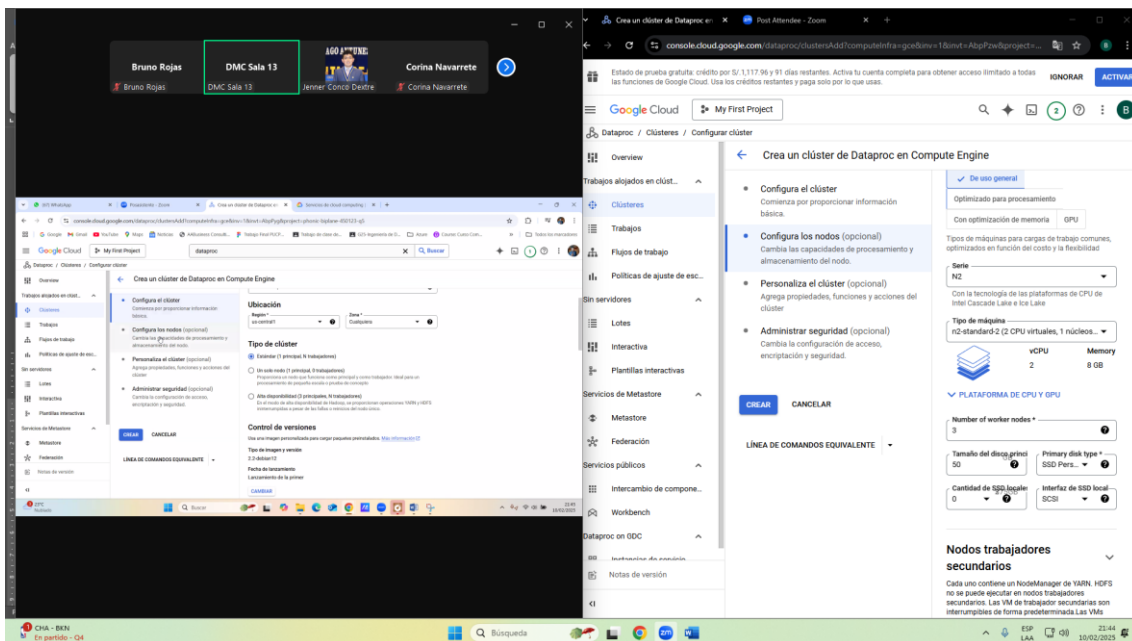
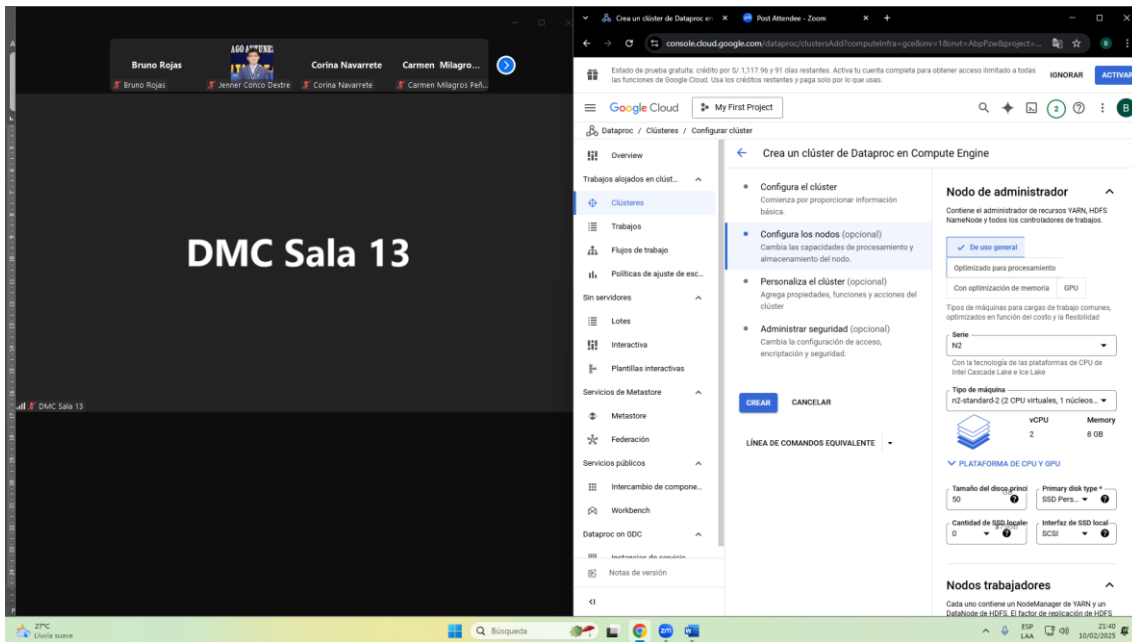


