International Journal of Academic Information Systems Research (IJAISR)
ISSN: 2643-9026
Vol. 4, Issue 5, May – 2020, Pages: 66-71

Visualization of Big Data in Terrorism Using Elasticsearch, Logstash, Kibana (ELK) Stack

Eko Handoyo*, Maman Somantri, Hermawan, Bambang Winardi, Agung Nugroho, Aris Triwiyatno, Susatyo Handoko, Sukiswo, Yuli Christiyono, Yosua Alvin Adi Soetrisno

Department of Electrical Engineering, Diponegoro University, Semarang, Indonesia

Email: eko.handoyo@undip.ac.id*

Abstract: *Field of information technology is developing so rapidly, as well as in terms of data processing. The amount of data available and stored (both at the global level) is almost unimaginable in number. The data will continue to grow without stopping. This means that big data has high potential to gather insights from business information. Unfortunately until now, only a small portion of the data has been analyzed. Big data in business is a good strategy in processing raw information into profits that continue to flow to business organizations every day. Data processing is demanded to be faster and visualized so well that big data can be useful and easily understood by its users. One of them is by using the ELK (Elasticsearch, Logstash, Kibana) stack. In the environment of PT. Bumi Manunggal Sinergi, in terms of visualizing big data, the use of the ELK method becomes very useful, a variety of varied and interesting visualization choices makes processing big data easier.*

Keywords— visualization; big data; ELK

1. Introduction

Growing technology demands fast data processing. This technology focuses on finding hidden threads, trends, or patterns from corporate data stacks. This is important information that can open new opportunities. The main reasons for using big data are to find competitive advantage, influence business performance, and drive innovation and improvement. Generally there is a 3V characteristic model used to describe big data, which are volume, velocity and variation. Volume relates to how much data will be stored. Businesses collect data from various sources such as sales transactions and social media where the data stored continuously. Nowadays, large data technology makes data storage easier but the data access is still limited to memory capacity and hard-disk input output capacity. Velocity is terminology for real time data. Data which is streaming in real time must be handled immediately and also needed to be transformed directly to fid data model. Finally, variation refers to different data formats such as text, video, images and audio. Businesses must be able to organize and store data in many formats and display the visualization that was easily understood by the users of the data. Therefore, in terms of presenting the big data, it can be used a searching collaboration technology such as the Elasticsearch, Logstash and Kibana stack. By using this stack presentation of data will be easier and simpler and can use various type of visualization styles which provided by the tools. In the term of research we used criminal dataset to show the characteristic of visualization of the shifting in terrorism case.

2. Theoritical Background

2.1 Data Visualization

Data visualization refers to techniques used to communicate data or information by making it a visual object (for example like a connected points, lines, or bars) in a graph. The goal of visualization is to communicate information clearly and efficiently to users. Data visualization is one stages of data analysis terminology in data science. According to Friedman [1], the main purpose of data visualization is to communicate information clearly and effectively in a graphical way. It does not mean that data visualization must look static and rigid in order to illustrate very sophisticated dataset but the attractively way is more preferred. Data representation must be designed in effectively, aesthetically, and concurrently manner by providing insights for complex and rare data sets. Data set must be represented to communicate main aspects in an intuitive way. Data engineer sometimes failed to strike a balance between data form and program function. Main problem of visualization is to create captivate data visualizations without failing to provide their primary purpose for providing information. Fernanda Viegas [2] suggested that an ideal visualization is not only to communicate clearly, but can stimulate the attention and involvement of the audience.

Data visualization is closely related to information graphs, information visualization, scientific visualization, exploration of data analysis and statistical graphs. In the new millennium, data visualization has become an active area of research, teaching and development. According to Post et al. [3], data visualization has combined information and scientific visualization.

