Usando Workflows Datacêntricos Para Analisar Tweets Sobre o *Aedes Aegypit*

Fillipe Dornelas[1,3], Sérgio Manuel Serra da Cruz[1,2]

[1] Departamento de Matemática –Universidade Federal Rural do Rio de Janeiro (UFRRJ)
[2] Programa PET Sistemas de Informação (PET-SI/UFRRJ)
BR 465, KM7 – UFRRJ – Seropédica – RJ – Brasil

[3] IBM Research do Brasil
Avenida Pasteur, 138 - Urca - Rio de Janeiro – RJ - Brasil

dornelasfillipe@gmail.com, serra@pet-si.ufrrj.br

Abstract. *Analyzing user messages in social media can offer different point of view of a given society, including public health issues. This work presents a strategy based on data-centric workflows able to collect, prepare and analyze large amounts of tweets evaluating the impact of the messages about the Aedes aegypti in the Brazilian public health scenario. Static and temporal analysis were performed by data-centric workflows enacted in Bluemix platform which has been shown as stable and scalable platform.*

Resumo. *A análise de mensagens de redes sociais pode oferecer diferentes perspectivas sobre como as populações se relacionam, incluindo áreas da saúde pública. Este trabalho apresenta um estudo inicial baseado no uso de workflows do tipo datacêntricos executados em nuvens de computadores capazes de coletar e preparar os dados e analisar tweets, avaliando o impacto das postagens acerca do mosquito Aedes aegypti no cenário de saúde pública brasileira. As análises ora apresentadas são de natureza estáticas e temporais e foram integralmente efetuadas na plataforma Bluemix.*

1. Introdução

Compreender com profundidade comportamentos e assuntos relacionados ao espalhamento dos *tweets* sobre a saúde pública ainda é um grande desafio em aberto na computação. Adicionalmente, se considerarmos as enormes quantidades de dados manipulados na área da saúde pública, a gravidade e a abrangência das doenças transmitidas pelo mosquito *Aedes aegypti* no Brasil, este problema se torna ainda mais crítico.

Vários estudos têm como objetivo avaliar o espalhamento das mensagens em redes sociais sobre eventos sociais, catástrofes e epidemias (Sprenger et al, 2013), (Dalmonte et al, 2014) e (Santos et al, 2015). O microblog Twitter é uma das redes sociais mais utilizadas no mundo e o Brasil ocupa a segunda posição entre os países com maior número de usuários. A rede emergiu como um dos meios de propagação mais profícuos de disseminação de informações sobre eventos sociais, catástrofes e epidemias (Chew e Eysenbach, 2010 e Kwak et al, 2010). Seu limite de postagem de poucos caracteres é

um facilitador para que os usuários, agências governamentais ou do terceiro setor realizem postagens de forma rápida e sucinta e que se tornam uma fonte importante de alertas de situações de emergências.

Nos últimos anos a disseminação do mosquito *Aedes aegypti* tem alcançados proporções alarmantes em diversas partes do Brasil e no mundo, favorecendo a disseminação de doenças virais até pouco tempo negligenciadas (por exemplo, Zika e Chikungunya). Neste trabalho apresentaremos um estudo baseado em workflows datacêntricos executados em nuvem de computadores que permitem analisar grandes volumes de tweets relacionados com as postagens relacionadas ao tema "*Aedes aegypti"*, utilizamos dados do Twitter coletados por um período de 6 meses em todo o Brasil.

Diferentemente dos trabalhos relacionados, concebemos uma estratégia baseada em workflows datacêntricos em um ambiente de nuvem do tipo PaaS cuja composição envolve atividades que variam desde a automação da coleta dos tweets, processamento e posterior análise/visualização e disseminação das mensagens. A estratégia foi materializada plataforma Bluemix (Kim et al, 2016), ela permite analisar questões relacionadas com o espalhamento de mensagens sobre *Aedes aegypti*. Para avaliar a abordagem, propomos um conjunto de questões (Q1, Q2, Q3 e Q4) para investigar o espalhamento dos tweets e testar a viabilidade da abordagem.

