Parallel Coordinates Visualization in the ELK Stack

Timofei Galkin[1,2][0000-0003-2859-6275] and Maria Grigorieva[2,3][0000-0002-8851-2187]

[1] National Research Nuclear University “MEPhI”, Kashirskoe shosse,
31, Moscow, 115409, Russia
[2] Scientific Research Computing Center, Lomonosov Moscow State University,
Leninskie Gory, 1, p.4, Moscow, 119991, Russian Federation
[3] Moscow Center of Fundamental and Applied Mathematics,
Leninskie Gory, 1, Moscow, 119991, Russian Federation
tpgalkin@mephi.ru, maria.grigorieva@cern.ch

Abstract. Modern large-scale distributed computing systems, processing large volumes of data, require mature monitoring systems able to control and track in resources, networks, computing tasks, queues and other components. In recent years, the ELK stack has become very popular for the monitoring of computing environment, largely due to the efficiency and flexibility of the ElasticSearch storage and wide variety of Kibana visualization tools. The analysis of computing infrastructure metadata often requires the visual exploration of multiple parameters simultaneously on one graphical image. Stacked bar charts, heatmaps, radar charts are widely used for the multivariate visual data analysis, but these methods have limitations on the number of parameters. In this research the authors propose to enhance the capacity of Kibana, adding Parallel Coordinates diagram - one of the most powerful method for visual interactive analysis of high-dimensional data. It allows to compare many variables together and observe correlations between them. This work describes the development process of Parallel Coordinates as a Kibana plugin, and demonstrates an example of visual data analysis based on the Nginx logs metadata.

Keywords: Parallel Coordinates, ELK Stack, Visualization, Log Analysis, Nginx.

1 Introduction

Nowadays the ELK (ElasticSearch-LogStash-Kibana) stack[1] [1] is increasingly used for the monitoring and analysis of data in a wide range of scientific and industrial applications. The integrated software components allow designing efficient and scalable monitoring systems. Kibana provides many plugins for graphical representation of

[1] https://www.elastic.co/

data, from tables to multi-layered maps and 3D images. Most of the plugins allow to explore a limited number of variables (or dimensions) in a single figure.

Thus, for the analysis of multidimensional data one of the most reasonable solutions is to use multiple graphs arranged together to display multiple variables. However, multivariate data analysis often lacks the ability to explore trends and correlations between many parameters (that might be of various types, units, scales, ranges) on a single graphical representation. We propose to expand the set of available Kibana plugins with Parallel Coordinates diagram designed specifically for the visual analysis of high-dimensional data.

2 ELK-stack in Data Analysis and Monitoring

ELK stack is a collection of open-source products: Logstash (and Beats) collects and filter data, ElasticSearch provides storing and searching through the collected dataset, and Kibana visualizes data in various graphical views.

Largest companies from many industries (LinkedIn, Fujitsu, eBay, Volvo, WLCG, InfoTrack, and many others)[2] chose the ELK stack for monitoring of the performance and security, log analysis and other use cases. CERN[3] with its large and powerful computing infrastructure utilizes the ELK for five different use cases for the WLCG[4]: messaging, job monitoring, data monitoring, infrastructure monitoring, cloud benchmarking [2]. ELK is also actively used to tackle a lot of analytical and monitoring tasks in the experiments at the LHC: logs processing and analysis, analysis of memory usage and CPU efficiencies at computing sites, exploration of timings of data processing steps, generation of alarms and alerts for network anomaly detection, optimizations of data and job brokering decisions. [3, 4]

3 Kibana Visualization Plugins

In the ELK stack, Kibana is responsible for data visualization. The visualizations are built on top of Elasticsearch queries and categorized into different types: Basic Charts, Data (Tables, Gauge, Metric), Maps and Time Series. Besides the standard means, it has many custom visualization plugins: scatter plots, 3D charts, calendar visualization, dendrograms, network graphs visualization, polar/radar charts, Cohort analysis chart, and many others[5].

