UNIVERSIDAD
COMPLUTENSE
MADRID

Proyecto de Innovación

Convocatoria 2021/2022

Nº de proyecto: 168

Computación interactiva en la nube con JupyterHub y Kubernetes

Responsable del proyecto: Sergio Pascual Ramírez

Facultad de Ciencias Físicas

Departamentos: Física de la Tierra y Astrofísica; Estructura de la Materia, Física Térmica y Electrónica

Objetivos propuestos en la presentación del proyecto

En la presente propuesta se pretende experimentar con el uso de cuadernos de computación interactiva del proyecto Jupyter, mediante sesiones de acceso y cálculo en la nube utilizando JupyterHub y tecnología de contenedores con Kubernetes en la nube.

Se trata de un modo de trabajo remoto en el que el alumno decide cuándo quiere llevarlo a cabo sin un horario fijo ni limitación de acceso, y lo más importante, haciendo uso de recursos del servidor remoto. Además, todo el proceso se beneficia del mantenimiento ofrecido por el servicio remoto, que cuenta con instalaciones y equipos profesionales dedicados. De esta manera se obtendrían los siguientes beneficios:

1. Aligerar la instalación y mantenimiento de software necesario para el trabajo.
2. Ahorro de inversión en recursos de espacio en disco y CPU. No es necesario duplicar los datos en cada ordenador.
3. Ahorro en mantenimiento y copias de respaldo.
4. Flexibilidad a la hora de invertir en recursos de CPU, memoria y espacio. En todo momento estos recursos se pueden dimensionar libremente en función de las necesidades, siempre que se cuente con presupuesto para ello.
5. Escalabilidad, por estar basado en tecnología de contenedores (Docker) que pueden desplegarse bajo demanda.
6. Portabilidad, uniendo el acceso de recursos en la nube con cuadernos Jupyter que solo requieren un navegador.
7. Flexibilidad con la autenticación de usuarios, que permite reutilizar las credenciales UCM (proporcionadas por Google) para acceder a los recursos.

Objetivos alcanzados

Los objetivos 1 a 7 se han alcanzado al desplegar JupyterHub con Kubernetes en la nube de Google Cloud.

Kubernetes proporciona escalabilidad y persistencia de los datos, al almacenarse en pequeños discos por usuario que se mantienen entre ejecuciones.

Se han creado también imágenes (basadas en Docker) personalizadas de JupyterHub, con los paquetes de Python específicos ya instalados.

Se han creado, así mismo, cuadernos de Jupyter específicos para probar las capacidades de JupyterHub con Kubernetes.

Tantos los cuadernos como las imágenes se encuentran disponibles de manera pública en Github, dentro de la organización https://github.com/proyecto168

Los contenedores de Docker están disponibles también en https://hub.docker.com/u/sergiopasra . Estas imágenes se basan en las imágenes canónicas de JupyterHub disponibles en https://hub.docker.com/u/jupyterhub

Cabe mencionar que estas imágenes son útiles por sí mismas, incluso sin Kubernetes, ya que permiten instalar JupyterHub, si bien como contenedor personal

El punto 7, referido a la autentificación de usuarios mediante credenciales de Google, es el único que no ha llegado a implementarse. Hemos preferido implementar una solución similar (ya que usa el mismo protocolo OAuth) pero basada en Github. Esta solución permite el acceso según la pertenencia del usuario a organizaciones o grupos y nos ha parecido más sencilla de implementar y con un control más granular.

La conexión con OAuth tiene que ir cifrada con certificados digitales; los certificados van ligados a nombres de dominio. La solución ha sido adquirir un dominio (proyecto168.es) y utilizar los servicios de LetsEncrypt (https://letsencrypt.org/) para proporcionar y renovar certificados digitales de manera automática.

Este mecanismo plantea algunas dificultades técnicas, por ejemplo:

- Una condición de carrera en las imágenes estables del cuadro de configuración (*helm chart*) de JupyterHub en Kubernetes hace que el certificado no llegue a adquirirse.
- Los servicios informáticos de la UCM bloquean los accesos desde o hacia nombres de dominio nuevos cerca de un mes (lo que bloqueó el uso de nuestra dirección https://cuaderno.proyecto168.es).