Este trabalho está organizado da seguinte forma. Na Seção 2 apresentamos uma visão geral da literatura relacionada sobre estudos de comportamento de usuários no Twitter em relação as doenças transmitidas pelo *Aedes aegypti*. Na Seção 3 descrevemos os materiais e a metodologia utilizada nos workflows centrados em dados, além da caracterização do dataset e as análises realizadas, Na Seção 4, discutimos os principais resultados obtidos. Finalmente, na Seção 5 apresentamos as conclusões, limitações e alguns direcionamentos para trabalhos futuros.

2. Trabalhos Relacionados

O Twitter tem sido usado em diversos contextos, possui canais de cidadania, saúde e emergências sociais que têm despertado grande importância no cenário de análise de dados sociais. Um dos usos do Twitter que mais vem despertando atenção diz respeito às questões ligadas à saúde pública, em especial aquelas relacionados com as doenças transmitidas por vírus (H1N1, SARS, Dengue, Zika, Malária, entre outros) que podem atingir grandes contingentes populacionais (Van Hilten et al, 2016).

Antunes et al. (2014) usaram os dados de tweets com a ocorrência do termo "dengue" para inferir quais os períodos onde mais se comentava sobre este assunto e onde mais se encontravam registros de casos da doença em uma determinada região analisada. Toriumi et al. (2013) usaram os tweets para elaborar mapas de projeção e abrangência de um determinado assunto, os autores desenvolveram uma aplicação que exibe mapas e informações sobre a provável infestação dos mosquitos *Aedes aegypti* no município de Cuiabá, no Mato Grosso. A partir da seleção e análise desses dados, os autores foram capazes de desenvolver uma ferramenta de fácil visualização e entendimento sobre possíveis infestações e disseminações das doenças virais.

Além dos estudos sobre a disseminação de epidemias, o Twitter também é largamente utilizado em desastres naturais. Por exemplo, existem trabalhos construídos com a perspectiva de analisar como a informação se propagada durante a ocorrência de

desastres naturais (Toriumi et al, 2013 e Thapa et al, 2016). O primeiro autor usou tweets para estudar como se comportava o compartilhamento das informações durante o terremoto no Leste do Japão de 2013. O segundo analisou o espalhamento de dados das redes sociais Twitter e Flicker relacionadas ao terremoto no Nepal de 2015. Apesar de serem trabalhos independentes, os autores concluíram que os usuários compartilharam tweets colaborativamente para disseminar as informações que consideram importantes acerca do desastre e também diminuíram o compartilhamento de informações não emergenciais para evitar interromper os fluxos das informações críticas.

Como relação a geolocalização dos tweets, verifica-se que grande parte destes não são localizados por opção própria de usuários ou por questões de privacidade; a maioria evita informar suas reais localizações. Segundo Leetaru et al., (2013) apenas 2% das mensagens são geolocalizadas. Com vistas a preencher essa lacuna, Davis Jr. et al., (2011) usaram dados de tweets não geolocalizados e de informações de relacionamentos entre os usuários do Twitter para enriquecer a tentativa de inferir a localização desses tweets a partir da técnica de validação cruzada de informações.

Até o momento, existem alguns trabalhos na literatura que associem o problema de extração e analise de tweets com uso de workflows científicos. Um workflow científico pode ser definido como sendo especificação formal de um processo científico que representa o encadeamento de fluxos de atividades e dados a serem conduzidas em um determinado experimento (Deelman et al, 2009.). Eles são executados por sistemas gerenciadores de workflows científicos (SGWfC) que fornecem o ferramental necessário para definir, modificar, gerenciar, executar e monitorar os workflows científicos. Os workflows do tipo datacêntricos seguem a mesma lógica dos workflows científicos tradicionais, são centrados em grandes volumes de dados complexos e podem ser executados por SGWfC ou não.

Um SGWfC é um sistema computacional que executa aplicações científicas compostas por atividades cuja ordem de execução é definida por uma representação digital da lógica do workflow científico (Goderis et al, 2006). Atualmente, existem dezenas de SGWfC (Kepler, VisTrails, Pegasus, Taverna, Panda, Galaxy, Swift, Knime, entre outros) (Deelman et al, 2009). Os SGWfC são produtos de diferentes motivações de desenvolvimento, públicos-alvo específicos e decisões técnicas particulares a cada projeto, o que faz com que suas funcionalidades se diferenciem consideravelmente um do outro e que representem diferentes aspectos relacionados à execução e à modelagem de workflows científicos.

