Journal of Engineering and Applied Sciences 12 (Special Issue 12): 9451-9454, 2017
ISSN: 1816-949X

An Aggregation Approach Based on Elasticsearch

[1]Sayar Ahmad Kuchy, [1]Syed K. Ahmed Khadri, [2]Manoj Mukherjee,
[3]Debabrata Samanta and [4]Dac-Nhuong Le
[1]ITC Infotech, Bangalore, India
[2]William O'Neil India, Bengalura, India
[3]DSCASC, Bangalore, India
[4]Faculty of Information Technology, Haiphong University, Haiphong, Vietnam

Abstract: Our focus here is to elaborate the functionality of the elasticsearch. To determine how it is having an edge over the previous demanding and very successful search engines like Lucene and solr search. A search engine which provides an advanced edge over the existing search engines by way of providing multiple features like analytical reporting and logging distribution, etc. in addition of the search results. Elasticsearch uses the regression mathematical model through which it is getting an edge over the rest of the search engines. Elasticsearch is having a great future in hand as it provides multi-dimensional approach in summarizing and analysing of huge data components. In this research study, we are trying to focus on the possibility of extending the feature of self-learning capability of elasticsearch engine.

Key words: Aggregation, bucket, facet filters, inverted indexing, Kibana, Logstash, Lucene, machine learning, search engine, solr, X-packs

INTRODUCTION

In 2004, Shay Banon created compass. Later, while thinking about that he can create a much better format of a new search engine. So, he used the Java as the technology and kept in mind the availability of output data format, hence, JSON is chosen (Pokorny, 2013; Allen, 2015). The first version of elasticsearch came into picture in 2010. Elasticsearch is most famous search engine right now in the market (Bhupathiraju and Ravuri, 2014). Elasticsearch is a Java based open source search engine under the terms of Apache license. Elasticsearch provides full text search, distributed and multitenant. It has an HTTP web interface. Elasticsearch provides the JSON output formatted results.

MATERIALS AND METHODS

Elasticsearch for analytics: Elasticsearch uses aggregation as a great methodology to have best in place results. Also, get a differential edge on top of other search engine available in market (Wu *et al*., 2014; Burger *et al*., 2011). There are various types of analytic reports prepared by the elasticsearch with the help of Logstash and Kibana integrations. Algorithm 1, demonstrate with sample code of elasticsearch.

Algorithm; Sample code of ealsticsearch:

Structure of aggregation-Code snippet

```
"aggregations": {
    "content_inserted": {
      "buckets": [          {
         "key": "HANA"
        "doc_count": 10
      },          {
 {
  "key": "clea"
  "doc_count": 25
},          {
  "key": "s4-HANA"
  "doc_count": 29
   "key": "mobile_apps"
    "doc_count": 21}
           {
          ]
      }    }
```

In this study, Fig. 1 describes elastic search full working package. Here are some of the analytical approches given by elasticsearch. Aggregations are very much powerful as many organizations have achieved best results after they build elasticsearch clusters for analytics

