

Using SAS Indexes

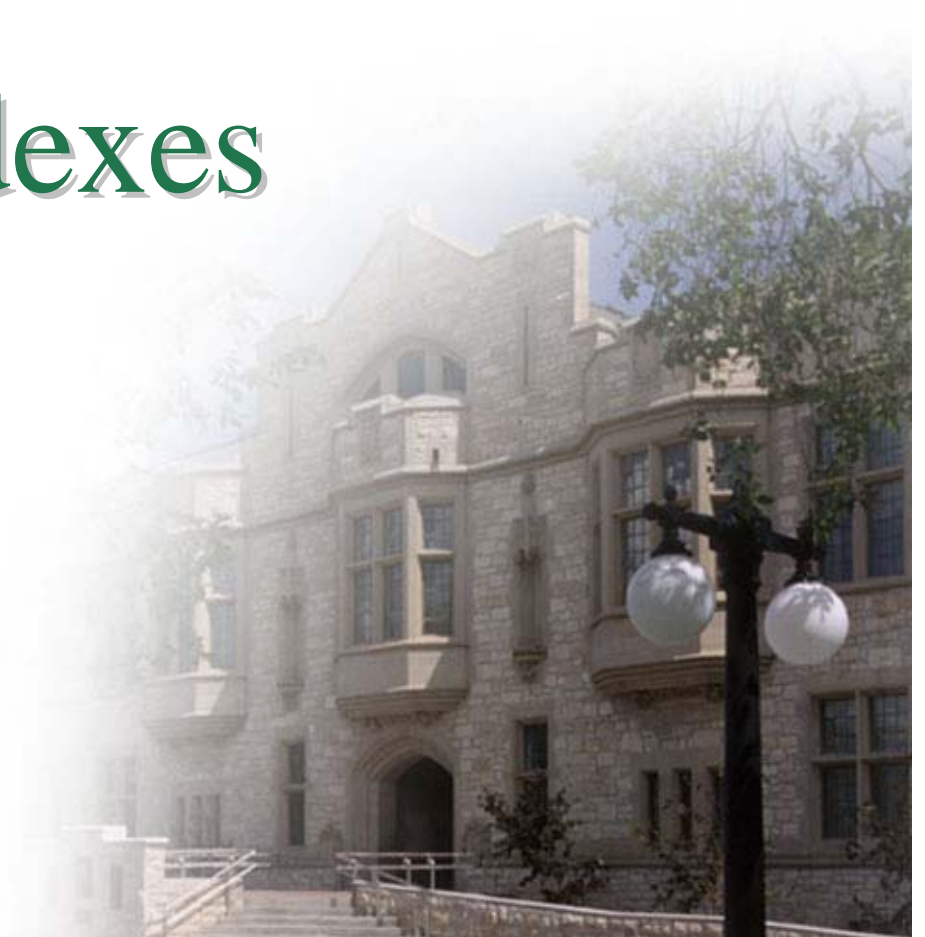
What, Why and How

Mark Lucas

University of Saskatchewan



Saskatoon, Saskatchewan, Canada.
Institutional Analysis www.usask.ca/ia



Background

- *Historically, all of the SAS reports I built processed an entire dataset and produced statistics or summary data based on most/all the rows in the dataset.*
- *Very limited use of WHERE or BY clauses as datasets were designed to meet specific needs*
- *Performance of those routines was exceptional or adequate.*

When did I run into trouble?

- *I started developing complex ETL routines using Base SAS.*
- *I reach a point where I needed to update the rows in one dataset (20,000 rows) with data stored in another dataset – 200,000 thousand rows).*
- *I needed to create summary statistics based on the larger dataset and the key to join the datasets could match 1 or N rows in the second larger dataset.*

When did I run into trouble?

Cont'd

- *Built an update query using PROC SQL making use of COUNT and SUM function.*
- *Submitted the program on a remote server. Still running 4 hours later.*
- *Ignored the problem, assumed server related, killed the program, submitted again to run over night.*
- *14 hours later query still running.*

What was the problem?

- *Number of reads required to match each row in the first dataset with 1 to N rows in the second larger dataset.*
- *Sorting the larger dataset would not actually solve the problem.*

What was the solution?

- *Learn about SAS Indexes*
- *Great resource*
 - *The Complete Guide to SAS Indexes, Michael A. Raithel, A SAS publication.*
 - *ISBN-13: 978-1-59047-849-3*
- *I created an index based on the key fields used to match records and procedure finished in less than 10 minutes.*
- *Saw similar performance gains with other SQL update transactions.*

What does an Index do?