Faz necessário ressaltar que até o momento da escrita deste trabalho não foram localizados na literatura SGWfC especificamente concebidos para modelar problemas comuns à área de análise de dados de redes sociais. Por esse motivo, investigamos uso de novas ferramentas de *data analytics* tais como plataforma Bluemix da IBM para modelar e executar os workflows datacêntricos.

O Bluemix da IBM é uma plataforma de serviços de nuvem (PaaS) elástica e escalável baseada no projeto de código aberto Cloud Foundry. Ela permite criar, implementar e gerenciar aplicativos na nuvem. O Bluemix é uma plataforma comercial que não foi concebida para atuar ou incorporar as funcionalidades de um SGWfC, porém ele oferece um ecossistema aplicativos, componentes e serviços em tempo de execução que permitem que um pesquisador encadeie atividades computacionais de modo análogo a um workflow científico. O encadeamento das atividades se dá por intermédio de

editores de workflows (nesta pesquisa utilizamos o editor Node-RED. O Node-RED é editor de workflows multiplataforma que possui interfaces ricas baseadas em Javascript e Node.js e que permite ao pesquisador modelar, executar, e monitorar a execução dos workflows datacêntricos que analisam os dados semiestruturados de oriundos de redes sociais.

Diferentemente dos trabalhos principais relacionados na literatura, neste trabalho propomos a adoção do paradigma dos workflows datacêntricos em ambientes elásticos de *data analytics* para analisar os tweets relacionados à disseminação do mosquito *Aedes Aegypti*. As análises dos tweets sobre o tema serão realizadas por workflows desenvolvidos e executados em uma plataforma genérica de serviços de computação em nuvem baseada em um projeto de código aberto.

3. Materiais e Métodos

Esta seção descreve os materiais, métodos e etapas propostas para a extração, análise, processamento e visualização dos dados do microblog Twitter na plataforma de serviços de nuvem.

3.1 Métodos

Neste trabalho propomos uma abordagem metodológica baseada em quatro fases para analisar os tweets. A representação gráfica das fases está ilustrada na Figura 1, elas foram utilizadas na modelagem do workflow datacêntrico.

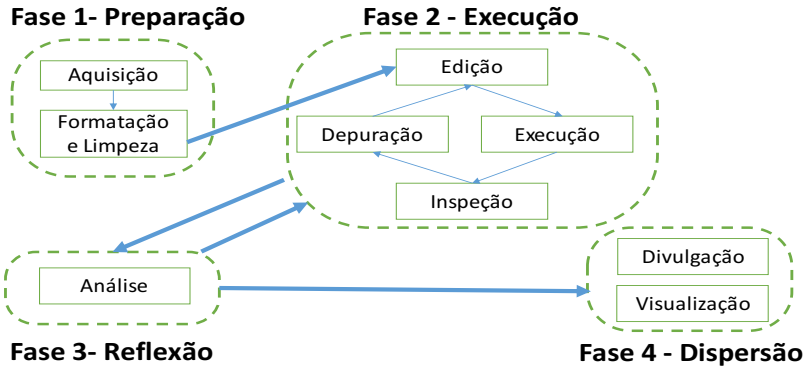

Figura 1. Representação simplificada e conceitual das fases de um workflow datacêntrico para análise de tweets.

A primeira fase (denominada preparação) é executada antes de qualquer tipo de processamento analítico, nela ocorrem a aquisição dos tweets e a preparação ou formatação/limpeza dos dados para serem analisados.

A segunda fase (denominada execução) é o elemento central no workflow datacêntrico. Nela, ocorrem a edição/codificação/encadeamento/execução de scripts. Além disso, ocorrem as análises parciais dos resultados intermediários do experimento, bem como a depuração dos scripts. Essa fase pode ser encarada com um laço, onde o pesquisador interage com a plataforma, realiza múltiplas execuções do workflow com parâmetros distintos para explorar as hipóteses do modelo computacional.

A terceira fase (denominada reflexão) é a eminentemente analítica (ou pós-execução) no processo de exploração dos tweets. Comumente, o pesquisador oscila entre as fases de

reflexão e execução até a finalização do seu experimento. Nesta fase ele analisa os resultados, inspeciona arquivos, faz anotações e comparações entre as múltiplas execuções dos workflows.

