

Big Query

Anna Salazar

5 de novembre 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 interície 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	7
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	9
3	Dades públiques i dades externes	11
3.1	Conjunts de dades públiques a BigQuery	11
3.2	Taules externes de BigQuery	13
4	Integració de BigQuery amb Data Studio	15
4.1	Ús de Data Studio des de BigQuery	15
4.2	Ús de Data Studio des de la pròpia plataforma	16
5	Connexió d'R a BigQuery	19
5.1	Connectem R a BigQuery	19
5.2	Consultes amb 'bigrquery'	20

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 mateix 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 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.

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 la 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. 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'emmagatzemament de dades tradicionals, i fan destacar BigQuery entre moltes.

Com en molts altres magatzems de dades, BigQuery és capaç de treballar amb moltes fonts de dades diferents. Es poden pujar les dades des del propi sistema d'arxius, des de Google Cloud Storage o des de Google Drive, entre moltes més fonts. Després de fer-ho, es poden consultar aquestes dades utilitzant SQL estàndard o SQL heretat. 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 operar 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 (UI) 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_proyecto` (Figura 1), i treballarem sobre aquest per il·lustrar el funcionament de la plataforma.

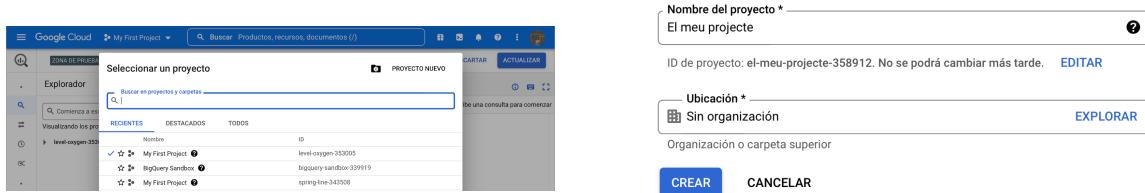


Figura 1: Creació d'un projecte

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

Això no obstant, 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 establerta en seixanta dies. Per tant, totes les taules, vistes o particions de les taules caducaran automàticament passats els seixanta 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 qual crear un conjunt de dades, ha arribat el moment de crear un nou conjunt de dades dins d'aquest 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 podrà 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.

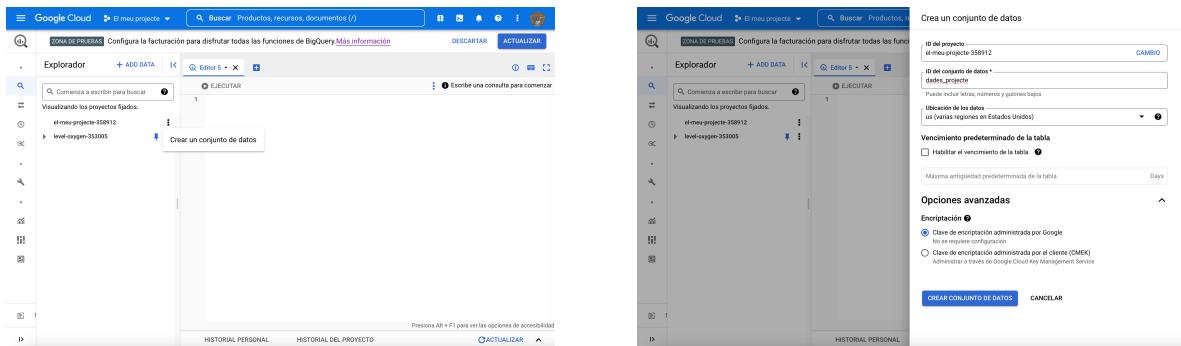


Figura 2: Creació d'un conjunt de dades

Tal com es pot veure a la figura 2, hem creat un nou conjunt de dades anomenat **dades_projecte** que estarà ubicat en el projecte **el_meu_proyecto**, la ubicació de les dades l'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 seixanta dies.

Un cop creat, el conjunt de dades **dades_projecte**, es pot comprovar que ara apareix dins de **el_meu_proyecto** a la UI de BigQuery 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 (Figura 3). 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 assenyala el projecte en el qual s'ha creat el conjunt de dades, i després altres detalls que inclouen les hores de creació i modificació, així com la ubicació d'aquest.

The screenshot shows the Google Cloud BigQuery interface. At the top, there's a navigation bar with 'Google Cloud' and 'El meu projecte'. Below it is a search bar 'Buscar Productos, recursos, documentos ()'. The main area is titled 'ZONA DE PRUEBAS' with the sub-instruction 'Configura la facturación para disfrutar todas las funciones de BigQuery'. There are buttons for 'DESCARTAR' and 'ACTUALIZAR'. On the left, there's an 'Explorador' sidebar with a search input 'Comienza a escribir para buscar'. The main content area shows a list of datasets: 'el-meu-projecte-358912' (selected), 'dades_projecte' (highlighted in blue), and 'level-oxygen-353005'. The 'dades_projecte' dataset details are shown on the right, including its ID, creation date (9 ago 2022, 15:03:23 UTC+2), expiration date (60 days), last modification date (9 ago 2022, 15:03:23 UTC+2), location (US), description, and default collation. Buttons for 'CREAR TABLA', 'USO COMPARTIDO', 'COPIAR', and 'BORRAR' are at the top right. Below the details are tabs for 'HISTORIAL PERSONAL' and 'HISTORIAL DEL PROYECTO', with an 'ACTUALIZAR' button.

