REPORT:

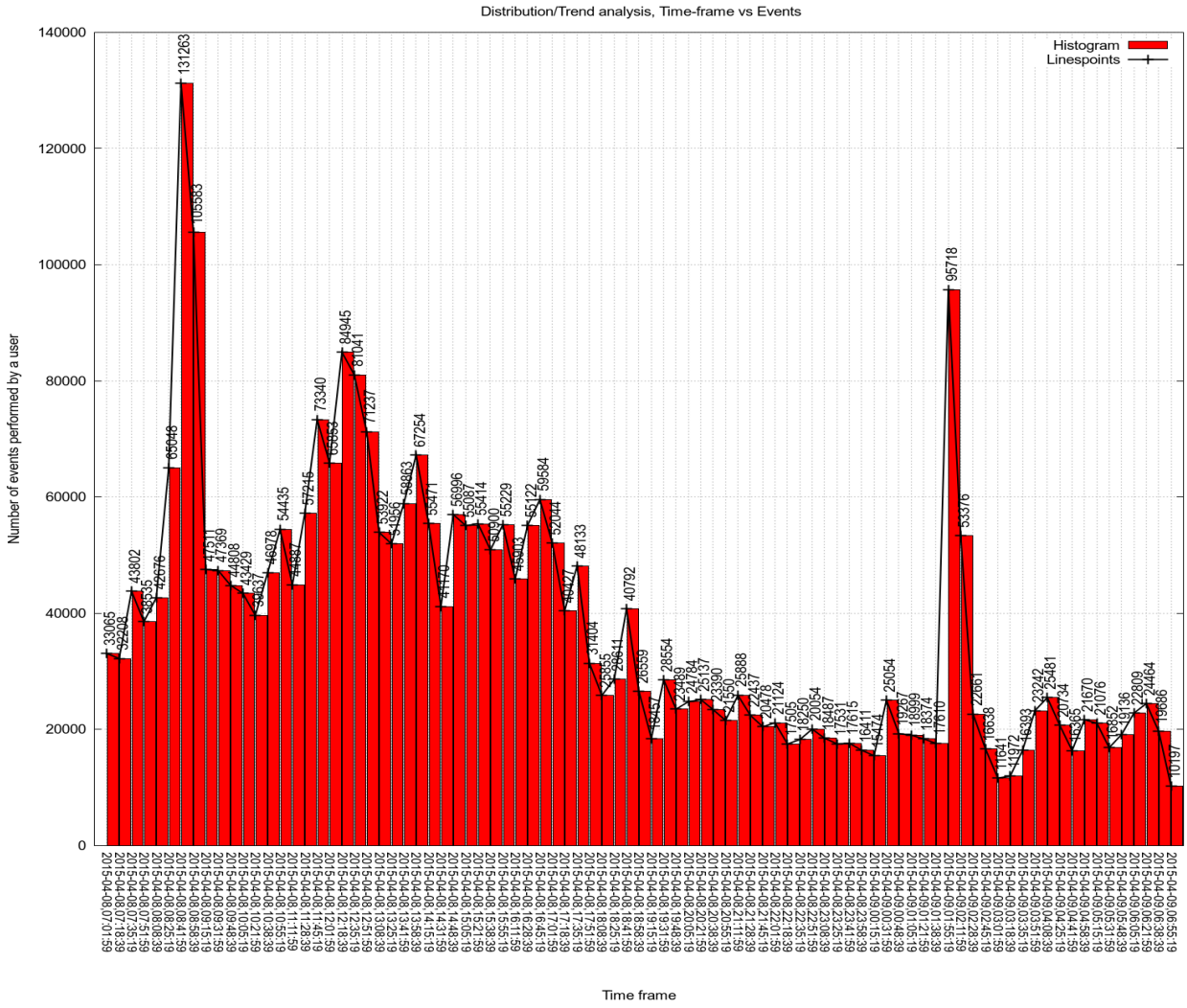
The data represents information about the actions performed using Egnyte cloud server. When performing data analysis, the data was taken as Time-series data.

Most time series patterns can be described in terms of two basic classes of components: trend and seasonality. The former represents a general systematic linear or (most often) nonlinear component that changes over time and does not repeat or at least does not repeat within the time range captured by our data. In the examined data, the trend of an event (action) occurring in the server changes over time. In the analysis the number of events were binned within a time frame. The approach was useful to understand the trend of actions performed on the cloud server. The latter may have a formally similar nature, however, it repeats itself in systematic intervals over time. Those two general classes of time series components may coexist in real-life data. .

