Data Project

Build Your Own Hate speech detection Engine

January 2021

www.tbs-education.fr

Definition

- Encyclopedia of the American Constitution: “Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity.”
- Facebook: “We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.
- Twitter: “Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.”

Definition

- Hate speech is to incite violence or hate
- Hate speech is to attack or diminish
- Hate speech has specific targets
- Whether humor can be considered hate speech

Motivation

- Hate crimes are nothing new in society;
- Social media have begun playing a major role in hate crimes;
- Suspects in several recent hate-related terror attacks had an extensive social media history of hate-related posts;
- Social media could contribute to their radicalization;
- Many online forums such as Facebook, YouTube, and Twitter consider hate speech harmful,

Your mission

- You are part of the TWITTER team of data scientists.
- This company offers microblogging and social networking service on which users post and interact with messages known as "tweets".
- She would like to develop a new application able to detect hate speech and offensive messages.

The objective

- For a given tweet/retweet, your application should be able to classify the comment as a hate message, an offensive message or neither hate speech nor offensive;
- “Rshiny” should be used to convert your R codes to an Application;
- Git and Github should be used as collaborative working tools,

Possible Datasets

- **The dataset1 is stored as a CSV. Each data file contains 5 columns:**
  - **count = number of experts who coded each tweet (min is 3, sometimes more experts coded a tweet when judgments were determined to be unreliable).**
  - **hate_speech = number of experts who judged the tweet to be hate speech.**
  - **offensive_language = number of experts who judged the tweet to be offensive.**
  - **neither = number of experts who judged the tweet to be neither offensive nor non-offensive.**
  - **class = class label for majority of experts: 0 - hate speech 1 - offensive language 2 - neither**

Possible Datasets

- The dataset2 contain a non-exhaustive list of Abuse words
    - 404 abuse terms used in:
        - "Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians", ICWSM 2018,
        - "Online abuse of uk mps in 2015 and 2017: Perpetrators, targets, and topics", extended version on arXiv, 2018.
    - 388 abuse terms used in:
        - "Online Abuse of UK MPs from 2015 to 2019: Working Paper", arXiv working paper.

Possible Datasets

- The dataset3 contain a list of 1528 annotated comments from Fox News website.
  - 435 of them are labeled as hateful;
  - 1093 of them are labeled as non-hateful.

Possible Datasets

- The dataset4 contain a list of abusive microposts that were considered to be explicitly abusive
  - The first line of each micropost represents the class label which is either "abusive" or "notAbusive"

Possible Datasets

- No restriction on using other datasets !!!!!

Methodology

- Text mining ?
- Logistic regression ?
- Naïve Bayes ?
- Decision trees ?
- Other Machine Learning Technics ?

Methodology

- **Some Related papers:**
  - **Automated Hate Speech Detection and the Problem of Offensive Language, Thomas Davidson et al. (2017);**
  - **Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter, Ziqi Zhang et al. (2018);**
  - **Hate speech detection: Challenges and solutions, Sean MacAvaney et al. (2019)**

Evaluation criteria: Groups of 05 students

- Evaluation criteria
  - Understanding of business context (identification of key points, understanding of sector) Mastery & Pertinence of statistical analysis (descriptive, modeling);
  - Link between data / analysis / business recommendations (logical flow, pertinence, completeness);
  - Business concepts (thoroughness, pertinence);
  - Thoroughness and quality of business recommendations (pertinence, relevance, professionalism, originality);
  - Understanding of big data (mastery of big data terms, confidence when using big data concepts);
  - Visuals (professionalism, slides support well the main arguments of the presentation, appropriate content);
  - Delivery (clear and logical organization, effective introduction and conclusion, creativity, transition between speakers, oral communication skills, eye contact);
  - Q&A session (ability to answer questions);
  - Report: Quality of data analysis,

Evaluation criteria: Groups of 04 students

- Very important:
  - Try to use materials you learned during all the Program
  - We want to be impressed !!!
  - Presentation during the last session:
    - 20 minutes per groups;
    - Presentation in front of jury members.

Use Github

- A collaborative tool to work together on projects;
- A code hosting platform for version control and collaboration;
- Basic services are free of charge;
- More advanced professional and enterprise services are commercial;
- Free GitHub accounts are commonly used to host open-source projects;

Use Github

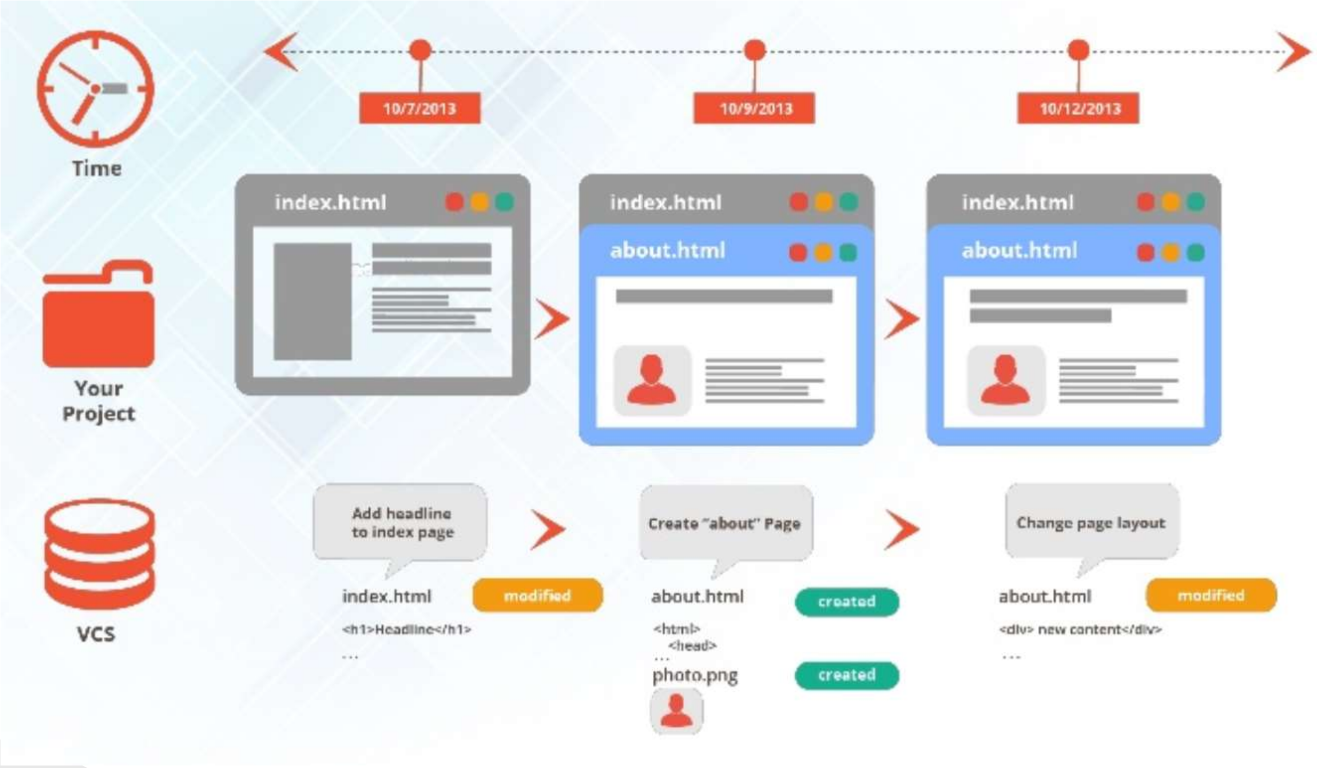
- Free plan: unlimited collaborators, private repositories restricted to 2,000 minutes of GitHub Actions per month;
- 40 million users and more than 190 million repositories;
- The largest host of source code in the world;
- Since October 26, 2018: Acquired by Microsoft.

Get started with Github

- Version control
  - The management system that manages the changes that you made in the project: addind new files, modifying older files, etc.
  - Every time you make a change on your project a different version is created and saved (called a snapshot);
  - All the versions are kept,

Get started with Github

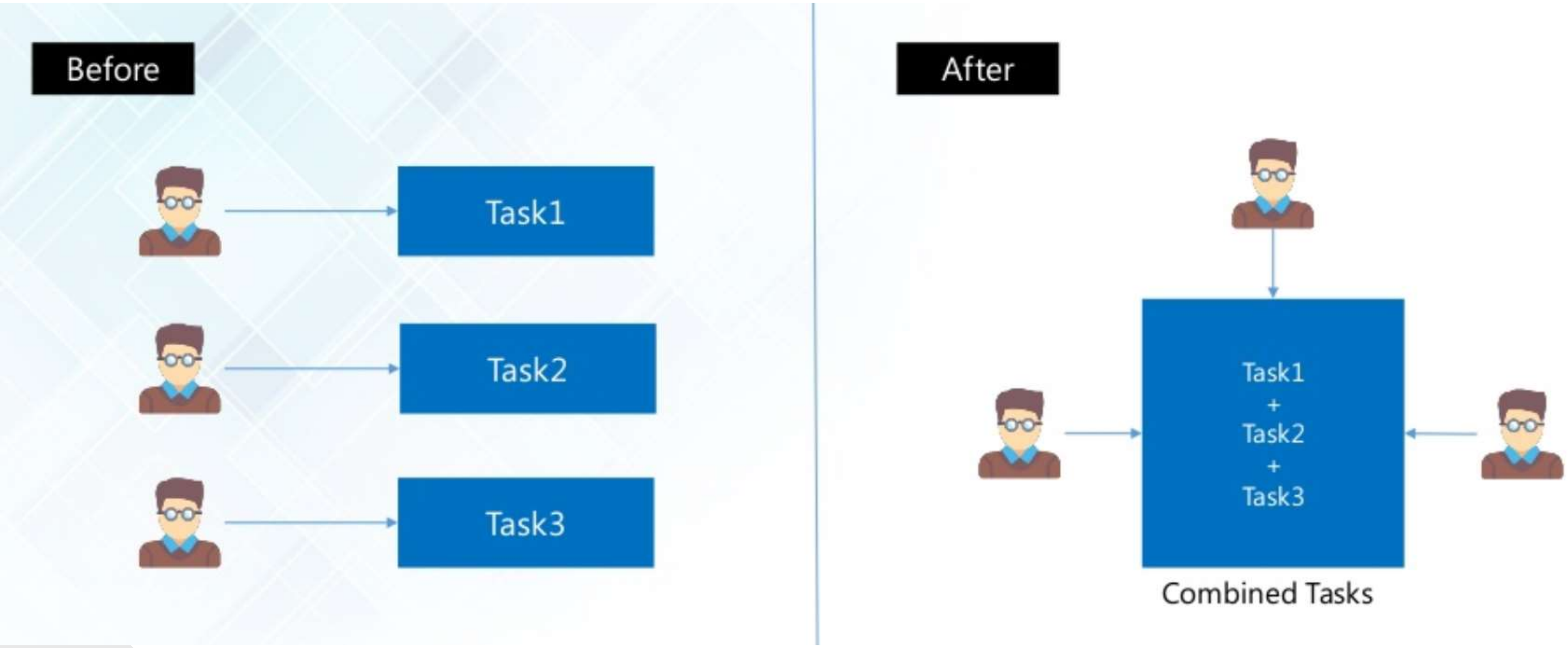
- **Version control**



- ➢ Version control is the management of changes to documents, computer programs, large web sites, and other collections of information.
- ➢ These changes are usually termed as "versions".

