巨資二 A 鄭雅綿 08170120

P1

1. 程式碼：awk -F":" '!a[$2]++{print$0}'user.txt

   說明： -F 指定輸入字元為分隔符，以”:”為分隔符，!表示否定，a[$2]查看第 2 行的值為 key，如果前面沒出現過則建立。!a[$0]++是將 a[$0]的值自動加 1 並返回他的值。以第 2 行為群組，印出所有筆資料的整筆但以第 2 行沒有重複為基準，只要第 2 行有重複出現過則只印出現的第一筆資料。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk -F":" '!a[$2]++{print$0}'
 user.txt
1:zhangsan
3:lisi
4:wangmazi
2:wangwu
```

2. 程式碼：awk -F":" 'ARGIND==1{a[$1]=$0;next}{if($1 in a){print a[$1]"\t"$0}}' user.txt consumer.txt

   說明：next 為下一筆 data，ARGIND 為指令中的檔案序號。 以”:”為分隔符，第 1 個檔案的第 1 行為單一值建立 (若第一行有重複出現過，以最後出現的為主)，若第 2 個檔案的第 1 行於前面建立的群組有出現過，則印出前面建立的群組加上”\t”及第 2 個檔案的整筆資料。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk -F":" 'ARGIND==1{a[$1]=$0
;next}{if($1 in a){print a[$1]"\t"$0}}' user.txt consumer.txt
1:zhangs1:15:20121213
2:wangwu2:20:20121213
3:lisi  3:100:20121213
4:wangma4:99:20121213
1:zhangs1:25:20121114
2:wangwu2:108:20121114
3:lisi  3:100:20121114
4:wangma4:66:20121114
1:zhangs1:15:20121213
1:zhangs1:115:20121114
```

3. 程式碼：awk -F":" '{a[$2]}END{asorti(a);for(i=1;i<=length(a);i++){print a[i]}}' user.txt

   說明：-F 指定輸入字元為分隔符，以”:”為分隔符，asorti 為對 data 的 key 值進行排序。第二行按序輸出且重複值刪除。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk -F":" '{a[$2]}END{asorti(
a);for(i=1;i<=length(a);i++){print a[i]}}' user.txt
lisi
wangmazi
wangwu
zhangsan
```

4. 程式碼：sed -e '1,2d' song.txt

說明：-e 直接在指令列模式上進行 sed 的動作編輯。1,2d 為刪除文件中開頭的 1~2 行。直接於 song.txt 檔案中編輯將 1~2 行刪除。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ sed -e '1,2d' song.txt
3, Mick Jagger, linuxlinux which one you choose, Price $7.90
4, Lady Gaga, unix is opensource., Price $6.30
5, Johnny Cash, unix is free os, Price $6.50
6, Elvis Presley, linux which one you choose linux, Price $6.30
7, John Lennon, learn linux system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
```

5. 程式碼：sed -e 's/6.30/7.30/g' song.txt>>song2.txt

說明：-e 直接在指令列模式上進行 sed 的動作編輯。s 為取代。/6.30/7.30/為要取代的正規表示法，將 6.30 取代成 7.30。g 為全域(每一行)。>> 是追加內容，增加內容至原本檔案(song2.txt)的底下。直接於 song.txt 檔案中編輯將 6.30 取代成 7.30 並追加至 song2.txt 檔案中。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ sed -e 's/6.30/7.30/g' song.t
xt>>song2.txt
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ cat song2.txt
1, Justin Timberlake, linux is great os, Price $7.30
2, Taylor Swift, learn operating system, Price $7.90
3, Mick Jagger, linuxlinux which one you choose, Price $7.90
4, Lady Gaga, unix is opensource., Price $7.30
5, Johnny Cash, unix is free os, Price $6.50
6, Elvis Presley, linux which one you choose linux, Price $7.30
7, John Lennon, learn linux system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
架構與技術檔案]$
```

6. 程式碼：sed 's/linux/Unix/' song.txt

說明：s 表示搜尋，也能夠進行取代的工作。/linux/Unix/為要取代的正規表示法，將 linux 取代成 Unix。為替換每一行中的第一個 linux 為 Unix。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ sed 's/linux/Unix/' song.txt
1, Justin Timberlake, Unix is great os, Price $6.30
2, Taylor Swift, learn operating system, Price $7.90
3, Mick Jagger, Unixlinux which one you choose, Price $7.90
4, Lady Gaga, unix is opensource., Price $6.30
5, Johnny Cash, unix is free os, Price $6.50
6, Elvis Presley, Unix which one you choose linux, Price $6.30
7, John Lennon, learn Unix system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
架構與技術檔案]$
```

