RWSH - Really Weird Shell

Un shell alternativ pentru sisteme UNIX-like

Tudor-Ioan Roman

Cuprins

Introducere - shell-ul

Codul poate fi găsit pe https://git.sr.ht/~tudor/rwsh

Shell-ul, sau linia de comandă, este interfața textuală a unui sistem de operare. Prin acesta, utilizatorul poate să execute programe sub formă de *comenzi*, sau să execute *script-uri*. Aceste comenzi pot fi ori scrise în timpul execuției shell-ului (modul interactiv), sau pot fi înșiruite într-un fișier, numit *script*. În cazul cel din urmă, shell-ul are rolul de interpretor pentru un limbaj special de programare.

Programele executate de către shell pot fi înlănțuite. În mod normal, un program citește date de la tastatură (intrarea standard) și scrie date pe ecran (ieșirea standard). Într-un lanț de programe, primul program citește de la tastatură, iar rezultatul care ar trebui scris pe ecran este în schimb transmis ca date de intrare pentru următorul program, ca și cum ar fi fost scrise la tastatură. Acest lanț se numește *pipe*, iar operația de înlănțuire se numește *piping*.

Comenzile date shell-ului pot coopera astfel pentru a ajunge la un rezultat final. Utilizatorul poate să prelucreze date și să administreze sistemul cu o eficiență ridicată.

Exemple de astfel de shell-uri sunt GNU `bash` (Bourne Again Shell), `csh` (C Shell), `ksh` (Korn Shell), `zsh` (Z Shell), `fish` (Friendly Interactive Shell) etc.

> În blocurile de cod care urmează, textul precedat de simbolul '#' formează comentariile.

Prelucrarea textului

Pe platformele descendente din UNIX, precum Linux și MacOS, programele care operează în modul text se bazează pe schimbul și prelucrarea informației de tip *text simplu*, fără alte formate binare (ca pe Windows). Fișierele de configurație pentru programe sunt exprimate în text simplu, cât și documentele cum ar fi manualul sistemului, care sunt exprimate într-un limbaj special, spre deosebire de programe precum *Microsoft Word* care stochează documentele într-un format binar.

Exemplu de cooperare între programe: afișarea tuturor fișierelor dintr-un folder care conțin `infoeducatie` în nume, în ordine inversă:

```
ls | grep infoeducatie | sort -r
```

Această linie de comandă conține trei comenzi: `ls`, `grep infoeducatie` și `sort -r`. Cele trei comenzi sunt legate între ele prin operatorul `|` (pipe). Operatorul pipe capturează rezultatul comenzii din stânga, și, în loc să îl afișeze, îl oferă ca date de intrare programului din dreapta, ca și cum ar fi datele date de la tastatură.

Comanda `ls` afișează fișierele din directorul curent. `grep infoeducatie` afișează șirurile de caractere de la citire care conțin subșirul "infoeducatie", iar `sort -r` ordonează în ordine lexicografică inversă. Când comanda `ls` "afișează" fișierele, textul este dat ca intrare comenzii `grep infoeducatie`, iar aceasta la rândul ei furnizează comenzii `sort -r` ca date de intrare fișierele care in denumirea lor conțin subșirul "infoeducatie". La final, rezultatul comenzii `sort -r` este afișat pe ecran.

Acest mod de funcționare al shell-ului (*piping*) se bazează pe faptul ca majoritatea programelor operează pe text, furnizând, filtrând și prelucrând text. `ls` furnizează text, `grep` filtrează, iar `sort` prelucrează (ordonează). Programele care operează pe text includ și uneltele de administrare a sistemului. Prin metoda *piping*-ului se pot realiza programe (*script-uri*) eficiente.

RWSH - prin ce diferă

RWSH include propriile facilități de prelucrare a textului, care operează într-un mod inedit, diferit de oricare alt shell sau program de prelucrare al textului, facilități inspirate din limbajul de prelucrare al textului folosit de `sam`, editorul de texte al sistemului de operare experimental *Plan 9*, dezvoltat în anii '80 de Laboratoarele Bell.

În mod tradițional, marea majoritate a programelor care operează pe text prelucrează datele linie cu linie. În unele cazuri, aceasta abordare poate fi ineficientă și pentru procesor, dar și pentru programator.

