Recruitment task report

The following report is presenting data calculated using attached script.

Generated plots and explanation:

1. Domain data bar plot presenting number of domains per data center.

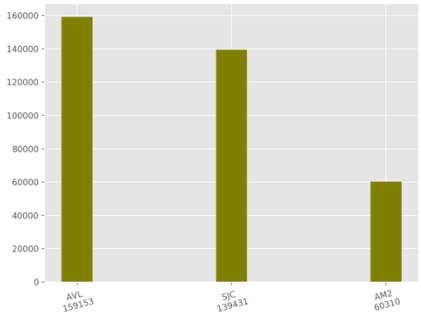
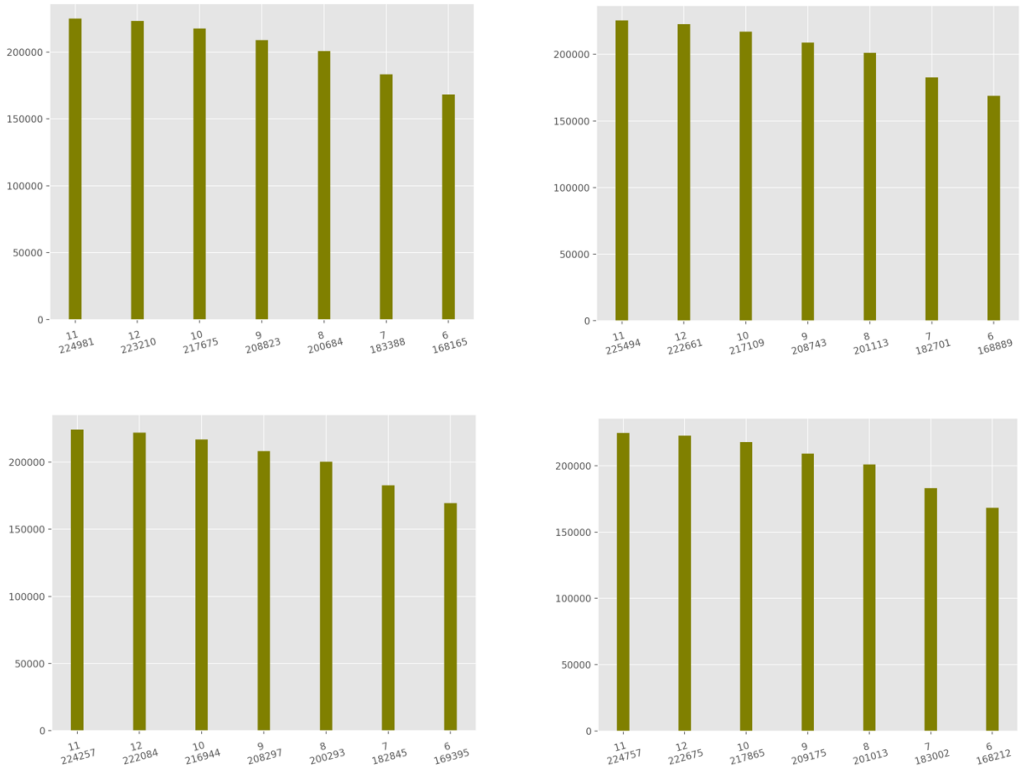


Chart is showing us how many domains from the json file are located in which data center. We can see that Asheville data center wins this coparison.