Figura 3: Informació del conjunt de dades

A més, des d'aquesta finestra podrem compartir el conjunt de dades amb altres usuaris. Hi ha opcions per a copiar i eliminar aquest conjunt de dades. I després, a l'opció *editar detalles*, 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. En acabant, quan guardem aquest conjunt de dades, les etiquetes apareixen a l'apartat d'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, i ens servirà per explorar una mica la plataforma. De fet, serà una taula buida anomenada **accidents** (Figura 4).

The screenshot shows the Google Cloud BigQuery interface. On the left, there's an 'Explorador' sidebar with a search input 'Comienza a escribir para buscar'. The main content area shows a list of datasets: 'el-meu-projecte-358912' (selected), 'dades_projecte' (highlighted in blue), and 'level-oxygen-353005'. A context menu is open over the 'dades_projecte' dataset, with options like 'Abrir en la pestaña actual', 'Abrir en una pestaña nueva', 'Abrir en una pestaña dividida', 'Borrar', 'Compartir', 'Crear tabla', and 'Copy ID'. The 'Crear tabla' option is highlighted. To the right, a 'Crear tabla' dialog box is open. It has sections for 'Origen' (Create table from), 'Destino' (Project: 'el-meu-projecte-358912', Dataset: 'dades_projecte', Table: 'accidents'), 'Esquema' (Schema: 'Nombre del campo * ST_CASE', 'Tipo * INTEGER', 'Modo * REQUIRED', 'Descripción * Codi de l'accident'), and a 'EXPLORAR' button. The 'Destino' section also includes dropdowns for 'Proyecto' and 'Conjunto de datos'.

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 definiré és l'identificador de l'accident, que s'anomenarà ST_CASE. Per al tipus de variable, podem triar d'entre menú d'opcions, 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 (com és el cas de l'identificador), i també podem establir que els valors siguin d'un tipus que es pugui repetir, marcant *indistint*. Finalment, es pot escriure una descripció per a la variable, 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 com es pot veure a continuació.

Variable	Tipus	Descripció
ST_CASE	Numèrica	Codi de l'accident
DAY	Numèrica	Dia de l'accident (de l'1 al 31)
HOUR	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	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	Numèrica	Nombre de ferits a l'accident
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident

Taula 1: Especificacions de la taula Accidents

En el transcurs del treball, farem ús d'aquesta taula, juntament amb dues més, que prenen de nom de **persones** i **vehicles**, per analitzar les dades que es van prendre d'un conjunt d'accidents que es van donar als Estats Units.

2.3.1 Afegir dades a una taula de BigQuery senzilla

Ara que hem creat una taula de consulta, podem centrar-nos a 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), on es troba a alguna informació interessant. Més enllà de la identificació de la taula, 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 sobre aquesta. La grandària d'emmagatzematge a llarg

termini assenyala les dades a les quals no s'ha accedit en els últims noranta dies, i després, per descomptat, tenim les hores de creació i modificació juntament amb la ubicació de les dades de la taula. Des d'aquesta interfície, també podem editar els detalls existents d'aquesta taula. Aquí podem establir un temps de caducitat en cas que vulguem anular 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: Details 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. Si féssim ús de SQL, en qualsevol altre context es podrien afegir dades a partir d'una simple consulta a la taula, que tindria l'estructura següent:

```
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);
```

A partir d'aquesta consulta afegiríem a la taula el cas d'un accident amb identificador 20055, que es va produir el dia 1 del mes a les 20:55 a una zona rural (**RUR_URB = 1**) un dimarts (**DAY_WEEK = 3**), i en el que hi ha 3 ferits i cap conductor begut involucrat en l'accident.

Això no obstant, quan intentem executar la consulta, BigQuery ens informa d'un error (Figura 7). Si recordem, prèviament s'han definit algunes de les limitacions per a l'ús de la zona de proves de BigQuery. Entre aquests s'hi troba que no podem utilitzar el llenguatge de manipulació de dades (DML), és a dir, que no podem modificar la taula amb sentències com **INSERT INTO**, **UPDATE** o **DELETE**, per exemple. Per aquest motiu, l'error ens avisa que no tenim el nostre projecte vinculat a un compte i, per tant, no ens avaluarà la nostra consulta.

Figura 7: Inserció de dades a la taula

2.4 Càrrega de dades per crear una taula de BigQuery

Hem vist que és possible crear una taula buida, però que la zona de proves (*sandbox*) no ens permet 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 nostre conjunt de dades, `dades_projecte`, i triarem crear una nova taula. Aquest cop, la font no serà una taula buida, sinó que carregarem un arxiu CSV del nostre propi sistema d'arxius. Un cop seleccionem importar les dades, es pot seleccionar diferents tipus d'arxiu com ara CSV, JSON, Avro o Parquet, principalment. Per a respectar les limitacions de la zona de proves, hi ha algunes restriccions quant a la grandària de l'arxiu que podem pujar, recordem que aquestes han de ser menors a 10 Gb. Procedim llavors a navegar pels nostres sistemes d'arxius per a l'arxiu a pujar. Una vegada que l'arxiu ha estat seleccionat, el format de l'arxiu s'ha establert automàticament en CSV (Figura 8).