Get started with Github

- Why a Version control ?
  - You are always notified to changes on files;
  - Avoid change conflicts;
  - Facilitates collaboration



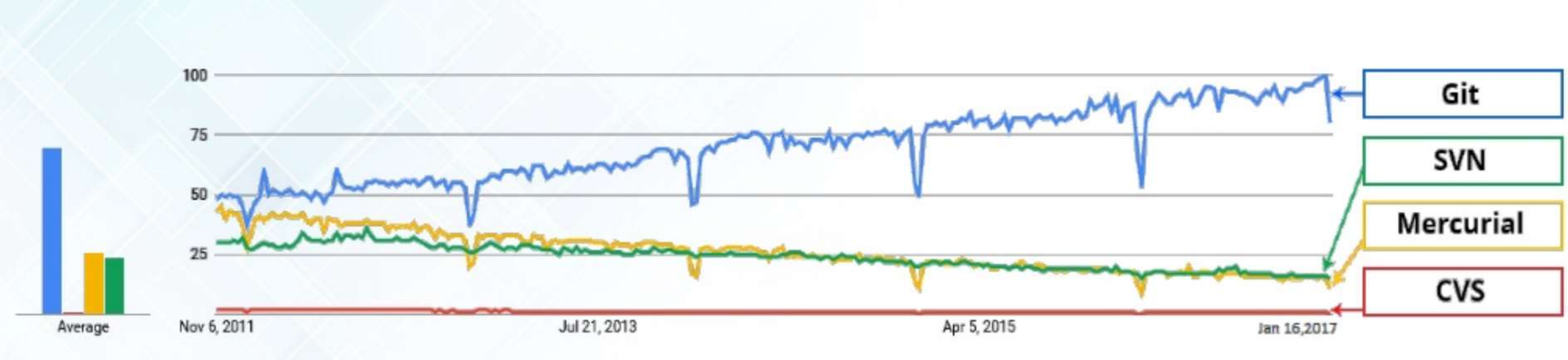
Get started with Github

- **Version control System Tools**

git
SUBVERSION
CVS
mercurial

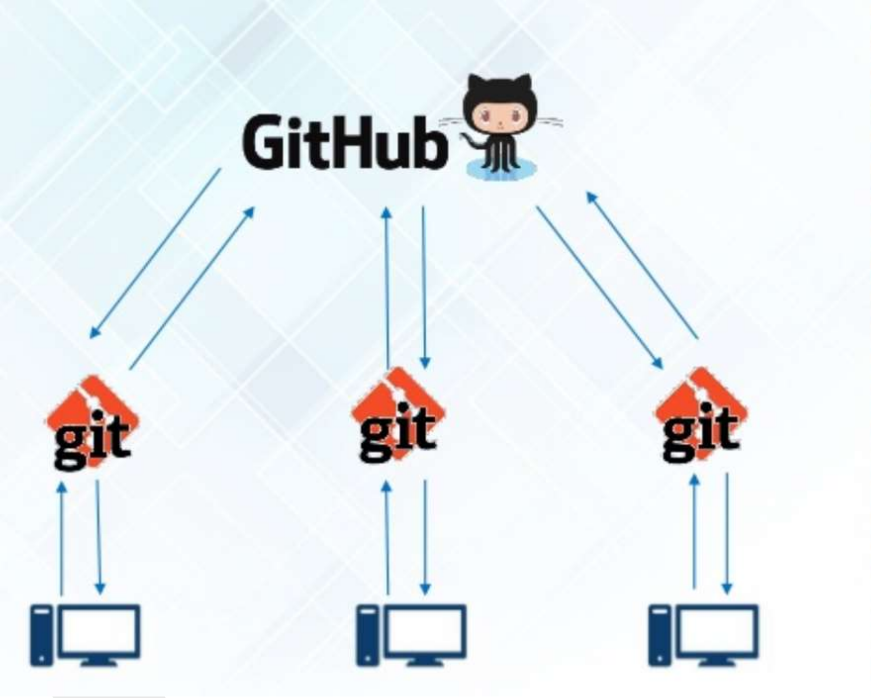
Get started with Github

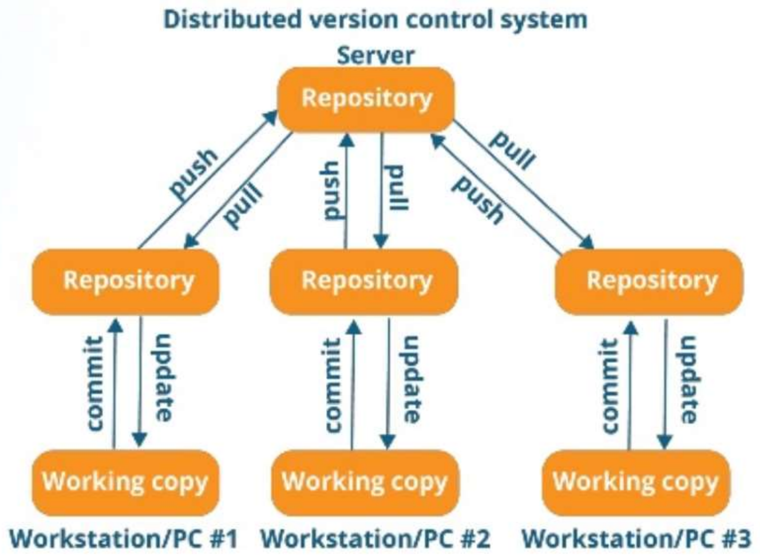
- **Version control System Tools**



Get started with Github

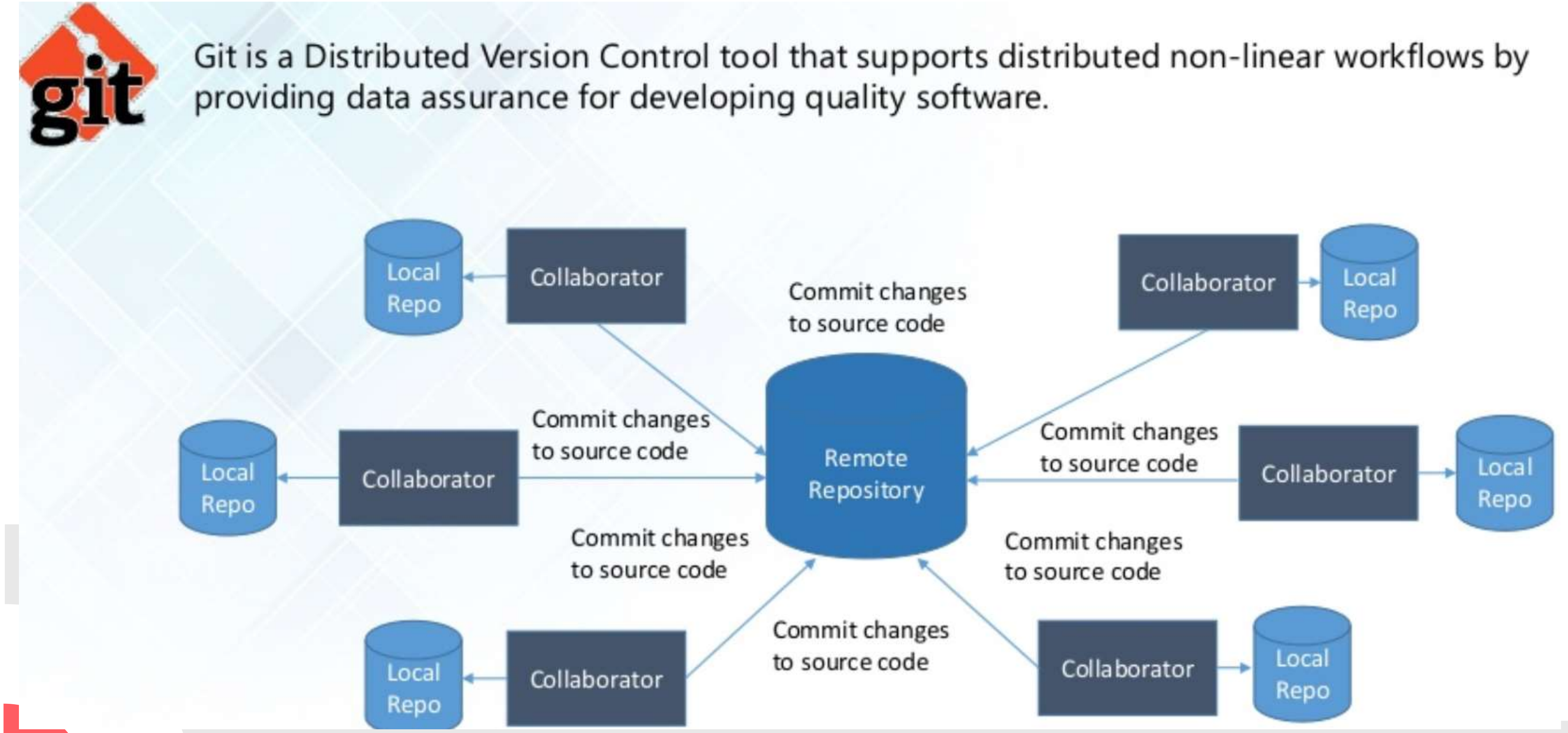
- Git and Github
  - Git: management version control tool; help to create local repository, pull data from the central server and push local data to the server;
  - Github: central repository; code hosting platform for version control collaboration.





Get started with Github

- Overview of Git
  - A repository is a directory or a storage space where your project can live; can be local on your computer or a storage space on Github; It keep every thing related to your project;

