Cut: Slitting a File Vertically

- The Features of the cut and paste commands will be illustrated with specific reference to the file shortlist,which stores the first five lines of emp.lst.
- In order To filter out First five Lines or Six Line from the file emp.lst we use this command
- $head -n 5 emp.lst | tee shortlist
- Or $head -n 6 emp.lst| tee shortlist
- Note that use of the tee facility saves the output in shortlist and also display it on the terminal.
- We can extract both columns and fields from this file with the cut command.
- Columns are specified with the –c option and fields with -f.

Cutting Columns (-c)

- To extract specific columns ,You need to follow the –c option with a list of column numbers,delimeted by comma.
- Ranges can also be used using the hyphen.
- Here’s how we extract the name and designation from shortlist
- cut –c 6-22, 24-32 shortlist
- a.k.Shukla g.m.
- jai sharma director
- sumit chakrobarty d.g.m.
- N.k.Gupta Chairman.
- .

Cutting Columns (-c)

- Note that there should be no whitespace in the column List.
- Moreover cut uses a special form for selecting a column from the beginning and upto the end of the Line.
- $cut -c -3 ,6-22,28-34,55- shortlist
- The expression 55- indicates column number 55 to the end of the Line.
- Similarly -3 is same as 1-3.
- The -c option is useful for fixed Length Lines.Most Unix Files (like \etc\passwd and \etc\group) don't contain fixed Length Lines.
- To extract useful data from these files you need to cut fields rather than columns.

Cutting Columns (-c)

- Cut uses the tab as the default field delimeter but can also work with a different delimeter.
- Two options need to be used here -d for the field delimeter and –f for the field List.This is how you cut the second and their fields of our sample file.
- $cut –d \| -f 2,3 shortlist | tee cutlist.
- The | was escaped to prevent the shell from interpreting it as the pipeline character; alternatively it can also be quoted (-d “|”).
- .Note:You must indicate cut whether you are extracting fields or columns.One of the options -f and -c must be specified.

Cutting Columns(-c)

- To cut out field numbers 1,4,5 and 6 and save the output in cutlist2,follow a simple procedure.
- $cut -d “|” -f 1,4 - shortlist > cutlist2
- Extracting User List from who output
- Cut can be used to extract the first word of a line by specifying the space as delimeter.For Eg.
- $who | cut –d “ “ -f 1.
- Cut is a powerful text manipulator often used in combination with other commands or filters.

Paste Command :Pasting Files

- What you cut with cut can be posted back with the paste command but vertically rather than horizontally.

- You can view two files side by side by passing them.
- In the previous topic ,cut was used to create the two files cutlist1 and cutlist2 containing two cut out portions of the same file.

$ paste cutlist1 cutlist2

- The original contents have been restored to some extent,except that the fields have different relative Locations and pasting has taken place on white space.
- Like cut ,paste also uses the tab as the default delimeter but you can specify one or more delimiters as –d.

Paste Command

- $paste -d “|” cutlist1 cutlist2
- Even though paste uses atleast two files for concatenating lines,the

data for one file can be supplied through the standard input.

- If for instance necessary fields from shortlist and piping

the output to paste. cutlist2 does’nt exist, you can provide the character stream by cutting out the

$cut –d \| -f 1,4- shortlist|paste –d”|” cutlist1-cutlist2

Paste Command

- Joining Lines(-s) :Paste is more useful than you might think.
- Consider the file that contains address of three persons,with three Lines for each.
- $ cat addressbook
- anup kumar Name
- anup_k@yahoo.com Email
- 24569083 Phone Number
- Vinod Sharma Name
- vinod_sharma@Hotmail.com Email
- 34586532 Phone Number.
- Etc

Paste Command

- Using the option of paste -s addressbook would join all of these nine lines to form a single Line.
- This won't be of much use,so we'll Learn to use the –d option with the

Multiple delimeters to join three Lines at a time.

.) If we specify the delimeter string as ||\n with -d,the delimeters are used in a circular manner .The first and second Lines would be joined with the | as delimeter and the same would be true for second and third Line.

-) The third and fourth line would be separated by a newline.After the List is exhausted it is reused.

