REPORT: Log files dataset analysis

Jacek Bera

8/07/2017

1. Introduction:

Report includes statistics of dataset which contains multiple log files from Egnyte Cloud Server. Each row of dataset is one performed action.

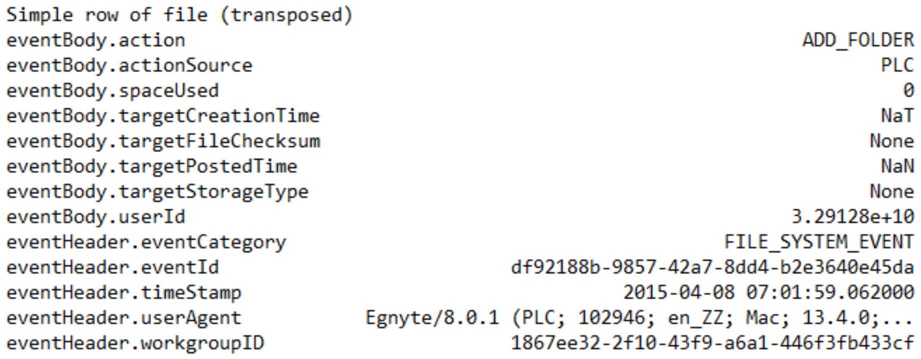
All calculations were made using Python and following modules and libraries:

a. json – used for encoding JSON objects
b. matplotlib – used for plotting
c. os – used for interfacing with folders and files
d. pandas – used for handling data, making calculations and statistics
e. tarfile – used for unpacking archive

Report also contains results of basic dataset tests and conclusions.

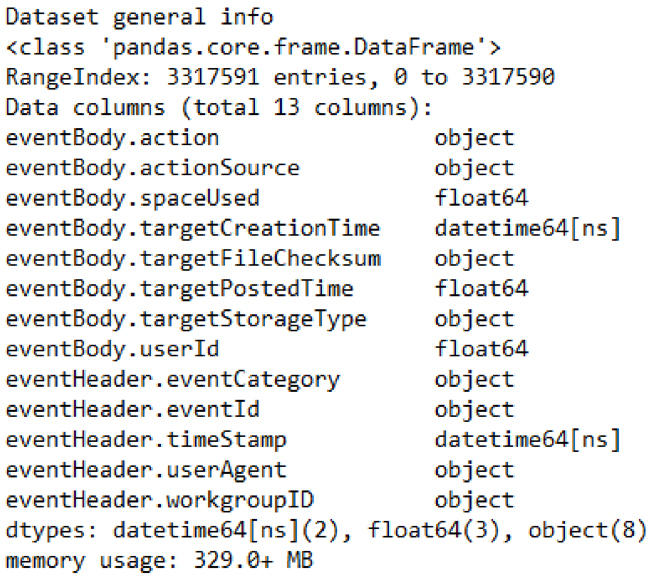
2. Statistics

Basic information about dataset is shown below.



```
Simple row of file (transposed)
eventBody.action                                                   ADD_FOLDER
eventBody.actionSource                                                    PLC
eventBody.spaceUsed                                                         0
eventBody.targetCreationTime                                              NaT
eventBody.targetFileChecksum                                             None
eventBody.targetPostedTime                                                NaN
eventBody.targetStorageType                                              None
eventBody.userId                                                  3.29128e+10
eventHeader.eventCategory                                   FILE_SYSTEM_EVENT
eventHeader.eventId                      df92188b-9857-42a7-8dd4-b2e3640e45da
eventHeader.timeStamp                              2015-04-08 07:01:59.062000
eventHeader.userAgent          Egnyte/8.0.1 (PLC; 102946; en_ZZ; Mac; 13.4.0;...
eventHeader.workgroupID                  1867ee32-2f10-43f9-a6a1-446f3fb433cf
```

