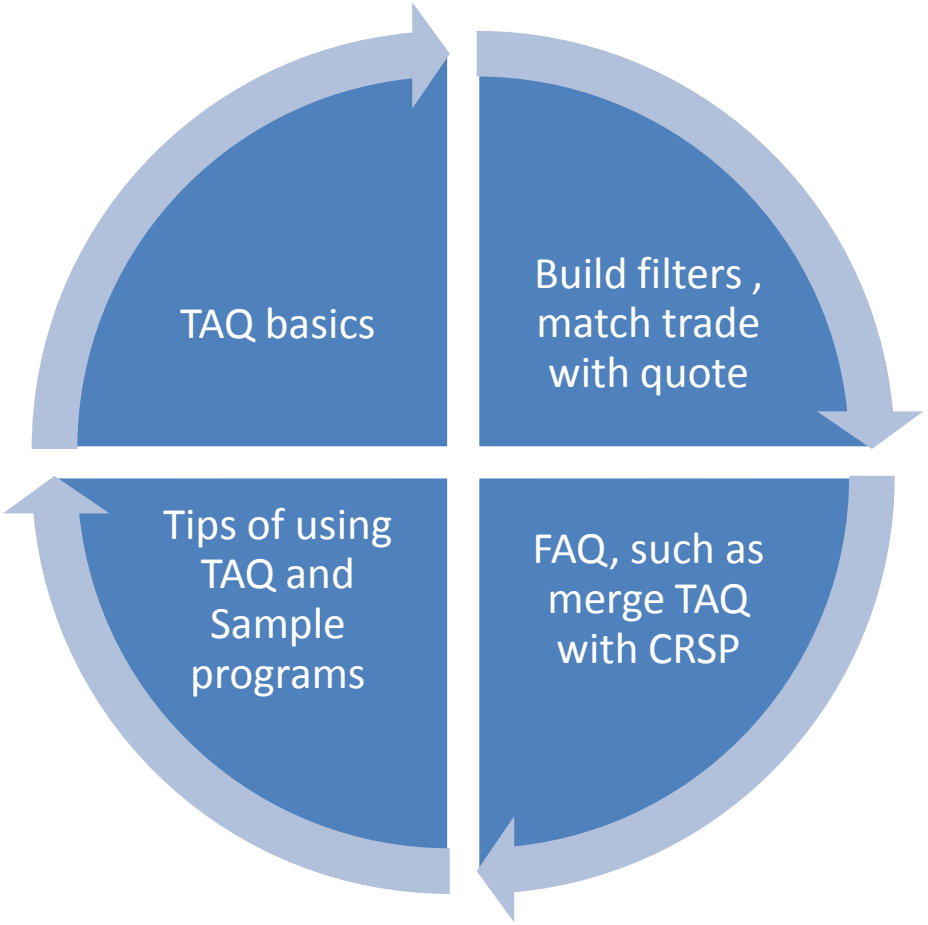
Introduction to TAQ

Yuxing Yan
Associate Director
Wharton Research Data Services
October 2, 2007

Agenda


TAQ basics
Build filters , match trade with quote
Tips of using TAQ and Sample programs
FAQ, such as merge TAQ with CRSP

