

Using Table Lookup Techniques Efficiently

Linda Jolley
Jane Stroupe
SAS Institute Inc.

Performing Table Lookups

Values stored in a data set may need to be compared to values in another data set to expand coded values for readability or to provide more information about observations.

Detail Table

Date	AreaCode	PhoneNum	ToAcronym	FromAcronym	SecretCode
21DEC2006	407	312-9088	AFAIK	DQMOT	103
21DEC2006	407	324-6674	BEG	TU	101
21DEC2006	407	324-6674	BFN	SYS	101
21DEC2006	407	312-5098	BTDT	IHU	102
22DEC2006	407	312-9088	C&G	AFAIK	103
23DEC2006	714	324-3452	GA	DQ	

Lookup Table

Acronym	Meaning
AFAIK	as far as I know
AFK	away from keyboard
ASAP	as soon as possible
B4N	bye for now
BBL	be back later
BBS	be back soon
BEG	big evil grin
.	.
.	.
.	.

Lookup Table

AreaCode	Location
407	Orlando, FL
714	Anaheim, CA

Lookup Table

SecretCode	Name	PhoneNum
101	Belle	324-6674
102	Snow White	312-5098
103	Cinderella	312-9088
104	Jasmine	324-3452
105	Ariel	312-7483
106	Aurora	324-8943
107	Mulan	324-4544
108	Pocahontas	312-4755
109	Mickey Mouse	312-7456
110	Minnie Mouse	312-5551

Table Lookups

Lookup values for a table lookup can be stored in the following ways in SAS:

- code
- array
- hash object
- format
- data set

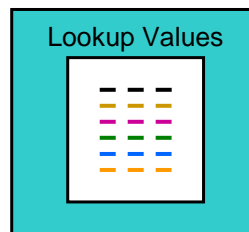
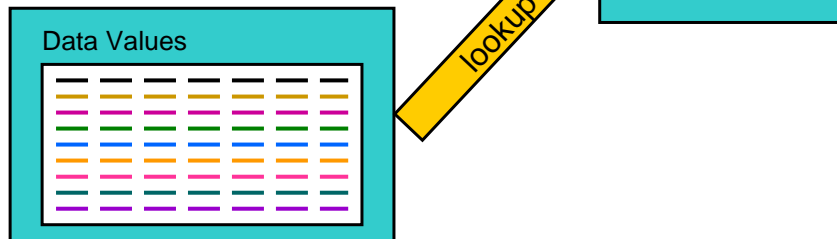


Table Lookups

Lookup techniques include the following:

- IF/THEN or SELECT/WHEN statements
- array index value
- hash object key value
- FORMAT statement, PUT function
- merge, join, KEY= option



4

...

Using IF/THEN Statements or SELECT/WHEN

```
data location;
  set mylist2;
  if AreaCode='407' then Location='Orlando, FL';
  else if AreaCode='714' then Location='Anaheim, CA';
  else Location='Unknown';
run;
```

```
data location;
  set mylist2;
  select (AreaCode);
  when('407') Location='Orlando, FL';
  when ('714') Location='Anaheim, CA';
  otherwise Location='Unknown';
end;
run;
```

5

Using IF/THEN Statements or SELECT/WHEN

Date	AreaCode	PhoneNum	ToAcronym	FromAcronym	Location
21DEC2006	407	312-5098	AFAIK	DQMOT	Orlando, FL
21DEC2006	407	324-6674	BEG	TU	Orlando, FL
21DEC2006	407	324-6674	BFN	SYS	Orlando, FL
21DEC2006	407	312-5098	BTDT	IHU	Orlando, FL
22DEC2006	407	312-5098	C&G	AFAIK	Orlando, FL
23DEC2006	714	389-0098	GA	DQMOT	Anaheim, CA
22DEC2006	714	389-0098	GMTA	YUP	Anaheim, CA
22DEC2006	714	324-6674	HTH	IHU	Anaheim, CA
23DEC2006	714	389-0098	IOW	YBS	Anaheim, CA
21DEC2006	407	312-5098	JTLYK	OIC	Orlando, FL
23DEC2006	714	389-0098	JTLYK	TAFN	Anaheim, CA
22DEC2006	407	312-5098	MLO	TPTB	Orlando, FL
23DEC2006	714	389-0098	NP	IMS	Anaheim, CA
21DEC2006	407	324-6674	OIC	IAC	Orlando, FL
21DEC2006	407	324-6674	OTOH	LOL	Orlando, FL
21DEC2006	407	312-5098	PM	GMTA	Orlando, FL
23DEC2006	714	389-0098	ROTFL	BBL	Anaheim, CA
23DEC2006	714	389-0098	WU?	WFM	Anaheim, CA