7. 程式碼：sed 's/linux/Unix/2' song.txt

說明：s 表示搜尋，也能夠進行取代的工作。/linux/Unix/為要取代的正規表示法，將 linux 取代成 Unix。2 為第 2 個出現。因此為把每一行的第二個 linux 取代為 Unix。

```
7, John Lennon, learn Unix system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
架構與技術檔案]$ sed 's/linux/Unix/2' song.txt

1, Justin Timberlake, linux is great os, Price $6.30
2, Taylor Swift, learn operating system, Price $7.90
3, Mick Jagger, linuxUnix which one you choose, Price $7.90
4, Lady Gaga, unix is opensource., Price $6.30
5, Johnny Cash, unix is free os, Price $6.50
6, Elvis Presley, linux which one you choose Unix, Price $6.30
7, John Lennon, learn linux system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
架構與技術檔案]$
```

8. 程式碼：sed 's/linux/Unix/g' song.txt

說明：s 表示搜尋，也能夠進行取代的工作。/linux/Unix/為要取代的正規表示法，將 linux 取代成 Unix。g 為全域(每一行)。因此為把全部行的 linux 取代為 Unix。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ sed 's/linux/Unix/g' song.txt

1, Justin Timberlake, Unix is great os, Price $6.30
2, Taylor Swift, learn operating system, Price $7.90
3, Mick Jagger, UnixUnix which one you choose, Price $7.90
4, Lady Gaga, unix is opensource., Price $6.30
5, Johnny Cash, unix is free os, Price $6.50
6, Elvis Presley, Unix which one you choose Unix, Price $6.30
7, John Lennon, learn Unix system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
```

9. 程式碼：sed -n 's/linux/Unix/p' song.txt

說明：-n 經過 sed 特殊處理的那一行(或者動作)才會被列出來。s 表示搜尋，也能夠進行取代的工作。/linux/Unix/為要取代的正規表示法，將 linux 取代成 Unix 。p 顯示匹配正則表達式

的行。將每一行有 linux 字串的轉換成 Unix，並印出，若沒有執行前面取代的工作則該行不印出。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ sed -n 's/linux/Unix/p' song.
txt
1, Justin Timberlake, Unix is great os, Price $6.30
3, Mick Jagger, Unixlinux which one you choose, Price $7.90
6, Elvis Presley, Unix which one you choose linux, Price $6.30
7, John Lennon, learn Unix system, Price $7.90[ec2-user@ip-172-31-31-142 巨量資料處理
架構與技術檔案]$
```

P2

1. 程式碼：awk '{ gsub(/:+/, ":");print }' info.txt

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk '{ gsub(/:+/, ":"); print }' info.txt
Mike Harrington:(510) 548-1278:250:100:175
Christian Dobbins:(408) 538-2358:155:90:201
Susan Dalsass:(206) 654-6279:250:60:50
Archie McNichol:(206) 548-1348:250:100:175
Jody Savage:(206) 548-1278:15:188:150
Guy Quigley:(916) 343-6410:250:100:175
Dan Savage:(406) 298-7744:450:300:275
Nancy McNeil:(206) 548-1278:250:80:75
John Goldenrod:(916) 348-4278:250:100:175
Chet Main:(510) 548-5258:50:95:135
Tom Savage:(408) 926-3456:250:168:200
Elizabeth Stachelin:(916) 440-1763:175:75:300
```

2. 程式碼：awk '{ gsub(/:+/, " "); print $3$4}' info.txt

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk '{ gsub(/:+/, " "); print $3$4}' info.txt
(510)548-1278
(408)538-2358
(206)654-6279
(206)548-1348
(206)548-1278
(916)343-6410
(406)298-7744
(206)548-1278
(916)348-4278
(510)548-5258
(408)926-3456
(916)440-1763
```

