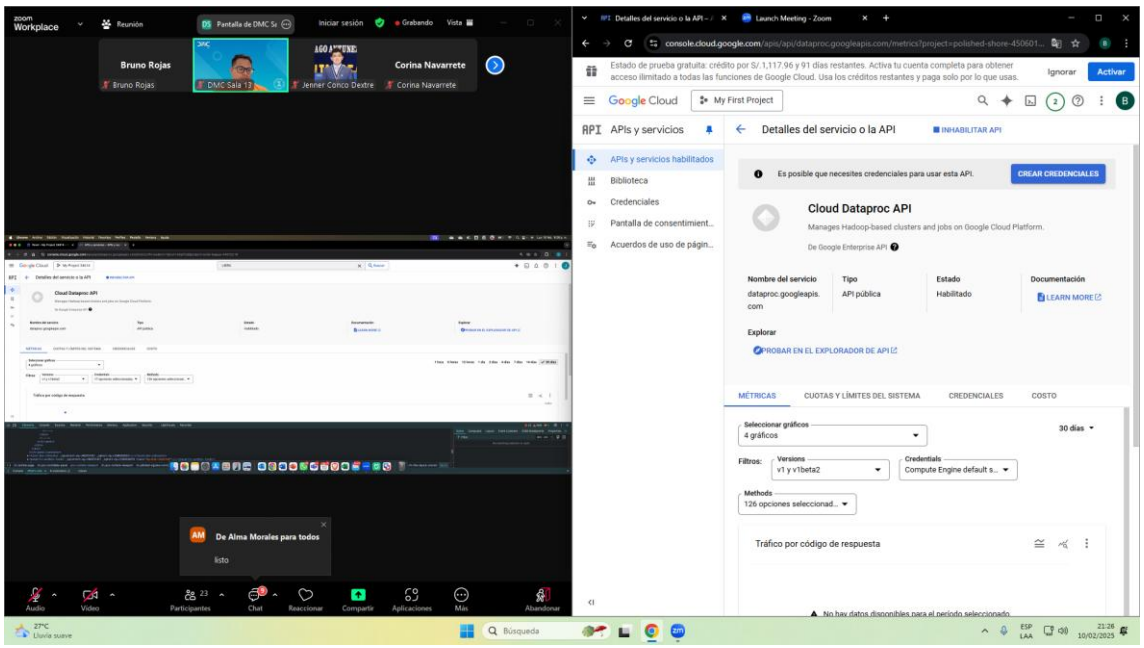
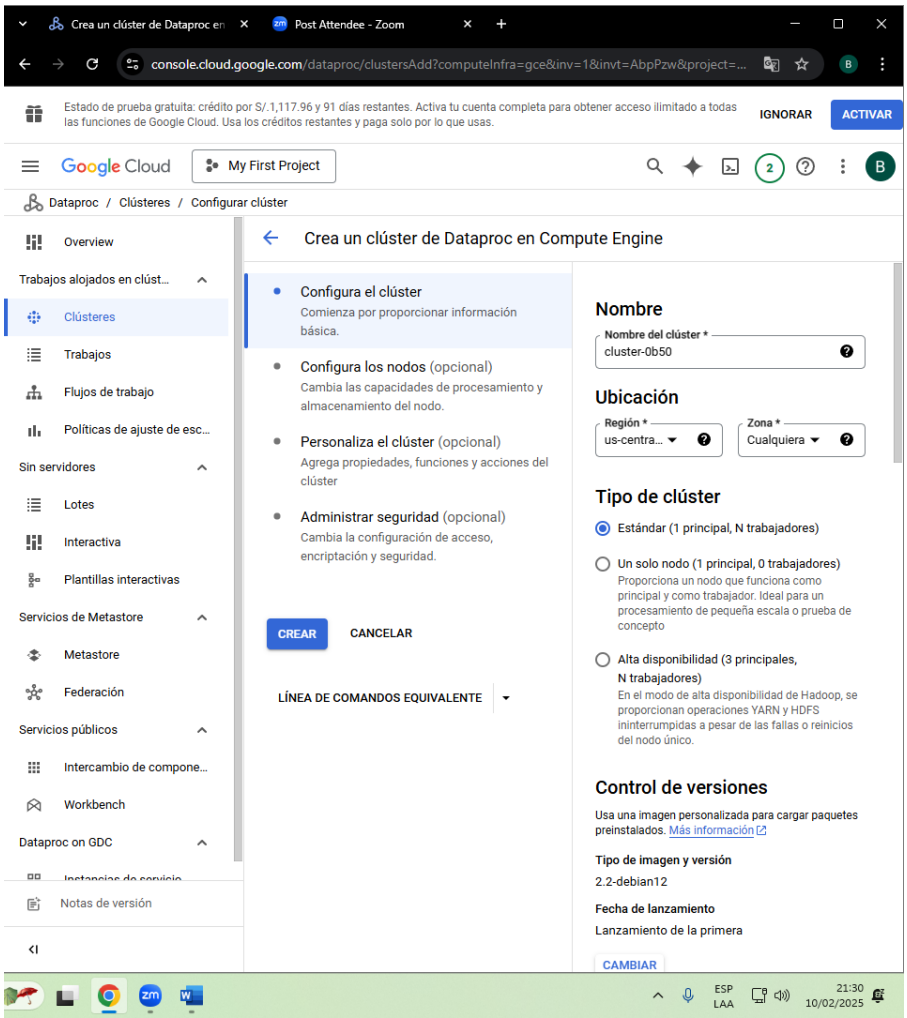
Trabajo Especialización Big Data

