

Big Query

Anna Salazar

13 d'agost de 2022



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística



UNIVERSITAT DE
BARCELONA

Índex

1	Què és BigQuery?	1
1.1	Per què hauríem d'utilitzar BigQuery en lloc d'altres eines?	1
2	Creació i treball amb conjunts de dades i taules	2
2.1	Configuració de la Plataforma de Google Cloud (GCP)	2
2.1.1	Limitacions	2
2.2	Creació d'un conjunt de dades	3
2.3	Definició d'una taula de BigQuery des de la interfície d'usuari	4
2.3.1	Afegir dades a una taula de BigQuery senzilla	5
2.4	Càrrega de dades per crear una taula de BigQuery	6
2.5	Consulta de dades i visualització d'estadístiques de consultes	8
2.6	Creació d'una taula a partir d'un resultat de consulta	8
2.7	Creació d'una consulta a partir d'un filtre	8
3	Execució de consultes i visualització de resultats	9
3.1	Conjunts de dades públics a BigQuery	9
3.2	Configuració i ús de memòria cau de BigQuery	9
3.3	Taules externes de BigQuery	9
3.4	Integració de BigQuery amb Data Studio	9

1 Què és BigQuery?

BigQuery és un motor d'anàlisi de macrodades (Big Data) que permet executar consultes SQL al núvol sobre les dades emmagatzemades en aquest, sense importar el volum de les dades ni el tipus de consultes que es volen fer. El motor de consulta és capaç de treballar sobre terabytes de dades en qüestió de segons, i sobre petabytes en pocs minuts. Avui en dia, les empreses estan adoptant cada cop més la presa de decisions basades en dades i fomentant una cultura oberta en la qual les dades no estan aïllades dins dels departaments. BigQuery, en proporcionar els mitjans tecnològics per a promoure un canvi cultural cap a l'agilitat i l'obertura, realitza un paper molt important en l'augment del ritme de la innovació.

Treballar amb dades a BigQuery implica 3 aspectes principals: l'emmagatzemament, la incorporació de les dades i la consulta d'aquestes, Google s'encarrega de tota la resta. Com BigQuery és un servei totalment gestionat, no és necessari configurar ni instal·lar res en el nostre ordinador i, pel mateix motiu, no necessitem un administrador de la base de dades. Simplement, podem entrar en el nostre projecte de Google Cloud des del nostre navegador i començar a analitzar.

Pel que fa a l'emmagatzemament, les dades es guarden en una taula estructurada, la qual cosa significa que es pot utilitzar l'SQL estàndard per a facilitar la consulta i l'anàlisi de dades. BigQuery és perfecta pel Big Data perquè gestiona tot aquest emmagatzemament i està proveïda d'operacions d'escalabilitat que funcionen de forma automàtica sense que l'usuari s'hagi d'involucrar, per la qual cosa mai haurem de preocupar-nos per la grandària de les dades amb els quals treballem. Part de la consideració de disseny darrere de BigQuery és animar als usuaris a centrar-se en els coneixements en lloc de la infraestructura. Quan s'introdueixen les dades a BigQuery no és necessari pensar en els diferents tipus d'emmagatzemament, ni en els seus avantatges pel que fa a velocitat i cost; l'emmagatzemament està totalment gestionat.

Per a més informació sobre BigQuery, es pot consultar la pàgina de [Google Cloud](#).

1.1 Per què hauríem d'utilitzar BigQuery en lloc d'altres eines?

Una de les característiques més rellevants que presenta BigQuery és que es tracta d'una plataforma sense servidor, és a dir, que els servidors s'executen en segon pla, sense l'intervenció de l'usuari. A més, presenta una alta disponibilitat, la qual cosa es tradueix en que no cal preocupar-se per la caiguda dels servidors, ja que la plataforma s'encarrega d'això. Per últim, BigQuery també té propietats d'escalabilitat automàtica que fan possible gestionar fins a petabytes de dades. Aquestes característiques no estan disponibles a la majoria de plataformes d'emmagatzament de dades tradicionals, cosa que fa destacar BigQuery entre moltes.