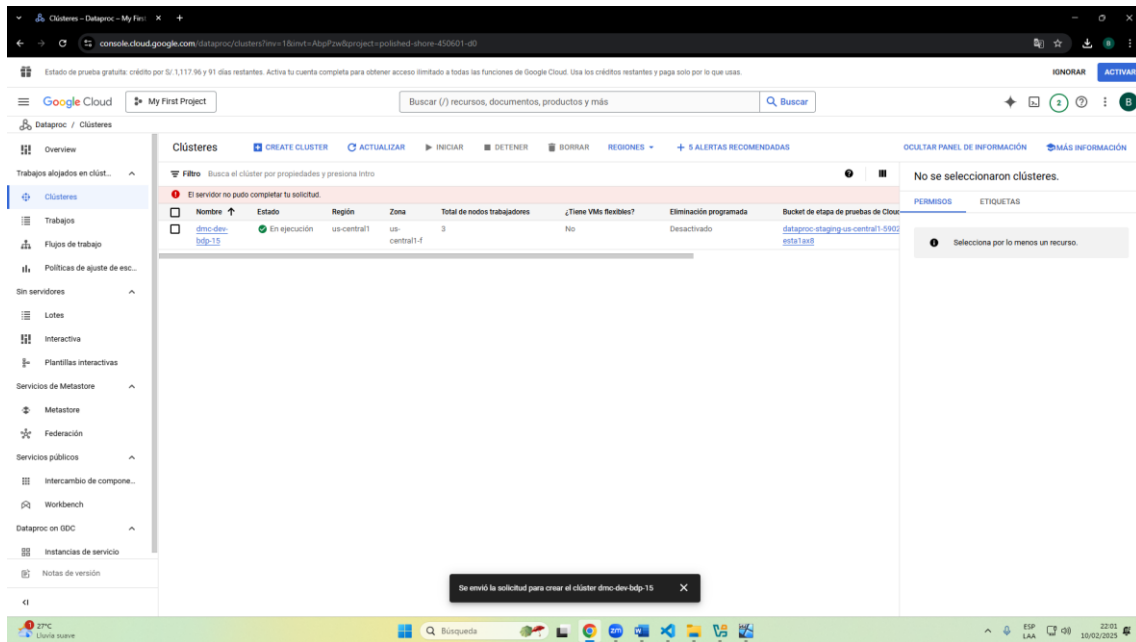
HDFS

Creación de Clúster Dataproc

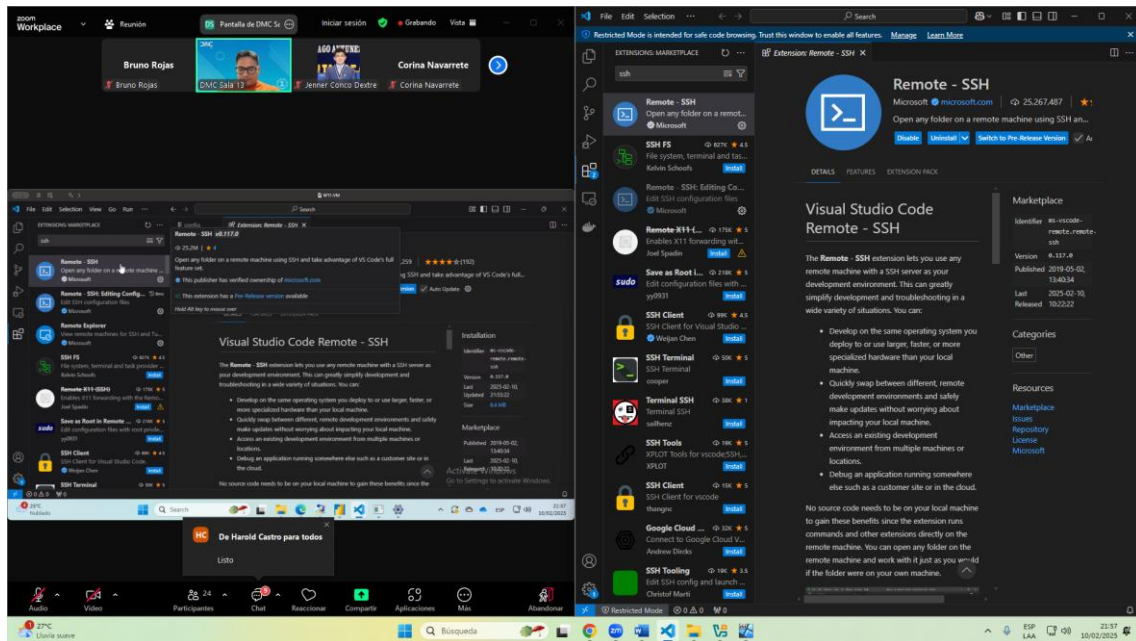
The screenshot shows a Zoom meeting window on the left and a Google Cloud console window on the right. The Zoom window displays a meeting with participants Bruno Rojas, DMC Sala 13, Jener Cortico Destre, and Corina Navarrete. The Google Cloud console window shows the 'Detalles del servicio o la API' (Service or API details) for the 'Cloud Dataproc API'. The API is listed as 'API pública' (Public API) and is 'Habilitado' (Enabled). The console also shows the 'Métricas' (Metrics) section, which includes a graph of API usage over the last 30 days. The graph shows a peak in usage around the 15th of the month.

The screenshot shows a Zoom meeting window on the left and a Google Cloud console window on the right. The Zoom window displays a meeting with participants Bruno Rojas, DMC Sala 13, Jener Cortico Destre, and Corina Navarrete. The Google Cloud console window shows the 'Crea un clúster de Dataproc en Compute Engine' (Create a Dataproc cluster in Compute Engine) wizard. The wizard is currently on the 'Configura el clúster' (Configure the cluster) step. The 'Nombre' (Name) field is set to 'cluster-0850'. The 'Ubicación' (Location) is set to 'us-central1'. The 'Tipo de clúster' (Cluster type) is set to 'Estándar (1 principal, N trabajadores)' (Standard (1 primary, N workers)). The 'Control de versiones' (Version control) section is also visible, showing the 'Tipo de imagen y versión' (Image type and version) as '2.2-debian12'.

