DataDiscrepancy.R

arcs

Tue Dec 5 14:25:56 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("OctVerification.csv")
data_hdfs <- fread(input = "Oct2017Efficiency_VO.csv", sep = ",", fill = TRUE)
```

```
##
Read 80.7% of 5876000 rows
Read 5876000 rows and 8 (of 8) columns from 0.193 GB file in 00:00:03
```

```
############### 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':   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>
```

```
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
```

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

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

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

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

```
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
##  Min.   :      1   Length:268
##  1st Qu.:     29   Class :character
##  Median :   2180   Mode  :character
##  Mean   : 133876
##  3rd Qu.:  27631
##  Max.   :3610994
##  NA's   :90
```

```
#########################################################################
############## 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: 10
```

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

```
##
## Total no of jobs from website: 23829919
```

```
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: 13067996
```

```
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: 10761923
```

```
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: 6717223
```

```
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: 4044700
```

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

```
##
## Total no of jobs from HDFS: 5876000
```

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

```
##
## No of missing jobs in HDFS System: 17953919
```

```
#################### 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] "na48"                 "lhcbt3"               "lhcb"
## [16] "l3"                   "isolde"               "ilc"
## [19] "harp"                 "geant4"               "engpara"
## [22] "dirac"                "delphi"               "default"
## [25] "dcms"                 "compass"              "cmst3"
## [28] "cmsphys"              "cmscomm"              "cmsalca"
## [31] "cms"                  "cast"                 "c3"
## [34] "atlaswisc"            "atlas"                "amsprod"
## [37] "amsp"                 "ams"                  "alice"
## [40] "vo.compass.cern.ch"   "te"                   "np04"
## [43] "np02"                 "next"                 "na62.vo.gridpp.ac.uk"
## [46] "na62"                 "it"                   "geant"
## [49] "fcc"                  "dteam"                "be"
## [52] "alpha"
```

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

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

```
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: 728786
## Number of observation from Website: 1707094
```

```
##
##
## ************ VO Name: atlas ***************
##
## Number of observation from HDFS: 1579459
## Number of observation from Website: 1891599
##
##
## ************ VO Name: vo.compass.cern.ch ***************
##
## Number of observation from HDFS: 1766354
## Number of observation from Website: 1983139
##
##
## ************ VO Name: lhcb ***************
##
## Number of observation from HDFS: 239651
## Number of observation from Website: 404512
##
##
## ************ VO Name: ilc ***************
##
## Number of observation from HDFS: 134236
## Number of observation from Website: 149640
##
##
## ************ VO Name: alice ***************
##
## Number of observation from HDFS: 1427509
## Number of observation from Website: 1663968
##
##
## ************ VO Name: None ***************
##
## Number of observation from HDFS: 5
## Number of observation from Website: 0
```