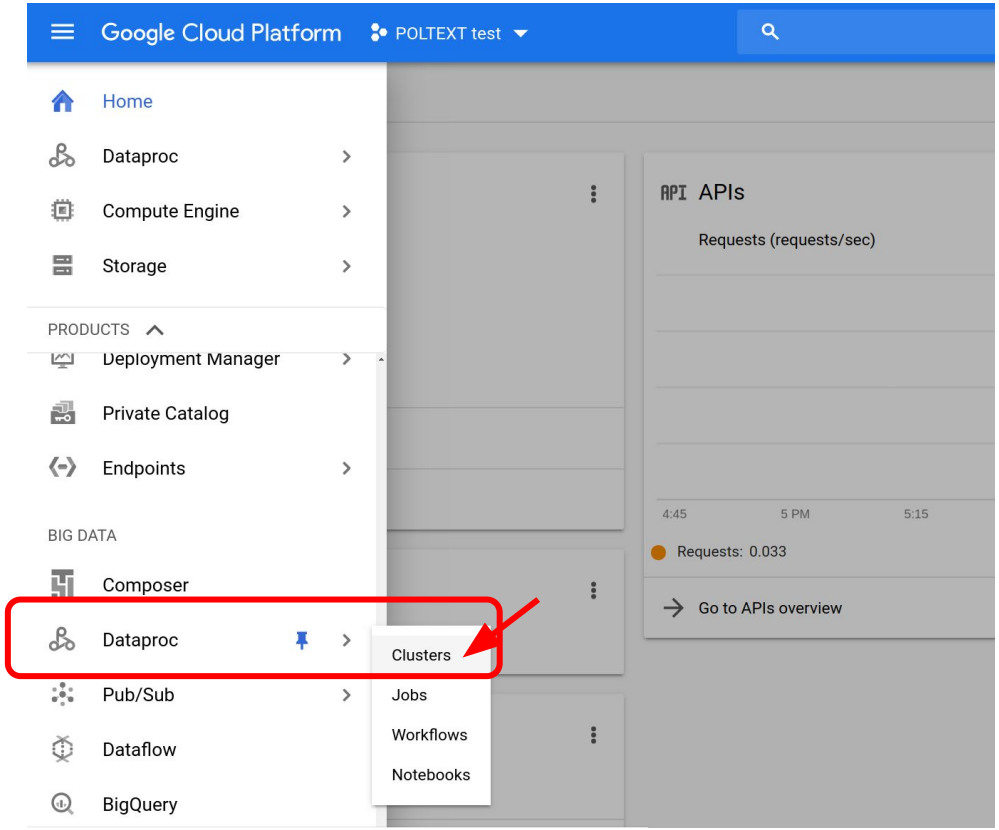
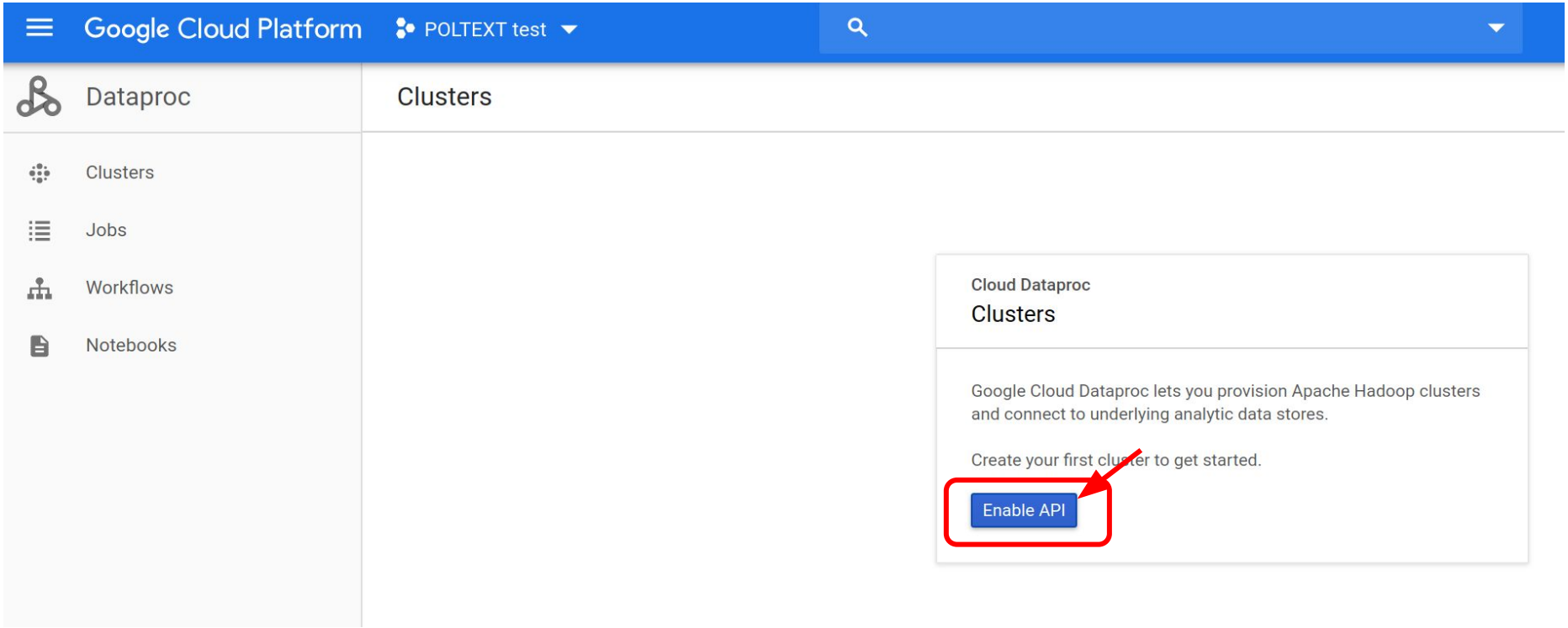
Create a Spark cluster in Google Cloud with RStudio Server

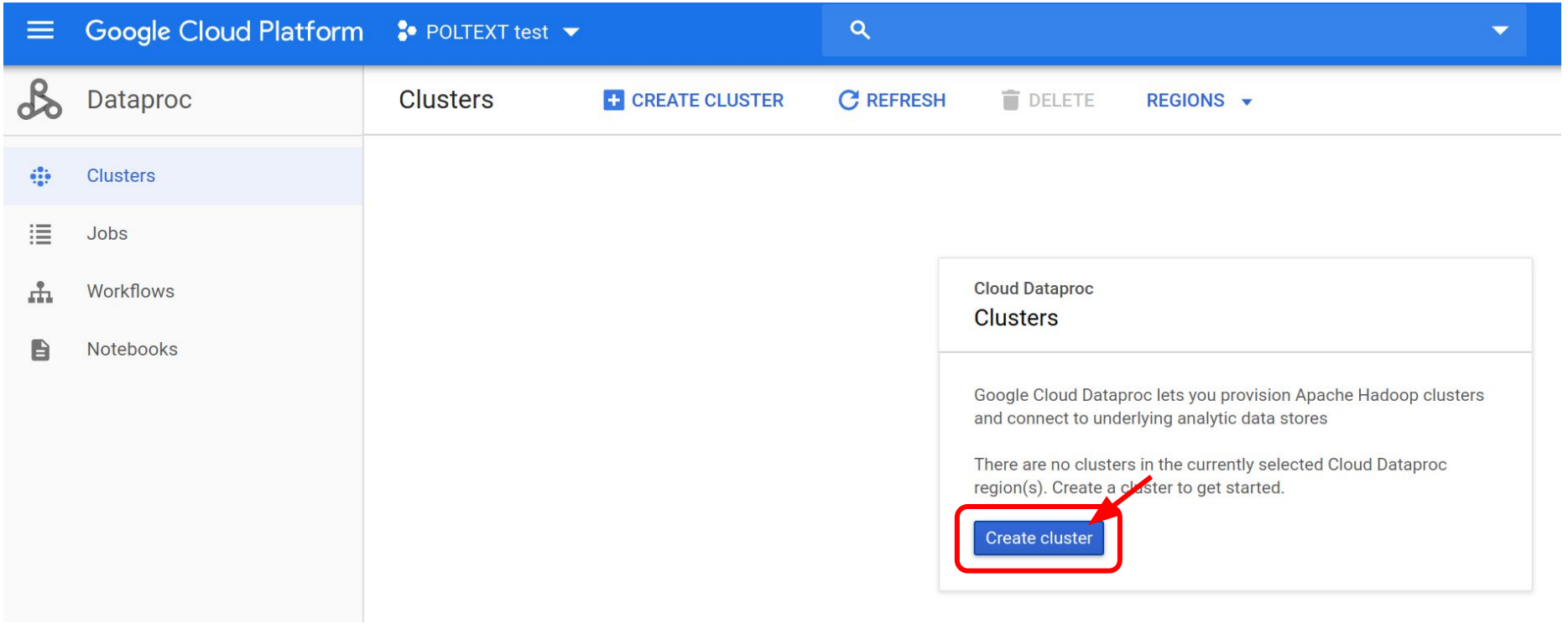
An illustrated guide to
https://cloud.google.com/solutions/running-rstudio-server-on-a-cloud-dataproc-cluster
with some modifications



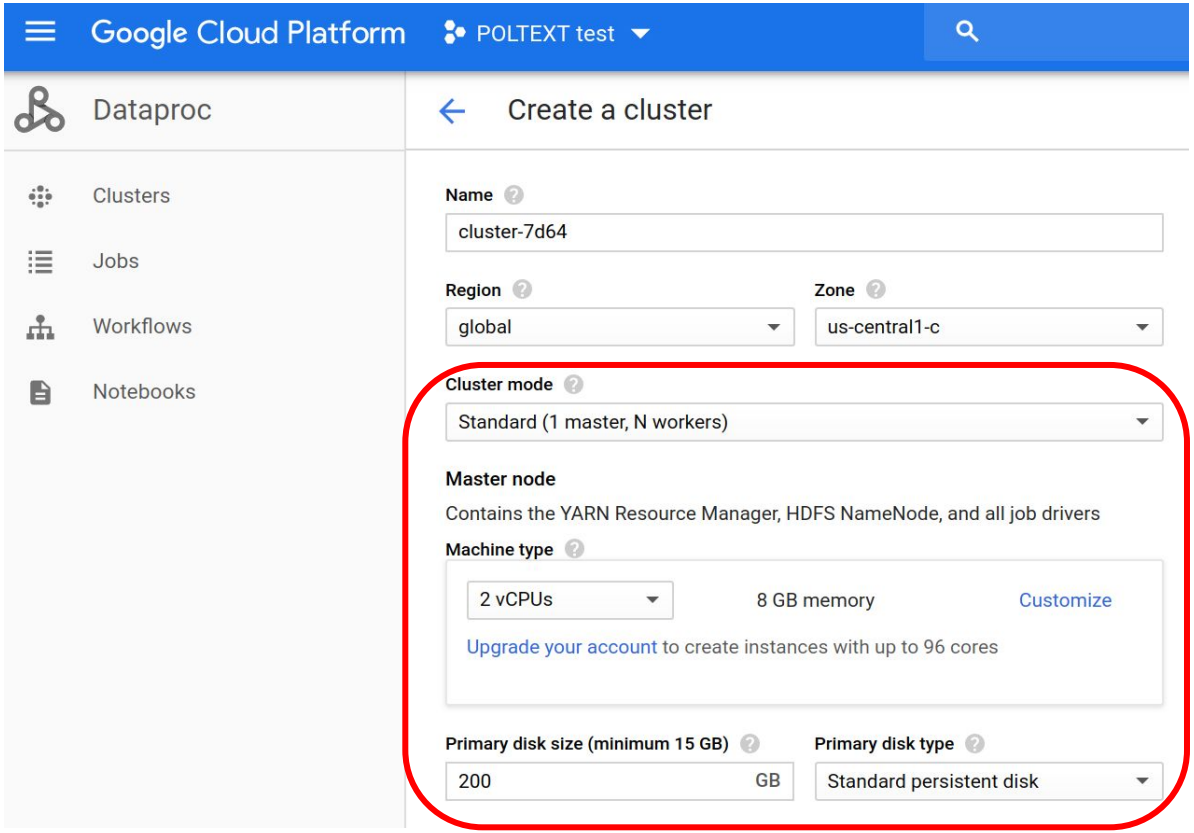
Select Dataproc under BIG DATA and choose Clusters