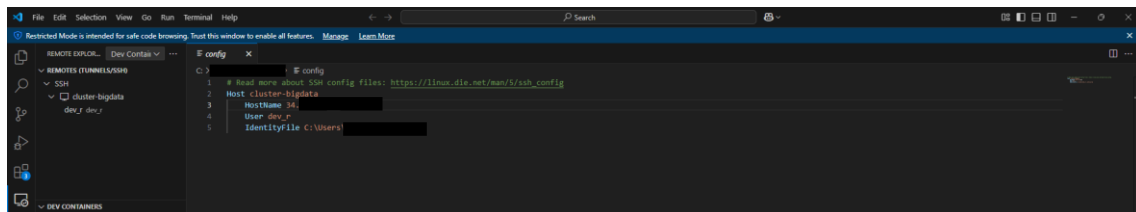




Instalación de Remote Explorer en Visual Studio Code

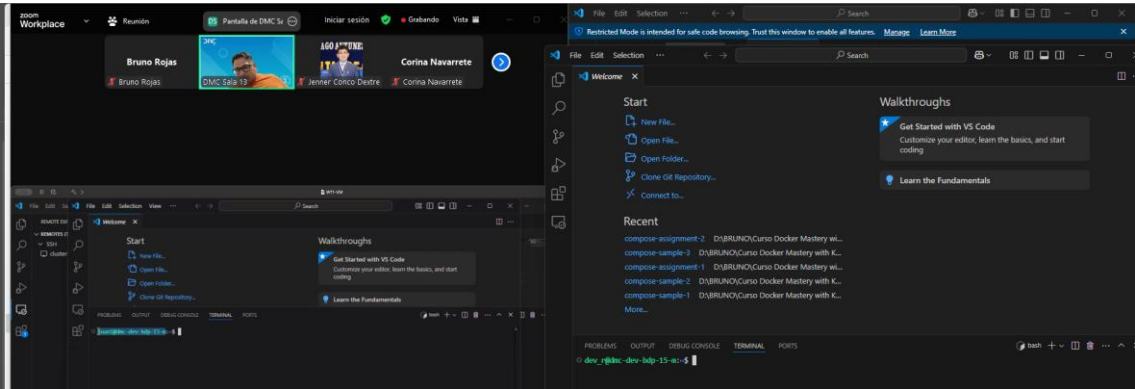


Configuración de Remote Explorer

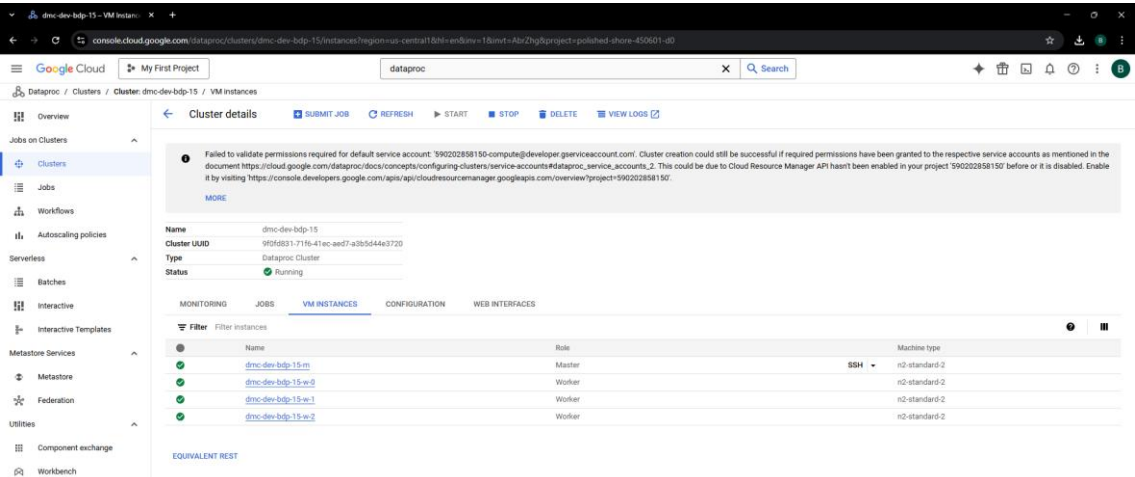
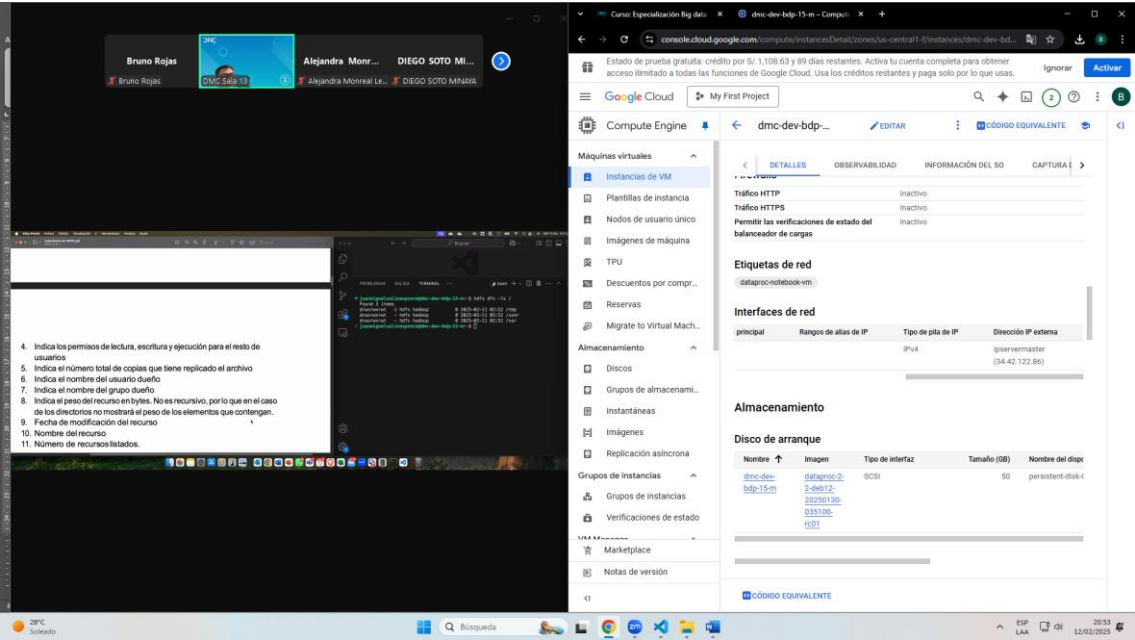


Node	LastWriteTime	Length	Name
-a----	10/02/2025 22:10	2610	id_rsa
-a----	10/02/2025 22:10	576	id_rsa.pub
-a----	2/12/2023 19:20	840	known_hosts
-a----	2/12/2023 18:45	96	known_hosts.old

Nos conectamos al cluster desde Visual Studio Code con Remote Explorer



Convertir IP Efímera en Estática o Permanente



Entramos al master y bajamos hasta network interfaces

Google Cloud console showing the details of a VM instance named 'dataproc-notebook-vm'. The 'Network interfaces' section is expanded, showing a table with columns: Name, Network, Subnetwork, Primary internal IP address, Alias IP ranges, IP stack type, External IP address, and Network. The table contains one entry: 'nic0' with network 'default', subnetwork 'default', primary IP '10.128.0.4', and external IP '34.42.122.86'. Below this, the 'Storage' section is expanded, showing the 'Boot disk' table with columns: Name, Image, Interface type, Size (GB), Device name, Type, Architecture, Encryption, Mode, and Wipe on destroy. The boot disk is 'dmc-dev-bdp-15-m' with image '2248612-20230319-035100-r001' and size '50 GB'.

La ip que normalmente aparece como ephemeral o efímera

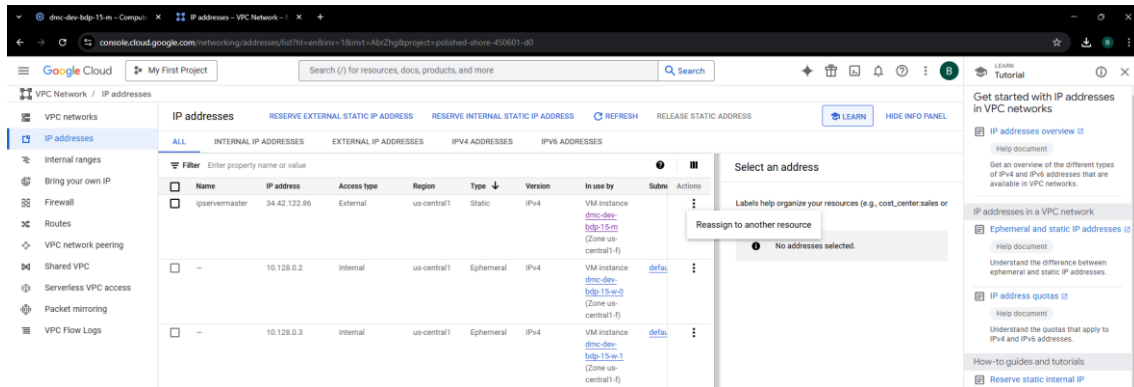
Click derecho sobre nic0 -> Abrir en una nueva ventana

Google Cloud console showing the 'Network interface details' for 'nic0'. The 'Selected network interface' is 'nic0'. The 'Network interface details' table shows: Name (nic0), Network (default), Subnetwork (default), Primary internal IP address (10.128.0.4), Alias IP ranges (none), IP stack type (IPv4), External IP address (34.42.122.86), and Network Service Tier (Premium). The 'VM instance details' table shows: Name (dmc-dev-bdp-15-m), Zone (us-central1-f), Network tags (dataproc-notebook-vm), Service account (590202858150-compute@developer.gserviceaccount.com), and IP forwarding (Off). The 'Firewall and routes details' section shows a table with columns: Name, Enforcement order, Type, Deployment scope, Rule priority, Targets, Source, Destination, Protocols and ports, Action, and Security profile group. The table contains one entry: 'vpc-firewall-rules' with enforcement order '1', type 'VPC firewall rules', and deployment scope 'Global'.