There are no proven "automatic" techniques to identify trend components in the time series data; however, as long as the trend is monotonous (consistently increasing or decreasing) that part of data analysis is typically not very difficult.

A) Distribution/Trend Analysis of Time-frame vs Events/Actions performed:

In the analysis, the number of actions performed in a time range were binned (script7.sh) as shown in figure1.



The X-axis represents the “time-range” with gradual increase. The time range was taken from the UNIX “timestamp” parameter of the data. The time range was obtained by converting the UNIX time (timestamp) to human readable date and time using UNIX converter. The functionality used in R is :

```
R --no-save << EOT
timestamps<-read.table("timestamp")
timestamps2 <- as.POSIXct(timestamps, origin="1970-01-01")
write.csv(timestamps2,file="timestamp.out")
quit()
EOT
```

The Y-axis represents the “Number of actions performed by a user” over a time-range. The information about actions was taken from the “action” parameter of the data. As we see in the graph, the number of actions performed in the time range “2015-04-08,08:41:59 2015-04-08,08:58:39” is significant high compared to other time range. From the data, it could be interpreted that, the usage of the cloud server is high in this time period and thus server monitoring in this time could be useful to maintain the server traffic. Similarly, the actions performed in the time range “2015-04-08,08:58:39 2015-04-08,09:15:19”, “2015-04-09,01:55:19 2015-04-09,02:11:59” and “2015-04-09,01:55:19 2015-04-09,02:11:59” is

considerably high. Figure1 is a “skewed right” distribution model of the data with an anomaly at the time frame “2015-04-09,01:55:19 2015-04-09,02:11:59” and “2015-04-09,01:55:19 2015-04-09,02:11:59”.

Information on, which action is performed the most during these peak time period could be useful for traffic analysis and maintenance. Figure2 is the graphical representation of the same.

