

Git za matematike

Abstract

Git je program, ki omogoča vodenje zgodovine različic datotek v nekem direktoriju. V glavnem se uporablja za upravljanje z izvorno kodo pri razvoju računalniških programov. Mnogi med nami pa ga uporabljajo tudi pri pisanju besedil v *LaTeX*-u. Poleg tega, da Git hrani zgodovino sprememb, tudi omogoča da več ljudi hkrati sodeluje pri urejanju istih datotek. Ogledali si bomo, kako Git deluje. Opisali bomo, kako Git uporabi *zgoščevalne funkcije*, *Merklejeva drevesa* in *usmerjene aciklične grafe*, da shrani zgodovino različic in omogoči hkratno urejanje vsebine. Matematični model, ki ga Git uporablja je v resnici zelo preprost in njegovo razumevanje nas lahko reši marsikatero zagato, ki nastane med njegovo uporabo.

1. Kaj je Git?

Git je kot **časovni stroj** za datoteke. Uporabniku omogoča, da vidi **pretekle različice** datotek, sprememinja datoteke, **brez skrbi, da bi kaj pokvaril** in datoteke **deli z drugimi**. Poleg časovnega stroja je Git **razpršeno skladišče datotek**. Omogoča, da datoteke **hkrati ureja več uporabnikov** na različnih računalnikih in kasneje spremembe **združi**.

Git hrani vsebino direktorija z datotekami in celotno zgodovino različic datotek iz preteklosti. Za vsako različico hrani Git zapis o avtorju, datumu in opis sprememb, ki so nastale v primerjavi s predhodno različico. Vse te informacije dajejo podroben pregled nad zgodovino sprememb.

Opomba

Git in GitHub nista eno in isto. Ljudje pogosto mešajo Git in GitHub. Git je program, ki si ga lahko vsakdo namesti in poganja na svojem računalniku. Program Git je ustvaril Linus Torvalds, da bi lažje upravljal z izvorno kodo za jedro operacijskega sistema Linux. GitHub je javno spletišče, ki je namenjeno skladiščenju Git repozitorijev.

Opomba

*Sisteme, ki omogočajo hranjenje preteklih različic datotek, imenujemo **sistemi za nadzor različic** (angl. version control system (VCS)) ali sistemi za upravljanje z izvorno kodo (angl. Source Code Management (SCM)).*

*Poleg nadzora različic Git omogoča hkratno spreminjanje datotek več uporabnikov na različnih računalnikih. Zato je Git **distribuiran sistem za nadzor različic** (angl. Distributed Version Control System (DVCS)).*

V nadaljevanju bomo obravnavali naslednje teme:

- **Podatkovno skladišče**: Kako Git uporablja **zgoščevalno funkcijo** in **Merklejeva drevesa** za hranjenje posnetkov vsebine direktorija.
- **Zgodovina sprememb**: Kako zgodovino predstavimo z **usmerjenim acikličnim grafom**, v katerem so vozlišča različice in ki povezuje različice z njihovimi neposrednimi predhodniki.
- **Reference**: Kako preproste reference (kazalci) na vsebino omogočajo bliskovito preklapanje med različicami in preprečijo popolno zmešnjavo, ko več ljudi hkrati spreminja iste datoteke.

2. Podatkovno skladišče

Ko ustvarimo nov Git repozitorij, Git ustvari direktorij z imenom `.git`, ki vsebuje vse podatke, ki jih Git potrebuje. Git v mapi `.git` hrani različne stvari:

- vsebino datotek, s ki smo jih dodali v repozitorij
- drevesno strukturo direktorija
- posnetke stanja v različnih trenutkih s podatki o avtoju, datumu in opisu sprememb
- kazalce na posamezne posnetke stanja

Git repozitorij je vsak direktorij, ki vsebuje poddirektorij `.git` z zgoraj navedenimi podatki. Podrobnosti o tem lahko preberete [poglavju o shrambi objektov](#) v knjigi o Gitu.