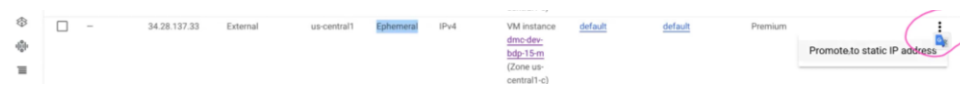
Click IP addresses

Google Cloud console showing the 'IP addresses' page. The 'IP addresses' table has columns: Name, IP address, Access type, Region, Type, Version, In use by, and Subnet. The table contains five entries: 'ipservermaster' (34.42.122.86, External, us-central1, Static, IPv4, VM instance dmc-dev-bdp-15-m (Zone us-central1-f)), and four ephemeral addresses (10.128.0.2, 10.128.0.3, 10.128.0.4, 10.128.0.5) all in us-central1, IPv4, used by VM instances dmc-dev-bdp-15-m-e0, dmc-dev-bdp-15-m-e1, dmc-dev-bdp-15-m-e2, and dmc-dev-bdp-15-m-e3 respectively. A 'Select an address' dialog is open on the right, showing 'No addresses selected.'

Click en los tres puntitos de la IP que está usando nuestro servidor y dar click en Promote to Static IP Address



Debería aparecer promover a ip estática

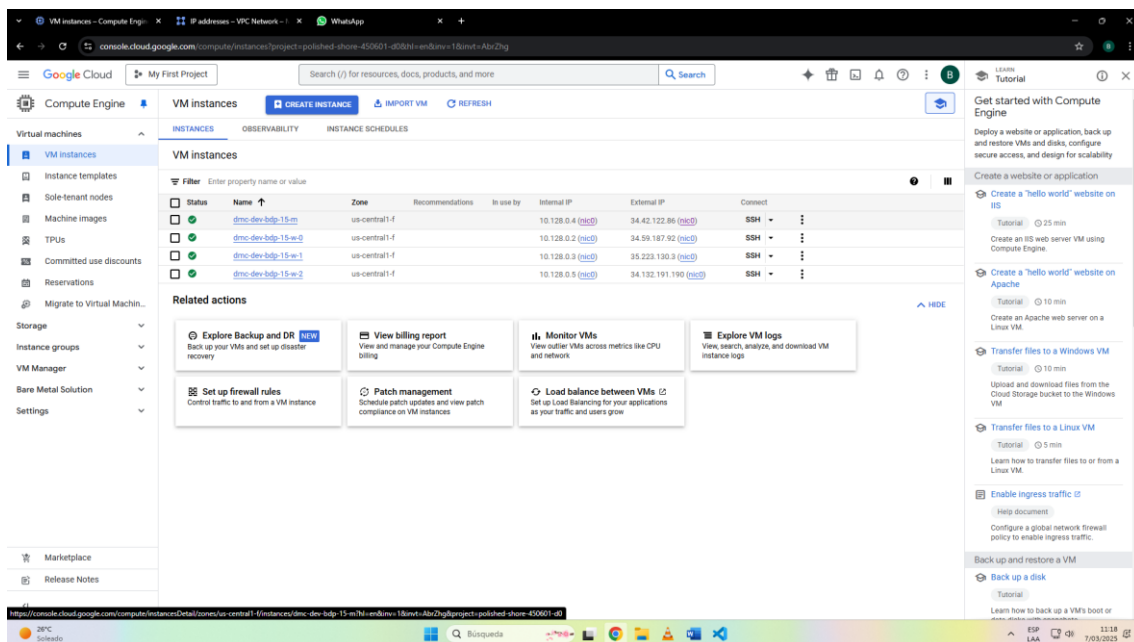


Le colocamos un nombre y descripción

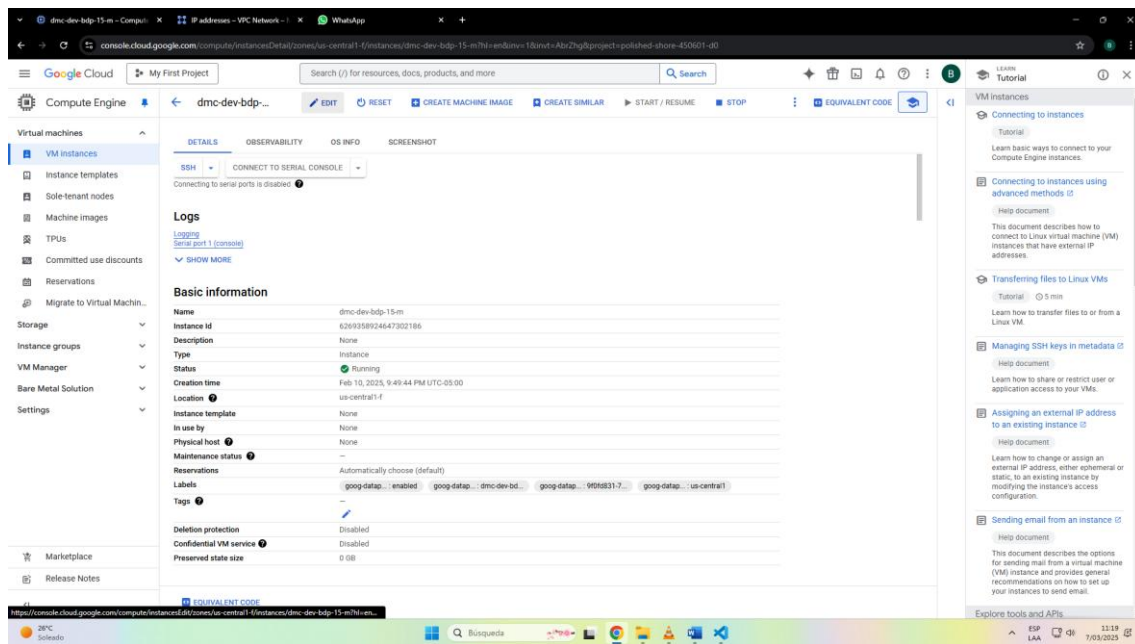


Luego la ip cambiaría de efímera a estática y es la que usaríamos siempre para conectarnos.

