

A Simple and Adaptable Automated Warehouse Design

Author

Philip R Holland, Consultant
Holland Numerics Limited

Abstract

The SAS® programs used to populate a data warehouse are likely to be complex and require expert support. The operational part of a data warehouse can, however, be designed to be small, data-driven, platform-independent and almost maintenance-free by taking advantage of features of Base SAS.

The automated warehouse design described in this paper can be used unchanged in multiple warehouses running on Windows NT-compatible and/or UNIX platforms by copying the operational files into a pre-defined directory structure. The basic automation allows for linear processing of SAS code using a weekly calendar, but could be enhanced to provide branched-linear processing using monthly, or even annual, calendars.

System and software requirements

- Runs on Windows NT-compatible or UNIX platforms.
- Windows NT Task Scheduler, or UNIX 'at' scheduler, used to schedule the start of each catalog.
- There are separate versions of the basic design for SAS versions 6 and 8, due to differences in the lengths of variable names.
- Operation requires only Base SAS to be licensed.

Warehouse operational code

- Each catalog can automatically schedule the execution of the next catalog, so the basic design operates as a linear warehouse network. The list of parent/child pairings is stored in a SAS table for each weekday, so there can be different daily schedules. All schedules should begin with the BEGIN catalog, which can be used to restrict user access to the data during the warehouse processing, and end with the END catalog, which can then be used to restore user access on the successful completion of all steps.
- Operational code is stored as SAS macros in SAS catalog source entries.
- The entire processing code, and operational code and data, are generally small enough to be backed up in compressed files to a single 1.44Mb floppy diskette! It is, therefore, unnecessary to take into account the space occupied by the code when calculating the data warehouse space requirements.

Processing code adaptations

- Processing code is stored in SAS catalog source entries beginning with 'STEP', e.g. STEP100, which are processed in numerical order.
- Sequential processing of catalog steps is controlled by setting the value of a retcode macro variable at the end of each entry. A retcode value greater than zero will prevent subsequent steps from executing, allowing remedial actions to be carried out, before the catalog is re-run.
- Each catalog runs in a single SAS session, so temporary tables are retained between catalog steps. Temporary tables are not retained between different catalog runs, so tables to be passed to subsequent catalogs must be stored in permanent SAS libraries.
- Errors in DATA steps, PROC steps, LIBNAME and FILENAME statements can be trapped using SAS macro calls and used to set the retcode macro variable, e.g.:

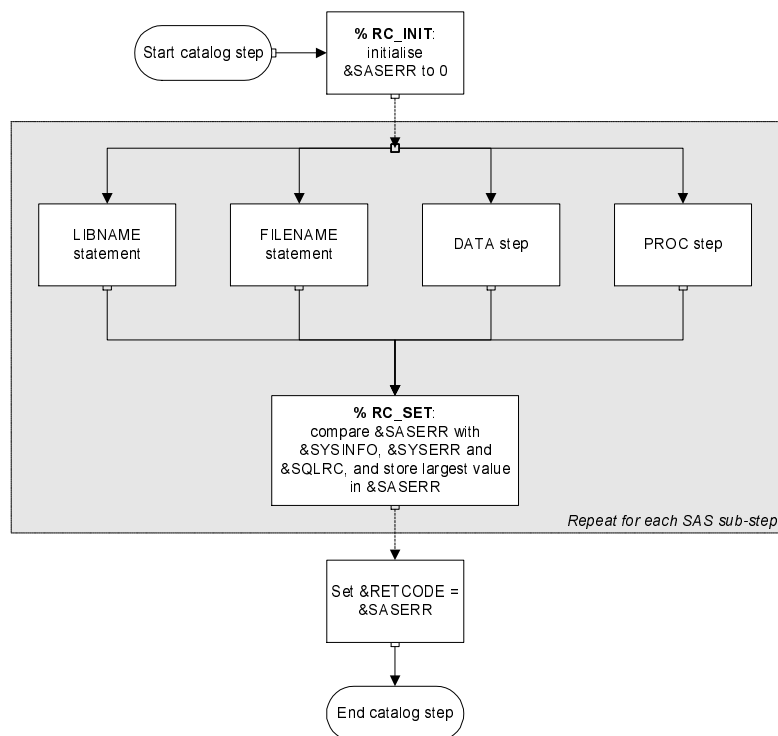
```
%rc_init
....code....;
%rc_set
....code....;
%rc_set
%let retcode=&saserr;
```

The saserr macro variable is assigned the largest error value recorded.