3. 程式碼：awk -F: '/Dan/' info.txtlawk '{ gsub(/:+/, " "); print $3$4}'

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk -F: '/Dan/' info.txt|awk '{ gsub(/:+/, " "); print $3$4}'
(406)298-7744
```

4. 程式碼：awk -F: '/^J/' info.txtlawk '{ gsub(/:+/, " "); print $1$3$4}'

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk -F: '/^J/' info.txt|awk '{ gsub(/:+/, " "); print $1$3$4}'

Jody(206)548-1278
John(916)348-4278
```

5. 程式碼：awk '{gsub(/:+/, " ")}{print "$"$5"$"$6"$"$7}' info.txt

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ awk '{gsub(/:+/, " ")}{print "$"$5"$"$6"$"$7}'  info.txt
$250$100$175
$155$90$201
$250$60$50
$250$100$175
$15$188$150
$250$100$175
$450$300$275
$250$80$75
$250$100$175
$50$95$135
$250$168$200
$175$75$300
```

6. 程式碼：sed 's/John/Joanthan/g' info.txt

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$  sed 's/John/Joanthan/g' info
.txt
Mike Harrington::(510) 548-1278::250::::100::175
Christian Dobbins::::(408) 538-2358::155:90::201
Susan Dalsass:::::(206) 654-6279:250:::60:::50
Archie McNichol::(206) 548-1348:250:::100:175
Jody Savage::(206) 548-1278:15:188:150
Guy Quigley::(916) 343-6410:::::250:100:::::175
Dan Savage:(406) 298-7744:::450:300:::275
Nancy McNeil::(206) 548-1278:250:::80:75
Joanthan Goldenrod::(916) 348-4278:250:100:175
Chet Main:(510) 548-5258:::50:95:::135
Tom Savage::(408) 926-3456:::250:::168:200
Elizabeth Stachelin:::(916) 440-1763:175:::75:300
```