2.1. Zgoščevalna funkcija

Git ne shranjuje datotek z običajnimi imeni, ampak za ime uporabi vrednost zgoščevalne funkcije njene vsebine. Git uporablja zgoščevalno funkcijo *SHA-1*. Funkcija *SHA1* je posebna implementacija zgoščevalne funkcije, ki se uporablja v kriptografiji.

Naj bo B množica vseh možnih podatkovnih nizov(besedil). Zgoščevalna funkcija *SHA1* je funkcija

$$H : B \rightarrow \{0, 1, \dots, 2^{160} - 1\},$$

ki vsakemu besedilu b priredi 160-bitno zgoščeno vrednost besedila $H(b)$. Funkcija H , je izbrana tako, da sprememba enega samega bita v besedilu $b \in B$ spremeni vrednost $H(b)$. Poleg tega zahtevamo, da je porazdelitev vrednosti $H(b)$ čim bližje enakomerni porazdelitvi. To pomeni, da so vse vrednosti $H(b)$ približno enako verjetne. Kljub temu, da zgoščevalna funkcija H ni injektivna, je verjetnost, da bi imela dva podatkovna niza isto vrednost H zelo majhna ($\approx 2^{-159}$). Zato lahko v praksi predpostavimo, da je z vrednostjo $H(b)$ niz b enolično določen.

Ko datoteko z vsebino b zabeležimo v Git repozitorij, git shrani vsebino v datoteko z imenom $H(b)$ v `git/objects`¹. Vsebina b je tako vedno dostopna pod imenom $H(b)$. Tako dobimo [vsebinsko naslovljivo shrambo objektov](#), ki je ena od bistvenih značilnosti Gita. Ta način shranjevanja omogoča, da lahko vedno preverimo, če ima shranjenjena vsebina isto vrednost zgoščevalne funkcije, kot je njeno ime. Lahko tudi shranimo več različic iste datoteke, saj ima vsaka različica drugačno zgoščevalno vrednost. Zgoščevalna vrednost služi tudi kot kontrola, če je prišlo do kvaritve podatkov, ki so shranjeni v Git repozitoriju.

2.2. Datotečna drevesa

V vsebinsko naslovljivo shrambo objektov lahko shranimo vsebino datotek in njihovih prejšnjih različic. A kako ohranimo informacijo o imenu datotek in drevesni strukturi direktorija? Git za to ustvari nov tip objekta *drevo* (angl. *tree*), ki hrani preprost seznam imen datotek in naslovov na vsebino datotek v direktoriju. Naslov na vsebino datoteke b je seveda zgoščena vrednost vsebine $H(b)$. Seznam imen datotek in zgoščenih vrednosti je preprosta tekstovna datoteka, za katero lahko izračunamo zgoščeno vrednost. Zgoščena vrednost datotečnega drevesa natanko določa tako imena datotek, kot tudi vsebino datotek, ki so vsebovane v direktoriju. Če se katerakoli datoteka ali ime datoteke v direktoriju spremeni, se bo spremenila tudi njena zgoščena vrednost in posledično zgoščena vrednost za drevo. Poleg posameznih datotek, lahko drevo vsebuje tudi poddrevesa. Tako lahko rekurzivno ustvarimo drevesno podatkovno strukturo, ki zajema direktorij z datotekami in poddirektoriji v poljubni globini.

¹V resnici Git shrani vsebino v datoteko z imenom $h_3h_4\dots h_{40}$ v mapi h_1h_2 , kjer je $h_1h_2h_3\dots h_{40}$ zapis $H(b)$ v 16-tišlkem sistemu. Datoteka, katere vsebina ima zgoščeno vrednost $H(b)$ enako `8dd6d4bdaeff93016bd49474b54a911131759648` bo shranjena v `.git/objects/8d/d6d4bdaeff93016bd49474b54a911131759648`

```

100644 blob bcc1382241e267cf790ca6b3afe9fde6dcf1072f    bla.txt
100644 blob 2ce22b4dc77442103f095503f1205937c1b0fcfc    blabla.txt
040000 tree 605f479464bebe4f7250ace49bab48e72855f84a    podmapa

```

Program 1: Vsebina direktorija v Gitu je preprost seznam datotek in poddirektorijev in zgoščenih vrednosti njihove vsebine

