Query på store datasæt –
kan jeg nå at hente kaffe eller male en carport inden den er færdig ?

Fans Netværksmøde November , programmering

Henrik Dorf , SAS Institute

Hvad er store Forespørgsler og hvordan kan de optimeres

- Store forespørgsler er queries på datamængder der enten har en størrelse eller en kompleksitet der medfører meget lange svartider.
- Normalt vil jeg anbefale at lave det simple program der fungerer – derefter måle og så optimere senere
- Her kommer en række tips til at gennemskue processerne

Først – at kende processen

- Information giver visdom…..
- Manglende information giver en mail….

```
option nonotes nosource ;

ERROR: BY variables are not properly sorted on data set DATA.CLASS_BIG.

OK=1 AAR=1984 AGE=14 Name=Alfred Sex=M Height=69 Weight=112.5 i=1000000 P=100000001
FIRST.AGE=0 LAST.AGE=0 _ERROR_=1

_N_=1000004

WARNING: The data set WORK.TESTJOIN_DS may be incomplete.  When this step was stopped there
were 0 observations and 8   variables.
```

Loggen er informativ – med de rette options

```
option notes nosource ;
NOTE: Libref DATA was successfully assigned as follows:
      Engine:        V9
      Physical Name: C:\Users\sdkhdo\AppData\Local\Temp\SAS Temporary
Files\_TD32936_kohdow102_

ERROR: BY variables are not properly sorted on data set DATA.CLASS_BIG.
OK=1 AAR=1984 AGE=14 Name=Alfred Sex=M Height=69 Weight=112.5 i=1000000 P=100000001
FIRST.AGE=0 LAST.AGE=0 _ERROR_=1
_N_=1000004
NOTE: The SAS System stopped processing this step because of errors.
NOTE: There were 6 observations read from the data set DATA.HIST01.
NOTE: There were 1000001 observations read from the data set DATA.CLASS_BIG.
WARNING: The data set WORK.TESTJOIN_DS may be incomplete.  When this step was stopped there
were 0 observations and 8
         variables.
NOTE: DATA statement used (Total process time):
      real time           0.10 seconds
      cpu time            0.07 seconds
```

```
option notes source ;
123  LIBNAME DATA "C:\Users\sdkhdo\AppData\Local\Temp\SAS Temporary Files\_TD32936_kohdow102_\" ;
NOTE: Libref DATA was successfully assigned as follows:
      Engine:        V9
      Physical Name: C:\Users\sdkhdo\AppData\Local\Temp\SAS Temporary Files\_TD32936_kohdow102_
124  %macro MERGE;
125  data testjoin_DS ;
126     MERGE DATA.HIST01(IN=OK) DATA.CLASS_BIG ;
127     BY AGE ;
128     IF AGE=16 and OK;
129  run;
130  %Mend;
131
132  %merge

ERROR: BY variables are not properly sorted on data set DATA.CLASS_BIG.
OK=1 AAR=1984 AGE=14 Name=Alfred Sex=M Height=69 Weight=112.5 i=1000000 P=100000001 FIRST.AGE=0 LAST.AGE=0
_ERROR_=1
_N_=1000004
NOTE: The SAS System stopped processing this step because of errors.
NOTE: There were 6 observations read from the data set DATA.HIST01.
NOTE: There were 1000001 observations read from the data set DATA.CLASS_BIG.
WARNING: The data set WORK.TESTJOIN_DS may be incomplete.  When this step was stopped there were 0 observations and 8
         variables.
NOTE: DATA statement used (Total process time):
      real time           0.10 seconds
      cpu time            0.07 seconds
```

FULLSTIMER – viser ressourcer brugt

```
option notes source NOFULLSTIMER;
NOTE: DATA statement used (Total process time):
      real time           0.08 seconds
      cpu time            0.07 seconds
```

```
option notes source FULLSTIMER;
NOTE: DATA statement used (Total process time):
      real time           0.08 seconds
      user cpu time       0.06 seconds
      system cpu time     0.01 seconds
      memory              1118.15k
      OS Memory           12532.00k
      Timestamp           28/11/2022 07.02.07 e.m.
      Step Count                        14  Switch Count  0
```