https://github.com/BrunoRojasB/curso-dmc-big-data-processing

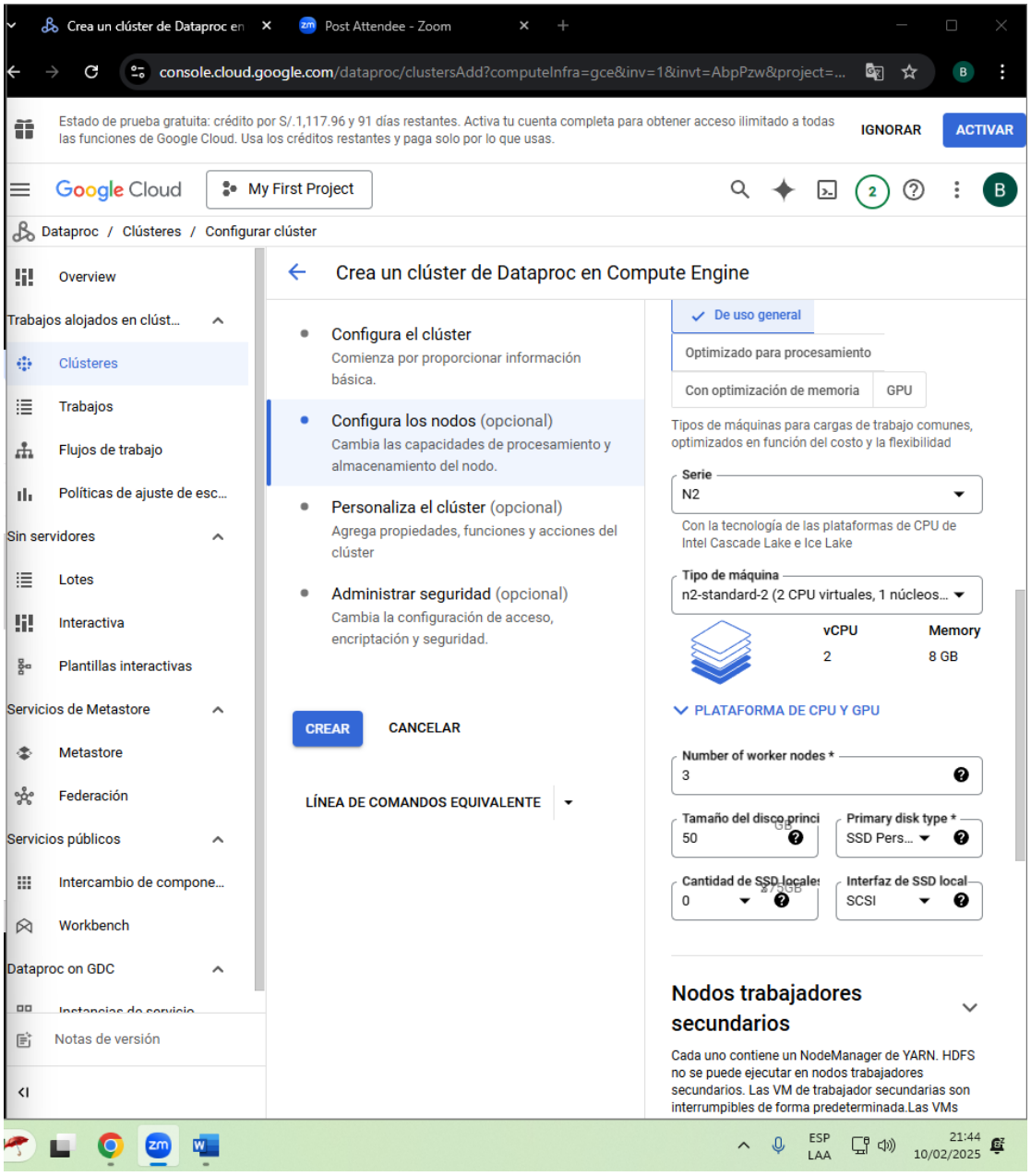
1.- HDFS

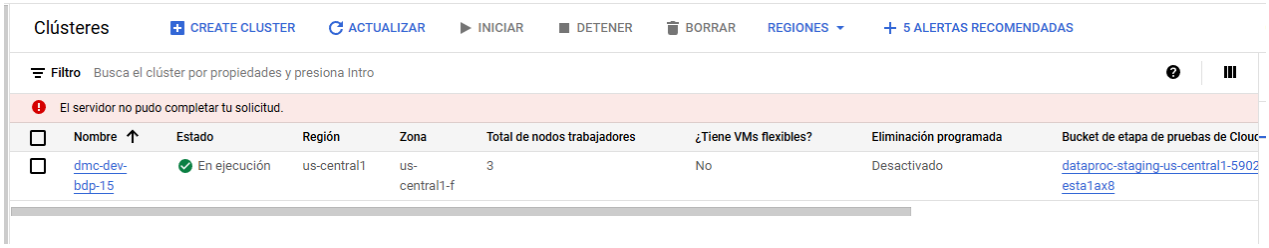
Creación de Clúster Dataproc



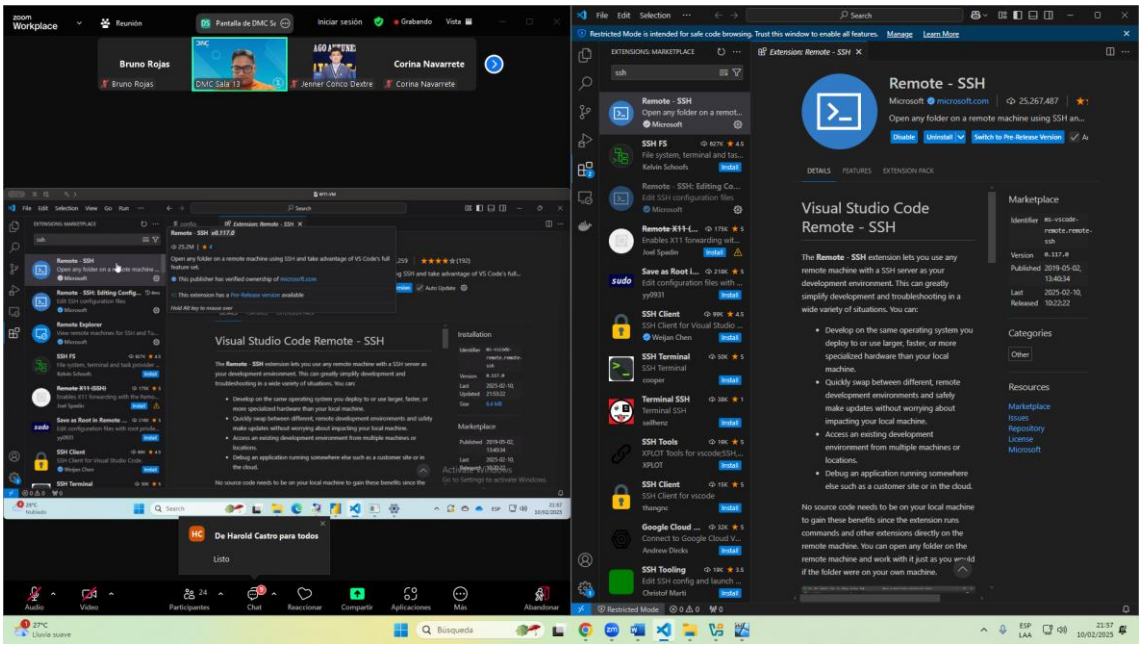


Configuración de Nodos
Tipo de Máquina
Número de nodos trabajadores
Espacio de disco

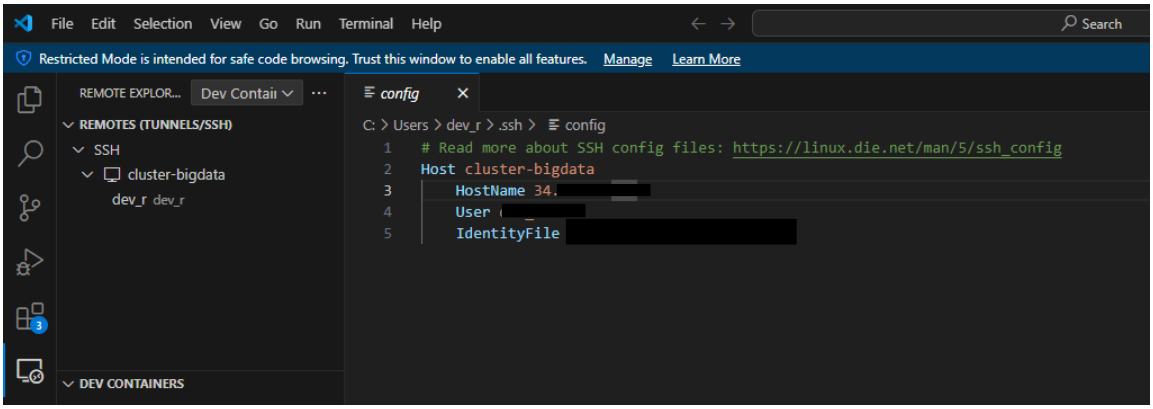




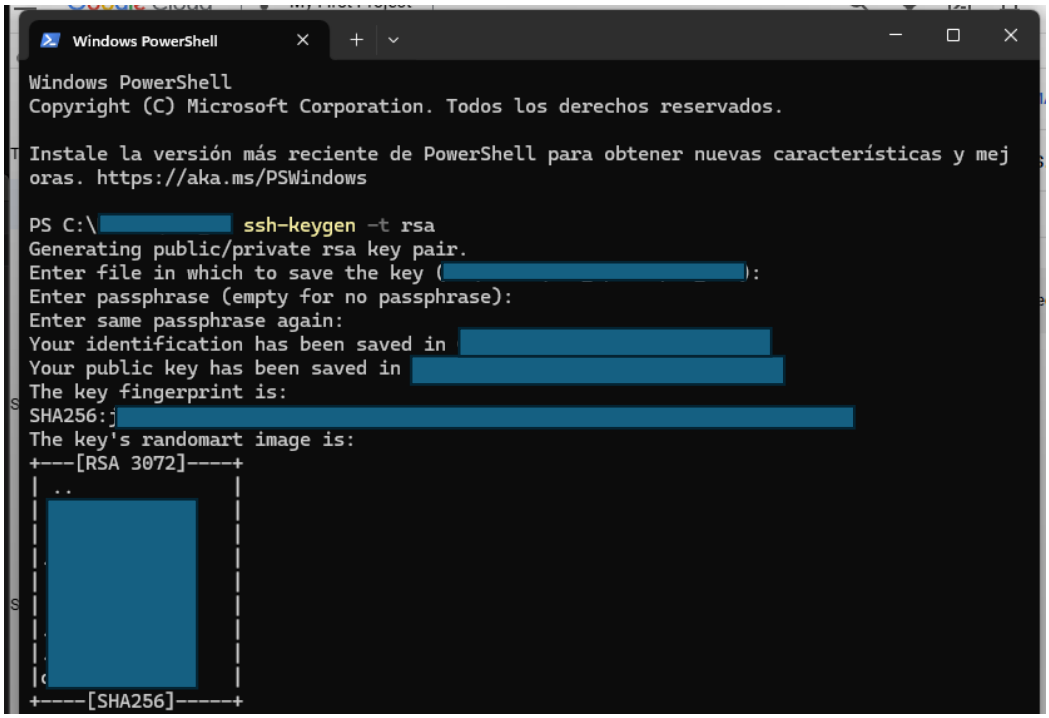
Instalación de Remote Explorer en Visual Studio Code

