International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

Data Analytics with Elk Stack and Custom Dashboard

Shetty Rasik Ramchandra[1], Shetty Sanath Santosh[2], Shetty Veekshith Krishna[3], Parikshith Adiga[4]

[1,2,3,4]*Student, Dept. of Computer Science and Engg., Alva's Institute of Engineering and Technology, Mijar, India*

***Abstract*: The project consists of two parts or problem statements 1. Increasing elastic search performance through creating compressed indices from daily indices in the production environment. 2. Creating a custom dashboard application for generating statistics based on live running application name. To increase the efficiency of Elastic search, number of shards in the Elastic search need to be reduced. In order to do this, we compress the Elasticsearch indices. It is done by using Python libraries for Elasticsearch. We write python scripts and merge the indices together to reduce the number of indices which are produced on a daily basis in a production environment. Another part of a project is to provide visualization for application wise usage statistical data, this application can be an alternative of Kibana, the user can select the applications and the time-period from the dashboard for visualization in the form graph or table. The visualization can be downloaded in the form of pdf or image**

***Keywords*: Elastic search, logstash, kibana, Data Analytics, python, api.**

1. Introduction

Log management system deals with large volumes of computer generated log records (also called as audit records, audit trails, event-logs, etc.). The process involves log collection, centralized aggregation, long-term retention, log analysis, log searching and reporting. A stream of messages in time sequence often contains log entity. Logs are either stored on a disk or supplied as a data stream to a log collector. Logs are being generated in order to help in debugging the operations in an application. The semantics and syntax of data in log messages are usually application or system specific. Log analysis systems must decode messages within the context of an application or system in order to make meaningful comparisons for messages within different log files. As every computing device is generating logs and if only web server logs case is considered then also it amounts to a very large data set. The analysis of these logs not only helps the companies in their decision making but also in improving their products and services. However, most of the small-scale companies cannot afford these log management systems. Another problem is the usability of the log management products as most of the available products require technical expertise in-order to process and interpret the logs data. In a nutshell, the three critical problem components for log analysis are Volume, Cost and Usability.

Elastic search is capable in indexing and searching. Elasticsearch has search and list information accessible REST fully as JSON over HTTP and can easily scale. It's under the Apache 2.0 permit and is based on top of Apache's Lucene project. Elasticsearch is a content indexing web search tool. The best representation to show Elasticsearch is the pages of a book. As we flip to the back of a book, and look for a word and afterward discover the reference page. This implies that instead of looking at the content strings straightforwardly, Elasticsearch makes an index of the content and it looks to the indexes as opposed to the contents. Thus, it is quick. Logstash is a tool for managing events and logs. It is used to collect logs, parse them, and store them for later use (like, for searching). If we store them in Elasticsearch, we can view and analyze them with Kibana.

The Kibana web interface is an adjustable dashboard that we can stretch out and change to suit our surroundings. It permits tables and diagrams creation and in addition complex representations. The Kibana web interface utilizes the Apache Lucene question linguistic structure to permit us to make inquiries. We can seek any of the fields contained in a Logstash, for instance, message, system log_program, and so on. We can utilize Boolean rationale with AND, OR, and NOT.

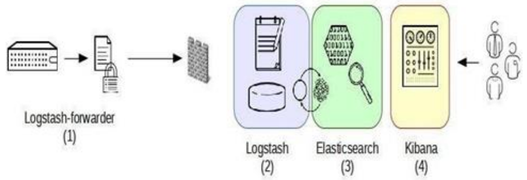


Fig. 1. ELK ecosystem

Data Analytics is the process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions and supporting decision making.

Python api of ELK stack is Elasticsearch - PY and its official low-level Python client for Elastic search. Why elastic search-py is cause of its Integrating Elasticsearch as a data storage and search component into your Python dominant infrastructure. Indexing data without worrying about translation of basic Python data types to json. Load balancing across all the Elastic search nodes and its thread safety. This paper demonstrates the application of ELK stack in Data analytics. As ELK stack is an open source technology, thus it doesn't involve paying high licensing fees.

