WRDS Overview of TAQ

TAQ is a historical data product that offers intraday, tick-by-tick trade and quote data of all activity within the U.S. National Market System.

General Description

The **Trade and Quote** (TAQ) database contains intraday transactions data (trades and quotes) for all securities listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), the Nasdaq National Market System (NMS), and all other U.S. equity exchanges.

TAQ is the **primary database** for **intraday** market research. With TAQ, a researcher can perform such analyses as daily volatility estimation, probability of informed trading, short-term impact of breaking news, back-testing of intraday trading strategies, and assessment of programmed trading protocols.

TAQ is the **only WRDS database** that records market information at the **intraday** level (to the microsecond), covering roughly 8,000 stock issues listed on major American exchanges. Because it provides a record of every trade and quote on these exchanges (since January 1993), it is by far the largest WRDS database, growing about 0.75 terabytes monthly. No other database has market information at this level of detail.

Top of Section

Coverage

TAQ is a historical data product that offers complete trade and quote data of all activity within the U.S. National Market System. **Trade & Quote Daily Product** (09/10/2003 - present), **Trade & Quote Monthly Product** (01/01/1993 - 12/31/2014), and **NYSE Reg Sho Data** (01/01/2005 - 07/31/2007) are available on the WRDS platform. This overview covers the TAQ Monthly and Daily products; for further information regarding the NYSE Reg Sho product, refer to the WRDS Reg Sho overview.

The TAQ Daily and Monthly products are nearly identical in terms of their content. **Monthly** is delivered a whole month at a time, typically 60-90 days after the last trading day of the month. **Daily** is delivered one day at a time, hours after trading stops, and is available on WRDS the next day. Timestamps for the Monthly product are shown as **whole seconds**. In the Daily product timestamps are provided at **millisecond** granularity through March 2015, and in **microseconds** starting in April 2015. See the Database Notes section below for more details about the differences between the two products.

The times for the TAQ data in the monthly/daily products are the times events occurred in New York (Eastern Time) and include changes between standard and daylight savings time.

TAQ does not include **transaction data** reported outside of the Consolidated Tape hours of operation. As of August 2000, Consolidated Tape hours were 8:00 a.m. until 6:30 p.m. Eastern Time (ET). The tape opens at 4:00 a.m. ET, as of March 4, 2004. Additionally, trading in NYSE-listed securities between 8:00 a.m and 9:30 a.m. ET by other markets are not available in TAQ.

Top of Section

Dataset Organization

File Location Description

NYSE products are sold by the year. Due to the very large size of the data, each year is located in a different physical directory: `/wrds/taq.YYYY` for Monthly and `/wrds/nyse/taq_msecYYYY` for Daily. Files from all years are compiled into the directory: `/wrds/taq/sasdata`for monthly; `/wrds/nyse/sasdata/taqms/ct`, and `/wrds/nyse/sasdata/taqms/cq` for millisecond using symbolic link files.

Symbolic link files allow WRDS to distribute these very large files appropriately while being able to display them all in one place. When downloading the data using the SSH secure file transfer method, WRDS **recommends** using physical files - files located under `/wrds/taq.YYYY`, `/wrds/nyse/taq_msecYYYY` and so on - not the symbolic link files. To locate the exact location of these files, use the UNIX command `ls -li`.

In the monthly TAQ product, each trade and quote file covers a **single trading date**. Trades are stored in files whose name begins with the prefix `ct_`, followed by the date in `YYYYMMDD` format. Quotes are stored using the `cq_` prefix and the same date format (e.g. `cq_YYYYMMDD`). Typically, trades files are 1 Gigabyte (GB) in size, while quotes files may range from 20 to 50 GB. There are a couple of additional files in these datasets. `div_YYYYMM` are the dividend files (according to the manual). Please note that dividend and stock split data included in TAQ are known to be inaccurate for some securities. For identifying information, refer to the master files, which are formatted as `mast_YYYYMM`. There is one dividend file and one master file per month.

In the Daily TAQ product, `ctm_YYYYMMDD` **refers to trades**, `cqm_YYYYMMDD` **refers to quotes**, `mastm_YYYYMMDD` refers to the **master** files, `luld_c*m_YYYYMMDD` refers to the **LULD** quotes and trades files, and `ix_*` refers to the corresponding compressed index files. Daily TAQ also includes **NBBO** files in the format `nbbom_YYMMDD`. There is one master file per day, and the dividend file is not available.

