

Documento Técnico

Adrián Otero Rodríguez

Contents

Infraestructura Cloud	3
Bucket en Google Cloud (Cloud Storage)	3
Máquina virtual en Google Cloud (Compute engine)	4
Base de datos MySQL en Google Cloud (Cloud SQL).....	6
Configurar redes en Google Cloud	10
Instalar Software en la VM (Compute Engine).....	12
Herramienta de visualización – Qlik Sense	15
Conectar Qlik Sense con nuestra base de datos	15
Redes Sociales – Twitter API	18
Repositorio GIT.....	20

Infraestructura Cloud

La infraestructura necesaria para resolver el problema será planteada sobre la plataforma de Google Cloud.

Esta plataforma ofrece diversos componentes que permiten acelerar la creación de recursos y el desarrollo de aplicaciones de ML.

Bucket en Google Cloud (Cloud Storage)

En primer lugar, vamos a crear un bucket en Cloud Storage donde almacenar los datos en bruto, sin procesar. Para ello, accedemos a través del menú de navegación a **“Storage”** y configuramos las características del bucket, incluyendo: nombre, tipo de almacenamiento, región, etc.

Google Cloud Platform ea-datascientist-test

Storage

← Crear un segmento

- Asigna un nombre a tu segmento**
Elige un nombre **global** único y permanente. [Directrices de nomenclatura](#)

Nota: No incluyas información sensible.
CONTINUAR
- Elige dónde quieres almacenar los datos**
Esta selección permanente define el emplazamiento geográfico de los datos y afecta al coste, el rendimiento y la disponibilidad. [Más información](#)
Tipo de ubicación
 - ☒ **Region**
Minima latencia en una sola región
 - ☐ **Multi-region**
Máxima disponibilidad en una zona más extensa
 - ☐ **Dual-region**
Alta disponibilidad y baja latencia en 2 regiones**Ubicación**

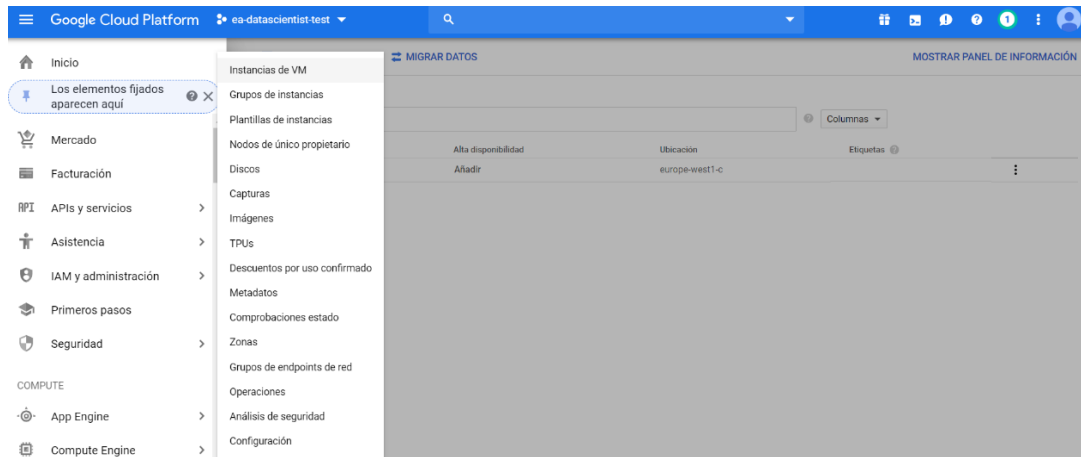
CONTINUAR
- Elige la clase de almacenamiento de datos predeterminada**
Las clases de almacenamiento establecen los costes del almacenamiento, la recuperación y las operaciones. Elige una clase de almacenamiento según la frecuencia con la que tienes previsto acceder a los datos. [Más información](#)
 - ☒ **Standard** ⓘ
Ideal para los datos a los que se accede con frecuencia
 - ☐ **Nearline**
Ideal para copias de seguridad y datos a los que se accede una vez al mes o menos
 - ☐ **Coldline**
Ideal para recuperaciones tras fallos y datos a los que se accede una vez al año o menos.

Una vez establecidos todos los parámetros como en la imagen anterior, tendremos nuestro bucket disponible.

Máquina virtual en Google Cloud (Compute engine)

A continuación, vamos a desplegar una máquina virtual en el entorno de Google Cloud. Sobre esta máquina virtual desarrollaremos los códigos en Python destinados a la extracción, transformación y carga de los datos, sobre los que posteriormente se realizarán distintos análisis.

Para ello seleccionamos **“Compute Engine -> Instancias de VM”** en el menú de navegación.



Y configuramos los detalles de la instancia, incluyendo: nombre, región y zona, tipo de máquina, distribución del SO, permisos, reglas del cortafuegos, redes, discos, etc.

← Crear una instancia

Para crear una instancia de VM, selecciona una de las opciones:

- Nueva instancia de VM**
Crea una instancia de VM desde cero
- Nueva instancia de VM a partir de una plantilla**
Crea una instancia de VM a partir de una plantilla disponible
- Marketplace**
Despliega una solución lista para usarse en una instancia de VM

Nombre ?
ea-vm-dev

Región ?
europe-west1 (Bélgica)

Zona ?
europe-west1-b

Configuración de la máquina ?

Familia de máquinas
Uso general Con memoria optimizada
Tipos de máquinas para cargas de trabajo habituales, optimizadas en cuanto al coste y a la flexibilidad

Generación
Primera
Con la tecnología de la plataforma de CPU de Skylake o de uno de sus predecesores

Tipo de máquina
n1-standard-4 (4 vCPU, 15 GB de memoria)


	vCPU	Memoria
	4	15 GB

⌄ Plataforma de CPU y GPU

Contenedor ?

☐ Desplegar una imagen de contenedor en esta instancia de VM. [Más información](#)

Disco de arranque ?



Nuevo disco persistente estándar de 10 GB
Imagen
Ubuntu 16.04 LTS

Cambiar

Identidad y acceso de API ?

Cuenta de servicio ?
Compute Engine default service account

Alcance del acceso ?

- ☒ Permitir el acceso predeterminado
- ☐ Permitir el acceso completo a todas las API de Cloud
- ☐ Definir acceso para cada API

Cortafuegos ?

Añade reglas de cortafuegos y etiquetas para permitir tráfico de red concreto de Internet

☒ Permitir el tráfico HTTP

☒ Permitir el tráfico HTTPS

?

Puedes usar Cloud DNS para servir tráfico desde esta VM a tu dominio.

Mostrar procedimiento

AdministraciónSeguridadDiscosRedesÚnico cliente

Disco de arranque

Regla de eliminación

☐ Eliminar el disco de arranque cuando se elimine la instancia

Encriptado

Los datos se encriptan automáticamente. Selecciona una solución para administrar claves de encriptado.