Los problemas anteriores se solucionaron, respectivamente:

- Usando una versión de desarrollo del cuadro de configuración.
- Esperando un mes hasta que los dominios nuevos pudieron usarse.

Metodología empleada en el proyecto

Durante el primer cuatrimestre del curso, se estuvo experimentando con el despliegue de JupyterHub con Kubernetes en la nube de Google, que es una actividad relativamente compleja por la gran cantidad de opciones disponibles.

A principios del segundo cuatrimestre se creó una organización en Github con el propósito de contener cuadernos de Jupyter y poder utilizar las capacidades de OAuth de Github.

Se crearon también diferentes repositorios con cuadernos de Jupyter y ficheros de datos, así como imágenes especializas de JupyterHub.

Se realizaron diferentes pruebas de despliegue de Kubernetes, con y sin nombre de dominio (utilizando directamente la dirección IP de Google Cloud o usando un nombre del dominio proyecto168.es).

Se llevó a cabo una clase práctica con 20 alumnos utilizando sus ordenadores para acceder a un *cluster* de JupyterHub especializado para la asignatura.

Recursos humanos

Sergio Pascual Ramírez (PDI Complutense)

Cristina Cabello González (Investigadora predoctoral FPI Complutense)

Nicolás Cardiel López (PDI Complutense)

José Luis Contreras González (PDI Complutense)

Jesús Gallego Maestro (PDI Complutense)

Daniel Sánchez Parcerisa (PDI Complutense)

Ainhoa Sánchez Penim (PAS Complutense)

Jaime Zamorano Calvo (PDI Complutense)

Desarrollo de las actividades

Durante el primer cuatrimestre del curso 21-22 se estableció el uso de Google Cloud y se estuvo afinando el uso de las diferentes herramientas disponibles.

Un esquema de la instalación de JupyterHub sería:

- JupyterHub se despliega sobre Kubenetes utilizando *helm.* El cuadro de despliegue *(helm chart)* se encuentra en https://artifacthub.io/packages/helm/jupyterhub/jupyterhub
- Con este *helm chart* se crea la estructura general de la aplicación, que incluye nodos redundantes, balanceadores de carga, almacenamiento persistente, etc.
- El usuario final decide qué imagen de JupyterHub se va a desplegar (que en la práctica difieren de los paquetes Python que contienen).

En el segundo cuatrimestre del curso 21-22 se han realizado la mayor parte de las actividades.

Se creó una organización (colección de recursos) en Github: https://github.com/proyecto168

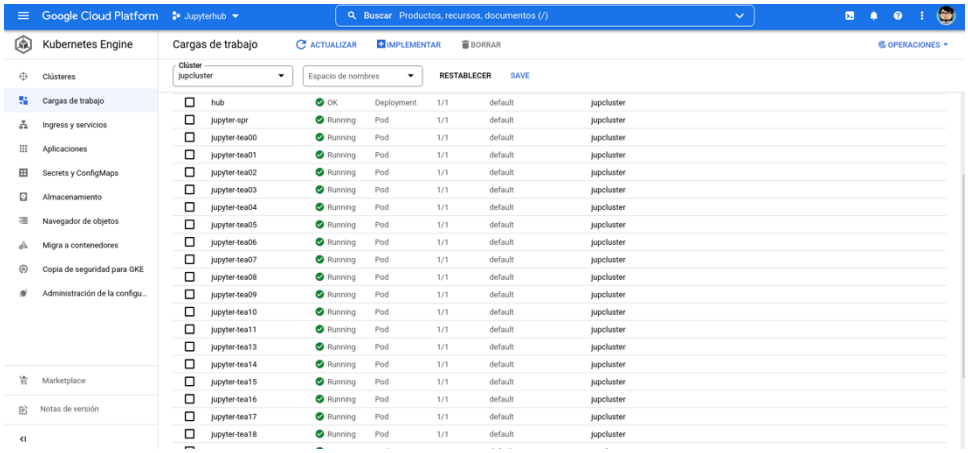
Esta organización de Github contiene diferentes repositorios que se han utilizado en el proyecto, tanto cuadernos de Jupyter como contenedores de JupyterHub especializados para diferentes asignaturas.