When it comes to the multivariate data analysis, there is a need not only to check out distributions but also to explore correlations between data attributes and carry out

[2] https://www.elastic.co/customers

[3] The European Organization for Nuclear Research

[4] Worldwide LHC Computing Grid

[5] https://www.elastic.co/guide/en/kibana/current/known-plugins.html

trend analysis. Below we provided a certain types of Kibana visualizations that are suitable for multivariate data analysis[6]:

- ***Bar Charts*** uses horizontal or vertical bars to show discrete, numerical comparisons across categories. One axis of the chart shows the specific category and the other axis - a discrete value scale, answering the question of "how many?" for each category. Stacked and Grouped Bar charts display information about the subgroups that make up the different categories, adding the third dimension to the visualization. Side-by-side charts allow to add fourth variable.
- ***Line Graph, Area Charts*** are used to display the trends of quantitative parameters over time period. It can be stacked as Bar and Area Charts and allow to show distribution of categories as parts of a whole.
  Bar Charts, Line Graphs and Area Charts are typically limited by quantity of parameters they can display: from 1 to 4.
- ***Heat Maps*** visualise multidimensional data as a matrix through variations in colouring. Categorical data is colour-coded, and numerical data requires a colour scale in order to represent the difference in high and low values. The gradients in the heatmap allow observing the strength of the correlations. But it does not allow to analyse trends.
- ***Radar/Polar Charts*** are used to compare multiple quantitative variables. Variables are represented as axis starting from the centre of a circle and arranged radially. Values of data items, connected across all the axis, form a polygon. The objective of the graph is to assess the symmetry of the values rather than to compare their magnitudes. But if it has many polygons it makes it too complicated. Radar Charts might often be misread due to the controversial interpretation of polygons shapes.

Parallel Coordinates diagram can be a useful compliment to the listed methods. Adding Parallel Coordinates to the list of Kibana plugins is not a new idea. It had already been proposed on Kibana GitHub as an issue in 2014[7], but was closed due to the lack of engagement. The issue was created again in 2017[8], but still hasn't been implemented. Nevertheless, as the proposed technique is used in many research projects across numerous application areas such as mathematics, statistics, bioinformatics, medicine and climate science, it would be helpful to have it in Kibana - one of the most widely used tool for visual data analysis [5].

4 Parallel Coordinates Visualization for Multivariate Data Analysis

Parallel Coordinates are a common way of visualizing high-dimensional data. It is ideal for comparing many variables together and observing the relationships between

[6] We do not take into account 3D visualization, and data having hierarchy or graph structure that imply specific methods of visualization

[7] https://github.com/elastic/kibana/issues/1936

[8] https://github.com/elastic/kibana/issues/12118

them. Springer's “Handbook of Data Visualization” [6] says: “No other statistical graphic can plot so much information (cases and variables) at a time. Thus parallel coordinate plots are an ideal tool to get a first overview of a data set. It escapes the limitation of the orthogonal coordinate system by placing the coordinate axes parallel to each other.”

In traditional Cartesian coordinates, axes are mutually perpendicular. In Parallel Coordinates, all axes, representing variables, are placed in parallel to each other that allows representing data in much more than three dimensions. Potentially, the number of dimensions in this type of visualization is unlimited. Each axis may have a different scale and data type.

Each item in the data set is represented as a polyline that traverses all axes at appropriate points, providing a general overview of the whole data set. It allows to make general observations concerning the scales and the distributions of variables.

Static Parallel Coordinates diagram might be complicated for visual perception and interpretation due to the overplotting[9]. That's where the interactivity can greatly improve a visibility [7]. A list of available interactivity options is provided below:

- ***Selections on axis:*** User can click and drag along any axis to allocate a group of data items. Selected polylines are then highlighted, while all others are blurred. The polylines related to the group visually traverse through all vertical axis providing a clear view of the parametric trends.
- ***Reordering of axis:*** The rearrangement of axis allows to avoid clutter in visualizations and can be a manual operation, or be driven by some algorithm.
- ***Linked data table:*** Parallel Coordinates may have direct and indirect interaction with the data table. A query can be specified both from the visual representation and from the table below it.

5 Parallel Coordinates as Kibana Plugin

Parallel Coordinates advanced browser was first implemented as a standalone JavaScript application based on the redesigned D3[10] Parallel Coordinates module[11], licenced under the GNU General Public License version 3, and DataTables[12] library. The source code can be found at GitHub[13]. The screenshot of the application is shown in Figure 1.

Variables (vertical axis) may be of any data type: numeric, string, date and time (automatically recognized).

