Building the TAQ file

FE-570

This note describes the construction of a TAQ-style file from tick level data. It is implemented as `JPM_TAQ_generator.R`. The raw file data obtained from Refinitiv has the form shown in Fig. 1.

Start by loading the `xts,highfrequency` packages. Work in GMT time zone (EST = GMT + 5). NYSE trading hours 9:30-16:00 convert to 14:30 - 21:00 in GMT time.

Time stamps. The time stamps are specified to nanosecond (9 decimal points) but I could not find a way to keep all these decimal points. Time handling at sub-second levels is related to the OS used. I can keep only microsecond data (3 decimal points). This requires adding the line `options(digits.secs=3)`

Refinitiv has a "T" in the time stamp, which confuses R, and has to be removed for processing in R.

```
> head(rawdata)
  X.RIC       Domain                     Date.Time GMT.Offset  Type
1   JPM Market Price 2021-01-13T00:00:00.025318659Z         -5 Trade
2   JPM Market Price 2021-01-13T00:01:00.856874169Z         -5 Trade
3   JPM Market Price 2021-01-13T00:03:08.100867734Z         -5 Quote
4   JPM Market Price 2021-01-13T00:05:10.428868317Z         -5 Trade
5   JPM Market Price 2021-01-13T00:05:35.688872153Z         -5 Trade
6   JPM Market Price 2021-01-13T00:05:35.689307853Z         -5 Trade
  Ex.Cntrb.ID  Price Volume Buyer.ID Bid.Price Bid.Size Seller.ID Ask.Price
1         NYS 140.22      0                NA       NA                   NA
2         PSE 140.29      5                NA       NA                   NA
3          NA     NA      NA      THM    140.12        1       PSE     140.3
4         PSE 140.29      7                NA       NA                   NA
5         PSE 140.28     10                NA       NA                   NA
6         PSE 140.13      1                NA       NA                   NA
  Ask.Size
1       NA
2       NA
```

Figure 1: Tick level data file from Refinitiv with JPM prices for 13-Jan-2021.

```
# remove the "T" from the Date.Time
longdate <- c('2021-01-13T00:00:00.015487583Z')

shortdate <- gsub("T", " ", longdate, perl=TRUE)
```

Step 1. Collect the trades and quotes data into two separate data frames `tdata` and `qdata`.

```
# R code which converts Refinitiv dataset into TAQ style file
# combines the trades and quotes data to form a TAQ style file

library(xts)
library(highfrequency)

Sys.setenv(TZ = "GMT")
options(digits.secs=3) # keep millisecond timestamps

# read in the datafile obtained from Thompson Reuters
rawdata <- read.csv("JPM_Jan_13_2021_EXCH.csv", header = TRUE)

class(rawdata) # is a data frame

names(rawdata)
#summary(rawdata)
length(rawdata$Price) #439,204 entries (trades+quotes)

head(rawdata)

tdata <- subset(rawdata, Type=="Trade")
qdata <- subset(rawdata, Type=="Quote")

length(tdata$Price) #129,392 trades
length(qdata$Bid.Price) #309,812 quotes

# filter the trades data with a subset of columns
tdata.small <- data.frame(TIME = gsub("T", " ", tdata$Date.Time, perl=TRUE),
                          SYMBOL = "JPM",
                          PRICE = tdata$Price,
```

```
                          SIZE = tdata$Volume,
                          EX = tdata$Ex.Cntrb.ID)

# filter the quotes for each product
qdata.small <- data.frame(TIME = gsub("T", " ", qdata$Date.Time, perl=TRUE),
                          SYMBOL = "JPM",
                          BID = qdata$Bid.Price,
                          BIDSIZ = qdata$Bid.Size,
                          OFR = qdata$Ask.Price,
                          OFRSIZ = qdata$Ask.Size)

head(qdata.small)
```