I: TAQ basics

TAQ stands for Trade And Quote database

- Data supplied by NYSE
- Data range: 1993 to 2007

Data is organized by month

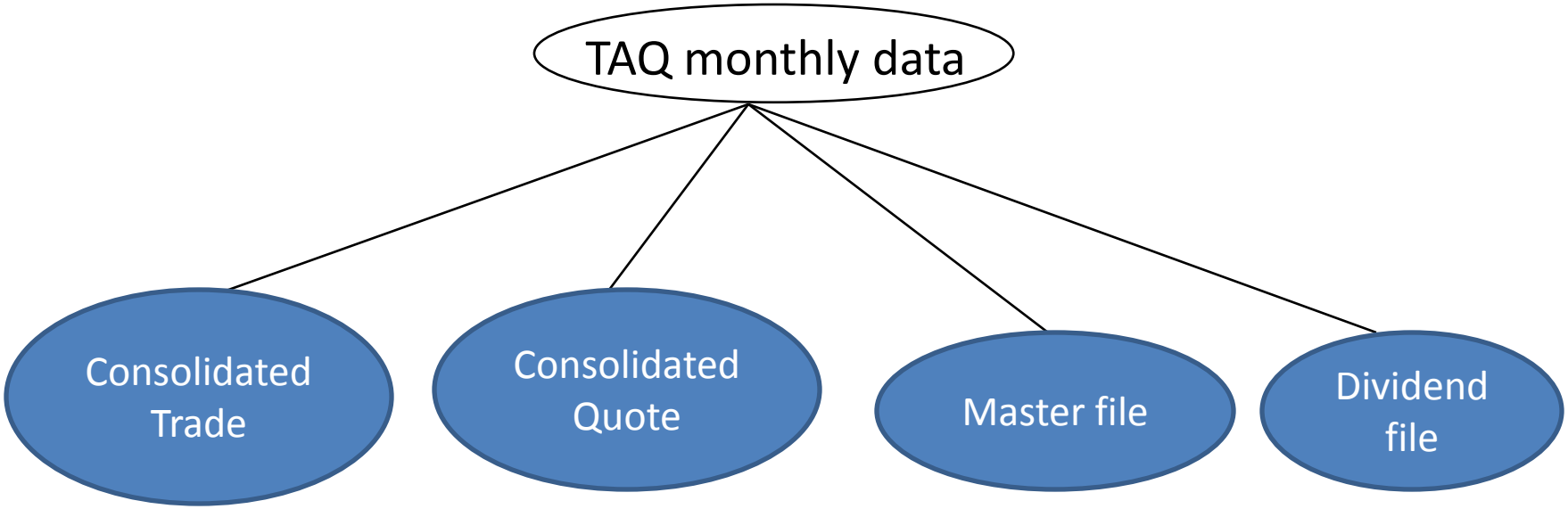
- 4 basic data sets for each month
- CT, CQ, Dividend and Master file

ISSM (Institute for the Study of Security Markets)

- Range: 1983 to 1992 (NYSE), 1987-1992 (NASDAQ)
- Data organized by year

I: TAQ basics

TAQ data is organized by month and each month has 4 related data sets: CT (Consolidated Trade), CQ (consolidated Quote), MAST (master file) and DIV (Dividend file).



e.g., January 2000, CT0001, CQ0001, MAST0001, DIV0001

Locations of TAQ data sets on WRDS

- When use SAS, the prefixed libname TAQ is referred to “/wrds/taq/sasdata/”;

```
%let yymm=0001;
Proc print data=taq.ct&yymm(obs=100);
run;
```

Physical locations of data sets

/wrds/taq93axs/taq93/sasdata

/wrds/taq94axs/taq94/sasdata

……

/wrds/taq00axs/taq00/sasdata

/wrds/taq/taq04axs/taq04a/sasdata

/wrds/taq/taq04axs/taq04b/sasdata

First 10 lines from Consolidated Trade

Obs	SYMBOL	DATE	TIME	PRICE	SIZE	G127	CORR	COND	EX	TSEQ
1	A	20000103	9:34:01	78.75	64700	40	0		N	807127
2	A	20000103	9:34:04	78.75	100	0	0		M	0
3	A	20000103	9:34:04	78.75	1000	0	0		M	0
4	A	20000103	9:34:04	78.75	100	0	0		M	0
5	A	20000103	9:34:04	78.75	200	0	0		M	0
6	A	20000103	9:34:04	78.75	100	0	0		M	0
7	A	20000103	9:34:04	78.75	100	0	0		M	0
8	A	20000103	9:34:04	78.75	100	0	0		M	0
9	A	20000103	9:34:04	78.75	100	0	0		M	0
10	A	20000103	9:34:04	78.75	100	0	0		M	0

Several variables for CT

SYMBOL this variable is not a permanent stock

G127 Combination of following 3 rules

G rule: trading for its own account

127 rule: executed as a block position

Stopped stock indicator

e.g., G127=0, does not qualify as “G”, Rule 12 or stopped stock trade

G127=40 A display book-reported trade

CORR Correction indicator

e.g, CORR=0 regular trade

COND Condition of a trade

e.g., COND=‘A’ Cash-only basis

EX : stock exchange code

A	AMEX
B	Boston
C	Cincinnati
D	NASD ADF and TRF (after 5/15/2006)
M	Chicago
N	NYSE
P	Pacific
X	Philadelphia
T/Q	NASD (no more after 6/28/2006)
W	CBOE