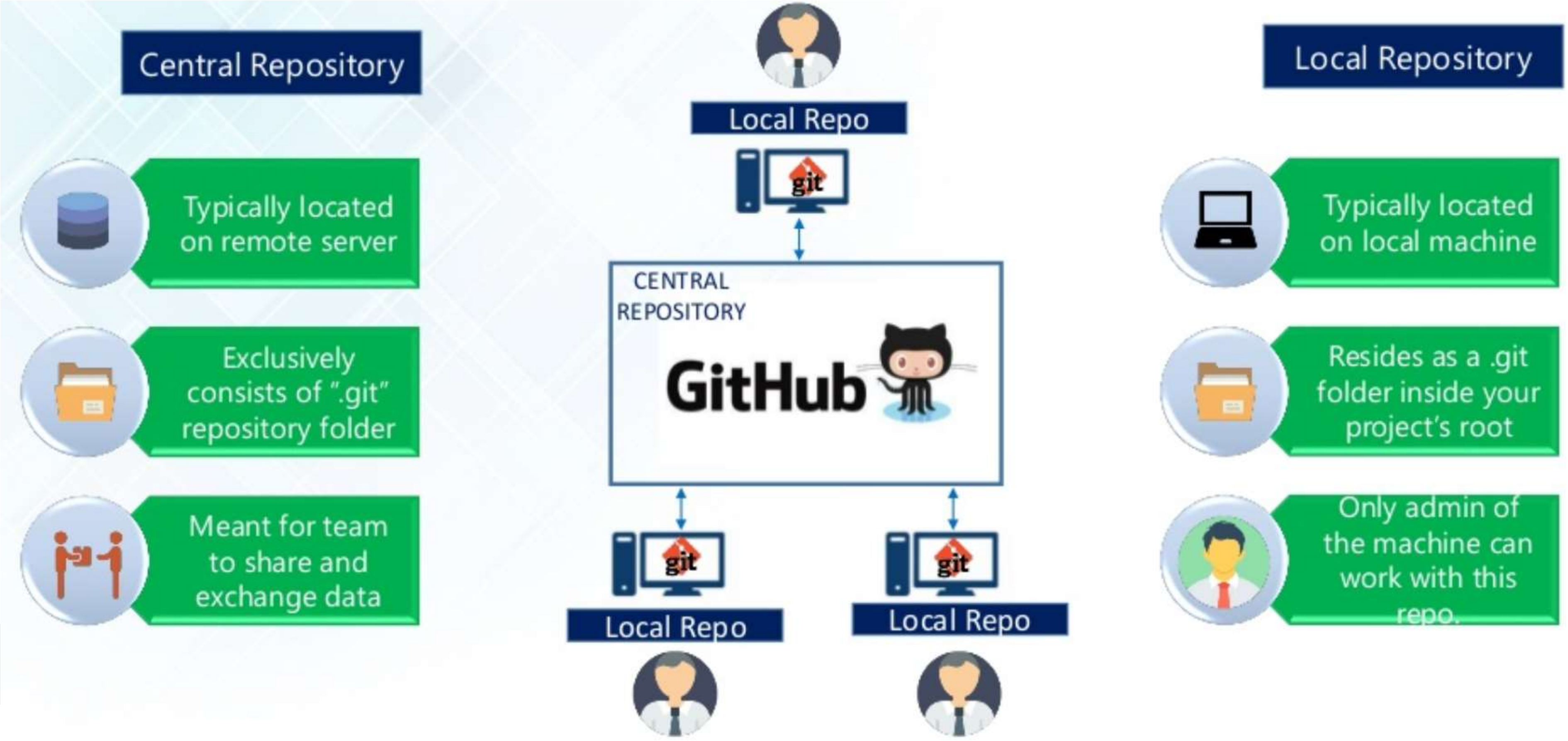


Git is a Distributed Version Control tool that supports distributed non-linear workflows by providing data assurance for developing quality software.

git
Local Repo
Collaborator
Commit changes to source code
Collaborator
Local Repo
Local Repo
Collaborator
Commit changes to source code
Remote Repository
Commit changes to source code
Collaborator
Local Repo
Commit changes to source code
Commit changes to source code
Commit changes to source code
Local Repo
Collaborator
Collaborator
Local Repo

Get started with Github

- **Central and local repository**



Central Repository

Typically located on remote server

Exclusively consists of ".git" repository folder

Meant for team to share and exchange data

Local Repository

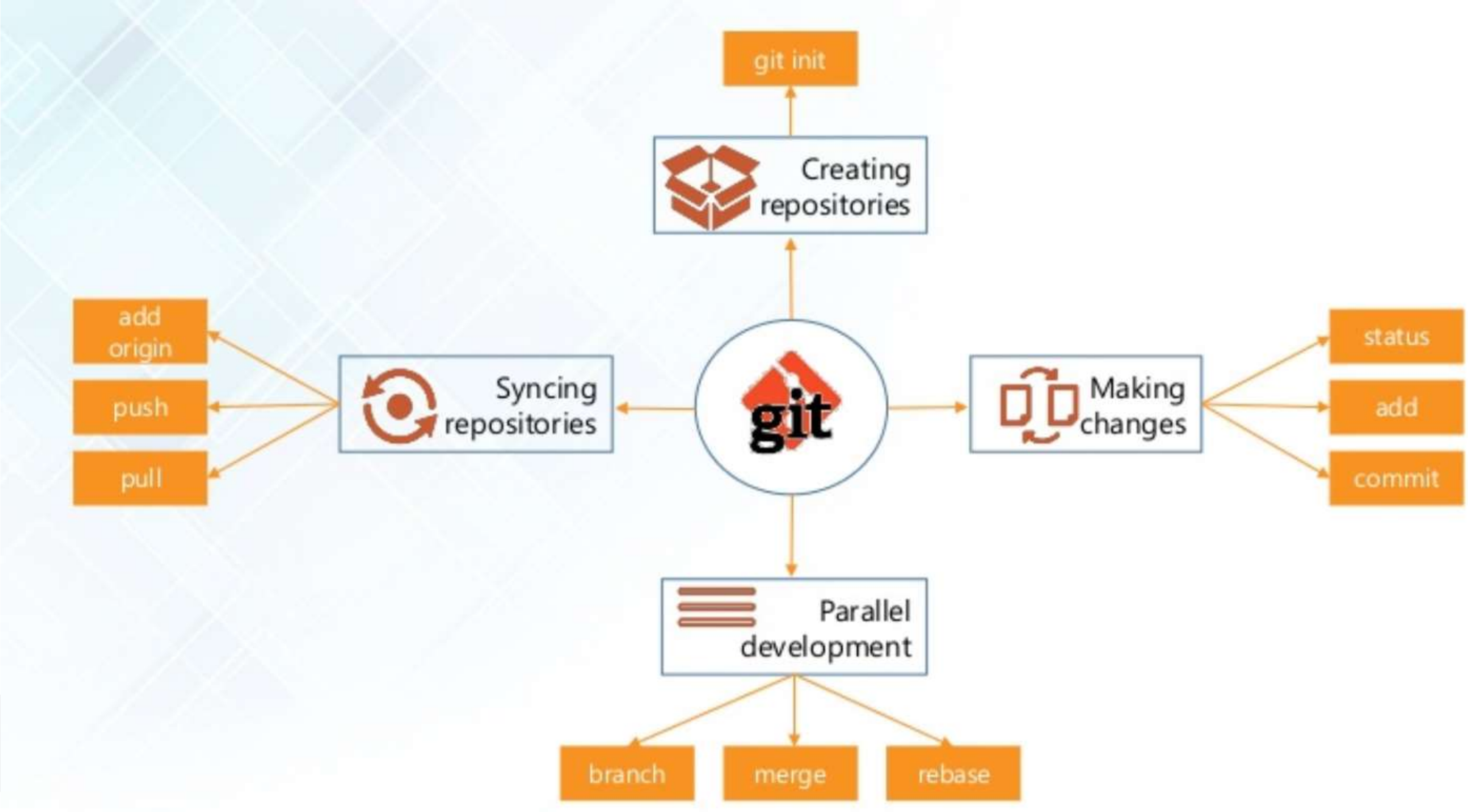
Typically located on local machine

Resides as a .git folder inside your project's root

Only admin of the machine can work with this repo.

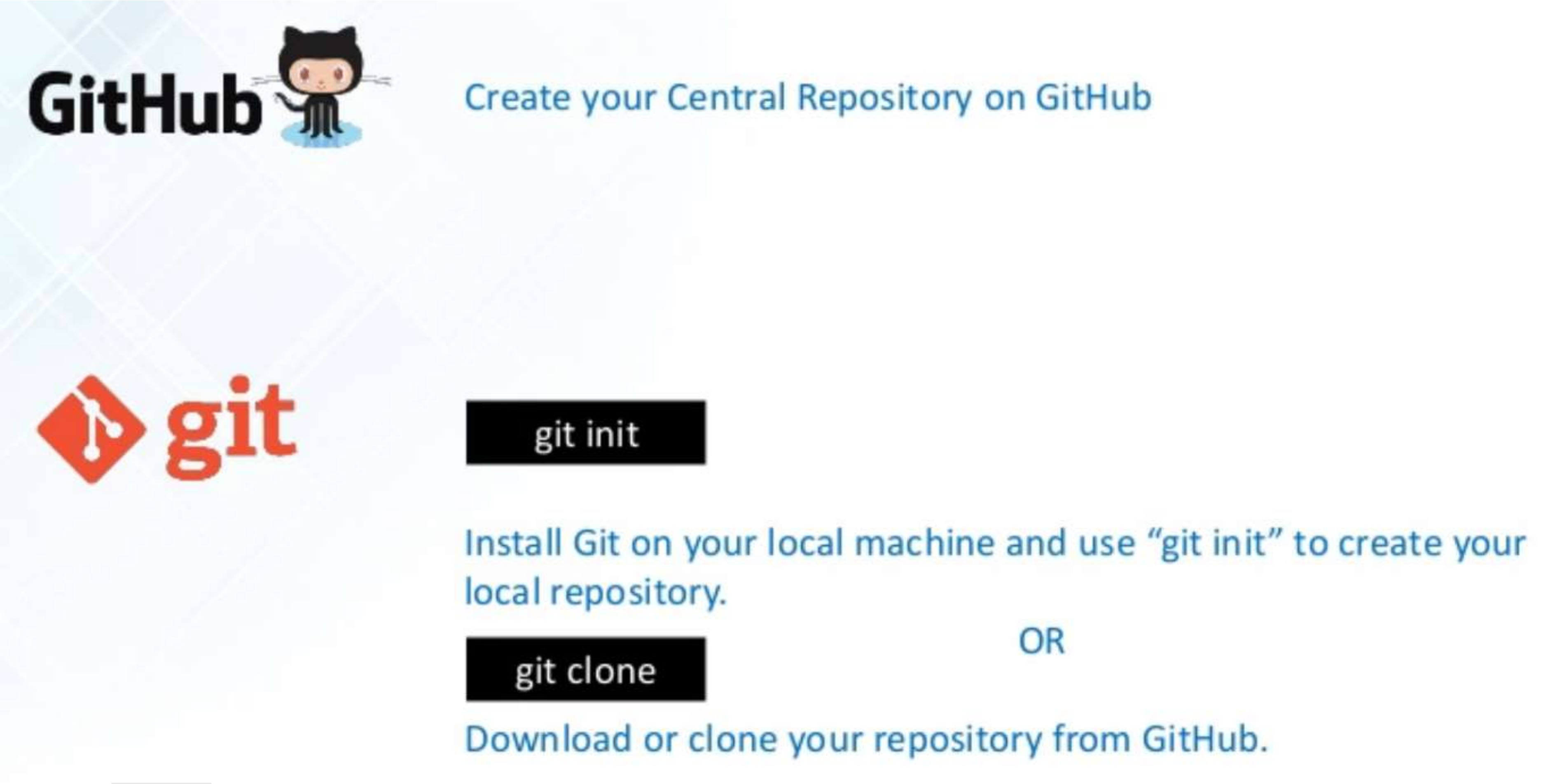
Get started with Github

- **Git Operations and Commands**



Get started with Github

- **Creating Repositories**



Create your Central Repository on GitHub

```
git init
```

Install Git on your local machine and use “git init” to create your local repository.