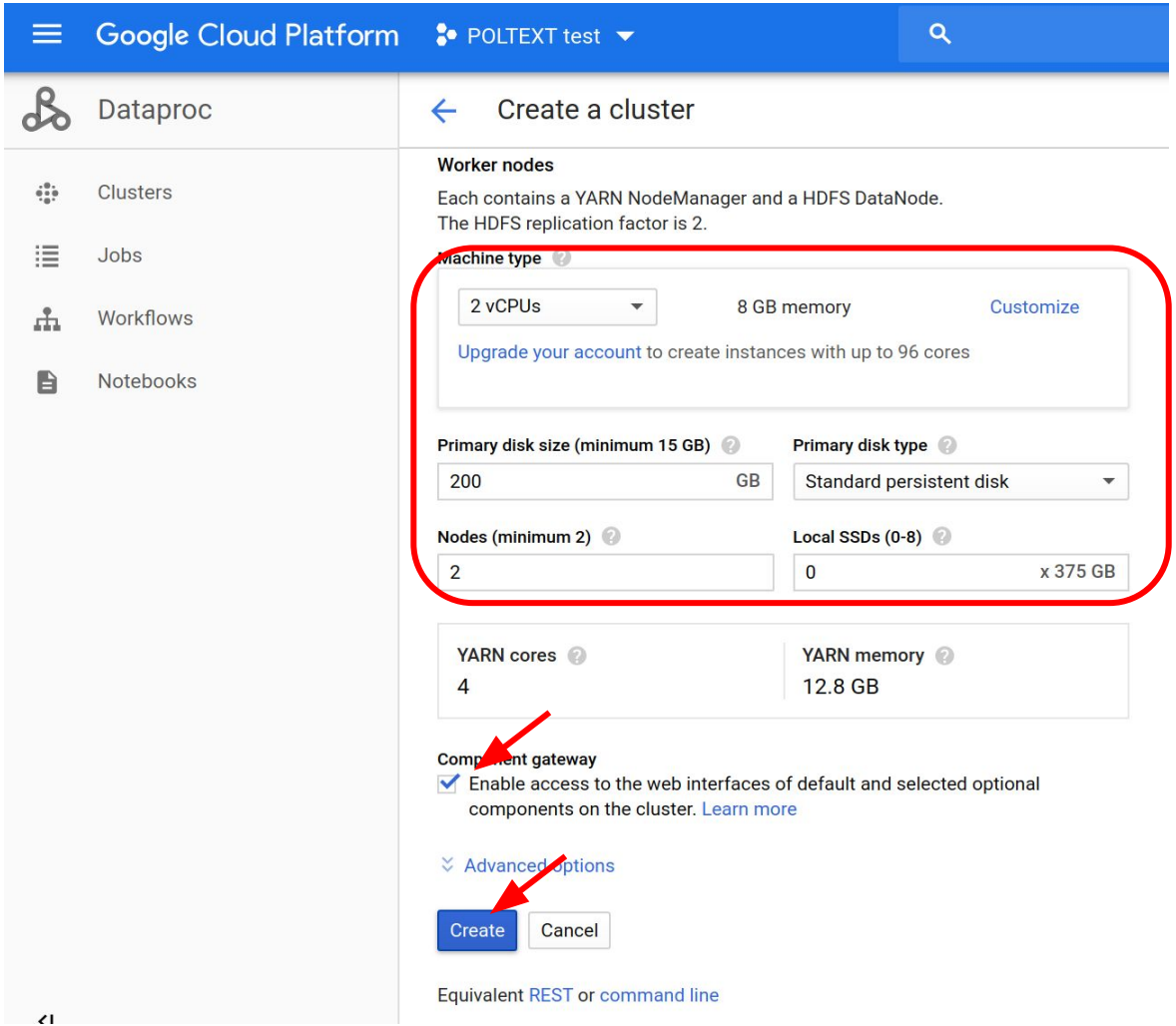
Enable the Dataproc API



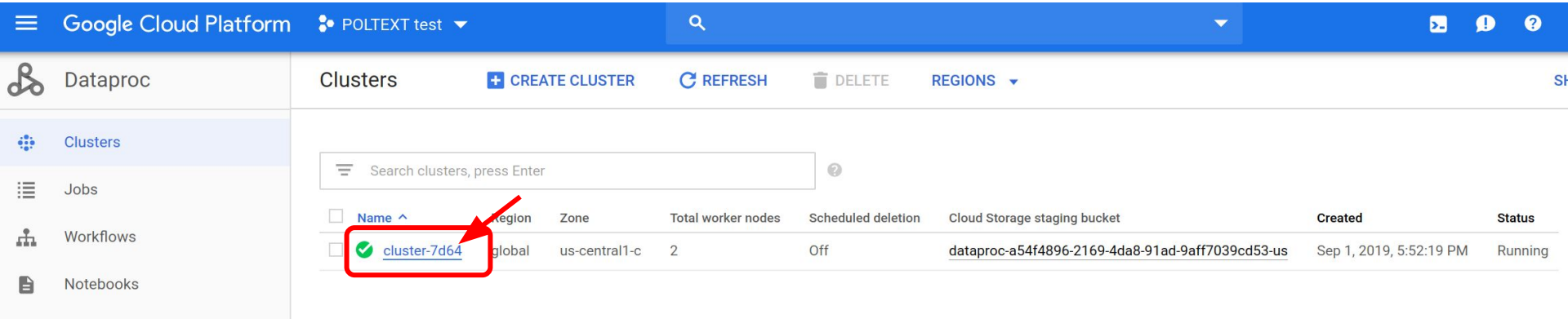
Start creating a cluster



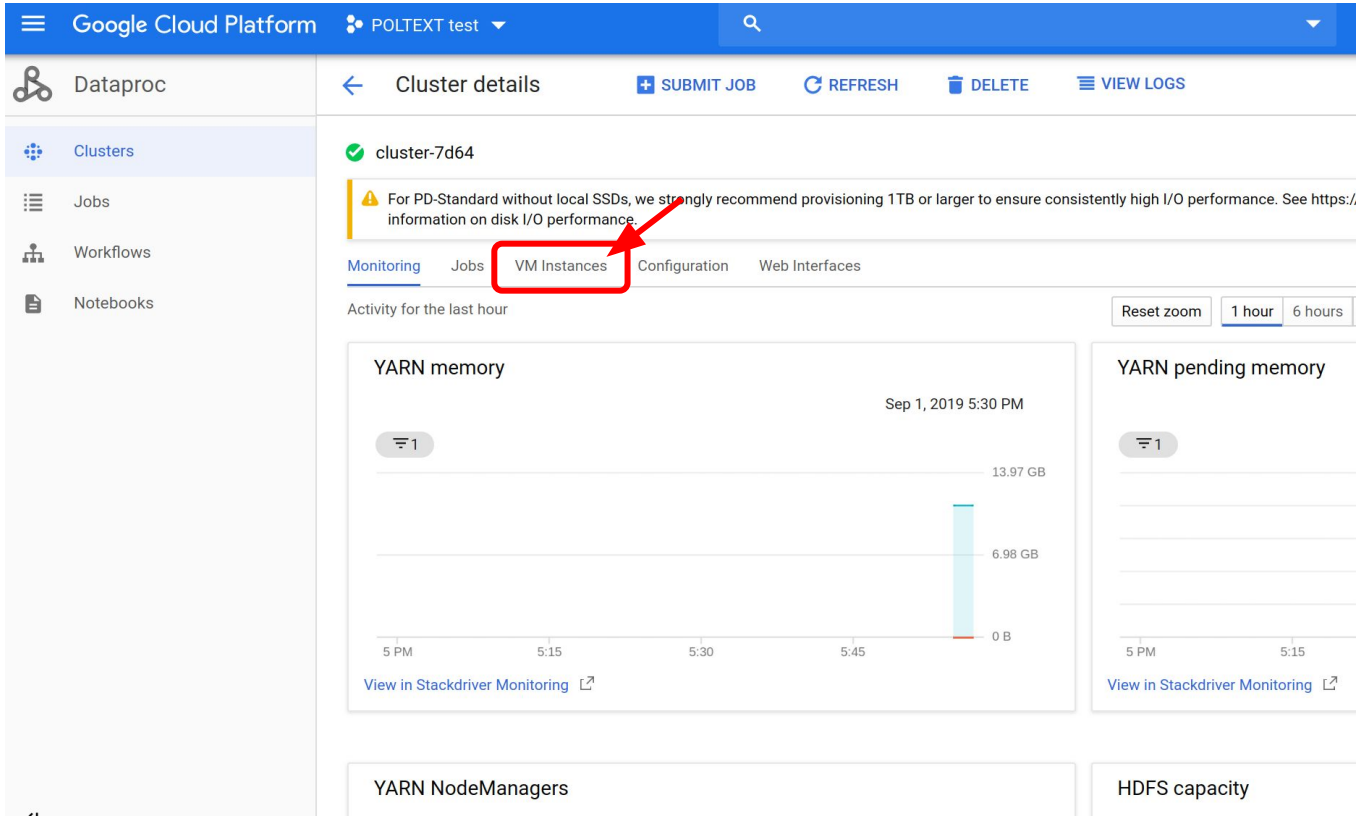
Master node settings (for this tutorial we can just accept name, region and zone)