2.2 Characteristics of Effective Data Appearance

Professor Edward Tufte [4] explains that users of data visualization have done certain analytical work such as making comparisons or determine causality. The design principle of information charts must support analytical work, showing comparison or causality. Edward Tufte [4] defines the appearance of graphs and the principles of effective graphic display as in statistical charts consists of communicating complex ideas with clarity, accuracy and efficiency. Criteria for a good graphic representation includes:

1. Showing data.
2. Encourage the viewer to think about substance rather than methodology, graphic design, technology from graphic production or other things.
3. Avoid deception of what is said by the data.
4. Give a lot of numbers in a small space (thorough).
5. Make a large, coherent data set.
6. Encourage the eye to compare different parts of the data displayed.
7. Open data at several levels of detail, from general description to the final structure
8. Serve a clear purpose: description, exploration, tabulation or decoration
9. Closely integrates with the statistics and verbal descriptions of a data set.

2.3 Quantitative Data Visualization Messages

Stephen Few [5] explains 8 types of quantitative messages that users try to understand or communicate from a collection of data and graphics that are used to help communicate messages

1. Time-series: a single variable is captured over a period of time, such as a 10-year poverty rate. A line chart can be used to show this trend.

2. Rating: categorical distribution is ranked in ascending or descending order, such as sales rank (size) by seller (category, with each seller as a categorical division) for a single period. A bar graph can be used to show comparisons between sellers.

3. Part-to-whole: categorical division is measured as a ratio to the whole (for example, a percentage of 100%). A circle chart or bar graph can show a ratio of ratios, such as the ownership of shares represented by competitors in a market.

4. Deviation: the division of categories is compared with a reference, such as the comparison of actual expenditure against the budget for several departments of a business in a certain time period. Bar graphs can show a comparison of actual values against the amount referenced.

5. Frequency distribution: shows the number of observations of certain variables over a certain time span, such as the number of years in which the stock market is profitable is between intervals such as 0-10%, 11-20%, etc. A histogram, a type of bar graph, can be used for this analysis.

6. Correlation: the comparison between observations is represented by two variables (X, Y) to determine whether they are inclined to move in the same or opposite direction. For example, plotting unemployment (X) and inflation (Y) for a sample of several months. A scatter plot is usually used to convey the message.

7. Nominal ratio: compares the division of categories in no particular order, such as the number of sales based on the product code. Bar charts can be used for this comparison.

8. Geographical or geospatial: comparison of a variable on the map or location, such as the unemployment rate based on the state or number of people on the floor in a building. The chart used is usually a cartogram. An analyst who reviews a set of data can consider whether some or all of the messages and graph types above can be applied to work or received by data users. The trial process to identify the relevance and meaning of messages to the data is part of the exploration of data analysis.

2.4 Visual Perception and Data Visualization

The main difference between the length of the two lines, the orientation of the shape, and the color (style) without significant processing effort, is called the "attribute of attention". For example, it might take time and effort ("attention processing") to identify the number of times that number "5" appears in a set of numbers, but if the numbers differ in size, orientation, or color, the instant of the number can be seen more quickly through processing attention.

Effective graphics use the advantages of attention processing and the attributes and relative strengths and attributes. For example, humans can easily process differences in line lengths rather than surface areas, it is more effective to use bar graphs (which take advantage of line lengths to show comparisons) than circle graphs (which use surface areas).

2.5 Data Presentation Architecture

Data Presentation Architecture (DPA) is a collection of expertise that tries to identify, place, manipulate, format and provide data with a way to optimally communicate meaning and provide knowledge.

The term DPA was first defined by Kelly Lautt. DPA is defined as an application of expertise that is rarely used for the success factor of business intelligence. Data presentation architecture combines the science of numbers, data and statistics in the term of finding valuable information. Data presentation extracted insight from data and making it useful, related and can be witnessed by the art of data visualization. Data presentation provide communication, organizational psychology and change management with the aim of providing business intelligence solutions. Business intelligence uses data scope, timing of delivery, format and visualization that will effectively support strategic behavior towards business goals (or organization) that could be understood. DPA is not the ability of information technology (IT) or business but is a separate part of science expertise. DPA is a broader expertise that includes determining what data and at what time and what format data will be presented, by considering the best way to display pre-selected data (ie data visualization). The ability of data visualization is one element of DPA.