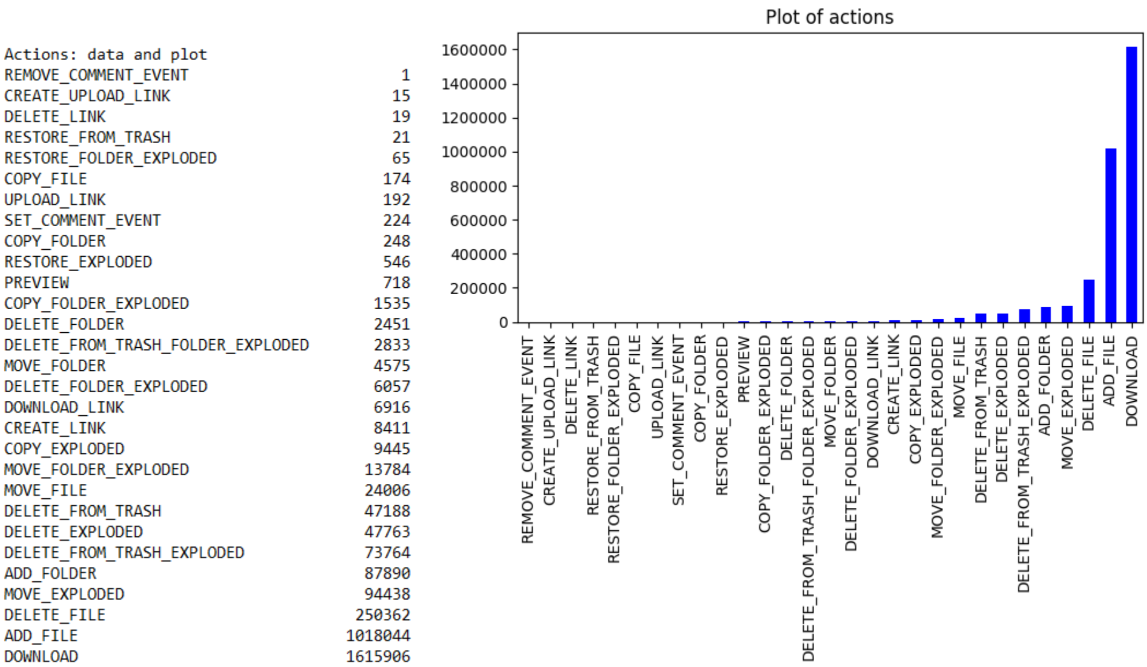
Row of dataset (transposed for more clear view) is showing parameters and simple values.



```
Dataset general info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3317591 entries, 0 to 3317590
Data columns (total 13 columns):
eventBody.action                object
eventBody.actionSource          object
eventBody.spaceUsed             float64
eventBody.targetCreationTime    datetime64[ns]
eventBody.targetFileChecksum    object
eventBody.targetPostedTime      float64
eventBody.targetStorageType     object
eventBody.userId                float64
eventHeader.eventCategory       object
eventHeader.eventId             object
eventHeader.timeStamp           datetime64[ns]
eventHeader.userAgent           object
eventHeader.workgroupID         object
dtypes: datetime64[ns](2), float64(3), object(8)
memory usage: 329.0+ MB
```

