# Internal

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.

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



* The commit object points to the root tree. The root tree points to subtrees and files.
* 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.

Git's object storage is a key-value storage, the key being the ID of the object and the value being the object itself. The key is an SHA-1 hash of the object

## Git's objects

There are four types of objects in Git

- Files, or blobs as they are also called in the Git context

- Directories, or trees in the Git context

- Commits

- Tags

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

### The commit object

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.

### 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

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.

### 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 signedtag. In the example repository, there are two annotated tags:

When a file is moved to the staging area, the SHA-1 hash of the file is created and the blob object is written to Git's database. This happens every time a file is added, but if nothing changes for a file, it means that it is already stored in the database.

We can use the git fsck command to check for dangling objects—that is, objects that are not referred to by other objects or references:

$ git fsck --dangling

=====

The following diagram describes the tree stages and the commands used to move between the stages:



let's limit the log to only look through the history since the last

tag (release). To find the last tag, we can use git describe:

$ git describe

v3.1.0.201310021548-r-96-gb14a939

The preceding output tells us three things:

The last tag was v3.1.0.201310021548-r

The number of commits since the tag was 96

The current commit in abbreviated form is b14a939

asaki@WW7FRDEV01ASI MINGW64 /d/work/marsng/source (BHAOrdinaryGroup)

$ git describe

MarsNG\_ContainerShip\_2.0-201-g70f486565

$ git log MarsNG\_ContainerShip\_2.0..HEAD --oneline --grep "BHA"

This gives us all the BHA commits, but we can format this to a

format that is easily parsed with the --pretty option.

First, we will print the abbreviated commit ID (%h), followed by a

separator of our choice (|), and then the commit subject (%s, the first

line of the commit message), followed by a new line (%n), and the

body (%b):

--pretty="%h|%s%n%b"

git log MarsNG\_ContainerShip\_2.0..HEAD --grep "BHA" -20 --pretty="%h|%s%n%b"

The following command lists all the files that have changed since the last release

$ git log MarsNG\_ContainerShip\_2.0..HEAD --name-only -5 --oneline

By specifying --name-only, Git will only give the paths of the files that were changed by the commits in the range specified as output.

If we only want to show which files have been deleted in the repository, we can use the --diff-filter switch with git diff:

There are also switches for the files that have been added (A), copied (C), deleted (D), modified (M), renamed (R), and so on.

asaki@WW7FRDEV01ASI MINGW64 /d/work/marsng/source (BHAOrdinaryGroup)

$ git diff HEAD~..HEAD --name-only --diff-filter=M

BV.Mars.Kernel/Domain/TransverseBulkhead/OrdinaryStiffener.cs

BVUtils

In the repository, run gitk --all & to bring up the gitk interface.

Finding commits in the history

$ git log --grep "Performance" --oneline –all

Note that the search is case sensitive—had we searched for "performance"

(all in lower case), the list of commits would have been very different:

Sometimes, it is not enough to list the commit messages. You may want to know which commits touched a specific method or variable

We would like to find all the commits that have had changes made to the lines that contain the "FindNextStructure" method.

$ git log -G"FindNextStructure" –oneline

This method is quite convenient to use in finding out when a

given string was introduced or deleted, and to get the full context and commit at that point in time.

There is another option that can be used with git log, namely -S, which will look through the difference in the patch text in a similar way to the -G option, but will only match commits where there is a change in the number of occurrences of the specified string

We can see the content of the commit with the git show command, and use grep -C4 to limit the output to just the four lines before and after the search string:

$ git show c9c57d3 | grep -C4 " FindNextStructure "

$ git config --list --system

# list the global configuration

$ git config --list –global

# list the configuration for this repository

$ git config --list –local

$ git config --global user.email

$ git config --local user.email john@example.com

You can set the editor to the editor of your choice either by changing the $EDITOR environment variable or with the core.editor configuration target, for example:

$ git config --global core.editor vim

To let Git know about our new commit message template, we can set

the configuration variable commit.template to point at the file we just

created with that template; we'll do it globally so it is applicable to all

our repositories:

$ git config --global commit.template $HOME/.gitcommitmsg.txt

When commit.template is set, Git simply uses the content of the template file as a starting point for all commit messages. This is quite convenient if you have a commit-message policy, as it greatly increases the chances of the policy being followed. You can even have different templates tied to different repositories since you can just set the configuration at the local level

Git aliases

An alias is a nice way to configure long and/or complicated Git commands to represent short useful ones. An alias is simply a configuration entry under the alias section. It is usually configured to --global to apply it everywhere.