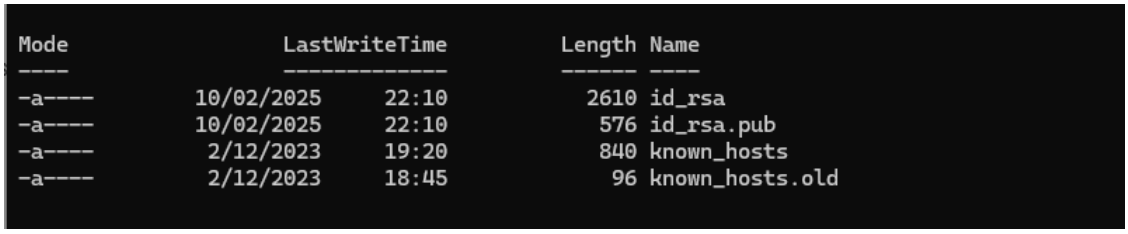


Configuración de Remote Explorer

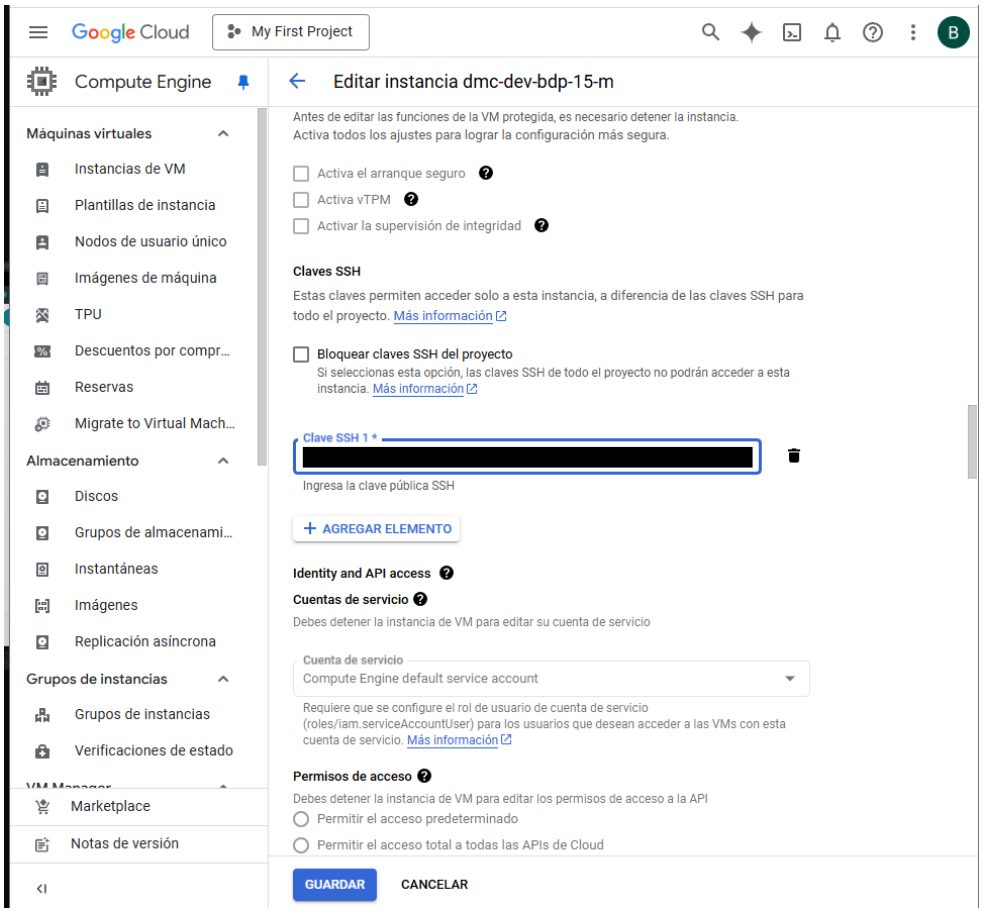


Generación de clave SSH .