Com en molts altres magatzems de dades, BigQuery és capaç de treballar amb moltes fonts de dades diferents. Podem pujar les dades des del nostre propi sistema d'arxius, de Google Cloud Storage, de Google Drive i de moltes fonts més. Després de fer-ho, es poden consultar aquestes dades utilitzant SQL estàndard o SQL heretat, el rendiment en qualsevol cas és excel·lent. Els resultats de les consultes solen emmagatzemar-se en la memòria cau durant 24 hores, de manera que les següents execucions d'aquesta consulta només hauran d'obtenir les dades de la memòria cau en lloc de fer-ho del disc.

2 Creació i treball amb conjunts de dades i taules

2.1 Configuració de la Plataforma de Google Cloud (GCP)

Per utilitzar aquesta eina d'anàlisi només ens caldrà crear un compte a Google Cloud i treballar a la zona de proves que ofereix Google per treballar de forma gratuïta. Per fer servir la zona de proves (*Sandbox*) seguirem els passos següents:

1. En primer lloc, ens dirigim a la interfície d'usuari de [BigQuery](#). Des d'aquesta interfície es poden realitzar la majoria de les operacions.

2. Accedim al nostre compte de Google o creem un nou compte si encara no en tenim cap. Si és el primer cop que iniciem sessió a Google Cloud, haurem de marcar el país on som i acceptar les condicions de servei.

3. Un cop dins, podem veure com és l'espai de treball SQL. Hi ha una secció de l'Explorador a l'esquerra que ens permet navegar en projectes, conjunts de dades i taules. Per tal de fer servir la zona de proves, haurem de crear un projecte.

Introdueix un nom al teu projecte i fes clic a Create. En el nostre cas, hem anomenat el projecte `el_meu_projecte`, i treballarem sobre aquest per il·lustrar el funcionament de la plataforma.

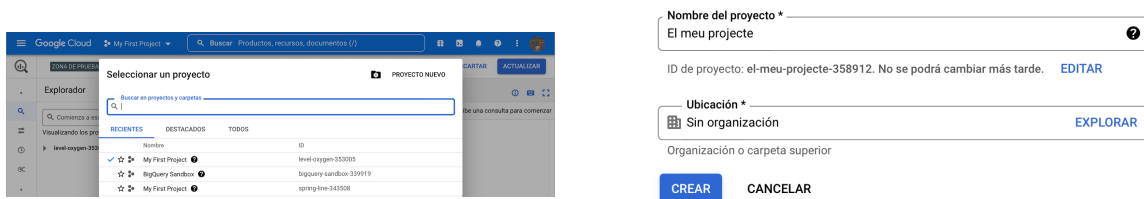


Figura 1: Creació d'un projecte

4. Un cop creat el projecte, el navegador ens redirigeix a la interfície web de BigQuery.

5. Ara ja podrem carregar o consultar dades en el nostre projecte sense cap compte de facturació adjunta.

2.1.1 Limitacions

Per a l'ús de la zona de proves gratuïta que ofereix Google, haurem de tenir en compte un seguit de limitacions.

En primer lloc, ens trobem amb un màxim de 10 Gb d'emmagatzemament i 10 Tb de consulta al mes. Al llarg d'aquest projecte no utilitzarem un volum de dades més gran ni sobrepassarem el límit d'espai de consulta, però s'han de tenir en compte aquestes limitacions si l'objectiu és treballar amb el format gratuït.

A més, ens trobem que tots els conjunts de dades tenen el temps de caducitat de la taula per defecte establert en 60 dies. Per tant, totes les taules, vistes o particions de les taules caducaran automàticament passats els 60 dies.

Una altra característica destacable és que els projectes de la zona de proves no són compatibles amb:

- La transmissió de dades
- Sentència de llenguatge de manipulació de dades (DML)
- Servei de transferència de dades de BigQuery

2.2 Creació d'un conjunt de dades

Ara que ja coneixem les limitacions de la plataforma i disposem d'un projecte en el que crear un conjunt de dades, ha arribat el moment de crear un nou conjunt de dades dins d'un projecte. Es pot pensar en un conjunt de dades a *BigQuery* com una agrupació lògica de taules. Alhora, diferents conjunts de dades s'integren en un mateix projecte.