2. Related work

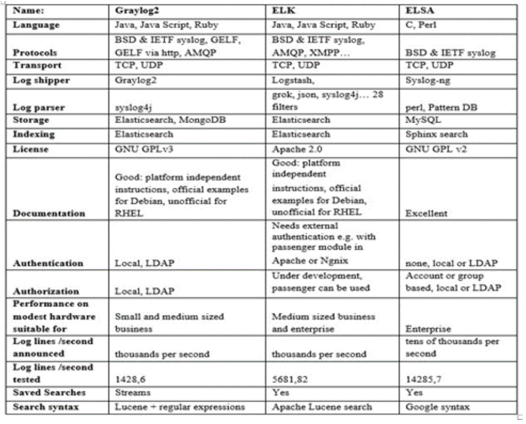
Various development works has been going on since the last decade to build log analytics systems. Artyom [1] did a comprehensive analysis on popular open source log management tools such as Graylog2, ELK, and ELSA. His study was based on performance testing, which showed ELSA was fastest and was able to handle 14000 logs per second but ELK secured impact on usability factor. Christopher [2] in his blog analyzed Yahoo's historical stock data [3] using ELK stack and build the dashboard showing the volume of stocks, if Average high and Average low. Aarvik [4] build a real time data analytics tool and tested logs of his Debian Operating system.

Tanmay [5] integrated apache log data with GeoCity [6] data, which transforms IP address into exact locations with attributes like longitude, altitude, city, state and country. William [7] in his blog post used Docker [8] container and deployed ELK stack in order to monitor infrastructure.

Jettro [9] In his project post scanned directory structure of jpg files and extracted meta data from images, utilizing the ELK stack he build a dashboard which shows details about aperture, focal length and iso. Bogdan [10] build a video card monitoring solution using ELK stack to have the information about the performance when gaming. In a real python [11] blog a real time twitter sentiment analysis dashboard utilizing Elasticsearch and Kibana was built.

Below Table 1 shows a comparative study of GrayLog2, ELK and ELSA.

Table 1
Feature Overview of Graylog2, ELK, ELSA [1]



Name:	Graylog2	ELK	ELSA
Language	Java, Java Script, Ruby	Java, Java Script, Ruby	C, Perl
Protocols	BSD & IETF syslog, GELF, GELF via http, AMQP	BSD & IETF syslog, AMQP, XMPP…	BSD & IETF syslog
Transport	TCP, UDP	TCP, UDP	TCP, UDP
Log shipper	Graylog2	Logstash,	Syslog-ng
Log parser	syslog4j	grok, json, syslog4j… 28 filters	perl, Pattern DB
Storage	Elasticsearch, MongoDB	Elasticsearch	MySQL
Indexing	Elasticsearch	Elasticsearch	Sphinx search
License	GNU GPLv3	Apache 2.0	GNU GPL v2
Documentation	Good: platform independent instructions, official examples for Debian, unofficial for RHEL	Good: platform independent instructions, official examples for Debian, unofficial for RHEL	Excellent
Authentication	Local, LDAP	Needs external authentication e.g. with passenger module in Apache or Ngnix	none, local or LDAP
Authorization	Local, LDAP	Under development, passenger can be used	Account or group based, local or LDAP
Performance on modest hardware suitable for	Small and medium sized business	Medium sized business and enterprise	Enterprise
Log lines /second announced	thousands per second	thousands per second	tens of thousands per second
Log lines /second tested	1428,6	5681,82	14285,7
Saved Searches	Streams	Yes	Yes
Search syntax	Lucene + regular expressions	Apache Lucene search	Google syntax

3. Methods and methodologies

A. *ELK Stack Working Overview*

- *Step 1:* Access logs – Apache access logs having the customized log formats.
- *Step 2:* Filebeat – Read Apache log lines from client servers and send them to Web Analytics Server
- *Step 3:* Logstash – Parses the log lines send by File beat, processes it to the corresponding fields and pushes data to elastic search
- *Step 4:* Kibana – Displays in different formats (charts, maps etc.) the search results from elastic search database

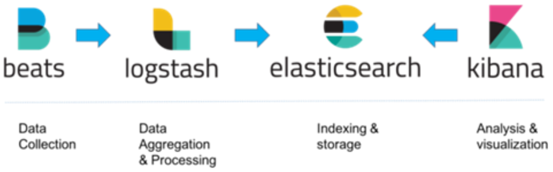


Fig. 2. ELK flowchart

In this study centralized ELK server was set up and File beat is used to send the logs to the ELK server.