OR

```
git clone
```

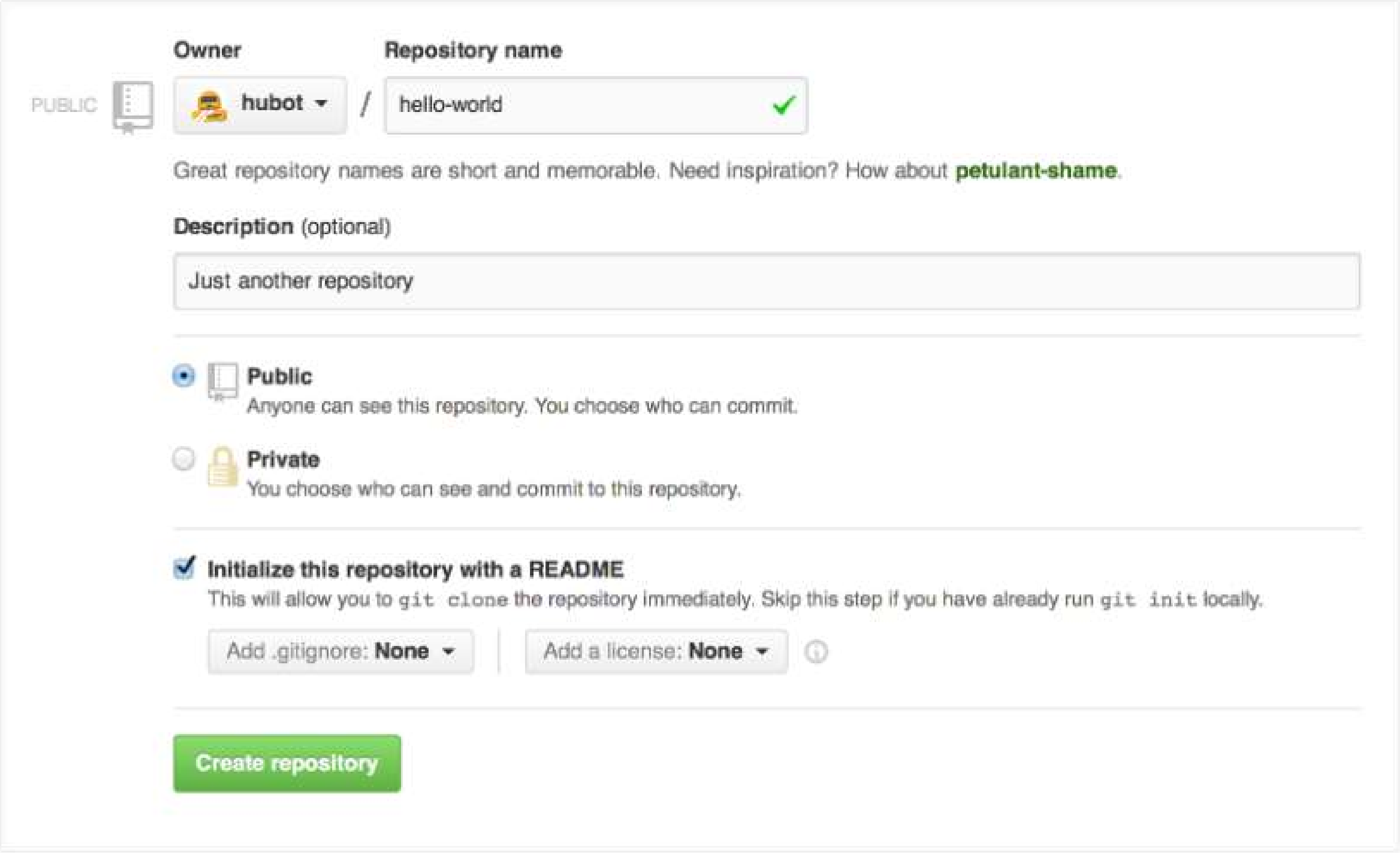
Download or clone your repository from GitHub.

Get started with Github

- **Creating a central Repository on Github**
  - **Create an account on https://github.com/**
  - Create a new repository
    - In the upper right corner, next to your avatar or identicon, click and then select New repository.
    - Name your repository.
    - Write a short description.
    - Select Initialize this repository with a README.

Get started with Github

- **Creating a central Repository on Github**



PUBLIC

Owner: hubot / Repository name: hello-world ✓

Great repository names are short and memorable. Need inspiration? How about **petulant-shame**.

Description (optional)

Just another repository

● **Public**
Anyone can see this repository. You choose who can commit.

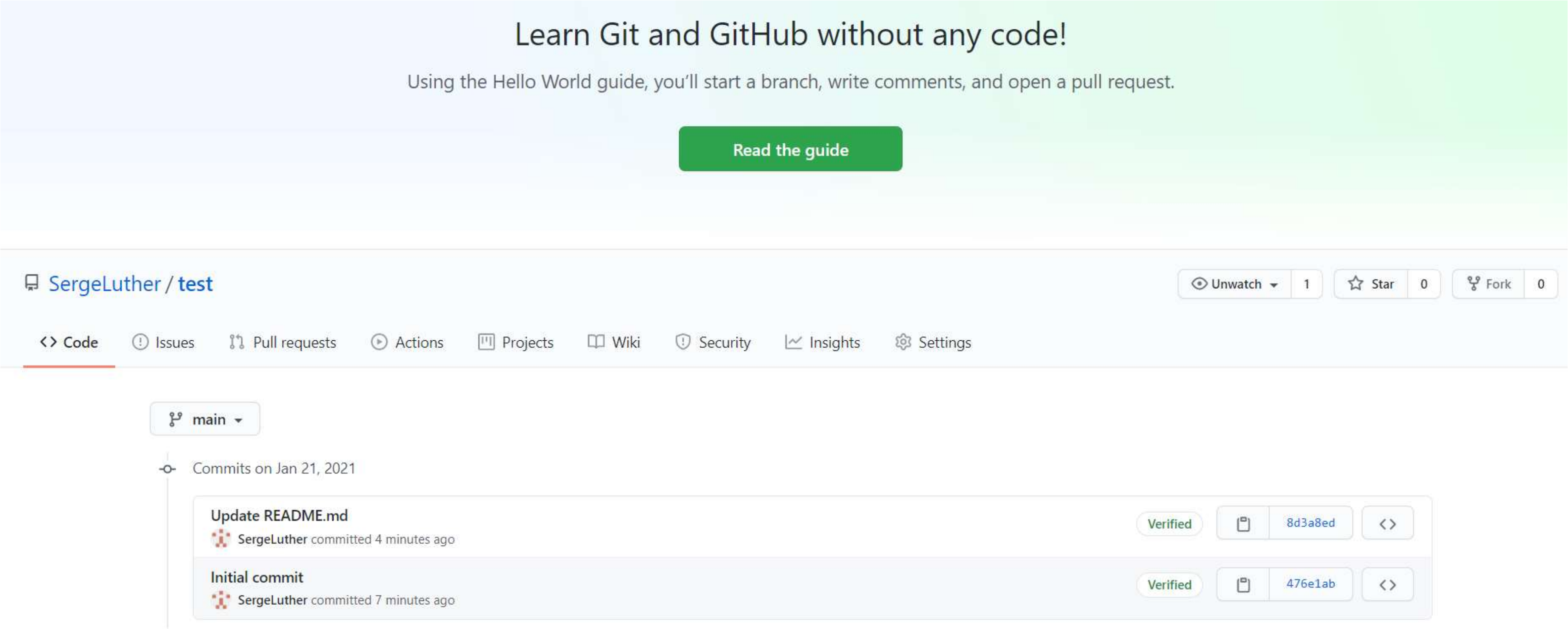
○ **Private**
You choose who can see and commit to this repository.

☑ **Initialize this repository with a README**
This will allow you to `git clone` the repository immediately. Skip this step if you have already run `git init` locally.

Add .gitignore: **None** | Add a license: **None**

Get started with Github

- **Creating a central Repository on Github**



Learn Git and GitHub without any code!

Using the Hello World guide, you'll start a branch, write comments, and open a pull request.

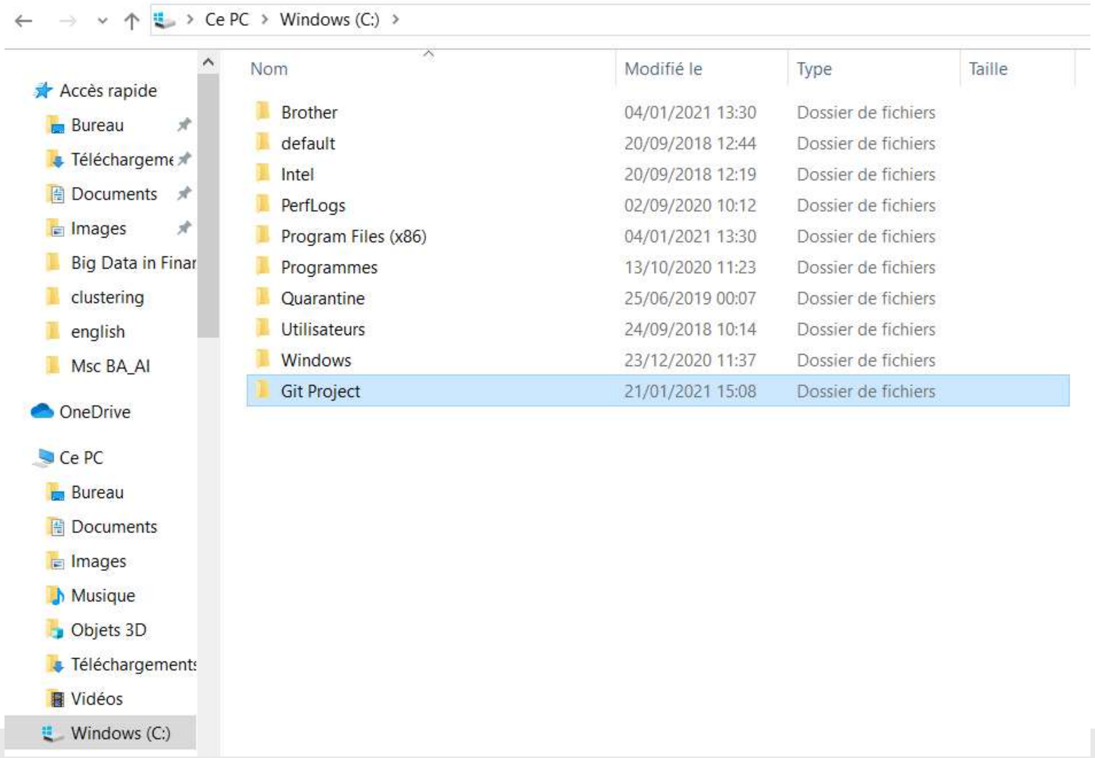
Commits on Jan 21, 2021

Update README.md
SergeLuther committed 4 minutes ago
Verified 8d3a8ed

Initial commit
SergeLuther committed 7 minutes ago
Verified 476e1ab

Get started with Github

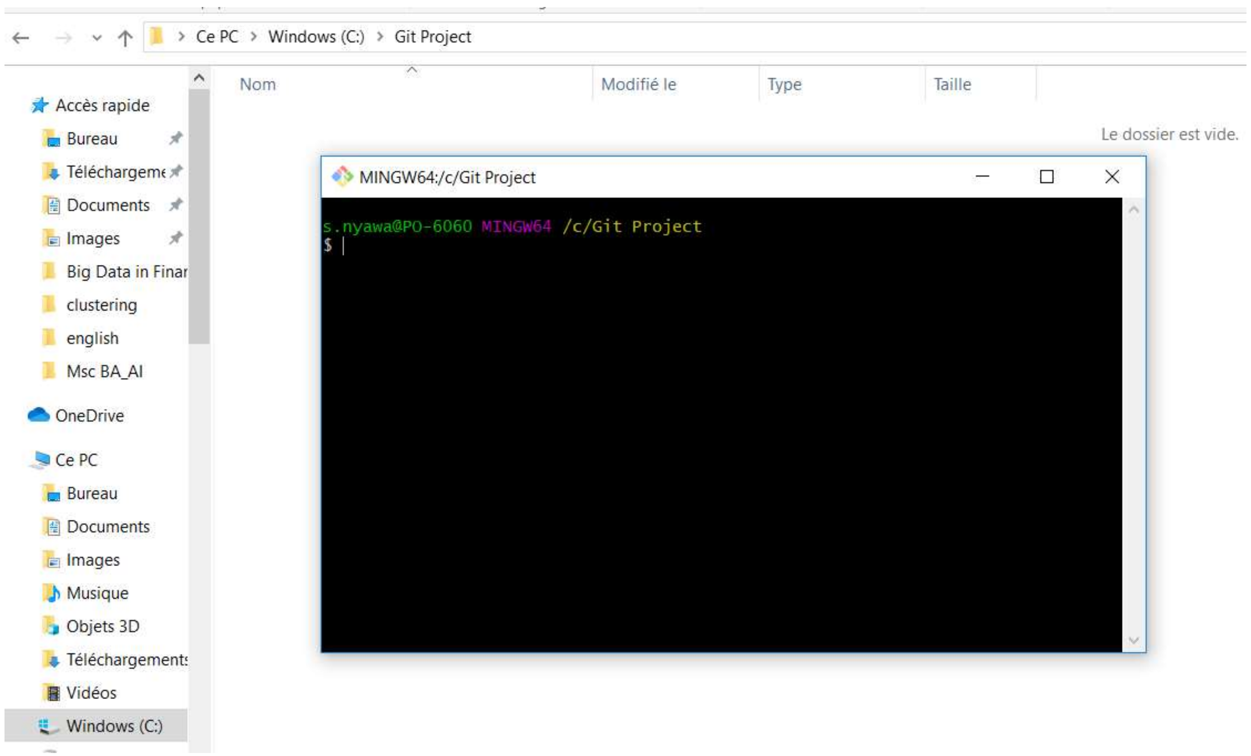
- **Creating a local Repository on my local machine**
  - **Install Git on your computer**
    - **For window: https://git-scm.com/download/win**
    - **For Mac: https://git-scm.com/download/mac**
  - **Create a project file in your "Local Disk (C)"**