Poglejmo si primer. Denimo, da imamo v naslednjo strukturo datotek in poddirektorijev

```

├─ bla.txt (vsebina: bla)
└─ mapa
   └─ bla.txt (vsebina: bla)
     └─ blabla.txt (vsebina: blabla)

```

Git bo shranil naslednje objekte v vsebinsko naslovljivo shrambo:

- vsebino datoteke bla.txt:

Zgoščena vrednost: bcc1382241e267cf790ca6b3afe9fde6dcf1072f
bla

- vsebino datoteke blabal.txt:

Zgoščena vrednost: 2ce22b4dc77442103f095503f1205937c1b0fcfc
blabla

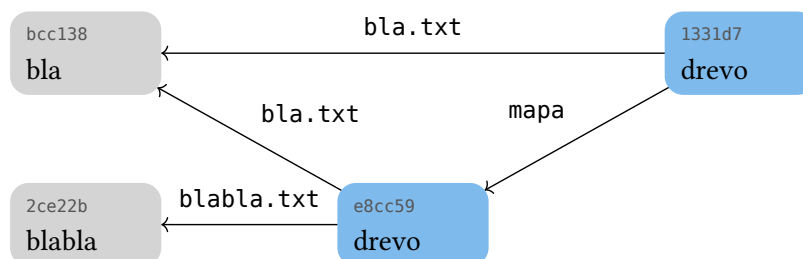
- seznam datotek v direktoriju mapa:

Zgoščena vrednost: e8cc593eddfb9cfdafb4f9c46ab7f5a05ea00b2b
100644 blob bcc1382241e267cf790ca6b3afe9fde6dcf1072f bla.txt
100644 blob 2ce22b4dc77442103f095503f1205937c1b0fcfc blabla.txt

- seznam datotek v korenskem direktoriju:

Zgoščena vrednost: 1331d77b31e40d6b470706b195f21244bb32cf21
100644 blob bcc1382241e267cf790ca6b3afe9fde6dcf1072f bla.txt
040000 tree e8cc593eddfb9cfdafb4f9c46ab7f5a05ea00b2b mapa

Git z uporabo zgoščene vrednosti kot kazalca na vsebino, vsebino direktorija postavi v podatkovno strukturo, ki jo matematično lahko opišemo z *usmerjenim acikličnim grafom*. Ko je vsebina datotek enaka (npr. bla.txt in mapa/bla.txt), Git shrani le eno kopijo, ki je dostopna v datoteki .git/objects/bc/c1382241e267cf790ca6b3afe9fde6dcf1072f. Zato datotečno drevo v Gitu ni nujno drevo, ampak usmerjen aciklični graf.



Slika 1: Primer datotečnega grafa povezanega z zgoščenimi vrednostmi

Posledično lahko celotno vsebino direktorija opišemo z eno samo zgoščeno vrednostjo. Če spremenimo vsebino, ime ali lokacijo datoteke, bo sprememba vplivala na zgoščeno vrednost spremenjene vsebine in sprememba bo splavala na površje do zgoščene vrednosti za korenski direktorij. Zgoščena vrednost služi tako kot identifikator vsebine, kot tudi kot kontrolna vsota, ki omogoča detekcijo sprememb.

Opomba

Podatkovna struktura objektov v Gitu je podobna Merklejevim drevesom. Razlika je v tem, da Gita hrani le eno kopijo datotek z identično vsebino, zato dobimo usmerjen aciklični graf in ne drevesa. Postopek je podoben veriženju blokov, ki se uporablja v kriptovalutah.

Opomba

Dostop do objekta je mogoč, če poznamo **zgoščeno vrednost** njegove vsebine. To pomeni, da je referenca na posamezen objekt v Gitu preprosto zgoščena vrednost (angl. hash) vsebine tega objekta. Po drugi strani je vsebina objekta določena z njegovo zgoščeno vrednostjo. To pomeni, da lahko enostavno preverimo verodostojnost vsebine, ki je shranjena v Gitu. Git hrani skladišče objektov v direktoriju `.git/objects`.