Expresiile regulate structurale (structural regular expressions)

RWSH folosește un sub-limbaj prin care poate fi exprimată structura textului pe care dorim să operăm. Un alt mecanism foarte important este cel al

expresiilor regulate (regular expressions, pe scurt *regex*). Acestea sunt șiruri de caractere, exprimate într-un limbaj special, care definesc un *șablon de căutare*. Aceste expresii regulate, extinse cu facilități de descriere a structurii, dau naștere *expresiilor regulate structurale* (structural regular expressions).

Acest sub-limbaj include operații de prelucrare a textului, care pot fi combinate cu programele convenționale.

Exemplu: înlocuirea numelui "Tudor" cu "Ioan" într-un document.

```
cat document.txt |> ,x/Tudor/ c/Ioan/ |> ,p
```

Modul de funcționare

O comandă *pizza* este formată dintr-o *adresă* și o *operație*. Adresa este o expresie regulată structurală, iar operația este identificată printr-o literă și poate avea parametri. Aceste comenzi sunt înlănțuite prin operatorul `|>`, numit *operatorul pizza* (pentru că seamănă cu o felie de pizza). Adresa poate fi omisă, fiind folosită adresa ultimei comenzi, numită *dot*. Intern, adresa este o pereche de numere: poziția de început, și poziția de sfârșit, în caractere de la începutul fișierului. Dot este setată atunci când se specifică în mod explicit adresa unei comenzi, și la finalul execuției comenzii. De exemplu, comanda `c`, care înlocuiește textul situat la adresa *dot* cu textul dat ca parametru, setează la final *dot* ca adresa la care se află textul cel nou.

În continuare, voi ilustra exemplul de mai sus:

`cat document.txt` invocă programul `cat`, care afișează pe ecran conținutul fișierului `document.txt`.

Comanda, fiind urmată de operatorul pizza (`|>`), conținutul fișierului, în loc să fie afișat pe ecran, va fi pasat comenzilor pizza care urmează.

Prima comandă, `,x/Tudor/ c/Ioan/` are adresa `,`, care se referă la datele de intrare în întregime. Operația este operația buclă, notată prin `x`. Ea primește doi parametri: primul este o expresie regulată, al doilea este o altă comanda. Comanda `x` executa comanda transmisă ca parametru pentru fiecare subșir care se potrivește cu expresia regulată dată. Comanda `c/Ioan/` schimbă subșirul cu textul "Ioan".

Următoarea comanda din șir este `,p`, care afișează textul dat în întregime.

Pe lângă operația `c`, mai există operațiile `a`, `d` și `i`, care adaugă după dot, șterg textul din dot și respectiv inserează înaintea dot-ului.

Analogul operației `x` este `y`, care execută o comandă pe subșirurile care nu se potrivesc cu expresia regulată.

O altă abilitate specială este cea de a executa comenzi *pizza* în paralel, adică se execută mai multe comenzi pe același text, iar rezultatele fiecăreia se aplică cumulat.

Exemplu practic: înlocuirea tuturor aparițiilor numelui "Tudor" cu "Andrei" și "Andrei" cu "Tudor":

```
cat text.txt |> ,x/Tudor|Andrei/ {
    g/Tudor/ c/Andrei/
    g/Andrei/ c/Tudor/
} |> ,p
```

Unde `text.txt` conține:

```
Tudor este prietenul lui Andrei. Tudor îi oferă
lui Andrei o bomboană. Alex vrea și el una, dar Tudor a rămas fără.
Andrei îi este recunoscător.
```

Programul va afișa pe ecran:

```
Andrei este prietenul lui Tudor. Andrei îi oferă
lui Tudor o bomboană. Alex vrea și el una, dar Andrei a rămas fără.
Tudor îi este recunoscător.
```

În cuvinte, programul de mai sus poate fi descris în următorul mod:

- Pentru fiecare apariție a cuvântului "Tudor" sau "Andrei"...
  - ... dacă este "Tudor", atunci schimbă-l cu "Andrei".
  - ... dacă este "Andrei", atunci schimbă-l cu "Tudor".
- Afișează rezultatul.

Comenzile aflate intre acolade sunt cele executate *în paralel*. Efectul fiecărei comenzi este înregistrat într-un jurnal sub formă de vector. Acestea sunt interclasate și la final efectele sunt aplicate.