- *Bypasses the need for sequential reads of very large datasets when looking for direct matches on a specified key field(s).*
- *Provides an address into the dataset matching the value of the key.*

Types of Indexs

→ *Simple Index*

→ *Index based on one field in the dataset – Person_id.*

→ *Composite Indexes*

→ *Index based on two or more fields in a dataset – Person_id, Account_Number*

When does SAS use an index?

- *WHERE* expressions in Data or Proc SQL
- *By* statement in Data or Proc Steps
- *Key* options on *Modify* and *SET* statement (I am not familiar with these)

Does SAS always use the index?

→ NO!

→ *Lots of things effect index usage*

→ *Like any relational database management system SAS evaluates, at run time, which is more efficient – index or sequential reads. SAS may make the wrong choice or other factors may impact the decision to use an index.*

What can break a Index

- *The book, referenced earlier, has a whole chapter on what could cause SAS not to use an index. The structure of the query, order of fields in the query, whether any of the fields in the index have missing values, if you use '=' instead of 'eq' or whether SAS thinks sequential reads are more efficient.*

How to tell if SAS is using the index?

- *A system option is used to monitor whether a program uses an index.*
 - *OPTIONS MSGLEVEL = I*
- *Warning – this option can quickly fill up a log file if the SAS routine uses statement(s) that trigger the index hundreds or thousands of times.*
- *Test using a small dataset and monitor and look for performance changes. The book recommends leaving it on all the time.*

How to define indexes

→ *Indexes can be defined within:*

→ *DATA step,*

→ *SAS procedure*

→ *DATASETS procedure*

→ *PROC SQL procedure*

→ *Basic Syntax*

→ *(INDEX = (index-name = (var1 var2)
</Unique> </NOMISS>));*

Indexes in DATA step

- *DATA work.test (index=(pid=(pid)));*
 - *This creates index at run time and applies to the newly created Dataset.*
 - *Pid is the name of the index and pid is the dataset field that the index is based on.*
- *DATA work.test (index=(person_act=(pid acct_num));*
- *You can create 1 to N indexes using this format.*

Indexes in SAS Procedures

- *An index can be created on a new dataset created by a SAS procedure if the procedure supports the 'OUTPUT' statement. For example: PROC SORT, PROC SUMMARY.*
- *Output `out=work.test (index=(pid=(pid)))`*
 - *pid is the index name and pid is the dataset field.*

Indexes in DATASET procedure

- *Can be used to add indexes to existing datasets. I find this syntax easier to follow.*
- *Proc datasets library = work;*
- *Modify test;*
 - *INDEX CREATE person =(pid);*
 - *INDEX CREATE person_acct = (pid acct_num);*
- *Run;*

Indexes in Proc SQL

- *You can use a separate PROC SQL statement or it can be built during a PROC SQL CREATE TABLE command. You can build simple or composite indexes.*
- *PROC SQL;*
 - *Create index pid on work.test (pid);*
- *Quit;*

Indexes in PROC SQL con't

→ PROC SQL;

→ Create index *person_acct* on *work.test* (*pid*
acct_num);

→ Quit;

→ Note: The *NOMIS INDEX* options is not
available in PROC SQL

Indexes can disappear

→ *You can accidentally make an index disappear through some very common types of manipulation.*

→ *If dataset is written over*

- Data work.test;
 - Set work.test;
 - Run;



Indexes can disappear con't

- *When creating a new dataset based on an existing dataset.*
 - *Data work.test1;*
 - Set work.test;
 - Run;
- *Sort an indexed dataset with an output to a new dataset.*
- *Sort onto itself using the FORCE option (not familiar with this option)*

Advanced Topics

→ *Indexes options available at create time:*

→ *UNIQUE*

- Only if key value is UNIQUE. If not specified DUPLICATE key values will be allowed in the INDEX.

→ *NOMISS*

→ *Indexes owned by IC (integrity constraints- which are a set of data validation rules. Another SUCCESS topic?)*

Advance Topics con't

- *Overriding SAS's use of indexes:*
 - *IDXNAME option in a SAS procedure.*
 - *IDXWHERE option in a SAS procedure*
- *Recovering and Repairing Indexes*
- *Designing indexes and controlling how /when SAS updates index file.*

Summary

- *Depending on the performance issue you are dealing with an index may, with minimal effort, provide huge RIO.*
- *But, if the index does not immediately improve performance you may have to delve into the advance topics to figure out why the index is not being used and may require experimentation with query design and index design.*

Contact info

- *Mark Lucas*
- *Senior Research Analyst*
- *Institutional Analysis*
- *University of Saskatchewan*
- *Phone – 306.966.7817*
- *E-mail – mark.lucas@usask.ca*