databricks records-link-example

```
#Attaching library for SparkR
library(SparkR)
library(data.table)
```

```
#Access the file system to see where all the files are stored
```

```
%fs ls FileStore/tables/9oyswmfm1505227191899/
```

path
dbfs:/FileStore/tables/9oyswmfm1505227191899/masterfile.csv

```
#Loading and looking at the masterfile
masterdf  <- read.df("FileStore/tables/9oyswmfm1505227191899/masterfile.csv",
                    source = "csv", header="true", inferSchema = "true")
str(masterdf)
```

```
'SparkDataFrame': 17 variables:
 $ GUID         : chr "dfc42b79-ac61-46c6-af2a-b08c5219dc36" "eedfeb26-5f83-
416f-923f-db0fae18fd78" "26f9e8da-45f6-4d72-
 $ NationalID   : chr "445-44-3892" "206-80-0666" "528-56-4973" "655-05-977
6" "306-11-6350" "647-20-1250"
 $ Gender       : chr "male" "female" "female" "female" "male" "male"
 $ Title        : chr "Mr." "Mrs." "Mrs." "Ms." "Mr." "Mr."
 $ GivenName    : chr "Raymond" "Sharon" "Misty" "Cornelia" "James" "Michae
l"
 $ MiddleInitial: chr "E" "C" "C" "A" "M" "E"
 $ Surname      : chr "Cole" "Hall" "Shook" "Nixon" "Walton" "Boyer"
 $ StreetAddress: chr "2233 Luke Lane" "1978 Tree Top Lane" "515 Austin Secr
et Lane" "1700 Emily Drive" "351 Duffy Stree
 $ City         : chr "Waurika" "Philadelphia" "Helper" "Columbia" "Michigan
City" "Layton"
 $ State        : chr "OK" "PA" "UT" "SC" "IN" "UT"
 $ ZipCode      : int 73573 19108 84526 29210 46360 84041
 $ Country      : chr "US" "US" "US" "US" "US" "US"
 $ Birthday     : chr "8/20/1959" "4/9/1951" "9/25/1946" "11/6/1934" "9/12/1
963" "10/30/1933"
 $ Age          : int 58 66 70 82 53 83
 $ Occupation   : chr "Pesticide vegetation" "Chemical technician" "Counter
 clerk" "Textile knitting and weaving machine
 $ Latitude     : num 34.276111 40.02431 39.70902 34.009534 41.599893 40.993
244
 $ Longitude    : num -97.953709 -75.245753 -111.014542 -81.133773 -86.77387
1 -111.943678
```

```
#Loading and looking at the linkfile
linkdf  <- read.df("FileStore/tables/6ov8poz81505268147137/linkfile.csv",
                   source = "csv", header="true", inferSchema = "true")
str(linkdf)
```

```
'SparkDataFrame': 10 variables:
 $ GivenName      : chr "James" "Renee" "Brian" "Fausto" "William" "Lois"
 $ MiddleInitial  : chr "S" "P" "G" "C" "A" "A"
 $ Surname        : chr "Sanroman" "Williams" "Iliff" "Larson" "Heaton" "Nai
l"
 $ StreetAddress  : chr "1381 Hardesty Street" "2251 Foley Street" "4705 Ham
pton Meadows" "4142 Stewart Street" "1934 Ro
 $ City           : chr "Albany" "Plantation" "Cambridge" "Indianapolis" "De
nver" "Chicago"
 $ State          : chr "NY" "FL" "MA" "IN" "CO" "IL"
 $ ZipCode        : int 12207 33324 2141 46241 80265 60620
 $ EmailAddress   : chr "JamesSSanroman@einrot.com" "ReneePWilliams@telewor
m.us" "BrianGIliff@jourrapide.com" "FaustoCLa
 $ TelephoneNumber: chr "518-427-5608" "954-693-8661" "978-319-5715" "317-66
0-1527" "303-945-7270" "773-723-3862"
 $ Birthday       : chr "4/8/1986" "9/19/1966" "9/13/1959" "5/11/1932" "8/3
0/1945" "11/13/1984"
```

```
#Checking to see if there are any duplicates in our Master File
masterdf <- as.data.frame(masterdf)
masterdf[duplicated(masterdf),4:9]
```

```
      Title GivenName MiddleInitial  Surname          StreetAddress        C
ity
50003   Ms.    Carmen           SanDiego  4600 Sliver Hill Road   Sweetland
50004   Mr.     Waldo             W    Waldo 1600 Candy Stripe Lane Picturel
and
50005   Ms.    Carmen           SanDiego  4600 Sliver Hill Road   Sweetland
50006   Mr.     Waldo             W    Waldo 1600 Candy Stripe Lane Picturel
and
50007   Ms.    Carmen           SanDiego  4600 Sliver Hill Road   Sweetland
