

Instalación y uso  
de los lenguajes  
**Julia**, Python y R  
y entornos  
RStudio JupyterLab  
*en la nube*  
Google Cloud

## Introducción

El presente documento es un instructivo para instalar la última versión de Julia, Python, R, RStudio Server y Jupyter Lab sobre [Ubuntu 22.04 LTS minimal](#) en el entorno Google Cloud corriendo sobre una máquina virtual *spot* que representa el costo más bajo de todas las opciones disponibles en la nube a septiembre-2023 para el tipo de procesamiento de generación de modelos predictivos.

Las principales "nubes" son :

Amazon AWS EC2	<a href="https://aws.amazon.com/ec2/instance-types/">https://aws.amazon.com/ec2/instance-types/</a>
Microsoft Azure	<a href="https://azure.microsoft.com/en-us/services/virtual-machines/">https://azure.microsoft.com/en-us/services/virtual-machines/</a>
Google Cloud	<a href="https://cloud.google.com/compute/docs/instances/">https://cloud.google.com/compute/docs/instances/</a>
Alibaba Cloud	<a href="https://www.alibabacloud.com/product/computing">https://www.alibabacloud.com/product/computing</a>
IBM Cloud	<a href="https://www.ibm.com/cloud/virtual-servers">https://www.ibm.com/cloud/virtual-servers</a>
Oracle Cloud	<a href="https://www.oracle.com/cloud/compute/">https://www.oracle.com/cloud/compute/</a>
Digital Ocean	<a href="https://www.digitalocean.com/products/droplets/">https://www.digitalocean.com/products/droplets/</a>

Google Cloud provee máquinas virtuales comunes, y a un precio que en promedio se reduce entre 4 y 5 veces las máquinas virtuales llamadas *spot* <https://cloud.google.com/compute/pricing> .

Una máquina virtual *spot* <https://cloud.google.com/compute/docs/instances/spot> , puede ser apagada por Google en cualquier momento. Esas máquinas *spot* pueden crearse y viven en la medida que Google tenga poder de cómputo ocioso en el datacenter donde está corriendo esa maquina virtual, cosa que suele suceder a la madrugada de ese datacenter, o los fines de semana.

Tenga en cuenta que lo más probable es que Google le mate su máquina *spot* cada 8/10 horas. Esto le demandará estar muy atento, verificando periódicamente cuales máquinas virtuales le han apagado. No subestime lo dicho en este párrafo.

Se puede ver las locaciones de los datacenters en <https://cloud.google.com/compute/docs/regions-zones/regions-zones> y una aplicación que muestra en tiempo real el día y la noche es <https://www.timeanddate.com/worldclock/sunearth.html>

Vale la pena leer esta documentacion <https://cloud.google.com/compute/docs/concepts>

Los alumnos más técnicos podrian leer sobre esta alternativa <https://blog.codinghorror.com/the-cloud-is-just-someone-elses-computer/>

**Se recomienda enfáticamente** que configure el idioma inglés en Google Cloud durante toda la instalación, ya que la versión en español de Google Cloud es, al menos ... ambigua.

Esta es la razón por la cual se elige usar máquinas virtuales con sistema operativo Linux de tipo *spot*:

Google Cloud, costo por hora en <u>dólares</u> ( al 2023-09-11 )				
maquina virtual N1 de 16 vCPU y 128 GB de RAM, 256GB HDD us-west4 (Las Vegas)				
	Linux			Windows
	Ubuntu	SUSE	Red Hat	
Spot ( <i>mortal</i> )	<b>0.15</b>	0.26	0.28	0.89
Normal ( <i>inmortal</i> )	0.76	0.87	0.89	1.50

Crear las spot virtual machines en Las Vegas hace una gran diferencia de costo !

#### Conceptualmente

- Todos los archivos que tomemos como input o generemos como output estarán siempre en el Google Storage Bucket , piénselo como un disco de red.
- En la imagen será donde están guardados el lenguaje R, todos sus paquetes, RStudio, el lenguaje Python, el lenguaje Julia, Jupyter Lab, etc
- Para cada proceso que corramos, crearemos una nueva máquina virtual, a partir de la imagen, procesaremos y eliminaremos la máquina virtual.
- Estaremos corriendo máquinas virtuales spot, que Google las apagará y en algunos casos desaparecerán automáticamente.

Al comienzo y por única vez vamos a crear una imagen del sistema operativo Ubuntu, con R y RStudio instalados, y todos los paquetes necesarios.

Esa imagen es la que van a utilizar siempre TODAS las máquinas virtuales.

Una máquina virtual esta definida por lo siguiente

1. La imagen del disco previamente creada ( SO, programas instalados, configuraciones)
2. La cantidad de vCPU
3. La cantidad de memoria RAM
4. El tamaño del disco local ( que debe ser igual o superior al tamaño de la imagen )
5. La disponibilidad de la máquina virtual ( Spot o normal )
6. La ubicación geográfica

Si se apaga y vuelve a encender una máquina virtual, sigue estando disponible lo que quedó en el disco local. Si se borra una máquina virtual, se pierde toda la información del disco local.

Por la arquitectura de la solución, el disco local de las máquinas virtuales solamente se utiliza para almacenar información temporal que selectivamente es copiada al bucket de datos.

El bucket de datos, sería algo como un “disco de red” al que todas las máquinas virtuales pueden acceder, leer y escribir al mismo tiempo. También podemos manualmente subir información desde nuestras PCs por medio del Google Cloud Console.

Utilizamos el bucket de datos para:

- Almacenar los datasets, que son leídos por los programas R que corren en las máquinas virtuales. ~/buckets/b1/datasets
- Almacenar los resultados intermedios y finales de cada una de las corridas, ~/buckets/b1/exp

Es decir, en nuestra arquitectura la entrada y la salida siempre queda en el bucket.

Se podría decir que la copia central del bucket está ubicada en USA.

La información del bucket de datos es persistente y totalmente independiente de las máquinas virtuales. Por más que se apague o incluso borre una máquina virtual, la información del bucket de datos permanece disponible, y es accesible por medio de Google Cloud Console.

No hace falta que una máquina virtual esté encendida para acceder al contenido del bucket.

Va a ser muy común tener varias maquinas virtuales encendidas al mismo tiempo, cada una corriendo un *experimento* distinto en distintos datacenters del planeta, y todas ellas leyendo y escribiendo simultáneamente del mismo bucket de datos.

Los pasos de este instructivo son :

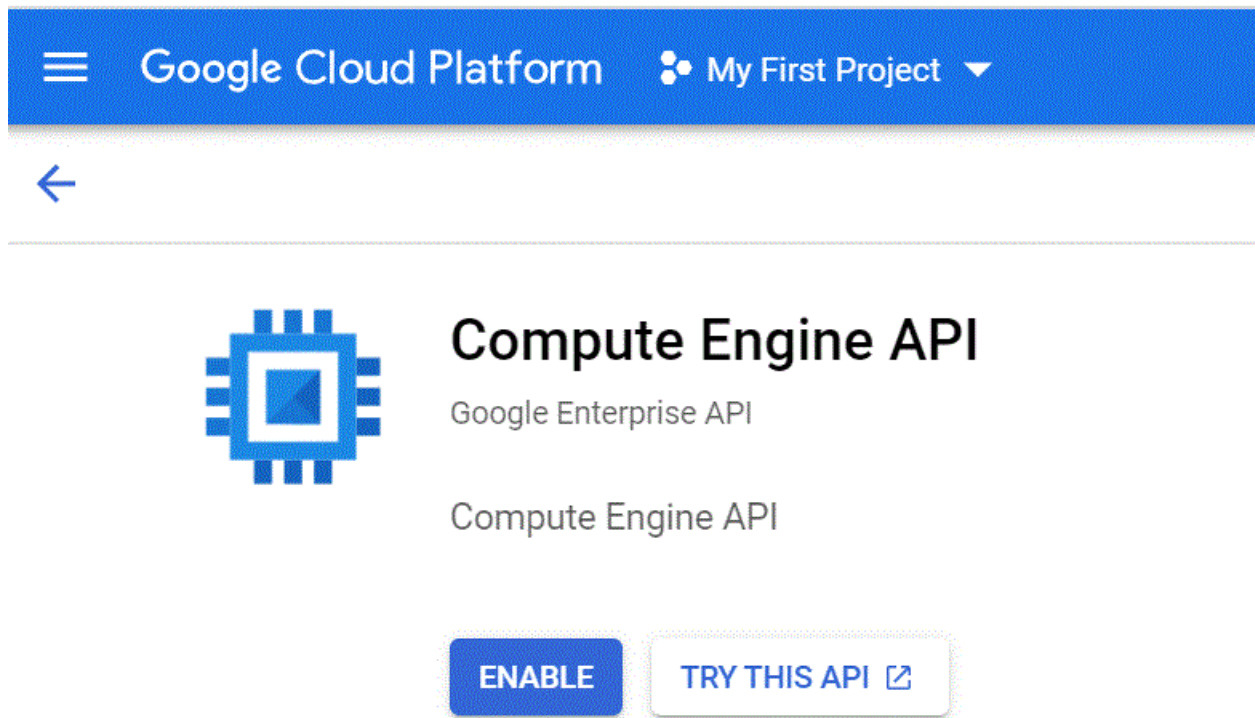
1. Instalar lenguajes y entornos (se hace por única vez)
  1. Habilitar Compute Engine API (1 min)
  2. Elegir idioma inglés (1 min)
  3. Crear máquina virtual inmortal (10 min)
  4. Ingresar a la terminal de la máquina virtual (1 min)
  5. Verificar versión de Ubuntu (1 min)
  6. Crear el bucket por línea de comando (3 min)
  7. Instalar ( 60 minutos, corre desatendido)
  
2. Crear Imagen, templates y desktop (se hace por única vez)
  1. Archivo `kaggle.json` (3 min)
  2. Cambiar la clave (1 min)
  3. Verificar datasets (1 min)
  4. Verificar R (1 min)
  5. Verificar Rstudio (1 min)
  6. Verificar Jupyter Lab (1 min)
  7. Clonar el repositorio personal (4 min)
  8. Crear la *imagen, templates y desktop* (10 min)
  9. Configuración inicial ambiente gráfico GNOME (15 min)
  
3. Metodología de trabajo en Google Cloud
  1. Entendiendo qué genera costos en GC
  2. Aumentando el poder de cómputo asumiendo riesgos
  3. Disciplina de trabajo
  4. Crear máquina virtual desde template
  5. RStudio en forma remota
  6. Usar Jupyter Lab en forma remota
  7. Desde una terminal Ubuntu con la consola de R correr scripts
  
4. Cuando finalice la materia
  1. Evitar que Google Cloud facture

# Instalación de lenguajes y entornos

## 1.1 Habilitar Compute Engine API

En a <https://console.cloud.google.com/compute/instances>

Y si le aparece



presione el botón de **Enable**

## 1.2 Elegir idioma inglés

Ir a <https://console.cloud.google.com/user-preferences/languages> y elegir el idioma English( United States )

## 1.3 Crear Maquina Virtual

Por favor, prestar atención a todo lo que esté encerrado en una **elipse roja**, **subrayado por una línea roja**, o **apuntado por una flecha roja** en las impresiones de pantalla, es VITAL cumplir a rajatabla esa instrucción.

Una vez que ya se dio de alta en Google Cloud, ir a la página <https://console.cloud.google.com/compute/instancesAdd>

Name \*  
instance-instalacion ?

