Caramel: Um Framework para Ecossistema de Big Social Data

Paulo Freitas Silva Júnior[1], **Tiago França**[2], **Jonice Oliveira**[1]

[1]PPGI – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ, Brasil

[2]DECOMP, – Universidade Federal Rural do Rio de Janeiro
Seropédica, RJ, Brasil

paulofreitas@macae.ufrj.br, tcruz.franca@gmail.com, jonice@dcc.ufrj.br

Resumo. *As pessoas se engajam nas mídias sociais, produzindo e propagando (em alta velocidade) grandes volumes de dados semanticamente ricos em informações. Essas características estão relacionadas a área e as pesquisas de Big Social Data (BSD). Mesmo diante das possibilidades, a extração de informação de trabalhos alinhados com o conceito de BSD está relacionada a retrabalho, falta de recursos técnicos e humanos e pouca colaboração. Esta proposta apresenta um Framework para Ecossistemas de BSD que orienta a criação de artefatos e o compartilhamento desses e de outros recursos úteis para lidar com a coleta, processamento, armazenamento, análise e visualização de dados sociais.*

1. Informações Gerais

- **Nível:** Mestrado
- **Orientando:** Paulo Freitas Silva Júnior
- **Orientadores:** Tiago França e Jonice Oliveira
- **Instituição:** PPGI – Universidade Federal do Rio de Janeiro (UFRJ)
- **Ingresso:** 03/2021
- **Defesa Prevista:** 03/2024
- **Etapas Concluídas:** Todos os créditos concluídos e artigo publicado.
- **Etapas Futuras:** Exame de Qualificação.
- **Publicações:** Silva Júnior, P. F., França, T. C., & Sampaio, J. O. (2023). CARAMEL: Ecosystem for Big Social Data. Proceedings of the XIX Brazilian Symposium on Information Systems, 1(1), 136–142. https://doi.org/10.1145/3592813.3592898

2. Formulação do Problema de Pesquisa

A análise de dados de mídias sociais tem se tornado cada vez mais relevante na compreensão de comportamentos sociais, tendências de mercado e opiniões públicas. São exemplos os estudos para prevenção do suicídio, a identificação de opinião e interesses dos usuários, a influência da mídia social no conteúdo consumido pelos usuários, discursos de ódio, entre outros [Lima Filho et al. 2020] [Rehem et al. 2016] [Hargreaves et al. 2020]. Toda essa interação acaba produzindo um grande volume de dados, os quais são semanticamente ricos em informação, de diferentes tipos e fontes, que precisam ser armazenados,

gerenciados e possivelmente analisados sob diferentes óticas para geração de novos conhecimentos [França et al. 2014] [B. Lima et al. 2022]. Todo esse contexto está relacionado com o conceito de Big Social Data (BSD) que observa a coleta, análise e modelagem de interações e comportamentos sociais a partir de grandes volumes de dados (gerados rapidamente, com alta variedade e grande valor semântico) gerados a partir de interações e ações sociais mediadas por tecnologia [Olshannikova et al. 2017].

Embora existam grandes possibilidades, os principais sistemas de apoio à análise dos dados possuem restrições, especialmente no que diz respeito à coleta e armazenamento de dados. Essas restrições podem impactar negativamente na qualidade da análise e interpretação dos resultados, limitando a eficácia das estratégias utilizadas [França et al. 2014].

Uma das principais limitações está relacionada ao esforço e tempo empregado na coleta (adequadamente) e demais etapas da gestão dos dados. Muitos pesquisadores e analistas, como estatísticos, sociólogos e jornalistas, podem não possuir o conhecimento e as habilidades técnicas necessárias para realizar uma coleta de dados eficiente e representativa. Isso pode inviabilizar as pesquisas ou levar a uma subutilização dos dados disponíveis e a análises superficiais ou até mesmo errôneas [França et al. 2014] [Wang et al. 2021].

Além disso, é comum que as estratégias de coleta de dados não sejam reaproveitadas, resultando em um desperdício de esforços e recursos. Ou seja, cria-se o mecanismo para servir apenas a uma coleta. Muitas vezes, não há um compartilhamento adequado dos mecanismos de coleta entre os pesquisadores, o que também acaba dificultando a replicação e validação dos resultados [Wang et al. 2021].

