## The Git Object Database

### Git Objects

It is important to note that this is very different from most SCM systems that you may be familiar with. Subversion, CVS, Perforce, Mercurial and the like all use Delta Storage systems - they store the differences between one commit and the next. Git does not do this - it stores a snapshot of what all the files in your project look like in this tree structure each time you commit. This is a very important concept to understand when using Git.

Now it's time to investigate another well-known difference between Git and other versioning systems. Take Subversion as an example: when you do a new commit, Subversion creates a new numbered revision that only contains deltas between the previous one; this is a smart way to archive changes to files, especially among big text files, because if only a line of text changes, the size of the new commit will be much smaller. Instead, in Git even if you change only a char in a big text file, it always stores a new version of the file: **Git doesn't do deltas** (at least not in this case), and **every commit is actually a snapshot of the entire repository**.

The data model of Git is different from other common **version control systems** (**VCSs**) in the way Git handles its data. Traditionally, a VCS will store its data as an initial file, followed by a list of patches for each new version of the file. Git is different: Instead of the regular file and patches list, Git records a snapshot of all the files tracked by Git and their paths relative to the repository root—that is, the files tracked by Git in the filesystem tree. Each commit in Git records the full tree state. If a file does not change between commits, Git will not store the file again.

Git is a version control system built on top of a key *value object store*. Git creates and stores a collection of objects when you commit. The object store is stored inside the Git *repository*. It exists entirely in a single .git directory in your project root. There is no central repository like in Subversion. The key is an SHA-1 hash of the object an the value is the object itself. The SHA1 hash is a cryptographic hash function. This *code*, or **hash**, as it is usually called, uniquely identifies the commit within the repository. What that means to us is that it is virtually impossible to find two different objects with the same name. This has a number of advantages; among others:

* Git can quickly determine whether two objects are identical or not, just by comparing names.
* Since object names are computed the same way in every repository, the same content stored in two repositories will always be stored under the same name.
* Git can detect errors when it reads an object, by checking that the object's name is still the SHA1 hash of its contents

Git uses objects to track changes throughout the history of a repository. To achieve this tracking, Git uses four types of objects:

* **Blobs**. Git uses blobs to store the contents of a file. A blob is a **Binary Large OBject (BLOB)**. A blob is created when we commence the tracking of a file by using the git add command. Since the blob is entirely defined by its data, if two files in a directory tree (or in multiple different versions of the repository) have the same contents, they will share the same blob object. The object is totally independent of its location in the directory tree, and renaming a file does not change the object that file is associated with.
* T**rees**. A tree object in Git can be thought of as a directory. It contains a list of blobs (files) and other tree objects (sub-directories).
* **Commits**:A commit object is essentially a pointer that contains a few pieces of important metadata.

However, the commit does not directly contain any changed files or data. Rather, it contains mostly metadata and pointers to other objects which contain the actual contents of the commit.

This means that each commit contains a snapshot of the entire project.

The commit itself has a hash, which is built from a combination of the metadata that it contains:

* + - The hash of the tree (the root tree object) at the time of the commit.
    - The hash of any parent commits. This is what gives a repository its history: every commit has a parent commit, all the way back to the very first commit.
    - The author’s name and email address, and the time that the changes were authored. The author is the name of the person responsible for this change.
    - The committer’s name and email address, and the time that the commit was made. The committer is the name of the person who actually created the commit, with the date it was done. This may be different from the author; for example, if the author wrote a patch and emailed it to another person who used the patch to create the commit
    - The commit message.

The parent line contains a hash of another commit object, and can be thought of as a "parent pointer" that points to the "previous commit". This implicitly forms a graph of commits known as the **commit graph**. Specifically, it's a directed acyclic graph (or DAG).

* **Annotated Tags** which point to a single commit object, and contain some metadata.it is a way to mark a specific commit as special in some way

Almost all of Git is built around manipulating this simple structure of four different object types. It is sort of its own little filesystem that sits on top of your machine's filesystem

So, the objects are tied together, blobs to trees, trees to other trees, and the root tree to the commit object, all connected by the SHA-1 identifier of the object.