Paste Command

- $paste -s –d “||\n” addressbook
- Anup kumar |anup_k@yahoo.com|24569083|
- vinod sharma|vinod_sharma@Hotmail.com|34586532
- Madhuri_bahl|madhuri@heavens.com |39034943
- Just see how paste works(with a single file this time) to concatenate lines in a specified manner.
-

Sort Options

- Unlike cut and paste ,sort uses one or more contiguous spaces as the default field separator(tab in cut and paste).
- We'll use the -t option to specify the delimeter.
- Sorting on Primary Key(-k)
- sort –t "|" –k 2 shortlist
- The sort order can be reversed with the –r option.The following sequence reverses a previous sorting order.
- $ sort -t"|" -r –k 2 shortlist.
- Sorting on secondary Key: You can sort on more than one key i,e you can provide a secondary key to start.

Sort Options

- $ sort -t “|” –k 3,3 ,-k 2,2 shortlist
- 5423|N.k.gupta| chairman | admin| 30/8/56 | 40000
- 5678|Sumit chakrobarty|D.g.M|Marketing |19/4/43|6000
- This sorts the file by designation and name -k 3,3 indicates that sorting starts on the field and ends on the same field.
- Sorting on Columns
- You can also specify a character Position within a field to be the beginning of sort.If you are to sort the file according to the year of birth then you need to sort on the seventh and eighth column positions within the field.

Sort Options

- $ sort -t “|” -k 5.7 5.8 shortlist
- The –k option also uses the form –k m.n where n is the character position in the mth field.So 5.7,5.8 means that sorting starts on column 7 of the fifth field.
- so 5.7,5.8 means that sorting starts on column 7 of the fifth field and ends on column 8.
- Numeric sort(-n) When sort acts on numerals ,strange things can happen.
- $sort numfile.
- 10
- 2
- 27
- 4

Sort : Ordering a File

- Sort is the ordering of data in ascending or descending sequence.The sort command orders a file.
- Like cut,It identifies fields and it can sort on specified fields.By default the entire Line is sorted.
- $Sort shortlist
- 2233|a.k.Shukla |g.m.| sales| 12/12/1952|7000

Sort Options

- This is probably not what you have expected,but the ASCII collating sequence places 1 above 2 ,and 2 above 4 .That's why 10 preceded 2 and 27 preceded 4.
- This can be overridden by the –k (numeric) option.
- $sort -n numfile
- 2 4 10 27
- Removing Repeated Lines(-u) : The -u option let's you remove repeated lines from a file .If you "cut" out the designation field from emp.lst, you can pipe it to sort to find out the unique designations that occur in the file.
- $cut –d "|" –f 3 emp.lst | sort –u |tee desig.lst

Sort Options

- Even though sort's output can be redirected to a file , we can use its –o option to specify the output filename.
- $sort -o sortedlist -k 3 shortlist.
- To check whether the file has actually been sorted in the default order,use the –c [check] option.
- $ sort –c shortlist.
- You can also add the –k option to the above to check whether a specified field is sorted
- $sort -t "|" –c –k 2 shortlist

Sort Options

Sort Options	Description
-t char	Uses delimeter char to identify fields sorts on Nth Field.
-k n	Sorts on nth field
-k m,n	Starts sort on mth Field and ends sort on nth field.
-k m.n	Starts sort on nth column of m th field
-u	Removes repeated Lines
-n	Sorts Numerically
-r	Reverses Sort Order
-f	Folds Lowercase to equivalent uppercase
-m list	Merges sorted Files in List.
-o flname	Places Output in file Filename

Sort Options

- When Sort is used with Multiple Filenames as arguments, it concatenates them and sorts them collectively.
- The –m (merge) option can merge two or more files that are sorted individually.
- sort -m foo1 foo2 foo3
- This command will run faster than the one used without the –m option only if the three files are sorted.
-

uniq :Locate Repeated Non-Repeated Lines

- When you concatenate or Merge Files ,You'll face the problem of duplicate entries creeping in.You saw how sort removes them with the –u option.
- Unix offers a special tool to handle these Lines the uniq command.
- $cat dept.lst
- uniq simply fetches one copy of line each and writes it to standard output.
- Since unique requires a sorted file as input the general procedure is to sort a file and pipe its output to uniq.
- The following pipeline also produces the same output,except that the output is saved to a file.

Uniq:Locate Repeated Non-Repeated Lines

- $sort dept.lst | uniq - uniqlist.
- unique is indeed unique;if provided with two filenames as arguments ,uniq will read the first file and write its output to the second.
- Hence it reads from the standard input and writes to unique list.
- Selecting Non-Repudiated Lines
- To determine the designation that occurs uniquely in emp.lst.cut out the third field,sort it and then pipe it to uniq.
- The –u (unique) option selects only lines that are not repeated.
- $cut -d “|” –f 3 emp.lst|sort|uniq -u
- Chairman

Selecting the Duplicate Lines(-d)

- The –d(duplicate) option selects only one copy of the repeated Lines.
- $cut –d “|” –f 3 emp.lst | sort |uniq -d
- D.g.m.
- Director
- Executive
- G.m.
- Manager
- This is an extremely useful option and we’ll make best use of it in an example that is taken up at the end of this chapter.
- It raises the possibility of printing a word-count list that displays the frequency of occurrence of each word.