7. 程式碼：sed '/Lane/d' info.txt

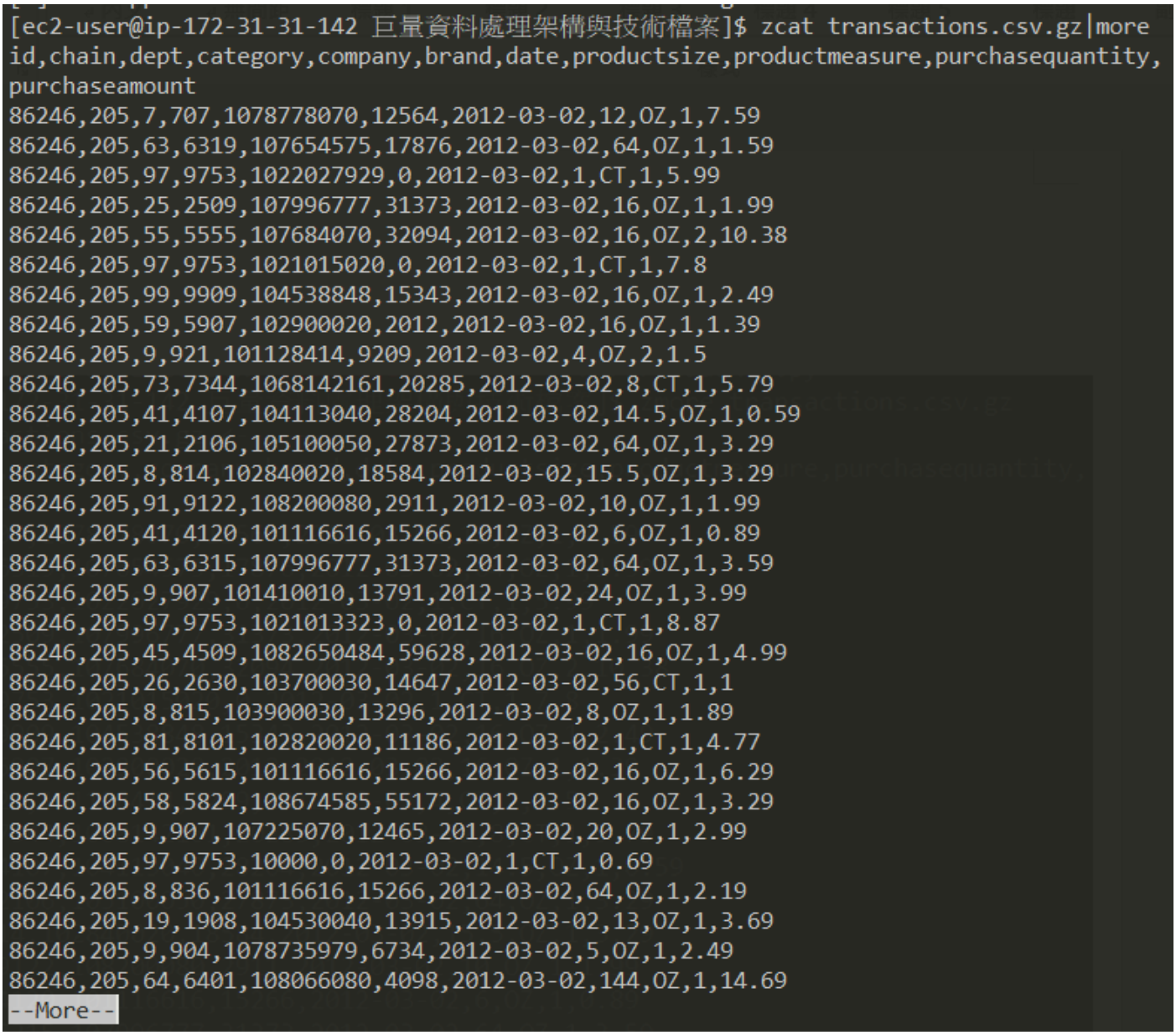
```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ sed '/Lane/d' info.txt
Mike Harrington::(510) 548-1278::250::::100::175
Christian Dobbins::::(408) 538-2358::155:90::201
Susan Dalsass:::::(206) 654-6279:250:::60:::50
Archie McNichol::(206) 548-1348:250:::100:175
Jody Savage::(206) 548-1278:15:188:150
Guy Quigley::(916) 343-6410:::::250:100:::::175
Dan Savage:(406) 298-7744:::450:300:::275
Nancy McNeil::(206) 548-1278:250:::80:75
Joanthan Goldenrod::(916) 348-4278:250:100:175
Chet Main:(510) 548-5258:::50:95:::135
Tom Savage::(408) 926-3456:::250:::168:200
Elizabeth Stachelin:::(916) 440-1763:175:::75:300
```

P3

1. 程式碼：zcat transactions.csv.gz|more

   說明：zcat 可察看壓縮檔內容，相當於一般檔案的 cat。more 可一次瀏覽一個頁面，該頁面是你的終端屏幕大小(可避免壓縮檔案資訊過多，終端機會一直執行載入的問題)。 想繼續瀏覽之後的行數可按 enter 或 space 繼續載入，要退出命令按 q 鍵。

   補充：相較於 less，more 載入的速度較慢一些，less 不會加載整個文件一次，且除了繼續瀏覽外，還可回推。



2. 程式碼：zcat transactions.csv.gzl wc －l

說明：wc 為壓縮檔的 wordcount ，–l 為 line，因此為讀取整份檔案後，計算壓縮檔內的行數。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz| wc
-l
349655790
```

3. 程式碼：time zcat transactions.csv.gzl wc -l

說明：可量測指令執行時間所需消耗的時間及系統資源等資訊。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ time zcat transactions.csv.gz
| wc -l
349655790

real    2m3.512s
user    1m57.409s
sys     0m5.766s
```