Nom	Modifié le	Type	Taille
Brother	04/01/2021 13:30	Dossier de fichiers	
default	20/09/2018 12:44	Dossier de fichiers	
Intel	20/09/2018 12:19	Dossier de fichiers	
PerfLogs	02/09/2020 10:12	Dossier de fichiers	
Program Files (x86)	04/01/2021 13:30	Dossier de fichiers	
Programmes	13/10/2020 11:23	Dossier de fichiers	
Quarantine	25/06/2019 00:07	Dossier de fichiers	
Utilisateurs	24/09/2018 10:14	Dossier de fichiers	
Windows	23/12/2020 11:37	Dossier de fichiers	
Git Project	21/01/2021 15:08	Dossier de fichiers	

Get started with Github

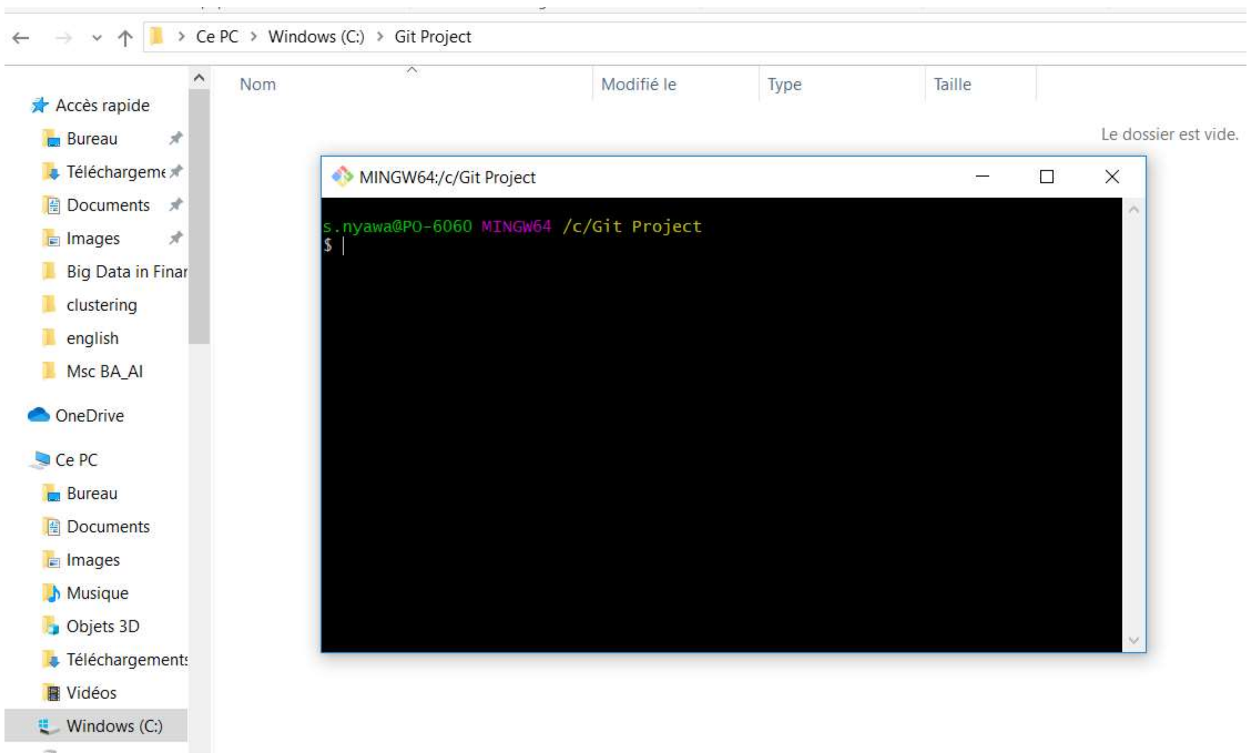
- Open the local created file and right click
- Click on “Git Bach Here”
- A Git Bach Emulator is opened



```
s.nyawa@PO-6060 MINGW64 /c/Git Project
$ |
```

Get started with Github

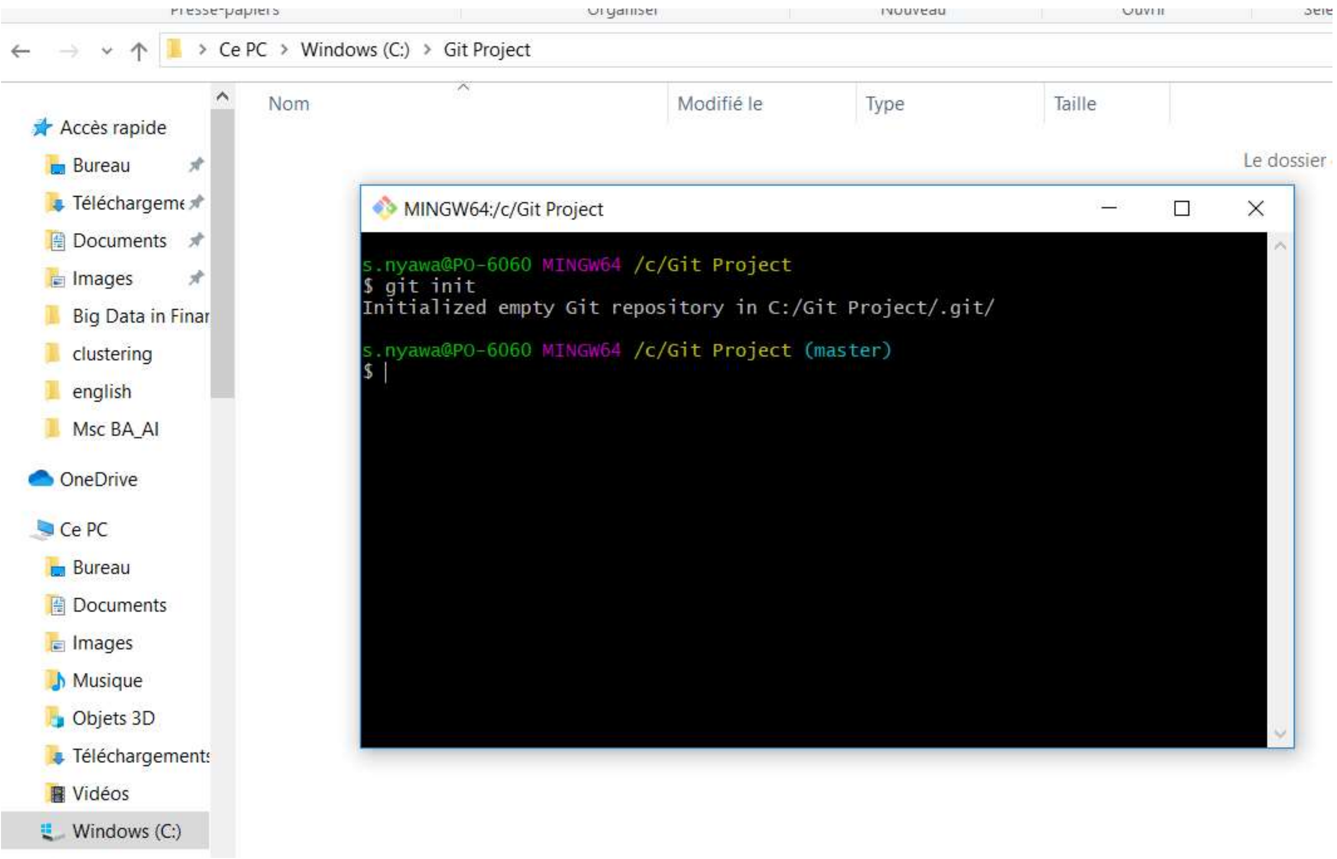
- Open the local created file and right click
- Click on “Git Bach Here”
- A Git Bach Emulator is opened
- Your commands are done in this bach



```
s.nyawa@PO-6060 MINGW64 /c/Git Project
$ |
```

Get started with Github

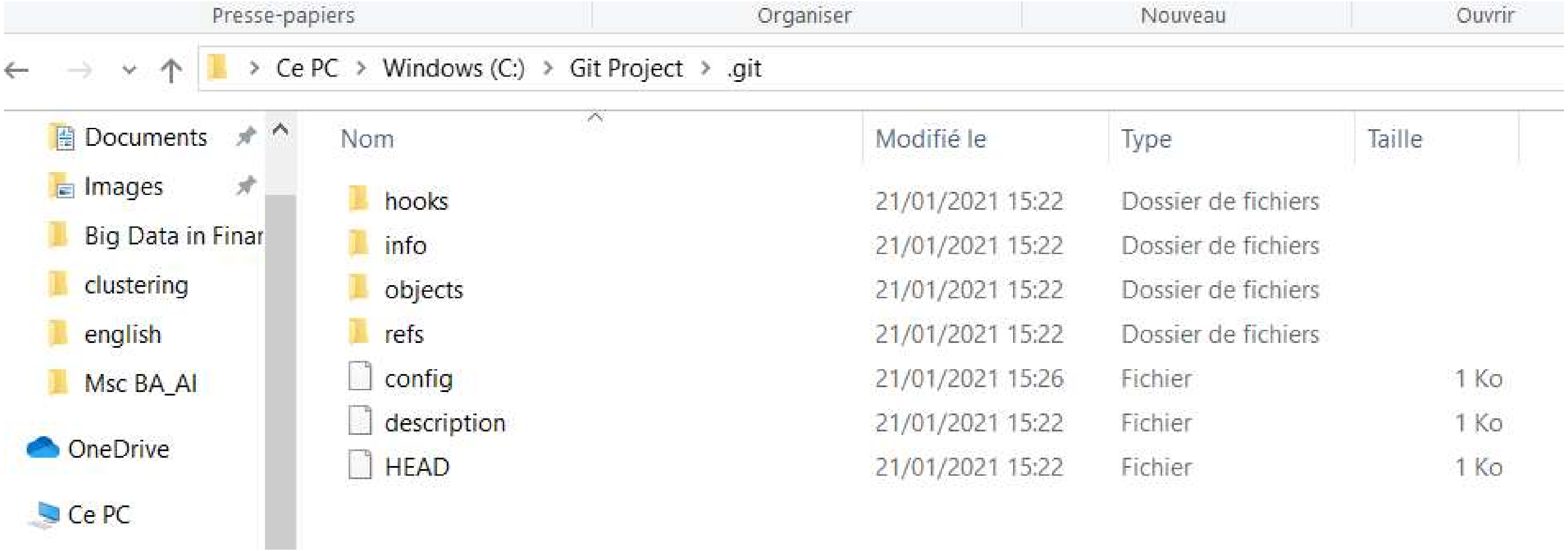
- To create your local repository type in this bach: “git init” and press “enter”



```
s.nyawa@PO-6060 MINGW64 /c/Git Project
$ git init
Initialized empty Git repository in C:/Git Project/.git/

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ |
```