First 10 lines from Consolidated Quote

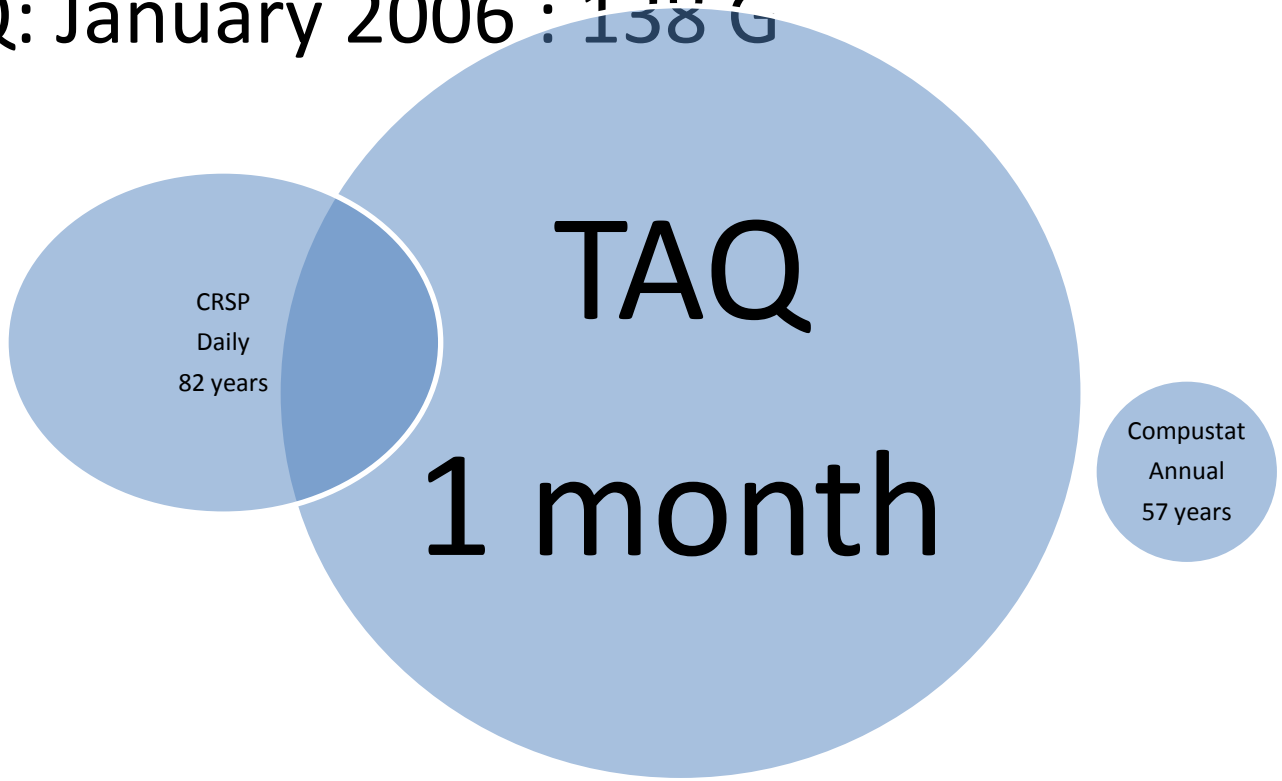
Obs	SYMBOL	DATE	TIME	BID	OFR	BIDSIZ	OFRSIZ	MODE	EX	MMID	QSEQ
1	A	20000103	8:59:07	0.000	0.000	0	0	12	T	PTRS	0
2	A	20000103	8:59:07	0.000	0.000	0	0	12	T	SWST	0
3	A	20000103	8:59:07	0.000	0.000	0	0	12	T	TRIM	0
4	A	20000103	8:59:07	0.000	0.000	0	0	12	T	MADF	0
5	A	20000103	9:34:02	0.000	0.000	0	0	12	C		0
6	A	20000103	9:34:08	78.625	78.875	10	10	10	N		807129
7	A	20000103	9:34:10	78.500	79.000	1	1	12	X		0
8	A	20000103	9:34:10	77.750	79.750	1	1	12	C		0
9	A	20000103	9:34:12	78.500	79.000	1	1	12	T	MADF	0
10	A	20000103	9:34:12	78.500	79.000	1	1	12	T	CAES	0

