

TS 592

***Tips and Techniques for Processing CRSPAccess
CDROM Data with PROC DATASOURCE***

Author: Karen B. Hoeve

Updated: November 2001 by Kurt W. Jones

***Technical Support Division
Base Procedures and Macro Facility***

***SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513***

Abstract

Many PROC DATASOURCE users call Technical Support with questions about reading CRSPAccess CDROM data. This brief paper is designed to help users successfully convert CRSPAccess data into a SAS data set.

Introduction

PROC DATASOURCE is documented in:

“SAS OnlineDoc®”, Version 8, “SAS/ETS® User’s Guide”.

“SAS/ETS® User’s Guide”, Version 8, Volume 1, pages 449-538

“SAS/ETS® User’s Guide”, Version 6, Second Edition, pages 323-390.

It is necessary to read and understand the information in the documentation before applying the following information specific to CRSPAccess data.

The CRSP Supplied Conversion Utilities

PROC DATASOURCE cannot read the CRSPAccess CDROM data directly. However, CRSP supplies conversion routines that will convert the data into a format that PROC DATASOURCE can read.

These utilities are called `stk_dump_bin` (for the securities file) and `ind_dump_bin` (for the calendar file). CRSP provides documentation detailing these utilities.

From the CRSP/ SFA Guide:

Usage is:

```
stk_dump_bin path outfile setid porttype1 porttype2 porttype3 keep/reverse  
vms/unix/none [yyyy/yy tsopt1 tsopt2] [permfile]
```

```
ind_dump_bin path outfile indno1 indno2 indno3 keep/reverse vms/unix/none [yyyy/yy]
```

The items in brackets are optional and all parameters must be on one command line.

Please refer to the CRSP documentation before proceeding with these utilities.

PROC DATASOURCE

Once the CRSPAccess data has been successfully converted to the UNIX binary format with the above utilities, then PROC DATASOURCE can read them with the following FILETYPEs:

daily data:	CRSPDUS	(Daily Securities)
	CRSPDUI	(Daily Calendar Indices)
	CRSPDUA	(Daily Annual)
monthly data:	CRSPMUS	(Monthly Securities)
	CRSPMUI	(Monthly Calendar/Indices)
	CRSPMUA	(Monthly Annual)

The following is an example of the PROC DATASOURCE syntax that can be used to read the daily data (Please note that the CALENDAR/INDICES file must be listed first on the INFILE= statement):

```
FILENAME CALFILE 'filename1';
FILENAME SECFILE 'filename2' LRECL=36000;

PROC DATASOURCE FILETYPE=CRSPDUS
  INFILE=( CALFILE SECFILE )
  INTERVAL=DAY
  OUTSELECT=OFF
  OUT=CRSPVAR
  OUTBY=CRSPBY
  OUTCONT=CRSPCONT
  OUTEVENT=CRSPEV;
  KEEP _ALL_;
  KEEPEVENT _ALL_;
  RUN;
```

The most frequent problem that users have is specifying the wrong LRECL= on the FILENAME statement for the securities file. CRSP reports that the 1996 CRSP daily securities file had an LRECL of 34,768 in UNIX format. Their data uses [(number of days in the calendar) * 4] = [8686 * 4] = [34,744] bytes to hold the largest timeseries. There is another 16 byte overhead for the CUSIP (8 bytes), PERMNO (4 bytes) and Record ID (4 bytes). This results in 34,760 bytes of CRSP data. The UNIX binary format, which is necessary for PROC DATASOURCE, adds 4 leading and 4 trailing bytes. Thus the unix binary format LRECL is 34,768. We recommend adding 1,000 to the LRECL for each year following 1996, or use the actual LRECL which is supplied to the user when the stk_dump_bin and ind_dump_bin utilities complete processing.

According to CRSP the **monthly** securities file had an LRECL of about 27,688 for 1996 and it grows by about 400 bytes a year.

RESOURCES

The CRSPAccess data is available on CDROM for Windows and UNIX systems. If you convert the entire daily file, the resulting UNIX binary security file and calendar file are about 1.06GB. If you convert the entire daily file to a SAS data set, it will be about 12.0GB. Extra space is also needed during processing for the utility files that are created by PROC DATASOURCE in the WORK library. Most of these utility files will be deleted at the end of the procedure, and the remainder will be deleted when the SAS session is terminated. Depending on the resources available, it is not feasible for many sites to convert the entire daily file to a SAS data set.

Some users will also run into a 2GB file size limit imposed by their operating system. In order for a data set to exceed the 2GB limit on Windows NT, the hard drive where the data will be stored must be formatted with the NTFS format. The latest version of Windows 95 (sometimes referred to as Revision B) should also support files larger than 2GB as long as the file format is FAT32). SAS 6.12 on Digital UNIX will automatically support SAS data sets larger than 2GB. The following SAS Note explains what is needed to support large files on Solaris 2.6 (SunOS 5.6):

V6 SAS Notes

E126 - SAS 6.12 on Solaris 2.6 (Sunos 5.6) and files larger than 2 gigabytes

In order to use files larger than 2 gigabytes with SAS on Solaris 2.6, users will need to download the sasvlfs module from our website/ftp server. This file can be found in the following locations:

Web site: <http://www.sas.com/techsup/download/unix>
ftp: ftp.sas.com (user anonymous) cd to /techsup/download/unix

Retrieve this file in binary mode and place it the !SASROOT/sasexe/base/ directory on your local installation.

Then invoke sas with the
-largefile sasvlfs
invocation option. This option can be placed in the config.sas612 file.
