DataDiscrepancy.R

arcs

Tue Dec 5 14:22:14 2017

```
############## Aim of this program is to check if number of jobs in HDFS and S3 #######################
##############          system match for the month of november                  #######################
##############                  as it did'nt match for Oct                      #######################

library(data.table)
setwd("/home/arcs/Oct14/DataCSV")
getwd()
```

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

```
data_web <- fread("NovWeb.csv")
data_hdfs <- fread(input = "Nov2017Efficiency_VO.csv", sep = ",", fill = TRUE)

############### Function to print values #########################
printf <- function(...) cat(sprintf(...))

###################################################################
############# 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"
```

```
str(data_web)
```

```
## Classes 'data.table' and 'data.frame':   245 obs. of  19 variables:
##  $ Site                            : chr  "CERN-PROD" "CERN-PROD" "CERN-PROD" "CERN-PROD" ...
##  $ Year                            : chr  "2017" "2017" "2017" "2017" ...
##  $ Month                           : chr  "11" "11" "11" "11" ...
##  $ 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  "15.33" "70.41" "29627.47" "39.31" ...
##  $ Wall Duration (d)               : chr  "28.00" "27.00" "37487" "226.00" ...
##  $ Quota (d)                       : chr  "null" "null" "null" "null" ...
##  $ Normalised CPU Duration (hs06d) : chr  "144.17" "680.10" "293813.13" "369.45" ...
##  $ Normalised Wall Duration (hs06d): chr  "277.63" "1092.32" "371961.02" "2160.79" ...
##  $ Normalised Quota (hs06d)        : chr  "null" "null" "null" "null" ...
##  $ Avg. Daily Wall Duration        : chr  "0.00" "0.00" "1249" "7.00" ...
##  $ Avg. Daily Quota                : chr  "null" "null" "null" "null" ...
##  $ Number of Jobs                  : chr  "1467" "29.00" "688835" "66374" ...
##  $ Notes                           : chr  "" "" "" "" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
summary(data_web)
```

```
##      Site               Year              Month          
##  Length:245         Length:245         Length:245        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##    Resource              VO            Project Type      
##  Length:245         Length:245         Length:245        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##     VORole          Infrastructure     Number of Cores   
##  Length:245         Length:245         Length:245        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  CPU Duration (d)   Wall Duration (d)   Quota (d)        
##  Length:245         Length:245         Length:245        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  Normalised CPU Duration (hs06d) Normalised Wall Duration (hs06d)
##  Length:245                      Length:245                      
##  Class :character                Class :character                
##  Mode  :character                Mode  :character                
##  Normalised Quota (hs06d) Avg. Daily Wall Duration Avg. Daily Quota  
##  Length:245               Length:245               Length:245        
##  Class :character         Class :character         Class :character  
##  Mode  :character         Mode  :character         Mode  :character  
##  Number of Jobs        Notes          
##  Length:245         Length:245        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character
```

```r
unique(data_web$Resource) # Tocheck the types of resources
```