Con este despliegue de JupyterHub, se llevó a cabo una sesión de prácticas de la asignatura *Técnicas Experimentales en Astrofísica* del Máster en Astrofísica en marzo de 2022. Los alumnos del grupo realizaron la práctica sin mayores incidencias, utilizando los navegadores de sus equipos, pero sin necesidad de instalar ningún software adicional y con los datos requeridos para la práctica preinstalados. En este caso se utilizó autenticación simple, es decir, un usuario y contraseña que se les proporcionó en el aula.

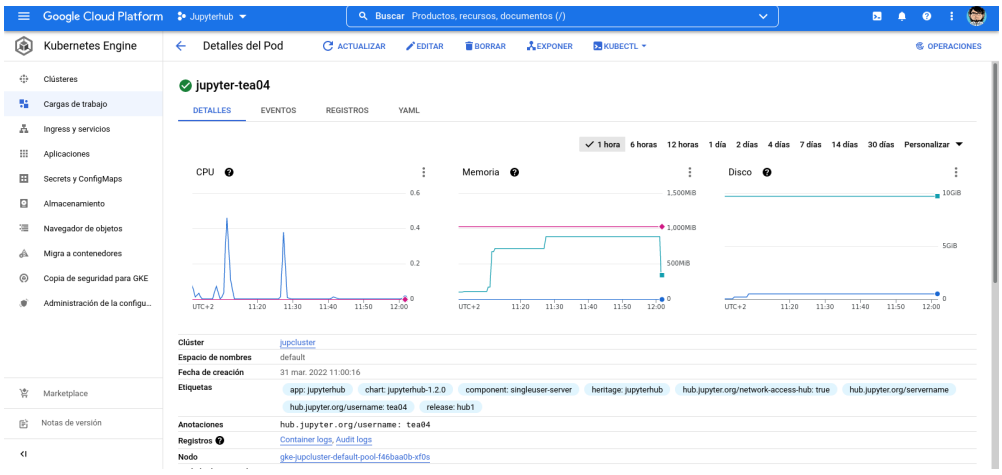
Una vez que las direcciones del dominio proyecto168.es fueron accesibles, a mediados de mayo de 202, los miembros del proyecto pudieron probar el sistema completo, con nombre de dominio, certificado digital y autentificación de usuarios mediante OAuth con Github, obteniéndose resultados favorables.

Anexos

Captura de pantalla de la lista de volúmenes asignados a los usuarios de la clase de Técnicas Experimentales. Cada volumen ocupa 10 Gb y es persistente entre diferentes ejecuciones de la instancia de JupyterHub.