6

Guidelines for Writing Efficient IF/THEN Logic

- Use IF-THEN/ELSE statements when the following circumstances exist:
 - There are few conditions to check.
 - The data values are not uniformly distributed.
 - The values are character or discrete numeric data.
 - There are bounded ranges of data (for example, $1 < x < 2$).
- For mutually exclusive conditions, use the ELSE-IF statement rather than an IF statement for all conditions except the first.
- Check the most frequently occurring condition first.
- When you execute multiple statements based on a condition, put the statements into a DO group.

7

Guideline for Using a SELECT Statement

Use a SELECT statement when you have a long series of mutually exclusive conditions.

The IF/THEN and SELECT statements typically use approximately the same computer resources.

8

Using an Array

```
data secret;  
  keep SecretPal AreaCode PhoneNum;  
  if _n_=1 then do i=1 to numobs;  
    set PhoneBook nobs=numobs;  
    array names{101:110} $ 22 _temporary_ ;  
    names{SecretCode}=Name;  
  end;  
  set mylist;  
  SecretPal=names{Code};  
run;
```

9

Using an Array

SecretPal	AreaCode	PhoneNum
Belle	407	324-6674
Snow White	407	312-5098
Cinderella	407	312-9088
Belle	407	324-6674
Snow White	407	312-5098
Snow White	407	312-5098
Belle	407	324-6674
Belle	407	324-6674
Cinderella	407	312-9088
Cinderella	407	312-9088
Belle	714	324-6674
Jasmine	714	324-3452
Jasmine	714	324-3452
Mulan	714	324-4544
Mulan	714	324-4544
Mickey Mouse	714	312-7456
Mickey Mouse	714	312-7456

10

Advantages of an Array

Advantages of using an array include the following:

- use of positional order
- can use multiple values to determine the array element to be returned
- ability to use a non-sorted and non-indexed base data set
- use of numeric mathematical expressions to determine which element of the array to be looked up; exact match not required

If you can use an array for the table lookup, it is the most efficient technique.

11

Disadvantages of an Array

Disadvantages of using an array include the following:

- memory requirements to load the entire array
- requirement that you must have a numeric value as pointer to the array elements
- return of only a single value from the lookup operation
- dimensions supplied at compile time by either hard-coding or macro variables

12

Using a DATA Step HASH Object

```
data Messages;
  keep ToAcronym FromAcronym ToMeaning
      FromMeaning PhoneNum;
  length Acronym $5 Meaning $50;
  if _n_ = 1 then do;
    declare hash wh(dataset:'Acronyms');
    wh.definekey('Acronym');
    wh.definedata('Meaning');
    wh.definedone();
  end;
  set mylist2;
  rc = wh.find(key:ToAcronym);
  if rc = 0 then ToMeaning = Meaning;
  else ToMeaning = 'Unknown';
  rc = wh.find(key:FromAcronym);
  if rc = 0 then FromMeaning = Meaning;
  else FromMeaning = 'Unknown';
run;
```

13

Using a DATA Step HASH Object

PhoneNum	ToAcronym	FromAcronym	ToMeaning	FromMeaning
324-6674	OTOH	LOL	on the other hand	laughing out loud
312-5098	JTLYK	OIC	just to let you know	oh, I see
312-9088	AFAIK	DQMOT	as far as I know	don't quote me on this
324-6674	OIC	IAC	oh, I see	in any case
312-5098	PM	GMTA	private message	great minds think alike
312-5098	BTDT	IHU	been there done that	I hear you
324-6674	BEG	TU	big evil grin	thank you
324-6674	BFN	SYS	bye for now	see you soon
312-9088	C&G	AFAIK	Unknown	as far as I know
312-9088	MLO	TPTB	Unknown	the powers that be
389-0098	GMTA	YUP	great minds think alike	Unknown
324-6674	HTH	IHU	hope this helps	I hear you

14

Advantages of Hash Objects

Advantages of using hash objects include the following:

- use of character and numeric keys
- use of composite keys
- ability for faster lookup
- ability to be loaded from a SAS data set
- fine level of control (flexibility)
- ability to do chained lookups

If your key is character or composite, then using the HASH object should be your choice.

