# Ingestão e Coleta de Dados na Alimentos LTDA

**Eduardo Germano**[1]

[1]Especialização em Data Science
UniRitter – Brazil

eduar.germano@gmail.com

A Alimentos LTDA é uma das maiores empresas do ramo alimentício Brasil. A Alimentos LTDA está na busca de agilidade, qualidade e economia nos projetos de inovação, de produtos e processos, em que cada vez mais orçamentos são enxutos, e informações instantâneas, seguras e de qualidade são essenciais para a tomada de decisão. Contudo, a empresa não consegue visualizar cenários prospectivos, tendo em vista a instabilidade econômica e de saúde no mundo. Diante desse cenário, são necessários grande volume, variedade, velocidade e veracidade de dados, que devem ser ingeridos pela organização. Como especialista em arquitetura e modelos de dados da Alimentos LTDA, tenho como responsabilidade auxiliar na construção de uma estratégia que contribua com o objetivo da empresa.

Para resolver este problema, devemos em um primeiro momento, elaborar ações estratégicas para auxiliar o negócio através da ingestão de dados em um *datalake* [Khine and Wang 2018]. Um *datalake* é um repositório de dados centralizado, que permite a ingestão de dados estruturados, semi-estruturados e não estruturados, de qualquer escala [Khine and Wang 2018]. A primeira ação para inserir dados em um *datalake* é mapear as fontes de dados da empresa, além de seus respectivos tipos de dados.

A cada dia surgem novas ferramentas com grande potencial para resolução de problemas em ambientes *big data* e *fast data* [Miloslavskaya and Tolstoy 2016]. Após um tempo entendendo o cenário da Alimentos LTDA, propomos a utilização de uma *stack* que vem ganhando bastante notoriedade nos últimos anos, a ELK Stack [Chhajed 2015]. Esta *stack* nasceu com o 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. Atualmente, a ELK Stack possui componentes que facilitam a tanto a ingestão de dados, quanto o processo de ETL [Vassiliadis et al. 2002], através dos componentes Logstash e Beats.

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 Alimentos LTDA atingir seu objetivo, pois possui ferramentas especificas para coleta, enriquecimento, ingestão, análise, busca e visualização dos dados. Por sua natureza distribuída, a ELK Stack utiliza *sharding* para particionamento horizontal, podendo trabalhar em *clusters*, o que garante a resiliência das informações armazenadas [Kononenko et al. 2014]. Por fim, adotar esta *stack* tecnológica pode ajudar a Alimentos LTDA a atingir seu objetivo de ter dados em real-time para auxiliar na tomada de decisão *data-driven* e, se manter como uma das principais empresas do ramo alimentício nacional.

## Referências


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

Khine, P. P. and Wang, Z. S. (2018). Data lake: a new ideology in big data era. In *ITM web of conferences*, volume 17, page 03025. EDP Sciences.

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]*.

Miloslavskaya, N. and Tolstoy, A. (2016). Big data, fast data and data lake concepts. *Procedia Computer Science*, 88:300–305.

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.

Vassiliadis, P., Simitsis, A., and Skiadopoulos, S. (2002). Conceptual modeling for etl processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, pages 14–21.