Step 2. Convert the `tdata` and `qdata` data frames to xts and combine them with `matchTradesQuotes`.

```
class(tdata.small) # is data.frame.
# must put it in xts format, to act on it with aggregateTrades

tdata.xts <- xts(tdata.small[,-1],
             order.by=as.POSIXct(tdata.small[,1],
             format = "%Y-%m-%d %H:%M:%OS"))

qdata.xts <- xts(qdata.small[,-1],
             order.by=as.POSIXct(qdata.small[,1],
             format = "%Y-%m-%d %H:%M:%OS"))


class(tdata.xts) # ok this is in xts format
head(tdata.xts) # usual TAQ format
head(qdata.xts)

> head(tdata.xts) # usual TAQ format
                        SYMBOL PRICE      SIZE      EX
2021-01-13 00:00:00.025 "JPM"  "140.2200" "      0" "NYS"
2021-01-13 00:01:00.856 "JPM"  "140.2900" "      5" "PSE"
2021-01-13 00:05:10.428 "JPM"  "140.2900" "      7" "PSE"
```

```
2021-01-13 00:05:35.688 "JPM"  "140.2800" "     10" "PSE"
2021-01-13 00:05:35.689 "JPM"  "140.1300" "      1" "PSE"
2021-01-13 00:05:35.689 "JPM"  "140.1100" "    106" "PSE"
> head(qdata.xts)
                        SYMBOL BID      BIDSIZ OFR      OFRSIZ
2021-01-13 00:03:08.100 "JPM"  "140.12" "  1"  "140.30" "  5"
2021-01-13 00:05:35.689 "JPM"  "140.11" "  8"  "140.30" "  5"
2021-01-13 00:17:21.964 "JPM"  "140.11" "  9"  "140.30" "  5"
2021-01-13 00:19:24.788 "JPM"  "140.11" "  8"  "140.30" "  5"
2021-01-13 00:19:24.884 "JPM"  "140.11" "  9"  "140.30" "  5"
2021-01-13 00:20:51.568 "JPM"  "140.12" "  1"  "140.30" "  5"

# merge trade and quote data
tqdata = matchTradesQuotes(tdata.xts, qdata.xts)

head(tqdata,10) # the first two rows have NA because no quotes available

> head(tqdata,10) # this is the TAQ file for JPM
                        SYMBOL BID      BIDSIZ OFR      OFRSIZ PRICE
2021-01-13 00:00:00.025 "JPM"  NA       NA     NA       NA     "140.2200"
2021-01-13 00:01:00.856 "JPM"  NA       NA     NA       NA     "140.2900"
2021-01-13 00:05:10.428 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.2900"
2021-01-13 00:05:35.688 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.2800"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1300"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1100"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1500"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1200"
2021-01-13 00:13:52.968 "JPM"  "140.11" "  8"  "140.30" "  5"  "140.2900"
2021-01-13 00:14:40.644 "JPM"  "140.11" "  8"  "140.30" "  5"  "140.3000"
                        SIZE
2021-01-13 00:00:00.025 "      0"
2021-01-13 00:01:00.856 "      5"
2021-01-13 00:05:10.428 "      7"
2021-01-13 00:05:35.688 "     10"
2021-01-13 00:05:35.689 "      1"
2021-01-13 00:05:35.689 "    106"
2021-01-13 00:05:35.689 "     22"
2021-01-13 00:05:35.689 "    110"
```

```
2021-01-13 00:13:52.968 "     25"
2021-01-13 00:14:40.644 "    199"

tqdata <- na.omit(tqdata) # do not delete before next step.

# this gives a clean TAQ-style data in xts format
length(tqdata$PRICE) # contains 129,390 rows, one for each trade

head(tqdata)

> head(tqdata)
                        SYMBOL BID      BIDSIZ OFR      OFRSIZ PRICE
2021-01-13 00:05:10.428 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.2900"
2021-01-13 00:05:35.688 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.2800"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1300"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1100"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1500"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1200"
                        SIZE
2021-01-13 00:05:10.428 "      7"
2021-01-13 00:05:35.688 "     10"
2021-01-13 00:05:35.689 "      1"
2021-01-13 00:05:35.689 "    106"
2021-01-13 00:05:35.689 "     22"
2021-01-13 00:05:35.689 "    110"
```

Step 3. Adding the exchange information. For some reason `matchTradesQuotes` drops the EX column, containing the exchange information. We have to add it back in.

```
# build a data frame with EX column at the end

tqdataEX <- data.frame(TIME = tdata.small$TIME,
                       SYMBOL = tqdata$SYMBOL,
                       BID = tqdata$BID,
                       BIDSIZ = tqdata$BIDSIZ,
                       OFR = tqdata$OFR,
```

```
                        OFRSIZ = tqdata$OFRSIZ,
                        PRICE = tqdata$PRICE,
                        SIZE = tqdata$SIZE,
                        EX = tdata$Ex.Cntrb.ID)

# before conversion to xts remove NA
tqdataEX <- na.omit(tqdataEX)

tqdataEX.xts <- xts(tqdataEX[,-1],
                    order.by=as.POSIXct(tqdataEX[,1], format = "%Y-%m-%d %H:%M:%OS"))
# this is a xts file

head(tqdataEX.xts)
> head(tqdata)
                         SYMBOL BID      BIDSIZ OFR      OFRSIZ PRICE
2021-01-13 00:05:10.428 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.2900"
2021-01-13 00:05:35.688 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.2800"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1300"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1100"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1500"
2021-01-13 00:05:35.689 "JPM"  "140.12" "  1"  "140.30" "  5"  "140.1200"
                         SIZE      EX
2021-01-13 00:05:10.428 "      7" "PSE"
2021-01-13 00:05:35.688 "     10" "PSE"
2021-01-13 00:05:35.689 "      1" "PSE"
2021-01-13 00:05:35.689 "    106" "PSE"
2021-01-13 00:05:35.689 "     22" "THM"
2021-01-13 00:05:35.689 "    110" "THM"
```