Por fim, a quarta fase (denominada dispersão) diz respeito a divulgação, visualização ou compartilhamento dos resultados consolidados obtidos na pesquisa. Nesta fase, ocorrem a publicações dos dados e resultados bem como dos workflows subjacentes.

3.2 Materiais

Nosso estudo se considerou apenas o termo "Aedes aegypti", não foram consideradas variações termo. A coleta dos dados levou em consideração os todos tweets postados por usuários de todo o mundo no período que variou entre junho de 2014 até junho de 2016. Durante esse intervalo, coletamos um total de 44.467 tweets.

Utilizamos a plataforma Bluemix e as ferramentas de Data&Analytics disponíveis no catálogo de serviços da plataforma para o desenvolvimento e execução dos workflows. Dentre as principais ferramentas utilizadas destacamos: API de extração e recuperação de dados do Twitter. O repositório de dados foi o dashDB. O dashDB oferece serviços de banco de dados SQL totalmente gerenciados para cargas de trabalho transacionais e de data warehousing, ele foi utilizado como área de armazenamento temporário dos dados (consumidos e produzidos) pelo workflow. Além disso, utilizamos a plataforma para executar as análises estatísticas sobre os tweets, provendo resultados textuais e visualizações gráficas dos resultados produzidos pelos workflows. A plataforma Bluemix e o dashDB foram instanciados em máquinas virtuais com 64GB de memória RAM com 20GB de espaço de armazenamento oferecidas pelo serviço de virtualização OpenStack.

3.3 Questões de pesquisa

Neste estudo se buscou investigar um pequeno conjunto de questões (Q1, Q2, Q3 e Q4) para verificar o espalhamento de tweets e testar a viabilidade da abordagem baseada em workflows datacêntricos na plataforma Bluemix. As questões de pesquisa são experimentos baseados no workflow (Figura 2). Q1: Qual foi a contribuição do tweets em termos de quantos foram os usuários mais influenciadores? Q2: Quais os períodos de maior postagem de tweets sobre o termo "Aedes aegypti"? Q3: Quais as hashtags foram mais postadas no período avaliado? Q4: Qual a predominância dos sentimentos dos tweets?

3.4 Representação conceitual do workflow

Para analisar os dados e verificar a abrangência dos tweets sobre o "Aedes aegypti", desenvolvemos um workflow datacêntrico composto por quatro fases apresentadas na subseção 3.1. A figura 2 ilustra uma representação conceitual simplificada das fases e recursos utilizados para a execução dos experimentos.

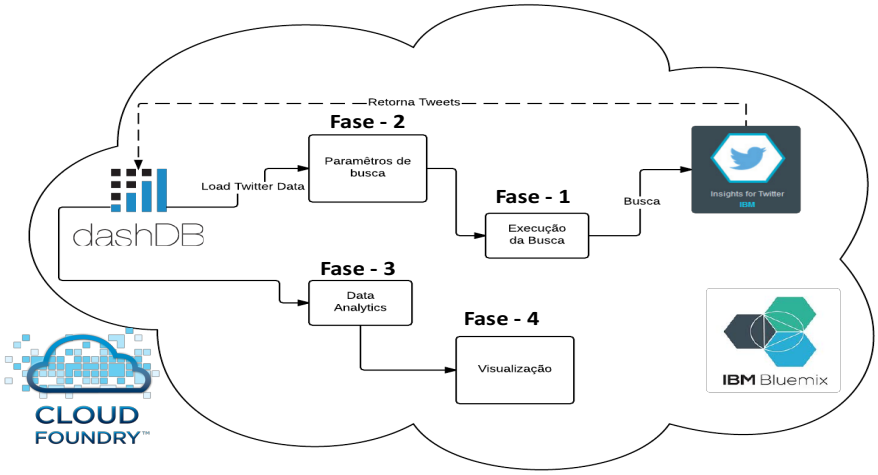

Figura 2. Representação conceitual de um workflow datacêntrico na plataforma Bluemix.