The alias method is also good for creating the Git commands you think are missing in Git. One of the common Git aliases is unstage, which is used to move a file out of the staging area, as shown in the following command:

$ git config --global alias.unstage 'reset HEAD --'

A common use case for aliases is to format the history of Git in specific ways. Let's say

you want the number of lines added and deleted for each file in the commit displayed

along with some common commit data. For this, we can create the following alias so we don't have to type everything each time:

$ git config --global alias.ll "log --pretty=format:"%C(yellow)%h%Cred%d %Creset%s %Cgreen(%cr)%C(bold blue)<%an>%Creset"

--numstat"

If you need a branch from a specific commit hash, you can create it

with the git branch command as follows:

$ git branch anotherBugFix 979e346

$ git log -1 anotherBugFix --format=format:%h

979e346

$ git log -1 anotherBugFix --format=format:%H

979e3467112618cc787e161097986212eaaa4533

2. As you can see, the abbreviated commit hash is shown when you use %h, and the full commit hash is shown when you use %H. You can see that the abbreviated commit hash is the same as the one used to create the branch. Most of the time, you want to create and start working on the branch immediately:

$ git checkout -b lastBugFix 979e346

Switched to a new branch 'lastBugFix'

Checking the difference between branches can show valuable information before merging. A regular git diff between two branches will show you all the information

Then, to list files that have changed between these branches, you can use the --nameonly

flag.

the command in this pattern, that is, git diff [options] <commit> <commit> <path>.

$ git diff --name-only origin/stable-3.1 origin/stable-3.2 src/org/transport

Let's try the same diff between branches, but this time we will diff the entire branches, not

just a sub-directory; however, we only want to show the deleted or added files between

the branches. This is done by using the --diff-filter=DA and --name-status options. The --namestatus

option will only show the filenames and the type of change. The --diff-filter=DA

option will only show the deleted and added files:

$ git diff --name-status --diff-filter=DA origin/stable-3.1 origin/stable-3.2

Orphan branches

However, in some situations, it is useful to have a branch with no parent.

One example would be an instance where you have your code base in two separate

Repositories

It is actually easy to create an orphan branch. The flag --orphan to checkout will do it.

$ git checkout --orphan fresh-start

the orphan branch does not share history with the master branch, and git

will not allow you to merge the branch. It shouldn't come as a surprise, since it is

basically what an orphan branch is all about. However, you can still merge an orphan

branch by allowing unrelated histories to be merge

$ git checkout master

$ git merge fresh-start

fatal: refusing to merge unrelated histories

$ git merge fresh-start --allow-unrelated-histories

Check out a new branch, rebaseExample, which tracks origin/stable-3.1:

$ git checkout -b rebaseExample --track origin/stable-3.1

$ git checkout -b resetAuthorRebase -t origin/master

Branch resetAuthorRebase set up to track remote branch 'master' from 'origin'.

Switched to a new branch 'resetAuthorRebase'

When you are working on a new feature and have branched from an old release into a feature branch, you might want to rebase this branch onto the latest release. When looking into the list of commits on the feature branch, you may realize that some of the commits are not suitable for the new release. In that case, when you want to rebase the branch onto a new release, you will need to remove some commits. This can be achieved with interactive rebasing, where Git gives you the option to pick the commits you wish to rebase.

If you need a branch from a specific commit hash, you can create it

with the git branch command as follows:

$ git branch anotherBugFix 979e346

$ git log -1 anotherBugFix --format=format:%h

979e346

$ git log -1 anotherBugFix --format=format:%H

979e3467112618cc787e161097986212eaaa4533

2. As you can see, the abbreviated commit hash is shown when you use %h, and the full commit hash is shown when you use %H. You can see that the abbreviated commit hash is the same as the one used to create the branch. Most of the time, you want to create and start working on the branch immediately:

$ git checkout -b lastBugFix 979e346

Switched to a new branch 'lastBugFix'

The remoteBugFix2 branch is just a local branch at the moment with no

tracking information; to set the tracking branch, we need to use --setupstream-to or -u as a flag to the git branch command:

$ git branch --set-upstream-to origin/stable-3.2

Branch remoteBugFix2 set up to track remote branch stable-3.2 from

origin.

$ git log -1 origin/stable-3.2 --format=format:%h

f839d383e

To force a merge commit, you need to use the --no-ff

$ git merge origin/stable-3.2 --no-ff --edit --quiet

Then, to create the merge and allow you to decide what will be part of the commit, you can use --no-commit:

$ git merge origin/master --no-ff --no-commit

Then, to list files that have changed between these branches, you can use the --name-only flag.