50008   Mr.     Waldo             W    Waldo 1600 Candy Stripe Lane Picturel
and
50009   Ms.    Carmen           SanDiego  4600 Sliver Hill Road   Sweetland
50010   Mr.     Waldo             W    Waldo 1600 Candy Stripe Lane Picturel
and
```

```
#Remove duplicates
masterdf<- masterdf[!duplicated(masterdf),]
print("duplicates removed")
print("Here's the proof:")
masterdf[duplicated(masterdf),4:9]
```

```
[1] "duplicates removed"
[1] "Here's the proof:"
[1] Title          GivenName      MiddleInitial Surname        StreetAddress
[6] City
<0 rows> (or 0-length row.names)
```

```
#See if there is any missing data in the Linkfile
linkdf <- as.data.frame(linkdf)
sapply(linkdf, function(x) sum(is.na(x)))
```

```
      GivenName   MiddleInitial         Surname   StreetAddress            C
ity
              0               0               0               0
  0
          State         ZipCode    EmailAddress TelephoneNumber        Birth
day
              0               0               0               0
15
```

```
#Merge the two datasets on common fields, but not on Birthday due to missing
data
combinedf<-
merge(masterdf,linkdf,c("GivenName","MiddleInitial","Surname","StreetAddress","
City","State","ZipCode"))
str(combinedf)
```

```
'data.frame':	15000 obs. of  20 variables:
```

```
#Assume that, if all other data match, we can "impute" the missing birthday using the one we have
colnames(combinedf)[which(names(combinedf) == "Birthday.x")] <- "Birthday"
combinedf$Birthday.y <- NULL
str(combinedf)
```

```
'data.frame':	15000 obs. of  19 variables:
 $ GivenName      : chr  "Aaron" "Aaron" "Aaron" "Aaron" ...
 $ MiddleInitial  : chr  "A" "B" "C" "C" ...
 $ Surname        : chr  "Conrad" "Pittman" "Briggs" "Jones" ...
 $ StreetAddress  : chr  "346 Tenmile" "3681 Froe Street" "3779 Gnatty Creek Road" "2444 Emma Street" ...
 $ City           : chr  "Newport News" "New Martinsville" "Garden City" "Kaneohe" ...
 $ State          : chr  "VA" "WV" "NY" "HI" ...
 $ ZipCode        : int  23608 26155 11530 96744 48075 67202 16803 72210 46625 48185 ...
 $ GUID           : chr  "0cdb7668-ac63-4eee-8f8f-269cc16c9031" "338ecfe9-7b9d-437a-ab7d-bd0e276768c5" "0433573b-48a1-478c-8885-85e08d60dd90" "6356845e-3e74-4009-a905-c11a6067611c" ...
 $ NationalID     : chr  "224-05-5991" "236-90-5104" "114-58-8417" "750-01-7461" ...
 $ Gender         : chr  "male" "male" "female" "male" ...
 $ Title          : chr  "Mr." "Mr." "Ms." "Mr." ...
 $ Country        : chr  "US" "US" "US" "US" ...
```

```
Classes ‘data.table’ and 'data.frame':  15000 obs. of  11 variables:
 $ GUID          : chr  "0cdb7668-ac63-4eee-8f8f-269cc16c9031" "338ecfe9-7b
9d-437a-ab7d-bd0e276768c5" "0433573b-48a1-478c-8885-85e08d60dd90" "6356845e-
3e74-4009-a905-c11a6067611c" ...
 $ GivenName     : chr  "Aaron" "Aaron" "Aaron" "Aaron" ...
 $ MiddleInitial : chr  "A" "B" "C" "C" ...
 $ Surname       : chr  "Conrad" "Pittman" "Briggs" "Jones" ...
 $ StreetAddress : chr  "346 Tenmile" "3681 Froe Street" "3779 Gnatty Creek
Road" "2444 Emma Street" ...
 $ City          : chr  "Newport News" "New Martinsville" "Garden City" "Ka
neohe" ...
 $ State         : chr  "VA" "WV" "NY" "HI" ...
 $ ZipCode       : int  23608 26155 11530 96744 48075 67202 16803 72210 466
25 48185 ...
 $ EmailAddress  : chr  "AaronAConrad@gustr.com" "AaronBPittman@armyspy.co
m" "AaronCBriggs@armyspy.com" "AaronCJones@rhyta.com" ...
 $ TelephoneNumber: chr  "757-988-0409" "304-321-1218" "516-369-1420" "808-2
34-1935" ...
 $ Birthday      : chr  "3/3/1948" "1/31/1991" "10/30/1961" "7/15/1972" ...
 - attr(*, ".internal.selfref")=
```