Query
Query Editor
UDF Editor
SELECT * FROM [konlper-168808:recom.chennel] LIMIT 100
Valid: This query will process 259
RUN QUERY
Save Query
Results
Query Failed
Error: Encountered " "SELECT" "
Job ID: konlper-168808:bquijob_3a
Note: You can also find errors in your queries before running them. Click the below the query composition box to enable real-time validation.
SELECT FROM [konlper-168808:recom.chennel] LIMIT 100
Confirm query
With this query, you will be billed for all the data in the table (even if your query contains a LIMIT clause). If you're using the free tier, this query still counts against your free quota.
You can use table preview instead to see records for free and without affecting quotas.
Don't show this again
Run query
Go to table preview
Cancel
오후 1:36

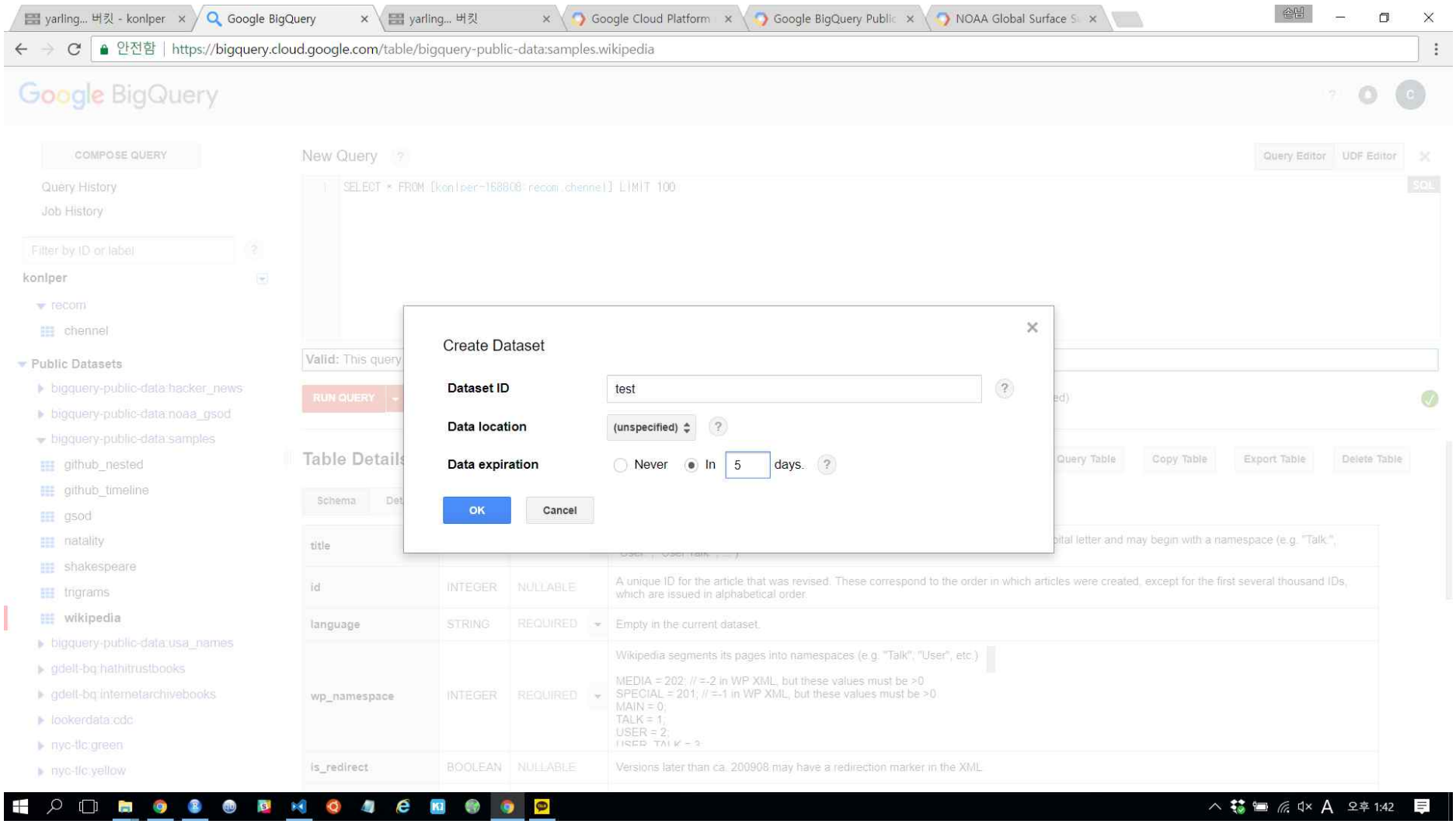
query 결과 확인


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
안전함 | https://bigquery.cloud.google.com/results/konlper-168808:bquijob_1f8d0e04_15cecfc9fa5
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
New Query
Query Editor
UDF Editor
SELECT * FROM [konlper-168808:recom.chennel] LIMIT 100
SQL
Valid: This query will process 259 KB when run.
RUN QUERY
Save Query
Save View
Format Query
Show Options
Query complete (2.1s elapsed, 259 KB processed)
Results
Explanation
Job Information
Download as CSV
Download as JSON
Save as Table
Save to Google Sheets
Row cusID chennel useCnt
1 14 A_MOBILE/APP 1
2 74 A_MOBILE/APP 1
3 241 A_MOBILE/APP 1
4 304 A_MOBILE/APP 1
5 326 A_MOBILE/APP 1
6 377 A_MOBILE/APP 1
7 448 A_MOBILE/APP 1
8 518 A_MOBILE/APP 1
9 549 A_MOBILE/APP 1
Table
JSON
First < Prev Rows 1 - 9 of 100 Next > Last
오후 1:36