$ git diff --name-only origin/stable-3.1 origin/stable-3.2 org.eclipse.source

We are building the command in this pattern, that is, git diff [options]

<commit> <commit> <path>

to

show the deleted or added files between the branches. This is done by using the --diff-filter=DA and --name-status options.

$ git diff --name-status --diff-filter=DA origin/stable-3.1

origin/stable-3.2

This shows the files that have been added and deleted while moving from origin/stable-3.1 to origin/stable-3.2.

You add additional information to the commits that can be displayed when running a git log command. You can also release the notes into a remote repository so that people can fetch the notes.

$ git notes add -m "Update MANIFEST files"

You can also append a current note with git notes append:

$ git notes append -m "Verified by John Doe"

But to show the note, we have to add --notes, which will

show the notes for the commits in the output:

$ git log -1 --notes –oneline

$ git log --format='format:%h %an <%ae>' origin/stable-3.2..HEAD

We will list all the commits from origin/stable-3.2 to HEAD and we will define a format with

%h as the abbreviated commit hash, %an for the author's name, and %ae for the author's email address.

There are two types of tags, a lightweight tag and an annotated tag. The lightweight tag is very similar to a branch, since it is just a named reference, such

as refs/tags/version123 .

$ git tag -l "v2.3.0.2\*"

We are using the git tag command with -l as a flag, since we want to

list the tags and not tag the current HEAD. Some repositories have a lot of tags; so to prevent the list from becoming too long, you can specify which tags you want to list and use a \* wildcard as we did previously.

So, in order to add more information, we

should use an annotated tag. An annotated tag is a tag where you have to add some information to the tag. To create an annotated tag, we use the --annotate tag for the git tag command:

$ git tag --annotate -m "Release Maturity rate 97%"

To delete a tag, you need to apply the -d flag:

$ git tag -d v1.3.0.201202121842-rc4

$ git tag -a -m "Local created tag" v1.3.0.201202121842-rc4

We have recreated the tag, and it points to HEAD because we did not

specify a commit hash at the end of the command

Once you have pushed a tag to a remote repository, you

should never change it, since the developers who are fetching from

the repository may never know about the changes unless they clone

again or delete the tags locally and fetch them again.

Extracting Data from the Repository

The git log command has different options, such as --numstat, that will show the number of files added and lines deleted for each file since each commit.

However, for finding the top committer in the repository, we can just use the git shortlog command. The shortlog Git command is very simple and does not have a lot of options or flags to use with it. To sort and find the top committer, we can use the -n or --numbered option to sort the output; the top committer is on top:

$ git shortlog -5 --numbered

$ git shortlog master..BHAOrdinaryGroup -sn

75 yu peng

44 adel saki

24 Baptiste Célérier

To list it for the repository, we need to add --all at the end of the command so as to execute the command for all branches as follows:

$ git shortlog --numbered --summary --all

So, to list the top committers for the last six months, we can add --since="6 months ago" to the git shortlog command as follows:

You can use "n weeks ago", "n days ago", "n months ago", " n hours ago", and so on

for specifying time periods. You can also use specific dates, such as "1 october 2013".

$ git shortlog --numbered –summary --all --since="6 months ago"

You can also list the top committer for a specific month using the -

-until option, where you can specify the date you wish to list the commit until. This can be done as follows:

$ git shortlog --numbered --summary --email --all --since="30 september

2013" --until="1 november 2013"

it is fairly simple to get some indication of who to go

to for the different files or directories in Git.

$ git shortlog --all -sn \*.sln

Finding bottlenecks in the source tree

The --dirstat option shows which directories have changed in the

commit and how much they have changed compared to each other.

The default setting is to count the number of lines added to or

removed from the commit.

$ git log -1 –dirstat

You can compensate for this slightly by using --dirstat=lines. This option

will look at each file line by line and see whether they have changed

compared to the previous version as follows:

$ git log -1 --dirstat=lines

If you would like to limit the output to only show directories with a certain percentage or higher, we can limit the output as follows:

$ git log -1 --dirstat=lines,10

we can add cumulative to the --dirstat=lines,10 command, and this will cumulate the changes and calculate a percentage.