✓ MANAGE TAGS AND LABELS

Region \*  
us-west4 (Las Vegas) ▼ ?  
Region is permanent

Zone \*  
us-west4-a ▼ ?  
Zone is permanent

### Machine configuration

✓ General purpose   Compute optimized   Memory optimized   GPUs


Machine types for common workloads, optimized for cost and flexibility

	Series ?	Description	vCPUs ?	Memory ?	Platform
<input type="radio"/>	C3	Consistently high performance	4 - 176	8 - 1,408 GB	Intel Sapphire Rapids
<input type="radio"/>	E2	Low cost, day-to-day computing	0.25 - 32	1 - 128 GB	Based on availability
<input type="radio"/>	N2	Balanced price & performance	2 - 128	2 - 864 GB	Intel Cascade and Ice Lake
<input type="radio"/>	N2D	Balanced price & performance	2 - 224	2 - 896 GB	AMD EPYC
<input type="radio"/>	T2A	Scale-out workloads	1 - 48	4 - 192 GB	Ampere Altra Arm
<input checked="" type="radio"/>	<u>T2D</u>	Scale-out workloads	1 - 60	4 - 240 GB	<u>AMD EPYC Milan</u>
<input type="radio"/>	N1	Balanced price & performance	0.25 - 96	0.6 - 624 GB	Intel Skylake

### Machine type

Choose a machine type with preset amounts of vCPUs and memory that suit most workloads.

t2d-standard-4 (4 vCPU, 16 GB memory) ▼

 **vCPU**  
4

**Memory**  
16 GB



## ▼ ADVANCED CONFIGURATIONS

### Availability policies

VM provisioning model

Standard

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#)

☐ Set a time limit for the VM ?

On VM termination

Choose what happens to your VM when it's preempted or reaches its time limit

On host maintenance

Terminate VM instance

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

Off

Compute Engine can automatically restart VM instances if they are terminated for non-user-initiated reasons (maintenance event, hardware failure, software failure and so on)

## ^ VM PROVISIONING MODEL ADVANCED SETTINGS

### Display device

Enable to use screen capturing and recording tools.

☐ Enable display device

### Confidential VM service ?

— Confidential Computing is disabled on this VM instance

ENABLE

### Container ?

Deploy a container image to this VM instance

DEPLOY CONTAINER

En la elección del Boot Disk deberá seleccionar estas opciones.

Por favor, siga estas instrucciones al pie de la letra, cualquier otra alternativa a [Ubuntu 22.04 LTS Minimal](#) no funcionará, un tamaño menor a 32 GB de disco, fallará y dará crípticos mensajes de error.


## Boot disk

Select an image or snapshot to create a boot disk; or attach an existing disk. Can't find what you're looking for? Explore hundreds of VM solutions in [Marketplace](#)

PUBLIC IMAGES	CUSTOM IMAGES	SNAPSHOTS	ARCHIVE SNAPSH
<div>Operating system</div> <div>Ubuntu</div>			
<div>Version *</div> <div>Ubuntu 22.04 LTS Minimal</div> <div>x86/64, amd64 jammy minimal image built on 2023-07-26</div>			
<div>Boot disk type *</div> <div>Standard persistent disk</div>			
<div>COMPARE DISK TYPES</div>			
<div>Size (GB) *</div> <div>32</div>			

Por favor, es MUY IMPORTANTE que el disco tenga **32GB** de espacio, en caso contrario no se podrá instalar todo e inexplicables errores lo devorarán.

## Boot disk ?

Name	<a href="#">instance-instalacion</a>
Type	<a href="#">New standard persistent disk</a>
Size	<a href="#">32 GB</a>
License type ?	<a href="#">Free</a>
Image	 <a href="#">Ubuntu 22.04 LTS Minimal</a>

[CHANGE](#)

## Identity and API access ?

### Service accounts ?

#### Service account

Compute Engine default service account

Requires the Service Account User role (roles/iam.serviceAccountUser) to be set for users who account. [Learn more](#)

### Access scopes ?

- ☐ Allow default access
- ☒ Allow full access to all Cloud APIs
- ☐ Set access for each API



## Firewall ?

Add tags and firewall rules to allow specific network traffic from the Internet

- ☒ Allow HTTP traffic
- ☒ Allow HTTPS traffic



## Observability - Ops Agent ?

Monitor your system through collection of logs and key metrics.

- ☐ Install Ops Agent for Monitoring and Logging

## Advanced options

Networking, disks, security, management, sole-tenancy



## 1.4 Ingresar a la terminal de la máquina virtual recién creada

<input type="checkbox"/> Name ^	Zone	Recommendation	Internal IP	External IP	Connect
<input type="checkbox"/> instance-1	us-west1-c		10.138.0.2	None	SSH ▾ ⋮
<input type="checkbox"/> instance-2	europa-west1-d		10.132.0.2	None	SSH ▾ ⋮
<input type="checkbox"/> instance-3	us-central1-c		10.128.0.3	None	SSH ▾ ⋮
<input checked="" type="checkbox"/> instance-instalacion	us-central1-c		10.128.0.4	35.192.64.53	SSH ▾ ⋮
<input type="checkbox"/> instance-itba	us-central1-c		10.128.0.2	None	SSH ⋮