Dataset general info

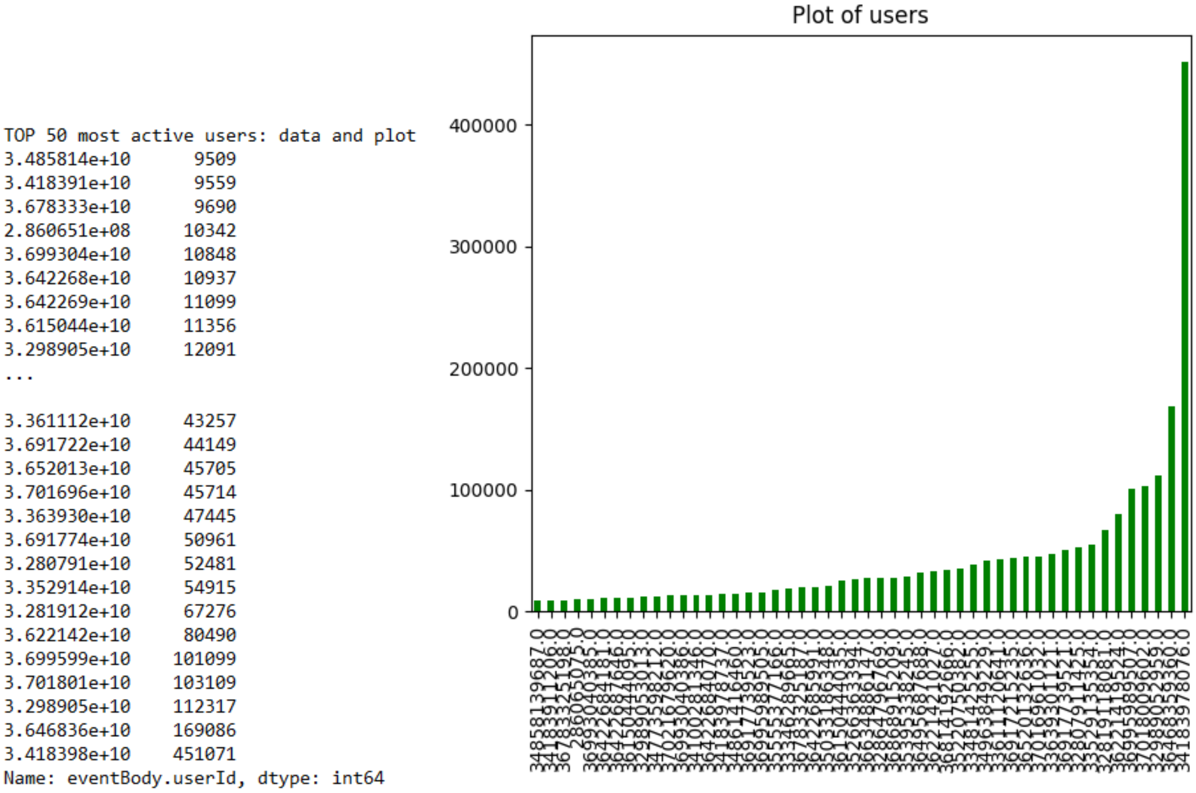
a. Statistics of actions performed by Egnyte Cloud Server.



```
Actions: data and plot
REMOVE_COMMENT_EVENT                          1
CREATE_UPLOAD_LINK                           15
DELETE_LINK                                  19
RESTORE_FROM_TRASH                           21
RESTORE_FOLDER_EXPLODED                      65
COPY_FILE                                   174
UPLOAD_LINK                                 192
SET_COMMENT_EVENT                           224
COPY_FOLDER                                 248
RESTORE_EXPLODED                            546
PREVIEW                                     718
COPY_FOLDER_EXPLODED                       1535
DELETE_FOLDER                              2451
DELETE_FROM_TRASH_FOLDER_EXPLODED          2833
MOVE_FOLDER                                4575
DELETE_FOLDER_EXPLODED                     6057
DOWNLOAD_LINK                              6916
CREATE_LINK                                8411
COPY_EXPLODED                              9445
MOVE_FOLDER_EXPLODED                      13784
MOVE_FILE                                 24006
DELETE_FROM_TRASH                         47188
DELETE_EXPLODED                           47763
DELETE_FROM_TRASH_EXPLODED                73764
ADD_FOLDER                                87890
MOVE_EXPLODED                             94438
DELETE_FILE                              250362
ADD_FILE                                1018044
DOWNLOAD                                1615906
```

Conclusion: *DOWNLOAD* action is the most popular action with big superiority, which is not surprising. Adding and deleting files are also very common. On the other hand copying files and folders is rare.

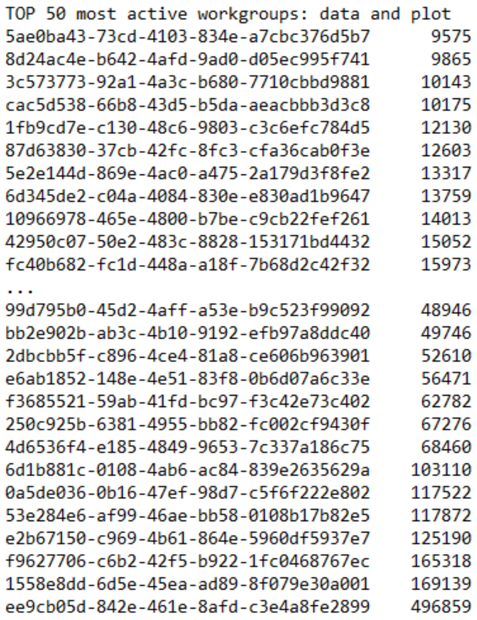
b. Statistics of most active users in Egnyte Cloud Server.