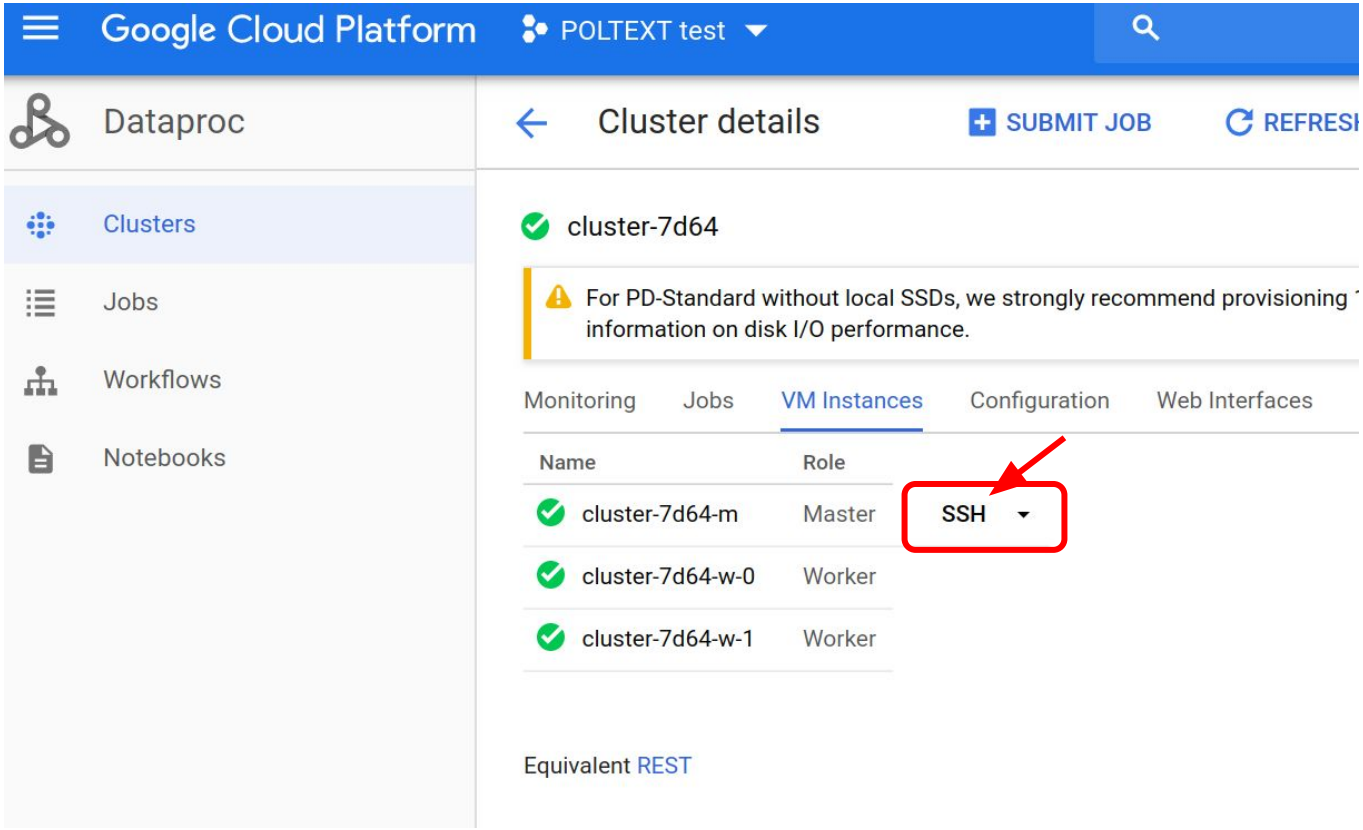
After adjusting worker node settings, check the Component gateway checkbox, and then press Create



Once the cluster is up and running, press the cluster name to go to the cluster management page



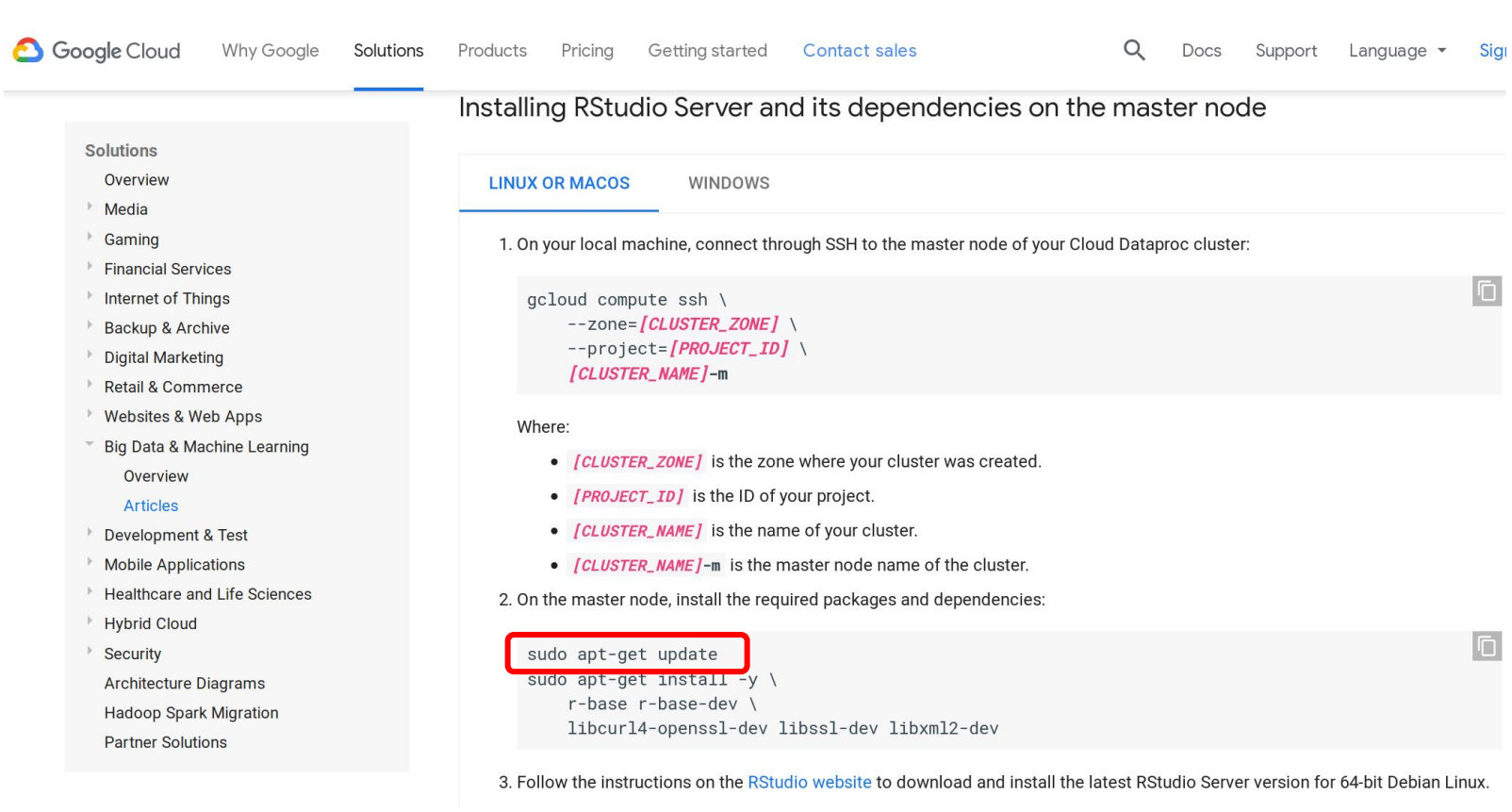
We are now inspecting different aspects of our cluster on the cluster management page, press VM Instances to see the virtual machines that make up our cluster



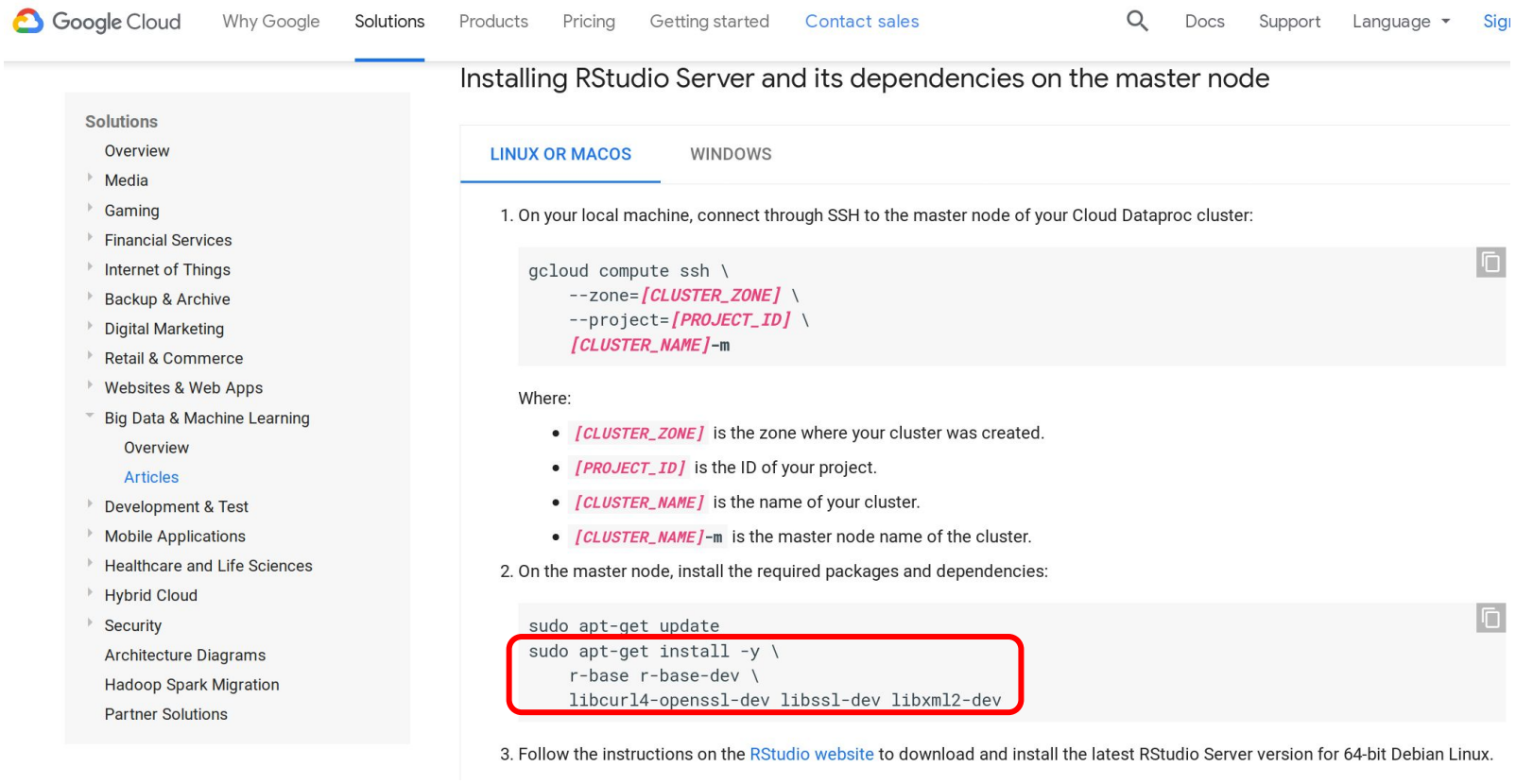
Open the SSH terminal to the master node by pressing the SSH button

You can follow the individual steps on the following slides or just copy the following line to the ssh terminal and press enter, in the latter case skip the next four slides and continue from adding the new user by choosing a password for the new user in the ssh terminal