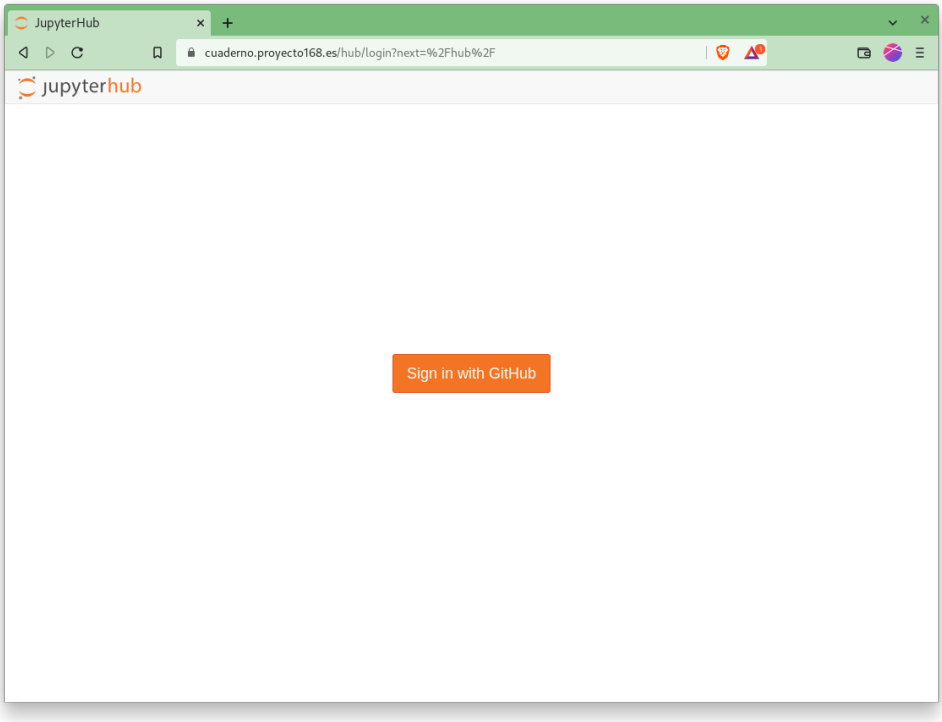


Captura de pantalla de algunas métricas de control del nodo del usuario *tea04* durante la realización de la práctica. Se incluyen uso de memoria, CPU y ocupación de disco.

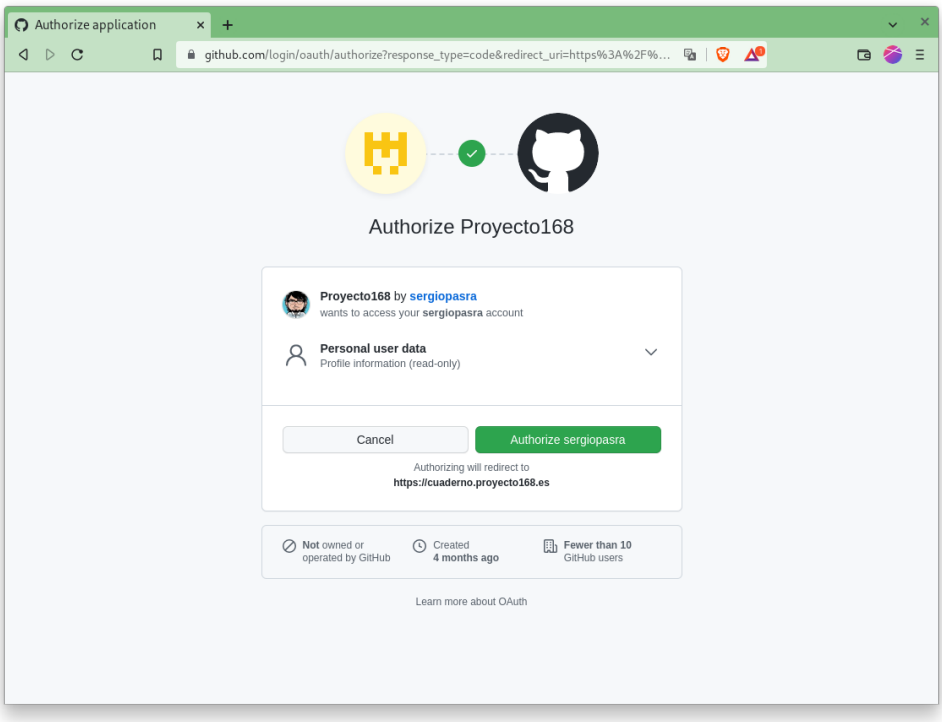


Autenticación con OAuth en Github

Captura de la pantalla de entrada cuando se utiliza autentificación con OAuth, con Github de proveedor.



Siguiente pantalla, en la que se solicita permiso para acceder a los datos de Github, utilizando una pequeña aplicación OAuth.



Tras aceptar la autorización, se introducen las credenciales de Github y ya es posible entrar en JupyterHub.