Xử lý Văn bản trong Linux

Giới thiệu

Linux cung cấp nhiều công cụ mạnh mẽ để xử lý văn bản, tuân theo triết lý Unix: "Viết chương trình để xử lý các dòng văn bản vì đây là giao diện phổ quát". Hãy tìm hiểu các công cụ quan trọng nhất.

Các Lệnh Cơ Bản

1. cat - Đọc và Hiển thị Tệp

Công dụng: Đọc và hiển thị nội dung tệp văn bản

```
# Hiển thị nội dung file
cat file.txt

# Kết hợp nhiều file
cat file1.txt file2.txt > combined.txt
```

Giải thích: `cat` là lệnh đơn giản nhất để xem nội dung file. Dấu `>` dùng để chuyển hướng đầu ra vào file mới.

2. grep - Tìm Kiếm Văn Bản

Công dụng: Tìm kiếm văn bản theo từ khóa hoặc mẫu

```
# Tìm từ "error" trong file log
grep "error" system.log

# Tìm kiếm không phân biệt hoa thường
grep -i "Error" system.log

# Hiển thị số dòng
grep -n "error" system.log
```

Giải thích:

- `-i`: Không phân biệt chữ hoa/thường
- `-n`: Hiển thị số dòng
- `-r`: Tìm kiếm đệ quy trong thư mục

3. sed - Chỉnh Sửa Văn Bản

Công dụng: Thay thế hoặc chỉnh sửa văn bản theo mẫu

```
# Thay thế từ đầu tiên trong mỗi dòng
sed 's/old/new/' file.txt

# Thay thế tất cả các từ trùng khớp
sed 's/old/new/g' file.txt

# Lưu thay đổi vào file
sed -i 's/old/new/g' file.txt
```

Giải thích:

- `s/`: Bắt đầu lệnh thay thế
- `g`: Thay thế tất cả các từ khớp (không chỉ từ đầu tiên)
- `-i`: Lưu thay đổi trực tiếp vào file

4. awk - Xử Lý Văn Bản Theo Cột

Công dụng: Xử lý dữ liệu có cấu trúc cột

```
# In cột thứ nhất
awk '{print $1}' data.txt

# Xử lý file CSV
awk -F',' '{print $1 " - " $2}' data.csv

# Tính tổng cột số
awk '{sum += $3} END {print sum}' numbers.txt
```

Giải thích:

- `-F','`: Định nghĩa dấu phân cách (ví dụ: dấu phẩy trong CSV)
- `$1, $2`: Đại diện cho cột 1, cột 2
- `sum += $3`: Cộng dồn giá trị cột 3

5. cut - Cắt Văn Bản

Công dụng: Trích xuất phần cụ thể từ mỗi dòng

```
# Lấy cột đầu tiên (phân cách bằng dấu :)
cut -d: -f1 /etc/passwd

# Lấy ký tự từ vị trí 1-10
cut -c1-10 file.txt
```

Giải thích:

- `-d`: Chỉ định ký tự phân cách

- `-f`: Chọn cột (field)
- `-c`: Chọn ký tự

6. head và tail - Xem Đầu và Cuối File

Công dụng: Xem một số dòng đầu hoặc cuối của file

```
# Xem 10 dòng đầu
head file.txt

# Xem 5 dòng đầu
head -n 5 file.txt

# Xem 10 dòng cuối và theo dõi thay đổi
tail -f log.txt
```

Giải thích:

- `-n`: Chỉ định số dòng
- `-f`: Theo dõi file realtime (thường dùng cho log)

7. sort và uniq - Sắp Xếp và Loại Bỏ Trùng Lặp

Công dụng: Sắp xếp dữ liệu và xử lý dòng trùng lặp

```
# Sắp xếp file
sort names.txt

# Sắp xếp số
sort -n numbers.txt

# Loại bỏ dòng trùng lặp
sort file.txt | uniq

# Đếm số lần xuất hiện
sort file.txt | uniq -c
```

Giải thích:

- `-n`: Sắp xếp số
- `-r`: Sắp xếp ngược
- `uniq -c`: Đếm số lần xuất hiện

8. wc - Đếm Từ

Công dụng: Đếm số dòng, từ, và ký tự

```
```

```
# Đếm tất cả
wc file.txt

# Chỉ đếm số dòng
wc -l file.txt
```

Giải thích:

- `-l`: Đếm dòng
- `-w`: Đếm từ
- `-c`: Đếm ký tự

Kết Hợp Các Lệnh

Linux cho phép kết hợp các lệnh bằng pipe (`|`):

```
# Tìm 5 file lớn nhất trong thư mục
ls -lh | sort -rh | head -n 5

# Đếm số lần xuất hiện của từ "error" trong log
grep -i "error" log.txt | wc -l

# Lọc và sắp xếp dữ liệu từ CSV
cat data.csv | cut -d',' -f1,2 | sort | uniq
```

Lời Khuyên

1. Luôn sao lưu file trước khi chỉnh sửa
2. Sử dụng `man` hoặc `--help` để xem hướng dẫn chi tiết
3. Thử lệnh với một phần nhỏ dữ liệu trước
4. Sử dụng pipe (`|`) để kết hợp các lệnh
5. Kiểm tra kết quả sau mỗi bước xử lý