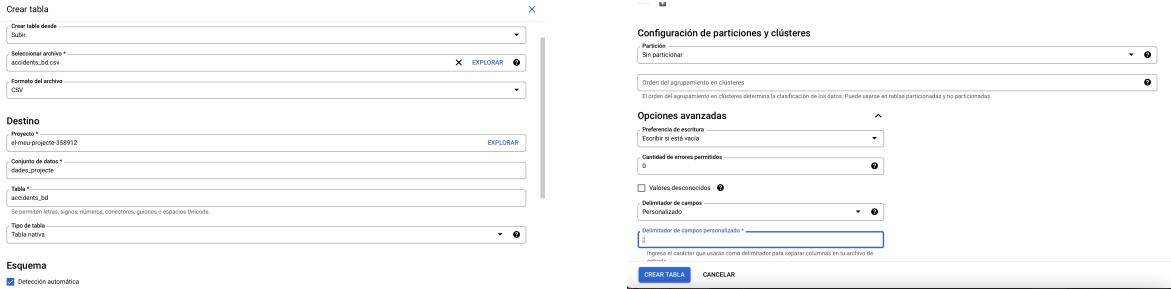


Figura 8: Lectura d'un arxiu extern

Quant al projecte i al conjunt de dades, els deixarem com estan. I escollirem el nom de la taula, `accidents_bd`, el qual 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 altres coses. Sabem que el delimitador de camps d'un CSV pot ser una tabulació o una coma entre 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 comprovar que `accidents_bd` apareix sota el nostre 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 (Figura 9).

Figura 9: Informació sobre la taula

2.5 Consulta de dades i visualització d'estadístiques de consultes

Per a executar consultes en aquesta taula, ens dirigim al botó de consulta i s'obrirà en una nova pestanya. Aquesta pot ser una pestanya completament nova que ocultarà aquesta vista de detalls, mentre que una pestanya dividida ens permetrà fer referència a aquesta vista de detalls per a la taula mentre construïm una consulta. Mitjançant aquest procés, ha aparegut una nova pestanya cap a la dreta, i la consulta que apareix per defecte inclou una clàusula **SELECT**, però no inclou cap camp (Figura 10). Precisament per això hi ha un error de sintaxi, com es mostra a la dreta de la figura.

Figura 10: Elaboració d'una consulta

Ara, per a completar la clàusula **SELECT**, podríem escriure els noms dels atributs que volem consultar, a més de condicions, per exemple, mitjançant la clàusula **WHERE**. Concretament, l'esquema que s'haurà de seguir per a consultar la base de dades té la forma següent:

```
SELECT atribut1, atribut2, ...
FROM '[nom_projecte].[nom_base_de_dades].[nom_taula]'
(WHERE condició)
```

Si escrivim a l'editor la nostra consulta, apareixerà un validador d'aquesta a la part superior dreta de la finestra. Aquest validador l'hem vist anteriorment quan ens indicava un error en voler executar una consulta fent servir llenguatge de manipulació de dades (DML), i pot agafar dues formes: - Si la consulta és vàlida, apareixerà una icona de verificació verd. - Si la consulta no és vàlida, apareixerà una icona d'exclamació vermella. A més, el validador també mostra la quantitat de dades que la consulta processarà quan s'executi. Per exemple, si demanem en una consulta que ens retorna la columna sencera **DAY**, el validador de la dreta ens marca que es processaran una quantitat de gairebé 22 kB, tal com es pot veure a la figura 11.

Executarem aquesta consulta prement **Executar**. Els resultats apareixen sota la finestra d'editor i mostra certs detalls com, per exemple, que s'ha executat en uns 0 segons per a mi. Per descomptat, podem desplaçar-nos i donar un cop d'ull a tots

The screenshot shows a Google Cloud BigQuery interface. At the top, there is a button labeled "EJECUTAR" (Execute) and a note stating "Esta consulta procesará 21.72 KB cuando se ejecute." (This query will process 21.72 KB when executed). Below this, the SQL query is displayed: "SELECT DAY FROM `el-meu-projecte-358912.dades_projecte.accidents_bd` LIMIT 1000". A message below the query says "Presiona Alt + F1 para ver las opciones de accesibilidad" (Press Alt + F1 to view accessibility options). The main area is titled "Resultados de la consulta" (Query Results) and includes tabs for "RESULTADOS" (Results), "JSON", and "DETALLES DE LA EJECUCIÓN" (Execution Details). The "DETALLES DE LA EJECUCIÓN" tab is selected, showing various metrics about the job execution, such as the job ID, user, location, creation and start times, duration, bytes processed, bytes billed, priority, and whether it's a legacy SQL job. At the bottom of this section, it says "Tabla de destino" (Destination Table) and "Temporary table".

Figura 11: Primera consulta

els resultats (a la dreta de la Figura 11). Entre els detalls que es mostren per a cada consulta s'hi troben la informació del treball, els resultats en forma de taula, els resultats en format JSON i certs detalls de l'execució de la consulta. Si posem el focus en la informació del treball, es troba l'identificador d'aquest, l'usuari que l'ha executat, la ubicació on s'emmagatzemen les dades, l'hora de la creació i l'execució d'aquesta, el temps d'execució i els bytes processats i facturats. Veiem que el nombre de bytes facturats és de 10 MB, aquesta és la quantitat mínima que surt per defecte per a cada consulta per *Google Cloud Platform*, i té en compte les despeses generals. Per a consultes a bases de dades més extenses, aquesta facturació serà major i ens impedirà l'ús de la zona de proves. L'última característica que crida l'atenció de la informació del treball és que els resultats s'emmagatzemem en una taula temporal. Això vol dir que aquesta taula resultant no es guardarà com una més en el nostre conjunt de dades i, per tant, no la podrem consultar.