Fig. 1: Elasticsearch full working package

Corresponding Author: Sayar Ahmad Kuchy, ITC Infotech, Bangalore, India

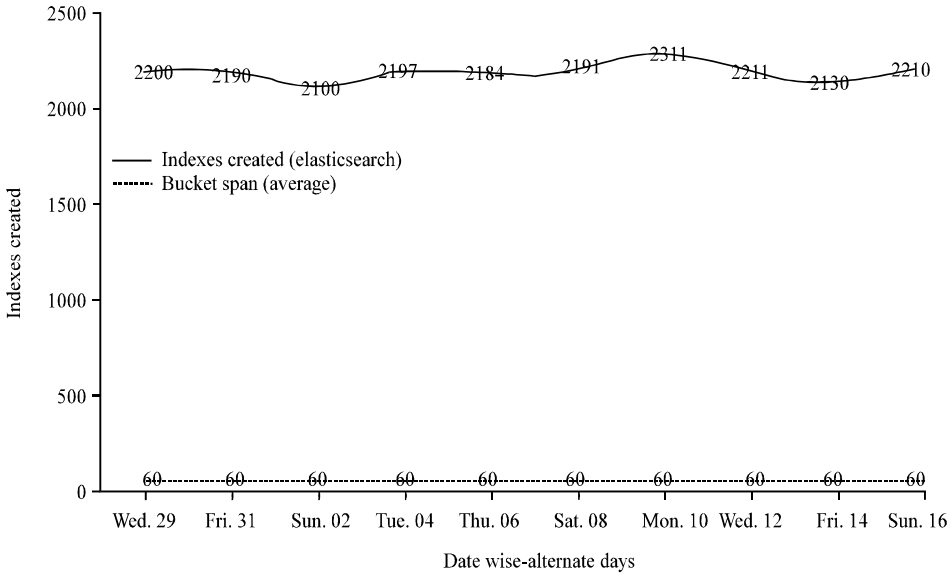

Fig. 2: Output data representation based on Table 1 input

Table 1: Input data to elasticsearch

Content inserted (No. of characters)	Indexes created (elasticsearch)	Bucket span (average m)	Date wise (alternate days)
5600-7800	2200	60	Wed. 29
5610-7800	2190	60	Fri. 31
5630-7800	2170	60	Sun. 02
5700-7800	2100	60	Tue. 04
5603-7800	2197	60	Thu. 06
5616-7800	2184	60	Sat. 08
5609-7800	2191	60	Mon. 10
5489-7800	2311	60	Wed. 12
5589-7800	2211	60	Fri. 14
5670-7800	2130	60	Sun. 16
5590-7800	2210	60	Wed. 29

(Cha *et al*., 2010; Paul *et al*., 2011; Khadri *et al*., 2013a, b). Table 1 describe the different types of input data to elasticsearch. Figure 2 shows out put data representation based on Table 1.

Production resource management for elasticsearch: Running our cluster in Productive environment, we need to take care of certain important points:

- Configurations gateways settings which are needed to be made at productive setup
- Hardware requirements and the recommendations of implementing elasticsearch guidelines for deployment at productive environment
- After deployment of cluster at production please take care of security, backups and indexing maximization

Performing of elasticsearch: It has been determined that without elasticsearch usability the time taken by the search operation was 11.5 msec and after using elasticsearch the same search resulted output came in 1.5 msec. Elasticsearch does not support SQL query type formats. It is having very uncommon type of data querying language (Khadri *et al*., 2013 a, b, 2014a-c). We need to know first tokenizes, analyzers to put any query for storage and retrieval of data. Here in case of elasticsearch everything is indexed in first hand or by default. Which in turn causes an extra index burden/load? The only output data format supported is JSON. Wordlist based spell checking is not available at the search input block (Wang and Tang, 2012). In elasticsearch, it is not possible to increase or decrease the number of primary shreds once the indexing part is completed.

RESULTS AND DISCUSSION

Applications of elasticsearch: Basically, it is based on our requirements that how our search should provide us results. Same is getting fulfilled by this search. We think

