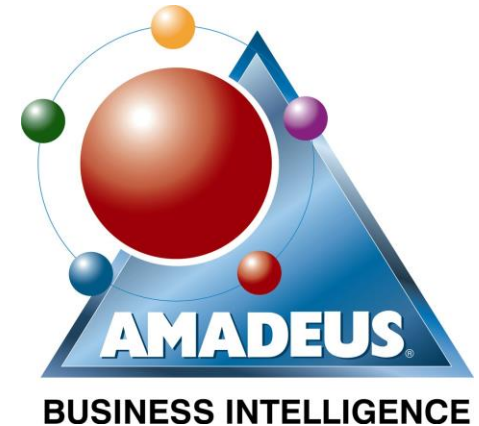




Which Merge?

Stephen Hanks
Amadeus Software Limited
20th May 2003



Introduction

- **SAS Merge**
- **SQL Merge**
- **Using Formats to Join Tables**
- **Indexed Key Merge**
- **Results**
- **Comparison**
- **SPD Server and SAS/ACCESS**

Example

- 2 SAS Data Sets Called A and B
- Add Variables in B to Dataset A

VIEWTABLE: Work.A		
	i	x
1	1	0.54028237
2	2	0.217615782
3	3	0.497206389
4	4	0.169914754
5	5	0.809005148
6		
7		
8		
9		

VIEWTABLE: Work.B		
	i	y
1	1	0.3797289908
2	2	0.7498091505
3	3	0.2843680597
4	5	0.7574904281
5	6	0.1879373589
		0.3894713402
		0.812994189
		0.2720566021
		0.1412464139

VIEWTABLE: Work.All				
	i	x	y	
1	1	0.5402823698	0.3797289908	
2	2	0.2176157819	0.7498091505	
3	3	0.4972063887	0.2843680597	
4	4	0.1699147537		.
5	5	0.8090051477	0.7574904281	

SAS Merge



Editor - Untitled2 *

```
- proc sort data=a;  
    by i;  
run;
```

```
- proc sort data=b;  
    by i;  
run;
```

```
- data all;  
    merge a(in=a) b(in=b);  
    by i;  
    if a;  
run;
```

SAS Merge

Advantages

- All combinations of; in both data sets, in either, or in just one data set
- Simple to code
- Pointers are much more efficient than SAS SQL

Disadvantages

- Can't do a many to many merge
- Data sets need to be sorted before merging

SQL Merge



Editor - Untitled2 *

```
- proc sql;  
  | create table all as  
  | select a.i, a.x, b.y  
  | from a left join b  
  | on a.i=b.i  
  | order i;  
quit;
```

SQL Merge

Advantages

- All combinations of; in both data sets, in either, or in just one data set
- Can do a many to many merge
- Simple to code particularly if already familiar with SQL
- No need to sort data set before merging
- Can easily join on variables with different names

Disadvantages

- Can be inefficient with large data sets

Formats To Join Files

Editor - Untitled2 *

Editor - Untitled2 *

```
- data c;  
    set a; (format=fmtlib1) ) end=eof;  
- proc format cntlin=c;  
- data all;  
    set a;  
    x=put(i,bfmt.);  
    if x=9999 then delete;  
run;
```

```
    if x=9999 then delete;  
run;
```

Formats To Join Files

Advantages

- No need to sort data set before merging
- Can be efficient particularly if adding details from a small data set to a large data set

Disadvantages

- Only observations common to both data sets, or alternatively in one data set can be joined
- Only use a single key, and single results (can be worked around by concatenating fields)
- Must not have duplicates in the data set used to create the format.



Indexed Key Merge



Editor - Untitled2 *

```
proc datasets lib=work;
    modify b;
    index create i/unique;
quit;

data all;
    set a;
    set b key=i/unique;
    array setmiss (*) y;
    if _iorc_ ne 0 then do;
        _error_=0;
        do i=1 to dim(setmiss);
            setmiss(i)=.;
        end;
    end;
end;

run;
```

Indexed Key Merge

VIEWTABLE: Work.A		
	i	x
1	1	1
2	2	12

VIEWTABLE: Work.B		
	i	y
1	1	11
2	3	15

X	I	Y (Setmiss {1})	J	_error_	_iorc_
12	2	.	1	0	1230015

Editor - Untitled2 *

VIEWTABLE: Work.All			
	i	x	y
1	1	1	11
2	2	12	.