Several variables for CQ

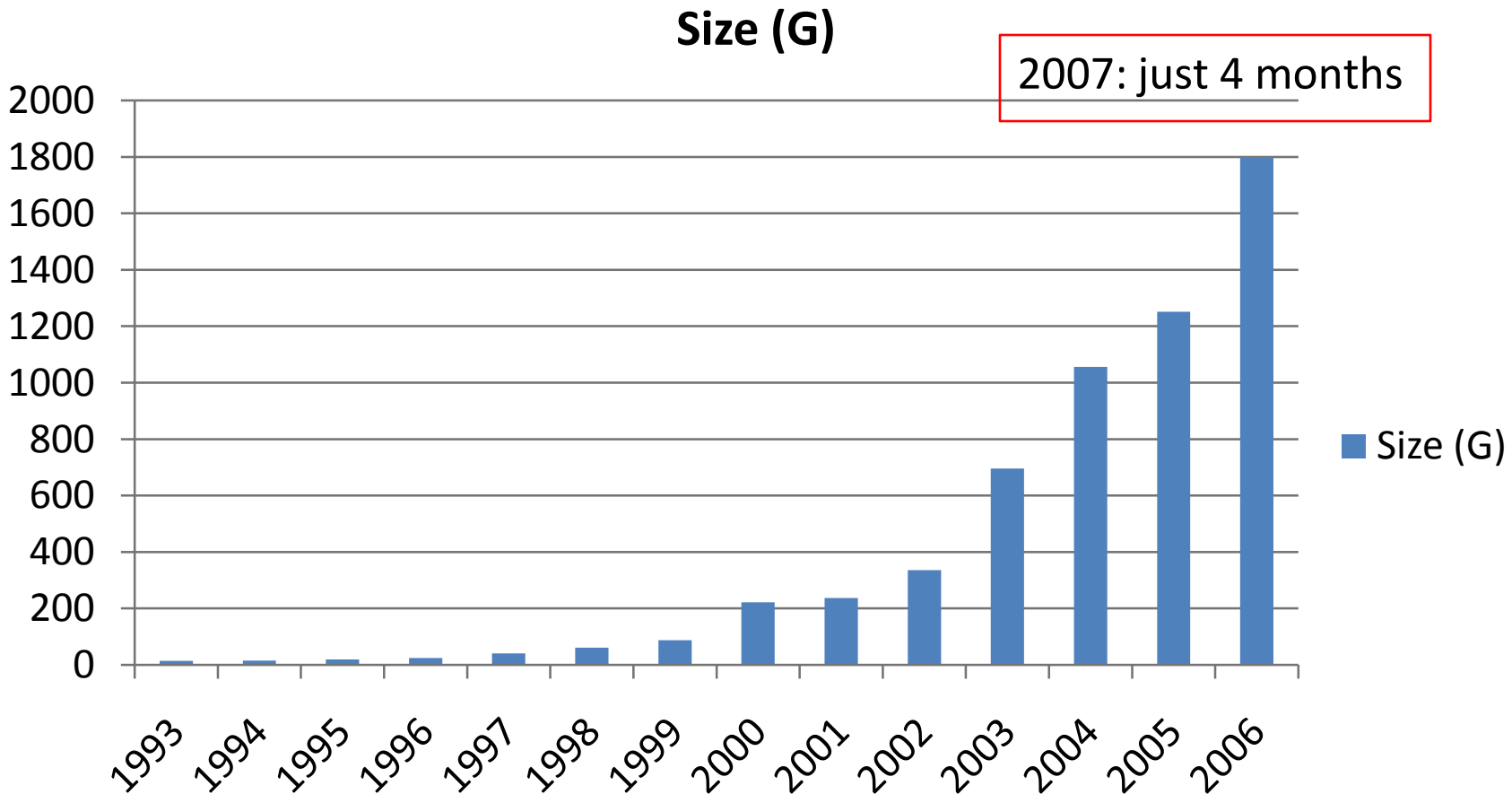
BID	Bid price
OFR	Offer price
BIDSIZ	Bid size (100 share units)
OFRSIZ	Offer size (100 share units)
MODE	Quote condition
	e.g. MODE=0 Invalid field
	MODE=4 News dissemination (regulatory halt)
MMID	NASDAQ market maker

Size matters!