Outro desafio enfrentado está relacionado à questão da proveniência e qualidade dos dados. Informações importantes, como os filtros utilizados, as restrições aplicadas e outras especificidades da coleta, acabam se perdendo ao longo do processo. Isso pode comprometer a confiabilidade e a validade dos dados para uma análise em uma pesquisa.

Além das dificuldades mencionadas, também se destaca a falta de instrumentos adequados para facilitar a colaboração e o compartilhamento de dados e artefatos entre os pesquisadores [França et al. 2014]. A ausência de uma infraestrutura eficiente para o compartilhamento e reuso dos dados limita a colaboração entre os pesquisadores e dificulta o avanço do conhecimento nessa área.

As restrições relacionadas aos recursos de infraestrutura, abordagens para coleta de dados, compartilhamento de recursos e colaboração estão relacionados ao conceito de ecossistemas de dados, definido por [S. Oliveira et al. 2019] como uma rede complexa que permite a interação entre atores, permitindo a colaboração em um ambiente para localizar, persistir, publicar, consumir, reutilizar e sustentar iniciativas de compartilhamento de dados. Então, a questão de pesquisa desta proposta é esta: quais características são necessárias para um ecossistema de dados sociais e como implementá-los de forma que se observe a colaboração e compartilhamento de recursos entre pesquisadores que trabalham com dados de mídias sociais?

3. Motivação e Justificativa

Esta proposta observa os desafios relacionados à coleta, gestão e análise de dados de BSD em uma perspectiva de software. Atualmente as principais ferramentas para coleta e gestão de dados apresentam limitações em termos de reaproveitamento de estratégias de coletas, compartilhamentos de recursos e bases de dados volumosas, bem como a garantia da proveniência (o processo de obtenção do dado gerado ou coletado) e qualidade dos dados e o processamento distribuído.

As pesquisas (e pesquisadores) que atuam com dados precisam de abordagens escaláveis que possibilitem a colaboração e compartilhamento de artefatos e recursos a fim de aumentar a possibilidade de maior produtividade ao reduzir o retrabalho. Proporcionar melhorias para esse cenário pode levar a comunidade científica e profissional a novas e mais produtivas práticas relacionadas a análise de dados e a colaboração. Possibilitar que as pesquisas com análise dos dados funcionem também de forma distribuída, suportando o reuso das soluções, pode levar a colaboração entre os pesquisadores e otimização do uso de recursos. Uma proposta como estas pode colaborar diretamente com as pesquisas de BSD e indiretamente com a sociedade de forma geral, visto que a área de BSD tem em vista entender e propor soluções para questões sociais.

4. Relevância da Pesquisa

Considerando o volume cada vez maior de dados de mídias sociais disponíveis, é fundamental superar as limitações atuais e fornecer uma infraestrutura tecnológica que permita a coleta, compartilhamento e análise eficazes desses dados. Essa pesquisa é relevante no contexto de banco de dados por ter em vista avançar na área de BSD, proporcionando uma abordagem abrangente para lidar com os desafios relacionados à coleta, análise, armazenamento e compartilhamento de dados de mídias sociais.

A aplicação de técnicas e abordagens tradicionais de banco de dados muitas vezes se mostra inadequada para lidar com a natureza complexa e em constante evolução dos dados de mídias sociais. A escalabilidade, a distribuição, a velocidade de ingestão e a heterogeneidade desses dados exigem soluções inovadoras no âmbito do banco de dados [França et al. 2014] [Laigner et al. 2021].

Ao superar as limitações existentes e propor abordagens mais eficazes para a gestão desses dados, espera-se melhorar significativamente a capacidade de armazenamento, recuperação e processamento de informações relevantes extraídas das mídias sociais.

5. Trabalhos Relacionados

Em [Perikos and Hatzilygeroudis 2018] foi proposto um framework para análise de dados de BSD, entretanto é apresentado um modelo genérico, sempre executando os mesmos processos, de análise de dados baseado em modelagem de tópicos, onde são apresentadas duas estruturas, com a primeira parte realizando à análise dos dados sociais textuais e a outra com ao reconhecimento do seu conteúdo.

[Al-Obeidat et al. 2021] propôs um mecanismo para realização de análises de mídias sociais com uma abordagem de microsserviços em ambientes de nuvem e toda a orquestração dos componentes de software, com considerações significativas para a arquitetura de software, virtualização, mensagens e padrões de computação.

