Práctica de PGITIC con Vagrant y Ansible. 2019/20.

Objetivo: aprovisionar un *box* de Vagrant, usando Ansible, con el software *Cloudera Distribution Hadoop*, CDH 5.

Crear una máquina virtual, basada en un *box* Ubuntu o Debian de Vagrant, donde, mediante un *playbook* de Ansible, se aprovisione la máquina con el software CDH 5. El *playbook* debe **automatizar** el procedimiento a seguir para **instalar CDH 5 e iniciar Hadoop.**

Debe seguirse el procedimiento explicado por Cloudera a partir de:

https://docs.cloudera.com/documentation/cdh/5-1-x/CDH5-Quick-Start/CDH5-Quick-Start.html

En particular, debe seguirse cuidadosamente el procedimiento consistente en descargar e instalar CDH 5 “1-click Install". Téngase en cuenta que:

- Hay que instalar “Java Development Kit” (JDK) para que funcione CDH 5. Si Oracle JDK da problemas, instalar en su lugar OpenJDK.
- El software se instala en un único nodo en modo pseudo-distribuido.
- Es suficiente instalar CDH 5 con MRv1 (no es necesario instalar YARN).
- Se debe escoger con cuidado el box Ubuntu o Debian que se usará (no sirve cualquiera).

Se debe sincronizar, al menos, un directorio de la máquina anfitrión (en el que ubicar, por ejemplo, programas MapReduce y archivos de entrada) con uno de la máquina virtual (sobre la que poder ejecutar esos programas en el HDFS).

Se valorará:

- Que funcione el software instalado.
  - El trabajo debe incluir un **documento** donde se explique **cómo se ha verificado que Hadoop funciona correctamente** en la máquina virtual.
- Que se siga el procedimiento indicado en la documentación de Cloudera adaptado a su automatización con Ansible.
- **Usar módulos de Ansible específicos** en vez de otros más genéricos (por ejemplo, evitar usar *command* cuando hay un módulo más específico para la tarea que se quiere realizar).
- Que el *playbook* y el *Vagrantfile* estén correctamente comentados.