Version Control (Git)

Version control systems (VCSs) are tools used to track changes to source code (or other collections of files and folders). As the name implies, these tools help maintain a history of changes; furthermore, they facilitate collaboration. VCSs track changes to a folder and its contents in a series of snapshots, where each snapshot encapsulates the entire state of files/folders within a top-level directory. VCSs also maintain metadata like who created each snapshot, messages associated with each snapshot, and so on.

Why is version control useful? Even when you're working by yourself, it can let you look at old snapshots of a project, keep a log of why certain changes were made, work on parallel branches of development, and much more. When working with others, it's an invaluable tool for seeing what other people have changed, as well as resolving conflicts in concurrent development.

Modern VCSs also let you easily (and often automatically) answer questions like:

- Who wrote this module?
- When was this particular line of this particular file edited? By whom? Why was it edited?
- Over the last 1000 revisions, when/why did a particular unit test stop working?

While other VCSs exist, **Git** is the de facto standard for version control. This XKCD comic captures Git's reputation:

Because Git's interface is a leaky abstraction, learning Git top-down (starting with its interface / command-line interface) can lead to a lot of confusion. It's possible to memorize a handful of commands and think of them as magic incantations, and follow the approach in the comic above whenever anything goes wrong.

While Git admittedly has an ugly interface, its underlying design and ideas are beautiful. While an ugly interface has to be *memorized*, a beautiful design can be *understood*. For this reason, we give a bottom-up explanation of Git, starting with its data model and later covering the command-line interface. Once the data model is understood, the commands can be better understood in terms of how they manipulate the underlying data model.

Git's data model

There are many ad-hoc approaches you could take to version control. Git has a well-thought-out model that enables all the nice features of version control, like maintaining history, supporting branches, and enabling collaboration.

Snapshots

Git models the history of a collection of files and folders within some top-level directory as a series of snapshots. In Git terminology, a file is called a "blob", and it's just a bunch of bytes. A directory is called a "tree", and it maps names to blobs or trees (so directories can contain other directories). A snapshot is the top-level tree that is being tracked. For example, we might have a tree as follows:

```
<root> (tree)
|
+- foo (tree)
|  |
|  + bar.txt (blob, contents = "hello world")
|
+- baz.txt (blob, contents = "git is wonderful")
```

The top-level tree contains two elements, a tree "foo" (that itself contains one element, a blob "bar.txt"), and a blob "baz.txt".

Modeling history: relating snapshots

How should a version control system relate snapshots? One simple model would be to have a linear history. A history would be a list of snapshots in time-order. For many reasons, Git doesn't use a simple model like this.

In Git, a history is a directed acyclic graph (DAG) of snapshots. That may sound like a fancy math word, but don't be intimidated. All this means is that each snapshot in Git refers to a set of "parents", the snapshots that preceded it. It's a set of parents rather than a single parent (as would be the case in a linear history) because a snapshot might descend from multiple parents, for example, due to combining (merging) two parallel branches of development.

Git calls these snapshots "commit"s. Visualizing a commit history might look something like this:

```
o <-- o <-- o <-- o
            ^
             \
              --- o <-- o
```

In the ASCII art above, the `o`s correspond to individual commits (snapshots). The arrows point to the parent of each commit (it's a "comes before" relation, not "comes after"). After the third commit, the history branches into two separate branches. This might correspond to, for example, two separate features being developed in parallel, independently from each other. In the future, these branches may be merged to create a

new snapshot that incorporates both of the features, producing a new history that looks like this, with the newly created merge commit shown in bold:

Commits in Git are immutable. This doesn't mean that mistakes can't be corrected, however; it's just that "edits" to the commit history are actually creating entirely new commits, and references (see below) are updated to point to the new ones.

Data model, as pseudocode

It may be instructive to see Git's data model written down in pseudocode:

```
// a file is a bunch of bytes
type blob = array<byte>

// a directory contains named files and directories
type tree = map<string, tree | blob>

// a commit has parents, metadata, and the top-level tree
type commit = struct {
    parents: array<commit>
    author: string
    message: string
    snapshot: tree
}
```

It's a clean, simple model of history.