Per a crear-ne un, només s'ha de desplegar el menú i triar l'opció de crear un nou conjunt de dades. Tot seguit hi ha diversos detalls per al conjunt de dades que es poden establir. En primer lloc, hi ha l'opció de canviar el projecte que l'encabirà. Això farà que aparegui un navegador on es pot especificar el projecte. Una altra possibilitat serà escollir la ubicació de les dades. Això determina on s'aprovisionaran els recursos subjacents, com la computació i l'emmagatzematge, per al servei *BigQuery*. Les consideracions a l'hora de triar una ubicació inclouran el rendiment per als usuaris finals, l'alta disponibilitat i també qualsevol restricció d'auditoria o compliment. I, per últim, es pot establir un temps d'expiració per defecte per a les taules dins d'un conjunt de dades.

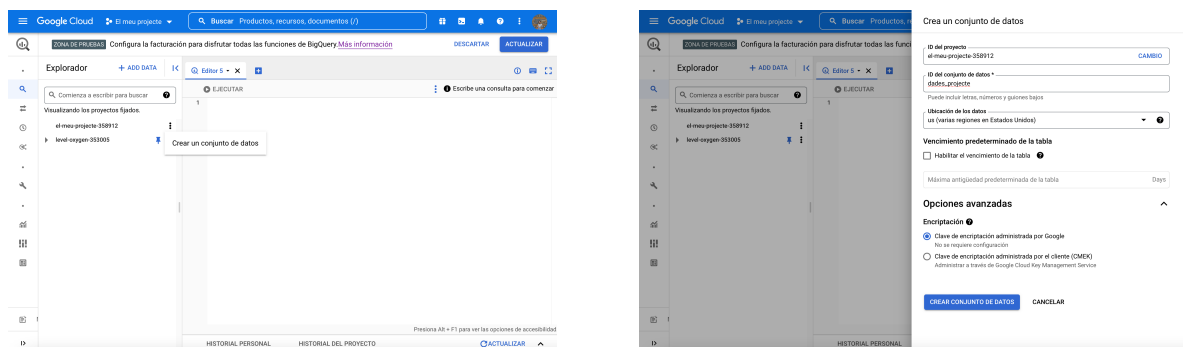


Figura 2: Creació d'un conjunt de dades

Tal com es pot veure a la figura anterior, hem creat un nou conjunt de dades anomenat `dades_projecte` que estarà ubicat en el projecte `el_meu_projecte`, la ubicació de les dades la hem posat a diverses regions dels Estats Units i, per últim, no hem habilitat un temps de venciment de la taula, sinó que per defecte *BigQuery* l'emmagatzemarà per 60 dies.

Un cop creat, el conjunt de dades `dades_projecte` ara apareix dins de `el_meu_projecte`, I ampliant això s'observa que no hi ha taules dins d'aquest. Ara, es pot donar un cop d'ull als detalls associats a aquest conjunt de dades. Des d'aquest menú, podem triar obrir-lo, moment en el qual la informació del conjunt de dades apareix a la dreta. Aquí podem confirmar l'identificador del conjunt de dades, que també assenyalava el projecte en el qual s'ha creat el conjunt de dades, i després altres detalls que inclouen les hores de creació i modificació.