- ☒ Clave administrada por Google
No se requiere configuración
- ☐ Clave administrada por el cliente
Administración a través de Google Cloud Key Management Service
- ☐ Clave proporcionada por el cliente
Administración fuera de Google Cloud

Nombre de dispositivo ?

Se usa para hacer referencia al dispositivo en procesos de activación o modificación de tamaño.

Basado en el nombre de la instancia (predeterminado)

ea-vm-dev

Discos adicionales ? (Opcional)

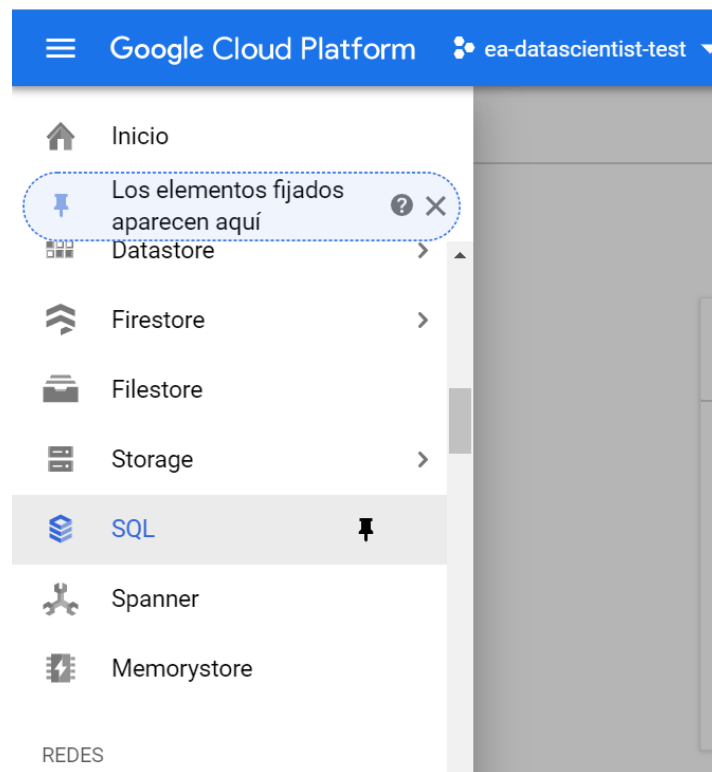
+ Añadir disco nuevo

+ Acoplar disco disponible

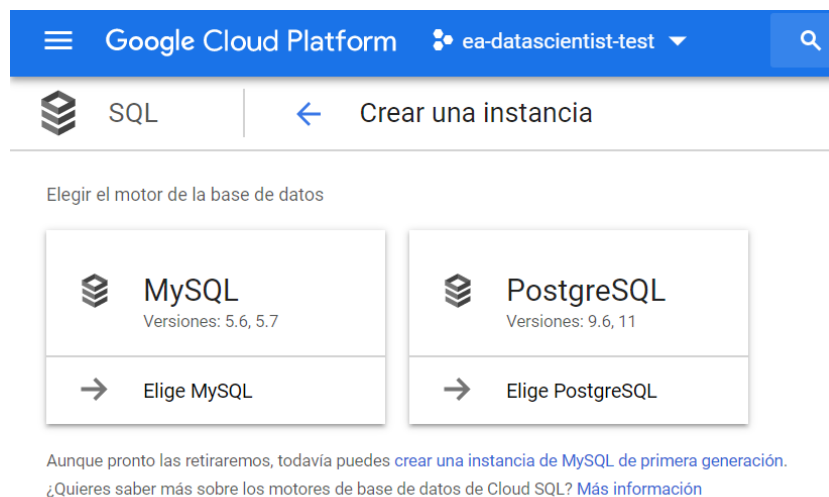
Finalmente, tras unos minutos, tenemos nuestra instancia de VM en funcionamiento.

Base de datos MySQL en Google Cloud (Cloud SQL)

A continuación, desplegaremos una base de datos MySQL en la que se cargarán los datos proporcionados. Para ello es necesario seleccionar Cloud “SQL” el menú de navegación.



Elegir una instancia MySQL.



Y configurar los detalles de la instancia, incluyendo: id, password del usuario root, región y zona, versión MySQL, conectividad, tipo de máquina y almacenamiento.



Información de la instancia

ID de instancia

La elección es permanente. Usa letras minúsculas, números y guiones, y empieza por una letra.

Contraseña "root"

Introduce una contraseña para el usuario raíz. [Más información](#)

[Generar](#)☐ Sin contraseña

Ubicación

Para mejorar el rendimiento, almacena los datos cerca de los servicios que los necesitan.

Región

La elección es permanente

Zona

Puede modificarse en cualquier momento

Versión de la base de datos

Opciones de configuración

☒ Conectividad



Elige cómo quieres conectarte a tu instancia de base de datos.

Como medida de seguridad adicional, puedes usar el proxy de Cloud SQL para conectarte a tus instancias después de crearlas. [Más información](#)

☐ IP privada

La conexión IP privada requiere otras API y permisos. Es posible que necesites ponerte en contacto con el administrador de tu organización para que te ayude a habilitar o utilizar esta función. Actualmente, la conexión IP privada no puede inhabilitarse una vez que se ha habilitado.

☒ IP pública

No has autorizado a ninguna red externa a conectarse a tu instancia de Cloud SQL. Las aplicaciones externas pueden conectarse a la instancia a través del proxy de Cloud SQL. [Más información](#)

Redes autorizadas

Autoriza una red o usa un proxy para conectarte a tu instancia. Las redes solo obtendrán autorización mediante las direcciones que indiques. [Más información](#)

[+ Añadir red](#)

✓ Tipo de máquina y almacenamiento ^

Tipo de máquina ?

Para mejorar el rendimiento, selecciona un tipo de máquina con suficiente memoria como para que quepa la tabla más grande



db-n1-standard-1

vCPU

1

Memoria

3,75 GB

Cambiar

Rendimiento de red (MB/s) ?

250 de 2.000



Tipo de almacenamiento ?

La elección es permanente.

☒ SSD (recomendado)

Opción más popular. Latencia más baja que HDD. Número de consultas por segundo y rendimiento de datos más altos.

☐ HDD

Rendimiento más bajo que SSD con niveles de almacenamiento inferiores.

Capacidad de almacenamiento ?