[Laigner et al. 2021] abordou o gerenciamento de dados em arquiteturas de microsserviços focado na necessidade de um controle de transação para ambientes de microsserviços e técnicas de ajuste orientadas a dados para melhorar o desempenho. Examinaram as ações típicas de ajuste de desempenho, discutindo as soluções disponíveis para dar suporte a algumas das atividades principais do processo de ajuste. Também foram analisadas implementações recentes de soluções de ajuste de desempenho baseadas em dados para plataformas de Big Data.

Em comparação com os trabalhos da literatura, que se concentram em mecanismos de persistência de dados e ajuste de desempenho no contexto de análise de dados de mídias sociais e ambientes de Big Data, nossa abordagem tem um enfoque diferenciado na otimização do processo de coleta e análise e compartilhamento de dados em Ecossistemas de BSD.

Nosso trabalho tem o intuito de criar um modelo de arquitetura para Ecossistemas de BSD que considere a construção colaborativa de mecanismos de extração, análise e compartilhamento de dados. Ao enfatizar o compartilhamento e a colaboração, nossa proposta tem em vista superar a falta de reuso de soluções e o compartilhamento limitado de dados presentes nos trabalhos anteriores.

Esses diferenciais tornam a pesquisa inovadora ao abordar a otimização do processo de coleta de dados em um ambiente de BSD e promover a construção colaborativa de mecanismos de extração, análise e compartilhamento de dados. Essa abordagem tem o potencial de melhorar significativamente a eficiência e a eficácia da análise de dados de mídias sociais, fornecendo visões mais relevantes e personalizados para os usuários.

6. O Projeto Caramel

Diante desses desafios, é necessário desenvolver uma framework que aborde as restrições na coleta e análise de dados oriundos de mídias sociais. Esse framework deve fornecer diretrizes, técnicas e instrumentos que permitam uma coleta eficiente, um compartilhamento adequado e uma análise robusta desses dados, contribuindo para a compreensão mais significativa e relevante dos dados.

Ao desenvolver um framework para Ecossistema de BSD, pretende-se superar esses desafios específicos, estabelecendo uma infraestrutura software escalável e distribuída para a gestão eficiente dos dados. Isso envolve a criação de mecanismos de coleta, armazenamento, indexação e consulta que possam lidar com o volume e a variabilidade dos dados de mídias sociais, proporcionando uma base sólida para análises subsequentes.

Além disso, essa pesquisa também tem implicações práticas significativas. Os pesquisadores poderão explorar e analisar de forma mais abrangente e eficiente os dados sociais, gerando novos conhecimentos e visões valiosos. Já as organizações que dependem da análise de dados de mídias sociais poderão aprimorar suas estratégias de políticas públicas, negócios, marketing e tomada de decisões, com base em informações mais precisas e relevantes.

Os principais requisitos levantados para atender as demandas comuns e características de um Ecossistema para BSD são:

- Funcionamento distribuído: permitirá que seja executado de forma distribuída em hardwares distintos e em serviços de nuvem convencionais.

- Padronização na construção de artefatos: fornecerá uma forma padronizada de construção de artefatos computacionais, considerando as diferentes competências e tecnologias utilizadas em ciências de dados, facilitando a colaboração e o compartilhamento entre os membros da comunidade.
- Reutilização e compartilhamento de artefatos: permitirá que os artefatos criados em um projeto de análise de dados sejam reutilizados e compartilhados entre a comunidade, evitando o retrabalho e a redundância de esforços.
- Execução concomitante de coletas e análises: possibilitará que coletas de dados e análises sejam executadas de forma simultânea, reduzindo o tempo necessário para obter resultados e visões.
- Coleta e compartilhamento de bases de dados de mídias sociais: possibilitará que os pesquisadores e profissionais tenham acesso a conjuntos de dados relevantes para suas análises e estudos.
- Suporte a soluções diversas: suportará a implementação de soluções diversas para o tratamento e gestão dos dados, observando os princípios FAIR (Findable, Accessible, Interoperable, Reusable). Isso garantirá que os dados sejam tratados adequadamente, possibilitando a interoperabilidade entre diferentes sistemas e a reutilização dos dados em diferentes contextos.

7. Metodologia

A pesquisa iniciou-se a partir da observação de um grupo de pesquisa multidisciplinar, ao notar o retrabalho e o esforço repetitivo nas atividades relacionadas à coleta e análise de dados de mídias sociais.