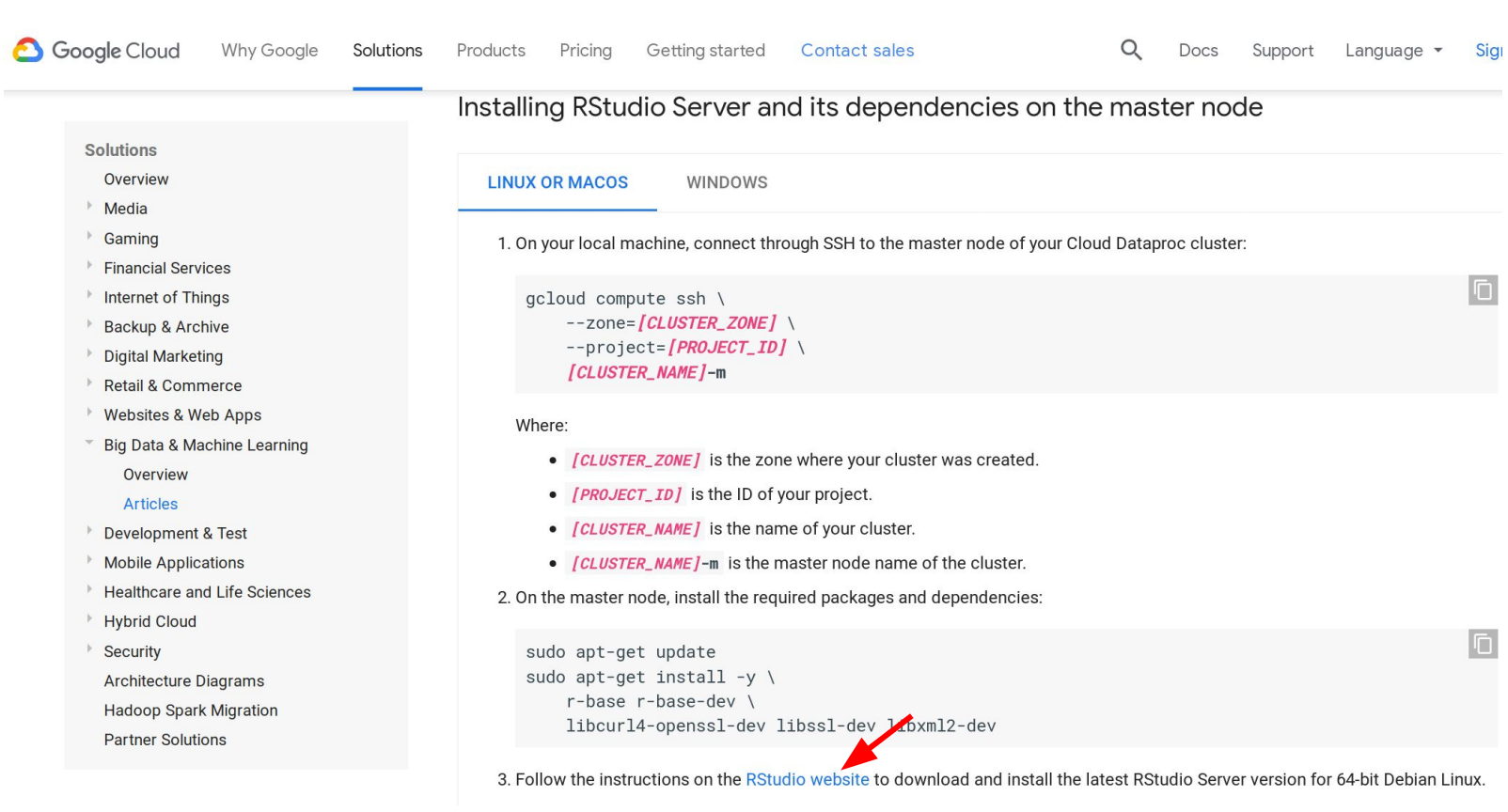
```
sudo apt-get update && sudo apt-get install -y r-base r-base-dev libcurl4-openssl-dev libssl-dev libxml2-dev gdebi-core && wget https://download2.rstudio.org/server/debian9/x86_64/rstudio-server-1.2.1335-amd64.deb && sudo gdebi --n rstudio-server-1.2.1335-amd64.deb && sudo su - -c "R -e \"install.packages('sparklyr', repos = 'http://cran.us.r-project.org')\"" && sudo adduser test
```



This is still https://cloud.google.com/solutions/running-rstudio-server-on-a-cloud-dataproc-cluster
Copy-paste or type sudo apt-get update into the SSH terminal (ctrl-v works) and press enter

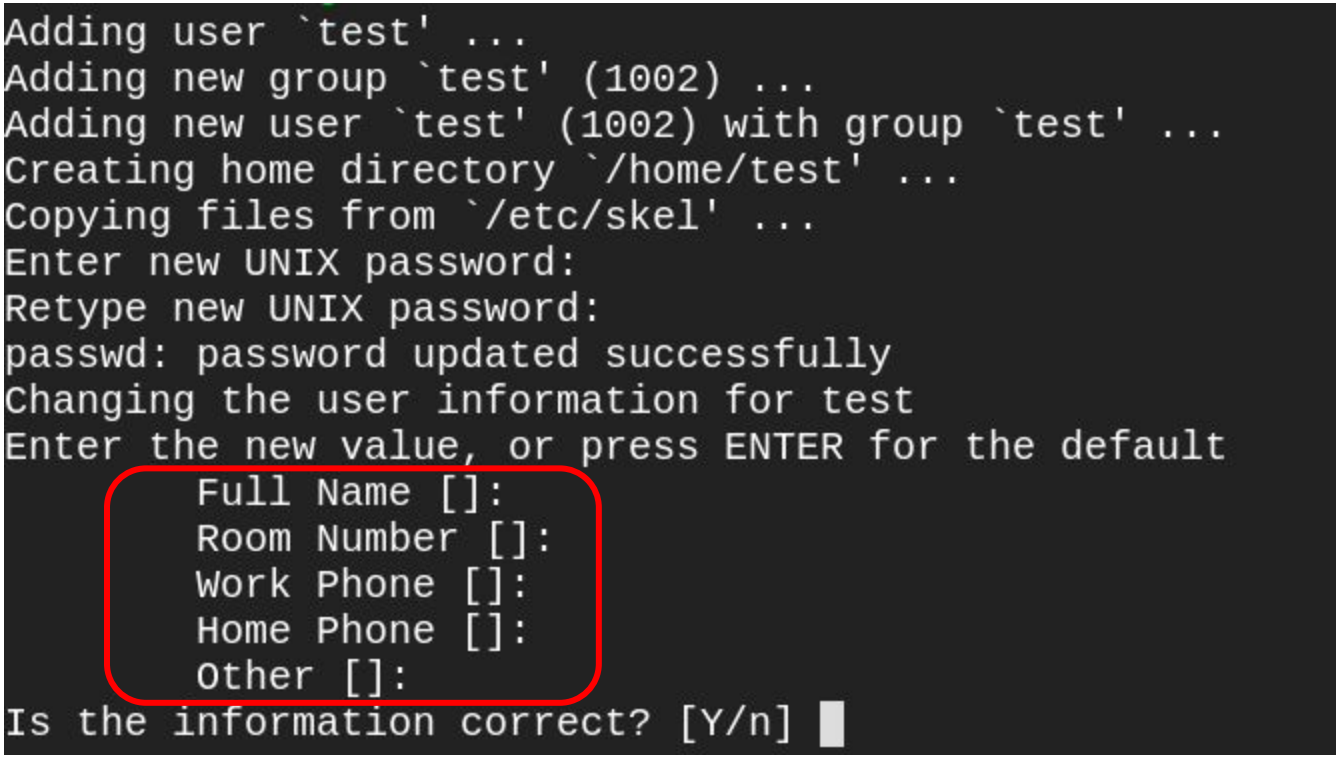


This is still https://cloud.google.com/solutions/running-rstudio-server-on-a-cloud-dataproc-cluster
Copy-paste the second command into the SSH terminal (ctrl-v works) and press enter



This is still https://cloud.google.com/solutions/running-rstudio-server-on-a-cloud-dataproc-cluster
Click the link to go to the RStudio Server install instructions website

This is https://www.rstudio.com/products/rstudio/download-server/
Scroll down to Debian 9, and copy and execute the three commands one by one in the SSH terminal like we did before (NOTE: don't copy the $ symbol at the beginning of the commands)



We need to create a user to access the RStudio Server web interface.
Type or copy from the instructions page into the SSH terminal: sudo adduser test
This will create the user “test”. Choose a password and leave the information blank (just press enter) and select y to finish the process.

We need to use the Google Cloud SDK (gcloud command line tool) for the next step. If you don’t have the SDK installed yet, follow the instructions for your OS here: https://cloud.google.com/sdk/docs/quickstarts

- To load the HDFS NameNode web UI, connect your specially configured browser to `http://[CLUSTER_NAME]-m:9870`.

Connect through SSH port forwarding

LINUX OR MACOS | WINDOWS

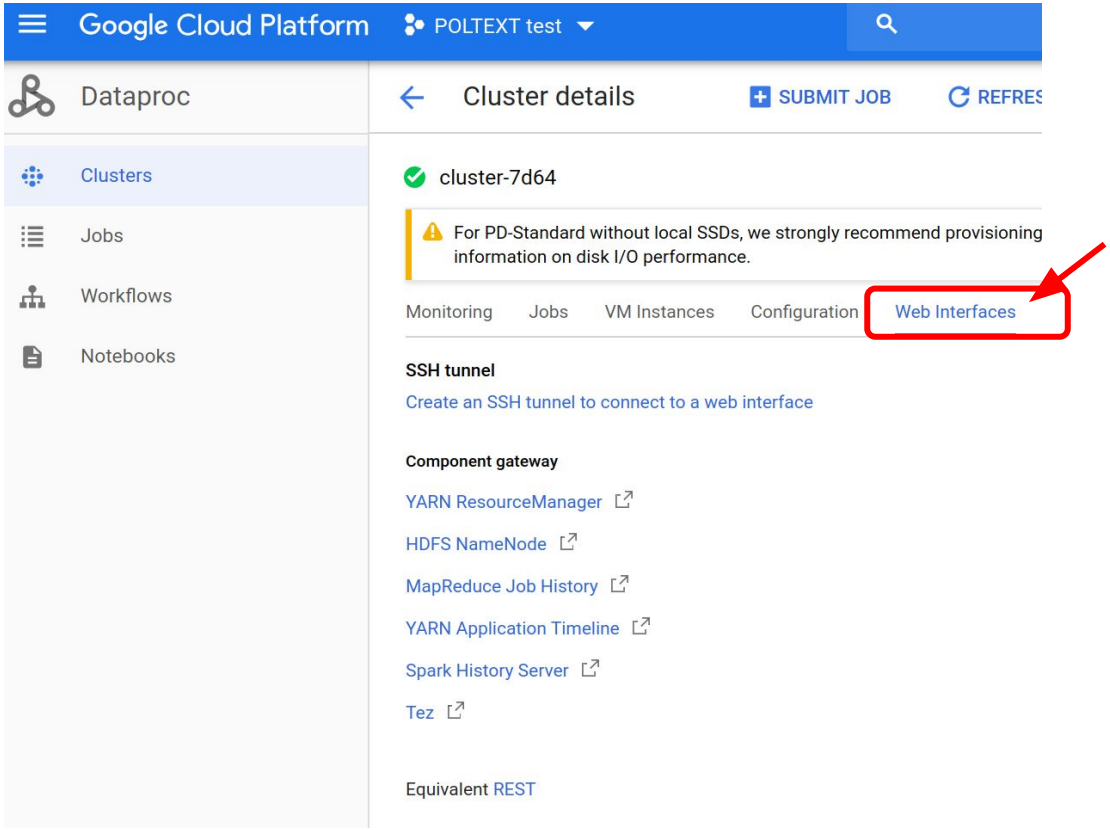
1. On your local machine, connect to the Cloud Dataproc master node:

```
gcloud compute ssh \
    --zone=[CLUSTER_ZONE] \
    --project=[PROJECT_ID] \
    [CLUSTER_NAME]-m -- \
    -L 8787:localhost:8787
```