MSGLEVEL=I viser ekstra Information
MSGLEVEL=n nulstiller

```
182
183  Data Teens ;
184       set data.CLASS_BIG_IA ;
185       where age=16 ;
INFO: Index Age selected for WHERE clause optimization.
186       run;
NOTE: There were 1000000 observations read from the data set DATA.CLASS_BIG_IA.
      WHERE age=16;
NOTE: The data set WORK.TEENS has 1000000 observations and 7 variables.



194  PROC SQL ;
195  create table Teens as select *
196       from data.CLASS_BIG_IA
197       where age=16 ;
INFO: Index Age selected for WHERE clause optimization.
NOTE: Table WORK.TEENS created, with 1000000 rows and 7 columns.
```

Proc SQL stimer _method;

```
230  PROC SQL stimer _method ;
231  create table Teens as select *
232      from data.CLASS_BIG_IA
233      where age=15
234      order by name;

NOTE: SQL execution methods chosen are:
      sqxcrta
          sqxsort
              sqxsrc( DATA.CLASS_BIG_IA )
NOTE: SAS threaded sort was used.
INFO: Index Age selected for WHERE clause optimization.
NOTE: Table WORK.TEENS created, with 4000000 rows and 7 columns.

NOTE: SQL Statement used (Total process time):
      real time           1.41 seconds
      cpu time            1.86 seconds
```

Testdata CLASS_BIG

Data Set Name	WORK.CLASS_BIG	Observations	19.000.000
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	27/11/2022 22:17:13	Observation Length	56
Last Modified	27/11/2022 22:17:13	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Versioner

Tabel	Obs	Size
CLASS_BIG	19.000.000	1Gb
CLASS_BIG_IA(index=(age))	19.000.000	1Gb
CLASS_BIG_IS(index=(sex))	19.000.000	1Gb
CLASS_BIG_IAS(index=(IX1=(age SEX)))	19.000.000	1Gb
CLASS_BIG_II(index=(p));	19.000.000	1Gb

Testdata HIST01

Data Set Name	WORK.HIST01	Observations	44
Member Type	DATA	Variables	2
Engine	V9	Indexes	1
Created	28/11/2022 13:56:05	Observation Length	16
Last Modified	28/11/2022 13:56:05	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Typisk udtræk

```
384  data testmerge_DS ;
385      MERGE HIST01(IN=OK) CLASS_BIG ;
386      BY AGE ;
387      IF AGE=16;
388  run;
```

Typisk udtræk

```
384  data testmerge_DS ;
385      MERGE HIST01(IN=OK) CLASS_BIG ;
386      BY AGE ;
387      IF AGE=16;
388  run;

ERROR: BY variables are not properly sorted on data set WORK.CLASS_BIG.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: There were 6 observations read from the data set WORK.HIST01.
NOTE: There were 1000001 observations read from the data set WORK.CLASS_BIG.
NOTE: DATA statement used (Total process time):
      real time           0.08 seconds
      user cpu time       0.07 seconds
      system cpu time     0.00 seconds
      memory              1110.59k
      OS Memory           29692.00k
```

Typisk udtræk – det fungerer

```
310     Proc Sql   stimer  _method ;
311      create table testjoin as
312          select a.*,B.* from HIST01 a LEFT join class_BIG b  on a.age = b.age
313          where a.age=16 ;

WARNING: Variable Age already exists on file WORK.TESTJOIN_SQL_I.
NOTE: SAS threaded sort was used.
NOTE: Table WORK.TESTJOIN created, with 1000000 rows and 8 columns.

NOTE: SQL Statement used (Total process time):
      real time           9.86 seconds
      user cpu time       3.84 seconds
      system cpu time     1.95 seconds
      memory              1.056.135.95k
      OS Memory           1.083.764.00k
```

Typisk udtræk – forsøg med index

```
329     Proc Sql   stimer  _method ;
330      create table testjoin_SQL_I as
331          select a.*,B.* from HIST01 a LEFT join class_BIG_IA b  on a.age = b.age
332          where a.age=16 ;

WARNING: Variable Age already exists on file WORK.TESTJOIN_SQL_I.
NOTE: SAS threaded sort was used.
NOTE: Table WORK.TESTJOIN_SQL_I created, with 1000000 rows and 8 columns.

NOTE: SQL Statement used (Total process time):
      real time           12.54 seconds
      user cpu time        3.70 seconds
      system cpu time      2.01 seconds
      memory              1.056.216.57k
      OS Memory           1.083.764.00k
```