[9] Overplotting is when the data or labels in a data visualization overlap, making it difficult to see individual data points in a data visualization

[10] https://d3js.org/

[11] https://bl.ocks.org/jasondavies/1341281

[12] https://datatables.net/

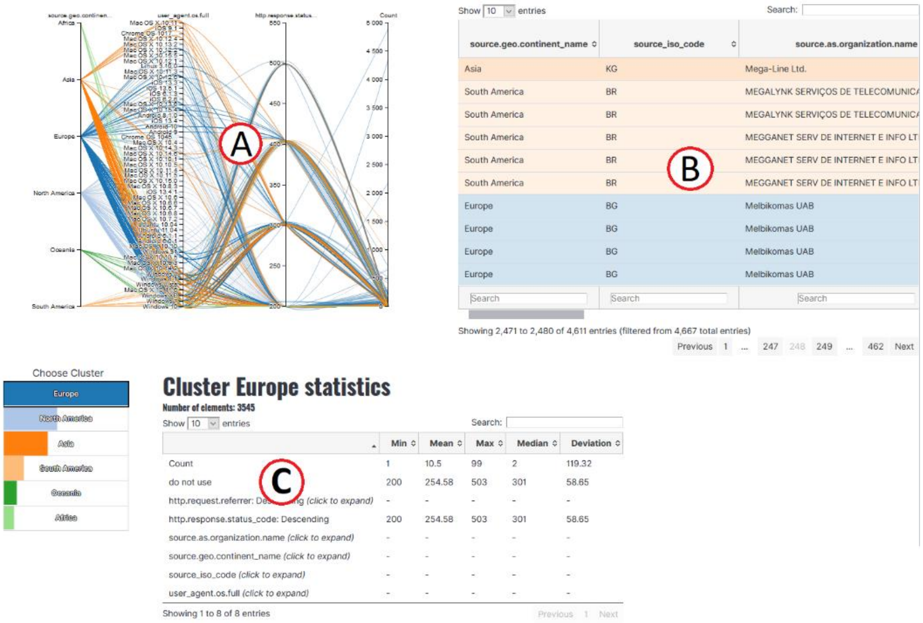
[13] https://github.com/PanDAWMS/InVEx-ParCoords-SA/tree/spa

To avoid overplotting, when too many overlapped lines make diagram unreadable, straight polyline segments are substituted with curves that do not stack. A slightly noticeable stroke opacity parameter is added to ease the overplotting impact.

The axis, visible on a diagram, as well as columns in a linked table, can be interactively selected by a user. Visible variables on the plot are not necessary the same as ones selected for the data table. It allows using the table as an additional source of data, providing auxiliary information about data objects.

Data items can be grouped by any categorical variable. Thus, all curves are highlighted in different colours, according to the group. Linked cluster panel in the section below allows to explore statistics of each group of items: min, max, mean, standard deviation values for numerical features, and distributions of unique values for categorical features. Size of the groups are demonstrated using the horizontal bar chart

A user can select segments on the Parallel Coordinates axis to limit the number of visible curves and explore trends of the selected group of items. This method is commonly referred as ‘brushing’ and described in [8]



Show 10 entries Search:

source.geo.continent_name	source_iso_code	source.as.organization.name
Asia	KG	Mega-Line Ltd.
South America	BR	MEGALYNK SERVIÇOS DE TELECOMUNICA
South America	BR	MEGALYNK SERVIÇOS DE TELECOMUNICA
South America	BR	MEGGANET SERV DE INTERNET E INFO LT
South America	BR	MEGGANET SERV DE INTERNET E INFO LT
South America	BR	MEGGANET SERV DE INTERNET E INFO LT
Europe	BG	Melbikomas UAB
Europe	BG	Melbikomas UAB
Europe	BG	Melbikomas UAB
Europe	BG	Melbikomas UAB

Showing 2,471 to 2,480 of 4,611 entries (filtered from 4,667 total entries)

Previous 1 ... 247 248 249 ... 462 Next

Cluster Europe statistics

Number of elements: 3545

Show 10 entries Search:

	Min	Mean	Max	Median	Deviation
Count	1	10.5	99	2	119.32
do not use	200	254.58	503	301	58.65
http.request.referrer: De[illegible]ng *(click to expand)*	-	-	-	-	-
http.response.status_code: Descending	200	254.58	503	301	58.65
source.as.organization.name *(click to expand)*	-	-	-	-	-
source.geo.continent_name *(click to expand)*	-	-	-	-	-
source_iso_code *(click to expand)*	-	-	-	-	-
user_agent.os.full *(click to expand)*	-	-	-	-	-

Showing 1 to 8 of 8 entries

Previous 1 Next

Fig. 1. Parallel coordinates plot with linked data table and clustering information. (A - Parallel Coordinates plot, B - linked data table, C - statistics of the selected groups)

The developed JavaScript tool was then implemented as a new Kibana visualization[14], allowing to visualize data as a Parallel Coordinates diagram using grouping features and filters available in Kibana. Plugin, as Kibana itself, is based on the React

[14] https://github.com/PanDAWMS/InVEx-ParCoords-Kibana

framework[15]. Node.js Packet Manager[16] modules (NPM modules) allows the installation of the necessary libraries in the project in a simple command, instead of the default way in pure JavaScript: download libraries one by one manually, insert them in the correct folder and link it in the html code.

A typical React repository structure consists of a set of folders and configuration files as shown in Figure 2.

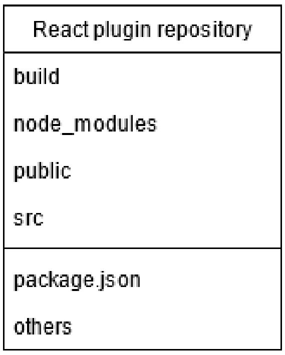


Fig. 2. React.js repository structure.

The *build* folder stores built archives of the application. *node_modules* contains all necessary Node.js modules. Other folders contain source code of the project. The repository structure of the Kibana repository itself, as well as the new Parallel Coordinates plugin, are similar to the shown one.

Kibana sources, necessary for the development of new plugins, are available for download at GitHub[17]. As repositories are nested in each other, the new plugins are placed in the *plugins* folder.

5.1 The Plugin Development

The development of a Kibana plugin starts with a configuration file called *package.json.* This file contains package name, its version, a specific Kibana version, necessary scripts and package dependencies.

It is important to know that this file necessarily holds only one very specific target Kibana version, without wildcard capabilities. It means that a plugin with version “7.6.*” will not work. Each version (and subversion) of Kibana requires its own set of archives, that limits plugin distribution capabilities.

A complete folder structure is shown in Figure 3.

[15] https://reactjs.org/

[16] https://www.npmjs.com/

[17] https://github.com/elastic/kibana

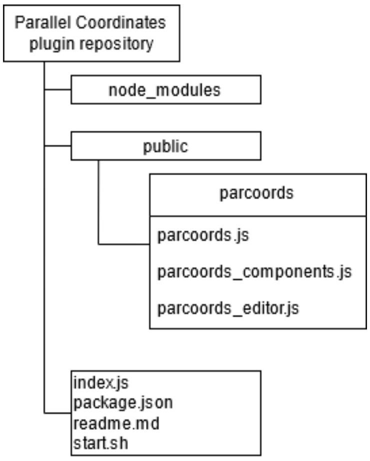


Fig. 3. Parallel Coordinates plugin file structure.

Initially, our Parallel Coordinates implementation was developed as an ordinary JavaScript code, which then was redesigned into an NPM module. The module is then added to the plugin project using the standard for Kibana packet manager - *yarn*. Based on our *package.json* installation scripts, it automatically creates necessary files in a single install command.

The new Kibana plugin is now available in "New Visualization" panel as "Parallel Coordinates", as shown in Figure 4.