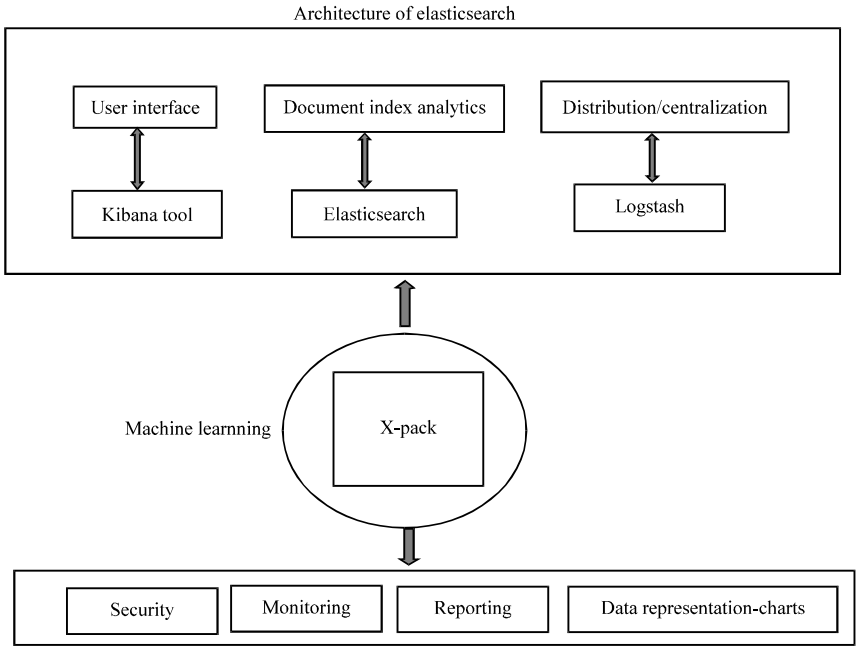

Fig. 3: Elasticsearch hierarchical structure

of a requirement and approach elasticsearch-resulted track is defined for you! Figure 3 describes the hierarchical structure of elasticsearch depends on machine learning (Kononenko *et al*., 2014). Some of the basic application uses of Elasticsearch areas.

Content search: Elasticsearch is a very fast text searching engine especially when we have a huge amount of distributed content either in the form of normal plain text or file content meaning document files. Elasticsearch provides quick results compared to other searches currently available in the market by way of how the 'inverted index' works. And the ways of how we use the dictionary and posting lists.

Analytics: There are ample analytical approaches available in elasticsearch which can help our organizations to grow our revenue multiple folds, i.e., marketing analytics, business analytics, customer feedback analytics, security analytics, log analytics. To overcome some of the limitations of facet searches elasticsearch brings the concept of new aggregation engineering. A huge data-maybe a 'big data' is summarized and analyzed by using elasticsearch aggregations. Which are getting generated by Kibana like tools. Kibana is a tool which provides a graphical view of the elasticsearch data-summarized one like tools.

Corporate or e-Commerce search: Elasticsearch is very much suitable for the product search in case of ecommerce applications. Example: when a customer is looking for buying a product Elasticsearch's analytical responsiveness provides relevant products which the buyer is interested based on the predictive data analysis (Bakshy *et al*., 2011).

Not only provides the normalized data but also the variations of price and features of the product comparison in an awesome way.

Machine learning: Elasticsearch with the integration of X-pack provides features of a normal learning by its own on the behavior of the data. Tremendous results of unsupervised machine learning outcomes have been obtained (Wijaya *et al*., 2013; Monarizqa *et al*., 2014).

This machine learning capability helps us in easily detecting the anomalies. Which intern help us in taking an action on time on top of those anomalies (Sakaki *et al*., 2010).

Multitenancy: User specific data search capability. Here, preference is given first to the users own specific data documents.

CONCLUSION

By using the aggregation, we came in to conclusion that elasticsearch has made itself a renowned and top class search engine. Elasticsearch uses a few wonderful technical advanced extensions Kibana, Logstash and x-pack. In most of its extensions it approaches the

behavior of the aggregation mathematical models. In the beginning of the era of searching concept. People used to put query on the top of indexed database tables and get the results. Then the solr search engine with facet features came into picture. Here, the facets drill down or carry forward search querying was not possible. The link was getting lost in between. Elasticsearch provides great features which one organization must have to be in the advancement race of technology. Now, the much-awaited search engine-elasticsearch is available which helps us in getting carry forward the facet drill down functionality, to reach to a maximum optimistic search result. By the implementation of the aggregation model it is possible to achieve above said functionality and provide value added feature of analytics and logging distribution. Be it the summarization, analytic reporting, distributed and communicating of logged data or be it the most exciting machine learning opportunistic facilities. All the said services/features are offered by the elasticsearch. The sample code of elasticsearch. The elasticsearch full working package. The different types of input data to elasticsearch. We got the output from the input data. The hierarchical structure of the elasticsearch with respect to machine learning.

