**Introduction**

In this note we will show how to speed up work in R by partitioning data and process-level parallelization. We will show the technique with three different R packages: rqdatatable, data.table, and dplyr. The methods shown will also work with base-R and other packages.

For each of the above packages we speed up work by using wrapr::execute\_parallel which in turn uses wrapr::partition\_tables to partition un-related data.frame rows and then distributes them to different processors to be executed. rqdatatable::ex\_data\_table\_parallel conveniently bundles all of these steps together when working with rquery pipelines.

The partitioning is specified by the user preparing a grouping column that tells the system which sets of rows must be kept together in a correct calculation. We are going to try to demonstrate everything with simple code examples, and minimal discussion.

Keep in mind: unless the pipeline steps have non-trivial cost, the overhead of partitioning and distributing the work may overwhelm any parallel speedup. Also data.table itself already seems to exploit some thread-level parallelism (notice user time is greater than elapsed time). That being said, in this note we will demonstrate a synthetic example where computation is expensive due to a blow-up in an intermediate join step.

**Our example**

First we set up our execution environment and example (some details: OSX 10.13.4 on a 2.8 GHz Intel Core i5 Mac Mini (Late 2015 model) with 8GB RAM and hybrid disk drive).

library("rqdatatable")

## Loading required package: rquery

library("microbenchmark")

library("ggplot2")

library("WVPlots")

suppressPackageStartupMessages(library("dplyr"))

## Warning: package 'dplyr' was built under R version 3.5.1

base::date()

## [1] "Sun Jul 8 09:05:25 2018"

R.version.string

## [1] "R version 3.5.0 (2018-04-23)"

parallel::detectCores()

## [1] 4

packageVersion("parallel")

## [1] '3.5.0'

packageVersion("rqdatatable")

## [1] '0.1.2'

packageVersion("rquery")

## [1] '0.5.1'

packageVersion("dplyr")

## [1] '0.7.6'

ncore <- parallel::detectCores()

print(ncore)

## [1] 4

cl <- parallel::makeCluster(ncore)

print(cl)

## socket cluster with 4 nodes on host 'localhost'

set.seed(2362)

mk\_example <- function(nkey, nrep, ngroup = 20) {

keys <- paste0("key\_", seq\_len(nkey))

key\_group <- sample(as.character(seq\_len(ngroup)),

length(keys), replace = TRUE)

names(key\_group) <- keys

key\_table <- data.frame(

key = rep(keys, nrep),

stringsAsFactors = FALSE)

key\_table$data <- runif(nrow(key\_table))

instance\_table <- data.frame(

key = rep(keys, nrep),

stringsAsFactors = FALSE)

instance\_table$id <- seq\_len(nrow(instance\_table))

instance\_table$info <- runif(nrow(instance\_table))

# groups should be no finer than keys

key\_table$key\_group <- key\_group[key\_table$key]

instance\_table$key\_group <- key\_group[instance\_table$key]

list(key\_table = key\_table,

instance\_table = instance\_table)

}

dlist <- mk\_example(10, 10)

data <- dlist$instance\_table

annotation <- dlist$key\_table

**rquery / rqdatatable**

rquery and rqdatatable can implement a non-trivial calculation as follows.

Library(rquery)

Library(rqdatatable)

# possible data lookup: find rows that

# have lookup data <= info

optree <- local\_td(data) %.>%

natural\_join(.,

local\_td(annotation),

jointype = "INNER",

by = "key") %.>%

select\_rows\_nse(., data <= info) %.>%

pick\_top\_k(.,

k = 1,

partitionby = "id",

orderby = "data",

reverse = "data",

keep\_order\_column = FALSE) %.>%

orderby(., "id")

cat(format(optree))

## table('data';

## key,

## id,

## info,

## key\_group) %.>%

## natural\_join(.,

## table('annotation';

## key,

## data,

## key\_group),

## j= INNER, by= key) %.>%

## select\_rows(.,

## data <= info) %.>%

## extend(.,

## row\_number := row\_number(),

## p= id,

## o= "data" DESC) %.>%

## select\_rows(.,

## row\_number <= 1) %.>%

## drop\_columns(.,

## row\_number) %.>%

## orderby(., id)

res1 <- ex\_data\_table(optree)

head(res1)

## data id info key key\_group

## 1: 0.9152014 1 0.9860654 key\_1 20

## 2: 0.5599810 2 0.5857570 key\_2 8

## 3: 0.3011882 3 0.3334490 key\_3 10

## 4: 0.3650987 4 0.3960980 key\_4 5

## 5: 0.1469254 5 0.1753649 key\_5 14

## 6: 0.2567631 6 0.3510280 key\_6 7

nrow(res1)