4. 程式碼：zcat transactions.csv.gzlhead -n 100ltail -n 50 > text.txt

說明：head 只看前幾行。Tail 只看後幾行。先解讀壓縮檔的前 100 列後暫時存取並再解讀倒數 50 列(51~100 列)，存成新的 text.txt 檔。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|head
 -n 100|tail -n 50 > text.txt
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ cat text.txt
86246,205,36,3630,104900040,23359,2012-03-02,144,OZ,1,5.79
86246,205,34,3410,107418272,16853,2012-03-02,7.5,OZ,1,2.19
86246,205,36,3611,108265787,4687,2012-03-02,405.6,OZ,1,6.29
86246,205,8,809,101300010,8164,2012-03-03,12,OZ,1,1.29
86246,205,71,7113,107940070,17521,2012-03-03,12,OZ,1,1
86246,205,60,6010,101116616,15266,2012-03-03,30,OZ,1,2.99
86246,205,81,8101,102820020,11186,2012-03-03,1,CT,2,9.54
86246,205,27,2702,102310020,13705,2012-03-03,17,LB,1,18.89
86246,205,81,8101,102610020,11320,2012-03-03,1,CT,1,3.59
86246,205,64,6408,101323212,40291,2012-03-03,0.75,LT,1,10.29
86246,205,36,3601,104900040,3809,2012-03-04,20,OZ,1,1.69
86246,205,56,5620,107592575,4368,2012-03-04,8,OZ,1,3.19
86246,205,41,4106,101116616,15266,2012-03-04,6.5,OZ,1,1.69
86246,205,63,6316,101116616,15266,2012-03-04,8,OZ,1,1.29
86246,205,33,3307,107641070,10160,2012-03-04,1.5,OZ,2,1
86246,205,23,2301,103340030,68309,2012-03-04,12,OZ,1,1.5
86246,205,44,4402,103100030,1358,2012-03-04,7,OZ,2,1.98
86246,205,26,2633,105400050,16048,2012-03-04,12,RL,1,8.39
86246,205,9,902,107641070,10160,2012-03-04,2.25,OZ,2,1.98
86246,205,56,5613,101116616,15266,2012-03-04,15,OZ,1,2.49
86246,205,8,811,104112949,3635,2012-03-04,24,OZ,1,3.19
86246,205,9,908,104980040,12592,2012-03-04,8,OZ,1,3.99
86246,205,64,6408,108289686,6276,2012-03-05,0.75,LT,1,8.29
86246,205,9,915,107312474,18613,2012-03-05,12,OZ,2,3.48
86246,205,63,6315,104138343,10091,2012-03-05,64,OZ,1,4.39
86246,205,16,1601,102200020,1827,2012-03-05,15,CT,1,1.29
86246,205,58,5899,107117474,22410,2012-03-05,16,OZ,5,13.05
86246,205,36,3611,108265787,4687,2012-03-05,202.8,OZ,1,4.49
86246,205,70,7002,1030813737,25162,2012-03-06,30,CT,1,2.39
86246,205,41,4105,101116616,15266,2012-03-06,14.75,OZ,6,8.34
86246,205,26,2634,103700030,3293,2012-03-06,6,RL,1,4.99
86246,205,41,4105,101116616,15266,2012-03-06,14.5,OZ,4,5.56
86246,205,73,7318,1030004535,10091,2012-03-06,32,CT,1,8.49
86246,205,41,4107,101116616,15266,2012-03-06,14.5,OZ,1,0.69
86246,205,56,5620,101116616,15266,2012-03-06,8,OZ,2,6.18
86246,205,41,4107,101116616,15266,2012-03-06,13.5,OZ,1,1.29
86246,205,75,7501,104589343,17521,2012-03-06,18,OZ,1,3.29
86246,205,37,3703,105100050,2820,2012-03-06,10.75,OZ,1,0.79
86246,205,26,2628,1077191373,13194,2012-03-06,1,RL,1,1
86246,205,29,2904,101600010,7860,2012-03-06,7.2,OZ,1,1.5
86246,205,72,7205,103500030,3830,2012-03-06,4.6,OZ,1,3.99
86246,205,33,3303,104166444,7179,2012-03-06,16,OZ,1,3.49
86246,205,63,6315,101116616,15266,2012-03-06,128,OZ,1,3.69
86246,205,29,2923,104100040,9814,2012-03-06,4.3,OZ,1,1.49
86246,205,21,2116,104300040,26420,2012-03-06,40.5,OZ,1,1.39
86246,205,97,9753,1022042424,0,2012-03-06,1,CT,3,34.99
86246,205,58,5823,107740070,2500,2012-03-06,2,OZ,1,0.89
86246,205,51,5134,107008575,26885,2012-03-06,7,OZ,1,2.69
86246,205,26,2628,104116343,8435,2012-03-06,1,RL,1,2.69
86246,205,97,9753,1022021020,0,2012-03-06,1,CT,3,27.3
```

5. 程式碼：zcat transactions.csv.gzlawk -F "," '{if(NR==1)print$0}'

說明：-F 指定輸入字元為分隔符，NR 為當前行數。讀取整份檔案後，以”,” 為分隔符，印出第 1 筆整列。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|awk -F "," '{if(NR==1)print$
0}'
id,chain,dept,category,company,brand,date,productsize,productmeasure,purchasequantity,purchaseamount
```