- *File beat:* It is a lightweight shipper for forwarding and centralizing log data. Installed as an agent on your servers, File beat monitors the log files or locations that you specify, collects log events, and forwards them to either to Elasticsearch or Logstash for indexing.
- *Logstash:* Its is an open source data collection engine with real- time pipelining capabilities. Logstash can dynamically unify data from disparate sources and normalize the data into destinations of your choice. Cleanse and democratize all your data for diverse advanced downstream analytics and visualization use cases. While Logstash originally drove innovation in log collection, its capabilities extend well beyond that use case. Any type of event can be enriched and transformed with a broad array of input, filter, and output plugins, with many native codecs further simplifying the ingestion process. Logstash accelerates your insights by harnessing a greater volume and variety of data.
- *Elastic search:* Its is a search engine based on the Lucene library. It provides distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. Elasticsearch is developed in Java. Elasticsearch divides indexes in physical spaces called shards. They allow you to easily split the data between hosts, but there's a drawback as the number of shards is defined at index creation.
- *Kibana:* It is an open source analytics and visualization platform designed to work with Elasticsearch. You use Kibana to search, view, and interact with data stored in Elasticsearch indices. You can easily perform advanced data analysis and visualize your data in a variety of charts, tables, and maps.

4. Experimental set up

For the purpose of this study - Elasticsearch -1.6.0, Logstash-

1.5.1, Elastic search-py python module,c3.js,pdf.js

Logs may be GeoIP, Request url, Referrer url, Response time, Response status

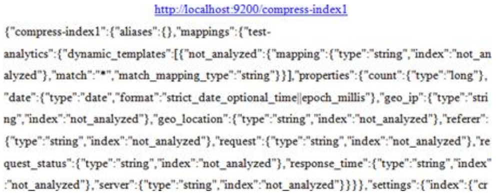
Elastic Search-py: it is a RESTFUL API and supports the CRUD operations (Create, Read, Update, Delete) over the HTTP without any client i.e. you can get the data using command-line tool (i.e. curl), or simply via your Internet browser, for example:curl -XGET 'http://localhost:9200/dummydata-*/_search?pretty'. It will return all the indexes have their name starting with ‘dummydata-‘in JSON data format.

5. Experimentation

First, we connect to the elastic search package in python which provides different modules to work with the elastic search engine. We read the date from elastic search indices which are obtained by the access log and through file beats and we create a new index to store the data in compressed format. Out of each index that is created for a day, we create a new index which stores the data of entire month. And once the data has compressed and stored using other python script we get the data according to the application and store it in csv format and the csv files loaded to dashboard using papa parse and using c3.js library we ge graphical representation of the csv data.

6. Results and discussions

ELK as a log management system is very useful in terms of usability as anyone can interpret the results and based on that they can have some insights. Below screen shows the Cluster health monitoring through a GUI. The figure 3 shows the compressed indexs. And in figure 4 Here we can see that number of ip’s particularly used on that specific date this data.



http://localhost:9200/compress-index1

{"compress-index1":{"aliases":{},"mappings":{"test-analytics":{"dynamic_templates":[{"not_analyzed":{"mapping":{"type":"string","index":"not_analyzed"},"match":"*","match_mapping_type":"string"}}],"properties":{"count":{"type":"long"},"date":{"type":"date","format":"strict_date_optional_time||epoch_millis"},"geo_ip":{"type":"string","index":"not_analyzed"},"geo_location":{"type":"string","index":"not_analyzed"},"referer":{"type":"string","index":"not_analyzed"},"request":{"type":"string","index":"not_analyzed"},"request_status":{"type":"string","index":"not_analyzed"},"response_time":{"type":"string","index":"not_analyzed"},"server":{"type":"string","index":"not_analyzed"}}}},"settings":{"index":{"cr