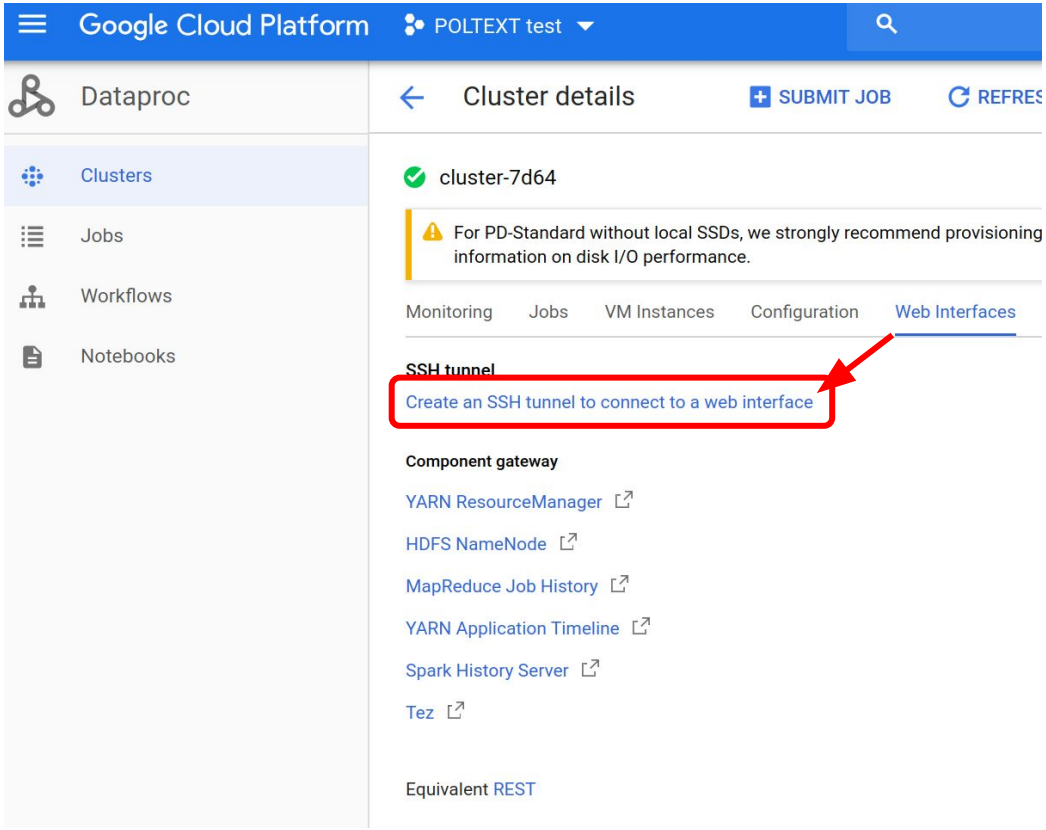
   The `--` parameter separates arguments for the `gcloud` command from arguments that are sent to the `ssh` command. The `-L` option sets up TCP port forwarding from port 8787 on the local machine to port 8787 on the cluster master node where RStudio Server is listening.

2. To load the RStudio web UI, connect your browser to `http://localhost:8787`.
3. Log in by using the username and password that you created.

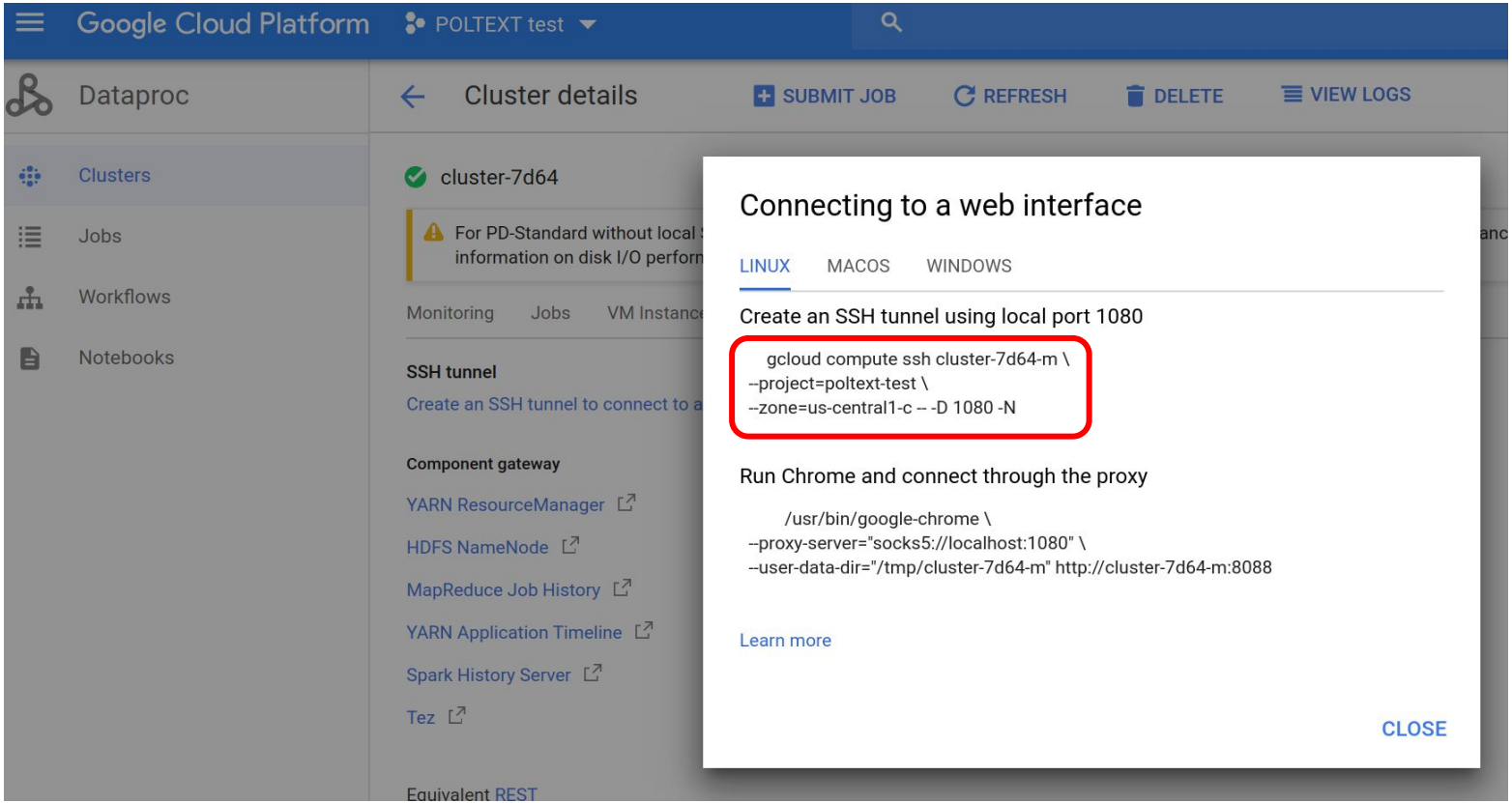
This is still https://cloud.google.com/solutions/running-rstudio-server-on-a-cloud-dataproc-cluster
We need the above information to create the gcloud ssh tunnel command. Luckily we can find it easily on our cluster management interface at Web Interfaces



Go to Web Interfaces on the cluster management page



Press Create an SSH tunnel...



Copy the SSH tunnel command from the pop-up to an empty file where you can edit it

Available interfaces

The following interfaces are available on a Cloud Dataproc cluster master node (replace `master-host-name` with the name of your master node).

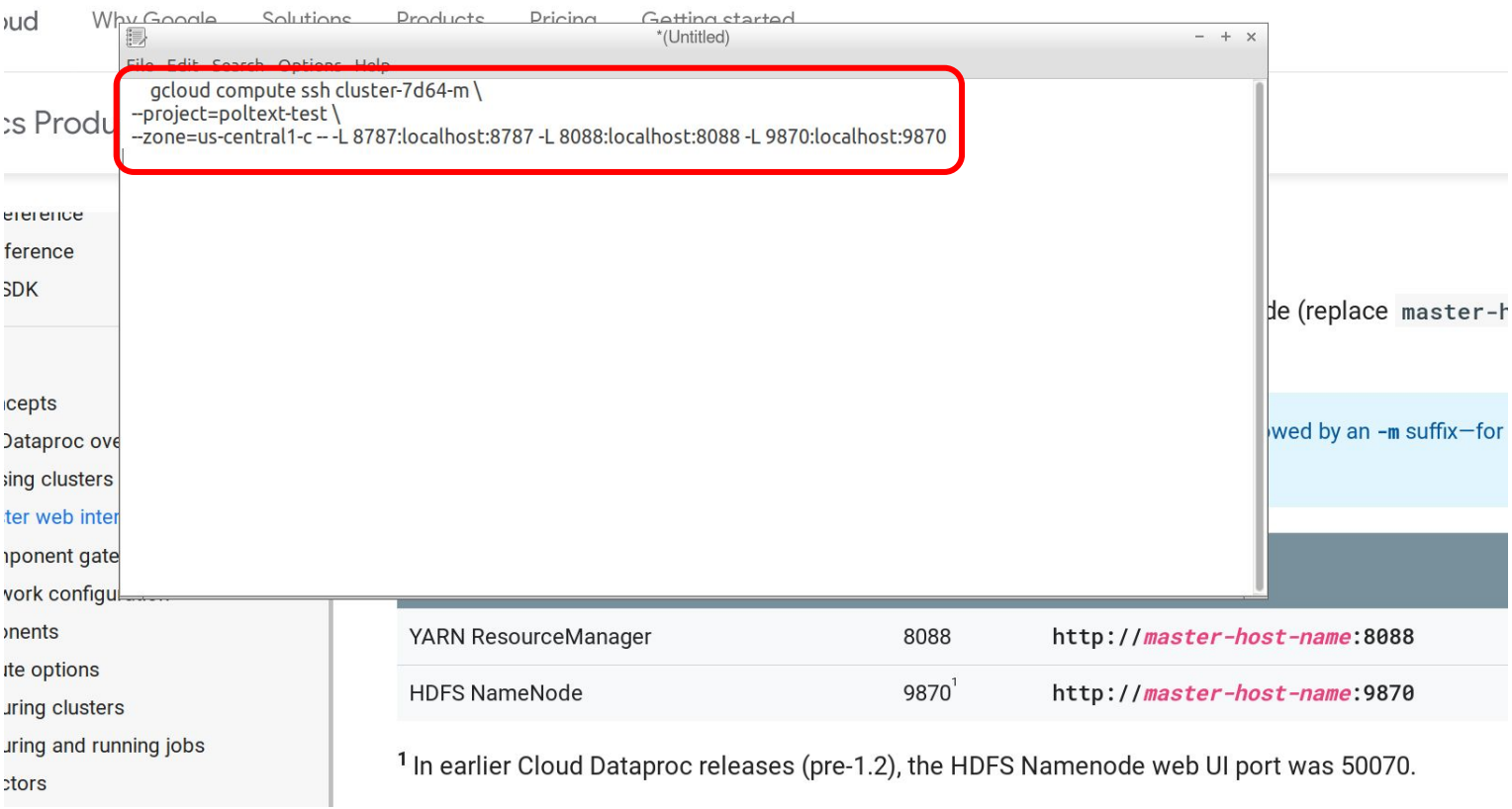
★ The cluster **master-host-name** is the name of your Cloud Dataproc cluster followed by an `-m` suffix—for example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".