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

```r
###################################################################
############# Conversion to numeric values ########################
###################################################################
data_web$`Number of Jobs` <- as.numeric(unlist(data_web[, data_web$`Number of Jobs`]))
```

```
## Warning: NAs introduced by coercion
```

```r
summary(data_web)
```

```
##      Site               Year              Month          
##  Length:245         Length:245         Length:245        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
```

```
##
##
##    Resource              VO            Project Type
##  Length:245         Length:245         Length:245
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     VORole          Infrastructure     Number of Cores
##  Length:245         Length:245         Length:245
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  CPU Duration (d)   Wall Duration (d)   Quota (d)
##  Length:245         Length:245         Length:245
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Normalised CPU Duration (hs06d) Normalised Wall Duration (hs06d)
##  Length:245                      Length:245
##  Class :character                Class :character
##  Mode  :character                Mode  :character
##
##
##
##
##  Normalised Quota (hs06d) Avg. Daily Wall Duration Avg. Daily Quota
##  Length:245               Length:245               Length:245
##  Class :character         Class :character         Class :character
##  Mode  :character         Mode  :character         Mode  :character
##
##
##
##
##  Number of Jobs       Notes
##  Min.   :      1   Length:245
##  1st Qu.:     45   Class :character
##  Median :   2496   Mode  :character
##  Mean   : 185836
##  3rd Qu.:  65556
##  Max.   :5361630
##  NA's   :85
```

```
###################################################################
############# Removing jobs with NA in      #######################
#############     Particular Col           ########################
```

```
###################################################################
data_web <- data_web[!is.na(data_web$`Number of Jobs`), ]


###################################################################
############# Comparing jobs from HDFS and Web data ###############
###################################################################

############ To check for the particular Month ####################

printf("\n Month of evaluation: %s", unique(data_web$Month))
```

```
##
##  Month of evaluation: 11
```

```
printf("\nTotal no of jobs from website: %s", sum(data_web$`Number of Jobs`))
```

```
##
## Total no of jobs from website: 29733752
```

```
data_lsf <- subset(data_web, Resource == "lsf")
printf("\nNo of lsf jobs from website: %s", sum(data_lsf$`Number of Jobs`))
```

```
##
## No of lsf jobs from website: 13746785
```

```
data_cloud <- subset(data_web, Resource == "cloud")
printf("\nNo of cloud jobs from website: %s", sum(data_cloud$`Number of Jobs`))
```

```
##
## No of cloud jobs from website: 0
```

```
data_condor <- subset(data_web, Resource == "condor")
web_condor_jobs = sum(data_web$`Number of Jobs`)
printf("\nNo of Condor jobs from website: %s", sum(data_condor$`Number of Jobs`))
```

```
##
## No of Condor jobs from website: 15986967
```

```
unique(data_condor$Infrastructure)
```

```
## [1] "grid"  "local"
```

```
web_condor_grid <- subset(data_condor, data_condor$Infrastructure == "grid")
printf("\nNo of Condor:grid jobs from website: %s", sum(web_condor_grid$`Number of Jobs`))
```

```
##
## No of Condor:grid jobs from website: 11858455
```

```
web_condor_local <- subset(data_condor, data_condor$Infrastructure == "local")
printf("\nNo of Condor:grid jobs from website: %s", sum(web_condor_local$`Number of Jobs`))
```

```
##
## No of Condor:grid jobs from website: 4128512
```

```
hdfs_condor_jobs = nrow(data_hdfs)
printf("\nTotal no of jobs from HDFS: %d", nrow(data_hdfs))
```

```
##
## Total no of jobs from HDFS: 3788263
```

```
diff = web_condor_jobs - hdfs_condor_jobs
printf("\nNo of missing jobs in HDFS System: %d", diff)
```

```
##
## No of missing jobs in HDFS System: 25945489
```

```
################### To check if all VOs are captured ######################
unique(data_web$VO)
```

```
##  [1] "wa105"                "va"                   "totem"
##  [4] "theory"               "sldiv"                "ship"
##  [7] "rd51"                 "parc"                 "ops"
## [10] "ntof"                 "nestor"               "na61"
## [13] "na49"                 "na48"                 "lhcbt3"
## [16] "lhcb"                 "itdc"                 "isolde"
## [19] "ilc"                  "harp"                 "geant4"
## [22] "engpara"              "delphi"               "default"
## [25] "compass"              "cmst3"                "cmsphys"
## [28] "cmscomm"              "cmsalca"              "cms"
## [31] "cast"                 "c3"                   "atlaswisc"
## [34] "atlas"                "amsprod"              "amsp"
## [37] "ams"                  "alice"                "vo.compass.cern.ch"
## [40] "te"                   "re18"                 "np04"
## [43] "np02"                 "next"                 "na62.vo.gridpp.ac.uk"
## [46] "na62"                 "it"                   "geant"
## [49] "fcc"                  "dune"                 "dteam"
## [52] "be"                   "alpha"
```

```
unique(data_hdfs$x509UserProxyVOName)
```

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

```
VO = unique(data_hdfs$x509UserProxyVOName)

for (vo in VO){
  printf("\n\n\n************ VO Name: %s ***************\n", vo)
  sub_Data <- subset(data_hdfs, x509UserProxyVOName == vo)
  printf("\nNumber of observation from HDFS: %d", nrow(sub_Data))
  sub_Data_web <- subset(data_condor, data_condor$VO == vo)
  printf("\nNumber of observation from Website: %d", sum(sub_Data_web$`Number of Jobs`))
}
```

```
##
##
##
## ************ VO Name: cms ***************
##
## Number of observation from HDFS: 273767
## Number of observation from Website: 1613805
##
##
## ************ VO Name: atlas ***************
##
## Number of observation from HDFS: 940922
## Number of observation from Website: 3059229
##
##
## ************ VO Name: lhcb ***************
##
## Number of observation from HDFS: 84424
## Number of observation from Website: 528939
##
##
## ************ VO Name: vo.compass.cern.ch ***************
##
## Number of observation from HDFS: 752332
## Number of observation from Website: 2075277
##
##
## ************ VO Name: ilc ***************
##
## Number of observation from HDFS: 21756
## Number of observation from Website: 73446
##
##
## ************ VO Name: alice ***************
##
## Number of observation from HDFS: 1714996
## Number of observation from Website: 5376366
##
##
## ************ VO Name: None ***************
##
## Number of observation from HDFS: 11
## Number of observation from Website: 0
##
##
## ************ VO Name: dune ***************
##
## Number of observation from HDFS: 54
## Number of observation from Website: 97
##
##
## ************ VO Name:  ***************
##
## Number of observation from HDFS: 1
## Number of observation from Website: 0
```