Get started with Github

- **A “.git” folder has been created with all information, objects, etc**



Nom	Modifié le	Type	Taille
hooks	21/01/2021 15:22	Dossier de fichiers	
info	21/01/2021 15:22	Dossier de fichiers	
objects	21/01/2021 15:22	Dossier de fichiers	
refs	21/01/2021 15:22	Dossier de fichiers	
config	21/01/2021 15:26	Fichier	1 Ko
description	21/01/2021 15:22	Fichier	1 Ko
HEAD	21/01/2021 15:22	Fichier	1 Ko

Get started with Github

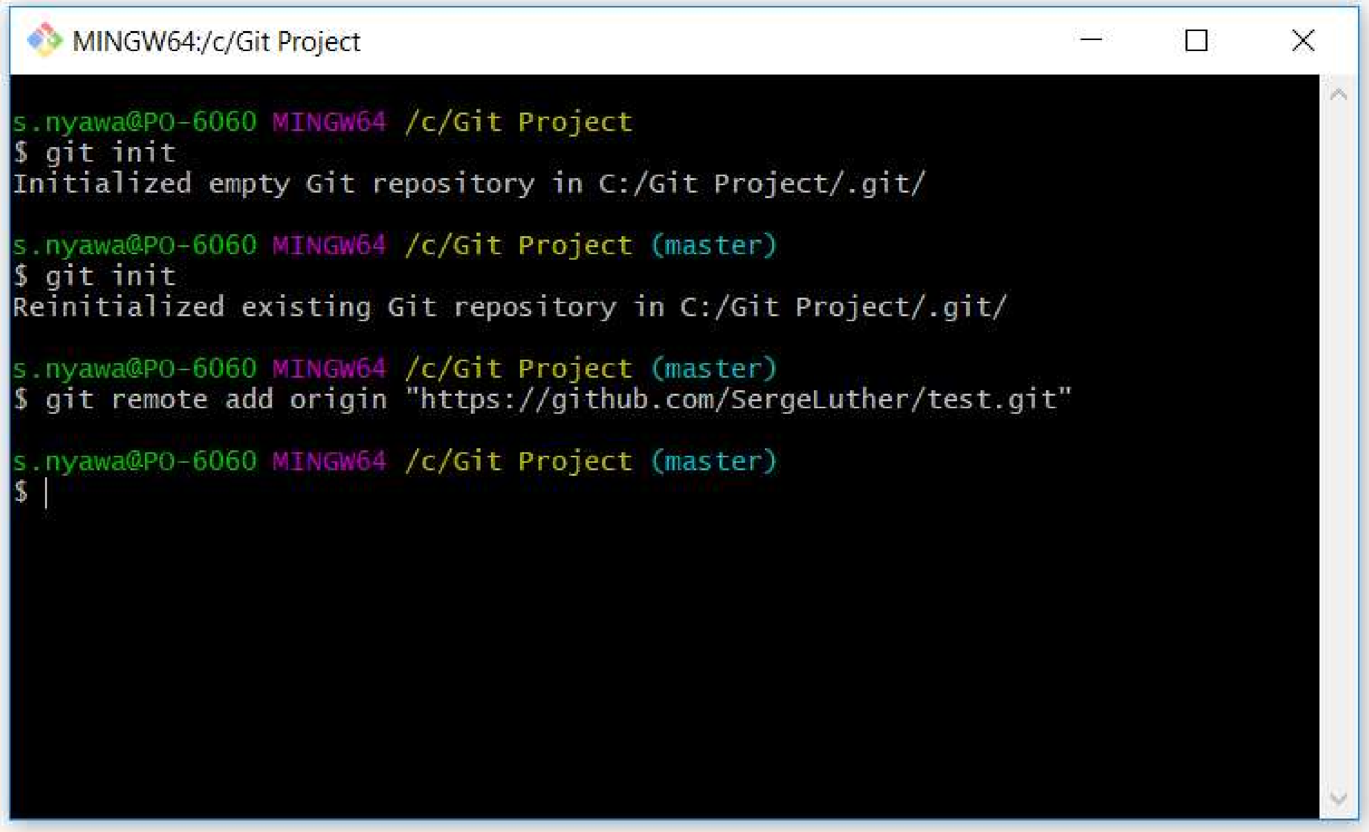
- **Link the local and the central repositories**

- Use ‘**git add origin <link>**’ to add remote repo.
- Pull files with ‘**git pull**’
- Push your own changes into central repo with ‘**git push**’

Get started with Github

- **Link the local and the central repositories**
  - **Add the remote repository as your origin by typing into the bach:**

    git remote add origin “http url of your Github "



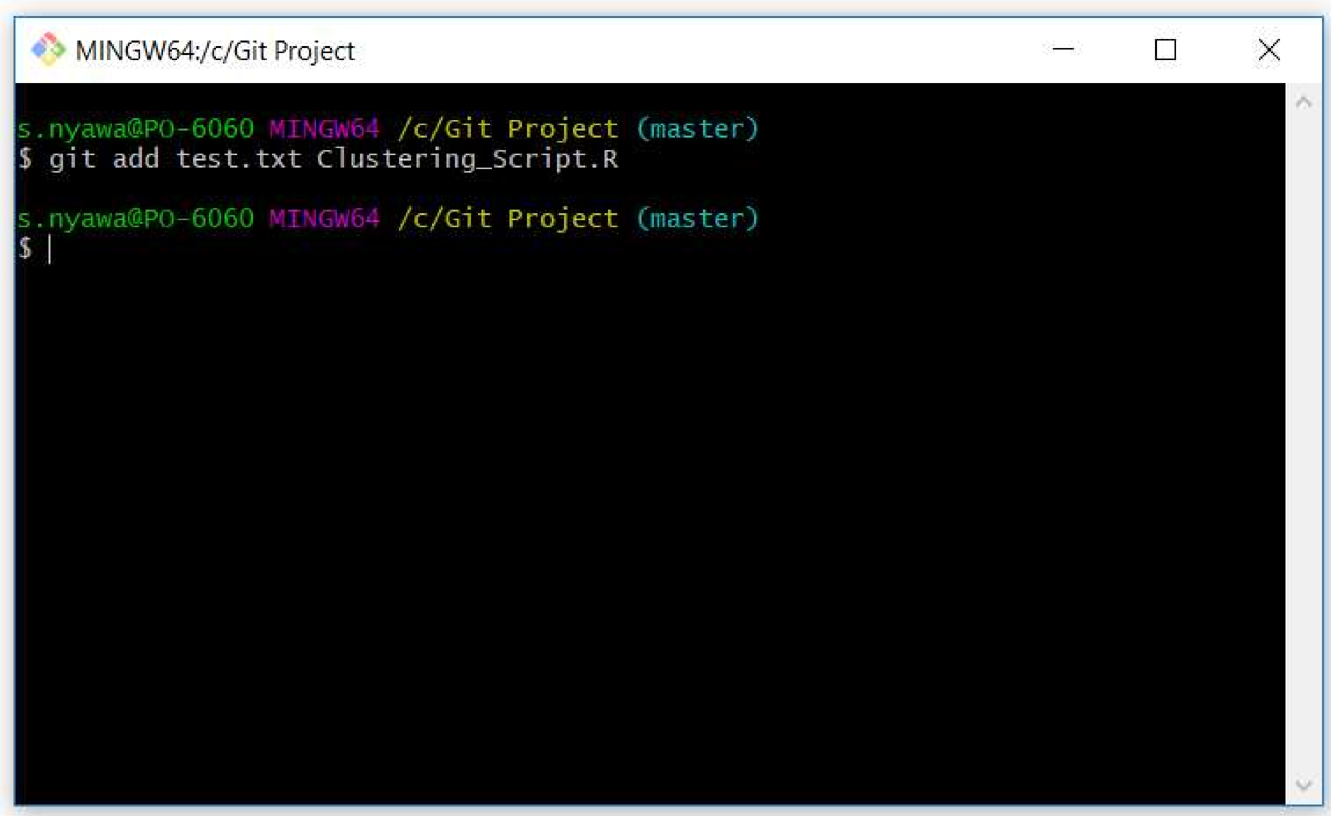
MINGW64:/c/Git Project

```
s.nyawa@PO-6060 MINGW64 /c/Git Project
$ git init
Initialized empty Git repository in C:/Git Project/.git/

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ git init
Reinitialized existing Git repository in C:/Git Project/.git/

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ git remote add origin "https://github.com/SergeLuther/test.git"

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ |
```

Get started with Github

- Pushing files from the local repository to the central repository
  - Add new files to your index by typing into the bach the command:

    git add filename1 filename2

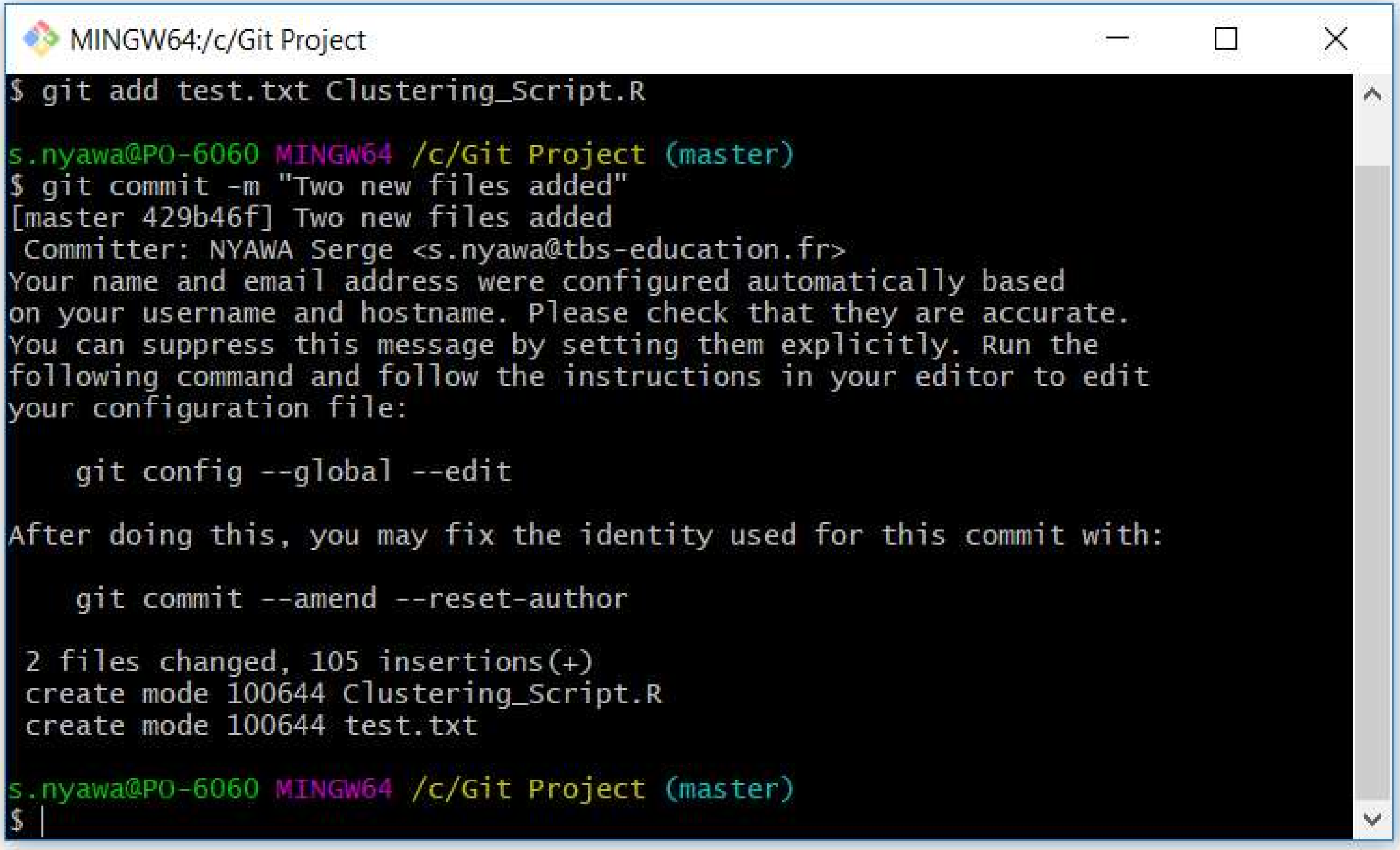


```
s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ git add test.txt Clustering_Script.R

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ |
```

