Um *Framework* de *Big Data* para Promoção da e-Ciência na Saúde

Adriana Benício Galvão[1]

[1]Laboratório de Inovação Tecnológica em Saúde – Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil

adriana.benicio@lais.huol.ufrn.br

Abstract. *Given the current context of increasing data production and strong potential for value creation in health through the exploration and analysis of data, this paper presents a framework designed with the objective of facilitating the implementation of software for analysis of large data volume and thus promote e-science in health. The general functioning of the framework is presented and technologies for its implementation are mentioned. However, it is worth mentioning that, for wide use of the framework, cooperation and data sharing initiatives between health systems are necessary.*

Resumo. *Diante do contexto atual de crescente produção de dados e forte potencial de criação de valor na saúde por meio da exploração e análise de dados, este artigo apresenta um framework projetado com o objetivo de facilitar a implementação de software para análise de grande volume de dados e, desse modo, promover a e-ciência na saúde. O funcionamento geral do framework é apresentado e são mencionadas tecnologias para sua implementação. Contudo, vale ressaltar que, para ampla utilização do framework, são necessárias iniciativas de cooperação e compartilhamento de dados entre sistemas de saúde.*

1. Introdução

Em meio ao quarto paradigma da ciência (Gray, 2009), que se baseia na computação intensiva de dados, há grandes perspectivas em relação aos benefícios que podem ser conseguidos com a exploração de grandes e heterogêneos conjuntos de dados. Na saúde, muito tem se falado em medicina baseada em evidências, na individualização dos tratamentos médicos, na agregação de diferentes bases de dados para obter diagnósticos mais precisos que os produzidos com dados isolados, em previsões de surtos de doenças, no monitoramento da eficácia de tratamentos e intervenções de saúde, etc. Entretanto, ainda há grandes barreiras a serem enfrentadas para a concretização dessas perspectivas. Para Gray (2009), é necessário produzir melhor as ferramentas para suportar todo o ciclo de pesquisa: desde a captura e o tratamento de dados até a análise e visualização de dados. De fato, nos últimos anos, tem havido um crescente e contínuo avanço no desenvolvimento de ferramentas promissoras. As tecnologias de *big data*

oferecem a possibilidade de trabalhar com um grande volume de dados, que podem ter origem em diversas fontes e estar em uma variedade de formatos, além de agregar maior velocidade de processamento e maior confiabilidade em relação à perda de dados, aumentando o potencial das análises na descoberta de evidências valiosas. Entretanto, o uso efetivo das tecnologias para *big data* na saúde é pouquíssimo significativo. Conforme Mehta & Pandit (2018), não há evidências mínimas sobre como a Análise de *Big Data* pode melhorar a qualidade dos cuidados na saúde e não há estudos de avaliação econômica sobre sua relação custo-benefício. Ademais, existem grandes dificuldades relacionadas à complexidade do uso das tecnologias disponíveis e à escolha do conjunto de ferramentas corretas para as finalidades pretendidas (Raghupathi & Raghupathi, 2014).

É razoável afirmar que *big data* têm grande papel na evolução da e-Ciência, uma vez que grande parte das intenções da e-Ciência pode ser implementada tecnicamente por tecnologias de *big data*, como o compartilhamento de dados utilizados e gerados por diferentes pesquisadores. Diante desse contexto, o objetivo deste artigo é apresentar um *framework* de experimentação para e-Ciência suportado por tecnologias de *big data,* com potencial contribuição para o surgimento de estudos práticos e quantitativos na saúde pública. A proposta do *framework* é permitir a colaboração no desenvolvimento de soluções em saúde entre epidemiologistas, praticantes da área e estatísticos, por exemplo, ao passo que abstrai complexidades relacionadas a competências técnicas específicas, como ao processamento do grande volume de dados. Desse modo, analistas podem realizar suas análises com foco nas questões de interesse, aplicando suas técnicas de domínio. Além de tudo, o *framework* viabiliza o rápido início no desenvolvimento de programas experimentais para análise de *big data* na saúde e permite incrementos progressivos de suas funcionalidades.