4. Prova de Conceito

A fim de avaliar as funcionalidades do workflow foi realizada uma extração de tweets entre os meses supracitados na subseção 3.2. Foram executados experimentos como prova de conceito com o workflow parametrizável no Bluemix. Os experimentos buscavam responder as questões Q1, Q2, Q3 e Q4.

Para responder a Q1, realizamos uma execução do workflow que produziu uma simples avaliação do tipo estatística. Dentre todos os tweets da base experimental, verificou-se que existiam apenas 25.370 usuários influenciadores que postaram tweets com o termo avaliado. A abrangência desses alcançaram 255.465.058 milhões seguidores. Como decorrência de Q1, refinamos as análises dos tweets da base experimental que explicitavam o termo "Aedes aegypti" e verificamos que apenas 1,83% (818 mensagens) possuíam informações de geolocalização.

A Figura 3 apresenta a distribuição dos tweets geolocalizados coletados e avaliados pelo workflow. Também se verificou que 43.649 (entre Tweets e Retweets) não são geolocalizados, sendo que 611 são confirmados do Brasil. Os resultados estão alinhados com as estimativas de geolocalização de tweets apresentada por (Leetaru et al, 2011).

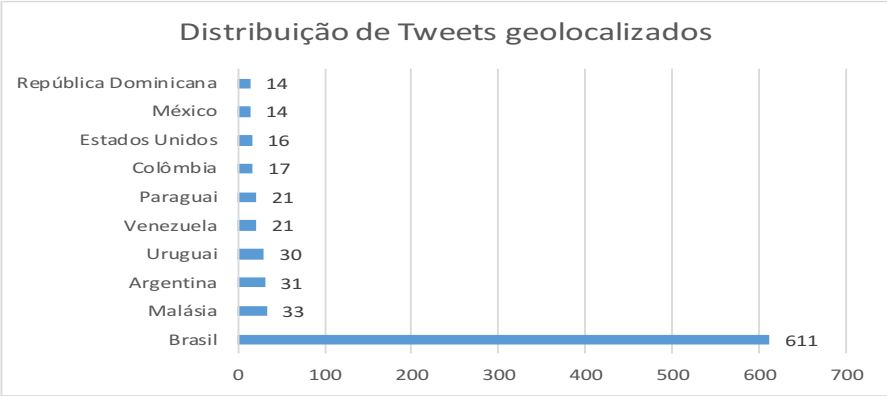

Figura 3. Distribuição de tweets por pais origem.

Para responder a Q2, realizamos uma segunda execução do workflow, dessa vez configurado para avaliar a frequência de mensagens. Obtivemos os resultados apresentadas da Figura 4.

Ressaltamos que a questão Q2 difere de Q1, a primeira apresenta apenas resultados estatísticos. Q2 analisa as mensagens em função da sua distribuição temporal e representa os quantitativos de tweets ao longo do período de tempo do estudo.

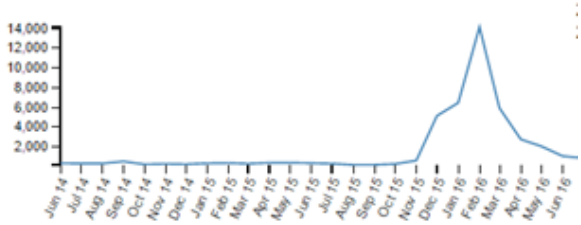

Figura 4. Distribuição temporal de tweets.

Verifica-se que ocorreu um aumento subido de mensagens sobre o tema avaliado nos períodos de novembro de 2015 até março de 2016. Estes períodos correspondem aos meses de verão no Brasil, onde ocorre um aumento natural dos casos das doenças transmitidas pelos mosquitos.

Para responder a Q3, houve uma terceira rodada do workflow, onde ele foi parametrizado para analisar a frequência e os períodos de postagens dos usuários sobre o tema. Os resultados gerados pelo workflow tiveram como saída a Tabela 1.

Tabela 1. Resumo do quantitativo das hashtags postadas no período avaliado.

Hashtag	Número total de ocorrências
#Zika	374
#ZikaZero	20
#G1	18
#Dengue	13
#CombateAedes	9
Outras	23.783

Para responder a Q4, houve uma última rodada do workflow no Bluemix, ele foi parametrizado para analisar os sentimentos das mensagens utilizando os algoritmos disponíveis na plataforma. Do total de mensagens originais, 42.697 não possui informações de sentimentos. Apenas 3,98% possuíam indicações, sendo 863 identificados como positivas e 575 com sentimentos negativos.