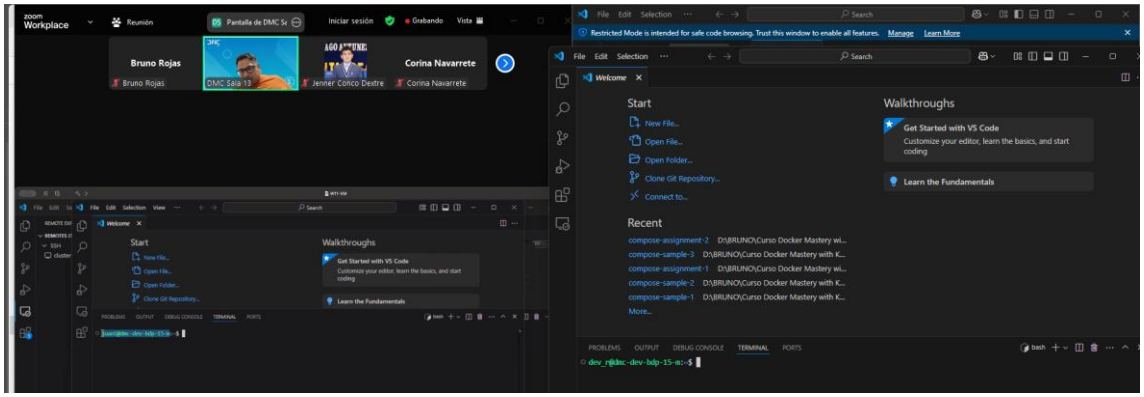




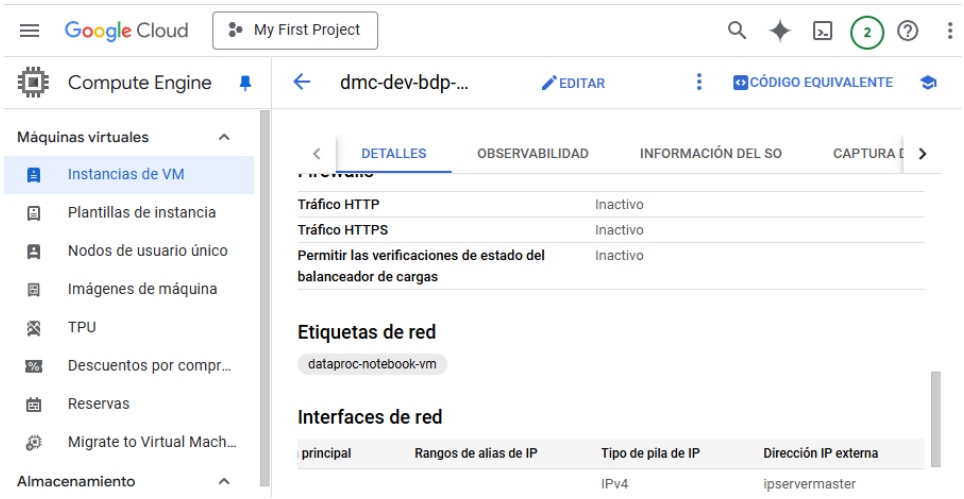
Agregar clave SSH a Dataproc

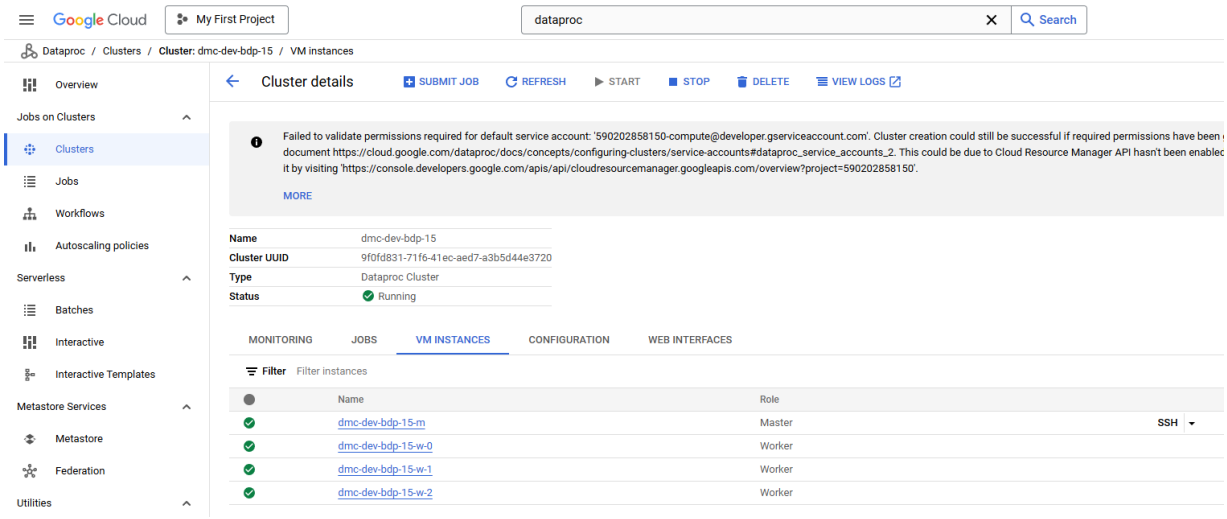


Nos conectamos al cluster desde Visual Studio Code con Remote Explorer

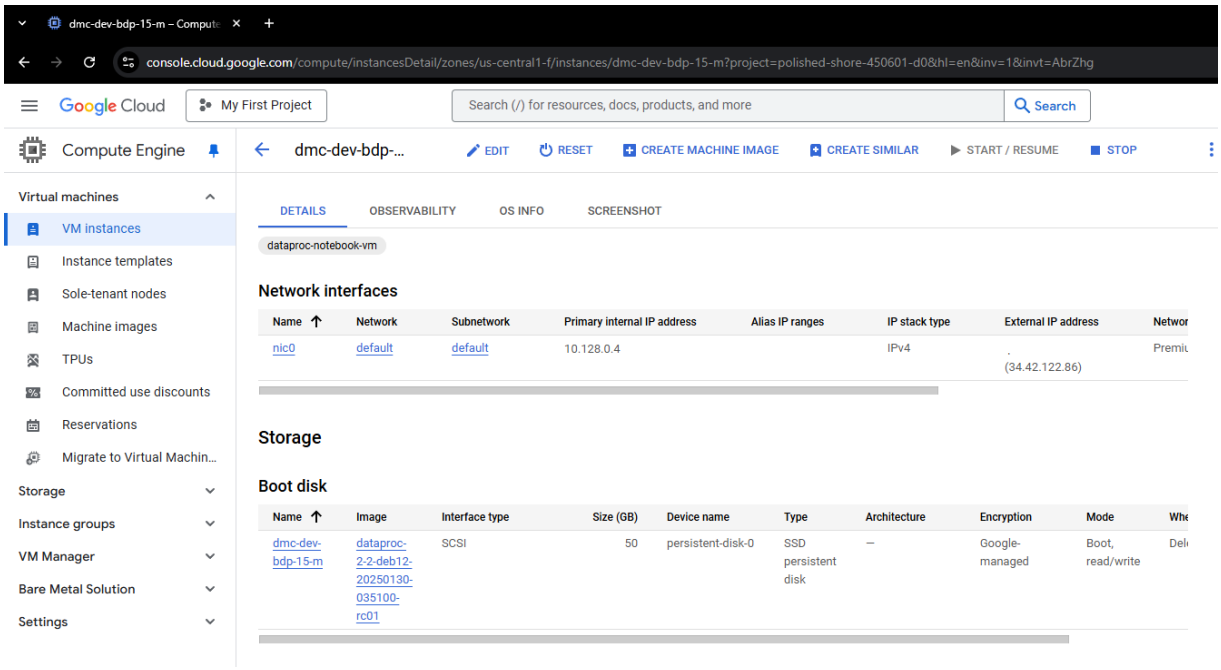


Convertir IP Efímera en Estática o Permanente