2. Trabalhos Relacionados

Alguns trabalhos buscaram elaborar ou utilizar *frameworks* arquiteturais de *big data* para aplicar em seus estudos. Raghupathi & Raghupathi (2014) esboçou um *framework* pioneiro no contexto de *big data* para o domínio da saúde, que consiste nas camadas de fonte de dados, transformação, plataforma e aplicações analíticas. Wang & Hajli (2017) encontrou que os recursos da análise de *big data*, com potenciais benefícios na indústria da saúde, são principalmente acionados por um componente de processamento de dados, seguido por uma agregação de dados e um componente de visualização de dados. De modo geral, os trabalhos mencionados identificaram os distintos e relevantes componentes de uma arquitetura de *big data*. O presente *framework* considera no seu projeto tais componentes consolidados na literatura. No entanto, acrescenta uma abordagem de implementação para os componentes de aplicações analíticas e processamento do grande volume de dados, no intuito de atenuar problemas relacionados ao fato de que, diferente das ferramentas tradicionais de análise de saúde, as ferramentas de análise de *big data* são complexas, requerem muita programação e exigem a aplicação de várias habilidades. Elas surgiram de maneira *ad hoc,* principalmente, como ferramentas e plataformas de desenvolvimento de código aberto e, portanto, não têm o suporte e a facilidade de uso das ferramentas proprietárias (Raghupathi & Raghupathi, 2014).

3. O *framework*

O *framework* de experimentação para e-Ciência na saúde foi idealizado tendo em vista a facilitação da análise de *big data* em estudos da saúde. É composto por três módulos principais (figura 1): Módulo de *Big Data*, Módulo de Processamento de Dados e Módulo de Transferência de Recursos. O Módulo de *Big Data* (MBD) lida diretamente com os dados em uma variedade de formatos, provenientes de distintos Sistemas de Informação em Saúde. Esses sistemas podem ter, por exemplo, dados biométricos, imagens médicas, registros de sinais vitais, prontuários digitalizados, etc. Como exemplos de sistemas de base nacional, podem ser citados: SIA (Sistema de Informações Ambulatoriais), GAL (Sistema de Gestão de Ambiente Laboratorial), Hórus (Sistema Nacional de Assistência Farmacêutica), SINAN (Sistema de Informação de Agravos de Notificação), entre outros. O objetivo do MBD é permitir o processamento do grande volume de dados em tempo viável, paralelizando e distribuindo o processamento entre vários servidores. Associada a ela, estão as ferramentas de monitoramento e gerenciamento de recursos computacionais e de ingestão dos dados. De modo geral, esta camada é composta por tecnologias do ecossistema Hadoop, como Hadoop-Yarn, HBase, Storm, Hive, Spark, Ambari, etc.

No Módulo de Transferências de Recursos (MTR), é implementada a *Application Programming Interface (*API) do *framework,* que pode ser utilizada por aplicações clientes por meio de REST (*Representational State Transfer).* A API permite obter metadados dos dados na Camada de *Big Data*; de modo que usuários possam selecionar os dados com que desejam trabalhar e aplicar as funções desejadas. Este módulo possui uma base de dados que contém os resultados dos processamentos finalizados, possibilitando que o serviço REST do *framework* retorne os mesmos resultados para requisições com parâmetros de entrada iguais. Isso permite, junto à fila de requisições, que as requisições de processamento feitas ao *framework* não fiquem necessariamente aguardando resposta; elas podem ser consultadas posteriormente, sem desencadear um novo processamento. Os resultados gerados em estudos pioneiros podem estar disponíveis para outros estudos no próprio banco de dados do MTR ou podem retroalimentar o sistema, ou seja, podem ser carregados de volta para o Módulo de *Big Data*.