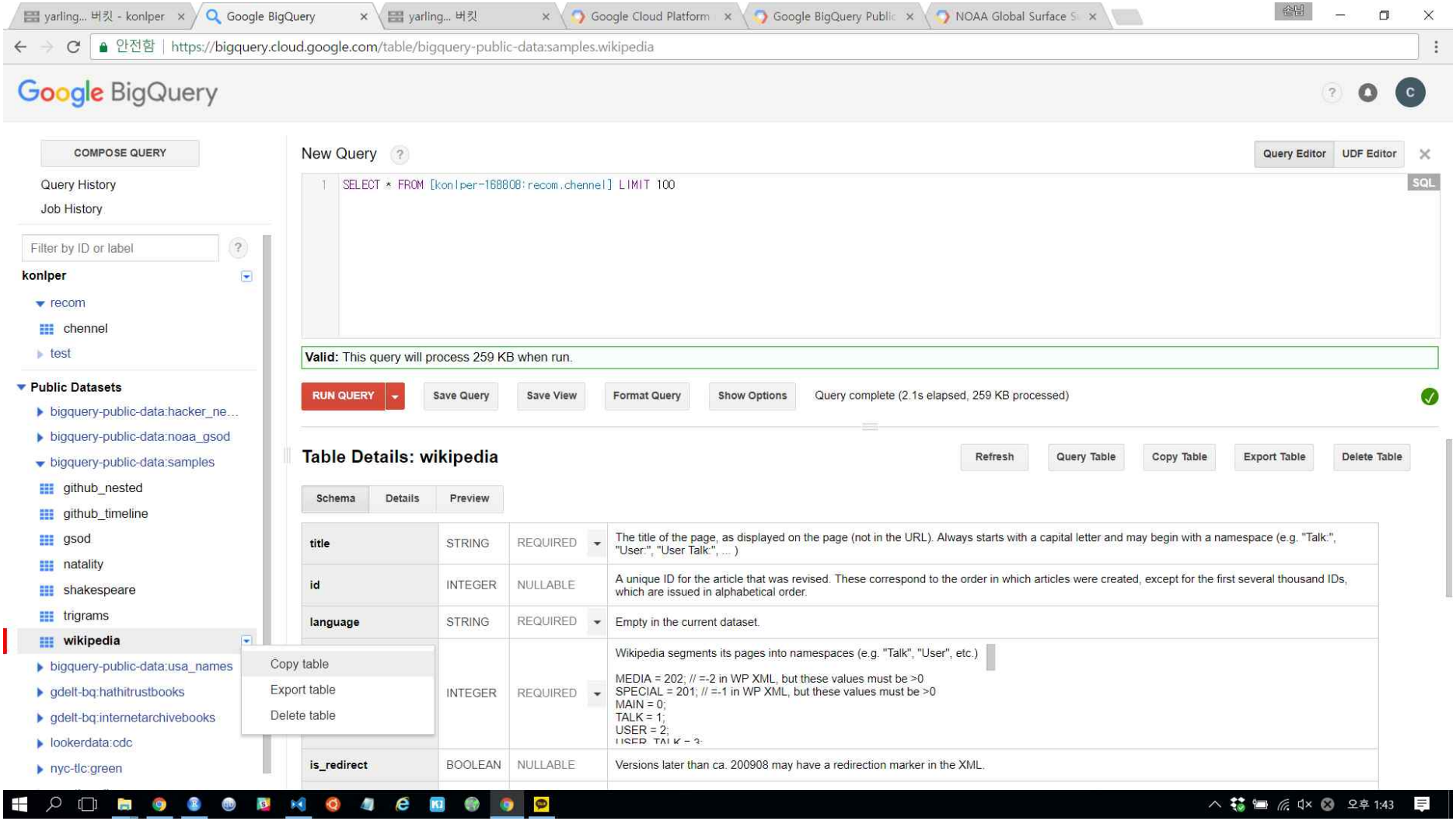
공개 데이터셋

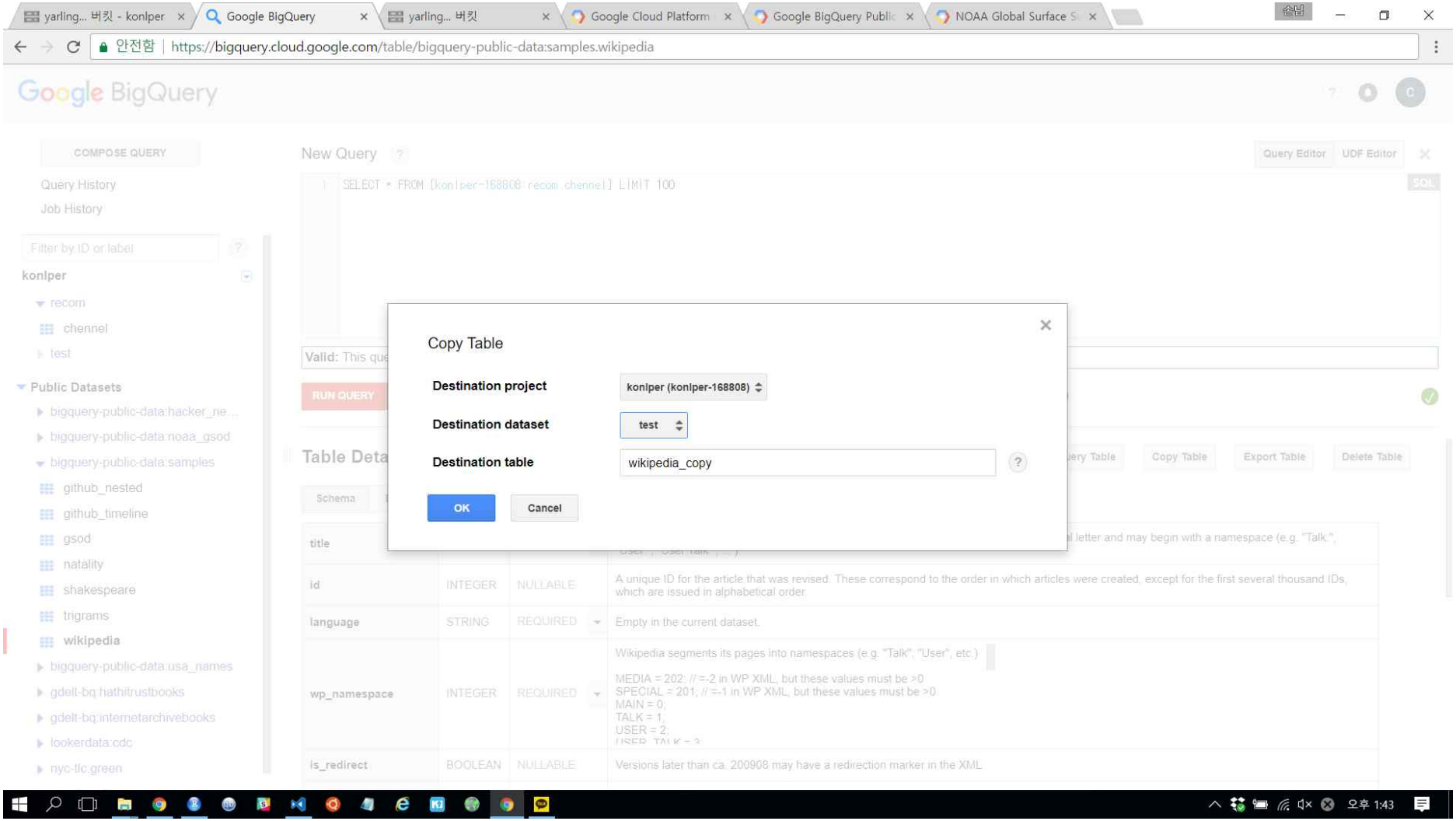

yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
Google Cloud Platform
Google BigQuery Public
https://cloud.google.com/bigquery/public-data/
Google Cloud Platform
검색
콘솔
Why Google
Products
Solutions
Launcher
Pricing
Customers
Documentation
Support
Partners
CONTACT SALES
All Resources
Pricing and Quotas
Release Notes
Support
Public Datasets
개요
1000 Cannabis Genomes Project
Bay Area Bike Share Trips Data
Chicago Crime Data
Chicago Taxi Trips
EPA Historical Air Quality Data
GDELT Books Corpus
GitHub Data
Hacker News
Healthcare Common Procedure Coding System (HCPCS) Level II
IRS 990 Data
Major League Baseball
Medicare