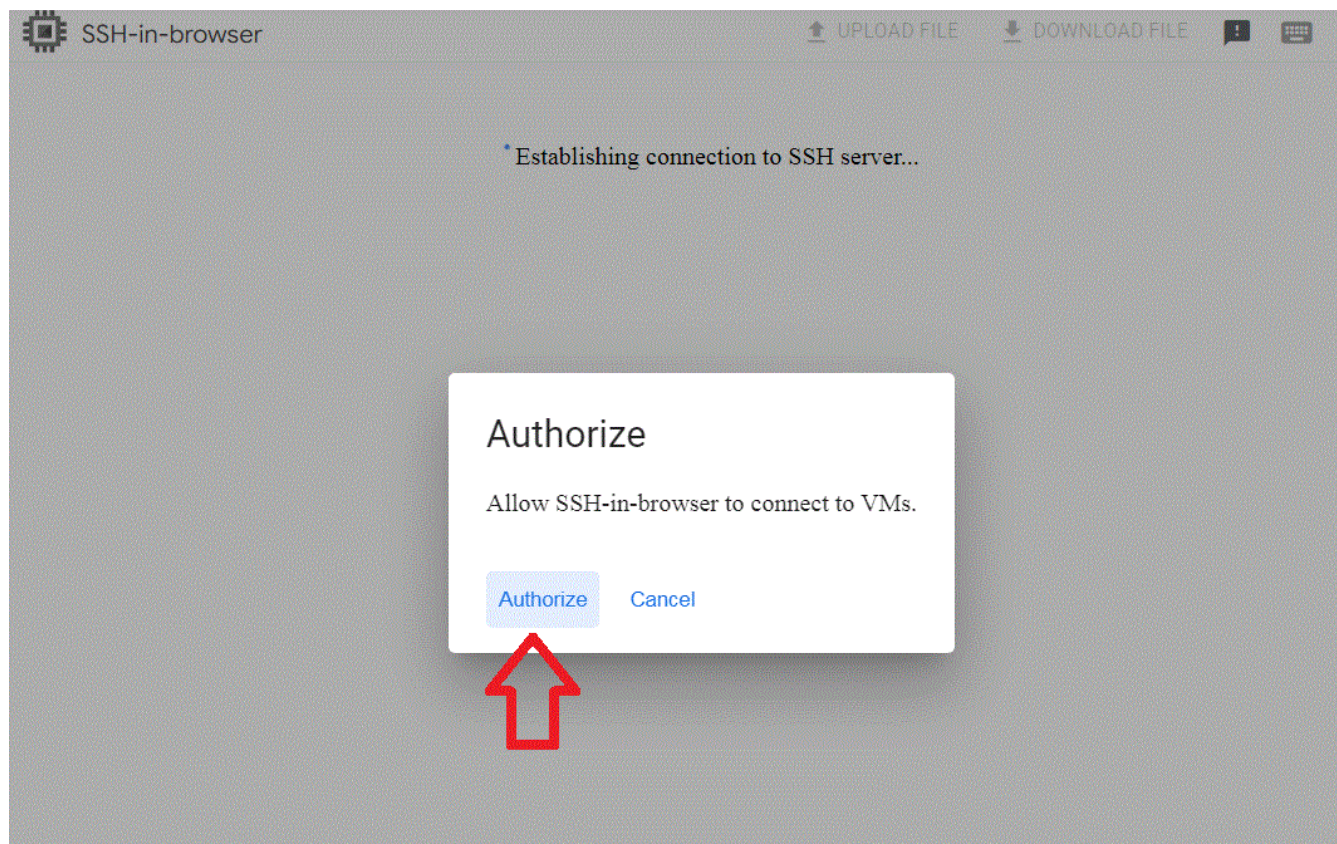
Open in browser window

Open in browser window on custom port

View gcloud command

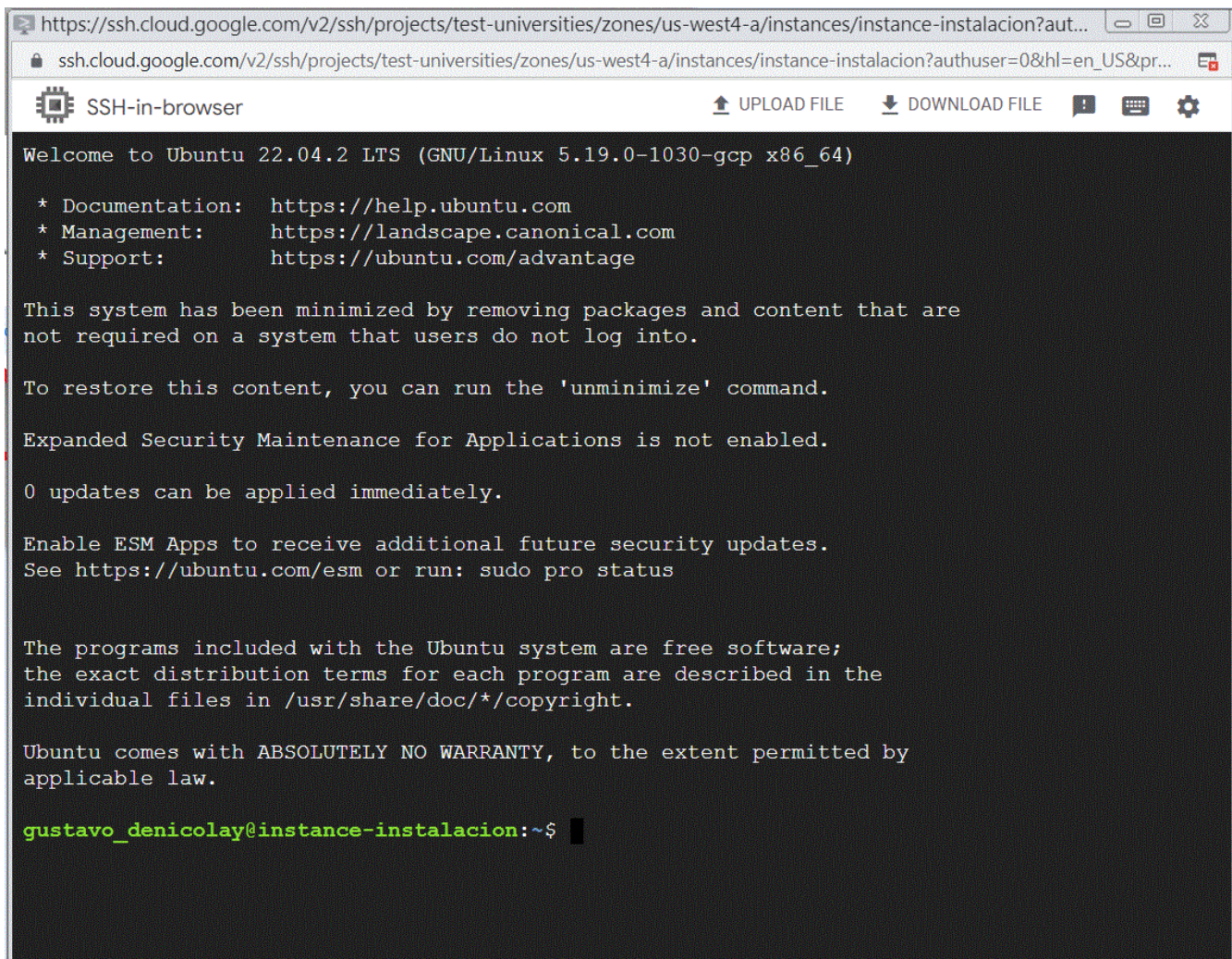
Use another SSH client

primero aparecerá una ventana como esta donde deberá presionar el botón de





luego aparecerá una ventana como esta (con **su** nombre de usuario)



```
https://ssh.cloud.google.com/v2/ssh/projects/test-universities/zones/us-west4-a/instances/instance-instalacion?aut...
ssh.cloud.google.com/v2/ssh/projects/test-universities/zones/us-west4-a/instances/instance-instalacion?authuser=0&hl=en_US&pr...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

Welcome to Ubuntu 22.04.2 LTS (GNU/Linux 5.19.0-1030-gcp x86_64)

* Documentation:  https://help.ubuntu.com
* Management:    https://landscape.canonical.com
* Support:       https://ubuntu.com/advantage

This system has been minimized by removing packages and content that are
not required on a system that users do not log into.

To restore this content, you can run the 'unminimize' command.

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

gustavo_denicolay@instance-instalacion:~$
```

para futuras referencias, su usuario Ubuntu que aparece antes del arroba @, en el caso de la imagen anterior el nombre de usuario es : **gustavo\_denicolay**

Atención, es posible por problemas de internet al momento de ingresar a la máquina virtual presionando el botón de SSH, o levantar el archivo, le salga el siguiente mensaje de error:

```
Code: 1006
Please ensure you can make a proper https connection to the IAP for TCP
hostname https://tunnel.cloudproxy.app
You may be able to connect without using the Cloud Identity-Aware Proxy
```

En ese caso, espere unos minutos y vuelva a intentarlo.

## 1.5 Verificar versión de Ubuntu

En la imagen anterior **debe** aparecer

```
This system has been minimized by removing packages and content that are
```

luego, desde la terminal Ubuntu escribir lo que está en color azul (dar Enter al final )

```
$ lsb_release -a
```

le deberá aparecer

```
No LSB modules are available.
Distributor ID: Ubuntu
Description:   Ubuntu 22.04.3 LTS
Release:      22.04
Codename:     jammy
```

cualquier cosa distinta que le aparezca indica que usted ha cometido un error en el paso 1.3 porque eligió otra versión de Ubuntu (un error usual); debe eliminar la virtual machine `instance-1` y crearla nuevamente

## 1.6 Crear bucket por línea de comandos

desde la terminal Ubuntu (dar enter al final )

```
$ gsutil mb gs://buko666
```

Donde **buko666** debe ser un nombre único, no repetido en todo Google Cloud

Si le aparece que ya existe ese nombre de bucket, volver a intentar el mismo comando con otro nombre de bucket

Si le sale el siguiente mensaje, es que debe elegir otro nombre y volver a correr el comando

```
ServiceException: 409 A Cloud Storage bucket named 'buko666' already exists. Try another name. Bucket names must be globally unique across all Google Cloud projects, including those outside of your organization
```

Si por alguna razón usted debe volver a reinstalar, y ya creó el bucket, NO vuelva a crear otro !

## 1.7 Instalar

desde la terminal Ubuntu ejecutar estos tres comandos

( el \$ inicial en color negro representa el prompt de Ubuntu, y **no** debe ser tipeado )

```
$ wget https://storage.googleapis.com/open-courses/itba2023b-52fa/instalar.sh -O instalar.sh
$ chmod u+x *.sh
$ ./instalar.sh
```

Este paso es completamente automático, puede llevar alrededor de **65 minutos**, se verán pasar decenas de miles de líneas por la pantalla; algunos alumnos manifestaron sentirse dentro de una Matrix pero es nada más que una humilde, aburrida y típica instalación Linux, la mayor parte del tiempo se compilan los paquetes de R.

Atención, ante cualquiera de estos desafortunados eventos

- se le corta internet durante la instalación
- se le apaga su PC
- su PC entra en modo ahorro de energía

se le cerrará la consola de Ubuntu y usted deberá volver a ingresar a la máquina virtual, botón SSH y una vez dentro de la terminal Ubuntu debe ejecutar el siguiente comando

```
$ tmux attach
```

e ingresará nuevamente a la instalación, la que habrá seguido corriendo por más que usted perdió conexión. NO ABANDONE JAMAS LA INSTALACION ! APAGUE ESE MAQUINA VIRTUAL SI VA A CONTINUAR AL DIA SIGUIENTE Y COMIENCE DE CERO.

Crear  
Imagen  
(del sistema operativo)  
templates  
y  
desktop

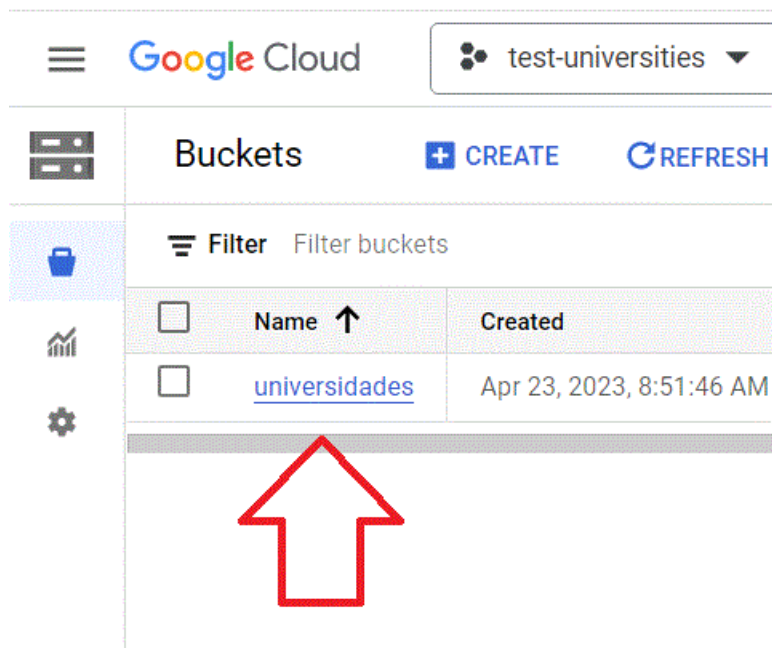


## 2.1 Archivo kaggle.json

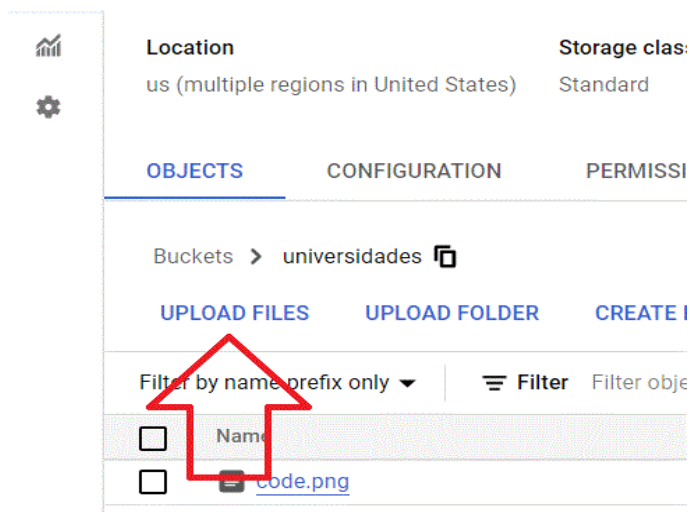
En el Libro de la Asignatura, usted se bajó el archivo kaggle.json a un lugar seguro en su laptop. Ahora debe subirlo al bucket.

Estando en la consola de Google Cloud vaya a <https://console.cloud.google.com/storage/browser>

Una vez allí dentro, haga click en el nombre de su bucket (le va a aparecer el nombre de su bucket, no la palabra universidades )



Dentro, vaya a **UPLOAD FILES** , busque el archivo en su laptop, y súbalo al bucket.



## 2.2 Cambiar la clave

desde la terminal

( el \$ inicial en color negro representa el prompt de Ubuntu, y **no** debe ser tipeado )

Cambiar la clave

```
$ ~/cambiar_claves.sh
```

aparecerá en la pantalla “New password:” , se debe tipear una password la que **no** es mostrada en pantalla mientras se va escribiendo (parece como que está todo “colgado”). Se debe presionar la tecla Enter al final. Luego solicita nuevamente la password.

Dado que es posible que en algún momento quiera compartir la máquina virtual con un compañero de equipo, se sugiere utilizar una password distinta a sus preferidas que utiliza en su HomeBanking. Se piden DOS passwords, una se utiliza para RStudio y la otra para Jupyter Lab, se sugiere que sean las mismas.

Con este usuario y password se ingresará al RStudio, por favor recordar ambos .  
Si comete un error, vuelva a correr el comando.

## 2.3 Verificar Datasets

corriendo desde la consola Ubuntu

```
$ ls -l ~/buckets/b1/datasets
```

se deberá observar los siguientes archivos

- dataset\_pequeno.csv
- competencia\_2023.csv.gz

Ir a la página <https://console.cloud.google.com/storage/browser> , hacer click en su bucket, luego en la carpeta datasets, y verificar que se ve el archivo.

## 2.4 Verificar R

Ejecutar el siguiente comando desde la terminal Ubuntu (la letra R debe estar en mayúsculas )

\$ **R**

Le deberá aparecer la consola de R , corroborar el `R version 4.3.1`

```
R version 4.3.1 (2023-06-16) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.

> █
```

para salir de esta pantalla, debe tipear `quit( save="no")` y luego la tecla Enter

## 2.5 Verificar RStudio

Ejecutar el siguiente comando desde la terminal Ubuntu

```
$ rstudio-server status
```

Le deberá aparecer algo de este estilo, donde es fundamental la parte en color verde **active (running)**

```
● rstudio-server.service - RStudio Server
   Loaded: loaded (/lib/systemd/system/rstudio-server.service; enabled; preset: enabled)
   Active: active (running) since Thu 2023-05-04 14:39:53 UTC; 1h 39min ago
     Main PID: 14734 (rserver)
        Tasks: 4 (limit: 19187)
       Memory: 2.4M
          CPU: 8.198s
       CGroup: /system.slice/rstudio-server.service
```

No de importancia a este tipo de mensajes, son simplemente hackers que intentan entrar a su máquina *y no lo lograron*

```
instance-instalacion rserver[42274]: ERROR Invalid method CONNECT requested for uri: google
```

Sin embargo, SI es MUY IMPORTANTE lo siguiente:

**Si no le aparece **active(running)** , deberá hacer de nuevo la instalación, desde cero, GAME OVER.**

Para salir presione la tecla "q"



## 2.6 Verificar Jupyter Lab

Ejecutar el siguiente comando desde la terminal Ubuntu

```
$ systemctl status jupyterlab
```

Le deberá aparecer algo de este estilo, donde es fundamental la parte en color verde **active (running)**

```
● jupyterlab.service - jupyterlab
   Loaded: loaded (/etc/systemd/system/jupyterlab.service; enabled; preset: enabled)
   Active: active (running) since Thu 2023-05-04 13:59:34 UTC; 2min 54s ago
     Main PID: 93538 (jupyter-lab)
        Tasks: 1 (limit: 19187)
      Memory: 62.1M
         CPU: 1.006s
        CGroup: /system.slice/jupyterlab.service
```

Si no le aparece **active(running)** , deberá hacer de nuevo la instalación, desde cero, GAME OVER.