- Raw data from flat files can be read using metadata via a single SAS macro call.
- Metadata describes:
 - (1) external data columns with filenames, informats and input transformations,
 - (2) internal SAS variables with tables and further transformations,
 - (3) global formats, field sizes and labels for SAS variables.

Catalog Step Template

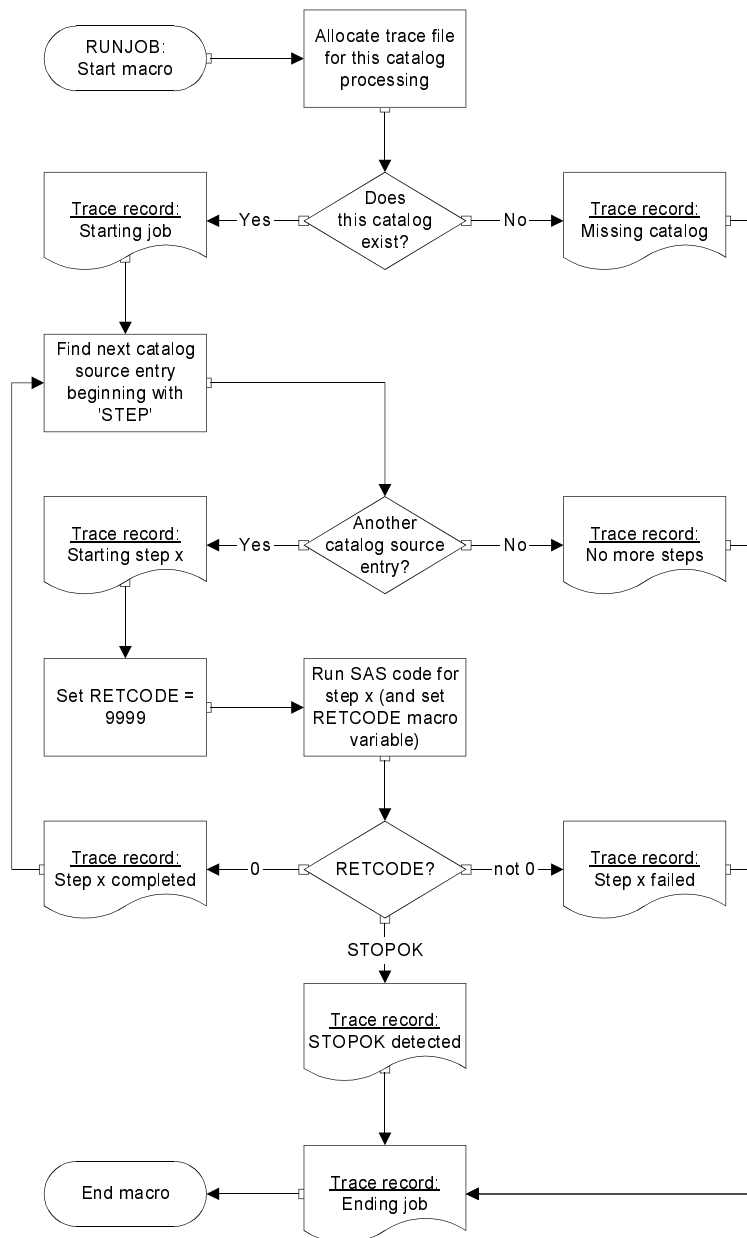
(xxx_JOBS.catalog.STEPnnn.SOURCE)



Operational monitoring

- SAS log files are stored for the most recent run of each catalog.
- Trace files have records appended to information stored during previous runs. Catalog entries generate start and end trace records, and additional records can be written from within the catalog steps via SAS macro calls.
- Empty zero-length files (*.fin) can also be created using SAS macro calls to provide external records of the time of completion of catalog steps.