Fig. 3. Compressed index’s

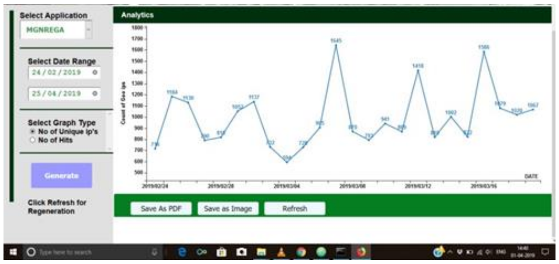


Fig. 4. Graphical representation of data

In this paper we performed the Compression of Index’s based on the access logs using open source stack ELK and By the use python module we compressed the indexs to one particular date and converted this index values to the csv files according to application wise and visualized the data for analytics, we have used a small log data set just to demonstrate the usability of ELK stack but the real usefulness of ELK stack will be on Big data sets as without much of installation steps and being the product from open source community, it is cost effective and had huge active contributor base which makes it more competitive as compared to other products for log management activities. Moreover, ELK stack can be used for tracking fraudulent activities, security breaches and it looks promising for Internet of things monitoring which is the future. ELK is much effective in building real time dashboard for analyzing tweets and other analytics.

7. Conclusion

The Project successfully helps in increasing the efficiency of the Elasticsearch by compressing the indices. All the indices are compressed properly and all the data items are included. The single index created contains all the data and is well suited for data storage. The Custom based dashboard works according to the user application and visualize the data based on the user aspirations. The dashboard provides a good alternative to the default dashboard in the ELK Stack. The Elasticsearch efficiency can be further increased by using different services to improve the congestion control. The services include apache kafka and other related services.

References

[1] Artyom Churilin, "Choosing an open-source log management system for small business," Master’s Thesis, Faculty of Information Technology, Tallin University of Technology, Tallinn, Estonia.

[2] Christopher, “Visualizing data with Elasticsearch, Logstash and Kibana,” Available: http://blog.webkid.io/visualize- datasets-with-elk/

[3] Yahoo Stock history data, Available: http://finance.yahoo.com

[4] Anders Aarvik (2014, April 04), “A bit on ElasticSearch + Logstash +Kibana (The ELK stack), Available:http://aarvik.dk/a-bit-on-elasticsearch-logstash-kibana-the-elk-stack/

[5] Tanmay Deshpande (2015, May 10), “Log Analytics using Elasticsearch, Logstash and Kibana Available:http://hadooptutorials.co.in/tutorials/elasticsearch/log-analytics-using-elasticsearch-logstash- kibana.html#

[6] GeoLite Legacy Downloadable Database Avaialble:http://dev.maxmind.com/geoip/leg acy/geolite/

[7] William Durand (2014, December 17), “Elasticsearch, Logstash & Kibana with Docker.” Available: http://williamdurand.fr/2014/12/17/elasticse arch-logstash-kibana-with-docker/

[8] DockerAvailable: https://www.docker.com/

[9] Jettro Coenradie (2013, November 28), “Use Kibana to analyze your images,” Avaialble: https://blog.trifork.com/2013/11/28/ use-kibana-to-analyze-your-images/

[10] Bogdan Dumitresc (2014, January 28), “Using Logstash, Elasticsearch and Kibana to monitor your video card Avaialble:http://blog.trifork.com/2014/01/28/using-logstash elasticsearch-and-kibana-to-monitor-your-video- card-a-tutorial/

[11] Real Python (2014, November 13), “Twitter Sentiment-Python, Docker,Elasticsearch, Kibana.” https://realpython.com/blog/python/twitter-sentiment-python-docker-elasticsearch-kibana/

[12] L.K. Joshila Grace, V. Maheswari, and Dhinaharan Nagamalai, “Web Log Data Analysis and Mining,” in Proc CCSIT-2011, Springer CCIS, Vol. 133, pp. 459-469, 2011.