A més, des d'aquesta finestra podrem compartir el conjunt de dades amb altres

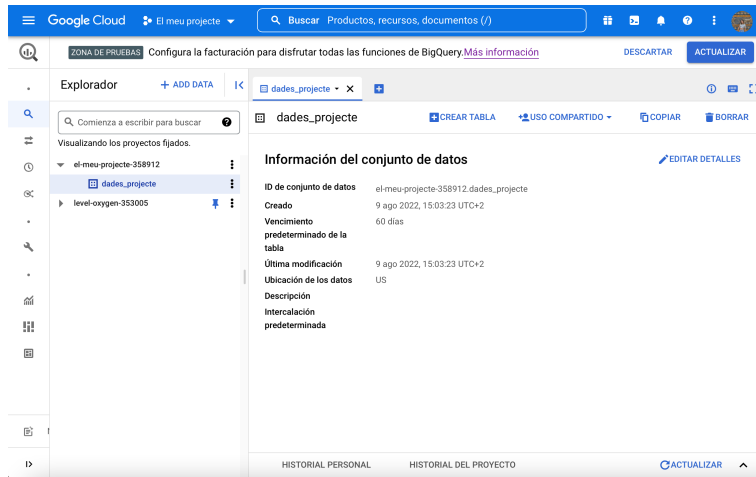


Figura 3: Informació del conjunt de dades

usuaris. Hi ha opcions per a copiar i eliminar aquest conjunt de dades. I després, a *editar detalls*, podem reconfigurar el temps de caducitat de les taules, establir una descripció o afegir etiquetes. Per exemple, si volem marcar aquest conjunt de dades com a pertanyent a un equip, podem establir una etiqueta amb la clau d'equip i el valor corresponent. Després, quan guardem aquest conjunt de dades, les etiquetes apareixen a l'apartat de informació.

2.3 Definició d'una taula de BigQuery des de la interfície d'usuari

Després d'haver creat un conjunt de dades en un projecte, ja es pot crear una taula dins d'aquest conjunt de dades. Si tenim la informació del conjunt de dades, hauríem de veure aquesta opció per a crear una nova taula des d'aquí. Alternativament, podem dirigir-nos al projecte, després al conjunt de dades i triar l'opció de crear una taula. Apareixerà un formulari i tindrem l'opció d'especificar una font per a la nostra taula. Això ens permetrà extreure dades de fonts ja existents, com l'emmagatzematge en el núvol de Google o bé un arxiu dels nostres propis sistemes. La primera taula que crearem serà bastant simple. De fet, serà una taula buida, i es pot observar la seva creació a la figura 4.

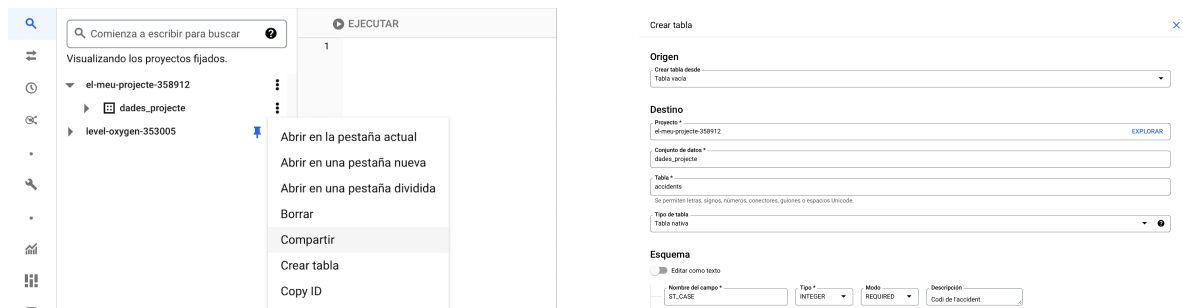


Figura 4: Creació d'una taula

A continuació, passem a la secció d'Esquema. Podem fer ús d'aquesta interfície per a establir les columnes de la nostra taula, incloent-hi els tipus i altres configuracions. La primera columna que definirà és l'identificador de l'accident, que s'anomenarà `ST_CASE`. Per al tipus de variable, podem triar d'un menú, que inclou tots els tipus amb els quals ja estem familiaritzats. Quant a la manera (columna *modo* a la figura 5), aquesta determinarà si els valors d'aquesta columna poden ser nuls o si es requereix un valor, així com també podem establir que els valors siguin d'un tipus repetit com *indistint*. Finalment, es pot escriure una descripció, que és opcional.

Figura 5: Esquema de la nostra taula

La taula 1 que acabem de crear està formada per 8 variables, 6 de les quals són numèriques i 2 categòriques, i es descriuen tal i com es pot veure a la taula següent:

Variable	Tipus	Descripció
ST_CASE	Númèrica	Codi de l'accident
DAY	Númèrica	Dia de l'accident (de l'1 al 31)
HOUR	Númèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Númèrica	Minut de l'accident (99 = desconegut)
RUR_URB	Categòrica	Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut)
DAY_WEEK	Categòrica	Dia de la setmana (1 = Diumenge, 2 = Dilluns, ..., 7 = Dissabte)
FATALS	Númèrica	Nombre de ferits a l'accident
DRUNK_DR	Númèrica	Nombre de conductors beguts involucrats a l'accident