So, every commit has a parent, and following these relations between commits, we can always navigate from a random one down to the first one, the already mentioned **root commit**

In short, the Git data model can be summarized as shown in the following diagram:



Git is amazingly smart and simple: to be quicker while searching through the filesystem, Git creates a set of folders where the name is two characters long, and those two characters represent the first two characters of a hash code; inside those folders, Git writes all the objects using as a name the other 38 characters of the hash, regardless of the kind of Git object. Git compresses them using the zlib library to reserve space on your disk.

This highlights once again the simplicity of Git: no metadata, no internal databases, or useless complexity, but simple files and folders are enough to make it possible to manage any repository.

All objects are stored as compressed contents by their sha values.

If the sha of your object is ab04d884140f7b0cf8bbf86d6883869f16a46f65, then the file will be stored in the following path:

.git/objects/ab/04d884140f7b0cf8bbf86d6883869f16a46f65

It pulls the first two characters off and uses that as the subdirectory, so that there are never too many objects in one directory. The actual file name is the remaining 38 characters.

**Git calculates the hash on the content of the file, not in the file itself.** This teaches us an important lesson: if you have two different files with the same content, even if they have different names and paths, in Git you will end up having only one blob.

This is why we use the git cat-file –p command, which decompresses them on the fly for us

The key **SHA-1** ([https://en.wikipedia.org/wiki/SHA-1)](https://en.wikipedia.org/wiki/SHA-1)s) is an alphanumeric sequence of 40 characters representing a hexadecimal number.

Git keeps all of these objects in the folder .git/objects. This is Git’s object database. Each object, regardless of type, is stored as a file, using its SHA-1 checksum as the filename (sort of). But, instead of storing all objects in a single folder, they are split up using the first two characters of their ID as a directory name, resulting in an object database that looks something like the following.

$ find .git/objects

.git/objects

.git/objects/00

.git/objects/00/11f080776acf2d04fb99b0d5c70f85747420a9

.git/objects/01

.git/objects/01/9da3ea8f032c4ebf7825cc13b5eeecc7cf017d

.git/objects/01/c3abfb09d4c4b2b306de4b20188574d4e02914

.git/objects/02

.git/objects/02/2d0352de4df1478f1f6571d0cf52ff22611f9f

.git/objects/03

.git/objects/03/1d9f1c82db42c05df688aed50bcea31bf7554b

For example, an object with the following ID:

022d0352de4df1478f1f6571d0cf52ff22611f9f

is stored in a folder called 02, using the remaining characters (2d0352...) as a filename.

### Git references

Another important directory are is .refs, where Git stores all of its references. .git stores:

* **References**, which are pointers to a single object (usually a commit or tag object).
* Branches and tags point to a commit object and the HEAD object points to the **branch** that is currently checked out. So, for every commit, the full tree state and snapshot are identified by the root tree.

In Git, **a branch is nothing more than a label**, a *mobile label* placed on a commit

Branches, remote-tracking branches, and tags are all references to commits. All references are named with a slash-separated path name starting with "refs"; the names we've been using so far are actually shorthand:

- The branch "test" is short for "refs/heads/test".

- The tag "v2.6.18" is short for "refs/tags/v2.6.18".

- "origin/master" is short for "refs/remotes/origin/master

Branches are movable labels

the commits are linked to each other by a parent-and-son relationship: each commit contains a reference to the previous commit.

So, branches are nothing but labels that are on the tip commit, the last one. This commit, our leaf, must always be identified by a label

Every time we make a commit to a branch, the **reference** that identifies that branch will move accordingly to always stay associated with the tip commit

In fact, every leaf on a Git branch has to be labeled with a meaningful name to allow us to reach it and then move around, go back, merge, rebase, or discard some commits when needed.