Para salir presione la tecla “q”

## 2.7 Clonar el repositorio personal

Esta es la parte más delicada de toda la instalación, proceda con gran cautela.

Usted debe clonar SU propio repositorio de GitHub, donde usted trabajará y hará modificaciones (no confundir con el repositorio oficial de la materia donde solo los profesores realizan cambios)

```
$ cd
$ git clone https://MariaLopez:token@github.com/MariaLopez/dm2023b.git
$ cd dm2023b
$ git config user.email "usted@gmail.com"
$ git config user.name "MariaLopez"
$ git remote add upstream https://github.com/itba-ecd/dm2023b.git
$ cd
```

es muy importante que luego del `cd dm2023b` verifique que efectivamente se encuentra en la carpeta `dm2023b` (se observa viendo el prompt), esto le indicará que la sentencia que comienza con `git clone https://` ha funcionado bien.

debe reemplazar

- `MariaLopez` por su nombre de usuario en GitHub
- `token` por el token que creó cuando accedió por primera vez a GitHub y que como decía el Libro de la Asignatura guardó en un lugar seguro . Si hubiera olvidado guardarlo deberá crear un nuevo token en GitHub, siguiendo los pasos de El Libro de la Asignatura, el que coexistirá con el primero que creó.
- `usted` por su dirección de gmail , ( va con las comillas como está )

## 2.8 Crear la imagen, templates y desktop

Esta instrucción creará la imagen, los templates, y apagará la máquina  
( el \$ inicial en color negro representa el prompt de Ubuntu, y **no** debe ser tipeado )


```
$ ~/final.sh
```

espere a que corra el comando unos 10 minutos, le apareciera “Presione la tecla Enter para finalizar...” presione Enter, se apagará la máquina virtual y su consola de Ubuntu se pondrá gris, con un mensaje de “Connection Failed”, cierre esa ventana.


Ha finalizado la instalación, ya está en condiciones de procesar en Google Cloud, en nuevas máquinas virtuales.

Espere a que el profesor le indique cuando es prudente eliminar `instance-instalacion`


Finalmente, usted tendrá 2 máquinas virtuales apagadas. Deberá ver algo parecido a esto:

 Google Cloud 


test-universities ▼




VM instances [+ CREATE INSTANCE](#) [IMPORT VM](#) [REFRESH](#)





[INSTANCES](#) [OBSERVABILITY](#) [INSTANCE SCHEDULES](#)






VM instances



 Filter Enter property name or value



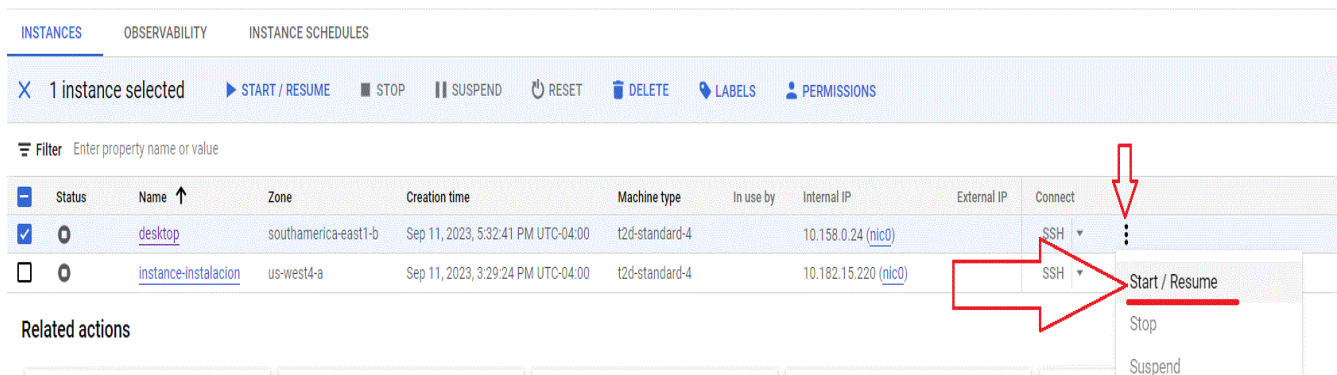
<input type="checkbox"/>	Status	Name ↑	Zone	Creation time
<input type="checkbox"/>		<a href="#">desktop</a>	southamerica-east1-b	Aug 1, 2023, 10:31:33 PM
<input type="checkbox"/>		<a href="#">instance-instalacion</a>	us-west4-a	Aug 1, 2023, 8:54:35 PM



Related actions

## 2.9 Configuración inicial ambiente gráfico GNOME

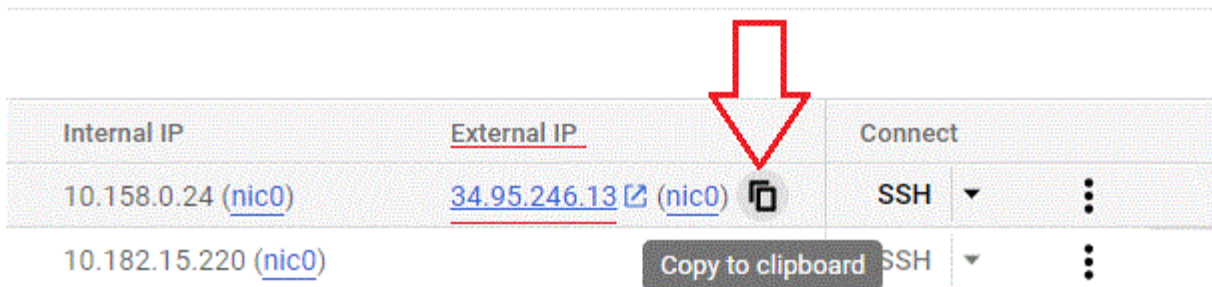
Vaya a <https://console.cloud.google.com/compute/instances> y de Start/Resume a la virtual machine desktop



The screenshot shows the Google Cloud Platform console with the 'Instances' tab selected. A table lists two instances: 'desktop' and 'instance-instalacion'. The 'desktop' instance is selected, and its context menu is open, showing options like 'Start / Resume', 'Stop', and 'Suspend'. A red arrow points to the 'Start / Resume' button.

Status	Name	Zone	Creation time	Machine type	In use by	Internal IP	External IP	Connect
On	desktop	southamerica-east1-b	Sep 11, 2023, 5:32:41 PM UTC-04:00	t2d-standard-4		10.158.0.24 (nic0)		SSH
Off	instance-instalacion	us-west4-a	Sep 11, 2023, 3:29:24 PM UTC-04:00	t2d-standard-4		10.182.15.220 (nic0)		SSH

Una vez encendida la virtual machine, copie la External IP al portapapeles (debe posicionar el mouse donde señala la flecha roja)

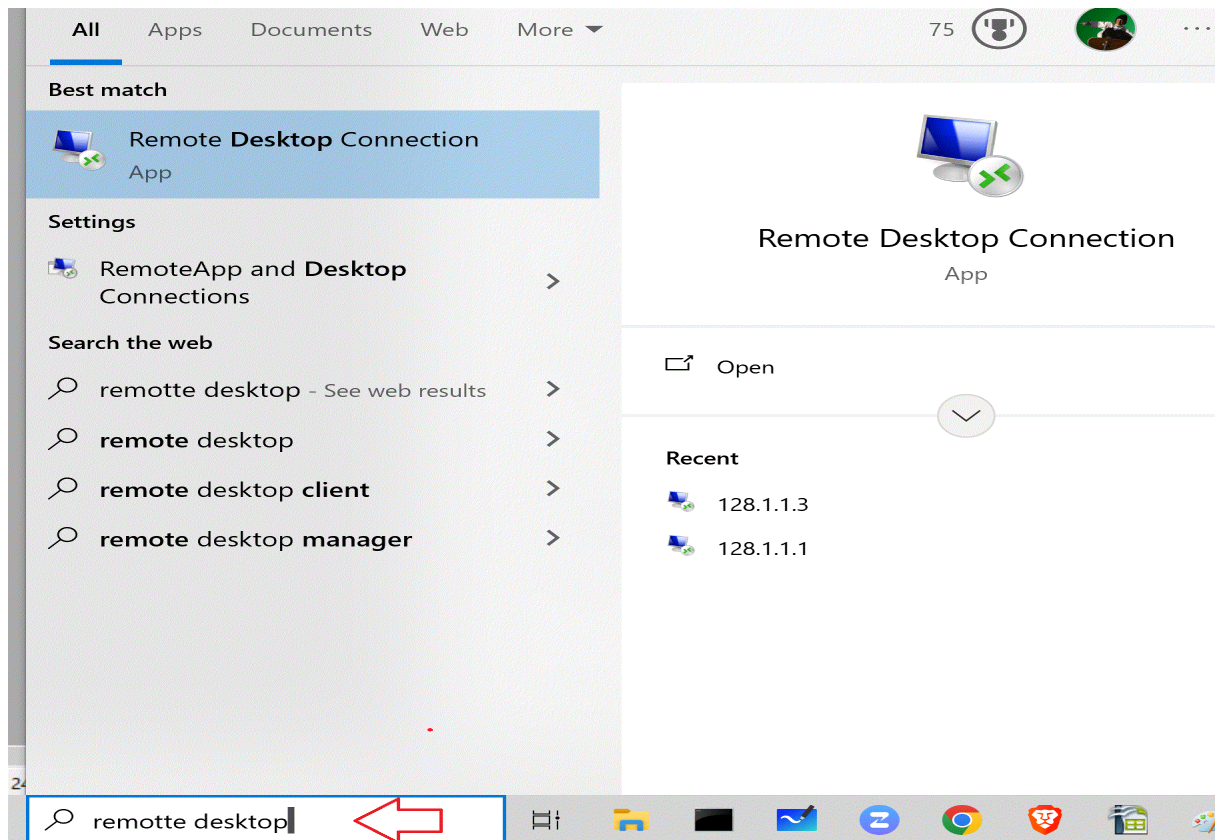


The screenshot shows a close-up of the 'Instances' table. A red arrow points to the 'Copy to clipboard' button next to the External IP of the 'desktop' instance, which is 34.95.246.13.