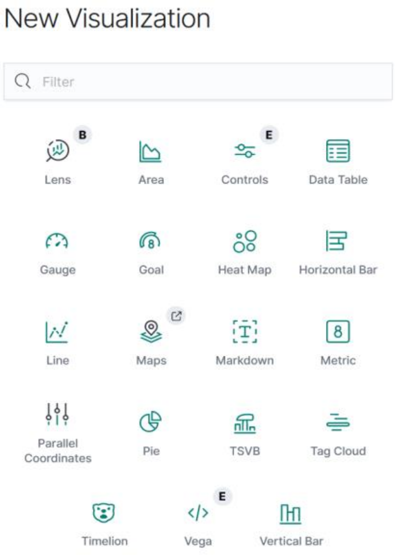
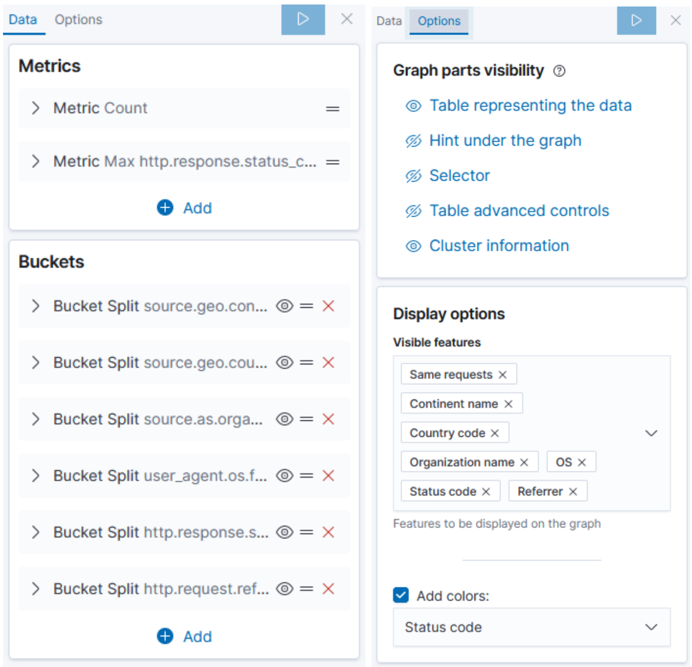


Fig. 4. New Visualization menu with the new plugin.

The visualization settings are provided in 'Data' and 'Options' tabs. An example of the settings window is shown in Figure 5.



Metrics

Metric Count

Metric Max http.response.status_c...

Buckets

Bucket Split source.geo.con...

Bucket Split source.geo.cou...

Bucket Split source.as.orga...

Bucket Split user_agent.os.f...

Bucket Split http.response.s...

Bucket Split http.request.ref...

Graph parts visibility

Table representing the data

Hint under the graph

Selector

Table advanced controls

Cluster information

Display options

Visible features

Same requests ×, Continent name ×, Country code ×, Organization name ×, OS ×, Status code ×, Referrer ×

Features to be displayed on the graph

☑ Add colors:

Status code

Fig. 5. Visualization Settings Window.

The 'Data' tab provides two types of ElasticSearch aggregations: Metrics and Buckets. Metrics aggregations calculate metrics over a set of documents (i.e., min, max, avg aggregation). Buckets group documents by a filter, criterion (i.e. terms aggregation) or range. Those options are used to construct an ElasticSearch request and finally visualize the result of the query. Buckets and Metrics parameters are converted into the vertical axis of Parallel Coordinates.

The 'Options' tab provides the ability to show or hide auxiliary parts of a diagram: data table, hints, feature selector and cluster information.

Next, the ElasticSearch request is constructed using the React Visualization procedure: user request is sent to the ElasticSearch, which process the request and returns the result in a form of a JavaScript object; then, the React visualization controller constructs a container as a Parallel Coordinates object and the respective visualization is rendered in a browser window.

6 Visual Analysis Using the Plugin

ELK stack is often used to collect and analyze log files from different applications. To demonstrate the application of the developed Kibana plugin in this paper we use data from an Nginx server node. The log files are configured to record data about client requests for a web page and access errors: request time, status, page address, command, client and host IP-addresses.

Below we will show the visual analysis of the server activity in May 2020 using Maps and Parallel Coordinates visualization.

The Map visualization (Figure 6) shows that the highest number of the requests come from Moscow (Russia) and France. Other regions send noticeably fewer requests. To explore the requests parameters in more detail we applied the Parallel Coordinates visualization plugin.