All data files for TAQ are stored in the SAS file format. WRDS does not make the original raw data available.

File Location Table

Table 1: Dataset Location

Product	Primary SAS Libname	Primary PostgreSQL Schema	Unix Server	Unix Location	Files
Monthly	taq	n/a (SAS-only)	wrds.wharton.upenn.edu	/wrds/taq/sasdata/	cq_YYYYMMDD, ct_YYYYMMDD, div_YYYYMM, mast_YYYYMM
Monthly	taq	n/a (SAS-only)	wrds-cloud.wharton.upenn.edu	/wrds/taq/sasdata, /wrds/nyse/sasdata/wrds_taqs, /wrds/nyse/sasdata/wrds_taqs_nbbo, /wrds/nyse/sasdata/wrds_taqs_ct	cq_YYYYMMDD, ct_YYYYMMDD, div_YYYYMM, mast_YYYYMM, nbbo_YYYYMMDD, wct_YYYYMMDD
Daily (Millisecond)	taqmsec	taqm_YYYY (or) taqmsec	wrds-cloud.wharton.upenn.edu	/wrds/nyse/sasdata/taqms/ct, /wrds/nyse/sasdata/taqms/cq, /wrds/nyse/sasdata/taqms/mast, /wrds/nyse/sasdata/taqms/luld_cq, /wrds/nyse/sasdata/taqms/luld_ct, /wrds/nyse/sasdata/taqms/nbbo	ctm_YYYYMMDD, ix_ctm_YYYYMMDD, cqm_YYYYMMDD, ix_cqm_YYYYMMDD, luld_cqm_YYYYMMDD, luld_ctm_YYYYMMDD, mastm_YYYYMMDD, nbbom_YYYYMMDD, ix_nbbom_YYYYMMDD

Datasets Created by WRDS

The following listed files are generated by WRDS from the original TAQ Trades and Quotes datasets. Using the WRDS-created NBBO and WCT files discussed below can save days - or even weeks - of processing time, in addition to reducing programming effort.

National Best Bid & Offer (NBBO)

Using the TAQ Monthly product, WRDS compiles, to the second, the best (highest) bid and the best (lowest) offer from all prevailing quotes issued by various market makers in national exchanges for each stock. Generated by WRDS, these monthly product files are labeled `nbbo_YYYYMMDD` and are located on the WRDS Cloud server in `/wrds/nyse/sasdata/wrds_taqs_nbbo`.

In the TAQ Daily product, `nbbo_` files are provided by NYSE and contain continuous National Best Bid and Offer updates, as well as consolidated trades and quotes for all listed and non-listed issues. These NBBO files should be used in conjunction with several NBBO fields in the quote files.

For additional information regarding NBBO, refer to the NBBO research application.

WRDS Consolidated Trades (WCT)

The NBBO's Bid and Ask midpoints are matched to each trade at seconds 0, -1, -2 and -5 relative to their trade time, and stored along with trades in the same datasets. These WRDS-generated trades files are labeled as `wct_YYYYMMDD`. This approach provides users with all necessary components to infer the trade directions, regardless of what specifications and assumptions are employed on the trade-quote lag or on trade filters. Please see the Lee and Ready research application for more information. Additionally the following article, "Matching TAQ Trades and Quotes in the Presence of Multiple Quotes", may be useful for understanding more about WCT files.

`wct_YYYYMMDD` files are generated by WRDS using the monthly TAQ product and are located on the WRDS Cloud server at `/wrds/nyse/sasdata/wrds_taqs_ct`.

Index Files

SAS users may notice datasets starting with `ix_` in the Daily data. These are datasets that are used like indexes to speed up processing by the web query and can be ignored. More information can be found in the paper, "A Faster Index for Sorted SAS Datasets", available externally at the SAS support website.

Top of Section

Linking to Other Products

Identifiers Used

The primary identifier used in the TAQ monthly product is `SYMBOL`, commonly referred to as `TICKER`, or `Trading Symbol` and `Official Ticker`in other databases.