## [1] 94

And we can execute the operations in parallel.

parallel::clusterEvalQ(cl,

library("rqdatatable"))

## [[1]]

## [1] "rqdatatable" "rquery" "stats" "graphics" "grDevices"

## [6] "utils" "datasets" "methods" "base"

##

## [[2]]

## [1] "rqdatatable" "rquery" "stats" "graphics" "grDevices"

## [6] "utils" "datasets" "methods" "base"

##

## [[3]]

## [1] "rqdatatable" "rquery" "stats" "graphics" "grDevices"

## [6] "utils" "datasets" "methods" "base"

##

## [[4]]

## [1] "rqdatatable" "rquery" "stats" "graphics" "grDevices"

## [6] "utils" "datasets" "methods" "base"

res2 <- ex\_data\_table\_parallel(optree,

"key\_group",

cl)

head(res2)

## data id info key key\_group

## 1: 0.9152014 1 0.9860654 key\_1 20

## 2: 0.5599810 2 0.5857570 key\_2 8

## 3: 0.3011882 3 0.3334490 key\_3 10

## 4: 0.3650987 4 0.3960980 key\_4 5

## 5: 0.1469254 5 0.1753649 key\_5 14

## 6: 0.2567631 6 0.3510280 key\_6 7

nrow(res2)

## [1] 94

**data.table**

data.table can implement the same function.

library("data.table")

##

## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':

##

## between, first, last

packageVersion("data.table")

## [1] '1.11.4'

data\_table\_f <- function(data, annotation) {

data <- data.table::as.data.table(data)

annotation <- data.table::as.data.table(annotation)

joined <- merge(data, annotation,

by = "key",

all=FALSE,

allow.cartesian=TRUE)

joined <- joined[joined$data <= joined$info, ]

data.table::setorderv(joined, cols = "data")

joined <- joined[, .SD[.N], id]

data.table::setorderv(joined, cols = "id")

}

resdt <- data\_table\_f(data, annotation)

head(resdt)

## id key info key\_group.x data key\_group.y

## 1: 1 key\_1 0.9860654 20 0.9152014 20

## 2: 2 key\_2 0.5857570 8 0.5599810 8

## 3: 3 key\_3 0.3334490 10 0.3011882 10

## 4: 4 key\_4 0.3960980 5 0.3650987 5

## 5: 5 key\_5 0.1753649 14 0.1469254 14

## 6: 6 key\_6 0.3510280 7 0.2567631 7

nrow(resdt)

## [1] 94

We can also run data.table in parallel using wrapr::execute\_parallel.

Library(wrapr)

parallel::clusterEvalQ(cl, library("data.table"))

## [[1]]

## [1] "data.table" "rqdatatable" "rquery" "stats" "graphics"

## [6] "grDevices" "utils" "datasets" "methods" "base"

##

## [[2]]

## [1] "data.table" "rqdatatable" "rquery" "stats" "graphics"

## [6] "grDevices" "utils" "datasets" "methods" "base"

##

## [[3]]

## [1] "data.table" "rqdatatable" "rquery" "stats" "graphics"

## [6] "grDevices" "utils" "datasets" "methods" "base"

##

## [[4]]

## [1] "data.table" "rqdatatable" "rquery" "stats" "graphics"

## [6] "grDevices" "utils" "datasets" "methods" "base"

parallel::clusterExport(cl, "data\_table\_f")

dt\_f <- function(tables\_list) {

data <- tables\_list$data

annotation <- tables\_list$annotation

data\_table\_f(data, annotation)

}

data\_table\_parallel\_f <- function(data, annotation) {

respdt <- wrapr::execute\_parallel(

tables = list(data = data,

annotation = annotation),

f = dt\_f,

partition\_column = "key\_group",

cl = cl) %.>%

data.table::rbindlist(.)

data.table::setorderv(respdt, cols = "id")

respdt

}

respdt <- data\_table\_parallel\_f(data, annotation)

head(respdt)

## id key info key\_group.x data key\_group.y

## 1: 1 key\_1 0.9860654 20 0.9152014 20

## 2: 2 key\_2 0.5857570 8 0.5599810 8

## 3: 3 key\_3 0.3334490 10 0.3011882 10

## 4: 4 key\_4 0.3960980 5 0.3650987 5

## 5: 5 key\_5 0.1753649 14 0.1469254 14

## 6: 6 key\_6 0.3510280 7 0.2567631 7

nrow(respdt)

## [1] 94

**dplyr**

dplyr can also implement the example.