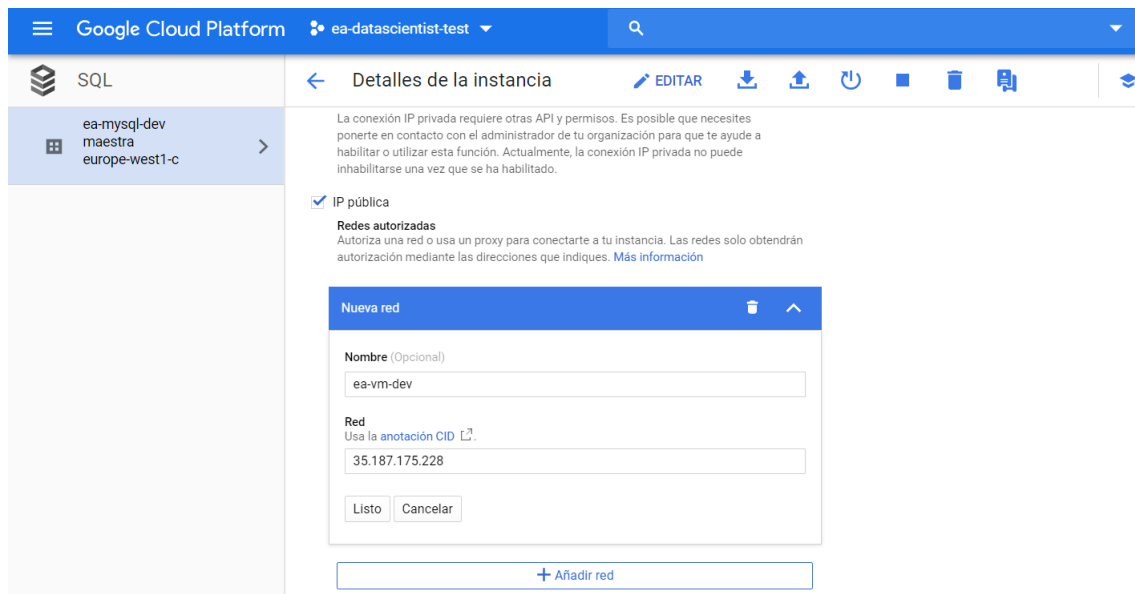
10 – 30720 GB. Cuanto mayor es la capacidad, mejor es el rendimiento (hasta el límite establecido por el tipo de máquina). La capacidad no se puede reducir más adelante.

10 GB

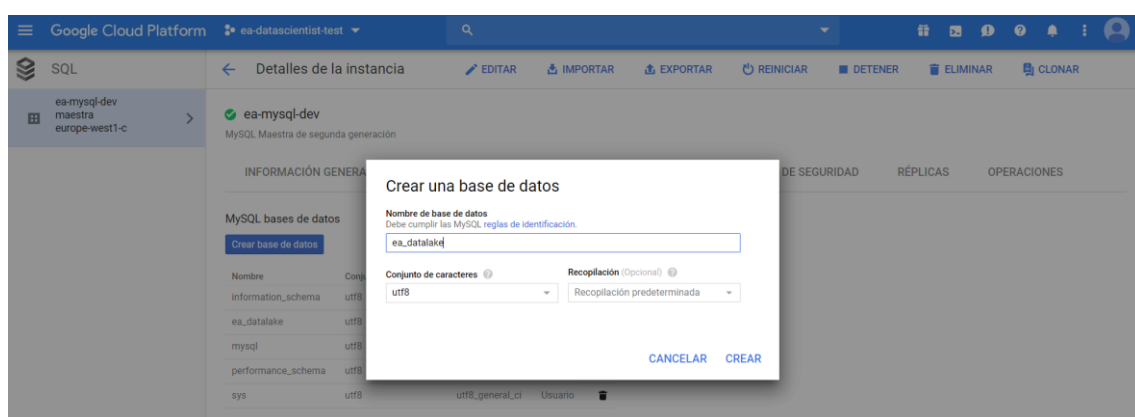
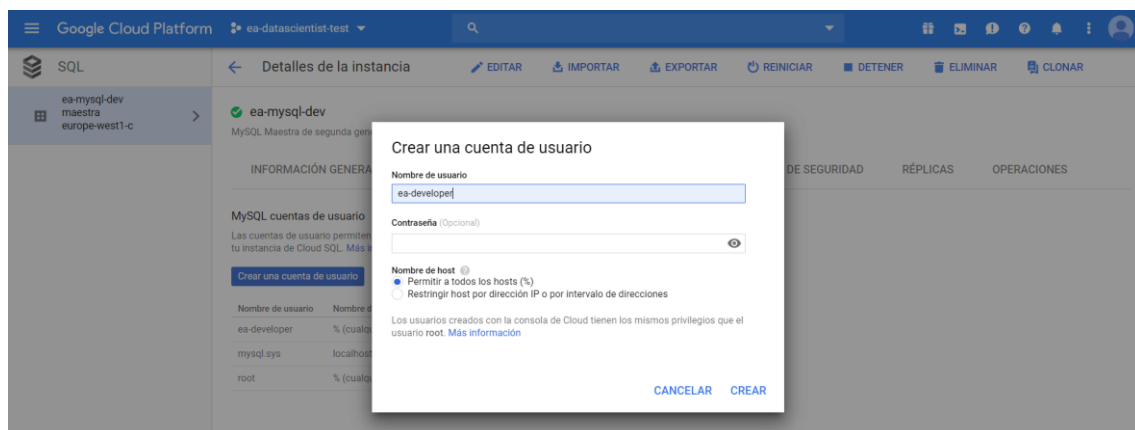
Finalmente, en unos minutos tendremos nuestra base de datos MySQL en funcionamiento:

Google Cloud Platform ea-datascientist-test				
SQL	Instancias	CREAR INSTANCIA	MIGRAR DATOS	MOSTRAR PANEL DE INFORMACIÓN
Filtrar instancias				
Columnas				
ID de instancia	Tipo	Alta disponibilidad	Ubicación	Etiquetas
<input checked="" type="checkbox"/> ea-mysql-dev	MySQL 2ª gen. 5.7	Añadir	eu-west1-c	

A continuación, vamos a autorizar las conexiones desde la instancia VM de Google Compute Engine que hemos creado.