5. Conclusão

O Twitter se mostra como um importante meio de comunicação em saúde pública. Neste trabalho desenvolvemos uma estratégia computacional baseada em workflows datacêntricos apoiados por uma PaaS comercial de data analytics para analisar o espalhamento de tweets relacionados com ao tema “Aedes aegypti”.

Verificou-se que a plataforma ainda é pouco conhecida pela comunidade científica, porém ofereceu um amplo suporte para o desenvolvimento do workflow e condução dos

experimentos. Ela permitiu responder as questões Q1, Q2, Q3 e Q4 com agilidade. Destacamos que, apesar de não ser o foco deste trabalho, avaliar o desempenho do workflow datacêntrico (mapeado através das quatro fases descritas na seção 3.1) informamos que ele produziu os resultados em tempo muito curto, aproximadamente cinco minutos para todas as execuções.

A abordagem baseada em workflow datacêntrico no Bluemix permitiu que se verificasse que o espalhamento dos tweets avaliados alcançaram milhões de retweets. Observou-se que os períodos de maior número de postagens coincidem com os momentos de maior enfoque do tema nas mídias (rádio, TV e Internet) e nas campanhas publicitárias que alertavam sobre os perigos e formas de prevenção das doenças relacionadas ao mosquito *Aedes aegypti*.

Como limitações encontramos dificuldades ao analisar como os tweets não geolocalizados. Como trabalhos futuros existem diversas possibilidades oferecidas pela plataforma e que por questões de escopo não foram exploradas neste trabalho, como por exemplo aprofundar as análises de sentimentos dos tweets sobre o tema e produzir visualizações dos dados.

Agradecimentos

Agradecemos ao FNDE e ao MEC/SeSU pelo financiamento concedido ao programa PET SI/UFRRJ e a IBM Research do Brasil pelo acesso gratuito ao Bluemix e seus recursos computacionais.

Referências

Antunes M. N, et al. 2014. "Monitoramento de informação em mídias sociais: o e-Monitor Dengue", In: TransInformação, Campinas, 26(1):9-18, jan./abr., Brasil.

Cloud Foundry, 2016. https://www.cloudfoundry.org/

Chew C, Eysenbach G. 2010. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One. 2010 Nov 29;5(11):e14118.

Dalmonte, E. 2014. Novos cenários comunicacionais no contexto das mídias interativas: o espalhamento midiático. *Revista Famecos*. DOI: http://dx.doi.org/10.15448/1980-3729.2015.2.19729.

DashDB. 2016. http://www.ibm.com/analytics/us/en/technology/cloud-data-services/dashdb/

Davis Jr, C. A. et al 2011. Inferring the Location of Twitter Messages Based on User Relationships", In: Transactions in GIS, Blackwell Publishing Ltd.

Deelman, E et al. 2009 Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems 25 (5), 528-540.

Goderis, A., Li, P., e Goble, C. 2006. Workflow discovery: the problem, a case study from e-science and a graph-based solution. In Int. Conf. on Web Services (ICWS), pp. 312–319.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is twitter, a Social Network or a News Media? In Proceedings of the 19th Int Conf. on World Wide Web, pp. 591–600.

Node-Red. 2016. A visual tool for wiring the Internet-of-Things. http://nodered.org/.

Santos, H.S et al. 2015. Uma Visão do Mercado Brasileiro de Ações a partir de Dados do Twitter, In: IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015), Brasil.

Thapa, L. 2016. Spatial-Temporal Analysis Of Social Media Data Related To Nepal Earthquake 2015. XXIII ISPRS Congress, July 2016, Prague, Czech Republic, pp. 567-571.

Toriumi, F. et al., 2013. Information Sharing on Twitter during the 2011 Catastrophic Earthquake" In: Proc. 22nd Int'l Conf. on World Wide Web, pp. 1025–1028.

Van Hilten, L. G. 2016. Debunking Zika virus pseudoscience: we need to respond fast, say researchers https://www.elsevier.com/connect/debunking-zika-virus-pseudoscience-we-need-to-respond-fast-say-researchers.