Un alt exemplu este să vedem de câte ori și unde apare cuvântul “linux” în jurnalul sistemului:

```
dmesg |> ,x/linux/ {
    =
    +-p
}
```

Date de ieșire:

```
#183,#188
[    0.000000] Command line: \\vmlinuz-linux-zen rw root=UUID=dba92c4c-35bc-4b73-9
#288,#293
[    0.000000] Command line: \\vmlinuz-linux-zen rw root=UUID=dba92c4c-35bc-4b73-9
#16485,#16490
[    0.000000] Kernel command line: \\vmlinuz-linux-zen rw root=UUID=dba92c4c-35bc
#16590,#16595
[    0.000000] Kernel command line: \\vmlinuz-linux-zen rw root=UUID=dba92c4c-35bc
#32278,#32283
[    0.540171] pps_core: Software ver. 5.3.6 - Copyright 2005-2007 Rodolfo Giomett
```

Operația `=` afișează valoarea percepută a lui *dot*. Mai concret, va afișa poziția cuvântului “linux”. Adresa `+-` a comenzii de afișare `p` înseamnă să meargă o linie în față și una în spate față de adresa subșirului găsit. Această tehnică este folosită pentru a afișa toată linia pe care am găsit subșirul. Altfel, s-ar fi afișat doar cuvântul “linux”.

Exemplu mai complex: afișarea liniilor care conțin cuvântul “linux”, dar fără timpul evenimentelor (textul dintre paranteze pătrate de la începutul fiecărei linii):

```
dmesg |> ,x/^\[.*\] /d |> ,x/linux/ {
    =
    +-p
} | lolcat
```

Rezultatul va fi pasat comenzii `lolcat` pentru a fi afișat în culorile curcubeului.

Avantajele abordării RWSH

Integrarea uneltelor de prelucrare a textului în cadrul shell-ului este inevitabilă. Uneltele convenționale, precum `grep`, `sed`, `cut`, `tr` etc. sunt folosite în aproape orice *shell script* din cauza funcțiilor elementare pe care le prestează. Majoritatea shell-urilor moderne prezintă unele astfel de funcționalități tocmai pentru că sunt indispensabile și sunt prea lungi de scris. Uitați care este diferența dintre eliminarea sufixului numelui unui fișier în mod tradițional vs. cu ajutorul sintaxei speciale din cel mai popular shell, GNU *bash*:

```
# Avem funcția proceseaza_fisier, care primeste numele fara extensie,
# si variabila filename, care tine minte numele complet al fisierului.

# Acestea sunt cele trei modalitati de apelare a functiei cu numele fisierului
# fara extensie.

proceseaza_fisier "$(echo "$filename" | cut -d'.' -f1-)" # in mod traditional

proceseaza_fisier "$(cut -d'.' -f1- <<< "$filename")" # modul traditional,
                                                      # dar folosind o scurtatura
                                                      # specifica bash


proceseaza_fisier "${filename%.in}" # folosind sintaxa speciala de eliminare
                                    # a sufixului din bash
```

În prima metodă, metoda pur tradițională, trebuie să folosim două procese de sistem pentru a afla denumirea fără extensie (un proces pentru `echo` și încă unul pentru `cut`). În cea de a doua, avem numai un singur proces, cel pentru `cut`, deoarece shell-ul are o sintaxa speciala pentru pasarea automata a unor date de intrare ca date de la tastatură (`<<<`).

Cea de a treia metodă nu folosește niciun proces.

Folosirea proceselor nu doar că încetinește mult programul, dar și este costisitoare pentru programatorul care scrie codul. Mai multă complexitate a codului poate duce la mai multe erori.

Un alt avantaj al abordării RWSH este că niciun alt program nu utilizează tehnica expresiilor regulate *structurale*, care fac procesarea textului mai eficientă și mai citibilă.

Funcționalitatea de limbaj de programare

Pentru moment, în limita timpului disponibil, am reușit să implementez variabilele, șirurile de caractere, operațiile matematice, blocurile de cod și blocurile decizionale (`if`).