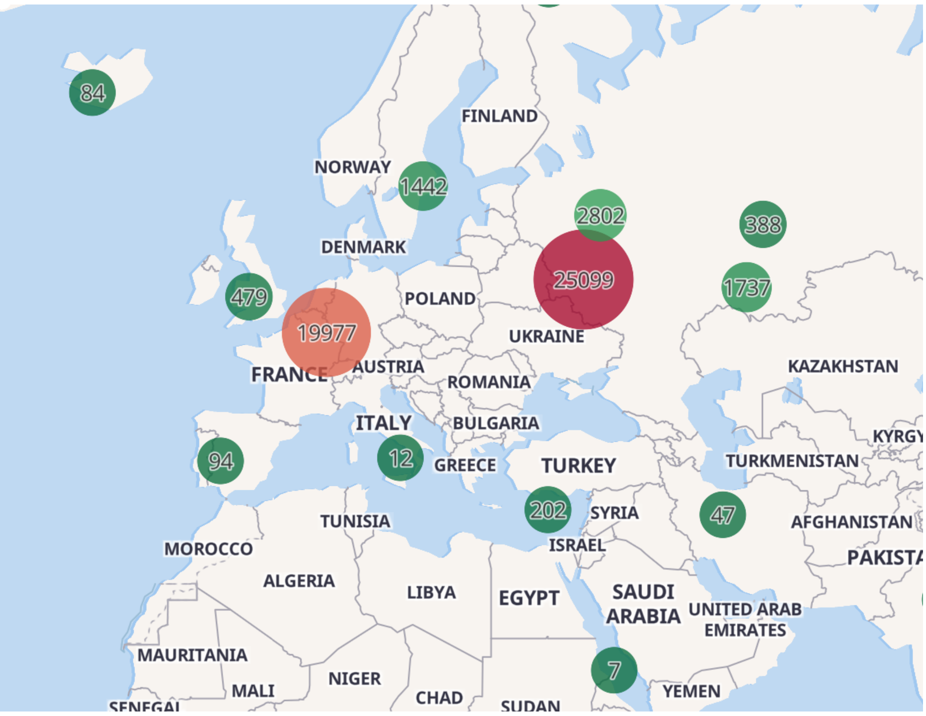


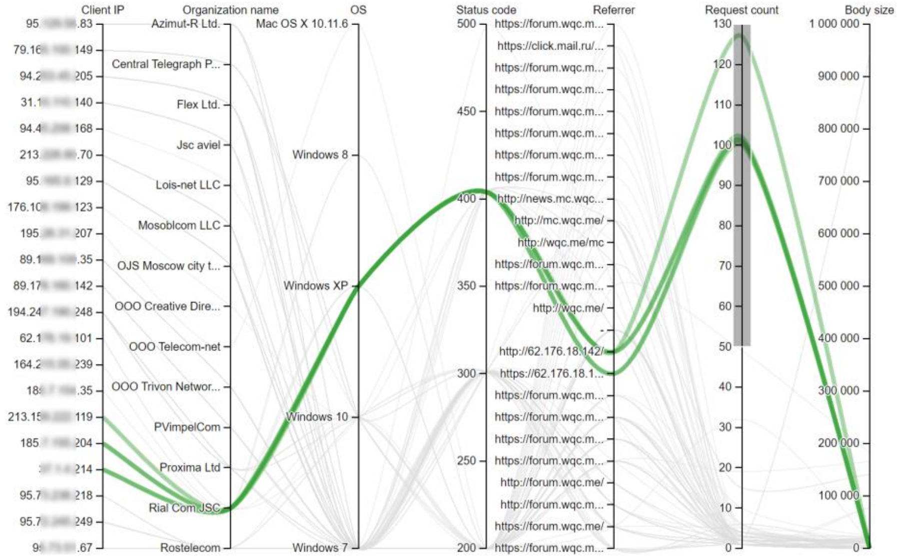
Fig. 6. Kibana Map visualization: distribution of the requests to the Nginx server node across the world.

Figure 7 demonstrates the visualization of the requests from Moscow region. The Buckets (groups), selected on the settings panel of the plugin, are the following:

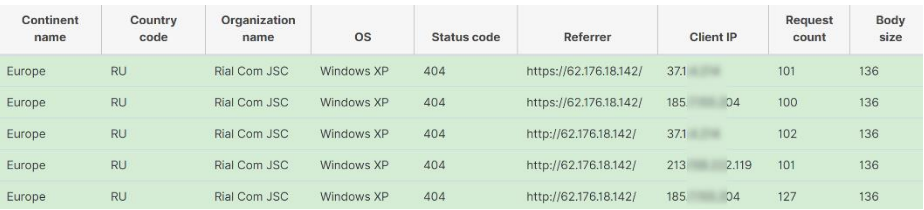
- ***IP*** - IP address of a sender
- ***Organization name*** - Internet Service Provider name
- ***OS*** - Operating system
- ***Status code*** - Status code of a request
- ***Referrer*** - Accessed web address
- ***Body size*** - size of a request message

Numerical metric is the ***Request count*** - number of requests.