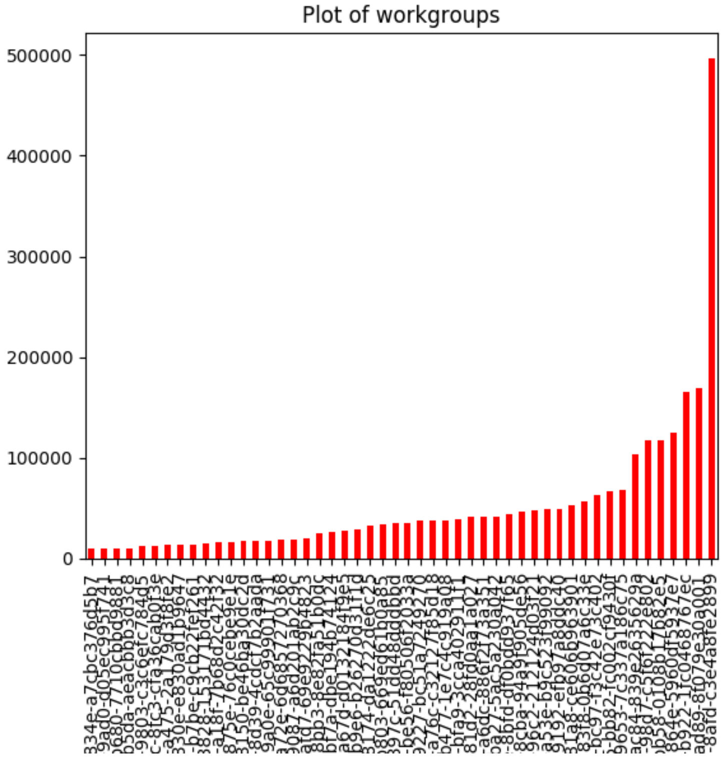
```
TOP 50 most active users: data and plot
3.485814e+10      9509
3.418391e+10      9559
3.678333e+10      9690
2.860651e+08     10342
3.699304e+10     10848
3.642268e+10     10937
3.642269e+10     11099
3.615044e+10     11356
3.298905e+10     12091
...

3.361112e+10     43257
3.691722e+10     44149
3.652013e+10     45705
3.701696e+10     45714
3.363930e+10     47445
3.691774e+10     50961
3.280791e+10     52481
3.352914e+10     54915
3.281912e+10     67276
3.622142e+10     80490
3.699599e+10    101099
3.701801e+10    103109
3.298905e+10    112317
3.646836e+10    169086
3.418398e+10    451071
Name: eventBody.userId, dtype: int64
```

Conclusion: Overall number of actions over users id shows that a small group of users is generating most of the traffic.

c. Statistics of most active workgroups in Egnyte Cloud Server.

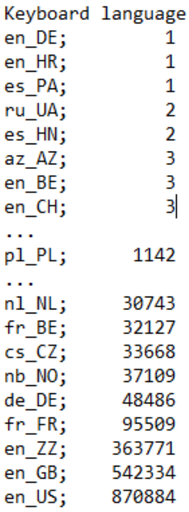


```
TOP 50 most active workgroups: data and plot
5ae0ba43-73cd-4103-834e-a7cbc376d5b7       9575
8d24ac4e-b642-4afd-9ad0-d05ec995f741       9865
3c573773-92a1-4a3c-b680-7710cbbd9881      10143
cac5d538-66b8-43d5-b5da-aeacbbb3d3c8      10175
1fb9cd7e-c130-48c6-9803-c3c6efc784d5      12130
87d63830-37cb-42fc-8fc3-cfa36cab0f3e      12603
5e2e144d-869e-4ac0-a475-2a179d3f8fe2      13317
6d345de2-c04a-4084-830e-e830ad1b9647      13759
10966978-465e-4800-b7be-c9cb22fef261      14013
42950c07-50e2-483c-8828-153171bd4432      15052
fc40b682-fc1d-448a-a18f-7b68d2c42f32      15973
...
99d795b0-45d2-4aff-a53e-b9c523f99092      48946
bb2e902b-ab3c-4b10-9192-efb97a8ddc40      49746
2dbcbb5f-c896-4ce4-81a8-ce606b963901      52610
e6ab1852-148e-4e51-83f8-0b6d07a6c33e      56471
f3685521-59ab-41fd-bc97-f3c42e73c402      62782
250c925b-6381-4955-bb82-fc002cf9430f      67276
4d6536f4-e185-4849-9653-7c337a186c75      68460
6d1b881c-0108-4ab6-ac84-839e2635629a     103110
0a5de036-0b16-47ef-98d7-c5f6f222e802     117522
53e284e6-af99-46ae-bb58-0108b17b82e5     117872
e2b67150-c969-4b61-864e-5960df5937e7     125190
f9627706-c6b2-42f5-b922-1fc0468767ec     165318
1558e8dd-6d5e-45ea-ad89-8f079e30a001     169139
ee9cb05d-842e-461e-8afd-c3e4a8fe2899     496859
```

