Kaggle Rain Training Set Timing Study

Dave Hurst

October 7, 2015

```
library(data.table)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.2
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:data.table':
##
##     between, last
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Sys.info()
```

```
##                      sysname                      release
##                    "Windows"                      "7 x64"
##                      version                     nodename
## "build 7601, Service Pack 1"                     "BRAZIL"
##                      machine                        login
##                    "x86-64"                      "Dave"
##                         user               effective_user
##                       "Dave"                      "Dave"
```

```
tx <- list( proc.time())
tcheck <- function(t=1) {
    # t=0 to reset counter, t=1 incremental time output,  t=n time difference from n intervals
    t <- min( t, length(tx))
    pt <- proc.time()
    if (t == 0) {
        tx <<- list( proc.time())
    } else {
        tx <<- c( tx, list(pt))
        tn <- length(tx)
        print ( tx[[tn]] - tx[[tn-t]])
    }
}

tcheck(0)
train <- fread("../train.csv")
```

```
##
Read 0.0% of 13765201 rows
Read 1.0% of 13765201 rows
Read 2.0% of 13765201 rows
Read 3.1% of 13765201 rows
Read 4.3% of 13765201 rows
Read 5.2% of 13765201 rows
Read 6.1% of 13765201 rows
Read 7.1% of 13765201 rows
Read 8.1% of 13765201 rows
Read 9.2% of 13765201 rows
Read 10.2% of 13765201 rows
Read 11.1% of 13765201 rows
Read 12.1% of 13765201 rows
Read 13.1% of 13765201 rows
Read 14.2% of 13765201 rows
Read 15.2% of 13765201 rows
Read 16.2% of 13765201 rows
Read 17.2% of 13765201 rows
Read 18.3% of 13765201 rows
Read 19.3% of 13765201 rows
Read 20.2% of 13765201 rows
Read 21.1% of 13765201 rows
Read 22.1% of 13765201 rows
Read 23.0% of 13765201 rows
Read 24.0% of 13765201 rows
Read 25.0% of 13765201 rows
Read 25.9% of 13765201 rows
Read 26.9% of 13765201 rows
Read 27.8% of 13765201 rows
Read 28.8% of 13765201 rows
Read 29.7% of 13765201 rows
Read 30.7% of 13765201 rows
Read 31.8% of 13765201 rows
Read 32.9% of 13765201 rows
Read 33.9% of 13765201 rows
Read 34.9% of 13765201 rows
Read 35.8% of 13765201 rows
Read 36.8% of 13765201 rows
Read 37.8% of 13765201 rows
Read 38.8% of 13765201 rows
Read 39.9% of 13765201 rows
Read 41.0% of 13765201 rows
Read 42.1% of 13765201 rows
Read 43.0% of 13765201 rows
Read 44.1% of 13765201 rows
Read 45.3% of 13765201 rows
Read 46.2% of 13765201 rows
Read 47.2% of 13765201 rows
Read 48.2% of 13765201 rows
Read 49.1% of 13765201 rows
Read 50.2% of 13765201 rows
Read 51.1% of 13765201 rows
Read 52.1% of 13765201 rows
```

```
Read 53.1% of 13765201 rows
Read 54.0% of 13765201 rows
Read 55.1% of 13765201 rows
Read 56.0% of 13765201 rows
Read 57.0% of 13765201 rows
Read 58.1% of 13765201 rows
Read 59.2% of 13765201 rows
Read 60.2% of 13765201 rows
Read 61.2% of 13765201 rows
Read 62.2% of 13765201 rows
Read 63.3% of 13765201 rows
Read 64.4% of 13765201 rows
Read 65.4% of 13765201 rows
Read 66.3% of 13765201 rows
Read 67.3% of 13765201 rows