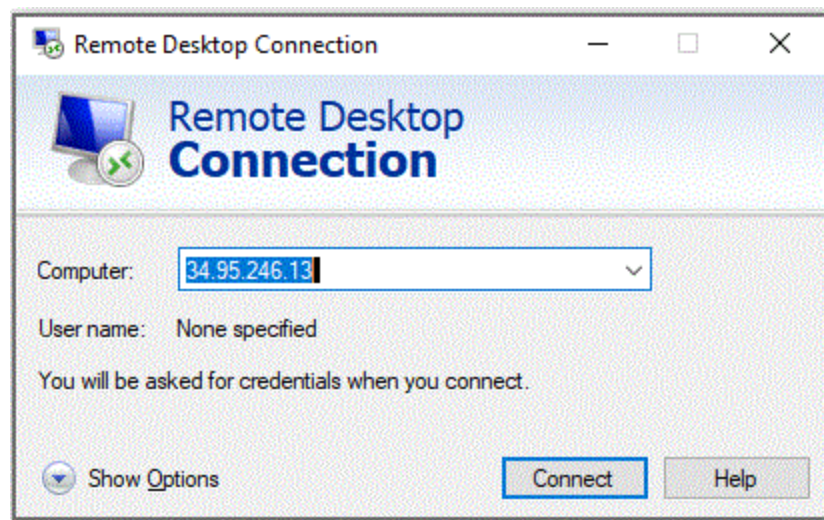
Internal IP	External IP	Connect
10.158.0.24 (nic0)	<a href="#">34.95.246.13</a> (nic0)	SSH
10.182.15.220 (nic0)		SSH



Luego, desde la Search Bar de Windows busque la aplicación Remote Desktop Connection

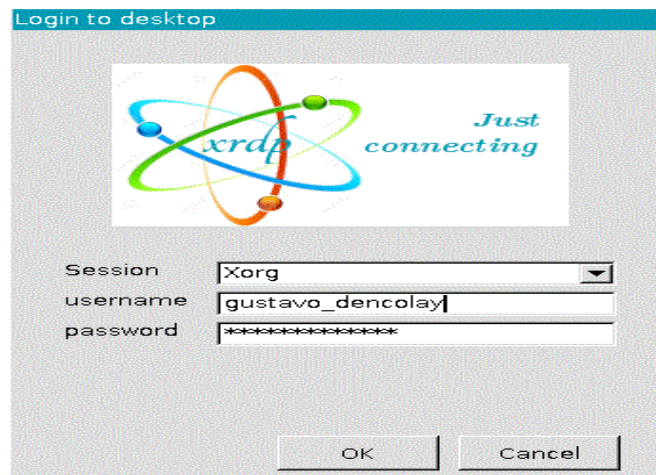


ingrese a la misma, y copie la External IP con Control + V

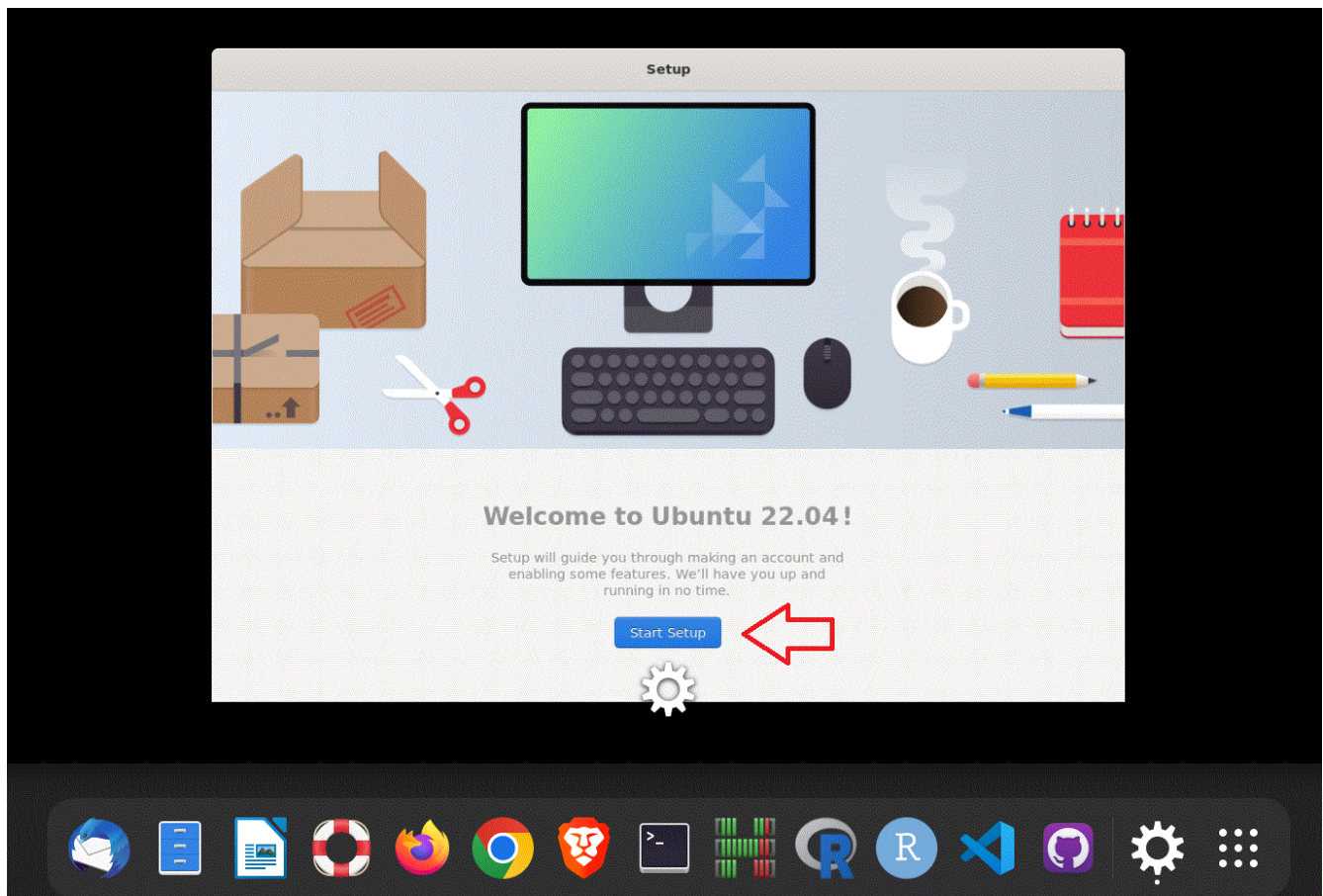


presione el boton connect en la siguiente pantalla pequeña que la aparece, hasta llegar a otra pantalla, donde deberá cargar el usuario y clave que especificó al momento de correr

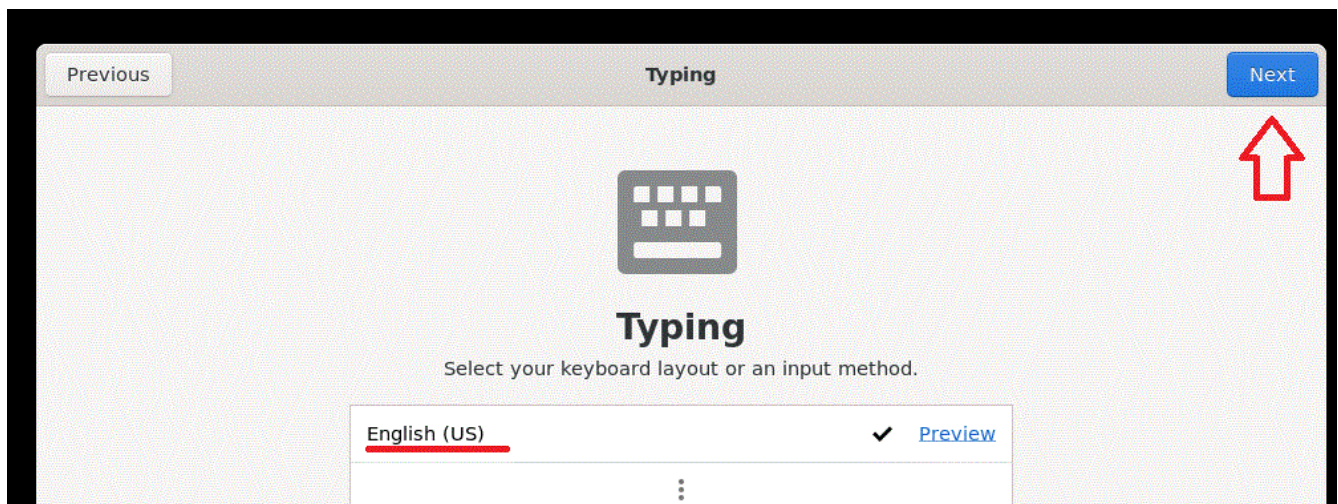
```
./cambiar_claves.sh
```



finalmente, ingresará a una pantalla como la que sigue, donde deberá presionar el boton azul Start Setup y recorrer las dos pantallas que siguen eligiendo su tipo de teclado



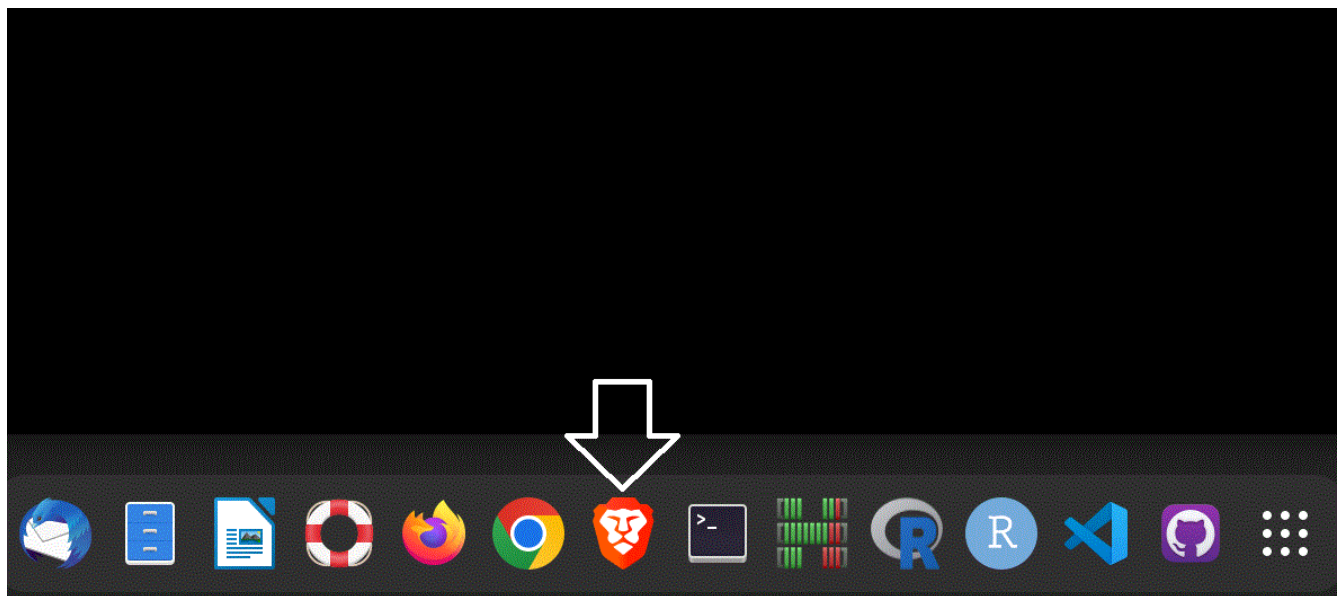




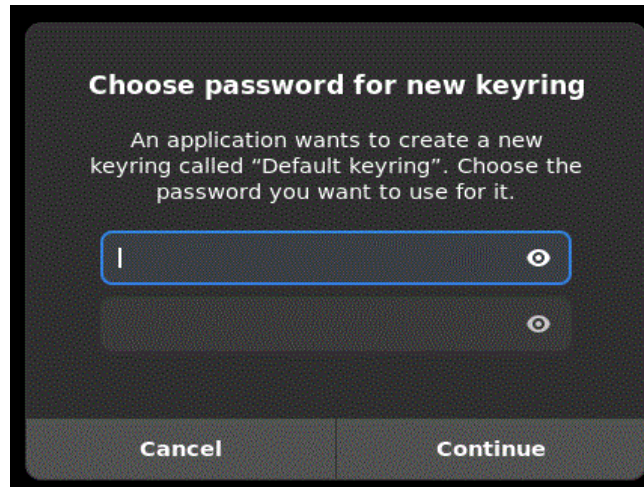
y de Next a la siguiente pantalla que le aparezca sobre Privacy sin cambiar nada (no se muestra la pantalla)