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

Una altra alternativa a les taules temporals serà crear una taula en el nostre conjunt de dades a partir d'una taula temporal o vista. En aquest cas, farem una sèrie temporal de 31 observacions que ens compti el nombre d'accidents ocorreguts cada dia del mes a partir de la consulta següent:

```
SELECT DAY, COUNT(*) AS FREQ
FROM `el-meu-projecte-358912.dades_projecte.accidents_bd`
GROUP BY DAY
```

Un cop realitzada la consulta, hem d'exportar totes aquestes dades a una nova taula i, per a fer-ho, revisarem algunes de les opcions d'exportació que es troben al menú **Guardar resultats**. En aquest, apareixen diverses opcions per a la manera de guardar els resultats (Figura 12). Podem guardar-los com un arxiu CSV en Google Drive o en un arxiu local. En el nostre cas, exportarem el contingut a una nova taula

de BigQuery. Una vegada feta aquesta selecció, podem decidir el nom del projecte i el conjunt de dades on s'aprovisionarà la taula i, a més, establir un nom de taula.

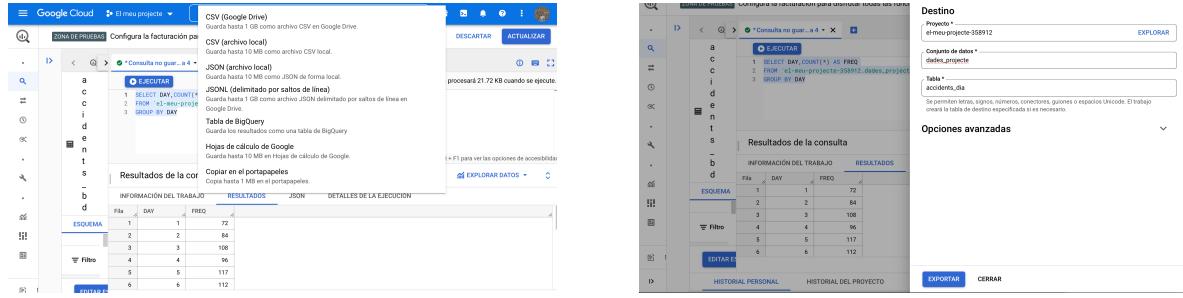


Figura 12: Creació d'una taula a partir d'una consulta

Llavors, quan guardem les coses, això haurà iniciat un nou treball per a guardar la nova taula i carregar-la amb dades. Una vegada que tanquem aquesta notificació, podem treure el pin de l'Explorador i baix, apareix com una taula. En obrir-la confirmem que l'esquema apunta a les mateixes columnes que havíem referenciat en la clàusula **SELECT** de la consulta que va crear aquesta taula. Des dels detalls podem confirmar que el nombre de files coincideix amb el dels resultats de la consulta, concretament 31. I després la vista prèvia ens mostrerà quines són les dades exactament.

Hom es pot preguntar: *Quina és la finalitat d'aquesta taula?* Bé, atès que només conté un subconjunt de la taula original, les consultes cap a aquesta taula tindran potencialment menys dades per a processar que les consultes que s'executen directament a la taula original. Potser en aquest cas no tornarem a necessitar aquesta consulta, però imaginem que només ens interessa estudiar els accidents que s'hagin produït a zones rurals i creem una consulta que filtri aquest tipus de casos. En aquesta situació, serà millor executar la resta de consultes sobre la taula petita, que inclou tota la informació que necessitem i probablement tindrà un cost de consulta menor.

3 Dades públiques i dades externes

3.1 Conjunts de dades públiques a BigQuery

Mentre continuem familiaritzant-nos amb la plataforma, podem explorar una altra opció que ofereix BigQuery, que són dades que estan disponibles públicament per a que qualsevol usuari pugui executar les seves consultes. Aquests conjunts de dades públiques es troben en un projecte anomenat BigQuery Public Data. Per a accedir a elles, farem servir al botó d'agregació de dades (**ADD DATA**) que es troba al costat de l'explorador, i escollirem l'opció **Explorar conjuntos de datos públicos**) (Figura 13).

