

Sorting, Merging and Duplicates Detection using the XL Audit Commander

Version 1.54 (March 2008)

XL Audit Commander

data analysis made easier ...

XL Audit Commander

Sort Merge

Sorting and merging files is a commonly performed operation in order to prepare data for certain other analysis. One example use of sorting and merging relates to the detection of certain types of fraud patterns. The primary purpose of the tool is to support forensic accounting and fraud investigations, although it may find other uses in research and as a general purpose accounting support tool. This is *free software*, there is no license cost, and it may be used without restriction for any purpose, including commercial and educational.

The tool is installed as an Excel add-in, and as such requires Excel 2000 or later. The tool works only on Windows operating systems.

The techniques described here require that the data to be analyzed be in “tab separated value” format and also be sorted. If the data to be analyzed is not in tab separated value format, there is an explanation of how to convert data in the white paper located at http://www.ezrstats.com/doc/Process_To_Convert_Data_To_TSV.pdf.

This paper is divided into four sections – 1) How to use the sort procedure, 2) How to use the data merge procedure, 3) Detection of duplicates, and 4) sub-totals.

XL Audit Commander

Sort Merge

Sort Procedure	1
Overview	1
Form	2
Meanings of form elements:	2
System Limitations	3
Sort Parameters	3
Example 1	3
Example 2	3
Example 3	4
A completed sort form	4
Sort Messages	4
Merge Procedure	5
Overview	5
Form	5
Column Meanings	6
System Limitations	6
Merge Parameters	6
Duplicates	6
Duplicates – single sort key specification	6
Duplicates – three sort key specification	7
Sub-Totals	8

XL Audit Commander

Sort Merge

Sort Procedure

Overview

The purpose of the sort procedure is to sort text files which are in tab separated value format, and whose first row consists of column names. If the file to be sorted is not in this format it can generally be converted to this format. An article on converting data to tab separated value format is available at

http://www.ezrstats.com/doc/Process_To_Convert_Data_To_TSV.pdf, and a tutorial on conversion of data in print file format is also available at http://ezrstats.com/doc/Data_Conversion_Data_Integrity.pdf.

Sorting a file is done using an input form, and specifying three pieces of information:

1. Name of the input file
2. Name of the output file where the sorted values are to be stored
3. The sort parameters to be used