2.6 ElasticSearch, Logstash, Kibana (ELK)

Elasticsearch, Logstash and Kibana are useful tools for collecting logs and also visualizing, Elasticsearch is useful for storing all logs originating from the server, Elasticsearch is one of the databases that use the NoSQL engine with a focus on database search engines. Elasticsearch is powered by Apache Lucene which is also a search engine for databases that have low-level queries. Elasticsearch has a query that is easier to use because it is based on RESTful.

Elasticsearch has a concept that is quite unique. Users can assume indexes as databases, types as tables and documents as records or rows. Whereas mapping can be assumed as a "table scheme". In Elasticsearch there are no transactions, so that the engineer could create an index structure according to user needs. Moreover, it can be arranged to be a distributed system for a number of servers.

Logstash is an open source software for collecting and parsing logs and also making indexes for logs, then stored in Elasticsearch. Kibana is a web interface that is useful for displaying logs both graphically and in other visualization type. File agent need a “filebeat” that is useful for sending logs from each server to logstash. A centralized log is very useful if one day a DevOps will identify a problem with the server or application. This makes them able to search historical data or logs. ELK component explained as logstash which is processing the log and indexes the log, elasticsearch which is storing all logs, kibana which is the web interface for searching and visualizing logs properly and displaying them in the desired graphic form. ELK diagram shown in Figure 1.

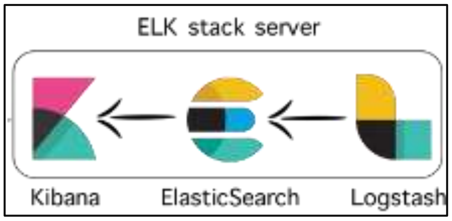


Fig. 1. *ELK Diagram*

3. Methodology

3.1 Installation and Preparation

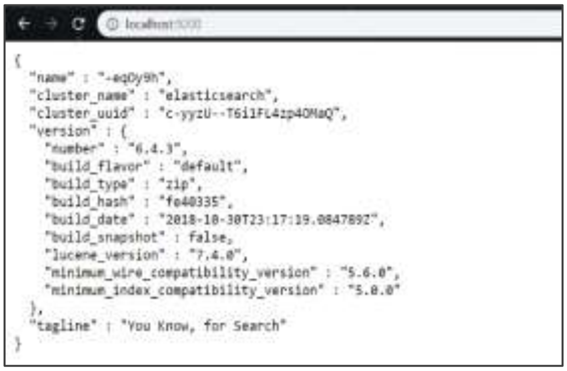
ElasticSearch, Logstash, and Kibana uses Java environment as a platform. The first process must introduce is ElasticSearch installation.

3.1.1 ElasticSearch Installation

1. Download and open the ElasticSearch file from the available site.

2. Run bin/ElasticSearch from Lmux operating system (or bin\elasticsearch.bat on Windows operating system).

3. Run curl http://localhost:9200 or invoke-RestMethod http://localhost:9200 using PowerShell. Figure 2 shows example of ElasticSearch running environment.



```
{
  "name" : "-eqOy9h",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "c-yyzU--T6i1FL4zp4OMaQ",
  "version" : {
    "number" : "6.4.3",
    "build_flavor" : "default",
    "build_type" : "zip",
    "build_hash" : "fe40335",
    "build_date" : "2018-10-30T23:17:19.084789Z",
    "build_snapshot" : false,
    "lucene_version" : "7.4.0",
    "minimum_wire_compatibility_version" : "5.6.0",
    "minimum_index_compatibility_version" : "5.0.0"
  },
  "tagline" : "You Know, for Search"
}
```

Fig. 2. *http://localhost:9200 Environment*

3.1.2 Logstash Installation

1. Download and open the Logstash file from the site.

2. Prepare a config file (logstash.conf). As an example that is done to index the file "terrorist.csv". Then the index created in the config file in logstash is:

```
input {
file{
path => "D:/gtd/teroris.csv"

start_position => "beginning"
sincedb_path => "NUL"
}
 beats {
    port => 5044
}

filter {
csv {
separator => ","
columns                                            =>
   ["eventid","iyear","country","country_txt","regi
   on","region_txt","city","latitude","longitude","
   success","suicide","attacktype1","attacktype1_tx
   t","targtype1","targtype1_txt","gname","weaptype
   1","weaptype1_txt"]
}
…
mutate{rename => {
"longitude"=>"[location][lon]"
"latitude"=>"[location][lat]"
}
}
  date {
          match => ["iyear", "yyyy"]
      }
  }
    output {
      elasticsearch {
      hosts => ["localhost:9200"]
      index => "teroris"
      document_type => "global"
      }
      stdout { codec => rubydebug }

}
```

3. Running command bin/logstash -f logstash.conf

The browser tab will display the contents of the big data that has been indexed before. Figure 3 shows the browser menu of ElasticSearch Head.