Library(dplyr)

dplyr\_pipeline <- function(data, annotation) {

res <- data %>%

inner\_join(annotation, by = "key") %>%

filter(data <= info) %>%

group\_by(id) %>%

arrange(-data) %>%

mutate(rownum = row\_number()) %>%

ungroup() %>%

filter(rownum == 1) %>%

arrange(id)

res

}

resd <- dplyr\_pipeline(data, annotation)

head(resd)

## # A tibble: 6 x 7

## key id info key\_group.x data key\_group.y rownum

##

## 1 key\_1 1 0.986 20 0.915 20 1

## 2 key\_2 2 0.586 8 0.560 8 1

## 3 key\_3 3 0.333 10 0.301 10 1

## 4 key\_4 4 0.396 5 0.365 5 1

## 5 key\_5 5 0.175 14 0.147 14 1

## 6 key\_6 6 0.351 7 0.257 7 1

nrow(resd)

## [1] 94

And we can use [wrapr::execute\_parallel](https://winvector.github.io/wrapr/reference/execute_parallel.html) to parallelize the dplyr solution.

parallel::clusterEvalQ(cl, library("dplyr"))

## [[1]]

## [1] "dplyr" "data.table" "rqdatatable" "rquery" "stats"

## [6] "graphics" "grDevices" "utils" "datasets" "methods"

## [11] "base"

##

## [[2]]

## [1] "dplyr" "data.table" "rqdatatable" "rquery" "stats"

## [6] "graphics" "grDevices" "utils" "datasets" "methods"

## [11] "base"

##

## [[3]]

## [1] "dplyr" "data.table" "rqdatatable" "rquery" "stats"

## [6] "graphics" "grDevices" "utils" "datasets" "methods"

## [11] "base"

##

## [[4]]

## [1] "dplyr" "data.table" "rqdatatable" "rquery" "stats"

## [6] "graphics" "grDevices" "utils" "datasets" "methods"

## [11] "base"

parallel::clusterExport(cl, "dplyr\_pipeline")

dplyr\_f <- function(tables\_list) {

data <- tables\_list$data

annotation <- tables\_list$annotation

dplyr\_pipeline(data, annotation)

}

dplyr\_parallel\_f <- function(data, annotation) {

respdt <- wrapr::execute\_parallel(

tables = list(data = data,

annotation = annotation),

f = dplyr\_f,

partition\_column = "key\_group",

cl = cl) %>%

dplyr::bind\_rows() %>%

arrange(id)

}

respdplyr <- dplyr\_parallel\_f(data, annotation)

head(respdplyr)

## # A tibble: 6 x 7

## key id info key\_group.x data key\_group.y rownum

##

## 1 key\_1 1 0.986 20 0.915 20 1

## 2 key\_2 2 0.586 8 0.560 8 1

## 3 key\_3 3 0.333 10 0.301 10 1

## 4 key\_4 4 0.396 5 0.365 5 1

## 5 key\_5 5 0.175 14 0.147 14 1

## 6 key\_6 6 0.351 7 0.257 7 1

nrow(respdplyr)

## [1] 94

**Benchmark**

We can benchmark the various realizations.

dlist <- mk\_example(300, 300)

data <- dlist$instance\_table

annotation <- dlist$key\_table

timings <- microbenchmark(

data\_table\_parallel =

nrow(data\_table\_parallel\_f(data, annotation)),

data\_table = nrow(data\_table\_f(data, annotation)),

rqdatatable\_parallel =

nrow(ex\_data\_table\_parallel(optree, "key\_group", cl)),

rqdatatable = nrow(ex\_data\_table(optree)),

dplyr\_parallel =

nrow(dplyr\_parallel\_f(data, annotation)),

dplyr = nrow(dplyr\_pipeline(data, annotation)),

times = 10L)

saveRDS(timings, "Parallel\_rqdatatable\_timings.RDS")

**Conclusion**

print(timings)

## Unit: seconds

## expr min lq mean median uq

## data\_table\_parallel 5.274560 5.457105 5.609827 5.546554 5.686829

## data\_table 9.401677 9.496280 9.701807 9.541218 9.748159

## rqdatatable\_parallel 7.165216 7.497561 7.587663 7.563883 7.761987

## rqdatatable 12.490469 12.700474 13.320480 12.898154 14.229233

## dplyr\_parallel 6.492262 6.572062 6.784865 6.787277 6.875076

## dplyr 20.056555 20.450064 20.647073 20.564529 20.800350

## max neval

## 6.265888 10

## 10.419316 10

## 7.949404 10

## 14.282269 10

## 7.328223 10

## 21.332103 10

# autoplot(timings)

timings <- as.data.frame(timings)

timings$seconds <- timings$time/1e+9

ScatterBoxPlotH(timings,

xvar = "seconds", yvar = "expr",

title="task duration by method")

The benchmark timings show parallelized data.table is the fastest, followed by parallelized dplyr, and parallelized rqdatatable. In the non-paraellized case data.table is the fastest, followed by rqdatatable, and then dplyr.