Uniq Command

- Counting Frequency of Occurrence(-c). The –c [count] option displays the frequency of occurrence of all Lines, along with the Lines.
- $ cut –d”|” –f 3 emp.lst|sort|uniq -c .
- This is an extremely useful option and we’ll make best use of it in an example that is taken up at the end of this chapter. It raises the possibility of printing a word-count list that displays the frequency of occurrence of each word.

tr command: Translating characters

- The tr(translate) filter manipulates individual characters in a Line.
- More specifically it translates characters using one or two compact expressions.
- tr options expression1 expression2 standard input.
- Note that tr takes input only from standard input ; it does’nt take a filename as argument.By default it translates each character in expression 1 to its mapped counterpart in expression 2.
- The first character in the first expression is replaced with the first character in the second expression, and similarly for the other characters.

Tr command : Translating Characters

- Let's use tr to replace | with a ~ (tilde) and the / with a - .Simply specify two expressions containing these charcaters in the proper sequence.
- $ tr '|/' '~-' < emp.lst | head –n 3
- 2233 ~ a.k.Shukla ~ g.m. ~sales ~12-12-52 ~ 6000
- Note that the Lengths of the two expressions should be equal.If they are not ,the longer expressions will have unmapped characters. Single quotes are used here because no variable evaluation or command substitution is involved.

tr command

- It’s Just as easy to define the two expressions as two separate variables, and then evaluate them in double quotes.
- $expr1 = ‘|/’ ; expr2=‘~-’
- tr “$exp1” “$exp2” <emp.lst
- Like wild cards tr also accepts ranges in the expressions.The same rule apply; the character on the right of the – must have an ASCII value higher than that of the character on the Left.
- The escaping rules should also be obivious.The character class [need to be escaped if the special meaning is removed from it.

Changing case of Text

- Since it does’nt accept a filename as an argument ,the input has to be redirected from a file or pipe.
- $head -n 3 emp.lst | tr ‘[a-z]’[A-Z]’
- 2233|A.K.Shukla |G.M.|SALES |12/12/52|6000
- 9876 |Jai sharma|DIRECTOR|Production|12/3/50|7000
- 5678 |Sumit chakrobarty|D.G.M|Marketing|19/4/43|6000

Changing case of characters

- Deleting characters (-d):
- The file emp.lst has fields separated by delimeters and the date formatted in readable form with a /.
- In non-database setup's ,delimeters are not used and the date is generally represented as a six-character field in the format ddmmyy.
- To convert this file to the traditional format,use the –d (delete) option to delete the characters | and / from the file.
- The following command does it for the first three Lines.
- tr –d '|/' <emp.lst | head –n 3

Compressing Multiple Consecutive characters

- Unix tools work best with fields rather than columns .So it's preferable to use files with delimeted fields.
- In that case ,Lines need not be of fixed length;You can eliminate all redundant spaces with the –s option, which squeezes multiple consecutive occurrence of its argument to a single character.
- $ tr -s ' ' < emp.lst |head –n 3
- 2233 |a.k.Shukla|g.m.|sales|12/12/52|6000

Compressing Multiple Consecutive characters

- Tip: If you use the –s option to compress all contiguous spaces in the output of several unix commands, and then Use cut to extract individual fields from this compressed output.
- Complementing Values of Expressions(-c)
- Finally the –c(complement) option complements the set of characters in the expression.Thus ,to delete all characters except the | and /,you can combine the –d and –c option.
-

Complementing values of Expression(-c)

- tr –cd ‘|/’ < emp.lst
- tr has deleted all characters except the | and / from the file .The appearance of the prompt at the immediate end of output shows that the newline character has also not been spared.

Using ASCII Octal Values and Escape Sequences.

tr also uses octal values and escape sequences to represent characters.This facility is specially suited for using non-printable characters in the expression.

tr command

- So to have each field on a separate line,replace the | with the LF character (Octal value 012).
- tr '|' '\012' < emp.lst | head –n 6
- 2233
- A.k.Shukla
- G.m.
- Sales
- If you reverse the two expressions ,You'll make the newline character visible.Study these tr options closely,and you'll discover many areas where you can apply them.