- CRSP daily stocks:13 (data) + 5 (index) = 18G
- Compustat: annual 1.6 +0.8=2.4 G
- TAQ: January 2006 : 138 G



Growth of size of TAQ over years



III: Data Issues in TAQ

1. Filtering out *invalid* trades
2. Filtering out *invalid* quotes
3. matching trades with quotes
4. Merging TAQ and CRSP
5. FAQs

Filtering out invalid trades

Keep if

1) price: price >0

2) size : size >0

3) CORR: Correction indicator

CORR = 0, 1 or 2

4) COND : Sale Condition

COND not in ("O" "Z" "B" "T" "L" "G" "W" "J" "K")

CORR-correction indicator (CT)

Good Trades

0 Regular trade
1 Trades were later corrected
2 Symbol correction

Original Trade Records

7 Trade cancelled due to error
8 Trade cancelled
9 Trade cancelled due to symbol correction

Correction Instructions

10 Cancel record (associated with 8)
11 Error record (associated with 7)
12 Correction record (associated with 1)

Correction Indicator (CT0001)

CORR	Frequency	Cumulative Percent	Cumulative Frequency	Percent
0	64511091	99.66	64511091	99.66
1	34701	0.05	64545792	99.71
7	2117	0.00	64547909	99.71
8	74313	0.11	64622222	99.83
10	74313	0.11	64696535	99.94
11	2117	0.00	64698652	99.95
12	34701	0.05	64733353	100.00

Example code=1 and 12

Obs	SYMBOL	DATE	TIME	PRICE	SIZE	G127	CORR	COND	EX	TSEQ
1	A	20000103	10:14:50	72.0000	100	0	1	Z	B	0
2	A	20000103	10:15:48	72.0000	100	0	12		B	0
3	A	20000103	11:19:03	69.2500	2800	0	1		B	0
4	A	20000103	11:24:58	69.1875	2800	0	12		B	0
5	A	20000103	11:33:14	70.7500	100	0	1		M	0
6	A	20000103	11:33:31	70.5000	100	0	12		M	0
7	A	20000103	12:55:12	71.8750	2000	0	1		B	0
8	A	20000103	13:00:57	71.8125	2000	0	12		B	0
9	A	20000103	15:31:06	71.3750	500	0	1		M	0
10	A	20000103	15:31:30	71.5625	500	0	12		M	0

Add filters for CT

```
data trades;
        set taq.ct0001;
         where price>0 and size>0 and
        corr in    (0,1,2) and  cond not in
       ("O" "Z" "B" "T" "L" "G" "W"    "J" "K" );
run;
```

COND: condition of sale

- COND='O'
  - an opening trade that occurs in sequence but is reported to the tape in a later time
- COND='B'
  - Bunched trade (aggregate of two or more regular trades executed within 60 seconds with same price
- COND='G'
  - A bunched trade not reported within 90 seconds

Filtering out invalid Quotes

Keep if

1) price: bid >0, ofr >0

2) size : bidsiz>0 , ofrsiz>0

3) mode: mode not in
(4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28)

e.g.,

mode=4: regulatory halt (news dissemination)

mode=7: non-regulatory halt (order imbalance)

mode=9: regulatory halt

Codes for filtering out invalid quotes

```
data quotes;
        set taq.cq0207;
        where bid>0 and ofr>0 and
         bidsiz>0 and ofrsiz >0 and
        mode not in
        (4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28);
run;
```

Matching trades with quotes

- Matching by SYMBOL, DATE and TIME

- Method 1: 5-second rule, Lee and Ready (1991)
  - Delay of the repor time for a trade
  - An isolated trade is a trade within a window just one trade in it.
    - Objective : identify the patterns of the delay of quotes entered the system.
    - An isolated trade is the first trade between 11:00am to 2:30pm with no other trades within 2-minute window (Lee and Ready,1991)

Matching trades with quotes – other rules

- Method 2: 0-second rule,
  - Peterson and Sirri (2003), and Bessembinder (2003)

- Method 3: 1-second rule
  - Henker and Wang (2005)

Q 1: how to retrieve data for certain dates?