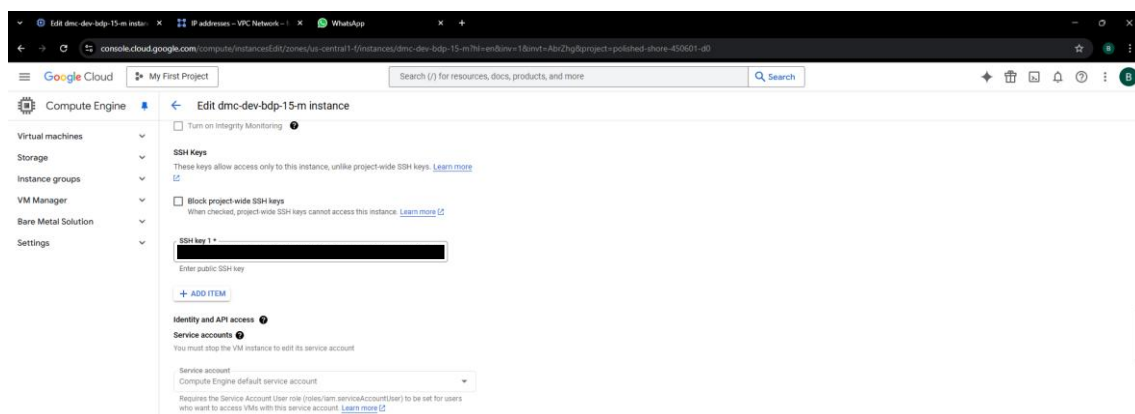
Para añadir una llave ssh



Entramos al servidor maestro dmc-dev-bdp-15-m



Damos click en edit. Bajamos hasta ssh keys. Y damos click en add item y se añade la llave pública.



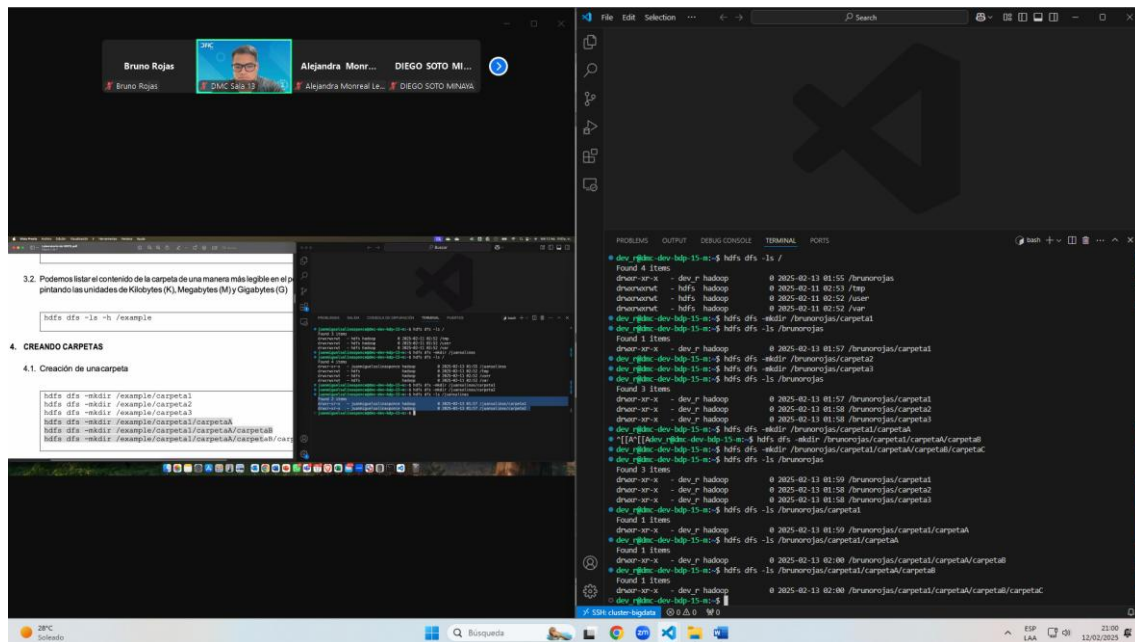
Comandos en HDFS

Listar carpetas en HDFS: `hdfs dfs -ls /`

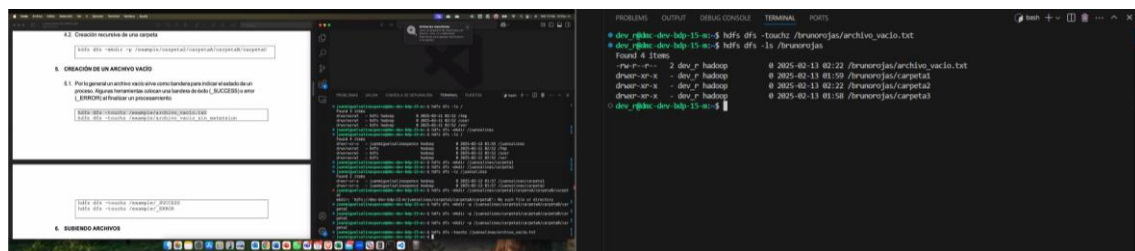
Crear directorios en HDFS: `hdfs dfs -mkdir /brunorojas`