Sorteret tabel = mulighed for merge

```
342     Proc Sql   stimer  _method ;
343      create table testjoin_SQL_S as
344          select a.*,B.* from HIST01 a LEFT join class_BIG_SORT_age b  on a.age = b.age
345          where a.age=16 ;

WARNING: Variable Age already exists on file WORK.TESTJOIN_SQL_S.
NOTE: Table WORK.TESTJOIN_SQL_S created, with 1000000 rows and 8 columns.

NOTE: SQL Statement used (Total process time):
      real time           3.05 seconds
      user cpu time       0.40 seconds
      system cpu time     0.15 seconds
      memory              5707.59k
      OS Memory           34.816.00k
```

Sortering for senere anvendelse

```
395  proc sort data=CLASS_BIG out=WORK.CLASS_BIG_SORT_AGE ;
396  BY AGE ; RUN ;

NOTE: There were 19000000 observations read from the data set WORK.CLASS_BIG.
NOTE: SAS threaded sort was used.
NOTE: The data set WORK.CLASS_BIG_SORT_AGE has 19000000 observations and 7 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           5.84 seconds
      user cpu time       3.89 seconds
      system cpu time     2.04 seconds
      memory              1061926.84k
      OS Memory           1.090.816.00k
```

Datastep merge igen -

```
469  data testjoin_DS ;
470      MERGE HIST01(IN=OK) CLASS_BIG_SORT_AGE ;
471      BY AGE ;
472      IF AGE=16 and OK;
473  run;

NOTE: There were 44 observations read from the data set WORK.HIST01.
NOTE: There were 19000000 observations read from the data set
WORK.CLASS_BIG_SORT_AGE.
NOTE: The data set WORK.TESTJOIN_DS has 1000000 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           1.43 seconds
      user cpu time       1.29 seconds
      system cpu time     0.14 seconds
      memory              1110.59k
      OS Memory           29.692.00k
```

Datastep med merge – og Where

```
475  data testjoin_DS ;
476      MERGE HIST01(IN=OK) CLASS_BIG_SORT_age ;
477      where age=16;
478      BY AGE ;
479  run;
```

```
NOTE: There were 1 observations read from the data set WORK.HIST01.
      WHERE age=16;
NOTE: There were 1000000 observations read from the data set WORK.CLASS_BIG_SORT_AGE.
      WHERE age=16;
NOTE: The data set WORK.TESTJOIN_DS has 1000000 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           0.70 seconds
      user cpu time       0.40 seconds
      system cpu time     0.26 seconds
      memory              1131.75k
      OS Memory           29692.00k
```

Ressourcer brugt

Metode	CPU	Memory
Data step merge u/sort	Error:	29.692.00k
SQL LEFT JOIN	9.36	1.083.764.00k
SQL Left join m Index	12.57	1.083.764.00k
SQL Left Join m/Sort	3.05	34.816.00k
Sortering	5.85	1.090.816.00k
Datastep Merge m Sort og if	1.43	29.692.00k
Datastep Merge m sort og where	0.70	29.692.00k

”Partnere”
mod og medspillere

- OS (windows/Unix)
- Windows cache
- SAS options
- Disksystemer
- Fysiske fil struktur
- Finurligheder

Check loggen / SAS Base index

Hvis begge , eller første variabel i index angives , bruges index

```
604  create table test as
605  select * from WORK.CLASS_BIG_IAS
606  where age=15 and sex="M";
NOTE: SQL execution methods chosen are:
      sqxcrta
          sqxsrc( WORK.CLASS_BIG_IAS )
INFO: Index IX1 selected for WHERE clause optimization.
NOTE: Table WORK.TEST created, with 2000000 rows and 7 columns.
```

Hvis anden variabel angives , bruges index ikke

```
590  create table test as
591  select * from WORK.CLASS_BIG_IAS
592  where  sex="M";
NOTE: SQL execution methods chosen are:
      sqxcrta
          sqxsrc( WORK.CLASS_BIG_IAS )
NOTE: Table WORK.TEST created, with 10000000 rows and 7 columns.
```

Check Loggen / SPDE

Spde kan anvende index efter behov

```
691  create table test as
692  select * from WSPDE.CLASS_BIG_IAS2
693  where sex="F" and age=15;
NOTE: SQL execution methods chosen are:
      sqxcrta
          sqxsrc( WSPDE.CLASS_BIG_IAS2 )
whinit: WHERE ((Sex='F') and (Age=15))
whinit: INDEX Sex uses 47% of segs (WITHIN maxsegratio 75%)
whinit: INDEX Age uses 21% of segs (WITHIN maxsegratio 75%)
whinit returns: ALL EVAL1(w/SEGLIST) EVAL2
NOTE: Table WORK.TEST created, with 2000000 rows and 7 columns.
```