**$ ls –al**

**$ ls -al .git/**

**$ ls -al .git/objects**

**$ ls -al .git/objects/63**

Therefore, any content in **git** can be looked up by it's hash:

git cat-file -p 4bb6f98

HEAD is a special ref. It always points to the current object. You can see where it's currently pointing by checking the .git**/**HEAD file. Normally, HEAD points to another ref:

$cat .git**/**HEAD

ref: refs**/**heads**/**mainline

A ref is essentially a pointer. It's a name that points to an object. For example,

"master" --**>** 1a410e...

They are stored in `.git/refs/heads/ in plain text files.

$ **cat** .git**/**refs**/**heads**/**mainline

4bb6f98a223abc9345a0cef9200562333

Now, it's possible to navigate **git** purely by jumping around to different objects directly by their hashes. But this would be terribly inconvenient. A ref gives you a convenient name to refer to objects by. It's much easier to ask **git** to go to a specific place by name rather than by hash

## Git Directory and Working Directory

The 'git directory' is the directory that stores all Git's history and meta information for your project - including all of the objects (commits, trees, blobs, tags), all of the pointers to where different branches are and more. There is only one Git Directory per project (as opposed to one per subdirectory like with SVN or CVS), and that directory is (by default, though not necessarily) '.git' in the root of your project. If you look at the contents of that directory, you can see all of your important files:

### The Working Directory

The Git 'working directory' is the directory that holds the current checkout of the files you are working on. Files in this directory are often removed or replaced by Git as you switch branches - this is normal. All your history is stored in the Git Directory; the working directory is simply a temporary checkout place where you can modify the files until your next commit.

### The Git Index

The Git index is used as a staging area between your working directory and your repository. You can use the index to build up a set of changes that you want to commit together. When you create a commit, what is committed is what is currently in the index, not what is in your working directory.

The easiest way to see what is in the index is with the git-status command. When you run git status, you can see which files are staged (currently in your index), which are modified but not yet staged, and which are completely untracked.

## Explore the Object Database

For Windows users, Git installation will install a special command shell called *Git Bash*. To test your installation, open a new command prompt and run

$ git --version.

**[15] ~/grocery (master)**

**$ git cat-file -p a57d7**

**tree a31c31cb8d7cc16eeae1d2c15e61ed7382cebf40**

this plumbing command lets you peek into the Git objects; with the -p option (which means *pretty-print* here), we ask Git to show an easier way to read what the contents of the object are.

**$ git cat-file -p 637a0**

**banana**

You can see the contents of any commit like this:

$ **git cat-file** commit 5bac93

if you have the hash of a blob, you can look at it's contents.

$ git cat-file -t 54196cc2

$ git cat-file -s 54196cc2

$ git cat-file -p 54196cc2

You can examine the contents of any tree using ls-tree

$ git ls-tree 92b8b694

All of these objects are stored under their SHA1 names inside the git directory and the contents of these files is just the compressed data plus a header identifying their length and their type. The type is either a blob, a tree, a commit, or a tag.

$ find .git/objects/

The simplest commit to find is the HEAD commit, which we can find from .git/HEAD:

$ cat .git/HEAD

ref: refs/heads/master

$ cat .git/refs/heads/master

c4d59f390b9cfd4318117afde11d601c1085f241

$ git cat-file -t c4d59f39

commit

We can list all the heads in this repository with linkgit:git-show-ref

$ git show-ref --heads

$ git show-ref --tags

Computes the object ID value for an object with specified type with the contents of the named file

$ git hash-object

The ever-versatile git-show command can also be used to examine tree objects, but : git-ls-tree will give you more details

The "commit" object links a physical state of a tree with a description of how we got there and why. You can use the --pretty=raw option to git-show or git-log to examine your favorite commit.

A commit is usually created by git-commit, which creates a commit whose parent is normally the current HEAD, and whose tree is taken from the content currently stored in the index.