6. 程式碼：zcat transactions.csv.gzlawk -F "," '{if(NR%1000==1)print$0}'|more

說明：-F 指定輸入字元為分隔符，NR 為當前行數。%取餘數。讀取整份檔案後暫時存取，以”,” 為分隔符，將所有行數每 1000 個一數的第 1 筆印出整列，且只顯示一個頁面的數量。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|awk -F "," '{if(NR%1000==1)p
rint$0}'|more
id,chain,dept,category,company,brand,date,productsize,productmeasure,purchasequantity,purchaseamount
86246,205,5,516,101116616,15266,2012-04-29,8.9,OZ,2,5.78
86246,205,41,4109,105200050,19212,2012-07-04,15,OZ,2,2.38
86246,205,16,1604,102200020,6021,2012-09-07,15,CT,1,1.29
86246,205,0,0,10000,0,2012-11-24,0,,1,-0.27
86246,205,58,5814,104470040,10837,2013-02-26,4.4,OZ,2,3.98
86246,205,26,2613,103700030,10861,2013-03-08,27,CT,1,8.99
86246,205,5,501,104240040,40317,2013-03-14,37,OZ,1,5.29
86246,205,58,5811,107196777,54258,2013-03-22,16,OZ,1,5.99
86246,205,9,906,101020010,19622,2013-03-28,20,OZ,1,1.99
86246,205,58,5812,101590010,1366,2013-04-04,16,OZ,5,7.95
86246,205,6,610,107084777,11910,2013-04-12,15,OZ,2,5.58
86246,205,45,4509,104113040,28168,2013-04-19,12,OZ,2,13.98
86252,205,64,6401,107231171,5150,2012-03-27,72,OZ,1,7.79
86252,205,30,3009,1087694181,13478,2012-05-17,16,OZ,1,1
86252,205,33,3303,102068525,2892,2012-07-21,8,OZ,2,7.58
86252,205,12,1206,108000080,17292,2012-09-29,5,OZ,1,2.39
86252,205,55,5552,104113040,84481,2012-11-13,48,OZ,2,8.58
86252,205,53,5307,101450010,20864,2013-02-03,10,OZ,2,5.98
86252,205,97,9753,1022027929,0,2013-03-04,1,CT,3,14.29
86252,205,56,5617,104130343,38922,2013-03-09,16,OZ,1,3.49
86252,205,21,2117,108768484,2903,2013-03-12,60,OZ,2,6.78
86252,205,51,5122,101590010,1366,2013-03-16,24,OZ,1,4.99
86252,205,26,2622,102570020,20361,2013-03-20,10,CT,1,2.79
86252,205,60,6012,101800010,14029,2013-03-24,7.3,OZ,1,1.99
--More--
```

7. 程式碼：zcat transactions.csv.gzlawk -F "," '{print$3}'

說明：-F 指定輸入字元為分隔符，NR 為行號。以”,” 為分隔符，印出所有筆的第 3 行，且只顯示一個頁面的數量。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|awk -F "," '{print$3}'|more
dept
7
63
97
25
55
97
99
59
9
73
41
21
8
91
41
63
9
97
45
26
8
81
56
58
9
97
8
--More--
```

8. 程式碼：zcat transactions.csv.gzlhead -n 10000000lawk -F"," '$10>100{print$0}'lmore