NHTSA Traffic Fatality Data
NOAA GHCN Weather
NOAA GSOD Weather
NOAA ICOADS
NYC 311 Service Requests
NYC Citi Bike Trips
NYC TLC Trips
BigQuery > Documentation
의견 보내기
Google BigQuery Public Datasets
A public dataset is any dataset that is stored in BigQuery and made available to the general public. This page lists a special group of public datasets that Google BigQuery hosts for you to access and integrate into your applications. Google pays for the storage of these data sets and provides public access to the data via BigQuery. You pay only for the queries that you perform on the data (the first 1 TB per month is free, subject to query pricing details).
Public datasets hosted by BigQuery
1000 Cannabis Genomes Project
Genomic open dataset of approximately 850 strains of Cannabis via the Open Cannabis Project.
Bay Area Bike Share Trips
This data includes all Bay Area Bike Share trips from August 2013 to the present, and is updated daily.
Chicago Crime Data
This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to the present.
목차
Public datasets hosted by BigQuery
1000 Cannabis Genomes Project
Bay Area Bike Share Trips
Chicago Crime Data
Chicago Taxi Trips
EPA Historical Air Quality Data
GDELT Book Corpus
GitHub Data
Hacker News
Healthcare Common Procedure Coding System (HCPCS) Level II
IRS Form 990 Data
Major League Baseball Data
Medicare Data
NHTSA Traffic Fatality Data
NOAA GHCN
NOAA GSOD
NOAA International Comprehensive Ocean-Atmosphere
오후 1:38