Get started with Github

- **Pushing files from the local repository to the central repository**
  - **Commit the changes into the bach the command:**

    git commit -m “your message”

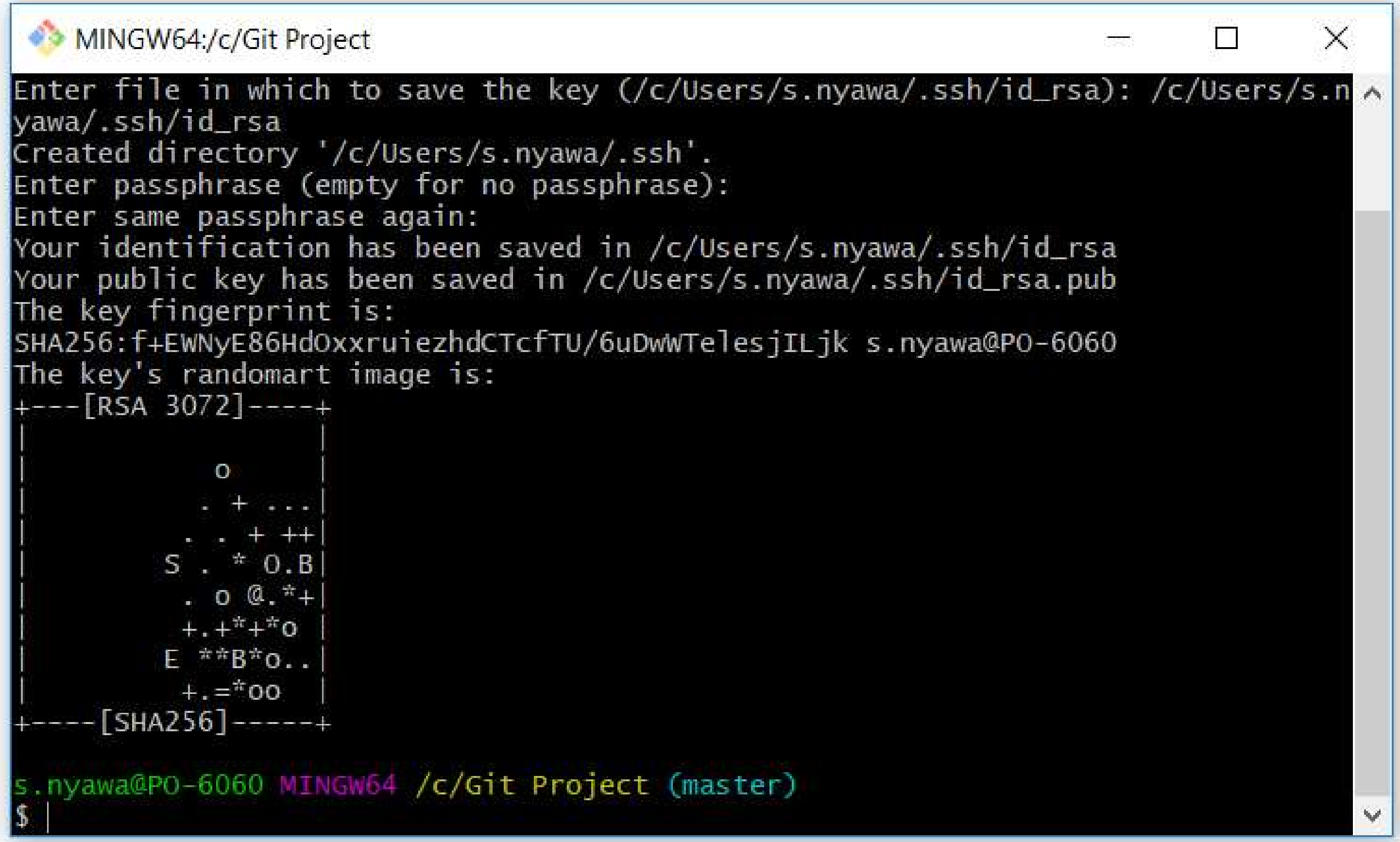


```
$ git add test.txt Clustering_Script.R

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ git commit -m "Two new files added"
[master 429b46f] Two new files added
 Committer: NYAWA Serge <s.nyawa@tbs-education.fr>
Your name and email address were configured automatically based
on your username and hostname. Please check that they are accurate.
You can suppress this message by setting them explicitly. Run the
following command and follow the instructions in your editor to edit
your configuration file:

    git config --global --edit

After doing this, you may fix the identity used for this commit with:

    git commit --amend --reset-author

 2 files changed, 105 insertions(+)
 create mode 100644 Clustering_Script.R
 create mode 100644 test.txt

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ |
```

Get started with Github

- **Pushing files from the local repository to the central repository**
  - **Generate an SSH public key from your bach:**

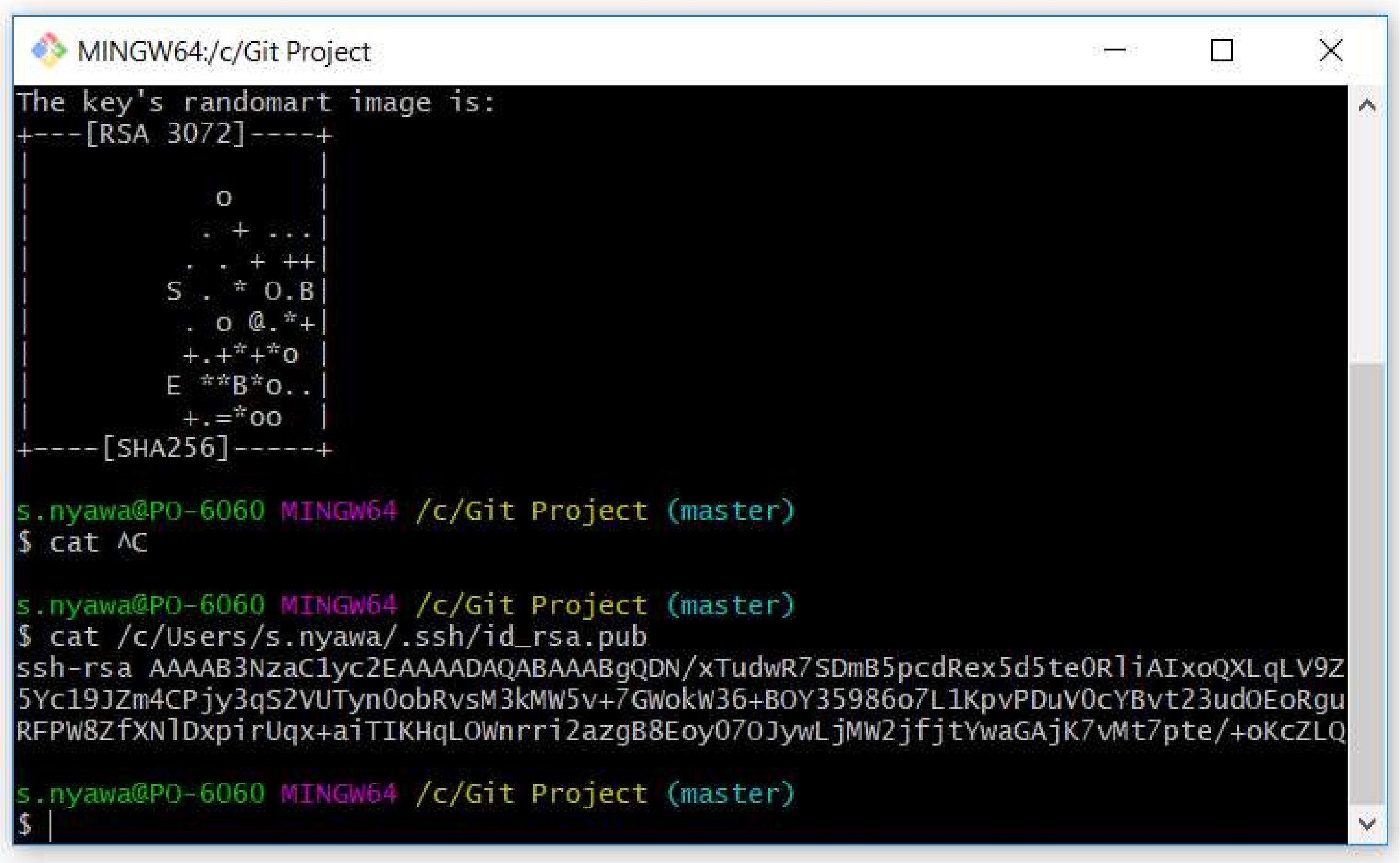
    ssh-keygen



```
Enter file in which to save the key (/c/Users/s.nyawa/.ssh/id_rsa): /c/Users/s.n
yawa/.ssh/id_rsa
Created directory '/c/Users/s.nyawa/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /c/Users/s.nyawa/.ssh/id_rsa
Your public key has been saved in /c/Users/s.nyawa/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:f+EWNyE86HdOxxruiezhdCTcfTU/6uDwWTelesjILjk s.nyawa@PO-6060
The key's randomart image is:
+---[RSA 3072]----+
|                 |
|           o     |
|          . + ...|
|         . . + ++|
|        S . * O.B|
|         . o @.*+|
|         +.+*+*o |
|        E **B*o..|
|         +.=*oo  |
+----[SHA256]-----+

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$
```