3. Zgodovinski graf sprememb

V prejšnjem poglavju smo videli, kako Git hrani vsebino direktorija in kako je mogoče do vsebine dostopati če poznamo zgoščeno vrednost korenskega direktorija. Zgodovinsko drevo sprememb je preprosta razširitev omenjene podatkovne strukture.

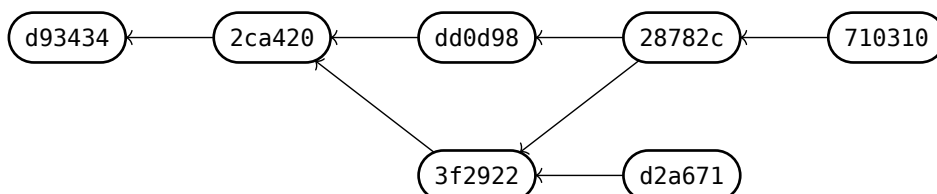
3.1. Posnetki stanja

Osnovna enota v Gitu je **Vnos** (angl. **commit**). Vnos je posnetek stanja zabeleženih datotek v trenutku, ko je bil ustvarjen. Poleg vsebine datotek vsak vnos vsebuje še metapodatke o avtorju, datumu vnosa in opisom sprememb. Podobno kot objekt tipa *drevo*, je tudi vnos objekt v vsebinsko naslovljivi shrambi, ki je ima določeno **zgoščeno vrednost vnosa**. Zgoščena vrednost vnosa je natanko določena z vsebino shranjenih datotek in metapodatkov vnosa.

Zgoščena vrednost: 8dd6d4bdaeff93016bd49474b54a911131759648
tree 65c47feec7465e80492620a48206793e078702e0
parent 16f2994757f1213935b8edb9ae7fee3a8e9ec98d
author MV <mv@example.com> 1765235698 +0100
committer MV <mv@example.com> 1765235698 +0100
Dodaj bla

Tabela 1: Primer vnosa v Gitu. Vnos vsebuje zgoščeno vrednost posnetka direktorija (tree), zgoščeno vrednost starševskega vnosa (parent) in metapodatke. Tudi sam vnos je natančno določen z zgoščeno vrednostjo.

Vsak vnos je povezan s točno določenim posnetkom vsebine korenskega datotečnega drevesa, ki ga identificira zgoščena vrednost. Poleg tega so posamezni vnosi so povezani v **usmerjen acikličen graf (DAG)**, ki predstavlja zgodovino sprememb. Vsak **vnos** je **vozlišče** v grafu. Vsak vnos izhaja iz enega ali več starševskih vnosov. Izjema je prvi vnos. **Povezave** v grafu povezujejo vnose z njihovimi starši.

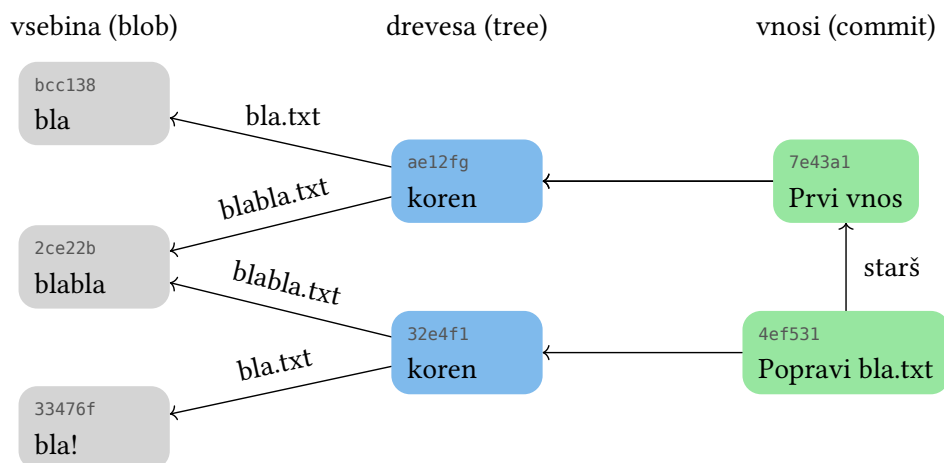


Slika 2: Vnosi v Gitu kot usmerjen graf. Vsak vnos (razen prvega) ima povezavo na vnose iz katerih izhaja.