REFERENCES

Allen, M., 2015. Relational databases are not designed for scale. MarkLogic, San Carlos, California. http://www.marklogic.com/blog/relational-databases-scale/.

Bakshy, E., J.M. Hofman, W.A. Mason and D.J. Watts, 2011. Everyone's an influencer: Quantifying influence on Twitter. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, February 09-12, 2011, ACM, Hong Kong, China, ISBN:978-1-4503-0493-1, pp: 65-74.

Bhupathiraju, V. and R.P. Ravuri, 2014. The dawn of big data-HBase. Proceedings of the 2014 Conference on IT in Business, Industry and Government (CSIBIG'14), March 8-9, 2014, IEEE, Indore, India, ISBN:978-1-4799-3065-4, pp: 1-4.

Burger, J.D., J. Henderson, G. Kim and G. Zarrella, 2011. Discriminating gender on Twitter. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011, Association for Computational Linguistics, Edinburgh, UK., ISBN:978-1-937284-11-4, pp: 1301-1309.

Cha, M., H. Haddadi, F. Benevenuto and P.K. Gummadi, 2010. Measuring user influence in Twitter: The million follower fallacy. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, George Washington University, Washington, D.C., USA., pp: 10-17.

Khadri, S.K.A., D. Samanta and M. Paul, 2013a. Message communication using Phase Shifting Method (PSM). Intl. J. Adv. Res. Comput. Sci., 4: 9-11.

Khadri, S.K.A., D. Samanta and M. Paul, 2013b. Secure approach for message communication. Intl. J. Adv. Res. Comput. Commun. Eng., 2: 3481-3484.

Khadri, S.K.A., D. Samanta and M. Paul, 2014c. Approach of message communication using Fibonacci series: In cryptology. Cryptology Notes Inf. Theor., 2: 168-171.

Khadri, S.K.A., D. Samanta and M. Paul, 2014b. Message encryption using text inversion plus n count: In cryptology. Intl. J. Inf. Sci. Intell. Syst., 3: 71-74.

Khadri, S.K.A., D. Samanta and M. Paul, 2014a. Novel approach for message security. Intl. J. Inf. Sci. Intell. Syst., 3: 47-52.

Kononenko, O., O. Baysal, R. Holmes and M.W. Godfrey, 2014. Mining modern repositories with elasticsearch. Proceedings of the 11th Working Conference on Mining Software Repositories, May 31-June 01, 2014, ACM, Hyderabad, India, ISBN:978-1-4503-2863-0, pp: 328-331.

Monarizqa, N., L.E. Nugroho and B.S. Hantono, 2014. [Application of sentiment analysis on Twitter speak Indonesia as a Ratinger (In Indonesian)]. J. Electr. Eng. Res. Inf. Technol., 1: 151-155.

Paul, M., D. Samanta and G. Sanyal, 2011. Dynamic job scheduling in cloud computing based on horizontal load balancing. Intl. J. Comput. Technol. Appl., 2: 1552-1556.

Pokorny, J., 2013. NoSQL databases: A step to database scalability in web environment. Intl. J. Web Inf. Syst., 9: 69-82.

Sakaki, T., M. Okazaki and Y. Matsuo, 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web, April 26-30, 2010, ACM, Raleigh, North Carolina, ISBN:978-1-60558-799-8, pp: 851-860.

Wang, G. and J. Tang, 2012. The NoSQL principles and basic application of cassandra model. Proceedings of the 2012 International Conference on Computer Science & Service System (CSSS'12), August 11-13, 2012, IEEE, Nanjing, China, ISBN:978-1-4673-0721-5, pp: 1332-1335.

Wijaya, H., A. Erwin, A. Soetomo and M. Galinium, 2013. Twitter sentiment analysis and insight for Indonesian mobile operators. Proceedings of the International Conference on Information Systems (ISICO'13), December 10, 2013, Institute of Technology Blanchardstown, Dublin, Republic of Ireland, pp: 2-4.

Wu, X., X. Zhu, G.Q. Wu and W. Ding, 2014. Data mining with big data. IEEE Trans. Knowledge Data Eng., 26: 97-107.