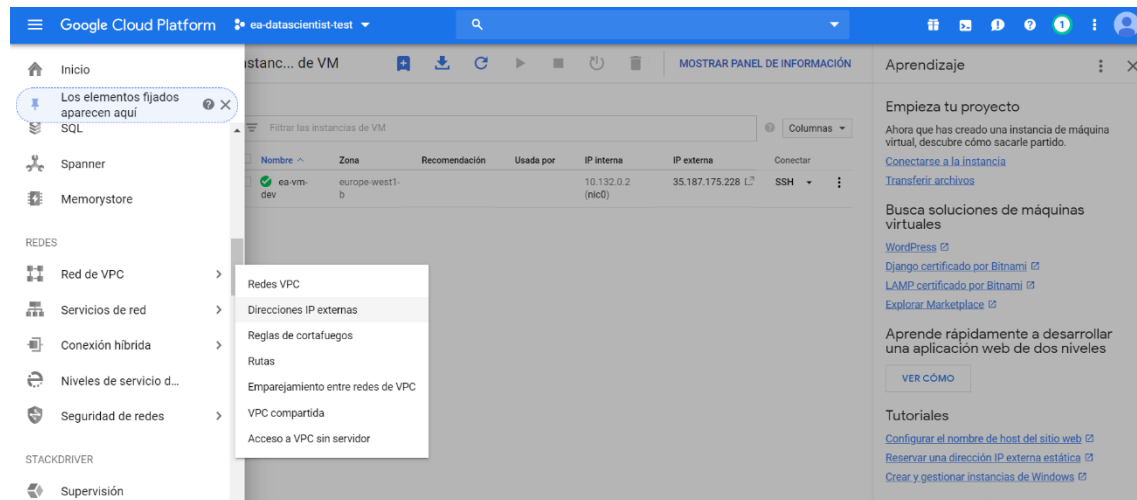


Por último, vamos a crear un usuario y una base de datos dentro de nuestra instancia Cloud SQL.

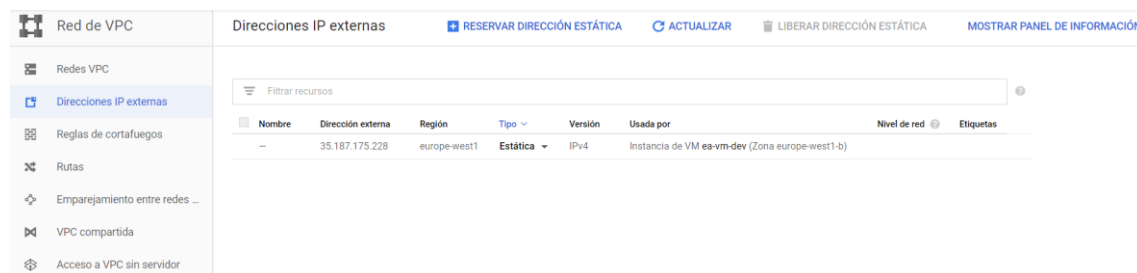


Configurar redes en Google Cloud

En el siguiente paso, vamos a hacer que la ip externa de la vm sea estática de cara a acceder posteriormente a nuestro Jupyter Notebook. Para ello, debemos acceder a “**Red de VPC -> Direcciones IP externas**” y reservar una nueva ip estática.



Finalmente, tenemos nuestra dirección IP externa estática: **35.187.175.228**



Por último, vamos a crear una regla que permita el acceso a través del puerto 8888 (para poder acceder posteriormente a nuestro Jupyter notebook). Por simplicidad, **aunque no es lo recomendable**, vamos a abrirlo a todos los rangos IP (0.0.0.0/0).

Red de VPC

Redes VPC

Direcciones IP externas

Reglas de cortafuegos

Rutas

Emparejamiento entre redes ...

VPC compartida

Acceso a VPC sin servidor

← Crear una regla de cortafuegos

Las reglas de cortafuegos controlan el tráfico entrante o saliente de una instancia. De forma predeterminada, se bloquea el tráfico entrante que sea ajeno a tu red. [Más información](#)

Nombre ?

Descripción (Opcional)

Registros
Activar los registros del cortafuegos puede generar un gran número de registros, lo que podría aumentar los costes de Stackdriver. [Más información](#)

☐ Activar
☒ Desactivar

Red ?

default

Prioridad ?
La prioridad puede estar entre 0 y 65535 [Comprobar la prioridad de otras reglas de cortafuegos](#)

Dirección del tráfico ?

☒ Entrada
☐ Salida

Red de VPC

Redes VPC

Direcciones IP externas

Reglas de cortafuegos

Rutas

Emparejamiento entre redes ...

VPC compartida

Acceso a VPC sin servidor

← Crear una regla de cortafuegos

Destinos ?

Todas las instancias de la red

Filtro de origen ?

Intervalos de IPs

Intervalos de IPs de origen ?

0.0.0.0/0

Filtro de origen secundario ?

Ninguno

Protocolos y puertos ?

☐ Permitir todos
☒ Protocolos y puertos especificados

☒ tcp :
☐ udp :
☐ Otros protocolos

[Inhabilitar regla](#)

Crear

Cancelar

Equivalent [REST](#) or [command line](#)

Instalar Software en la VM (Compute Engine)

Por último, vamos a configurar todo el software necesario en nuestra instancia, incluyendo tanto programas (Anaconda + Python) como librerías.

En primer lugar, es necesario acceder a la VM en Google Cloud mediante SSH.

Instancias de VM						
<div>Filtrar las instancias de VM</div>						
<input type="checkbox"/>	Nombre ^	Zona	Recomendación	Usada por	IP interna	IP externa
<input type="checkbox"/>	ea-vm-dev	europe-west1-b			10.132.0.2 (nic0)	35.187.175.228
						Conectar
						SSH

Se deben ejecutar los siguientes comandos para instalar y configurar **Anaconda** y preparar la instalación para ejecutar nuestros Jupyter Notebooks:

wget http://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh

```
adrian_otero_ea_case@ea-vm-dev:~$ wget http://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh
--2019-10-14 09:01:49-- http://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh
Resolving repo.continuum.io (repo.continuum.io)... 104.18.201.79, 104.18.200.79, 2606:4700::6812:c94f, ...
Connecting to repo.continuum.io (repo.continuum.io)|104.18.201.79|:80... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: https://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh [following]
--2019-10-14 09:01:49-- https://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh
Connecting to repo.continuum.io (repo.continuum.io)|104.18.201.79|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 417798602 (398M) [application/x-sh]
Saving to: 'Anaconda3-4.0.0-Linux-x86_64.sh'

Anaconda3-4.0.0-Linux-x86_64 100%[=====>] 398.44M  89.3MB/s   in 4.4s

2019-10-14 09:01:54 (91.0 MB/s) - 'Anaconda3-4.0.0-Linux-x86_64.sh' saved [417798602/417798602]
```

bash Anaconda3-4.0.0-Linux-x86_64.sh

```
adrian_otero_ea_case@ea-vm-dev:~$ bash Anaconda3-4.0.0-Linux-x86_64.sh

Welcome to Anaconda3 4.0.0 (by Continuum Analytics, Inc.)

In order to continue the installation process, please review the license
agreement.
Please, press ENTER to continue
>>>
=====
Anaconda License
=====

Copyright 2016, Continuum Analytics, Inc.
All rights reserved under the 3-clause BSD License:

installing: anaconda-4.0.0-mp10py35_0 ...
installing: conda-4.0.5-py35_0 ...
installing: conda-build-1.20.0-py35_0 ...
installing: conda-env-2.4.5-py35_0 ...
Python 3.5.1 :: Continuum Analytics, Inc.
creating default environment...
installation finished.
Do you wish the installer to prepend the Anaconda3 install location
to PATH in your /home/adrian_otero_ea_case/.bashrc ? [yes|no]
[no] >>> yes

Prepending PATH=/home/adrian_otero_ea_case/anaconda3/bin to PATH in /home/adrian_otero_ea_case/.bashrc
A backup will be made to: /home/adrian_otero_ea_case/.bashrc-anaconda3.bak

For this change to become active, you have to open a new terminal.

Thank you for installing Anaconda3!

Share your notebooks and packages on Anaconda Cloud!
Sign up for free: https://anaconda.org
```