Crear una carpeta dentro de otra: `hdfs dfs -mkdir /brunorojas/carpeta2`

Listar el contenido de la carpeta: `hdfs dfs -ls /brunorojas`



Crear archivo vacío: `hdfs dfs -touchz /brunorojas/archivovacio.txt`



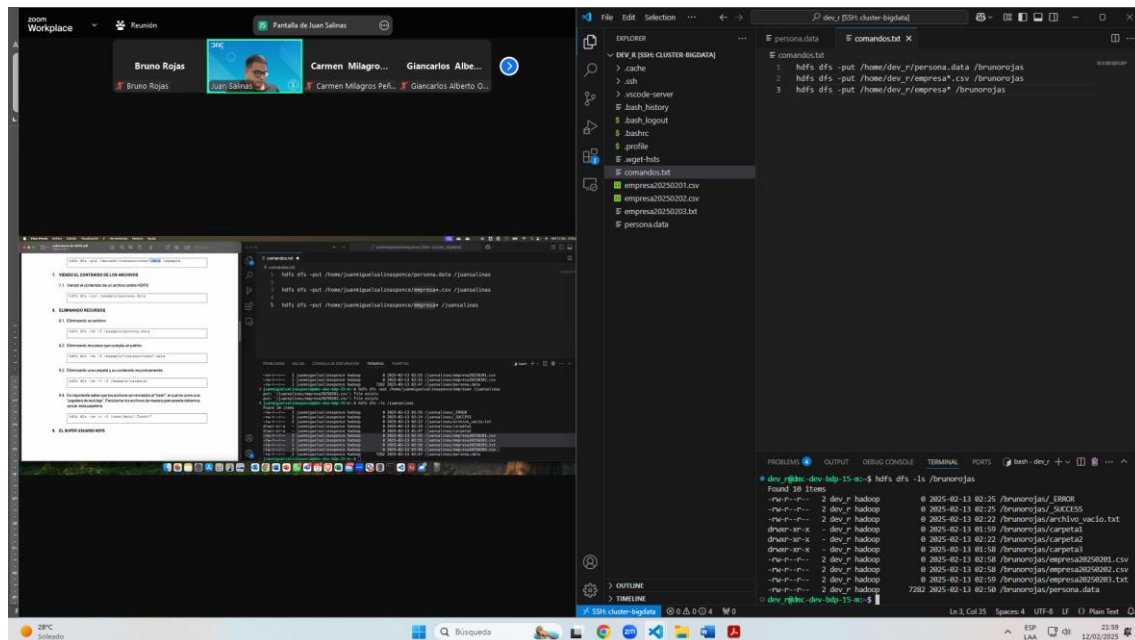
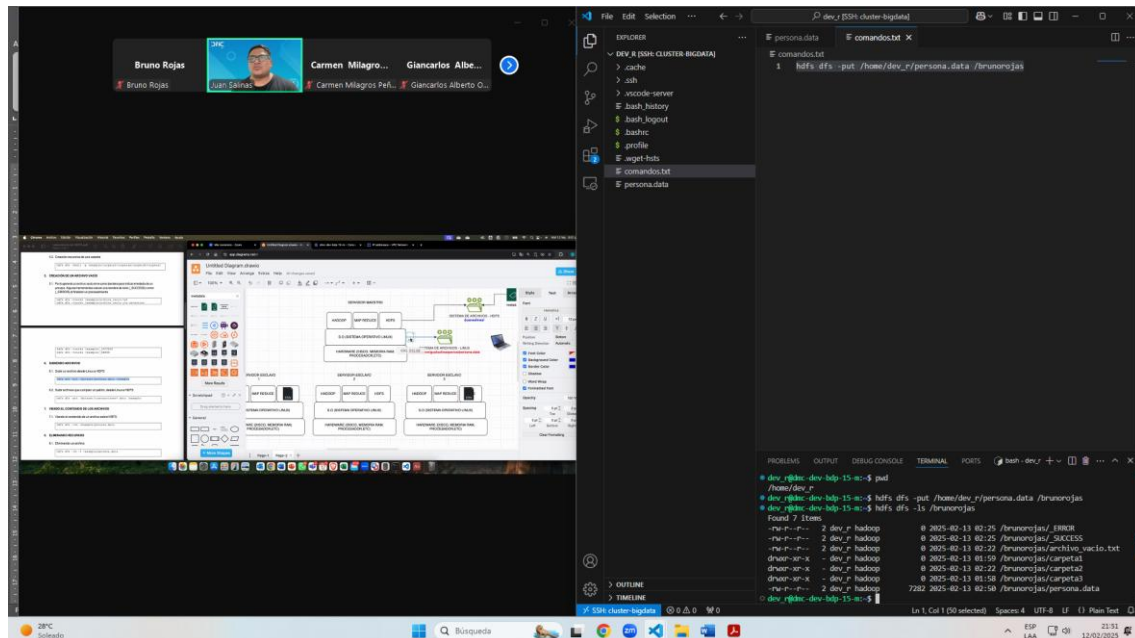
Creamos un archivo txt dentro del servidor para ir guardando nuestros comandos.

Para llevar un archivo de Linux a hdf se usa el comando put

```
hdfs dfs -put /home/dev_r/persona.data /brunorojas
```

```
hdfs dfs -put /home/dev_r/empresa*.csv /brunorojas
```

```
hdfs dfs -put /home/dev_r/empresa* /brunorojas
```



Usamos el comando cat para leer el contenido de un archivo.

```
hdfs dfs -cat /brunorojas/persona.data
```

Usamos el commando rm para eliminar un archivo. La f es para forzar el borrado.

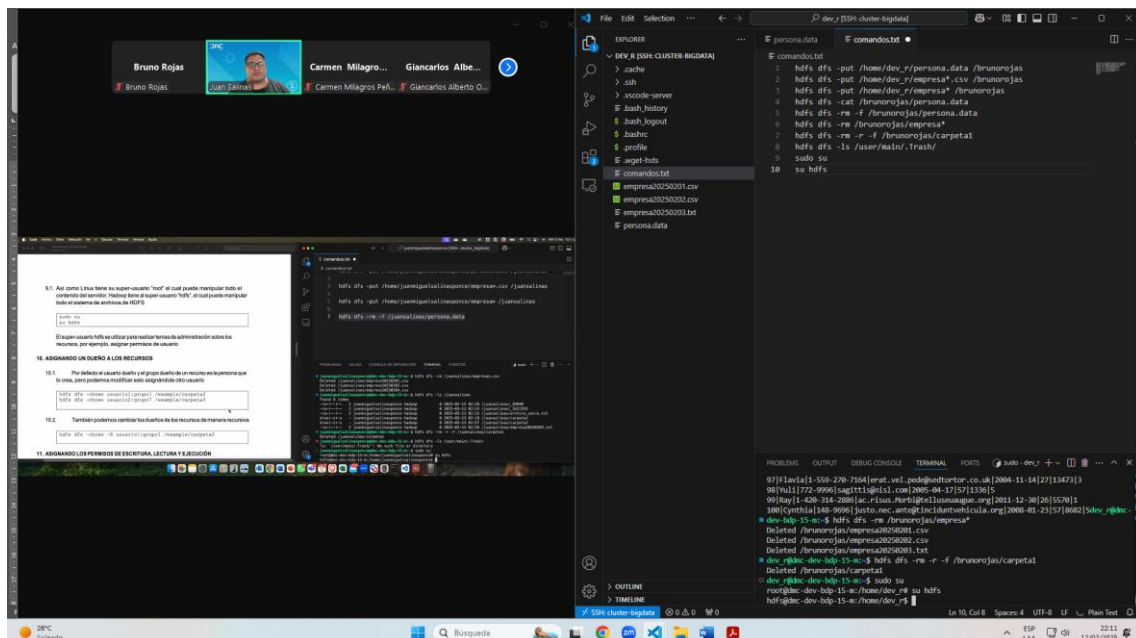
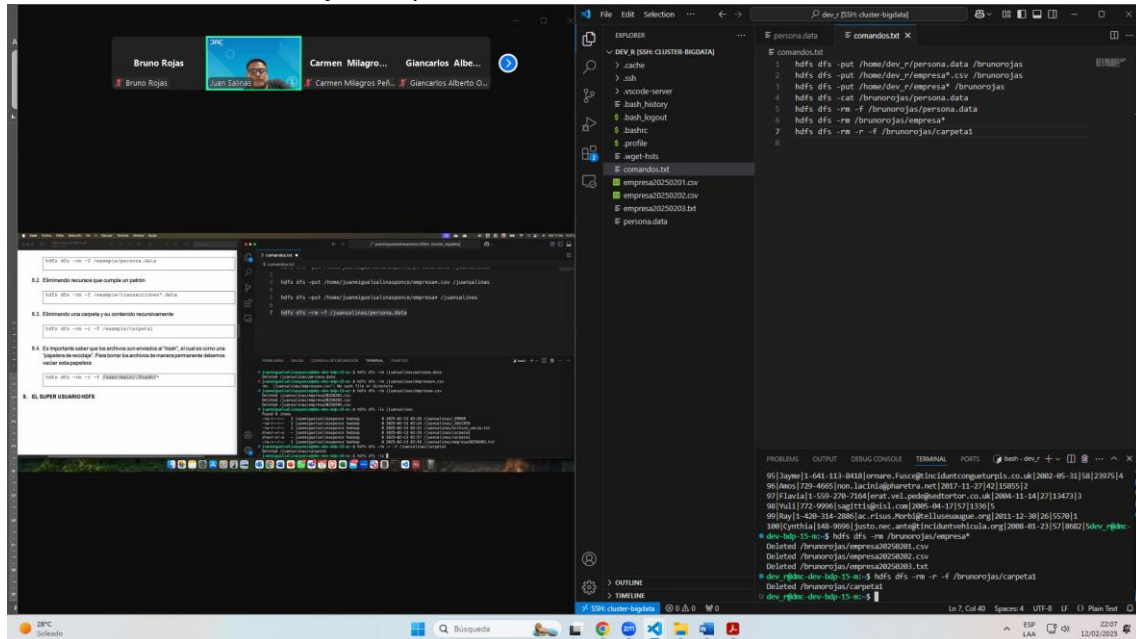
```
hdfs dfs -rm -f /brunorojas/persona.data
```