For example, below if the input file

A 07SEP2002
AA 02JAN2001
IBM 07SEP2002
GE 01JUN1999

```
data temp;
        infile 'symbol_date.txt';
        informat date date9.;
       format date date9.;
       input symbol $ date;
run;
```

```
%macro get_data;
    %do i=1 %to 4;
        data temp2;
            set temp(obs=&i firstobs=&i);
            y=mod(year(date),100);
            m=month(date);
           call symput('year',put(y,z2.));
           call symput('month',put(m,z2.));
         run;
        proc sql;
            create table temp3 as select bb.*   from temp2 ,taq.ct&year&month as bb
            where temp2.symbol = bb.symbol;
        quit; run;
       proc append base=final data=temp3;run;
%end;
%mend get_data;

%get_data;
```

Q 2: how to merge TAQ with CRSP?

Unlike PERMNO in CRSP, GVKEY in Compustat, SYMBOL is not a permanent stock ID

Obs	YYYYMM	SYMBOL	cusip8	NAME
1	199302	A	04987020	ATTWOODS PLC ADS REP5 ORD/5PN
2	199605	A	04987020	ATTWOODS PLC ADS REP5 ORD/5PN
3	199605	A	04629810	ASTRA AB CL-A ADS 1CL-ASEK2.5
4	199911	A	04629810	ASTRA AB CL-A ADS 1CL-ASEK2.50
5	199911	A	00846U10	AGILENT TECHNOLOGIES INC
6	200612	A	00846U10	AGILENT TECHNOLOGIES, INC

Master files are useful

Master file could be used to get CUSIP

SYMBOL Stock Symbol
CUSIP 9-digit CUSIP + 3-digit NSCC exchange id
FDATE Effective Date

Obs	SYMBOL	CUSIP	FDATE
1	A	00846U101000	20020412
2	A	00846U101000	20020710
3	AA	013817101000	20010423
4	AAA	02143N103000	20020529
5	AAA	02143N103000	20020710
6	AAAB	007231103002	20010813
7	AAAB	007231103002	20020701
8	AABC	00431F105002	19980917
9	AABC	00431F105002	20020701
10	AAC	00371F206001	20010129

A useful data set called TAQNAMES

SYMBOL	Stock Symbol
CUSIP	CUSIP (9-digit +3-digit exchange id)
begin	1st Month & Year for this SYMBOL
end	Last Month & Year for this SYMBOL
NAME	Company Name

Obs	SYMBOL	NAME	CUSIP	begin	end
1	A	AGILENT TECHNOLOGIES INC	00846U101000	NOV1999	AUG2006
2	A	ASTRA AB CL-A ADS 1CL-ASEK2.50	046298105000	MAY1996	NOV1999
3	A	ATTWOODS PLC ADS REP5 ORD/5PN	049870207000	JAN1993	MAY1996
4	AA	ALCOA INC	013817101000	JAN1999	AUG2006
5	AA	ALUMINUM CO AMERICA	022249106000	JAN1993	JAN1999
6	AAA	ALTANA AKTIENGESELLSCHAFT SPON	02143N103000	MAY2002	AUG2006
7	AAA	ASCO PLC ADS REP 5 ORD SHS	04363R103000	JUN2000	JUN2000
8	AAA	US ALCOHOL TESTING OF AMER IN	91154J101001	JAN1993	OCT1996
9	AAAB	ADMIRALTY BANCORP INC CL B	007231103002	AUG2001	JAN2003
10	AAABB	ADMIRALTY BANCORP INC CL B	007231103002	SEP1998	AUG2001

Q3: how to identify preferred stocks?