The using of the brushing method is shown in Figure 7a. The number of requests to the server is interactively limited to be more than 50 using the selection of a relevant segment on the axis "Request count". The relevant group of curves is highlighted, while others become blurred. It helps to visually observe that several computers with the relatively outdated Windows XP have made over 500 requests that resulted with 404 (Not Found) error. Those requests do not come from any known search bot and might be initiated from computers with malicious reasons. The linked table, shown in Figure 7b, is used as the auxiliary source of information and may have additional data, that is not visualized on Parallel Coordinates diagram.



(a)

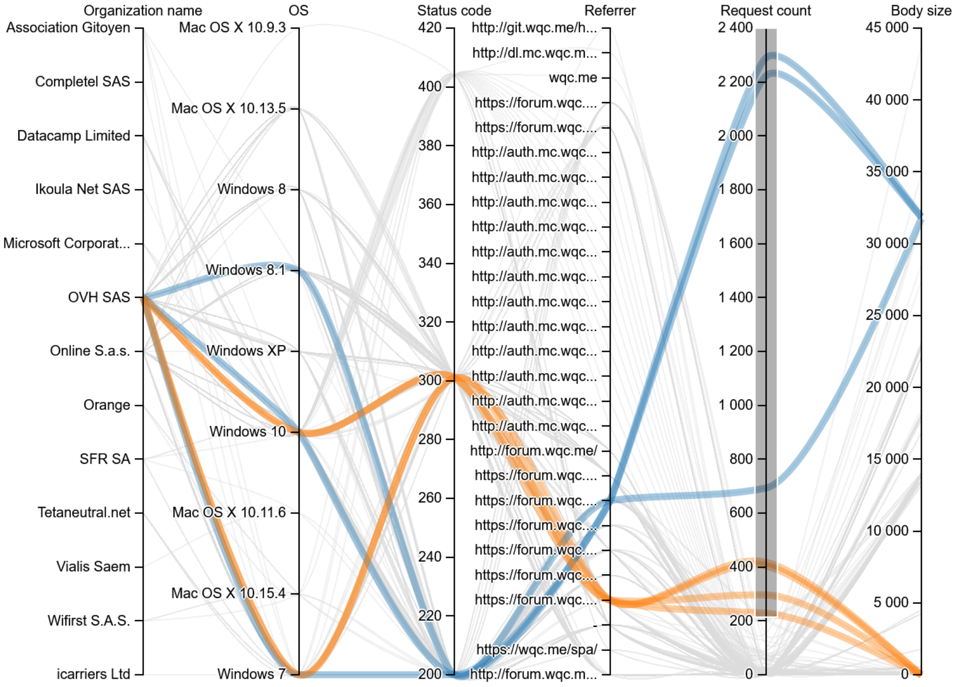


Continent name	Country code	Organization name	OS	Status code	Referrer	Client IP	Request count	Body size
Europe	RU	Rial Com JSC	Windows XP	404	https://62.176.18.142/	37.1[redacted]	101	136
Europe	RU	Rial Com JSC	Windows XP	404	https://62.176.18.142/	185[redacted]04	100	136
Europe	RU	Rial Com JSC	Windows XP	404	http://62.176.18.142/	37.1[redacted]	102	136
Europe	RU	Rial Com JSC	Windows XP	404	http://62.176.18.142/	213[redacted]2.119	101	136
Europe	RU	Rial Com JSC	Windows XP	404	http://62.176.18.142/	185[redacted]04	127	136

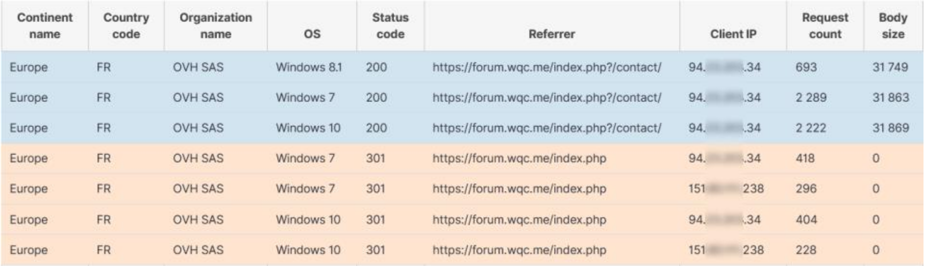
(b)

Fig. 7. Visualization of requests from Moscow, Russia.

