# VerificationOCT.R

*arcs*

*Thu Nov 30 13:59:46 2017*

```r
#library(ggplot2)
#library(scales)
library(data.table)
setwd("/home/arcs/Oct14/DataCSV")
getwd()
```

```
## [1] "/home/arcs/Oct14/DataCSV"
```

```r
data_web <- fread("OctVerification.csv")
data_condor <- fread("Oct2017Efficiency_VO.csv")
```

```
##
Read 90.2% of 5876000 rows
Read 5876000 rows and 8 (of 8) columns from 0.193 GB file in 00:00:03
```

```r
###################################################################
############# Studying the structure of Data #####################
###################################################################
names(data_web)
```

```
##  [1] "Site"                             "Year"
##  [3] "Month"                            "Resource"
##  [5] "VO"                               "Project Type"
##  [7] "VORole"                           "Infrastructure"
##  [9] "Number of Cores"                  "CPU Duration (d)"
## [11] "Wall Duration (d)"                "Quota (d)"
## [13] "Normalised CPU Duration (hs06d)"  "Normalised Wall Duration (hs06d)"
## [15] "Normalised Quota (hs06d)"         "Avg. Daily Wall Duration"
## [17] "Avg. Daily Quota"                 "Number of Jobs"
## [19] "Notes"
```

```r
str(data_web)
```

```
## Classes 'data.table' and 'data.frame':   268 obs. of  19 variables:
##  $ Site                            : chr  "CERN-PROD" "CERN-PROD" "CERN-PROD" "CERN-PROD" ...
##  $ Year                            : chr  "2017" "2017" "2017" "2017" ...
##  $ Month                           : chr  "10" "10" "10" "10" ...
##  $ Resource                        : chr  "lsf" "lsf" "lsf" "lsf" ...
##  $ VO                              : chr  "wa105" "va" "va" "totem" ...
##  $ Project Type                    : chr  "null" "null" "null" "null" ...
##  $ VORole                          : chr  "" "" "" "" ...
##  $ Infrastructure                  : chr  "local" "local" "local" "local" ...
##  $ Number of Cores                 : chr  "1" "4" "1" "1" ...
##  $ CPU Duration (d)                : chr  "12.35" "244.05" "25484.41" "40.83" ...
##  $ Wall Duration (d)               : chr  "23.00" "61.00" "32833" "154.00" ...
##  $ Quota (d)                       : chr  "null" "null" "null" "null" ...
##  $ Normalised CPU Duration (hs06d) : chr  "117.14" "2352.2" "250474.04" "387.57" ...
##  $ Normalised Wall Duration (hs06d): chr  "227.37" "2353.54" "323055.86" "1462.17" ...
##  $ Normalised Quota (hs06d)        : chr  "null" "null" "null" "null" ...
##  $ Avg. Daily Wall Duration        : chr  "0.00" "1.00" "1059" "4.00" ...
##  $ Avg. Daily Quota                : chr  "null" "null" "null" "null" ...
##  $ Number of Jobs                  : chr  "1414" "110.00" "700299" "12914" ...
##  $ Notes                           : chr  "" "" "" "" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
summary(data_web)
```

```
##      Site               Year              Month          
##  Length:268         Length:268         Length:268        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##    Resource              VO            Project Type      
##  Length:268         Length:268         Length:268        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##     VORole          Infrastructure     Number of Cores   
##  Length:268         Length:268         Length:268        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  CPU Duration (d)   Wall Duration (d)   Quota (d)        
##  Length:268         Length:268         Length:268        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  Normalised CPU Duration (hs06d) Normalised Wall Duration (hs06d)
##  Length:268                      Length:268                      
##  Class :character                Class :character                
##  Mode  :character                Mode  :character                
##  Normalised Quota (hs06d) Avg. Daily Wall Duration Avg. Daily Quota  
##  Length:268               Length:268               Length:268        
##  Class :character         Class :character         Class :character  
##  Mode  :character         Mode  :character         Mode  :character  
##  Number of Jobs        Notes          
##  Length:268         Length:268        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character
```

```r
unique(data_web$Resource)
```

```
## [1] "lsf"    "condor" "cloud"
```

```r
data_web <- subset(data_web, Resource == "condor")
unique(data_web$VO)
```

```
##  [1] "vo.compass.cern.ch"   "theory"               "te"                  
##  [4] "ntof"                 "np04"                 "np02"                
##  [7] "next"                 "na62.vo.gridpp.ac.uk" "na62"                
## [10] "na61"                 "lhcb"                 "it"                  
## [13] "ilc"                  "geant"                "fcc"                 
## [16] "dteam"                "default"              "compass"             
## [19] "cms"                  "be"                   "atlas"               
## [22] "ams"                  "alpha"                "alice"
```

```r
alice_web <- subset(data_web, VO == "alice")



names(data_condor)
```

```
## [1] "RequestCpus"         "MATCH_HEPSPEC"       "MATCH_TotalCpus"
## [4] "RemoteWallClockTime" "ExitCode"            "RemoteSysCpu"
## [7] "RemoteUserCpu"       "x509UserProxyVOName"
```

```
str(data_condor)
```