source ~/.bashrc

jupyter notebook --generate-config

vi ~/.jupyter/jupyter_notebook_config.py

```
# CFileContentsManager.save_script = False

# Configuration file for jupyter-notebook.
#-----
# Configurable configuration
#-----
c = get_config()
c.NotebookApp.ip = '*'
c.NotebookApp.open_browser = False
c.NotebookApp.port = 8888

#-----
# LoggingConfigurable configuration
#-----

# A parent class for Configurables that log.
#
# Subclasses have a log trait, and the default behavior is to get the logger
# from the currently running Application.
#-----
# SingletonConfigurable configuration
```

Añadir al fichero de configuración las 4 líneas de la imagen anterior, incluyendo el puerto configurado previamente en la regla del firewall (8888).

A continuación, vamos a instalar algunos paquetes y librerías de Python usando los siguientes comandos:

sudo apt update

sudo apt upgrade

sudo apt install zip

sudo apt-get install mysql-client

sudo apt-get install libsm6 libxext6 libxrender-dev

sudo apt-get install texlive texlive-latex-extra pandoc

conda update pandas

pip install --upgrade pip

pip install mysql-connector

pip install mysql-connector-python-rf

pip install pandas_profiling

pip install kaggle

pip install google-cloud-storage

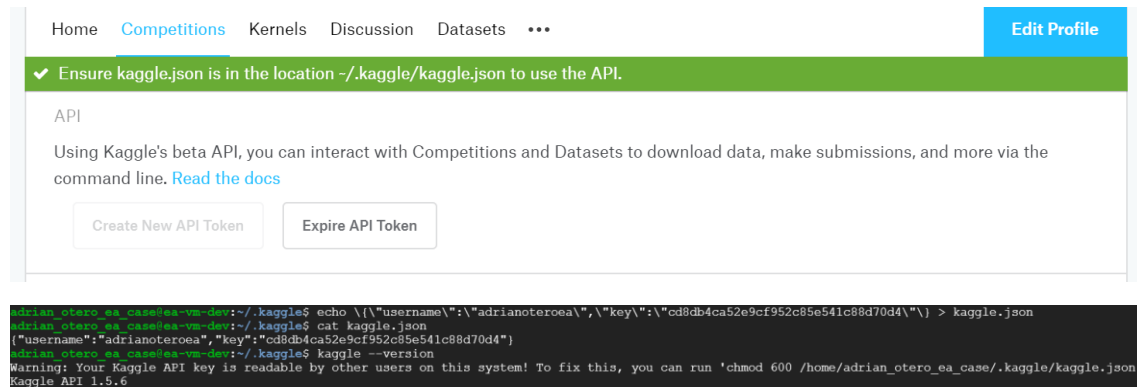
pip install google-cloud-bigquery

pip install google-cloud-language

pip install pyarrow

pip install tweepy

Por último, vamos a generar las credenciales de Kaggle que más adelante necesitaremos para acceder a algunos datos:



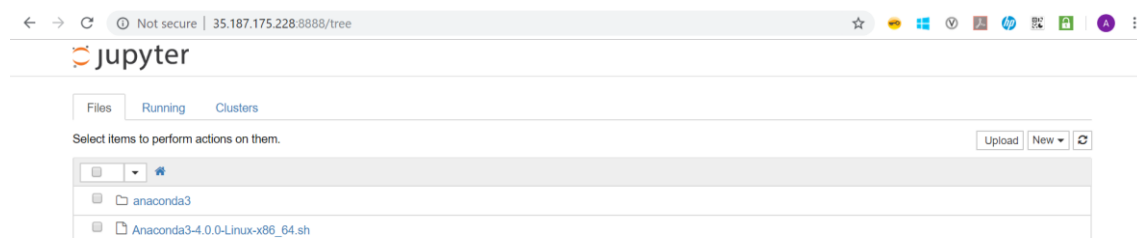
The image shows the Kaggle website's API setup page. At the top, there are navigation links: Home, Competitions, Kernels, Discussion, Datasets, and an Edit Profile button. A green banner states: "Ensure kaggle.json is in the location ~/.kaggle/kaggle.json to use the API." Below this, the "API" section explains that using Kaggle's beta API allows interaction with Competitions and Datasets via the command line, with a link to "Read the docs". Two buttons are present: "Create New API Token" and "Expire API Token".

Below the website interface is a terminal window showing the following commands and output:

```
adrian_otero_ea_case@ea-vm-dev:~/.kaggle$ echo "{\"username\":\"adrianoteroea\",\"key\":\"cd8db4ca52e9cf952c85e541c88d70d4\"}" > kaggle.json
adrian_otero_ea_case@ea-vm-dev:~/.kaggle$ cat kaggle.json
{"username":"adrianoteroea","key":"cd8db4ca52e9cf952c85e541c88d70d4"}
adrian_otero_ea_case@ea-vm-dev:~/.kaggle$ kaggle --version
Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 /home/adrian_otero_ea_case/.kaggle/kaggle.json'
Kaggle API 1.5.6
```

Con todo listo, ejecutemos el notebook de Python mediante el comando: **jupyter notebook**

Este Jupyter Notebook será accesible a través de la ip estática y puerto configurados en los pasos previos, en este caso: <http://35.187.175.228:8888>



The image shows the Jupyter Notebook web interface in a browser. The address bar shows "Not secure | 35.187.175.228:8888/tree". The Jupyter logo is at the top left. Below it are tabs for "Files", "Running", and "Clusters". A message says "Select items to perform actions on them." with "Upload" and "New" buttons. A list of files is shown:

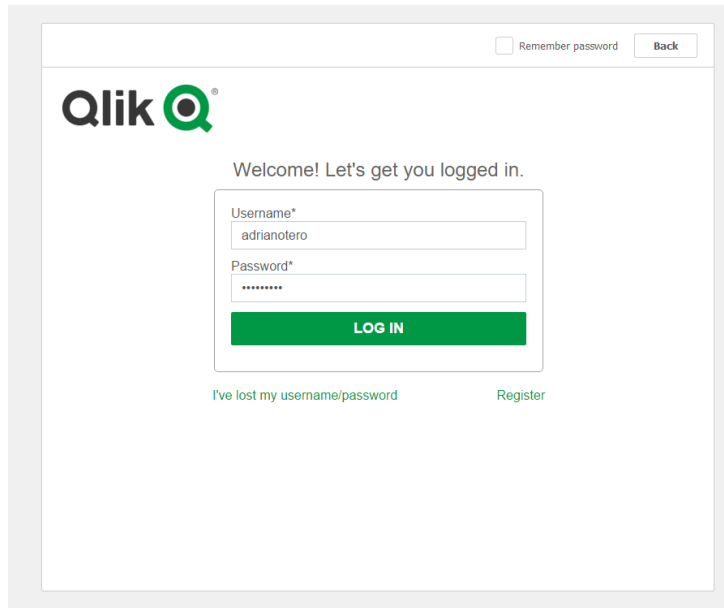
File Name
anaconda3
Anaconda3-4.0.0-Linux-x86_64.sh

Herramienta de visualización – Qlik Sense

Como herramienta de visualización se utilizará Qlik Sense. En primer lugar, es necesario descargar e instalar la aplicación. Es posible descargarla de forma gratuita a través de su web oficial:

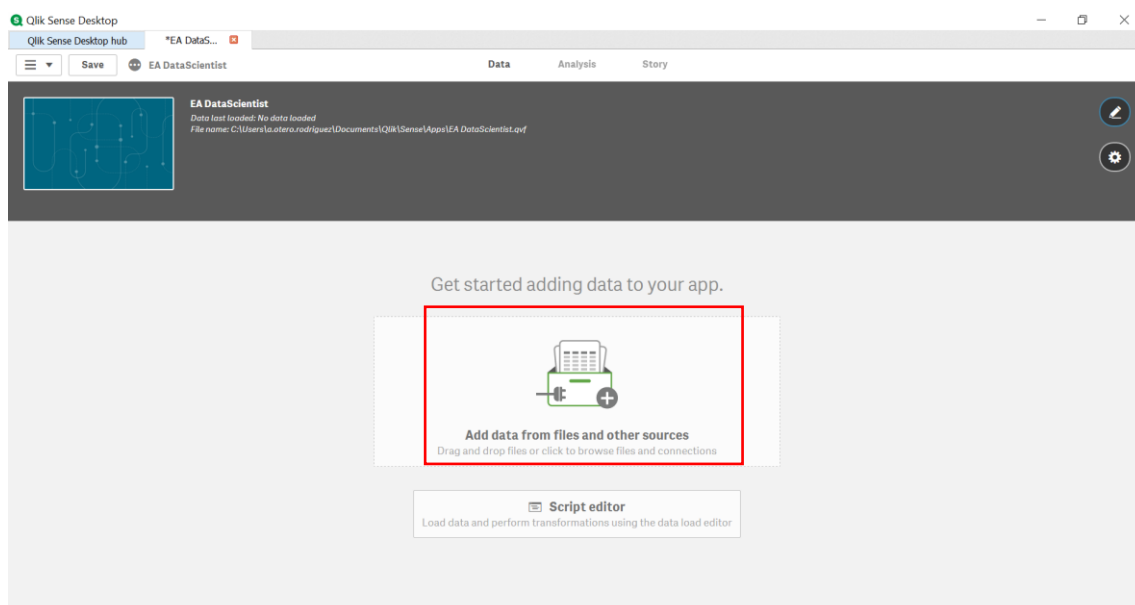
<https://www.qlik.com/es-es/try-or-buy/download-qlik-sense>

Para poder obtener una copia, es necesario registrarse en su web y aceptar el correo de activación. Por otro lado, para instalar Qlik se aceptarán todas las opciones configuradas por defecto.

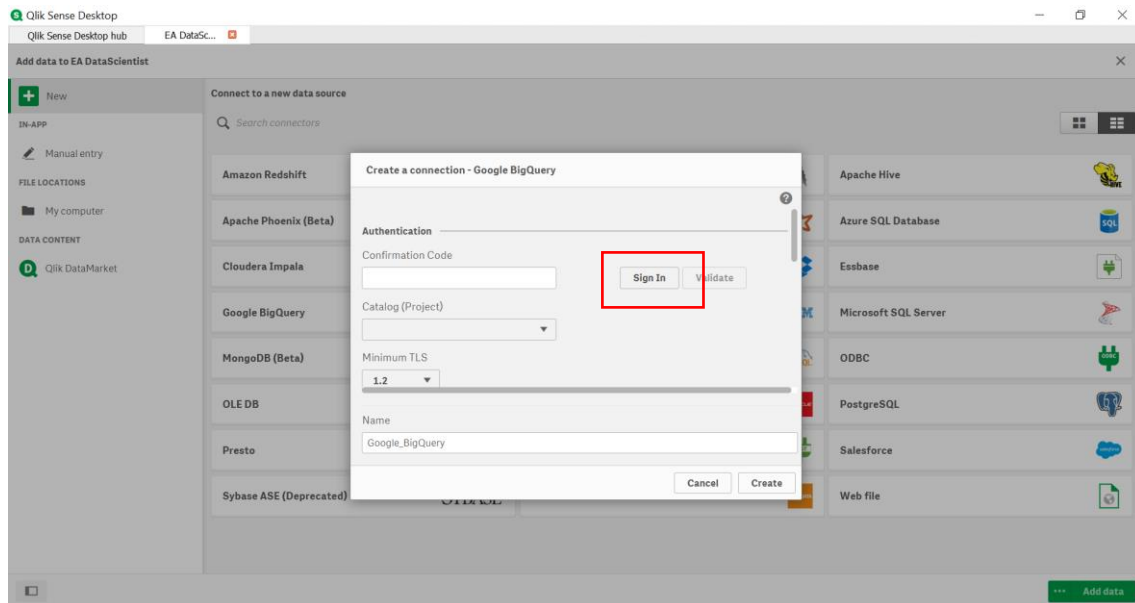
The image shows the Qlik Sense login interface. At the top right, there is a checkbox for "Remember password" and a "Back" button. The Qlik logo is prominently displayed on the left. The main heading says "Welcome! Let's get you logged in." Below this, there are two input fields: "Username*" with the text "adrianotero" and "Password*" with masked characters. A green "LOG IN" button is positioned below the password field. At the bottom, there are two links: "I've lost my username/password" and "Register".