Finalmente presione el boton azul que dice Start Using Ubuntu (no se muestra la pantalla)

Maximice la pantalla del Remote Desktop y presione la tecla Windows de su teclado, se le desplegará en la parte inferior de la pantalla el menu de GNOME llamado dash, presione el botón del browser Brave con el mouse (un solo click)



luego de unos segundos se desplegará una pantalla que le pedirá, por única vez, una clave para el keyring, escriba la misma que eligió para Ubuntu

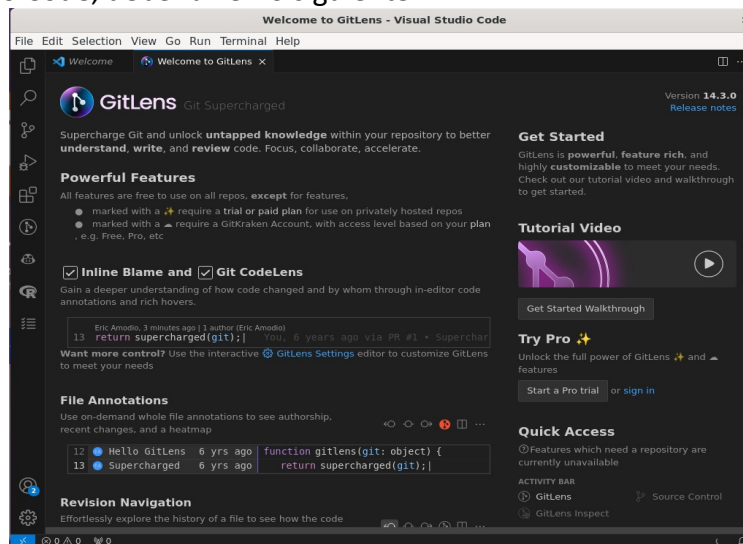


una vez que se abra el browser Brave ingrese, en este orden a

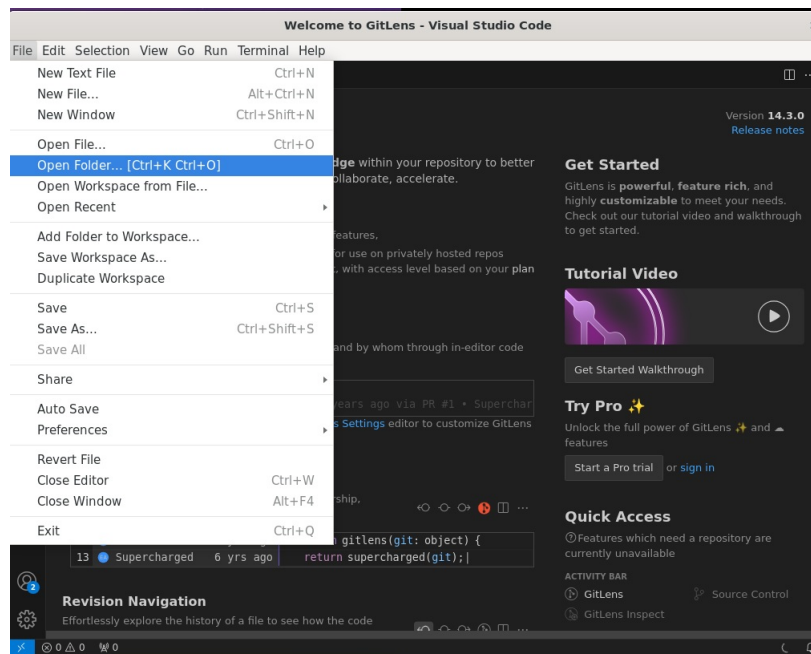
1. Su gmail
2. A su cuenta GitHub
3. A su cuenta Kaggle

tenga presente que seguramente Google le envíe un mensaje a su teléfono móvil, ya que estará sorprendido que estando usted en Buenos Aires quiera conectarse de Sao Paulo (no se muestran las pantallas)

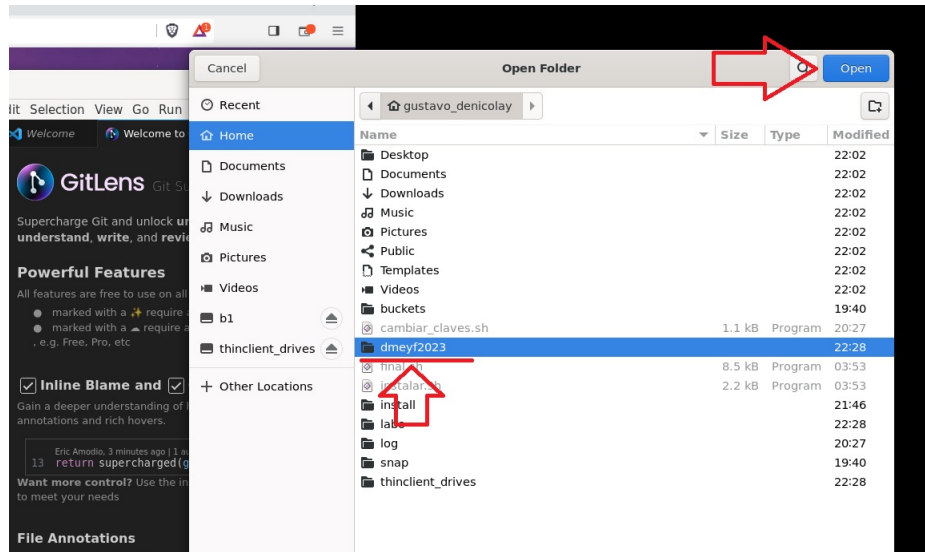
Vuelva a presionar el botón de Windows de su teclado para que se desplique el dash de GNOME, e ingrese al Visual Studio Code, debería ver lo siguiente



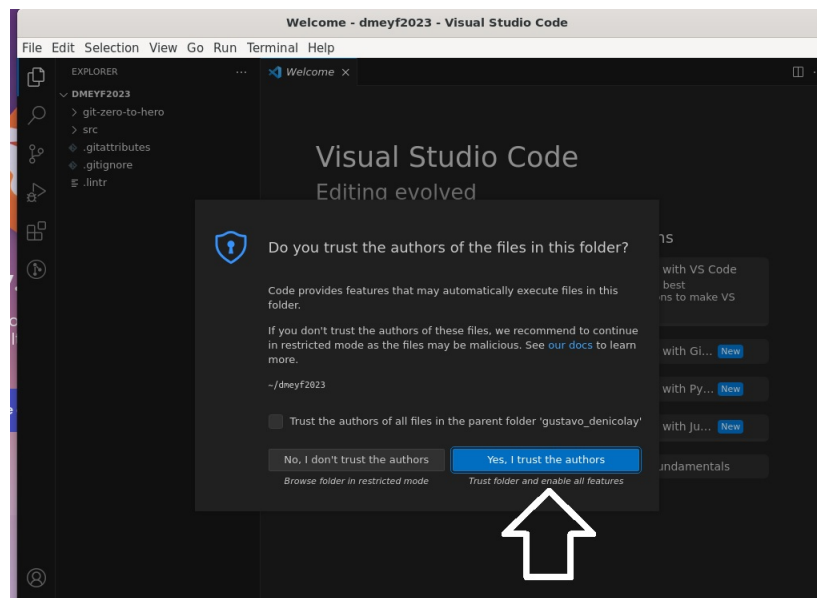
Vaya a File, Open Folder



Elija la carpeta que es el nombre del repositorio , y presione el boton azul Next



En la siguiente pantalla presione el boton  
**Yes, I trust the authors**



Enhorabuena, usted ya está en condiciones de trabajar.

Por ahora, no haga nada más, la virtual machine desktop entrará a los 15 minutos en modo Suspensión

En la clase presencial se le mostrará como trabajar en el ambiente con VSCode.

# Metodología de trabajo en Google Cloud

### 3.1 Entendiendo qué genera costos en GC

#### ¿ Qué genera costo en Google Cloud ?

Máquina Virtual ENCENDIDA, en función de la cantidad de vCPU, memoria RAM, tipo de disco local SSD o normal, tamaño disco local reservado, trafico de red saliente real, accesos a disco reales, costo del Sistema Operativo si este es pago ( Windows, Linux Red Hat ), zona donde es creada .

**Se paga aunque no se esté procesando nada .**

**Se paga aunque el proceso haya terminado.**

Máquina Virtual APAGADA, en función del espacio en disco alocado del sistema operativo y datos locales. Es adicional a la imagen.

Almacenar en discos SSD es muy costoso, aunque no se estén utilizando.

Bucket, espacio en disco realmente utilizado, GB por segundo en los que existieron los archivos.

Transferencias salientes del Bucket a otros lugares, ya sea lecturas desde máquinas virtuales en otras regiones de Google, o hacia PCs que residen fuera de Google.

Imágenes de discos según la cantidad de GB de disco utilizada, por más que ninguna máquina virtual esté utilizando dicha imagen.

[https://cloud.google.com/compute/all-pricing?hl=en\\_US&\\_ga=2.249918022.-1998581785.1564797469](https://cloud.google.com/compute/all-pricing?hl=en_US&_ga=2.249918022.-1998581785.1564797469)

En cualquier momento usted puede verificar su consumo accediendo a

<https://console.cloud.google.com/billing> y luego hacer click en el link que muestra la cuenta de facturación.



## 3.2 Aumentando el poder de cómputo asumiendo riesgos

Google busca constantemente su propio beneficio tentando a sus clientes para luego facturar más de lo que ellos esperan.

Esta es la historia detrás de los USD 300 gratuitos que duran un máximo de 3 meses, y una forma de "ganarle" a Google.

El uso de los USD 300 tiene ciertas restricciones, que son estas

<https://cloud.google.com/free/docs/gcp-free-tier>

Es decir, tiene la limitación que solo se pueden tener corriendo al mismo tiempo 8 vCPU, ya sean todas en una sola máquina virtual o distribuidas a lo largo de varias.

Aquí hay dos alternativas :

1) Crear una máquina virtual con solo 8 vCPU y vivir en un mundo completamente seguro y feliz, donde jamás Google les va a cobrar nada, y lo único que va a pasar es que una vez consumidos los USD 300 se les borren las máquinas virtuales creadas y particularmente el bucket con todos los datos que tengan dentro, con todas las salidas de las Optimizaciones Bayesianas.