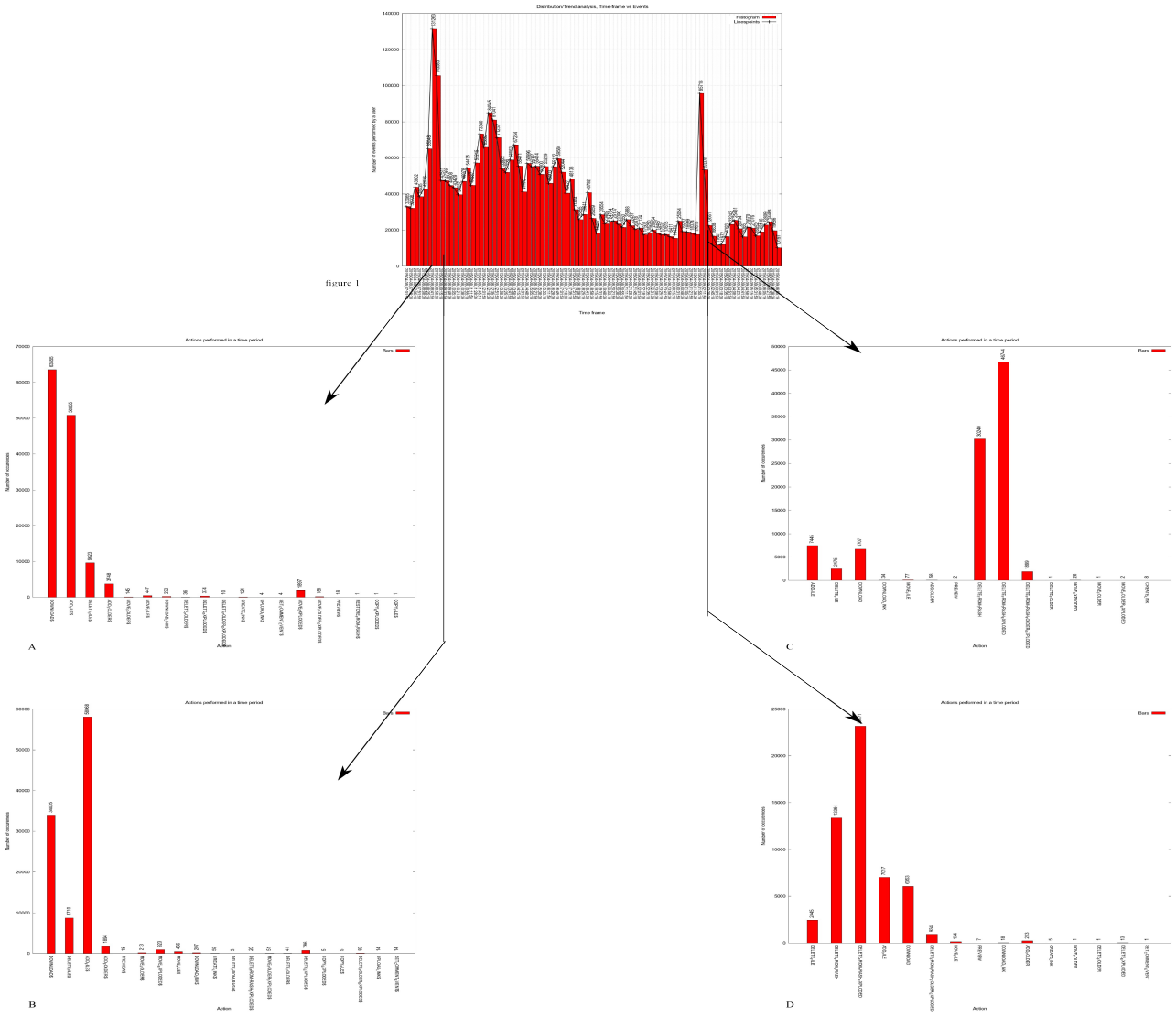


figure 1

Graph A. gives information about the type of actions performed in the time-frame “2015-04-08,08:41:59 2015-04-08,08:58:39”. As we see most of the actions performed by users are “DOWNLOAD” followed by “ADD FILE”. Similarly, in graph B. in the time frame “2015-04-08,08:58:39 2015-04-08,09:15:19” most actions performed are “ADD FILE” followed by “DOWNLOAD”. The interpretation derived form the observation is that in the early hours users perform the actions of downloading and adding files to the server. Graph C and D. give relatively

similar information about the actions performed in the time frame “2015-04-09,01:55:19 2015-04-09,02:11:59” and “2015-04-09,01:55:19 2015-04-09,02:11:59” repectively. In Graph C. most of the actions performed within the corresponding time frame are “DELETE FROM TRASH” and in graph D. most of the actions performed are again “DELETE FROM TRASH”. Based on the statistics the actions performed in the latter time period, graph C. denote that may be maintenance work is going on on the server. The analysis performed in figure1 and the related graphs is to visualize the trend followed by users in a time-range.

B) Linear Regression Analysis/Curve fitting:

Linear regression analysis is useful to understand and quantify the strength of the relation between two variables in which one variable is dependent on the other variable. Let say y=f(x) where y is dependent on function of x.
In the data, actions performed depends on the users. Hence, it is worthwhile to see if the correlation between these two variables can help to predict an anomaly.