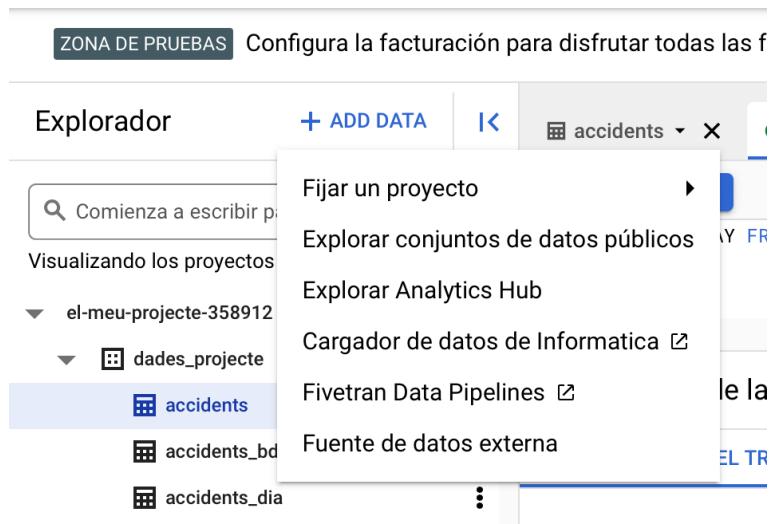


Figura 13: Accés als conjunts de dades públiques

Podem donar un cop d'ull als diferents conjunts de dades i taules d'aquest projecte i navegar entre les diferents opcions per a poder triar aquells que cridin la nostra atenció i, en definitiva, amb els que treballarem. Per exemple, entre totes aquestes taules s'hi troba una que conté informació sobre la població catalana recopilada per l'aplicació *GenCat Mobile Coverage* (Figura 14). Dins d'aquesta taula s'hi troben dades recollides mitjançant crowdsourcing¹ i tenen informació sobre l'estat de la cobertura de la telefonia mòbil a Catalunya. La plataforma utilitza una aplicació Android per a registrar les dades dels ciutadans a través dels seus dispositius mòbils sobre el nivell de cobertura per operador, la xarxa utilitzada (2G, 3G i 4G) i la ubicació del dispositiu. Aquestes dades en concret van ser recopilades durant els anys 2015-2017.

Si volguéssim descarregar la taula i situar-la dins del nostre projecte, BigQuery ens permet aquesta operació si premem **Visualizar conjunt de dades** seguit de **crear taula**. Això no obstant, també és possible consultar la taula sense necessitat de descarregar-la dins del nostre projecte, fent servir el projecte **bigquery-public-data**, que és el que farem en aquest cas.

¹La pràctica d'obtenir informació o aportacions a una tasca o projecte recorrent als serveis d'un gran nombre de persones, remunerades o no, normalment a través d'Internet.

Figura 14: Creació d'una taula a partir d'una consulta

Dins la informació de la taula que ens proporciona la plataforma, apareix una secció amb consultes suggerides (les podem trobar a la dreta de la Figura 14). D'entre aquestes, la primera fa referència als 10 pobles catalans amb pitjor cobertura de mòbil. Si premem el botó **Run this query**, el navegador ens redirigeix a l'editor amb la consulta preparada, i nosaltres la podem executar sense cap cost, ja que tant la taula com el resultat de la consulta estan emmagatzemats a BigQuery i les dades no s'han de processar (Figura 15).

Figura 15: Consulta a un conjunt de dades públic

Els resultats d'aquesta consulta ens diu que els pobles catalans amb menys cobertura mòbil són Canejan, Boadella i les Escaules, Cabó, la Vajol, Gaià, Farrera, Gisclareny, Viver i Serrateix, Savallà del Comtat i Torroja del Priorat. A més, els bytes processats en aquesta consulta apareixen com un paràmetre indefinit, ja que aquests resultats estan emmagatzemats a la memòria cau de BigQuery.

Tal com hem vist, BigQuery emmagatzema a la seva memòria cau els resultats de la consulta perquè les dades puguin ser recuperades més ràpidament la pròxima vegada que s'executi una mateixa consulta. No obstant això, cal tenir en compte que només s'accedeix a les dades de la memòria cau quan s'executa la mateixa consulta després de

la seva creació. Per exemple, si es modifiqués una mica aquesta consulta i, per exemple, demanés tan sols el nom del poble, en comptes d'aquest i la mitjana del senyal mòbil, podríem arribar a pensar que els resultats de la consulta d'aquesta execució haurien de retornar-nos un subconjunt de les dades que ja són presents en la memòria cau (ja que és un subconjunt de la nostra consulta anterior). Això no obstant, per la forma en què funciona la memòria cau de BigQuery, quan executem això, observarem que la informació no s'ha recuperat d'aquesta. En canvi, quan aquesta consulta es torna a executar, és quan la memòria cau s'activa, i és d'on es recuperen les dades.

Per tant, la memòria cau només funciona si és la mateixa consulta la que es torna a executar. Aquestes característiques són de gran interès, ja que l'emmagatzematge en memòria cau és una gran manera de reduir el cost d'execució de les consultes, i també millora el rendiment d'aquestes.

3.2 Taules externes de BigQuery

Una altra funció que presenta la plataforma és la lectura d'arxius externs que s'actualitzen de forma periòdica. Això és d'especial utilitat en els casos en què la informació de la qual es disposa és a temps real, que és una característica prou habitual quan es treballa amb volums de dades molt elevats. Per a il·lustrar el funcionament de BigQuery en aquests casos, crearem una nova taula que estarà vinculada, en aquest cas, a Google Drive, concretament a **Hojas de cálculo** (Figura 16). És molt important que el propietari de la taula sigui el mateix compte que està vinculat a BigQuery, perquè en cas contrari sorgeix un missatge d'error i no és capaç de vincular la taula externa.

Figura 16: Connexió a una taula externa

Si ens dirigim als detalls, aquí és on veiem quelcom interessant. La grandària de la taula és de zero bytes, atès que les dades són externes a BigQuery (Figura 17). Si ens desplaçem, podem veure els detalls de les dades externes. Això significa que quan actualitzem el full de càlcul, qualsevol consulta cap a aquesta taula recollirà automàticament les dades més recents. Ja que la nostra consulta de la taula gran no és només una còpia del full de càlcul, sinó que és de fet una referència a ella. Parlant de consultar les dades, ens dirigirem a Query, i a obrir un editor de consultes en una nova pestanya. Quan una consulta s'executa, totes les dades són retornades a nosaltres,

i podem accedir a elles com ho faríem amb qualsevol dada que resideixi en una taula nativa de BigQuery.