Taula 1: Especificacions de la taula Accidents

2.3.1 Afegir dades a una taula de BigQuery senzilla

Ara que hem creat una taula de consulta, podem centrar-nos en treballar amb ella. Per a això, ens desplaçarem cap avall i donarem un cop d'ull al primer esquema de la taula (a la figura 6). I aquí és on podem donar un cop d'ull a alguna informació interessant. Més enllà de la identificació de la taula, que es troba a l'esquerra de la figura, també podem comprovar la grandària de la taula a la dreta, que ens donarà una indicació de la quantitat de dades que es processaran, si anéssim a executar consultes contra ella. La grandària d'emmagatzematge a llarg termini assenjala les dades a les quals no s'ha accedit en els últims 90 dies, i després, per descomptat, tenim les hores de creació i modificació juntament amb la ubicació de les dades de la taula, que correspon a la ubicació de l'alimentador de dades que va ser reemplaçat. Des d'aquesta interfície,

també podem editar els detalls existents d'aquesta taula. Aquí podem establir un temps de caducitat en cas que vulguem anul·lar el que s'ha establert en el nivell del conjunt de dades. També tenim l'opció d'establir una descripció o afegir etiquetes.

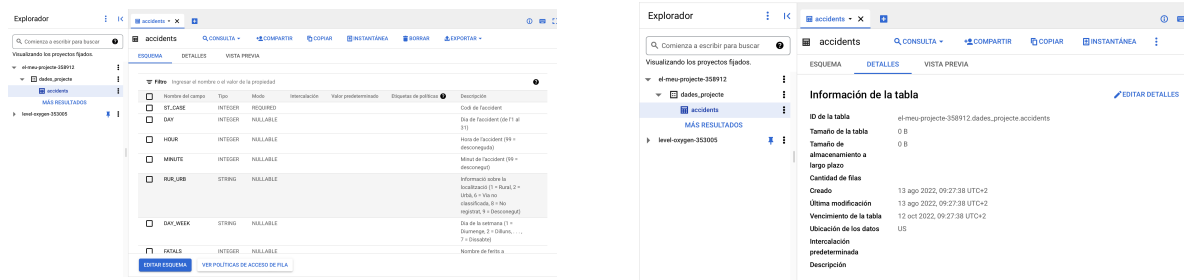


Figura 6: Detalls de la taula

Una altra característica que podem consultar és la vista prèvia de la taula, i com és lògic, veurem que aquesta encara no conté dades, ja que simplement hem creat l'esquema de la taula, sense inserir cap dada en aquesta. Podrem afegir dades a partir d'una simple consulta a la taula.

```
INSERT INTO 'el-meu-projecte-358912.dades_projecte.accidents'
(ST_CASE, DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR)
VALUES (20055, 1, 20, 55, "1", "3", 3, 0);
```

Comencem amb 'insert into', seguit de la taula en la qual s'ha de realitzar la inserció. Per a això, utilitzem parèntesi, i dins dels parèntesis, utilitzem la sintaxi, data set *name *dot *able *name, que en el meu cas és, *loony* underscore university.students. A continuació, en la línia número dos, podem especificar els noms de les columnes per a les quals s'afegiran valors, així que, com s'ha especificat, cadascuna de les cinc columnes presents en la taula, i després, seguit de la paraula clau values, especifiquem entre parèntesi, els propis valors. Observi's que en les consultes grans, podem utilitzar cometes dobles en especificar cadenes. També hi ha una altra informació interessant aquí per a la que minimitzaré el panell de l'explorador. Notaràs que hi ha un missatge que assenyalava que es processaran zero bytes, quan executem aquesta consulta. Ara, seguirem endavant i executar aquesta consulta, i sembla que la inserció va ser un èxit, atès que una fila s'ha afegit a la taula dels estudiants. Per a confirmar-ho, executem una consulta més, concretament, un select from loony university.students. Significativament, obtenim un missatge que diu que, en executar aquesta consulta, es processaran 60 bytes de dades, assenyalant el cost que es produirà. Efectivament, l'única fila ha estat retornada, i fins i tot en els resultats de la consulta, obtenim la confirmació que s'han processat 60 bytes. El temps d'execució de la consulta també apareix aquí.