One variable called TYPE in master file

Code	Description
0	common
1	preferred
2	warrant
3	right
4	other
5	derivative

III: Sample programs related to TAQ

Location : /wrds/taq/samples/

taq0.sas
taq1.sas
taq1old.sas
taq2.sas
taq3.sas
taq4.sas
taq4a.sas
taq4b.sas
taq5.sas
taq6.sas
taq6.sas~
taq_old.sas
taqquote.sas
taqtrade.sas
tradequote.sas

TAQ1.SAS (sample program)

- For example, taq1.sas defines a macro to retrieve data

- %getdata(file=CT,begdate='29APR1997'd,enddate='03MAY1997'd,query="IBM" "DELL",vars=price size ex,outlib=mylib,outds=my_taq_ct);

- The output is a SAS data set called my_taq_ct

TAQ2.sas (sample program)

1) With a macro called TAQQ()
2) Read a input file called tick.txt (give SYMBOL)

%taqq(CT,'11apr1997'd,'14apr1997'd,~/tick.txt,label,price size ex);

Tick.txt has to entries:

A

AA

- %taqq(CT,'11apr1997'd,'14apr1997'd,./tick.txt,label,price size ex);

TRADEQUOTE.SAS (sample program)

Match Trade with Quote

Consider 5-second rule (Lee and Ready ,1991)

```
        qtime_reported=time;
        time=time+5;
        qtime_adjusted=time;
        format time qtime_reported qtime_adjusted time9.;
        rename ex=q_ex;
        type='Q';
    run;
```

Generate a SAS data set called tr_qt, see several lines on the next slide

Obs	SYMBOL	PRICE	SIZE	t_ex	ttime_ reported	prior_ q_ex	prior_ mmid	prior_ qtime adjusted
1	GE	85.625	60500	N	9:30:47	T		9:00:10
2	GE	85.625	100	M	9:30:52	N		9:30:52
3	GE	85.625	200	M	9:30:54	N		9:30:52
4	GE	85.625	100	M	9:30:54	N		9:30:52
5	GE	85.625	100	M	9:30:54	N		9:30:52
6	GE	85.625	100	C	9:30:56	X		9:30:56
7	GE	85.625	500	B	9:31:01	B		9:31:01
8	GE	85.625	800	B	9:31:01	B		9:31:01

Obs	prior_ qtime_ reported	prior_ bid	prior_ ofr	prior_ bidsiz	prior_ ofrsiz	prior_ mode
1	9:00:05	0.000	0.000	0	0	12
2	9:30:47	85.500	85.750	100	100	12
3	9:30:47	85.500	85.750	100	100	12
4	9:30:47	85.500	85.750	100	100	12
5	9:30:47	85.500	85.750	100	100	12
6	9:30:51	85.375	85.875	1	1	12
7	9:30:56	85.375	85.875	1	1	12
8	9:30:56	85.375	85.875	1	1	12

~

Tips about programming using TAQ data

1) use a small sample to debug your program

```
data temp;
        set taq.ct0207(obs=500000);
run;
```

2) retrieve only needed variables

```
data temp;
        set taq.ct0207(keep=symbol date time price corr cond);
         where corr in (1,2, 12);
run;
```

Tips on programming (2)

3) Use loops to get all periods

```
%macro all_period;

   %do j=1993 %to 2006;

                %do i=1 %to 12;
                        *(add main program here) ;
                %end;
    %end;

%mend all_period;

%all_period;
```

```
%let ds=ct;    * CT for trade, CQ for Quote, DIV for dividend and MAST for master;
%macro all_period;

    %do j=1993 %to 2006;
        %let year=%sysfunc(substr(&j,3,2));

        %do i=1 %to 12;
            %let prefix=0;
            %if &i>=10 %then %let prefix=;
            title " trade for yera=&j month=&i";
            data temp;
                set taq.&ds&year&prefix.&i(obs=5000);
            run;
            proc append base=final data=temp;run;
        %end;

    %end;
%mend all_period;
%all_period;
```

Tips on programming (3)

4) Use index to speed up the data retrieval.

a) TAQ is sorted by SYMBOL, DATE and TIME

b) use “where” statement instead of “if”

```
data temp;
        set taq.ct0207;
        where symbol="IBM";
run;
proc print data=temp(obs=10);run;
```

Tips on programming (4)

5) Generate intermediate permanent data sets

6) Use UNIX background process

nohup sas t.sas &

7) Use shell language (or C)

Conclusions

- Users need to use SAS to process TAQ data
- Pay attention to filters
- Using 5-second, 1-seoncd or zero-second rule
- When merge TAQ with CRSP using taqnames
- Some sample programs are available on WRDS