do;

```
do j=1 to dim(setmiss);
    setmiss(i)=.;
end;
end;
```

Indexed Key Merge

Advantages

- No need to sort data set before merging (need to index one data set)
- Can be very efficient

Disadvantages

- Only observations common to both data sets, or alternatively in one data set can be joined
- Need to index one of the data sets
- Needs good understanding of the SAS LPDV

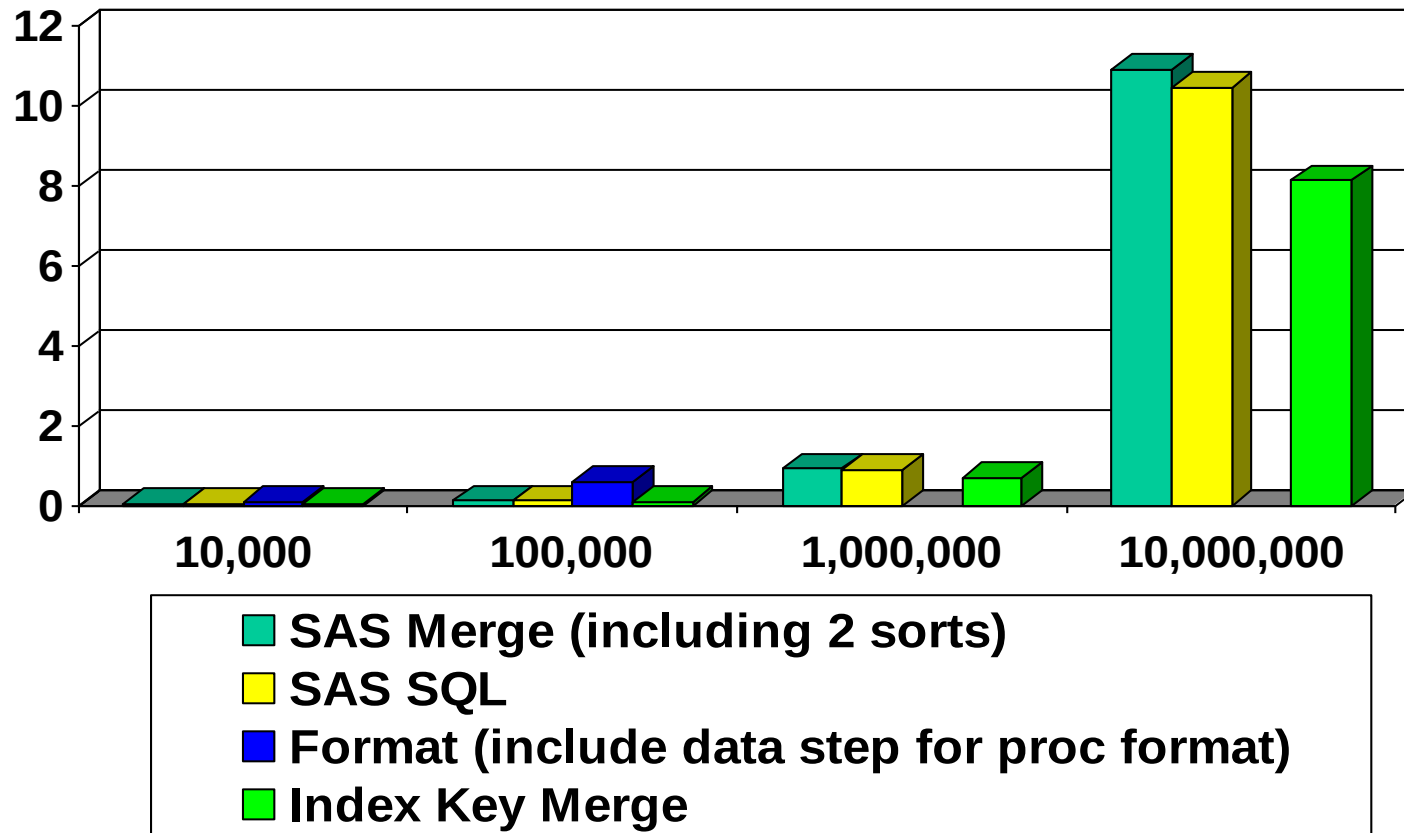
Results

Total Time

	Size A	10,000	100,000	1,000,000	10,000,000
SAS Merge (including 2 sorts)	10,000	0.03	0.12	0.93	10.90
SAS SQL		0.05	0.11	0.90	10.45
Format (include data step for proc format)		0.09	0.60	N/A	N/A
Index Key Merge		0.03	0.09	0.70	8.13
SAS Merge (including 2 sorts)	100,000	0.15	0.24	1.04	11.06
SAS SQL		0.19	0.37	1.19	11.60
Format (include data step for proc format)		0.15	0.71	N/A	N/A
Index Key Merge		0.08	0.15	0.76	8.55
SAS Merge (including 2 sorts)	1,000,000	1.25	1.37	2.16	12.50
SAS SQL		1.50	1.69	3.88	16.46
Format (include data step for proc format)		0.61	1.13	N/A	N/A
Index Key Merge		0.54	0.57	1.36	8.74
SAS Merge (including 2 sorts)	10,000,000	13.90	14.06	14.85	25.63
SAS SQL		16.28	17.61	20.18	51.87
Format (include data step for proc format)		6.06	6.41	N/A	N/A
Index Key Merge		5.91	5.04	5.84	17.53