The screenshot shows the 'DETAILS' tab of a table named 'exemple' in the 'el-meu-projecte-358912.dades_proyecto' dataset. The table has an ID of 'el-meu-projecte-358912.dades_proyecto.exemple', a size of 0 B, and was created on 14 ago 2022, 15:46:09 UTC+2. It also shows the last modification on 14 ago 2022, 15:46:10 UTC+2, and an expiration date of 13 oct 2022, 15:46:09 UTC+2. The location is set to US. The 'DETAILS' tab also displays external data configuration, including the source URI (https://docs.google.com/spreadsheets/d/1d8JY1QH4GKjc6N8JFu_ueVJb6C9rl3qNrDnrDTeQEWM/edit#gid=0), automatic schema detection (true), and ignoring unknown values (false). The origin format is set to GOOGLE_SHEETS.

Figura 17: Característiques d'una taula externa

4 Integració de BigQuery amb Data Studio

Ara que hem cobert els diferents tipus de taules de BigQuery, ens centrarem en com podem visualitzar les nostres consultes.

Data Studio és una eina gratuïta que permet convertir les dades de les quals es disposa en panells o informes complets, fàcils de llegir, fàcils de compartir i totalment personalitzables. Algunes de les seves funcionalitats són:

- Descriure les dades amb gràfiques, que inclouen gràfics de línies, de barres i circulares, mapes geogràfics, gràfics d'àrea i de bombolles, taules de dades dinàmiques i molt més.
- Permet que els nostres informes siguin interactius amb filtres de visualització.
- Inclou enllaços i imatges en les quals es pot clicar per crear catàlegs de productes, biblioteques de vídeo i altres continguts amb URL.
- Facilita l'anotació i descripció dels informes amb text i imatges.

A més de presentar totes aquestes característiques, amb Data Studio es poden elaborar fàcilment informes sobre dades procedents d'una gran varietat de fonts, sense necessitat de programar. En tan sols uns instants, ens podem connectar a conjunts de dades com BigQuery.

4.1 Ús de Data Studio des de BigQuery

Imaginem que volem tornar a consultar el nombre d'accidents de trànsit que es van donar cada dia durant aquell mes als Estats Units. Per a fer aquesta consulta, farem servir la *query* anterior:

```
SELECT DAY, COUNT(*) AS FREQ
FROM `el-meu-projecte-358912.dades_projecte.accidents_bd`
GROUP BY DAY
```

The screenshot shows the Data Studio interface. At the top, there is a code editor window containing the SQL query:

```
1 SELECT DAY, COUNT(*) AS FREQ
2 FROM `el-meu-projecte-358912.dades_projecte.accidents_bd`
3 GROUP BY DAY
```

Below the code editor, there is a message: "Presiona Alt + F1 para ver las opciones de acceso".

The main area displays the results of the query:

INFORMACIÓN DEL TRABAJO	Resultados de la consulta	
Fila	DAY	FREQ
1	1	
2	2	
3	3	108
4	4	96

On the right side of the results table, there are two buttons: "GUARDAR LOS RESULTADOS" and "EXPLORAR DATOS". A tooltip for "EXPLORAR DATOS" says: "Analiza macrodatos con una conexión en vivo en una herramienta de hoja de cálculo conocida." Another tooltip for "Explorar con Data Studio" says: "Visualiza resultados y crea paneles en vivo a partir de tus datos."

Figura 18: Visualització d'una consulta a Data Studio

Per a visualitzar la taula resultant, podem ampliar el menú **Explorar dades**, tot seguit d'**Explorar amb Data Studio** (Figura 18).

Quan fem aquesta selecció, notaràs que ha sorgit una interfície, que ja té una taula que conté alguna informació agregada i un histograma amb la informació d'aquesta (Figura 19).

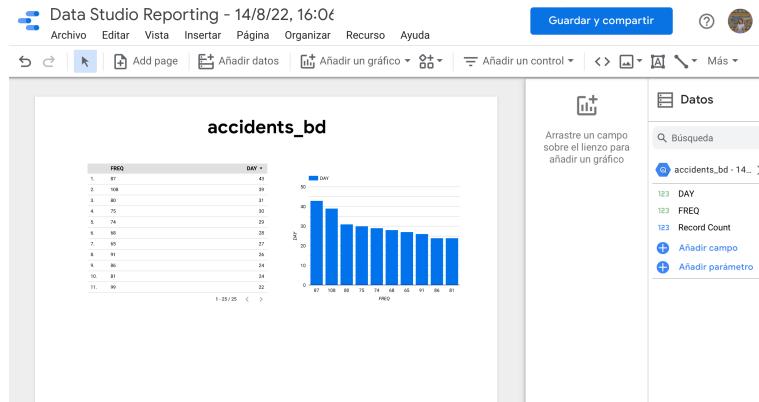


Figura 19: Consulta des de Data Studio

Data Studio, per defecte, ha entès que una taula de recompte es visualitza normalment a partir d'un *bar plot* o histograma, i per això l'ha creat sense que nosaltres ho hagim especificat. Així i tot, nosaltres podem seleccionar una visualització abans de configurar-la per a presentar la informació que necessitem. Pose'm-nos en el cas que preferim un gràfic de barres per a la visualització d'aquestes dades. Premem l'opció **Afegir un gràfic** i ens assegurem que la dimensió, que en aquest cas són els dies, i la mètrica, la freqüència dels accidents, estiguin seleccionades segons el que vulguem representar (Figura 20).