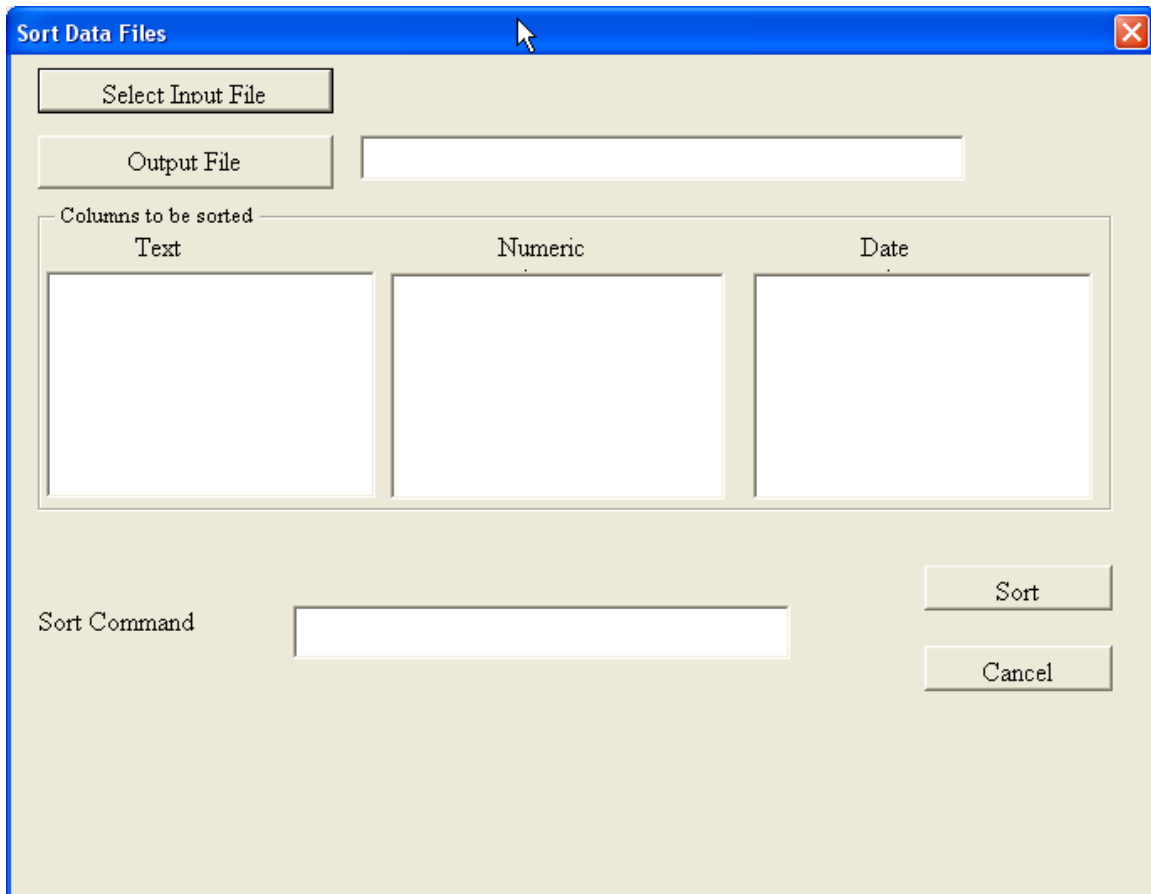
The name of the input file is obtained through a directory search to locate the file, whose name is then displayed as part of the form caption. The name of the output file can either be selected from browsing the directories and specifying a file name, or else typing in a file name directly. The sort parameters must be typed in, and are based on the column names displayed. There is a limit of 200 columns maximum which may be specified as sort columns. Each may be sorted in either ascending order (specified using a “+”) or descending order (“specified with a “=”). The default sort order, if none is specified, is ascending. Each sort column may also be specified as to its type, which may be either alphabetic (“A”), numeric “N” or date “D”. The default value for column type is alphabetic, if no column type was specified. The column type and sort order specifications are optional, but if specified, must immediately follow the column name separated by spaces. Order of these two sort parameters does not matter. Examples of valid sort parameters are provided in the section below “Sort Parameters”.

XL Audit Commander

Sort Merge

Form

A blank input form for sorting is shown below.



The screenshot shows a dialog box titled "Sort Data Files". It contains the following elements:

- A "Select Input File" button.
- An "Output File" button next to a text input field.
- A section titled "Columns to be sorted" containing three empty text boxes labeled "Text", "Numeric", and "Date".
- A "Sort Command" label next to a text input field.
- "Sort" and "Cancel" buttons.

Meanings of form elements:

The form contains four elements which must be used to sort a file:

1. The "select input file" button, which when clicked allows the selection of the input file to be sorted
2. The "output file" button which when clicked allows the selection of the output file name (where the sorted file is stored). The name of the file may also be typed in.
3. The "sort command" which is the specification of the sort parameters (more details in the section "Sort Parameters" below)
4. The "Sort" button, which when clicked will validate the parameters specified, and if valid, sort the file and store the results in the output file name.

XL Audit Commander

Sort Merge

System Limitations

The maximum number of columns which may be specified as columns for sorting is 200.

The maximum number of rows in a file which can be sorted is approximately 100,000,000.

Each numeric value to be sorted must be between -920 trillion and positive 920 trillion.

Only dates between AD 100 and AD 9999 are handled.

Sort Parameters

Sort parameters specify how the file is to be sorted. For each column to be sorted, up to three sort elements may be specified:

1. (Required) – the name of the column
2. (Optional) – the sort sequence for this column, either ascending “+” or descending “-“. If not specified, ascending is assumed
3. (Optional) – the column type of this column, either alphanumeric “A”, numeric “N” or date “D”. If not specified, alphanumeric is assumed.