Es un mundo completamente protegido, que no requiere de atención, solo es necesaria paciencia para esperar a que terminen los procesos.

Una máquina de 16vCPU no funciona 2 veces más rápido que una de 8 vCPU, el escalamiento del procesamiento NO ES LINEAL, ya que los programas por más que estén escritos para correr en paralelo, hay partes inevitables que corren en forma secuencial.

2) Hacer algo más arriesgado que necesita un monitoreo diario:

2.1) Darle autorización a Google para que cuando se pasen de los USD 300, empiece a cobrar. De esa forma podrán crear tener corriendo más de 8 vCPU al mismo tiempo.

2.2) Controlar en forma diaria cuántos USD van gastando, y al llegar a los USD 270, apagar todas las máquinas virtuales y borrarlas, bajarse los datos que quieran del bucket, borrar el Bucket, borrar la imagen del sistema operativo, es decir dejar sin nada cosa que lo que google Cloud quiera cobrarles por mes sea CERO .

En este caso les quedará la cuenta quizás con USD 15 a favor , y no les cobrará nada jamás.

El umbral de USD 270 es porque hay un delay de alrededor de 24 horas entre el consumo que muestra Google Cloud en la página versus el real que terminará facturando.

El punto 2) implica estar alerta, y revisar todo el tiempo el consumo, en particular cuando se pone una máquina grande a funcionar.

Si el alumno no tiene un computer literacy adecuado y no tiene una férrea disciplina de control es preferible que opte por la opción 1) (lo que tiene actualmente), para vivir más tranquilo en un

mundo seguro y feliz. Después de todo, se puede perfectamente hacer todo con 8 vCPU , a lo sumo se esperan más horas.

Aclaración importante, Google corre un algoritmo para detectar a personas que probablemente luego de terminado el periodo gratis jamás paguen por nuevos servicios, y en esos casos NO permite pasar de las 8 vCPU. No está claro cómo funciona este algoritmo.

Adicionalmente, alumnos han reportado haber sido contactados por Atención al Cliente de Google Cloud, en donde les hacen una campaña preventiva de retención de clientes, intentando lograr que luego de agotar los USD 300 pasen al modo pago y se transformen en clientes reales.

Jamás ha sido reportado por un alumno el evento que Google les bloquee definitivamente una cuenta.

Todos los años, sucede con probabilidad 0.02 ( 1 en 50 ), que un alumno pasó al modo desprotegido, no controla diariamente si se olvidó una máquina virtual inmortal encendida, se pasa de los USD 300 gratuitos, y Google le factura a su tarjeta. En la gran mayoría de estos casos han podido revertirlo implorando piedad a Google haciéndose pasar por desvalidos estudiantes de un país del tercer mundo que no entiende muy bien que es la nube y que solo obedecía ordenes de una cátedra...

### 3.3 Disciplina de trabajo

- Usted tiene su propio repositorio *hosteado* en GitHub
- Si usted tiene un avanzado computer literacy y ya conoce GitHub, se las arreglará solo, y no siga leyendo este capítulo de Disciplina de trabajo.
- En caso contrario, usted trabajará en una máquina virtual Ubuntu con interfase gráfica llamada **desktop** que se encuentra en el datacenter de Google Cloud en San Pablo, Brasil.
- Todos los scripts y archivos de experimentos se desarrollan (escriben y/o modifican) desde VSCode en la máquina virtual **desktop**, y deberá mantener constantemente sincronizado su repositorio en desktop con su repositorio en GitHub
- Se hará una demostración completa en clase de la forma de trabajo.
- La máquina virtual solamente se utiliza para desarrollar, para correr scripts grandes deberá hacerlo en nuevas máquinas virtuales
- Para cada nueva corrida, cada optimización bayesiana, se procede de esta forma:
  - Se modifican los scripts en la máquina virtual gráfica **desktop**
  - Sincroniza los repositorios
  - Se crea una nueva máquina virtual a partir de los templates
  - Se ingresa al RStudio de esa máquina virtual
  - Se pone a correr el experimento desde Rstudio de la máquina virtual
  - De ninguna forma usted puede verse tentado a hacer un cambio de último momento en el script y que eso no quede reflejado en el repositorio GitHub, hay decenas de casos de buenas corridas que no se han podido volver a reproducir por no haber guardado la versión del script que corrió.
  - Se espera unos minutos para ver que levantó bien el dataset y empezó a escribir en los logs
  - Un tema delicado es saber si usted le asignó la suficiente cantidad de memoria RAM a su máquina virtual. Luego de lanzar a correr su experimento desde el RStudio, puede ingresar a la terminal Ubuntu, correr el comando **htop** y monitoree el uso de memoria RAM los primeros minutos.
  - Toda optimización Bayesiana sobre varios meses demandará al menos 128GB de memoria RAM, y si trabaja con datasets con mas de mil atributos necesitará 256 GB de RAM. Si la memoria RAM es insuficiente, deberá BORRAR esa máquina virtual y volver a crear otra con EL DOBLE DE MEMORIA; no sea tacaño, ir subiendo de apenas unos pocos GB para probar es caminar sobre espinas.
  - Monitorear periódicamente que la máquina virtual spot no fue apagada por Google y que además de estar encendida nuestro experimento continúa corriendo (se puede monitorear desde la app de Google Cloud Console para móviles en caso que desee hacerlo un sábado a la noche desde un local bailable)
  - Si la máquina se apagó y eliminó, usted debe volver a crear la misma máquina, y poner a

correr nuevamente el experimento, los experimentos están diseñados para retomar la Optimización Bayesiana desde donde había quedado.

- Podría llegar a suceder que sea de día en ese datacenter, se estén utilizando todos los recursos, y usted vea que la máquina virtual no se enciende, o se apaga enseguida. En este caso, va a tener que eliminarla, y crear una nueva en otro datacenter de Google.
  - Si la máquina está encendida pero el experimento ya no está corriendo verifique si ha sido tan afortunados que su proceso ya terminó.
  - Si el proceso abortó por algún error, primero debe prepararse un café bien cargado y luego empezará la batalla por entender el motivo y como corregirlo.
  - Si dejó corriendo un experimento durante la noche, al despertar verifique su estado.
  - Una vez que está todo terminado, se verifican que existan los archivos de salida.  
apagar la máquina virtual
  - verificar que realmente quedó apagada, que no está el tilde verde
  - **eliminar la máquina virtual. Por favor, no reutilice las máquinas virtuales, solo encontrará angustia y sufrimiento; para cada corrida debe crear una nueva máquina virtual.**
  - Verifique que la máquina virtual ya no aparece en VM Instances. Una máquina virtual apagada continúa consumiendo dinero, ya que Google cobra por el espacio de disco local.
- 
- Desestime la recomendación de Google de aumentar la cantidad de vCPU, ellos ven que su proceso está al 100% de uso de CPU y solo quieren facturarle más. Toda corrida suya de optimización de hiperparámetros siempre estará al 100% independientemente de cuantas vCPU asigne a esa máquina virtual.

## No le de importancia a estos warnings

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

(simplemente está diciendo que no puede construir un nuevo árbol de decisión en el ensemble ya que no son lo suficientemente buenos)

```
[LightGBM] [Warning] Auto-choosing col-wise multi-threading, the overhead of testing was 0.962394 seconds.
```

```
You can set `force_col_wise=true` to remove the overhead.
```

(simplemente está informando sobre una optimización)

```
[LightGBM] [Warning] verbosity is set=-1, verbose=-1 will be ignored. Current value: verbosity=-1
```

(no hay ningún tipo de problema)

aprenda a convivir con warnings, **NO TEMA**

### 3.4 Crear una Máquina Virtual desde template

Ingresa al link <https://console.cloud.google.com/compute/instancesAdd?creationFlow=fromTemplate>

verá algo similar a esto:

Google Cloud CRANR001 Search (/) for resources, docs, products, and more

← Create an instance

To create a VM instance, select one of the options:

- New VM instance**  
Create a single VM instance from scratch
- New VM instance from template**  
Create a single VM instance from an existing template
- New VM instance from machine image**  
Create a single VM instance from an existing machine image
- Marketplace**  
Deploy a ready-to-go solution onto a VM instance

**1 Select template** — **2 Customize VM instance**

**Filter** Filter instance templates

	Name ↑	Machine type	Disk type	Creation time
<input type="radio"/>	instance-template-proxy-immortal	f1-micro	Standard persistent disk	Dec 2, 2017, 3:0
<input type="radio"/>	instance-template-proxy-mortal	f1-micro	Standard persistent disk	Dec 2, 2017, 3:0
<input type="radio"/>	temp-08vcpu-016ram	custom-8-16384-ext	Standard persistent disk	Apr 6, 2023, 2:15
<input type="radio"/>	temp-08vcpu-032ram	custom-8-32768-ext	Standard persistent disk	Apr 6, 2023, 2:15
<input type="radio"/>	temp-08vcpu-064ram	custom-8-65536-ext	Standard persistent disk	Apr 6, 2023, 2:16
<input type="radio"/>	temp-08vcpu-128ram	custom-8-131072-ext	Standard persistent disk	Apr 6, 2023, 2:16
<input type="radio"/>	temp-08vcpu-256ram	custom-8-262144-ext	Standard persistent disk	Apr 6, 2023, 2:16

CONTINUE

Elija la máquina virtual que se adapte a la memoria RAM que requiere el script que va a correr. Para todos los scripts diseñados para correr en Google Cloud la cátedra ha comentado al comienzo del mismo la cantidad de memoria RAM que requiere, en la medida que utilice un dataset de dimensiones similares. Si usted le ha agregado al dataset cientos de atributos adicionales, deberá elegir una máquina virtual con más memoria RAM.

Luego de elegir un template, presione el botón **CONTINUE**

en la siguiente pantalla debe elegir

- **Name** nombre de la máquina virtual (generalmente el script que va a correr)
- **Region**
- **Zone**

Usted debe buscar una Region de forma que el costo, ese día a esa hora, sea lo menor posible. Generalmente Las Vegas es el menor costo.

La **Zone** depende de la Region, y todas las Zone de la misma Region tienen el mismo costo. Una Region es un edificio distinto dentro del mismo predio al que llaman Zone, que es donde está el datacenter.

✓ Select template — 2 Customize VM instance

#### Source template

temp-08vcpu-016ram

[CHANGE TEMPLATE](#)

Name \*

z723

Labels ?

[+ ADD LABELS](#)

Region \*

us-west4 (Las Vegas)

Region is permanent

Zone \*

us-west4-a

Zone is permanent

#### Monthly estimate

\$30.51

That's about \$0.04 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
8 vCPU + 16 GB memory	\$27.69
64 GB standard persistent disk	\$2.82
Total	\$30.51

[Compute Engine pricing](#)

[^ LESS](#)

### Machine configuration



General purpose

Compute optimized

Memory optimized

GPUs

Machine types for common workloads, optimized for cost and flexibility

Finalmente dar al boton **Create** que está al final del formulario, y volver a la página <https://console.cloud.google.com/compute/instances> que es donde verá encenderse su máquina virtual.

Para ingresar a RStudio o Jupyter Lab pase a los siguientes capítulos

Nota para alumnos avanzados : los templates que están creados tienen todos 8 vCPU y la memoria RAM va de 16GB a 256GB . Es posible que en las últimas semanas usted se vea tentado a crear máquinas virtuales con más vCPU para acelerar, aunque sea algo, los procesos.