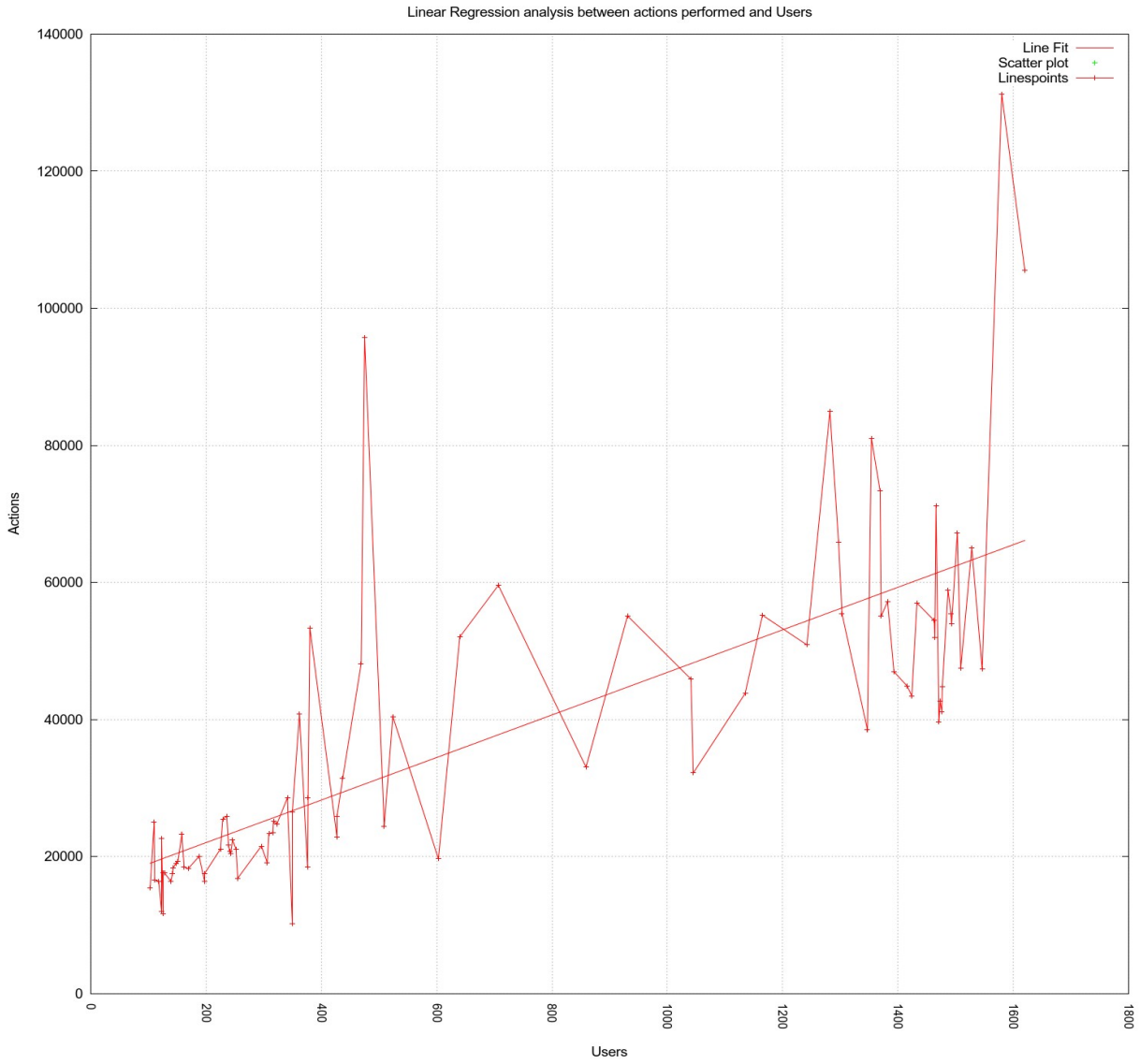
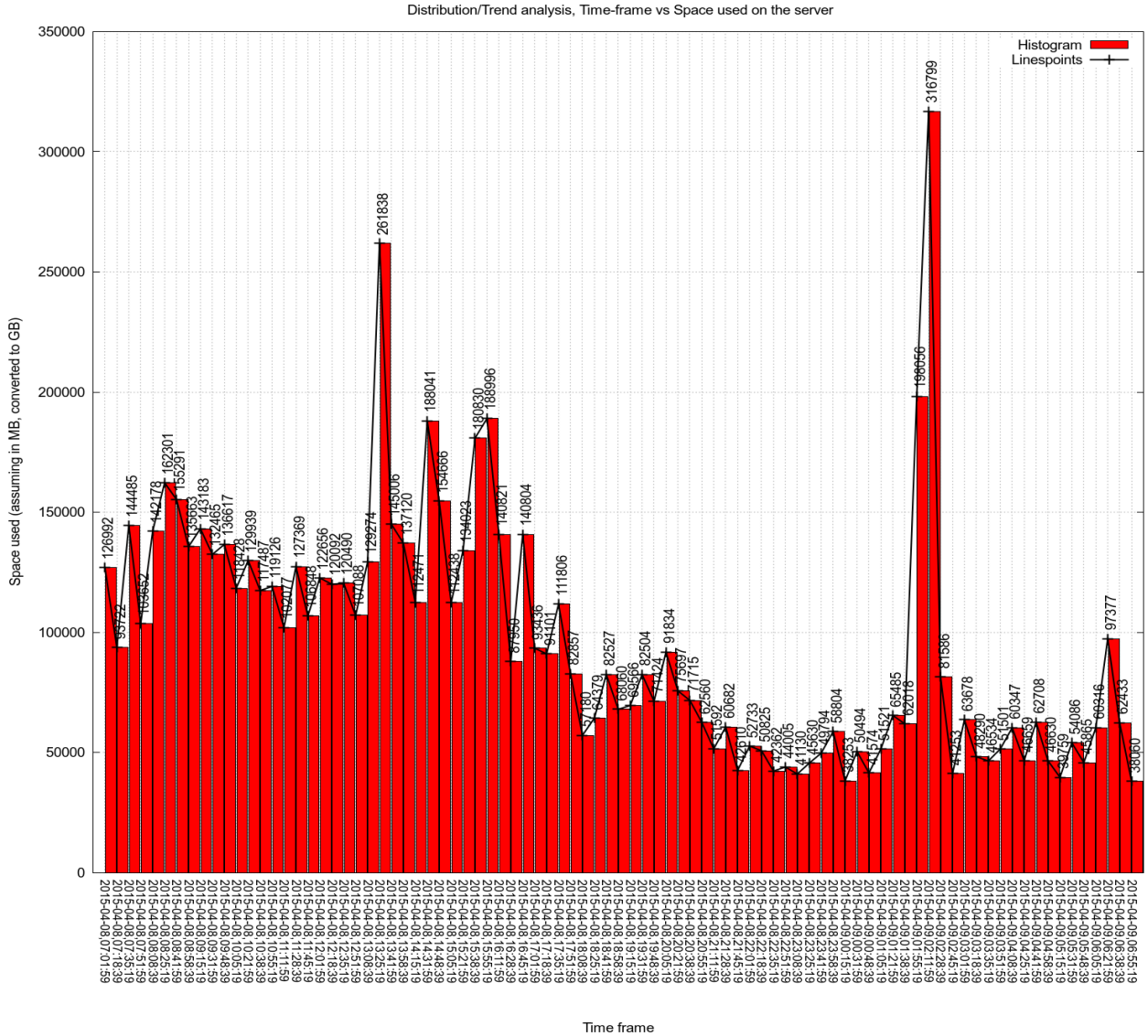