The complete lists of securities (firms) in TAQ datasets are available in the TAQ Master files. The TAQ monthly product master file (`mast_YYYYMM`) contains, for each `SYMBOL`, its `SHROUT`, `CUSIP`, company name, and `FDATE` (i.e. the effective date of those characteristics). TAQ monthly master files use 12-character `CUSIP`s. The first 9 digits are assigned by the Committee on Uniform Security Identification Procedure (CUSIP). Digits and characters 1 through 6 identify the issuer; 7-9 identify the issue. The final three digits are applied by the NSCC to distinguish between NYSE, AMEX, and NASD issues.

Table 2: Issue Type Extension

SYMBOL	NAME
NYSE	000
NYSE when issued	100
AMEX	001
AMEX when issued	101
NASD	002
NASD when issued	102

The following is an example taken from `MAST_200307` for Dell Computer Corporation. On July 22nd, 2003, Dell changed its `CUSIP` from 247025109 to be 24702R101.

Table 3: Information taken from MAST_200307 for Dell

SYMBOL	NAME	CUSIP	FDATE
DELL	DELL COMPUTER CORP	247025109001	20030703
DELL	DELL COMPUTER CORP	247025109002	20030602
DELL	DELL INC	24702R101001	20030722
DELL	DELL INC	24702R101002	20030723

To retrieve the 9-character `CUSIP`, use `cusip9 = substr(cusip,1,9)` in SAS.

To link CRSP with TAQ, refer to the article "Matching CRSP and TAQ Data" in the WRDS Knowledgebase. Additionally, the "tclink" research macro may be helpful.

The primary identifiers in the TAQ Daily product are `symbol_root` - the symbol that indicates the root of the security - and `symbol_suffix`, which is the NYSE or NASDAQ stock symbol suffix. See Appendix B and Appendix C of the TAQ Daily manual). TAQ Daily master files also provide the Trading Symbol (`symbol_15`) and 9-character CUSIP information besides `symbol_root` and `symbol_suffix`.

Below is an example for Dell in the `mastm_20100104` file. UOT is the unit of trade in Round-Lot value.

Table 4: Information taken from MAST_20100104 for Dell

SYMBOL_ROOT	SYMBOL_SUFFIX	SYMBOL_15	SEC_DESC	CUSIP	DATE	UOT
DELL		DELL	DELL INC	24702R101	20100104	100

Top of Section

Database Notes

WRDS Cloud

As shown in Table 1, Daily TAQ files are only available through the WRDS Cloud server, which is accessible at wrds-cloud.wharton.upenn.edu. WRDS Cloud is part of a grid server environment; running SAS and other code on this server is different than the main WRDS server (wrds.wharton.upenn.edu). To learn more about using the WRDS Cloud, refer to the "Getting Started" guide.

Efficient Programming Techniques

TAQ data is known to be extremely large. For example, the quote dataset for April 30, 2014 (`cq_20140430.sas7bdat`), occupies 32GB; its index `cq_20140430.sas7bndx` is 7GB, for a total of 39GB - this is for a single day of data.

Inefficient programming techniques in processing TAQ data sets can cause a system slowdown for all WRDS users, not just TAQ users. Using "Dow-Loop" data steps and SAS Views are the two of the most commonly used techniques in the SAS community to keep programs running efficiently when using TAQ data. WRDS recommends reading the documents, "SAS Dow Loop Approach" and "TAQ SAS Programming Issues" to gain deeper insight into creating more robust and effective programs.

To create a list of Daily TAQ file names, or to use both TAQ Trades and Quotes datasets and merge to create inferences on buyer or seller initiated trades, refer to the "TAQ SAS Programming Issues" document as well. To retrieve open and closing price, or transaction price for a fixed interval, the TAQ Sample Programs may be helpful.

Top of Section

Monthly and Daily (Millisecond) Product Comparisons

Trades

TAQ Monthly and Daily trades files have the same number of rows and the same values for price and other key variables.

TAQ Monthly product: Only one sale condition is displayed. Please see the latest TAQ User's Guide.

TAQ Daily product: One trade can have as many as four sale condition codes. This applies to all exchanges. Please refer to the latest available Daily TAQ Client Specification, and to the following table:

Table 5: Comparison of Sale Condition, Monthly and Daily

STOCK SYMBOL	TRANSACTION DATE	TRADE TIME	TAQ MONTHLY: SALE CONDITION	TAQ DAILY: TRADE SALE CONDITION (up to four codes)
IBM	20131202	8:00:00	F	FT
IBM	20131202	8:00:11		T
IBM	20131202	8:00:27		T
IBM	20131202	8:02:19		TB
IBM	20131202	9:10:10	F	FT
IBM	20131202	9:19:21	F	FT
IBM	20131202	9:30:00	F	FT
IBM	20131202	9:30:00	F	FT
IBM	20131202	9:30:00	F	FT
IBM	20131202	9:30:56	4	4B

IBM	20131202	8:00:27		B
IBM	20131202	8:02:19	4	4B

Quotes

Similar to the trade files, the Monthly and Daily quotes files have the same number of rows and the same values for bid, offer, and other key variables. Different codes are used for the quote condition variable.