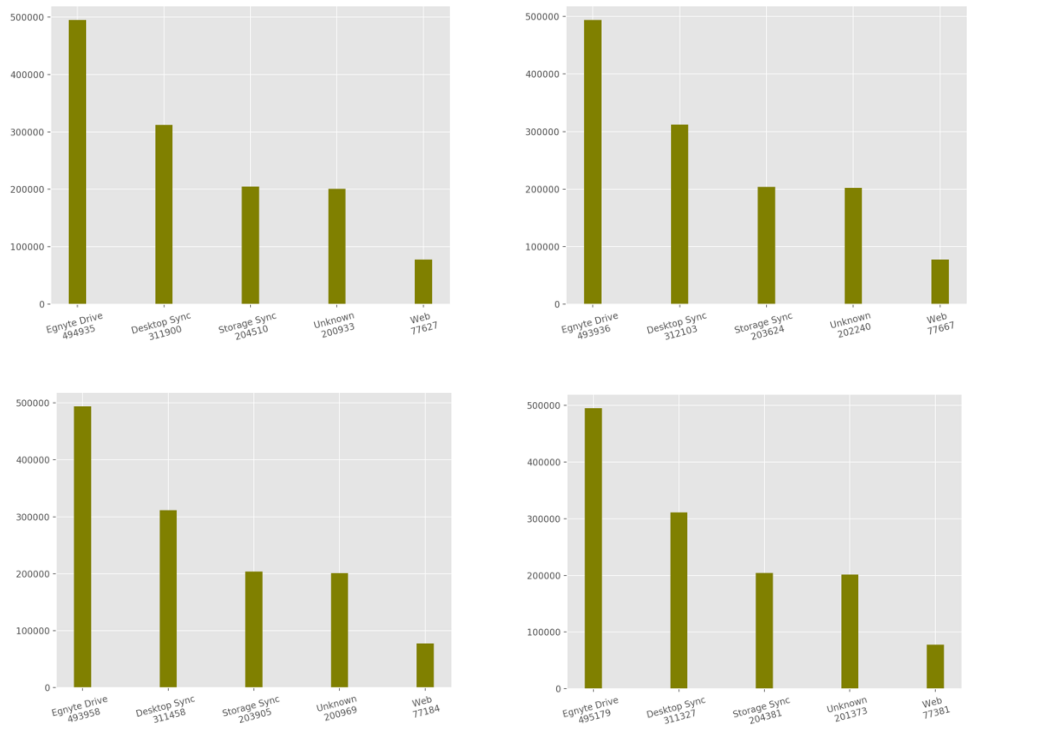
2. File System Event(First/Second/Third/Fourth) bar plot presenting most activities per hour.



First: Upper left, Second: Upper right
Third: Lower left, Fourth: Lower right

Charts are showing us activity on the domains per hour, we can see from the report that most actions are done between 6(almost million actions) and 12(almost half a million).

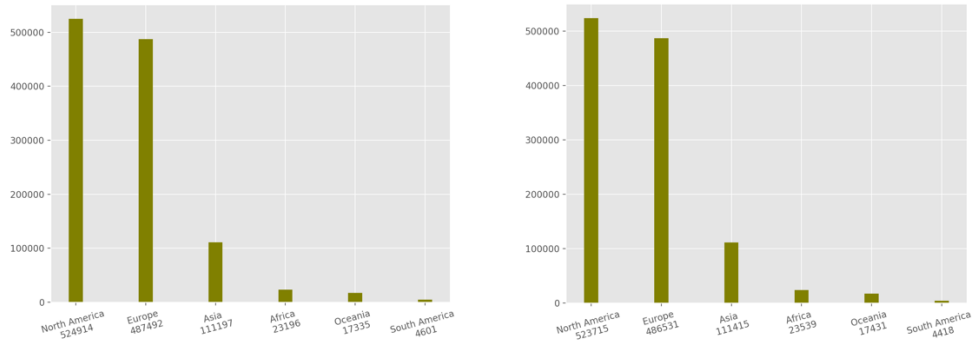
3. File System Event(First/Second/Third/Fourth) bar plot presenting most used apps.

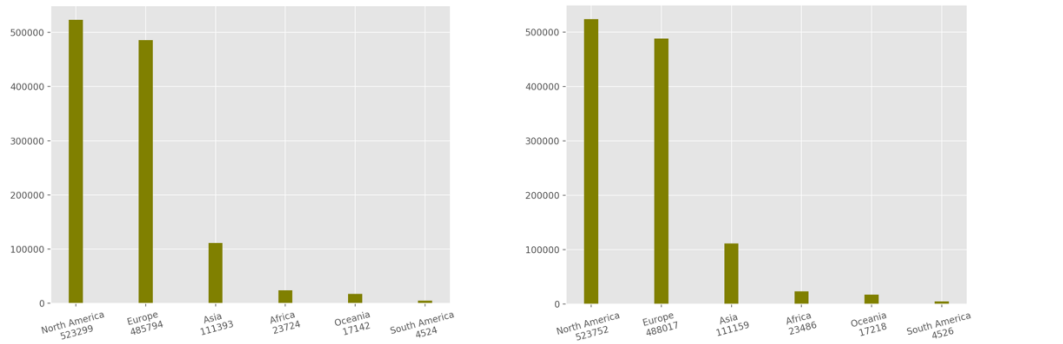


First: Upper left, Second: Upper right
Third: Lower left, Fourth: Lower right

Chart is showing us most used apps, as we can see most of the users are using Desktop App(almost 2 million actions).