Check Loggen /SPDSserver - trivielt false

```
700  create table test as
701  select * from WSPDS.CLASS_BIG_IAS
702  where age=10 and sex="M";
NOTE: SQL execution methods chosen are:
      sqxcrta
          sqxsrc( WSPDS.CLASS_BIG_IAS )
whinit: WHERE ((Age=10) and (Sex='M'))
whinit: wh-tree presented
                      /-NAME = [Age]
           /-CEQ----|
          |           \-LITN = [10]
 --LAND---|
          |           /-NAME = [Sex]
           \-CEQ----|
                      \-LITC = ['M']
whinit: wh-tree after split
 --<empty>
whinit: INDEX IX1 uses 0% of segs (WITHIN maxsegratio 75%)
whinit: INDEX tree after split
                      /-NAME = [Age] <1>INDEX IX1 (Age,Sex)
           /-CEQ----|
          |           \-LITN = [10]
 --LAND---|
          |           /-NAME = [Sex] COMPOSITE INDEX IX1
           \-CEQ----|
                      \-LITC = ['M']
whinit costing: detects no TRUE segments (TRIVIALLY FALSE)
whinit returns: FALSE
```

Check loggen / SAS Base – trivielt FALSE også her

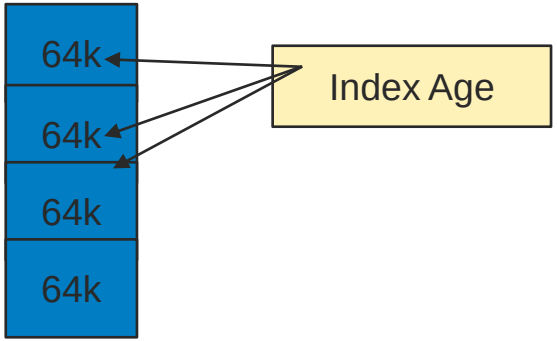
```
733  create table test as
734  select * from WORK.CLASS_BIG_IAS
735  where age=10 and sex="M";

NOTE: SQL execution methods chosen are:
      sqxcrta
          sqxsrc( WORK.CLASS_BIG_IAS )
INFO: Index IX1 selected for WHERE clause optimization.
NOTE: Table WORK.TEST created, with 0 rows and 7 columns.

NOTE: SQL Statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
```

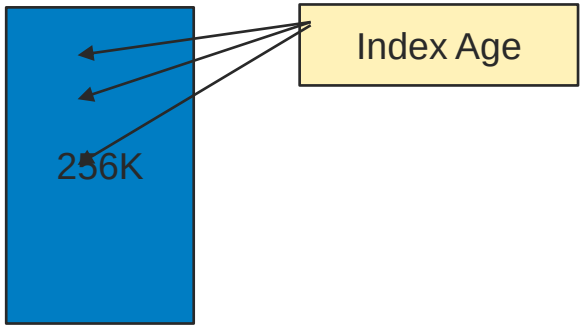
BUFSIZE=64K(default) / 256k
File blokke til data

Cardinalitet:
Antallet af elementer i et sæt data
Påvirker index størrelsen og effektivitet

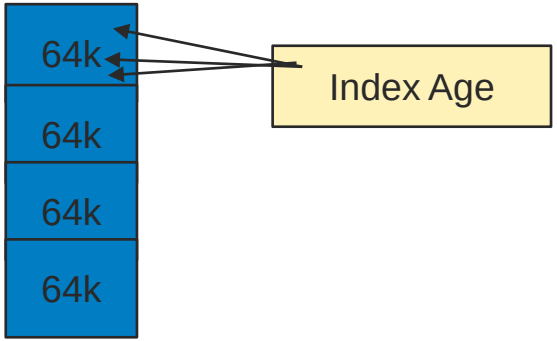
Variabel	Værdier	Cardinalitet
AGE	11,12,13,14,15,16	6
Sex	”F”,”M”	2
Name	”Alfred”,”Alice”,…”William”	19
Timer	0 – 23	24
Dage i et år	1JAN-31Dec	365 (366)
Dage i en periode		365.25* antal År