Conectar Qlik Sense con nuestra base de datos

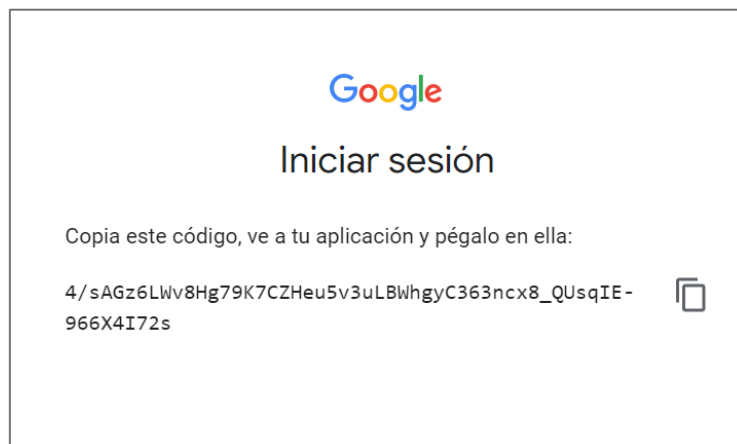
En primer lugar, debemos crear una nueva APP y seleccionar la opción de añadir una nueva fuente, desde aquí configuraremos una nueva configuración con Google Bigquery:



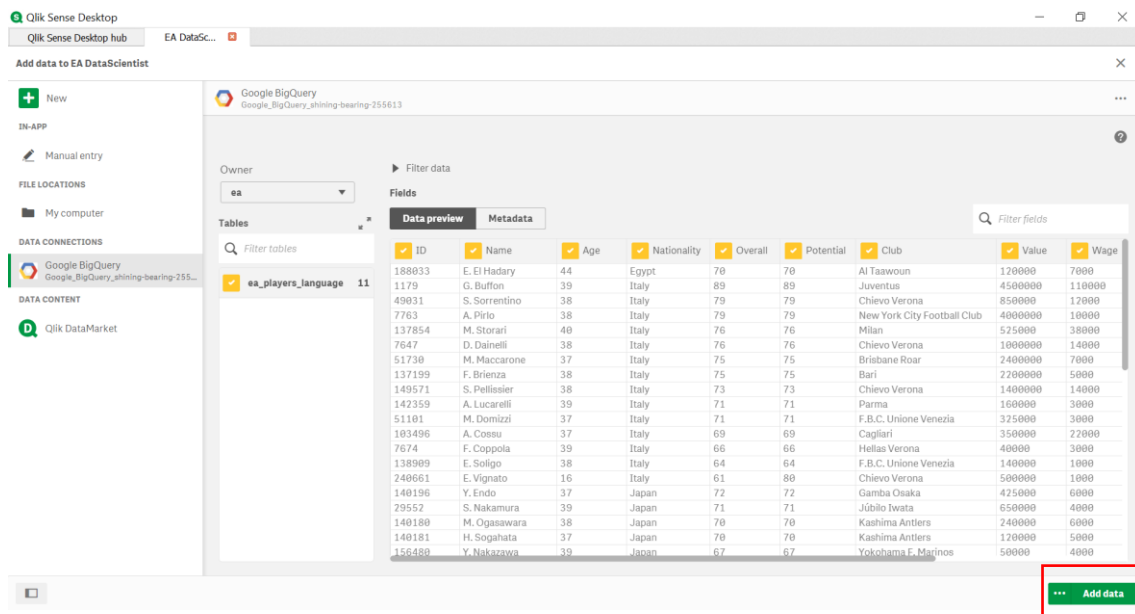
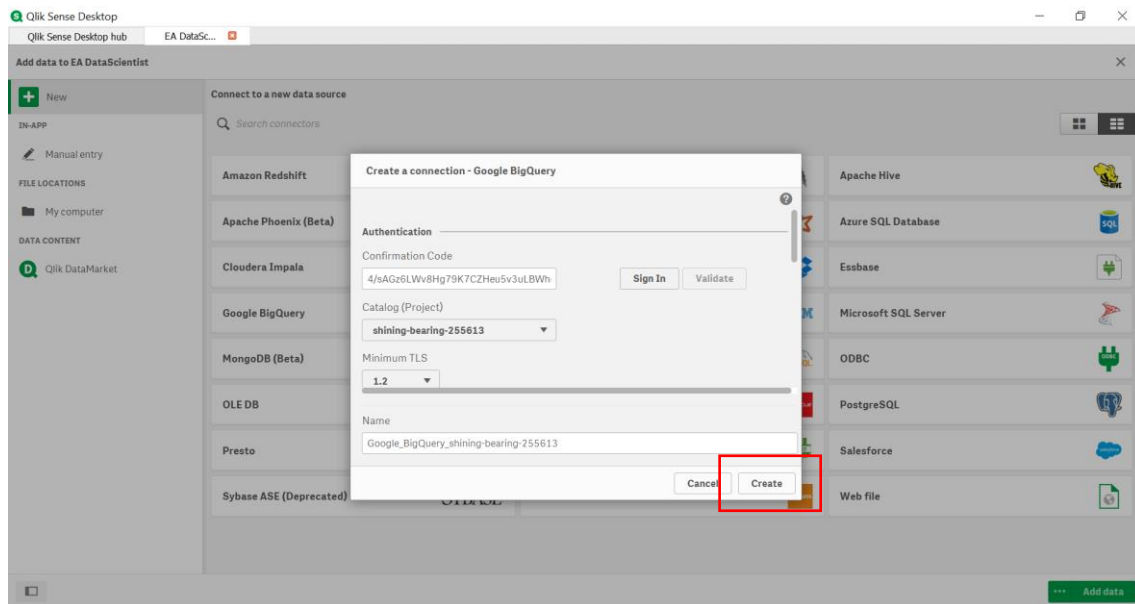
Qlik cuenta con conectores nativos con Google Biquery que permiten acceder a la información almacenada en nuestra base de datos de forma sencilla.



Durante la configuración, necesitaremos autorizar el acceso a nuestra cuenta en Google Cloud y descargar las credenciales recibidas:

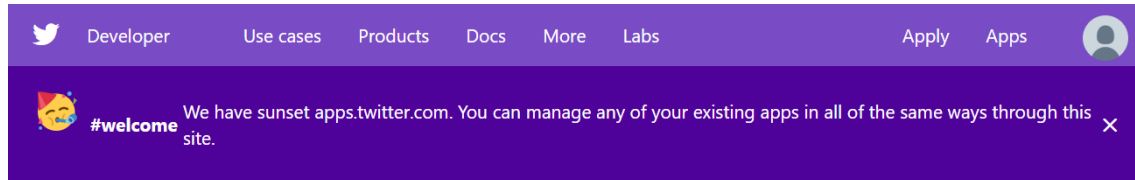


Con estas credenciales podremos acceder a los distintos datasets y tablas creadas en Google Bigquery y seleccionar aquella de la queramos extraer los datos en Qlik Sense.