If a column name has embedded spaces then the column name must be enclosed in quotation marks. Column names which are the same as sort parameter names are not allowed, i.e. “+”, “-“, “A”, “N” and “D”.

Example 1

The file is to be sorted on a single column named “Vendor” in ascending order. “Vendor” is an alphanumeric column. All of the sort parameters below are valid and equivalent:

Vendor
Vendor +
Vendor A
Vendor A +
Vendor + A

Example 2

Same sort parameter, except the vendor column should be sorted as a numeric value. Any of the sort parameters below could be used and are equivalent:

Vendor N

XL Audit Commander

Sort Merge

Vendor + N
Vendor N +

Example 3

The file is to be sorted first by Due Date (a date field) and then by Invoice Number, a numeric value. Due date is to be sorted ascending, invoice number is to be sorted descending. Any of the sort parameters below could be used and are equivalent:

“Due Date” + D “Invoice Number” N -
“Due Date” D “Invoice Number” - N
“Due Date” + D “Invoice Number” - N
“Due Date” D “Invoice Number” N -

A completed sort form

Sort Data in File: C:\Program Files\EZS\XLAC\data\t_CMADData.txt

Select Input File

Output File: C:\Program Files\EZS\XLAC\data\t_CMADData.srt

Columns to be sorted

Text	Numeric	Date
Payee	Check Number Amount	Invoice Date Due Date

Sort Parameters: "Check Number" + n "Invoice Date" d -

Sort

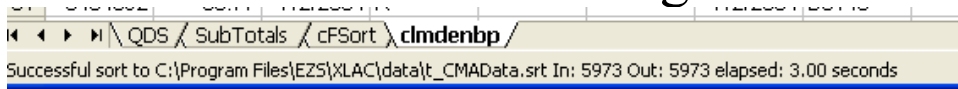
Cancel

Sort Messages

A status bar message is displayed which shows the success/failure of the sort, number of rows processed, elapsed time in seconds, etc. An example is shown below:

XL Audit Commander

Sort Merge



Merge Procedure

Overview

The purpose of the merge procedure is to combine two files, both in tab separated value format with the first row containing the column names into one output file. Both input files must be sorted in the same sequence. The output file will combine (merge) the two files into the same sorted order.

Form

The screenshot shows the 'Merge Data Files' dialog box. It has a blue title bar with a timer icon and a close button. The dialog contains three file selection fields: 'Select File 1', 'Select File 2', and 'Output File'. Below these is a section titled 'Columns to base merge on' with three columns: 'Text', 'Numeric', and 'Date'. At the bottom, there is a 'Merge columns' field, a 'Merge' button, and a 'Cancel' button.

XL Audit Commander

Sort Merge

Column Meanings

There are five form elements which must be used to merge the files:

1. “Select File 1” button, is used to specify the name of the first file to be merged.
2. “Select File 2” button, is used to specify the name of the second file to be merged.
3. “Output File” button, is used to specify the name of the file where the merged data is to be stored.
4. “Merge columns” button is used to provide the names of the columns which specify the sort order that the input files are sorted in.
5. “Merge” button is used to validate the input parameters, and if successful, merge the two files using the merge sequence specified.

System Limitations

The maximum number of columns which may be specified for merging is 200. Each numeric value in the specified merge sequence must be between -920 trillion and positive 920 trillion.

For date values specified as part of the merge sequence, only dates between AD 100 and AD 9999 are handled.

Merge Parameters

Merge parameters are the same as the sort parameters used when the input files were sorted.