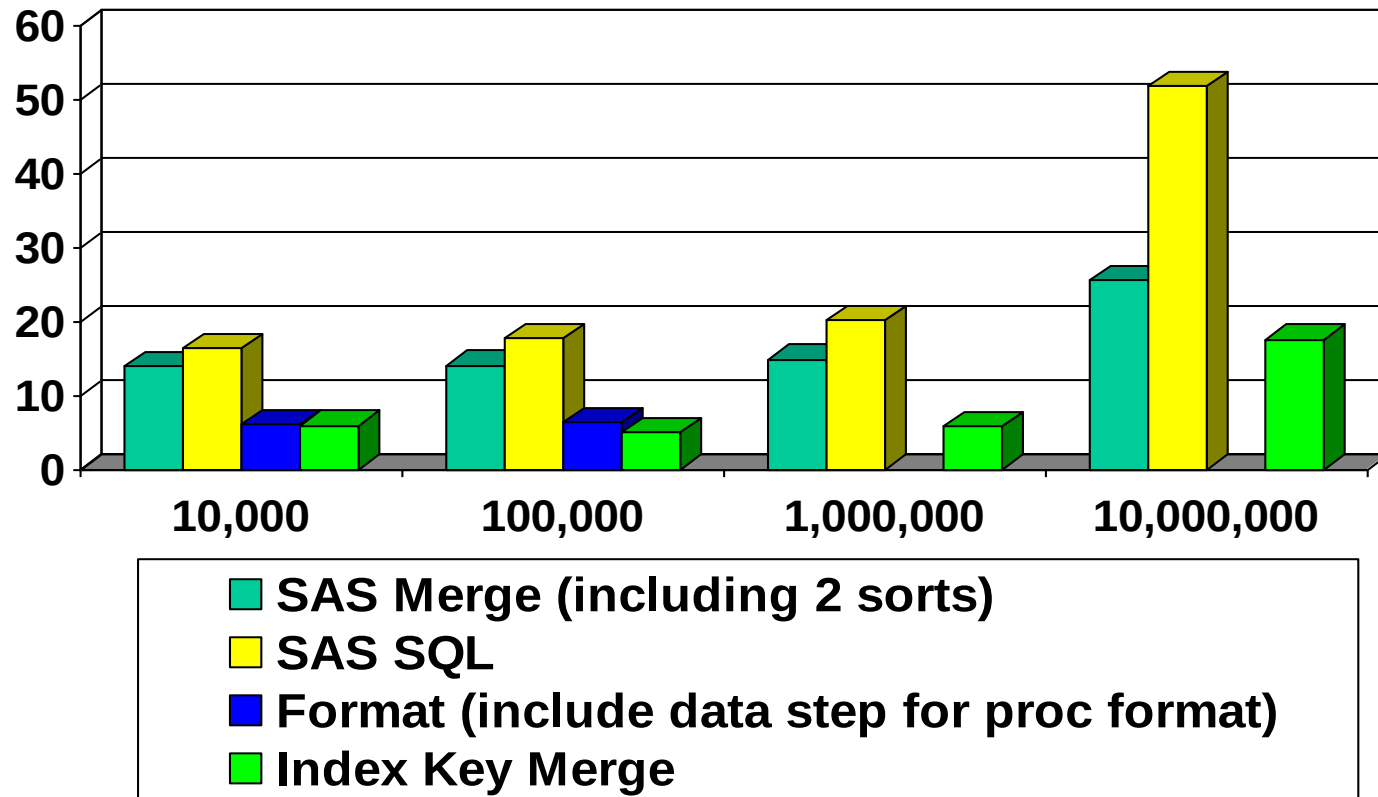
Results

Size of A 10,000 obs



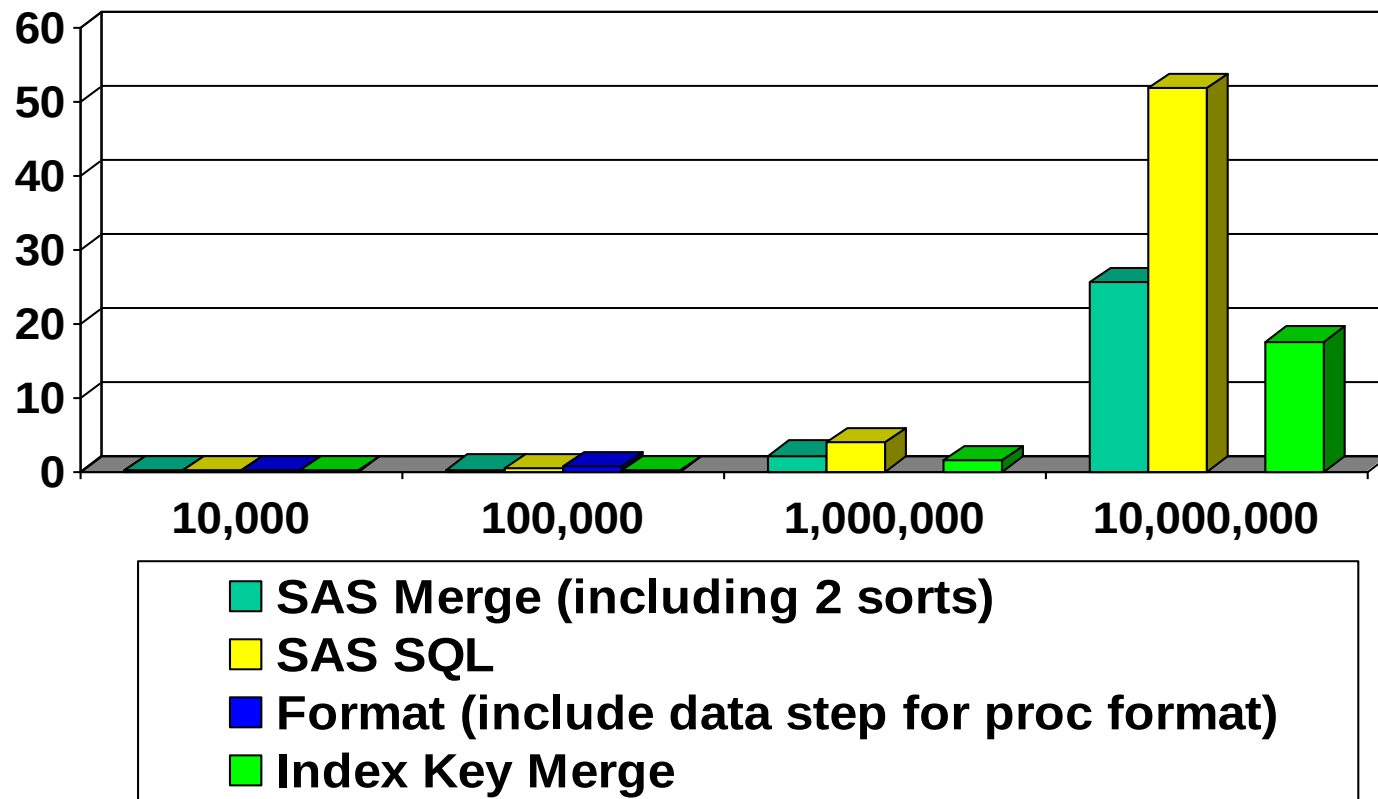
Results

Size of A 10,000,000 obs



Results

Size A = Size B



Results

Total Time

	Size A	10,000	100,000	1,000,000	10,000,000
SAS Merge (including 2 sorts)	10,000	0.03	0.12	0.93	10.90
SAS SQL		0.05	0.11	0.90	10.45
Format (include data step for proc format)		0.09	0.60	N/A	N/A
Index Key Merge		0.03	0.09	0.70	8.13
SAS Merge (including 2 sorts)	100,000	0.15	0.24	1.04	11.06
SAS SQL		0.19	0.37	1.19	11.60