Web UI	Port	URL
YARN ResourceManager	8088	`http://master-host-name:8088`
HDFS NameNode	9870[1]	`http://master-host-name:9870`

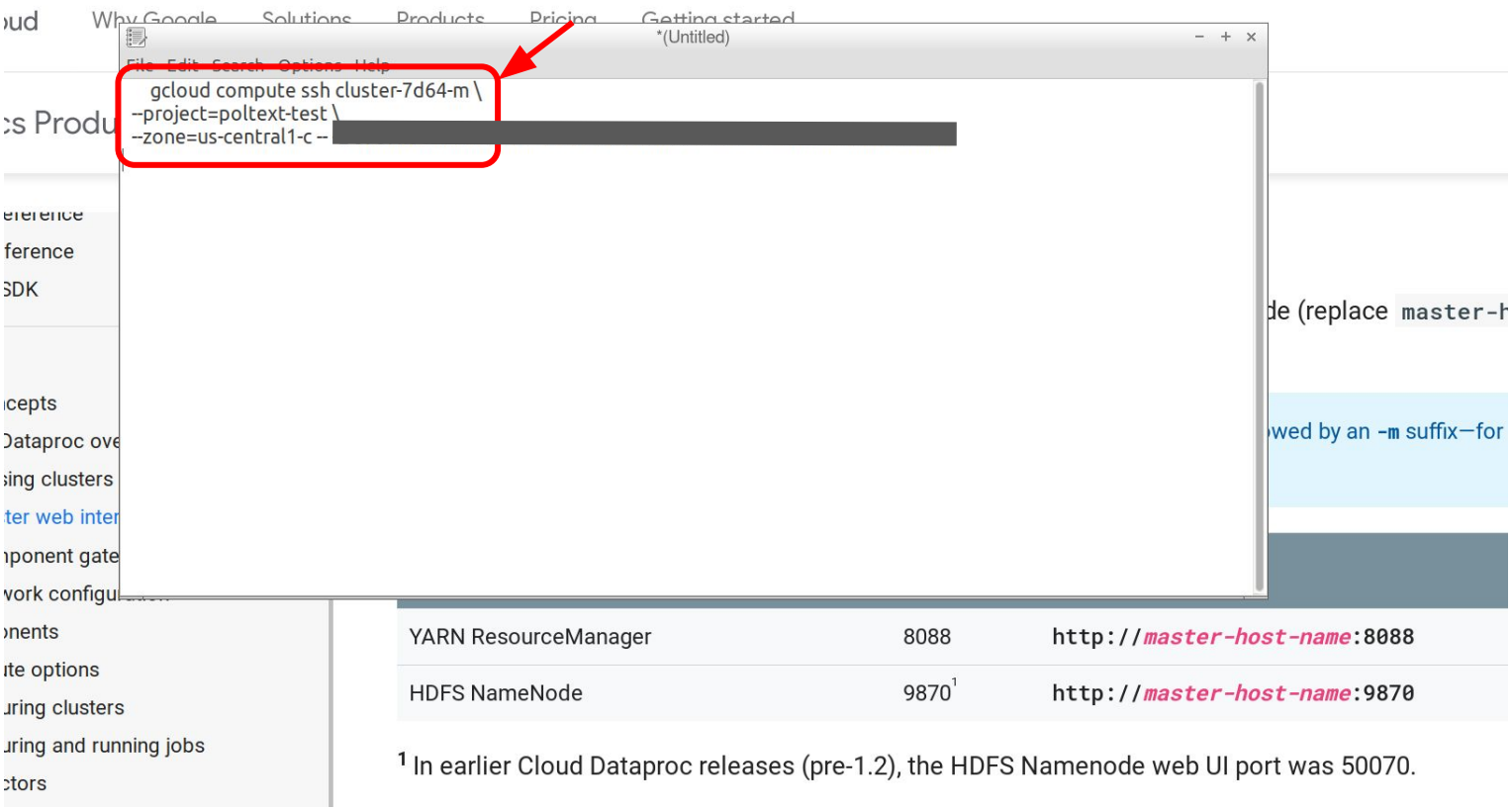
[1] In earlier Cloud Dataproc releases (pre-1.2), the HDFS Namenode web UI port was 50070.

The YARN ResourceManager has links for all currently running and completed MapReduce and Spark Applications web interfaces under the "Tracking UI" column.

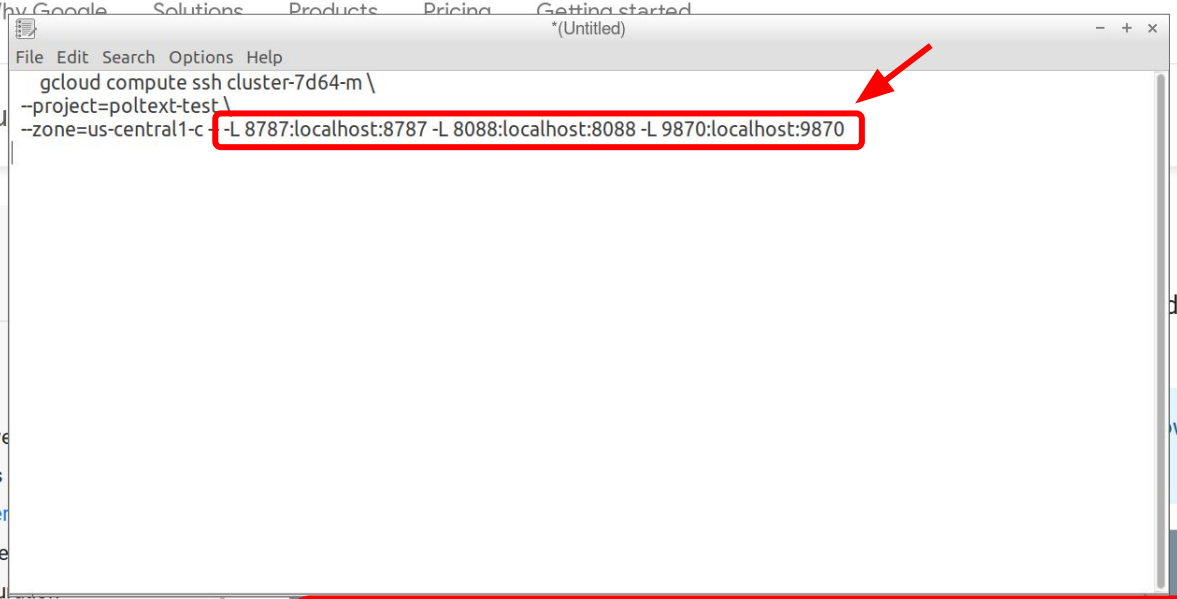
We can find the default port numbers for the HDFS NameNode and the cluster manager web UIs at: https://cloud.google.com/dataproc/docs/concepts/accessing/cluster-web-interfaces



Now we can put together the command we need to set up a simple ssh tunnel to access the web UIs. NOTE: you can also set up a SSH SOCKS tunnel instead: https://cloud.google.com/solutions/running-rstudio-server-on-a-cloud-dataproc-cluster



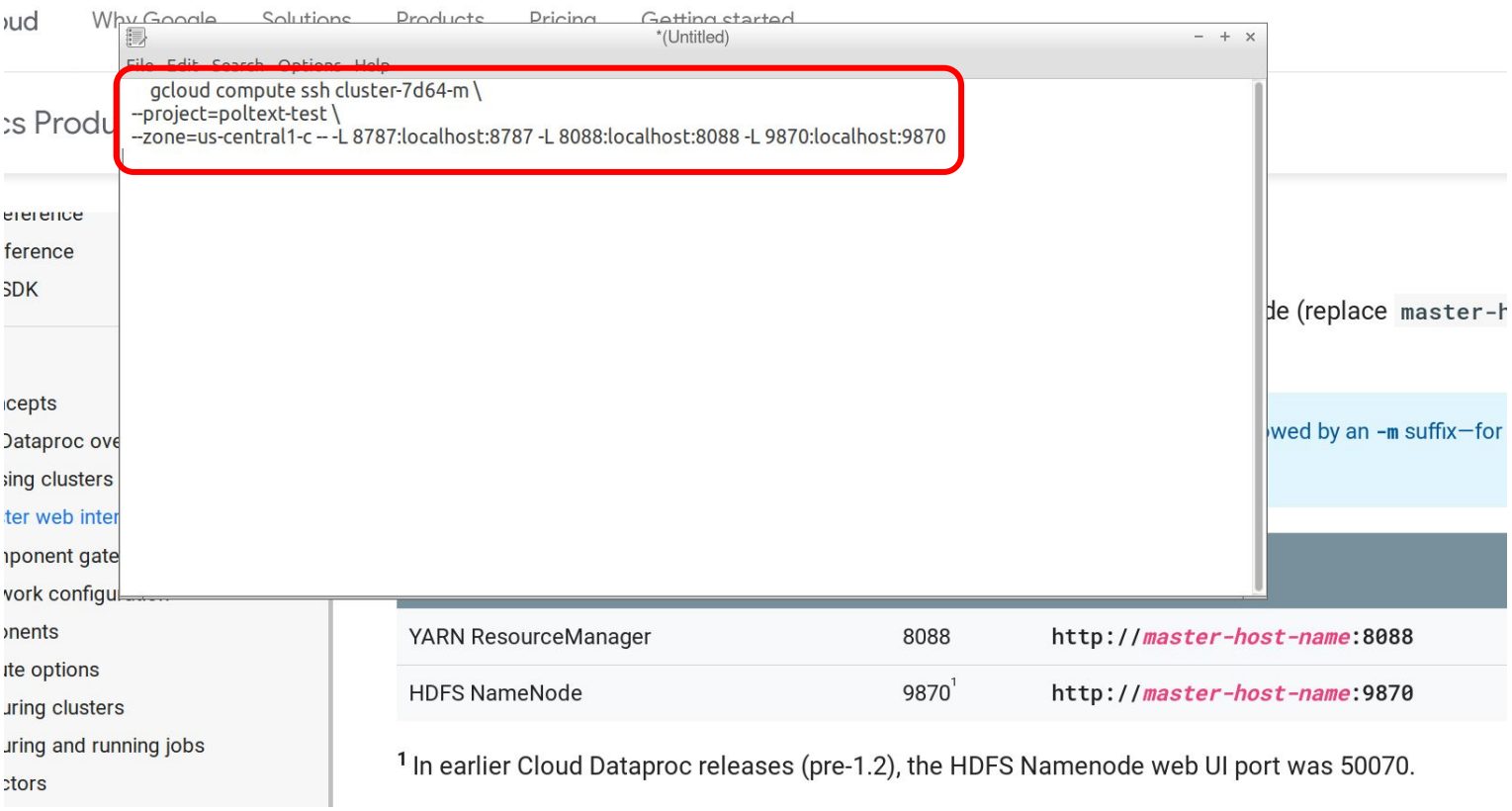
This is the part we copied from the cluster management Web Interfaces pop-up, minus the “-D 1080 -N” at the end. The -- marks the end of the gcloud ssh address, and the start of normal ssh command parameters.