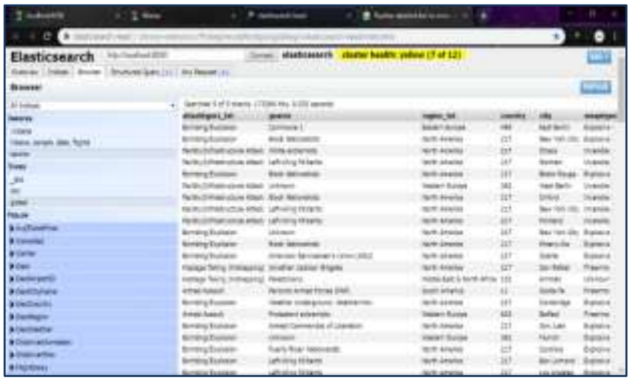


Fig. 3. *Browser menu in Extensions Elasticsearch Head.*

In the "terrorist.csv" data there is a latitude and longitude column which is allowed to be displayed in the form of maps (coordinates

3.1.3 Kibana Installation

1. Download and open the Kibana file from the site.
2. Open config/kibana.yml in the editor.
3. Link elasticsearch.url to points as did Elasticsearch.
4. Run bin/kibana or bin\kibana.bat on Windows.
5. Enter the access point in the browser at http:// localhost:5601
6. To visualize big data in Kibana, the menu that will be used is the "visualize" menu. In Kibana there are various graphs that can present data to be visualized. Figure 4 shows the menu view in Kibana.

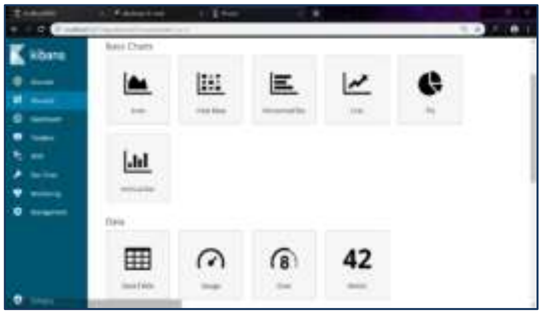


Fig. 4. *Kibana view http://localhost:5601*

4. Result

4.1 Visualization using Heat Map Graphic

By using the Heat Map chart, we can display data on the map (maps) by utilizing the latitude and longitude coordinates that exist in the data in accordance with the coordinates that exist on the map. Figure 5 shows the latitude and longitude in Kibana view.

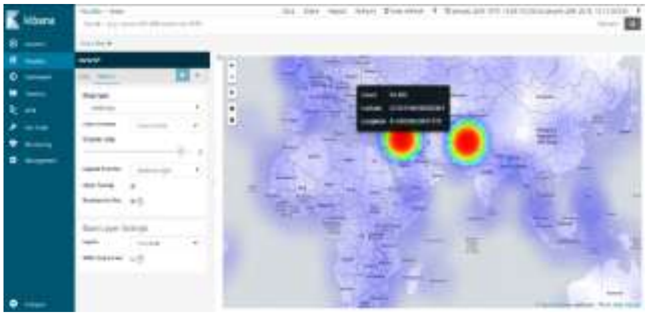


Fig. 5. *Heat Map Graphic*

The regions that experienced the most terrorism incidents (red to green scale) are shown in red, namely at latitude = 32.851246280 and longitude = 41.0503952 with 38,450 events from the total data counted from 1 January 1970 to 1 January 2017. With Thus it can be concluded that the area that is prone to experiencing terrorism is the Middle East or Iraq.

4.2 Visualization using Heat Map Graphic

By using a Vertical Bar chart we can display the highest amount of data based on variations in bar size.

