Linux系统基础

第三天

陈健

2024年

数据处理

- 数据处理需要你知道哪些工具可以被用来达成特定数据处理的目的，并且明白如何组合使用这些工具
- 系统日志处理
  - ssh myserver journalctl
  - ssh myserver journalctl | grep ssh
  - ssh myserver 'journalctl | grep sshd | grep "Disconnected from"' | less
  - ssh myserver 'journalctl | grep sshd | grep "Disconnected from“’ > ssh.log
  - cat ssh.log | sed 's/.*Disconnected from //’ | less

正则表达式的常见匹配模式

. 除换行符之外的任意单个字符

* 匹配前面字符零次或多次

+ 匹配前面字符一次或多次

[abc] 匹配 a、b、c 中的任意一个

(RX1|RX2) 任何能够匹配RX1或RX2的结果

^ 行首

$ 行尾

正则表达式匹配示例

```
$ echo "aba" | sed 's/[ab]//'
ba

$ echo "abaczbda" | sed 's/[ab]//g'
czd

$ echo 'abcaba' | sed 's/(ab)*//g’
abcaba

$ echo 'abcaba' | sed -E 's/(ab)*//g'
ca
```

正则表达式匹配示例

```
$ echo 'abcababc' | sed -E 's/(ab|bc)*//g'
cc

$ echo 'abcabbc' | sed -E 's/(ab|bc)*//g'
c
```

正则表达式的贪婪模式及解决方法

$ **echo 'Disconnected from invalid user Disconnected from 192.168.1.2' | sed 's/.*Disconnected from //'**

192.168.1.2

$ **cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating)?user .* [^]+ port [0-9]+(\[preauth\])?$//'**

$ **cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating)?user (.*) [^]+ port [0-9]+(\[preauth\])?$/\2/'**

$ **cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating)?(user)?(.*)?[ ]?[0-9.]+ port [0-9]+(\[preauth\])?$/\3/'**

正则表达式在线调试工具

https://regex101.com

日志数据的进一步处理
——找出最常出现的用户名

```
$ cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating )?(user )?(.* )?[ ]?[0-9.]+ port [0-9]+( \[preauth\])?$/\3/'| sort | uniq –c
```

```
$ cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating )?(user )?(.* )?[ ]?[0-9.]+ port [0-9]+( \[preauth\])?$/\3/'| sort | uniq -c | sort -nk1,1 | tail -n 10
```

日志数据的进一步处理
——给出最常出现的用户名列表

```
$ cat ssh.log | sed -E
's/.*Disconnected from (invalid
|authenticating )?(user )?(.* )?[ ]?[0-
9.]+ port [0-9]+(
\[preauth\])?$/\3/'| sort | uniq -c |
sort -nk1,1 | tail -n 10 | awk '{print
$2}' | paste -sd,
```

日志数据的进一步处理
——给出以a开头以u结尾只出现一次的用户

```
$ cat ssh.log | sed -E
's/.*Disconnected from (invalid
|authenticating )?(user )?(.* )?[ ]?[0-
9.]+ port [0-9]+(
\[preauth\])?$/\3/'| sort | uniq -c |
sort -nk1,1 | awk '$1 == 1 && $2 ~
/^a[^ ]*u$/ { print $2 }' | paste -sd,
```

日志数据的进一步处理
——用awk语言实现wc –l的功能

```
$ cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating )?(user )?(.* )?[ ]?[0-9.]+ port [0-9]+( \[preauth\])?$/\3/'| sort | uniq -c | sort -nk1,1 | awk '$1 == 1 && $2 ~ /^a[^ ]*u$/ { print $2 }' | wc –l
```

```
$ cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating )?(user )?(.* )?[ ]?[0-9.]+ port [0-9]+( \[preauth\])?$/\3/'| sort | uniq -c | sort -nk1,1 | awk 'BEGIN { rows = 0 } $1 == 1 && $2 ~ /^a[^ ]*u$/ { rows += $1 } END { print rows }'
```

日志数据的进一步处理
——用bc命令实现数学计算

```
$ cat ssh.log | sed -E 's/.*Disconnected from (invalid |authenticating )?(user )?(.* )?[ ]?[0-9.]+ port [0-9]+( \[preauth\])?$/\3/'| sort | uniq -c | sort -nk1,1 | awk { print $1 }' | paste –sd+ | bc -l
```

课堂练习

统计words文件中包含至少四个a且不以s 结尾的单词个数。这些单词中，出现频率前三的末尾三个字母是什么？

作业3提交方法和截止期限

- 实验报告的文件名命名统一为：学号_lab03.pdf
- 提交截止期限：2024年7月27日零点
- 实验报告通过电子邮件发送给chenj@nju.edu.cn