Read 68.3% of 13765201 rows
Read 69.2% of 13765201 rows
Read 70.2% of 13765201 rows
Read 71.3% of 13765201 rows
Read 72.3% of 13765201 rows
Read 73.2% of 13765201 rows
Read 74.1% of 13765201 rows
Read 75.0% of 13765201 rows
Read 76.0% of 13765201 rows
Read 77.0% of 13765201 rows
Read 78.0% of 13765201 rows
Read 79.0% of 13765201 rows
Read 80.0% of 13765201 rows
Read 80.9% of 13765201 rows
Read 81.9% of 13765201 rows
Read 82.9% of 13765201 rows
Read 83.8% of 13765201 rows
Read 84.9% of 13765201 rows
Read 85.9% of 13765201 rows
Read 86.9% of 13765201 rows
Read 87.9% of 13765201 rows
Read 88.8% of 13765201 rows
Read 89.9% of 13765201 rows
Read 91.0% of 13765201 rows
Read 92.0% of 13765201 rows
Read 93.0% of 13765201 rows
Read 94.0% of 13765201 rows
Read 95.0% of 13765201 rows
Read 95.9% of 13765201 rows
Read 96.8% of 13765201 rows
Read 98.0% of 13765201 rows
Read 99.2% of 13765201 rows
Read 13765201 rows and 24 (of 24) columns from 1.163 GB file in 00:02:03
```

```r
object.size(train)
```

```
## 2532802144 bytes
```

```
tcheck()
```

```
##    user  system elapsed
##  104.63    3.18  123.00
```

```
summary(train); tcheck()   #benchmark operation
```

```
##        Id           minutes_past    radardist_km        Ref
##  Min.   :      1   Min.   : 0.00   Min.   : 0.00   Min.   :-31
##  1st Qu.: 296897   1st Qu.:15.00   1st Qu.: 9.00   1st Qu.: 16
##  Median : 592199   Median :30.00   Median :11.00   Median : 22
##  Mean   : 592337   Mean   :29.52   Mean   :11.07   Mean   : 23
##  3rd Qu.: 889582   3rd Qu.:44.00   3rd Qu.:14.00   3rd Qu.: 30
##  Max.   :1180945   Max.   :59.00   Max.   :21.00   Max.   : 71
##                                                    NA's   :7415826
##   Ref_5x5_10th      Ref_5x5_50th      Ref_5x5_90th      RefComposite
##  Min.   :-32       Min.   :-32       Min.   :-28       Min.   :-32
##  1st Qu.: 14       1st Qu.: 16       1st Qu.: 18       1st Qu.: 18
##  Median : 20       Median : 22       Median : 26       Median : 24
##  Mean   : 20       Mean   : 23       Mean   : 26       Mean   : 25
##  3rd Qu.: 26       3rd Qu.: 29       3rd Qu.: 34       3rd Qu.: 32
##  Max.   : 62       Max.   : 69       Max.   : 72       Max.   : 92
##  NA's   :8481213   NA's   :7408719   NA's   :6213920   NA's   :7048858
##  RefComposite_5x5_10th RefComposite_5x5_50th RefComposite_5x5_90th
##  Min.   :-31           Min.   :-28           Min.   :-25
##  1st Qu.: 16           1st Qu.: 18           1st Qu.: 20
##  Median : 22           Median : 24           Median : 27
##  Mean   : 22           Mean   : 24           Mean   : 27
##  3rd Qu.: 28           3rd Qu.: 32           3rd Qu.: 35
##  Max.   : 66           Max.   : 71           Max.   : 94
##  NA's   :8009528       NA's   :7053538       NA's   :5935998
##      RhoHV         RhoHV_5x5_10th    RhoHV_5x5_50th    RhoHV_5x5_90th
##  Min.   :0         Min.   :0         Min.   :0         Min.   :0
##  1st Qu.:1         1st Qu.:1         1st Qu.:1         1st Qu.:1
##  Median :1         Median :1         Median :1         Median :1
##  Mean   :1         Mean   :1         Mean   :1         Mean   :1
##  3rd Qu.:1         3rd Qu.:1         3rd Qu.:1         3rd Qu.:1
##  Max.   :1         Max.   :1         Max.   :1         Max.   :1
##  NA's   :8830285   NA's   :9632047   NA's   :8828633   NA's   :7859617
##       Zdr           Zdr_5x5_10th      Zdr_5x5_50th      Zdr_5x5_90th
##  Min.   :-8        Min.   :-8        Min.   :-8        Min.   :-8
##  1st Qu.: 0        1st Qu.:-1        1st Qu.: 0        1st Qu.: 1
##  Median : 0        Median :-1        Median : 0        Median : 2
##  Mean   : 1        Mean   :-1        Mean   : 0        Mean   : 2
##  3rd Qu.: 1        3rd Qu.: 0        3rd Qu.: 1        3rd Qu.: 3
##  Max.   : 8        Max.   : 8        Max.   : 8        Max.   : 8
##  NA's   :8830285   NA's   :9632047   NA's   :8828633   NA's   :7859617
##       Kdp           Kdp_5x5_10th       Kdp_5x5_50th      Kdp_5x5_90th
##  Min.   :-96       Min.   :-81        Min.   :-79       Min.   :-100
##  1st Qu.: -1       1st Qu.: -5        1st Qu.: -1       1st Qu.:   2
##  Median :  0       Median : -3        Median :  0       Median :   4
##  Mean   :  0       Mean   : -3        Mean   :  0       Mean   :   4
##  3rd Qu.:  2       3rd Qu.: -2        3rd Qu.:  0       3rd Qu.:   6
```

```
##  Max.   :180       Max.   :  4        Max.   : 13       Max.   : 145
##  NA's   :9582566   NA's   :10336419   NA's   :9577920   NA's   :8712425
##     Expected
##  Min.   :    0.01
##  1st Qu.:    0.25
##  Median :    1.02
##  Mean   :  108.63
##  3rd Qu.:    3.81
##  Max.   :33017.73
##
```

```
##    user  system elapsed
##   32.52    3.18   35.94
```

```
#data table way
train[, median(Expected)] ; tcheck()   #median of entire set
```

```
## [1] 1.016001
```

```
##    user  system elapsed
##    0.81    0.05    0.88
```

```
train[,mean(Expected),Id][,median(V1)]; tcheck()  #median of station means
```

```
## [1] 1.016001
```

```
##    user  system elapsed
##    0.39    0.10    0.53
```

```
#dplyr way (no pipe)
summarise( summarise( group_by( train, Id ), mean = mean(Expected)), median(mean)); tcheck()
```

```
## Source: local data table [1 x 1]
##
##   median(mean)
##          (dbl)
## 1     1.016001
```

```
##    user  system elapsed
##    2.02    4.88   20.00
```

```
#dplyr way (with pipe)
train %>% group_by(Id) %>%
    summarise( V1 = mean(Expected)) %>%
    summarise( median(V1))
```

```
## Source: local data table [1 x 1]
##
##   median(V1)
##        (dbl)
## 1   1.016001
```

```
tcheck()
```

```
##    user  system elapsed
##    1.61    1.37   59.87
```

```
#dplyr way after converting to tbl_df
tbl_df(train) %>% group_by(Id) %>%
    summarise( V1 = mean(Expected)) %>%
    summarise( median(V1))
```

```
## Source: local data frame [1 x 1]
##
##   median(V1)
##        (dbl)
## 1   1.016001
```

```
tcheck()
```

```
##    user  system elapsed
##    4.49    0.47    7.75
```

```
#remove NA's
object.size(train)
```

```
## 2532802144 bytes
```

```
na_obs <- train %>%
    select( starts_with("Ref"), starts_with("Rho"), starts_with("Zdr"), starts_with("Kdp")) %>%
    .[, is.na(.SD)] %>%
    rowSums() == 20
tcheck()
```

```
##    user  system elapsed
##   15.49    9.25 1351.99
```

```
sum(na_obs) / length(na_obs)