The next Figure 8 shows the inspection of requests from France. A lot of requests (more than 6000) come from the same IP address (that is shown in linked table), but various operating systems. This could mean an attempt to harm the server. Diagram shows that more than 5000 requests were sent to one particular web page. This page is a 'Contact Administration' form, which in this case is used to send unwanted emails to the server administration.



(a)



Continent name	Country code	Organization name	OS	Status code	Referrer	Client IP	Request count	Body size
Europe	FR	OVH SAS	Windows 8.1	200	https://forum.wqc.me/index.php?/contact/	94.[redacted].34	693	31 749
Europe	FR	OVH SAS	Windows 7	200	https://forum.wqc.me/index.php?/contact/	94.[redacted].34	2 289	31 863
Europe	FR	OVH SAS	Windows 10	200	https://forum.wqc.me/index.php?/contact/	94.[redacted].34	2 222	31 869
Europe	FR	OVH SAS	Windows 7	301	https://forum.wqc.me/index.php	94.[redacted].34	418	0
Europe	FR	OVH SAS	Windows 7	301	https://forum.wqc.me/index.php	151[redacted]238	296	0
Europe	FR	OVH SAS	Windows 10	301	https://forum.wqc.me/index.php	94.[redacted].34	404	0
Europe	FR	OVH SAS	Windows 10	301	https://forum.wqc.me/index.php	151[redacted]238	228	0

(b)

Fig. 8. Visualization of requests from France.

7 Conclusion

General strength of Parallel Coordinates is the ability to quickly provide an overview of the multivariate data. For cluster and trends analysis this type of visualization may perform better than other methods used in Kibana. Currently, the plugin is being actively tested against various analytic tasks in ATLAS experiment at the LHC: slow tasks analysis, analysis of computing resources, data popularity and networks analysis.

8 Acknowledgements

The development of the standalone JavaScript application - Parallel Coordinates browser - was supported by the Russian Science Foundation grant No.18-71-10003. Kibana plugin was implemented with the support by the Moscow Center of Fundamental and Applied Mathematics.

References

1. Son, S. J., Kwon, Y.: Performance of ELK stack and commercial system in security log analysis. In: 2017 IEEE 13th Malaysia International Conference on Communications (MICC), Johor Bahru, 2017, pp. 187-190. https://doi.org/10.1109/MICC.2017.8311756
2. Sanz, P.: Grid Monitoring at CERN with the Elastic Stack. https://www.elastic.co/elasticon/conf/2016/sf/grid-monitoring-at-cern-with-the-elastic-stack. Last accessed 28 Jun 2020
3. Alekseev, A., Korchuganova, T., Padolski, S.: The BigPanDA self-monitoring alarm system for ATLAS. In: Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018). https://cds.cern.ch/record/2649752/
4. Vukotic, I., Robert, G., Lincoln, B.: Getting the Most from Distributed Resources: an Analytics Platform for ATLAS Computing Services. In: 38th International Conference on High Energy Physics (ICHEP2016), vol. 282. https://doi.org/10.22323/1.282.0192
5. Johansson, J., Forsell, C.: Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research. IEEE Transactions on Visualization and Computer Graphics. (2015). Vol. 22. 1-1. https://doi.org/10.1109/TVCG.2015.2466992
6. Chen, C., Härdle, W., Unwin, A.: Handbook of Data Visualization. Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-33037-0
7. Siirtola, H.: Interactive Visualization of Multidimensional Data. PhD thesis, University of Tampere, 2007.
8. Roberts, R. C., Laramee, R. S., Smith, G. A., Brookes, P., D'Cruze, T.: Smart Brushing for Parallel Coordinates, In IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 3, pp. 1575-1590, 1 March 2019, https://doi.org/10.1109/TVCG.2018.2808969