Objects and content-addressing

An "object" is a blob, tree, or commit:

```
type object = blob | tree | commit
```

In Git data store, all objects are content-addressed by their SHA-1 hash.

```
objects = map<string, object>

def store(object):
    id = sha1(object)
    objects[id] = object

def load(id):
    return objects[id]
```

Blobs, trees, and commits are unified in this way: they are all objects. When they reference other objects, they don't actually *contain* them in their on-disk representation, but have a reference to them by their hash.

For example, the tree for the example directory structure above (visualized using `git cat-file -p 698281bc680d1995c5f4caaf3359721a5a58d48d`), looks like this:

```
100644 blob 4448adbf7ecd394f42ae135bbeed9676e894af85    baz.txt
040000 tree c68d233a33c5c06e0340e4c224f0afca87c8ce87    foo
```

The tree itself contains pointers to its contents, `baz.txt` (a blob) and `foo` (a tree). If we look at the contents addressed by the hash corresponding to baz.txt with `git cat-file -p 4448adbf7ecd394f42ae135bbeed9676e894af85`, we get the following:

```
git is wonderful
```

References

Now, all snapshots can be identified by their SHA-1 hashes. That's inconvenient, because humans aren't good at remembering strings of 40 hexadecimal characters.

Git's solution to this problem is human-readable names for SHA-1 hashes, called "references". References are pointers to commits. Unlike objects, which are immutable, references are mutable (can be updated to point to a new commit). For example, the `master` reference usually points to the latest commit in the main branch of development.

```
references = map<string, string>

def update_reference(name, id):
    references[name] = id

def read_reference(name):
    return references[name]

def load_reference(name_or_id):
    if name_or_id in references:
        return load(references[name_or_id])
    else:
        return load(name_or_id)
```

With this, Git can use human-readable names like "master" to refer to a particular snapshot in the history, instead of a long hexadecimal string.

One detail is that we often want a notion of "where we currently are" in the history, so that when we take a new snapshot, we know what it is relative to (how we set the `parents` field of the commit). In Git, that "where we currently are" is a special reference called "HEAD".

Repositories

Finally, we can define what (roughly) is a Git *repository*: it is the data `objects` and `references`.

On disk, all Git stores are objects and references: that's all there is to Git's data model. All `git` commands map to some manipulation of the commit DAG by adding objects and adding/updating references.

Whenever you're typing in any command, think about what manipulation the command is making to the underlying graph data structure. Conversely, if you're trying to make a particular kind of change to the commit DAG, e.g. "discard uncommitted changes and make the 'master' ref point to commit `5d83f9e`", there's probably a command to do it (e.g. in this case, `git checkout master; git reset --hard 5d83f9e`).

Staging area

This is another concept that's orthogonal to the data model, but it's a part of the interface to create commits.

One way you might imagine implementing snapshotting as described above is to have a "create snapshot" command that creates a new snapshot based on the *current state* of the working directory. Some version control tools work like this, but not Git. We want clean snapshots, and it might not always be ideal to make a snapshot from the current

state. For example, imagine a scenario where you've implemented two separate features, and you want to create two separate commits, where the first introduces the first feature, and the next introduces the second feature. Or imagine a scenario where you have debugging print statements added all over your code, along with a bugfix; you want to commit the bugfix while discarding all the print statements.

Git accommodates such scenarios by allowing you to specify which modifications should be included in the next snapshot through a mechanism called the "staging area".

Git command-line interface

To avoid duplicating information, we're not going to explain the commands below in detail. See the highly recommended Pro Git for more information, or watch the lecture video.