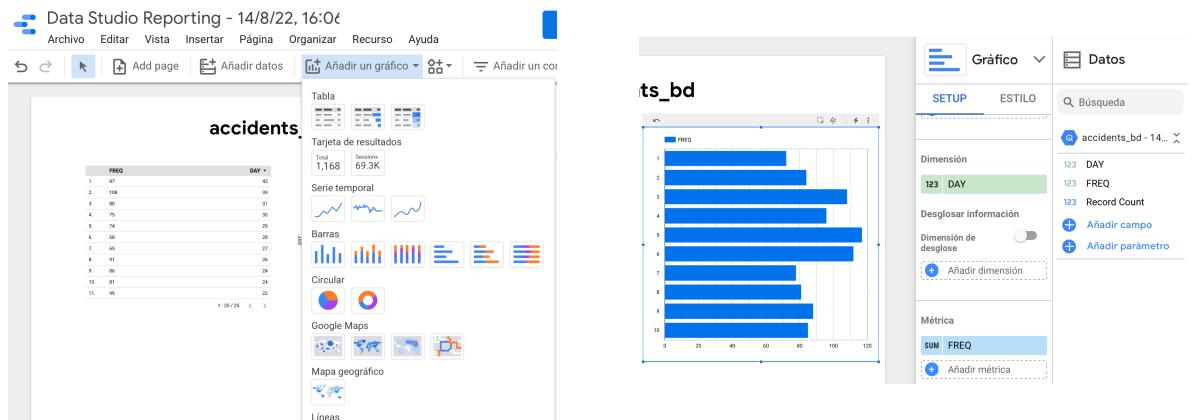


Figura 20: Creació d'un gràfic a partir de Data Studio

4.2 Ús de Data Studio des de la pròpia plataforma

Per a utilitzar aquesta eina és necessari disposar d'un compte a Google. Per accedir a la pàgina, naveguem [a aquesta pàgina](#) i iniciem sessió amb el nostre compte de Google. Un

Un cop dins de Data Studio, es troba una pantalla d'inici amb les característiques següents:



Figura 21: Inici de Data Studio

A la pàgina d'inici hi ha una sèrie de plantilles, que són una forma entretinguda d'explorar les capacitats de Data Studio. En el nostre cas, com el que ens interessa és crear un informe des de zero, clicarem a **Informe vacío**. Un cop dins l'informe en blanc, s'hauran d'afegir les dades que volem representar en aquest (Figura 22). Tenim moltes opcions a l'hora d'escol·lar la font de les dades, però per a fer-ho més senzill vincularem l'informe a BigQuery, específicament al nostre conjunt de dades `accidents_bd`.

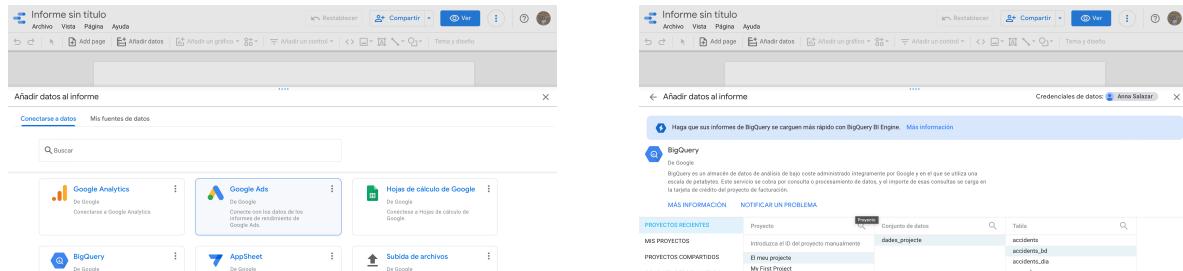


Figura 22: Afegir dades a Data Studio

Per a crear un gràfic, podem clicar a **Añadir un gráfico** i, d'aquesta manera, escol·lar el que ens interessa representar en el panell de la dreta, ajustant la dimensió i la mètrica del nostre interès. Un cop creat el gràfic, es pot editar la seva mida fent-lo més gran o més petit, segons la nostra preferència, i es pot moure de lloc dins la pàgina.

A més de gràfiques, es poden decorar les pàgines per fer-les més boniques, o per afegir informació. També podem afegir un nom a la pàgina, i crear-ne tantes com en necessitem per al nostre informe. Més enllà d'això, també podem establir un nom per a tot l'informe, perquè a l'hora de guardar-lo i compartir-lo sigui més fàcilment interpretable.

Per exemple, un informe de dues pàgines pot tenir un aspecte semblant a la Figura 23.