```
## Classes 'data.table' and 'data.frame':   5876000 obs. of  8 variables:
##  $ RequestCpus        : int  8 8 8 8 8 8 8 1 1 8 ...
##  $ MATCH_HEPSPEC      : chr  "None" "None" "None" "None" ...
##  $ MATCH_TotalCpus    : chr  "None" "None" "None" "None" ...
##  $ RemoteWallClockTime: chr  "None" "None" "None" "None" ...
##  $ ExitCode           : chr  "None" "None" "None" "None" ...
##  $ RemoteSysCpu       : int  0 0 0 0 0 0 0 97 182 25311 ...
##  $ RemoteUserCpu      : int  0 0 0 0 0 0 0 49122 663 1323662 ...
##  $ x509UserProxyVOName: chr  "cms" "cms" "cms" "cms" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(data_condor)
```

```
##   RequestCpus    MATCH_HEPSPEC      MATCH_TotalCpus    RemoteWallClockTime
##  Min.   :1.000   Length:5876000     Length:5876000     Length:5876000
##  1st Qu.:1.000   Class :character   Class :character   Class :character
##  Median :1.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2.018
##  3rd Qu.:1.000
##  Max.   :8.000
##    ExitCode          RemoteSysCpu      RemoteUserCpu
##  Length:5876000     Min.   :     0.0   Min.   :      0
##  Class :character   1st Qu.:     0.0   1st Qu.:      2
##  Mode  :character   Median :     2.0   Median :      5
##                     Mean   :   294.6   Mean   :  15690
##                     3rd Qu.:   110.0   3rd Qu.:   9335
##                     Max.   :298748.0   Max.   :1989119
##  x509UserProxyVOName
##  Length:5876000
##  Class :character
##  Mode  :character
##
##
##
```

```
unique(data_condor$x509UserProxyVOName)
```

```
## [1] "cms"                "atlas"              "vo.compass.cern.ch"
## [4] "lhcb"               "ilc"                "alice"
## [7] "None"
```

```
data_condor <- subset(data_condor, ExitCode == "0")
alice_hdfs <- subset(data_condor, data_condor$x509UserProxyVOName == "alice")
unique(data_condor$x509UserProxyVOName)
```

```
## [1] "atlas"              "cms"                "vo.compass.cern.ch"
## [4] "lhcb"               "ilc"                "alice"
## [7] "None"
```

```
###################################################################
############# Conversion to numeric values ########################
###################################################################
alice_web$NCPU <- as.numeric(unlist(alice_web[, "Normalised CPU Duration (hs06d)"]))
alice_web$NWall <- as.numeric(unlist(alice_web[, "Normalised Wall Duration (hs06d)"]))
TotalCPU_web <- sum(alice_web$NCPU)
TotalWall_web <- sum(alice_web$NWall)

Efficiency_web <- TotalCPU_web/TotalWall_web



alice_hdfs[,"RemoteWallClockTime"] <- as.numeric(unlist(alice_hdfs[,"RemoteWallClockTime"])) #RemoteWal
alice_hdfs[, "ExitCode"] <- as.numeric(unlist(alice_hdfs[, "ExitCode"]))
alice_hdfs[, "MATCH_HEPSPEC"] <- as.numeric(unlist(alice_hdfs[, "MATCH_HEPSPEC"]))
alice_hdfs[, "MATCH_TotalCpus"] <- as.numeric(unlist(alice_hdfs[, "MATCH_TotalCpus"]))


###################################################################
##################### Data Cleansing ##############################
###################################################################



#alice_hdfs <- subset(alice_hdfs, alice_hdfs$CPUTime > 0)
#alice_hdfs <- subset(alice_hdfs, alice_hdfs$WallTime > 0) # Removing the failed Jobs

str(alice_hdfs)
```

```
## Classes 'data.table' and 'data.frame':   16135 obs. of  8 variables:
##  $ RequestCpus        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MATCH_HEPSPEC      : num  35 104 104 104 104 ...
##  $ MATCH_TotalCpus    : num  4 10 10 10 10 8 8 8 8 12 ...
##  $ RemoteWallClockTime: num  158 2 2 2 2 2 14 11 2 6 ...
##  $ ExitCode           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RemoteSysCpu       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RemoteUserCpu      : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ x509UserProxyVOName: chr  "alice" "alice" "alice" "alice" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
alice_hdfs <- na.omit(alice_hdfs)


alice_hdfs$CPUTime <- alice_hdfs$RemoteSysCpu + alice_hdfs$RemoteUserCpu
alice_hdfs$WallTime <- alice_hdfs$RemoteWallClockTime #- alice_hdfs2$CumulativeSuspensionTime
str(alice_hdfs)
```

```
## Classes 'data.table' and 'data.frame':   16135 obs. of  10 variables:
##  $ RequestCpus        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MATCH_HEPSPEC      : num  35 104 104 104 104 ...
##  $ MATCH_TotalCpus    : num  4 10 10 10 10 8 8 8 8 12 ...
##  $ RemoteWallClockTime: num  158 2 2 2 2 2 14 11 2 6 ...
##  $ ExitCode           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RemoteSysCpu       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RemoteUserCpu      : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ x509UserProxyVOName: chr  "alice" "alice" "alice" "alice" ...
##  $ CPUTime            : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ WallTime           : num  158 2 2 2 2 2 14 11 2 6 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
alice_hdfs$HEPSPEC_TotalCpus <- alice_hdfs$MATCH_HEPSPEC/ alice_hdfs$MATCH_TotalCpus
alice_hdfs$NWallTime <- alice_hdfs$WallTime * alice_hdfs$RequestCpus * alice_hdfs$HEPSPEC_TotalCpus
alice_hdfs$NCPUTime <- alice_hdfs$CPUTime * alice_hdfs$HEPSPEC_TotalCpus
TotalWallTime_hdfs <- sum(alice_hdfs$NWallTime)
TotalCPUTime_hdfs <- sum(alice_hdfs$NCPUTime)
Efficiency_hdfs <- TotalCPUTime_hdfs/TotalWallTime_hdfs
```