wikipedia dataset

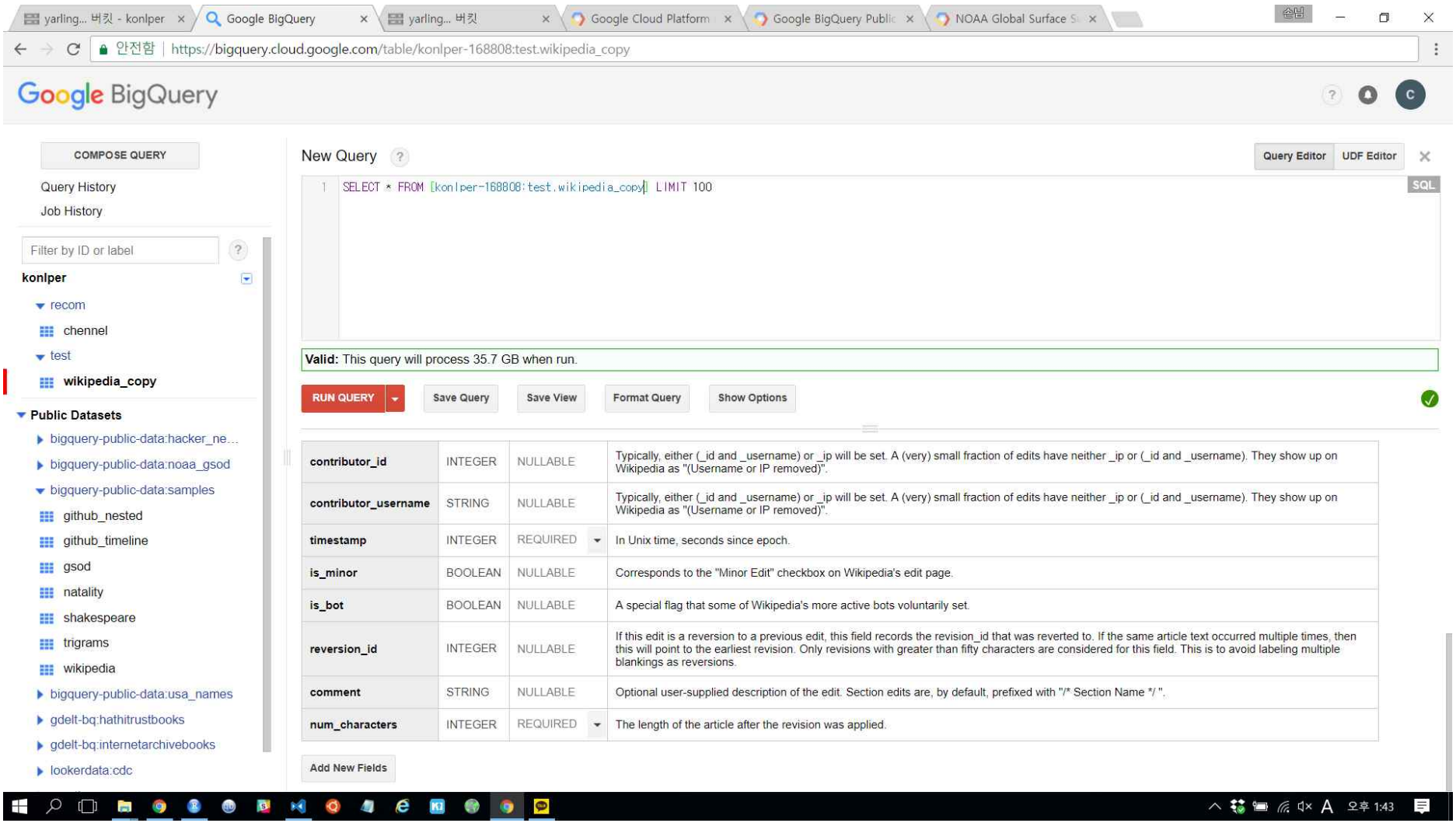


dataset 복사


Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
test
Public Datasets
bigquery-public-data:hacker_ne…
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
github_nested
github_timeline
gsod
natality
shakespeare
trigrams
wikipedia
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
Copy table
Export table
Delete table
New Query
Query Editor
UDF Editor
SELECT * FROM [konlper-168808:recom.chennel] LIMIT 100
SQL
Valid: This query will process 259 KB when run.
RUN QUERY
Save Query
Save View
Format Query
Show Options
Query complete (2.1s elapsed, 259 KB processed)
Table Details: wikipedia
Refresh
Query Table
Copy Table
Export Table
Delete Table
Schema
Details
Preview
title STRING REQUIRED The title of the page, as displayed on the page (not in the URL). Always starts with a capital letter and may begin with a namespace (e.g. "Talk:", "User:", "User Talk:", ...)
id INTEGER NULLABLE A unique ID for the article that was revised. These correspond to the order in which articles were created, except for the first several thousand IDs, which are issued in alphabetical order.
language STRING REQUIRED Empty in the current dataset.
INTEGER REQUIRED Wikipedia segments its pages into namespaces (e.g. "Talk", "User", etc.)