Git hrani zgodovino sprememb v *vsebinsko naslovljivi shrambi objektov*, ki hrani tri vrste objektov:

- blob: vsebina datotek,
- tree: vsebina direktorijev,
- commit: posnetek vsebine v določenem trenutku.

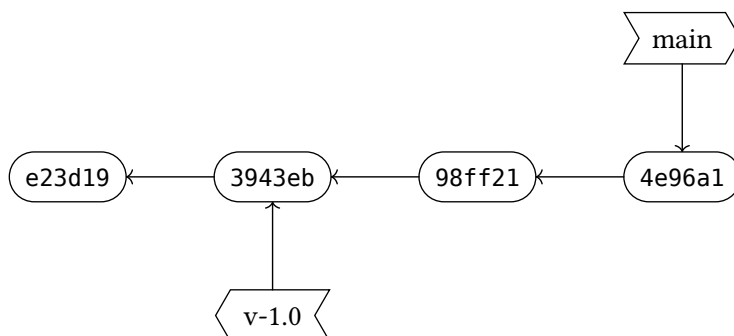
Objekti so povezani v *usmerjen aciklični graf*. Podgraf na vnosih določa zgodovino sprememb. Naslovi objektov so *zgoščene vrednosti* vsebine objekta, zato je zagotovljena verodostojnost shranjenih podatkov.



Slika 3: Vsebinsko naslovljiva shramba objektov v Gitu. Naslovi so zgoščene vrednosti vsebine. Shramba vsebuje dva vnosa. V prvem vnosu smo dodali dve datoteki bla.txt in blabla.txt, v drugem vnosu pa smo spremenili le vsebino datoteke bla.txt.

4. Kazalci: veje in značke

Poleg objektov kot so *vnosi*, *posnetki direktorijev* in *posnetki datotek* pozna git še reference. Reference so kazalci z določenim imenom na posamezen vnos.

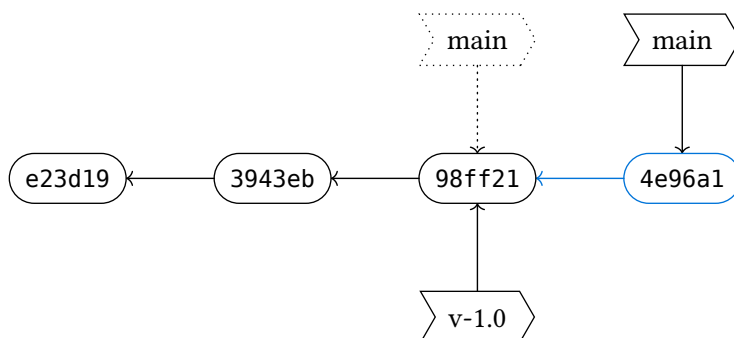


Slika 4: Veja (angl. branch) ali značka(angl. tag) je preprost kazalec na posamezen vnos(angl. commit).

Referenc git ne hrani v skladišču objektov, temveč posebej v direktoriju `.git/refs`. Zato so reference vezane na posamezen repozitorij in se lahko razlikujejo med različnimi kloni določenega repozitorija.

Veja (angl. **branch**) je posebne vrste referenca, ki se premika, ko dodajamo nove vnose. Vsakič ko ustvarimo nov vnos, se trenutno aktivna veja premakne na novo ustvarjeni vnos.

Značka (angl. **tag**) je referenca, ki je statična in se ne premika več, ko jo enkrat ustvarimo.



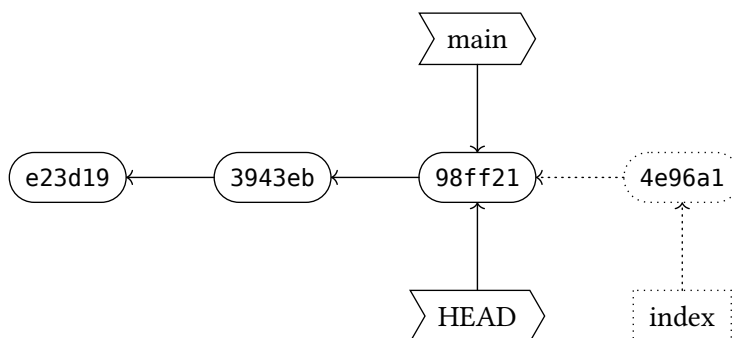
Slika 5: Ko ustvarimo nov vnos, se aktivna veja main premakne naprej, značka v-1.0 pa ostane tam, kjer je bila.