The Git command git cat-file -p will print the object given as an input. Normally, it is not used in everyday Git commands, but it is quite useful to investigate how it ties the objects together

$ git cat-file -p HEAD

We can now see the commit object, consisting of the root tree (tree), the parent commit object's ID (parent), the author and timestamp information (author), the committer and timestamp information (committer), and the commit message.

A tag object contains an object name (called simply 'object'), object type, tag name, the name of the person ("tagger") who created the tag, and a message, as can be seen using git-cat-file.

$ git cat-file tag v1.5.0

git-tag can also be used to create "lightweight tags", which are not tag objects at all, but just simple references whose names begin with "refs/tags/"

$ ls -al

**$ git cat-file -t 11b8b15**

**$ git cat-file -p 11b8b15**

**$ git checkout -**

**Switched to branch 'berries**

New trick: using the dash (-), you actually are saying to Git: "*Move me to the branch I was before switching*"; and Git obeys, moving us to the berries branch

The index is a binary file (generally kept in .git/index) containing a sorted list of path names, each with permissions and the SHA1 of a blob object; linkgit:git-ls-files[1] can show you the contents of the index:

$ git ls-files --stage

The **git commit** command does a few things:

1. Create blobs and trees to represent your project directory - stored in .git**/**objects

2. Creates a new commit object with your author information, commit message, and the root **tree** from step 1 - also stored in .git**/**objects

3. Updates the HEAD ref in .git**/**HEAD to the hash of the newly-created commit

We can ask git about particular objects with the cat-file command. Note that you can shorten the shas to only a few characters to save yourself typing all 40 hex digits:

I made a commit without first making git add; the *trick* is in the -a (--add) option added to the git commit command, which means *add to this commit all the modified files that I have already committed at least one time before*. In our case, this option allowed us to go faster and skip the git add command.

**$ ls -al .git/**

**$ ls -al .git/refs**

**$ ls -al .git/refs/heads**

**$ cat .git/refs/heads/master**

**0e8b5cf1c1b44110dd36dea5ce0ae29ce22ad4b8**

Git manages all this articulated reference system... with a trivial text file! It contains the hash of the last commit made on the branch

As branches are, HEAD is a **reference**. It represents a pointer to the place on where we are right now, nothing more, nothing less. In practice instead, it is just another plain text file:

**$ cat .git/HEAD**

**ref: refs/heads/berries**

The difference between the HEAD file and branches text file is that the HEAD file usually refers to a branch, and not directly to a commit as branches do. The ref: part is the convention Git uses internally to declare a pointer to another branch, while refs/heads/berries is of course the relative path to the berries branch text file

*In Subversion, we usually have different folders for each different branch.* When you switch a branch, Git goes to the commit the branch is pointing to, and following the parent relationship and analyzing trees and blobs, rebuilds the content on the **working directory** accordingly, getting hold of that files and folders

This gives us an object ID, but before we can inspect items in the object database, we need to know what type of object it is. Again, we can use the -t flag:

$ git cat-file -t 022d0352de4

tree

Of course, change the object ID to an object from your database (don’t forget to combine the folder name with the filename to get the full ID). This will output the type of commit, which we can then pass to a normal call to git cat-file.

git cat-file blob 7a52bb8

My object was a blob, but yours may be different. If it’s a tree, remember to use git ls-tree to turn that ugly binary data into a pretty directory listing.

$ git ls-tree 022d0352de4

100644 blob 1d09ca3ac33e045ccde753b47f81a9e980c90774 .gitconfig-template

100644 blob 139597f9cb07c5d48bed18984ec4747f4b4f3438 .gitignore

100644 blob 812e4df6163374ffb1ffbd1dac2cf8ec5460684e Basic-Tutorials-master.zip

040000 tree b9e065e8380804fcc424dc08128b3163ecf3ae6b Chapters

100644 blob 42ca1cef8e65effe2ad5bae228bcfc406c058f85 Links.txt

100644 blob 3aa637263b63eaaa57a2f34419374cfe7fc0701a Separating Collated Code with Branching Strategies.docx