Figure3 is the plot representing the correlation between the number of users and the number of actions performed by the users in the complete dataset.

X-axis represents the number of users performing actions and Y-axis represents the number of actions performed. As we see there is a considerable correlation between the events with two outliers /anomalies (peaks). These abrupt increase in the number of users and actions gives a hint on the peak point of the server usage and business. Similar to figure1, we can say that special attention should be given in maintaining the server traffic during this period of time.

C) Distribution/ trend analysis of time-frame vs space used on the server:

This analysis was performed to understand the trend of the space usage on the server. In the analysis the space used on the server in an action were binned for a particular time-frame.



In Figure4, the X-axis represents the “time-range” with gradual increase. The time range was taken from the UNIX “timestamp” parameter of the data. The time range was obtained by converting the UNIX time (timestamp) to human readable date-time using UNIX converter. The functionality used in R is :

```
R --no-save << EOT
timestamps<-read.table("timestamp")
timestamps2 <- as.POSIXct(timestamps, origin="1970-01-01")
write.csv(timestamps2,file="timestamp.out")
quit()
EOT
```

The Y-axis represents the “Space used by a user” over a time-range. The information about space usage was taken from the “spaceused” parameter of the data. As we see in the graph, the amount of space used in the time range “2015-04-08,13:25:19 2015-04-08,13:41:59” is significant high compared to other time range. From the data, it could be interpreted that, the space usage of the cloud server is high in this time period and thus server monitoring in this time could be useful to maintain the server traffic. Similarly, the space used in the time range “2015-04-09,02:11:59 2015-04-09,02:28:3”, is considerably high. Figure4 is a “skewed right” distribution model of the data with an anomaly at the time frame “2015-04-08,13:25:19 2015-04-08,13:41:59” and “2015-04-09,02:11:59 2015-04-09,02:28:3”.

In summary, we can see that there are few anomalies found in the data which could be useful in maintaining the server and worth checking. High traffic during these peak time frame needs to be carefully monitored to sustain a hitch free server usage by the users.