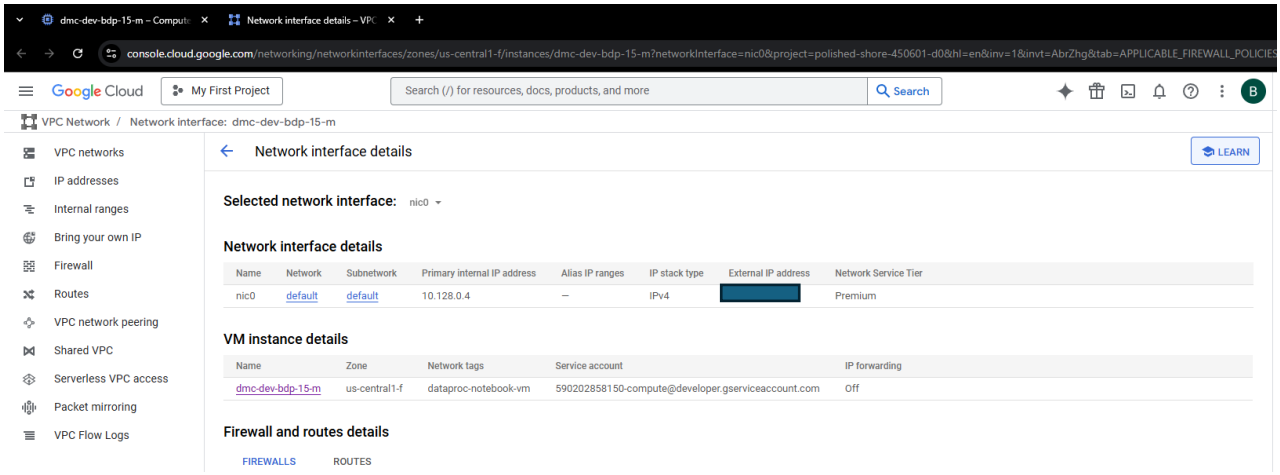


Entramos al master y bajamos hasta network interfaces

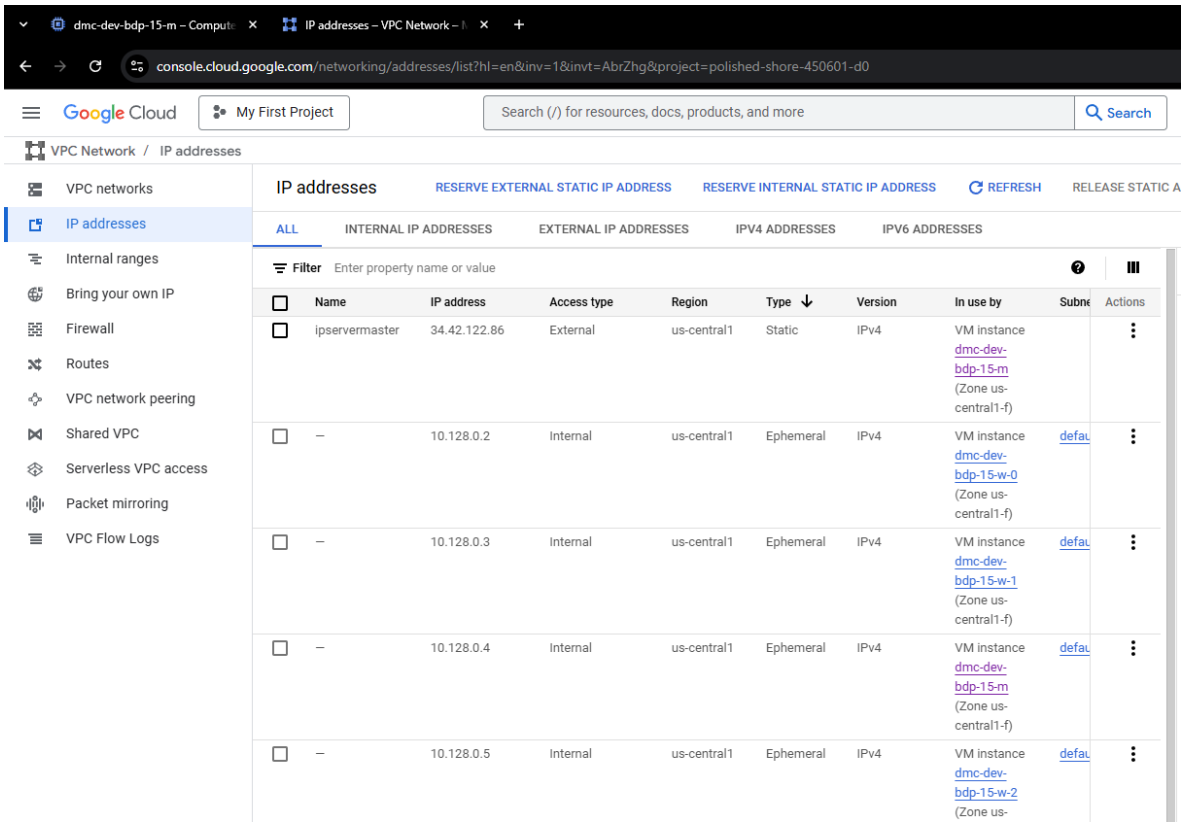


La ip que normalmente aparece como ephimeral o efímera

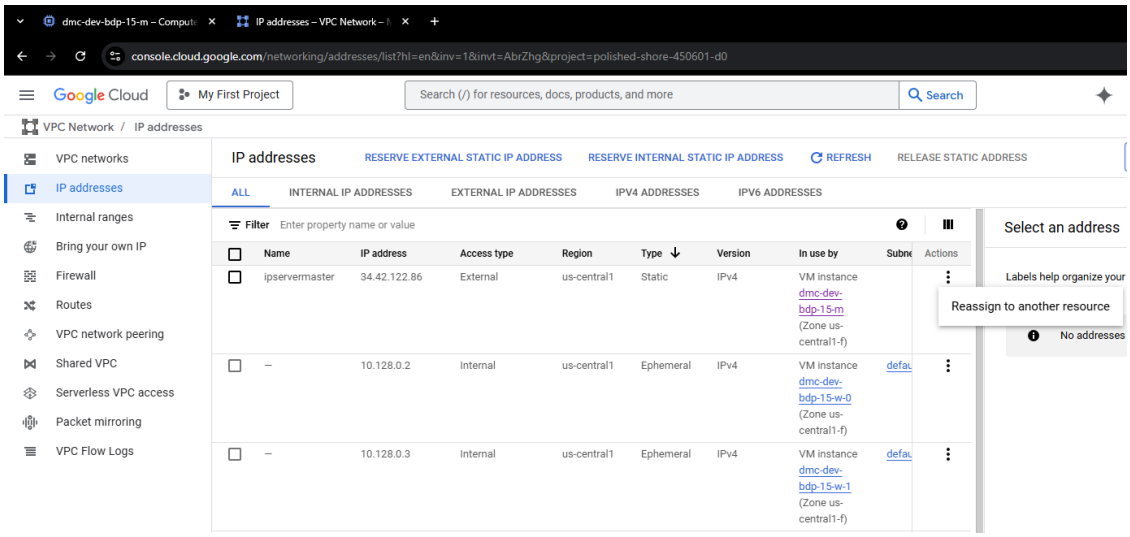
Click derecho sobre nic0 -> Abrir en una nueva ventana