Variabilele

Valorile sunt atribuite variabilelor cu comanda `let`. Variabilele sunt declarate automat la atribuire.

```
let nume Tudor
echo "Salut, $nume!" # Va afisa "Salut, Tudor!"

let nume Andei
echo "Salut, $nume!" # Va afisa "Salut, Andrei!"
```

Pentru a folosi o variabilă, numele ei va fi precedat de simbolul `$`.

Pentru a șterge variabila, se va folosi comanda `unset`: `unset nume`.

Notă: există o variabilă specială, numită `?`. Ea ține minte "exit code"-ul ultimei comenzi executate. Comanda precedentă se consideră executată cu succes dacă `?` va fi egal cu 0.

Șirurile de caractere

Parametrii comenzilor date sunt exprimați ca șiruri de caractere separate prin spațiu. Pentru a putea folosi șiruri de caractere cu caractere speciale și spații, acestea vor fi înconjurate de ghilimele (") sau apostrofuri (').

Apostrofurile diferă de ghilimele prin faptul că în șirurile de caractere cu apostrofuri, cuvintele precedate de $ nu vor fi tratate ca variabile.

```
let nume Tudor
echo Salut, $nume! # Va afisa "Salut, Tudor!"
                   # Comanda echo primeste doi parametri: "Salut," si "Tudor!"

echo "Salut, $nume!" # Va afisa tot "Salut, Tudor!"
                     # De data asta, echo primeste un singur parametru:
                     # "Salut, Tudor!"

echo 'Salut, $nume!' # Va afisa "Salut, $nume!"
                     # echo primeste un singur parametru.
```

Șirurile de caractere pot fi lipite pentru a forma unul singur, respectând regulile fiecăruia:

```
let nume Tudor

echo Salut", $nume"'!' # Va afisa tot "Salut, Tudor!"
                       # Comanda echo primeste un singur parametru
```

Șirurile de caractere simple și cele între ghilimele pot de asemenea să conțină rezultatul unei comenzi:

```
echo Este ora $(date +%H:%M) # Va afisa "Este ora 11:27"
                             # Comanda echo primeste 3 parametri
```

Notă: dacă un șir de caractere simplu (fără ghilimele sau apostrofuri) conține la început o cale de fișier care începe cu caracterul ~, tilda va fi înlocuită de calea către directorul utilizatorului. Exemplu:

```
ls ~/src # Afiseaza conținutul folder-ului /home/tudor/src
```

Blocurile de cod și blocurile `if`

Sintaxa pentru un bloc `if` este `if (condiție) comandă_de_executat`.

Condiția este o comandă. Dacă "exit code"-ul comenzii din condiție este 0, condiția este validă, iar comanda se va executa.

Dacă vrem să executăm mai multe comenzi, vom folosi blocul de cod, scris între acolade:

```
if (condiție) {
    o_comanda
    a_doua_comanda
}
```

Putem și să executăm cod dacă condiția nu se verifică, folosind blocul `else`, și să punem condiții în plus cu `else if`:

```
if (condiție) fa_ceva
else if (alta_condiție) fa_altceva
else nu_avem_incotro
```

Cum condiția este o comandă, putem sa folosim pipe-uri, operatori pizza, etc.

Operațiile matematice

Operațiile matematice se fac cu comanda `calc`. Putem stoca rezultatul într-o variabila astfel:

```
let a 2
let b 3
let c $(calc $a + $b)

echo "Rezultatul este $c" # Va afisa "Rezultatul este 5"
```

Detalii tehnice

Singura cerință de sistem este un sistem de operare UNIX-like, precum Linux, MacOS, FreeBSD etc.

Programul este scris în limbajul de programare Rust, un limbaj similar cu C++ care pune accent pe corectitudinea programului și a modelului de memorie. Cum shell-ul este un program cheie în orice sistem de calcul, acesta nu trebuie să aibă erori de memorie sau probleme de securitate (a se vedea: Shellshock)

Pentru a asigura siguranța codului și sănătatea minții, folosesc teste automate pentru a detecta bug-uri în cod. Acestea se execută cu comanda `cargo test` din Rust și cu script-ul `run_examples.sh` din folder-ul `examples`.

Librăriile folosite includ `nix` pentru funcțiile de bibliotecă pentru sistemul de operare, `regex` pentru expresiile regulate *simple*, și `calculate` pentru funcția de calculator.