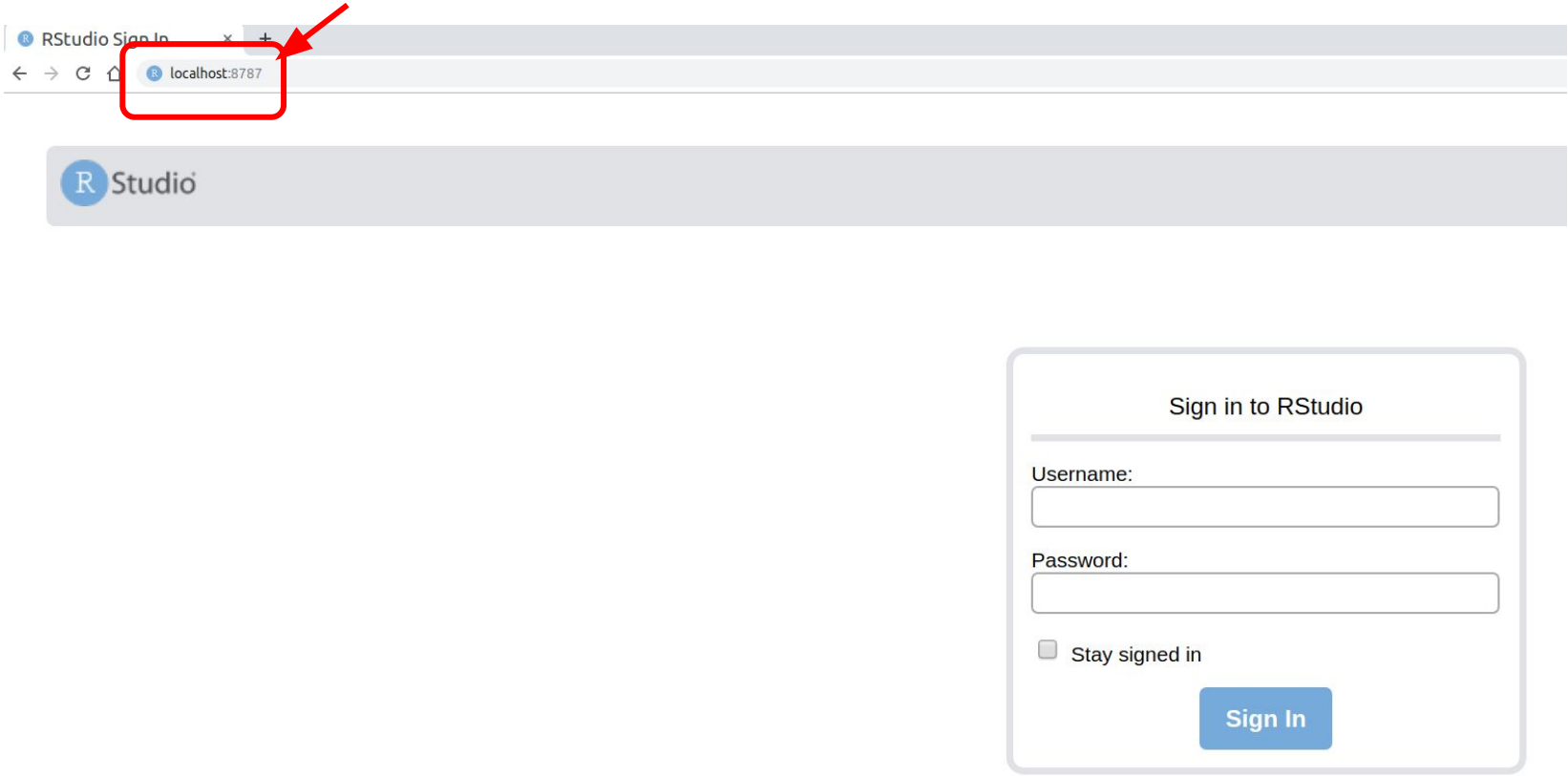
YARN ResourceManager	8088	http://*master-host-name*:8088
HDFS NameNode	9870[1]	http://*master-host-name*:9870

[1] In earlier Cloud Dataproc releases (pre-1.2), the HDFS Namenode web UI port was 50070.

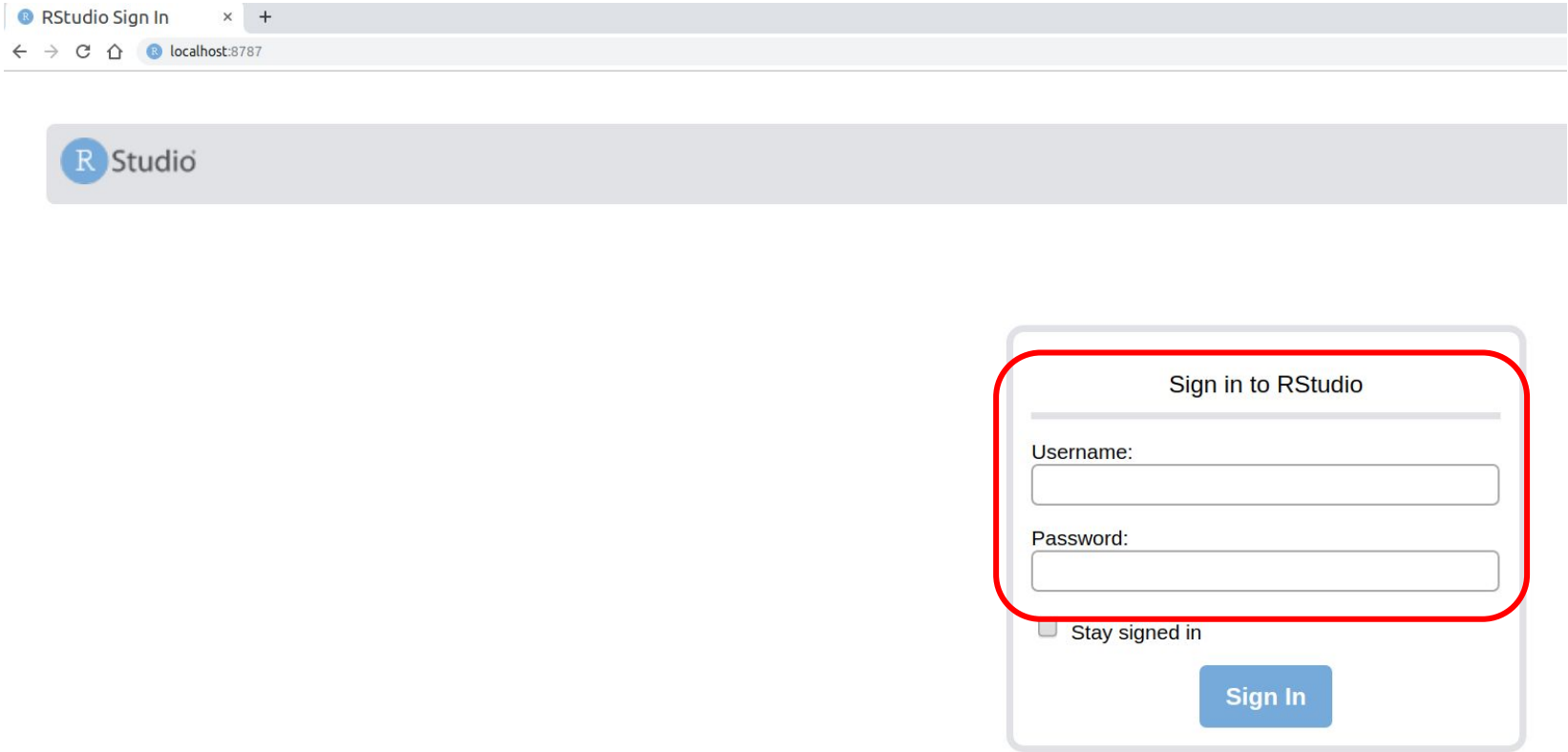
This is the part we are adding ourselves. -L is the flag for setting up the port forwarding, and localhost is our machine. We can set forwarding for more than one port in the same ssh command. 8787 is the default port number for RStudio Server, and the other two port numbers are for the cluster manager and the HDFS NameNode, from this page.



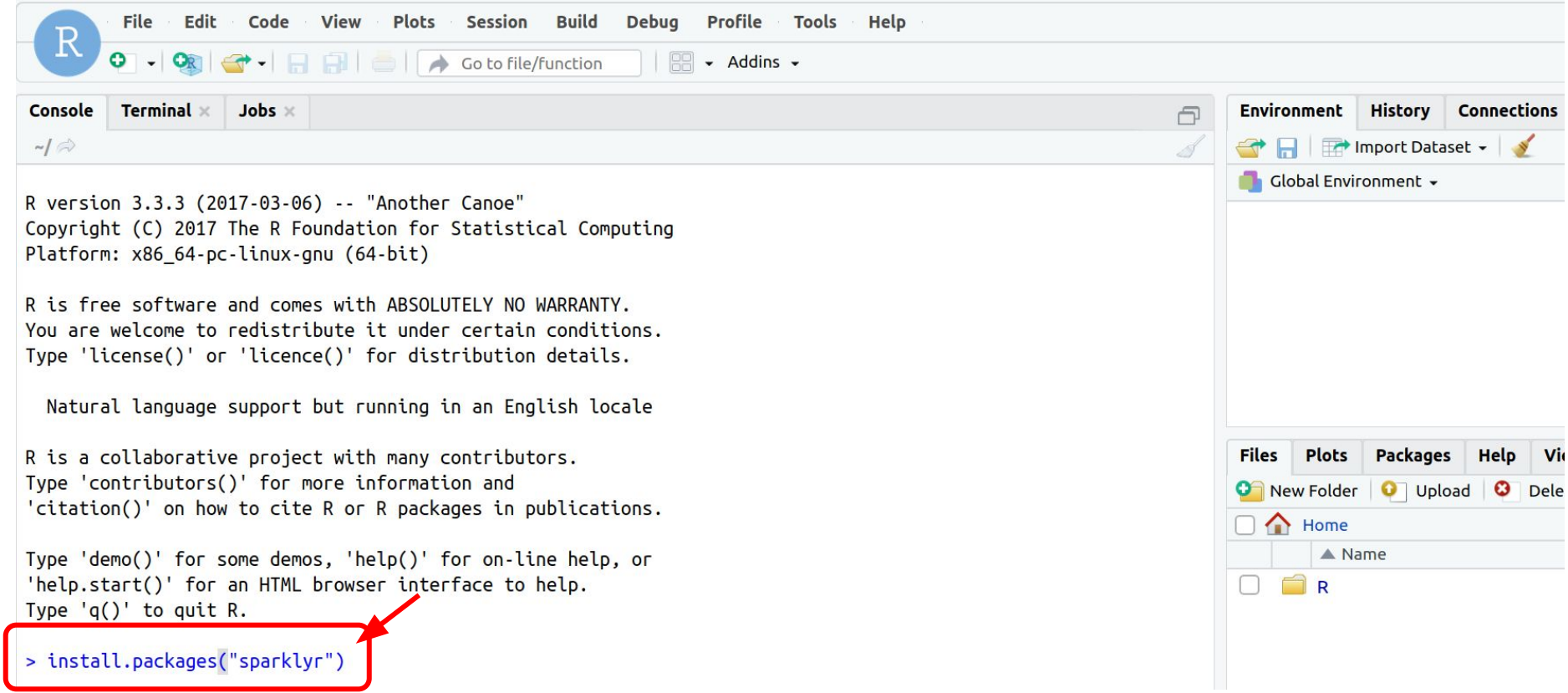
Copy this command into a command shell on your own computer where you have already set up the SDK and have connected with gcloud init, and press enter.