Click IP addresses



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

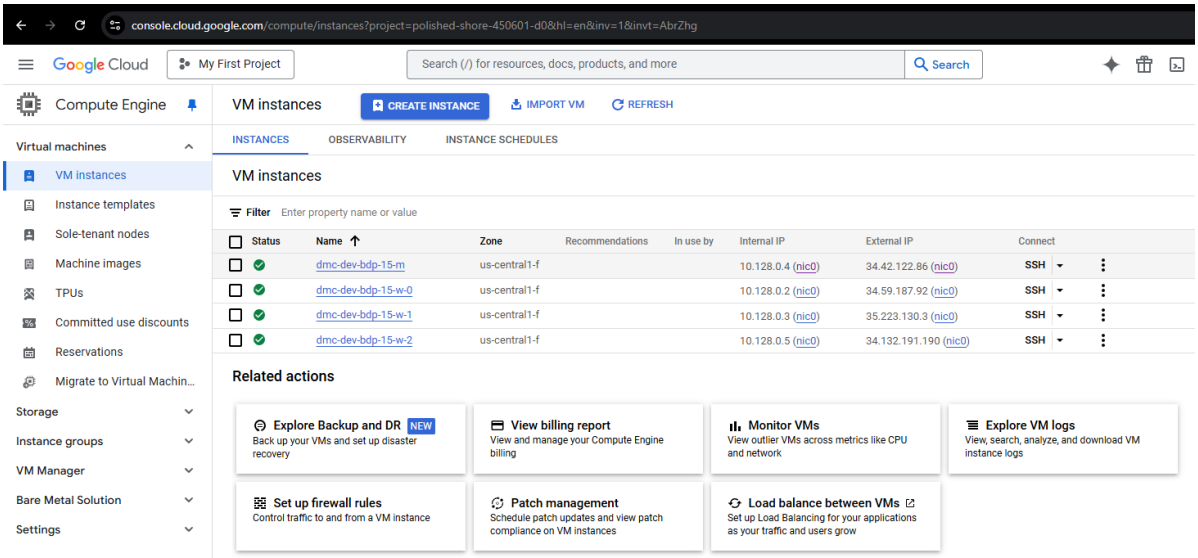


Debería aparecer promover a ip estática

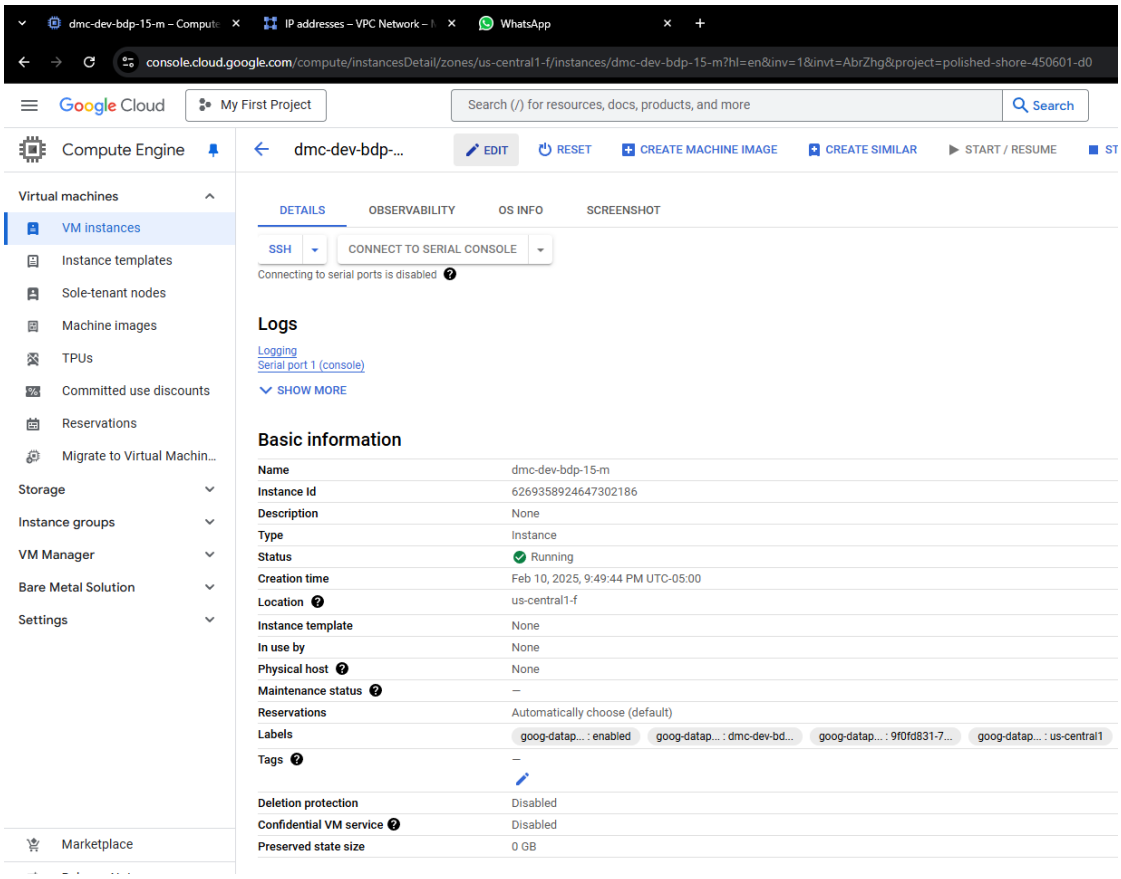
Le colocamos un nombre y descripción

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

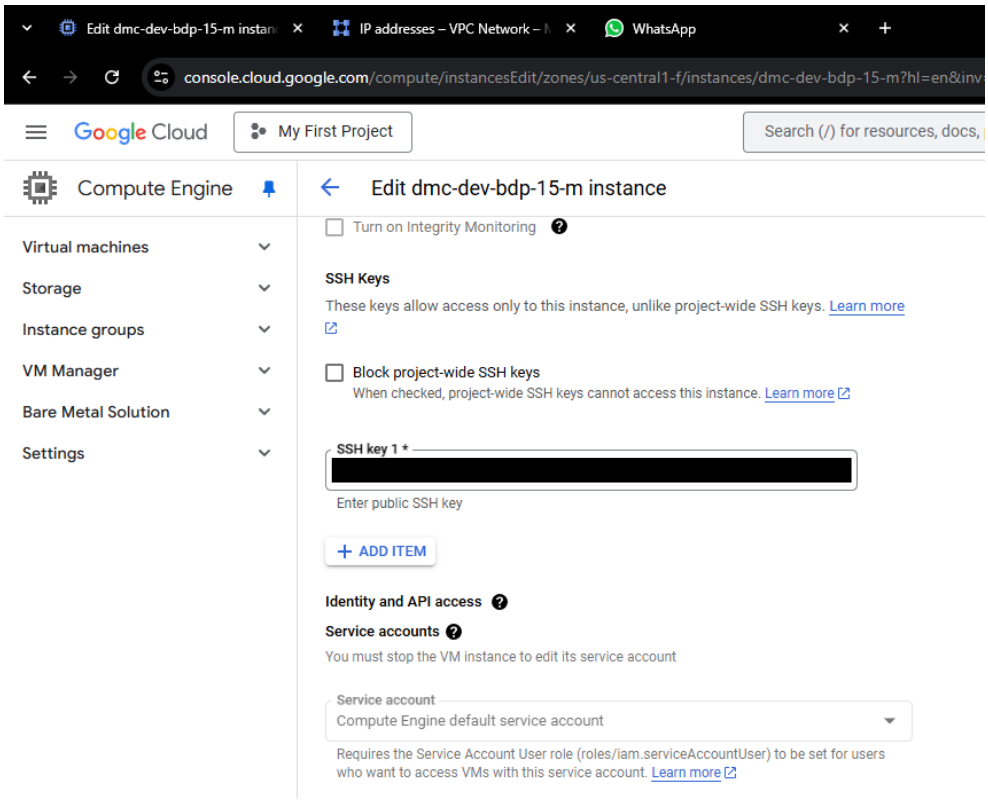
Para añadir una llave ssh



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.



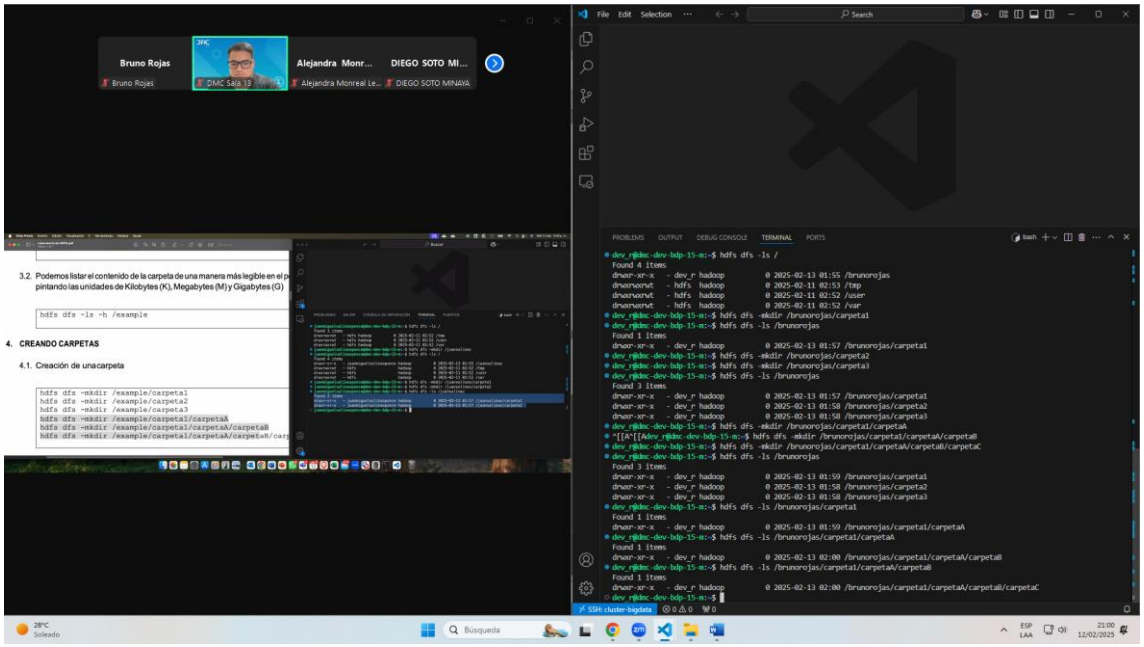
Comandos en HDFS

Listar carpetas en HDFS: hdfs dfs -ls /

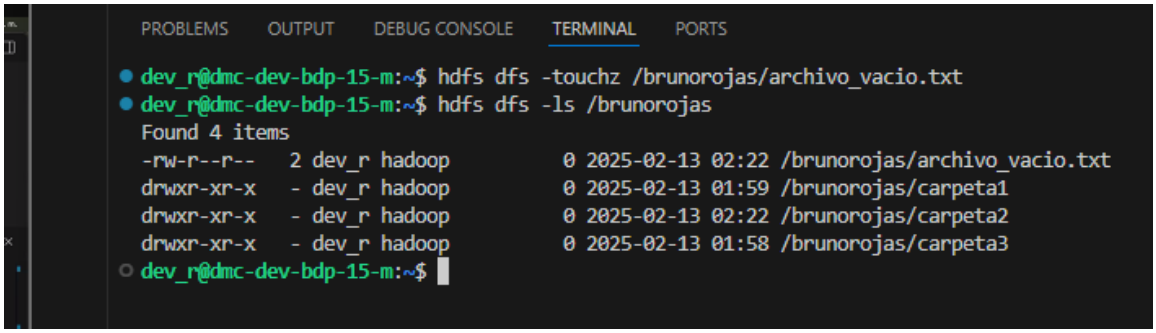
Crear directorios en HDFS: hdfs dfs -mkdir /brunorojas

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

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



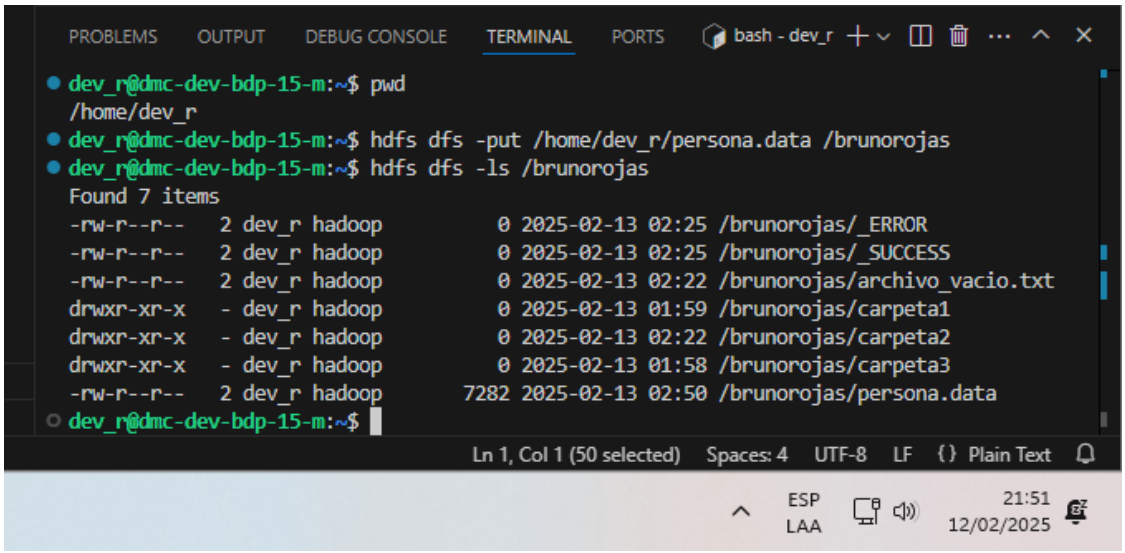
Crear archivo vacío: hdfs dfs -touchz /brunorojas/archivovacio.txt