2.4 Càrrega de dades per crear una taula de BigQuery

Hem vist que és possible crear una taula buida i després emplenar-la amb dades amb sentències INSERT, ara explorarem un cas d'ús més comú per als usuaris de BigQuery en el qual es crea una taula a partir de dades existents. Per a això, ens dirigirem al conjunt de dades `dades_projecte` i triarem crear una nova taula. I aquesta vegada, la font no serà una taula buida, sinó que carregarem un arxiu CSV del nostre propi

sistema d'arxius per al qual es pot seleccionar l'opció de càrrega. Això no obstant, els tipus d'arxiu que podem importar han de ser del tipus CSV, JSON, Avro o Parquet, principalment. Hi ha algunes restriccions quant a la grandària de l'arxiu que podem pujar. Per tant, aquest ha de romandre dins dels 100 MB. Procedim llavors a navegar pels nostres sistemes d'arxius per a l'arxiu a pujar. I després navegar fins a `accidents.csv`. Una vegada que l'arxiu ha estat seleccionat, el format de l'arxiu s'ha establert automàticament en CSV.

La imatge mostra dues captures de pantalla de la interfície d'usuari. La captura de l'esquerra, titulada 'Crear tabla', mostra els passos per crear una nova taula: seleccionar l'arxiu 'accidents_bd.csv', confirmar el format 'CSV', especificar el projecte 'el meu projecte 358912' i el nom de la taula 'accidents_bd'. La captura de la dreta, titulada 'Configuración de particiones y clústeres', mostra les opcions avançades de configuració, incloent-hi la preferència d'escriptura, la quantitat de errors permesos i el delimitador de camps personalitzat, configurat com a punt i coma (;).

Figura 7: Lectura d'un arxiu extern

Quant al projecte i al conjunt de dades, els deixarem com estan. I després, el nom de la taula, `accidents_bd`, farà saber que inclou informació sobre diversos accidents de trànsit. A continuació, tenim l'opció de definir explícitament l'esquema. No obstant això, atès que es tracta d'un arxiu CSV amb múltiples columnes, podem triar l'opció de detectar automàticament. D'aquesta manera, BigQuery donarà un cop d'ull al contingut de cada columna i determinarà quin ha de ser l'esquema. Més enllà d'això, al final de la finestra de creació de la taula ens apareixeran unes Opcions avançades. Aquestes opcions permeten la lectura de diferents tipus de CSV, entre d'altres coses. Sabem que el delimitador de camps d'un CSV pot ser una tabulació o una coma entre d'altres possibilitats. En el nostre cas, cal especificar que el nostre tabulador és el punt i coma “;”. Si hem establert totes aquestes especificacions, ja podrem començar amb l'anàlisi.

Ara es pot comprobar que `accidents_bd` apareix sota el conjunt de dades `dades_projecte`. A continuació, podem accedir a la informació de la taula i al seu contingut desplegant el menú i triant Obrir. En l'esquema de la taula, sabrà que s'ha detectat automàticament el tipus dels diferents camps. Donem un cop d'ull als detalls de la taula. Aquí notaràs que la grandària total és de poc més de 280 KB. El nombre de files és d'unes 2.780. I després, quan ens dirigim a la Vista Prèvia, obtenim un cop d'ull als continguts.

2.5 Consulta de dades i visualització d'estadístiques de consultes

2.6 Creació d'una taula a partir d'un resultat de consulta

2.7 Creació d'una consulta a partir d'un filtre

3 Execució de consultes i visualització de resultats

3.1 Conjunts de dades públics a BigQuery

3.2 Configuració i ús de memòria cau de BigQuery

3.3 Taules externes de BigQuery

3.4 Integració de BigQuery amb Data Studio

Índex de figures

1	Creació d'un projecte	2
2	Creació d'un conjunt de dades	3
3	Informació del conjunt de dades	4
4	Creació d'una taula	4
5	Esquema de la nostra taula	5
6	Detalls de la taula	6
7	Lectura d'un arxiu extern	7