A reason dplyr sees greater speedup relative to its own non-parallel implementation (yet does not beat data.table) is that data.table starts already multi-threaded, so data.table is exploiting some parallelism even before we added the process level parallelism (and hence sees less of a speed up, though it is fastest).

In this note we have demonstrated simple procedures to reliably parallelize any of rqdatatable, data.table, or dplyr.

Note: we did not include alternatives such as multidplyr or dtplyr in the timings, as they did not appear to work on this example.

Exhibit superior performance on Big Data Systems

Running on a C4.8xlarge EC2 intance.

Define and demonstrate pipelines:

DBI::dbGetQuery(my\_db, paste("SELECT COUNT(1) FROM", d$table\_name))

## count(1)

## 1 4e+05

length(column\_names(d))

## [1] 1003

sparklyr::sdf\_nrow(dT)

## [1] 4e+05

sparklyr::sdf\_ncol(dT)

## [1] 1003

scale <- 0.237

tng <- wrapr::mk\_tmp\_name\_source("tmptab")

rquery\_run <- function(collect) {

dq <- d %.>%

extend\_nse(.,

probability :=

exp(assessmentTotal \* scale)/

sum(exp(assessmentTotal \* scale)),

count := count(1),

partitionby = 'subjectID') %.>%

extend\_nse(.,

rank := rank(),

partitionby = 'subjectID',

orderby = c('probability', 'surveyCategory')) %.>%

rename\_columns(., 'diagnosis' := 'surveyCategory') %.>%

select\_rows\_nse(., rank == count) %.>%

select\_columns(., c('subjectID',

'diagnosis',

'probability')) %.>%

orderby(., 'subjectID')

sql <- to\_sql(dq, my\_db)

if(collect) {

dR <- DBI::dbGetQuery(my\_db, sql)

} else {

# count and throw away the data

sql <- paste("SELECT COUNT(1) FROM (", sql, ") ctab")

dR <- DBI::dbGetQuery(my\_db, sql)

}

dR

}

dplyr\_run <- function(narrow, collect) {

dR <- dT

if(narrow) {

dR <- dR %>%

select(subjectID, surveyCategory, assessmentTotal)

}

dR <- dR %>%

group\_by(subjectID) %>%

mutate(probability =

exp(assessmentTotal \* scale)/

sum(exp(assessmentTotal \* scale), na.rm = TRUE)) %>%

arrange(probability, surveyCategory) %>%

filter(row\_number() == n()) %>%

ungroup() %>%

rename(diagnosis = surveyCategory) %>%

select(subjectID, diagnosis, probability) %>%

arrange(subjectID)

if(collect) {

dR <- collect(dR)

} else {

# count and throw away the data (force calculation)

dR <- as.data.frame(tally(dR))

}

dR

}

head(rquery\_run(collect=TRUE))

## subjectID diagnosis probability

## 1 1 withdrawal behavior 0.6163301

## 2 2 positive re-framing 0.6163301

## 3 3 withdrawal behavior 0.8694381

## 4 4 withdrawal behavior 0.5589742

## 5 5 positive re-framing 0.7207128

## 6 6 withdrawal behavior 0.5000000

rquery\_run(collect=FALSE)

## count(1)

## 1 2e+05

head(dplyr\_run(narrow=FALSE, collect=TRUE))

## # A tibble: 6 x 3

## subjectID diagnosis probability

## <int> <chr> <dbl>

## 1 1 withdrawal behavior 0.616

## 2 2 positive re-framing 0.616

## 3 3 withdrawal behavior 0.869

## 4 4 withdrawal behavior 0.559

## 5 5 positive re-framing 0.721

## 6 6 withdrawal behavior 0.500

dplyr\_run(narrow=FALSE, collect=FALSE)

## n

## 1 2e+05

head(dplyr\_run(narrow=TRUE, collect=TRUE))

## # A tibble: 6 x 3

## subjectID diagnosis probability

## <int> <chr> <dbl>

## 1 1 withdrawal behavior 0.616

## 2 2 positive re-framing 0.616

## 3 3 withdrawal behavior 0.869

## 4 4 withdrawal behavior 0.559

## 5 5 positive re-framing 0.721

## 6 6 withdrawal behavior 0.500

dplyr\_run(narrow=TRUE, collect=FALSE)