Duplicates

For each sequence of items specified by the sort sequence, if two or more consecutive rows contain the same values for the specified sort sequence, then they are included in the report, along with a count of the number of duplicates identified. Note that “duplicate” means two consecutive rows in the input sorted file which contain the same sort key values, even though the remainder of the information in the row may be different. An example of a report for a single sort specification is shown below:

Duplicates – single sort key specification

XL Audit Commander

Sort Merge

BILLPROV	N
3403054	8
3404390	17
3404392	45
3404410	149
3404420	25
3404424	88
3404425	24
3404426	28
3404427	93
3404428	25
3404429	18
3404430	12
3404431	34
3404432	47
3404434	115
3404440	66
3404441	60

Duplicates – three sort key specification

This report lists all duplicates from a file which had been sorted by billing provider (“BILLPROV”) and identifies all such records which had a duplicate billing provider, along with a count of the total number of records which are duplicates. Note that the minimum value for the duplicate count is 2, meaning there are two records with the same sort key value.

The following test for duplicates was based upon the same provider number, same claim type and same date of service. The input file was sorted using the parameter “billprov + a clmtype + a tdos + d”.

The sorted output file was then analyzed for duplicates specifying the same sort parameters, i.e. “billprov + a clmtype + a tdos + d”. The output file from the analysis is as follows:

BILLPROV	CLMTYPE	FDOS	N
3403054	K	7/2/2004	2
3404390	K	7/1/2004	2
3404390	K	8/18/2004	2
3404390	K	9/8/2004	3
3404390	K	10/6/2004	2
3404390	K	11/30/2004	2
3404392	K	7/15/2004	4
3404392	K	8/17/2004	2
3404392	K	9/3/2004	2
3404392	K	10/27/2004	2

XL Audit Commander

Sort Merge

3404392	K	11/15/2004	2
3404410	K	7/21/2004	3
3404410	K	7/22/2004	8
3404410	K	8/3/2004	9
3404410	K	8/4/2004	5
3404410	K	8/13/2004	10

Sub-Totals

For each sequence of items specified by the sort sequence, if two or more consecutive rows contain the same values for the specified sort sequence, then they are included in the report, along with a sum of the values specified for the sub total column.

The following sub-total report was based upon the same provider number, same claim type and same date of service. The input file was sorted using the parameter “billprov + a clmtype + a tdos + d”.

The sorted output file was then sub-totaled, specifying the same sort parameters, i.e. “billprov + a clmtype + a tdos + d”. The sub-total report is as follows:

BILLPROV	CLMTYPE	FDOS	N	Sum	Avg	Min	Max	Std Dev	CV
3403054	K	7/2/2004	2	57.01	28.505	27.01	30	2.114249	7.4%
3403054	K	8/6/2004	2	57.01	28.505	27.01	30	2.114249	7.4%
3403054	K	9/17/2004	2	57.01	28.505	27.01	30	2.114249	7.4%
3403054	K	9/20/2004	2	57.01	28.505	27.01	30	2.114249	7.4%
3404390	K	7/1/2004	3	278.18	92.72667	51.1	149.7	51.06006	55.1%
3404390	K	8/18/2004	3	106.36	35.45333	27.01	41.35	7.501535	21.2%
3404390	K	9/8/2004	4	136.72	34.18	27.01	41.35	8.279203	24.2%
3404390	K	9/28/2004	2	193.45	96.725	77.38	116.07	27.35796	28.3%
3404390	K	10/6/2004	3	88.4	29.46667	27.01	31.39	2.238176	7.6%
3404390	K	11/30/2004	3	188.34	62.78	62.78	62.78	0	0.0%
3404392	K	6/8/2004	2	48.91	24.455	13.14	35.77	16.00183	65.4%
3404392	K	6/17/2004	1	78.11	78.11	78.11	78.11	16.00183	20.5%
3404392	K	7/2/2004	2	48.91	24.455	13.14	35.77	16.00183	65.4%
3404392	K	7/6/2004	1	62.78	62.78	62.78	62.78	16.00183	25.5%
3404392	K	7/15/2004	5	464.28	92.856	77.38	116.07	21.19139	22.8%
3404392	K	7/16/2004	2	232.14	116.07	116.07	116.07	0	0.0%
3404392	K	7/20/2004	1	30	30	30	30	0	0.0%
3404392	K	7/29/2004	2	154.76	77.38	77.38	77.38	0	0.0%