O PROCESSO DE EXTRAÇÃO E ANÁLISE DE DADOS DE BIG DATA: UM EXPERIMENTO

Viviane Etges Gomes[1], Daniel Penz,[2] Carlos Roberto De Rolt[3]

[1] Acadêmica do Curso de Administração Empresarial bolsista PROBIC/UDESC
[2] Acadêmico do Curso de Doutorado em Administração - ESAG/UDESC
[3] Orientador, Departamento de Administração Empresarial da ESAG, rolt@udesc.br

Palavras-chave: Extração de dados. Análise de Big data. Twitter

Este paper tem como objetivo apresentar testes de um motor de análise de dados para avaliar a eficiência e praticidade das ferramentas escolhidas (ElasticSearch Logstah e Kibana - ELK), para realizar o processo de seleção, extração e processamento para análise de uma base em dados coletados no Twitter. O estudo é motivado pela necessidade de escolher ferramentas de análise de dados adequadas ao projeto Participact Brasil, onde pesquisas científicas são realizadas sobre problemas urbanos em cidades inteligentes.

O ParticipAcT Brasil é uma extensão da plataforma de crowdsensing ParticipACT da Universidade de Bolonha. Este projeto foi desenvolvido pela Universidade do Estado de Santa Catarina (UDESC) em parceria com a Universidade Federal de Santa Catarina (UFSC) e conta com suporte financeiro e de dados de entidades públicas e privadas. O projeto está sendo desenvolvido na cidade de Florianópolis, capital do estado de Santa Catarina, Brasil. O objetivo do ParticipACT Brasil é usar dados de empresas de serviços públicos e campanhas de crowdsensing para conduzir pesquisas e estudos e, eventualmente, direcionar gerentes para resolver problemas urbanos. Também a criação de um portal aberto de dados, que pode ser usado por qualquer pessoa para melhorar e descobrir mais sobre os hábitos das pessoas em Florianópolis e fazer inferências sobre como podemos viver melhor na sociedade. A infraestrutura do ParticipAcT Brasil, tem três componentes: Big Data, Crowdsensing e Website.

A intensificação da urbanização e a intensa disseminação de tecnologias de comunicação e sensoriamento criam novas oportunidades para a aplicação do conhecimento, a fim de melhorar a qualidade de vida nas cidades. A coleta voluntária e colaborativa de dados usando smartphones de cidadãos em um sistema de crowdsensing associado a uma plataforma de big data caracteriza o ambiente de computação usado neste trabalho. É composto por ferramentas de coleta, catalogação, armazenamento, organização, análise e publicação de dados. Nos últimos meses, estamos tentando usar uma grande quantidade de dados e, durante o processo, enfrentamos alguns desafios relacionados à escalabilidade, indexação, funcionalidade, redução de custos, desempenho e alguns outros, enfrentados por outros pesquisadores Langi et. al (2016) que define (ELK) como um conjunto confiável de ferramentas de armazenamento de dados, mecanismo de busca, análise e visualização. Por outro lado, temos algumas redes sociais em todo o mundo, como o twitter, que desempenham um papel importante na distribuição de informações.

Portanto, o foco deste paper está na etapa de análise de dados da plataforma computacional Participact e envolve a avaliação da viabilidade do uso de ELK, desde sua instalação, facilidade de uso, treinamento de pesquisadores e integração com outras ferramentas por um grupo de pesquisadores da área de ciências sociais. Os testes foram realizados a partir de dados coletados

no Twitter, de uma região e período específicos, a fim de conhecer os assuntos de interesse desses usuários. O foco está em entender como o ELK funciona e não no comportamento dos usuários de redes sociais.

A experiência demonstrou que instalar e usar o ELK requer um monitoramento especializado e um poder computacional não presente na maioria dos desktops. Recomenda-se combinar no laboratório de análise de dados ferramentas complexas e poderosas com outras que tenham características e interfaces mais amigáveis para que a cultura de análise de dados possa ser disseminada, especialmente entre pesquisadores da área de ciências sociais.

O estudo de caso, uma análise dos dados capturados e apontamos alguns desafios encontrados e aprendizado durante o processo de instalação e uso das ferramentas. Instalamos o ecossistema Elasticsearch em uma máquina com sistema operacional Windows 7, 8GB de memória e processador AMD Phenom (tm) II X4 B93. Nos testes, optamos por capturar dados do Twitter de uma região geográfica específica da cidade de Florianópolis. O período de busca foi de 20 a 23 de abril de 2018. Durante esse período, 16.446 tweets foram capturados. Não utilizamos nenhuma palavra-chave, ou seja, fizemos uma busca livre, apenas com restrição territorial. A escolha da área geográfica para a coleta de dados foi aleatória, pois o objetivo era testar a pilha ELK.

Os resultados dos dados coletados e analisados apenas confirmam que obtivemos sucesso na instalação das ferramentas computacionais para coleta, análise de dados e geração de gráficos. Os dados foram classificados para permitir a contagem de assuntos preferidos pelos usuários do Twitter naquele período. Para instalação e uso do ELK, podemos encontrar algumas vantagens como: O ELK é um sistema multiplataforma e pode ser instalado no Windows, Linux e MacOS; existe uma grande quantidade de documentação técnica que explica como instalar e usar as ferramentas, considerando que elas são ferramentas de código aberto; e sem custos de licenças. A instalação das ferramentas requer um bom entendimento em documentos técnicos e habilidades computacionais, o que dificulta o processo de instalação e configuração das ferramentas para pesquisadores com pouca experiência na área computacional. Além de problemas de software e humanos, o ELK consome muita memória que dificulta a instalação em máquinas simples e comuns com 4 GB de RAM. Em ambientes acadêmicos, os instrutores gastam muito tempo preparando as ferramentas e ensinando os alunos a usá-las. Este período inercial só é compensado depois da curva de aprendizado, quando um nível razoável de produtividade é obtido.

Como resultado, podemos verificar que o ELK fornece ferramentas eficientes. Já que o Logstash e o Elasticsearch fornecem pesquisa e análise de dados rápida e de qualidade, o Kibana oferece uma visão robusta com gráficos, mapas e tabelas. Esses recursos encontrados no ecossistema Elasticsearch atendem às expectativas de busca e análise de dados do projeto ParticipACT Brasil, em especial no que diz respeito aos custos de licenciamento. No entanto, um esforço considerável terá que ser feito para preparar material didático e treinamento de estudantes e pesquisadores em estágios mais avançados. O laboratório de análise de dados deve ser pré-preparado e receber cuidados intensivos de atualização tecnológica das ferramentas, aumentar a capacidade de processamento, RAM das estações de trabalho e armazenamento na nuvem.

Referência:

P.P.Langi,Widyawan,W.Najib,andT.B.Aji,“AnevaluationofTwitter river and Logstash performances as elasticsearch inputs for social media analysis of Twitter,” Proceedings of 2015 International Conference on Information and Communication Technology and Systems, ICTS 2015, pp. 181–186, 2016.