Basics

- `git help <command>`: get help for a git command
- `git init`: creates a new git repo, with data stored in the `.git` directory
- `git status`: tells you what's going on
- `git add <filename>`: adds files to staging area
- `git commit`: creates a new commit
  - Write good commit messages!
  - Even more reasons to write good commit messages!
- `git log`: shows a flattened log of history
- `git log --all --graph --decorate`: visualizes history as a DAG
- `git diff <filename>`: show changes you made relative to the staging area
- `git diff <revision> <filename>`: shows differences in a file between snapshots
- `git checkout <revision>`: updates HEAD and current branch

Branching and merging

- `git branch`: shows branches
- `git branch <name>`: creates a branch
- `git checkout -b <name>`: creates a branch and switches to it
  - same as `git branch <name>; git checkout <name>`
- `git merge <revision>`: merges into current branch
- `git mergetool`: use a fancy tool to help resolve merge conflicts
- `git rebase`: rebase set of patches onto a new base

Remotes

- `git remote`: list remotes
- `git remote add <name> <url>`: add a remote

- `git push <remote> <local branch>:<remote branch>`: send objects to remote, and update remote reference
- `git branch --set-upstream-to=<remote>/<remote branch>`: set up correspondence between local and remote branch
- `git fetch`: retrieve objects/references from a remote
- `git pull`: same as `git fetch; git merge`
- `git clone`: download repository from remote

Undo

- `git commit --amend`: edit a commit's contents/message
- `git reset HEAD <file>`: unstage a file
- `git checkout -- <file>`: discard changes

Advanced Git

- `git config`: Git is highly customizable
- `git clone --depth=1`: shallow clone, without entire version history
- `git add -p`: interactive staging
- `git rebase -i`: interactive rebasing
- `git blame`: show who last edited which line
- `git stash`: temporarily remove modifications to working directory
- `git bisect`: binary search history (e.g. for regressions)
- `.gitignore`: specify intentionally untracked files to ignore

Miscellaneous

- **GUIs**: there are many GUI clients out there for Git. We personally don't use them and use the command-line interface instead.
- **Shell integration**: it's super handy to have a Git status as part of your shell prompt (zsh, bash). Often included in frameworks like Oh My Zsh.
- **Editor integration**: similarly to the above, handy integrations with many features. fugitive.vim is the standard one for Vim.
- **Workflows**: we taught you the data model, plus some basic commands; we didn't tell you what practices to follow when working on big projects (and there are many different approaches).
- **GitHub**: Git is not GitHub. GitHub has a specific way of contributing code to other projects, called pull requests.
- **Other Git providers**: GitHub is not special: there are many Git repository hosts, like GitLab and BitBucket.

Resources

- Pro Git is **highly recommended reading**. Going through Chapters 1–5 should teach you most of what you need to use Git proficiently, now that you understand the data model. The later chapters have some interesting, advanced material.
- Oh Shit, Git!?! is a short guide on how to recover from some common Git mistakes.
- Git for Computer Scientists is a short explanation of Git's data model, with less pseudocode and more fancy diagrams than these lecture notes.
- Git from the Bottom Up is a detailed explanation of Git's implementation details beyond just the data model, for the curious.
- How to explain git in simple words
- Learn Git Branching is a browser-based game that teaches you Git.

Exercises

1. If you don't have any past experience with Git, either try reading the first couple chapters of Pro Git or go through a tutorial like Learn Git Branching. As you're working through it, relate Git commands to the data model.
2. Clone the repository for the class website.
   1. Explore the version history by visualizing it as a graph.
   2. Who was the last person to modify `README.md`? (Hint: use `git log` with an argument).
   3. What was the commit message associated with the last modification to the `collections:` line of `_config.yml`? (Hint: use `git blame` and `git show`).
3. One common mistake when learning Git is to commit large files that should not be managed by Git or adding sensitive information. Try adding a file to a repository, making some commits and then deleting that file from history (you may want to look at this).
4. Clone some repository from GitHub, and modify one of its existing files. What happens when you do `git stash`? What do you see when running `git log --all --oneline`? Run `git stash pop` to undo what you did with `git stash`. In what scenario might this be useful?
5. Like many command line tools, Git provides a configuration file (or dotfile) called `~/.gitconfig`. Create an alias in `~/.gitconfig` so that when you run `git graph`, you get the output of `git log --all --graph --decorate --oneline`. Information about git aliases can be found here.
6. You can define global ignore patterns in `~/.gitignore_global` after running `git config --global core.excludesfile ~/.gitignore_global`. Do this, and set up your global gitignore file to ignore OS-specific or editor-specific temporary files, like `.DS_Store`.
7. Fork the repository for the class website, find a typo or some other improvement you can make, and submit a pull request on GitHub (you may want to look at this).