```

```
## [1] 0.3841742
```

```
train <- train[ ! na_obs, ]
object.size(train)
```

```
## 1559766896 bytes
```

```
tcheck()
```

```
##    user  system elapsed
##    1.67    1.98 1637.54
```

```
summary(train); tcheck()   #benchmark operation
```

```
##        Id           minutes_past    radardist_km         Ref
##  Min.   :      2   Min.   : 0.00   Min.   : 0.000   Min.   :-31.0
##  1st Qu.: 294470   1st Qu.:14.00   1st Qu.: 7.000   1st Qu.: 16.0
##  Median : 591485   Median :29.00   Median :10.000   Median : 22.5
##  Mean   : 591071   Mean   :29.34   Mean   : 9.635   Mean   : 22.9
##  3rd Qu.: 889742   3rd Qu.:44.00   3rd Qu.:12.000   3rd Qu.: 29.5
##  Max.   :1180945   Max.   :59.00   Max.   :21.000   Max.   : 71.0
##                                                     NA's   :2127591
##   Ref_5x5_10th      Ref_5x5_50th      Ref_5x5_90th     RefComposite
##  Min.   :-32       Min.   :-32.0     Min.   :-28.5    Min.   :-32.0
##  1st Qu.: 14       1st Qu.: 16.0     1st Qu.: 18.0    1st Qu.: 17.5
##  Median : 20       Median : 22.5     Median : 25.5    Median : 24.0
##  Mean   : 20       Mean   : 22.6     Mean   : 25.9    Mean   : 24.7
##  3rd Qu.: 26       3rd Qu.: 29.0     3rd Qu.: 33.5    3rd Qu.: 31.5
##  Max.   : 62       Max.   : 69.0     Max.   : 72.5    Max.   : 92.5
##  NA's   :3192978   NA's   :2120484   NA's   :925685   NA's   :1760623
##  RefComposite_5x5_10th RefComposite_5x5_50th RefComposite_5x5_90th
##  Min.   :-31.0         Min.   :-27.5         Min.   :-25.0
##  1st Qu.: 16.0         1st Qu.: 17.5         1st Qu.: 19.5
##  Median : 22.0         Median : 24.0         Median : 27.0
##  Mean   : 22.2         Mean   : 24.4         Mean   : 27.4
##  3rd Qu.: 28.5         3rd Qu.: 31.5         3rd Qu.: 35.0
##  Max.   : 66.0         Max.   : 71.0         Max.   : 93.5
##  NA's   :2721293       NA's   :1765303       NA's   :647763
##      RhoHV         RhoHV_5x5_10th    RhoHV_5x5_50th    RhoHV_5x5_90th
##  Min.   :0         Min.   :0         Min.   :0         Min.   :0.2
##  1st Qu.:1         1st Qu.:1         1st Qu.:1         1st Qu.:1.0
##  Median :1         Median :1         Median :1         Median :1.0
##  Mean   :1         Mean   :1         Mean   :1         Mean   :1.0
##  3rd Qu.:1         3rd Qu.:1         3rd Qu.:1         3rd Qu.:1.1
##  Max.   :1         Max.   :1         Max.   :1         Max.   :1.1
##  NA's   :3542050   NA's   :4343812   NA's   :3540398   NA's   :2571382
##       Zdr           Zdr_5x5_10th      Zdr_5x5_50th      Zdr_5x5_90th
##  Min.   :-8        Min.   :-8        Min.   :-8        Min.   :-7.9
##  1st Qu.: 0        1st Qu.:-1        1st Qu.: 0        1st Qu.: 1.1
##  Median : 0        Median :-1        Median : 0        Median : 1.7
##  Mean   : 1        Mean   :-1        Mean   : 0        Mean   : 2.1
##  3rd Qu.: 1        3rd Qu.: 0        3rd Qu.: 1        3rd Qu.: 2.6
##  Max.   : 8        Max.   : 8        Max.   : 8        Max.   : 7.9
##  NA's   :3542050   NA's   :4343812   NA's   :3540398   NA's   :2571382
##       Kdp           Kdp_5x5_10th      Kdp_5x5_50th      Kdp_5x5_90th
##  Min.   :-96       Min.   :-81       Min.   :-79       Min.   :-100
##  1st Qu.: -1       1st Qu.: -5       1st Qu.: -1       1st Qu.:   2
##  Median :  0       Median : -3       Median :  0       Median :   4
##  Mean   :  0       Mean   : -3       Mean   :  0       Mean   :   4
##  3rd Qu.:  2       3rd Qu.: -2       3rd Qu.:  0       3rd Qu.:   6
##  Max.   :180       Max.   :  4       Max.   : 13       Max.   : 145
##  NA's   :4294331   NA's   :5048184   NA's   :4289685   NA's   :3424190
##     Expected
##  Min.   :    0.01
##  1st Qu.:    0.35
```

```
##  Median :    1.27
##  Mean   :   24.32
##  3rd Qu.:    3.56
##  Max.   :33017.73
##
```

```
##    user  system elapsed
##   26.40    2.26   34.66
```

```
train %>% group_by(Id) %>%
    summarise( V1 = mean(Expected)) %>%
    summarise( median(V1))
```

```
## Source: local data table [1 x 1]
##
##   median(V1)
##        (dbl)
## 1   1.016001
```

```
tcheck()
```

```
##    user  system elapsed
##    1.22    1.92   38.41
```

```
save( train, file="train.Rdata ");tcheck()
```

```
##    user  system elapsed
##   83.25    0.81  102.87
```

```
load( "train.Rdata"); tcheck()
```

```
##    user  system elapsed
##    7.51    0.42   14.24
```

```
summary(train); tcheck()   #benchmark operation
```

```
##        Id           minutes_past    radardist_km         Ref
##  Min.   :      2   Min.   : 0.00   Min.   : 0.000   Min.   :-31.0
##  1st Qu.: 294470   1st Qu.:14.00   1st Qu.: 7.000   1st Qu.: 16.0
##  Median : 591485   Median :29.00   Median :10.000   Median : 22.5
##  Mean   : 591071   Mean   :29.34   Mean   : 9.635   Mean   : 22.9
##  3rd Qu.: 889742   3rd Qu.:44.00   3rd Qu.:12.000   3rd Qu.: 29.5
##  Max.   :1180945   Max.   :59.00   Max.   :21.000   Max.   : 71.0
##                                                     NA's   :2127591
##   Ref_5x5_10th      Ref_5x5_50th      Ref_5x5_90th     RefComposite
##  Min.   :-32       Min.   :-32.0     Min.   :-28.5    Min.   :-32.0
##  1st Qu.: 14       1st Qu.: 16.0     1st Qu.: 18.0    1st Qu.: 17.5
##  Median : 20       Median : 22.5     Median : 25.5    Median : 24.0
##  Mean   : 20       Mean   : 22.6     Mean   : 25.9    Mean   : 24.7
##  3rd Qu.: 26       3rd Qu.: 29.0     3rd Qu.: 33.5    3rd Qu.: 31.5
```

```
##  Max.   : 62       Max.   : 69.0     Max.   : 72.5    Max.   : 92.5
##  NA's   :3192978   NA's   :2120484   NA's   :925685   NA's   :1760623
##  RefComposite_5x5_10th RefComposite_5x5_50th RefComposite_5x5_90th
##  Min.   :-31.0         Min.   :-27.5         Min.   :-25.0
##  1st Qu.: 16.0         1st Qu.: 17.5         1st Qu.: 19.5
##  Median : 22.0         Median : 24.0         Median : 27.0
##  Mean   : 22.2         Mean   : 24.4         Mean   : 27.4
##  3rd Qu.: 28.5         3rd Qu.: 31.5         3rd Qu.: 35.0
##  Max.   : 66.0         Max.   : 71.0         Max.   : 93.5
##  NA's   :2721293       NA's   :1765303       NA's   :647763
##      RhoHV         RhoHV_5x5_10th    RhoHV_5x5_50th    RhoHV_5x5_90th
##  Min.   :0         Min.   :0         Min.   :0         Min.   :0.2
##  1st Qu.:1         1st Qu.:1         1st Qu.:1         1st Qu.:1.0
##  Median :1         Median :1         Median :1         Median :1.0
##  Mean   :1         Mean   :1         Mean   :1         Mean   :1.0
##  3rd Qu.:1         3rd Qu.:1         3rd Qu.:1         3rd Qu.:1.1
##  Max.   :1         Max.   :1         Max.   :1         Max.   :1.1
##  NA's   :3542050   NA's   :4343812   NA's   :3540398   NA's   :2571382
##       Zdr           Zdr_5x5_10th      Zdr_5x5_50th      Zdr_5x5_90th
##  Min.   :-8        Min.   :-8        Min.   :-8        Min.   :-7.9
##  1st Qu.: 0        1st Qu.:-1        1st Qu.: 0        1st Qu.: 1.1
##  Median : 0        Median :-1        Median : 0        Median : 1.7
##  Mean   : 1        Mean   :-1        Mean   : 0        Mean   : 2.1
##  3rd Qu.: 1        3rd Qu.: 0        3rd Qu.: 1        3rd Qu.: 2.6
##  Max.   : 8        Max.   : 8        Max.   : 8        Max.   : 7.9
##  NA's   :3542050   NA's   :4343812   NA's   :3540398   NA's   :2571382
##       Kdp           Kdp_5x5_10th      Kdp_5x5_50th      Kdp_5x5_90th
##  Min.   :-96       Min.   :-81       Min.   :-79       Min.   :-100
##  1st Qu.: -1       1st Qu.: -5       1st Qu.: -1       1st Qu.:   2
##  Median :  0       Median : -3       Median :  0       Median :   4
##  Mean   :  0       Mean   : -3       Mean   :  0       Mean   :   4
##  3rd Qu.:  2       3rd Qu.: -2       3rd Qu.:  0       3rd Qu.:   6
##  Max.   :180       Max.   :  4       Max.   : 13       Max.   : 145
##  NA's   :4294331   NA's   :5048184   NA's   :4289685   NA's   :3424190
##     Expected
##  Min.   :    0.01
##  1st Qu.:    0.35
##  Median :    1.27
##  Mean   :   24.32
##  3rd Qu.:    3.56
##  Max.   :33017.73
##
```

```
##    user  system elapsed
##   26.49    1.97   30.50
```

```
train %>% group_by(Id) %>%
    summarise( V1 = mean(Expected)) %>%
    summarise( median(V1))
```

```
## Source: local data table [1 x 1]
##
##   median(V1)
```

```
##        (dbl)
## 1   1.016001
```

```
tcheck()
```

```
##    user  system elapsed
##    0.75    0.42    3.98
```

```
#total script time
tcheck(999)
```

```
##    user  system elapsed
##  309.25   32.26 3462.16
```

Discussion

data.table is faster than dplyer, which converts to a data.frame (I think), but its a lot less intuitive to work with (for me anyway). Removing the NA's reduces the size of the train data from 2.5 GB, to 1.5 GB and dplyr seems to perform okay on the latter.

On my PC (12GB on and AMD Athlon II X4 630 @2.8 GHz) working with the entire dataset eventually brings it to its knees. Removing the NA's makes the dataframe much more manageable and things like `summary` operations are much more manageable.

Recommendation

Run at least this part of the script one time

```
library(data.table)
library(dplyr)

#load the data
train <- fread("../train.csv")

#remove the lines that are all NA
object.size(train)
na_obs <- train %>%
    select( starts_with("Ref"), starts_with("Rho"), starts_with("Zdr"), starts_with("Kdp")) %>%
    .[, is.na(.SD)] %>%
    rowSums() == 20
tcheck()
sum(na_obs) / length(na_obs)
train <- train[ ! na_obs, ]
object.size(train)

save( train, file="train.Rdata ")
```

Then load the data into your scripts using

```
load( "train.Rdata"); tcheck()
```