Per acabar, podem compartir el nostre informe de manera que altres usuaris puguin tenir accés. Això ho podem fer des de la pantalla d'inici de Data Studio, seleccionant l'opció de **Compartir** l'informe i afegint les direccions de correu dels qui vulguem fer



Figura 23: Informe a Data Studio

lectors de l'informe (aqueells qui poden veure l'informe, però no tenen permisos d'edició) o bé *editors* (poden veure i editar l'informe). Figura 24

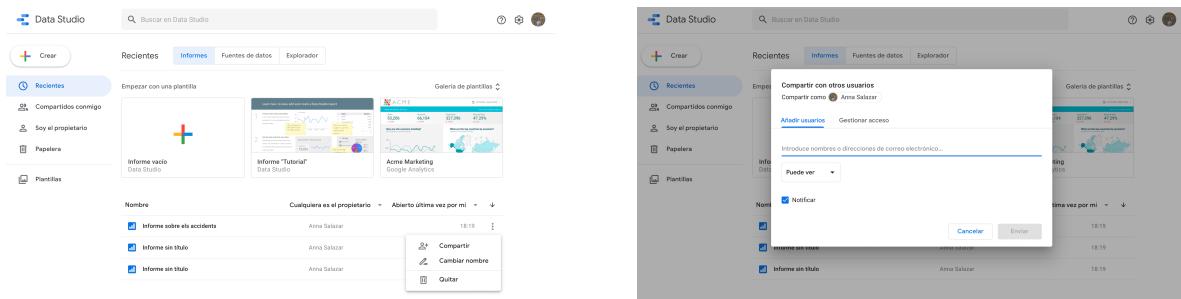


Figura 24: Compartir un informe

5 Connexió d'R a BigQuery

5.1 Connectem R a BigQuery

Per a entendre com funciona la connexió entre BigQuery i R, primer s'ha de definir el concepte de *API* o interfície de programació d'aplicacions.

Una API és un conjunt de processos, funcions i mètodes que ofereix una determinada biblioteca de programació que serveix com a capa d'abstracció perquè sigui emprada per un altre programa informàtic. Dit d'una altra manera, podem entendre les APIs com un codi que indica a les aplicacions com comunicar-se entre elles, de manera que pugui interactuar.

En el nostre cas, volem que R sigui capaç d'interactuar amb BigQuery, perquè així ens permeti realitzar les nostres consultes a la base de dades sense haver d'importar les dades a R. Haurem d'utilitzar les llibreries `bigrquery`¹ i `DBI`² per aconseguir-ho.

```
install.packages("bigrquery")
library(bigrquery)
library(DBI)
```

Asignem a l'objecte `projecte` el nom del nostre projecte, que a l'exemple correspon a *el-meu-projecte-358912*

```
projecte <- "el-meu-projecte-358912"
```

I per últim, mitjançant la funció `dbConnect()` obrirem la connexió amb BigQuery indicant el nom del nostre projecte i la base de dades a la que volem accedir dins d'aquest, `dades_projecte`. A l'apartat `billing` s'haurà d'introduir l'identificador del projecte amb la font de facturació. Nosaltres indiquem el nom del nostre projecte, que està sotmès a les limitacions de la *Sandbox* o zona de proves gratuïta.

```
dades <- dbConnect(
  bigrquery::bigrquery(),
  project = projecte,
  dataset = "dades_projecte",
  billing = projecte
)
```

Executant aquest codi no es produeix gran cosa, tret que es crea una variable de connexió. Però la primera vegada que intentem fer ús d'aquesta connexió (per exemple, fent una consulta a una de les taules de la base de dades), se'ns demana que ens autentifiquem a través del nostre compte de Google en una finestra del navegador. Un cop fet això, ja podrem començar a consultar les nostres dades de BigQuery, així com els conjunts de dades públics.

¹Una interfície per a la API de 'BigQuery' de Google

²Una definició d'interfície de base de dades per a la comunicació entre R i els sistemes de gestió de bases de dades relacionals. Totes les classes d'aquest paquet són virtuals i han de ser esteses per les diferents implementacions de R/SGBD.

5.2 Consultes amb ‘bigrquery’

Per a mostrar com seria una consulta a la nostra taula des de R buscarem el llistat d'accidents que van ocórrer un cap de setmana a les 21:00.

```
query1 <- "SELECT ST_CASE  
          FROM `el-meu-projecte-358912.dades_projecte.accidents_bd`  
          WHERE (DAY = 1 or DAY = 7) and HOUR = 21 and MINUTE = 0"
```

Un cop hem formulat la nostra *query*, a través de la funció `dbGetQuery()`, ens comunicarem amb la API de Google.

```
dbGetQuery(connexió, consulta, n)      dbGetQuery(dades, query1, n = 10)
```

Essent *n* el nombre de casos a mostrar (si calen restriccions).

En executar aquesta consulta, R retorna la següent taula:

ST_CASE
62353

Aquesta correspon a que únicament va ocórrer un accident un cap de setmana a les 21:00, i aquest té l'identificador 62353.

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	Inserció de dades a la taula	6
8	Lectura d'un arxiu extern	7
9	Informació sobre la taula	8
10	Elaboració d'una consulta	8
11	Primera consulta	9
12	Creació d'una taula a partir d'una consulta	10
13	Accés als conjunts de dades públiques	11
14	Creació d'una taula a partir d'una consulta	12
15	Consulta a un conjunt de dades públic	12
16	Connexió a una taula externa	13
17	Característiques d'una taula externa	14
18	Visualització d'una consulta a Data Studio	15
19	Consulta des de Data Studio	16
20	Creació d'un gràfic a partir de Data Studio	16
21	Inici de Data Studio	17
22	Afegir dades a Data Studio	17
23	Informe a Data Studio	18
24	Compartir un informe	18