## n

## 1 2e+05

Get timings:

timings <- microbenchmark(rquery\_run(collect=TRUE),

rquery\_run(collect=FALSE),

dplyr\_run(narrow=FALSE, collect=TRUE),

dplyr\_run(narrow=FALSE, collect=FALSE),

dplyr\_run(narrow=TRUE, collect=TRUE),

dplyr\_run(narrow=TRUE, collect=FALSE))

saveRDS(timings, "PerfTest\_timings.RDS")

Present results:

print(timings)

## Unit: milliseconds

## expr min lq mean

## rquery\_run(collect = TRUE) 833.7267 929.3922 974.8569

## rquery\_run(collect = FALSE) 669.4242 734.8816 774.6747

## dplyr\_run(narrow = FALSE, collect = TRUE) 2437.2860 2561.7170 2603.2164

## dplyr\_run(narrow = FALSE, collect = FALSE) 2295.2626 2399.2467 2467.7334

## dplyr\_run(narrow = TRUE, collect = TRUE) 1205.1181 1276.6383 1333.4385

## dplyr\_run(narrow = TRUE, collect = FALSE) 1027.0068 1114.8148 1173.5855

## median uq max neval

## 976.3468 1004.0608 1537.182 100

## 762.9225 791.5005 1169.712 100

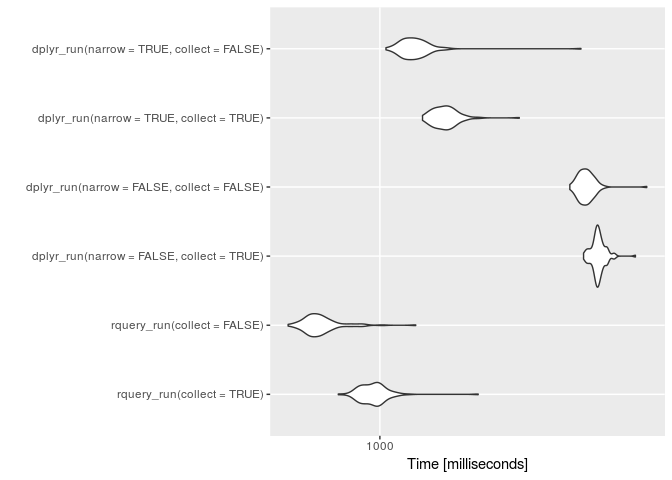
## 2593.1621 2638.7540 3058.012 100

## 2457.8133 2514.6530 3213.354 100

## 1328.4738 1374.1824 1839.873 100

## 1159.7749 1204.2670 2409.451 100

autoplot(timings)

[](https://github.com/WinVector/rquery/blob/master/extras/PerfTest_files/figure-markdown_github/present-1.png)

tdf <- as.data.frame(timings)

# order the data

tdf <- tdf %.>%

group\_by\_se(., "expr") %.>%

mutate\_se(., qae(mtime := median(time))) %.>%

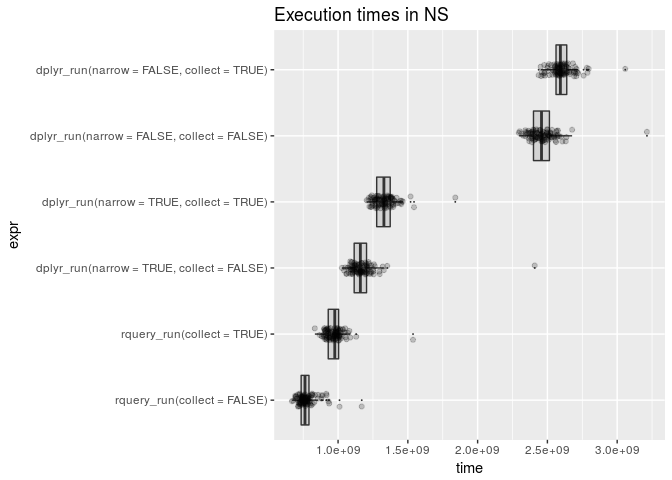
ungroup(.)

tdf$expr <- reorder(tdf$expr, tdf$mtime)

WVPlots::ScatterBoxPlotH(tdf, "time", "expr",

pt\_alpha=0.2,

title="Execution times in NS")

[](https://github.com/WinVector/rquery/blob/master/extras/PerfTest_files/figure-markdown_github/present-2.png)

sparklyr::spark\_disconnect(my\_db)