Opomba

Veje in značke nimajo v Gitu nobenega posebnega pomena, razen tega, da so reference na vnose. Pomen posameznih vej je stvar dogovora med uporabniki. Tako se pogosto uporablja različne veje za različne namene: `main` ali `master` je navadno glavna veja razvoja, veje z imeni `stable`, `production`, `development` in podobno označujejo različne stopnje razvoja programske opreme, veje s predpono `feature-` označujejo razvoj novih funkcionalnosti.

Vse te pomene damo vejam ljudje, ki sodelujemo v nekem Git repozitoriju. Za Git so vse veje in značke zgolj preprosti kazalci na določen vnos.

HEAD je posebna referenca, ki kaže na trenutno aktiven vnos. Vnos, na katerega kaže **HEAD** bo starševski vnos naslednjega vnosa, ki ga bomo dodali.



Slika 6: **HEAD** je referenca na trenutno aktiven vnos. **Index** vsebuje spremembe, ki bodo zabeležene v naslednjem vnosu.

4.1. Povzetek

Povzemimo sedaj, kaj smo spoznali o podatkovnem modelu Gita. V vsebinsko naslovljivi shrambi hrani Git posnetke stanja direktorija, ki ga vodimo v repozitoriju skupaj z metapodatki o spremembah.

Najpomembnejša pojma sta:

- **Vnos** (angl. **commit**) je posnetek trenutnega stanja projekta, shranjen kot vozlišče v zgodovinskem grafu, ki vsebuje posnetek stanja datotek ter metapodatke (avtor, čas, sporočilo).
- **Zgoščena vrednost vnosa** (angl. **commit hash**) je 40-mestna heksadecimalna vrednost, izračunana s SHA-1, ki enolično identificira vnos na podlagi vsebine posnetka in metapodatkov.

Izven shrambe objektov hrani Git še reference na posamezne vnose. Poznamo dve vrsti referenc:

- **Veja** (angl. **branch**) je premična reference, ki kaže na določen vnos v zgodovini in se samodejno premakne naprej, ko dodajamo nove vnose. Veje omogočajo vzporedne razvojne linije ki so med sabo neodvisne.
- **Oznaka** (angl. **tag**) je statična referenca, ki trajno kaže na določen vnos. Za razliko od veje se oznaka, nikoli ne premika samodejno, zato se uporablja predvsem za označevanje pomembnih točk v zgodovini, kot so izdaje ali stabilne verzije.
- **HEAD** je posebna oznaka, ki kaže na trenutno aktiven vnos v delovni kopiji.

Omenimo še dva pojma, ki jih uporabljamo pri delu z Gitom:

- **Delovna kopija** (angl. **workout copy**) je direktorij v katerem urejamo datoteke, ki jih nato vnesemo v Git. V delovni kopiji imajo na začetku datoteke isto vsebino kot je vsebina trenutno aktivnega vnosa (**HEAD**). Spremembe, ki jih nauredimo na delovni kopiji lahko zabeležimo v nov vnos.
- **Oddaljen repozitorij** (angl. **remote**) je povezava(url) na isti repozitorij na drugem računalniku(ponavadi strežniku), s katerim lahko izmenjujemo vsebino.

5. Git ukazi kot operacije na grafu

Ko smo opremljeni z razumevanjem podatkovnega modela Gita, razložimo kaj pomenijo posamezne operacije, ki jih Git omogoča. Ukazov ne bom prevajal, ampak jih bom navedel kot jih pozna program `git`.