15

Disadvantages of Hash Objects

Disadvantages of using a hash object include the following:

- unique keys required
- DATA step only

16

Using a Format

```
data fmt;
  set acronyms(rename =(acronym = start meaning = label));
  retain fmtname '$Acronym';
run;

proc format cntlin=fmt library=sasuser;
run;

proc format library=sasuser;
  value $Extra 'C&G','MLO','YUP' = 'Unknown'
              other = [$acronym.];
run;

options fmtsearch =(sasuser);

data combine;
  keep ToAcronym FromAcronym ToMeaning FromMeaning
      PhoneNum;
  set mylist;
  ToMeaning = put(ToAcronym, $Extra.);
  FromMeaning = put(FromAcronym, $Extra.);
run;
```

Using a Format

PhoneNum	ToAcronym	FromAcronym	ToMeaning	FromMeaning
324-6674	OTOH	LOL	on the other hand	laughing out loud
312-5098	JTLYK	OIC	just to let you know	oh, I see
312-9088	AFAIK	DQMOT	as far as I know	don't quote me on this
324-6674	OIC	IAC	oh, I see	in any case
312-5098	PM	GMTA	private message	great minds think alike
312-5098	BTD	IHU	been there done that	I hear you
324-6674	BEG	TU	big evil grin	thank you
324-6674	BFN	SYS	bye for now	see you soon
312-9088	C&G	AFAIK	C&G	as far as I know
312-9088	MLO	TPTB	MLO	the powers that be
324-6674	HTH	IHU	hope this helps	I hear you
324-3452	GA	DQMOT	go ahead	don't quote me on this
324-3452	IOW	YBS	in other words	you'll be sorry
324-4544	NP	IMS	no problem	I am sorry
324-4544	WU?	WFM	what's up?	works for me
312-7456	JTLYK	TAFN	just to let you know	that's all for now
312-7456	ROTFL	BBL	rolling on the floor laughing	be back later

18

Advantages of Formats

Advantages of using formats include the following:

- familiarity
- no need to create additional data
- can be used with procedures
- range search for both character and numeric
- binary search through lookup table
- centralize maintenance
- use of multiple PUT functions to create multiple variables

A format can be stored permanently, as in the example, and that makes using the lookup table much simpler, since the code to create the lookup table only needs to be run one time.

19

Disadvantages of Formats

Disadvantages of using formats include the following:

- memory requirements to load the entire format for the binary search
- use of only one variable for the table lookup
- requires more disk space to store a format than to store the equivalent SAS data

20

Merging

```
proc sort data = acronyms;
  by Acronym;
run;

proc sort data = mylist;
  by ToAcronym;
run;

data ToMsg;
  keep ToAcronym FromAcronym ToMeaning PhoneNum;
  merge acronyms(in = a rename =(Acronym = ToAcronym))
        mylist(in = m);
  by ToAcronym;
  if m;
  if a then ToMeaning = Meaning;
  else ToMeaning = 'Unknown';
run;
```

21

Continued...

Merging

```
proc sort data = ToMsg;
  by FromAcronym;
run;

data BothMsg;
  keep ToAcronym FromAcronym ToMeaning FromMeaning
      PhoneNum;
  merge acronyms(in=a rename =(Acronym = FromAcronym))
        ToMsg(in = t);
  by FromAcronym;
  if t;
  if a then FromMeaning = Meaning;
  else FromMeaning = 'Unknown';
run;
```