說明：-F 指定輸入字元為分隔符。先解讀壓縮檔的前 10000000 列後暫時存取，以”,”為分隔符，若該筆第 10 行大於 100 則印出該列整列，且只顯示一個頁面的數量。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|head -n 10000000|awk -F"," '
$10>100{print$0}'|more
id,chain,dept,category,company,brand,date,productsize,productmeasure,purchasequantity,purchaseamount
59150399,20,21,2105,105200050,7054,2013-03-20,32,OZ,120,105.6
75671702,15,27,2701,101113212,491,2013-04-15,22,OZ,124,136.4
102750414,20,56,5614,102113020,10786,2013-02-28,1,OZ,384,96
105125900,95,21,2105,105200050,7054,2012-09-16,32,OZ,360,540
107682575,15,23,2301,102924323,10022,2012-08-25,7,OZ,140,46.67
113319656,3,27,2706,105000050,6132,2012-03-28,3,OZ,134,92.53
113319656,3,27,2706,105000050,6132,2012-11-08,3,OZ,106,74.2
119316259,15,36,3634,1078616272,7237,2012-06-17,20,OZ,132,128.18
119369344,4,27,2706,105000050,6132,2012-03-23,3,OZ,144,100.8
119369344,4,27,2706,105000050,6132,2012-04-06,3,OZ,120,84
119369344,4,27,2706,105000050,6132,2012-05-21,3,OZ,144,100.8
119369344,4,27,2706,105000050,6132,2012-07-14,3,OZ,120,84
119369344,4,27,2706,105000050,6132,2012-09-13,3,OZ,120,84
119369344,4,27,2706,105000050,6132,2012-10-07,3,OZ,144,90.02
119369344,4,27,2706,105000050,6132,2012-10-26,3,OZ,120,60.47
119369344,4,27,2706,105000050,6132,2012-12-07,3,OZ,120,98.82
119369344,4,27,2706,105000050,6132,2013-03-01,3,OZ,120,90
119369344,4,27,2706,105000050,6132,2013-03-21,3,OZ,120,90
120891520,4,27,2705,105000050,6775,2012-03-26,5.5,OZ,144,86.36
120891520,4,27,2705,105000050,6775,2012-05-21,5.5,OZ,144,84.19
120891520,4,27,2705,105000050,6775,2012-06-29,5.5,OZ,120,94.8
120891520,4,27,2705,105000050,6775,2012-09-28,5.5,OZ,120,72
120947783,4,27,2713,102113020,15704,2012-04-06,5.5,OZ,120,76.12
120947783,4,27,2713,102113020,15704,2012-08-24,5.5,OZ,120,77.64
120947783,4,27,2713,102113020,15704,2012-10-12,5.5,OZ,120,78.04
120947783,4,27,2713,102113020,15704,2012-11-17,5.5,OZ,120,78
124666310,15,36,3634,1078616272,7237,2012-06-29,20,OZ,144,141.59
```

9. 程式碼：zcat transactions.csv.gzlhead -n 10000000lawk -F"," '$7~/^2013-06-*/{print$0}'lmore

說明：-F 指定輸入字元為分隔符。~包含。先解讀壓縮檔的前 10000000 列後暫時存取，

以”,”為分隔符，輸出第 7 行為”2013-06-“開頭的該筆整列，且只顯示一個頁面的數量。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|head -n 10000000|awk -F"," '
$7~/^2013-06-*/{print$0}'|more
12262064,95,16,1604,101254616,4806,2013-06-03,16,CT,1,0.99
12262064,95,36,3618,104900040,23359,2013-06-03,20,OZ,1,1.79
12262064,95,91,9110,102113020,19478,2013-06-03,0.75,LT,1,4.89
12262064,95,18,1898,104133343,5379,2013-06-04,4,CT,1,10.99
12262064,95,63,6311,108259282,12297,2013-06-04,15.2,OZ,1,2.99
12262064,95,16,1604,102200020,6021,2013-06-04,15,CT,1,1.39
12262064,95,36,3618,104900040,23359,2013-06-05,20,OZ,1,1.79
12262064,95,4,421,101070010,9275,2013-06-05,10,OZ,1,3
12262064,95,4,421,107097070,56002,2013-06-05,8.3,OZ,1,0.99
12262064,95,21,2118,1073951070,16807,2013-06-05,20,OZ,2,2
12262064,95,4,421,107046272,16934,2013-06-05,8,OZ,1,2.19
12262064,95,59,5902,103450030,10165,2013-06-06,16,OZ,1,3
12262064,95,63,6315,102113020,10786,2013-06-06,128,OZ,1,3.19
12262064,95,63,6317,102529323,16410,2013-06-06,64,OZ,1,3.19
12262064,95,38,3802,102113020,15704,2013-06-06,2,LB,1,2.19
12262064,95,30,3010,104100040,19975,2013-06-06,16,OZ,1,2.5
12262064,95,18,1828,101980010,15181,2013-06-06,7.5,OZ,2,1.98
12262064,95,72,7212,1076687878,875,2013-06-06,1,CT,1,7.99
12262064,95,26,2634,103700030,3293,2013-06-06,9,RL,1,11.99
12262064,95,13,1306,103997838,2082,2013-06-06,1.38,LB,1,4.39
12262064,95,56,5611,104175747,10216,2013-06-06,6,OZ,1,4.99
12262064,95,57,5710,1089470080,21065,2013-06-06,5.3,OZ,4,4
12262064,95,57,5710,1068954464,6086,2013-06-06,5.3,OZ,2,2
12262064,95,72,7214,101254616,18738,2013-06-06,18,CT,1,0.99
12262064,95,63,6325,102113020,16397,2013-06-08,12,OZ,1,4.99
12262064,95,63,6328,102113020,16397,2013-06-08,9,OZ,1,3.49
12262064,95,36,3638,1085375989,204,2013-06-08,16,OZ,1,2.29
12262064,95,99,9904,1081204989,27736,2013-06-08,1,CT,1,2.99
12262064,95,63,6313,1082676686,33170,2013-06-08,9,OZ,1,3.99
12262064,95,21,2105,105200050,7054,2013-06-08,16.9,OZ,2,3
12262064,95,58,5834,107454575,58655,2013-06-08,8,OZ,1,5.99
12262064,95,21,2105,105200050,7054,2013-06-08,4,OZ,5,6.25
--More--
```