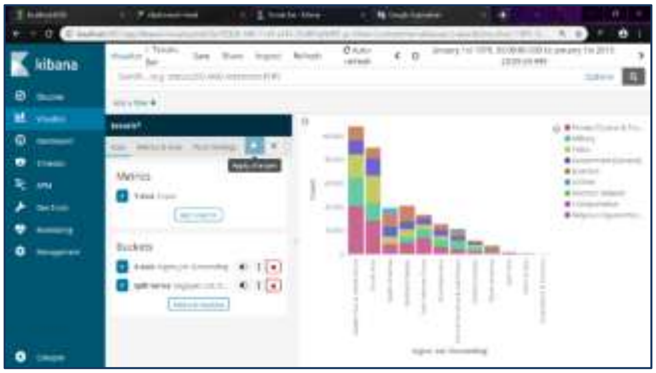


Fig. 6. *Vertical Bar Graphic*

Figure 6 shows the region that experienced the most terrorism incidents, namely "Middle East & North Africa" is shown with the largest bar size and with the most targeted terrorism targets is "private citizen & property" (citizens & private property). So it can be concluded that the motives and objectives of terrorism activities are to harm citizens and facilities of the country.

By using the Pie chart we can display the most data based on variations in the size of the pie / pie chart.

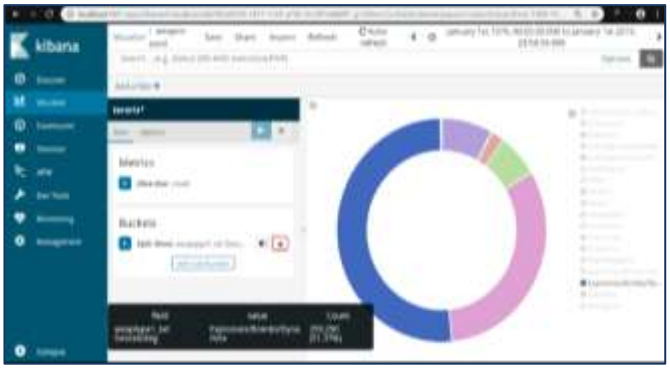


Fig. 7. *Pie graphic visualization*

From Figure 7 it can be seen that the most widely used weapons in acts of terrorism are "Explosives / Bombs / Dynamite" with a total of 259,290 or around 51.37% of the total terrorism events. So it can be concluded that the use of bombs and dynamite has the greatest chance of damage in the event of terrorism.

By using the "Horizontal" graph we can display the most data based on variations in the size of the bar chart.

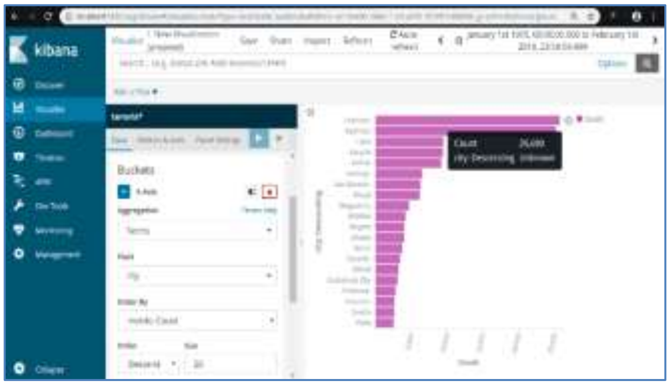


Fig. 8. *Horisontal graphic visualization*

From Figure 8 the graph can be seen the cities that have experienced the most terrorism incidents. So it can be concluded that the city that experienced the most terrorism incidents is the city of Baghdad. About 25,000 more events have occurred in the city.

By using a Goal chart we can display data with different achievements. Use logic 1 for true events and logic 0 for false events. In the data "terrorist.csv" this type of graph can be used to present data on terrorism incidents with suicide motives (logic 1) and those without suicide motives (logic 0).