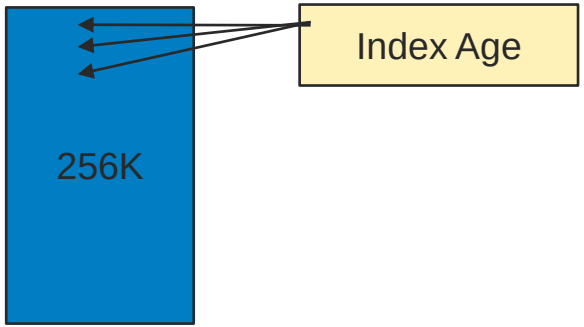
BUFSIZE=64K(default) / 256k





BUFSIZE=64K(default) / 256k
Sortering før index forbedrer performance ved læs





SAS Views og index

```
481  data HIST01 (index=(aar)) HIST02 (index=(aar)) HIST03 (index=(aar));
482  DO AAR=1980 to 2023 ;
483      AGE=AAR-1970;
484      output;
485  END;
486  RUN;

NOTE: The data set WORK.HIST01 has 44 observations and 2 variables.
NOTE: Simple index aar has been defined.
NOTE: The data set WORK.HIST02 has 44 observations and 2 variables.
NOTE: Simple index aar has been defined.
NOTE: The data set WORK.HIST03 has 44 observations and 2 variables.
NOTE: Simple index aar has been defined.
NOTE: DATA statement used (Total process time):
      real time           0.06 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              1036.12k
      OS Memory           29952.00k
```

Datastep

```
488  DATA UD1 ;
489  SET HIST01  HIST02 HIST03 ;
490  where Aar in (2020 2011 1998 2022);
INFO: Index AAR selected for WHERE clause optimization.
INFO: Index AAR selected for WHERE clause optimization.
INFO: Index AAR selected for WHERE clause optimization.
491  run;

NOTE: There were 4 observations read from the data set WORK.HIST01.
      WHERE Aar in (1998, 2011, 2020, 2022);
NOTE: There were 4 observations read from the data set WORK.HIST02.
      WHERE Aar in (1998, 2011, 2020, 2022);
NOTE: There were 4 observations read from the data set WORK.HIST03.
      WHERE Aar in (1998, 2011, 2020, 2022);
NOTE: The data set WORK.UD1 has 12 observations and 2 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1419.75k
      OS Memory           29692.00k
```

Data step View – men index anvendes ikke

```
501  data HIST  / VIEW=HIST ;
502      SET HIST01 HIST02 HIST03 ;
503  RUN;
NOTE: DATA STEP view saved on file WORK.HIST.

505  DATA UD1 ;
506  SET HIST;
507  where Aar =2020;
508  run;

NOTE: View WORK.HIST.VIEW used (Total process time):

NOTE: There were 44 observations read from the data set WORK.HIST01.
NOTE: There were 44 observations read from the data set WORK.HIST02.
NOTE: There were 44 observations read from the data set WORK.HIST03.
NOTE: There were 3 observations read from the data set WORK.HIST.
      WHERE Aar=2020;
NOTE: The data set WORK.UD1 has 3 observations and 2 variables.
```

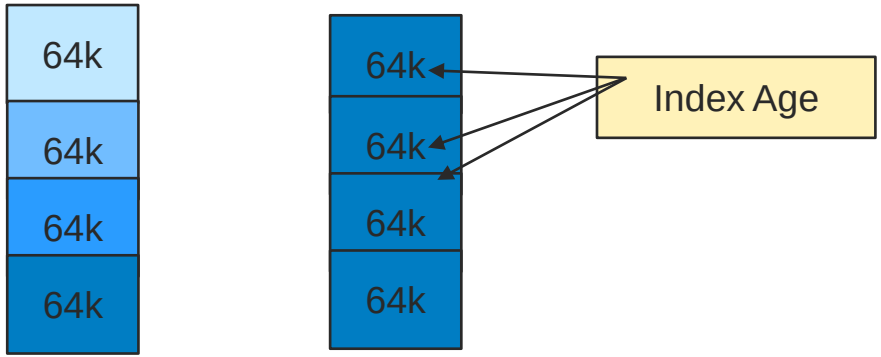
SQL View

```
510  proc sql ;
511       create view HIST_SQLV as select * from HISTO1
512           union ALL select * from HISTO2
513           union ALL select * from HISTO3  ;
NOTE: SQL view WORK.HIST_SQLV has been defined.
514  run;
```