To access the RStudio Server web UI now you can enter the following address in your browser: localhost:8787



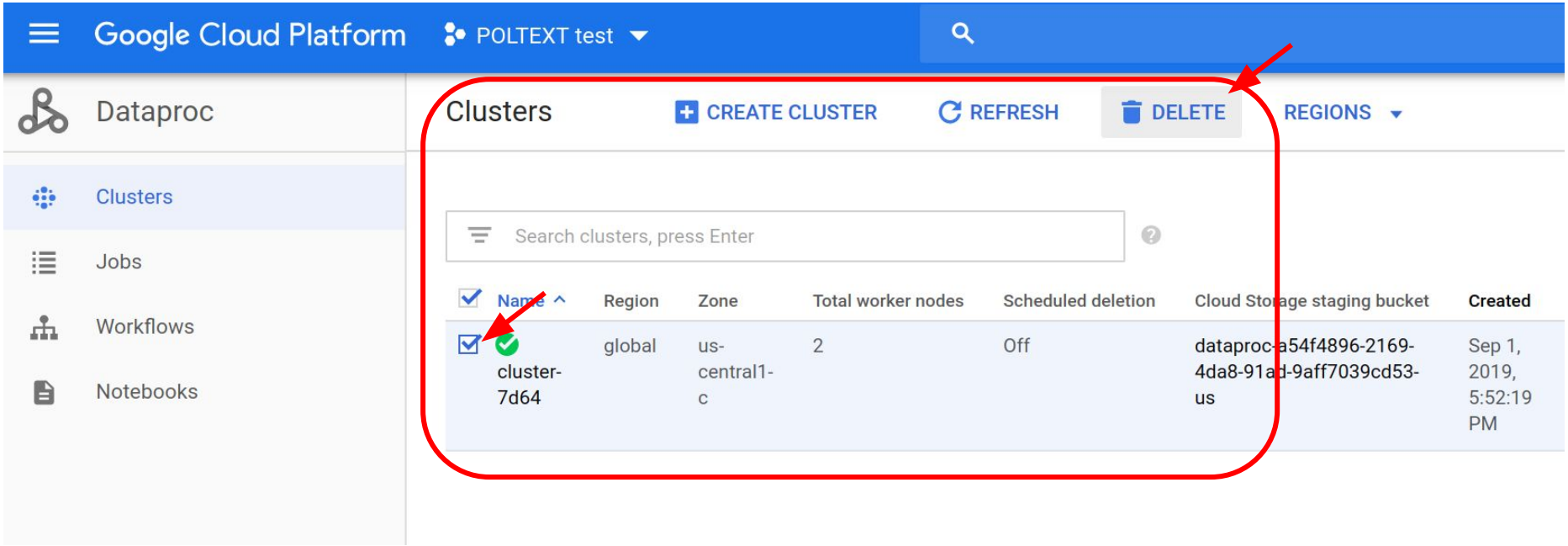
Sign in as “test” user with the password you set up the user profile with.



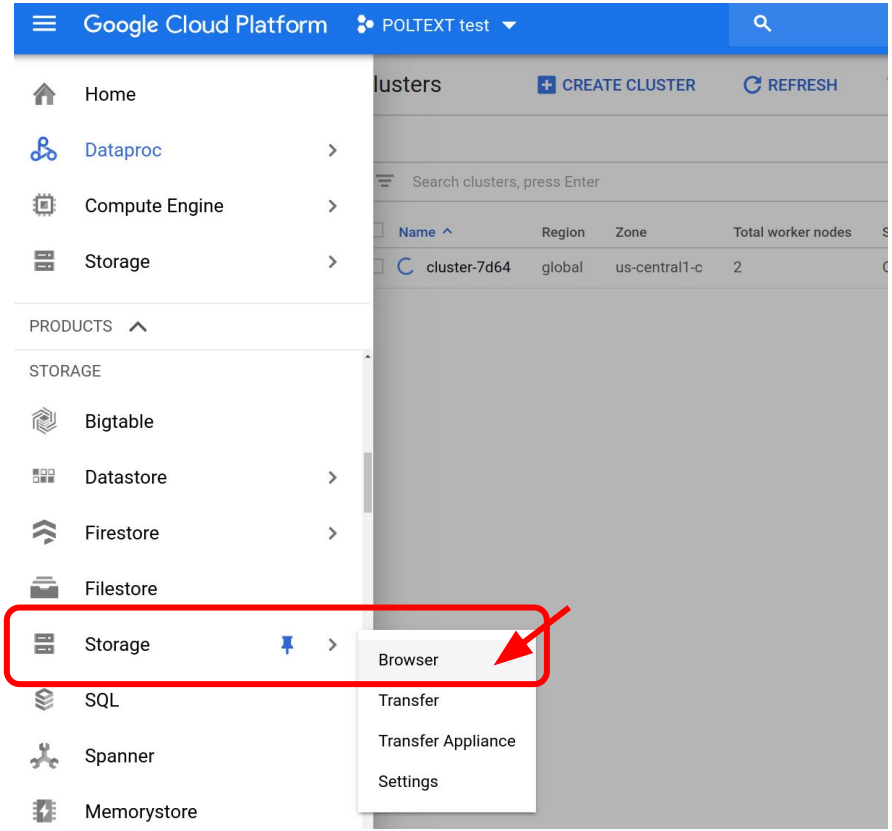
If you didn't run the install script on slide 10, then you still have to install the sparklyr package in the usual way: install.packages("sparklyr")

Great, we are ready with setting up our system.
But, please don’t forget, we will also have to clean up the cluster and storage once we are done.

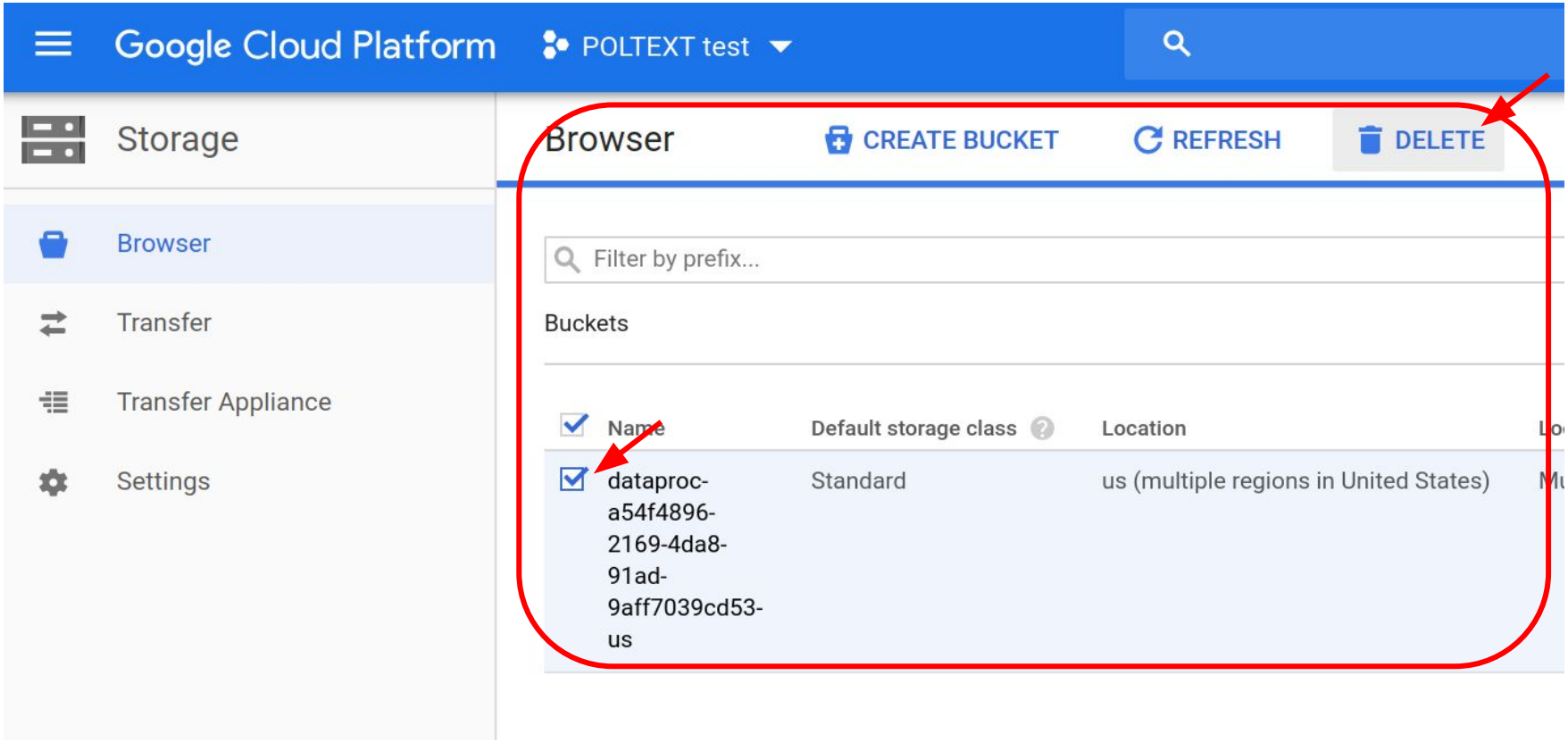
How to clean up your cluster and cloud storage after use to prevent unnecessary billing



To delete the cluster: Go back to the clusters overview page and mark the checkbox next to the cluster's name, then press delete



To delete the cloud storage: First go to the Storage Browser page



Then mark the checkbox next to the storage name and press delete