O Módulo de Processamento de Dados (MPD) submete funções ao Módulo de *Big Data*. Um determinado grupo de funções do Módulo de Processamento de Dados compõe a Interface de *Big Data*, que diz respeito às funções específicas que conectam a linguagem de programação e a tecnologia do ecossistema Hadoop utilizadas. O *framework* recebe suas funções por meio de *plugins*, adicionados por diferentes analistas com objetivos específicos, que podem auxiliar, por exemplo, na descoberta de padrões de comportamento epidemiológicos, correlações, periodicidades, intervenções não programadas e seus efeitos, em novas formas de ação preventiva de doenças, entre outras questões. Uma versão do MPD está sendo desenvolvida no Laboratório de Inovação Tecnológica em Saúde (LAIS) da Universidade Federal do Rio Grande do Norte (UFRN), como parte do projeto Sífilis Não. A versão contará com funções para processamento de grandes séries temporais, as quais serão aplicadas no estudo de séries temporais correspondentes às notificações dos casos de Sífilis, registradas no Sistema de

Informação de Agravos de Notificação (SINAN) durante um período específico, com cruzamento de dados do Sistema de Informações Ambulatoriais do SUS (SIA).

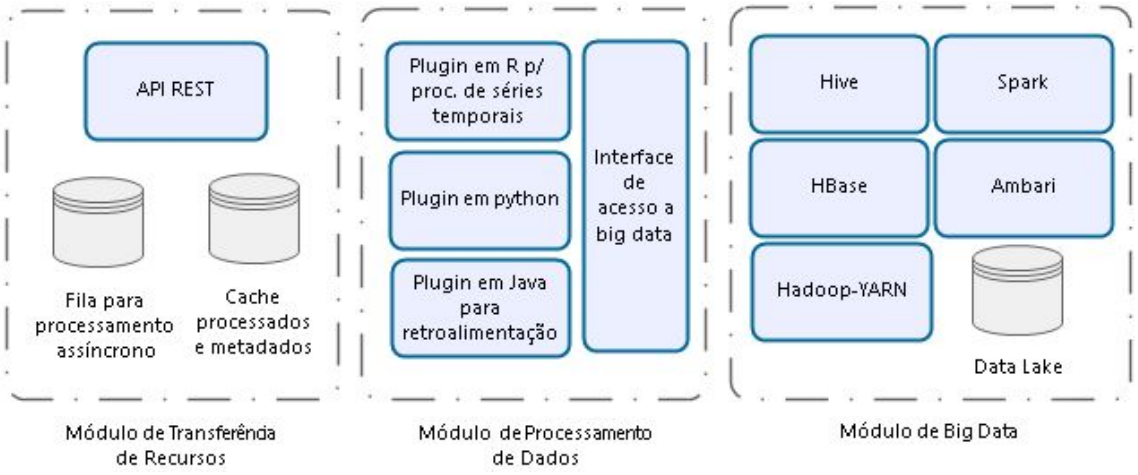

Figure 1. Representação do *framework* experimental para e-Ciência com exemplos do uso de tecnologias de *big data* e plugins.

4. Conclusões

O *framework* apresentado, de forma geral, propõe a interação com *big data* de forma transparente para as aplicações clientes e uma sistematização do uso de tecnologias para análise de grande volume de dados, como forma de facilitar e, consequentemente, promover a e-Ciência na área da saúde, além de contribuir para uma adoção mais ampla de *big data* na área. Foi planejado para atender características da e-Ciência, como compartilhamento de dados diversos e reuso. Além disso, é implementado como um serviço *online* que permite selecionar dados e funções e iniciar processamentos sob demanda e de forma escalável, de modo a viabilizar análises para atender necessidades de negócio, assim como avaliações técnicas sobre a execução de determinados métodos muito utilizados em estudos de saúde, considerando um ambiente de grande volume de dados. Vale destacar que, para a ampla utilização de *frameworks* como o apresentado neste trabalho, é necessário vencer obstáculos relacionados ao compartilhamento de dados críticos, à cooperação entre os diversos sistemas de saúde pública, entre outros.

Referências

Gray, J. (2009). Jim Gray on eScience: A Transformed Scientific Method. Microsoft Research. <http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf>

Mehta, Nishita & Anil Pandit (2018). Concurrence of big data analytics and healthcare: A systematic review. International Journal of Medical Informatics, 57-65.

Raghupathi, Wullianallur & Viju Raghupathi (2014). Big data analytics in healthcare: promise and potential, Health Information Science and Systems.

Wang, Yichuan & Nick Hajli (2017). Exploring the path to big data analytics success in healthcare, Journal of Business Research.