22

Merging

FromAcronym	ToAcronym	PhoneNum	ToMeaning	FromMeaning
AFAIK	C&G	312-9088	Unknown	as far as I know
BBL	ROTFL	312-7456	rolling on the floor laughing	be back later
DQMOT	AFAIK	312-9088	as far as I know	don't quote me on this
DQMOT	GA	324-3452	go ahead	don't quote me on this
GMTA	PM	312-5098	private message	great minds think alike
IAC	OIC	324-6674	oh, I see	in any case
IHU	BTDT	312-5098	been there done that	I hear you
IHU	HTH	324-6674	hope this helps	I hear you
IMS	NP	324-4544	no problem	I am sorry
LOL	OTOH	324-6674	on the other hand	laughing out loud
OIC	JTLYK	312-5098	just to let you know	oh, I see
SYS	BFN	324-6674	bye for now	see you soon
TAFN	JTLYK	312-7456	just to let you know	that's all for now
TPTB	MLO	312-9088	Unknown	the powers that be
TU	BEG	324-6674	big evil grin	thank you
WFM	WU?	324-4544	what's up?	works for me
YBS	IOW	324-3452	in other words	you'll be sorry

23

Advantages of DATA Step MERGE

- Multiple values can be returned.
- There is no limit to the size of the table, other than disk space.
- Multiple BY variables enable lookups that depend on more than one variable.
- Multiple data sets can be used to provide access to different tables.
- A merge enables complex business logic to be incorporated into the new data set by using DATA step processing, such as arrays and DO loops, in addition to merging features.

continued...

24

Advantages of DATA Step MERGE

- The IN= data set option and subsequent IF-THEN/ELSE logic afford comprehensive control over whether to accept, reject, or process differently depending on which data set contributed each observation.
- Observations with duplicate BY values are joined one-to-one instead of being expanded into a Cartesian product, as SQL does.

25

Disadvantages of DATA Step MERGE

- Data sets must be sorted by or indexed based on the BY variable(s).
- An exact match on the key value(s) must be found.
- The BY variable(s) must be present in all data sets.
- When more than one data set contributes variables with the same name, the values from the variable in the rightmost data set overwrite the other like-named variables, and no warning is printed.

26

Using the SQL Procedure

```
proc sql;  
  create table SQLBothMsg as  
    select ToAcronym,  
           FromAcronym,  
           a1.Meaning as ToMeaning,  
           a2.Meaning as FromMeaning,  
           PhoneNum  
    from Mylist,  
         Acronyms as a1,  
         Acronyms as a2  
    where a1.Acronym = ToAcronym and  
          a2.Acronym = FromAcronym;  
quit;
```

27

Using the SQL Procedure

ToAcronym	FromAcronym	ToMeaning	FromMeaning	PhoneNum
ROTFL	BBL	rolling on the floor laughing	be back later	312-7456
AFAIK	DQMOT	as far as I know	don't quote me on this	312-9088
GA	DQMOT	go ahead	don't quote me on this	324-3452
PM	GMTA	private message	great minds think alike	312-5098
OIC	IAC	oh, I see	in any case	324-6674
HTH	IHU	hope this helps	I hear you	324-6674
BTDT	IHU	been there done that	I hear you	312-5098
NP	IMS	no problem	I am sorry	324-4544
OTOH	LOL	on the other hand	laughing out loud	324-6674
JTLYK	OIC	just to let you know	oh, I see	312-5098
BFN	SYS	bye for now	see you soon	324-6674
JTLYK	TAFN	just to let you know	that's all for now	312-7456
BEG	TU	big evil grin	thank you	324-6674
WU?	WFM	what's up?	works for me	324-4544
IOW	YBS	in other words	you'll be sorry	324-3452

28

Advantages of PROC SQL Joins

- Multiple data sets can be joined without having common variables in all data sets.
- Data sets do not have to be sorted or indexed.
- Inequality joins can be performed.
- You can create data files (tables), views, or reports.
- PROC SQL follows ANSI standard language definitions, so that you can use knowledge gained from other languages.
- Duplicate BY values are combined into a Cartesian product.

29

Disadvantages of PROC SQL Joins

- The maximum number of tables that can be joined at one time is 32. (256 in SAS 9.1.3 SP4)
- PROC SQL might require more resources than the DATA step with the MERGE statement for simple joins.
- Complex business logic is difficult to incorporate into the join.
- Duplicate BY values are combined into a Cartesian product, which can produce an extremely large output data set.

30

Efficiency: DATA Step MERGE or PROC SQL?