MEDIA = 202; // =-2 in WP XML, but these values must be >0
SPECIAL = 201; // =-1 in WP XML, but these values must be >0
MAIN = 0;
TALK = 1;
USER = 2;
is_redirect BOOLEAN NULLABLE Versions later than ca. 200908 may have a redirection marker in the XML.


Google BigQuery
https://bigquery.cloud.google.com/table/bigquery-public-data:samples.wikipedia
COMPOSE QUERY
Query History
Job History
konlper
recom
chennel
test
Public Datasets
bigquery-public-data:hacker_ne
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
github_nested
github_timeline
gsod
natality
shakespeare
trigrams
wikipedia
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
New Query
RUN QUERY
Copy Table
Destination project
konlper (konlper-168808)
Destination dataset
test
Destination table
wikipedia_copy
OK
Cancel
Table Deta
Schema
title
id
INTEGER
NULLABLE
A unique ID for the article that was revised. These correspond to the order in which articles were created, except for the first several thousand IDs, which are issued in alphabetical order.
language
STRING
REQUIRED
Empty in the current dataset.
wp_namespace
INTEGER
REQUIRED
Wikipedia segments its pages into namespaces (e.g. "Talk", "User", etc.)
is_redirect
BOOLEAN
NULLABLE
Versions later than ca. 200908 may have a redirection marker in the XML
오후 1:43

