Flujo de trabajo en Ciencia de Datos

Manejo de herramientas computacionales

Daniel Jiménez M.

UNAL

02 -11 -2021

Contenido

- Roles en el trabajo de ciencia de datos;
- ¿Cómo hacer limpieza de datos?;
- Git & Github para el trabajo colaborativo;
- Distribución de sistemas;
- Deployment;
- Ejemplo del objetivo del curso

Roles en el trabajo de ciencia de datos

Empecemos por definir la ciencia de datos: Es una ciencia en desarrollo, que busca extraer conocimiento sobre los datos disponibles y con base a ello desarrollar programas que puedan clasificar, pronosticar o clusterizar algún objetivo.

Roles en el trabajo de ciencia de datos

- Ingenieria de datos;
- Científicos de datos;
- Machine Learning Engineer;
- Deep Learning Engineer;
- AI Engineer

Cuando crees que todo se resuelve con ciencia de datos

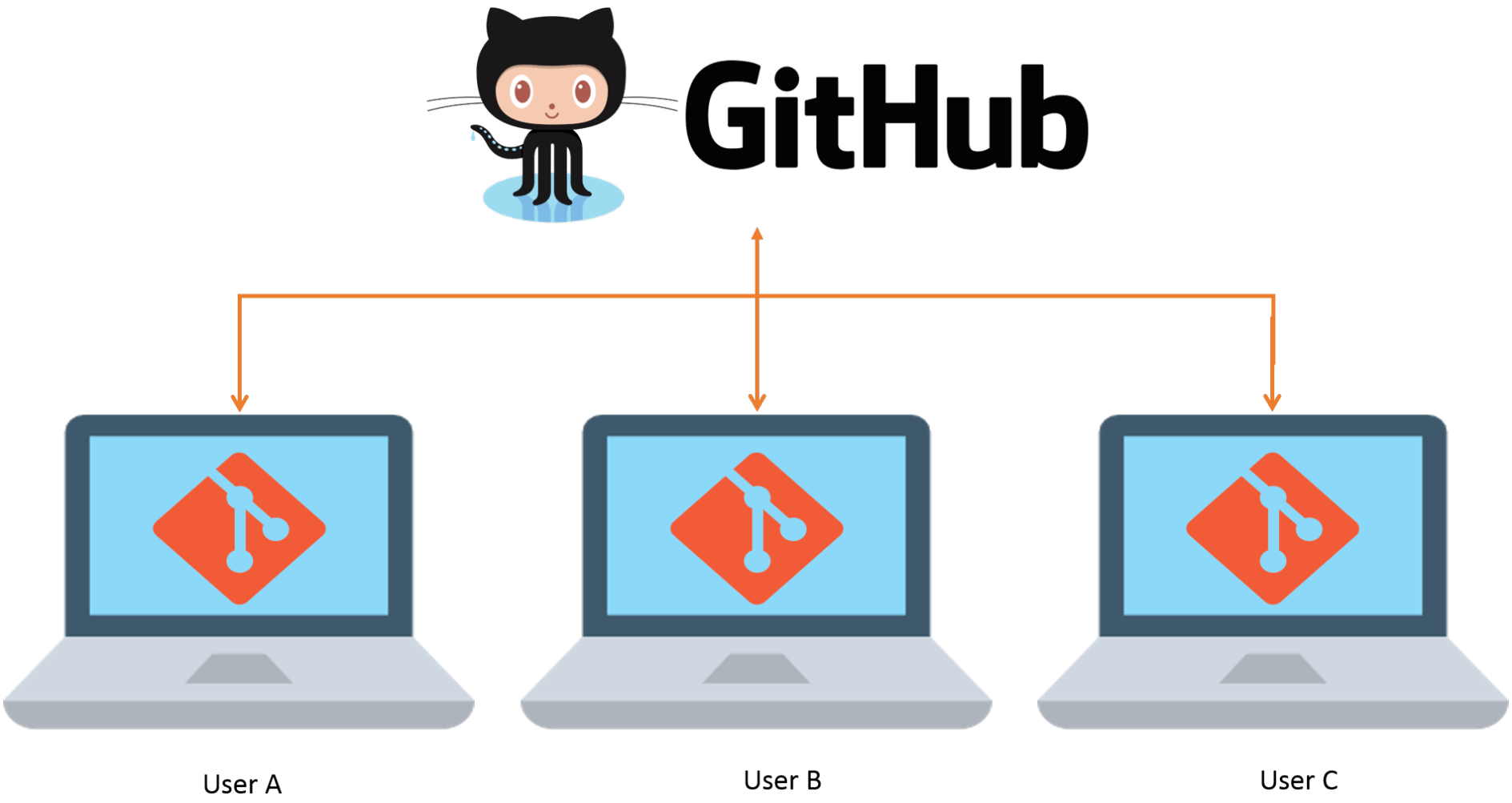
Roles en el trabajo de ciencia de datos

1. Es importante trabajar la parte de ingenieria de datos en el total que la conforma
2. Es necesario trabajar en CLOUD por temas de disponibilidad de recursos.

¿Cómo hacer limpieza de datos?

- SQL;
- Regex;
- Normalización de data

Git & Github para el trabajo colaborativo



Git & Github para el trabajo colaborativo

1. Instale git haciendo click aquí;
2. Abrá una cuenta en github, haciendo click aquí;
3. Cree una llave ssh:
   - Si es **Windows**: ssh-keygen -t rsa -b 4096 -C + "su correo"
   - Si es **MAC** :ssh-keygen -t rsa -b 4096 -C + "su correo"
   - Adjunte la clave en **Windows**: eval $(ssh-agent - s) ssh-add ~/.ssh/id_rsa
   - Si es Mac : eval "$(ssh-agent -s)"

Distribución de sistemas

Esta parte es la que garantiza que se pueda desplegar todo el trabajo desarrollado en cualquier sistema.

Generalmente se trabaja con Docker

Deployment

Es la forma de automatizar rutinas, generalmente se hace con jenkins

Ejemplo del objetivo del curso

Trabajaremos en el siguiente notebook

BERT