Processamento Massivo de Dados na Expertise PMD

Eduardo Germano[1]

[1]Especialização em Data Science
UniRitter – Brazil

eduar.germano@gmail.com

A Expertise PMD é uma das maiores empresas de desenvolvimento de soluções de tecnologia no Brasil. Esta empresa passa por um grande crescimento e tem como objetivo aprimorar seus serviços, no que se refere a gerência e processamento massivo de dados em *clusters* para então, oferecer aos seus clientes orientações na tomada de decisão em diferentes áreas. Como consultor de tecnologia da Expertise PMD, tenho como responsabilidade auxiliar na construção de uma estratégia que contribua com o objetivo da empresa.

Existem muitas ferramentas para processamento massivo de dados, das quais podemos citar Hadoop, Spark [Hazarika et al. 2017], Hive, Pig [Fuad et al. 2014], Yarn, Zookeper [Frampton 2015]. Além disso, a cada dia surgem novas ferramentas com grande potencial para resolução de problemas em ambientes distribuídos. Após um tempo entendo o cenário da Expertise PMD, propomos a utilização de uma *stack* que vem ganhando bastante notoriedade nos últimos anos, a ELK Stack [Chhajed 2015]. Esta *stack* para processamento massivo de dados nasceu como ElasticSearch, onde o foco era a busca e análise de dados distribuídos. Desenvolvido sobre o Apache Lucene [Lucene 2010], a Elastic incorporou novos componentes ao seu produto, e o renomeou para ELK Stack.

A ELK Stack se tornou popular por possuir APIs de simples utilização. Além disso, esta *stack* foi concebida para atuar de forma escalável, veloz e distribuída [Son and Kwon 2017]. Apesar do ElasticSearch ser o componente principal da ELK Stack, ela também possui ferramentas gratuitas e *open-source* para ingestão, enriquecimento, armazenamento, análise e visualização de dados. Os componentes da ELK Stack são:

- **ElasticSearch**: Mecanismo de armazenamento e busca que utiliza uma estrutura de dados baseado em índice invertido, projetada para buscas de texto rápidas;
- **Logstash**: Pipeline para ingestão de dados do lado do servidor. O Logstash pode receber, tratar e enriquecer dados de diversas fontes;
- **Kibana**: Ferramenta da ELK Stack responsável pela visualização e gerenciamento dos dados. Com o Kibana é possível criar histogramas, mapas e diferentes tipos de gráficos com base nos dados armazenados no ElasticSearch;
- **Beats**: Por fim, os Beats são microagentes de coleta de dados que ficam em aplicações clientes. Os Beats podem enviar dados de diversos dispositivos ou sistemas para o Logstash ou ElasticSearch.

A ELK Stack é uma boa opção para a Expertise PMD atingir seu objetivo, pois possui ferramentas especificas para coleta, enriquecimento, ingestão, análise, busca e visualização dos dados. Por sua natureza distribuída, a ELK Stack pode trabalhar em clusters, o que garante a resiliência das informações armazenadas [Kononenko et al. 2014].

Por fim, adotar esta stack tecnologica pode ajudar a Expertise PMD a atingir seu objetivo de oferecer aos seus clientes orientações na tomada de decisão *data-driven* e, se manter como uma das principais empresas de tecnologia do Brasil.

Referências

Chhajed, S. (2015). *Learning ELK stack*. Packt Publishing Ltd.

Frampton, M. (2015). Storing and configuring data with hadoop, yarn, and zookeeper. In *Big Data Made Easy*, pages 11–56. Springer.

Fuad, A., Erwin, A., and Ipung, H. P. (2014). Processing performance on apache pig, apache hive and mysql cluster. In *Proceedings of International Conference on Information, Communication Technology and System (ICTS) 2014*, pages 297–302. IEEE.

Hazarika, A. V., Ram, G. J. S. R., and Jain, E. (2017). Performance comparision of hadoop and spark engine. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 671–674. IEEE.

Kononenko, O., Baysal, O., Holmes, R., and Godfrey, M. W. (2014). Mining modern repositories with elasticsearch. In *Proceedings of the 11th working conference on mining software repositories*, pages 328–331.

Lucene, A. (2010). Apache lucene-overview. *Internet: http://lucene. apache. org/iava/docs/[Jan. 15, 2009]*.

Son, S. J. and Kwon, Y. (2017). Performance of elk stack and commercial system in security log analysis. In *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, pages 187–190. IEEE.