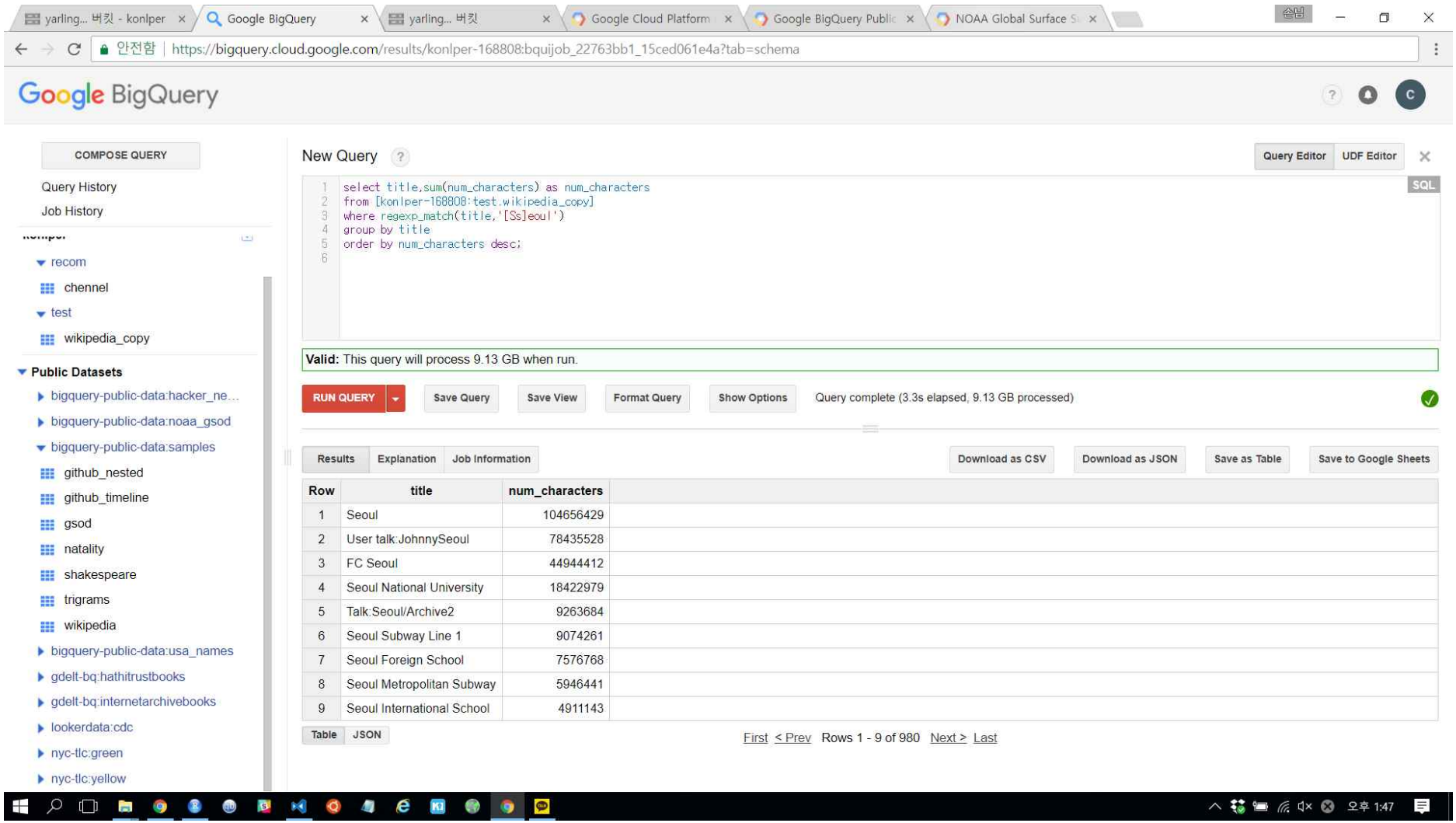
데이터 구조 보기


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
Google Cloud Platform
Google BigQuery Public
NOAA Global Surface
손님
안전함 | https://bigquery.cloud.google.com/table/konlper-168808:test.wikipedia_copy
Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
konlper
recom
chennel
test
wikipedia_copy
Public Datasets
bigquery-public-data:hacker_ne...
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
github_nested
github_timeline
gsod
natality
shakespeare
trigrams
wikipedia
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
New Query
Query Editor
UDF Editor
SQL
1 SELECT * FROM [konlper-168808:test.wikipedia_copy] LIMIT 100
Valid: This query will process 35.7 GB when run.
RUN QUERY
Save Query
Save View
Format Query
Show Options
contributor_id INTEGER NULLABLE Typically, either (_id and _username) or _ip will be set. A (very) small fraction of edits have neither _ip or (_id and _username). They show up on Wikipedia as "(Username or IP removed)".
contributor_username STRING NULLABLE Typically, either (_id and _username) or _ip will be set. A (very) small fraction of edits have neither _ip or (_id and _username). They show up on Wikipedia as "(Username or IP removed)".
timestamp INTEGER REQUIRED In Unix time, seconds since epoch.
is_minor BOOLEAN NULLABLE Corresponds to the "Minor Edit" checkbox on Wikipedia's edit page.
is_bot BOOLEAN NULLABLE A special flag that some of Wikipedia's more active bots voluntarily set.
reversion_id INTEGER NULLABLE If this edit is a reversion to a previous edit, this field records the revision_id that was reverted to. If the same article text occurred multiple times, then this will point to the earliest revision. Only revisions with greater than fifty characters are considered for this field. This is to avoid labeling multiple blankings as reversions.
comment STRING NULLABLE Optional user-supplied description of the edit. Section edits are, by default, prefixed with "/* Section Name */ ".
num_characters INTEGER REQUIRED The length of the article after the revision was applied.
Add New Fields
오후 1:43