SQL View - bruger index ..

```
528  DATA UD2 ;
529  SET HIST_SQLV;
530  where Aar IN (2020,2019);
531  run;

INFO: Index AAR selected for WHERE clause optimization.
INFO: Index AAR selected for WHERE clause optimization.
INFO: Index AAR selected for WHERE clause optimization.
NOTE: There were 2 observations read from the data set WORK.HIST01.
      WHERE AAR in (2019, 2020);
NOTE: There were 2 observations read from the data set WORK.HIST02.
      WHERE AAR in (2019, 2020);
NOTE: There were 2 observations read from the data set WORK.HIST03.
      WHERE AAR in (2019, 2020);
NOTE: There were 6 observations read from the data set WORK.HIST_SQLV.
      WHERE Aar in (2019, 2020);
NOTE: The data set WORK.UD2 has 6 observations and 2 variables.
NOTE: DATA statement used (Total process time):
      real time           0.06 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              6445.81k
      OS Memory           34560.00k
```

SAS Byprocessing



- Erstat Sortering af data med et index der leverer data i den ønskede rækkefølge
- Anvendes på tabeller der ikke kan forblive i den ønskede sorteringsorden

By group processing

```
538  proc freq data=WORK.CLASS_BIG ;
539  BY AGE ;
540  TABLE SEX / LIST ;
541  RUN ;

ERROR: Data set WORK.CLASS_BIG is not sorted in ascending sequence.
The current BY group has Age = 14 and the next BY
       group has Age = 13.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: There were 1000001 observations read from the data set
WORK.CLASS_BIG.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.30 seconds
      user cpu time       0.09 seconds
      system cpu time     0.00 seconds
      memory              668.87k
      OS Memory           29436.00k
```

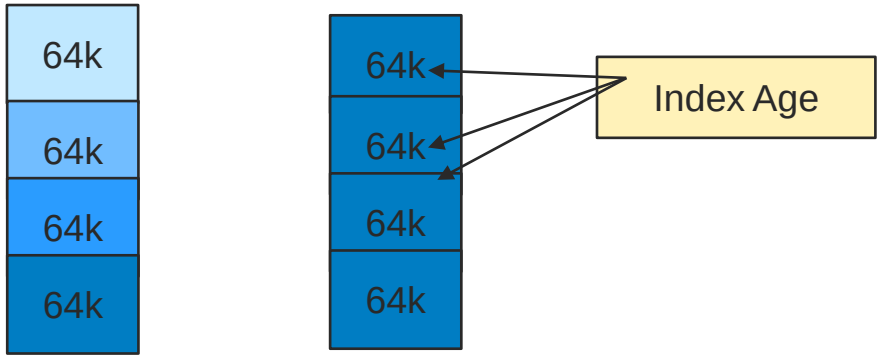
Bygruppe med sorteret input

```
543  proc freq data=WORK.CLASS_BIG_SORT_AGE ;
544  BY AGE ;
545  TABLE SEX / LIST ;
546  RUN ;

NOTE: There were 19000000 observations read from the data set WORK.CLASS_BIG_SORT_AGE.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           4.80 seconds
      user cpu time       1.37 seconds
      system cpu time     0.21 seconds
      memory              779.50k
      OS Memory           29436.00k
```

Bygruppe via index

```
548  proc freq data=WORK.CLASS_BIG_IA ;
549  BY AGE ;
INFO: Index Age selected for BY clause processing.
NOTE: An index was selected to execute the BY statement.
      The observations will be returned in index order rather than in physical order.  The
selected index is for the
      variable(s):
 Age
550  TABLE SEX / LIST ;
551  RUN ;

NOTE: There were 19000000 observations read from the data set WORK.CLASS_BIG_IA.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           6.19 seconds
      user cpu time       5.20 seconds
      system cpu time     0.56 seconds
      memory              841.46k
      OS Memory           29436.00k
```

Nøgletal



Bygruppe	Sorteret	Via Index
Real time	4.90 seconds	6.19 seconds
User CPU	1.37 seconds	5.20 seconds

Nyttige SAS Options

- Options FULLSTIMER MSGLEVEL=i NOTES SOURCE;
- Proc SQL STIMER _METHOD ;

- Ved RDBMS:
  options sastraceloc=saslog sastrace=",,d";

- Ved SPDServer:
  %let spdswdeb=YES;

- Tak for i dag

- Spørgsmål ?
- Henrik.Dorf@SAS.com