$ git log -1 --dirstat=files,10,cumulative

we will be using git diff, as Git will show the accumulated diff between the two releases, and git log will show dirstat for each commit between the releases

git diff master..BHAOrdinaryGroup --dirstat=lines,cumulative

We can then dig a little deeper using --stat or --numstat for the directory, and again use git diff. We will also use -- relative="org.eclipse.jgit.test/tst/org/eclipse/", which will show the relative path of the files from BV.Mars.UI.

git diff master..BHAOrdinaryGroup --numstat --relative=BV.Mars.UI

asaki@WW7FRDEV01ASI MINGW64 /d/work/marsng/source (BHAOrdinaryGroup)

$ git diff master..BHAOrdinaryGroup --pretty --numstat --relative=BV.Mars.UI/

31 0 BV.Mars.UI.csproj

7 2 UserInterface/ApplicationShell.cs

5 1 UserInterface/Context/DesignContext/BhDesignContext/BhOrdinaryStiffenerDesignContext.cs

1 1 UserInterface/Controls/CalcTorsionModelControlPresenter.cs

45 0 UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/BhaOrdinaryGroupInputCustomController.cs

6 1 UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Data/AddStiffener.cs

30 0 UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Data/AddStiffenerGroupe.cs

34 39 UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Data/DataReferencePointType.cs

The first number is the number of lines added, and the second number is the lines removed from the files between the two branches

We have used git log, git diff, and git shortlog to find information about the

repository, but there are so many options for those commands on how to find bottlenecks in the source code.

If we want to find the files with the most commits, and these are not

necessarily the files with the most line additions or deletions, we can use git log:

We can use git log between the origin/stable-3.1 and origin/stable-3.2 branches and list all the files changed in each commit.

$ git log origin/stable-3.1..origin/stable-3.2 --format=format: --nameonly

You can see from the extensive output that you only see file

names and nothing else. This is due to the options used. The --

format=format: option tells Git to not display any commit-messagerelated

information, and --name-only tells Git to list the files for each

commit

Then, we just need to sort and accumulate the result with some Bash tools as follows:

asaki@WW7FRDEV01ASI MINGW64 /d/work/marsng/source (BHAOrdinaryGroup)

$ git log master..BHAOrdinaryGroup --format=format: --name-only | sed '/^$/d' | sort | uniq -c | sort -r | head -10

21 BV.Mars.PresentationLogic/PresentationLogic/Presenters/FormsPresenters/InputFormsPresenters/CrossSectionPresenters/CrossSectionInputPresenters/EndConnections/BhStiffEndConnectionInputPresenter.cs

21 BV.Mars.Kernel/Domain/TransverseBulkhead/Bulkhead.cs

18 BV.Mars.PresentationLogic/Applicative/Controllers/CtrlBulkhead.cs

14 BV.Mars.Kernel/Domain/TransverseBulkhead/OrdinaryStiffenerGroup.cs

12 BV.Mars.UI/UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Forms/BhaOrdinaryStiffenerInputForm.resx

12 BV.Mars.UI/UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Forms/BhaOrdinaryStiffenerInputForm.Designer.cs

11 BV.Mars.UI/UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Forms/BhaOrdinaryStiffenerInputForm.cs

10 BV.Mars.Kernel/Domain/TransverseBulkhead/SmallOrdinaryStiffener.cs

10 BV.Mars.Kernel/Domain/TransverseBulkhead/OrdinaryStiffener.cs

9 BV.Mars.UI/UserInterface/Forms/InputForms/BhaForms/BhaOrdinaryStiffener/Forms/EndConnectionInputUc.cs

We got the list of files, and we used sed '/^$/d' to remove empty lines from the output. After this, we used sort to sort the list of files. Then, we used uniq -c, which counts the occurrences of each item in the files

and adds the number from the output. Finally, we sorted in reverse

order using sort -r and displayed only the top ten results using head 10.

To proceed from here, we should list all the commits between the branches that are changing the top file as follows

By adding the file to the end of the git log command, we will see the commits between the two branches

$ git log origin/stable-3.1..origin/stable-3.2 org.eclipse.jgit/src/org/eclipse/jgit/api/RebaseCommand.java

Let's see how many commits in the repository are referring to a bug:

$ git log --all --grep="^[bB][uU][gG]: [0-9]+"

So what was the grep doing? The ^[Bb][Uu][gG]: part matches any combination of lowercase and uppercase bugs. The ^ character means from the beginning

of the line. The : character is matching :. Then, we have [0-9]+, which will match any

number between zero and nine, and the + part means one or more occurrences.

but for now, we just want to count the commits. We can do this by piping it to wc -l (word count –l is to count the lines):

$ git log --all --oneline --grep="^[bB][uU][gG]: [0-9]+" | wc -l

Before piping it to wc, remember to use --oneline to limit the output to one line for each

commit.