4. File System Event(First/Second/Third/Fourth) bar plot presenting activity per continent.

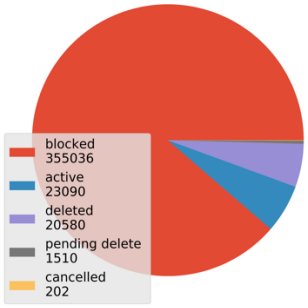




First: Upper left, Second: Upper right
Third: Lower left, Fourth: Lower right

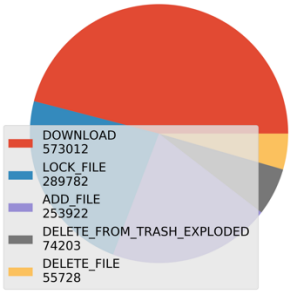
As we can see from the chart, most actions are done from USA(more than 2 million). Interesting is amount of actions done in Africa.

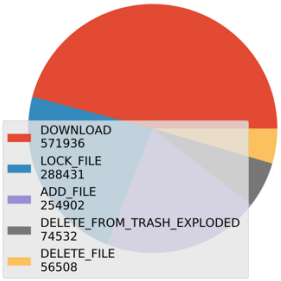
5. Domain data pie plot presenting domain statuses.

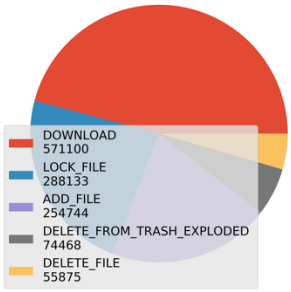


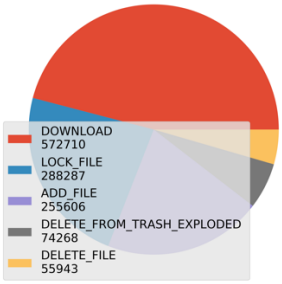
We see from the pie plot that most of the domains we get from the json file are blocked.

6. File System Event(First/Second/Third/Fourth) pie plot presenting top actions.





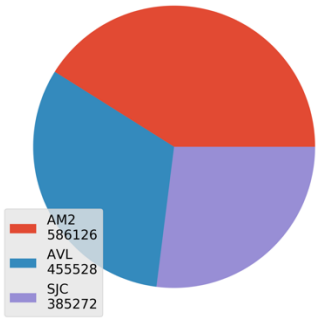


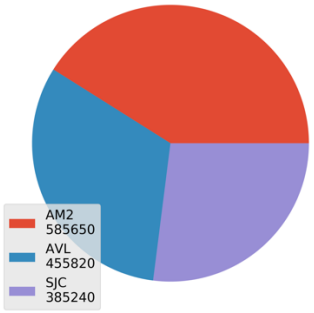


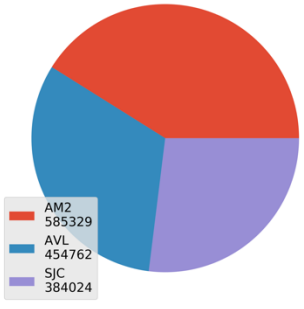
First: Upper left, Second: Upper right
Third: Lower left, Fourth: Lower right

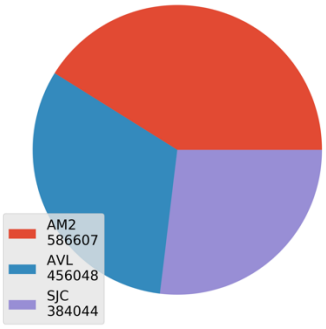
Charts are showing us that most of the actions are download actions, what is interesting in the top are delete actions (almost half a million).

7. File System Event(First/Second/Third/Fourth) pie plot presenting actions done per data center.









First: Upper left, Second: Upper right
Third: Lower left, Fourth: Lower right

Here are actions done per data center, interesting is fact that everything is divided almost in the same way, but comparing to the chart that is showing actions per continent and domain per data center, we can assume that a lot of US customers are using Europe data center(AM2) to store their data.

File values.txt contains mean value of domains per data center and summed used storage in MB.

All of the data was processed by reading the json files into date frames(stored previously in lists). I took mostly the head of counted values.

Improvements:

- Automatically changing filenames of saved plots
- Recognizing .tar.gz files and taking it automatically
- Excluding NaN rows or filtering them out during processing stage
- Counting data for given timestamp

Unit test suggestion:

- Checking filenames
- Checking file extensions
- Checking if the files from package are divided correctly
- Checking if data is counted correctly
- Checking if the files are not empty
- Checking if the files are containing correct data
- Checking the correct data type provided to plotting class
- Checking if the data written in .txt file is correct

Sources used during building the script:

- StackOverflow
- Libraries documentation