Se puede eliminar usando comodines también.

```
hdfs dfs -rm /brunorojas/empresa*
```

Si se quiere eliminar de forma recursiva es decir los archivos o carpetas dentro de una carpeta se usa -r

```
hdfs dfs -rm -r -f /brunorojas/carpeta1
```



Para poder realizar el cambio de permisos de ejecución, escritura y lectura así como del propietario de un archivo o carpeta se debe realizar con el superusuario.

Cambiamos a superusuario en Linux con : sudo su

Cambiamos a superusuario en hdfs con: su hdfs

Estructura del comando chown: chown usuario:grupo ruta

Cambiar de propietario a una carpeta

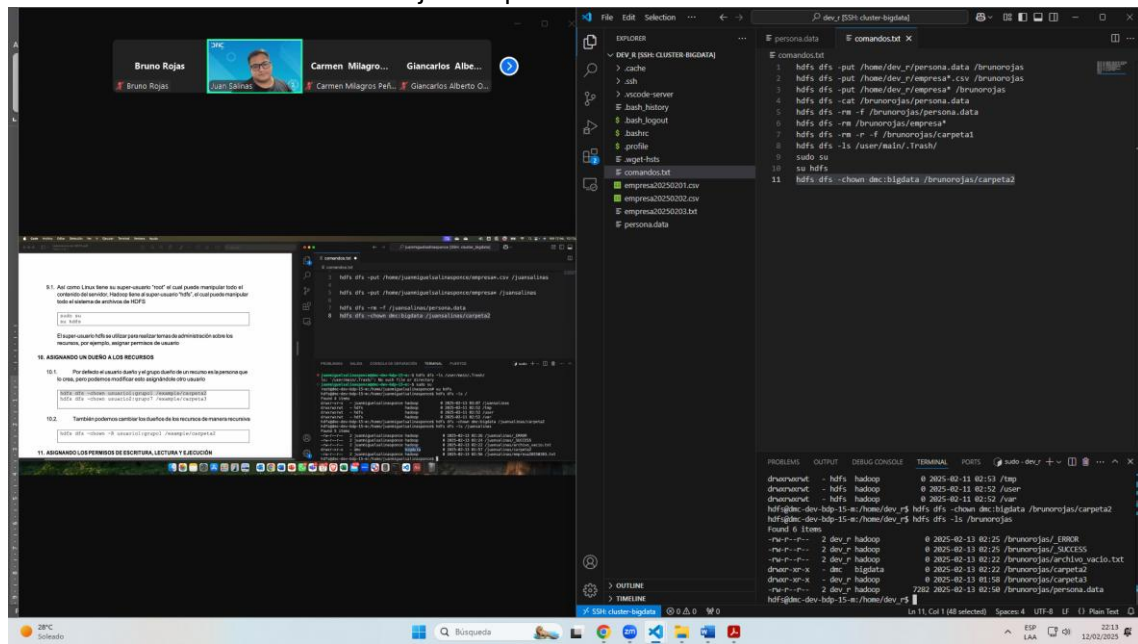
hdfs dfs -chown dmc:bigdata /brunorojas/carpeta2

Cambiar de propietario de forma recursiva

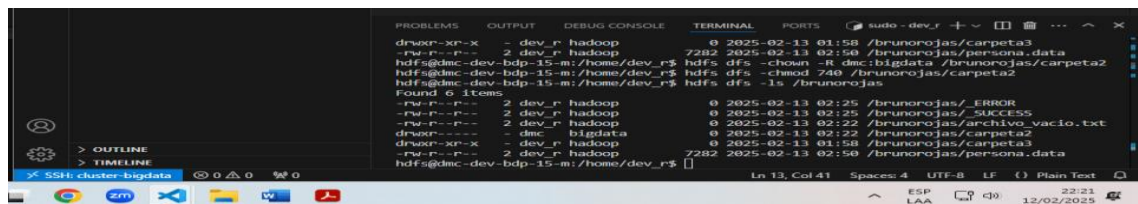
hdfs dfs -chown -R dmc:bigdata /brunorojas/carpeta2

hdfs dfs -chmod 740 /brunorojas/carpeta2

hdfs dfs -chmod -R 750 /brunorojas/carpeta2



```
hdfs dfs -put /home/dev_r/persona.data /brunorojas
hdfs dfs -put /home/dev_r/empresa*.csv /brunorojas
hdfs dfs -cat /brunorojas/persona.data
hdfs dfs -rm -f /brunorojas/persona.data
hdfs dfs -rm /brunorojas/empresa*
hdfs dfs -rm -f /brunorojas/carpeta1
hdfs dfs -ls /user/main/.Trash/
sudo su
hdfs
hdfs dfs -chown dmc:bigdata /brunorojas/carpeta2
```