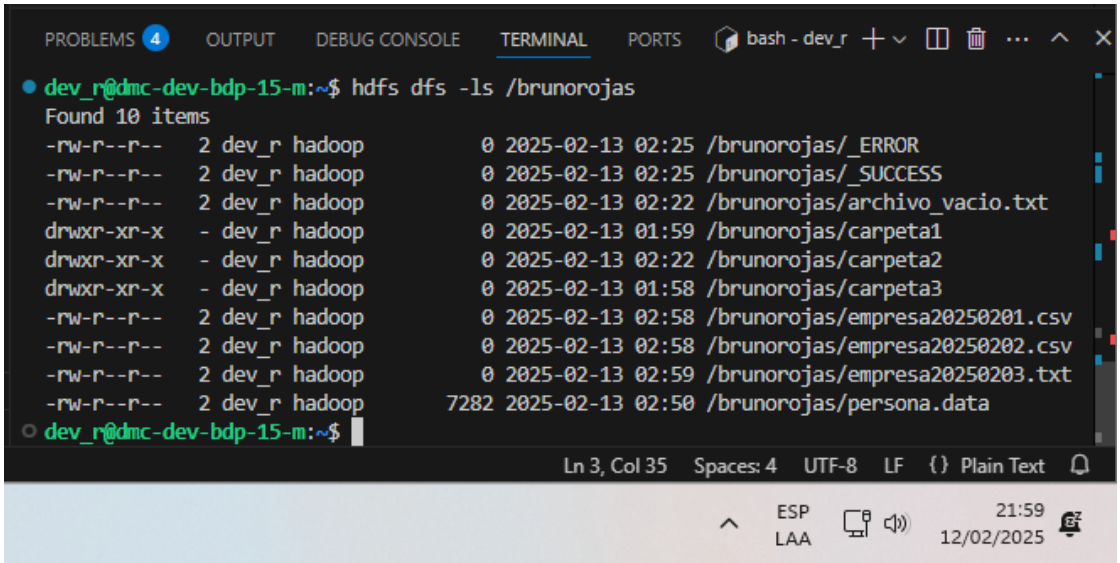


Creamos un archivo txt dentro del servidor para ir guardando nuestros comandos.

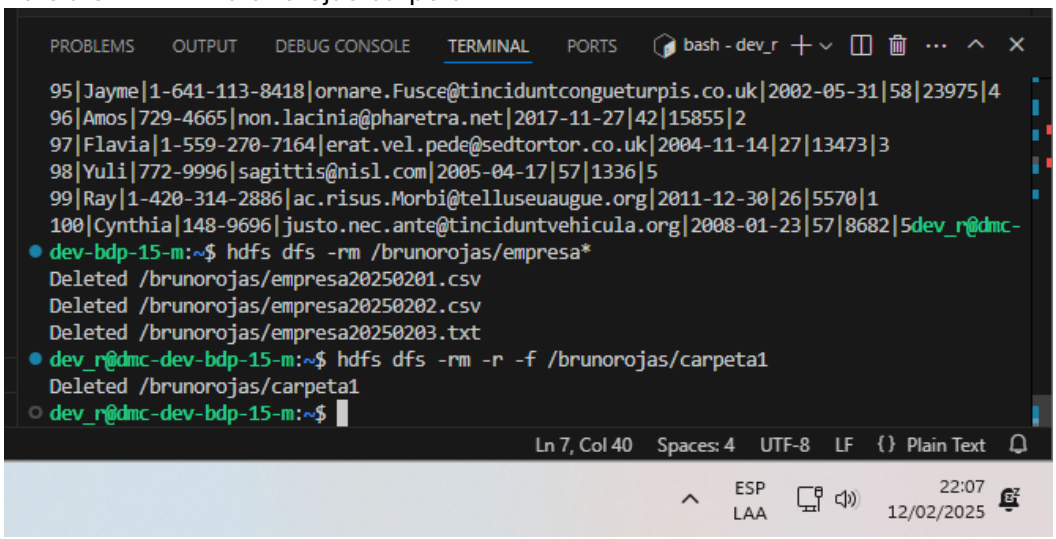
Para llevar un archivo de Linux a hdf se usa el comando put

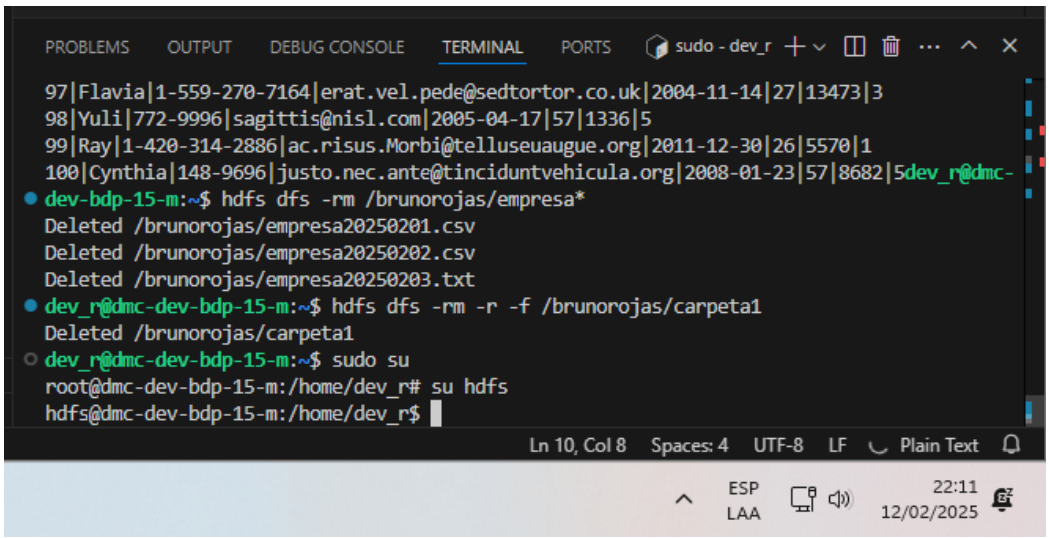
hdfs dfs -put /home/dev_r/persona.data /brunorojas
hdfs dfs -put /home/dev_r/empresa*.csv /brunorojas
hdfs dfs -put /home/dev_r/empresa* /brunorojas





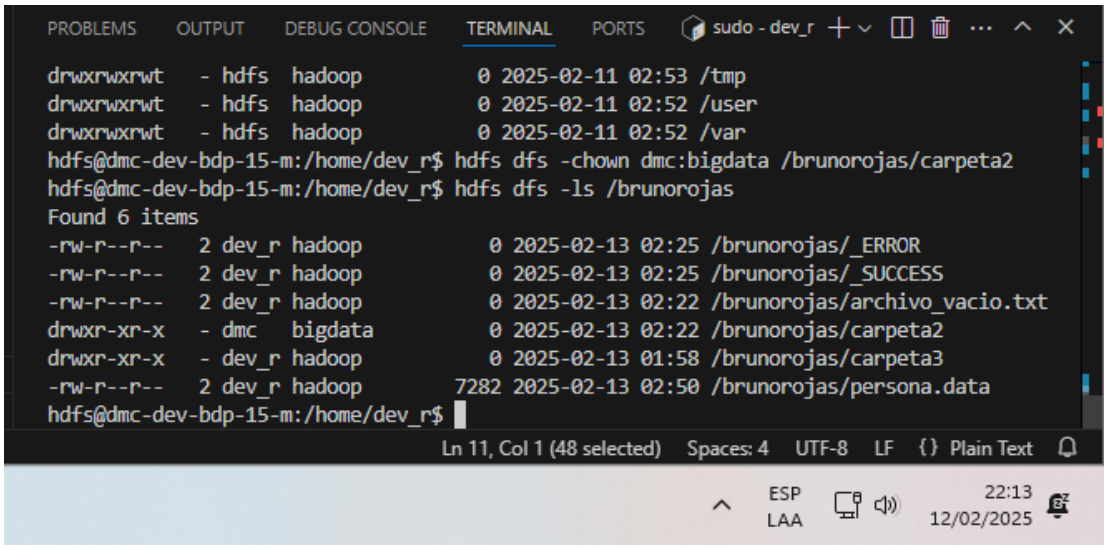
Usamos el comando cat para leer el contenido de un archivo.
hdfs dfs -cat /brunorojas/persona.data
Usamos el commando rm para eliminar un archivo. La f es para forzar el borrado.
hdfs dfs -rm -f /brunorojas/persona.data
Se puede eliminar usando comodines también.
hdfs dfs -rm /brunorojas/empresa*
Si 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