5.1. Checkout

Ukaz

```
git checkout referenca
```

spremeni datoteke v delovni kopiji tako, da se ujemajo z vnosom, na katerega kaže referenca. Poleg tega prestavi oznako HEAD na isti vnos. Če je referenca veja, jo nastavi, kot aktivno vejo. Če je referenca oznaka ali zgoščena vrednost vnosa, priredimo v stanje brez aktivne veje (angl. *deteached HEAD*).

5.2. Commit

Ukaz

```
git commit -m "Sporočilo za vnos"
```

ustvari nov vnos, ki kaže na stanje v čakalnici (angl. staging area ali index). V zgodovinskem grafu ustvari novo vozlišče, ki je povezano s prejšnjim vnosom. Poleg tega prestavi aktivno vejo in oznako HEAD na novo ustvarjeni vnos.

5.3. Add

Ukaz

```
git add bla.txt
```

doda vsebino spremenjene datoteke `bla.txt` v čakalnico. Ukaz ne spreminja zgodovinskega grafa, pač pa doda novo vsebino in datotečna drevesa, ki vsebujejo spremembe v shrambo objektov. Vsebina čakalnice bo zabeležena v naslednjem vnosu.

5.4. Pull

Ukaz

```
git pull
```

pobere vsebino(objekte in reference) iz oddaljenega repozitorija in uskladi lokalno vejo z oddaljeno. Shrambi objektov se preprosto doda nove objekte, ki so v oddaljeni veji. Če je lokalna veja prednik oddaljene, se lokalna veja enostavno prestavi, da kaže na isti vnos, kot oddaljena veja. V nasprotnem primeru, mora uporabnik posredovati in razrešiti morebitne konflikte.

5.5. Push

Ukaz

```
git push
```

potisne novo vsebino na oddaljeni repozitorij. Push deluje obratno kot `pull`. Ukaz je uspešno izveden le, če je oddaljena veja prednica lokalne veje in ni konfliktov.

5.6. Fetch

Ukaz

```
git fetch
```

pobere novo vsebino (vnose, veje in oznake) iz oddaljenega repozitorija. Pri tem ne more priti do konfliktov, ker git preprosto doda nove objekte v shrambo in obstoječih objektov nikakor ne spreminja. Oddaljenim vejam in oznakam preprosto doda predpono z imenom oddaljenega repozitorija.

6. Reset

Ukaz

```
git reset referenca
```

spremeni kam kaže trenutno izbrana veja. Ukaz ne spremeni zgodovinskega drevesa, ampak le to, na kateri vnos kaže trenutno izbrana veja.

6.1. Merge

Ukaz

```
git merge referenca
```

ustvari nov vnos, ki združi dve ločeni veji v eno (trenutno izbrano in referenco). Nov vnos ima dva starša: vnos na katerega kaže trenutna veja in vnos, na katerega kaže referenca. Če pride do konfliktov, jih mora uporabnik sam razrešiti, preden se ustvari nov vnos.

6.2. Rebase

Ukaz

```
git rebase referenca
```

prestavi vnose v trenutno izbrani veji tako, da so potomci vnosa, na katerega kaže referenca. Med ukazi, ki smo jih spoznali, je ta ukaz edini, ki lahko povzroči izgubo podatkov. Običajno ukazi le dodajajo nove vnose in predstavljajo reference. Zato je večina ukazov v Gitu varna, v smislu, da jih lahko kasneje prekličemo in pridemo nazaj na prejšnje stanje. Ukaz rebase pa spremeni zgodovino in ga ne moremo preklicati, saj trenutne vnose nadomesti z novimi in stare vnose pobriše².

6.3. Delo z Git

Opis dela z Gitom presega namen tega dokumenta. Zato vas raje preusmerimo na uradno dokumentacijo:

<https://git-scm.com/cheat-sheet>

Pri pripravi dokumenta sem uporabil Gemini 3. Vse odgovore sem preveril in uredil po svoje.

Sledi še skica, ki povzame vse komponente Git repozitorija.

²Obstaja enostaven način, da tudi rebase lahko prekličemo. Na zadnji vnos, ki ga želimo prestaviti preprosto postavimo novo referenco(vejo ali oznako). To povzroči, da se stari vnosi ne pobrišejo tudi, ko se izvede ukaz rebase.