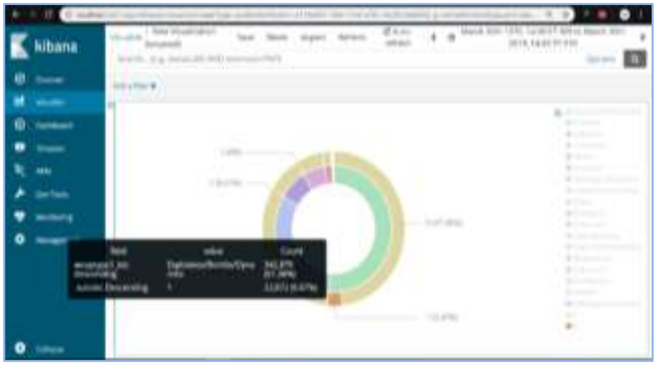


Fig. 9. *Goal graphic visualization*

In Figure 9, it can be seen that the combination of graphs of the use of weapons such as bombs, explosives or dynamite with the motives of the events behind them. The incidence of terrorism with suicide motives that have occurred is as many as 22,872 or about 6.67% of the total incidents. So it can be concluded that suicide bombing is not the most common motive for terrorism.

By using a metric chart the user can display totals from different data. Logic 1 is for true events and logic 0 is for false events. In "terrorist.csv" data this type of graph can be used to present the total number of successful terrorist incidents (logic 1) and the total number of terrorism incidents that can be foiled (logic 0).

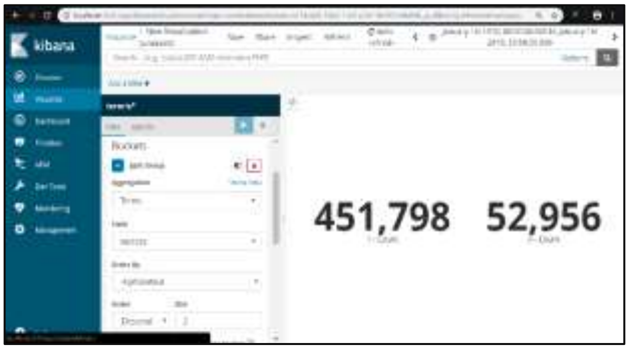


Fig. 10. *Metric visualization*

In Figure 10, it can be seen that the total number of terrorism incidents that have occurred is 451,798 and the number of terrorism events that can be foiled is 52,926. So it can be concluded that the number of terrorism incidents that occurred is difficult to anticipate beforehand.

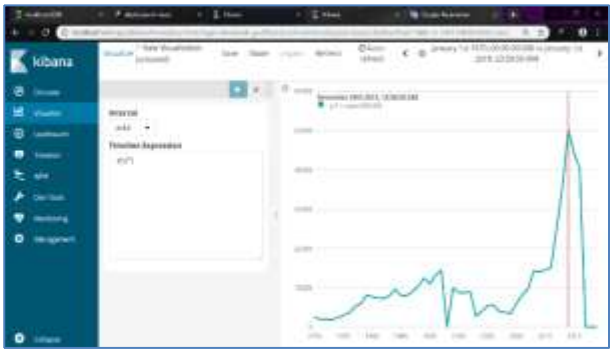


Fig. 11. *Timelion data visualization*

From Figure 11, it can be seen that there are fluctuations from the data on terrorism incidents and the highest peak of terrorism incidents that have occurred, namely in the period 2010-2015 precisely in December 2013 amounted to 50,253 events.

5. Conclusion

The conclusions obtained from the big data visualization that have been carried out are as follows:

1. In visualizing data using Elasticsearch, Logstash, Kibana is quite easy to use.

2. The use of various charts can be adjusted to the data to be visualized.

3. Big data that has latitude and longitude columns allows to be visualized in the form of maps.

4. The combined graphic display makes it possible to make the information conveyed more complex.

Reference

[1] Eason, G., Noble, B., & Sneddon, I. N. (1995). On certain integrals of Lipschitz-Hankel type involving products of Bessel functions, Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551.

[2] Maxwell, J. C. (1892). A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, pp.68-73.

[3] Nicole, R. (2016). Title of paper with only first word capitalized, Journal Name Stand. Abbrev., in press.

[4] Media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] Young, M. (1989). The Technical Writer's Handbook. Mill Valley, CA: University Science.