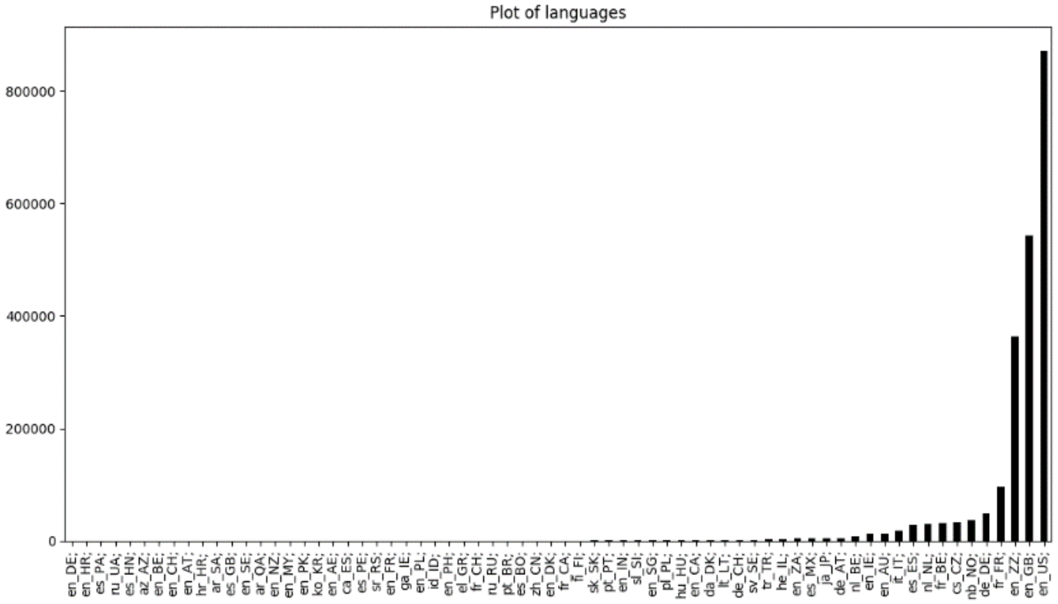


Conclusion: Similar to previous chart number of actions over workgroups shows that a small group of workgroups is generating most of the traffic.

d. Statistics of user keyboard



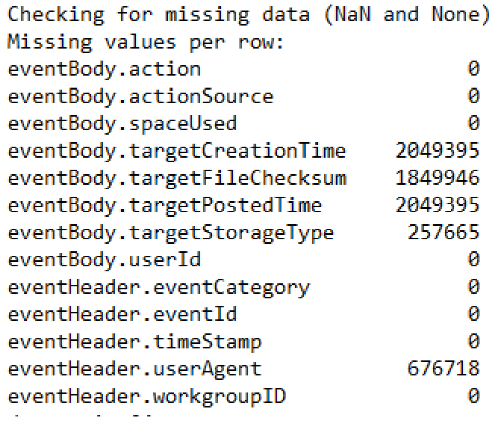
```
Keyboard language (users nationality) data and plot
en_DE;         1
en_HR;         1
es_PA;         1
ru_UA;         2
es_HN;         2
az_AZ;         3
en_BE;         3
en_CH;         3
...
pl_PL;      1142
...
nl_NL;     30743
fr_BE;     32127
cs_CZ;     33668
nb_NO;     37109
de_DE;     48486
fr_FR;     95509
en_ZZ;    363771
en_GB;    542334
en_US;    870884
```



Thanks to the *userAgent* parameter it is possible to extract data what keyboard languages are used by users. That allows to evaluate which languages the users speak. As a chart shows the most popular is English used in US and GB. Other main languages are: French, German and Norwegian.

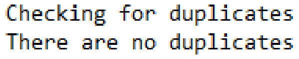
3. Dataset tests

The script is capable of verifying the data in terms of completeness and duplicates.
As tests show there are a lot of missing data in parameters:



```
Checking for missing data (NaN and None)
Missing values per row:
eventBody.action                    0
eventBody.actionSource              0
eventBody.spaceUsed                 0
eventBody.targetCreationTime  2049395
eventBody.targetFileChecksum  1849946
eventBody.targetPostedTime    2049395
eventBody.targetStorageType    257665
eventBody.userId                    0
eventHeader.eventCategory           0
eventHeader.eventId                 0
eventHeader.timeStamp               0
eventHeader.userAgent          676718
eventHeader.workgroupID             0
```

Looking for duplicates finished with empty result, so all records are unique.



```
Checking for duplicates
There are no duplicates
```

4. Conclusion

Dataset consists of 3317591 logs which can bring useful information about usage, most popular actions, users and groups. Time aspects can also be analyzed thanks to time stamps. Each log is unique but some of them have missing values probably because of the type of action or the type of client application.

5. Bibliography

- Pandas documentation: https://pandas.pydata.org/pandas-docs/stable/index.html
- Python documentation: https://docs.python.org/3/
- Matplotlib documentation: https://matplotlib.org/contents.html
- *Python dla każdego*, Michale Dawson, Helion 2010, Gliwice