RUNJOB: Processing Flow (SASTOOL.SASMACR.RUNJOB.SOURCE)

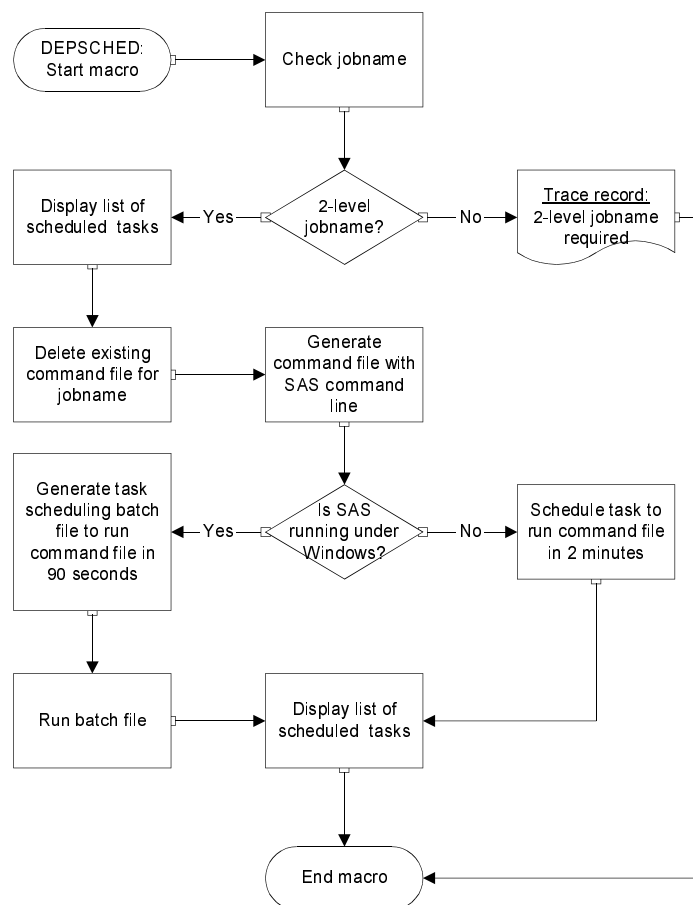


Converting existing data warehouse programs

- **Step 1:** Separate the existing code into discrete independent programs.
- **Step 2:** Split these programs into simple sub-programs.
- **Step 3:** Copy these sub-programs into individual catalog entries, with each catalog storing an independent program, naming each entry STEP nnn in processing order. Note the ' nnn ' values need not be contiguous.
- **Step 4:** Edit each catalog entry to add
 `%rc_init`
at the top,
 `%let retcode=&saserr;`
at the end, and
 `%rc_set`
after each DATA or PROC step, or LIBNAME or FILENAME statement.
- **Step 5:** If another catalog is to be scheduled automatically after this one, create a new catalog entry with the largest sequence number, e.g. STEP999, so that it will be processed after all the other steps.

DEPSCHED: Processing Flow

(SASTOOL.SASMACR.DEPSCHED.SOURCE)



How to obtain the data warehouse

- The operational programs that control the data warehouse are freeware.
- Holland Numerics Limited provides on-site or off-site consultancy for customizing, installing and developing the data warehouse, as well as staff training on its development and maintenance, on a per day, or fixed price, basis.
- Holland Numerics Limited also operates a low-cost email support service, called [sas.answers @ Holland Numerics], which will provide on-going assistance to internal operations and development staff. The same service also provides general SAS software advice and assistance through a service level agreement to up to 5 specified email addresses.
- More information about consulting and support services can be found on the company web site.

Summary

Disadvantages

- Only available for Windows NT-compatible and UNIX platforms.
- Basic design only incorporates linear scheduling.
- Basic design only uses a weekly scheduling calendar.
- Existing data warehousing code must be adapted to be run within this data warehouse design.
- Upgrading from SAS version 6 to 8 requires the installation of a new version of the operational code and metadata.

Advantages

- Minimal SAS System specifications: only Base SAS is required to be licensed.
- Operational programs are freeware.
- Platform-independent operation.
- Automatic scheduling of programs using SAS macro calls and native scheduling software forming part of Windows NT or UNIX.
- Automatic error-trapping, tracing and logging.
- Stable operational code.
- The storage space occupied by the processing code, with the operational code and metadata, is insignificant compared to that occupied by the data warehouse.
- Flat files can be read, with any data transformations, using simple SAS macro calls and metadata tables.
- Processing code is stored in SAS catalogs as small and structured SAS programs for easy maintenance.
- Individual catalog processing can be controlled and scheduled by SAS/Warehouse Administrator[®] software, in place of the internal scheduler.
- Fully supported by Holland Numerics Limited.

Contact details

- **Contact:** Philip R Holland
- **Company:** Holland Numerics Limited
- **Address:** 94 Green Drift, Royston, Herts SG8 5BT, UK
- **Tel/fax:** +44-(0)1763-244497
- **Email:** sales@hollandnumerics.com
- **Web:** www.hollandnumerics.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.