Foi realizada uma busca na literatura científica para identificar propostas e soluções existentes para coleta de dados e reuso de artefatos computacionais nesse contexto. Foram analisadas as características de ferramentas amplamente utilizadas, como Gephi, Cytoscape, NodeXL, SocioViz, entre outras, considerando os seguintes aspectos; distribuição, código-fonte, arquitetura, formatos de entrada e saída de dados, análises disponíveis, visualização de informações e integração com mídias sociais.

A partir da revisão da literatura e da análise das ferramentas existentes, foi constatado que não havia uma solução que atendesse ao conjunto de características desejadas para resolver o problema identificado anteriormente. Essa lacuna na literatura e nas ferramentas existentes motivou a proposta deste trabalho de pesquisa.

Com base no levantamento acadêmico e na experiência dos proponentes, foram elucidados os principais requisitos comuns para a solução do problema. Esses requisitos incluíam coleta e análise de dados, reuso de artefatos, compartilhamento de dados e funcionamento distribuído.

Com os requisitos estabelecidos, foi realizado o desenvolvimento do framework proposto, considerando as necessidades identificadas e buscando preencher a lacuna na literatura e nas ferramentas existentes.

8. Resultados Preliminares

Os resultados obtidos nesta pesquisa proporcionaram uma percepção clara dos benefícios que o Ecossistema proposto pode trazer para os pesquisadores que trabalham com dados

de mídias sociais. A principal contribuição do trabalho está relacionada à extração de conhecimento a partir desses dados.

Em termos tecnológicos, o trabalho resultou na implementação de uma arquitetura para a nuvem, onde os coletores de dados de mídia social são disponibilizados como microsserviços conteinerizados. Isso permite a distribuição e escalabilidade desses coletores, facilitando a coleta de dados em larga escala.

Ao comparar as ferramentas analisadas no estudo, foi possível perceber o potencial da solução proposta em termos de compartilhamento de dados entre diferentes Pipelines (fluxos de análises), reutilização dos artefatos criados, integração com diferentes ferramentas de mídias sociais e processamento instantâneo dos dados coletados e gerados. Essas características oferecem maior flexibilidade e eficiência na análise de dados de mídias sociais, possibilitando a criação de fluxos de trabalho mais eficientes e a obtenção de resultados mais rápidos.

Esses resultados demonstram o valor e as vantagens do Ecossistema proposto em relação às ferramentas existentes, fornecendo uma solução abrangente que atende às necessidades dos pesquisadores que lidam com dados de mídias sociais.

Referências

Al-Obeidat, F., Bani-Hani, A., Adedugbe, O., Majdalawieh, M., and Benkhelifa, E. (2021). A microservices persistence technique for cloud-based online social data analysis. *Cluster Computing*, 24(3):2341–2353.

B. Lima, G. d. F., S. Oliveira, M. I., and Farias Lóscio, B. (2022). FASED: A Framework for Data Ecosystems Health Evaluation. *Journal of Information and Data Management*, 13(3).

França, T. C., de Faria, F. F., Rangel, F. M., de Farias, C. M., and Oliveira, J. (2014). *Big Social Data: Princípios sobre Coleta, Tratamento e Análise de Dados Sociais*, volume d.

Hargreaves, E., Mangabeira, E. F., Oliveira, J., Franca, T. C., and Mcnasché, D. S. (2020). Facebook news feed personalization filter: a case study during the brazilian elections. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 615–618, Netherlands. IEEE, IEEE.

Laigner, R., Zhou, Y., Salles, M. A. V., Liu, Y., and Kalinowski, M. (2021). Data management in microservices: State of the practice, challenges, and research directions. *CoRR*, abs/2103.00170:70–75.

Lima Filho, S. P., Oliveira, J., and da Silva, M. F. (2020). Detection of depression symptoms using social media data. *Simpósio Brasileiro de Banco de Dados (SBBD)*, 2020:3–8.

Olshannikova, E., Olsson, T., Huhtamäki, J., and Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, 4(1):3.

Perikos, I. and Hatzilygeroudis, I. (2018). A Framework for Analyzing Big Social Data and Modelling Emotions in Social Media. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 80–84. IEEE.

Rehem, D., Oliveira, J., França, T., Brito, W., and Motta, C. (2016). News recommendation based on tweets for understanding of opinion variation and events. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1182–1185, New York, NY, USA. ACM.

S. Oliveira, M. I., Barros Lima, G. d. F., and Farias Lóscio, B. (2019). Investigations into Data Ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61(2):589–630.

Wang, X., Duan, Q., and Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, 72(9):1161–1182.