That depends on your data:

- relationship between the tables
- sparseness or denseness of matches
- size of the tables
- availability of an index or sort flag

31

Efficiency: DATA Step MERGE or PROC SQL?

- When data sets are large and unsorted, the SQL inner join might out perform SORT and MERGE.
- If you have a long series of SORT and DATA steps, the SQL inner join might be easier to code and read.
- In most cases, a DATA step MERGE statement generally out performs an SQL outer join, even taking sort resources into account.
- One exception is a very sparse match join when you only want the observations with matching key values.

32

continued...

Efficiency: DATA Step MERGE or PROC SQL?

- Keep in mind that the SQL procedure and the DATA step MERGE do not provide the same results if you have a many-to-many match.
- Since there are no hard and fast rules about the efficiency of MERGE vs. SQL, you'll just have to benchmark them to find out how they perform with your data. Remember that benchmarking involves multiple submissions of the two sets of code against the data, each submission running in a separate SAS session.

33

Using SET/SET with KEY=

```

proc datasets library = work;
  modify Acronyms;
  index create Acronym/unique;
run;

data KEYMsg1(rename =(Acronym = ToAcronym));
  keep Acronym FromAcronym ToMeaning PhoneNum;
  set mylist(rename =(ToAcronym = Acronym));
  set Acronyms key = Acronym;
  if _iorc_ = 0 then ToMeaning = Meaning;
  else ToMeaning = 'Unknown';
run;

data KEYMsgs(rename =(Acronym = FromAcronym));
  keep ToAcronym Acronym ToMeaning FromMeaning PhoneNum;
  set KEYMsg1(rename =(FromAcronym = Acronym));
  set Acronyms key = Acronym;
  if _iorc_ = 0 then FromMeaning = Meaning;
  else FromMeaning = 'Unknown';
run;

```

34

Using SET/SET with KEY=

PhoneNum	ToAcronym	FromAcronym	ToMeaning	FromMeaning
324-6674	OTOH	LOL	on the other hand	laughing out loud
312-5098	JTLYK	OIC	just to let you know	oh, I see
312-9088	AFAIK	DQMOT	as far as I know	don't quote me on this
324-6674	OIC	IAC	oh, I see	in any case
312-5098	PM	GMTA	private message	great minds think alike
312-5098	BTDT	IHU	been there done that	I hear you
324-6674	BEG	TU	big evil grin	thank you
324-6674	BFN	SYS	bye for now	see you soon
312-9088	C&G	AFAIK	Unknown	as far as I know
312-9088	MLO	TPTB	Unknown	the powers that be

35

Advantages of SET/SET with the KEY= Option

- Only the necessary observations are read.
- An existing index is used.
- Multiple values can be returned.
- Availability of DATA step syntax provides the full power of the DATA step.
- Exact matches are returned.
- `_IORC_` can be used to control non-matching data.
- Only one page from the master data set is stored in memory, thus conserving memory.

36

Disadvantages of SET/SET with the KEY= Option

- An index on one data set is required.
- Creating and maintaining an index uses resources.
- Useful only for data with exact matches.

When the indexed data set is not sorted by the key variable(s), there can be considerable increase in I/O due to the random access of the master data set and the I/O required to access the index.

37

Comparison of Techniques

In Memory Techniques			
	Array	Hash Object	Format
Where used?	DATA step only	DATA step only	DATA steps and procedures
Permanently Stored?	No; you have to repeat the code every time you use the lookup table.	No; you have to repeat the code every time you use the lookup table.	Can be stored permanently so the code to create the lookup table only has to be run one time.
When loaded into memory?	When array is created.	When hash object is created.	When the format is used.
Can load from a SAS data set?	Yes.	Yes.	Yes.

38

Comparison of Techniques

Other Techniques				
	IF/ THEN	MERGE	SQL	SET/SET & KEY=
Processing Method	sequential	sequential	Cartesian product	direct access using index
Memory Used	data set page(s)	data set page(s)	utility work space to form Cartesian product	data set page(s) including all page(s) for index

39

Questions?

Linda Jolley

linda.jolley@sas.com

913-491-1166



Jane Stroupe

jane.stroupe@sas.com

847-367-7216