Redes Sociales – Twitter API

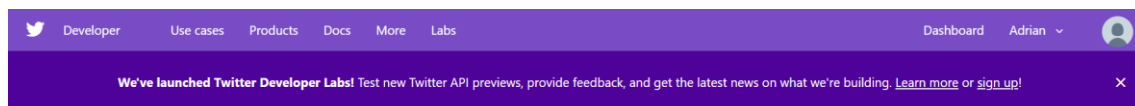
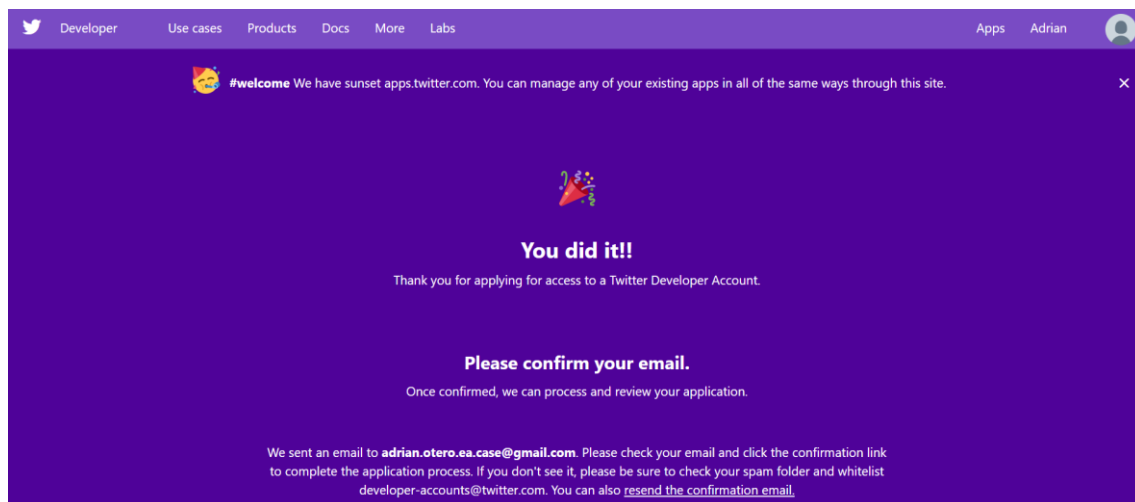
Para acceder a la información de redes sociales vamos a utilizar Twitter. Para ello es necesario registrarse en Twitter y acceder a la plataforma de desarrolladores (<https://apps.twitter.com/>), para generar una nueva APP y conseguir las credenciales que nos permitirán acceder a la API desde nuestro sistema.



No apps here.

You'll need an app and API key in order to authenticate and integrate with most Twitter developer products. Create an app to get your API key.

Presionar en “**Create an app**” y registrarse como desarrollador, rellenando los campos indicados en los diferentes formularios de la plataforma.



Apps > [Create an app](#)

Understanding apps

What is an app? ▾

Why register an app? ▾

Which products require an API key? ▾

App details

The following app details will be visible to app users and are required to generate the API keys needed to authenticate Twitter developer products.

App name (required) ⓘ

Maximum characters: 32

Application description (required)

Share a description of your app. This description will be visible to users so this is a good place to tell them what your app does.

Apps > ea-test-case

App details

Keys and tokens

Permissions

App details

Details and URLs

Edit



App icon

App icon is default, click edit to upload.

App Name

ea-test-case

Description

Extract information from Twitter to do Sentiment Analysis based on tweet locations and tweet languages

Por último, necesitaremos identificar nuestras claves secretas, para incluirlas en nuestra APP y poder acceder a los datos de Twitter de forma programática.

Apps > ea-test-case

App details

Keys and tokens

Permissions

Important notice about access tokens and secrets

To make your API integration more secure, we will no longer show your access token and access token secret beyond the first time that you generate it starting **January 20, 2020**. You will be able to regenerate it at anytime here, which will invalidate your current access token and secret. Please save this information if you need to access it. This does not affect your consumer API keys, which will still be shown here as they are below.

Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

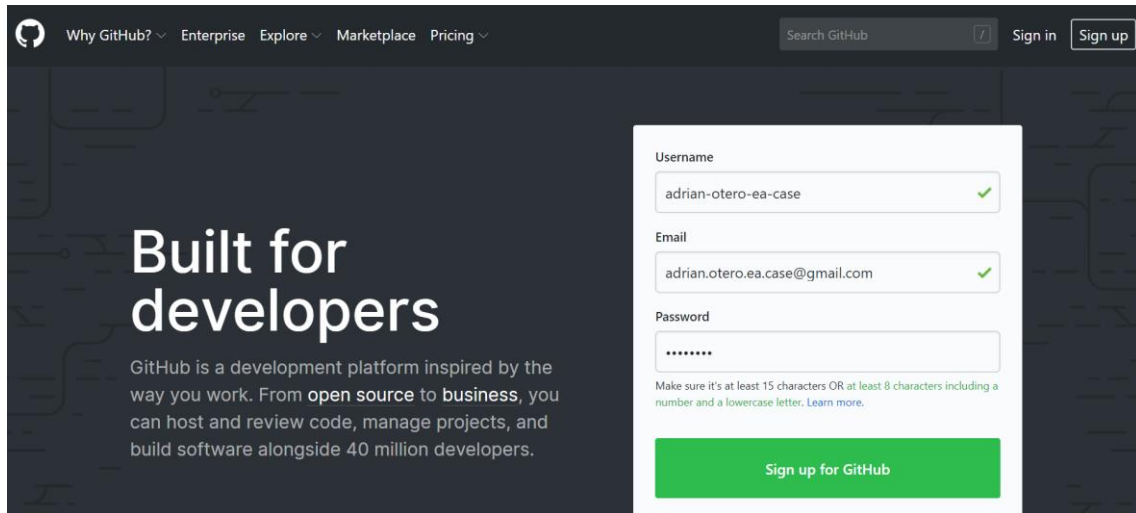
JHi9h2byuCWorokMLClbh20KI (API key)

[REDACTED] (API secret key)

Regenerate

Repositorio GIT

Por último, todos los códigos desarrollados serán almacenados en un repositorio GIT. Para configurar este repositorio es necesario acceder a Github a través del siguiente enlace: <https://github.com/> y crear una nueva cuenta.



Why GitHub? ▾ Enterprise ▾ Explore ▾ Marketplace ▾ Pricing ▾

Search GitHub [7] Sign in Sign up

Built for developers

GitHub is a development platform inspired by the way you work. From **open source** to **business**, you can host and review code, manage projects, and build software alongside 40 million developers.

Username

adrian-otero-ea-case ✓

Email

adrian.otero.ea.case@gmail.com ✓

Password

Make sure it's at least 15 characters OR at least 8 characters including a number and a lowercase letter. [Learn more.](#)

Sign up for GitHub

Una vez lo hayamos configurado, podemos acceder al repositorio a través del siguiente enlace:

https://github.com/adrian-otero-ea-case/ea_test