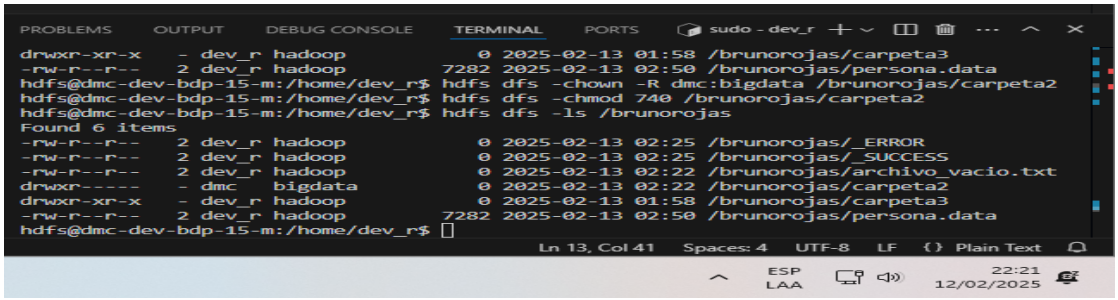




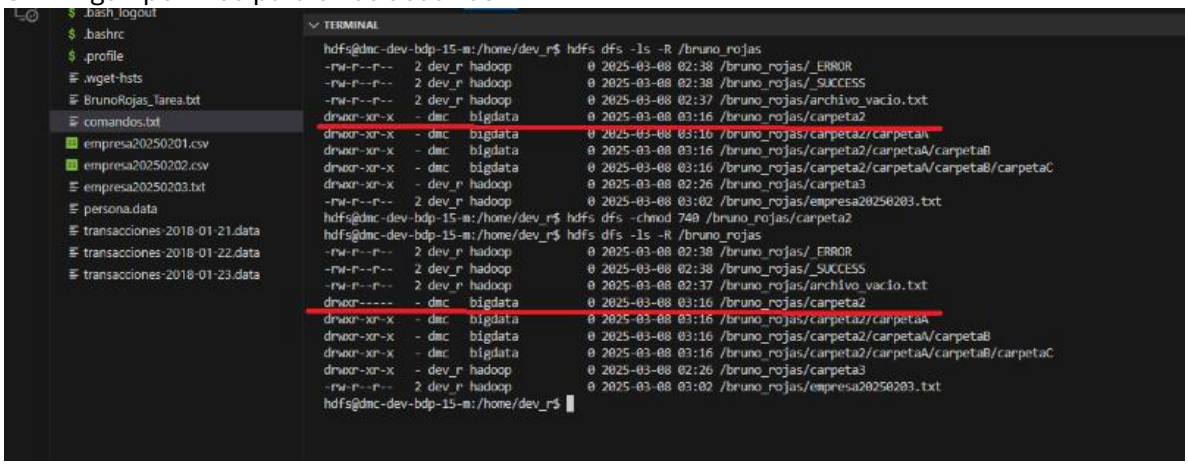
Para poder realizar el cambio de permisos de ejecución, escritura y lectura así como del propietario de un archivo o carpeta se debe realizar con el superusuario.

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



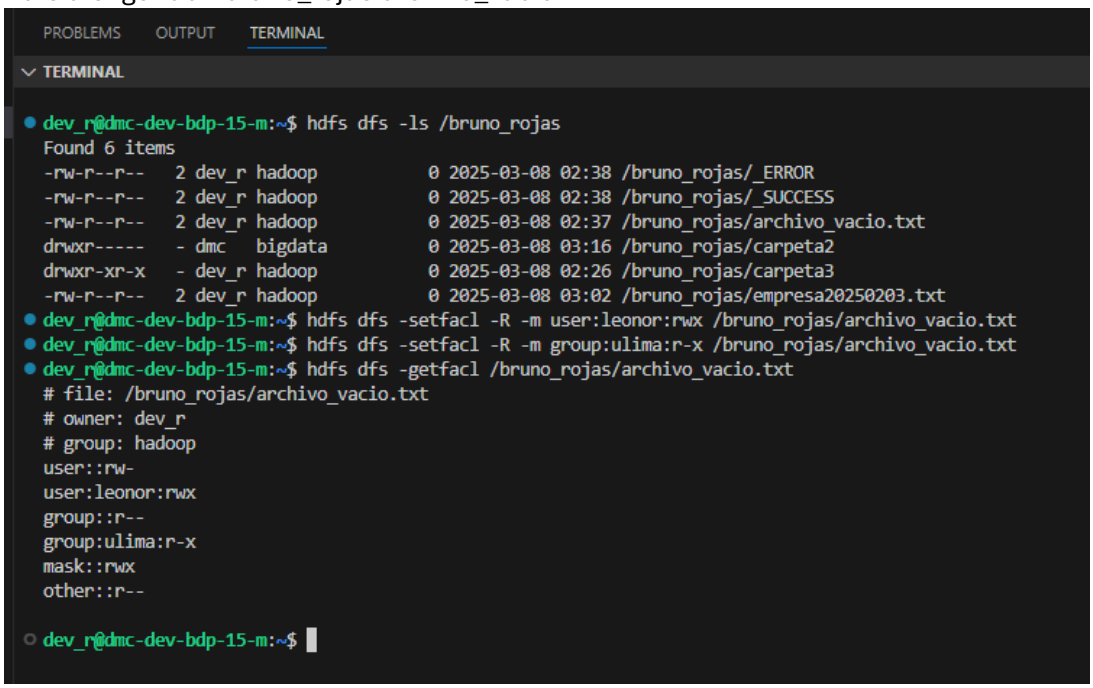


Para asignar permisos de lectura, escritura y ejecución se usa chmod.
hdfs dfs -chmod 740 /bruno_rojas/carpeta2
Permisos de lectura escritura y ejecución para DMC
Permisos de lectura para el grupo bigdata
Sin ningún permiso para otros usuarios

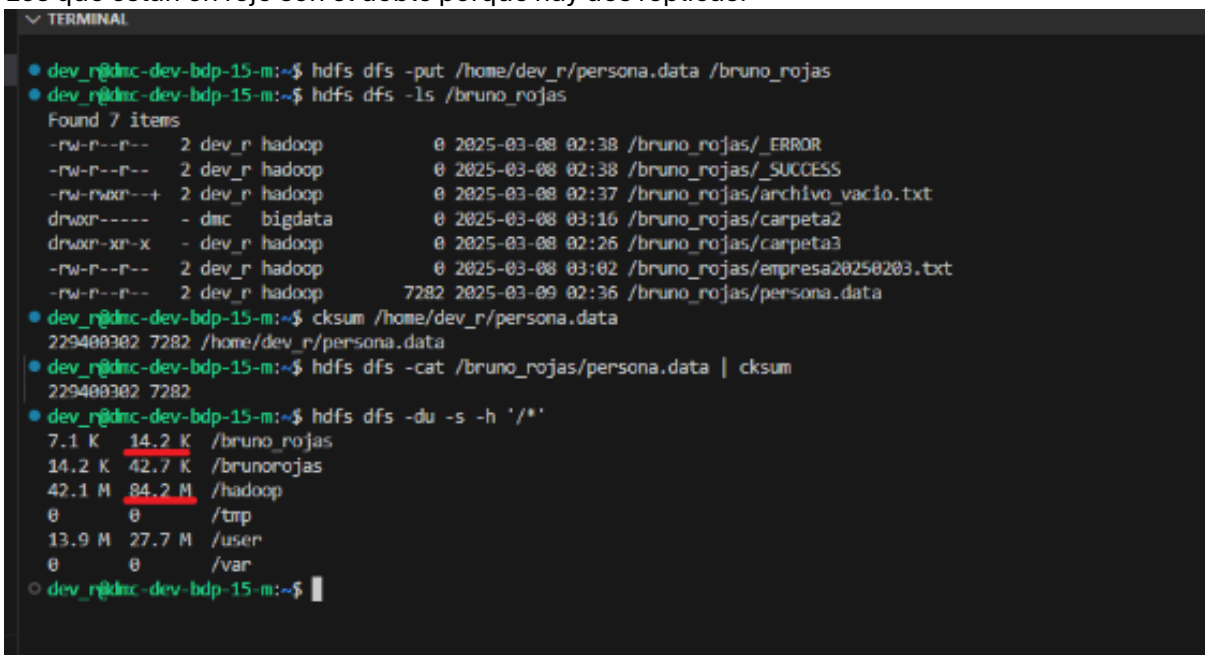


Permisos con ACL

Estructura del comando setfacl: hdfs dfs -setfacl -R -m user:nombreusuario:wwx /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.