Step 4. Convert to EST time and keep only exchange hours 9:30 - 16:00.

```
Sys.setenv(TZ = "EST")
head(tqdataEX.xts)
tail(tqdataEX.xts)

###############################################
tradhrsEST <- '2021-01-13 09:30:00::2021-01-13 16:00:00'
```

```
length(tqdataEX.xts[tradhrsEST]$PRICE) # 127,720 trades

tqdataMktHrs <- tqdataEX.xts[tradhrsEST]

######## Save only MktHrs on EST ###################
head(tqdataMktHrs)

> head(tqdataMktHrs)
                        SYMBOL BID      BIDSIZ OFR      OFRSIZ PRICE
2021-01-13 09:30:00.398 "JPM"  "138.44" "  1"  "138.66" "  1"  "138.6500"
2021-01-13 09:30:00.401 "JPM"  "138.44" "  1"  "138.66" "  1"  "138.6400"
2021-01-13 09:30:01.237 "JPM"  "138.33" "  1"  "138.66" "  1"  "138.9500"
2021-01-13 09:30:01.269 "JPM"  "138.33" "  1"  "138.66" "  1"  "138.5900"
2021-01-13 09:30:01.273 "JPM"  "138.33" "  1"  "138.66" "  1"  "138.9000"
2021-01-13 09:30:01.285 "JPM"  "138.33" "  1"  "138.66" "  1"  "138.7900"
                        SIZE      EX
2021-01-13 09:30:00.398 "      3" "PSE"
2021-01-13 09:30:00.401 "    100" "THM"
2021-01-13 09:30:01.237 "      1" "BAT"
2021-01-13 09:30:01.269 " 319239" "NYS"
2021-01-13 09:30:01.273 "    200" "IEX"
2021-01-13 09:30:01.285 "     11" "THM"
> tail(tqdataMktHrs)
                        SYMBOL BID      BIDSIZ OFR      OFRSIZ PRICE
2021-01-13 15:59:59.177 "JPM"  "140.35" "  2"  "140.38" " 39"  "140.3498"
2021-01-13 15:59:59.177 "JPM"  "140.35" "  2"  "140.38" " 39"  "140.3491"
2021-01-13 15:59:59.372 "JPM"  "140.33" " 14"  "140.35" " 10"  "140.3500"
2021-01-13 15:59:59.784 "JPM"  "140.34" " 12"  "140.36" " 17"  "140.3500"
2021-01-13 15:59:59.853 "JPM"  "140.34" " 12"  "140.36" "  1"  "140.3500"
2021-01-13 16:00:00.025 "JPM"  "140.36" " 14"  "140.38" " 25"  "140.3400"
                        SIZE      EX
2021-01-13 15:59:59.177 "     22" "ADF"
2021-01-13 15:59:59.177 "    550" "ADF"
2021-01-13 15:59:59.372 "    100" "NYS"
2021-01-13 15:59:59.784 "     35" "NYS"
2021-01-13 15:59:59.853 "    577" "NYS"
2021-01-13 16:00:00.025 "     51" "NYS"
```

```
# save the TAQ file for later processing
save(tqdataMktHrs, file = "taqdata_JPM_20210113_ESTMktHrs.RData")

write.csv(tqdataMktHrs, file = "taqdata_JPM_20210113_ESTMktHrs.csv")
##############################################
```

This leads to a major size reduction: while the original Refinitiv .csv data file is 36.6 MB, the .RData file is 0.8MB (the .csv file is somewhat larger 9.2MB)

The dataset can be further cut down into 1-hr or less slices.

```
# extract only one hr of data. subsetting with xts files
tqdata.1 <- tqdata["T10:00/T11:00"]
tqdata.2 <- tqdata["T14:30/T15:00"]   #GMT = ET + 5

length(tqdata.2$PRICE)

head(tqdata.2,10)
```

Visualizing the data.

```
# Plot prices the first hour after market open

plot(as.numeric(tqdata.2$PRICE),col="red", type="l", ylab="Trade price",
     xlab="Trade #", main="JPM price", ylim=c(137,142))
#lines(mids, type="l", col="blue")

plot(as.numeric(tqdata.2$SIZE),col="red", type="l", ylab="Trade price",
     xlab="Trade #", main="Trade volume", ylim=c(0,100000))
```