Si usted desea crear templates con otras configuraciones, debe ir a

<https://console.cloud.google.com/compute/instanceTemplates/add> y tener en cuenta lo siguiente

- En name elija un nombre nemotécnico para su template
- En Series debe elegir **N1** , en Machine type **Custom**, debe marcar el recuadro **Extended Memory**
- Elija los **Cores** y **Memory** que desea
- En Boot disk debe ir a la solapa CUSTOM IMAGES y elegir **image-dm** , en Boot disk type elija **Standard Persistent Disk**, en Size el doble de la memoria RAM que eligió ( se va a encontrar con el limite de 300GB )
- En Access scopes elija **Allow full access to all Cloud APIs**
- En Firewall marque ambas opciones **Allow HTTP traffic** y **Allow HTTP traffic**
- Lo MAS IMPORTANTE, en Advanced Options / Management / VM provisioning Model elegir **Spot** , y en Advanced Options / Management / On VM termination elegir **Delete** .

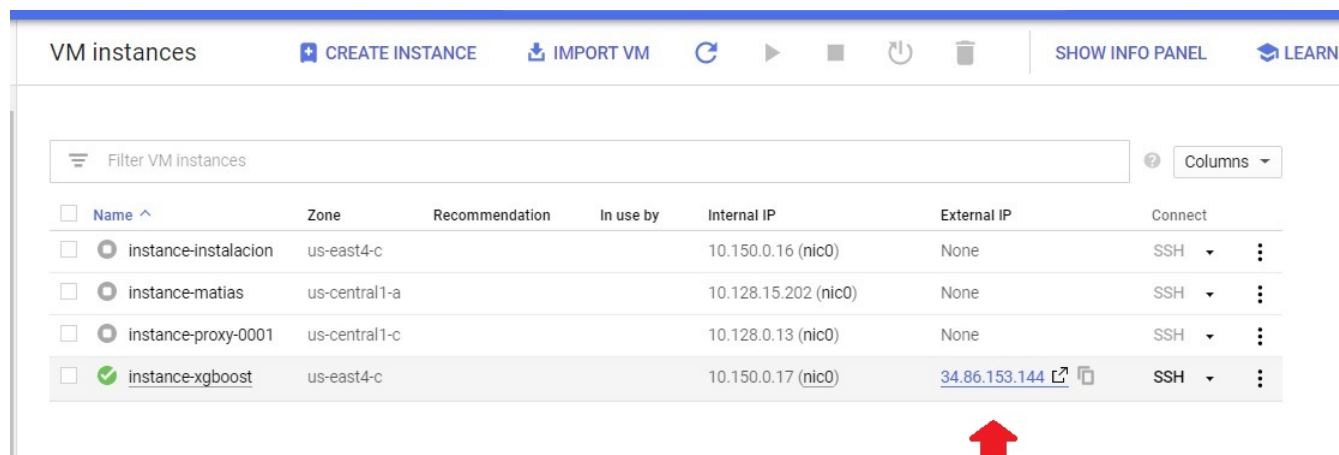


### 3.5 RStudio en forma remota

RStudio fue configurado para trabajar con el puerto 80, de forma de poder ser utilizado detrás del firewall de una universidad, empresa, cafetería, etc

Para RStudio se debe utilizar el usuario y claves vistos en el punto "2.1 asignar password al usuario".


La forma de ingresar al RStudio copiando la dirección External IP (pública)



<input type="checkbox"/>	Name ^	Zone	Recommendation	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	instance-instalacion	us-east4-c			10.150.0.16 (nic0)	None	SSH ▾ ⋮
<input type="checkbox"/>	instance-matias	us-central1-a			10.128.15.202 (nic0)	None	SSH ▾ ⋮
<input type="checkbox"/>	instance-proxy-0001	us-central1-c			10.128.0.13 (nic0)	None	SSH ▾ ⋮
<input checked="" type="checkbox"/>	instance-xgboost	us-east4-c			10.150.0.17 (nic0)	<a href="http://34.86.153.144">34.86.153.144</a> ↗ 📄	SSH ▾ ⋮

y en el browser poner <http://34.86.153.144/> notar que dice [http](http://34.86.153.144/) SIN la letra "s" al final (en este ejemplo utilicé mi IP pública, reemplazarla por la que le aparece a usted)

y se abrirá en el browser de la PC local la página de ingreso al RStudio



### Sign in to RStudio

Username:

Password:

☐ Stay signed in

[Sign In](#)

Es muy importante notar, que el acceso a RStudio es mediante una conexión insegura del tipo `http://` y NO mediante la conexión segura de `https://` , hay alumnos que utilizan laptops corporativas y esto les ha generado imposibilidad de acceso.

Una vez dentro, navegar a la carpeta del repositorio [labo](#) buscar el script, y comenzar a ejecutarlo.

Dentro de Rstudio, ir a la solapa Terminal , que es la terminal de Ubuntu, y allí correr el comando `htop` el que mostrará el uso de CPU y memoria RAM. Lo importante a controlar es el Swap , aparece debajo de la memoria como `Swp` , si pasa de cero y las CPU estan con muy poco uso por mas de un minuto, entonces quiere decir que se creó una máquina virtual con insuficiente memoria RAM para ese <script, dataset> ; debe eliminarse esa máquina virtual , crear una nueva con el doble de memoria y volver a correr el script.

### 3.6 Jupyter Lab en forma remota

Desde el navegador de la PC local, utilizando la External IP (pública) de la máquina virtual escribir `http://35.44.55.66:8888/lab`

(reemplazar lo rojo el la IP que corresponda)

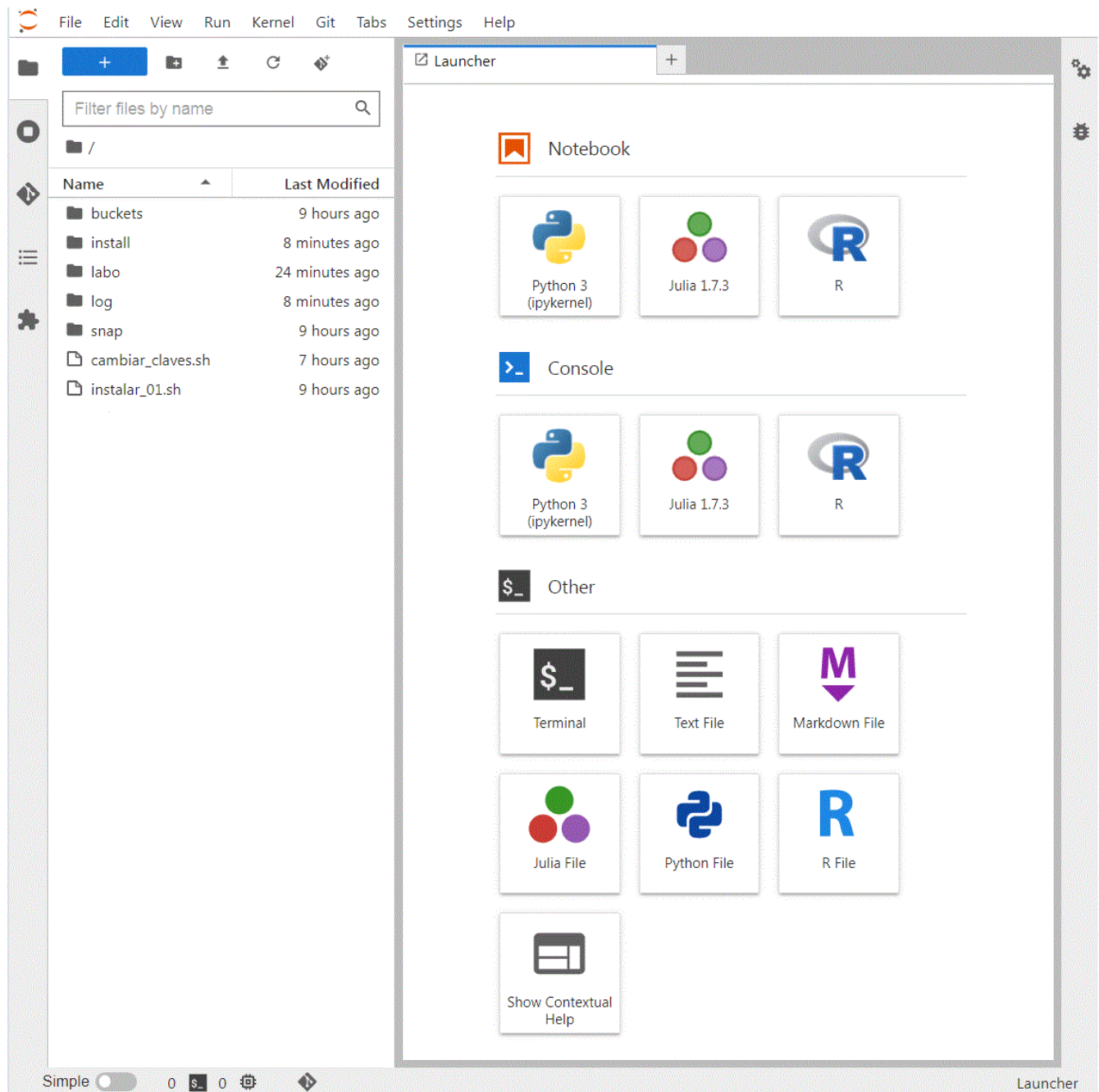
notar aquí lo mismo, es también una conexión insegura del tipo `http://` ya que NO funciona con la letra "s" al final, o sea `https://` . Lamentablemente la conexión a Jupyter Lab se realiza por el puerto 8888 el que podría estar bloqueado por su empleador.

La primer pantalla que aparece le solicita la clave de Jupyter Lab, que la definió en `cambiar_claves` y casi con seguridad es la misma que la de RStudio



The image shows a web interface for Jupyter Lab. At the top, there is a horizontal line, followed by the Jupyter logo (an orange circle with a white 'j' inside) and the word 'jupyter' in a sans-serif font. Below this, there is another horizontal line. Further down, the text 'Password:' is displayed to the left of a rectangular text input field. To the right of the input field is a button labeled 'Log in'.

luego de unos 20 segundos le aparecerá



Dentro de JupyterLab navegar dentro de la carpeta **labo** hasta donde este el notebook que se desea ejecutar, cargarlo, y ejecutarlo

### 3.7 Desde una terminal Ubuntu con la consola de R correr scripts

No vaya por este camino a menos que usted tenga más de 3000 ( tres mil ) horas de experiencia en programación generadas en los últimos 5 años.

Esta es una forma de muy bajo nivel para correr un script, se le recomienda permanecer en la zona de confort del RStudio y su hermosa y segura interfase gráfica.

En primera instancia se debe abrir la terminal Ubuntu

luego, para correr desde la terminal se debe tipear (es una sola línea )

```
$ nohup Rscript --vanilla ~/labo/src/rpart/z101_PrimerModelo.R &
```

( decididamente [vanilla](#) ni es un script ni un paquete de R )

## 4 Cuando Finalice la Materia

### 4.1 Evitar que Google Cloud facture

Para que Google Cloud jamás le facture nada, al finalizar la materia se debe proceder de la siguiente forma :

- Borrar todas las imágenes de disco creadas
- Borrar todos los datos dentro de los Buckets creados ( o el único Bucket que se creó )
- Eliminar todos los Buckets creados
- Apagar todas las máquinas virtuales creadas
- Eliminar todas las máquinas virtuales creadas
  
- En caso de haber creado algún otro objeto en Google Cloud que no fue mencionado en la materia, eliminarlo .