큰 데이터 query 진행

```
select title,sum(num_characters) as num_characters
from [konlper-168808:test.wikipedia_copy]
where regexp_match(title,'[Ss]eoul')
group by title
order by num_characters desc;
```

query 결과


yarling... 버킷 - konlper
Google BigQuery
yarling... 버킷
Google Cloud Platform
Google BigQuery Public
NOAA Global Surface
손님
안전함 | https://bigquery.cloud.google.com/results/konlper-168808:bquijob_22763bb1_15ced061e4a?tab=schema
Google BigQuery
COMPOSE QUERY
Query History
Job History
recom
chennel
test
wikipedia_copy
Public Datasets
bigquery-public-data:hacker_ne...
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
github_nested
github_timeline
gsod
natality
shakespeare
trigrams
wikipedia
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow
New Query
Query Editor
UDF Editor
SQL
select title,sum(num_characters) as num_characters
from [konlper-168808:test.wikipedia_copy]
where regexp_match(title,'[Ss]eoul')
group by title
order by num_characters desc;
Valid: This query will process 9.13 GB when run.
RUN QUERY
Save Query
Save View
Format Query
Show Options
Query complete (3.3s elapsed, 9.13 GB processed)
Results
Explanation
Job Information
Download as CSV
Download as JSON
Save as Table
Save to Google Sheets
Row | title | num_characters
1 | Seoul | 104656429
2 | User talk:JohnnySeoul | 78435528
3 | FC Seoul | 44944412
4 | Seoul National University | 18422979
5 | Talk:Seoul/Archive2 | 9263684
6 | Seoul Subway Line 1 | 9074261
7 | Seoul Foreign School | 7576768
8 | Seoul Metropolitan Subway | 5946441
9 | Seoul International School | 4911143
Table
JSON
First < Prev Rows 1 - 9 of 980 Next > Last
오후 1:47

과제

1. RSQLite와 DBI를 활용해서 nycflights13 데이터를 db Table로 만들고, 제출해 주세요.
    1. 'nycflights13' 패키지를 설치하고 5개 데이터를 확인하세요.
    2. dbConnect 명령으로 SQLite 파일을 sql_[이름].db 으로 생성하세요.
    3. 5개의 데이터를 각각의 이름으로 table을 생성하세요.
    4. 1)~3)의 과정을 모두 코드로 남기고 run_[이름].R로 저장하세요.
    5. run_[이름].R과 sql_[이름].db 두 개의 파일을 class2assignment 폴더에 저장하세요.
    6. github에서 pull request로 제출해 주세요.
    7. recomen 폴더에 있는 데이터 6개도 같은 과정을 진행하고, 제출은 하지 마세요.
    8. .gitignore를 이용하면 push의 범위에서 제외할 수 있습니다.

과제

1. recomen 폴더에 있는 데이터 6개를 bigquery에 업로드 하고 질의를 실행해 보세요.
    1. "bigrquery" 패키지의 기능으로 작은 용량 5개의 데이터를 업로드해 주세요.
    2. Storage 서비스를 이용해서 `tran.csv` 파일을 bigquery에 테이블로 생성하세요.
    3. query_exec 함수와 "select * from [tran] limit 10" 을 실행하고 결과를 받으세요.
        - [tran]은 각자 해당하는 이름으로 변경하셔야 합니다.
    4. 1)~3)의 과정을 big_[이름].R로 저장하세요.
    5. big_[이름].R 파일을 class2assignment 폴더에 저장하세요.
    6. github에서 pull request로 제출해 주세요.