040000 tree ba6aabea140dfdd02d0b4d84844dbe95410a78dd book-svg

100644 blob 2c6500fd68bb673ba4a5f571ced1e9cbdee31bf9 borderl.txt

100644 blob 6f6305ebd02adcfc9967d396246d5b0bc0183e37 instaLL.rtf

040000 tree e950201e1a88e43226e5d497656125936a514286 ppt

100644 blob 87701586a37307d58549e39dcdfc70bd8db51936 script.sh

100644 blob 0fe633e30c461e9e5e08d545fe06d909243955dd txt1.txt

Any file is compressed and transformed into a blob before archiving it into a Git repository. Each file is marked with a *hash*; this hash uniquely identifies the file within our repository, and it is thanks to this ID that Git can then retrieve it when needed, and detect any changes when the same file is altered (files with different content will have different hashes).

SHA-1 hashes are unique

**$ echo "banana" | git hash-object --stdin**

**637a09b86af61897fb72f26bfb874f2ae726db82**

The git hash-object command is the plumbing command to calculate the hash of any object; in this example, we used the --stdin option to pass as a command argument the result of the preceding command, echo "banana"; in a few words, we calculated the hash of the string "banana",

You can initialize a Git repository anywhere with the git init command. Take a look inside the .git folder to get a glimpse of what a repository looks like.

$ git init

Initialized empty Git repository in C:/temp/demo/.git/

$ ls -la .git

total 11

drwxr-xr-x 1 asaki 1049089 0 Jun 22 13:50 ./

drwxr-xr-x 1 asaki 1049089 0 Jun 22 13:49 ../

-rw-r--r-- 1 asaki 1049089 130 Jun 22 13:50 config

-rw-r--r-- 1 asaki 1049089 73 Jun 22 13:49 description

-rw-r--r-- 1 asaki 1049089 23 Jun 22 13:49 HEAD

drwxr-xr-x 1 asaki 1049089 0 Jun 22 13:49 hooks/

drwxr-xr-x 1 asaki 1049089 0 Jun 22 13:49 info/

drwxr-xr-x 1 asaki 1049089 0 Jun 22 13:50 objects/

drwxr-xr-x 1 asaki 1049089 0 Jun 22 13:49 refs/

The special HEAD pointer that refers to the branch/commit currently being checked out

* The tree object:

There are many ways to see the objects in the Git database. The git ls-tree command can easily show the content of trees and subtrees, and git show can show the Git objects, but in a different way.

We can also specify that we want the tree object from the commit pointed to by HEAD by specifying:

$ git cat-file -p HEAD^{tree}

The special notation HEAD^{tree} means that from the reference given, HEAD recursively dereferences the object at the reference until a tree object is found. The first tree object is the root tree object found from the commit pointed to by the master branch, which is pointed to by HEAD.

A generic form of the notation is <rev>^<type>, and will return the first object of <type>, searching recursively from <rev>.

* The blob object:

The branch object

we can take a look at the branch inside the .git folder where the whole Git repository is stored. If we open the text file .git/refs/heads/master, we can actually see the commit ID that the master branch points to. We can do this using cat, as follows:

**$ cat .git/refs/heads/master**

**13dcada077e446d3a05ea9cdbc8ecc261a94e42d**

We can also see that HEAD is pointing to the active branch by using cat with the .git/HEAD file:

**$ cat .git/HEAD**

The branch object is simply a pointer to a commit, identified by its SHA-1 hash.

The tag object

There are three different kinds of tag: a lightweight (just a label) tag, an annotated tag, and a signed tag. In the example repository, there are two annotated tags:

**$ git cat-file -p v1.0**

As you can see, the tag consists of an object—which, in this case, is the latest commit on the master branch—the object's type (commits, blobs, and trees can be tagged), the tag name, the tagger and timestamp, and finally the tag message.

There are many ways to see the objects in the Git database. The git ls-tree command can easily show the content of trees and subtrees, and git show can show the Git objects, but in a different way.