Master Files

Master files have **different variables** between the two products. Please refer to the latest TAQ User's Guide for details.

TAQ Monthly Product: The master file is at the monthly level (`mast_YYYYMM`).

TAQ Daily Product: The master file is provided at the daily level (`mastm_YYYYMMDD`). As of April 2015, the earliest available date for `mastm_YYYYMMDD` files is January 2010.

Dividend Files

TAQ Monthly product provides dividend information (`div_*`), whereas the TAQ Daily product does not.

NBBO Files and Fields

TAQ Monthly product: NYSE does not provide NBBO files. Instead, these files are generated by WRDS. Please refer to "Datasets created by WRDS" in this overview for details.

TAQ Daily product: NYSE provides NBBO files. However, the `nbbom_YYYYMMDD` datasets provided by NYSE in the millisecond product do not have complete NBBO history. For more information, please see this article. In 2013 NYSE added the following extra items to NBBO files: LULD indicator (`LULD_INDICATOR`) (LULD Indicator), LULD NBBO indicator (`LULD_NBBO_INDICATOR`), and a SIP-generated Message Identifier (`SIP_MESSAGE_ID`).

Orders and Ties

TAQ Monthly Product: Data files are sorted by Symbol and Time.

TAQ Daily Product: Data files are sorted by symbol root, symbol suffix, time, and sequence number. Within a security, Daily and Monthly TAQ trades and quotes are in the same order, while the addition of the symbol suffix occasionally changes the location of one security relative to another within the dataset.

For both TAQ products, the trades and quotes are in a specific order even if the timestamps are the same. There are many ties at the second level, as well as a lot of ties at the millisecond level.

TAQ Master File Issues

From time to time, users will report inaccurate `CUSIP`s in the master files. Usually, inaccurate `CUSIP`s are fixed in later releases of the files. To resolve this issue, use the TCLink macro, which finds unmatched cases using the Exchange Ticker, and then back up `CUSIP`s from CRSP.

Top of Section

Technical Advice

Downloading and Converting TAQ Data

WRDS does not recommend downloading large amounts of data via a web query, unless it is for a small number of days or select companies.

As mentioned earlier, TAQ datasets are incredibly large in their file size. It is not possible for most users to download a complete month to their desktop machines.

The most efficient way to deal with large amounts of TAQ data is by using SAS via a UNIX connection. Work with subsets of data, or download the analysis of the data, rather than the raw data itself.

The following instructions provide users with the necessary steps to download and convert TAQ data into a variety of file formats.

1. Connect to WRDS using SSH.
2. Run `%df -h /sastemp*` and select the directory with the most available space.
3. Extract the data and output it in the desired format using the sample SAS code below. In the sample code below, *sastemp1* is selected and returns a comma separated file (.csv) for quotes from IBM between 9:30am and 9:45am on January 2014.

4. Transfer the output file using SSH or scp.

```
data vmydata / view=vmydata
    set taq.cq_201401:  open=defer;
        where symbol in ('IBM') and time between '09:30:00't and '09:45:00't;
    run;

    proc export data=vmydata outfile="/sastemp1/data_201401.csv" dbms=csv replace;
    run;
```

Tips

- In many cases, more space will be needed than is available in your home directory. Files older than 48 hours are automatically removed, among other restrictions. The WRDS Cloud provides extra scratch space, which is shared among all users at your institution, and is located at `/scratch/[group name]`.
- The colon at the end of the data set name tells SAS to read all data sets whose names begin with `taq.cq_201401`.
- Users can also use the `%taq_daily_dataset_list` macro, as mentioned in TAQ SAS Programming Issues, which provides more flexible control of the date range.
- The `open=defer` command tells SAS not to allocate a memory buffer for all the data sets simultaneously. Instead it will re-use the same memory buffer for each one in succession. This saves time and memory.

Top of Section

Top