10. 程式碼：zcat transactions.csv.gzlawk -F"," '{print$7}'luniq -clmore

說明：-F 指定輸入字元為分隔符。uniq -c 忽略重複行後，在每行旁邊顯示重複的次數。解讀壓

縮檔後暫時存取，以”,”為分隔符，印出全部的第 7 行並刪除重複文字行，標示出每一資料的

重複次數(第 1 行)，且只顯示一個頁面的數量。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|awk -F"," '{print$7}'|uniq -
c|more
      1 date
     52 2012-03-02
      7 2012-03-03
     12 2012-03-04
      6 2012-03-05
     44 2012-03-06
     18 2012-03-07
      1 2012-03-08
     13 2012-03-09
     22 2012-03-10
     24 2012-03-11
      6 2012-03-12
      3 2012-03-13
     12 2012-03-14
      7 2012-03-15
      6 2012-03-16
     19 2012-03-17
     21 2012-03-18
     21 2012-03-19
      5 2012-03-20
     52 2012-03-21
      8 2012-03-22
      3 2012-03-23
     31 2012-03-24
     26 2012-03-25
     10 2012-03-26
     12 2012-03-27
      5 2012-03-28
     13 2012-03-30
     20 2012-03-31
     37 2012-04-01
     55 2012-04-02
      3 2012-04-03
     75 2012-04-04
     55 2012-04-05
     29 2012-04-06
--More--
```

11. 程式碼：zcat transactions.csv.gzlawk -F"," 'NR>1{print$11}'lhead -n10000000lawk 'BEGIN{max=0}{if($1>max)max=$1}END{print max}'

說明：-F 指定輸入字元為分隔符，NR 為行號。BEGIN 在 awk 讀取紀錄之前被執行，並執行一次。END 在讀取了所有記錄之後才執行，並執行一次。解讀壓縮檔後暫時存取。以”,”為分隔符，除了第一筆，印出所有筆的第 11 行後暫時存取。只執行前一個結果的前 10000000 筆後暫時存取。首先預設參數 max 為 0，若該比第一行資料大於 max 參數則覆蓋 max 參數，最後印出 max 參數。

```
[ec2-user@ip-172-31-31-142 巨量資料處理架構與技術檔案]$ zcat transactions.csv.gz|awk
-F"," 'NR>1{print$11}'|head -n 10000000|awk 'BEGIN{max=0}{if ($1>max) max=$1} END{pri
nt max}'
3000
```