Get started with Github

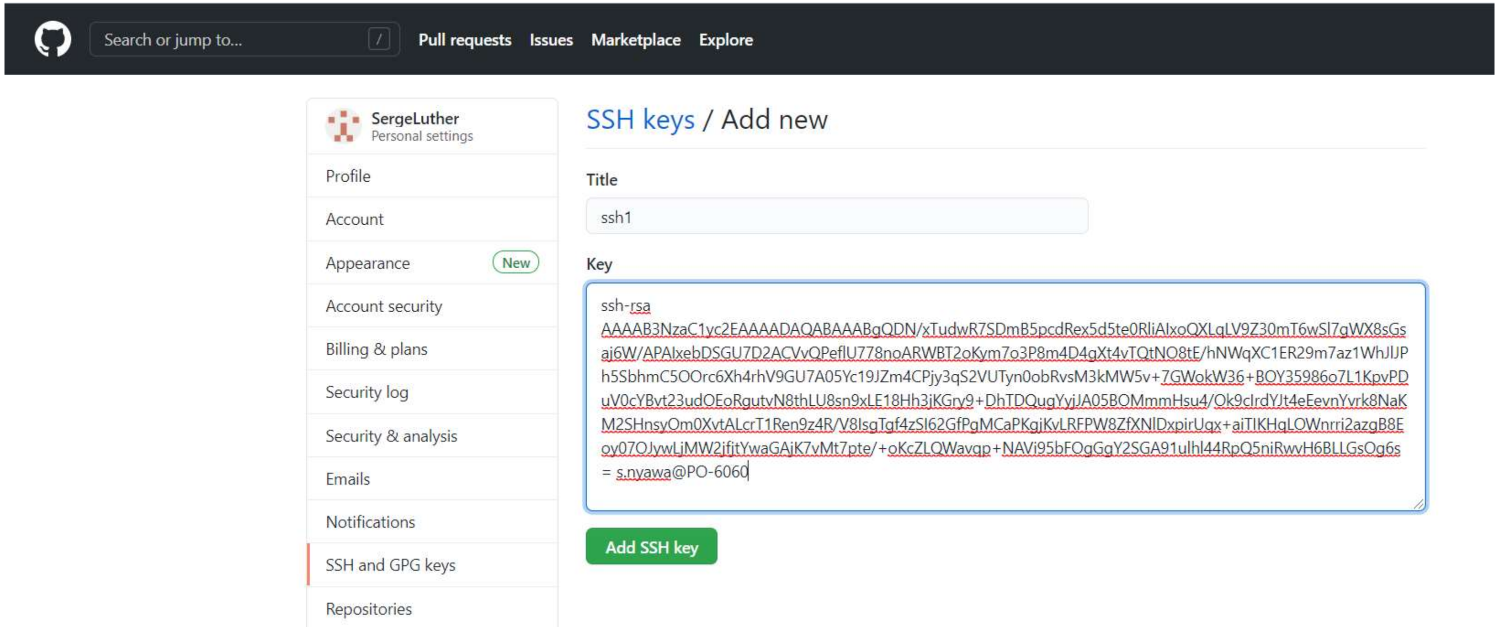
- **Pushing files from the local repository to the central repository**
  - **Print your SSH public key from your bach:**



```
The key's randomart image is:
+---[RSA 3072]----+
|                 |
|           o     |
|          . + ...|
|         . . + ++|
|        S . * O.B|
|         . o @.*+|
|         +.+*+*o |
|        E **B*o..|
|         +.=*oo  |
+----[SHA256]-----+

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ cat ^C

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ cat /c/Users/s.nyawa/.ssh/id_rsa.pub
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQDN/xTudwR7SDmB5pcdRex5d5teORliAIxoQXLqLV9Z
5Yc19JZm4CPjy3qS2VUTyn0obRvsM3kMW5v+7GWokW36+BOY35986o7L1KpvPDuV0cYBvt23udOEoRgu
RFPW8ZfXNlDxpirUqx+aiTIKHqLOWnrri2azgB8Eoy07OJywLjMW2jfjtYwaGAjK7vMt7pte/+oKcZLQ

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ 
```

Get started with Github

- **Pushing files from the local repository to the central repository**
  - **Add your SSH public key in your Github account:**
    Settings>SSH and GPG Keys>New SSH key
  - **Give a name to your key, paste the key code, and add a SSH Key**



SSH keys / Add new

Title

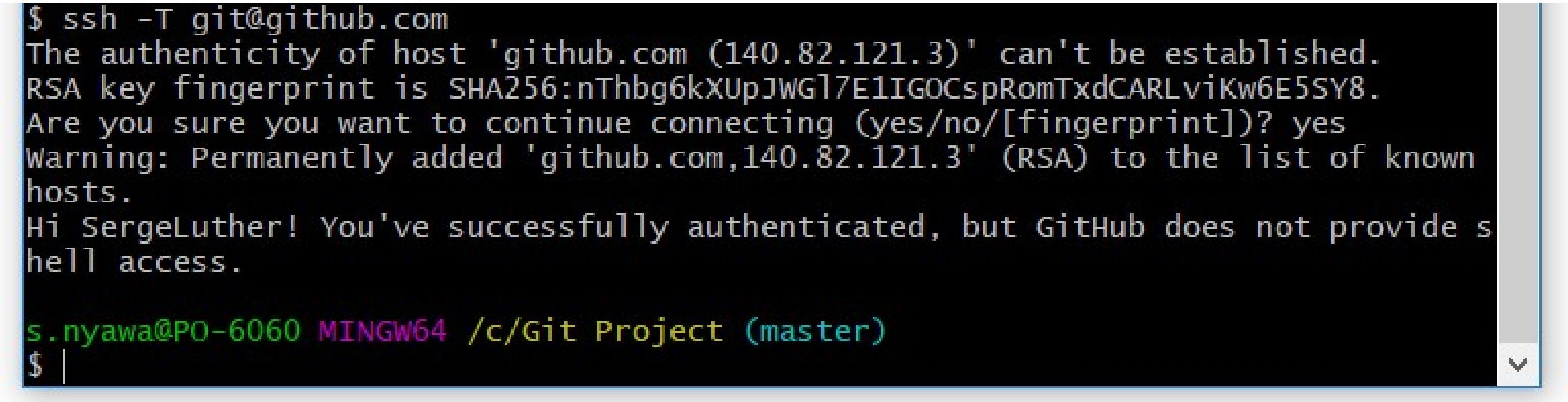
ssh1

Key

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABgQDN/xTudwR7SDmB5pcdRex5d5te0RliAlxoQXLqLV9Z30mT6wSl7gWX8sGs
aj6W/APAlxebDSGU7D2ACVvQPeflU778noARWBT2oKym7o3P8m4D4gXt4vTQtNO8tE/hNWqXC1ER29m7az1WhJlJP
h5SbhmC5OOrc6Xh4rhV9GU7A05Yc19JZm4CPjy3qS2VUTyn0obRvsM3kMW5v+7GWokW36+BOY35986o7L1KpvPD
uV0cYBvt23udOEoRgutvN8thLU8sn9xLE18Hh3jKGry9+DhTDQugYyjJA05BOMmmHsu4/Ok9clrdYJt4eEevnYvrk8NaK
M2SHnsyOm0XvtALcrT1Ren9z4R/V8lsgTgf4zSl62GfPgMCaPKgjKvLRFPW8ZfXNlDxpirUqx+aiTlKHqLOWnrri2azgB8E
oy07OJywLjMW2jfitYwaGAjK7vMt7pte/+oKcZLQWavqp+NAVi95bFOgGgY2SGA91ulhl44RpQ5niRwvH6BLLGsOg6s
= s.nyawa@PO-6060
```

Get started with Github

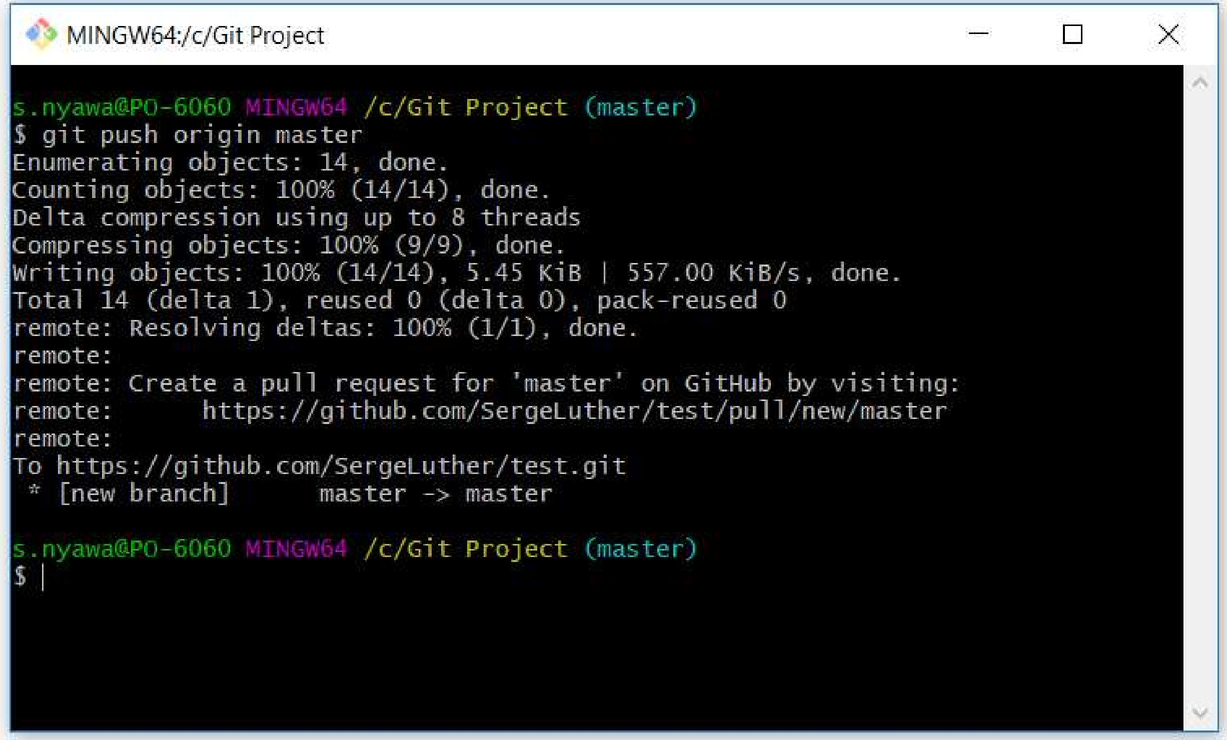
- **Pushing files from the local repository to the central repository**
  - **Authentify your key from your bach:**



```
$ ssh -T git@github.com
The authenticity of host 'github.com (140.82.121.3)' can't be established.
RSA key fingerprint is SHA256:nThbg6kXUpJWGl7E1IGOCspRomTxdCARLviKw6E5SY8.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'github.com,140.82.121.3' (RSA) to the list of known
hosts.
Hi SergeLuther! You've successfully authenticated, but GitHub does not provide s
hell access.

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ 
```

Get started with Github

- **Pushing files from the local repository to the central repository**
  - **Push the files from your bach:**



```
s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ git push origin master
Enumerating objects: 14, done.
Counting objects: 100% (14/14), done.
Delta compression using up to 8 threads
Compressing objects: 100% (9/9), done.
Writing objects: 100% (14/14), 5.45 KiB | 557.00 KiB/s, done.
Total 14 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), done.
remote:
remote: Create a pull request for 'master' on GitHub by visiting:
remote:      https://github.com/SergeLuther/test/pull/new/master
remote:
To https://github.com/SergeLuther/test.git
 * [new branch]      master -> master

s.nyawa@PO-6060 MINGW64 /c/Git Project (master)
$ |
```

tbs Think & Create

Toulouse
Business School

www.tbs-education.fr