链滴

awk 将大文件按照某一列的值快速划分到不同文件

作者：xiaowuzi0214
原文链接：https://ld246.com/article/1482239673961
来源网站：链滴

有一个1.5亿行的文件data，每一行都是固定的几列：

```
userid columnA columnB columnC
```

不同行的userid可能一样。

现在有个需求，需要把这个文件里面的每一行按照userid划分到不同的文件，如有一行的内容是：

```
123 A B C
```

则需要将这一行写入文件data_123里面。

最开始的方法比较暴力，如下：

```
while read line;do
    userid=`cat ${line} | awk '{print $1}'`
    echo ${line} >> data_${userid}
done<data
```

这个方法执行了一下，发现半天还没结束。后来计算了一下，大概一秒才处理600行，1.5亿行大概得理70个小时。。。

果断换个方案。后来在Linux命令大全上看了一下awk命令的说明，改成下面实现方案：

```
# 先划分成小文件,一个文件200w行
split -d -a3 -l2000000 data d_
for file in d_*;do
    awk '{print $0 >> "data_"$1}' ${file}
done
```

重新跑了一下，大概40分钟跑完了，速度提高100来倍。虽然还是很慢，但是对于shell菜鸟的我，已很满足了哈哈~~