Format (include data step for proc format)		0.15	0.71	N/A	N/A
Index Key Merge		0.08	0.15	0.76	8.55
SAS Merge (including 2 sorts)	1,000,000	1.25	1.37	2.16	12.50
SAS SQL		1.50	1.69	3.88	16.46
Format (include data step for proc format)		0.61	1.13	N/A	N/A
Index Key Merge		0.54	0.57	1.36	8.74
SAS Merge (including 2 sorts)	10,000,000	13.90	14.06	14.85	25.63
SAS SQL		16.28	17.61	20.18	51.87
Format (include data step for proc format)		6.06	6.41	N/A	N/A
Index Key Merge		5.91	5.04	5.84	17.53

Comparison

SAS Merge

In either data set with
both large data set

Both data sets of similar
size

Simplicity > Efficiency

Formats

Adding Small to Large

Excluding

SQL Merge

Many-to-Many

Small data sets

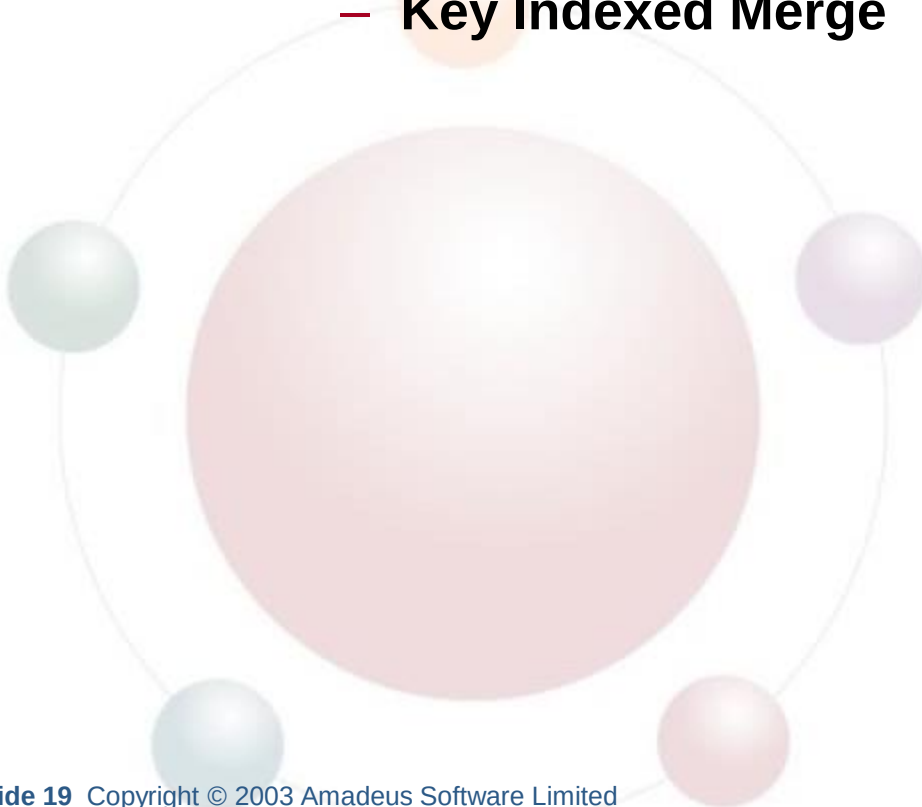
Index Merge

Efficiency is only
consideration

Vary Large files

SPD Server

- **No need to sort data sets**
 - SAS Merge
 - Key Indexed Merge



SAS/ACCESS

- Machine Speed or Priorities
- Network Speed
- Either join in database then move to SAS, or
- Move to SAS then join in SAS
- Know your own set up

Conclusion

- **Think About the Options**
- **No One Best Solution**