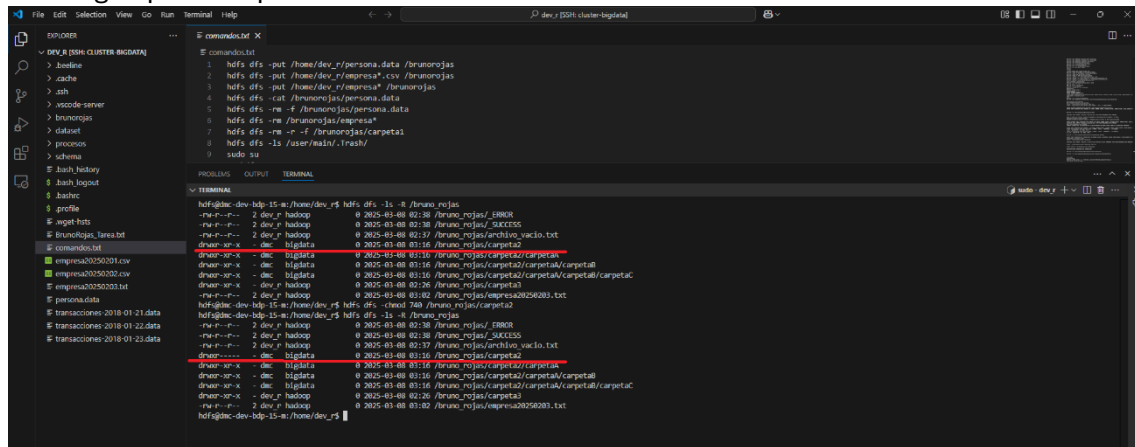
```
hdfs dfs -chown -R dmc:bigdata /brunorojas/carpeta2
hdfs dfs -chmod 740 /brunorojas/carpeta2
hdfs dfs -ls /brunorojas
Found 6 items
-rw-r--r-- 2 dev_r hadoop 0 2025-02-13 02:25 /brunorojas/ERROR
-rw-r--r-- 2 dev_r hadoop 0 2025-02-13 02:25 /brunorojas/SUCCESS
-rw-r--r-- 2 dev_r hadoop 0 2025-02-13 02:22 /brunorojas/archivo_vacio.txt
drwxr-x-- dmc bigdata 0 2025-02-13 02:22 /brunorojas/carpeta2
drwxr-x-- dev_r hadoop 7282 2025-02-13 02:50 /brunorojas/carpeta3
-rw-r--r-- 2 dev_r hadoop 7282 2025-02-13 02:50 /brunorojas/persona.data
```

```
hdfs dfs -chmod 740 /bruno_rojas/carpeta2
```

Permisos de lectura escritura y ejecución para DMC

Permisos de lectura para el grupo bigdata

Sin ningún permiso para otros usuarios



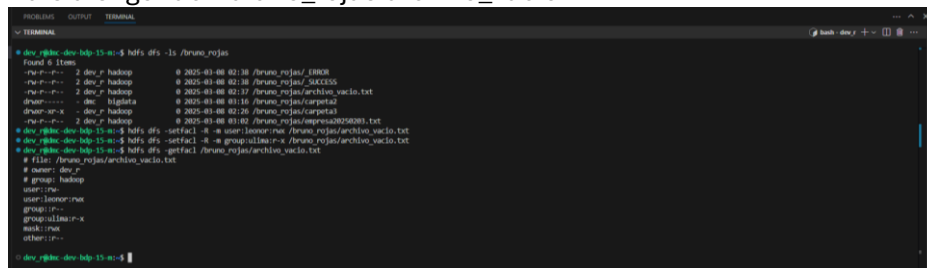
Permisos con ACL

Estructura del comando setfacl: `hdfs dfs -setfacl -R -m user:nombreusuario:wxw /ruta`

```
hdfs dfs -setfacl -R -m user:leonor:rw- /bruno_rojas/archivo_vacio.txt
```

```
hdfs dfs -setfacl -R -m group:ulimar:r-x /bruno_rojas/archivo_vacio.txt
```

```
hdfs dfs -getfacl /bruno_rojas/archivo_vacio.txt
```



Verificación de la integridad de los datos

```
cksum /home/dev_r/persona.data
```

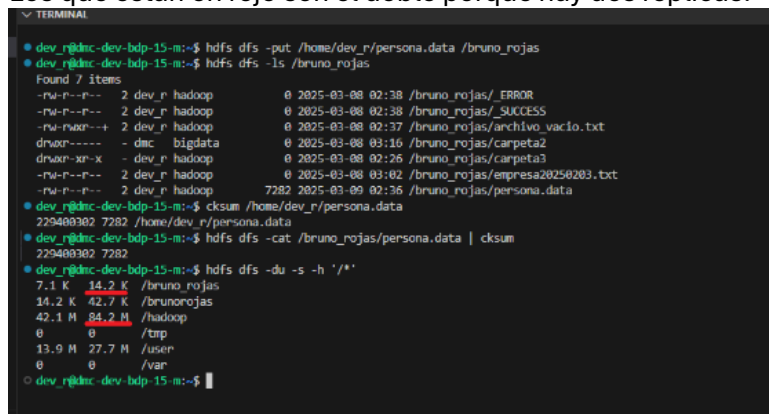
```
hdfs dfs -cat /brunorojas/persona.data | cksum
```

Listando pesos recursivamente

```
hdfs dfs -du -s -h '/*'
```

#14.2 K 28.4 K /brunorojas

Los que están en rojo son el doble porque hay dos réplicas.



Cambiando el número de replicas a 3

`hdfs dfs -setrep -w 3 -R /brunorojas`

Ahora los pesos se multiplan por 3 porque hay tres réplicas.

```
dev_r@dnc-dev-bdp-15-m:~$ hdfs dfs -setrep -w 3 -R /bruno_rojas
setrep: '-R': No such file or directory
Replication 3 set: /bruno_rojas/_ERROR
Replication 3 set: /bruno_rojas/_SUCCESS
Replication 3 set: /bruno_rojas/archivo_vacio.txt
Replication 3 set: /bruno_rojas/empresa20250203.txt
Replication 3 set: /bruno_rojas/persona.data
Waiting for /bruno_rojas/_ERROR ... done
Waiting for /bruno_rojas/_SUCCESS ... done
Waiting for /bruno_rojas/archivo_vacio.txt ... done
Waiting for /bruno_rojas/empresa20250203.txt ... done
Waiting for /bruno_rojas/persona.data .... done
dev_r@dnc-dev-bdp-15-m:~$ hdfs dfs -du -s -h '/'
7.1 K 21.3 K /bruno_rojas
14.2 K 42.7 K /brunorojas
42.1 M 84.2 M /hadoop
0 0 /tmp
13.9 M 27.7 M /user
0 0 /var
dev_r@dnc-dev-bdp-15-m:~$
```