If you are used to using regular expressions in another scripting or programming language, you will see that using --grep does not support everything. You can enable a more extensive regular expression support using the --extended-regexp option for git log; however, the pattern still has to be used with --grep as follows:

$ git log --all --oneline --extended-regexp --grep="^[bB][uU][gG]: [0-9]{6}"

I have used a slightly different expression, and have now added {6} instead of +;

the {6} searches for six occurrences of the associated pattern. In our case, it is six digits as it is next to the [0-9] pattern.

To shrink the regular expression even more, we can use --regexpignore-case, which will ignore the case for the pattern:

$ git log --all --oneline --regexp-ignore-case --extended-regexp --grep="^bug: [0-9]{6}"

While extracting information from Git, one of the natural things to do is to generate release notes. To generate a release note, you need all the valid information from the repository between this release and the previous release

We start by listing the commits between two tags, v2.3.1.201302201838-r and v3.0.0.201305080800-m7,

and then we build on that information:

By using git log with v3.0.0.201305080800-m7.. v3.0.0.201305080800-m7, we will get the commits between the tags. As we have a lot of commits between these two tags, let's count them using wc -l:

$ git log --oneline v2.3.1.201302201838-r..v3.0.0.201305080800-m7 | wc –l

Now, we will show the most modified files between the releases:

$ git log v2.3.1.201302201838-r..v3.0.0.201305080800-m7 --format=format: --name-only | sed '/^$/d' | sort | uniq -c | sort -r | head -10

This information is useful as we now have an overview of where the majority of the

changes are

Sometimes it's useful to be able to extract what has been achieved in a specific range of time. Let's see how git log numerous arguments can help with this task.

we want to know everything that has been done in the last 30 days in the jgit repository that we have been analyzing so far:

$ git log --all --since="30 days ago"

Here, we use --all in order to see the commits in all the branches and not only the

current one.

let's only show the commits by David Pursehouse:

$ git log --all --since="30 days ago" --oneline --author="David Pursehouse"

It looks like some merge commits are present. These are not really useful to describe the activity of the last month, so let's get rid of those with --no-merges:

$ git log --all --since="30 days ago" --oneline --author="David Pursehouse" --no-merges

The next alias is a little different, as it will count the number of commits in the repository, and this can be done with the wc (wordcount) tool

However, since this is not a built-in Git tool, we have to use the exclamation mark and also specify Git:

$ git config alias.count '!git log --all --oneline | wc -l'

This also means you can execute external tools as if they were Git tools just by creating a Git alias; for instance, if you are using Windows, Mac, or Linux, you can create an alias as follows:

$ git config alias.wa '!explorer .' # Windows

This alias will open up an Window Explorer at the path you are currently at. The next one shows what changed in the HEAD commit. It executes this with the --name-status option for git log:

$ git config alias.gl1 'log -1 --name-status'

So, you can try and open the .git/config configuration file, or you can list the configuration with git config -list:

$ git config --list | grep alias

The alias feature is very strong, and the idea behind it is that you should use it to shorten those long one-liners that you often use.

To configure the template, we need to use git config commit.template <pathtofile> to set it, and, as soon as it is set, we can try to create a commit and see how it works:

1. Start by configuring the template as follows:

$ git config commit.template ~/committemplate

2. Now list the config file to see that it has been set:

$ git config --list | grep template

As we predicted, the configuration was a success. The template, just like any other configuration, can be set at a global level using git config --global, or it can be set at a local repository level by leaving out the --global option. We configured our commit template for this repository only. Let's try and make a commit:

$ git commit --allow-empty

The --amend option to git commit is roughly equivalent to performing git reset --soft HEAD^,

followed by fixing the files needed and adding those to the staging area. Then, we will run git commit reusing the commit message from the previous commit (git commit -c ORIG\_HEAD).

Now, you can amend the latest commit with git commit --amend. The command will include files in the index in the new commit and you can, as with the last example, reword the commit message if needed. It is not needed in this example, so we'll pass the --no-edit option to the command:

$ git commit --amend --no-edit

Revert can be used to undo a commit in history that has already been published (pushed),whereas this can't be done with the amend or reset options without rewriting history. Revert works by applying the anti-patch introduced by the commit in question

It's possible to revert more than one commit in a single revert, for example, git revert

master~6..master~2 will revert the commits from the sixth commit from the bottom in the master to the third commit from the bottom in the master (both included). It is also possible not to create a commit while reverting; passing the -n option to git revert will

apply the needed patched, but only to the working tree and the staging area.

Merge commits are a special case when it comes to revert. In order to be able to revert a merge commit, you'll have to specify which parent side of the merge you want to keep.